A linguagem R: um ambiente para explorar dados e aprender com eles

Conferência *Hello World*

3 de Maio de 2017, 14h30

Luís Borges Gouveia

Auditório da UFP, Porto

Universidade Fernando Pessoa

A linguagem R: um ambiente para explorar dados e aprender com eles

Luis Borges Gouveia

Hello World conf, 3 de Maio

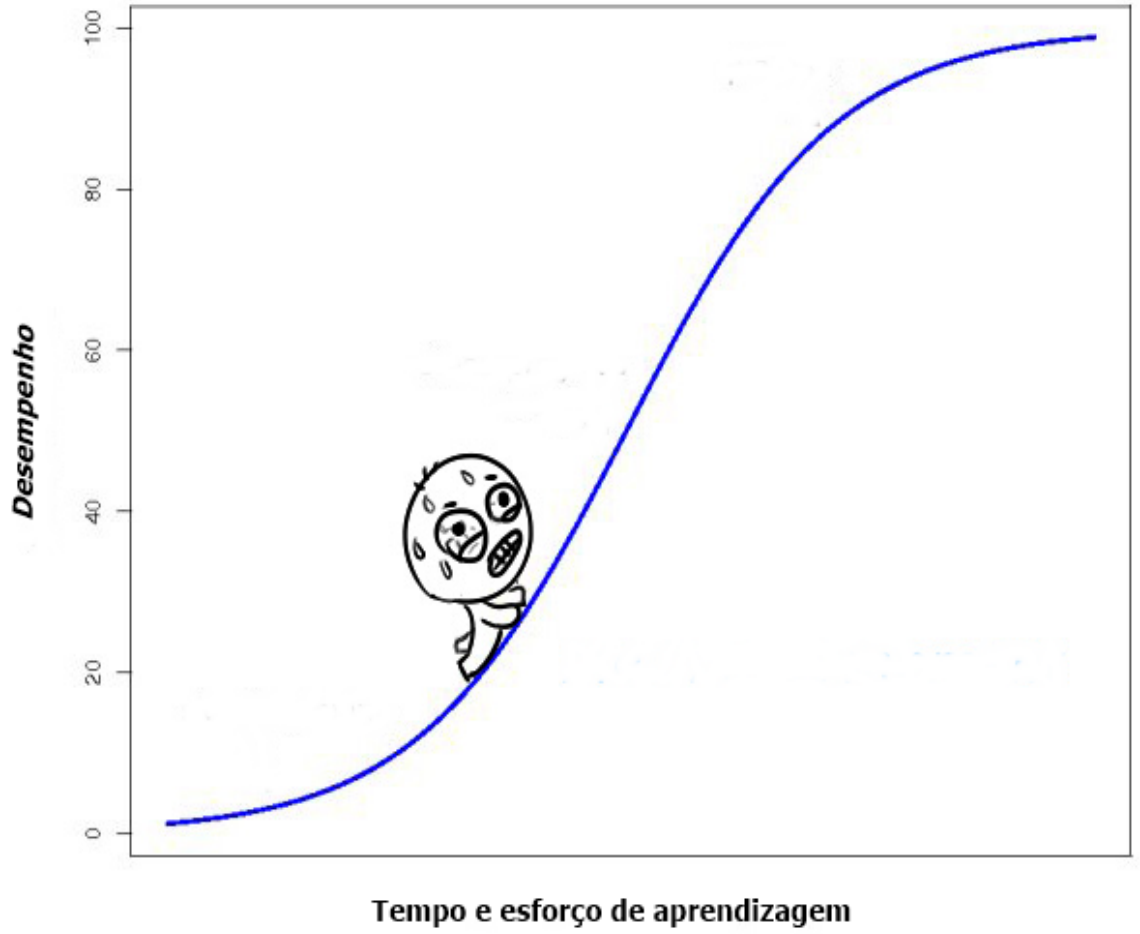
1. ***Introdução***
2. ***Explorar dados e descobrir informação***
3. ***R stuff***

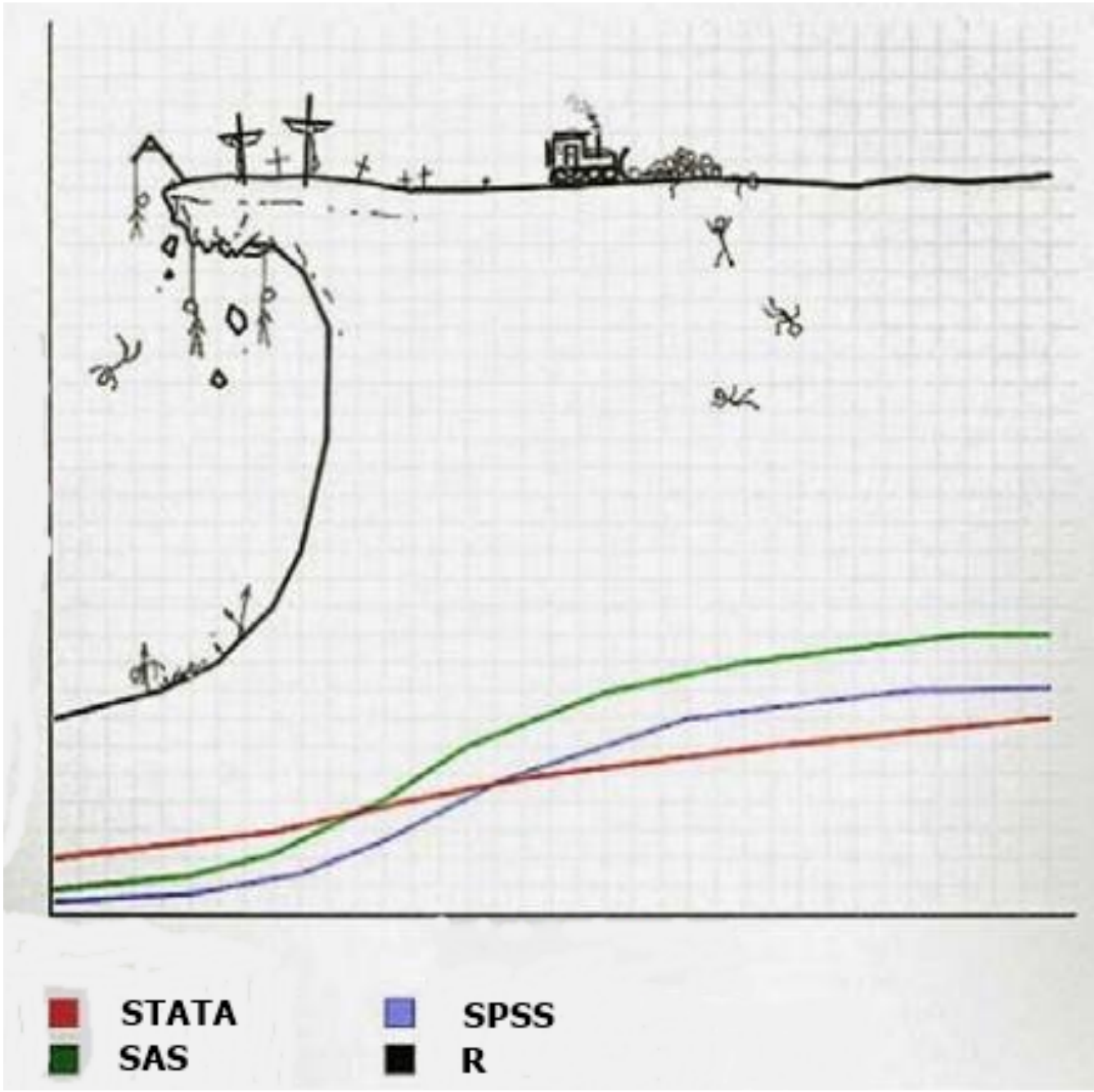
Mensagem

*A linguagem R é uma **experiência** Séc XXI: exige conhecimento de estatística e matemática, programação, criatividade, orientação para o mundo real e uma perspetiva orientada aos problemas*
Obriga a lidar com dados e a descobrir informação neles, sendo muito visual.
*Saber R é uma boa marca para os desafios resultantes dos dados que existem em quantidade, diversidade e múltipla qualidade (**competência** do Séc XXI)*

Porque estes slides?

