Programming and frameworks for ML

R – Additional Resources

1

Presentation

Big Data Consultant at Indra / Big Data Lecturer

- More than 20 years of experience in different environments, technologies, customers, countries ...
- Passionate data and technology
- Enthusiastic Big Data world and NoSQL

Daniel Villanueva Jiménez

Consultor BigData en Indra

NDRA • Universidad Pontificia de Salamanca

Tidyverse

The tidyverse

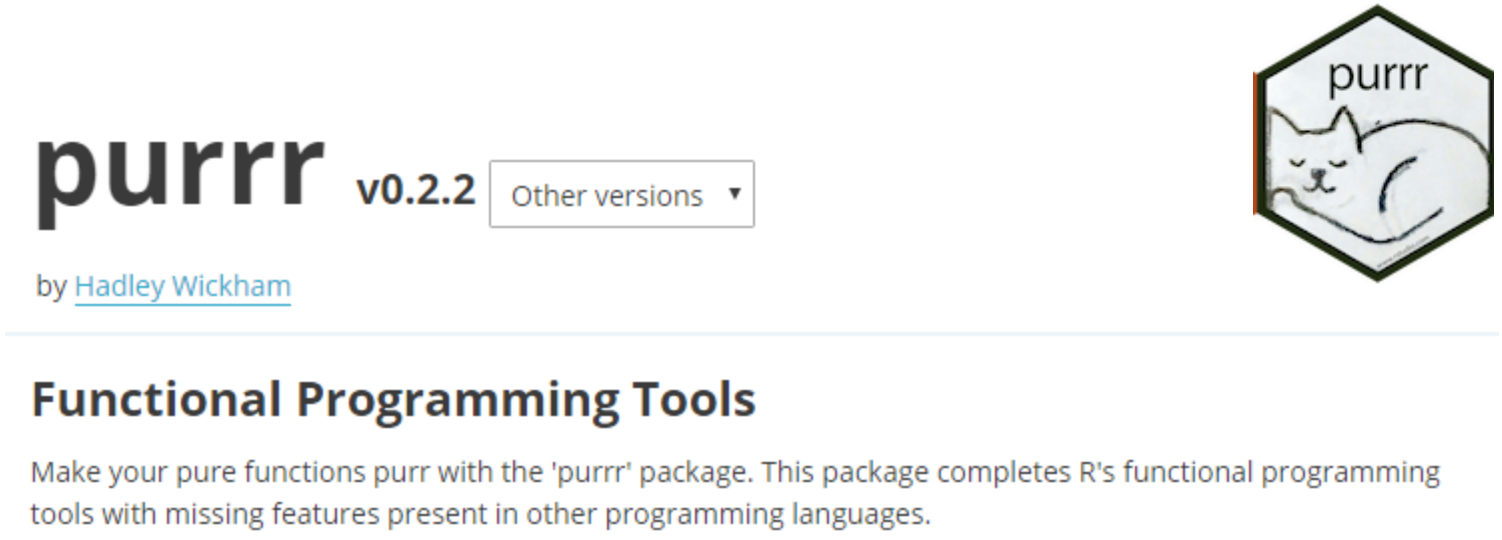
Components

The tidyverse is a collection of R packages that share common philosophies and are designed to work together. This site is a work-in-progress guide to the tidyverse and its packages.

If you are new to the tidyverse, the best place to learn the complete philsophy and how everything fits together is the R for data science book. This book is available for free online, and can you order a physical copy from Amazon (currently taking pre-orders, the book should be out by the end of the year).

Programing in R

- Procedures and functions in R
- Functional programming
- Implicit loops using map (), filter () functions, and so on.



Management of dates and times

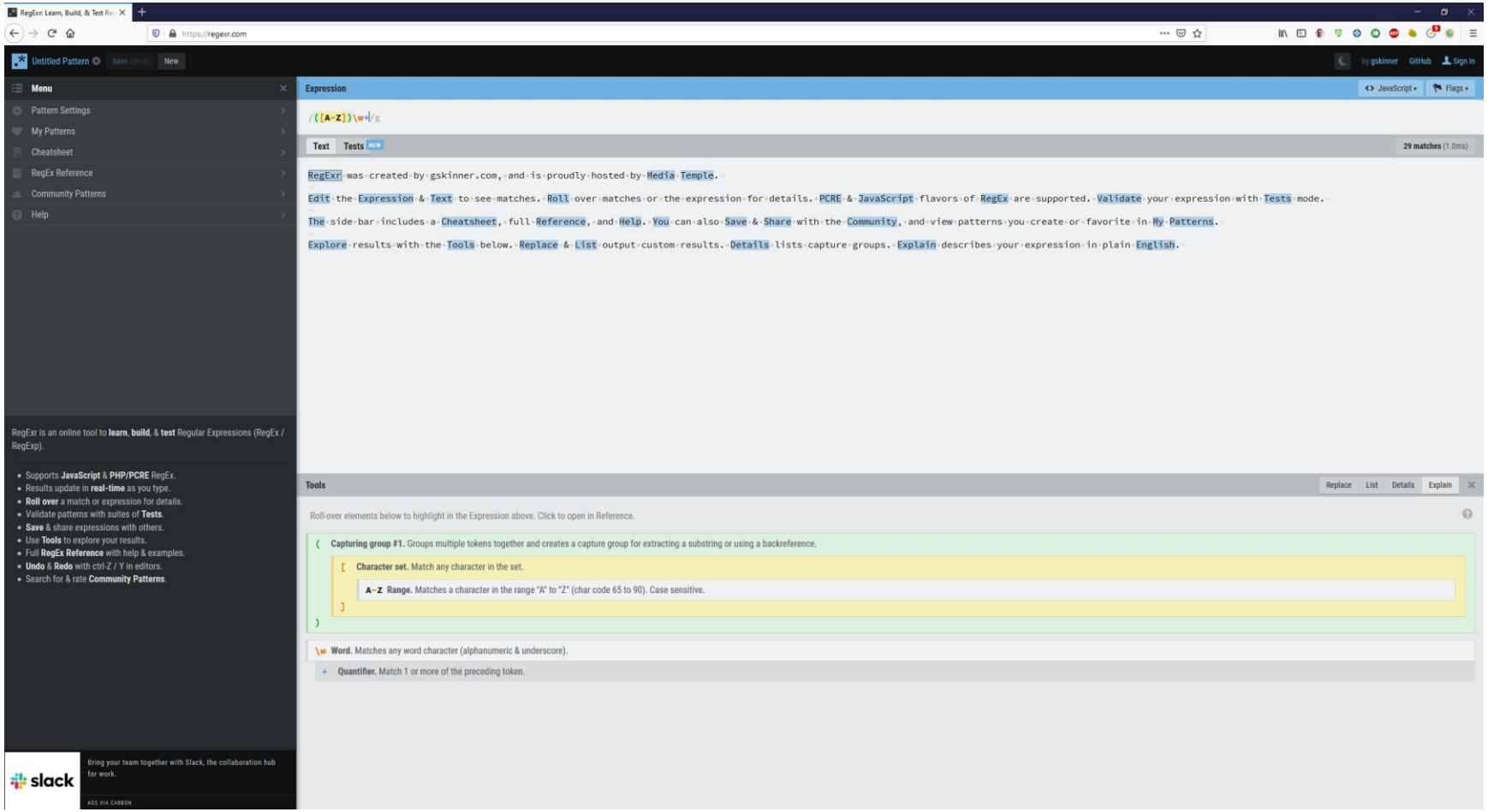
- Special types to store dates and times
- Operations with dates
- Input and output formats
- Temporal analysis in R

