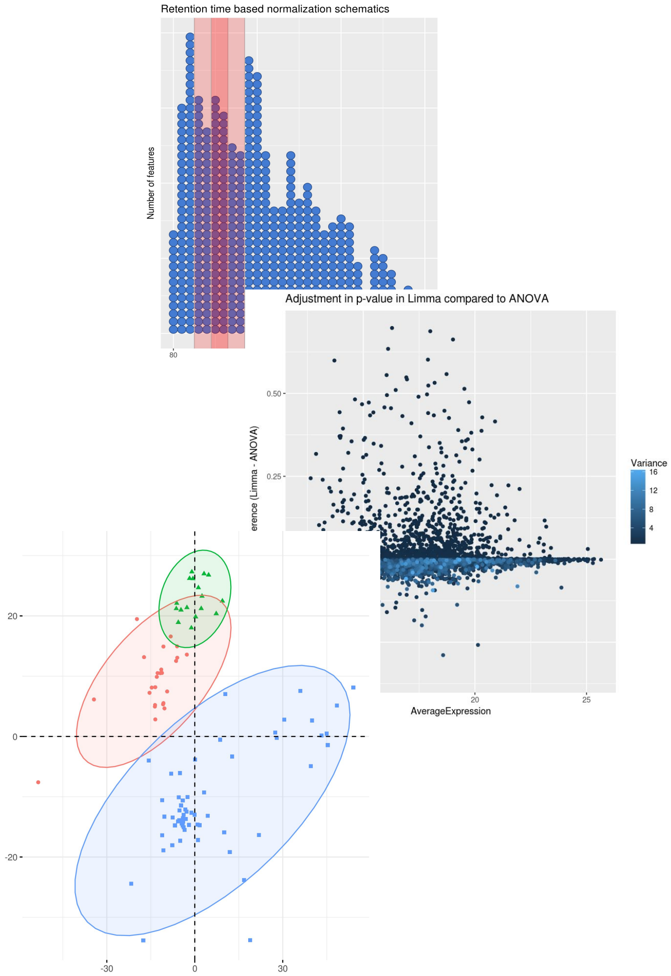
Git and version control in data science

Jakob Willforss, Department of Immunotechnology, LTH

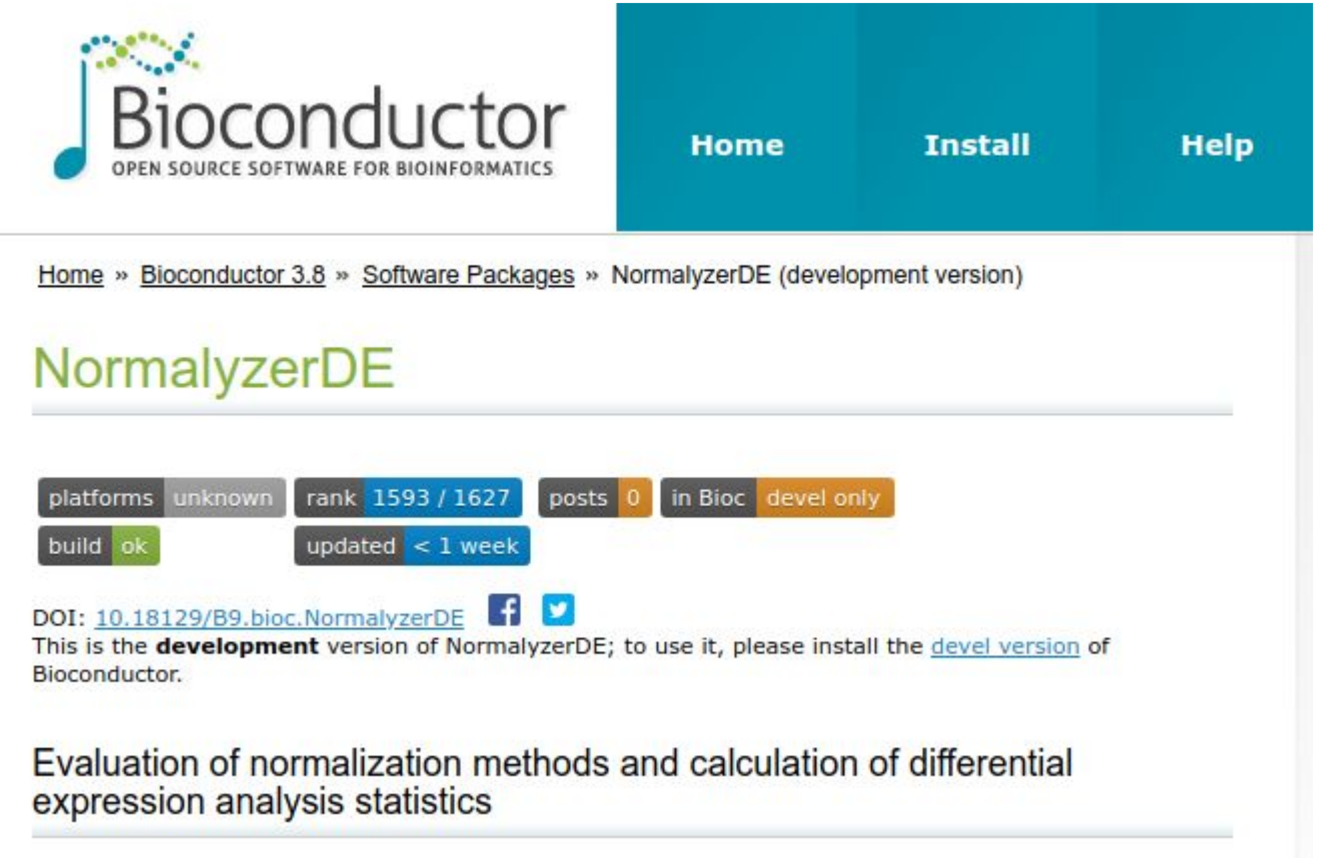
What do I do?



Using statistics, machine learning and R to analyze molecular biology in potato, bull, oat and other

Photos from Unsplashed.com

What do I do?



Home » Bioconductor 3.8 » Software Packages » NormalyzerDE (development version)

NormalyzerDE

platforms unknown | rank 1593 / 1627 | posts 0 | in Bioc devel only
build ok | updated < 1 week

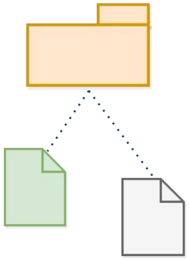
DOI: 10.18129/B9.bioc.NormalyzerDE

This is the **development** version of NormalyzerDE; to use it, please install the devel version of Bioconductor.

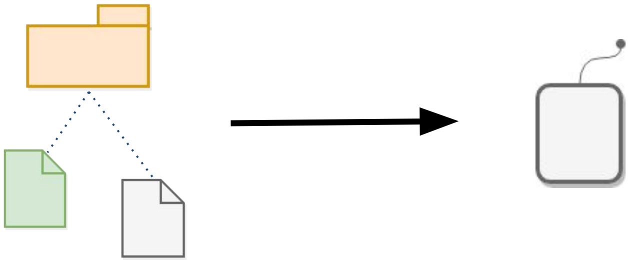
Evaluation of normalization methods and calculation of differential expression analysis statistics

