Intro to R: Part 2

IBIO 851

Sept 8 2016

Suggested reading:

Chapter 2 in **Biostatistical Design & Analysis Using R** (Logan)

Goals for today

- Review key basics
- Importing data
- Viewing data
- Basics of data manipulation
- Reproducibility in science
- Intro to R Markdown

Data types

- R has a wide variety of data types
  - Scalars, vectors, matrices, dataframes, lists
- Vectors

```
a <- c(1,2,5.3,6,-2,4) # numeric
b <- c("one","two","three") # character
d <- c(TRUE,TRUE,TRUE,FALSE,TRUE) #logical
```

  - Refer to elements of a vector using subscripts

```
a[2:4] # 2nd through  4th elements of vector
```

Data types

- Matrices: all columns in a matrix must have the same mode (numeric, character, etc.) and same length

```
mymatrix <- matrix(vector, nrow=r, ncol=c, byrow=FALSE,dimnames=list(char_vector_rownames, char_vector_colnames))
```

  - byrow=TRUE indicates that the matrix should be filled by rows
  - byrow=FALSE indicates that the matrix should be filled by columns (the default)
  - dimnames provides optional labels for the columns and rows

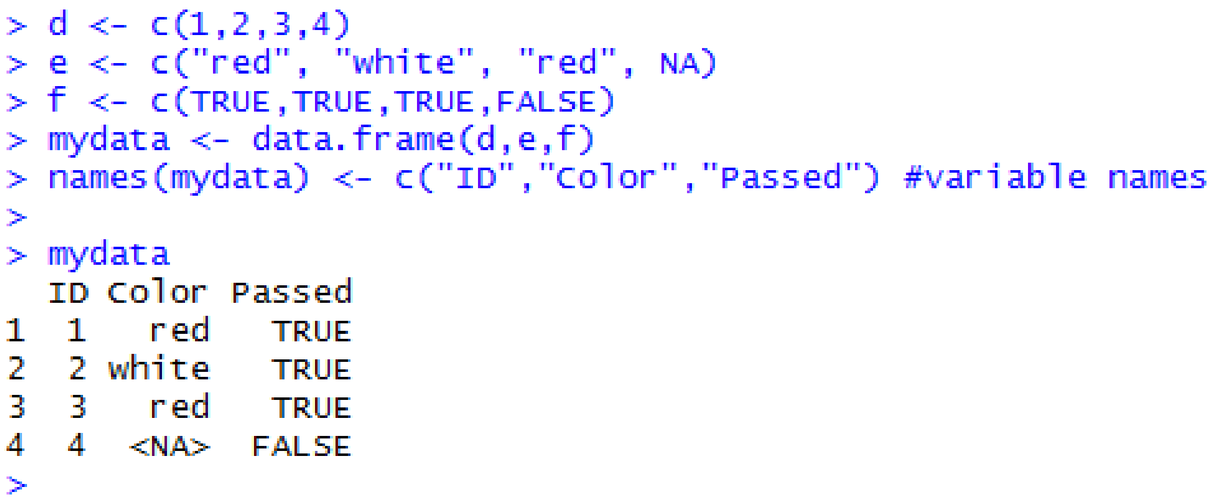
Data types

- Matrices
- Generates 5 x 4 numeric matrix:
  y<-matrix(1:20, nrow=5,ncol=4)
- Identify rows, columns or elements using subscripts.

```
x[,4] # 4th column of matrix
x[3,] # 3rd row of matrix
x[2:4,1:3] # rows 2,3,4 of columns 1,2,3
```

Data types

- Arrays: similar to matrices, but can have more than 2 dimensions
  - See help(array) for details
- Data frames: more general than a matrix in that it can include non-numbers



```
> d <- c(1,2,3,4)
> e <- c("red", "white", "red", NA)
> f <- c(TRUE,TRUE,TRUE,FALSE)
> mydata <- data.frame(d,e,f)
> names(mydata) <- c("ID","Color","Passed") #variable names
>
> mydata
  ID Color Passed
1  1   red   TRUE
2  2 white   TRUE
3  3   red   TRUE
4  4  <NA>  FALSE
>
```

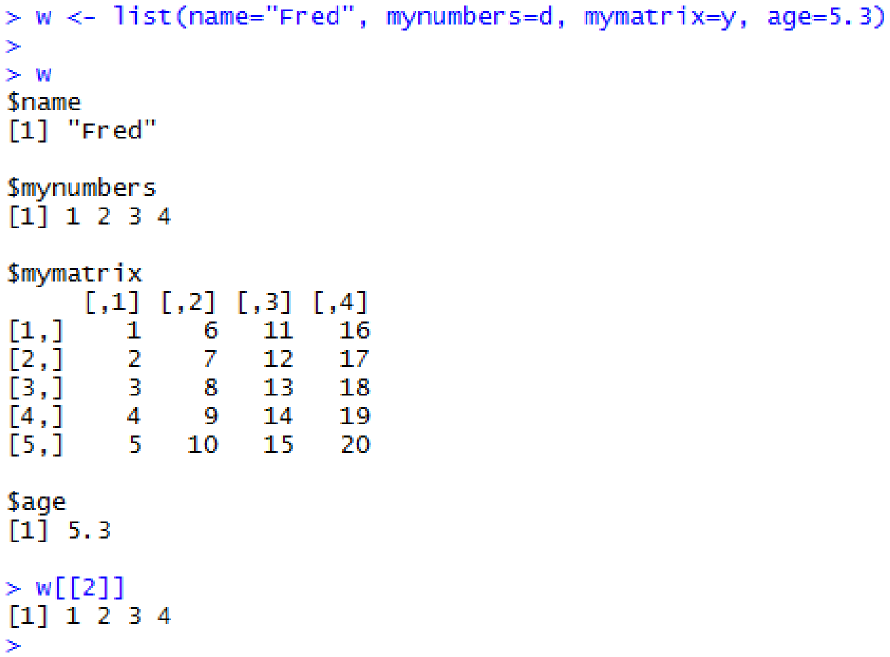
Data types

- Dataframes
  - There are a variety of ways to identify elements of a dataframe

```
myframe[3:5] # columns 3,4,5 of dataframe
myframe[c("ID","Age")] # columns ID & Age
myframe$X1 # variable x1
```

