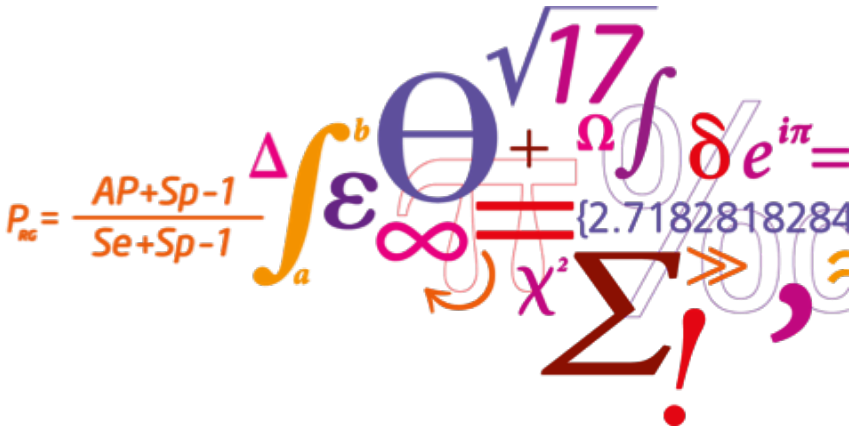
How to organize a project

The most important talk you never heard!

DTU Bioinformatics hackinar February 8th 2018

Leon Eyrich Jessen
Postdoc
Immunoinformatics and Machine Learning
Department of Bio and Health Informatics
Technical University of Denmark



Immunoinformatics and Machine Learning
Department of Bio and Health Informatics

The most important talk you never heard!

- Think about it...
  - How many courses have you attended?
  - How many classes have you taken?
  - How many talks have you been to?
  - Etc.

- Has anyone ever talked to you about the underlying machinery?

- Has anyone ever presented you to or with a project organisation plan?

- How were the results you are being presented to produced?

- I know there are supposed to be a materials and methods section in papers – Have you ever been tasked with deciphering and repeating such a section?

The Corner Stone of Research

- In essence - What is that we do?
  - We produce knowledge!
  - We disseminate knowledge!

- But…
  - You cannot simply say I found ‘Z’
  - You HAVE to be able to account for how you got from ‘A’ to ‘Z’

- Reproducible Research!
  - Being able to (easily) reproduce every single result from a paper

- Why?
  - Basically, we need to be able to see if you are cheating
  - Others need to stand on your shoulders

So, how are we doing?

So, how are we doing?

So, how are we doing?

- ”More than 70% of researchers have tried and failed to reproduce another scientist's experiments”

- ”More than half have failed to reproduce their own experiments”

So, what can we do?

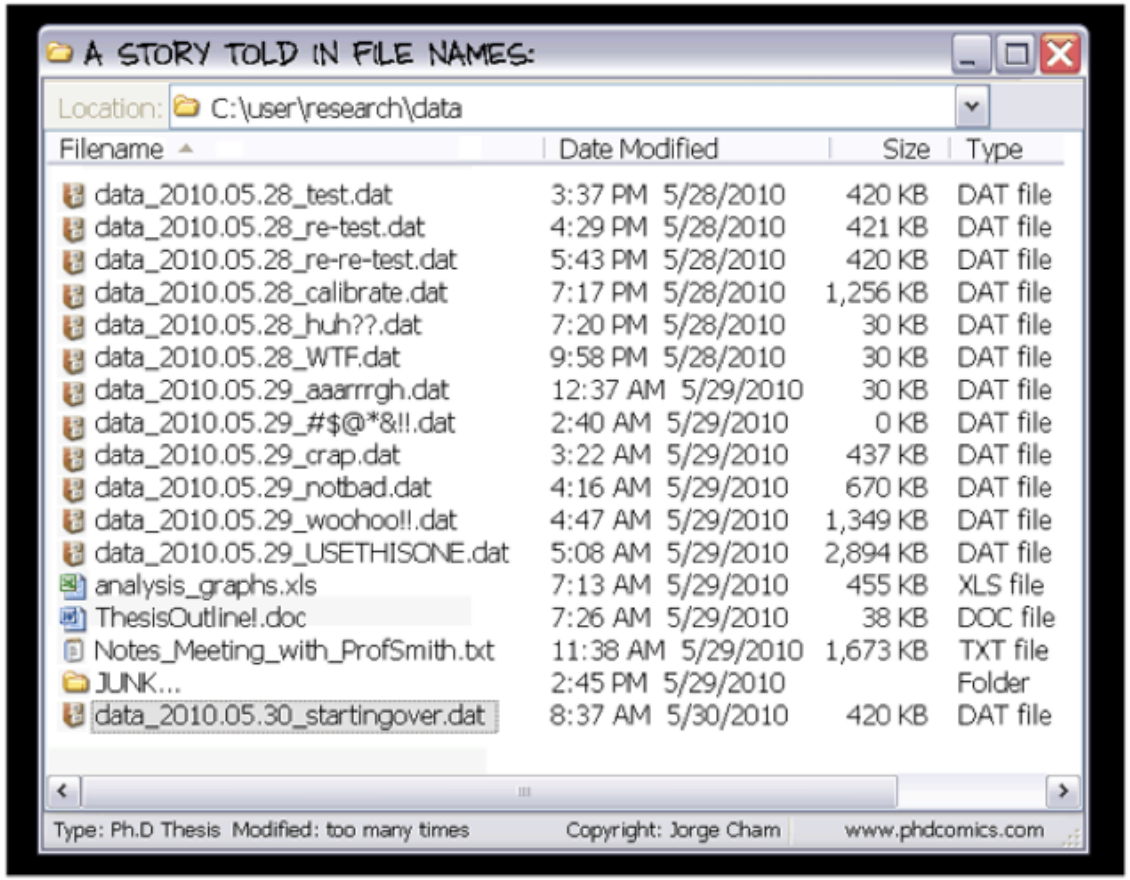
- Granted, some of the reasons for the reproducibility crisis is beyond our control
  - Biology is notoriously messy
  - False positives
  - Etc.

