

Big Data: Data Wrangling Boot Camp
Web Crawling with R and CSS

Chuck Cartledge, PhD

25 February 2018

Table of contents (1 of 1)

1 Intro.

2 CSS

- Background
- Examples

3 Hands-on

- US Senate Bills
- NASA Reports

4 Q & A

5 Conclusion

6 References

7 Files

What are we going to cover?

- Look to the future
- **Data wrangle** static Web pages from different sources
- Explore a few of the mysteries of CSS
- Understand how to download web pages



What is HTML?[3]

HTML is the standard markup language for creating Web pages.

- HTML stands for Hyper Text Markup Language
- HTML describes the structure of Web pages using markup
- HTML elements are the building blocks of HTML pages
- HTML elements are represented by tags
- HTML tags label pieces of content such as "heading", "paragraph", "table", and so on
- Browsers do not display the HTML tags, but use them to render the content of the page

A Simple HTML Document



Image from [3].

HTML tags have attributes

Attributes help the browser to display tag related information “correctly.”

Correctly can be dependent on which browser is being used.

All HTML Attributes

Attribute	Belongs to	Description
accept	<code><input></code>	Specifies the types of files that the server accepts (only for type="file")
accept-charset	<code><form></code>	Specifies the character encodings that are to be used for the form submission
accesskey	Global Attributes	Specifies a shortcut key to activate/focus an element
action	<code><form></code>	Specifies where to send the form-data when a form is submitted
align	Not supported in HTML 5.	Specifies the alignment according to surrounding elements. Use CSS instead
alt	<code><area></code> , <code></code> , <code><input></code>	Specifies an alternate text when the original element fails to display
async	<code><script></code>	Specifies that the script is executed

https://www.w3schools.com/tags/ref_attributes.asp

How to manage the “look” of a company’s web pages?

HTML attributes can be used for “branding.” [1]

- Website Content
- Overall Design and Layout
- Use of Innovation



Image from [2].

We’re going to focus on design and layout.

Managing HTML attributes

- HTML attributes on a few web pages can be managed by hand.
- HTML attributes on a few 10s of web pages might be managed by hand.
- HTML attributes on 1,000s of web pages can not be managed by hand.
- Cascading Style Sheets (CSS) are designed to work with HTML pages.
- HTML attributes can be written into the page, or can take values from a CSS.

The browser renders the page based on the HTML attribute values

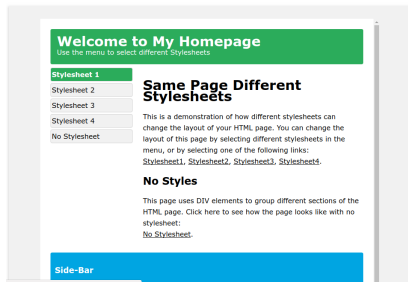
Same HTML, different CSSs

To do:

- 1 Load the page
- 2 Locate and press the "Stylesheet ..." button
- 3 Repeat as desired

CSS Demo - One HTML Page - Multiple Styles!

Here we will show one HTML page displayed with four different stylesheets. Click on the "Stylesheet 1", "Stylesheet 2", "Stylesheet 3", "Stylesheet 4" links below to see the different styles:



https://www.w3schools.com/css/css_intro.asp

The same HTML tags can be rendered differently based on the CSS.

Sample stylesheet 1.

Welcome to My Homepage

Use the menu to select different Stylesheets

Stylesheet 1

Stylesheet 2

Stylesheet 3

Stylesheet 4

No Stylesheet

Same Page Different Stylesheets

This is a demonstration of how different stylesheets can change the layout of your HTML page. You can change the layout of this page by selecting different stylesheets in the menu, or by selecting one of the following links:

[Stylesheet1](#), [Stylesheet2](#), [Stylesheet3](#), [Stylesheet4](#).

No Styles

This page uses DIV elements to group different sections of the HTML page. Click [here](#) to see how the page looks like with no stylesheet:

[No Stylesheet](#).

Side-Bar

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi.

Sample stylesheet 2.