A curva de aprendizagem do R

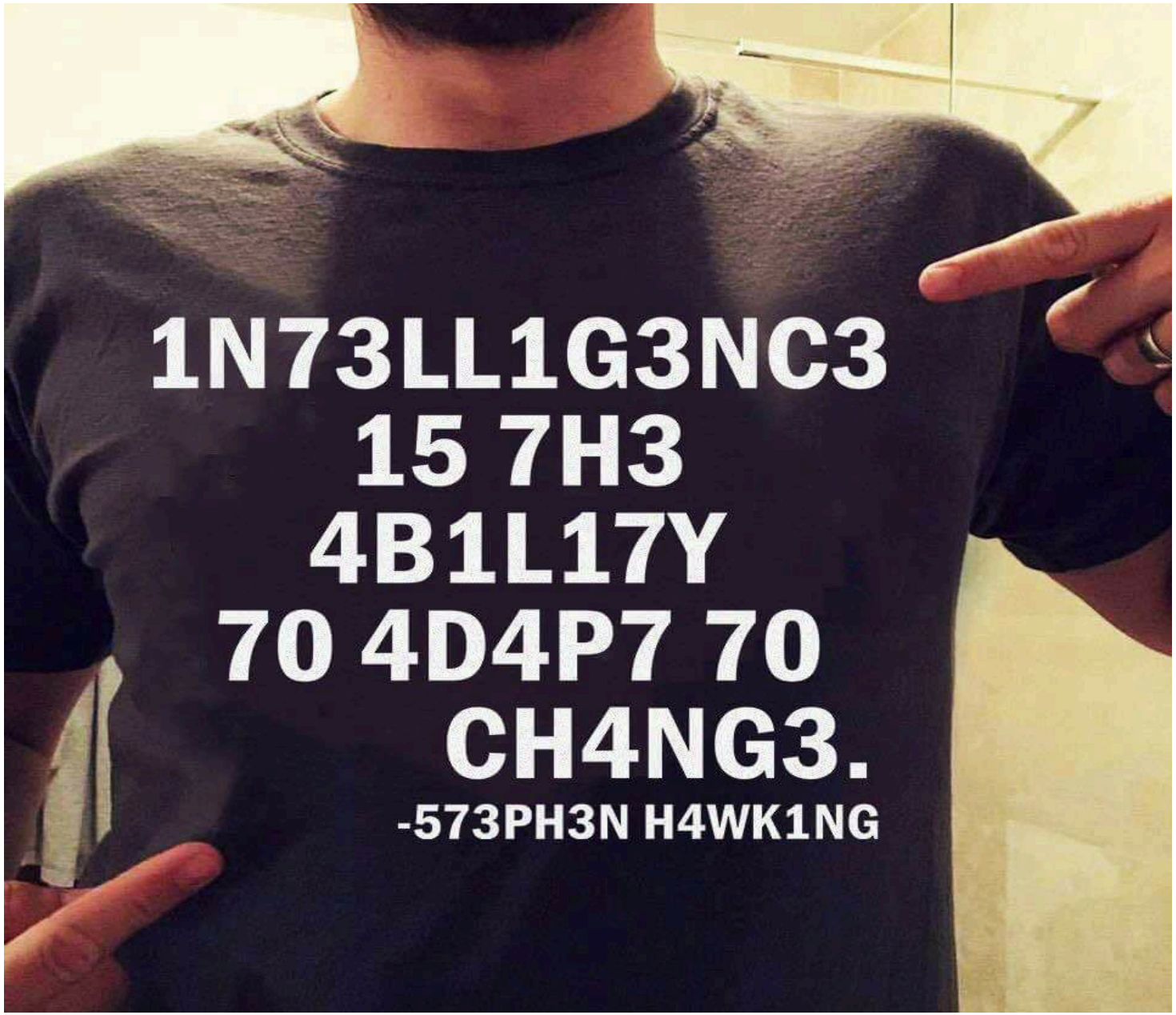




Luis Borges Gouveia, lmbg@ufp.edu.pt

Being smart !

inteligência
adaptação
flexibilidade
evolução
mudança
resposta



Luis Borges Gouveia, lmbg@ufp.edu.pt

A importância do conhecimento

dos dados ao conhecimento

Luis Borges Gouveia, lmbg@ufp.edu.pt

Os dados são o novo capital!
Muitos desafios…

- Com impacto na segurança e defesa
  - Sustentabilidade: Económica, ambiental e social
  - Emprego / Sociedade
- E nas STI? (*Sistemas e Tecnologias de Informação*)
  - Segurança e privacidade
  - Interoperabilidade
  - Transformação digital
  - Mobilidade e adaptação: BYOD / BYOA
  - Novas plataformas: da colaboração à integração
  - Automatização da atividade humana: da IA à robótica
  - …e claro, redes sociais, jogos, realidade aumentada, IoT, *cloud*…

2017

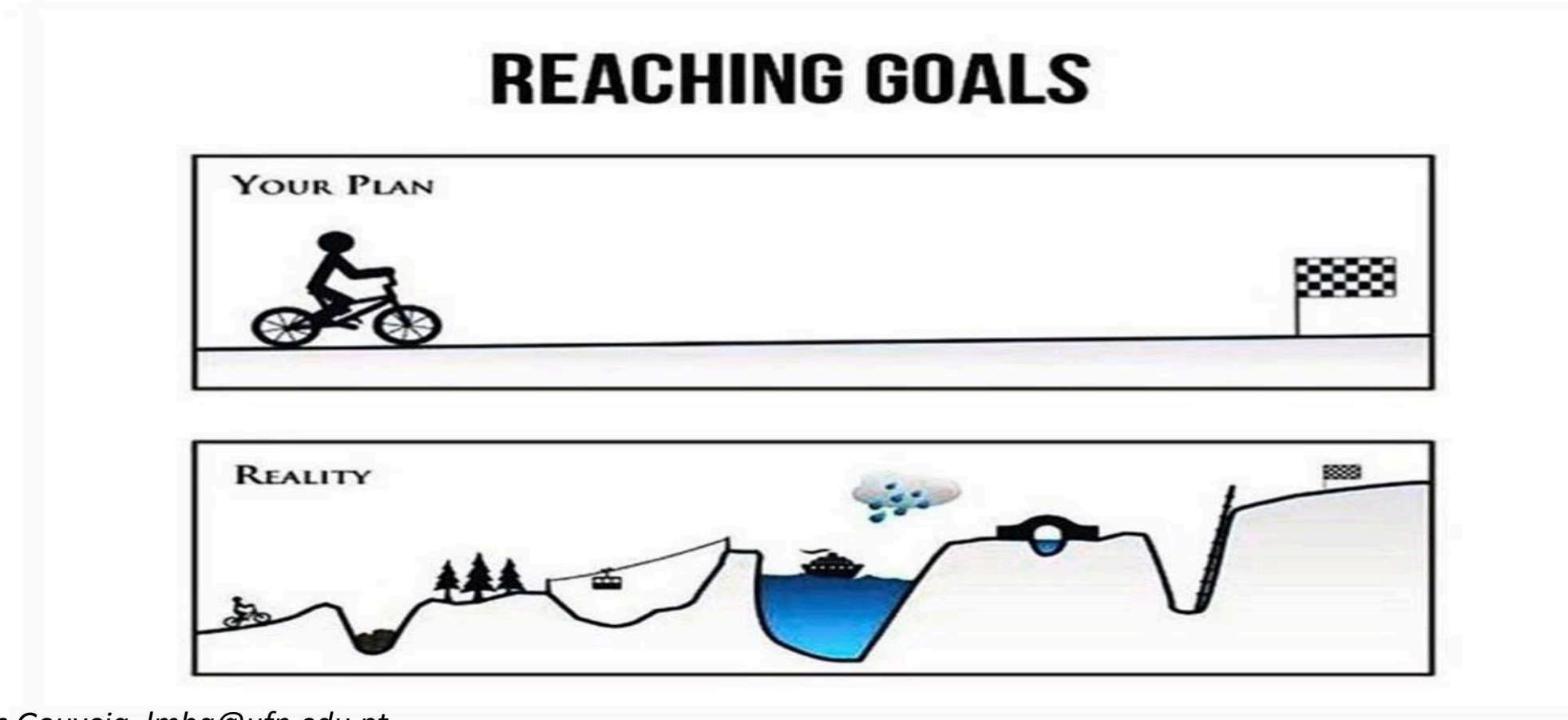
O exemplo do Facebook…

*HTTP **Status Code 201**: The request has been fulfilled and has resulted in one or more new resources being created.*

2017

2 Aprender (***ok***) e explorar (***ver imagem***) o R…

What we think and what really happens.



3, *daqui para frente…*

R *stuff*

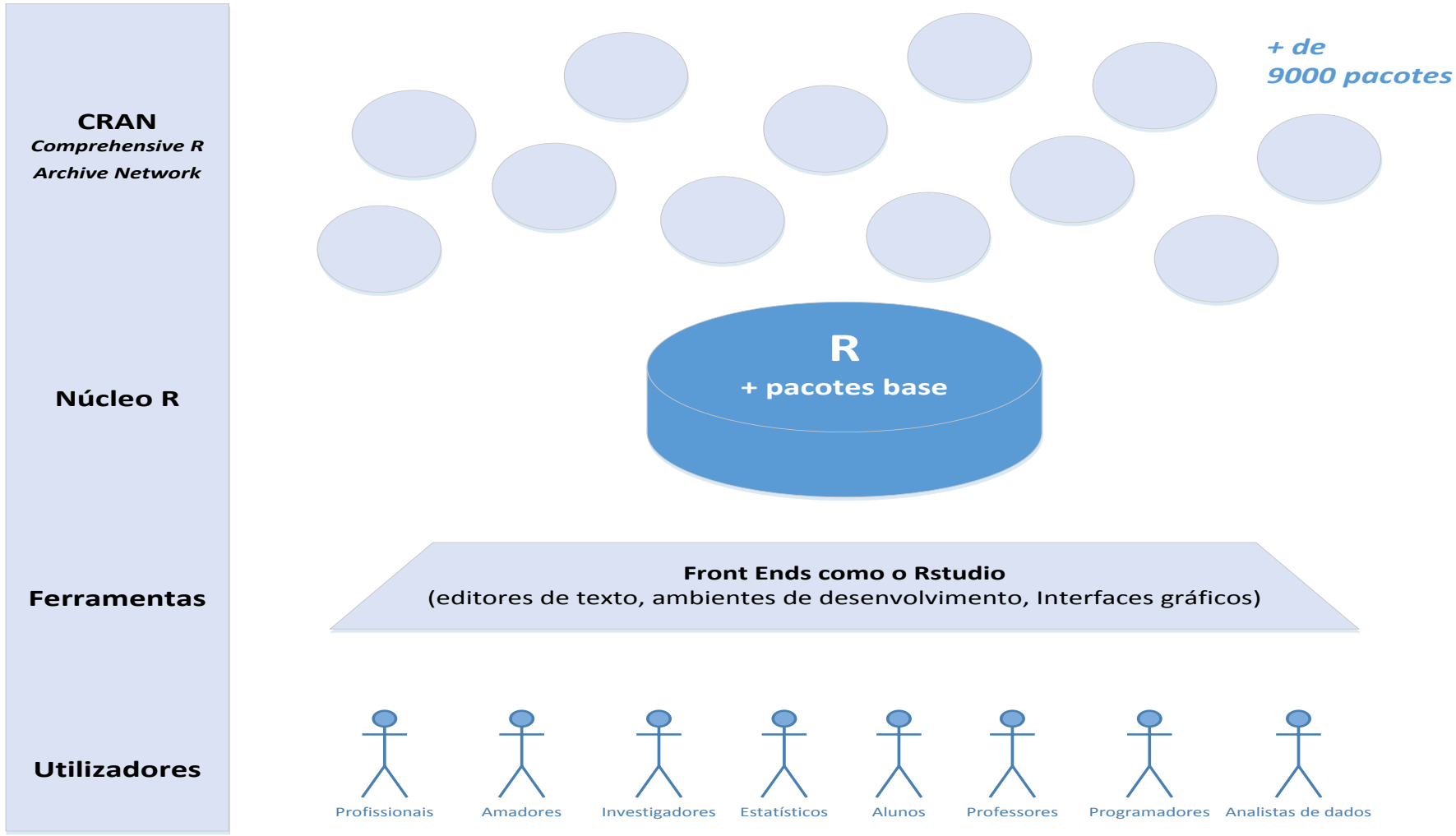
Em que consiste o R?