Developing software to help analysis of biomolecular data

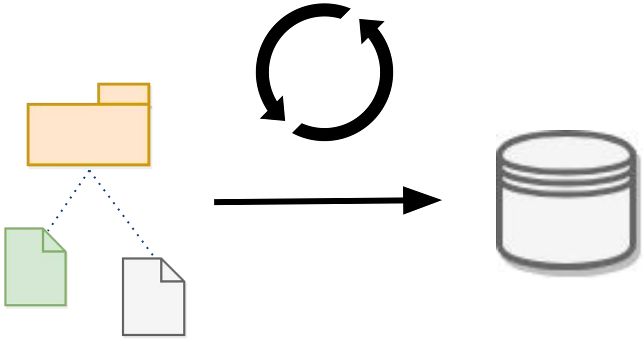
How we manage our files



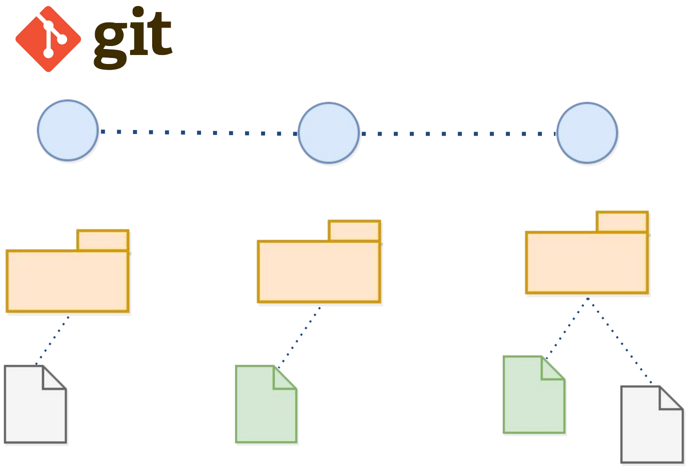
No backup



Manual backup

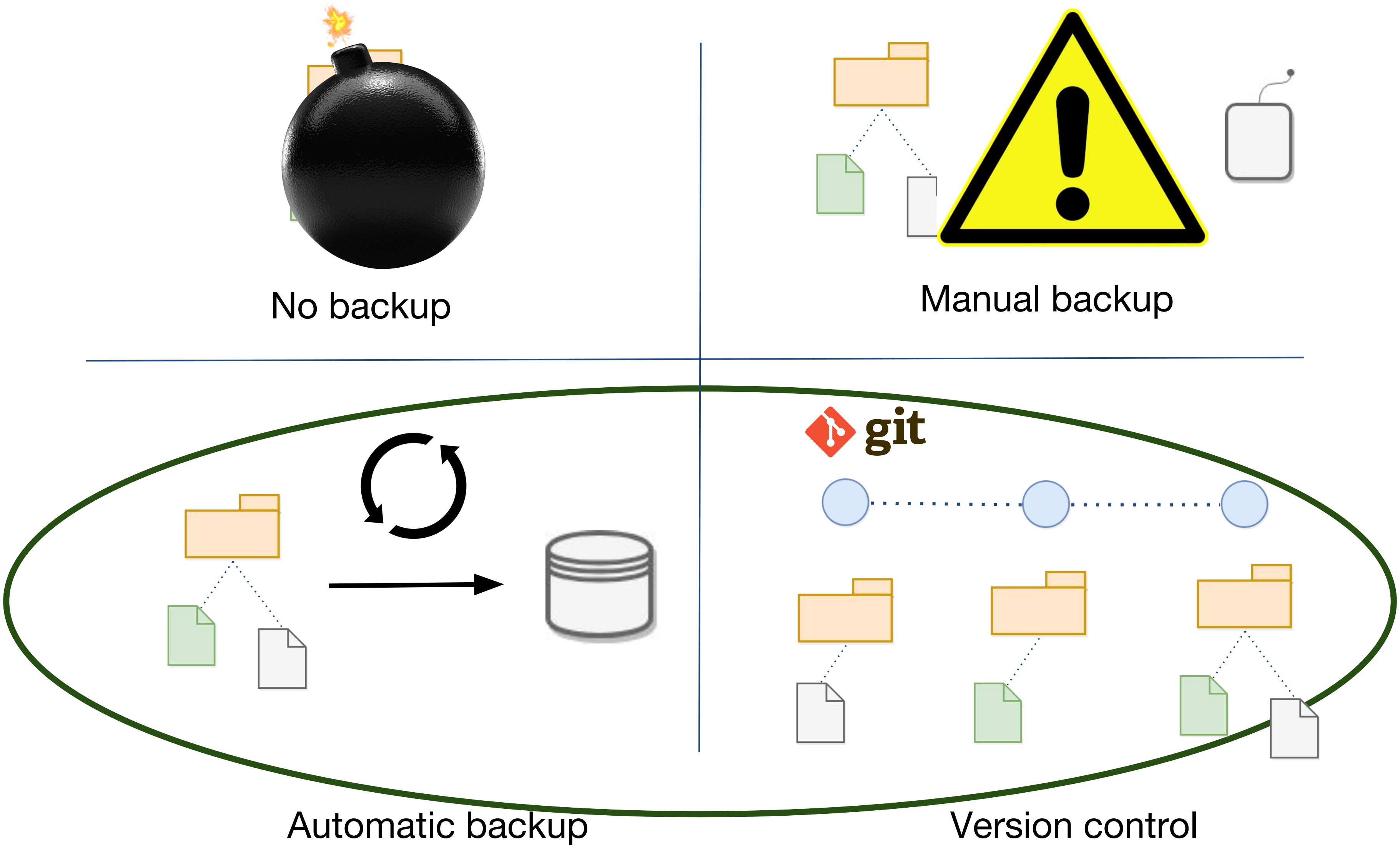


Automatic backup



Version control

How we manage our files



No backup

Manual backup

Automatic backup

Version control

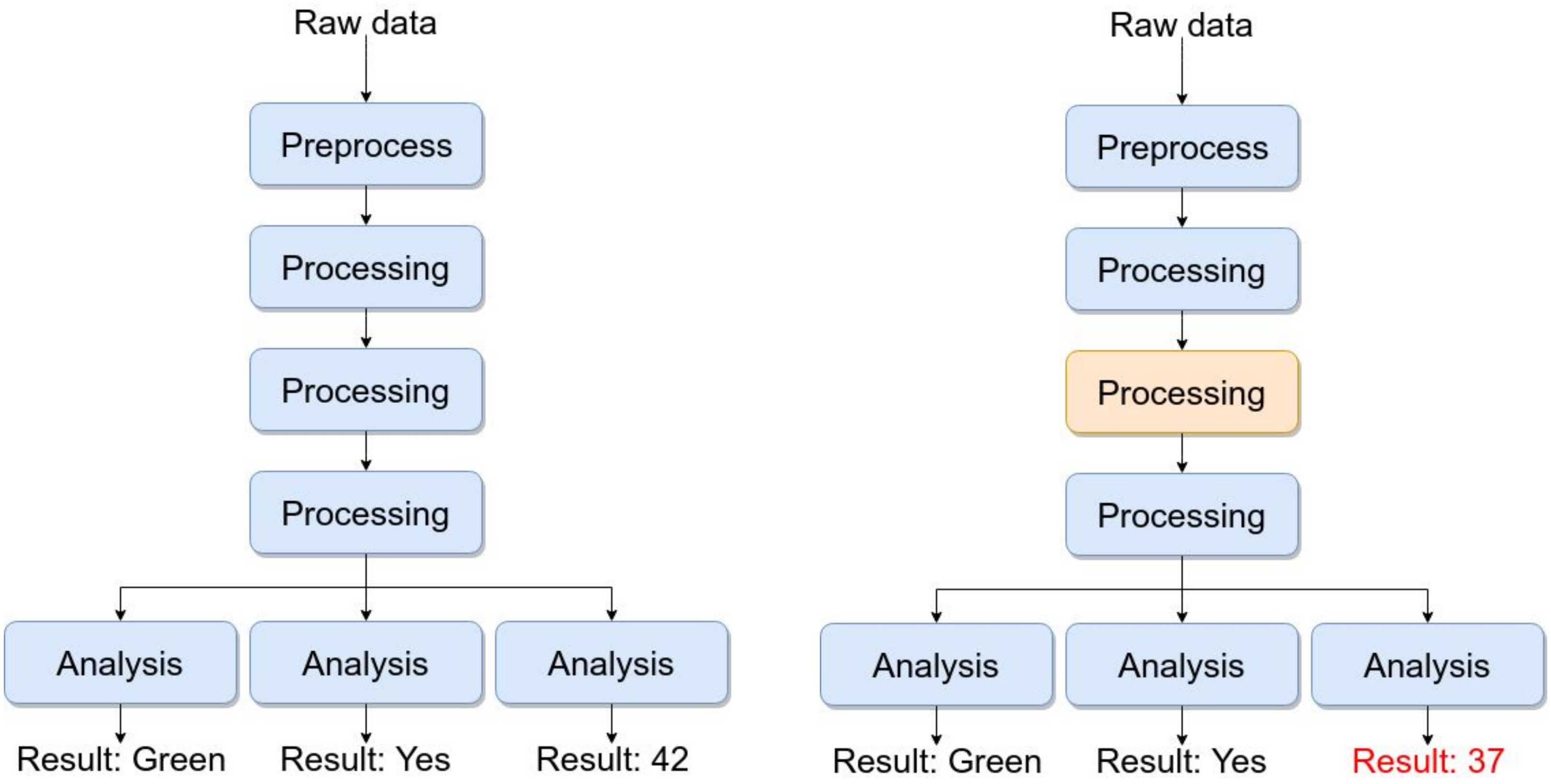
Version control vs. backup

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

git

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

Reproducible research



Raw data

Preprocess

Processing

Processing

Processing

Analysis

Analysis

Analysis

Result: Green

Result: Yes

