

# Bioinformatics Data Skills

## Git and Version Control in data science



**Ταμπόσης Ιωάννης**

Researcher & Software Development Engineer

PhD Candidate, University of Thessaly

Department of Computer Science and Biomedical Informatics

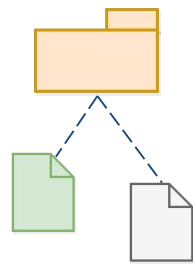
# Τι θα μάθουμε

- Τι είναι το git
- Βασική χρήση git
- Δουλεύοντας τοπικά με git
- Δουλεύοντας απομακρυσμένα με git
- Συνεργασία μέσω git{,hub}

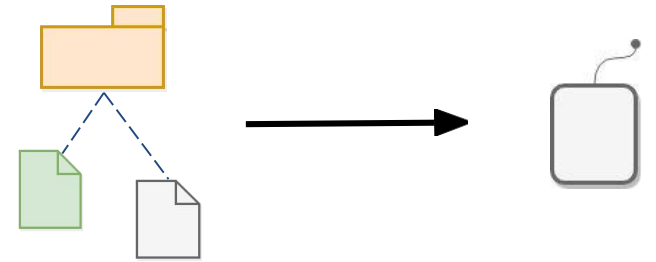
# Γιατί Version Control;

- Ως προγραμματιστές έχουμε ανάγκες
  - Νέος κώδικας μερικές φορές είναι buggy
  - Δουλεύουμε πολλοί ταυτόχρονα στον ίδιο κώδικα
  - Διαγράφουμε κώδικα που μπορεί να χρειαστεί ξανά
  - Χρειαζόμαστε back-ups για τη δουλειά μας
- Πώς κρατάμε πολλές εκδόσεις ενός αρχείου;
- Πώς επιστρέφουμε σε μία παλιά έκδοση;

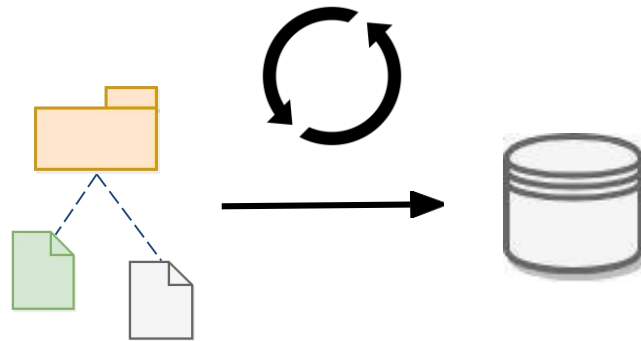
# Γιατί Version Control;



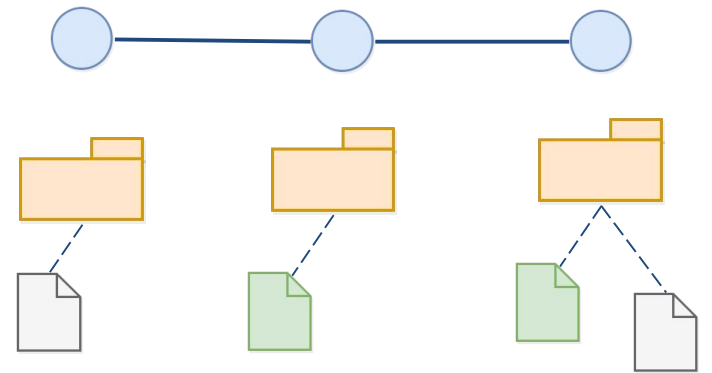
No backup



Manual backup



Automatic backup



Version control

# Γιατί Version Control;



- κρατάμε εκδόσεις στα αρχεία
- κάνουμε undo αλλαγές
- συνεργαζόμαστε με άλλους
- κρατάμε backups των αρχείων μας
- μοιραζόμαστε εύκολα τον κώδικα με την ομάδα
- ξέρουμε ποια είναι η «τελευταία» έκδοση

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

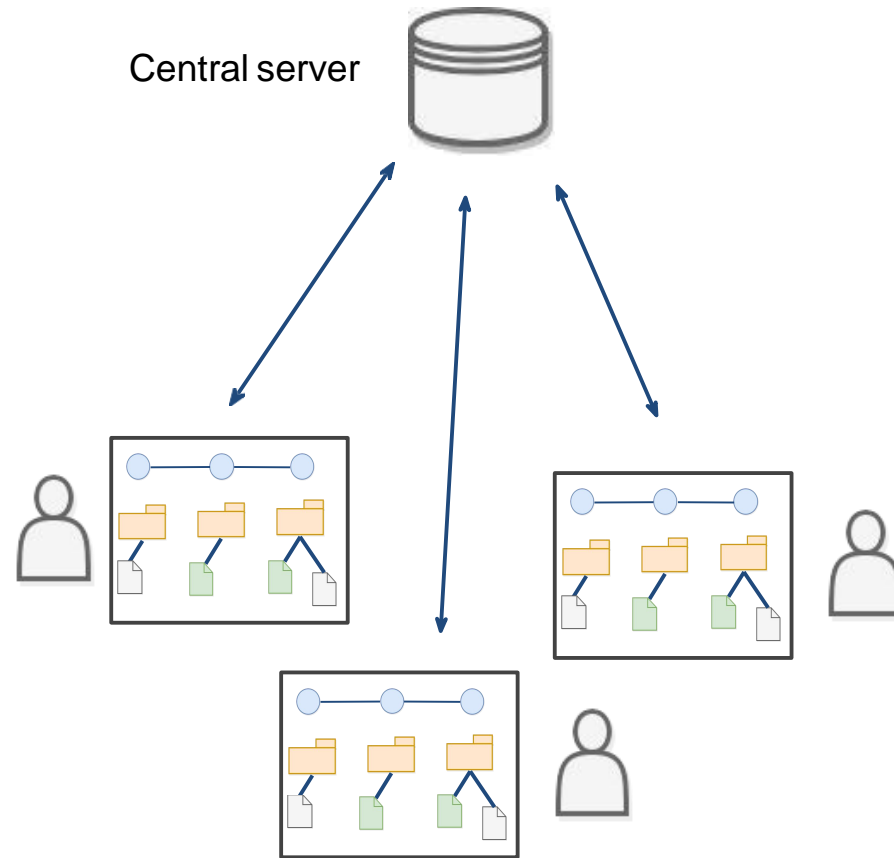
# Ιστορικά

- CVS – 1990
  - Από τα πρώτα πλήρη version control systems
- Subversion (SVN) – 2000
  - Διορθωμένο CVS για project-wide management
- git – 2005
  - Distributed version control system
- GitHub – 2008
  - Συνεργατικό περιβάλλον version control

