

Web Scraping for Sports Data with R

Yaqiong Yao

10/1/2020

Outline

- Introduction
- Using R
 - ▶ Import files downloaded from websites
 - ▶ Static data
 - ▶ Dynamic data
- Summary

Introduction

- Web scraping technique is used for capturing data from websites.
- Motivation of Web Scraping
 - ▶ Need to extract data from websites
 - ▶ A reproducible way of capturing data online
- Prerequisite
 - ▶ Having experience with R
 - ▶ A laptop with R and R studio installed

Example

College basketball school index

- These data can be obtained by copy and paste manually.
- Web scraping technique helps capture the data efficiently.

Sports Reference

Baseball

Football (college)

Basketball (college)

Hockey

Calcio

Blog

Stathead

Widgets

Create Account

Login

Questions or Comments?

Enter Person, Team, Section, etc

Search

Players

Schools

Seasons

Leaders

Scores

NCAA Tournaments

Play Index

Newsletter

Full Site Menu Below

School Index

Schools that were classified as a major school (i.e., Division I or equivalent) for at least one season.

480 Schools

SRS back to 1949-50

Share & more

Glossary

Rk	School	City, State	From	To	Yrs	G	W	L	W-L%	SRS	SOS	AP	CREG	CTRN	NCAA	FF	NC
1	Abilene Christian Wildcats	Abilene, Texas	1971	2020	10	294	148	146	.503	-10.99	-6.93	0	0	1	1	0	0
2	Air Force Falcons	USAF Academy, Colorado	1958	2020	62	1703	741	962	.435	-2.63	1.34	0	1	0	4	0	0
3	Akron Zips	Akron, Ohio	1902	2020	69	1593	942	651	.591	-0.24	-1.47	0	9	4	4	0	0
4	Alabama A&M Bulldogs	Normal, Alabama	2000	2020	21	610	232	378	.380	-16.99	-11.31	0	1	1	1	0	0
5	Alabama Crimson Tide	Tuscaloosa, Alabama	1913	2020	107	2756	1693	1062	.615	7.27	4.58	15	10	7	21	0	0
6	Alabama State Hornets	Montgomery, Alabama	1983	2020	38	1128	540	588	.479	-12.96	-10.02	0	4	4	4	0	0
7	Alabama-Birmingham Blazers	Birmingham, Alabama	1980	2020	41	1320	820	500	.621	6.37	2.62	2	7	5	15	0	0
8	Albany (NY) Great Danes	Albany, New York	2000	2020	21	658	326	332	.495	-6.75	-6.08	0	2	5	5	0	0
9	Alcorn State Braves	Alcorn State, Mississippi	1978	2020	43	1275	552	723	.433	-13.57	-8.87	0	10	6	6	0	0
10	Allegheny Gators	Meadville, Pennsylvania	1896	1916	21	234	191	41	.823			0	0	0			
11	American Eagles	Washington, D.C.	1967	2020	54	1533	755	778	.492	-5.55	-3.79	0	7	3	3	0	0
12	Amherst Lord Jeffs	Amherst, Massachusetts	1901	1902	2	12	12	0	1.000			0	0	0			
13	Appalachian State Mountaineers	Boone, North Carolina	1974	2020	47	1385	675	710	.487	-5.90	-3.30	0	10	2	2	0	0
14	Arizona State Sun Devils	Tempe, Arizona	1912	2020	105	2570	1368	1202	.532	4.94	3.91	7	8	0	16	0	0
15	Arizona Wildcats	Tucson, Arizona	1905	2020	113	2754	1808	945	.657	8.92	5.03	27	24	7	35	4	1
16	Arkansas Razorbacks	Fayetteville, Arkansas	1924	2020	97	2675	1708	967	.639	7.52	3.09	16	26	7	32	6	1
17	Arkansas State Red Wolves	State University, Arkansas	1971	2020	50	1452	743	709	.512	-3.58	-2.94	0	7	1	1	0	0
18	Arkansas-Pine Bluff Golden Lions	Pine Bluff, Arkansas	1999	2020	22	667	199	468	.298	-19.55	-8.53	0	0	1	1	0	0
19	Armstrong Pirates	Savannah, Georgia	1987	1987	1	28	6	22	.214	-21.60	-4.78	0	0	0	0	0	0
20	Army Black Knights	West Point, New York	1903	2020	118	2516	1250	1266	.497	-9.23	-4.68	0	0	0	0	0	0

Rk	School	City, State	From	To	Yrs	G	W	L	W-L%	SRS	SOS	AP	CREG	CTRN	NCAA	FF	NC
21	Auburn Tigers	Auburn, Alabama	1906	2020	114	2599	1389	1209	.535	5.54	4.17	8	4	2	10	1	0

Web Scraping Using R

- Different web scraping techniques are required to deal with different situations of data in R.
- Data have been organized into files.
 - ▶ Directly download it and read it in R
- Data are contained in HTML pages.
 - ▶ Static data
 - ▶ Dynamic data

Import Data Files from Websites

- These files that can be read by **read.csv** or related functions.
- They can be directly imported from a URL.
- Example: we extract the most recent Australian Open Tennis Championships match (AUS Open):

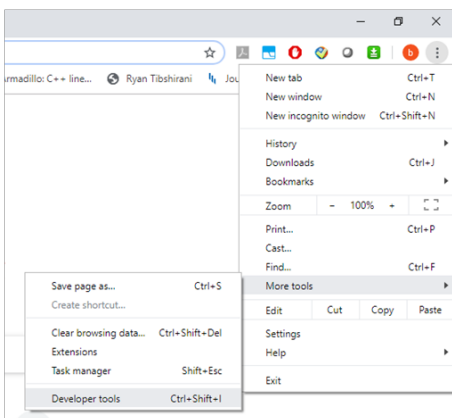
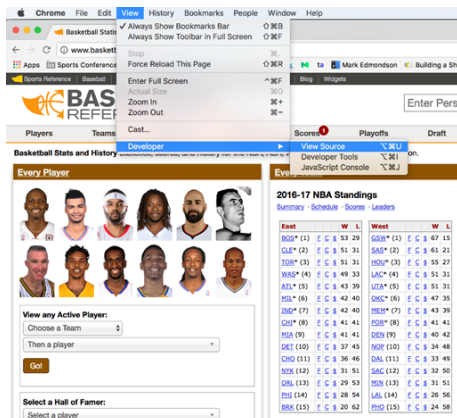
```
url <- "http://www.tennis-data.co.uk/2020/ausopen.csv"
tennis_au <- read.csv(url)
str(tennis_au)
```

Static Data and Dynamic Data

- Most of data in the web are not organized into files, which can be directly imported into R.
- Before we capture these data, we need to determined whether the data is static or dynamic based on the source code.
- Static data is the data that can be seen in the source code.
- We cannot see the dynamic data in the source code.

Static Data and Dynamic Data

- The source code can be accessed by View → Developer → View Source in Chrome. Or right click the website and choose “View Page Source”.



Static Data and Dynamic Data

Exercise: Determine what kind of the data are in the following examples, static or dynamic.

