Preregistration, Open Science & GitHub

Judith Degen & Sebastian Schuster
LSA 2021 minicourse

Jan 7, 2021

A typical psycholinguistics study

Hypothesis

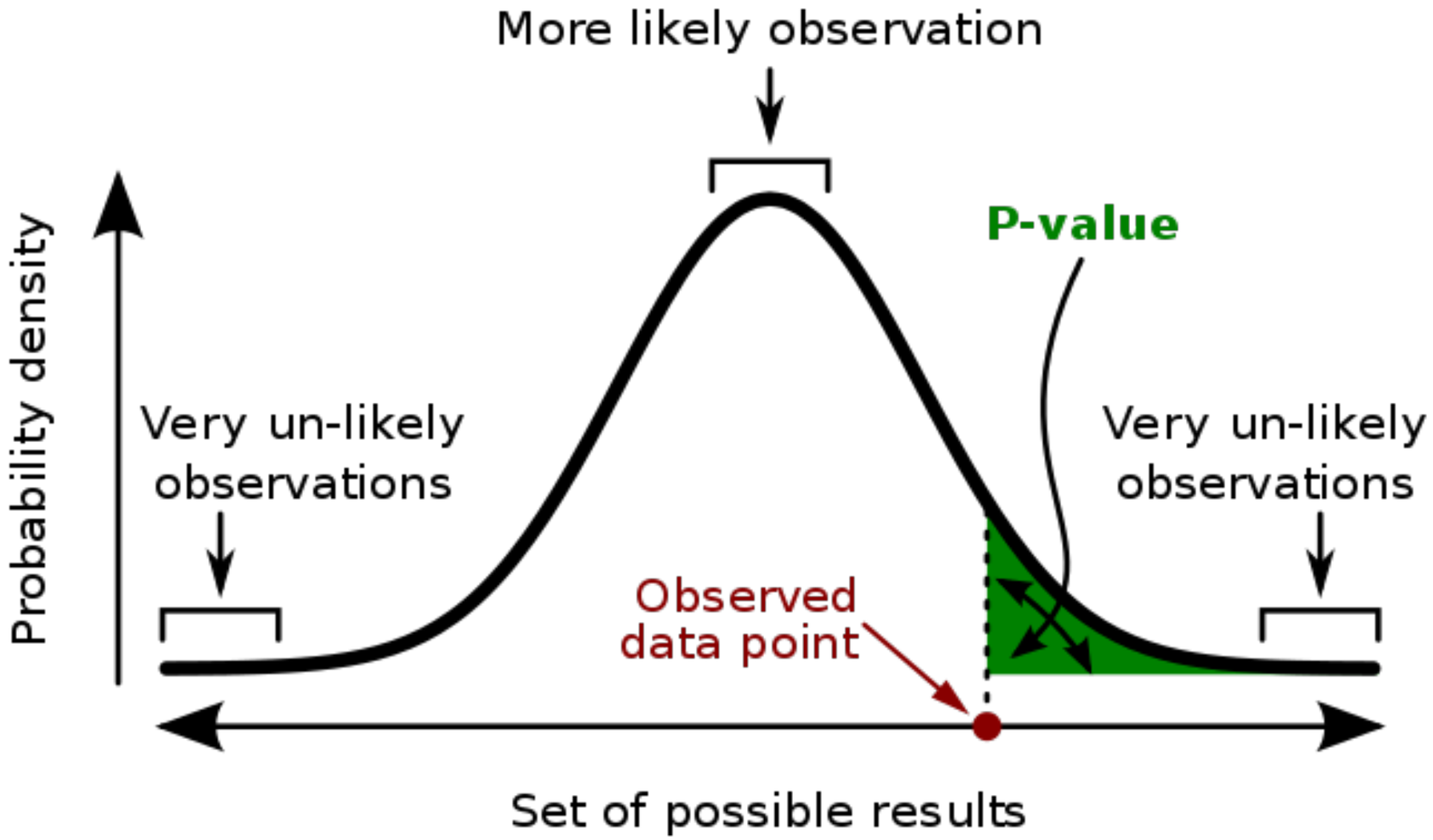
Reading of sentences with reduced relative clauses is slower than reading of sentences with overt complementizer

H_0: Average reading times of both sentence types are equal

The horse raced past the barn fell into a ditch

The horse that was raced past the barn fell into a ditch

p-values



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

Sketchy things you can do to get a significant *p*-val

1. Number of subjects per condition

    a. Run 10 subjects per condition

    b. Perform a t-test

    c. If $p < .05$: Publish paper!

    Otherwise: Go to step a.

