Reproducible Research

Automate Your Science

Robert M. Porsch

What is reproducibility and Why is it important?

Some general statistics

A recent (2016) survey by Nature found that:

- 70% of scientist fail to reproduce other scientist's experiments
- 50% fail to reproduce their own experiments
- 2% admitted to falsify study results
- 14% say they know at least one person who falsified study results

Replication Problems in Medicine

- FDA found flaws in 10-20% of studies (1977-90) ()
- in a sample of 49 studies)
  - 16% of studies investigating therapy effects were contradicted by later studies
  - 44% were replicated
  - 24% were never challenged

How can reproducibility help?

Some Steps you can do:

- Publish your data
- Publish your analysis script
- Enable others to build on to your data/analysis
- make your data citeable
- have a public repository

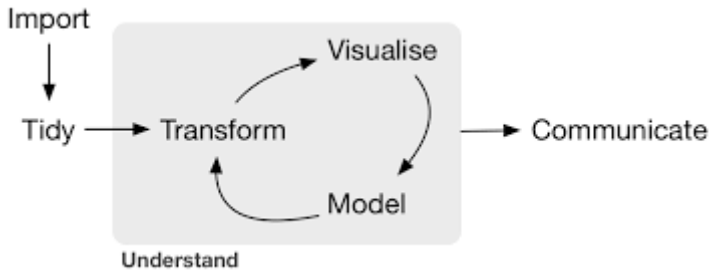
Some additional perks:

- you automate your whole analysis
- you automate your paper
- you add credibility and accountability

Make yourself as redundant as possible

Automation of Data Preparation and Analysis

A Pipeline


Import
Tidy
Transform
Visualise
Model
Communicate
Understand

Download and Formating

Step 1:

```
download.file("www.some_data.com/data.csv")
```

Step 2:

```
dat  <- read.csv("data.csv")
dat[,"Sum"] <- sum(dat[,1:2], na.rm=T)
write.csv(dat, "data_modified.csv")
```

Step 3:

```
dat <- read.csv("data_modified.csv")
analysis(dat)
```

- step 1 should be before step 2
- we should not repeat step 1 when we modify step 2

What is a Makefile?

- `make` is an application which can execute make files
- Makefiles contain collection of commands
- keeps track of the last time files have been updated
- it contains a list of rules

```
RULE: DEPENDENCY LINE
  ACTION LINE(S)
```

DEPENDENCY LINE:

TARGET FILES : SOURCE FILES

Makefiles

```
all: data.csv data_modidified.csv results.txt

clean:
  rm data.csv data_modidified.csv results.txt

data.csv:
  Rscript -e 'download.file("www.some_data.com/data.csv")'

data_modidified.csv: formating.R data.csv
  Rscript $<

results.txt: analysis.R data_modidified.csv
  Rscript $<
```

Some Advanced Makefiles Stuff

Define variables

```
data.csv: get_data.r
  R CMD BATCH get_data.r

RR := R CMD BATCH

data.csv: get_data.r
  $(RR) get_data.r
```

Some other operators

```
histogram.png: histogram.tsv
    Rscript -e 'library(ggplot2);
     qplot(Length, Freq, data=read.delim("$<")); ggsave("$@")'
```

Some more stuff

```
.PHONY: all clean
.DELETE_ON_ERROR:
.SECONDARY:
```

- `.PHONY` tells make that those are not files/directories
- `.DELETE_ON_ERROR` delete files of a target if it fails
- `.SECONDARY` prevents the deletion of intermediate files

Some more stuff

```
all: $(Main)
  @echo Its done!!!
```

Some other stuff

```
plotting.txt: formating.r function.r
  Rscript $^ $@
```

- `$^` right side of `:`
- `$@` left side of `:`

Dynamic Documents

What are dynamic documents

- documents which change based on the underlying code
- Written not in word but in "code"
- `markdown` or `latex`
- compiling of documents is necessary

Pro and Cons

Pro:

- no copy pasting
- less errors
- coherent pipeline
- version controlled

Con:

- harder to collaborate
- you don't "see" what you write
- some stuff is harder

What is rmarkdown

- an R flavoured version of markdown
- enables integration of R code (or a number of other languages into markdown)
- dynamic documents

Example