Fachhochschule Kiel
University of Applied Sciences

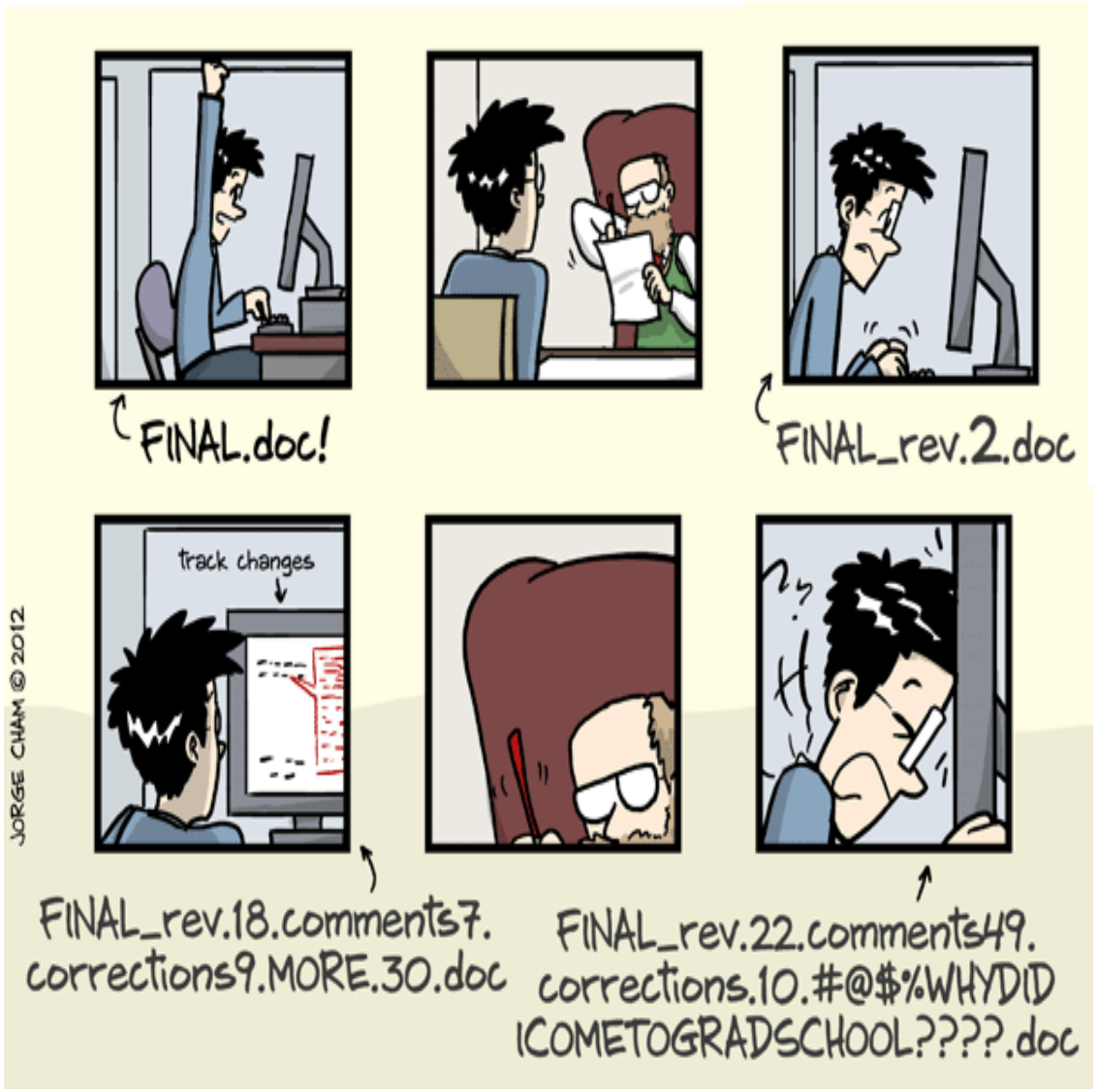
Tools and Programming for Data Science
Version Control with Git and GitHub

Study Program Data Science
Prof. Dr. Tillmann Schwörer

What is a version control system?

Fachhochschule Kiel
University of Applied Sciences

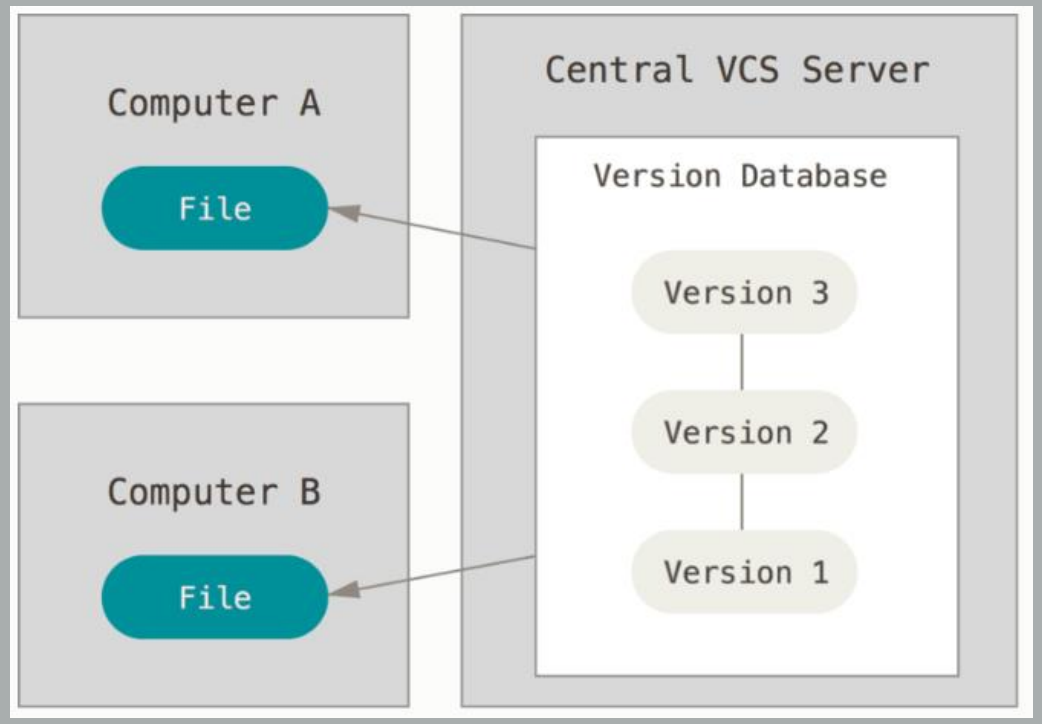
- System to manage changes in code or other documents
  - Backup: undo changes, restore files, safely experiment
  - Transparency: who changed? what? when?
  - Collaboration: work with others on the same project
- Git: open source, distributed version control system



