

GIT AS A GENERAL TOOL FOR DATA-MANAGEMENT

Harald von Waldow <harald.vonwaldow@thuenen.de>

First Thünen Code Café

2022-09-27

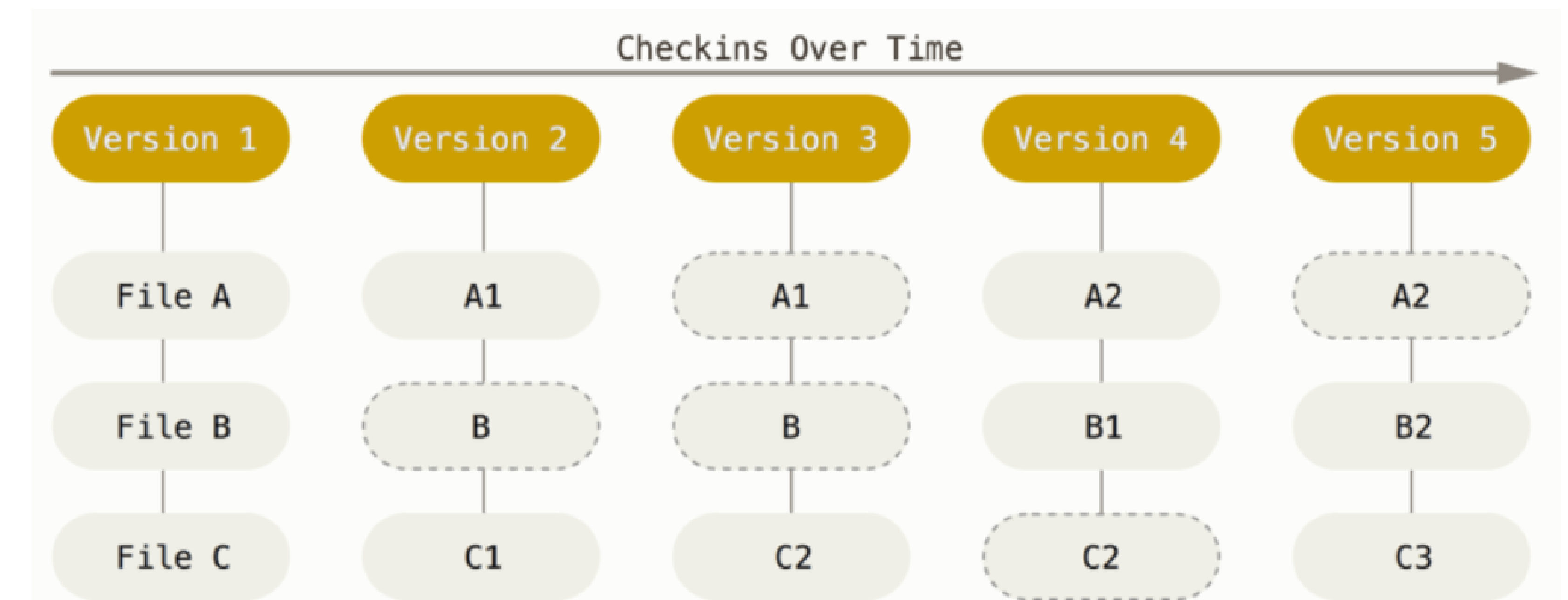
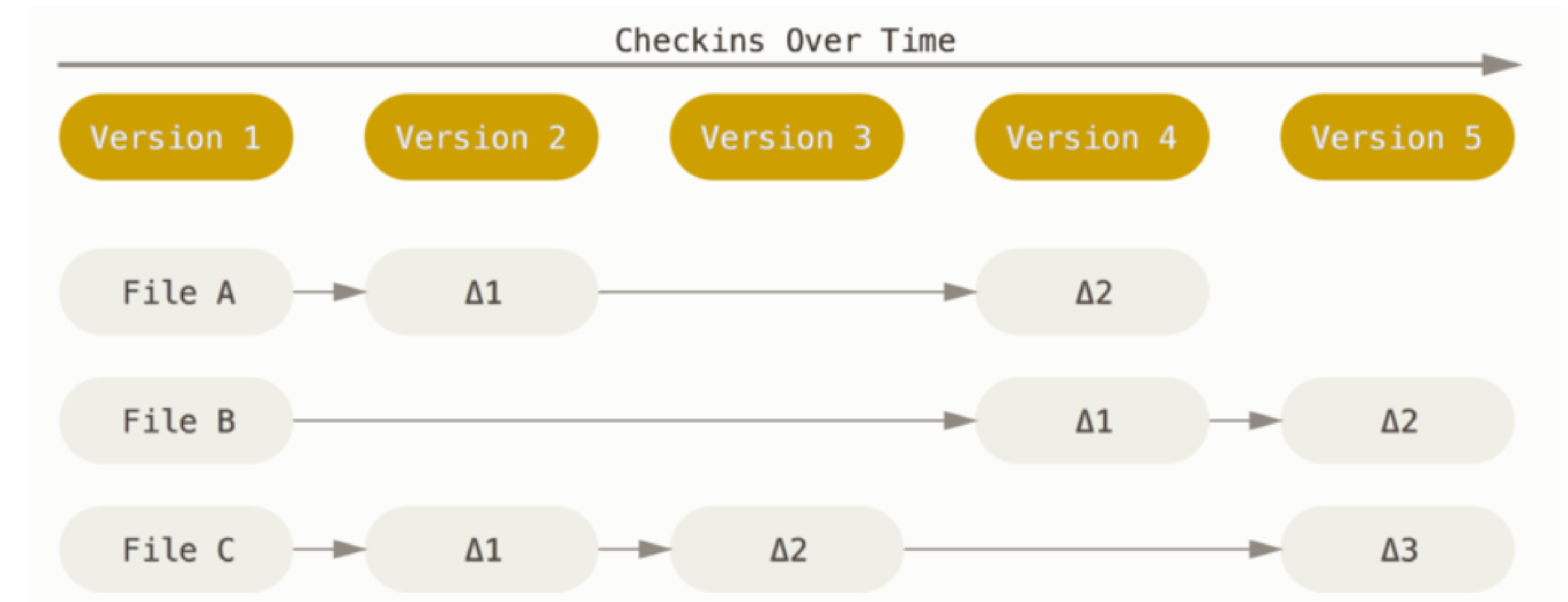


