Data Analysis Projects
MoBi 4. FS - SoSe 2023

19.04.2023

Carl Herrmann

Institut für Pharmazie und Molekulare Biotechnologie | Abt. Bioinformatik

- Allgemeine Vorstellung des Moduls (C. Herrmann ~15-20 Minuten)
- Vorstellung der 5 Themen (ca. 10 Minuten/Thema)
- Kurze Einführung in GitHub

Konzept

- Fortsetzung der **Vorlesung Datenanalyse** aus dem 3. FS
- **Projekt-orientiertes Lernen**: Erfahrung in der konkreten programmatischen Umsetzung der Methoden der Datenanalyse
- **Ziele**
  - Erfahrung in der **Anwendung der Methoden der Datenanalyse** anhand von reellen Datensätzen und wissenschaftlichen Fragen!
  - Erfahrung in der **Teamarbeit**
  - Erlernen des **Umgangs mit modernen Werkzeugen** der Datenanalyse (R / Python / Markdown / GitHub / ...)

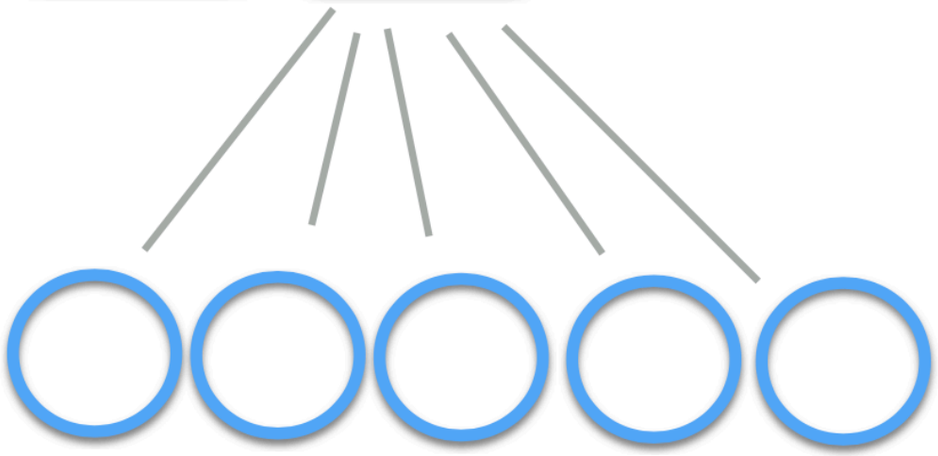
Learning by doing

Themen / Projekte

- **5 Forschungsthemen**
- Für jedes Thema gibt es bis zu 5 sub-Projekte
- Jedes Projekt wird durch ein Team von 4 Studierenden bearbeitet
- **Pro Thema: wissenschaftliche Betreuer und Tutor/Tutorin**
- **Aufgabe der TutorInnen:**
  - wöchentliche Treffen mit den Gruppen die an den Projekten eines Themas arbeiten
    (z.B. Mittwochs 10-13h)
- TutorInnen:
  Hannah Winter; Benedikt Wolf; Luise Nottmeyer; Ana Luisa Costa

Topic 01 | Topic 02 | Topic 03 | Topic 04

Sup. Tutor | Sup. Tutor | Sup. Tutor | Sup. Tutor



5 groups of 4 students

jupyter

GitHub

R Markdown from RStudio

Themen / Projekte

- **Topic 01**: ***Biomedical Image Analysis***
  (Karl Rohr / Leonid Kostrykin; Tutor: Hannah Winter)
  - Data types: MNIST images / cell nuclei images

- **Topic 02**: ***Deep Mutational Scans***
  (Dominik Niopek; Tutor: Benedikt Wolf)
  - Data types: mutation data

- **Topic 03**: ***Proteome screen***
  (Maiwen Caudron-Herger; Tutor: Fabio Rauscher)
  - Data types: mass spectrometry data

- **Topic 04**: ***Climate impact on dengue infection***
  (Marina Treskova; tutor Luise Nottmeyer)
  - Data types: Climate data / infection data

- **Topic 05**: ***Drug repurposing in cancer treatment***
  (Carl Herrmann; tutor Ana Luisa Costa)
  - Data types: Drug screen data / expression data / genetic data

Zeitplan

Was wir machen...

19/04 Vorstellung der Projekte
Intro zu GitHub

26/04 Plenum lineare Regression

Was ihr macht!

21/04 Auswahl der Projekte und Teams (Google)

17/05 Präsentation des project proposal (10 + 10 min)

17/07 (8pm) ***Schließung der GitHub repos***

19/07 Finale Präsentation und Bericht (15+10 minutes)

Project proposal (17/05)

- In der Präsentation des **project proposals** solltet ihr...
  - einige des angegebenen Referenzen im Kontext des Projektes vorstellen
  - Die allgemeine Fragestellung / Herausforderungen erklären
  - Die Daten beschreiben
  - Die Ziele eures Projektes definieren
  - Ungefährer Zeitplan
    - milestones = important steps in the analysis
    - when these milestones should be achieved

- Mündliche Präsentation vor Betreuer / Tutoren
  - 10 Minuten Präsentation
  - 10 Minuten Diskussion / Fragen
- *Alle Mitglieder des Teams sollten aktiv beitragen!*

Finale Präsentation und Bericht

- **Finale Präsentation** am 19.07
  - 15 Minuten Präsentation
  - 10 Minuten Fragen
  - Vorstellung der wichtigsten Ergebnisse
- **Bericht (auf Englisch!)**
  - pdf Bericht (im GitHub repo als pdf)
  - **10 Seiten max.**
  - Aufbau: Introduction / material and methods / results/ discussion
  - Wichtig: sorgfälltige Auswahl der Plots; Beschriftung der Plots wichtig!
- **Github repo**
  - Repo vor deadline aufräumen !!
  - Bitte ein klares README erstellen

GitHub repo schließt am Montag 17.07 um 20h
Bericht sollte bis dahin fertig sein!

Benotung

- Project proposal = 30%
- Finale Präsentation = 30%
- Bericht = 40%
- **Kriterien**
  - Qualität der Einführung in die wissenschaftliche Frage (insbesondere Vorstellung der Literatur)
  - Klare Definition der Forschungsfragen
  - Qualität der Plots und Beschriftungen
  - Qualität der Besprechung der Ergebnisse

Themen / Projekte

- **Topic 01**: ***Biomedical Image Analysis***
  (Karl Rohr / Leonid Kostrykin; Tutor: Hannah Winter)
  - Data types: MNIST images / cell nuclei images

- **Topic 02**: ***Deep Mutational Scans***
  (Dominik Niopek; Tutor: Benedikt Wolf)
  - Data types: mutation data

- **Topic 03**: ***Proteome screen***
  (Maiwen Caudron-Herger; Tutor: Fabio Rauscher)
  - Data types: mass spectrometry data

- **Topic 04**: ***Climate impact on dengue infection***
  (Marina Treskova; tutor Luise Nottmeyer)
  - Data types: Climate data / infection data

- **Topic 05**: ***Drug repurposing in cancer treatment***
  (Carl Herrmann; tutor Ana Luisa Costa)
  - Data types: fitness data / ...

Getting started...

Auswahl der Projekte / Registrierung

- Vorstellung der Projekte/Themen; Beschreibung der Projekte lesen
- Webseite: https://www.hdsu.org/teaching/data2023.html
- Wenn ihr das Projekt und das Team zusammengestellt habt, bitte hier registrieren: https://docs.google.com/spreadsheets/d/1AX28mauyWvg0_7wTtYKxHwI2MN2PVUcK6uyULB8g8uo/edit?usp=sharing

- ***Auswahl der Projekte***
  - ***startet Freitag 21.04 um 13 h***
  - ***Sollte bis Montag 24.04 abgeschlossen sein.***

Wie sollte man den/die TutorIn (nicht) benutzen?

Institute for Pharmacy and Molecular Biotechnology

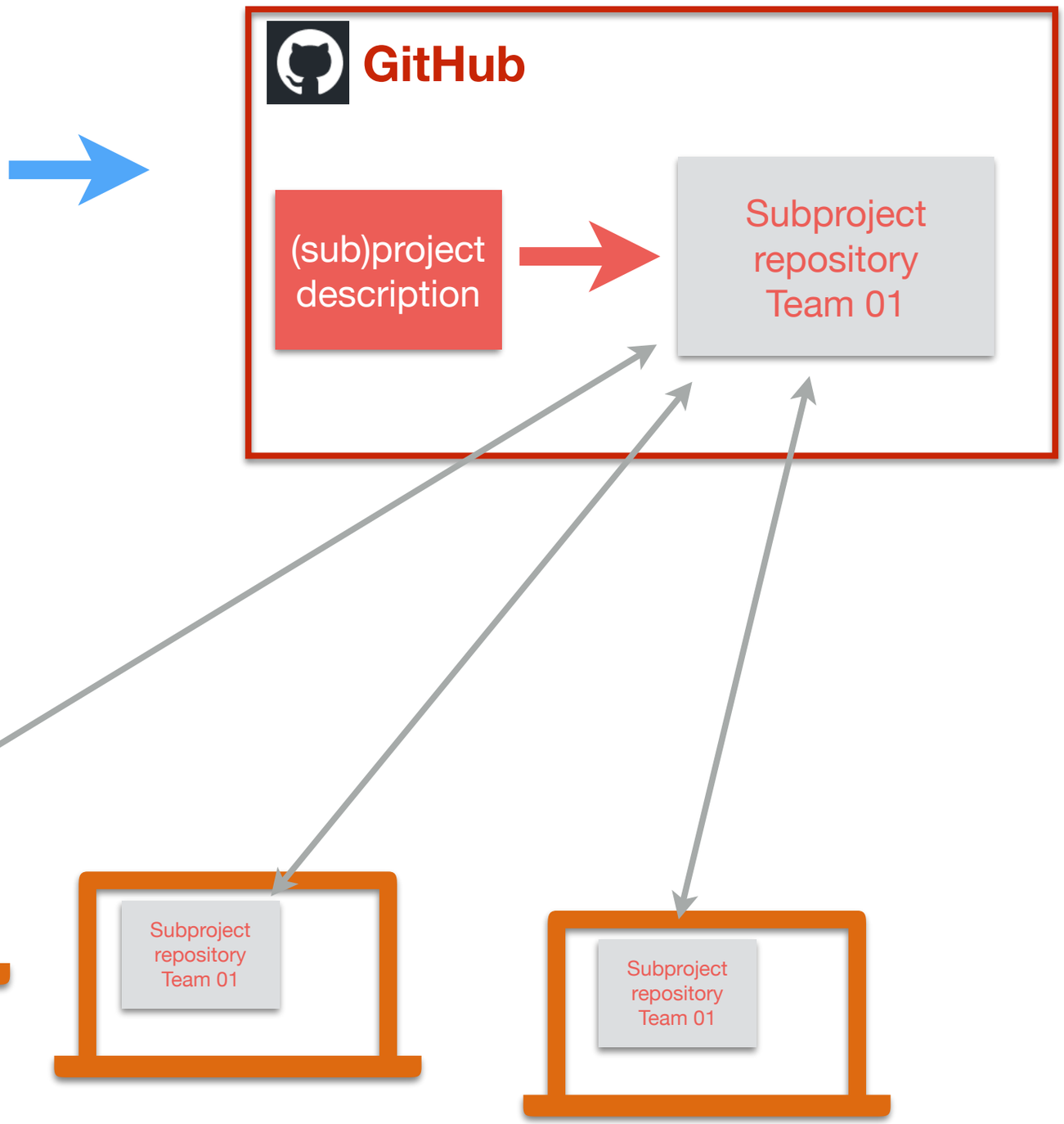
UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

- Zeit für das wöchentliche Treffen ist **beschränkt (30 Minuten/Woche)**;
  → nutzt die Zeit sinnvoll!!
- TutorInnen sind NICHT dazu da, euer Code zu debuggen!
- Bereitet das wöchentliche Treffen sorgfälltig vor:
  - was haben wir seit letzter Woche erreicht?
  - welche Probleme/Fragen haben wir?
  - Ziele für die kommende Woche

TutorInnen sind nicht dazu da, WhatsApp Nachrichten um Mitternacht zu beantworten...