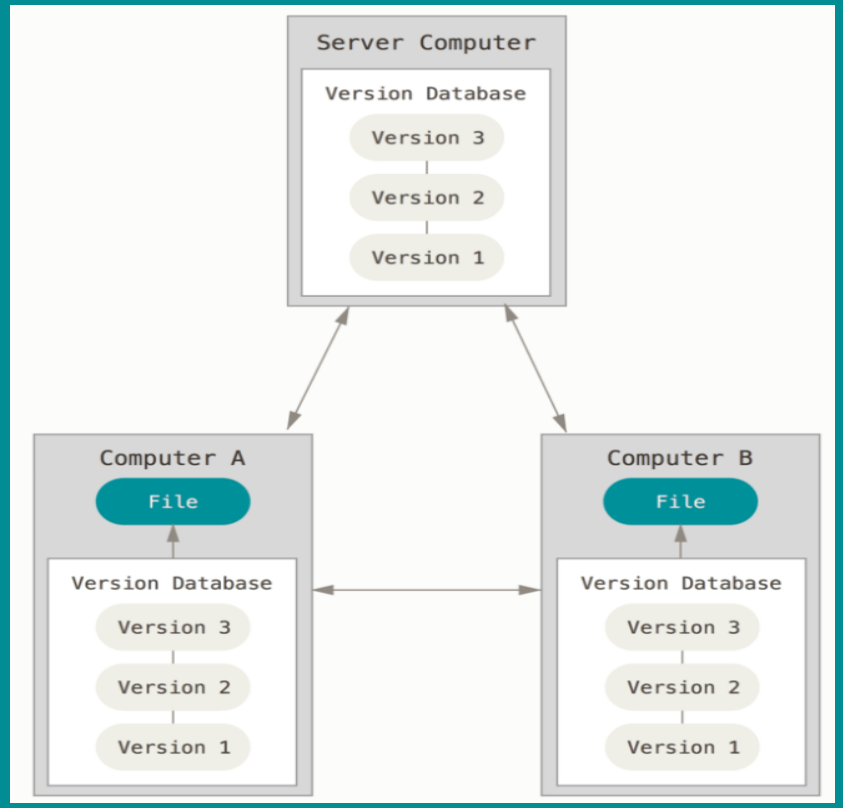
Centralized vs distributed

Fachhochschule Kiel
University of Applied Sciences

Centralized: single files are pulled from or pushed to central repository (→ **Subversion**)



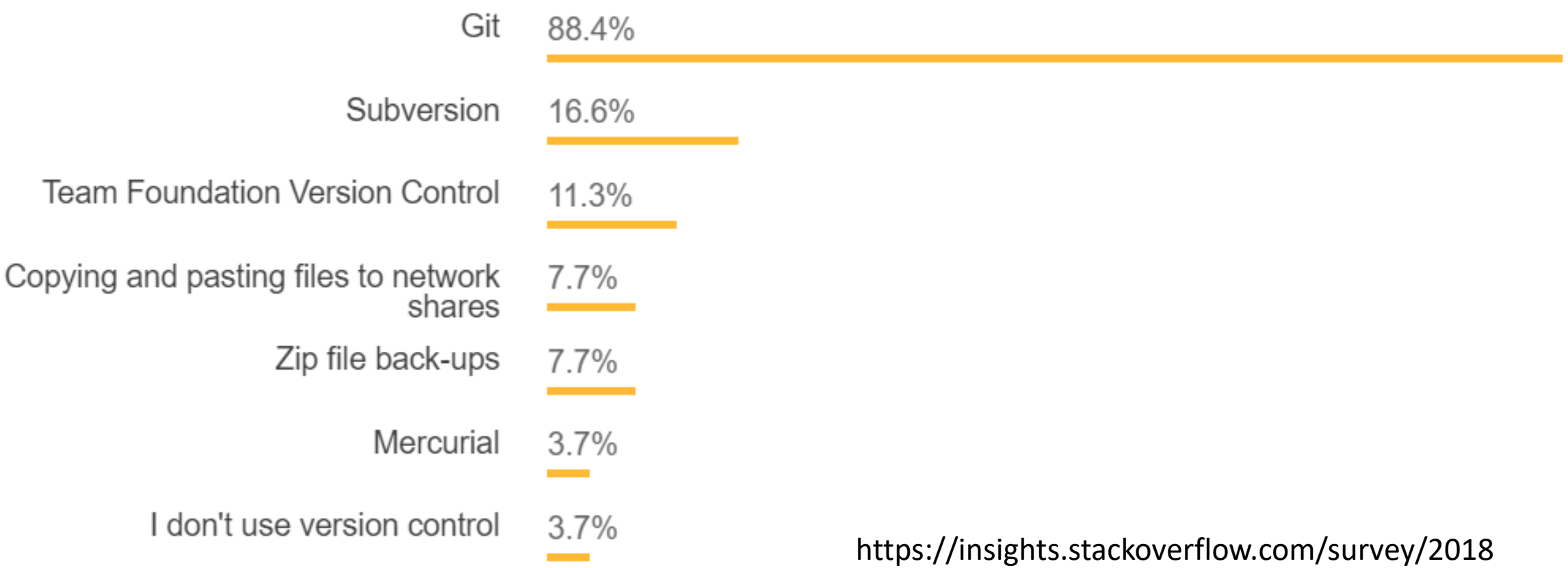
Distributed: Each client has full copy of entire repository (→ **git**)



