Stat 240 Week 9
Data from the web and putting analytics onto the web

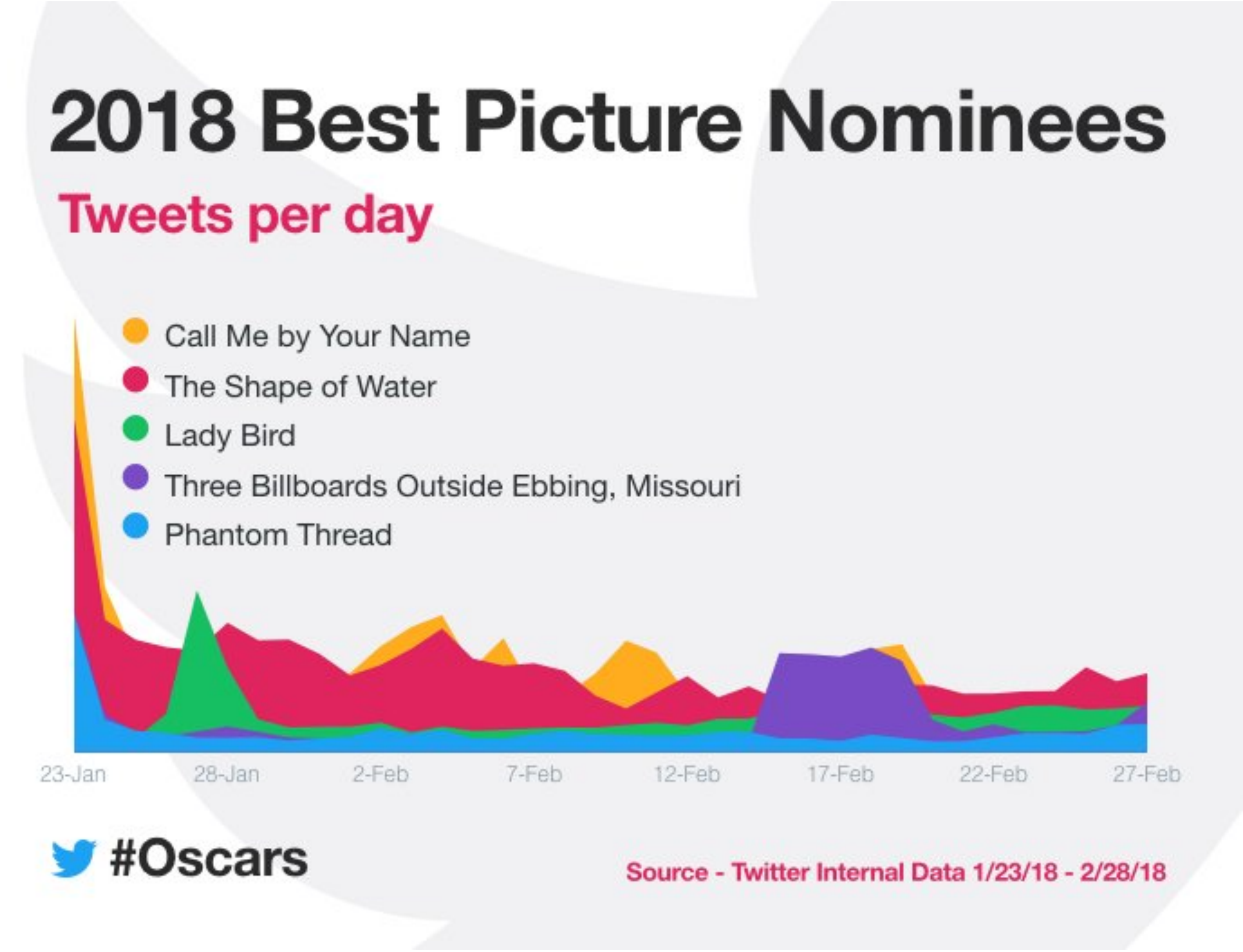
Week 9
Dr. Dave Campbell

Twitter and the Oscars
https://twitter.com/TwitterData

Winner: “Shape of water”

Oscars were held March 4th

Image posted March 1



Source - Twitter Internal Data 1/23/18 - 2/28/18

Local conference
bctechsummit.ca

Hashtag usage: #bctechsummit

```
plot(UniqueTweetsHashtag1$created)
```

Retweet count per vs time

```
plot(UniqueTweetsHashtag1$created,UniqueTweetsHashtag1$retweetCount,    main = "retweet count over time",ylab="count of retweets",xlab="date")
```

Wordclouds from a specific day

Undergraduate Mathematics and Statistics Conference

https://www.bcumsc.org

SSC Case study competition

Meeting: https://ssc.ca/en/meeting/annual/2019

Competition: https://ssc.ca/en/meeting/annual/2019/case-studies-data-analysis-competition

Show me some insights about Canada’s National Parks

What is the lifetime of a university textbook?