R-statistics blog

Statistics with R, and open source stuff (software, data, community)

Do more with dates and times in R with lubridate 1.1.0

Regular expressions



- http://regexr.com/

Text strings

- Manipulating and transforming text strings with **strintr**

Source on Save

Run

```
1  library(stringr)
2
3  str_
```

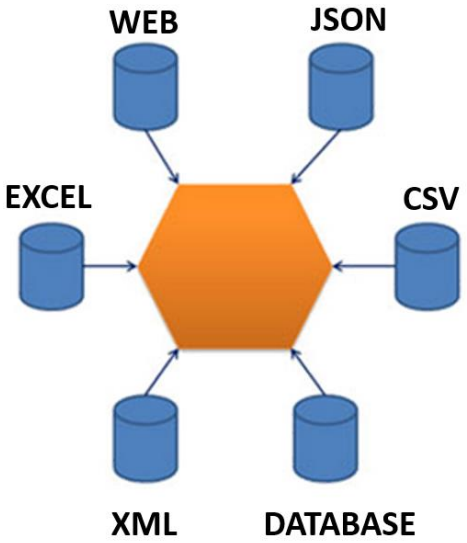
```
str_detect {stringr}
str_dup {stringr}
str_extract {stringr}
str_extract_all {stringr}
str_join {stringr}
str_length {stringr}
```

```
str_c(..., sep = "", collapse = NULL)

To understand how str_c works, you need to imagine that you
are building up a matrix of strings. Each input argument forms a
column, and is expanded to the length of the longest argument,
using the usual recyling rules. The sep string is inserted between
each column. If collapse is NULL each row is collapsed into a
single string. If non-NULL that string is inserted at the end of each

Press F1 for additional help
```

Data entry

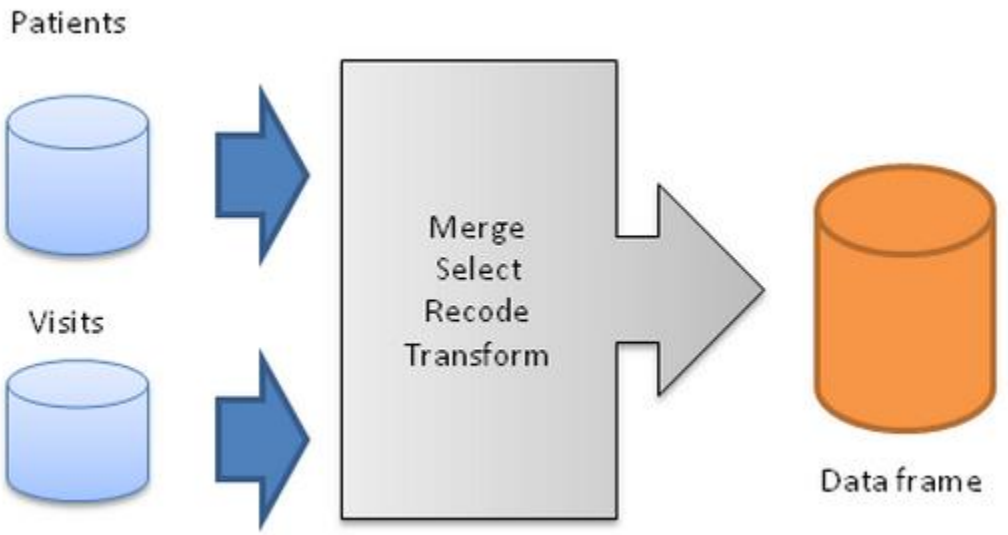

Google
R package import excel
Web
Vídeos
Imágenes
Noticias
Maps
Más
Herramientas de búsqueda
Aproximadamente 114.000.000 resultados (0,55 segundos)
R Data Import/Export - The Comprehensive R Archive Network
cran.r-project.org/doc/manuals/r.../R-data.html Traducir esta página
(See the rJava package from CRAN and the SJava, RSPerl and RSPython ... The easiest form of data to import into R is a simple text file, and this will often be ... such files directly from R. For Excel spreadsheets, the available methods are ...
Acknowledgements - 1 Introduction - 2 Spreadsheet-like data


readr
readxl
haven


WEB
JSON
EXCEL
CSV
XML
DATABASE

Data transformation

- Create new variables
- Shape data
- Link
- Order
- Mix
- Add
- Filter



Data transformation

- **dplyr** is a package that provides a series of tools to manipulate datasets
- Very easy for those who come from the SQL world
- Very intuitive once you know the basics
- Very easy to read and maintain code
- Very fast


```
movies %>%
  group_by(rating) %>%
  summarize(n()) %>%
  plot() # plots the histogram of movies by Each value of rating
```

