

# Data Science for Economists

## Lecture 2: Version control with Git(Hub)

---

Grant McDermott, adapted by Kyle Coombs

Bates College | [ECON 368](#)

# Table of contents

1. Prologue
2. Git and GitHub
3. Git(Hub) + RStudio
4. Git from the shell
5. Merge conflicts
6. Branches and forking
7. Other tips
8. Summary
9. Appendix: FAQ

# Prologue

---

# Student presentation: R Markdown

Resources:

- Website: <https://rmarkdown.rstudio.com>
- Cheatsheet: <https://github.com/rstudio/cheatsheets/raw/main/rmarkdown-2.0.pdf>
- Book: *R Markdown: The Definitive Guide* (Yihui Xie, JJ Allaire, and Garrett Grolemund)

Other points:

- You will be submitting all of your assignments as R Markdown documents.
- FWIW, my lecture slides and notes are all written in R Markdown too. (E.g. This slide deck is built using the **xaringan** package with the metropolis theme.)

# Before we start

We went through a software installation check during the previous lecture. By now you should have:

- ☑ Installed [R](#).
- ☑ Installed [RStudio](#).
- ☑ Installed [Git](#).
- ☑ Created an account on [GitHub](#)
- ☑ Accepted an invitation to ECON 368 course repo.

If in doubt about software, please consult Jenny Bryan's amazing guide: <http://happygitwithr.com>.

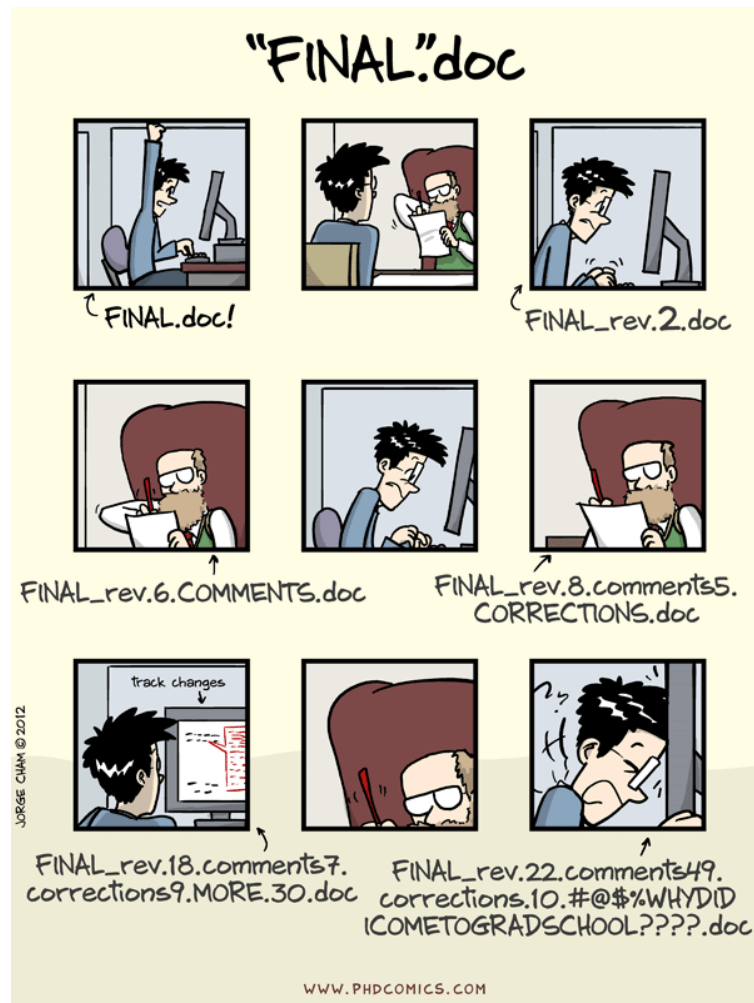
# Class updates

- The [final project replication](#) instructions have been posted. Please read through.
  - Replication proposal is due 1/25
  - It is a group project, please select a partner by 1/18
- I added my policy on posting to GitHub to the [Discussions page](#)
  - tl;dr: Post thoughtfully
- How are folks doing with problem set 0?
- Questions on:
  - GitHub course page?
  - Syllabus and course policies? (Make note of regrading policy.)
  - Grading?
  - GitHub Education?

# Git and GitHub

---

# Why bother?





# Git(Hub) solves this problem

## Git

- Git is a distributed version control system. (Wait, what?)
- Okay, try this: Imagine if Dropbox and the "Track changes" feature in MS Word had a baby. Git would be that baby.
- In fact, it's even better than that because Git is optimised for the things that economists and data scientists spend a lot of time working on (e.g. code).
- There is a learning curve, but I promise you it's worth it.

## GitHub

- It's important to realise that Git and GitHub are distinct things.
- GitHub is an online hosting platform that provides an array of services built on top of the Git system. (Similar platforms include Bitbucket and GitLab.)
- Just like we don't *need* Rstudio to run R code, we don't *need* GitHub to use Git... But it will make our lives so much easier.

# Git(Hub) for scientific research

## From software development...

Git and GitHub's role in global software development is not in question.

- There's a high probability that your favourite app, program or package is built using Git-based tools. (RStudio is a case in point.)

## ... to scientific research

Scientists and academic researchers are cottoning on too.

- Benefits of VC and collaboration tools aside, Git(Hub) helps to operationalise the ideals of open science and reproducibility.
- Journals have increasingly strict requirements regarding reproducibility and data access. GH makes this easy (DOI integration, off-the-shelf licenses, etc.)
- I host all of the code and data for my [papers](#) on GH. Same for teaching materials. I even use it to host and maintain my [website](#).
- *Nature*: "[Democratic databases: science on GitHub](#)" (Perkel, 2016).

# Git(Hub) + RStudio

---

# Seamless integration

One of the (many) great features of RStudio is how well it integrates version control into your everyday workflow.

- Even though Git is a completely separate program to R, they feel like part of the same "thing" in RStudio.
- This next section is about learning the basic Git(Hub) commands and the recipe for successful project integration with RStudio.

# Seamless integration

One of the (many) great features of RStudio is how well it integrates version control into your everyday workflow.

- Even though Git is a completely separate program to R, they feel like part of the same "thing" in RStudio.
- This next section is about learning the basic Git(Hub) commands and the recipe for successful project integration with RStudio.

I also want to bookmark a general point that we'll revisit many times during this course:

- The tools that we're using all form part of a coherent data science ecosystem.
- Greatly reduces the cognitive overhead ("aggregation") associated with traditional workflows, where you have to juggle multiple programs and languages at the same time.

# First create an SSH-key

Before we get started, you need to create an SSH key on your local machine.

- SSH keys are a way to identify trusted computers, without involving passwords.
- A private key on your computer that matches a public key to put on servers, GitHub, etc.
  - Never reveal the private key, only the public

## SSH-Keys

- A key has a type of encryption, two examples:
  - RSA (Rivest-Shamir-Adleman)
  - ED25519 (Edwards-curve Digital Signature Algorithm)
- We will use ED25519, which is smaller, more secure and faster than RSA
  - GitHub also recommends it
- We will generate a key, then add the public key to GitHub SSH keys.
  - I will show you the "hard" way first, then an "easier" way second (guess which we'll do?)
- You can add a passphrase to your SSH-key, but you will need to remember that password every time you push/pull. (I often don't use a passphrase for my SSH-keys unless it is for a highly sensitive use case.)

