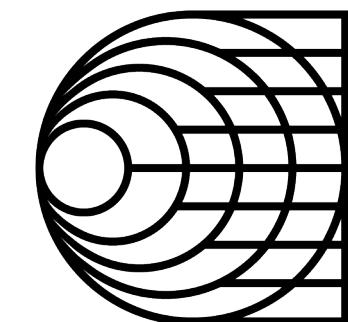
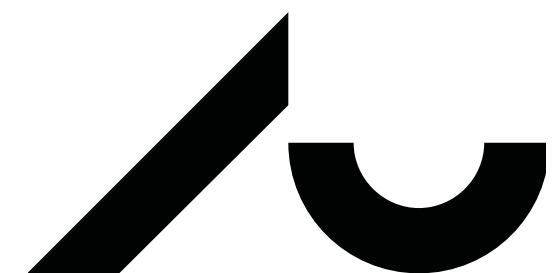


# NLP Class

## Class 1: Tech Stack & Demo of Foundation Models

Jan & Sara | 2024



CENTER FOR  
HUMANITIES  
COMPUTING

# Agenda

Technical concepts and tools we will need for this course

- cloud computing, version control
- computing resources available (TBD)

Demo of using a generative model to solve a problem

- we will try to run a LLM locally & give it some instructions
- a little sentiment classification experiment



