

# (A bit of) Advanced R

## Part 1 - R-base programming

Julien Chiquet

<http://github/jchiquet/CourseAdvancedR>

Université Paris Dauphine, Juin 2018



# Outline

- ① Control Statements
- ② Functions
- ③ Functionals

# References

Advanced R (Wickham, 2014), <http://adv-r.had.co.nz/>



A Language and Environment for Statistical Computing (R Core Team, 2017), <https://www.R-project.org/>



→ Many ideas/examples inspired/stolen there.

# Prerequisites

## Data Structure in base R

- ① Atomic vector (integer, double, logical, character)
- ② Recursive vector (list)
- ③ Factors
- ④ Matrices and array
- ⑤ Data Frame


→ Creation, Basic Operation, Manipulation, Representation

## Resources

- Advanced R, chapters 1.2, 1.3 (Wickham, 2014, <http://adv-r.had.co.nz/>)
- An introduction to R programming  
[http://julien.cremeriefamily.info/teachings\\_L3BI\\_ISV51.html](http://julien.cremeriefamily.info/teachings_L3BI_ISV51.html)

# Development environment I

## The Rstudio API

- A full API with code, interpreter, workspace and plots
- Package development and external code integration are easier
- Notebooks integration with Rmarkdown
- Interface with github  required tool for efficient development in R

## My favorites shortcuts

- `ctrl + return`: execute current selection in console
- `ctrl + 1/2/3/4`: navigate between panels
- `ctrl + down/up`: navigate between tabs
- `ctrl + shift + k`: knit current document

# Development environment II

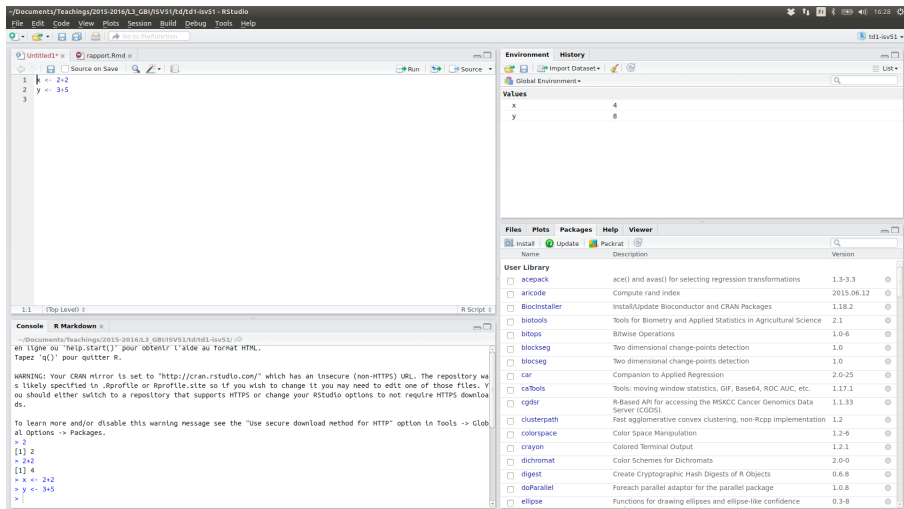


Figure 1: Screenshot of the Rstudio API

# Outline

① Control Statements

② Functions

③ Functionals

# Grouped expressions

## Syntax

```
{expr_1; expr_2; ...; expr_n }  
{  
  expr_1  
  ...  
  expr_n  
}
```

## Remarks

- the last value is sent back
- un group statement can be passed to a function



# Grouped expressions: examples

## Example 1

```
expr1 <- {a<-3; b<-5; a*b}  
expr1
```

```
## [1] 15
```

## Example 2

```
tmp <- 12  
expr2 <- {a<-3; b<-5; tmp<-a*b+tmp}
```

```
expr2
```

```
## [1] 27
```

```
tmp
```

```
## [1] 27
```

# Grouped expressions: examples

## Example 1

```
expr1 <- {a<-3; b<-5; a*b}  
expr1
```

```
## [1] 15
```

## Example 2

```
tmp <- 12  
expr2 <- {a<-3; b<-5; tmp<-a*b+tmp}
```

```
expr2
```

```
## [1] 27
```

```
tmp
```

```
## [1] 27
```