Cleaning the data

- Detection and localization of errors
- Error correction
- Gap filling
- Duplicate rows
- Impossible Values
  - Inconsistent dates
  - Negative sales

Good data scientists understand, in a deep way, that the heavy lifting of cleanup and preparation isn't something that gets in the way of solving the problem – it *is* the problem.

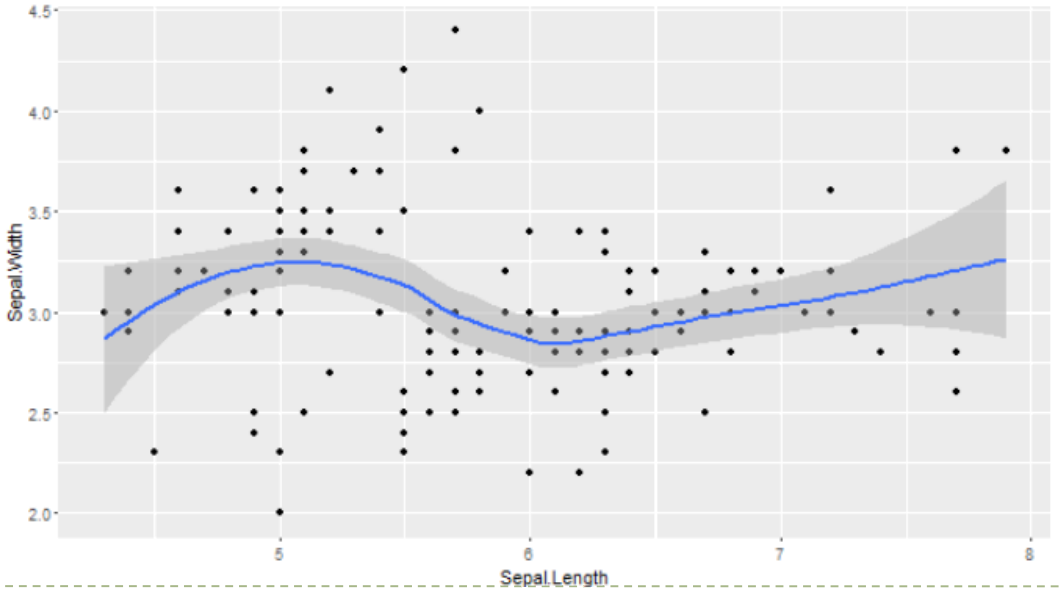
DJ Patil, *Building Data Science Teams*

Change the shape of the data

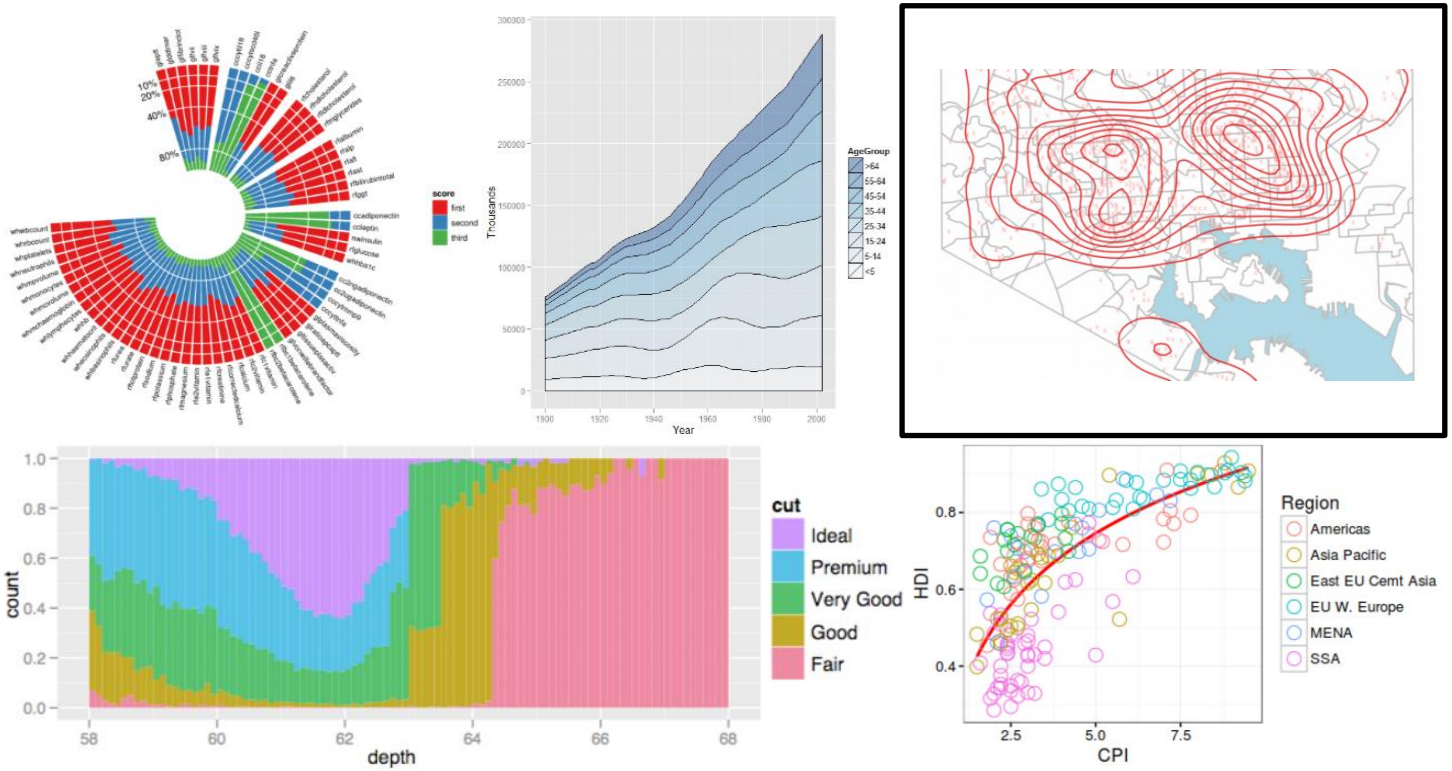
Advanced graphics in R

- **ggplot2** is a package that is based on the grammar of graphics
- Provides a language to create complex graphs in a very simple way

```
> ggplot(data=iris, aes(x=Sepal.Length, y=Sepal.Width))  + geom_point()  + stat_smooth()
```