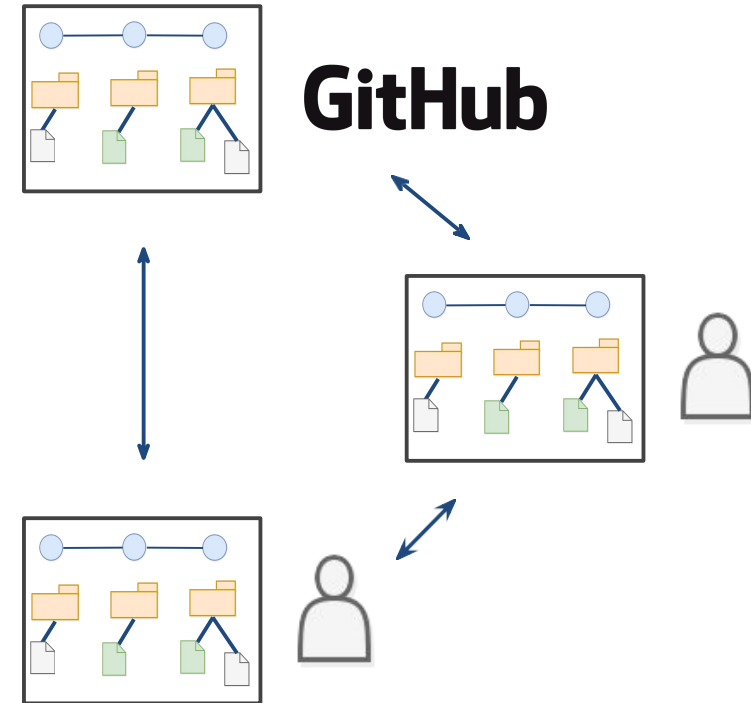
# Γιατί git;

## Centralized version control

SVN



## Distributed version control



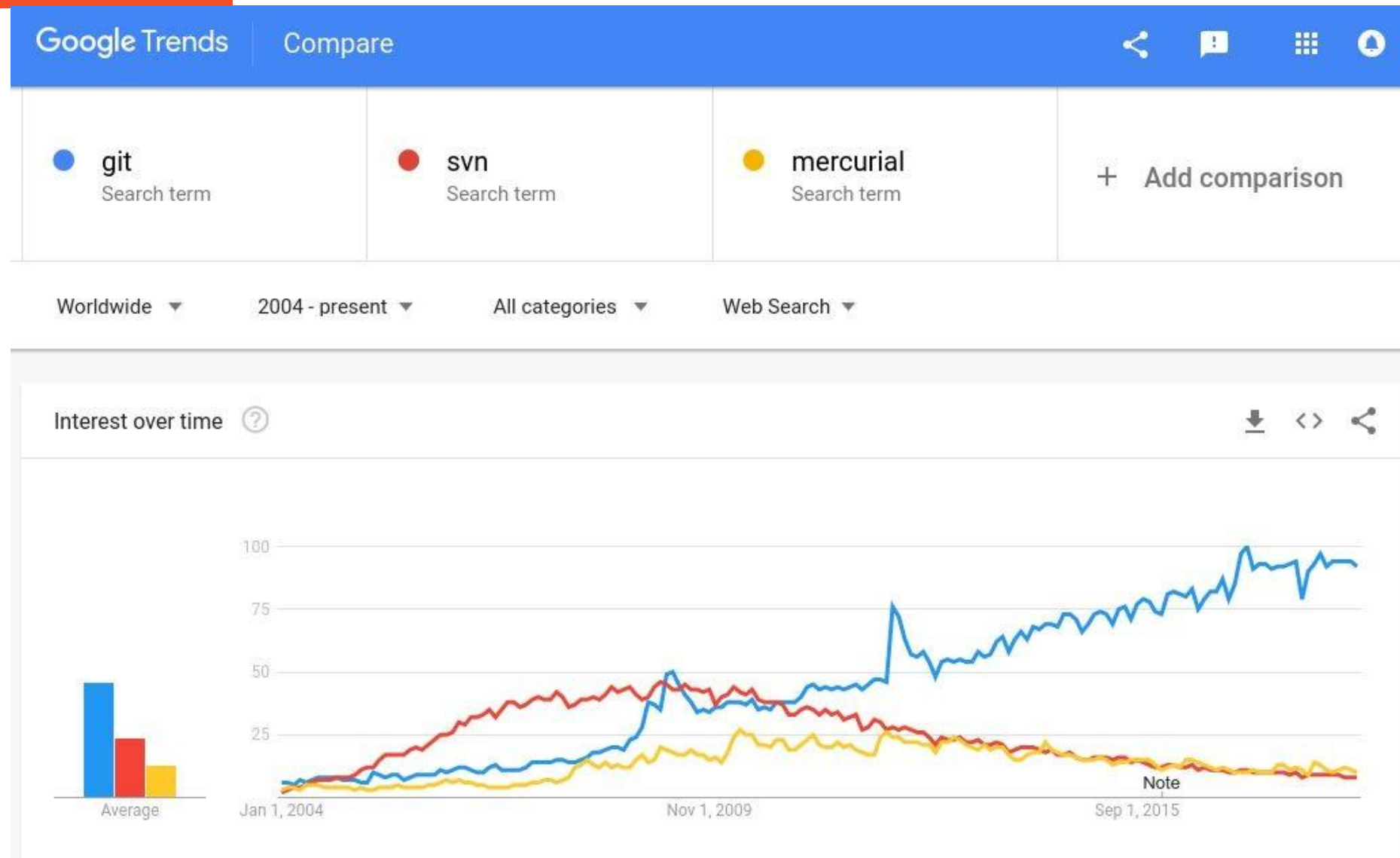
# Τι είναι το git?

- Πρόγραμμα που τρέχεις στον υπολογιστή σου
- Εργαλείο στο command line
- Το αφήνεις να χειριστεί τον κώδικά σου





# Γιατί git;



# Εγκατάσταση

- Linux (Debian, Ubuntu)
  - `apt-get install git`
- Mac
  - Τρέξε git και ακολούθα οδηγίες εγκατάστασης
- Windows
  - Κατέβασμα από το <https://git-scm.com/download>
  - Εγκατάσταση
  - Τρέξε το git CMD

