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

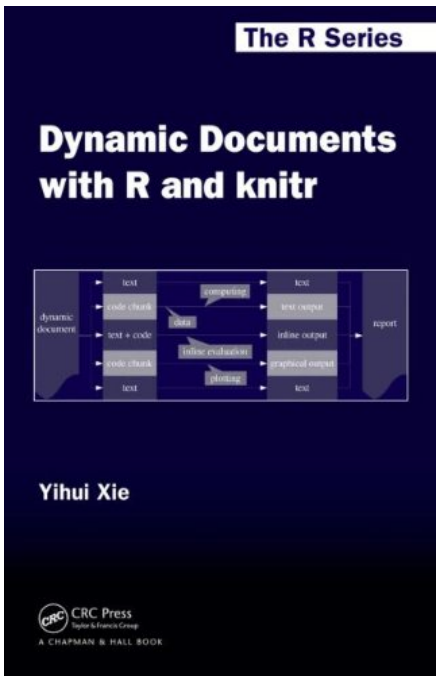
- Ceci est une **introduction** au outils pour la recherche reproductible
- Le niveau de cet atelier est ”intermédiaire” et suppose des connaissances de base en R ainsi que de l’environnement RStudio
- N’hésitez pas à posez des questions

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

Recherche Reproductible (RR)

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**...

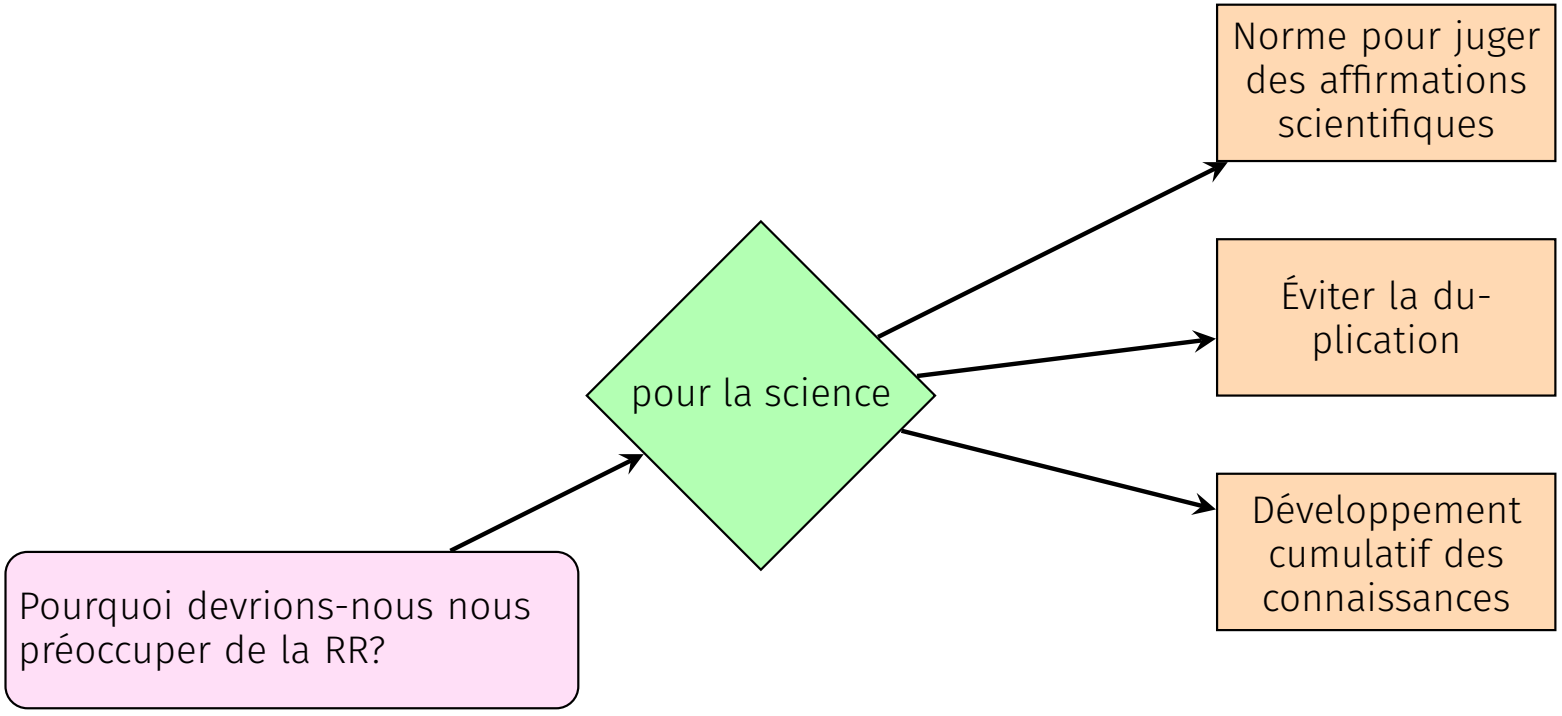
Le **succès et la crédibilité de la science** sont fondés sur la volonté des scientifiques **d'exposer leurs idées** et **leurs résultats** à des **tests indépendants** et à leur **reproduction** par d'autres scientifiques.

