The basics of Git & GitHub

2022-09-29 - Joran Martijn - Tips & Tricks ICG

LINUX

- servers
- mainframes
- embedded devices
- desktops (`ubuntu`, `debian` etc)
- computer clusters
- mobile devices (`android`)
- video game devices (`steamOS`)

Git

Linus Torvalds

Origin of the name 'git'

Git can mean 'unpleasant person' in British English slang

"I am an egotistical bastard and name all my products after myself. First Linux, now Git"

"I'd like to be a nice person and curse less and encourage people to grow rather than telling them they are idiots. I'm sorry. I tried, it's just not in me."

README of the source code:

"git" can mean anything, depending on your mood

- **Random three-letter combination that is pronounceable**, and not actually used by any common UNIX command. The fact that it is a mispronunciation of "get" may or may not be relevant
- **Stupid. Contemptible and despicable**. Simple. Take your pick from the dictionary of slang
- "**Global information tracker**": you're in a good mood, and it actually works for you. Angels sing, and a light suddenly fills the room
- "**Goddamn idiotic truckload of sh*t**": when it breaks

Git

- A **Source Control Management** system
- Linus and team used **BitKeeper** - a proprietary software - to maintain the LINUX code base since 2002
- In 2005, Larry Mcvoy **discontinued free use of BitKeeper** after claiming that Andrew Tridgell created SourcePuller by reverse engineering BitKeeper
- Linus tried to find another solution that met his criteria, but found none
- He **started development of git** in April 3rd 2005
  On June 16th, git managed the release of the next LINUX version
- **git 1.0** was released on 21 December 2005
- git is now the dominant source control management system out there

What does it do?

git keeps track of changes of plain text files

- *What* was changed?
- *When* was it changed?
- *Why* was it changed?
- *Who* changed it?

Example of tracked changes

Showing 1 changed file with 8 additions and 12 deletions.

```
12     U1=${base}/eb_dna_fw_unp.fastq.gz
13     U2=${base}/eb_dna_rv_unp.fastq.gz
14   - ASSEMBLY='canu_medaka_polish.fasta'
15
16     # output
17   - PREFIX=dnaseq_vs_canu_medaka_polish
18
19     # load bwa-mem2 software
20     source activate bwa-mem2
       @@ -23,19 +23,15 @@ source activate bwa-mem2
23     ## first index
24     bwa-mem2 index -p ${ASSEMBLY%.fasta} $ASSEMBLY
25     ## then align paired reads
26   - bwa-mem2 mem -t 20 ${ASSEMBLY%.fasta} $R1 $R2 > ${PREFIX}-pe.sam
27     ## and unpaired reads
28   - bwa-mem2 mem -t 20 ${ASSEMBLY%.fasta} $U1 > ${PREFIX}-se1.sam
29   - bwa-mem2 mem -t 20 ${ASSEMBLY%.fasta} $U2 > ${PREFIX}-se2.sam
30
31     conda deactivate
32
33     # convert to bam
34   - for SAM in *.sam; do
35   -     samtools view --threads 20 -b $SAM > ${SAM/sam/bam}
36   -     samtools sort --threads 20 ${SAM/sam/bam} > ${SAM/sam/sorted.bam}
37   -     samtools index -@ 20 ${SAM/sam/sorted.bam} ${SAM/sam/sorted.bam.bai}
38   -     rm $SAM ${SAM/sam/bam}
39     done
40
```

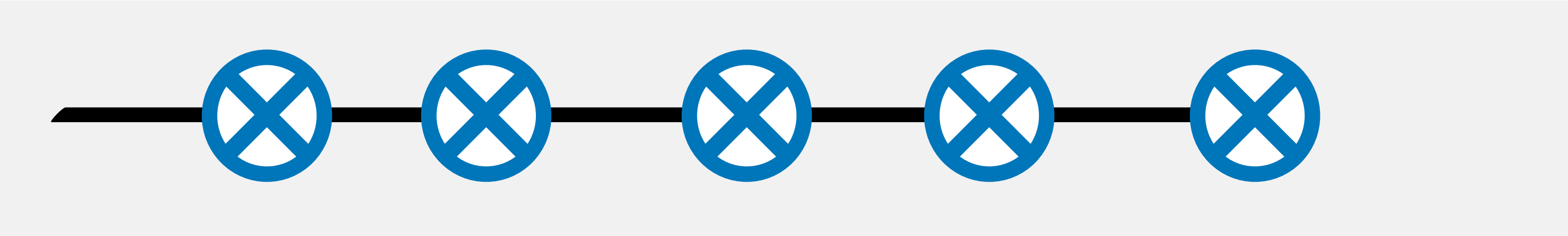
```
12     U1=${base}/eb_dna_fw_unp.fastq.gz
13     U2=${base}/eb_dna_rv_unp.fastq.gz
14   + ASSEMBLY='canu_ergo_contigs.fasta'
15
16     # output
17   + PREFIX=dnaseq_vs_canu_medaka_pilon_polish
18
19     # load bwa-mem2 software
20     source activate bwa-mem2

23     ## first index
24     bwa-mem2 index -p ${ASSEMBLY%.fasta} $ASSEMBLY
25     ## then align paired reads
26   + bwa-mem2 mem -t 40 ${ASSEMBLY%.fasta} $R1 $R2 | samtools sort --threads 40 -o
       ${PREFIX}-pe.sort.bam -
27     ## and unpaired reads
28   + bwa-mem2 mem -t 40 ${ASSEMBLY%.fasta} $U1 | samtools sort --threads 40 -o ${PREFIX}-
       se1.sort.bam -
29   + bwa-mem2 mem -t 40 ${ASSEMBLY%.fasta} $U2 | samtools sort --threads 40 -o ${PREFIX}-
       se2.sort.bam -
30
31     conda deactivate
32
33     # convert to bam
34   + for BAM in *.bam; do
35   +     samtools index -@ 40 $BAM $BAM.bai

36     done
37
```

old version

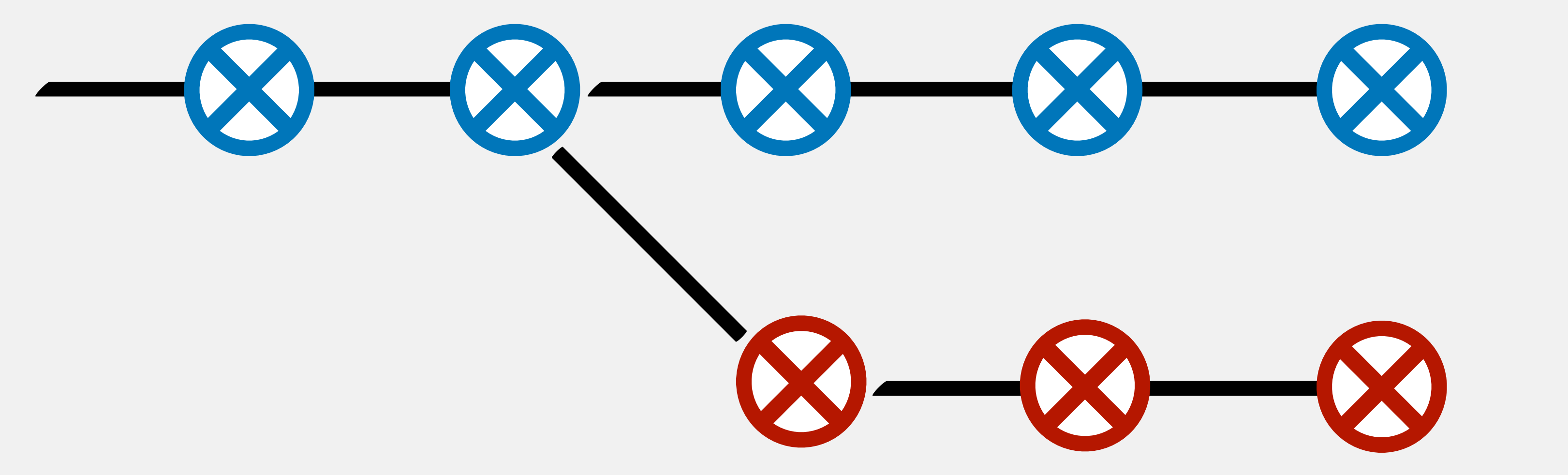
new version

"Commits"	When	What	Why	Who
`b58bf20`	August 1	Addition of a new gene	Software missed a gene	Jason
`1d6271b`	August 2	End position of gene A	No support for original end position	Jason
`4acf3e4`	August 5	Removal of an intron	Intron was not supported	Jason
`d28027f`	August 6	Addition of intron	Intron was supported after all	Jason
`6d7f098` `(HEAD)`	August 7	Coordinates of many genes	Script to fix many software mistakes	Joran