Data types

- Lists: An ordered collection of objects (components)
- A list allows you to gather a variety of (possibly unrelated) objects under one name
- Example of a list with 4 components:



```
> w <- list(name="Fred", mynumbers=d, mymatrix=y, age=5.3)
>
> w
$name
[1] "Fred"

$mynumbers
[1] 1 2 3 4

$mymatrix
     [,1] [,2] [,3] [,4]
[1,]    1    6   11   16
[2,]    2    7   12   17
[3,]    3    8   13   18
[4,]    4    9   14   19
[5,]    5   10   15   20

$age
[1] 5.3

> w[[2]]
[1] 1 2 3 4
>
```

Useful functions

- length(object) #number of elements
- str(object) #structure of an object
- names(object) #names (e.g. variables)
- cbind(object, object) #combine objects as columns
- rbind(object, object) #combine objects as rows
- ls() #list current objects stored in memory

Importing data

- From a comma delimited text file:
  - First row contains variable names, comma is separator

(1) mydata <- read.table("c:/mydata.txt", header=TRUE, sep=",", row.names="id")

(2) mydata <- read.csv("c:/mydata.csv", header=TRUE, sep=",", row.names="id")

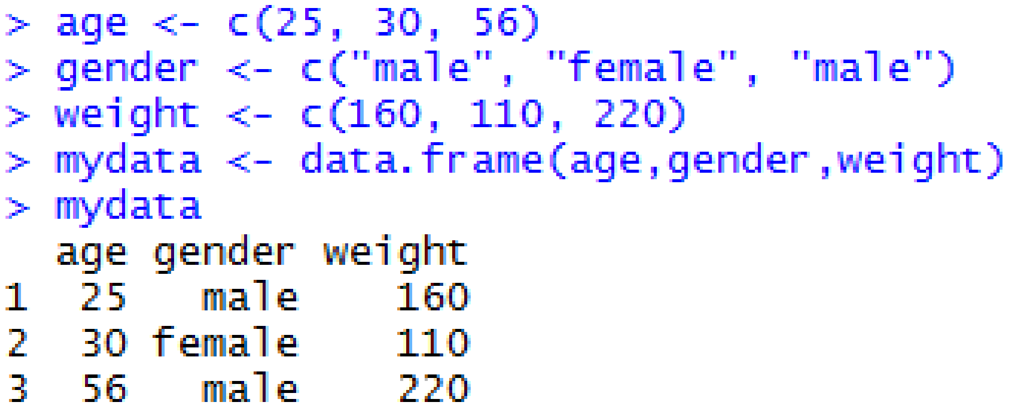
- header=TRUE means that you have column headings in the first row
- sep=“,” means that the data is comma-delimited
- row.names=“id” means that the number of each row is now the ID

Importing data

- From Excel:
  - The best way to read an Excel file is to export it to a comma delimited file and import it using the previous method

Importing data

- Manually in R:
  - Create a dataframe from scratch
  - Only possible with very small datasets



```
> age <- c(25, 30, 56)
> gender <- c("male", "female", "male")
> weight <- c(160, 110, 220)
> mydata <- data.frame(age,gender,weight)
> mydata
  age gender weight
1  25   male    160
2  30 female    110
3  56   male    220
```

Viewing data

- There are a number of functions for listing the contents of an object or dataset
- Here are commonly used ones (some you've seen already):

```
# list objects in the working
    environment
    ls()
# list the variables in mydata
    names(mydata)
# list the structure of mydata
    str(mydata)
# list levels of factor v1 in mydata
    levels(mydata$v1)
# dimensions of an object
    dim(object)
```

Viewing data

```
# class of an object (numeric, matrix, dataframe, etc)
    class(object)
# print mydata (this shows up in the console)
    mydata
# print first 10 rows of mydata
    head(mydata, n=10)
# print last 5 rows of mydata
    tail(mydata, n=5)
```

Data manipulation

- Once you have imported your data, you will want to manipulate it into a useful form
- This includes creating new variables, sorting & merging datasets, aggregating data, reshaping data, and subsetting datasets

Data manipulation

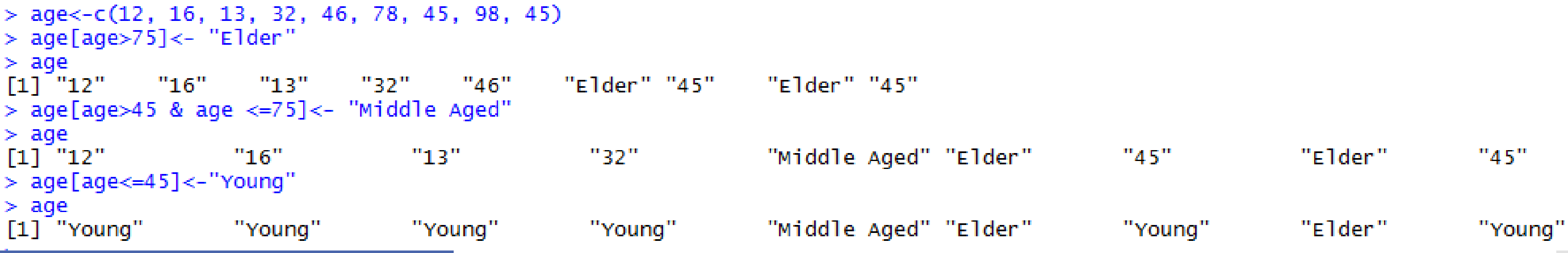
- Each of these activities usually involves the use of R’s built-in functions & operators
- Sometimes you have to write your own function!
- Other times you will need to convert variables or datasets to another type (e.g. numeric to character or matrix to dataframe)

Creating new variables

- Use the assignment operator <- to create new variables

```
mydata$sum <- mydata$x1 + mydata$x2
mydata$mean <- (mydata$x1 + mydata$x2)/2
```

- Sometimes you will need to recode variables (i.e. create another variable from input data)



```
> age<-c(12, 16, 13, 32, 46, 78, 45, 98, 45)
> age[age>75]<- "Elder"
> age
[1] "12"    "16"    "13"    "32"    "46"    "Elder" "45"    "Elder" "45"
> age[age>45 & age <=75]<- "Middle Aged"
> age
[1] "12"          "16"          "13"          "32"          "Middle Aged" "Elder"       "45"          "Elder"       "45"
> age[age<=45]<-"Young"
> age
[1] "Young"       "Young"       "Young"       "Young"       "Middle Aged" "Elder"       "Young"       "Elder"       "Young"
```

