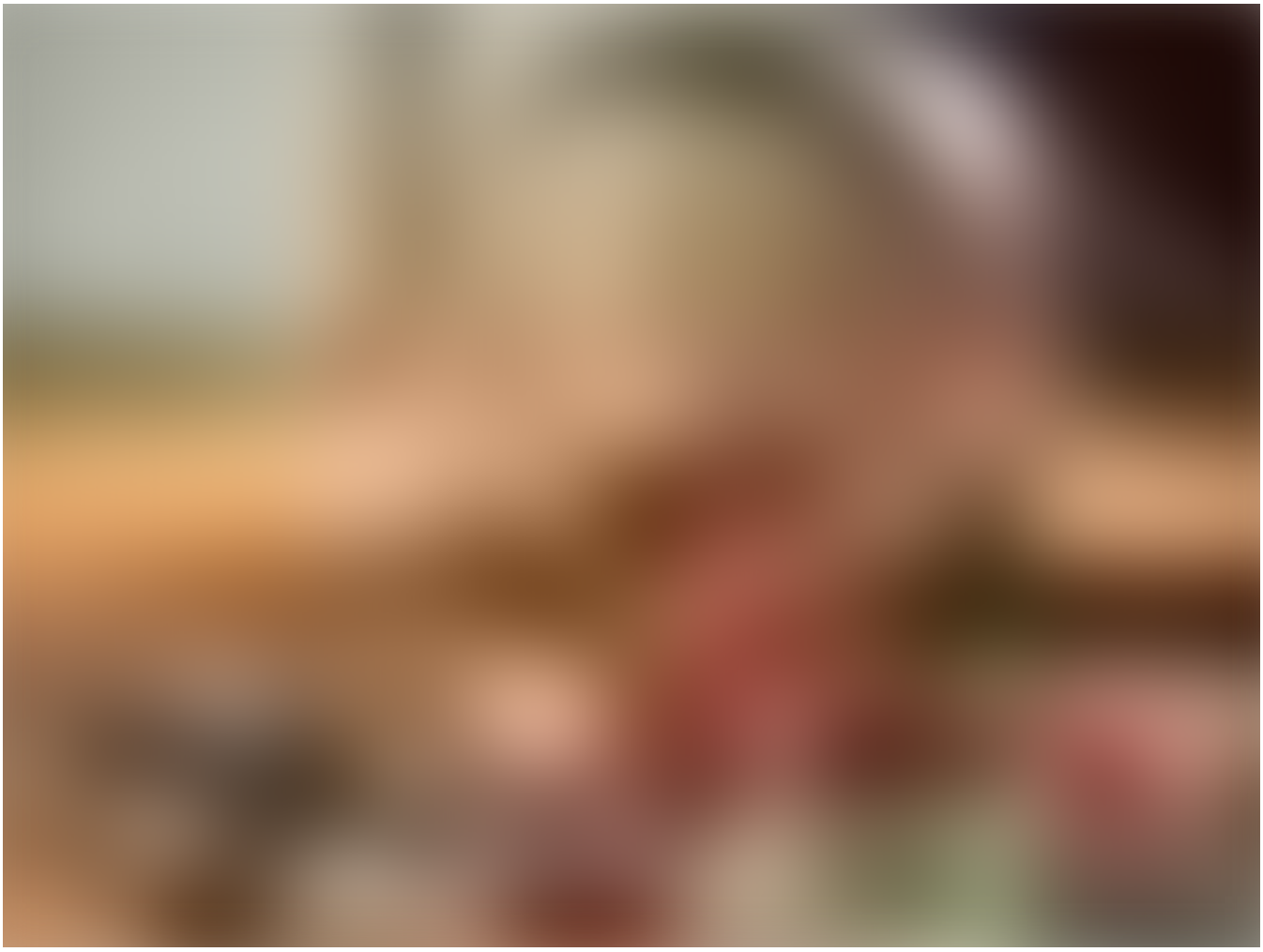


Photo by Josh Appel on Unsplash

How to Learn Data Science for Free

A complete learning path including resources that won't cost you a cent

Rebecca Vickery
Sep 30, 2019 · 10 min read ★

The average cost of obtaining a masters degree at traditional bricks and mortar institutions will set you back anywhere between $30,000 and $120,000. Even online data science degree programs don't come cheap costing a minimum of $9,000. So what do you do if you want to learn data science but can't afford to pay this?

I trained into a career as a data scientist without taking any formal education in the subject. In this article, I am going to share with you my own personal curriculum for learning data science if you can't or don't want to pay thousands of dollars for more formal study.

