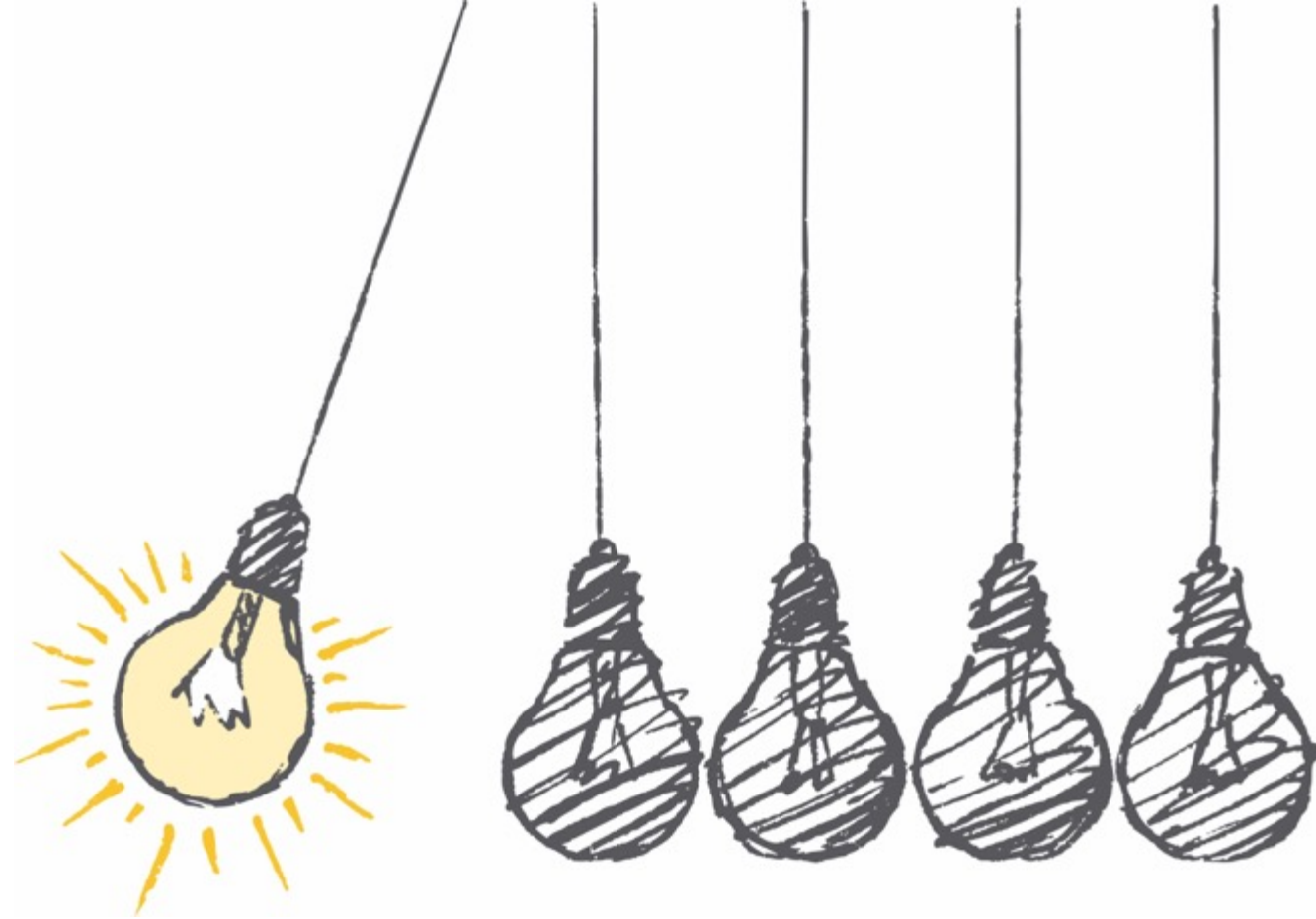


# Data Day:

## Introduction to Git and GitHub

Prepared by **DIME Analytics**  
[dimeanalytics@worldbank.org](mailto:dimeanalytics@worldbank.org)

Presented by **Benjamin Daniels**  
[bdaniels@worldbank.org](mailto:bdaniels@worldbank.org)



# What is Git?

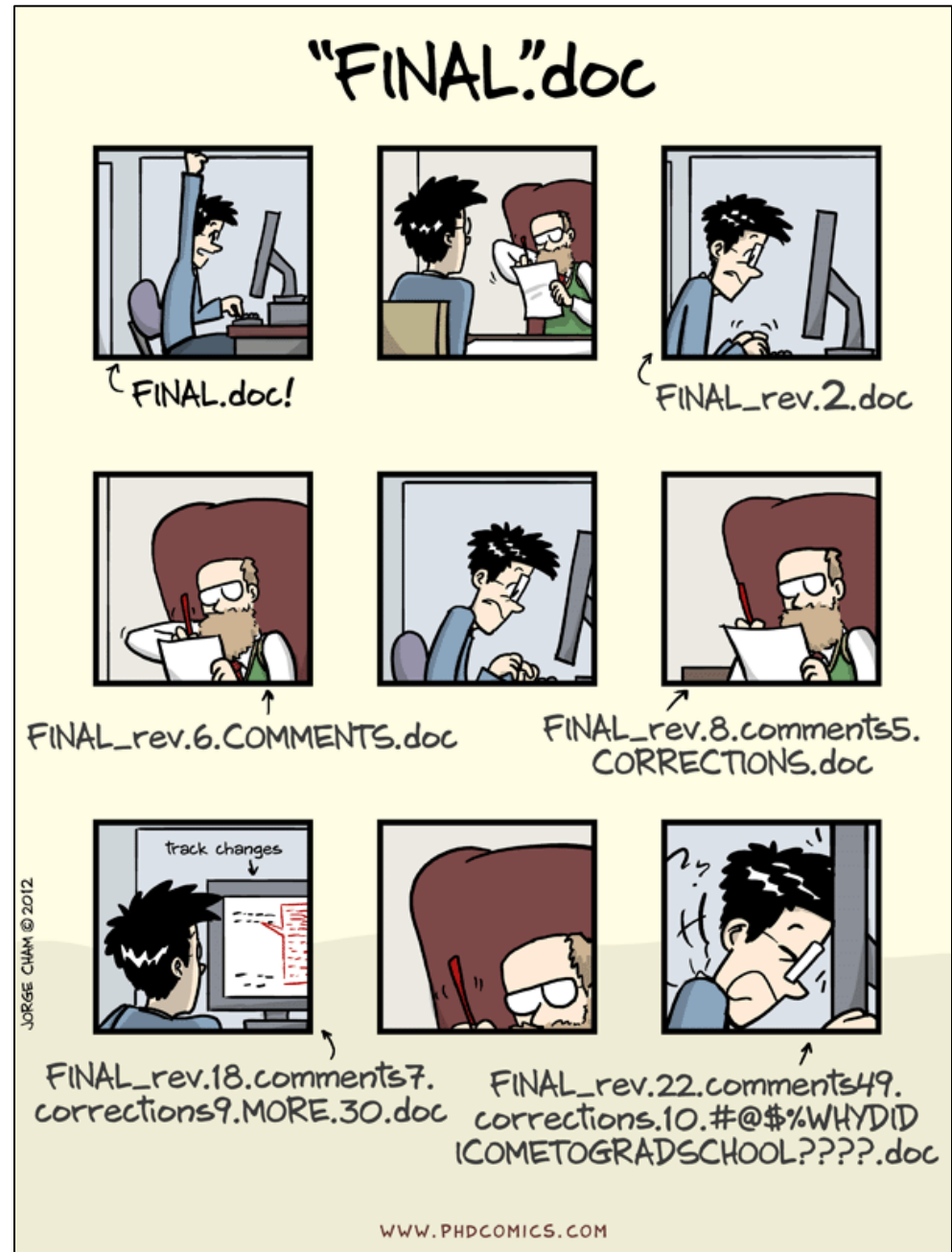
## What is GitHub?

First, Git is not GitHub.  
Second, GitHub is not Git.

Git is a **version control software**.  
It organizes *versions* of every file in a project in an orderly way. It is approved for use on WB machines.

GitHub is a **development platform**.  
It manages *contributions* to projects in an orderly way. It is a standard tool in code development and the Bank has a subscription.

Together, Git and GitHub allow you to avoid the "FINAL".doc problem.  
(For code, not Word documents)



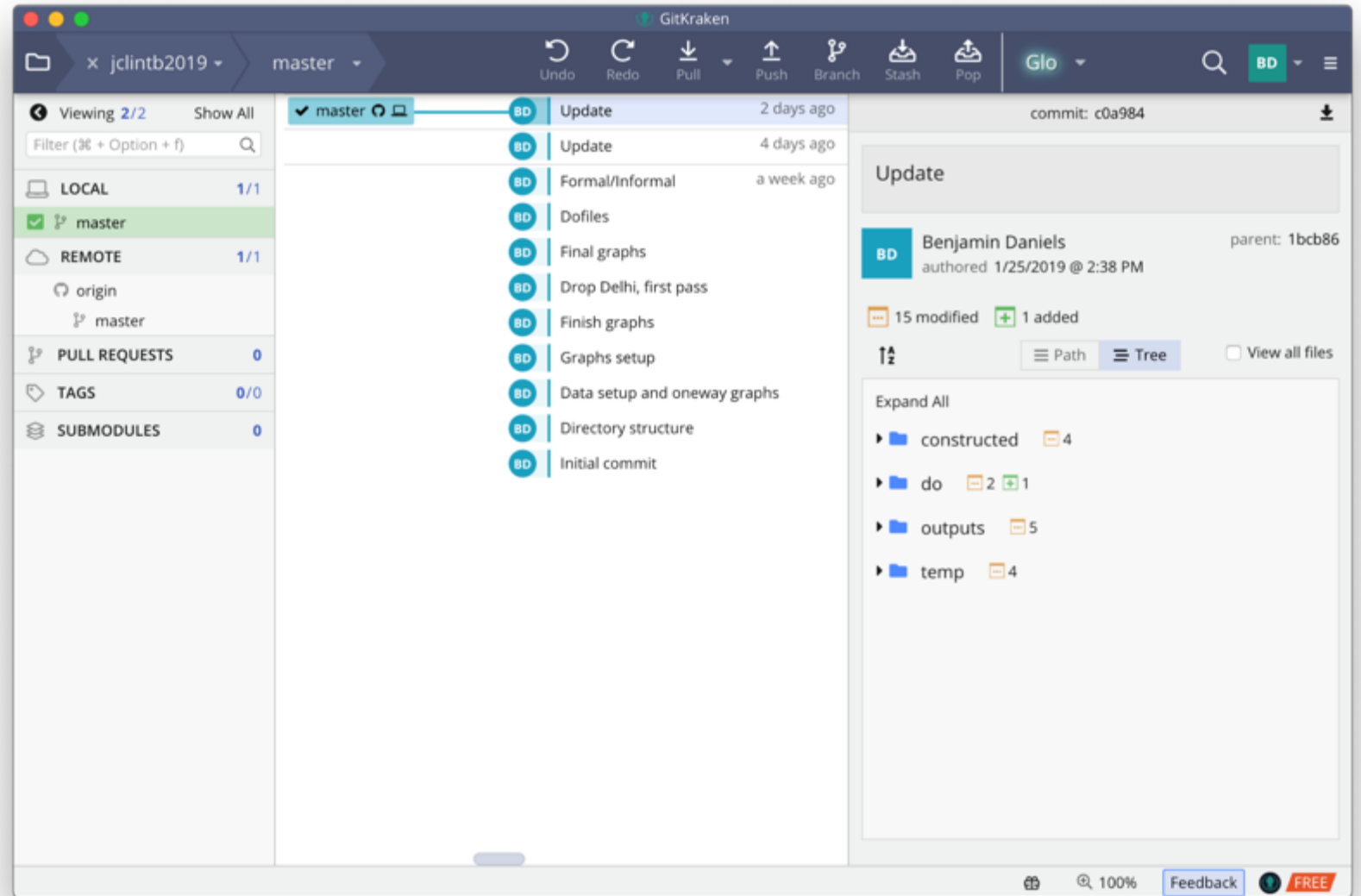
# Git:

## The basics

Git is a software that runs on your computer. It stores “version histories” of your files in a way that is really useful, but that you mainly don’t need to understand.

The complete history of a project is called a “repository”. A repository can be viewed in a desktop “client”, such as [GitKraken](#), shown here.

A very simple repository would look much like this one: it has a series of versions called *commits* that make up its history. Each commit is a record of the changes between itself and the previous commit.



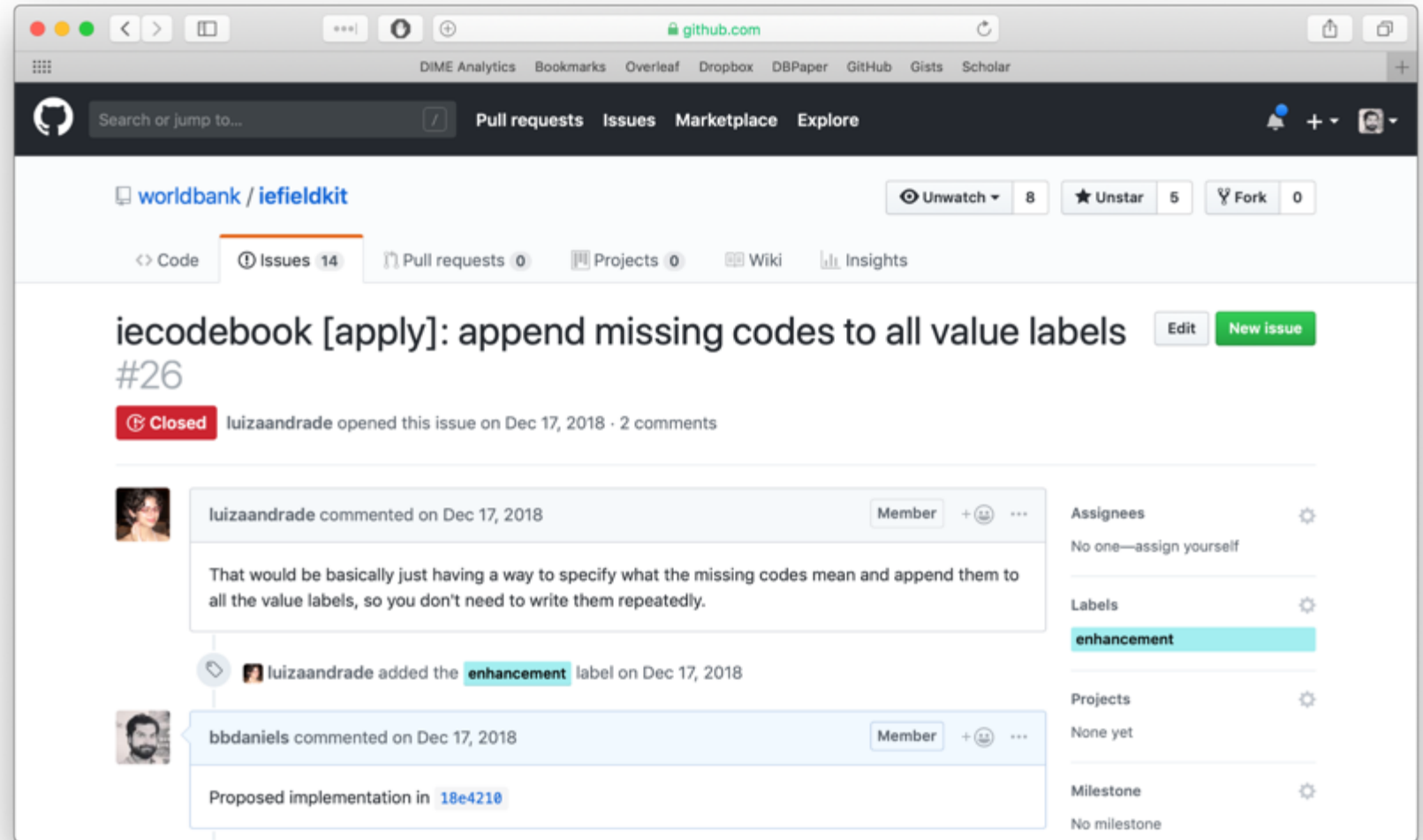
# GitHub:

## The basics

GitHub is a website that you can use to manage collaboration with other people.

This has two main forms: first, managing and tracking tasks (known generally as “issues”), and second, allowing people to submit new materials to the repository.

A very simple issue looks like this one: A problem is posed, and after some discussion a solution is submitted as a “pull request” to the repository.

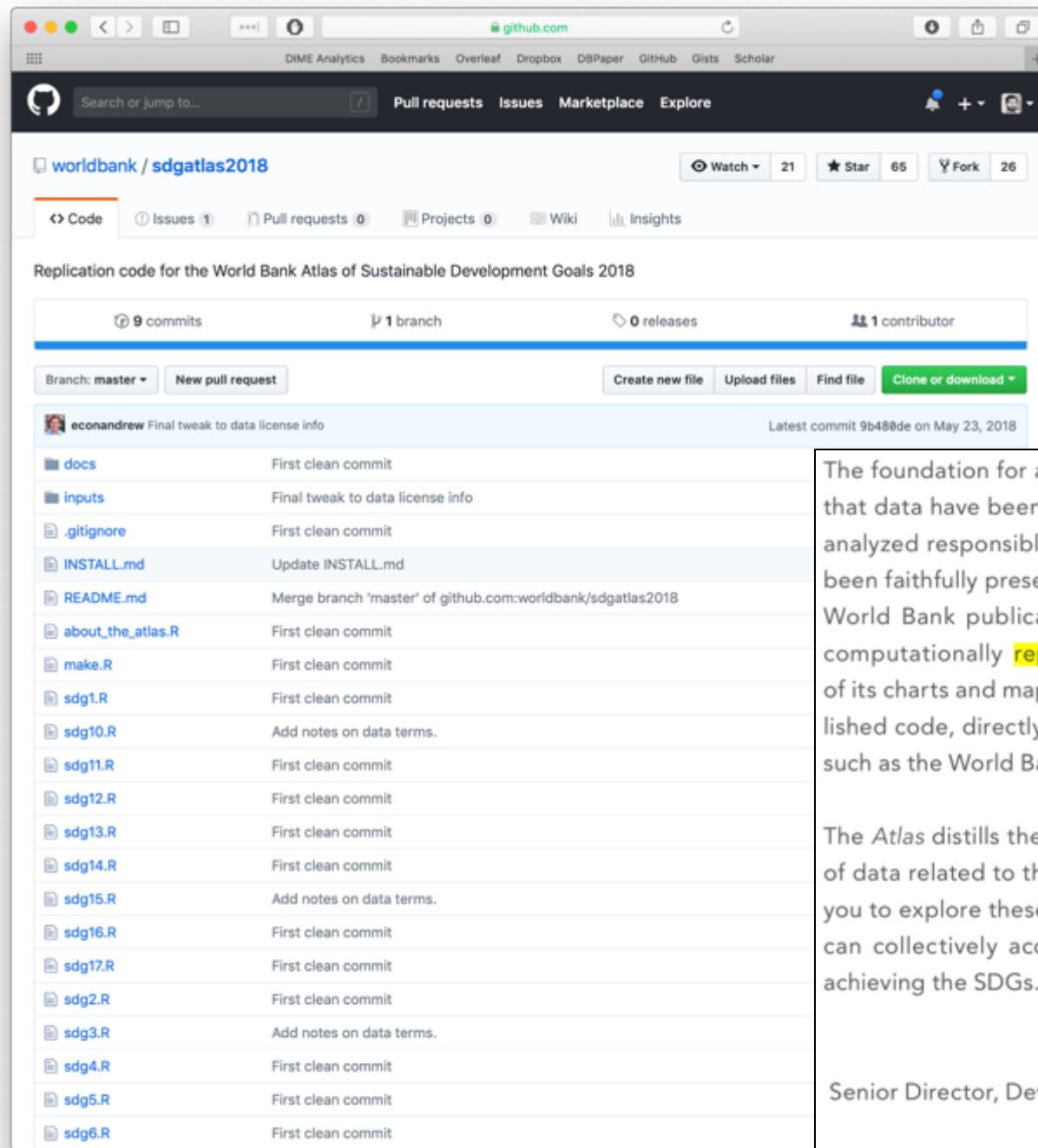


# Git & GitHub: The basics

Mastering Git with GitHub gives you an awesome tool for managing your own team's workflow. Having both your data, code, and *code history* available for all authors makes an incredibly useful tool.

Plus, Git protects all your files much better than other software.

Repositories can optionally be published publicly, which is increasing popular for process transparency. Shown here: The World Bank Atlas of Sustainable Development Goals.



The foundation for any evidence is trust: trust that data have been collected, managed, and analyzed responsibly and trust that they have been faithfully presented. The *Atlas* is the first World Bank publication that sets out to be computationally **reproducible**—the majority of its charts and maps are produced with published code, directly from public data sources such as the World Bank's Open Data platform.

