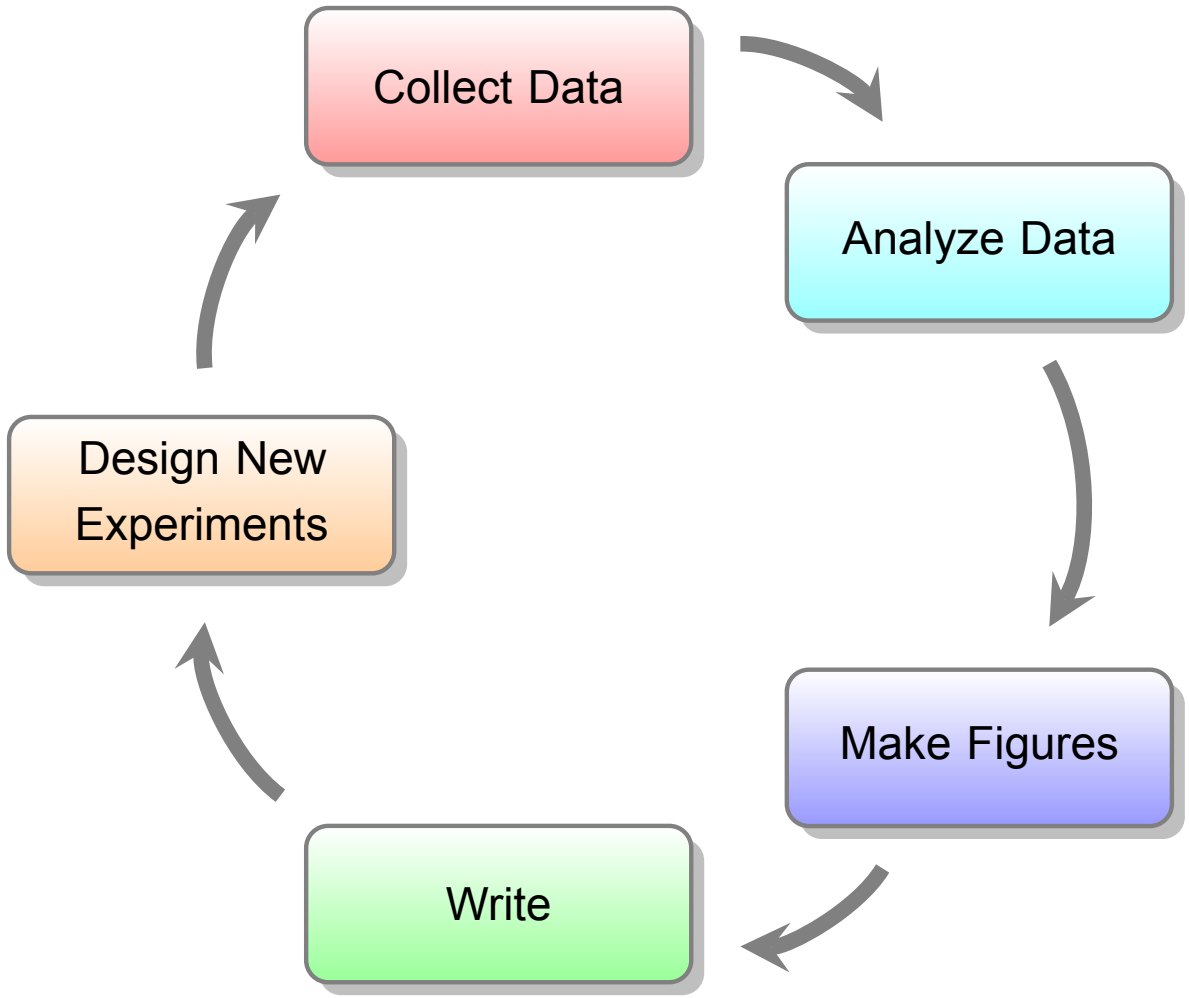
Presenting Data with R Markdown

PHAR G8012: Statistics for the Basic Sciences

Jared Sampson

2/20/2018

A typical scientific workflow



Things we shouldn't have to do when writing a paper

- Scour over a data table to make sure it matches the latest version of the Excel worksheet where the calculations were done, because we can't copy it again or we'd ruin the formatting.
- Completely redo the data analysis if we add a couple more samples.
- Re-check the statistics reported in the paper and any labels added to plots (e.g. significance of results, R^2 values) after tweaking the analysis procedure.
- Feel like you're spending more time formatting the paper than writing it.

