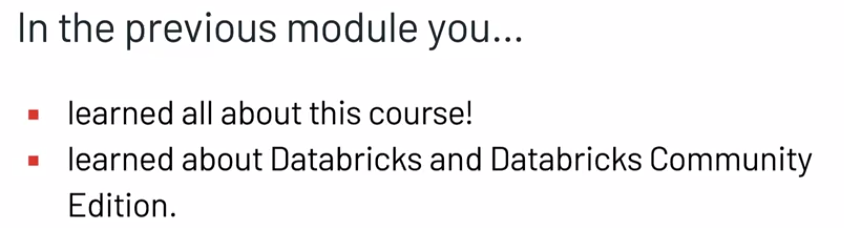
**An Introduction to Data Science**

**Learning Objectives**

* Design a data science project using the information presented in this course about the scientific method [Bilimsel yöntem hakkında bu derste sunulan bilgileri kullanarak bir veri bilimi projesi tasarlayın]
* Explain the expectations behind the Data Science Process Activity [Veri Bilimi Süreç Etkinliğinin ardındaki beklentileri açıklayın]
* Examine given examples of real-world data science applications [Gerçek dünyadaki veri bilimi uygulamalarının verilen örneklerini inceleyin]
* Formulate a definition for data science using concepts presented in the Data Science Fundamentals course [Veri Biliminin Temelleri kursunda sunulan kavramları kullanarak veri bilimi için bir tanım formüle edin]
* Describe the role of domain knowledge within the applied scientific process [Uygulamalı bilimsel süreç içinde alan bilgisinin rolünü açıklayın]
* Differentiate between the fields of applied statistics and computer science [Uygulamalı istatistik ve bilgisayar bilimi alanlarını ayırt edin]
* Describe the skills necessary to complete the applied scientific process at scale [Uygulanan bilimsel süreci uygun ölçekte tamamlamak için gerekli becerileri tanımlayın]
* Describe the role that data plays in the scientific method [Bilimsel yöntemde verilerin oynadığı rolü açıklayın]
* Explain each of the five steps in the scientific method [Bilimsel yöntemdeki beş adımın her birini açıklayın]

# **Module and Lesson Introduction**

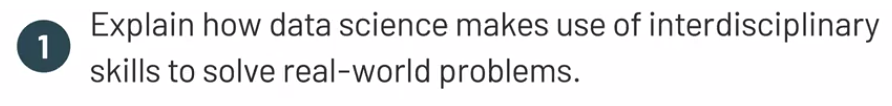
Hello and welcome to module two of data science fundamentals for data analysts. My name is Astrid and if we didn't meet before in the first course in the specialization, nice to meet you. I work on the curriculum team here in Databricks now, every time you see me, it means that you're getting ready to start a brand new module in this course, and I'll help you get familiar with some of the things that you'll be doing in that module. In the previous module, you learned all about this course.



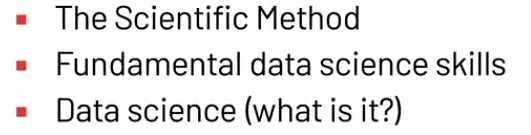
You also have the option of learning more about Databricks and Databricks Community Edition. And that's the free tool that you use to complete the hands on exercises in this course. So if you're not familiar with using database community edition, it's probably a good idea that you go back and review those materials now before moving forward. Next, we're going to start talking about data science.



In this module, an introduction to data science, we have a single lesson with just one objective, and that is that by the end of this module you'll be able to explain how data science makes use of interdisciplinary skills to solve real world problems.



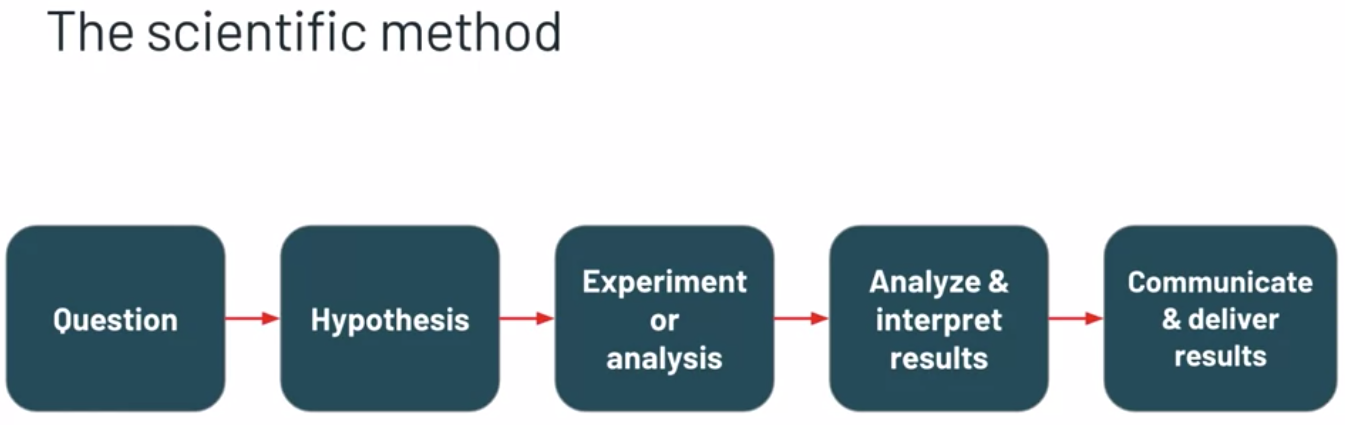
Along the way, we'll explore the scientific method, the fundamental skills of data science and data science itself.