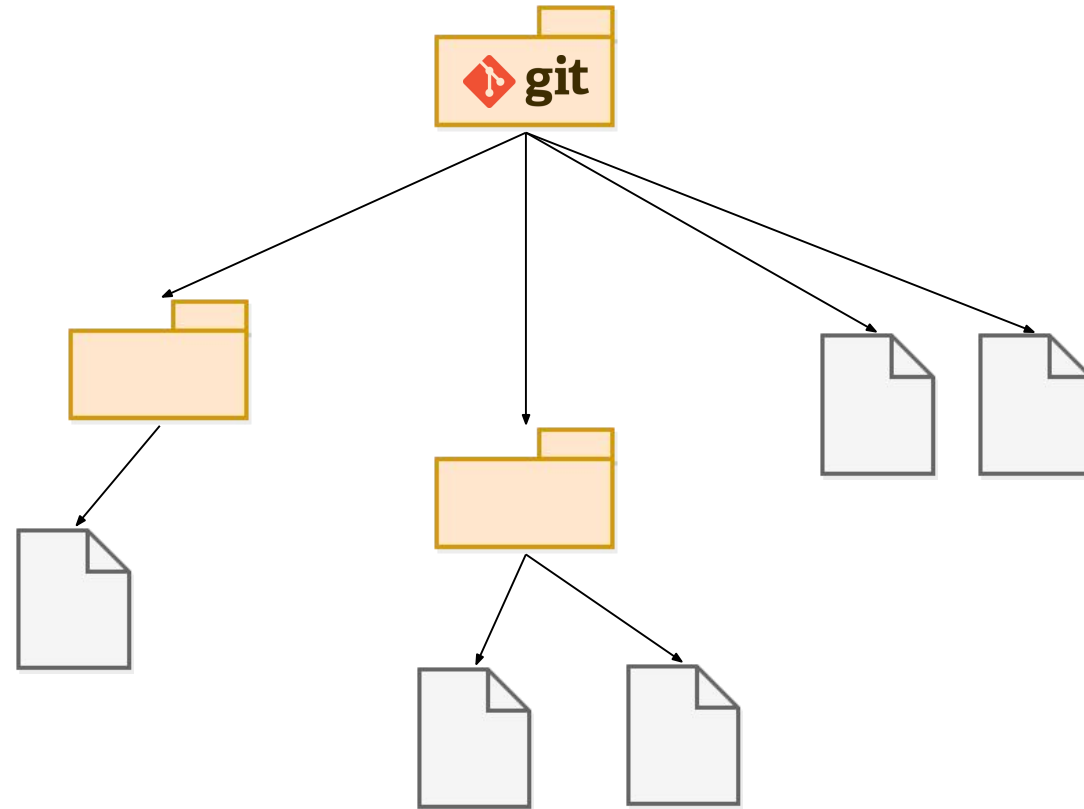


# Τι κάνει το Git

- Προσδιορίζει τι άλλαξε, ποιος το άλλαξε και γιατί
- Οργανωμένος τρόπος συνεργασίας στον κώδικα
- Τρόπος παρουσίασης κώδικα
- επιτρέπεται η πλοήγηση στο ιστορικό αρχείων