# Grouped expressions: examples

## Example 1

```
expr1 <- {a<-3; b<-5; a*b}  
expr1
```

```
## [1] 15
```

## Example 2

```
tmp <- 12  
expr2 <- {a<-3; b<-5; tmp<-a*b+tmp}
```

```
expr2
```

```
## [1] 27
```

```
tmp
```

```
## [1] 27
```

# Conditional statements: if,if/else,ifelse

## Standard syntax

```
if (condition) {  
  expr_1  
} else {  
  expr_2  
}
```

## Vectorial form

```
ifelse(condition, a, b)
```

## Remarks

- condition is logical, so use &, |, !, etc.
- else is optional
- elseif allows imbricating statements

# Conditional statements: example

```
partiel <- 11
DS <- 14
if (partiel > 6 & mean(DS,partiel) > 10) {
  cat("\nreçu(e).")
} else {
  cat("\nrecalé(e).")
}
```

```
##
## reçu(e).
```

## Exercice

Use the vectorial ifelse to send the full vector of results

```
partiel <- c(11,5,6,12,9,8,14)
DS <- c(14,16,12,12,19,12,7)

ifelse(partiel > 6 & rowMeans(cbind(DS,partiel)) > 10, "reçu(e)", "recalé(e)")

## [1] "reçu(e)" "recalé(e)" "recalé(e)" "reçu(e)" "reçu(e)" "recalé(e)"
## [7] "reçu(e)"
```

# Conditional statements: example

```
partiel <- 11
DS <- 14
if (partiel > 6 & mean(DS,partiel) > 10) {
  cat("\nreçu(e).")
} else {
  cat("\nrecalé(e).")
}
```

```
##
## reçu(e).
```

## Exercice

Use the vectorial ifelse to send the full vector of results

```
partiel <- c(11,5,6,12,9,8,14)
DS <- c(14,16,12,12,19,12,7)
```

```
ifelse(partiel > 6 & rowMeans(cbind(DS,partiel)) > 10, "reçu(e)", "recalé(e)")

## [1] "reçu(e)" "recalé(e)" "recalé(e)" "reçu(e)" "reçu(e)" "recalé(e)"
## [7] "reçu(e)"
```

# Conditional statement: switch

## Syntax

```
switch (expr,  
    expr_1 = do_1,  
    ...,  
    expr_n = do_n,  
    do_default  
)
```

## Remarks

- `expr` is either an integer or a character
- if an integer with value  $i$ , the  $i$ th expression `do_i` is evaluated
- if a string the expression `do_i` so that `expr == expr_i` is evaluated

# switch: examples

## integer form

```
expr <- 2  
switch(expr, cat("My value is 1"), cat("My value is 2"))
```

```
## My value is 2
```

```
expr <- 3  
switch(expr, cat("My value is 2"), cat("My value is 2"))
```

## character form

```
stat <- "variance"  
f_stat <- switch(stat,  
  "mean" = mean,  
  "variance" = var,  
  NULL)  
f_stat(1:10)
```

```
## [1] 9.166667
```



# loop statement: for

## Syntax

```
for (var in set) {  
    expr(var)  
}  
for (var in set) # avoid this syntax!!  
    expr(var)  
for (var in set) expr(var)
```

## Remarks

- var is the incremented variable
- set is a vector of the successive values
- generally slow compared to matricial/vectorize operation

# for loop: examples

## Example: C/C++ like

```
for (i in sample(1:5)) cat(i)
```

```
## 54231
```

```
v <- numeric(7)
```

```
for (i in seq_along(v)) v[i] <- i*3
```

## Exercise:

Use a for loop to display the colnames of the data frame `iris` which are not a factor, by completing the following piece of code

```
data(iris)
for (nom in colnames(iris)) {
  if (!is.factor(iris[,nom])) cat("",nom)
}
```

```
## Sepal.Length Sepal.Width Petal.Length Petal.Width
```

