



ENVIRONMENTAL DATA ANALYTICS: WEEK 2

Spring 2023

Nicholas School of the Environment - Duke University

Introductions: John Fay (he, him, his)

1991 BA Biology & Environmental Studies

Bowdoin

1996 MSc Conservation Biology &
Ecosystems Management



1997 GIS Manager at Jasper Ridge Biological Station
GIS Manager at Stanford's Ctr. for Conservation Biology



2005 Research associate/Instructor at NSOE

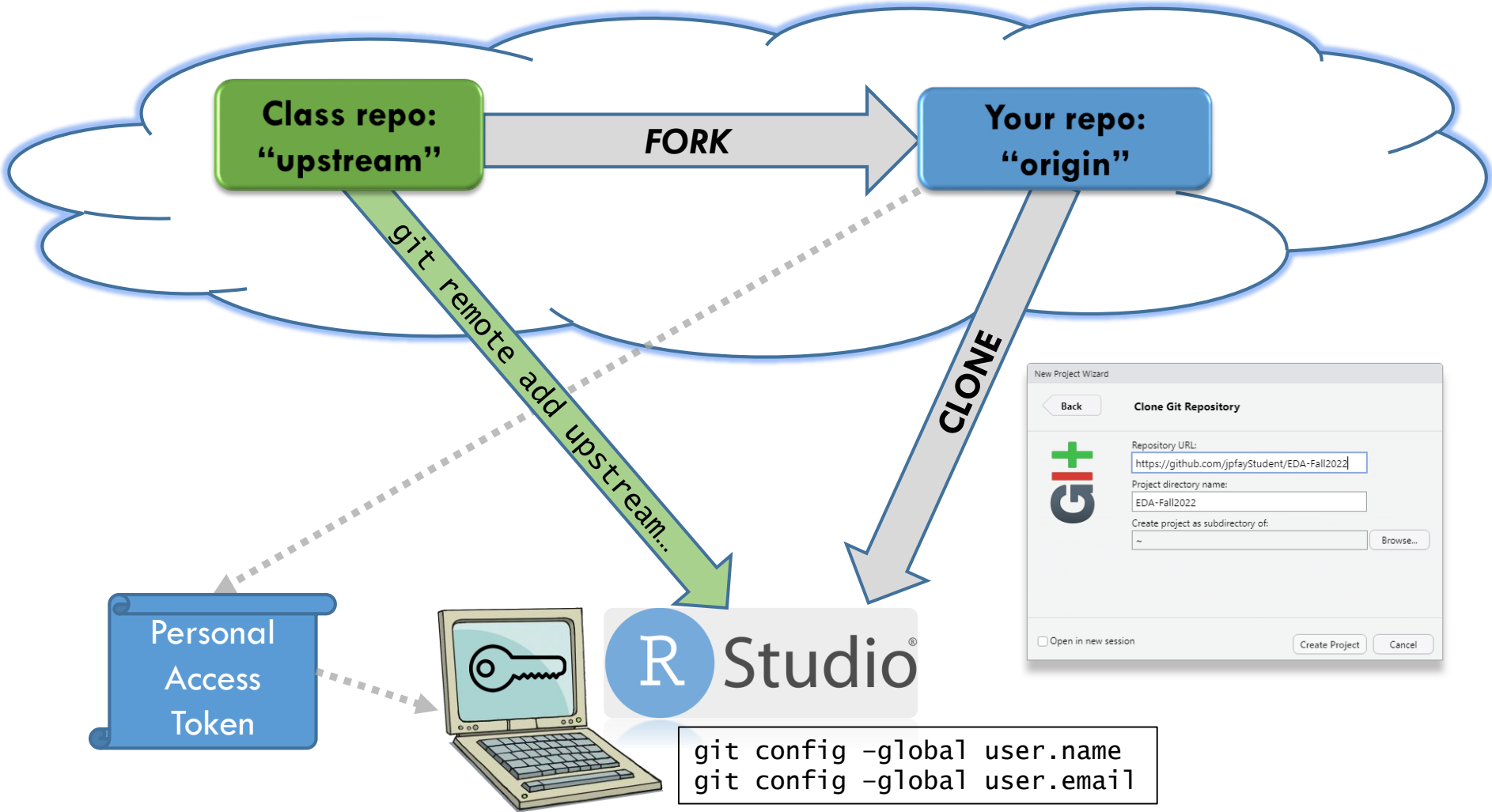


Tech Roundup

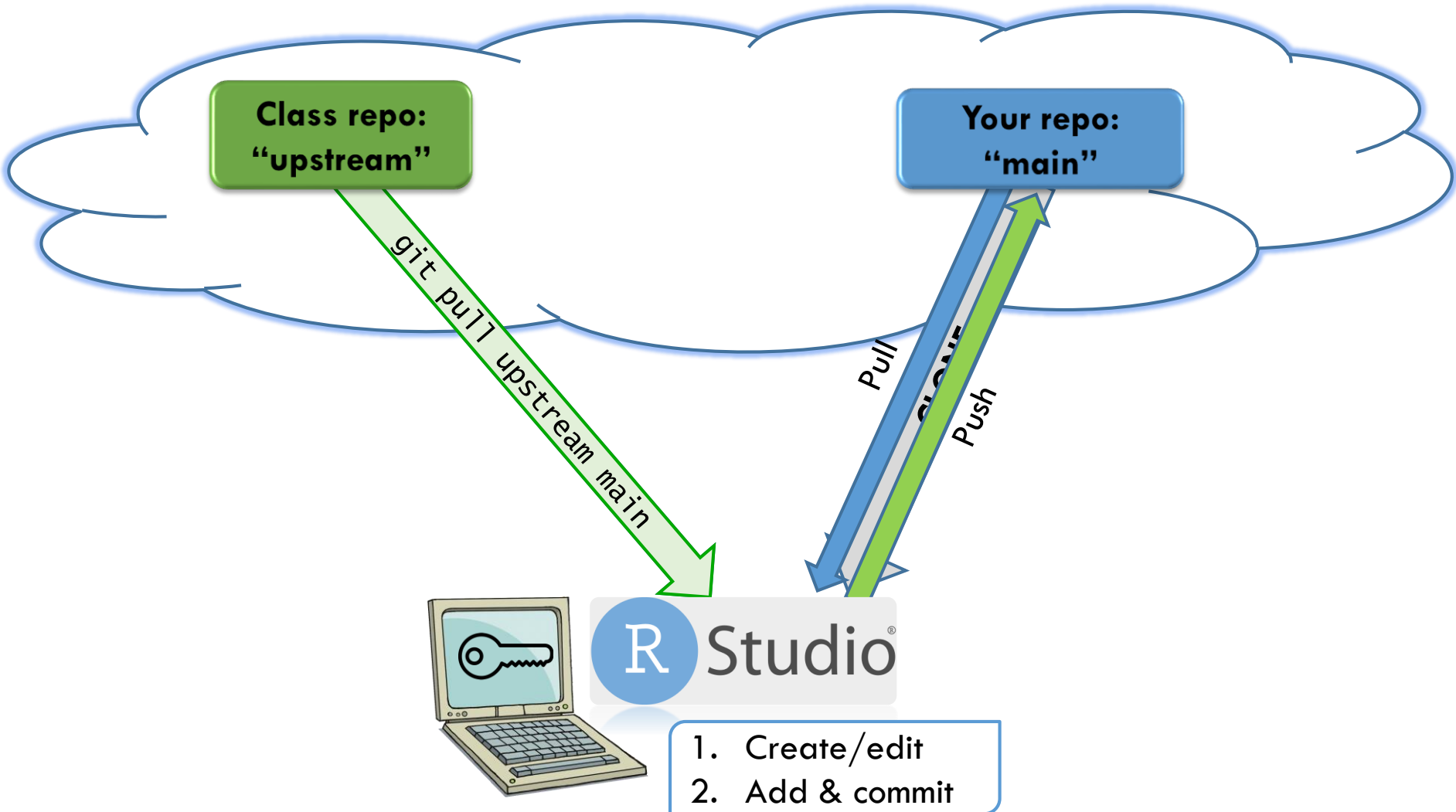


- Issues joining Slack? Navigating Slack?
- Issues installing applications:
R... | RStudio... | Git...
- Issues creating your R project?
Forking... | Cloning... | PAT... | Committing... | Pushing...
- Accessing/navigating web site?