# Technical Concepts and Tools

---

## Cloud computing

- This is what you can use to run stuff
- Industry standard

## Version control

- We are sharing course resources that way
- Industry standard



Sources  
& Notes

# Cloud: What problems are we trying to solve?

---

## Standard Environment

- Focus on coding, not the setup

## Scalability and Resources

- Run jobs that are too big for your own computer

## Tech Support

- Is easier on cloud



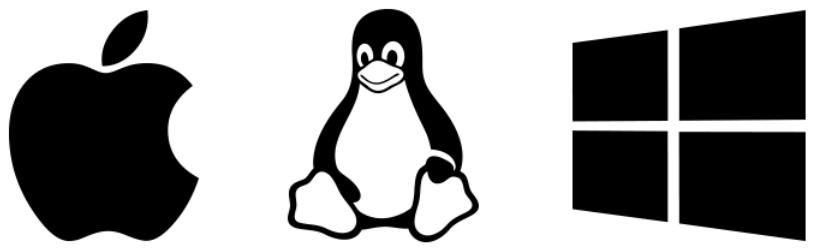
Sources  
& Notes

# Cloud: Standard Environment

---



Python version



OS

numpy 2.1.1

torch 2.4.1

Package versions



CUDA®

System dependencies



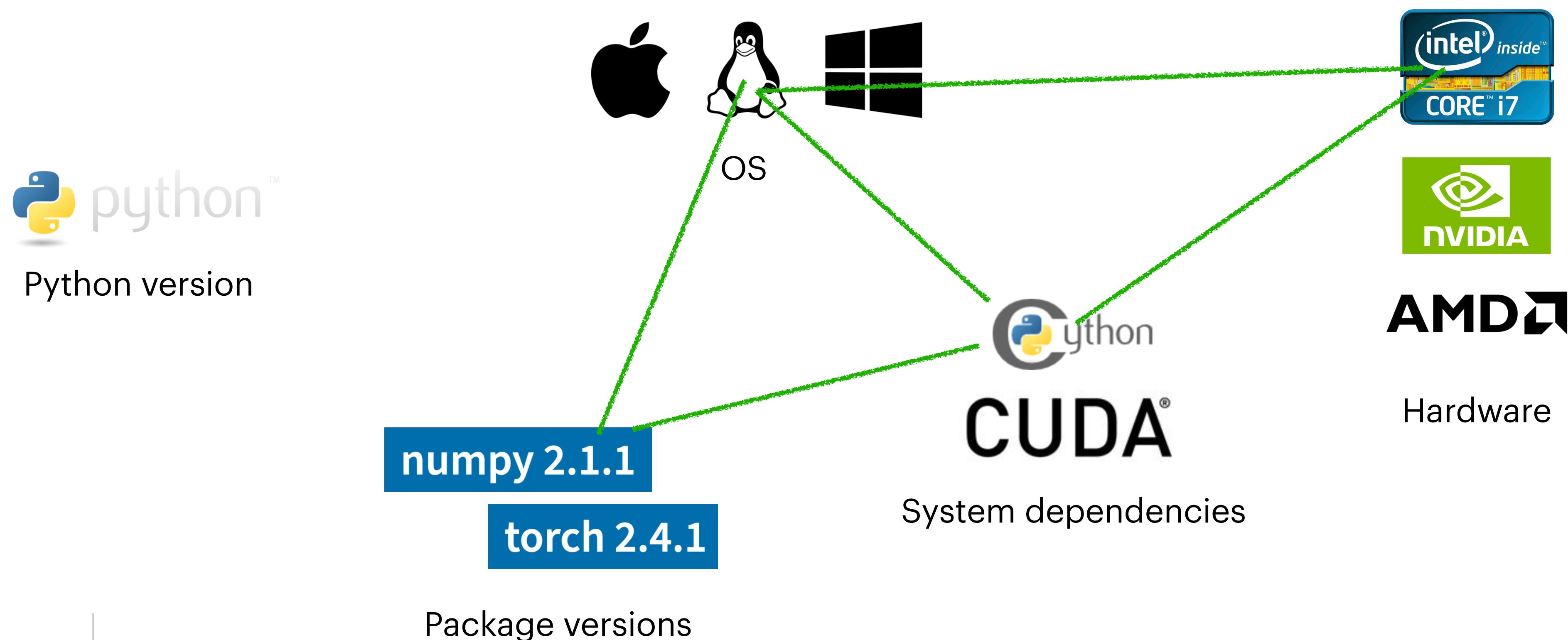
Hardware



Sources  
& Notes

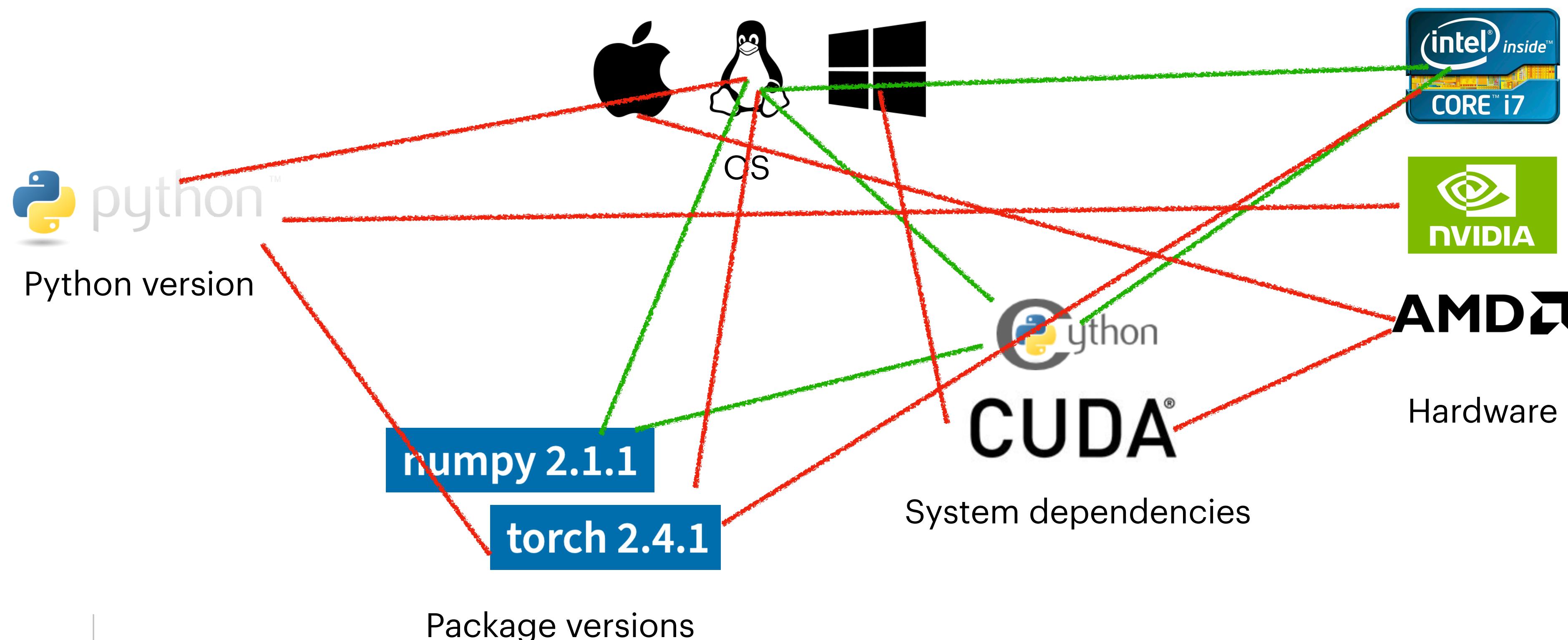
# Cloud: Standard Environment

---



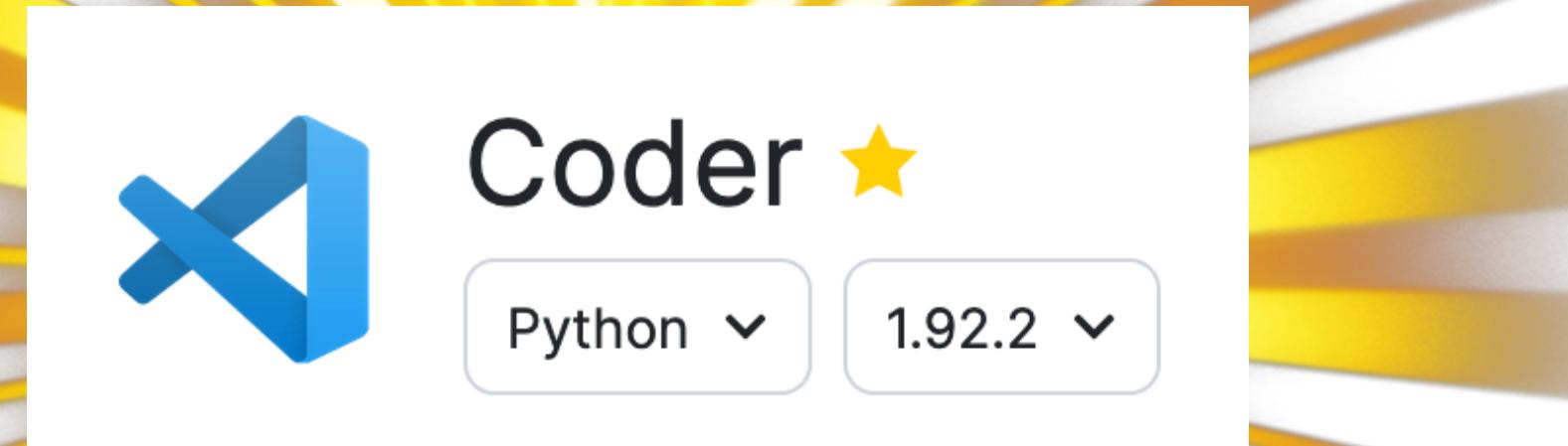
# Cloud: Standard Environment

---



Sources  
& Notes

# Cloud: Standard Environment

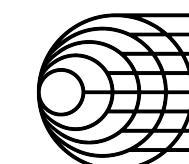


- **Operating System:** Ubuntu 24.04
- **Terminal:** tini 0.19.0 tmux 3.4
- **Shell:** Bash 5.2.21 fish 3.7.0 zsh 5.9
- **Editor:** GNU nano 7.2 Vim 9.1
- **Package Manager:** apt 2.7.14 dpkg 1.22.6 npm 9.2.0 pip 24.0
- **Programming Language:** GCC 13.2.0 Python 3.12.3
- **Utility:** EasyBuild 4.9.2 Lmod 8.7
- **Extension:** Open MPI 4.1.6
- **VSCode Extensions:** Live Share 1.0.5936 vscode-remote-extensionpack 0.25.0



Sources  
& Notes

<https://docs.cloud.sdu.dk/Apps/coder.html>



CENTER FOR  
HUMANITIES  
COMPUTING

# Cloud: Scalability and Resources

---

Too big to run locally.