A more R-style way to do that is

```
cat(colnames(iris)[!sapply(iris, is.factor)])
```

```
## Sepal.Length Sepal.Width Petal.Length Petal.Width
```

# for loop: examples

## Example: C/C++ like

```
for (i in sample(1:5)) cat(i)
```

```
## 54231
```

```
v <- numeric(7)
```

```
for (i in seq_along(v)) v[i] <- i*3
```

## Exercise:

Use a for loop to display the colnames of the data frame `iris` which are not a factor, by completing the following piece of code

```
data(iris)
for (nom in colnames(iris)) {
  if (!is.factor(iris[,nom])) cat("",nom)
}
```

```
## Sepal.Length Sepal.Width Petal.Length Petal.Width
```

A more R-style way to do that is

```
cat(colnames(iris)[!sapply(iris, is.factor)])
```

```
## Sepal.Length Sepal.Width Petal.Length Petal.Width
```

# for loop: examples

## Example: C/C++ like

```
for (i in sample(1:5)) cat(i)
```

```
## 54231
```

```
v <- numeric(7)
```

```
for (i in seq_along(v)) v[i] <- i*3
```

## Exercise:

Use a for loop to display the colnames of the data frame `iris` which are not a factor, by completing the following piece of code

```
data(iris)
for (nom in colnames(iris)) {
  if (!is.factor(iris[,nom])) cat("",nom)
}
```

```
## Sepal.Length Sepal.Width Petal.Length Petal.Width
```

A more R-style way to do that is

```
cat(colnames(iris)[!sapply(iris, is.factor)])
```

```
## Sepal.Length Sepal.Width Petal.Length Petal.Width
```

# Loop statement: while, repeat

## Syntax

```
while (condition) {  
    expr  
}  
repeat {  
    expr  
}
```

## Remarks

- avoid imbrication (slow)
- can be interrupted/controlled with with break/next

```
repeat {  
    expr  
    if (condition) {break}  
}  
while (condition1){  
    expr_1  
    if (condition2) {next}  
    expr_2  
}
```

# Outline

① Control Statements

② Functions

③ Functionals

# Functions definition

## Syntax

```
my_func <- function(arg1,arg2, ...) {  
  expression  
}
```

## Remarks

- The last value of the expression is returned
- One must use a list to send back several objects
- `return()` is used only when you need to send a value at an early stage in the expression
- In R, functions are object like any others and can manipulated as such

# Function components I

Most of functions have three parts

- the `body()` (code inside the function)
- the `formals()` (list of arguments)
- the `environment()` (a set of bindings between symbols and objects, i.e, a place to store variables)

```
environment(var)
```

```
## <environment: namespace:stats>
```

```
formals(var)
```

```
## $x  
##  
##  
## $y  
## NULL  
##  
## $na.rm  
## [1] FALSE  
##  
## $use
```



# Function components II

`body(var)`

```
## {  
##   if (missing(use))  
##     use <- if (na.rm)  
##       "na.or.complete"  
##     else "everything"  
##   na.method <- pmatch(use, c("all.obs", "complete.obs", "pairwise.complete.obs",  
##     "everything", "na.or.complete"))  
##   if (is.na(na.method))  
##     stop("invalid 'use' argument")  
##   if (is.data.frame(x))  
##     x <- as.matrix(x)  
##   else stopifnot(is.atomic(x))  
##   if (is.data.frame(y))  
##     y <- as.matrix(y)  
##   else stopifnot(is.atomic(y))  
##   .Call(C_cov, x, y, na.method, FALSE)  
## }
```

# Lexical Scoping I

## Definition

Set of rule that governs how R looks up the value of a symbol

## Name masking

If a name is not defined inside a function, R looks a level up

```
y <- 2
func <- function(x) c(x,y)
func(4)
```

```
## [1] 4 2
```

This applies to function defined in another function

```
x <- 2
func <- function(y) {
  sub_func <- function(z) c(x,y,z)
  sub_func(5)
}
func(3)
```

```
## [1] 2 3 5
```

