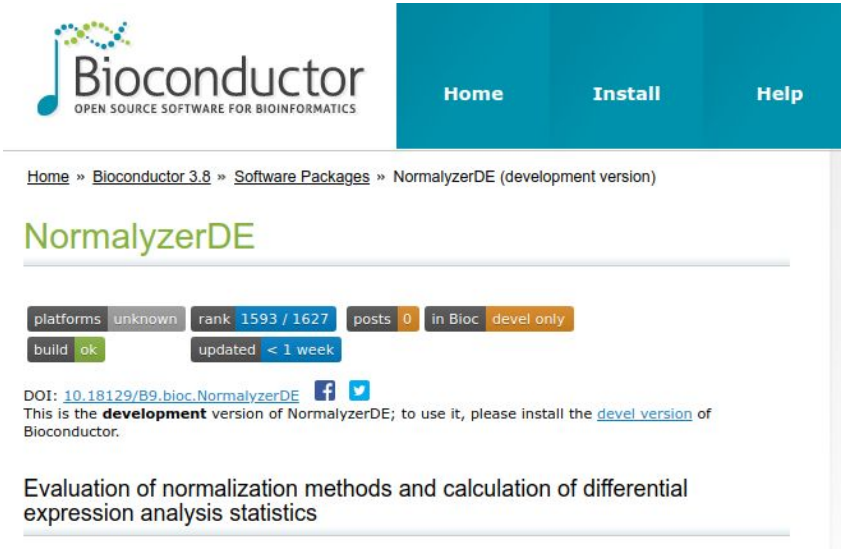
Git and version control in data science

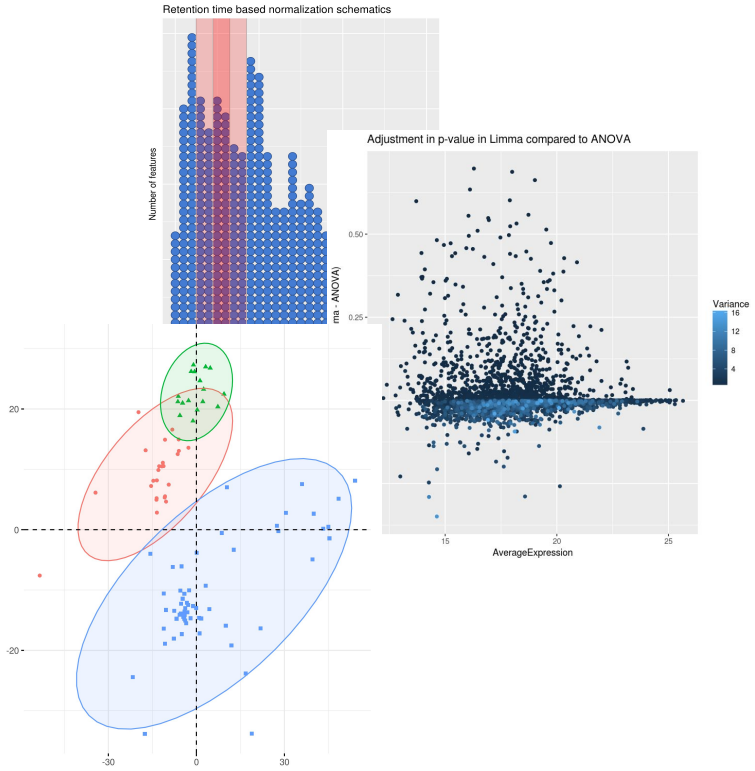
Jakob Willforss, Department of Immunotechnology, LTH
24th April 2020

I use Git when I...

