Advanced R

September 2016

Hadley Wickham
@hadleywickham
Chief Scientist, **RStudio**

HELLO

my name is

Hadley

Sean Hackett

Postdoc,
Storey Lab,
Princeton

Jared Lander

Chief Data Scientist,
Lander Analytics

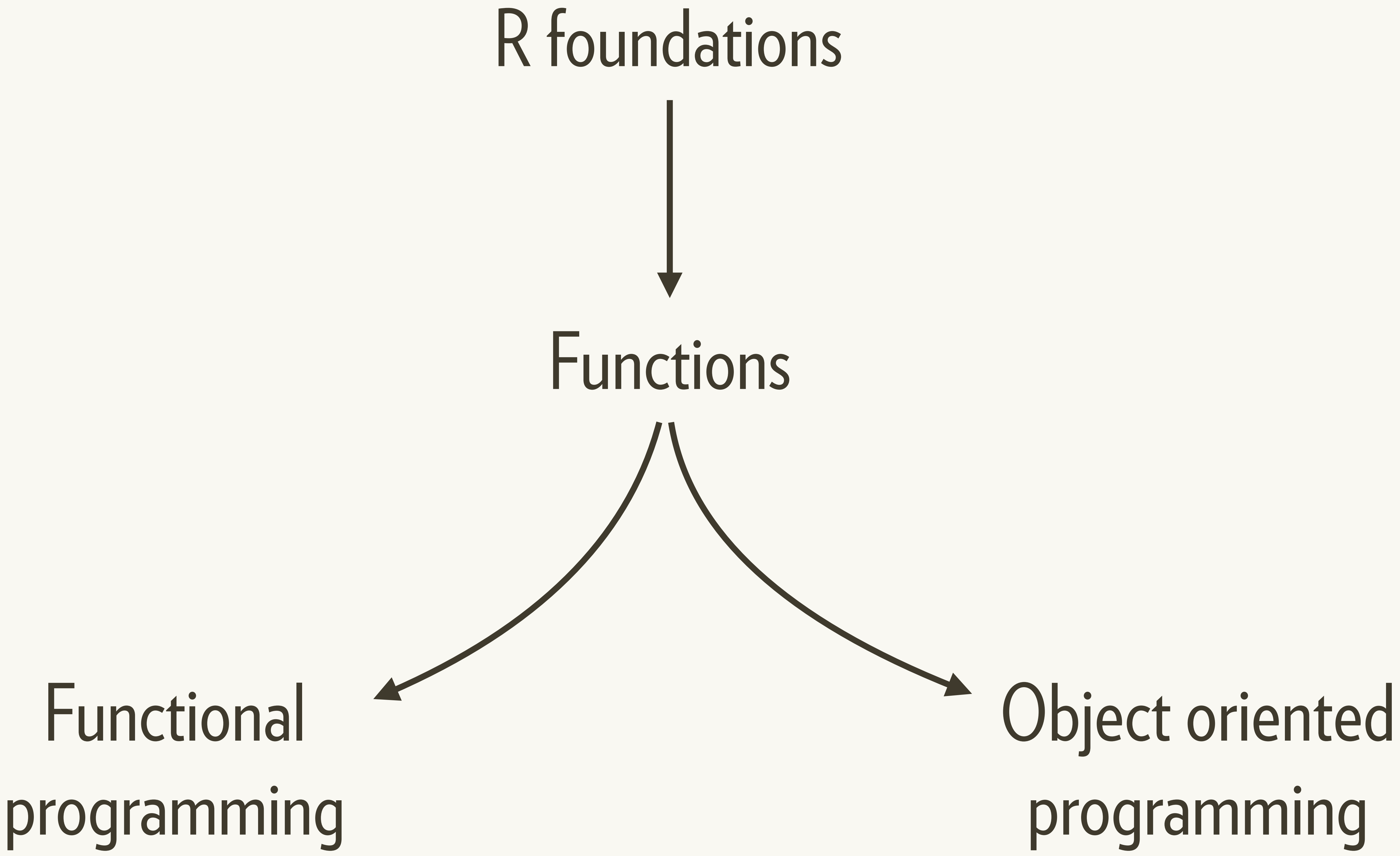
Your turn

This means that you have to work!

This course is very hands on, and while we're here to help you, the best resource is often the person sitting next to you.

Introduce yourself to your neighbours. Who are you and what are you using R for?

Today you’ll learn



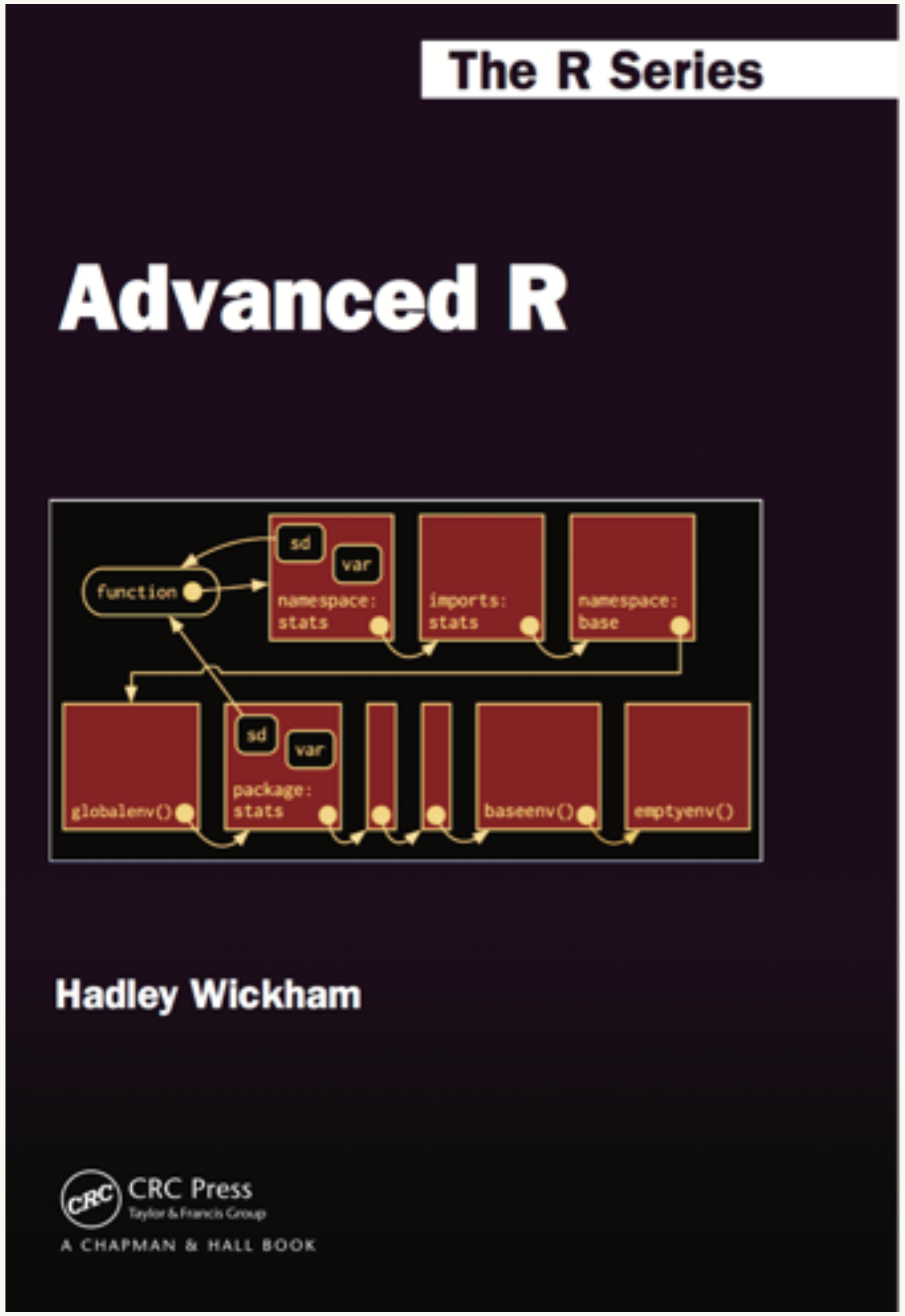
R foundations

Functions

Functional programming

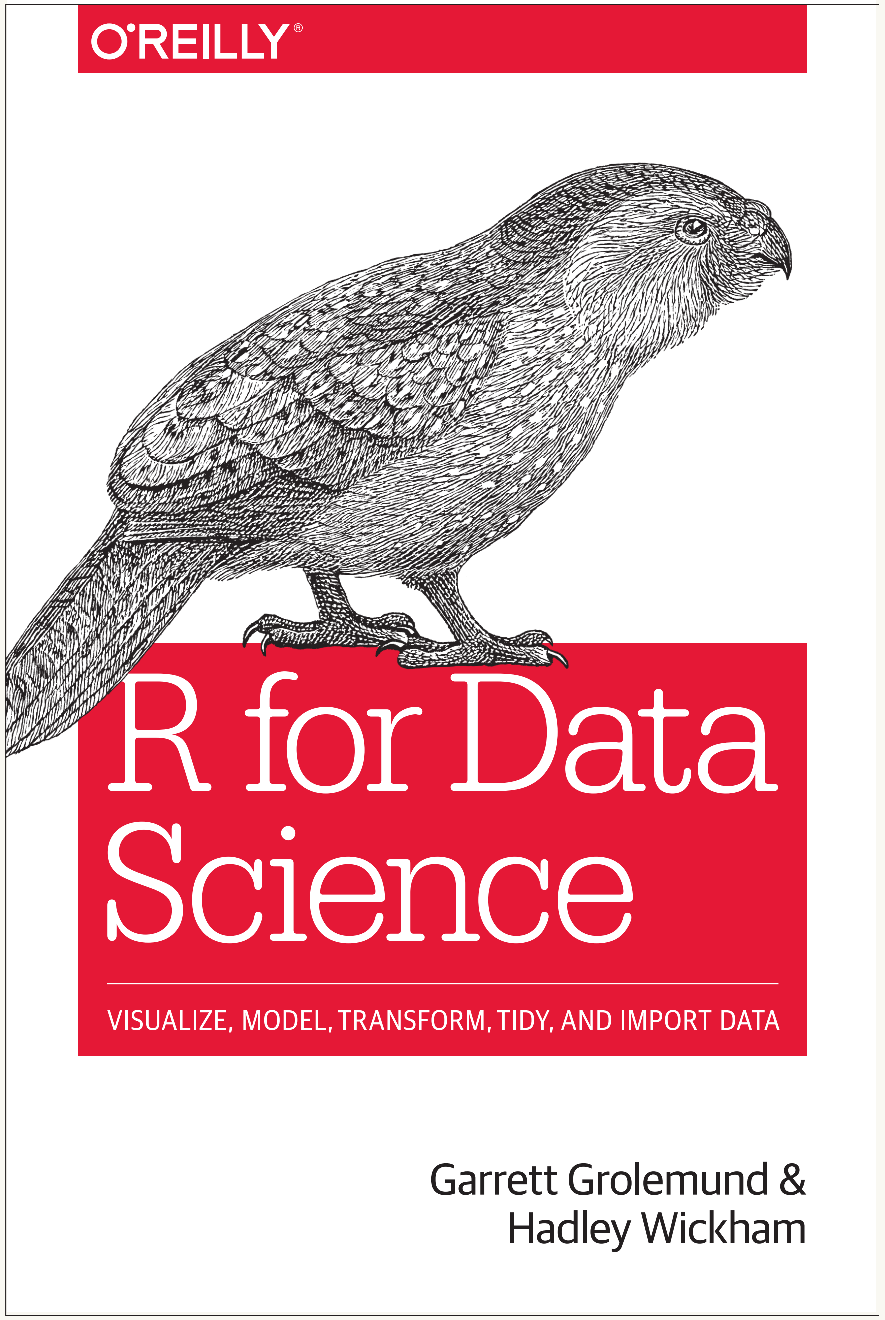
Object oriented programming

The material is mostly drawn from:



