Curating Research Assets in Behavioral Sciences: A Tutorial on the Git Version Control System

Matti Vuorre[1] & James P. Curley[1]

[1] Department of Psychology, Columbia University, New York, USA

Abstract

Recent calls for improving the reproducibility of behavioral sciences have increased attention to the ways in which researchers curate, share and collaborate on their research assets. However, these discussions have largely failed to provide practical instructions on how to do so. In this tutorial paper, we explain why version control systems, such as the popular Git program, are especially suitable for these challenges to reproducibility. We then present a tutorial on how to use Git from the computer's command line, and with the popular graphical interface contained in the R Studio development environment. This tutorial is especially written for behavioral scientists with no previous experience with version control systems. Git is easy to learn, presents an elegant solution to challenges to reproducibility related to curating research assets, facilitates multi-site collaboration and productivity (by allowing multiple collaborators to work on the same source simultaneously) and can be implemented on common behavioral science workflows with little extra effort. Git may also offer a suitable solution to transparent data (and material) sharing through popular online services, such as GitHub and Open Science Framework.

Keywords: reproducibility; version control; git; research methods; open science

Word count: X

Correspondence concerning this article should be addressed to Matti Vuorre , 406 Schermerhorn Hall, 1190 Amsterdam Avenue MC 5501, New York, NY 10027 . E-mail: mv2521@columbia.edu

Introduction

The lack of reproducibility is increasingly recognized as a problem across scientific disciplines, and calls for changing the scientific workflow to enhance reproducibility have been published in a wide range of research areas, including Biology (Markowetz, 2015), Ecology (Ihle, Winney, Krystalli, & Croucher, 2017), Neuroscience (Eglen et al., 2017) and Psychology (Munafò et al., 2017). However, although calls to focus on reproducibility are now commonplace on the pages of leading scientific journals (e.g. Baker, 2016), only a small minority of researchers are familiar with the tools and practices that enable implementing reproducibility in their workflows. Therefore, although there now is a broad consensus that efforts to improve reproducibility are important, materials instructing researchers in using them are lacking.

Reproducibility

Consider for a while if you have ever struggled with the following questions (Ihle et al., 2017, p. 2): Have you ever found a mistake in your results without knowing what caused it? Forgot what analyses you have already done and how (and possibly, why)? Lost datasets or information about the cases or variables in a dataset? Struggled in redoing statistical or computational analyses when new data became available? Had difficulty understanding what data to use or how in a project that you inherited from another researcher? Answering any of these questions in the affirmative suggests that your work might benefit from improving reproducibility (Ihle et al., 2017).

But what exactly is reproducibility? Reproducibility can be defined as follows: "A research project is computationally reproducible if a second investigator (including you in the future) can recreate the final reported results of the project, including key quantitative findings, tables, and figures, given only a set of files and written instructions." (Kitzes, Turek, & Deniz, 2017). In the context of experimental Psychology, for example, a "project" could be an experiment or a set of experiments investigating a theory or hypothesis, "reported results" could be figures or statistical analyses reported in a conference presentation or a manuscript.

This definition makes an important distinction between *reproducibility* and *replicability*. Methods and results sections in traditional journal articles have focused on ensuring replicability by giving detailed instructions on how to repeat the experiment and collect a data set similar to the original one. However, little or no emphasis is placed on issues related to reproducibility, such as details of the data processing pipeline and statistical model exploration, although these issues are often more important to the reliability of the scientific results than are their end results (e.g. reported *p*-values; Leek and Peng (2015)). To put it more generally, replicability is a broad conceptual issue concerning the epistemology of scientific claims. Reproducibility, on the other hand, is a technical requirement for allowing others (including researchers themselves) to assess the reliability of a specific set of empirical or computational results.

Further, although they are often erroneously conlated (Collaboration, 2015; Peng,

2011) reproducibility and replicability are independent of one another. Although the distinction between the two may be somewhat blurred in purely computational areas (e.g. simulation studies; Peng (2011)), it is clear that an experiment can be replicated even if the materials of the original study are not available for replicability. Likewise, the materials of an experiment might allow its results to be reproduced, yet the underlying scientific hypothesis might not be replicable. Therefore, it is important to keep these two concepts distinct; in the remainder of the manuscript we explicitly only discuss reproducibility.

Given this broad definition of reproducibility, it is easy to recognize that many projects fall between the extremes of completely replicable and not replicable at all. A completely reproducible project would provide consumers the entire raw data set, and detailed instructions to analyze it (usually as unambigous computer code). This degree of reproducibility is very rare in Psychology (Vanpaemel, Vermorgen, Deriemaecker, & Storms, 2015; Wicherts, Borsboom, Kats, & Molenaar, 2006), although data sharing has become more common in recent years in journals that award authors who do so with special badges (Kidwell et al., 2016). However, even when authors (report to) share data, the shared data sets sometimes turn out to be incorrect, not usable, incomplete, or not available at all (Kidwell et al., 2016). This suggests that reproducibility requires more than good intentions. What, then, are some specific challenges to reproducibility?

Challenges to reproducibility

Evidence suggests that the level of reproducibility within Psychology is not very high. One survey of leading Psychology journals found that 34-58% of articles published between 1985 and 2013 had at least one inconsistently reported *p*-value (Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2015): Upon recalculation, the authors couldn't obtain the same *p*-values from the reported test statistics, and therefore these 34-58% of articles were, to some degree, non-reproducible (the inferential statistics could not be reproduced).

