The Division of Labour

https://github.com/sahirbhatnagar/knitr-tutorial

https://sahirbhatnagar.com/CSSC2018

Sahir Bhatnagar, @syfi_24

July 5, 2018

McGill Univeristy
sahirbhatnagar.com

The Division of Labour

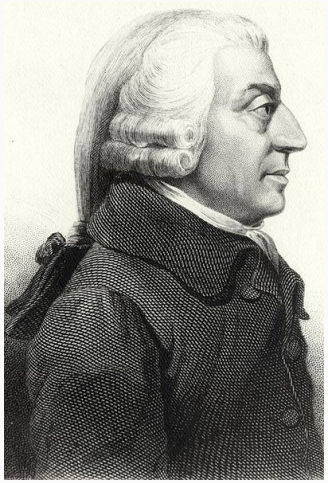


Figure 1: Adam Smith, author of *The Wealth of Nations* (1776), in which he conceptualizes the notion of the division of labour

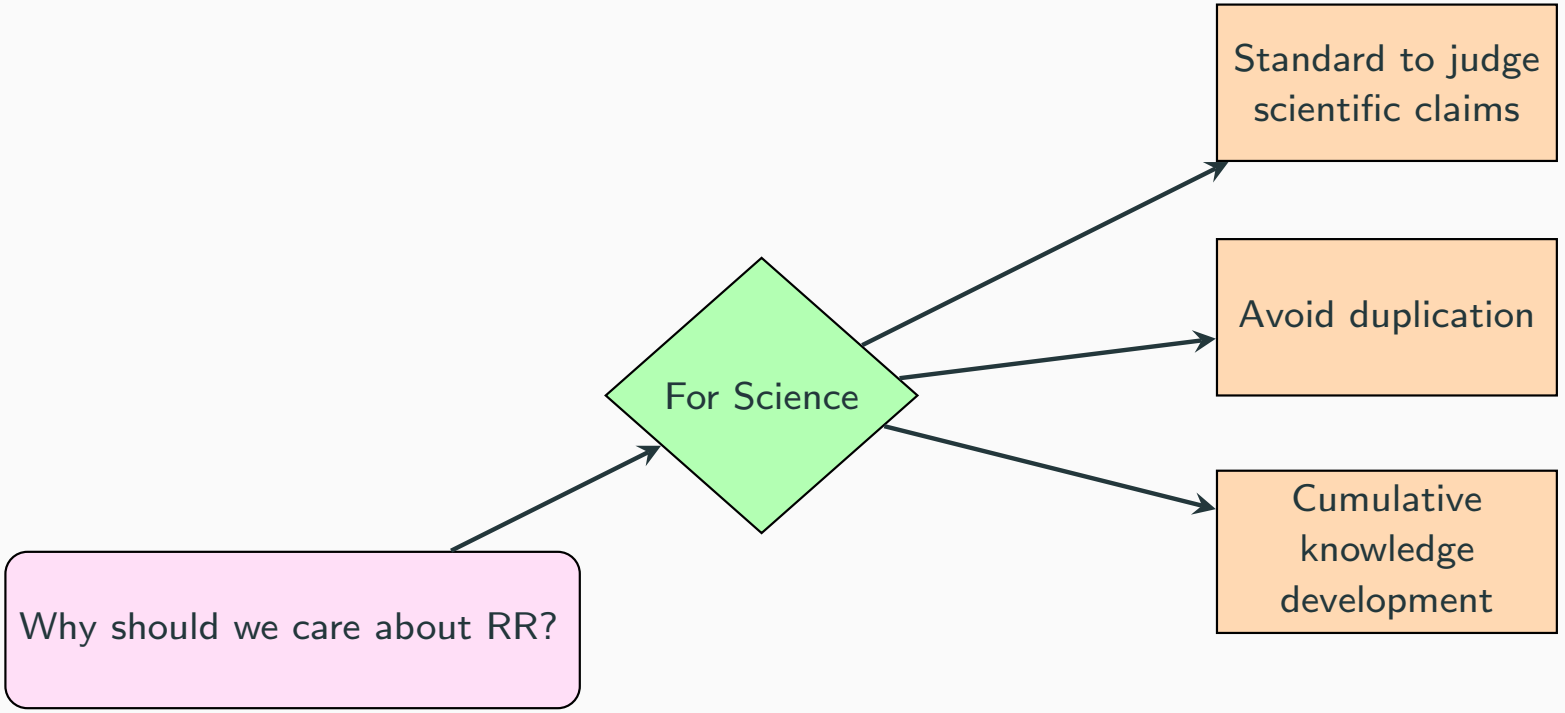