Result: 42

Raw data

Preprocess

Processing

Processing

Processing

Analysis

Analysis

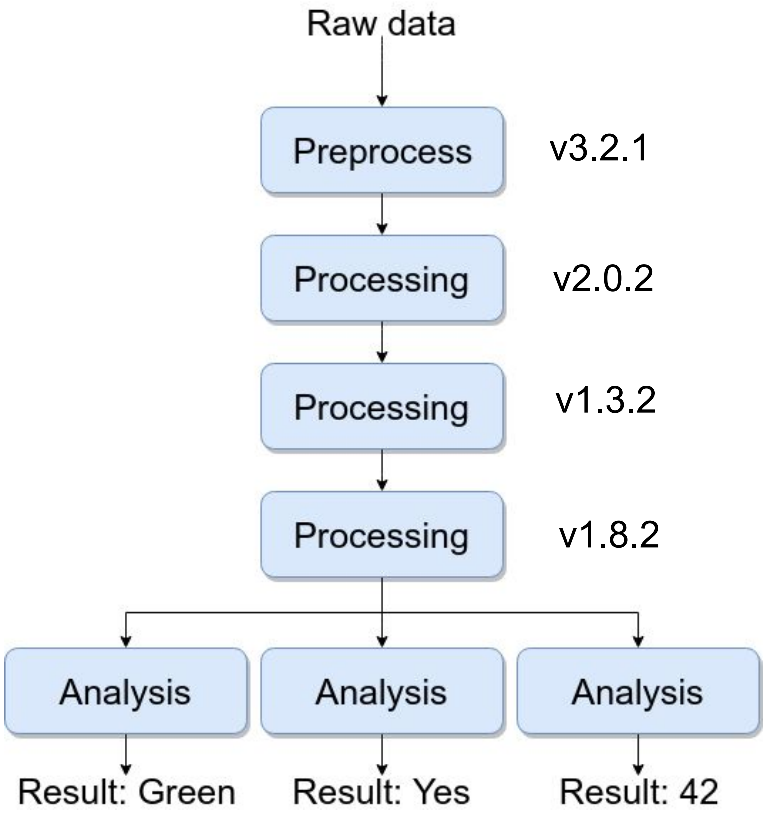
Analysis

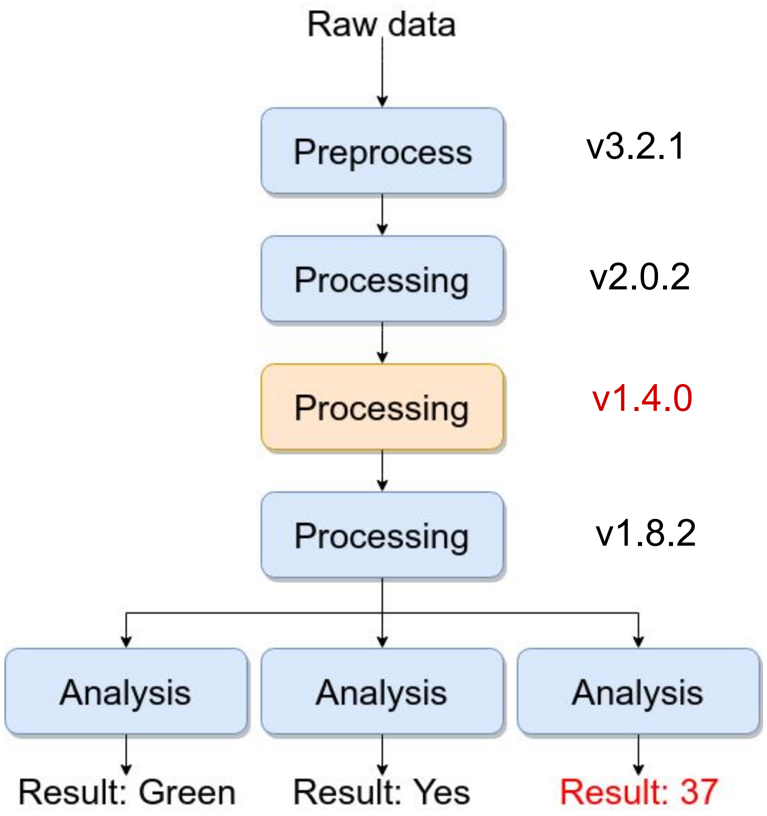
Result: Green

Result: Yes

Result: 37

Reproducible research



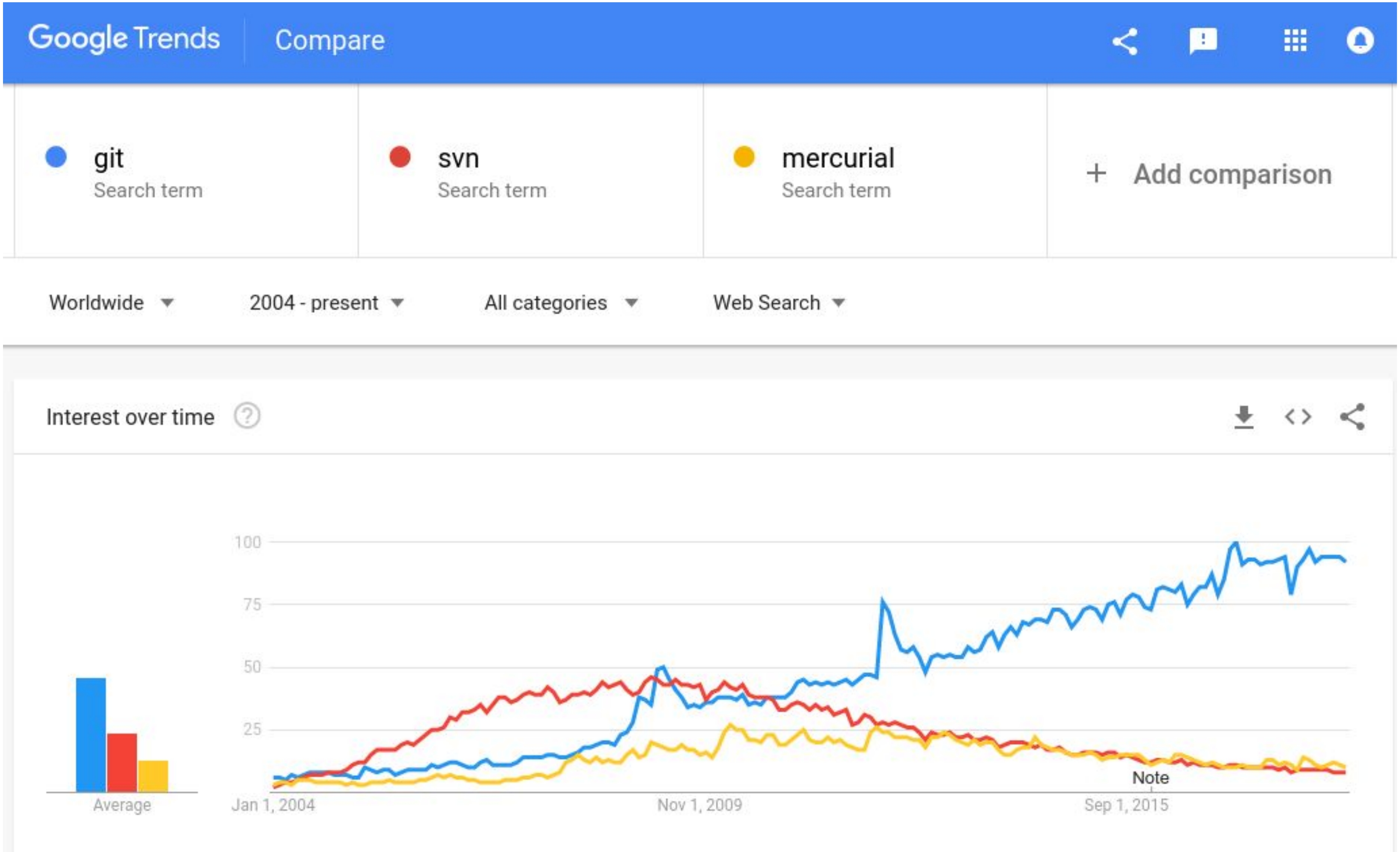


Git

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 ▾

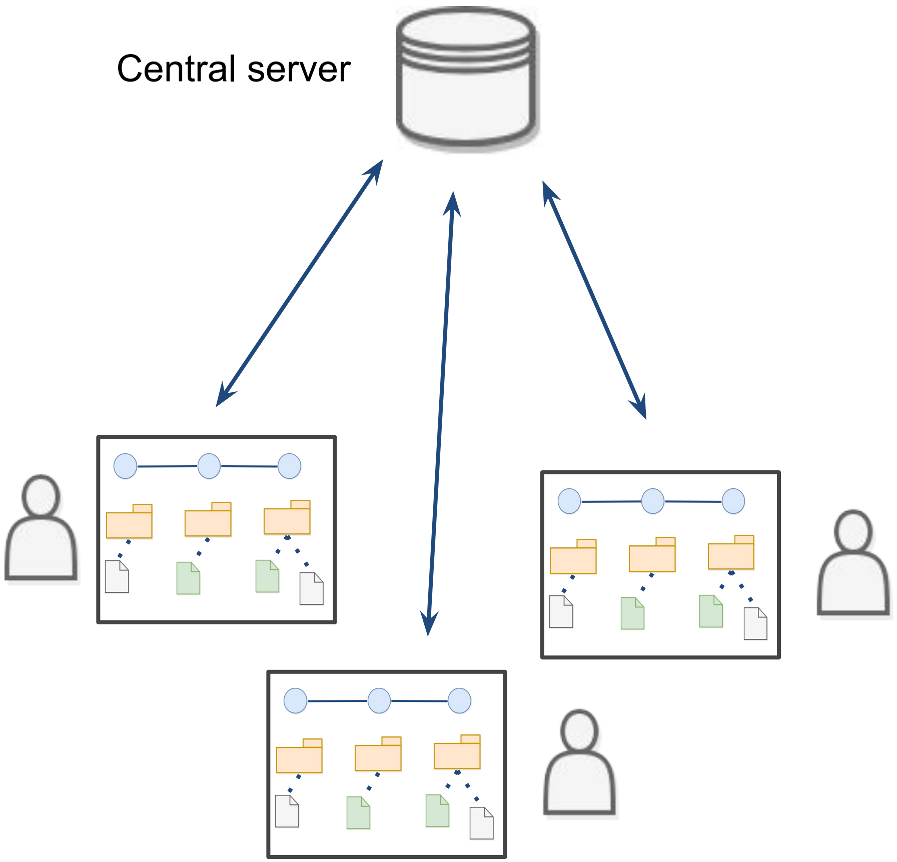
Interest over time

2004 2009 2015 Now

Centralized version control

SVN

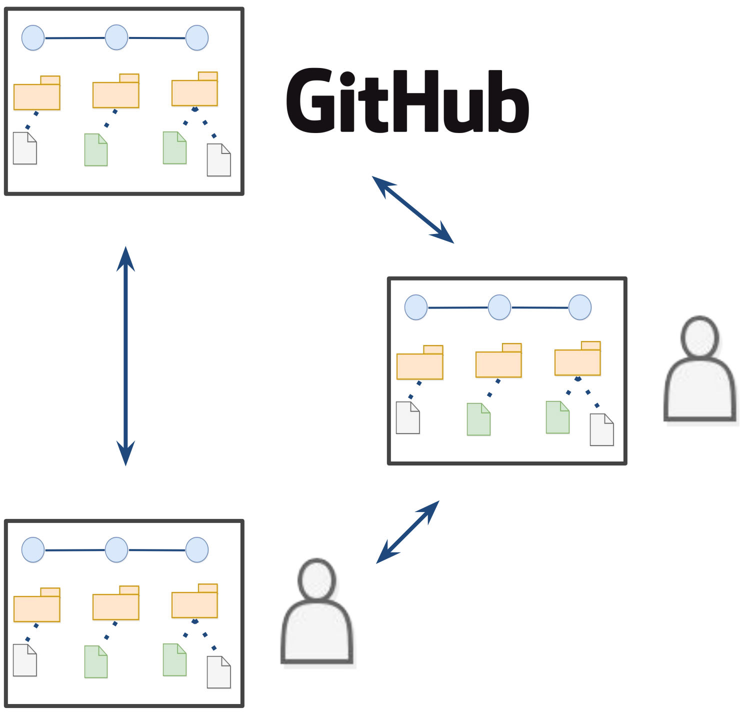
SUBVERSION



Distributed version control

git

mercurial



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

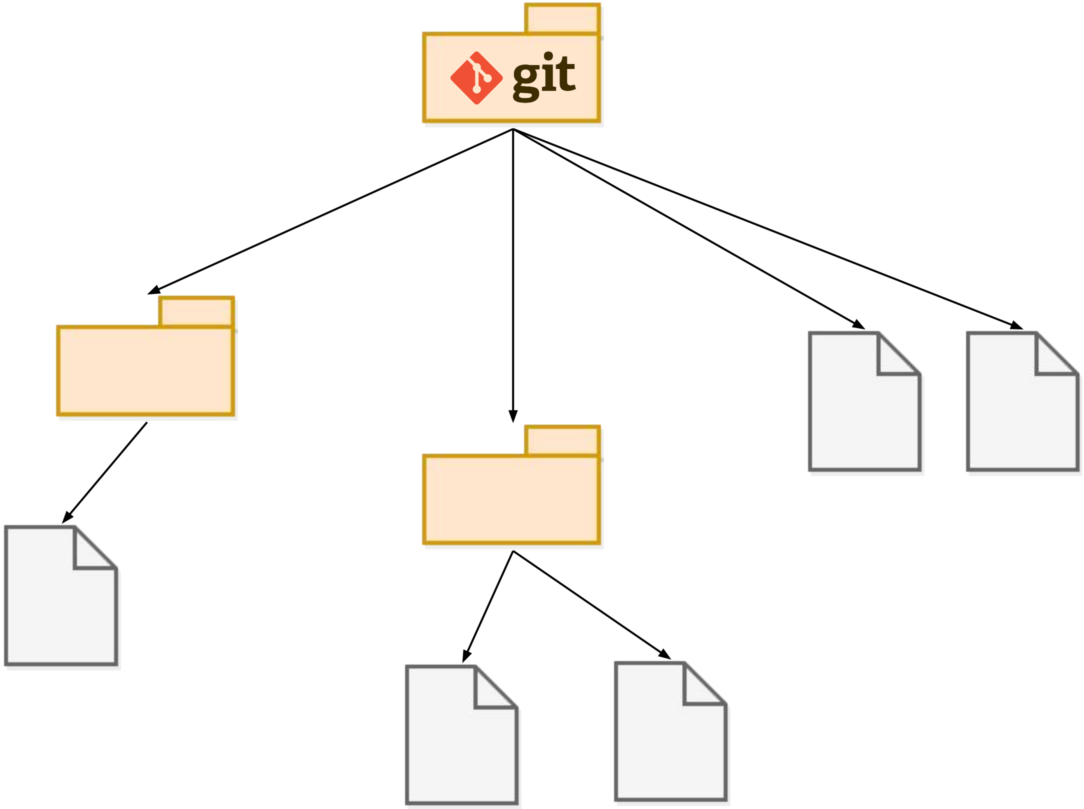
Why don’t everyone use it?