- http://tennisabstract.com/reports/atp_elo_ratings.html
- <https://www.espn.com/sports/cricket/ci/content/stats/>

Static Data and Dynamic Data

tennisabstract.com

Current Elo ratings for the ATP tour. This list includes only those players who have completed 10 or more tour-level, round qualifying, men's challenger, or ITF \$50K+ matches in the last 52 weeks.

Unlike the official rankings, Elo ratings give credit for who you play, not the round or tournament in which you play them. I've written an extensive introduction to tennis Elo ratings [here](#).

A 100-point difference in Elo ratings implies that the favorite has a 64% chance of winning; 200 points implies 76%, 300 points implies 85%, 400 points implies 91%, and 500 points implies 95%. The overall rating ("Elo") doesn't consider surface, and the surface-specific ratings ("Hard" etc.) are based solely on matches played on a single surface.

To generate forecasts for a specific matchup, use a 50/50 blend of overall Elo and surface-specific Elo. These 50/50 blends are shown in the table as "Helo", "Elo", and "gelo". The "default" match type is best-of-three, so in a best-of-five match, the favorite will have a better chance of winning, by a factor that depends on the best-of-three odds.

Updated weekly(ish). Last update: 2020-09-28

Rank	Player	Age	Elo	Hard	Clay	Grass	Helo	Elo	gelo	Peak Match	Peak Age
1	Novak Djokovic	33.3	2255.4	2142.9	2085.6	2153.9	2191	2170.5	2154.7	2016 Miami F	28.8
2	Rafael Nadal	34.3	2186.9	2045.2	2115.2	1971.9	2115	2148	1951.4	2009 Madrid SF	22.8
3	Roger Federer	36.5	2170.0	2051.7	1924.3	1933.8	2115	1907	1951.9	2007 Dubai F	25.8
4	Dominic Thiem	27.1	2079.8	1988.9	2009.2	1914.3	2034.4	2044.5	1847.0	2016 Halle R16	22.8
5	Andrey Rublev	22.9	2023.5	1910.0	1786.6	1516.4	1967.4	1904.5	1706.9	2020 Hamburg F	22.8
6	Stefanos Tsitsipas	22.1	2022.2	1900.0	1989.9	1975.1	1980.8	1960.5	1977.6	2020 Cincinnati R16	22.0
7	Dani Medvedev	24.8	2020.0	1954.1	1921.7	1942.5	1971	1930.9	1831.2	2019 Shanghai F	23.7
8	Kirill Andreiev	23.4	1984.8	1904.6	1912.2	1836.0	1944.4	1979.4	1810.3	2017 Corvallis F	20.3
9	Matteo Berrettini	24.7	1963.2	1901.1	1952.2	1915.3	1973.2	1792.7	1797.8	2017 Cincinnati SF	22.3
10	Francesca Schiavone	32.4	1953.9	1964.9	1798.3	1913.4	1909.4	1871.1	1853.8	2016 Rotterdam R16	27.8
11	Denis Morfilla	34.1	1945.5	1911.2	1670.0	1548.1	1925.5	1605.2	1744.3	2015 Monte Carlo GP	24.8
12	Flavia Pennetta	28.5	1935.4	1841.7	1947.6	1399.3	1888.5	1891.5	1744.3	2020 Rome R32	28.5
13	Dimitri Pavlou	21.4	1931.6	1932.0	1706.2	1339.1	1881.8	1819.8	1830.3	2020 Australian R16	20.7
14	Blanca Zvereva	25.5	1928.1	1907.1	1859.4	1832.5	1924.2	1891.2	1790.3	2016 Australian Open R32	25.8
15	Dimitri Scheremeta	28.1	1923.3	1743.8	1863.4	1667.3	1827	1905.5	1711.1	2019 Beijing R12	27.1
16	Matteo Berrettini	24.4	1921.9	1771.6	1862.9	1722.3	1848.8	1852.4	1847.1	2019 Vienna GP	23.3
17	Alexa De Minaur	21.6	1920.7	1863.1	1363.2	1350.9	1891.9	1614.1	1740.7	2020 Alp Cup RRR	20.8
18	Flavia Pennetta	29.2	1912.0	1940.6	1712.4	1340.8	1797.3	1812.2	1620.4	2017 Roland Garros R16	29.8
19	Dimitri Scheremeta	28.8	1888.1	1791.1	1757.9	1722.3	1844.8	1820.8	1810.2	2017 Rome R32	28.4
20	Blanca Zvereva	24.7	1888.2	1791.1	1742.1	1625.1	1847.1	1812.2	1797.8	2009 Indian Wells R64	23.9
21	Blanca Zvereva	28.1	1888.1	1790.5	1656.0	1611.3	1844.8	1750.1	1757.7	2016 Wimbledon SF	28.4
22	Matteo Berrettini	23.2	1868.8	1770.0	1796.9	1769.3	1830.3	1841.1	1835.1	2010 Dubai R16	21.4
23	Dimitri Scheremeta	28.0	1863.0	1633.0	1612.9	1445.4	1843.0	1747.9	1635.5	2020 Cincinnati R16	23.0
24	Andrey Rublev	23.3	1872.0	1786.8	1798.9	1629.4	1922	1830.9	1701.7	2017 Dubai SF	28.6
25	Novak Djokovic	30.2	1871.3	1804.8	1280.4	1403.4	1839.4	1575.9	1862.2	2020 US Open R32	30.2
26	Novak Djokovic	21.7	1868.0	1598.0	1839.3	1398.9	1732.9	1852.2	1832.4	2020 Hamburg GP	21.7

```
26 <table width="100%" align="left">
27 <tr>
28 <td align="left" class="headline">
29 <div class="class" class="blackheader">
30 <a href="http://www.tennisabstract.com/tennis.aspx" style="color: blue">abstract</a></div>
31 <div align="right" align="right" style="vertical-align: top">
32 </div>
33 </table>
34 <table width="600px">
35 <tr>
36 <td align="right">
37 <div align="right">
38 <div align="right">
39 <div align="right">
40 <div align="right">
41 <div align="right">
42 <div align="right">
43 <div align="right">
44 <div align="right">
45 <div align="right">
46 <div align="right">
47 <div align="right">
48 <div align="right">
49 <div align="right">
50 <div align="right">
51 <div align="right">
52 <div align="right">
53 <div align="right">
54 <div align="right">
55 <div align="right">
56 <div align="right">
57 <div align="right">
58 <div align="right">
59 <div align="right">
60 <div align="right">
61 <div align="right">
62 <div align="right">
63 <div align="right">
64 <div align="right">
65 <div align="right">
66 <div align="right">
67 <div align="right">
68 <div align="right">
69 <div align="right">
70 <div align="right">
71 <div align="right">
72 <div align="right">
73 <div align="right">
74 <div align="right">
75 <div align="right">
76 <div align="right">
77 <div align="right">
78 <div align="right">
79 <div align="right">
80 <div align="right">
81 <div align="right">
82 <div align="right">
83 <div align="right">
84 <div align="right">
85 <div align="right">
86 <div align="right">
87 <div align="right">
88 <div align="right">
89 <div align="right">
90 <div align="right">
91 <div align="right">
92 <div align="right">
93 <div align="right">
94 <div align="right">
95 <div align="right">
96 <div align="right">
97 <div align="right">
98 <div align="right">
99 <div align="right">
100 <div align="right">
101 <div align="right">
102 <div align="right">
103 <div align="right">
104 <div align="right">
105 <div align="right">
106 <div align="right">
107 <div align="right">
108 <div align="right">
109 <div align="right">
110 <div align="right">
111 <div align="right">
112 <div align="right">
113 <div align="right">
114 <div align="right">
115 <div align="right">
116 <div align="right">
117 <div align="right">
118 <div align="right">
119 <div align="right">
120 <div align="right">
121 <div align="right">
122 <div align="right">
123 <div align="right">
124 <div align="right">
125 <div align="right">
126 <div align="right">
127 <div align="right">
128 <div align="right">
129 <div align="right">
130 <div align="right">
131 <div align="right">
132 <div align="right">
133 <div align="right">
134 <div align="right">
135 <div align="right">
136 <div align="right">
137 <div align="right">
138 <div align="right">
139 <div align="right">
140 <div align="right">
141 <div align="right">
142 <div align="right">
143 <div align="right">
144 <div align="right">
145 <div align="right">
146 <div align="right">
147 <div align="right">
148 <div align="right">
149 <div align="right">
150 <div align="right">
151 <div align="right">
152 <div align="right">
153 <div align="right">
154 <div align="right">
155 <div align="right">
156 <div align="right">
157 <div align="right">
158 <div align="right">
159 <div align="right">
160 <div align="right">
161 <div align="right">
162 <div align="right">
163 <div align="right">
164 <div align="right">
165 <div align="right">
166 <div align="right">
167 <div align="right">
168 <div align="right">
169 <div align="right">
170 <div align="right">
171 <div align="right">
172 <div align="right">
173 <div align="right">
174 <div align="right">
175 <div align="right">
176 <div align="right">
177 <div align="right">
178 <div align="right">
179 <div align="right">
180 <div align="right">
181 <div align="right">
182 <div align="right">
183 <div align="right">
184 <div align="right">
185 <div align="right">
186 <div align="right">
187 <div align="right">
188 <div align="right">
189 <div align="right">
190 <div align="right">
191 <div align="right">
192 <div align="right">
193 <div align="right">
194 <div align="right">
195 <div align="right">
196 <div align="right">
197 <div align="right">
198 <div align="right">
199 <div align="right">
200 <div align="right">
201 <div align="right">
202 <div align="right">
203 <div align="right">
204 <div align="right">
205 <div align="right">
206 <div align="right">
207 <div align="right">
208 <div align="right">
209 <div align="right">
210 <div align="right">
211 <div align="right">
212 <div align="right">
213 <div align="right">
214 <div align="right">
215 <div align="right">
216 <div align="right">
217 <div align="right">
218 <div align="right">
219 <div align="right">
220 <div align="right">
221 <div align="right">
222 <div align="right">
223 <div align="right">
224 <div align="right">
225 <div align="right">
226 <div align="right">
227 <div align="right">
228 <div align="right">
229 <div align="right">
230 <div align="right">
231 <div align="right">
232 <div align="right">
233 <div align="right">
234 <div align="right">
235 <div align="right">
236 <div align="right">
237 <div align="right">
238 <div align="right">
239 <div align="right">
240 <div align="right">
241 <div align="right">
242 <div align="right">
243 <div align="right">
244 <div align="right">
245 <div align="right">
246 <div align="right">
