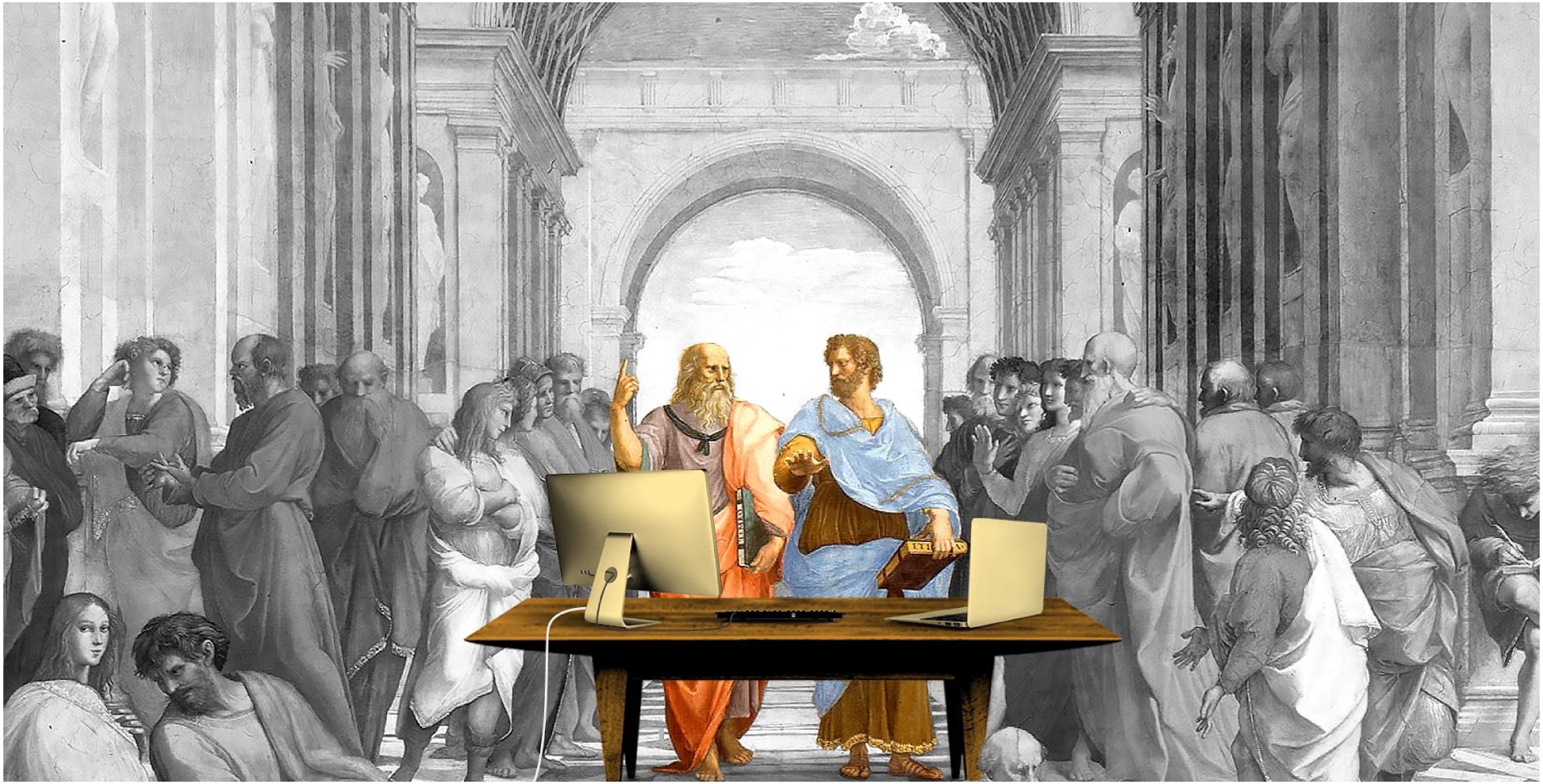
A Beginner’s Guide To Data Science

Start your self-learning journey into the world of data right now.