WHAT IS GIT?

1. A program for **version-control** in software-development
 - Developed in 2005 to manage Linux kernel development (~5e6 LOC at the time)
2. A **distributed, decentralized** database
 - Every copy of the database has the full information
3. A **content-addressable** file-system
 - Everything is stored by an **SHA1-hash**.
 - *Commits, trees and blobs* form a *Merkle Tree*

VERSION CONTROL

- Snapshots whole folders recursively
- Provides a history
- Attached metadata: who, what, when



VERSION CONTROL

- Snapshots whole folders recursively
- Provides a history
- Attached metadata: who, what, when

```
commit 8d0e7f557bbcf2d152c4e984d5866bddf504c0d6
Author: Harald von Waldow <harald.vonwaldow@thuenen.de>
Date:   Fri Aug 26 14:30:02 2022 +0200

    added exceptions (reading)

commit 587d98289cb9c7b7bdf46d7c366c338ea287dd45
Author: Harald von Waldow <harald.vonwaldow@thuenen.de>
Date:   Fri Aug 26 14:16:17 2022 +0200

    created skiplist

commit 8a3c797875b0b2e9f6eaa2e9364a2a178faeff45
Author: Harald von Waldow <harald.vonwaldow@thuenen.de>
Date:   Thu Aug 25 16:28:24 2022 +0200

    adapted test to new entryform-template

commit ba779d0601efa856ee468ab8b6979beed8e29de9
Author: Harald von Waldow <harald.vonwaldow@thuenen.de>
Date:   Wed Aug 24 13:11:17 2022 +0200

    new entryform in fix_entryform dir

commit 1b563b55c7ea590dfb7130dc1d47c36e91a7f152
Author: Harald von Waldow <harald.vonwaldow@thuenen.de>
Date:   Wed Aug 24 13:10:12 2022 +0200

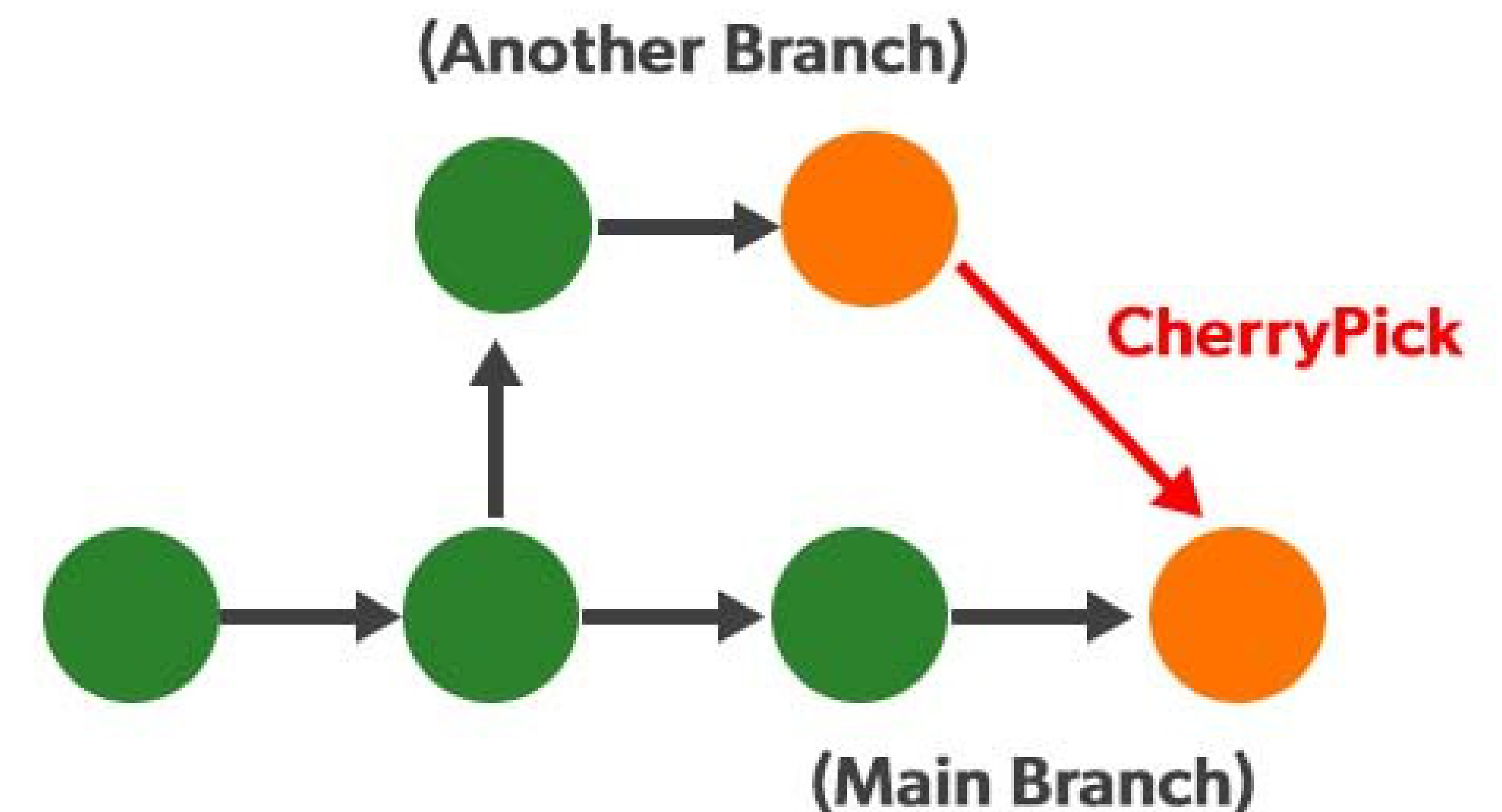
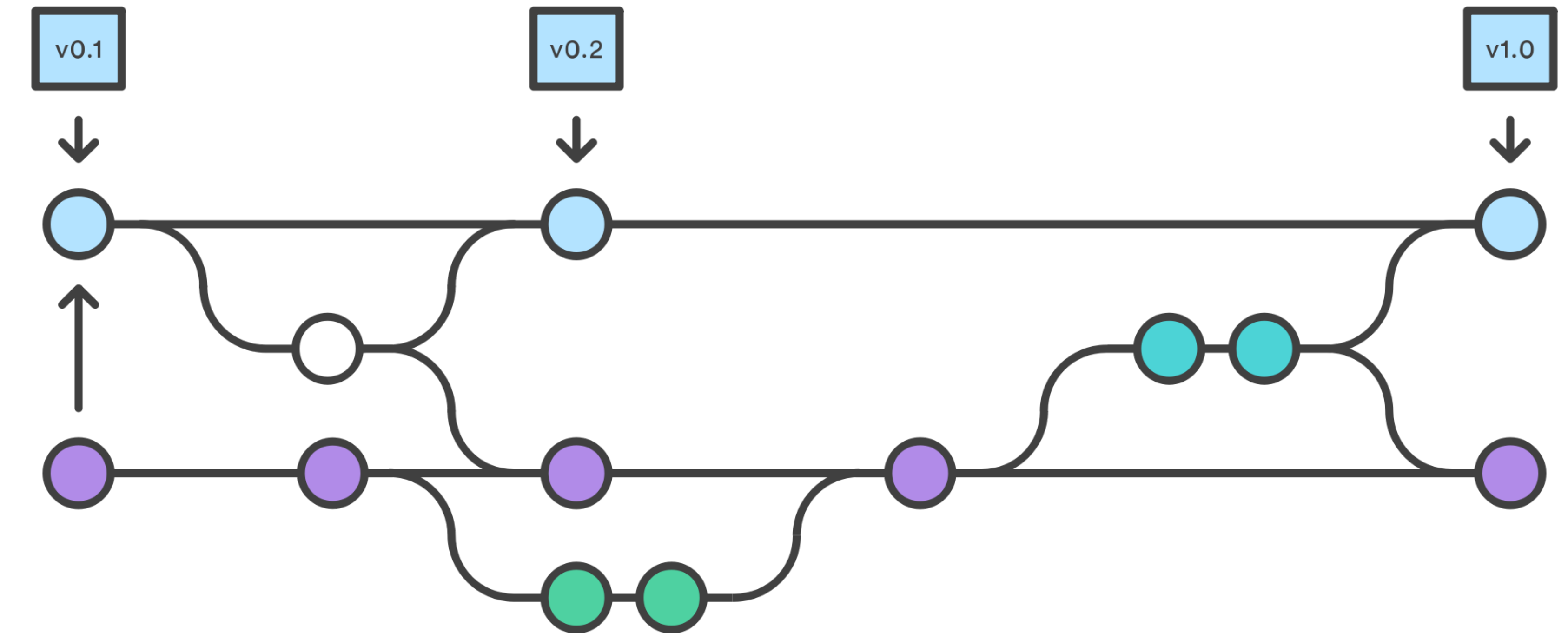
    PoC (one trapnest) works

commit 3d93154093bd9651d781af578dda907509bcb5e6
Author: Harald von Waldow <harald.vonwaldow@thuenen.de>
Date:   Wed Aug 24 13:07:59 2022 +0200

    new entryform: mod for single image
```