http://adv-r.had.co.nz/

http://amzn.com/1466586966



http://r4ds.had.co.nz

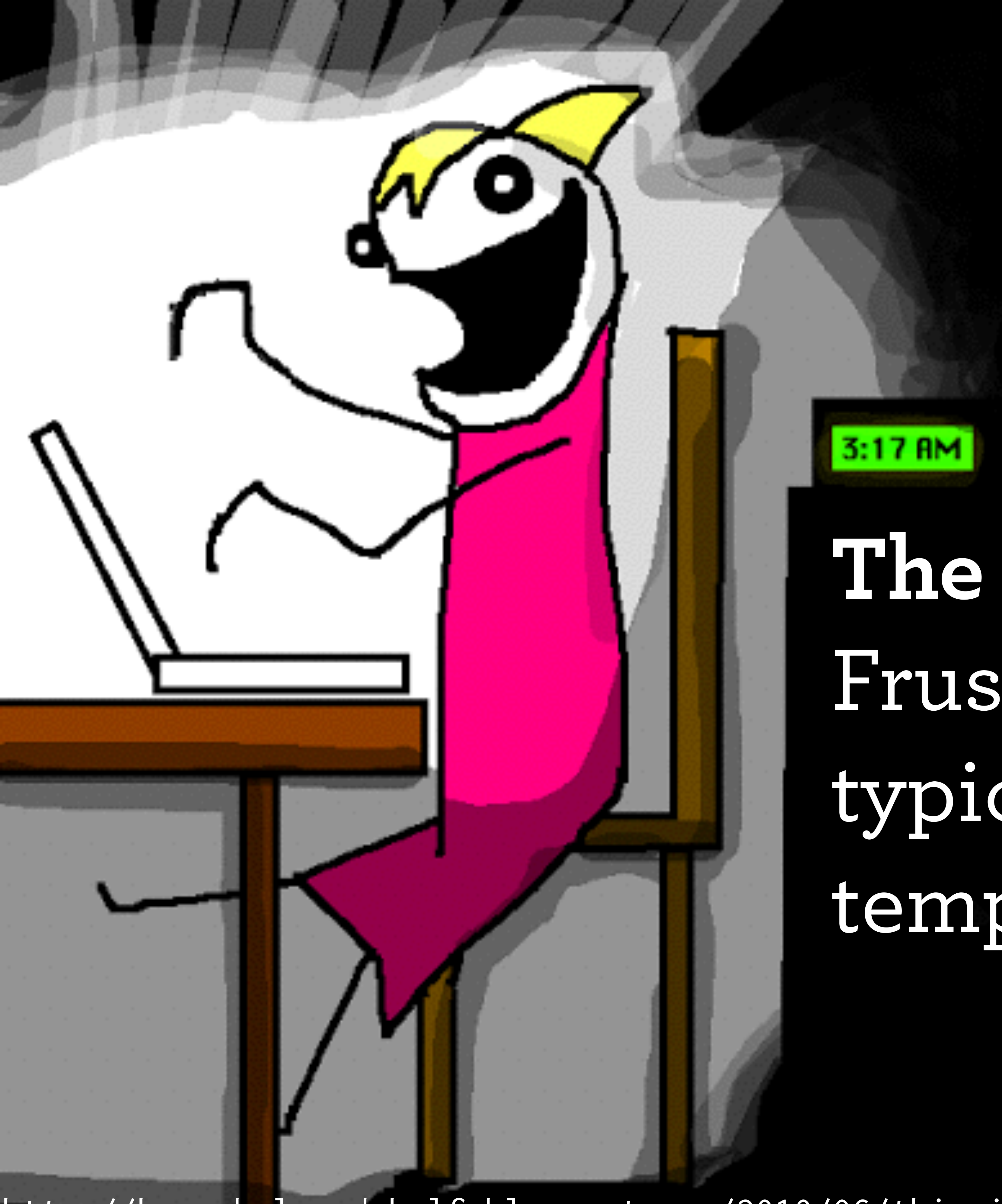
https://amzn.com/1491910399

© Allie Brosh

The bad news:

It's going to be frustrating

http://hyperboleandahalf.blogspot.com/2010/09/four-levels-of-social-entrapment.html



© Allie Brosh

The good news:
Frustration is typical and temporary

http://hyperboleandahalf.blogspot.com/2010/06/this-is-why-ill-never-be-adult.html

Warmups

Don’t expect to know all the answers!

Your turn

What are the four common types of atomic vectors?
(Bonus points for the two uncommon types)

What are the three properties of a vector?

```
typeof(TRUE)
typeof(1L)
typeof(1.5)
typeof(”a”)

# We’ll talk about this (S3) later:
typeof(factor(1:10))
typeof(Sys.Date())
```

Every vector has three properties:

```
x <- 1:5

# 1. Type:
typeof(x)

# 2. Length
length(x)

# 3. Attributes
attributes(x)
# (we’ll come back to those later)
```

~~`mode()`~~

~~`storage.mode()`~~

Designed for S compatibility!

Beware built-in is.*() functions

```
# What does is.numeric() do?
x <- 1
y <- 1L
is.integer(x)
is.integer(y)
is.numeric(x)
is.numeric(y)

# What does is.vector() do?
is.vector(1:3)
is.vector(factor(1:3))

# Assume a is.*() does something surprising
```

Purrr provides unsurprising predicates

```
library(purrr)

is_atomic(NULL)
is_vector(factor(1:3))
is_numeric(Sys.Date())

# All purrr predicates are of the form
# function(x) typeof(x) %in% y
```

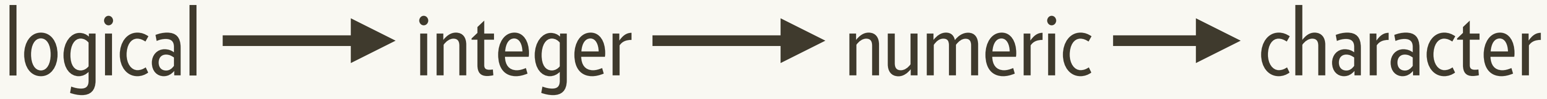
Coercions

What happens when you try to combine multiple atomic vectors with `c()`? Rank the four types according to which “wins”.

What happens when use a logical vector in a numeric context?

Most useful coercion? logical → numeric

```
x <- c(FALSE, TRUE, FALSE, TRUE, TRUE)
sum(x)
mean(x)

x <- sample(1e4, 1e3)
sum(x > 500)
mean(x > 500)
```



Missing values

What does `NA == NA` return? Why? What should you use instead?

What is `NA_real_`? `NA_integer_`? `NA_character_`?

There isn’t a single unknown value

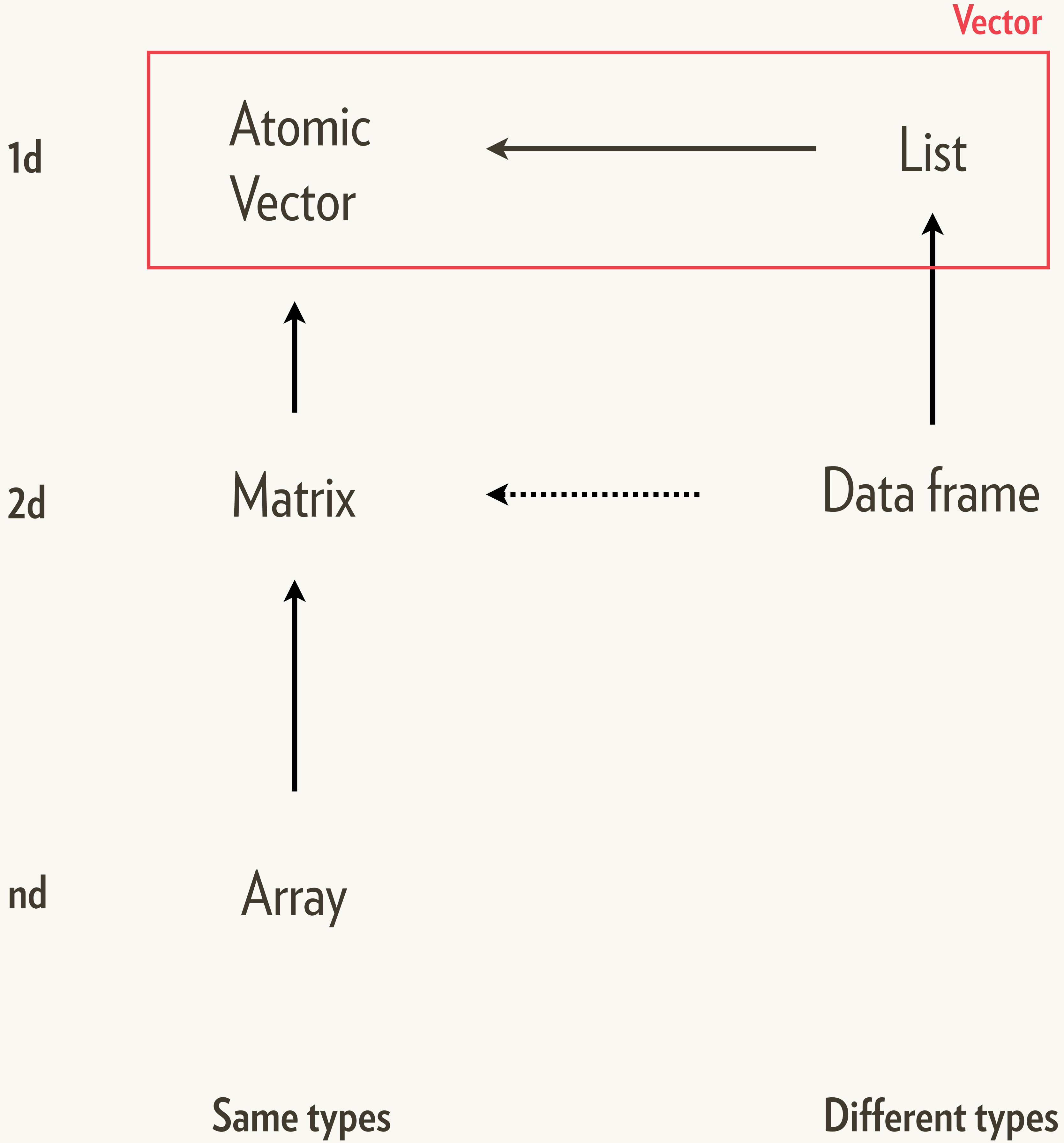
```
age_john <- NA
age_mary <- NA
age_john == age_mary

is.na(x)
!is.na(x) & !is.na(y)

# One NA for each basic atomic vector
typeof(NA)
typeof(NA_real_)
typeof(NA_integer_)
typeof(NA_character_)
```

