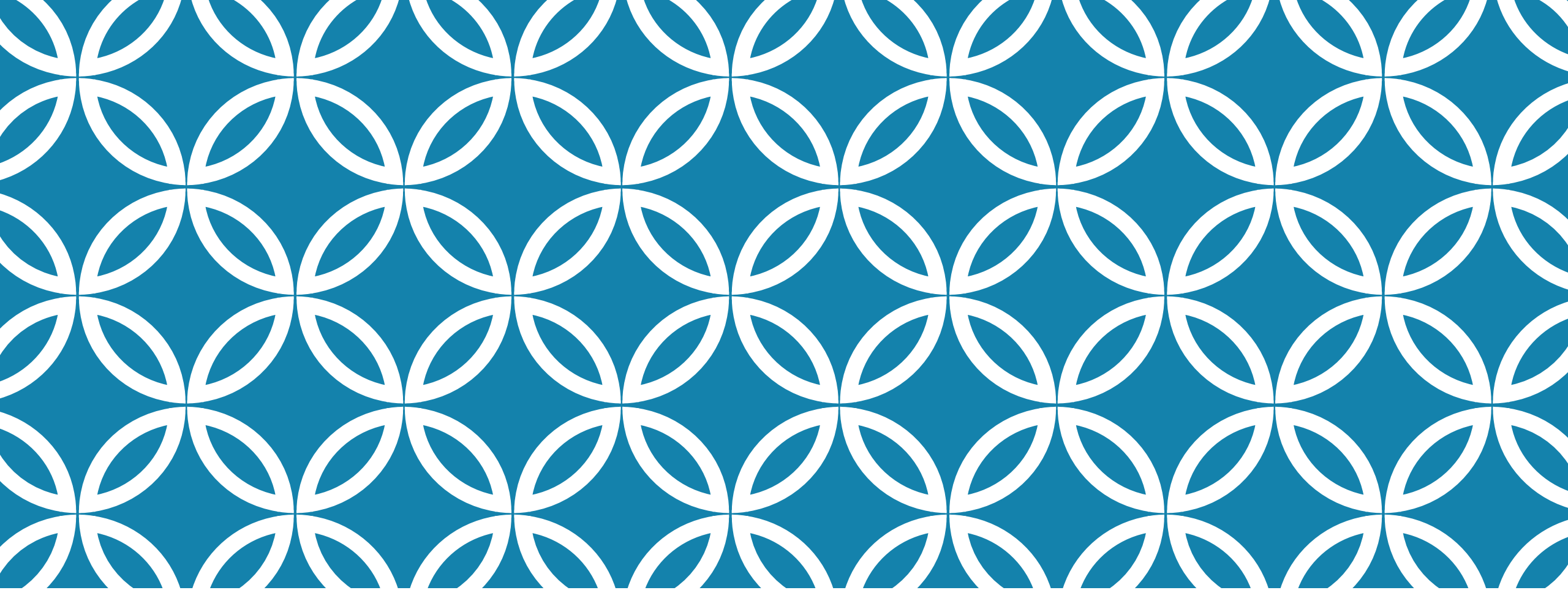
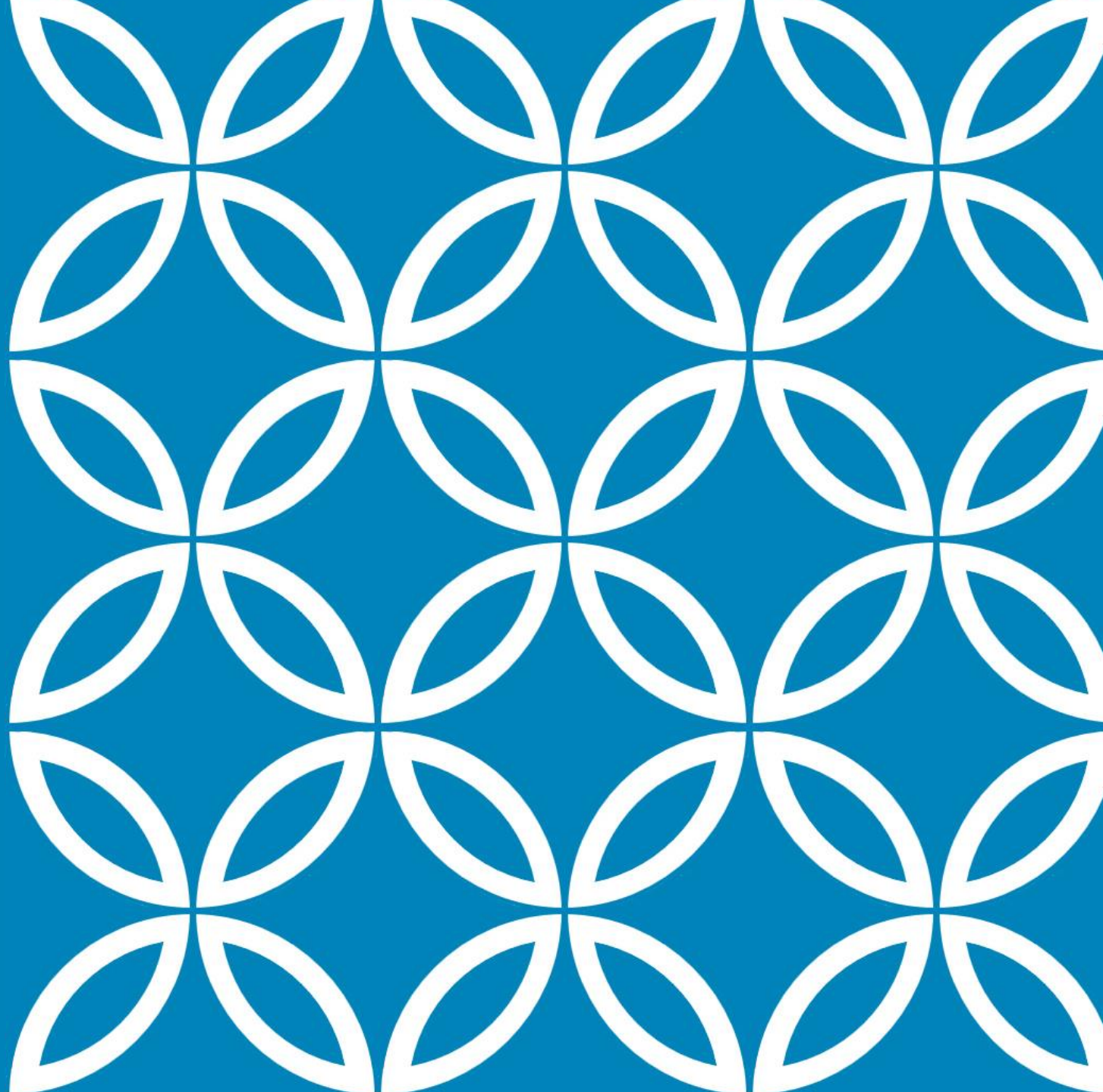