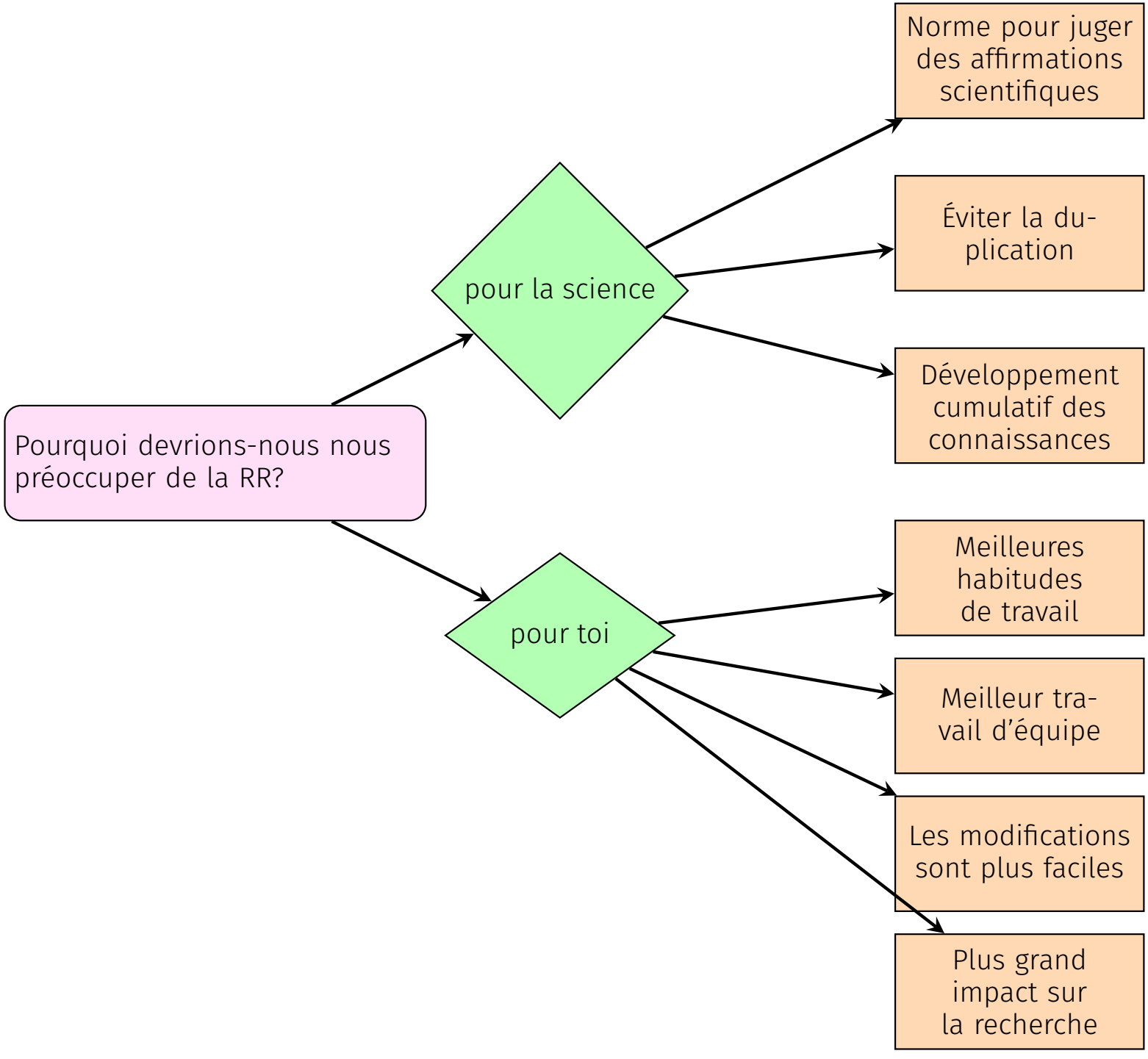
RR: Une norme minimale pour vérifier les résultats scientifiques

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

Démo: 001-motivating-example

Commencer

Outils pour la recherche reproductible[2]

Logiciel gratuit et « open source »

- **RStudio**: Créer, gérer, compiler des documents
- LaTeX: langage de balisage pour la composition d'un document
- **R**: Langage d'analyse statistique
- **knitr**: Intègre le code LaTeXet le code **R**. La version moderne de Sweave du professeur Friedrich Leisch
- **RMarkdown**: Intègre le code Markdown et le code **R**

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

Comparaison

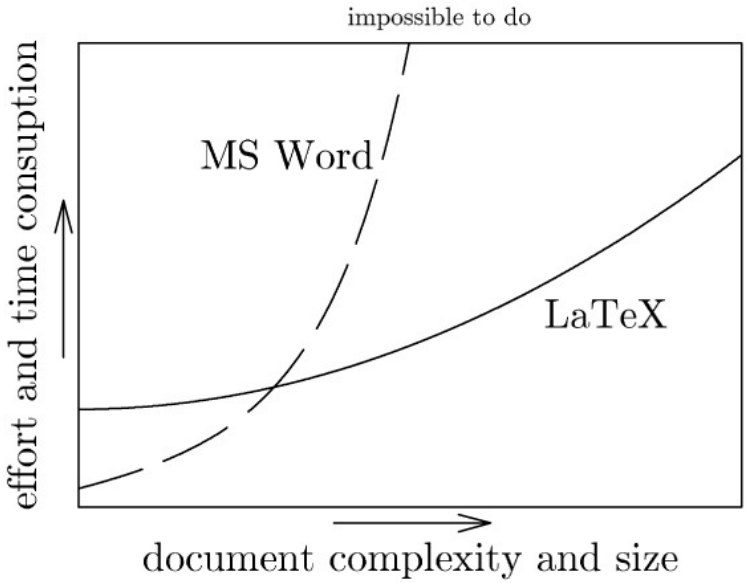


Figure 1: Comparison

- LaTeX a une plus grande courbe d'apprentissage
- De nombreuses tâches sont très difficiles ou impossibles (la plupart des cas) à effectuer dans MS Word ou Libre Office

La philosophie derrière LaTeX

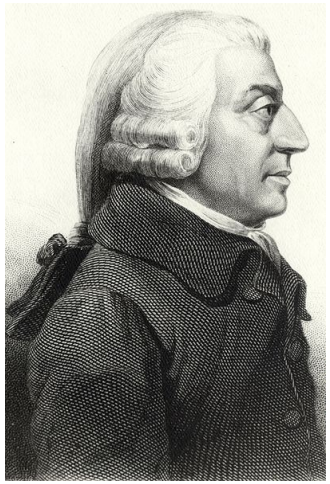


Figure 2: Adam Smith, l'auteur de *The Wealth of Nations* (1776), dans lequel il conceptualise la notion de division du travail

Division du travail

La **composition** et la structuration logique du texte constituent la contribution spécifique de **l'auteur** à la production d'un texte imprimé.

