Introducción a la ciencia reproducible con R

Document Freedom Day - OSLUGR

Francisco Charte (@fcharte, fcharte.com)

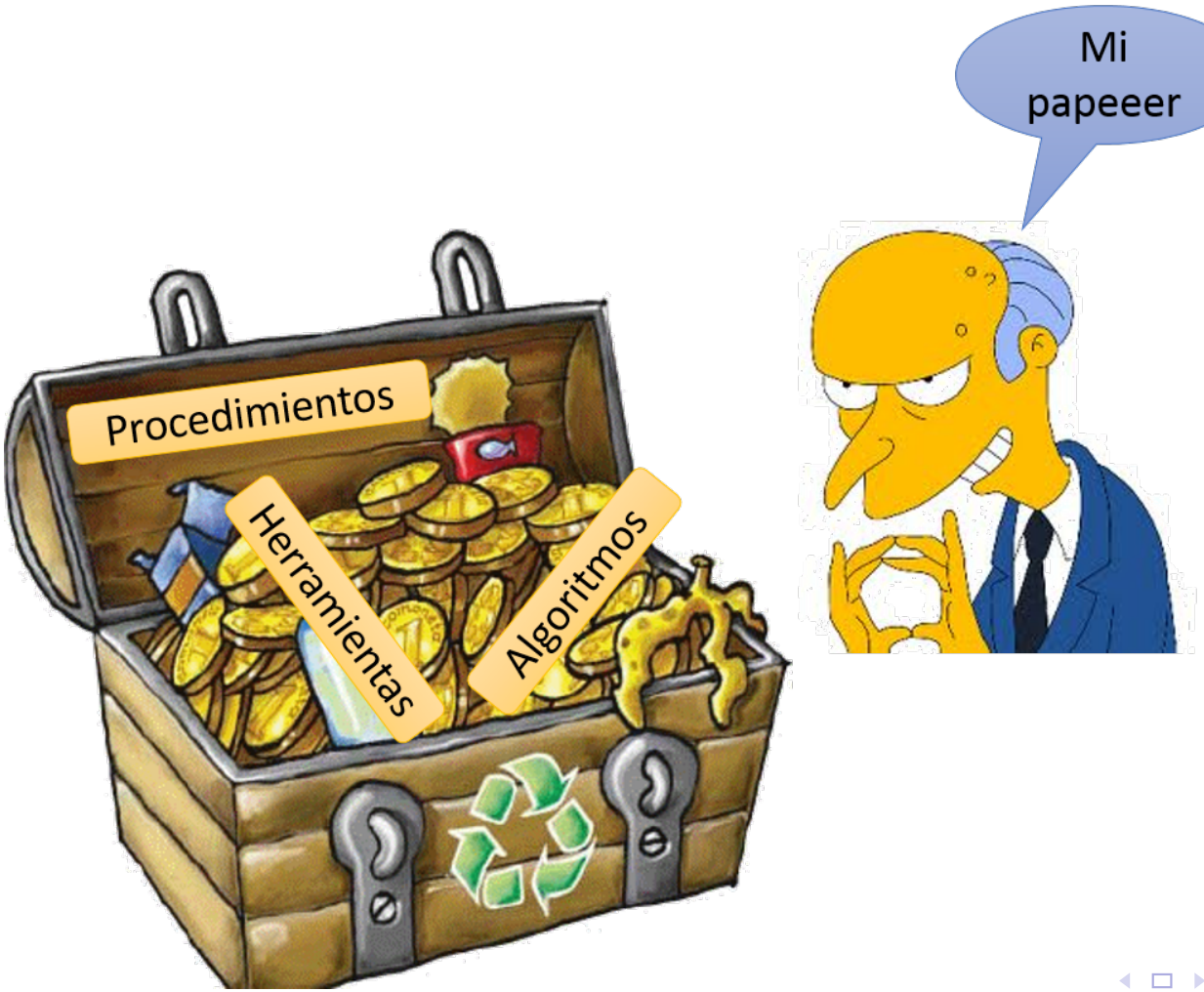
25 marzo 2015

En qué consiste la ciencia reproducible

Datos disponibles y accesibles