- However, a step in the right direction is to think about organising and documenting your research

- I have seen many times people revisiting old projects only to find that they cannot figure the project out or even reproduce it or understanding the project is so time consuming, that repeating it is more time efficient

- Why does this happen?
  - Admitted, we're all storm chasers – Always on the hunt for the next publication
  - Many see documentation as a waste of valuable time

Familiar?


A STORY TOLD IN FILE NAMES:
Location: C:\user\research\data
Filename | Date Modified | Size | Type
data_2010.05.28_test.dat 3:37 PM 5/28/2010 420 KB DAT file
data_2010.05.28_re-test.dat 4:29 PM 5/28/2010 421 KB DAT file
data_2010.05.28_re-re-test.dat 5:43 PM 5/28/2010 420 KB DAT file
data_2010.05.28_calibrate.dat 7:17 PM 5/28/2010 1,256 KB DAT file
data_2010.05.28_huh??.dat 7:20 PM 5/28/2010 30 KB DAT file
data_2010.05.28_WTF.dat 9:58 PM 5/28/2010 30 KB DAT file
data_2010.05.29_aaarrrgh.dat 12:37 AM 5/29/2010 30 KB DAT file
data_2010.05.29_#$@*&!!.dat 2:40 AM 5/29/2010 0 KB DAT file
data_2010.05.29_crap.dat 3:22 AM 5/29/2010 437 KB DAT file
data_2010.05.29_notbad.dat 4:16 AM 5/29/2010 670 KB DAT file
data_2010.05.29_woohoo!!.dat 4:47 AM 5/29/2010 1,349 KB DAT file
data_2010.05.29_USETHISONE.dat 5:08 AM 5/29/2010 2,894 KB DAT file
analysis_graphs.xls 7:13 AM 5/29/2010 455 KB XLS file
ThesisOutline!.doc 7:26 AM 5/29/2010 38 KB DOC file
Notes_Meeting_with_ProfSmith.txt 11:38 AM 5/29/2010 1,673 KB TXT file
JUNK... 2:45 PM 5/29/2010 Folder
data_2010.05.30_startingover.dat 8:37 AM 5/30/2010 420 KB DAT file
Type: Ph.D Thesis Modified: too many times
Copyright: Jorge Cham
www.phdcomics.com

Who has taught you to organise a project?

- "In practice, the principles behind organizing and documenting … are often learned on the fly"

PLoS Comput Biol. 2009 Jul; 5(7): e1000424.
Published online 2009 Jul 31. doi: 10.1371/journal.pcbi.1000424
PMCID: PMC2709440

A Quick Guide to Organizing Computational Biology Projects

William Stafford Noble [1,2,*]

Let’s dive in…

- The following is inspired by this paper

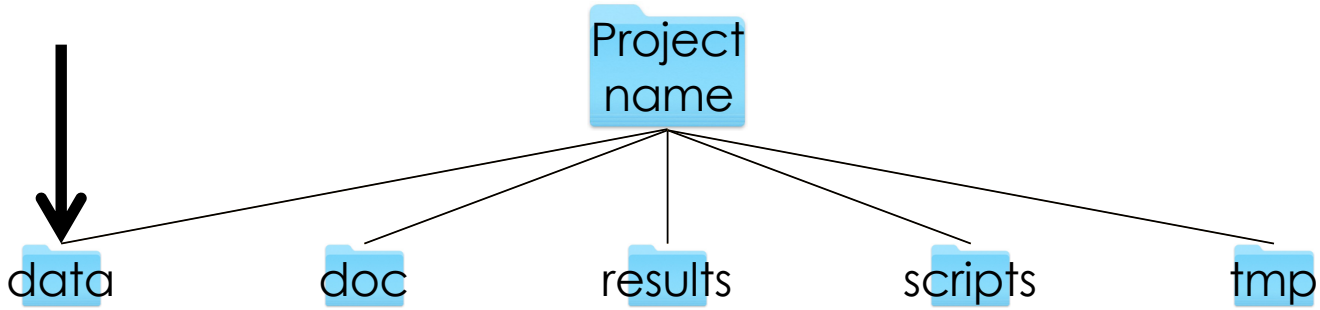
PLoS Comput Biol. 2009 Jul; 5(7): e1000424.
Published online 2009 Jul 31. doi: 10.1371/journal.pcbi.1000424

PMCID: PMC2709440

A Quick Guide to Organizing Computational Biology Projects

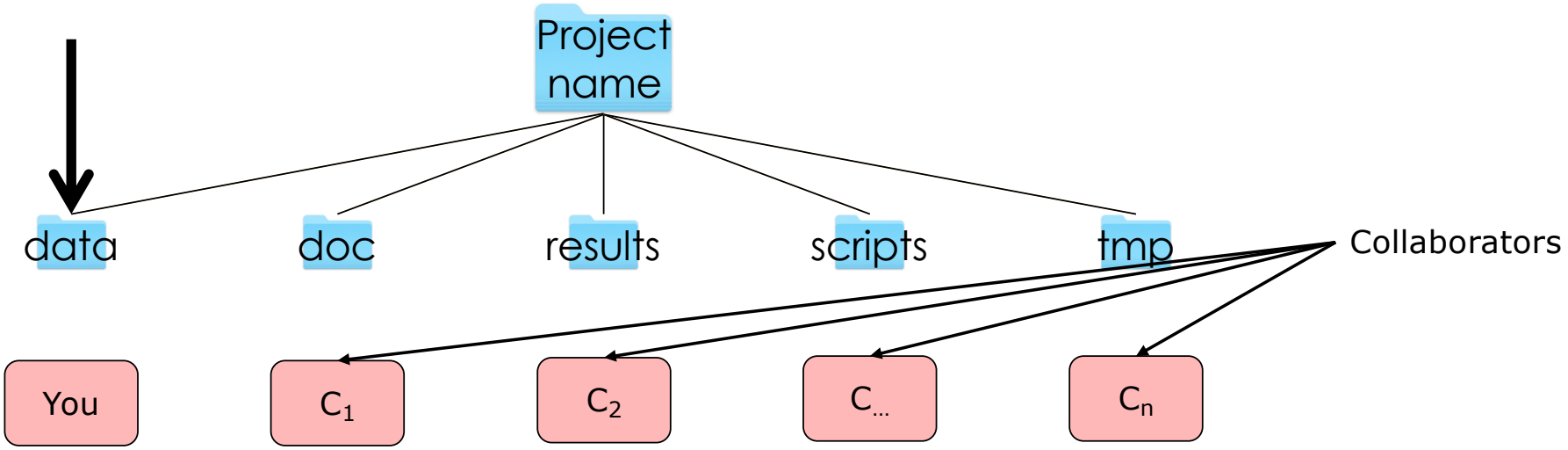
William Stafford Noble [1,2,*]