Your turn

How is a list different from an atomic vector?

How is a data frame different from a matrix?

How do you examine the structure of an object?



Vector

1d

Atomic Vector

List

2d

Matrix

Data frame

nd

Array

Same types

Different types

http://adv-r.had.co.nz/Data-structures.html str()

Your turn

What are the six types of thing that you can put inside []?

What does drop = FALSE do when subsetting a data frame or matrix?

blank	include all
integer	**+ve**: include **0**: drop all **-ve**: exclude
logical	keep TRUEs
character	lookup by name

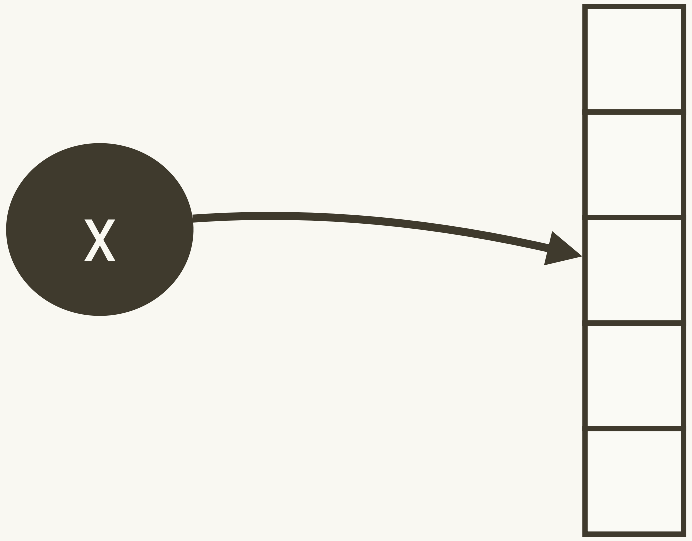
	Simplifying	Preserving
Vectors	`x[[1]]`	`x[1:4]`
Matrices/ Data frame	`x[, 1]`	`x[, 1, drop = F]`

Lists	`x[[1]]`
	`x$name`
		`x[1]`

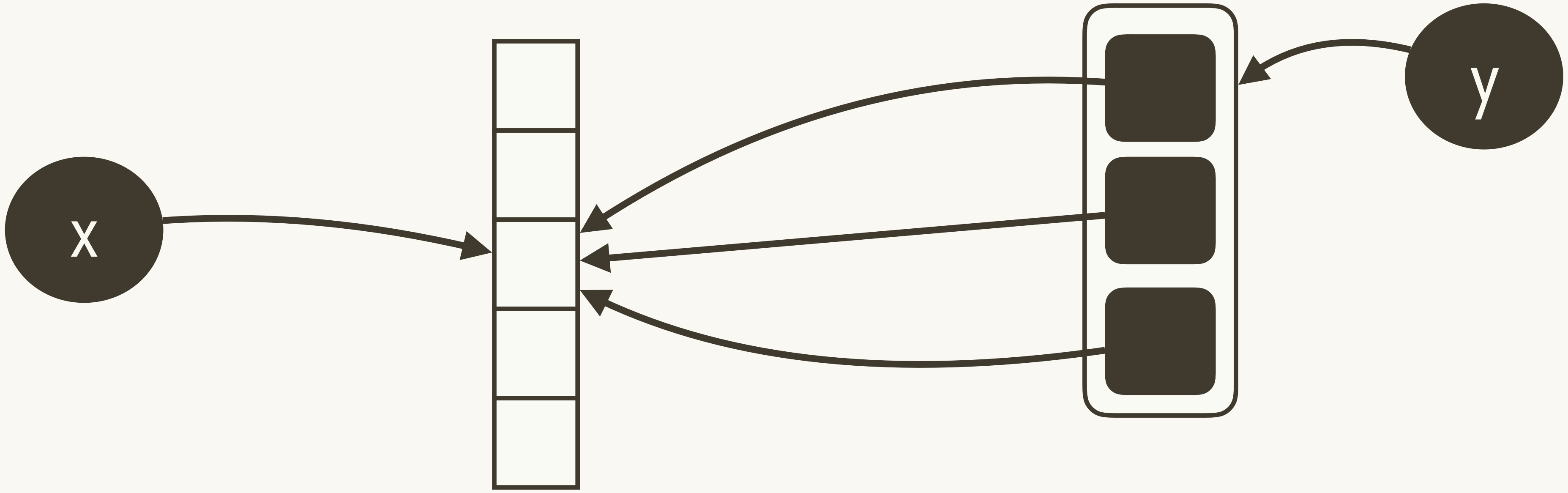
Your turn

```
x <- rnorm(1e6)
pryr::object_size(x)
#> 8 MB

# How big will y be?
y <- list(x, x, x)
pryr::object_size(y)

# What if we change a value?
y[[1]][[1]] <- NA
pryr::object_size(y)
```