Explaining Git/GitHub: *Setup*



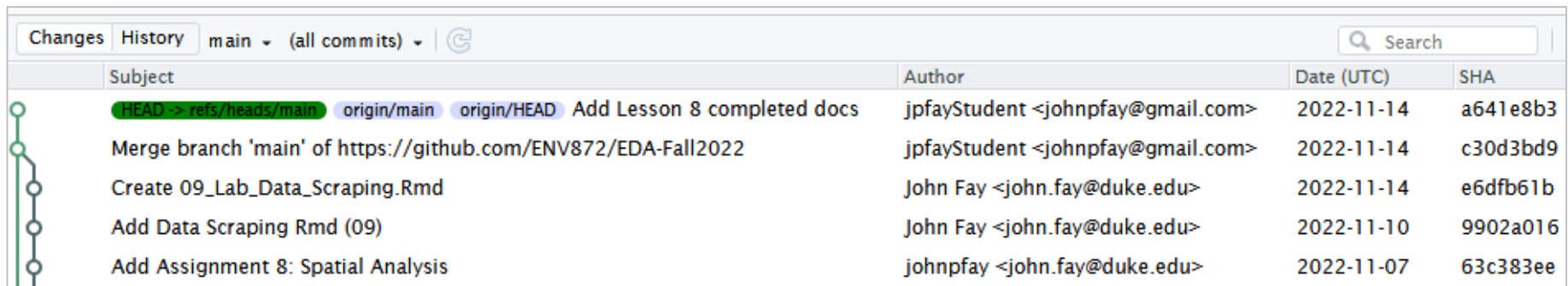
Explaining Git/GitHub: *Use*



Explaining Git/GitHub: Workflow

<https://git-school.github.io/visualizing-git>

- Single repository...
- Remotes repositories...
- Upstream changes...
- Explaining merges & divergent branches



