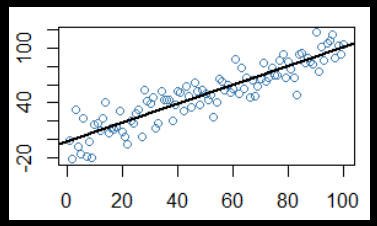
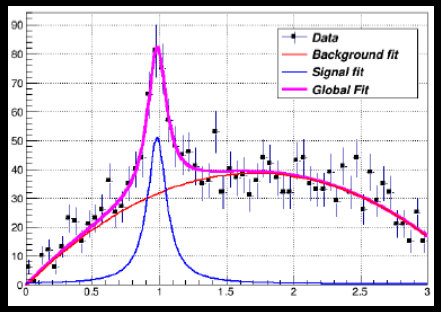
Temas avanzados en física computacional
Análisis de datos





Semestre

2016-I

Clase-2

José Bazo

jbazo@pucp.edu.pe

Contenidos del curso

✓ Introducción al análisis de datos y data science

2. Lenguaje de programación R

3. ROOT Data Analysis Framework

4. Manipulación y visualización de datos

5. Modelamiento estadístico

6. Machine Learning

7. TMVA (Toolkit for Multivariate Data Analysis)

Dinámica inicial

En grupos de 3-4 discutir y responder la siguiente pregunta:

¿Qué temas se deberían estudiar para aprender un lenguaje de programación?

Referencias

Coursera: R Programming at Johns Hopkins University:
https://www.coursera.org/learn/r-programming/