# The hard way -- ssh-keygen

- You can generate an SSH key via the terminal with the following command:

```
ssh-keygen -t ed25519 -C "YOUR-EMAIL-ADDRESS OR OTHER COMMENT"
```

- You will be prompted to enter a file in which to save the key. Just hit enter to accept the default location.
- You will then be prompted to enter a passphrase. You can either enter a passphrase or leave it blank. (If you leave it blank, you will not be prompted for a passphrase when you use the key.)
- You should see something like this:

```
Generating public/private ed25519 key pair.  
Enter file in which to save the key (/Users/you/.ssh/id_ed25519):  
Enter passphrase (empty for no passphrase):  
Enter same passphrase again:  
Your identification has been saved in /Users/you/.ssh/id_ed25519  
Your public key has been saved in /Users/you/.ssh/id_ed25519.pub.
```

# Add the SSH-key to ssh-agent, then

- Then you'll need to add the key to the ssh-agent:

```
eval "$(ssh-agent -s)"  
ssh-add ~/.ssh/id_ed25519
```

**Note:** This is OS-specific, see instructions [here](#)

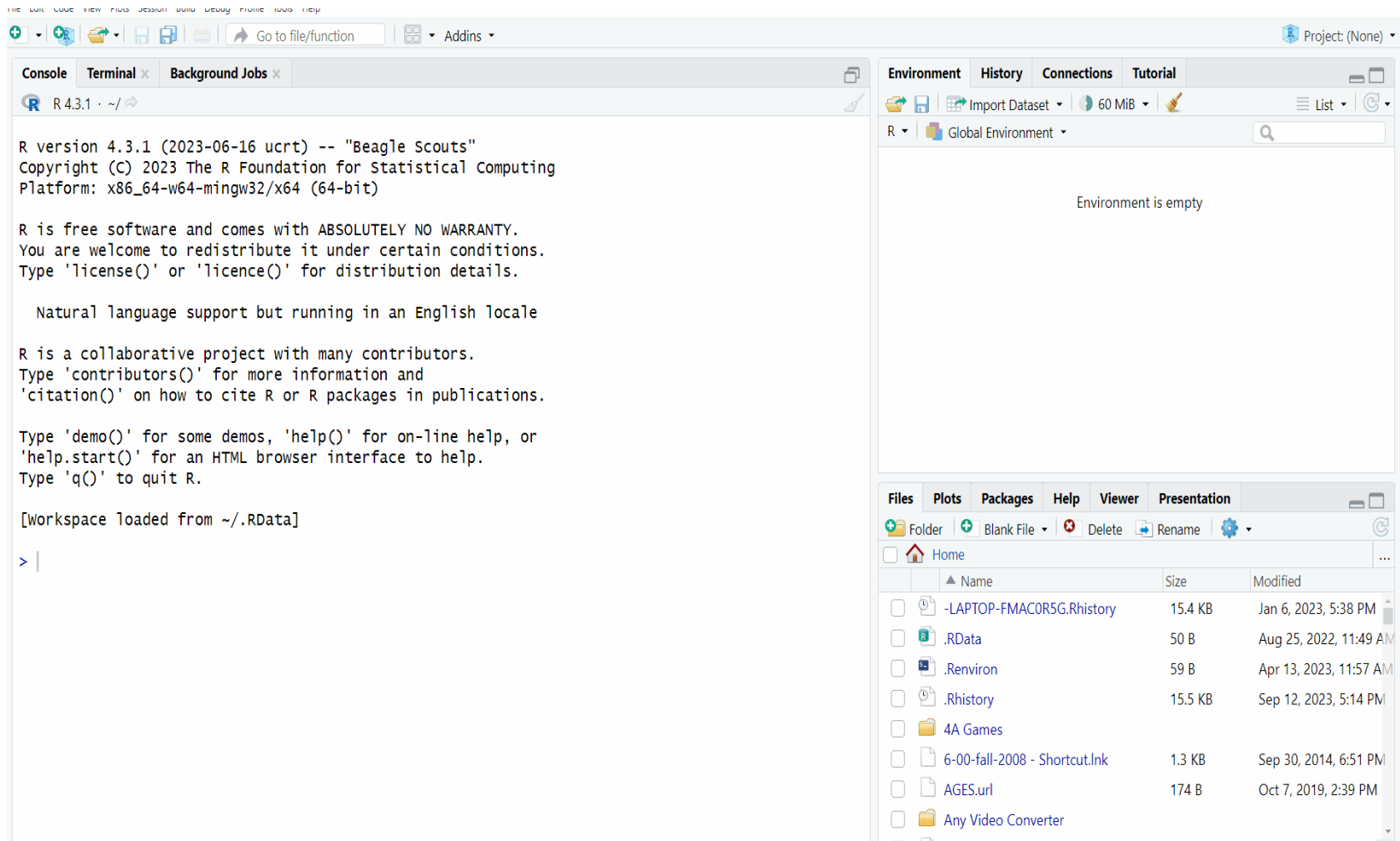
- Last you'll navigate to your ssh-key in your file system and copy the contents of the public key (id\_ed25519.pub) to your clipboard.
- Then you'll go to your GitHub account settings and add the public key to your account under SSH keys.



# Point-and-click with RStudio

- You can also point-and-click with RStudio: Tools -> Global Options -> Git/SVN -> Create RSA Key...
  - Guess which we're gonna do today?
- If you already have an SSH key on your local machine, then you can skip this step.
- Instead, RStudio will already see your key and you can click "View public key" to copy it to your clipboard and add to GitHub

# A gif how-to



# Link a GitHub repo to an RStudio Project

The starting point for our workflow is to link a GitHub repository (i.e. "repo") to an RStudio Project. Here are the steps we're going to follow:

1. Create the repo on GitHub and initialize with a README.
2. Copy the HTTPS/SSH link (the green "Clone or Download" button).<sup>1</sup>
3. Open up RStudio.
4. Navigate to **File -> New Project -> Version Control -> Git**.
5. Paste your copied link into the "Repository URL:" box.
6. Choose the project path ("Create project as subdirectory of:") and click **Create Project**.

<sup>1</sup> It's easiest to start with HTTPS, but **SSH** is advised for more advanced users.

# Link a GitHub repo to an RStudio Project

The starting point for our workflow is to link a GitHub repository (i.e. "repo") to an RStudio Project. Here are the steps we're going to follow:

1. Create the repo on GitHub and initialize with a README.
2. Copy the HTTPS/SSH link (the green "Clone or Download" button).<sup>1</sup>
3. Open up RStudio.
4. Navigate to **File -> New Project -> Version Control -> Git**.
5. Paste your copied link into the "Repository URL:" box.
6. Choose the project path ("Create project as subdirectory of:") and click **Create Project**.

Now, I want you to practice by these steps by creating your own repo on GitHub — call it "test" — and cloning it via an RStudio Project.

- See Grant's GIF walkthrough on the next slide...

<sup>1</sup> It's easiest to start with HTTPS, but [SSH](#) is advised for more advanced users.

# Link a GitHub repo to an RStudio Project

The screenshot shows the GitHub profile of Grant McDermott. The profile includes a bio, location, email, website, and a list of popular repositories. The repositories listed are: **rstudio-compute-engine** (RStudio Server on Google Compute Engine, 18 stars, 5 forks), **R-intro** (Introduction to R, HTML, 2 stars, 1 fork), **sfg-bayes-intro** (A very high-level introduction to Bayesian data analysis (using R), HTML, 1 star, 1 fork), **blueparadox** (Code and data for "The blue paradox: Preemptive overfishing in marine reserves" (McDermott, Meng, et al., 2018; PNAS), Stata, 1 star, 1 fork), **cmip5-models** (Comparing the CMIP5 suite of climate models to observed global temperatures, HTML, 9 forks), and **sceptic-priors** (Code and data for "Sceptic priors and climate policy" paper, R, 1 fork). Below the repositories, there is a contribution graph showing 453 contributions in the last year.

GitHub, Inc. [US] | <https://github.com/grantmcdermott>

Search or jump to... Pull requests Issues Marketplace Explore

**Grant McDermott**  
grantmcdermott

★ PRO

Assistant Professor of Economics at the University of Oregon. Mostly working on natural resources and the environment, but easily distracted.

University of Oregon  
Eugene, Oregon  
✉ [grantmcd@uoregon.edu](mailto:grantmcd@uoregon.edu)  
🌐 [www.grantmcdermott.com](http://www.grantmcdermott.com)

Edit

**Organizations**

**Overview** Repositories 33 Stars 1 Followers 18 Following 0

**Popular repositories** Customize your pinned repositories

**rstudio-compute-engine**  
RStudio Server on Google Compute Engine  
★ 18 🍴 5

**R-intro**  
Introduction to R  
HTML ★ 2 🍴 1

**sfg-bayes-intro**  
A very high-level introduction to Bayesian data analysis (using R)  
HTML ★ 1 🍴 1

**blueparadox**  
Code and data for "The blue paradox: Preemptive overfishing in marine reserves" (McDermott, Meng, et al., 2018; PNAS)  
Stata ★ 1 🍴 1

**cmip5-models**  
Comparing the CMIP5 suite of climate models to observed global temperatures  
HTML 🍴 9

**sceptic-priors**  
Code and data for "Sceptic priors and climate policy" paper  
R 🍴 1

453 contributions in the last year Contribution settings

Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec  
Mon  
Wed

# Make some local changes

Look at the top-right panel in your RStudio IDE. Do you see the "Git" tab?

