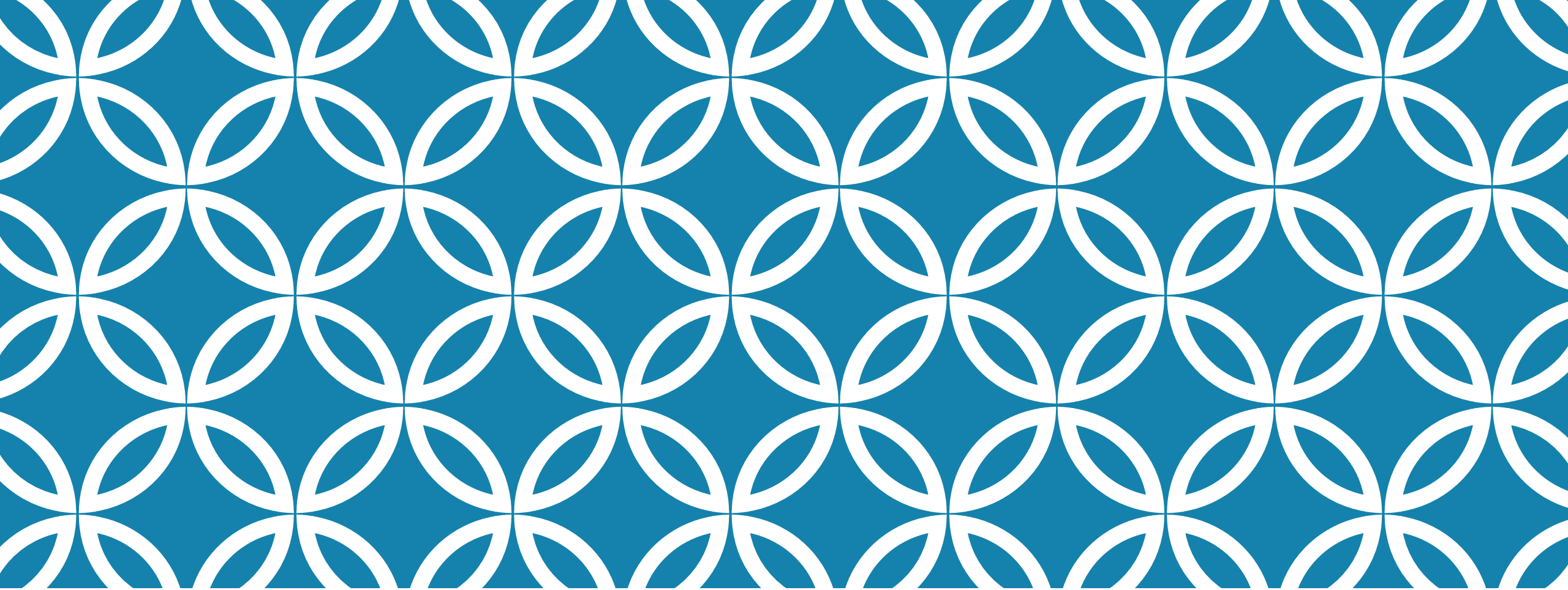


Reproducible Clinical Data Analysis with R and RStudio

Session 1

Introduction

March 27, 2018

Course Structure

Goals

1. Appreciate Reproducibility in Data Analysis
2. Learn a Practical Way to Analyze Clinical Data Reproducibly

Objectives

1. Define “Reproducibility” and Explain its Importance
2. Learn How to Use R/RStudio to Import Data from Files and Databases; Transform Data; and Visualize Data
3. Create a Reproducible Report About Some Aspect of Quality Improvement

Introduction	Early: 3/27, 4 PM Late: 3/27, 5:30 PM
Getting Data	Early: 3/28, 4 PM Late: 3/28, 5:30 PM
Exploring Data	Early: 4/3, 4 PM Late: 4/2, 5:30 PM
Reproducible Reports	Early: 4/5, 4 PM Late: 4/5, 5:30 PM
Course Project Presentations	TBA

Reproducibility

Replication vs Reproduction

- Replication: other people collect new data
  - Scientific gold standard
  - Difficult and time-consuming

- Reproduction: other people analyze the same data
  - Does not by itself validate the analysis …
  - Has been proposed as a minimal standard

The Duke Cancer Scandal

- Chemo sensitivity from microarrays
- Errors first, then misconduct
- Clinical trials based on flawed models
- Papers retracted, lawsuits settled

“Common errors are simple, Simple errors are common”

Theirs	Ours
"1881_at"	"1882_g_at"
"31321_at"	"31322_at"
"31725_s_at"	"31726_at"
"32307_r_at"	"32308_r_at"
...	

https://youtu.be/tN31vZj0LZY

Point-and-Click Is Not Reproducible

- Interactive tools do not record user actions
- Manual documentation is error-prone
- Manual analyses cannot be repeated on new data sets or shared with collaborators

Computer code can precisely document each step of the analysis

Your Turn #1

Complete the section “Reproducibility” on Handout 1.

