Licenses; human subjects data

Tools for Reproducible Research

Karl Broman

Biostatistics & Medical Informatics, UW–Madison

kbroman.org
github.com/kbroman
@kwbroman
Course web: kbroman.org/Tools4RR

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 and talks and posters
- Use licenses to make reusability clear

Karl -- this is very interesting, however you used an old version of the data (n=143 rather than n=226).

I'm really sorry you did all that work on the incomplete dataset.

Bruce

Intellectual property

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

IP protection

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

Copyright

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

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."*

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

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

Breaking copyright $\longleftrightarrow$ plagiarism

Breaking copyright $\longleftrightarrow$ plagiarism

These are totally different things.

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!

Pick a license, any license

– Jeff Atwood

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.
```

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.

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/>.
```

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)

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 usable within a course

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`

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

Human subjects research

- Avoid human subjects research

Human subjects research

- Avoid human subjects research

  (just kidding!)

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

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.

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