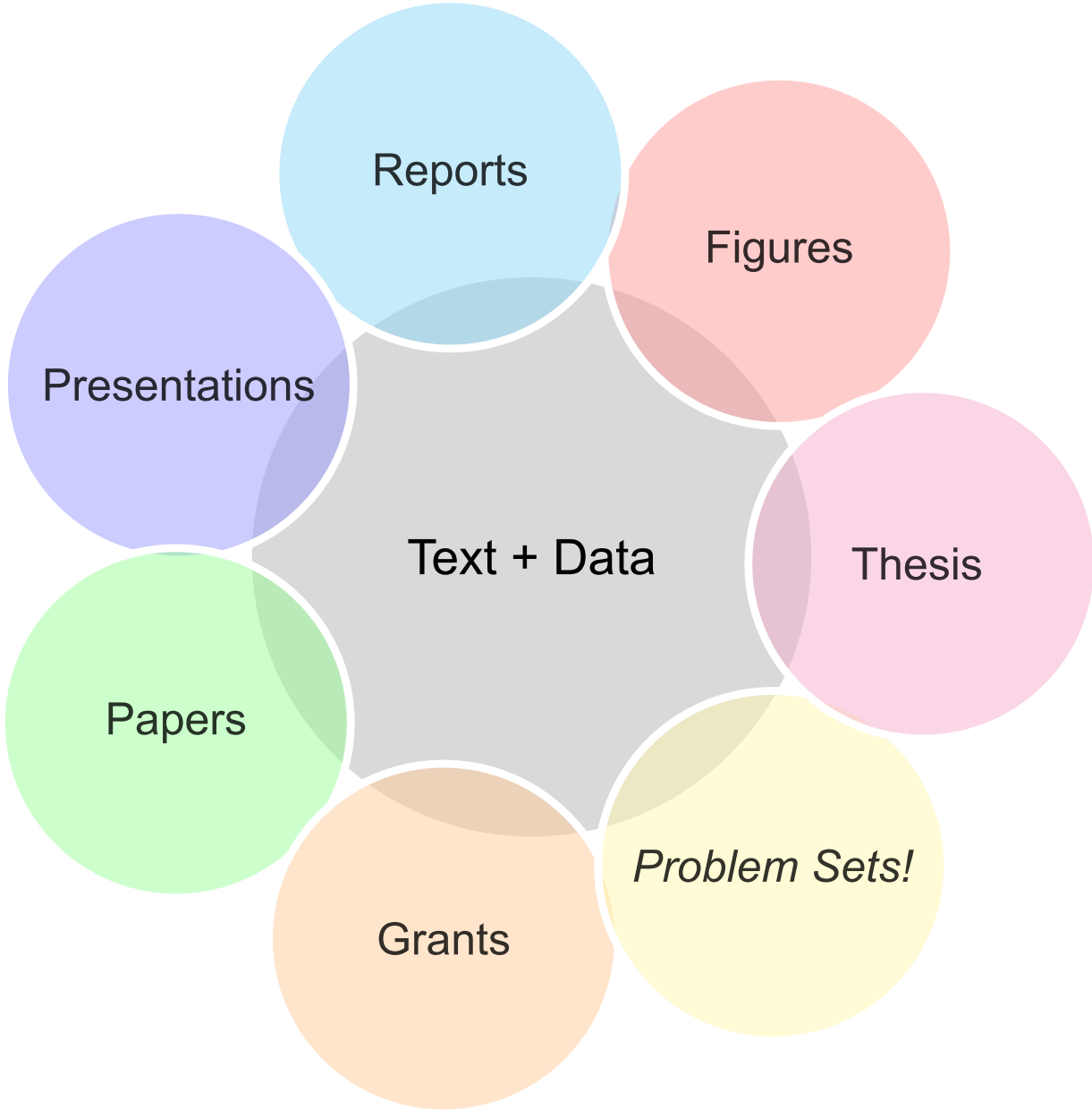
R Markdown

- Plain-text format, use any text editor.
- Embedded R code is evaluated in-place at compile time.
- Go from raw data to final figures in a single step.
- Less copy-pasting/scouring/redoing work you've already done.
- Based on the following projects:
  - Markdown (John Gruber)
  - knitr (Yihui Xie)
  - pandoc (John McFarlane)