Merging

- To merge two dataframes (datasets) horizontally, use the merge function
- In most cases, you join two dataframes by one or more common key variables (i.e., an inner join)

```
# merge two dataframes by ID
total <- merge(dataframeA, dataframeB, by="ID")

# merge two dataframes by ID & Country
total <- merge(dataframeA, dataframeB,
by=c("ID","Country"))
```

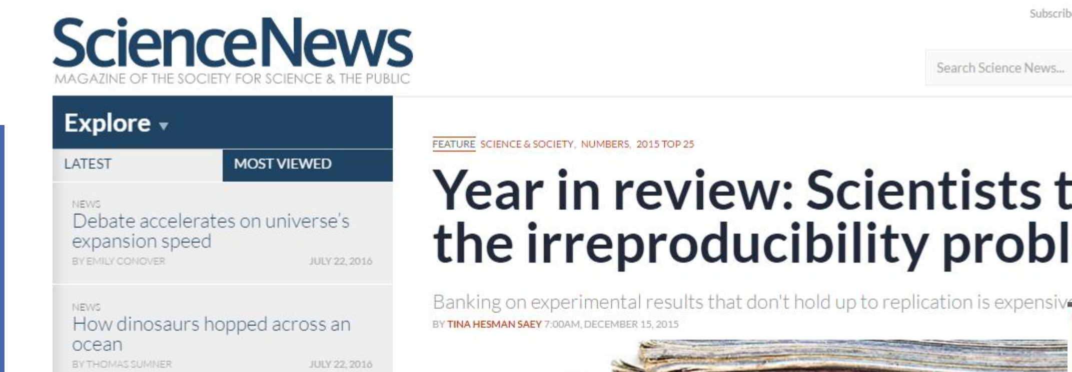
Practicing in R

- In class lab 2 & vector practice

Reproducibility in science

Reproducibility

- Reproducibility is the ability of an entire experiment (or statistical analysis) to be duplicated, either by the same researcher or by someone else working independently
- It is one of the main principles of the scientific method
- It has come to the forefront recently



ScienceNews
MAGAZINE OF THE SOCIETY FOR SCIENCE & THE PUBLIC

Subscribe | Advertise
Archive

Search Science News...

Explore ▾

LATEST | MOST VIEWED

NEWS
Debate accelerates on universe's expansion speed
BY EMILY CONOVER JULY 22, 2016

NEWS
How dinosaurs hopped across an ocean
BY THOMAS SUMNER JULY 22, 2016

SCIENCE STATS
U.S. lags in ro
BY ALEX MADDON

NEWS
Yeasts hide ir partnerships
BY SUSAN MILIUS

SCIENCE TICKER
Getting rid of stopping snai
BY AMY MCDERMOTT

SCIENCE NEWS FOR ST
The shocking

NEWS
Humans, bird collaborate
BY BRUCE BOWER

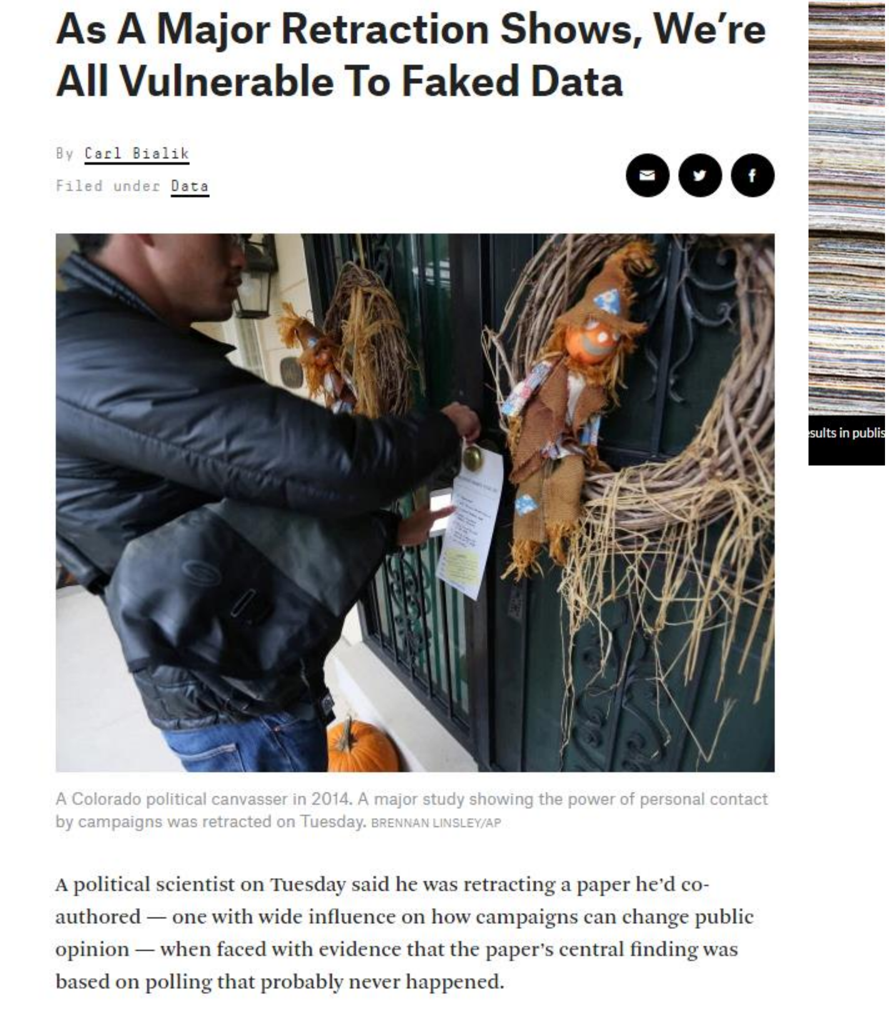
FEATURE SCIENCE & SOCIETY, NUMBERS, 2015 TOP 25