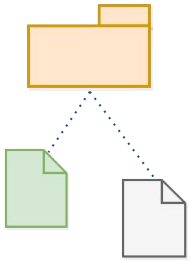
...develop software



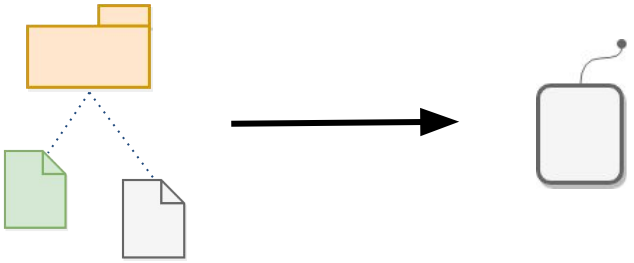
...analyze data



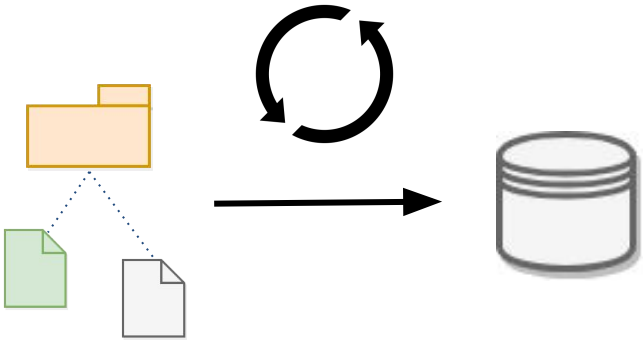
How we manage our files



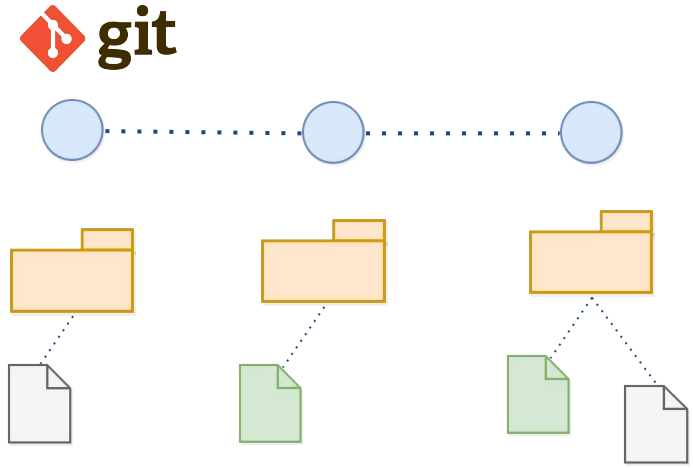
No backup



Manual backup

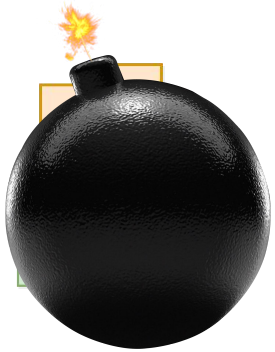


Automatic backup



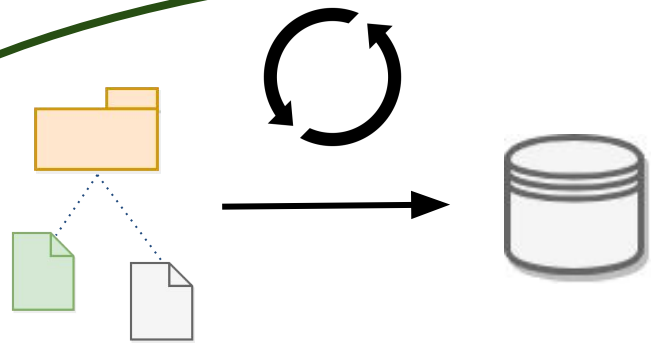
Version control

How we manage our files

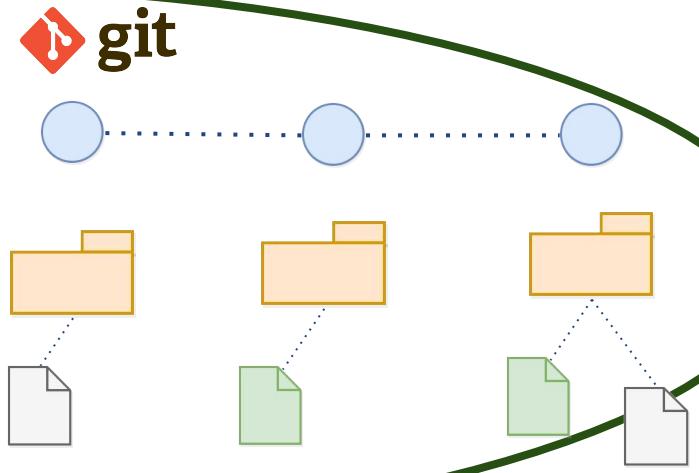


No backup

Manual backup

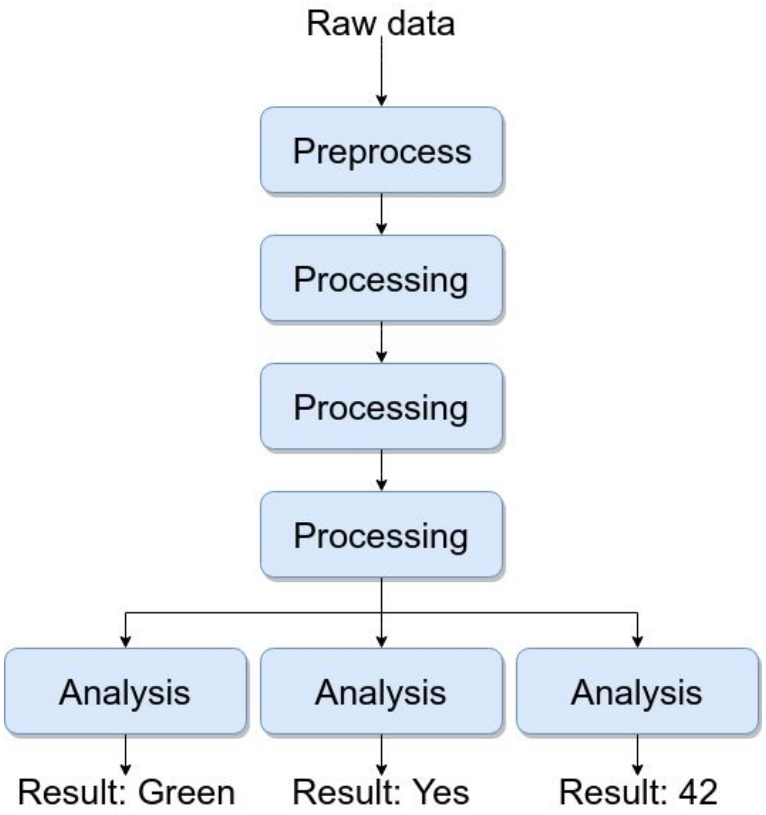


Automatic backup

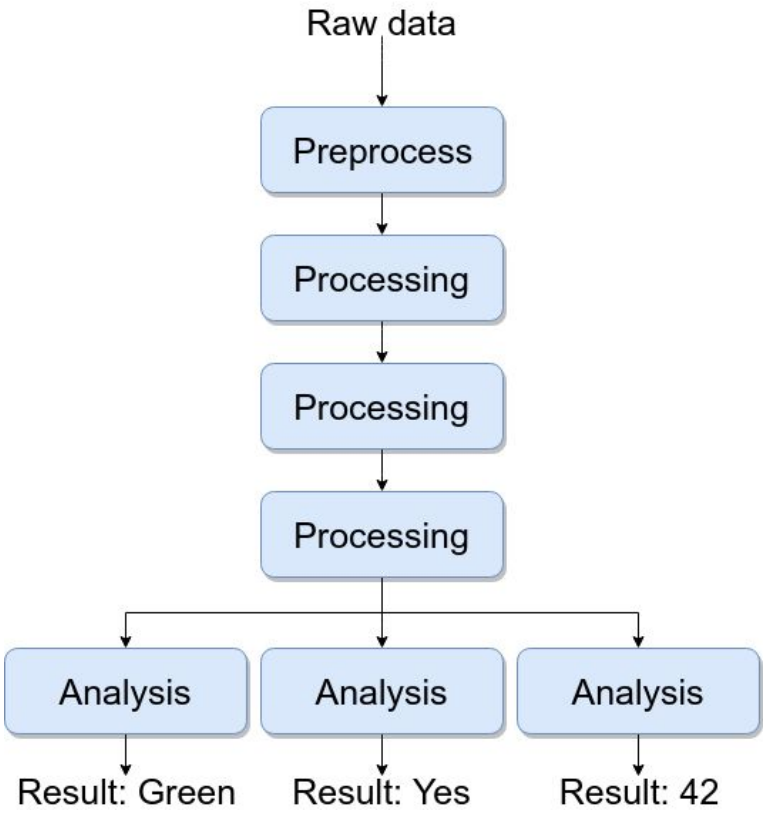


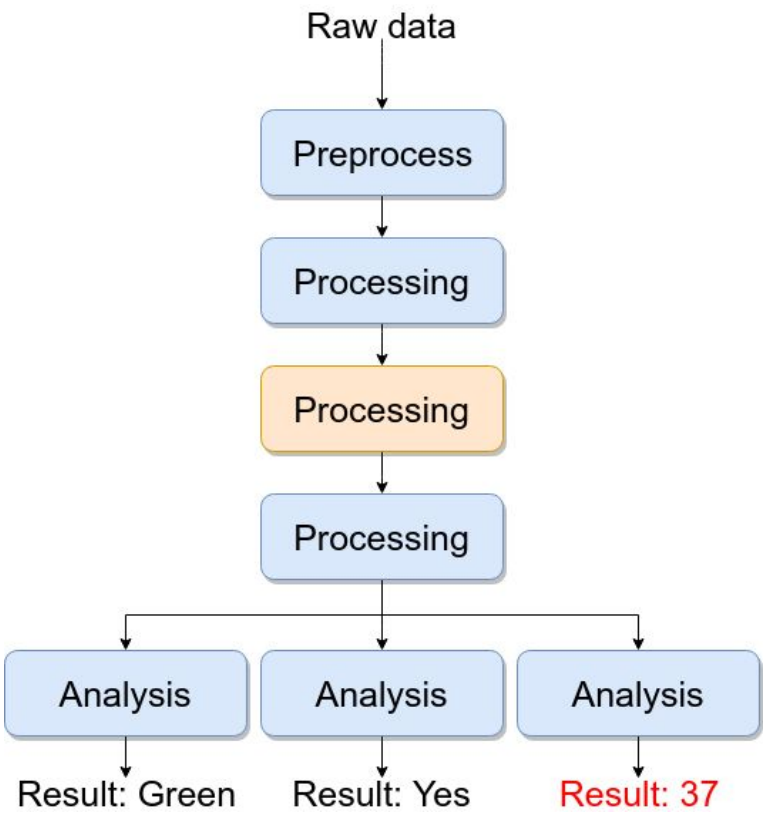
Version control

Reproducible research

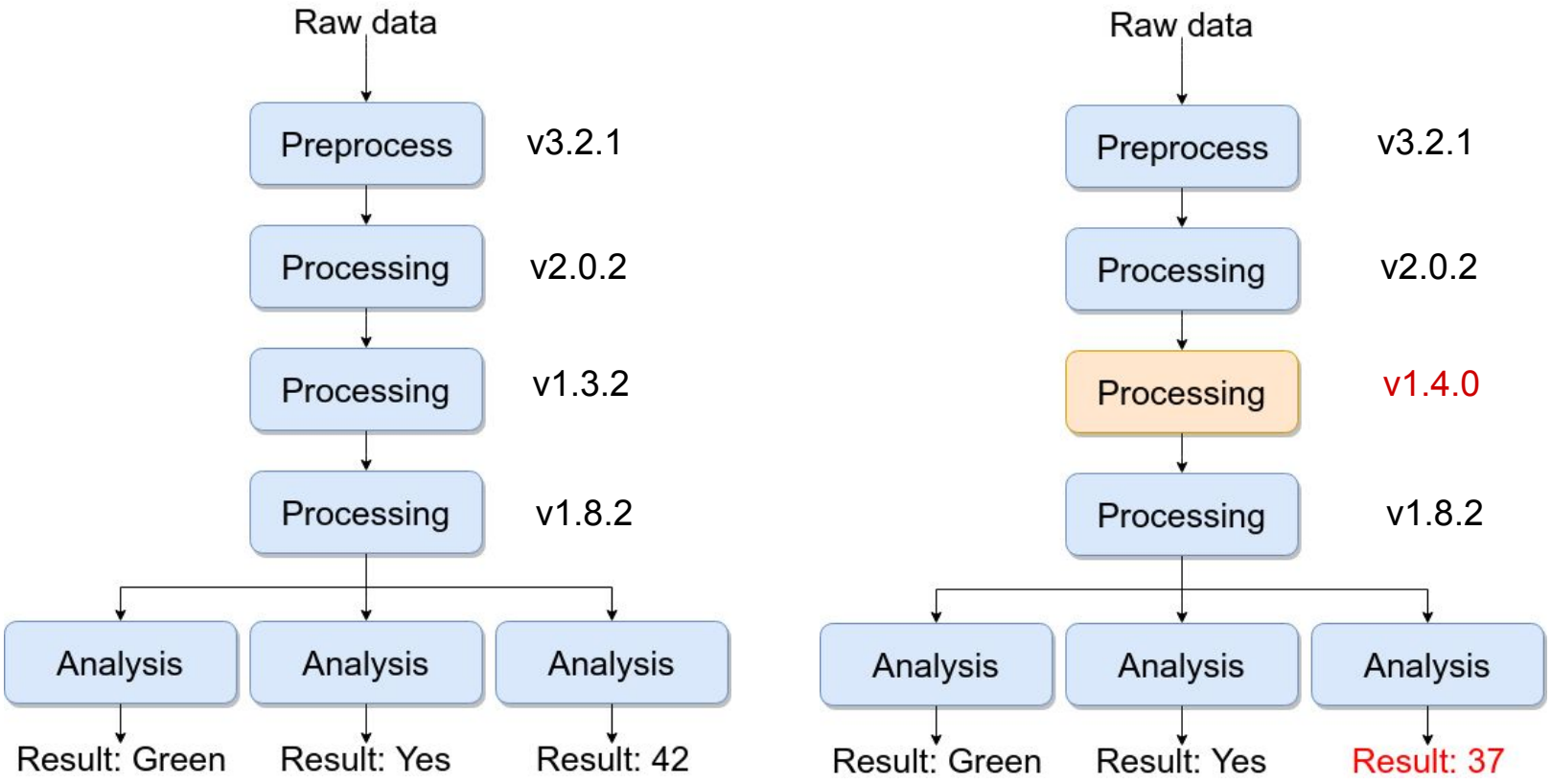


Reproducible research

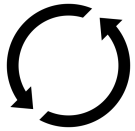




Reproducible research



Version control vs. backup



26th sept	1 file, 2 mb
27 sept	1 file, 2 mb
28 sept	2 files, 3 mb
29 sept	2 files, 3.5 mb
30 sept	2 files, 3.5 mb
1 oct	3 files, 3.8 mb
2 oct	5 files, 5.2 mb
3 oct	5 files, 5.2 mb
4 oct	5 files, 5.5 mb
5 oct	5 files 5.5 mb

