**The Simple Things a Data Science Beginner Needs to Know**

**Are you curious about data science? Are you interested in the work of your data scientist coworker? Do you want to become a data scientist? This is your primer into this exciting field!**

**Table Of Contents**

1. [A (Very Brief) History](https://towardsdatascience.com/the-simple-things-you-need-to-know-as-a-data-science-beginner-c4b9ffd0b516#6935)
2. [How Data Scientists Create Value With Data](https://towardsdatascience.com/the-simple-things-you-need-to-know-as-a-data-science-beginner-c4b9ffd0b516#9b92)
3. [Tools Data Scientists Use](https://towardsdatascience.com/the-simple-things-you-need-to-know-as-a-data-science-beginner-c4b9ffd0b516#c7f6)
4. [What Problems Do Data Scientists Solve?](https://towardsdatascience.com/the-simple-things-you-need-to-know-as-a-data-science-beginner-c4b9ffd0b516#5f9a)
5. [How Do Machines Learn?](https://towardsdatascience.com/the-simple-things-you-need-to-know-as-a-data-science-beginner-c4b9ffd0b516#eddc)
6. [What Is A Data Science Project Like?](https://towardsdatascience.com/the-simple-things-you-need-to-know-as-a-data-science-beginner-c4b9ffd0b516#a662)
7. [Conclusion](https://towardsdatascience.com/the-simple-things-you-need-to-know-as-a-data-science-beginner-c4b9ffd0b516#a6d9)

This article is for anyone who wants a no-nonsense, easy explanation of what data science is, how it works, and what it is used for. Maybe you heard of data science and you wanted to learn more. Maybe you work with a data scientist and want to better understand their role, or you even have the goal of becoming one. This article, featuring uncomplicated term definitions and depicting examples, was made for you.

My name is Ken Jee, and I am a data scientist and content creator. In general, my mission is to help make this field more accessible for everyone (to be honest, I’m also kind of writing this article for my parents so they will stop asking me to explain what I do every week). With that being said, let’s jump right into it.

If you prefer a video format, watch here:

**A (Very Brief) History**

To understand what data science is now, it is important to understand where it started. Data science has been around for longer than most of us realize. In 1974, the famous computer scientist, Peter Naur, proposed data science as an alternative name for computer science. And funnily enough, in 1985, C.F. Jeff Wu used the term as an alternative name for a completely different field, statistics, in one of his lectures. If it isn’t obvious, there is some pretty killer foreshadowing here.

The official title of data scientist was first popularized by DJ Patil at LinkedIn who would go on to become the first U.S. Chief Data Scientist under Barack Obama.

While the origins of the field are actually quite old, the true evolution of the field is relatively new. There have been dramatic changes in the way data science has been done over the last 10 years due to the massive advancements in storage and computing capacity. These rapid changes over a short period of time are why people generally still consider data science a new or evolving discipline.

Although it may have been coincidence that both computer science and statistics were both called data science by a couple professors, it has become a reality. Data science is now a beautiful hybrid of these two domains. We also should throw in a little bit of business and subject area domain to even things out.

Many people still insist that data science is just calling statistics by a different name. In 1985, that could have been true during a C.F. Jeff Wu’s lecture. However, I don’t believe that to still be the case. Because of the huge volumes of data and the increased complexity of computing, many of the problems a data scientist faces today cannot be done without the help of computer science and some advanced understanding of the unique domain they are operating in.

So, in real terms, what is data science now? Enough with this abstract stuff.

Data science is an aptly named field. Data science is a domain where we work with data to generate some form of value. And we use scientific techniques to extract this value.

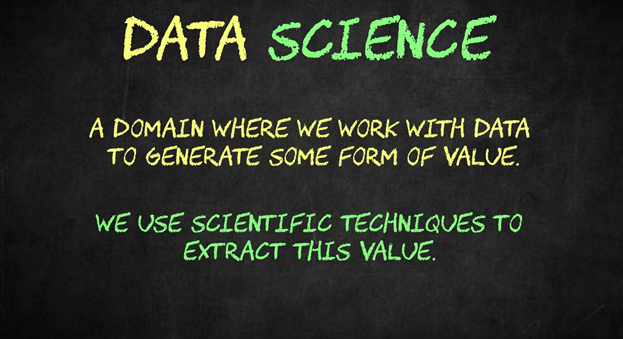


Image by the author

**How Data Scientists Create Value With Data**

Let’s take that apart a bit farther; how do we work with data to generate some value? The ways that data scientists drive value from data is for the most part derived from the data science lifecycle. All organization’s data usually follows this path.

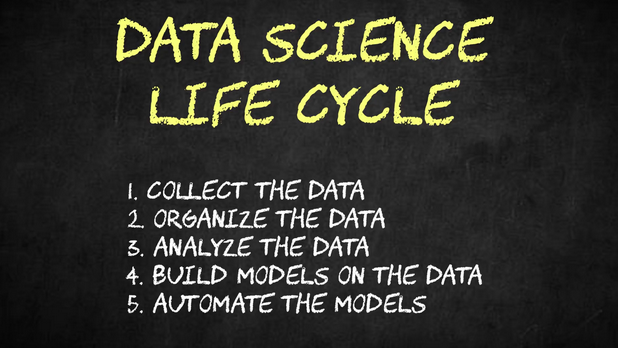


Image by the author

The first way that we generate value from data is through collecting it. While it isn’t necessarily a core role of all data scientists. Some data scientists use their skillset to collect data. This can be done through building systems for data intake like webpages or surveys or it can be done through writing code that collects data from different places online.