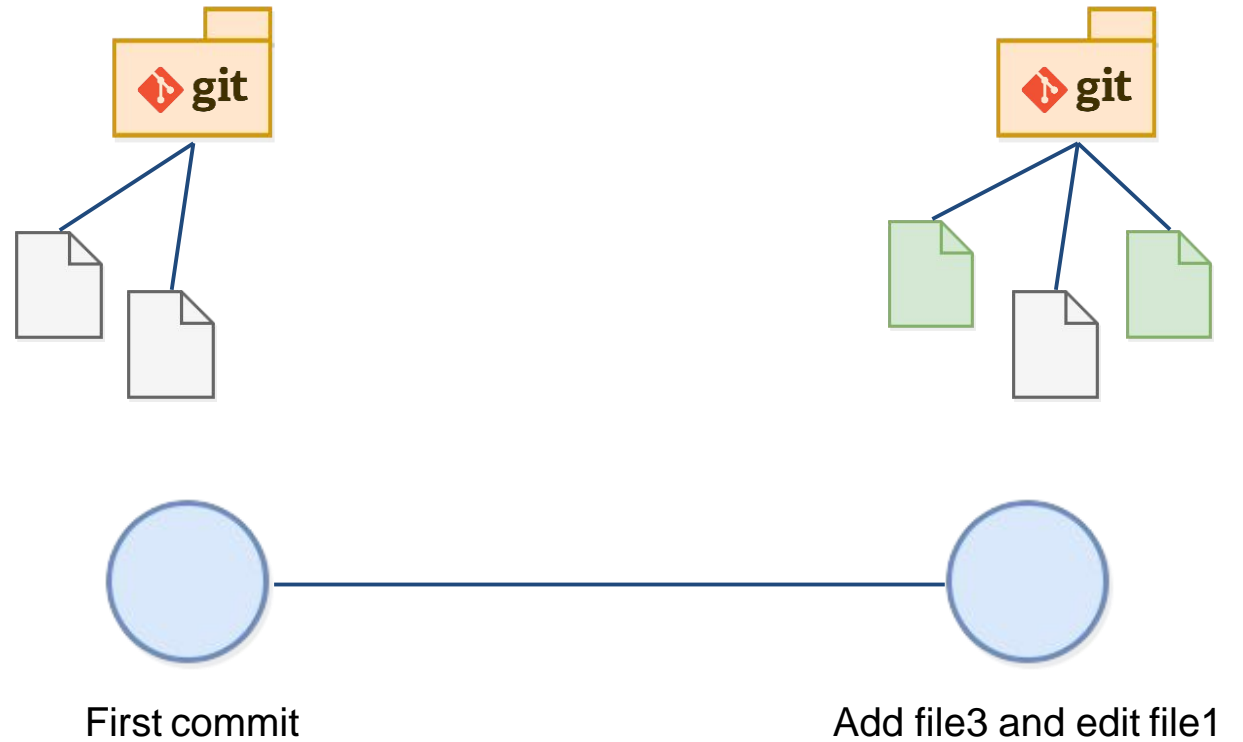


# The 'repository' and the 'file tree'



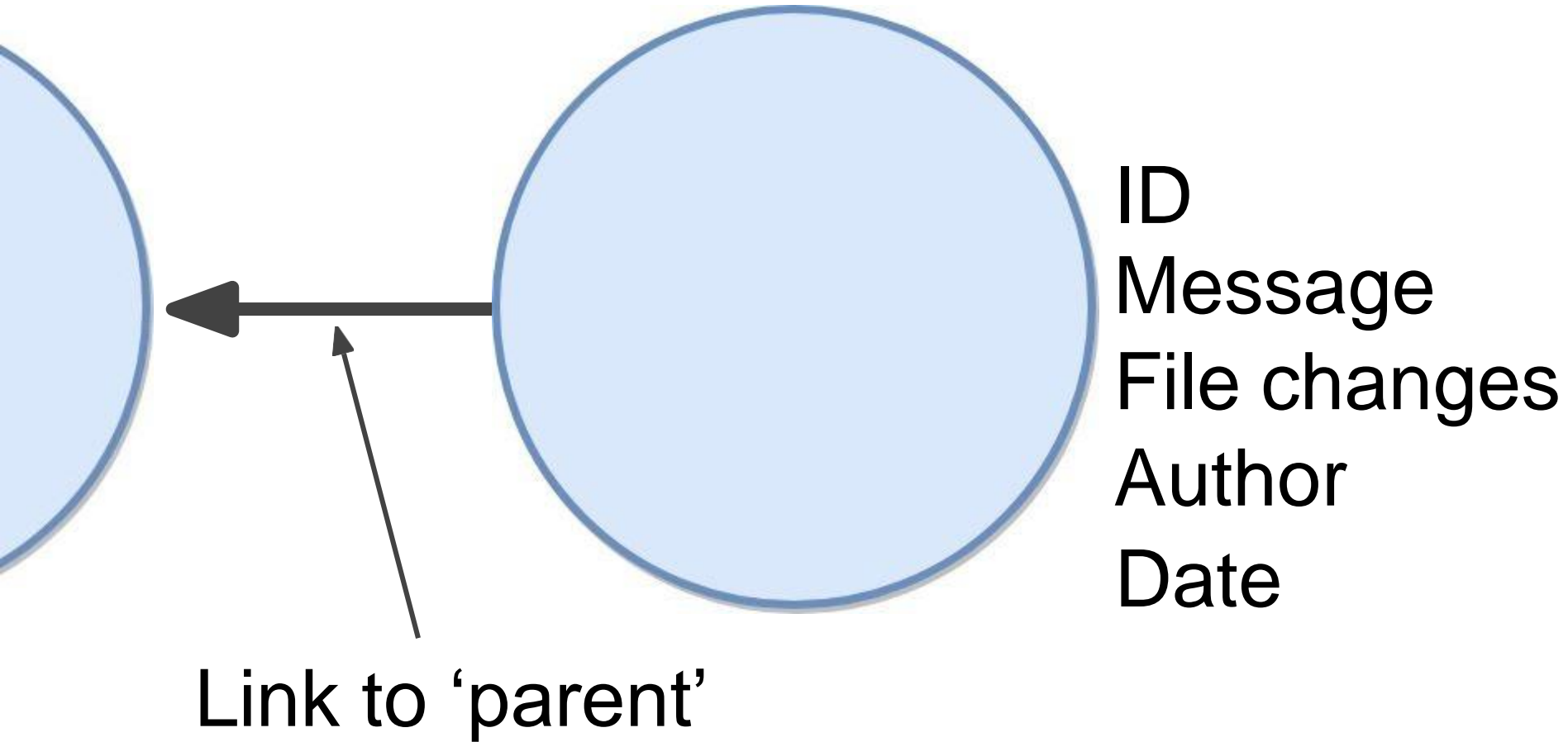
The 'repository' παρακολουθεί τις αλλαγές στο 'file tree'

# What is a commit?



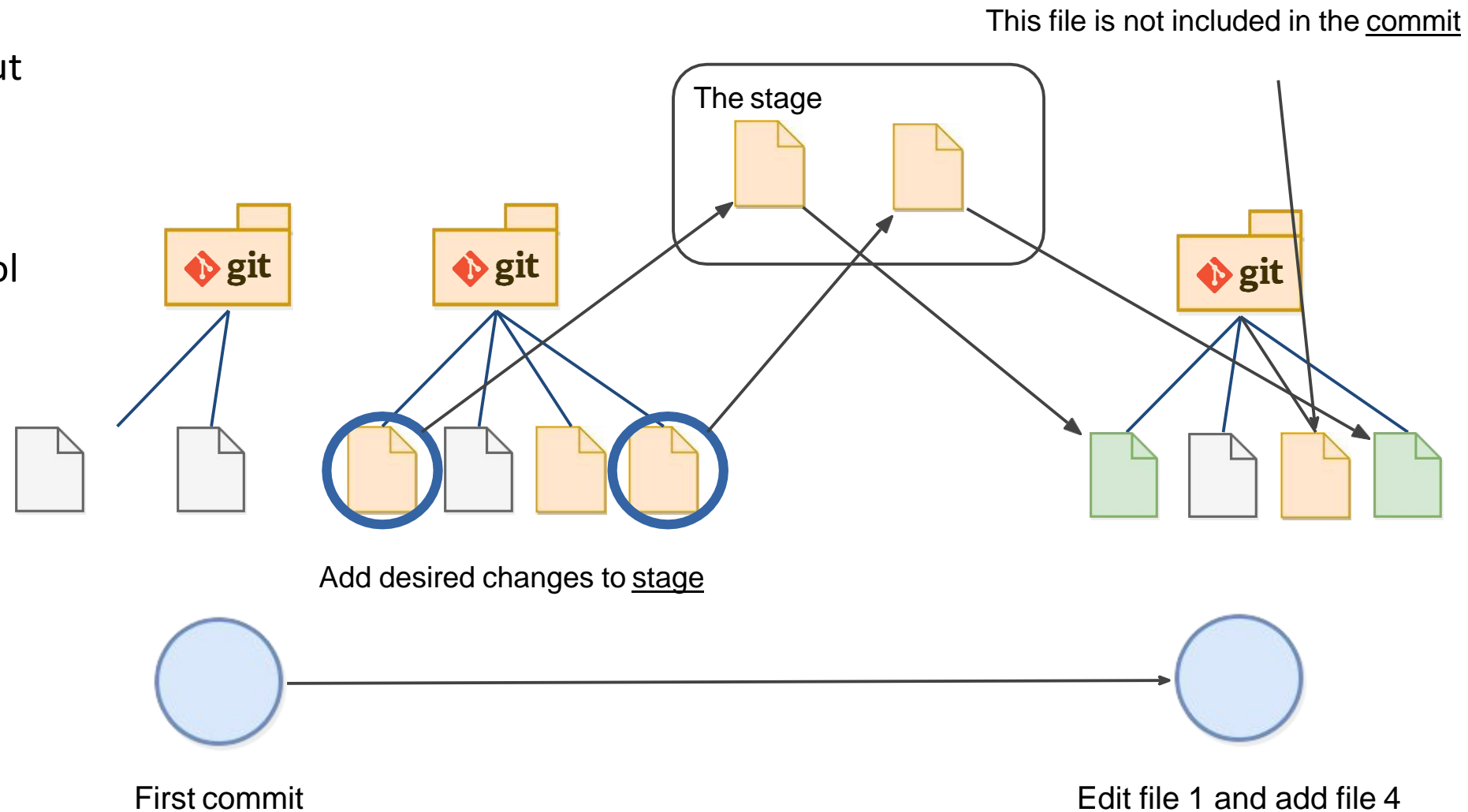
Ένα στιγμιότυπο συγκεκριμένης κατάστασης στο file tree