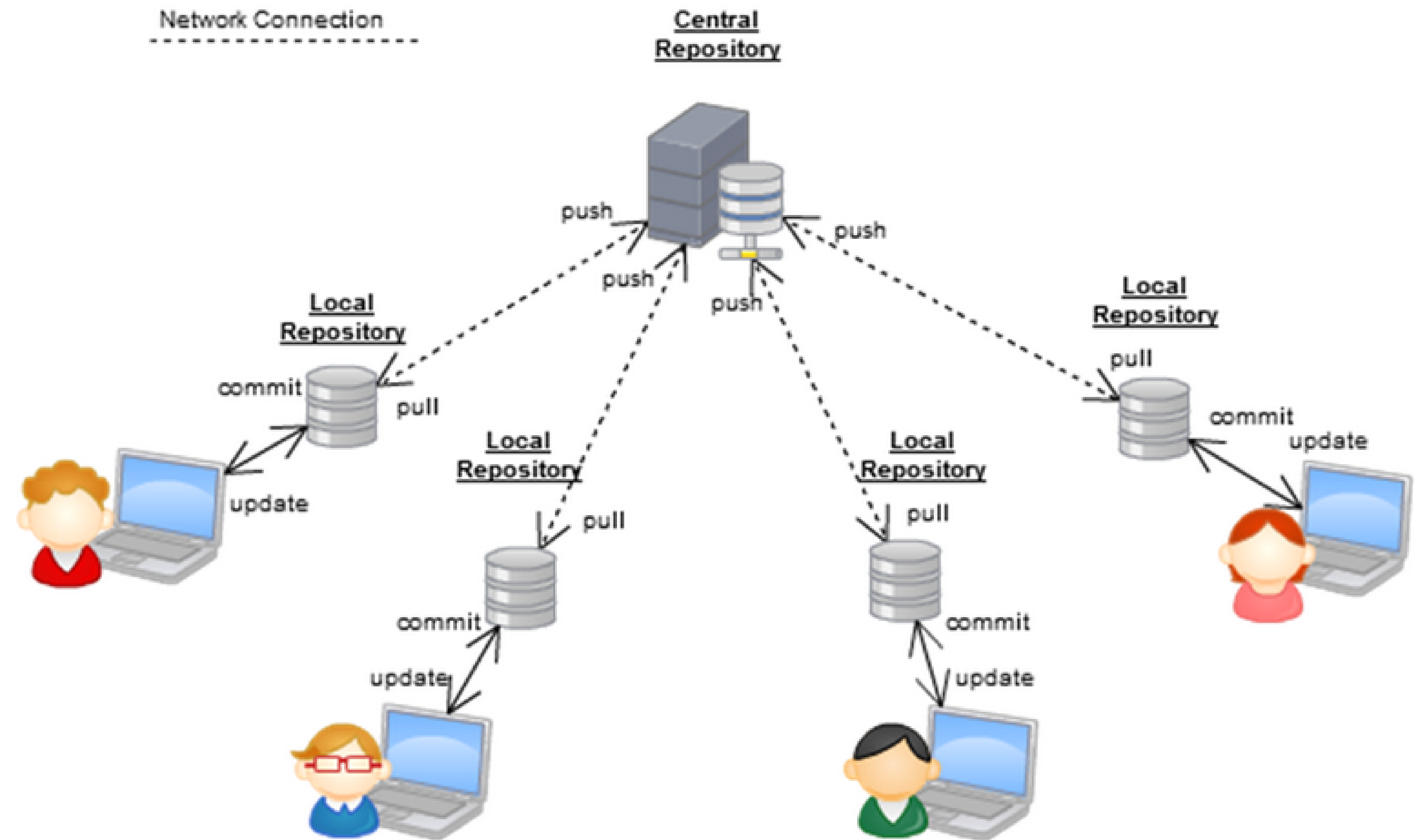
VERSION CONTROL

- Anti-data-loss design
- Allows for massive non-linear development in *branches*
- Easily see differences between files, branches, commits
- Track development through different branches
- Sophisticated mechanisms to pick & choose individual changes and sequences of changes