Another basic way a data scientist can create value from data is by organizing it. The vast majority of the data in the world is unstructured, meaning that it hasn’t been organized into a database. Some data scientists can transform this unorganized data into a structured format making it far easier to analyze. As part of this process, they can also “clean” the data by correcting misspellings, fixing errors, identifying duplicate records, and identifying missing values. A lot of these tasks are handled by data engineers these days, but they still fall under the data science umbrella. If you want more details on the specific data science roles, [here’s a video about that.](https://www.youtube.com/watch?v=BZFfNwj7JhE)

The next way that we create value from data is through analysis. Simple analysis starts with basic statistics. For example, we may want to look at the average spending of an online customer vs. an in-store customer. Insights like these can help us to make better informed decisions about how we merchandise or market. Often the best way to convey these insights is through beautiful data visualizations. We also may want to see if an ad campaign is effective or not. We could run an A/B test to see which of two ad placements drives the most sales.



Image by the author

This is where a lot of the science starts to pop up. For something like this we would want to use the scientific method and concepts like hypothesis testing to evaluate differences between groups and campaigns.

After data analysis, we start to get into what most people consider the sexy stuff: building predictive models. From past data, data scientists can often build models that predict future outcomes at better than random chance. This allows businesses to (hopefully) make better decisions about how they allocate their resources. For example, if we owned a farm we could build a model to predict how much fertilizer we need to purchase each month. Especially if fertilizer has a shelf life, it could save us money if we predict this very accurately. Another example would be if we were looking to franchise a new restaurant, we could in theory build a model that would predict the return on investment based on the geography, traffic patterns, and demographics of the new location.

The final main way that data scientists create value from data is through automation. If we put some of these models that we build into production, often they make recommendations at a pace that far exceeds humans. A great example of this is on Netflix. They have machine learning algorithms that recommend you videos in real time. For a real person to do that same service, it would take thousands and thousands of people and thousands and thousands of hours. In this case, it really only takes a few algorithms to do it almost instantaneously. And these things pay off. According to Comparitech, apparently [Netflix’s recommendation algorithm is worth over 1 billion dollars per year to them](https://www.comparitech.com/blog/vpn-privacy/netflix-statistics-facts-figures/).

**Tools Data Scientists Use**

Now you should hopefully understand where data scientists create value. But what tools do they use and what does their work look like?

In my mind, the most important tool in a data scientist’s toolkit is programming. Most data scientists use either Python or R with Python being the more popular of the two. Other languages are used, but it is usually for a specific domain or use case. Data scientists are able to access the data, manipulate it, create visualizations, build models, and productionize their models all through coding. Programming is a data scientist’s all purpose tool.

Data scientists also have specialist tools that they use. For getting and manipulating data, data scientists will often use SQL. This allows data scientists to communicate easily with databases where the data is stored. Another specialist tool would be something like Tableau or Power BI which provide a graphic interface for creating data visualizations and dashboards.

For some projects, there is so much volume that data scientists need to use more computing power. In these cases, they will access virtual computers owned by Amazon, Google or Microsoft, also known as cloud providers, to run their analysis.

The last tool that is becoming increasingly popular is Git. Git is a versioning tool for people who write code. Here’s a [full video on this for those who want a deeper dive.](https://www.youtube.com/watch?v=_0rHU6qAQe0)

**What Problems Do Data Scientists Solve?**

We get it, data scientists are using fancy tools. But what problems can they actually help solve? I think this is really important for non-data scientists to understand. There are 2 main types of problems that data science and machine learning are good at taking on: supervised learning and unsupervised learning.

The first is supervised learning. Supervised learning serves to predict specific outcomes. We want to predict things like if someone is wearing a mask or exactly how tall someone is. Supervised learning means that we have data where the outcomes we want to predict are labeled. Let’s say that we wanted to predict if a papaya was ripe or not. If our data had some papaya characteristics like length, softness, mass, and sugar content and if someone had labeled it “ripe”, then this would be a supervised learning problem. More specifically, this would be a classification problem. We are trying to classify if a papaya falls into a ripe or a not ripe category. If we were trying to predict how heavy a papaya is in grams, that would be a regression problem because we are trying to predict a continuous numeric value.

Unsupervised learning is another story. With unsupervised learning, we often don’t have pre-defined categories that things neatly fall into. Instead we see what data naturally groups together and we create new categories based on its similarities or differences. An example would be simple customer segmentation. Maybe you group customers together based on buying patterns and then name these groups based on similar characteristics. Another form of unsupervised learning is generative where we are creating text or images from a model trained on a massive corpus.

