

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

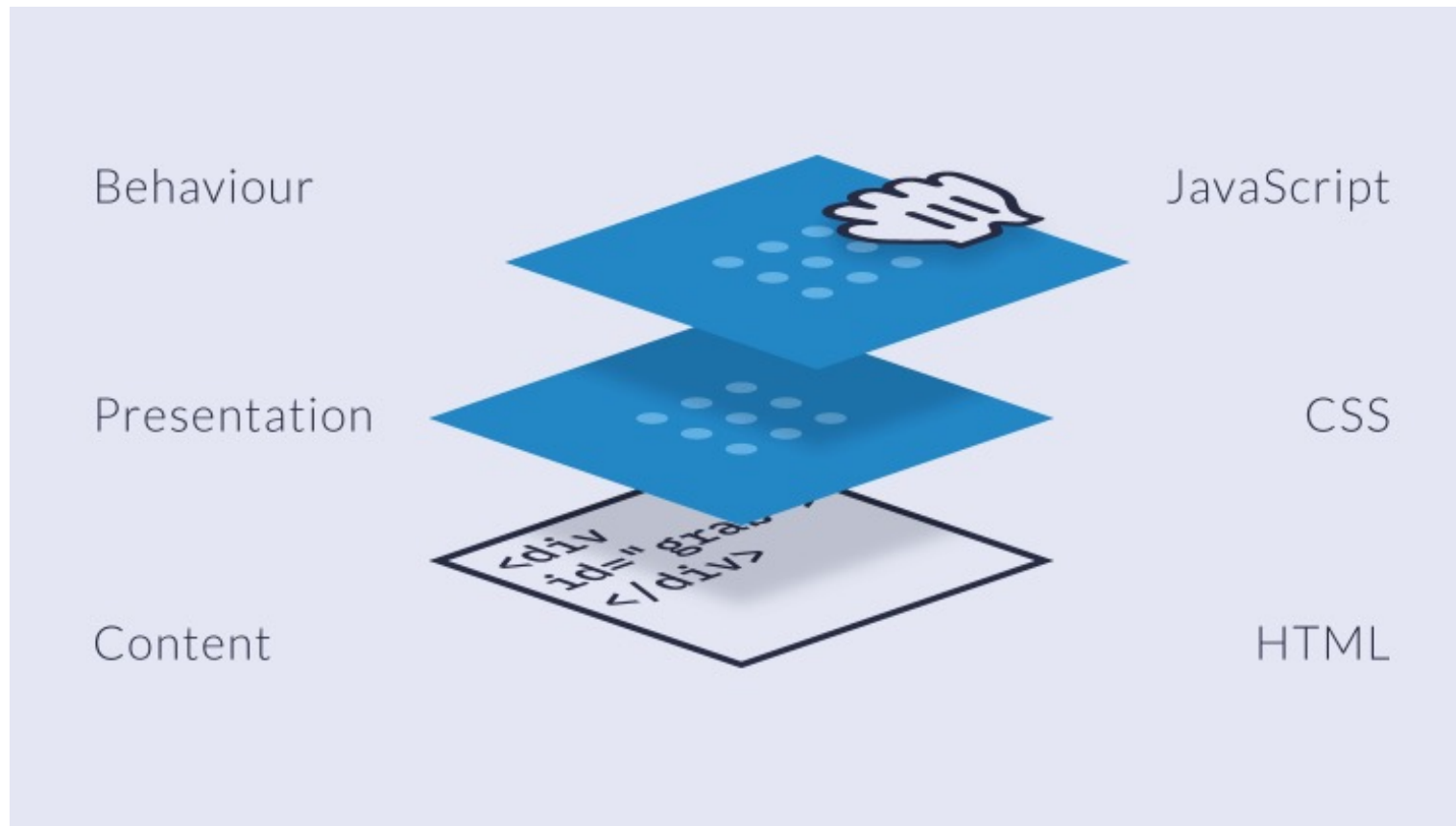
- 基礎爬蟲
- 網頁基本架構 (html 、 CSS 、 xpath)
- 多網址爬蟲
- 需認證網頁爬取

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

本網站已依網站內容分級規定處理

警告：您即將進入之看板內容需滿十八歲方可瀏覽。

若您尚未年滿十八歲，請點選離開。若您已滿十八歲，亦不可將本區之內容派發、傳閱、出售、出租、交給或借予年齡未滿18歲的人士瀏覽，或將本網站內容向該人士出示、播放或放映。

我同意，我已年滿十八歲
進入

未滿十八歲或不同意本條款
離開

如果需要先「點選年滿18歲」，這樣的網站要怎麼爬取呢？