Advanced graphics in R

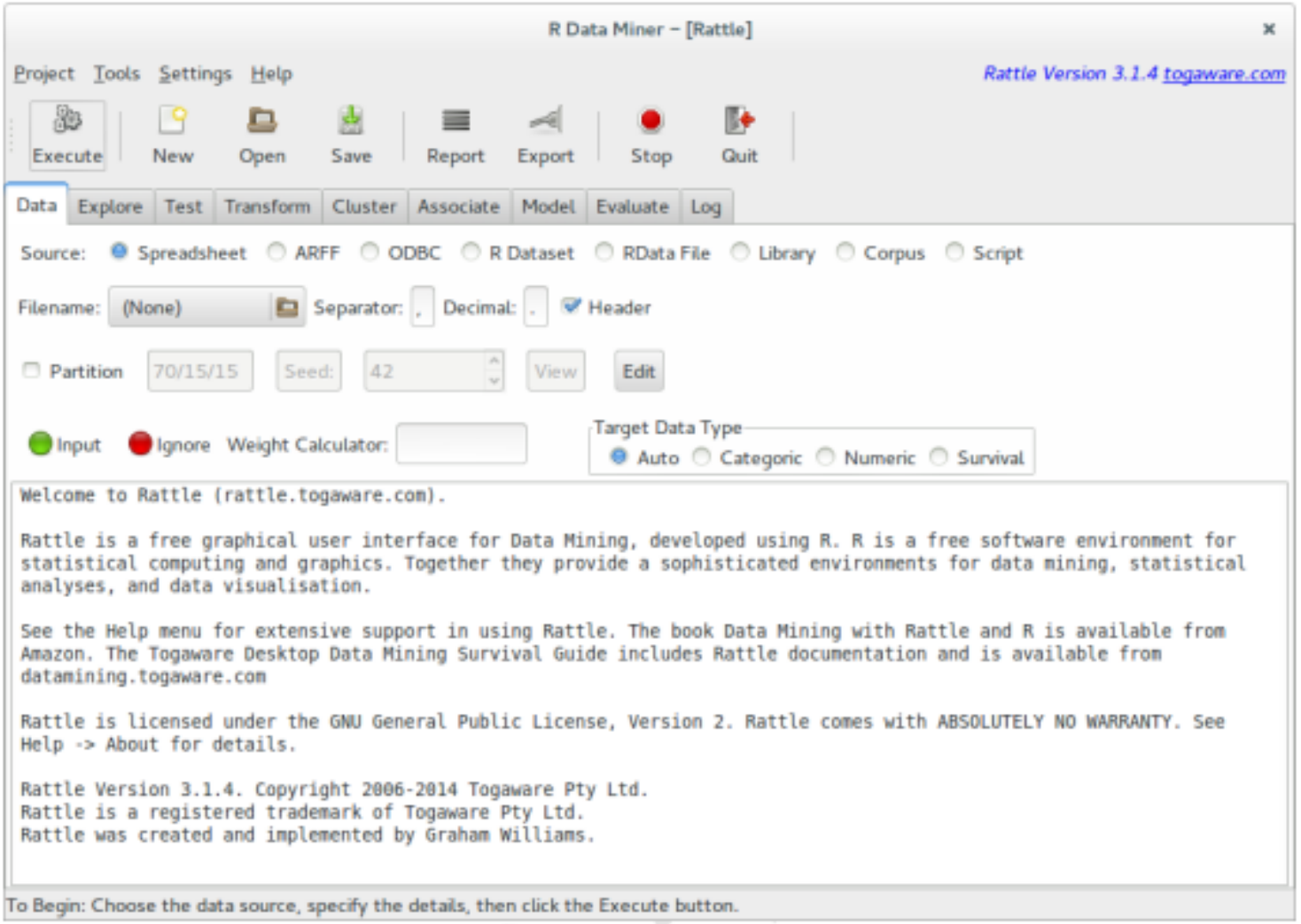


Machine learning

	Stata	SAS	SPSS	Mplus	R
Regression Models					
Robust Regression	Stata	SAS			R
Models for Binary and Categorical Outcomes					
Logistic Regression	Stata	SAS	SPSS	Mplus	R
Exact Logistic Regression	Stata	SAS			R
Multinomial Logistic Regression	Stata	SAS	SPSS	Mplus	R
Ordinal Logistic Regression	Stata	SAS	SPSS	Mplus	R
Probit Regression	Stata	SAS	SPSS	Mplus	R
Count Models					
Poisson Regression	Stata	SAS	SPSS	Mplus	R
Negative Binomial Regression	Stata	SAS	SPSS	Mplus	R
Zero-inflated Poisson Regression	Stata	SAS		Mplus	R
Zero-inflated Negative Binomial Regression	Stata	SAS		Mplus	R
Zero-truncated Poisson	Stata	SAS			R
Zero-truncated Negative Binomial	Stata	SAS		Mplus	R
Censored and Truncated Regression					
Tobit Regression	Stata	SAS		Mplus	R
Truncated Regression	Stata	SAS			R
Interval Regression	Stata	SAS			R
Multivariate Analysis					
One-way MANOVA	Stata	SAS	SPSS		
Discriminant Function Analysis	Stata	SAS	SPSS		