Simmons et al. (2011)

Sketchy things you can do to get a significant *p*-val

2. Have multiple dependent variables

    a. Run tests to predict each of the variables

    b. Pick the dependent variable that gives you a significant *p*-value

The horse raced past the barn fell into a ditch
The horse that was raced past the barn fell into a ditch

Simmons et al. (2011)

Sketchy things you can do to get a significant *p*-val

3. Run models with many different independent variables

    a. Have a set of many independent variables

    b. Run models with various combinations and interactions until your manipulation is significant

Simmons et al. (2011)

Sketchy things you can do to get a significant *p*-val

4. Have conditions that you don't report on

    a. Run n > 2 conditions

    b. Pick 2 conditions which differ significantly and don't tell anybody about the other conditions

DON'T DO ANY OF THESE THINGS!!!

Simmons et al. (2011)

Simulated false-positive rates

	Significance level		
Researcher degrees of freedom	$p < .1$	$p < .05$	$p < .01$
Situation A: two dependent variables ($r = .50$)	17.8%	9.5%	2.2%

Simmons et al. (2011)

Simulated false-positive rates

Researcher degrees of freedom	Significance level		
	$p < .1$	$p < .05$	$p < .01$
Situation A: two dependent variables ($r = .50$)	17.8%	9.5%	2.2%
Situation B: addition of 10 more observations per cell	14.5%	7.7%	1.6%

Simmons et al. (2011)

Simulated false-positive rates

Researcher degrees of freedom	Significance level		
	$p < .1$	$p < .05$	$p < .01$
Situation A: two dependent variables ($r = .50$)	17.8%	9.5%	2.2%
Situation B: addition of 10 more observations per cell	14.5%	7.7%	1.6%
Situation C: controlling for gender or interaction of gender with treatment	21.6%	11.7%	2.7%

Simmons et al. (2011)

Simulated false-positive rates

Researcher degrees of freedom	Significance level		
	$p < .1$	$p < .05$	$p < .01$
Situation A: two dependent variables ($r = .50$)	17.8%	9.5%	2.2%
Situation B: addition of 10 more observations per cell	14.5%	7.7%	1.6%
Situation C: controlling for gender or interaction of gender with treatment	21.6%	11.7%	2.7%
Situation D: dropping (or not dropping) one of three conditions	23.2%	12.6%	2.8%

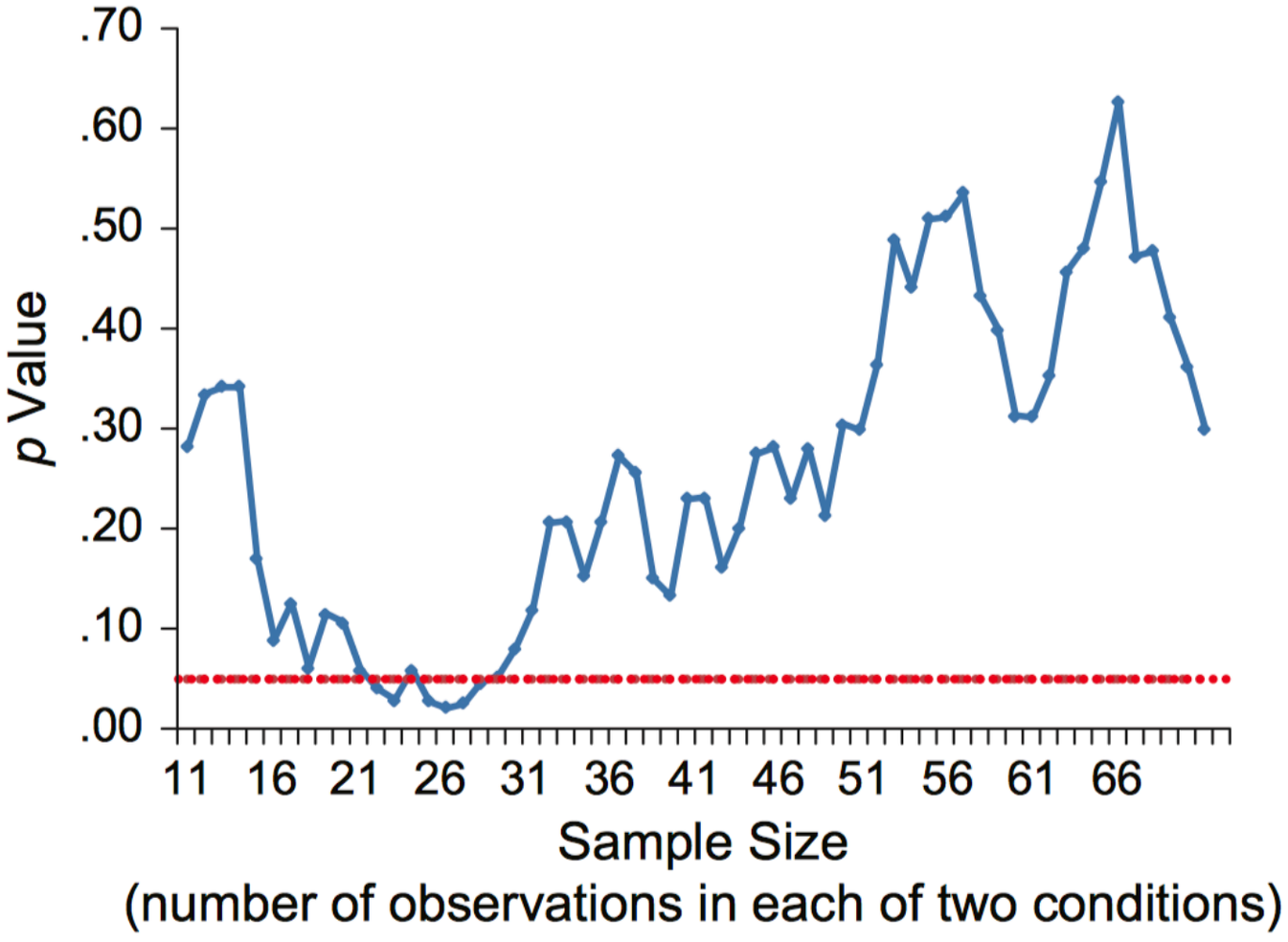
Simmons et al. (2011)