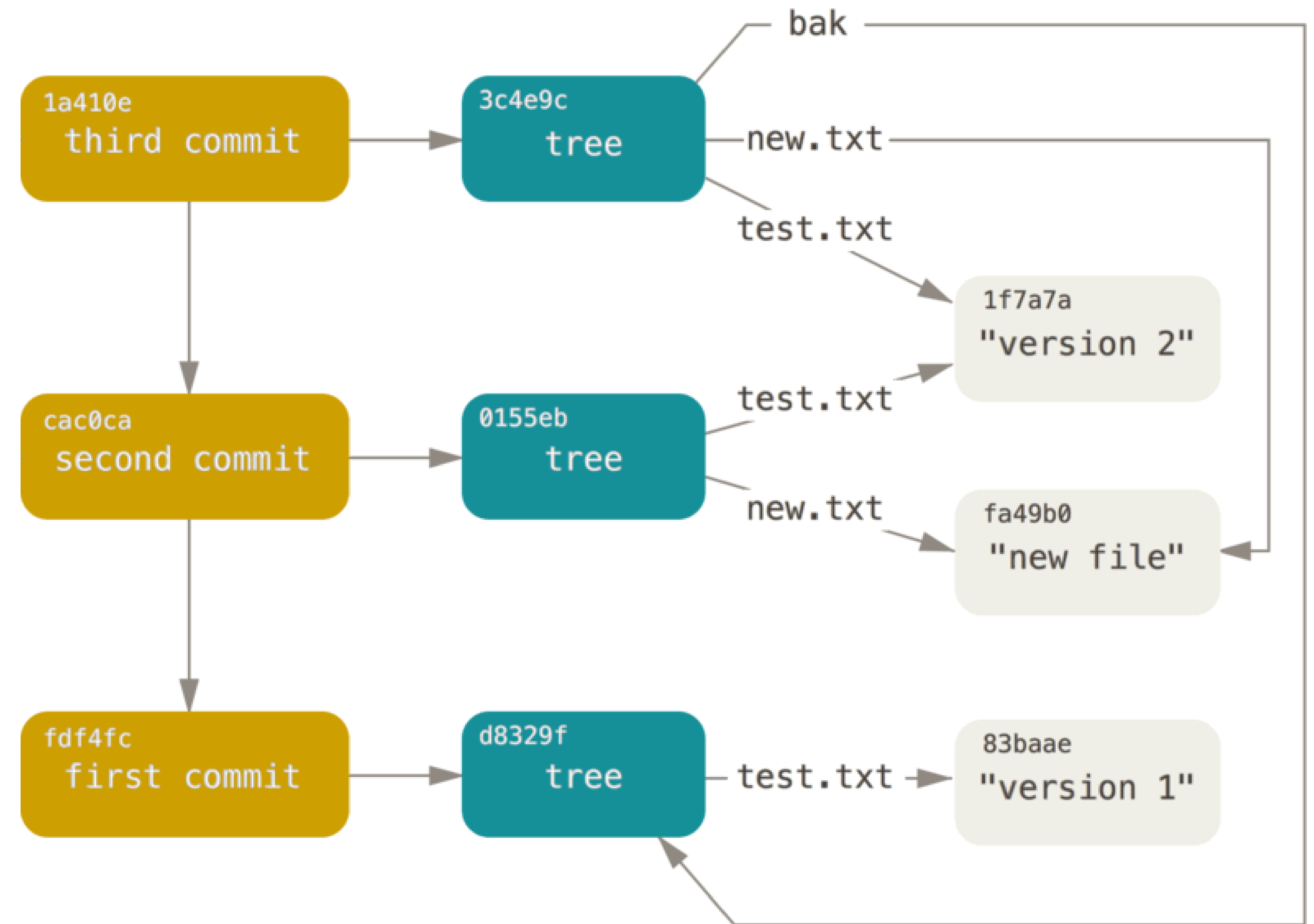
DISTRIBUTED & DECENTRALIZED

- Multiple complete copies of everything
- Work offline, e.g. no interruptions with shaky VPN or while travelling
- No "owner" of a central repository
→ forks are a non-issue



CONTENT-ADDRESSABLE

- Diffing across a large number of files & directories is **fast**.
- Changes in everything (files, tree, commits) is detected reliably and automatically.
- Identical blobs are only stored once



WHAT DOES IT DO TO YOU?