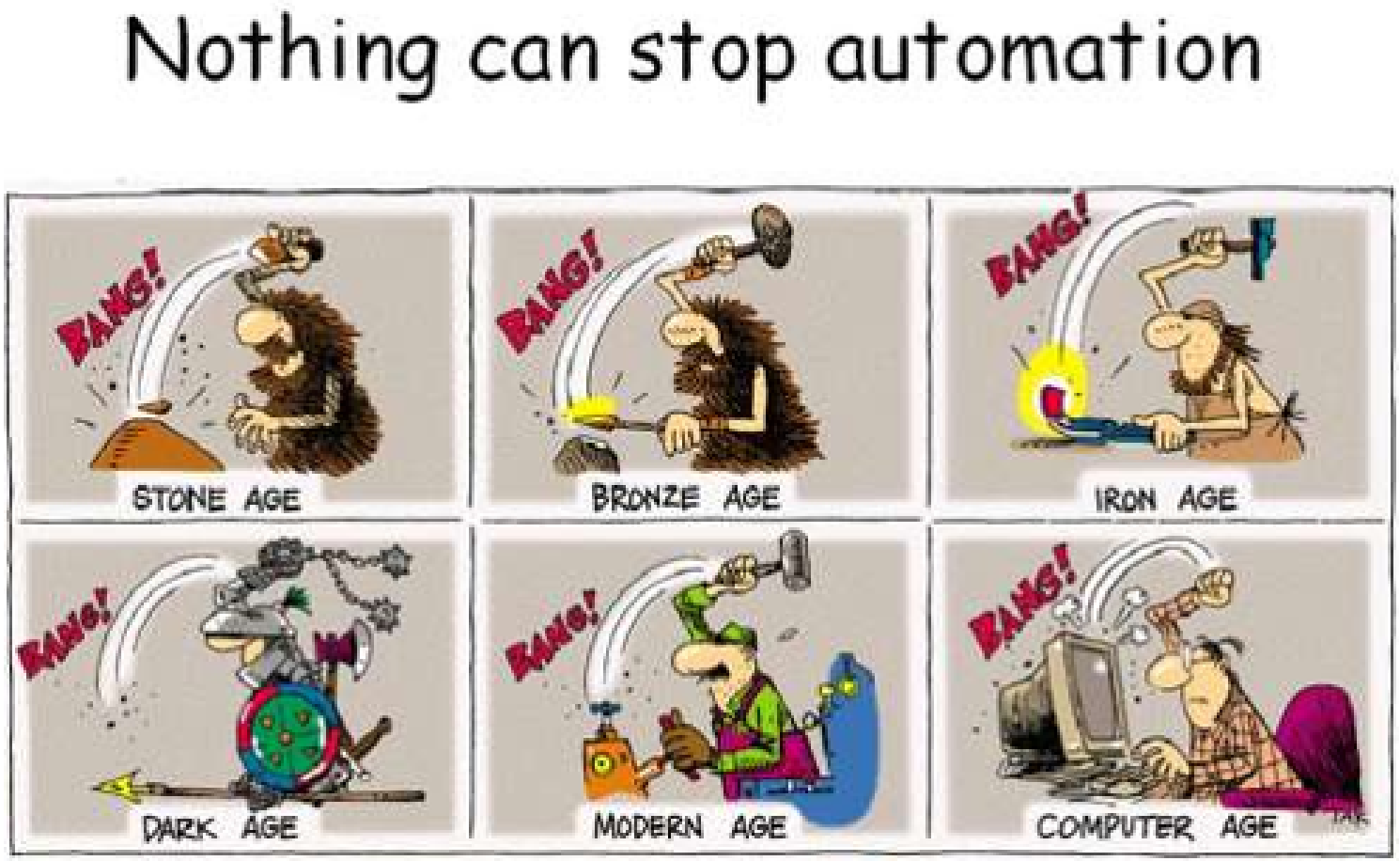
En qué consiste la ciencia reproducible

Procedimientos, algoritmos y herramientas disponibles y accesibles

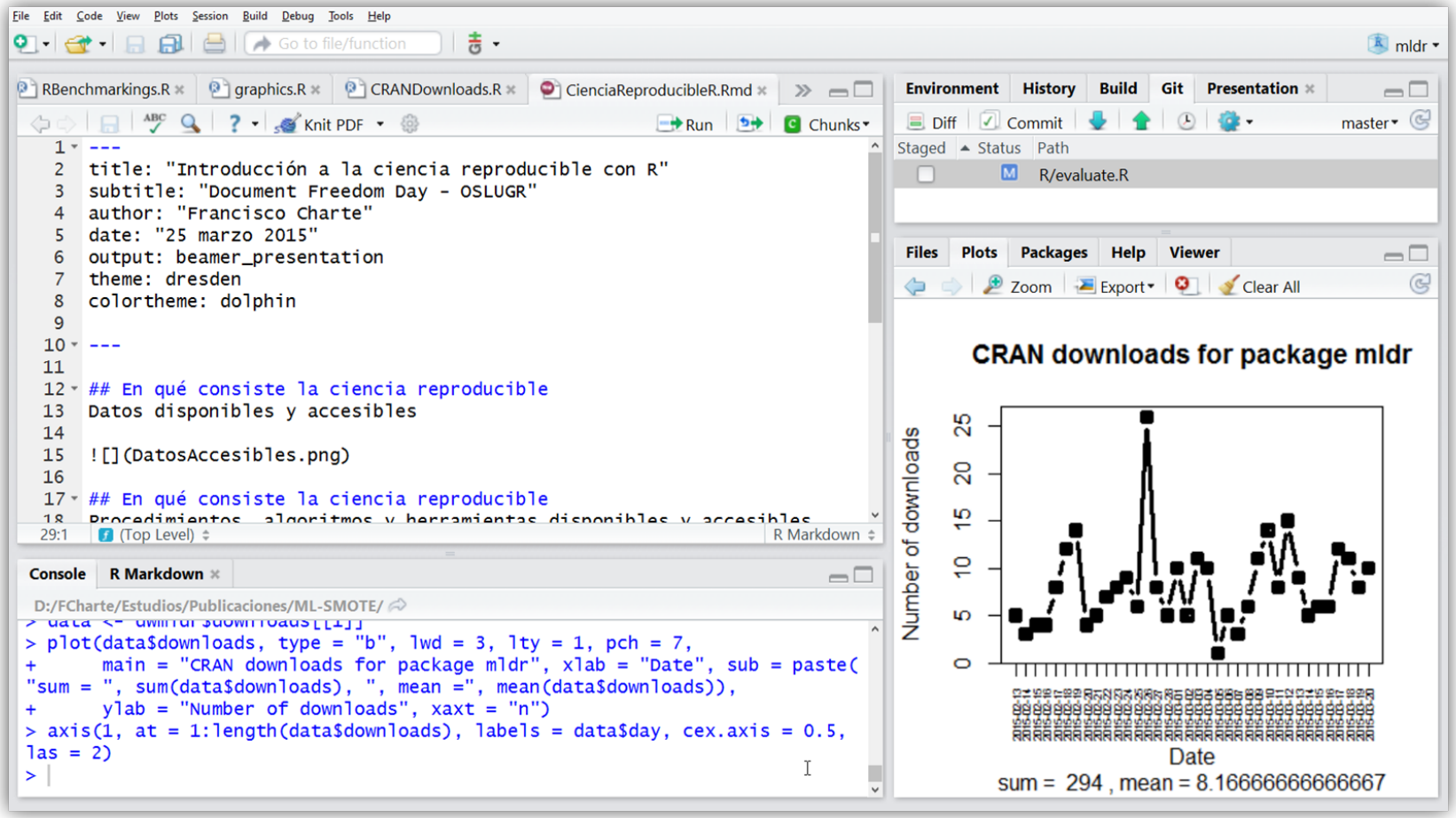


En qué consiste la ciencia reproducible