Why, do so many research products fall short of the basic scientific standard of reproducibility? We suggest one reason is the poor level at which researchers organize, curate, and collaborate on the research assets—e.g. data and code used to analyze it—which support the product's conclusions. In one study, researchers asked authors of 141 Psychology articles (with a total of 249 studies) to share their data for reanalysis. 73% of the authors refused to share their data, leading the study's authors to suggest that the considerable efforts involved in cleaning datasets at the final stages of publication may make data sharing unattractive to many (see also Vanpaemel et al., 2015; Wicherts et al., 2006). Vanpaemel et al. highlight some reasons provided by authors for not sharing their data:

> "To our surprise, some authors are apparently willing to share, but have no easy access to their own data or have lost their data altogether, due to

> computer crashes or collaborators having left the university. Many authors cite a lack of time as a reason not to share, and note that sharing their data would take too much effort, which is probably due to poor documentation and storage practices." (p. 2-3)

In short, scientists usually excel at *producing* research assets, but commonly fail at *curating* them (a topic commonly known as Digital Asset Management; Jacobsen, Schlenker, and Edwards (2012)). This state of affairs is not surprising because researchers are trained to do research, but receive little or no training in formal methods or accepted gold standards for curating their research assets. For simple[1] projects, such as writing a solo-author manuscript, this has not been a problem. However, with increasing complexity of research designs, number of collaborators, research sites, and technical skills required in modern scientific workflows, not knowing how to keep track of materials can lead to serious problems, as detailed above.

Fortunately for the empirical sciences, challenges related to organizing and curating materials ("digital assets") across time, space, and personnel have been solved to a high standard in computer science with Version Control Systems (VCS). In the remainder of this article, we introduce a popular VCS called Git, and illustrate its use in the scientific workflow with a hypothetical example project. In the tutorial below, we show how to use the Git VCS by using text commands from the computer's command line. After introducing the fundamentals of Git's VCS functions from the command line, we also show how to use Git easily with a Graphical User Interface (GUI) implemented in the popular R Studio Integrated Development Environment (IDE).

Version Control Systems

Version Control Systems (VCS; also known as Source Control Systems) are computer programs designed for tracking changes to computer code, and collaborating on the code with others. Although initially developed for writing code collaboratively[2], it is easy to recognize that VCS can also be adopted for scientific production and collaboration. For example, behavioral experiments are often written in a programming language, and can have multiple authors and versions. Keeping track of the versions of the program, and allowing many authors to contribute to it (without breaking the code) are problems that VCSs were specifically designed for. By extension, VCS is easily adopted to other aspects of the research cycle, such as curating data across computers and laboratories, or writing manuscripts with multiple authors, versions, and sources of results.

[1]By "simple", we only mean that the technical aspects of the project, related to curating materials related to it, are simple. We of course do not suggest that there might be anything theoretically or scientifically simple in such projects.

[2]The software we present below is used used by major software developers such as Microsoft, Google, and Facebook on code bases with hundreds of contributors.

Box 1. Common problems and their VCS solutions

- How to try different ways of visualizing and modeling data while keeping track of the different versions—and the differences between them?
  - VCS saves each version of the analysis, allowing testing new features without losing previous versions or proliferating files in the project's directories. All the past versions can be directly compared with each other.
- How to work on the same code or manuscript files simultaneously with collaborators?
  - VCS were built for distributed collaborative work: Collaborators work on their local copies and "push" and "pull" material to and from each other or a central "repository". VCS keeps precise track of who has done what, when, and (possibly) why.
- How to share my work in an organized manner with others?
  - VCS requires committing to a common organization scheme with collaborators, making it easier to keep everyone "on the same page" with what goes where and how to contribute to specific parts of the project. Sharing the project with others is built into VCS, and can be facilitated with online services such as GitHub.

The core concepts of version control are that contributors to a project create small checkpoints of the changes they make to the source code (or manuscript text, data, etc.), and then submit those changes to the VCS. The VCS maintains a history of all these little checkpoints, and the exact state of the project at each of these checkpoints. See Figure 1 for a diagram of the typical VCS workflow.

Some popular Version Control Systems are SVN, Mercurial, and Git. In this tutorial, we focus on the most widely used (in science) of these, called Git. Git is especially good for scientific collaboration because of online tools (GitHub) that allow seamless collaboration even for very large research teams.

The Git Version Control System

It is important to understand that unlike an operating system's (OS) default file viewer, such as Finder (Apple computers) or File Explorer (?? Windows), Git[3] is not a standalone program for navigating files and folders on a computer. Instead, it adds functionality to the OSs existing file system, by making available a specific set of functions–either executed from the command line, or through a graphical user interface (GUI) / integrated development environment (IDE)–that allow taking snapshots of the project and its files, and distributing the work across multiple computers and users. To begin using Git, users must first download and install the Git software on their computers. Git is free and open source, works on Windows, Macintosh, and Linux OSs (among others). It can be freely downloaded at https://git-scm.com/. Before we

[3]The creator of Git, Linus Torvalds, named Git after himself as "the stupid content tracker" (McMillan (2005); https://git-scm.com/docs/git.html)

explain the basics of using Git, we present brief installation instructions for Windows and Apple users.

