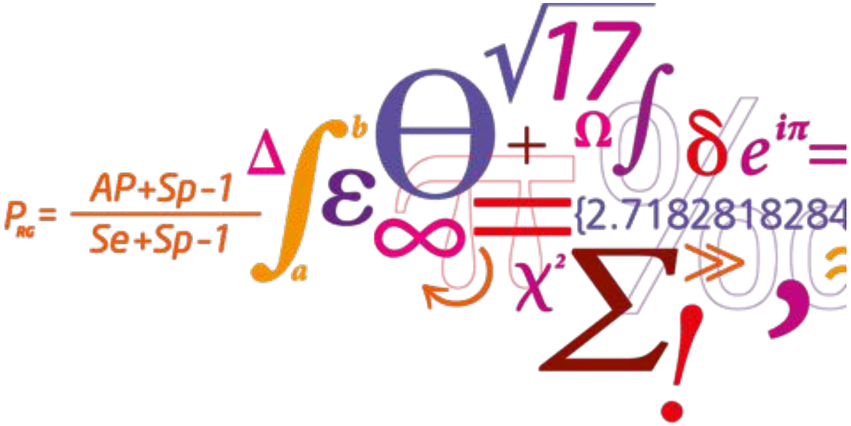
Leon Eyrich Jessen



Immunoinformatics and Machine Learning
Department of Bio and Health Informatics

- *Eur J Heart Fail: "Our findings do not support a major role for fish intake in the prevention of heart failure"*

- Eur J Clin Nutr: *"Moderate consumption of fatty fish … were associated with a lower rate of first HF hospitalization or death"*

1. Eur J Heart Fail. 2009 Oct;11(10):922-8. doi: 10.1093/eurjhf/hfp126
2. Eur J Clin Nutr. 2010 Jun;64(6):587-94. doi: 10.1038/ejcn.2010.50. Epub 2010 Mar 24

- *Eur J Heart Fail: “Our findings do not support a major role for fish intake in the prevention of heart failure”*

- Eur J Clin Nutr: *“Moderate consumption of fatty fish … were associated with a lower rate of first HF hospitalization or death”*

FAKE

NEWS?

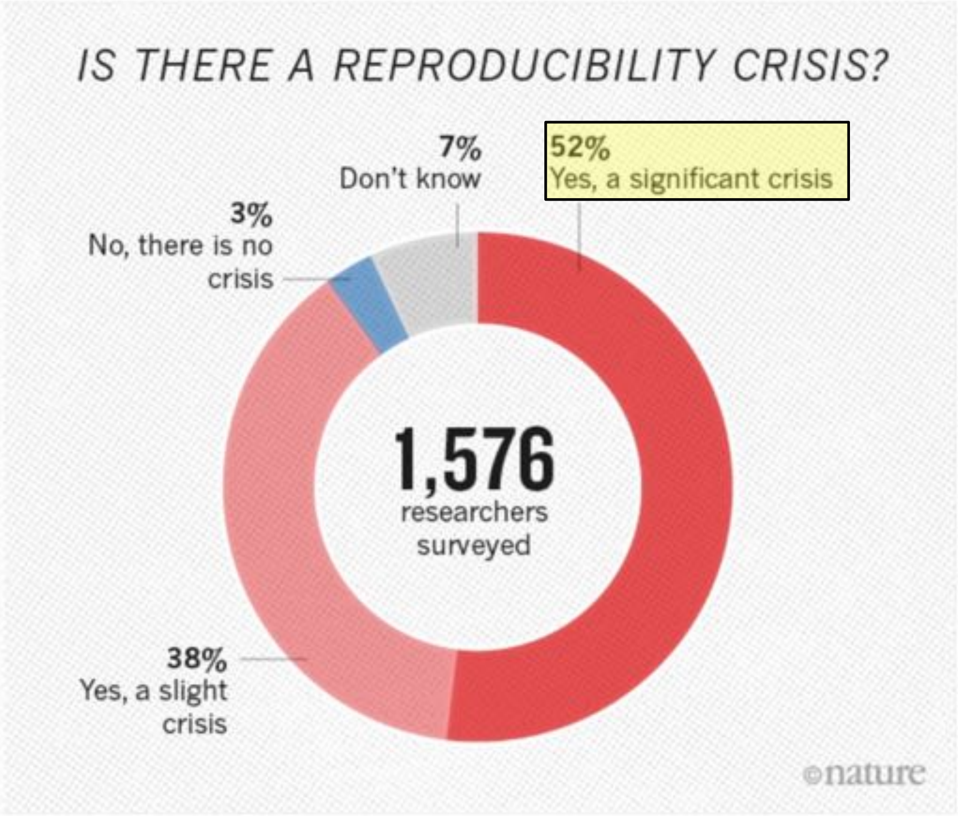
1. Eur J Heart Fail. 2009 Oct;11(10):922-8. doi: 10.1093/eurjhf/hfp126
2. Eur J Clin Nutr. 2010 Jun;64(6):587-94. doi: 10.1038/ejcn.2010.50. Epub 2010 Mar 24

Plenum

- *Take 1 minute to discuss with your neighbour:*
  - *Possible consequences of contradictory research results?*

Nature | News Feature (May 2016)

- 1,500 scientists lift the lid on reproducibility
- Based on questionnaire
- 52% “Yes, a significant crisis”
- “More than 70% of researchers have tried and failed to reproduce another scientist's experiments”
- “More than 50% have failed to reproduce their own experiments”



Baker, M., Nature 533, 452–454 (26 May 2016) doi:10.1038/533452a

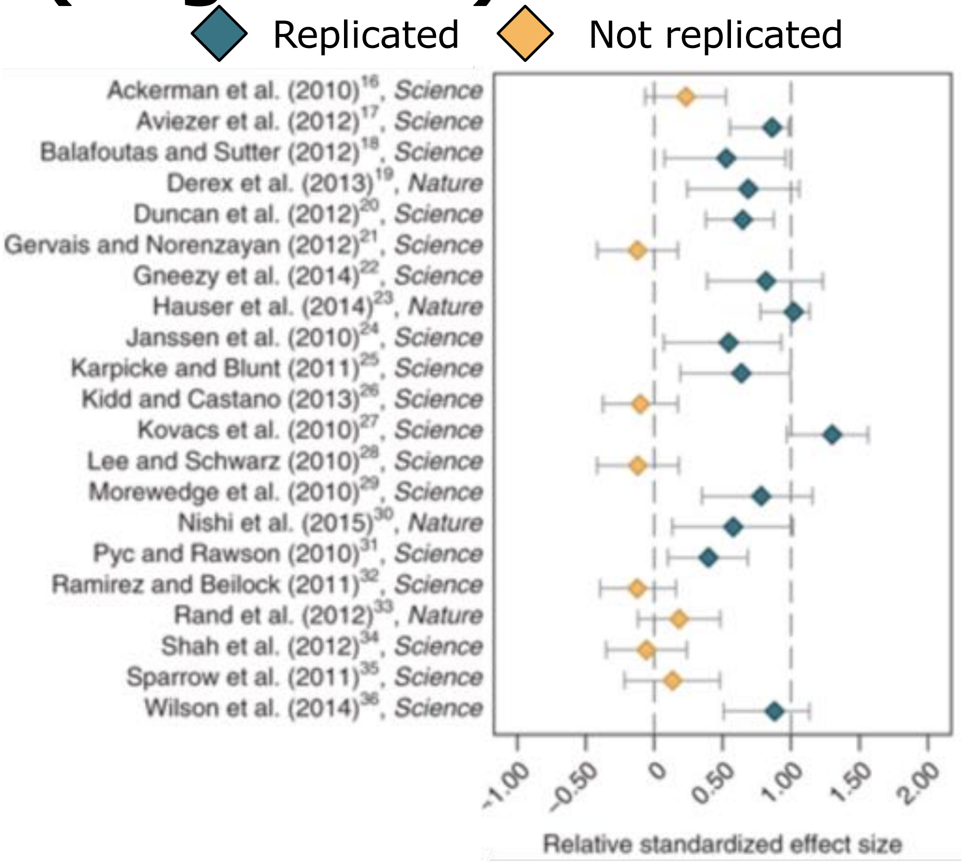
Nature | Human Behaviour (Aug. 2018)

- Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015

- Based on re-analysis of published results

- "There is a significant effect in the same direction as in the original study for 13 out of 21 replications" (62%)

- I.e. 38% of the studies failed to replicate



Camerer, C.F. et al., Nature Human Behaviour, 2397-3374 (27 Aug 2018) 10.1038/s41562-018-0399-z

Plenum

- *Take 1 minute to discuss with your neighbour:*
  - *Possible <u>reasons</u> for irreproducible research results?*

Reasons for lack of reproducibility

- Many!

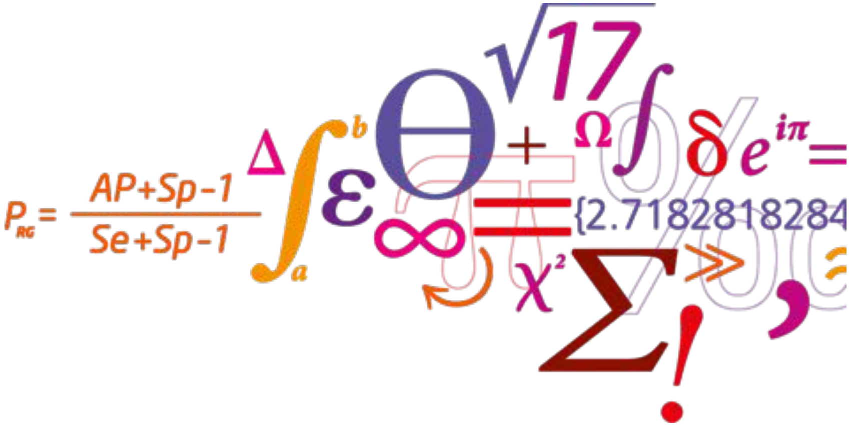
- Many of which cannot be controlled

- Focus for this talk is on one aspect we can control

- Namely
  - Reproducible data analysis workflow

Topic: Reproducible Data Analysis Workflow

Leon Eyrich Jessen



Immunoinformatics and Machine Learning

Department of Bio and Health Informatics

Motivation - Why bother?

- You are obliged to make reproducible research, it's the cornerstone of what we do
- During a review process, you might be asked to redo and/or expand your analysis
- If you revisit an unorganised project, 2 years later, you will loose *a lot* of time redoing everything
- Even worse, if you revisit someone else's project 2 years later…
- So, spend time to save time
- …and once again, we *really have to be able to redo an analysis* as effortlessly as possible and it's your responsibility, not your supervisors
- IMHO: Every detail of the analysis in a paper should be open source (100% transparency)

Learning Objectives

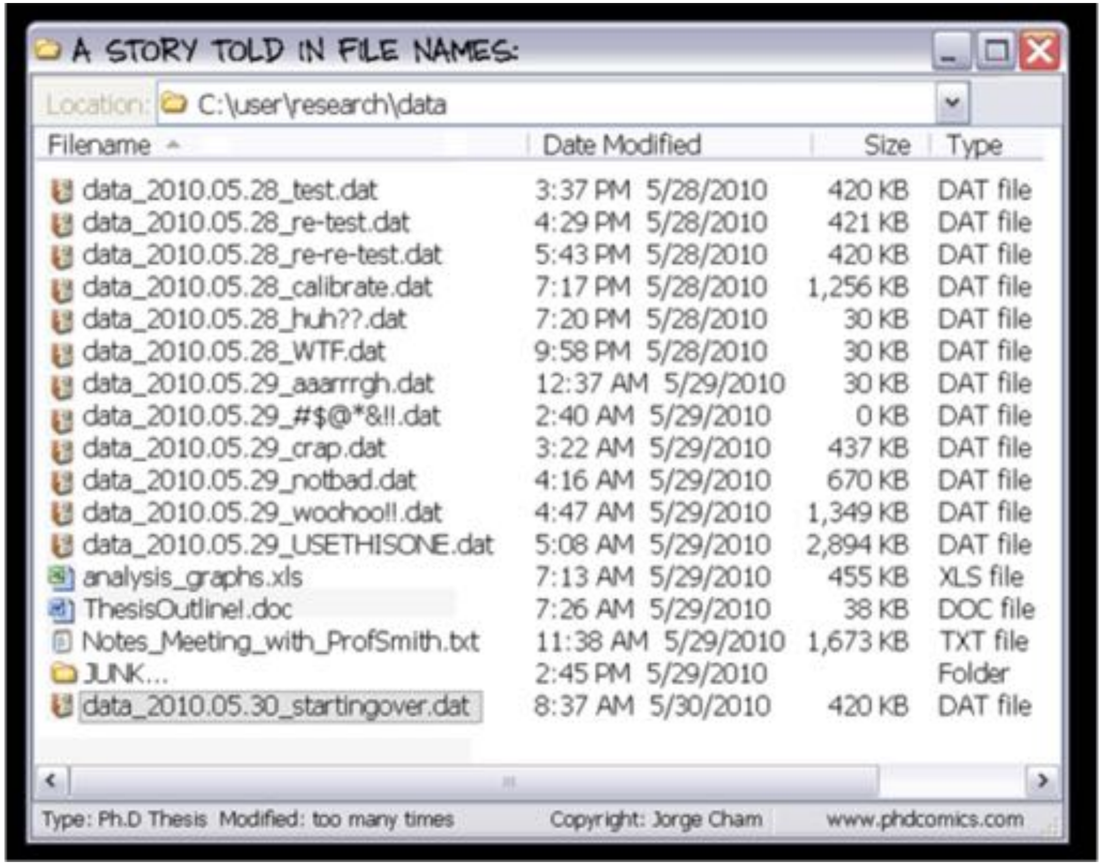
- After this session, you should be able to:

  1. *Define what a reproducible data analysis workflow is*
  2. *List the elements of a reproducible data analysis workflow*
  3. *Explain the meaning and purpose of each of the elements in a reproducible data analysis workflow*
  4. *When presented with a pre-made workflow, determine if it constitutes a reproducible data analysis workflow*

