Mengubah Web Menjadi Data

(Turning web into Data)

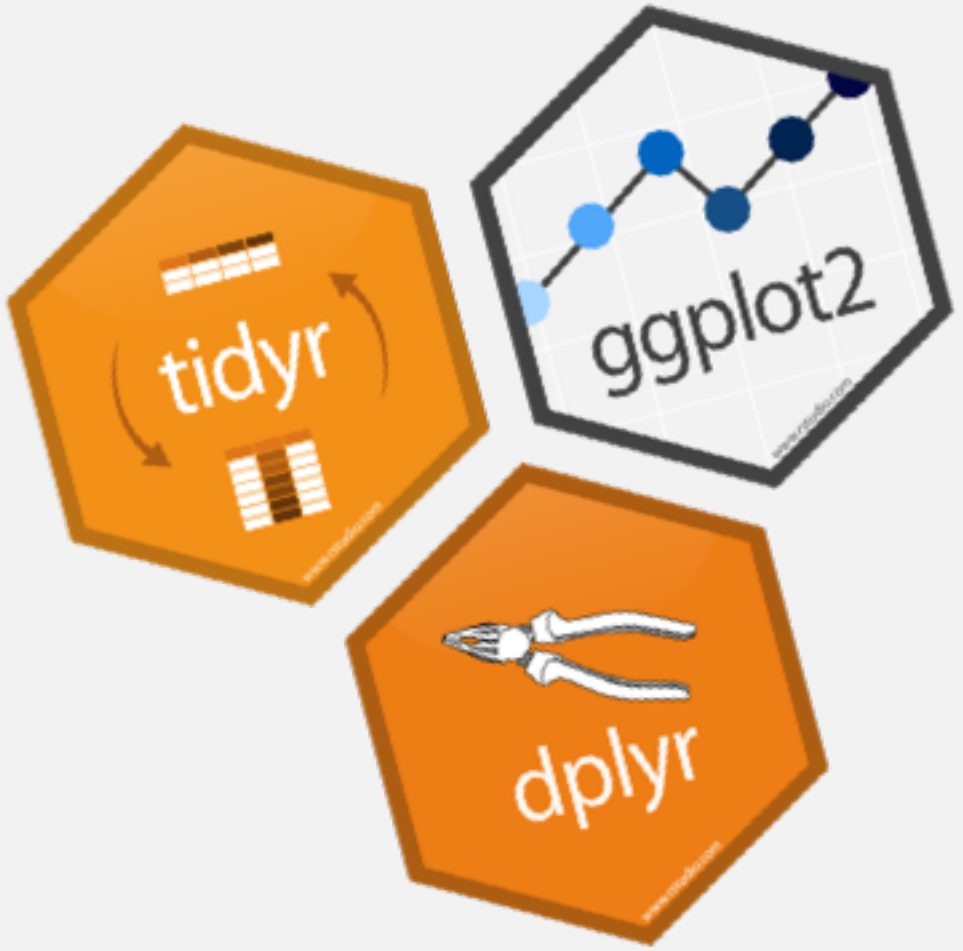
By: Ujang Fahmi

11 April 2018

sagasitas | Data Scrapping Workshop organized by MKP

Tujuan

- Mengenal Bahasa pemrograman R (skip jika sudah)
- Tidyverse
- Scrapping media sosial dengan API (Twitter)
- Scrapping static web page
- Scrapping *multipages* (jika memungkinkan waktunya)



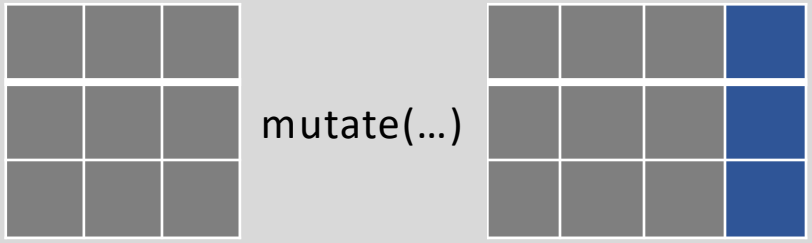
Intro

- R adalah Bahasa pemrograman statistik
- R dan R Studio dua hal yang berbeda tapi punya fungsi yang mirip
- Pengguna R didukung oleh keberadaan packages yang berisi fungsi, algoritme, dan terkadang kombinasi algoriteme
- Tidyverse sebagai pedoman