Example: LaTeX

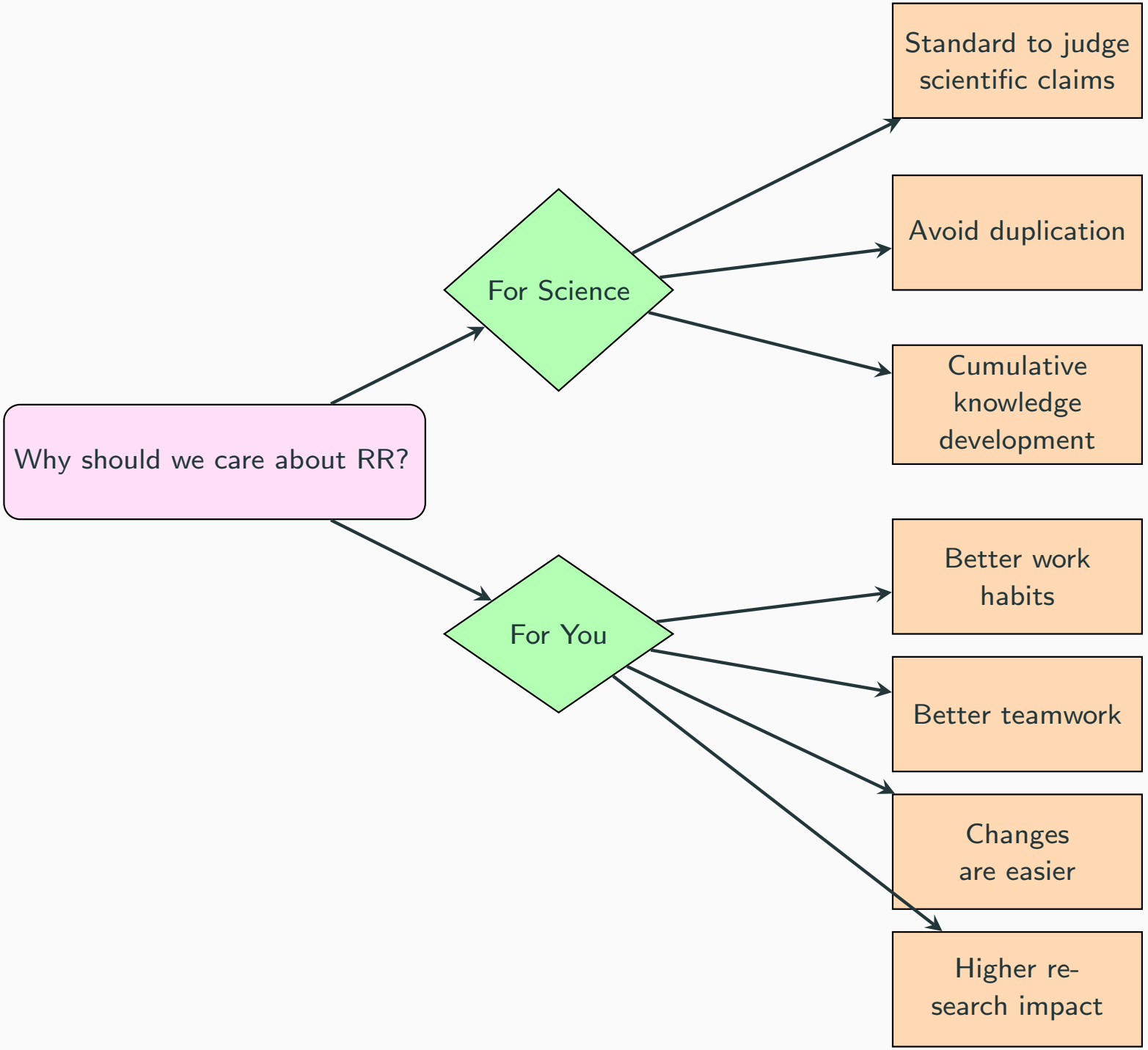
Composition and logical structuring of text is the author's specific contribution to the production of a printed text. Matters such as the choice of the font family, should section headings be in bold face or small capitals? Should they be flush left or centered? Should the text be justified or not?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