Welcome to My Homepage

Use the menu to select different Stylesheets

Same Page Different Stylesheets

This is a demonstration of how different stylesheets can change the layout of your HTML page. You can change the layout of this page by selecting different stylesheets in the menu, or by selecting one of the following links:

[Stylesheet1](#), [Stylesheet2](#), [Stylesheet3](#), [Stylesheet4](#).

No Styles

This page uses DIV elements to group different sections of the HTML page. Click here to see how the page looks like with no stylesheet:

[No Stylesheet.](#)

Side-Bar

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et justo odio dignissim qui blandit praesent luptatum zzril delenit augue dui dolore te feugiat nulla facilis.

[Stylesheet 1](#)[Stylesheet 2](#)[Stylesheet 3](#)[Stylesheet 4](#)[No Stylesheet](#)

Sample stylesheet 3.

Welcome to My Homepage

Use the menu to select different Stylesheets

[Stylesheet 1](#)[Stylesheet 2](#)[Stylesheet 3](#)[Stylesheet 4](#)[No Stylesheet](#)

Same Page Different Stylesheets

This is a demonstration of how different stylesheets can change the layout of your HTML page. You can change the layout of this page by selecting different stylesheets in the menu, or by selecting one of the following links:

[Stylesheet1](#), [Stylesheet2](#), [Stylesheet3](#), [Stylesheet4](#).

No Styles

This page uses DIV elements to group different sections of the HTML page. Click here to see how the page looks like with no stylesheet:

[No Stylesheet](#).

Side-Bar

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi.

Sample stylesheet 4.

Welcome to My Homepage
Use the menu to select different Stylesheets

- [Stylesheet 1](#)
- [Stylesheet 2](#)
- [Stylesheet 3](#)
- **[Stylesheet 4](#)**
- [No Stylesheet](#)

Side-Bar

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

Same Page Different Stylesheets

This is a demonstration of how different stylesheets can change the layout of your HTML page. You can change the layout of this page by selecting different stylesheets in the menu, or by selecting one of the following links:

[Stylesheet1](#), [Stylesheet2](#), [Stylesheet3](#), [Stylesheet4](#).

No Styles

This page uses DIV elements to group different sections of the HTML page. [Click here](#) to see how the page looks like with no stylesheet:

[No Stylesheet](#).

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et justo odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugiat nulla facilisis.

Without a stylesheet.

Welcome to My Homepage

Use the menu to select different Stylesheets

- Stylesheet 1
- Stylesheet 2
- Stylesheet 3
- Stylesheet 4
- No Stylesheet

Same Page Different Stylesheets

This is a demonstration of how different stylesheets can change the layout of your HTML page. You can change the layout of this page by selecting different stylesheets in the menu, or by selecting one of the following links:

[Stylesheet1](#), [Stylesheet2](#), [Stylesheet3](#), [Stylesheet4](#).

No Styles

This page uses DIV elements to group different sections of the HTML page. Click here to see how the page looks like with no stylesheet:

[No Stylesheet](#).

Side-Bar

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.

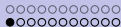
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exercitation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi.

Terminology

- HTML uses tags and attributes. These are sometimes called “elements.”
- CSS uses selectors and declarations.
- CSS declarations have two fields: property and value.
- CSS selectors are used to “find” (or select) HTML elements based on their name, id, class, attribute, and more¹.

Amazon reports over 6,000 CSS books.

¹https://www.w3schools.com/css/css_syntax.asp



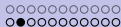
Examples

Start somewhere easy:

Bring up the ODU College of Continuing Education and Professional Development boot camp web page.

The screenshot shows a web browser displaying the Old Dominion University website. The page is titled "College of Continuing Education & Professional Development Bootcamps". The navigation menu includes "Future Students", "Current Students", "More...", "Media", and "Faculty & Staff". The main content area features a large image of a globe and the text "College of Continuing Education & Professional Development Bootcamps". Below this, there is a section for "ODU Weekend Bootcamps" with a description of the program and contact information for the College of Continuing Education & Professional Development. A yellow callout box with the text "LET'S TALK!" is overlaid on the page.

