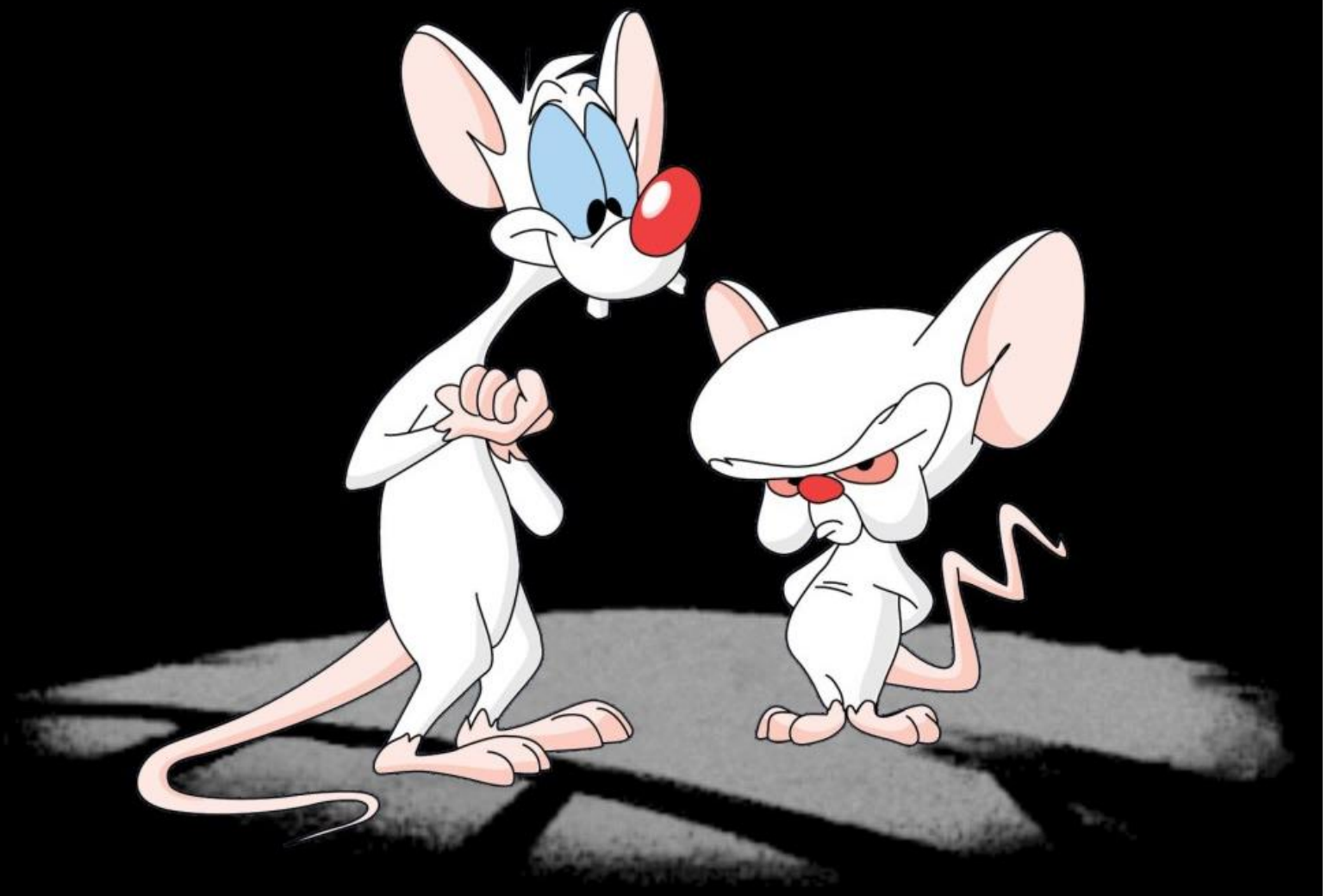
Tell me which packages you use and I'll tell you who you are

tidyverse vs data.table vs base R

Anna Skrzydło & Bartosz Kowalski

WhyR? 2018

04 | 07 | 2018



MEDIACOM
BUSINESS SCIENCE
WARSAW

What happens when you use packages from 10th page of google results...

ywność koloru oznacza wielkość inwestycji me

Error: could not find function "ggplotify"

If you know only base R, tidyverse might surprise you...