Simulated false-positive rates

Researcher degrees of freedom	Significance level		
	$p < .1$	$p < .05$	$p < .01$
Situation A: two dependent variables ($r = .50$)	17.8%	9.5%	2.2%
Situation B: addition of 10 more observations per cell	14.5%	7.7%	1.6%
Situation C: controlling for gender or interaction of gender with treatment	21.6%	11.7%	2.7%
Situation D: dropping (or not dropping) one of three conditions	23.2%	12.6%	2.8%
Combine Situations A and B	26.0%	14.4%	3.3%
Combine Situations A, B, and C	50.9%	30.9%	8.4%
Combine Situations A, B, C, and D	81.5%	60.7%	21.5%

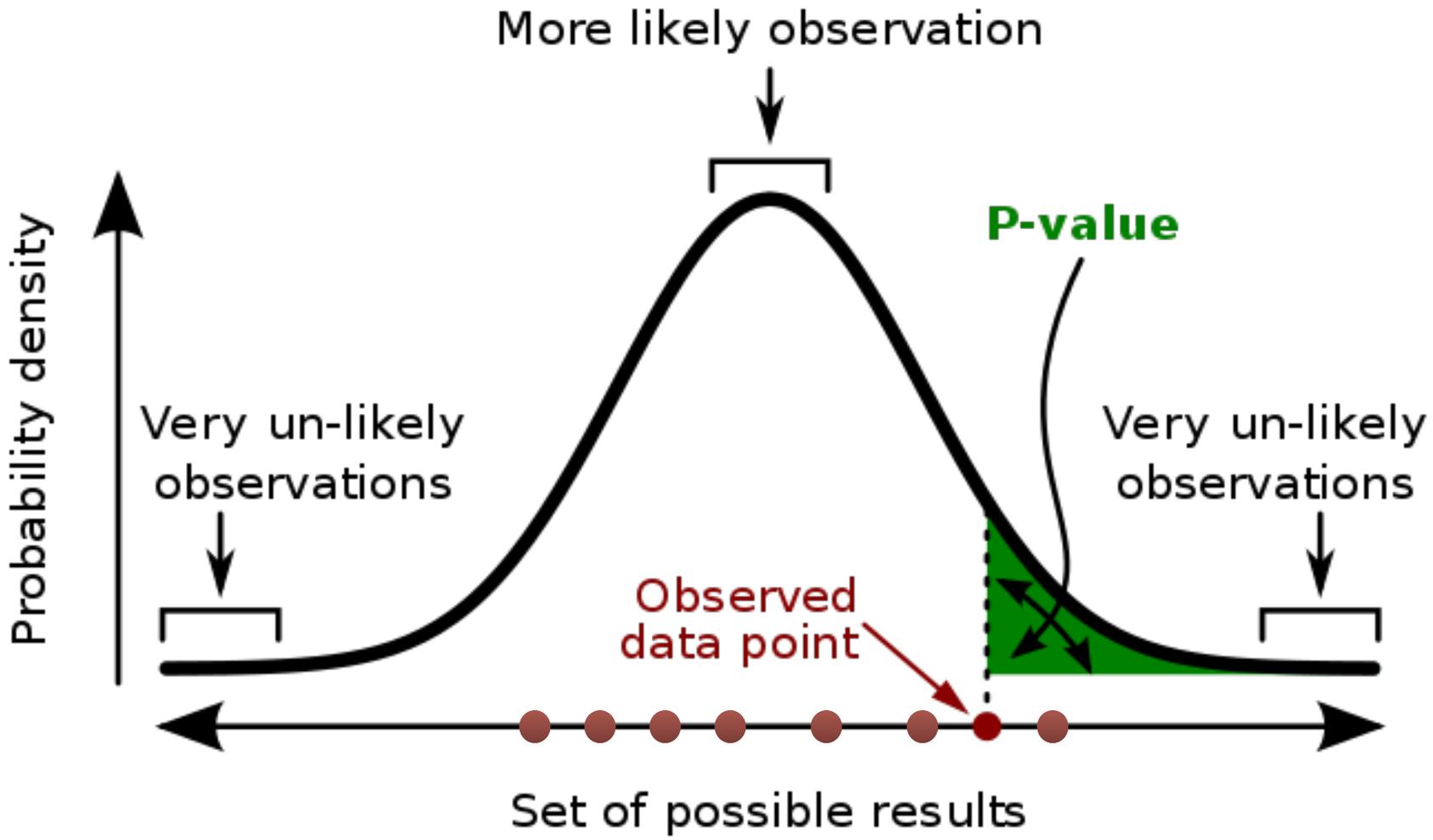
Simmons et al. (2011)