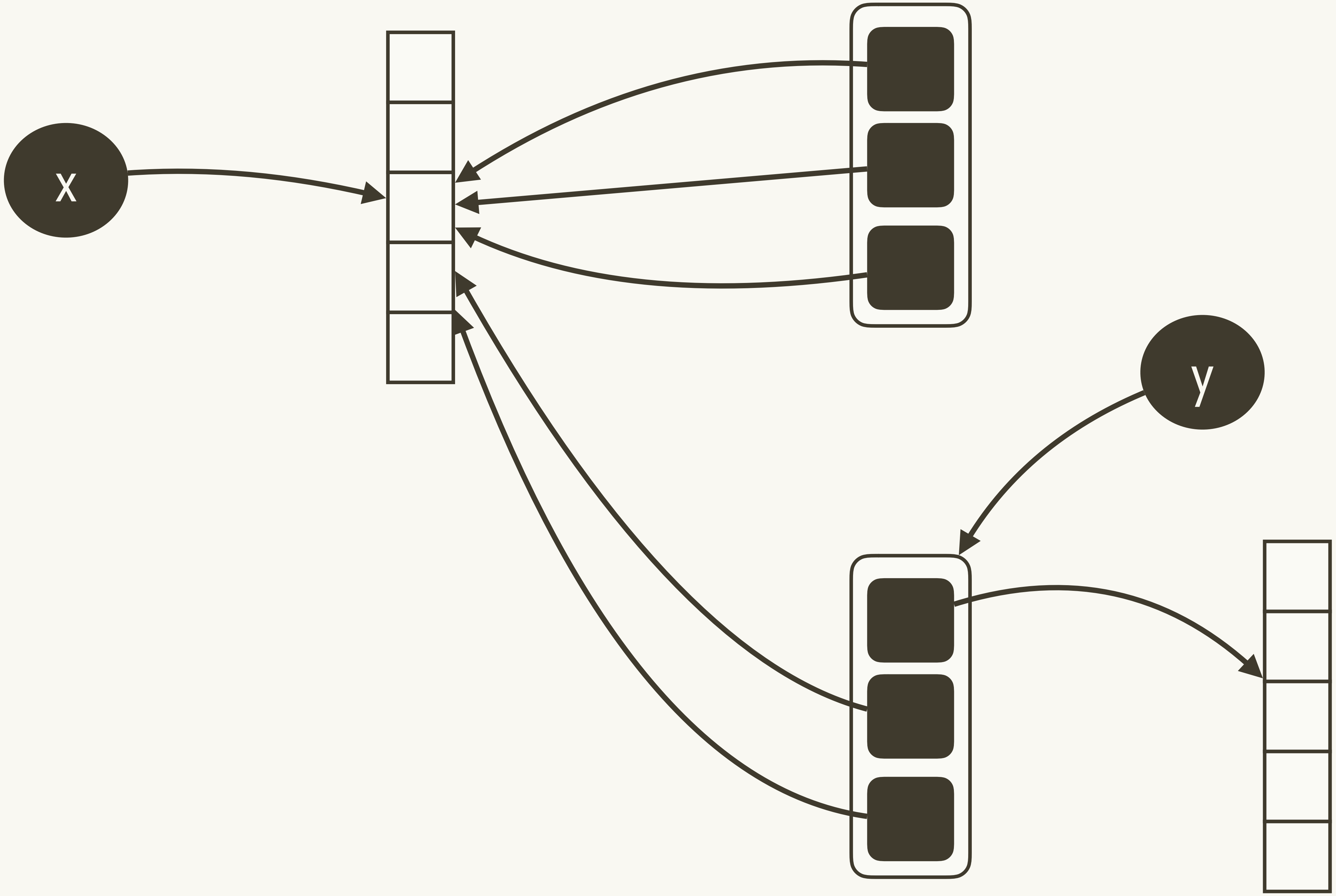
A name points to an object



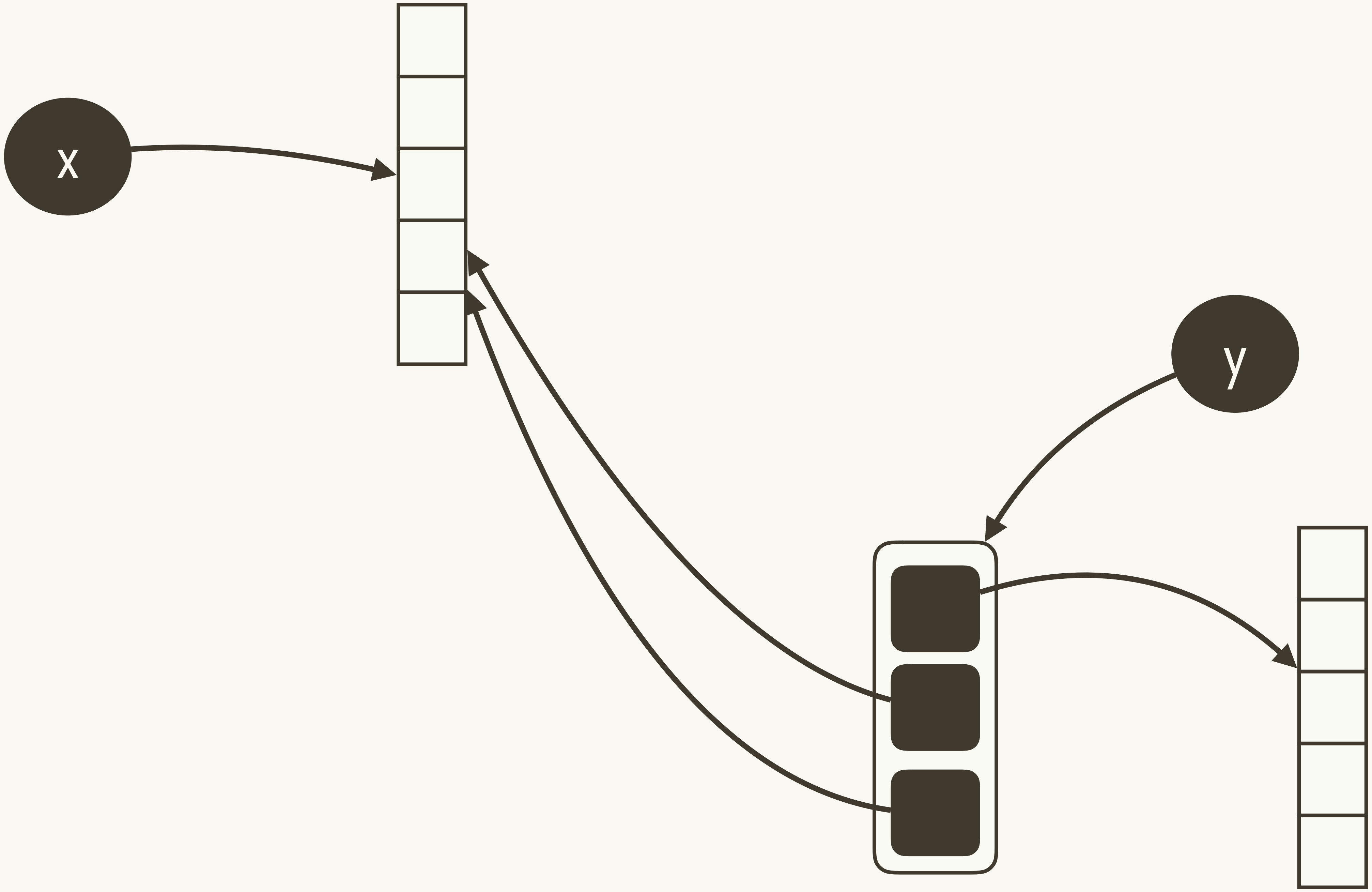
An object can be referenced in multiple places



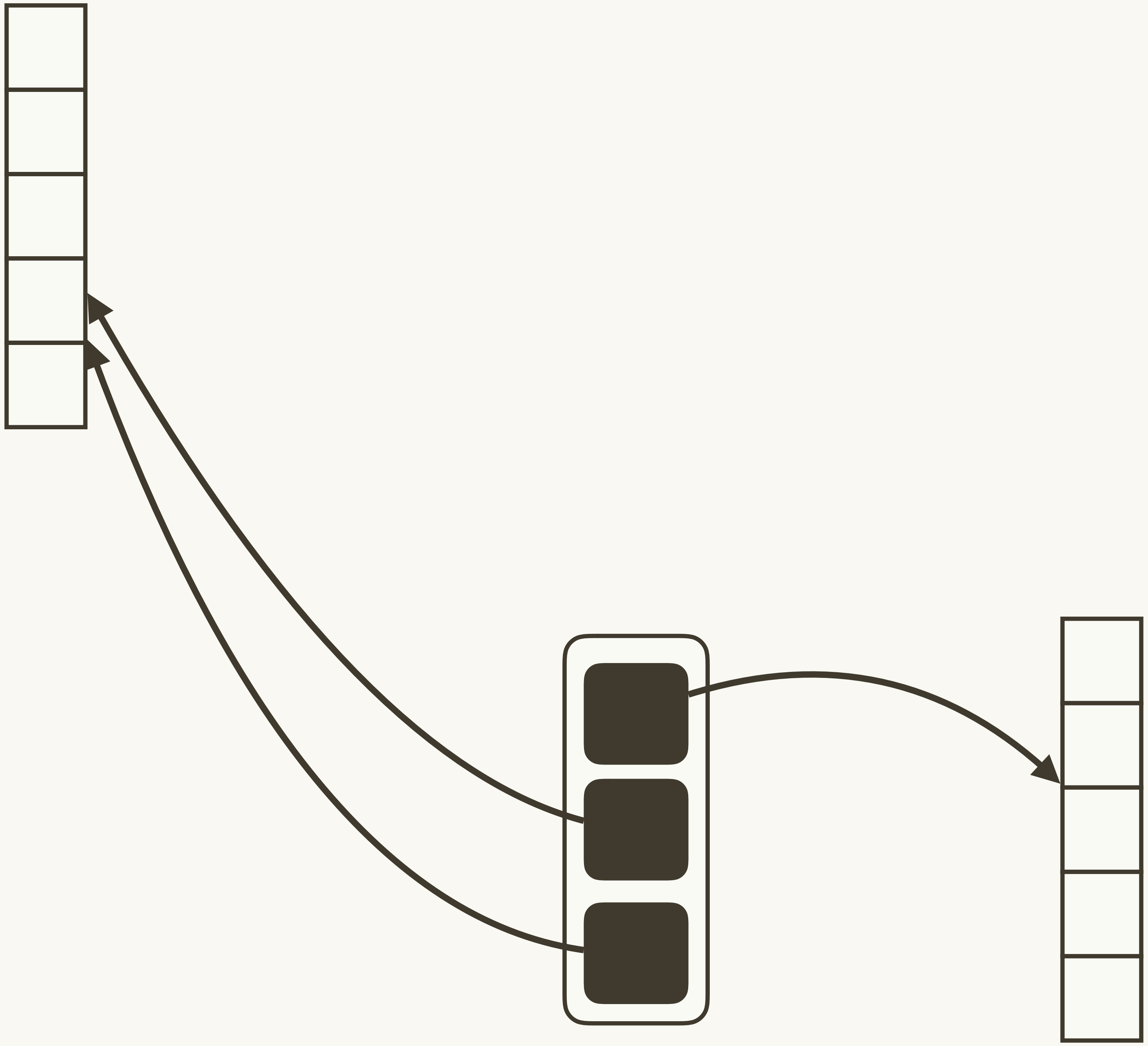
Modifying an object creates a copy



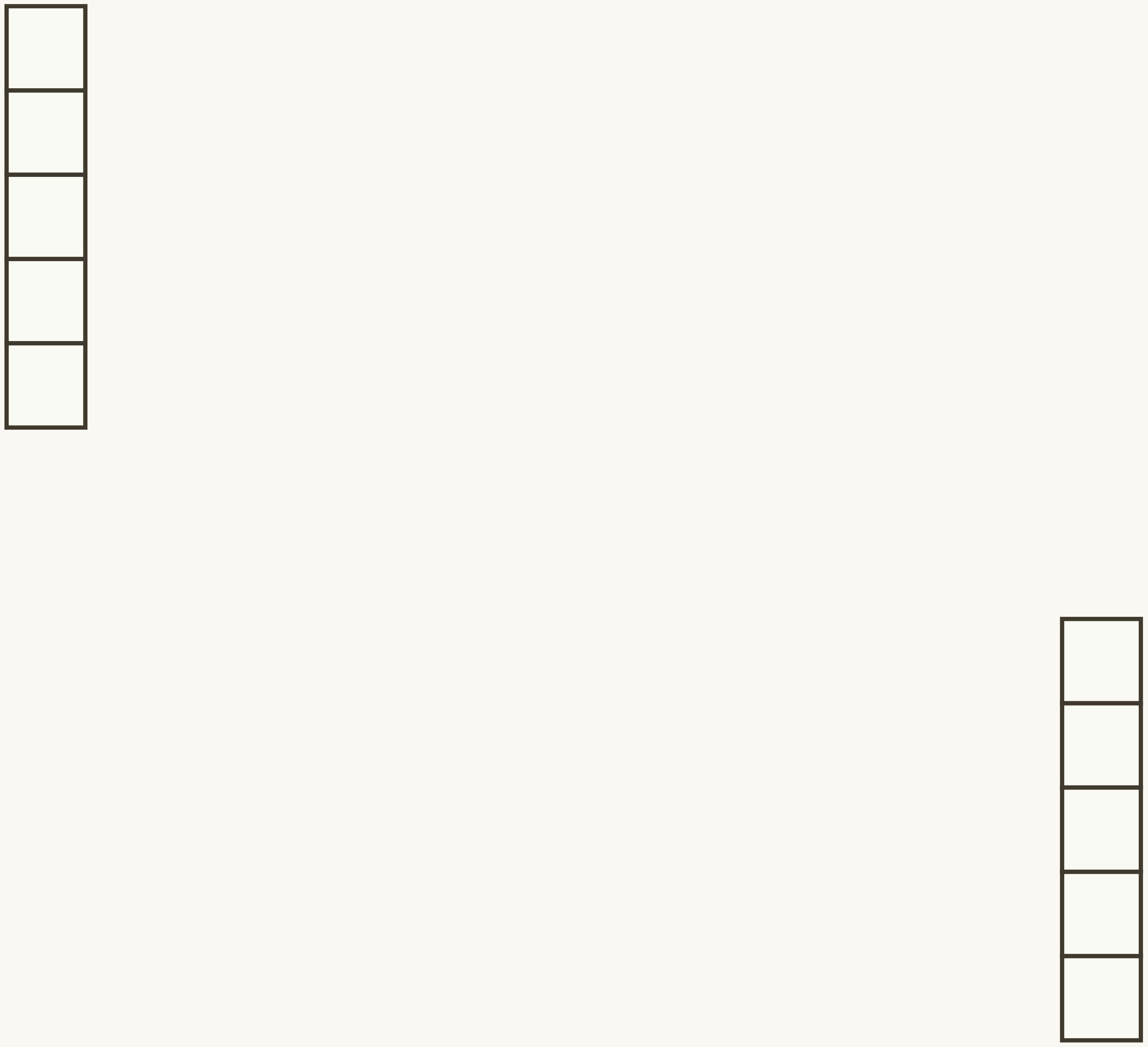
What happens after rm(x) and rm(y)?