Most popular version control systems

Fachhochschule Kiel
University of Applied Sciences

All Respondents | Professional Developers

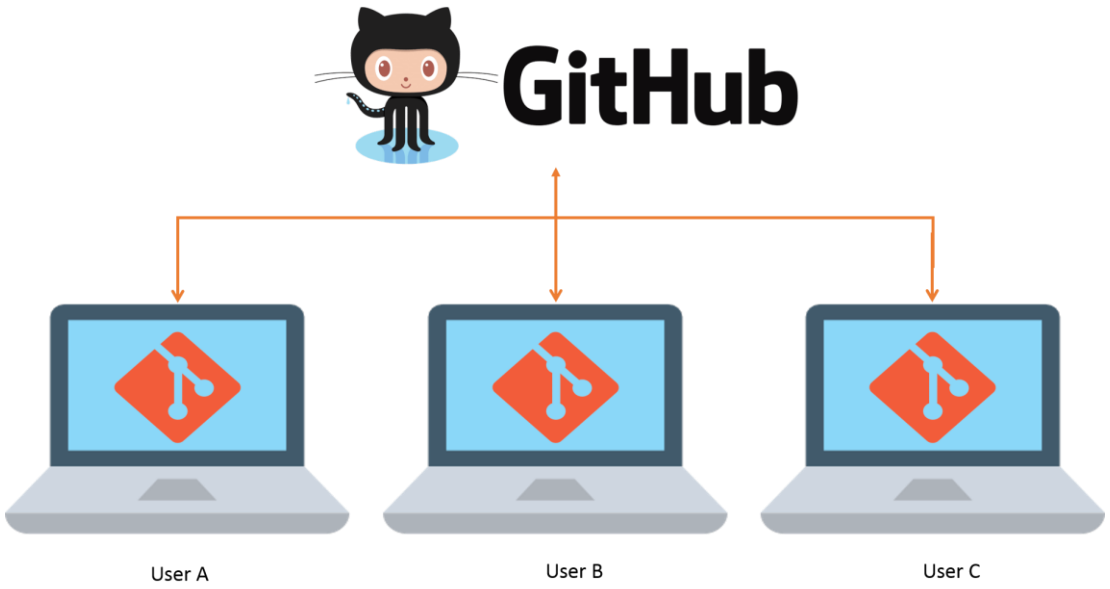


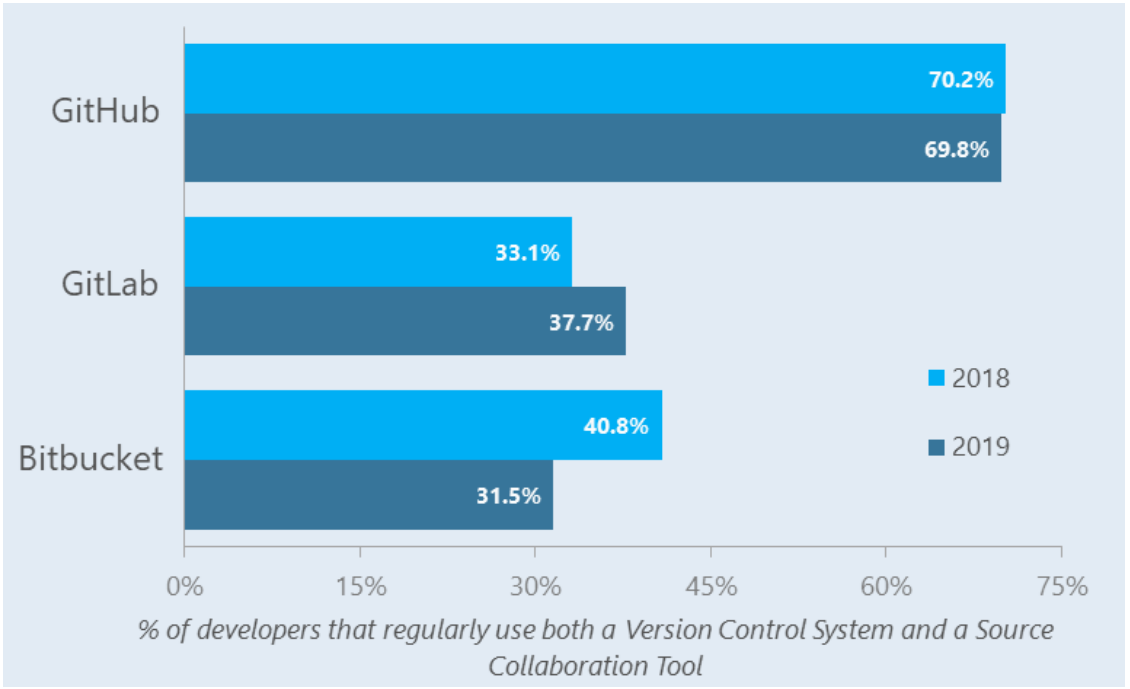
https://insights.stackoverflow.com/survey/2018

How is GitHub different from Git?

Fachhochschule Kiel
University of Applied Sciences

- Hosting service for Git repositories
- Collaboration plattform: bug tracking, feature requests, task management, wikis, etc.





Example

Fachhochschule Kiel
University of Applied Sciences

Chaos Computer Club: Entdecke unsere Verfassung

Johannes Rau, Bundespräsident committed on 26 Jul 2002

1 parent dc4dd4d commit b17a4b2848dac62d37618ebc92c06cccce5385ff

Showing 1 changed file with 1 addition and 1 deletion.

Unified Split

2 020a.md

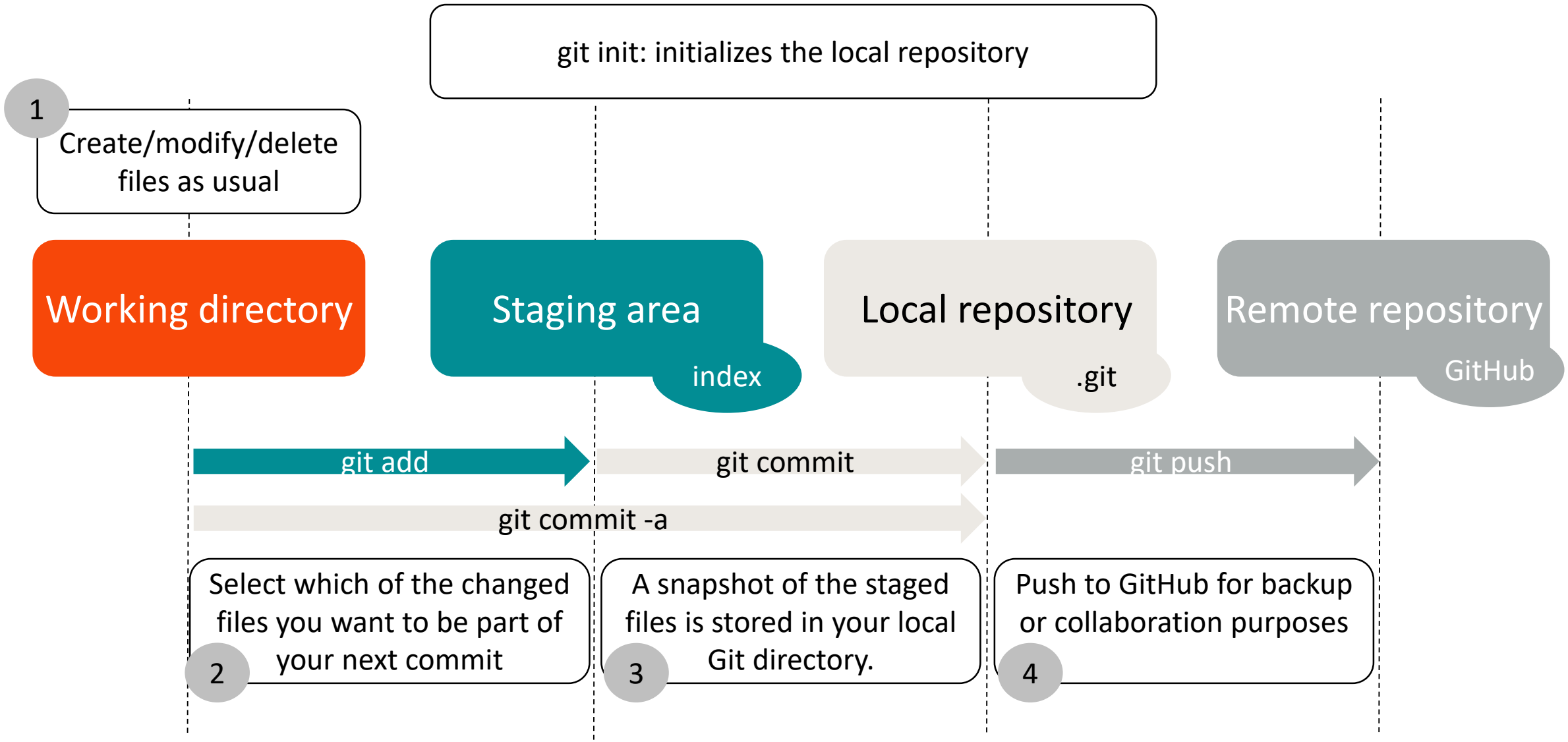
```
@@ -1,4 +1,4 @@
 ## Artikel 20a

- Der Staat schützt auch in Verantwortung für die künftigen Generationen die natürlichen Lebensgrundlagen im Rahmen der verfassungsmäßigen Ordnung durch die Gesetzgebung und nach Maßgabe von Gesetz und Recht durch die vollziehende Gewalt und die Rechtsprechung.
+ Der Staat schützt auch in Verantwortung für die künftigen Generationen die natürlichen Lebensgrundlage und die Tiere m Rahmen der verfassungsmäßigen Ordnung durch die Gesetzgebung und nach Maßgabe von Gesetz und Recht durch die vollziehende Gewalt und die Rechtsprechung.
```