You'll build a better understanding of what data science is and what skills are necessary to perform modern an effective data science. All of these concepts will be taught through a series of videos, knowledge checks and a a peer-reviewed activity at the end of the module. In just so you're aware, each of the videos will be focused on teaching you new concepts and the activities, that's where you'll be able to apply what you've learned. So when you are ready click on to the next video where my colleague is waiting for you.

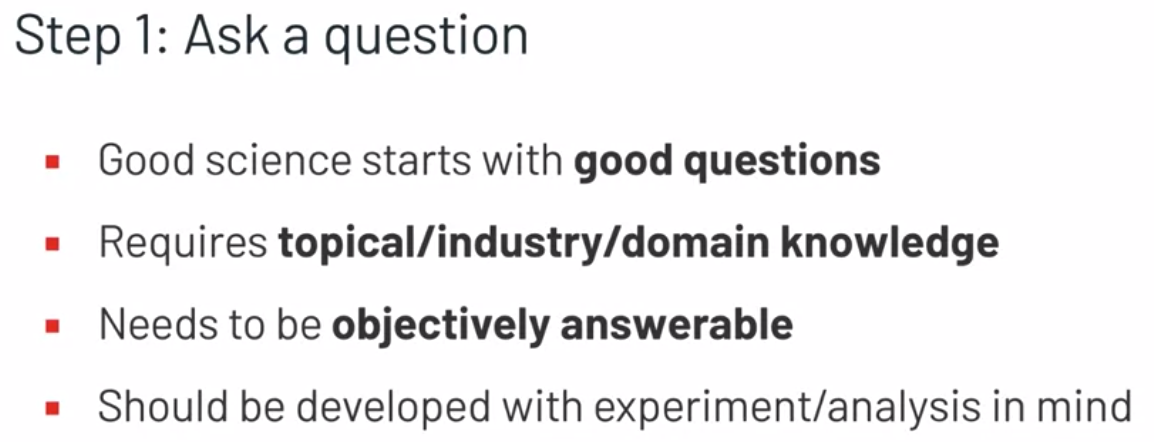
# **The Scientific Method**

Hi and welcome back, my name is Mark Roepke and I'm a technical curriculum developer here at Databricks. Where I focus on designing and developing data science and machine learning engineering courses and certification exams. In my free time I like to learn to cook new things, even though I mostly make pizza, and find ways to stay active outside, whether that be kayaking or hiking. As we mentioned in the module introduction, the first step of our learning journey is to understand what data science actually is. There a lot of definitions of data science out there, but we've seen through our experience that data science at large boils down to one simple idea, performing science with data. Let's take a look at how scientists perform with data playing a central role. You might have learned about the scientific method at some point. Perhaps like me you were sitting in a grade school classroom as your teacher talked about experiments. Maybe you have another memory of learning the scientific process or this could be your first time hearing the term. Whatever your experience, the scientific method is fundamental to quality data science.

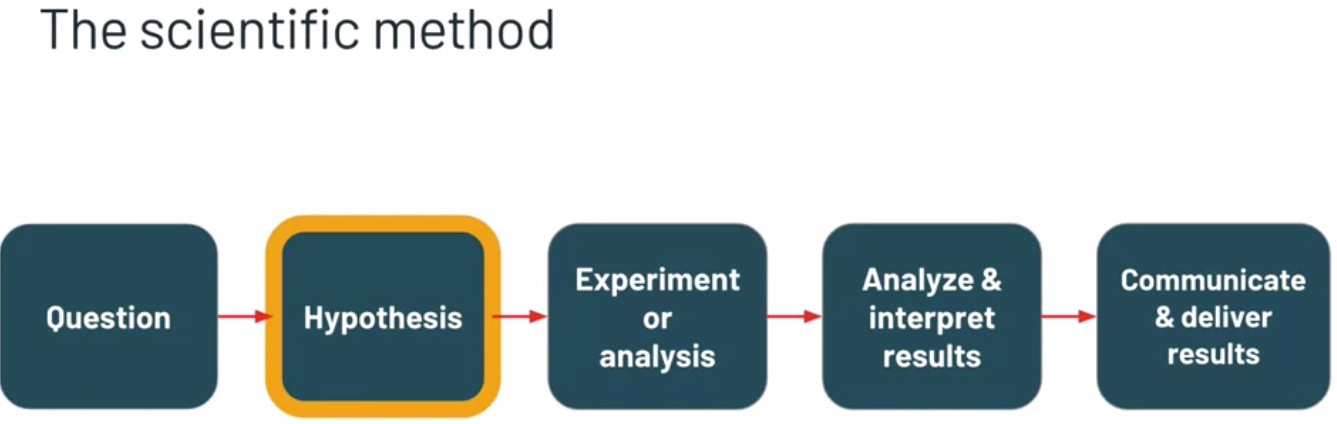


Generally speaking, the scientific method is the process of asking a question. Hypothesising an answer to that question, carrying out an experiment to assess that hypothesis. Analyzing and interpreting the results, and finally communicating those results. This process can vary in application across industries, but it's usually based on something that looks similar to this. The basic structure allows scientific practitioners a framework for learning objective truths about the world around them. Each of the steps in this process are vital to its success, so it's important to understand what makes each step successful.