Las acciones manuales no son fácilmente reproducibles



Introducción a R



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`

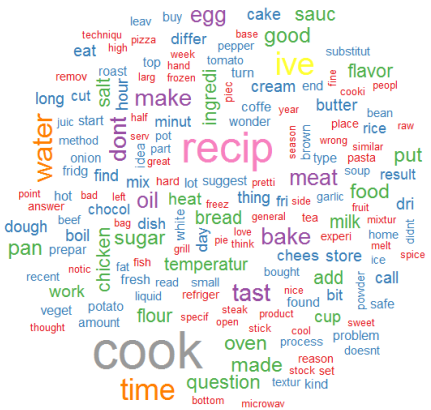
Importación de datos - Otras fuentes

- 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

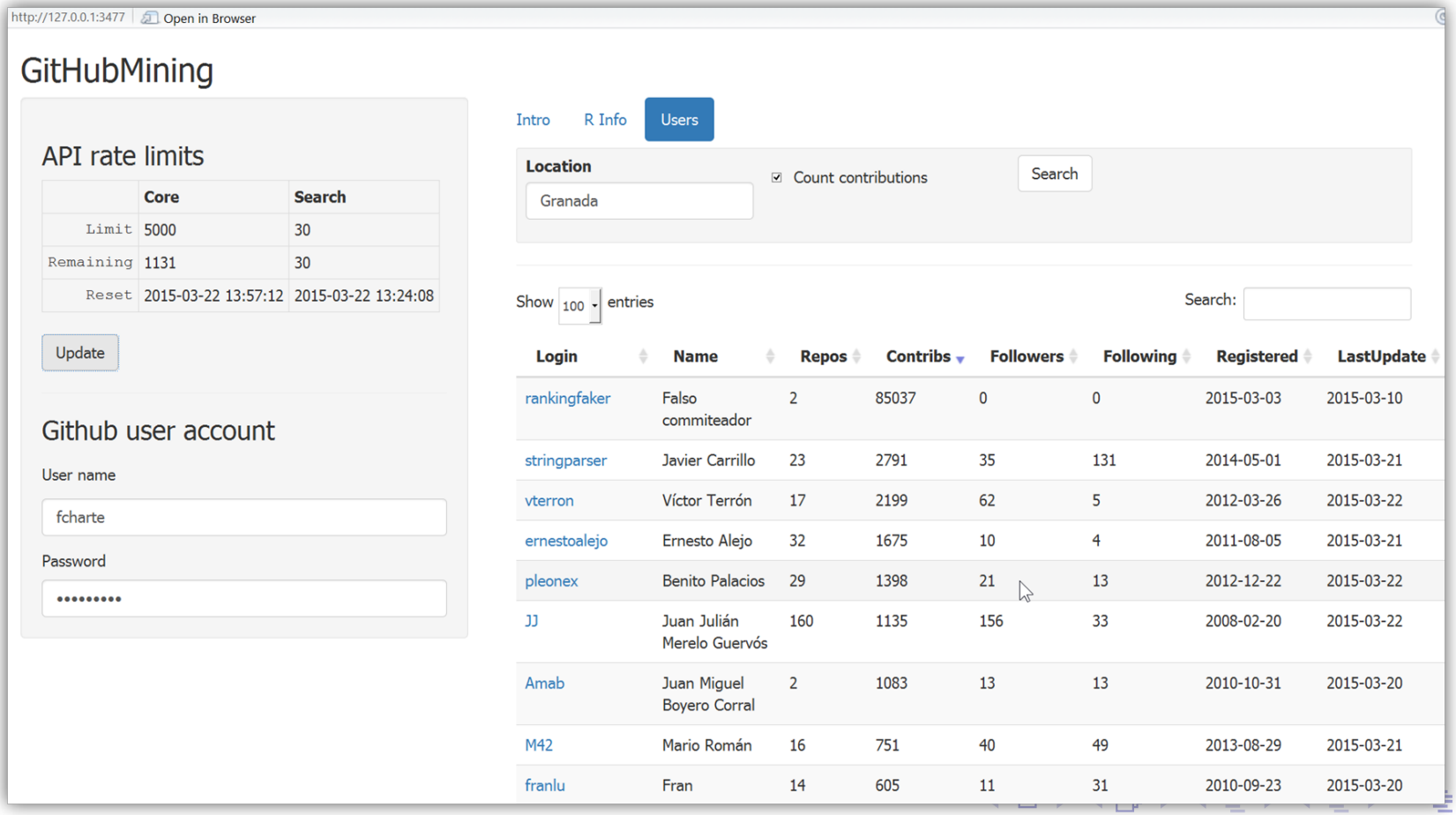
```
library(XML)
library(tm)
library(wordcloud)
content<-xmlTreeParse(filename)
...
docs<-Corpus(VectorSource(content))
docs<-tm_map(docs,
content_transformer(tolower))
...
wordcloud(DocumentTermMatrix(docs))
```