- Click on it.
- There should already be some files in there, which we'll ignore for the moment.<sup>1</sup>

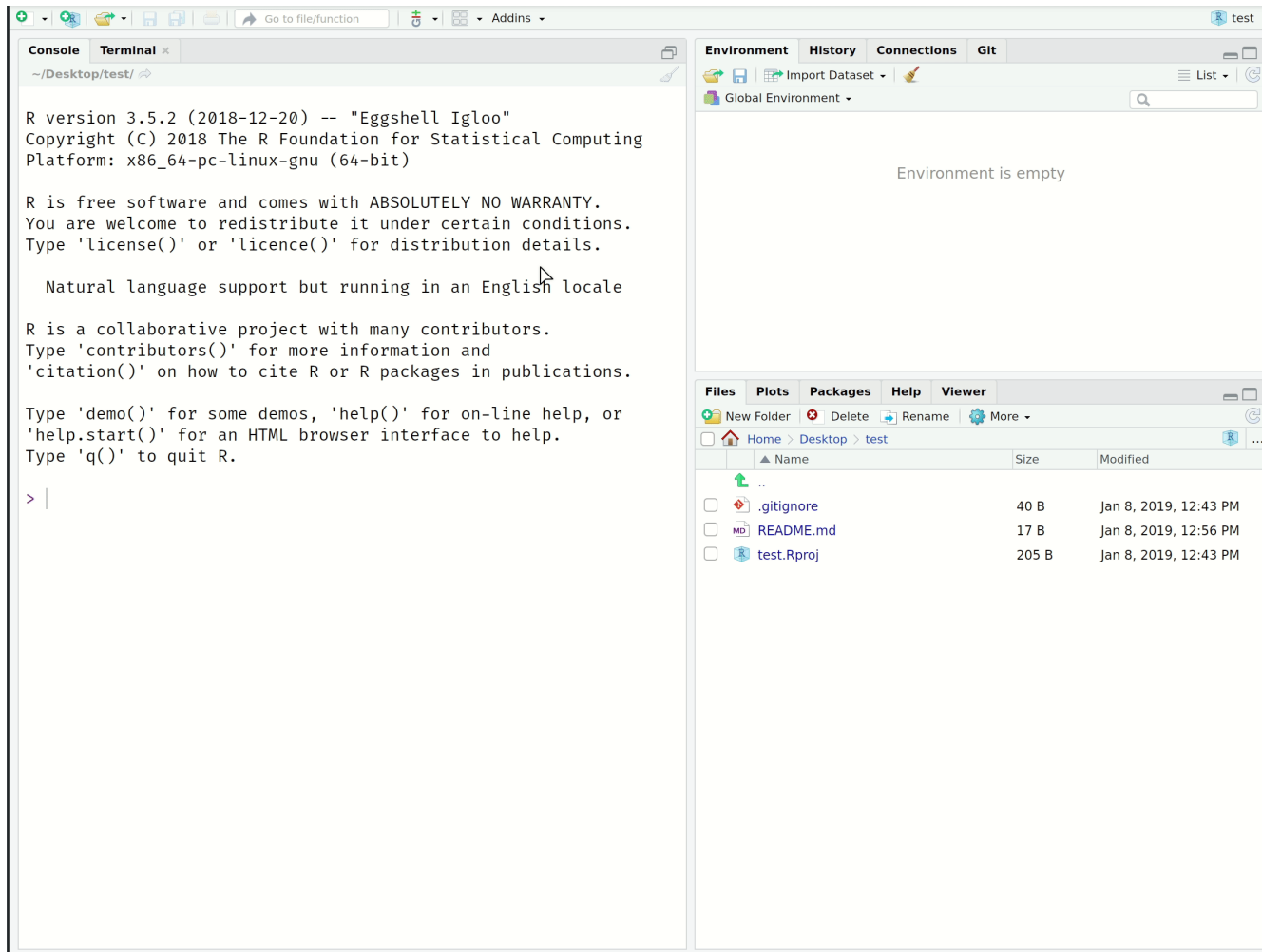
Now open up the README file (see the "Files" tab in the bottom-right panel).

- Add some text like "Hello World!" and save the README.
- Do you see any changes in the "Git" panel? Good. (Raise your hand if not.)

Again, see Grant's GIF walkthrough on the next slide...

<sup>1</sup> They're important, but not for the purposes of this section.

# Make some local changes



The screenshot displays the RStudio IDE interface. The top-left panel is the 'Console' terminal, showing the R version 3.5.2 (2018-12-20) -- "Eggshell Igloo" and copyright information. The top-right panel is the 'Environment' pane, which is currently empty. The bottom-right panel is the 'Files' pane, showing the file explorer for the project directory. The file explorer shows a list of files: '..', '.gitignore', 'README.md', and 'test.Rproj'.

**Console Terminal Output:**

```
R version 3.5.2 (2018-12-20) -- "Eggshell Igloo"
Copyright (C) 2018 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

**Environment Pane:**

Environment is empty

**Files Pane:**

Name	Size	Modified
..		
.gitignore	40 B	Jan 8, 2019, 12:43 PM
README.md	17 B	Jan 8, 2019, 12:56 PM
test.Rproj	205 B	Jan 8, 2019, 12:43 PM

# Main Git operations

Now that you've cloned your first repo and made some local changes, it's time to learn the four main Git operations.

1. **Stage** (or "add")
  - Tell Git that you want to add changes to the repo history (file edits, additions, deletions, etc.)
2. **Commit**
  - Tell Git that, yes, you are sure these changes should be part of the repo history.
3. **Pull**
  - Get any new changes made on the GitHub repo (i.e. the upstream remote), either by your collaborators or you on another machine.
4. **Push**
  - Push any (committed) local changes to the GitHub repo



# Main Git operations

Now that you've cloned your first repo and made some local changes, it's time to learn the four main Git operations.

1. **Stage** (or "add")

- Tell Git that you want to add changes to the repo history (file edits, additions, deletions, etc.)

2. **Commit**

- Tell Git that, yes, you are sure these changes should be part of the repo history.

3. **Pull**

- Get any new changes made on the GitHub repo (i.e. the upstream remote), either by your collaborators or you on another machine.

4. **Push**

- Push any (committed) local changes to the GitHub repo

For the moment, it will be useful to group the first two operations and last two operations together. (They are often combined in practice too, although you'll soon get a sense of when and why they should be split up.)

# Main Git operations

Now that you've cloned your first repo and made some local changes, it's time to learn the four main Git operations.

1. **Stage** (or "add")

- Tell Git that you want to add changes to the repo history (file edits, additions, deletions, etc.)

2. **Commit**

- Tell Git that, yes, you are sure these changes should be part of the repo history.

3. **Pull**

- Get any new changes made on the GitHub repo (i.e. the upstream remote), either by your collaborators or you on another machine.

4. **Push**

- Push any (committed) local changes to the GitHub repo

For the moment, it will be useful to group the first two operations and last two operations together. (They are often combined in practice too, although you'll soon get a sense of when and why they should be split up.)

Ready for more GIFs?

# Stage and Commit

The screenshot shows the RStudio IDE interface. The main editor window displays a file named `README.md` with the following content:

```
1 # My first test repo
2
3 Hello world!
4
```

The Environment pane on the right shows the current environment with the following files:

Staged	Status	Path
<input type="checkbox"/>	?	.gitignore
<input type="checkbox"/>	M	README.md
<input type="checkbox"/>	?	test.Rproj

The Files pane on the right shows the file explorer for the `test` directory:

Name	Size	Modified
..		
.gitignore	40 B	Jan 8, 2019, 12:43 PM
test.Rproj	205 B	Jan 8, 2019, 12:43 PM
README.md	35 B	Jan 8, 2019, 1:18 PM

The Console pane at the bottom shows the R terminal output:

```
~/Desktop/test/
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

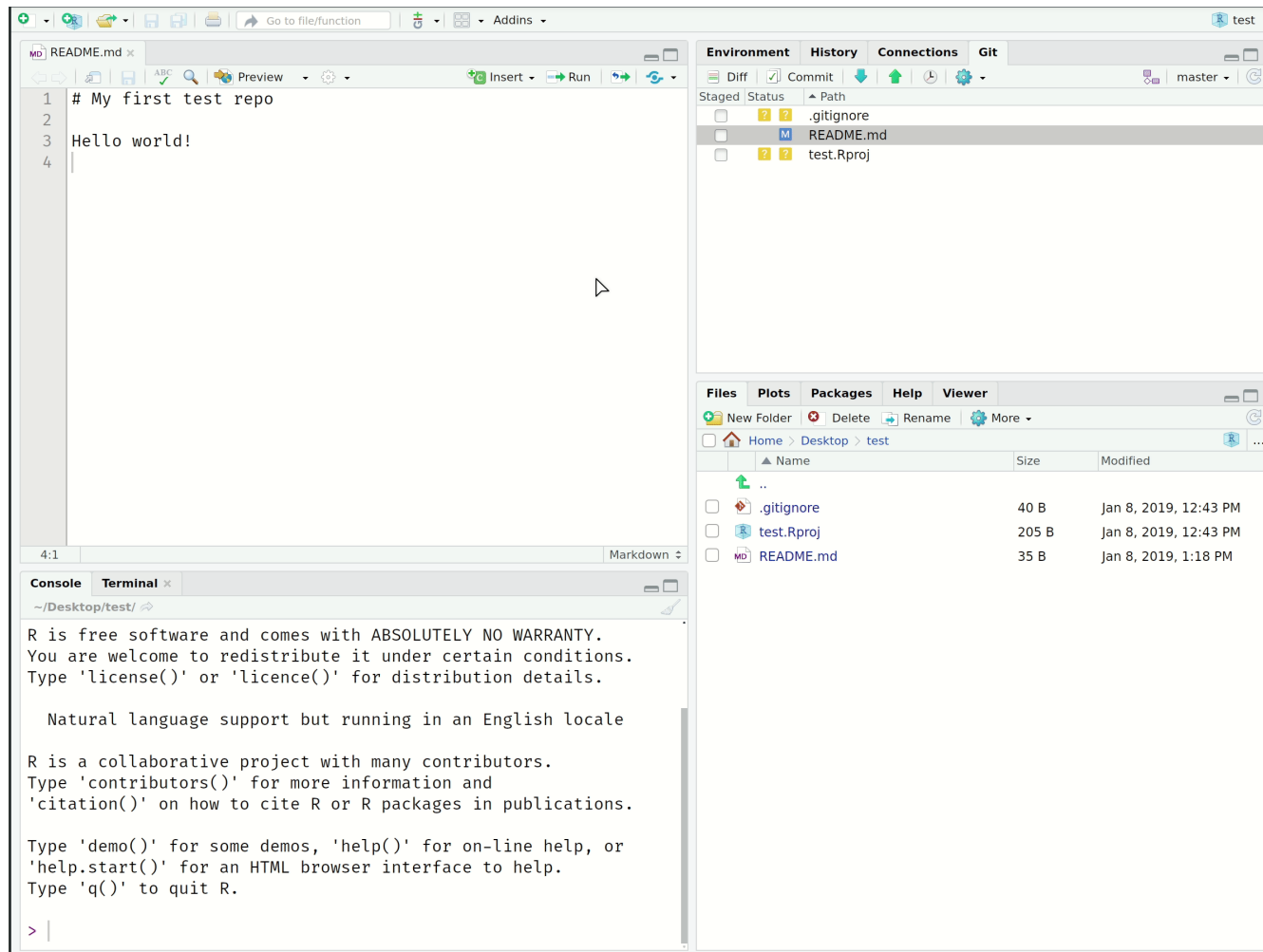
Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

