Got R? Catch 'em all!

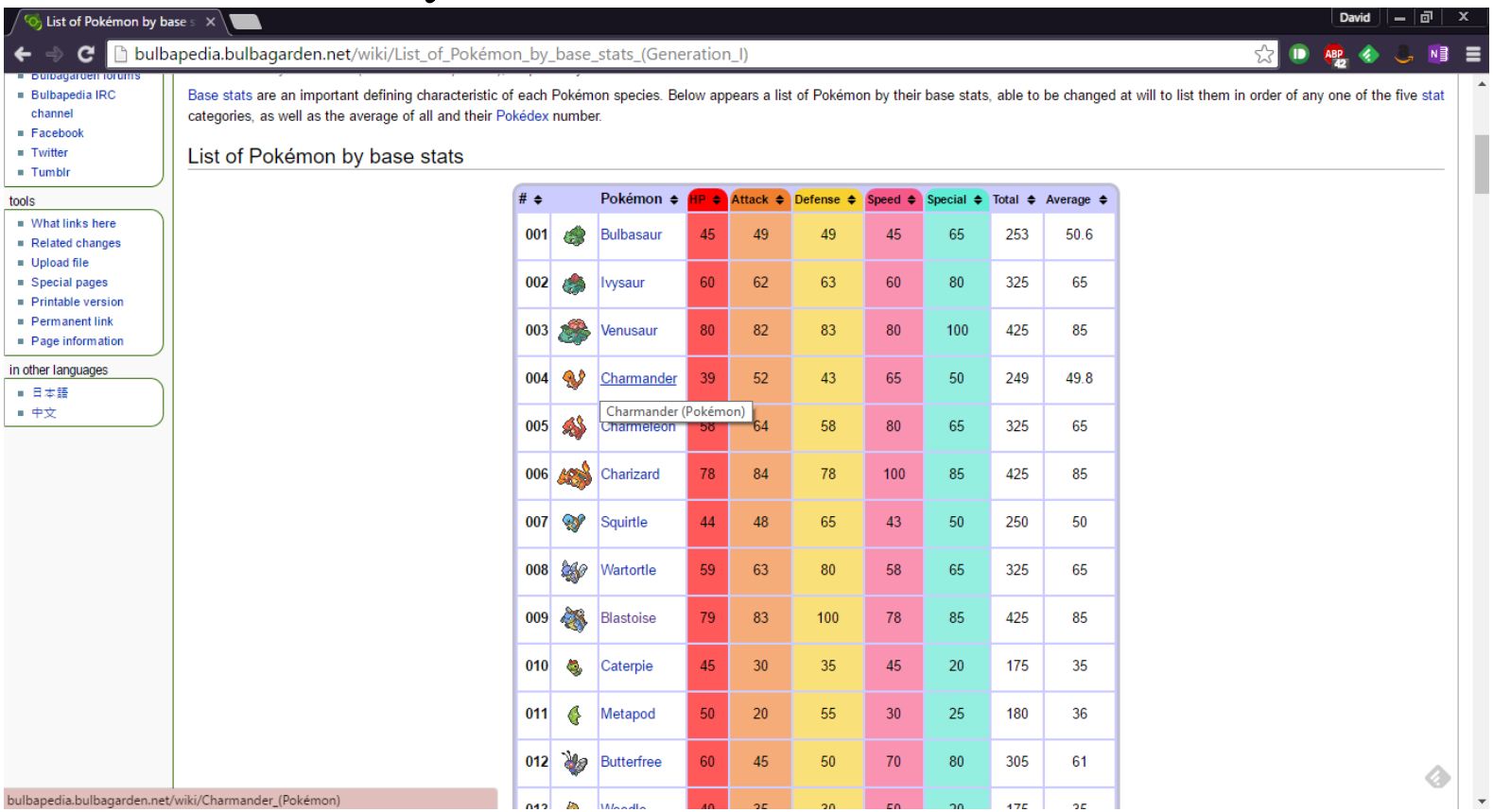
3 steps

- We're gonna' do 3 different web scraping tasks in 5 minutes from a single site

1. Scrape a table of original 151 pokemon stats from one webpage
2. Scrape 151 images of pokemon from seperate 151 webpages
3. Build a plot that scrapes the .pngs of each pokemon from 151 webpages **by itself**