Year in review: Scientists tackle the irreproducibility problem

Banking on experimental results that don't hold up to replication is expensive

BY TINA HESMAN SAEY 7:00AM, DECEMBER 15, 2015

sults in publis

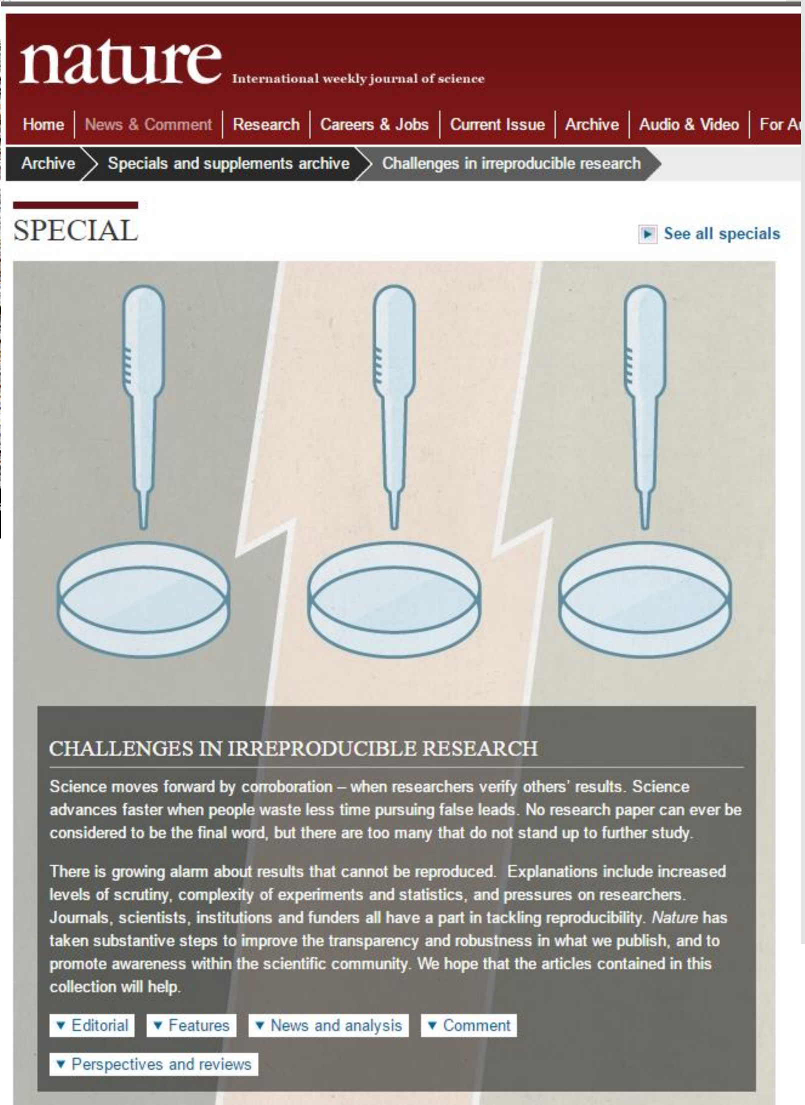


As A Major Retraction Shows, We're All Vulnerable To Faked Data

By Carl Bialik
Filed under Data

A Colorado political canvasser in 2014. A major study showing the power of personal contact by campaigns was retracted on Tuesday. BRENNAN LINSLEY/AP

A political scientist on Tuesday said he was retracting a paper he'd co-authored — one with wide influence on how campaigns can change public opinion — when faced with evidence that the paper's central finding was based on polling that probably never happened.



nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Au

Archive > Specials and supplements archive > Challenges in irreproducible research

SPECIAL

See all specials

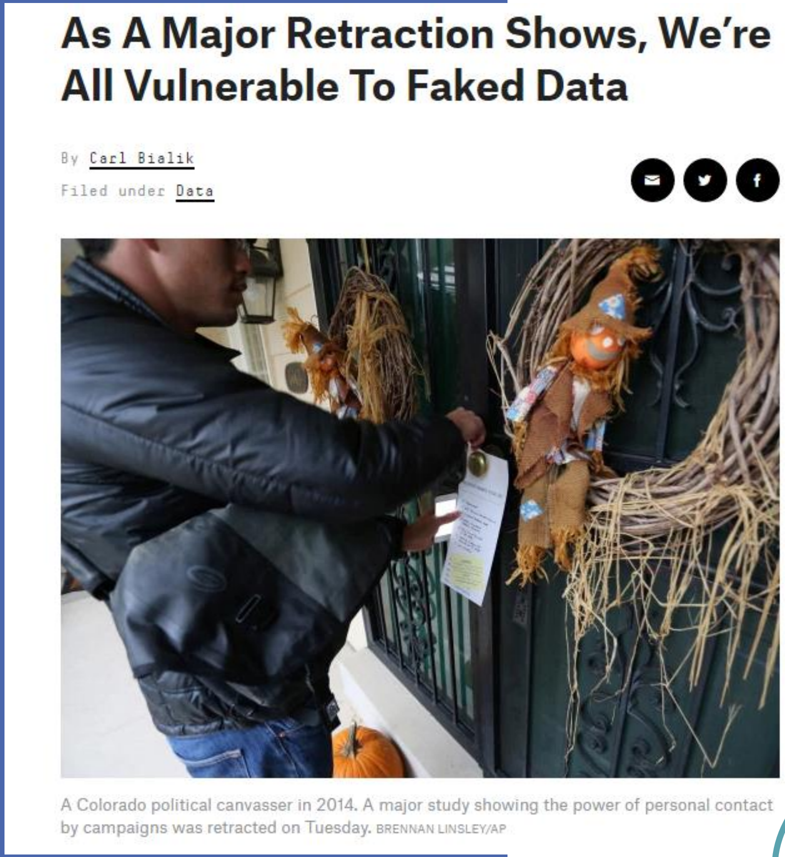
CHALLENGES IN IRREPRODUCIBLE RESEARCH

Science moves forward by corroboration – when researchers verify others' results. Science advances faster when people waste less time pursuing false leads. No research paper can ever be considered to be the final word, but there are too many that do not stand up to further study.