Canonical Correlation Analysis	Stata	SAS	SPSS		R
Multivariate Multiple Regression	Stata	SAS		Mplus	
Mixed Effects Models					
Generalized Linear Mixed Models	Introduction to GLMMs				
Mixed Effects Logistic Regression	Stata				R
Other					
Latent Class Analysis				Mplus	

	Stata	SAS	SPSS	Mplus	R	G*Power
Power Analysis / Sample Size						
Single-sample t-test	Stata	SAS			R	G*Power
Paired-sample t-test	Stata	SAS			R	G*Power
Independent-sample t-test	Stata	SAS			R	G*Power
Two Independent Proportions	Stata	SAS				G*Power
One-way ANOVA	Stata	SAS				G*Power
Multiple Regression	Stata	SAS				G*Power
Accuracy in Parameter Estimation	Stata					

Rattle - R's analytical tool



Publication of reports in R - Knitr

Which checks out with our value. This suggests that X_1 and X_2 are pretty colinear. Let's visualize this by using a scatter plot:

```
plot(x1, x2, main = "Correlation of X1 and X2", xlab = "X1", ylab = "X2")
```

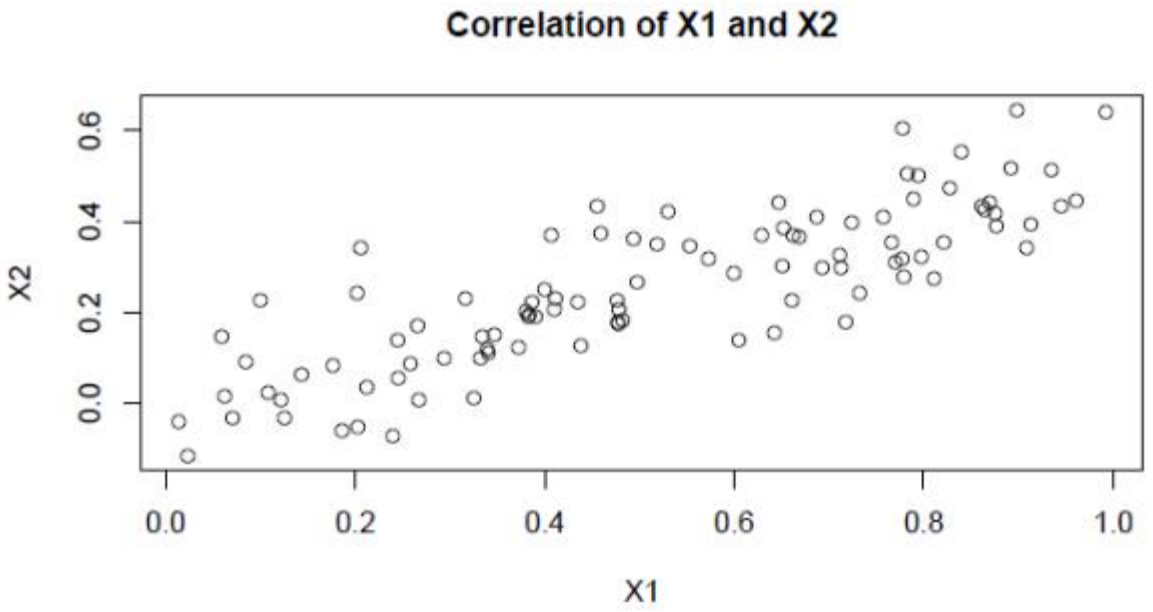
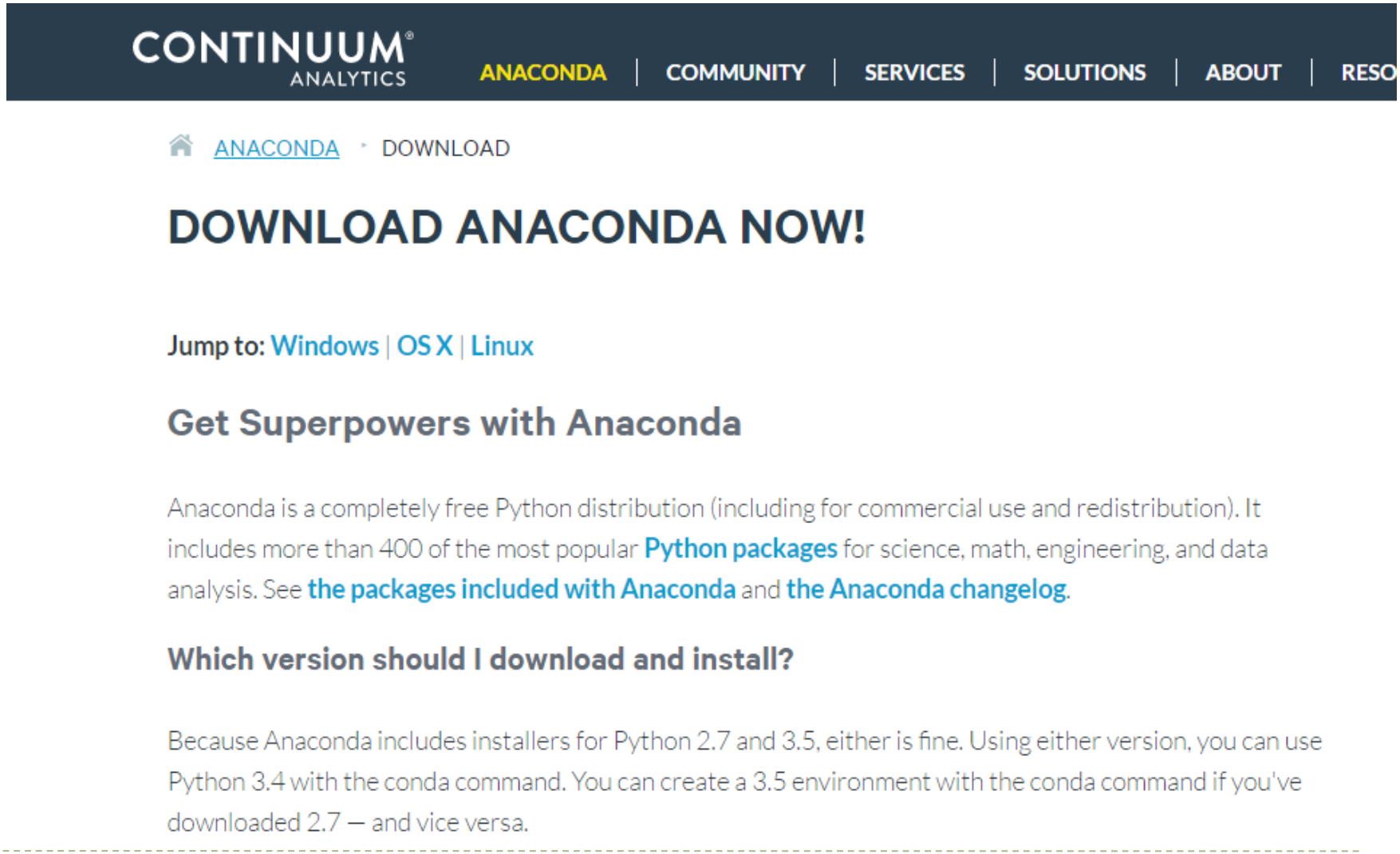