La philosophie derrière LaTeX

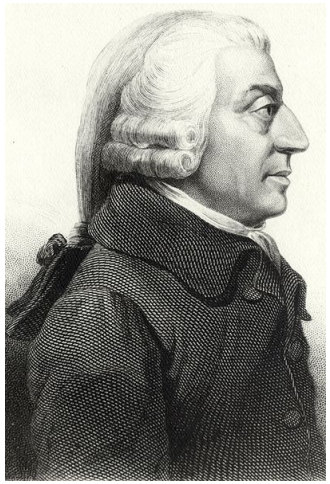


Figure 2: Adam Smith, l'auteur de *The Wealth of Nations* (1776), dans lequel il conceptualise la notion de division du travail

Division du travail

La **composition** et la structuration logique du texte constituent la contribution spécifique de **l'auteur** à la production d'un texte imprimé.

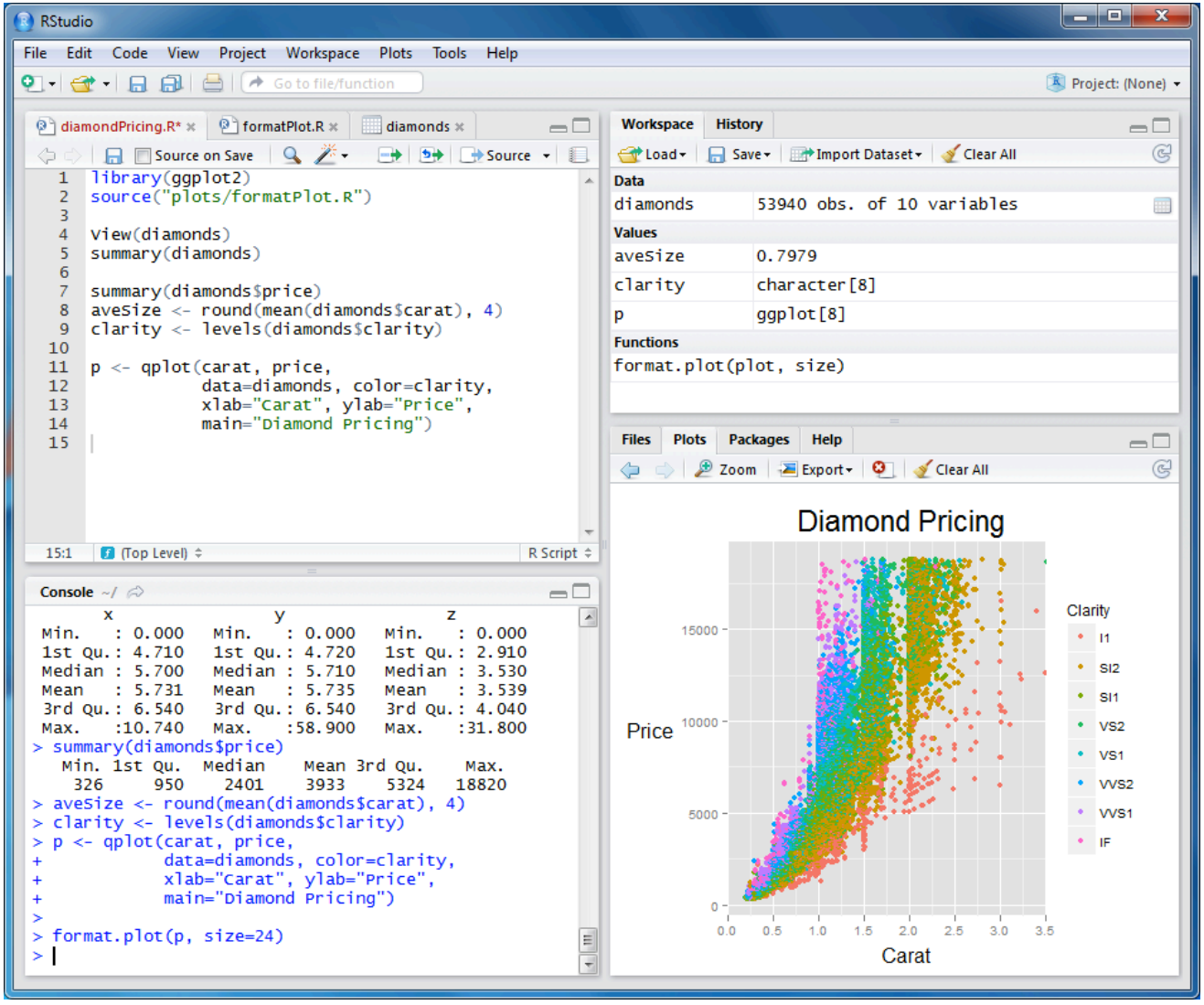
Des questions telles que le choix de la famille de polices, les **en-têtes de section doivent-ils être en caractères gras ou en petites capitales**? Doivent-ils être alignés à gauche ou centrés? Le texte doit-il être justifié ou non? Les notes doivent-elles apparaître au bas de la page ou à la fin? Le texte doit-il être placé dans une colonne ou deux? et ainsi de suite, est l'affaire de la **typographe**

Le génie derrière 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)