- O R é uma linguagem de computador interpretada, orientada aos objetos
  - O R, é desenvolvido em R, à exceção de um pequeno conjunto de primitivas internas
  - É possível integrar código em C, C++, FORTRAN ou Python, para maior eficiência ou reutilização de trabalho existente
  - Os comandos de sistema são chamados do interior do R
- O R é utilizado para a manipulação de dados, estatística e geração de gráficos e é constituido por:
  - Operadores: (+ - <- * %, entre outros) para cálculos em vetores e matrizes
  - Oferece múltiplos conjuntos, coerentes e integrados, de funções
  - Possui funcionalidades para produzir gráficos de elevada qualidade
  - Permite funções escritas pelo utilizador e conjuntos de funções (pacotes), com uma extensa lista já existente (quase 10 000)

Origem do R

- S: linguagem para a análise de dados desenvolvida nos laboratórios Bell, por volta de 1976
  - Licenciada pela *AT&T/Lucent* à *Insightful, que criou o S-plus*, em 2004
  - Em 2008, a TIBCO adquiriu a *Insightful* (*http://www.tibco.com/*)
- R: proposto como software livre, por Ross Ihaka (1954) e Robert Gentleman (1959) na Universidade de Auckland (*https://www.auckland.ac.nz*, Nova Zelândia) em Agosto de 1993 (o R está surge da letra inicial do nome dos seus autores e é também a letra anterior a S…)
  - R foi inspirado na linguagem S e no LISP
  - Desde 1997 o núcleo internacional do R (cerca de 20 pessoas mas uma comunidade de milhares de programadores) transformou o R num software de estatística e análise de dados dos mais sofisticados e completos

A arquitetura do R

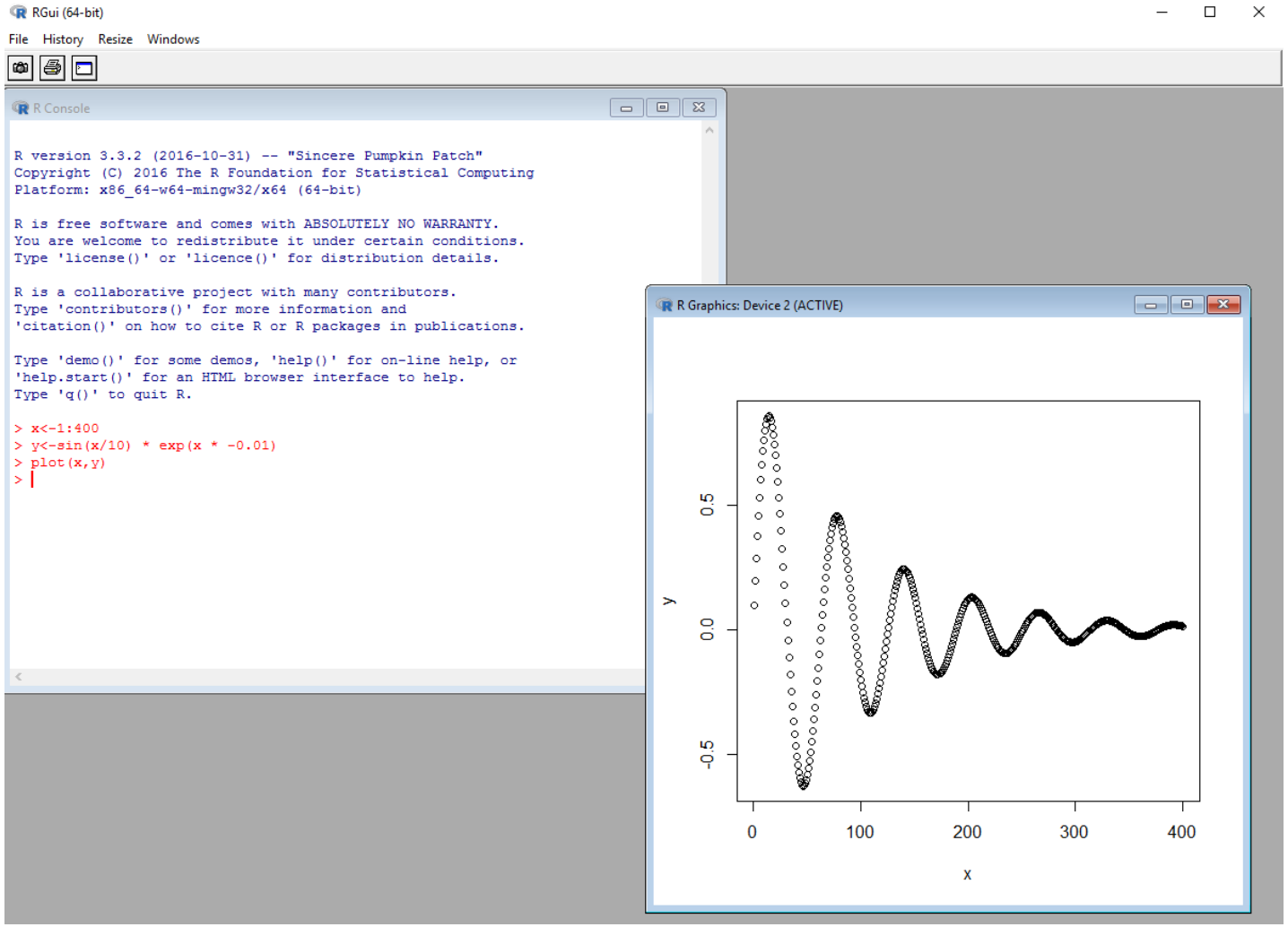


Luis Borges Gouveia, lmbg@ufp.edu.pt

O Interface do R – https://www.r-project.org/

R (v 3.4.0 de 21 de Abril de 2017 – *You Stupid Darkness*)

- A maior parte das coisas no R são objetos
  - Funções, conjuntos de dados, resultados, …
  - Os gráficos são produzidos e não são guardados como objetos
- Um *script* pode ser pensado como um modo para produzir objetos
  - O objetivo é obter os resultados dos dados (novos dados) e os gráficos pretendidos