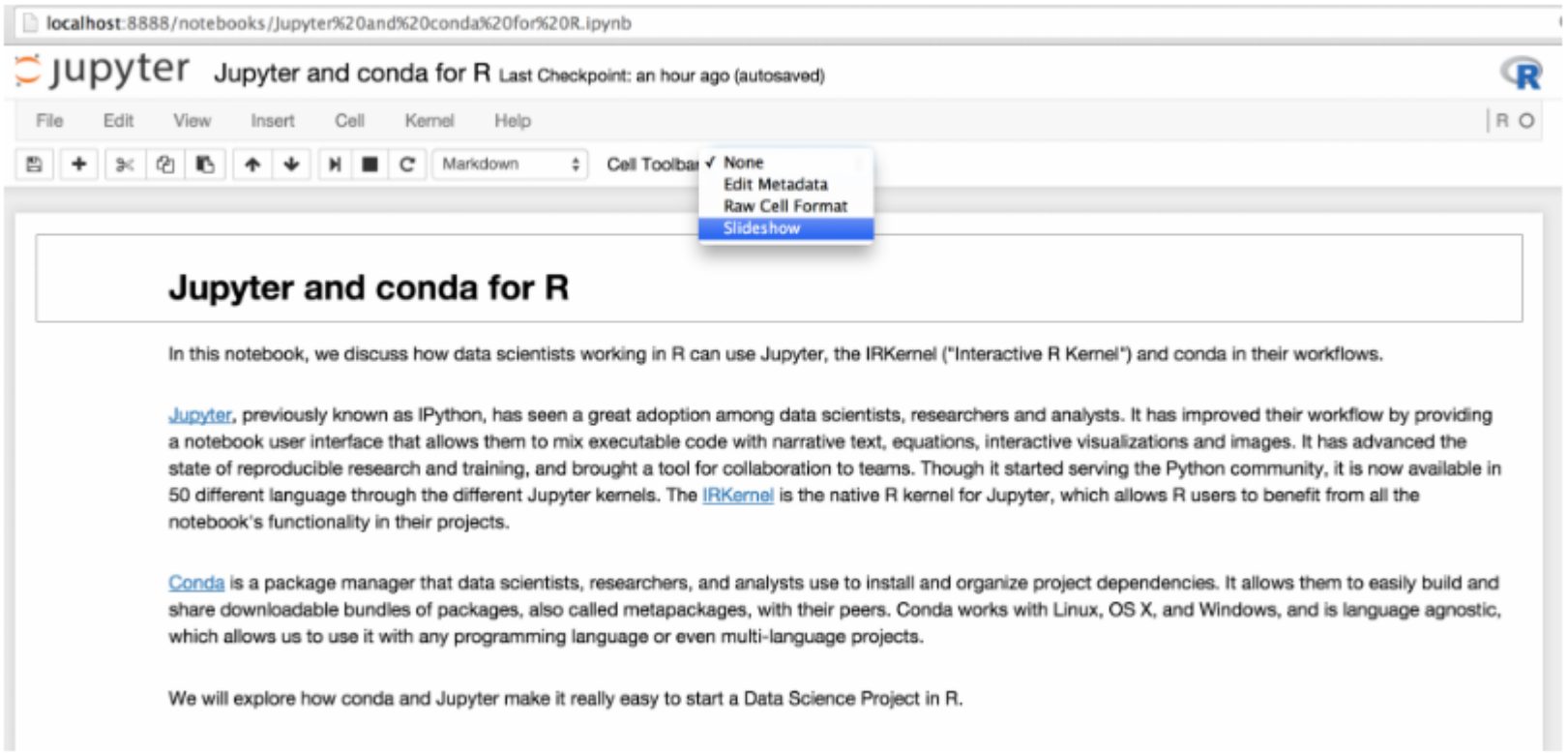
Figure 1: Correlation of given predictors.

