How I learned Linear Algebra, Probability and Statistics for Data Science

How I failed to learn math for data science and then what I did to understand Linear Algebra, Probability, Bayes’ Theorem, Probability Density Function, and basic Statistics

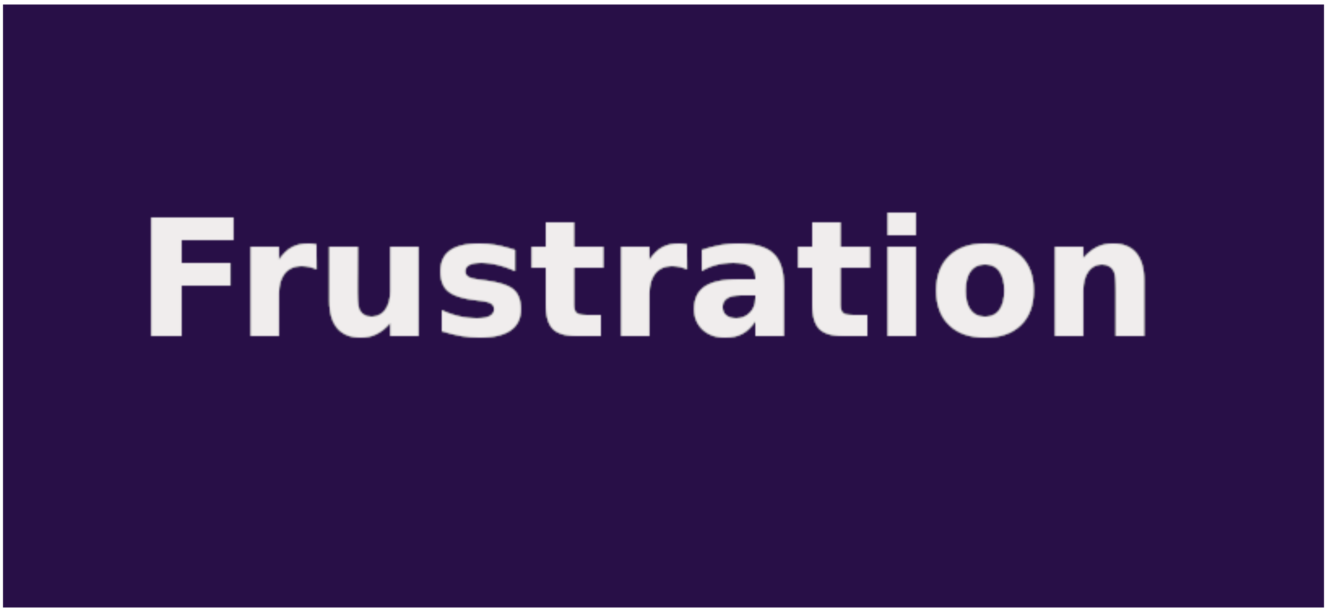
Accident ferroviaire de la gare Montparnasse, Source : Wikimedia

Background

That is exactly how my data science journey looked like after a year.

And yeah, Happy New Year :-)

When it comes to learning math for data science, everything starts and ends with failure. I am sure I am not alone, and it is the same story of many of those who started in data science. If you want one word to describe my efforts to learn math for data science, it is:



Frustration

Image by Author

Failing to learn what you need, especially when there is no clear path to data science, leads to frustration. At least I had my goals cleared-up:

1. Learn whatever math I need and nothing more
2. It does not matter what my background is, what experience I have, or lack. If all I have is a desire to learn math for data science then I should be able to do it
3. Focus more on behavioral characteristics, specifically **attitude** and **persistence** rather than mastering a particular math topic.

Math is a scary subject. We humans have emotions and desires while math is based on logic and methods. With artificial intelligence, it might be possible to put some emotions into machines but math got no place for them. Our heads, our feelings get hurt while learning math. *At least this is how I always felt about math*. An ancient and dreaded thing.

And what do we do when we don't know any math and when we are not gifted with that genius-level IQ. We do these things:

1. buy a book
2. begin a MOOC
3. work hard
4. if you fail, you work harder
5. if you fail again, you work even harder
6. you fail time and again. You work harder and harder and exhaust your will-power and one day start believing that you are not a "math type"

I did that.

Working harder resulted in more disappointments, more frustrations, and in the end, **anger** and **low self-esteem**. And when it comes to math for data science, I repeated this story for every topic I needed to learn, Linear Algebra, Statistics, Probability, Linear Regression, and Gradient Descent. This was "my story of learning math". Until now.

