OFICINA CIS-USP

INTRODUÇÃO À PROGRAMAÇÃO E WEB SCRAPING

INSTRUTOR: NATHAN GIACHETTA (NECI-USP)

DATAS: 19, 21 E 28 DE SETEMBRO DE 2017

O QUE VAMOS APRENDER...

Objetivo: Capacitar pesquisadoras(es) do CIS em coletar dados provenientes da internet por meio de Web Scraping.

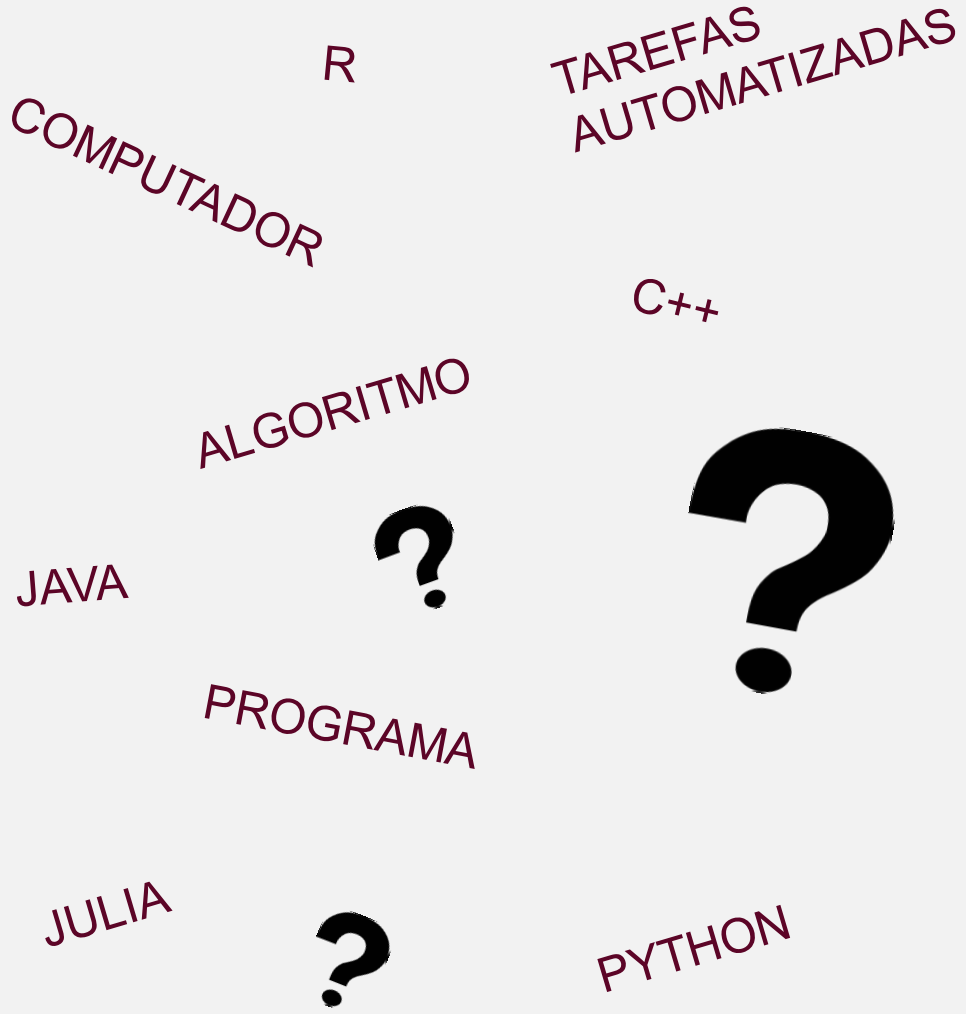
- Introdução à programação → **AULA 1**
- Conhecendo alguns pacotes: *magrittr, stringr* e *purrr*
- Introdução à estrutura HTML