How do we answer these question?

What kind of information do we need?

Webscraping

Best: find data already nicely formatted

Next Best: use API

Next next best: Parse HTML

Next next next best: Web scraping, it is fragile

Worst: visit webpage by hand and write things on paper, then input them into Excel

robots.txt

Some sites do not want you to scrape them

http://www.robotstxt.org/robotstxt.html

http://www.sfu.ca/robots.txt

https://en.wikipedia.org/robots.txt

https://twitter.com/robots.txt

https://postsecret.com/robots.txt

robots.txt

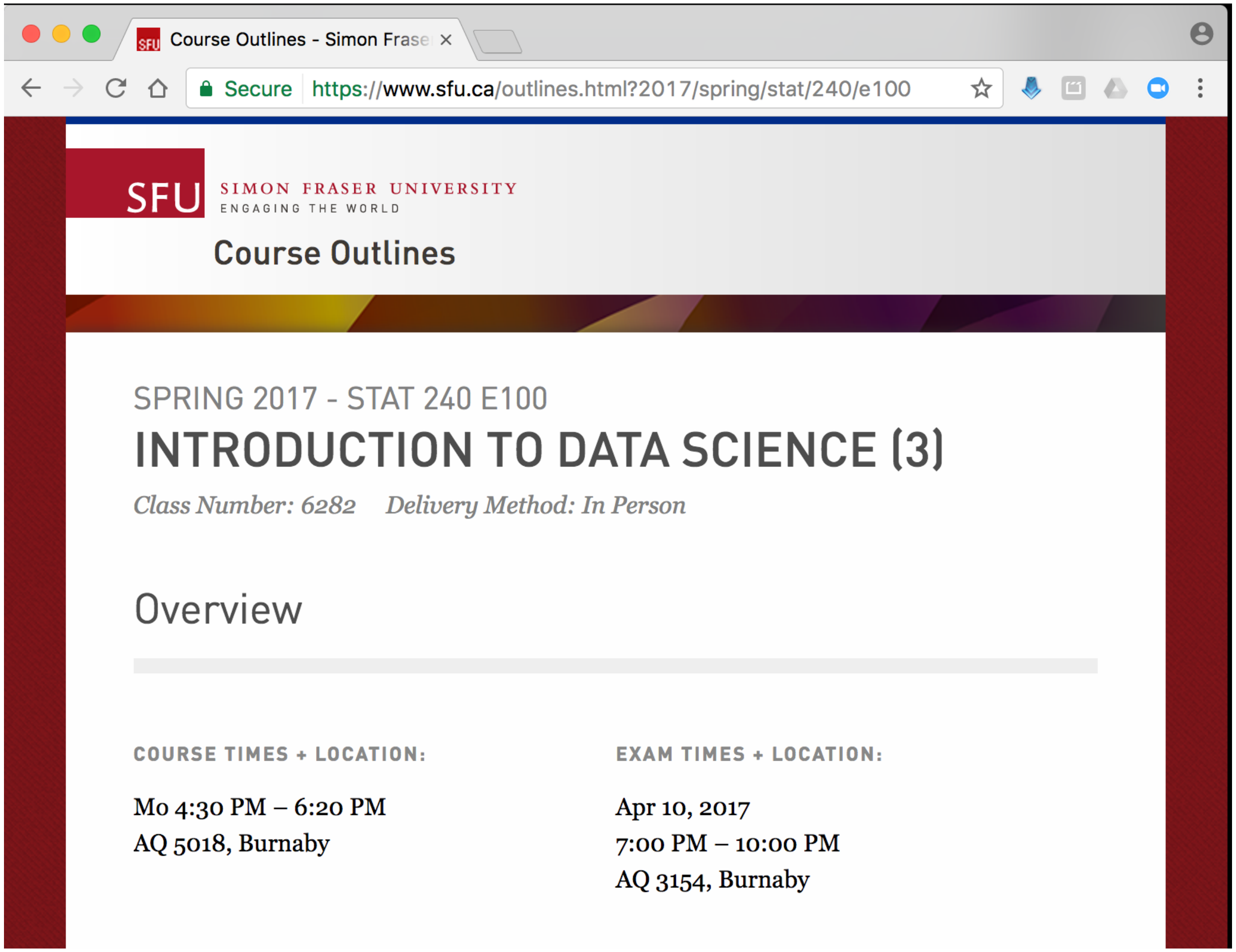
```
User-agent: *   <—- For all user agents

Disallow: /     <—- All pages are disallowed


User-agent: *   <—- For all user agents

Disallow:       <—- No pages are disallowed
```