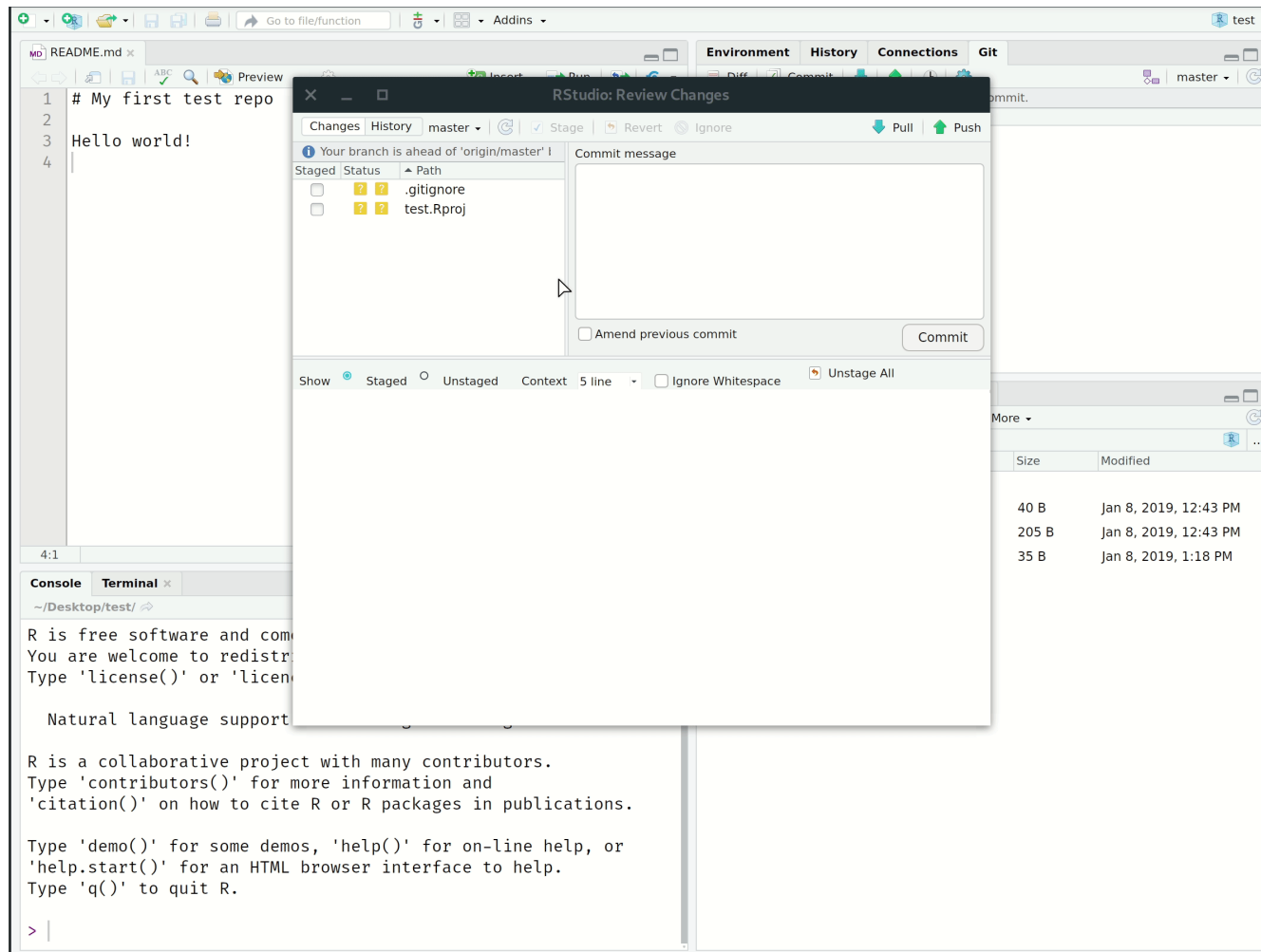
> |
```

# Stage and Commit

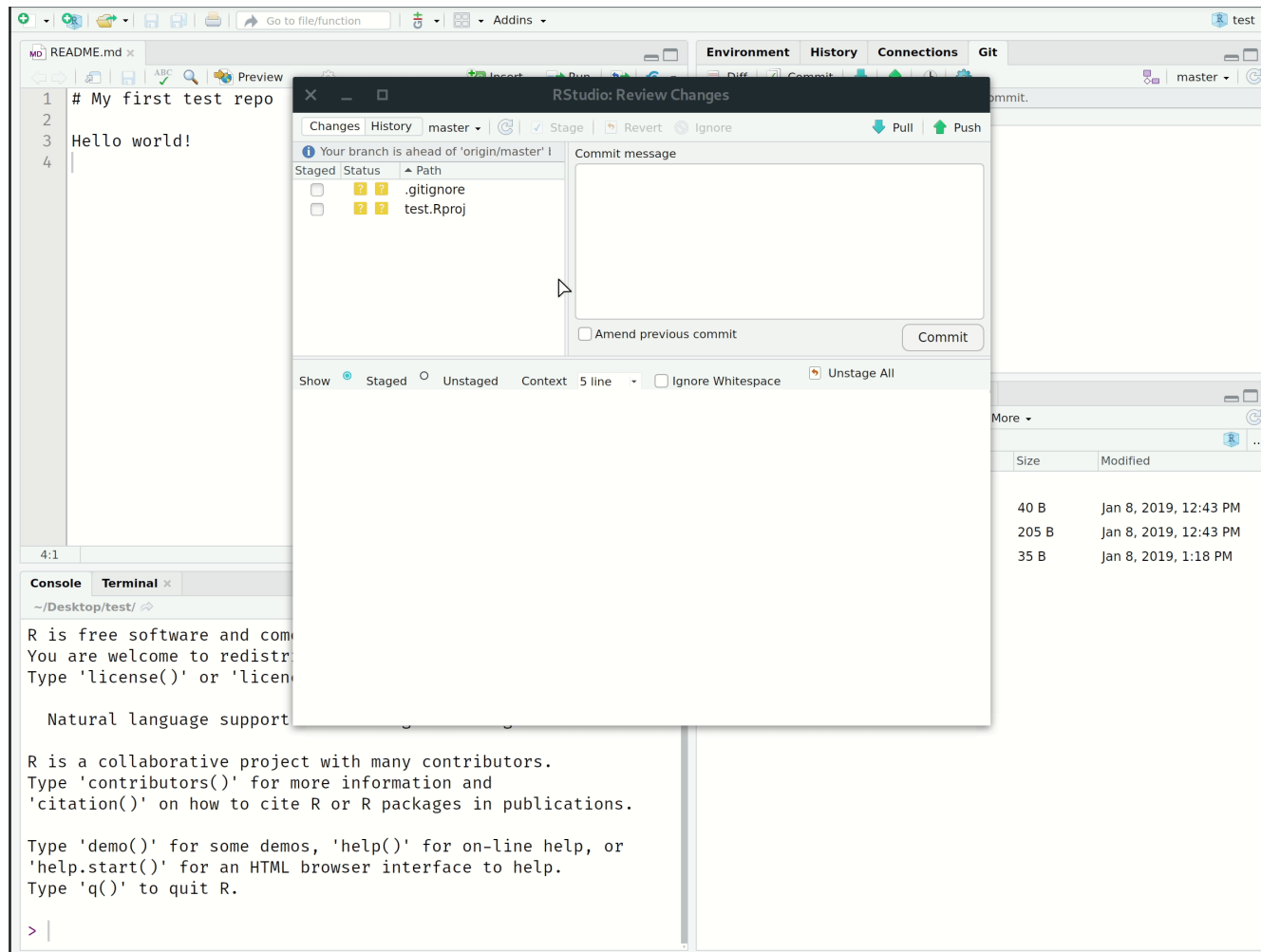


Note the helpful commit message to ourselves.

# Push and Pull



# Push and Pull



See [here](#) if you get Error: unable to read askpass response from 'rpostback-askpass'.

# Recap

Here's a step-by-step summary of what we just did.

- Made same changes to a file and saved them locally.
- *Staged* these local changes.
- *Committed* these local changes to our Git history with a helpful message.
- *Pulled* from the GitHub repo just in case anyone else made changes too (not expected here, but good practice).
- *Pushed* our changes to the GitHub repo.

**Note:** Always pull from the upstream repo *before* you push any changes. Seriously, do this even on solo projects; making it a habit will save you headaches down the road.

# Why this workflow?

Creating the repo on GitHub first means that it will always be "upstream" of your (and any other) local copies.

- In effect, this allows GitHub to act as the central node in the distributed VC network.
- Especially valuable when you are collaborating on a project with others — more on this later — but also has advantages when you are working alone.
- If you would like to move an existing project to GitHub, my advice is still to create an empty repo there first, clone it locally, and then copy all your files across.

RStudio Projects are great.

- Again, they interact seamlessly with Git(Hub), as we've just seen.
- They also solve absolute vs. relative path problems, since the **.Rproj file acts as an anchor point for all other files in the repo.**<sup>1</sup>

<sup>1</sup> You know that calling files from `YourComputer/YourName/Documents/Special-Subfolder/etc` in your scripts makes you a bad person, right?



# Git from the shell

---

# Why bother with the shell?

The GitHub + RStudio Project combo is ideal for new users.

- RStudio's Git integration and built-in GUI cover all the major operations.
- RStudio Projects FTW.

However, I want to go over Git **shell** commands so that you can internalise the basics.

- The shell is more powerful and flexible. Does some things that the RStudio Git GUI can't.
- Potentially more appropriate for projects that aren't primarily based in R. (Although, no real harm in using RStudio Projects to clone a non-R repo.)
- Also, I don't want to screen record more

# Why bother with the shell?

The GitHub + RStudio Project combo is ideal for new users.

- RStudio's Git integration and built-in GUI cover all the major operations.
- RStudio Projects FTW.

However, I want to go over Git **shell** commands so that you can internalise the basics.

- The shell is more powerful and flexible. Does some things that the RStudio Git GUI can't.
- Potentially more appropriate for projects that aren't primarily based in R. (Although, no real harm in using RStudio Projects to clone a non-R repo.)
- Also, I don't want to screen record more

(Yes, we're going to cover the shell in depth next lecture, but think of this as a sneak preview.)

# Main Git shell commands

Clone a repo.

```
$ git clone git@REPOSITORY-URL
```

See the commit history (hit spacebar to scroll down or q to exit).

```
$ git log
```

What has changed?

```
$ git status
```

# Main Git shell commands (cont.)

Stage ("add") a file or group of files.

```
$ git add NAME-OF-FILE-OR-FOLDER
```

You can use **wildcard** characters to stage a group of files (e.g. sharing a common prefix). There are a bunch of useful flag options too:

- Stage all files.

```
$ git add -A
```

- Stage updated files only (modified or deleted, but not new).

```
$ git add -u
```

- Stage new files only (not updated).

```
$ git add .
```

# Main Git shell commands (cont.)

Commit your changes.

```
$ git commit -m "Helpful message"
```

Pull from the upstream repository (i.e. GitHub).

```
$ git pull
```

Push any local changes that you've committed to the upstream repo (i.e. GitHub).

```
$ git push
```

# Merge conflicts

---

# Collaboration time

Turn to the person next to you. You are now partners. (Congratulations.)

- P1: Invite P2 to join you as a collaborator on the "test" GitHub repo that you created earlier. (See the *Settings* tab of your repo.)
- P2: Clone P1's repo to your local machine.<sup>1</sup> Make some edits to the README (e.g. delete lines of text and add your own). Stage, commit and push these changes.
- P1: Make your own changes to the README on your local machine. Stage, commit, pull from the GitHub repo, and then try to push them.

<sup>1</sup> Change into a new directory first or give it a different name to avoid conflicts with your own "test" repo. Don't worry, Git tracking will still work if you change the repo name locally.



# Collaboration time

Turn to the person next to you. You are now partners. (Congratulations.)

- P1: Invite P2 to join you as a collaborator on the "test" GitHub repo that you created earlier. (See the *Settings* tab of your repo.)
- P2: Clone P1's repo to your local machine.<sup>1</sup> Make some edits to the README (e.g. delete lines of text and add your own). Stage, commit and push these changes.
- P1: Make your own changes to the README on your local machine. Stage, commit, pull from the GitHub repo, and then try to push them.

Did P1 encounter a `merge conflict` error?

- Good, that's what we were trying to trigger.
- Now, let's learn how to fix them.

<sup>1</sup> Change into a new directory first or give it a different name to avoid conflicts with your own "test" repo. Don't worry, Git tracking will still work if you change the repo name locally.

# Merge conflicts

Let's confirm what's going on.