Infrastructure for your exam project.



Sources  
& Notes



# Cloud: UCloud

---

We will have shared resources for the whole course.

We will have some GPUs



Sources  
& Notes



# **Cloud: Alignment of Expectations**

---

You can use whatever works for you.

But, we can do tech support only for UCloud.

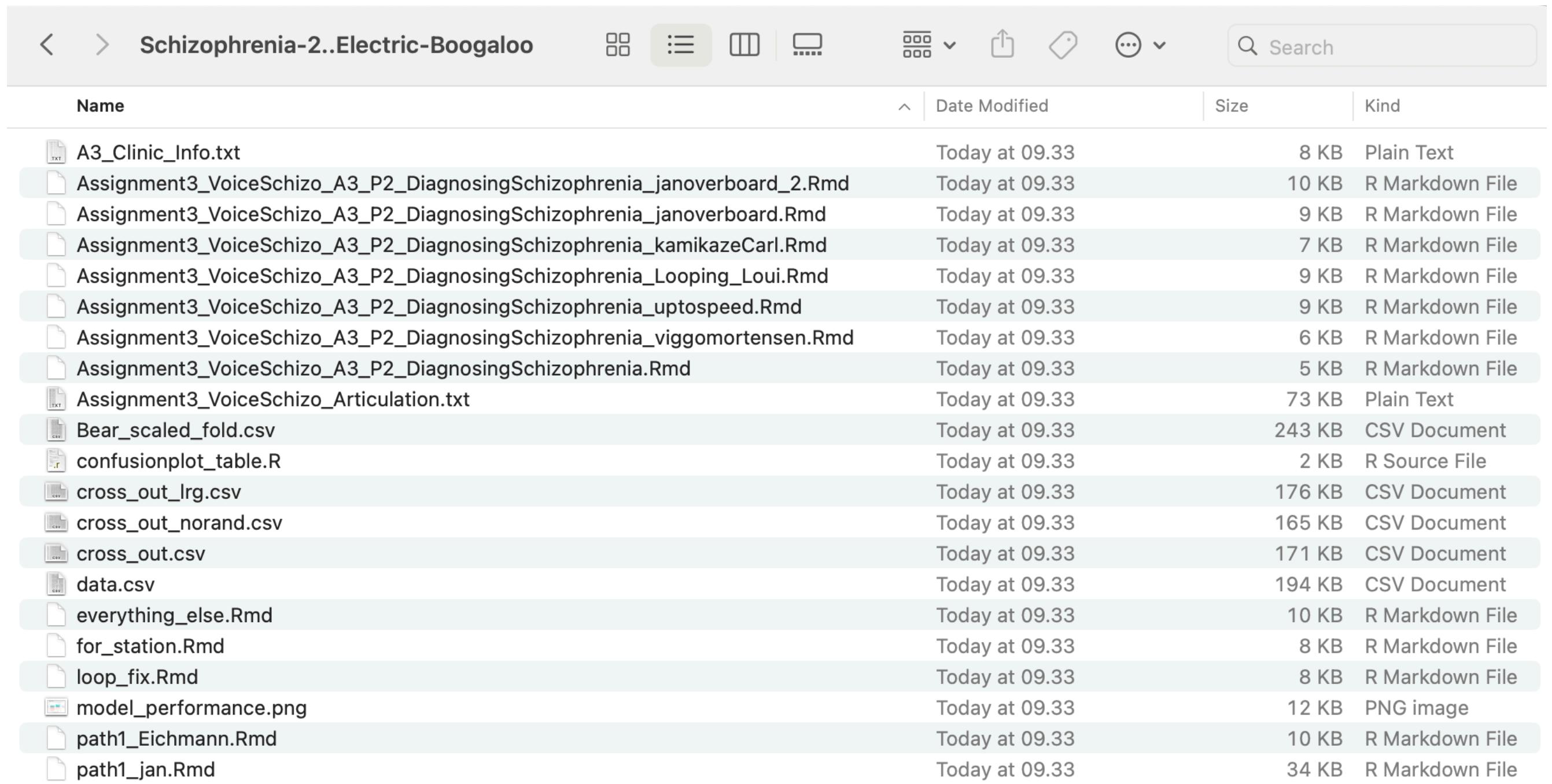


Sources  
& Notes

# Version Control: What problems?

---

Have you ever had a million versions of the same file?