Peng. R Programming for Data Science. 2015

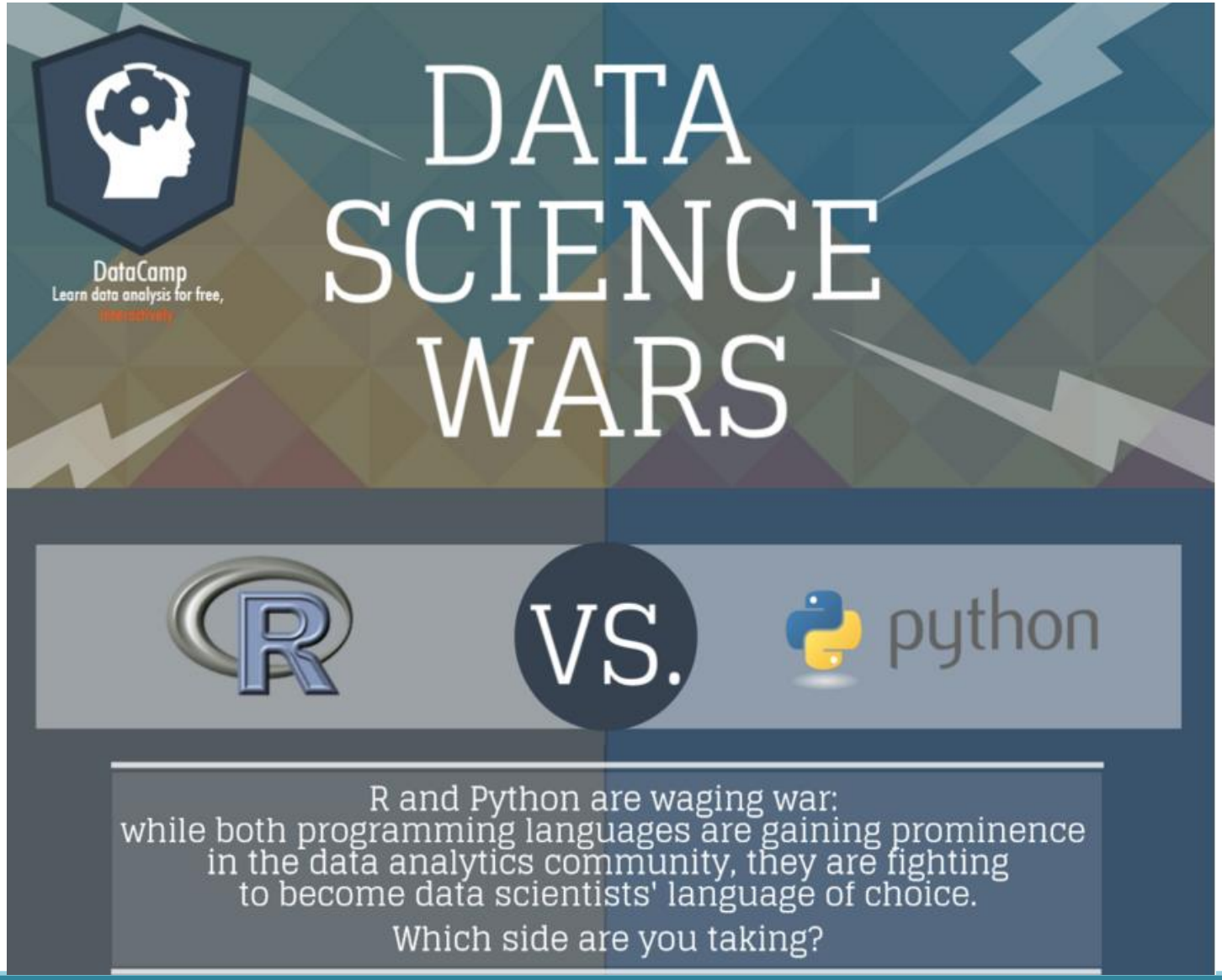
2. Lenguaje de programación R

Contenido

- Ingresar expresiones
- Tipos de objetos
- Atributos
- Datos categóricos
- Valores ausentes
- Formato de datos
- Nombres
- Leer y escribir datos, archivos
- Subconjuntos
- Operaciones vectoriales
- Estructuras de control
- Funciones y argumentos
- Asignación de valores a variables (scoping)
- Estándares de programación
- Fechas y tiempo
- Funciones con loop
- Debugging
- Números aleatorios
- Optimización de código

R vs Python

link



R vs Python


#1
Introducing The Opponents
Current Version
R version 3.2.4
March 2016
R version 3.5.1
Dec 2015
History
Creators
Ross Ihaka and Robert Gentleman
Release Year
1995
Must Knows
1. R is an implementation of S programming language (Bell Labs).
2. R's design and evolution is handled by the R-core group and R foundation.
3. R's software environment was written primarily in C, Fortran and R.
Creator
Guido Van Rossum
Release Year
1991
Must Knows
1. Python was inspired by C, Modula-3, and particularly ABC.
2. Python gets its name from the "Monty Python's Flying Circus" comedy series.
3. Python Software Foundation (PSF) takes care of Python's advances.

R vs Python

Purpose

R focuses on better, user friendly data analysis, statistics and graphical models.

Python emphasizes productivity and code readability.

Used By?

R has been used primarily in academics and research. However, R is rapidly expanding into the enterprise market.

"The closer you are to statistics, research and data science, the more you might prefer R."

Python is used by programmers that want to delve into data analysis or apply statistical techniques, and by developers that turn to data science.

"The closer you are to working in an engineering environment, the more you might prefer Python."

Community

Huge community with support coming in the form of:
- Mailing lists
- User-contributed documentation
- Active Stackoverflow members

More adoption from researchers, data scientists, statisticians, quants.

Overall good support for general purpose coding. Python support is found at:
- Stackoverflow
- Mailing lists
- User-contributed code and documentation

More adoption from developers and programmers.

R vs Python

Usability	
Statistical models can be written with only a few lines.	Coding and debugging is easier to do in Python, mainly because of the "nice" syntax.
There are R stylesheets but not everyone uses them.	The indentation of the code affects its meaning.
The same piece of functionality can be written in several ways in R.	Any piece of functionality is always written the same way in Python.
Flexibility	
It is easy to use complex formulas in R. All kinds of statistical tests and models are readily available and easily used.	Python is flexible for doing something novel that has never been done before. Developers can also use it for scripting a website or other applications.
Ease of Learning	
R has a steep learning curve at start. Once you know the basics, you can easily learn advanced stuff.	Python's focus on readability and simplicity makes that its learning curve is relatively low and gradual.
R is not hard for experienced programmers.	Python is considered a good language for starting programmers.