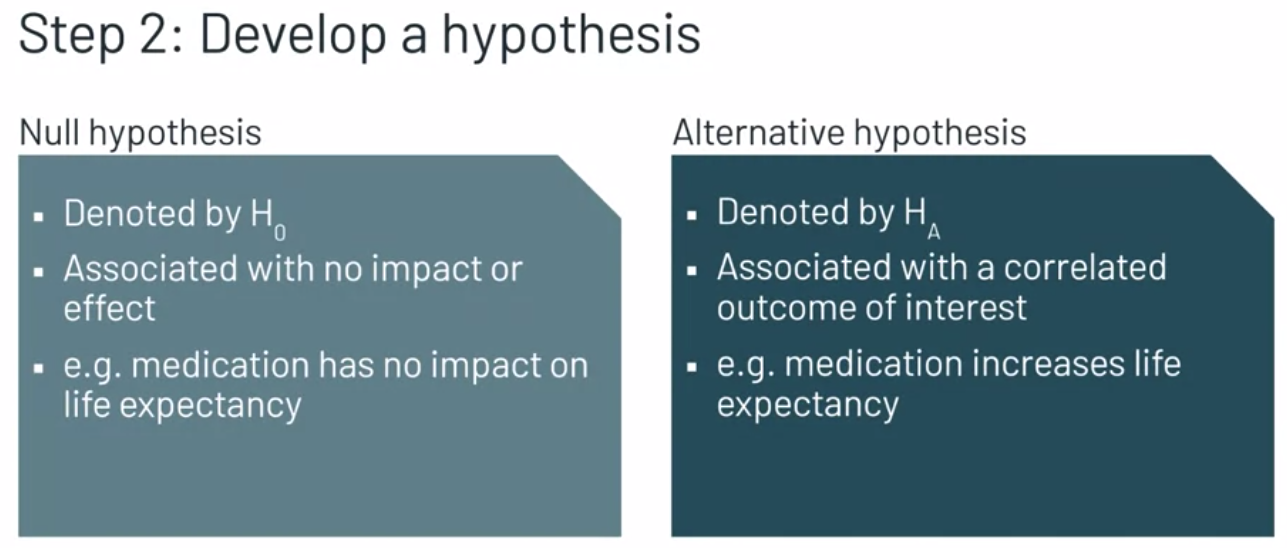
Asking a question is the root of the scientific process, so good science must begin with a good question.



In order to ask a good question, the scientific practitioner must have knowledge of the topic they're researching. They must understand whether or not their question is relevant, but they also need to determine whether or not their question can be objectively answered. This frequently requires an understanding of how well an experiment can be designed for a particular question, and how well the result of the experiment can be analyzed and interpreted. To start with a good question sets the rest of the scientific process out on the right foot.



Next, a quality hypothesis should be constructed to allow the design of the quality experiment. While it might be common to assume that the hypothesis is a prediction of what the practitioner thinks will happen, that's not necessarily always correct. Oftentimes the hypothesis should be developed in two parts, the null hypothesis and the alternative hypothesis.

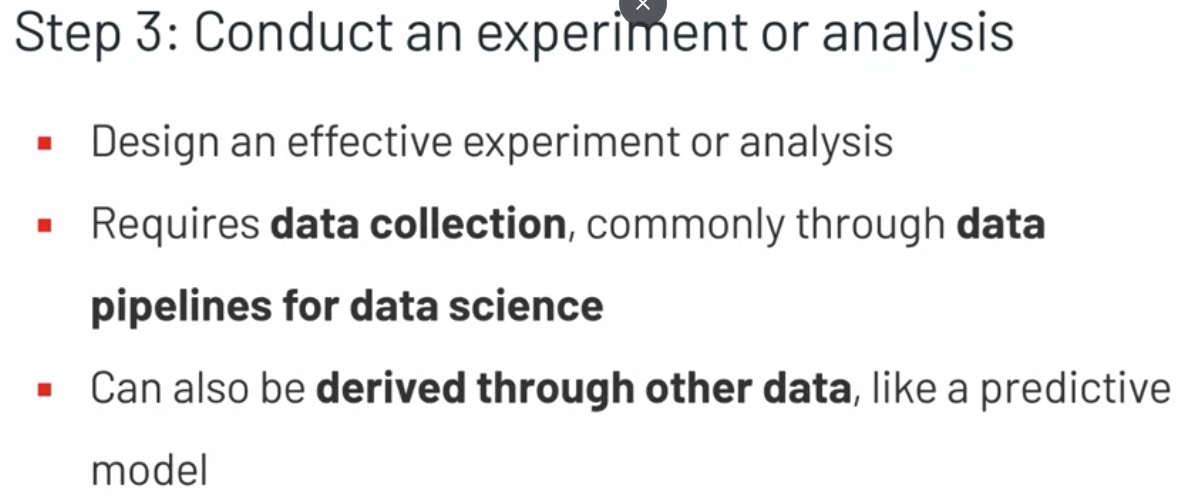


The null hypothesis is usually associated with an expectation that something will have no impact or no effect. An example of this would be a new medication having no impact on the life expectancy of those with a specific medical condition. On the other hand, the alternative hypothesis is usually associated with an expectation that something will have some type of correlated outcome. An example of this would be a new medication increasing the life expectancy for those with the specific medical condition.

Developing a hypothesis in this format allows the experiment to be assessed using various statistical tools, and we'll cover those later in this course.



Carrying out an experiment to assess a hypothesis can be a large task, but the validity of the scientific process depends on it.



To start, the practitioner must first design an effective experiment or analysis. Next, the collection of data must begin. While this can happen in a traditional scientific field with measurement tools and surveys, it's common in modern data science within industry for this to happen through automated data pipelines, things that Databricks facilitates very well.

It's also common for this data to be derived from other data. This could take the form of a model that uses data to make predictions which themselves are data.



