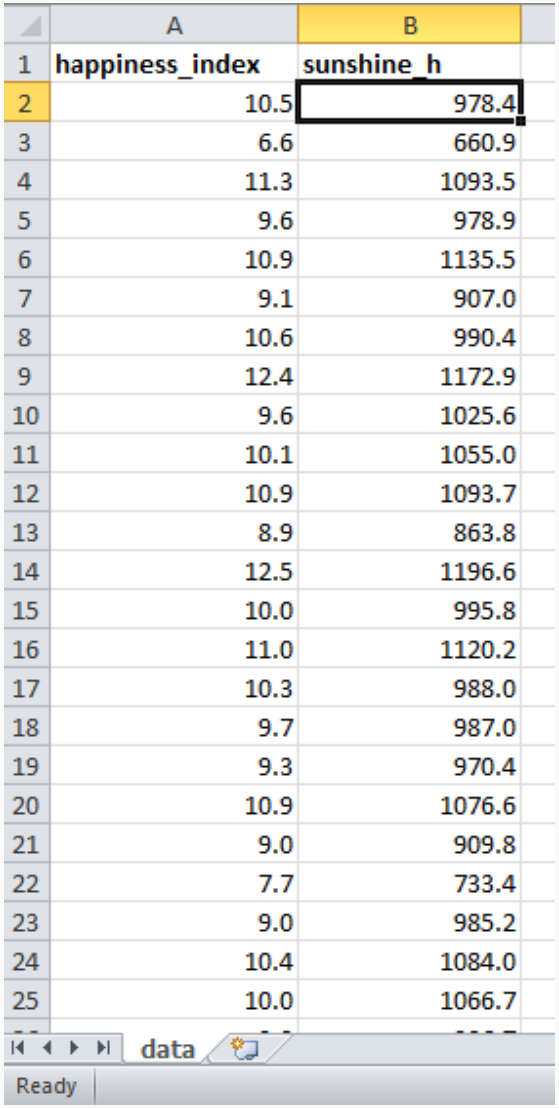
Doing reproducible science: from your hard-won data to a publishable manuscript without going mad

Francisco Rodriguez-Sanchez (@frod_san)

November 2016

A typical research workflow

1. Prepare data (**EXCEL**)



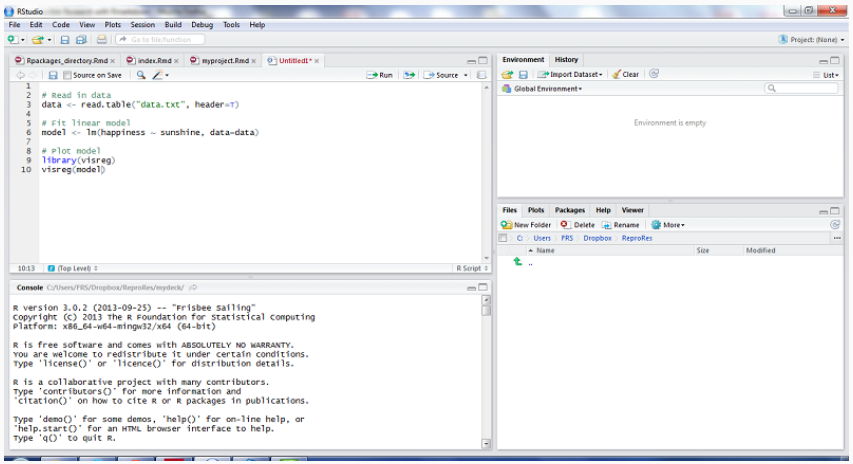
	A	B
1	happiness_index	sunshine_h
2	10.5	978.4
3	6.6	660.9
4	11.3	1093.5
5	9.6	978.9
6	10.9	1135.5
7	9.1	907.0
8	10.6	990.4
9	12.4	1172.9
10	9.6	1025.6
11	10.1	1055.0
12	10.9	1093.7
13	8.9	863.8
14	12.5	1196.6
15	10.0	995.8
16	11.0	1120.2
17	10.3	988.0
18	9.7	987.0
19	9.3	970.4
20	10.9	1076.6
21	9.0	909.8
22	7.7	733.4
23	9.0	985.2
24	10.4	1084.0
25	10.0	1066.7

data

Ready

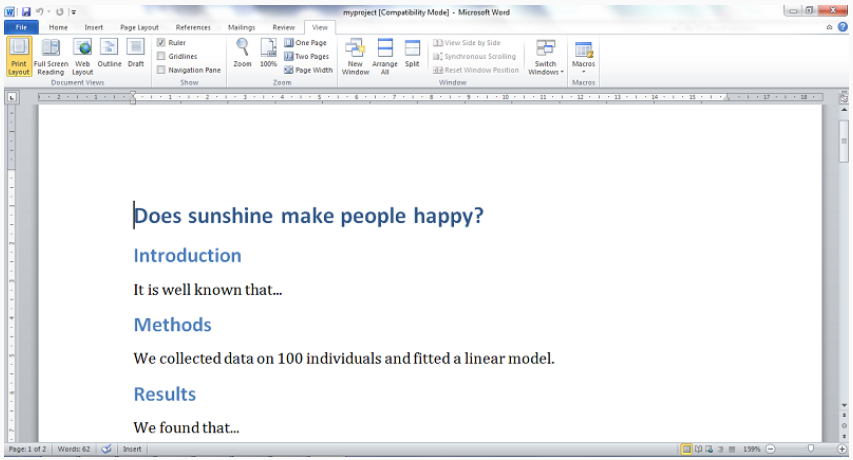
A typical research workflow

1. Prepare data (**EXCEL**)
2. Analyse data (**R**)



A typical research workflow

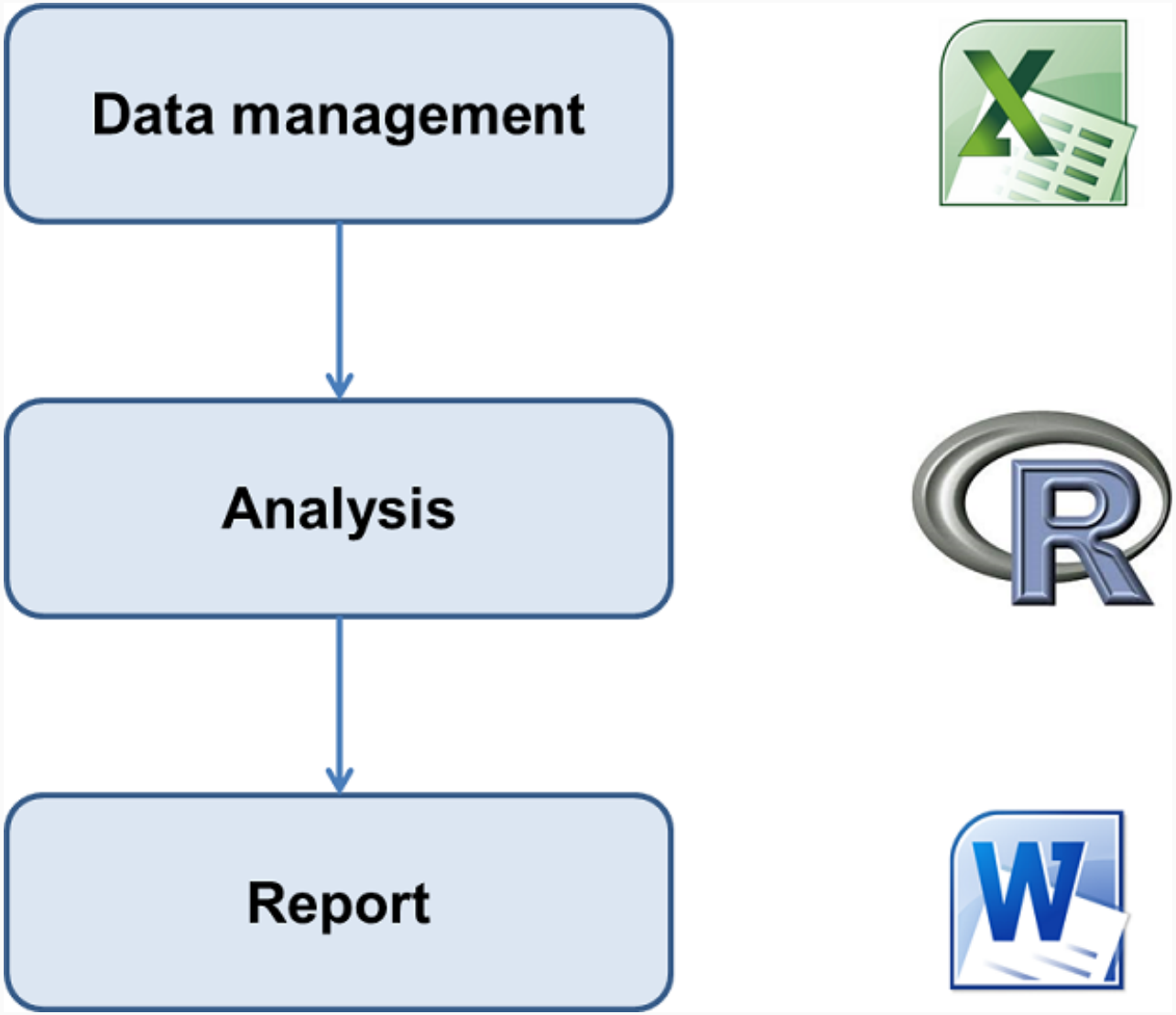
1. Prepare data (**EXCEL**)
2. Analyse data (**R**)
3. Write report/paper (**WORD**)