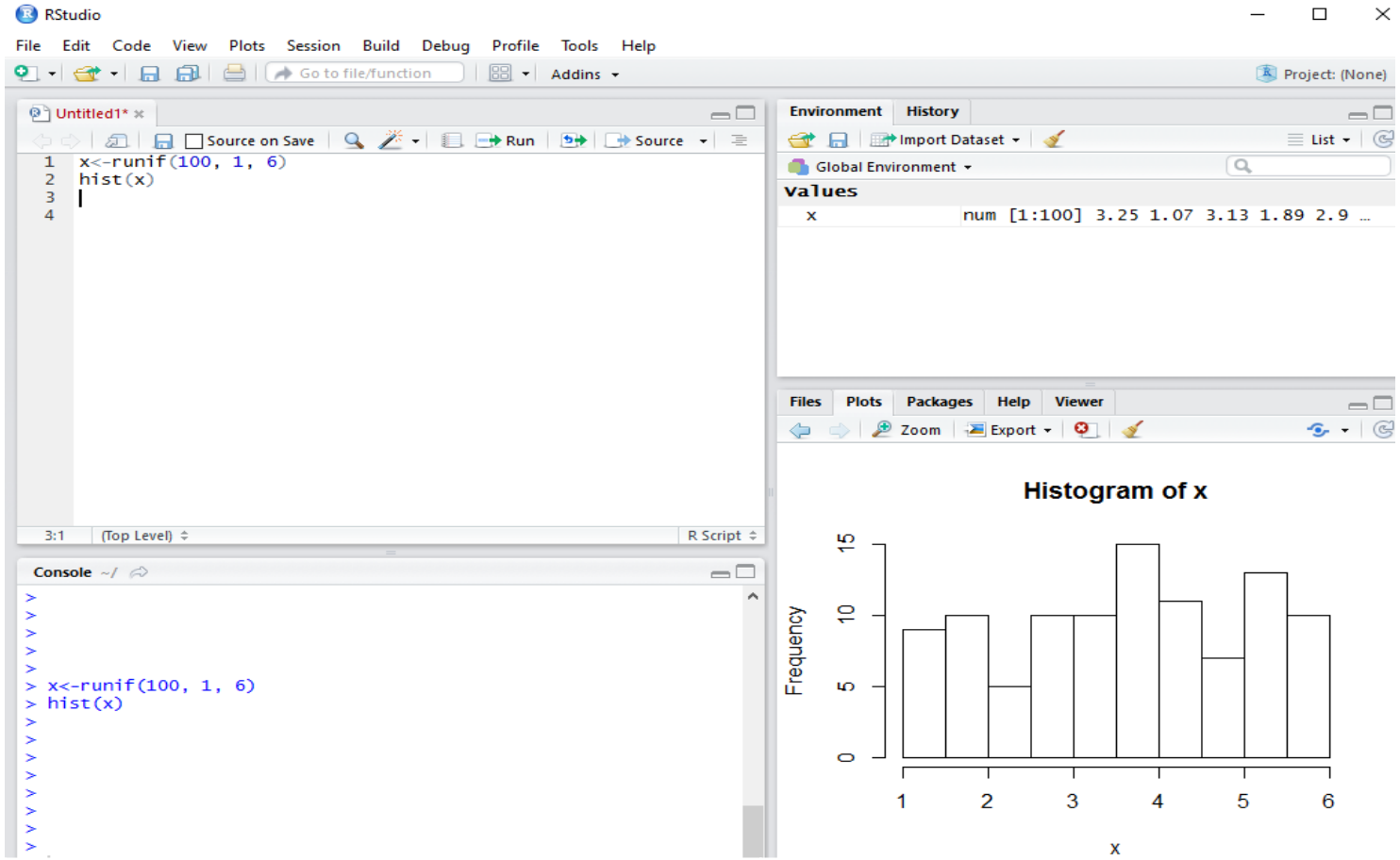


Luis Borges Gouveia, lmbg@ufp.edu.pt

Interfaces de desenvolvimento e suporte
o R-Studio como ambiente de trabalho

Muitos editores e ambientes que facilitam o desenvolvimento em R

- o mais conhecido e usado é o R-Studio
  - https://www.rstudio.com/
- Eclipse com StaET for R
  - http://www.walware.de/goto/statet
- ESS, *Emacs Speaks statistics*
  - http://ess.r-project.org/
- Togaware
  - http://rattle.togaware.com/
- Rgedit
  - http://rgedit.sourceforge.net/



Editor de código	Ambiente (objetos) e histórico
Consola do R	Ficheiros, Gráficos, Pacotes e Ajuda

Luis Borges Gouveia, lmbg@ufp.edu.pt

Operação básica em R

Numa sessão R

1. **Dados**: Ler os dados de outras fontes
2. **Ferramentas**: Utilizar pacotes, bibliotecas e funções
3. **Problema**: Escrever funções e código, quando necessário
4. **Análise**: Conduzir uma análise de dados estatística ou outra
5. **Visualização**: Produzir os gráficos necessários para evidenciar os resultados
6. **Reporte**: Gravar os resultados para ficheiros, as tabelas de dados e os gráficos gerados ou produzir relatórios
7. **Preservação**: Gravar o espaço de trabalho R, se necessário

Ler os dados: alternativas

- como tabelas (*data frames*), a partir de:
  - ficheiros de texto, txt (delimitados por *tab*), csv (com informação separada por vírgulas)
  - diretamente dos conjuntos de dados (origem)
  - via *clipboard* (*copy* e *paste*) ou introduzido por teclado
- dados em formato texto que podem ser lidos de:
  - páginas Web (*webscraping*)
  - ficheiros de texto
  - ficheiros em formato pdf (*portable document format*)
- dados de som e de imagem podem ser lidos e processados

Categorias de objetos em R

(*mode*): como os objetos que são armazenados no R
(*class*): como os objetos são tratados pelas funções

```
> M <- matrix(c(2, 3, 5, 6, 4, 2, 1, 8, 5), nrow=3, ncol=3)
> M
     [,1] [,2] [,3]
[1,]    2    6    1
[2,]    3    4    8
[3,]    5    2    5
```

- O **modo** da matriz M é determinado de forma automática pelos tipos de valores guardados em M, neste caso números inteiros (caso sejam uma mistura de tipos, o modo é lista)
- A **classe** da matriz M pode ser definida por defeito (dependendo de como foi criada) ou de forma explícita pelo utilizador. Podemos verificar a classe de um objeto e modificá-la. A classe determina como as funções vão lidar com M

Pequenos exercícios em R…

1. Gerar uma distribuição normal de 100 valores, com uma média de 62 e um desvio padrão de 25
   - x <- rnorm(100, mean=62, sd=25)
2. Gerar dados que simulam 20 lançamentos de uma moeda equilibrada
   - x <- sample(1:2,20,TRUE,prob=c(1/2,1/2))
3. Gerar dados que simulam 100 lançamentos de um dado equilibrado e com valores de 1 a 6, nas faces
   - x <- sample(1:6,100,TRUE, prob=c(1/6,1/6,1/6, 1/6, 1/6, 1/6))

Mais sete exercícios…

1. Qual é o maior valor? $\log \sqrt{\pi}$ ou $\sqrt{\log \pi}$
2. O que é que a função *rep* faz?
3. Crie um vetor contendo 50 vezes o valor 1 e chame ao vetor *grupo2*
4. Utilize um ciclo *for* para calcular os primeiros 50 números de Fibonacci. Armazene estes valores no vetor *grupo2*
   (obs: Fn = F(n-1) + F(n-2), F1 = 1, F2 = 1
5. O que faz a função *table*?
6. Quantos dos 50 números de Fibonacci são divisíveis por 3?
   (obs: a %% b)
7. Qual é a média dos primeiros 15 números de Fibonacci?

Luis Borges Gouveia, lmbg@ufp.edu.pt

Exercícios… resolução do 1

1. Qual é o maior valor? $\log \sqrt{\pi}$ **ou** $\sqrt{\log \pi}$

```
> log(sqrt(pi))
[1] 0.5723649
> sqrt(log(pi))
[1] 1.069921
```

Resposta: O valor associado à segunda expressão é maior

Exercícios… resolução do 2 e 3

2. O que é que a função *rep* faz?

?rep

Resposta: repete os elementos num vetor ou lista

3. Crie um vetor contendo 50 vezes o valor 1 e chame ao vetor *grupo2*

```
grupo2 <- rep(1, 50)
grupo2
[1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[26] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

Luis Borges Gouveia, lmbg@ufp.edu.pt

Exercícios… resolução do 4

4. **Utilize um ciclo *for* para calcular os primeiros 50 números de Fibonacci. Armazene estes valores no vetor grupo2**
(obs: Fn = F(n-1) + F(n-2), F1 = 1, F2 = 1

Resposta:

```
grupo2 <- rep(1,50)
for(i in 3:50) {
  grupo2[i] <- grupo2[i-1] + grupo2[i-2]
}
```

