UiT
THE ARCTIC
UNIVERSITY
OF NORWAY

Version control and you

Einar Holsbø
mail: einar@cs.uit.no — web: 3inar.github.io
November 16th, 2016

Take-home messages

- You’re one out of four types of analyst: try to get to the better class
- You should think about the people you collaborate with
- You’re a programmer, so you may as well act like one
- Programmers use version control

Thesis of these talks: reproducible code -> reproducible analysis -> reproducible research

Classification of analysts

1. Does most work in Excel. R, stata, etc. considered “specialist tools”
2. Uses R or other real programming environment, keeps code in files, file names for version control
3. Like #2, with real version control, a programming style guide, peer review
4. Like #3, but more software engineering-savvy. Automated testing, modular code, proper documentation, packages, etc., etc.

http://ellisp.github.io/blog/2016/09/16/version-control

Classification of analysts

+++ complexity

oes most work in Excel. R, stata, etc. considered “specialist tools”

ses R or other real programming environment, keeps code in files,
le names for version control

2, with real version control, a programming style guide, peer
ew

ke #3, but more software engineering-savvy. Automated testing,
nodular code, proper documentation, packages, etc., etc.

http://ellisp.github.io/blog/2016/09/16/version-control

Classification of analysts

+++ complexity

+++ professionalism

oes most work in Excel. R, stata, etc. considered “specialis ols”

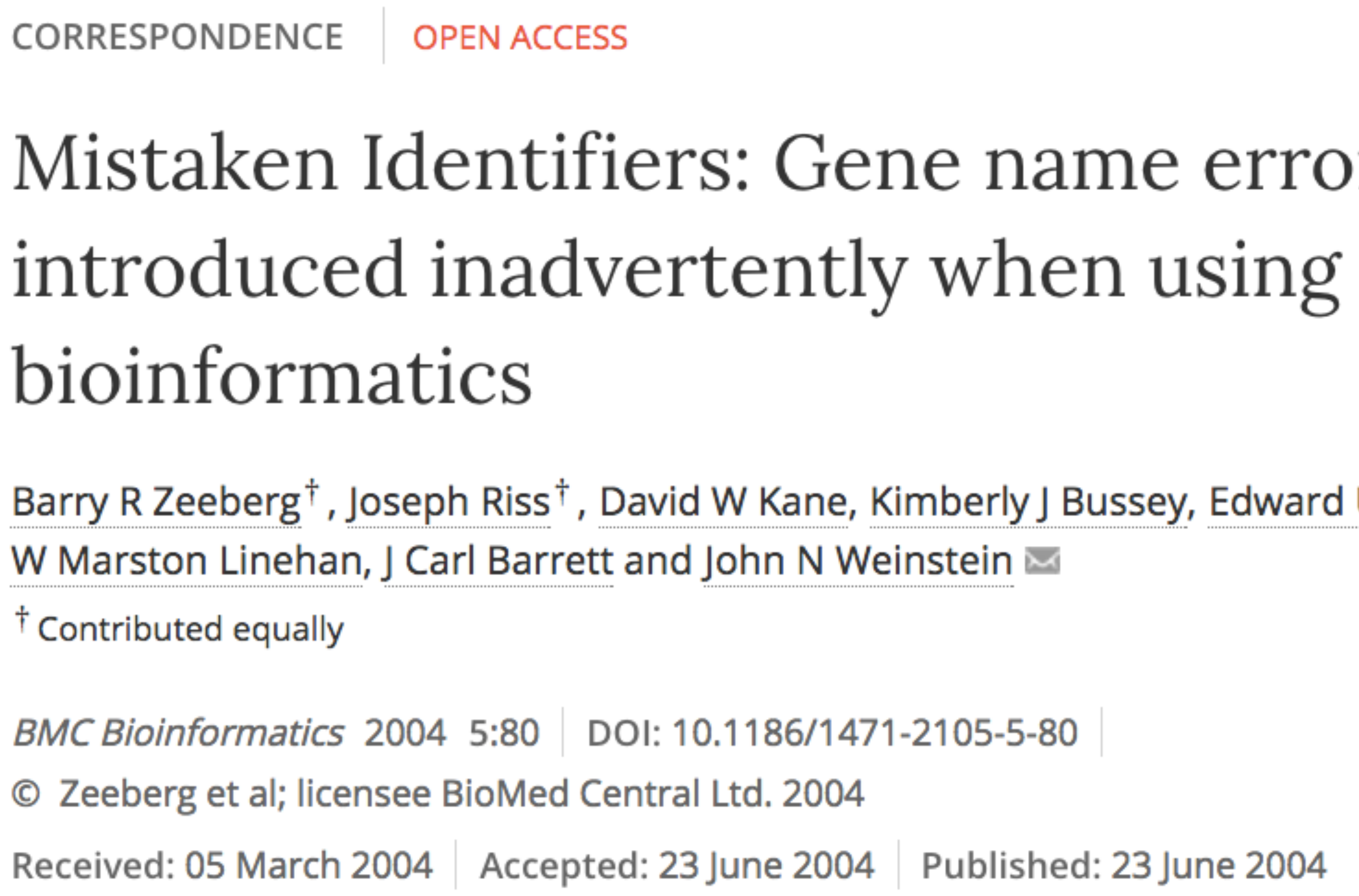
ses R or other real programming environment, keeps code les,
le names for version control

2, with real version control, a programming style guide er
ew

ke #3, but more software engineering-savvy. Automated
nodular code, proper documentation, packages, etc., etc.

http://ellisp.github.io/blog/2016/09/16/version-control

Class 1: the excel-analyst



CORRESPONDENCE | OPEN ACCESS

Mistaken Identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics

Barry R Zeeberg[†], Joseph Riss[†], David W Kane, Kimberly J Bussey, Edward Uchio, W Marston Linehan, J Carl Barrett and John N Weinstein

[†] Contributed equally

BMC Bioinformatics 2004 5:80 | DOI: 10.1186/1471-2105-5-80

© Zeeberg et al; licensee BioMed Central Ltd. 2004

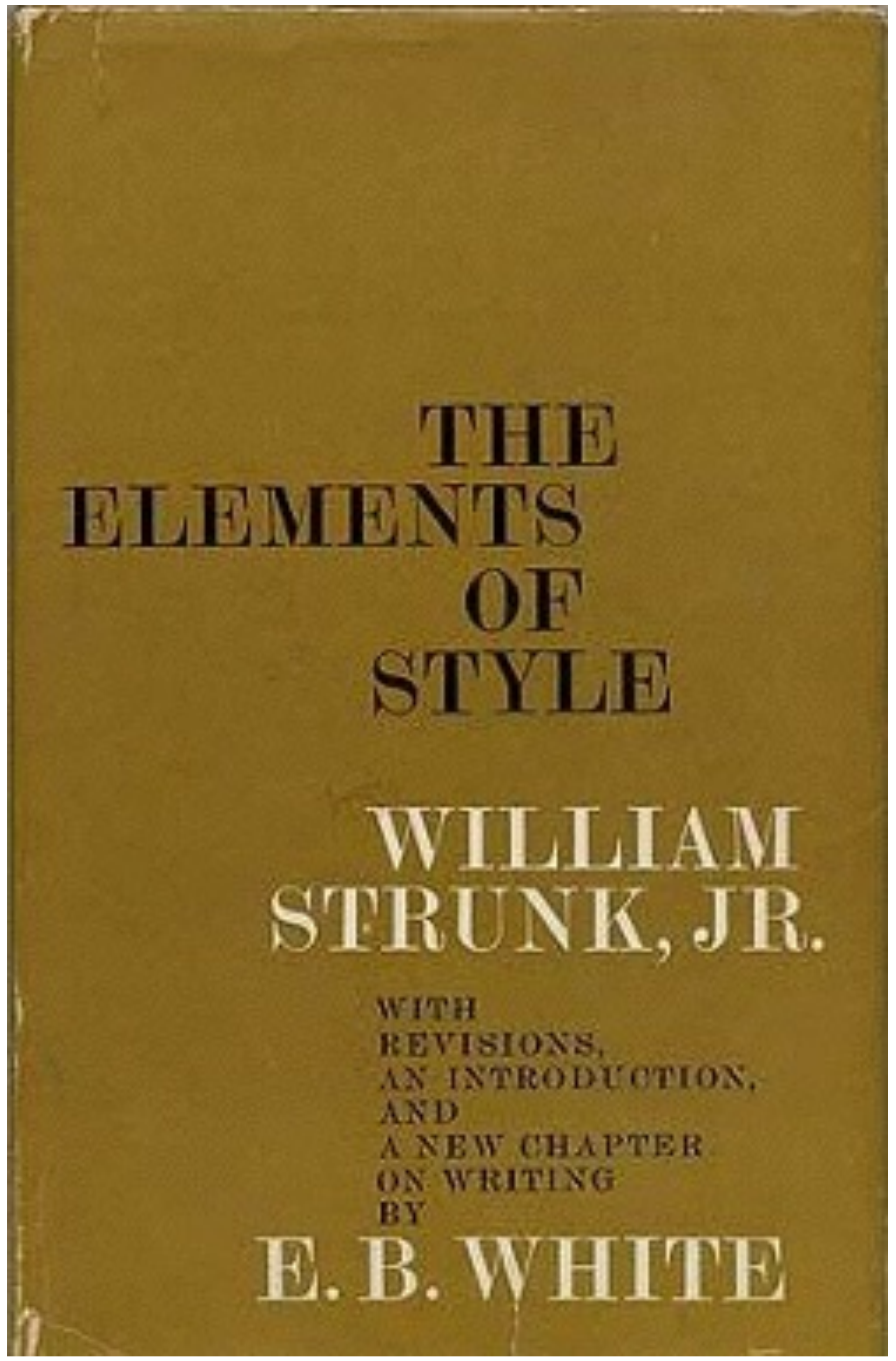
Received: 05 March 2004 | Accepted: 23 June 2004 | Published: 23 June 2004

Class 2: uses R

"FINAL".doc
FINAL.doc!
FINAL_rev.2.doc
FINAL_rev.6.COMMENTS.doc
FINAL_rev.8.comments5.CORRECTIONS.doc
track changes
FINAL_rev.18.comments7.corrections9.MORE.30.doc
FINAL_rev.22.comments49.corrections.10.#@$%WHYDIDICOMETOGRADSCHOOL????.doc
JORGE CHAM © 2012
WWW.PHDCOMICS.COM

Class 3: version control, style guide, peer review

https://git-scm.com/about


THE
ELEMENTS
OF
STYLE
WILLIAM
STRUNK, JR.
WITH
REVISIONS,
AN INTRODUCTION,
AND
A NEW CHAPTER
ON WRITING
BY
E. B. WHITE

Class 4: basically a software engineer

https://wiki.teamfortress.com/wiki/File:Engineertaunt1.PNG

Class 1 & 2: hard to work with

Class 3 & 4: play well with others

I originally planned to reproduce some of the results on gene expression, but decided not to.

(1) First I had to find out, by guessing and asking, who had written the R scripts. It turned out that several people had been involved. No authors, no dates, no mention of who had written what, several versions writing to the same file, etc.

(2) Some of the authors are no longer employed by ISM, others are on leave, making it a harder job to get answers.

I concluded that it would probably take less time for me to write my own scripts, despite that I have to learn R first.

–Kajsa Møllersen, distributed on nowac mailing list

LACK OF PROVENANCE

+

LACK OF DOCUMENTATION

=

UNPROFESSIONAL

& UNREPRODUCIBLE

I originally planned to reproduce some of the results on gene expression, but decided not to.