Importación de datos - Ejemplo

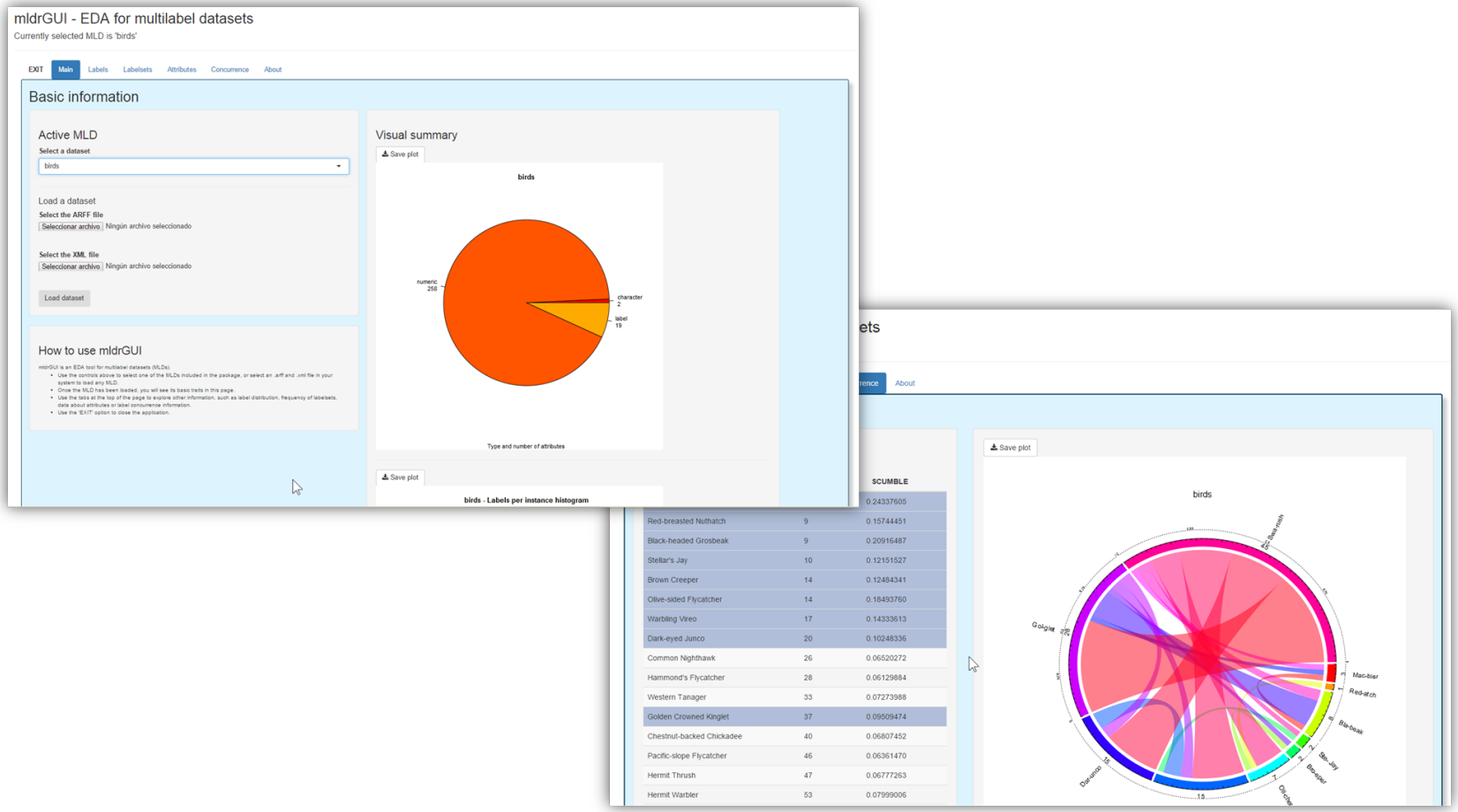
Datos obtenidos de la API de GitHub en formato JSON