Installing Git

Even if you already have Git installed (some computers do) it is a good idea to install the latest version[4], which can be downloaded as a standalone program from https://git-scm.com/download. Here, we provide more detailed instructions on installing Git for OS X and Windows, but Linux (and other OS) users can find instructions at the Git website (https://git-scm.com/book/en/v2/Getting-Started-Installing-Git). After installation instructions, we detail how to configure the Git program so that it is ready to be used.

OS X. Many OS X computers already have Git installed, especially those operating OS X 10.9 or higher, but it is good practice to install the latest version. The easiest way to update Git to the latest version is to download the installer from http://git-scm.com/download/mac, and install it like any other program. Once Git is installed, it can be used from the command line (the OS X command line is available through the Terminal app, which is included by default with the OS X install), or through various GUIs. Although it might at first appear intimidating because of the archaic-looking interface, we encourage using Git from the command line.

Windows. Windows users can download the Git software installer from http://git-scm.com/download/win. [TODO]

Git setup

To operate properly, Git needs to be able to identify its user. You can set these by entering a few basic commands in the command line. To verify the current user information, type `git config --global user.name`, and hit return. This should not return anything, unless a previous user of the computer has set the global Git user name.

To ensure that Git knows who you are, type `git config --global user.name "User Name"` (where User Name is your name) in the command line, and hit return. If you now re-run the first command, the command line will return the name you entered as `"User Name"`. The username can be anything you'd like, but it is probably a good idea to use your real name so that potential collaborators know who you are. The second piece of information is your email, which can be entered by `git config --global user.email "email@address.com"` (where email@address.com is your email address). You can verify that the correct email address was saved by typing `git config --global user.email` in the Terminal, and hit return. Once this information is entered, Git will know who you are, and is thus able to track who is doing what within a project, which is especially helpful when you are collaborating with other people, or when you are working on multiple computers.

[4]As of the writing of this article, the current version of Git is 2.13.1.

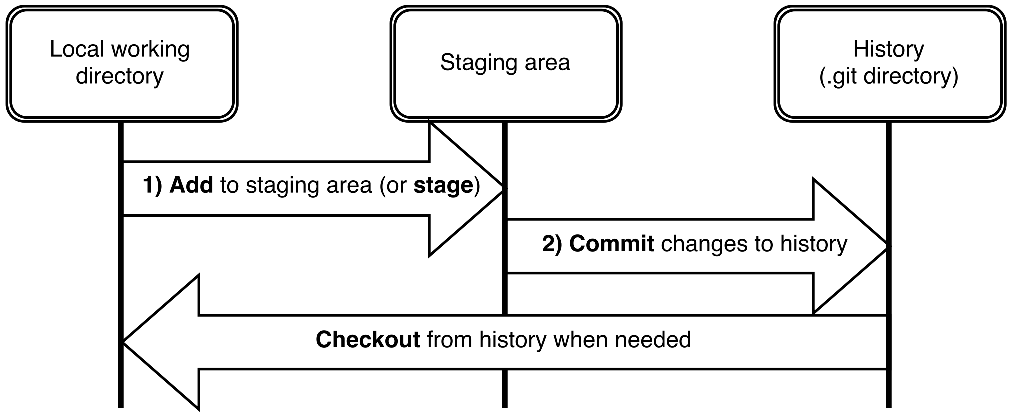

Local working directory
Staging area
History (.git directory)
1) Add to staging area (or stage)
2) Commit changes to history
Checkout from history when needed

Figure 1. A diagram illustrating the typical Git workflow. Verbs indicated in bold text are Git operations and explained in detail in the main text.

Using Git

The first operating principle of Git is that your work is organized into independent projects, which Git calls *repositories*. At its core, a repository is a folder on your computer, which is version controlled by Git (you can tell if a folder is "monitored" by git by checking if the folder, or any of its parent folders, contains a hidden `.git` directory). Everything that happens inside a repository is tracked by Git, but the user has full control of what is tracked (everything, by default) and when. Because the user(s) has full control of what and when is tracked, there is a small set of operations the users need to execute every time they wish to log something in Git.

Briefly, when you work in a Git repository, Git stores the state of each file that it monitors, and when any of these files change, Git indicates that they differ from the previously logged state. If the user then is happy with the current changes, she can **add** the changed files to Git's "staging area". If the user then is certain that the files added to the staging area are in good shape, they can **commit** the changes. These two operations are the backbone of using Git to store the state of the project whenever meaningful changes are made.

Most importantly, each state in Git's log contains the full state of the files at that point in time. Users can always go back to an earlier version by "checking out" a previous state of Git's log. That's why these programs are called Version Control Systems.

To establish the operating principles of Git in practice, we now turn to a practical example using a hypothetical project that, we feel, reflects the components of a typical project in experimental psychology. To most clearly show the use of Git, we begin with an empty project with no components.

Organizing files and folders

Implementing reproducibility into the scientific workflow is less time-consuming if it is planned from the onset of the project, rather than attempting to add reproducibility to the project after it has been completed. It is therefore important to organize the project keeping a few key goals in mind (here, we broadly follow established guidelines such as Project TIER recommendations): The files and folders should have easy to understand names (avoid idiosyncratic naming schemes), and the names should indicate the purposes of the files and folders.