For A Complete Beginner

If you are a complete beginner, then I suggest, as per my experience, to go in this order:

- Learn Python programming
- Learn idiomatic pythonic methods (e.g. list comprehensions, generators, etc.)
- Learn Pandas
- Clean and Wrangle some datasets using pandas
- Learn Matplotlib
- Plot some datasets
- Combine your knowledge of Pandas and Matplotlib. Wrangle datasets and plot them
- For machine learning, do 2–3 small projects like Iris flower, Boston housing, wine classification, and Titanic dataset, etc.

You see, no math needed for beginning in data science.

This will take good 3–4 months of your time (some people can do it in one month but I am friends with Sloths)

A Sloth (Bicho-preguiça 3) by Daniella Maraschiello , Source : Wikimedia

You Don’t Need A Lot Of Math For Data Science

I guess the experience of one person should not be enough to convince you that we don’t need much math for data science.

Hence I found a good blog-post by Josh Ebner of Sharp Sight Labs. he explains the difference between Junior and Senior data scientists, the math you need for data science foundational skills, the difference between data science theory and practice, etc. You should read it:

You don't need to know much math for data science

There's a lot of misinformation about how much math you need for data science. Most people tell you that you need way...

How about a bit of advice from Tim Hopper. He was a math major and was also a Ph.D. math student for a year before he became a data scientist. Surely he knows how much math we need for data science. Short answer: not much, less than 10%:

How I Became a Data Scientist Despite Having Been a Math Major - tdhopper.com

Caution: the following post is laden with qualitative extrapolation of anecdotes and impressions. Perhaps ironically...

Here is his YouTube talk.

No, you don’t need much math and you do need some, only certain topics. You can do one bulleted point here per week:

- Learn basic Algebra (only certain topics)
- Learn Probability (only certain topics)
- Learn Statistics (only certain topics)
- Learn Linear algebra (only certain topics)
- Learn Linear Regression

has a list of math topics you need to learn for data science:

Maths and Statistics, A Complete Roadmap for Learning Data Science — Part 3

Key concepts in maths and statistics for data science, and where to learn them.

This was the **what** of this post. Next, we will talk about the **Why** and **How**.

Linear Algebra

I learned how to clean datasets using Pandas. I learned how to use matplotlib to create visualizations. Then I did Iris and Boston housing projects and then instead of picking up machine learning, I went ahead and directly started Practical Deep Learning for Coders, and it was an amazing, awesome book. I decided I will read and re-read this book will follow fast.ai for eternity. I am a fan of practical learning methods that don’t waste time and don’t expect you to have your graduate school syllabus memorized. teaches in this practical way. After successful 10 days with the book and building the bear detection model and deploying it on binder, I hit a snag when I came across this:

> (train_x[0]*weights.T).sum() + bias

What is that **.T**?

That was my first reaction. I posted the same on Fast AI Forums and got a good reply. The vector was stacked vertically, but it was needed to be in a horizontal position. It should have solved my problem, but it complicated matters more for me when I searched for vectors. Usually, when we want a vector, we simply do this:

> *import numpy as np*
>
> *np.array([1,2,3])*

And I got a vector, simple. Right?

Nope.

When I checked how a vector is presented in mathematics, it was strange. That day, almost all of my search on "vectors in mathematics" showed a vector that looked like this:

From Matrix at Wikipedia

You can see for yourself here. Now, this hit me on the head. I thought if this is how a vector looks by default in math then why don't we do the same, by default, in computer science. To get the default look of math, why I have to do this:

> *np.array([1,2,3]).reshape(3,1)*

Next question: *why this creation of a vector and that **.T** look the same?* What is transpose anyway? Why most of the vectors were vertically stacked in math but not in programming libraries?

I did not know how to do matrix multiplication either. So instead of struggling through the book in the next sections, I decided to learn what is transpose, matrix multiplication, etc before moving ahead. All the Linear Algebra you need for data science can be learned from these good places:

1. Linear Algebra from Ritchie Ng
2. Linear Algebra from Dive Into Deep Learning
3. Linear Algebra from Pablo Caceres. (most comprehensive. I did 70% of it because I wanted to learn certain topics. It has a lot of theory and I think it contains more than enough of whatever you need to know for even for deep learning)
4. Linear Algebra from Deep Learning Book

The 4th one is where I got stuck. It was too advanced for me. So, I have kept it for later.