[https://www.odu.edu/cepd/
bootcamps](https://www.odu.edu/cepd/bootcamps)



Examples

Look at the HTML tags (elements)

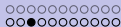
Right click on the page and select “View page source”

```

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
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
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000

```

[https://www.odu.edu/cepd/
bootcamps](https://www.odu.edu/cepd/bootcamps)



Examples

Find the style sheet references

Execute a CTRL-F to locate the string: stylesheet

```

<!--script type="text/javascript" src="/ajax/jquery.js" -->
<script src="https://code.jquery.com/jquery-3.2.1.min.js"></script>
<script src="/etc/designs/odu/clientlibs/libs/jquery-migrate-3.0.0.js" type="text/javascript"></script>
<script type="text/javascript">
  if (typeof jQuery == "undefined") {
    (function ($) {
      $.fn.jquery = "3.2.1";
      $.migrate = "3.0.0";
    })(jQuery);
  }
</script>
<script type="text/javascript">
  (function ($) {
    $.browser = {};
    function ua() {
      $.browser.msie = false;
      $.browser.version = 0;
      if (/msie ([0-9]+)/.test(ua())) {
        $.browser.msie = true;
        $.browser.version = RegExp.$1;
      }
    }
    ua();
  })(jQuery);
</script>
<link rel="stylesheet" href="/etc/designs/odu/clientlibs/libs/slick.min.css" type="text/css">
<link rel="stylesheet" href="/etc/designs/odu/clientlibs/fonts/awesome.min.css" type="text/css">
<link rel="stylesheet" href="/etc/designs/odu/clientlibs/min.css" type="text/css">
<script type="text/javascript" src="/etc/designs/odu/clientlibs/libs/sockjs.min.js"></script>
<script type="text/javascript" src="/etc/designs/odu/clientlibs/min.js"></script>
<link href="/etc/designs/odu/css" rel="stylesheet" type="text/css">
<link rel="stylesheet" type="text/css" href="/fonts/fontawesome/css/
<link rel="stylesheet" type="text/css" href="/maxcdn.bootstrapcdn.com/font-awesome/4.6.2/css/font-awesome.min.css" />
<!--[if lt IE 9]>
<script src="/cdnjs.cloudflare.com/ajax/libs/html5shiv/3.7.3/html5shiv.js"></script>
</endif-->
<link rel="icon" type="image/vnd.microsoft.icon" href="/etc/designs/odu/favicon.ico">
<link rel="shortcut icon" type="image/vnd.microsoft.icon" href="/etc/designs/odu/favicon.ico">
<title>College of Continuing Education & Professional Development Bootcamps - Old Dominion University</title>
</head>
<body>
<script>
  var names = new Array();
  var imgurls = new Array();
  var generatedLinks = new Array();
  </script>

```

[https://www.odu.edu/cepd/
bootcamps](https://www.odu.edu/cepd/bootcamps)



