Reproducible Research

An Introduction to `knitr`

Sahir Rai Bhatnagar[1]

May 28, 2014

[1] https://github.com/sahirbhatnagar/knitr-tutorial

Acknowledgements

- Dr. Erica Moodie
- Maxime Turgeon, Kevin McGregor, Greg Voisin
- You

Disclaimer #1

- Feel free to Ask questions
- Interrupt me often
- You don't need to raise your hand to speak

Disclaimer #2

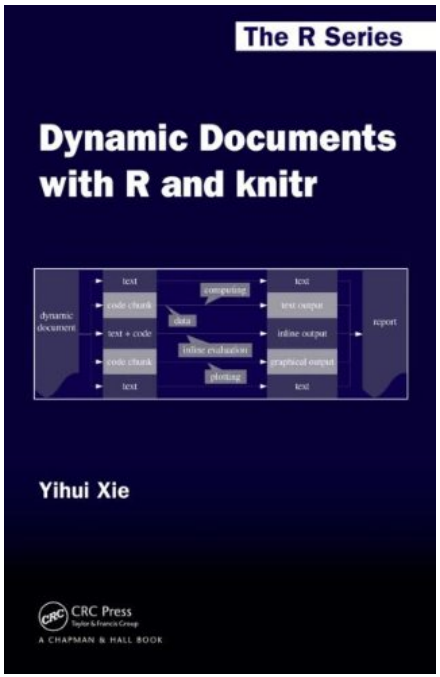
R Markdown v2

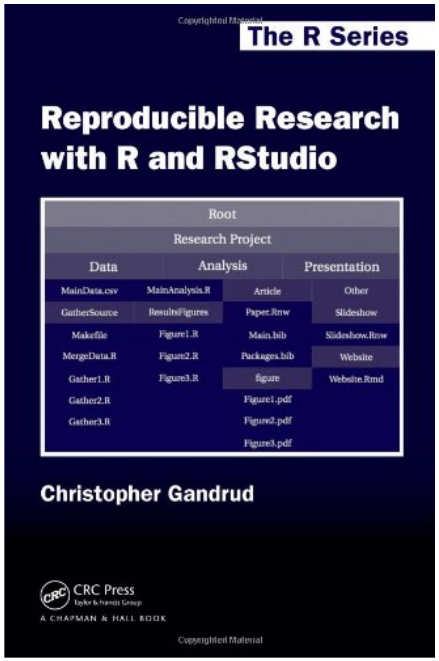
LaTeX

I don't work for, nor am I an author of any of these packages. I'm just a messenger.

Disclaimer #3

- Material for this tutorial comes from many sources. For a complete list see: https://github.com/sahirbhatnagar/knitr-tutorial
- Alot of the content in these slides are based on these two books





Eat Your Own Dog Food

- These slides are reproducible
- Source code: https://github.com/sahirbhatnagar/knitr-tutorial

Main objective for today

What is Science Anyway?

What is Science Anyway?

According to the American Physical Society:

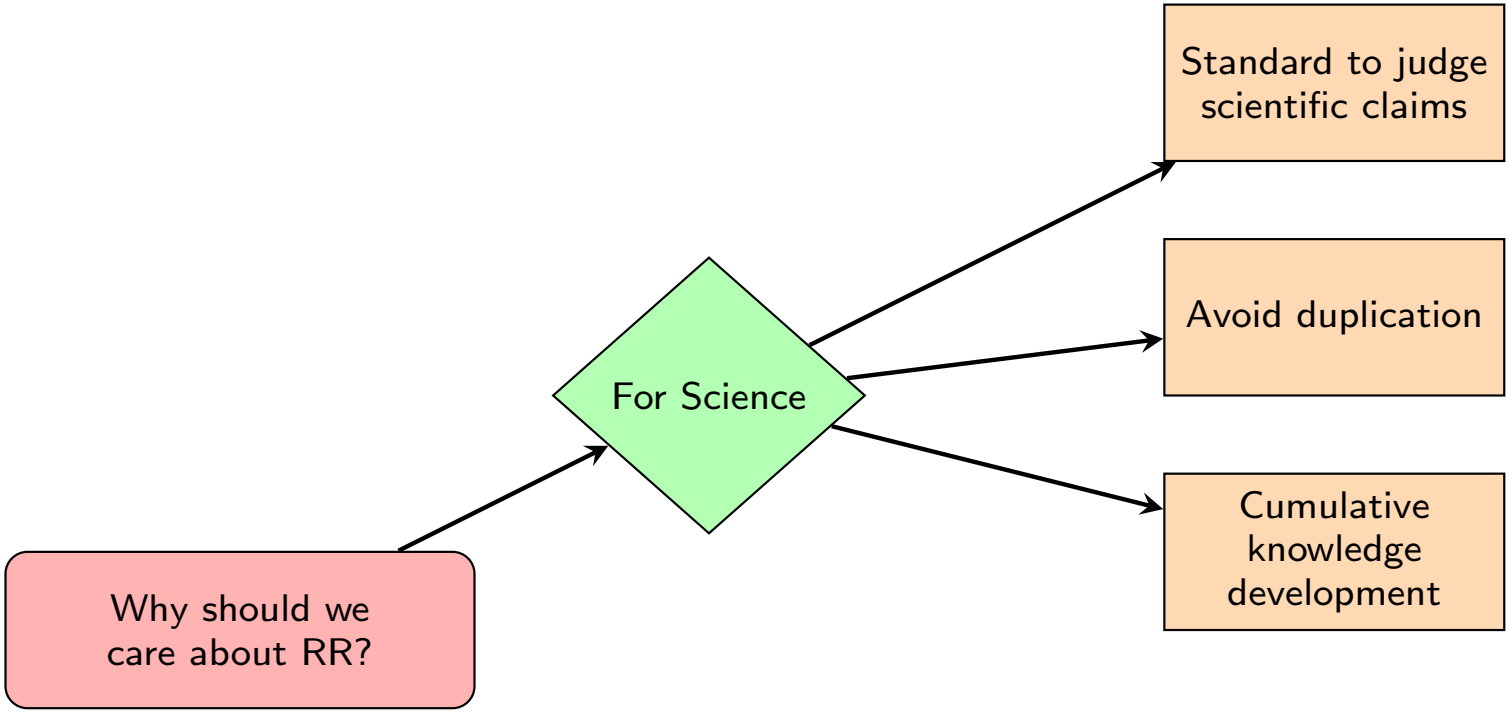
Science is the systematic enterprise of gathering knowledge about the universe and organizing and condensing that knowledge into **testable** *laws and theories. The* **success and credibility of science** *are anchored in the* **willingness** *of scientists to* **expose their ideas** *and results to* **independent testing** *and* **replication** *by other scientists*