# Lexical Scoping II

## function vs variable

R makes the distinction between variable and function names

```
n <- function(x) x/2
f <- function() {n <- 10 ; n(n)}
f()
```

```
## [1] 5
```

## Fresh star

An environment is created at *each time a function is called*

```
f <- function() {
  a <- ifelse(exists("a"), a + 1, 1)
  print(a)
}
f()
```

```
## [1] 1
```

```
f()
```

```
## [1] 1
```

```
a <- 4
f()
```

```
## [1] 5
```

```
f()
```

```
## [1] 5
```

# Function arguments

## Calling function

Arguments can be specified 1. by name 2. by partial name 3. by position  
Here are some stupid (but correct) call to `mean(x=,trim=,na.rm=)`

```
mean(1:10, n = T)
```

```
## [1] 5.5
```

```
mean(1:10, , FALSE)
```

```
## [1] 5.5
```

```
mean(1:10, 0.05, FALSE)
```

```
## [1] 5.5
```

```
mean(, TRUE, x = c(1:10, NA))
```

```
## [1] 5.5
```

## Exercise

Clarify the following function calls

```
x <- sample(replace = TRUE, 20, x = c(1:10, NA))  
y <- runif(min = 0, max = 1, 20)  
cor(m = "k", y = y, u = "p", x = x)
```

# Default arguments

## Arguments can have default values in R

```
f <- function(a = 1, b = 2) c(a,b)
f()
```

```
## [1] 1 2
```

## The missing function

You can check whether an argument was passed or not with `missing`:

```
f <- function(a = 1, b = 2) c(missing(a),missing(b))
f(a)
```

```
## [1] FALSE TRUE
```

Hence, you can assign a default value *a posteriori*

```
f <- function(a, b = 2) {
  if (missing(a)) a <- 3
  c(a, b)
}
f(a)
```

```
## [1] 4 2
```

# Default arguments

Arguments can have default values in R

```
f <- function(a = 1, b = 2) c(a,b)
f()
```

```
## [1] 1 2
```

The missing function

You can check whether an argument was passed or not with `missing`:

```
f <- function(a = 1, b = 2) c(missing(a),missing(b))
f(a)
```

```
## [1] FALSE TRUE
```

Hence, you can assign a default value *a posteriori*

```
f <- function(a, b = 2) {
  if (missing(a)) a <- 3
  c(a, b)
}
f(a)
```

```
## [1] 4 2
```

# Lazy evaluation

Arguments are evaluated only if they are used, which is known as “lazy evaluation”

```
f <- function(a = 1, b = 4*a) c(a,b)
f()
```

```
## [1] 1 4
```

```
f(43)
```

```
## [1] 43 172
```

Even better (or worse...)

```
f <- function(a = 1, b = d) {
  d <- 4 + 2 * a; c(a,b)
}
f()
```

```
## [1] 1 6
```

```
f(4)
```

```
## [1] 4 12
```

# The ... argument

The argument ... matches any argument not otherwise matched

- useful when collecting argument to call another function
- do not need to specify the name of required argument
- the counterpart is that any misspelled argument is passed to ... and show no warning

## Example: plot

Many argument in plot are passed to the par function that manages the graphical parameters:

```
plot(1:5, col = "red")  
plot(1:5, lty = "dotted")
```

## Capturing ...

list() can be used to easily capture arguments passed with ...

```
f <- function(...) names(list(...))  
f(a = 1, b = 2)
```

```
## [1] "a" "b"
```



## Calling function with a list

The `do.call` function constructs and executes a function call from a name or a function and a list of arguments to be passed to it:

```
do.call(mean, list(x = 1:10, trim = 0.05, na.rm = FALSE))
```

```
## [1] 5.5
```

### Exercise

Suppose the outputs of 100 simulations are stored in a list like that

```
class(res)
```

```
## [1] "list"
```

```
res[[1]]
```

```
##   method      mse      timing
## 1  lasso 0.8222197  0.05823235
## 2  ridge 0.6912516  0.91282404
## 3  bayes 0.9756149 123.53501949
```

```
length(res)
```

```
## [1] 100
```