First x and y are removed

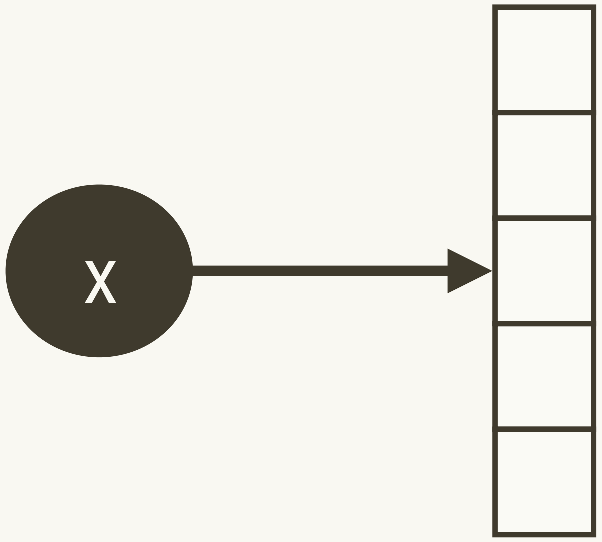


Then the objects they point to

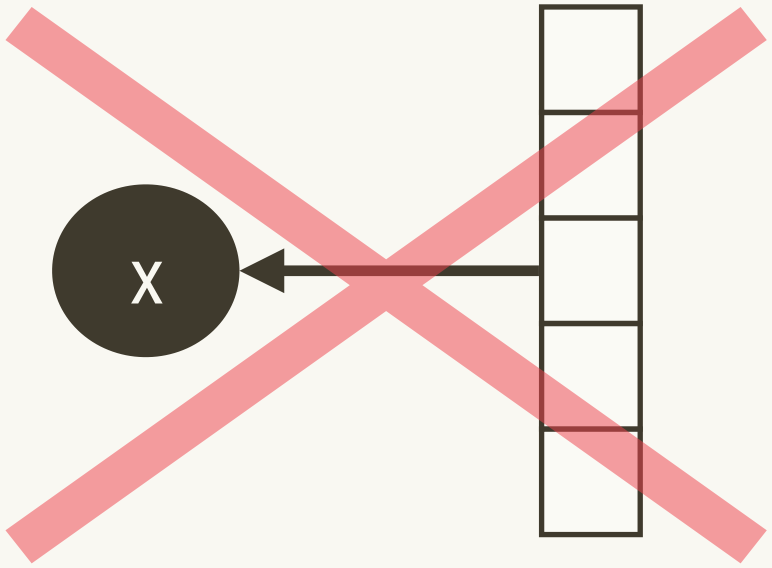


Until nothing is left

The direction of the arrow is important

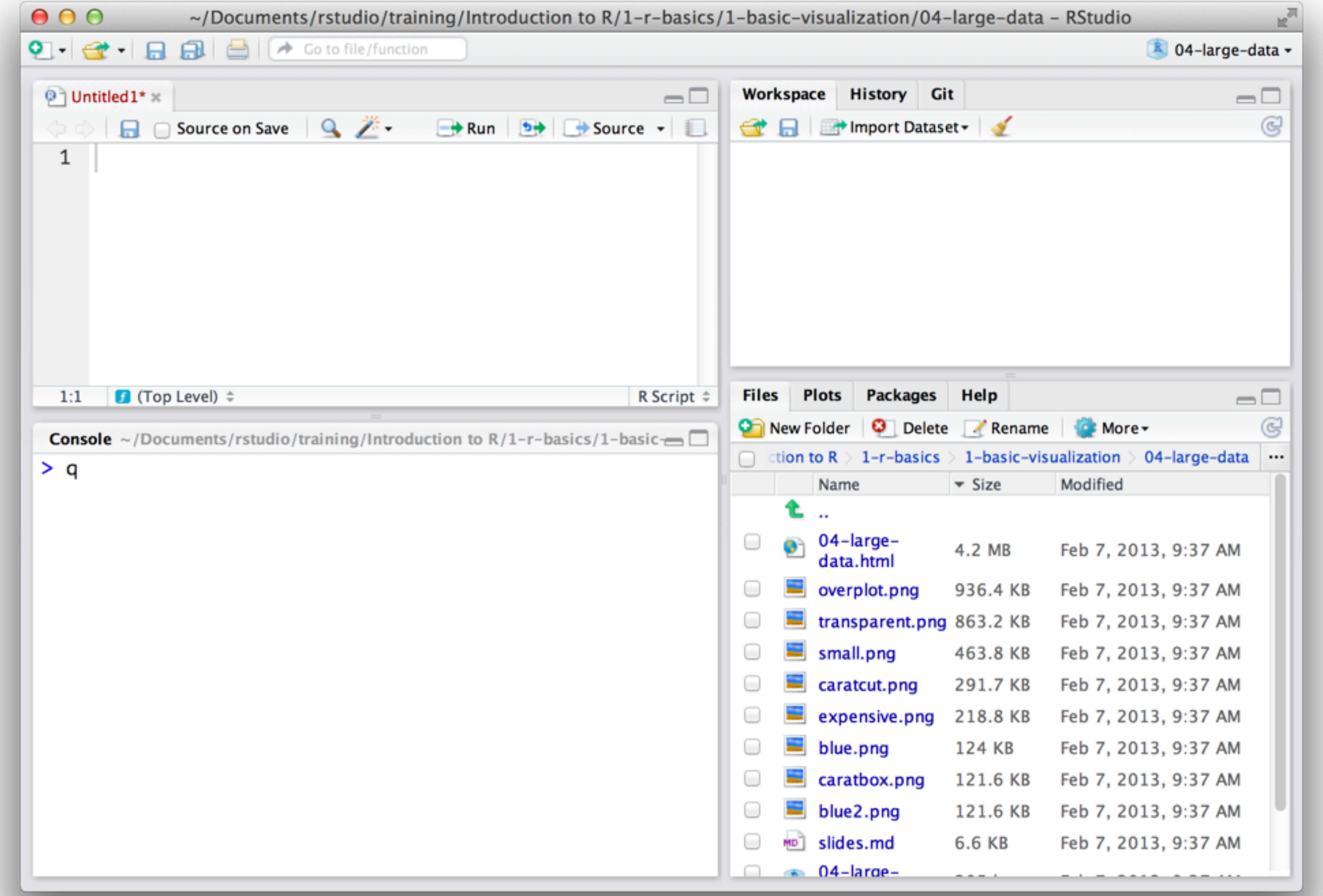


A name “has” an object



An object doesn’t have a name

RStudio

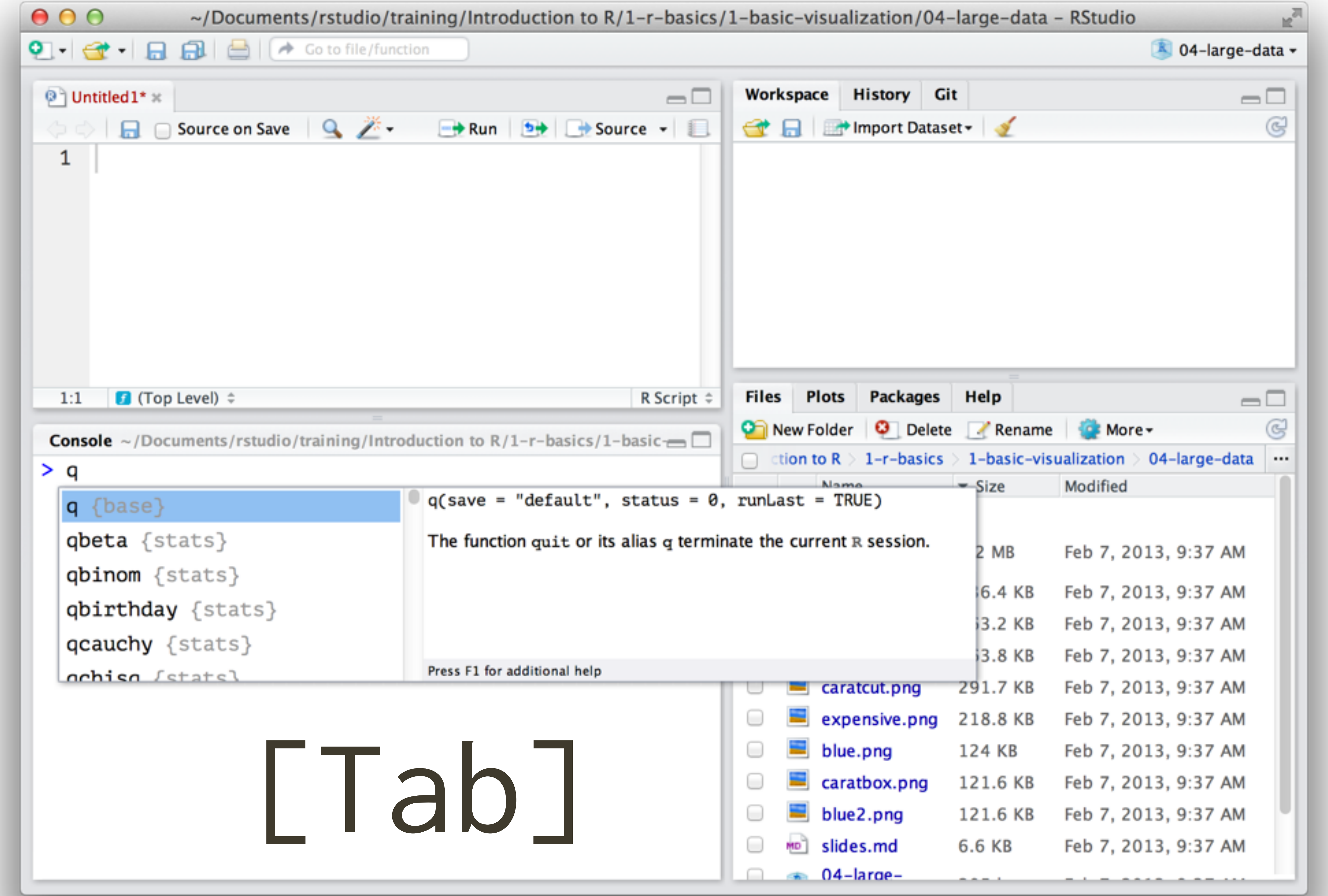


~/Documents/rstudio/training/Introduction to R/1-r-basics/1-basic-visualization/04-large-data – RStudio