There is growing alarm about results that cannot be reproduced. Explanations include increased levels of scrutiny, complexity of experiments and statistics, and pressures on researchers. Journals, scientists, institutions and funders all have a part in tackling reproducibility. *Nature* has taken substantive steps to improve the transparency and robustness in what we publish, and to promote awareness within the scientific community. We hope that the articles contained in this collection will help.

▾ Editorial ▾ Features ▾ News and analysis ▾ Comment

▾ Perspectives and reviews

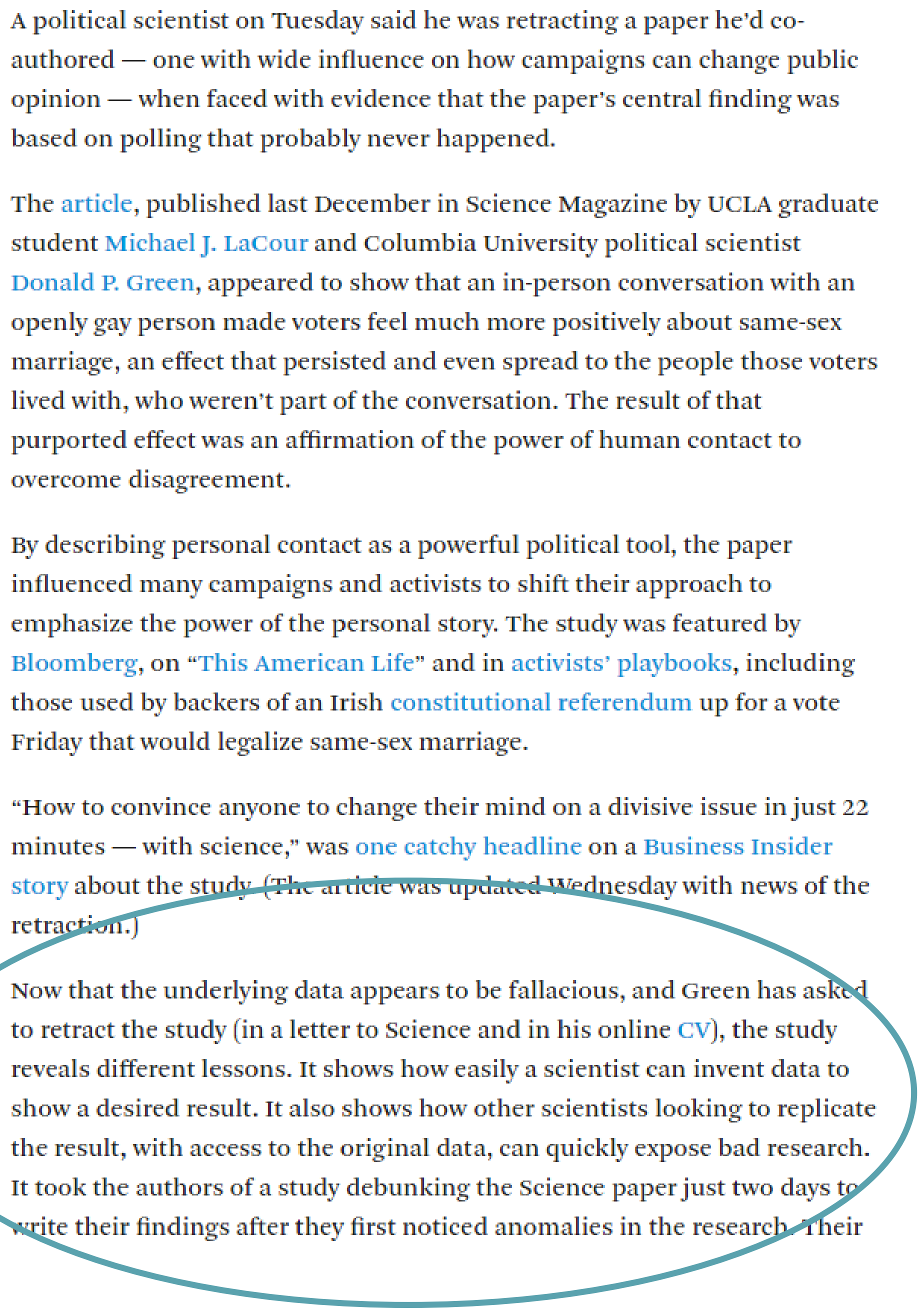


As A Major Retraction Shows, We're All Vulnerable To Faked Data

By Carl Bialik

Filed under Data

A Colorado political canvasser in 2014. A major study showing the power of personal contact by campaigns was retracted on Tuesday. BRENNAN LINSLEY/AP



A political scientist on Tuesday said he was retracting a paper he'd co-authored — one with wide influence on how campaigns can change public opinion — when faced with evidence that the paper's central finding was based on polling that probably never happened.

The article, published last December in Science Magazine by UCLA graduate student Michael J. LaCour and Columbia University political scientist Donald P. Green, appeared to show that an in-person conversation with an openly gay person made voters feel much more positively about same-sex marriage, an effect that persisted and even spread to the people those voters lived with, who weren't part of the conversation. The result of that purported effect was an affirmation of the power of human contact to overcome disagreement.

By describing personal contact as a powerful political tool, the paper influenced many campaigns and activists to shift their approach to emphasize the power of the personal story. The study was featured by Bloomberg, on "This American Life" and in activists' playbooks, including those used by backers of an Irish constitutional referendum up for a vote Friday that would legalize same-sex marriage.

"How to convince anyone to change their mind on a divisive issue in just 22 minutes — with science," was one catchy headline on a Business Insider story about the study. (The article was updated Wednesday with news of the retraction.)

