Avoir une présence en ligne

Outils pour la diffusion rapide et reproductible de la recherche

Sahir Rai Bhatnagar[1]

May 14, 2019

[1] https://github.com/sahirbhatnagar/raqc

Remerciements

- La comité organisateur
- Pierre Racine et Sophie Baillargeon
- Don Knuth (TeX)
- Friedrich Leisch (Sweave)
- Yihui Xie (knitr)
- Vous

UNIVERSITÉ LAVAL

Avis #1

- Ceci est une **introduction** au outils pour la recherche reproductible

Avis #1

- Ceci est une **introduction** au outils pour la recherche reproductible
- Le niveau de cet atelier est ”intermédiaire” et suppose des connaissances de base en R ainsi que de l’environnement RStudio

Avis #1

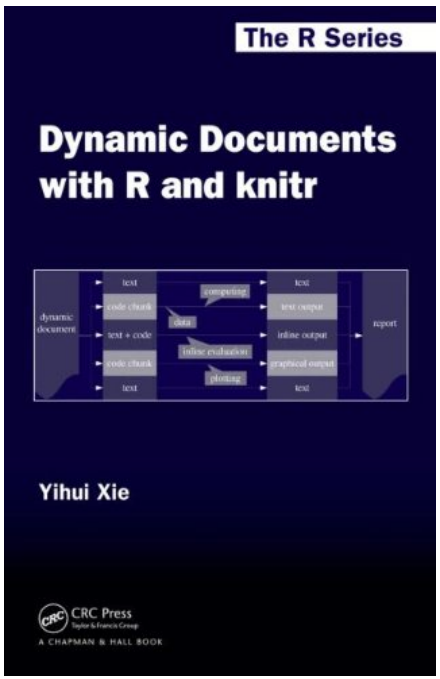
- Ceci est une **introduction** au outils pour la recherche reproductible
- Le niveau de cet atelier est ”intermédiaire” et suppose des connaissances de base en R ainsi que de l’environnement RStudio
- N’hésitez pas à posez des questions

Avis #2

Je n'ai aucune relation commerciale avec ces logiciels.

Avis #3

- Le matériel pour cet atelier est basé sur plusieurs ressources
- Voir ce lien pour une liste complète de références: https://github.com/sahirbhatnagar/raqc
- Une grande partie du contenu de ces diapositives est basée sur ces deux livres:



Eat Your Own Dog Food

- Ces diapositives sont reproductibles
- Voir `raqc-slides.Rnw`: https://github.com/sahirbhatnagar/raqc/tree/master/slides

Le programme de l'atelier

- **8h30 à 10h00**: Introduction aux raports reproductibles avec knitr et RMarkdown
- **10h00 à 10h30**: Pause
- **10h30 à 12h00**:
- **13h30 à 15h00**: Créer un siteweb avec blogdown
- **15h00 à 15h30**: Pause
- **15h30 à 17h**:
- **17h**: Fin de l'atelier

C'est quoi la science?

C'est quoi la science?

Selon l'American Physical Society:

La science est l'entreprise systématique consistant à rassembler des connaissances sur l'univers et à les organiser et les condenser en lois et **théories vérifiables**...

C'est quoi la science?

Selon l'American Physical Society:

La science est l'entreprise systématique consistant à rassembler des connaissances sur l'univers et à les organiser et les condenser en lois et **théories vérifiables**...