Looking at a CSS

Click on one of the links near the string “stylesheet”

```

/*!
 * Font Awesome 4.7.0 by @davegandy - http://fontawesome.io - @fontawesome
 * License - http://fontawesome.io/license (Font: SIL OFL 1.1, CSS: MIT License)
 */
@font-face {font-family: 'Font Awesome'; src:url('fontawesome-webfont.woff2?v=4.7.0') format('embedded-
opentype'),url('fontawesome/fonts/fontawesome-webfont.woff?v=4.7.0')
format('woff2'),url('fontawesome/fonts/fontawesome-webfont.woff?v=4.7.0')
format('woff'),url('fontawesome/fonts/fontawesome-webfont.ttf?v=4.7.0')
format('truetype'),url('fontawesome/fonts/fontawesome-webfont.svg?v=4.7.0fontawesome-regular')} font-
weight:normal;font-style:normal;
-@{display:inline-block;font:normal normal normal 14px/1 FontAwesome;font-size:inherit;text-rendering:auto;-webkit-font-
smoothing:antialiased;-ms-font-smoothing:grayscale}
.fg-lg{font-size:1.3333333em;line-height:1.2em;vertical-align:middle}
.fg-2x{font-size:2em}
.fg-3x{font-size:3em}
.fg-4x{font-size:4em}
.fg-5x{font-size:5em}
.fg-fw{width:1.28571428em;text-align:center}
.fg-l{padding-left:0;margin-left:1.285714em;list-style-type:none}
.fg-ul{position:relative}
.fg-li{position:absolute;left:2.1428574em;width:1.285714em;top:1428574em;text-align:center}
.fg-li.fg-li-left{-1.8571428em}
.fg-border{padding:2em 20em .15em;border:solid .40em #000;border-radius:1em}
.fg-pull-left{float:left}
.fg-pull-right{float:right}
.fg-pull-left{margin-right:.3em}
.fg-pull-right{margin-left:.3em}
.pull-right{float:right}
.pull-left{float:left}
.fg-pull-left{margin-right:.3em}
.fg-pull-right{margin-left:.3em}
.fg-spin{-webkit-animation:fa-spin 2s infinite linear;animation:fa-spin 2s infinite linear}
.fg-pulse{-webkit-animation:fa-spin 1s infinite steps(8);animation:fa-spin 1s infinite steps(8)}
@-webkit-keyframes fa-spin{0%{-webkit-transform:rotate(0);transform:rotate(0)}
30%{-webkit-transform:rotate(35deg);transform:rotate(35deg)}
60%{-webkit-transform:rotate(70deg);transform:rotate(70deg)}
90%{-webkit-transform:rotate(105deg);transform:rotate(105deg)}
}
.fg-rotate-90{-ms-filter:'progid:DXImageTransform.Microsoft.BasicImage(rotation=1)';-webkit-transform:rotate(90deg);ms-
transform:rotate(90deg);transform:rotate(90deg)}
.fg-rotate-180{-ms-filter:'progid:DXImageTransform.Microsoft.BasicImage(rotation=2)';-webkit-transform:rotate(180deg);ms-
transform:rotate(180deg);transform:rotate(180deg)}
.fg-rotate-270{-ms-filter:'progid:DXImageTransform.Microsoft.BasicImage(rotation=3)';-webkit-transform:rotate(270deg);ms-
transform:rotate(270deg);transform:rotate(270deg)}
.fg-flip-horizontal{-ms-filter:'progid:DXImageTransform.Microsoft.BasicImage(rotation=0, mirror=1)';-webkit-
transform:scale(1, -1);-ms-transform:scale(1, -1);transform:scale(1, -1)}
.fg-flip-vertical{-ms-filter:'progid:DXImageTransform.Microsoft.BasicImage(rotation=2, mirror=1)';-webkit-
transform:scale(1, 1);-ms-transform:scale(1, 1);transform:scale(1, 1)}
.fg-stack{-display:inline-block;vertical-align:middle}
.fg-stack-1{-font-size:2em}
.fg-stack-2{-font-size:3em}

```

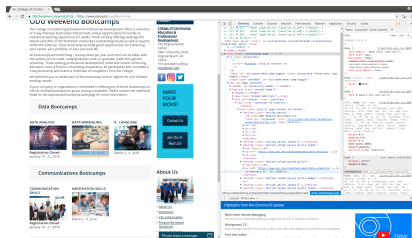
[https://www.odu.edu/etc/
designs/odu/clientlibs/
libs/fontawesome4.min.css](https://www.odu.edu/etc/designs/odu/clientlibs/libs/fontawesome4.min.css)

CSS are meant for computer consumption, not humans.

We'll be picking apart a page to find the CSS selectors

A few steps using Google Chrome.

- 1 Open this url:
<https://www.odu.edu/cepd/bootcamps>
- 2 Press F12 (the debug key)
- 3 Inside the HTML source, drop down and expand until our class date is selected on the left



This is doable, but there are easier ways.



Examples

Same image.

ODU weekend bootcamps

The College of Continuing Education & Professional Development offers a collection of 3-day intensive bootcamps that provide unique opportunities for hands-on interactive learning experiences for adults. These exciting offerings leverage the talents and skills of Old Dominion University's outstanding faculty as well as experts within the industry. These bootcamps provide great opportunities for enhancing your career, your portfolio, or your personal life.

All bootcamps are held Friday-Sunday twice per year and most can be taken with the options of noncredit, undergraduate credit, or graduate credit through the university. Those seeking professional development credit will receive Continuing Education Units (CEUs) for completing a bootcamp. All participants who complete a 3-day bootcamp will receive a certificate of completion from the College.

We welcome you to review each of the bootcamps and to register for one of these exciting classes.

If your company or organization is interested in offering any of these bootcamps on-site for multiple participants, group pricing is available. Please contact the individual listed on the appropriate bootcamp web page for more information.