Simple figure

Anaconda


CONTINUUM ANALYTICS
ANACONDA | COMMUNITY | SERVICES | SOLUTIONS | ABOUT | RESO
ANACONDA · DOWNLOAD
DOWNLOAD ANACONDA NOW!
Jump to: Windows | OS X | Linux
Get Superpowers with Anaconda
Anaconda is a completely free Python distribution (including for commercial use and redistribution). It includes more than 400 of the most popular Python packages for science, math, engineering, and data analysis. See the packages included with Anaconda and the Anaconda changelog.
Which version should I download and install?
Because Anaconda includes installers for Python 2.7 and 3.5, either is fine. Using either version, you can use Python 3.4 with the conda command. You can create a 3.5 environment with the conda command if you've downloaded 2.7 — and vice versa.

Jupyter - R interactive



https://www.continuum.io/blog/developer/jupyter-and-conda-r

Jupyter - R interactive

- The version of R that comes in Anaconda already comes with most of the necessary packages
- In any case, new packages can be installed using the install.packages () function.

```
install.packages('openxlsx', repos='http://cran.us.r-project.org')
```

```
package 'openxlsx' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\dvillanuevaj\AppData\Local\Temp\Rtmp6Z49XS\downloaded_packages
```

Markdown

Syntax Guide

Phrase Emphasis

```
*italic*   **bold**
_italic_   __bold__
```

Links

Inline:

```
An [example](http://url.com/ "Title")
```

Reference-style labels (titles are optional):

```
An [example][id]. Then, anywhere
else in the doc, define the link:

  [id]: http://example.com/  "Title"
```

Images

Inline (titles are optional):

```
![alt text](/path/img.jpg "Title")
```

Reference-style:

```
![alt text][id]

[id]: /url/to/img.jpg "Title"
```

Headers

Setext-style:

```
Header 1
========

Header 2
--------
```

atx-style (closing #'s are optional):

```
# Header 1 #

## Header 2 ##

###### Header 6
```

Code Spans

```
`<code>` spans are delimited
by backticks.

You can include literal backticks
like `` `this` ``.
```

Lists

Ordered, without paragraphs:

```
1.  Foo
2.  Bar
```

Unordered, with paragraphs:

```
*   A list item.

    With multiple paragraphs.

*   Bar
```

You can nest them:

```
*   Abacus
    * answer
*   Bubbles
    1.  bunk
    2.  bupkis
        * BELITTLER
    3. burper
*   Cunning
```

Blockquotes

```
> Email-style angle brackets
> are used for blockquotes.

> > And, they can be nested.

> #### Headers in blockquotes
>
> * You can quote a list.
> * Etc.
```

Preformatted Code Blocks

Indent every line of a code block by at least 4 spaces or 1 tab.

```
This is a normal paragraph.

    This is a preformatted
    code block.
```

Horizontal Rules

Three or more dashes or asterisks:

```
---
* * *
- - - -
```

Learn R with R

Learn R, in R.

swirl teaches you R programming and data science
interactively, at your own pace, and right in the R console!

http://swirlstats.com/

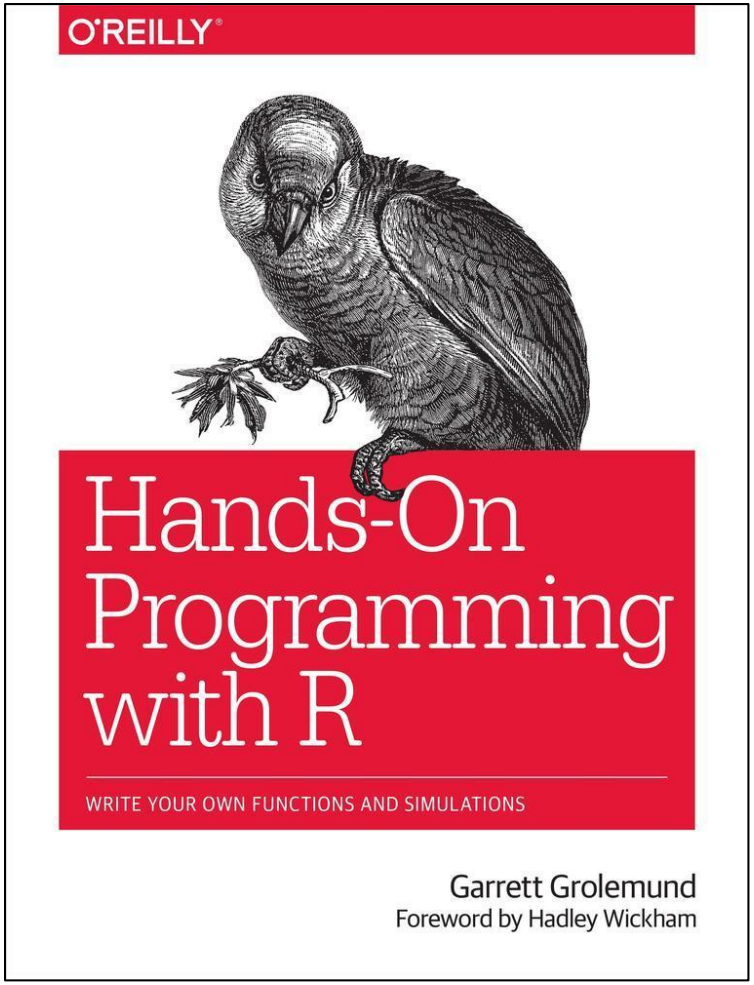
Learn R with R

```
>
> library(swirl)
> swirl()

| Welcome to swirl!

| Please sign in. If you've been here before, use the same name as you did then. If you
| are new, call yourself something unique.

What shall I call you? |
```