Oleksii... Follow
Jun 10, 2019 · 12 mi...



When Aristotle and Plato were passionately debating whether the world is material or the ideal, they did not even guess about the power of data. Right now, Data rules the world and Data Science increasingly picking up traction accepting the challenges of time and offering new algorithmic solutions. No surprise, it's becoming more attractive not only to observe all those movements but also be a part of them.

However, if you are a newcomer in this stuff, does this mean that you have a long way to go through to be an expert? Is it necessary to go through a whole slew of tries and fails before reaching total confidence of this job? Probably, yes. But with this post, I will try to ease this task for you. Today I will draw the most effective way of learning with exceptionally the most necessary steps.

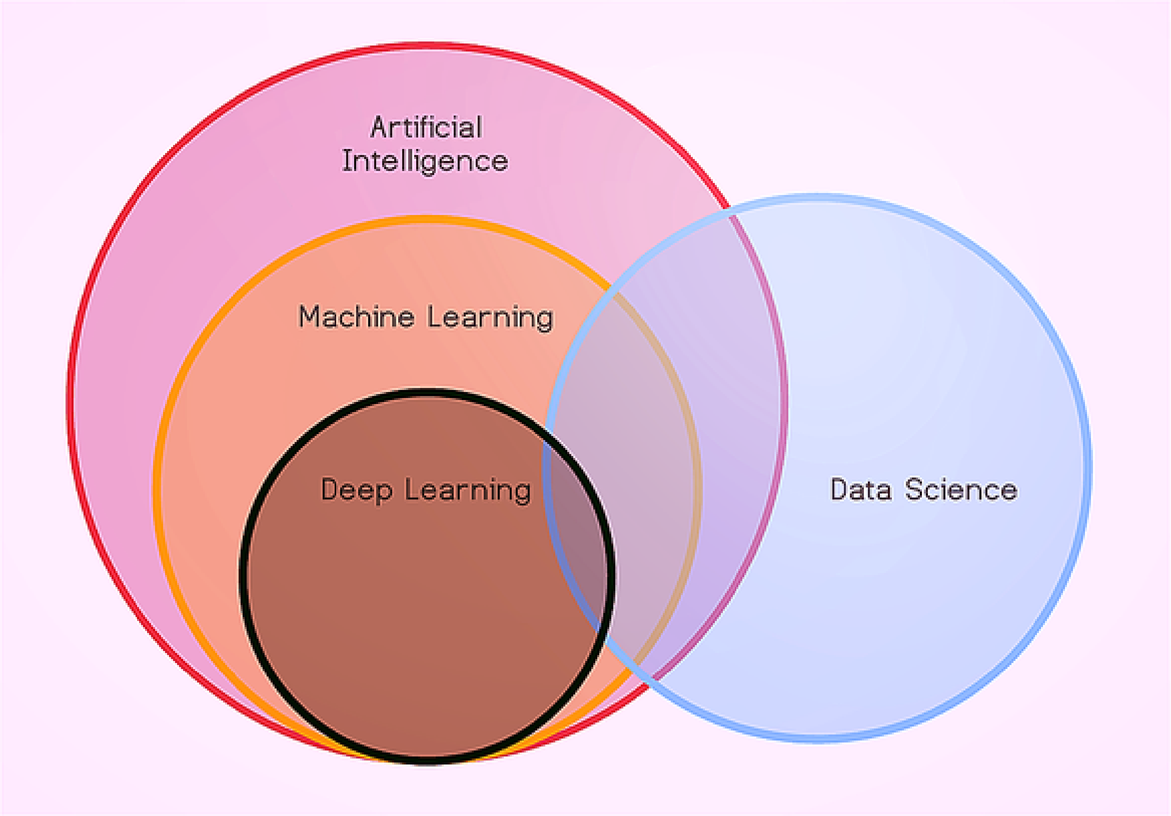
Step 0. What is What

Well, generally speaking, Data Science is not a certain or a single one realm, it's like a combination of various disciplines that are focusing on analyzing data and finding the best solutions based on them. Initially, those tasks were held by math or statistics specialists, but then data-experts began to use machine learning and artificial intelligence, which added optimization and computer science as a method for analyzing data.