Arbeit organisieren

Website
www.hdsu.org/teaching/data2023.html



jupyter

R Markdown
from RStudio

Für R-Projekte

- Es gibt viele tolle Software Libraries in R (Bioconductor / CRAN / ...)
- **Bitte keine Methoden/Tools benutzen, die ihr nicht absolut versteht!**
- Uns sind einfach Lösungen "per Hand" lieber als "black-boxes"!
- Wenn ihr externe libraries benutzt solltet ihr auch imstande sein, diese genau zu erklären!!

IPMB
Institute for Pharmacy and Molecular Biotechnology

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

Brief intro to Git(Hub)

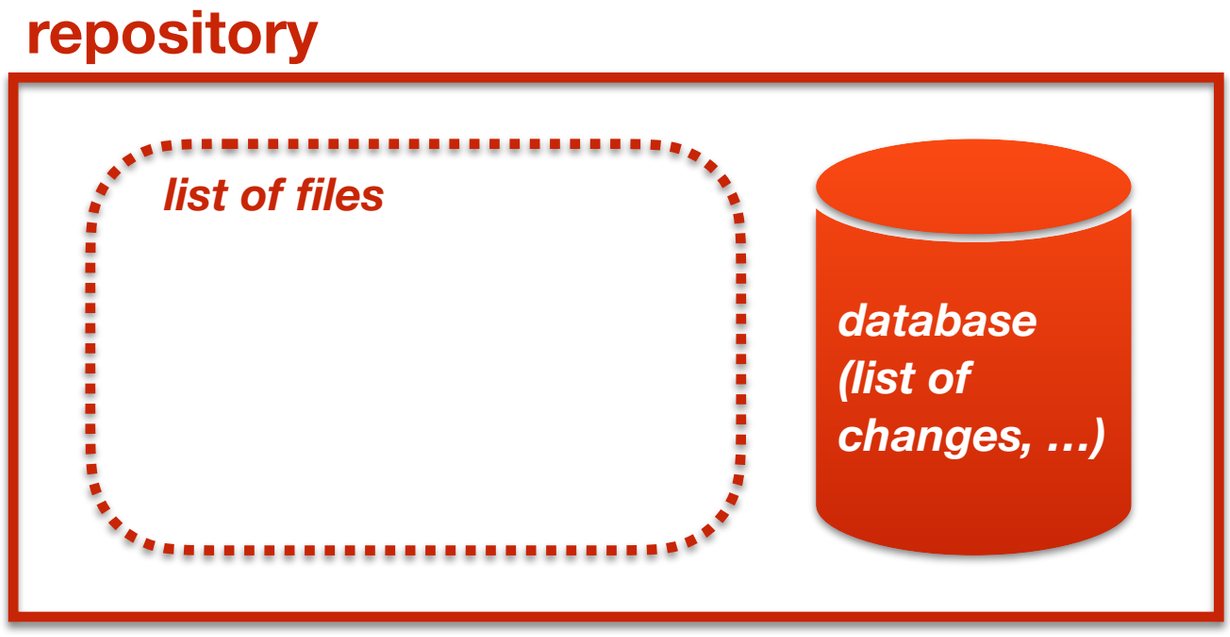
Carl Herrmann

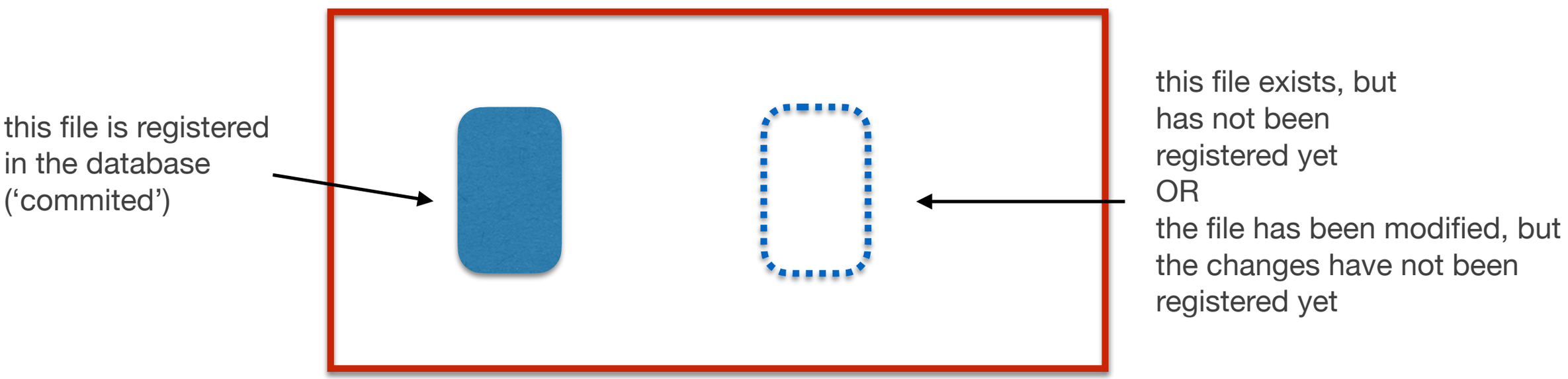
IPMB - BioQuant

Git(Hub)

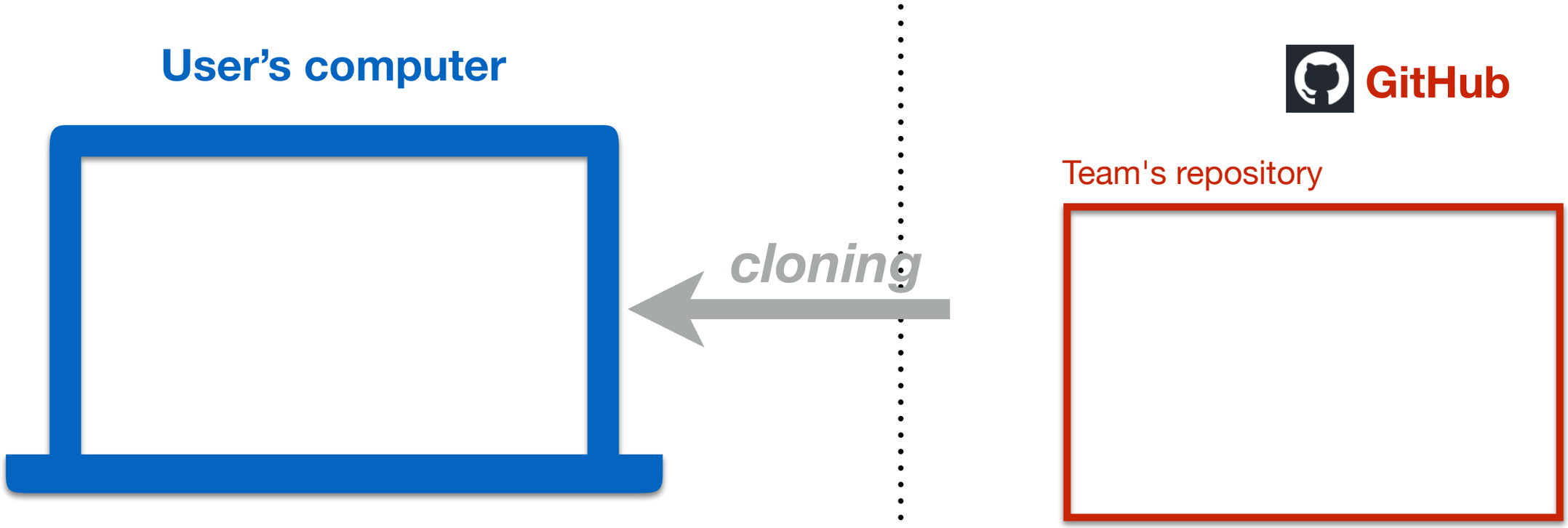
- Git is a **version control system**:
  - allows simultaneous work of different people on the same project
  - tracks the changes (‘**commits**’) made by each member
  - helps solve the **conflicts** between various versions
- GitHub is a plateform which hosts Git projects (‘repositories’)
  - is free to use
  - required to create a (free) account
  - can be used in command line or using GUI tools (‘GitHub Desktop’)