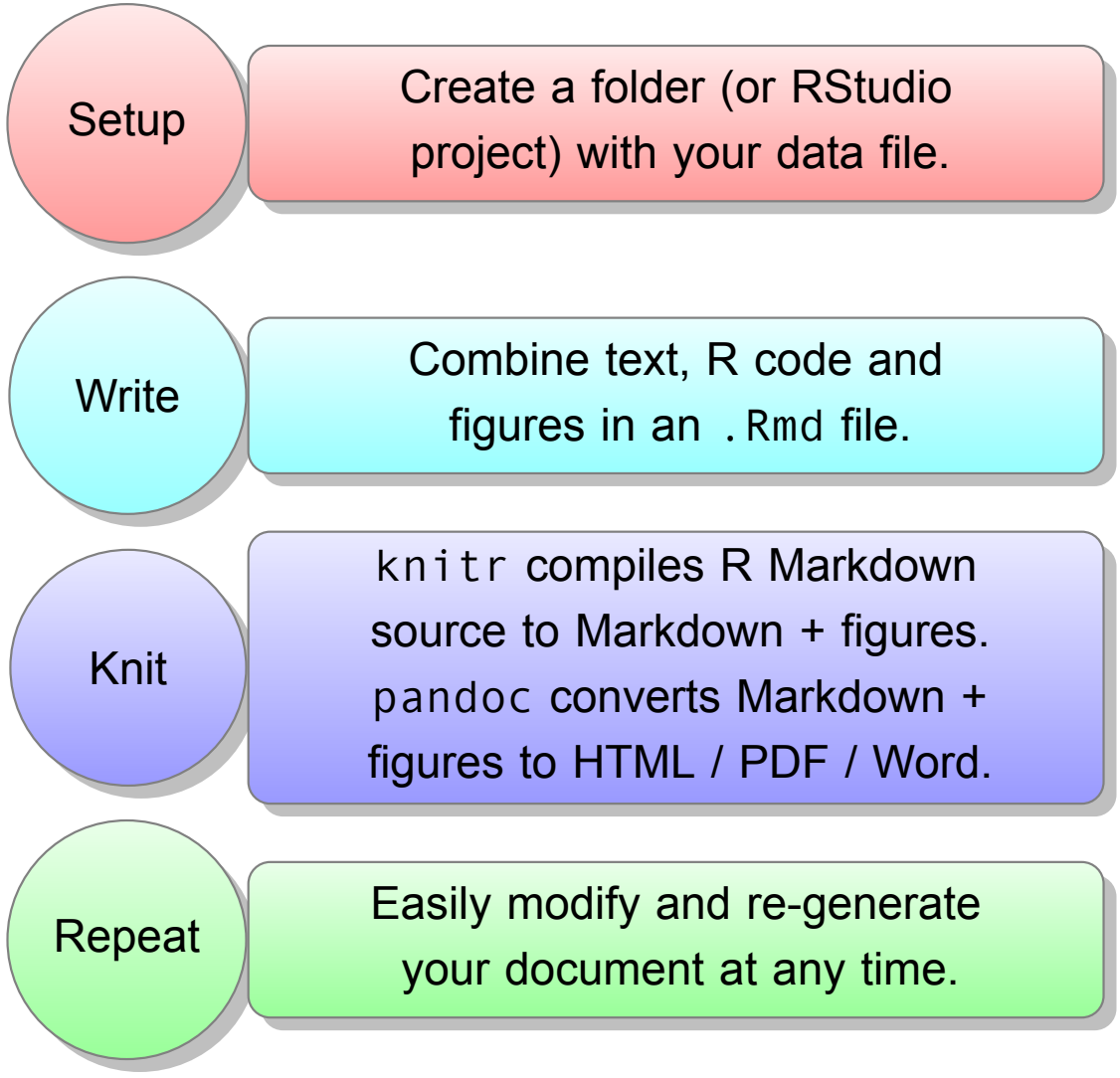
Why use R Markdown?

1. Reproducibility
   - Each analysis step is documented.
   - Source code travels with the text.
2. Automation
   - Save time.
   - Avoid tedious formatting and transcription errors.
3. Durability
   - Plain text is universal.
   - Multiple supported output formats.