Machine Learning Failed

Data science as a field is not mature yet, hence there is no direct path to data science yet. This is in contrast to fields like computer programming, software engineering, and web development. These three fields are quite mature and if you need help to build a career in any then there is plenty of help around. All you need to do is to just look for it. Data science is not as developed that yet. One has to keep updating himself by reading articles, blog posts, and watching videos. I do the same. And because of this, I came to know that doing deep learning before you understand machine learning can lead to a “disaster waiting to happen”. You need to comprehend the difference b/w linear regression, and logistic regression, and why you will prefer one over the other for a specific problem. If you don’t know it, don’t get into building deep learning models. It made sense to me.

So, I started machine learning.

Source: Wikimedia

I started learning linear regression and then got hit hard by the Statistics. I could not connect different parts of the topics. Just like SQL, I have learned what is “mode” and what is A/B testing, five times over, and then forgotten both the same five times. I took a STAT100 from Penn State online (a week) and when I tried to learn a bit more Statistics from other places, I got hit again with something called…

My Feelings for Probability

Probability is something that is used very much in real life (just like Statistics). And probabilistic thinking is not easy to come by. I tried to learn Bayes’ Theorem three times in 2020 and I gave it up all those three times. I have spent days and nights and my weekends trying to get a grip on Bayes’ Theorem, but it was like a mystery I could never solve. This time I was more hungry because everything was depending on it:

- I started with Practical Deep Learning for Coders and got stuck at **.T**
- Learned Linear Algebra
- Started machine learning and got stuck in linear regression
- Statistics was the answer and that got me stuck with Probability
- Probability is frustrating (think Bayes’ Theorem)
- Back to where I started. Can’t understand math. Frozen

> *5 construction workers build a 10x10 feet wall in 10 days. Given that 3 workers take 7 days to paint it with the color yellow, what is the probability that the price of tomatoes in Berlin is exactly the same as in Stockholm?*

Photo by yang miao on Unsplash

Yup, this is how I saw Bayes' Theorem whenever I failed. Frustration, when not given an empowering meaning, makes you behave dumb. 's blog-post saved me here. I highly recommend you read his blog-post. He has accurately portrayed the psychological condition of someone with average intelligence trying to learn data science on his own:

How to Learn Data Science: Staying Motivated.

Advice on how to be more consistent in your educational journey.

A New Approach

So you see, I am carrying all that baggage on my back. I needed to have a lot more hunger and drive to break through all these chains of “can’t learn math” and “I always got stuck in Bayes’ Theorem” etc. The only way I saw was to push the limits of my capabilities. In the age of AI, the boundaries between personal and professional lives have blurred. Whatever we do in one affects the other a whole lot more than you can imagine. I needed a new thinking, a new approach to learning. I needed to give a new meaning to learning. **I wanted every aspect of data science learning an enjoyable experience, an experience that I could cherish as loving memories in the future.** I asked myself:

> *In my personal life, how did I entertain myself in the last few months? Where did I find joy?*

I loved watching House of Cards, SUITS, Ghost in the Shell, Billions, and Star Wars. I binge-watched many seasons/volumes of these. I decided to binge-watch, binge-read, and binge-practice Probability for one whole week: Monday to Sunday. But before that, I wanted to see if I could learn Statistics and Probability together at the same time. This is how I got…

The Udacity Experience

I started Udacity’s Intro to Statistics free MOOC because it had all the Statistics and Probability one needs. It looked short and to-the-point which fits a data scientist’s approach to math. It was good but after chapter 16 (33% in) I gave it up. The problem is that even though this MOOC is to-the-point, it assumes a natural mathematical intuition. Even though Udacity says it is an intro level MOOC, I found one needs to be pretty smart and have a really good mathematical intuition to get through it. Like I said in the beginning, I am not a genius, I am just another guy you find on the street. So with every section of the MOOC, I had to spend double/quadruple the time searching and learning from multiple resources outside of the MOOC. This is one of the experiences that pushed me towards the "binge-watch" idea. By all means, you go ahead try that MOOC, it has less theory and some really good problem to be solved at the end of each video. I highly recommend you do it if you can

Adapting My New Approach

I came up with a new plan:

I will not read a mathematics textbook. *I will not do any MOOC either*. The reason is: both of these come from academic standards designed for graduate studies (3+ years). People in academics are already experts on their subjects, they had been teaching those for years and hence the MOOCs/books written on the same are one or two semester-long at minimum. What about a guy who does not know anything about those subjects and does not have a semester or two to learn?