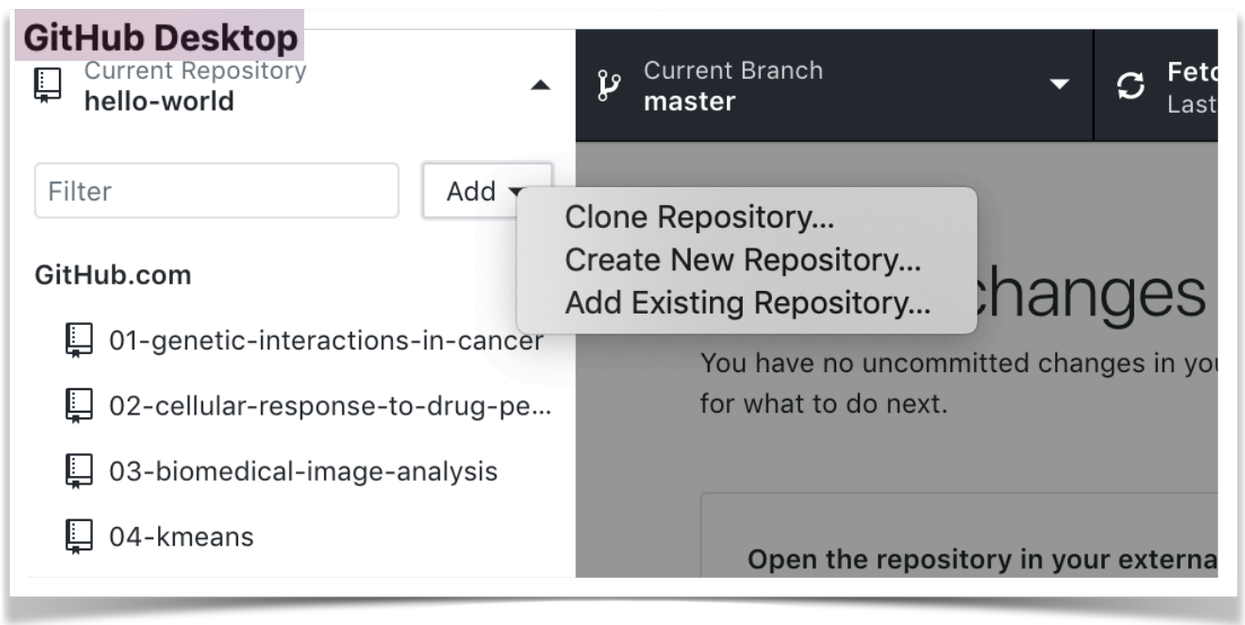
Git repository

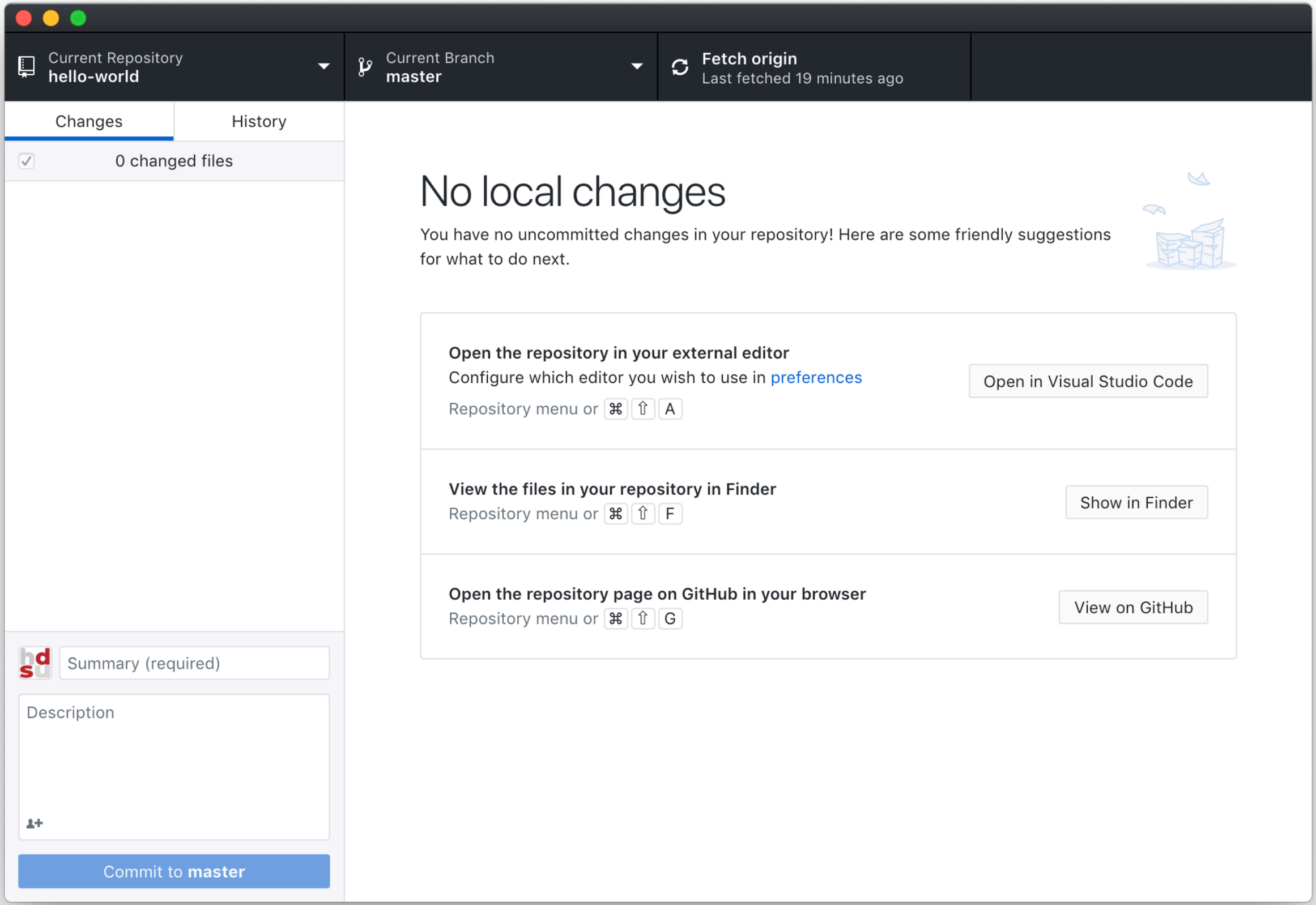




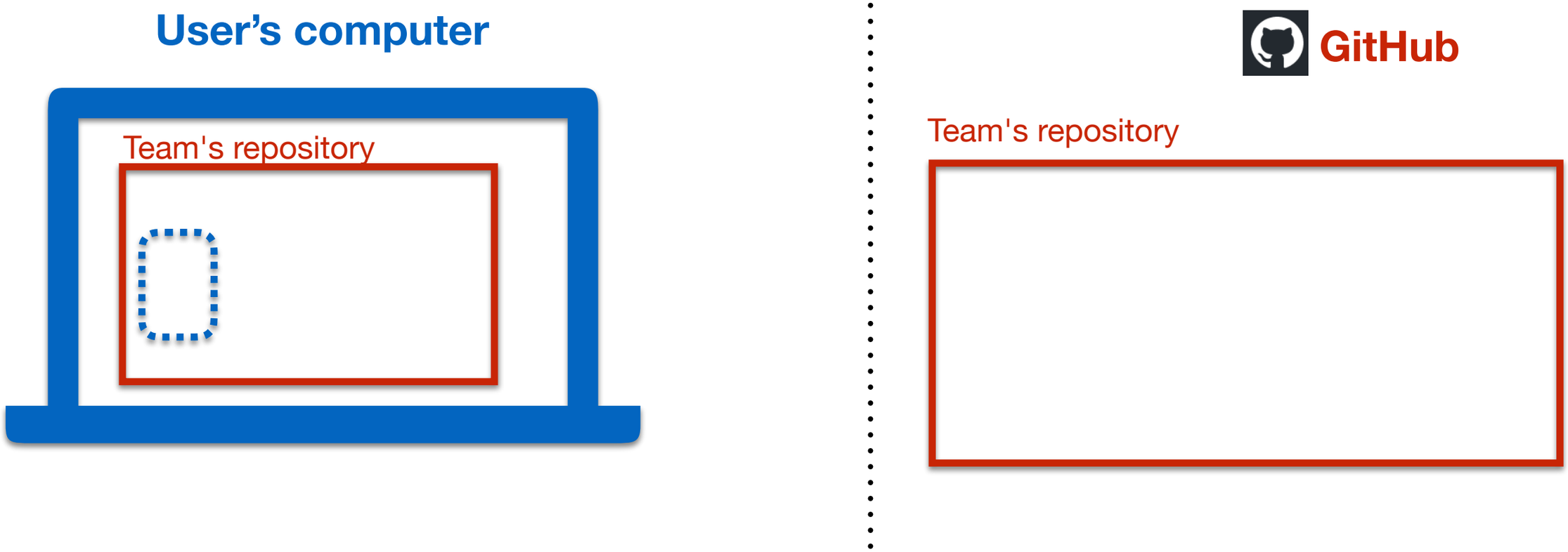
1. Cloning an existing repository






Current Repository
hello-world
Current Branch
master
Fetch origin
Last fetched 19 minutes ago
Changes
History
0 changed files
No local changes
You have no uncommitted changes in your repository! Here are some friendly suggestions for what to do next.
Open the repository in your external editor
Configure which editor you wish to use in preferences
Repository menu or ⌘ ⇧ A
Open in Visual Studio Code
View the files in your repository in Finder
Repository menu or ⌘ ⇧ F
Show in Finder
Open the repository page on GitHub in your browser
Repository menu or ⌘ ⇧ G
View on GitHub
Summary (required)
Description
Commit to master

2. creating a local file



- When a new file is added / modified in the local folder, it is not yet registered in the git database!
- it first needs to be **commited**

Current Repository
hello-world

Current Branch
master

Fetch origin
Last fetched 22 minutes ago

Changes 1 | History

1 changed file

my_markdown.Rmd

new file created locally

Create my_markdown.Rmd

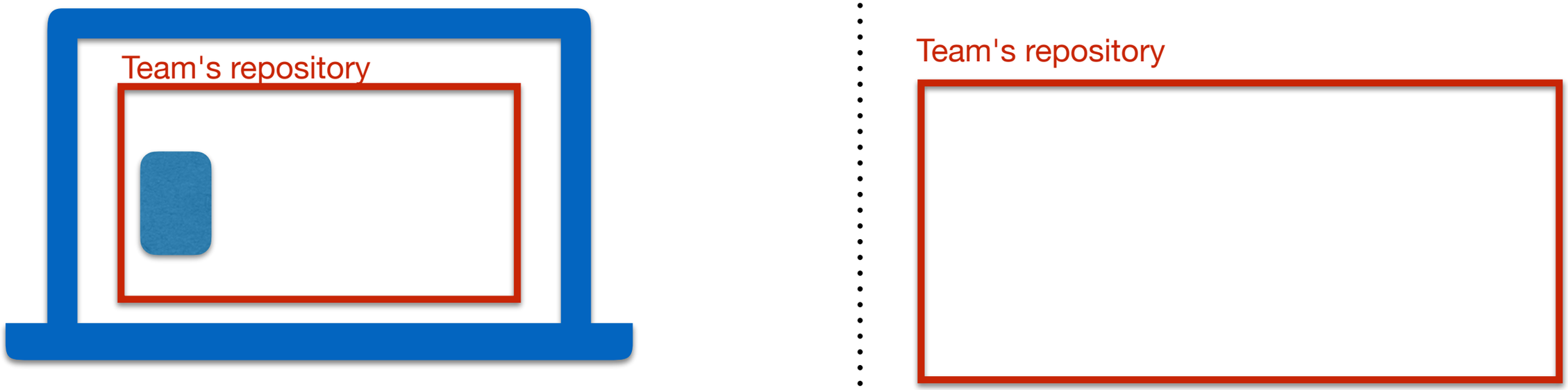
Description

Commit to **master**

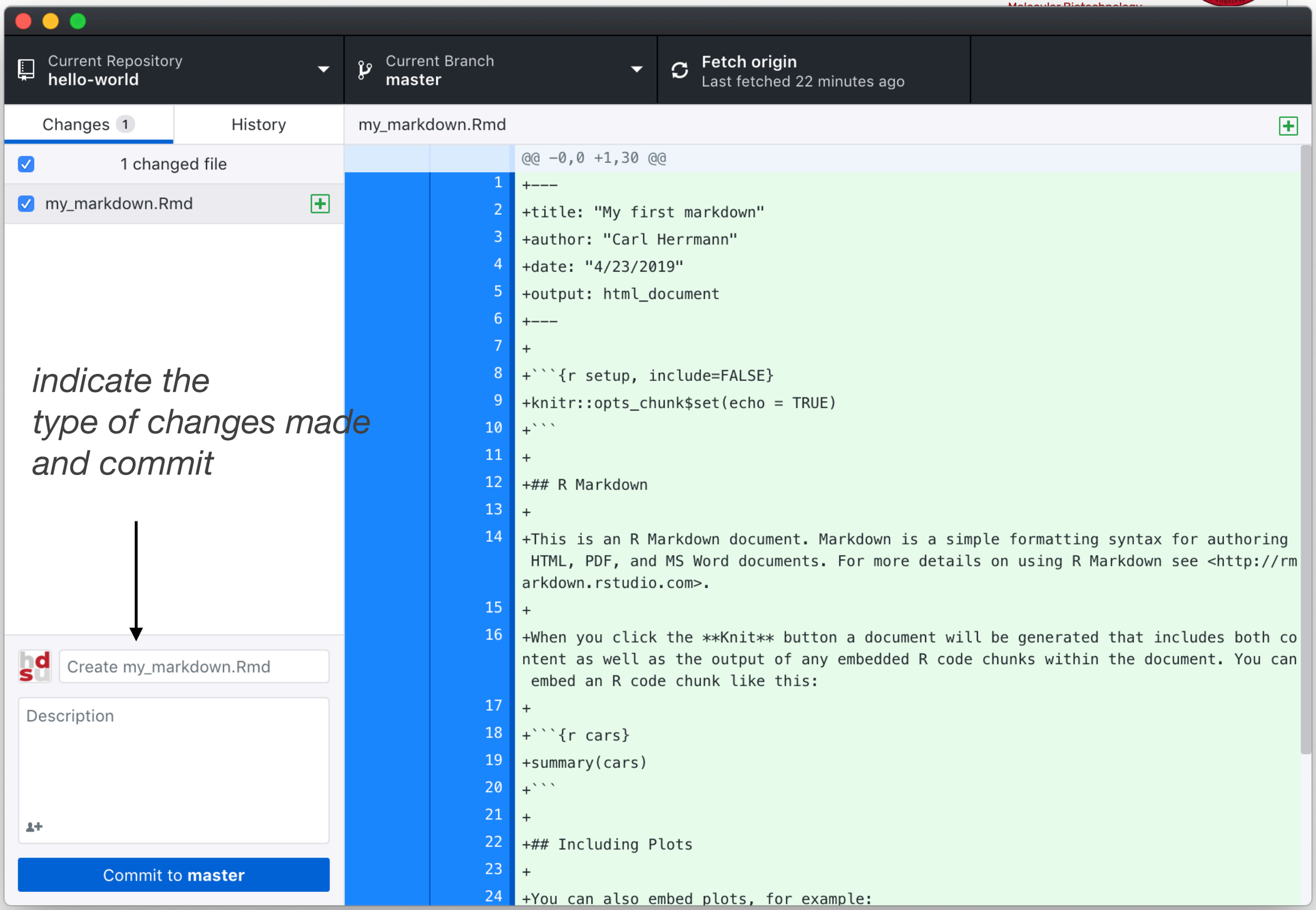
my_markdown.Rmd

```
@@ -0,0 +1,30 @@
1  +---
2  +title: "My first markdown"
3  +author: "Carl Herrmann"
4  +date: "4/23/2019"
5  +output: html_document
6  +---
7  +
8  +```{r setup, include=FALSE}
9  +knitr::opts_chunk$set(echo = TRUE)
10 +```
11 +
12 +## R Markdown
13 +
14 +This is an R Markdown document. Markdown is a simple formatting syntax for authoring
    HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rm
   arkdown.rstudio.com>.
15 +
16 +When you click the **Knit** button a document will be generated that includes both co
   ntent as well as the output of any embedded R code chunks within the document. You can
    embed an R code chunk like this:
17 +
18 +```{r cars}
19 +summary(cars)
20 +```
21 +
22 +## Including Plots
23 +
24 +You can also embed plots, for example:
```