Simulated false-positive rates



Simmons et al. (2011)

p-values



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

Pre-registration

- To keep *p*-value meaningful, fix the following things **before collecting data**
    1. number of subjects you'll run
    2. exclusion criteria:
       Which data points are you going to exclude from your analysis
    3. dependent variable
    4. independent variables
    5. experimental conditions

Preregistration

- Preregistering provides you (and reviewers and readers of your paper) with proof that you actually fixed all these things

- Only requires filling out a short questionnaire which is permanently stored on a pre-registration platform

A preregistration workflow

1. Come up with and implement experiment
2. Run pilot study with 2-4 subjects
3. Write analysis scripts and test them with pilot data
4. Preregister study and upload analysis scripts to OSF
5. Run actual study
6. Analyze data with pre-registered analysis script
7. (optional) Do **exploratory** post-hoc analyses

What you'll do today

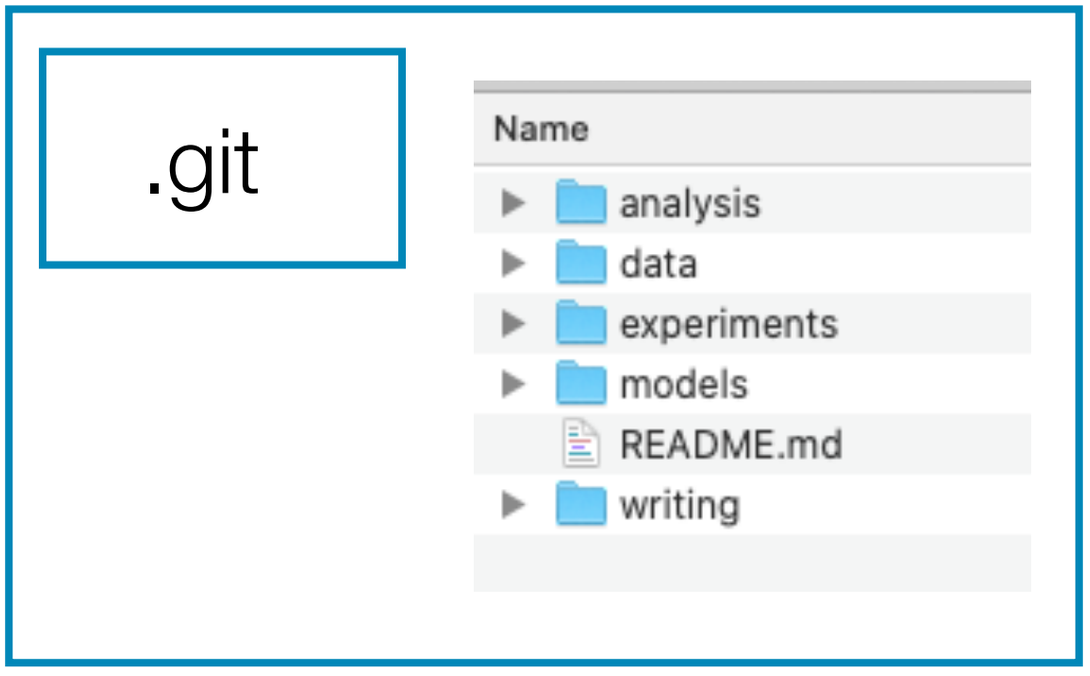
1. Implement experiment
2. Preregister study
3. Test experiment
4. Write analysis scripts
5. Analyze pooled data from the class