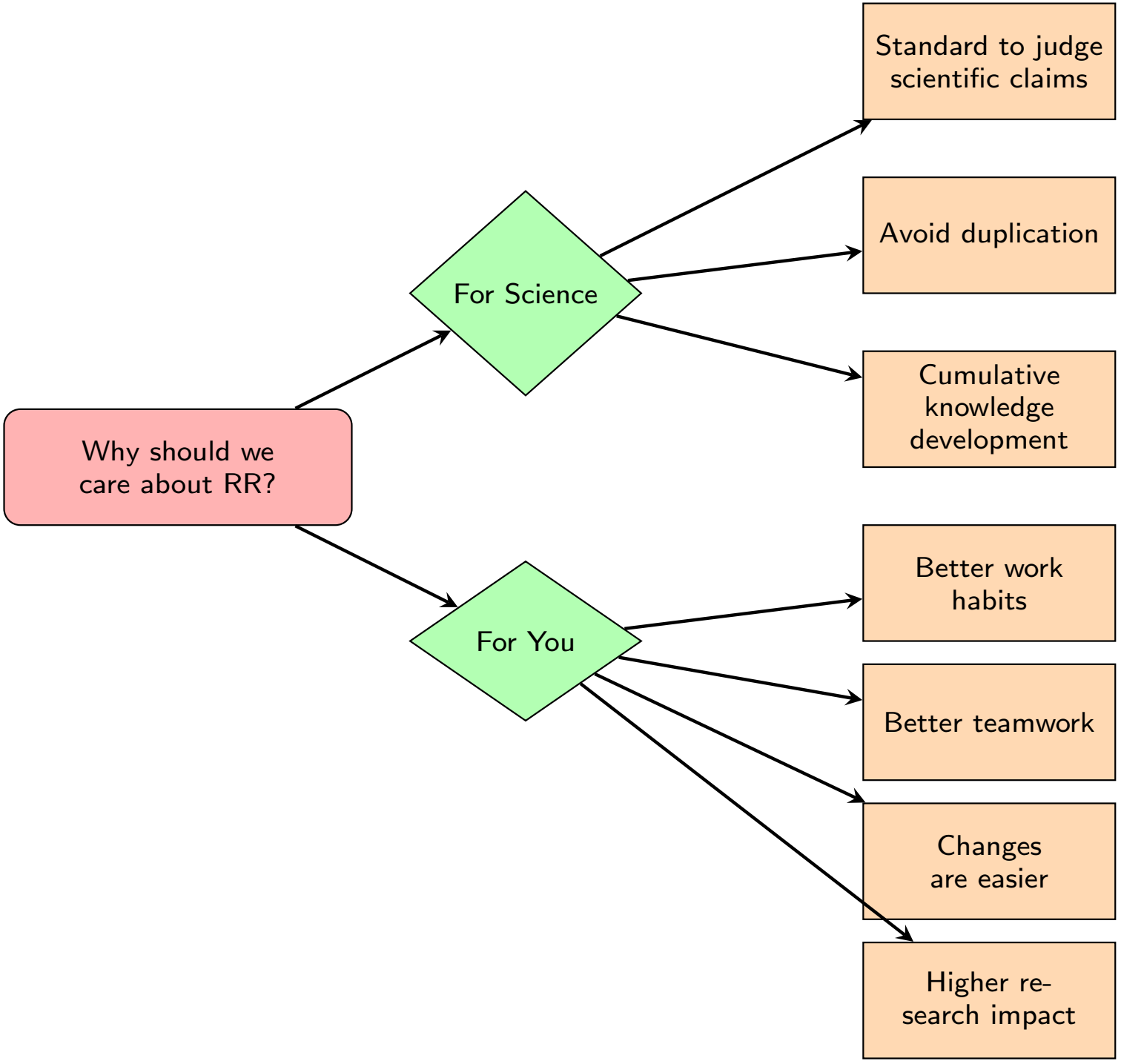
RR: A Minimum Standard to Verify Scientific Findings

RR: A Minimum Standard to Verify Scientific Findings

Reproducible Research (RR) in Computational Sciences

The data and the code used to make a finding are available and they are sufficient for an independent researcher to recreate the finding





A Motivating Example

Demonstrate: 001-motivating-example

Tools for Reproducible Research[2]

Free and Open Source Software

- `RStudio`: Creating, managing, compiling documents
- LaTeX: Markup language for typesetting a document
- `R`: Statistical analysis language
- `knitr`: Integrate LaTeXand `R` code. Based on Prof. Friedrich Leisch's `Sweave`

[2] http://onepager.togaware.com/

Comparison

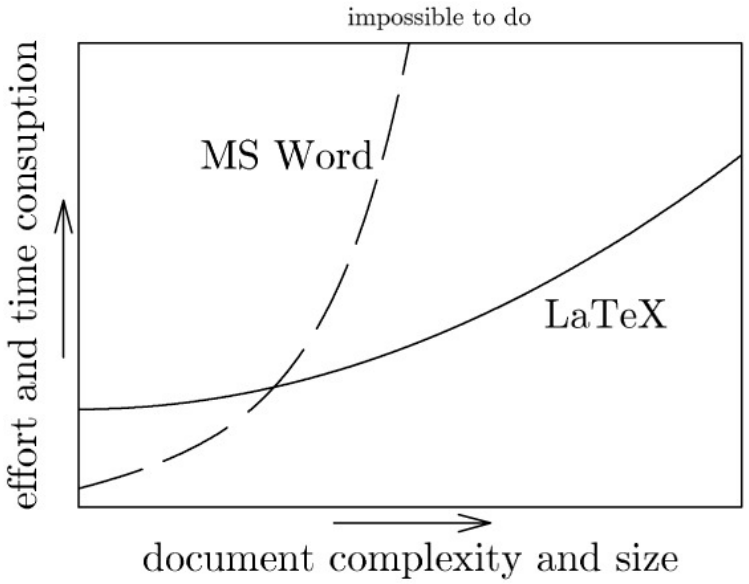


Figure 1 : Comparison

- LaTeX has a greater learning curve
- Many tasks are very tedious or impossible (most cases) to do in MS Word or Libre Office

