Adopting open source practicesfor better science

Pierce Edmiston

Outline

Open source practices that make for more reproducible science:

1. Version control
2. Dynamic documents
3. Building from source

Conclusion: It's worth it!

Why I care about reproducibility

1. I want my research to be reproducible.
2. I want to attract collaborators.

Steps toward reproducible science

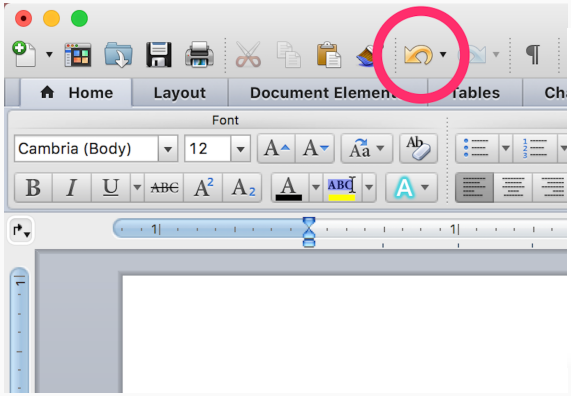
- Blind data analysts to experimental conditions.
- Improve statistics education (adapted for web).
- Hire methodological consultants.
- Seek collaboration for scalability.

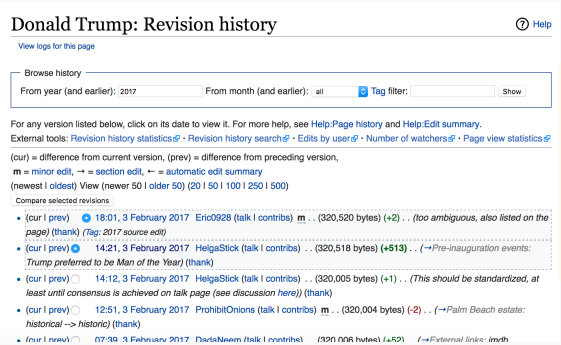
Why I think open source is the answer

Compare these two goals of reproducibility in science and in open source:

1. Fellow researchers should be able to reproduce my work.
2. Anyone should be able to use and contribute to this project.

Version control





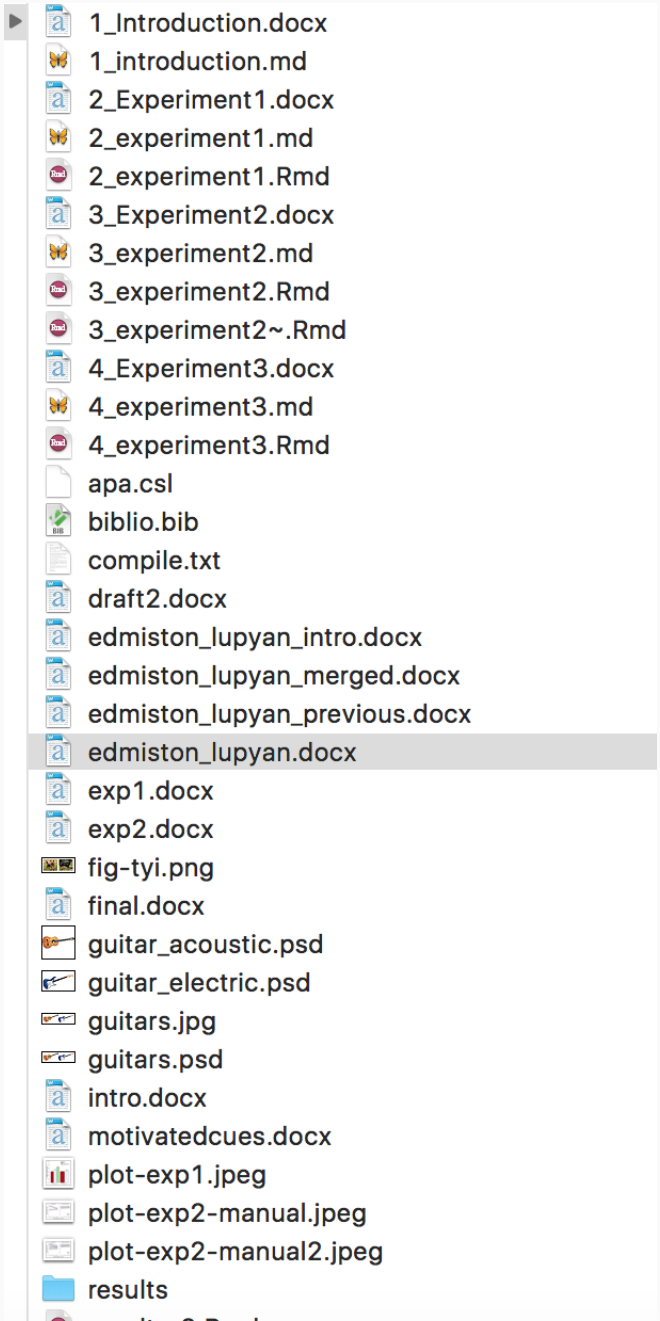

```
reproducible-research/$ git log --oneline -n 10
4a4fe06 Add rock climbing ref for version control
46e5c02 A version of the abstract that makes sense
a0cbafe Replace updated topic list with abstract
07ea24a Settled on a title and abstract
6dbf6e4 Start with dynamic documents
c847dc4 Thoughts on topics for UW RDS talk
92482d4 Merge branch 'master' of github.com:pedmiston/reproducible-research
bfbb256 Last minute stuff
c0b9d36 Fixed problem with multiple list indicators (- and +)
1f4dfd9 Cleaning up broken links in the presentation.
reproducible-research/$
```