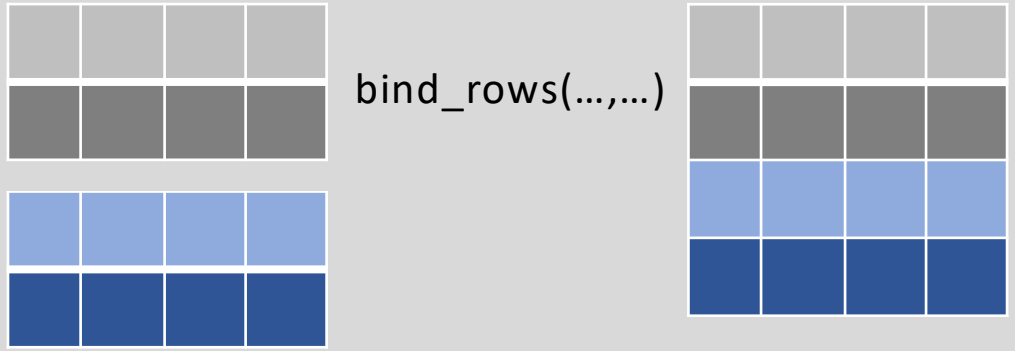
Code	Read
#	Comment
= or <-	Equals to
%>%	Pipe or then
"..."	The object is string
[... , ...]	Rows , column
$ (e.g. tweets$retweet)	Part of (e.g. read column retweet from data with name tweet)
str()	Data structure
?help	Get some help from documentation
summary()	Summary of data