(1) First I had to find out, by guessing and asking, who had written the R scripts. It turned out that several people had been involved. No authors, no dates, no mention of who had written what, several versions with the same name, etc.

(2) Some of the authors are no longer employed by ISM, others are on leave, making it a harder job to get answers.

I concluded that it would probably take less time for me to write my own scripts, despite that I have to learn R first.

–Kajsa Møllersen, distributed on nowac mailing list

```
%EM algorithm

function[cls, probs, mus, sigmas] = EMmachine(X, num_classes)
    % i) init all prog_m, mu_m, sigma_m
    probs = zeros(1, num_classes);
    mus = zeros(1, num_classes);
    sigmas = zeros(1, num_classes);
    cls = 0;
    l = 1;

    [s, samples] = size(X);

    max_iterations = 6000;

    covariance = cov(X);

```

```
%EM algorithm

function[cls, probs, mus, sigmas] = EMmachine(X, num_classes)
    % i) init all prog_m, mu_m, sigma_m
    probs = zeros(1, num_classes);
    mus = zeros(1, num_classes);
    sigmas = zeros(1, num_classes);
    cls = 0;
    l = 1;

    [s, samples] = size(X);

    max_iterations = 6000;

    covariance = cov(X);
```

Your closest collaborator is you six months ago, but you don't reply to emails.

https://github.com/kbroman/datasciquotes

Version control: Git and friends

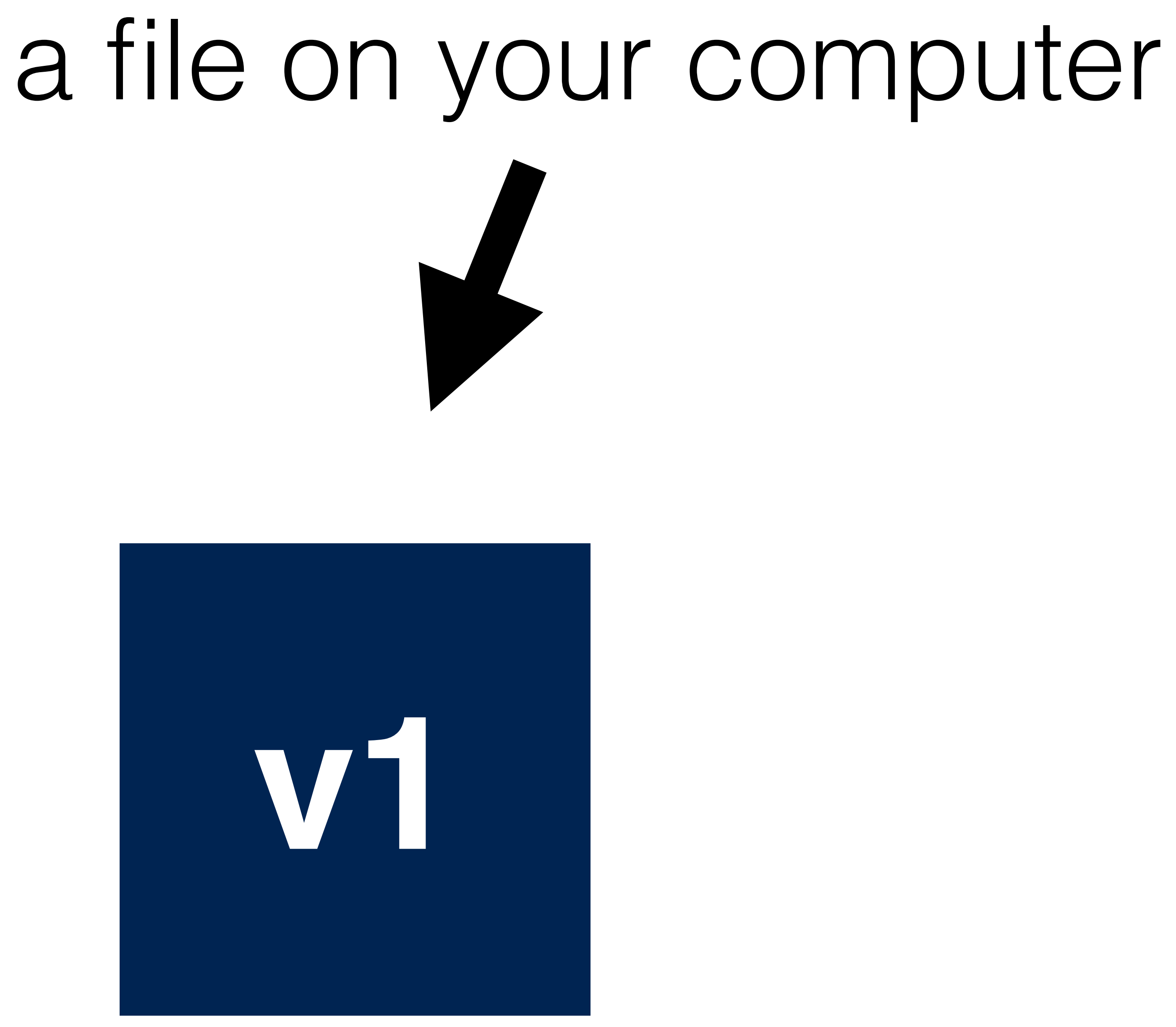
- A sane way to keep track of changes to your analysis
- A structured way to collaborate on code
- Manages provenance

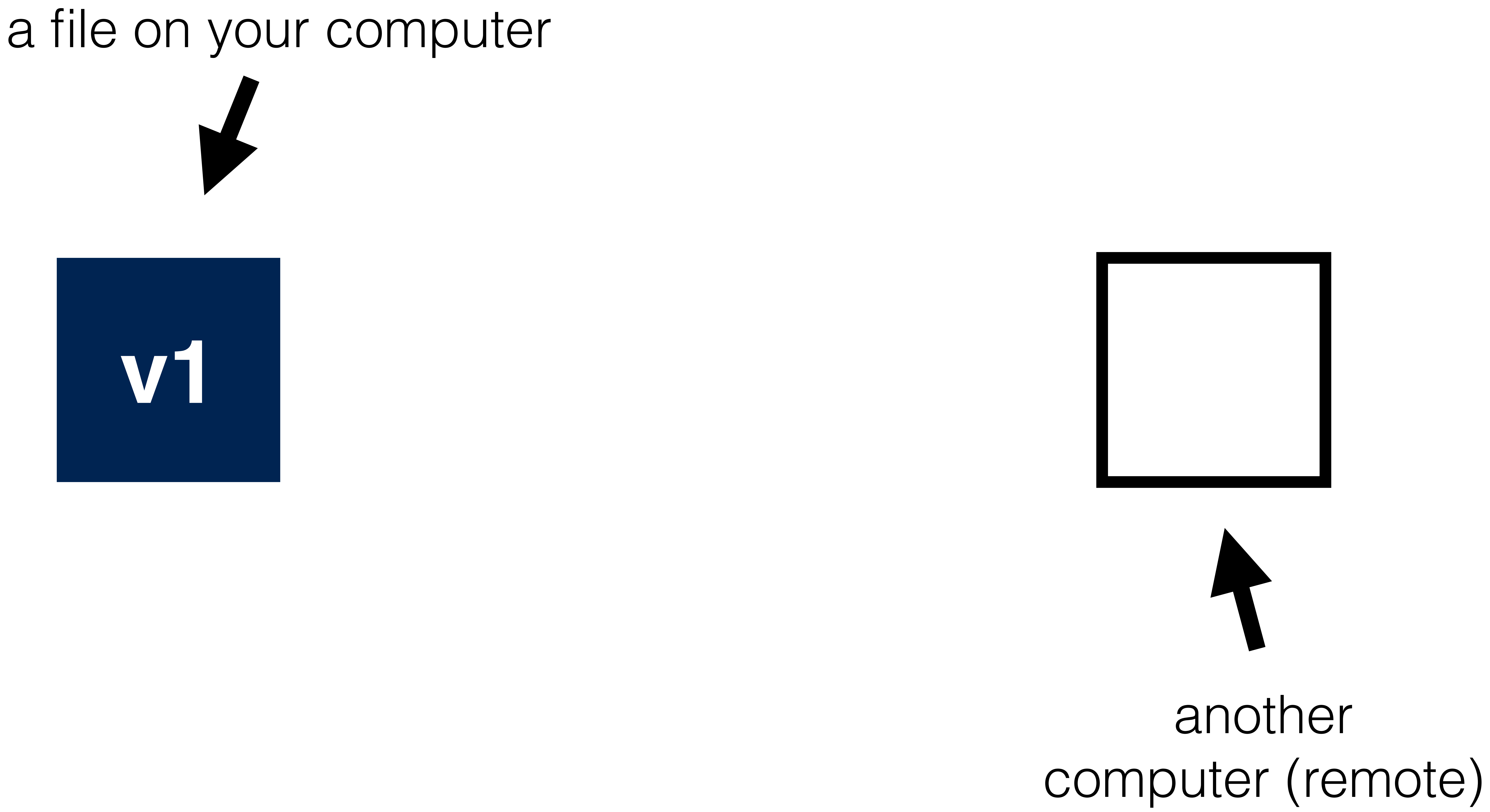
How to Git locally

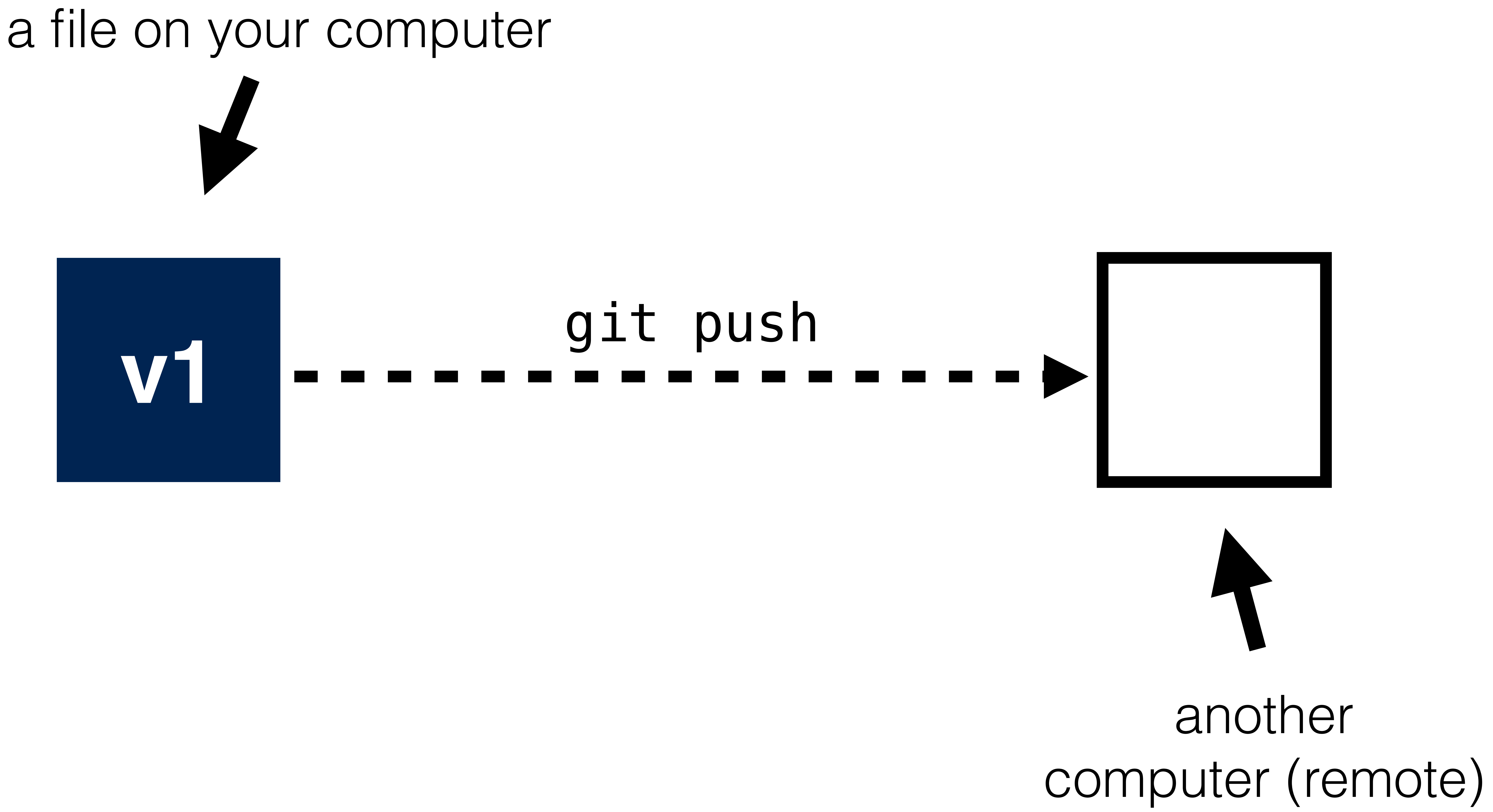
- `git init`
- `git status`
- `git add`
- `git commit`