Pick your poison

- **git**
- mercurial
- subversion
- gitless

Tools for climbing

Conquer clutter



Version control's dirty little secret

(It only really works on plaintext files.)

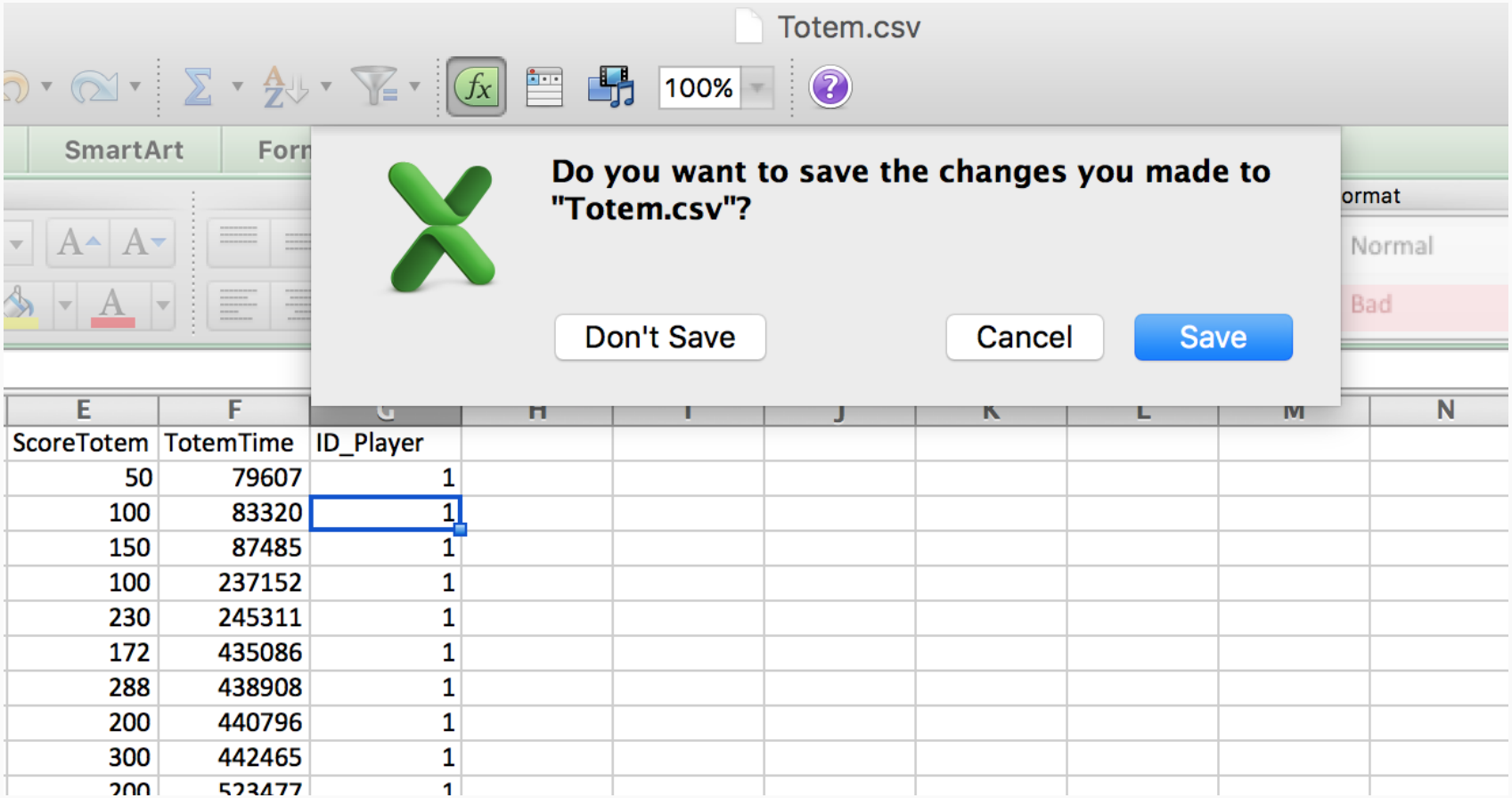


Figure 1: Excel panic. Well, did you make changes or didn't you??