How would you store them in a single data frame?

```
head(do.call(rbind, res), 2)
```

```
##   method      mse      timing
## 1  lasso 0.8222197 0.05823235
## 2  ridge 0.6912516 0.91282404
```

# Calling function with a list

The `do.call` function constructs and executes a function call from a name or a function and a list of arguments to be passed to it:

```
do.call(mean, list(x = 1:10, trim = 0.05, na.rm = FALSE))
```

```
## [1] 5.5
```

## Exercise

Suppose the outputs of 100 simulations are stored in a list like that

```
class(res)
```

```
## [1] "list"
```

```
res[[1]]
```

```
##      method      mse      timing
## 1  lasso 0.8222197  0.05823235
## 2  ridge 0.6912516  0.91282404
## 3  bayes 0.9756149 123.53501949
```

```
length(res)
```

```
## [1] 100
```

How would you store them in a single data frame?

```
head(do.call(rbind, res), 2)
```

```
##      method      mse      timing
## 1  lasso 0.8222197  0.05823235
## 2  ridge 0.6912516  0.91282404
```

# Calling function with a list

The `do.call` function constructs and executes a function call from a name or a function and a list of arguments to be passed to it:

```
do.call(mean, list(x = 1:10, trim = 0.05, na.rm = FALSE))
```

```
## [1] 5.5
```

## Exercise

Suppose the outputs of 100 simulations are stored in a list like that

```
class(res)
```

```
## [1] "list"
```

```
res[[1]]
```

```
##      method      mse      timing
## 1  lasso 0.8222197  0.05823235
## 2  ridge 0.6912516  0.91282404
## 3  bayes 0.9756149 123.53501949
```

```
length(res)
```

```
## [1] 100
```

How would you store them in a single data frame?

```
head(do.call(rbind, res), 2)
```

```
##      method      mse      timing
## 1  lasso 0.8222197  0.05823235
## 2  ridge 0.6912516  0.91282404
```

# Infix functions

## Definition

Infix function (contrary to prefix functions) are function where the name comes between the argument (like “-” or “+”).

R comes with the following infix functions predefined: `%%`, `%*%`, `%/%`, `%in%`, `%o%`, `%x%`, `:`, `::`, `:::`, `$`, `@`, `^`, `*`, `/`, `+`, `-`, `>`, `>=`, `<`, `<=`, `==`, `!=`, `!`, `&`, `&&`, `|`, `||`, `~`, `<-`, `<<-`

## Example

Can be use to define operator

```
`%+%` <- function(x,y) paste(x,y)
"Université" %+% "Paris" %+% "Dauphine"
```

```
## [1] "Université Paris Dauphine"
```

## Exercise

Create infix functions for intersection, union and setdiff and test it on simple vectors.

# Primitive functions: definition

- Primitive functions are functions from the base package that call C code directly
- Primitive functions do not contain R code, as so

```
sum
```

```
## function (... , na.rm = FALSE) .Primitive("sum")
```

```
formals(sum)
```

```
## NULL
```

```
body(sum)
```

```
## NULL
```

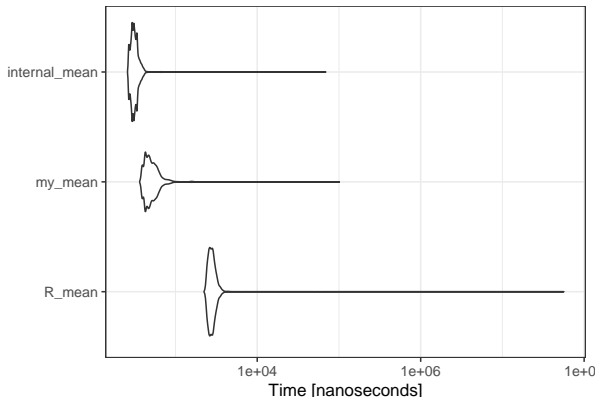
```
environment(sum)
```

```
## NULL
```

# Primitive functions: performance

Function defined internally (either with `.Primitive` either called via `.Internal`) are sometimes incredibly faster (written in C), but cannot be called directly in packages submitted to CRAN.

```
x <- rnorm(100)
R_mean <- function(x) mean(x)
my_mean <- function(x) sum(x)/length(x)
internal_mean <- function(x) .Internal(mean(x))
```