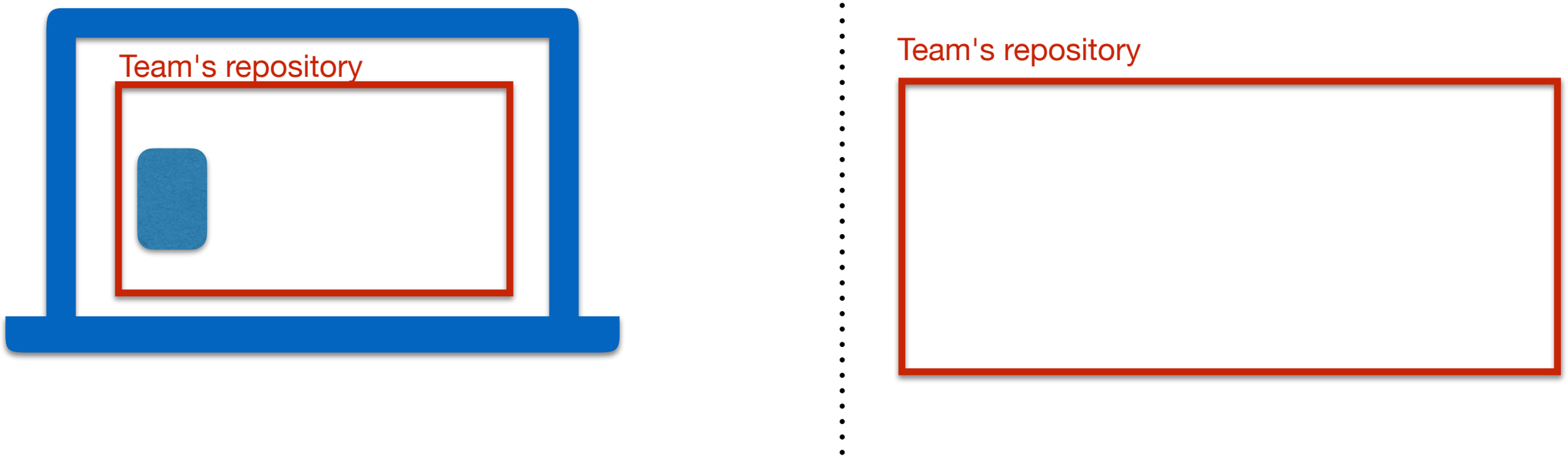
2. Adding a file



- When a new file is added / modified in the local folder, it is not yet registered in the git database!
- it first needs to be **commited**


Current Repository
hello-world
Current Branch
master
Fetch origin
Last fetched 22 minutes ago
Changes 1
History
my_markdown.Rmd
1 changed file
my_markdown.Rmd
indicate the type of changes made and commit
Create my_markdown.Rmd
Description
Commit to master
@@ -0,0 +1,30 @@
+---
+title: "My first markdown"
+author: "Carl Herrmann"
+date: "4/23/2019"
+output: html_document
+---
+
+```{r setup, include=FALSE}
+knitr::opts_chunk$set(echo = TRUE)
+```
+
+## R Markdown
+
+This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.
+
+When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:
+
+```{r cars}
+summary(cars)
+```
+
+## Including Plots
+
+You can also embed plots, for example:

2. Adding a file



- the file is now commited to the local git repository
- it needs to be pushed to the remote repository on GitHub

Current Repository
hello-world

Current Branch
master

Push origin
Last fetched 30 minutes ago
1

Changes | History

0 changed files

Summary (required)

Description

Commit to **master**

Committed just now
Create my_markdown.Rmd
Undo

No local changes

You have no uncommitted changes in your repository! Here are some friendly suggestions for what to do next.

Push 1 commit to the origin remote
You have one local commit waiting to be pushed to GitHub
Always available in the toolbar when there are local commits waiting to be pushed or ⌘ P

Push origin

Open the repository in your external editor
Configure which editor you wish to use in preferences
Repository menu or ⌘ ⇧ A

Open in Visual Studio Code

View the files in your repository in Finder
Repository menu or ⌘ ⇧ F

Show in Finder

Open the repository page on GitHub in your browser
Repository menu or ⌘ ⇧ G

View on GitHub

test repository

Edit

Manage topics

2 commits | 1 branch | 0 releases | 1 contributor

Branch: master | New pull request | Create new file | Upload files | Find File | Clone or download

carlherrmann Create my_markdown.Rmd — Latest commit b2b0bdf 2 minutes ago

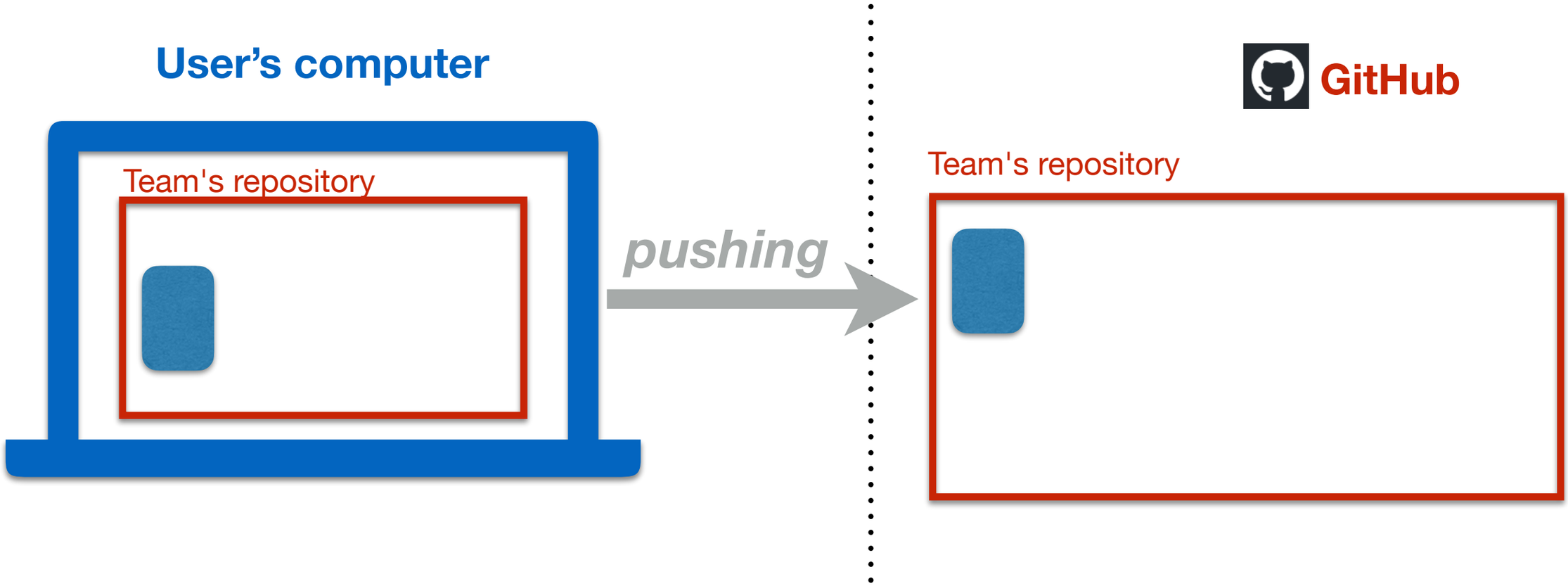
README.md	Initial commit	37 minutes ago
my_markdown.Rmd	Create my_markdown.Rmd	2 minutes ago