Now that the underlying data appears to be fallacious, and Green has asked to retract the study (in a letter to Science and in his online CV), the study reveals different lessons. It shows how easily a scientist can invent data to show a desired result. It also shows how other scientists looking to replicate the result, with access to the original data, can quickly expose bad research. It took the authors of a study debunking the Science paper just two days to write their findings after they first noticed anomalies in the research. Their

Noteworthy irreproducible results

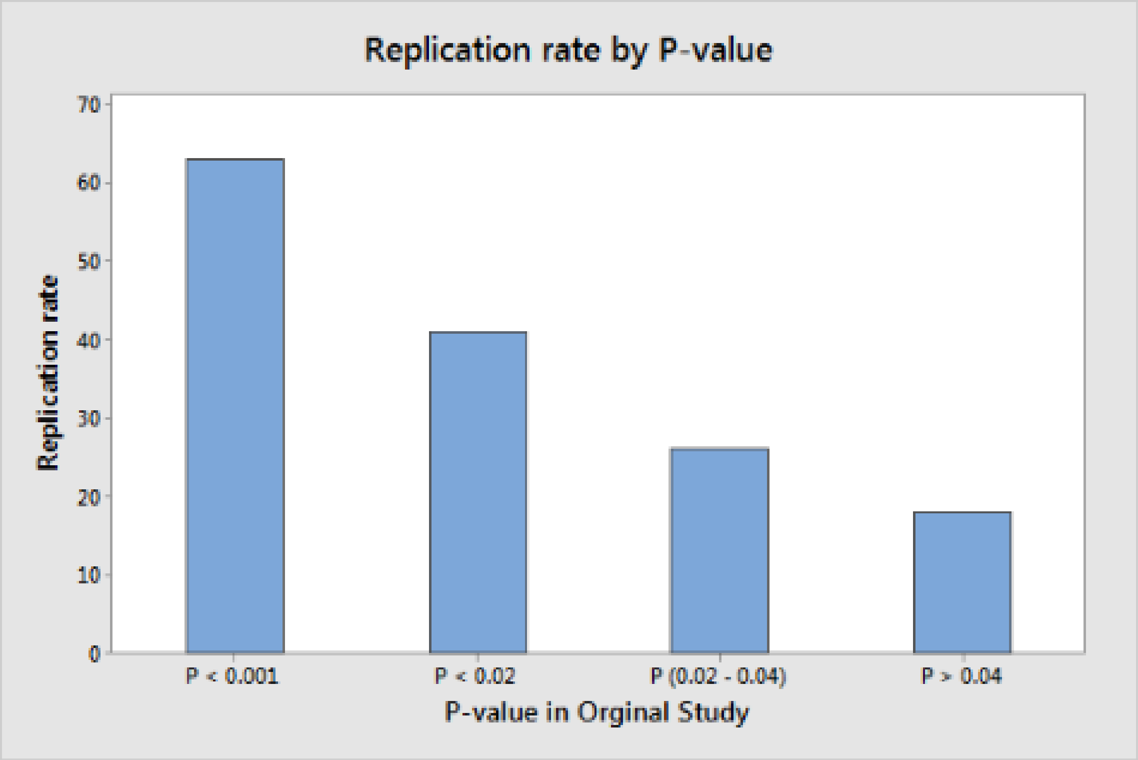
- **Hideyo Noguchi** became famous for correctly identifying the bacterial agent of syphilis, but also claimed he could culture this agent in the lab
  - Nobody has been able to reproduce this result
- **MMR vaccine** controversy
  - A study in the Lancet (peer-reviewed med journal) claiming the MMR vaccine caused autism was revealed to be fraudulent

Going back to p-values

- A 2015 study wanted to assess the rate & predictors of reproducibility in the field of psychology (but this can apply to any field!)
- The authors looked at whether the following factors were predictive of the likelihood that a replication study would be statistically significant (given that the original study already obtained stat. sig. [SS] results):
  - Characteristics of investigator
  - Hypotheses
  - Analytical methods
  - P-value in original study

Going back to p-values

- Most factors did not predict reproducibility



P-values & reproducibility take-home messages

- The exact p-value matters—not just whether a result is significant or not
- A p-value near 0.05 isn't worth much by itself
- Replication is crucial

What does this have to do with us?

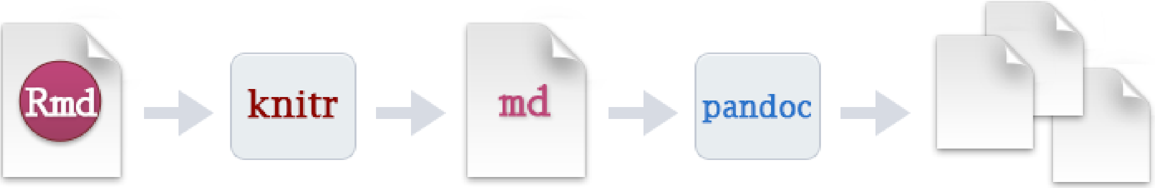
- To maintain the reproducibility of research, certain protocols should be followed like second nature
  - Maintaining well-annotated code
  - Noting final version of codes and models
  - Knowing which code produces the final figures that appear in your papers
  - Saving workspaces so that you can reproduce the same parameter estimates you report

What does this have to do with us?

- One method to achieve these goals is R Markdown
- R Markdown is a way to embed R code chunks in a shareable document, making it easy to create reproducible web-based reports

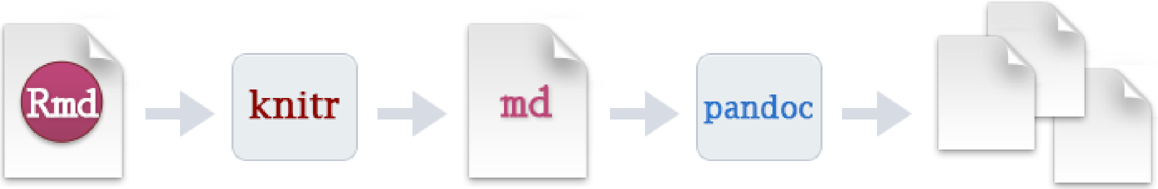
R Markdown Tour

What is R Markdown?



- Creating docs with R Markdown starts with a .Rmd file (not .R)
  - Contains a combination of R code chunks and simple text formatting
- .Rmd file fed to knitr, which executes all R code chunks and creates new .md doc including R code and output

What is R Markdown?