The first step is to create a home folder for the project. This folder should have an immediately recognizable name, and should be placed somewhere on your computer where you can find it. We call the example project `git-example`. All the materials related to this project will be placed in subfolders of the home directory, but for now `git-example` is just an empty folder waiting to be filled with data, code, and documents.

Initializing a Git repository

Next, you need to navigate to the folder and initialize it as a Git repository. Here, the `git-example` project is in the users `Documents` folder, and we can use `cd ..` in the OS X terminal to navigate up in the folder hierarchy, and `cd <folder>` to navigate into `<folder>`. So assuming that the Terminal opens up in `Documents`, we navigate to `git-example` with (to execute code in Terminal, enter the text into the command line and hit Return):

```
cd git-example
```

Then, to turn this folder into a Git repository, execute

```
git init
```

This command initializes the folder as a Git repository, and the only change so far has been the addition of a hidden `.git` folder inside `git-example` (and a `.gitattributes` file; but users can ignore these hidden files and folders).

Adding a file to Git

Every project, and therefore every repository, should contain a brief not that explains what the project is about and who to contact about it. This note is usually called a readme, and therefore our first contribution to this project will be a README file. The README file should be a plain text file (i.e. not created with Microsoft Word) that can be read with a simple text editor. This file can now be seen in the `git-example` folder by using the system file viewer. Because we added this file to a Git repository, Git is also aware of it. To see what files have changed since the last status change in the repository (there clearly has been only this one), you can ask for Git's **status**:

```
git status
```

Which in this case would return

```
Matti [~/Documents/git-example]$ git status
On branch master
Initial commit
Untracked files:
  (use "git add <file>..." to include in what will be committed)
    README
nothing added to commit but untracked files present (use "git add" to track)
Matti [~/Documents/git-example]$
```

The relevant output returned from executing this command is the "Untracked files:" part. There, Git tells the user that there is an untracked file (README) in the repository. To keep track of the status of this file, we **add** it to Git by using the command `git add` followed by either a `.` (for adding all untracked files) or `README` for only tracking the `README` file.

```
git add README
```

We've now added this file to the staging area, and if we are happy with changes to the file's status (it has been created), we can **commit** the file to Git's versioning system. Commits are the most important Git operation: They signify any meaningful change to the repository, and can be later browsed and compared to one another. As such, it is helpful to attach a small message to each commit, describing why that commit was made. Here, we've created a README file, and our commit command would look as follows:

```
git commit -m "Add README file."
```

The quoted text after the `-m` flag is the "commit message". Entering this command to the command line returns a brief description of the commit, such as how many files changed, and how many characters inside those files were inserted and deleted.

Keeping track of changes to a file with Git

Most importantly, the `git-example` project now keeps track of all and any changes to README. To illustrate, we now add some text to README to describe the project (title, what the project is about, who are involved, who to contact), save the file, and then ask for Git's status with `git status` on the command line:

```
Matti [~/Documents/git-example]$ git status
On branch master
Changes not staged for commit:
  (use "git add <file>..." to update what will be committed)
  (use "git checkout -- <file>..." to discard changes in working directory)

    modified:   README

no changes added to commit (use "git add" and/or "git commit -a")
```

Git can now tell that the README file has changed, and we can repeat the add and commit steps to permanently record the current state of the project to Git's log:

```
git add .  # We use the '.' shortcut
git commit -m "Populate README with project description."
```

What does Git know?

The real importance of these somewhat abstract steps becomes apparent when we consider the Git **log**. We can call

```
git log
```

To reveal the commit log of this repository. The main things to notice in the output of this command are that each commit is associated with a unique hash code (long alphanumeric string), which we can use to call for further information (see below); an author (this is where the earlier setup is apparent); a date and time; and the short commit message. Until now, the `git-example` log looks like this:

```
Matti [~/Documents/git-example]$ git log
commit 60cbe5c9b4a78e500314f791080381030577a035
Author: Matti Vuorre <mv2521@columbia.edu>
Date:   Tue Jun 13 17:20:27 2017 -0400

    Populate README with project description.

commit 16c475023ecbc99446164187eeaaab10647ac550
Author: Matti Vuorre <mv2521@columbia.edu>
Date:   Tue Jun 13 17:14:14 2017 -0400

    Add README file.
```

To see what exactly changed in the last commit (the log has latest commits at the top), we can call `git show` with the commit's hash (only relevant parts of output shown below):

```
Matti [~/Documents/git-example]$ git show 60cbe5c9b4a78e500314f791080381030577a035
commit 60cbe5c9b4a78e500314f791080381030577a035
Author: Matti Vuorre <mv2521@columbia.edu>
Date:   Tue Jun 13 17:20:27 2017 -0400

    Populate README with project description.

diff --git a/README b/README
--- a/README
+++ b/README
@@ -0,0 +1,8 @@
+# Example Git Project
+This example project illustrates the use of Git.
+authors:
+Matti Vuorre <mv2521@columbia.edu>
+James Curley
+2017
```

This output can be investigated for a detailed log of all changes created by that commit. From top, it lists the author of the commit, the commit message, and then the commit's "**diff**" (i.e. what differs in the new version vs the old version of the file.) The current diff shows that 1 file received eight additional lines of text (the part wrapped in @ symbols), and then the additions themselves (the lines prepended with +s).

Although we now understand the fundamentals of using Git to track the states of (and therefore changes to) a repository, this contrived and overly simplistic example doesn't allow full appreciation of the benefits of using Git for version control. To better illustrate Git's functioning, we now fast-forward in the hypothetical example project to a stage in the project where more files and materials have been created.