A typical research workflow

1. Prepare data (**EXCEL**)
2. Analyse data (**R**)
3. Write report/paper (**WORD**)
4. Start the email attachments nightmare...

This workflow is broken



Problems of a broken workflow

- How did you do this? What analysis is behind this figure? Did you account for . . . ?

Problems of a broken workflow

- How did you do this? What analysis is behind this figure? Did you account for . . . ?
- What dataset was used? Which individuals were left out? Where is the clean dataset?

Problems of a broken workflow

- How did you do this? What analysis is behind this figure? Did you account for . . . ?
- What dataset was used? Which individuals were left out? Where is the clean dataset?
- Oops, there is an error in the data. Can you repeat the analysis? And update figures/tables in Word!

Trevor A. Branch
@TrevorABranch

Follow

My rule of thumb: every analysis you do on a dataset will have to be redone 10–15 times before publication. Plan accordingly. #Rstats

Our everyday scary movie

https://youtu.be/s3JldKoA0zw

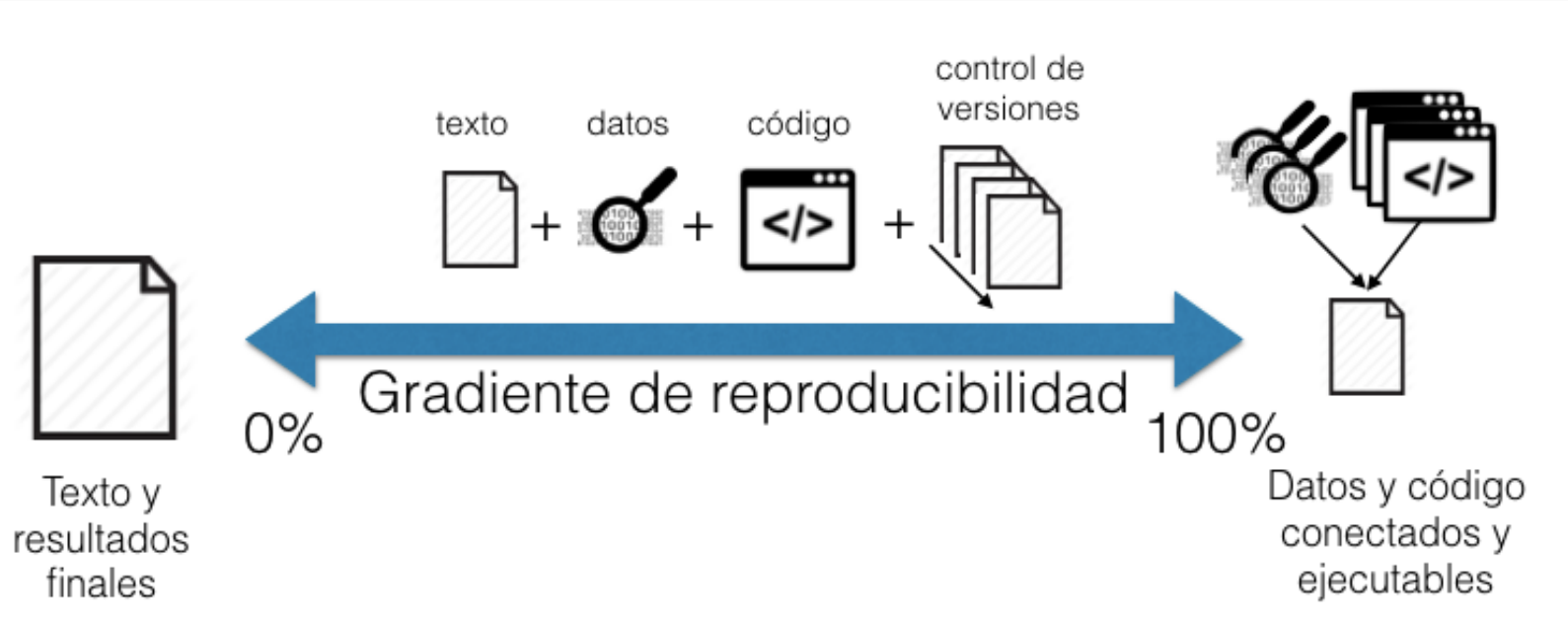
WHAT is Reproducible Science?

Reproducible Science: WHAT

A scientific article is **reproducible** if there is computer **code** that can **regenerate** all results and figures from the original data.

- Transparent
- Traceable
- Comprehensive
- Useful

Most science is not reproducible



Even **you** will struggle to reproduce **your own results** from a few weeks/months ago.

You can't reproduce if you don't understand where a number came from.
You can't reproduce what you don't remember. And trust me: you won't.
You can't reproduce what you've lost. What if you need access to a file as it existed 1, 10, 100, or 1000 days ago?

Ben Bond-Lamberty

WHY Reproducible Science?

Reproducible Science: WHY

- Fundamental pillar of scientific method

Reproducible Science: WHY

- Fundamental pillar of scientific method
- Much less prone to errors

Reproducible Science: WHY

- Fundamental pillar of scientific method
- Much less prone to errors
- Automatically regenerate results

Reproducible Science: WHY

- Fundamental pillar of scientific method
- Much less prone to errors
- Automatically regenerate results
- Code reuse & sharing accelerates scientific progress

Reproducible Science: WHY

- Fundamental pillar of scientific method
- Much less prone to errors
- Automatically regenerate results
- Code reuse & sharing accelerates scientific progress
- Increasingly required by journals

Reproducible Science: WHY

- Fundamental pillar of scientific method
- Much less prone to errors
- Automatically regenerate results
- Code reuse & sharing accelerates scientific progress
- Increasingly required by journals
- Higher publication impact (citations, future collaborations, etc)

HOW TO DO Reproducible Science?

Reproducible Science: HOW

1. File organisation.
2. Data management. Spreadsheet good practices.
3. Code-based data analysis. Rmarkdown
4. Software dependencies.
5. Version control & collaborative writing.

File organisation

- All files in same directory (Rstudio project).

File organisation

- All files in same directory (Rstudio project).
- Raw data untouched in independent folder.

File organisation

- All files in same directory (Rstudio project).
- Raw data untouched in independent folder.
- Derived, clean data in another folder.

File organisation

- All files in same directory (Rstudio project).
- Raw data untouched in independent folder.
- Derived, clean data in another folder.
- Figures, code, etc also have their own folder.

File organisation example

```
myproject

|- README        # general info about the project

|- analysis.R    # master script that executes everything

|- data-raw/     # original raw data

|- data/         # clean data (produced w/ script)

|- R/            # functions definitions

|- doc/          # manuscript files

|- figs/         # final figures

|- output/       # other code output
```

Data management

Data management

1. Planification (e.g. DMPTool)
2. Collection
3. Metadata description (EML, Morpho)
4. Quality control
5. Storage

Storage

Use the cloud: safe, persistent, easy to share

- Dropbox
- OSF
- Figshare, etc
- See all data repositories in www.re3data.org

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- Avoid spaces, numbers, and special characters in column names.

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- Avoid spaces, numbers, and special characters in column names.
- Always write zero values, to distinguish from blank/missing data.