Project Directory Structure

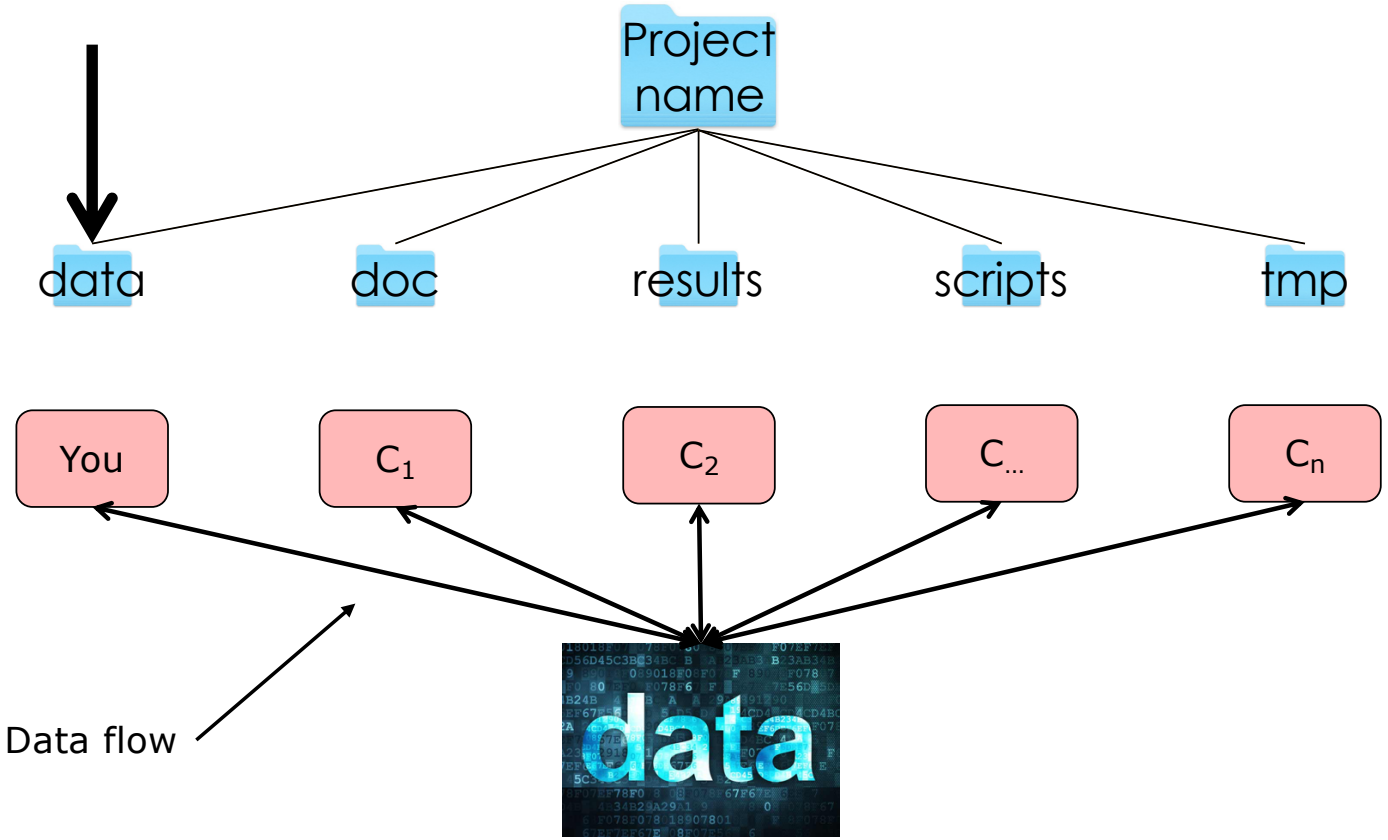


- Raw Data should always be pulled from central source, **never** from an excel sheet someone sent to you
- You are not allowed to touch or alter the original raw data
- Make sure that every step from the raw data, to the data you use for analysis can be repeated
- Save the cleaned data and proceed from that