The curriculum will consist of 3 main parts, technical skills, theory and practical experience. I will include links to free resources for every element of the learning path and will also be including some links to additional 'low cost' options. So if you want to spend a little money to accelerate your learning you can add these resources to the curriculum. I will include the estimated costs for each of these.

. . .

Technical skills

The first part of the curriculum will focus on technical skills. I recommend learning these first so that you can take a practical first approach rather than say learning the mathematical theory first. Python is by far the most widely used programming language used for data science. In the Kaggle Machine Learning and Data Science survey carried out in 2018 83% of respondents said that they used Python on a daily basis. I would, therefore, recommend focusing on this language but also spending a little time on other languages such as R.

Python Fundamentals

Before you can start to use Python for data science you need a basic grasp of the fundamentals behind the language. So you will want to take a Python introductory course. There are lots of free ones out there but I like the Codeacademy ones best as they include hands-on in-browser coding throughout.

I would suggest taking the introductory course to learn Python. This covers basic syntax, functions, control flow, loops, modules and classes.

Data analysis with python

Next, you will want to get a good understanding of using Python for data analysis. There are a number of good resources for this.

To start with I suggest taking at least the free parts of the data analyst learning path on dataquest.io. Dataquest offers complete learning paths for data analyst, data scientist

and data engineer. Quite a lot of the content, particularly on the data analyst path is available for free. If you do have some money to put towards learning then I strongly suggest putting it towards paying for a few months of the premium subscription. I took this course and it provided a fantastic grounding in the fundamentals of data science. It took me 6 months to complete the data scientist path. The price varies from $24.50 to $49 per month depending on whether you pay annually or not. It is better value to purchase the annual subscription if you can afford it.

The Dataquest platform

Python for machine learning

If you have chosen to pay for the full data science course on Dataquest then you will have a good grasp of the fundamentals of machine learning with Python. If not then there are plenty of other free resources. I would focus to start with on scikit-learn which is by far the most commonly used Python library for machine learning.

When I was learning I was lucky enough to attend a two-day workshop run by Andreas Mueller one of the core developers of scikit-learn. He has however published all the material from this course, and others, on this Github repo. These consist of slides, course notes and notebooks that you can work through. I would definitely recommend working through this material.

Then I would suggest taking some of the tutorials in the scikit-learn documentation. After that, I would suggest building some practical machine learning applications and learning the theory behind how the models work — which I will cover a bit later on.

SQL

SQL is a vital skill to learn if you want to become a data scientist as one of the fundamental processes in data modelling is extracting data in the first place. This will more often than not involve running SQL queries against a database. Again if you haven't opted to take the full Dataquest course then here are a few free resources to learn this skill.

Codeacamdemy has a free introduction to SQL course. Again this is very practical with in-browser coding all the way through. If you also want to learn about cloud-based database querying then Google Cloud BigQuery is very accessible. There is a free tier so you can try queries for free, an extensive range of public datasets to try and very good documentation.

Codeacademy SQL course

R

To be a well-rounded data scientist it is a good idea to diversify a little from just Python. I would, therefore, suggest also taking an introductory course in R.

Codeacademy have an introductory course on their free plan. It is probably worth noting here that similar to Dataquest Codeacademy also offers a complete data science learning plan as part of their pro account (this costs from $31.99 to $15.99 per month depending on how many months you pay for up front). I personally found the Dataquest course to be much more comprehensive but this may work out a little cheaper if you are looking to follow a learning path on a single platform.

Software engineering

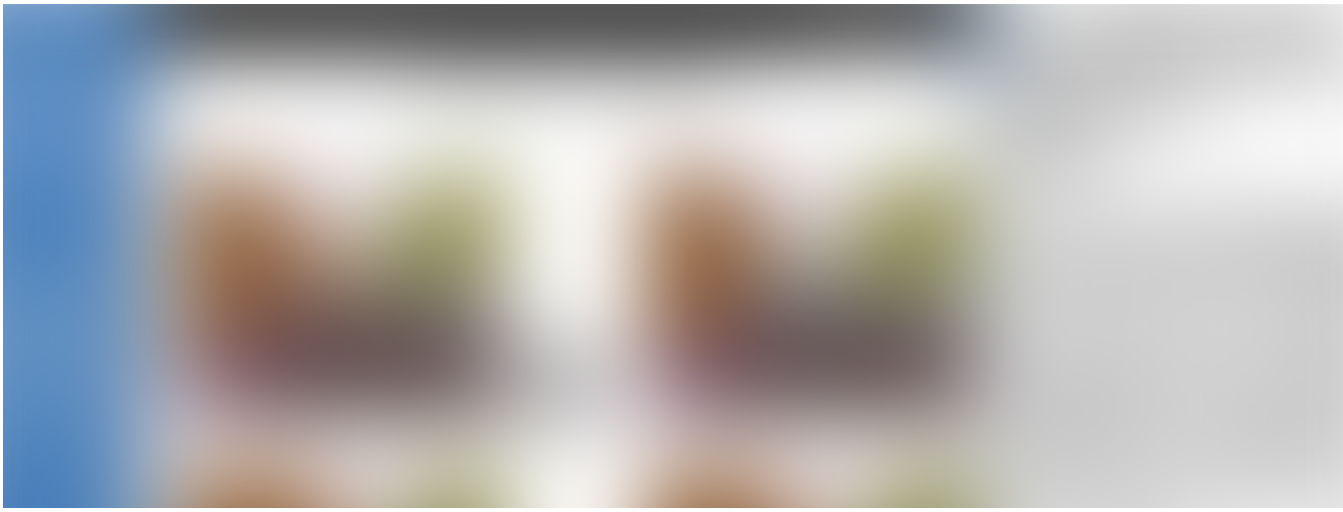
It is a good idea to get a grasp of software engineering skills and best practices. This will help your code to be more readable and extensible both for yourself and others. Additionally, when you start to put models into production you will need to be able to write good quality well-tested code and work with tools like version control.