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- Avoid spaces, numbers, and special characters in column names.
- Always write zero values, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- Avoid spaces, numbers, and special characters in column names.
- Always write zero values, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.
- Input dates as year, month, day in separate columns. Or YYYY-MM-DD as text.

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- Avoid spaces, numbers, and special characters in column names.
- Always write zero values, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.
- Input dates as year, month, day in separate columns. Or YYYY-MM-DD as text.
- Use 'Data validation' in Excel to constrain data entry to accepted values.

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- Avoid spaces, numbers, and special characters in column names.
- Always write zero values, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.
- Input dates as year, month, day in separate columns. Or YYYY-MM-DD as text.
- Use 'Data validation' in Excel to constrain data entry to accepted values.
- Don't combine multiple pieces of information in one cell.

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- Avoid spaces, numbers, and special characters in column names.
- Always write zero values, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.
- Input dates as year, month, day in separate columns. Or YYYY-MM-DD as text.
- Use 'Data validation' in Excel to constrain data entry to accepted values.
- Don't combine multiple pieces of information in one cell.
- Don't touch raw data. Do all data manipulation with R code.

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- Avoid spaces, numbers, and special characters in column names.
- Always write zero values, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.
- Input dates as year, month, day in separate columns. Or YYYY-MM-DD as text.
- Use 'Data validation' in Excel to constrain data entry to accepted values.
- Don't combine multiple pieces of information in one cell.
- Don't touch raw data. Do all data manipulation with R code.
- Export data as plain text (txt, csv)

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- Avoid spaces, numbers, and special characters in column names.
- Always write zero values, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.
- Input dates as year, month, day in separate columns. Or YYYY-MM-DD as text.
- Use 'Data validation' in Excel to constrain data entry to accepted values.
- Don't combine multiple pieces of information in one cell.
- Don't touch raw data. Do all data manipulation with R code.
- Export data as plain text (txt, csv)
- http://www.datacarpentry.org/spreadsheet-ecology-lesson/

Spreadsheet good practices

- Put **variables** in **columns** (things you are measuring: height, weight, sex)
- Each **observation** in one **row** (e.g. individuals).
- Avoid spaces, numbers, and special characters in column names.
- Always write zero values, to distinguish from blank/missing data.
- Use blank/empty cells, or NA, for missing data.
- Input dates as year, month, day in separate columns. Or YYYY-MM-DD as text.
- Use 'Data validation' in Excel to constrain data entry to accepted values.
- Don't combine multiple pieces of information in one cell.
- Don't touch raw data. Do all data manipulation with R code.
- Export data as plain text (txt, csv)
- http://www.datacarpentry.org/spreadsheet-ecology-lesson/
- http://kbroman.org/dataorg/

Common spreadsheet errors

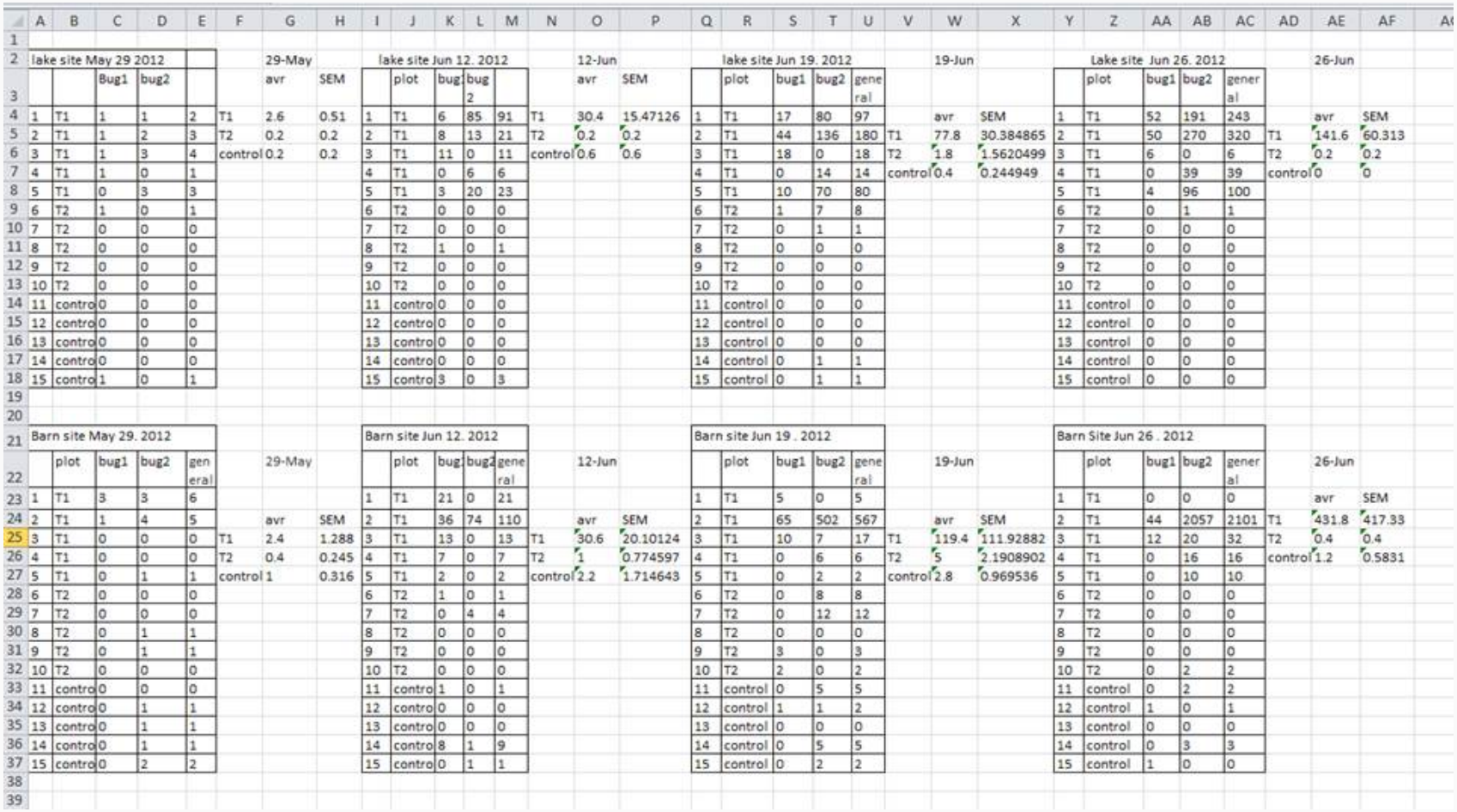
More than one variable per column

Date collected	Plot	Species-Sex	Weight
1/9/78	1	DM-M	40
1/9/78	1	DM-F	36
1/9/78	1	DS-F	135
1/20/78	1	DM-F	39
1/20/78	2	DM-M	43
1/20/78	2	DS-F	144
3/13/78	2	DM-F	51
3/13/78	2	DM-F	44
3/13/78	2	DS-F	146

Source: Data Carpentry

Date collected	Plot	Species	Sex	Weight
1/9/78	1	DM	M	40
1/9/78	1	DM	F	36
1/9/78	1	DS	F	135
1/20/78	1	DM	F	39
1/20/78	2	DM	M	43
1/20/78	2	DS	F	144
3/13/78	2	DM	F	51
3/13/78	2	DM	F	44
3/13/78	2	DS	F	146

Multiple tables



lake site May 29 2012						29-May	
		Bug1	bug2			avr	SEM
1	T1	1	1	2	T1	2.6	0.51
2	T1	1	2	3	T2	0.2	0.2
3	T1	1	3	4	control	0.2	0.2
4	T1	1	0	1			
5	T1	0	3	3			
6	T2	1	0	1			
7	T2	0	0	0			
8	T2	0	0	0			
9	T2	0	0	0			
10	T2	0	0	0			
11	contro	0	0	0			
12	contro	0	0	0			
13	contro	0	0	0			
14	contro	0	0	0			
15	contro	1	0	1			

lake site Jun 12. 2012						12-Jun	
	plot	bug1	bug2			avr	SEM
1	T1	6	85	91	T1	30.4	15.47126
2	T1	8	13	21	T2	0.2	0.2
3	T1	11	0	11	control	0.6	0.6
4	T1	0	6	6			
5	T1	3	20	23			
6	T2	0	0	0			
7	T2	0	0	0			
8	T2	1	0	1			
9	T2	0	0	0			
10	T2	0	0	0			
11	contro	0	0	0			
12	contro	0	0	0			
13	contro	0	0	0			
14	contro	0	0	0			
15	contro	3	0	3			

lake site Jun 19. 2012						19-Jun	
	plot	bug1	bug2	general		avr	SEM
1	T1	17	80	97			
2	T1	44	136	180	T1	77.8	30.384865
3	T1	18	0	18	T2	1.8	1.5620499
4	T1	0	14	14	control	0.4	0.244949
5	T1	10	70	80			
6	T2	1	7	8			
7	T2	0	1	1			
8	T2	0	0	0			
9	T2	0	0	0			
10	T2	0	0	0			
11	control	0	0	0			
12	control	0	0	0			
13	control	0	0	0			
14	control	0	1	1			
15	control	0	1	1			