A series of commits is called a **branch**



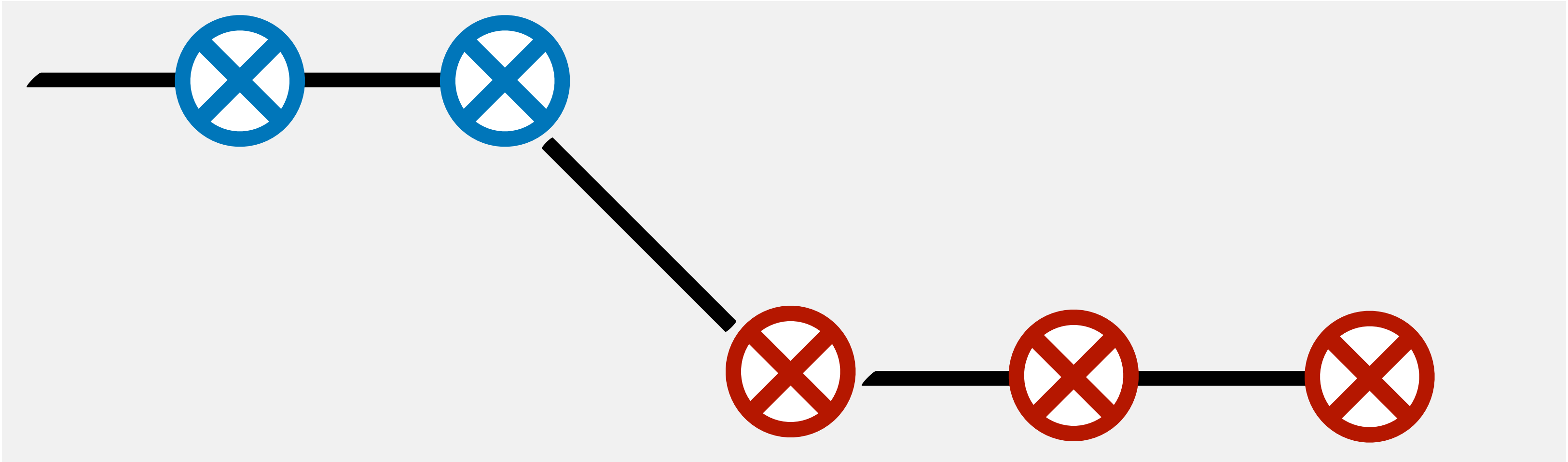
Git allows you to create additional branches



Commits in the **alternative branch** do not affect the original branch, i.e. the **master branch** or main branch

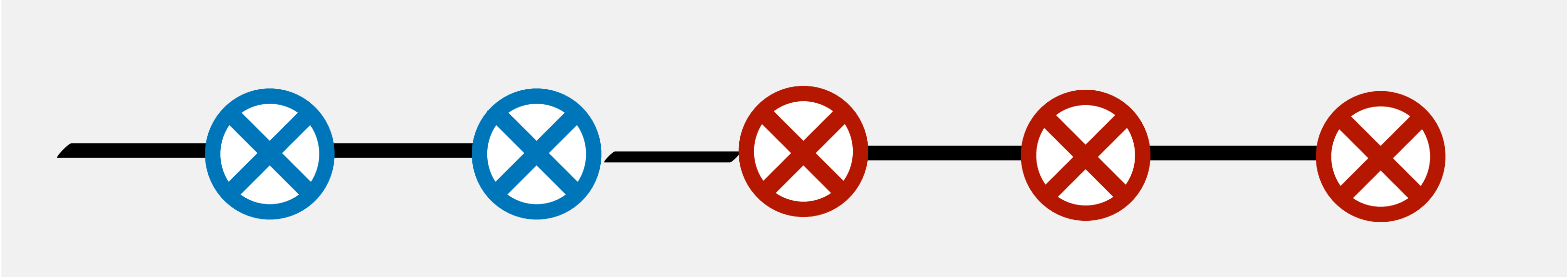
Alternative branches are useful for developing experimental features of code, or doing something potentially risky