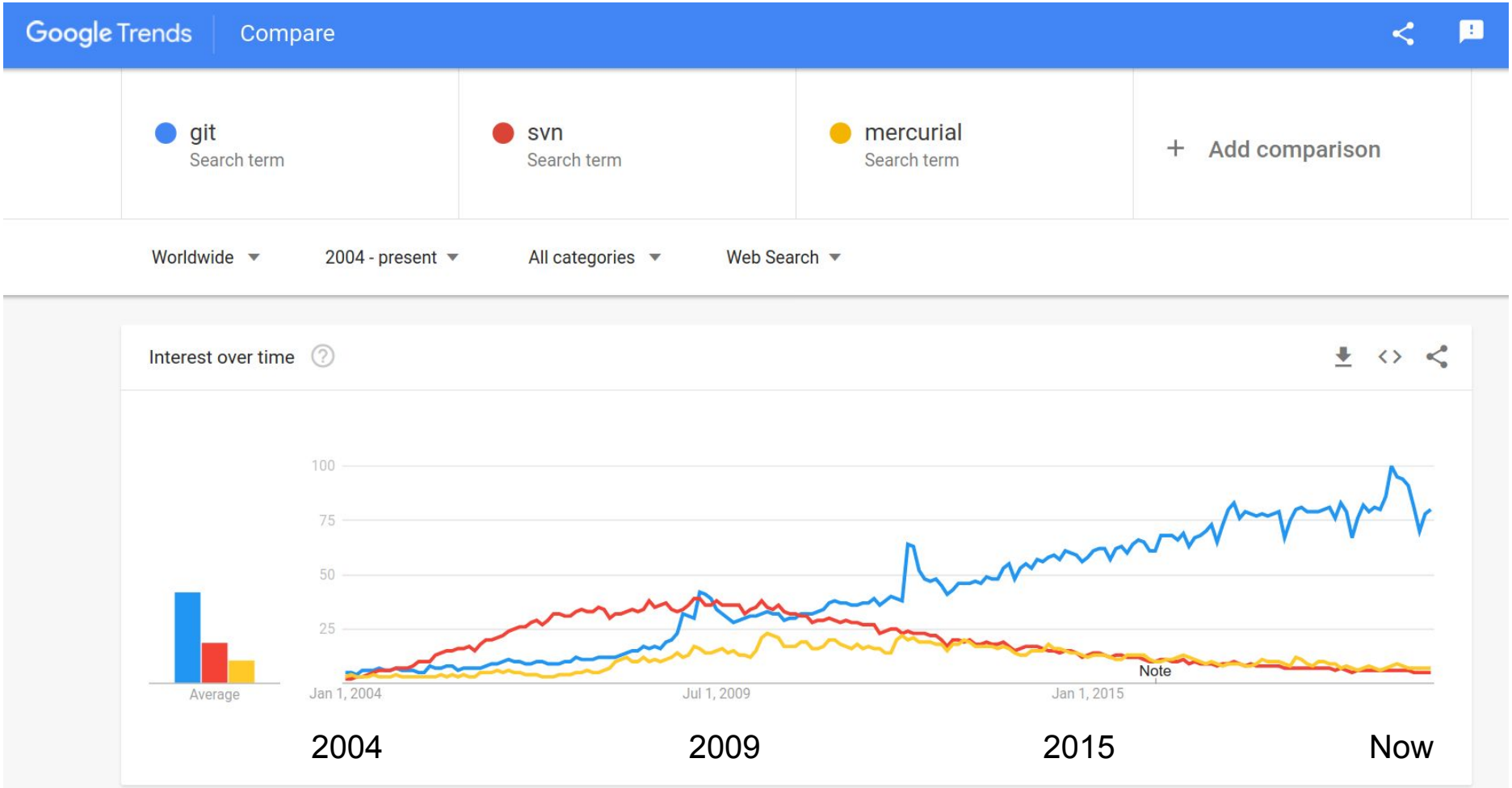
26th sept	Add first file	Tags
26 sept	Make important edits	
29 sept	Include second analysis	v2.0 Analysis 2
30 sept	Extend second analysis	
30 sept	Add visualizations for first analysis	
2 oct	Found errors in ANOVA, corrected	
2 oct	Preparing third analysis	
4 oct	Start third analysis	v3.0 Analysis 3

Git

Developed 2005 by Linus Torvald to maintain Linux source code

Linus Torvald

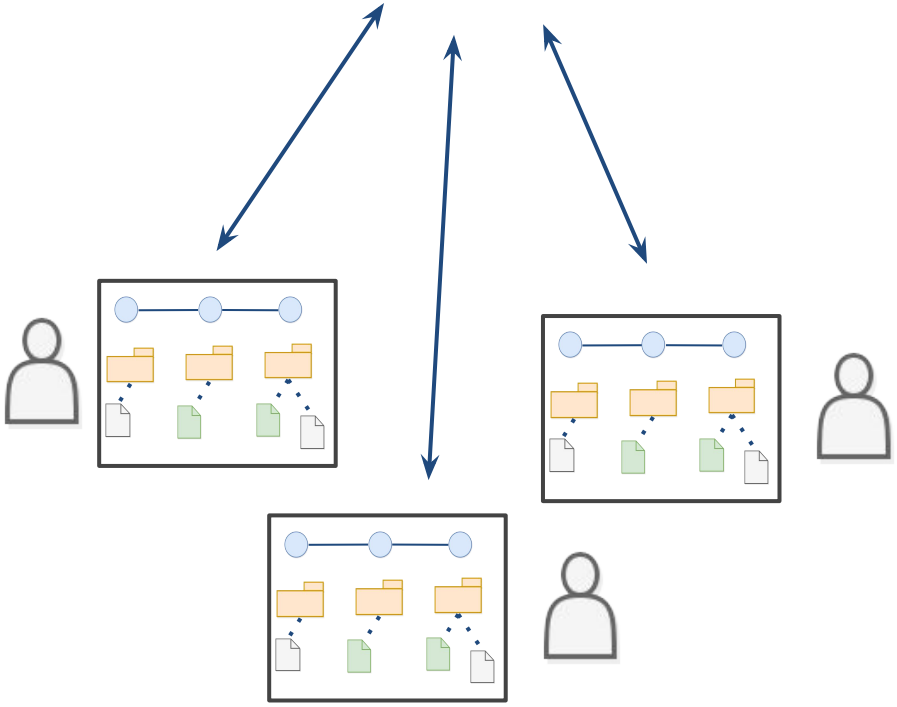
From wikimedia commons


Google Trends
Compare
git
Search term
svn
Search term
mercurial
Search term
Add comparison
Worldwide
2004 - present
All categories
Web Search
Interest over time
100
75
50
25
Note
Average
Jan 1, 2004
Jul 1, 2009
Jan 1, 2015
2004
2009
2015
Now

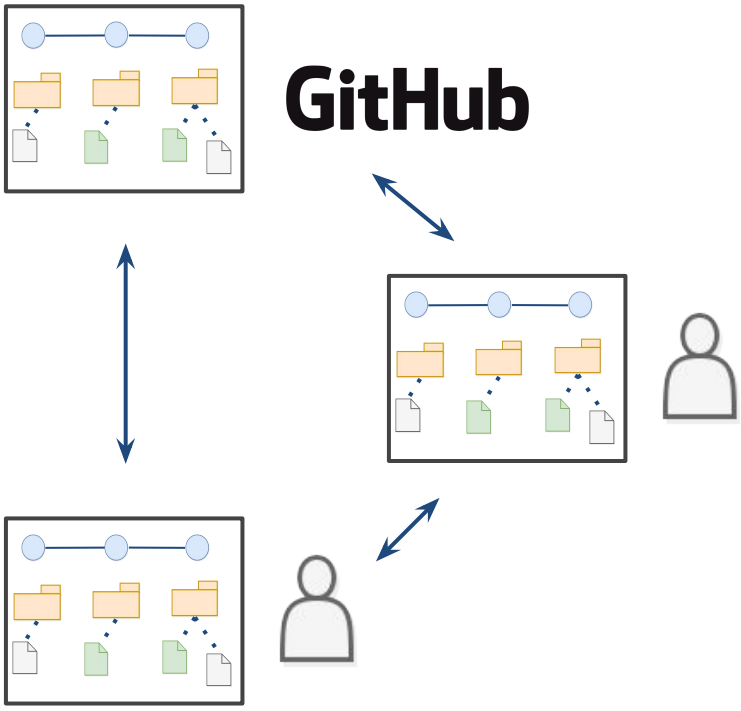
Centralized version control

SVN

Central server



Distributed version control



What does version control do?

Determine *what* changed, *who* changed it and *why*

Organized way of collaborating in code

Way of presenting code

Allow navigating in history of files

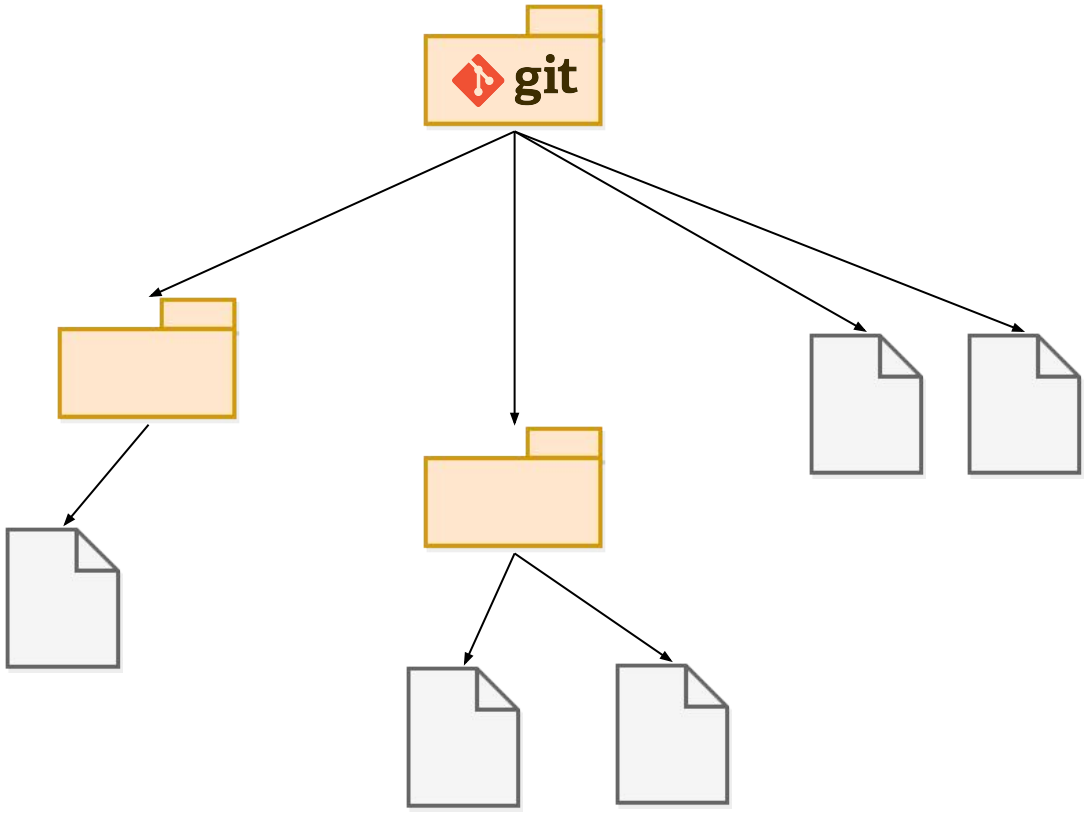
Why don’t everyone use it?