Lake site Jun 26. 2012						26-Jun	
	plot	bug1	bug2	general		avr	SEM
1	T1	52	191	243			
2	T1	50	270	320	T1	141.6	60.313
3	T1	6	0	6	T2	0.2	0.2
4	T1	0	39	39	control	0	0
5	T1	4	96	100			
6	T2	0	1	1			
7	T2	0	0	0			
8	T2	0	0	0			
9	T2	0	0	0			
10	T2	0	0	0			
11	control	0	0	0			
12	control	0	0	0			
13	control	0	0	0			
14	control	0	0	0			
15	control	0	0	0			

Barn site May 29. 2012						29-May	
	plot	bug1	bug2	general		avr	SEM
1	T1	3	3	6			
2	T1	1	4	5		avr	SEM
3	T1	0	0	0	T1	2.4	1.288
4	T1	0	0	0	T2	0.4	0.245
5	T1	0	1	1	control	1	0.316
6	T2	0	0	0			
7	T2	0	0	0			
8	T2	0	1	1			
9	T2	0	1	1			
10	T2	0	0	0			
11	contro	0	0	0			
12	contro	0	1	1			
13	contro	0	1	1			
14	contro	0	1	1			
15	contro	0	2	2			

Barn site Jun 12. 2012						12-Jun	
	plot	bug1	bug2	general		avr	SEM
1	T1	21	0	21			
2	T1	36	74	110		avr	SEM
3	T1	13	0	13	T1	30.6	20.10124
4	T1	7	0	7	T2	1	0.774597
5	T1	2	0	2	control	2.2	1.714643
6	T2	1	0	1			
7	T2	0	4	4			
8	T2	0	0	0			
9	T2	0	0	0			
10	T2	0	0	0			
11	contro	1	0	1			
12	contro	0	0	0			
13	contro	0	0	0			
14	contro	8	1	9			
15	contro	0	1	1			

Barn site Jun 19 . 2012						19-Jun	
	plot	bug1	bug2	general		avr	SEM
1	T1	5	0	5			
2	T1	65	502	567		avr	SEM
3	T1	10	7	17	T1	119.4	111.92882
4	T1	0	6	6	T2	5	2.1908902
5	T1	0	2	2	control	2.8	0.969536
6	T2	0	8	8			
7	T2	0	12	12			
8	T2	0	0	0			
9	T2	3	0	3			
10	T2	2	0	2			
11	control	0	5	5			
12	control	1	1	2			
13	control	0	0	0			
14	control	0	5	5			
15	control	0	2	2			

Barn Site Jun 26 . 2012						26-Jun	
	plot	bug1	bug2	general		avr	SEM
1	T1	0	0	0			
2	T1	44	2057	2101	T1	431.8	417.33
3	T1	12	20	32	T2	0.4	0.4
4	T1	0	16	16	control	1.2	0.5831
5	T1	0	10	10			
6	T2	0	0	0			
7	T2	0	0	0			
8	T2	0	0	0			
9	T2	0	0	0			
10	T2	0	2	2			
11	control	0	2	2			
12	control	1	0	1			
13	control	0	0	0			
14	control	0	3	3			
15	control	1	0	0			

Multiple tabs

Could you avoid new tab by adding a column to original spreadsheet?

Using formatting, comments, etc to convey information

Plot: 2			
Date collecte	Species	Sex	Weight
1/8/14	NA		
1/8/14	DM	M	44
1/8/14	DM	M	38
1/8/14	OL		
1/8/14	PE	M	22
1/8/14	DM	M	38
1/8/14	DM	M	48
1/8/14	DM	M	43
1/8/14	DM	F	35
1/8/14	DM	M	43
1/8/14	DM	F	37
1/8/14	PF	F	7
1/8/14	DM	M	45
1/8/14	OT		
1/8/14	DS	M	157
1/8/14	OX		
2/18/14	NA	M	218
2/18/14	PF	F	7
2/18/14	DM	M	52

measurement device not calibrated

Plot: 2				
Date collecte	Species	Sex	Weight	Calibrated
1/8/14	NA			
1/8/14	DM	M	44	Y
1/8/14	DM	M	38	Y
1/8/14	OL			
1/8/14	PE	M	22	Y
1/8/14	DM	M	38	Y
1/8/14	DM	M	48	Y
1/8/14	DM	M	43	Y
1/8/14	DM	F	35	Y
1/8/14	DM	M	43	Y
1/8/14	DM	F	37	Y
1/8/14	PF	F	7	Y
1/8/14	DM	M	45	Y
1/8/14	OT			
1/8/14	DS	M	157	N
1/8/14	OX			
2/18/14	NA	M	218	N
2/18/14	PF	F	7	Y
2/18/14	DM	M	52	Y

Your turn: tidy up this messy dataset

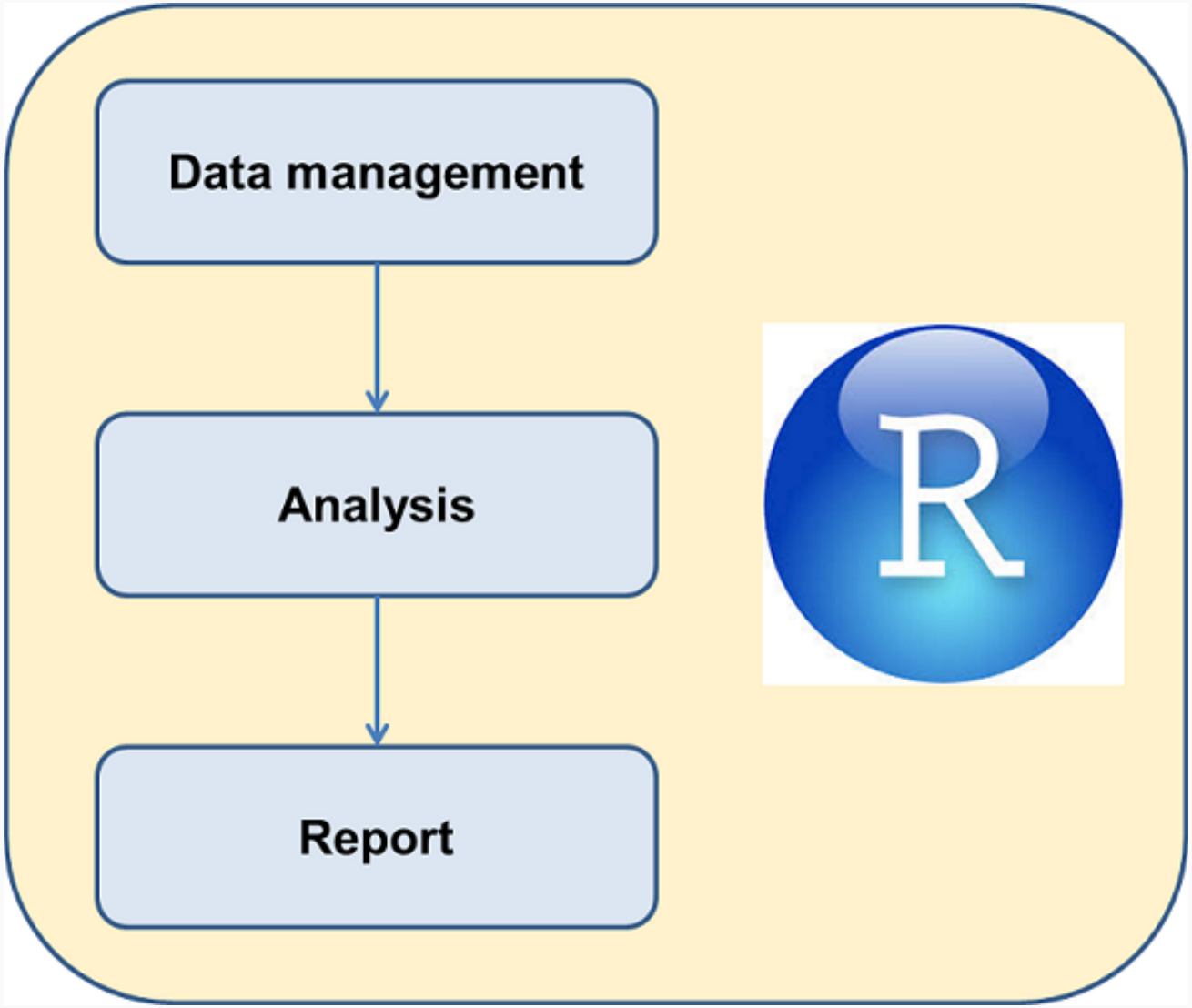
https://ndownloader.figshare.com/files/2252083

