

Git for researchers

Matt Mulvahill

CU Anschutz - Colorado School of Public Health - Dept. of Biostatistics and Informatics

2017-09-19

About me

M.S. in Biostatistics from CSPH

Research Instructor with Dept. of Biostatistics and Informatics

Affiliated with CCTSI-BERD and the CBC

Specialize in:

- **Reproducible research**
- **Computing for 'data science' (R, integrating languages, HTC/HPC, big data management)**
- **Statistical software development**

<http://www.github.com/mmulvahill>

Reproducible Research Toolkit

1. Computing basics

2. Reproducible reports

3. Version control

**Version control is the
lab notebook of the
digital world**

Software Carpentry
swcarpentry.github.io/git-novice/

1. Who

2. What

3. Why

4. How

1. Who is version control for?

- **Software developers**
- **Analysts**
- **Designers**
- **Researchers**
- **NY Times**
- **Novelists**
- **Taco lovers**

You!



This repository

Search

Pull requests

Issues

Gist



NYTimes / gunsales

Watch

13

Star

94

Fork

31

Code

Issues 0

Pull requests 0

Projects 0

Pulse

Graphs

Statistical analysis of monthly background checks of gun purchases <http://www.nytimes.com/interactive/20...>

107 commits

1 branch

0 releases

3 contributors

Branch: master

New pull request

Create new file

Upload files

Find file

Clone or download



eddelbuettel release 0.1.2

Latest commit 3fbc759 on Jan 30



R

update URL, make canonical for CRAN

3 months ago



data

updated population RData file, increased version to mark interim version

a year ago



inst

population

a year ago



man

update URL, make canonical for CRAN

3 months ago



out

added png charts (and ggplot pdf)

a year ago



vignettes

update URL, make canonical for CRAN

3 months ago

github.com/nytimes/gunsales

Project

Repository

Issues 1

Merge Requests 0

Pipelines

Wiki

Settings

Files

Commits

Branches

Tags

Contributors

Graph

Compare

Charts










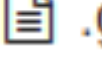
master ▾

p013_dominguez_kawasaki /

+

Find File

 ▾

Name	Last commit > fcbff1e1  a month ago - Merge branch 'develop' History	Last Update
 R	Many changes – only abnormal arteries used in rq5, added combined outcome ana...	a month ago
 Rmd	Many changes – only abnormal arteries used in rq5, added combined outcome ana...	a month ago
 data	working on cleaning NEW aneurysm dataset.	2 months ago
 docs	minor formatting changes to docs w/ description of aims/research questions.	5 months ago
 lib	separated out rq5 and rq6, add floating TOC to html doc, finished methods update a...	a month ago
 output	Many changes – only abnormal arteries used in rq5, added combined outcome ana...	a month ago
 products	added abstract draft.	4 months ago
 sas	finished rq4 analysis.	5 months ago
 .gitignore	initial commit -plus proj outline in README.	8 months ago

Where to eat, if you like tacos.

22 commits

1 branch

0 releases

5 contributors

Unlicense

Branch: master

New pull request

Create new file

Upload files

Find file

Clone or download

hunterowens committed on GitHub Merge pull request #6 from matikin9/patch-1 ...			Latest commit 98fc476 on Dec 23, 2016
chicago	acommit		a year ago
denver	Rename denver_tacos to denver_tacos.md		5 months ago
los_angeles	added more taco places		4 months ago
.gitignore	adding gitignore		4 years ago
Cedar City Utah	Create Cedar City Utah		2 years ago
LICENSE	initial		4 years ago
README.md	change readme		4 years ago

README.md

the_great_taco_hunt

This is a community-driven taco repo, inspired by [TacoFancy](#). Please fork and submit your own places, because that is awesome.

Each city should be its own markdown document. If content warrents it, we can switch to states [or country] naming conventions.

