

Introducción a la ciencia reproducible con R

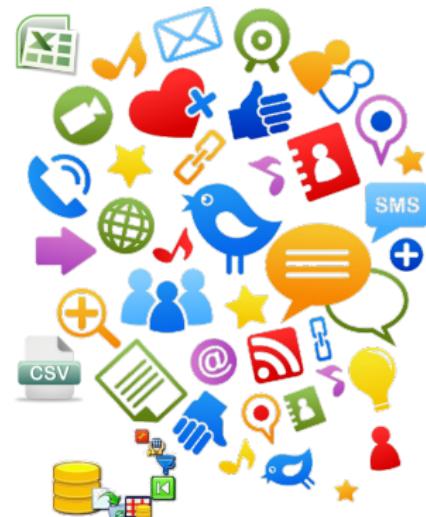
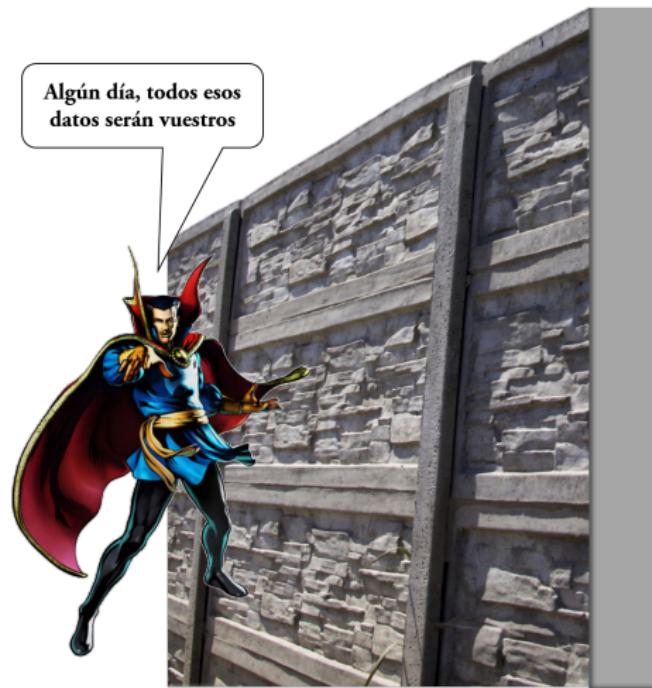
Document Freedom Day - OSLUGR

Francisco Charte (@fcharte, fcharte.com)

25 marzo 2015

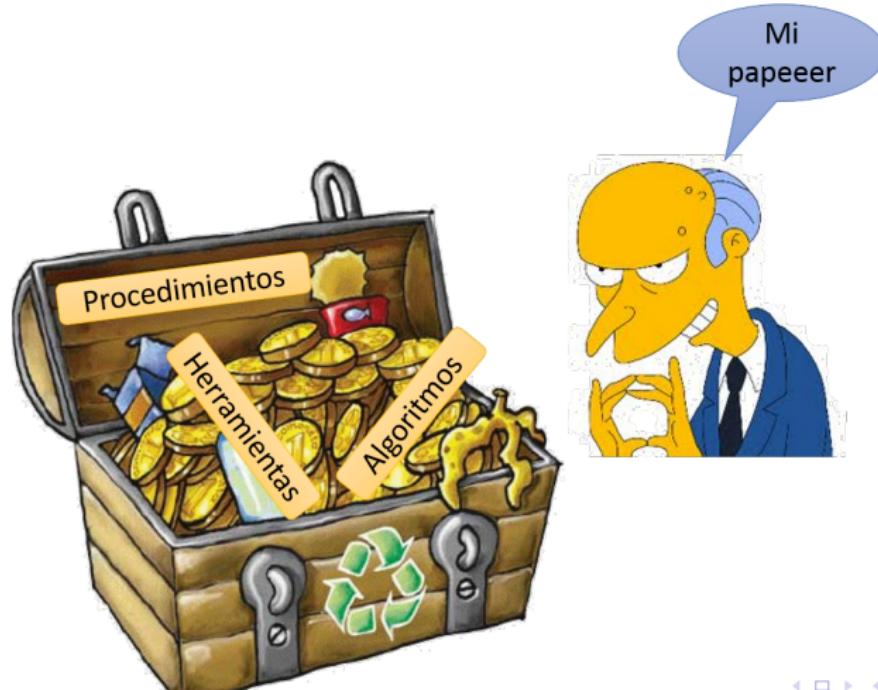
En qué consiste la ciencia reproducible

Datos disponibles y accesibles



En qué consiste la ciencia reproducible

Procedimientos, algoritmos y herramientas disponibles y accesibles



En qué consiste la ciencia reproducible

Las acciones manuales no son fácilmente reproducibles

Nothing can stop automation



Introducción a R

The screenshot shows the RStudio interface with the following components:

- Code Editor:** Displays R code in the `RBenchmarkings.R` file. The code includes sections for a presentation, a reproducibility check, and a plot generation.
- File Browser:** Shows the project structure with files like `graphics.R`, `CRANDownloads.R`, and `ClenciaReproducible.Rmd`.
- Plot Viewer:** Displays a line graph titled "CRAN downloads for package mldr". The Y-axis is "Number of downloads" (0 to 25) and the X-axis is "Date" (from 2013-01-01 to 2015-01-01). The plot shows several peaks, with a major peak around January 2015 reaching approximately 25 downloads.

Importación de datos - Desde archivos

- ▶ CSV: `read.table()`
- ▶ Paquete `foreign`
 - ▶ Stata: `read.dta()`
 - ▶ SPSS: `read.spss()`
 - ▶ SAS: `read.ssd()`
 - ▶ dBase: `read.dbf()`
 - ▶ ARFF: `read.arff()`
 - ▶ Octave: `read.octave()`
- ▶ Excel: `loadWorkbook()` - Paquete `XLConnect`
- ▶ OpenOffice/LibreOffice: `read.ods()` - Paquete `ROpenOffice`
- ▶ XML: `xmlParse()` - Paquete `XML`

Importación de datos - Desde bases de datos

- ▶ SQLite: `RSQLite`
- ▶ Oracle: `ROracle`
- ▶ MySQL: `RMySQL`
- ▶ PostgreSQL: `RPostgreSQL`
- ▶ ODBC/JDBC: `RODBC/RJDBC`
- ▶ MongoDB: `rmongodb`
- ▶ CouchDB: `R4CouchDB`
- ▶ Cassandra: `RCassandra`

Importación de datos - Otras fuentes

- ▶ Hadoop: RHadoop
- ▶ Spark: SparkR
- ▶ JSON: jsonlite
- ▶ Web: parseHTML() y readHTMLTable() - Paquete XML
- ▶ Web: GET() - Paquete httr
- ▶ Web: getURL() y getForm() - Paquete RCurl

Importación de datos - Ejemplo

Tags de los posts del foro de cocina de Stack Exchange

```
library(XML)
library(tm)
library(wordcloud)
content<-xmlTreeParse(filename)
...
docs<-Corpus(VectorSource(content))
docs<-tm_map(docs,
content_transformer(tolower))
...
wordcloud(DocumentTermMatrix(docs))
```



Importación de datos - Ejemplo

Datos obtenidos de la API de GitHub en formato JSON
runGitHub('GitHubMining', 'fcharte')