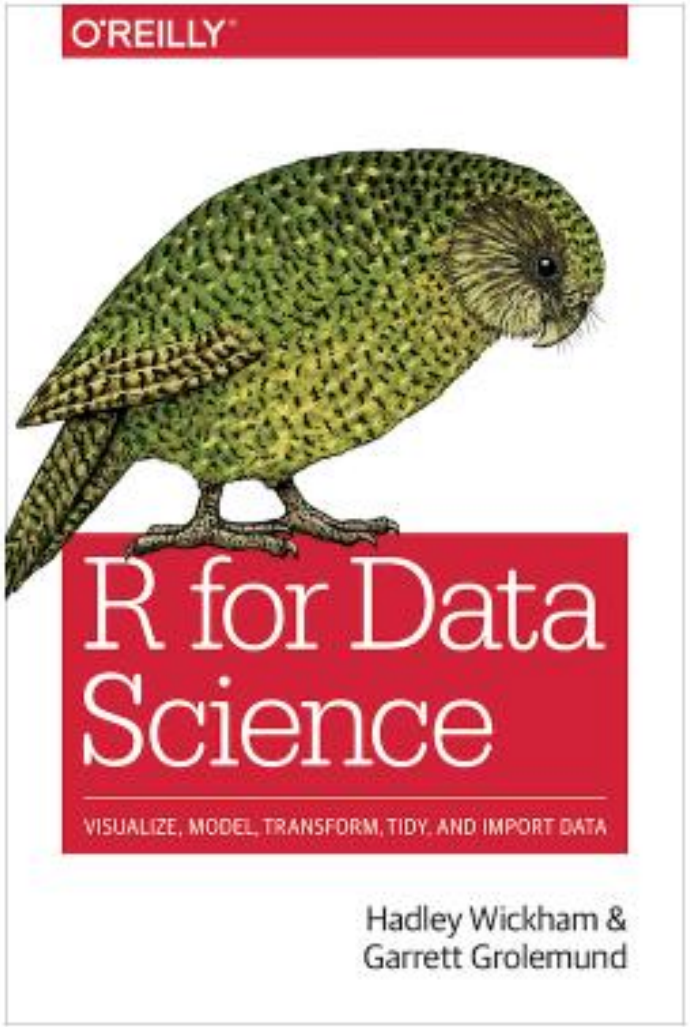
```
install_from_swirl("R Programming")
install_from_swirl("Getting and Cleaning Data")
install_from_swirl("Exploratory Data Analysis")
install_from_swirl("Open Intro")
install_from_swirl("Regression Models")
install_from_swirl("Statistical Inference")
```

```
install.packages("swirl")
library(swirl)
```

Bibliography

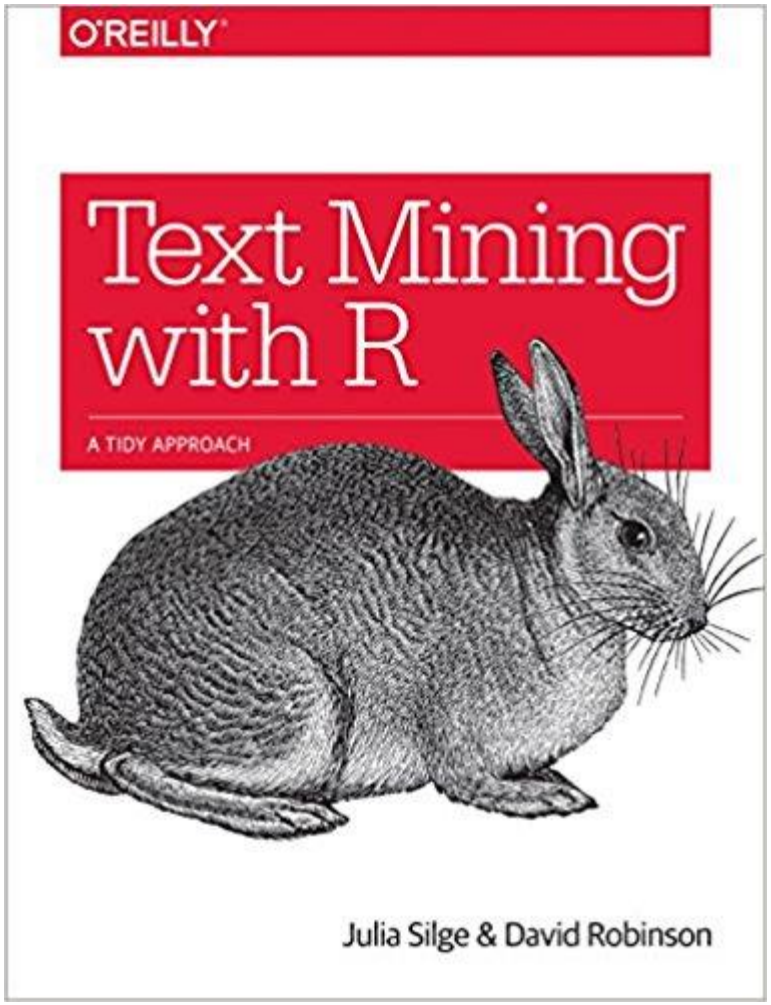


Bibliography



- http://r4ds.had.co.nz/

Bibliography



- http://r4ds.had.co.nz/

THANKS FOR YOUR ATTENTION

Daniel Villanueva Jiménez
daniel.villanueva@immune.institute
@dvillaj