Licenses; human subjects data

Tools for Reproducible Research

Karl Broman

Biostatistics & Medical Informatics, UW–Madison

biostat.wisc.edu/~kbroman
github.com/kbroman
@kwbroman
Course web: bit.ly/tools4rr

An often neglected aspect in discussions of reproducible research: software and data need to be licensed. If you want your software and data to be reused, you need to provide an explicit license that explains exactly how the software and data may be reused.

I'm going to try to explain the issues and give suggestions about licenses to consider. But I'm no expert, and I'm definitely not a lawyer. I don't guarantee that this is entirely correct.

If you will be sharing data on human subjects, or, for that matter, just working with data on human subjects, you need to be extra careful. I'll try to sketch the basic concerns.

Course summary

- Make everything you do script-based
  - code + data $\rightarrow$ product
- Use version control (git and GitHub/Bitbucket)
- Take your time; organize
- Write clear code; make R packages
- Write unit tests
- Capture exploratory data analysis
  - what you did, saw, and thought (and why)
- KnitR + Markdown for reports
- KnitR + LaTeX for papers, talks, and posters
- Use licenses to make reusability clear

This is the last lecture in the course, so I thought I should summarize what we've talked about. And I thought I should put this first, because otherwise I might not get to it.

The central element in ensuring reproducibility: take your time and organize your work.

Invest effort now to save effort (or embarrassment) later.

Use of Markdown for talks and papers will continue to improve. A year from now, it will likely be sufficient.

Intellectual property

- Manuscripts/journal articles
- Books
- Software
- Data sets
- Ideas, inventions
- Lab/research notebooks
- Instructional materials
- Web sites

Intellectual property is property (ie, someone can own it) that is not an actual thing but more the idea of the thing. For example, it's not the actual physical book, but the text in the book. It's not an actual physical art work, but any depiction of the content of that artwork. This can get pretty complicated; it's best to move on.

Most of what academics produce is intellectual property.

IP protection

- Copyright
- Patents
- Trademarks, Trade "dress"
- Trade secrets

Different kinds of intellectual property is protected in different ways. I'm going to focus on copyright.

An important point to mention here is that an idea, fact or algorithm can't be copyrighted. Ideas and algorithms can be protected with a patent, but facts (including individual data points) can be neither copyrighted nor patented.

So, for example, copyright protection says that you may not be able to copy and use someone's exact code, but you can grab all of the ideas and re-implement them in your own way. Unless the ideas or algorithms are patented, and then you have to get their permission to use them, even if it's your own implementation.

Copyright

- Copyright is automatic
- In "works for hire," the employer holds the copyright
- In academics, it is customary that researchers control copyright
- At UW-Madison:

> *"Except as required by funding agreements or other university policies, the university does not claim ownership rights in the intellectual property generated during research by its faculty, staff, or students."*

Since 1978, works you create are automatically copyrighted. That includes data, software, papers, books, talks, posters, course syllabi, and lecture notes. Since 1989, you don't need to include a copyright notice.

In many job situations, your employer will own the copyright on works that you create as part of your employment. But in academic settings, it is traditional that researchers retain copyright on the works they create. But there are exceptions, and universities are increasingly interested in generating income from the intellectual property of their faculty.

Also, students may be treated somewhat differently from other employees. Academic staff may be treated differently from faculty, for that matter. Ideally, the exact rules are written down somewhere. Find the rules and read them, and ask questions about them.

At UW-Madison, faculty, staff, and students own the intellectual property of the works they produce except in some special cases, such as instructional materials produced with significant university resources. And if you're going to patent something, it must be through the Wisconsin Alumni Research Foundation (WARF).

Exclusive rights under copyright

- To make copies of the work
- To distribute/sell copies of the work
- To create derivative works
- To perform the work
- To display the work publicly

Copyright protection gives the author exclusive rights to the work and to derivatives of the work.

Thus, the default is that no one can copy, modify, or redistribute your code.

Fair use

Reproduction for criticism/commentary, teaching, and research

- For non-commercial or nonprofit educational purposes
- Can't be a substantial portion of the work
- Can't affect the value/market of the original work

There are important limitations to copyright protection.

We are allowed to reproduce portions of a work as part of a criticism or commentary, in teaching, or for research.

The rules aren't precise.

Quoting from a work is okay. Posting the full thing on the web is not.

Breaking copyright $\longleftrightarrow$ plagiarism

These are totally different things.

Just in case it's not clear: the tradition of citing one's sources is really totally different from following copyright law.

Works in the public domain should still be appropriately cited.

Software licenses

- Critical if you want your code to be reused.
- Also important to protect yourself from lawsuits.
- I choose between the MIT license and the GPL.
- Don't use Creative Commons licenses for software!

If you don't indicate a license for your software, others can't reuse it. You need to be explicit about whether and how your software may be reused, by providing a license.

I choose between the MIT license and the GNU General Public License (GPL). The MIT license is as open as possible: do whatever you want, just don't sue me. The GPL is "viral" (they say "copyleft") in that extends to derivative works: software that incorporates code under the GPL must also be under the GPL.

The Creative Commons licenses may work for data, but they're more for text, music, video, and such things. They should not be used for code, as they are not compatible with other standard software licenses, such as the GPL. That means that you wouldn't be able to mix in code that was licensed under the GPL.

Pick a license, any license

– Jeff Atwood

I can't emphasize this enough. If you release your software without a license, no one can modify it or incorporate it into their own software, as it's under copyright protection.

If you want your software to be reused, pick a license, and make the licensing absolutely clear.