Tools for Reproducible Research

Free and Open Source Software

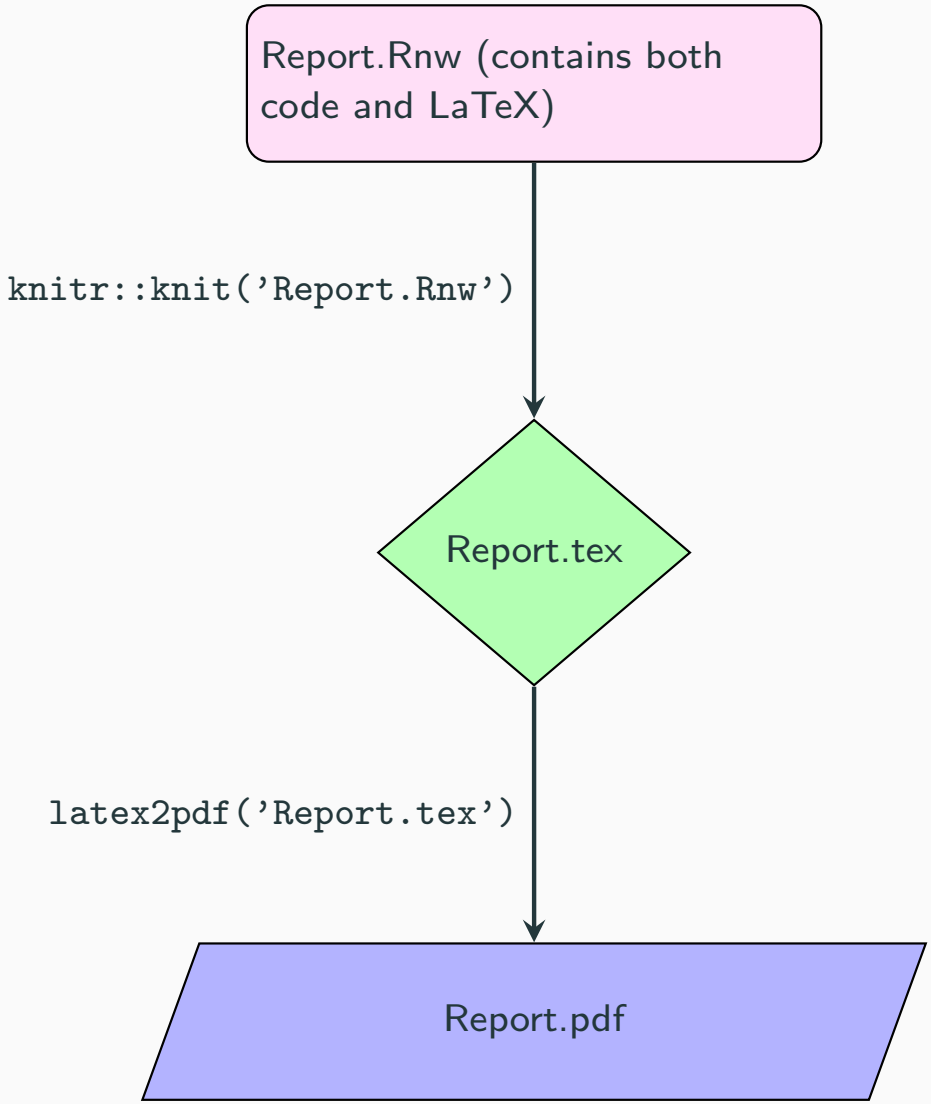
- `RStudio`: Creating, managing, compiling documents
- LaTeX: Markup language for typesetting a pdf
- Markdown: Markup language for typesetting an html
- `R`: Statistical analysis language
- `knitr`: Integrate LaTeXand `R` code. Based on Prof. Friedrich Leisch's `Sweave`