247 <div align="right">
248 <div align="right">
249 <div align="right">
250 <div align="right">
251 <div align="right">
252 <div align="right">
253 <div align="right">
254 <div align="right">
255 <div align="right">
256 <div align="right">
257 <div align="right">
258 <div align="right">
259 <div align="right">
260 <div align="right">
261 <div align="right">
262 <div align="right">
263 <div align="right">
264 <div align="right">
265 <div align="right">
266 <div align="right">
267 <div align="right">
268 <div align="right">
269 <div align="right">
270 <div align="right">
271 <div align="right">
272 <div align="right">
273 <div align="right">
274 <div align="right">
275 <div align="right">
276 <div align="right">
277 <div align="right">
278 <div align="right">
279 <div align="right">
280 <div align="right">
281 <div align="right">
282 <div align="right">
283 <div align="right">
284 <div align="right">
285 <div align="right">
286 <div align="right">
287 <div align="right">
288 <div align="right">
289 <div align="right">
290 <div align="right">
291 <div align="right">
292 <div align="right">
293 <div align="right">
294 <div align="right">
295 <div align="right">
296 <div align="right">
297 <div align="right">
298 <div align="right">
299 <div align="right">
300 <div align="right">
301 <div align="right">
302 <div align="right">
303 <div align="right">
304 <div align="right">
305 <div align="right">
306 <div align="right">
307 <div align="right">
308 <div align="right">
309 <div align="right">
310 <div align="right">
311 <div align="right">
312 <div align="right">
313 <div align="right">
314 <div align="right">
315 <div align="right">
316 <div align="right">
317 <div align="right">
318 <div align="right">
319 <div align="right">
320 <div align="right">
321 <div align="right">
322 <div align="right">
323 <div align="right">
324 <div align="right">
325 <div align="right">
326 <div align="right">
327 <div align="right">
328 <div align="right">
329 <div align="right">
330 <div align="right">
331 <div align="right">
332 <div align="right">
333 <div align="right">
334 <div align="right">
335 <div align="right">
336 <div align="right">
337 <div align="right">
338 <div align="right">
339 <div align="right">
340 <div align="right">
341 <div align="right">
342 <div align="right">
343 <div align="right">
344 <div align="right">
345 <div align="right">
346 <div align="right">
347 <div align="right">
348 <div align="right">
349 <div align="right">
350 <div align="right">
351 <div align="right">
352 <div align="right">
353 <div align="right">
354 <div align="right">
355 <div align="right">
356 <div align="right">
357 <div align="right">
358 <div align="right">
359 <div align="right">
360 <div align="right">
361 <div align="right">
362 <div align="right">
363 <div align="right">
364 <div align="right">
365 <div align="right">
366 <div align="right">
367 <div align="right">
368 <div align="right">
369 <div align="right">
370 <div align="right">
371 <div align="right">
372 <div align="right">
373 <div align="right">
374 <div align="right">
375 <div align="right">
376 <div align="right">
377 <div align="right">
378 <div align="right">
379 <div align="right">
380 <div align="right">
381 <div align="right">
382 <div align="right">
383 <div align="right">
384 <div align="right">
385 <div align="right">
386 <div align="right">
387 <div align="right">
388 <div align="right">
389 <div align="right">
390 <div align="right">
391 <div align="right">
392 <div align="right">
393 <div align="right">
394 <div align="right">
395 <div align="right">
396 <div align="right">
397 <div align="right">
398 <div align="right">
399 <div align="right">
400 <div align="right">
401 <div align="right">
402 <div align="right">
403 <div align="right">
404 <div align="right">
405 <div align="right">
406 <div align="right">
407 <div align="right">
408 <div align="right">
409 <div align="right">
410 <div align="right">
411 <div align="right">
412 <div align="right">
413 <div align="right">
414 <div align="right">
415 <div align="right">
416 <div align="right">
417 <div align="right">
418 <div align="right">
419 <div align="right">
420 <div align="right">
421 <div align="right">
422 <div align="right">
423 <div align="right">
424 <div align="right">
425 <div align="right">
426 <div align="right">
427 <div align="right">
428 <div align="right">
429 <div align="right">
430 <div align="right">
431 <div align="right">
432 <div align="right">
433 <div align="right">
434 <div align="right">
435 <div align="right">
436 <div align="right">
437 <div align="right">
438 <div align="right">
439 <div align="right">
440 <div align="right">
441 <div align="right">
442 <div align="right">
443 <div align="right">
444 <div align="right">
445 <div align="right">
446 <div align="right">
447 <div align="right">
448 <div align="right">
449 <div align="right">
450 <div align="right">
451 <div align="right">
452 <div align="right">
453 <div align="right">
454 <div align="right">
455 <div align="right">
456 <div align="right">
457 <div align="right">
458 <div align="right">
459 <div align="right">
460 <div align="right">
461 <div align="right">
462 <div align="right">
463 <div align="right">
464 <div align="right">
465 <div align="right">
466 <div align="right">
467 <div align="right">
468 <div align="right">
469 <div align="right">
470 <div align="right">
471 <div align="right">
472 <div align="right">
473 <div align="right">
474 <div align="right">
475 <div align="right">
476 <div align="right">
477 <div align="right">
478 <div align="right">
479 <div align="right">
480 <div align="right">
481 <div align="right">
482 <div align="right">
483 <div align="right">
484 <div align="right">
485 <div align="right">
486 <div align="right">
487 <div align="right">
488 <div align="right">
489 <div align="right">
490 <div align="right">
491 <div align="right">
492 <div align="right">
493 <div align="right">
494 <div align="right">
495 <div align="right">
496 <div align="right">
497 <div align="right">
498 <div align="right">
499 <div align="right">
500 <div align="right">
501 <div align="right">
502 <div align="right">
503 <div align="right">
504 <div align="right">
505 <div align="right">
506 <div align="right">
507 <div align="right">
508 <div align="right">
509 <div align="right">
510 <div align="right">
511 <div align="right">
512 <div align="right">
513 <div align="right">
514 <div align="right">
515 <div align="right">
516 <div align="right">
517 <div align="right">
518 <div align="right">
519 <div align="right">
520 <div align="right">
521 <div align="right">
522 <div align="right">
523 <div align="right">
524 <div align="right">
525 <div align="right">
526 <div align="right">
527 <div align="right">
528 <div align="right">
529 <div align="right">
530 <div align="right">
531 <div align="right">
532 <div align="right">
533 <div align="right">
534 <div align="right">
535 <div align="right">
536 <div align="right">
537 <div align="right">
538 <div align="right">
539 <div align="right">
540 <div align="right">
541 <div align="right">
542 <div align="right">
543 <div align="right">
544 <div align="right">
545 <div align="right">
546 <div align="right">
547 <div align="right">
548 <div align="right">
549 <div align="right">
550 <div align="right">
551 <div align="right">
552 <div align="right">
553 <div align="right">
554 <div align="right">
555 <div align="right">
556 <div align="right">
557 <div align="right">
558 <div align="right">
559 <div align="right">
560 <div align="right">
561 <div align="right">
562 <div align="right">
563 <div align="right">
564 <div align="right">
565 <div align="right">
566 <div align="right">
567 <div align="right">
568 <div align="right">
569 <div align="right">
570 <div align="right">
571 <div align="right">
572 <div align="right">
573 <div align="right">
574 <div align="right">
575 <div align="right">
576 <div align="right">
577 <div align="right">
578 <div align="right">
579 <div align="right">
580 <div align="right">
581 <div align="right">
582 <div align="right">
583 <div align="right">
584 <div align="right">
585 <div align="right">
586 <div align="right">
587 <div align="right">
588 <div align="right">
589 <div align="right">
590 <div align="right">
591 <div align="right">
592 <div align="right">
593 <div align="right">
594 <div align="right">
595 <div align="right">
596 <div align="right">
597 <div align="right">
598 <div align="right">
599 <div align="right">
600 <div align="right">
601 <div align="right">
602 <div align="right">
603 <div align="right">
604 <div align="right">
605 <div align="right">
606 <div align="right">
607 <div align="right">
608 <div align="right">
609 <div align="right">
610 <div align="right">
611 <div align="right">
612 <div align="right">
613 <div align="right">
614 <div align="right">
615 <div align="right">
616 <div align="right">
617 <div align="right">
618 <div align="right">
619 <div align="right">
620 <div align="right">
621 <div align="right">
622 <div align="right">
623 <div align="right">
624 <div align="right">
625 <div align="right">
626 <div align="right">
627 <div align="right">
628 <div align="right"&
```