README.md

hello-world

test repository

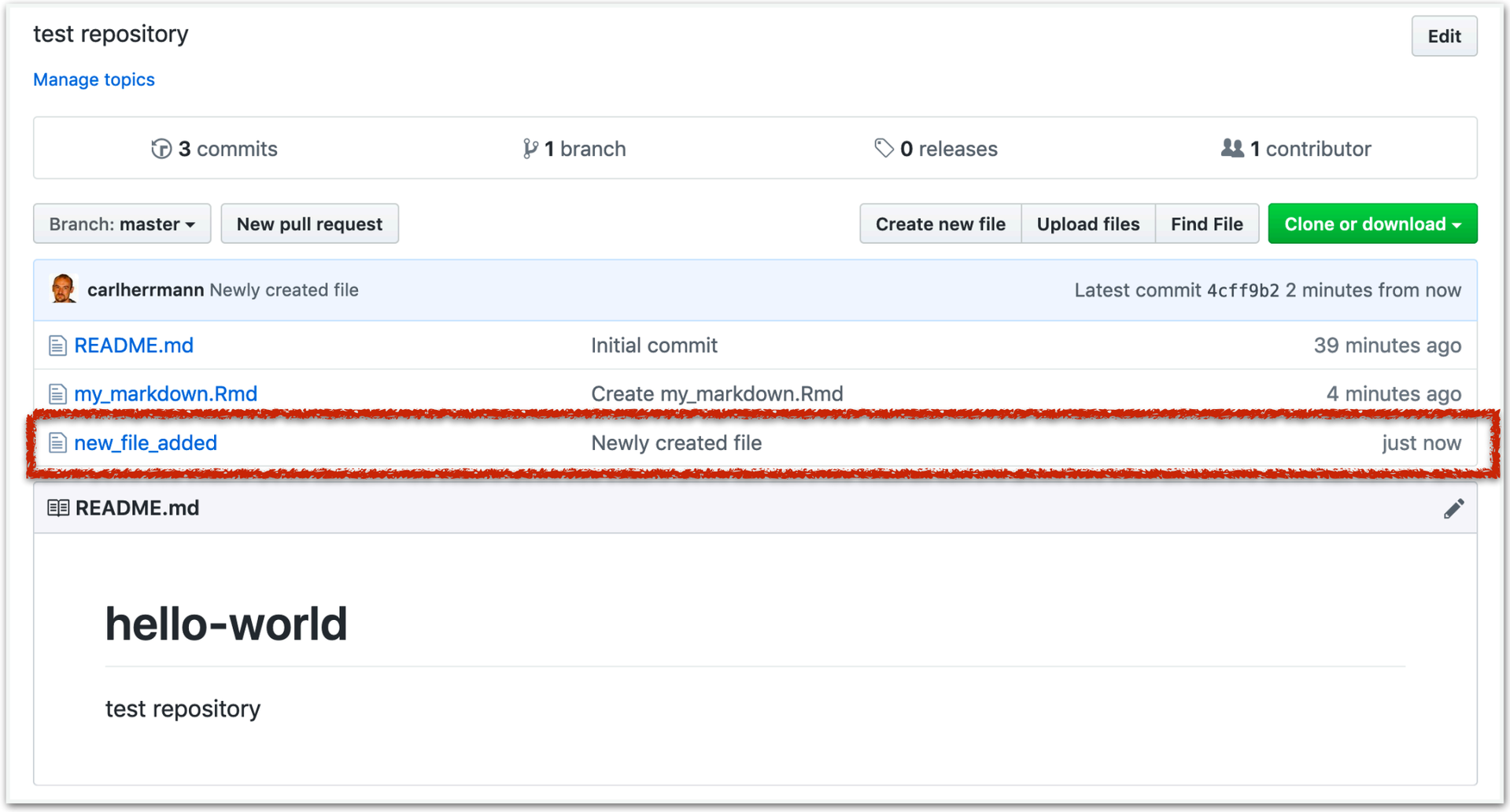
2. Adding a file



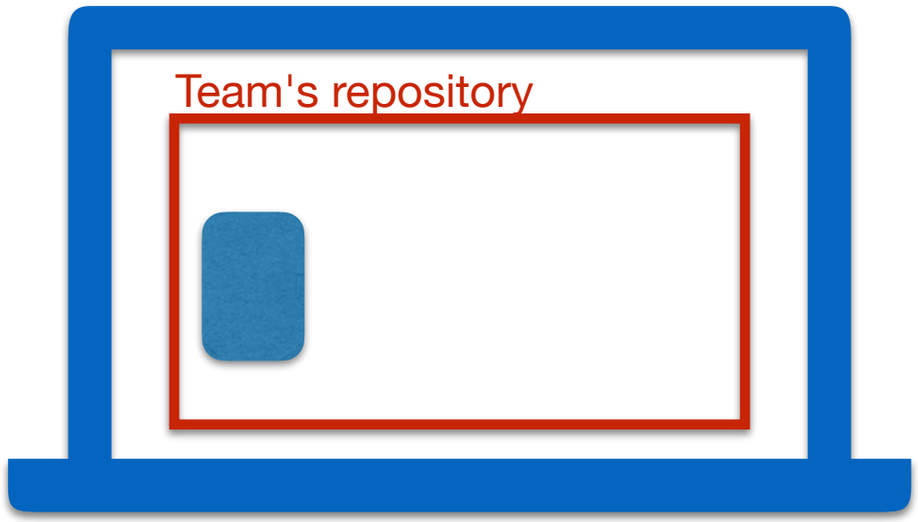
- the file is no commited to the local git repository
- it needs to be pushed to the remote repository on GitHub

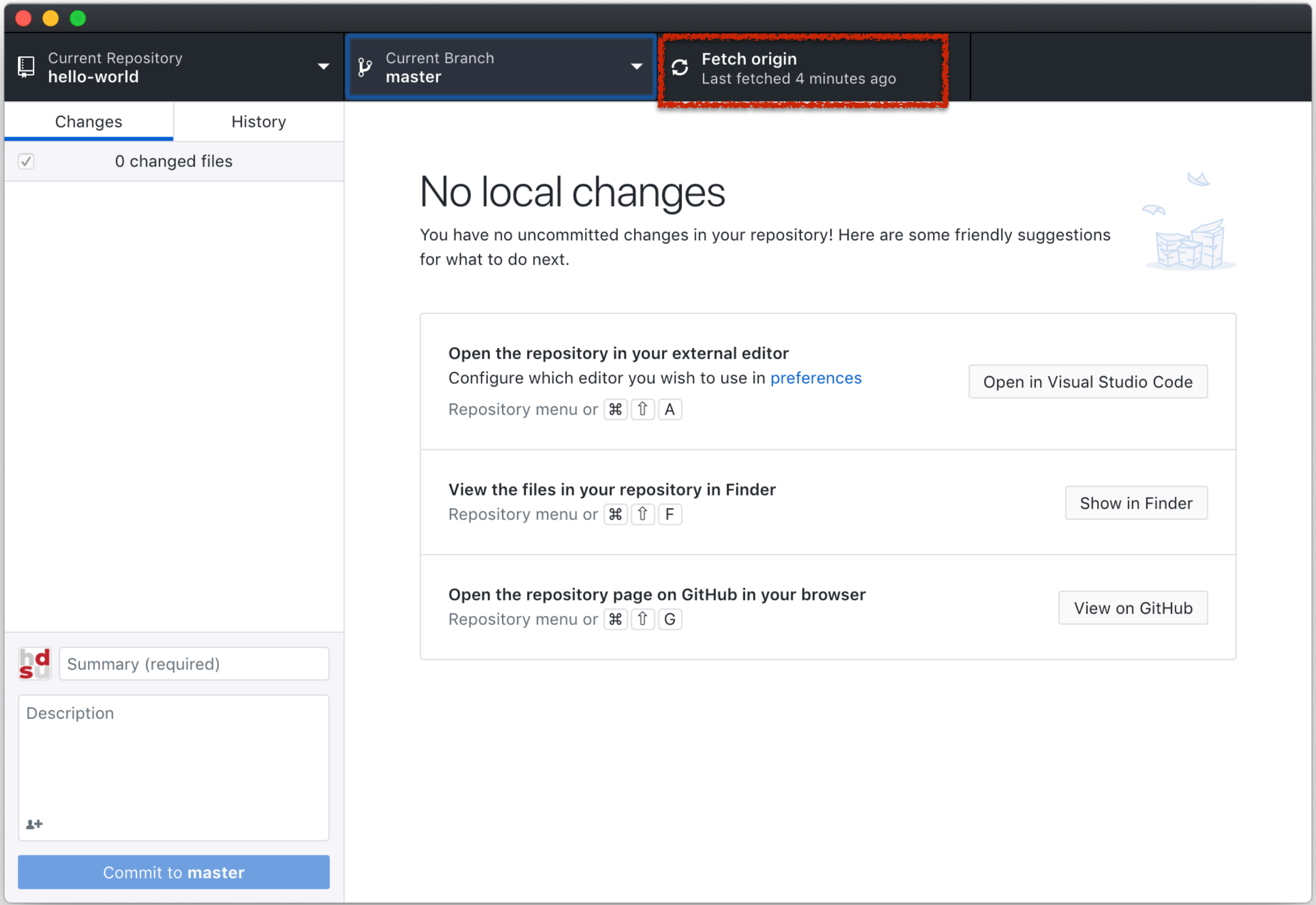
3. Pulling from the remote repository

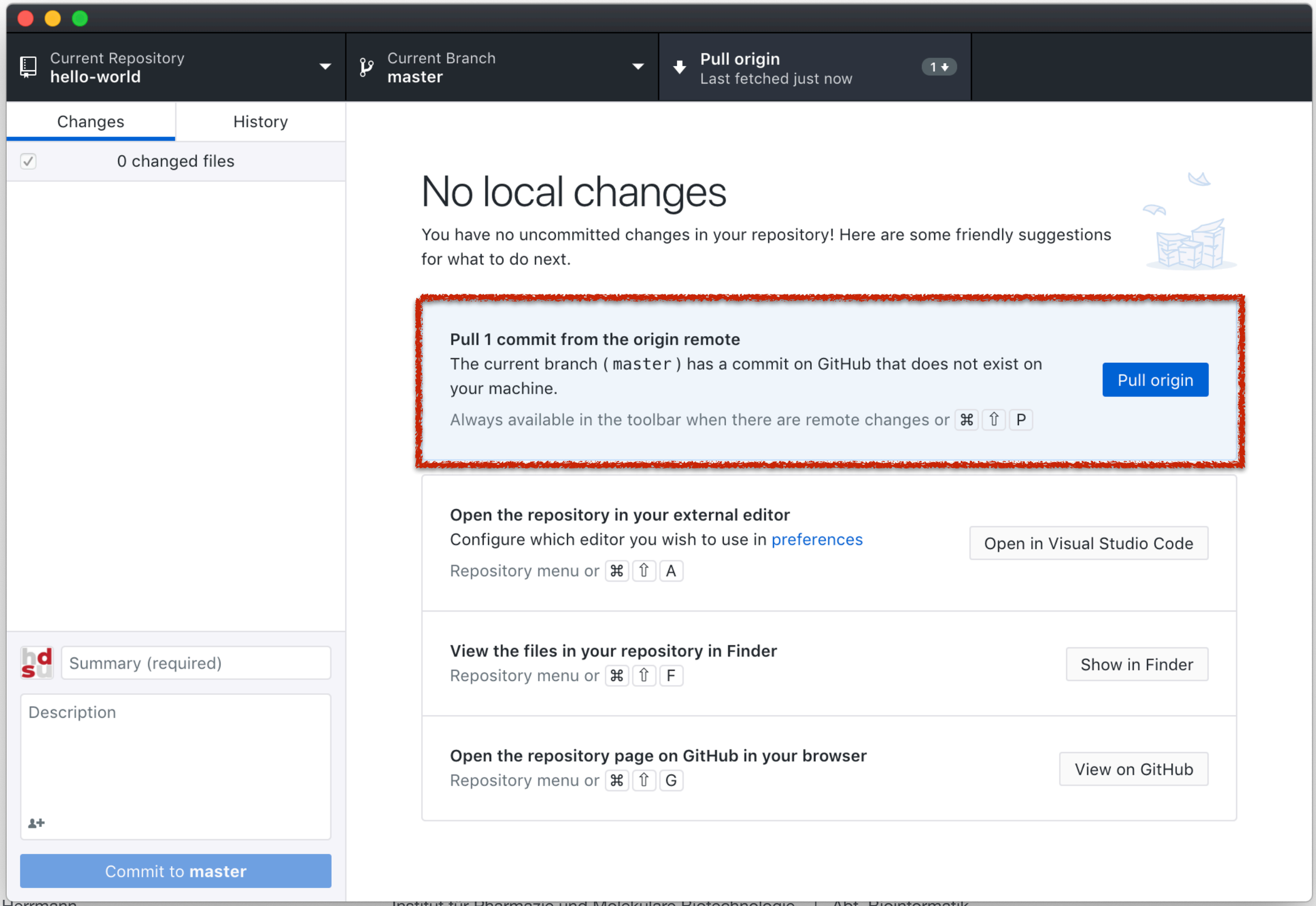
- Someone (probably one of your team mates) has added a new file into the remote repository
- It is not yet in your local repository and need to be **pulled**