Data Bootcamps

DATA ANALYSIS	DATA WRANGLING	R - LANGUAGE
 Registration Closed January 19 - 21, 2018	 February 23 - 25, 2018	 March 2 - 4, 2018

Communications Bootcamps

COMMUNICATION SKILLS	NEGOTIATION SKILLS
 Registration Closed January 19 - 21, 2018	 March 2 - 4, 2018

MAKE YOUR MOVE!

Contact Us

Join Our E-Mail List

```

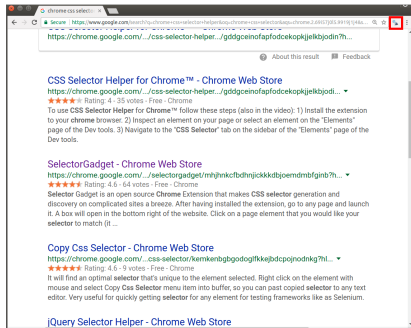
[1] Elements Console Sources Network Performance Memory Application
[2] Styles Computed Event Listeners M
Filter: new .css +
#content.style {
  body {
    font-family: "sans-serif";
    font-size: 14px;
    margin: 0;
    padding: 0;
    text-align: right;
  }
  body, select, input, #content {
    color: #444;
  }
  body {
    font: 13px/1.231 sans-serif;
  }
  html, body, div, span, #content {
    border: 1px solid #ccc;
    padding: 5px;
    margin: 0;
    text-align: right;
    font-family: "sans-serif";
    font-size: 14px;
    margin: 0;
    padding: 0;
    text-align: right;
  }
  body {
    font: 13px/1.231 sans-serif;
  }
  user-agent stylesheet
  margin: 0;
}
#content {
  font-family: "sans-serif";
  font-size: 14px;
  margin: 0;
  padding: 0;
  text-align: right;
}

```

This is doable, but there are easier ways.

Chrome has a free plug-in: SelectorGadget

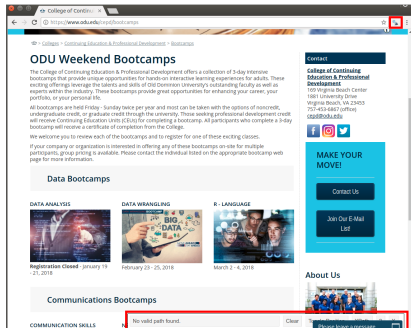
- The plug-in is free.
- A new icon is added to the browser



See the red rectangle in the image.

How to use the SelectorGadget? (1 of 4)

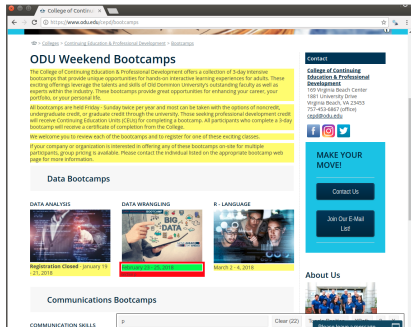
- 1 Load the page of interest.
- 2 Click the icon (small square at top of page).
- 3 See the pop-up windows (wide rectangle at bottom of page)



Press the gadget icon a second time to turn it off.

How to use the SelectorGadget? (2 of 4)

- Click on the “item” of interest. The item will turn green. The bottom pop-up will have the CSS selector that identifies what was clicked.



All other HTML elements that match the CSS selector will turn yellow.

How to use the SelectorGadget? (3 of 4)

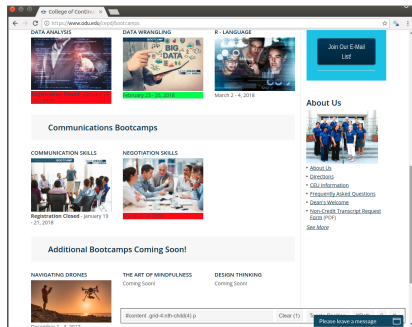
- 1 Click on the “item” **NOT** of interest. The item will be outlined in red.
- 2 Repeat the above step until only the “item” of interest is highlighted. This may take several iterations, and scrolling up and down the page to ensure only the items of interest are highlighted.