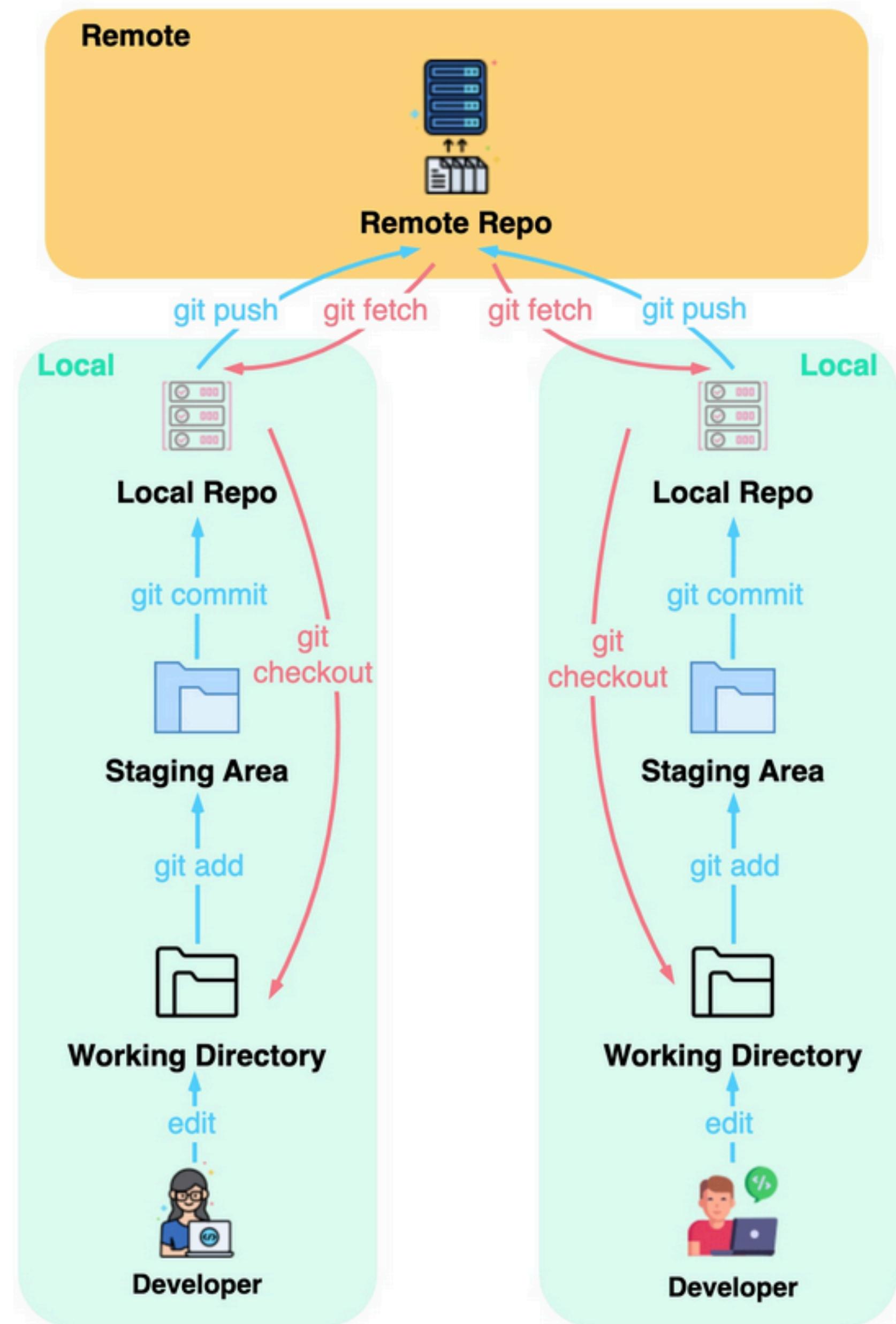
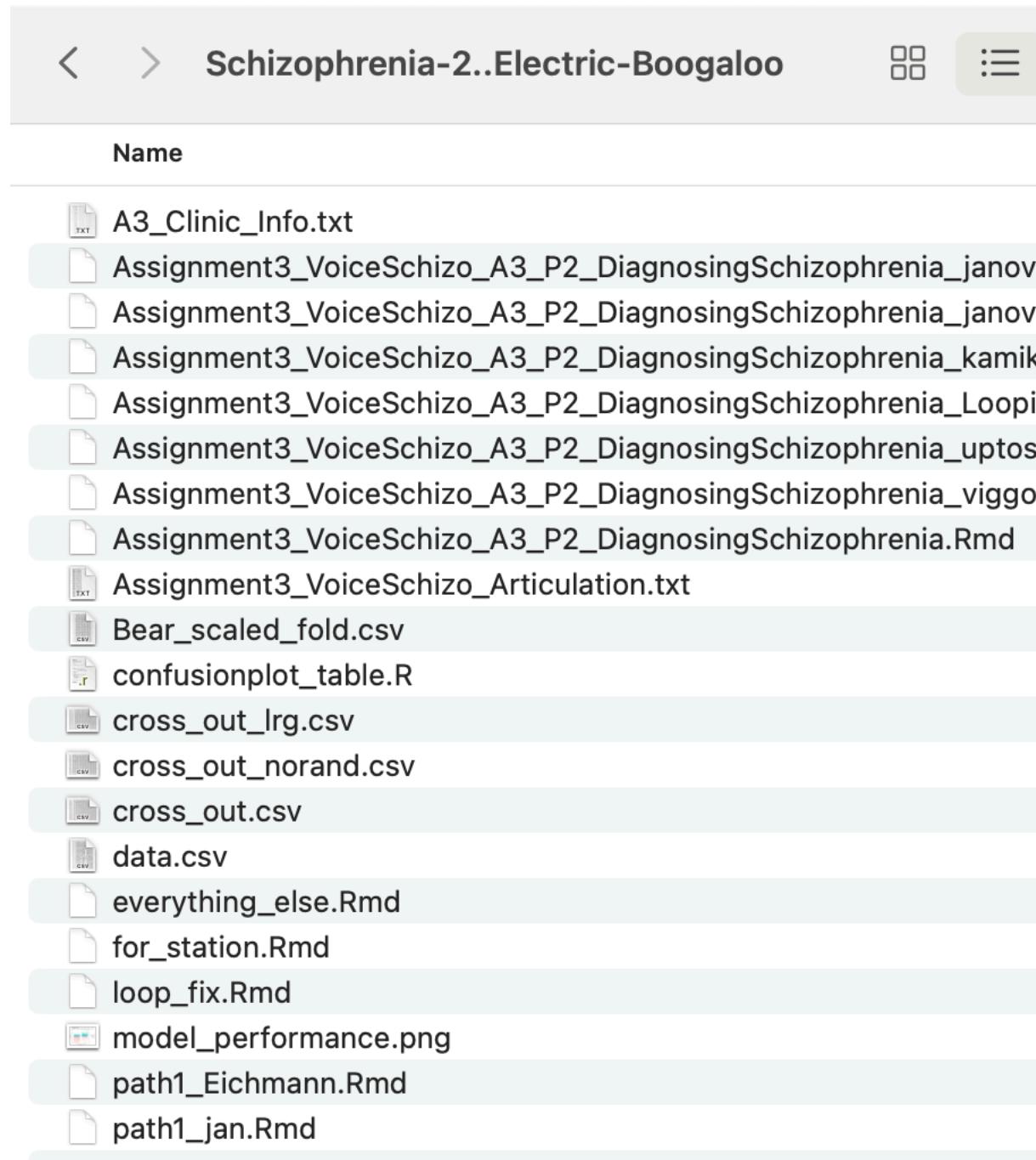