Console ~/Documents/rstudio/training/Introduction to R/1-r-basics/1-basic-

```
> q
```

	Name	Size	Modified
	..		
□	04-large-data.html	4.2 MB	Feb 7, 2013, 9:37 AM
□	overplot.png	936.4 KB	Feb 7, 2013, 9:37 AM
□	transparent.png	863.2 KB	Feb 7, 2013, 9:37 AM
□	small.png	463.8 KB	Feb 7, 2013, 9:37 AM
□	caratcut.png	291.7 KB	Feb 7, 2013, 9:37 AM
□	expensive.png	218.8 KB	Feb 7, 2013, 9:37 AM
□	blue.png	124 KB	Feb 7, 2013, 9:37 AM
□	caratbox.png	121.6 KB	Feb 7, 2013, 9:37 AM
□	blue2.png	121.6 KB	Feb 7, 2013, 9:37 AM
□	slides.md	6.6 KB	Feb 7, 2013, 9:37 AM
□	04-large-		



```
> q
```

```
q {base}
qbeta {stats}
qbinom {stats}
qbirthday {stats}
qcauchy {stats}
qchisq {stats}
```

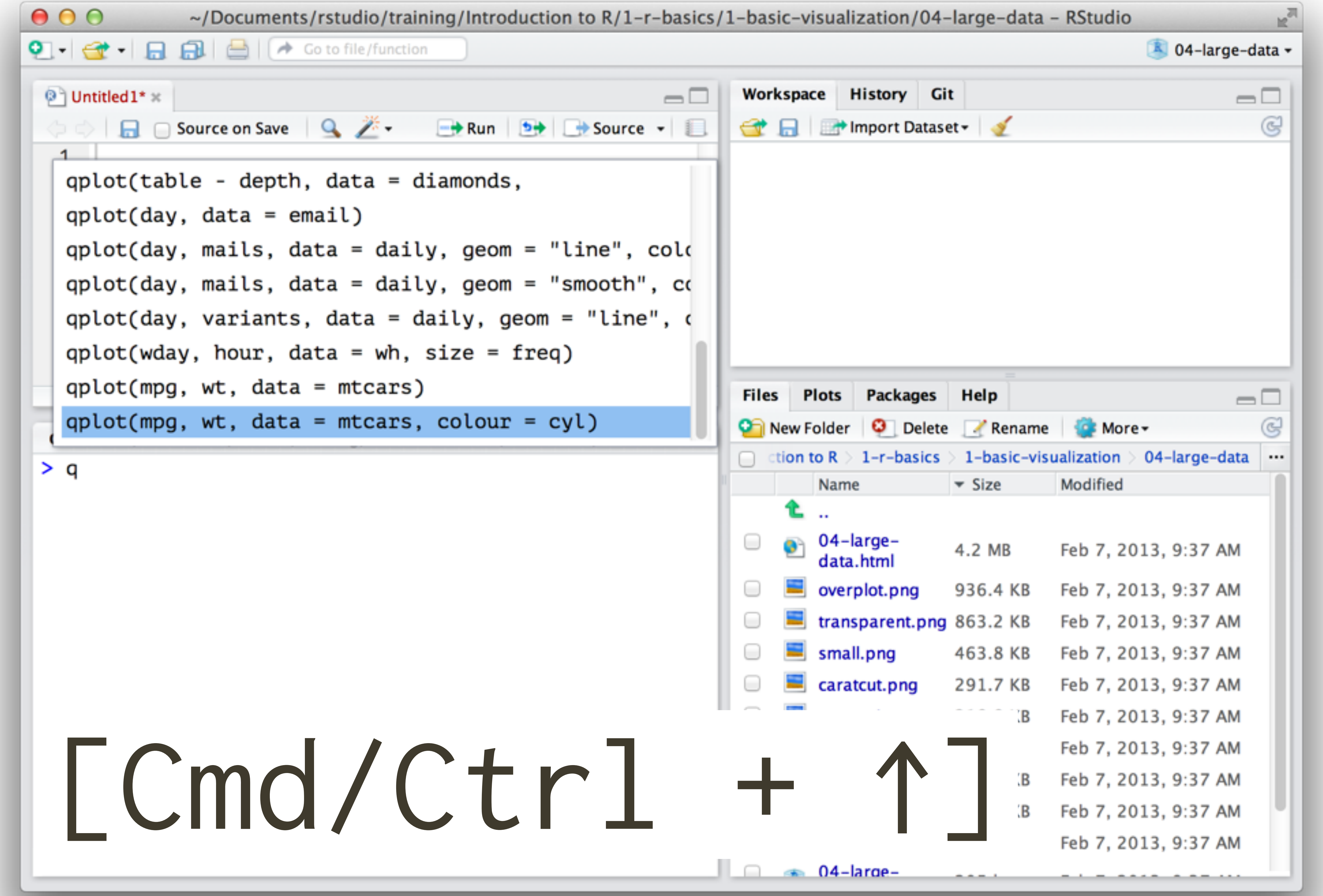
```
q(save = "default", status = 0, runLast = TRUE)
```

The function quit or its alias q terminate the current R session.

Press F1 for additional help

Name	Size	Modified
	2 MB	Feb 7, 2013, 9:37 AM
	6.4 KB	Feb 7, 2013, 9:37 AM
	3.2 KB	Feb 7, 2013, 9:37 AM
	3.8 KB	Feb 7, 2013, 9:37 AM
caratcut.png	291.7 KB	Feb 7, 2013, 9:37 AM
expensive.png	218.8 KB	Feb 7, 2013, 9:37 AM
blue.png	124 KB	Feb 7, 2013, 9:37 AM
caratbox.png	121.6 KB	Feb 7, 2013, 9:37 AM
blue2.png	121.6 KB	Feb 7, 2013, 9:37 AM
slides.md	6.6 KB	Feb 7, 2013, 9:37 AM
04-large-		

[Tab]

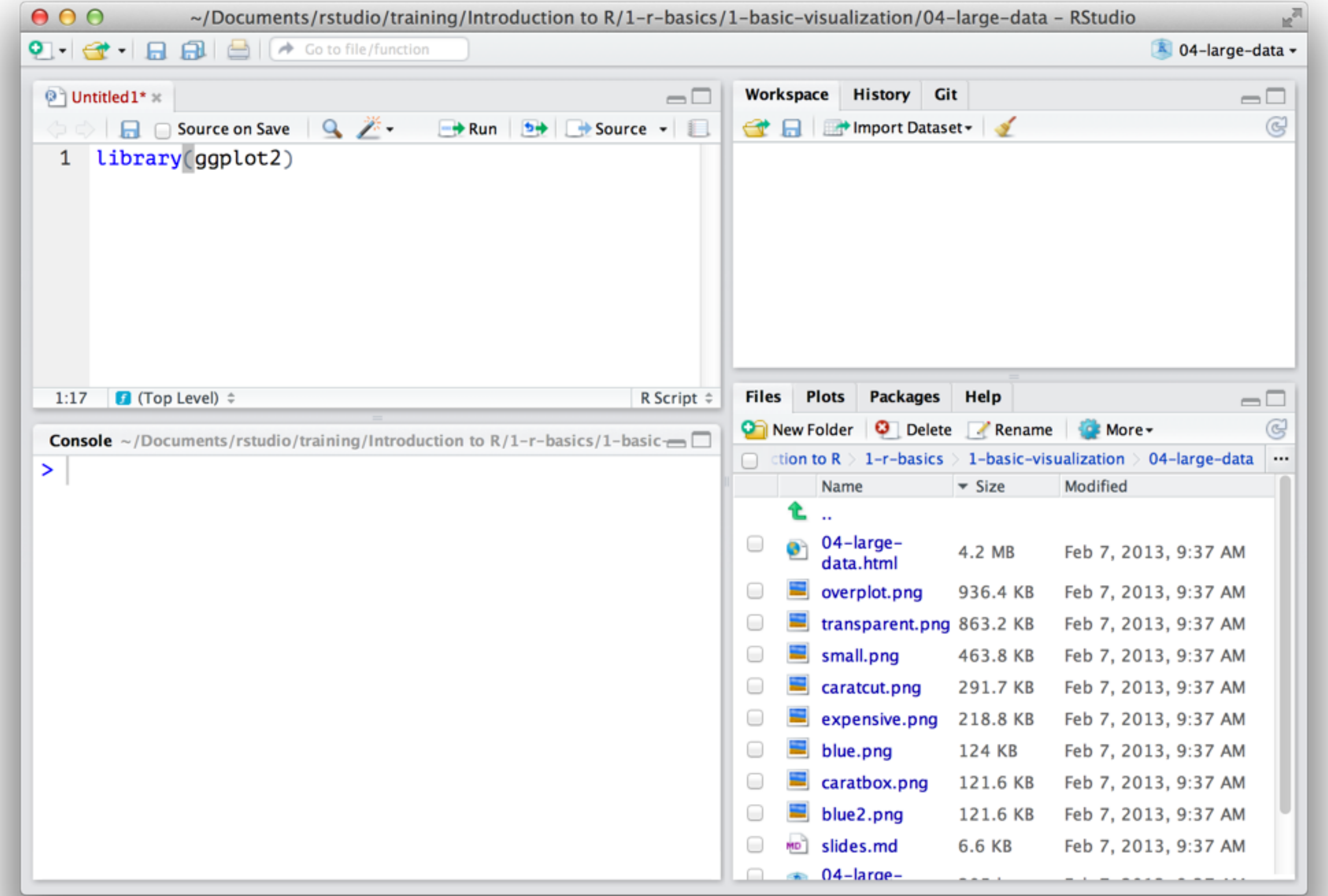


```
qplot(table - depth, data = diamonds,
qplot(day, data = email)
qplot(day, mails, data = daily, geom = "line", colo
qplot(day, mails, data = daily, geom = "smooth", co
qplot(day, variants, data = daily, geom = "line", c
qplot(wday, hour, data = wh, size = freq)
qplot(mpg, wt, data = mtcars)
qplot(mpg, wt, data = mtcars, colour = cyl)
```

```
> q
```