Alternative branches can be **merged** back into the *master branch*



master branch

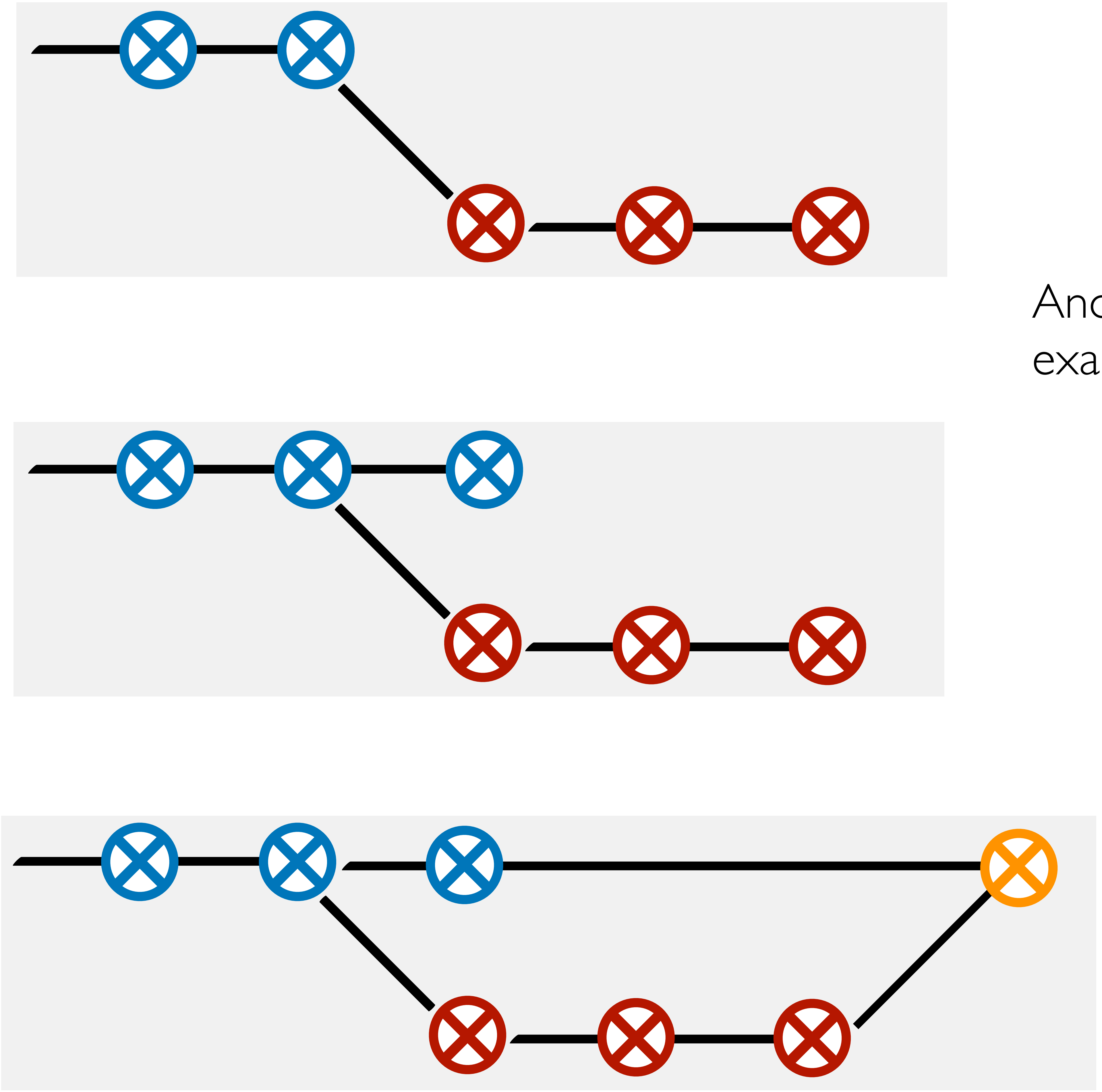
alternative branch



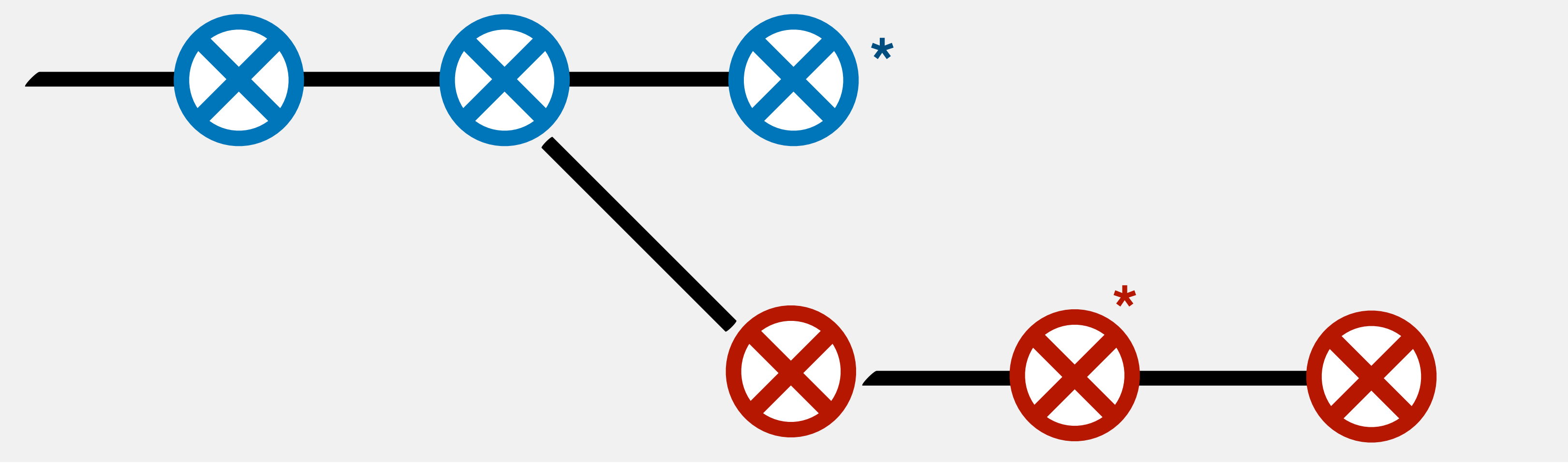
master branch

This is an example of a **fast-forward** merge

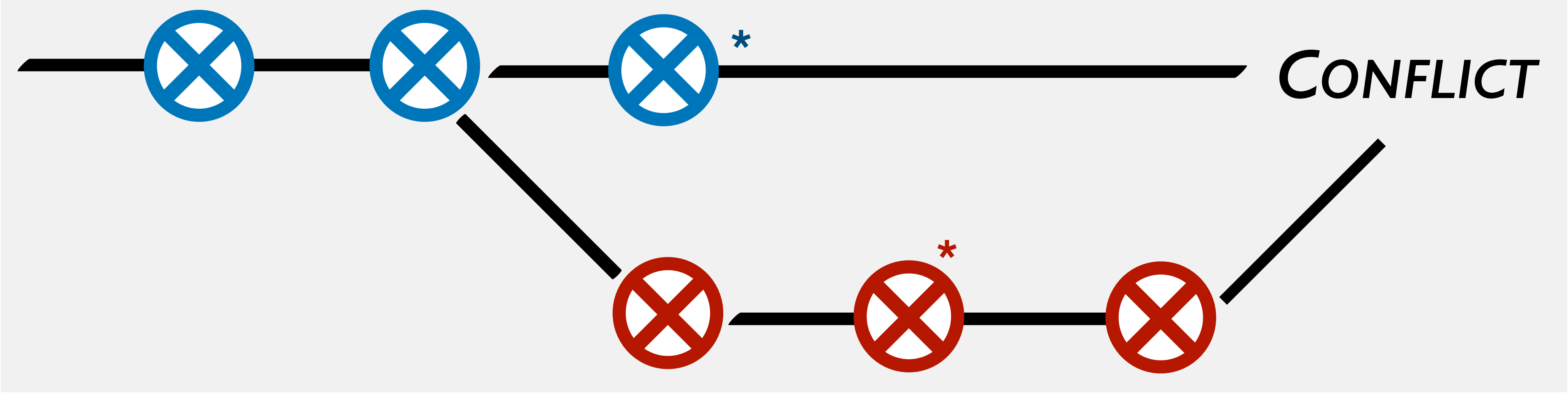
Another merging example



A **'merge commit'**

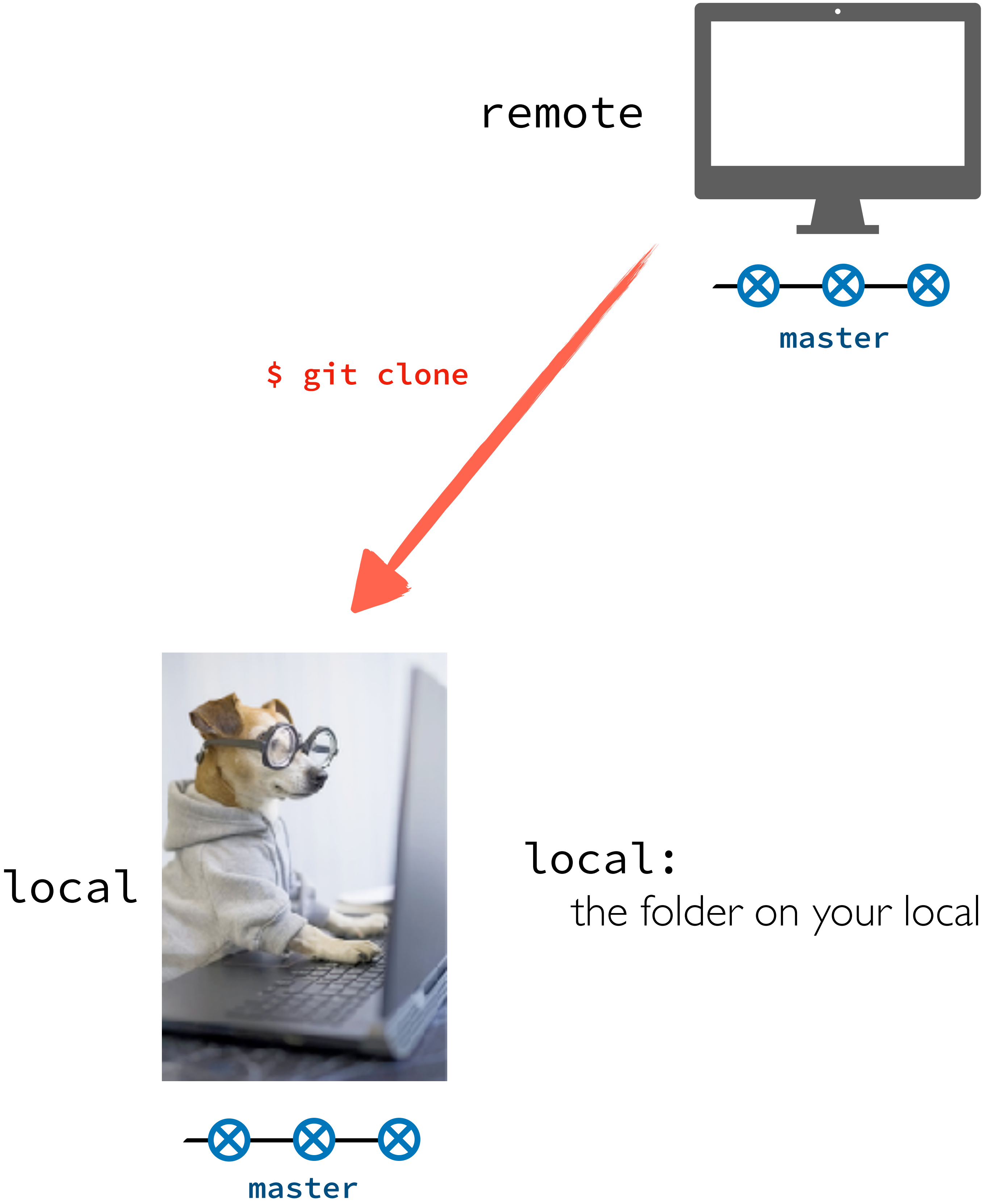


Commits ⊗* and ⊗* changed *the same file on the same line(s)*.
What happens if we merge both branches?



git reports a ***CONFLICT*** and asks you to **resolve** the conflict.
In other words, which of the conflicting changes do we choose?