Our Git and Github Agenda

Fachhochschule Kiel
University of Applied Sciences

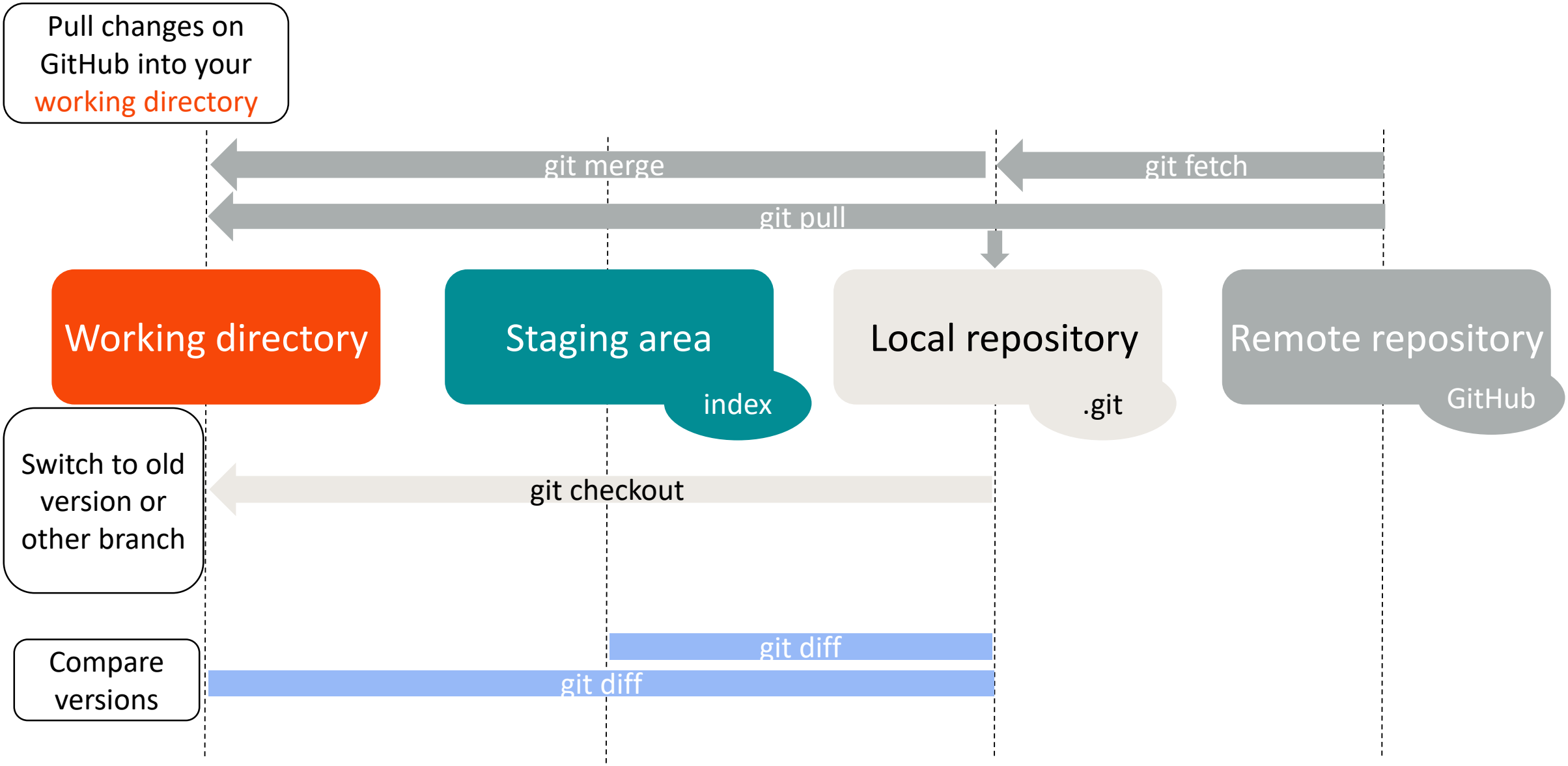
Standard Git Workflow

Fachhochschule Kiel
University of Applied Sciences



Standard Git Workflow

Fachhochschule Kiel
University of Applied Sciences

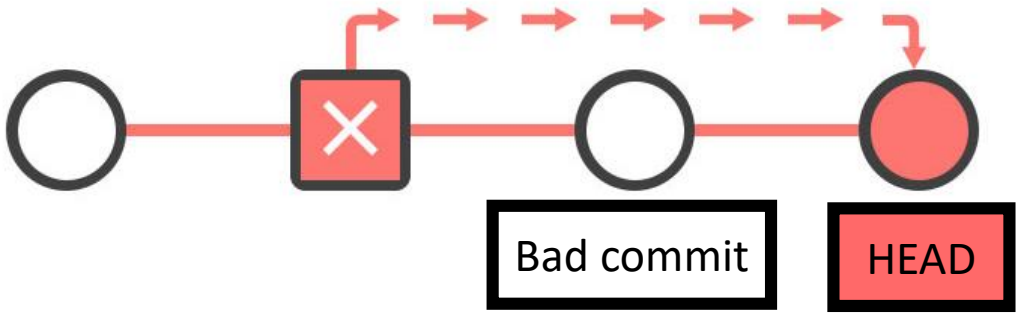


Undo changes

Fachhochschule Kiel
University of Applied Sciences

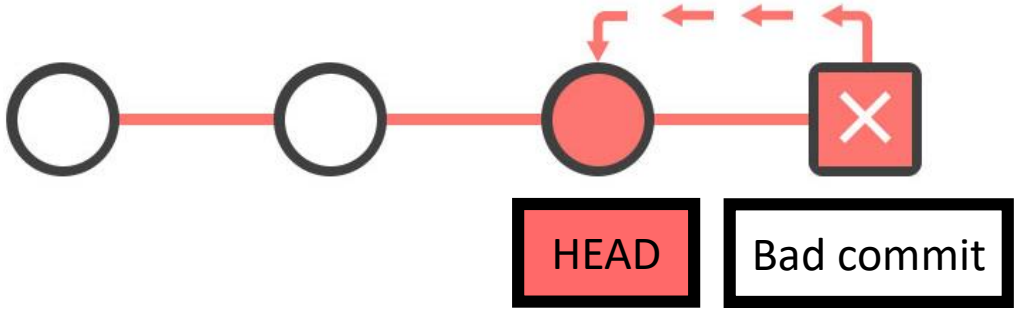
Revert (*#bad-commit*)

- Adds a new commit which reverts the „bad commit“
- Leaves history intact
- Can be safely done, even if you have already pushed to GitHub



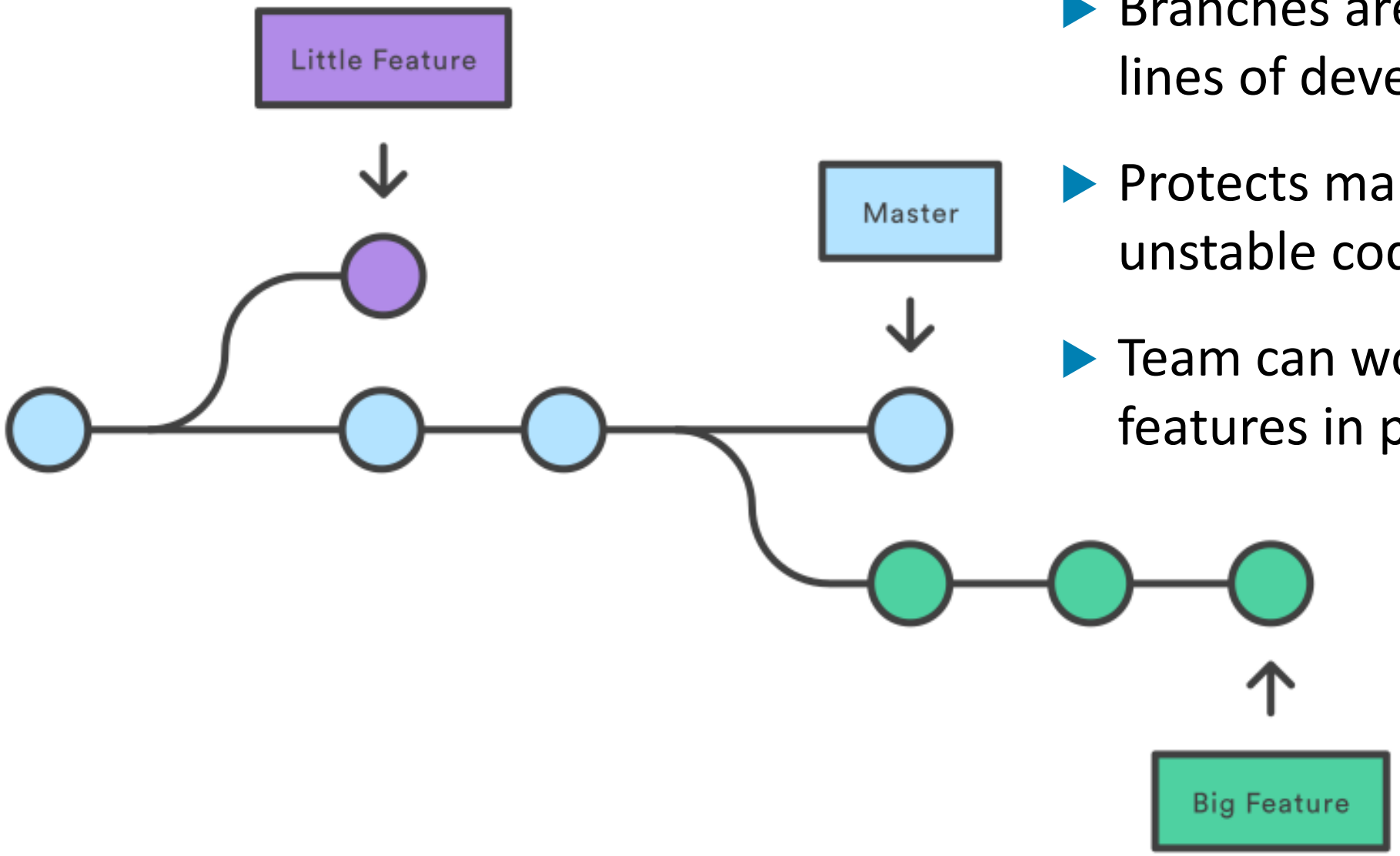
Reset (*#last-good-commit*)

