R Basics and Differences from Other Languages

Basic Types

```
Numeric:    x = 2.5

Integer:    y = as.integer(value)        v <- c(1L, 2L, 3L)

Complex:    z = 1 + 2i
    >   sqrt(-1)  Error!
    >   sqrt(-1 + 0i)
    >   sqrt(as.complex(-1))

Logical:    v = True       t = 7 > 9

Character:  name = “Obama”
```

Data Structures in R

Vector

(atomic type)

- Collections of cells with elements of the same type
- Default mode is logical

```
>   v <- vector()
>   v <- c(#, #, #, … , #)
```

Matrix

- Vector with an added dimension
- 2 dimensional rectangular layout

```
>   m <- matrix(data, nrow, ncol, byrow, dimnames)
>   z <- 1:10
>   dim(z) <- c(2,5)     #2 rows, 5 columns
```

Array

(homogenous type)

- Storing data in **n** dimensions
- Vectors with a dimension attribute

```
>   a <- array(1:10, c(2, 5))

>   m <- matrix(1:10, 2)
```

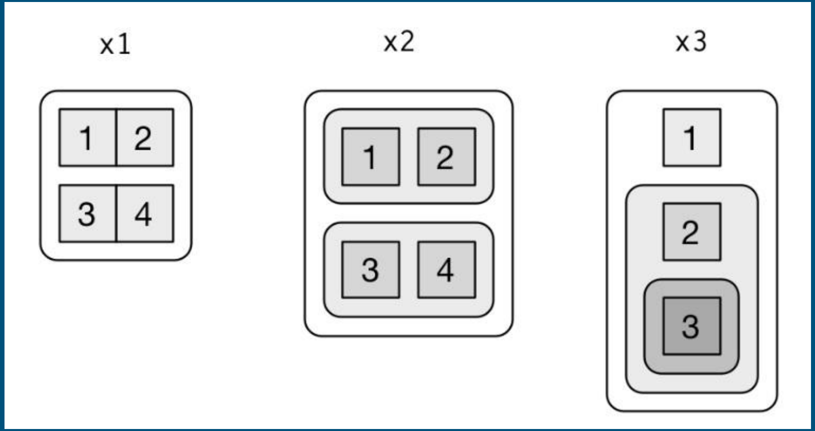
Vector is not the same as a 1 dimensional array

Matrix is the same as a 2 dimensional array

List

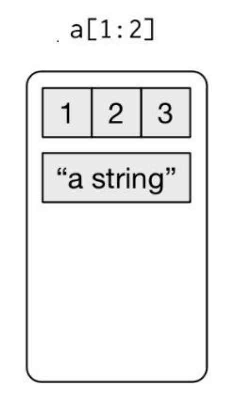
- Mixture of data types
- Can store other lists

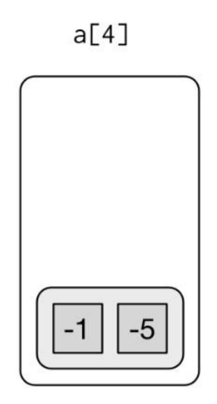
```
>   x1 <- list(c(1,2), c(3,4))
>   x2 <- list(list(1,2), list(3,4))
>   x3 <- list(1, list(2, list(3)))
```

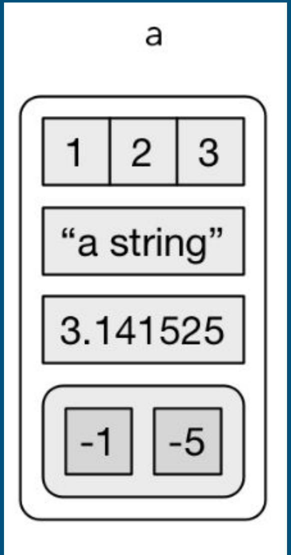


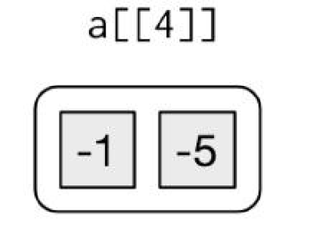
List exercise

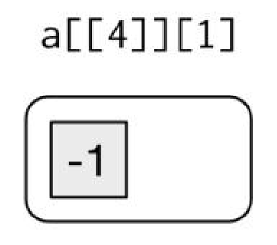
```
>   a <- list(1.3, “a string”, pi, list(-1, -5))
>   a[1:2]
>   a[4]
>   a[[4]]
>   a[[4]][1]
>   a[[4]][[1]]
```

