Let's `Git' started: An introduction to version control

Daiva & Luke

2015-03-29

Welcome to our Data-related workshop

Purpose:

To teach a few tips and tricks for more efficiently managing your data, tracking your computer files, understanding appropriate analytical approaches, and speeding up the process from code to tables.

Welcome to our Data-related workshop

Purpose:

To teach a few tips and tricks for more efficiently managing your data, tracking your computer files, understanding appropriate analytical approaches, and speeding up the process from code to tables.

Significance:

Topics we cover will help you get more comfortable with data, reduce the chance of overlooked errors, and give you more control over your work. They are also all important parts of a science movement gaining increasing attention – Reproducible Research.

Caveat: We aren't here to teach statistics

Need help with stats? Use these resources!

- U of T Statistical Consulting Services (click here)
- `http://www.stackoverflow.com`
- `http://stats.stackexchange.com`

Overview of some future workshops

- Know your data and how to handle it correctly: Statistical assumptions
- Fighting chaos: Coding tricks to keep your analysis – and mind – sane
- Visual exploration (of data): Techniques and code to better understand your data using plots and graphs

Overview of some future workshops

- Know your data and how to handle it correctly: Statistical assumptions
- Fighting chaos: Coding tricks to keep your analysis – and mind – sane
- Visual exploration (of data): Techniques and code to better understand your data using plots and graphs

- Code Review Club. . . ?

Notes and help during this workshop

Go to this website:

`https://etherpad.mozilla.org/dnsWorkshops`

Reproducible research


Essay

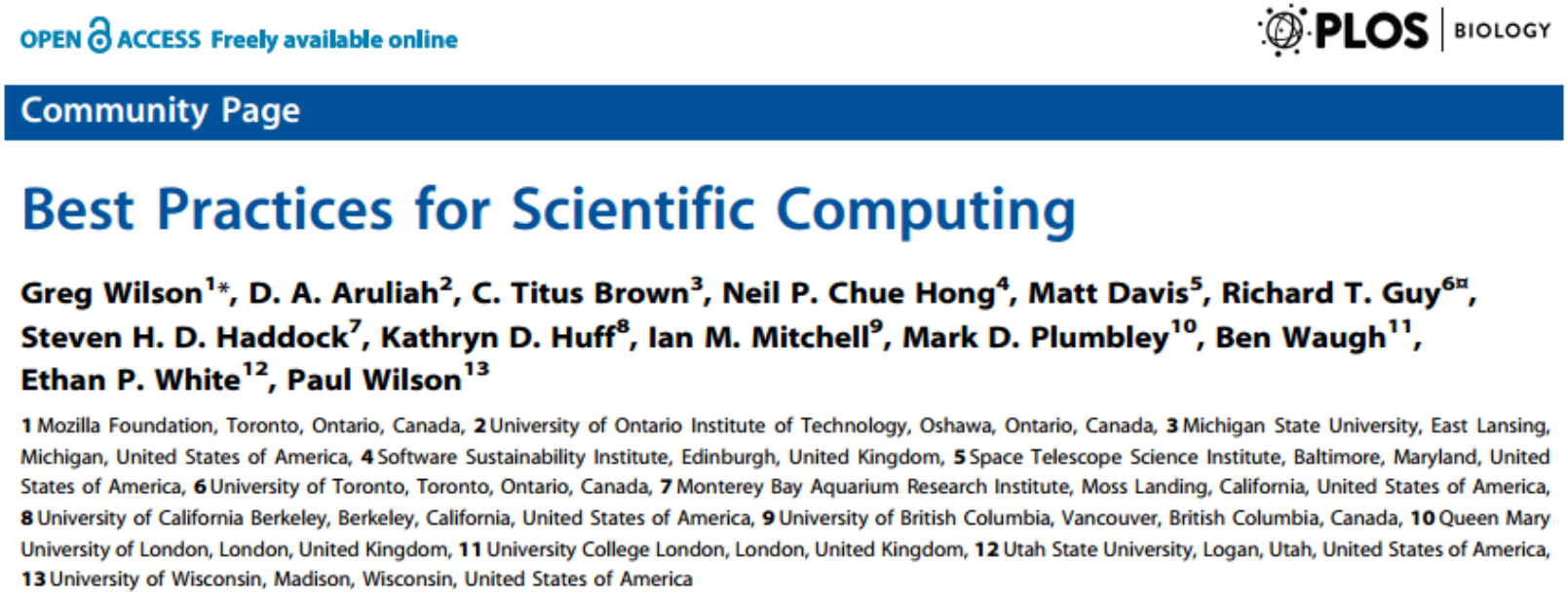
Why Most Published Research Findings Are False