(Slightly more) advanced Git

The state of the project now looks like this:

```
git-example/
    admin/
        ethics-info.pdf
    experiment-1/
        analysis/
```

```
            plan_n.R
    manuscript/
    README
```

Running `git status` now tells that there are two new files (empty folders are ignored). That is, since the previous commit, two files have been added to the project; one a pdf with some administrative information (`ethics-info.pdf`), the other one is an R script for a pre-planned power analysis (`plan_n.R`). We would certainly like to track any changes to the power analysis script, but the ethics information file isn't really something that we need to keep track of—we can't make changes to it anyway. It would be laborious to constantly keep ignoring it when committing changes to Git (especially if we'd like to commit multiple files simultaneously with `git add .`) Git has an elegant solution to specifying which files to keep track of. Because by default all files are tracked, Git uses a special file for instructing which files are to be *ignored* from tracking.

Make Git Ignore Files. To make Git ignore files, we simply add a plain text file called `.gitignore` to the home folder of the repository. Each row of this file should specify a file or a folder (or a regular expression) that Git should ignore. In our example we want to make Git ignore the `admin/` folder entirely, and any file with the .pdf extension inside `manuscript/. The example .gitignore file looks like this (lines beginning with`#‘ are comments):

```
# Ignore everything inside the 'admin' folder
admin/

# Ignore all .pdf files in the 'manuscript' folder
manuscript/*.pdf
```

Re-running `git status` now only shows the `plan_n.R` file (and the newly created .gitignore file, which is also under version control, naturally.)

Because there are now two untracked files, and we would like to keep a clean history of what has happened in the project, we create two separate commits: One for the .gitignore file, and one for the power analysis file.

```
git add .gitignore
git commit -m "Added .gitignore file"

git add .
git commit -m "Completed power analysis"
```

After this last commit, if we ever want to know in the future what our planned sample size was, we can come back to this commit with `git log` or `git show` and see what was inside the newly created power analysis file. For instance, if new information suggests that we should change assumed effect size in the power analysis, we can simply edit and save the file, then add and commit the changes to Git with a helpful message that logs this important event in the Git memory of the repository.

This possibility of "rewinding history" is especially useful for files that might go under multiple revisions, or if we might be interested when and in what order the files were created. One might consider committing a power analysis as a small personal pre-registration of a part of the research plan.

Collaborating

The true advantages of using a VCS such as Git become apparent when we consider projects with more than one contributor. For example, consider a project where data is collected at multiple sites, and these files are then saved onto a centralized server, or shared through a service that automatically merges files from multiple sources (such as the popular Dropbox service). If two or more sites accidentally save a data file with the same name (e.g. `data-001.csv`), and these changes are automatically merged, the later file will simply overwrite the earlier file. Disaster!

Or consider a data analysis where two or more people work simultaneously on some complicated analysis. If user A and B have the same file open on multiple computers, after a while when user B saves the file, user A's version of the file will be overwritten. Disaster!

Git and other VCSs were specifically designed to allow (and facilitate) multi-site collaboration on extremely complex projects, as for example required by leading software companies such as Apple and Microsoft. Next, we'll illustrate how collaborating on a Git project works in practice.

There are many ways in which a team could collaborate on a Git project (e.g. https://www.atlassian.com/git/tutorials/comparing-workflows), and here we focus on a common one, called the "Centralized Workflow".

Overview of the Centralized Git Workflow

In this workflow, a central repository hosted on a research team's server, or an online service like GitHub (https://github.com/), is used to pull and push (see Table 1) information to and from local repositories. First, one user creates this central repository, then other users clone a local copy (see Table 1). All changes are first created, added, and committed locally as detailed above, and only then **pushed** to the central repository. To get changes that other users have pushed to the central repository, a user **pulls** changes from it.

Setting up a centralized Git repository on a research team's private server is relatively straightforward, but because the details vary from team to team, here we illustrate the Centralized Workflow using GitHub.

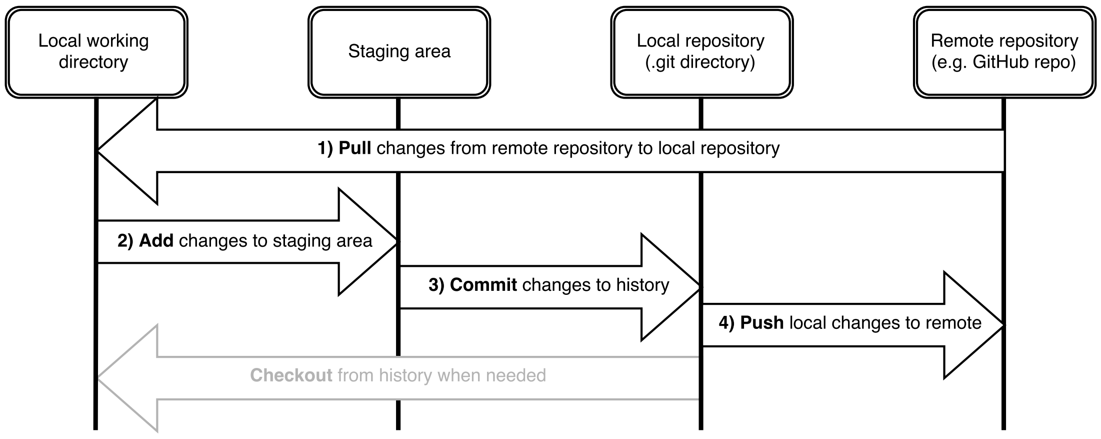

Local working directory
Staging area
Local repository (.git directory)
Remote repository (e.g. GitHub repo)
1) Pull changes from remote repository to local repository
2) Add changes to staging area
3) Commit changes to history
4) Push local changes to remote
Checkout from history when needed

Figure 2. A diagram illustrating the typical collaborative Git workflow with a remote repository (e.g. GitHub). Verbs indicated in bold text are Git operations and explained in detail in the main text.