Data Wrangling

Add new column

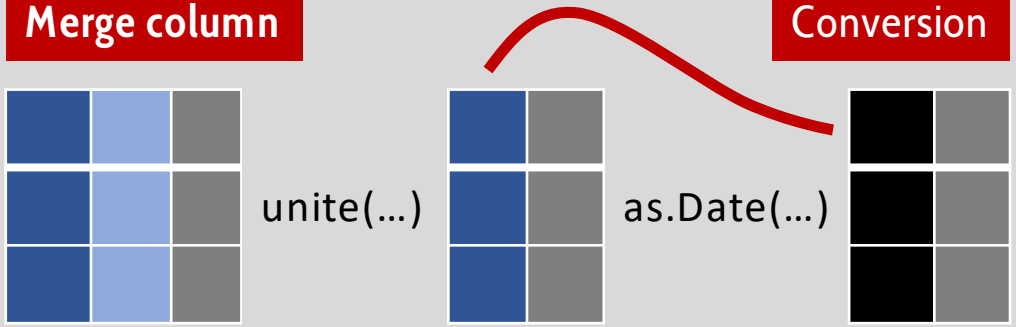


Combine rows from two data frame

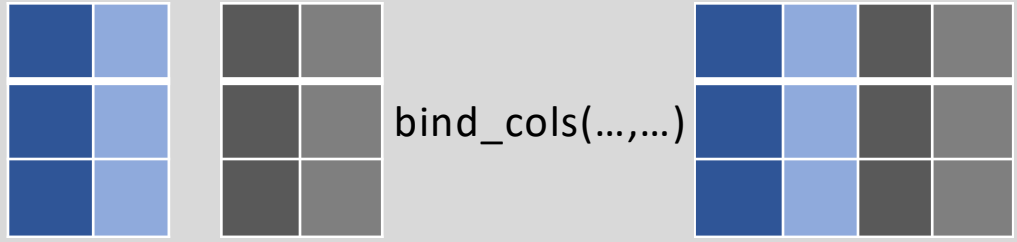


Merge column

Conversion



Combine column from two data frame



Scrapping dengan API

Package yang bisa digunaka

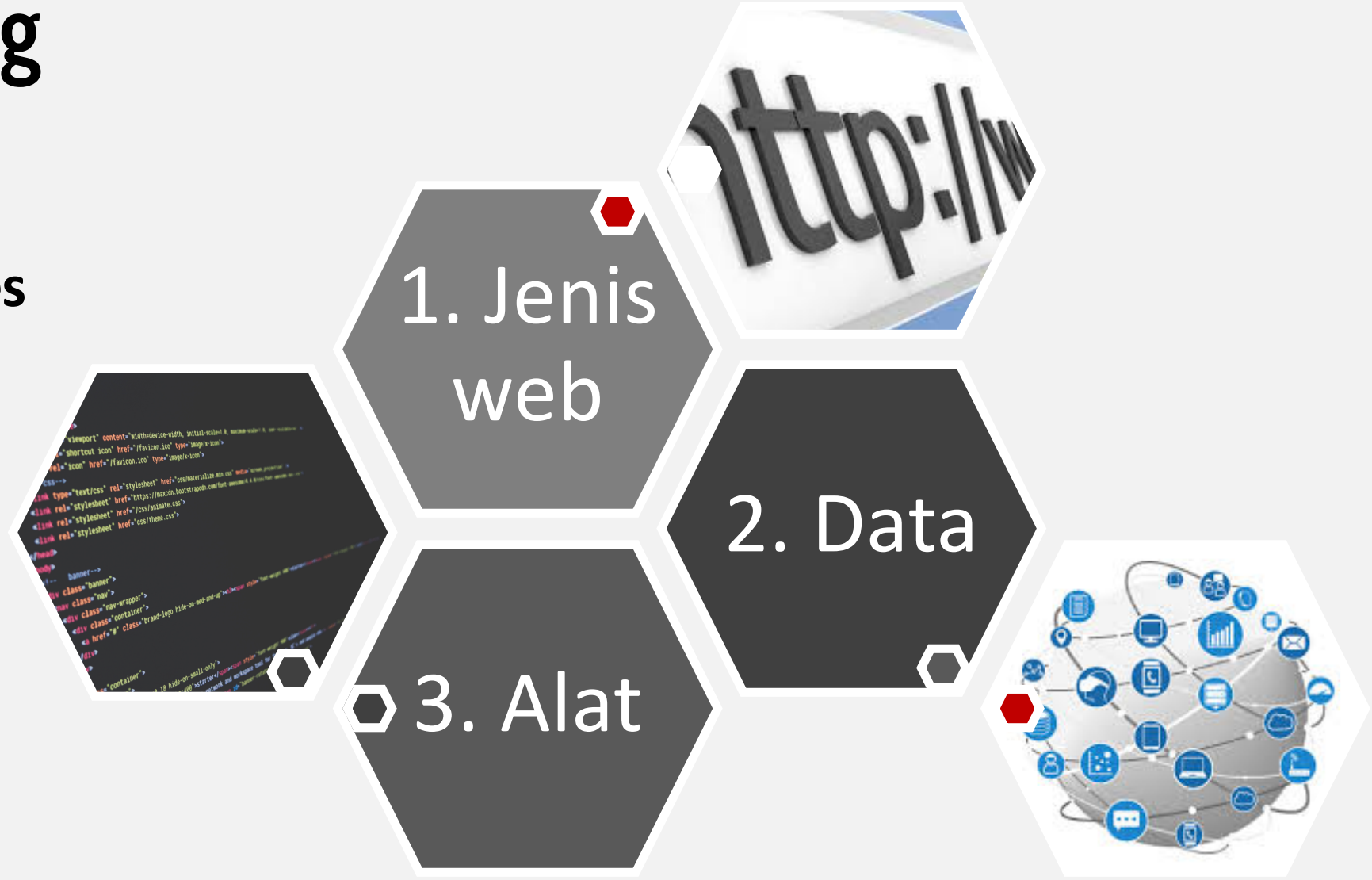
Nama Package	Keterangan	Console
twitter Package	Digunakan untuk mendapatkan data dari twitter dengan focus pada kontentnya	Install.package('twitter') library(twitter)
Social media lab	Digunakan untuk mendapatkan data dari twitter dengan focus interaksi antar akun (cocok untuk SNA scara instan)	Install.package('SocialMediaLab') (SocialMediaLab)

Mendaptakan API Twitter

- API yang dapat dimanfaatkan secara gratis adalah api basic. Dengan jenis API tersebut kita bisa mendapatkan data maksimal ± 50.000 twit yang dikirim pada 7 hari terakhir dihitung pada saat pengambilan data kebelakang.
- Untuk membuat API Twitter, kita harus memiliki akun Twitter dan mengikuti petunjuk di sini atau https://apps.twitter.com/app/new
- Hal yang bisa di dapat: Id, username, jumlah retweet, favorite, reply, twit

Web Scrapping

- Menentukan Web
- Menemukan nodes yang konsisten
- Atribut data (text/links)
- Memahami urls (multiplepage)
- Eksperimen



LATIHAN

Latihan 1: Menggunakan html

- Script: Menambang Data Web 2.R
- Menginstall packages yang dibutuhkan
- Loading packages
- Dokumentasi nodes
- Eksperimen 1
- Eksperimen 2
- Eksperimen 3
- Wrangling & Cleaning

Rvest

Fungsi	Keterangan
html_nodes(...)	Nodes
html_text(...)	Atribut teks
html_attrs(...)	Atribut num/int

Dplyr/tidyverse

Fungsi	Keterangan
bind_colss(...)	Combine columns
bind_rows(...)	Combine rows
%>%	Pipelining

Latihan 2: Menggunakan API

- Script: Menambang Data Web 3.R
- Membuat aplikasi API di twitter
- Install packages
- Loading packages
- Membuat script
- Running (scrapping)
- Wrangling data

twitteR

Fungsi	Keterangan
searchTwitter(…)	Mencari tweets
twListToDF(….)	Tweet to data frame
%>%	Pipelining

Rujukan untuk belajar mandiri

- **Buku:** Hadley Wickahm (Advance R/R data scientist) -> https://adv-r.hadley.nz/
- **Buku:** Julia Silge and David Robinson (Text Mining with R) -> https://www.tidytextmining.com/
- **Kursus:** Datacamp -> https://www.datacamp.com/
- **Q&A:** Forum -> https://stackoverflow.com/questions/tagged/r
- **Blog Penelitian:** Oxford internet institute -> https://blogs.oii.ox.ac.uk/policy/