All of us are trying to break into data science in this 4th industrial revolution. We are hard-pressed for time, we don’t have 3 years. We need to get up and start producing stuff as soon as possible (it’s been already a year since my last job). So that is why we have to come up with new ways of learning for the 21st century’s business requirements. has explained it better in his blog-post “Don’t Do Data Science”:

Don’t Do Data Science, Solve Business Problems

The term ‘Data Scientist’ has become colloquialized in modern business speak to signify an individual with almost every…

To be honest, I have the uttermost respect for academia. Some of the greatest discoveries have come from academic institutions. In fact, I still have a desire to become a researcher in academia, primarily because it is not run by commercial interests. I want to do a master’s and then a Ph.D. in machine learning (maybe even two Ph.D.). I think the development of humanity depends on academia as much it depends on businesses bringing technology to solve problems. That said, everything has a time and a place, and right now my need is for quick but fundamental learning.

Don’t Follow your passion

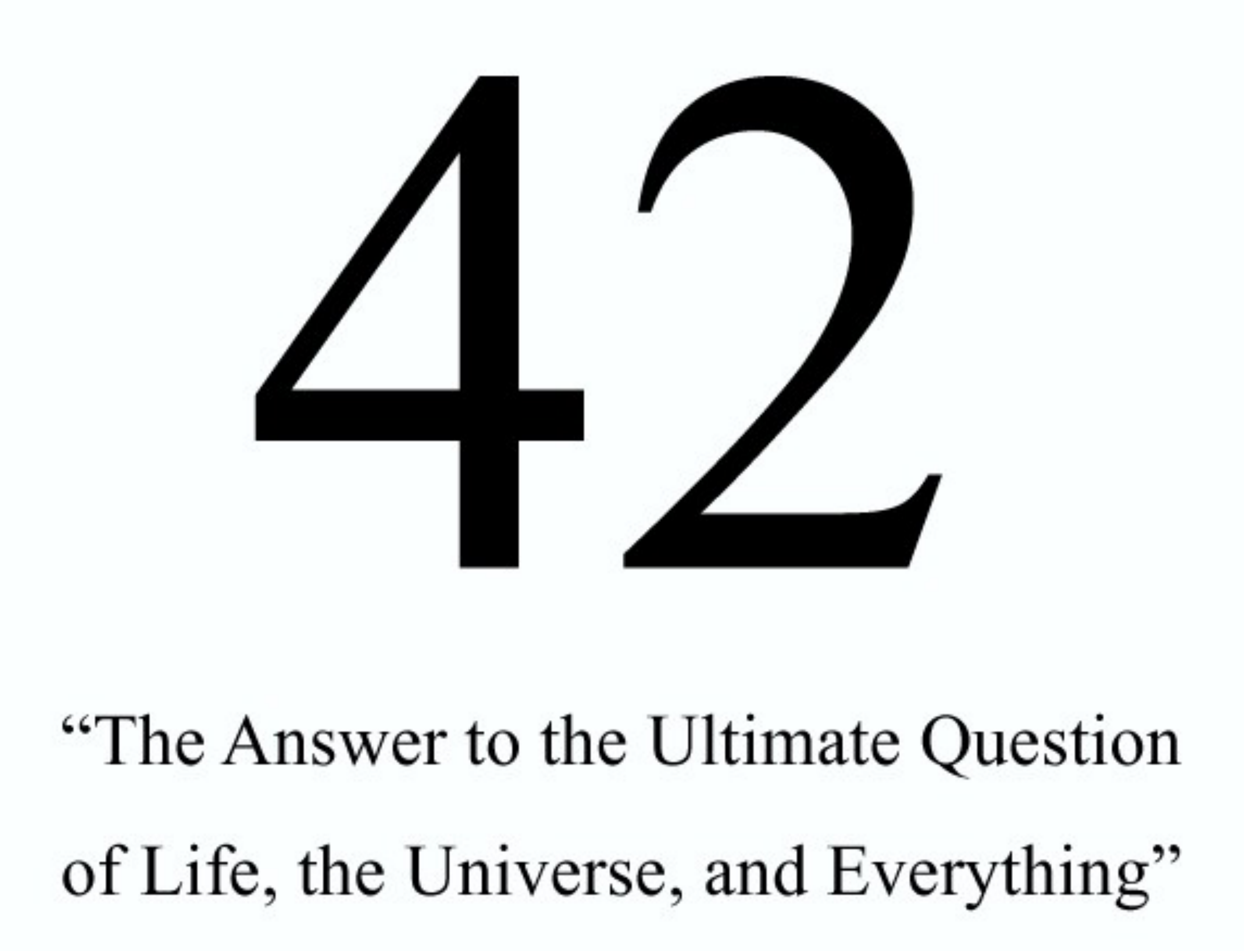
I do not follow my passion. I have spent years trying to find my passion. After failing over and over again, I learned one must not choose a career purely based on his passion. It was a hard and bitter lesson, and it goes opposite to the usual motivational posts and common-sense. So what you do then. Read what says:

How to Figure Out What Your Passionate About

Finding fulfillment and success through micro-motives.

How I Learned

1. If I don’t understand something from one place, I quit and go to a second place. Rather than working hard on the same article or blog-post or video for hours, I focus on working hard for the topic at hand and that makes me flexible. I use another resource and then another till I get the concept/idea.
2. I practice problems. We can’t learn math by reading and understanding. We need to apply it to the problems. mathisfun.com has a list of problems with answers. This is what I used to practice Bayes’ Theorem


42
“The Answer to the Ultimate Question of Life, the Universe, and Everything”

42 Answer by Mbartelsm, Source : Wikimedia

I think I found my 42. This method worked for me. It might or might not work for you but you won’t know this unless you try it for a week. I have read hundreds of blog posts on how to learn math for data science and many did not work but some did. In the end, I found my own path. I did not find my path to learning by just thinking. I tried many and failed many times. So, you gotta keep on trying till you succeed. Give a new approach a few days or more but not a few weeks or months. One week is fine.

Probability

I started with Bayes’ Theorem but I ended up binge-watching, binge-reading, and binge-practicing many concepts from Statistics and Probability. Here are the resources I used for learning conditional probability and Bayes′ Theorem:

1. Eddie Woo's discrete random variable. Total 3 videos (this includes expected value)
2. Permutations and combinations from Eddie Woo
3. Permutations and combinations from Mario's Math Tutoring
4. Bayes′ Theorem from Math is fun
5. Conditional probability, Bayes′ Theorem and others from Investopedia
6. Probability distributions from zedstatistics (it explains in terms of gradient)
7. Probability density function (PDF) from Explained by Michael (explains the same in terms of algebra and graphs)
8. Cumulative Distribution Function (CDF) from Explained by Michael
9. Discrete probability distributions from Jason Gibson of mathtutordvd.com (the best video on what is a discrete probability distribution)
10. An excellent StackExchange post on PDF vs PMF
11. Math insight link on the idea of PDF from StackExchange post I mentioned above
12. MIT OCW lecture on PDF (it is mentioned in the StackExchange post)

Now I can explain all about PDF with the Feynman technique :-)

I am not the only one who understood this principle of learning. has come up with a similar plan in his YouTube video:

Ken Jee on YouTube

An excellent video if you are beginning your data science journey. Might save you many months of suffering. Go ahead, watch it and then come back here

Statistics and Linear Regression

Lastly, while I was writing this blog post during the Xmas holidays, I also binge-watched these Statistics Fundamentals by of StatQuest fame.

His Linear Regression and Linear Models playlist is what I am watching right now. The guy is great at explaining stuff, he does not waste any time, he keeps things to-the-point and makes sure he reviews before moving ahead, and all that with hardly any code. He strives for clarity and fundamentals which is the whole point of learning anything anyway. Josh has the best introduction to logarithms and linear regression I have come across so far. And you will love his BAMs and tiny bam and triple BAMs :-)

To remind you, An Introduction to Statistical Learning mentions Linear Regression as a prerequisite. So I thought it will be a good idea to get through this before I pick-up the book. That book is known as almost the bible of machine learning algorithms.

