

# Web Scraping for Sports Data

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

# 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):

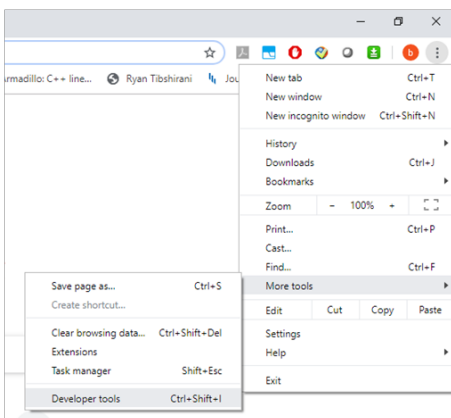
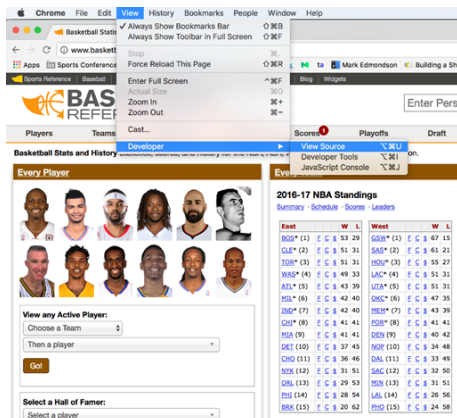
```
url <- "http://www.tennis-data.co.uk/2020/ausopen.csv"
tennis_australia <- read.csv(url)
str(tennis_australia)
```

# 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](http://tennisabstract.com/reports/atp_elo_ratings.html)
- <https://www.espn.com/cricinfo.com/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, 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 they 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 "Elo," "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.

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

Rank	Player	Age	Elo	Hard	Clay	Grass	ITF	cElo	gElo	Peak Match	Peak Age
1	Novak Djokovic	33.3	2228.4	2142.0	2066.6	2013.8	2191.1	2170.3	2104.7	2016 Miami F	28.4
2	Rafael Nadal	33.3	2160.0	2045.2	2111.2	1977.9	2115.1	2148.1	1991.4	2019 Madrid SF	22.8
3	Roger Federer	36.5	2170.0	2051.7	1924.3	1933.8	2119.8	1907.1	1851.9	2017 Dubai F	25.8
4	Dominic Thiem	27.0	2079.8	1988.9	2009.2	1914.3	2034.4	2004.4	2007.0	2016 Halle R16	22.8
5	Andreas Seppi	22.9	2033.5	1910.0	1765.6	1516.4	1967.2	1904.5	1709.9	2020 Hamburg F	22.8
6	Stefanos Tsitsipas	21.2	2022.2	1900.9	1989.9	1975.1	1980.4	1965.0	1977.6	2020 Cincinnati R16	22.2
7	Dani Medvedev	24.6	2002.0	1954.1	1942.7	1942.5	1987.1	1920.0	1871.2	2019 Shanghai F	23.7
8	Alexander Zverev	23.4	1984.8	1904.6	1912.2	1936.0	1944.0	1919.4	1810.3	2017 Corvallis F	20.3
9	Nick Pietrangeli	24.7	1963.2	1901.1	1922.2	1912.3	1927.2	1792.7	1797.8	2017 Cincinnati SF	22.3
10	Roberto Bautista Agut	32.4	1953.9	1964.9	1799.3	1513.4	1909.4	1871.8	1853.6	2016 Rotterdam R16	27.8
11	Denis Shapovalov	24.1	1945.5	1911.2	1970.0	1548.1	1925.5	1905.2	1740.4	2015 Monte Carlo GP	24.6
12	Diego Schwartzman	28.5	1934.4	1941.7	1947.6	1399.3	1888.5	1891.5	1847.3	2020 Rome R32	28.1
13	Denis Shapovalov	21.4	1931.6	1932.0	1706.2	1339.1	1811.4	1918.0	1650.3	2020 Auckland R16	20.7
14	Alan Hingray	34.6	1928.1	1907.1	1958.4	1932.8	1982.6	1902.2	1709.3	2016 Australian Open R32	30.4
15	Owen Seppelt	26.1	1923.5	1733.8	1903.4	1468.7	1837.3	1908.2	1711.1	2019 Beijing R32	27.1
16	Matteo Berrettini	24.4	1921.9	1771.6	1962.9	1737.2	1848.8	1906.4	1847.1	2019 Vienna GP	23.3
17	Adrian Panatta	21.6	1920.7	1963.1	1363.2	1590.8	1891.9	1816.1	1740.7	2020 Alp Cup R16	20.9
18	Flavia Pennetta	29.2	1912.0	1940.2	1712.4	1340.9	1879.3	1912.2	1620.4	2017 Roland Garros R16	25.8
19	Daniil Medvedev	28.6	1888.1	1714.1	1757.9	1273.3	1844.4	1820.0	1710.2	2017 Rome R32	26.4
20	Lucretia Glushko	24.7	1886.2	1791.9	1744.1	1626.1	1841.0	1817.2	1757.6	2019 Indian Wells R64	21.9
21	Maria Sanchez	20.7	1880.1	1795.6	1660.0	1661.3	1844.4	1730.1	1757.2	2016 Wimbledon SF	25.5
22	Maria Sanchez	22.8	1868.8	1770.9	1796.9	1785.4	1830.8	1841.8	1805.1	2010 Dubai R16	21.4
23	Bailey Donnell	22.0	1863.0	1903.0	1422.5	1443.9	1830.4	1747.9	1693.5	2020 Cincinnati R16	23.0
24	Andy Murray	33.3	1827.0	1786.4	1798.8	1629.4	1922.2	1830.9	1700.7	2017 Dubai SF	29.6
25	Yves Faudry	30.2	1871.3	1760.8	1280.4	1400.4	1839.4	1753.9	1662.2	2020 US Open R32	30.2
26	Grigor Dimitrov	21.7	1868.0	1959.0	1839.3	1398.9	1732.5	1852.2	1932.4	2020 Hamburg GP	21.7