Learning threshold

Tricky to manage data

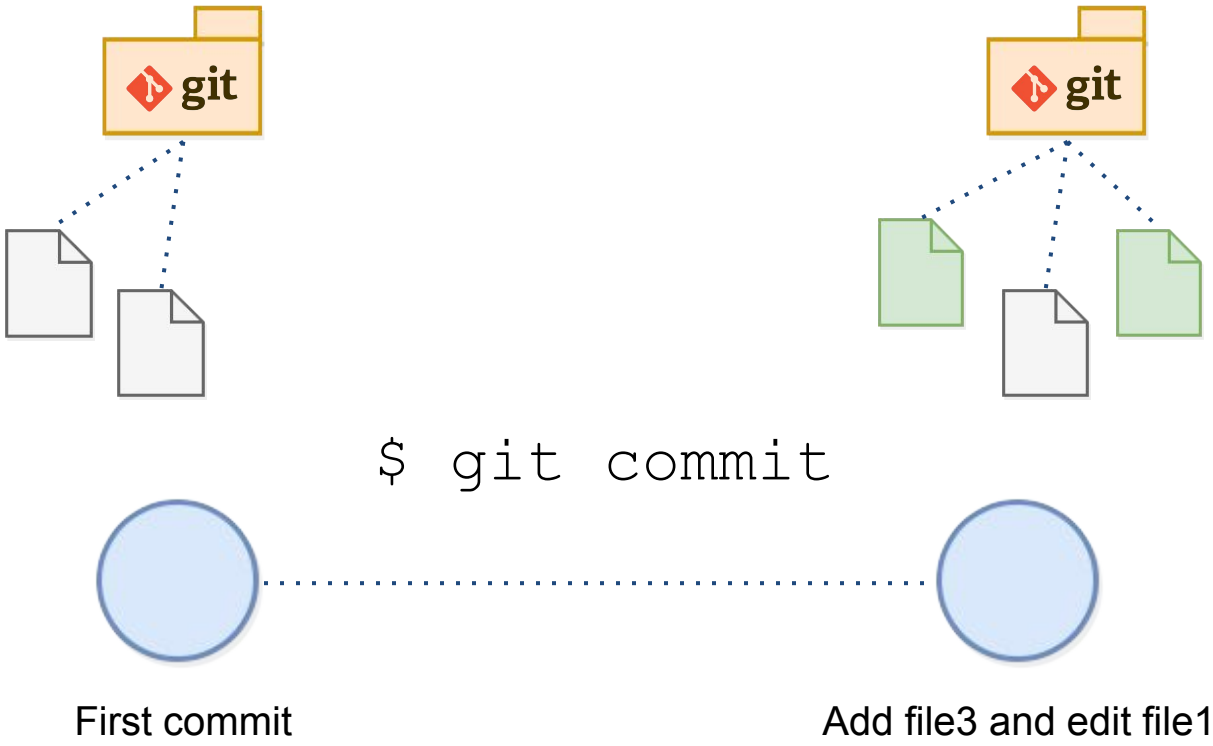
Let’s dive in

The ‘repository’ and the ‘file tree’



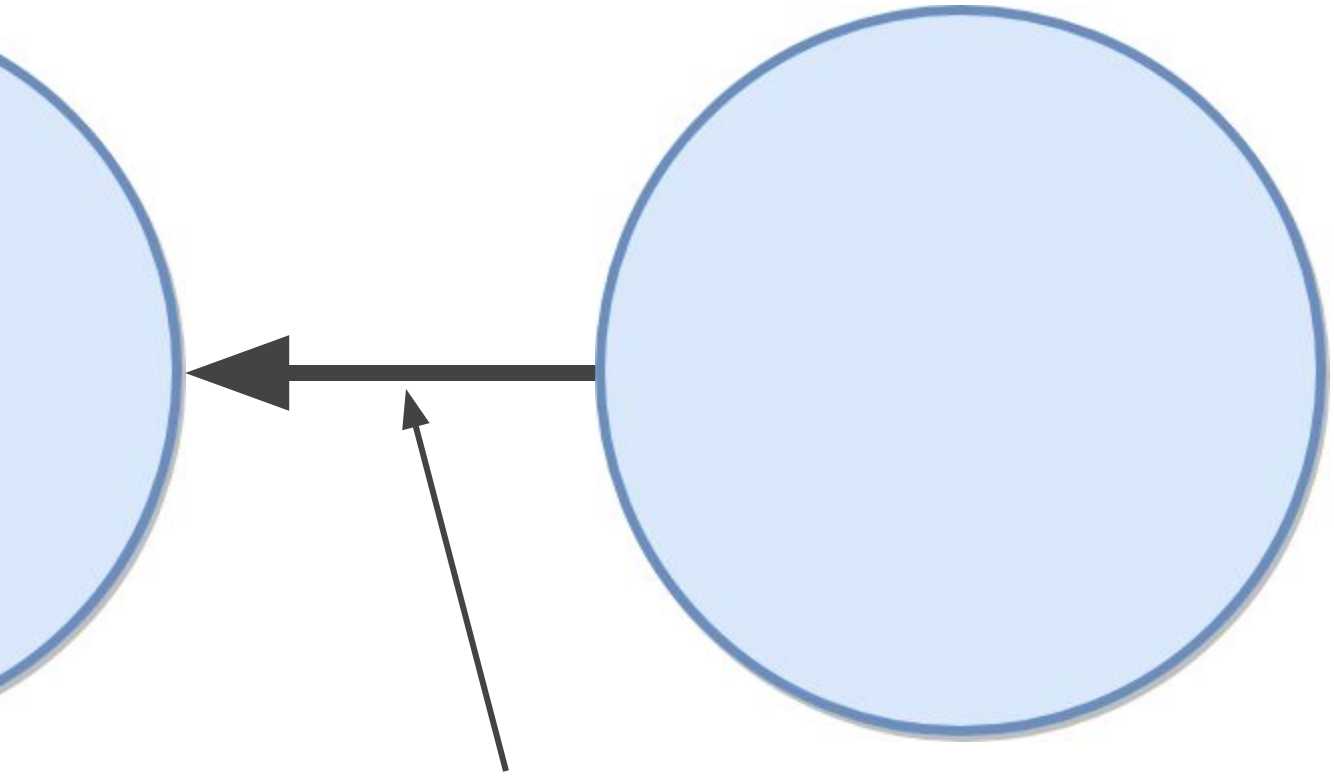
The ‘repository’ keeps track of changes in the ‘file tree’

What is a commit?



A snapshot of particular state in the file tree

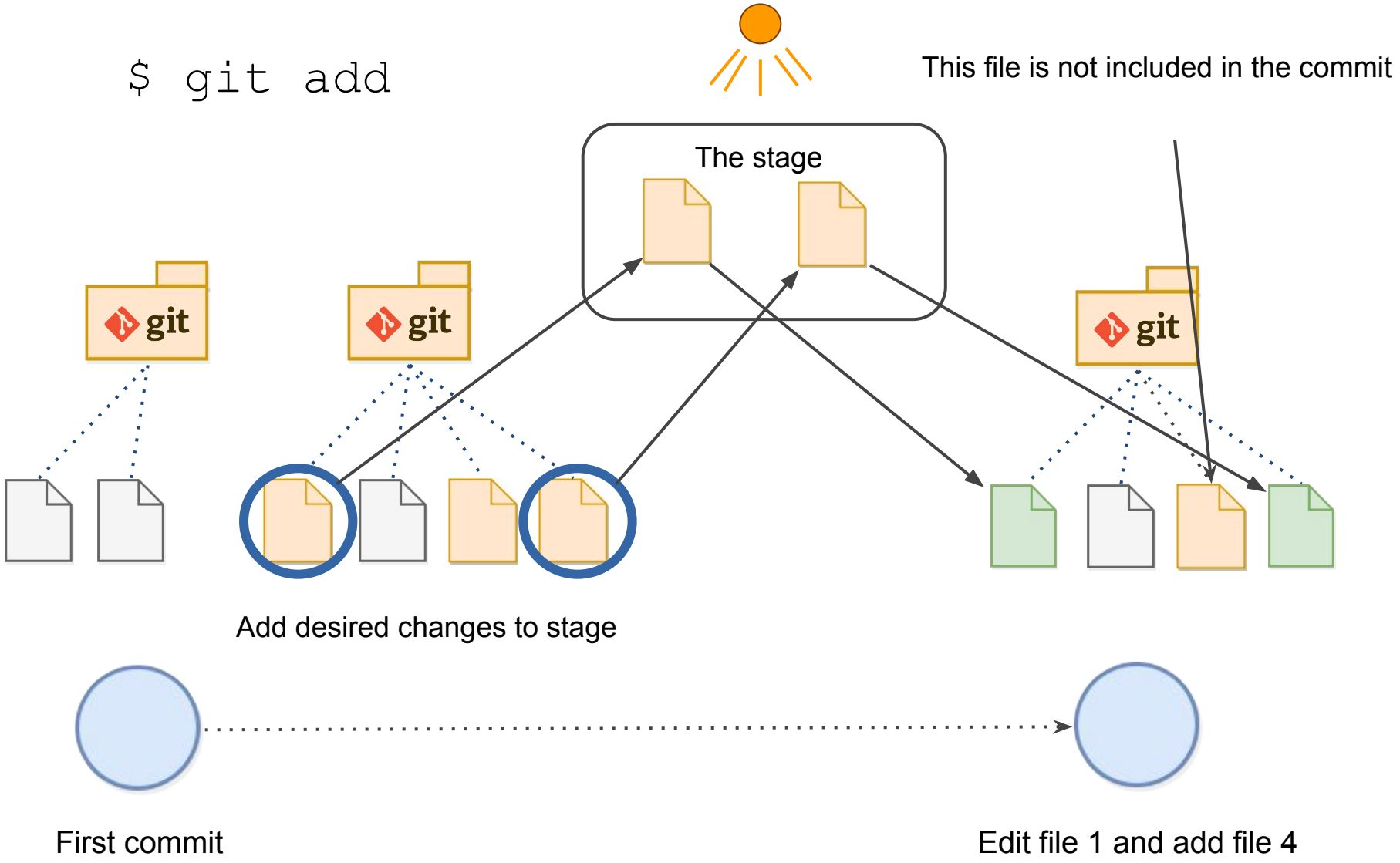
What is a commit?