This new approach turned out to be much faster and effective, and so extremely popular.

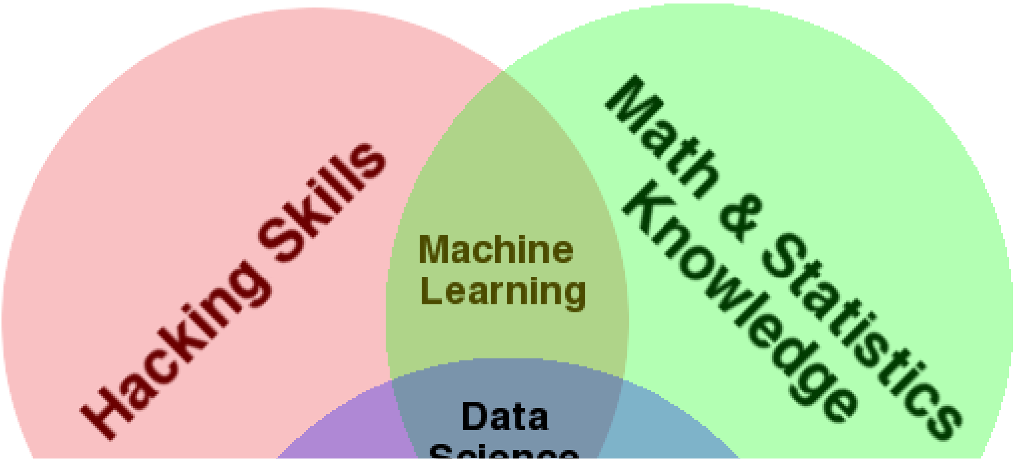
So all-in-all, the popularity of Data Science lies in the fact it encompasses the collection of large arrays of structured and unstructured data and their conversion into human-readable format, including visualization, work with statistics and analytical methods — machine and deep learning, probability analysis and predictive models, neural networks and their application for solving actual problems.

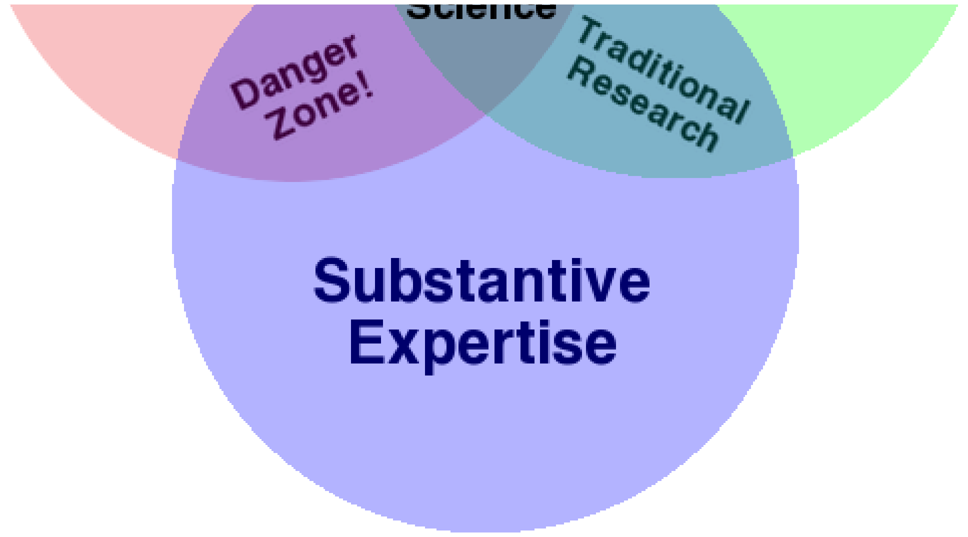
Artificial Intelligence, Machine Learning, Deep Learning, and Data Science — undoubtedly, these major terms are the most popular today. And although they are somehow related, they are not the same. So, before jumping into any of those realms, it is mandatory to feel the difference.