Player Search

```

26 <table width="100%" align="left">
27 <tr>
28 <td align="left" class="headline">
29 <div align="center" class="background">
30 <a href="http://www.tennisabstract.com/tennis-spn.asp?style=blue">abstract</a><a href="http://www.tennisabstract.com/spn.asp?style=blue">spn</a></div>
31 <td align="right" class="headline">
32 <a href="http://www.tennisabstract.com/playersearch.asp?style=vertical-align=top">playersearch</a></td>
33 </tr>
34 </table>
35 <table width="100%" align="left">
36 <tr>
37 <td align="left" class="headline">
38 <div align="center" class="background">
39 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></div>
40 <td align="right" class="headline">
41 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></td>
42 </tr>
43 </table>
44 <table width="100%" align="left">
45 <tr>
46 <td align="left" class="headline">
47 <div align="center" class="background">
48 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></div>
49 <td align="right" class="headline">
50 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></td>
51 </tr>
52 </table>
53 <table width="100%" align="left">
54 <tr>
55 <td align="left" class="headline">
56 <div align="center" class="background">
57 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></div>
58 <td align="right" class="headline">
59 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></td>
60 </tr>
61 </table>
62 <table width="100%" align="left">
63 <tr>
64 <td align="left" class="headline">
65 <div align="center" class="background">
66 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></div>
67 <td align="right" class="headline">
68 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></td>
69 </tr>
70 </table>
71 <table width="100%" align="left">
72 <tr>
73 <td align="left" class="headline">
74 <div align="center" class="background">
75 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></div>
76 <td align="right" class="headline">
77 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></td>
78 </tr>
79 </table>
80 <table width="100%" align="left">
81 <tr>
82 <td align="left" class="headline">
83 <div align="center" class="background">
84 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></div>
85 <td align="right" class="headline">
86 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></td>
87 </tr>
88 </table>
89 <table width="100%" align="left">
90 <tr>
91 <td align="left" class="headline">
92 <div align="center" class="background">
93 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></div>
94 <td align="right" class="headline">
95 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></td>
96 </tr>
97 </table>
98 <table width="100%" align="left">
99 <tr>
100 <td align="left" class="headline">
101 <div align="center" class="background">
102 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></div>
103 <td align="right" class="headline">
104 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></td>
105 </tr>
106 </table>
107 <table width="100%" align="left">
108 <tr>
109 <td align="left" class="headline">
110 <div align="center" class="background">
111 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></div>
112 <td align="right" class="headline">
113 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></td>
114 </tr>
115 </table>
116 <table width="100%" align="left">
117 <tr>
118 <td align="left" class="headline">
119 <div align="center" class="background">
120 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></div>
121 <td align="right" class="headline">
122 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></td>
123 </tr>
124 </table>
125 <table width="100%" align="left">
126 <tr>
127 <td align="left" class="headline">
128 <div align="center" class="background">
129 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></div>
130 <td align="right" class="headline">
131 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></td>
132 </tr>
133 </table>
134 <table width="100%" align="left">
135 <tr>
136 <td align="left" class="headline">
137 <div align="center" class="background">
138 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></div>
139 <td align="right" class="headline">
140 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></td>
141 </tr>
142 </table>
143 <table width="100%" align="left">
144 <tr>
145 <td align="left" class="headline">
146 <div align="center" class="background">
147 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></div>
148 <td align="right" class="headline">
149 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></td>
150 </tr>
151 </table>
152 <table width="100%" align="left">
153 <tr>
154 <td align="left" class="headline">
155 <div align="center" class="background">
156 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></div>
157 <td align="right" class="headline">
158 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></td>
159 </tr>
160 </table>
161 <table width="100%" align="left">
162 <tr>
163 <td align="left" class="headline">
164 <div align="center" class="background">
165 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></div>
166 <td align="right" class="headline">
167 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></td>
168 </tr>
169 </table>
170 <table width="100%" align="left">
171 <tr>
172 <td align="left" class="headline">
173 <div align="center" class="background">
174 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></div>
175 <td align="right" class="headline">
176 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></td>
177 </tr>
178 </table>
179 <table width="100%" align="left">
180 <tr>
181 <td align="left" class="headline">
182 <div align="center" class="background">
183 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></div>
184 <td align="right" class="headline">
