Introduction to R

Indian Institute of
Information Technology
Kottayam

Types of statistical software

- command-line software
  - requires knowledge of syntax of commands
  - reproducible results through scripts
  - detailed analyses possible
- GUI-based software
  - does not require knowledge of commands
  - not reproducible actions
- hybrid types (both command-line and GUI)

Indian Institute of
Information Technology
Kottayam

Well-known statistical software

- SAS
- SPSS
- Minitab
- Statgraphics
- S-Plus
- R
- …

R

- free
- language almost the same as S
- maintained by top quality experts
- available on all platforms
- continuous improvement

Available through www.r-project.org

SPSS =

muggle

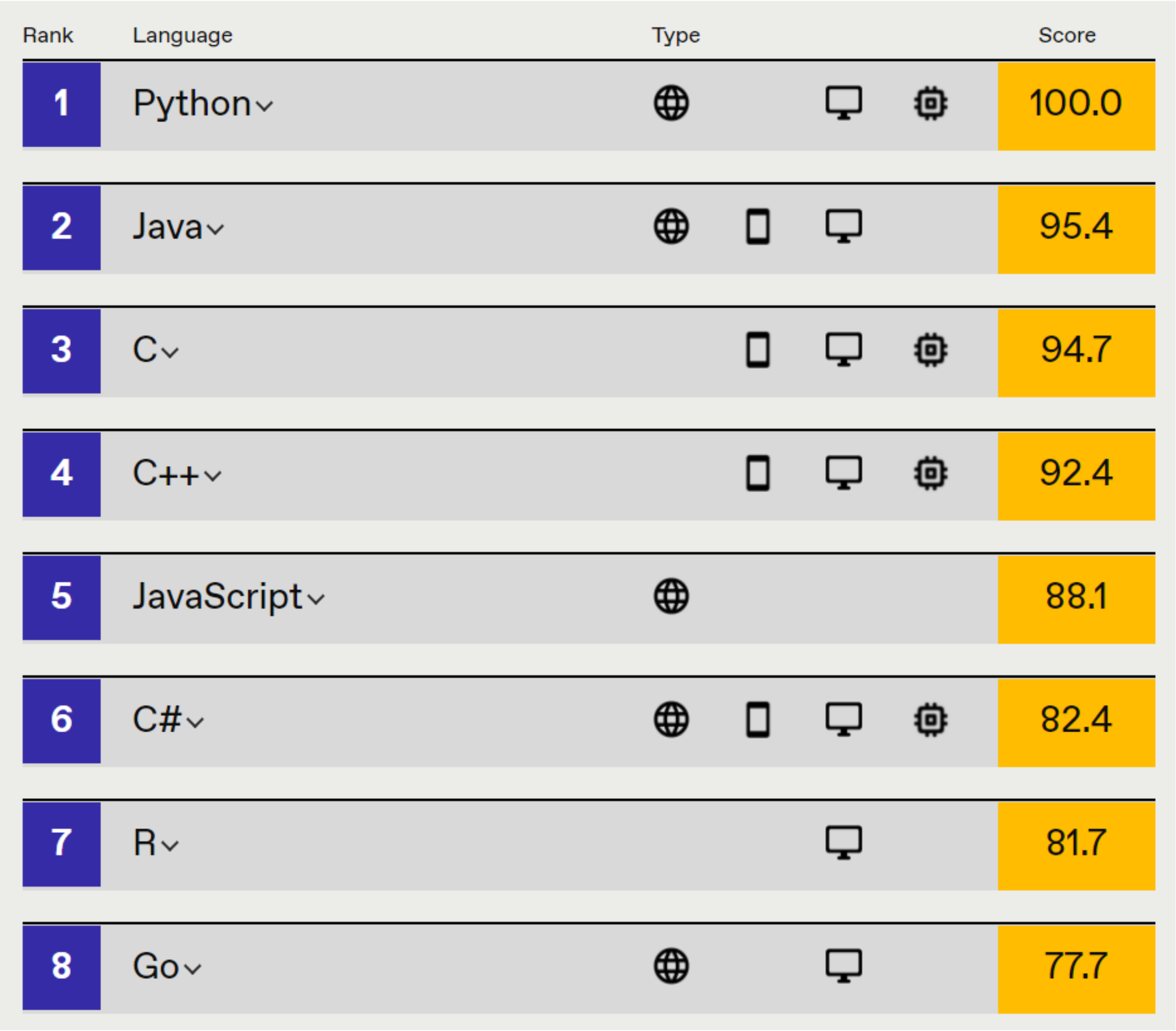
SPSS and SAS users are like muggles. They are limited in their ability to change their environment. They have to rely on algorithms that have been developed for them. The way they approach a problem is constrained by how SAS/SPSS employed programmers thought to approach them. And they have to pay money to use these constraining algorithms.