There are two great free resources for this. Python like you mean it covers things like the PEP8 style guide, documentation and also covers object-oriented programming really well.

The scikit-learn contribution guidelines, although written to facilitate contributions to the library, actually cover the best practices really well. This covers topics such as Github, unit testing and debugging and is all written in the context of a data science application.

Deep learning

For a comprehensive introduction to deep learning, I don't think that you can get any better than the totally free and totally ad-free fast.ai. This course includes an introduction to machine learning, practical deep learning, computational linear algebra and a code-first introduction to natural language processing. All their courses have a practical first approach and I highly recommend them.



Fast.ai platform

. . .

Theory

Whilst you are learning the technical elements of the curriculum you will encounter some of the theory behind the code you are implementing. I recommend that you learn the theoretical elements alongside the practical. The way that I do this is that I learn the code to be able to implement a technique, let's take KMeans as an example, once I have something working I will then look deeper into concepts such as inertia. Again the scikit-learn documentation contains all the mathematical concepts behind the algorithms.

In this section, I will introduce the key foundational elements of theory that you should learn alongside the more practical elements.

The khan academy covers almost all the concepts I have listed below for free. You can tailor the subjects you would like to study when you sign up and you then have a nice tailored curriculum for this part of the learning path. Checking all of the boxes below will give you an overview of most elements I have listed below.

The Khan Academy

Maths

Calculus

Calculus is defined by Wikipedia as "the mathematical study of continuous change." In other words calculus can find patterns between functions, for example, in the case of derivatives, it can help you to understand how a function changes over time.

Many machine learning algorithms utilise calculus to optimise the performance of models. If you have studied even a little machine learning you will probably have heard of Gradient descent. This functions by iteratively adjusting the parameter values of a model to find the optimum values to minimise the cost function. Gradient descent is a good example of how calculus is used in machine learning.

What you need to know:

Derivatives

- Geometric definition
- Calculating the derivative of a function
- Nonlinear functions

Chain rule

- Composite functions
- Composite function derivatives
- Multiple functions

Gradients

- Partial derivatives
- Directional derivatives
- Integrals

Linear Algebra

Many popular machine learning methods, including XGBOOST, use matrices to store inputs and process data. Matrices alongside vector spaces and linear equations form

the mathematical branch known as Linear Algebra. In order to understand how many machine learning methods work it is essential to get a good understanding of this field.

What you need to learn:

Vectors and spaces

- Vectors
- Linear combinations
- Linear dependence and independence
- Vector dot and cross products

Matrix transformations

- Functions and linear transformations
- Matrix multiplication
- Inverse functions
- Transpose of a matrix

Statistics

Here is a list of the key concepts you need to know:

Descriptive/Summary statistics

- How to summarise a sample of data
- Different types of distributions
- Skewness, kurtosis, central tendency (e.g. mean, median, mode)
- Measures of dependence, and relationships between variables such as correlation and covariance

Experiment design

- Hypothesis testing
- Sampling

- Significance tests
- Randomness
- Probability
- Confidence intervals and two-sample inference

Machine learning

- Inference about slope
- Linear and non-linear regression
- Classification

. . .

Practical experience

The third section of the curriculum is all about practice. In order to truly master the concepts above you will need to use the skills in some projects that ideally closely resemble a real-world application. By doing this you will encounter problems to work through such as missing and erroneous data and develop a deep level of expertise in the subject. In this last section, I will list some good places you can get this practical experience from for free.