Project Directory Structure

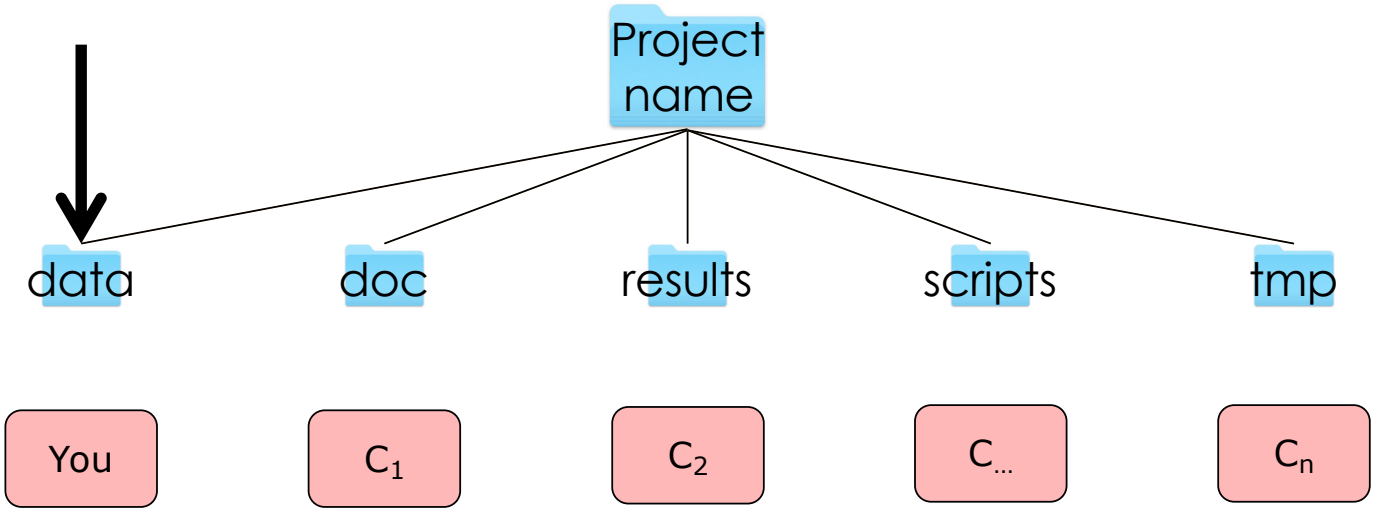


Project Directory Structure

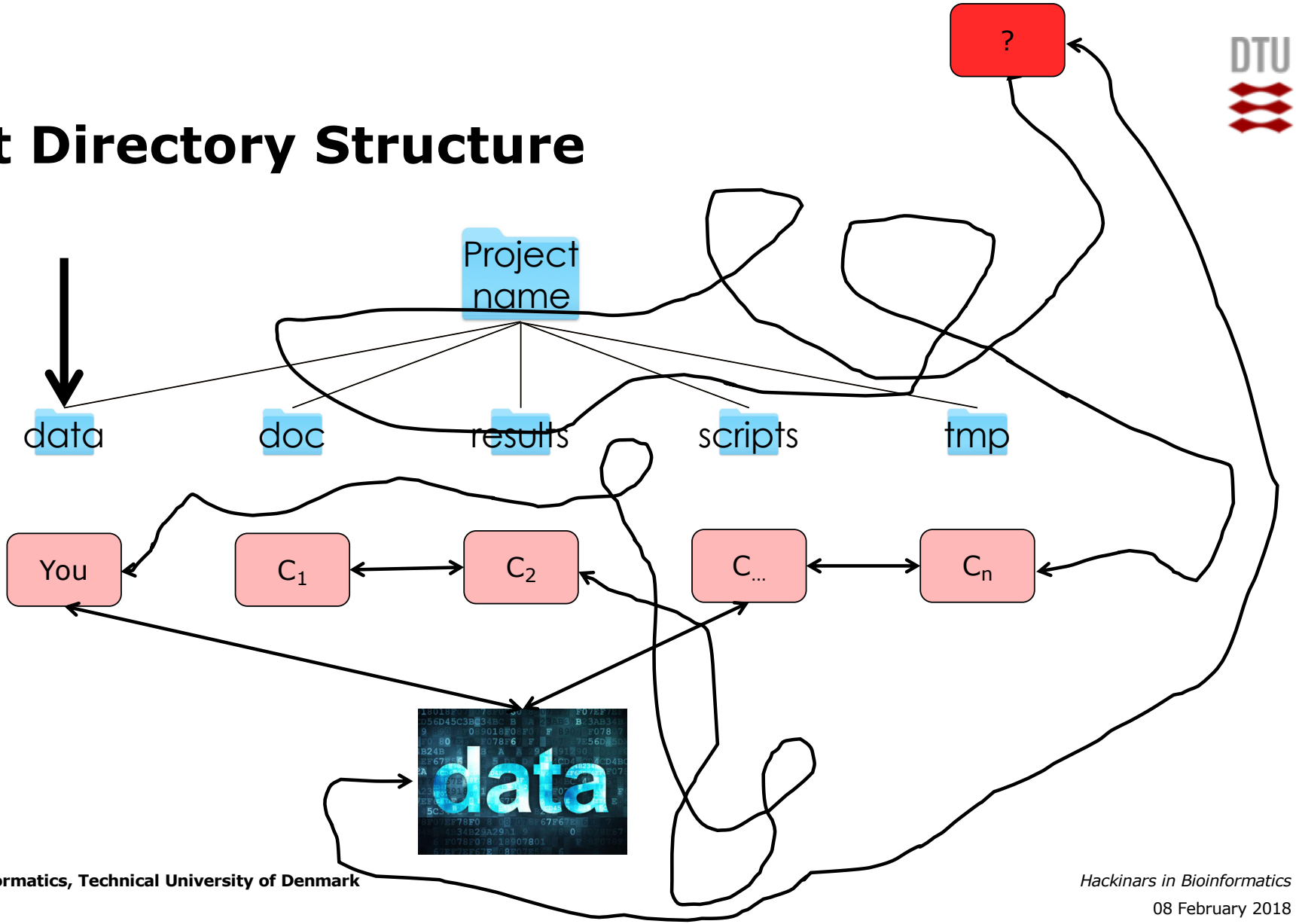


Project Direct

Project Directory Structure

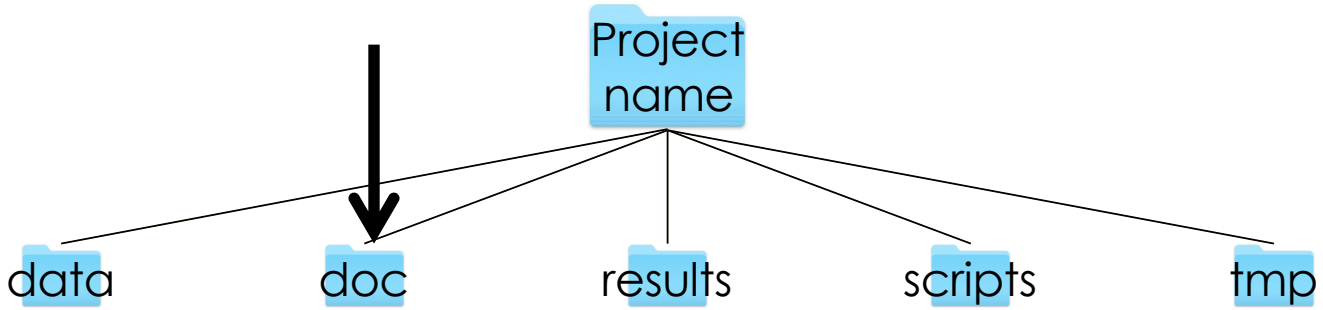


Project Directory Structure



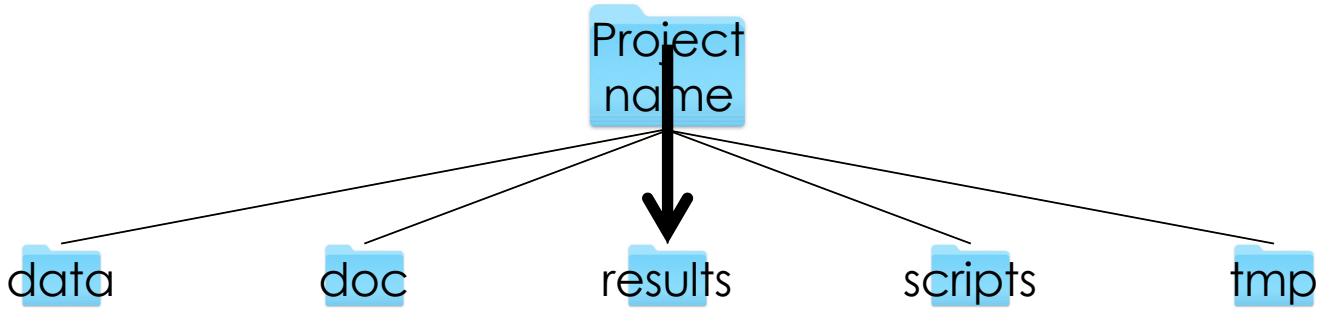
Project Dire