- **Artificial Intelligence** is the realm focusing on the creation of intelligent machines that work and react like humans. AI as a study dates back to 1936 when Alan Turing build first AI-powered machines. Despite quite a long history, today AI in most areas is not yet able to completely replace a human. And the competition of AI with humans in chess, and data encryption are two sides of the same coin.

- **Machine learning** is a creating tool for extracting knowledge from data.In ML models can be trained on data independently or in stages: training with a teacher, that is, having human-prepared data or training without a teacher, working with spontaneous, noisy data.
- **Deep learning** is the creation of multi-layer neural networks in areas where more advanced or fast analysis is needed and traditional machine learning cannot cope. "Depth" provides more than one hidden layer of neurons in the network that conducts mathematical calculations.
- **Big Data** — work with huge amounts of often unstructured data. The specifics of the sphere are tools and systems capable of withstanding high loads.
- **Data Science** is the addition of meaning to arrays of data, visualization, collection of insights, and making decisions based on these data. The field specialists use some methods of machine learning and Big Data — cloud computing, tools for creating a virtual development environment and much more. Data Science's tasks summed up well by this Venn diagram created by Drew Conway:





So what does Data Scientist do? Here is all you need to know about it:

- detection of anomalies, for example, abnormal customer behavior, fraud;
- personalized marketing — personal e-mail newsletters, retargeting, recommendation systems;
- Metric forecasts — performance indicators, quality of advertising campaigns and other activities;
- scoring systems — process large amounts of data and help to make a decision, for example, on granting a loan;
- basic interaction with the client — standard answers in chat rooms, voice assistants, sorting letters into folders.

To do any of the above tasks you need to follow certain steps:

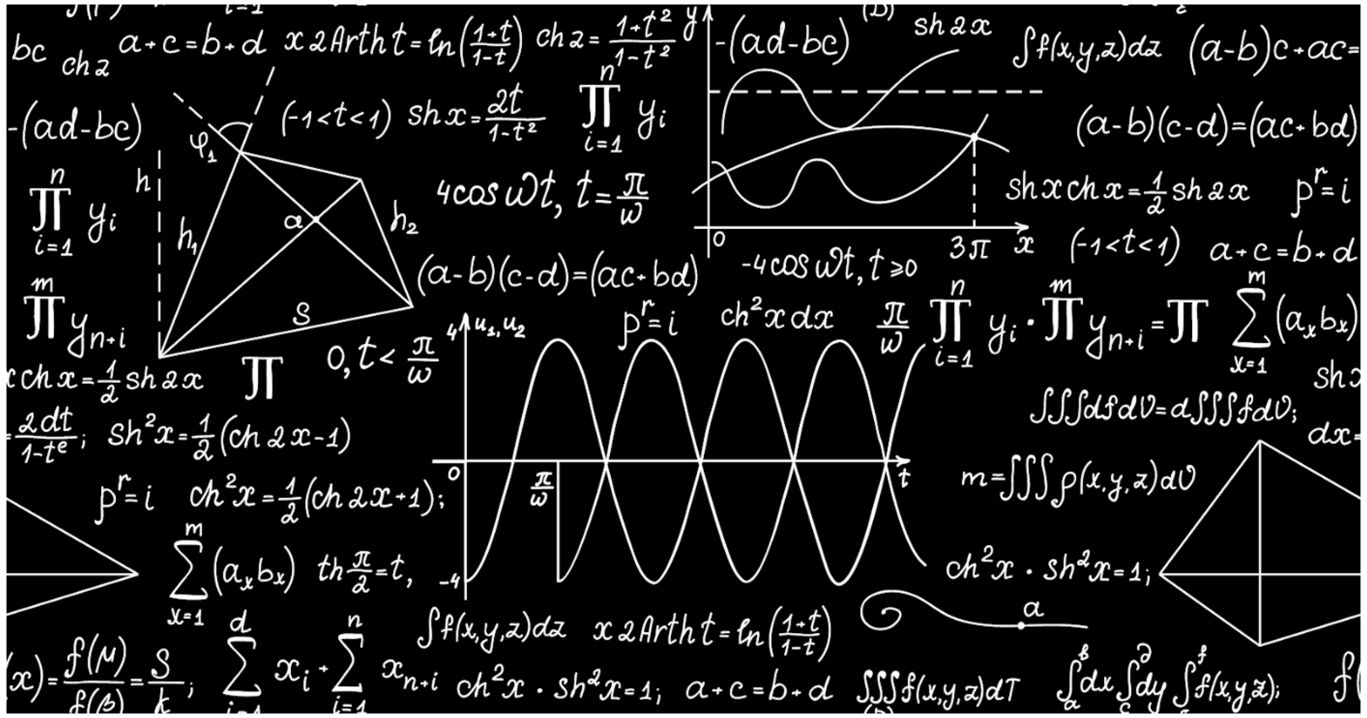
- Collection Search for channels where you can collect data, and how to get it.
- Check. Validation, pruning anomalies that do not affect the result and confuse with further analysis.