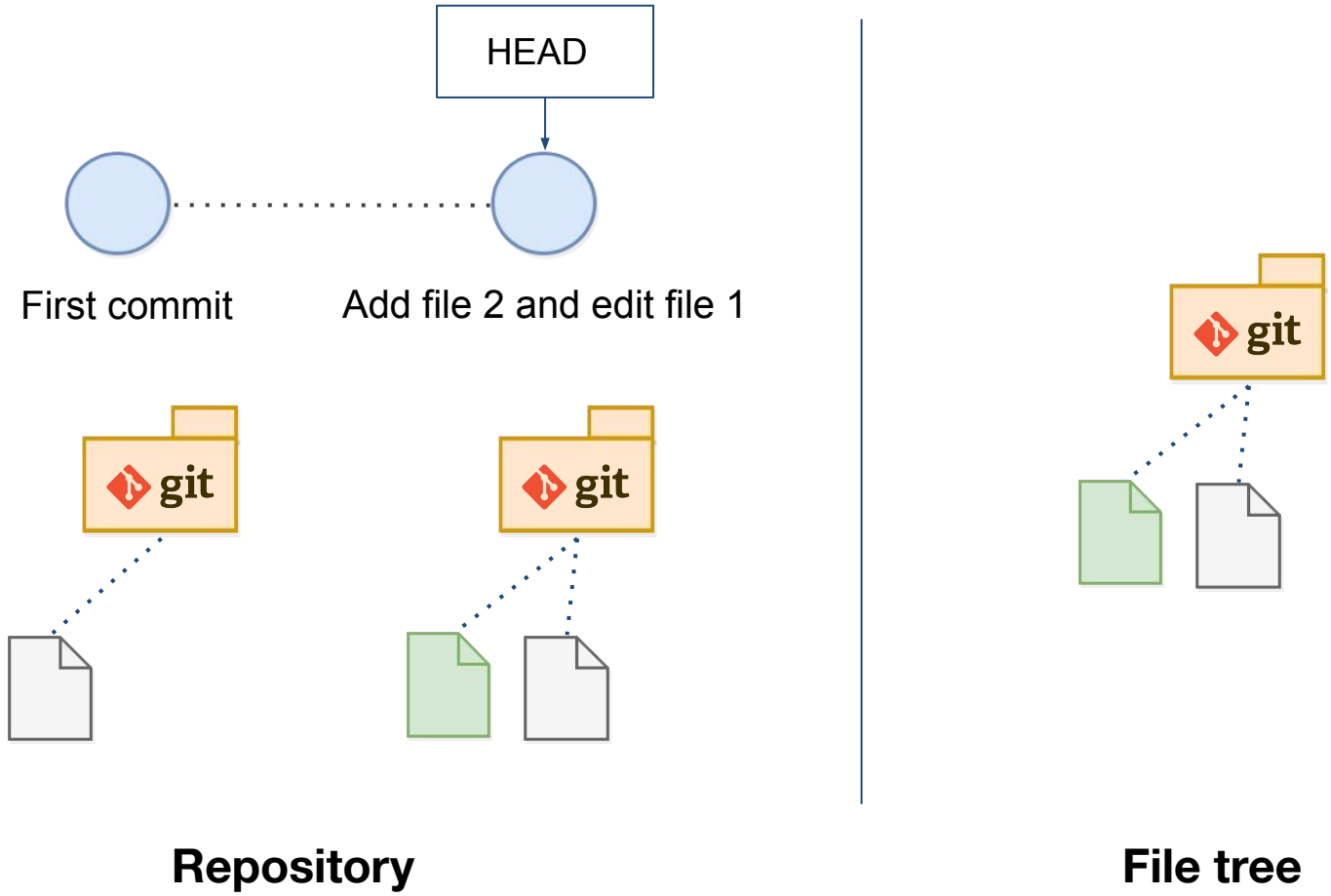
ID
Message
File changes
Author
Date

Link to ‘parent’

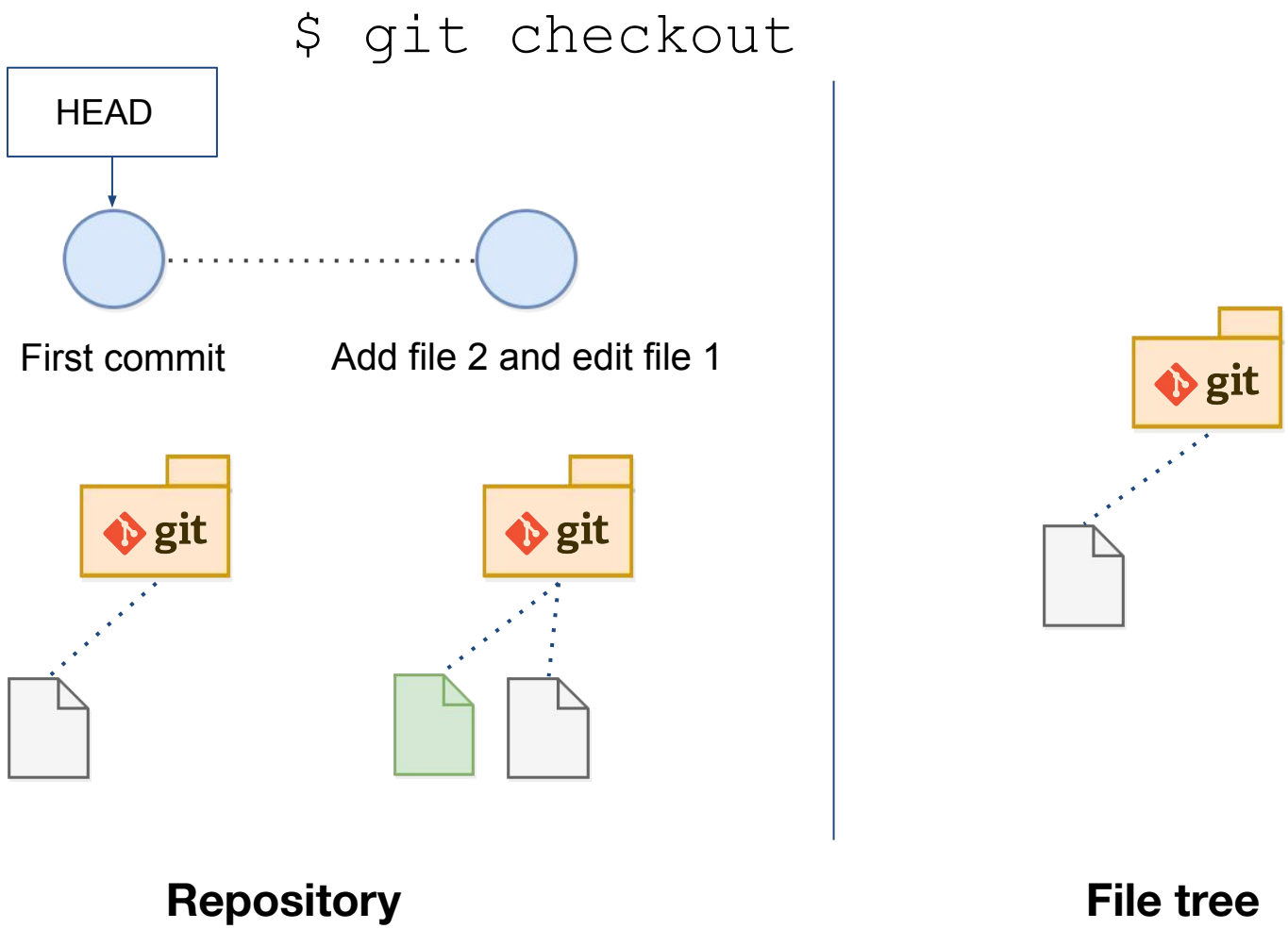
What is the ‘stage’?