Project Directory Structure



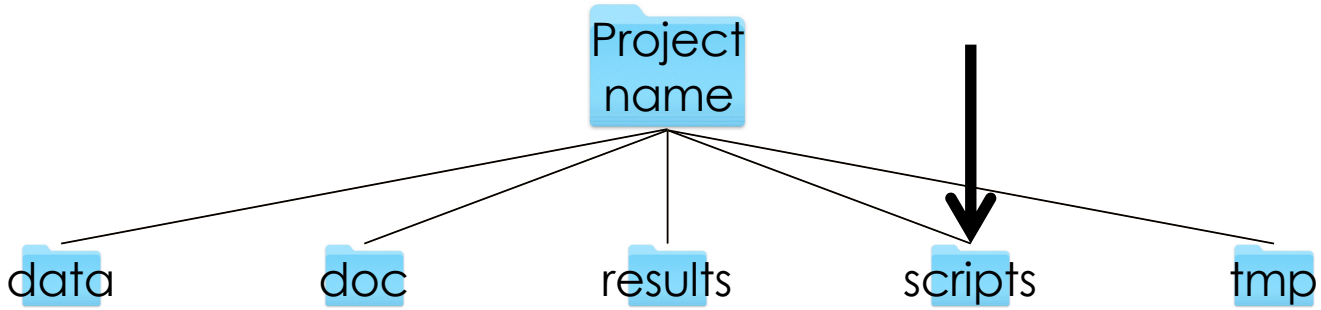
- This is where your manuscript lives
- Notes, presentations, pdfs and alike pertaining to the project
- Etc.

Project Directory Structure



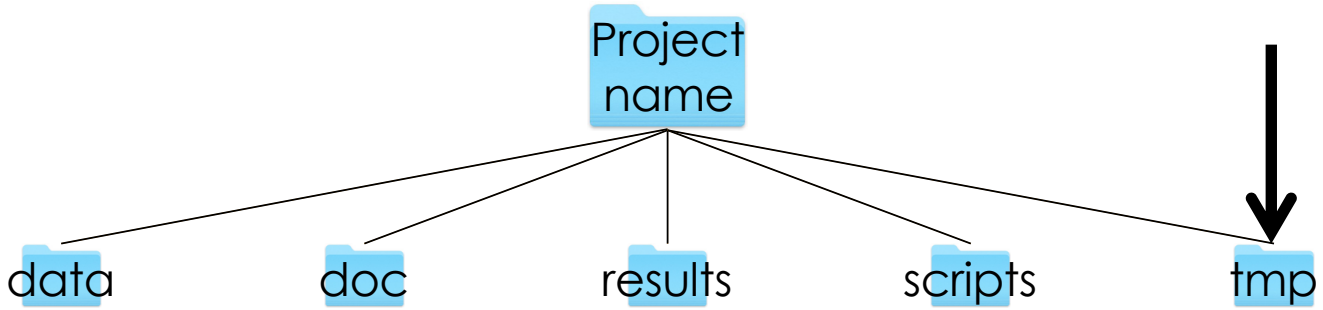
- Anything considered a results
- Plots
- Text file with p-value tables
- Etc.

