Intellectual Property Rights for Data Scientists

IP Law 101, so to say

Andreas Stöffelbauer Sep 9, 2020 · 9 min read ★

Photo by Luke Michael on Unsplash

Data Scientists use software they didn't write and data they don't own pretty much all the time. It is only thanks to open source that they can use programming languages like Python and R, or libraries like Scikit-Learn and TensorFlow, or databases like SQLite and MongoDB. This should not at all be taken for granted. In fact, given how important and ubiquitous intellectual property is in the data science world, it is not being discussed enough I believe. This is why I wrote this blog post.

I'll try to answer questions such as:

- Which intellectual property rights apply to data science?
- Can you copy random public GitHub code, and can others use yours?
- Who owns data and the databases it is stored in?
- Why did nobody patent neural networks?

Let's start with a different question, Why do intellectual property rights even exist? This is not uncontroversial. In fact, there is debate about whether they even should exist to such a large extent.

> The main social purpose of the protection of intellectual property is to encourage and reward creative work. It is believed that IP rights stimulate investment (time and money) and lead to innovation.

The three main concepts of intellectual property are *copyrights*, *trademarks*, and *patents*.

A **copyright** grants the author of an original piece of work ownership and exclusive rights for (commercial and non-commercial) distribution. The right exists from the moment of creation, meaning that it need not be registered, although it can be. The work, however, must be fixed in a tangible medium (eg on paper or as bits on a computer), so a thought running through your head is not protected. Besides obviously books, paintings, and music, computer software is also protected by copyright. When it

comes to data, things get more complicated (but more on that later). Importantly, however, it is only the expression itself that is protected and never the idea behind it. Copyrights are valid until at least 50 years after the death of the author.

A **trademark** is a sign that distinguishing one company (or goods thereof) from others. An organization's logo is protected by trademark, but it can also be a phrase like Nike's *Just Do It*. There is no registration required for trademarks and they may last indefinitely. It is not of much relevance to data science, however.

Patents, in contrast, do not exist automatically. They require an application. And unlike copyrights and trademarks, patents protect ideas and inventions rather than the form of expression. Which inventions receive a patent and which not is generally more art than science. A patent grants the owner the exclusive right to commercially exploit the invention, generally for 20 years. A famous patent is Swiffer's Wet Floor sheet (see the patent application here). Patenting software is possible too, but more on that later.

In addition to those three, **trade secrets** are often also considered a method for IP protection, although not a formal one. Code or data that is not made public may be considered a trade secret.

Open Source Licensing

Intellectual property can not only be commercialized by marketing the products they protect but also by giving away the right itself through **license agreements**. In other words, the owner of IP can confer their right to another party, often in exchange for a royalty fee (or sometimes even for free). Such an agreement can be given exclusively to one other party, or to many (non-exclusive licensing).

In the data science world, licensing is hugely important and basically everywhere. Remember that every piece of software is protected by copyright. That means, principally only their respective authors should be allowed to use each of Python, R, Scikit-Learn, TensorFlow, or any other software. Thankfully, open-source licensing is very common today. For an exact definition of *open source,* see here. Roughly, it's three pillars are

- the source code is publicly accessible
- the software can be used for free (there is no royalty)

- derivative works can be made of it

While everyone can simply write their own open source license for a piece of software, there are a number of standard licenses that are broadly used. The benefit is that there is no (or less) confusion about what rights they include. Open-source licenses generally fall into three categories, depending on the degree of rights the author grants:

- Public Domain
- Permissive Licenses
- Copyleft Licenses

The Public Domain (License)

When a work's copyright expires, it falls into the public domain, meaning that it has no owner anymore. Anyone and everyone can do with it what they want. There is debate about whether existing copyrights can be transferred into the public domain if an author wants to. Most say no. Hence so-called **Public Domain Equivalent Licenses** have been designed to grant the same degree of rights as the public domain. These are the most dismissive licenses, as you can imagine. They waive as many rights as possible; for example they usually don't even require you to attribute the author. The most common public domain equivalent licenses are the **Unilicense**, the **Zero Clause BSD License**, the **Creative Commons Zero License** (CC0), and the **Do What the Fuck You Want To Public License** (WTFPL). The latter pretty much sums it all up. The only thing you can't do with such a license is to claim the work is yours since it belongs to the public domain (or kind of). An example of modern software in the public domain is SQLite.

Here is the CC0 license text as an example.

> ***CC0 License***
>
> *No Copyright*
> *The person who associated a work with this deed has dedicated the work to the public domain by waiving all of his or her rights to the work worldwide under copyright law, including all related and neighboring rights, to the extent allowed by law.*

> *You can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission.*

Permissive Software Licensing

Licenses that only include minimal restrictions such as an attribution clause are called permissive software licenses. Between the various licenses that fall in this category, there are subtle (and some notable) differences. But roughly said, they all allow you free use and redistribution of the software as long as you attribute the original authors in any derivative work (the attribution clause). However, one major restriction of permissive licenses is that they do not guarantee that future versions of the software remain publicly available, i.e. the software can be later made proprietary by the author. The most common permissive software license is the **MIT License**. Also very common are the **BSD License** and the **Apache 2.0 License**.

Much of the software that data scientists use has some sort of permissive open source license, including Python, Julia, PyTorch, and many more. The projects maintained by the Apache Software Foundation deserve particular attention as they include TensorFlow, Hadoop, and Spark. Here is an example.