The Philosophy behind LaTeX

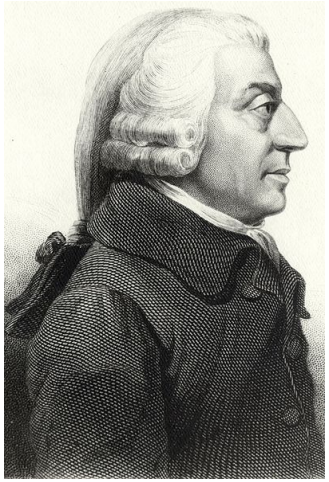


Figure 2 : Adam Smith, author of *The Wealth of Nations* (1776), in which he conceptualizes the notion of the division of labour

Division of Labour

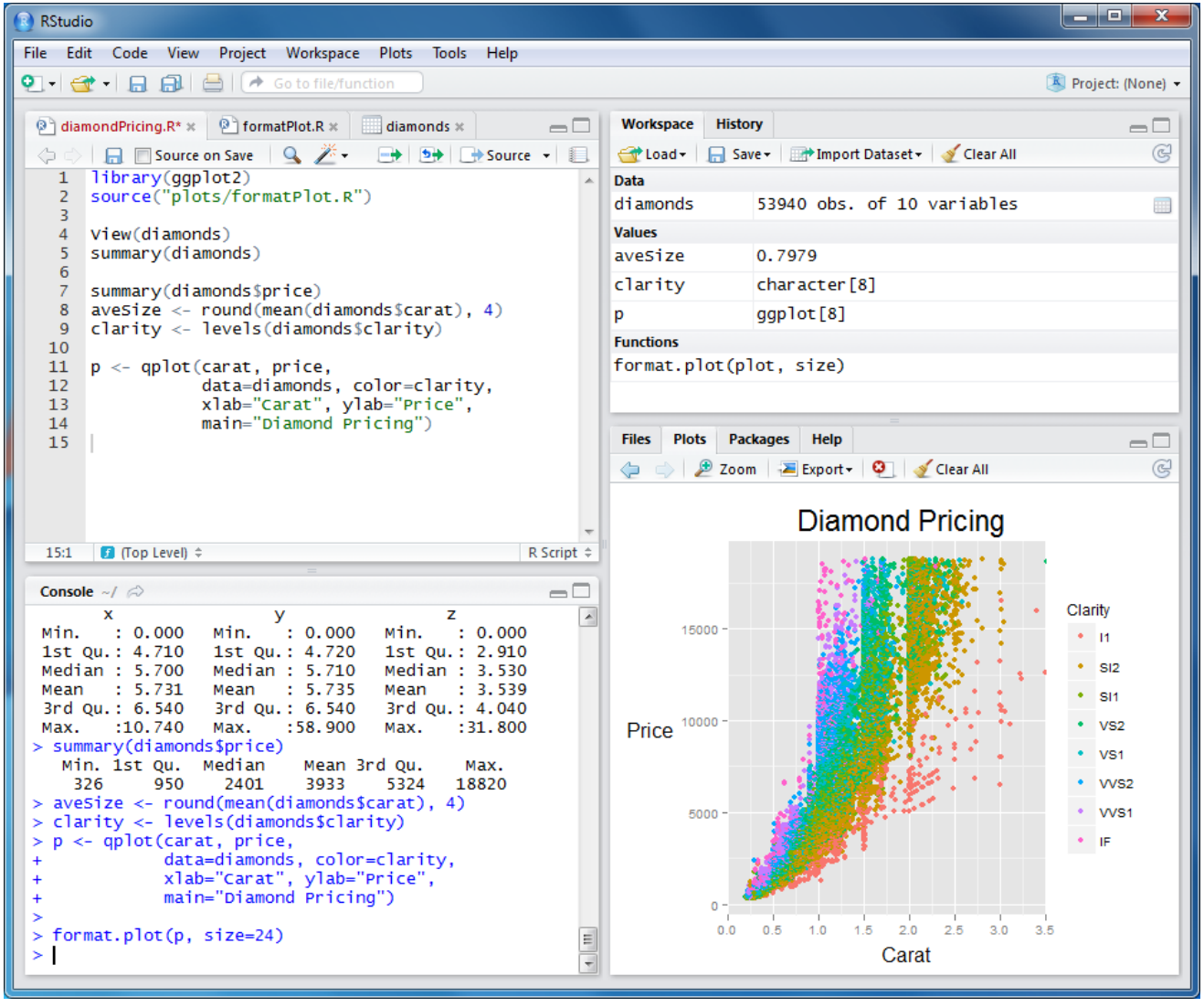
Composition and logical structuring of text is the author's specific contribution to the production of a printed text. Matters such as the choice of the font family, should section headings be in bold face or small capitals? Should they be flush left or centered? Should the text be justified or not? Should the notes appear at the foot of the page or at the end? Should the text be set in one column or two? and so on, is the typesetter's business