Why should we care about RR?
For Science
Standard to judge scientific claims
Avoid duplication
Cumulative knowledge development


Why should we care about RR?
For Science
For You
Standard to judge scientific claims
Avoid duplication
Cumulative knowledge development
Better work habits
Better teamwork
Changes are easier
Higher research impact

What `knitr` does

LaTeX example:



Example 1: Show code and results

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

produces

```
x <- rnorm(50)
mean(x)

## [1] 0.12
```

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')])
fit <- lm(mpg ~ disp , data = mtcars)
abline(fit,lwd=2)
@
```

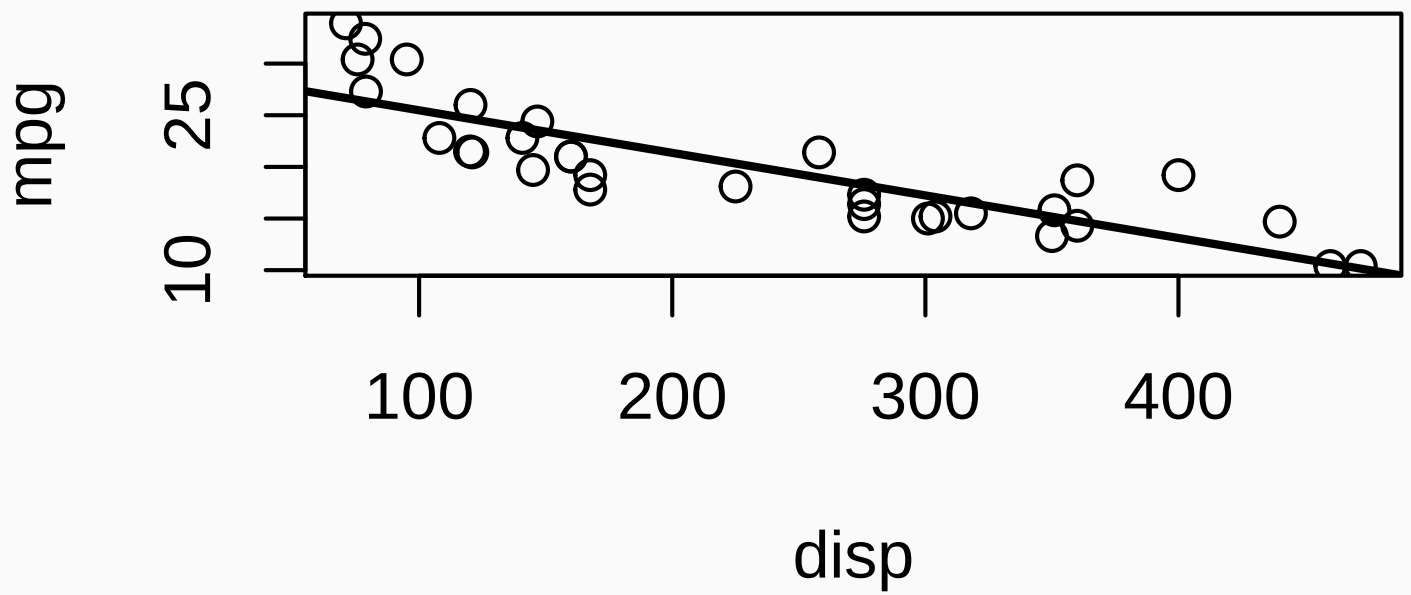


Figure 2: Linear regression

Include a Table

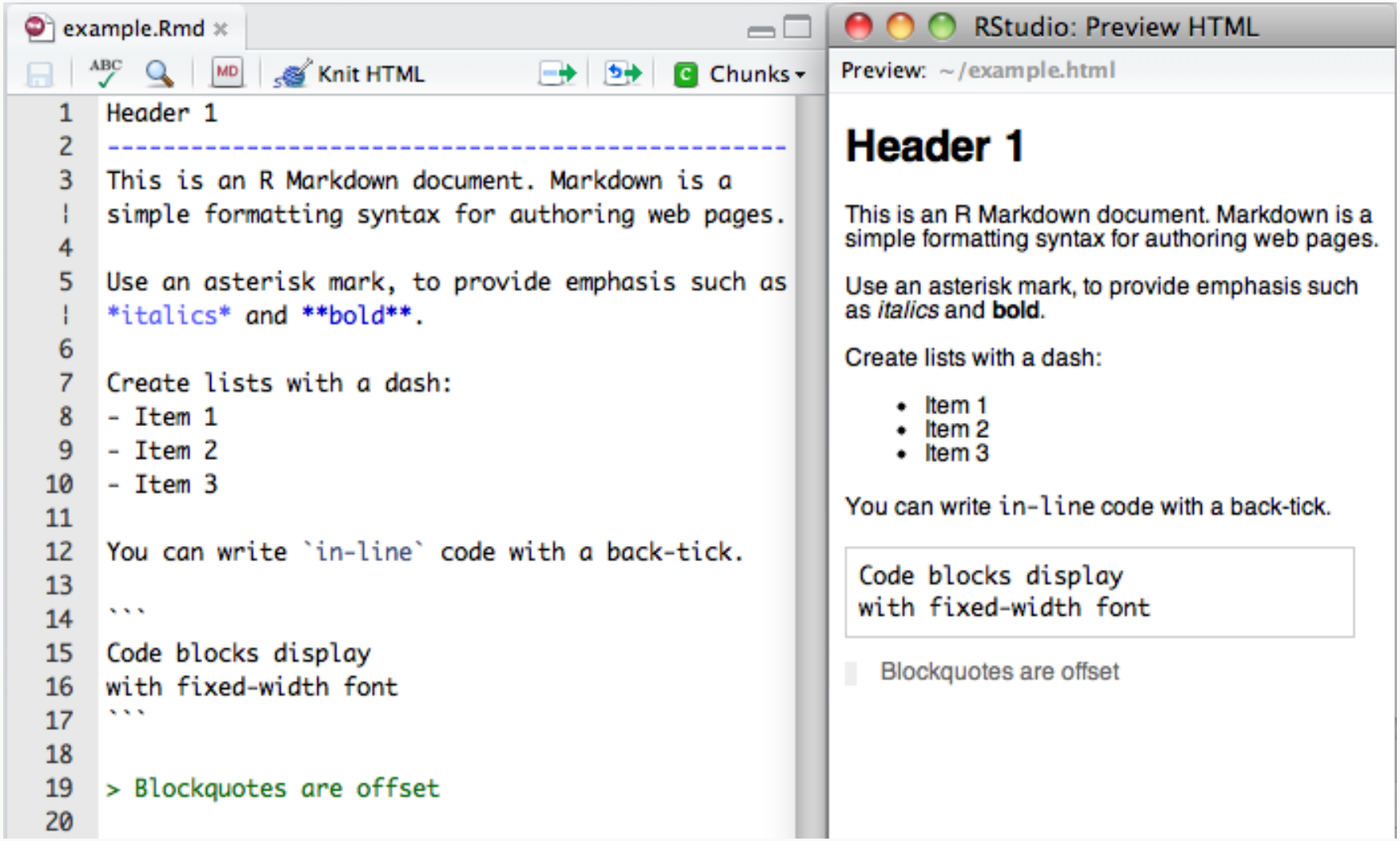
```
<<table.ex, results='asis'>>=
library(xtable)
tab <- xtable(iris[1:5,1:5],caption='Sample of Iris data')
print(tab, include.rownames=FALSE)
@
```