R vs Python

Code Repositories

CRAN stands for the Comprehensive R Archive Network: it is a huge repository of R packages to which users can easily contribute.

Packages are collections of R functions, data, and compiled code. They can be installed in R with one line.

PyPi is the Python Package Index: it is a repository of Python software, consisting of libraries. Users can contribute to Pypi, but it is a bit complicated in practice.

Watch out with dependencies and installing Python libraries!

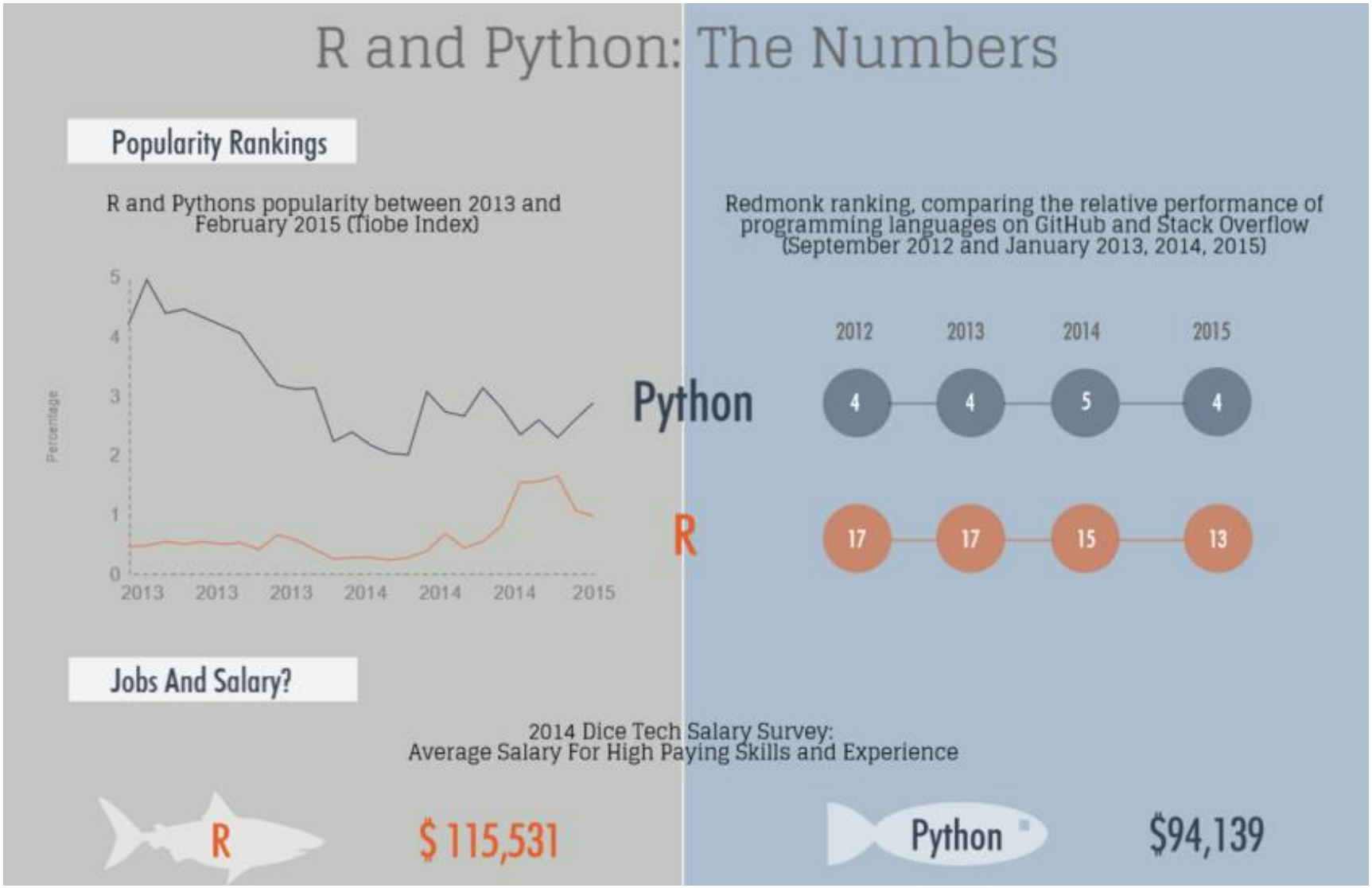
"I don't see Python [...] building up a huge code repository comparable to CRAN. [R has] a gigantic head start, [and] [...] statistics simply is not Python's central mission;"
- Norm Matloff, professor of computer science

