Tidy working with R

Mike K Smith
Canterbury R Users-group
June 2018

Does this sound familiar?

- You get data
- You push and pull data into shape for analysis or visualisation
- You create analysis (formulas), graphs, tables
- You write a report

Does this sound familiar - Scenario 1

- Late breaking changes to the data
- PANIC!
- Recreate graphs, tables, report
- HAVE YOU CHANGED EVERYTHING?
- **PANIC PANIC!**

Does this sound familiar - Scenario 2

- You sent your report
- Your collaborator / supervisor / head of department liked it ***BUT***…
- “Can you do the same but for this other outcome?”
- “I’ve just got some new data, can you update it?”
- “Nice graphs and tables, but we really should combine these two groups in any analysis.”

Does this sound familiar - Scenario 2 continued

- ***so***…
- You get the next set of data (like dataset 1)
- You push and pull data into shape for analysis (AGAIN!)
- You recreate formulae, graphs, tables (AGAIN!)
- You write (ANOTHER!) report that is much like the last one

Housekeeping!

- A lot of energy goes into basic housekeeping of analysis
  - Data preparation: checks, reshaping, etc.
  - Analysis: assumption checks, linking results to data
  - Graphs, tables: keeping up to data with data, results.
  - Version control: How is today's data different from the last set?
  - Logging: What did I do last time?
- A tidy project helps keep YOU organised, focused
  - It also helps others see what you're doing.

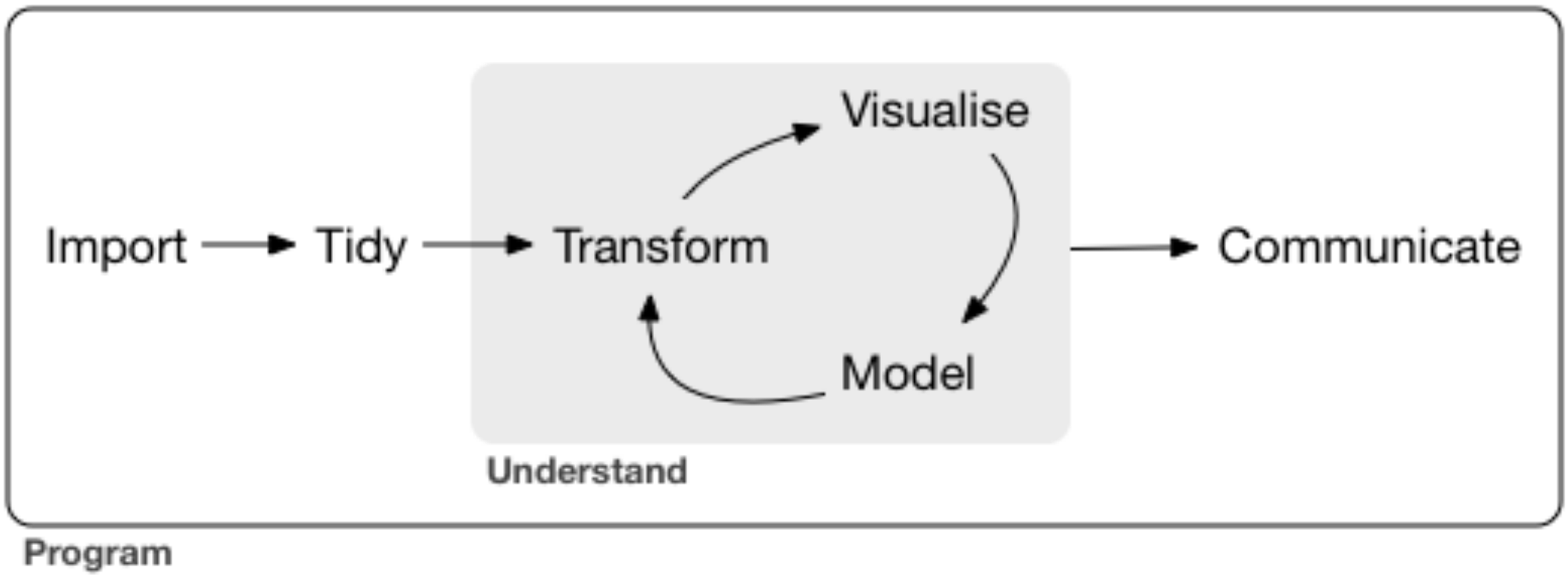
Reproducibility is a thing

- Reproducibility = being able to get from your data to the report content again and again and again.
- Better still, OTHER people being able to recreate what YOU did.
- Why is this a good thing?
    - Credibility in your results (I didn't HACK this, honest!)
    - When new data comes in you can "push a button" and get the new report.

Workflow

- A flow of work?
- Yup.
    - From data to report.
- Scripts help
    - You can see what changes you made
    - Others can see too
    - If you break work into chunks then you can build new analysis like Lego.

Workflow



https://github.com/hadley/r4ds/blob/master/diagrams/data-science.png

How does R help?