Project Directory Structure



- This is where your analysis scripts are places
- All scripts shall be able to run from start-to-end

Project Directory Structure



- Anything you can delete without thinking about it
  - Tests
  - Stuff you want to check
  - Temporary exploratory files
  - Etc…

Building your 'scripts' directory

- Load-clean-func-do philosophy

- First scripts takes your raw data from raw to analysis-ready
  - Raw data is loaded and cleaned
  - Clean data and versions hereof are saved for subsequent use

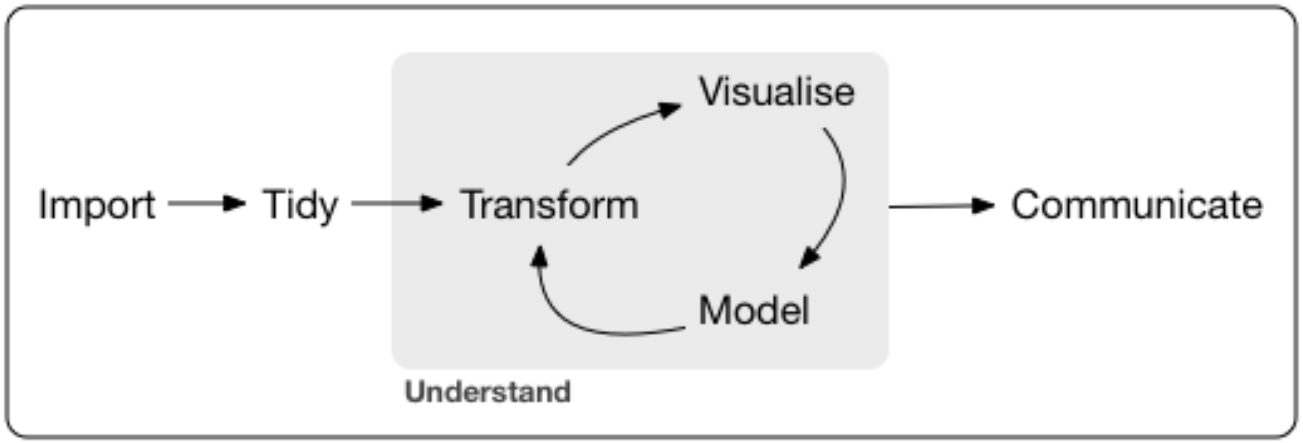
- Project specific functions are put in a separate file

- A single do file is defined capable of running the ENTIRE project and produce ALL results

- Collect the results in a markdown file

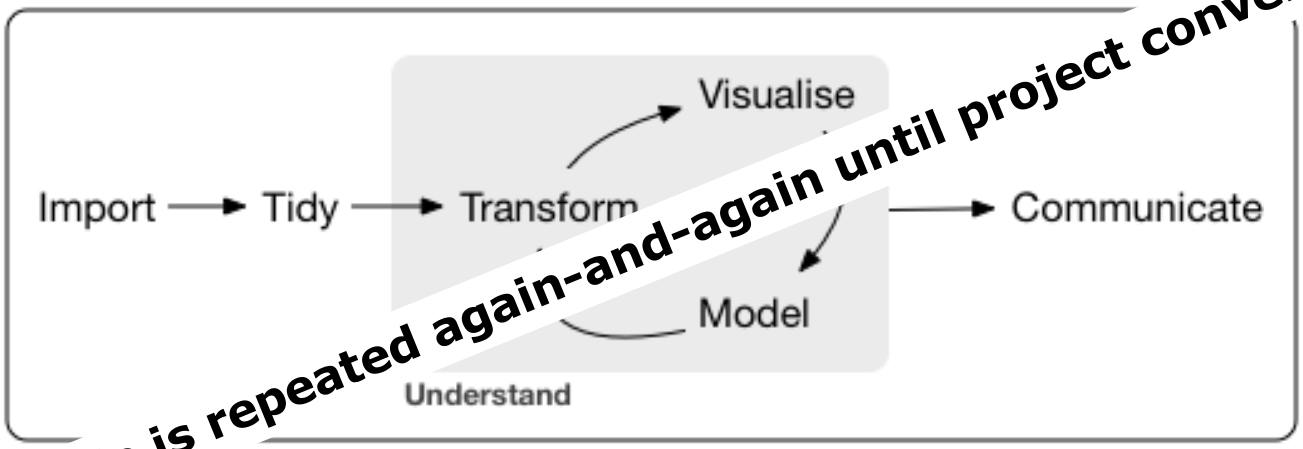
- Use GitHub for sharing, version control and backup