- Analysis. The study of data, confirmation of assumptions, conclusions.
- Visualization. Presentation in a form that will be simple and understandable for perception by a person — in graphs, diagrams.
- Act. Making decisions based on the analyzed data, for example, about changing the marketing strategy, increasing the budget for any activity of the company.

Right now is the time to move towards more complicated things. All of the steps below will probably seem too hard, time and energy consuming and blah blah. Well, yes, this path is hard if you perceive it as something you can learn in a month or even in a year. You should admit the fact of constant learning, the fact of making baby steps every day and be ready to see mistakes, be ready to try again and count on a long period of mastering this field.

So, are you really ready for this stuff? If so, let’s roll.

Step 1. Statistics, Math, Linear Algebra



> “Data Scientist is a person who is better at statistics than any programmer and better at programming than any statistician.”
>
> Josh Wills

If we talk in general about Data Science, then for a serious understanding and work we need a fundamental course in probability theory (and therefore, mathematical analysis as a necessary tool in probability theory), linear algebra and, of course, mathematical statistics. Fundamental mathematical knowledge is important in order to be able to analyze the results of applying data processing algorithms. There are examples of relatively strong engineers in machine learning without such a background, but this is rather the exception.

If university education has left many gaps, I recommend the book The Elements of Statistical Learning by Hastie, Tibshirani, and Friedman. In this book, the classic sections of machine learning are presented in terms of mathematical statistics with rigorous mathematical calculations. Despite the abundance of mathematical formulations and evidence, all methods are accompanied by practical examples and exercises.

The best book at the moment to understand the mathematical principles underlying neural networks — Deep Learning by Ian Goodfellow. In the introduction, there is a whole section about all the math that is needed for a good understanding of neural networks. One more good reference is Neural Networks and Deep Learning by Michael Nielsen — this may not be a fundamental work, but it will be very useful for understanding the basic principles.

Additional resources:

- A Complete Guide To Math And Statistics For Data Science: cool and not boring walkthrough to help you become well-oriented in the realms of math and statistics
- Introduction to Statistics for Data Science: This tutorial helps explain the central limit theorem, covering populations and samples, sampling distribution, intuition, and contains a useful video so you can continue your learning.
- A comprehensive beginners guide to Linear Algebra for Data Scientists: Everything you need to know about Linear Algebra

- Linear Algebra for Data Scientists: Amazing article to dive into a quick run through of the basics.