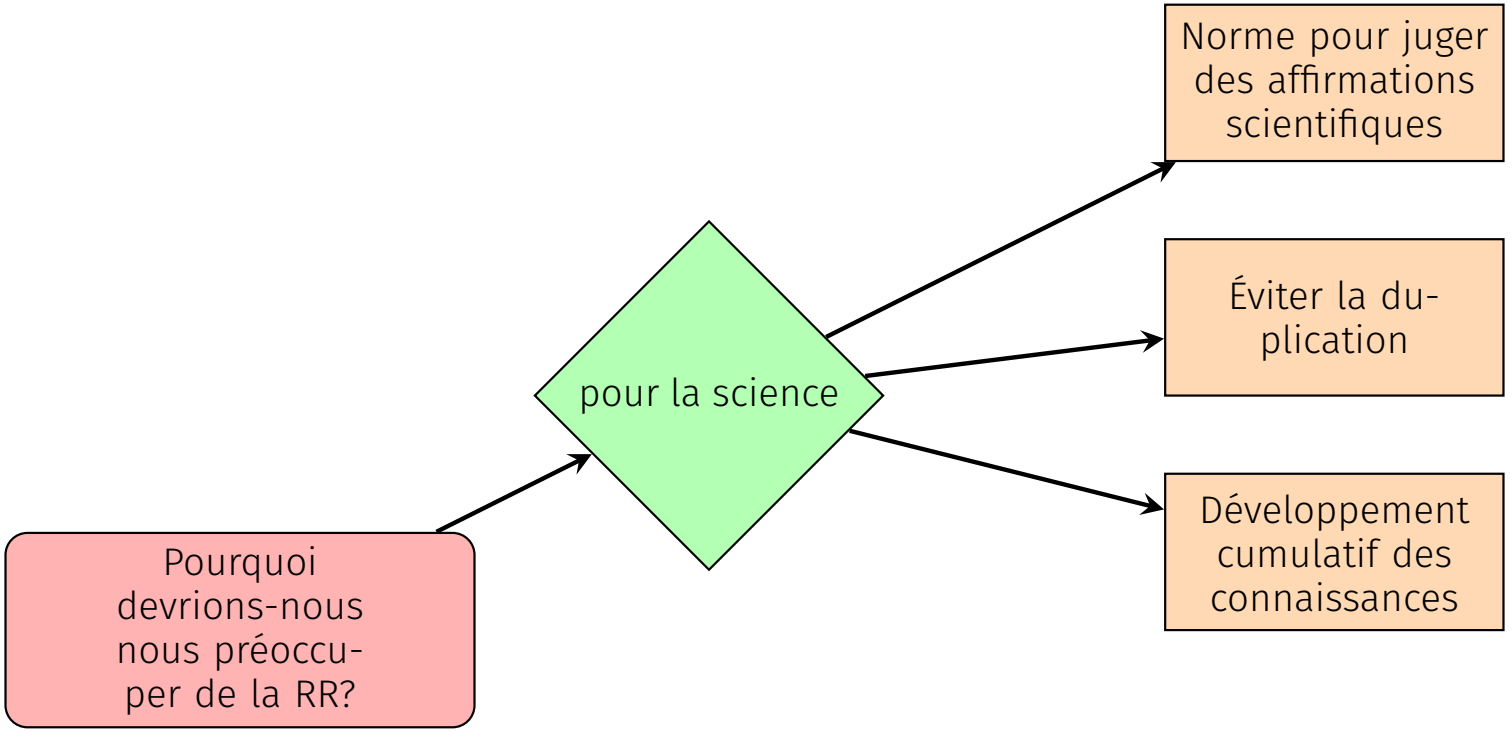
Le **succès et la crédibilité de la science** sont fondés sur la volonté des scientifiques **d'exposer leurs idées** et **leurs résultats** à des **tests indépendants** et à leur **reproduction** par d'autres scientifiques.