Combating The Fear Of Math

This is a biggie. There are a lot of learners who fear math. Even though we need not know much math but the fear of math still does not let us understand and grasp whatever topics we need to learn. They think they don’t have a mathematical mind. **Being a genius like Georg Cantor and creating mathematical entities and being able to understand and use math as a tool/model for solving problems are two very different things, the former is a gift from the Universe (or God) while the latter is a skill-set.** I do understand that neither of us is a genius and nor we are excellent Harvard or Oxford graduates. We can’t do anything about this limitation. And definitely, we can do something about “attitude” and “capacity” to acquire math as a skill set. We can inculcate mathematical thinking as a part of our character. Check these videos out to change your beliefs about math and what you can or can not learn:

List A:

Any 10 videos you like from The Math Sorcerer (I have watched 30+). Start with these:

1. Three Tips For Learning Math on Your Own
2. 6 Little Known Reasons Why Self Study is the Key to Success in Math

3. Why Do Some People Learn Math So Fast
4. How to Overcome Failure in Math

List B:

1. What does it take to learn math? To live a life? | Miroslav Lovric
2. Anyone Can Be a Math Person Once They Know the Best Learning Techniques | Po-Shen Loh
3. How you can be good at math, and other surprising facts about learning | Jo Boaler
4. The interesting story of our educational system | Adhitya Iyer

Watch the last one only if you are interested in knowing how the Indian education system works. That is what I studied, so I have a bit of bias to include it here. It is an interesting video by the way.

List C:

Pick up a math topic you always wanted to learn, go to math is fun and read it, work through all the exercises. Trust me you will lose half of the fear right away by doing this. The explanations have been made so simple, easy, and basic that you see through math no matter what is your age or background.

How Not To Forget What You Learn

While you learn all the above by reading and watching and solving problems. You will soon forget 80–90% of it in a week or so. The only ways to make learning permanent are:

1. Use it daily in your work
2. Revise with a fixed schedule

While #1 may not be possible for you are busy building data science projects. Same I can say for #2. yes, you can revise like I put up a fixed schedule in my school years:

- Revise what you learned this week by end of the week
- Revise every week + previous weeks
- Revise whatever you learned at the end of the month
- Revise every month + previous months

This worked in school but now in professional work environments, it does not. The only method that works with me now is method #1. The curse of self-study to become a data scientist is that you can't use everything you have learned. So I had to devise method #3 using **the Feynman Technique**.

- Once you learn a topic. Use the Feynman technique next
- Put the heading/title of the topic on a list
- At end of the week, check your list and use the Feynman technique to explain all topics on the list

Image by John Hain from Pixabay

I don’t think you will need monthly revisions.

Benefits Of This Approach

This approach of **Binge-* + Feynman technique** has several benefits:

- You don’t need to wait long. You save a lot of time because you are not reading an entire book on math or doing a MOOC, both of which require months.
- You learn only what you need to. Data science is not mathematics. Don’t forget the industry, business value, portfolio preparation, GitHub presence, business stakeholders, and storytelling using data.

You can’t afford to replace those with “learning mathematics comprehensively”.

- Your focus remains on the real-work
- You learn how to explain. A very useful skill in being able to put your point across in your workplace while respecting everyone around you. It is beneficial in interviews too.
- Since you got the fundamental idea behind certain math topics, you can explore and learn in detail later in your free time when you are not hard-pressed for a deadline. After you are employed, you can even make a 3 or 5-year plan to master Calculus if it sparks your curiosity or if your domain demands expertise in it.

These benefits look tiny but these can be major factors in whether you are going to succeed or not.

Image by Montanasuffragettes, Source: Wikimedia

Epilogue

I wish you luck in your learning and I hope you keep on persisting. Data science is hard but certainly within your reach. It might take time but all worthwhile careers take time.

The year 2020 will go down in history as the year of the pandemic, the year of lockdowns and masks, the year which shook the building blocks of nations across this planet. It spared no one, neither employees nor employers, neither governments nor the public, neither black nor white, neither God-loving nor an atheist. It was the first time in my life I have seen such fear and havoc at an international level. It reminded me of an ancient Chinese saying **“under the heavens, we all are but one family”**. We suddenly were thrown into a dark age. It was as if some dystopian science-fiction was coming to life.

While this is a bleak picture, for the first time in history, such a dark event brought scientists all across the world under one united front: to bring out humanity from this peril. Numerous scientific minds across the globe worked tirelessly to create a vaccine. Finally, not one but two vaccines have been produced, and more to come in 2021. If we have the capacity to bear and come out of this pandemic, then the fear of math is just a tiny thing to handle for the capacities of the human mind. Let’s forge into 2021 with a conviction that *“I will break-down any obstacle when it comes to learning data science”*. You need to own this. There are very few things in this world that are impossible, learning math for data science, honing your soft-skills, and crafting an impressive data science portfolio are not among those. **May The Force Be With You**