Learning threshold

Tricky to manage big datasets

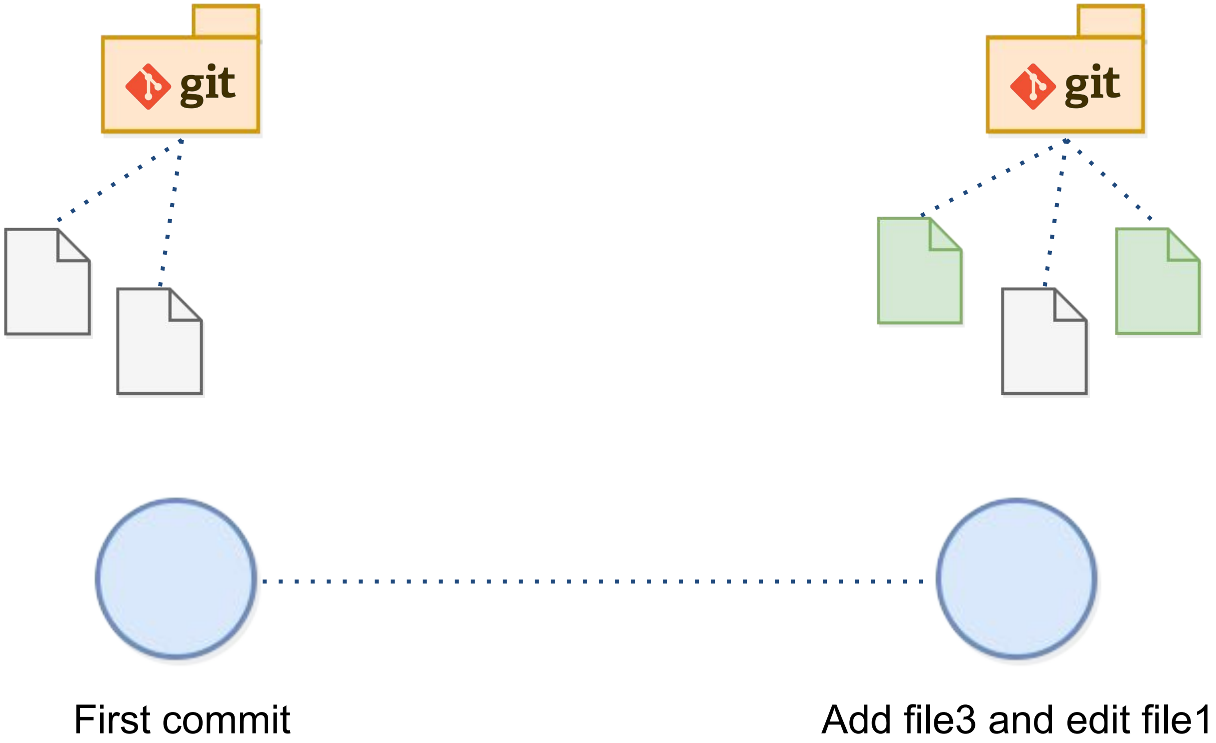
Let’s dive in

The ‘repository’ and the ‘file tree’



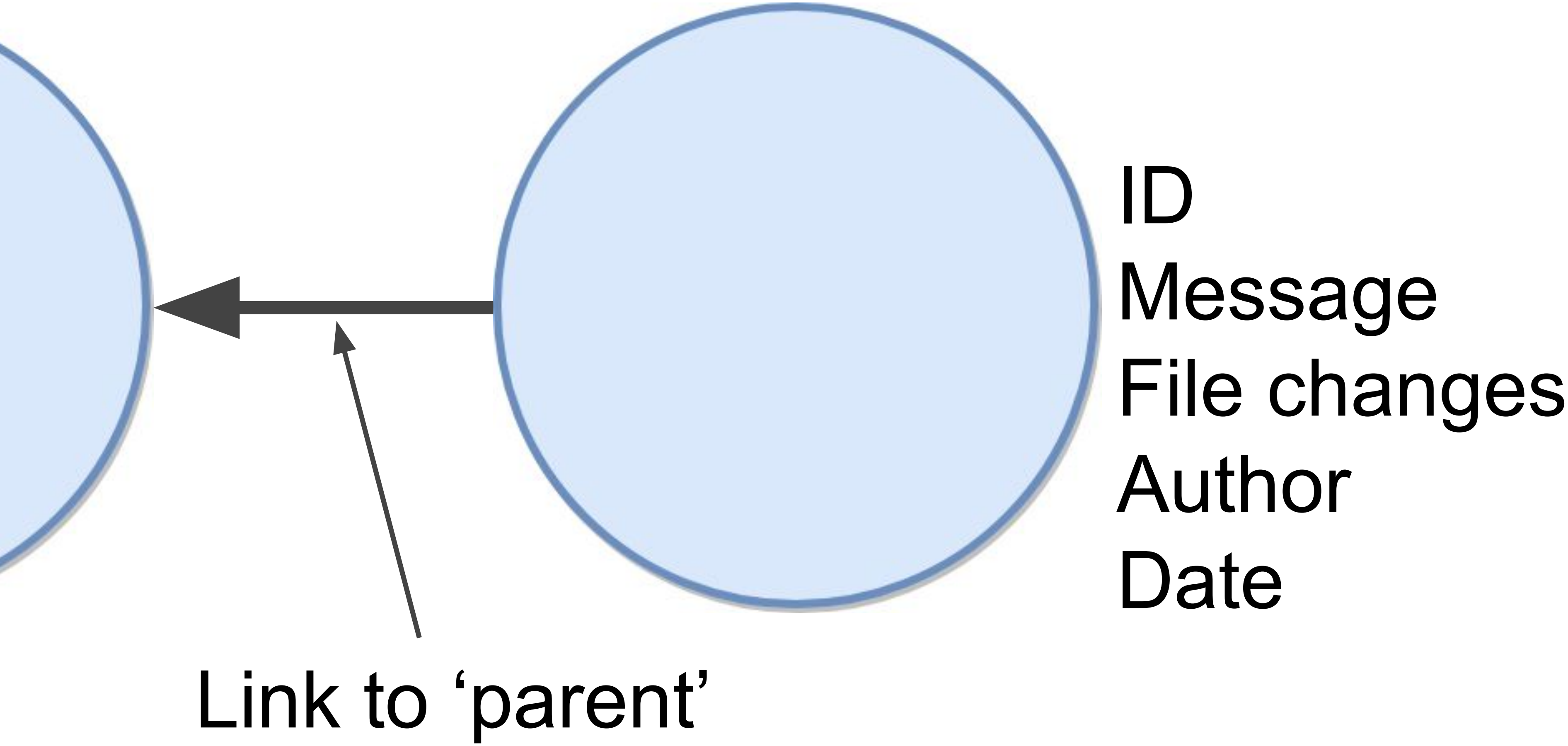
The ‘repository’ keeps track of changes in the ‘file tree’

What is a commit?



A snapshot of particular state in the file tree

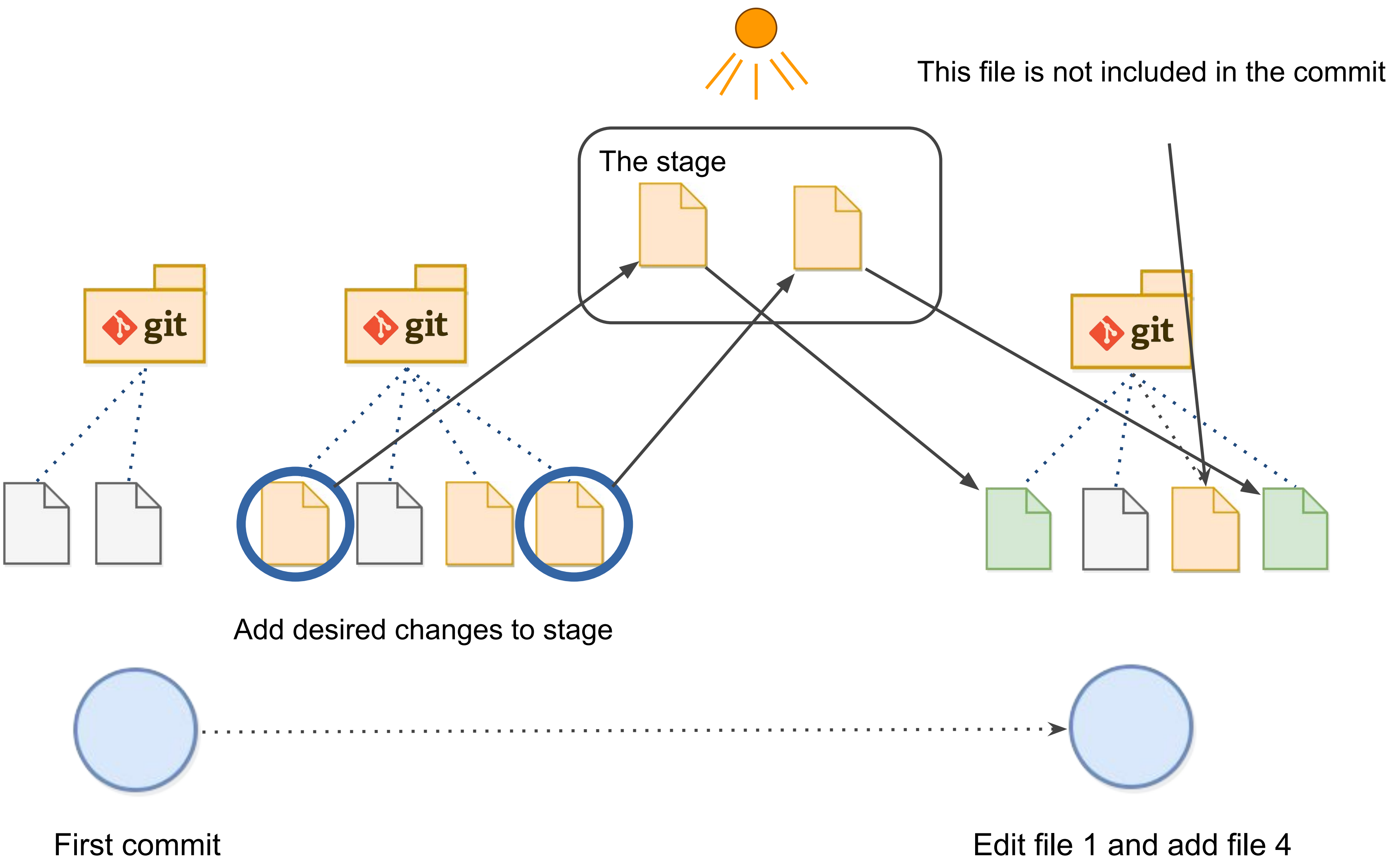
What is a commit?