Git & GitHub

What is Git?

- A version control (VC) system that keeps a history of all previous versions of files
    - allows you to go back to previous versions of a file anytime

Repository:



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

sebschu / my-project Template

Unwatch 2 Star 0 Fork 0

Code Issues Pull requests Actions Projects Wiki Security Insights

master 1 branch 0 tags Go to file Add file Code Use this template

thegrican added analysis script ✓ 7846e97 8 hours ago 10 commits

analysis	added analysis script	8 hours ago
data	added analysis script	8 hours ago
experiments/01_implicat...	updated experiment template	9 hours ago
.gitignore	rename experiment	3 days ago
README.md	Initial commit	3 days ago

README.md

my-project

Template repository for web-based experiment projects.

About

Template repository for web-based experiment projects.

Readme

Releases

No releases published
Create a new release

Packages

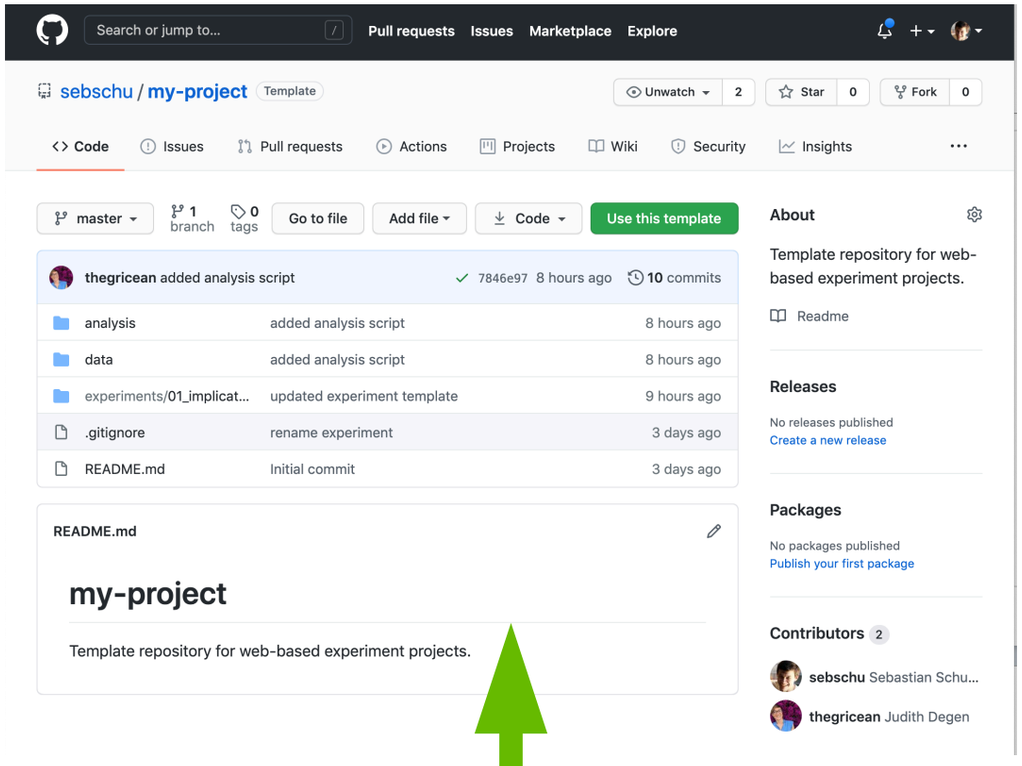
No packages published
Publish your first package