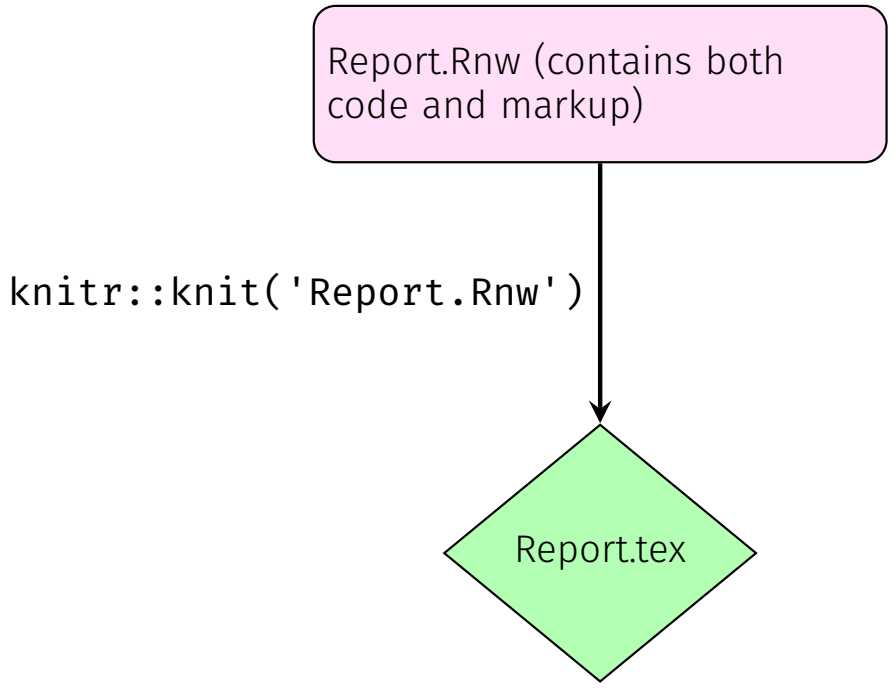
Integrated Development Environment (IDE)



Demonstrate: Explore **RStudio**

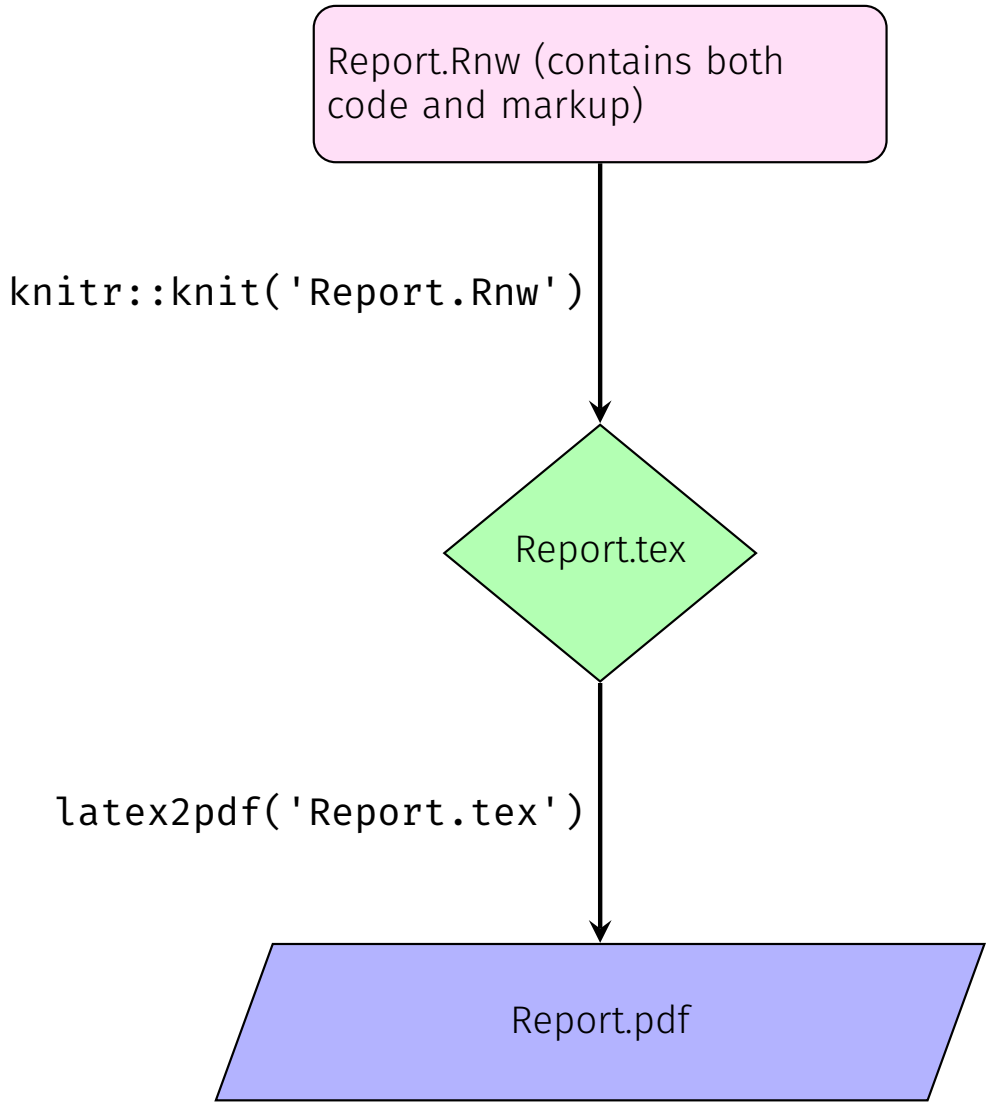
Que fait `knitr`

Exemple LaTeX:



Que fait `knitr`

Exemple LaTeX:

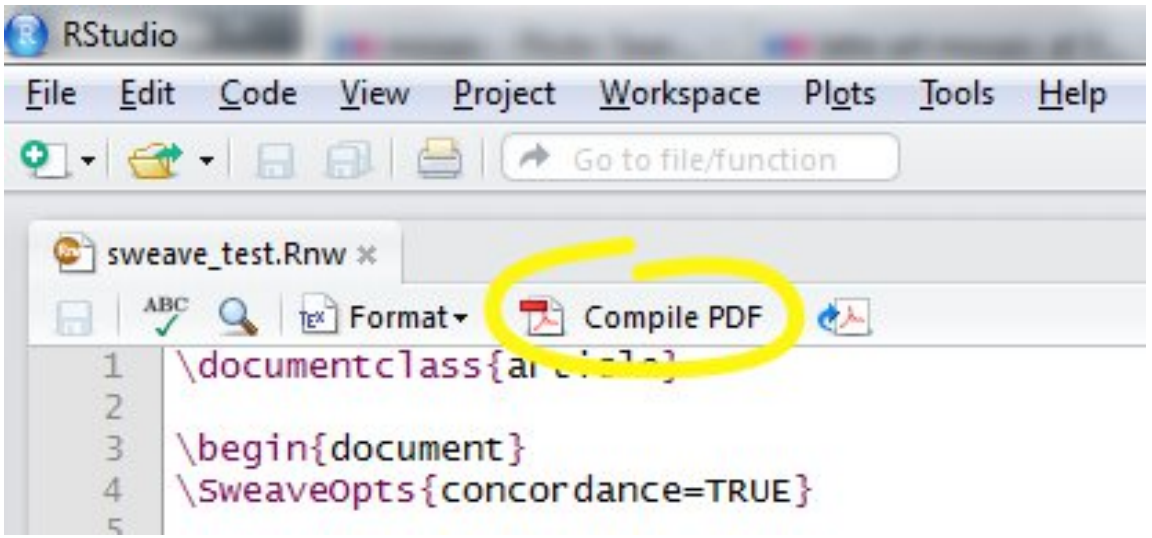


Compiler un document `.Rnw`

Les deux étapes de la diapositive précédente peuvent être exécutées en une seule commande:

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

ou dans `RStudio`:



Incorporer le code R

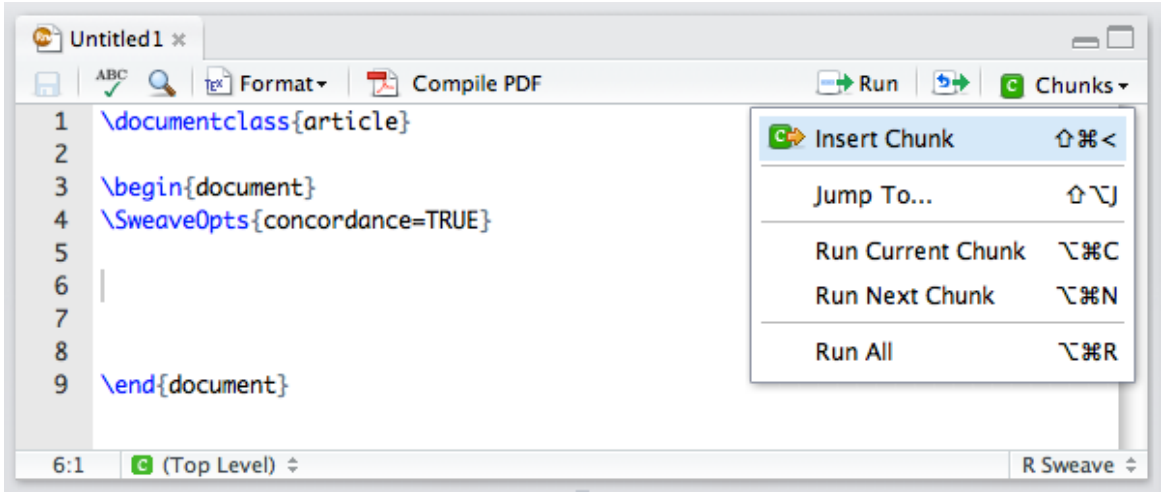
- Insérer le code R dans un **morceau de code** commençant par

<< >>=

et se terminant par

@

Dans RStudio:



Exemple 1

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

produces

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

## [1] 0.12
```

Exemple 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

Démo: Essayez vous-même

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

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

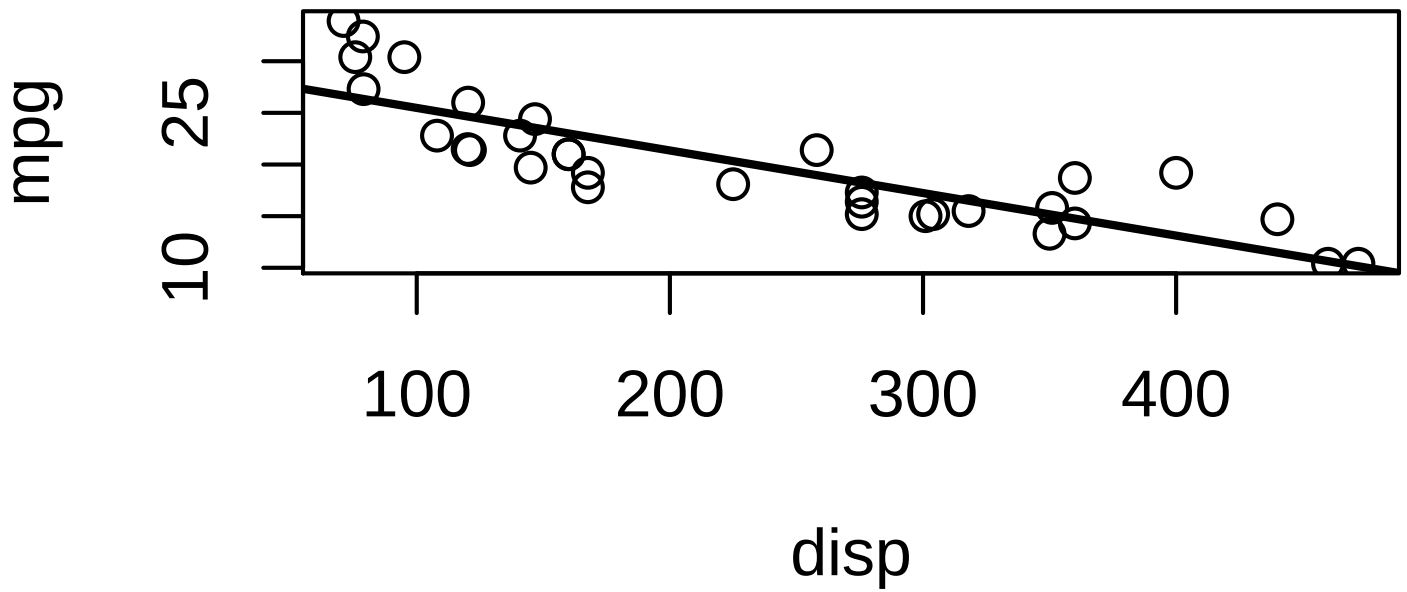


Figure 4: Linear regression

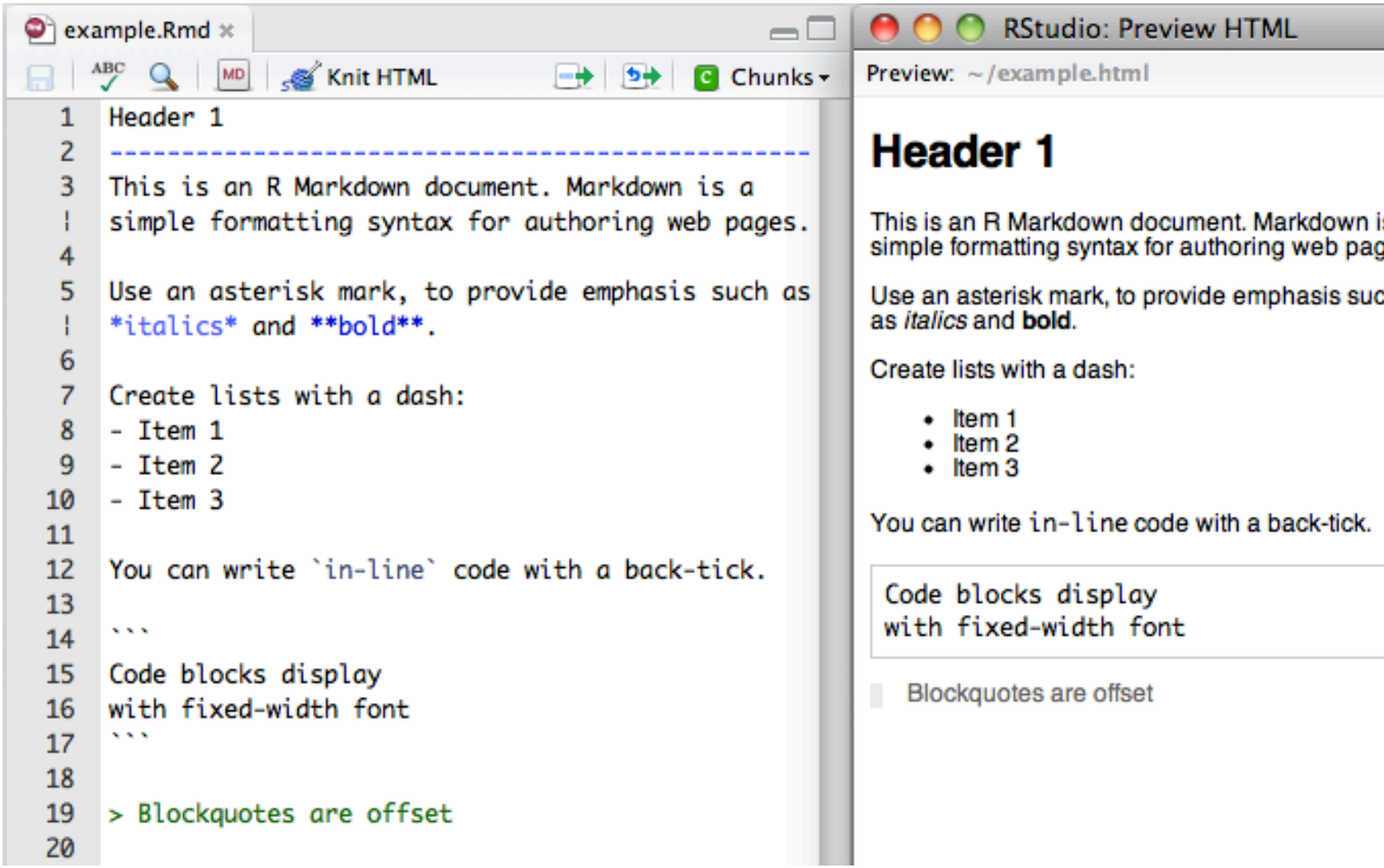
Include a Table

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

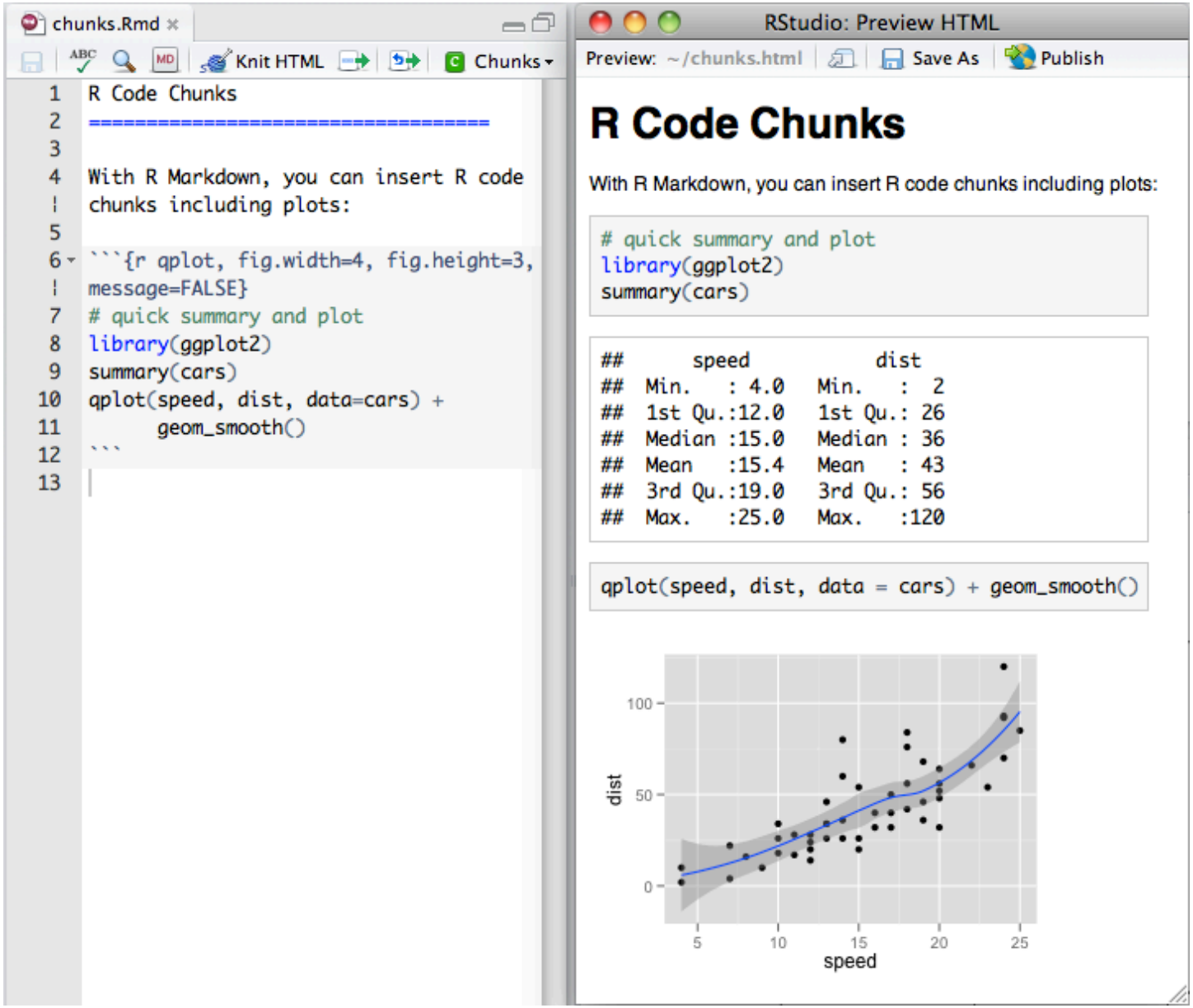
Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.10	3.50	1.40	0.20	setosa
4.90	3.00	1.40	0.20	setosa
4.70	3.20	1.30	0.20	setosa
4.60	3.10	1.50	0.20	setosa
5.00	3.60	1.40	0.20	setosa