```
> grupo2
 [1] 1 1 2 3
 [5] 5 8 13 21
 [9] 34 55 89 144
[13] 233 377 610 987
[17] 1597 2584 4181 6765
[21] 10946 17711 28657 46368
[25] 75025 121393 196418 317811
[29] 514229 832040 1346269 2178309
[33] 3524578 5702887 9227465 14930352
[37] 24157817 39088169 63245986 102334155
[41] 165580141 267914296 433494437 701408733
[45] 1134903170 1836311903 2971215073 4807526976
[49] 7778742049 12586269025
```

Exercícios… resolução 5, 6 e 7

5. **O que faz a função *table*?**

?table

Resposta: usa os fatores de classificação cruzada para a criação de uma tabela de contingência que conta cada combinação dos níveis de fatores

6. **Quantos dos 50 números de Fibonacci são divisíveis por 3?** (obs: a %% b)

> table(grupo2 = grupo2 %% 3 ==0)

grupo2

FALSE TRUE

38 12

Resposta: até ao quinquagésimo número de Fibonacci, existem 12 valores divisíveis por 3

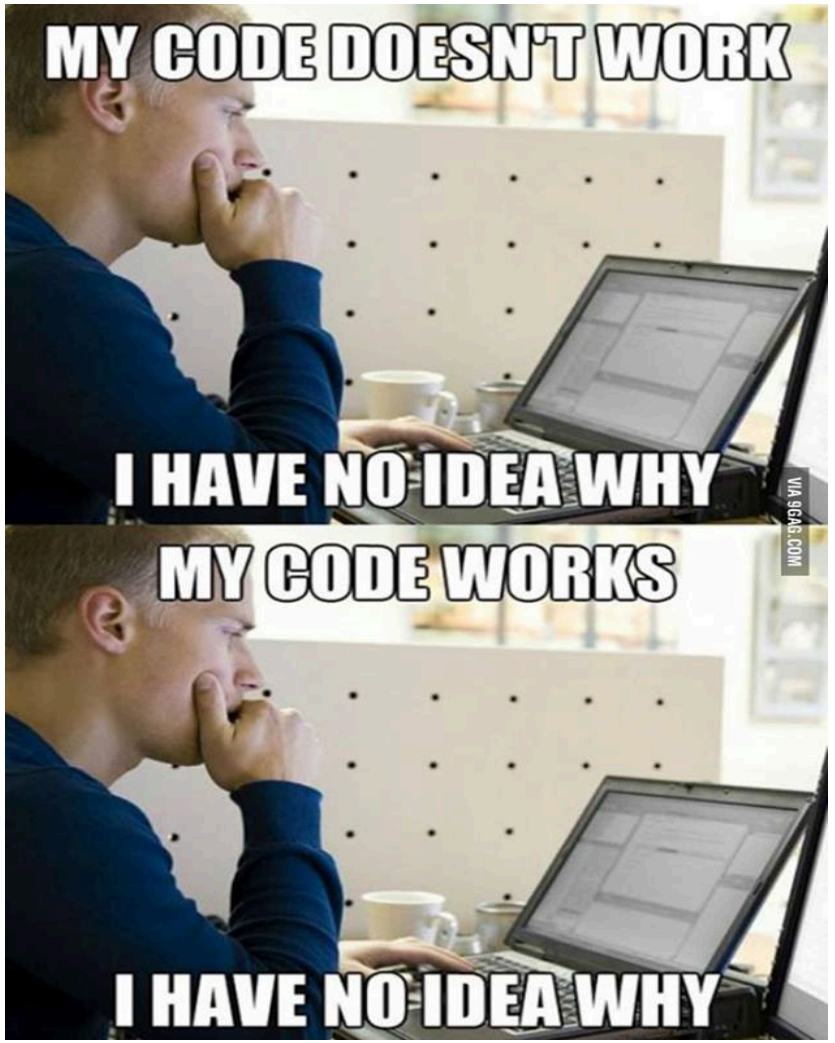
7**. Qual é a media dos primeiros 15 números de Fibonacci?**

> mean(grupo2[1:15])

[1] 106.4

Típico de uma linguagem de programação… Tempo e experiência

- usar um caderno (*logbook*) para apoio (Jupyter notebook: *https://ipython.org/notebook.html*)
- recorrer a uma carta de referência da linguagem R: *https://cran.r-project.org/doc/contrib/Short-refcard.pdf*
- consultar índice de referência do R (*R: A Language and Environment for Statistical Computing*), manual com 3518 páginas: *https://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf*
- Recorrer a apoios especializados como o catálogo de cores em R: *http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf*



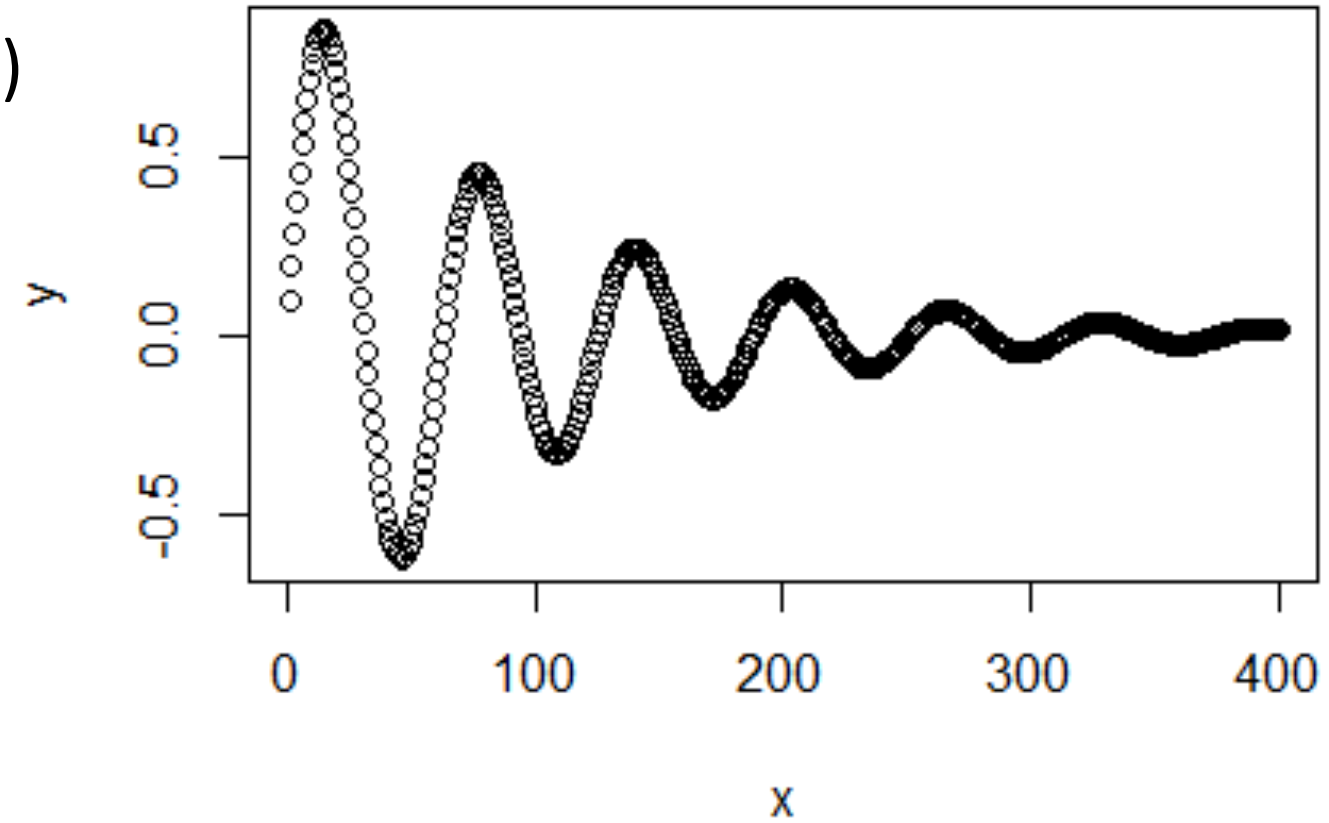
Luis Borges Gouveia, lmbg@ufp.edu.pt

Exemplos do uso de R

Uso de gráficos em R

- Um exemplo de mapeamento de uma função em 2D

```
x<-1:400
y<-sin(x/10) * exp(x * -0.01)
plot(x,y)
```

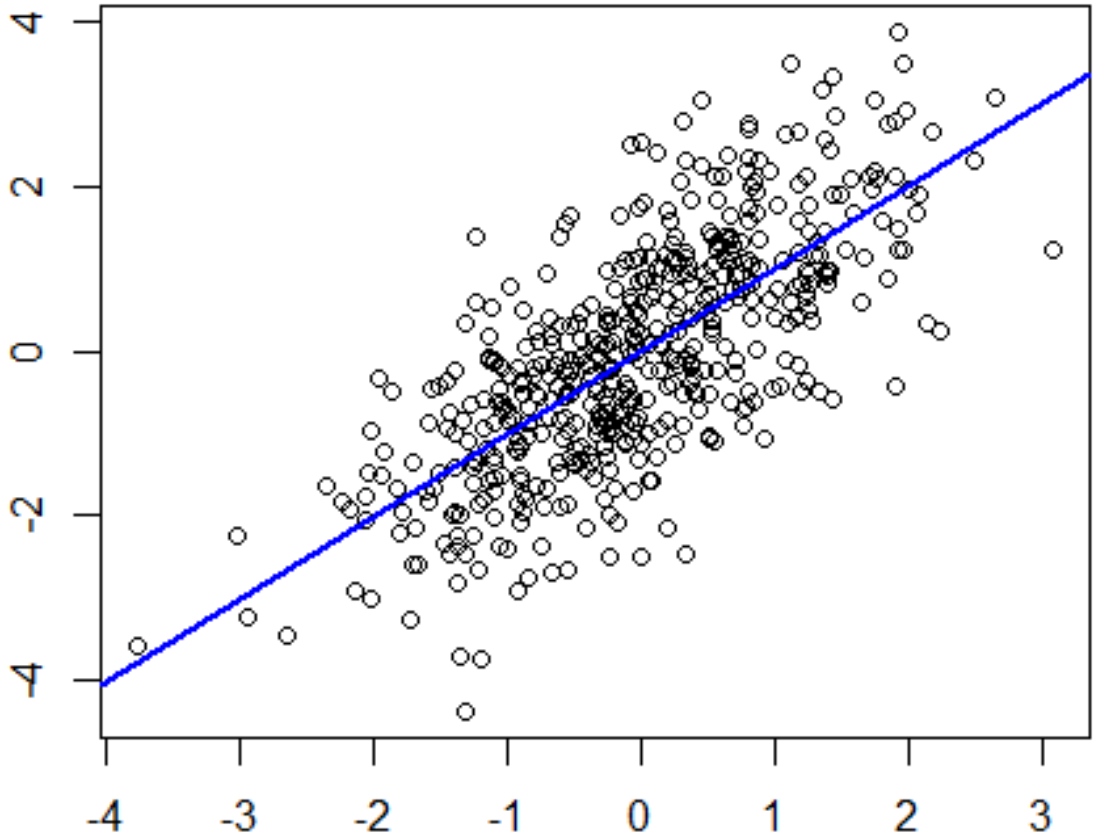


Luis Borges Gouveia, lmbg@ufp.edu.pt

Uso de gráficos em R

- Um exemplo de um gráfico de pontos com visualizacão da linha de regressão entre as variáveis X e Y