GitHub

GitHub is one of the 100 most popular website worldwide, and hosts over 60 million software projects with a total of over 20 million users (https://github.com/about). We chose GitHub for collaboration because it is already the de-facto standard in VCS collaboration among scientists, it offers free private (see below) repositories for students, and has many attractive features, some of which we will detail below.

Anyone wanting to use GitHub must first create a free user account at https://github.com. After the user account has been set up, and the user has logged in, creating a new repository at GitHub is self evident (there is a big green button labeled "New repository"). You must create the GitHub account with the same email address as you used above when configuring your local Git (or re-configure your local Git to use your the email address that you used to register for GitHub).

Create a new GitHub repository. Although we are creating a new repository here, the goal is to create an online "remote" for the running example within this tutorial, and therefore the specific steps are slightly different than what they would be if a brand new repository was needed.

For the current example, after clicking "New repository", we simply enter a name, which can be anything, but for consistency we call the GitHub repository `git-example`. After entering the name, click "Create repository", and the next step is to link the new remote to the local repository. The required steps are detailed on screen when you do this at GitHub, but here we would execute the following command in the command line:

```
git remote add origin https://github.com/<username>/<reponame>.git
```

Where `<username>` and `<reponame>` should be replaced with the user's GitHub username, and the GitHub repository name. Once the connection is set up, we can **push** local changes to the GitHub remote:

```
git push -u origin master
```

The `-u origin master` are only required for the first push, as they set up the connection. For pushing any changes collowing this, simply type `git push` after adding and committing locally.

The team has now created the remote Central repository, and other users can start contributing to it.

Contributing to a Central (GitHub) repository. Once another team member has set up Git on their own computer, and signed up for GitHub, they need to **clone** the remote GitHub repository onto their local computer. [TODO actually check that these steps are correct with a "fresh" computer.]

To clone a repository, the new user must first navigate to an appropriate location on the computer where they would like to create the repository on their computer, for example their `username/Documents` folder. Once the appropriate location is found, cloning will create a new folder for the repo inside this folder.

```
git clone https://github.com/<username>/<reponame>.git
```

These new contributors can then work on their local copies as detailed in earlier parts of this tutorial; making changes, adding, and committing. After committing their changes, they can update the status of the Central Git repo by pushing their changes to it:

```
git push
```

Obtaining other's changes from Central repo. Just as you must manually push your own local changes to the remote repository, you must also obtain others' changes by **pulling** them from the central repo. Before starting to work on your proposed change, pull the remote changes with [TODO see if it needs to be rebased]:

```
git pull
```

The way in which users, and their local repositories, interact with the central repository by pushing and pulling is the cornerstone of collaboration on GitHub, and thoughtful use of these commands allows for complex workflows without any important code (data, ideas in manuscript, analysis code) being ever overwritten. However, there is no automatic way for a computer to tell what changes to prioritize: If two users have worked on the same code, and they both attempt to push their changes, the user to do this later will have to resolve the **conflict**.

Resolving a conflict. If user A has pushed changes to the central repository while user B was working on the same code, user B's push will result in a merge conflict. That is just a natural consequence of two individuals working simultaneously on the same idea. User B simply needs to first merge the changes in the remote repository to her own code by pulling from the remote. Once the conflict has been manually resolved (and user B's code updated with the pull), user B can push changes from her local repository to the central repository. How these potential conflicts appear depends on how users collaborate with one another, and a detailed explanation of all potential scenarios is outside the scope of this tutorial. For more information, see e.g. https://www.atlassian.com/git/tutorials/comparing-workflows. The GitHub customer service is also very responsive to users' help requests.

Private or public collaboration?

By default, all GitHub repositories are public, meaning that anyone with an internet connection can use their web browser to inspect the contents of your repository, or even clone it to their computer. This may sound intimidating to researchers who are used to working in a more private context, and clearly necessitates planning and thought with respect to issues such as data privacy and sharing sensitive materials. However, for many projects—including the writing of this tutorial—we see very few downsides to working "in the open".

There are two alternatives to working in a public GitHub repository: One, which we won't cover here, is to not use GitHub but instead place the central repository on the research team's shared server. The second option is to change the repository's settings on GitHub (this can be done when the repository is first created or afterwards) to private. Private repositories, and their contents, are only accessible to invited team members, and are therefore ideal for small teams who would like to work without revealing their master plans to the public just yet.

To make a GitHub repository private, navigate to the repository's website with a web browser (e.g. https://github.com/%3Cusername/>), and click "Settings", then "Make this repository private". Once one user has set the central GitHub repository to as private, anyone wishing to clone, push, pull, or view the repo must provide their GitHub username and password. Only if they match an invited team members username and password can the user access the repository.