=

wizard

R users are like wizards. They can rely on functions (spells) that have been developed for them by statistical researchers, but they can also create their own. They don’t have to pay for the use of them, and once experienced enough (like Dumbledore), they are almost unlimited in their ability to change their environment.

Language Ranking : IEEE Spectrum



Rank	Language	Type				Score
1	Python	🌐		🖥	▣	100.0
2	Java	🌐	📱	🖥		95.4
3	C		📱	🖥	▣	94.7
4	C++		📱	🖥	▣	92.4
5	JavaScript	🌐				88.1
6	C#	🌐	📱	🖥	▣	82.4
7	R			🖥		81.7
8	Go	🌐		🖥		77.7

Learning R....

Indian Institute of
Information Technology
Kottayam

R

Advantages

Disadvantages

- Fast and free.
- State of the art: Statistical researchers provide their methods as R packages. SPSS and SAS are years behind R!
- Mx, WinBugs, and other programs use or will use R.
- Active user community
- Excellent for simulation, programming, computer intensive analyses, etc.
- Forces you to *think* about your analysis.
- Interfaces with database storage software (SQL)

Indian Institute of Information Technology Kottayam

R

Advantages

- Fast and free.
- State of the art: Statistical researchers provide their methods as R packages. SPSS and SAS are years behind R!
- Mx, WinBugs, and other programs use or will use R.
- Active user community
- Excellent for simulation, programming, computer intensive analyses, etc.
- Forces you to *think* about your analysis.
- Interfaces with database storage software (SQL)

Disadvantages

- Not user friendly @ start - steep learning curve, minimal GUI.
- No commercial support; figuring out correct methods or how to use a function on your own can be frustrating.
- Easy to make mistakes and not know.
- Working with large datasets is limited by RAM
- Data prep & cleaning can be messier & more mistake prone in R vs. SPSS or SAS.

Tools to work efficiently with R

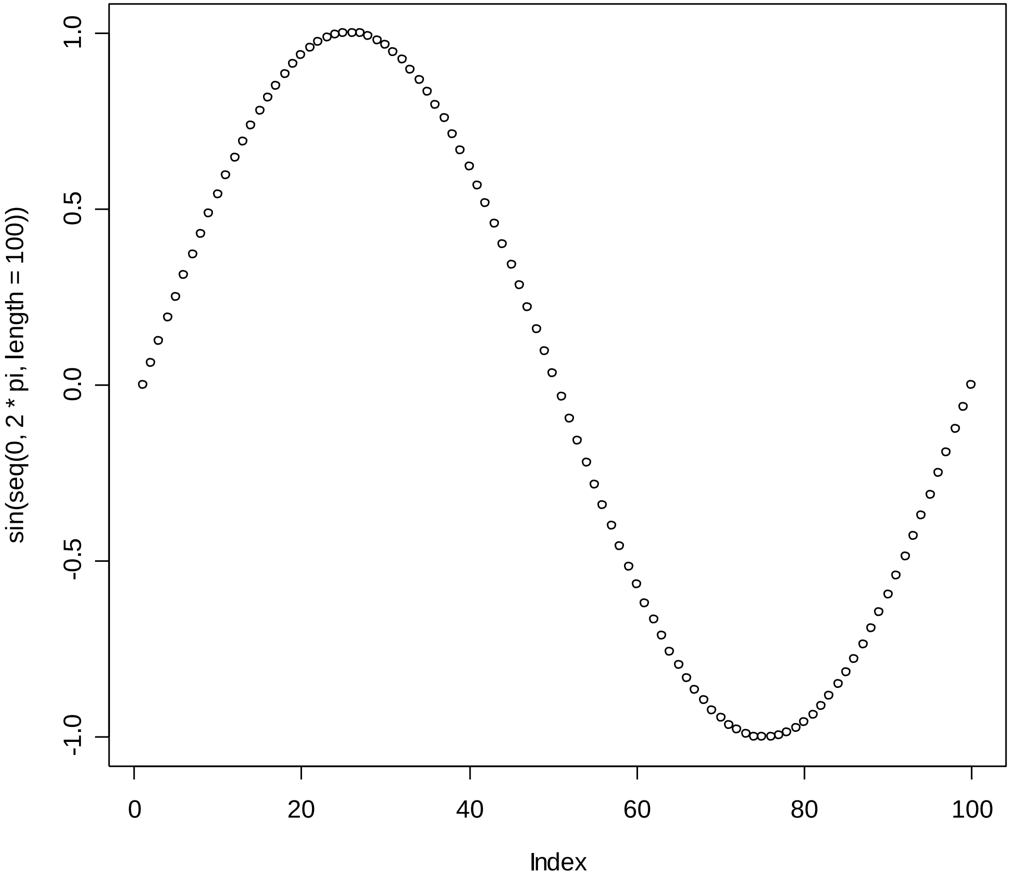
- Pick an IDE or a powerful editor
- The most popular IDE for R is Rstudio www.rstudio.com
  - RStudio offers tabs to navigate files, browse installed packages, visualize plots, among other features, as well as a large amount of configuration options under the top menu dropdowns.
- For this session you may use :
  - https://rstudio.iiitkottayam.ac.in/
  - Sign Up using your Roll Number and the password send for first login of email id