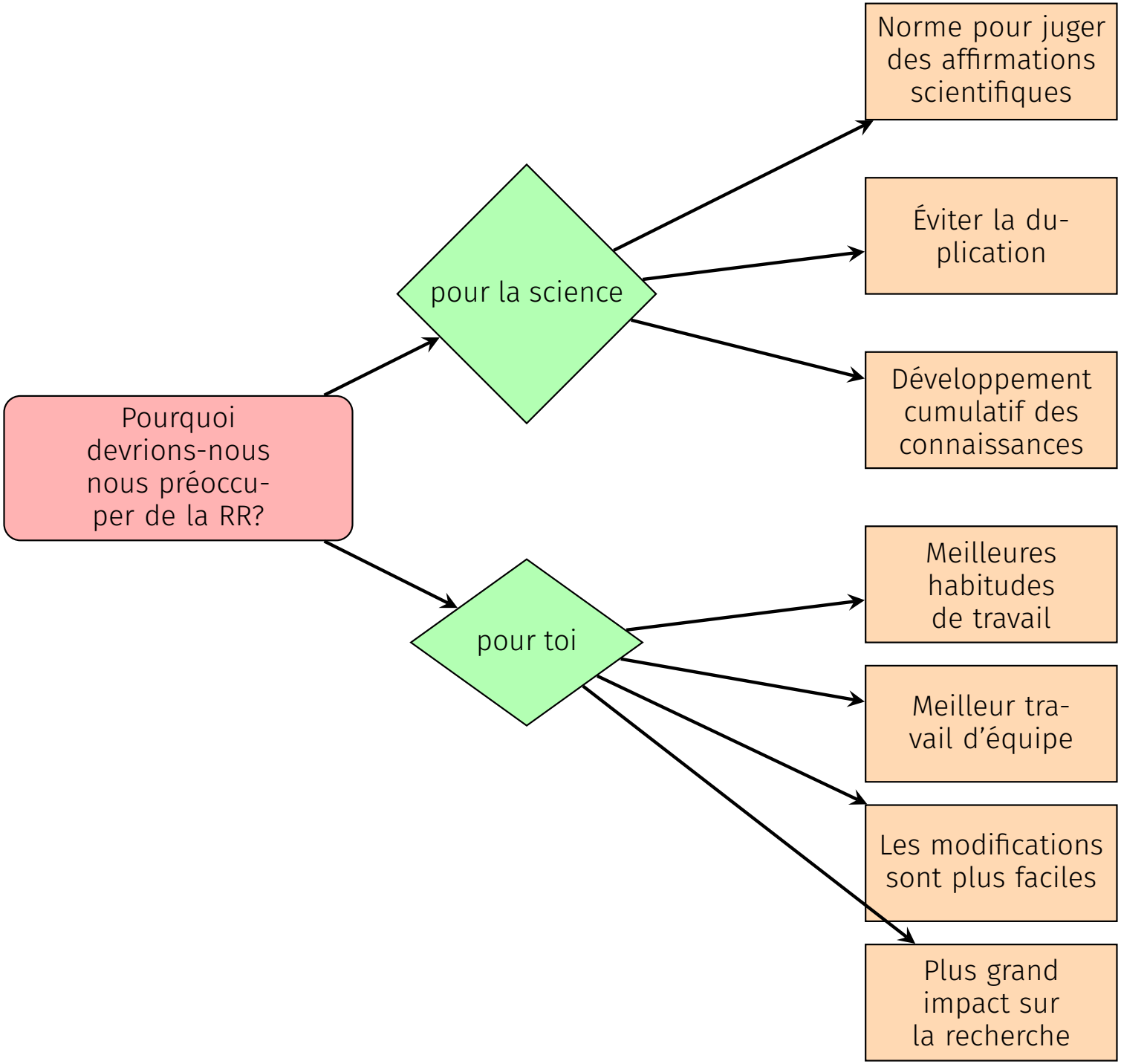
RR: Une norme minimale pour vérifier les résultats scientifiques

RR: Une norme minimale pour vérifier les résultats scientifiques

Recherche reproductible (RR) dans la science des données

Les données et le code utilisés pour effectuer une constatation sont disponibles et suffisent à un chercheur indépendant pour recréer la constatation.