# What is a commit?



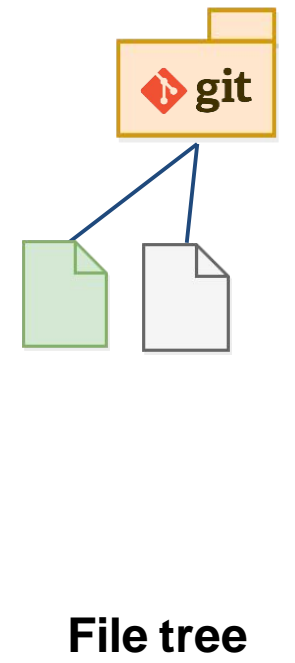
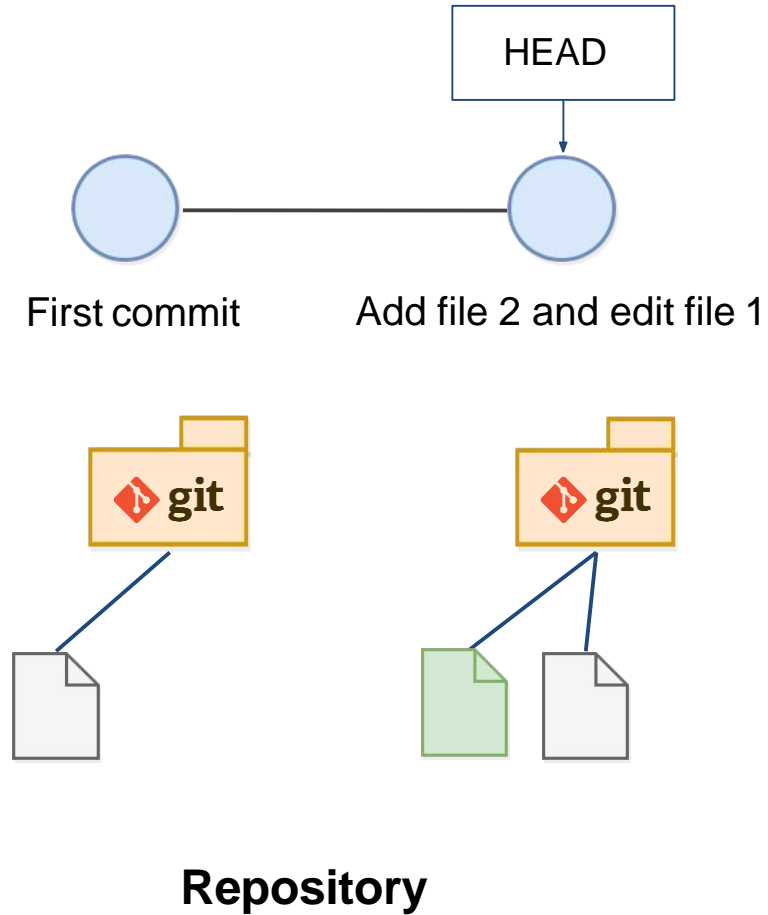
# What is the 'stage'?

- Untracked: the file exists, but is not part of git's version control
- Staged: the file has been added to git's version control but changes have not been committed
- Committed: the change has been committed