R as a Calculator

```
> log2(32)
[1] 5

> print(sqrt(2))
[1] 1.414214

> pi
[1] 3.141593

> seq(0, 5, length=6)
[1] 0 1 2 3 4 5

> 1+1:10
 [1]  2  3  4  5  6  7  8  9 10 11
```

R as a Graphics Tool

```
> plot(sin(seq(0, 2*pi, length=100)))
```



Indian Institute of
Information Technology
Kottayam

Variables

```
> a <- 49
> sqrt(a)
[1] 7
```

numeric

```
> b <- "The dog ate my homework"
> sub("dog","cat",b)
[1] "The cat ate my homework"
```

character
string

```
> c <- (1+1==3)
> c
[1] FALSE
> as.character(b)
[1] "FALSE"
```

logical

Vectors

vector: **an ordered collection of data of the same type**

```
> a <- c(1,2,3)
> a*2
[1] 2 4 6
```

Example: the mean spot intensities of all 15488 spots on a microarray is a numeric vector

In R, a single number is the special case of a vector with 1 element.

Other vector types: character strings, logical

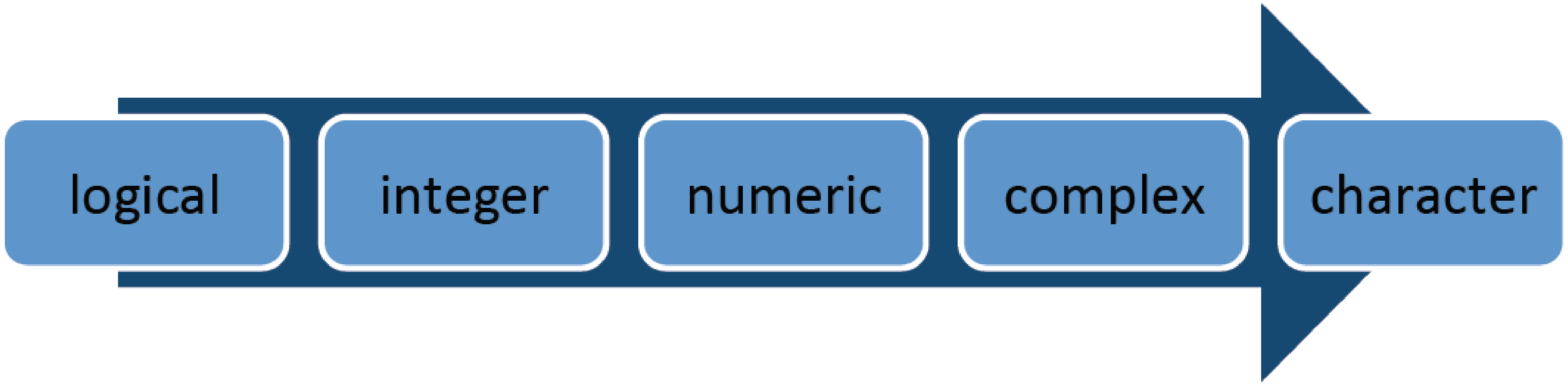
Vectors

Data type conversion

Vectors may only have one type

When combining different types, R will **coerce** a vector to the most flexible type

Coercion rule in R:



logical

integer

numeric

complex

character

Matrices and Arrays

matrix: rectangular table of data of the same type

A matrix is a two-dimensional vector (fixed size, all cell types the same).

array: 3-,4-,..dimensional matrix

An array is a vector with one or more dimensions.

So, an array with one dimension is (almost) the same as a vector. An array with two dimensions is (almost) the same as a matrix. An array with three or more dimensions is an n-dimensional array.