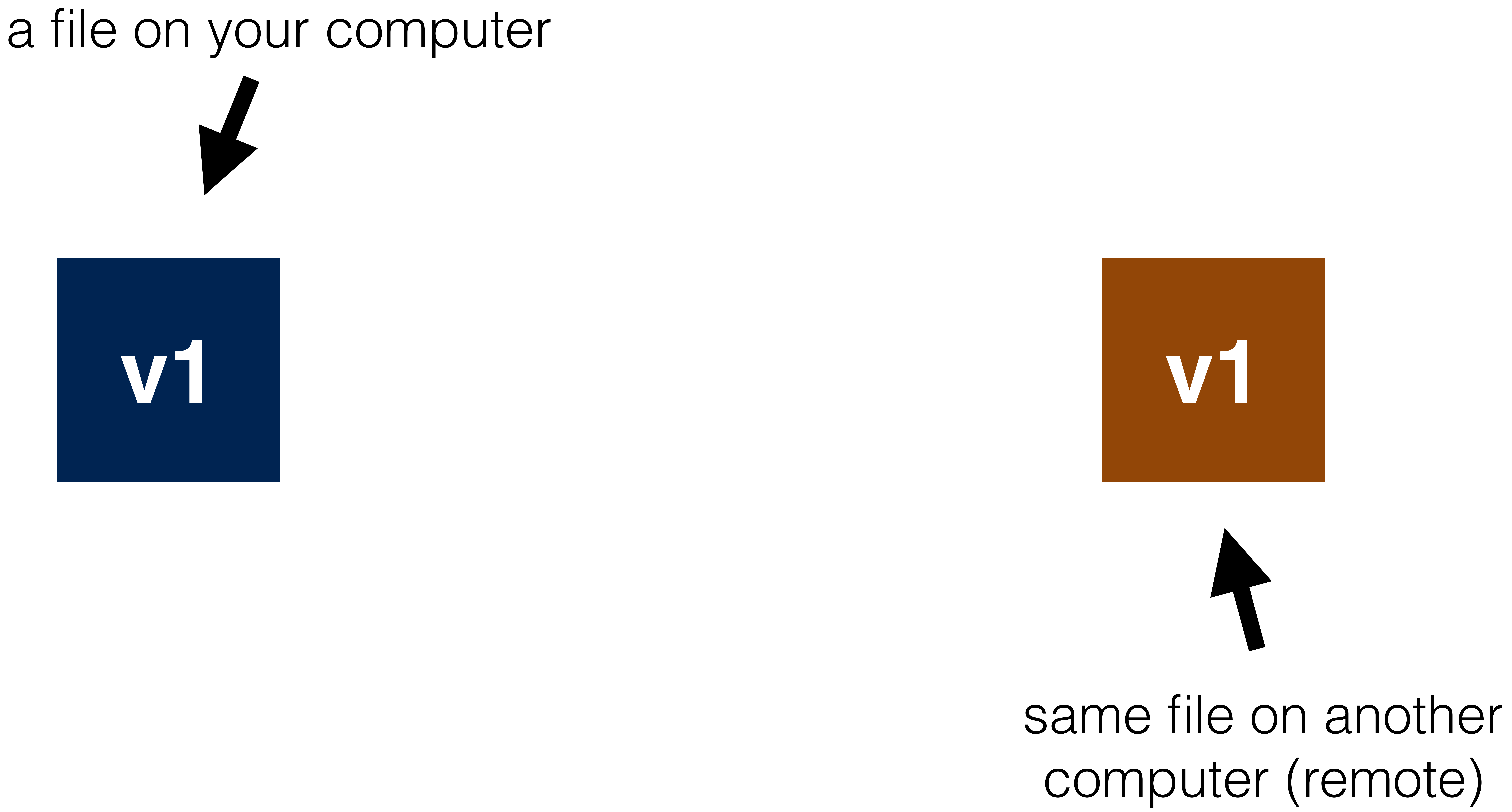
Git demo

Git with friends: `push` and `pull`

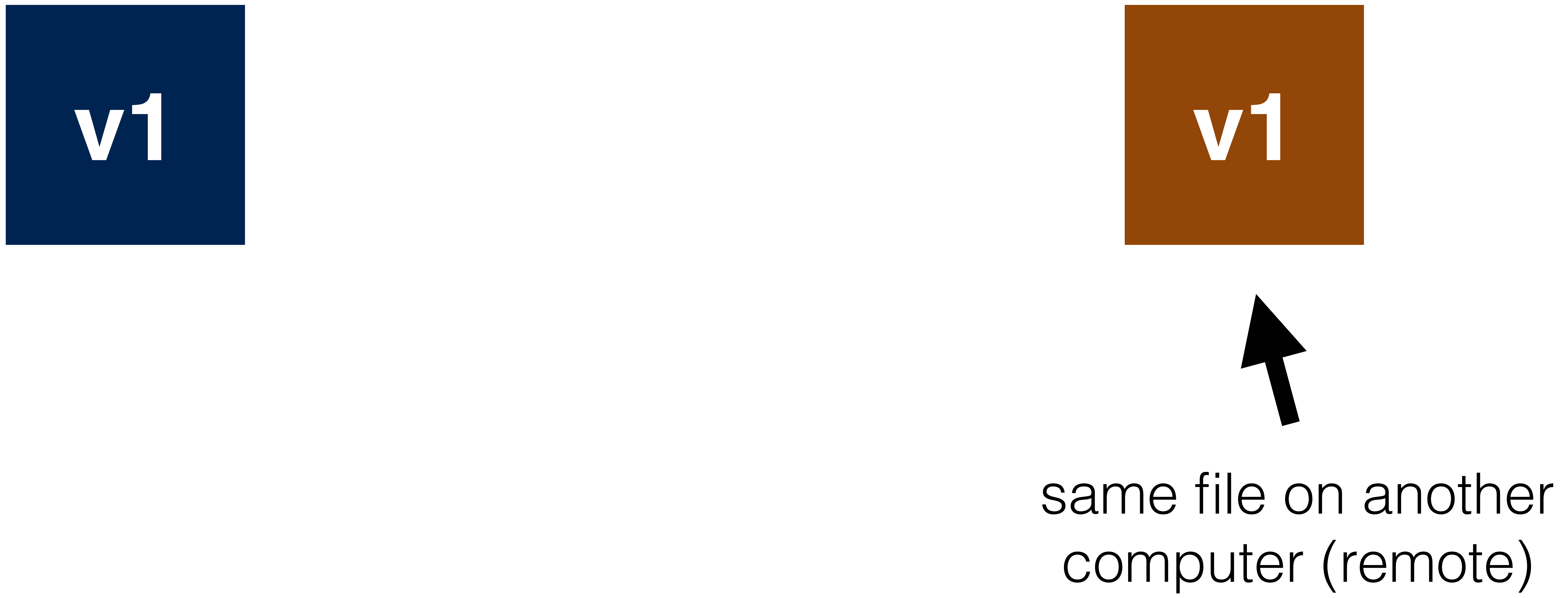

a file on your computer
v1

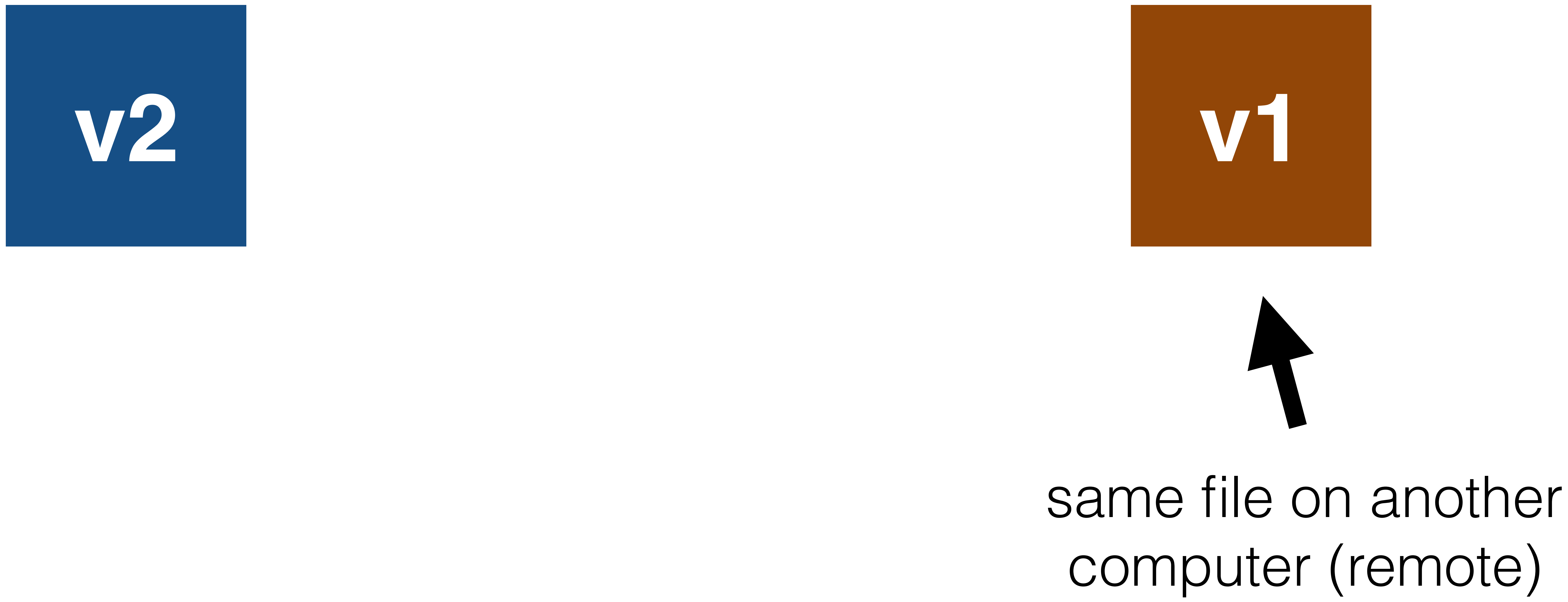

a file on your computer
v1
another
computer (remote)


a file on your computer
v1
git push
another
computer (remote)


a file on your computer
v1
v1
same file on another computer (remote)