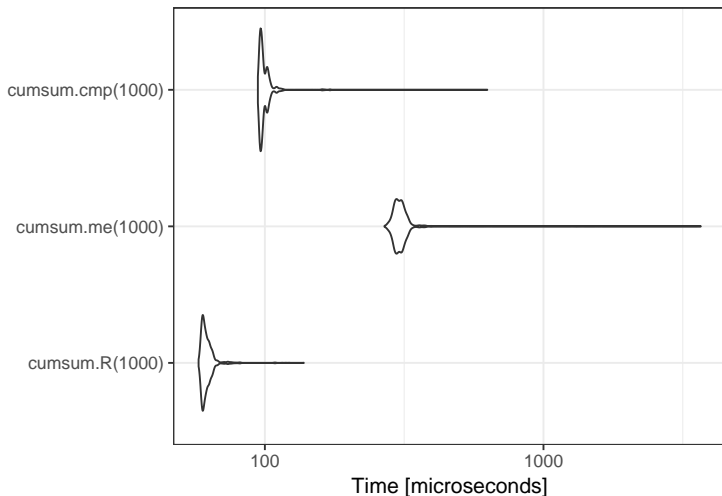
# Compile your functions with `base::compiler` I

## The R byte code compiler

The function `cmpfun` compiles the body of your function and returns a new function with the same formals and the body replaced by the compiled body expression.

```
cumsum.R <- function(n) {  
  x <- rnorm(n)  
  cumsum(x)  
}  
  
cumsum.me <- function(n) {  
  x <- rnorm(n)  
  res <- 0  
  for (i in 1:length(x))  
    res <- res + x[i]  
  res  
}  
  
cumsum.cmp <- compiler::cmpfun(cumsum.me)
```

## Compile your functions with `base::compiler` II



- If you cannot avoid a loop, you will save some time
- Can be set automatically with `compiler::enableJIT(3)`



# Outline

- ① Control Statements
- ② Functions
- ③ Functionals**

# Functional for matrix and arrays: apply

## Definition

Applies a function along a dimension of an array (row/columns of matrix).

```
apply(array, dim, FUN, ...)
```

## Example

```
mat <- matrix(1:6, 2, 3)
apply(mat, 2, max)
```

```
## [1] 2 4 6
```

```
arr <- array(1:12, c(2,3,2))
apply(arr, 3, colMeans)
```

```
##      [,1] [,2]
## [1,]  1.5  7.5
## [2,]  3.5  9.5
## [3,]  5.5 11.5
```

# Other array functionals: sweep

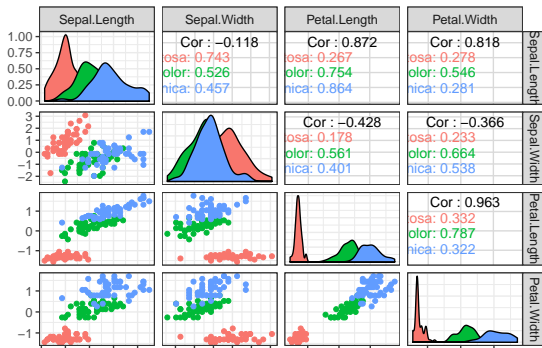
## Definition

'Sweep out' a summary statistic along a dimension of an array for a given function

```
sweep(array, dim, stat, FUN, ...)
```

## Example

```
out <- sweep(iris[, 1:4], 2, colMeans(iris[, 1:4]), "-") ## center
out <- sweep(out, 2, sqrt(colSums(out^2)/(nrow(out) - 1)), "/") ## scale
iris_sc <- data.frame(out, Species = iris$Species)
GGally::ggpairs(iris_sc, columns = 1:4, ggplot2::aes(colour = Species))
```