- Data safety
- Peace of mind, fosters creativity
- Provides a practical beat for your work (commit-messages)
- Ease of sharing (`clone`, `push`, `pull`)



WHAT IS NOT GIT

GITHUB.COM, GITLAB.COM



- **Commercial** SaaS offers
- Web-interface for Git **plus**
 - **Project management** (Milestones, Kanban Boards, ...)
 - **Collaboration** features (Issues, Discussion, Merge Requests, ...) → **Florian's presentation**
 - **DevOps** features (Cont. Integration, Cont. Delivery, ..) → **Winnie's presentation**

WHAT IS NOT GIT

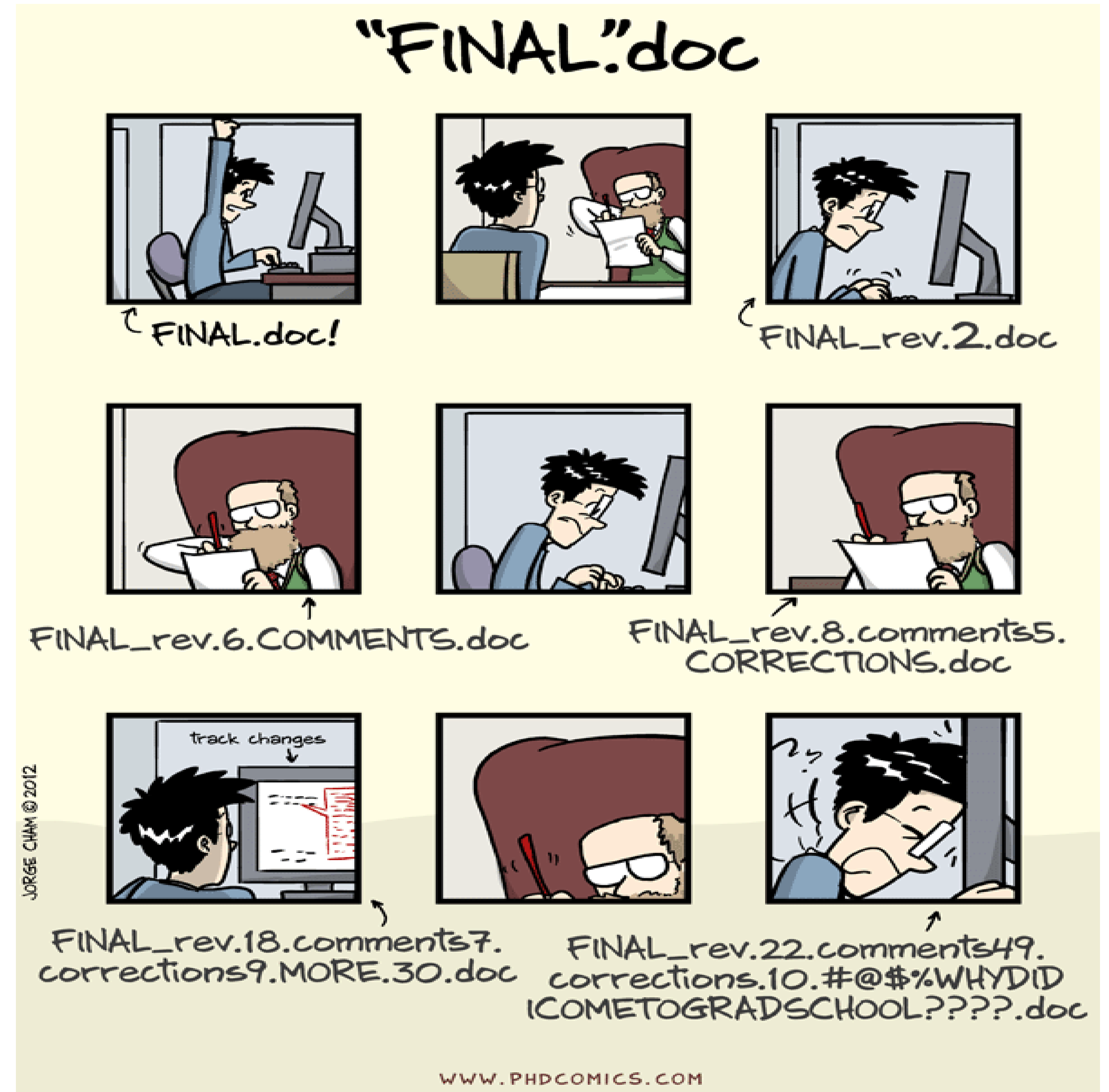
GITLAB COMMUNITY EDITION, GITEA



- Open Source products, on-premise installation
- GitLab is extremely widespread. Thünen Institute is running 2 instances (`git-int` and `git-dmz`)
- GitLab is becoming the de-facto standard platform for agile software development.
- GiTea has less features, is much more lightweight uses less ressources and is easier to configure and run.