Git allows you to work with branches from remote repositories

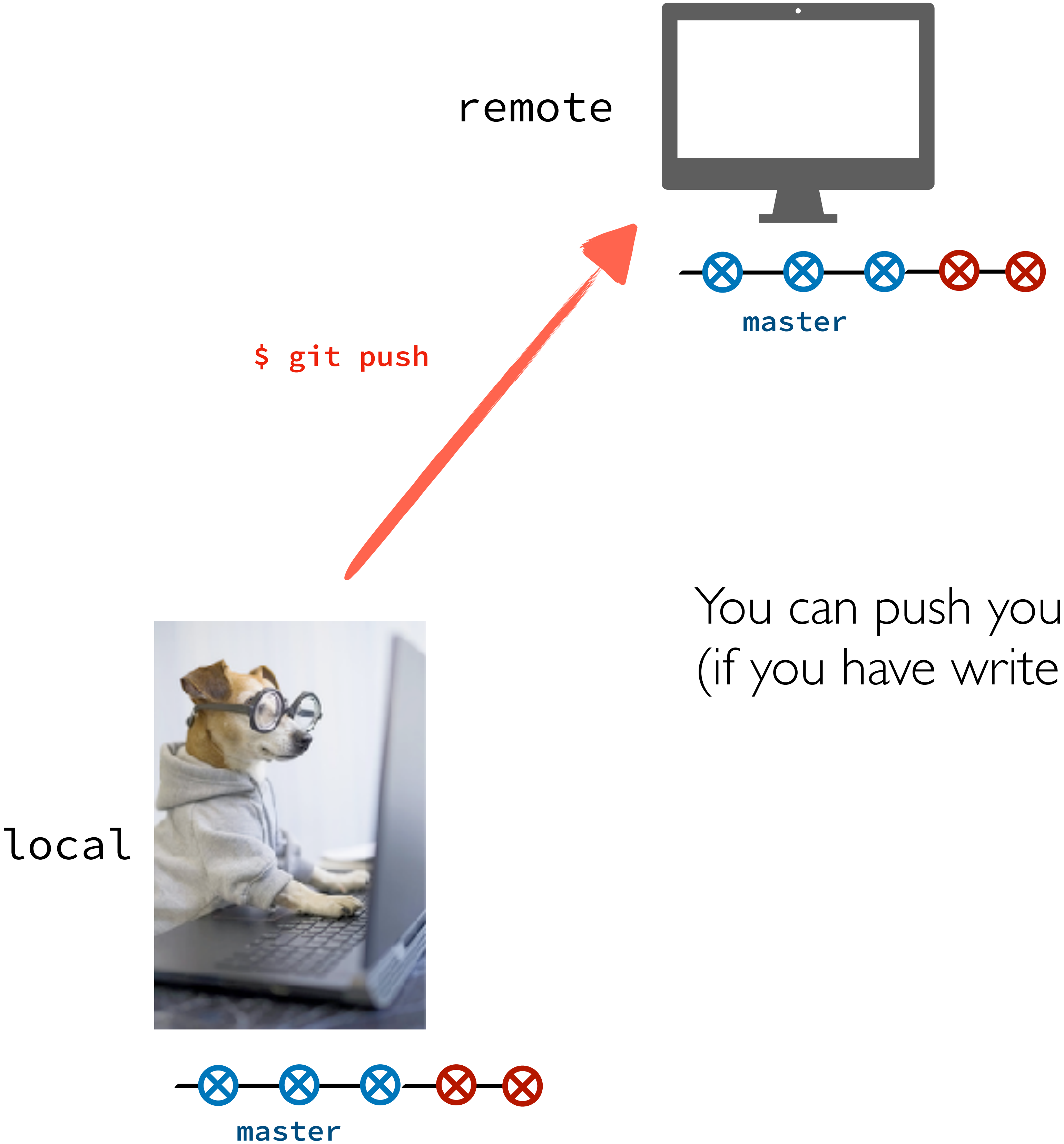


`remote:`

a far away server on the internet
a different computer on your local network
a different folder on the same computer

`local:`

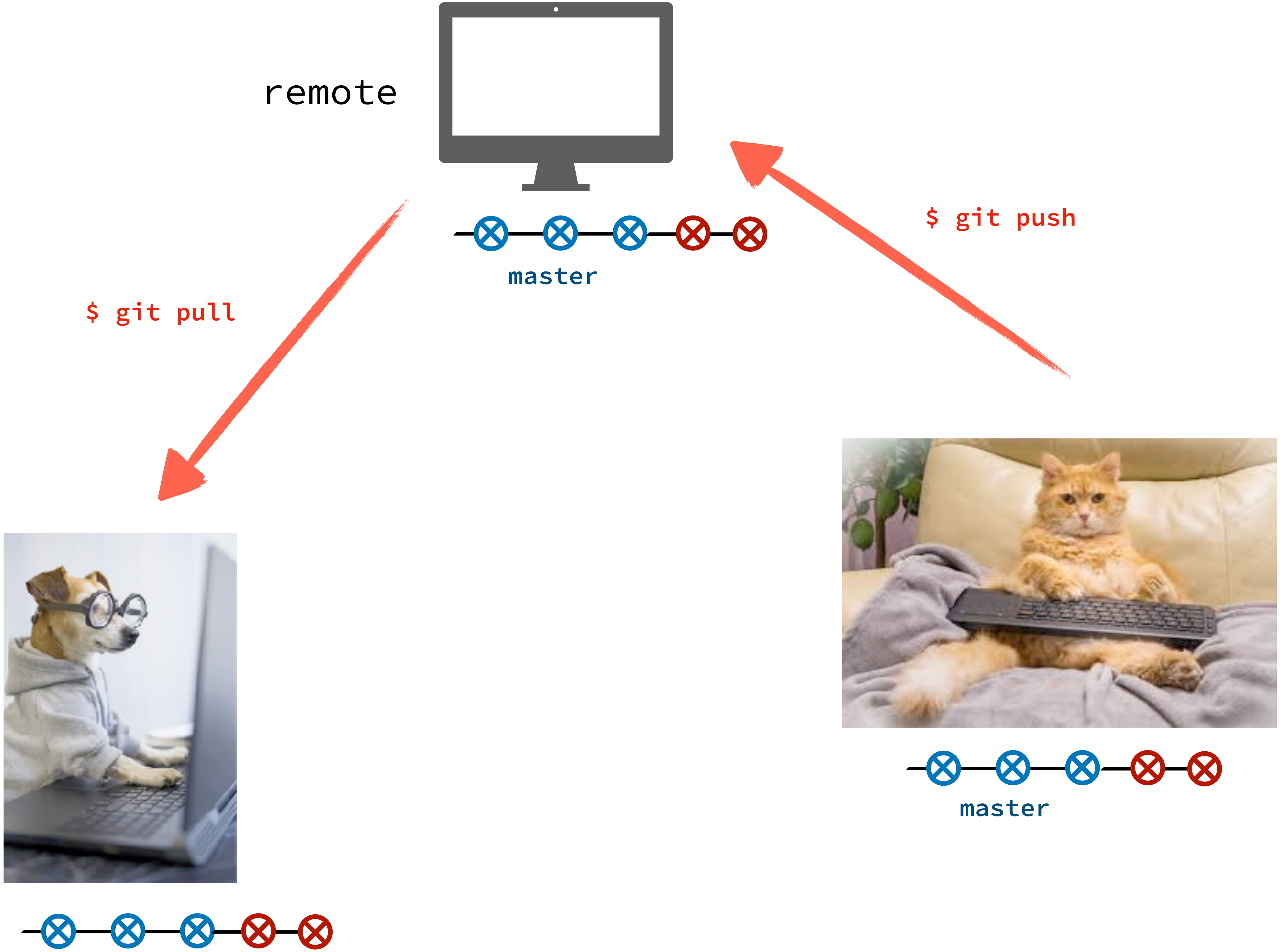
the folder on your local computer



```
$ git push
```

You can push your new commits to the remote (if you have write access)