# Other array functionals: outer

## Definition

```
outer(array1, array2, FUN, ...)
```

## Example

The more basic example is the kronecker product, but FUN can be anything!

```
outer(1:3, 1:4, "*")
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    1    2    3    4
## [2,]    2    4    6    8
## [3,]    3    6    9   12
```

```
1:3 %o% 1:4
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    1    2    3    4
## [2,]    2    4    6    8
## [3,]    3    6    9   12
```

# Grouped functional: `tapply`

## Definition

Applies a function on a vector partitioned by a factor: combine a `split` + `lapply` operation

## Example

```
with(iris, tapply(Sepal.Length, Species, mean)) # readable
```

```
##      setosa versicolor  virginica  
##      5.006      5.936      6.588
```

```
with(iris, lapply(split(Sepal.Length, Species), mean)) # still ok (I think)
```

```
## $setosa  
## [1] 5.006  
##  
## $versicolor  
## [1] 5.936  
##  
## $virginica  
## [1] 6.588
```

```
stat <- c() # less readable (and naming is lost)  
for (l in levels(iris$Species))  
  stat <- c(stat, mean(iris$Sepal.Length[iris$Species == l]))  
stat
```

```
## [1] 5.006 5.936 6.588
```

# Functionals for list: Reduce and co

## Definition

'Reduce' uses a binary function to successively combine the elements of a given vector

```
Reduce(FUN, vector, init, right = FALSE, accumulate = FALSE)  
Reduce(f, 1:3) <-> f(f(1,2))
```

## Exercise

Consider a list of vectors of integer.

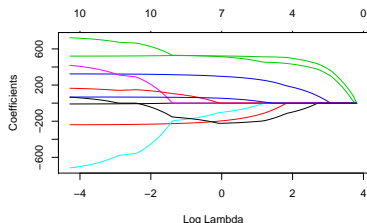
```
my_list <- replicate(10, sample.int(20, 10, replace = TRUE), simplify = FALSE)
```

Find which values occur in every element of the list visited so far, starting from the end.

# Example: “jackknifing” a lasso solution path

A single Lasso fit of the diabetes data set

```
library(glmnet)
library(lars) # the diabetes data set (part of the lars package)
data(diabetes)
x <- diabetes$x; y <- diabetes$y; n <- length(y)
lasso <- glmnet(x,y)
plot(lasso, xvar = "lambda")
```



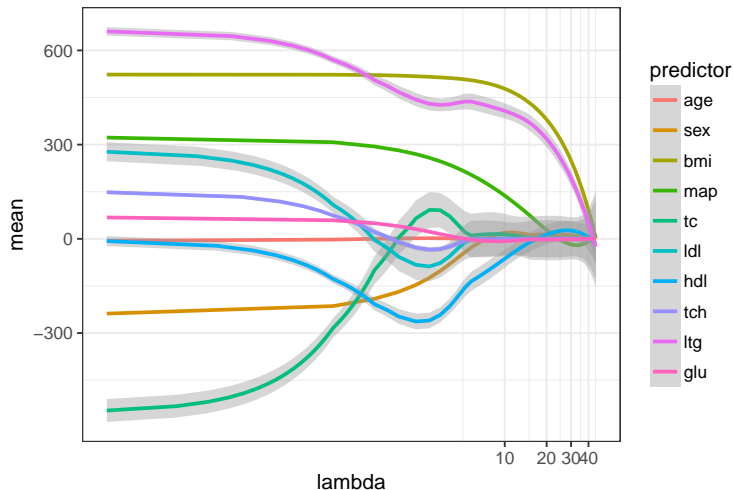
Compute the regularization paths for all subsets, removing one individual at once

```
paths <- lapply(1:n, function(i) {
  glmnet(x[-i, ], y[-i], lambda = lasso$lambda)$beta
})
```

# A Reduce example: “jackknifing” a lasso solution path II

Computing the envelop around the average regularization path with Reduce

```
mean.path <- Reduce("+", paths)/n  
sdev.path <- sqrt(Reduce("+", lapply(paths, function(path) path**2))/n -  
                  mean.path**2)
```





# References I

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>

Wickham, H. (2014). *Advanced r*. CRC Press. Retrieved from <http://adv-r.had.co.nz/>