Navigating the history and the HEAD



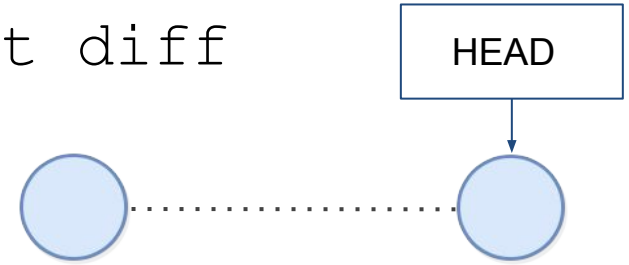
Navigating the history and the HEAD



By moving the HEAD, the file tree changes

Comparing changes

```
$ git diff
```



Differences between commits or branches can be easily visualized

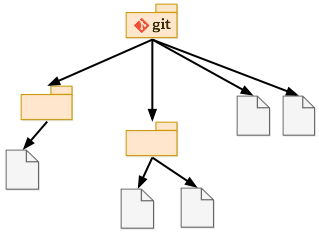
Diff

37 matching.R → algorithm.R — View

```
...  ...  @@ -1,6 +1,6 @@
 1    1
 2    2
 3       - update_character_stats <- function(session, dict, character_stats, cur_ind) {
      3  + update_character_stats <- function(session, dict, character_stats, cur_ind, debug=T) {
 4    4
 5    5       input_english <- session$input$english
 6    6       input_pinying <- session$input$pinying
      @@ -9,7 +9,8 @@ update_character_stats <- function(session, dict, character_stats, cur_ind) {
 9    9           dict[[cur_ind]],
10   10           input_pinying,
11   11           input_english,
12       -       type=session$input$practice_type)
     12  +       type=session$input$practice_type,
     13  +       debug=debug)
13   14
```

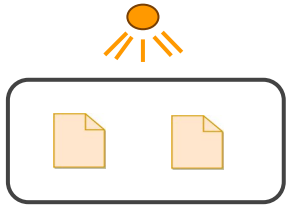
Recap

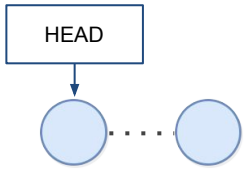
Repository and file tree



Commit

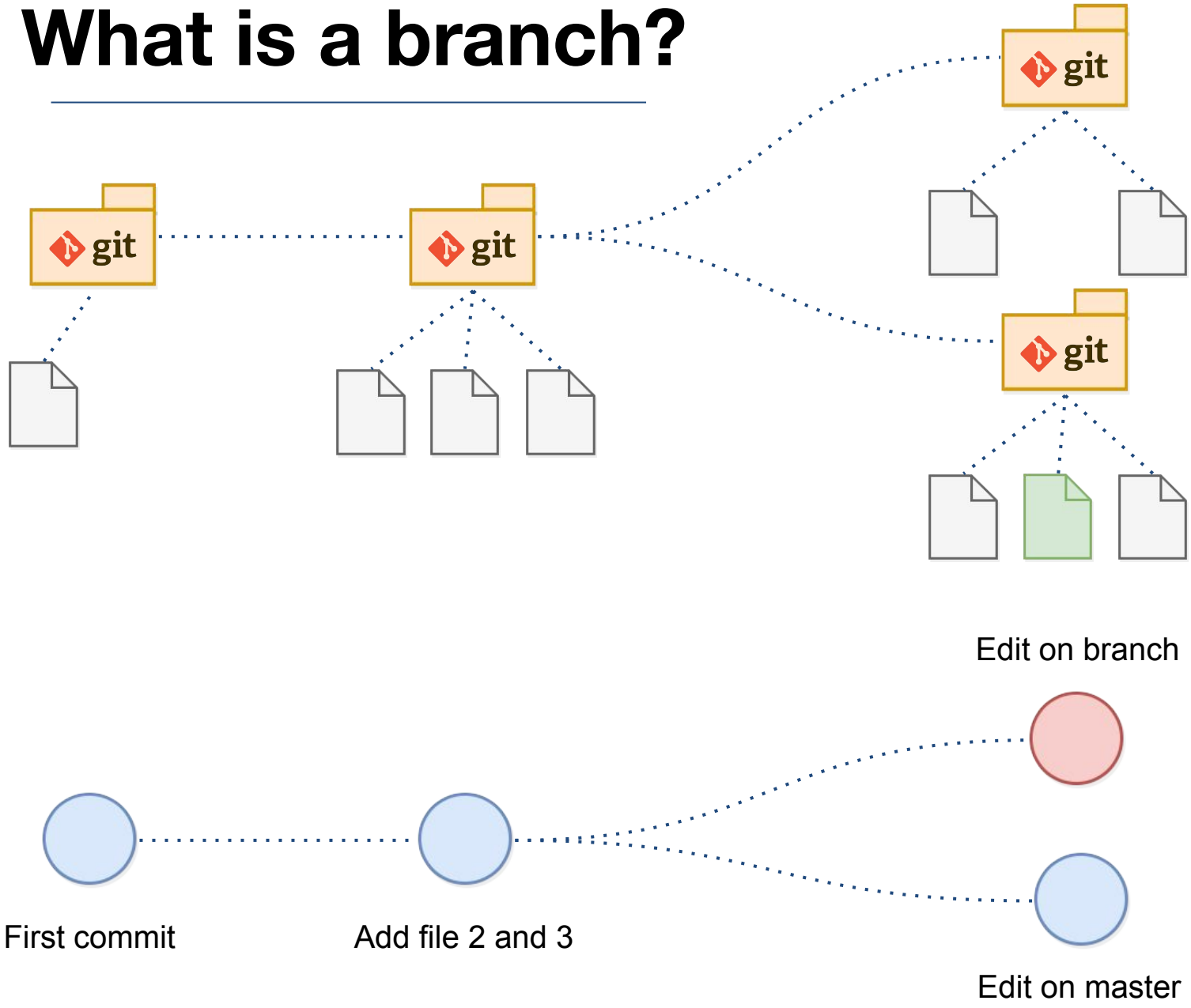
Stage



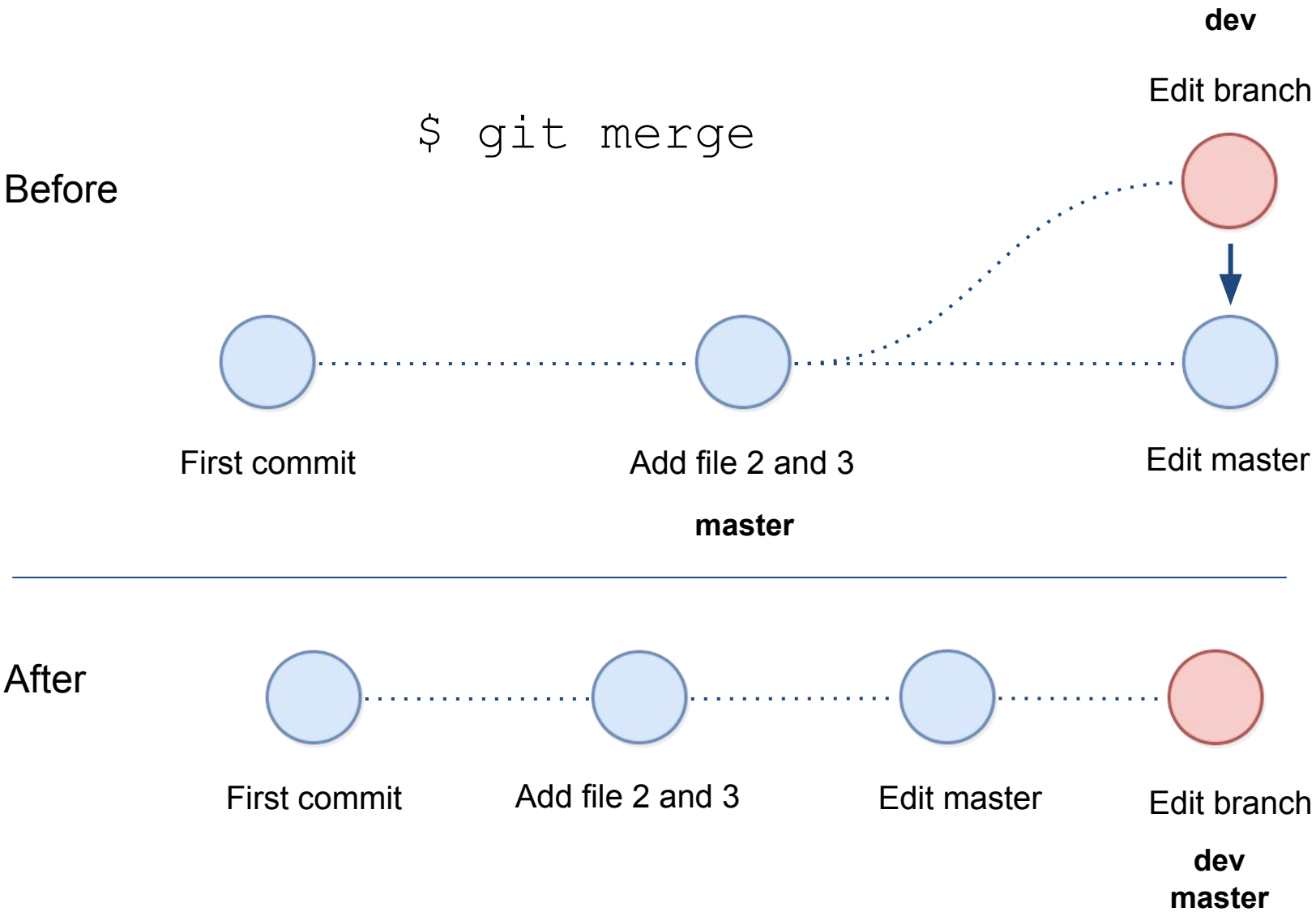


The HEAD

What is a branch?