- Undo the „bad commit“
- Rewrites history → use only for local changes!
- Versions:
  - reset hard: resets irreversably
  - reset mixed (default): rewrites commit history, but keeps changes in working directory



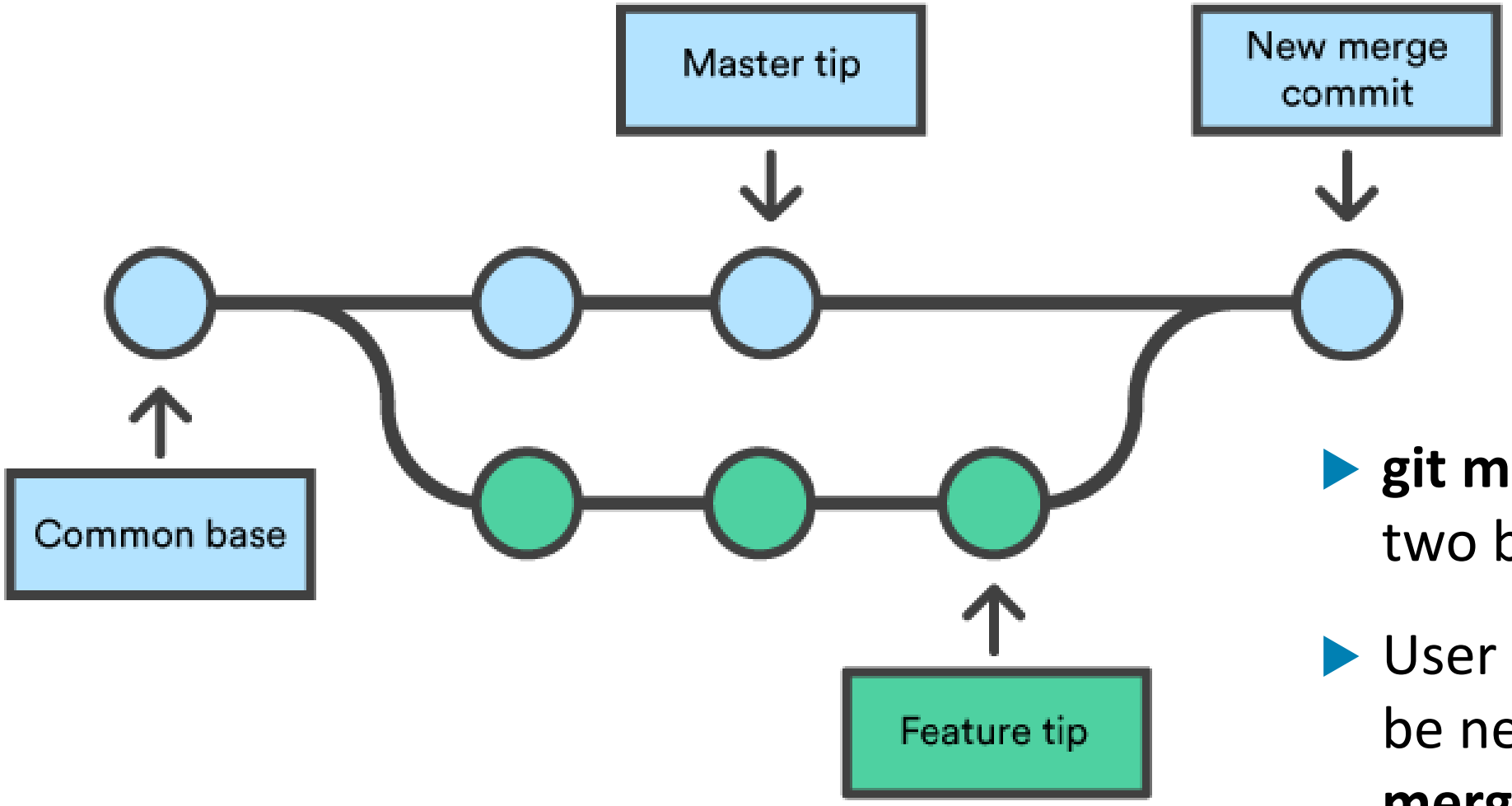
Branching

Fachhochschule Kiel
University of Applied Sciences



- Branches are independent lines of development
- Protects main branch against unstable code
- Team can work on separate features in parallel

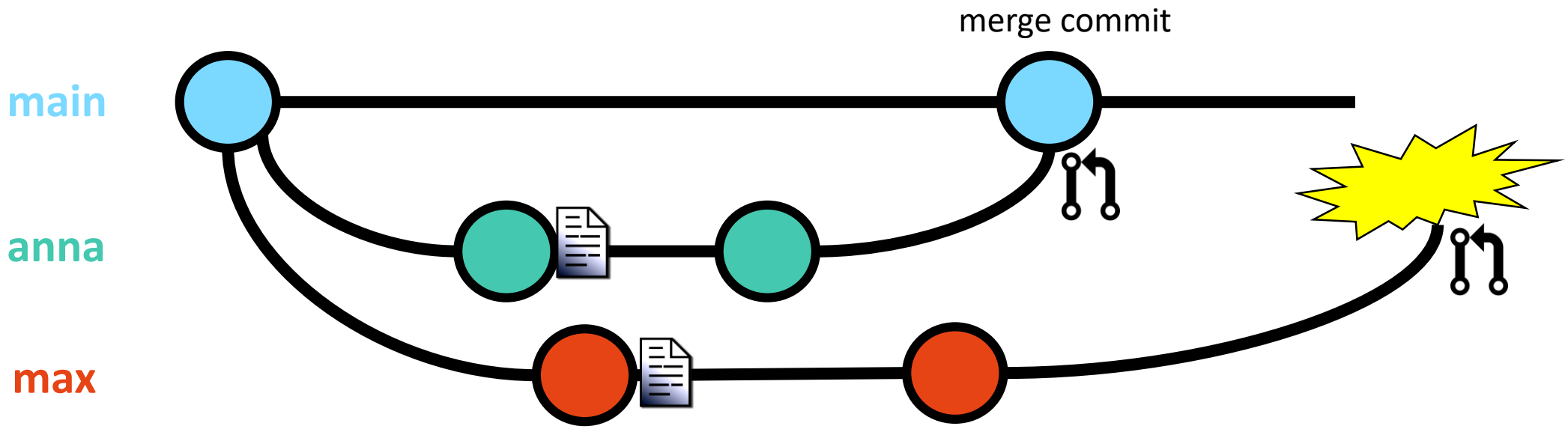
Merging



- **git merge** combines two branches into one.
- User intervention may be needed to resolve **merge conflicts**