185 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></td>
186 </tr>
187 </table>
188 <table width="100%" align="left">
189 <tr>
190 <td align="left" class="headline">
191 <div align="center" class="background">
192 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></div>
193 <td align="right" class="headline">
194 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></td>
195 </tr>
196 </table>
197 <table width="100%" align="left">
198 <tr>
199 <td align="left" class="headline">
200 <div align="center" class="background">
201 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></div>
202 <td align="right" class="headline">
203 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></td>
204 </tr>
205 </table>
206 <table width="100%" align="left">
207 <tr>
208 <td align="left" class="headline">
209 <div align="center" class="background">
210 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></div>
211 <td align="right" class="headline">
212 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></td>
213 </tr>
214 </table>
215 <table width="100%" align="left">
216 <tr>
217 <td align="left" class="headline">
218 <div align="center" class="background">
219 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></div>
220 <td align="right" class="headline">
221 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></td>
222 </tr>
223 </table>
224 <table width="100%" align="left">
225 <tr>
226 <td align="left" class="headline">
227 <div align="center" class="background">
228 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></div>
229 <td align="right" class="headline">
230 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></td>
231 </tr>
232 </table>
233 <table width="100%" align="left">
234 <tr>
235 <td align="left" class="headline">
236 <div align="center" class="background">
237 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></div>
238 <td align="right" class="headline">
239 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></td>
240 </tr>
241 </table>
242 <table width="100%" align="left">
243 <tr>
244 <td align="left" class="headline">
245 <div align="center" class="background">
246 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></div>
247 <td align="right" class="headline">
248 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></td>
249 </tr>
250 </table>
251 <table width="100%" align="left">
252 <tr>
253 <td align="left" class="headline">
254 <div align="center" class="background">
255 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></div>
256 <td align="right" class="headline">
257 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></td>
258 </tr>
259 </table>
260 <table width="100%" align="left">
261 <tr>
262 <td align="left" class="headline">
263 <div align="center" class="background">
264 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></div>
265 <td align="right" class="headline">
266 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></td>
267 </tr>
268 </table>
269 <table width="100%" align="left">
270 <tr>
271 <td align="left" class="headline">
272 <div align="center" class="background">
273 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></div>
274 <td align="right" class="headline">
275 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></td>
276 </tr>
277 </table>
278 <table width="100%" align="left">
279 <tr>
280 <td align="left" class="headline">
281 <div align="center" class="background">
282 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></div>
283 <td align="right" class="headline">
284 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></td>
285 </tr>
286 </table>
287 <table width="100%" align="left">
288 <tr>
289 <td align="left" class="headline">
290 <div align="center" class="background">
291 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></div>
292 <td align="right" class="headline">
293 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></td>
294 </tr>
295 </table>
296 <table width="100%" align="left">
297 <tr>
298 <td align="left" class="headline">
299 <div align="center" class="background">
300 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></div>
301 <td align="right" class="headline">
302 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></td>
303 </tr>
304 </table>
305 <table width="100%" align="left">
306 <tr>
307 <td align="left" class="headline">
308 <div align="center" class="background">
309 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></div>
310 <td align="right" class="headline">
311 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></td>
312 </tr>
313 </table>
314 <table width="100%" align="left">
315 <tr>
316 <td align="left" class="headline">
317 <div align="center" class="background">
318 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></div>
319 <td align="right" class="headline">
320 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></td>
321 </tr>
322 </table>
323 <table width="100%" align="left">
324 <tr>
325 <td align="left" class="headline">
326 <div align="center" class="background">
327 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></div>
328 <td align="right" class="headline">
329 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></td>
330 </tr>
331 </table>
332 <table width="100%" align="left">
333 <tr>
334 <td align="left" class="headline">
335 <div align="center" class="background">
336 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></div>
337 <td align="right" class="headline">
338 <a href="http://www.tennisabstract.com/bio/2015/09/18/the-case-for-novak-djokovic-and-roger-federer-and-rafael-nadal/">here</a></td>
339 </tr>
340 </table>
341 <table width="100%" align="left">
342 <tr>
343 <td align="left" class="headline">
344 <div align="center" class="background">
345 <a href="http://www.tennisabstract.com/bio/2015/09/18
```