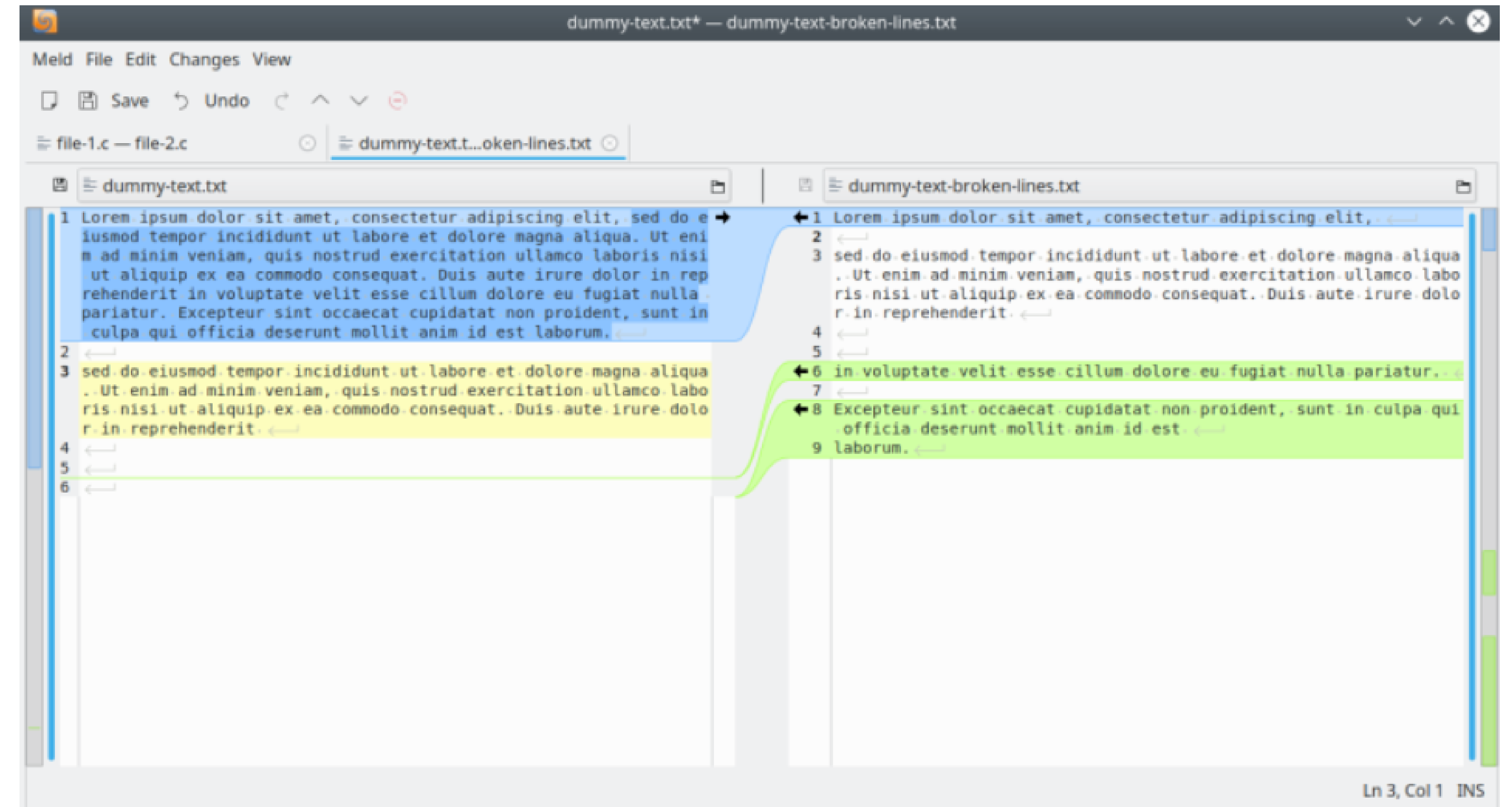
GIT FOR OTHER TASKS

- Can be used by any knowledge-worker who produces digital output
- E.g. to write articles, books, contracts
- E.g. as a backbone of sophisticated RDM-system (Renku)
- E.g. to power a distributed password-manager (pass)