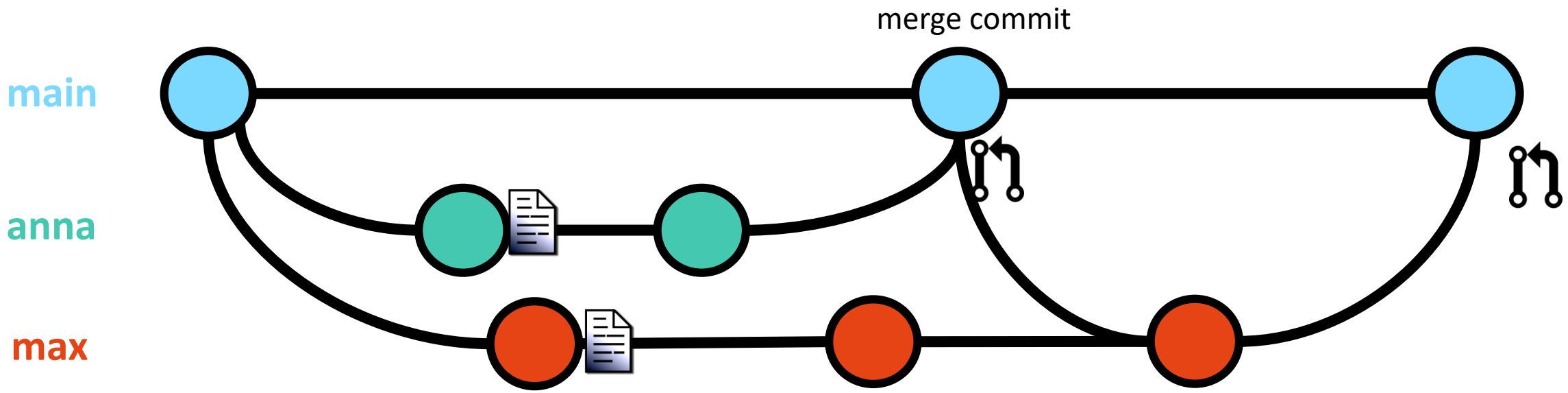
Handling merge conflicts

Fachhochschule Kiel
University of Applied Sciences



Handling merge conflicts

Fachhochschule Kiel
University of Applied Sciences



Option 1: Before max opens a pull request, he checks for changes in main and merges them into his feature branch (dealing manually with any conflict)

Handling merge conflicts

Fachhochschule Kiel
University of Applied Sciences

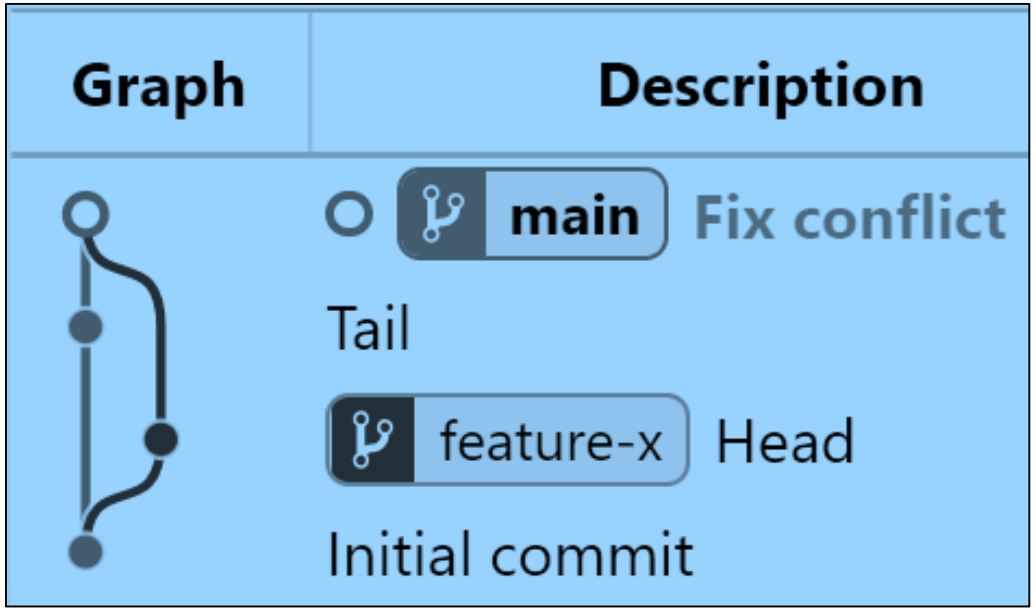
```
git merge feature-x
```

```
Auto-merging analysis.py
CONFLICT (content): Merge conflict in analysis.py
Automatic merge failed; fix conflicts and then commit the result.
```

```
Accept Current Change | Accept Incoming Change | Accept Both Changes | Compare Changes
<<<<<<< HEAD (Current Change)
df.tail()
=======
df.head()
>>>>>>> feature-x (Incoming Change)
```

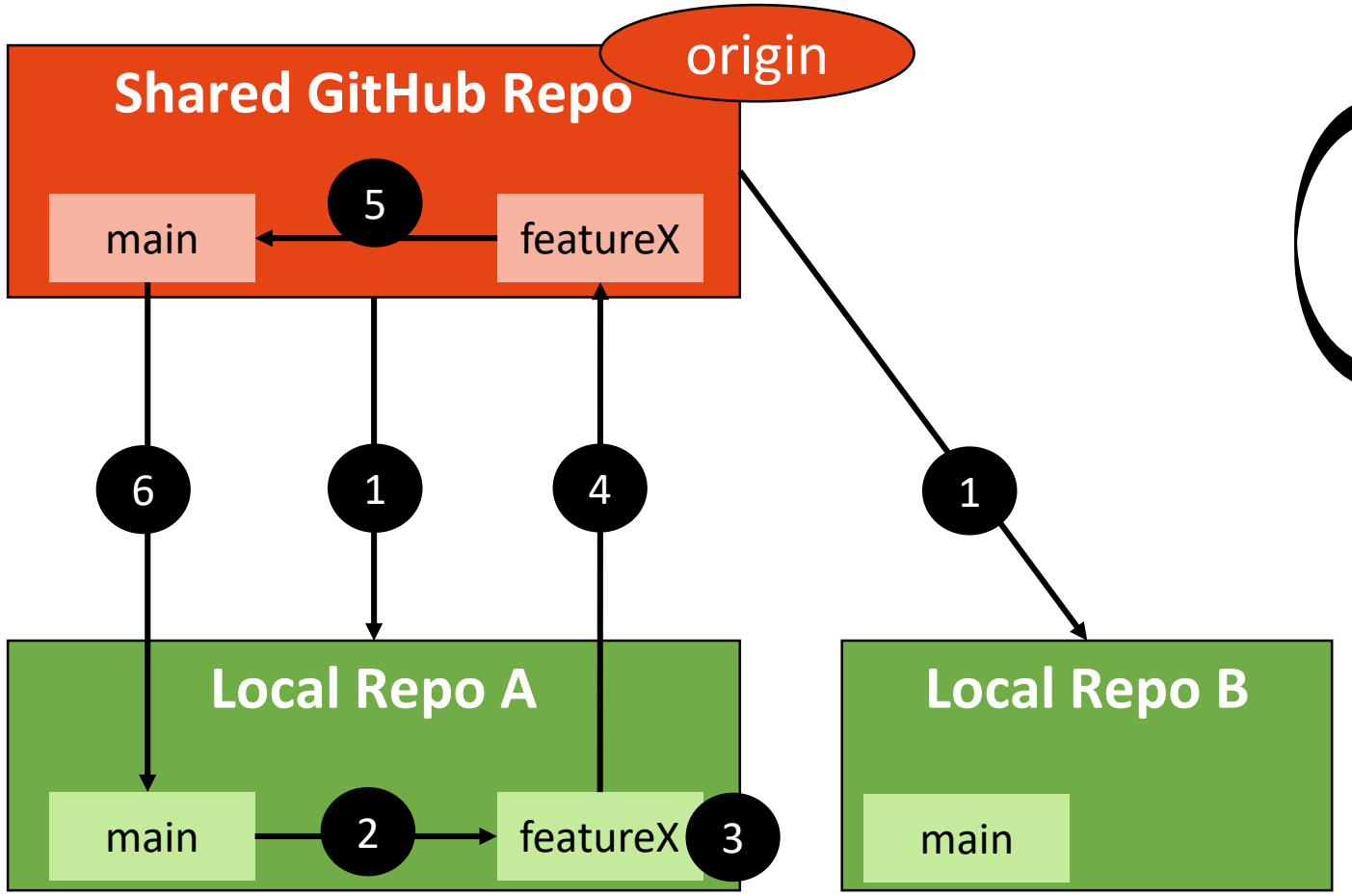
```
df.head()
df.tail()
```

```
git commit -m "Fix conflict"
```