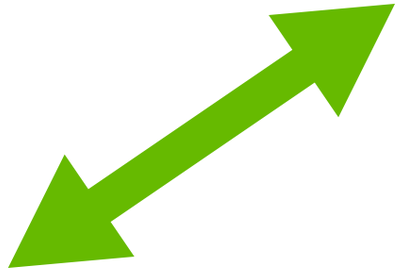
Contributors 2

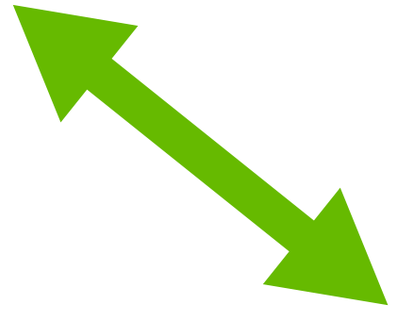
sebschu Sebastian Schu...

thegrican Judith Degen

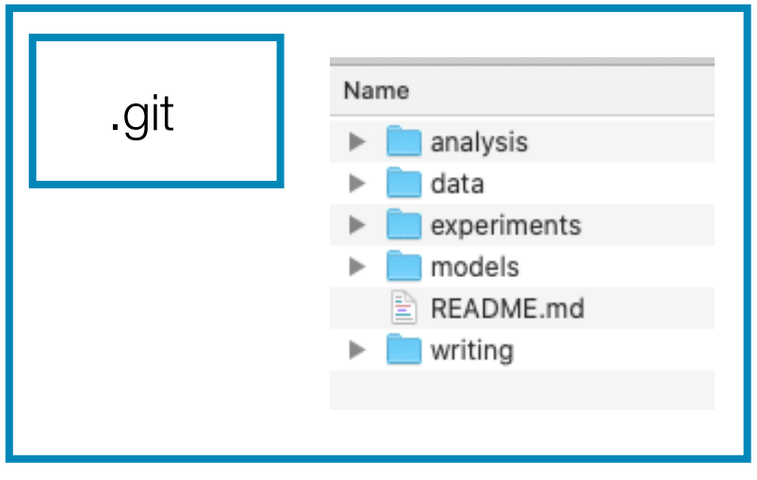
Git and GitHub



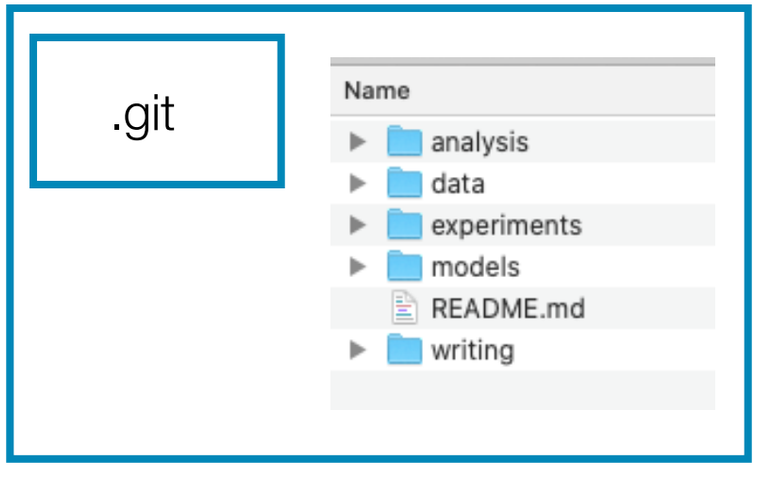




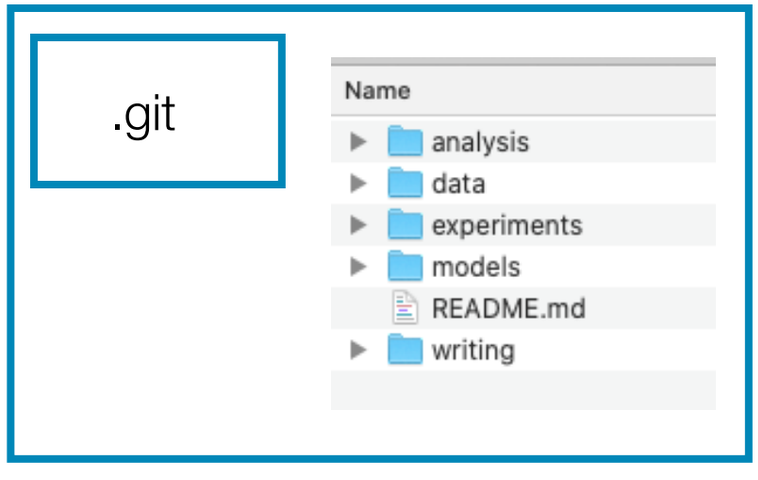
Computer A



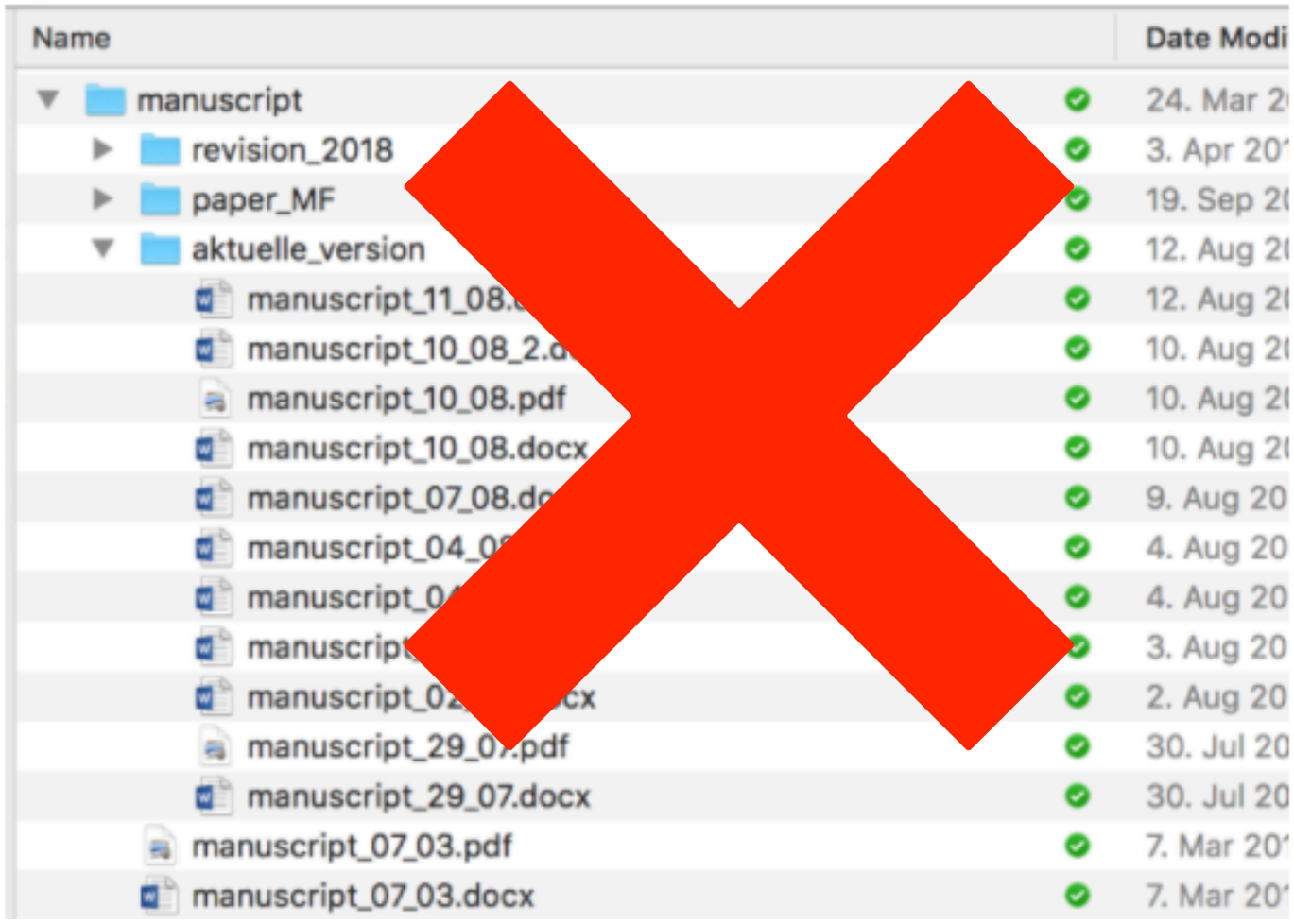
Computer B



Computer C



Why use Git and GitHub?



No more manuscript_v2_oct15_final_revisions_v3_seb.pdf!

Why use Git and GitHub?

- Easy synchronization of project folder between people: makes it easy to **collaborate** (no more emailing of files)

- Provides **backups** of your data and project files

- Promotes **open science**: Makes all your project data and experiments publicly available

- You can set up **automatic hosting** of your experiments on GitHub

Basic structure of a research repository

- **my-project**/
  - **experiments**: code necessary to run experiments
  - **data**: anonymized (!) data from experiments
  - **analysis:** scripts to analyze and visualize data
  - **writing:** manuscripts, papers, etc.