Definition

- *A reproducible data analysis workflow is when you can go from the raw data to recreating all the figures, tables and numbers in your paper automatically and consistently*

I’ve seen several data dirs like this



A STORY TOLD IN FILE NAMES:

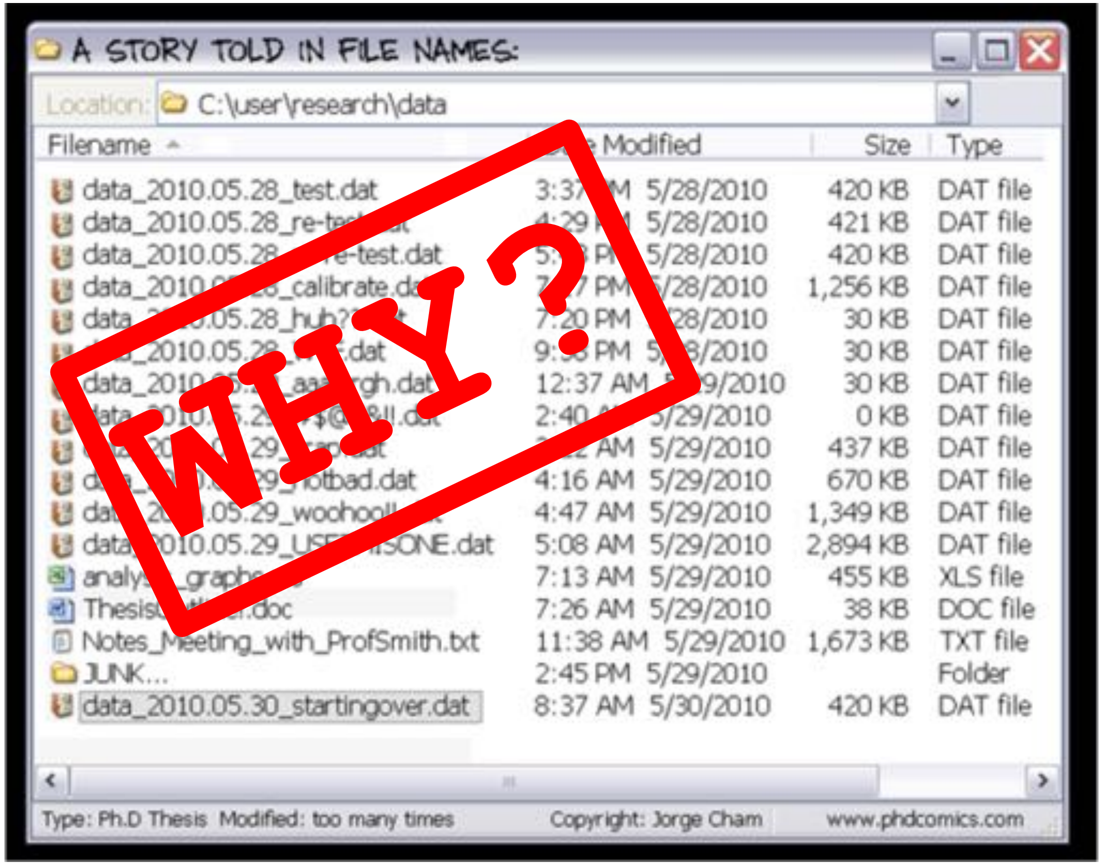
Location: C:\user\research\data

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

http://phdcomics.com/comics/archive.php?comicid=1323

Most research deals with data, so…



A STORY TOLD IN FILE NAMES:

Location: C:\user\research\data

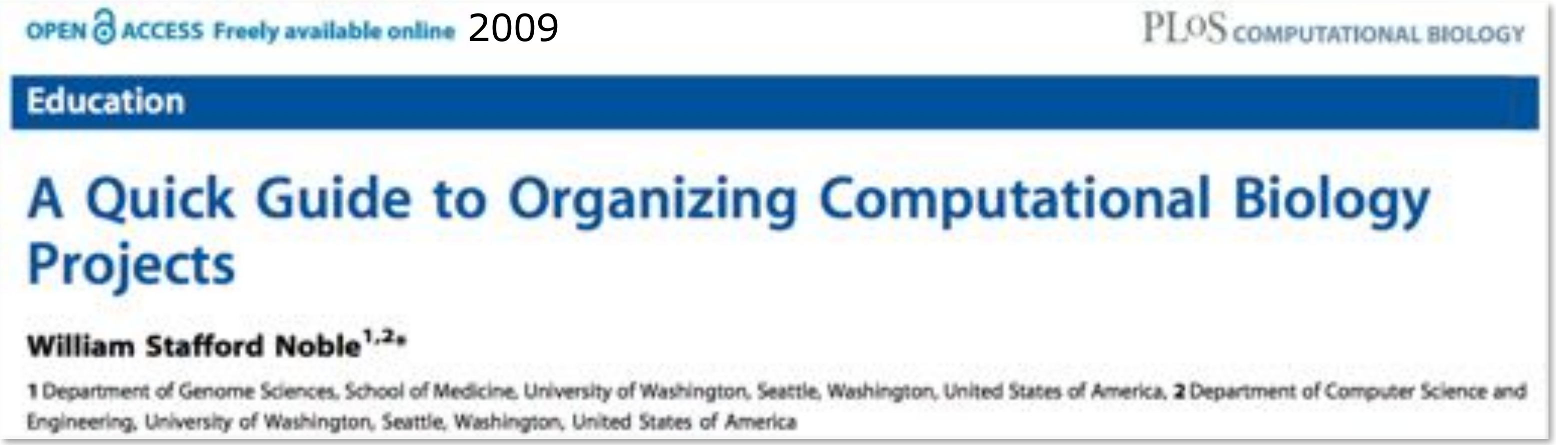
Filename	[illegible] Modified		Size	Type
data_2010.05.28_test.dat	3:37 [illegible]M	5/28/2010	420 KB	DAT file
data_2010.05.28_re-te[illegible]	4:29 [illegible]	5/28/2010	421 KB	DAT file
data_2010.05.28[illegible]e-test.dat	5:[illegible] P[illegible]	5/28/2010	420 KB	DAT file
data_2010[illegible]_calibrate.d[illegible]	7:[illegible]7 PM	[illegible]/28/2010	1,256 KB	DAT file
data[illegible].05.28_huh?[illegible]	7:20 PM	[illegible]28/2010	30 KB	DAT file
[illegible]_2010.05.2[illegible].dat	9:[illegible] PM	5[illegible]8/2010	30 KB	DAT file
data_2010[illegible]aa[illegible]gh.dat	12:37 AM	[illegible]9/2010	30 KB	DAT file
[illegible]ata[illegible]010[illegible]5.2[illegible]$@[illegible]&!!.d[illegible]	2:40 [illegible]	[illegible]/29/2010	0 KB	DAT file
[illegible]20[illegible].29[illegible]	[illegible] AM	5/29/2010	437 KB	DAT file
d[illegible].29_[illegible]otbad.dat	4:16 AM	5/29/2010	670 KB	DAT file
da[illegible]2[illegible].05.29_woohoo!![illegible]	4:47 AM	5/29/2010	1,349 KB	DAT file
data[illegible]010.05.29_U[illegible]SONE.dat	5:08 AM	5/29/2010	2,894 KB	DAT file
analy[illegible]_graph[illegible]	7:13 AM	5/29/2010	455 KB	XLS file
Thesis[illegible].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