Name	Date Modified	Size	Kind
A3_Clinic_Info.txt	Today at 09.33	8 KB	Plain Text
Assignment3_VoiceSchizo_A3_P2_DiagnosingSchizophrenia_janoverboard_2.Rmd	Today at 09.33	10 KB	R Markdown File
Assignment3_VoiceSchizo_A3_P2_DiagnosingSchizophrenia_janoverboard.Rmd	Today at 09.33	9 KB	R Markdown File
Assignment3_VoiceSchizo_A3_P2_DiagnosingSchizophrenia_kamikazeCarl.Rmd	Today at 09.33	7 KB	R Markdown File
Assignment3_VoiceSchizo_A3_P2_DiagnosingSchizophrenia_Louie.Rmd	Today at 09.33	9 KB	R Markdown File
Assignment3_VoiceSchizo_A3_P2_DiagnosingSchizophrenia_uptospeed.Rmd	Today at 09.33	9 KB	R Markdown File
Assignment3_VoiceSchizo_A3_P2_DiagnosingSchizophrenia_viggomortensen.Rmd	Today at 09.33	6 KB	R Markdown File
Assignment3_VoiceSchizo_A3_P2_DiagnosingSchizophrenia.Rmd	Today at 09.33	5 KB	R Markdown File
Assignment3_VoiceSchizo_Articulation.txt	Today at 09.33	73 KB	Plain Text
Bear_scaled_fold.csv	Today at 09.33	243 KB	CSV Document
confusionplot_table.R	Today at 09.33	2 KB	R Source File
cross_out_lrg.csv	Today at 09.33	176 KB	CSV Document
cross_out_norand.csv	Today at 09.33	165 KB	CSV Document
cross_out.csv	Today at 09.33	171 KB	CSV Document
data.csv	Today at 09.33	194 KB	CSV Document
everything_else.Rmd	Today at 09.33	10 KB	R Markdown File
for_station.Rmd	Today at 09.33	8 KB	R Markdown File
loop_fix.Rmd	Today at 09.33	8 KB	R Markdown File
model_performance.png	Today at 09.33	12 KB	PNG image
path1_Eichmann.Rmd	Today at 09.33	10 KB	R Markdown File
path1_jan.Rmd	Today at 09.33	34 KB	R Markdown File



# Version Control

Have you ever had a hard time understanding how version control works?



# Is?

?