`runGitHub('GitHubMining', 'fcharte')`



Exploración y análisis - Herramientas interactivas

Fáciles de utilizar - Útiles para explorar



Exploración y análisis - Scripts

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

Source on Save

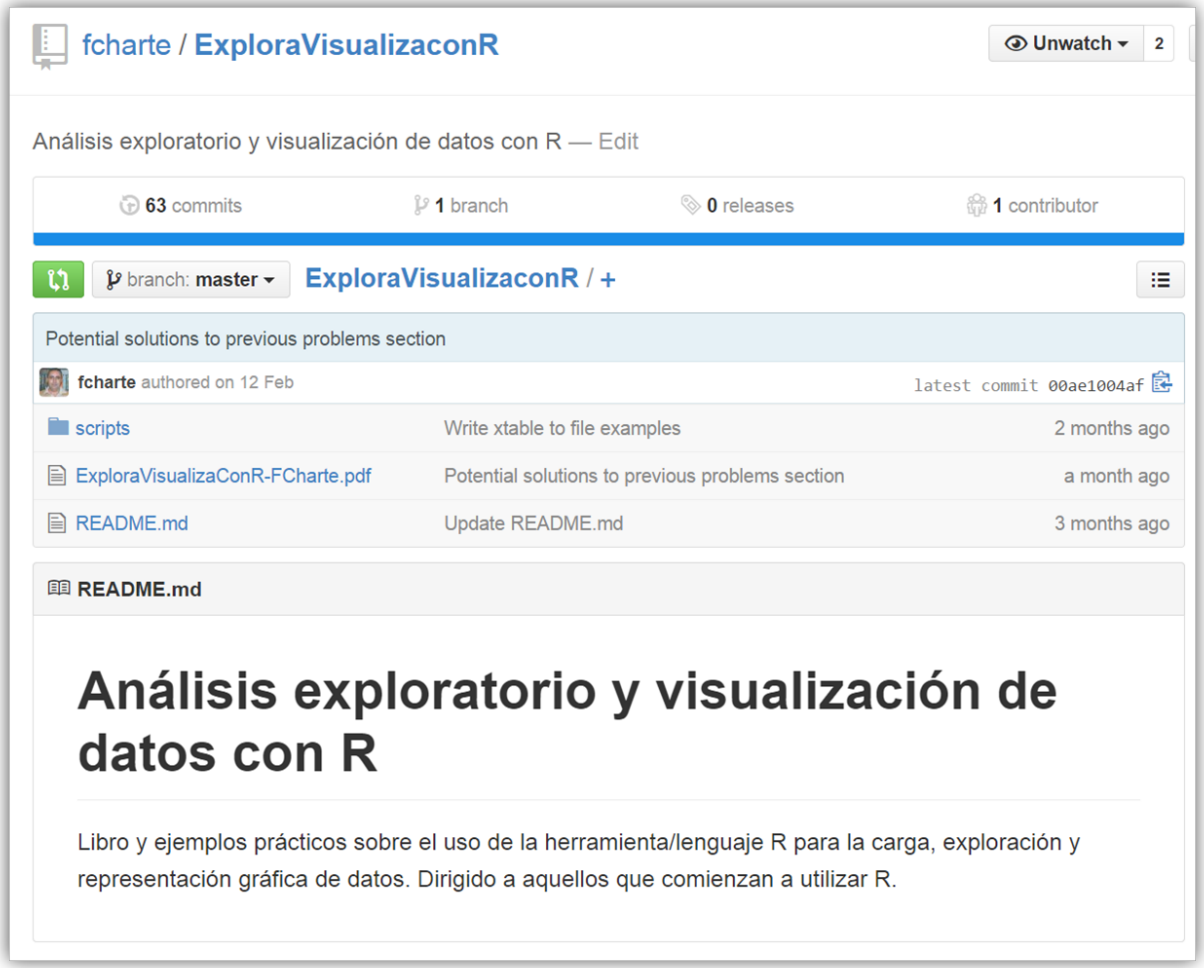
```
pdf(file = 'D:/FCharte/Estudios/Publicaciones/mldrPackage/plots.pdf', width = 10, height = 10)
layout(matrix(c(1,1,2,1,1,3,4,5,6), 3, 3, byrow = TRUE))
plot(emotions, type = "LC")
plot(birds, type = "LH")
plot(emotions, type = "LB")
plot(birds, type = "CH")
plot(emotions, type = "AT")
plot(emotions,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