What can you do with R Markdown?



General workflow



A simple R Markdown document

````
## A Heading

This is some text in **bold** and *italic*.  With a
[link](http://www.example.com/).

```{r cars}
summary(cars)
```
````

A simple R Markdown document (knitted)

A Heading

This is some text in **bold** and *italic*. With a link.

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

YAML “frontmatter” (header)

```
---
title: "Hello World"
author: "Jared Sampson"
date: "2/20/2018"
output: pdf_document
---
```

- YAML: “a human friendly data serialization standard”.
- Goes at the top between 2 lines with `---`.
- Holds document metadata and `pandoc` output settings.

`knitr` global configuration options

````
```{r setup, include=FALSE}
knitr::opts_chunk$set(echo=TRUE)
```
````

- Set default options to be applied to all code blocks (“chunks”) with `opts_chunk$set(...)`.
- Note the `include=FALSE`: this code does not appear in the document.

Evaluate R code in code blocks or inline

````
First let's do some calculations in a block:

```{r blocks_or_inline}
mean_speed <- mean(cars$speed)
first_100 <- sum(1:100)
```

Then let's explain them in a paragraph.  The mean speed
was `r mean_speed`, and the sum of the first 100
positive integers is `r first_100`.

We can also do calculations directly inline.  For
example, the first 10 numbers of the Fibonacci series
are `r fib <- numeric(10); fib[1] <- 1; fib[2] <- 1;
for (i in 3:10) {fib[i] <- fib[i-2] + fib[i-1]}; fib`.
````

Evaluate R code in code blocks or inline (formatted)

First let's do some calculations in a block:

```
mean_speed <- mean(cars$speed)
first_100 <- sum(1:100)
```

Then let's explain them in a paragraph. The mean speed was 15.4, and the sum of the first 100 positive integers is 5050.

We can, of course, also do calculations directly inline. For example, the first 10 numbers of the Fibonacci series are 1, 1, 2, 3, 5, 8, 13, 21, 34, 55.