Spend one minute writing down as many reasons as you can think of for why you might want your **own** data analyses to be reproducible.

Then compare your list with your group mates’.

Why YOU Should Do Data Analysis Reproducibly

"Can we redo the analysis with this month's data?"

"Why do the data in Table 1 not seem to agree with Figure 2?"

"Why did I decide to omit these six samples?"

YOUR CLOSEST COLLABORATOR IS **YOU** FROM 6 MONTHS AGO
(BUT YOU DON'T ANSWER E-MAILS)

The Tools We'll Use

“R” Does Not Stand for “Reproducible”… But It Might As Well

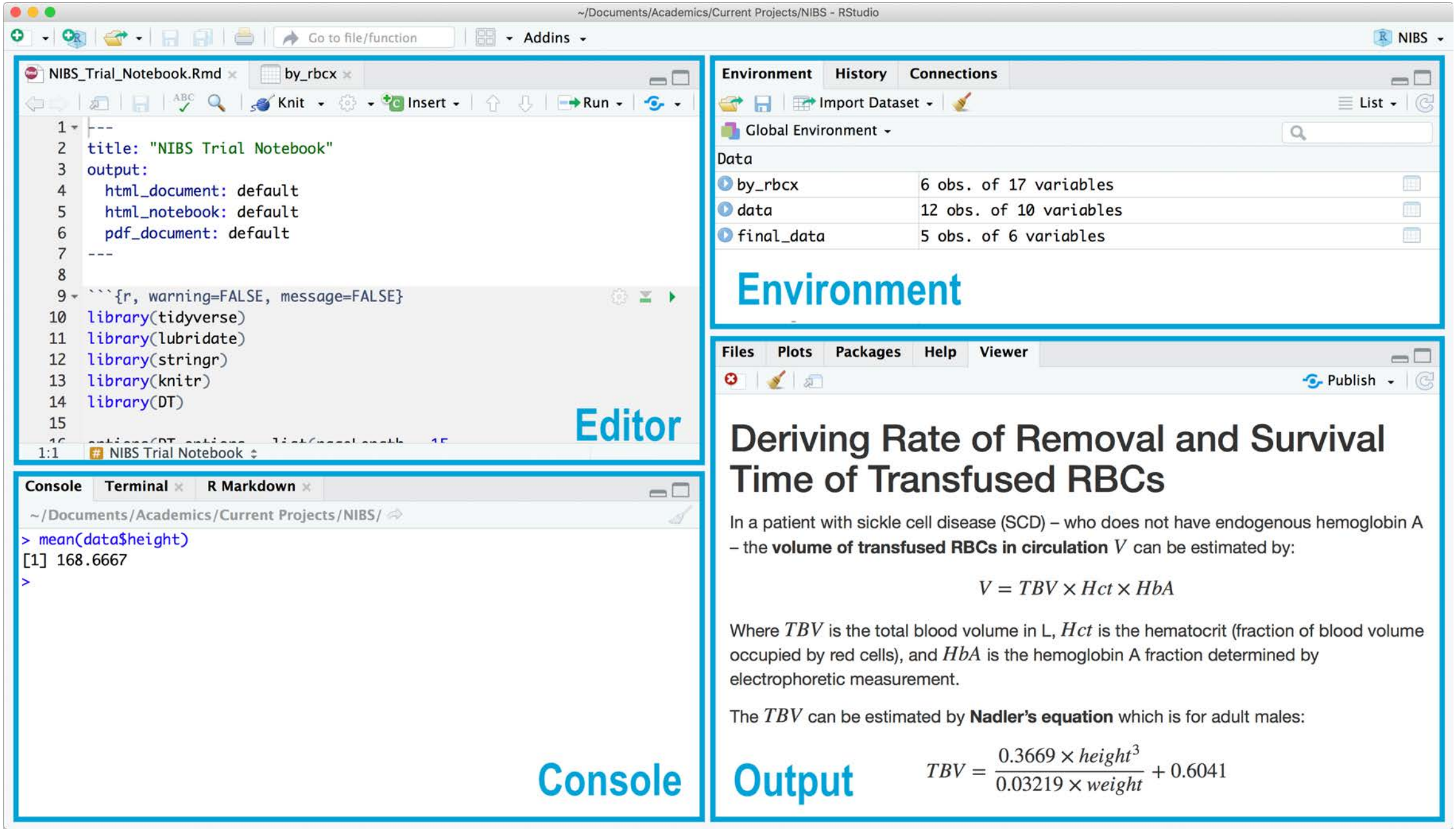
- *RStudio* programming environment
- Concise human-readable code for data transformation and graphics
- Reproducible reporting