Un exemple justificatif

Demonstrate: 001-exemple-justificatif

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

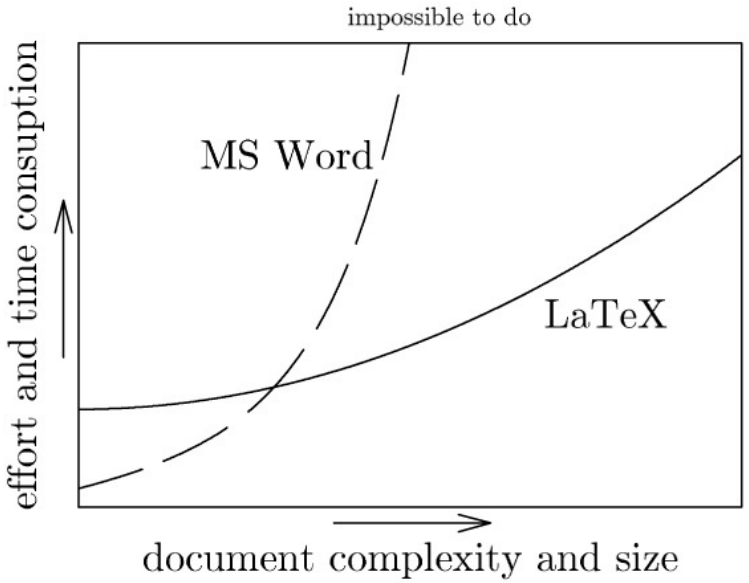


Figure 1: Comparison

- 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

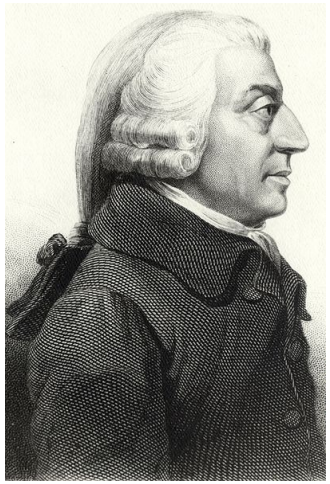


Figure 2: Adam Smith, author of *The Wealth of Nations* (1776), in which he conceptualizes the notion of the division of labour

Division of Labour

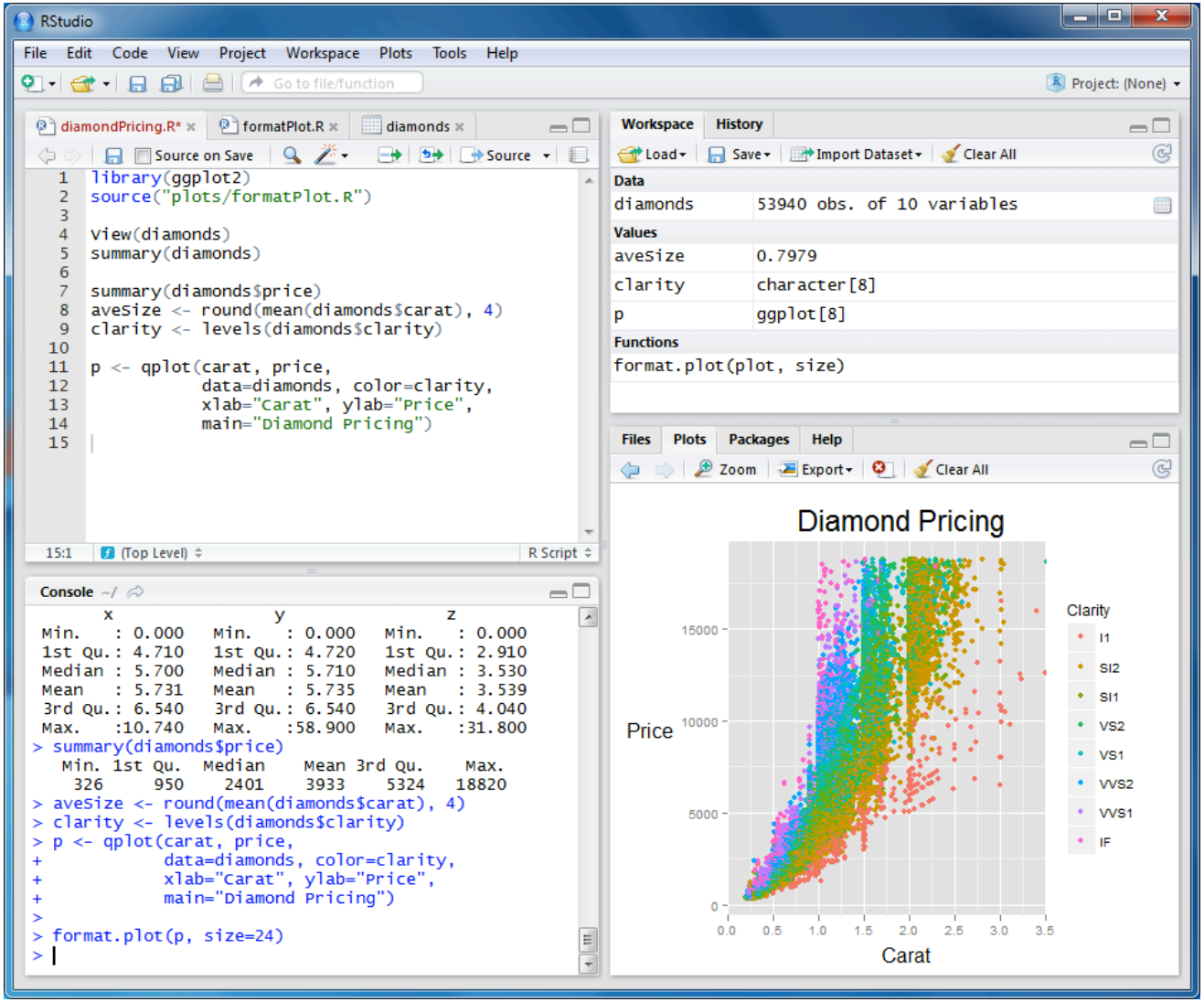
Composition and logical structuring of text is the author's specific contribution to the production of a printed text. Matters such as the choice of the font family, should section headings be in bold face or small capitals? Should they be flush left or centered? Should the text be justified or not? 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)