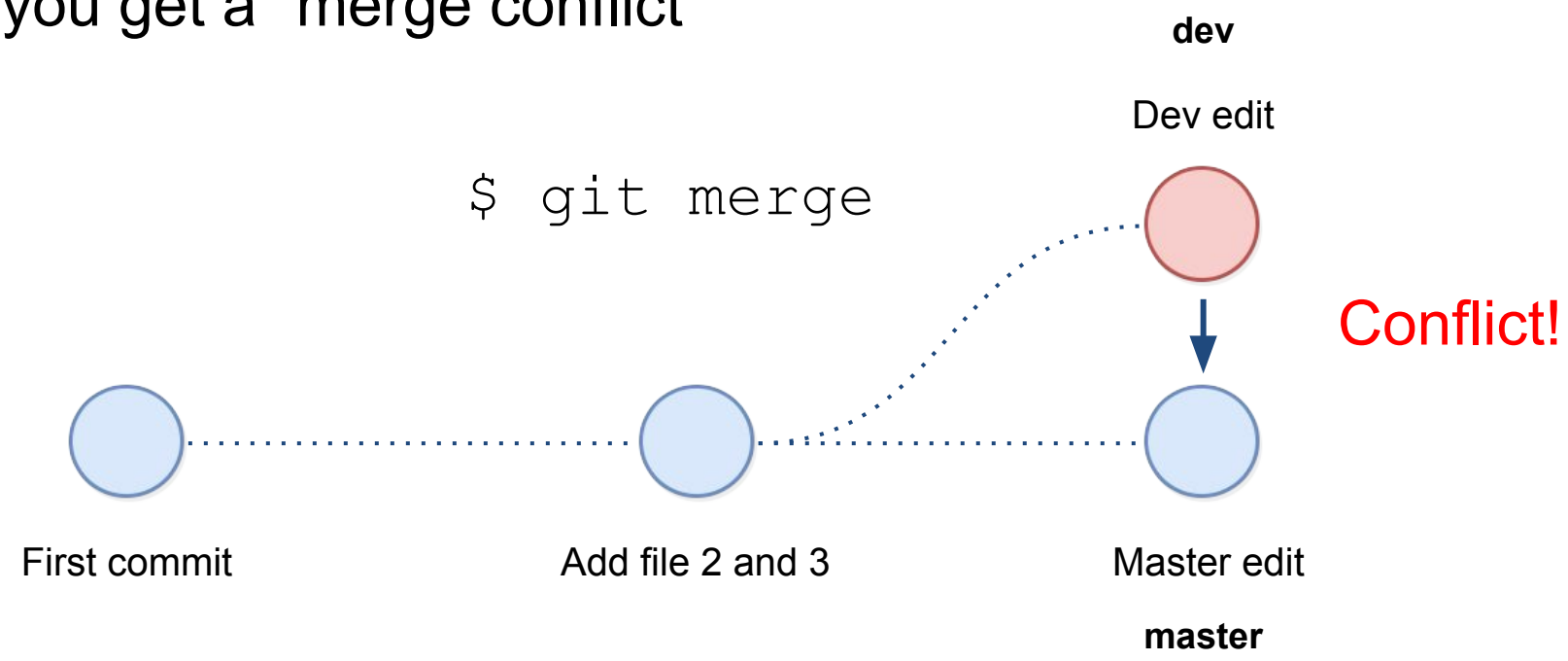


Merge branches



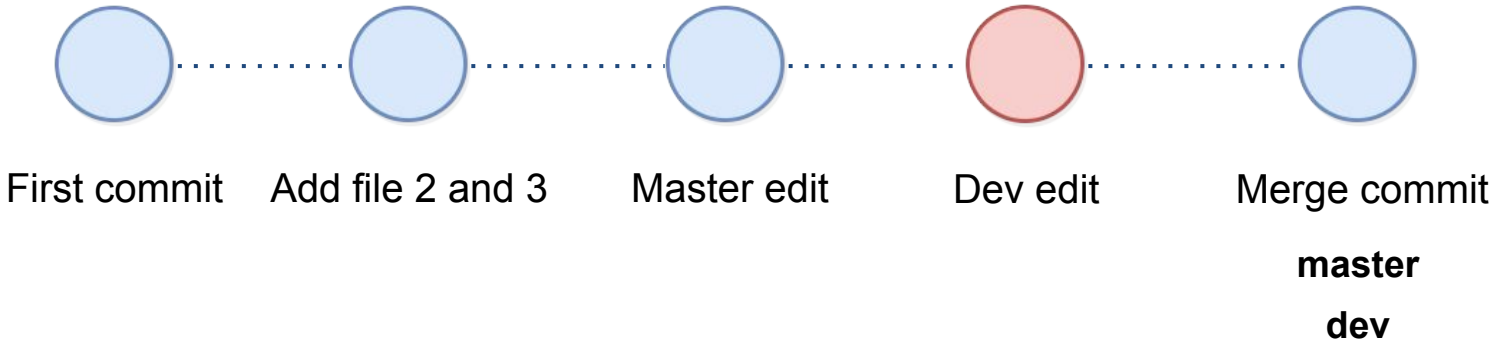
Merge conflicts

If merging conflicting changes,
you get a "merge conflict"



Git stores both changes - You decide what to keep!

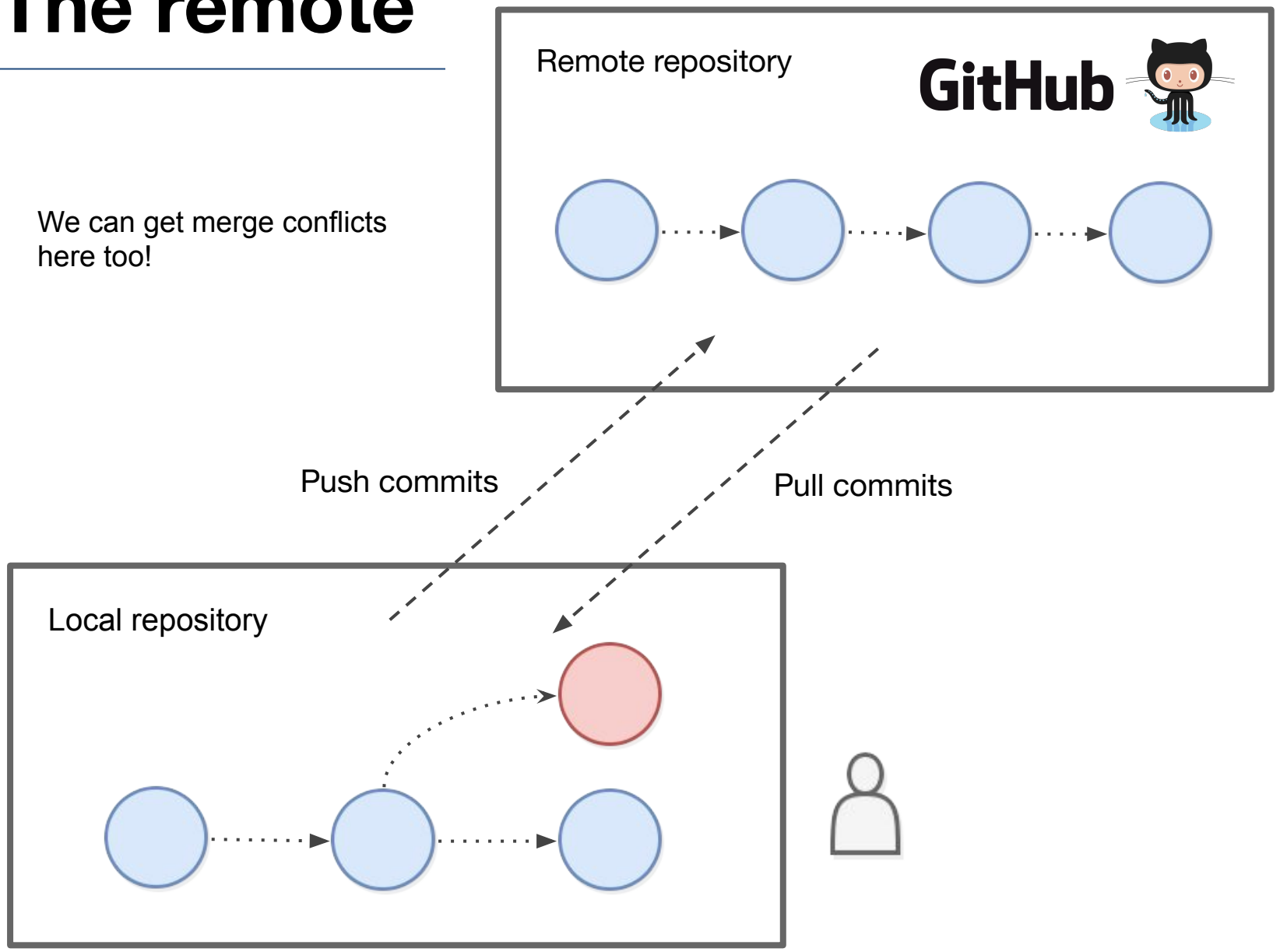
Resolved merge conflict



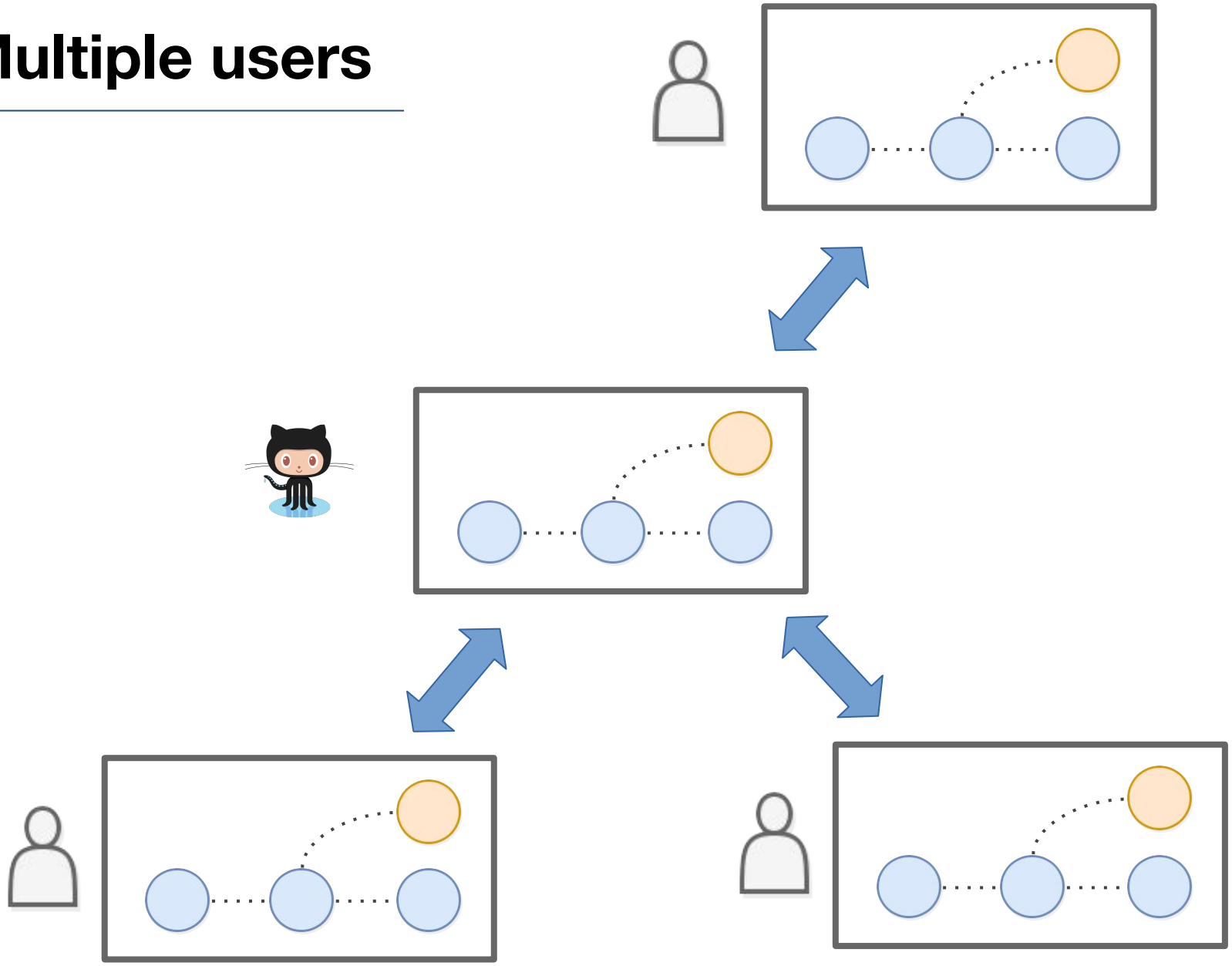
After resolving all commits are there, followed by a merge commit