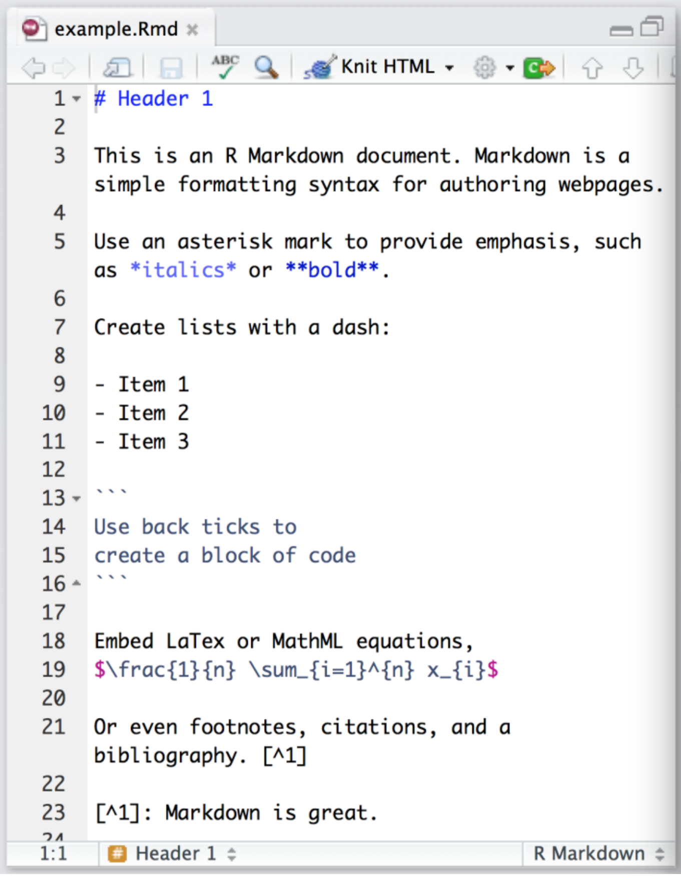
- The .md file is then processed by pandoc which creates a finished webpage, PDF, Word doc, etc.
- Rstudio includes a ‘Knit’ button that enables you to render a .Rmd file and preview it with a single click

R Markdown basics

- Installation
  - install.packages(“rmarkdown”)
- Simply go to File, New File, R Markdown to start a new .Rmd file

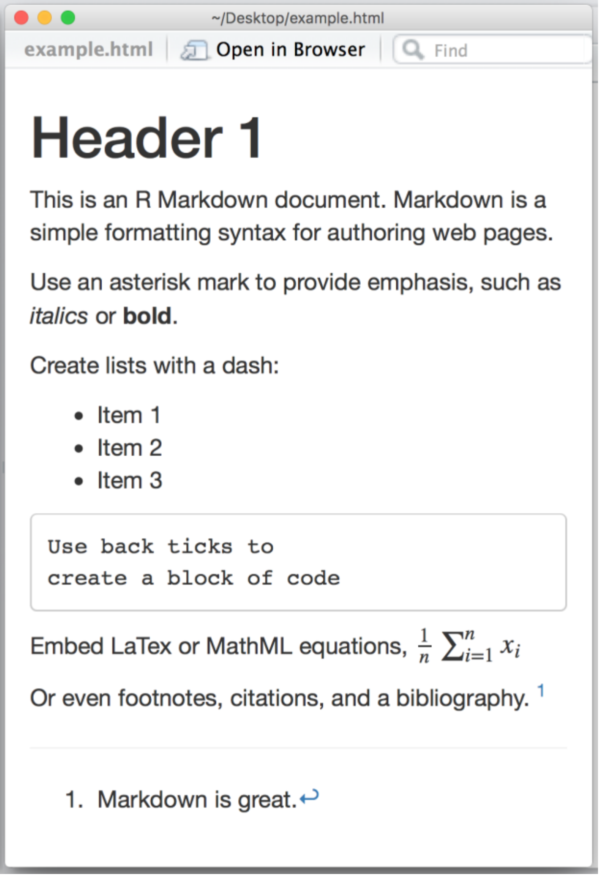
R Markdown basics

Before



```
# Header 1

This is an R Markdown document. Markdown is a simple formatting syntax for authoring webpages.

Use an asterisk mark to provide emphasis, such as *italics* or **bold**.

Create lists with a dash:

- Item 1
- Item 2
- Item 3

```
Use back ticks to
create a block of code
```

Embed LaTex or MathML equations,
$\frac{1}{n} \sum_{i=1}^{n} x_{i}$

Or even footnotes, citations, and a bibliography. [^1]

[^1]: Markdown is great.
```

After



Header 1

This is an R Markdown document. Markdown is a simple formatting syntax for authoring web pages.

Use an asterisk mark to provide emphasis, such as *italics* or **bold**.

Create lists with a dash:

- Item 1
- Item 2
- Item 3

```
Use back ticks to
create a block of code
```

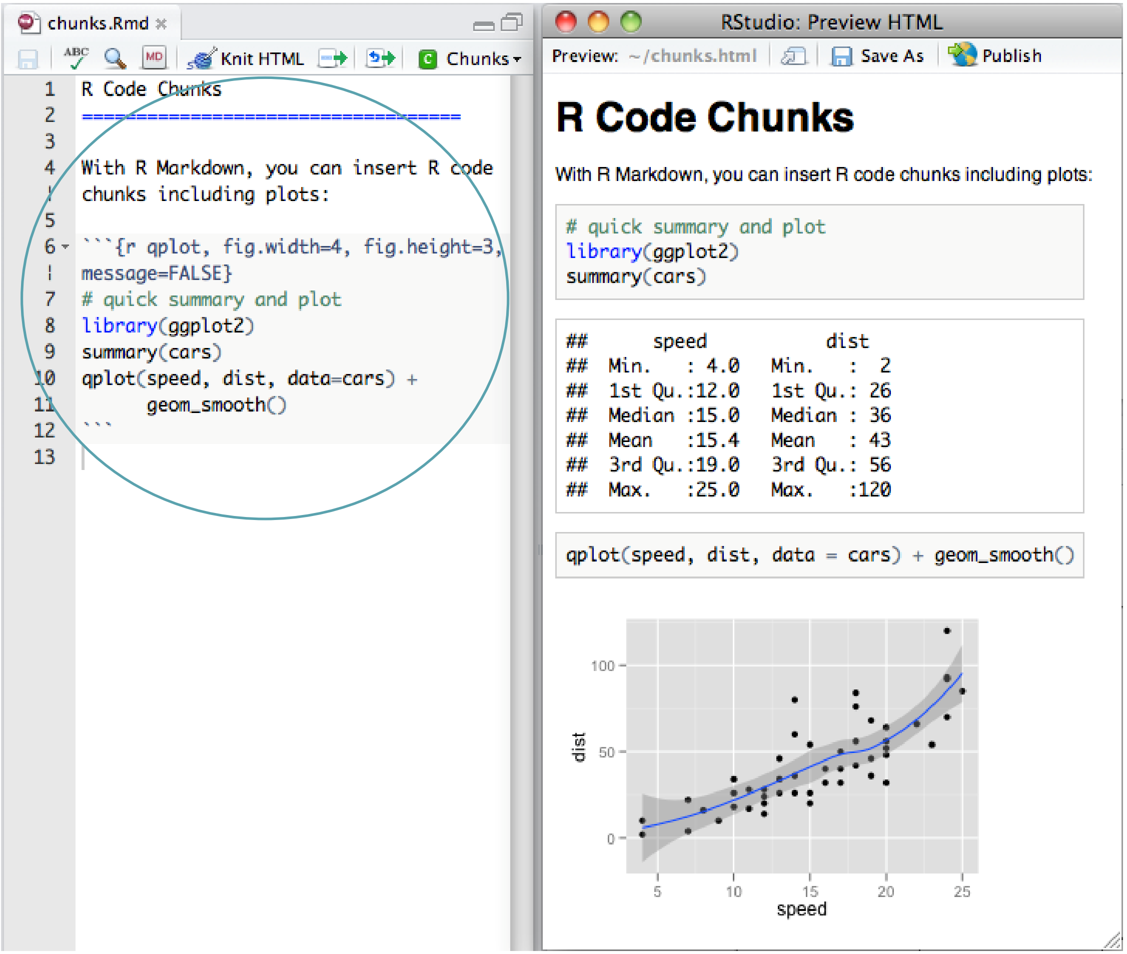
Embed LaTex or MathML equations, $\frac{1}{n} \sum_{i=1}^{n} x_i$

Or even footnotes, citations, and a bibliography. [1]

1. Markdown is great.↩

R Markdown basics

- R code chunks
  - Can be embedded within fenced code regions



~~~
R Code Chunks
=====================================

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

```{r qplot, fig.width=4, fig.height=3, message=FALSE}
# quick summary and plot
library(ggplot2)
summary(cars)
qplot(speed, dist, data=cars) +
    geom_smooth()
```
~~~

R Code Chunks

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

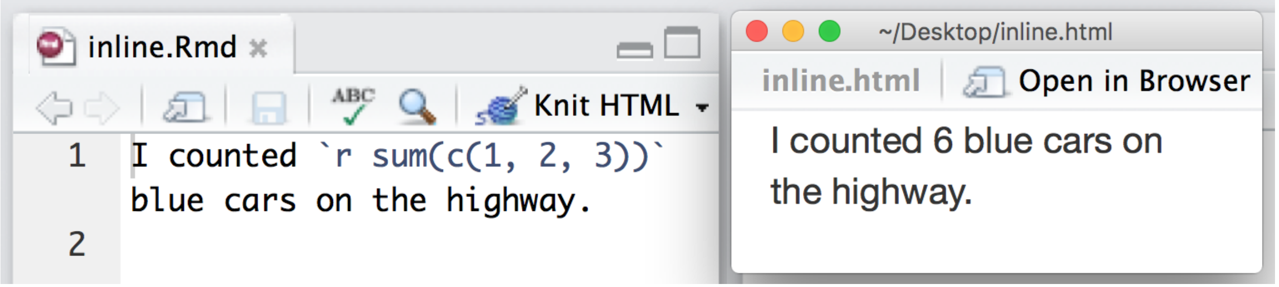
```
# quick summary and plot
library(ggplot2)
summary(cars)
```

```
##      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()
```

R Markdown basics

- Inline R code
  - You can evaluate R expressions inline by enclosing the expression within a single back-tick qualified with ‘r’

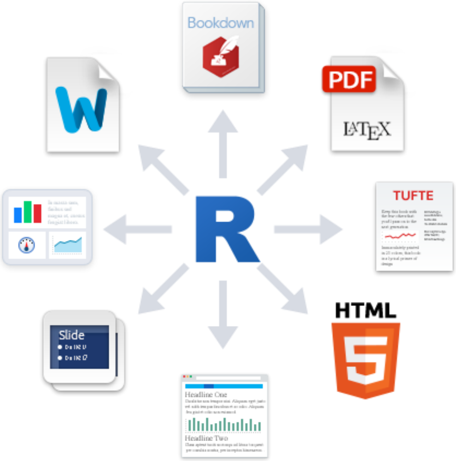


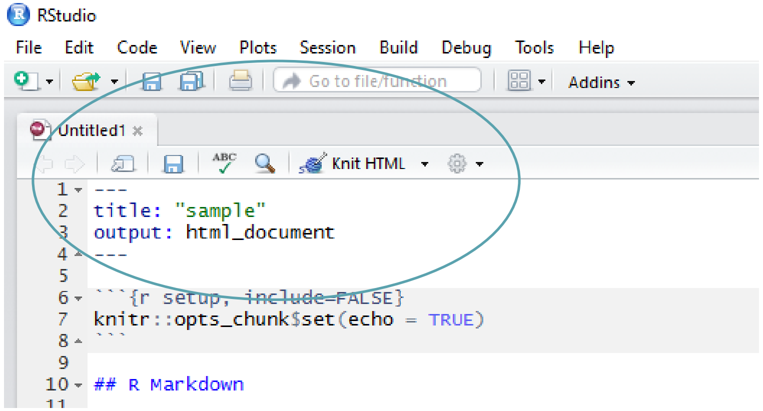
```
I counted `r sum(c(1, 2, 3))`
blue cars on the highway.
```

I counted 6 blue cars on the highway.

R Markdown basics

- Rendering output
  - Use the Knit HTML button in upper left to finalize document
  - Specify what format you prefer when dialog box pops open under File > New
    - You can also specify format under output style at top of doc





~~~
---
title: "sample"
output: html_document
---

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

## R Markdown
~~~

R Markdown basics

- Let’s open RStudio to try a .Rmd tutorial
- Assignment #1 due next Thurs (9/15)—can get started on it today