Data analysis

Always use code

- Reproducible
- Reusable

Dynamic reports



Rmarkdown documents

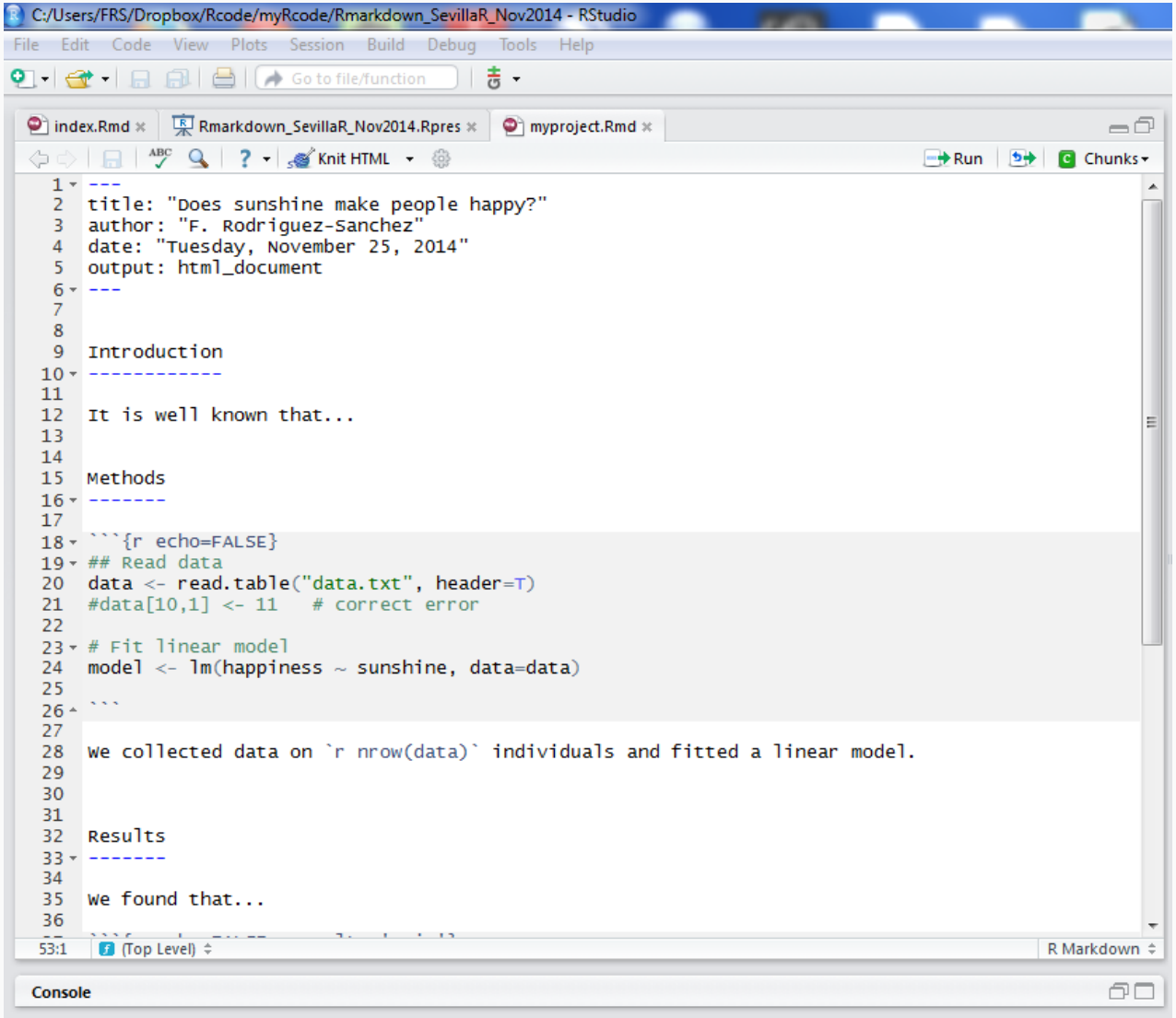
- Fully reproducible (trace all results inc. tables and plots)
- Dynamic (regenerate with 1 click)
- Suitable for
  - documents (Word, PDF, etc)
  - presentations
  - books
  - websites
  - ...

Let's see Rmarkdown in action

In Rstudio, create new Rmarkdown document and click on `Knit HTML`.

Example: Does sunshine influence happiness?

See `myproject.Rmd`



HTML output

Does sunshine make people happy?

F. Rodriguez-Sanchez

Tuesday, November 25, 2014

Introduction

It is well known that individual well-being can be influenced by climatic conditions. However, ...

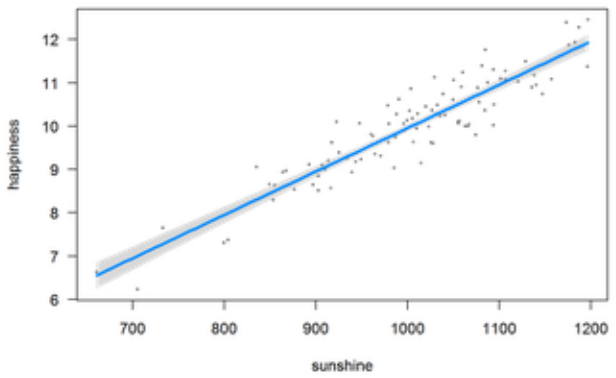
Methods

We collected data on 100 individuals and fitted a linear model.

Results

We found that...

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0651657	0.4264970	-0.1527928	0.8788758
sunshine	0.0100228	0.0004232	23.6833264	0.0000000



Discussion

These results confirm that sunshine is good for happiness (slope = 0.0100228).

Acknowledgements

Y. Xie, J. MacFarlane, Rstudio...

Spotted error in the data? No problem!

Make changes in Rmarkdown document, click `knit` and report will update automatically!

Other formats: PDF, Word

Does sunshine make people happy?

F. Rodrigues-Sanchez

Tuesday, November 25, 2014

Introduction

It is well known that individual well-being can be influenced by climatic conditions. However, ...

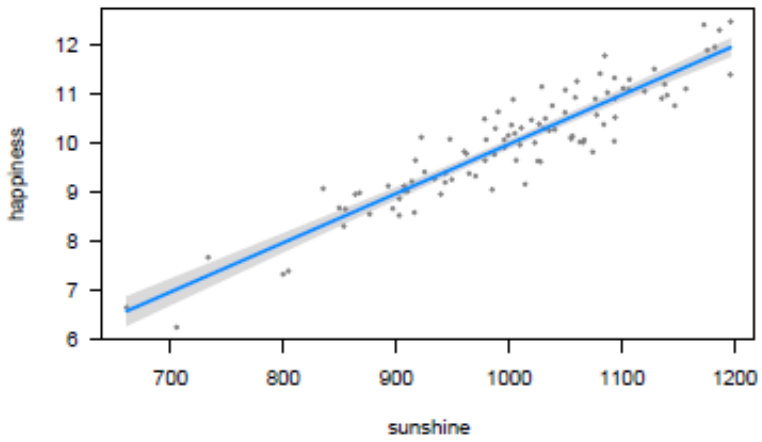
Methods

We collected data on 100 individuals and fitted a linear model.

Results

We found that...