```
library(xtable)
tab <- xtable(iris[1:5,1:5], caption = 'Sample of Iris data')
print(tab, include.rownames = F)
```

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

chunks.Rmd

Knit HTML | Chunks

```
1  R Code Chunks
2  ==========
3
4  With R Markdown, you can insert R code
   chunks including plots:
5
6  ```{r qplot, fig.width=4, fig.height=3,
   message=FALSE}
7  # quick summary and plot
8  library(ggplot2)
9  summary(cars)
10 qplot(speed, dist, data=cars) +
11       geom_smooth()
12 ```
13
```

RStudio: Preview HTML

Preview: ~/chunks.html | Save As | Publish

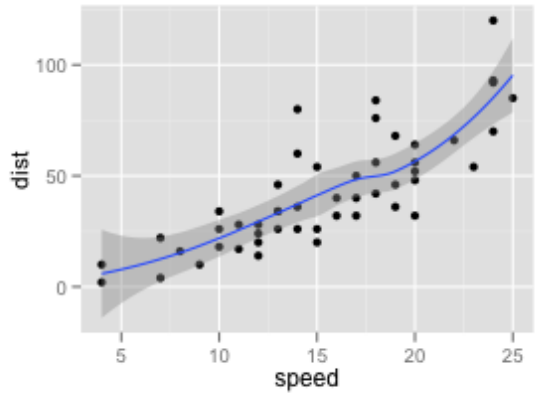
R Code Chunks

With R Markdown, you can insert R code chunks including plots:

```
# quick summary and plot
library(ggplot2)
summary(cars)
```

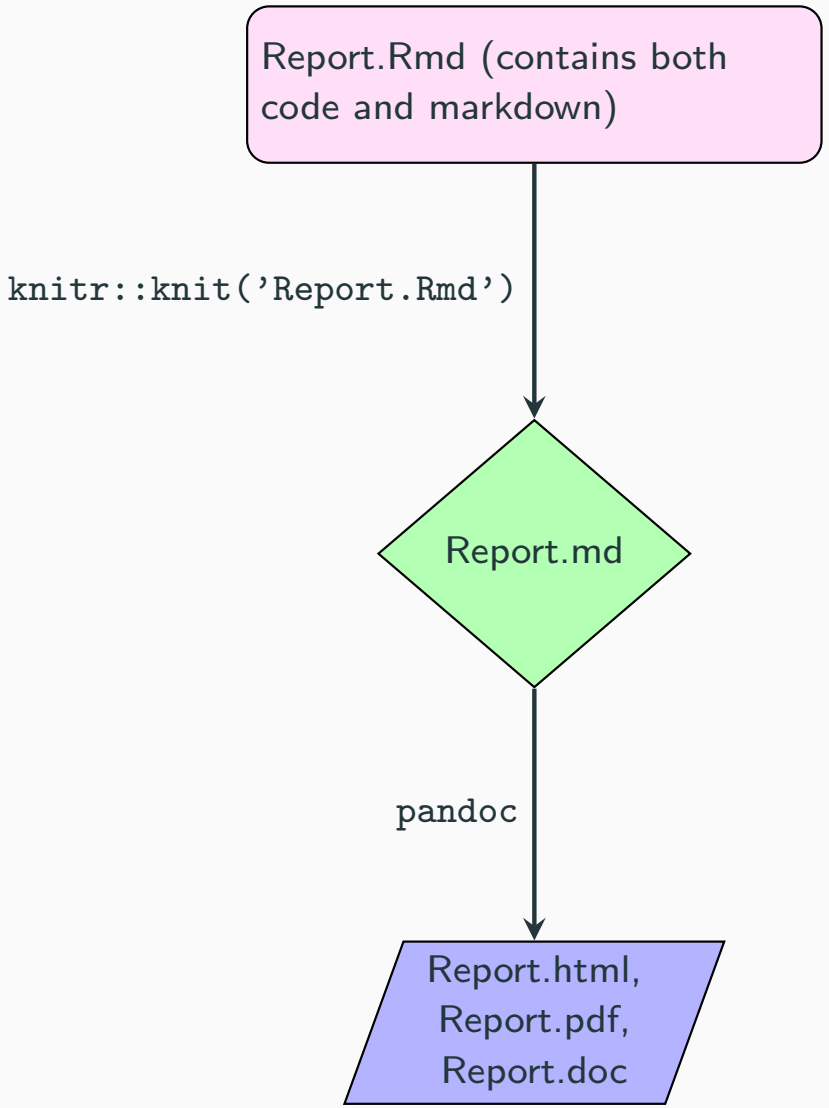
```
##      speed           dist
##  Min.   : 4.0   Min.   :  2
##  1st Qu.:12.0   1st Qu.: 26
##  Median :15.0   Median : 36
##  Mean   :15.4   Mean   : 43
##  3rd Qu.:19.0   3rd Qu.: 56
##  Max.   :25.0   Max.   :120
```

```
qplot(speed, dist, data = cars) + geom_smooth()
```



What `rmarkdown` does

`RMarkdown` example:



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

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.

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