Static Data and Dynamic Data

STATISTICS
STATSGURU
RECORDS
SUPERSTATS
ASK STEVEN
NUMBER GAME
WORLD CUP
WORK

Number Crunching
229*

The score by Australia's Belinda Carr, in a World Cup game against Denmark in 1997. It is the highest score in women's ODI, and was the highest score in all ODIs till Rohit Sharma made 264 against Sri Lanka in 2014.

Imperious Maxwell sets up thumping win

Runs: 21 7 0 14 0 9 149 64

Readers recommend

Curated Tweets by @ESPNCricket

Arvind Mathesh
@ArvindMathesh

I enjoyed reading that I was being transported along with the story - Rahul Tewatia and the romance of the struggle! [ESPNCricket.com](#) [@ESPNCricket](#) [espnocricket.com/content/articles...](#)

Rahul Tewatia and the romance of the str...
He was 5 of 13. He finished 53 of 31. This in... [espnocricket.com](#)

Danish Sat
@DanishSat

This is a good read Thanks for placing this Sidharth Moriya [@espnocricket.com/story/...](#) [@GS95](#)

Rahul Tewatia and the romance of the str...
He was 5 of 13. He finished 53 of 31. This in... [espnocricket.com](#)

Features and Analysis

Duckless Implosions, and more wickets than runs

Ask Stevens: What is the record for most wickets in a T20 match?

Moseen and Stokes show England's all-round might

Numbers Game: Unlike most other teams, England have a very successful lower middle order

When Moseen trumped Sobers

Also, was Joe Root's 190 the highest score by a batsman in his first Test as captain?

Stevens Lynch answers your trivia questions

When all 11 players had a bowl

Ask Stevens: Also, what's the highest score made on T20 debut?

What score have you not got?

Ask Stevens: Sachin Tendulkar was never dismissed or left out either - on which number?

The boundary-rider effect

Numbers Game: Has the extra outfielder affected scoring rates in the last ten years of ODIs?