## Static Data and Dynamic Data

ESPN

STATISTICS
STATSGURU
RECORDS
SUPERSTATS
ASK STEVEN
NUMBERS GAME
WORLD CUP
WORLD

Number  
Crunching

229\*

The score by Australia's Belinda Clark, in a World Cup game against Denmark in 2007. It is the highest score in women's ODI's, and was the highest score in all ODI's (8 Right Sharma made 264 against Sri Lanka in 2014).

### Imperious Maxwell sets up thumping win

### Readers recommend

Curated Tweets by @ESPNcricinfo

**Arvind Mahesh**  
@arvindmahesh

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/cricket/articles...](#)

### Features and Analysis

#### Duckless Implosions, and more wickets than runs

**Ask Steven:** What is the record for most ducks in a T20 match?

#### Mooen and Stokes show England's all-round might

**Numbers Games:** Unlike most other teams, England have a very successful lower middle order.

#### When Moen trumped Sobers

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

**Steven Lynch** answers your trivia questions

#### When all 11 players had a bowl

**Ask Steven:** Did, maybe, the highest score made on T20's debut?

#### What score have you not got?

**Ask Steven:** Sachin Tendulkar was never dismissed or left unbatting on which number?

#### The boundary-free effort

**Numbers Game:** Has the extra outfielder affected scoring rates in the last ten overs of ODI's?

### Did you know?

Number of single-digit scores by the Australia batsmen against South Africa in Hobart was 16 - their joint-most in any Test. The last time before this they had 16 such scores in a Test was in 1932 at the Oval.

### Records

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

Player	Swi	Bowl	Inns	Team	Opposition	Ground	Match Date	Scorecard
CTB Turner	6	5/44	1	Australia	v England	Sydney	10 Feb 1888	<a href="#">Test # 27</a>
		7/43	2	Australia	v England	Sydney	10 Feb 1888	<a href="#">Test # 28</a>
		5/27	2	Australia	v England	Leam's	14 Jul 1888	<a href="#">Test # 28</a>
		5/26	4	Australia	v England	Leam's	16 Jul 1888	<a href="#">Test # 28</a>
		6/112	2	Australia	v England	The Oval	13 Aug 1888	<a href="#">Test # 29</a>

### Readers recommend

Curated Tweets by @ESPNcricinfo

**Danish Sal**  
@danihsal

This is such a good read! Thanks for placing this [Saharsh Monge](#) | [espn.cricinfo.com/story/1Jd59n](#)