John P. A. Ioannidis

Figure 1:

- http://www.nature.com/nature/focus/reproducibility/
- http://ropensci.org/blog/2014/06/09/reproducibility/
- http://biostatistics.oxfordjournals.org/content/10/3/405.full
- http://www.sciencemag.org/content/314/5807/1856.full
- http://stats.stackexchange.com/questions/14999/how-are-we-defining-reproducible-research

Reproducible research



OPEN ACCESS Freely available online

PLOS | BIOLOGY

Community Page

Best Practices for Scientific Computing

Greg Wilson[1]*, D. A. Aruliah[2], C. Titus Brown[3], Neil P. Chue Hong[4], Matt Davis[5], Richard T. Guy[6¤], Steven H. D. Haddock[7], Kathryn D. Huff[8], Ian M. Mitchell[9], Mark D. Plumbley[10], Ben Waugh[11], Ethan P. White[12], Paul Wilson[13]

1 Mozilla Foundation, Toronto, Ontario, Canada, **2** University of Ontario Institute of Technology, Oshawa, Ontario, Canada, **3** Michigan State University, East Lansing, Michigan, United States of America, **4** Software Sustainability Institute, Edinburgh, United Kingdom, **5** Space Telescope Science Institute, Baltimore, Maryland, United States of America, **6** University of Toronto, Toronto, Ontario, Canada, **7** Monterey Bay Aquarium Research Institute, Moss Landing, California, United States of America, **8** University of California Berkeley, Berkeley, California, United States of America, **9** University of British Columbia, Vancouver, British Columbia, Canada, **10** Queen Mary University of London, London, United Kingdom, **11** University College London, London, United Kingdom, **12** Utah State University, Logan, Utah, United States of America, **13** University of Wisconsin, Madison, Wisconsin, United States of America

Figure 2:

by Greg Wilson, founder of Software Carpentry (click here)