The Genius Behind LaTeX

Figure 3 : The TeX project was started in 1978 by Donald Knuth (Stanford). He planned for 6 months, but it took him nearly 10 years to complete. Coined the term "Literate programming": mixture of code and text segments that are "human" readable. Recipient of the Turing Award (1974) and the Kyoto Prize (1996).

Integrated Development Environment (IDE)

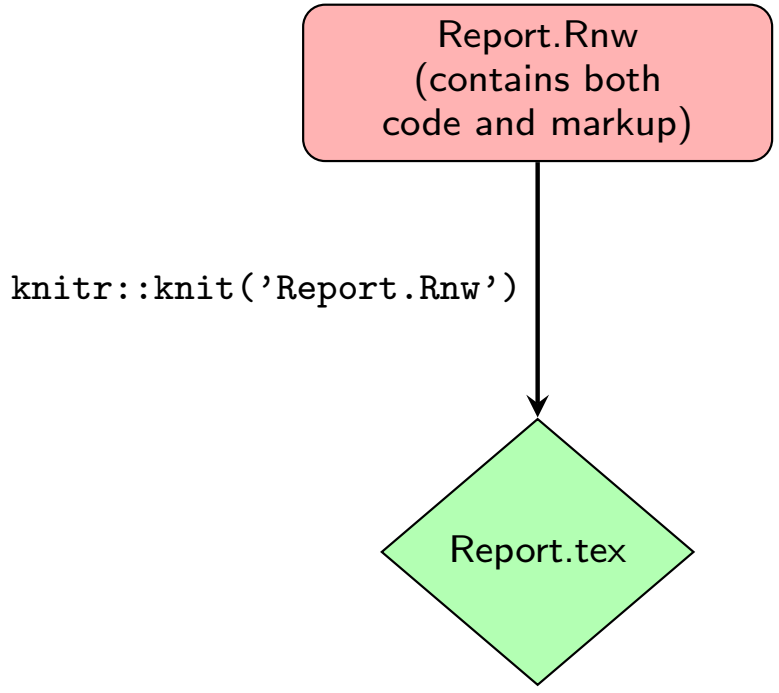
Integrated Development Environment (IDE)