Pluck the data from Bulbapedia

- Bulbapedia is a community website with tonnes of info on every pokemon
- `http://bulbapedia.bulbagarden.net/wiki/List_of_Pok%C3%A9mon_by_base_stats_(Generation_I)`



#	Pokémon	HP	Attack	Defense	Speed	Special	Total	Average
001	Bulbasaur	45	49	49	45	65	253	50.6
002	Ivysaur	60	62	63	60	80	325	65
003	Venusaur	80	82	83	80	100	425	85
004	Charmander	39	52	43	65	50	249	49.8
005	Charmeleon	58	64	58	80	65	325	65
006	Charizard	78	84	78	100	85	425	85
007	Squirtle	44	48	65	43	50	250	50
008	Wartortle	59	63	80	58	65	325	65
009	Blastoise	79	83	100	78	85	425	85
010	Caterpie	45	30	35	45	20	175	35
011	Metapod	50	20	55	30	25	180	36
012	Butterfree	60	45	50	70	80	305	61

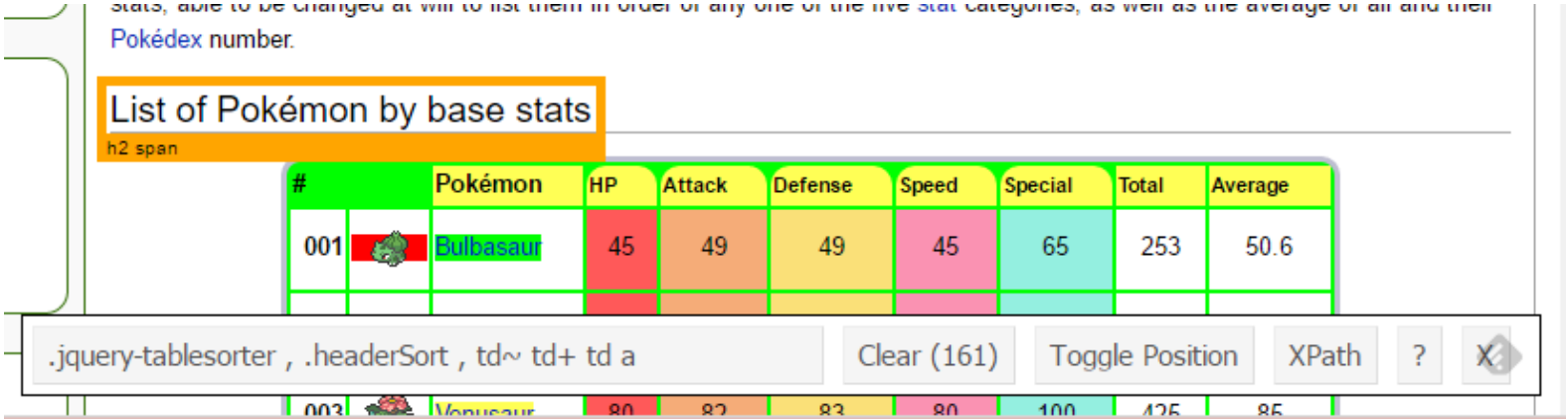
PokedexR

```
library(rvest)
library(magrittr)

bulbagarden <- "http://bulbapedia.bulbagarden.net/wiki/List
baseStats <-
  # Read HTML or XML
  xml2::read_html(
    x = bulbagarden
    ) %>% # This is a pipe operator from magrittr
  # Extract pieces out of HTML using css selectors
  rvest::html_node(
    # rvest recommends using 'Selector Gadget'
    css = "div table"
    ) %>%
  # Parse an html table into a data frame
  rvest::html_table()
```

SelectorGadget (like a silph scope but for web pages)

- `rvest` recommends SelectorGadget is a chrome extension for CSS selector generation.
- It "Makes the Invisible Plain to See!" by exposing which parts of the html correspond to which bits of the user facing webpage.