Add folders for each experiment into experiments/data/ analysis folders: e.g., 1_pilot, 2_implicature_strength, …

Interacting with a repository

1. **Create and edit files** as you would in a regular folder

2. To preserve the current state of files, **commit the current version** to the repository

3. To make your recent commits available on GitHub and to you collaborators, **push** the repository to GitHub

4. To update your local repository with changes from other people, **pull** the repository from GitHub

BASIC CONCEPTS & TERMINOLOGY

pulling: download changes from server

tracked files: marked to be under VC

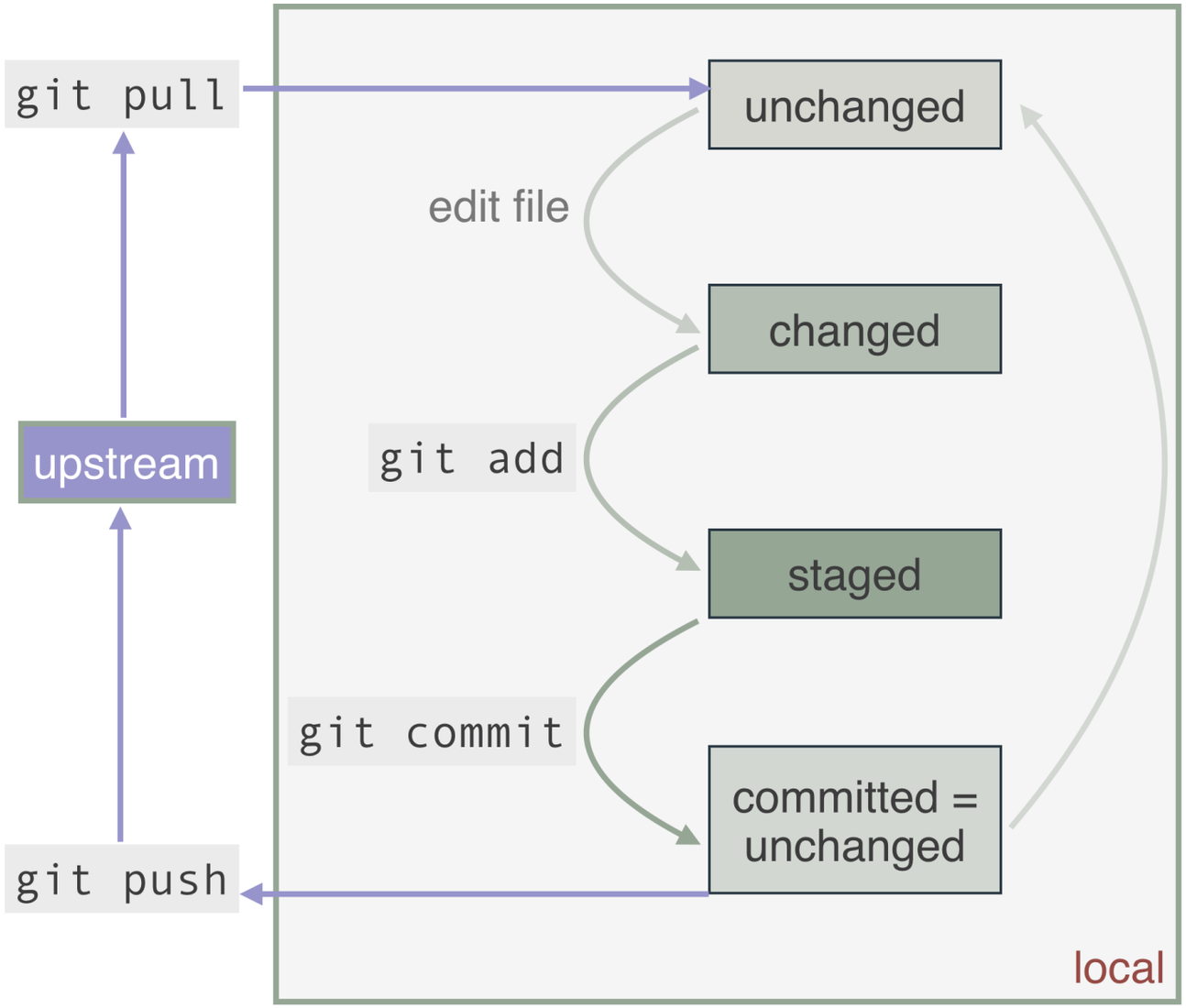
ignores: file types excluded from VC

adding/staging: mark local changes as to be committed

commit: save local changes locally

stage area (aka index): everything that is locally committed but not yet pushed

pushing: upload local changes to server



Open Science book by Michael Franke

Now…

1. Create your own repository and edit files
2. Modify an existing experiment
3. Preregister the experiment