Table 1: Sample of Iris data

Markdown: HTML without knowing HTML

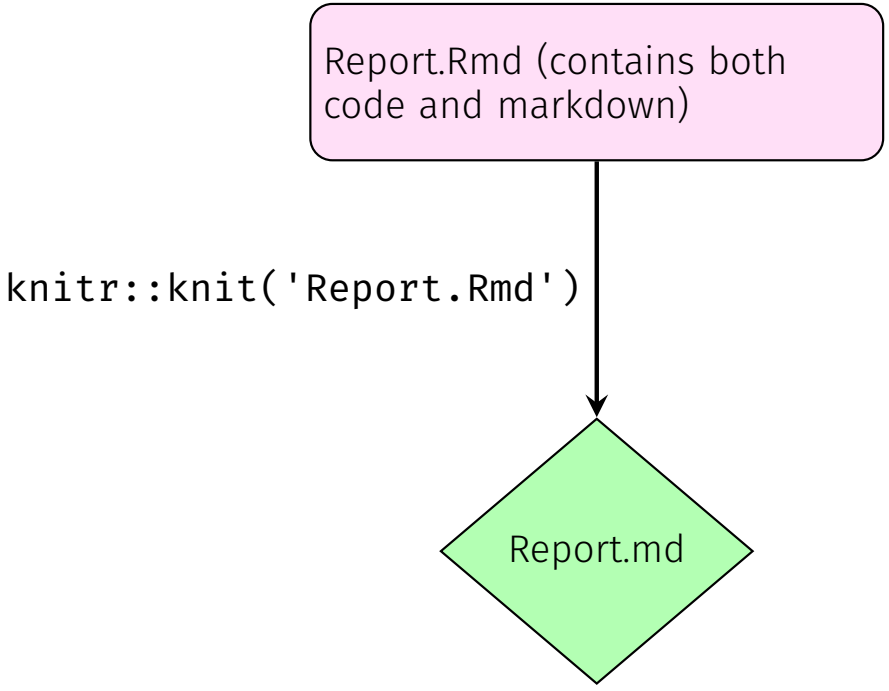


R + Markdown = RMarkdown



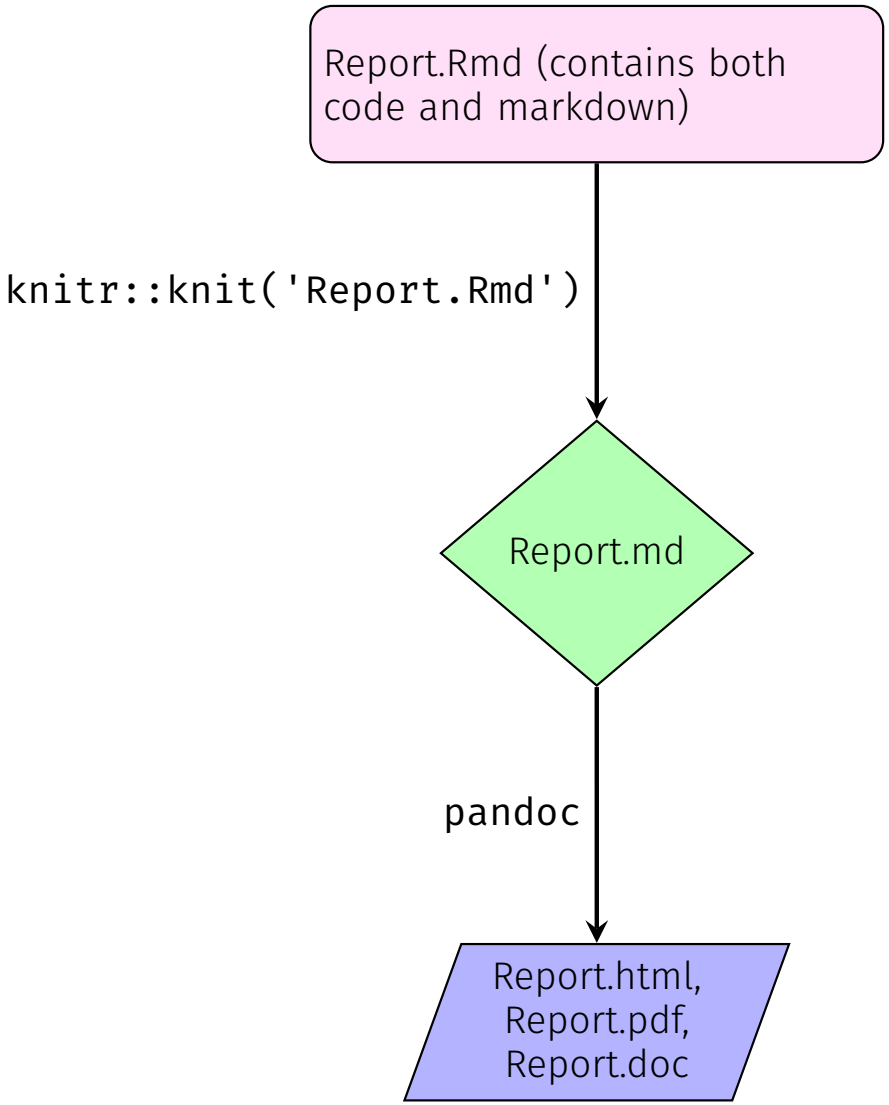
What `rmarkdown` does

`RMarkdown` example:



What `rmarkdown` does

RMarkdown example:

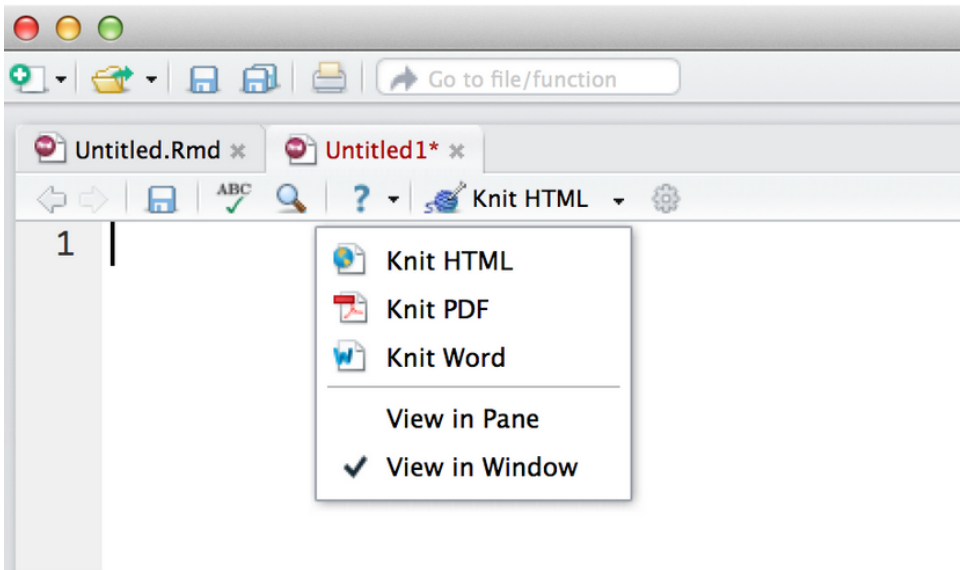


Compiling a `.Rmd` document

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

```
rmarkdown::render()
```

or in RStudio:



How to choose between LaTeX and Markdown ?

LaTeX

- math/stat symbols
- beamer presentations
- customized documents
- publish to journals,
- arXiv

Markdown

- quick and easy reports
- use javascript libraries
- interactive plots
- publish to websites

Exemples

Minimum Working Example

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

Extracting output from Regression Models

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

Figures

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

Beamer Presentations

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

Changing one Parameter in an Analysis

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

Changing Many Parameters in an Analysis

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

Large Documents

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

HTML Reports

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

HTML Presentations

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

Final Remarks

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?