```
$ git status
```

As part of the response, you should see something like:

```
Unmerged paths:
  (use "git add <file> ..." to mark resolution)

* both modified:   README.md
```

Git is protecting P1 by refusing the merge. It wants to make sure that you don't accidentally overwrite all of your changes by pulling P2's version of the README.

- In this case, the source of the problem was obvious. Once we start working on bigger projects, however, `git status` can provide a helpful summary to see which files are in conflict.

# Merge conflicts (cont.)

Okay, let's see what's happening here by opening up the README file. RStudio is a good choice, although your preferred text editor is fine.<sup>1</sup>

You should see something like:

```
# README
Some text here.
<<<<<<< HEAD
Text added by Partner 2.
=====
Text added by Partner 1.
>>>>>> 814e09178910383c128045ce67a58c9c1df3f558.
More text here.
```

<sup>1</sup> Other good choices are [VS Code](#) or [Atom](#), which both support native Git(Hub) integration. You can set your preferred default editor with `$ git config --global core.editor "PREFERRED_EDITOR"`.

# Merge conflicts (cont.)

What do these symbols mean?

```
# README
Some text here.
<<<<<< HEAD
Text added by Partner 2.
=====
Text added by Partner 1.
>>>>>> 814e09178910383c128045ce67a58c9c1df3f558.
More text here.
```

# Merge conflicts (cont.)

What do these symbols mean?

```
# README
Some text here.
<<<<<< HEAD
Text added by Partner 2.
=====
Text added by Partner 1.
>>>>>> 814e09178910383c128045ce67a58c9c1df3f558.
More text here.
```

- <<<<<< HEAD Indicates the start of the merge conflict.

# Merge conflicts (cont.)

What do these symbols mean?

```
# README
Some text here.
<<<<<< HEAD
Text added by Partner 2.
=====
Text added by Partner 1.
>>>>>> 814e09178910383c128045ce67a58c9c1df3f558.
More text here.
```

- <<<<<< HEAD Indicates the start of the merge conflict.
- ===== Indicates the break point used for comparison.

# Merge conflicts (cont.)

What do these symbols mean?

```
# README
Some text here.
<<<<<< HEAD
Text added by Partner 2.
=====
Text added by Partner 1.
>>>>>> 814e09178910383c128045ce67a58c9c1df3f558.
More text here.
```

- <<<<<< HEAD Indicates the start of the merge conflict.
- ===== Indicates the break point used for comparison.
- >>>>>> <long string> Indicates the end of the lines that had a merge conflict.

# Merge conflicts (cont.)

Fixing these conflicts is a simple matter of (manually) editing the README file.

- Delete the lines of the text that you don't want.
- Then, delete the special Git merge conflict symbols.

Once that's done, you should be able to stage, commit, pull and finally push your changes to the GitHub repo without any errors.



# Merge conflicts (cont.)

Fixing these conflicts is a simple matter of (manually) editing the README file.

- Delete the lines of the text that you don't want.
- Then, delete the special Git merge conflict symbols.

Once that's done, you should be able to stage, commit, pull and finally push your changes to the GitHub repo without any errors.

## Caveats

- P1 gets to decide what to keep because they fixed the merge conflict.
- OTOH, the full commit history is preserved, so P2 can always recover their changes if desired.
- A more elegant and democratic solution to merge conflicts (and repo changes in general) is provided by Git **branches**. We'll get there next.

# Advice

- Merge conflicts can create a bit of a headache if you're not careful.
- Often, you can avoid merge conflicts by pulling from the upstream repo before you start working on your local copy of code.
  - This is especially true if you are working on a project with multiple collaborators.
  - Before you start editing code, always pull from the upstream repo to make sure that you have the latest version.
- If you do encounter a merge conflict, don't panic. It's not the end of the world.
  - Git is designed to preserve your work and the work of your collaborators.
  - You can always revert to a previous commit if you need to.
- Jenny Bryan has a great section on [Git Workflows](#) to help you use Git productively

# Aside: Line endings and different OSs

## Problem

During your collaboration, you may have encountered a situation where Git is highlighting differences on seemingly unchanged sentences.

- If that is the case, check whether your partner is using a different OS to you.

The "culprit" is the fact that Git evaluates an invisible character at the end of every line. This is how Git tracks changes. (More info [here](#) and [here](#).)

- For Linux and MacOS, that ending is "LF"
- For Windows, that ending is "CRLF" (of course it is...)

## Solution

Open up the shell and enter

```
$ git config --global core.autocrlf input
```

(Windows users: Change `input` to `true`).

# Branches and forking

---

# What are branches and why use them?

Branches are one of Git's coolest features.

- Allow you to take a snapshot of your existing repo and try out a whole new idea *without affecting* your main branch.<sup>1</sup>
- Only once you (and your collaborators) are 100% satisfied, would you merge it back into the main branch.<sup>2</sup>
  - This is how most new features in modern software and apps are developed.
  - It is also how bugs are caught and fixed.
  - But researchers can easily — and should! — use it to try out new ideas and analysis (e.g. robustness checks, revisions, etc.)
- If you aren't happy, then you can just delete the experimental branch and continue as if nothing happened.

<sup>1</sup> Github used to call the main branch "master", but has now switched to "main."

<sup>2</sup> You can actually have branches of branches (of branches). But let's not get ahead of ourselves.

# What are branches and why use them?

Branches are one of Git's coolest features.

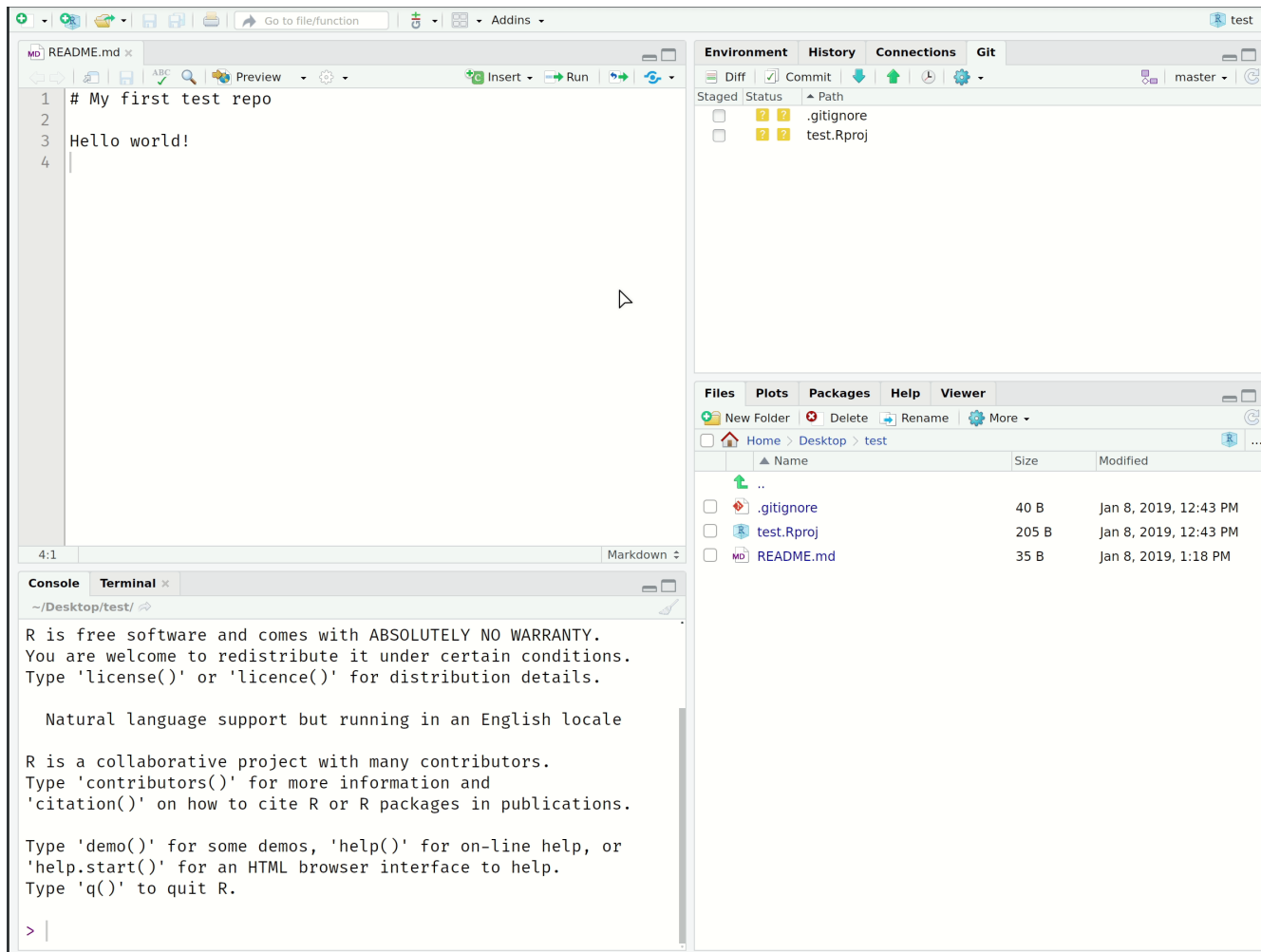
- Allow you to take a snapshot of your existing repo and try out a whole new idea *without affecting* your main branch.<sup>1</sup>
- Only once you (and your collaborators) are 100% satisfied, would you merge it back into the main branch.<sup>2</sup>
  - This is how most new features in modern software and apps are developed.
  - It is also how bugs are caught and fixed.
  - But researchers can easily — and should! — use it to try out new ideas and analysis (e.g. robustness checks, revisions, etc.)
- If you aren't happy, then you can just delete the experimental branch and continue as if nothing happened.

I use branches all the time for my own research projects.

<sup>1</sup> Github used to call the main branch "master", but has now switched to "main."

<sup>2</sup> You can actually have branches of branches (of branches). But let's not get ahead of ourselves.

# Create a new branch in RStudio



# Branch shell commands

Create a new branch on your local machine and switch to it:

```
$ git checkout -b NAME-OF-YOUR-NEW-BRANCH
```

Push the new branch to GitHub:

```
$ git push origin NAME-OF-YOUR-NEW-BRANCH
```

List all branches on your local machine:

```
$ git branch
```

Switch back to (e.g.) the main branch:

```
$ git checkout main
```

Delete a branch

```
$ git branch -d NAME-OF-YOUR-FAILED-BRANCH  
$ git push origin --delete NAME-OF-YOUR-FAILED-BRANCH
```



# Merging branches + Pull requests

You have two options:

## 1. Locally

- Commit your final changes to the new branch (say we call it "new-idea").
- Switch back to the main branch: `$ git checkout main`
- Merge in the new-idea branch changes: `$ git merge new-idea`
- Delete the new-idea branch (optional): `$ git branch -d new-idea`

## 2. Remotely (i.e. *pull requests* on GitHub)

- PRs are a way to notify collaborators — or yourself! — that you have completed a feature.
- You write a summary of all the changes contained in the branch.
- You then assign suggested reviewers of your code — including yourself potentially — who are then able to approve these changes ("Merge pull request") on GitHub.
- Let's practice this now in class...

# Your first pull request

You know that "new-idea" branch we just created a few slides back? Switch over to it if you haven't already.

- Remember: `$ git checkout new-idea` (or just click on the branches tab in RStudio)

Make some local changes and then commit + push them to GitHub.

- The changes themselves don't really matter. Add text to the README, add some new files, whatever.

After pushing these changes, head over to your repo on GitHub.

- You should see a new green button with "Compare & pull request". Click it.
- Add a meta description of what this PR accomplishes. You can also change the title if you want.
- Click "Create pull request".
- (Here's where you or your collaborators would review all the changes.)
- Once satisfied, click "Merge pull request" and then confirm.

# Your first pull request (cont.)

The screenshot shows the RStudio IDE interface for a new project named 'test'. The main editor window displays a file named 'README.md' with the following content:

```
1 # My first test repo
2
3 Hello world!
4
```

The console window at the bottom shows the R prompt and the following output:

```
~/Documents/Projects/test/
you are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

The right-hand pane shows the 'Files' tab, displaying a list of files in the project directory:

Name	Size	Modified
..		
.gitignore	40 B	Jan 9, 2020, 4
.Rhistory	0 B	Jan 9, 2020, 4
README.md	35 B	Jan 9, 2020, 9
test.Rproj	205 B	Jan 9, 2020, 9

# Forks

Git forks lie somewhere between cloning a repo and branching from it.

- In fact, if you fork a repo then you are really creating a copy of it.

Forking a repo on GitHub is [very simple](#); just click the "Fork" button in the top-right corner of said repo.

- This will create an independent copy of the repo under your GitHub account.
- Try this now. Use one of [the class repos](#) if you can't think of anyone else's.

Once you fork a repo, you are free to do anything you want to it. (It's yours.) However, forking — in combination with pull requests — is actually how much of the world's software is developed. For example:

- Outside user *B* forks *A*'s repo. She adds a new feature (or fixes a bug she's identified) and then [issues an upstream pull request](#).
- *A* is notified and can then decide whether to merge *B*'s contribution with the main project.

# Forks (cont.)

Creating forks is super easy as we've just seen. However, maintaining them involves some more leg work if you want to stay up to date with the original repo.

- GitHub: "[Syncing a fork](#)"
- OTOH, this isn't going to be an issue for completed projects. E.g. Forking the repo that contains the code and data of a published paper.

# Forks (cont.)

Creating forks is super easy as we've just seen. However, maintaining them involves some more leg work if you want to stay up to date with the original repo.

- GitHub: "[Syncing a fork](#)"
- OTOH, this isn't going to be an issue for completed projects. E.g. Forking the repo that contains the code and data of a published paper.

## Open Source Software contribution

Lots of Open Source Software (OSS) projects are hosted on GitHub. They rely on forks, branches, etc. to manage contributions from the community.

I will give anyone who successfully makes a contribution to an OSS project during the semester a bonus 2.5% on their final grade. (Link me to the PR and I'll verify.)

- Grades aside, I want to encourage you to start thinking about contributing to software projects in general.
- Seriously, it can be something as simple as correcting typos or language. Many great programmers and data scientists are not English first-language speakers. Helping to improve package documentation is a small way to say thanks. (More [here](#).)

# Other tips

---

# README

README files are special in GitHub because they act as repo landing pages.

- For a project tied to a research paper, this is where you should be explicit about the goal of the research paper, the software requirements, how to run the analysis, and so forth (e.g. [here](#)).
- On the other end of the scale, many GitHub repos are basically standalone README files. Think of these as version-controlled blog posts (e.g. [here](#)).

README files can also be added to the *sub-directories* of a repo, where they will act as a landing pages too.

- Particularly useful for bigger projects. Say, where you are using multiple programming languages (e.g. [here](#)), or want to add more detail about a dataset (e.g. [here](#)).

READMEs should be written in Markdown, which GH automatically renders.

- We'll learn more about [Markdown](#) (and its close relation, [R Markdown](#)) during the course of our homework assignments.



# .gitignore

A .gitignore file tells Git what to — *wait for it* — ignore.

This is especially useful if you want to exclude whole folders or a class of files (e.g. based on size or type).

- Proprietary data files should be ignored from the beginning if you intend to make a repo public at some point.
- Very large individual files (>100 MB) exceed GitHub's maximum allowable size and should be ignored regardless. See [here](#) and [here](#).

I typically add compiled datasets to my .gitignore in the early stages of a project.

- Reduces redundant version control history, where the main thing is the code that produces the compiled dataset, not the end CSV in of itself. ("Source is real.")
- Simple to remove from my .gitignore once the project is being finalised (e.g. paper is being submitted).

# .gitignore (cont.)

You can create a .gitignore file in multiple ways.

- A .gitignore file was automatically generated if you cloned your repo with an RStudio Project.
- You could also have the option of adding one when you first create a repo on GitHub.
- Or, you can create one with your preferred text editor. (Must be saved as ".gitignore".)

Once the .gitignore file is created, simply add in lines of text corresponding to the files that should be ignored.