DATA FRAMES PART 1

mgr Maria Kubara, WNE UW

DATA FRAME

Structure used to store data in a table form (the most common structure in statistical data analysis and machine learning). It can be seen as a list of vectors of equal length (commonly with unique names). The most important basic data structure in the *tidyverse* environment.



CREATING A DATA FRAME

Data frame object is created with data.frame() function, which takes vectors of equal length as its input (can be of different types). Data frame can be also created by conversion from a matrix with as.data.frame().

Vectors in data.frame are stored vertically and the name of the original vectors are stored as column names. Data frame is a tabular data storage (like in Excel).

```
> ### Creating data frame ########################################
>
> # Vectors must have equal length, but can have different types
> column1 <- c(1:3)
> column2 <- c("Anna", "Tom", "Sue")
> column3 <- c(T, T, F)
>                                    Creating data.frame
> dataset1 <- data.frame(column1, column2, column3)
> dataset1
  column1 column2 column3
1       1    Anna    TRUE
2       2     Tom    TRUE
3       3     Sue   FALSE
>
> colnames(dataset1) # names of vectors are stored as column names
[1] "column1" "column2" "column3"
> colnames(dataset1)[2] <- "name"
> dataset1                                  Changing name of
  column1 name column3                      the second column
1       1 Anna    TRUE
2       2  Tom    TRUE
3       3  Sue   FALSE
```