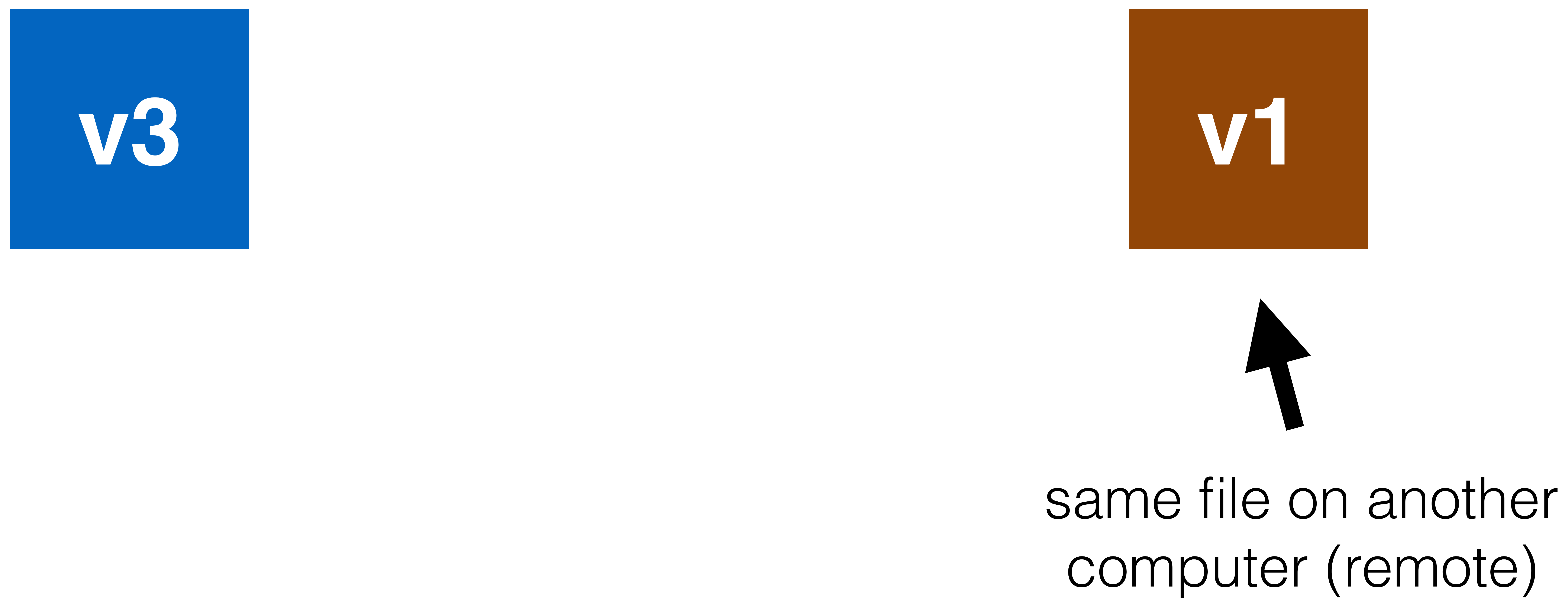
change + `commit`

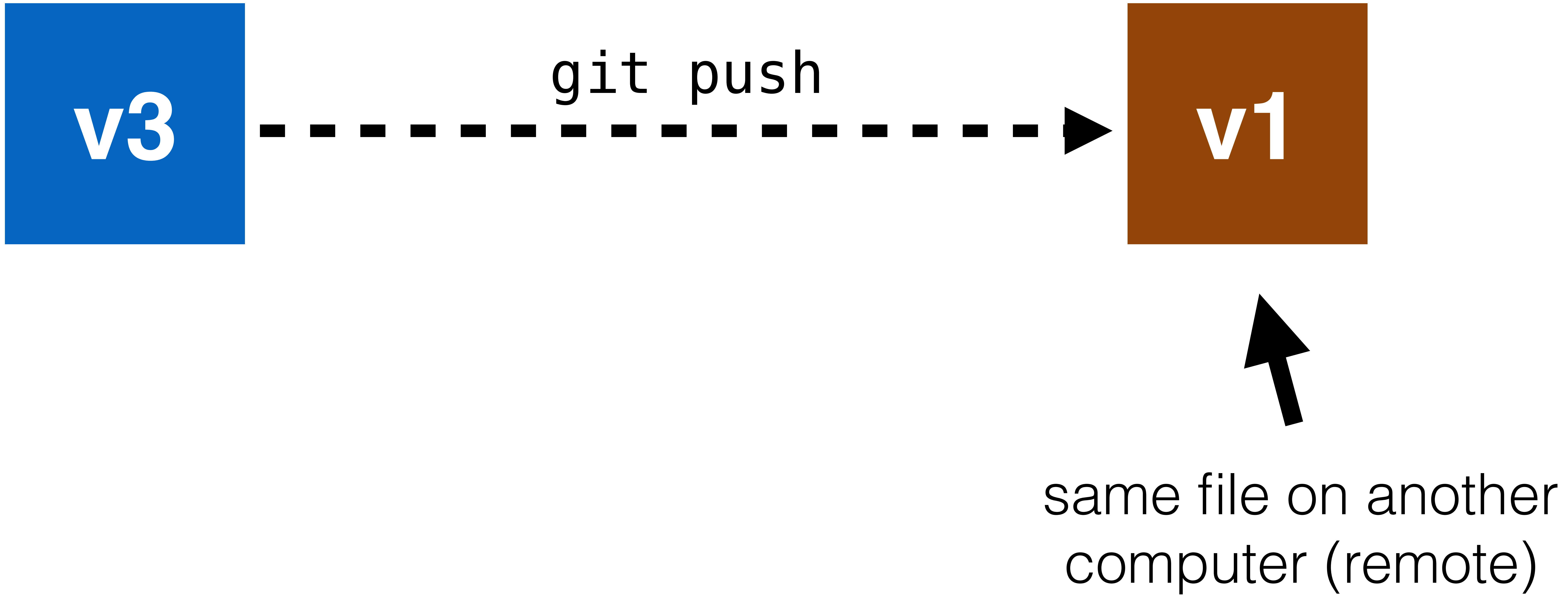


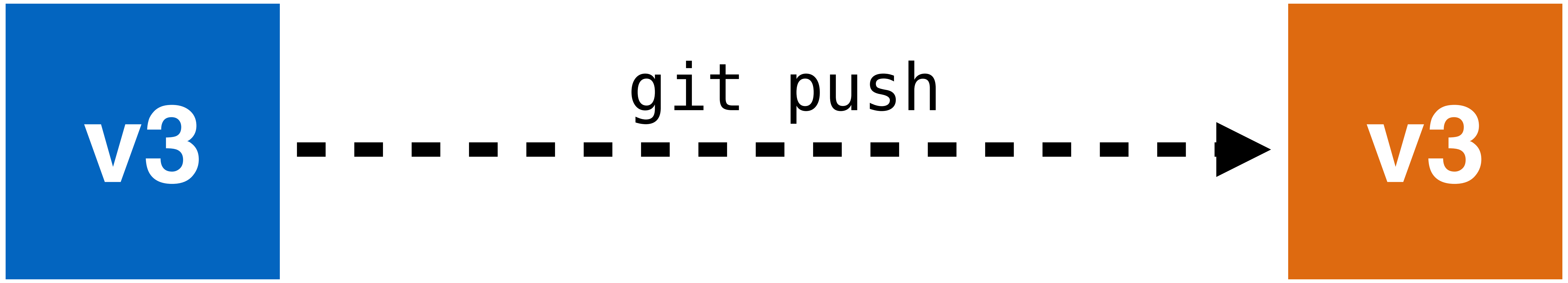

v2
v1
same file on another computer (remote)

change + `commit`

same file on another computer (remote)


v3
v1
same file on another computer (remote)

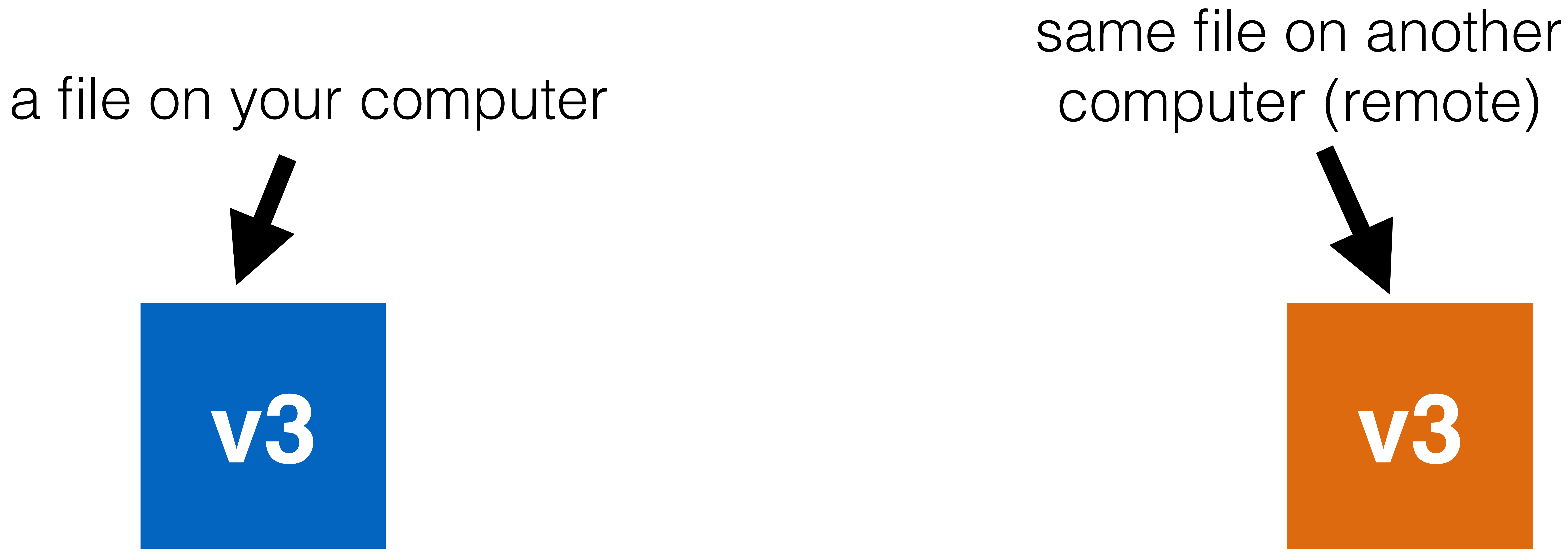

v3
git push
v1
same file on another computer (remote)