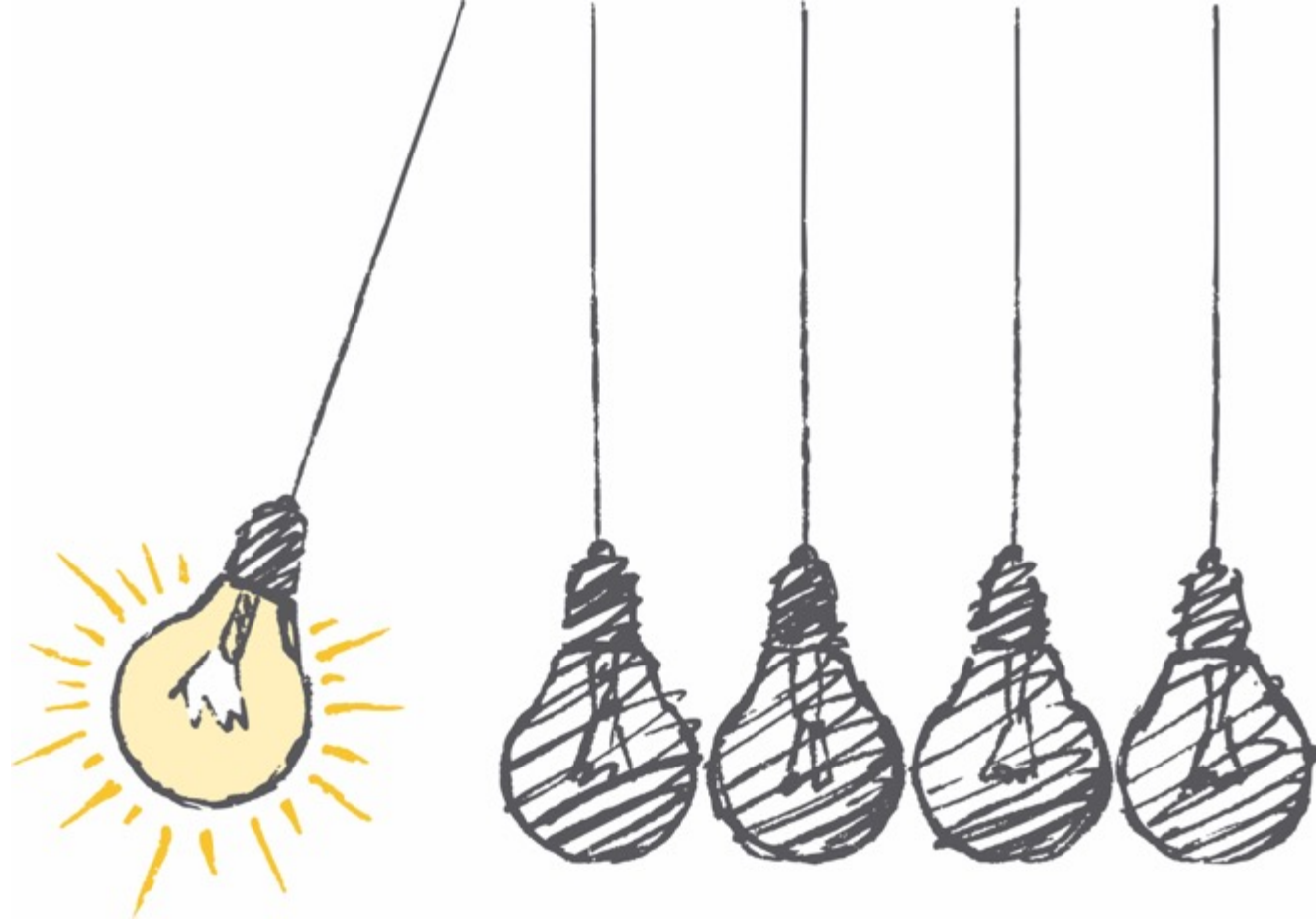
The *Atlas* distills the World Bank's knowledge of data related to the SDGs. I hope it inspires you to explore these issues further so that we can collectively accelerate progress toward achieving the SDGs.

Shanta Devarajan  
Senior Director, Development Economics and  
Acting Chief Economist  
World Bank Group



So how does it *work*?

## Part 1: Creating a repository and making commits



# Create a new repository on GitHub

The screenshot shows the GitHub 'Create a new repository' page. Four red arrows with numbers 1 through 4 point to the following elements:

- 1** Points to the 'Repository name' input field, which contains 'my-new-repo' and has a green checkmark.
- 2** Points to the 'Public' radio button, which is selected.
- 3** Points to the 'Initialize this repository with a README' checkbox, which is checked.
- 4** Points to the 'Create repository' button at the bottom.

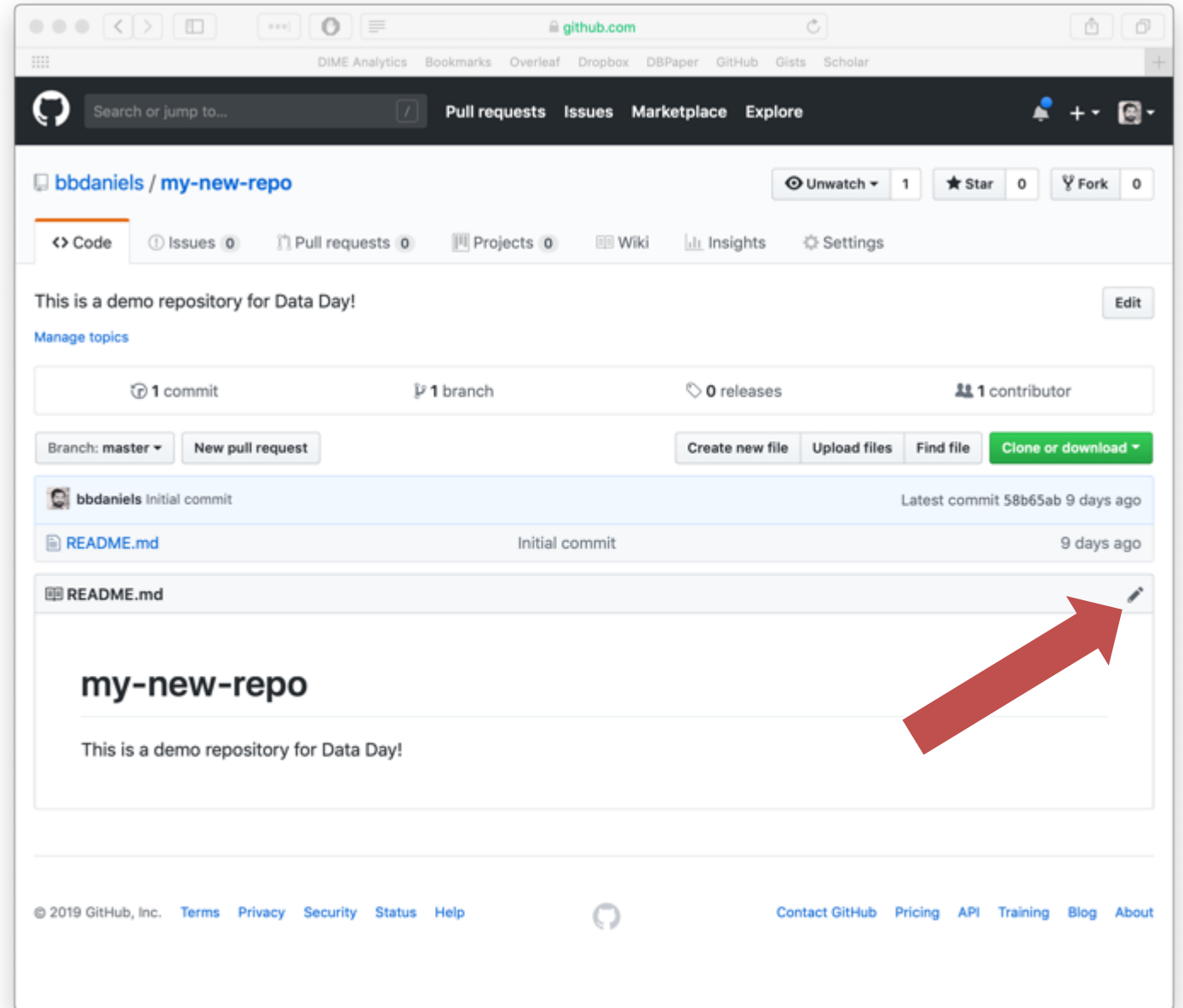
The page also includes a description field with the text 'This is a demo repository for Data Day!', a section for repository visibility (Public/Private), and a section for adding a .gitignore and a license.

# Your repository has a README

README.md is a common filename. In fact, it's so common that if you have one in a folder, GitHub will automatically display its contents when you are browsing that folder.

You should write short READMEs for every folder. Let's edit this one.

Click on the edit button, and add something like your name here.



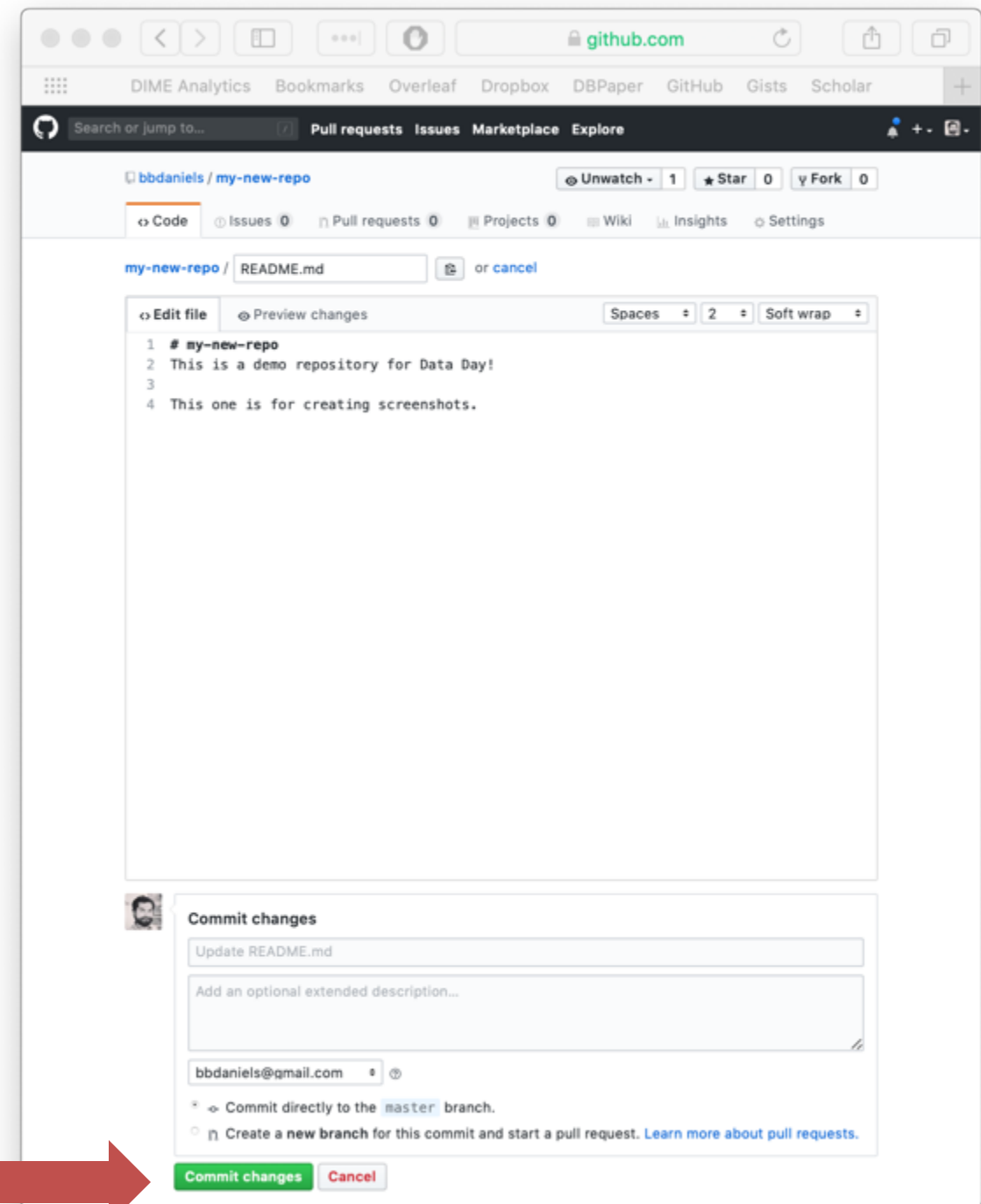


# Add some text and “Commit changes”

Unlike a regular editor, the built-in editor on GitHub has no “save” button. “Saving” is not something Git or GitHub know how to do.

Git and GitHub are NOT file managers, or code editors. Therefore you will rarely, if ever, make file changes from within Git or GitHub.

Most people already have a preferred way to edit their files. You can stick to yours without any trouble.



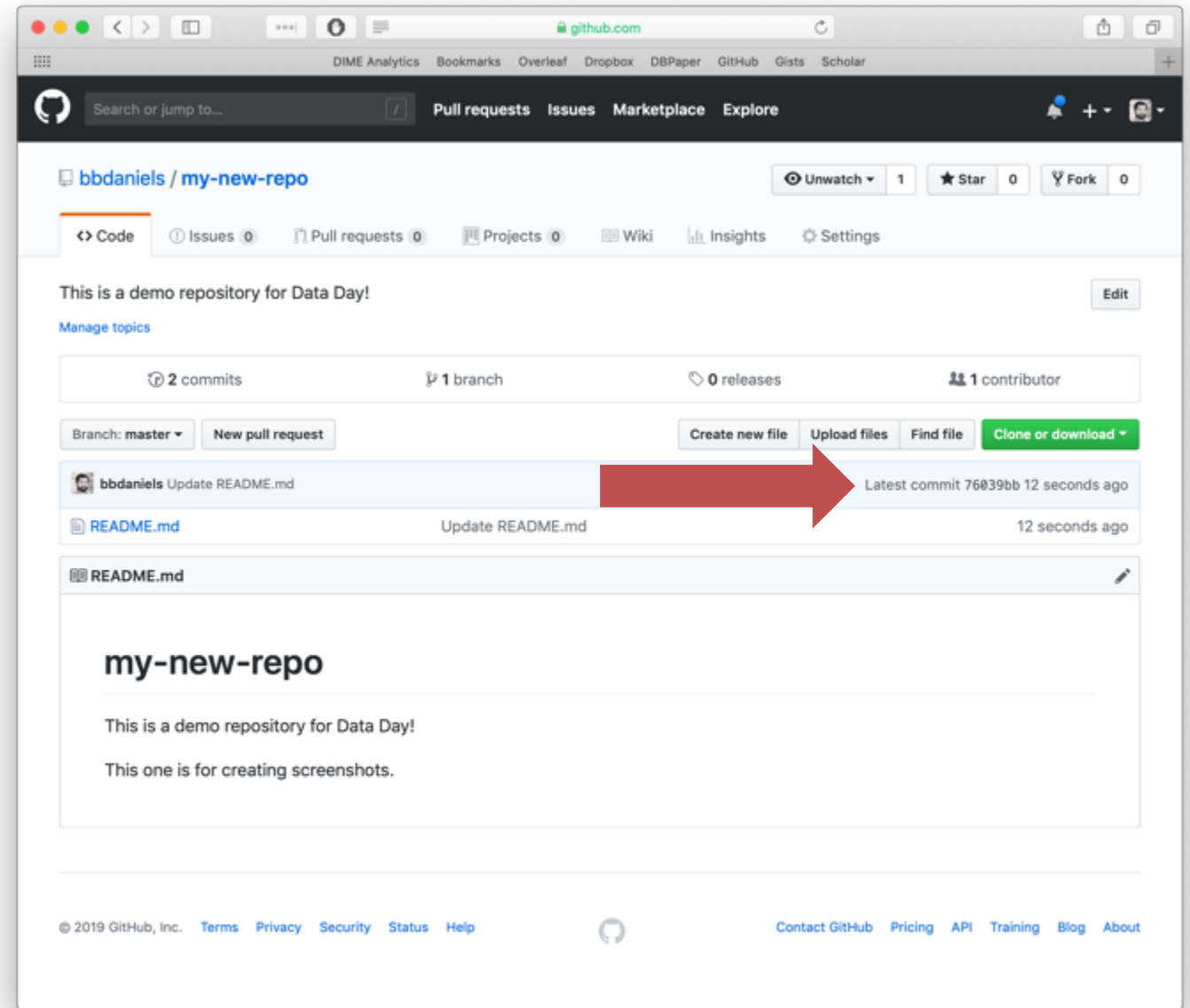
# The repo records changes for you

Instead of “saving” the file, you have done something called “committing changes”.

By doing that, you have simultaneously:

- Changed the contents of the file reflected by GitHub;
- Recorded when you did it; and
- Noted what you did (“Update README.md”)

This is Git fundamental feature #1.

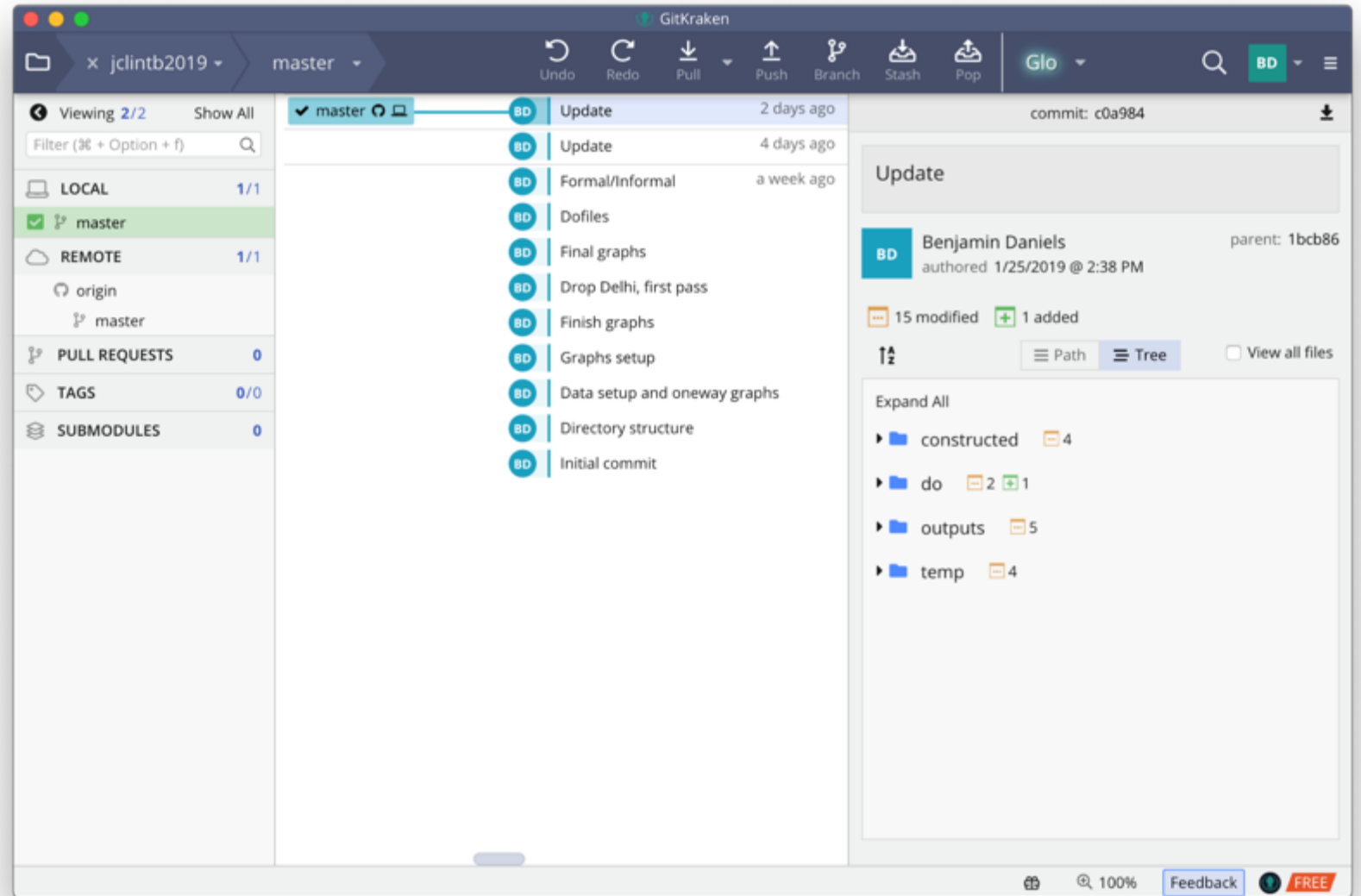


# Git creates commits

The most important thing Git does is create “commits”.

A commit is a snapshot of the entire working directory. Each commit has a *name* and a *timestamp*.

You have to tell Git when to create new commits. This is probably the biggest difference from whatever you do now.