- However, I had to fudge it a bit, as it didn't pick up the table properly.

All the data

```
head(baseStats)
```

```
##   #      Pokémon HP Attack Defense Speed Special Total Avera
## 1 1 NA  Bulbasaur 45     49      49    45      65   253    50.6
## 2 2 NA    Ivysaur 60     62      63    60      80   325    65.0
## 3 3 NA   Venusaur 80     82      83    80     100   425    85.0
## 4 4 NA Charmander 39     52      43    65      50   249    49.8
## 5 5 NA Charmeleon 58     64      58    80      65   325    65.0
## 6 6 NA  Charizard 78     84      78   100      85   425    85.0
```

Muky Data

```
library(data.table)

baseStats <- data.table::as.data.table(baseStats)

# Remove second col with only "NA"
baseStats[, "" := NULL]
# Rename cols 1 and 2 to something workable
data.table::setnames(baseStats,
                     1:2,
                     c("DexNo", "Pokemon"))
```

Master Ballin'

```
library(stringr)

baseStats[, imgURL := read_html(
    x = bulbagarden# set x to be bulbagarden url
    ) %>%
  rvest::html_nodes(
    # css to identify EVWERY string for pokemon image urls
    css = "#mw-content-text img"
    ) %>% # split string at point
  stringr::str_split_fixed(
    "src=\"",
    n = 2
    ) %>% # use second part of string
    .[,2] %>% # split string at point again
  stringr::str_split_fixed(
    "\" width=",
    n = 2) %>% # use first part of string
    .[,1]]
```

Congratulations, you caught all 151 image urls!

```
head(baseStats)
```

```
##   DexNo   Pokemon HP Attack Defense Speed Special Total Ave
## 1:     1 Bulbasaur 45     49      49    45      65   253   50.6
## 2:     2   Ivysaur 60     62      63    60      80   325   65.0
## 3:     3  Venusaur 80     82      83    80     100   425   85.0
## 4:     4 Charmander 39     52      43    65      50   249   49.8
## 5:     5 Charmeleon 58     64      58    80      65   325   65.0
## 6:     6  Charizard 78     84      78   100      85   425   85.0
##                                                          imgURL
## 1: http://cdn.bulbagarden.net/upload/e/ec/001MS.png
## 2: http://cdn.bulbagarden.net/upload/6/6b/002MS.png
## 3: http://cdn.bulbagarden.net/upload/d/df/003MS.png
## 4: http://cdn.bulbagarden.net/upload/b/bb/004MS.png
## 5: http://cdn.bulbagarden.net/upload/d/dc/005MS.png
## 6: http://cdn.bulbagarden.net/upload/0/01/006MS.png
```

Your Pokemon have been moved to “Someones PC”!

```
dir.create("./someonesPC/")

# Use for loop as data.table only
# returns last item if this is run within DT function
for(pkmn in baseStats[, DexNo]){
  download.file(baseStats[pkmn == DexNo,
                          imgURL],
                destfile = paste0("./someonesPC/",
                                  pkmn, ".png"),
                # Makes this work for windows
                mode = "wb")
}
```

DaveRGP checked “Someones PC”!