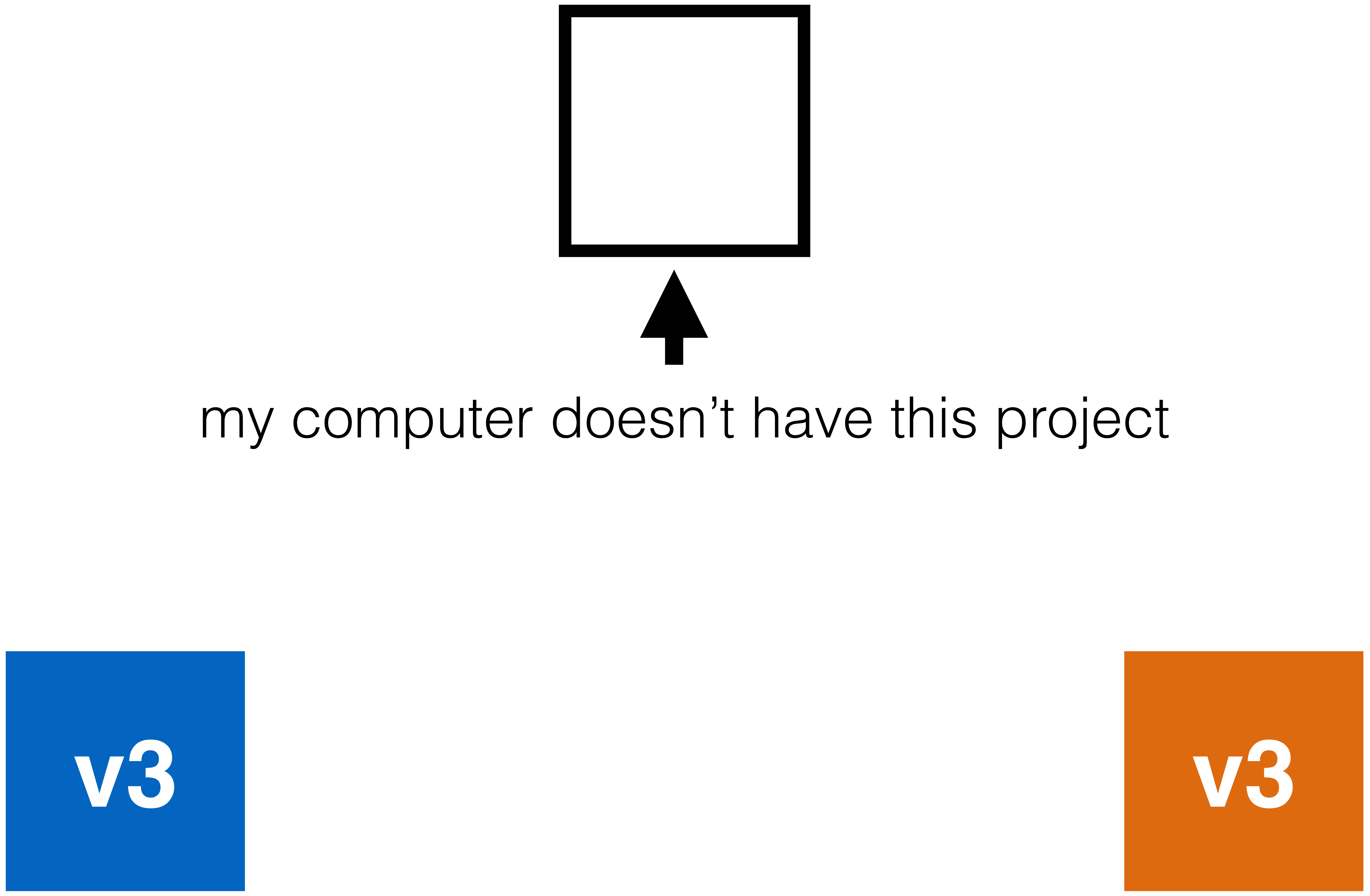

v3
git push
v3

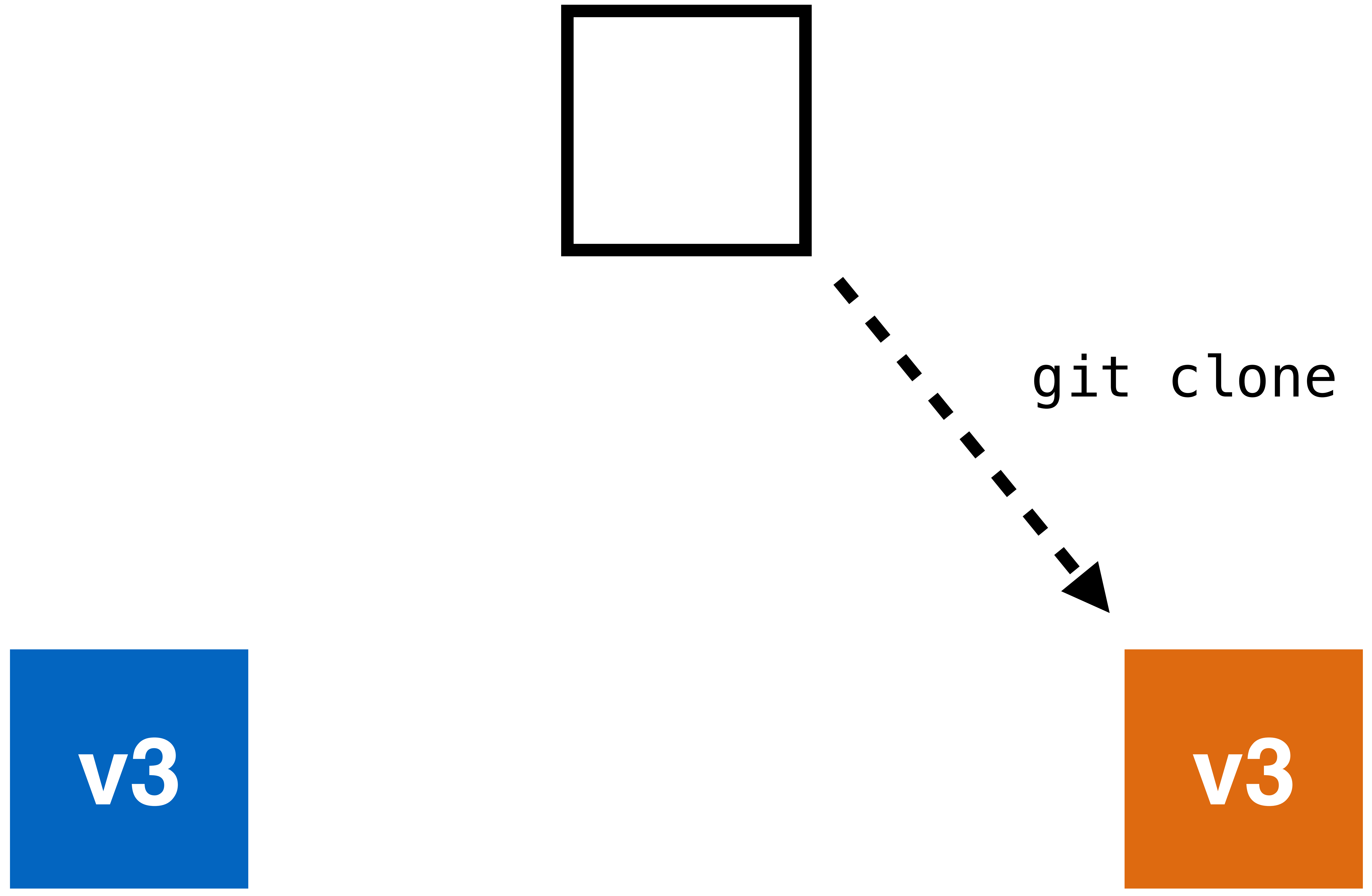
Remote is up-to-date

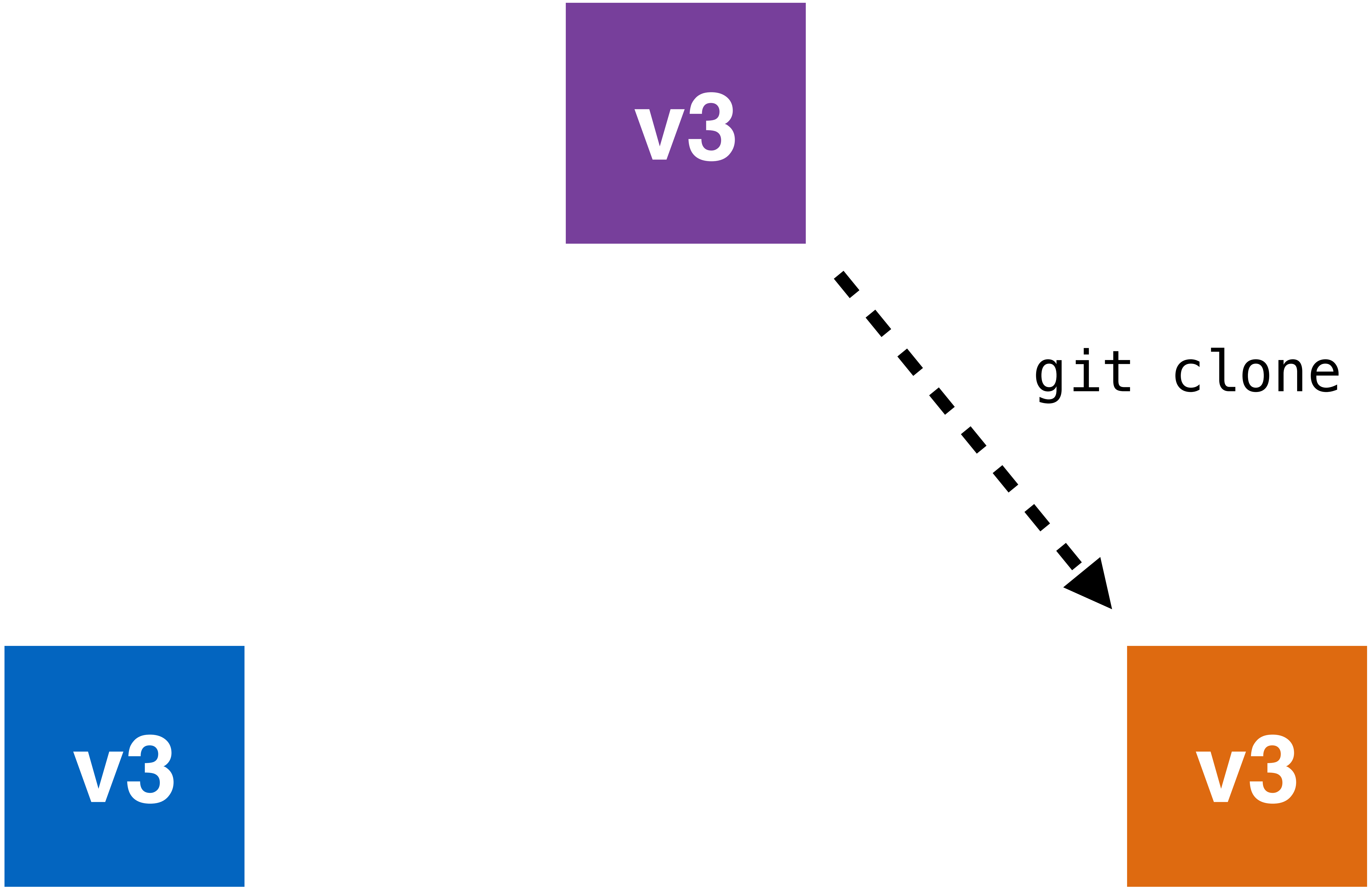
Remote is up-to-date

Also has version history (`git log`)