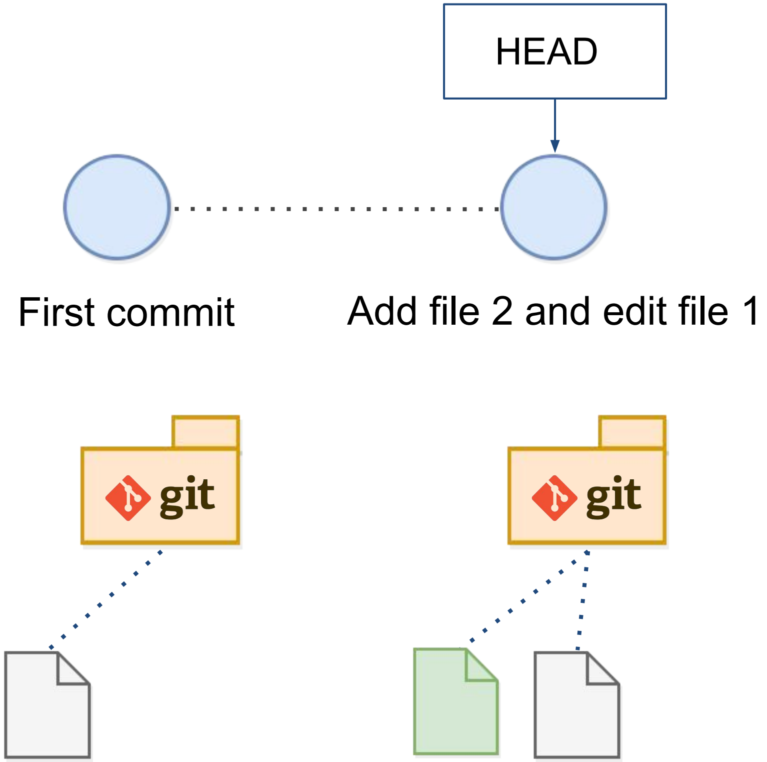
ID
Message
File changes
Author
Date

Link to ‘parent’

What is the ‘stage’?


This file is not included in the commit
The stage
git
git
git
Add desired changes to stage
First commit
Edit file 1 and add file 4