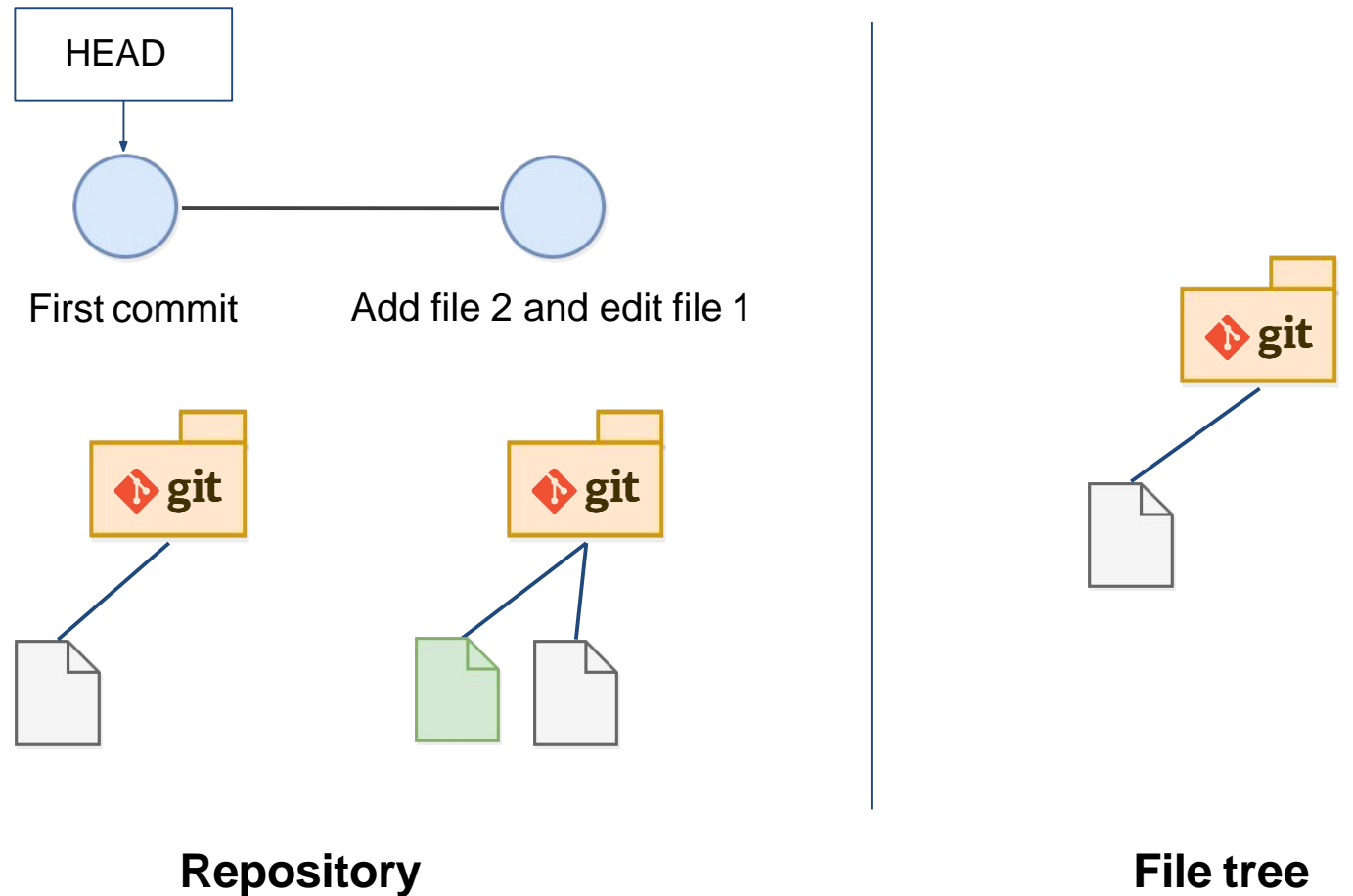
>>git status

# What is the 'HEAD'?



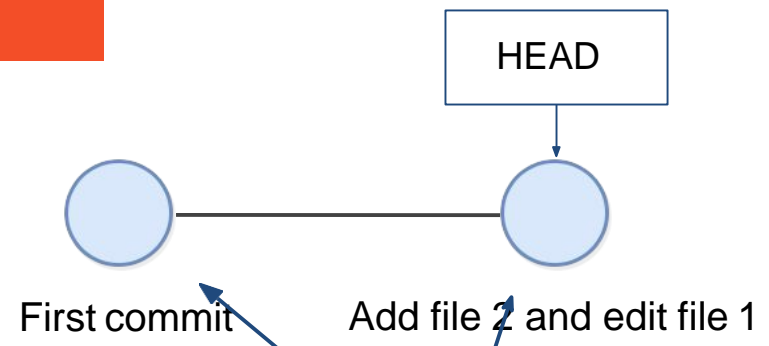


# What is the 'HEAD'?



By moving the HEAD, the file tree changes

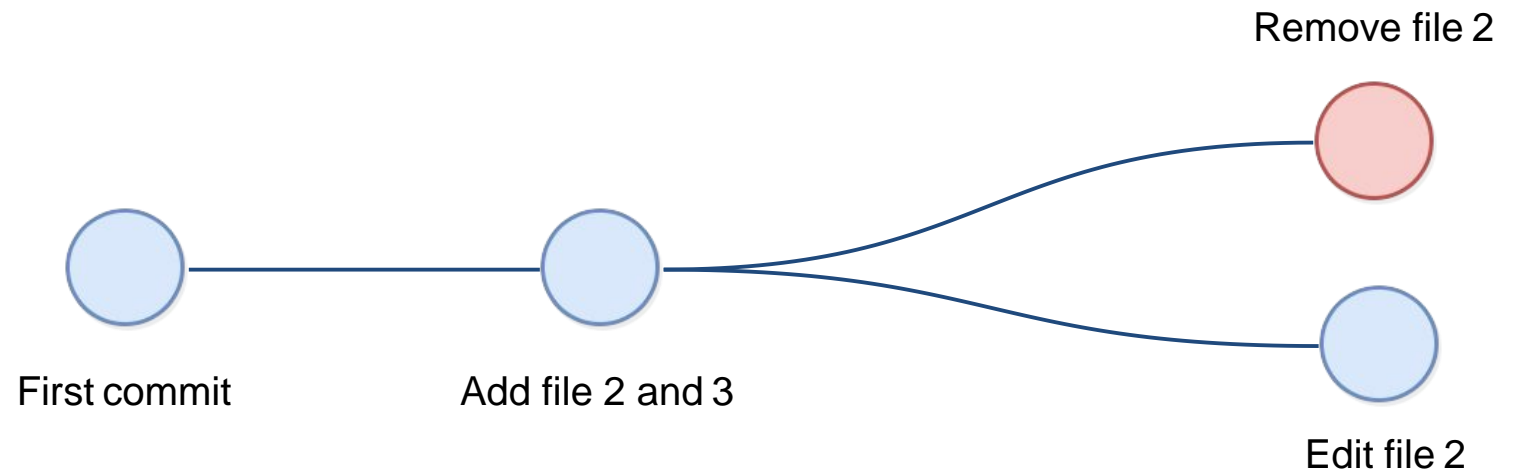
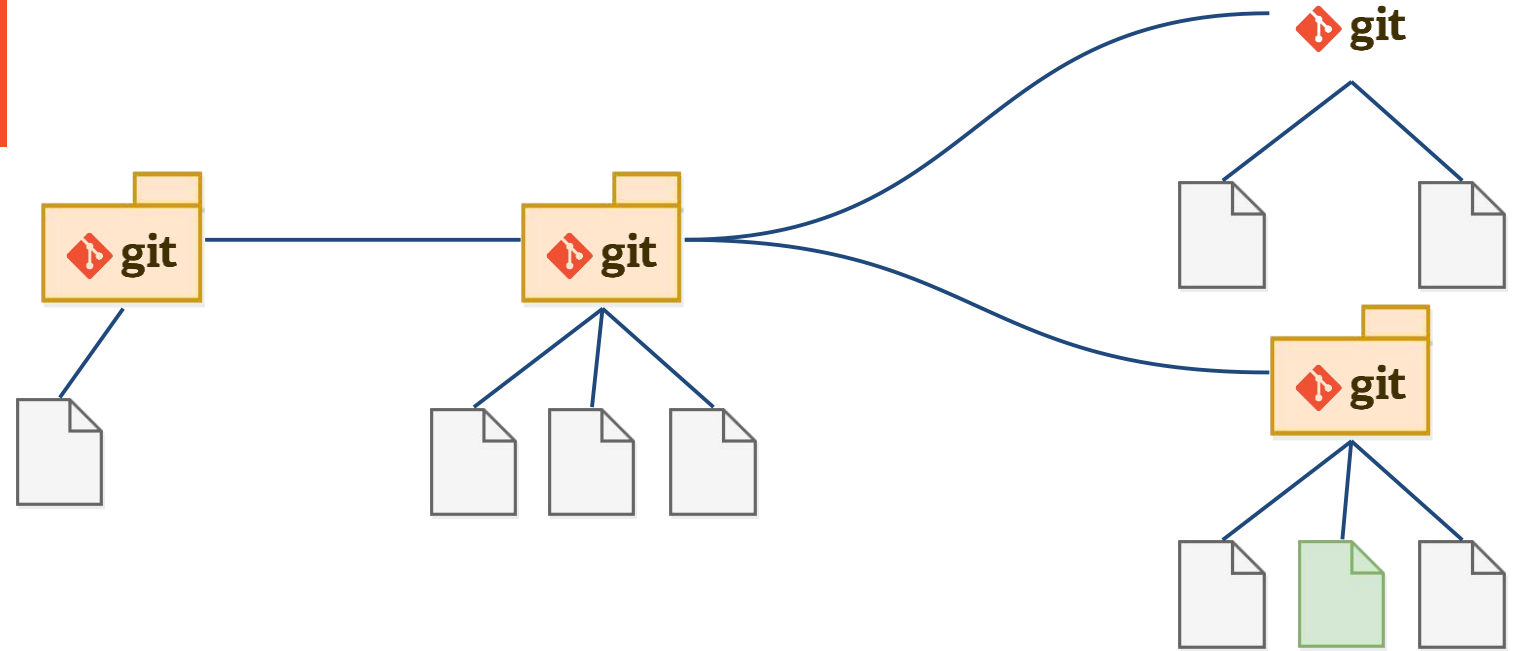
# Comparing changes