Community assistance in growing the number of cities, or in better organizing or whatever is more more more than

2. What is version control?

Version control is
an **application**
that **documents**
and **tracks**
changes to files

presentation.Rmd Unstaged Staged Stage File Ignore white space X

```
@@ -14,7 +14,11 @@ output:
  14      14      ---
  15      15
  16      16
-17
+17      # Version control with git
+18
+19      # "The lab notebook of the digital world" <br> -Software Carpentry
+20
+21      # Who <br> What <br> Why <br> How
  18      22
  19      23      <div class="notes">
  20      24          *Topics to cover*
```

```
@@ -25,33 +29,38 @@ output:
  25      29      - .gitignore
  26      30      - SSH keys
  27      31      - where to put data. HPAAs issues,
-28      <\div>
-29
-30
-31      # "The lab notebook of the digital world" <br> -Software Carpentry
+32      </div>
  32      33
-33      # 1. Who <br> 2. What <br> 3. Why <br> 4. How <br> 5. Extras
  34      34
  35      35
  36      36
  37      37      <!-------
  38      38      - Section 1 - Who is version control for?
  39      39      ----->
-40      # 1. Who?
+40      # Who is version control for?
```

git

- Ubiquitous (as far as version control systems go)
- Created in 2005 by Linus Torvalds
- Distributed, lightweight, free
- Ideal for collaboration, but you always have one collaborator — **your future self.**