Navigating the history and the HEAD

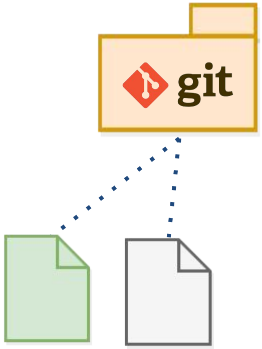


HEAD

First commit

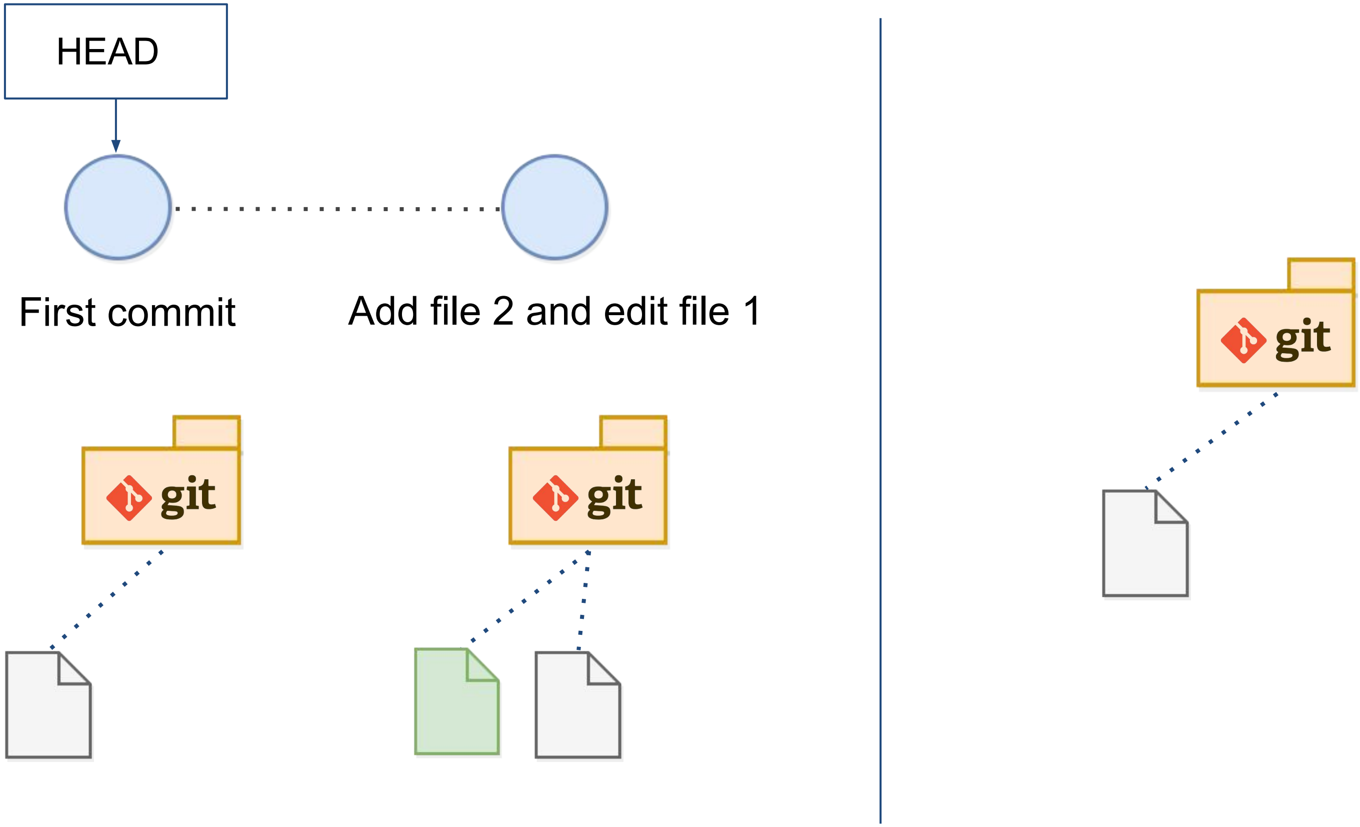
Add file 2 and edit file 1

Repository



File tree

Navigating the history and the HEAD



HEAD

First commit

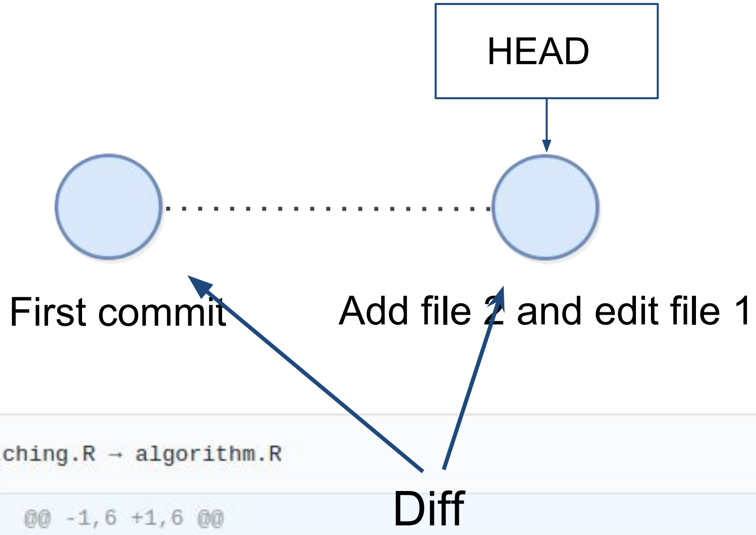
Add file 2 and edit file 1

Repository

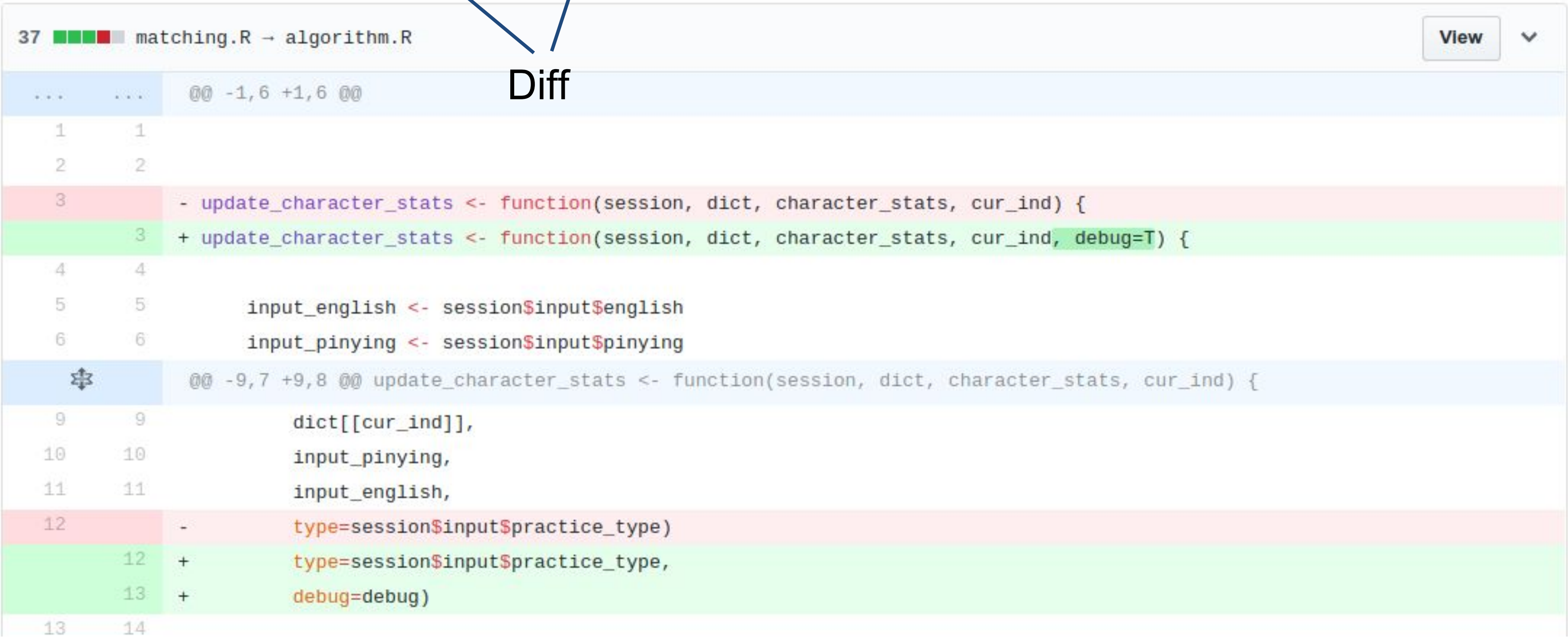
File tree

By moving the HEAD, the file tree changes

Comparing changes



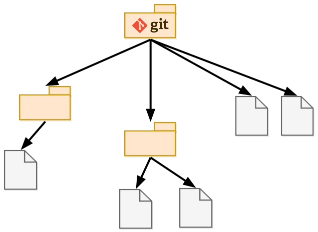
Differences between commits or branches can be easily visualized



```
matching.R → algorithm.R

@@ -1,6 +1,6 @@


- update_character_stats <- function(session, dict, character_stats, cur_ind) {
+ update_character_stats <- function(session, dict, character_stats, cur_ind, debug=T) {

      input_english <- session$input$english
      input_pinying <- session$input$pinying
@@ -9,7 +9,8 @@ update_character_stats <- function(session, dict, character_stats, cur_ind) {
          dict[[cur_ind]],
          input_pinying,
          input_english,
-         type=session$input$practice_type)
+         type=session$input$practice_type,
+         debug=debug)
```