Using Git from IDE

Finally, we have above detailed how to use Git from the computer's command line, but not all users are comfortable—although there is no reason not to be!—with this mode of interacting with their computers. We chose to introduce Git from the command line because eventually users will need to use it for more complicated operations (and it is required for setting the user's information); however, there are various graphical user interfaces (GUIs) available for Git, that allow interacting with Git by pointing and clicking buttons with the mouse.

In this section, we introduce how to use the popular R Studio Integrated Development Environment (IDE) for the R programming language for interacting with Git. An IDE is an interface that bundles together many necessary features of software development—some users may be familiar with R Studio for conducting statistical analyses. Version control is an essential feature of software development, and R Studio provides a good GUI within its IDE for controlling Git.

For this example, we will begin a completely new repository using only R Studio, and assume that readers are familiar with the concepts from the above tutorial.

R Studio

[TODO download and install R Studio, cite]

Managing projects (and Git repositories) with R Studio is centered on the idea of R Projects. To start a new project with R Studio, open the R Studio application, and click File -> New Project. This brings up a dialog asking whether to create a project in a new directory, existing directory, or checkout an existing project from a version control repository. Here, we create a new project in a new directory, and choose "New Project" in the following screen. Then we'll give a name to the project's home folder and choose where to save it on the computer. Importantly, we'll also check the "Create a git repository" box, which will automatically set up a new repository for the project (provided you have set Git up as detailed above). Clicking "Create project" creates the folder in the specified location, and two files inside the project's main folder.

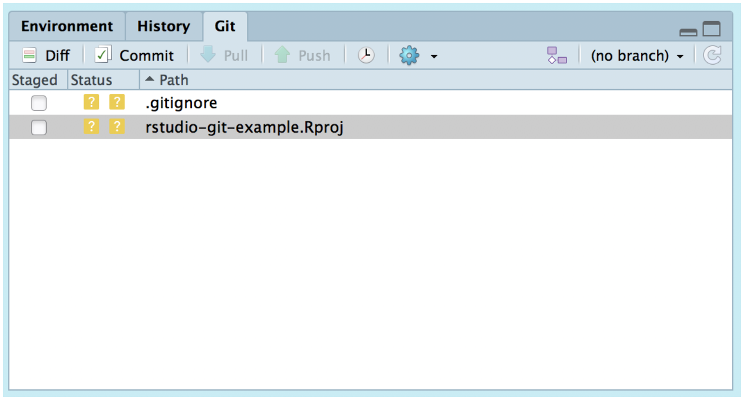
[TODO put some screenshots in here]

One of these files is `.gitignore` which we covered above. The other file is and .Rproj file, which indicates that the folder is the home folder for an R (Studio) project. Users don't interact with this file, but when opened, it is simply a plain text file containing the project's settings (these can be modified through Tools -> Project Options in R Studio).

Using Git with R Studio

Once the R Project has been created, R Studio has a "Git" tab in the top-right panel of the GUI. At first, this tab shows the two new files in the repository, and some buttons with familiar looking names ("Commit", "Push", etc.)

Add and Commit changes. To mark this milestone of creating a project, let's commit these changes to Git. The workflow is exactly the same as when using Git from the terminal, but we now have some visually appealing helpers. To begin committing these changes, click on "Commit" in the Git tab. This brings up another window where we can select files to add into the staging area (the `.gitignore` file in Figure 4 is staged by checking the box). Once all the desired files (here we chose



.gitignore
rstudio-git-example.Rproj

Figure 3. R Studio's Git tab for a newly created project.

both) are staged, we write a short commit message as detailed above, and click the "Commit" button.

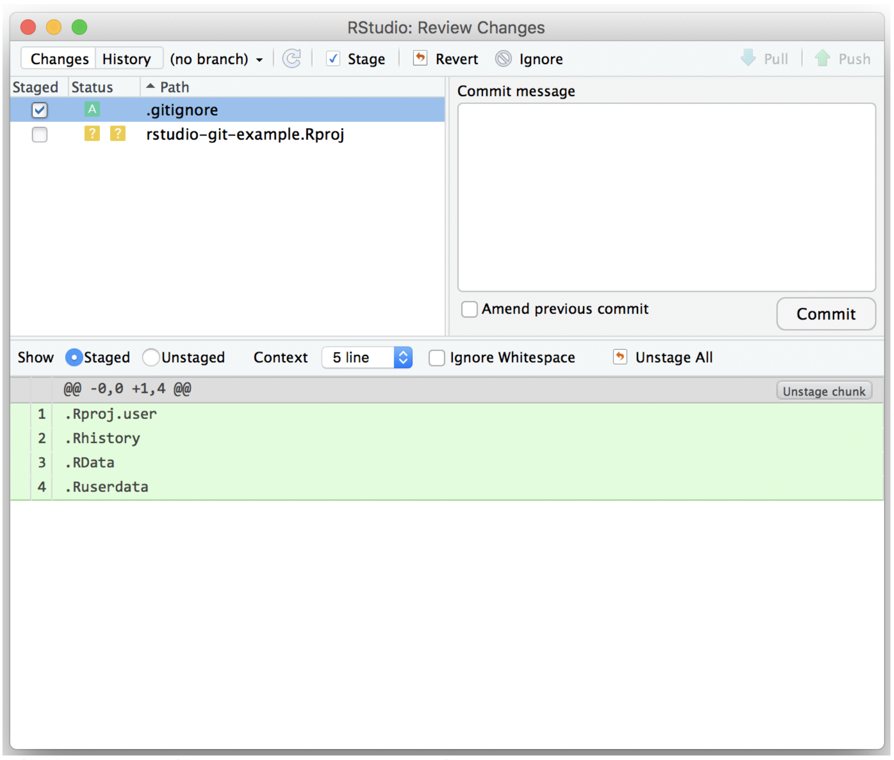
Add a remote GitHub repository. When creating the R Project, there was an option for creating the project from an existing Git repository. Because we didn't do this, we must now manually instruct Git that this repo should have a remote in GitHub. This is done exactly as above, and we must use command line functions. After you've created a new repository on GitHub, navigate to the project's folder on your computer with the command line navigator of your choice, and execute the following commands:

```
git remote add origin https://github.com/<username>/<reponame>.git
git push -u origin master
```