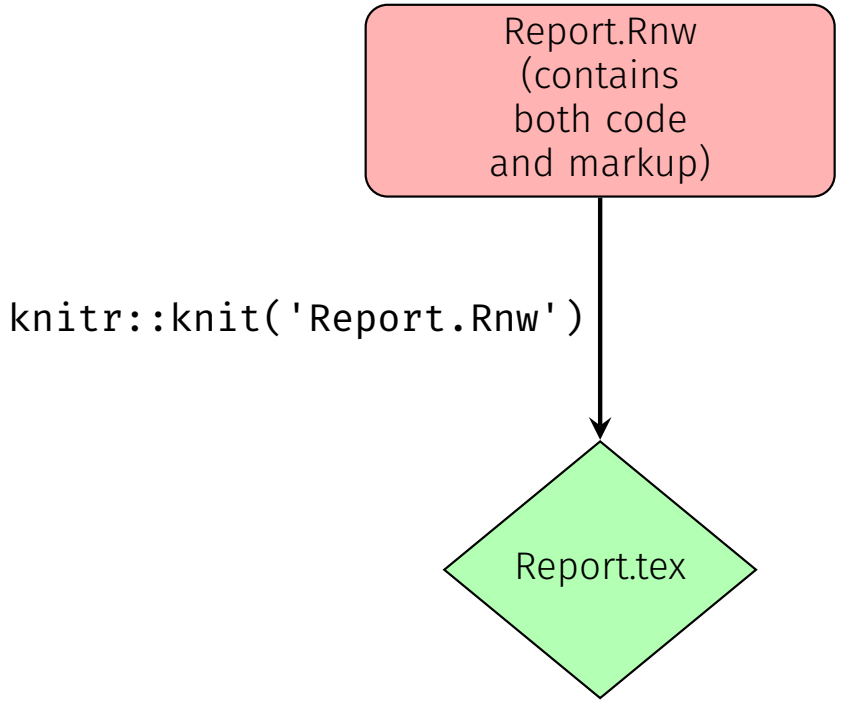
Integrated Development Environment (IDE)