Miscellaneous

Use the rPython package to run Python code from R. Pass or get data from Python, call Python functions or methods.

Use the RPy2 library to run R code from within Python. It provides a low-level interface from Python to R.

R vs Python


R and Python: The Numbers
Popularity Rankings
R and Pythons popularity between 2013 and February 2015 (Tiobe Index)
Percentage
5
4
3
2
1
0
2013
2013
2013
2014
2014
2014
2015
Redmonk ranking, comparing the relative performance of programming languages on GitHub and Stack Overflow (September 2012 and January 2013, 2014, 2015)
2012
2013
2014
2015
Python
4
4
5
4
R
17
17
15
13
Jobs And Salary?
2014 Dice Tech Salary Survey:
Average Salary For High Paying Skills and Experience
R
$ 115,531
Python
$94,139

R vs Python

#2 The Data Analysis Battlefield

Usage

R is mainly used when the data analysis tasks require standalone computing or analysis on individual servers.

Python is generally used when the data analysis tasks need to be integrated with web apps or if statistics code needs to be incorporated into a production database.

Task

For exploratory work, R is easier for beginners. Statistical models can be written with a few lines of code.

As a full-fledged programming language, Python is a good tool to implement algorithms for production use.

Data Handling Capabilities

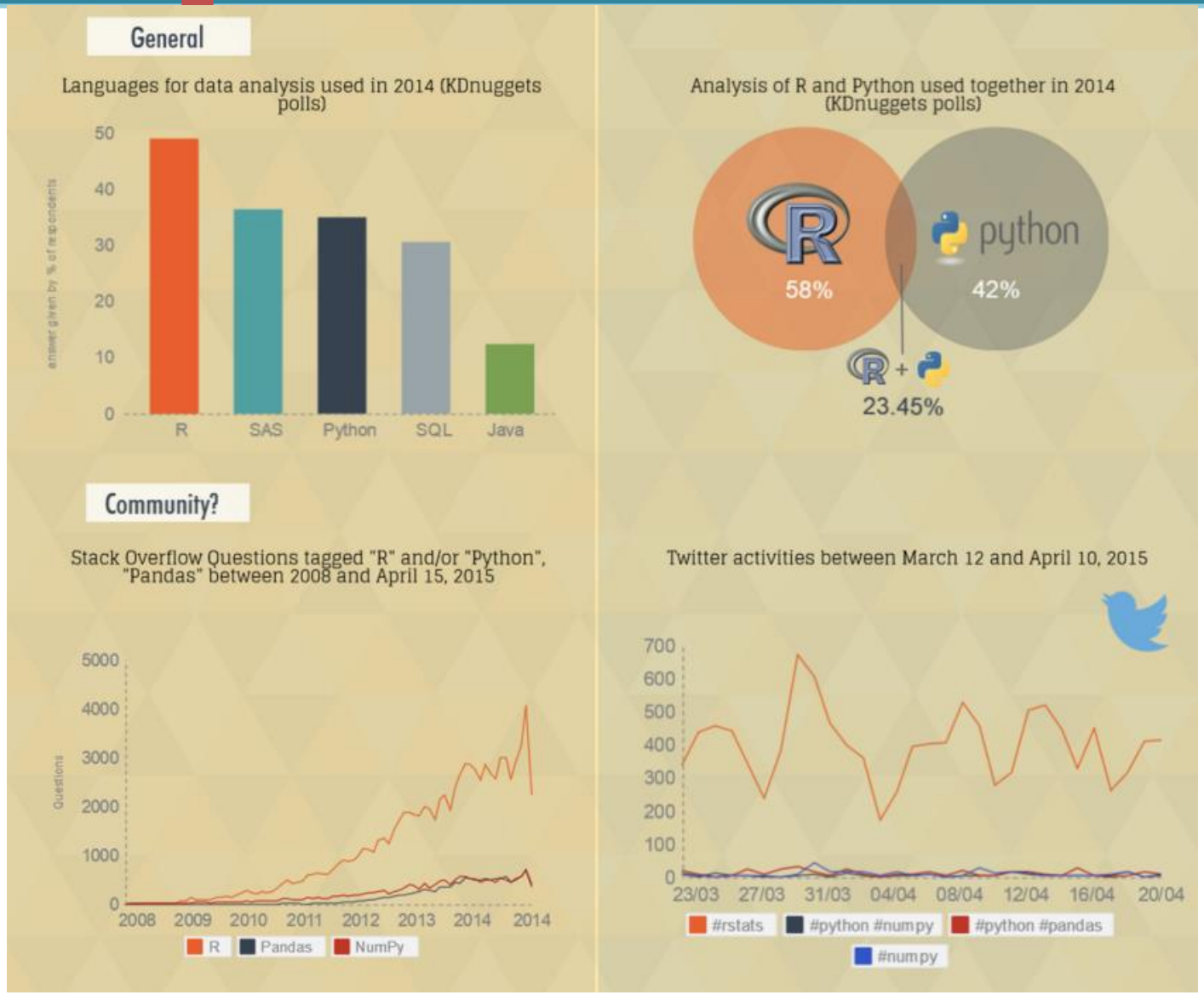
R is handy for data analysis because of the huge number of packages, readily usable tests and the advantage of using formulas.

R is usable for basic data analysis without the installation of packages. Big datasets require the use of packages such as data.table and dplyr, though.