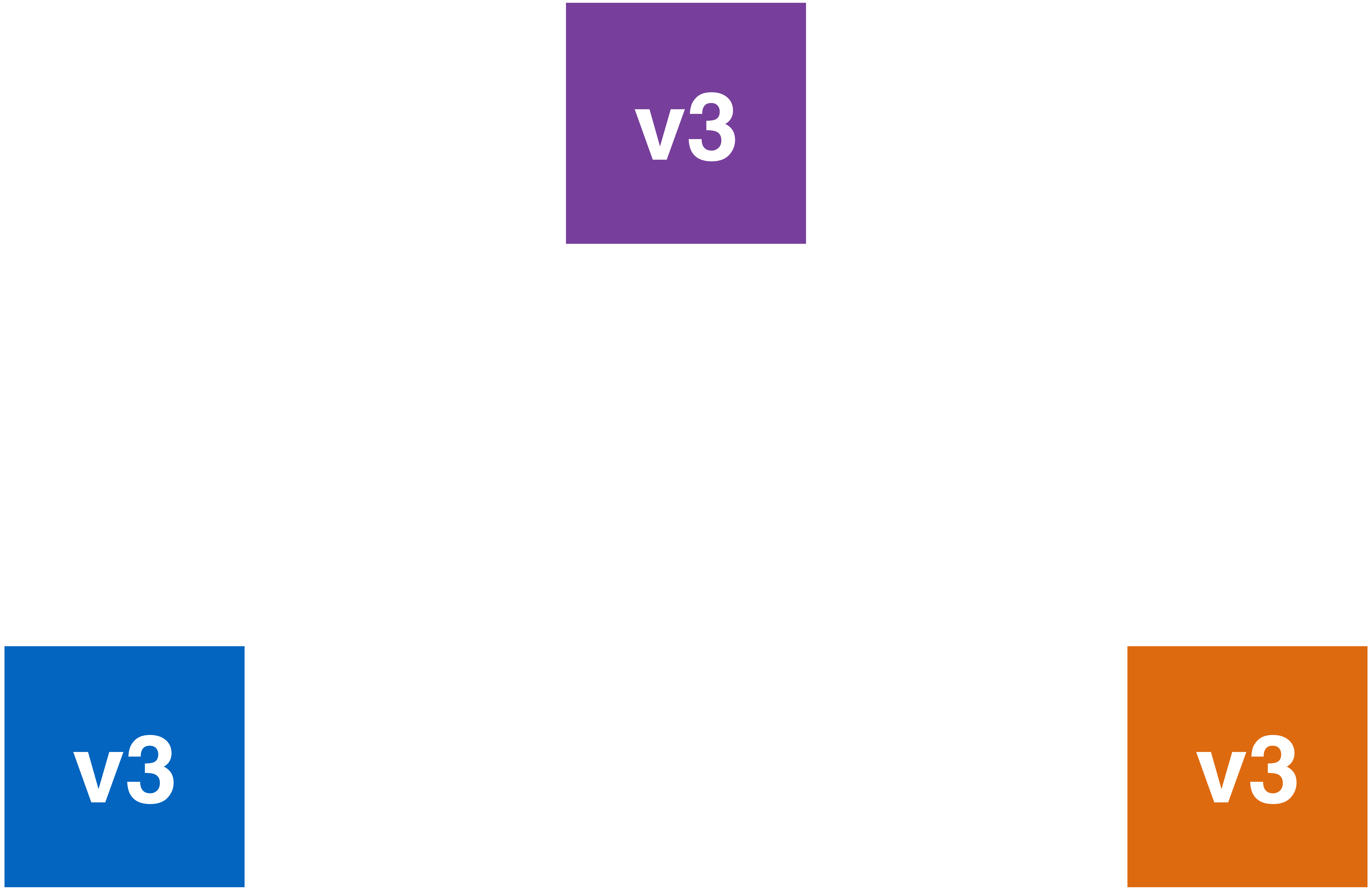

a file on your computer
v3
same file on another computer (remote)
v3

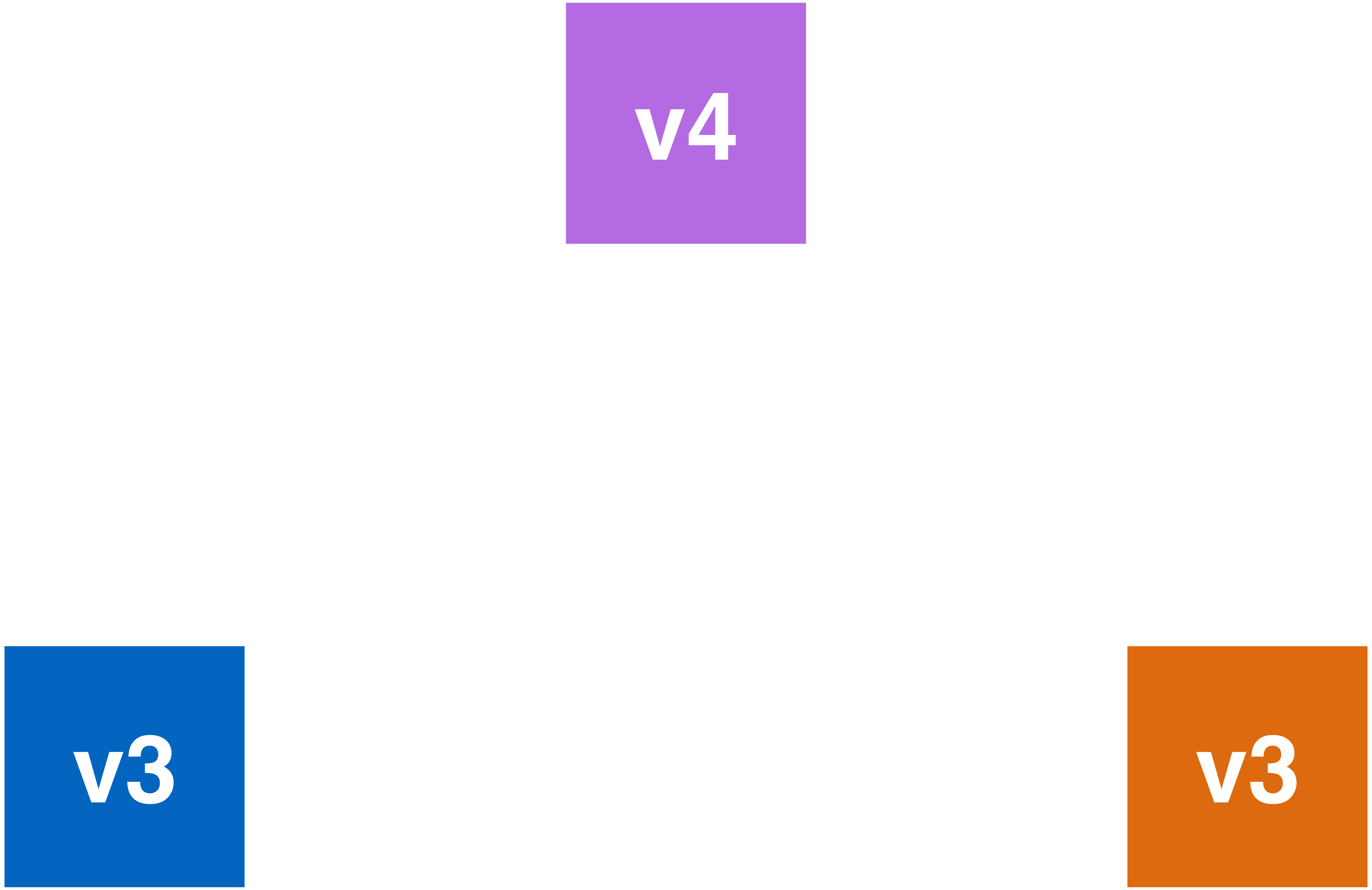

my computer doesn’t have this project
v3
v3

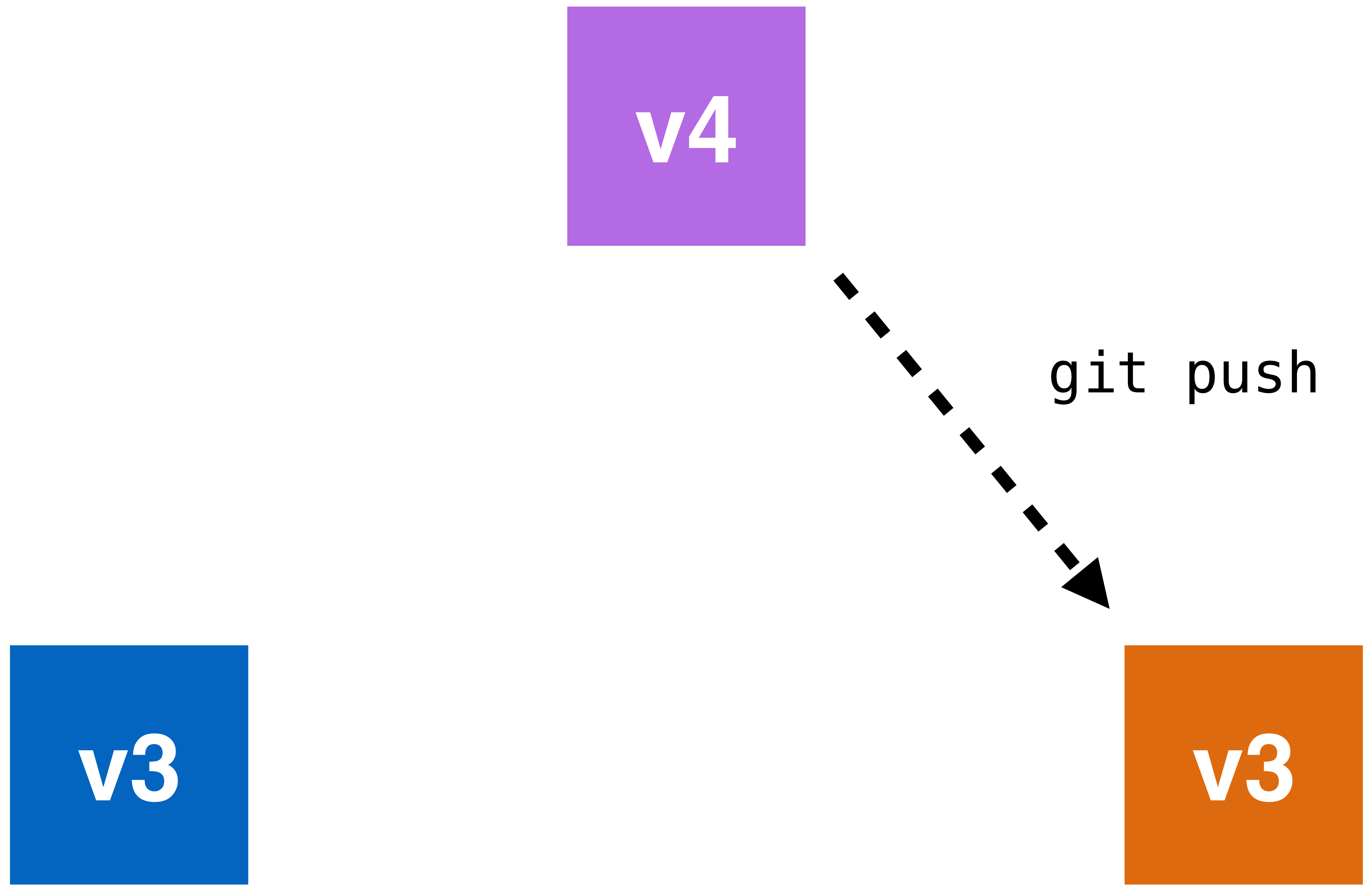

git clone
v3
v3

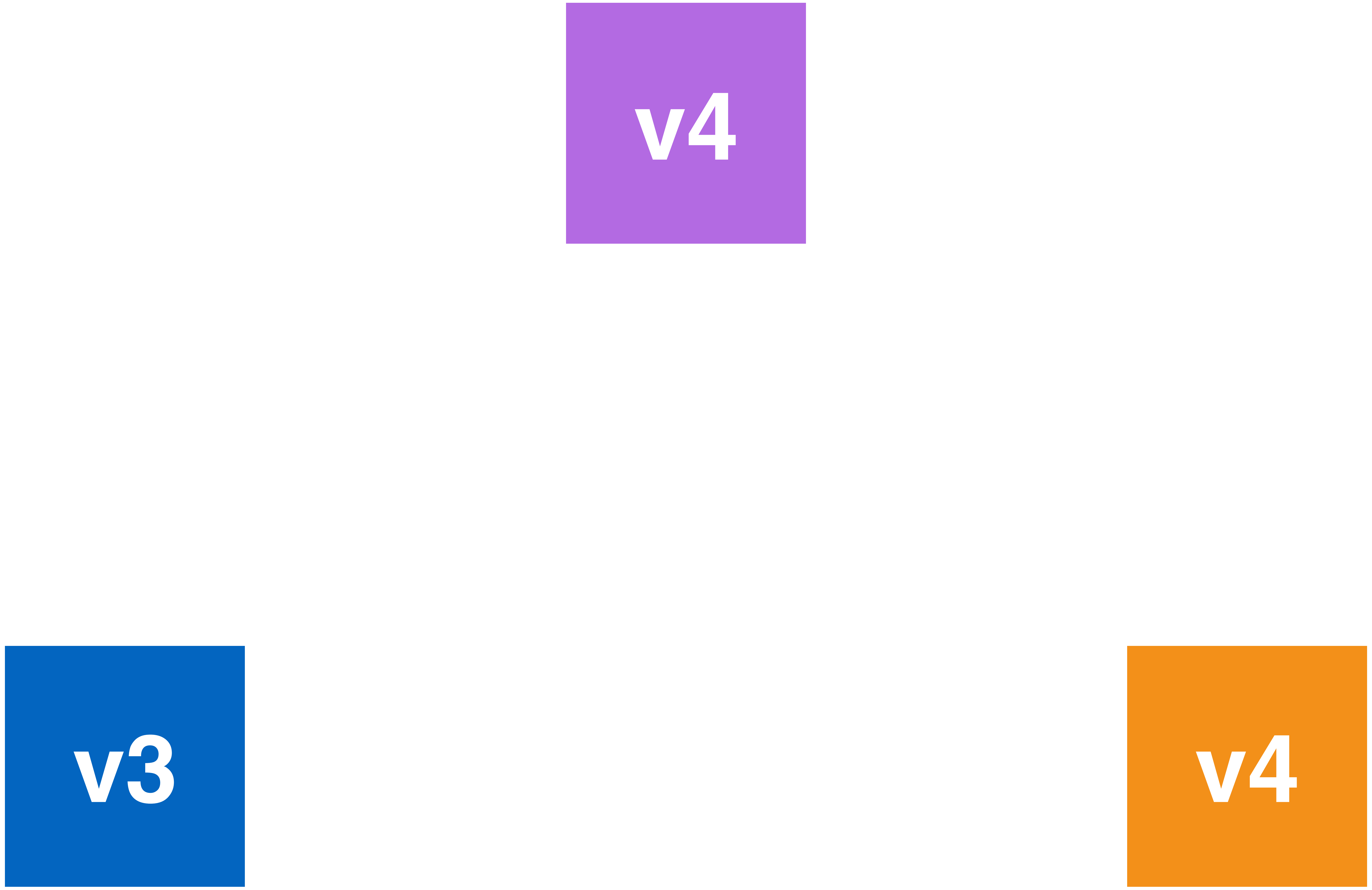

v3
git clone
v3
v3

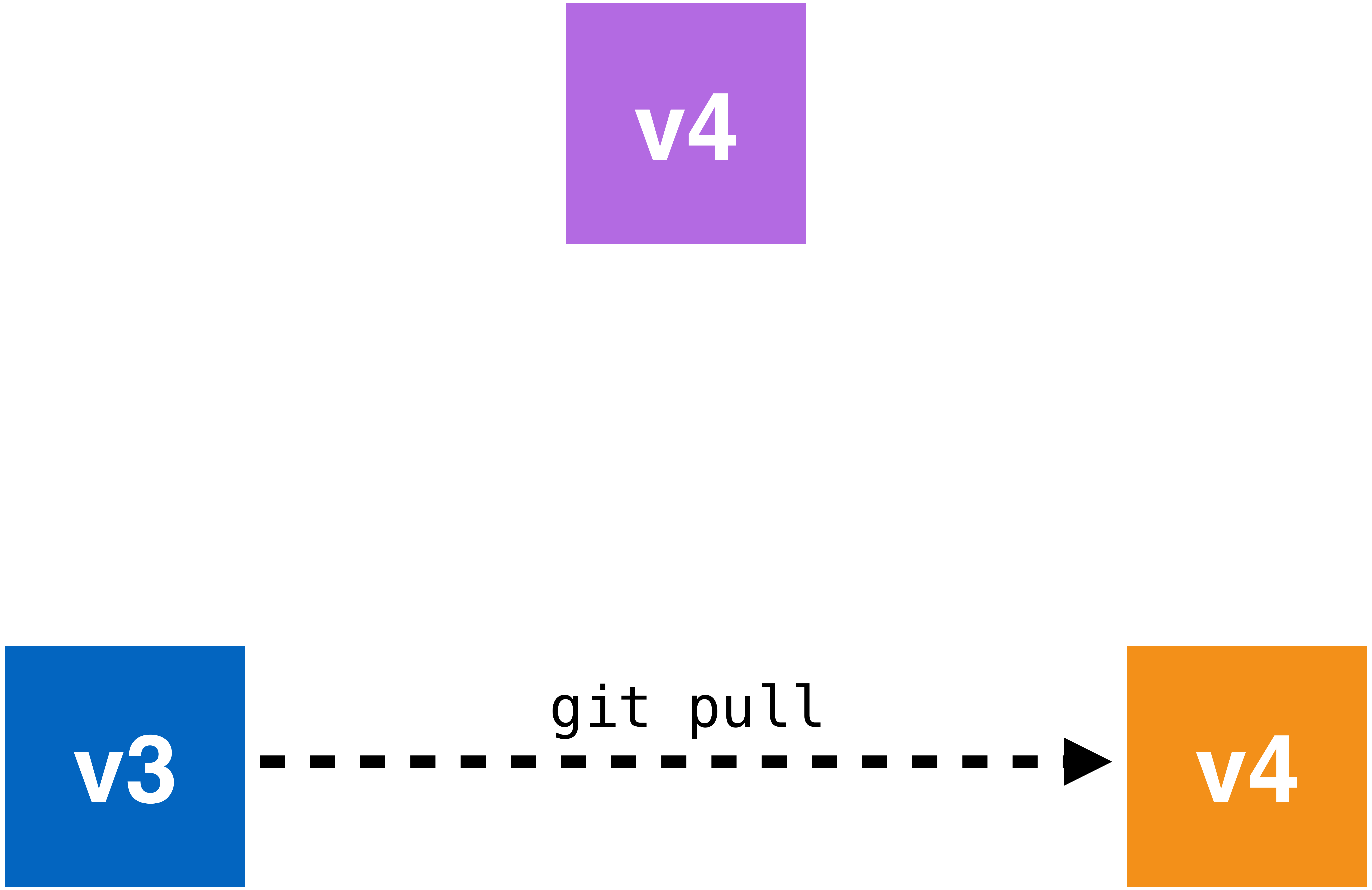
change + `commit`

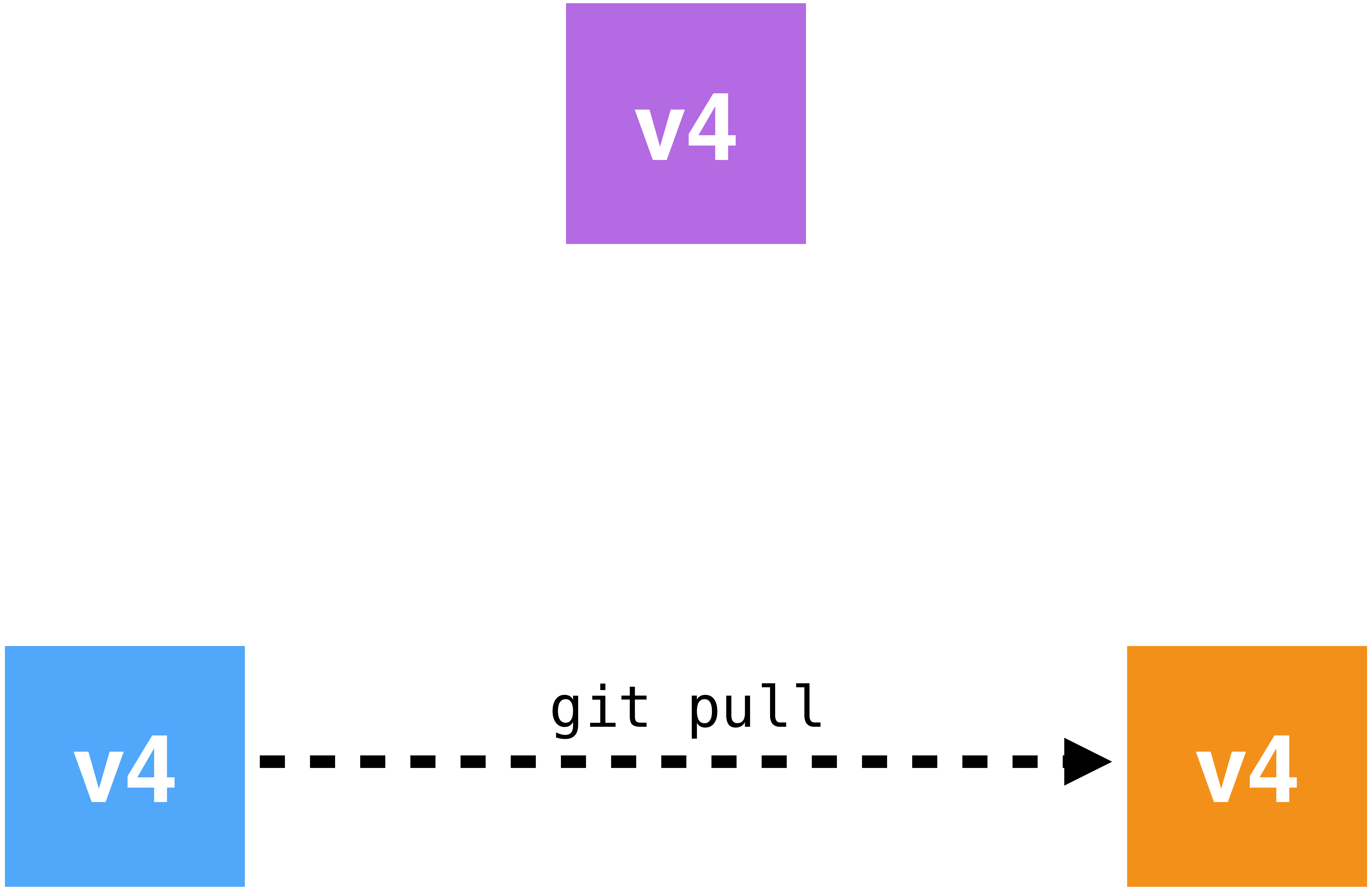


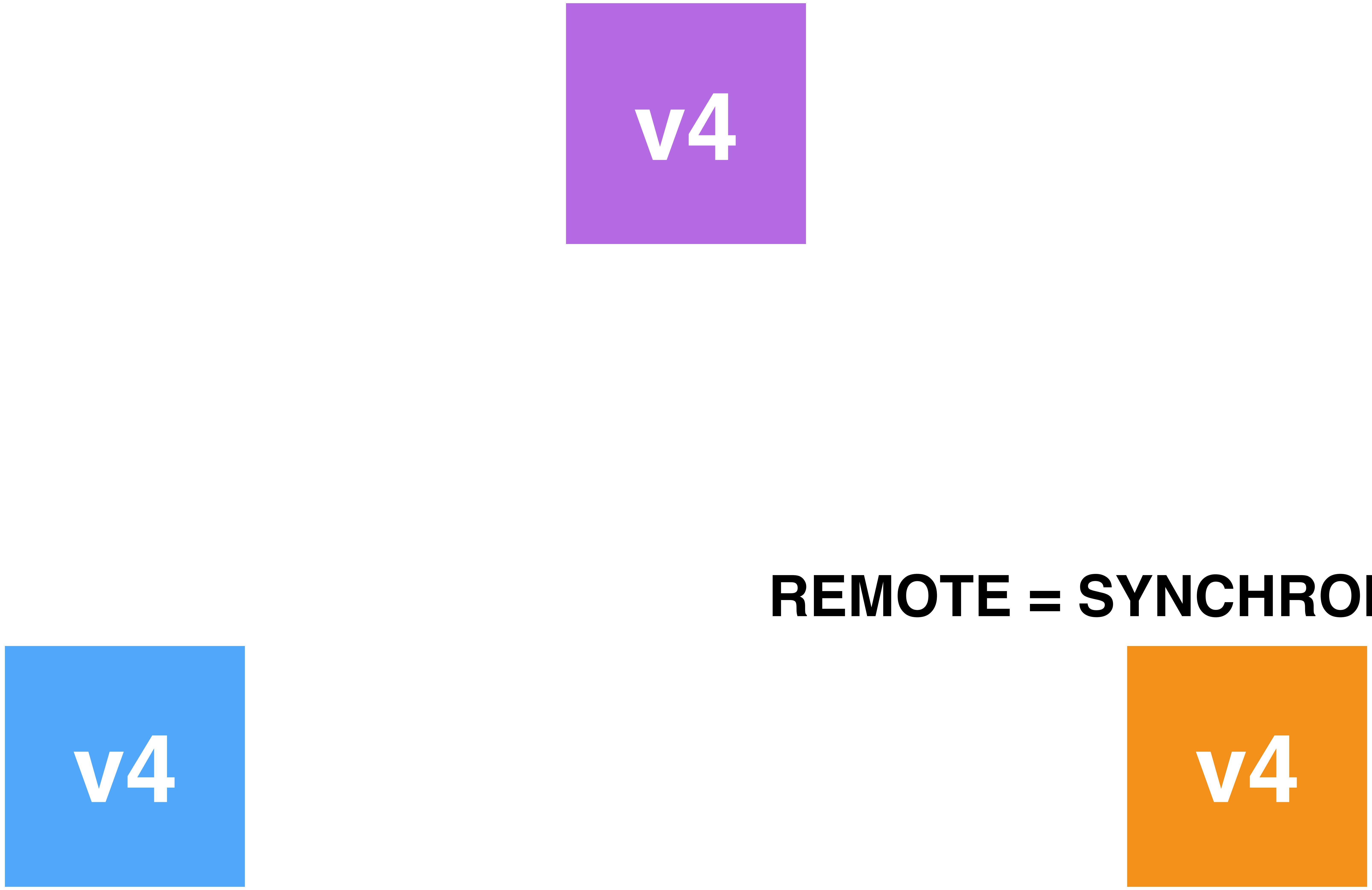

v4
v3
v3


v4
git push
v3
v3


v4
v3
v4


v4
v3
git pull
v4


v4
v4
git pull
v4



REMOTE = SYNCHRONIZATION PT.

Local vs remote copy

- Local copy is yours, no-one can see it unless you publish it by `push` to remote
- Remote as synchronization point, get changes by `pull`
- Etiquette note: **never push to remote if your code is broken**

http://gunshowcomic.com/648

Some different remotes

- GitHub: 100% the most popular one. It’s a very nice website. I host most of my code here. Public with possibility for private repos (free with academic license)
- Ice-git: located at the UiT dep’t of computer science. It’s private. The NOWAC documentation package lives here.

Honorable git mentions

- `log`
- `clone`
- `checkout`
- `branch`
- `merge`
- & so on

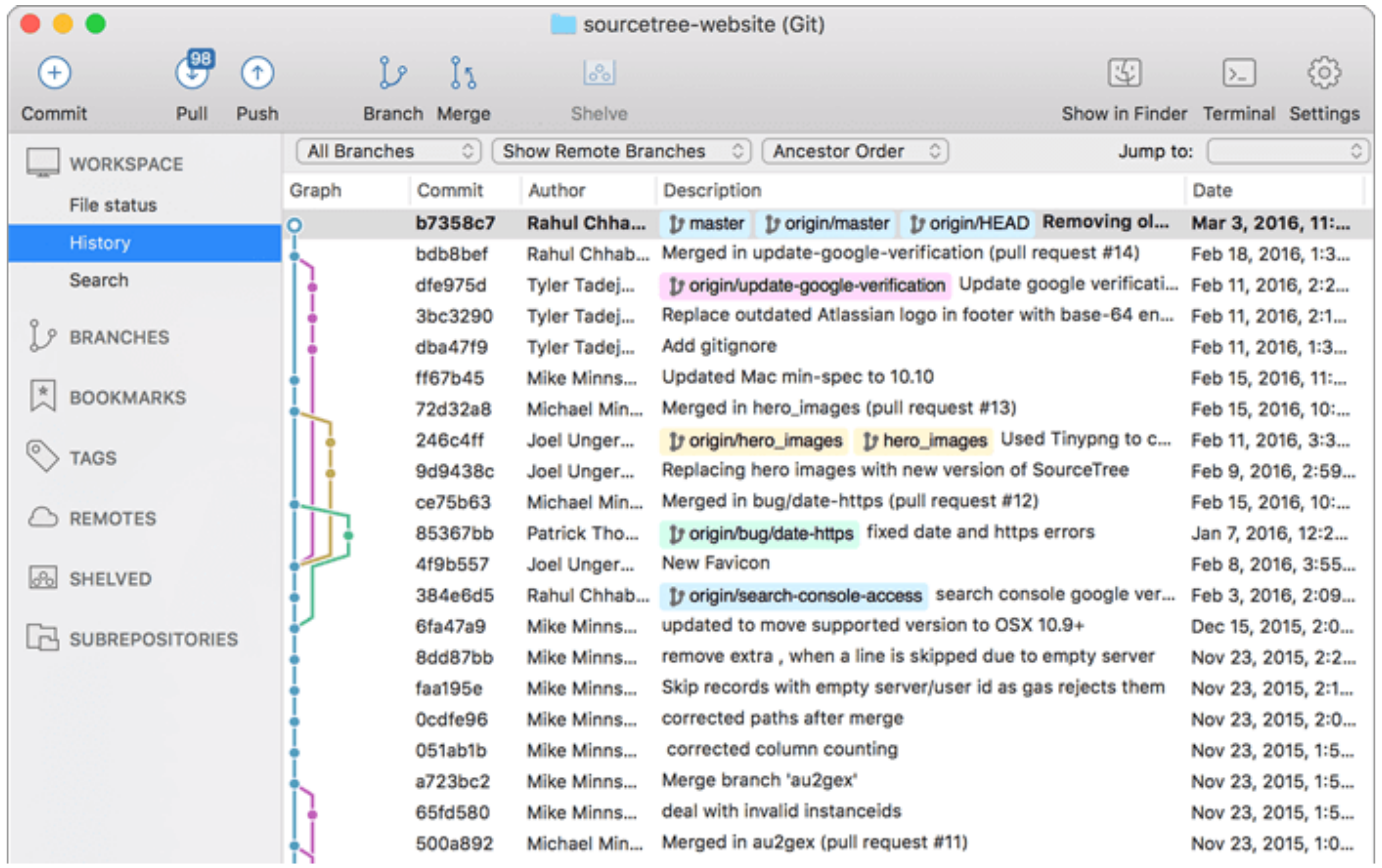
Branching/merging

- For keeping several versions or Branches of your code
- Eg. “development” and “release”
- Makes collaboration easier.
- You should learn about it, but learn the basics 1st.