git is powerful,
and powerful
software can be
complicated.

```
tmux:1.git: (tmux) #1 x tmux:1.bash (tmux) #2
* 62738a0 - Mon, 12 Sep 2016 01:23:23 -0600 (8 months ago) (origin/develop)
| switched the %>% import to dplyr from magrittr - Matt Mulvahill (peregrine)
* 858b91d - Mon, 12 Sep 2016 01:22:21 -0600 (8 months ago)
| added packages to import and load %>% from magrittr for use w/in package. - Matt Mulvahill (peregrine)
* 741364c - Sun, 11 Sep 2016 23:31:58 -0600 (8 months ago) (origin/revision-pulse_spec)
| started writing an 'update.pulse_spec' function, but am abandoning it. Also am abandoning moving priors to fit.
|/
peregrine)
* 3531ddd - Sun, 11 Sep 2016 19:21:23 -0600 (8 months ago)
| changed package name to biompulsatile - Matt Mulvahill (peregrine)
* 610644a - Sun, 11 Sep 2016 03:18:35 -0600 (8 months ago)
| added function to create pdf of plots from data. - Matt Mulvahill (peregrine)
* 777f5d2 - Sun, 11 Sep 2016 01:14:25 -0600 (8 months ago)
| switched variable separator to _ from . in simulate.R and added simulate_many_pulsets function. - Matt Mulvahill (peregrine)
* ebd0c0c - Sat, 10 Sep 2016 17:49:25 -0600 (8 months ago)
| started filling out DESCRIPTION file. - Matt Mulvahill (peregrine)
* 100d131 - Sat, 10 Sep 2016 17:49:09 -0600 (8 months ago)
| added simulation function. - Matt Mulvahill (peregrine)
* 2b53bbb - Fri, 2 Sep 2016 22:41:31 -0600 (8 months ago) (HEAD -> master, tag: v0.0.1.0, origin/master, origin/HEAD)
| Reproducible, working function. Merge branch 'develop' - Matt Mulvahill (peregrine)
* df96dfb - Wed, 31 Aug 2016 13:35:47 -0600 (8 months ago)
| removed test text from readme. - Matt Mulvahill (annapurna)
* 6f30dc7 - Wed, 31 Aug 2016 12:25:52 -0600 (8 months ago)
| testing gitkraken's conflict resolution - Matt Mulvahill
* 65d89b3 - Tue, 6 Sep 2016 14:24:02 -0600 (8 months ago) (origin/vectors)
| started pseudo-coding the new pulse object. - Matt Mulvahill (peregrine)
* 10bee77 - Fri, 2 Sep 2016 22:40:55 -0600 (8 months ago)
| removed no longer applicable comment. - Matt Mulvahill (peregrine)
* 1b00457 - Mon, 22 Aug 2016 19:07:22 -0600 (8 months ago)
| Merging repeatability fix -- got rid of global vars for counters. - Matt Mulvahill (MtEvans)
* 95bcdea - Mon, 22 Aug 2016 18:28:39 -0600 (8 months ago) (origin/fix-repeatability)
| Removed extraneous Get/PutRNGstate() calls and old Rprintf() debugging calls. - Matt Mulvahill (MtEvans)
* 2403558 - Mon, 22 Aug 2016 18:17:21 -0600 (8 months ago)
| FIXED: issue was the global variables for counting accept ratio in MH algos. Switched to pointers passed to c
* fdf20ae - Sun, 21 Aug 2016 20:08:56 -0600 (8 months ago)
| added get/put around more RNGs that used the rnorm/runif call -- also switched to Rf_* format. - Matt Mulvahill (MtEvans)
* cd4d9d8 - Sun, 21 Aug 2016 18:44:58 -0600 (8 months ago)
| For travis.yml - Merge branch 'master' into fix-repeatability - Matt Mulvahill (MtEvans)
|/
* 526504f - Sun, 21 Aug 2016 18:19:04 -0600 (8 months ago)
| added travis.yml for testing builds. - Matt Mulvahill (MtEvans)
* a32cbba - Sat, 20 Aug 2016 14:59:39 -0600 (8 months ago)
| Closes #1 - Merge branch 'develop' - Matt Mulvahill (MtEvans)
* 63774ac - Sun, 21 Aug 2016 18:34:16 -0600 (8 months ago)
| put Get/PutRNGstate() before/after all rng calls. - Matt Mulvahill (peregrine)
|/
* e9d119c - Sun, 21 Aug 2016 20:14:20 -0600 (8 months ago)
| Added Rf_ to runif and rnorm calls. - Matt Mulvahill (MtEvans)
* 35d87cf - Sat, 20 Aug 2016 14:58:49 -0600 (8 months ago)
| mcmc is working - closes #1; results reproducible on OSX MBP, but not MtEvans linux Mint. - Matt Mulvahill (MtEvans)
* 368ad68 - Sat, 20 Aug 2016 11:16:14 -0600 (8 months ago)
| clean up testing file. - Matt Mulvahill (MtEvans)
* d1db450 - Sat, 20 Aug 2016 00:06:57 -0600 (8 months ago)
| Chain mixing looks good; locations match w/ visual id of pulses! - Matt Mulvahill (MtEvans)
* 6924ca8 - Fri, 19 Aug 2016 23:48:30 -0600 (8 months ago)
| Alpha v1.1 -- working! Merge branch 'labeled-output' into develop - Matt Mulvahill (MtEvans)
* 535a8cc - Fri, 19 Aug 2016 23:48:00 -0600 (8 months ago) (origin/labeled-output)
| checked pulse and common chains --- look good! - Matt Mulvahill (MtEvans)
* ce2c6dd - Fri, 19 Aug 2016 23:44:54 -0600 (8 months ago)
| Took renamed propsd (formerly prcpvar) and pulse_chains (formerly parm1 and parm); - Matt Mulvahill (MtEvans)
* 6b03cba - Fri, 19 Aug 2016 23:28:37 -0600 (8 months ago)
| Fixed colnames for pulse-specific chain; ***Fixed output of chains to R*** - Matt Mulvahill (MtEvans)
|/
1:git* 2:NvimR-
```


**But the
basics
are
simple.**

Key Concepts

Repository

A folder that is monitored by git

Commit

A record of file changes, accompanied by a unique ID (hash), description of the changes, date they were recorded and name of the person who recorded them


Remote






An external location (server) where your repository is backed-up (GitHub). A project's primary remote is typically named 'origin.'