```
n <- 500
x <- rnorm(n)
y <- x + rnorm(n)
plot(x, y)
abline( lm(y ~ x), col = "blue ", lwd=2)
```



Luis Borges Gouveia, lmbg@ufp.edu.pt

Uso de gráficos em R

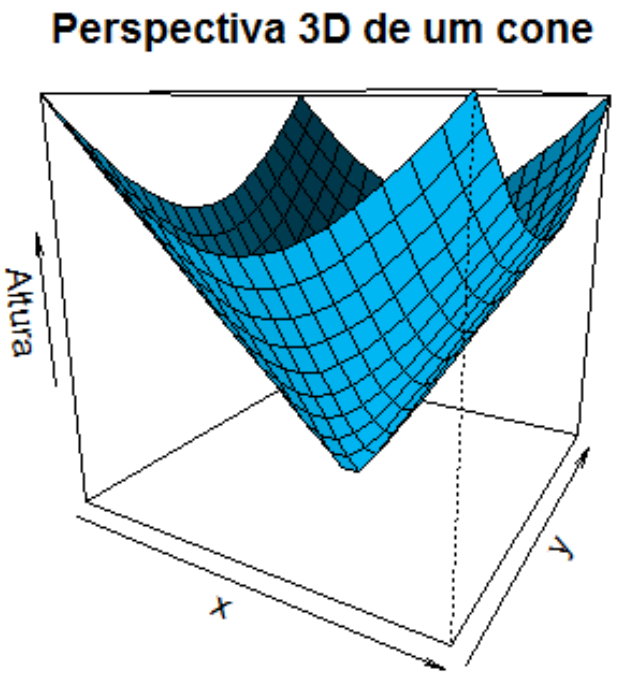
- Um exemplo de mapeamento de uma função em 3D

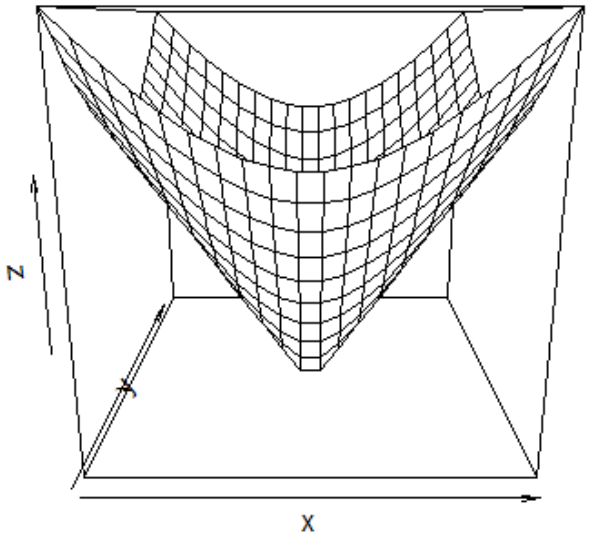
```
cone <- function(x, y){
  sqrt(x^2+y^2)
}

x <- y <- seq(-1, 1, length= 20)
z <- outer(x, y, cone)

persp(x, y, z)

persp(x, y, z,
      main="Perspectiva 3D de um cone",
      zlab = "Altura",
      theta = 30, phi = 15,
      col = "deepskyblue1", shade = 0.5)
```



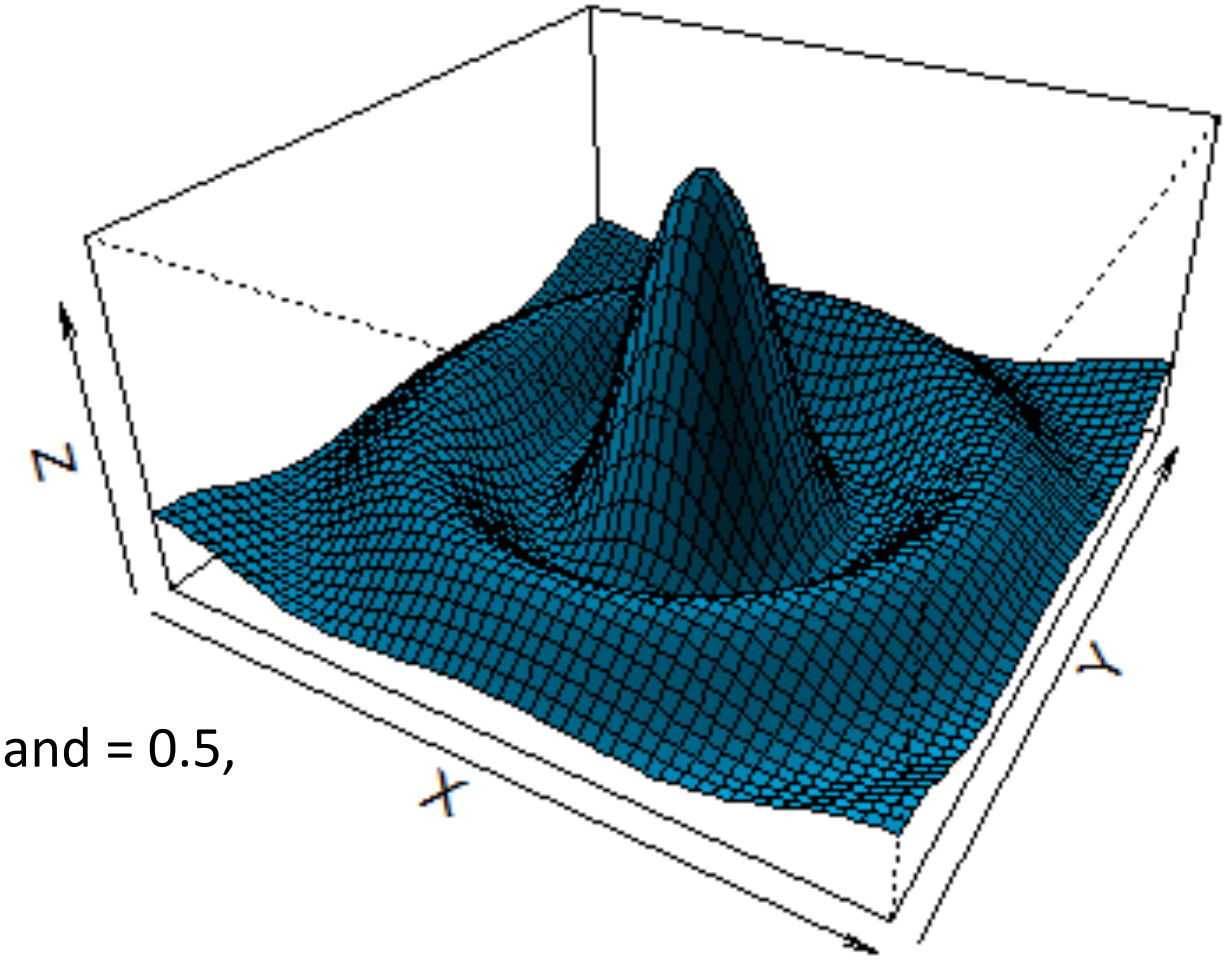


Luis Borges Gouveia, lmbg@ufp.edu.pt

Uso de gráficos em R

- Um clássico…

```
x <- seq(-10, 10, length=50)
y <- x
f <- function(x,y) {
  r <- sqrt(x^2+y^2)
  10 * sin(r)/r
}
z <- outer(x, y, f)
z[is.na(z)] <- 1
persp(x, y, z, theta = 30, phi = 30, expand = 0.5,
      col = “deepskyblue",
      shade=.5,
      xlab = "X", ylab = "Y", zlab = "Z")
```



Luis Borges Gouveia, lmbg@ufp.edu.pt

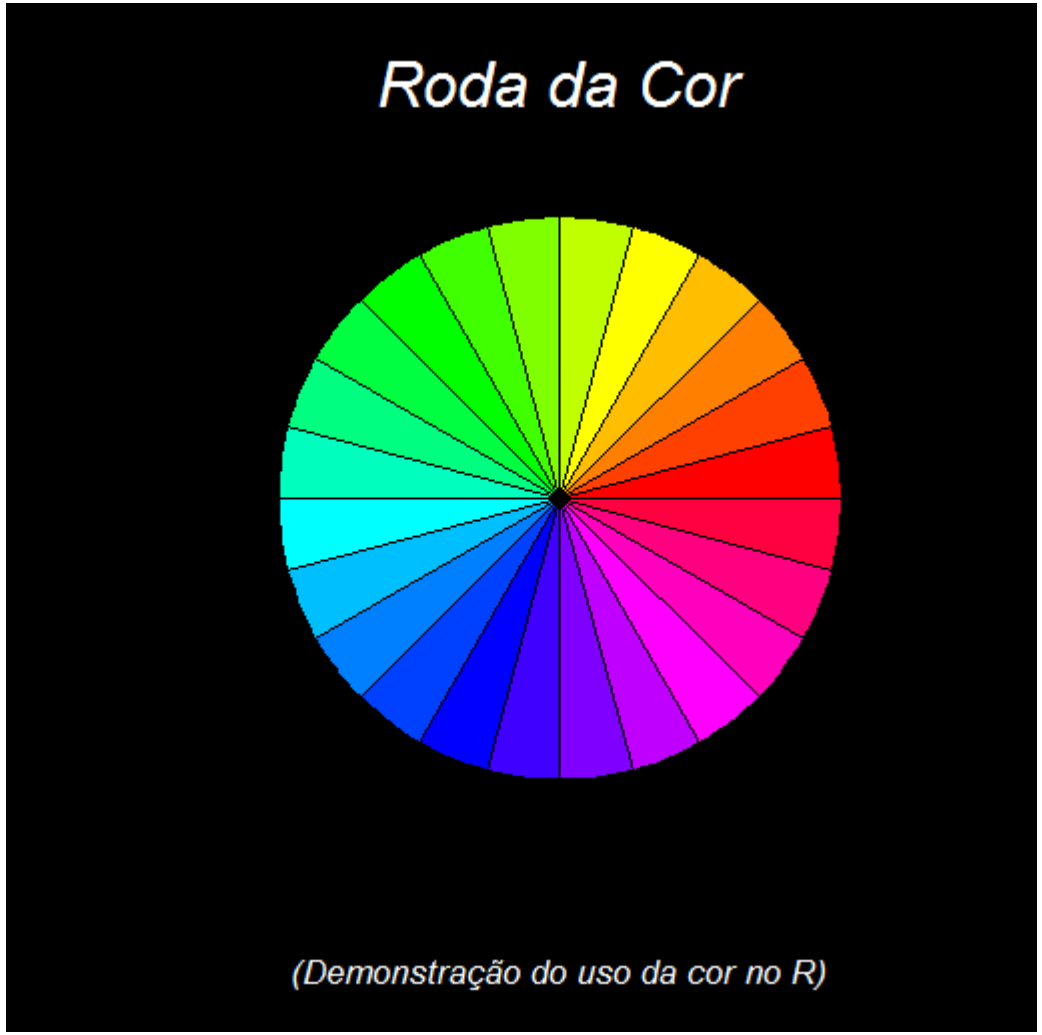
O uso de cor