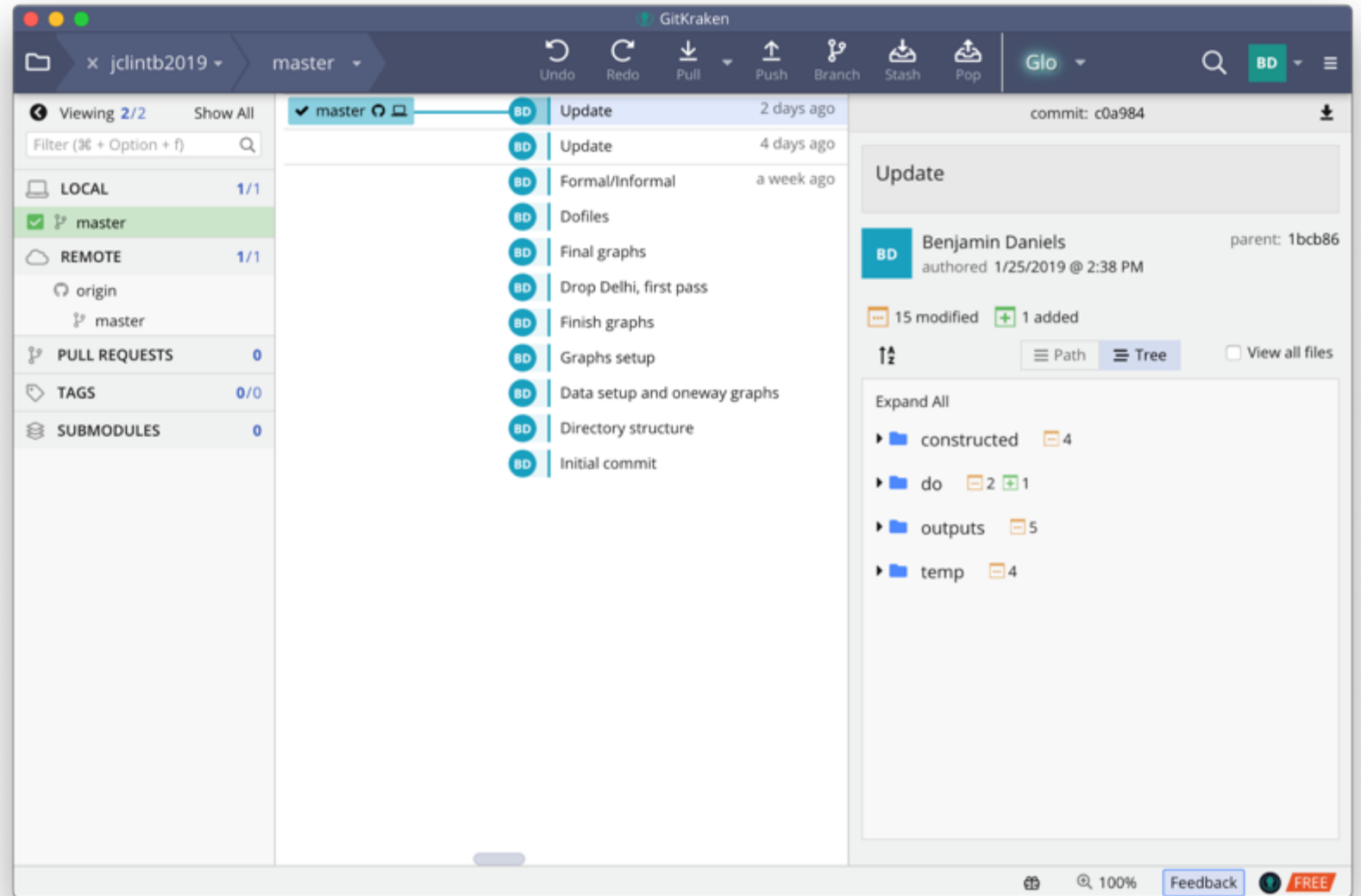
# Commits name past versions

Git *watches* your saved changes

But it only *stores* them when you tell it to: this is a “commit”

You can name these so that they make sense and are easy to find when you need them in the future (you *always* need them when you least expect, right?)

Required Reading: Git vs Dropbox  
<https://michaelstepner.com/blog/git-vs-dropbox>



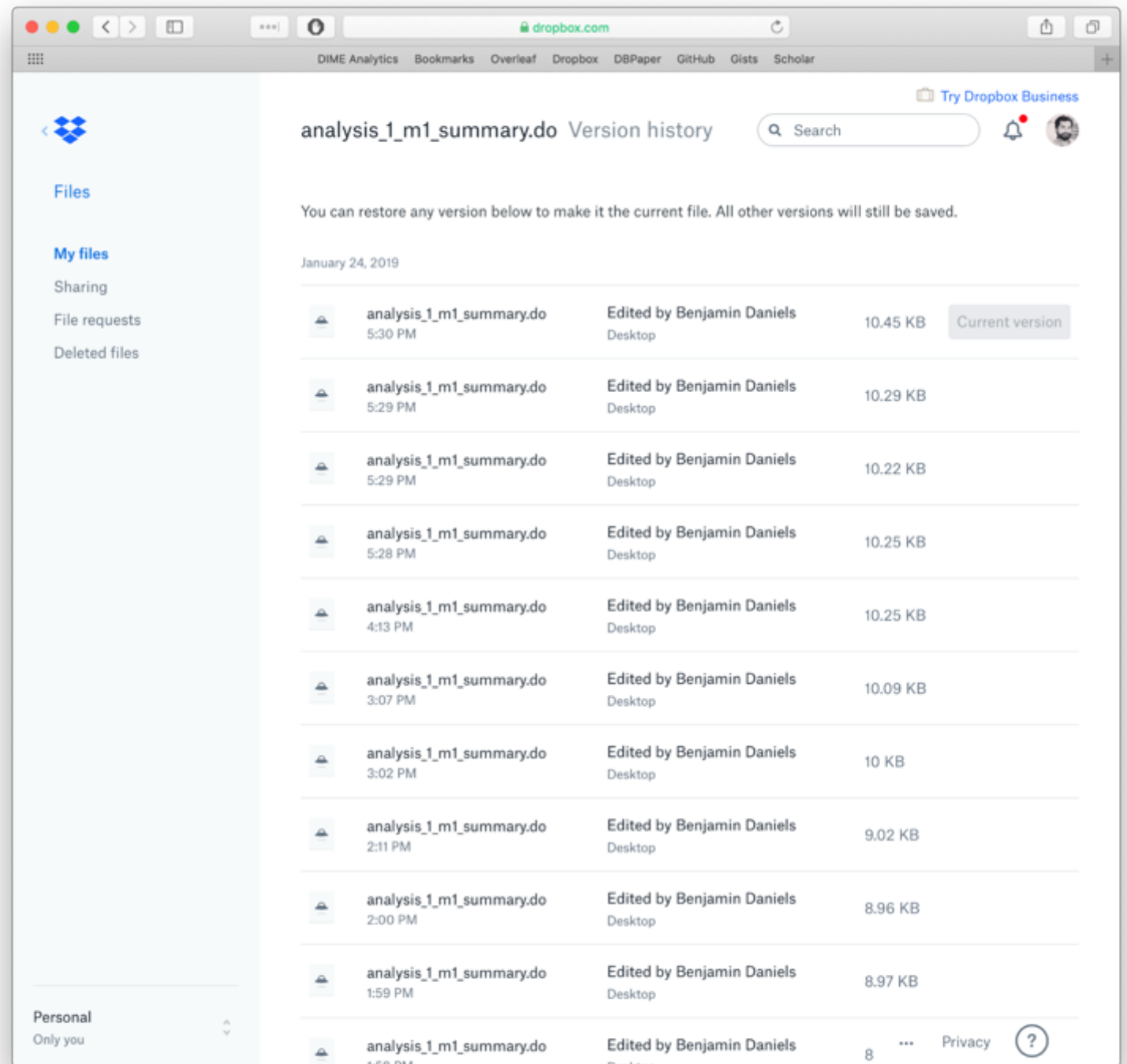
# How is that better than Dropbox?

Dropbox stores a new snapshot *every time you save* each file.

You save your work *often* (right?)

So Dropbox stores *many* similar images of your files

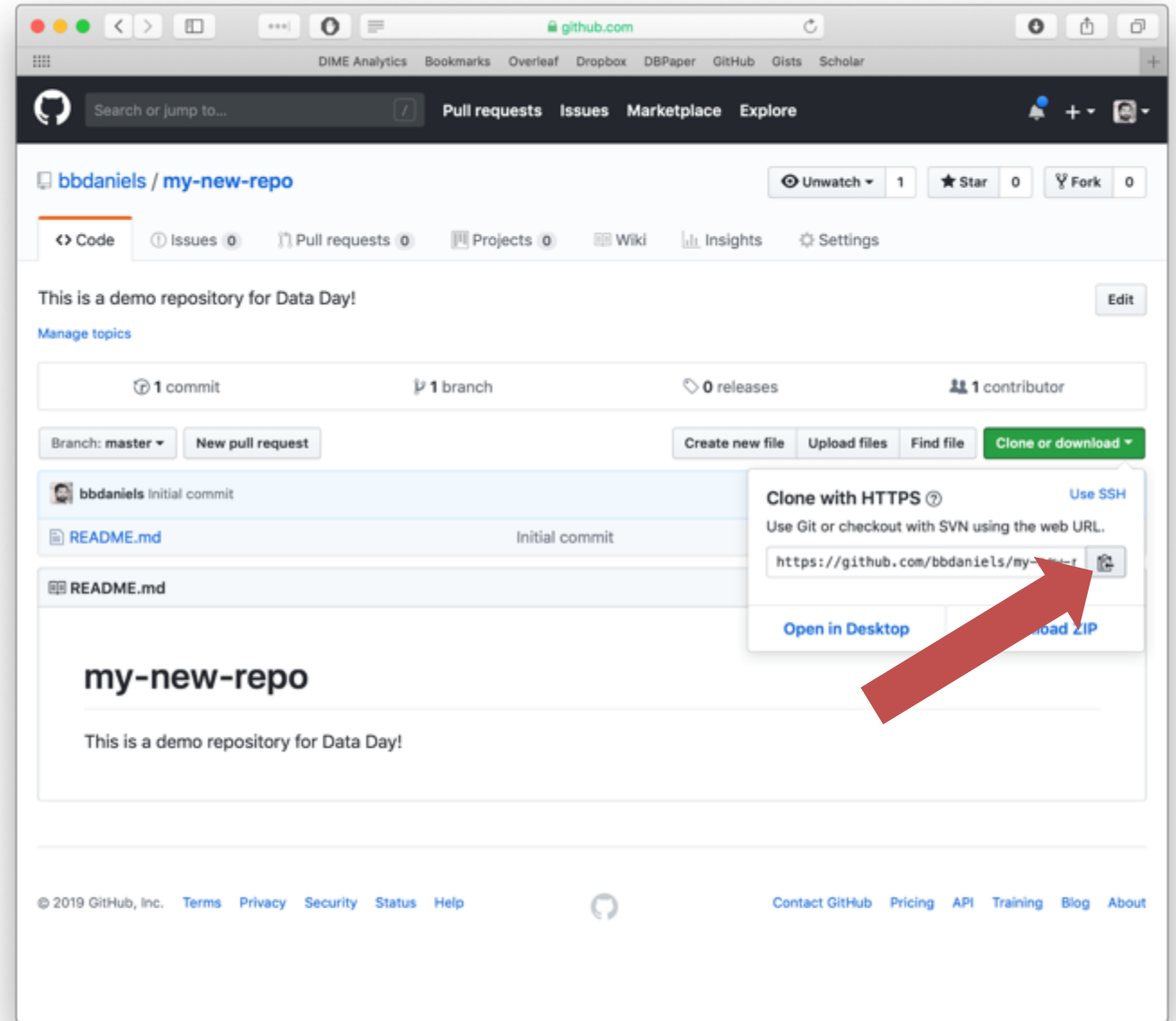
But you can't tell which is which!



# Clone the repo with your desktop client

Cloning the repository makes a complete copy of it at the new location, including the entire history (remember, the repository is the history).

Cloning from GitHub to your local computer is a good way to start a repository.



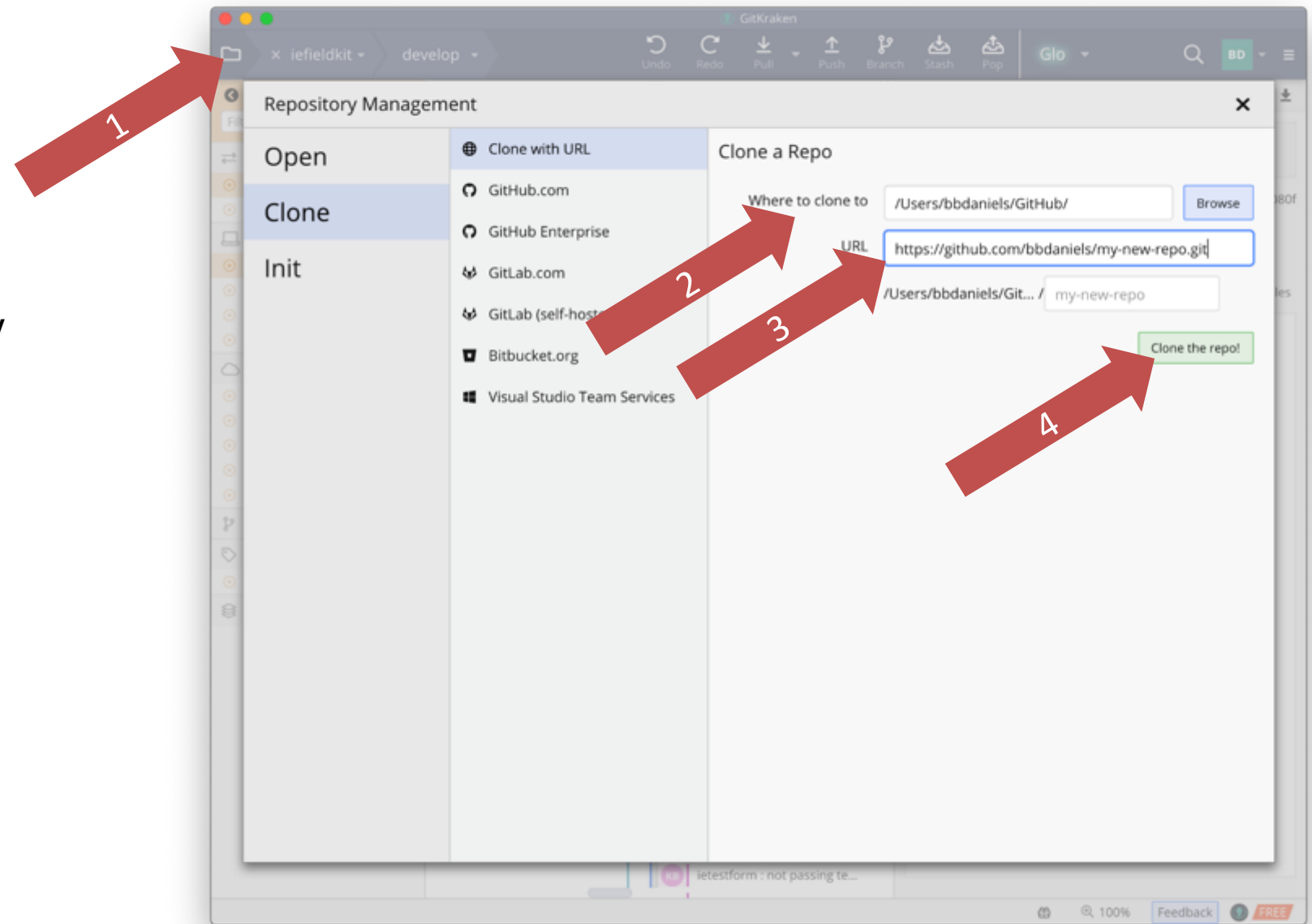


# Clone the repo with your desktop client

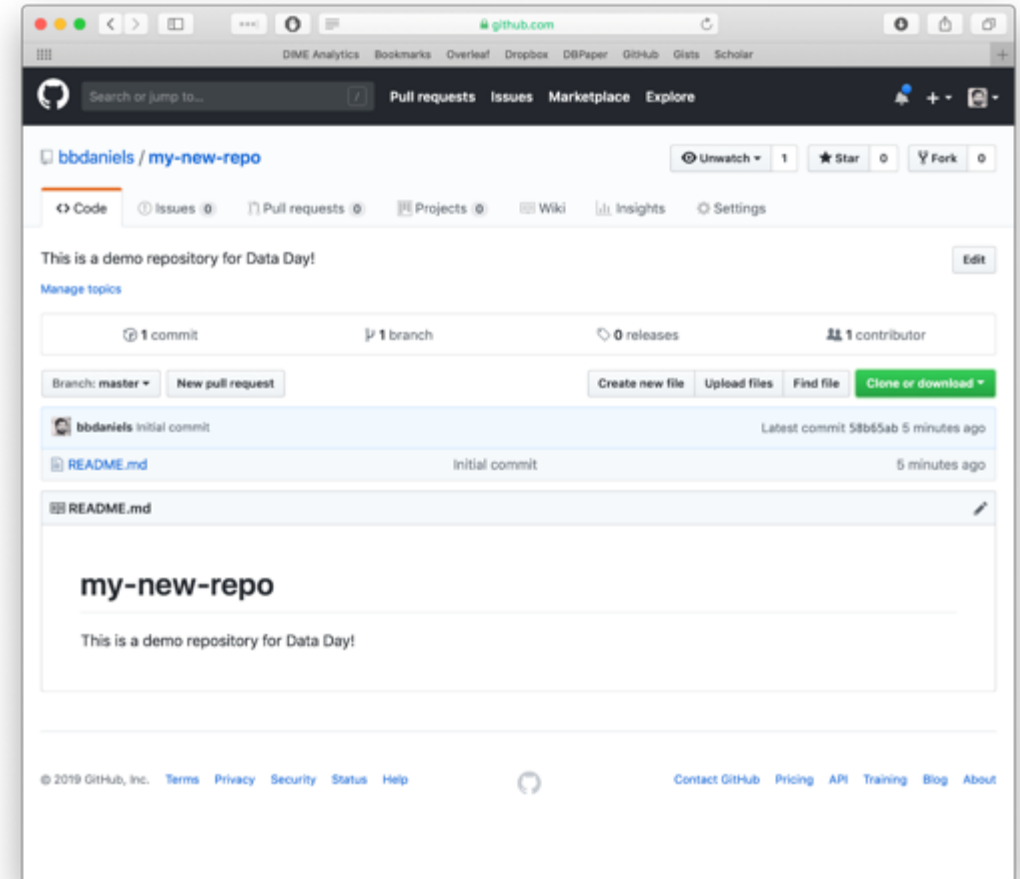
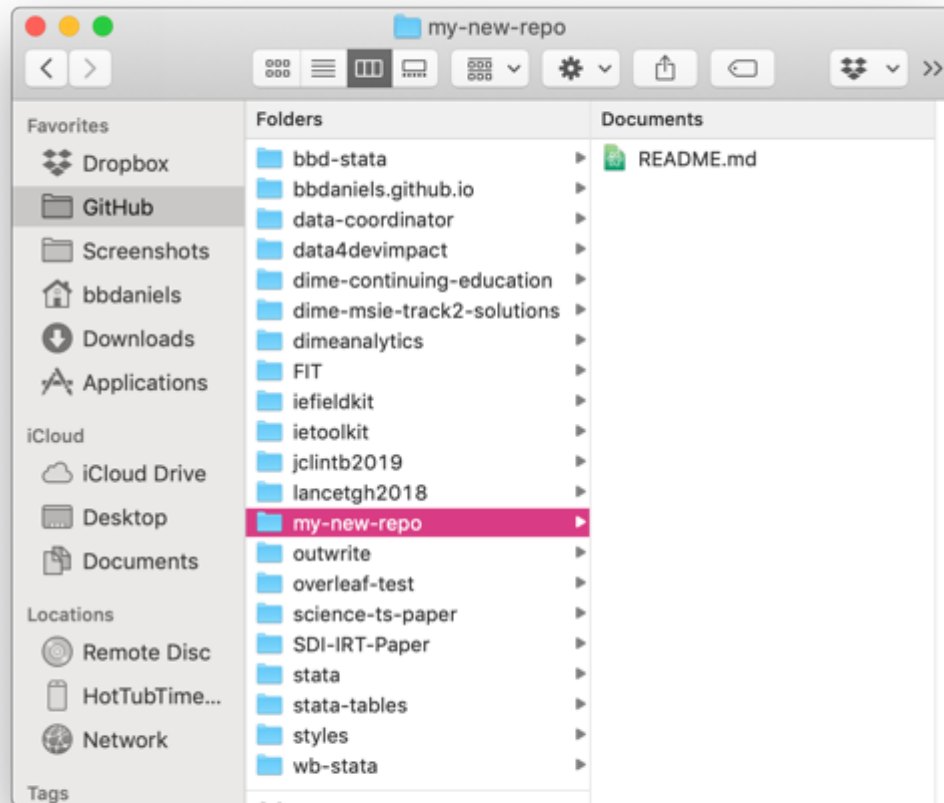
Cloning the repository makes a complete copy of it at the new location, including the entire history (remember, the repository is the history).