Remotes facilitate **collaboration**



remote

master

`$ git pull`

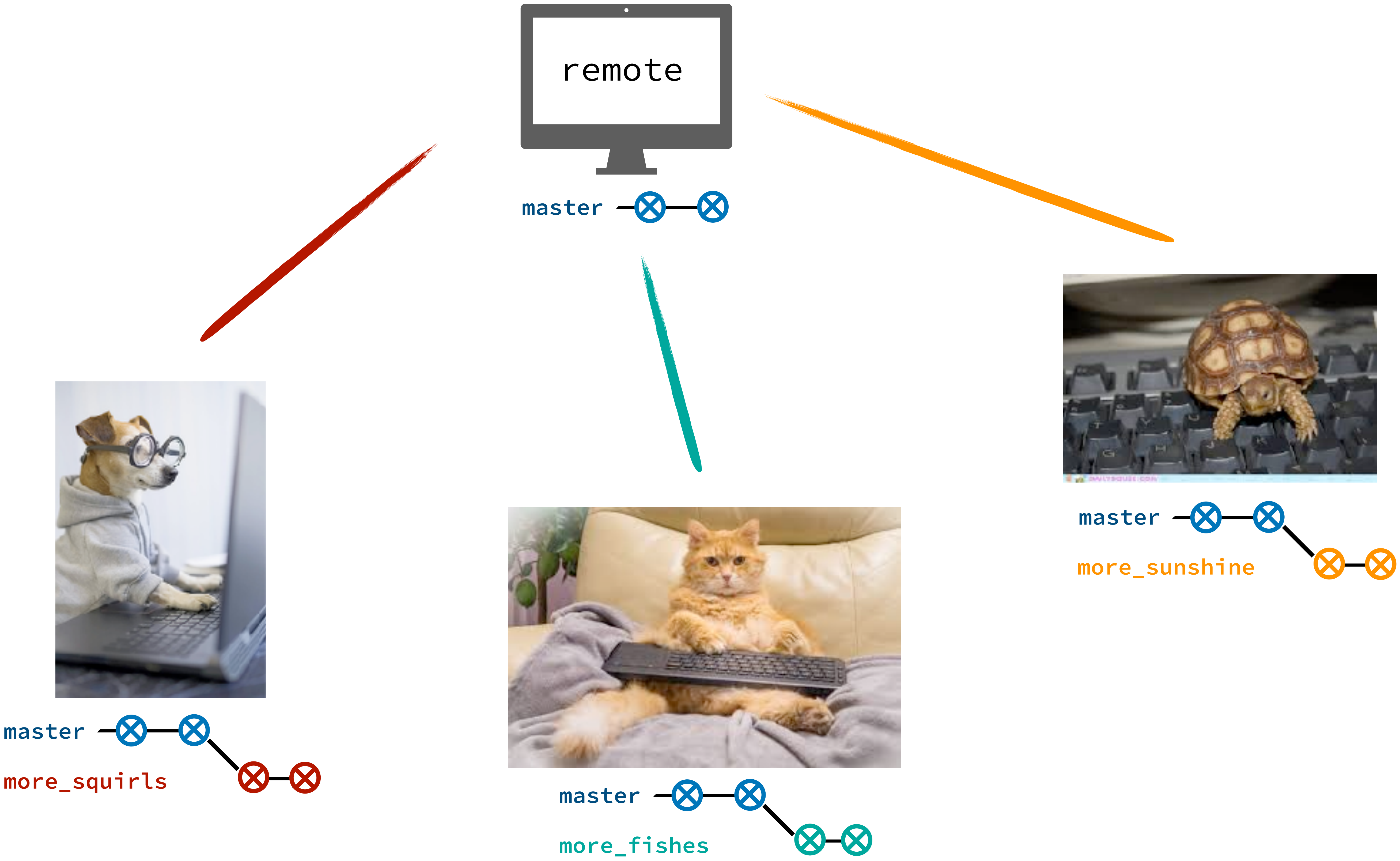
`$ git push`

local

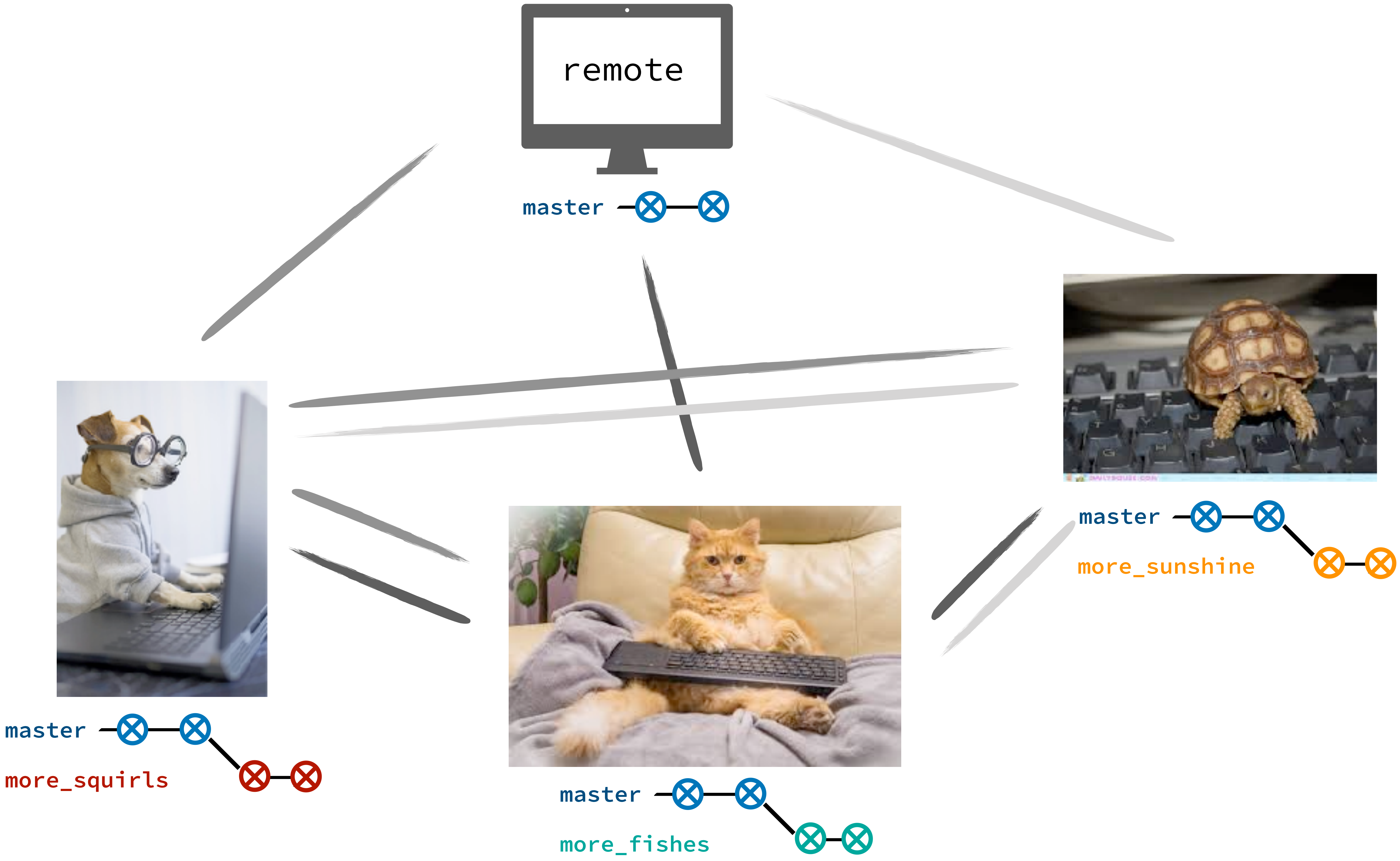
master

master

Collaborators can use branches to work on separate features



Collaborators can inspect and interact with each others branches directly
Git is a decentralized system