Step 2. Programming (Python)

In fact, a great advantage would be to immediately get acquainted with the basics of programming. But since this is a very time-consuming process, you can simplify this task a bit. How? Everything is simple. Start learning one language and focus on all the nuances of programming through the syntax of that language.

But still, it is difficult to do without some kind of general guide. For this reason, I recommend paying attention to this article: Software Development Skills for Data Scientists: Amazing article about important soft skills for programming practice.

For example, I would advise you to pay attention to Python. Firstly, it is perfect for beginners to learn, it has a relatively simple syntax. Secondly, Python combines the demand for specialists and is multifunctional.

But if these statements don't tell you anything, read more about it here: Python vs R. Choosing the Best Tool for AI, ML & Data Science.

Time is a precious resource, so it's better not to disintegrate at once and not just waste it. So how to learn Python?

If you don't have any programming understanding, I recommend reading Automate the Boring Stuff With Python. The book offers to explain practical programming for total beginners and teach from scratch. Read Chapter 6, "String Manipulation," and complete the practical tasks for this lesson. That will be enough.

Here are some other great resources to explore:

- Codecademy — teaches good general syntax
- Learn Python the Hard Way — a brilliant manual-like book that explains both basics and more complex applications.
- Dataquest — this resource teaches syntax while also teaching data science
- The Python Tutorial — official documentation

After you learn the basics of Python, you need to spend time getting to know the main libraries.

Here is a list of recommendations for studying libraries. Here I have divided all the necessary libraries for their intended purpose, and also provided all the necessary links for mastering (documentation and guides):

Main libraries:

- Numpy — documentation — tutorial
- Scipy — documentation — tutorial
- Pandas — documentation — tutorial

Visualization:

- Matplotlib — documentation — tutorial
- Seaborn — documentation — tutorial

Machine learning and deep learning:

- SciKit-Learn — documentation — tutorial
- TensorFlow — documentation — tutorial
- Theano — documentation — tutorial
- Keras — documentation — tutorial