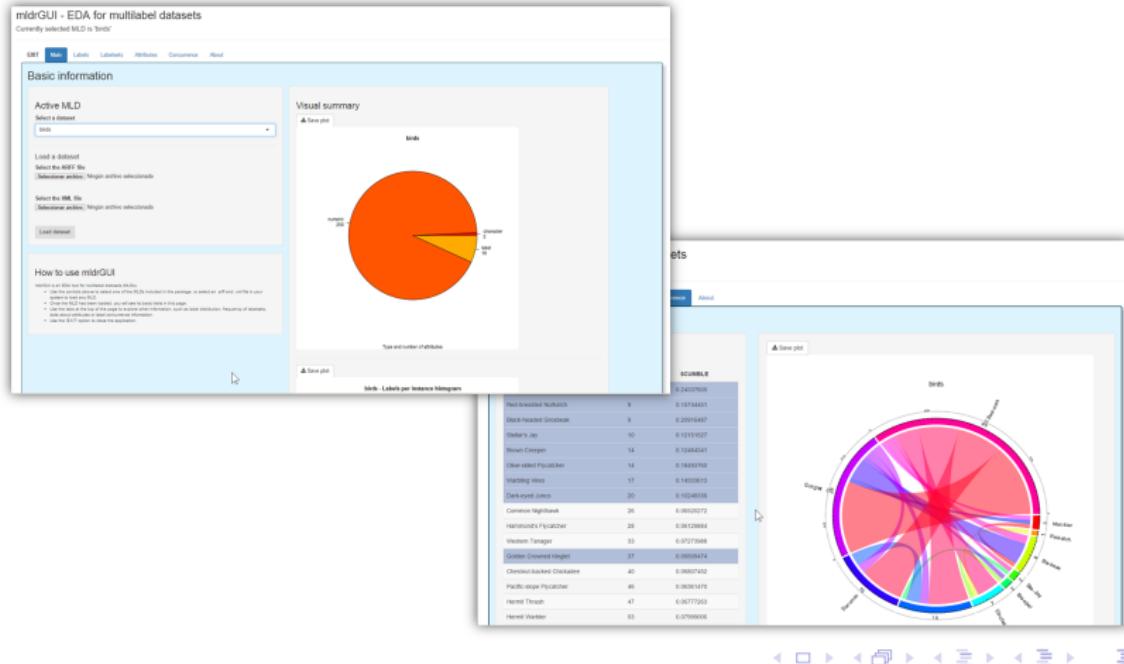
The screenshot shows the GitHubMining R package interface. It includes:

- API rate limits:** Shows Core Limit (5000), Search Limit (30), Remaining (1131), and Reset (2015-03-22 13:57:12).
- Github user account:** Fields for User name (fcharte) and Password.
- Users:** A table listing GitHub users with columns: Login, Name, Repos, Contribs, Followers, Following, Registered, and LastUpdate.

Login	Name	Repos	Contribs	Followers	Following	Registered	LastUpdate
rankingfaker	Falso commiteador	2	85037	0	0	2015-03-03	2015-03-10
stringparser	Javier Carrillo	23	2791	35	131	2014-05-01	2015-03-21
vterrón	Víctor Terrón	17	2199	62	5	2012-03-26	2015-03-22
ernestoalejo	Ernesto Alejo	32	1675	10	4	2011-08-05	2015-03-21
pleonex	Benito Palacios	29	1398	21	13	2012-12-22	2015-03-22
JJ	Juan Julián Merelo Guervós	160	1135	156	33	2008-02-20	2015-03-22
Amab	Juan Miguel Boyero Corral	2	1083	13	13	2010-10-31	2015-03-20
M42	Mario Román	16	751	40	49	2013-08-29	2015-03-21
frenlu	Fran	14	605	11	31	2010-09-23	2015-03-20

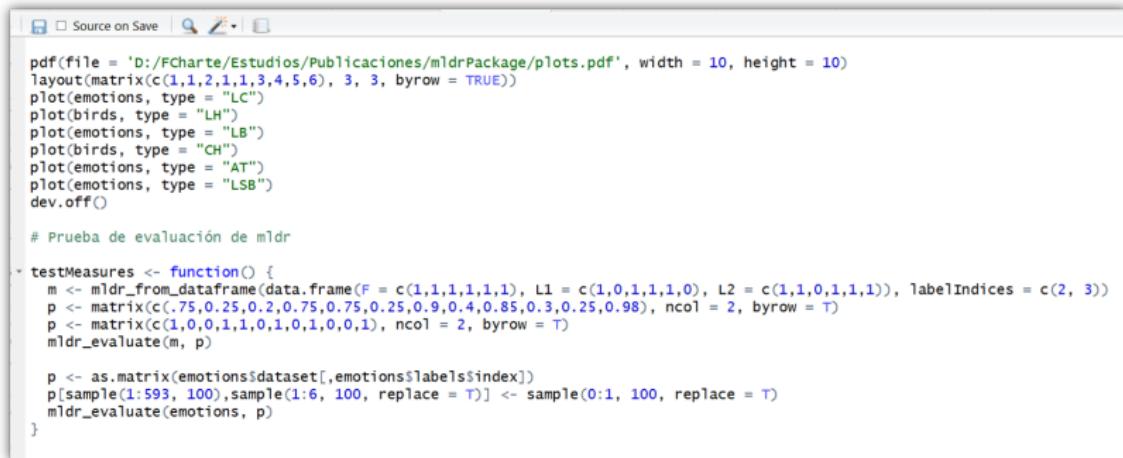
Exploración y análisis - Herramientas interactivas