Name	Size	Modified
..		
04-large-data.html	4.2 MB	Feb 7, 2013, 9:37 AM
overplot.png	936.4 KB	Feb 7, 2013, 9:37 AM
transparent.png	863.2 KB	Feb 7, 2013, 9:37 AM
small.png	463.8 KB	Feb 7, 2013, 9:37 AM
caratcut.png	291.7 KB	Feb 7, 2013, 9:37 AM
	KB	Feb 7, 2013, 9:37 AM
		Feb 7, 2013, 9:37 AM
	KB	Feb 7, 2013, 9:37 AM
	KB	Feb 7, 2013, 9:37 AM
		Feb 7, 2013, 9:37 AM
04-large-		

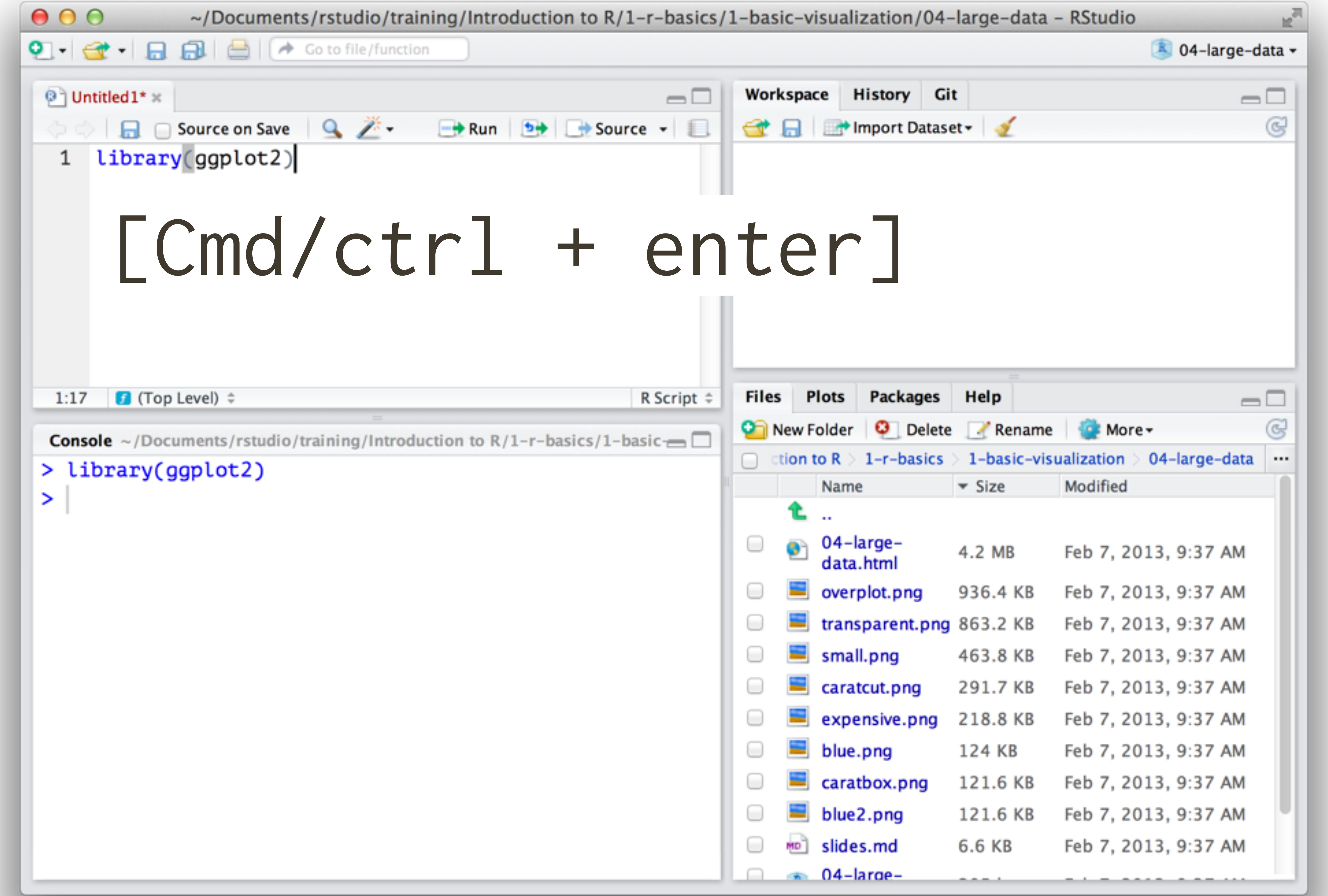
[Cmd/Ctrl + ↑]


~/Documents/rstudio/training/Introduction to R/1-r-basics/1-basic-visualization/04-large-data - RStudio
Go to file/function
04-large-data
Untitled1*
Source on Save
Run
Source
1:17
(Top Level)
R Script
Workspace
History
Git
Import Dataset
Console ~/Documents/rstudio/training/Introduction to R/1-r-basics/1-basic-
Files
Plots
Packages
Help
New Folder
Delete
Rename
More
ction to R › 1-r-basics › 1-basic-visualization › 04-large-data

```
1  library(ggplot2)
```

```
>
```

	Name	Size	Modified
	..		
□	04-large-data.html	4.2 MB	Feb 7, 2013, 9:37 AM
□	overplot.png	936.4 KB	Feb 7, 2013, 9:37 AM
□	transparent.png	863.2 KB	Feb 7, 2013, 9:37 AM
□	small.png	463.8 KB	Feb 7, 2013, 9:37 AM
□	caratcut.png	291.7 KB	Feb 7, 2013, 9:37 AM
□	expensive.png	218.8 KB	Feb 7, 2013, 9:37 AM
□	blue.png	124 KB	Feb 7, 2013, 9:37 AM
□	caratbox.png	121.6 KB	Feb 7, 2013, 9:37 AM
□	blue2.png	121.6 KB	Feb 7, 2013, 9:37 AM
□	slides.md	6.6 KB	Feb 7, 2013, 9:37 AM
□	04-large-		

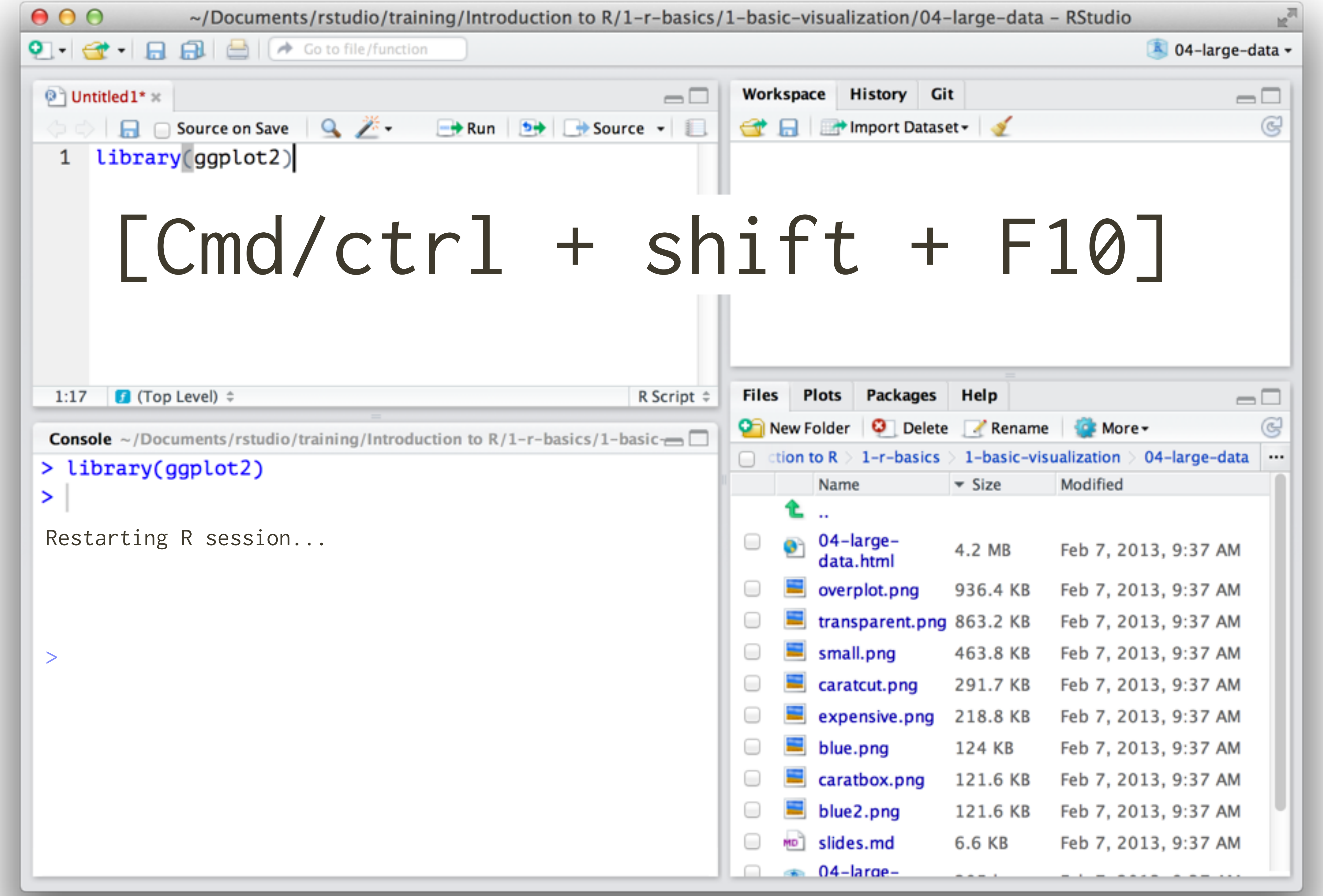


```
library(ggplot2)
```

[Cmd/ctrl + enter]

```
> library(ggplot2)
>
```

	Name	Size	Modified
	..		
□	04-large-data.html	4.2 MB	Feb 7, 2013, 9:37 AM
□	overplot.png	936.4 KB	Feb 7, 2013, 9:37 AM
□	transparent.png	863.2 KB	Feb 7, 2013, 9:37 AM
□	small.png	463.8 KB	Feb 7, 2013, 9:37 AM
□	caratcut.png	291.7 KB	Feb 7, 2013, 9:37 AM
□	expensive.png	218.8 KB	Feb 7, 2013, 9:37 AM
□	blue.png	124 KB	Feb 7, 2013, 9:37 AM
□	caratbox.png	121.6 KB	Feb 7, 2013, 9:37 AM
□	blue2.png	121.6 KB	Feb 7, 2013, 9:37 AM
□	slides.md	6.6 KB	Feb 7, 2013, 9:37 AM
□	04-large-		