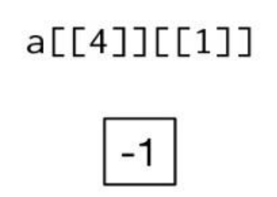












Dataframe

- Most common way of storing data
- Heterogeneous

```
> firstNames = c("Emma", "Olivia", "Sophia", "Ava", "Isabella", "Mia")

> df <- data.frame(ID = letters[1:6], HOME = sample(state.name, 6), NAME=firstNames)

> df

  ID       HOME     NAME
1  a   Illinois     Emma
2  b   Nebraska   Olivia
3  c    Indiana   Sophia
4  d New Jersey      Ava
5  e   Maryland Isabella
6  f    Alabama      Mia
```

Base R & Tidyverse

Base R	Tidyverse
built in “out of the box” R	independent packages
standard object types	dataframes and tibbles
indexing based (e.g. df[df$group == “a”,])	verb based (e.g. filter(df, group == “a”))
no method chaining	Lots of method chaining (%>%

R vs. Python

R and Python are waging war:
while both programming languages are gaining prominence in the data analytics community, they are fighting to become data scientists' language of choice.

Which side are you taking?

Introducing The Opponents

https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis

R & Python

DataCamp
Learn data analysis for free.
R
&
python

DATA SCIENCE ~~WARS~~

R and Python are waging war:
while both programming languages are gaining prominence in the data analytics community, they are fighting to become data scientists' language of choice.
Which side are you taking?

#1 Introducing The Options

https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis

R & Python

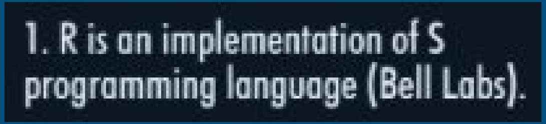
DATA SCIENCE WARS

R & python

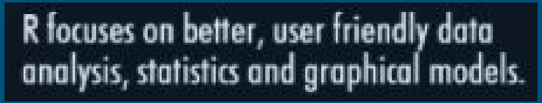
R and Python are waging war: while both programming languages are gaining prominence in the data analytics community, they are fighting to become data scientists' language of choice. Which side are you taking?

#1 Introducing The Options

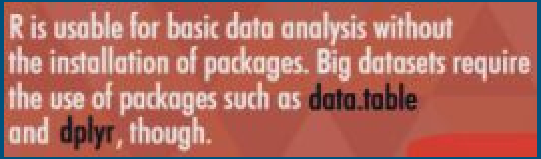
R



1. R is an implementation of S programming language (Bell Labs).

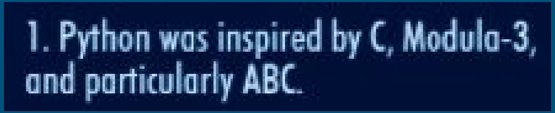


R focuses on better, user friendly data analysis, statistics and graphical models.

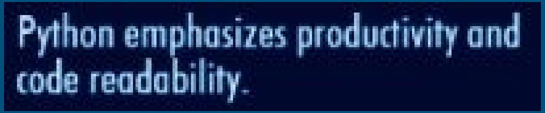


R is usable for basic data analysis without the installation of packages. Big datasets require the use of packages such as data.table and dplyr, though.