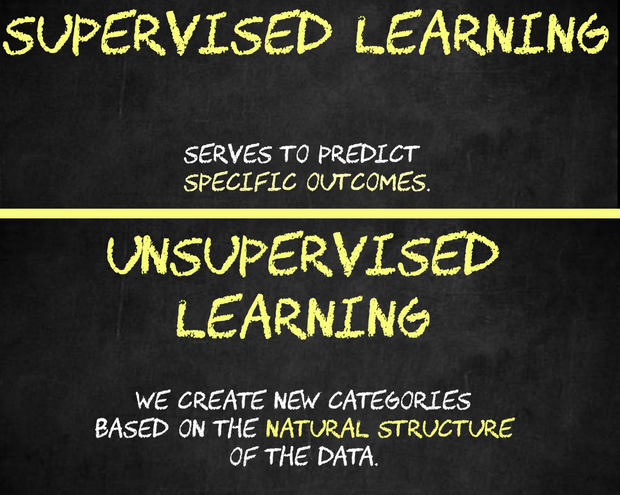


Image by the author

Again, I think it is really important to understand the limitations of machine learning and data science. Often I hear non-data scientists speaking about this field like a cure all. Yes, we can do some incredible stuff, but there are still limitations based on the specific business case. [Here’s a great example of how data science can go wrong with poor assumptions.](https://www.youtube.com/watch?v=FqNpDNmpcEo)

**How Do Machines Learn?**

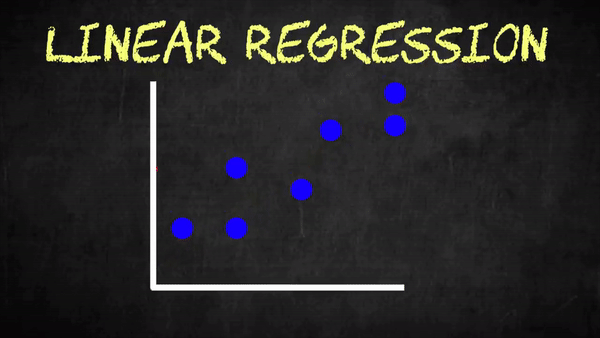
I realize I’ve mentioned machine learning quite a few times in this article. I’ve probably said it interchangeably with data science in fact. Questions I get a lot are what is machine learning versus what is data science? And how do machines actually learn?

While [I have a whole video that goes more into the specifics of this topic,](https://www.youtube.com/watch?v=q8cEt8gj3zY) machine learning algorithms are mainly what data scientists use to build their models. The supervised & unsupervised learning techniques that I listed above are all machine learning. On the other hand, most of the data analysis, data collection, and data cleaning techniques don’t fall into the machine learning bucket.

For the most part, whenever we are making a model that predicts future outcomes, groups data points algorithmically, or generates new material, what we are doing would be considered machine learning.

But where does the learning really take place?

Whenever a data scientist builds a model, they will split their data into a train set and a test set. They will use the train data to “teach the model” and then they will see how their model does at predicting the outcomes in the test set. Let’s do a very simple example. Our data is predicting someone’s YouTube views based on how many videos they’ve made. With a simple linear regression, training our model is fitting a line to this data. That is how our model would learn from this data. In order to make predictions on our test data, we just have to see the value of our line based on the data point we’re selecting in our training set.



One way of training a model is to use a flat line and then change the slope and intercept of the model to reduce the error or how wrong we were with each of the estimations. Image by the author

If our model does a good job predicting these test values, then we may consider it ready to apply to new data. With model training, you will hear terms like overfitting and underfitting and bias and variance trade off. I think they’re probably outside the scope of this article, but let me know in the comments section if you want me to make separate articles on these.

**What Is A Data Science Project Like?**

We build these models but what does the end-product of a data scientist’s work look like? Honestly, this varies pretty greatly by the role. Data science deliverables generally come in three flavors: 1) a dashboard that guides business stakeholders to their own insights, 2) a deliverable that makes a recommendation or a prediction on a specific problem, and 3) a trained model that users can get real-time predictions from.

I think it is really important to understand that within this domain, there aren’t really ever clear right and wrong answers. There are just shades of certainty and uncertainty. What I mean by that is that a model that we build gives us an estimate about what will likely happen. The confidence of our model helps us to decide if we should take action or not. In theory, any model can be wrong even if it is predicting if the sun will rise the next day.

Most models that we build, especially if they pertain to real-time predictions, need to be constantly maintained, retrained, and updated with new data so they are as accurate as possible.



Image by the author

**Conclusion**

Hopefully this article helped you to better understand data science and some of the types of problems data scientists can help solve. If you think this article would be helpful to one of your friends, someone you work with, or someone who is looking to become a data scientist, I would appreciate it if you forwarded it along.

If you enjoyed this article, remember to follow me on Medium for more content like this and [sign up for my newsletter](https://www.kennethjee.com/newsletter) to get weekly updates on my content creation and on additional learning resources in the data science industry! Also, consider supporting me and thousands of other writers [by signing up for a membership.](https://medium.com/@kenneth.b.jee/membership)

Thank you so much for reading and good luck on your data science journey.