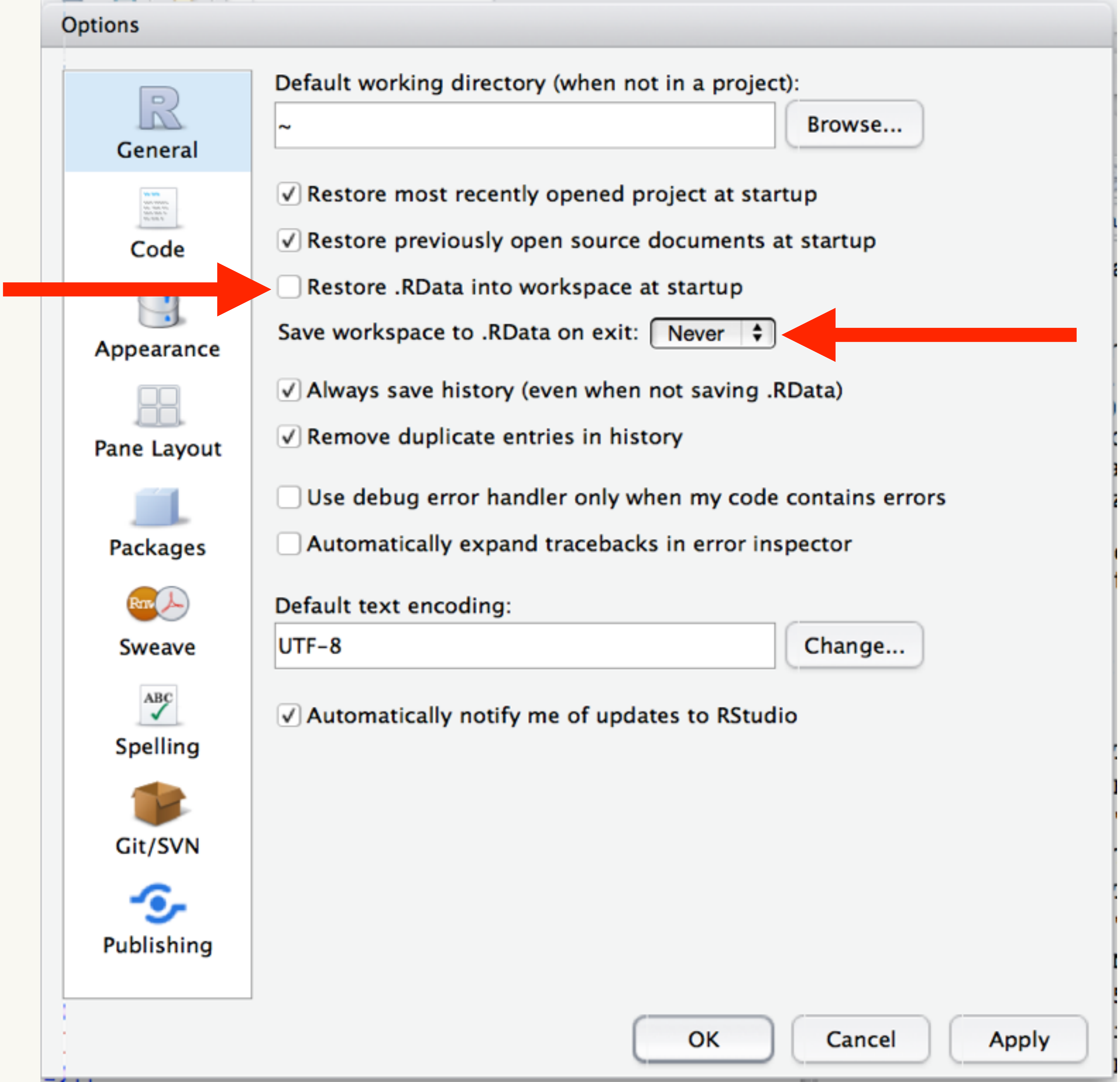


[Cmd/ctrl + shift + F10]

```
library(ggplot2)
```

```
> library(ggplot2)
> 

Restarting R session...

>
```

	Name	Size	Modified
	..		
□	04-large-data.html	4.2 MB	Feb 7, 2013, 9:37 AM
□	overplot.png	936.4 KB	Feb 7, 2013, 9:37 AM
□	transparent.png	863.2 KB	Feb 7, 2013, 9:37 AM
□	small.png	463.8 KB	Feb 7, 2013, 9:37 AM
□	caratcut.png	291.7 KB	Feb 7, 2013, 9:37 AM
□	expensive.png	218.8 KB	Feb 7, 2013, 9:37 AM
□	blue.png	124 KB	Feb 7, 2013, 9:37 AM
□	caratbox.png	121.6 KB	Feb 7, 2013, 9:37 AM
□	blue2.png	121.6 KB	Feb 7, 2013, 9:37 AM
□	slides.md	6.6 KB	Feb 7, 2013, 9:37 AM
□	04-large-		



Default working directory (when not in a project):

~

☑ Restore most recently opened project at startup

☑ Restore previously open source documents at startup

☐ Restore .RData into workspace at startup

Save workspace to .RData on exit: Never

☑ Always save history (even when not saving .RData)

☑ Remove duplicate entries in history

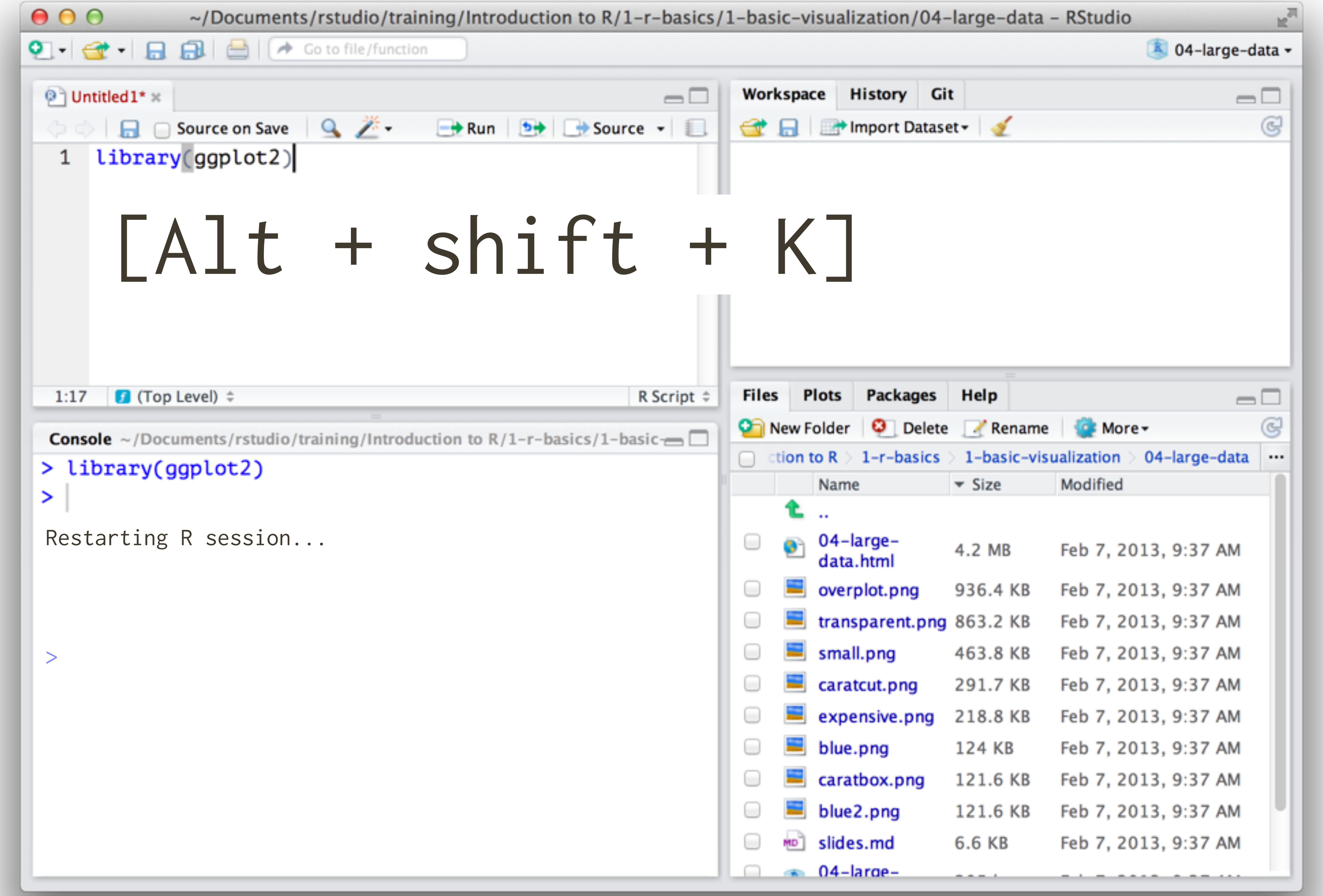
☐ Use debug error handler only when my code contains errors

☐ Automatically expand tracebacks in error inspector

Default text encoding:

UTF-8

☑ Automatically notify me of updates to RStudio



```
library(ggplot2)
```

```
> library(ggplot2)
> 

Restarting R session...



> 
```

Name	Size	Modified
..		
04-large-data.html	4.2 MB	Feb 7, 2013, 9:37 AM
overplot.png	936.4 KB	Feb 7, 2013, 9:37 AM
transparent.png	863.2 KB	Feb 7, 2013, 9:37 AM
small.png	463.8 KB	Feb 7, 2013, 9:37 AM
caratcut.png	291.7 KB	Feb 7, 2013, 9:37 AM
expensive.png	218.8 KB	Feb 7, 2013, 9:37 AM
blue.png	124 KB	Feb 7, 2013, 9:37 AM
caratbox.png	121.6 KB	Feb 7, 2013, 9:37 AM
blue2.png	121.6 KB	Feb 7, 2013, 9:37 AM
slides.md	6.6 KB	Feb 7, 2013, 9:37 AM
04-large-		

[Alt + shift + K]

Your turn

Check that you're familiar with the basic workflow outlined in http://r4ds.had.co.nz/workflow-scripts.html

Skim https://twitter.com/rstudiotips – which one do you think is most useful?