Module 2: Collaboration and Workflow

Rebecca C. Steorts

Agenda

- Organization
- Reproducibility
- Version Control

Organization

One asset of being an efficient data scientist is being organized.

Organization

Why is organization an important skill?

Organization

Messy files and poorly documented code mean that:

- you will need to often dig through your code
- you will have trouble communicating with your collaborative team
- you will have a difficult workflow for a industry/academic project
- you will likely have trouble reproducing your results

These can all be disasterous if you're working on product development for a large company and they are depending on your skills, expertise, and analysis!

Self-Assessment

Think about a recent programming assignment from the past year.

- Do you think you can reproduce the results exactly?
- Do you think that anyone in the class could easily understand your code (i.e., is it well documented)?

Give a **short assessment** regarding one strength of your coding and one weakness that you plan to work on in this course.

Collaborative Programming

How can we work collaboratively on a project involving code?

1. Create a strong team with a team leader to keep the group **organized** and **on track**
2. Making sure that our code is **reproducible** and using software such as R markdown as one such option
3. Using **version control** (github) for our collaborative work in order to **effectively communicate** and **work together**

Reproducibility

What is reproducibility and why is this important?

The Reproducibility Crisis

"The terms "reproducibility crisis" and "replication crisis" gained currency in conversation and in print over the last decade (Pashler & Wagenmakers 2012), as disappointing results emerged from large scale reproducibility projects in various medical, life and behavioural sciences (Open Science Collaboration, OSC 2015)."

The Reproducibility Crisis

What does this mean?

- Researchers are findings that we they try and replicate papers, they are not finding the same results as in the published work.
- This has led to the above crisis, and thus, pushing us as data scientists to publish work that is fully transparent and can be checked when we pass it to someone else.

Reproducibility

- Work is considered reproducible if the entire process can be replicated from start to finish.
- This includes the data, code, figures — everything.

Reproducibility

- Why is reproducibility important for our community?
- Why is reproducibility hard to achieve?

Reproducibility

We just talked about **why** it's important for our work to be both organized and reproducible.

Let's talk about one tool we can use to make our work reproducible.

Rmarkdown

- R is a statistical programming language
- RStudio is a convenient interface for R (an integrated development environment, IDE)

Put simply:

- R is like a car's engine
- RStudio is like a car's dashboard

R packages

To make our work the most reproducible as possible, best practice is writing our finished product into an R package available on CRAN and making this accessible to the community.

R packages

1. CRAN packages have strict criteria for acceptance so this force you to write your code in best practice
2. You can tie in your R packages to your paper/project to make it reproducible from start to finish

This is the most idealized way of writing software and working with a team.

To learn more about R packages, you can check out https://r-pkgs.org/.

Rmarkdown

[DEMO]

Let's perform a demo regarding how we can make a simple .Rmd file reproducible in markdown.

Question: What are the benefits over just writing an R script?

Version Control

What is version control?

Version Control

Version control is the practice of tracking and managing changes to software code.

Version control systems (github) are software tools that help software teams manage changes to source code over time.

Version Control

- GitHub is a platform for collaboration
- It's really designed for version control