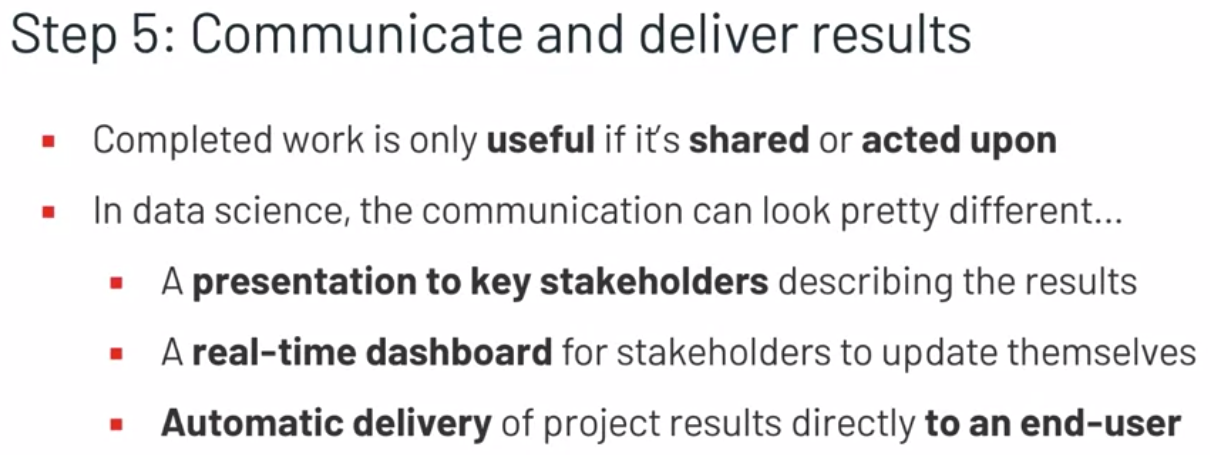
Once the data is collected, the practitioner can begin analyzing and interpreting the results to draw a conclusion on the experiment or analysis.



This includes looking at the data collected in a meaningful way to try to determine the validity of the hypotheses. The outcome of this step is usually a rejection of the null hypothesis or a failure to reject the null hypothesis. For reasons we'll go over later in this course, this terminology is incredibly important. The outcome of this step should always be an objective decision on the extent to which the alternative hypothesis is true.



While it might seem like reaching a conclusion should be the end of the scientific process, it's not. It's important and vital to communicate and deliver the results effectively.



In an academic scientific setting, this might look like a written paper or a journal article. But in a lot of data science settings, this communication could look pretty different. It could be a presentation to key stakeholders, a realtime dashboard for stakeholders to view whenever they want. Or results might be directly delivered to an end user, like a customer. Without this final step, much of the work done prior to this point in a data science project could fail to have a meaningful impact.

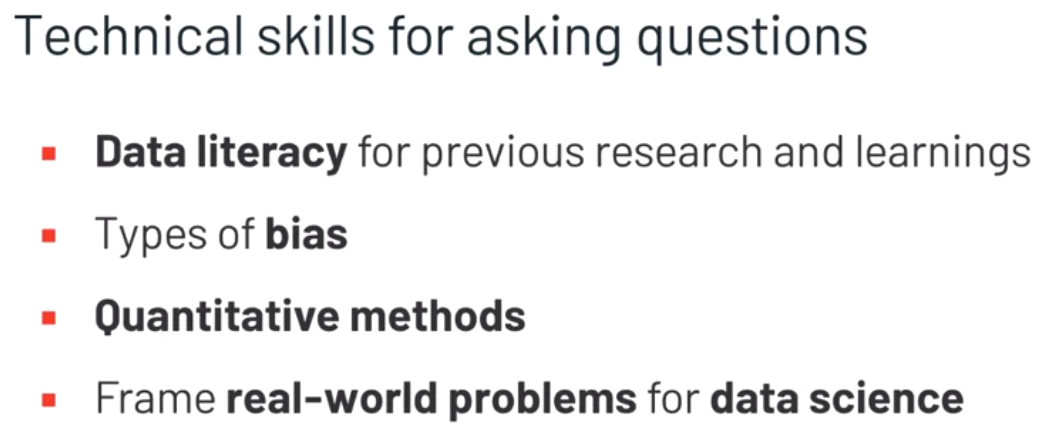
These steps together make up the scientific method. Data-related skills play an important role in each one of these steps, and we'll define these skills in the next video.

# **Skills of Data Science**

In the previous video, we described each of the steps of the scientific method, and the role data plays in that process. In this video, we'll go into a little more depth and specificity on the technical skills necessary to complete each of the steps in the data science workflow. If you're not familiar with all of the skills and techniques we outlined in this video, that's perfectly okay. One of the purposes of this course is to develop some of these skills. The key thing here is that you understand the importance of developing these skills, in order to perform data science well.



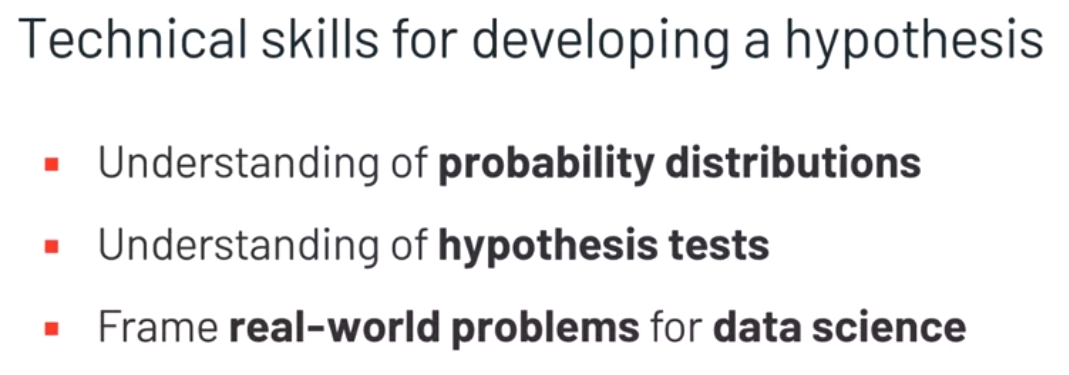
Let's begin with question development. Remember that quality questions must be relevant and answerable, and in objectively measurable manner. But how do we know if that's the case? Well, it requires some fundamental data science skills.