Demonstrate: Explore `RStudio`, projects and `.Rprofile`

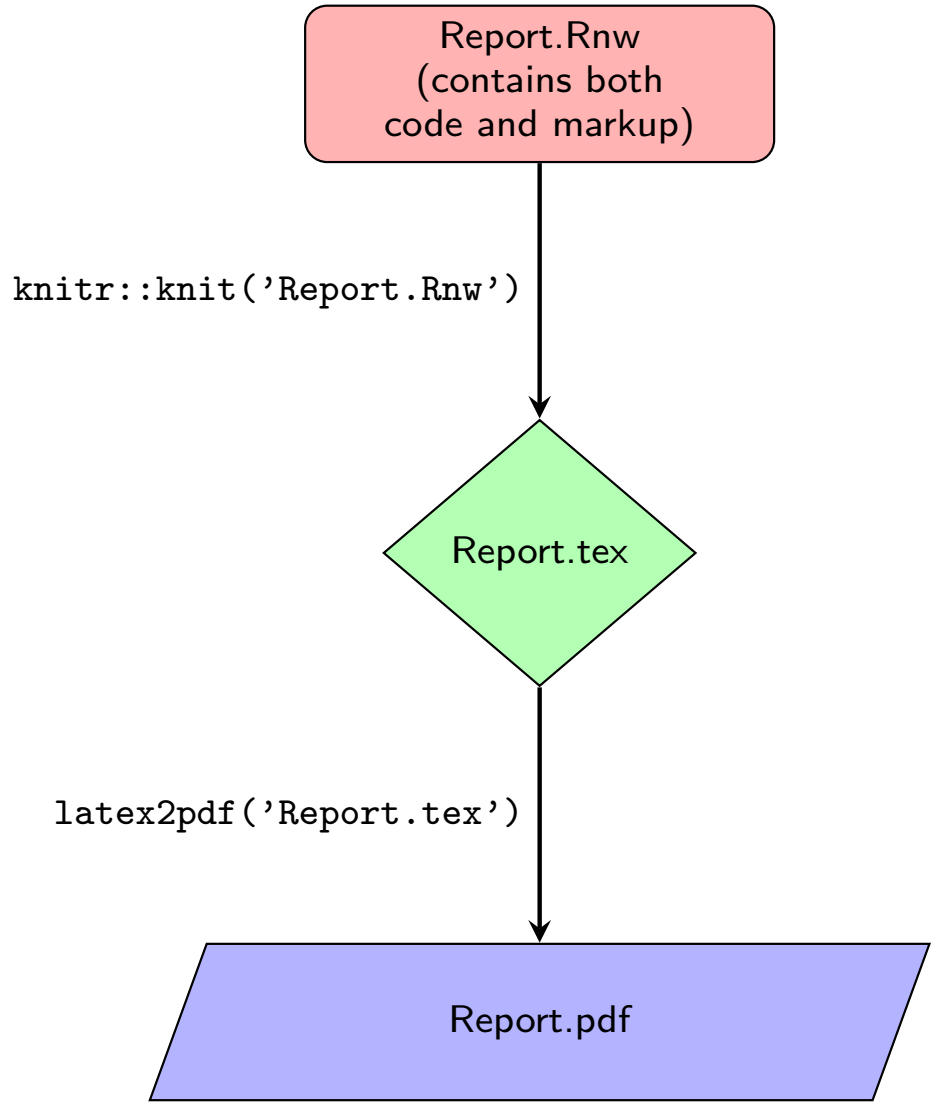
What knitr does

LaTeX example:



What knitr does

LaTeX example:

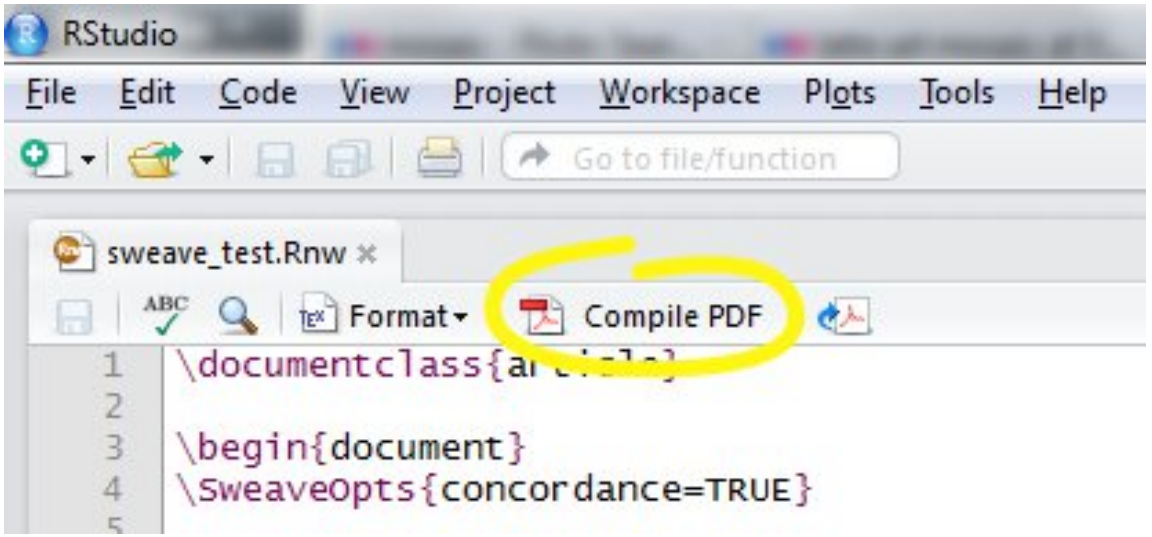


Compiling a `.Rnw` document

The two steps on previous slide can be executed in one command:

```
knitr::knit2pdf()
```

or in RStudio:



Incorporating R code

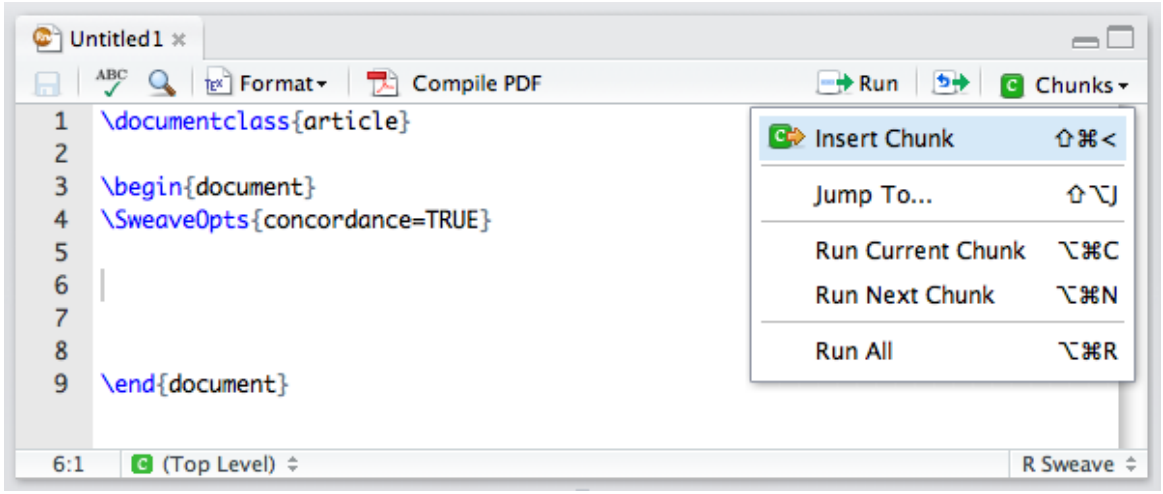
- Insert R code in a **Code Chunk** starting with

  << >>=

  and ending with

  @

In RStudio:



Example 1

```
<<example-code-chunk-name, echo=TRUE>>=
library(magrittr)
rnorm(50) %>% mean
@
```

produces

```
library(magrittr)
rnorm(50) %>% mean

## [1] 0.031
```

Example 2

```
<<example-code-chunk-name2, echo=TRUE, tidy=TRUE>>=
for(i in 1:5){ (i+3) %>% print}
@
```

produces

```
for (i in 1:5) {
    (i + 3) %>% print
}
```

```
## [1] 4
## [1] 5
## [1] 6
## [1] 7
## [1] 8
```

Example 2.2


```
<<example-code-chunk-name3, echo=FALSE>>=
for(i in 1:5){ (i+3) %>% print}
@
```

produces

```
## [1] 4
## [1] 5
## [1] 6
## [1] 7
## [1] 8
```

Example 2.3

```
<<example-code-chunk-name4, echo=FALSE, eval=FALSE>>=
for(i in 1:5){ (i+3) %>% print}
@
```

produces

R output within the text

- Include R output within the text
- We can do that with "S-expressions" using the command `\Sexpr{...}`

Example:

The iris dataset has `\Sexpr{nrow(iris)}` rows and `\Sexpr{ncol(iris)}` columns

produces

The iris dataset has 150 rows and 5 columns

Include a Figure


```
<<fig.ex, fig.cap='Linear Regression',fig.height=3,fig.width=3>>=
plot(mtcars[ , c('disp','mpg')])
lm(mpg ~ disp , data = mtcars) %>%
abline(lwd=2)
@
```