Cloning from GitHub to your local computer is a good way to start a repository.

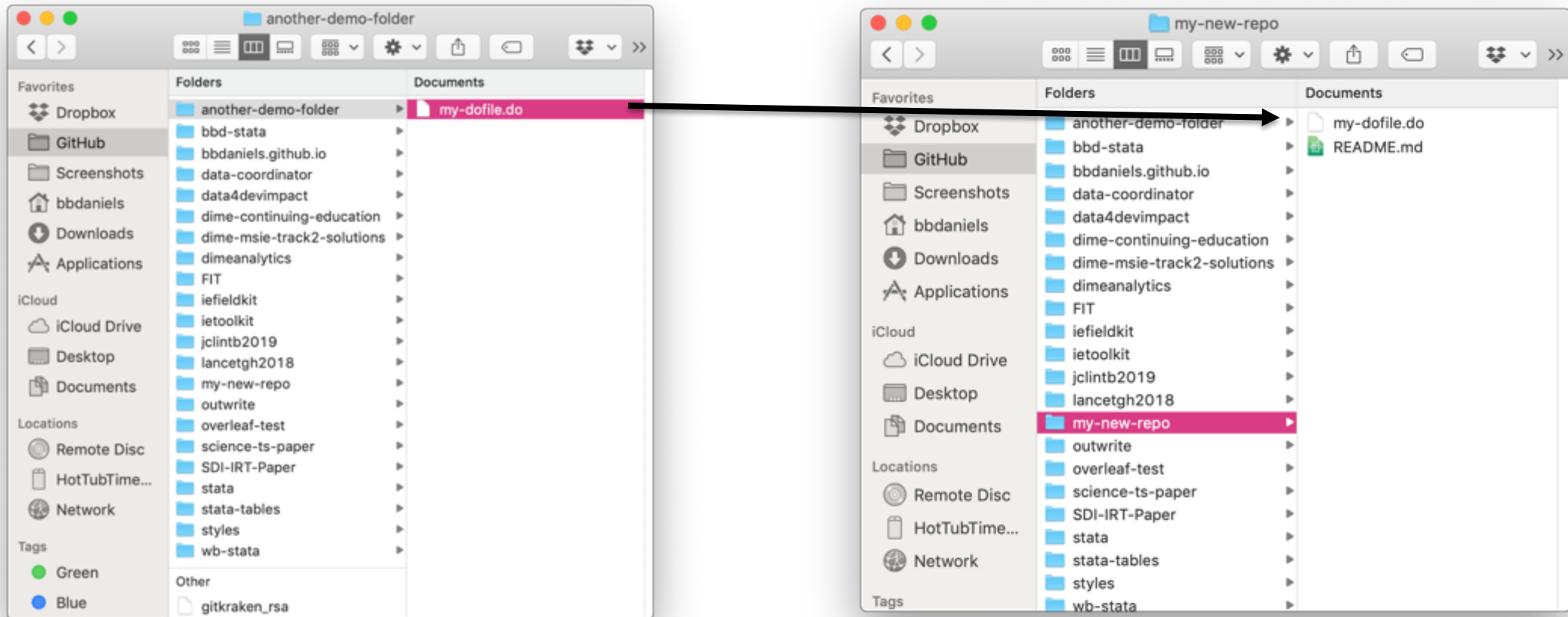
(I know it seems like a lot of steps now, but this becomes second nature. Bear with me, and use this presentation as a guide!)



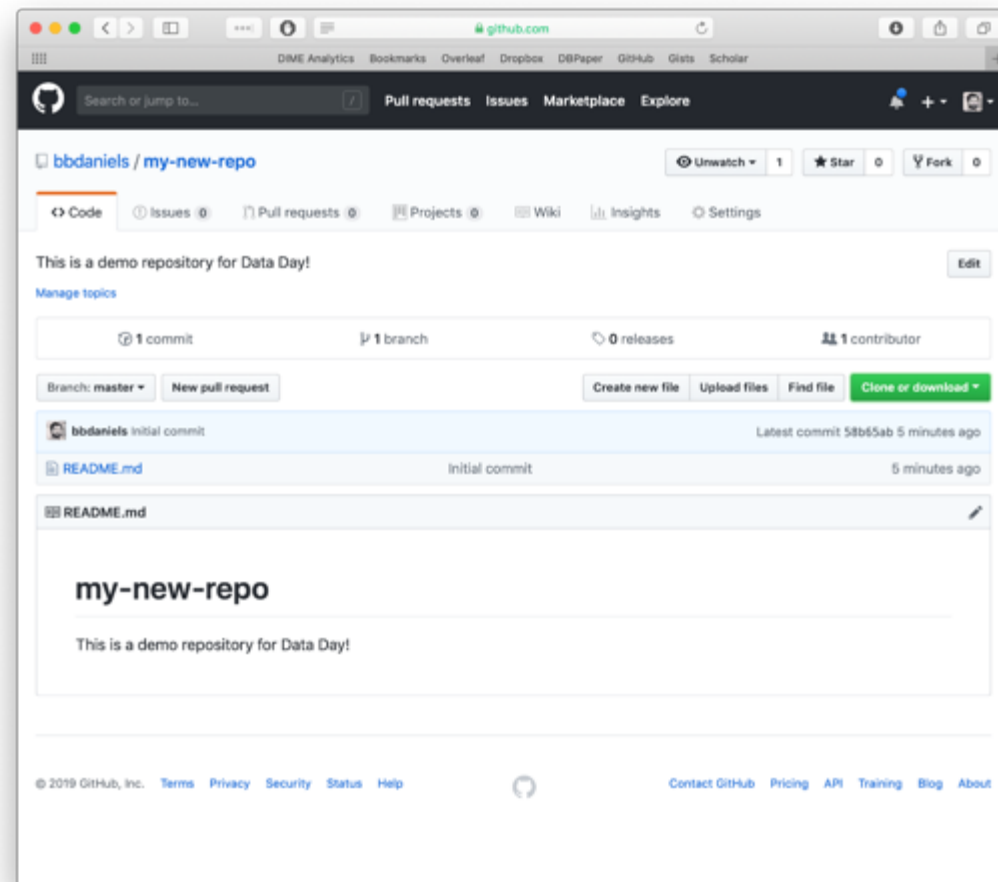
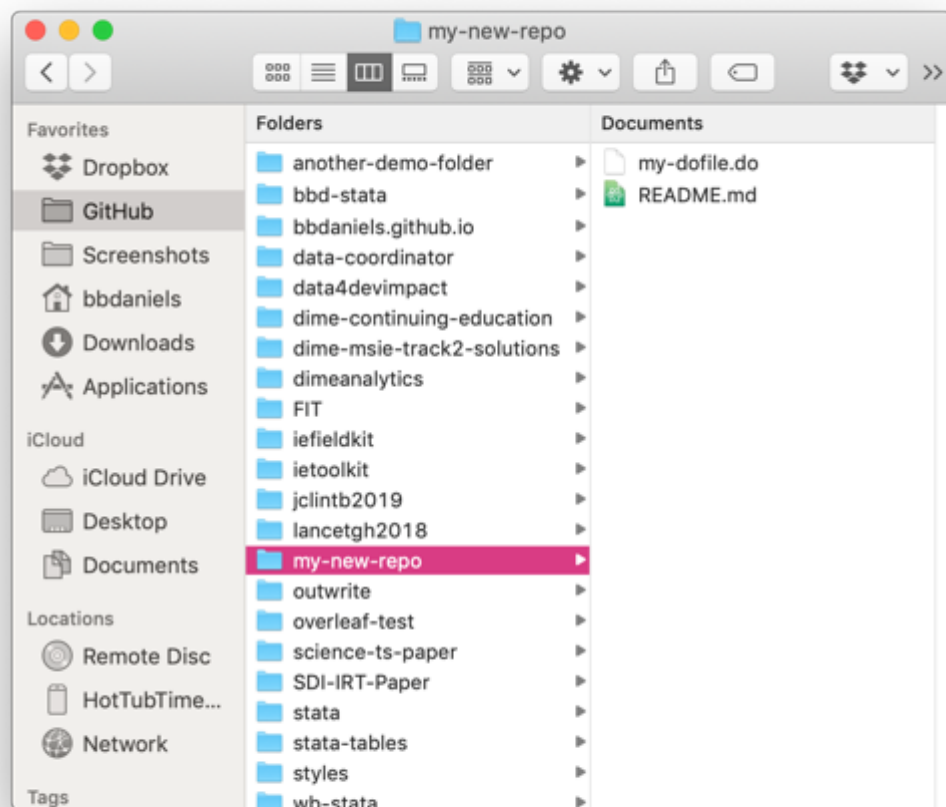
# The “local” and “remote” instances are identical



# Add some kind of file locally

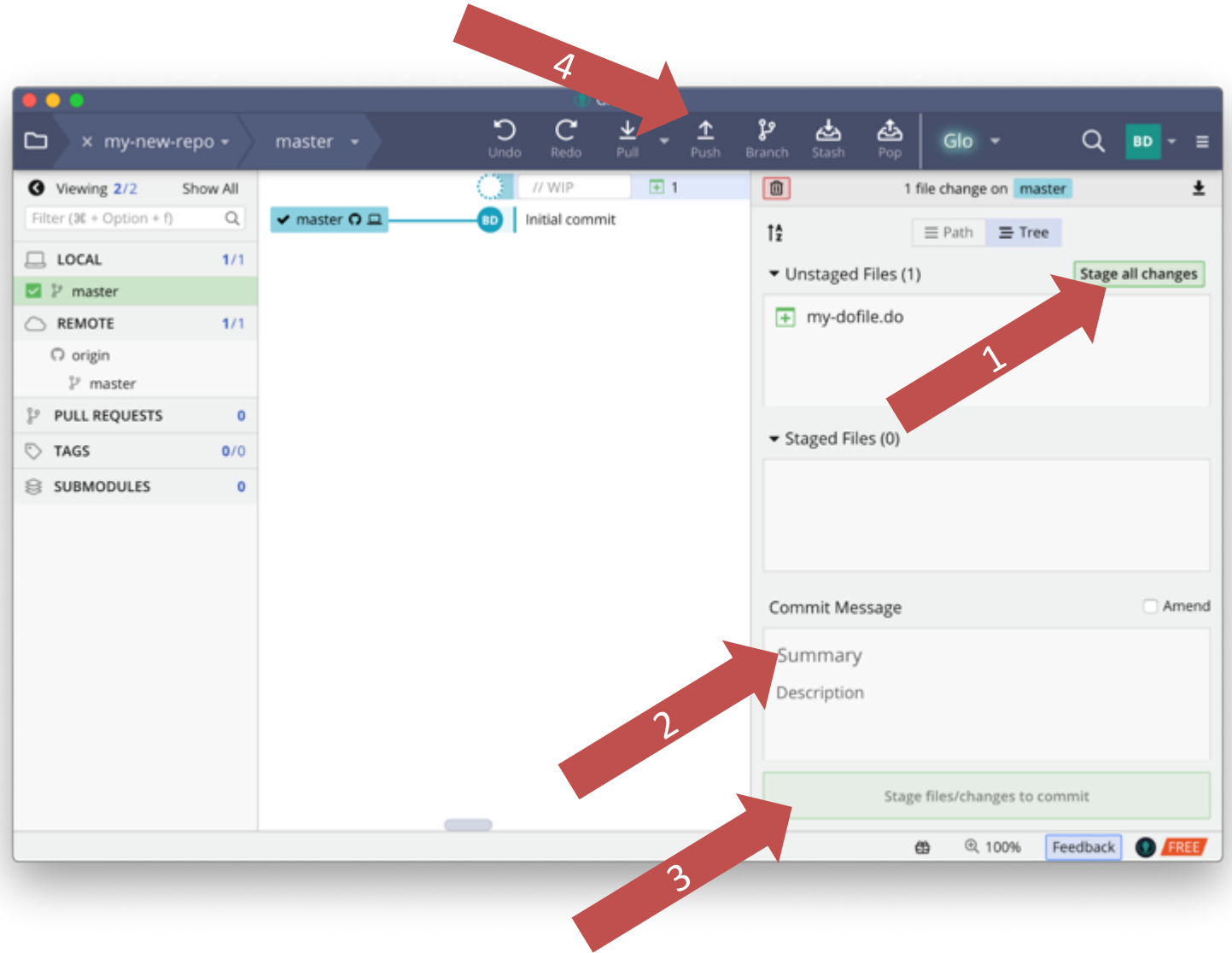


# The local and remote are *not* in sync



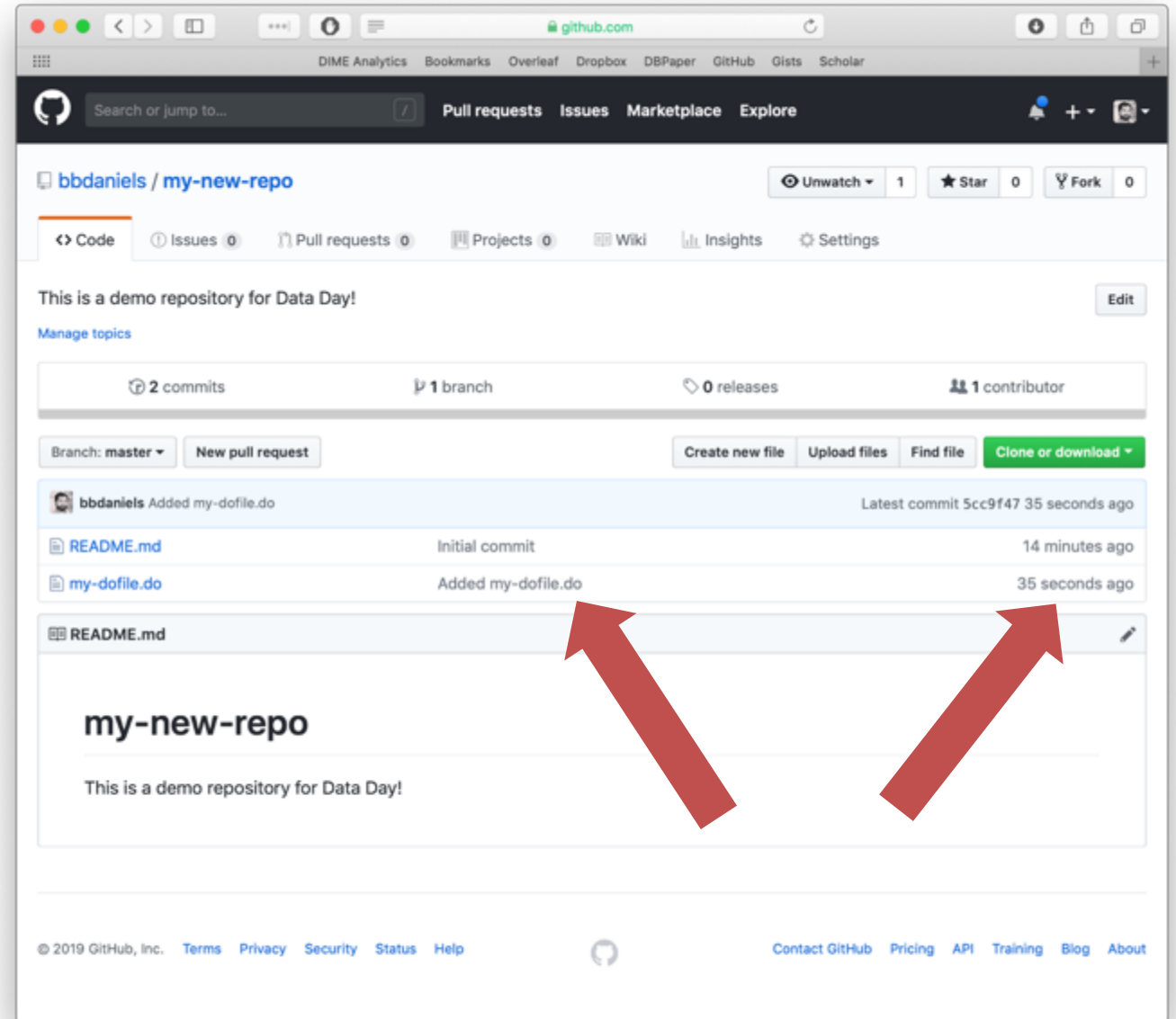
# Git notes changes as “work in progress”

1. “Stage” your changes to add them to the queue to be committed.
2. “Name” your changes informatively – a short sentence will do nicely.
3. “Commit” your changes to add them to the version history. You will see the local copy of the repository move ahead of the GitHub (remote) copy.
4. “Push” your changes to sync the remote (GitHub) repository with your local copy.



# The local and remote are in sync

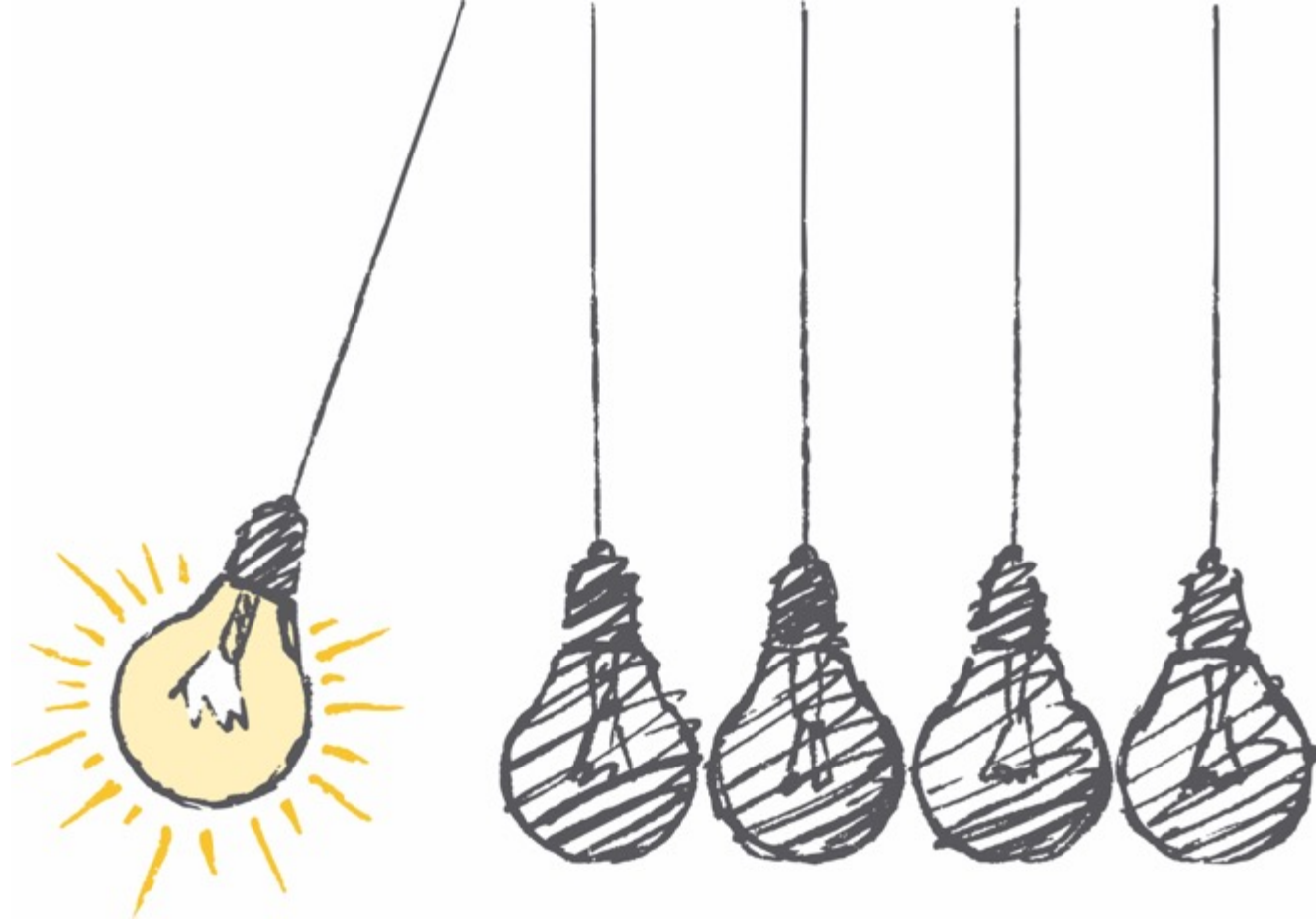
Each item records when it was last modified – using the corresponding name and timestamp from the commit that modified it.





So how does it *work*?