HTML

```
<p>Directly parsing Canada's <a href="https://
en.wikipedia.org/wiki/
List_of_National_Parks_of_Canada"
title="National Parks of Canada">National
Parks</a> table should be much easier than
what I did in class.  But that’s step 2</p>
```

Visit a page

https://www.sfu.ca/outlines.html?2017/spring/stat/240/e100



Course Outlines - Simon Frase
Secure
https://www.sfu.ca/outlines.html?2017/spring/stat/240/e100
SFU
SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

Course Outlines

SPRING 2017 - STAT 240 E100

INTRODUCTION TO DATA SCIENCE (3)

Class Number: 6282 *Delivery Method: In Person*

Overview

COURSE TIMES + LOCATION:

Mo 4:30 PM – 6:20 PM
AQ 5018, Burnaby

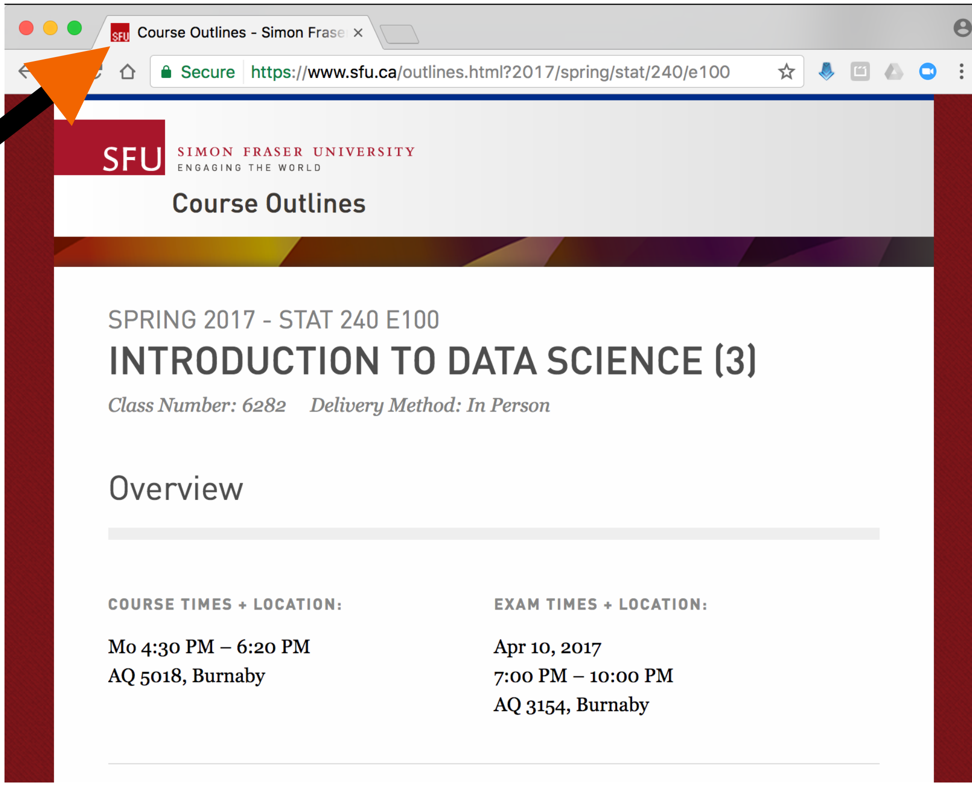
EXAM TIMES + LOCATION:

Apr 10, 2017
7:00 PM – 10:00 PM
AQ 3154, Burnaby

https://www.sfu.ca/outlines.html?2018/spring/stat/240/d100

Chrome: view —> Developer —> view source (on a mac: command+option+u)

```
<!DOCTYPE html>

<html>

<head>

     <meta http-equiv="X-UA-Compatible"
content="IE=Edge, chrome=1">

<meta http-equiv="content-type" content="text/
html; charset=UTF-8" />

<meta name="viewport" content="width=device-
width, initial-scale=1.0, maximum-scale=1.0">

<title>Course Outlines - Simon Fraser
University</title>
```



Course Outlines - Simon Frase ×
Secure
https://www.sfu.ca/outlines.html?2017/spring/stat/240/e100
SFU
SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

Course Outlines

SPRING 2017 - STAT 240 E100

INTRODUCTION TO DATA SCIENCE (3)

Class Number: 6282   *Delivery Method: In Person*

Overview

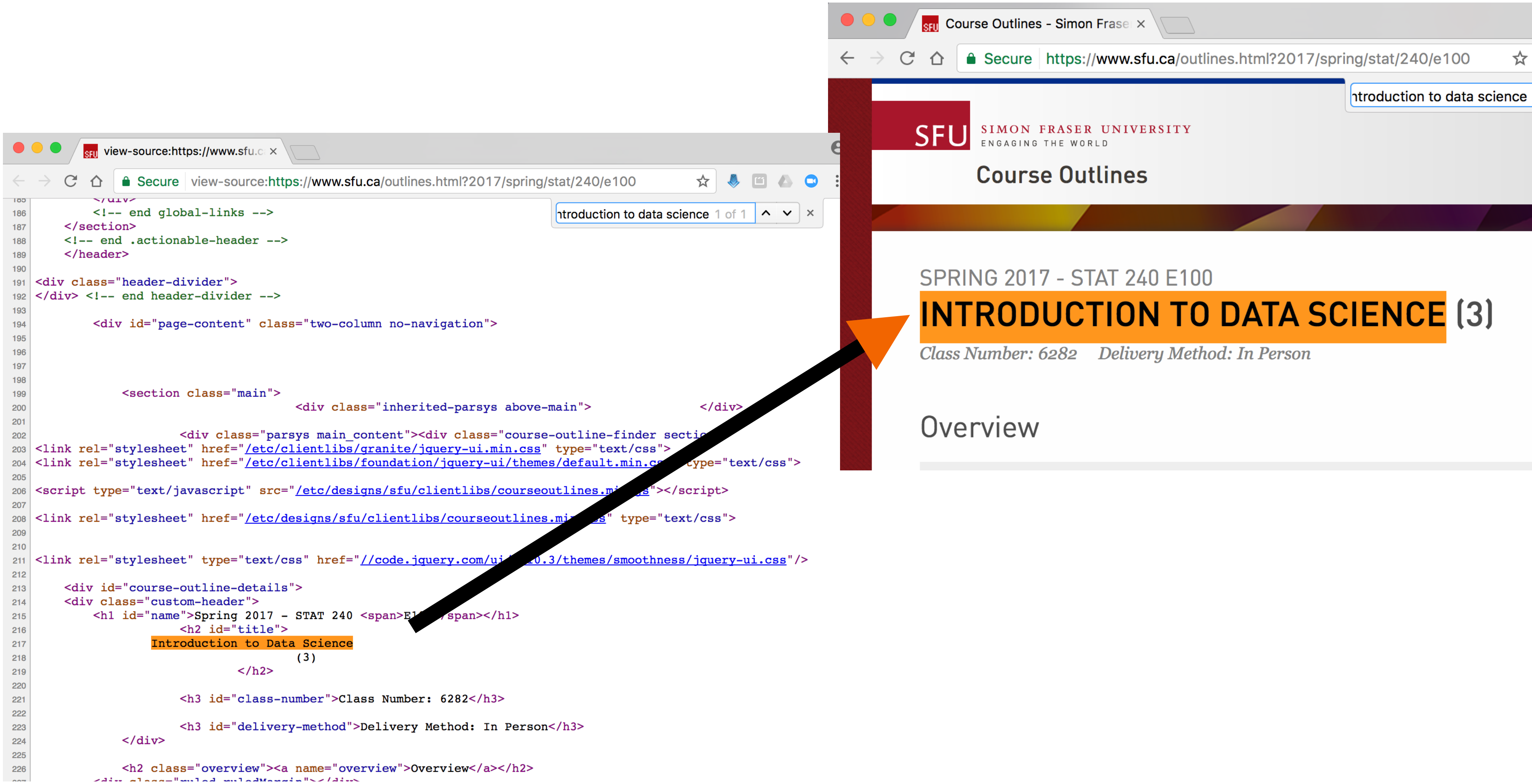
COURSE TIMES + LOCATION:

Mo 4:30 PM – 6:20 PM
AQ 5018, Burnaby

EXAM TIMES + LOCATION:

Apr 10, 2017
7:00 PM – 10:00 PM
AQ 3154, Burnaby

Search the source code and the webpage for reference points


view-source:https://www.sfu.c
Secure
view-source:https://www.sfu.ca/outlines.html?2017/spring/stat/240/e100
ntroduction to data science
1 of 1
Course Outlines - Simon Frase
Secure
https://www.sfu.ca/outlines.html?2017/spring/stat/240/e100
ntroduction to data science
SFU
SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

```
        </div>
        <!-- end global-links -->
    </section>
    <!-- end .actionable-header -->
    </header>

<div class="header-divider">
</div> <!-- end header-divider -->

        <div id="page-content" class="two-column no-navigation">




            <section class="main">
                                    <div class="inherited-parsys above-main">              </div>

                    <div class="parsys main_content"><div class="course-outline-finder sectio
<link rel="stylesheet" href="/etc/clientlibs/granite/jquery-ui.min.css" type="text/css">
<link rel="stylesheet" href="/etc/clientlibs/foundation/jquery-ui/themes/default.min.cs   type="text/css">

<script type="text/javascript" src="/etc/designs/sfu/clientlibs/courseoutlines.mi   s"></script>

<link rel="stylesheet" href="/etc/designs/sfu/clientlibs/courseoutlines.mi   s" type="text/css">


<link rel="stylesheet" type="text/css" href="//code.jquery.com/ui   0.3/themes/smoothness/jquery-ui.css"/>

    <div id="course-outline-details">
    <div class="custom-header">
        <h1 id="name">Spring 2017 - STAT 240 <span>E1   /span></h1>
                    <h2 id="title">
                Introduction to Data Science
                                        (3)
                                </h2>

                    <h3 id="class-number">Class Number: 6282</h3>

                    <h3 id="delivery-method">Delivery Method: In Person</h3>
            </div>

            <h2 class="overview"><a name="overview">Overview</a></h2>
        <div class="ruled ruledMargin"></div>
```

Course Outlines

SPRING 2017 - STAT 240 E100

INTRODUCTION TO DATA SCIENCE (3)