Feature Branch Workflow

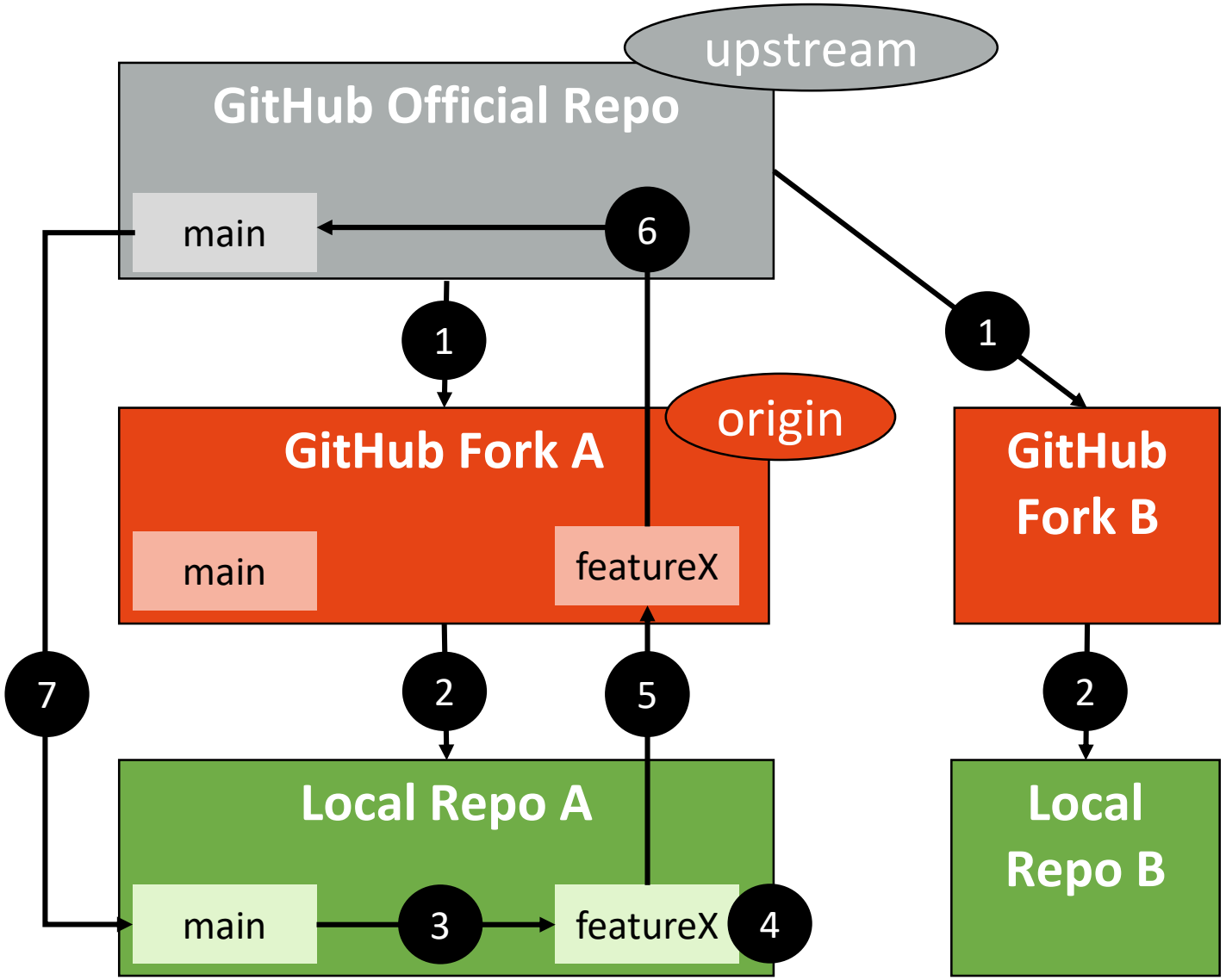
Company Setting



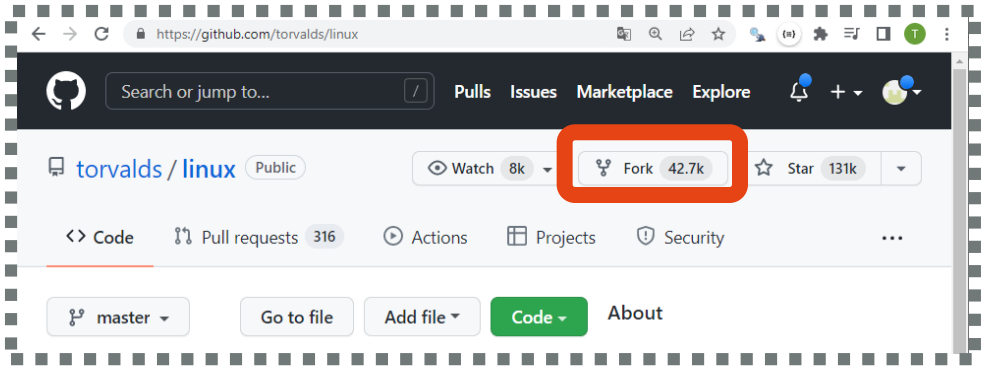
1. git clone <url>
2. git branch featureX
   git checkout featureX
3. edit, git add, git commit
4. git push
5. Pull Request → git merge
6. git pull