To understand if a question is relevant, we need data literacy to read and analyze previous research on a topic. This allows us to develop our own knowledge and ask meaningful questions whose answers and solutions can make a real impact. In order to determine what's objectively measurable, we need an understanding of bias and quantitative methods. Related to this, practitioners need to be able to ask questions in a way that can allow them to be answerable using data science methods. These skills will save us time and energy at this stage by letting us frame our questions in a way that can be easily solved using data. Together, these technical skills allow us to ask questions that will enable successful science.



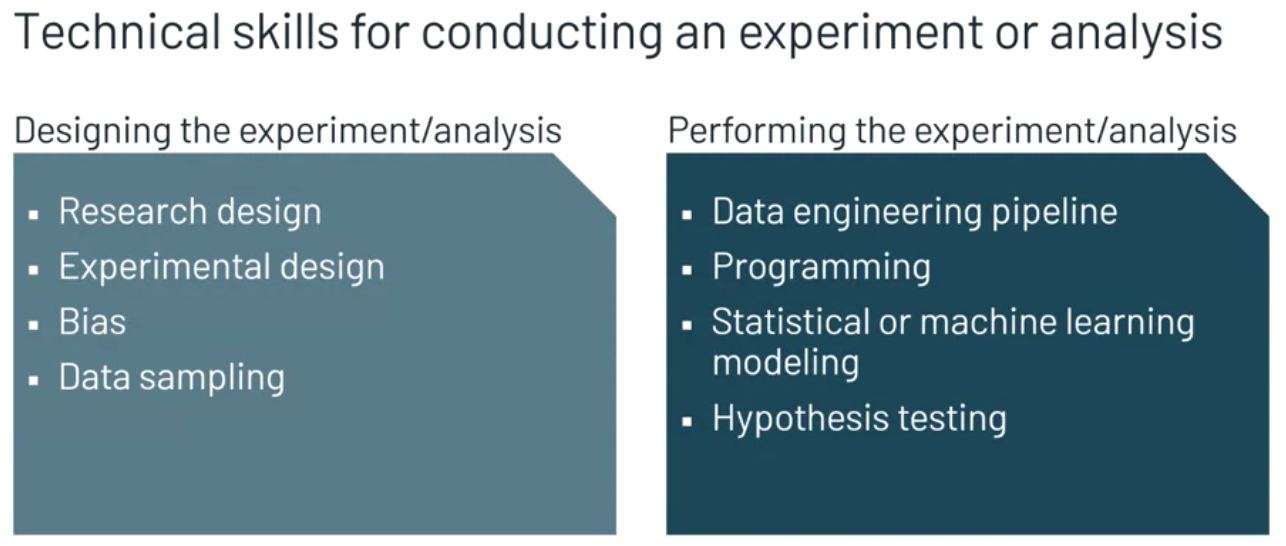
Next, let's talk about hypothesis development. When constructing hypotheses, remember that we want to construct a testable hypothesis set, one null hypothesis, and one alternative hypothesis.



In order to do that successfully, the scientific practitioner needs to have a fundamental understanding of probability distributions, relevant hypothesis tests, and the framing of real-world problems for data science. While these statistical techniques will not be directly used in this step, they can inform the development of hypotheses so that they're easily assessed later on in the data science process. This will help us scale out data science solutions so that they can be successful with limited manual intervention.



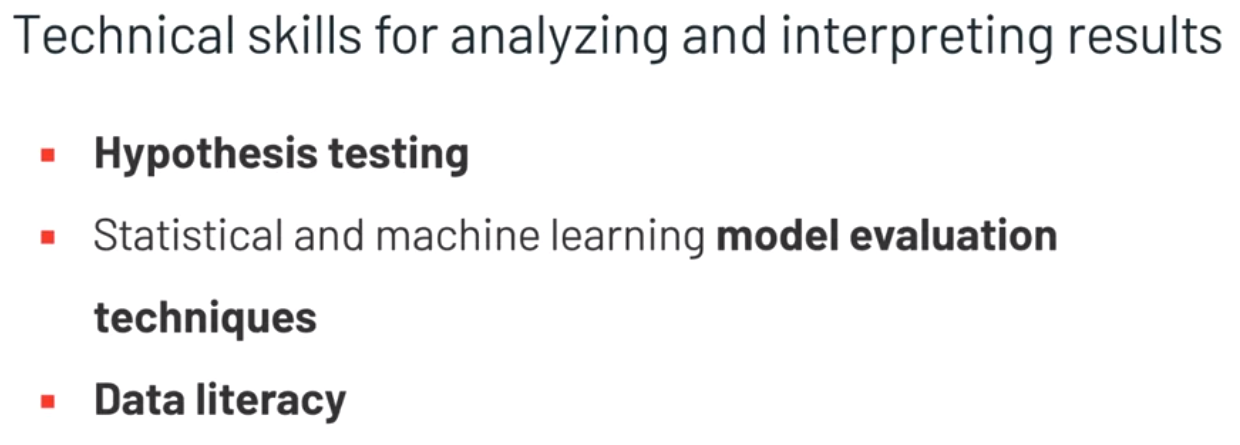
The next step in the scientific process is carrying out an experiment and there are a lot of technical skills necessary to do this effectively.