http://phdcomics.com/comics/archive.php?comicid=1323

Something Essential You (most likely) Were Not Taught

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



2009

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Education

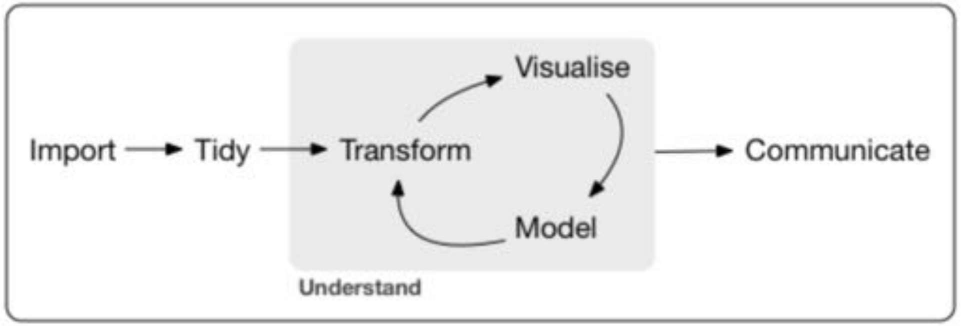
A Quick Guide to Organizing Computational Biology Projects

William Stafford Noble[1,2*]

[1] Department of Genome Sciences, School of Medicine, University of Washington, Seattle, Washington, United States of America, [2] Department of Computer Science and Engineering, University of Washington, Seattle, Washington, United States of America

The Elements of Data Analysis

- Import
- Tidy
- Transform
- Visualise
- Model
- Communicate



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

The Elements of Data Analysis

- Import
  - Import and combine your raw data from (potentially) multiple sources
- Tidy
  - Clean variables (e.g. missing data), setup observations as rows and variables as columns
- Transform
  - Compute new variables (augment the data), reduce data to desired focus
- Visualise
  - Explore and understand your data by seeing it, generate questions
- Model
  - Answer initial and generated questions, extract value, gain insight
- Communicate
  - Condense and communicate gained insight via essential, well defined and focused plots

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

The Elements of Data Analysis

- Import
  - Import and combine your raw data from (potentially) multiple sources
- Tidy
  - Clean variables (e.g. missing data), setup observations as rows and variables as columns
- Transform
  - Compute new variables (augment the data), reduce data to desired focus
- Visualise
  - Explore and understand your data by seeing it, generate questions
- Model
  - Answer initial and generated questions, extract value, gain insight
- Communicate
  - Condense and communicate gained insight via essential, well defined and focused plots

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

The Elements of Data Analysis

- Import
  - Import and combine your raw data from (potentially) multiple sources
- Tidy
  - Clean variables (e.g. missing data), setup observations as rows and variables as columns
- Transform
  - Compute new variables (augment the data), reduce data to desired focus
- Visualise
  - Explore and understand your data by seeing it, generate questions
- Model
  - Answer initial and generated questions, extract value, gain insight
- Communicate
  - Condense and communicate gained insight via essential, well defined and focused plots

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

The Elements of Data Analysis

- Import
  - Import and combine your raw data from (potentially) multiple sources
- Tidy
  - Clean variables (e.g. missing data), setup observations as rows and variables as columns
- Transform
  - Compute new variables (augment the data), reduce data to desired focus
- Visualise
  - Explore and understand your data by seeing it, generate questions
- Model
  - Answer initial and generated questions, extract value, gain insight
- Communicate
  - Condense and communicate gained insight via essential, well defined and focused plots

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

The Elements of Data Analysis

- Import
  - Import and combine your raw data from (potentially) multiple sources
- Tidy
  - Clean variables (e.g. missing data), setup observations as rows and variables as columns
- Transform
  - Compute new variables (augment the data), reduce data to desired focus
- Visualise
  - Explore and understand your data by seeing it, generate questions
- Model
  - Answer initial and generated questions, extract value, gain insight
- Communicate
  - Condense and communicate gained insight via essential, well defined and focused plots