→ **AULA 2**

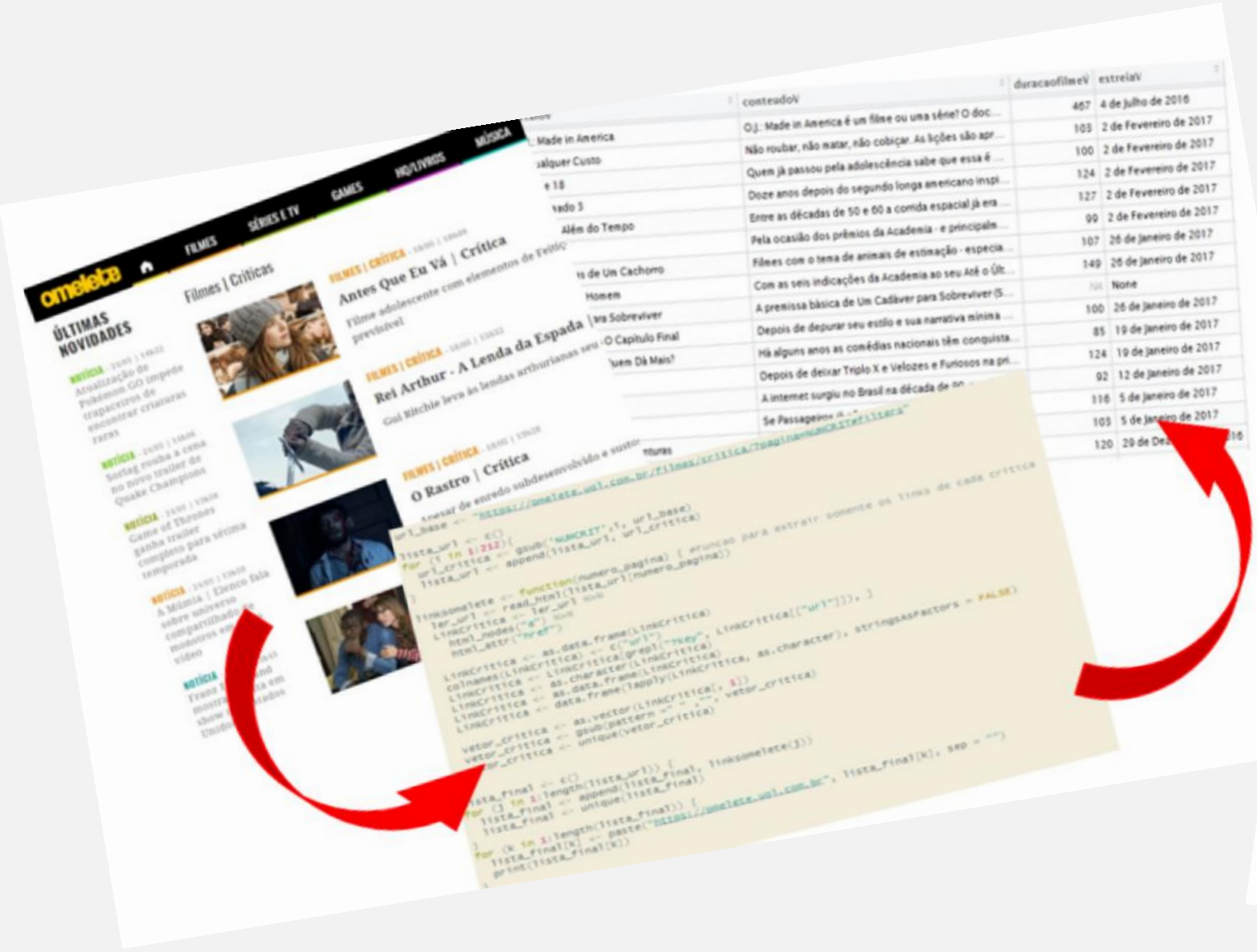
- Web Scraping com *rvest* → **AULA 3**

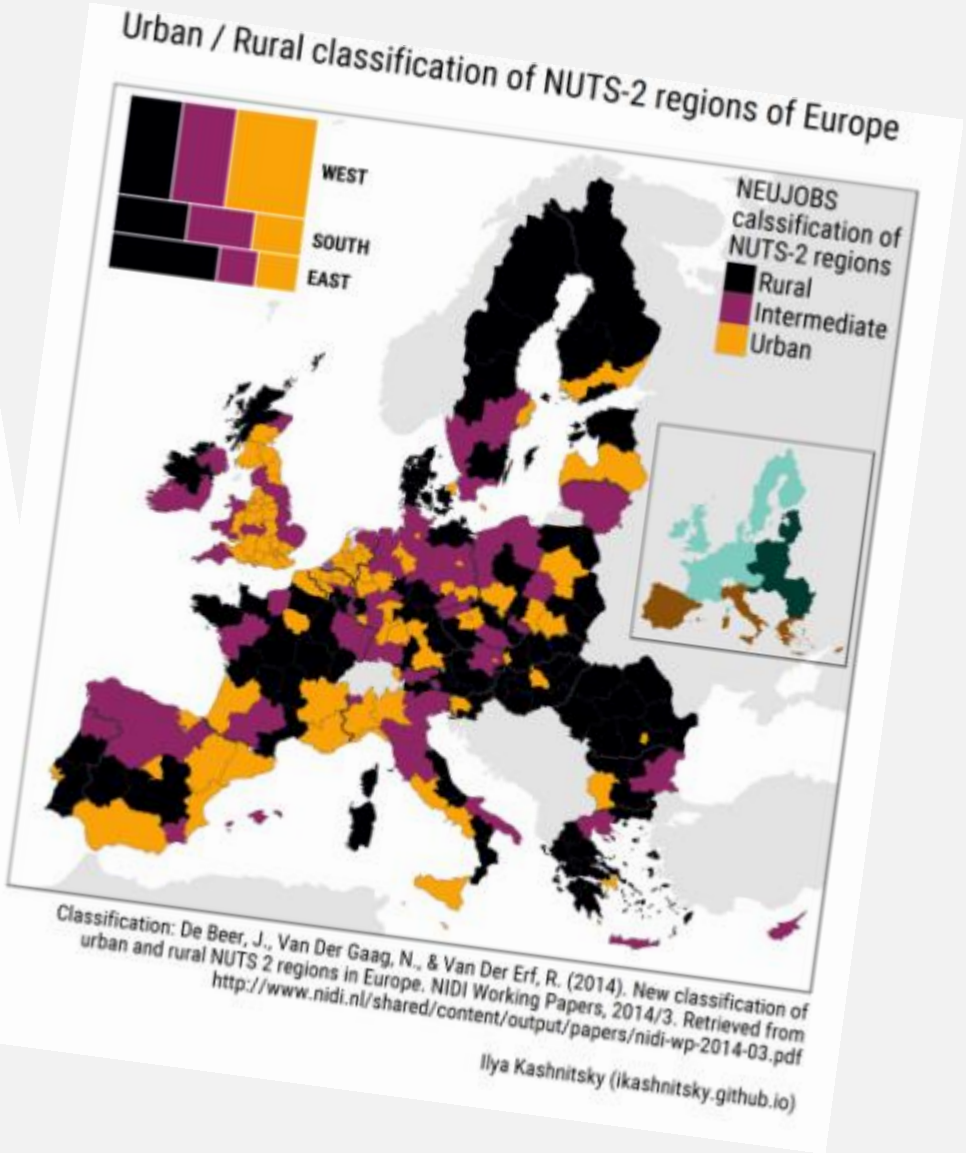
O QUE É PROGRAMAR?

```
#Extraindo uma lista com todas as críticas do omelete
url_base <- "https://omelete.uol.com.br/filmes/critica/?pagina=NUMCRIT#filters"

lista_url <- c()
for (i in 1:212){
  url_critica <- gsub('NUMCRIT',i, url_base)
  lista_url <- append(lista_url, url_critica)
}

linksomelete <- function(numero_pagina) { #Funcao para extrair somente os links de cada critica
  ler_url <- read_html(lista_url[numero_pagina])
  LinkCritica <- ler_url %>%
    html_nodes("a") %>%
    html_attr("href")

  LinkCritica <- as.data.frame(LinkCritica)
  colnames(LinkCritica) <- c("url")
  LinkCritica <- LinkCritica[grepl("?key", LinkCritica[["url"]]), ]
  LinkCritica <- as.character(LinkCritica)
  LinkCritica <- as.data.frame(LinkCritica)
  LinkCritica <- data.frame(lapply(LinkCritica, as.character), stringsAsFactors = FALSE)

  vetor_critica <- as.vector(LinkCritica[, 1])
  vetor_critica <- gsub(pattern =" " ,"", vetor_critica)
  vetor_critica <- unique(vetor_critica)
}

lista_final <- c()
for (j in 1:length(lista_url)) {
  lista_final <- append(lista_final, linksomelete(j))
```