The good news

Once you're working in plaintext, you can do lots of cool things.

- Full power of VCS (merge, blame, etc).
- Use free and open source tools (Unix).
- **Write dynamic documents.**

Dynamic documents

- Philosophy: DRY, Literate Programming
- Tools: Sphinx, Jupyter, Knitr, Pandoc

Philosophy

Don't Repeat Yourself (DRY) Every piece of knowledge must have a single, unambiguous, authoritative representation within a system.

Literate programming (LP) Intermingle prose and code for better understanding of the program. The explanation of a program does not need to resemble the program structure.

Knitr

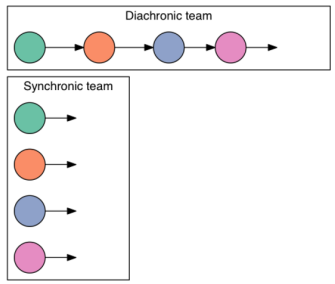
Elegant, flexible and fast dynamic report generation with R.

`Participants in condition A outperformed participants in condition B, `report_model_results(mod, param = "condition"`

```
43
44 # 1 Team structure
45
46 Experiment 1 is designed to test the hypothesis that diachronic collaboration is more
   effective than synchronic collaboration at solving adaptive problems. The type of adaptive
   problems solved by teams in Experiment 1 are classification problems. The solution to a
   classification problem involves generating predicted labels for unlabeled data using
   statistics and machine learning to improve the accuracy of the predictions. For example, a
   typical classification problem is to predict whether or not passengers on the Titanic
   survived based on their age, gender, ticket price, cabin location, and other features of
   the passenger. Classification problems are good examples of adaptive problems because there
   are many possible solutions, all varying in degree of success. For example, teams can use
   decision trees, linear regression, neural network models, and combinations of these, each
   varying in which features they are trained on and the parameters of the models. Given the
   vast space of possible solutions, a useful strategy for solving classification problems is
   to iteratively try out different solutions and incrementally improve classification
   accuracy. Although both diachronic and synchronic teams are able to iteratively develop
   solutions to classification problems, diachronic teams are hypothesized to be more
   effective than synchronic teams at utilizing this feedback to improve classification
   accuracy.
47
48 ```{r team-structure, engine = "dot", fig.cap = "Team structures. Teams of four are
   allotted the same number of total labor hours to solve a problem. Synchronic teams work all
   at the same time on a single solution. Diachronic teams work one at a time, each inheriting
   the previous solution and improving it."}
49 ```
50
51 **Procedure**. Skilled and motivated participants are randomly assigned team and condition.
   Each team is given training data and instructed to write a program that accepts unlabeled
   test data and generates predicted labels. Synchronic teams are given two hours with all
```

1 Team structure

Experiment 1 is designed to test the hypothesis that diachronic collaboration is more effective than synchronic collaboration at solving adaptive problems. The type of adaptive problems solved by teams in Experiment 1 are classification problems. The solution to a classification problem involves generating predicted labels for unlabeled data using statistics and machine learning to improve the accuracy of the predictions. For example, a typical classification problem is to predict whether or not passengers on the Titanic survived based on their age, gender, ticket price, cabin location, and other features of the passenger. Classification problems are good examples of adaptive problems because there are many possible solutions, all varying in degree of success. For example, teams can use decision trees, linear regression, neural network models, and combinations of these, each varying in which features they are trained on and the parameters of the models. Given the vast space of possible solutions, a useful strategy for solving classification problems is to iteratively try out different solutions and incrementally improve classification accuracy. Although both diachronic and synchronic teams are able to iteratively develop solutions to classification problems, diachronic teams are hypothesized to be more effective than synchronic teams at utilizing this feedback to improve classification accuracy.



Team structures. Teams of four are allotted the same number of total labor hours to solve a problem. Synchronic teams work all at the same time on a single solution. Diachronic teams work one at a time, each inheriting the previous solution and improving it.

Procedure. Skilled and motivated participants are randomly assigned team and condition. Each team is given training data and instructed to write a program that accepts unlabeled test data and generates predicted labels. Synchronic teams are given two hours with all four team members in the same room. Each team member is provided a personal work station and network access to the team's program files, but there are no constraints on how the members of the synchronic teams interact or delegate labor over the course of the two hours. Diachronic teams work one at a time, each spending two hours on the program, and are not allowed to interact with their whole team all at

Dynamic documents in practice

- Handouts
- Homework
- Supplemental materials
- Conference proceedings
- Journal papers

A litmus test for reproducible research

Can you build the published paper without the original data?