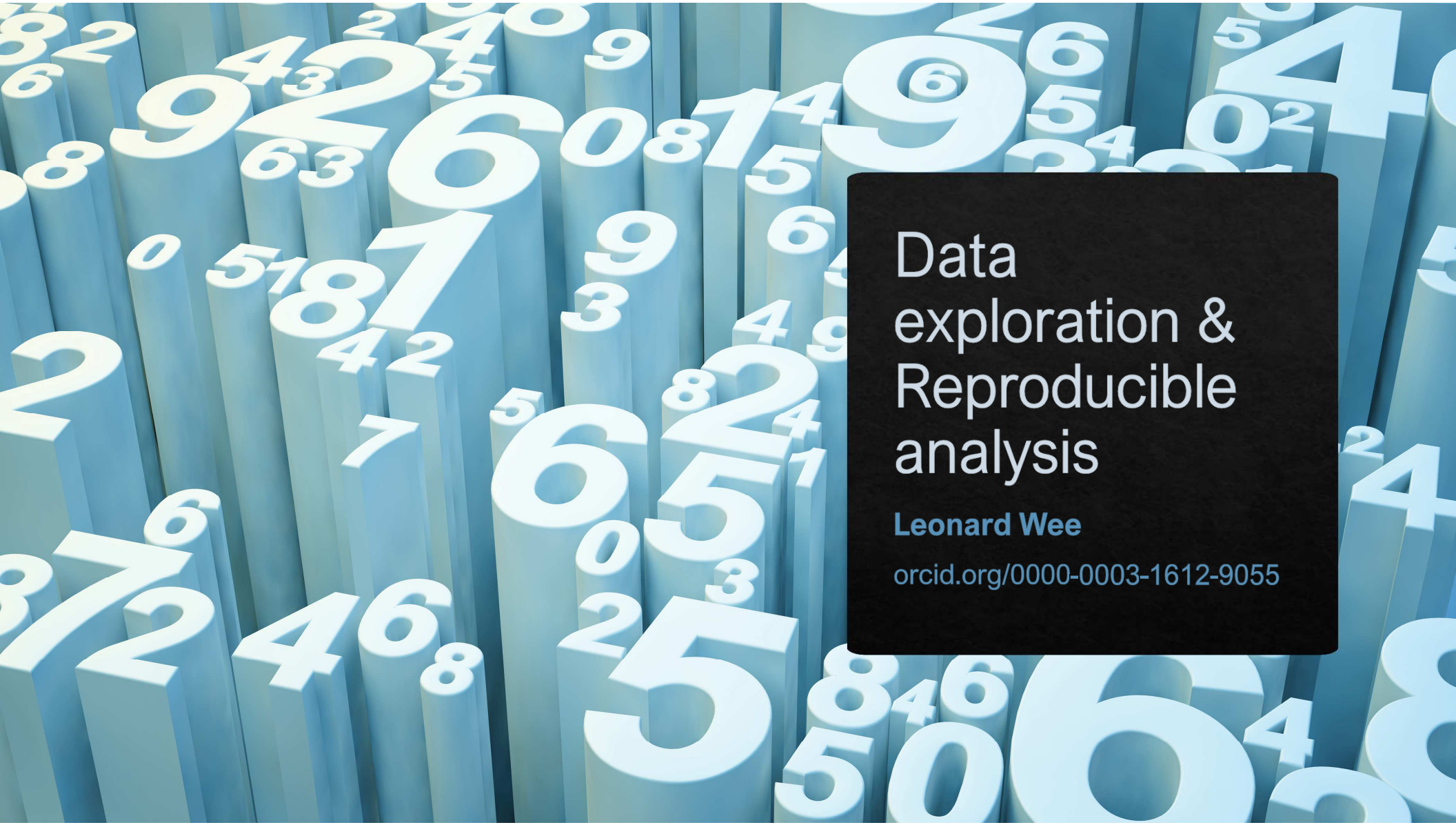
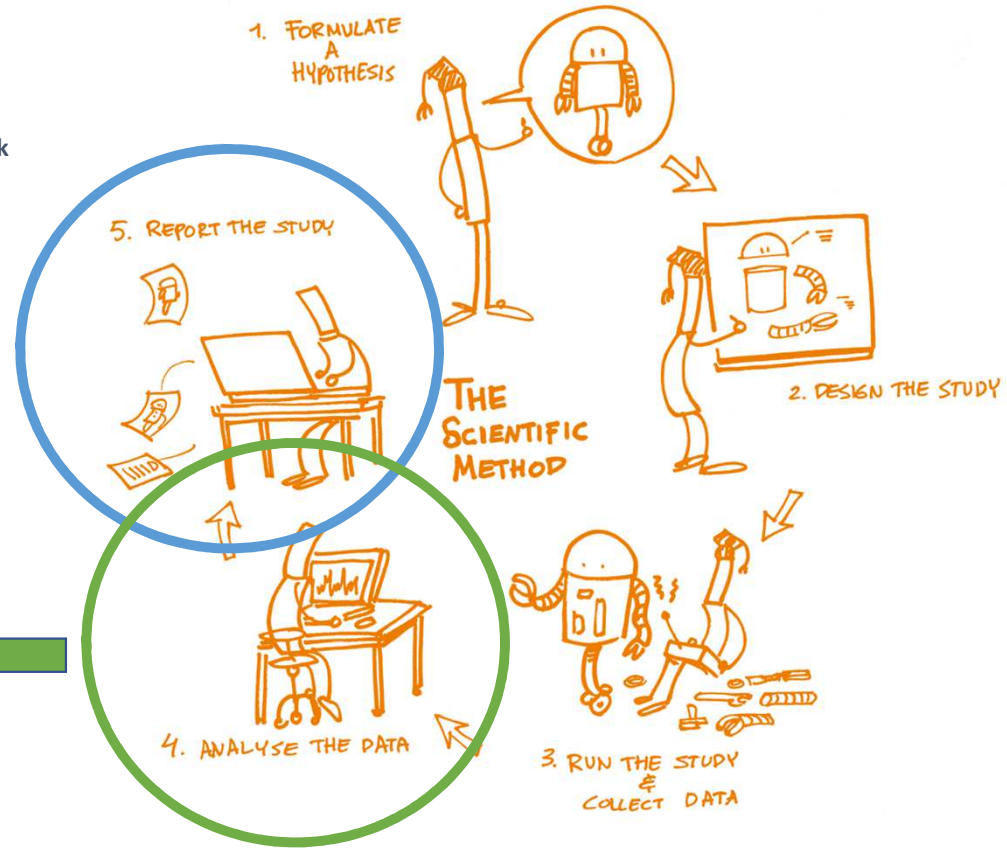