- `http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1003285`

Version control

Version control

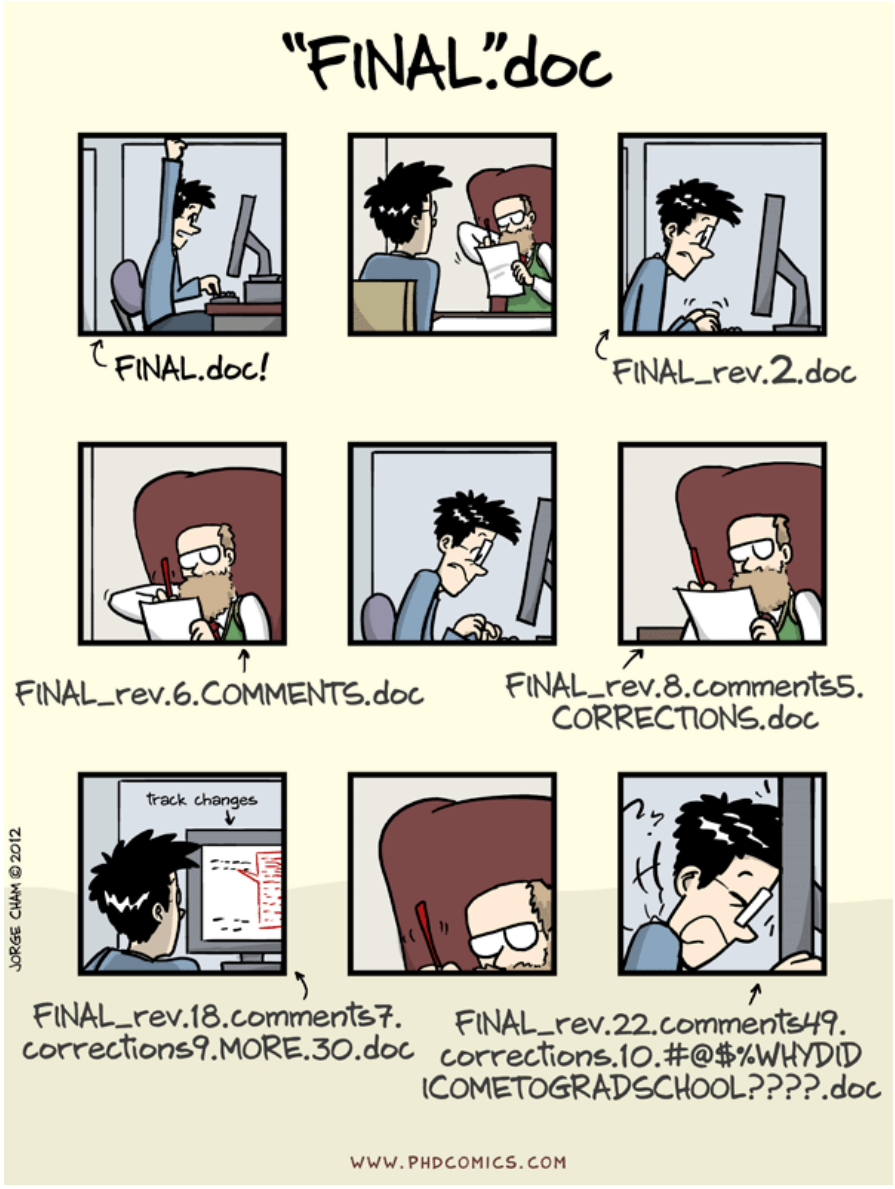


Figure 3:

What is version control[1] (VC)

- Keeps history of all changes done to files in a folder
- Like a big track changes
- Like your experiment logbook/journal (basic science)
- Can revert to previous change
- Don't have to worry about losing what you wrote!

[1] See the Git website (click here) for more detail.

What is version control[1] (VC)

- Keeps history of all changes done to files in a folder
- Like a big track changes
- Like your experiment logbook/journal (basic science)
- Can revert to previous change
- Don't have to worry about losing what you wrote!

Importance of VC!

- Future of journals and retractions
- Requests for data and code
- Transparency, scientific rigor
- Protect against accusations of fraud

[1] See the Git website (click here) for more detail.

Visualization of VC[2]