### Readers recommend

Curated Tweets by @ESPNcricinfo

**Danish Sal**  
@danihsal

This is such a good read! Thanks for placing this [Saharsh Monge](#) | [espn.cricinfo.com/story/1Jd59n](#)

```

1  <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
2
3  <i!-- hostname= web03, edition= view.asp?articleid=68, country= us, cluster= us, created= 2020-10-01 21:56:52 -->
4
5  <html xmlns="http://www.w3.org/1999/xhtml" xmlns-h="http://www.facebook.com/2008/html"
6    xmlns-g="http://opengraphprotocol.org/schema/" xmlns-fs="http://developers.facebook.com/schema/">
7
8  <head>
9
10   <script type="text/javascript" var="af_startpt=[new Date()].getTime()"</script>
11
12   <meta name="application-name" content="EaglesBlogs" id="eagleblogs31105d6fbc">
13
14   <title>Cricket Statistics | Statguru | ESPNcricinfo</title>
15
16   <meta http-equiv="Content-Type" content="text/html;charset=utf-8">
17
18   <meta name="keywords" content="cricket statistics, statguru, cricket stats, cricket records, batting average,
19     cricket centuries" />
20
21   <meta name="new_keywords" content="cricket statistics, statguru, cricket stats, cricket records, batting
22     average, cricket centuries" />
23
24   <meta name="description" content="Cricket statistics and stories around cricket numbers on ESPN Cricinfo" />
25
26   <!-- IE 9 -->
27
28   <script language="javascript" type="text/javascript">
29
30     function fncLoadJsmpList(iScenario) {
31       fncLoadJsmpList();
32       window.external.msiSetMethodAddJsmpList("Highlights",
33         "https://www.espncricinfo.com/ci/content/current/url/1094910.html",
34         "https://a.espnodn.com/espcricinfo/favicon.ico")|window.external.msiSetMethodAddJsmpListItem("Run Order",
35         "https://www.espncricinfo.com/ci/content/current/url/1094910.html",
36         "https://www.espncricinfo.com/espcricinfo/favicon.ico")|window.external.msiSetMethodAddJsmpListItem("Both Ends",
37         "https://www.espncricinfo.com/ci/content/current/url/1094910.html",
38         "https://a.espnodn.com/espcricinfo/favicon.ico")|window.external.msiSetMethodAddJsmpListItem("MATCH DAT",
39         "https://www.espncricinfo.com/ci/content/current/url/1095560.html",
40         "https://a.espnodn.com/espcricinfo/favicon.ico")|window.external.msiSetMethodAddJsmpListItem("Inns",
41         "https://www.espncricinfo.com/ci/content/current/url/1093416.html",
42         "https://a.espnodn.com/espcricinfo/favicon.ico")|window.external.msiSetMethodAddJsmpListItem("WI V IND",
43         "https://www.espncricinfo.com/west-indies-vs-india-2017/content/current/series/1094920.html",
44         "https://www.espncricinfo.com/espcricinfo/favicon.ico")|window.external.msiSetMethodAddJsmpListItem("ENG v SA",
45         "https://www.espncricinfo.com/england-v-south-africa-2017/content/current/series/10931417.html",
46         "https://a.espnodn.com/espcricinfo/favicon.ico")|window.external.msiSetMethodAddJsmpListItem("MCC",
47         "https://www.espncricinfo.com/mcc-west-indies-tour-2017/content/current/series/1095393.html",
48         "https://a.espnodn.com/espcricinfo/favicon.ico")|window.external.msiSetMethodAddJsmpListItem("ZIM v ZIM",
49         "https://www.espncricinfo.com/zimbabwe-albania-2017/content/current/series/1094447.html",
50         "https://a.espnodn.com/espcricinfo/favicon.ico")|window.external.msiSetMethodShowJsmpList();
51     }
52
53     function fncClearJsmpList() {
54       window.external.msiSetMethodClearJsmpList();
55     }
56   </script>
57
58   <meta name="application-task" content="name=Live Scores;action=
59     https://www.espncricinfo.com/ci/engine/content/match/scores/live.html;icon=
60     https://a.espnodn.com/espcricinfo/favicon.ico"/>
61
62   <meta name="application-task" content="name=Latest News;action=
63     https://www.espncricinfo.com/ci/content/current/story/news.html;icon=
64     https://a.espnodn.com/espcricinfo/favicon.ico"/>
65
66   <meta name="application-task" content="name=Fixtures;action=
67     https://www.espncricinfo.com/ci/content/current/match/fixtures/index.html;icon=
68     https://a.espnodn.com/espcricinfo/favicon.ico"/>
69
70   <meta name="application-task" content="name=Results;action=
71     https://www.espncricinfo.com/ci/engine/content/match/scores/recent.html;icon=
72     https://a.espnodn.com/espcricinfo/favicon.ico"/>
73
74   <meta name="application-task" content="name=Photos;action=
75     https://www.espncricinfo.com/ci/content/current/image/index.html;icon=
76     https://a.espnodn.com/espcricinfo/favicon.ico"/>
77
78   <meta name="application-task" content="name=Audio/Video;action=
79     https://www.espncricinfo.com/ci/content/video/audio/index.html;icon=
80     https://a.espnodn.com/espcricinfo/favicon.ico"/>