Data exploration & Reproducible analysis
Leonard Wee
orcid.org/0000-0003-1612-9055

What do we cover in this elective module?

Figure from :
The Open Science Training Handbook



R Markdown

Basics of reproducible data analysis

Learning Objectives

Reproducibility

A consistent coding style

Read-able / Re-usable code

Using Markdown

Exploration

Potential data issues

Variation within variables

Co-variation between variables

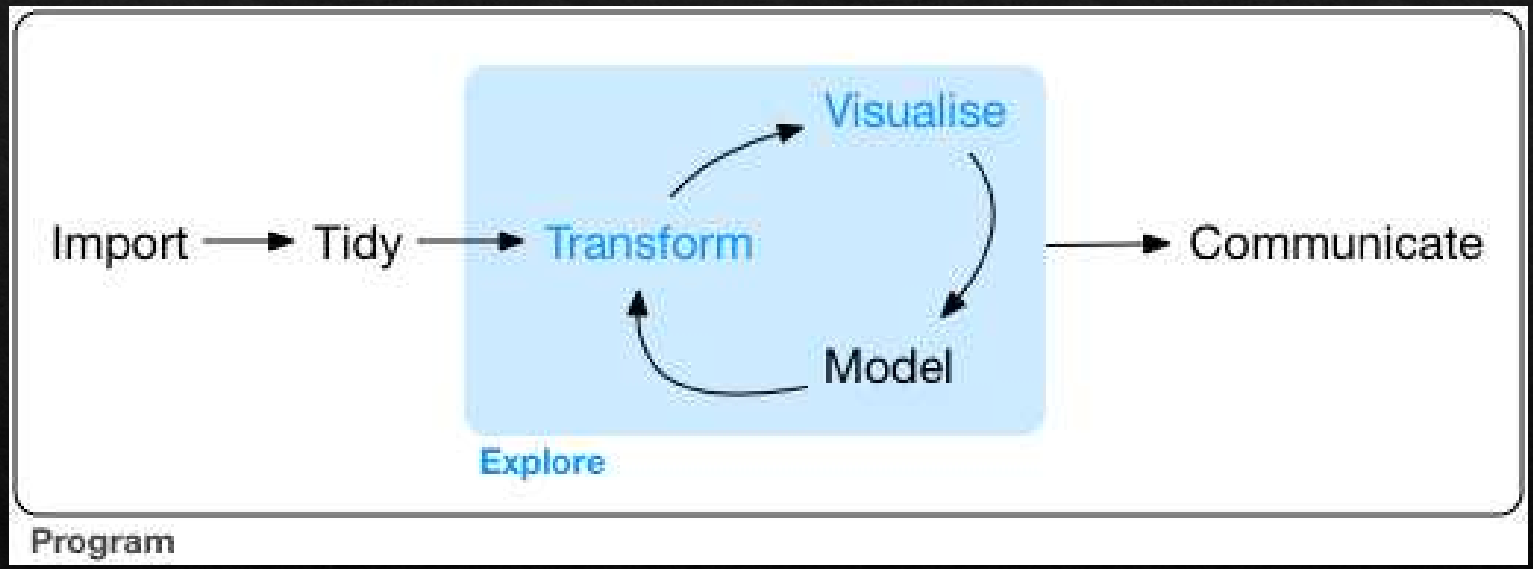
Visualization

Build on top of recipes

Keep it simple!