The screenshot shows a web browser window displaying the 'ODU Weekend Bootcamps' page. The page title is 'ODU Weekend Bootcamps'. Below the title, there is a description of the bootcamps and a list of bootcamps. The 'Data Bootcamps' section is expanded, showing three bootcamps: 'DATA ANALYSIS', 'DATA WRANGLING', and 'R - LANGUAGE'. The 'DATA ANALYSIS' bootcamp is highlighted with a red border, and the 'DATA WRANGLING' bootcamp is highlighted with a green border. The 'DATA ANALYSIS' bootcamp has a red box around the text 'Registration Closed (January 11, 2018)'. The 'R - LANGUAGE' bootcamp has a date 'March 2 - 4, 2018'. Below the 'Data Bootcamps' section, there is a 'Communications Bootcamps' section. The page also includes a 'Contact Us' button, a 'Join Our E-Mail List' button, and an 'About Us' section.

How to use the SelectorGadget? (4 of 4)

Once only the “items” of interest are green:

The gadget contains the CSS selector. In this case, it is:
`#content .grid-4:nth-child(4) p`
 Where spaces, and case are important.



The selector is what we are after.

Getting a field of data.

Copy and paste these lines into RStudio:

```
library(xml2)
library(rvest)

url <- "https://www.odu.edu/cepd/bootcamps"
selector <- "#content .grid-4:nth-child(4) p"

webPage <- read_html(url)

data <- html_text(html_nodes(webPage, selector))

data
```

The result is: February 23 - 25, 2018

An interesting question.

Our numerically inclined supervisor has an interest in politics. Specifically, can bi-partisan support be quantified?



How can we use R, CSS selectors, and a little math to answer the question?

Where/how to begin?

We'll need to do the following:

- 1 Get a list of all the Senate bills
- 2 Get each bill's sponsor (there will only be one)
- 3 Get each bill's cosponsor (there may be none, or a bunch)
- 4 Get each member's party (there are three)
- 5 See how often members of different parties are on the same bill

All of this leading to a data structure, indexed by bill number, where each bill has a list containing the sponsor and any cosponsors in that order.

Where/how to get a list of Senate bills?

This is the first step in our search:
<https://www.congress.gov/>
 and pressing the “Introduced” link.

The screenshot shows the homepage of CONGRESS.GOV. At the top, there are navigation links for 'Legislation', 'Congressional Record', 'Committees', and 'Members'. Below this is a search bar with 'Congress, H.R., search, "health care"' entered. A dropdown menu for 'Current Legislation' is open, showing a list of bills: H.R. 11086 (110th) Am. Act to provide for reauthorization pursuant to H.R. 5 and V of the Consolidated Appropriation Act for the fiscal year 2015, H.R. 10110 (110th) Federal Register Printing Schedule Act of 2017, and H.R. 10109 (110th) Consolidated Carry Reciprocity Act of 2017. On the right side of the dropdown, there are links for 'All Bills', 'House', 'Senate', 'Amend. Legislation (Democrat)', and 'Introduction'. The 'Introduced' link is highlighted with a red box. Below the search bar, there are sections for 'Most Viewed Bills', 'Current Legislative Activities', 'House of Representatives', 'Senate', 'Recent', and 'Contact Your Member'.

Limiting the search to Senate bills.

Using the filters on the left, limit the search to the 115th Congress, originated in the Senate, and is a Bill.

There are 2,328 bills introduced. Far too many to fit on a screen.

The screenshot shows the CONGRESS.GOV legislative search interface. The search filters on the left are set to '115th Congress', 'Senate', and 'Bill'. The search results show 1,108 of 2,328 bills per page. The first two results are highlighted with red boxes:

- Bill 1** (S. 2337) - 115th Congress (2017-2018)
 - Congress:** 115th Congress (2017-2018)
 - Check at:** 115 (2017-2018) 0/100
 - 114 (2015-2016)** 0/100
 - Bill Type:**
 - Check at:**
 - 115 (2017-2018)** 0/100
 - Amendments (P.A.M.C. or H.A.M.C.)** 0/100
 - Resolutions (H. Res. or S. Res.)** 0/100
 - Joint Resolutions (J. Res. or S. J. Res.)** 0/100
- Bill 2** (S. 2338) - 115th Congress (2017-2018)
 - Congress:** 115th Congress (2017-2018)
 - Check at:** 115 (2017-2018) 0/100
 - 114 (2015-2016)** 0/100
 - Bill Type:**
 - Check at:**
 - 115 (2017-2018)** 0/100
 - Amendments (P.A.M.C. or H.A.M.C.)** 0/100
 - Resolutions (H. Res. or S. Res.)** 0/100
 - Joint Resolutions (J. Res. or S. J. Res.)** 0/100

Looking at the page:

Some details about the page:

- The maximum number of results per page is 250.
- There are navigation controls at the bottom of the page.
- The only apparent difference between pages, is the value of the last argument.
- Changing the value of “page” to something extreme results in the last valid page



Sample URLs

URLs to get pages 3 and 5.
Lines broken for readability

```
https://www.congress.gov/search?searchResultViewType=expanded&
pageSort=dateOfIntroduction%3Adesc&q=%7B%22congress%22%3A%22115%22
%2C%22chamber%22%3A%22Senate%22%2C%22type%22%3A%22bills%22%7D&page=3
https://www.congress.gov/search?searchResultViewType=expanded&
pageSort=dateOfIntroduction%3Adesc&q=%7B%22congress%22%3A%22115%22
%2C%22chamber%22%3A%22Senate%22%2C%22type%22%3A%22bills%22%7D&page=5
```

So:

- 1 We can use SelectorGadget to identify the Bill number
- 2 We can loop across all pages, starting at 1 and continuing until we get the same page twice
- 3 Collect all the Bill numbers as we go along



Lets open the attached file, and have some fun.

The screenshot shows the RStudio environment with the following components:

- Source Editor:** Contains an R script named 'politicalPartyBills.R' with the following code:


```

1 rm(list=ls())
2
3 library(bltops)
4 library(RCurl)
5 library(xml2)
6 library(rvest)
7 library(tools)
8
9 crossParties <- function(sponsors)
10 {
11   partyAffiliation <- function(s)
12     {
13       regexpr("[", s, fixed=TRUE)[1] + 1
14     }
15   originators <- c()
16   party <- c()
17 }
18
19
```
- Environment Pane:** Currently empty.
- Console:** Shows the R startup message:


```

~Teaching/DataWrangling/2018-Spring/Presse
locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publi
cations.

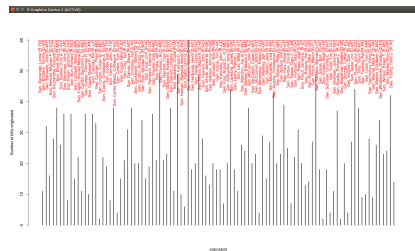
Type 'demo()' for some demos, 'help()' for on-line h
elp, or
'help.start()' for an HTML browser interface to help
.
Type 'q()' to quit R.

> |
```

How a bill originator histogram may look

Things of interest:

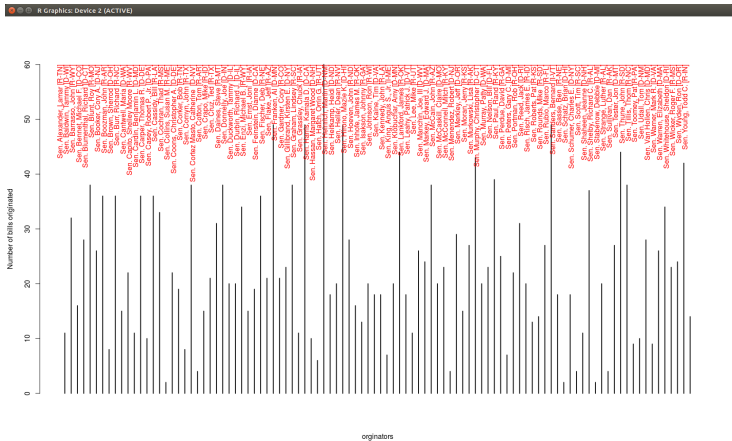
- Orrin Hatch's office is really, really busy.
- Cochran's and Schalz's office, not so much. (Makes you wonder.)



Data is time sensitive, and
deserves greater examination.