remote

master

master

more_sunshine

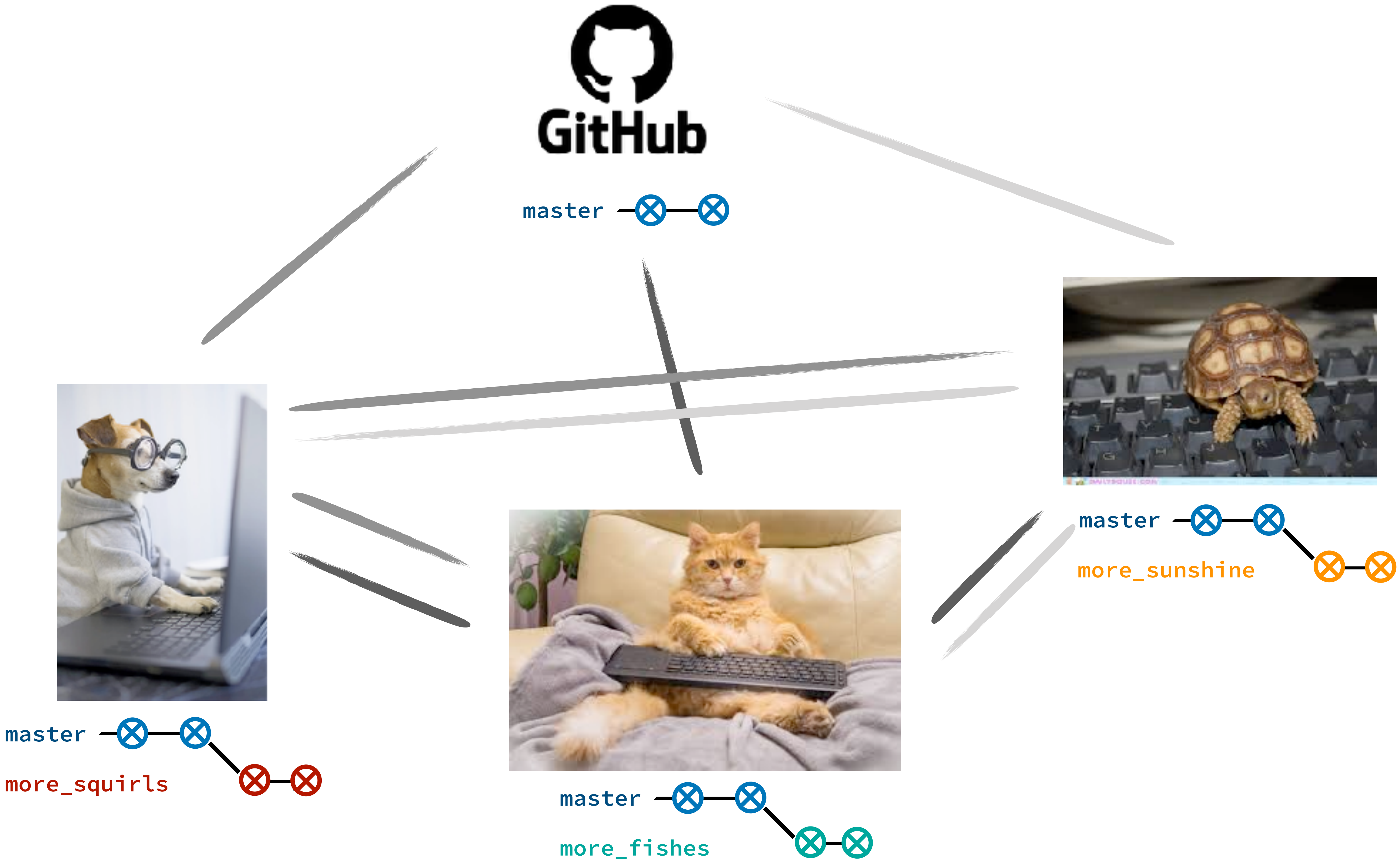
master

more_squirls

master

more_fishes

GitHub commonly acts as a remote



master

master
more_sunshine

master
more_squirls

master
more_fishes

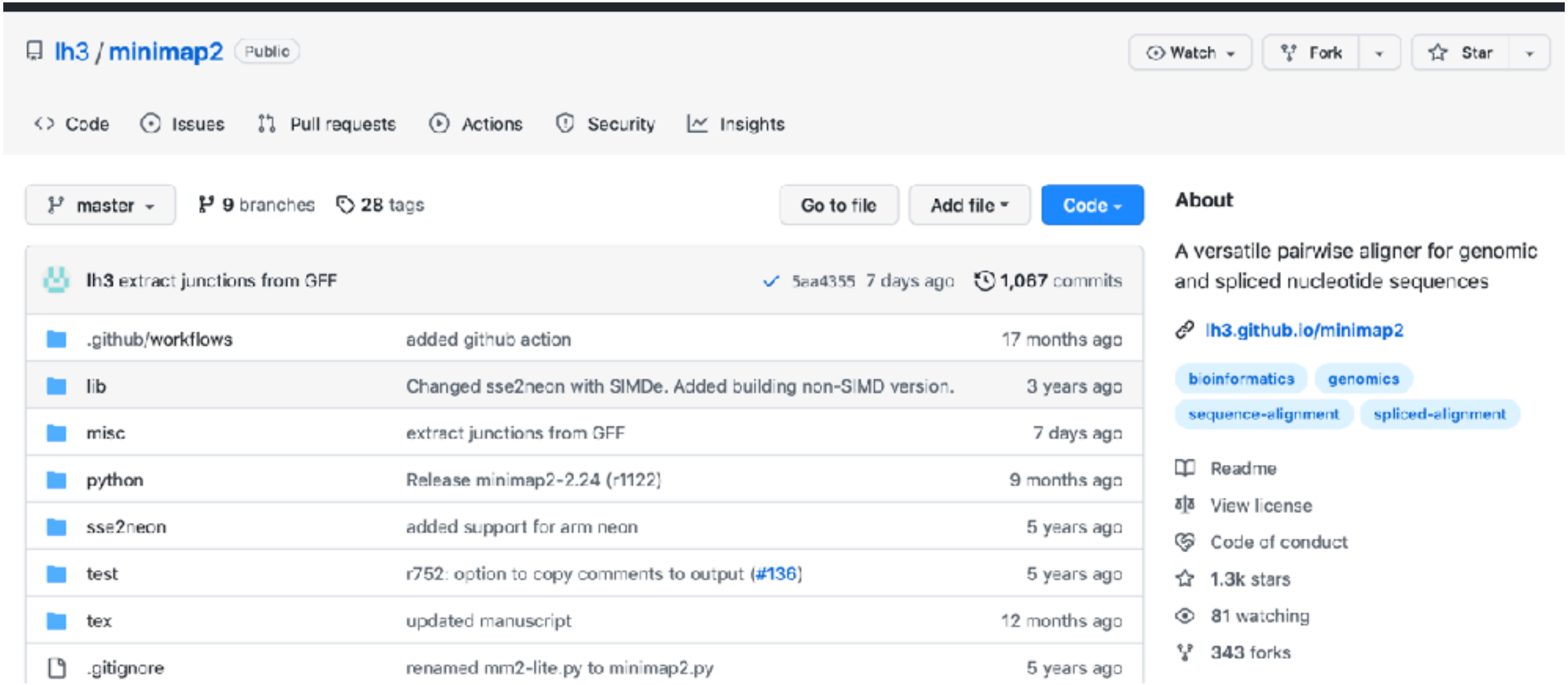
GITHUB

Internet hosting service for git

- publish your code
- graphical user interface
- project wiki and README
- issue tracking
- pull requests
- release version support
- search engine
- see who did what when
- *and much more*

- Launched on April 2008
- Bootstrapped start-up business
- 46000 repositories in the first year
  1 million in 2010
- Acquired by Microsoft for 7.5 billion dollars in 2018, so now a subsidiary of Microsoft. This led to a lot of projects moving to competitors GitLab, BitBucket and SourceForge
- June 2022
  83 million users
  200 million repositories, of which 28 million public

A tour of a publicly available repository



lh3 extract junctions from GFF ✓ 5aa4355 7 days ago 1,067 commits

.github/workflows	added github action	17 months ago
lib	Changed sse2neon with SIMDe. Added building non-SIMD version.	3 years ago
misc	extract junctions from GFF	7 days ago
python	Release minimap2-2.24 (r1122)	9 months ago
sse2neon	added support for arm neon	5 years ago
test	r752: option to copy comments to output (#136)	5 years ago
tex	updated manuscript	12 months ago
.gitignore	renamed mm2-lite.py to minimap2.py	5 years ago