```

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/"
## [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.

# Web Scrapping for Static Data in R

Use [http://tennisabstract.com/reports/atp\\_elo\\_ratings.html](http://tennisabstract.com/reports/atp_elo_ratings.html) as an example

- CSS selector: id = "reportable", class = "tablesorter"

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.

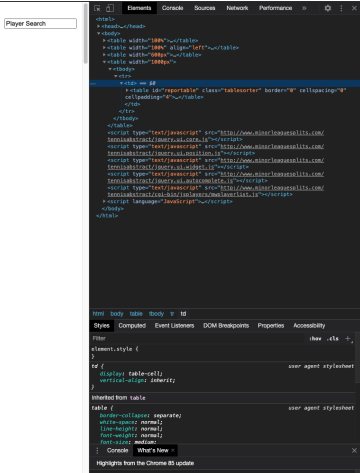
Unlike the official rankings, Elo ratings give credit for *who* you play, not the round or tournament in which you play them. I've written an extensive introduction to tennis Elo ratings [here](#).

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 "eElo," "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.

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

Row	Player	Age	Size	Handoff	Passes	Goalshare	NetR	cRR	gRR	Peak Match	Peak Age
1	Danish Dinkola	33	225L4	214.9	205.6A	203.9	218.1	217.3	213.4*	2019 Miami F	28.8
2	Harsh Patel	34	218L5	204.5	212.1	187.7	219.1	214.8	191.3	2009 Mumbai SP	29.2
3	Ravi Patel	38.5	217.0L	200.1	184.3	193.8	219.9	190.7	203.9	2007 Dubai F	25.6
4	Aravind Chinnai	38.9	209.2	198.9	198.9	198.9	216.4	196.4	197.9	2019 RIR	26.6
5	Aravind Chinnai	22.3	223.3	191.6	176.8A	157.6*	187.2	194.4	179.9	2020 Hamburg F	22.2
6	Aravind Chinnai	22.1	202.2	190.0	198.9	197.1	198.6	190.5	179.7	2020 Cincinnati R16	22.0
7	Daniel Mendez	24.8	203.0	195.4	187.1	194.25	188.1	182.8	181.3	2019 Shanghai F	23.3
8	Alexander Zverev	23.4	198.4	190.6	197.2	193.8	194.4	197.6	191.3	2017 Canada F	20.7
9	Aravind Chinnai	34.4	198.1	182.2	184.7	197.1	197.1	197.1	197.1	2017 Hamburg F	20.7
10	Aravind Chinnai	32.4	193.9	184.6	179.0	193.4	190.4	187.1	186.6	2015 Rotterdam R16	27.3
11	David Martell	34.1	194.5	181.2	170.0	154.1	182.9	169.5	179.3	2015 Monte Carlo GP	26.8
12	Felix Auger-Aliassime	28.3	183.6	184.1	164.7	139.3	188.5	181.9	184.3	2020 Rome R16	28.6
13	Danai Sotgiu	21.1	181.6	183.2	163.0	133.1	188.1	181.6	183.5	2020 Australian R16	20.7
14	Aravind Chinnai	35.2	181.1	169.4	169.4	169.4	186.1	186.1	179.1	2019 Wimbledon Open R16	32.3
15	Danai Sotgiu	35.1	182.3	174.3	160.4	149.7	183.7	190.6	171.1	2017 Beijing R16	37.1
16	Matteo Berrettini	24.4	182.1	179.1	162.9	177.2	184.6	184.6	184.1	2019 Vienna GP	23.9
17	Aravind Chinnai	24.1	182.0	169.5	173.2	150.8	189.9	184.1	174.9	2020 Abu Dhabi Cup R16	20.9
18	Aravind Chinnai	29.2	182.1	168.6	172.4	134.9	189.3	182.1	182.4	2017 Roland Garros R16	25.6
19	Aravind Chinnai	27.1	172.9	172.9	172.9	172.9	184.4	184.4	184.4	2017 Wimbledon SP	27.1
20	Aravind Chinnai	34.7	189.0	171.9	174.1	152.1	184.1	187.2	179.7	2009 Indian Wells H16	29.3
21	Mika Sestro	29.7	180.1	170.6	160.6	161.3	184.8	170.1	177.5	2016 Wimbledon SP	25.5
22	Danai Sotgiu	30.2	188.8	177.0	170.9	170.4	183.9	184.1	183.5	2017 Dubai R16	31.4
23	Mark Ochoa	23.1	183.5	170.0	162.9	144.5	184.5	174.9	183.5	2020 Cincinnati R16	23.0
24	Aravind Chinnai	35.6	172.2	176.4	169.6	159.2	182.2	186.2	175.1	2017 Wimbledon SP	35.6
25	Aravind Chinnai	30.2	167.3	160.4	160.4	143.5	183.8	173.9	168.2	2020 US Open R32	30.2
26	Aravind Chinnai	27.1	184.6	169.0	158.8	139.6	172.9	195.2	183.2	2020 Hamburg GP	27.1