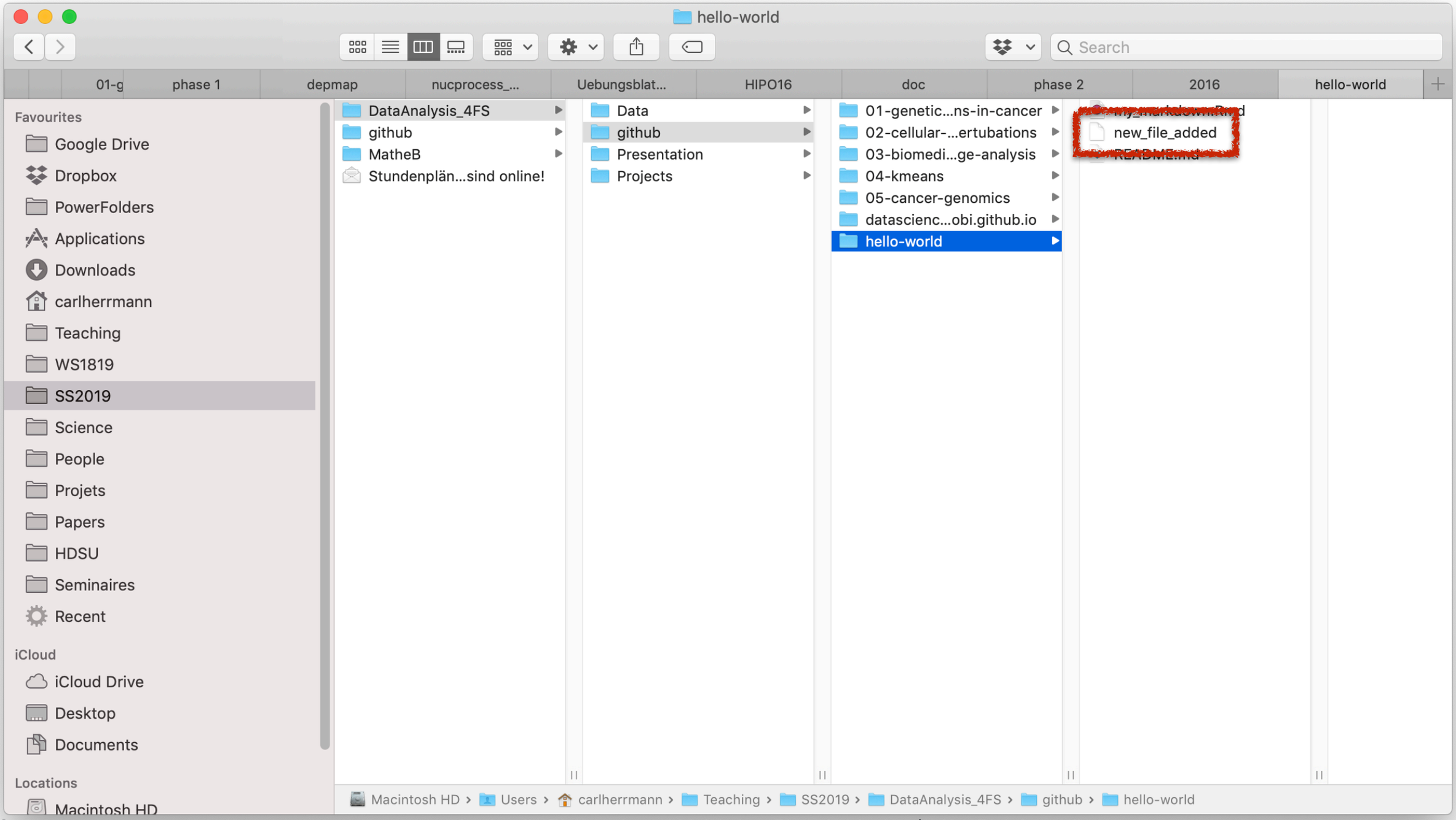
3. Pulling from the remote repository



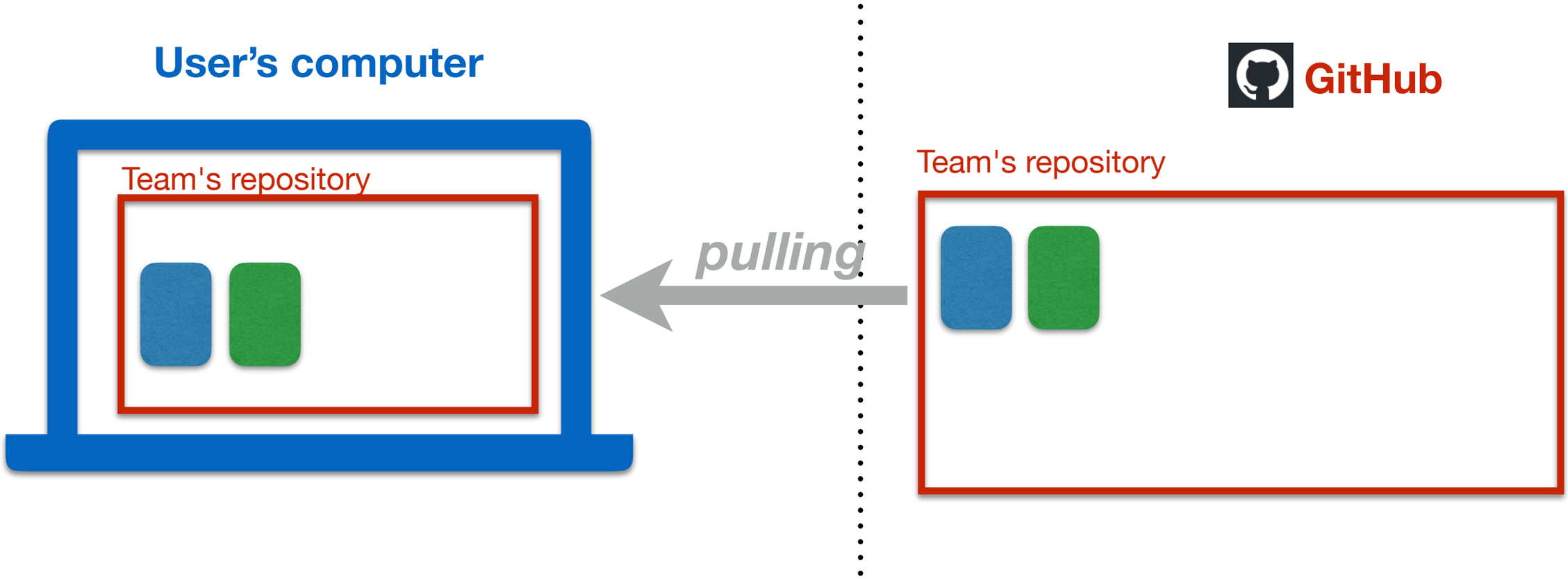

Current Repository
hello-world
Current Branch
master
Fetch origin
Last fetched 4 minutes ago
Changes
History
0 changed files
No local changes
You have no uncommitted changes in your repository! Here are some friendly suggestions for what to do next.
Open the repository in your external editor
Configure which editor you wish to use in preferences
Repository menu or ⌘ ⇧ A
Open in Visual Studio Code
View the files in your repository in Finder
Repository menu or ⌘ ⇧ F
Show in Finder
Open the repository page on GitHub in your browser
Repository menu or ⌘ ⇧ G
View on GitHub
Summary (required)
Description
Commit to master


Current Repository
hello-world
Current Branch
master
Pull origin
Last fetched just now
1
Changes
History
0 changed files
Summary (required)
Description
Commit to master
No local changes
You have no uncommitted changes in your repository! Here are some friendly suggestions for what to do next.
Pull 1 commit from the origin remote
The current branch (master) has a commit on GitHub that does not exist on your machine.
Pull origin
Always available in the toolbar when there are remote changes or ⌘ ⇧ P
Open the repository in your external editor
Configure which editor you wish to use in preferences
Repository menu or ⌘ ⇧ A
Open in Visual Studio Code
View the files in your repository in Finder
Repository menu or ⌘ ⇧ F
Show in Finder
Open the repository page on GitHub in your browser
Repository menu or ⌘ ⇧ G
View on GitHub