```
par(bg = "black", col.main= "white ", col.lab= " white")
pie(rep(1,24), col = rainbow(24), radius = 0.9)
title(main = "Roda da Cor", cex.main = 2.0, font.main = 3)
title(xlab = "(Demonstração do uso da cor no R)",
                cex.lab = 1.1, font.lab = 3)
```

Este exemplo foi retirado de *demo(graphics)*

Os códigos de cor no R, estão disponíveis num catálogo com os nomes e cores, em *http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf*

Para ajudar a saber os códigos Hexadecimais de cor: *http://www.color-hex.com/*

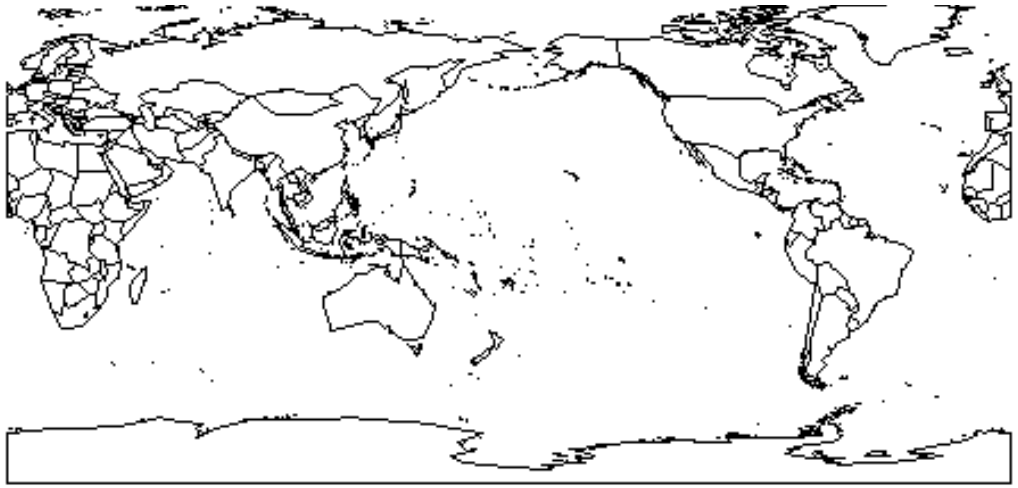


Luis Borges Gouveia, lmbg@ufp.edu.pt

Package maps

```
install.packages("maps")
library("maps")
map("world", "Portugal")

map("world")
map("mworld2")
```

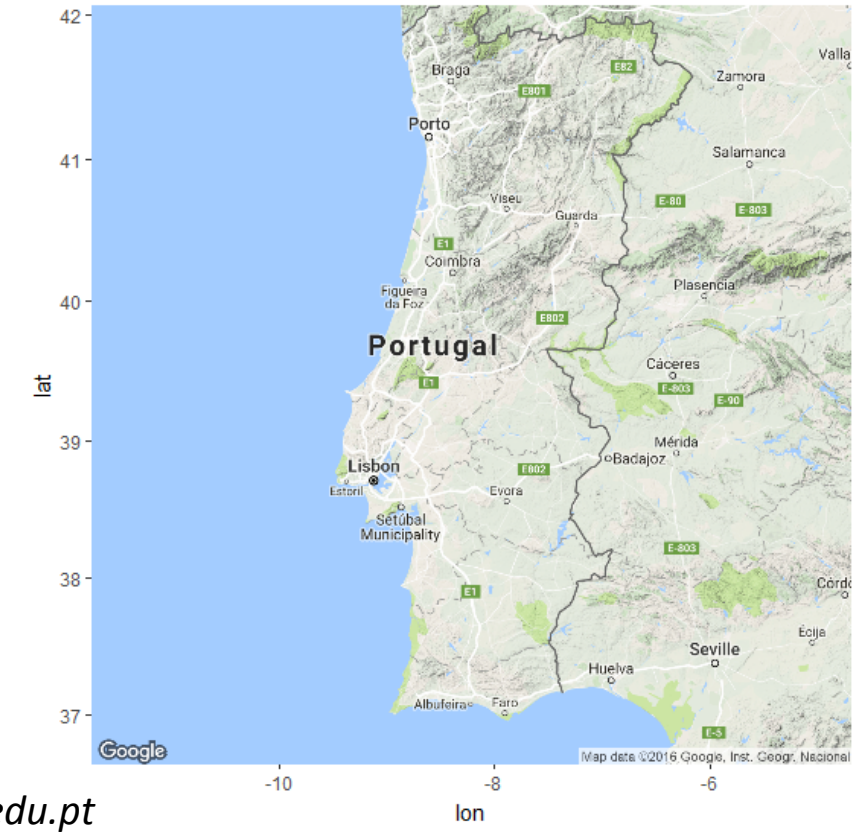


Luis Borges Gouveia, lmbg@ufp.edu.pt

Package ggmap
usa como fonte, o Google Maps

```
library(ggmap)
library(mapproj)
map <- get_map(location = ‘Europe?', zoom = 4)
map2 <- get_map(location = 'Portugal?', zoom = 7)
ggmap(map)

ggmap(map2)
```



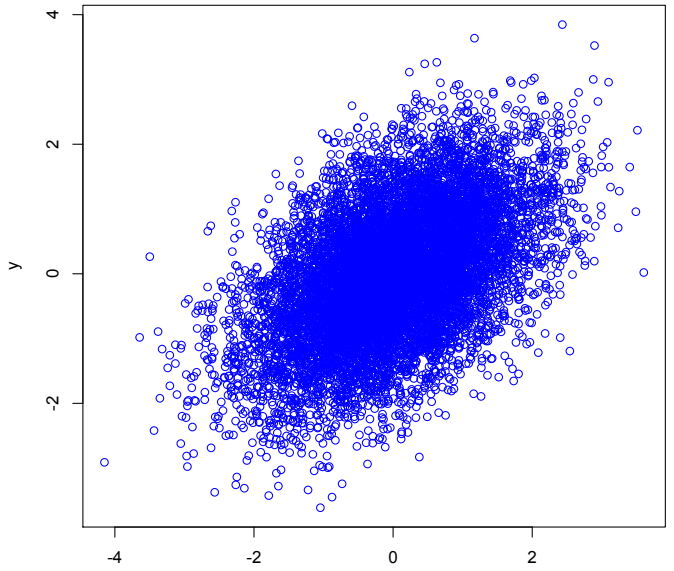
Luis Borges Gouveia, lmbg@ufp.edu.pt

Gerar dados com duas variáveis correlacionadas
(desenvolvimento de funções próprias)

```
cor(x,y)
[1] 0.4945964
```

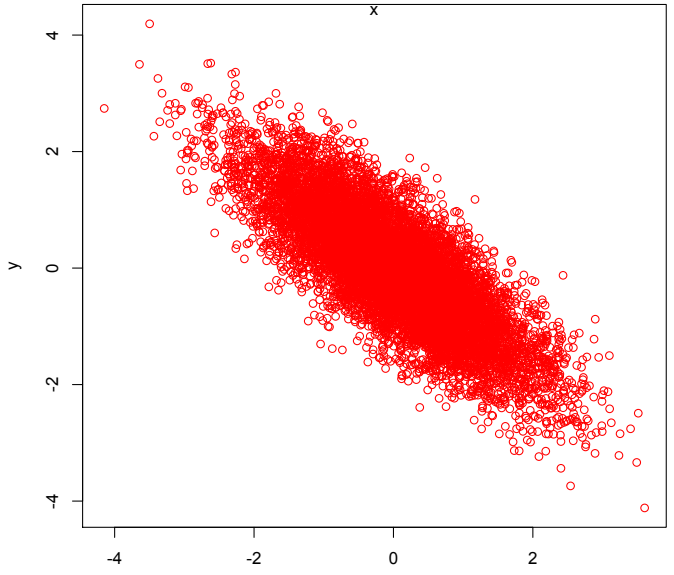
```
vCorrelacionado = function(x, r){
  r2 = r**2
  ve = 1-r2
  SD = sqrt(ve)
  e  = rnorm(length(x), mean=0, sd=SD)
  y  = r*x + e
  return(y)
}
```