Class Number: 6282 *Delivery Method: In Person*

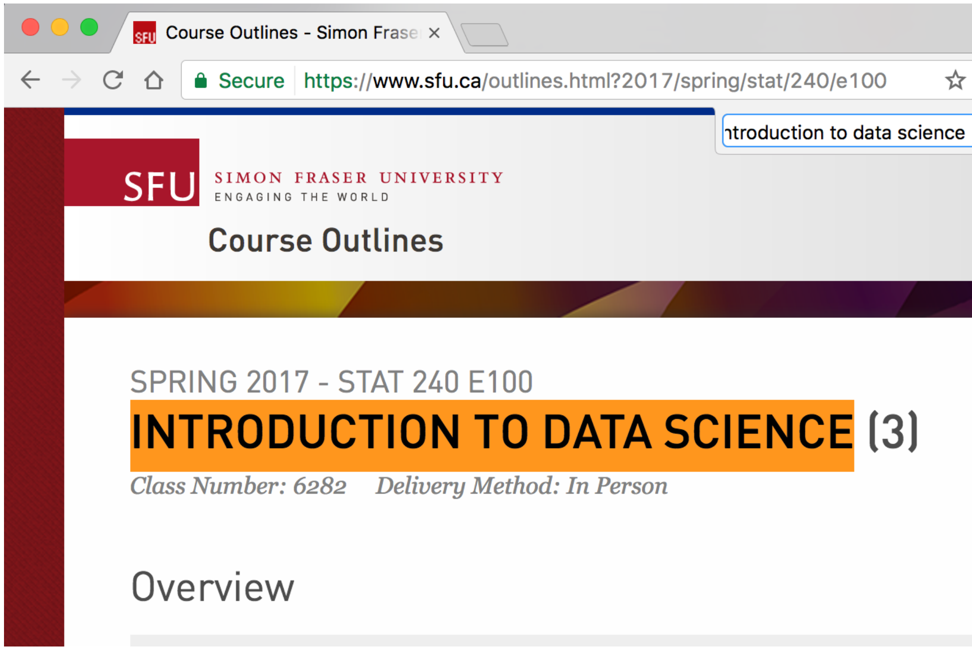
Overview

Search the source code and the webpage for reference points

Look for the html tags

tags define content

```
<h1 id="name">Spring 2017 - STAT
240 <span>E100</span></h1>

<h2 id="title">

Introduction to Data Science

(3)

</h2>
```



Course Outlines - Simon Frase
Secure
https://www.sfu.ca/outlines.html?2017/spring/stat/240/e100
ntroduction to data science
SFU
SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

Course Outlines

SPRING 2017 - STAT 240 E100

INTRODUCTION TO DATA SCIENCE (3)

Class Number: 6282 *Delivery Method: In Person*

Overview

tag start	tag end
<h1>	</h1>

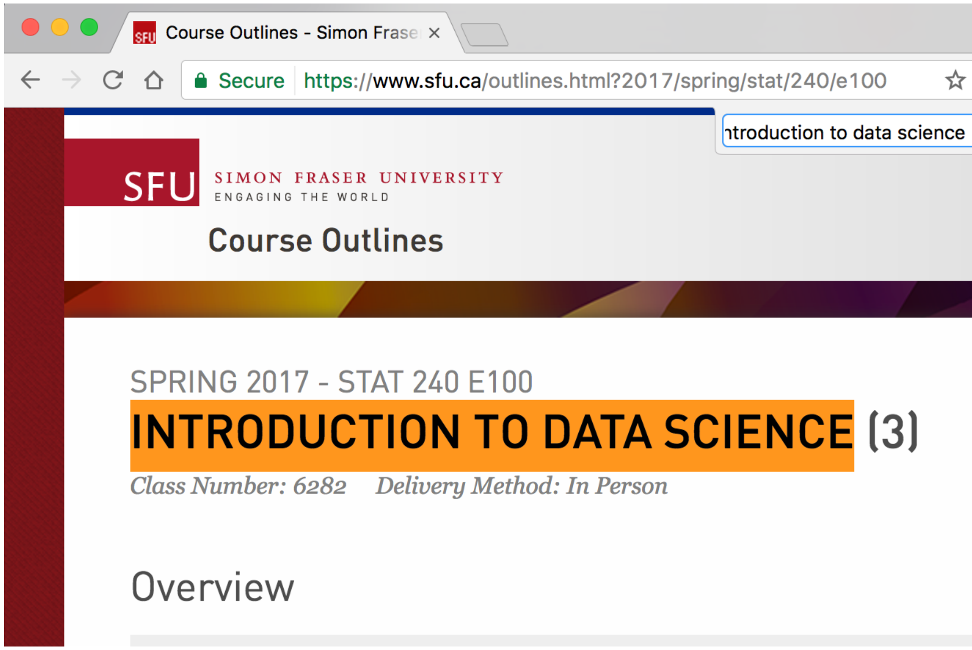
everything between is treated the same

Some tags are generic and define a text style
<h1>, <h2>, <p>, <em>, <strong>,…

Some tags are generic

Some tags define specific content

```
 <h1 id="name">Spring 2017 - STAT
240 <span>E100</span></h1>

                <h2 id="title">

            Introduction to Data Science

                        (3)

                  </h2>
```


Course Outlines - Simon Frase
Secure
https://www.sfu.ca/outlines.html?2017/spring/stat/240/e100
ntroduction to data science
SFU
SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