- Not so much R, as all the tools that go with R.
- You CAN do this with other tools (python, Excel even)
    - Script-based tools like python & R are more suited to building workflow
    - It's POSSIBLE, just harder in Excel
- R has a mountain of tools to help...

Data Workflow
- The Tidyverse

- Tidyverse
  - Get data (readr, readxl, haven, rvest, jsonlite, xml2)
  - Check data (tibble, dplyr, stringr, lubridate, forcats)
  - Reshape data (tidyr, dplyr)
  - Analyse data (purrr, ...)
  - Graph data (ggplot2)
  - Report data (rmarkdown, RStudio Notebooks)

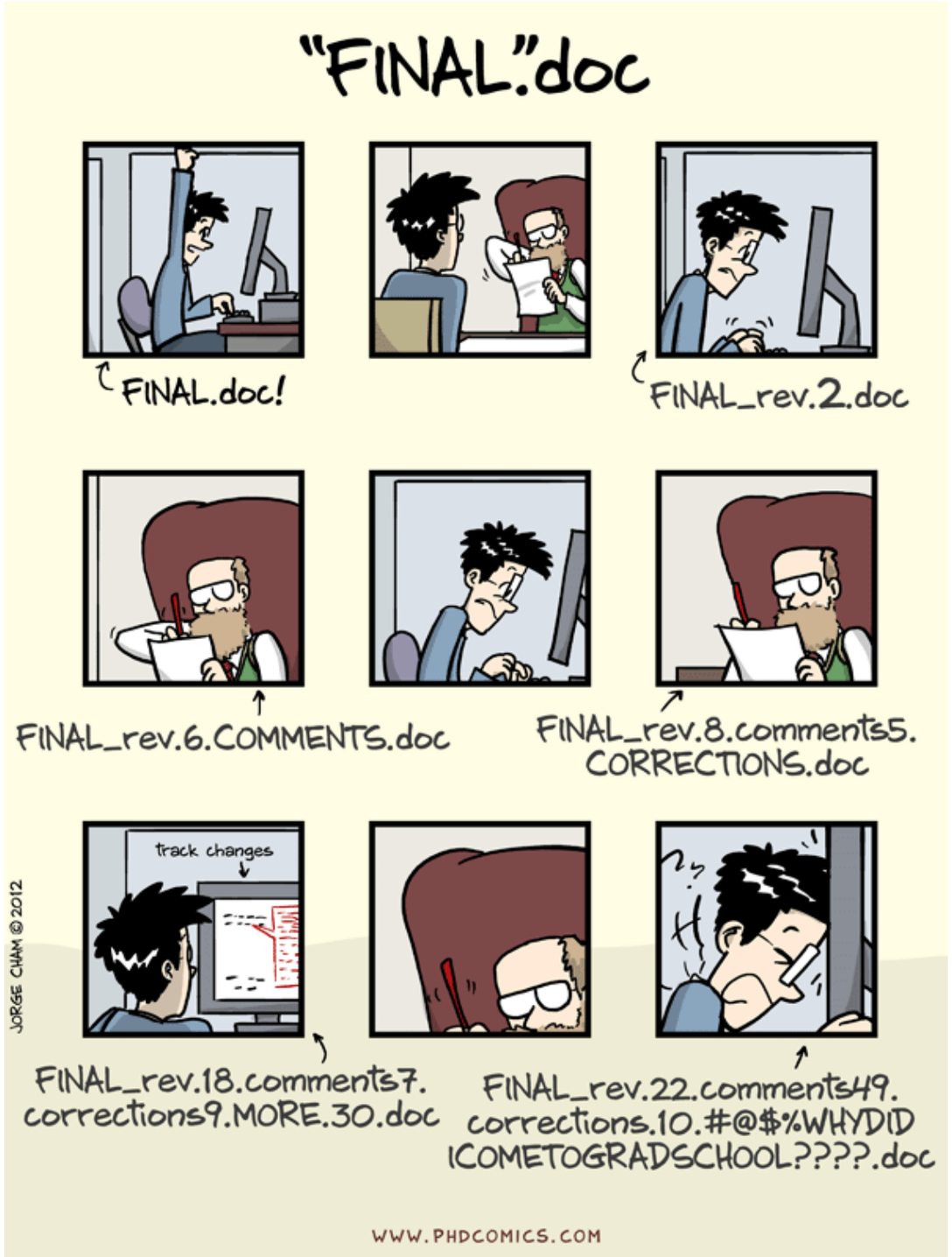
Project Housekeeping

- ProjectTemplate
  - Basic project structure. TIDY!
- RStudio Projects
  - Does a fair amount of workflow and housekeeping for you.
  - MANY tools within RStudio to help…

R Notebooks?

- In RStudio, you can create a new file "R Notebook"
  - Text description of WHAT you're doing (and WHY!)
  - Inline R code
    - Can show code, or hide and only show results
  - Output is stored WITH the text and code
  - Allows anyone else to see (and review) WHAT, WHY, HOW
  - REPRODUCIBILITY baked in!