```
set.seed(5)
x = rnorm(10000)
y = vCorrelacionado (x=x, r=.5)
plot(x,y, col="blue")
```




```
cor(x,y)
[1] -0.8029628
```

```
set.seed(5)
x = rnorm(10000)
y = vCorrelacionado (x=x, r=-.8)
plot(x,y, col="red")
```



Luis Borges Gouveia, lmbg@ufp.edu.pt

Recursos disponíveis sobre R

Existe sempre o (santo) Google…

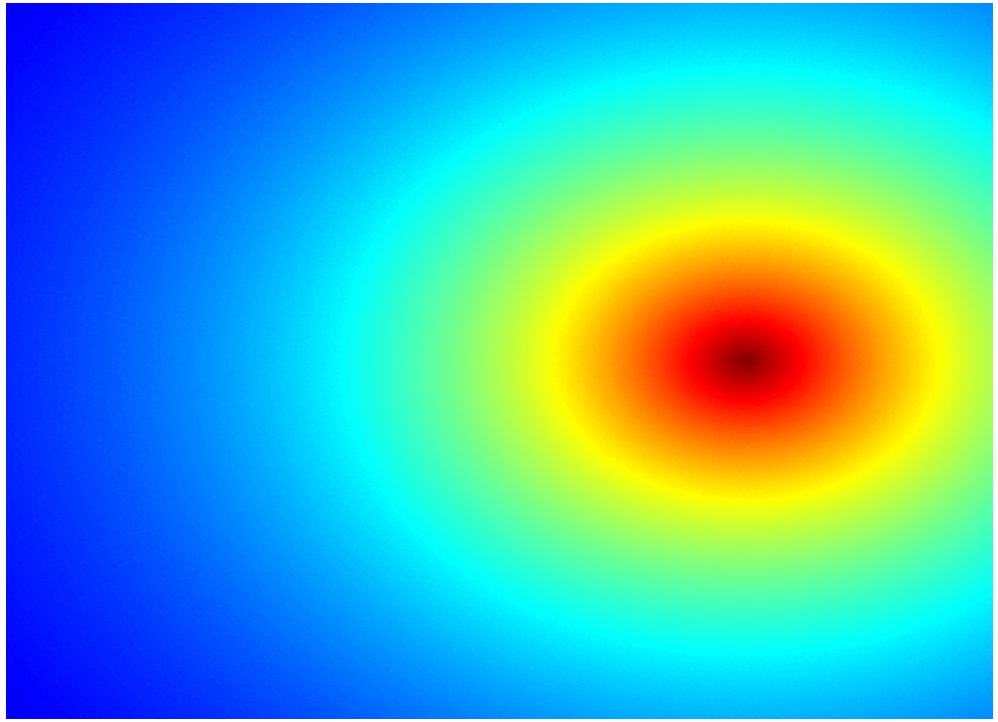
- O uso do Google pode complicar mais, do que ajudar: excesso de informação que exige um esforço considerável para o seu entendimento
- Aplica-se a lei da distância ao conhecimento…
  - ***A Web possui informação útil em função do quadrado do conhecimento de quem o usa e da sua capacidade de fazer as perguntas corretas***

Nem sempre o código funciona (Wikipedia)

https://pt.wikipedia.org/wiki/R_(linguagem_de_programa%C3%A7%C3%A3o)

```
library(caTools)
jet.colors<-colorRampPalette(c("#00007f", "#0000ff", "#007fff", "#00ffff", "#7fff7f", "#ffff00", "#ff7f00", "#ff0000",
"#7f0000"))
m<-1200
C <-complex(real=rep(seq(-1.8, 0.6, length.out=m), each=m), imag=rep(seq(-1.2, 1.2,length.out=m), m))
C<-matrix(C, m, m)
Z<-0
X<-array(0, c(m, m, 50))
for (k in 1:50) {
    Z<-Z^2+C
    X[,,k]<-exp(-abs(Z))
}
write.gif(X, "Mandelbrot.gif", col=jet.colors, delay=100)

# cria 2 matrizes com 1,44 Milhões de elementos (22 MB cada)
# cria um array com 72 Milhões de elementos (549 MB)
# cria um Gif animado com 11 MB
```



Luis Borges Gouveia, lmbg@ufp.edu.pt

7 motores de pesquisa para recursos em R

- **RSeek *http://www.rseek.org/***
- R *Documentation*: *https://www.rdocumentation.org/*
- R *Site Search*: *http://finzi.psych.upenn.edu/search.html*
- *Search the R statistical language*: *http://www.dangoldstein.com/search_r.html*
- **R-Bloggers: *https://www.r-bloggers.com/***
- Nabble R Forum *http://r.789695.n4.nabble.com/*
- R *mailing lists archive*: *http://tolstoy.newcastle.edu.au/R/*

Luis Borges Gouveia, lmbg@ufp.edu.pt

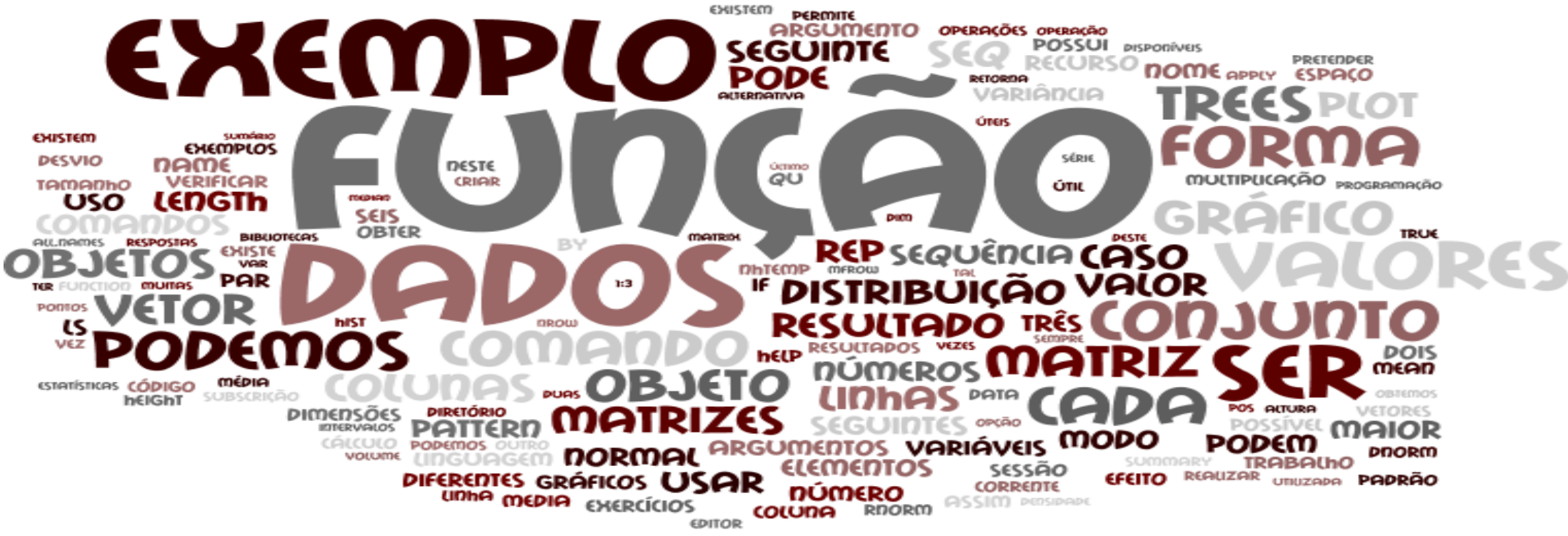
Aprender e usar o R

- Instalar o R (*http://www.r-project.org/*) e o RStudio (*https://www.rstudio.com/*)
- Experimentar, resolver problemas, partilhar e experimentar de novo…
- Explorar o CRAN (*The Comprehensive R Archive Network*): *https://cran.r-project.org/*
- Participar em grupos de discussão, aprender com os outros e voltar a experimentar... O mais completo é o R-Bloggers (*https://www.r-bloggers.com/*)
- Usar o motor de pesquisa RSeek *http://www.rseek.org/* em vez do Google
- Para saber mais dos objetos: class(x)
- Para saber mais das funções: ?x ou help(x)
- Para correr demonstrações ou exemplos de funções: demo(x) ou example(x)
- Como o R é interativo, os erros constituem uma ajuda para o código a criar
- Quanto mais experimentar o R, mais confortável é o seu uso. A experiência adquire-se essencialmente da realização de projetos concretos

Luis Borges Gouveia, lmbg@ufp.edu.pt

Uma breve introdução ao R
exploração prática e exercícios

- Pequeno manual para explorar o R, de 68 páginas, com exercícios resolvidos (Luís Borges Gouveia, 2016) - *http://hdl.handle.net/10284/5936*



Luis Borges Gouveia, lmbg@ufp.edu.pt

A linguagem R: um ambiente para explorar dados e aprender com eles

Luís Borges Gouveia

- Professor Catedrático da Faculdade de Ciência e Tecnologia da Universidade Fernando Pessoa, fervoroso adepto do FCPorto e Nortenho assumido, amigo da natureza humana e amante do digital. Gosta de computadores, mas mais ainda de os usar para melhorar a qualidade de vida das pessoas.

- Possui página Web: *http://homepage.ufp.pt/lmbg*
- Email: *lmbg@ufp.edu.pt*