Course Outlines

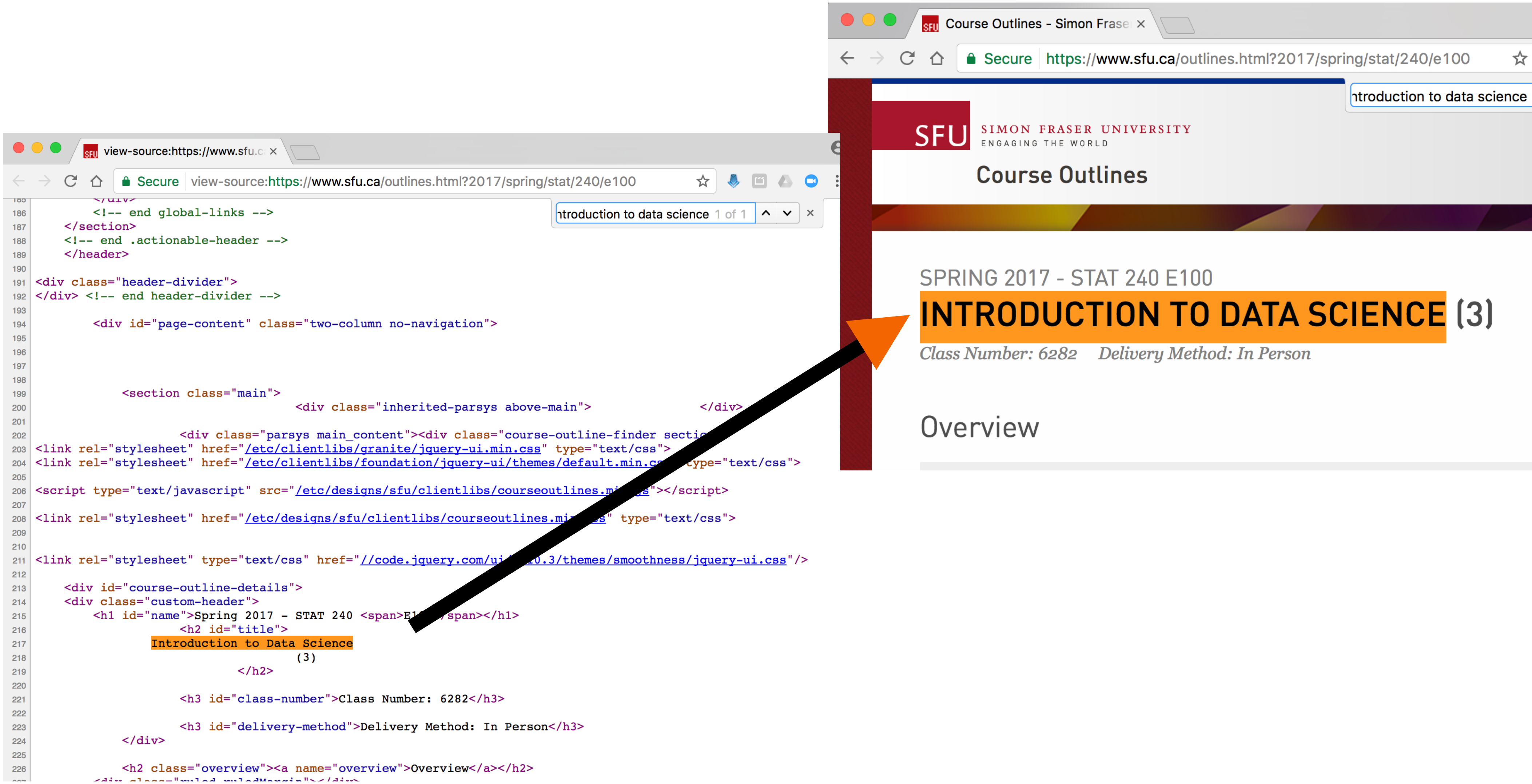
SPRING 2017 - STAT 240 E100

INTRODUCTION TO DATA SCIENCE (3)

Class Number: 6282 *Delivery Method: In Person*

Overview

Ideally, the information we want on multiple pages can be found based on a consistent location or id tag within a page.