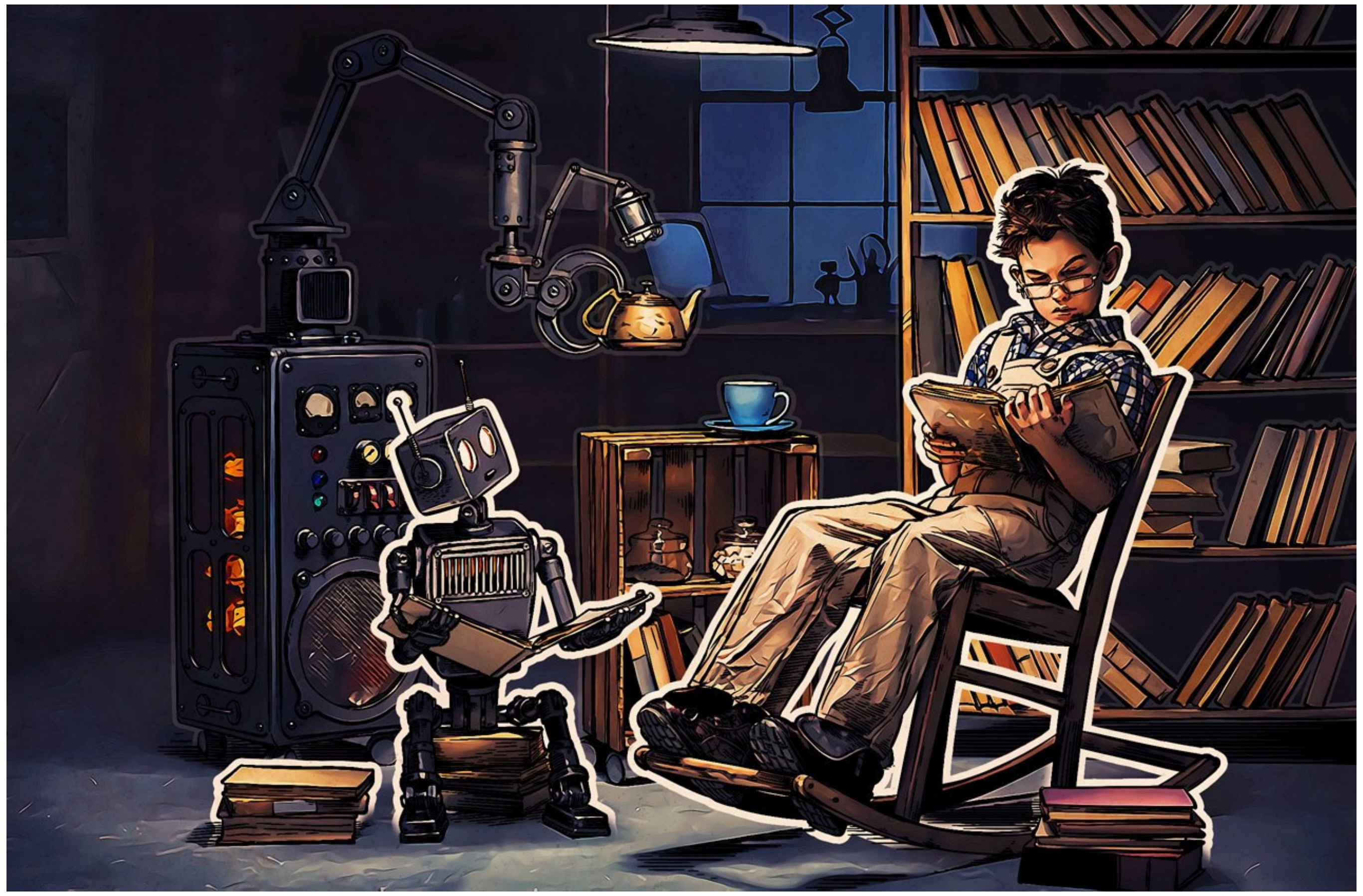
Natural language processing:

- NLTK — documentation — tutorial

Web scraping:

- BeautifulSoup 4 — documentation — tutorial

Step 3. Machine Learning



Machine learning allows you to train computers to act independently so that we do not have to write detailed instructions for performing certain tasks. For this reason, machine learning is of great value for almost any area, but first of all, of course, it will work well where there is Data Science.

First thing or the first step in learning ML is its three main groups:

1) Supervised Learning is now the most developed form of ML. The idea here is that you have historical data with some notion of the output variable. Output Variable is meant for recognizing how you can a good combination of several input variables and corresponding output values as historical data presented to you and then based on that you try to come up with a function which is able to predict an output given any input. So, the key idea is that historical data is labeled. Labeled means that you have a specific output value for every row of data, that is presented to it

PS. in the case of the output variable, if the output variable is discreet, it is called CLASSIFICATION. And if it is continuous it is called REGRESSION

2) Unsupervised learning doesn't have the luxury of having labeled historical data input-output. Instead, we can only say that it has a whole bunch of input data, RAW INPUT DATA. It allows us to identify what is known as patterns in the historical input data and interesting insights from the overall perspective. So, the output here is absent and all you need to understand is that is there a pattern being visible in the unsupervised set of input. The beauty of unsupervised learning is that it lends itself to numerous combinations of patterns, that's why unsupervised algorithms are harder.

3) Reinforcement learning occurs when you present the algorithm with examples that lack labels, as in unsupervised learning. However, you can accompany an example with positive or negative feedback according to the solution the algorithm proposes. RL is connected to applications for which the algorithm must make decisions, and the decisions bear consequences. It is just like learning by trial and error. An interesting example of RL occurs when computers learn to play video games by themselves.