First, some type of experiment or analysis must be designed. This requires skills in experimental design and research design, like mitigating bias by sampling data in a responsible way. Then the experiment or analysis needs to actually be performed. In data science, this usually involves some type of data pipeline. This requires an ability to write data engineering code to import, clean, and manipulate data to make it useful. There are varying degrees of quality to these tasks, and it's not uncommon for them to be completed by dedicated data engineers working alongside data scientists on a data team. Data scientists can also create data in this step by using statistical modeling, machine learning modeling, or other statistical or machine learning techniques. This is a common practice in modern data science.



Following the conduction of an experiment, an analysis of the results needs to be completed and interpreted.

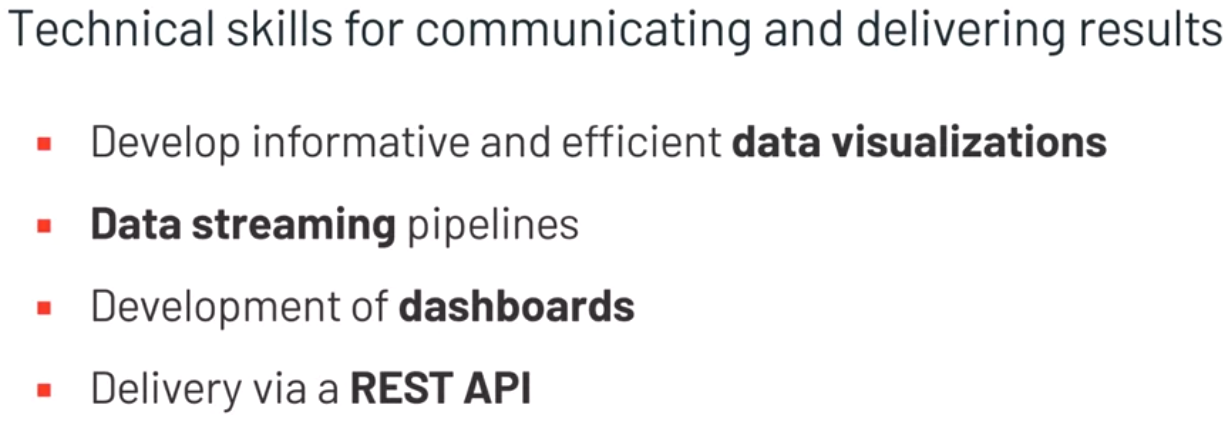


In order to analyze the results, data scientists need to have an ability to apply hypothesis testing methods, and machine learning model evaluation techniques.

Then based on these results, data scientists need data literacy to interpret the outcome and arrive at a conclusion. These are vitally important skills to a data scientist because this is how the value of data science work is assessed. In other words, this is how to determine if data science projects are making an impact, or if they might need to be reworked in some way.



Data science projects can be ineffective if the results aren't communicated, shared, or delivered and technology can help us complete these tasks.

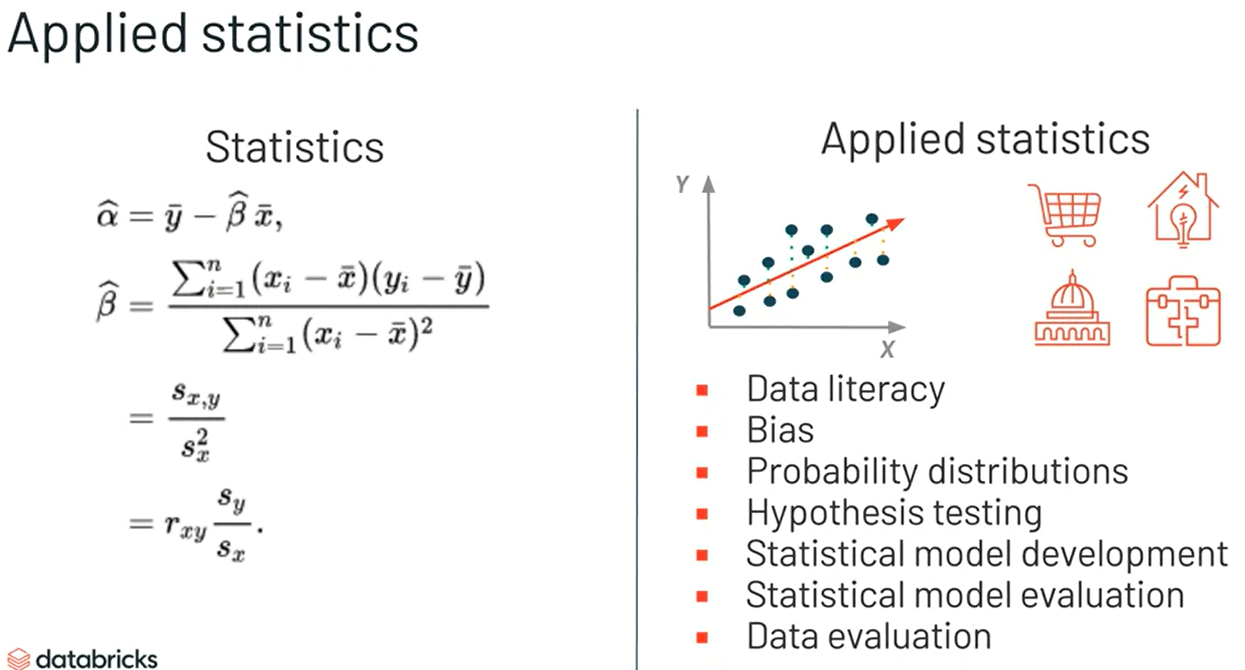


To deliver results to stakeholders as a presentation, a data scientist must be able to develop data visualizations that are informative and easy to understand. This skill, along with the ability to engineer data pipelines, is also important in the development of real-time dashboards, a common request from key project stakeholders. If the result that needs to be communicated is a model or prediction, a live REST API or an interface used by computers to get results on request, might need to be developed to support that communication.