Demonstrate: Explore **RStudio**

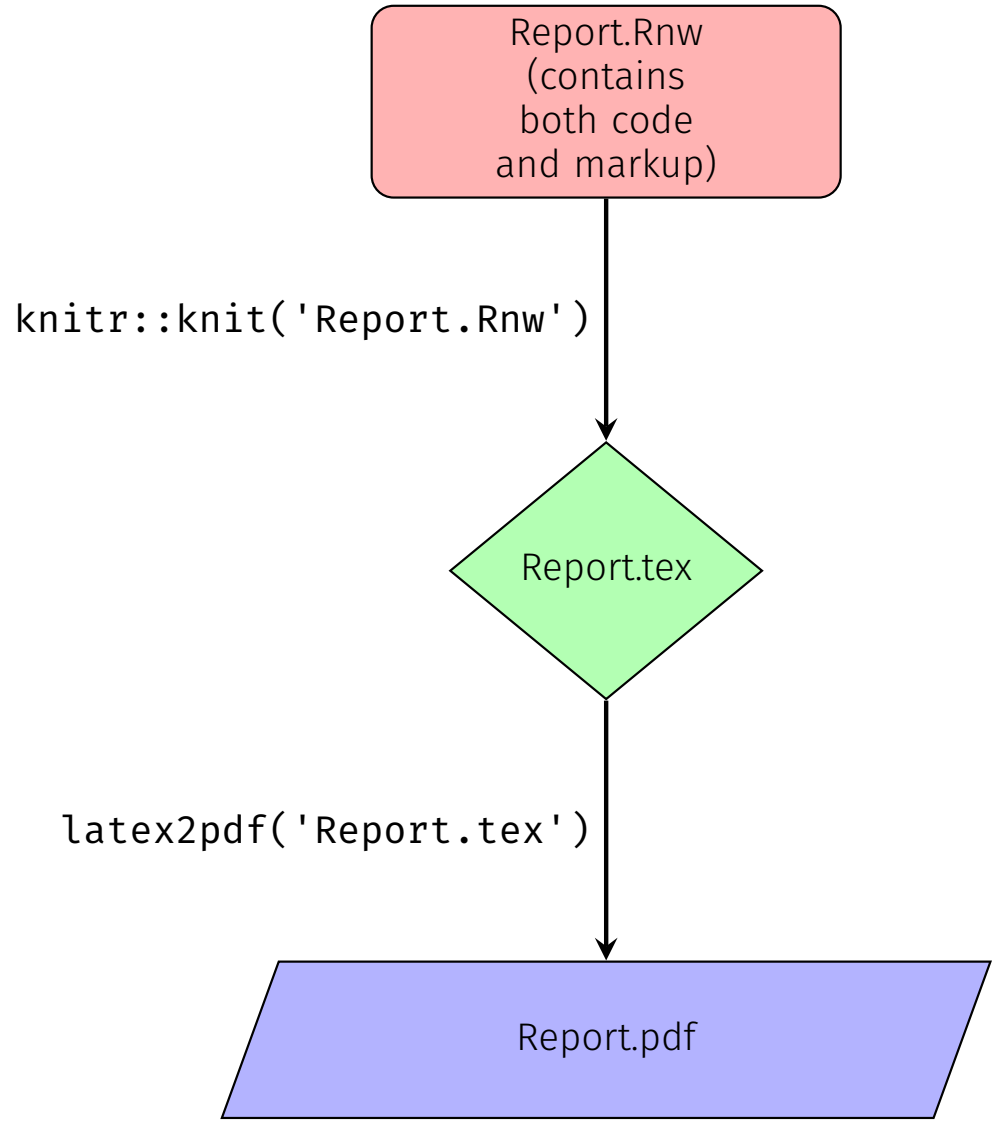
What `knitr` does

LaTeX example:



What `knitr` does

LaTeX example:

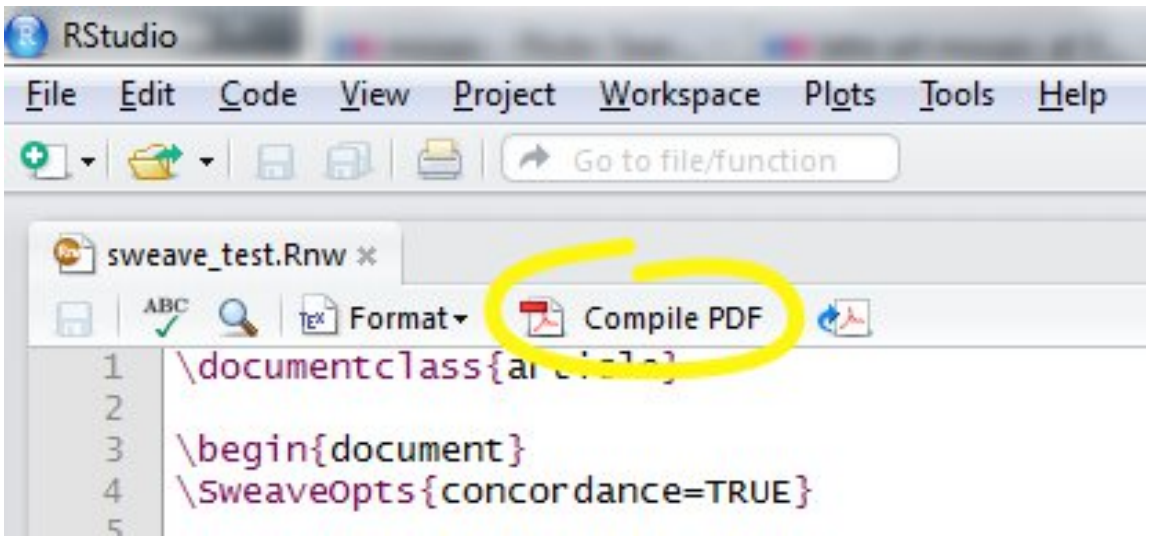


Compiling a `.Rnw` document

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

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

or in RStudio:



Incorporating R code

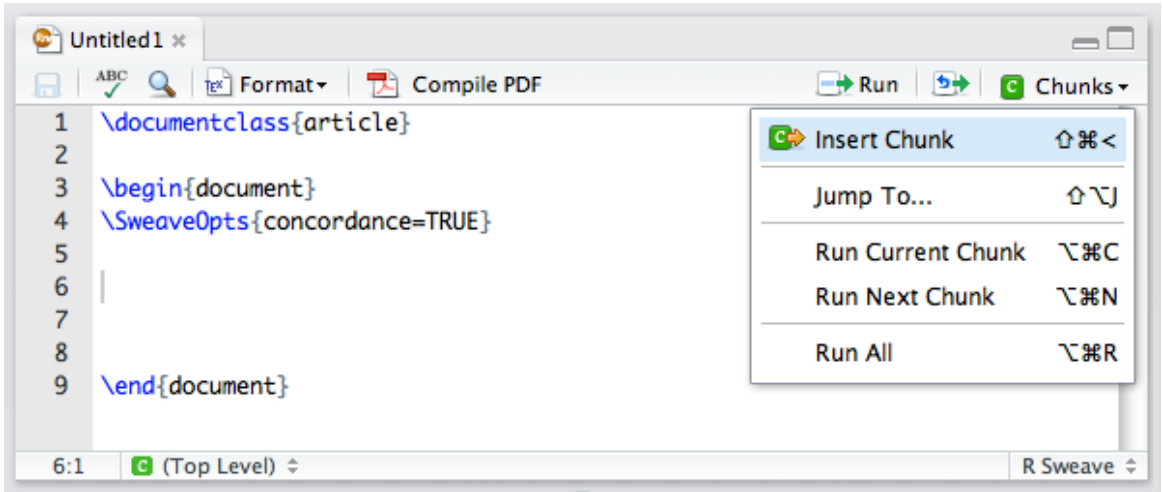
- Insert R code in a **Code Chunk** starting with

<< >>=

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] 0.12
```

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

Demonstrate: Try it yourself

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)
@
```

