The Two Most Important Software That Top Data Scientists Must Know

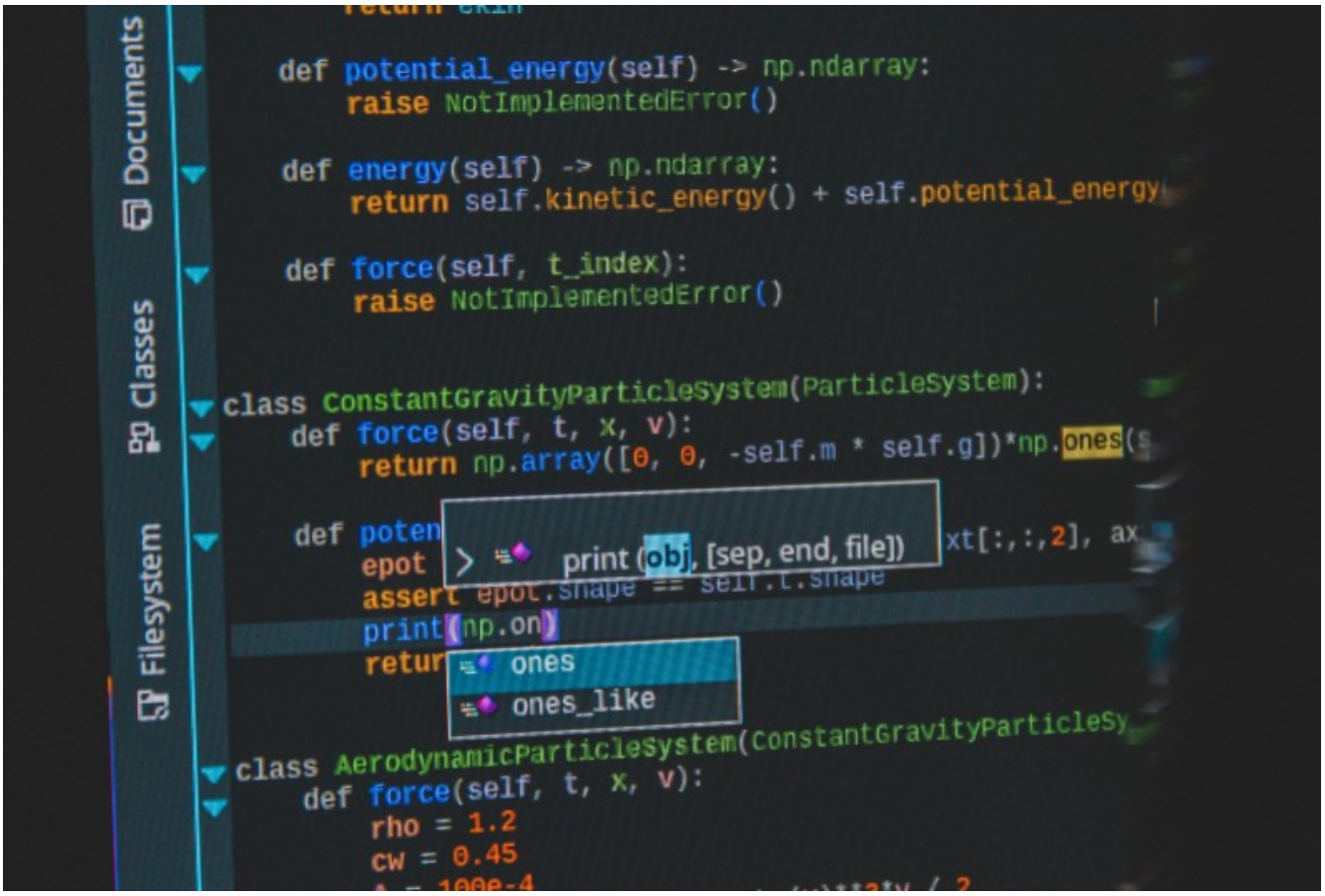

Photo by Árpád Czapp on Unsplash

Some years ago when I was a university student, I was discussing with some friends and professors which are the best software out there for scientific multitasking, writing, and presentation.

After discussing for a while we arrived almost unanimously at the conclusion that if someone has the money to pay for Wolfram Mathematica (or simply Mathematica) license, then there is no reason to use other software other than Mathematica for scientific multitasking. Almost Ever! On the other hand, if someone wants to write and present scientific results at high-quality level, then LaTeX is the number one choice in every respect.