We hope that these examples of concrete skills demonstrate the inseparability of the scientific process and performing data science at scale. Next, we'll work to make more sense of these skills by organizing them into common fields associated with data science.

# **Defining the Skills of Data Science**

Welcome back. In the previous video we identified some technical skills that can enable data practitioners at each step of the data science process. In this video, we're going to organize those skills into two core fields, applied statistics and computer science. Along the way well define, compare and contrast those two fields.



The field of statistics is usually associated with mathematics and mathematicians. But the field of statistics is broad and at its heart applied statistics is specifically concerned with the use of statistical techniques to solve real world data problems. In other words, applied statistics is focused on how we can use statistics in other fields. So applied statistics is more associated with people like data scientists and social scientists than it is with mathematicians. It's no coincidence, then, that we can classify many of the techniques we previously discussed as applied statistics techniques. Things like data literacy, sampling bias, probability distributions, hypothesis testing, statistical modeling and data evaluation are all part of the field of applied statistics. These topics are rooted in complex mathematics, but the way they are used in data science is frequently more applied in nature. Within the field of data science, computer science can be a broad term, but usually refers to a series of computer related tasks. Computer science for data science just like applied statistics is rooted in a more theoretical and academic version of itself. But when it comes to its applications, computer science is immensely valuable to the scaling of reproducible quality data science work. These tenants of scale reproducibility in quality are key to data science. So being able to apply the tools of computer science is vital to success as a data scientist. Things like programming, working with big data, real time streaming, creating data pipelines and developing REST APIs are all applications of computer science. Each of these topics have very complex computer science theory back in them, but data practitioners usually only need to know how to apply them to obtain their benefits. While applied statistics in computer science or separate fields, the intersection of the two is usually referred to as machine learning. Machine learning is concerned with using statistical and computer science principles together to learn from data. It's common for machine learning to be used to develop some type of predictive model. Aside from being combined in a way that can be considered machine learning, applied statistics and computer science have a few similarities and how they relate to data science. Both applied statistics and computer science are vital to performing quality data science at scale. And they're also both technical fields that require dedicated time to learn. At the root of both of these fields is mathematics. So any quantitative skill development or experience will be important in learning these fields. And while they do have these similarities, applied statistics and computer science go about contributing to data science in different ways. Applied statistics is mostly concerned with the techniques and the methodologies associated with extracting some type of information from the data. And computer science is primarily used to automate that process in a scalable, reliable and reproducible way. Developing technical skills in applied statistics in computer science will enhance any data practitioner ability to do data science well. Next you will assess your understanding of these fields with a short knowledge check. And then we'll begin describing the importance of domain knowledge as we continue to introduce data science.

# **Domain Knowledge**

Throughout this lesson so far, we've spent a lot of time talking about the technical details of data science. The scientific process, applied statistics, and various applications in computer science, but we've left out an imperative part of data science thus far, domain knowledge. Domain knowledge is the knowledge a practitioner has in the industry in which data science is being applied. And it's really important for data scientists to have a level of domain knowledge to do their job effectively. As an example, if a data scientist is working for a large grocery retail organization, they should have some knowledge of the grocery retail industry. And if a data scientist is working for a financial investment company, they should have some knowledge of the financial investment industry. Having knowledge in the domain in which data science is being done can help data scientists at various points in the scientific process. One of the most impactful places for domain knowledge is the question development stage. Data practitioners that have industry knowledge can ask better questions and more pointed questions. They know what types of industry problems need solutions and the start of their scientific processes reflect that. Domain knowledge is also helpful at other stages in the data science process. Understanding the details of the industry can inform how data scientists collect data, perform analysis, and interpret and communicate the results. As an example, a data scientist working in supply chain management would need to understand that the latency of data is very important in order to make the ordering decisions as quickly as possible. Also, this data scientists would likely choose to develop predictive models to estimate the demand of certain products in certain locations. They would know that they should evaluate their predictive models based on some balance between stockroom availability and out of stock issues. And finally, this information would need to be delivered in near real time to product ordering devices and warehouses so they're always aware of any results from the data science process. While the technical skill sets of applied statistics in computer science are core to the skill set of a data scientist, it's particularly important to build and develop domain knowledge as well. Incorporating domain knowledge into the data science process can transition good technical work into work that's truly impactful to an organization. In the next video we'll use what we've learned about all of these skills to drive a formal definition on data science that we'll use throughout this course.

# **Defining Data Science**

In the last few videos we've spent some time learning about the data centric scientific process, applied statistics, applications of computer science, and the importance of domain knowledge and data science. In this video we'll use what we know about each of those things to derive a definition for data science that we'll use throughout the rest of this course. So what do we know about data science? Well, we know that it starts with a scientific process, and we know that the scientific process has data involved at each step in some capacity, whether data is informing our questions or hypothesis or being analyzed to arrive at a conclusion. We could reasonably connect these dots to define data science as a scientific process in which data plays a central role in each step. But we know more about data science than that. We know that data science uses an interdisciplinary set of technical skills to complete this process. From applied statistic techniques like hypothesis testing, and statistical modeling, to computer science applications like programming and real time data streaming. Any definition of data science should include these components. It's also probably best to include the intersection of these two disciplines. Machine learning, because it plays such an important role in modern data science. So if we wanted to incorporate these skills into our working definition, we could define data science as a data centric scientific process using applied statistics, computer science, and machine learning. But again, this definition would be incomplete. It doesn't mention why we do data science. Generally speaking, why does science exist? Well, partly to learn more about specific topics and that's why we do data science. Our goal is to further our understanding and apply that understanding within a specific domain. And we can do this by approaching the problem with a level of domain knowledge. With domain knowledge as a central purpose and knowledge of that domain as an important skill, we can now define data science appropriately. Data science is an interdisciplinary field made up of domain knowledge, applied statistics in computer science, along with machine learning with the goal of using data within a specific process to develop and apply knowledge in a specific domain. This definition incorporates the scientific method, the importance of data or technical and domain skills in a purpose for data science. It can be applied to understand a variety of data science projects and problems across a variety of industries. And we encourage you to keep this definition in mind as we progress through the course. It will help you understand how each of the fundamental topics we will cover fit into data science as a whole. In the next video we'll use this definition of data science as a model to explain various examples of data science. After that, we'll introduce an activity for you to apply your learnings to design a data science project in an industry of your choice.