Data Frames

data frame: rectangular table with rows and columns; data within each column has the same type (e.g. number, text, logical), but different columns may have different types.

Represents the typical data table that researchers come up with – like a spreadsheet.

Example:

```
> a <-
data.frame(localization,tumorsize,progress,row.names=p
atients)
> a
      localization tumorsize progress
XX348     proximal       6.3    FALSE
XX234       distal       8.0     TRUE
XX987     proximal      10.0    FALSE
```

Lists

list: ordered collection of data of arbitrary types.

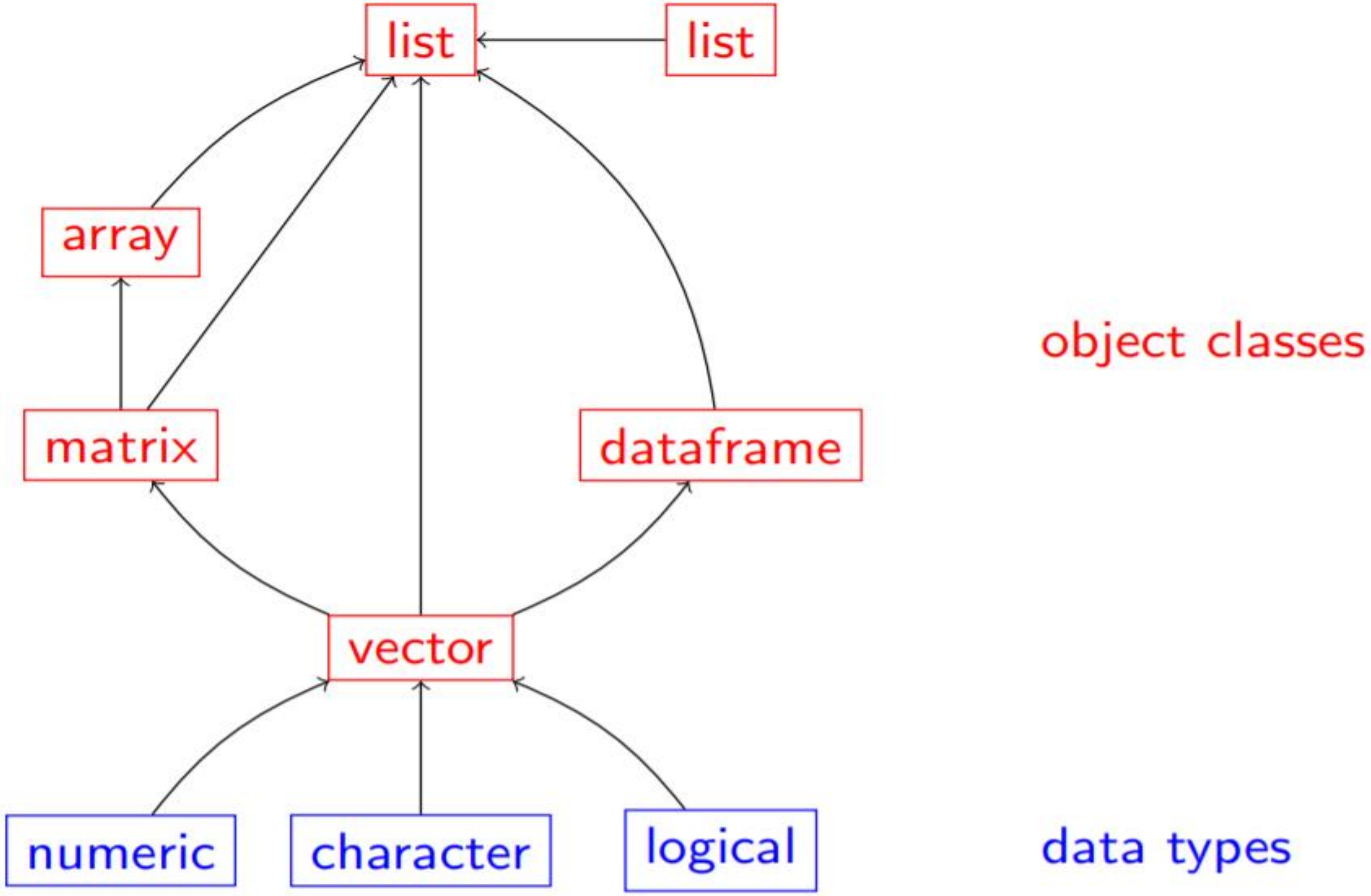
A list can hold items of different types and the list size can be increased on the fly. List contents can be accessed either by index (like mylist[[1]]) or by name (like mylist$age).