- To ignore a single a file: `FILE-I-WANT-TO-IGNORE.csv`
- To ignore a whole folder (and all of its contents, subfolders, etc.): `FOLDER-NAME/**`
- The standard shell commands and special characters apply.
  - E.g. Ignore all CSV files in the repo: `*.csv`
  - E.g. Ignore all files beginning with "test": `test*`
  - E.g. Don't ignore a particular file: `!somefile.txt`

# GitHub Codespaces

- GitHub Codespaces is a new feature that allows you to code directly in the browser
- All of you have access to this as members of the class GitHub organization and/or if you registered with GitHub Education
- You can access it by clicking the `Code` dropdown menu > `Codespaces` tab > `Create codespace on main` (or whatever branch you prefer)
- It is a great tool to fiddle with a repository in a browser without having to install anything on your local machine
- It is also a great tool to test drive a coding project in a controlled environment
- In this class, there is a port to a server version of RStudio under the `Ports` tab in the bottom panel
  - The port is labeled `Rstudio`
  - The username/password are "rstudio" and "rstudio"
  - I show a gif of this use case on the next slide

# GitHub Codespace/RStudio launch

The screenshot shows the GitHub repository page for 'big-data-and-economics/big-data-class-materials'. The repository is public and has 1 branch and 0 tags. The commit history shows a recent commit by kgcsport updating github slides. The file list includes folders like .devcontainer, Class Notes, People, img, lectures, literature, mcdermott, ransom, rubin/notes, syllabus, tools, and files like .gitattributes and .gitignore. The right sidebar shows repository details like README, MIT license, Activity, Custom properties, 5 stars, 1 watching, 20 forks, and a list of releases and packages.

github.com/big-data-and-economics/big-data-class-materials

big-data-and-economics / big-data-class-materials

Code Issues Pull requests Actions Projects Security Insights Settings

big-data-class-materials Public

Edit Pins Unwatch 1 Fork 20 Star 5

main 1 Branch 0 Tags

Go to file Add file Code

kgcsport updated github slides deda3cc · yesterday 970 Commits

File	Commit	Time
.devcontainer	dockerrifle fix	last week
Class Notes	Revert "Merge branch 'big-data-and-economics:main' into ..."	3 months ago
People	Revert "Merge branch 'big-data-and-economics:main' into ..."	3 months ago
img	adding the codespaces pieces	3 months ago
lectures	updated github slides	yesterday
literature	causal forest example	2 months ago
mcdermott	Revert "Merge branch 'main' into main"	3 months ago
ransom	upload planning branch	3 months ago
rubin/notes	Revert "Merge branch 'main' into main"	3 months ago
syllabus	charlies hours	yesterday
tools	update folder creator	2 months ago
.gitattributes	Initial commit	5 months ago
.gitignore	Trump lies csv	4 months ago

About

Bates ECON 368 Big Data and Economics syllabus and lectures

Readme MIT license Activity Custom properties 5 stars 1 watching 20 forks Report repository

Releases

No releases published Create a new release

Packages

No packages published Publish your first package

Contributors 46

# Git source control in VS Code

- VS Code has built-in [Git source control integration](#).
- This is what you will use in GitHub Codespaces
- The "Git" vocabulary is the same as what we've learned in RStudio, but the UI is different.
- You will need to `git fetch` to pull from the upstream repository (i.e. GitHub)
- It is built a little smarter and will automatically pull, then push when you select "sync" from the UI
- This is not exactly "best practice," but it does make it easier to get started with Git
- I show a `pull` and `sync` in the GIF on the next slide, but the documentation is linked above
- ALWAYS ALWAYS ALWAYS make sure you use source control to push/pull your changes after coding in the RStudio port

# Git source control gifs

The screenshot shows a GitHub Codespace environment. The top bar displays the browser address bar with the URL `studious-disco-76q4w76v7xh49x.github.dev`. Below the browser bar, a navigation bar shows various icons and a search bar. The main editor area is divided into two panes. The left pane, titled "SOURCE CONTROL", shows a commit message input field with the text "Message (Ctrl+Enter to commit on 'm...'" and a green "Commit" button. The right pane, titled "README.md", displays the content of the file. The content includes a title "# Class Materials for Bates ECON/DCS 368: Big Data and Economics", a link to the syllabus, a table of contents, a feedback link, office hours, and a link to make an appointment. The bottom pane is a terminal window showing the command prompt `@kcsport →/workspaces/big-data-class-materials (main) $`.

```
1 # Class Materials for Bates ECON/DCS 368: Big Data and Economics
2
3 [Full syllabus with official policies](https://github.com/big-data-and-economics/big-data-class-materials/blob/main/syllabus/syllabus.pdf)
4
5 ['Lectures'](#lectures) | ['Goals'](#goals-for-this-course) |
6 ['Other details'](#other-details) | ['FAQ'](#faq) | ['License'](#license)
7
8 # Feedback
9
10 I am constantly trying to improve this course. Provide [feedback](https://docs.google.com/forms/d/e/1FAIpQLScZyphM1fwb6GBH7HtGx4H_hwM6sGGVfZ3MXFnFLN1awQo0sQ/viewform?usp=sf_link).
11
12 # Office hours:
13 My office hours are:
14 - Tuesdays 4pm-5pm
15 - Wednesdays 10:30am-11:30am
16
17 You can make an appointment at [here](https://calendar.google.com/calendar/u/0/appointments/schedules/AcZss702UMZxGreYvp2MnV15VxKrIQN0XpFuu6v0I-1oX3ZIJ1E141M14Qh05FCPbw73KVLUj5FCRHm61).
18
19 ## Getting in touch
20
21 In this course, I ask that you use GitHub Discussions and 'Issues' to ask questions about the problem sets, final projects, presentation clarifications, and other class specifics. This is so that everyone can benefit from the answer. Also, it will encourage collaboration (and declutter my inbox). A portion of the grade is based on
```

# GitHub Issues

**GitHub Issues** are another great way to interact with your collaborators and/or package maintainers.

- If you spot any problems with these lecture notes, please file an issue [here!](#) (Keep in mind that is public!)

# Summary

---



# Recipe (shell commands in grey)

1. Create a repo on GitHub and initialize with a README.
2. Clone the repo to your local machine. Preferably using an RStudio Project, but as you wish. (E.g. Shell command: `$ git clone REPOSITORY-URL`)
3. Stage any changes you make: `$ git add -A`
4. Commit your changes: `$ git commit -m "Helpful message"`
5. Pull from GitHub: `$ git pull`
6. (Fix any merge conflicts.)
7. Push your changes to GitHub: `$ git push`

# Recipe (shell commands in grey)

1. Create a repo on GitHub and initialize with a README.
2. Clone the repo to your local machine. Preferably using an RStudio Project, but as you wish. (E.g. Shell command: `$ git clone REPOSITORY-URL`)
3. Stage any changes you make: `$ git add -A`
4. Commit your changes: `$ git commit -m "Helpful message"`
5. Pull from GitHub: `$ git pull`
6. (Fix any merge conflicts.)
7. Push your changes to GitHub: `$ git push`

Repeat steps 3–7 (but especially steps 3 and 4) often.

# Coming up

## Get the course materials

If you haven't done so already, now is a good time to clone/fork the course materials to your local computer.

## Problem Set 0

Problem Set 0 is due on Thursday.

## Problem Set 1

This will be posted later this week and is due at the end of the month.

## Next lecture

Learning to love the shell...

# Appendix: FAQ

---

**Q: When should I commit (and push) changes?**

**A: Early and often.**

- It's not quite as important as saving your work regularly, but it's a close second.
- You should certainly push everything that you want your collaborators to see.

**Q: Do I need branches if I am working on a solo project?**

**A: You don't *need* them, but they offer big advantages in maintaining a sane workflow.**

- Experiment without any risk to the main project!
- If you combine them with pull requests, then you can compress significant additions to your project (which may comprise many small edits) into a single branch.

# FAQ (cont.)

**Q: What's the difference between cloning and forking a repo?**

**A: Cloning directly ties your local version to the original repo, while forking creates a copy on your GitHub (which you can then clone).**

- **Cloning** makes it easier to fetch updates (and is often the best choice for new GitHub users), but **forking** has advantages too.

**Q: What happens when something goes wrong?**

**A: Think: "Oh shit, Git!"**

- Seriously: <http://ohshitgit.com/>.

**Q: What happens when something goes *horribly* wrong?**

**A: Burn it down and start again.**

- <http://happygitwithr.com/burn.html>
- This is a great advantage of Git's distributed nature. If something goes horribly wrong, there's usually an intact version somewhere else.

# FAQ (cont.)