The essence of data science



http://r4ds.had.co.nz/

The essence of data science



This cycle is repeated again-and-again until project converges

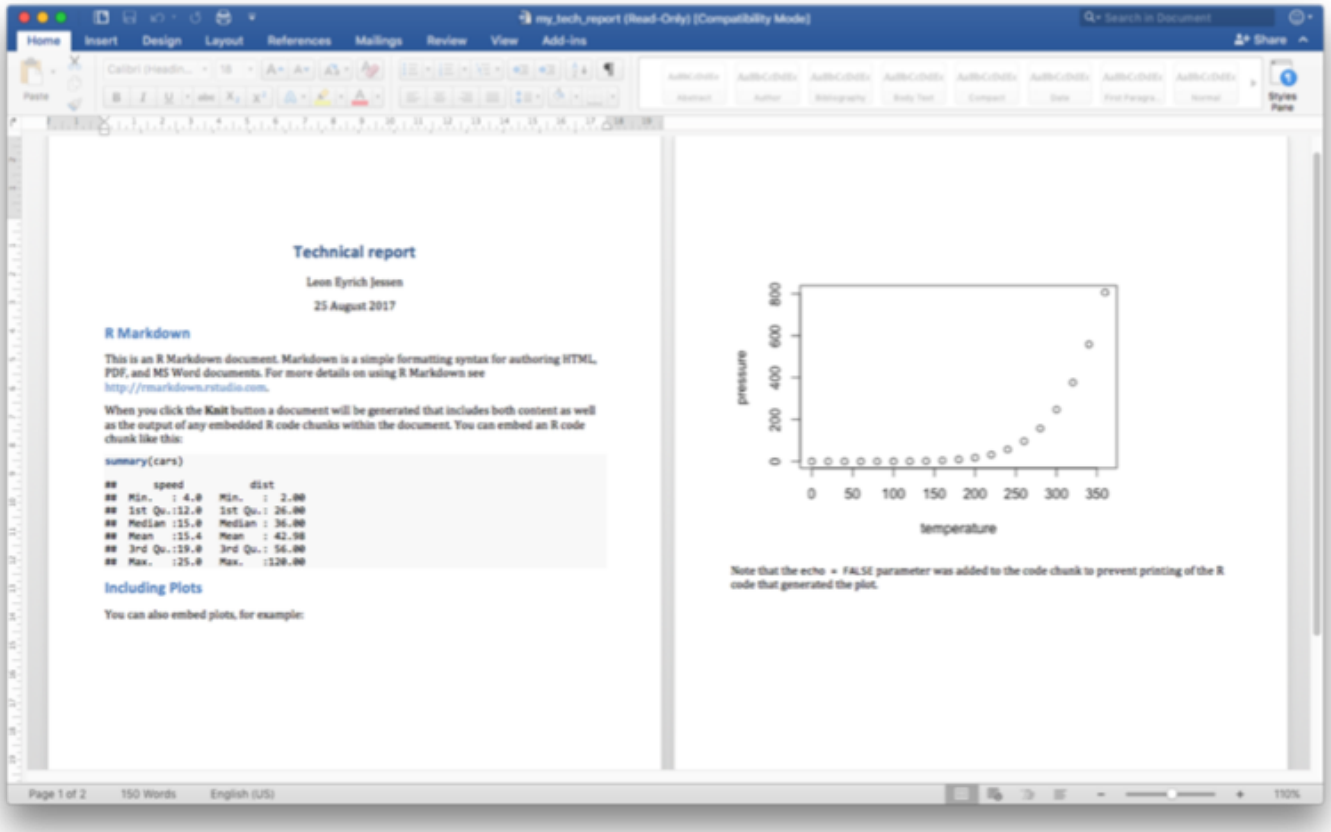
Project finalisation

- Ideally,
  - Once analysis has converged, a technical report should be created using markdown
  - Once the paper is published the project directory should be frozen as read-only
  - The directory should contain everything needed to recreate all the exact figures and tables in the paper

Briefly on markdown

- Dynamic generation of technical documents
- All numbers, pictures, tables etc. are dynamic
- If input changes, so does output
- Can be generated to output word, pdf and html