# **Examples of Data Science**

Hello. In this video, we'll use our definition of data science to model various examples of data science. Recall our definition of data science. Data science is an interdisciplinary field made up of domain knowledge, applied statistics, computer science, and machine learning with the goal of using data within a scientific process to develop and apply knowledge in a specific domain. This definition can be used to model almost any data science project. But let's take a look at our supply chain management example. We can represent this visually and use the framework to design our data science project. Let's start with a data-centric scientific process. First, we need to determine our question. Given our domain knowledge and customer survey data, we see that our stores have been having out-of-stock problems recently. As a result, we want to know if we can reduce the number of out-of-stock items in our store. Next, we need to determine our hypothesis set. Our null hypothesis could be that the number of out-of-stock items in the store remains constant after our treatment and the alternative hypothesis and in this case, the one we're hoping for is that the number of out-of-stock items in the store decreases after our treatment. These hypotheses represent opposite answers to our original question. So we'll conclude with only one. Next, we need to carry out our experiment or analysis. Given our technical and domain knowledge, we know that using a predictive model to predict when an item is likely to go out of stock could be a good way to reduce the number of out-of-stock items. It would provide our ordering system with an estimate of when to order more of each particular item in the store. We'll need to collect historical data on item inventory and items sales for our model to learn to use the relationship between the two. We can also use any other data we might think has a relationship with item inventory or items sales. Once we have all the data we'd like, we can go ahead and build a model to predict when an item is likely to go out of stock. Then, we need to evaluate the model. We might see according to pass data that it's predictions are pretty accurate and will reduce the number of out-of-stock items if used correctly. At this point, if it's possible, it would be best to release these predictions to a small sample of stores, as in market tests to see if the model truly does reject the null hypothesis by reducing the number of out-of-stock items in the store in a real-world setting, rather than just in a original model evaluation phase. Once we get our definitive result from a real-world setting, we can communicate and deliver our results. If we conclude that the model did reduce the number of out-of-stock items, we will likely want to describe the model and process to key stakeholders in some type of presentation. Next, we should set up some way to deliver these predictions to those responsible for ordinary new items, whether that's an automated ordering system or storm integers. Finally, it would be wise to set up a real-time monitoring dashboard to understand how our model is performing at all times. After a while, it might need to be refreshed due to changing relationships between inventory and sales. On the other hand, if we conclude that a model did not reduce the number of out-of-stock items, we still need to create some type of presentation describing why it did not work as intended and any possible recommendations for changes that can make the project more successful. It can be really helpful to fill out the decisions, the need to be made for a specific data science problem at each of these steps of the scientific process. But we also need to identify which technical skills will be necessary along the way. Throughout this process, we will need to use all of the skills in our data science toolkit, domain knowledge, data literacy, data manipulation and transformation, statistics, machine learning, model evaluation, programming, dashboarding, and working with REST APIs and real-time streaming. It might not be reasonable to expect that a junior level data scientists have all of these skills necessary to complete and deliver a project, but outlining which skills will be necessary at different stages can help entire data team's plan where other data practitioners like data engineers or machine learning engineers might need to be involved with this project. Well, this example looked at supply chain management. You can complete this process for just about any data science project. A few examples include targeted advertisements toward customers that are more likely to buy a product, predicting when a stock is likely to rise or fall in a financial market, determining whether a new vaccine is effective in preventing the spread of a disease and predicting which candidate might be likely to win an election, among many others across many other industries. For each of these projects, it can be helpful to design the data science process ahead of time by identifying the decisions and actions that need to be taken at each step, and identifying the skills necessary to complete those actions. At this point, we hope you have a solid understanding of what data science is, including its purpose, the skills required to perform data science, and the process that should be followed. You should also be comfortable with the idea of framing real-world problems in this data science process. To close out this lesson, you will complete an activity where you will design your own data science project using the process exemplified in this lesson. Thanks for joining us.

# **Design a Data Science Process Activity**

In this video, we'll describe the Designing Data Science Process activity. In this activity, you will be completing the data science process template for a project of your choice. It's your responsibility to pick an industry or domain that you're interested in or have an expertise in. Develop a question, hypotheses and some type of experiment or analysis, and you plan for how you'll nterpret and communicate your results. In addition, you'll want to identify the skills that the data scientists will need to have for each stage. Refer to the skills we listed in the previous videos for ideas. A completed project will look familiar to the data science process map that we created in the previous video for the supply chain management problem. Once you've completed your activity submitted to Coursera, and then you will have the opportunity to assess another student's submission. A detailed grading rubric will be available in the activity instructions. At that point, you will have finished the introduction to data science module. Be sure to join us in the next module to learn more about [inaudible] statistics methodologies that can be used to answer real-world questions using data.