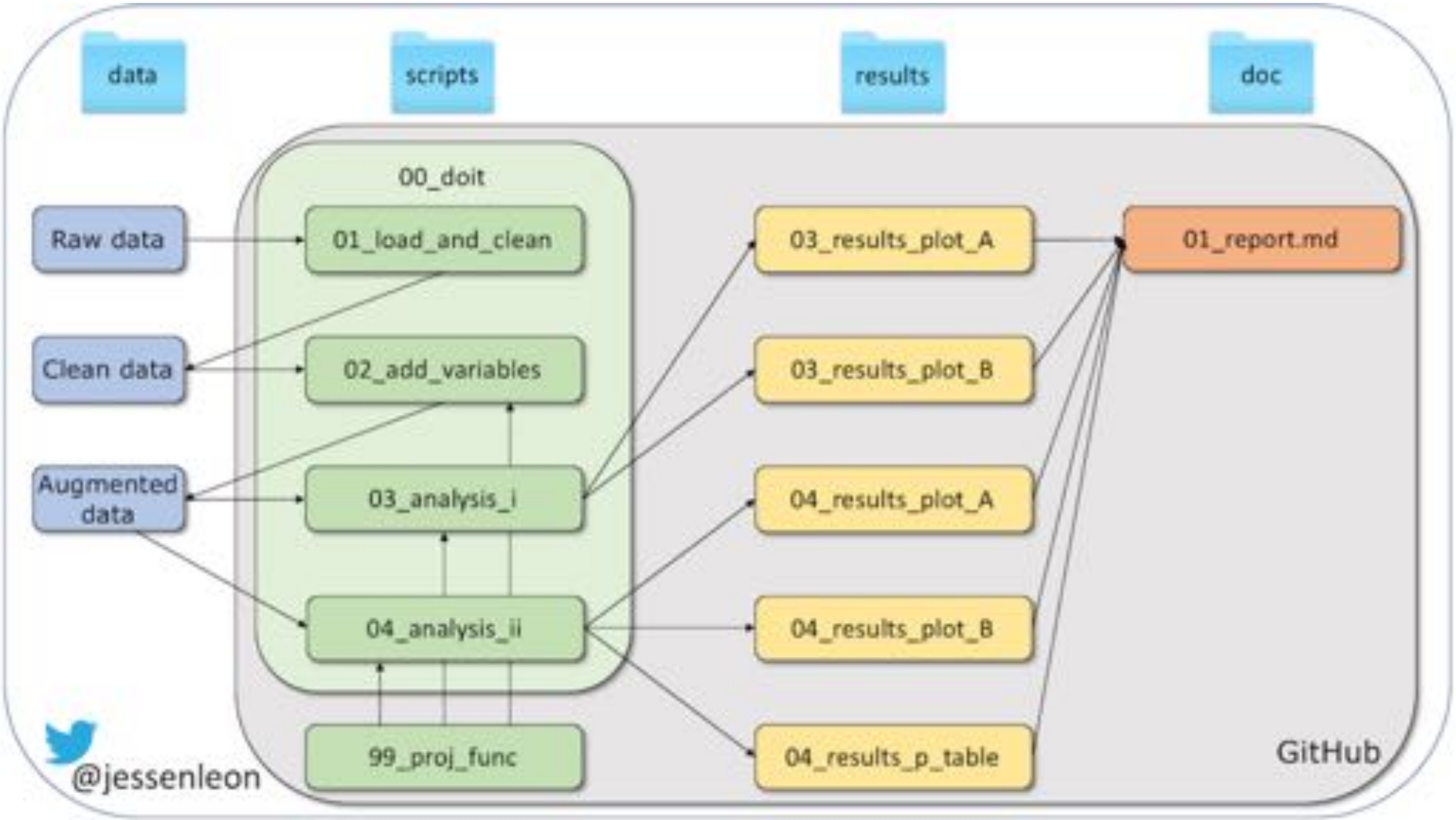
http://r4ds.had.co.nz/introduction.html

The Elements of Data Analysis

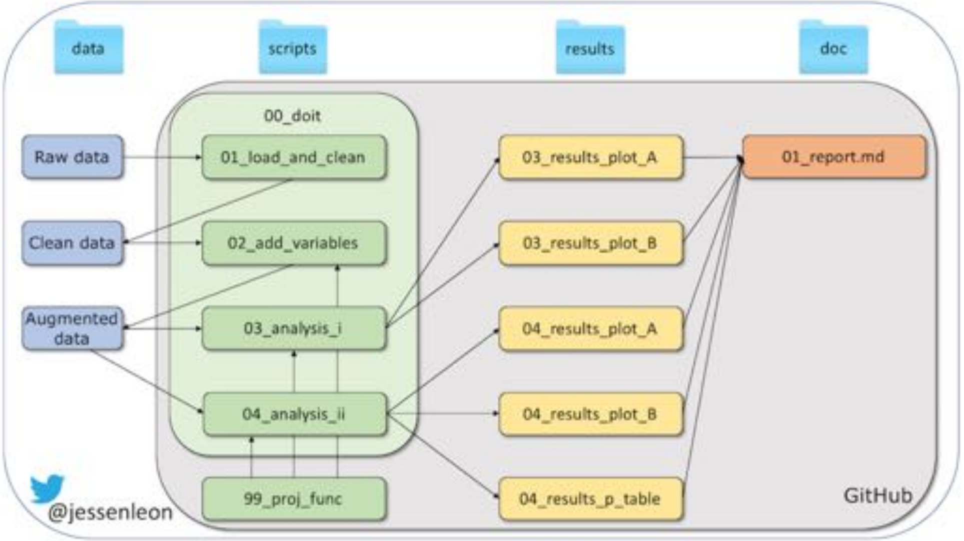
- Import
  - Import and combine your raw data from (potentially) multiple sources
- Tidy
  - Clean variables (e.g. missing data), setup observations as rows and variables as columns
- Transform
  - Compute new variables (augment the data), reduce data to desired focus
- Visualise
  - Explore and understand your data by seeing it, generate questions
- Model
  - Answer initial and generated questions, extract value, gain insight
- Communicate
  - Condense and communicate gained insight via essential, well defined and focused plots