- Once the remote repository is pulled, the new file(s) are available locally

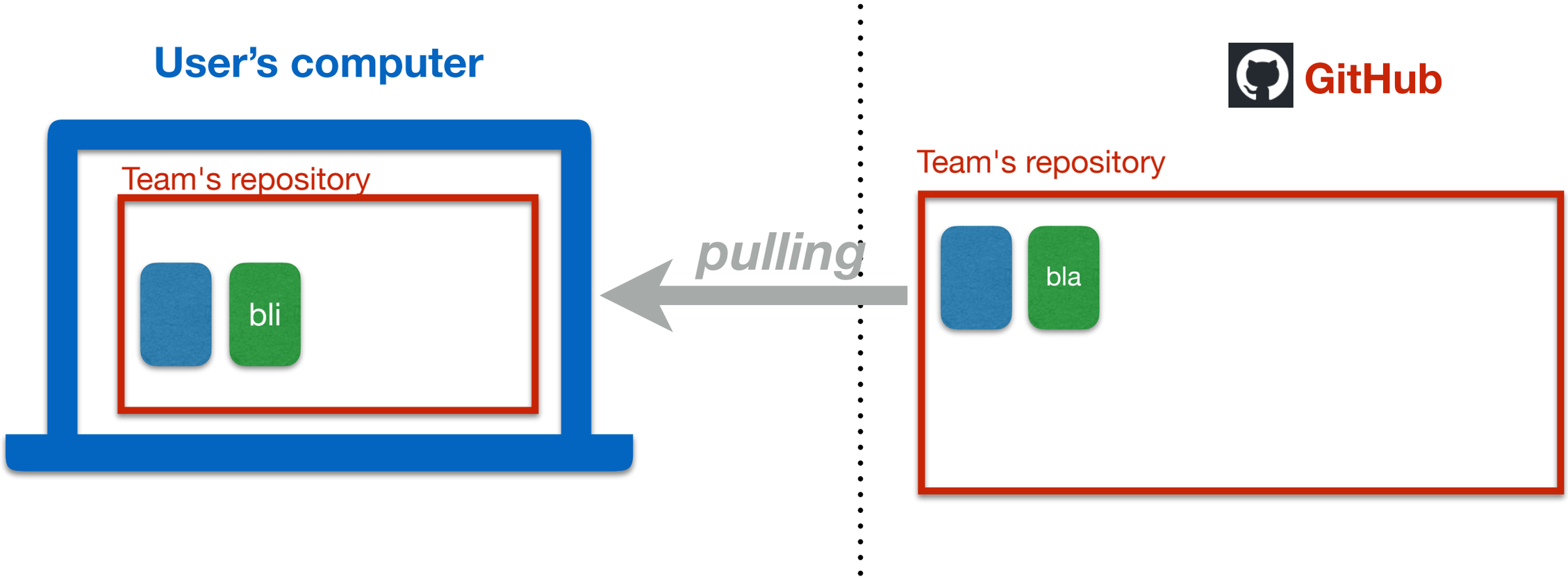


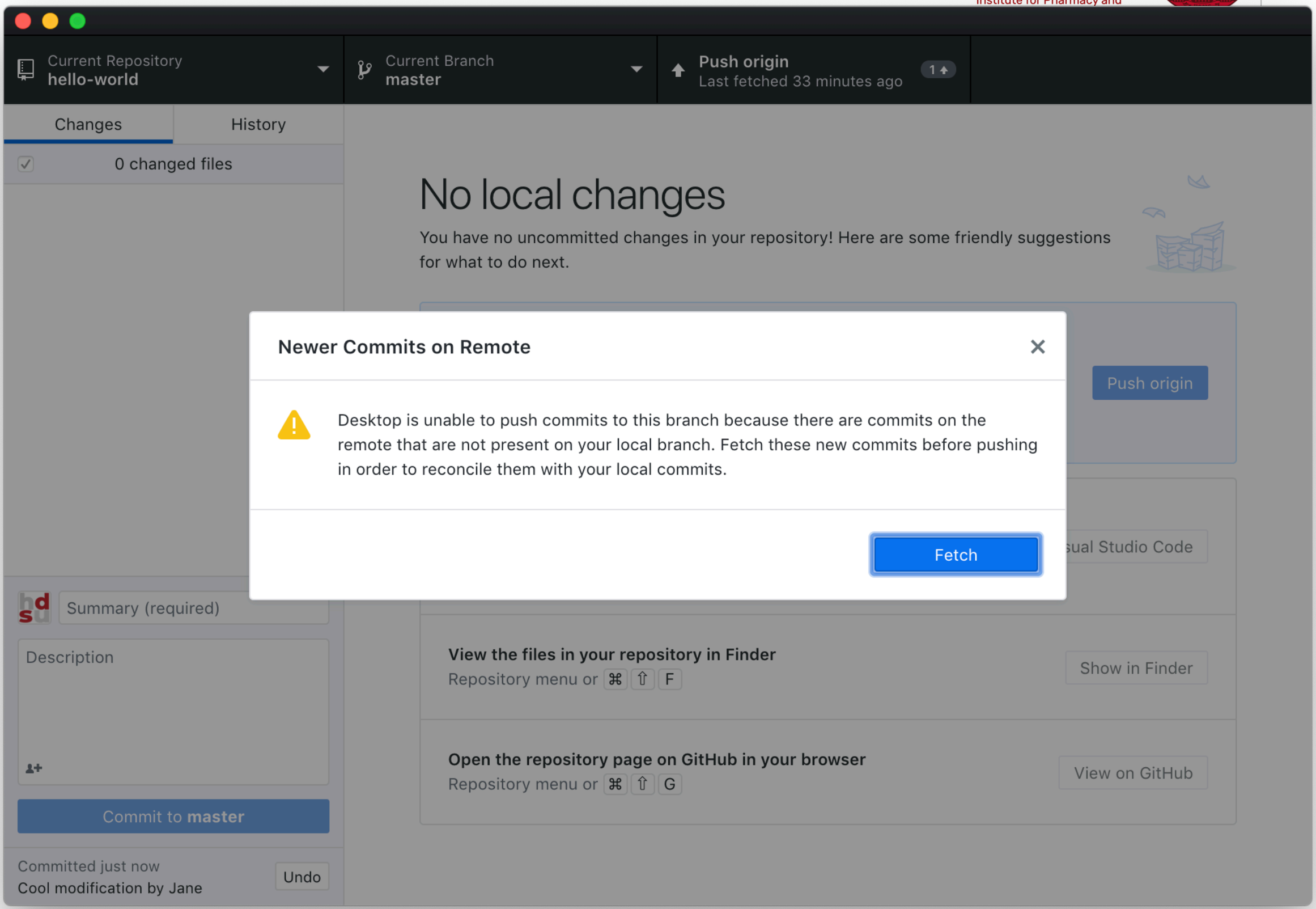
3. Pulling from the remote repository

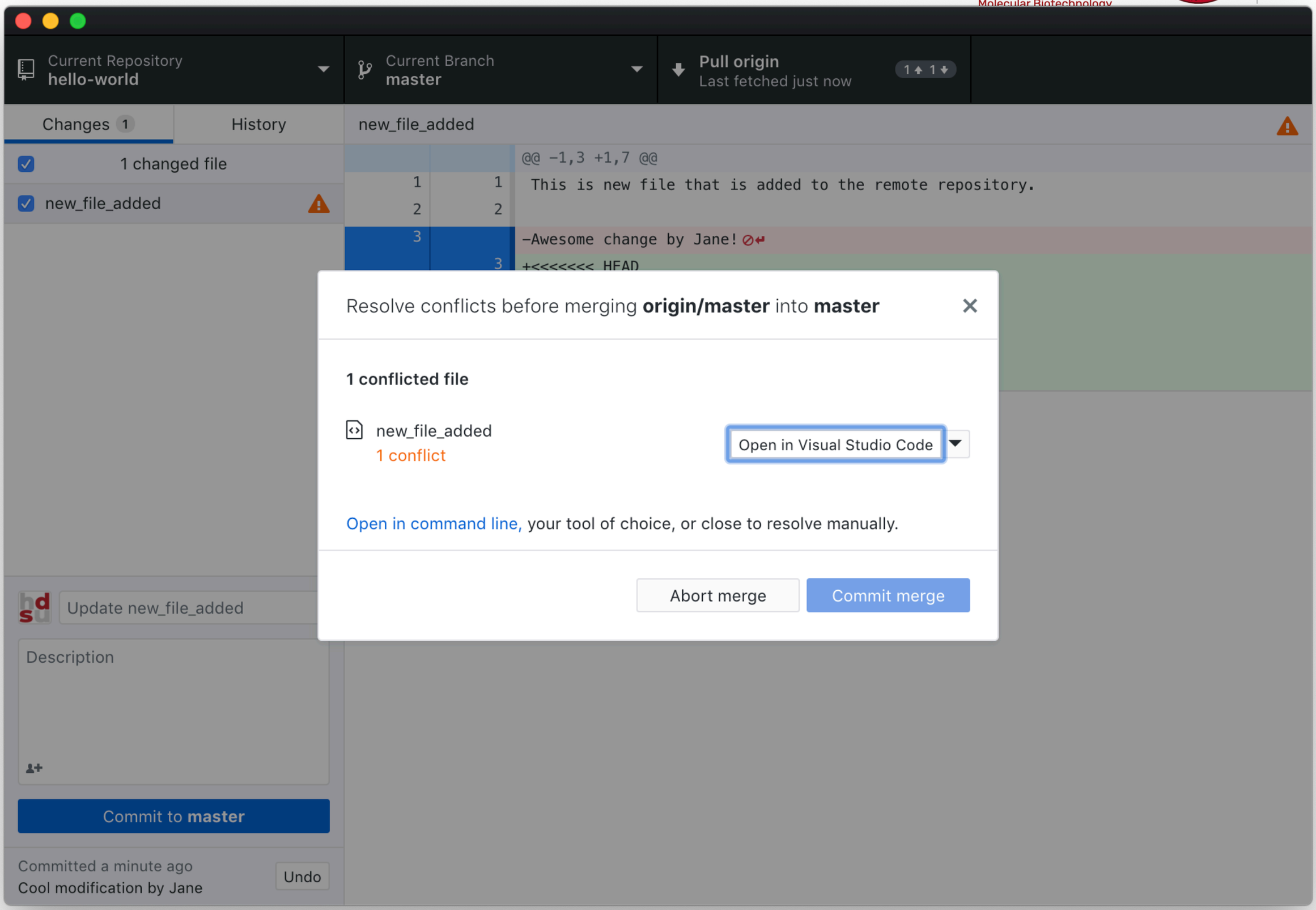


4. conflicting changes

IPMB
Institute for Pharmacy and Molecular Biotechnology
UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386




Current Repository
hello-world
Current Branch
master
Push origin
Last fetched 33 minutes ago
1
Changes
History
0 changed files
No local changes
You have no uncommitted changes in your repository! Here are some friendly suggestions for what to do next.
Newer Commits on Remote
Desktop is unable to push commits to this branch because there are commits on the remote that are not present on your local branch. Fetch these new commits before pushing in order to reconcile them with your local commits.
Fetch
Push origin
sual Studio Code
Summary (required)
Description
Commit to master
Committed just now
Cool modification by Jane
Undo
View the files in your repository in Finder
Repository menu or ⌘ ⇧ F
Show in Finder
Open the repository page on GitHub in your browser
Repository menu or ⌘ ⇧ G
View on GitHub


Current Repository
hello-world
Current Branch
master
Pull origin
Last fetched just now
1 ↑ 1 ↓
Changes 1
History
new_file_added
1 changed file
new_file_added
@@ -1,3 +1,7 @@
1 1 This is new file that is added to the remote repository.
2 2
3 -Awesome change by Jane!
3 +<<<<<<< HEAD
Resolve conflicts before merging origin/master into master
1 conflicted file
new_file_added
1 conflict
Open in Visual Studio Code
Open in command line, your tool of choice, or close to resolve manually.
Abort merge
Commit merge
Update new_file_added
Description
Commit to master
Committed a minute ago
Cool modification by Jane
Undo

4. conflicting changes

- Conflicting changes can be resolved with a text editor
- options depend on which editor is used

```
You, a few seconds ago | 2 authors (You and others)
This is new file that is added to the remote repository.

Accept Current Change | Accept Incoming Change | Accept Both Changes | Compare Changes
<<<<<<< HEAD (Current Change)
Awesome change by Jane!
=======
This is a great new modification by Joe!
>>>>>>> af5e9c9981b21a39ed11d09f468bea576d669191 (Incoming Change)
```

local change

changes in the remote file

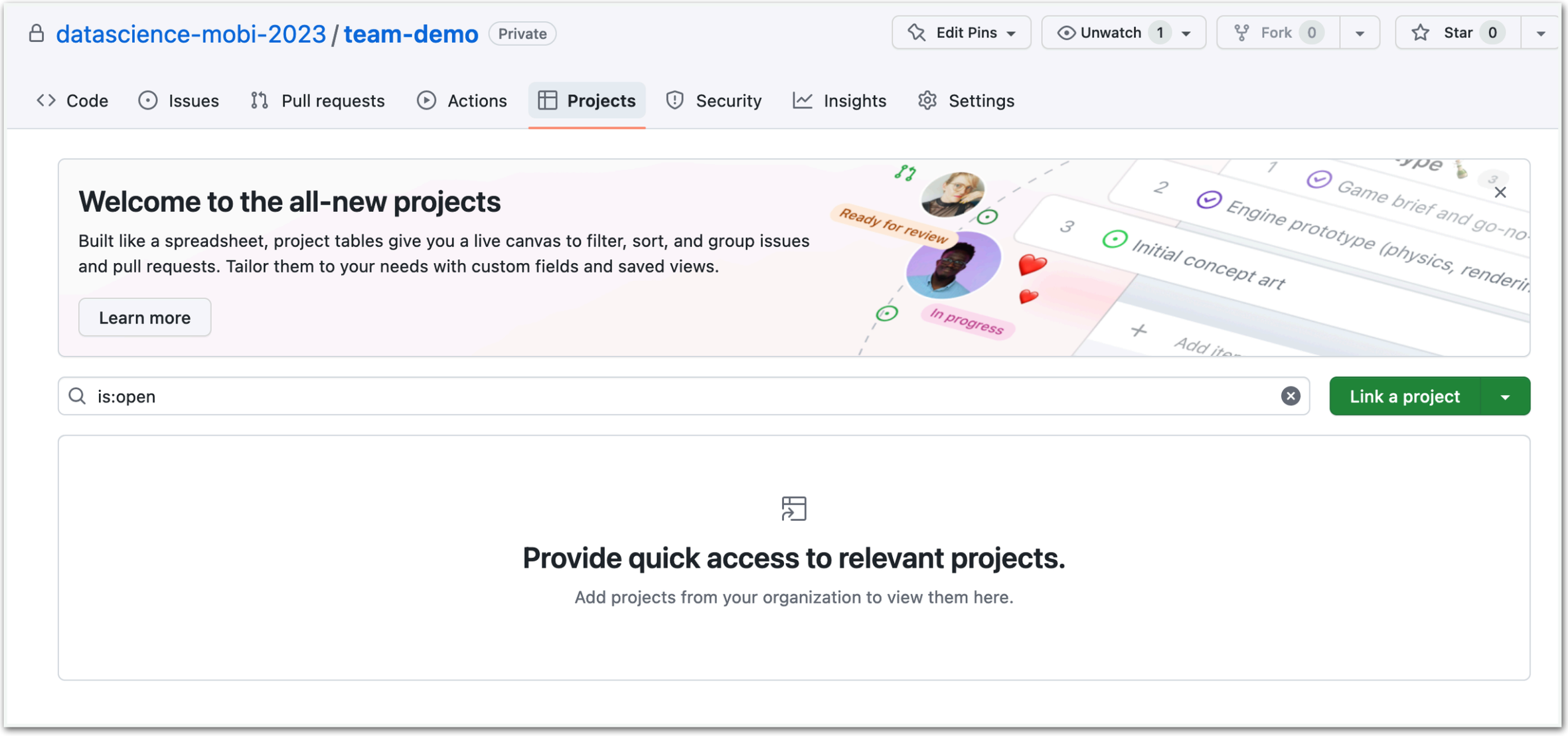
To do

- Create your own personal GitHub account

- Register your Github user name into the Google Sheet
  - all team members will be added to the corresponding GitHub repo
    - Project 03 - Team 02 → **`project-03-group-02`**