tourism_processing.R

```r
library(tidyverse)
library(MBSWarsawEconometricTools)

# Read dictionaries
country.dict.df <- read_csv2('02_Working/country_of_residence_dictionary.csv')
type.dict.df <- read_csv2('02_Working/type_dictionary.csv')

# Prepare monthly data
tourism.df <- read_csv('01_Input/DCSC_TUR_28062018103715289.csv') %>%
    filter(`Country of residence of guests` %in% c('Foreign countries', 'Italy', 'All countries
    filter(TIPO_ALLOGGIO2 == 'ALL') %>%
    select(Indicators, `Country of residence of guests`, TIME, Value) %>%
    mutate(`Code var type` = 'TU') %>%
    left_join(country.dict.df, by = "Country of residence of guests") %>%
    left_join(type.dict.df, by = "Indicators") %>%
    mutate(Variable = paste(`Code var type`, `Code type`, `Code country`, sep = '_')) %>%
    select(TIME, Variable, Value) %>%
    spread(Variable, Value) %>%
    mutate(Start.date = as.Date(paste0(TIME, '-01'))) %>%
    mutate(End.date = as.Date(lead(Start.date) - 1)) %>%
    select(Start.date, End.date, everything(), -TIME)

tourism.df$End.date[nrow(tourism.df)] <- as.Date('2018-03-31')
```

tourism_processing.R | tourism_processing_in_base.R | tourism.monthly.df | current.var.df | tourism.df

```r
library(MBSWarsawEconometricTools)

# Read dictionaries
country.dict.df <- read.csv2('02_Working/country_of_residence_dictionary.csv', stringsAsFactors = FALSE)
type.dict.df <- read.csv2('02_Working/type_dictionary.csv', stringsAsFactors = FALSE)

tourism.raw.df <- read.csv('01_Input/DCSC_TUR_28062018103715289.csv', stringsAsFactors = FALSE)

tourism.df <- tourism.raw.df[tourism.raw.df$Country.of.residence.of.guests %in% c('Foreign countries', 'Italy', 'All countries of the world'), ]
tourism.df <- tourism.raw.df[tourism.raw.df$TIPO_ALLOGGIO2 == 'ALL', ]
tourism.df <- tourism.df[, c('Indicators', 'Country.of.residence.of.guests', 'TIME', 'Value')]
tourism.df$Code.var.type <- 'TU'
tourism.df <- merge(tourism.df, country.dict.df, by = "Country.of.residence.of.guests")
tourism.df <- merge(tourism.df, type.dict.df, by = "Indicators")
tourism.df$Variable <-  paste(tourism.df$Code.var.type, tourism.df$Code.type, tourism.df$Code.country, sep = '_')
tourism.df <- tourism.df[, c('TIME', 'Variable', 'Value')]

tourism.monthly.df <- data.frame(TIME = unique(tourism.df$TIME))
tourism.monthly.df <- tourism.monthly.df[order(tourism.monthly.df$TIME), , drop = FALSE]
tourism.monthly.df$TIME <- as.character(tourism.monthly.df$TIME)

for (var in unique(tourism.df$Variable)) {
    current.var.df <- tourism.df[tourism.df$Variable == var, c('TIME', 'Value')]
    colnames(current.var.df) <- c('TIME', var)
    tourism.monthly.df <- merge(tourism.monthly.df, current.var.df, by = 'TIME')
}

tourism.monthly.df$Start.date <- as.Date(paste0(tourism.monthly.df$TIME, '-01'))
tourism.monthly.df$End.date <- c(as.Date(tourism.monthly.df$Start.date[2:nrow(tourism.monthly.df)]) - 1, as.Date('2018-03-31'))

tourism.monthly.df <- tourism.monthly.df[, c('Start.date', 'End.date', setdiff(colnames(tourism.monthly.df), c('Start.date', 'End.date', 'TIME')))]
```

Tidyverse is faster than base R

Reading a data.frame with 468 observations and 5 575 variables

```
library(microbenchmark)
file <- "speed_case/Sales_Nielsen.csv”

microbenchmark(base=read.csv2(file),
               tidy=suppressMessages(readr::read_csv2(file)), times = 25L)
```

```
Unit: seconds
 expr     min      lq    mean  median      uq     max neval
 base    5.06    5.26    5.56    5.47    5.60    7.29    25
 tidy    2.25    2.65    2.99    3.07    3.23    4.10    25
```

If speed is a criterion...
Go for data.table!