# 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 "allo", "cilo", and "gilo". 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.

table#reportable.tablesorter 994x7042 99.28													
Rank	Player	Age	Elo	Hardflow	Clayflow	Grassflow	Hard	cElo	gElo	Peak Match	Peak Age	Peak Elo	
1	Nicola Pietrangeli	33.3	2255.4	2142.9	3005.6	2013.9	2199.1	2175.0	2134.7	2016 Miami F	28.6	2483.7	
2	Andrei Panatier	34.3	2186.0	2045.2	2111.2	1977.9	2176.1	2148.1	1817.4	2009 Madrid SF	29.6	2384.6	
3	Benjamin Boncompagni	36.5	2174.9	2051.7	1624.5	1933.8	2116.9	1807.1	2201.9	2007 Dubai F	25.8	2379.4	
4	Guillaume Lhuissier	27.0	2079.8	1989.5	2009.2	1814.3	2034.5	2044.5	1947.5	2016 Halle R16	22.9	2122.9	
5	Andreas Schick	22.8	2023.5	1910.8	1705.6	1516.4	1967.2	1904.5	1709.9	2020 Hamburg F	22.9	2023.5	
6	Barbara Schick	22.1	2023.2	1899.0	1899.9	1975.1	1980.6	1960.6	1797.6	2020 Cincinnati R16	22.9	2046.1	
7	Daniel Medvedev	24.6	2020.0	1954.1	1621.7	1842.5	1987.1	1820.8	1831.2	2016 Shanghai F	23.7	2128.7	
8	Alexander Zverev	23.4	1984.6	1904.6	1972.2	1936.0	1844.8	1978.4	1870.3	2017 Canada F	20.3	2147.0	
9	Jack Kruger	24.7	1983.2	1891.1	1622.2	1812.3	1837.2	1805.5	1717.5	2017 Cincinnati SF	22.3	2047.8	
10	Francesca Schiavone	32.4	1953.9	1984.9	1799.3	1815.4	1929.4	1971.6	1883.8	2016 Rotterdam R16	27.9	2073.8	
11	David Goffin	34.1	1940.5	1911.2	1070.0	1549.1	1825.6	1862.2	1744.3	2016 Monte Carlo GP	24.6	2007.2	
12	Elena Korneichuk	28.5	1935.4	1941.7	1847.6	1394.3	1888.5	1891.5	1847.3	2020 Rome R32	28.5	1935.4	
13	Daria Sharmayeva	21.4	1931.6	1832.0	1706.2	1336.1	1881.8	1818.8	1835.3	2020 Auckland R16	20.7	1973.5	
14	Sean Wenbin	36.5	1928.1	1907.1	1898.4	1932.5	1892.8	1893.2	1780.3	2016 Australian Open R32	30.8	2145.5	
15	Evgeny Korneichuk	28.1	1923.5	1742.8	1862.4	1448.7	1837.2	1805.5	1717.5	2016 Beijing R32	27.1	1945.1	
16	Matteo Berrettini	24.4	1923.9	1777.6	1862.9	1772.3	1943.8	1900.4	1847.5	2016 Vienna GP	23.5	2079.9	
17	Alex de Minaur	21.6	1920.7	1863.1	1363.2	1560.8	1891.6	1841.9	1740.7	2020 Atp Cup R8	20.9	1949.8	
18	Polina Garmonova-Buda	29.2	1912.6	1940.6	1712.4	1340.9	1879.3	1912.2	1826.4	2017 Roland Garros R16	26.9	1947.8	
19	Daniel Goffin	29.8	1888.1	1791.1	1757.9	1722.3	1844.6	1820.8	1810.2	2017 Rome R32	28.4	2037.7	
20	Jo-Wilfried Tsonga	34.7	1880.2	1791.9	1744.1	1626.1	1841.0	1817.2	1757.9	2009 Indian Wells R64	23.9	2126.2	
21	Maria Sharapova	29.7	1880.1	1759.5	1698.0	1561.3	1844.8	1770.1	1757.9	2016 Wimbledon SF	25.5	2153.3	
22	Maria Clay	32.9	1848.4	1775.0	1798.9	1785.4	1835.2	1841.5	1835.1	2010 Osaka R16	21.4	2046.0	
23	David Goffin	22.0	1843.0	1803.0	1612.9	1443.5	1843.5	1747.9	1653.2	2020 Cincinnati R16	23.0	1898.1	
24	Andriy Medvedev	33.3	1872.0	1798.4	1798.8	1629.4	1829.2	1809.0	1807.7	2017 Dubai SF	29.8	2346.5	
25	Yevgeny Pavlov	30.2	1871.3	1808.4	1280.4	1493.0	1839.4	1875.9	1682.2	2020 Us Open R32	30.2	1879.6	
26	Constant Huet	21.7	1868.0	1698.0	1536.5	1396.9	1732.0	1802.2	1832.4	2020 Hamburg GP	21.7	1876.0	