https://try.github.io

Learn Git in 15 minutes: Go try all this, including branching, for yourself

SourceTree: a graphical user interface to git



Graph	Commit	Author	Description	Date
	b7358c7	Rahul Chha...	master origin/master origin/HEAD Removing ol...	Mar 3, 2016, 11:...
	bdb8bef	Rahul Chhab...	Merged in update-google-verification (pull request #14)	Feb 18, 2016, 1:3...
	dfe975d	Tyler Tadej...	origin/update-google-verification Update google verificati...	Feb 11, 2016, 2:2...
	3bc3290	Tyler Tadej...	Replace outdated Atlassian logo in footer with base-64 en...	Feb 11, 2016, 2:1...
	dba47f9	Tyler Tadej...	Add gitignore	Feb 11, 2016, 1:3...
	ff67b45	Mike Minns...	Updated Mac min-spec to 10.10	Feb 15, 2016, 11:...
	72d32a8	Michael Min...	Merged in hero_images (pull request #13)	Feb 15, 2016, 10:...
	246c4ff	Joel Unger...	origin/hero_images hero_images Used Tinypng to c...	Feb 11, 2016, 3:3...
	9d9438c	Joel Unger...	Replacing hero images with new version of SourceTree	Feb 9, 2016, 2:59...
	ce75b63	Michael Min...	Merged in bug/date-https (pull request #12)	Feb 15, 2016, 10:...
	85367bb	Patrick Tho...	origin/bug/date-https fixed date and https errors	Jan 7, 2016, 12:2...
	4f9b557	Joel Unger...	New Favicon	Feb 8, 2016, 3:55...
	384e6d5	Rahul Chhab...	origin/search-console-access search console google ver...	Feb 3, 2016, 2:09...
	6fa47a9	Mike Minns...	updated to move supported version to OSX 10.9+	Dec 15, 2015, 2:0...