MIT license

```
Copyright (C) <year> <copyright holders>

Permission is hereby granted, free of charge, to any person
obtaining a copy of this software and associated documentation
files (the "Software"), to deal in the Software without
restriction, including without limitation the rights to use,
copy, modify, merge, publish, distribute, sublicense, and/or
sell copies of the Software, and to permit persons to whom the
Software is furnished 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, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE
WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE
AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT
HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY,
WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING
FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR
OTHER DEALINGS IN THE SOFTWARE.
```


The MIT license basically says: do whatever you want with the software, but be sure to include this notice, and don't sue me.

Those are the key things you want: protect yourself from liability, and make plain that people can be free to reuse the software.

GPL-3

- Use, modify, distribute, …
- Don't hold the author liable.
- Distributions must include the source code.
- Software incorporating the work must also be under GPL-3.

There is an older GPL-2. Use the GPL-3. It was updated to close some loopholes.

Key additions vs MIT license: distributions of the work or derivatives must include source code, and derivatives must also be licensed under GPL-3.

For GPL-3, include this

```
<line with the program's name and a brief idea of what it does.>
Copyright (C) <year>  <name of author>

This program is free software: you can redistribute it and/or
modify it under the terms of the GNU General Public License as
published by the Free Software Foundation, either version 3 of
the License, or (at your option) any later version.

This program is distributed in the hope that it will be useful,
but WITHOUT ANY WARRANTY; without even the implied warranty of
MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
GNU General Public License for more details.

You should have received a copy of the GNU General Public
License along with this program.  If not, see
<http://www.gnu.org/licenses/>.
```


To license your software under the GPL, include a notice like this.

Creative Commons licenses

- CC0 (Public Domain)
- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

The Creative Commons licenses are really useful for things like manuscripts, data files, videos, web sites, and such.

You shouldn't use them for software, as they can't be mixed with the GPL, and they make no explicit mention of source or object code. In the FAQ at Creative Commons, they explicitly recommend against the use of CC licenses for software.

BY means people must cite you as the originator.

SA means that derivative works must be distributed under the same license (like the GPL).

ND means the work must be distributed in its entirety, without any changes.

NC means the work can't be used in a commercial setting.

CC licenses: issues to consider

- BY may be an unnecessary hassle.
- CC-BY on a paper would allow a company to include it in a book
  - but maybe you don't care
- ND is really restrictive
  - all or none
  - no modifications at all
- NC means people in a company can't use it at all
  - might not be useable within a course

There are a lot of issues to consider.

I'd recommend avoiding ND and probably also NC.

Personally, I'm going with CC0 (by academic tradition, people should still cite you) or CC-BY. It means that a company could grab my stuff and make money off of it. But I'm fine with that. I'd rather see the results of my efforts put to further use.

Data copyright

- Individual data points are generally considered facts
  - Can't be copyrighted
- Compilations of data can be copyrighted
  - Involves some creativity, so an "original work of authorship"
- But someone can just extract and reformat the data
- Can assign a license to the data files to prevent extraction and redistribution
- See `bitlaw.com/copyright/database.html`

Data are viewed as facts and so they can't be copyrighted.

Your data file or database, though, can be copyrighted, if its compilation involves some creativity, and that would generally be true for scientific data files.

So people can't redistribute your data files unless you say it's okay. But they may be able to extract and reformat the data and then distribute that.

If you want to prevent extraction and redistribution, the data files need a license, which would say the end user is prohibited from extracting data for uses other than intended.

Keep data open

- Cite the source; cite the relevant papers
- Talk to the originator of the data
  - Even if redistribution is legal, don't piss them off.
- For your own data, use CC0 (public domain)
- If you want more control, talk to a lawyer

Statisticians, in particular, should want data to be openly available. And so you should cite the source of data and any relevant papers, not just because that's the academic tradition, but also because we want to reward, as much as possible, people who make data accessible.

Even if it's perfectly legal for you to re-distribute data, you should talk to the originator of the data before doing so. You don't want them to get annoyed and then stop distributing data in the future.

If you want data to be reused, just put it in the public domain. Don't add any extra complexities, like CC-BY.

If you want to control reuse or redistribution, talk to a lawyer. It seems really complicated.

Human subjects research

- If there are humans involved, they're human subjects
  - e.g., surveys
- Human subjects research must be reviewed by an Institutional Review Board (IRB)
- Not everything is research
  - e.g., data used solely in a course
- Most things are research
  - If you publish a paper about it, it's research
- Anonymized data may be exempt
  - But the IRB wants to make that determination

Any research on human subjects must be reviewed by an IRB.

If you're considering publishing a paper about it, and if humans are involved, it's human subjects research. That includes things like surveys. So informed consent, and review by IRB, with a clearly defined protocol and protection of data.

You can do a survey within a class and it may not be research, but then you can't publish the results.

The NIH considers the analysis of anonymized human data to be "not human subjects research," and it may be exempt from full IRB review, but IRBs generally want to make such determinations themselves: you need to fill out some amount of paperwork.

HIPAA

- HIPAA = Health Insurance Portability and Accountability Act of 1996
- Special rules about medical data with any identifying information
  - Private
  - Secure
- Full zip code may be considered identifying information.
- Dates of test results are considered identifying information.

HIPAA is really important, but it's also a real pain.

The key thing is that medical data with any identifing information needs a whole bunch of paperwork if transferred/disclosed, and there need to be special security measures.

And the definition of "identifying information" is surprisingly broad.

Summary

- Pick a license, any license
- Use MIT or GPL for software
- Use CC0 for data
- Cite sources of software and data
- Talk to the source of data
- Be careful with human data
  - If you're unsure, ask for help

If you don't license your software, it can't be modified or reused.

Make data open, and be sure to reward those who make data and software accessible.

Be careful with human data, particularly if there's anything remotely identifiable.