Sources  
& Notes

# Version Control: Git & Github

---

Data

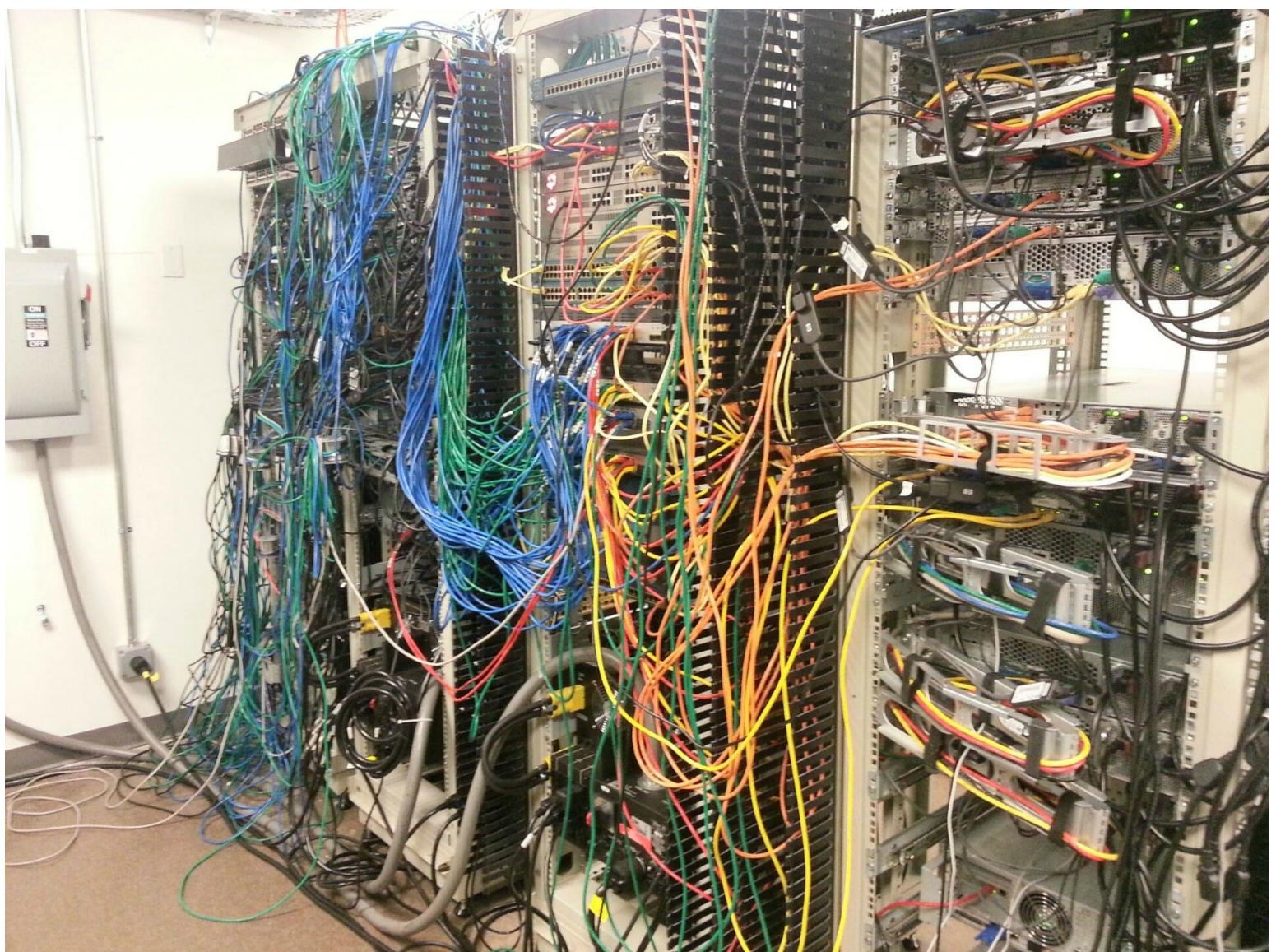
There's a starman waiting in the sky



Commits

There's a **spaceman** waiting in the sky

**Git**



**GitHub**



Sources  
& Notes

# Version Control: Alignment of Expectations

---

How much git you learn is up to you.

We only expect you to be able to access the course materials\*.

\*edit: access the course materials and submit a link to your repository as a part of the exam.



Sources  
& Notes



# **Let's get practical, practical**

---

Go to UCloud:

We will start a job, pull the exercise from GitHub and get coding.



Sources  
& Notes

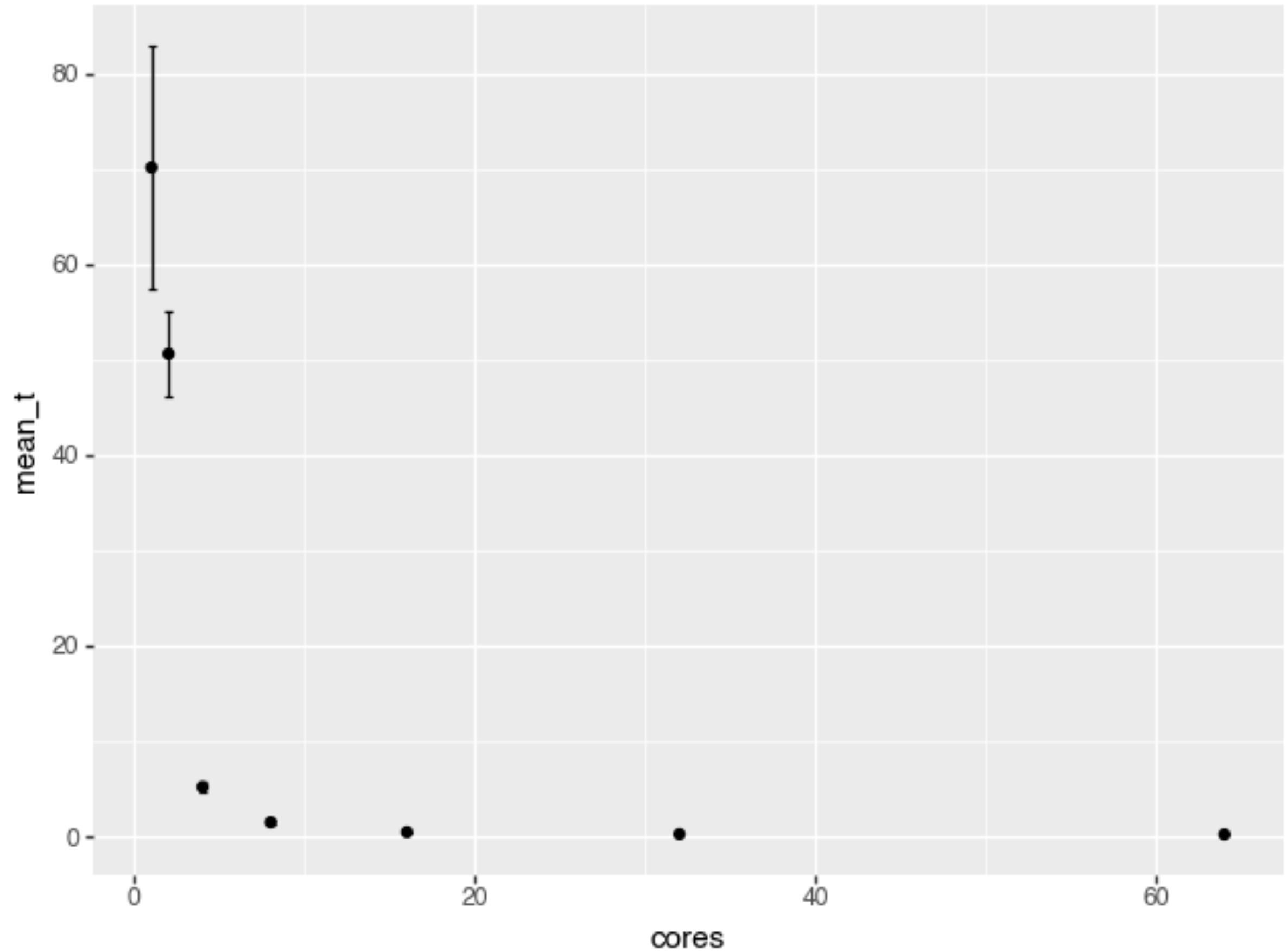


# Intuition about resources needed

---

NLP will need at least 4 cores.

Start small, then go up.