**Repositories
are folders,
with some
extra stuff.**

test_project			+
Name	^	Date Modified	Size
▶ .git		Today, 4:01 PM	--
.gitignore		Today, 3:58 PM	267 bytes
KD_abstract.docx		Today, 3:09 PM	19 KB
Macintosh HD > Users > matt > Project > BERD > test_project			
3 items, 121.07 GB available			

A commit is a record of a change

Name
 KD_abstract.docx

	My responses and edits to first round of revisions	
	Co-I revision	
	PI added background	
	Added initial abstract draft	
	Initial commit	12 hours ago



Matt Mulvahill (peregrine)

authored 4/25/2017 @ 3:08 PM

commit: ca99c9

parent: 7670c9



This repository

Search

Pull requests

Issues

Gist



mmulvahill / test_project

Unwatch

1

★ Star

0

Fork

0

Code

Issues 0

Pull requests 0

Projects 0

Wiki

Pulse

Graphs

Settings

For use in git seminar

Edit

Add topics

6 commits

1 branch

0 releases

1 contributor

Branch: master

New pull request

Create new file

Upload files

Find file

Clone or download



mmulvahill removed readme and license files for simplicity.

Latest commit 127716b an hour ago

.gitignore

Initial commit

a day ago

KD_abstract.docx

My responses and edits to first round of revisions

2 hours ago

Help people interested in this repository understand your project by adding a README.

Add a README

And a remote is an external backup (with some fancy features)





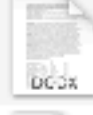

3. Why?

Reasons abound.


Have you ever:


- Made a change to code, realized it was a mistake and wanted to revert back?
- Lost code or had a backup that was too old?
- Had to maintain multiple versions of an analysis?
- Wanted to see the difference between two (or more) versions of your code?
- Wanted to prove that a particular change broke or fixed a piece of code?
- Wanted to review the history of some code?
- Wanted to submit a change to someone else's code?
- Wanted to share your code, or let other people work on your code?
- Wanted to see how much work is being done, and where, when and by whom?
- Wanted to experiment with a new analysis approach without interfering with working code?

But mostly, to keep yourself organized.


Name	
	KD abstract 01-10-17 mjm sd pnjVers2 mjm sd .docx
	KD abstract 12-10-16 mjm.docx
	KD abstract 12-14-16 mjm sd.docx
	KD abstract 12-20-16 mjm sd pnj.docx
	KD abstract 12-22-16 mjm sd pnj Vers2.docx
	KD abstract 12-28-16 mjm sd png Vers2 mjm.docx

versus


Name
 KD_abstract.docx




My responses and edits to first round of revisions




Co-I revision



PI added background




Added initial abstract draft




Initial commit

12 hours ago



Co-I revision

Co-investigator revised some text in background section. Added comments and questions to results.



Matt Mulvahill (peregrine)

authored 4/25/2017 @ 3:08 PM

commit: ca99c9

parent: 7670c9

3. How?

3. How? — git locally

Prerequisites

- Install git (option 1)
- Install a git client (GitKraken)
- Introduce yourself to git

- **git init**
- **git status**
- **git add**
- **git commit**

- **git log**
- **git show**
- **git diff**
- **git checkout (cli)**

.gitignore

Tells git what files it should not track

- **.DS_Store**
- **~\$my_word_file.docx**
- **.Rhistory**
- **data files containing PHI**

3. How? — git remotely

Prerequisites

- An account with a remote
- Connect to remote (SSH or HTTPS)

- **git remote add origin**
- **git push**
- **git pull**

Always pull at the start of the day

“ push “ “ end “ “ “

Implementation issues in research

- **PHI — where to put data?**
- **medium to big data**
- **learning curve + collaboration**
- **standards + collaboration**
- **binary files (output, docs)**

Git for class!

- **Bios 6623 organization and webpage**
- **Resources:**
 - **Happy Git with R by Jenny Bryan**
 - **stat545.com**
 - **GitHub for Education**

Git for methods development and dissemination!

- KechrisLab organization
- Resources:
 - GitHub for Education
 - Travis CI and AppVeyor
 - CodeCov
 - Depsy

References and Resources

- Happy Git with R by Jenny Bryan
- Git for Humans by Alice Bartlett
- Line endings in git (for cross-platform teams/users)
- Advance git cheatsheet
- What is Version Control? by Atlassian (BitBucket)
- Large file storage with Git