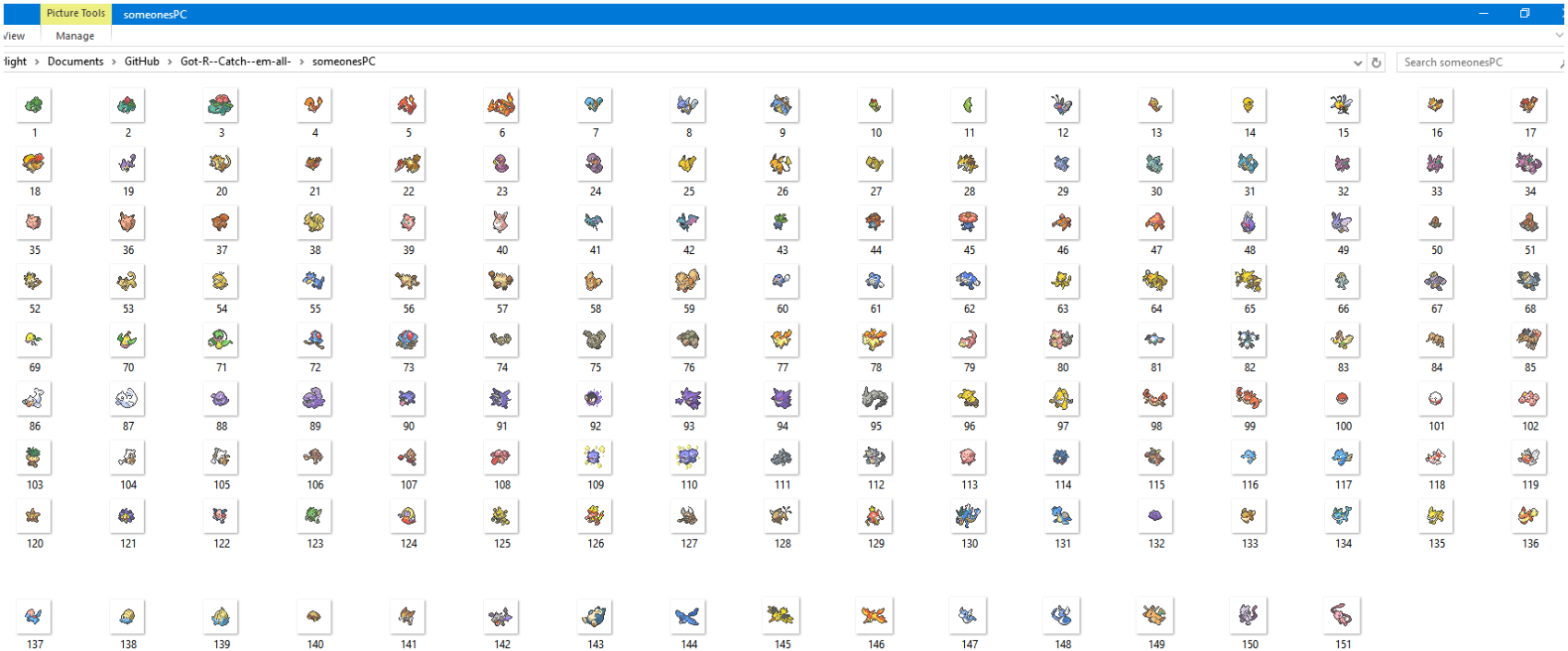


Figure 1:someones PC

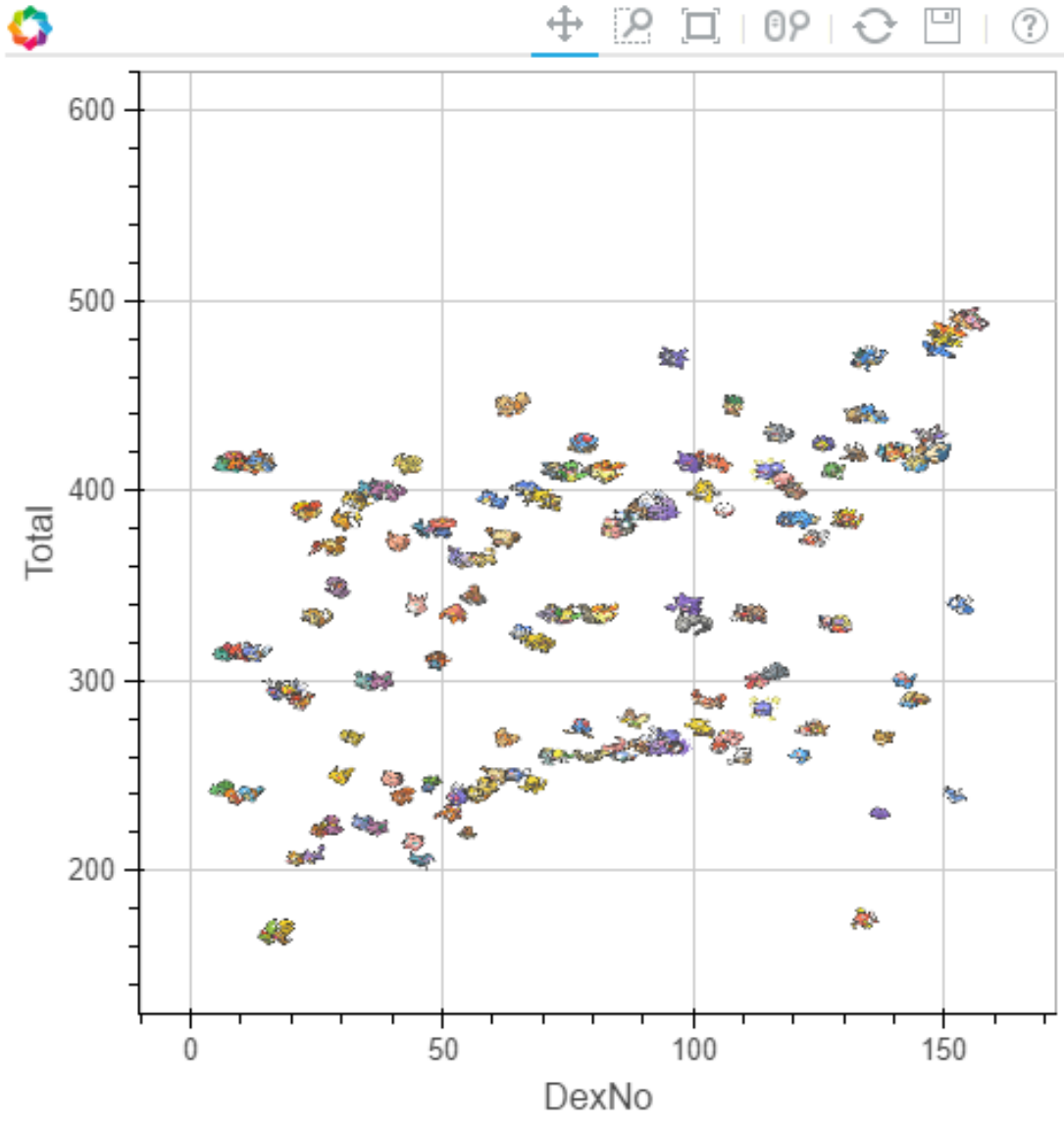
I'm gonna be the very best...

```
library(rbokeh)

P <- figure() %>%
  # layer to get .png from url as points
  ly_image_url(
    data = baseStats,
    x = DexNo,
    y = Total,
    w = 10,
    h = 20,
    image_url = imgURL
  )
```

...like no one ever was!

```
P
```



Trainer Tips

- Use `xml2::read_html()` to read the whole page into memory
- Use `rvest::html_node()` to find individual parts of the page
- Use `rvest::html_nodeS()` to return multiple items
- Return a data.frame with `rvest::html_table()`
- `data.table` can be finickity about making connections
- using for loops might solve that
- Use `stringr` for manipulating urls
- `rbokeh` is an interactive graphics package with an argument for using urls to source .png icons

Trainer Card

David Parr

github: DaveRGP

twitter: [@biomimicron](https://twitter.com/biomimicron)

website: davergp.github.io

Pokemon Field Studies: davergp.github.io/Pokemon_FieldStudies/