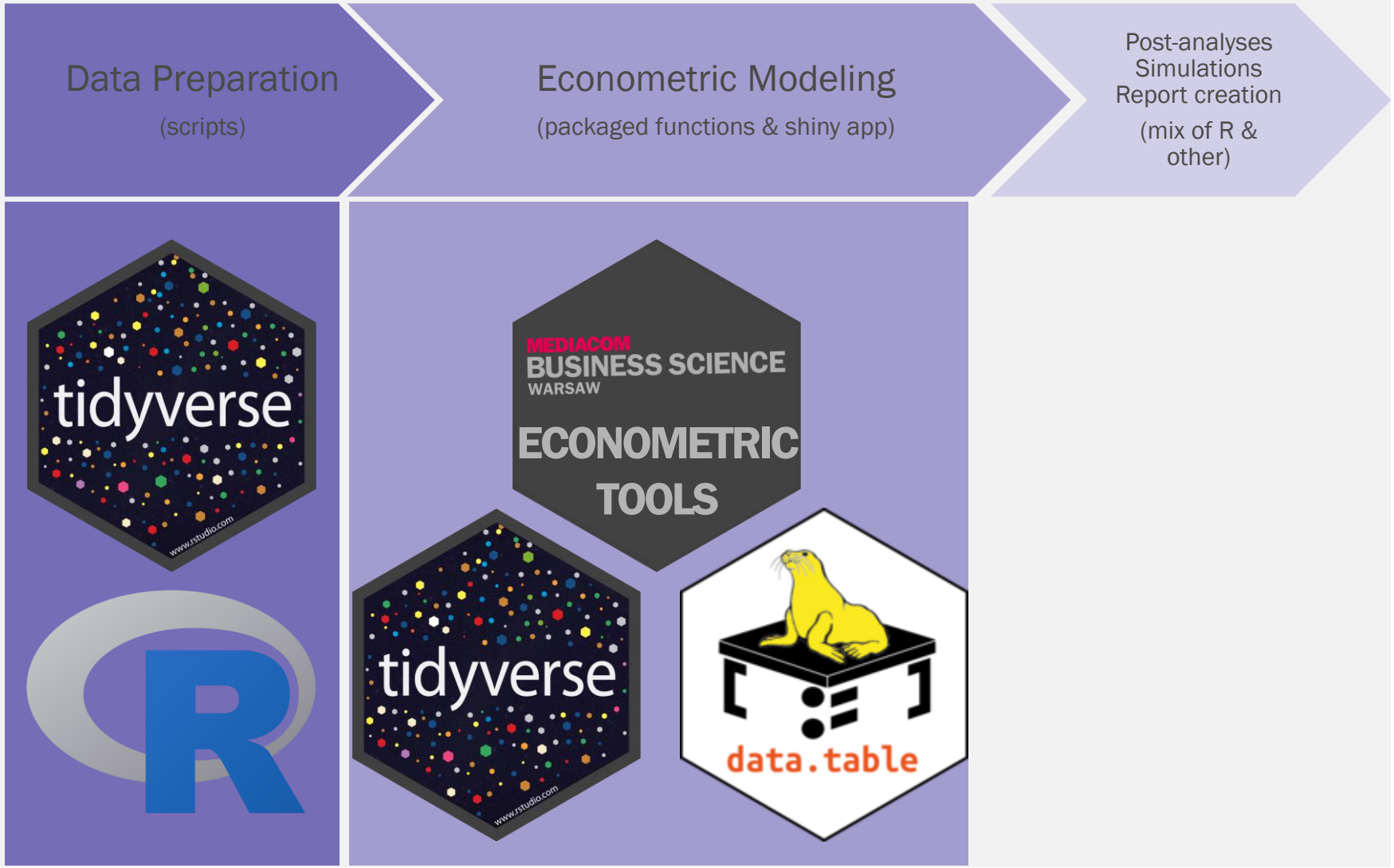
Reading a data.frame with 468 observations and 5 575 variables

```
microbenchmark(base=read.csv2(file),
               tidy=suppressMessages(readr::read_csv2(file)),
               dt=data.table::fread(file, sep = ";", dec = ",", showProgress=FALSE), times = 25L)
```

```
Unit: milliseconds
 expr      min       lq     mean   median       uq      max  neval
 base     5049     5175     5487     5315     5569     6705     25
 tidy     2324     2460     2894     2795     3050     4035     25
   dt      181      186      228      196      269      393     25
```

Where we are today

R framework for Markeing Mix Modeling at MediaCom Business Science



Our take on base vs Tidyverse vs data.table

	+	-
R	▪ Everyone in the community knows it ▪ Stable, no sudden changes to functions structure	▪ Outdated default arguments ▪ Slow to run
tidyverse	▪ Structures and orders thinking by providing a limited number of options (functions) ▪ When combined with meaningful naming convention for R objects, the code is extremely easy to undestand ▪ Quick to learn by a non-coder	▪ The syntax is quite verbose. Thus, it takes longer to run ▪ Even though speed is getting better over time, it still underperforms in some cases, even on smaller datasets vs data.table. This includes reading data (read_csv vs fread)
data.table	▪ It's FAST* ▪ Memory efficiency ▪ Compact syntax;	▪ The syntax is not self-explaining ▪ We find data.table more timely to learn

* Source: grouping operations benchmarking: dplyr vs data.table vs pandas: https://github.com/Rdatatable/data.table/wiki/Benchmarks-%3A-Grouping

MEDIACOM
BUSINESS SCIENCE
WARSAW

@MBSWarsaw
anna.skrzydlo@mediacom.com
bartosz.kowalski@mediacom.com

An investment in knowledge
Always pays the best interest

Benjamin Franklin