Data Acquisition & Organization: Introduction

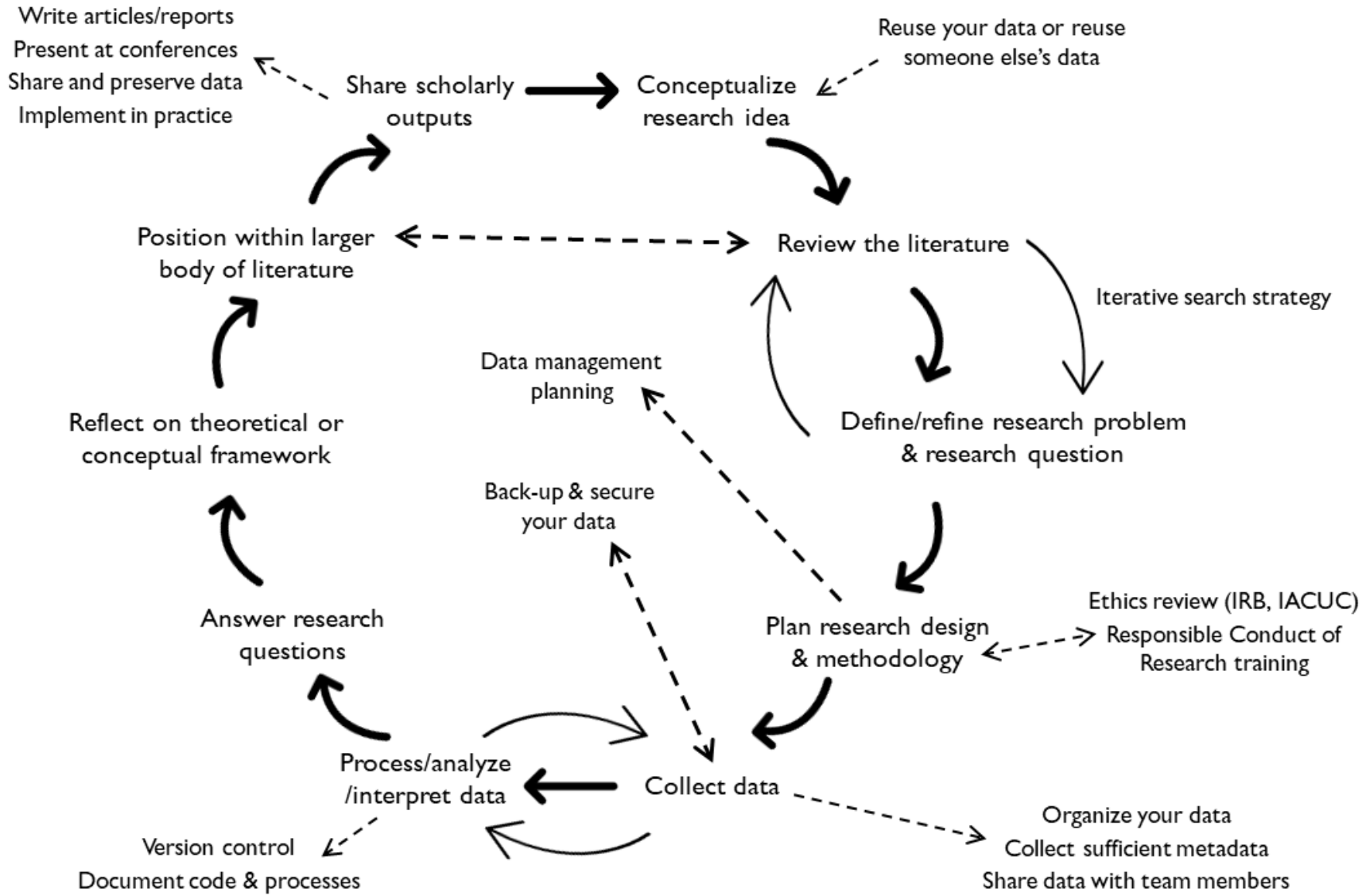
DATA 3101

Elizabeth Stregger

September 7, 2023

Learning Objectives

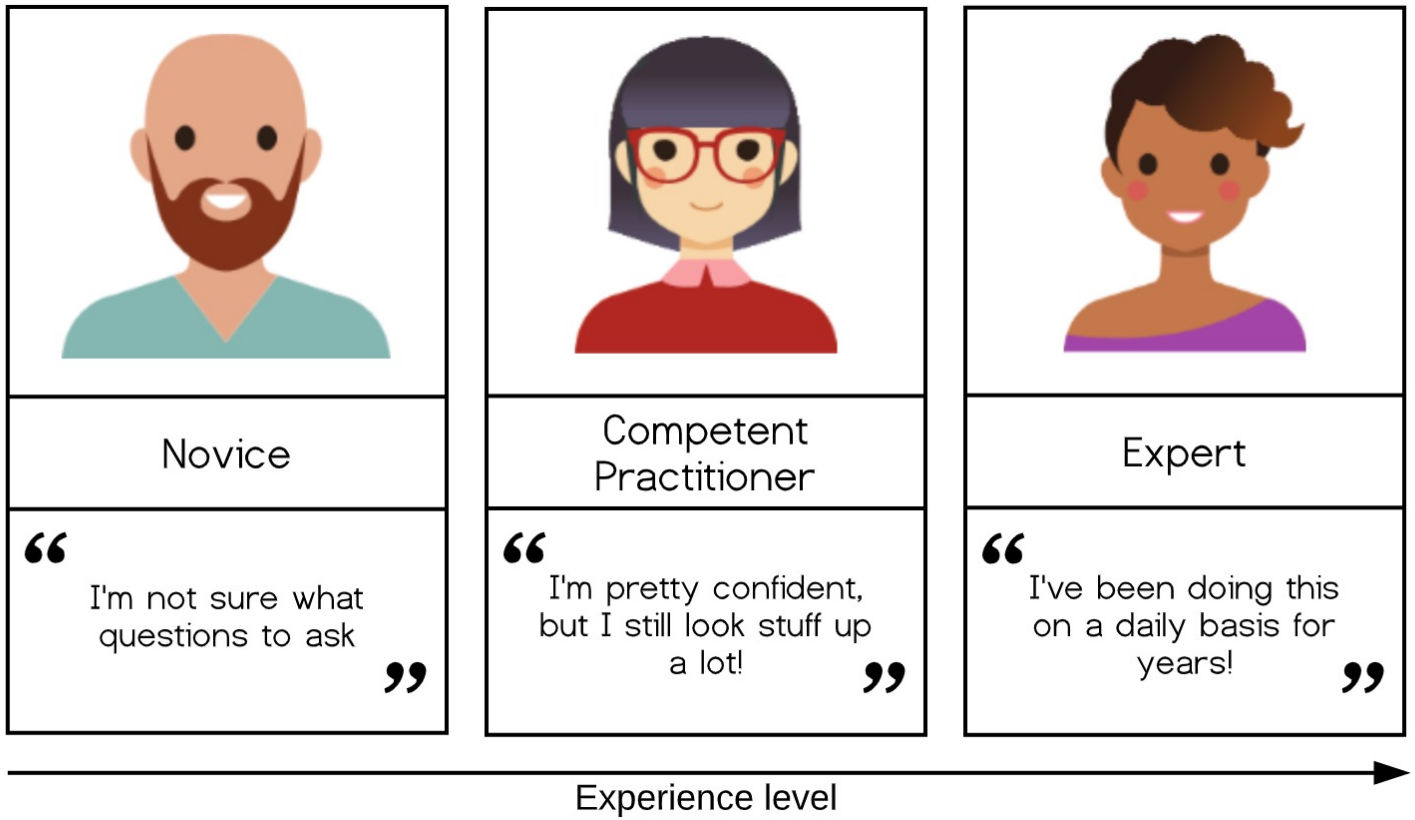
- Demonstrate in depth understanding of the principles, motivations and goals for reproducible, ethical, and open data
- Use Git for communication and reproducible version control
- Import and tidy diverse data sources in R
- Explore data to identify potential research questions or problems in the dataset
- Identify best practices for research data management, including data organization, storage, security, sharing, and ethical reuse
- Demonstrate what you have learned about data acquisition, data organization, and data tools through self-reflection



Teaching and learning: inspired by the Carpentries

- Approximately half of our class time will be participatory live coding and problem solving
- I will ask questions to track your progress and adjust my teaching
- At the end of each week, you can give me feedback on teaching:
  - What is one thing you'd like us to spend more time on or to review?
  - What is one thing you found useful or are excited about?

Practice and Learning



From The Carpentries, Building Skill with Practice:
https://carpentries.github.io/instructor-training/02-practice-learning.html

Teaching and learning: Ungrading

- Focus on learning and interaction
- Submit – get feedback – improve & resubmit
- Tracking engagement, contributions, and progress towards goals
- Self-evaluation

What do you want to learn?

- What topics, concepts, or skills do you want to get from this course?
- For the research data management sections, you'll be able to choose an openly available research dataset that is of interest to you (think about your program, research methods, or personal interests
- Please complete the Student Information Survey in Moodle

Tools: Git & RStudio

- One of the themes throughout this course is the importance of reproducibility and open data
- Git and RStudio are a reproducible research environment
- RStudio: import, analyze, communicate data using a reproducible workflow
- RStudio can be linked with Git for version control
- Many of the good practices in research data management can be demonstrated using these tools (importance of file naming, file structures, documentation, version control)

GitHub as Course Platform

- Doing your work is integrated with organizing it
- Collaboration is highly structured
- We can exchange working code
- Practicing openness (but with guardrails)
- We’ll use GitHub issues as a to-do list