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0651657	0.4264970	-0.1527928	0.8788758
sunshine	0.0100228	0.0004232	23.6833264	0.0000000



Does sunshine make people happy?

F. Rodriguez-Sanchez

Tuesday, November 25, 2014

Introduction

It is well known that individual well-being can be influenced by climatic conditions. However, ...

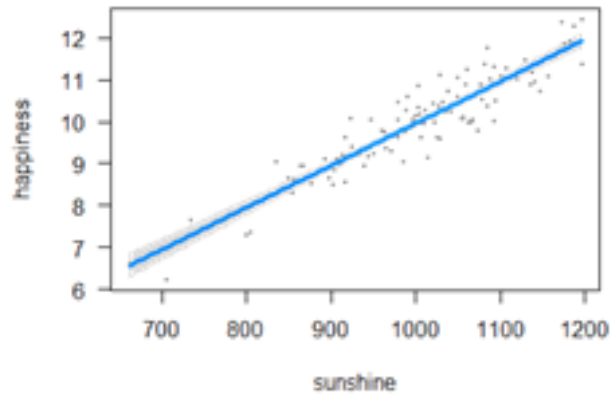
Methods

We collected data on 100 individuals and fitted a linear model.

Results

We found that...

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0651657	0.4264970	-0.1527928	0.8788758
sunshine	0.0100228	0.0004232	23.6833264	0.0000000



Adding citations

```
bibliography: references.bib
---

```{r citations_setup, include=FALSE}

library(knitcitations)
cite_options(citation_format = "pandoc")

```

Introduction

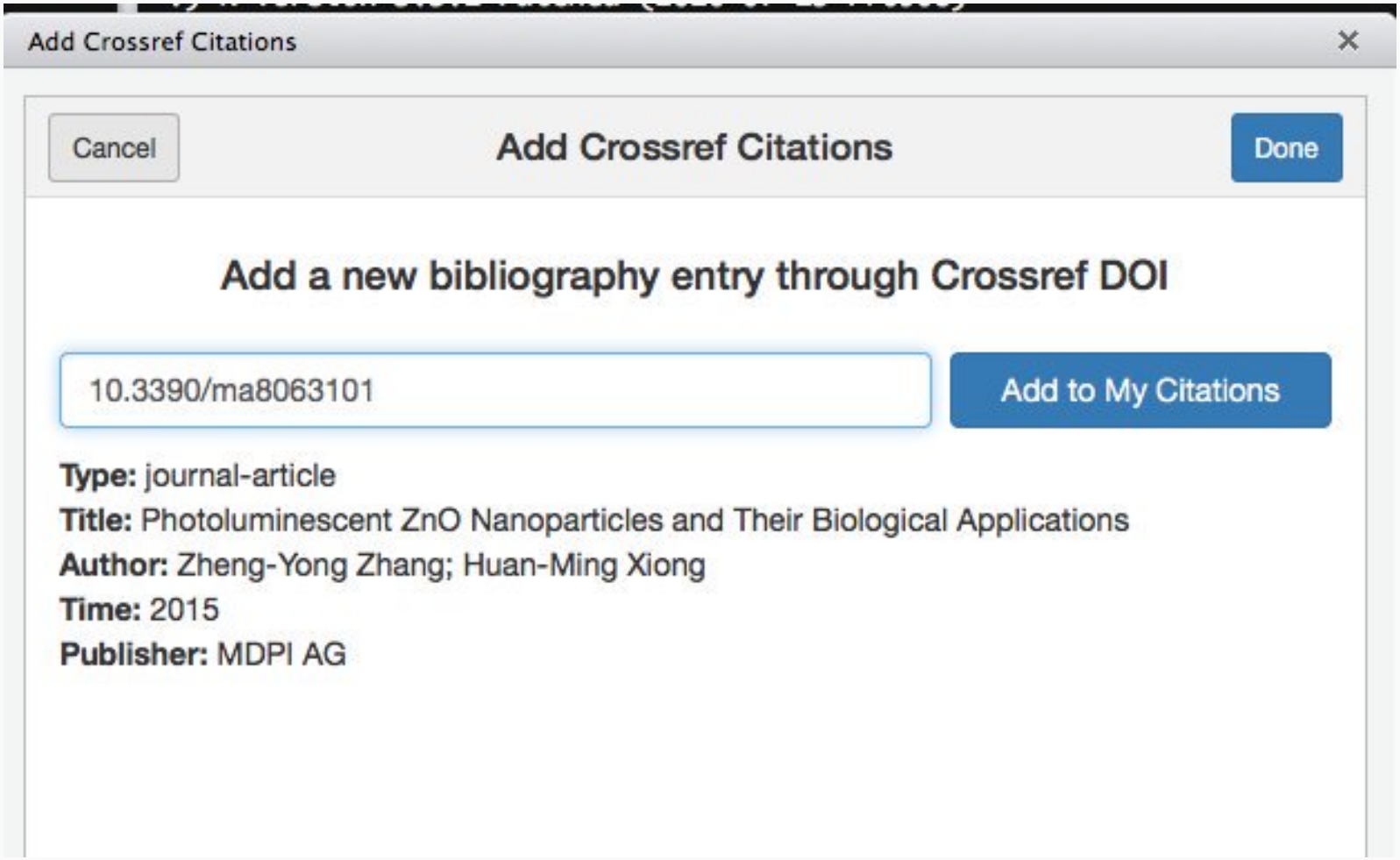
It is well known that individual well-being can be influenced by climatic conditions `r citep("10.1016/j.ecolecon.2004.06.015")`. However, ...
```

See output

References

Rehdanz, Katrin, and David Maddison. 2005. "Climate and Happiness." *Ecological Economics* 52 (1). Elsevier BV: 111–25. doi:10.1016/j.ecolecon.2004.06.015.

Useful: rcrossref addin

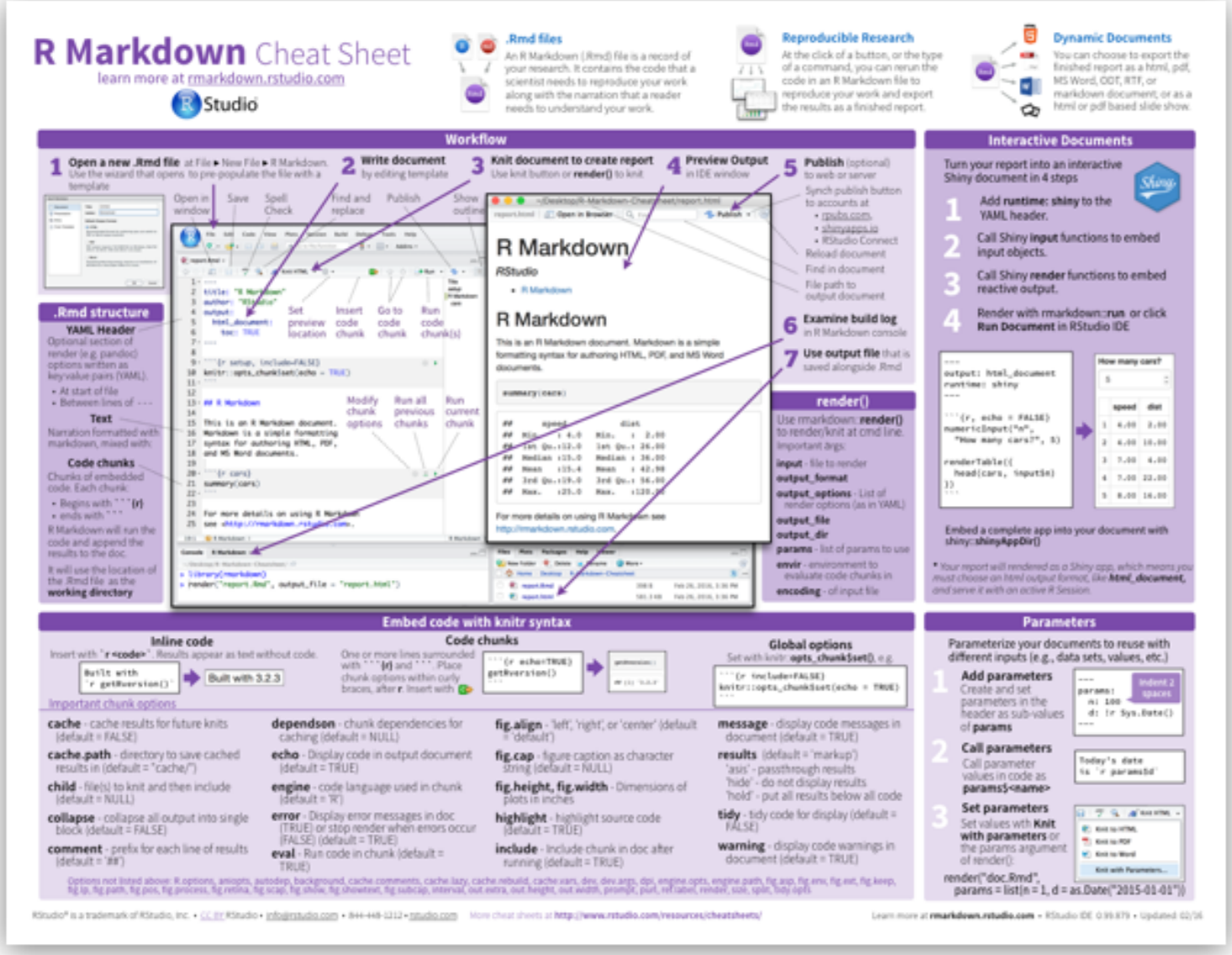


Can write full thesis in Rmarkdown!