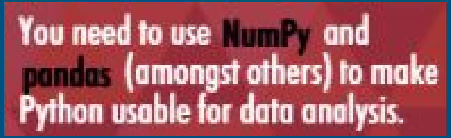
Python



1. Python was inspired by C, Modula-3, and particularly ABC.



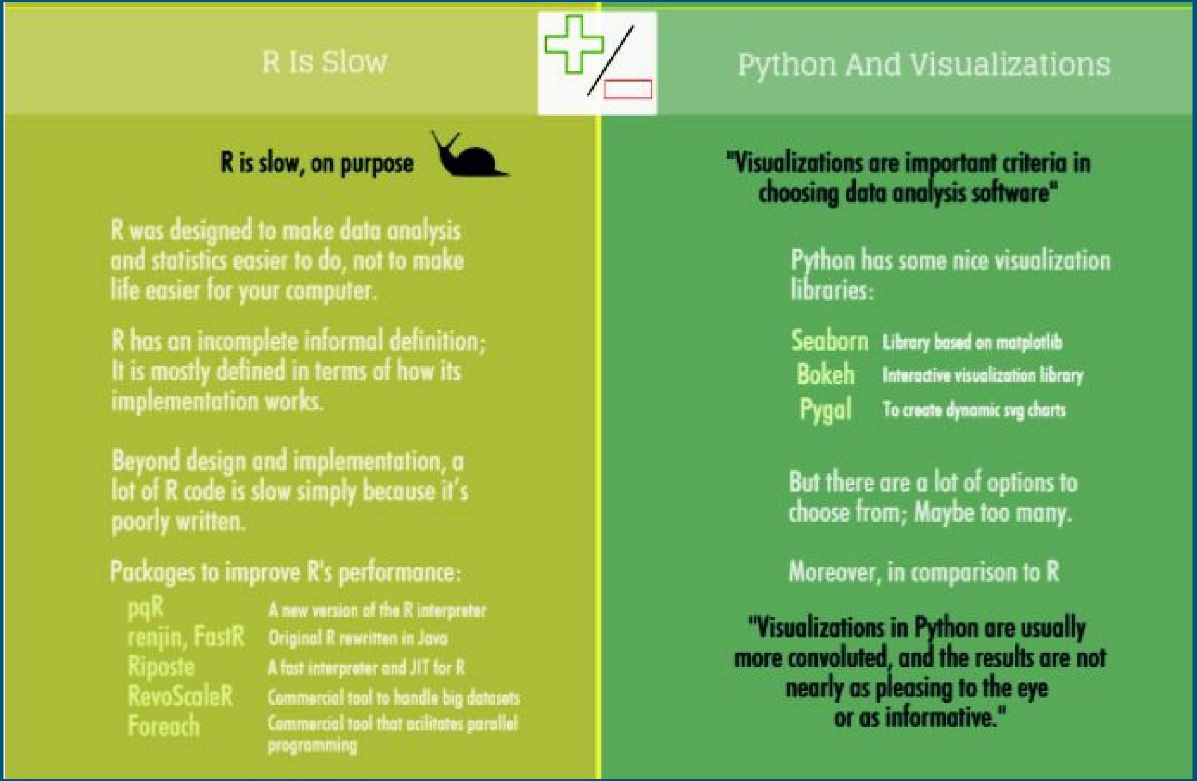
Python emphasizes productivity and code readability.



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

https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis

R & Python



R Is Slow

R is slow, on purpose

R was designed to make data analysis and statistics easier to do, not to make life easier for your computer.

R has an incomplete informal definition; It is mostly defined in terms of how its implementation works.

Beyond design and implementation, a lot of R code is slow simply because it's poorly written.

Packages to improve R's performance:

pqR	A new version of the R interpreter
renjin, FastR	Original R rewritten in Java
Riposte	A fast interpreter and JIT for R
RevoScaleR	Commercial tool to handle big datasets
Foreach	Commercial tool that acilitates parallel programming

Python And Visualizations

"Visualizations are important criteria in choosing data analysis software"

Python has some nice visualization libraries:

Seaborn	Library based on matplotlib
Bokeh	Interactive visualization library
Pygal	To create dynamic svg charts

But there are a lot of options to choose from; Maybe too many.

Moreover, in comparison to R

"Visualizations in Python are usually more convoluted, and the results are not nearly as pleasing to the eye or as informative."