## Part 2: Creating branches in a repository

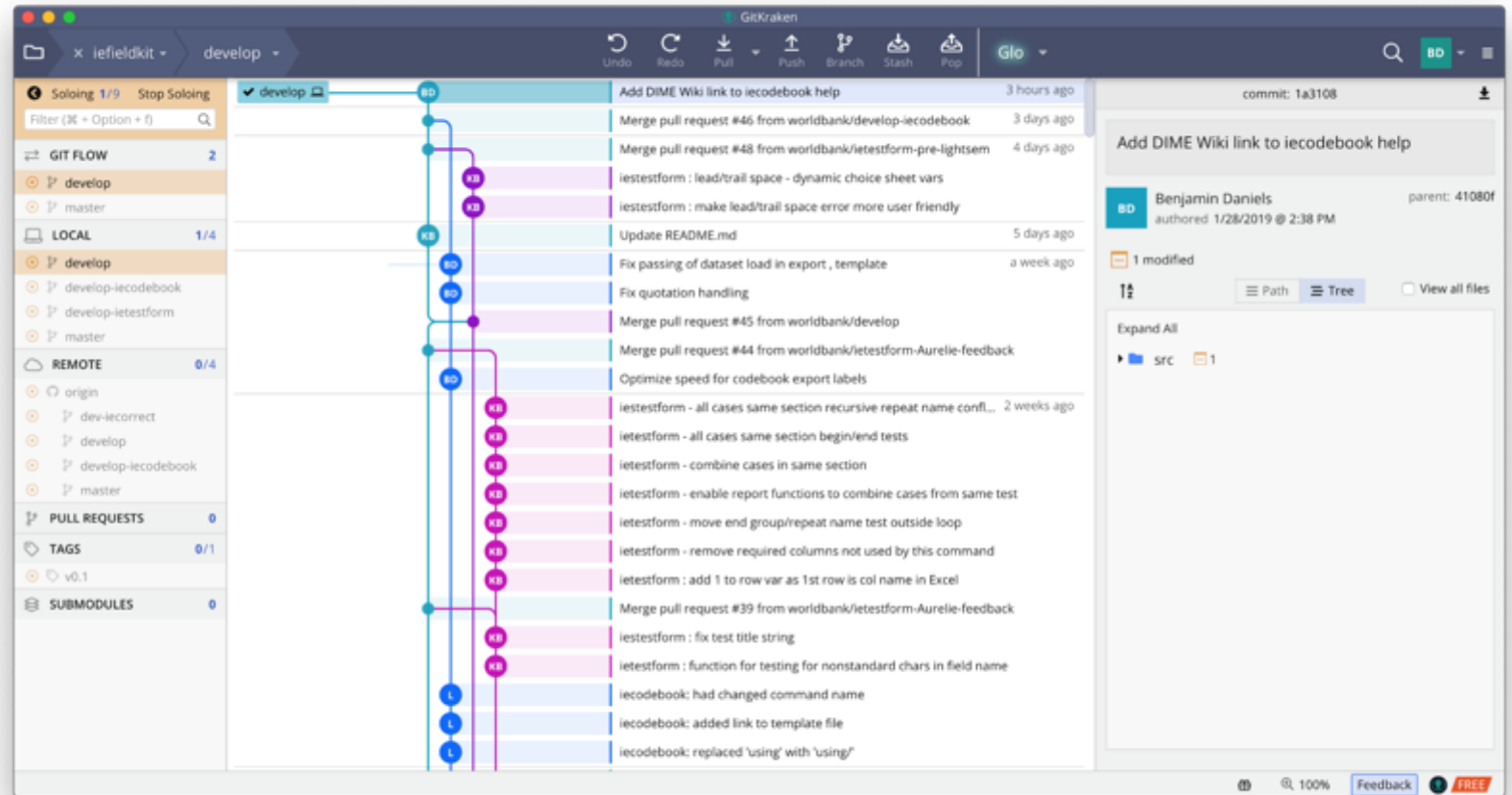


# Git creates branches

Branches are the “killer feature” of Git. Nearly every advanced Git usage you learn will be about how to manage branches.

Branches enable different people to work on the same thing at the same time, and they enable you to view different versions of your files.

Branching allows you to move forwards or backwards in time; and to move “horizontally” in time through various concurrent versions.



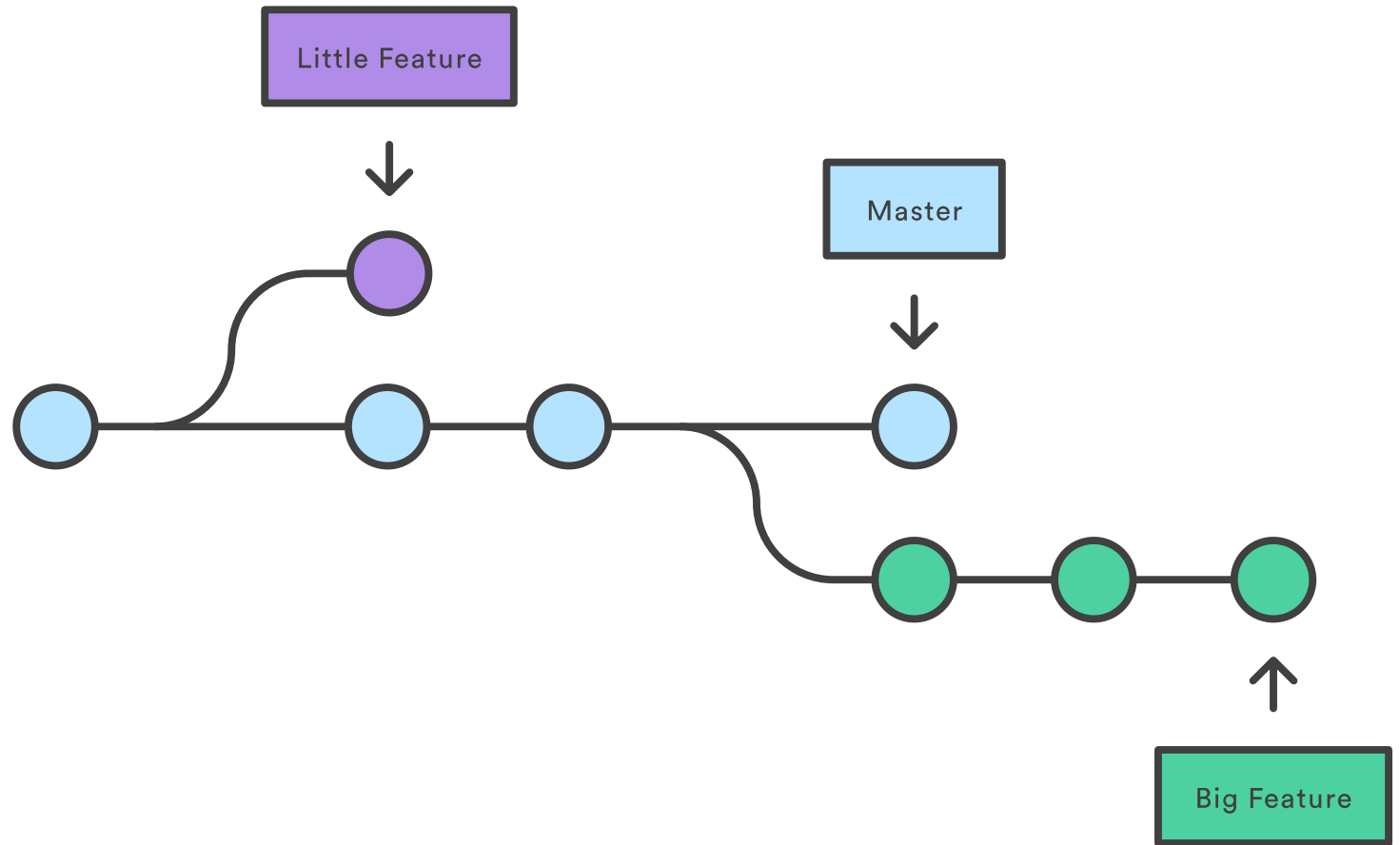
# Branches name

## current versions

Each “branch” points to a commit, usually the latest version in some development workflow. You can switch between branches *locally*. When you do, your working directory will look exactly as it did when that commit was made.

Every contributor can be on any branch they like at any time – past or current.

This means various experimental changes can be made simultaneously without affecting the current edition of a product.

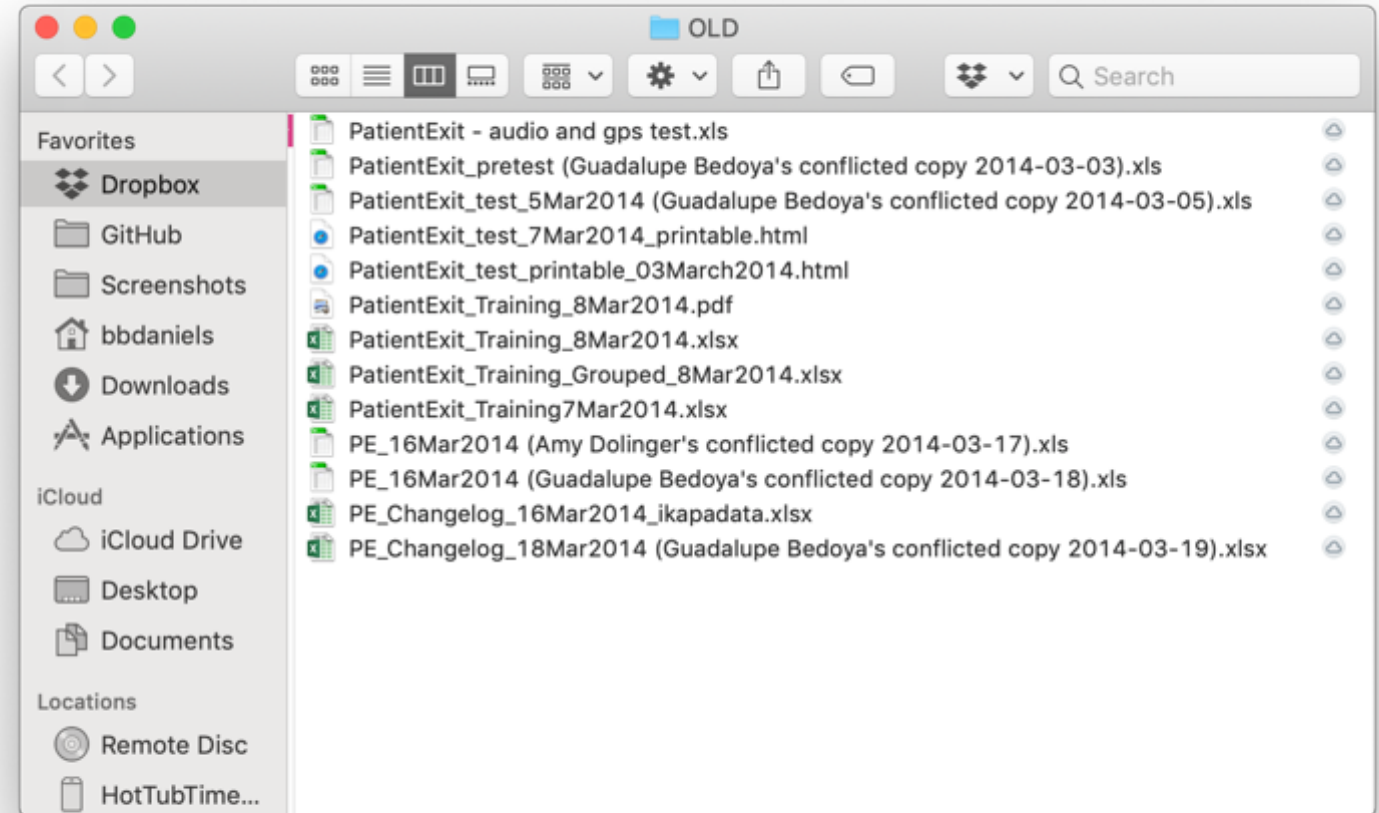


# How is that better than Dropbox?

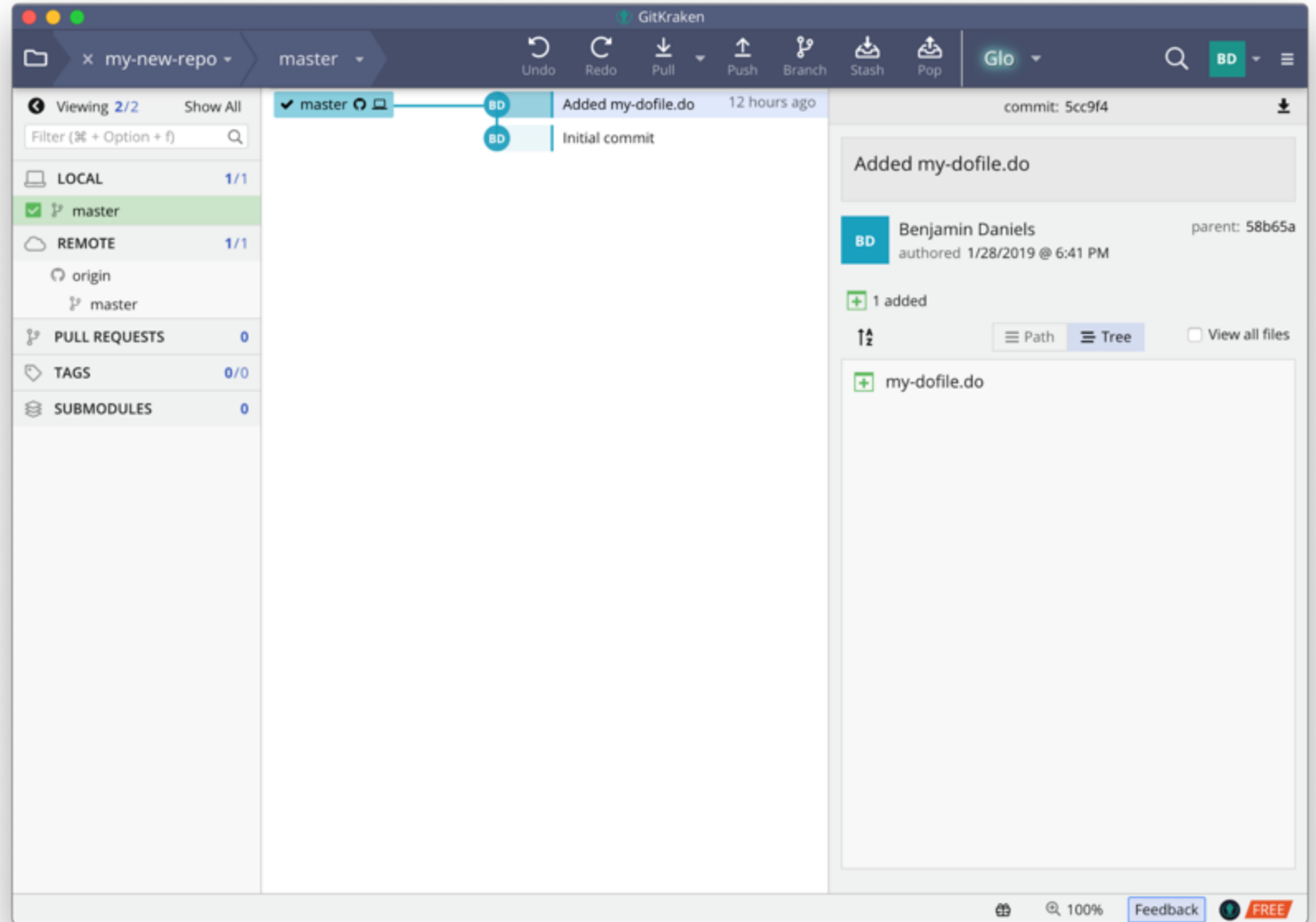
In Dropbox, there can be only *one* living version of a file at any time – otherwise a “conflicted copy” is created. This means nobody can edit the same file at the same time: Dropbox has no concurrency.

Worse, if you “roll back” a file to a previous version to see what it looked like, this affects everyone’s *current* version, even if you don’t want it to.

Finally, Dropbox versions are costly (computationally) – so they delete them often without telling you.



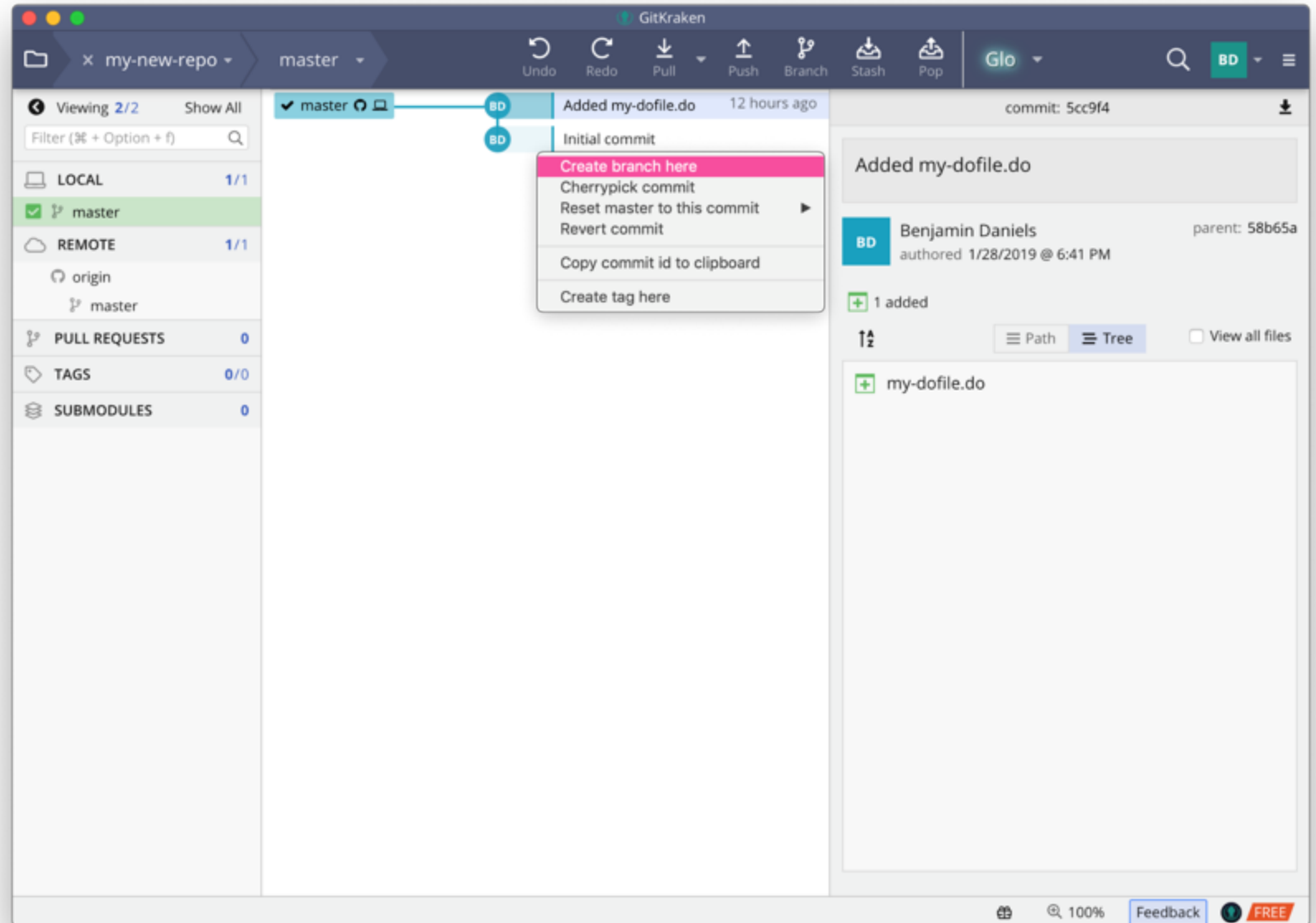
# Open your repository in GitKraken



# Create a branch called *develop*

The name *develop* is a common name for a second branch. In one popular workflow model, *master* always holds a released product, and *develop* is the workstream for the next version.

Right-click on the commit named “Initial commit”, and select “Create branch here”.





