# Why R?
# (CMP595 PPGC/INF/UFRGS)

Lucas Mello Schnorr, Jean-Marc Vincent

INF/UFRGS

Porto Alegre, Brazil – October 2018



# Why R?

R is a great language for data analysis and statistics

- Open-source and multi-platform
- Very expressive with high-level constructs
- Excellent graphics
- Widely used in academia and business
- Very active community
  - Documentation, FAQ on `http://stackoverflow.com/questions/tagged/r`
- Great integration with other tools

# Why is such R a pain for computer scientists?

- R is not really a programming language
- Documentation is for statisticians
- Default plots are *cumbersome* (meaningful)
- Summaries are *cryptic* (precise)
- Steep learning curve even for us, computer scientists whereas we generally switch seamlessly from a language to another! That's frustrating! ;)

# Do's and dont's

*R is high level, I'll do everything myself*

- CTAN comprises 4,334 TEX, LATEX, and related packages and tools. Most of you do not use plain TEX.
- Currently, the CRAN package repository features 4,030 available packages.
- How do you know which one to use??? Many of them are highly exotic (not to say useless to you).

  I learnt with `http://www.r-bloggers.com/`
- Lots of introductions but not necessarily what you're looking for so I'll give you a short tour.

  You should quickly realize though that you need proper training in statistics and data analysis if you do not want tell nonsense.
- Again, you should read Jain's book on The Art of Computer Systems Performance Analysis
- You may want to follow online courses:
  - `https://www.coursera.org/course/compdata`
  - `https://www.coursera.org/course/repdata`

# Install and run R on debian

```
apt-cache search r
```

Err, that's not very useful :) It's the same when searching on google but once the filter bubble is set up, it gets better...

```
sudo apt-get install r-base
```

```
R
```

```
R version 3.2.0 (2015-04-16) -- "Full of Ingredients"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>
```

# Install a few cool packages

R has its own package management mechanism so just run R and type the following commands:

- Metapackage `tidyverse` by H. Wickham (`http://had.co.nz/`)
  - Which includes many packages `dplyr`, `tidyr`, `ggplot2`, `tibble`, `readr`, `purrr`, `stringr`, and `forcats`

```
install.packages("tidyverse")
```

- `knitR` by (Yihui Xie) `http://yihui.name/knitr/`

```
install.packages("knitr")
```

# IDE

Using R interactively is nice but quickly becomes painful so at some point, you'll want an IDE.

Emacs is great but you'll need *Emacs Speaks Statistics*

```
sudo apt-get install ess
```

In this tutorial, I will briefly show you rstudio (`https://www.rstudio.com/`) and later how to use `org-mode`

# Rstudio screenshot

RStudio

File Edit Code View Project Workspace Plots Tools Help

markdown-introduction.rmd | example-r-markdown.rmd

Knit HTML | Run | Chunks

```
28
29  ```{r basicconsole}
30  x <- 1:10
31  y <- round(rnorm(10, x, 1), 2)
32  df <- data.frame(x, y)
33  df
34  ```
35
36  ## Plots
37  Images generated by `knitr` are saved in a figures folder. However,
    they also appear to be represented in the HTML output using a [data
    URI scheme]( http://en.wikipedia.org/wiki/Data_URI_scheme). This
    means that you can paste the HTML into a blog post or discussion
    forum and you don't have to worry about finding a place to store the
    images; they're embedded in the HTML.
38
39  ### Simple plot
40  Here is a basic plot using base graphics:
41
42      ```{r simpleplot}
43      plot(x)
44      ```
45
46  ```{r simpleplot}
47  plot(x)
```

47:1 Chunk 3: simpleplot R Markdown

Workspace | History

Load | Save | Import Dataset | Clear All

| Data | |
|---|---|
| df | 10 obs. of 2 variables |
| **Values** | |
| x | integer[10] |
| y | numeric[10] |

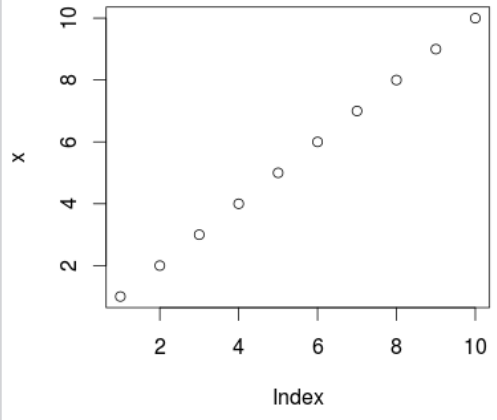
Console ~/research/statistics/rmarkdown-rmeetup-2012/

```
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> set.seed(1234)
> library(ggplot2)
> library(lattice)
> x <- 1:10
> y <- round(rnorm(10, x, 1), 2)
> df <- data.frame(x, y)
> df
    x    y
1   1 1.31
2   2 2.31
3   3 3.36
4   4 3.27
5   5 5.04
6   6 6.11
7   7 8.43
8   8 8.98
9   9 8.38
10 10 9.27
> plot(x)
>
```

Files | Plots | Packages | Help

Zoom | Export | Clear All



# Reproducible analysis in Markdown + R

- Create a new R Markdown document (Rmd) in rstudio
- R chunks are interspersed with "‘{r} and "‘
- Inline R code: ‘r sin(2+2)‘
- You can knit the document and share it via rpubs
- R chunks can be sent to the top-level with Alt-Ctrl-c
- I usually work mostly with the current environment and only knit in the end
- Other engines can be used (use rstudio completion)

  ‘‘‘{r engine='sh'}
  ls /tmp/
  ‘‘‘

- Makes reproducible analysis as simple as one click
- Great tool for quick analysis for self and colleagues, homeworks, ...