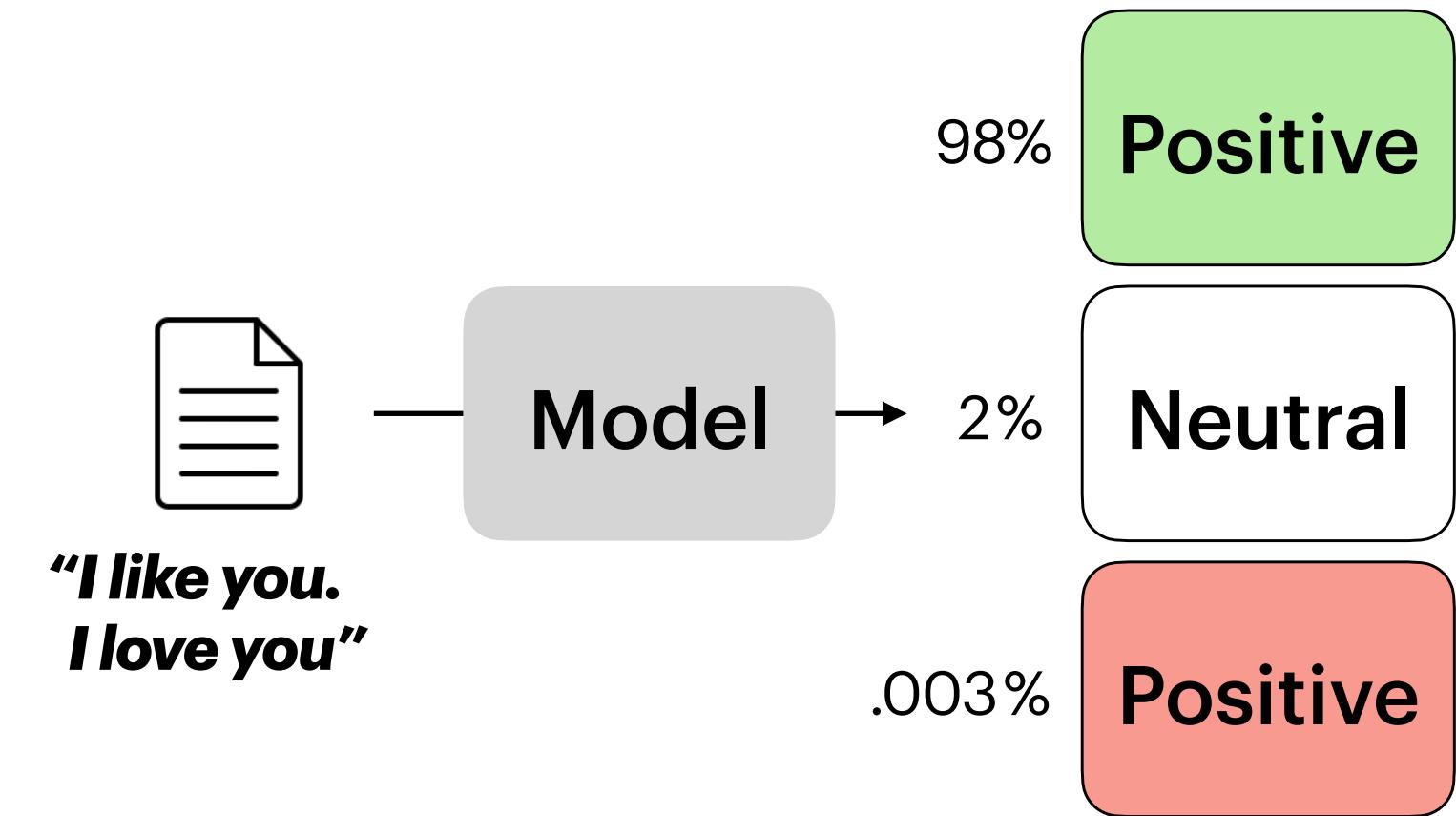


Sources  
& Notes

# Examples of Predictive Tasks

---

- **Sentiment Classification**
- Named Entity Recognition (NER)
- Text Summarization
- Machine Translation
- Question-answering



**Question: How could we do this?**



Sources  
& Notes

Demo on an open source model can be found here: <https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment>

# Where are we at now?

---

- **flexibly performs multiple tasks**
- and can generalize to new tasks (few- or zero-shot generalization\*)
- extremely **fluent** language generation
- does this model have any “deep” understanding of the linguistic task it is performing? does it have any command of “factual” information, any notion of truth?