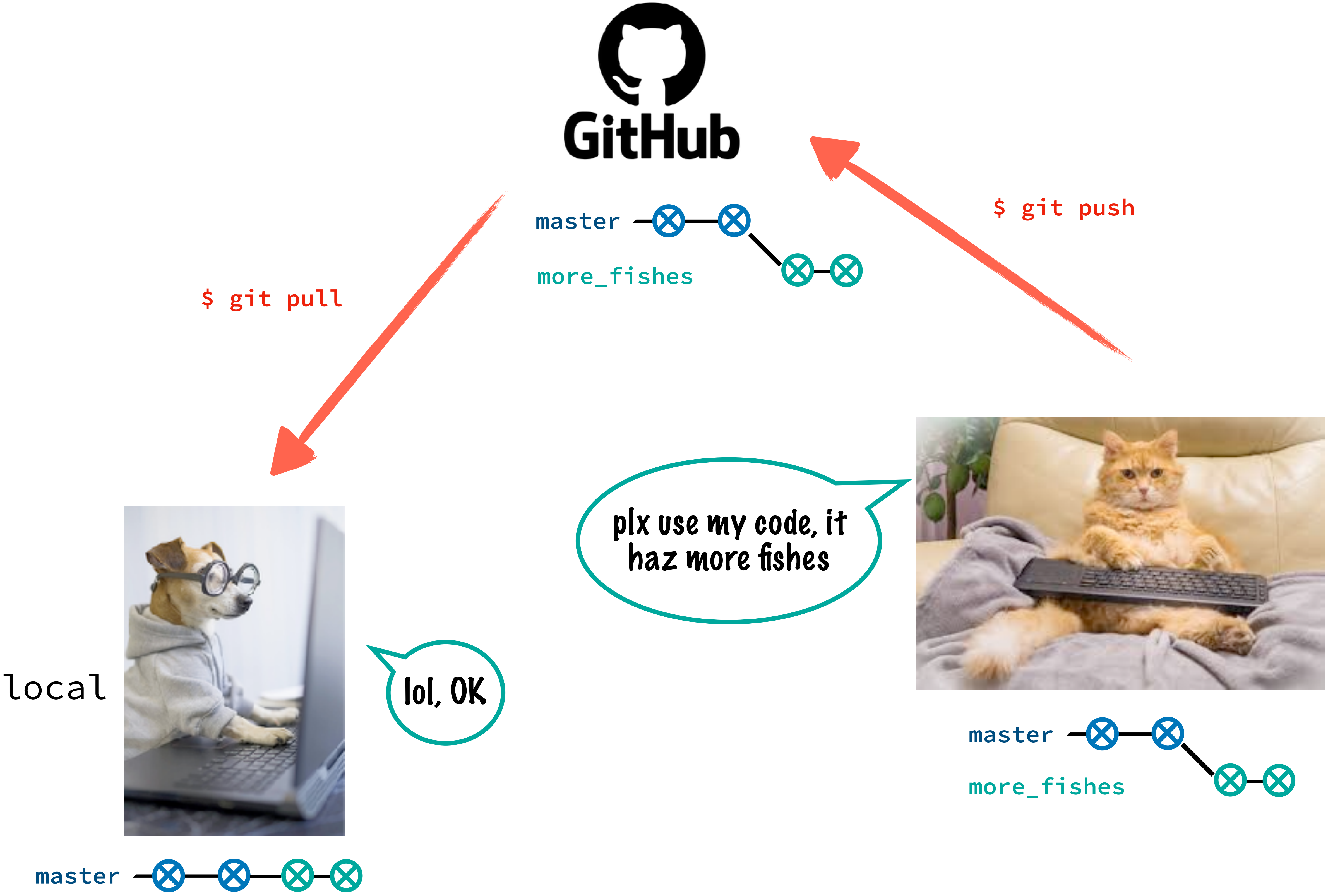
About

A versatile pairwise aligner for genomic and spliced nucleotide sequences

lh3.github.io/minimap2

- The code
- Getting the code with `$ git clone`
- Commit history
- The branches
- README
- Tags and release history
- Issues, open and closed
- Contributors & Blame
- Pull requests
- Other repositories of the same user