Exploración y análisis - Documentos reproducibles

RBenchmarkings

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

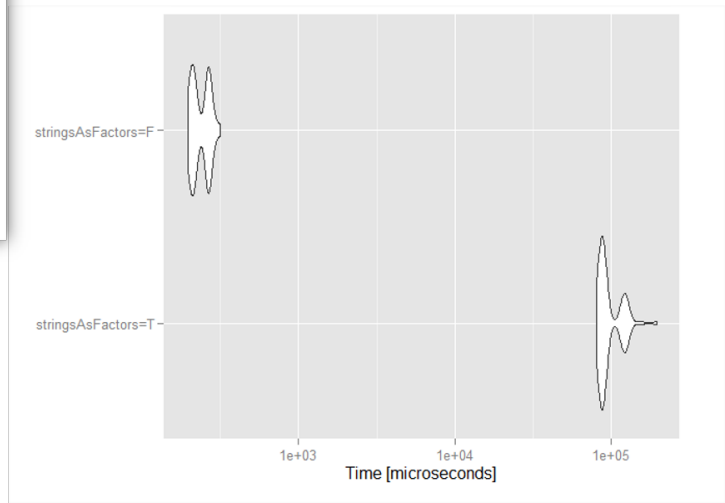
With this code I want to test the difference between using `stringsAsFactors = TRUE` versus `stringsAsFactors = FALSE` while creating a new data.frame.

```
numElements <- 1e6
someStrings <- sapply(1:25, function(x) paste(sample(c(letters, 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 378.7762
##  stringsAsFactors=F   1.000   1.0000   1.0000   1.000   1.0000   1.0000
##  neval
##    100
##    100
```



Publicación de datos - Presentaciones