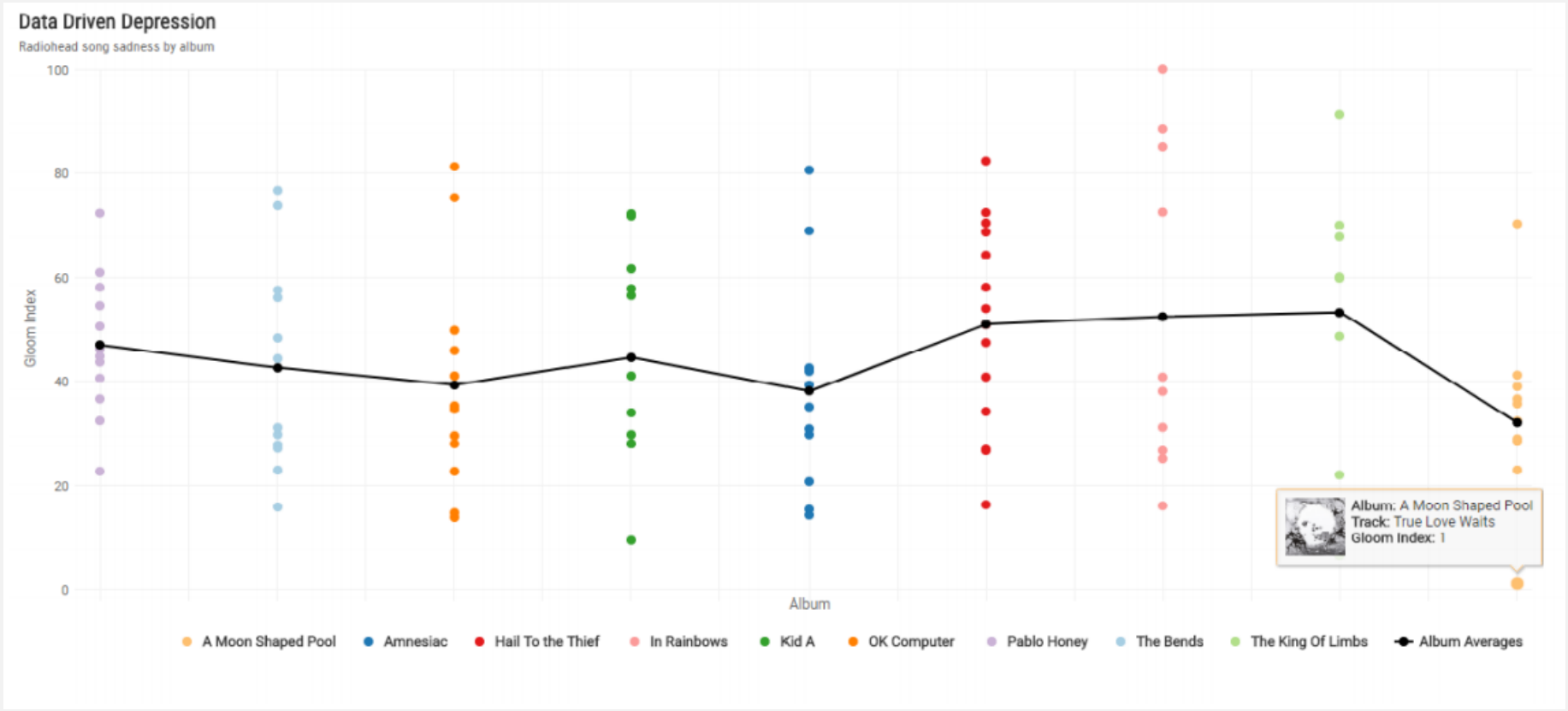


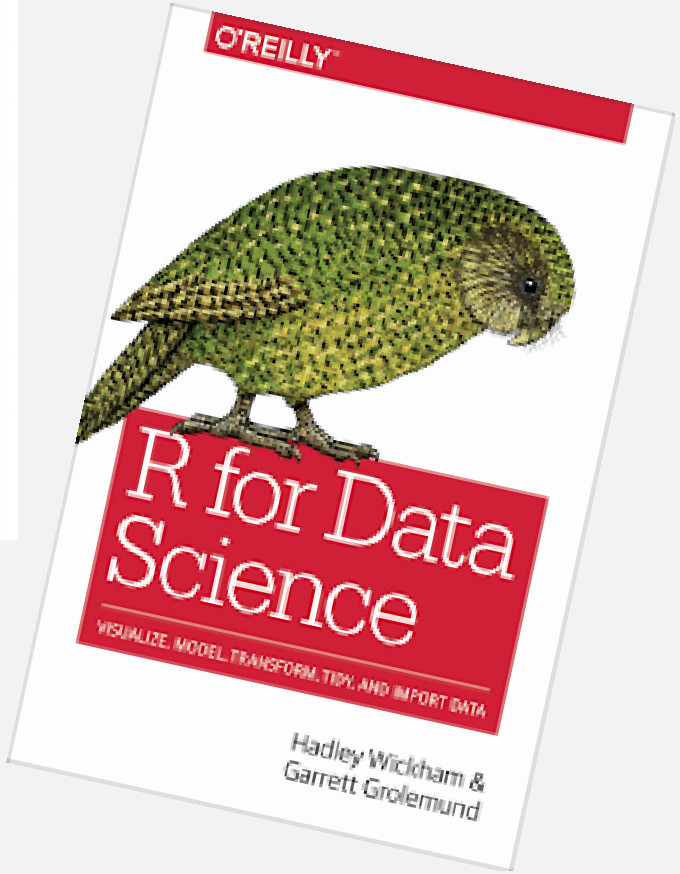
LINGUAGEM R? O QUE PODEMOS FAZER?





LINGUAGEM R? O QUE PODEMOS FAZER?





VAMOS PROGRAMAR!