RStudio for Developing Data Analyses

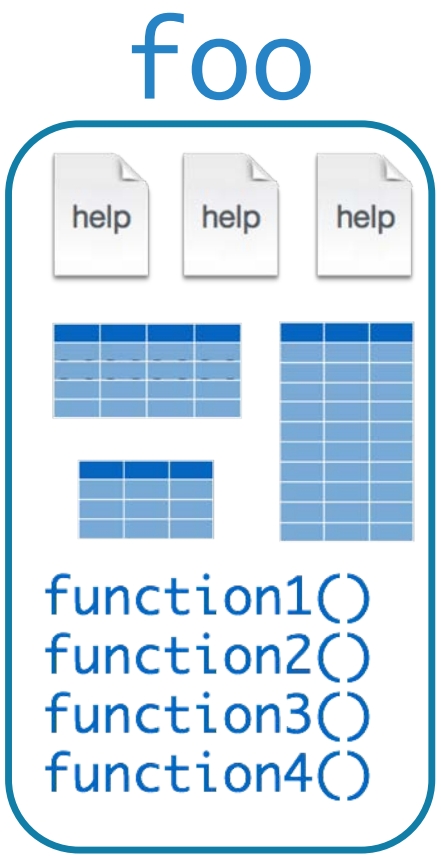
- Free and open-source state-of-the-art integrated development environment

- Tightly integrated with R Markdown for reproducible reporting

- Built-in support for modern machine learning applications including Google CloudML, TensorFlow, and Keras


~/Documents/Academics/Current Projects/NIBS - RStudio
Go to file/function
Addins
NIBS
NIBS_Trial_Notebook.Rmd
by_rbcx
Knit
Insert
Run
1 ---
2 title: "NIBS Trial Notebook"
3 output:
4 html_document: default
5 html_notebook: default
6 pdf_document: default
7 ---
8
9 ```{r, warning=FALSE, message=FALSE}
10 library(tidyverse)
11 library(lubridate)
12 library(stringr)
13 library(knitr)
14 library(DT)
15
1:1
NIBS Trial Notebook
Editor
Environment
History
Connections
Import Dataset
List
Global Environment
Data
by_rbcx
6 obs. of 17 variables
data
12 obs. of 10 variables
final_data
5 obs. of 6 variables
Environment
Files
Plots
Packages
Help
Viewer
Publish
Deriving Rate of Removal and Survival Time of Transfused RBCs
In a patient with sickle cell disease (SCD) – who does not have endogenous hemoglobin A – the volume of transfused RBCs in circulation V can be estimated by:
$$V = TBV \times Hct \times HbA$$
Where TBV is the total blood volume in L, Hct is the hematocrit (fraction of blood volume occupied by red cells), and HbA is the hemoglobin A fraction determined by electrophoretic measurement.
The TBV can be estimated by Nadler's equation which is for adult males:
$$TBV = \frac{0.3669 \times height^3}{0.03219 \times weight} + 0.6041$$
Output
Console
Terminal
R Markdown
~/Documents/Academics/Current Projects/NIBS/
> mean(data$height)
[1] 168.6667
>
Console

A Word About Packages



1

```
install.packages("foo")
```

Downloads files to computer

1 x per computer

2

```
library("foo")
```

Loads package

1 x per R Session

Your Turn #2

Open RStudio and install the `tidyverse` package by typing in the Console:

```
install.packages("tidyverse")
```

Tidyverse for “Tidy” Data Analysis

- A consistent way to organize data
- Human readable, concise, consistent code
- Build pipelines from atomic data analysis steps