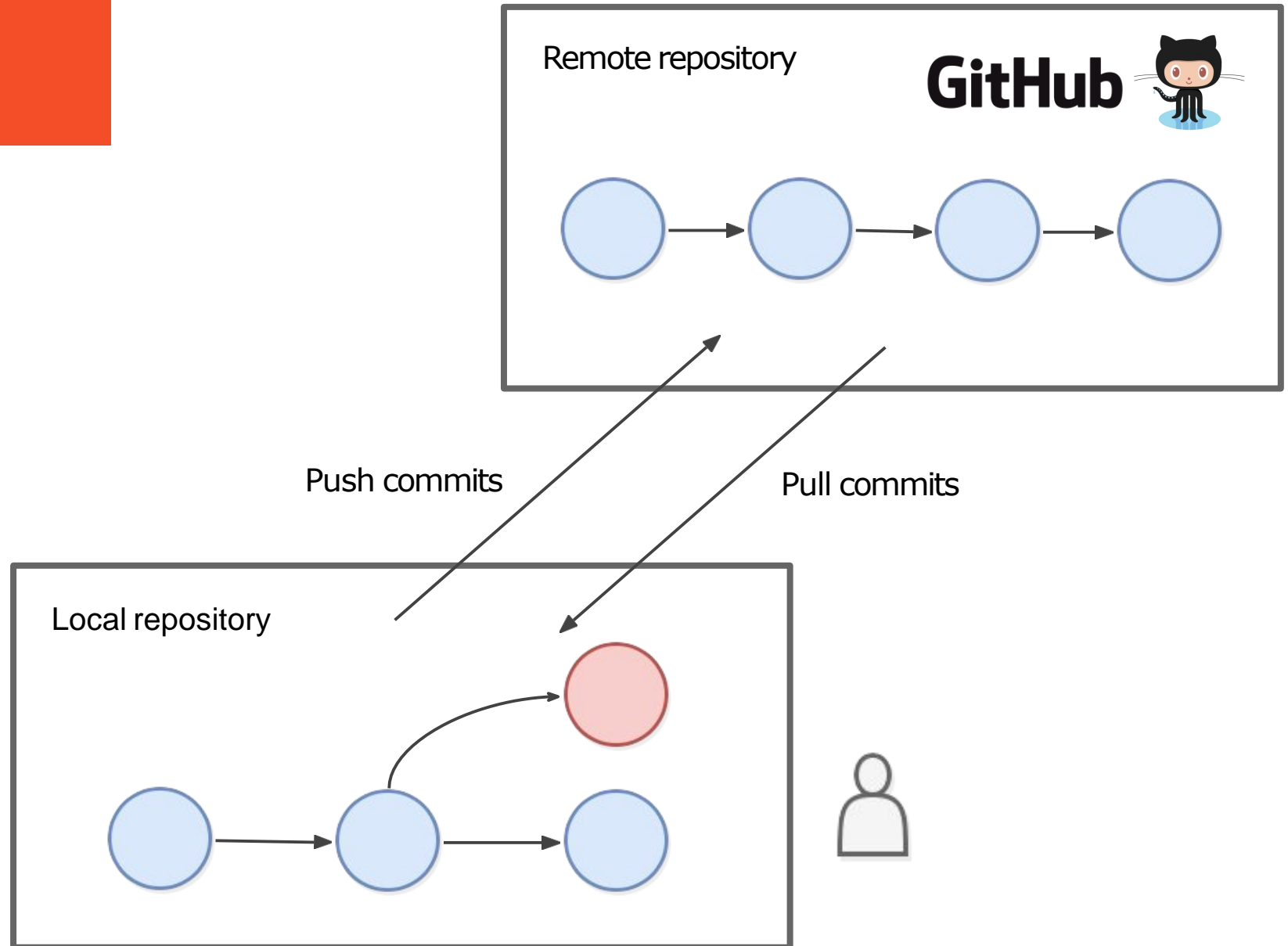
Differences between commits or branches can be easily visualized

```
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) {
+ 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)
+     type=session$input$practice_type,
+     debug=debug)
13 14
```

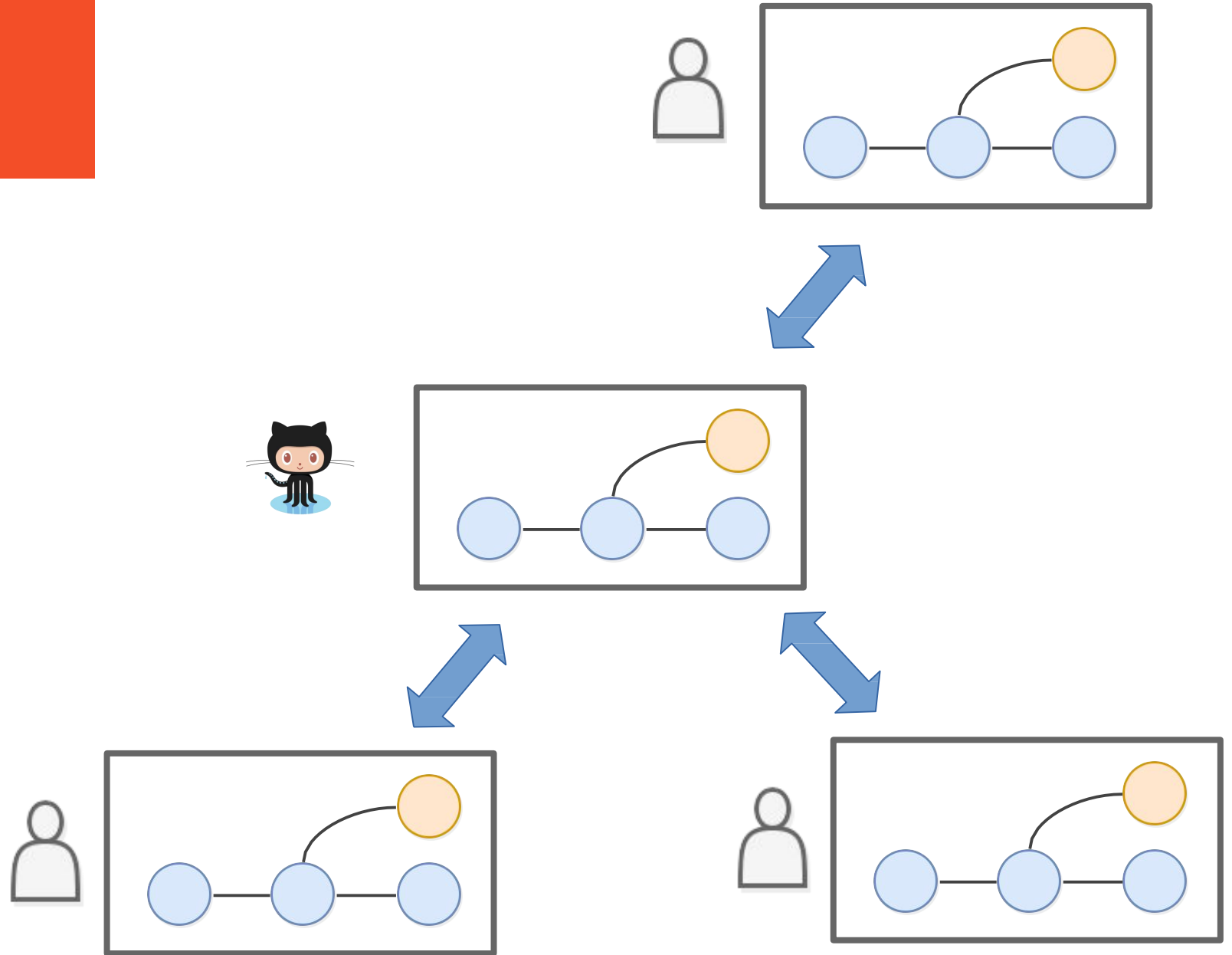
# What is a branch?