Player Search

The screenshot shows a web browser window with the URL `tennisabstract.com`. A search bar labeled "Player Search" is at the top. Below it is a table of player ratings. A right-click context menu is open over the table, showing options like "Copy", "Copy XPath", and "Copy full XPath". The "Copy XPath" option is highlighted. The menu also shows "Cut element", "Copy element", "Paste element", "Hide element", "Force state", "Break on", "Expand recursively", "Copy children", "Scroll into view", "Focus", "Store as global variable", and "Speech".



# 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](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.

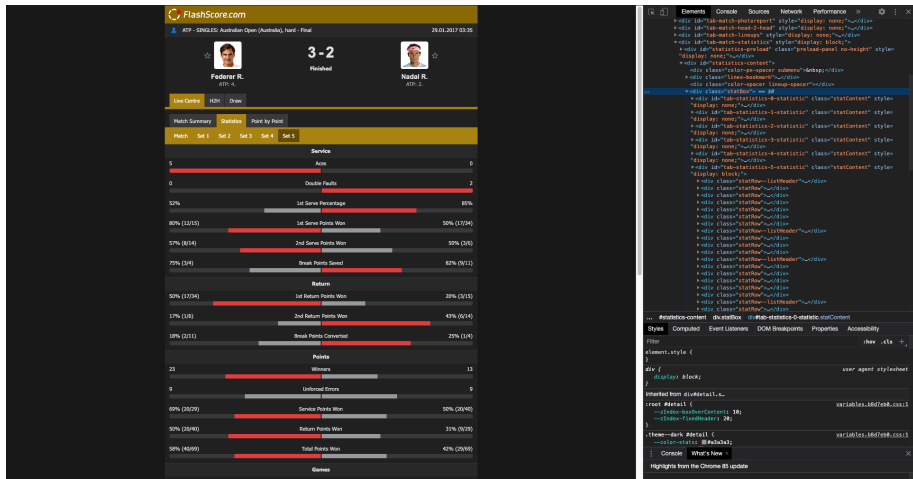
# Web Scraping for Dynamic Data in R

- 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 package

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

# Web Scrapping for Dynamic Data in R

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



# 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')))
```

```
## [1] "Service"      "20"           "Aces"         "4"
## [5] "3"            "Double Faults"
```

```
remDr$close()
```

# Web Scraping for Dynamic Data in R