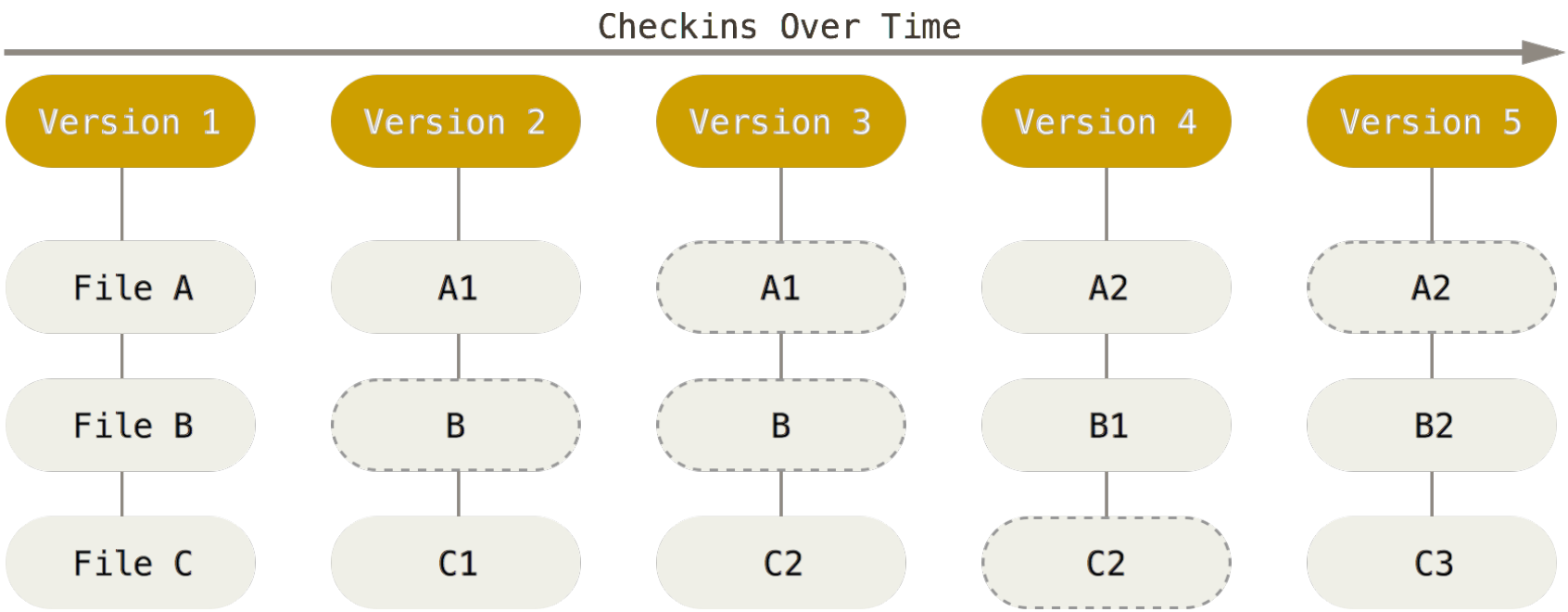


Figure 4:

[2]Taken from the Git site (click here)

Slight tangent: Filenaming rules

Slight tangent: Filenaming rules

1. Keep names short, meaningful. Remove "the", "and", "a", etc.
2. Don't include spaces.
3. Use hyphens to separate important parts of the filename.
4. Avoid redundancy in file names and folder names (e.g. `folderName/fileName-folderName.txt`, instead use `folderName/fileName.txt`).
5. If a number is included, such as for the version number, use two digits not one (e.g. V01, not V1).
6. When including a date, include it at the end of the filename and in the international standard format YYYY-MM-DD.

Before getting into Git, any questions?

Using Git

Before we start. . . The command line is **not** something to be afraid of!! Open up the terminal (Mac/Linux) or Git Bash (Windows).

Using Git

Before we start...The command line is **not** something to be afraid of!! Open up the terminal (Mac/Linux) or Git Bash (Windows).

Open up your terminal (Mac or Linux) or Git-Bash (Windows).

```
git config --global user.name "Your Name"
git config --global user.email "you@some.domain"
git config --global color.ui "auto"
git config --global core.editor "your_editor"
git config --list
```

Download our GitHub repository

GitHub[3] is a place to store your git repo for several reasons:

1. As a backup
2. To use across computers
3. To share with others

[3] Or `http://BitBucket.org`

Download our GitHub repository

GitHub[3] is a place to store your git repo for several reasons:

1. As a backup
2. To use across computers
3. To share with others

In your terminal/Git-Bash, run:

```
cd ~
git clone https://github.com/codeasmanuscript/gitWorkshop.git
cd gitWorkshop
```

[3] Or `http://BitBucket.org`

Download our GitHub repository

GitHub[3] is a place to store your git repo for several reasons:

1. As a backup
2. To use across computers
3. To share with others

In your terminal/Git-Bash, run:

```
cd ~
git clone https://github.com/codeasmanuscript/gitWorkshop.git
cd gitWorkshop
```

Check out the `cheatsheet.html` file.

[3] Or `http://BitBucket.org`

Live coding

Main Exercise

Main Exercise

1. Create a git repository in a new folder to practice in
2. Create a SAS (or R) file to run analyses on the dataset `sashelp.fish` (SAS) or `airquality` (R)
3. Output the dataset into a `csv` file
4. Save your work to the git repository
5. Find the means and run an ANOVA on the dataset
6. Commit your changes to git
7. Make a fake report on your findings
8. Commit
9. Make a change to your report
10. Commit
11. Revert to the older version
12. Make a branch to experiment