Many software products allow the user to do scientific multitasking and presentations, but the software that I discuss below are, in my opinion, the best for these tasks. Clearly, in this article I present my own opinion about these software and why they are the best, so, I am fully aware that other people might not agree with my shortlist. On the other hand, my opinion is based on observations of many people during the years that are at top positions in their research careers.

In this article, I want to discuss why the software below are important in the context of data science and machine learning and why top data scientists and machine learning researchers use these software in the majority of cases (80%-90%). In this article I am giving my personal thoughts as a user of these software and I am not promoting them but rather suggesting to integrate them in your data science projects.

I. Wolfram Mathematica

Wolfram Mathematica was conceived and developed in the mid ’80 by theoretical physicist Stephen Wolfram. The main reason that motivated Stephen Wolfram was that at that time there ware not any software capable of performing symbolic and numerical calculations in particle physics, where usually one has to simulate and calculate different quantities related to particle interaction. Since then, Mathematica has become the standard software in the scientific community in academia and research institutes.

If you visit the official website of Wolfram Mathematica, you will immediately read on the first page the following statement:

> Wolfram Mathematica—The world’s definite system for modern technical

computing

To make the above statement, namely, by claiming that **Mathematica is the world's definite system**, it means that Wolfram Mathematica researchers have the guts to make such a claim. In reality, what Wolfram Mathematica company claims about its product is not just a claim but it is a fact.

Suppose that you are a researcher in a STEM field or a data scientist or a machine learning researcher and ask yourself the question: What characteristics does a top multitasking scientific software must have? While I don't know what is your personal answer, but for me, the main characteristics must be:

1. All in one—All integrated.
2. High efficiency in performing symbolic and numerical calculations.
3. Minimal and easy code writing.
4. Easily accessible documentation.
5. High-quality data, figures, and image visualisation.

Now let me explain why Mathematica satisfies all five characteristics that have been enumerated above. Independently what is your scientific expertise, one of the most important features of scientific software is the all-integrated feature, namely the software has built-in a place where you can perform almost everything you need. In the case of Mathematica, that place is the Mathematica notebook as shown in Fig. 1.

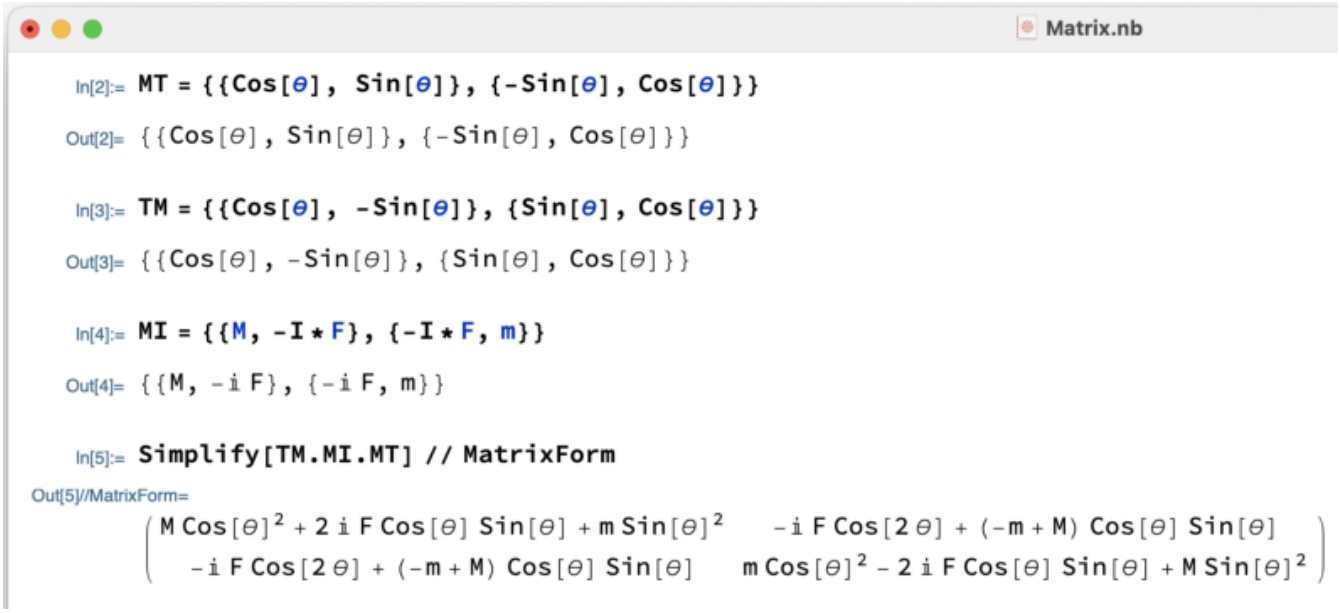

Fig. 1. Typical appearance of a Mathematica notebook while performing calculations. Figure created by the author for educational purposes.