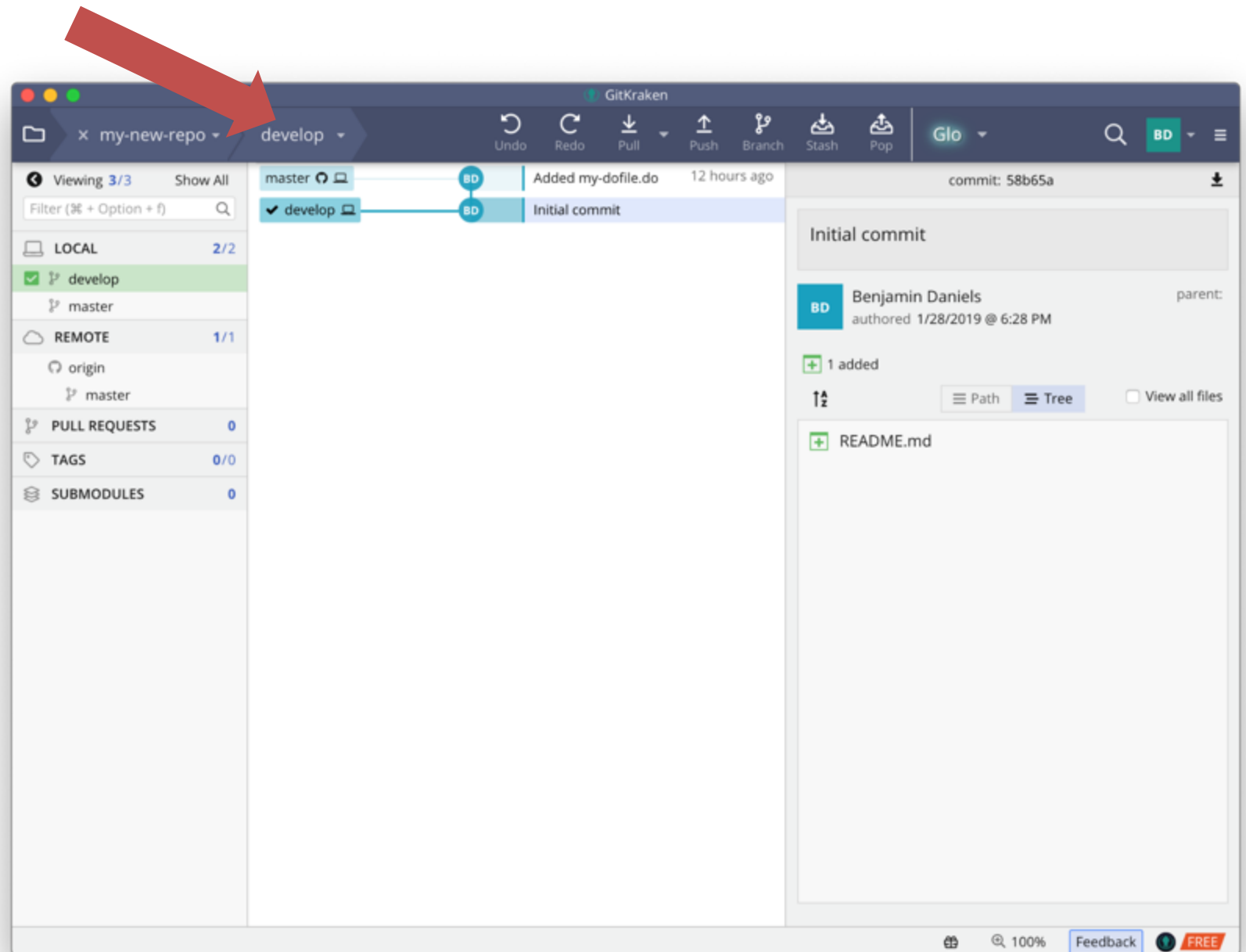
# Open your repository in GitKraken

You should now see that the *develop* branch is “checked out” in this repo.

You should also see that the *develop* branch “points to” the “Initial commit”.

What does that mean?

It means that you have used Git to set your working directory to reflect the state recorded in the “Initial commit”. Go check!

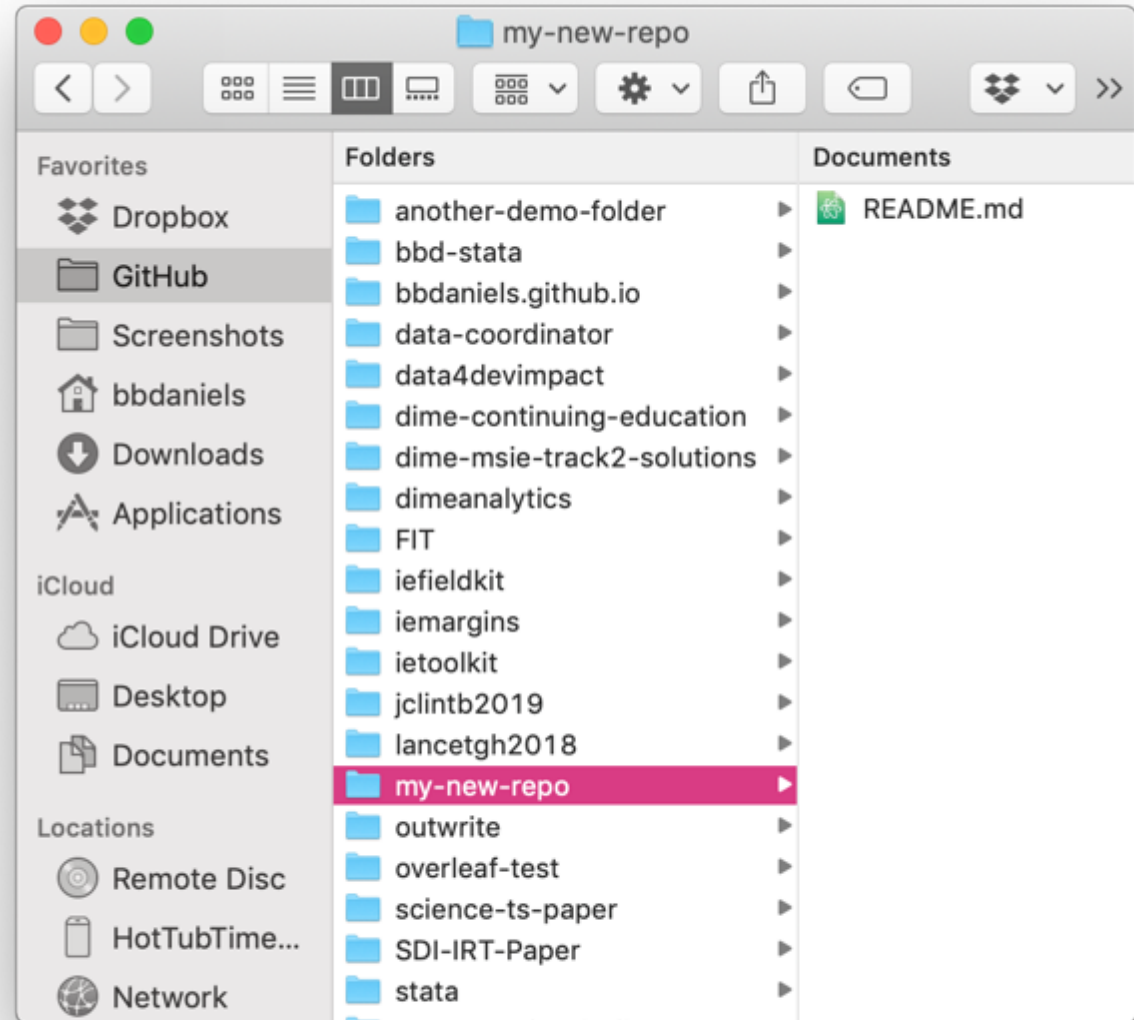


# Where's my file?

If you look in the working directory associated with “my-new-repo”, you will see that it is exactly in the state it was when you initialized it.

Don't ask where your file “is”. You really don't need to know.

What you do need to know is that by “checking out” the *master* branch, it will be in the working directory again. Try it!

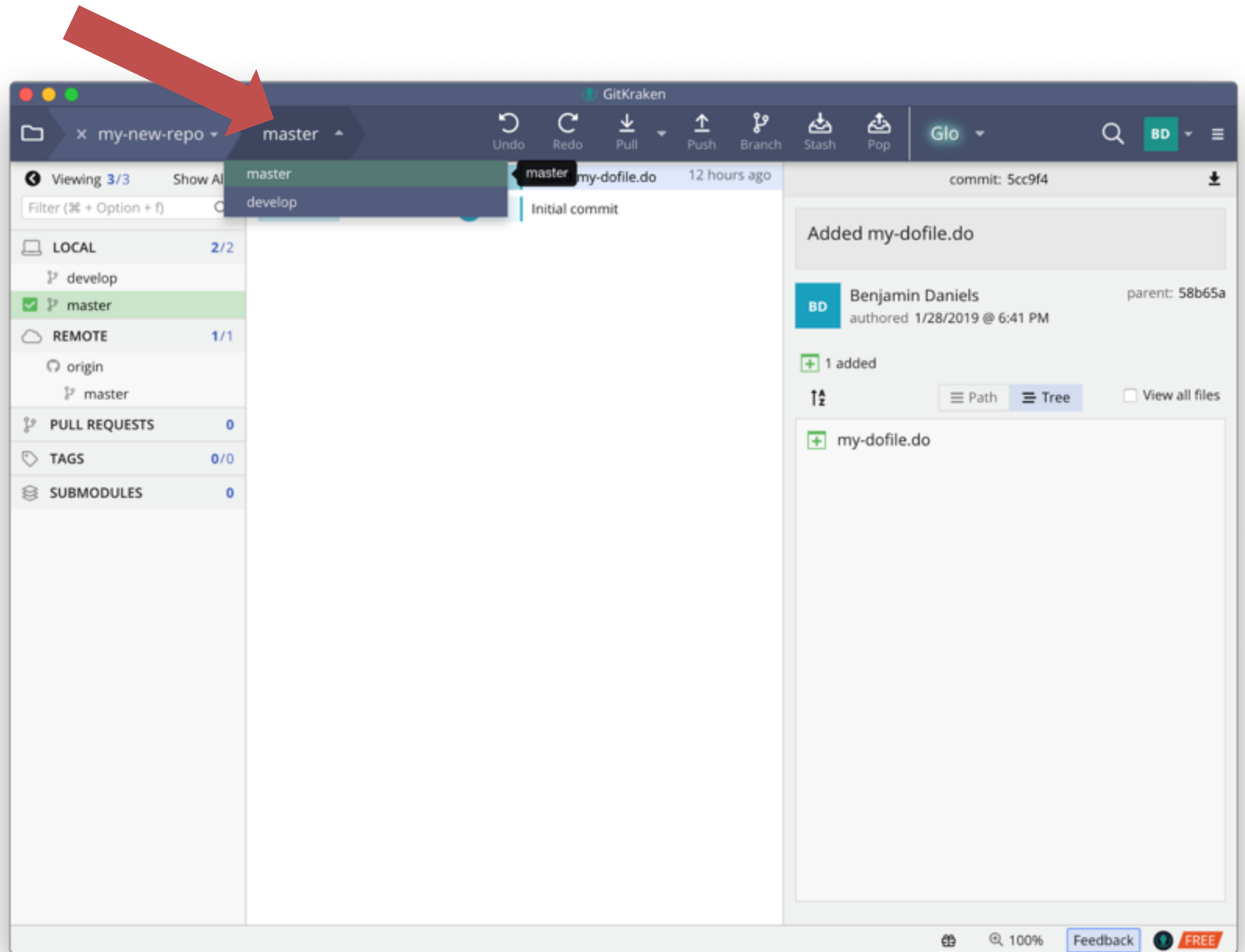


# Check out the *master* branch

This will return your working directory to the state recorded in the commit named “Added my-dofile.do”. This is why naming your commits is important!

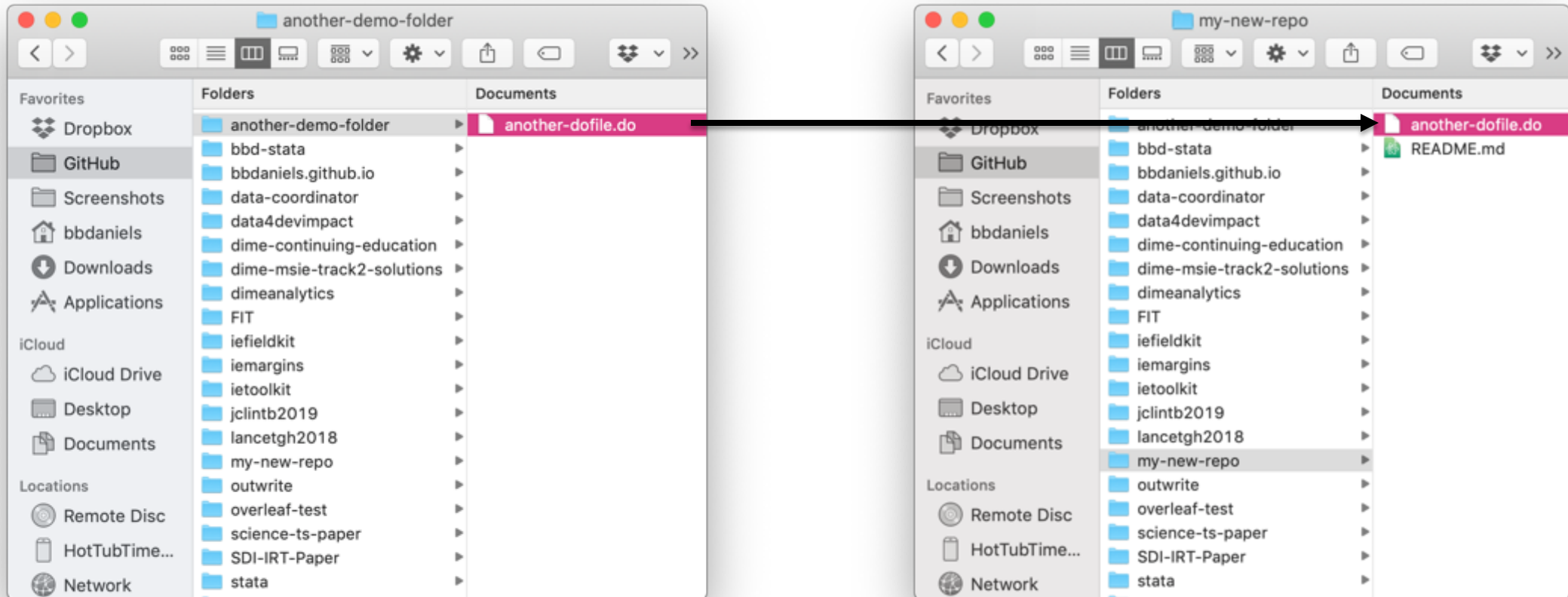
Finally, go ahead and check out the *develop* branch again, so we can do some work there. Go back and forth a couple times if you like, to convince yourself how this works.

Make sure you know which branch you have checked out when you are done! You can only know this information from the client, not from the working directory.



# Add a *different* file to your working directory

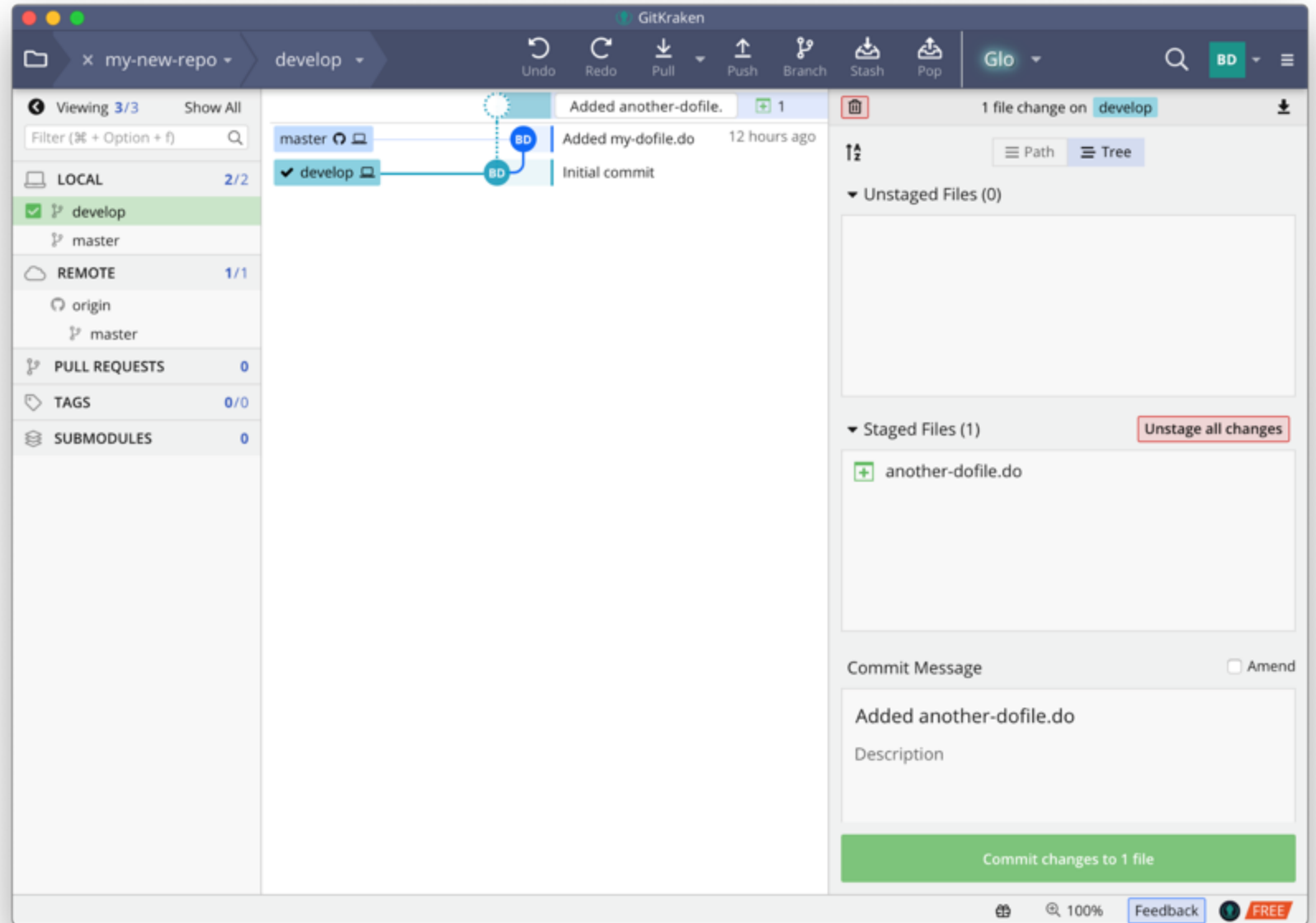
---



# Stage and commit your changes

Navigate back to the client and view the “work in progress” at the top of the tree.

Stage, name, and commit your changes.



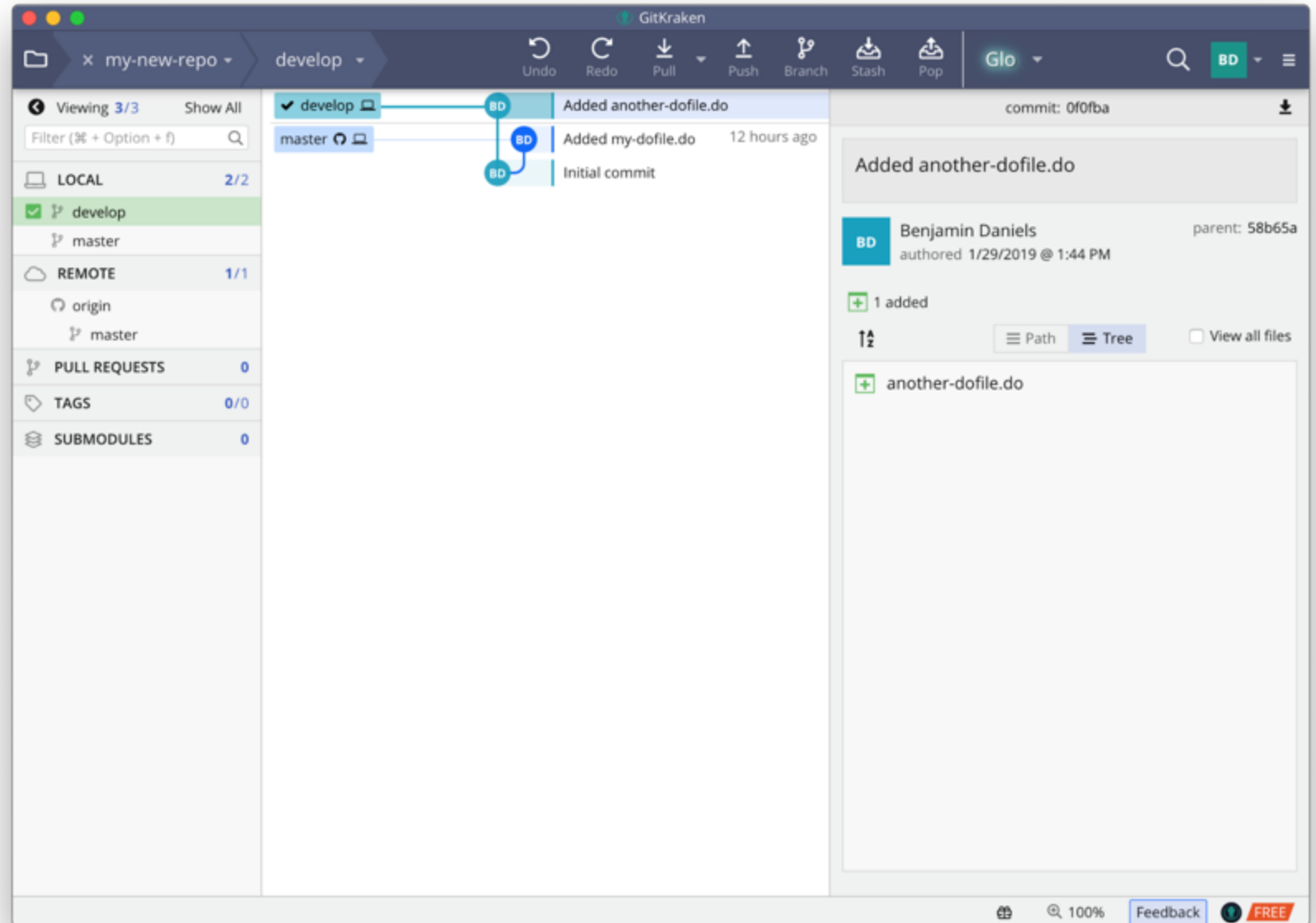
# Stage and commit your changes

Navigate back to the client and view the “work in progress” at the top of the tree.