Example:

```
> doe <- list(name="john",age=28,married=F)
> doe$name
[1] "john‶
> doe$age
[1] 28
> doe[[3]]
[1] FALSE
```

Typically, vector elements are accessed by their index (an integer) and list elements by $name (a character string). But both types support both access methods. Slots are accessed by @name.

Data Types and Objects in R

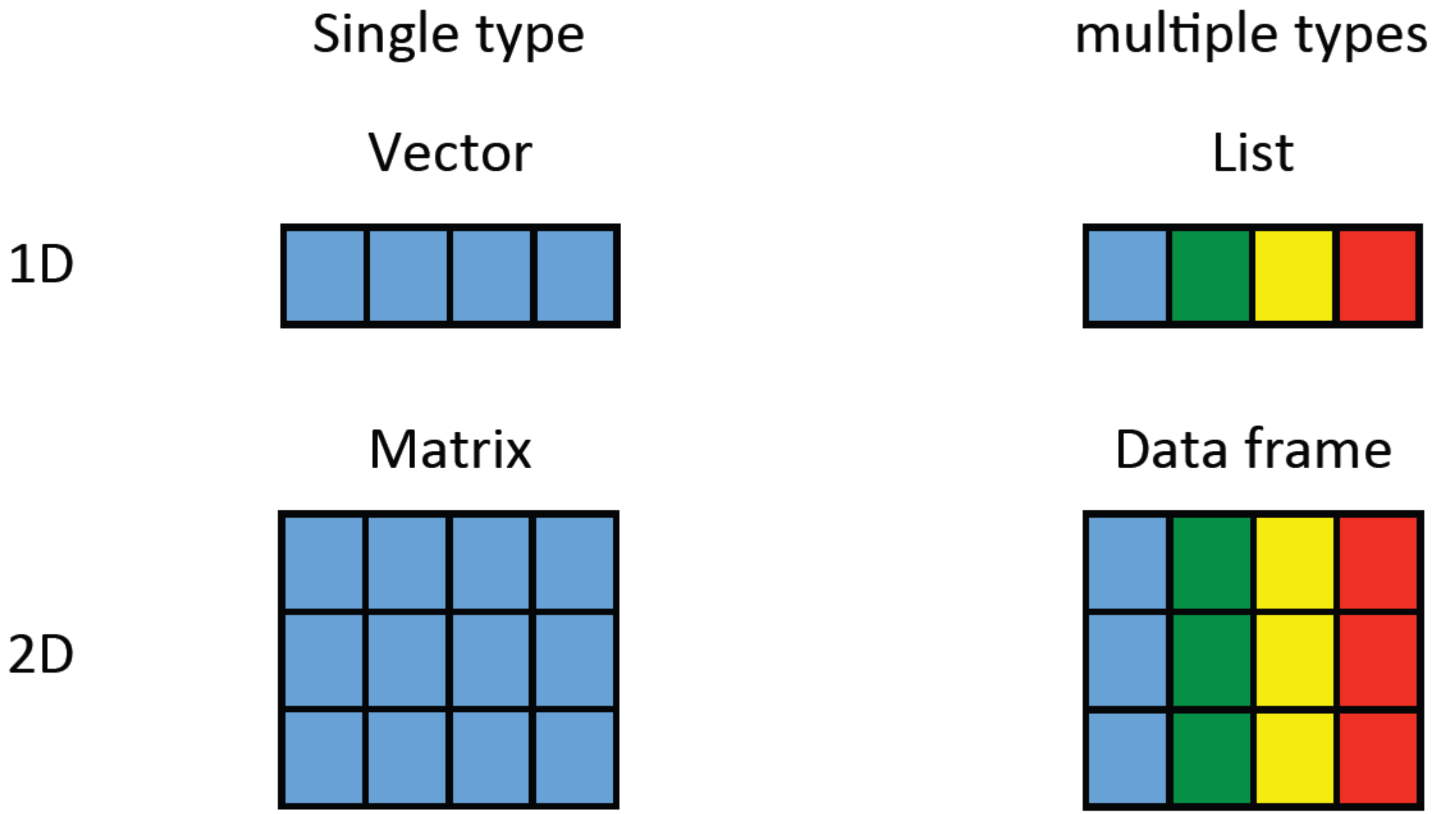


Indian Institute of
Information Technology
Kottayam

Factors

- Factors usually look like character data, but are typically used to represent categorical information.

Summary of R data structures



	Single type	multiple types
1D	Vector	List
2D	Matrix	Data frame