Version Control

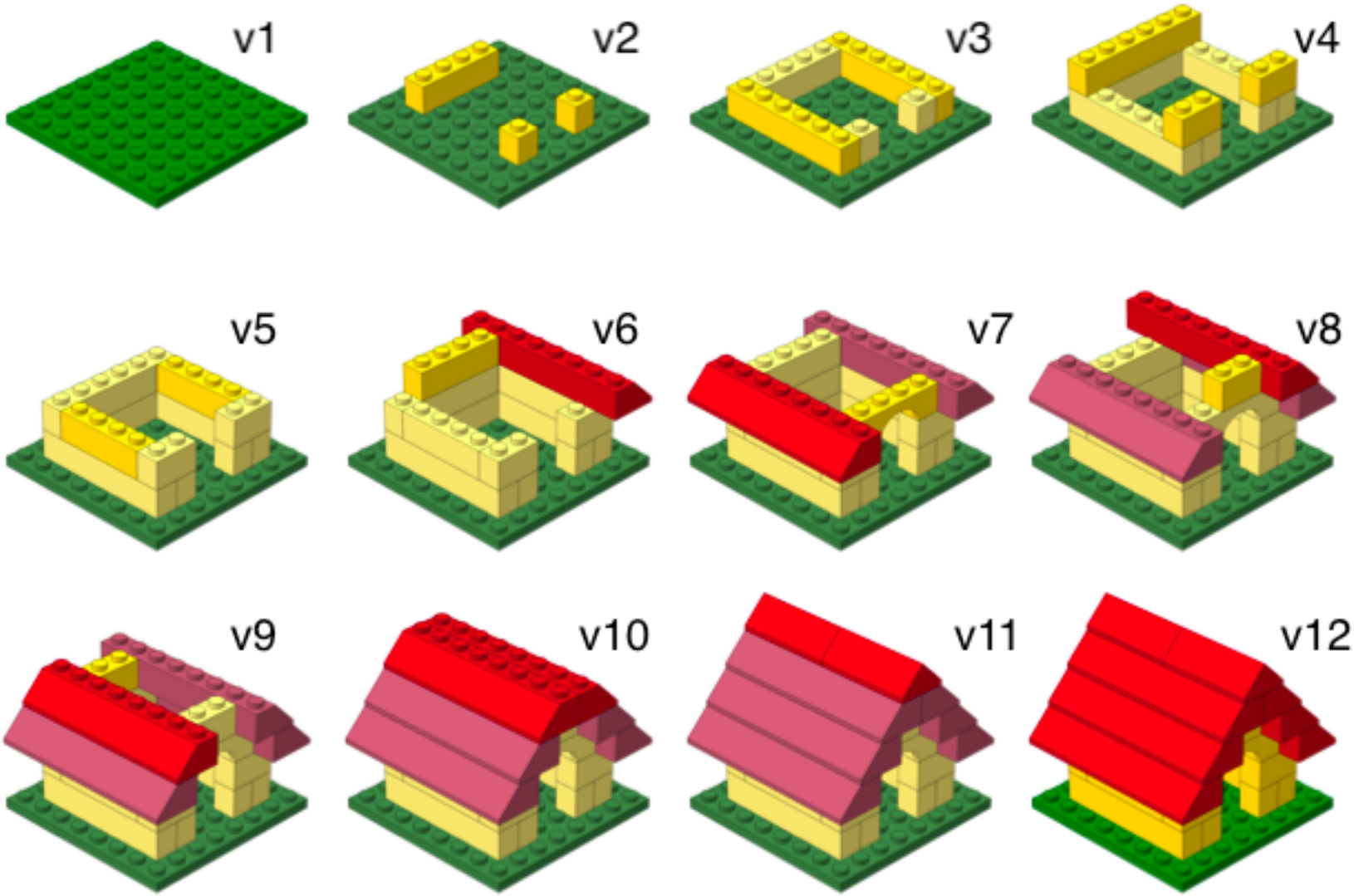


Figure 1: Lego steps

Version Control with commit messages

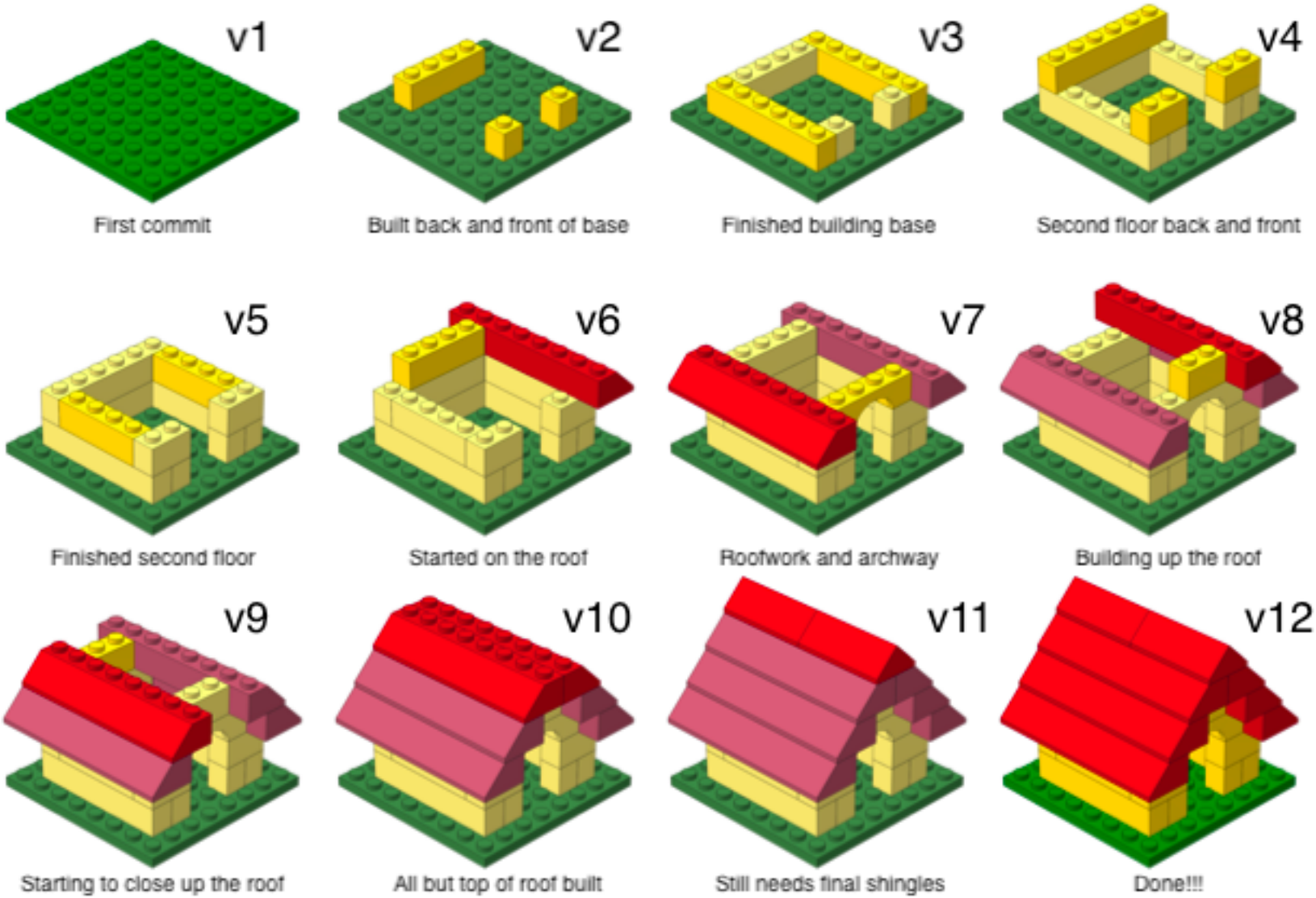
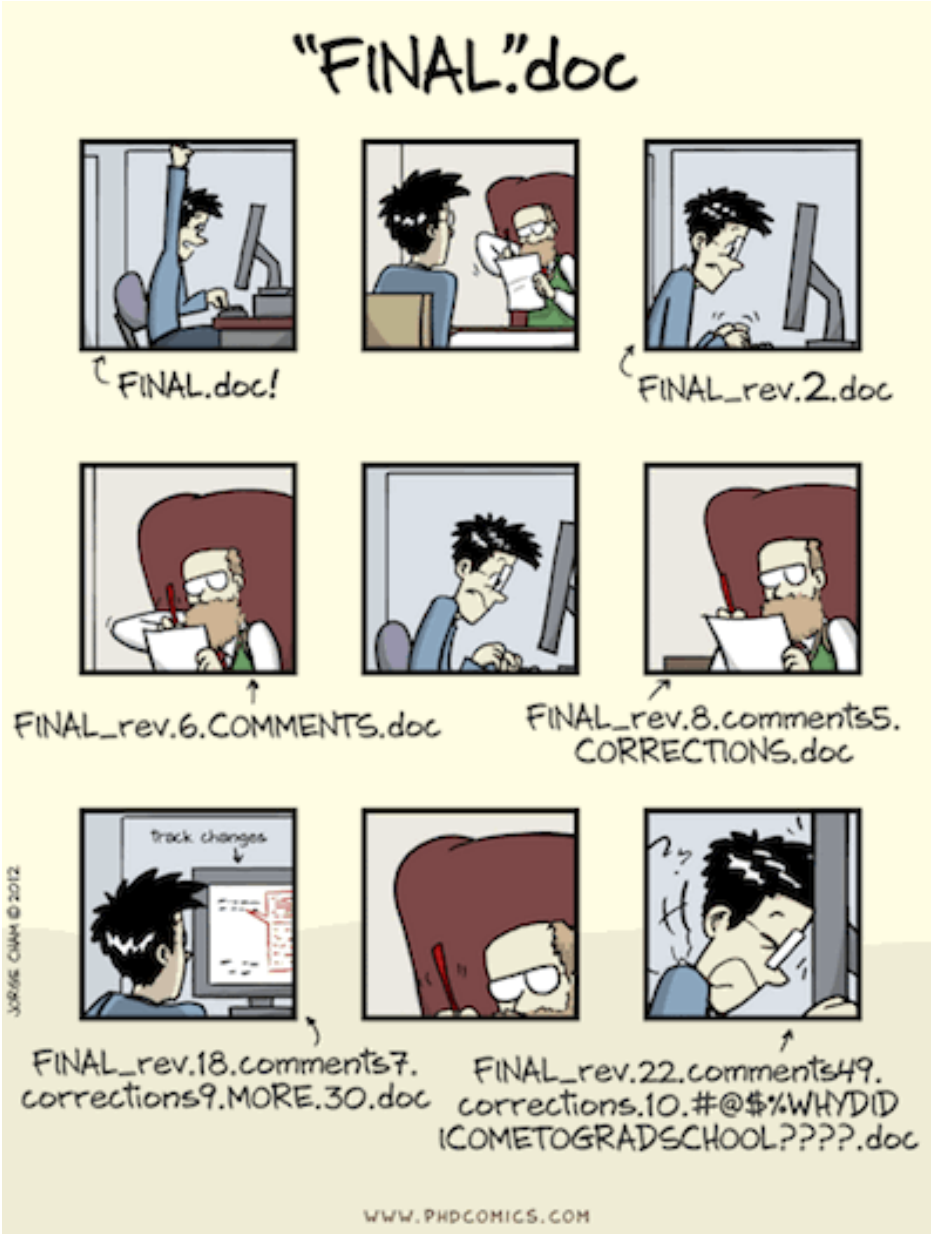


Figure 2: Lego steps with commit messages

Why do we need version control?



Github and Tips

- Git is a version control system – like "Track Changes" features from Microsoft Word.
- GitHub is the home for your Git-based projects on the internet (like DropBox but much better).
- There are a lot of Git commands and very few people know them all. 99% of the time you will use git to add, commit, push, and pull.

Git and Github Tour

[DEMO]

- Connect an R project to GitHub repository
- Working with a local and remote repository
- Making a change locally, committing, and pushing
- Making a change on GitHub and pulling

Recap

Can you answer these questions?