Stage, name, and commit your changes.

After doing this, you will see that the *develop* branch has “diverged” from the *master* branch.

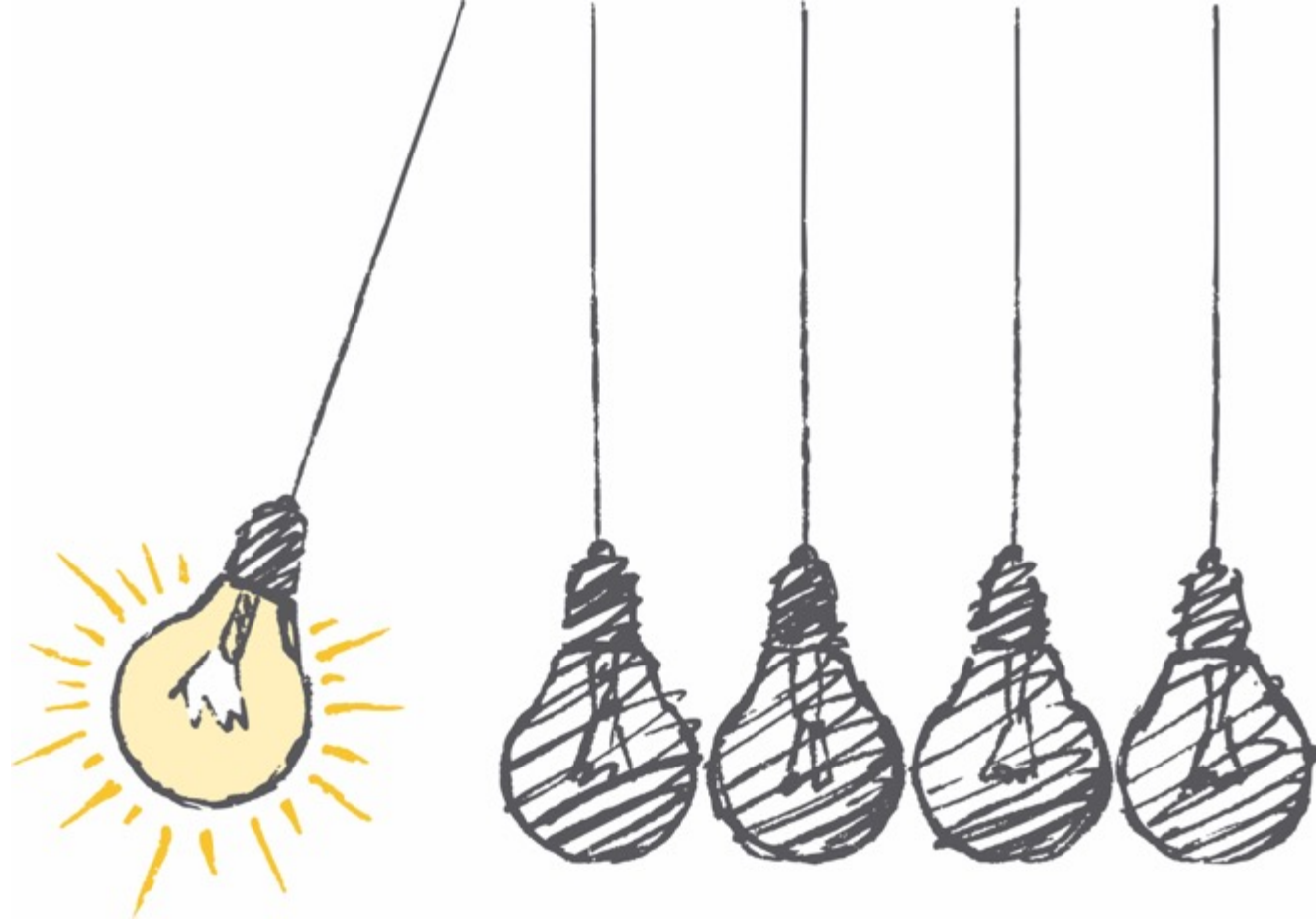
Try switching back and forth between them and see what happens in the working directory.





So how does it *work*?

## Part 3: Merging branches using GitHub

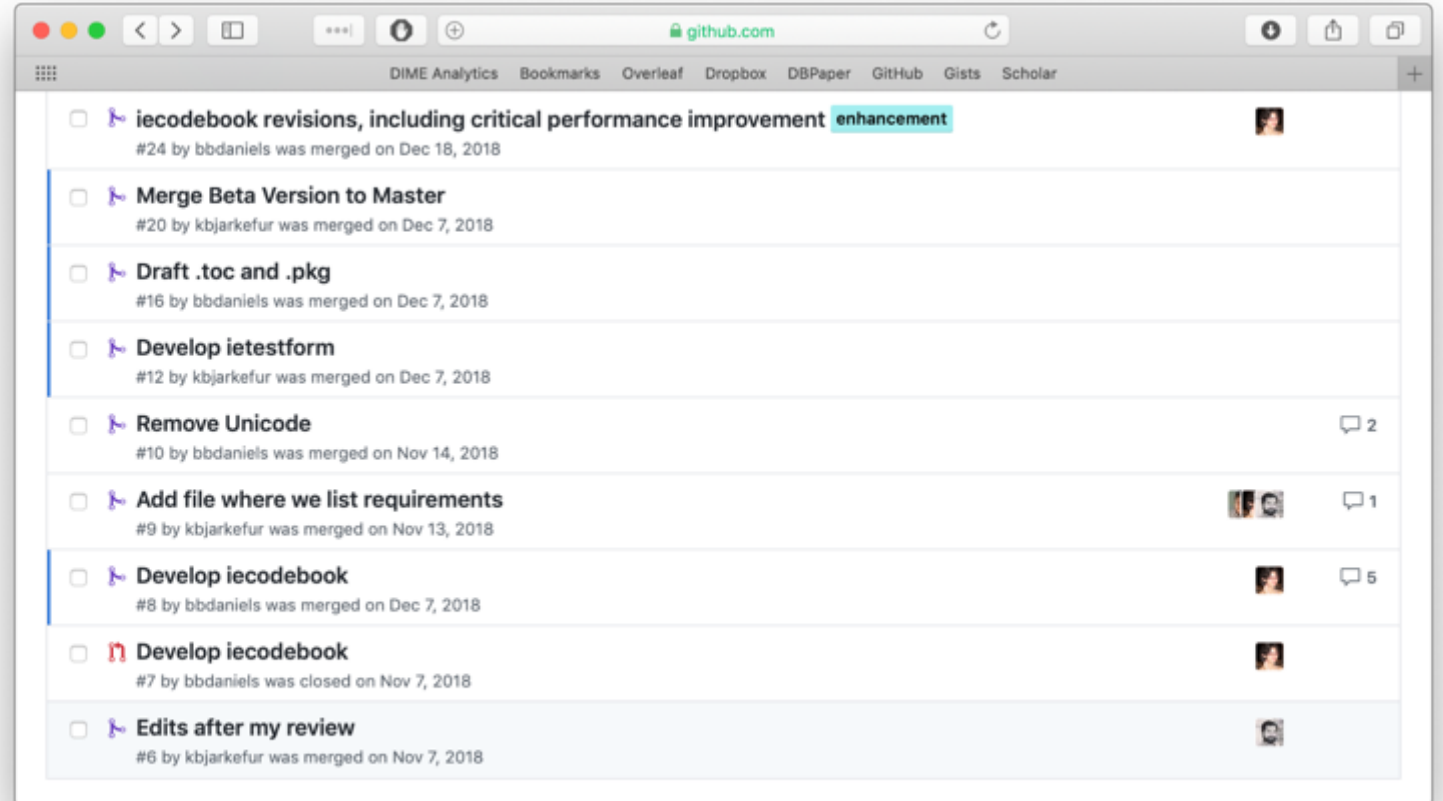


# GitHub manages contributions

GitHub provides interfaces for assigning tasks, submitting updates, and approving and accepting contributions into a project.

Like Git, everything you ever do is saved in an orderly way, so you can always look back and see when and why you did (or undid) something.

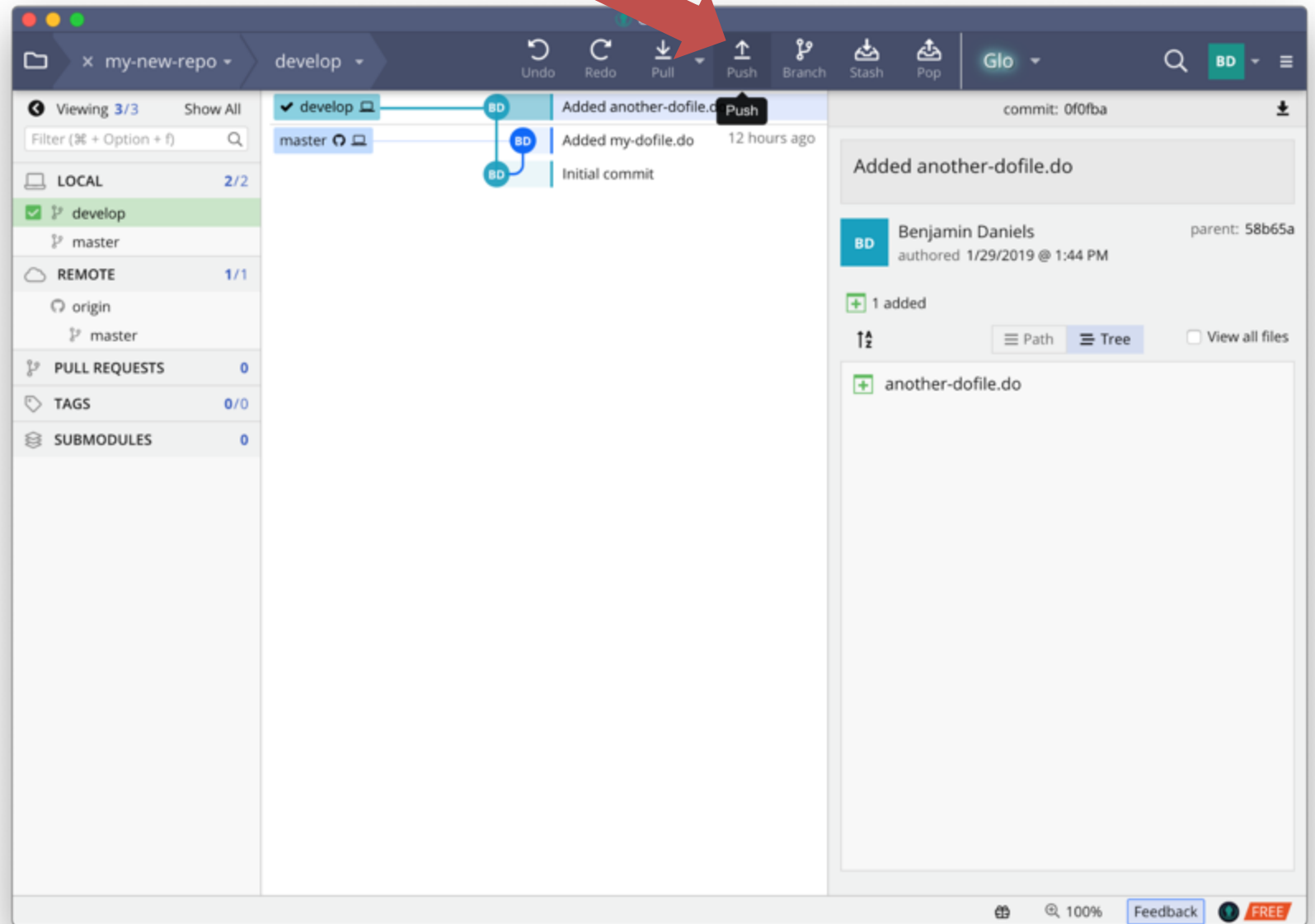
It is designed to make project record-keeping and task-management as part of your workflow.



# Push the *develop* branch to GitHub

Navigate back to the client and make sure the *develop* branch is checked out.

“Push” this branch to GitHub (the “origin”). You will get a notification that the branch does not yet exist there, and ask you to name it. The default is *develop*; stick with that.

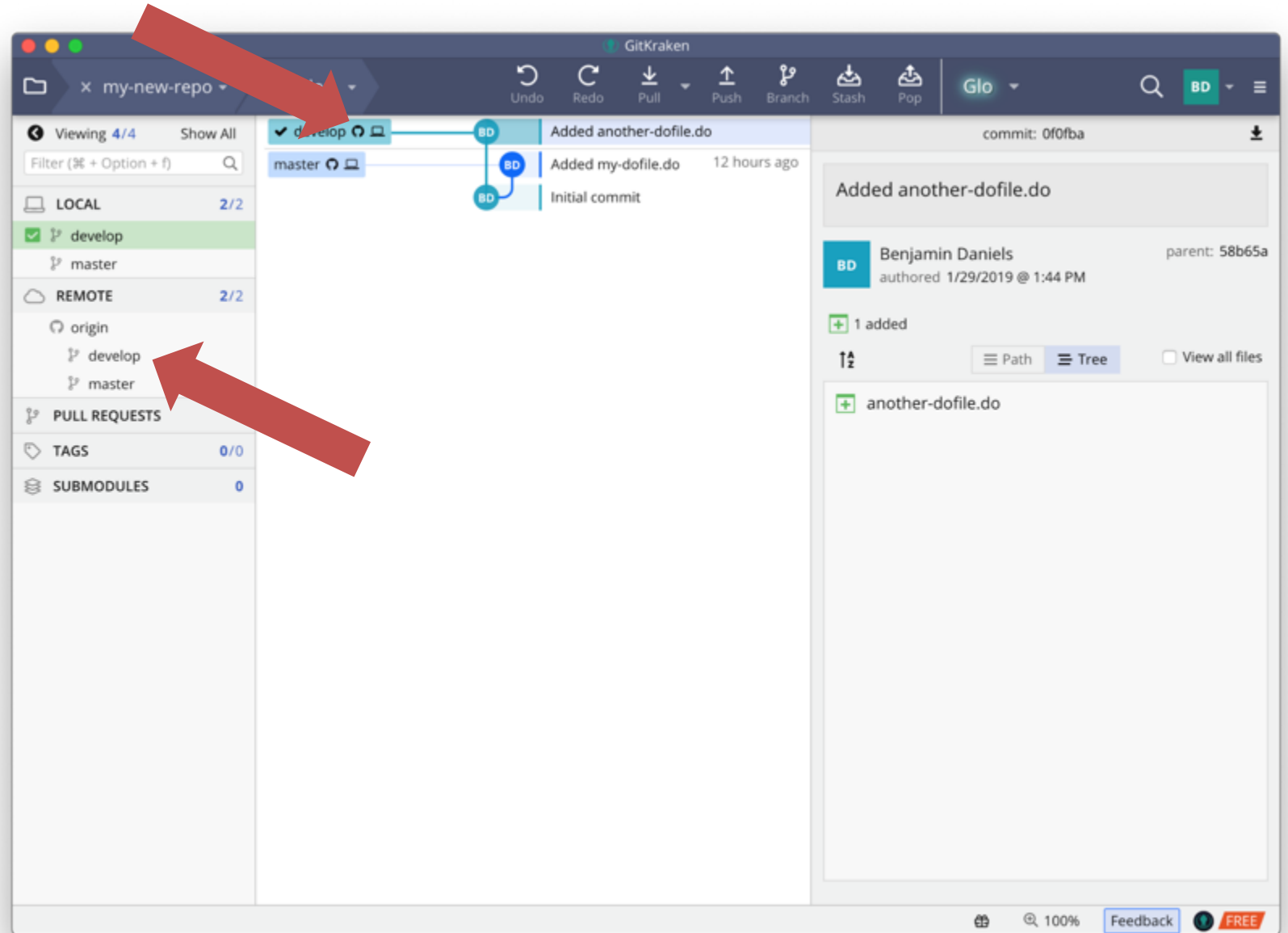


# Push the *develop* branch to GitHub

Navigate back to the client and make sure the *develop* branch is checked out.

“Push” this branch to GitHub (the “origin”). You will get a notification that the branch does not yet exist there, and ask you to name it. The default is *develop*; stick with that.

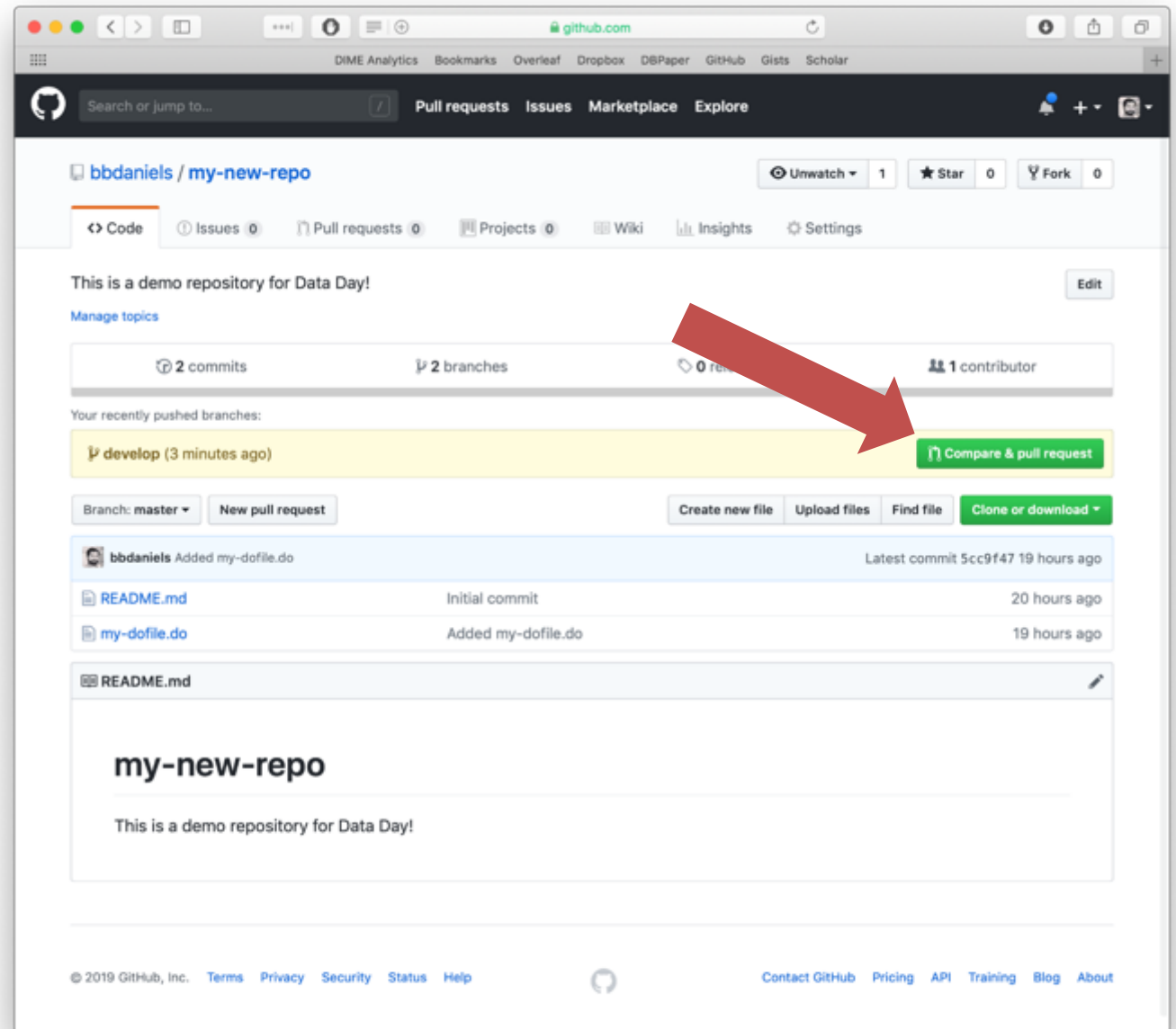
You’ll know it’s worked when the client reflects the existence and location of the *develop* branch on *origin*.