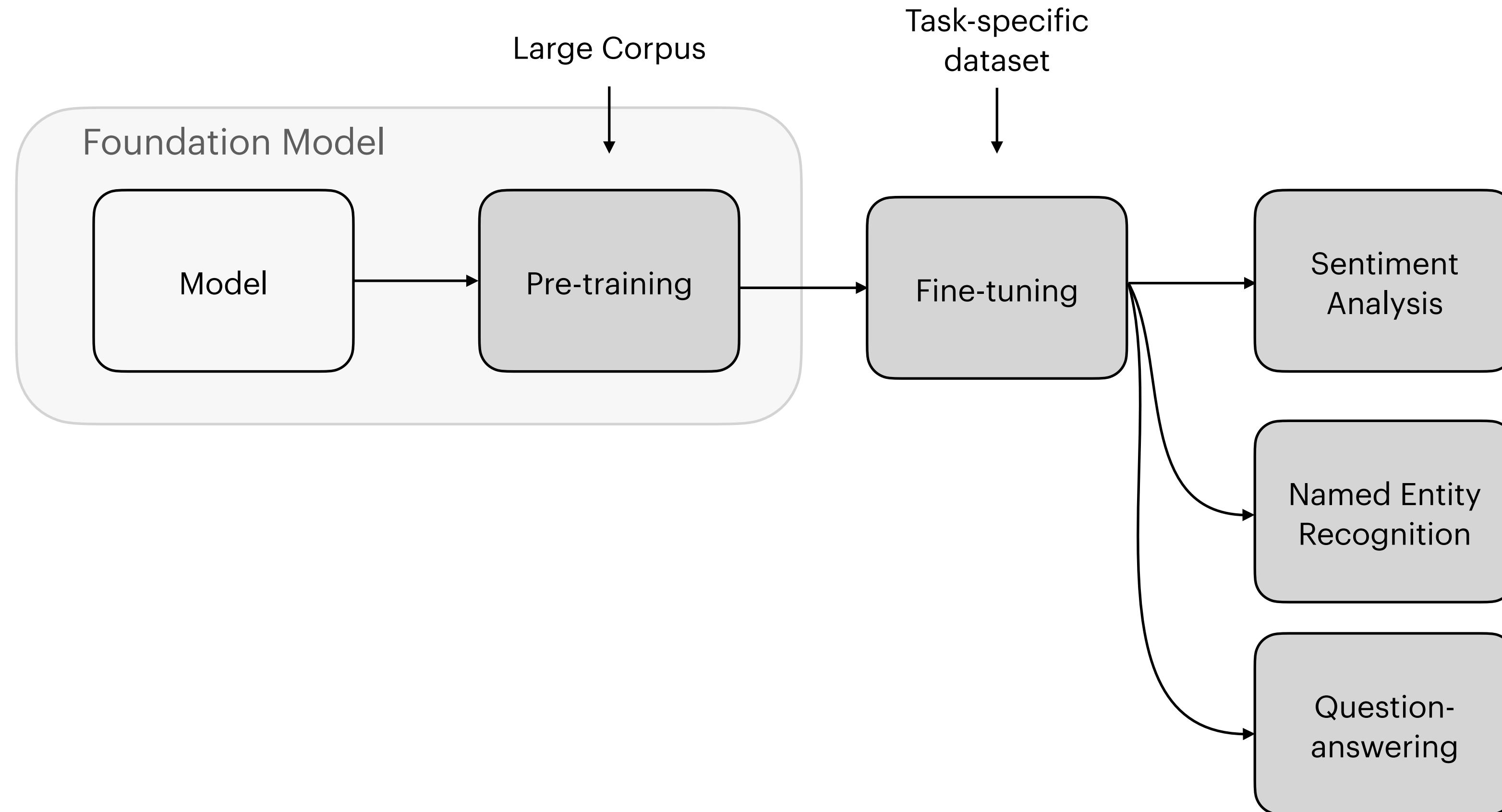


Sources  
& Notes

Whether it is completely unseen is actively discussed as it likely represented within the training data

# What lies behind?

---



# Familiar prompts?

---

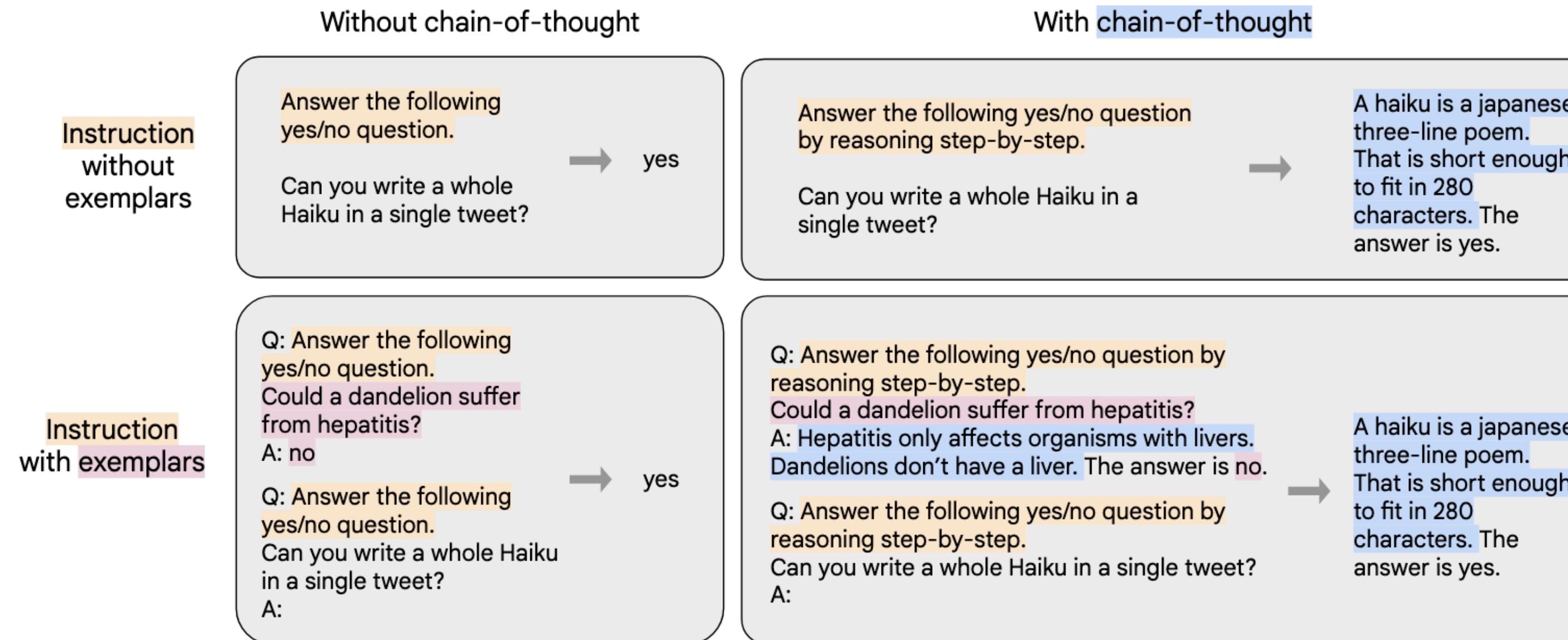
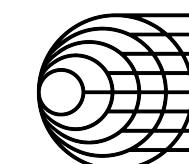


Figure 3: Combinations of finetuning data formats in this work. We finetune with and without exemplars, and also with and without chain-of-thought. In addition, we have some data formats without instructions but with few-shot exemplars only, like in Min et al. (2022) (not shown in the figure). Note that only nine chain-of-thought (CoT) datasets use the CoT formats.



Sources  
& Notes

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., ... & Wei, J. (2024). Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70), 1-53. <https://doi.org/10.48550/arXiv.2210.11416>



CENTER FOR  
HUMANITIES  
COMPUTING