R 語言
網路爬蟲

王貿

國立臺灣大學行為與資料科學研究中心助理研究員

國立臺灣大學政治學系博士、兼任講師

maowang01@gmail.com

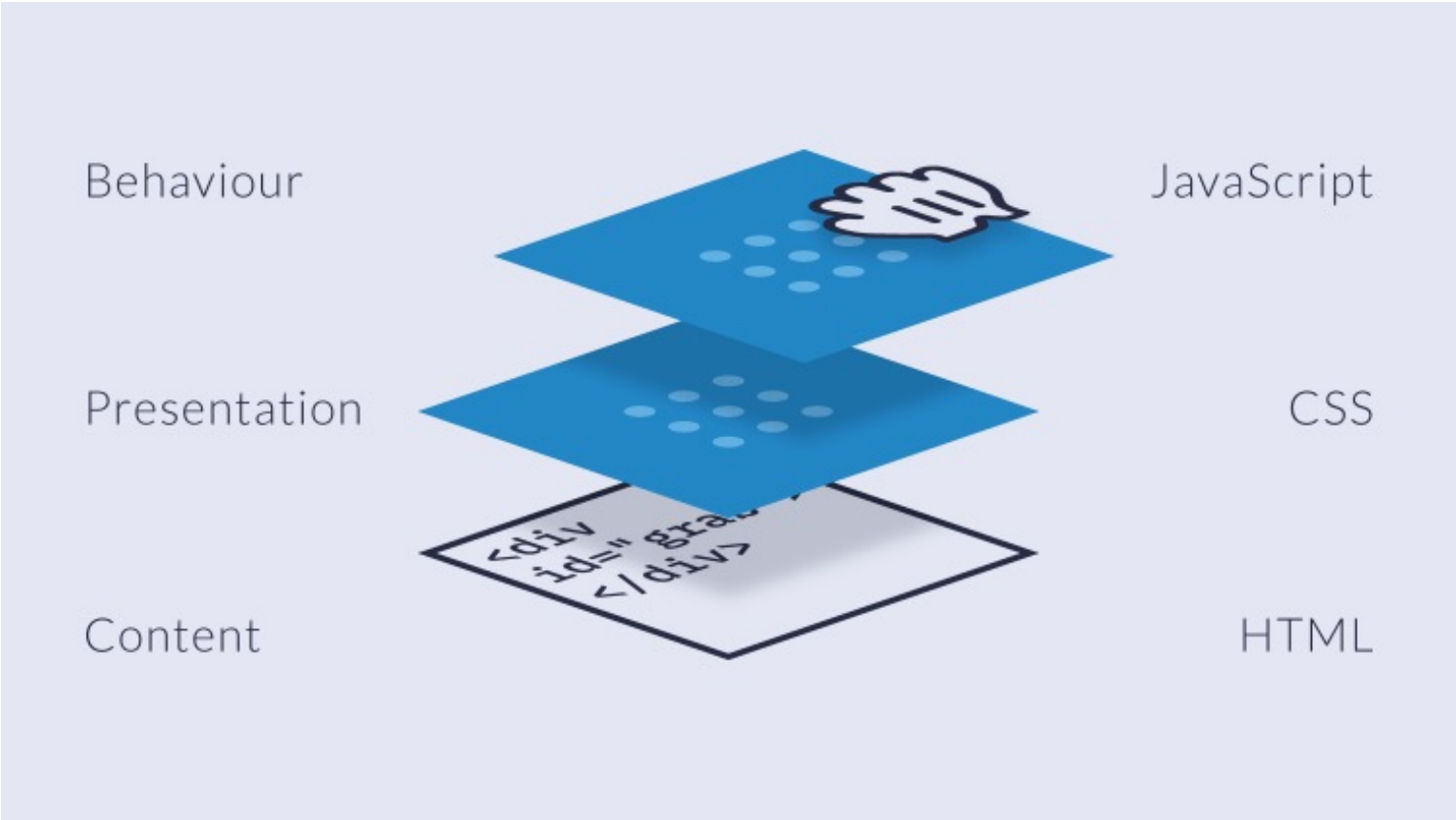
課程主題重點

- 基礎爬蟲
- 網頁基本架構（html、CSS、xpath）
- 多網址爬蟲
- 需認證網頁爬取

rvest

- Simple web scraping for R
  （ https://github.com/tidyverse/rvest ）
- 屬於 tidyverse 內的套件。

網頁基本架構



https://blog.rsquaredacademy.com/web-scraping/

基本爬蟲步驟

- read_html(path) %>% #網址

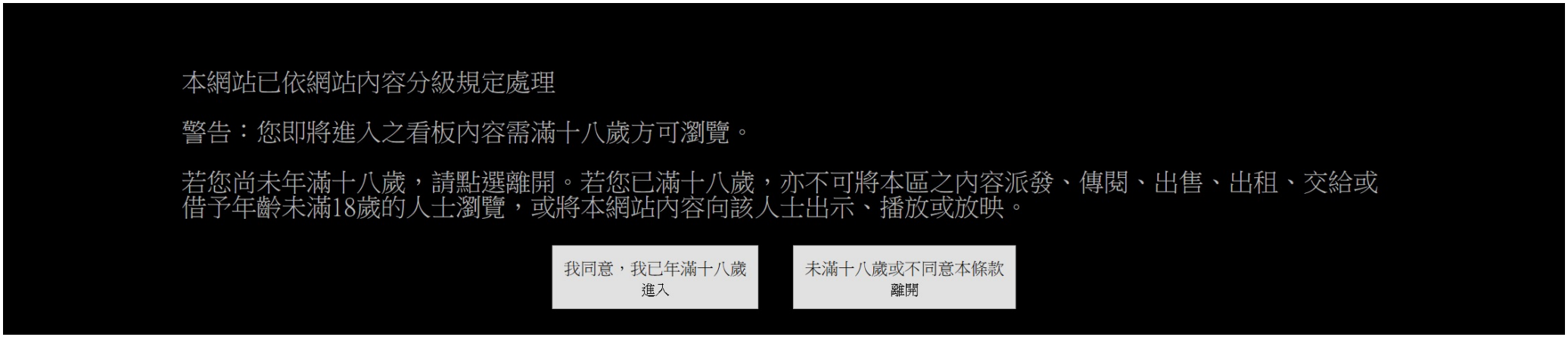
  html_nodes(css = “.before”) %>% #定位點

  html_text() #擷取什麼類型資料

SelectorGadget

- 協助取得css、xpath定位點：SelectorGadget

需認證網站



如果需要先「點選年滿18歲」，這樣的網站要怎麼爬取呢？