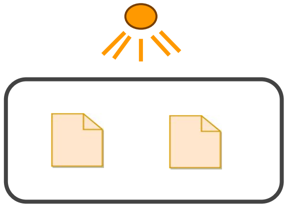
Recap

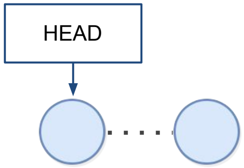
Repository and file tree



Commit

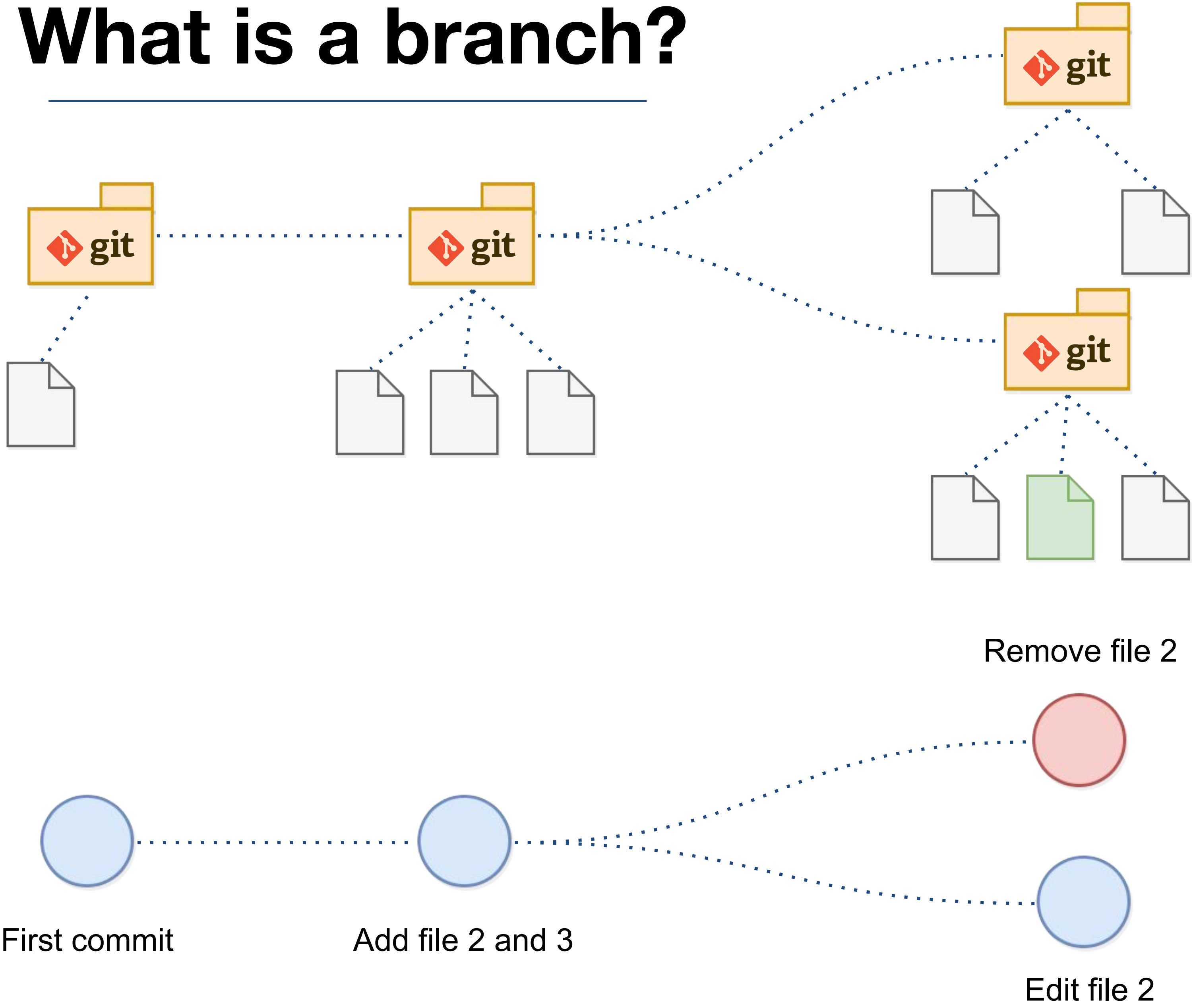
Stage



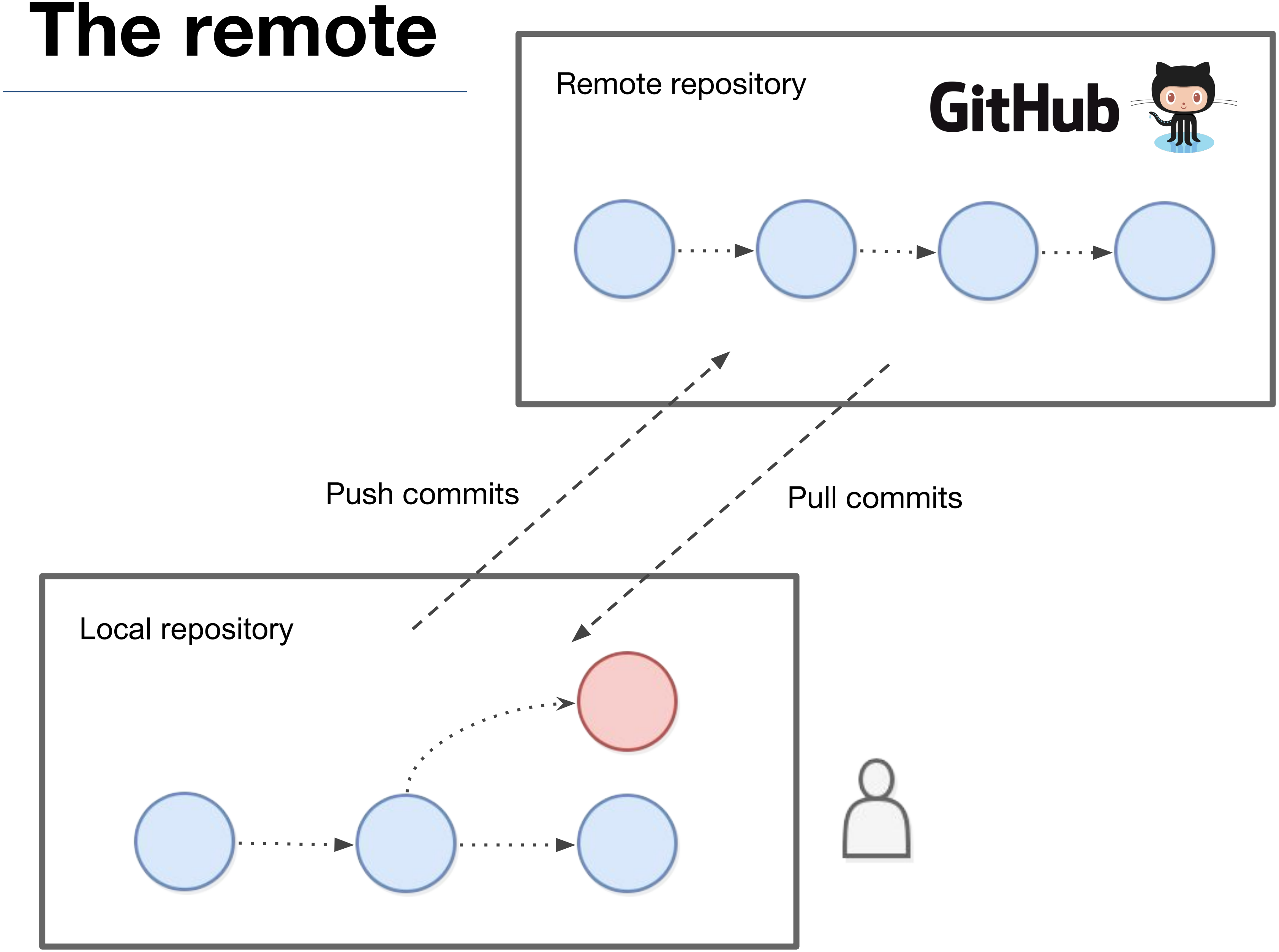


The HEAD

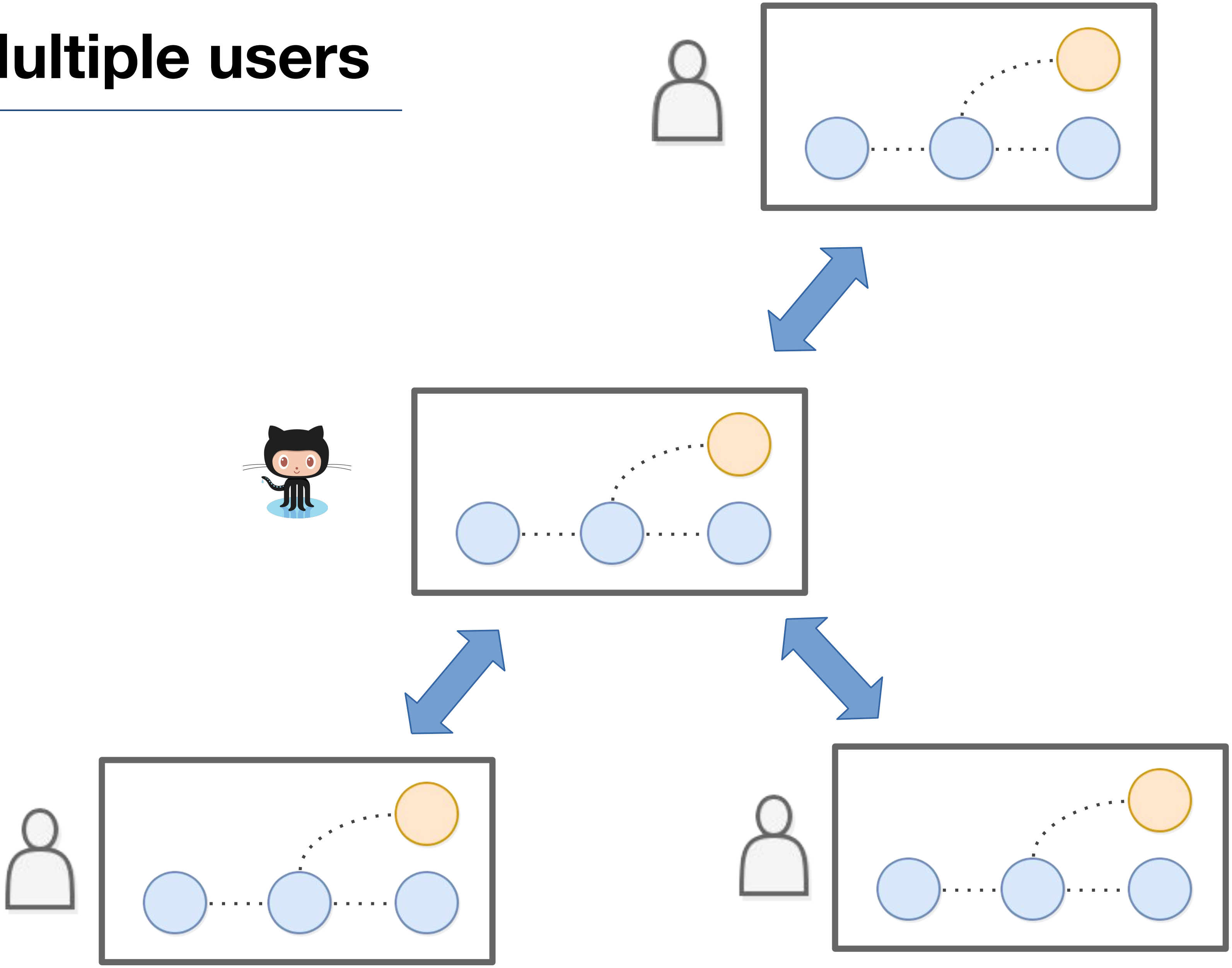
What is a branch?



The remote

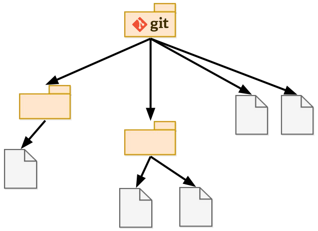

Remote repository
GitHub
Push commits
Pull commits
Local repository

Multiple users



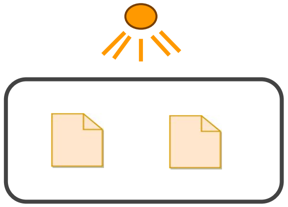
Recap

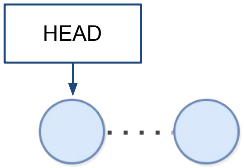
Repository and file tree



Commit

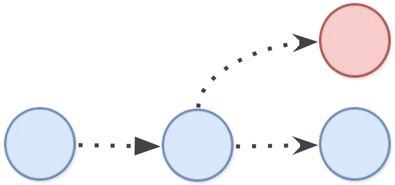
Stage

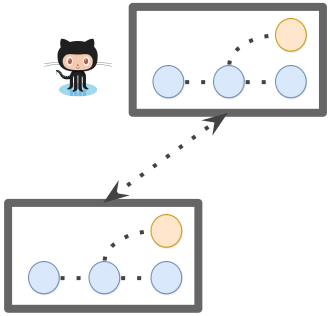




The HEAD

Branch





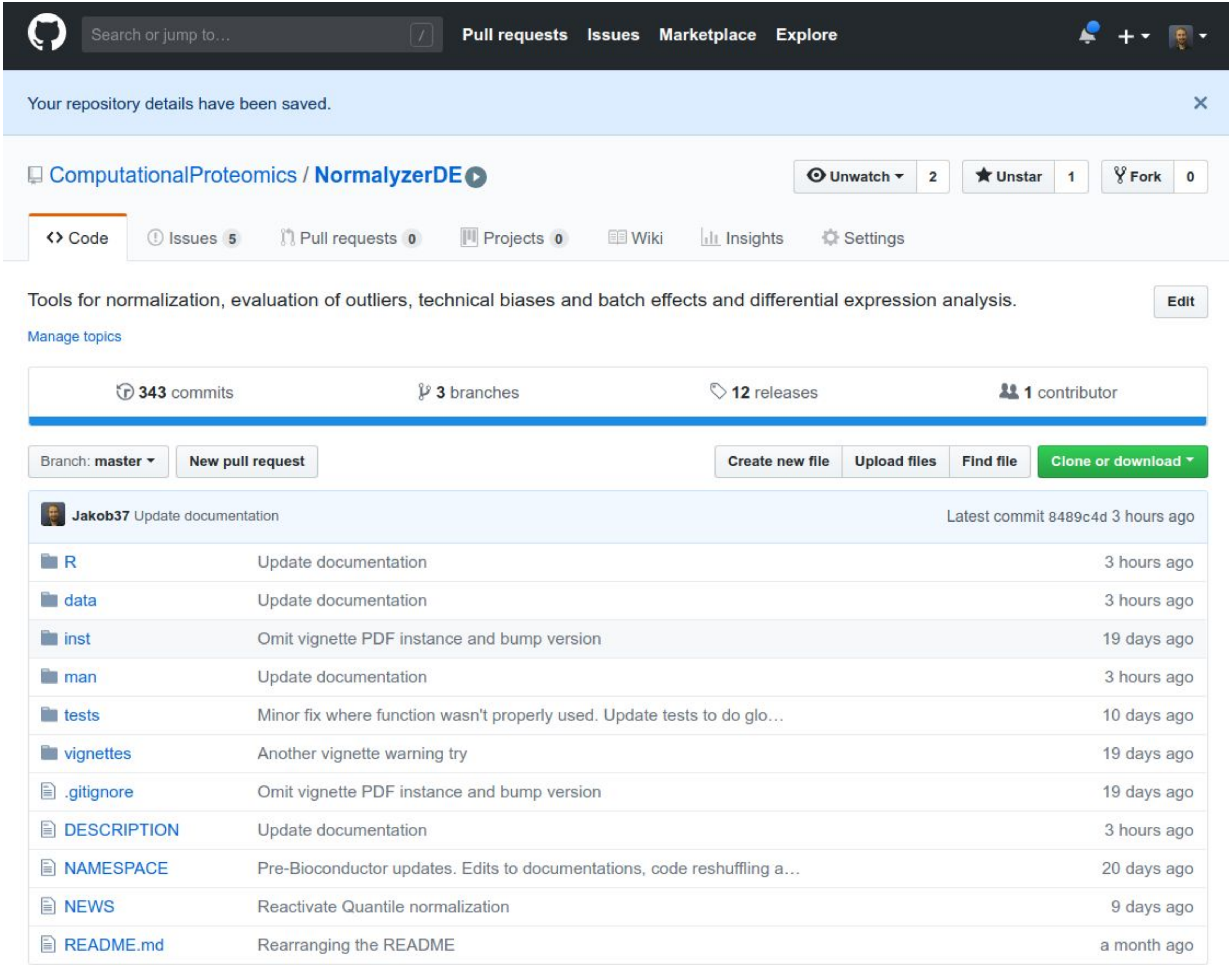
Remote

GitHub

Social platform for code

Common way to make code public

Allows interacting with other peoples code



Your repository details have been saved.