# The remote



# Multiple users



# Remote repositories

# GitHub



# GitLab



- Κοινωνική πλατφόρμα για κώδικα
- Κοινός τρόπος για να γίνει δημόσιος ο κώδικας
- Επιτρέπει την αλληλεπίδραση με τον κώδικα άλλων λαών

# Remote repositories

The screenshot shows the GitHub interface for the repository 'pbagos / juchmme'. At the top, there is a search bar and navigation links for 'Pull requests', 'Issues', 'Marketplace', and 'Explore'. The repository name is followed by 'Watch' (3), 'Star' (1), and 'Fork' (0) buttons. Below this, there are tabs for 'Code', 'Issues' (0), 'Pull requests' (0), 'Projects' (0), 'Wiki', 'Insights', and 'Settings'. The repository description is 'Java Utility for Class Hidden Markov Models and Extensions' with a link to 'http://www.compgen.org/tools/juchmme' and an 'Edit' button. A 'Manage topics' link is also present. A summary bar shows '14 commits', '1 branch', '2 releases', '2 contributors', and 'GPL-3.0' license. Below this, there are buttons for 'Branch: master', 'New pull request', 'Create new file', 'Upload files', 'Find File', and 'Clone or download'. The commit history shows a commit by 'itamposis' titled 'Modified Parallelization' with a 'Latest commit 9c4d059 8 days ago'. The file list includes 'conf' (Model Files, 25 days ago), 'models' (Modified Parallelization, 8 days ago), 'src' (Modified Parallelization, 8 days ago), 'tables' (Modified Parallelization, 8 days ago), '.gitignore' (Initial commit, 3 months ago), 'LICENSE' (Initial commit, 3 months ago), and 'README.md' (Update README.md, 2 months ago). At the bottom, there is a preview of the 'README.md' file.

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

pbagos / juchmme Watch 3 Star 1 Fork 0

<> Code Issues 0 Pull requests 0 Projects 0 Wiki Insights Settings

Java Utility for Class Hidden Markov Models and Extensions <http://www.compgen.org/tools/juchmme> Edit

Manage topics

14 commits 1 branch 2 releases 2 contributors GPL-3.0

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

itamposis Modified Parallelization ... Latest commit 9c4d059 8 days ago

conf	Model Files	25 days ago
models	Modified Parallelization	8 days ago
src	Modified Parallelization	8 days ago
tables	Modified Parallelization	8 days ago
.gitignore	Initial commit	3 months ago
LICENSE	Initial commit	3 months ago
README.md	Update README.md	2 months ago

README.md





# Google Drive

The screenshot displays the Google Drive web interface. On the left, the navigation sidebar includes options like 'New', 'My Drive', 'Computers', 'Shared with me', 'Recent', 'Starred', 'Bin', 'Backups', and 'Storage' (2 GB of 17 GB used). The main area shows a list of files under 'My Drive'. A context menu is open over the file 'LogistisPayment.xlsx', listing actions such as 'Preview', 'Open with', 'Share', 'Get shareable link', 'Move to', 'Add to Starred', 'Rename', 'View details', 'Manage versions', 'Make a copy', 'Report abuse', 'Download', and 'Remove'. The 'Manage versions' option is circled in red. A notification at the bottom states 'You are offline. Some functionality may be unavailable.'

Name	Owner	Last modified	File size
ΠΡΟΚΥΡΗΞΕΙΣ	me	20 Mar 2017	—
biomarker_Kidney_pro	me	20 Mar 2019	13 KB
EAL30-03-15.pdf	me	30 Mar 2015	667 KB
ECCB Volunteer Regist	me	21 Nov 2018	—
karta-igeias-athliti154	me	1 Aug 2018	423 KB
<b>LogistisPayment.xlsx</b>	me	14 Jan 2019	10 KB
MediaMarkt_Singer_O	me	6 Nov 2015	75 KB
OEEK_Aithsh9055.pdf	me	27 Apr 2018	76 KB
OEEK_Aithsh9064.pdf	me	27 Apr 2018	76 KB
safari oreilly info.txt	me	16 Nov 2008	54 bytes
	me	9 Mar 2019	342 bytes
signal ML cros VITER	me	9 Mar 2019	340 bytes

# Google Drive

The screenshot shows the Google Drive web interface. A modal dialog titled "Manage versions" is open, displaying a list of file versions for "LogistisPayment.xlsx". The dialog includes an "UPLOAD NEW VERSION" button at the top and a "CLOSE" button at the bottom right. The background shows a file list with columns for Name, File size, and other details. The "LogistisPayment.xlsx" file is highlighted in the list, corresponding to the version information in the dialog.

**Manage versions**

Drive keeps older versions of 'LogistisPayment.xlsx' for 30 days or 100 versions, whichever happens first. Versions are displayed in the order they were uploaded to Drive. [Learn more](#)

**UPLOAD NEW VERSION**

Version	File Name	Timestamp	Owner	File Size
Current version	LogistisPayment.xlsx	14 Jan, 12:44	Ioannis Tamposis	10 KB
Version 3	LogistisPayment.xlsx	15 Sep 2018, 14:16	Ioannis Tamposis	423 KB
Version 2	LogistisPayment.xlsx	15 Sep 2018, 14:14	Ioannis Tamposis	75 KB
Version 1	LogistisPayment.xlsx			76 KB

**CLOSE**

# Getting started

## Πώς να το χρησιμοποιήσετε

Από γραμμή εντολών:

<https://git-scm.com>

Από γραφικό περιβάλλον:

<https://desktop.github.com>

## Πώς να το μάθετε

Codecademy - Interactive online tutorial (free):

<https://www.codecademy.com/learn/learn-git>