Did you know?

Number of single-digit scores by the Australia batsmen against South Africa in Hobart was 16 - their joint-record in any Test. The last time before this they had 16 such scores in a Test was in 1952 at the Oval.

Records

Most consecutive five-wickets-in-an-innings - Test matches

Player	Sp	Row	Opp	Year
CTB Turner	6	5/44	1 Australia	v England
	7/43	3 Australia	v England	
	5/27	2 Australia	v England	
	5/36	4 Australia	v England	
	6/112	2 Australia	v England	

Ground	Match Date	Scorecard
Sydney	10 Feb 1888	Test # 27
Sydney	10 Feb 1888	Test # 27
Sydney	30 Jul 1889	Test # 28
Lord's	30 Jul 1889	Test # 28
The Oval	13 Aug 1888	Test # 29

Readers recommend

Curated Tweets by [@ESPNorinfo](#)

Arvind Mahesth
@arvindmahesth

I enjoyed reading this! I was being transported along with the story - Rahul Tewatia and the romance of the struggle! ESPNcricinfo.com @ESPNcricinfo espn.cricinfo.com/ci/content/stp...



Rahul Tewatia and the romance of the stru...
He was 5 off 13. He finished 53 off 31. This in...
espnindia.com

Sep 28, 20

 Danish Salt

This is such a good read! Thanks for piecing this
Sidharth Monga 🙌 espn.cricinfo.com/story/_/id/299

Rahul Tewatia and the romance of the stru...
He was 5 off 13. He finished 53 off 31. This in...
espn.cricinfo.com

```

1 <DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-
2 transitional.dtd">
3 <!--
4 <!-- hostname: web03, edition: view, espcricinfo-en-us, country: us, cluster: us, created: 2020-10-01 21:56:52 -
5 -->
6 <html xmlns="http://www.w3.org/1999/xhtml" xmlns:fb="http://www.facebook.com/2008/html"
7 xmlns:og="http://opengraphprotocol.org/schema/" xmlns:fb="http://developers.facebook.com/schema/">
8 <head>
9 <script type="text/javascript" var aj_startp=<new Date()>.getTime()</script>
10 <meta name="google-site-verification" content="Eadg3Kj3U9W9e-No2R03E4n53Jf0d0efxnt">
11 <!--Crickets Statistics / Statguru | ESPNcricinfo.com-->
12 <meta http-equiv="Content-Type" content="text/html; charset=utf-8">
13 <meta name="Keywords" content="cricket statistics, statguru, cricket stats, cricket records, batting average,
14 centuries">
15 <meta name="new_keywords" content="cricket statistics, statguru, cricket stats, cricket records, batting
16 average, cricket centuries" />
17
18
19 <!--
20 <!-- description: Content="Cricket statistics and stories around crickets numbers on ESPN Cricinfo" />
21 <!-- if IE 9.5 -->
22 <script language="javascript" type="text/javascript">
23 function fncCreateJsमित्त(IconTitle)
24 fncIcresJsमित्त();
25 window.external.mditWeb0addJsमित्त("Quick Links",
26 "https://a.espcricinfo.com/ci/content/current/url/1049413.html",
27 "https://a.espcricinfo.com/ci/content/current/url/1049413.html",
28 "https://a.espcricinfo.com/ci/content/current/url/1049413.html",
29 "https://a.espcricinfo.com/espcricinfo/taivcon.ieo")window.external.mditWeb0addJsमित्त("Both Ends",
30 "https://a.espcricinfo.com/ci/content/current/url/1049413.html",
31 "https://a.espcricinfo.com/espcricinfo/taivcon.ieo")window.external.mditWeb0addJsमित्त("MATCH Dats",
32 "https://a.espcricinfo.com/ci/content/current/url/1049546.html",
33 "https://a.espcricinfo.com/espcricinfo/taivcon.ieo")window.external.mditWeb0addJsमित्त("Tasndhs",
34 "https://a.espcricinfo.com/ci/content/current/url/349455.html",
35 "https://a.espcricinfo.com/espcricinfo/taivcon.ieo")window.external.mditWeb0addJsमित्त("WI v IND",
36 "https://a.espcricinfo.com/west-indies-vs-india-2017/content/current/series/1098203.html",
37 "https://a.espcricinfo.com/espcricinfo/taivcon.ieo")window.external.mditWeb0addJsमित्त("ENG v SA",
38 "https://a.espcricinfo.com/espcricinfo/taivcon.ieo")window.external.mditWeb0addJsमित्त("IND vs BAN",
39 "https://a.espcricinfo.com/espcricinfo/taivcon.ieo")window.external.mditWeb0addJsमित्त("WOM",
40 "https://a.espcricinfo.com/espcricinfo/taivcon.ieo")window.external.mditWeb0addJsमित्त("WTC 2019",
41 "https://a.espcricinfo.com/espcricinfo/taivcon.ieo")window.external.mditWeb0addJsमित्त("SL v ZIM",
42 "https://a.espcricinfo.com/ri-lanka-vs-zimbabwe-2017/content/current/series/1104471.html",
43 "https://a.espcricinfo.com/espcricinfo/taivcon.ieo")window.external.mditWeb0addJsमित्त();
44
45
46 function fncIcresJsमित्त()
47 window.external.mditWeb0clearJsमित्त();
48
49 </script>
50
51 <meta name="application-task" content="name=Live Scores;action=
52 uri=https://a.espcricinfo.com/ci/engine/current/match/scores/live.html;icon=
53 uri=https://a.espcricinfo.com/espcricinfo/taivcon.ieo"/>
54 <meta name="application-task" content="name=Latest News;action=
55 uri=https://a.espcricinfo.com/ci/content/current/story/news.html;icon=
56 uri=https://a.espcricinfo.com/espcricinfo/taivcon.ieo"/>
57 <meta name="application-task" content="name=Match Fixtures;action=
58 uri=https://a.espcricinfo.com/ci/content/current/match/fixtures/index.html;icon=
59 uri=https://a.espcricinfo.com/espcricinfo/taivcon.ieo"/>
60 <meta name="application-task" content="name=Recent Match;action=
61 uri=https://a.espcricinfo.com/ci/engine/current/match/scores/recent.html;icon=
62 uri=https://a.espcricinfo.com/espcricinfo/taivcon.ieo"/>
63 <meta name="application-task" content="name=Photo gallery;action=
64 uri=https://a.espcricinfo.com/ci/content/current/image/index.html;icon=
65 uri=https://a.espcricinfo.com/espcricinfo/taivcon.ieo"/>
66 <meta name="application-task" content="name=Audio/Video;action=
67 uri=https://a.espcricinfo.com/ci/content/Video/audio/index.html;icon=
68 uri=https://a.espcricinfo.com/espcricinfo/taivcon.ieo"/>
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
```

This is dynamic data.

Web Scraping for Static Data in R

R provides several approaches for web scraping static data. Two of them will be discussed in this workshop.

- **readLines** function: Read the source code of the HTML page.
- **rvest** package: Capture useful data by identifying the elements contains the data in the source code.