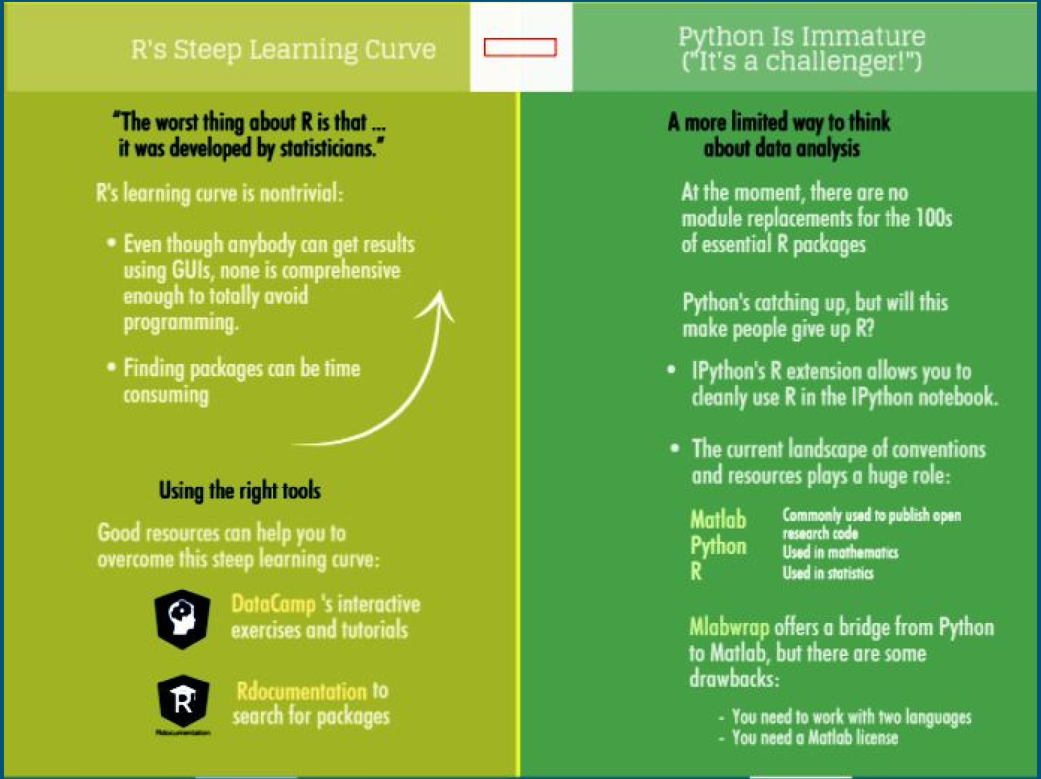
DATA SCIENCE ~~WARS~~

R and Python are waging war:
while both programming languages are gaining prominence in the data analytics community, they are fighting to become data scientists' language of choice.
Which side are you taking?

#1 Introducing The Options

https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis

R & Python



R's Steep Learning Curve

Python Is Immature ("It's a challenger!")

"The worst thing about R is that ... it was developed by statisticians."

R's learning curve is nontrivial:

- Even though anybody can get results using GUIs, none is comprehensive enough to totally avoid programming.
- Finding packages can be time consuming

Using the right tools

Good resources can help you to overcome this steep learning curve:

DataCamp 's interactive exercises and tutorials

Rdocumentation to search for packages

A more limited way to think about data analysis

At the moment, there are no module replacements for the 100s of essential R packages

Python's catching up, but will this make people give up R?

- IPython's R extension allows you to cleanly use R in the IPython notebook.
- The current landscape of conventions and resources plays a huge role:

Matlab	Commonly used to publish open research code
Python	Used in mathematics
R	Used in statistics

Mlabwrap offers a bridge from Python to Matlab, but there are some drawbacks:

- You need to work with two languages
- You need a Matlab license

DATA SCIENCE ~~WARS~~

R & python

R and Python are waging war: while both programming languages are gaining prominence in the data analytics community, they are fighting to become data scientists' language of choice. Which side are you taking?

#1 Introducing The Options

https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis

R & Python: Data Structures

R	**Python**
atomic vector, list	list, tuple
named list	dictionary
matrix, array	matrix, array (via numpy)
dataframe	dataframe (via pandas)