See `thesis.Rmd`.

See `thesis.pdf`.

Rmarkdown cheat sheet



https://www.rstudio.com/wp-content/uploads/2016/03/

Rmarkdown reference guide

R Markdown Reference Guide

Learn more about R Markdown at rmarkdown.rstudio.com

Learn more about Interactive Docs at shiny.rstudio.com/articles

Contents:

1. Markdown Syntax
2. **Knitr chunk options**
3. Pandoc options

Syntax	Becomes

Make a code chunk with three back ticks followed by an r in braces. End the chunk with three back ticks:
 	
``{r}
paste("Hello", "World!")
``
	Make a code chunk with three back ticks followed by an r in braces. End the chunk with three back ticks:
	
	`paste("Hello", "World!")`
	
	`## [1] "Hello World!"`

Place code inline with a single back ticks. The first back tick must be followed by an R, like this `r paste("Hello", "World!")`.	Place code inline with a single back ticks. The first back tick must be followed by an R, like this Hello World!.

Add chunk options within braces. For example, `echo=FALSE` will prevent source code from being displayed:
 	
``{r eval=TRUE, echo=FALSE}
paste("Hello", "World!")
``
	Add chunk options within braces. For example, `echo=FALSE` will prevent source code from being displayed:
	
	`## [1] "Hello World!"`

Learn more about chunk options at http://yihui.name/knitr/options

Chunk options

option	default value	description
Code evaluation		
child	NULL	A character vector of filenames. Knitr will knit the files and place them into the main document.
code	NULL	Set to R code. Knitr will replace the code in the chunk with the code in the code option.
engine	'R'	Knitr will evaluate the chunk in the named language, e.g. **engine = 'python'**. Run **names(knitr::knit_engines$get())** to see supported languages.
eval	TRUE	If **FALSE**, knitr will not run the code in the code chunk.
include	TRUE	If **FALSE**, knitr will run the chunk but not include the chunk in the final document.
purl	TRUE	If **FALSE**, knitr will not include the chunk when running **purl()** to extract the source code.
Results		
collapse	FALSE	If **TRUE**, knitr will collapse all the source and output blocks created by the chunk into a single block.
echo	TRUE	If **FALSE**, knitr will not display the code in the code chunk above it's results in the final document.
results	'markup'	If **'hide'**, knitr will not display the code's results in the final document. If **'hold'**, knitr will delay displaying all output pieces until the end of the chunk. If **'asis'**, knitr will pass through results without reformatting them (useful if results return raw HTML, etc.)
error	TRUE	If **FALSE**, knitr will not display any error messages generated by the code.
message	TRUE	If **FALSE**, knitr will not display any messages generated by the code.
warning	TRUE	If **FALSE**, knitr will not display any warning messages generated by the code.
Code Decoration		
background	'#F7F7F7'	A background color for chunks in LaTex output.
comment	'##'	A character string. Knitr will append the string to the start of each line of results in the final document.
highlight	TRUE	If **TRUE**, knitr will highlight the source code in the final output.
prompt	FALSE	If **TRUE**, knitr will add > to the start of each line of code displayed in the final document.
size	'normalsize'	Fontsize for LaTex output.
strip.white	TRUE	If **TRUE**, knitr will remove white spaces that appear at the beginning or end of a code chunk.
tidy	FALSE	If **TRUE**, knitr will tidy code chunks for display with the **tidy_source()** function in the **formatR** package.

RStudio Updated 10/30/2014 © 2014 RStudio, Inc. All rights reserved.

2

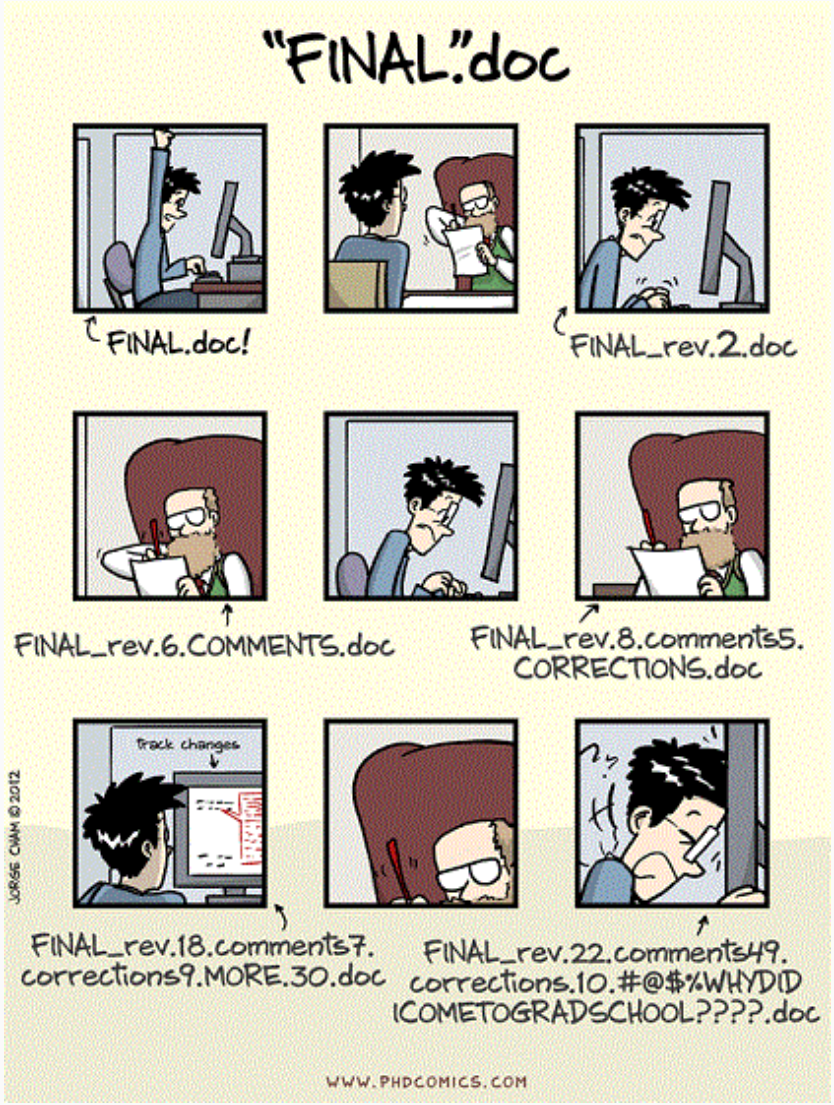
https://www.rstudio.com/wp-content/uploads/2015/03/

Managing software dependencies

Managing package dependencies in R

- sessionInfo (or session_info)
- switchr
- rctrack
- checkpoint
- packrat
- docker