Web Scrapping for Static Data in R

Use **readLines** function for College basketball school index.

```
web_page <- readLines("https://www.sports-reference.com/cbb/schools/")
head(web_page, n = 10L)
```

```
## [1] ""
## [2] "<!DOCTYPE html>"
## [3] "<html data-version=\"klecko-\" data-root=\"/home/cbb/build\" itemscope itemtype=\"https://schema.org/W"
## [4] "<head>"
## [5] "    <meta charset=\"utf-8\">"
## [6] "    <meta http-equiv=\"x-ua-compatible\" content=\"ie=edge\">"
## [7] "    <meta name=\"viewport\" content=\"width=device-width, initial-scale=1.0, maximum-scale=2.0\" />"
## [8] "    <link rel=\"dns-prefetch\" href=\"https://d2p3byggnzw9w3.cloudfront.net/req/202009101\" />"
## [9] ""
## [10] "    <title>School Index | College Basketball at Sports-Reference.com</title>"
```

- Gives the source code.
- Needs data cleaning and organization.

Web Scraping for Static Data in R

Before we talk about web scraping by **rvest** package, we need to know how to locate the elements containing the data in the source code.

- Right click the page and choose “Inspect”.
- Click “Select an element in the page to inspect it”.
- We can locate the element by CSS selector or XPATH.

Use http://tennisabstract.com/reports/atp_elo_ratings.html as an example

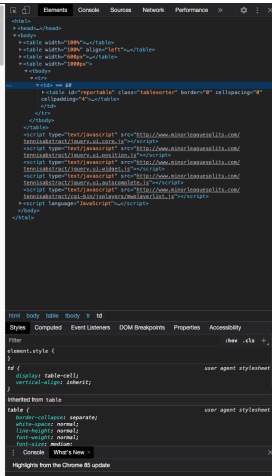
- tennisabstract.com**

Unlike the official rankings, Elo ratings give credit for *who* you play, not the round or tournament in which you play them. I've written an extensive introduction to tennis Elo ratings [here](#).

To generate forecasts for a specific matchup, use a 50/50 blend of overall Elo and surface-specific Elo. These 50/50 blends are shown in the table as "bElo," "cElo," and "gElo." The 'default' match type is best-of-three, so in a best-of-five match, the favorite will have a better chance of winning, by a factor that depends on the best-of-three odds.

896 x 7069 (ish). Last update: 2020-09-28

Rank	Player	Age	ESL	HeadRate	ClayRate	GrassRate	IRL	cESL	gESL	Peak Month	Peak Age	Peak ES
1	Robert Dinkelds	33.3	247.6	200.6	200.6	200.6	217.6	214.6	173.6	2020 March	22.6	248.6
2	Robert Dinkelds	34.3	185.6	205.6	177.6	177.6	217.6	214.6	173.6	2020 March	22.6	248.6
3	Robert Dinkelds	36.6	179.6	201.6	150.4	153.6	215.6	190.1	201.6	2017 October	16.6	237.6
4	Dominic Thoms	27.0	207.6	199.6	209.2	161.4	203.4	200.5	1947.0	2019 Hubs R16	22.8	219.6
5	Andrew Rubins	23.8	202.6	191.6	176.6	151.4	197.2	190.4	176.6	2020 Hamburg F	22.9	203.6
6	Hendrik Thoms	22.1	202.2	190.6	189.6	191.6	188.6	190.5	179.6	2020 Cincinnati R1	22.2	208.6
7	Daniel Melnikov	34.8	200.6	190.4	161.7	146.5	180.1	160.6	161.6	2017 Shanghai F	23.7	198.6
8	Alexander Jansen	34.6	200.6	186.6	177.2	163.6	189.6	191.6	191.6	2017 Berlin F	23.9	217.6
9	Nick Smeets	24.7	194.6	191.1	160.2	161.2	176.2	176.2	176.2	2017 Cincinnati F	23.2	247.6
10	Robert Dinkelds	32.4	193.6	186.6	190.6	191.4	190.4	191.6	186.6	2019 Rotterdam R16	27.8	213.6
11	Gianni Miala	34.1	194.6	191.2	167.0	154.1	162.6	160.2	174.6	2015 Monte Carlo GP	28.6	207.6
12	Pete Kravtsov	28.5	193.6	194.7	164.6	139.6	188.6	191.1	164.7	2020 Rome R2	28.9	193.6
13	Denis Shmelev	24.1	193.6	183.2	167.6	139.1	184.1	191.6	193.6	2020 Auckland R16	26.7	192.6
14	Alan Vavrin	35.6	192.6	190.1	168.4	152.6	182.6	190.5	173.6	2019 Australian Open R2	30.6	194.6
15	Daniel Melnikov	34.6	192.6	167.6	160.4	146.6	186.6	191.6	171.6	2019 Australian R2	30.6	194.6
16	Matteo Bartalis	24.4	192.6	177.6	160.2	177.2	164.6	186.2	194.7	2017 Vienna GP	23.6	191.6
17	Alex De Minor	21.6	192.6	183.1	156.2	156.6	181.9	161.4	190.1	2020 Aig GP	20.9	164.6
18	Patrick Corbin-Buffe	29.2	191.6	190.6	171.4	134.6	167.6	161.2	160.6	2017 Roland Garros R16	29.9	194.6
19	Daniel Sjöberg	28.8	186.6	171.1	175.9	172.6	184.6	162.6	181.2	2017 Rome R2	26.4	207.6
20	de-Wittow Thoms	34.7	186.6	171.9	144.1	162.1	184.1	191.7	175.6	2009 Indian Wells R64	29.6	212.6
21	Matteo Bartalis	24.7	180.1	176.6	160.6	161.3	184.4	170.1	175.6	2016 Wimbledon SF	23.6	212.6
22	Matteo Bartalis	23.2	184.6	176.6	176.6	176.6	176.6	184.6	161.6	2019 Davis Cup	23.6	212.6
23	Danilo Smeets	23.0	183.6	160.2	161.2	144.5	140.4	174.9	193.6	2020 Cincinnati R16	23.0	198.6
24	Andre Mueck	33.1	182.6	176.6	176.6	176.6	162.4	182.6	160.6	2017 Osaka SF	29.6	206.6
25	James Pugh	30.2	181.3	180.4	160.4	140.3	183.6	197.6	166.2	2020 US Open R32	30.2	167.6
26	Gianni Miala	32.7	184.6	160.6	160.6	139.6	173.2	166.2	163.4	2020 Hamburg GP	27.7	166.6



Web Scrapping for Static Data in R

- XPATH: `'//*[@id="reportable"]'`

tennisabstract.com

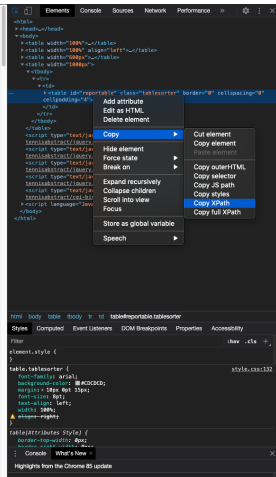
Current Elo ratings for the ATP tour. This list includes only those players who have completed 10 or more tour-level, tour-level qualifying, men's challenger, or ITF \$50K+ matches in the last 52 weeks.