R Markdown for Reproducible Reports

- Code remains attached to its documentation and output

- Text and code compile into a single HTML or PDF document that can be shared

- Reproducible reports can be automated and turned into analytic dashboards

```
# One Hashtag = Large Header


## Two Hashtags = Smaller Header

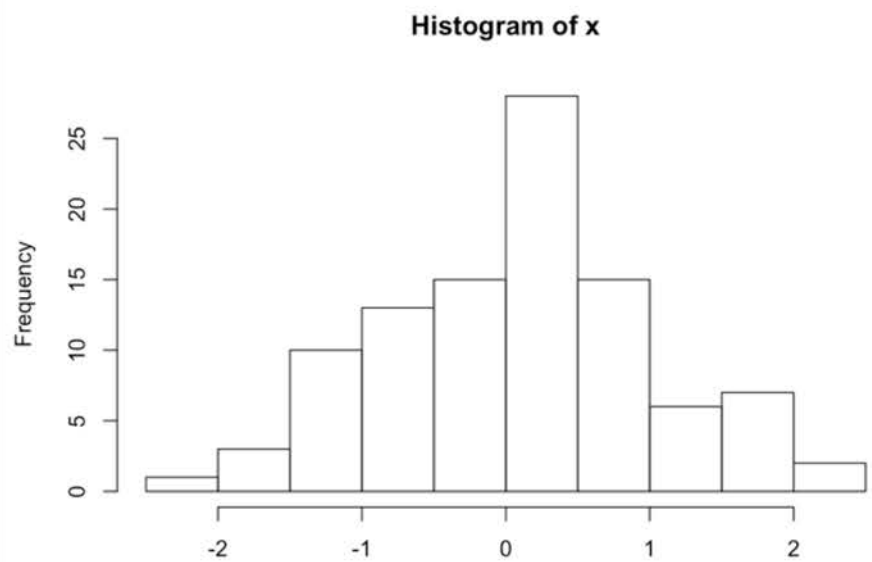
Here is some text.

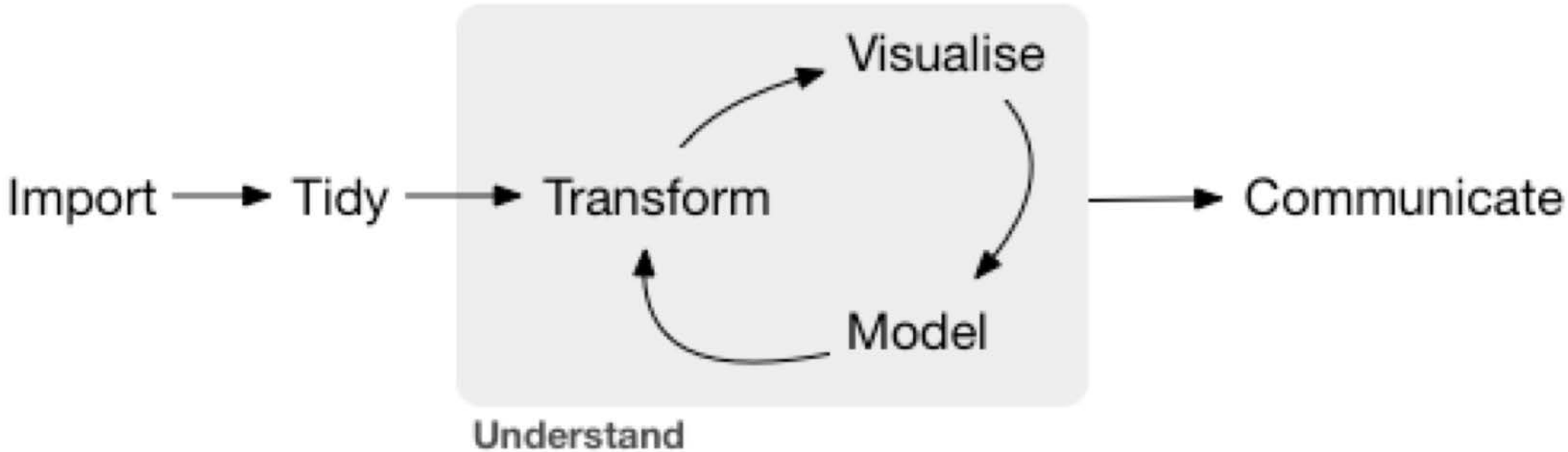
* It's easy to make a list.
* Here's how you style text *cursive* or **bold**.
* Let's add a [link](https://www.massgeneral.org).

```{r}
x <- rnorm(100)
summary(x)
```
```

One Hashtag = Large Header

Two Hashtags = Smaller Header

Here is some text.

- It's easy to make a list.
- Here's how you style text *cursive* or **bold**.
- Let's add a link.

```
x <- rnorm(100)
summary(x)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -2.0053 -0.6084  0.1125  0.0747  0.6100  2.1437
```

```
## Including plots

```{r, echo = FALSE}
hist(x)
```
```

Including plots




Import
Tidy
Transform
Visualise
Model
Communicate
Understand

Your Turn #3

Open `01-introduction.Rmd`.

Read through the R Notebook and do everything it tells you to do.

When you are done, complete the remaining sections of Handout 1.