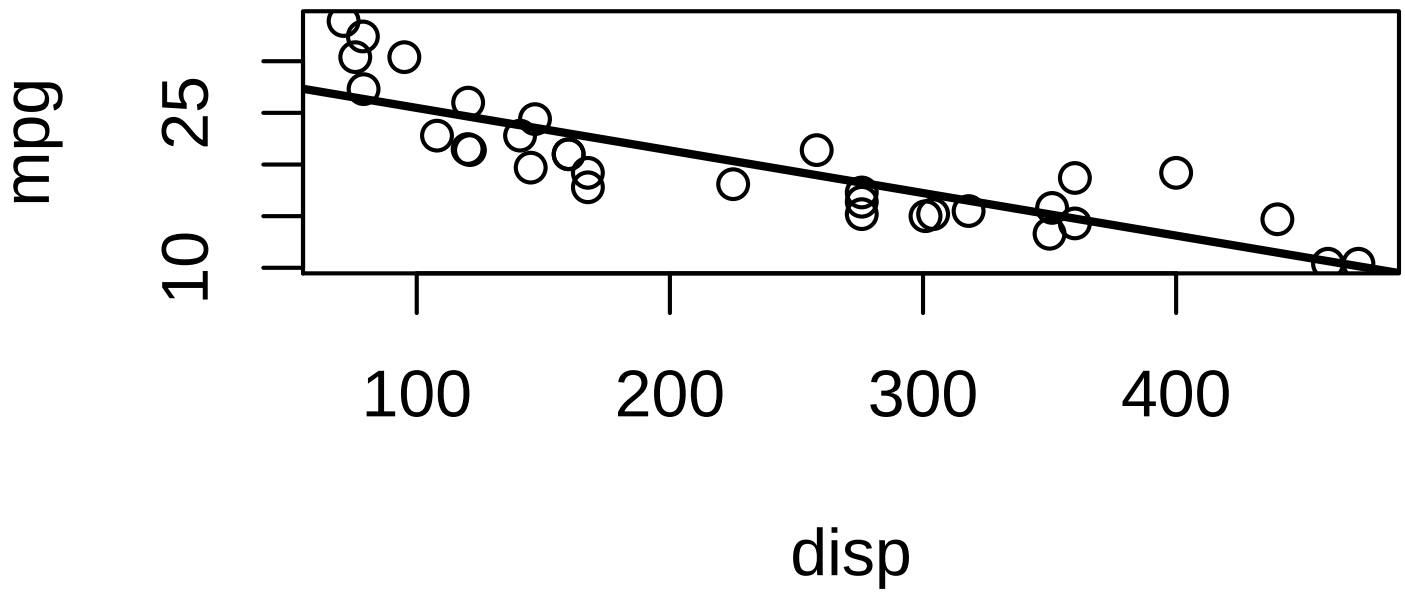


Figure 4: Linear regression

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

Minimum Working Example

https://github.com/sahirbhatnagar/knitr-tutorial/tree/master/002-minimum-working-example

Extracting output from Regression Models

https://github.com/sahirbhatnagar/knitr-tutorial/tree/master/003-model-output

Figures

https://github.com/sahirbhatnagar/knitr-tutorial/tree/master/004-figures

Beamer Presentations

https://github.com/sahirbhatnagar/knitr-tutorial/tree/master/005-beamer-presentation

Changing one Parameter in an Analysis

https://github.com/sahirbhatnagar/knitr-tutorial/tree/master/006-sensitivity-analysis-one-parameter

Changing Many Parameters in an Analysis

https://github.com/sahirbhatnagar/knitr-tutorial/tree/master/007-sensitivity-analysis-many-parameters

Large Documents

https://github.com/sahirbhatnagar/knitr-tutorial/tree/master/008-large-documents

HTML Reports

https://github.com/sahirbhatnagar/knitr-tutorial/tree/master/009-rmarkdown

HTML Presentations

https://github.com/sahirbhatnagar/knitr-tutorial/tree/master/010-rmarkdown-presentation

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.

Always Remember ...

$$\text{Reproducibility} \propto \frac{1}{\text{copy paste}}$$

Is the juice worth the squeeze?