Unlike the official rankings, Elo ratings give credit for *who* you play, not the round or tournament in which you play them. I've written an extensive introduction to tennis Elo ratings [here](#).

A 100-point difference in Elo ratings implies that the favorite has a 64% chance of winning; 200 points implies 76%, 300 points implies 85%, 400 points implies 91%, and 500 points implies 95%. The overall rating ("Elo") doesn't consider surface, and the surface-specific ratings ("Hard" etc.) are based solely on matches played on a single surface.

To generate forecasts for a specific matchup, use a 50/50 blend of overall Elo and surface-specific Elo. These 50/50 blends are shown in the table as "nElo," "cElo," and "gElo." The 'default' match type is best-of-three, so in a best-of-five match, the favorite will have a better chance of winning, by a factor that depends on the best-of-three odds.

TableReportable tablesorter 694 x 7042													
Rank	Player	Age	elo	Handicap	Time/Rate	Green/Red	Wb10	cELO	gELO	Peak Match	Peak Age	Peak ELO	
1	Norvik Dinesen	33.3	2255.4		2142.9	2058.5	2013.9	2189.1	2175.0	2314.7	2015 Miami F	28.8	2400.0
2	Ruben Hult	34.3	2161.0		2068.2	2112.1	1877.9	2150.1	2148.1	1931.4	1990 Madrid SF	22.8	2300.0
3	Bojan Radovic	34.2	2115.0		2021.7	1854.9	1923.7	2141.8	2018.1	2051.7	2017 London F	22.1	2250.0
4	Dimitris Theodor	27.0	2079.8		1888.8	2009.2	1614.3	2054.8	2044.5	1947.0	2016 Hialeah R16	22.8	2122.0
5	Andreas Rubens	22.9	2023.5		1910.0	1785.6	1516.4	1967.2	1904.0	1759.9	2020 Hamburg F	22.9	2023.5
6	Stefanos Tsitsipas	22.1	2022.2		1900.0	1688.0	1515.1	1886.6	1860.5	1790.6	2020 Cincinnati R16	22.0	1988.0
7	Dimitri Michailidis	24.6	2020.0		1904.1	1621.7	1542.5	1927.1	1902.8	1931.2	2017 Stuttgart F	22.7	2128.0
8	Alexander Zverev	24.2	1984.4		1854.6	1642.4	1512.2	1844.8	1844.4	1910.1	2019 Stuttgart F	22.7	2128.0
9	Nick Kyrgios	24.7	1963.3		1861.1	1602.2	1512.3	1827.2	1760.7	1791.9	2017 Cincinnati F	22.3	1967.0
10	Roberto Bautista Agut	32.4	1963.9		1849.6	1739.0	1515.4	1809.4	1871.0	1869.0	2016 Rotterdam R16	27.9	2013.0
11	Sam Querrey	34.1	1945.5		1891.2	1670.0	1546.1	1825.9	1805.2	1744.3	2015 Miami Open SF	28.0	2007.0
12	Evan Kravitch	28.6	1935.0		1845.7	1647.8	1390.3	1888.5	1861.5	1747.3	2020 Roma R32	26.5	1998.0
13	Denis Shapovalov	22.5	1905.2		1786.2	1706.2	1522.0	1841.8	1841.0	1826.3	2020 Auckland R16	22.7	1927.0
14	Sam Querrey	35.5	1928.1		1867.1	1668.4	1532.6	1852.8	1865.2	1740.3	2019 Australian Open R32	38.0	2143.0
15	Dimitri Schenker	28.1	1923.5		1774.0	1682.4	1408.7	1827.3	1908.0	1711.1	2017 Beijing SF	27.1	1940.0
16	Matteo Berrettini	28.4	1912.9		1713.9	1682.0	1772.9	1848.8	1862.4	1847.1	2019 Vienna R32	23.8	1918.0
17	Alex de Minaur	21.6	1902.7		1863.1	1632.7	1590.0	1881.9	1814.0	1740.2	2020 Abu Dubi R32	20.8	1848.0
18	Flavia Pennetta / Naomi Osaka	34.6	1891.6		1846.6	1715.2	1446.6	1841.8	1841.8	1902.9	2020 Cincinnati R16	28.7	1988.0
19	Dimitri Skovrt	29.8	1868.1		1791.1	1757.0	1722.3	1844.6	1828.0	1810.7	2017 Rome R32	28.4	2027.0
20	Jo-Wilfried Tsonga	34.7	1860.2		1791.0	1641.4	1626.1	1841.1	1917.2	1797.8	1990 Indian Wells R64	32.9	2125.0
21	Mina Rezaei	28.7	1860.1		1789.9	1638.0	1661.3	1844.8	1701.1	1725.7	2016 Wimbledon SF	25.5	1925.0
22	Marin Cilic	32.0	1848.8		1776.6	1706.0	1705.4	1830.9	1841.6	1835.1	2017 Dubai R16	21.4	2048.0
23	Bojan Radovic	32.3	1848.0		1802.0	1671.9	1485.5	1843.5	1747.0	1803.2	2020 Cincinnati R16	22.5	1888.0
24	Andy Murray	33.3	1827.0		1746.6	1799.0	1620.4	1822.2	1836.9	1701.0	2017 China SF	37.6	2284.0
25	Nicole Pietrangeli	30.2	1817.3		1808.4	1682.4	1443.0	1833.9	1875.9	1882.0	2020 Las Vegas Open R32	30.2	1879.0
26	Grigor Dimitrov	27.1	1846.0		1690.0	1639.3	1398.9	1732.0	1752.2	1802.2	2020 Hamburg SF	21.7	1948.0



Web Scraping for Static Data in R

Next, we are going to talk about how to use **rvest** for web scraping by using an example.

- Install **rvest** package from cran.

```
install.packages("rvest", repos = "http://cran.us.r-project.org")  
require("rvest")
```

Web Scraping for Static Data in R

- Web scraping data from
http://tennisabstract.com/reports/atp_elo_ratings.html

```
url_elo <- "http://tennisabstract.com/reports/atp_elo_ratings.html"
webpage <- read_html(url_elo)
elo_class <- webpage %>%
  html_nodes(".tablesorter") %>%
  html_table()
elo_id <- webpage %>%
  html_nodes("#reportable") %>%
  html_table()
identical(elo_class, elo_id)
```

```
## [1] TRUE
```

Web Scrapping for Static Data in R

```
elo_xpath <- webpage %>%  
  html_nodes(xpath = '//*[@id="reportable"]') %>%  
  html_table()  
identical(elo_class, elo_xpath)
```

```
## [1] TRUE
```

```
head(elo_class[[1]])
```