After executing these commands, you can use all Git and GitHub features from R Studio without having to use the command line.

Further examples.

- Born open data (Rouder, 2015)



Staged	Status	Path
✓	A	.gitignore
	? ?	rstudio-git-example.Rproj

```
@@ -0,0 +1,4 @@
.Rproj.user
.Rhistory
.RData
.Ruserdata
```

Figure 4. Creating a Git commit with R Studio.

Table 1
Main Git commands.

Operation	What it does	Command
Initialize	Turns a local folder into a Git repository	git init
Clone	Create a local copy of an existing remote repository	git clone <url>
Add	Adds files (or changes to file) to Git's staging area	git add <file / .>
Commit	Save changes in staging area to version control	git commit -m '<message>'
Show log	Reveal a timeline of the repository's past commits	git log
Compare	Compare two versions of a file to each other	git diff
Push	Send committed local changes to a remote repository	git push
Pull	Download changes from remote repository to local	git pull

Note. Source: https://help.github.com/articles/github-glossary/ and https://services.github.com/on-demand/downloads/github-git-cheat-sheet.pdf

Git Summary

Overview of Git Commands

Box 2. Further resources for learning Git

- Basic VCS workflow, an infographic explaining how VCS works (https://www.git-tower.com/blog/workflow-of-version-control).
- GitHub's Git cheat sheet is available in multiple languages and contains the most used Git commands (https://services.github.com/resources/cheatsheets/).
- TryGit, an interactive website for learning the basics of Git (https://try.github.io).
- Git + GitHub in an R programming context (http://r-pkgs.had.co.nz/git.html).
- Pro Git book, a complete manual of Git (https://git-scm.com/book/en/v2).

References

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature News*, *533*(7604), 452. doi:10.1038/533452a

Collaboration, O. S. (2015). Estimating the reproducibility of psychological. *Science*, *349*(6251), aac4716. doi:10.1126/science.aac4716

Eglen, S. J., Marwick, B., Halchenko, Y. O., Hanke, M., Sufi, S., Gleeson, P., … Poline, J.-B. (2017). Toward standard practices for sharing computer code and programs in neuroscience. *Nature Neuroscience*, *20*(6), 770–773. doi:10.1038/nn.4550

Ihle, M., Winney, I. S., Krystalli, A., & Croucher, M. (2017). Striving for transparent and credible research: Practical guidelines for behavioral ecologists. *Behavioral Ecology*. doi:10.1093/beheco/arx003

Jacobsen, J., Schlenker, T., & Edwards, L. (2012). *Implementing a Digital Asset Management System: For Animation, Computer Games, and Web Development*. CRC Press.

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L.-S., … Nosek, B. A. (2016). Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. *PLOS Biology*, *14*(5), e1002456. doi:10.1371/journal.pbio.1002456

Kitzes, J., Turek, D., & Deniz, F. (2017). *The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences*. University of California Press. Retrieved from https://www.practicereproducibleresearch.org/

Leek, J. T., & Peng, R. D. (2015). Statistics: *P* values are just the tip of the iceberg. *Nature News*, *520*(7549), 612. doi:10.1038/520612a

Markowetz, F. (2015). Five selfish reasons to work reproducibly. *Genome Biology*, *16*(1), 274. doi:10.1186/s13059-015-0850-7

McMillan, R. (2005, April 20). After controversy, Torvalds begins work on "Git". *PC World*. Retrieved from https://www.pcworld.idg.com.au/article/129776/after_controversy_torvalds_begins_work_git_/

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Sert, N. P. du, … Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, 0021. doi:10.1038/s41562-016-0021

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2015). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, 1–22. doi:10.3758/s13428-015-0664-2

Peng, R. D. (2011). Reproducible Research in Computational Science. *Science*, *334*(6060), 1226–1227. doi:10.1126/science.1213847

Rouder, J. N. (2015). The what, why, and how of born-open data. *Behavior Research Methods*, 1–8. doi:10.3758/s13428-015-0630-z

Vanpaemel, W., Vermorgen, M., Deriemaecker, L., & Storms, G. (2015). Are We Wasting a Good Crisis? The Availability of Psychological Research Data after the Storm. *Collabra: Psychology*, *1*(1). doi:10.1525/collabra.13

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*(7), 726–728. doi:10.1037/0003-066X.61.7.726