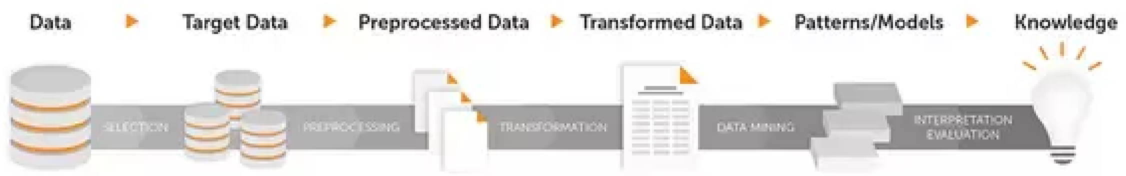
So okay, now you know the basics of ML. After this, you obviously need to learn more. Here are great resources to explore for this purpose:

- Supervised and Unsupervised Machine Learning Algorithms: Clear, concise explanations of the types of machine learning algorithms.

- Visualization of Machine Learning: Excellent visualization that walks you through exactly how machine learning is used.

Step 4. Data Mining and Data Visualization

Data Mining is an important analytic process designed to explore data. It is the process of analyzing hidden patterns of data according to different perspectives for categorization into useful information, which is collected and assembled in common areas, such as data warehouses, for efficient analysis, data mining algorithms, facilitating business decision making and other information requirements to ultimately cut costs and increase revenue.



Resources to master Data Mining:

- How data mining works — great video with the best explanation I found so far
- ‘Janitor Work’ is Key Hurdle to Insights: Interesting article that goes into detail regarding the importance of data mining practices in the field of data science.

Data Visualization is a general term that describes an effort to help people understand the significance of data by placing it in a visual context.

Resources to master Data Visualization:

- Data visualization beginner's guide
- What Makes a Good Data Visualization

Step 5. Practical Experience

Studying only the theory is not very interesting, you need to try your hand at practice. Data Scientist's beginner has a few good options for this:

Use Kaggle, a website dedicated to Data Science. It constantly hosts data analysis competitions in which you can take part. There are also a large number of open data sets that you can analyze and publish your results. In addition, you can watch scripts published by other participants (on Kaggle, such scripts are called Kernels) and learn from successful experience.

Step 7. Qualification Confirmation

After you have studied everything you need to analyze the data and try your hand at open tasks and contests, then start looking for a job. Of course, you will say only good things, but you have the right to doubt your words. Then you will demonstrate independent confirmations, for example:

Advanced profile on Kaggle. Kaggle has a ranks system, you can go through the steps from beginner to grandmaster. For successful participation in competitions, the publication of scripts and discussions, you can get points that allow you to raise the rating. In addition, the site shows in what competitions you participated, and what are your results.

Data analysis programs can be published on GitHub or other open repositories, then all interested can get acquainted with them. Including representatives of the employer, who will conduct an interview with you.

Final Advice: Don’t Be a Copy of a Copy, Find Your Own Way

Now anyone can become Data Scientist. There is everything you need for this in the public domain: online-courses, books, competitions for gaining practical experience and so on. It’s good for the first glance, but you shouldn’t learn it just because of hype. All we hear about Data Science it is unbelievably cool and it’s the sexiest job of the 21st century. If these things are the main motivation for you, nothing ever will work. Sad truth yes and maybe I’m exaggerating a little bit but that’s kind of how I feel about it.

What I’m going to say right now is becoming a self-taught Data Scientist is possible. However, the key to your success is a high motivation to regularly find time to study data analysis and its practical application. Most importantly, you have to learn to get satisfaction in the process of learning and working. Think about it.

Good luck!

Feel free to share your ideas and thoughts.

Check out my blog on Medium and Instagram.

Data Science
Machine Learning
Data
Data Scientist
Towards Data Science