Markdown - Word



Technical report

Leon Eyrich Jessen

25 August 2017

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

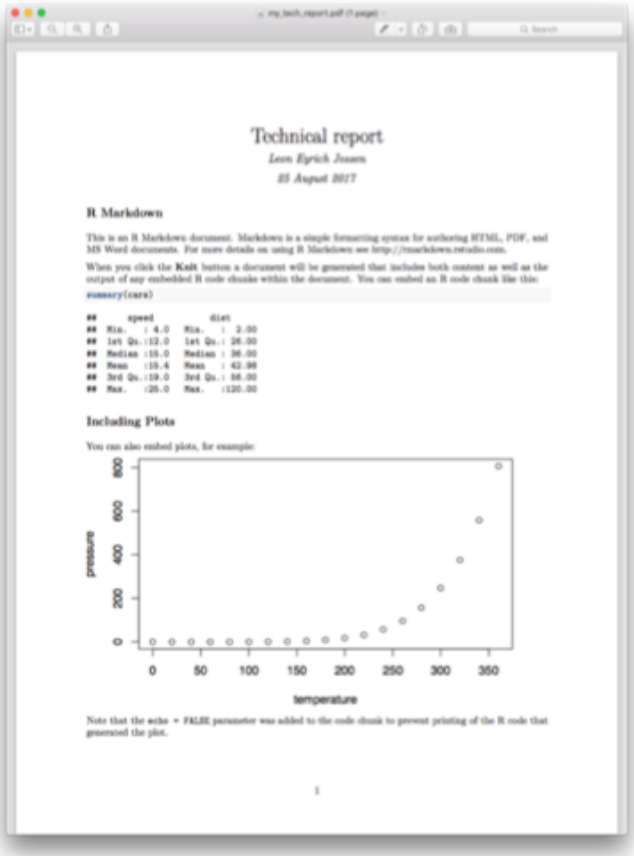
```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:

Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Markdown - PDF



Technical report

Leon Eyrich Jessen

05 August 2017

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

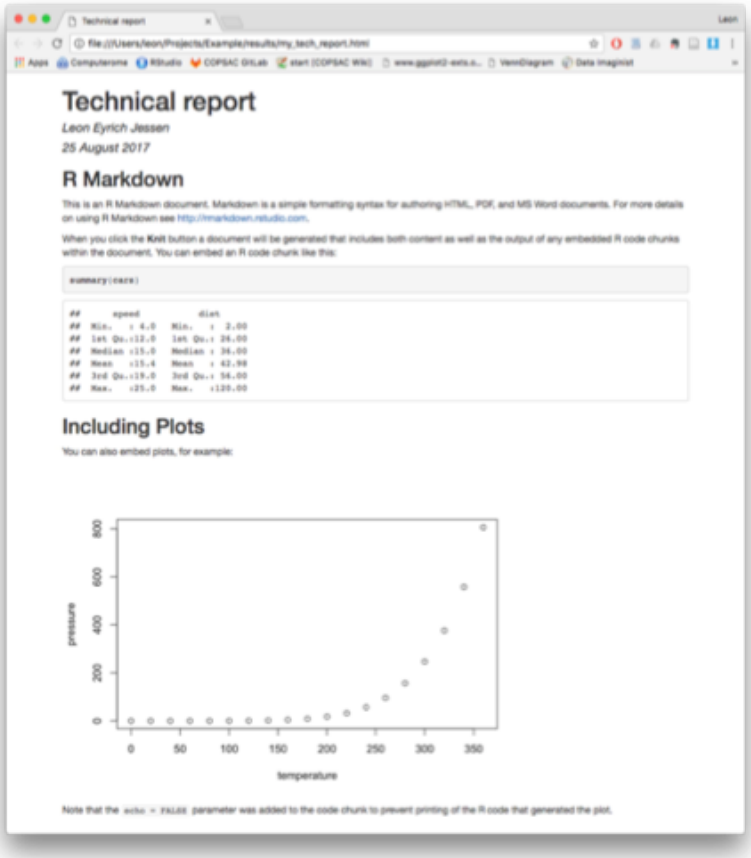
Including Plots

You can also embed plots, for example:

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

1

Markdown - HTML



Technical report

Leon Eyrich Jessen

25 August 2017

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

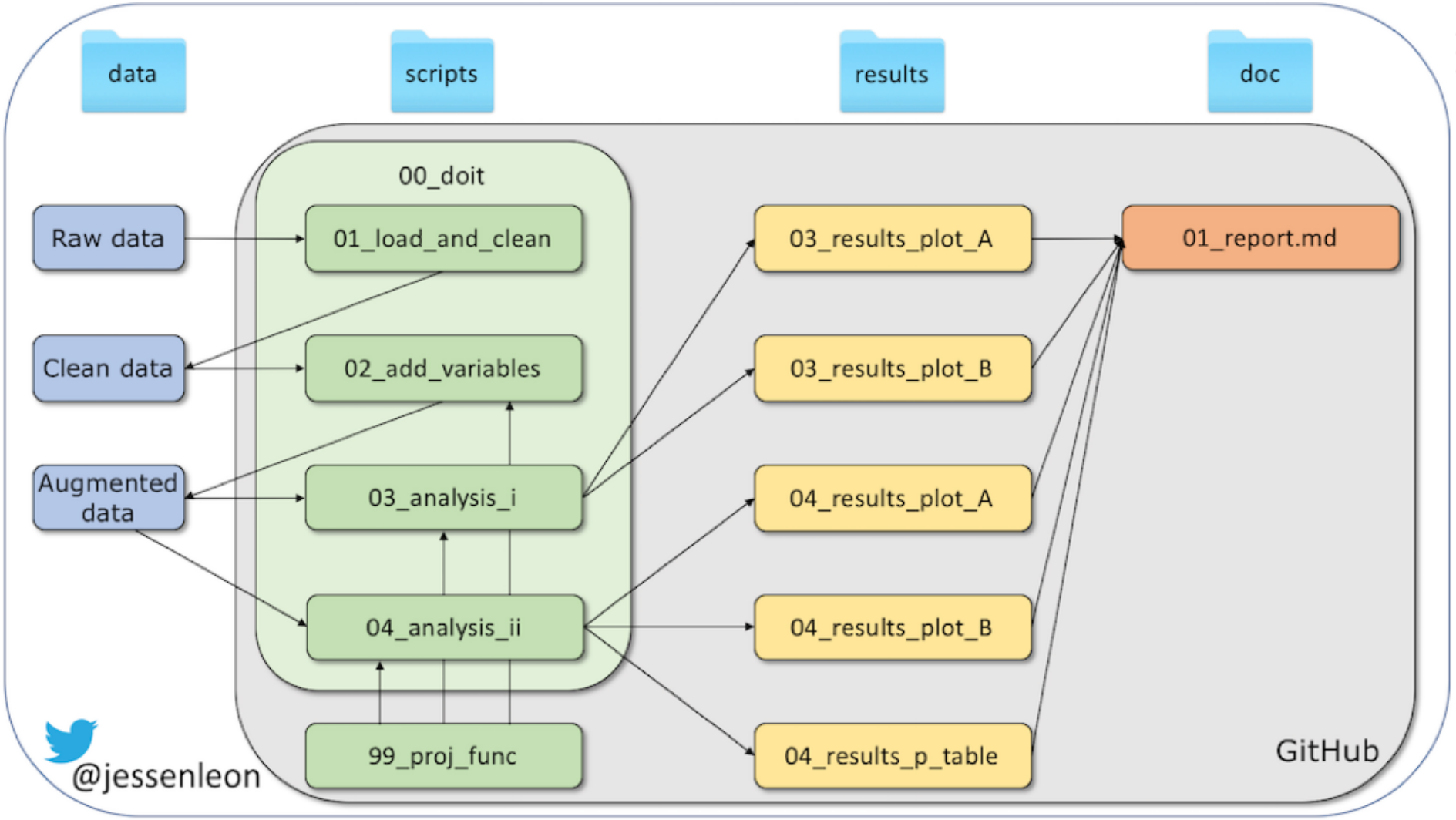
Including Plots

You can also embed plots, for example:

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

In conclusion

- This is not the absolute truth
- This is my current take on a how-to data science
- Structure takes time in order to save time
- Am I adhering 100% to this always and forever? No, but…
- I strongly believe that striving for structure is better than abandoning it
- A picture speaks a thousand words - Let’s try to visualise it!


data
scripts
results
doc
00_doit
Raw data
Clean data
Augmented data
01_load_and_clean
02_add_variables
03_analysis_i
04_analysis_ii
99_proj_func
03_results_plot_A
03_results_plot_B
04_results_plot_A
04_results_plot_B
04_results_p_table
01_report.md
@jessenleon
GitHub

> Think about readability of your code. Every project you work on is fundamentally collaborative. Even if you are not working with any other person, you are always working with future you and you really do not want to be in a situation where future you has no idea what past you was thinking, because past you will not respond to any emails!
> - Hadley Wickham