Mathematica notebook essentially has all functionalities that the user needs to perform scientific calculations. It allows the user to directly write and manipulate text, write and run code, create and visualise dynamical graphics, etc. All these functionalities are just in one place, the Mathematica notebook.

The next two important characteristics in my list above are efficient symbolic and numerical calculations and minimal and easy code writing. Well, if you see the calculations presented in Fig. 1, then you can easily realise that Mathematica satisfies both of these characteristics.

For example, in the code shown in Fig. 1, I performed symbolic calculations to evaluate the product of three 2x2 matrices and showed the result in Out[5]. All these calculations are performed very fast and efficiently and in just four code lines! In addition, if you put the cursor on the name of the function which you want some information (MatrixForm for example), a pop-up window will appear to give the user the possibility to click on the information button. This functionality

explains my fourth point in my list above about easily accessible documentation.

What can someone do with Mathematica in Data Science and Machine Learning? In practice, one can do almost everything in these fields in few code lines. For example, with Mathematica, one can do the symbolic computation, mathematical computation, numerics, algebraic manipulation, data analysis, graph computation, interactive computation, etc. In connection with data science, Mathematica has all necessary functionally for statistics and data analysis.

At this point, I believe it is better to give a specific example to show the simplicity and efficiency of Mathematica in **machine learning**. In a previous article, I discussed various ways of how someone can perform linear regression with different Python modules. If I use the same data as used in my previous article, and perform linear regression with Mathematica, I get the results shown in Fig. 2 below.