The remote

We can get merge conflicts here too!

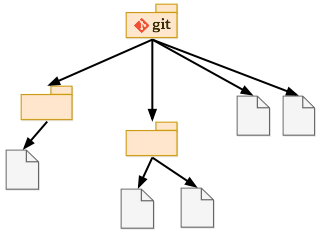


Multiple users



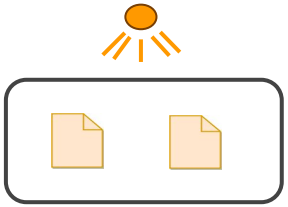
Recap

Repository and file tree

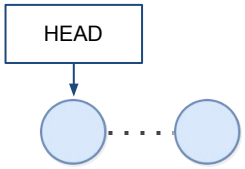


Commit

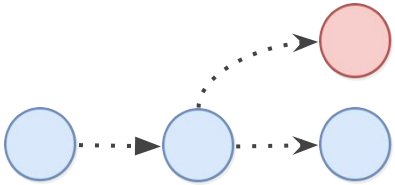
Stage



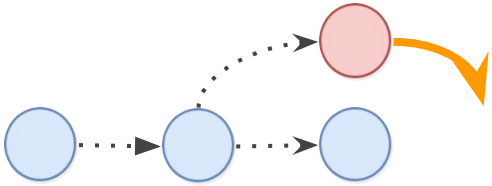
The HEAD



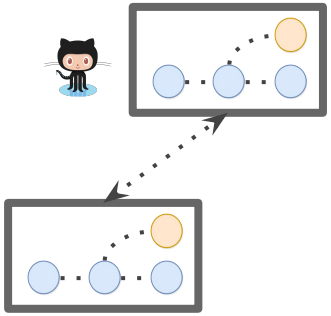
Branch



Merging



Remote

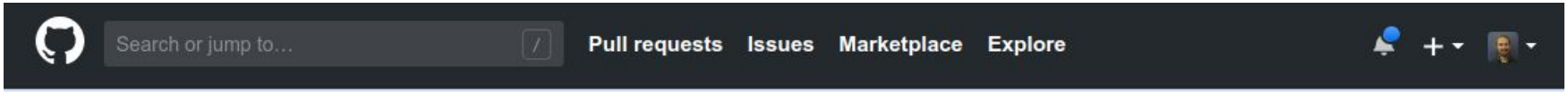


GitHub

Social platform for code

Common way to make code public

Allows interacting with other peoples code



Your repository details have been saved.

ComputationalProteomics / **NormalyzerDE**

Unwatch 2 | Unstar 1 | Fork 0

Code | Issues 5 | Pull requests 0 | Projects 0 | Wiki | Insights | Settings

Tools for normalization, evaluation of outliers, technical biases and batch effects and differential expression analysis. Edit

Manage topics

343 commits | **3** branches | **12** releases | **1** contributor

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

Jakob37 Update documentation — Latest commit 8489c4d 3 hours ago

R	Update documentation	3 hours ago
data	Update documentation	3 hours ago
inst	Omit vignette PDF instance and bump version	19 days ago
man	Update documentation	3 hours ago
tests	Minor fix where function wasn't properly used. Update tests to do glo…	10 days ago
vignettes	Another vignette warning try	19 days ago
.gitignore	Omit vignette PDF instance and bump version	19 days ago
DESCRIPTION	Update documentation	3 hours ago
NAMESPACE	Pre-Bioconductor updates. Edits to documentations, code reshuffling a…	20 days ago
NEWS	Reactivate Quantile normalization	9 days ago
README.md	Rearranging the README	a month ago

Jakob37 / **LearnChineseWithRShiny**

Watch 0 | Star 0 | Fork 0

Code | Issues 2 | Pull requests 0 | Projects 0 | Wiki | Insights | Settings

Branch: **master**

Commits on Sep 9, 2018

Basic table interaction running using handsontable
Jakob37 committed 16 days ago — `c9be19d`

Show data as table
Jakob37 committed 16 days ago — `147f783`

Database interactions extended. Visualizing different parts of charac…
Jakob37 committed 16 days ago — `ffc8cfe`

Commits on Sep 8, 2018

Partially running DB system for word entries
Jakob37 committed 17 days ago — `5357193`

Primitive MySQL backend setup
Jakob37 committed 17 days ago — `a94878f`

Draft data classes and non functional group/character management inte…
Jakob37 committed 17 days ago — `2200b0d`

Basic dummy group iteration functional
Jakob37 committed 17 days ago — `c776f0c`

Updated working algorithm

master

Browse files

Jakob37 committed on Aug 2 — 1 parent 85aa4de — commit c9a8d66b39c8a247ffd9dc87b7ad512724761637

Showing **4 changed files** with **57 additions** and **32 deletions**.

Unified | Split

37 matching.R → algorithm.R — View

```
...  ...  @@ -1,6 +1,6 @@
 1    1
 2    2
 3       - update_character_stats <- function(session, dict, character_stats, cur_ind) {
      3  + update_character_stats <- function(session, dict, character_stats, cur_ind, debug=T) {
 4    4
 5    5       input_english <- session$input$english
 6    6       input_pinying <- session$input$pinying
      @@ -9,7 +9,8 @@ update_character_stats <- function(session, dict, character_stats, cur_ind) {
 9    9           dict[[cur_ind]],
10   10           input_pinying,
11   11           input_english,
12      -        type=session$input$practice_type)
      12 +        type=session$input$practice_type,
      13 +        debug=debug)
13   14
14   15       if (correct) {
15   16           character_stats[[cur_ind]]$right <- character_stats[[cur_ind]]$right + 1
      @@ -18,10 +19,15 @@ update_character_stats <- function(session, dict, character_stats, cur_ind) {
18   19           character_stats[[cur_ind]]$wrong <- character_stats[[cur_ind]]$wrong + 1
19   20       }
```

Getting started

How to use it

Use it from terminal:
https://git-scm.com

Or graphical interface:
https://desktop.github.com

How to learn it

Visualizing Git - See the commands in the browser
https://git-school.github.io/visualizing-git

Codecademy - Interactive online tutorial (free):
https://www.codecademy.com/learn/learn-git

DataCamp - Comprehensive online tutorials (not free):
https://www.datacamp.com/courses/introduction-to-git-for-data-science

For deeper understanding:
https://www.sbf5.com/~cduan/technical/git

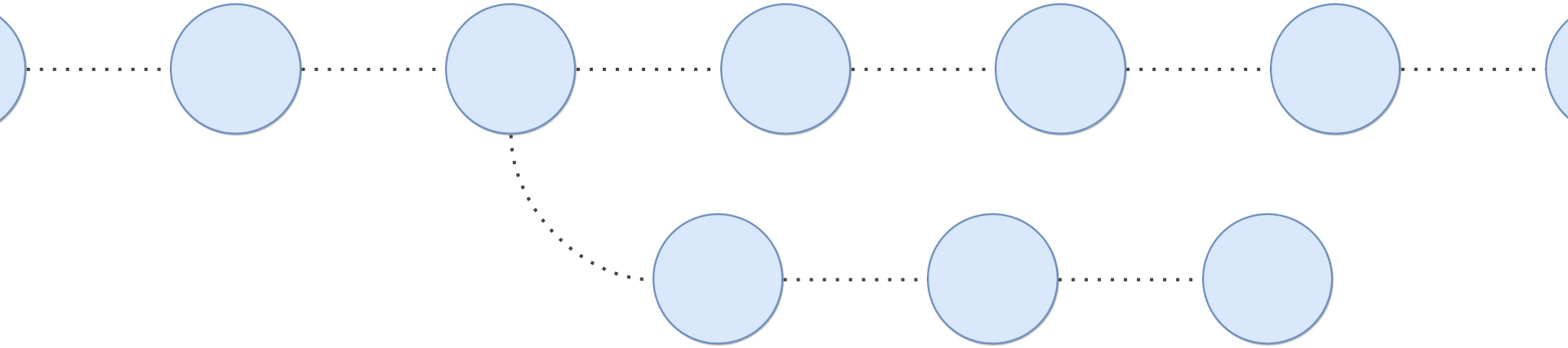
My materials:
http://ponderomatics.com/git.html

Handling large data

Git Large File Storage
https://git-lfs.github.com/

Quilt
https://quiltdata.com/

Thank you for listening!



Jakob Willforss, jakob.willforss@immun.lth.se