Fáciles de utilizar - Útiles para explorar



Exploración y análisis - Scripts

Automatizan tareas - Fácil reproducción (`pdf()`, `xtable()`, etc.)



```
pdf(file = 'D:/FCharte/Estudios/Publicaciones/mldrPackage/plots.pdf', width = 10, height = 10)
layout(matrix(c(1,1,2,1,1,3,4,5,6), 3, 3, byrow = TRUE))
plot(emotions, type = "LC")
plot(birds, type = "LH")
plot(emotions, type = "LB")
plot(birds, type = "CH")
plot(emotions, type = "AT")
plot(emotions, type = "LSB")
dev.off()

# Prueba de evaluación de mldr

testMeasures <- function() {
  m <- mldr_from_dataframe(data.frame(F = c(1,1,1,1,1,1), L1 = c(1,0,1,1,1,0), L2 = c(1,1,0,1,1,1)), labelIndices = c(2, 3))
  p <- matrix(c(.75,0.25,0.2,0.75,0.75,0.25,0.9,0.4,0.85,0.3,0.25,0.98), ncol = 2, byrow = T)
  p <- matrix(c(1,0,0,1,1,0,1,0,1,0,0,1), ncol = 2, byrow = T)
  mldr_evaluate(m, p)

  p <- as.matrix(emotions$dataset[, emotions$labels$index])
  p[sample(1:593, 100), sample(1:6, 100, replace = T)] <- sample(0:1, 100, replace = T)
  mldr_evaluate(emotions, p)
}
```

Exploración y análisis - Libro y ejemplos

fcharte / ExploraVisualizaconR

Análisis exploratorio y visualización de datos con R — Edit

63 commits · 1 branch · 0 releases · 1 contributor

branch: master

Potential solutions to previous problems section

File	Description	Last Commit
scripts	Write xtable to file examples	2 months ago
ExploraVisualizaConR-FCharte.pdf	Potential solutions to previous problems section	a month ago
README.md	Update README.md	3 months ago

README.md

Análisis exploratorio y visualización de datos con R

Libro y ejemplos prácticos sobre el uso de la herramienta/lenguaje R para la carga, exploración y representación gráfica de datos. Dirigido a aquellos que comienzan a utilizar R.

Exploración y análisis - Documentos reproducibles

RBenchmarkings

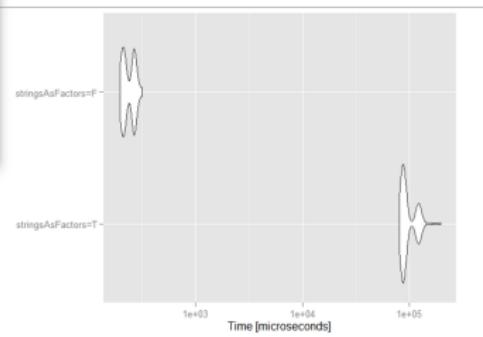
Generating a data.frame containing character data with and without stringsAsFactors

With this code I want to test the difference between using `stringsAsFactors = TRUE` VERSUS `stringsAsFactors = FALSE` while creating a new data.frame.

```
numElements <- 1e6
someStrings <- sapply(1:10, function(x) paste(sample(c(LETTERS), 10, replace = TRUE), collapse = ""))
aNumericVector <- runif(numElements)
aStringVector <- sample(someStrings, numElements, replace = TRUE)
bStringVector <- sample(someStrings, numElements, replace = TRUE)

result <- microbenchmark(
  data.frame(aNumericVector, aStringVector, bStringVector, stringsAsFactors = TRUE),
  data.frame(aNumericVector, aStringVector, bStringVector, stringsAsFactors = FALSE)
)

## Unit: relative
##      expr      min       lq     mean   median       uq      max
## stringsAsFactors=T 320.012 307.7241 304.4763 255.215 364.2376 370.7762
## stringsAsFactors=F  1.000  1.0000  1.0000  1.000  1.0000  1.0000
##      nval
##      100
##      1000
```



Publicación de datos - Presentaciones