TAKING VALUES FROM DATA FRAME

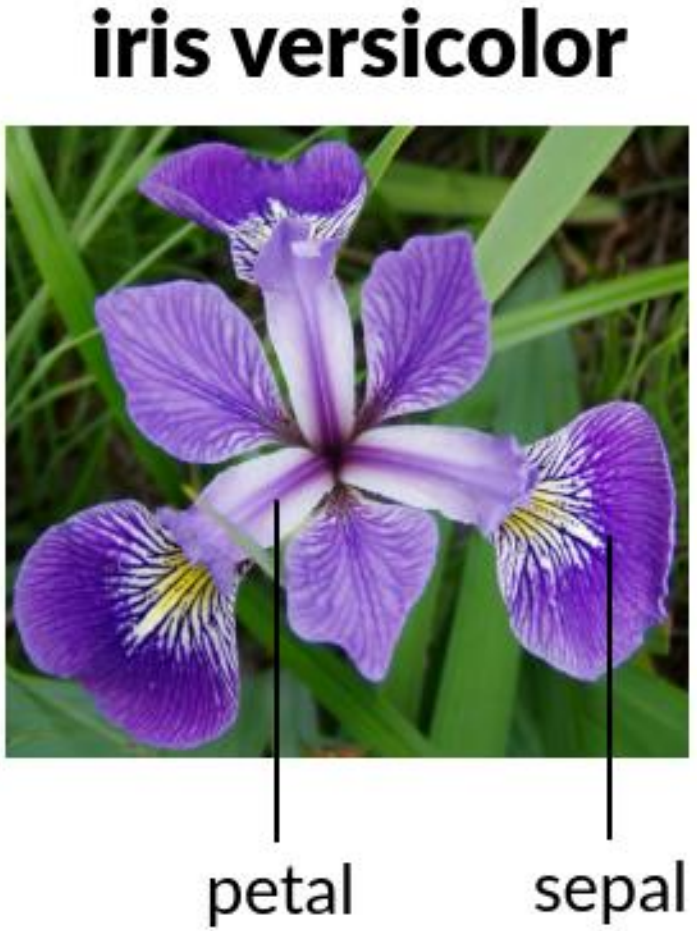
```
> ### Getting data from data frame ################
>
> # by index - like in matrix
> dataset1[3,2] # 3rd row, 2nd column
[1] "Sue"
>
> # by colum names
> dataset1["name"] # the whole name vector
  name
1 Anna
2  Tom
3  Sue
> dataset1[, "name"] # alternative notation
[1] "Anna" "Tom"  "Sue"
> dataset1$name # convinient notation
[1] "Anna" "Tom"  "Sue"
>
> dataset1[3, "name"] # only name from the 3rd row
[1] "Sue"
>
> # by row names
> rownames(dataset1) <- c("girl", "boy", "teacher")
> dataset1
        column1 name column3
girl          1 Anna    TRUE
boy           2  Tom    TRUE
teacher       3  Sue   FALSE
> dataset1["teacher", "name"]
[1] "Sue"
```

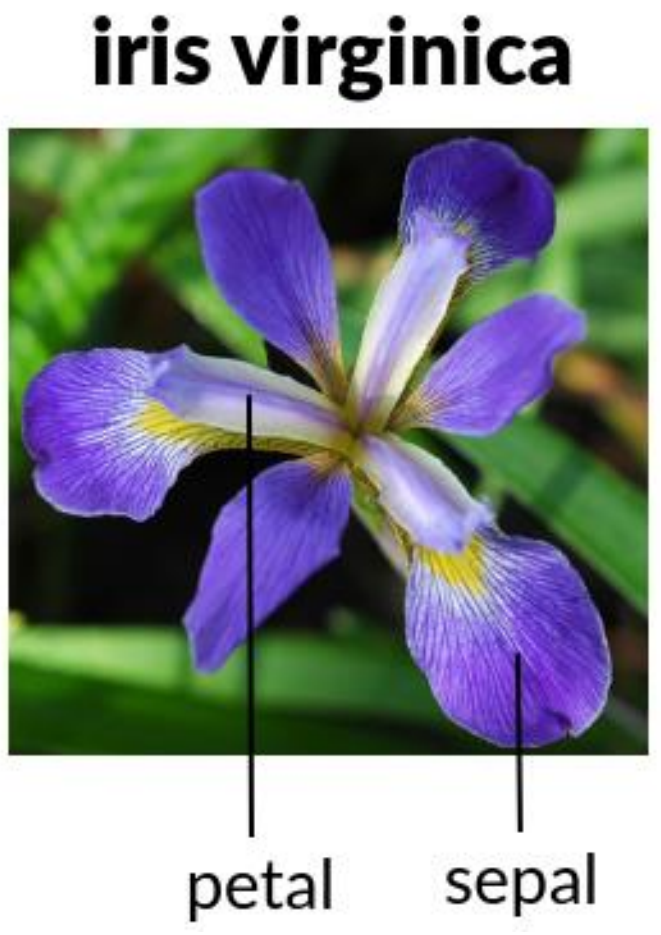
First possibility – using a vector of indexes like in matrices

Second possibility – choosing data by column names

The name vector needs to be compatible with the dimension of named data

Third possibility – using row names





BUILT-IN DATASETS

In R we have example datasets, which are helpful in testing new functions and operations before we will move to the empirical datasets.

Źródło: https://i.imgur.com/PQqYGaW.png

THE MOST IMPORTANT FUNCTIONS FOR DATA INSPECTION