> "With deliberate practice, however, the goal is not just to reach your potential but to build it, to make things possible that were not possible before. This requires challenging homeostasis — getting out of your comfort zone — and forcing your brain or your body to adapt.", Anders Ericsson, Peak: Secrets from the New Science of Expertise

. . .

Kaggle, et al

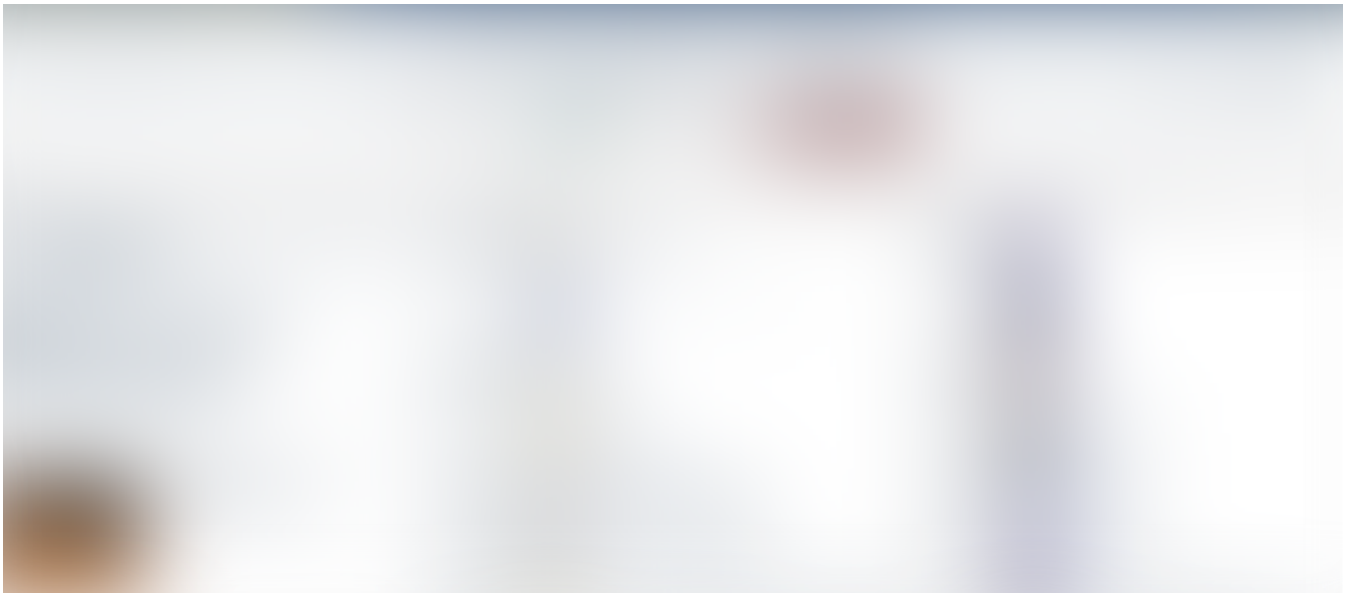
Machine learning competitions are a good place to get practice with building machine learning models. They give access to a wide range of data sets, each with a specific problem to solve and have a leaderboard. The leaderboard is a good way to benchmark how good your knowledge at developing a good model actually is and where you may need to improve further.

In addition to Kaggle, there are other platforms for machine learning competitions including Analytics Vidhya and DrivenData.

Driven data competitions page

UCI Machine Learning Repository

The UCI machine learning repository is a large source of publically available data sets. You can use these data sets to put together your own data projects this could include data analysis and machine learning models, you could even try building a deployed model with a web front end. It is a good idea to store your projects somewhere publically such as Github as this can create a portfolio showcasing your skills to use for future job applications.



UCI repository

Contributions to open source

One other option to consider is contributing to open source projects. There are many Python libraries that rely on the community to maintain them and there are often hackathons held at meetups and conferences where even beginners can join in. Attending one of these events would certainly give you some practical experience and an environment where you can learn from others whilst giving something back at the same time. Numfocus is a good example of a project like this.

. . .

In this post, I have described a learning path and free online courses and tutorials that will enable you to learn data science for free. Showcasing what you are able to do in the form of a portfolio is a great tool for future job applications in lieu of formal qualifications and certificates. I really believe that education should be accessible to everyone and, certainly, for data science at least, the internet provides that opportunity. In addition to the resources listed here, I have previously published a recommended reading list for learning data science available here. These are also all freely available online and are a great way to complement the more practical resources covered above.

Thanks for reading!

I send out a monthly newsletter if you would like to join please sign up via this link. Looking forward to being part of your learning journey!

Data Science Machine Learning Programming Learning Education