##	Rank	Player	Age	Elo	HardRaw	ClayRaw	GrassRaw	hElo
## 1	1	Novak Djokovic	33.3	2255.4	NA	2142.9	2085.6	2013.9 NA 2199.1
## 2	2	Rafael Nadal	34.3	2185.0	NA	2045.2	2111.2	1677.9 NA 2115.1
## 3	3	Roger Federer	38.5	2170.0	NA	2051.7	1824.3	1933.8 NA 2110.9
## 4	4	Dominic Thiem	27.0	2079.8	NA	1989.8	2009.2	1614.3 NA 2034.8
## 5	5	Andrey Rublev	22.9	2023.5	NA	1910.8	1785.6	1516.4 NA 1967.2
## 6	6	Stefanos Tsitsipas	22.1	2022.2	NA	1939.0	1898.9	1573.1 NA 1980.6
##	cElo	gElo	Peak	Match	Peak	Age	Peak	Elo
## 1	2170.5	2134.7	NA	2016 Miami	F	28.8	2469.7	
## 2	2148.1	1931.4	NA	2009 Madrid	SF	22.9	2368.4	
## 3	1997.1	2051.9	NA	2007 Dubai	F	25.6	2379.4	
## 4	2044.5	1847.0	NA	2016 Halle	R16	22.8	2122.5	
## 5	1904.5	1769.9	NA	2020 Hamburg	F	22.9	2023.5	
## 6	1960.5	1797.6	NA	2020 Cincinnati	R16	22.0	2069.1	

Web Scraping for Static Data in R

- Except **html_nodes** and **html_table**, there are many other frequently used functions in **rvest**.
 - ▶ **html_node** : extract element
 - ▶ **html_text** : extract text
 - ▶ **html_attrs** : extract attributes
 - ▶ **html_form** : extract forms
- Please look up rvest cran for more information.
- SelectorGadget is a convenient tool to identify CSS selector.

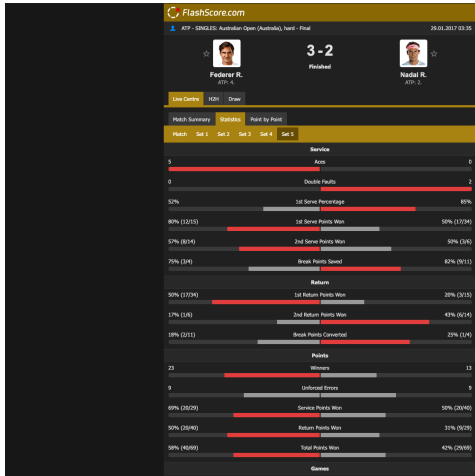
Web Scraping for Dynamic Data in R

- The dynamic data displayed in the website can be different because the website may provide user interaction.
- We need to automate the web browsing process in R for the dynamic data.
- **RSelenium** package helps this automating process by providing connection to Selenium Server.
- Install **RSelenium** package.

```
devtools::install_github("ropensci/RSelenium")  
require("RSelenium")
```

Web Scraping for Dynamic Data in R

- Use **RSelenium** to extract data on 2017 Australian Open Final

[illegible]

Web Scrapping for Dynamic Data in R

- Connect to a selenium server and open browser.

```
rD <- rsDriver(port = 5560L, chromever = "85.0.4183.87")  
remDr <- rD$client
```

- Extract Information and organize data.

```
url <- "http://www.flashscore.com/match/Cj6I5iL9/#match-statistics;0"  
remDr$navigate(url)  
webElem <- remDr$findElements(using = 'class', "statBox")  
webElem <- unlist(lapply(webElem, function(x){x$getElementText()}))[[1]]  
head(unlist(strsplit(webElem, split = '\n')))
```

```
## character(0)
```

```
remDr$close()
```

Web Scraping for Dynamic Data in R

- Frequently used functions of **RSelenium**:
 - ▶ `rsDriver()` : start a selenium server
 - ▶ `navigate()` : navigate web pages
 - ▶ `findElements()` : find elements by CSS selector or XPATH
 - ▶ `getPageSource()` : get current page source
 - ▶ `clickElement()` : click element
- Please go to RSelenium cran for more details.

Web Scraping for Dynamic Data in R

Exercise: Web Scraping for the history basketball recording of UConn

<https://www.flashscore.com/team/connecticut-huskies/8rqVf3Tj/results/>

- Start a selenium server and open web browser.

```
require("RSelenium")
rD <- rsDriver(port = 5533L, chromever = "85.0.4183.87")
remDr <- rD$client
url <- "https://www.flashscore.com/team/connecticut-huskies/8rqVf3Tj/results/"
remDr$navigate(url)
```

Web Scraping for Dynamic Data in R

- Automate to click all “show more results”.

```
repeat{
  x <- try(webElemMore <-
            remDr$findElement(using = 'xpath',
                              '//*[@id="live-table"]/div[1]/div/div/a'),
          silent=T)
  if (inherits(x, "try-error")) break
  webElemMore$clickElement()
}
```

- Extract data, such as time, home/away, score and result.

```
webElemTime <- remDr$findElements(using = 'xpath',
                                   '//*[@class="event__time"]')
webElemTime <-
  unlist(lapply(webElemTime, function(x){x$getElementText()}))
webElemTime <- gsub("\\n", " ", webElemTime)
```

Web Scrapping for Dynamic Data in R

```
webElemHome <-  
  remDr$findElements(using = 'class',  
                      'event__participant')  
  
webElemHome <-  
  unlist(lapply(webElemHome, function(x){x$getElementText()}))  
  
webElemScore <-  
  remDr$findElements(using = 'class', 'event__score')  
webElemScore <-  
  unlist(lapply(webElemScore, function(x){x$getElementText()}))  
  
webElemResult <-  
  remDr$findElements(using = 'class', 'wld')  
webElemResult <-  
  unlist(lapply(webElemResult, function(x){x$getElementText()}))
```

Web Scraping for Dynamic Data in R

- Organize dataset.

```
n <- length(webElemHome)
basketball <-
  data.frame(time = webElemTime,
             Home = webElemHome[seq(n) %% 2 == 1],
             Away = webElemHome[seq(n) %% 2 == 0],
             HomeS = webElemScore[seq(n) %% 2 == 1],
             AwayS = webElemScore[seq(n) %% 2 == 0],
             Result = webElemResult)

head(basketball)

remDr$close()
```

	time	Home	Away	HomeS	AwayS	Result
1	08.03. 16:00	Tulane	UConn	76	80	W
2	05.03. 19:00	UConn	Houston	77	71	W
3	29.02. 14:00	East Carolina	UConn	63	84	W
4	26.02. 19:00	UConn	UCF Knights	81	65	W
5	23.02. 14:00	UConn	South Florida	78	71	W
6	20.02. 19:00	AOT Temple	UConn	93	89	L

Summary

- For different kinds of data, we need to use different web scraping techniques with R.
- One can simply use **read.csv** or related functions to directly import organized files from web pages.
- The static data can be extract with the help of **rvest**.
- We could use **RSelenium** to parse the dynamic data.

Resources

- CSS and HTML crash course
- rvest
- RSelenium
- R task view: web technology

Acknowledgement

The slides are modified from Dr. Kovalchik's material and Wanwan Xu's slides.