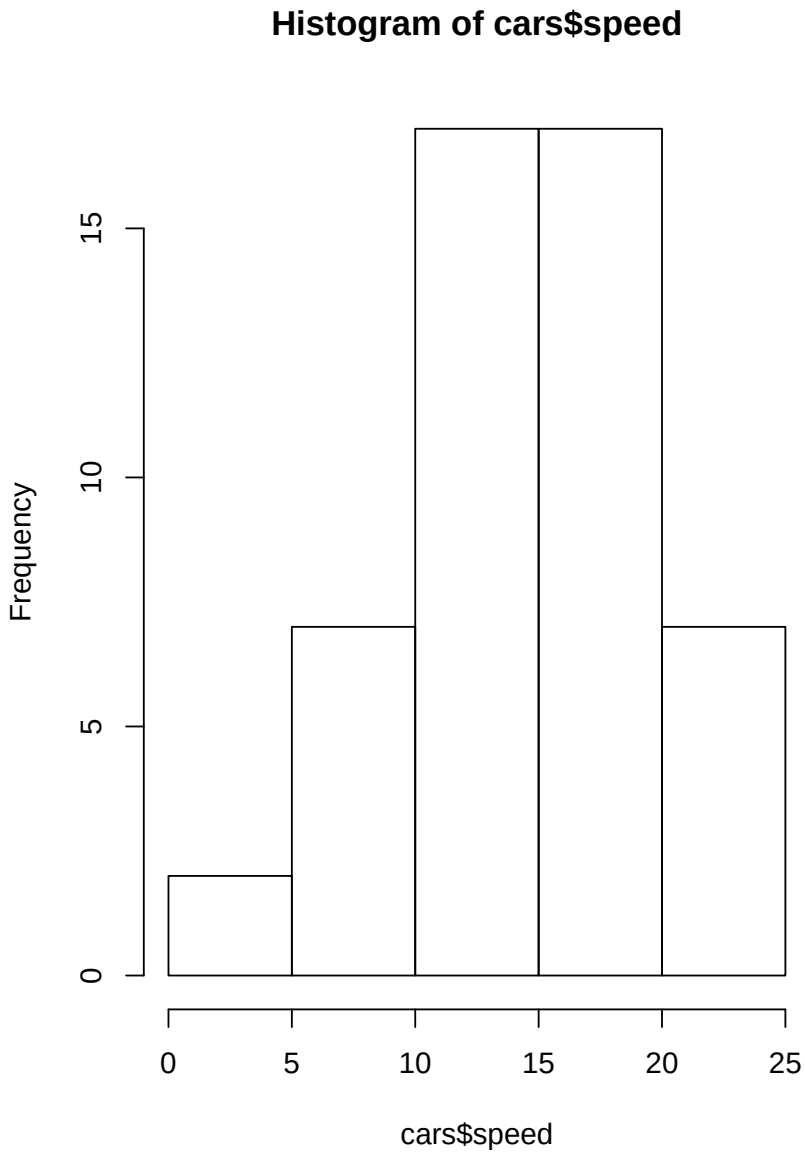
Include plots

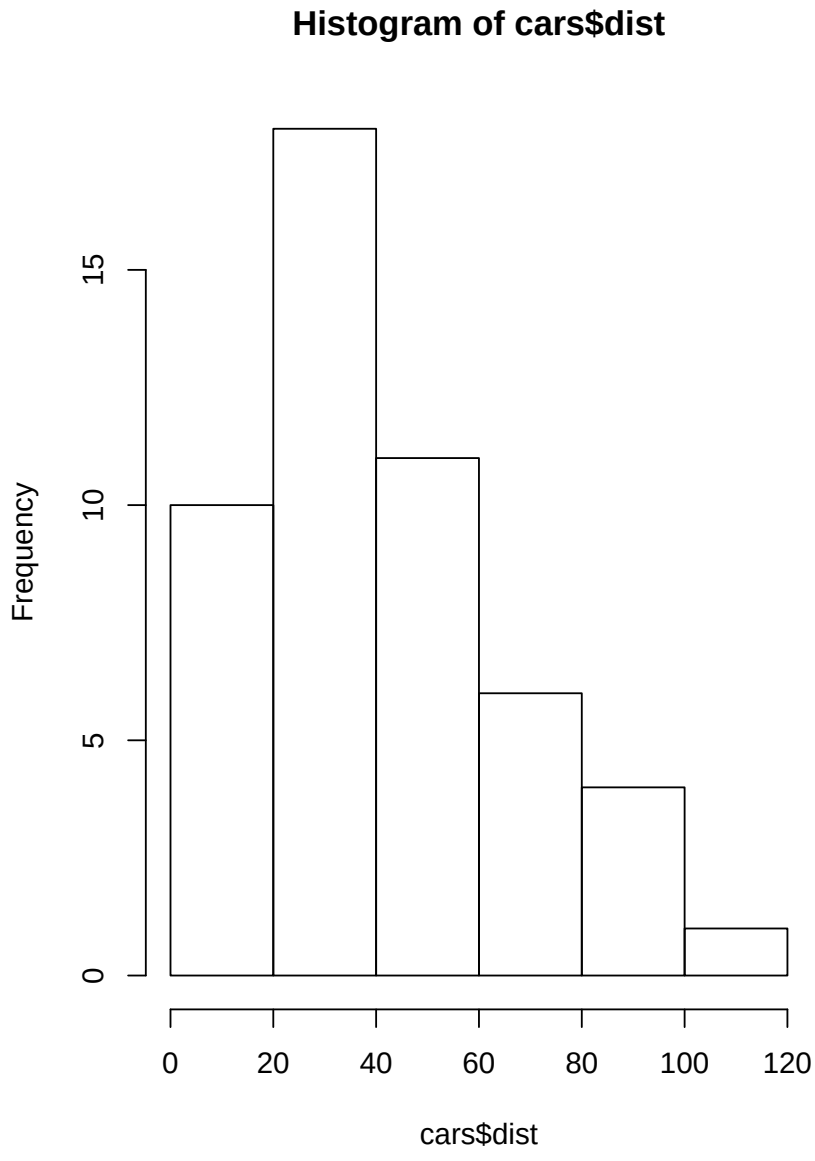
````
Here are 2 histograms from some car data.

```{r hist, echo=FALSE}
par(mfrow=c(1, 2))
hist(cars$speed)
hist(cars$dist)
```
````

Include plots (formatted)

Here are 2 histograms from some car data.





Demo

Let’s dive in!