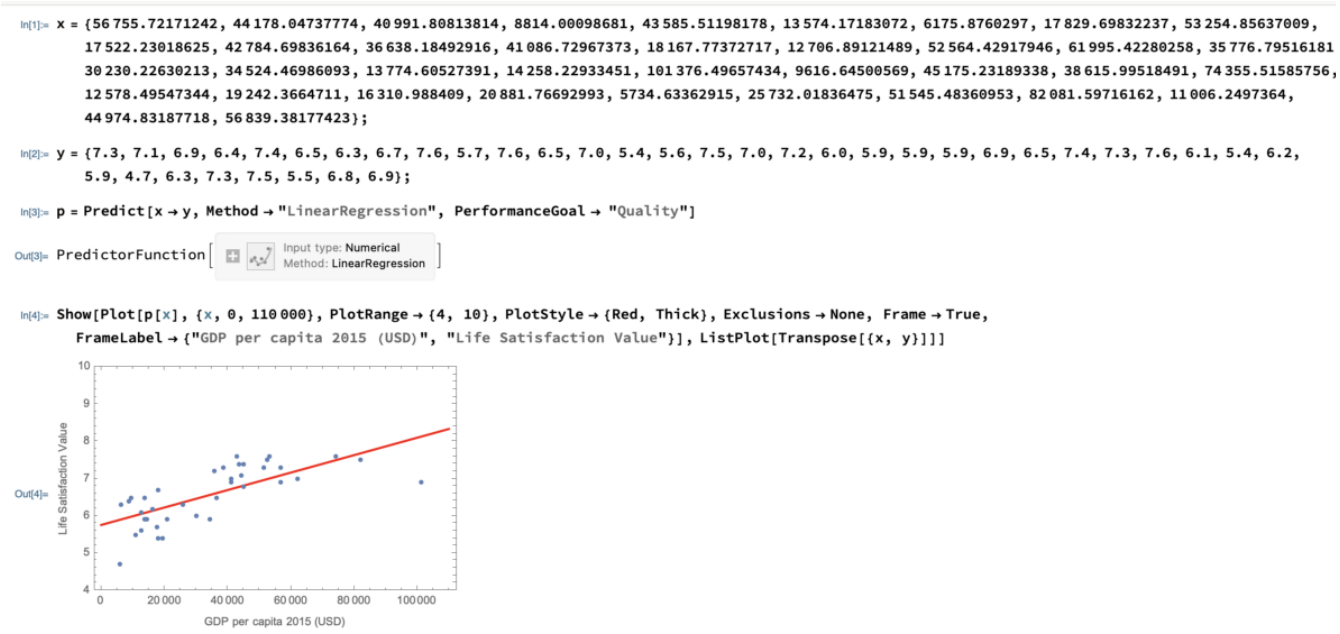

Fig. 2. Linear regression with Mathematica for the data used in previous article. Figure created by the author for educational purposes.

You can see how compact and minimal code lines are required to perform linear regression with Mathematica. To plot the result, It was necessary only four code lines. The quality of the plots is extremely high when created with the Mathematica notebook and figures can be saved in different formats. This functionality of Mathematica reflects the fifth point of my list above.

II. LaTeX

When I was an undergraduate university student, I prepared my Bachelor thesis with Microsoft Word. It was a pain! I wasted so much time just to number equations and display them correctly and after some minimal error had to restart again. Well, that situation changed when I started learning LaTeX.

LaTeX is open-source software for document preparation, and it was created in early ’80. What can I say about LaTex? Well, LaTeX is just perfect! It is very difficult to find any other software close to LaTeX for document preparation and presentation. LaTeX can be used to prepare any type of document, like articles, thesis, books, letters, memoirs, slide presentations, etc.

Since LaTeX is open-source software, during the years it has evolved a lot and it has, in my opinion, reached almost perfection. LaTeX can also be used to make high-quality presentations for meetings and conferences. This is made possible by using the Beamer package. This package allows the user to prepare high-quality and stable presentations in pdf formats. With Beamer presentations, you will forget those embarrassing moments during a conference where your Powerpoint

presentation gets completely messed up because the conference computer or the slide projector is somehow not compatible with something unknown.

Discussion and remarks

Why you should use the above software in data science and machine learning projects? The answer is simple: you want them because they save you a lot of time and give you the best results.

One of the most important things to remember about the terms “data science” and “machine learning researcher” are the “science” and “research” components of these two overlapping fields.

If you are a researcher and a scientist, you want to use the data to achieve scientific findings and present them in a clear and concise way without losing too much time on the technicalities to achieve these results. You do not want to waste the majority of the time by writing long lines of code to achieve and present your findings. You want to work with software that help minimise your working time by achieving excellent results and both software that have been mentioned above allow you to do this.

Unfortunately, many “data science” jobs nowadays are just writing code mostly in Python and have very little to do with science and research. If you would be using Mathematica in your data science job, you would realise very soon that you wasted a lot of time by using other software like for example Python. Yes, Python is a very good software but it is not Mathematica. I am pretty much confident that many Python-based data science books out there would be thinner if they were written based on

Mathematica.

If you are a data scientist and/or machine learning researcher, I would recommend learning Mathematica and use it for your data science projects and use LaTeX for document preparation and presentation. In this case, you will waste less time in code writing and appreciate the science part of data science.

Why is not Mathematica that much used in data science industry jobs? There are essentially two reasons for this phenomena. **The first reason** is due to laziness and lack of innovation of people working in most data science jobs. The majority of people working in data science have just a bachelor's degree and some of them might not have a degree at all.

During their bachelor studies, these people *had not much chance* to learn Mathematica in academia because they had not many problems to solve at this early stage. In my experience, one usually appreciates the full capabilities of Mathematica during their Ph.D. studies but not before. Because of this fact, these future data scientists go out there and start working in data science with software that lose them a lot of time in writing code, etc. The same situation very likely happened to senior data scientists that those future data scientists work with.

The second reason is related to the fact that to get Mathematica you need to pay for it, while many other software out there like Python and R are free. In fact, Mathematica is quite expensive and its price might be an issue, but you really get what you pay for.

So by concluding, my recommendation to data scientists and/or machine learning researches is to start using Mathematica and LaTeX in their projects and I am pretty sure that you hardly will look back.

If you liked my article, please share it with your friends that might be interested in this topic and cite/refer to my article in your research studies. Do not forget to subscribe for other related topics that will post in the future.

By Damian Ejlli on September 8, 2021.

Canonical link

Exported from Medium on October 8, 2021.