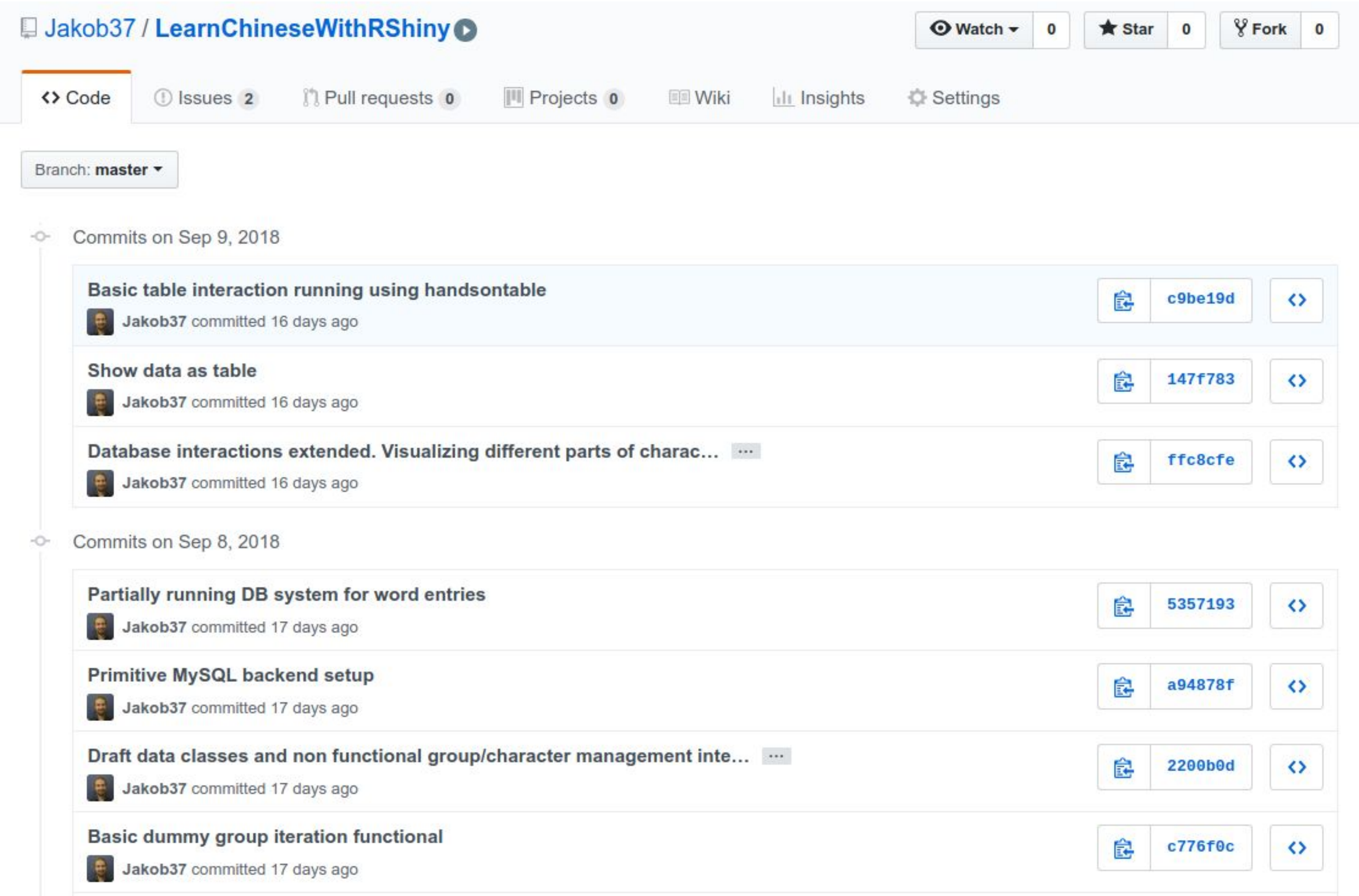
ComputationalProteomics / NormalyzerDE

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

343 commits · 3 branches · 12 releases · 1 contributor

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



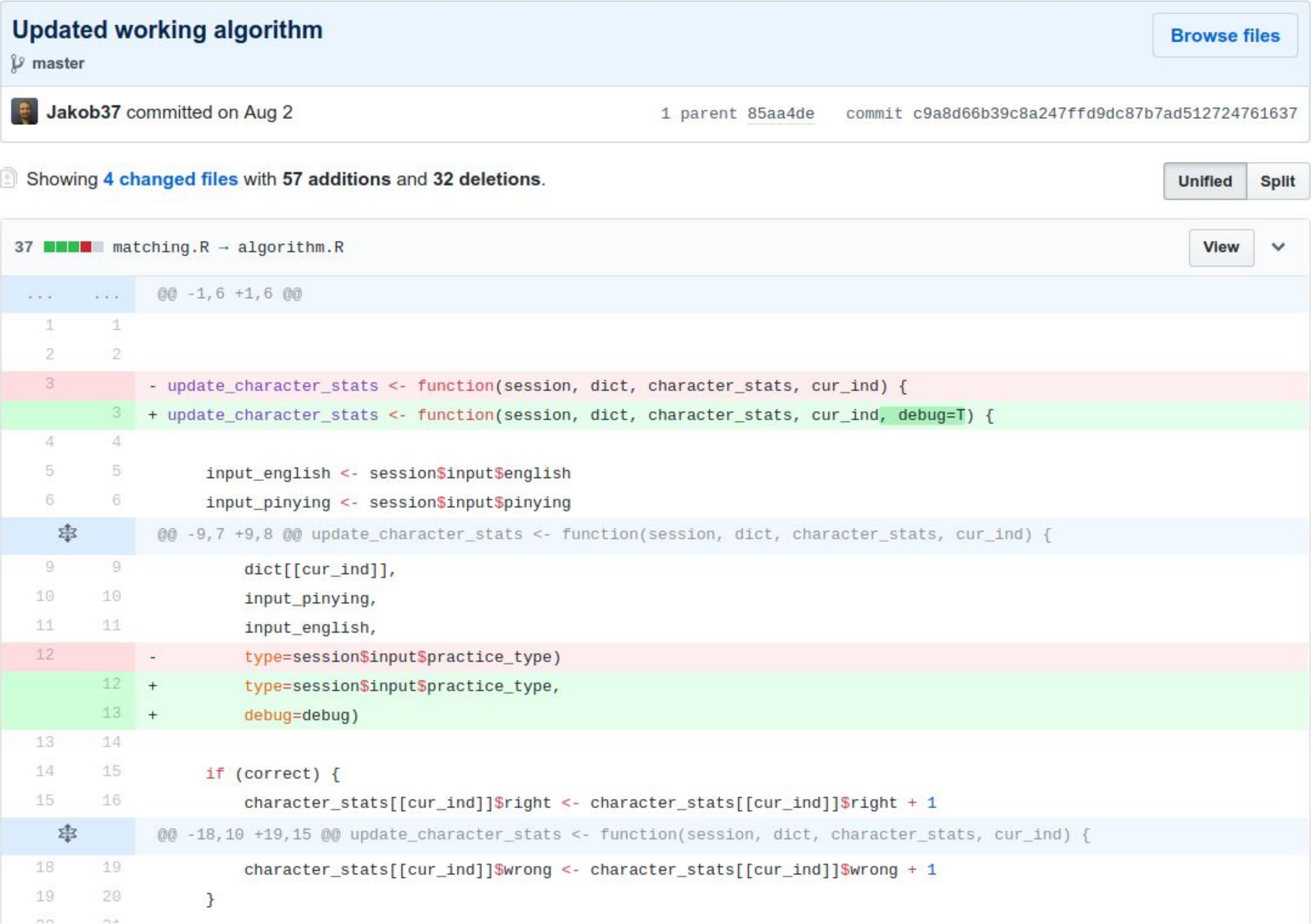
Jakob37 / LearnChineseWithRShiny

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

Jakob37 committed on Aug 2 1 parent 85aa4de commit c9a8d66b39c8a247ffd9dc87b7ad512724761637

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

37 🟩🟩🟩🟥⬜ matching.R → algorithm.R

```diff
@@ -1,6 +1,6 @@


- update_character_stats <- function(session, dict, character_stats, cur_ind) {
+ update_character_stats <- function(session, dict, character_stats, cur_ind, debug=T) {

    input_english <- session$input$english
    input_pinying <- session$input$pinying
@@ -9,7 +9,8 @@ update_character_stats <- function(session, dict, character_stats, cur_ind) {
      dict[[cur_ind]],
      input_pinying,
      input_english,
-     type=session$input$practice_type)
+     type=session$input$practice_type,
+     debug=debug)

    if (correct) {
      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) {
      character_stats[[cur_ind]]$wrong <- character_stats[[cur_ind]]$wrong + 1
    }
```

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

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

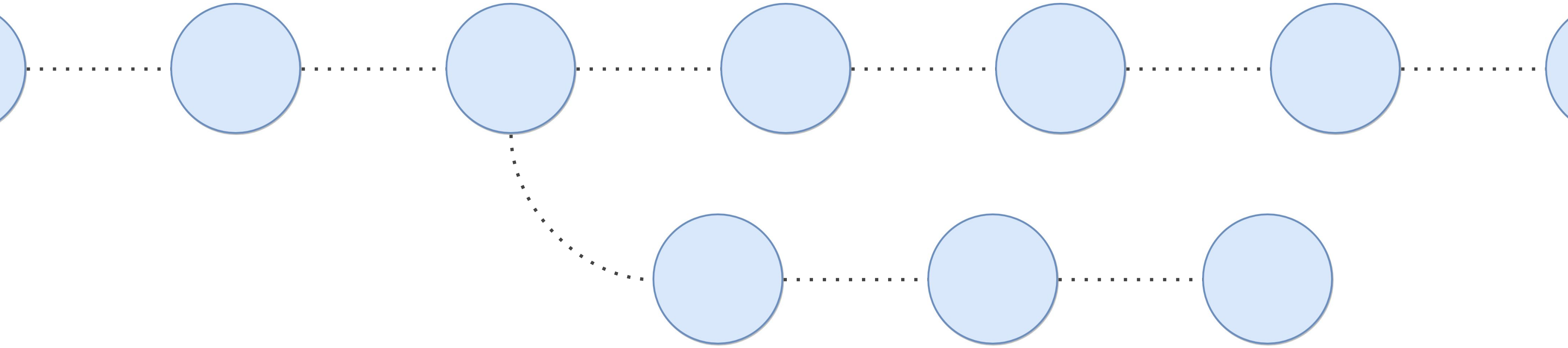
I will teach a 1.5 day introduction during spring 2019, contact me:
jakob.willforss@immun.lth.se

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