Version control



http://phdcomics.com/comics.php?f=1531

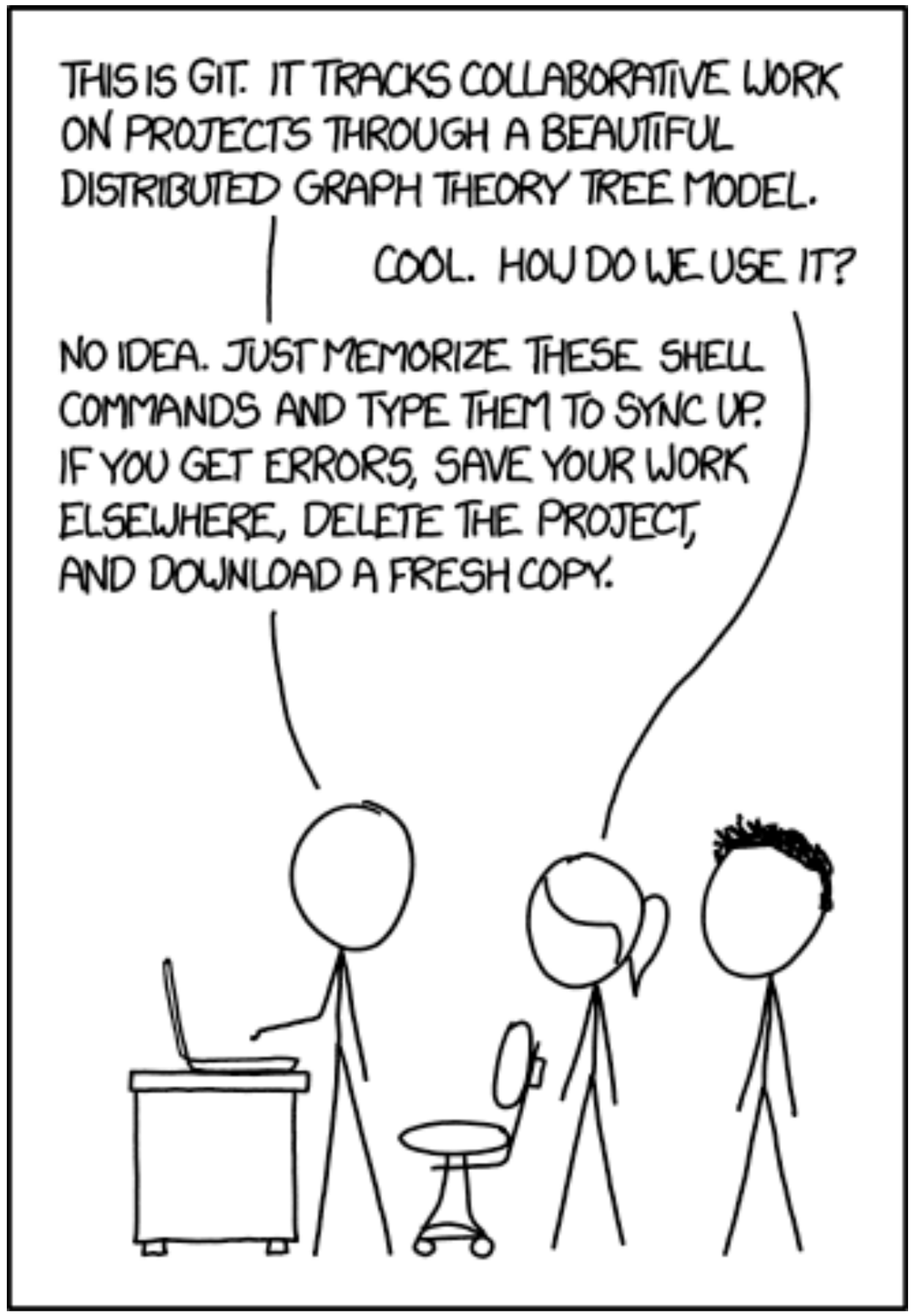
Version control

- code v0.1 + data v1.0 + analysis v0.5 = never shared
- code v1.0 + data v1.0 + analysis v1.0 = report v1.0
- code v1.0 + data v2.0 + analysis v1.1 = report v1.999
- code v1.9 + data v2.1 + analysis v1.7 = report v2.1.3.57
- code v1.9 + **data v1.0** + analysis v1.7 = **AAARGH!!!**
- WHAT HAS CHANGED!?!?!?!!?!

Version control

- More housekeeping!
- Tools like GitHub help keep things neat and tidy
    - Branch to create tomorrow's analysis
    - Commit messages to keep track of WHY things changed
        - Diffs take care of WHAT changed
    - Helps review
        - Let's discuss these changes
    - Collaborate!
- Git (and SVN) is baked into RStudio

Caveat



https://xkcd.com/1597/

Good resources

- RStudio
  - Webinars
- Jenny Bryan
  - Workflow - you should have one
- Hadley Wickham
  - R for Data Science (R4DS)
- Mara Averick
  - R Workflow fun