> ***MIT License***
>
> *Copyright © 2020 Andreas Stöffelbauer*
>
> *Permission is hereby granted, free of charge, to any person obtaining a copyof this software and associated documentation files (the "Software"), to dealin the Software without restriction, including without limitation the rightsto use, copy, modify, merge, publish, distribute, sublicense, and/or sellcopies of the Software, and to permit persons to whom the Software isfurnished to do so, subject to the following conditions: The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.*
>
> *THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND […]*

Copyleft Licensing

In contrast to permissive software licenses, so-called copyleft licenses are reciprocal because they require any derivative work to have the same license terms, i.e. to have the very same copyright license. One major motivation could be to ensure that the software

remains free, which is not guaranteed under a permissive license. The **GNU General Public License** (GPT) and the **AGPL** are the most common copyleft licenses. The most famous software under a copyleft license is Linux.

Data Scientists and Software Licenses

As you have probably realized by now, you cannot simply copy and use any code you find on GitHub. If there is a license included in a repository, check what you are allowed to do with the code. But if there is no license, the code is still protected under copyright law and you are not allowed to copy it — the fact that the code is public makes no difference. That being said, however, you are not infringing copyright for an insignificant and obvious piece of code; say a simple for-loop; often there is simply just one obvious way to code it. In addition, remember that it is only the code that is protected by copyright, not the idea behind it. So there is nothing preventing you from building a new TensorFlow (as PyTorch did) as long as you do it your own way, for example.

Usually, data scientists don't have to be overly careful when it comes to common data science tools. As long as they have some sort of open source license (which they typically do), you are good to go. Of course, there is also proprietary data science software such as MATLAB, SAS, and Tableau, but you won't find any of those on GitHub anyway. Licensing for proprietary software is typically more individualized.

How to open source your own code? To open source your own repository or project, you first have to think about which license fits your need. GitHub has a dedicated help page for that (see here). The main step is very easy: all you need to do is include your terms and condition as a license file in your repository. Again, GitHub lets you add the most common licenses with a few clicks. That's why you'll find *license.md* or *license.txt* files in many GitHub repositories.

Photo by Markus Winkler on Unsplash

Software Patenting for Data Scientists

As mentioned previously, copyright only protects the literal expression of a computer program but never the idea behind it, which is often more valuable. Therefore, many firms want the opposite of open source, i.e. to protect their code. A patent would be the logical solution. However, it's relatively uncommon to patent software. For one thing, think about the fact that a patent requires full disclosure of the source code, which would allow others to reverse-engineer it. In contrast, simply keeping code private (trade secret) is probably more effective than a patent in many cases.

What about algorithms such as neural networks? Patenting algorithms is an even more delicate issue and will become more important as AI is making progress. It is not only a question of law but particularly one of ethics.

Principally, facts are not patentable, and since machine learning algorithms like neural networks are basically mathematical methods, they are exempt from protection. However, applied to a certain problem, an algorithm may become part of a patent. So yes, if framed it in the right way, patenting an algorithm is possible. For example, a deep learning algorithm generating a certain kind of audio may be eligible. But that would not prevent the network from being applied to any other problem.

In addition, I think that the enforcement of software patents is very difficult, and the field of machine learning is moving too fast for many patents to be worth a lengthy application. Perhaps once more powerful AI algorithms arise will patents to play a bigger role. But again, there are ethical questions.

Intellectual Property and Data

Data ownership is obviously an important topic for data scientists, but it has not really been discussed a lot. Generally, copyright applies to data. However, data is distinct from software or other creative expressions and therefore deserves a closer look.

To begin with, anything protected by copyright must be creative. That excludes data that simply represent facts, such as data on weather, sports events, or stock prices — perhaps most data that exist are exempt from protection for that reason.

In addition to creativity, copyright protection always requires fixation in some tangible medium. With respect to data, this naturally means that **datasets** and **databases** play a central role. Indeed, copyrights on a database must be distinguished from copyrights on its content (i.e. the data). It is perfectly imaginable (even common) that the creative way data is combined in a database enjoys copyright protection while the data itself does not. For example, there is a copyright on the famous MNIST dataset (but the author made it available to the public domain).

However, like with software, the most effective way to protect the intellectual property of data may be simply to keep it a secret. Copyright or not, that means others are prevented from using it. Ethical issues arise again, especially when it comes to personal data.

In short, while most data is not protected by copyright, the way data is organized sometimes is. However, trade secrets probably play a more important role than copyrights when it comes to data.

Conclusion

As much as I would love to write more about intellectual property for data scientists, it's a vast topic and this blog post was only meant as an introduction. Especially the parts on software patenting and data ownership would require much more explanation. Indeed, there's so much more to say about how IP relates to machine learning and artificial intelligence, for instance. In addition, I didn't talk about whether copyrights and patents apply beyond borders. And I only touched upon some of the ethical issues. These are just some of the topics I would still like to learn more about myself.

Sign up for The Variable

By Towards Data Science

Every Thursday, the Variable delivers the very best of Towards Data Science: from hands-on tutorials and cutting-edge research to original features you don't want to miss. Take a look.

Get this newsletter

You'll need to sign in or create an account to receive this newsletter.

Data Science Intellectual Property Open Source Copyright Patents

About Help Legal

Get the Medium app