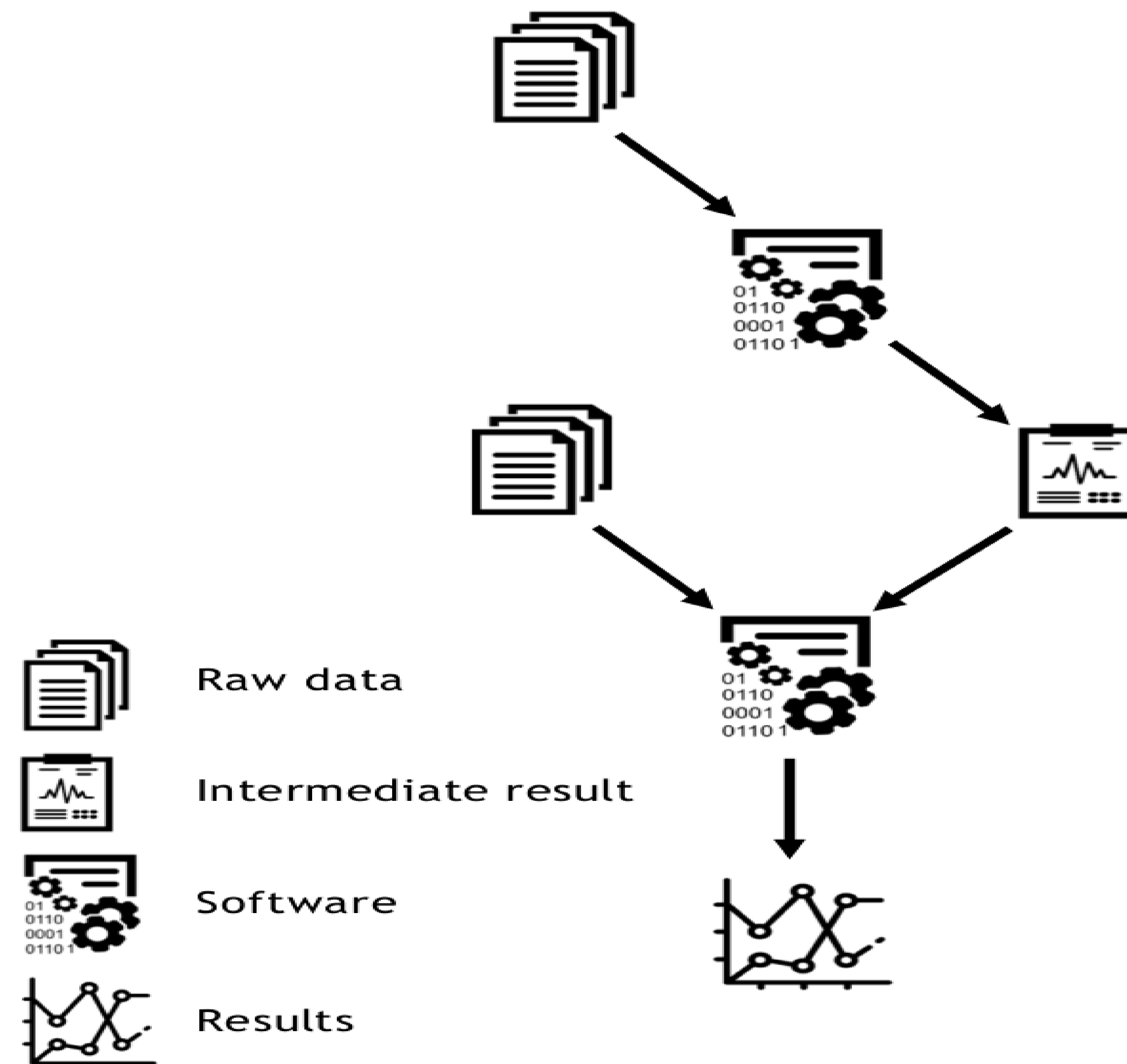
GIT FOR OTHER TASKS

- Can be used by any knowledge-worker who produces digital output
- E.g. to write articles, books, contracts
- E.g. as a backbone of sophisticated RDM-system (Renku)
- E.g. to power a distributed password-manager (pass)



GIT FOR OTHER TASKS

- Can be used by any knowledge-worker who produces digital output
- E.g. to write articles, books, contracts
- E.g. as a backbone of sophisticated RDM-system (Renku)
- E.g. to power a distributed password-manager (pass)



THIS IS GIT. IT TRACKS COLLABORATIVE WORK
ON PROJECTS THROUGH A BEAUTIFUL
DISTRIBUTED GRAPH THEORY TREE MODEL.

COOL. HOW DO WE USE IT?

NO IDEA. JUST MEMORIZE THESE SHELL
COMMANDS AND TYPE THEM TO SYNC UP.
IF YOU GET ERRORS, SAVE YOUR WORK
ELSEWHERE, DELETE THE PROJECT,
AND DOWNLOAD A FRESH COPY.



THÜNEN