Be convincing, not fancy!

Ch 2 - 8 : R for Data Science



By Hadley Wickham
https://r4ds.had.co.nz/explore-intro.html

Reproducible analysis

- Reproducibility (Ioannidis 2016) :
  - the ability to **duplicate the results of a prior study using the same materials and methods as were used by the original investigator**.
- Replication :
  - the ability to **obtain the same conclusions as a prior study** if the same procedures are followed **with newly or fully-independently sourced data**.
- **Reproducibility puts emphasis on materials and methods.**
- **Replication puts emphasis on validity of results.**

Where to begin working reproducibly?

- Make a persistent, accessible and auditable repository **e.g. GitHub**
- Use version control **e.g. Git**
- Use universally accessible tools and code templates **e.g. R and RStudio**
- "Literate programming" that integrates analysis code with understandable narrative and good documentation **e.g. R markdown**
- **Organize analysis work to be readable, reportable and re-usable.**
- **Data exploration** – understand what data is already available/attainable.
- Design a **good analysis plan** **BEFORE** the plunging into models, predictions, etc.

Switch to markdown and html "data_exploration_heart_fail.Rmd"

- If you have not already done so :
- Rstudio create new project from Git (first time) or **Rstudio git checkout main and then pull**
- **Rstudio git checkout –b [give your branch a name]**
- **Check where you are : git branch**
- **Rstudio Select "data_exploration_heart_fail.Rmd"**

Recap : Reproducibility

- A consistent coding style
  - Keep closely-related actions in **code blocks**
  - Numerical, factorial and binary **data types made explicitly known**
- Readability and reusability
  - **Include data** in your code, or code the import from a stable and persistent source
  - **Small reusable functions** instead of repeating hard-coded steps
  - “Source” helpful functions from separate R file(s)
- Exploit R Markdown for keeping code with clean documentation of work
  - Rmd and PDF/HTML does not always have to be the exact same!

Data exploration

- Also referred to as ‘Exploratory Data Analysis’ (EDA)
  - What would be the usual (or unusual) amount of variation within my variables?
  - What, if any, is the degree of co-variation between my variables?
- The primary goal of EDA is to understand **what potential insights might be inside** the data
- And one of the primary means of deriving understanding is to LOOK at it in multiple different ways – ie transformations and data visualizations.

Essential quotes from the textbook

- EDA is fundamentally a **creative** process.
- And like most creative processes, the key to asking *quality* questions is to **generate a large *quantity* of questions**. It is difficult to ask revealing questions at the start of your analysis because you do not know what insights are contained in your dataset.
- On the other hand, each new question that you ask will expose you to a new aspect of your data and increase your chance of making a discovery.
- You can quickly drill down into the most interesting parts of your data—and develop a set of thought-provoking questions—if you follow up each question with a new question based on what you find.

Visualization

- Paraphrasing John Tukey : "**A simple picture has brought more information to the data analyst's mind** than any other device".
- Or my crusty old handyman : "Measure twice cut once."
- "Cognitive" computing -
  - **PATTERN cognition and recognition.**
  - Before engaging a machine "brain", first use the wet squishy one in your skull.
  - The endpoint of visualization is **UNDERSTANDING what is in the data** and (later on) deciding which question(s) you can or cannot answer with it.

Some extra excellent resources :

- R for Data Science textbook (absolutely must have on your desk)
- **R Graphics Cookbook** (recipes for "baking" most figures you may think of)
- **https://www.data-to-viz.com/** summarises all the figures you could make using R, simply beautiful and beautifully simple
- **https://informationisbeautiful.net/** speechlessly stunning (for slightly more advanced data science skills)
- Disclaimer : the aim of this part of the lecture is to **start you on reading, understanding and re-using pieces of R (here and online),** then re-organize it and change it to your own purpose – try not to have to write all code from scratch.
- **Keep your visualizations simple and insightful**, fancy is definitely not better!

Switch to markdown and html "data_visualization_examples.Rmd"

- If you have not already done so :
- Rstudio create new project from Git (first time) or **Rstudio git checkout main and then pull**
- **Rstudio git checkout –b [give your branch a name]**
- **Check where you are : git branch**
- **Rstudio Select "data_visualization_examples.Rmd"**

Recap : Exploration & Visualization

- Essential part of EDA that engages the inbuilt human intuition by seeing
  - **Explore variation within variables**
  - **Explore co-variation between variables**
- Understand patterns and trends, do not try to model or predict at this moment
- **Complicated visualization is NOT better :**
  - Explore a lot, but show your reader only what is important
  - Present a small number of relevant and insightful visuals
  - Keep it simple – no extra marks for pointless complexity!
  - **Each visual with an ESSENTIAL part of the story to tell**

Recap Learning Objectives

Reproducibility

Consistent coding style

Read-able / Re-usable

Markdown

Exploration

Potential data issues

Variation within variables

Co-variation between variables

Visualization

Build on top of recipes

Keep it simple!

Tell a pictorial story