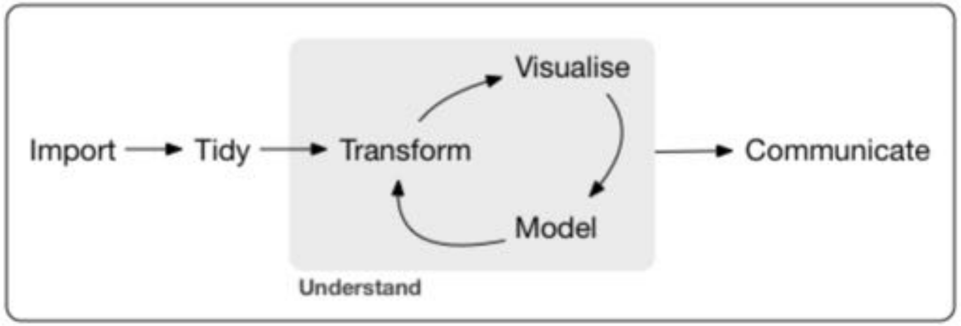
http://r4ds.had.co.nz/introduction.html

Organising Your Data Analysis Project

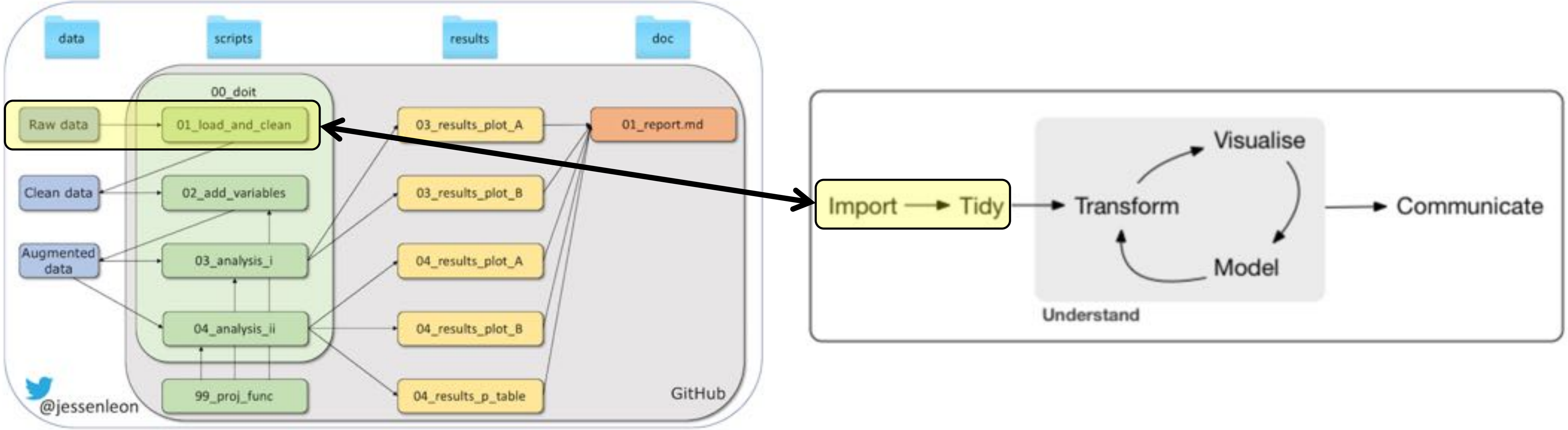


Combining Project Organisation with DA elements

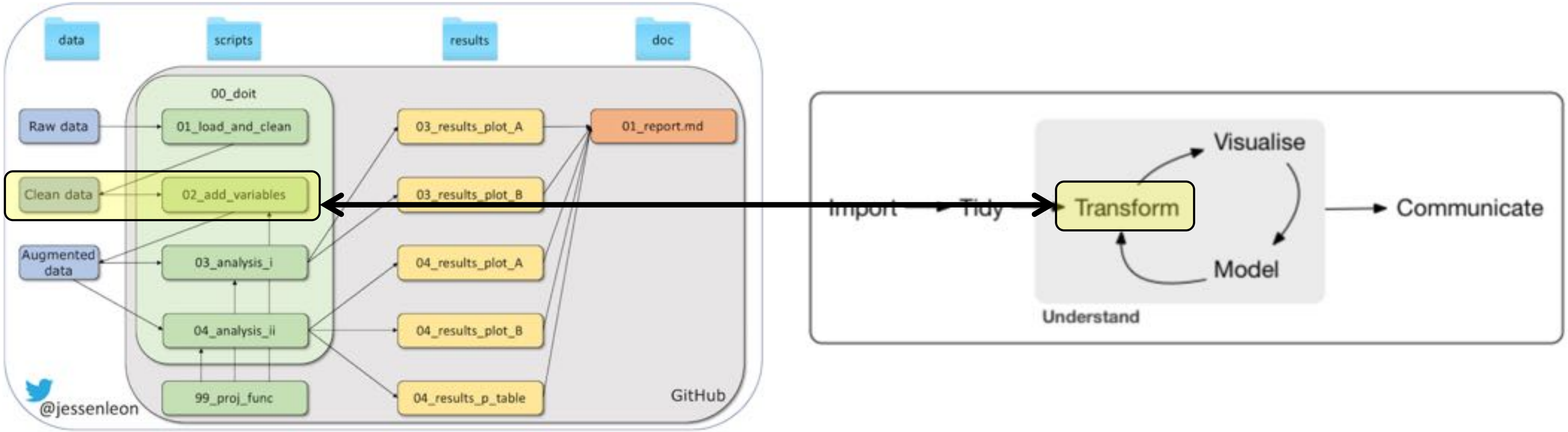




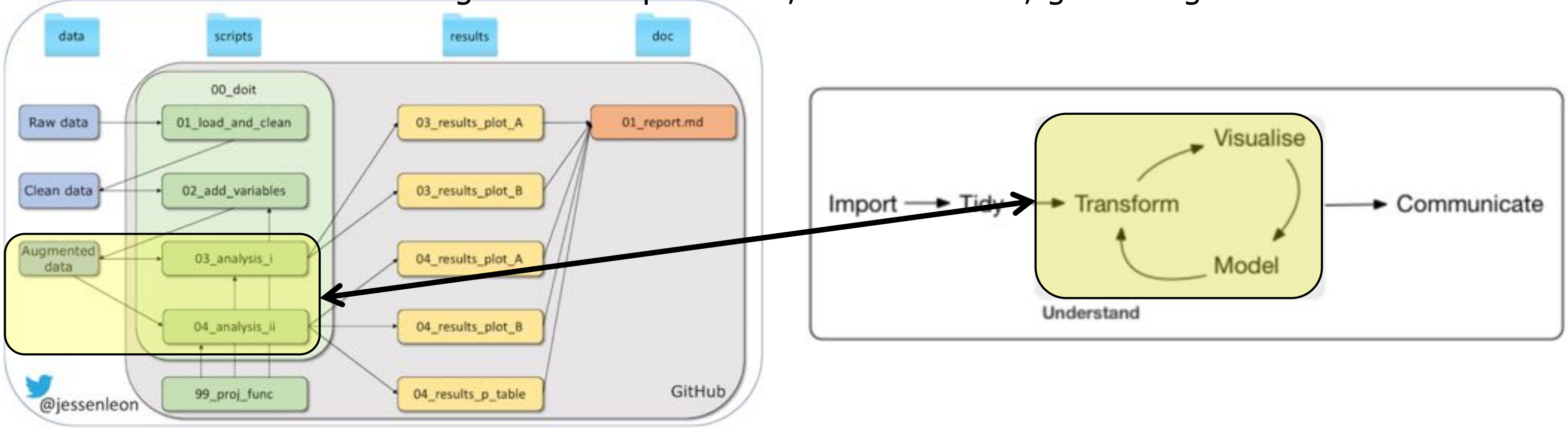
- Import: Import and combine your raw data from (potentially) multiple sources
- Tidy: Clean variables (eg. missing data), setup observations as rows and variables as columns



- Transform: Compute new variables (augment the data), reduce data to desired focus



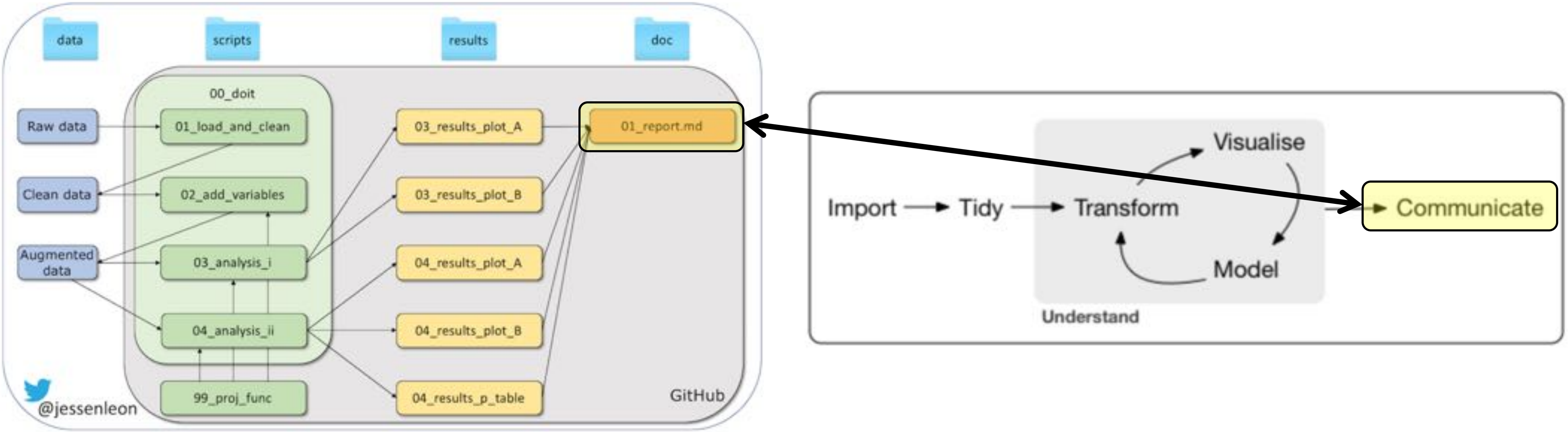
- Transform: Compute new variables (augment the data), reduce data to desired focus
- Visualise: Explore and understand your data by seeing it, generate questions
- Model: Answer initial and generated questions, extract value, gain insight



data
scripts
results
doc
00_doit
Raw data
01_load_and_clean
Clean data
02_add_variables
Augmented data
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
GitHub
@jessenleon