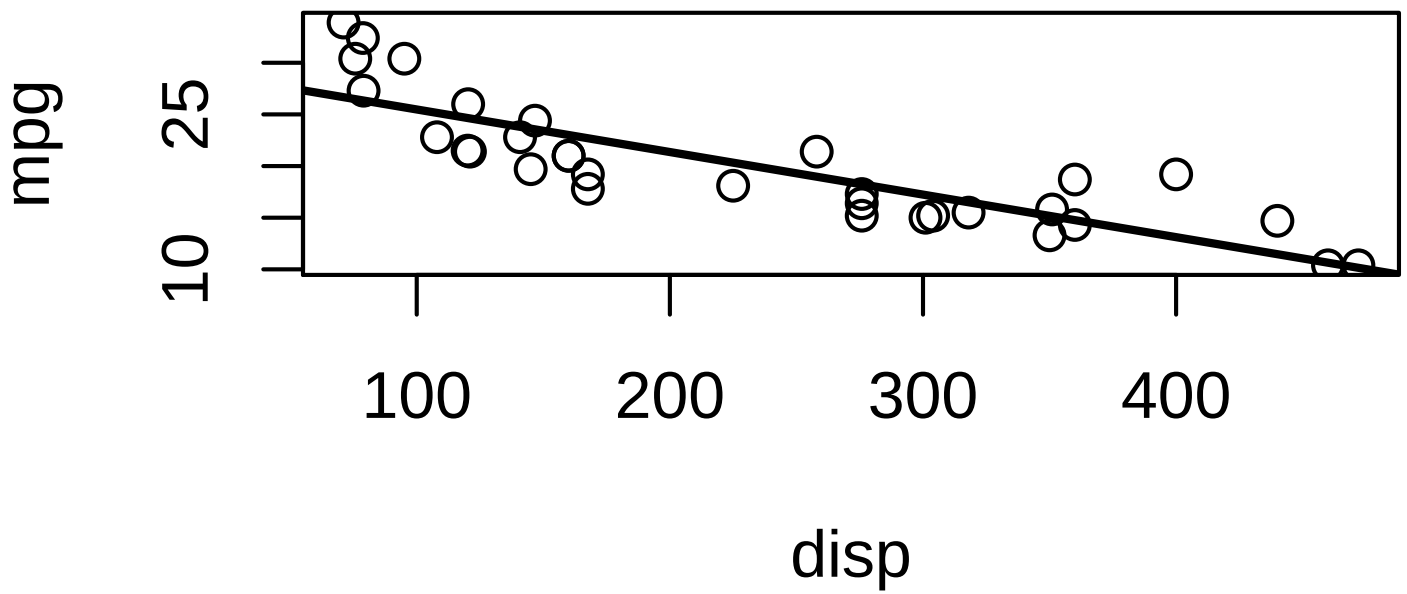


Figure 4 : Linear regression

Include a Table

```
<<table.ex, results='asis'>>=
library(xtable)
iris[1:5,1:5] %>%
xtable(caption='Sample of Iris data') %>%
print(include.rownames=FALSE)
@
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.10	3.50	1.40	0.20	setosa
4.90	3.00	1.40	0.20	setosa
4.70	3.20	1.30	0.20	setosa
4.60	3.10	1.50	0.20	setosa
5.00	3.60	1.40	0.20	setosa

Table 1 : Sample of Iris data

A selection of `knitr` code chunk options

content...

Set global chunk options

content...

Option Aliases

see page 109 yihui

Option Templates

see page 110 yihui

Chunk References

see page 79 yihui

Code in Appendix

see page 110 yihui

A selection of `knitr` code chunk options

content...

A selection of `knitr` code chunk options

see 83

Example Environment

see 120

hello world

hello world

hello world

hello world

hello world

hello world

hello world

Opinion: Reproducible research can still be wrong: Adopting a prevention approach

Jeffrey T. Leek[a,1] and Roger D. Peng[b]
[a]*Associate Professor of Biostatistics and Oncology and* [b]*Associate Professor of Biostatistics, Johns Hopkins University, Baltimore, MD*

computational tools such as knitr, iPython notebook, LONI, and Galaxy (8) have simplified the process of distributing reproducible data analyses.

Is the juice worth the squeeze?