Using projects

- You can define tasks using the GitHub project tool



IPMB
Institute for Pharmacy and Molecular Biotechnology

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

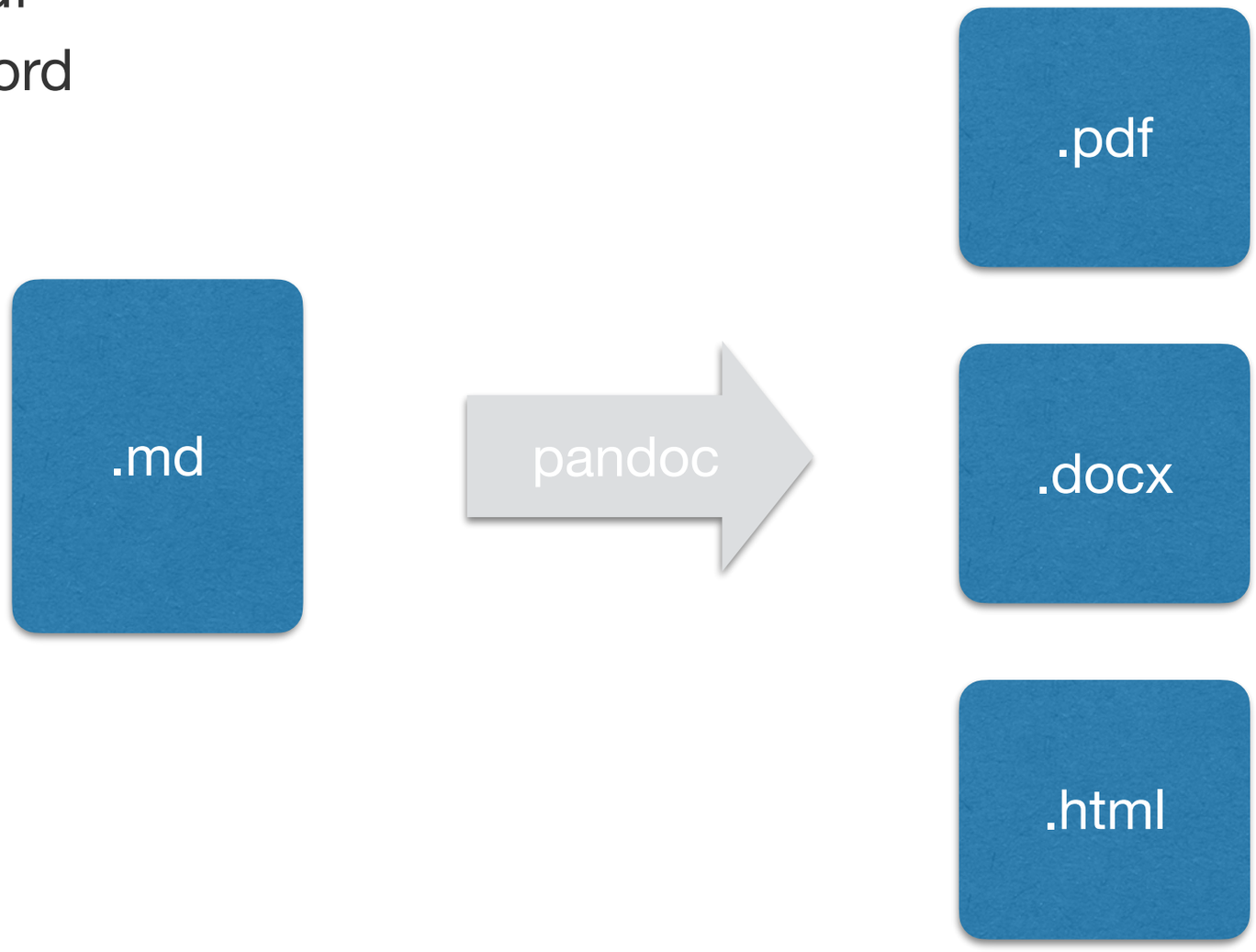
(R)markdown

Carl Herrmann

IPMB - BioQuant

Markdown

- Markdown is a way to format plain text with a simple text editor
- Markdown documents can be converted with a **renderer** into
  - html
  - pdf
  - word



Rendering markdown

Institute for Pharmacy and Molecular Biotechnology

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

markdown

```
# My document

## this is a header

In the text we can *highlight* or put in **bold**.

## making lists

We can make **numbered lists**:

1. first item
2. second item

or unordered lists

* first item
* second item
  + subitem
  + subitem
* third item

This is `code` which can be put inline

```bash
this is bash code
```

```python
this is python code
```
```

pdf

My document

this is a header

In the text we can *highlight* or put in **bold**.

making lists

We can make **numbered lists**:

1. first item
2. second item

or unordered lists

- first item
- second item
- subitem
- subitem
- third item

This is `code` which can be put inline

```
this is bash code
```

```
this is python code
```

html

My document

this is a header

In the text we can *highlight* or put in **bold**.

making lists

We can make **numbered lists**:

1. first item
2. second item

or unordered lists

- first item
- second item
- subitem
- subitem
- third item

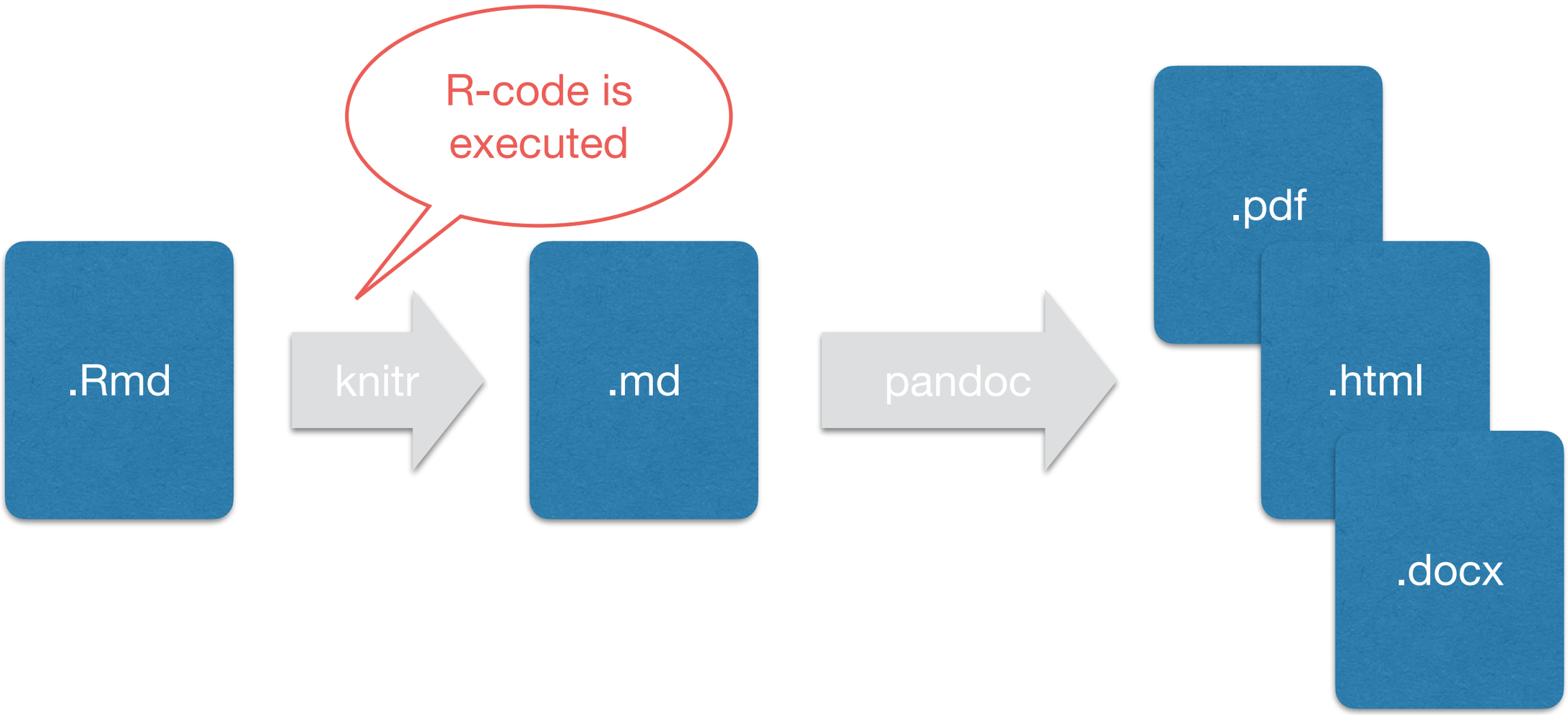
This is `code` which can be put inline

```
this is bash code
```

```
this is python code
```

Rmarkdown

- With Rmarkdown, R-code parts can be included into the markdown document
- the R-code will be executed, the result integrated into markdown



Rmarkdown format

IPMB
Institute for Pharmacy and Molecular Biotechnology
UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

```
---
title: "Project 01"
author: "Carl Herrmann"
date: "4/17/2019"
output:
  html_document:
    keep_md: yes
  pdf_document: default
---
# A Rmarkdown tutorial

This is a brief tutorial on how to use Rmarkdown to create dynamic documents

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_knit$set(root.dir='/Users/carlherrmann/Teaching/SS2019/DataAnalysis_4FS')
```

Load the dataset

```{r read_data}
allDepMapData = readRDS('Data/depmap/DepMap19Q1_allData.RDS')
```

Now plot the distribution of the cell lines according to the tissue type

```{r plot_data}
freq = sort(table(allDepMapData$annotation$Primary.Disease))
par(las=2,mar=c(3,8,3,3));barplot(freq,horiz=TRUE, col='lightgrey')
```
```

header: set options

R code chunks

text in markdown

Rmarkdown chunk options

- Display options can be set for each chunk individually, or for all chunks at the beginning of the document

```
```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
knitr::opts_chunk$set(cache = TRUE)
```

valid for all chunks
- echo=TRUE : R-code is displayed in final document
- cache = TRUE : results of all chunks are cached

```
```{r plot_data,fig.height=12,fig.width=12}
freq = sort(table(allDepMapData$annotation$Primary.Disease))
par(las=2,mar=c(3,8,3,3));barplot(freq,horiz=FALSE, col='lightgrey')
```
```

valid for **this** chunks
- set height and width of output figure
```

Reference

- https://rmarkdown.rstudio.com/
- https://www.rstudio.com/wp-content/uploads/2016/03/rmarkdown-cheatsheet-2.0.pdf