Import → Tidy → Transform → Visualise → Model → Communicate
Understand

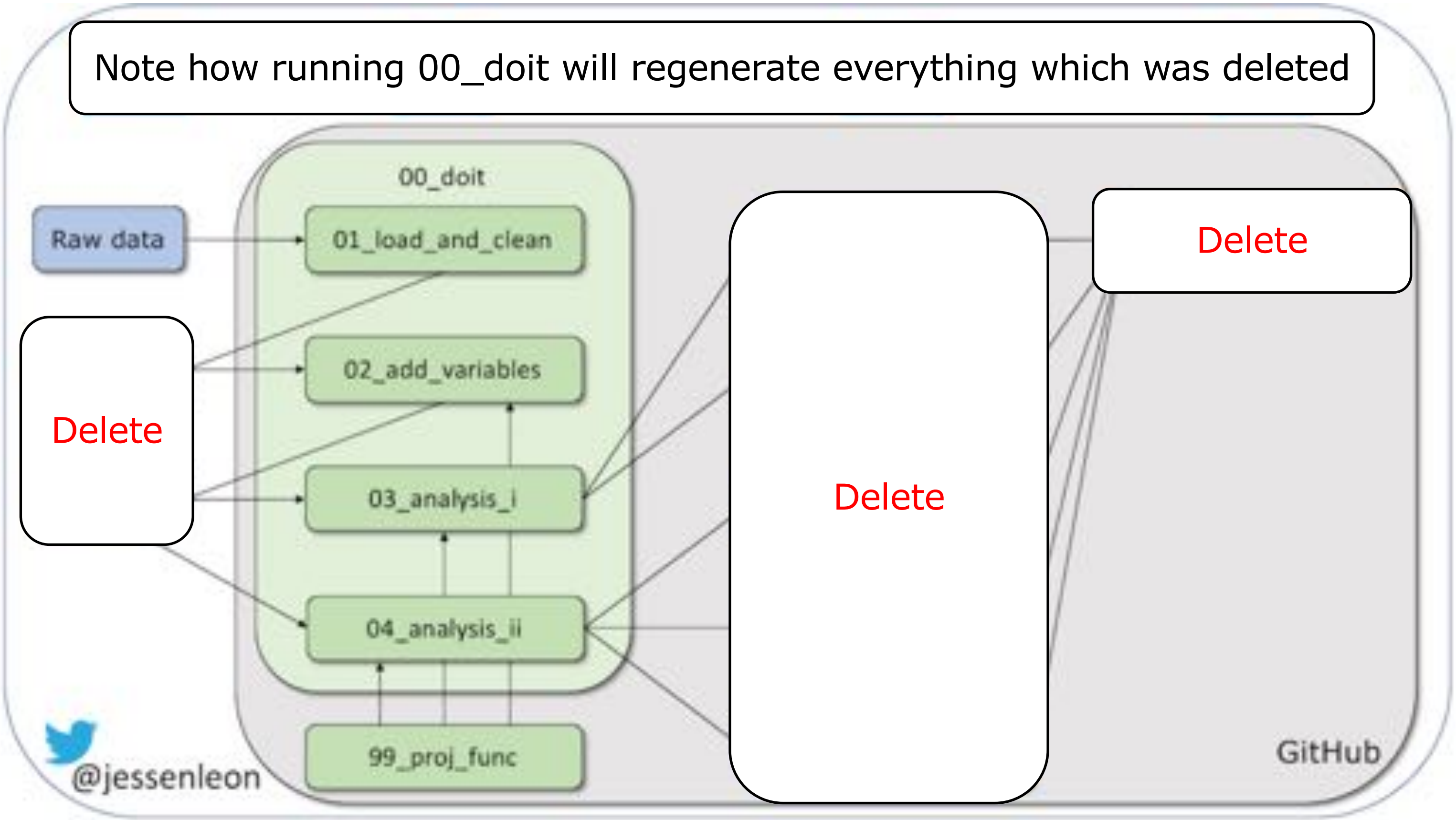
- Communicate: Condense and communicate gained insight via essential, well defined and focused plots



- Now, your paper is published, what then to do with all the files you generated?

Cleaning Up Your Data Analysis Project

Note how running 00_doit will regenerate everything which was deleted



Plenum

- *Take 1 minute to discuss with your neighbour*
  - *What are possible factors which may influence the reproducibility of a workflow, which we have not touched upon?*

Is this a reproducible workflow?

- When looking at a workflow, ask your self:
  - Can the entire workflow be run without manual intervention?
  - Is the workflow start data static or dynamic?
  - What are the dependencies of the workflow?

Summary - Learning Objectives Revisited

- Define what a reproducible data analysis workflow is
  - *“A reproducible data analysis workflow is when you can go from the raw data to recreating all the figures, tables and numbers in your paper automatically and consistently”*

- List the elements of a reproducible data analysis workflow
  - “Import → Tidy → Transform → Visualise → Model → Communicate”

- Explain the meaning and purpose of each of the elements in a reproducible data analysis workflow

- When presented with a pre-made workflow, determine if it constitutes a reproducible data analysis workflow
  - Run without manual intervention? Workflow start data static or dynamic? Workflow dependencies?

Reasons for lack of reproducibility

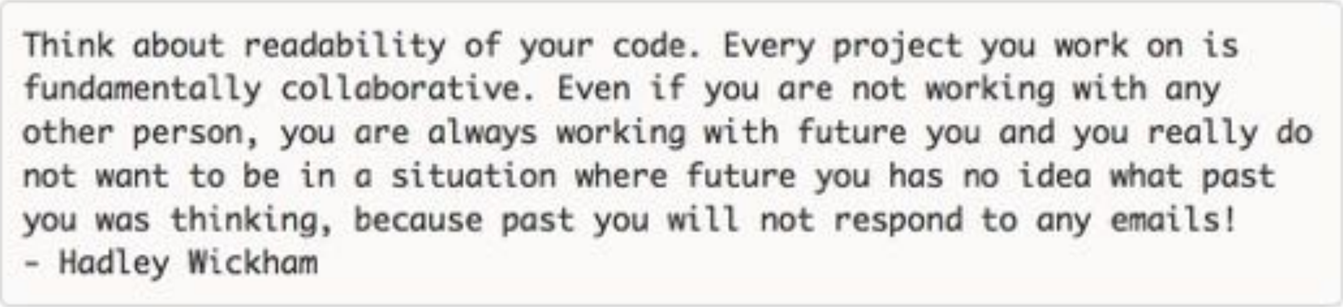
- Many!