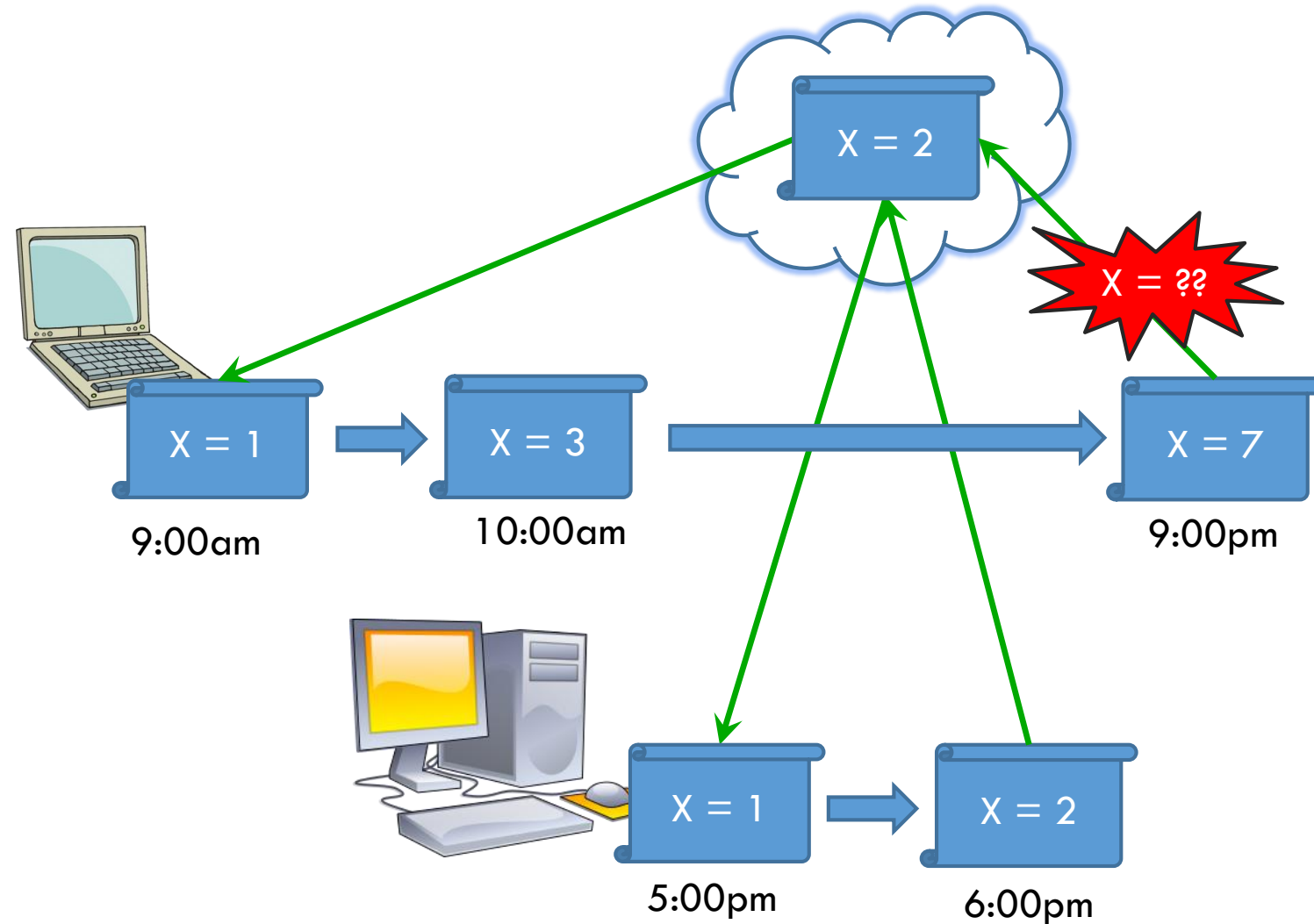
The screenshot shows a Git commit history table with the following columns: Subject, Author, Date (UTC), and SHA. The table lists five commits, with the top one being a merge of the 'main' branch from a remote repository. The commit subjects are: 'HEAD -> refs/heads/main Merge branch 'main' of https://github.com/ENV872/EDA-Fall2022', 'Create 09_Lab_Data_Scraping.Rmd', 'Add Data Scraping Rmd (09)', and 'Add Assignment 8: Spatial Analysis'. The authors are listed as 'jpfayStudent <johnpfay@gmail.com>' and 'John Fay <john.fay@duke.edu>'. The dates range from 2022-11-07 to 2022-11-14. The SHA values are a641e8b3, c30d3bd9, e6dfb61b, 9902a016, and 63c383ee.

Subject	Author	Date (UTC)	SHA
HEAD -> refs/heads/main Merge branch 'main' of https://github.com/ENV872/EDA-Fall2022	jpfayStudent <johnpfay@gmail.com>	2022-11-14	a641e8b3
Create 09_Lab_Data_Scraping.Rmd	John Fay <john.fay@duke.edu>	2022-11-14	e6dfb61b
Add Data Scraping Rmd (09)	John Fay <john.fay@duke.edu>	2022-11-10	9902a016
Add Assignment 8: Spatial Analysis	johnpfay <john.fay@duke.edu>	2022-11-07	63c383ee

Explaining Git/GitHub: *Terms*

- Repository
- Git vs GitHub
- Fork vs Clone
- Staging vs Committing
- SHA vs Commit message
- Push vs Pull
- `git pull` vs `git pull upstream main`
- Merge