Pull Requests



`$ git pull`

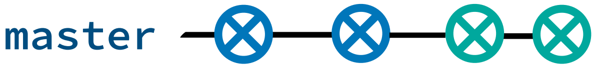
`$ git push`

GitHub: `master`, `more_fishes`

local: `master`

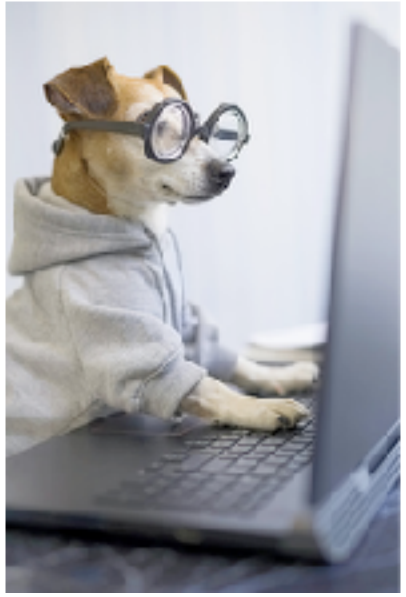
cat: `master`, `more_fishes`

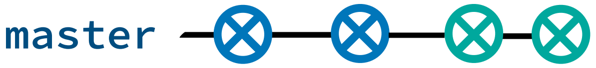
GitHub

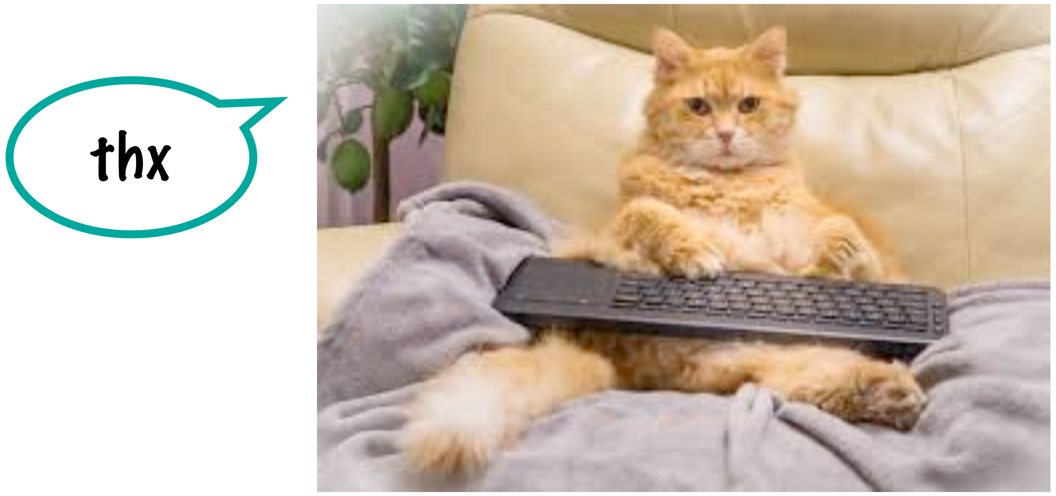


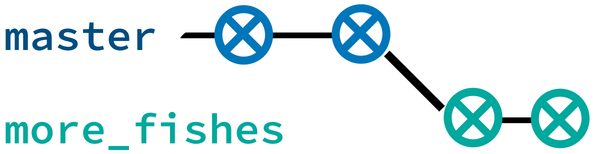
$ git push

local

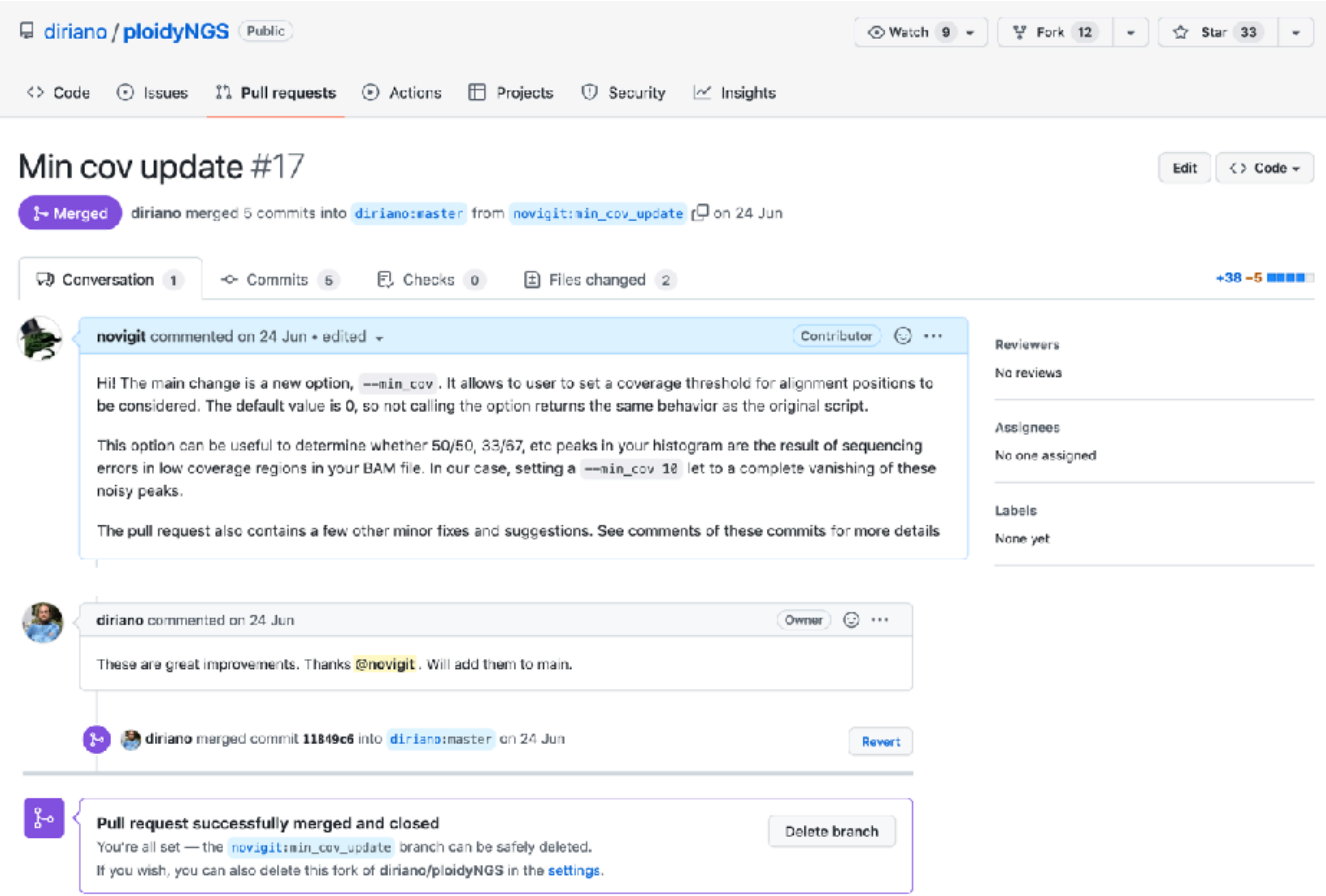








A pull request of my own



Min cov update #17

Merged diriano merged 5 commits into `diriano:master` from `novigit:min_cov_update` on 24 Jun

novigit commented on 24 Jun • edited

Hi! The main change is a new option, `--min_cov`. It allows to user to set a coverage threshold for alignment positions to be considered. The default value is 0, so not calling the option returns the same behavior as the original script.

This option can be useful to determine whether 50/50, 33/67, etc peaks in your histogram are the result of sequencing errors in low coverage regions in your BAM file. In our case, setting a `--min_cov 10` let to a complete vanishing of these noisy peaks.

The pull request also contains a few other minor fixes and suggestions. See comments of these commits for more details

diriano commented on 24 Jun

These are great improvements. Thanks @novigit. Will add them to main.

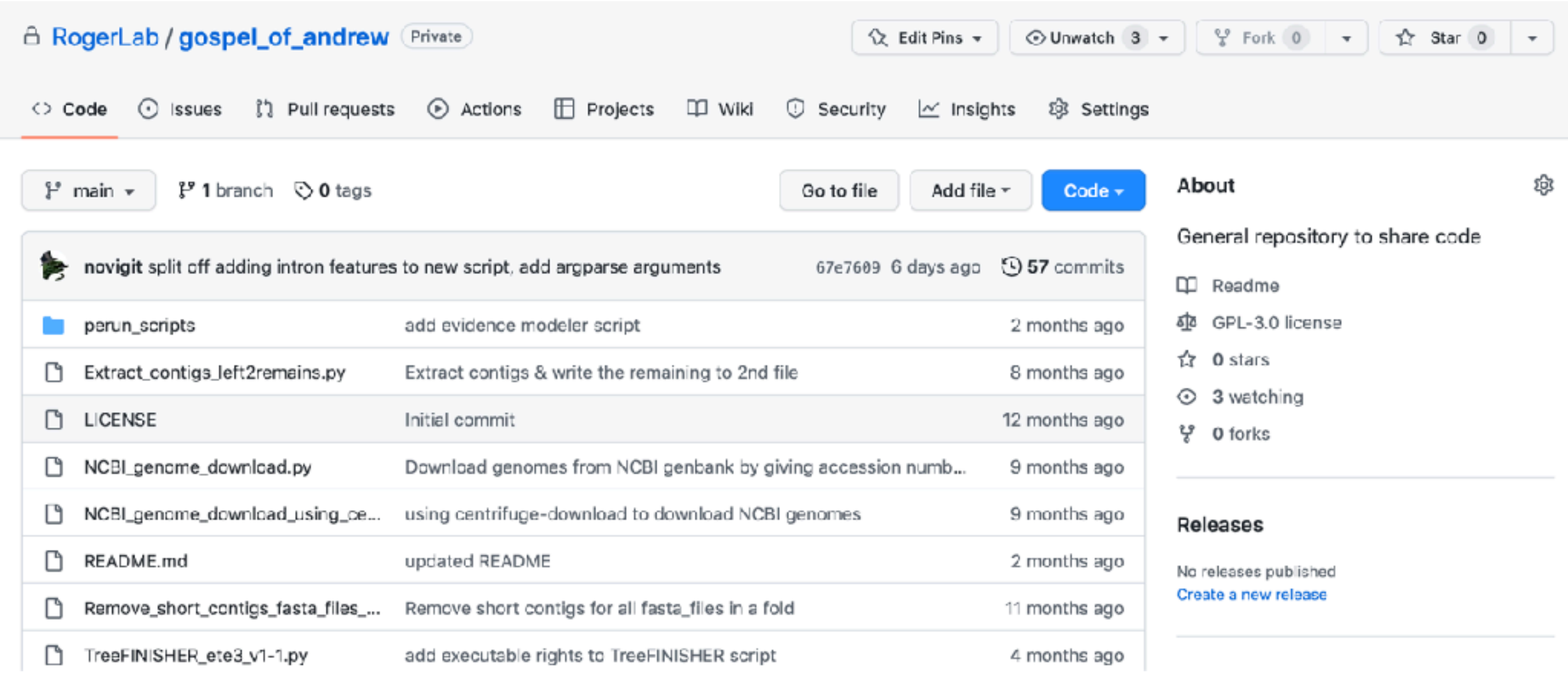
diriano merged commit **11849c6** into `diriano:master` on 24 Jun

Pull request successfully merged and closed

You're all set — the `novigit:min_cov_update` branch can be safely deleted.

If you wish, you can also delete this fork of **diriano/ploidyNGS** in the settings.

Our own private repository for the Roger Lab



RogerLab / **gospel_of_andrew** Private

novigit split off adding intron features to new script, add argparse arguments 67e7609 6 days ago 57 commits

perun_scripts	add evidence modeler script	2 months ago
Extract_contigs_left2remains.py	Extract contigs & write the remaining to 2nd file	8 months ago
LICENSE	Initial commit	12 months ago
NCBI_genome_download.py	Download genomes from NCBI genbank by giving accession numb...	9 months ago
NCBI_genome_download_using_ce...	using centrifuge-download to download NCBI genomes	9 months ago
README.md	updated README	2 months ago
Remove_short_contigs_fasta_files_...	Remove short contigs for all fasta_files in a fold	11 months ago
TreeFINISHER_ete3_v1-1.py	add executable rights to TreeFINISHER script	4 months ago

About

General repository to share code

- Readme
- GPL-3.0 license
- 0 stars
- 3 watching
- 0 forks

Releases

No releases published
Create a new release

- The code
- Commit history
- README
- Contributors & Blame
- RogerLab "organization"

We are not software developers, but 'mere' mortal biologists
Lots of git functionality is probably not useful for us

Still, the basics of Git and GitHub are very useful:

- `gospel_of_andrew` is an **easy-to-access central location to store and share scripts** useful for the lab

- track updates of scripts with git

- **interact directly with developers** of bioinformatics softwares
    - request features
    - report issues
    - provide your own bug fixes / code enhancements

- **publish your code** and/or pipelines on GitHub along with your papers
    - example: ergobibamus project
    - example: johanneson lab

- track manual curation of gene models with git