- Many of which cannot be controlled

- Focus for this talk is on one aspect we can control

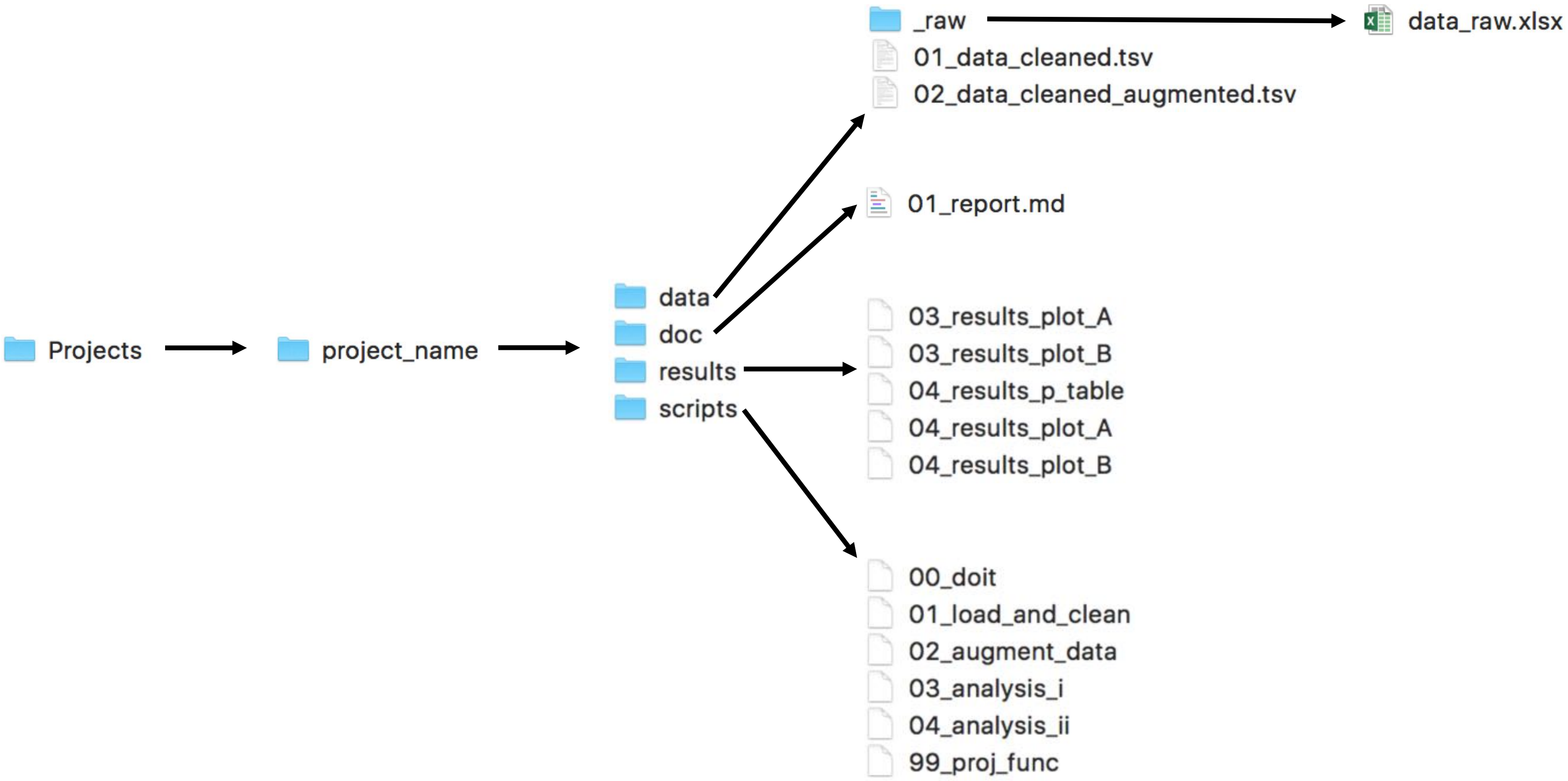
- Namely
  - Reproducible data analysis workflow

- Hopefully, now you have an idea about how you can control your data analysis workflow



> 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

Summary and open discussion



Projects

project_name

data

doc

results

scripts

_raw

01_data_cleaned.tsv

02_data_cleaned_augmented.tsv

data_raw.xlsx

01_report.md

03_results_plot_A

03_results_plot_B

04_results_p_table

04_results_plot_A

04_results_plot_B

00_doit

01_load_and_clean

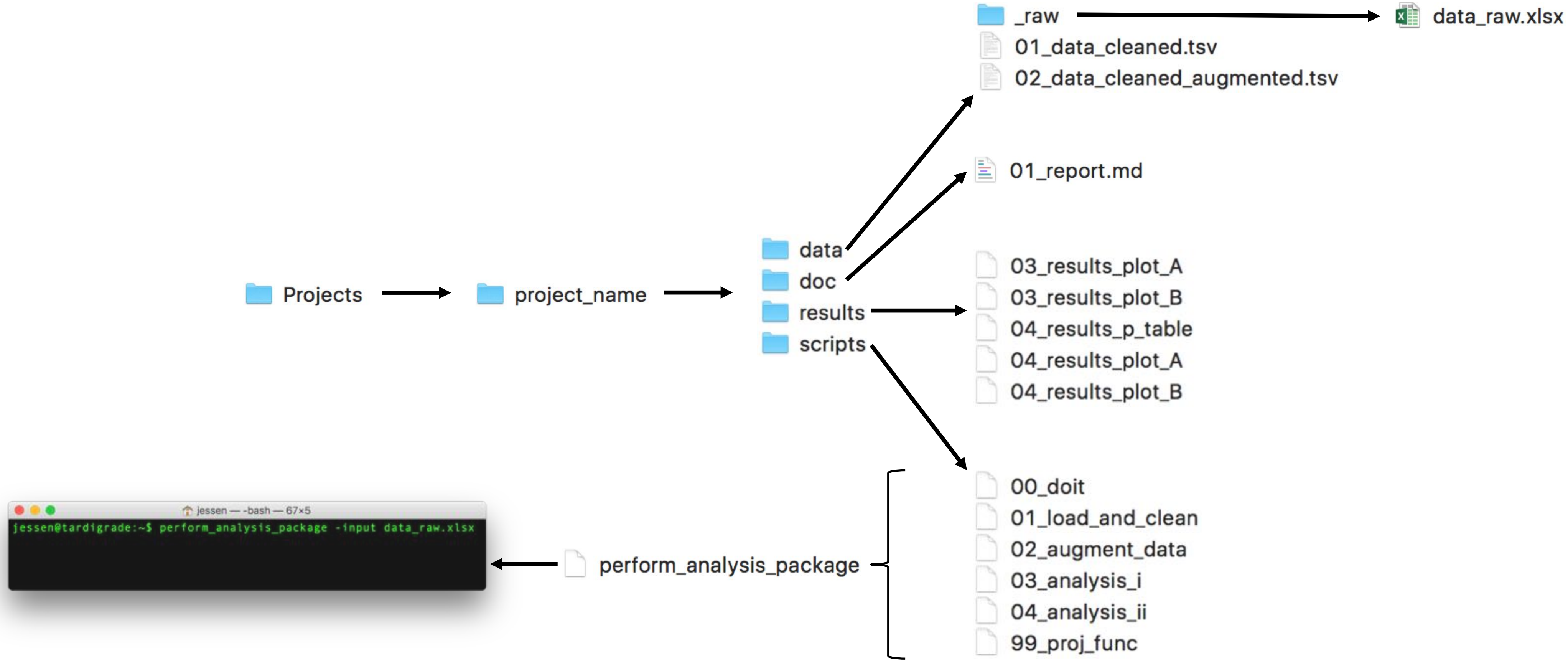
02_augment_data

03_analysis_i

04_analysis_ii

99_proj_func

Redo analysis across projects



```
jessen@tardigrade:~$ perform_analysis_package -input data_raw.xlsx
```