Same image.



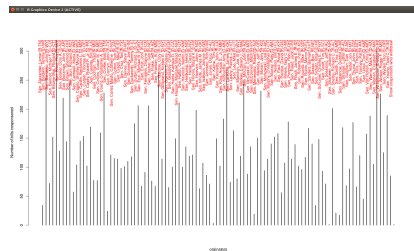
Data is time sensitive, and deserves greater examination.

Senate bill cosponsors.

Things of interest:

- There are some senators who are very popular.
- There are others, not so much.

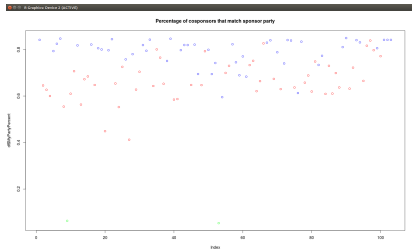
It would be interesting to “normalize” sponsorship by length of time in the congress.



Data is time sensitive, and deserves greater examination.

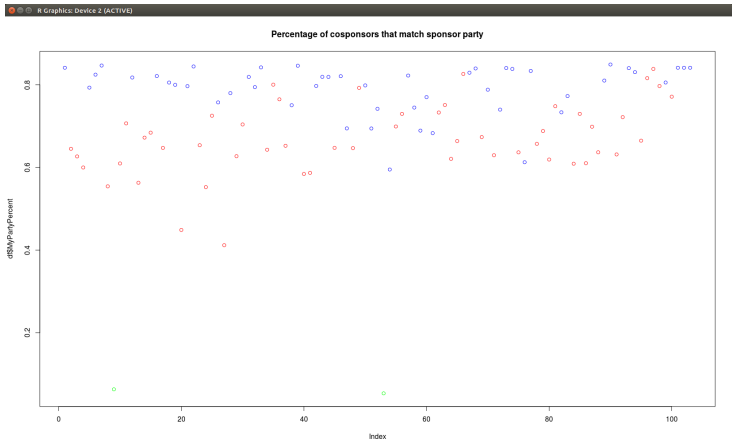
Given my party, what percentage of cosponsors match my party?

There is almost a division.
Republicans are red.
Democrats are blue.
Independents are green.



By and large, people stay within their party.

Same image.



By and large, people stay within their party.

Now that we have looked at the data, what other questions could be asked and answered?

In no particular order:

- 1 Who are the most influential senators? Where influential is measured by how many bills they are directly, or indirectly associated with.
- 2 Are senators from the same region more likely to work together than those not from the same region? (Think about senators around the Chesapeake Bay.)
- 3 What are individual senators really interested in? (Think about looking at the text of their sponsored and cosponsored bills to understand what they put their names to.)
- 4 Should be lots of others.

Q & A time.

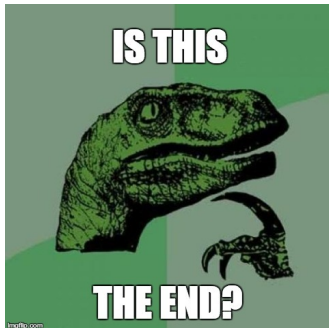
Q: How many bureaucrats does it take to screw in a light bulb?

A: Two. One to assure everyone that everything possible is being done while the other screws the bulb into the water faucet.



What have we covered?

- Explored a little bit of cascading style sheets (CSS)
- Used R to download web pages
- Used R to extract data based on CSS selectors



Next: Exploring the wild and woolly Web world.

References (1 of 1)

- [1] Corey Smith, [HOW TO USE YOUR WEBSITE AS A BRANDING TOOL](https://www.tributemedia.com/blog/using-your-website-branding-tool), <https://www.tributemedia.com/blog/using-your-website-branding-tool>, 2013.
- [2] OBsurvey Staff, [Most effective distribution methods to get answers](http://obsurvey.com/blog/most-effective-distribution-methods-to-get-answers/), <http://obsurvey.com/blog/most-effective-distribution-methods-to-get-answers/>, 2014.
- [3] W3 Staff, [What is HTML?](https://www.w3schools.com/html/html_intro.asp), https://www.w3schools.com/html/html_intro.asp, 2017.

Files of interest

- 1 Political parties 