Forking Workflow

Open Source Setting



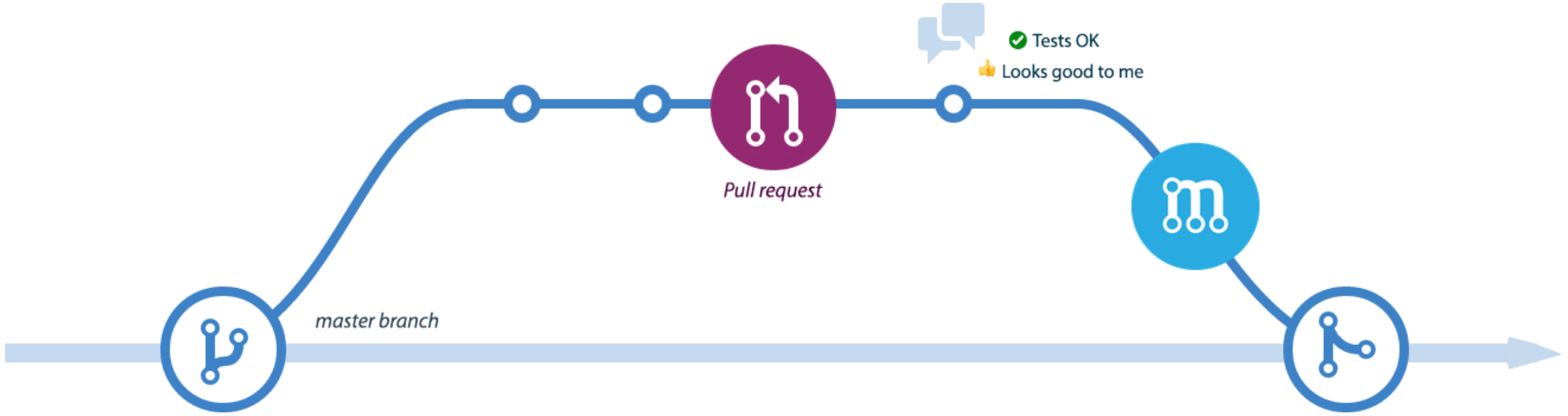
1. Fork
2. git clone <url>
3. git branch featureX
   git checkout featureX
4. edit, git add, git commit
5. git push
6. Pull Request → git merge
7. (git remote add upstream <url>)
   git pull upstream



Pull Request

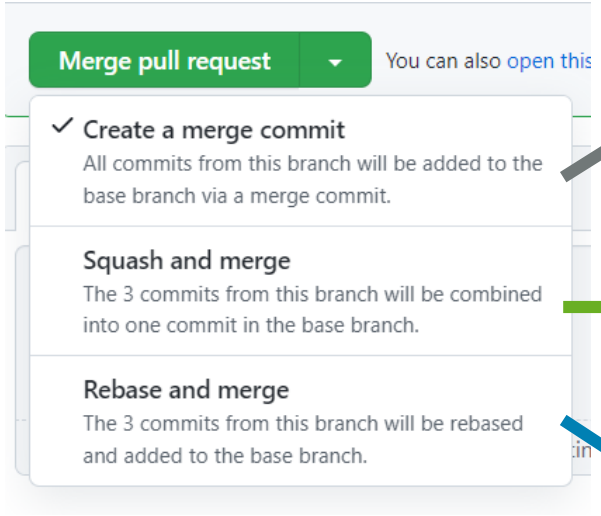
Fachhochschule Kiel
University of Applied Sciences

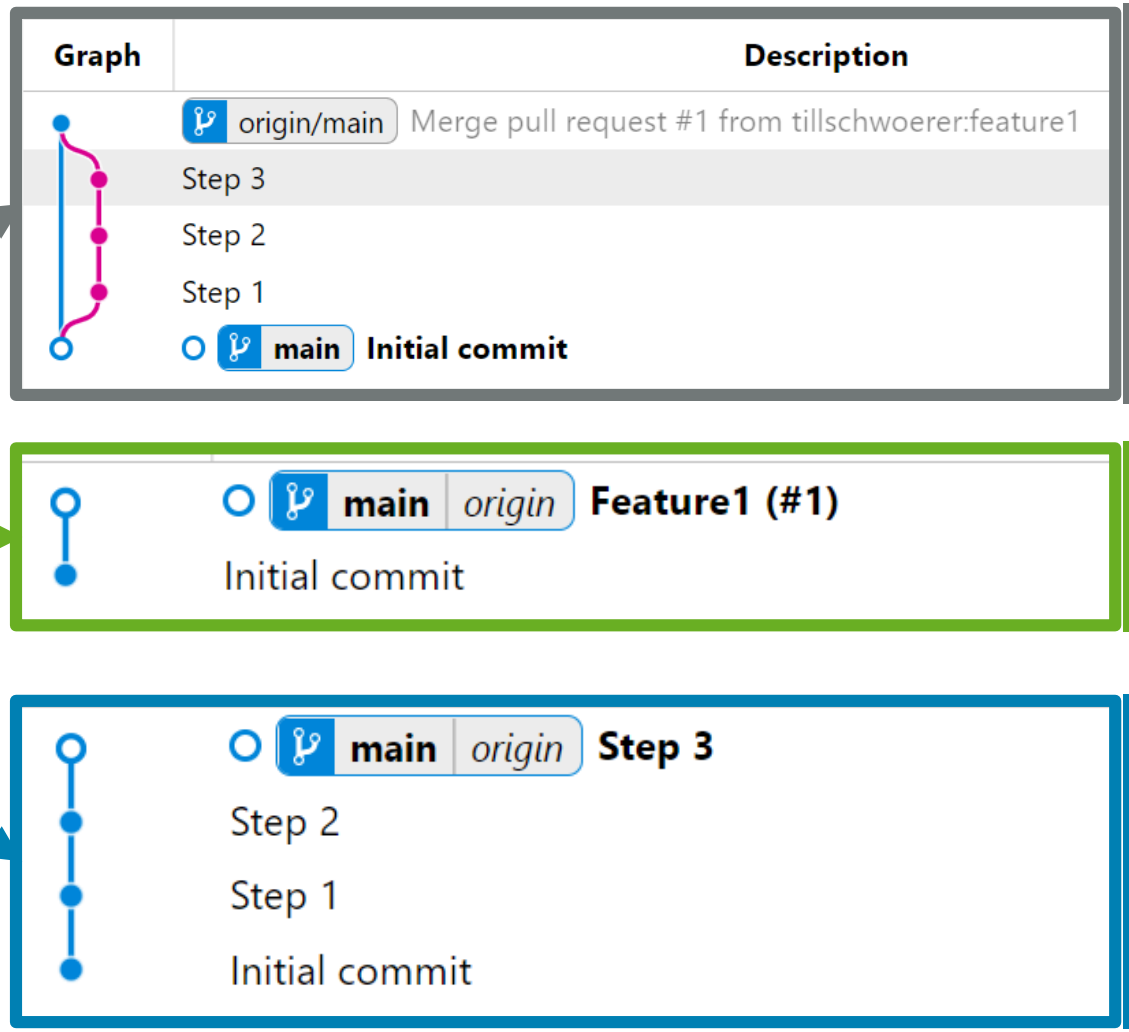
- **A Pull request is a GitHub feature**, not a git command: you request to pull changes from your feature branch (“merge request” on Gitlab)
- May involve multiple iterations of discussions, code reviews, and follow-up commits, before the commit is merged into the main branch



3 Ways of Merging in a Pull Request

Fachhochschule Kiel
University of Applied Sciences





+ Truthful history of actions
- May pollute project history

+ Short and linear project history
+ Pull requests can still be identified
- Actual single commits are not logged anymore

+ Linear project history
+ Single commits are all part of the history
- Feature branches / pull requests cannot be identified any more

Further aspects

Fachhochschule Kiel
University of Applied Sciences

- **.gitignore:**
  - define files that you don’t want to track
  - e.g. password files (.env), .ipynb_checkpoints, very large binary files
- **Large files**
  - GitHub file size limit: 100 MB
  - Git Large File Storage: tracking files up to 2 GB
  - Git LFS stores references to the file in the Github repository. The actual file is stored separately.

Further Resources

Fachhochschule Kiel
University of Applied Sciences

- Tutorial on Git (Workflows) (available in German)
- Git reference book (available in German)
- Git commands: cheat sheet
- 10-minute reads mostly on GitHub topics (by GitHub)
- Glossary of Git and GitHub terms (by Github)