	8dd87bb	Mike Minns...	remove extra , when a line is skipped due to empty server	Nov 23, 2015, 2:2...
	faa195e	Mike Minns...	Skip records with empty server/user id as gas rejects them	Nov 23, 2015, 2:1...
	0cdfe96	Mike Minns...	corrected paths after merge	Nov 23, 2015, 2:0...
	051ab1b	Mike Minns...	corrected column counting	Nov 23, 2015, 1:5...
	a723bc2	Mike Minns...	Merge branch 'au2gex'	Nov 23, 2015, 1:5...
	65fd580	Mike Minns...	deal with invalid instanceids	Nov 23, 2015, 1:5...
	500a892	Michael Min...	Merged in au2gex (pull request #11)	Nov 23, 2015, 1:0...

https://www.sourcetreeapp.com/

Take-home messages again

- You’re one out of four types of analyst: try to get to the better class
- You should think about the people you collaborate with
- You’re a programmer, so you may as well act like one
- Programmers use version control

Next time: Packages, GitHub

Links/References

- Github is a nice place to keep code: https://github.com/
- Github's interactive guide to git: https://try.github.io
- We use Hadley Wickham's style guide: http://adv-r.had.co.nz/Style.html
- Why you need version control: http://ellisp.github.io/blog/2016/09/16/version-control

Links/References

- Karl Broman tutorials: http://kbroman.org/pages/tutorials
- A successful Git branching model: http://nvie.com/posts/a-successful-git-branching-model/

UiT THE ARCTIC UNIVERSITY OF NORWAY

Thank you!

Einar Holsbø
einar@cs.uit.no

Slides available at
3inar.github.io/talks/