The infancy of Python packages for data analysis was an issue in the past, but this has improved a lot!

You need to use NumPy and pandas (amongst others) to make Python usable for data analysis.

R vs Python


General
Languages for data analysis used in 2014 (KDnuggets polls)
answer given by % of respondents
50
40
30
20
10
0
R
SAS
Python
SQL
Java
Analysis of R and Python used together in 2014 (KDnuggets polls)
58%
42%
23.45%
Community?
Stack Overflow Questions tagged "R" and/or "Python", "Pandas" between 2008 and April 15, 2015
Questions
5000
4000
3000
2000
1000
0
2008
2009
2010
2011
2012
2013
2014
2014
R
Pandas
NumPy
Twitter activities between March 12 and April 10, 2015
700
600
500
400
300
200
100
0
23/03
27/03
31/03
04/04
08/04
12/04
16/04
20/04
#rstats
#python #numpy
#python #pandas
#numpy

R vs Python

"My current strategy is to leverage the best of both worlds — do early stage data analysis in R, then switch to Python when it's time to get serious, be a team player, and ship some real code and data products."

• • •

"I use R to conduct statistical tests, graph data, and inspect large data sets. If I actually have to write an algorithm, I prefer Python..."

• • •

"I'd rather do math in a general-purpose language than try to do general-purpose programming in a math language."

link

Para practicar

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

> library(swirl)

> swirl()
```

link

Seguir las instrucciones y practicar en 15 lecciones con ejercicios:
1: R Programming: The basics of programming in R