Version control

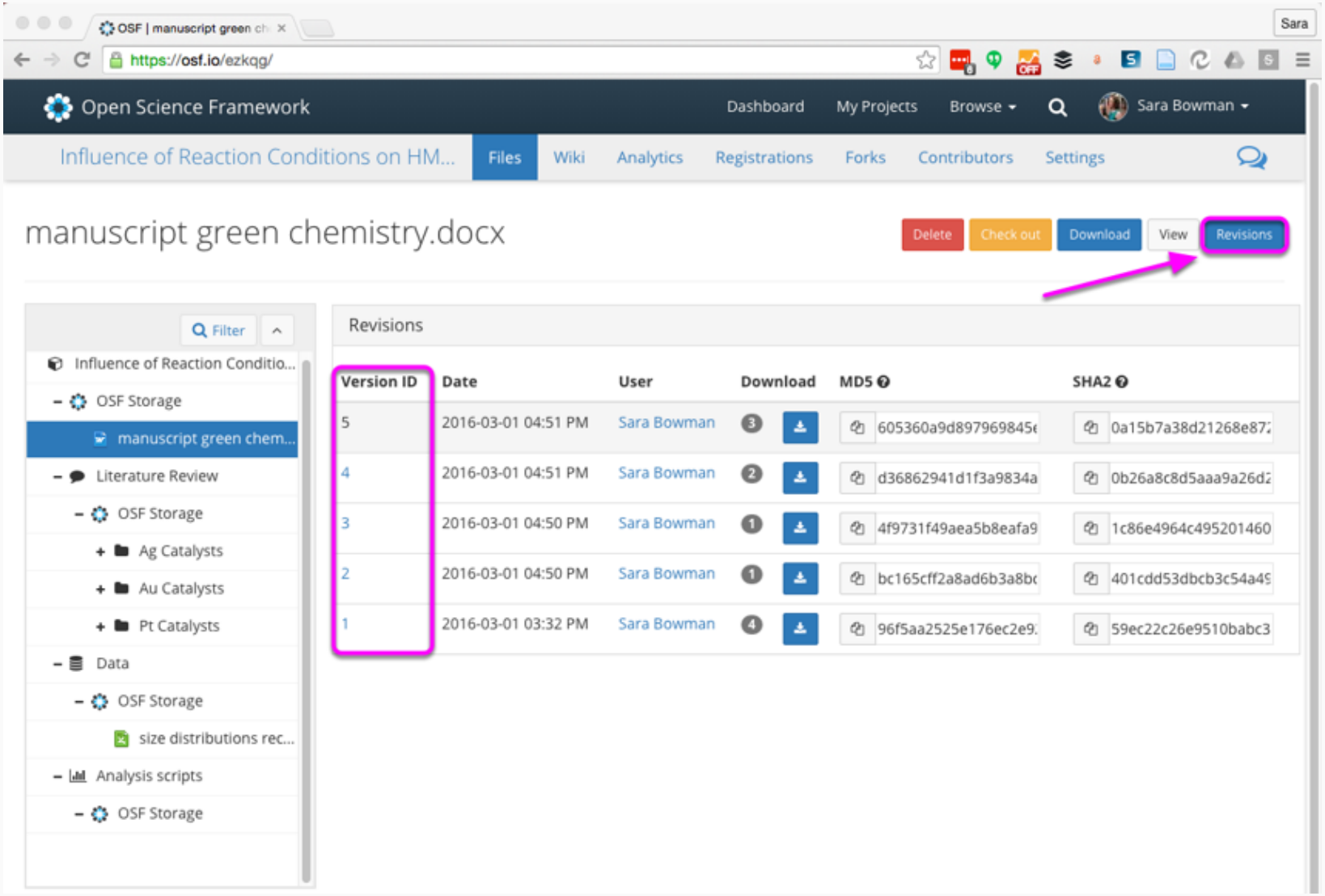

"FINAL".doc
FINAL.doc!
FINAL_rev.2.doc
FINAL_rev.6.COMMENTS.doc
FINAL_rev.8.comments5.
CORRECTIONS.doc
track changes
FINAL_rev.18.comments7.
corrections9.MORE.30.doc
FINAL_rev.22.comments49.
corrections.10.#@$%WHYDID
ICOMETOGRADSCHOOL????.doc
JORGE CHAM © 2012
WWW.PHDCOMICS.COM

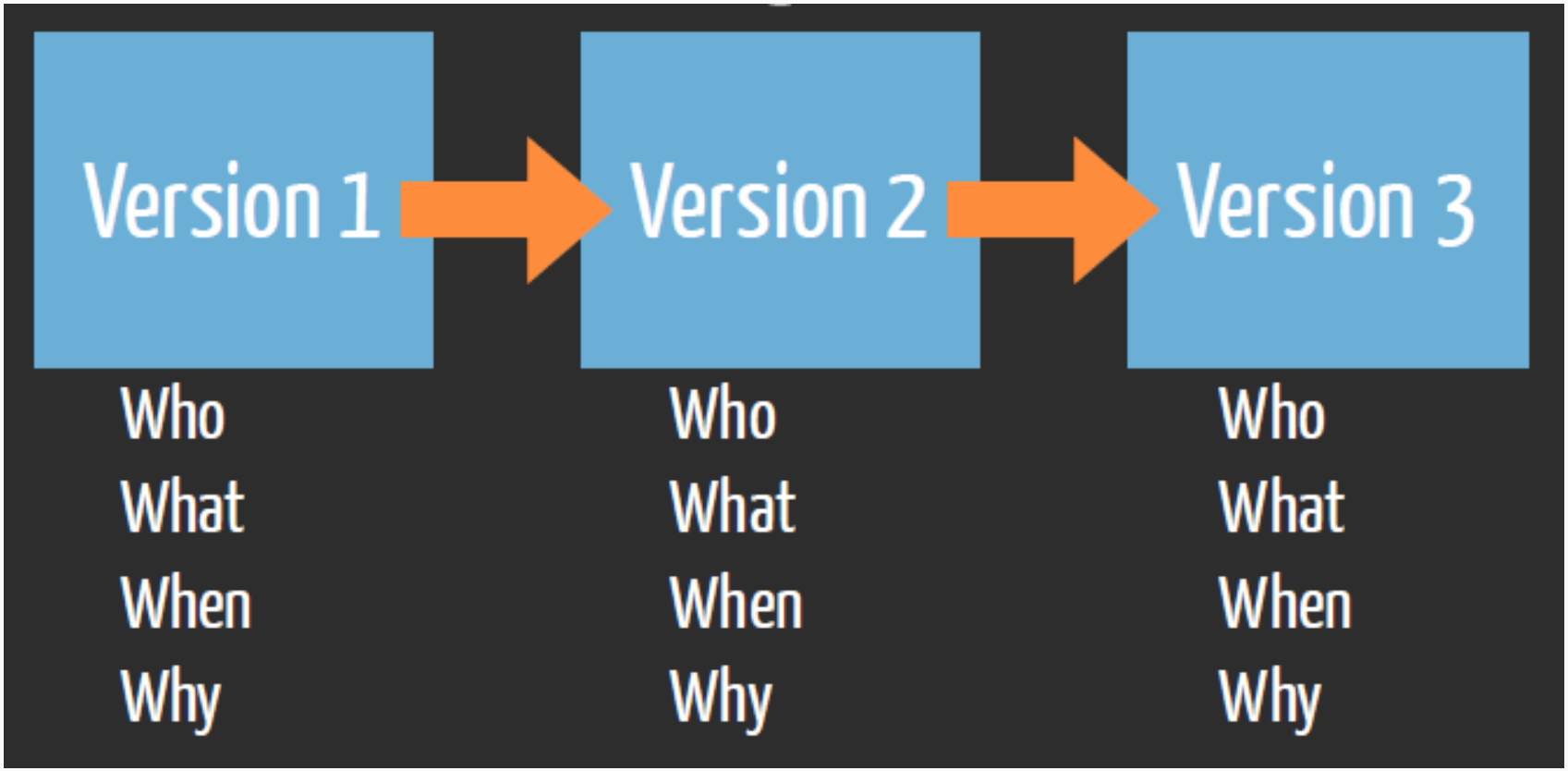
Dropbox

Dropbox keeps record of deleted/edited files for 30 days

Open Science Framework

Automatic version control, no time limit.





R. Fitzjohn (https://github.com/richfitz/reproducibility-2014)

Git & GitHub

- Sign up for GitHub
- Install Git
- Introduce yourself
- Create repo on GitHub
- Clone repo in Rstudio
- Make changes, push, pull
- Collaboration

Collaborative writing

Many alternatives

- Rmarkdown + GitHub
- Word + Dropbox
- Google Docs
- Overleaf
- Authorea
- ...

Happy writing!

Slides and source code available at

https://github.com/Pakillo/ReproducibleScience