Git: Merge conflicts

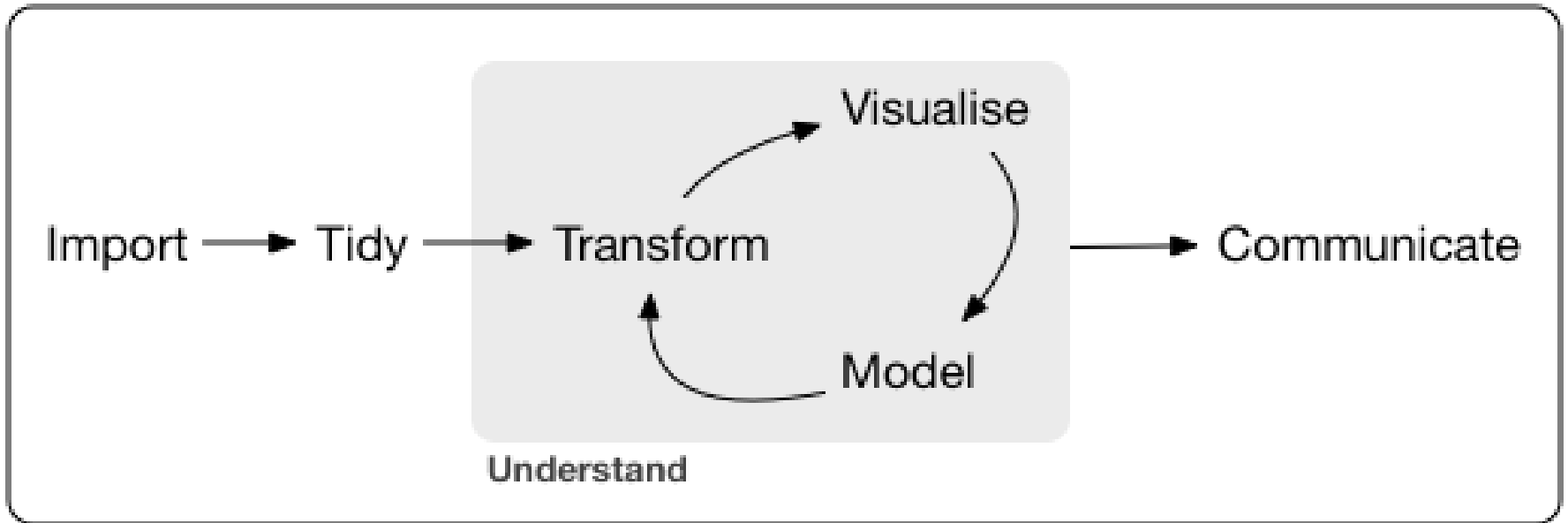


Exercise...

If Git goes totally sideways...

- Make a back up (e.g. zip) of you project folder
- Close RStudio project and rename project folder
- Create a new project linked to your forked repo
- Link to upstream remote, as before
- Copy over any missing items from your renamed folder to your newly cloned repository

Q&A: What is Data Analytics



<https://vita.had.co.nz/papers/tidy-data.pdf>

Data Types

- Understand what it means to “tidy” data
- Differentiate “primary” and “secondary” data
- Differentiate “qualitative” and “quantitative” data
- Identify different file types used in data analytics and discuss why some formats are better than others in terms of transparency and reproducibility
- Describe the various data structures used in data analytics and what each are used for: *Vectors, matrices, arrays, data frames, lists*
- Understand the difference between *R* and *RStudio*
- Become familiar with the typical layout of an *RStudio session*

Up next: Module 2



Reproducibility & Coding Basics

- What is “reproducibility”? Why is it important?
- Working with RMarkdown files...
- Writing R code...

The class “rhythm”

Each week = 1 module = {recordings + exercise + assignment}

□ Recordings:

- Watch recordings prior to class

□ In class:

- Ask questions about recordings and about assignment
- Group exercises to re-inforce concepts (+ some new concepts)

□ Assignments:

- Made available after class discussion
 - A01 made available after class session last week; A02 available now
- Due on Friday following class discussion
 - A01 is due this Friday, A02 is due next Friday
- *No late penalty if submitted by following Monday*
However, instructors/TA's not likely to answer questions on weekend!