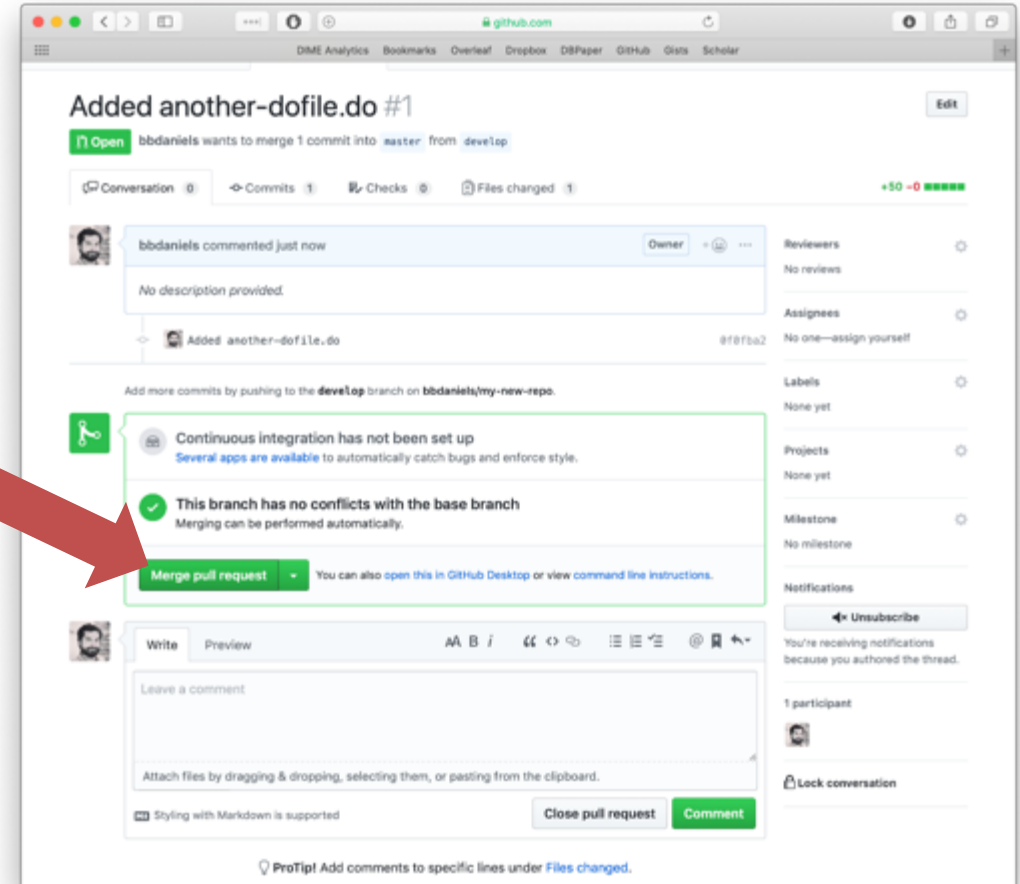
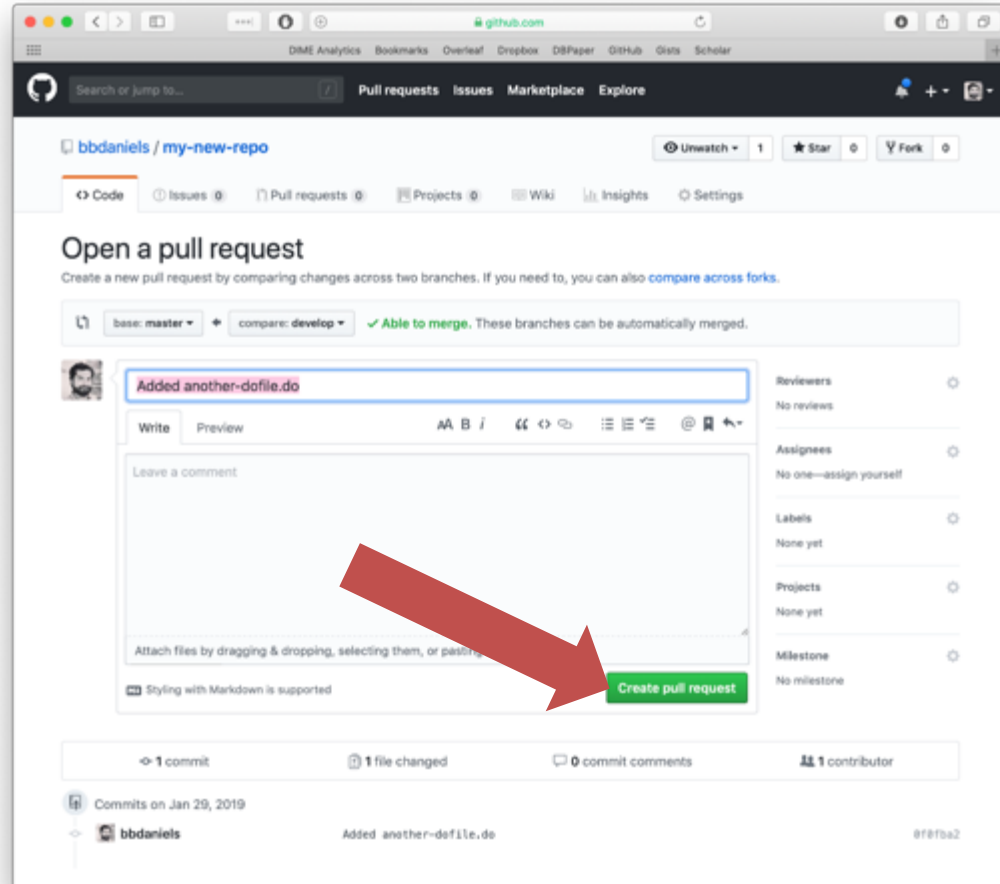
# Go to GitHub and open a pull request

When there are recent changes, GitHub will notify you and ask you if you want to merge the changes.

This is not the most common way to start a pull request, but it will do for now.



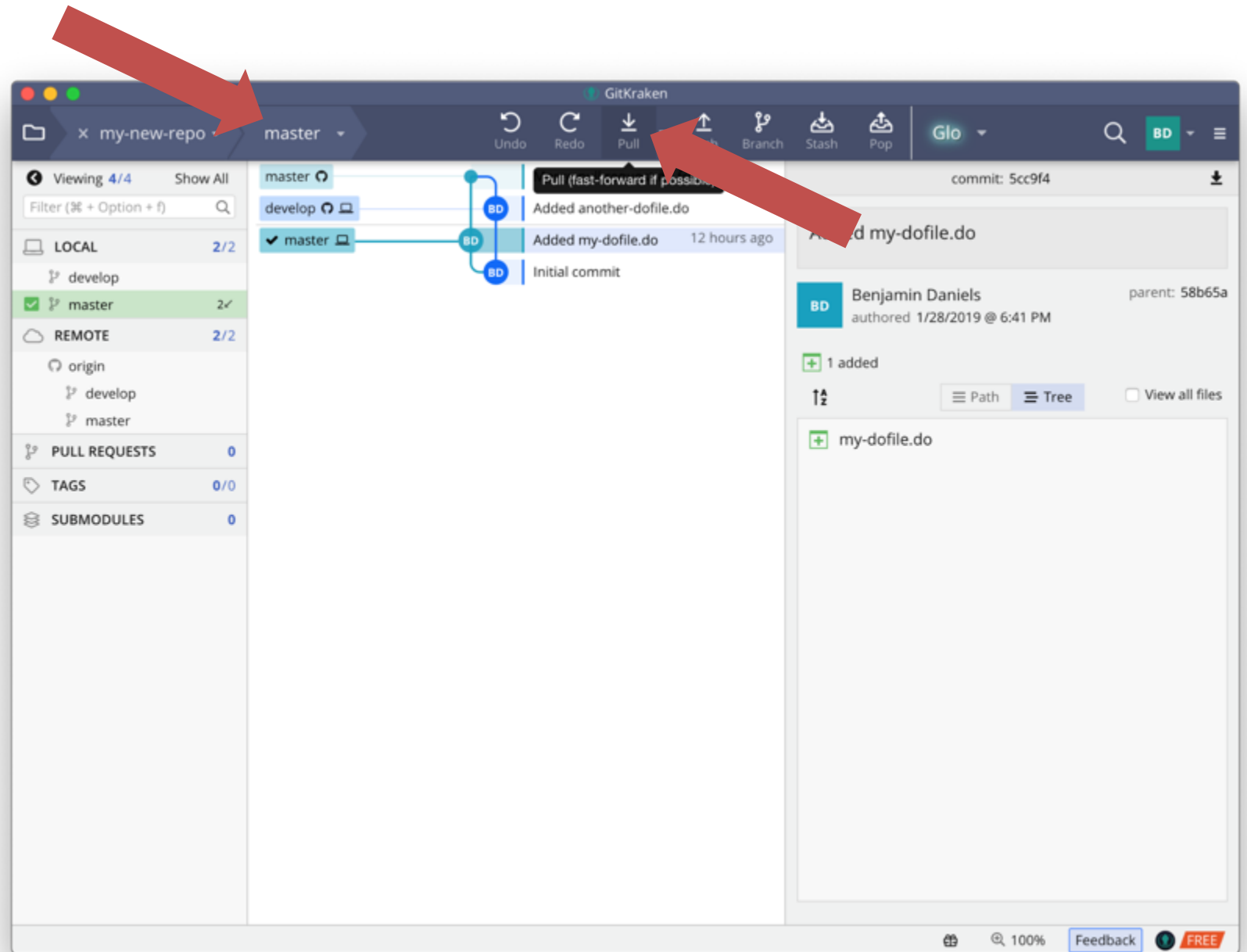
# Create and merge the pull request



# Check out and pull the *master* branch

In your client, you will at first see that the “origin” copy of the *master* branch is ahead of the local copy.

That’s because you merged it on GitHub, and all operations in Git and GitHub are manual.





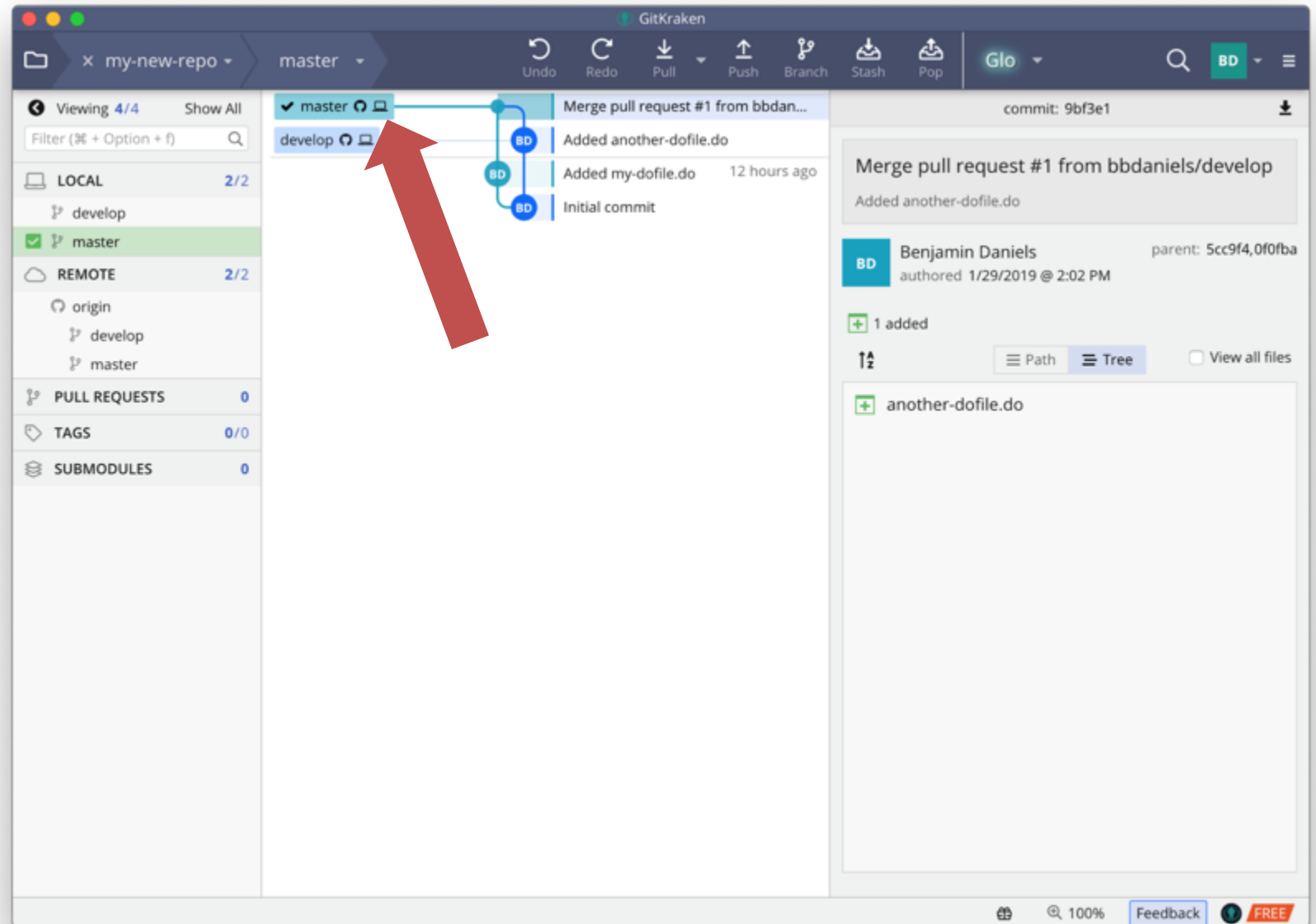
# Check out and pull the *master* branch

In your client, you will at first see that the “origin” copy of the *master* branch is ahead of the local copy.

That’s because you merged it on GitHub, and all operations in Git and GitHub are manual.

When you “pull” the branch, you update your local copy to reflect the origin copy, and you will see the local pointer for the branch move forward.

What did this do?

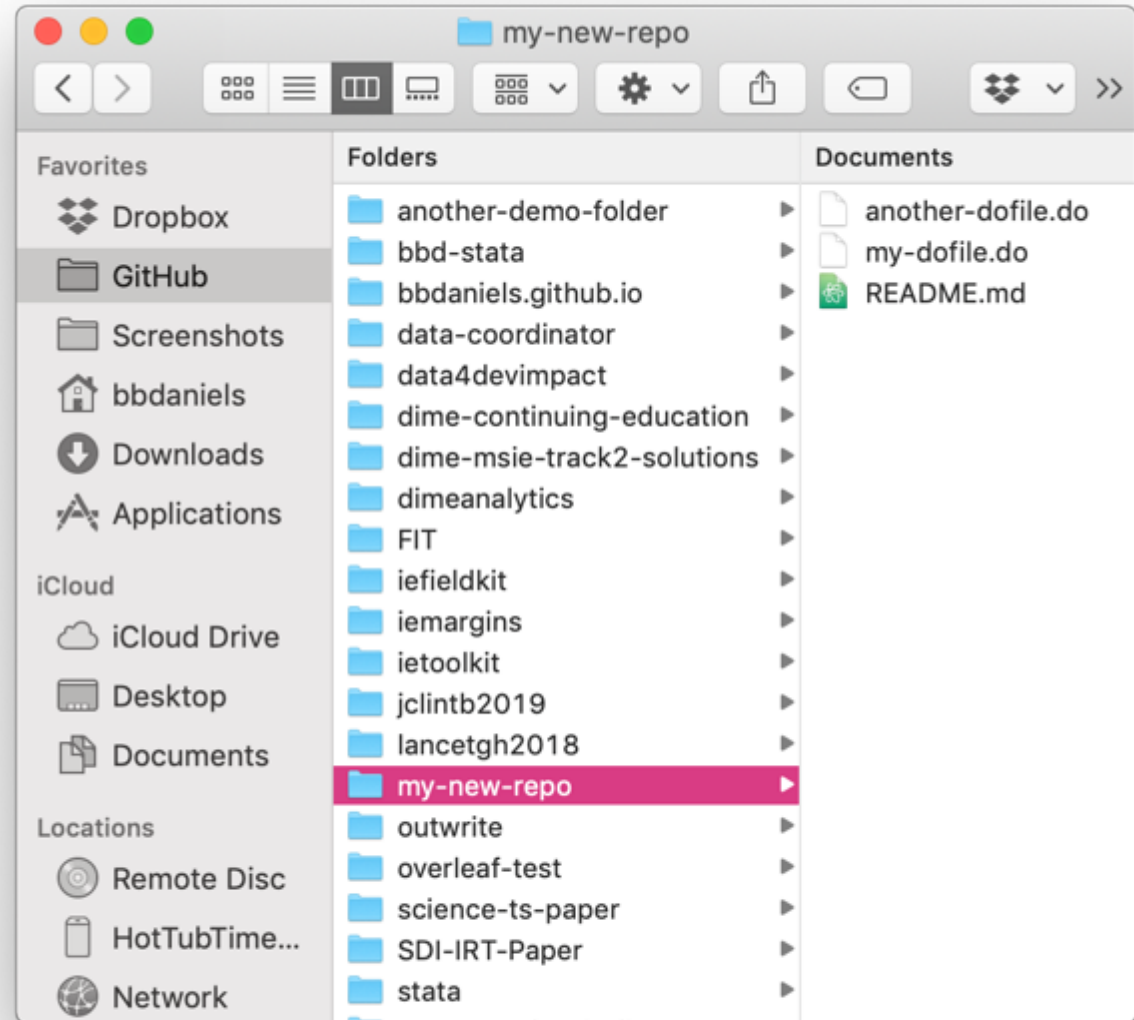


# Merging consolidates changes

In the local working directory, you will now see that *both* files now exist in the same commit, “Merge pull request #1 from bbdaniels/develop”, on the *master* branch.

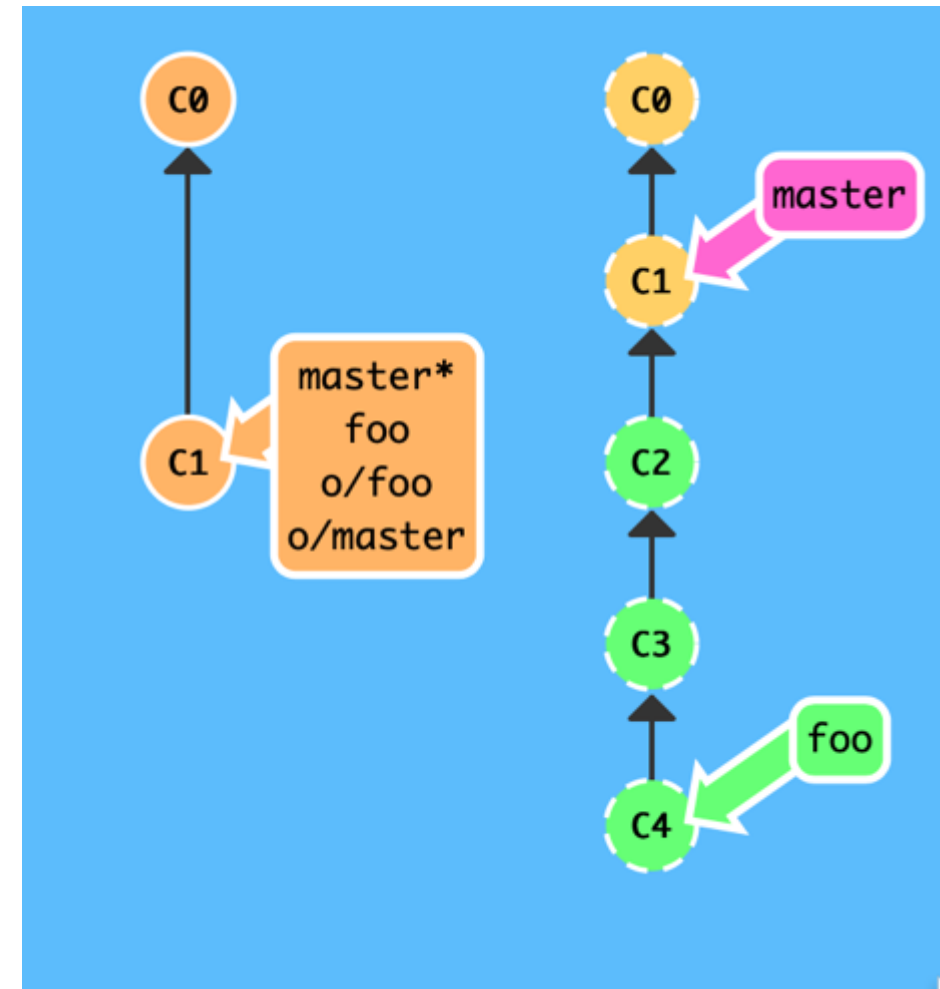
If people made *conflicting* changes to the same files, this process also gives you a chance to resolve those, and this and other workflows will be covered in a later session.

That’s all for now!



# Homework: play with toys

- Practice Git branching at: <https://learngitbranching.js.org>
- Type *[levels]*
- Complete Level 1-4 of “Introduction Sequence” on the Main tab
- Complete all levels on the Remote tab



**Thank you!**