Course Outlines - Simon Frase ×
Secure
https://www.sfu.ca/outlines.html?2017/spring/stat/240/e100
ntroduction to data science
view-source:https://www.sfu.c ×
Secure
view-source:https://www.sfu.ca/outlines.html?2017/spring/stat/240/e100
ntroduction to data science 1 of 1

SFU SIMON FRASER UNIVERSITY

ENGAGING THE WORLD

Course Outlines

SPRING 2017 - STAT 240 E100

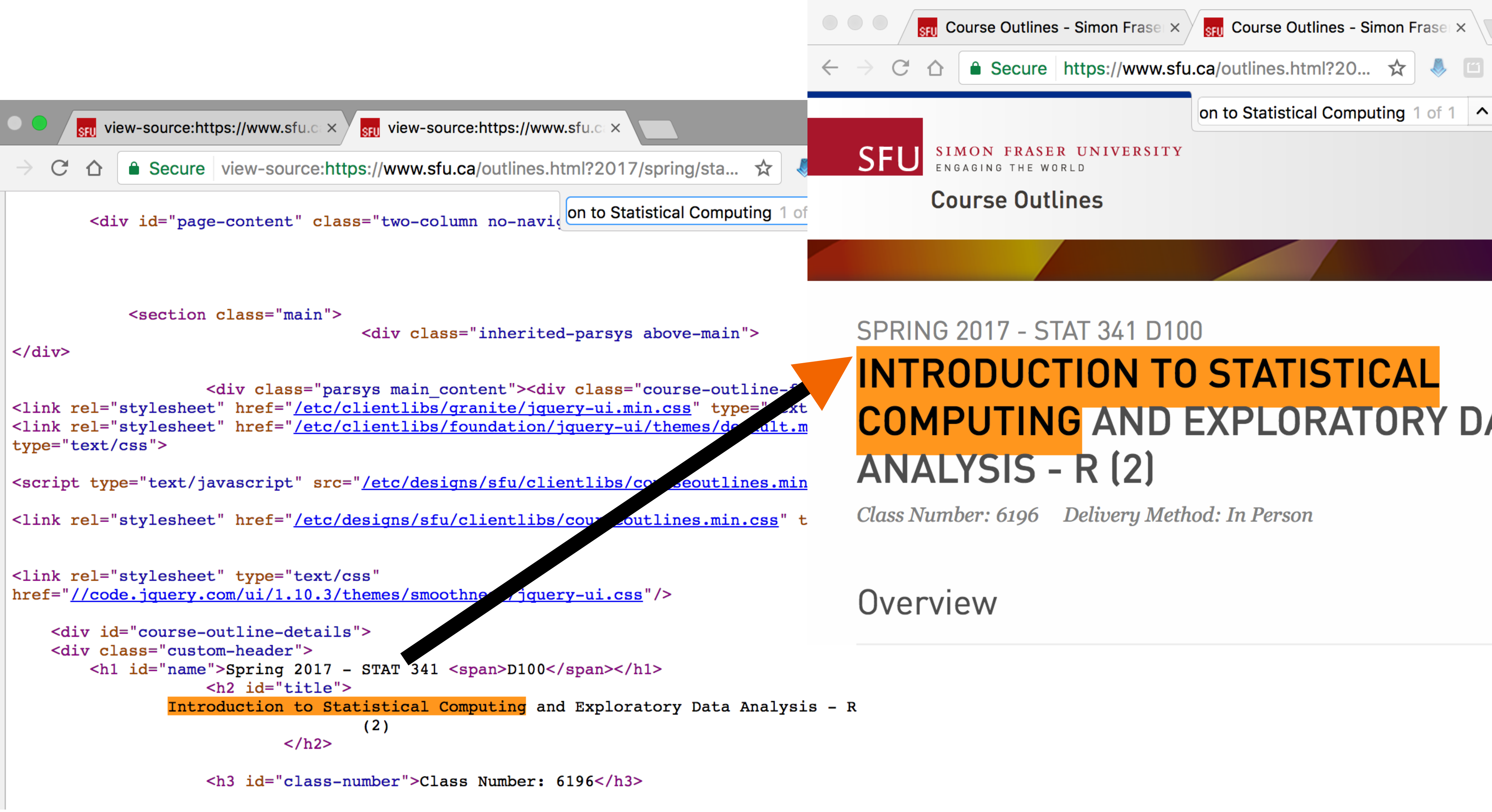
INTRODUCTION TO DATA SCIENCE (3)

Class Number: 6282 *Delivery Method: In Person*

Overview

```
        </div>
        <!-- end global-links -->
    </section>
    <!-- end .actionable-header -->
    </header>

<div class="header-divider">
</div> <!-- end header-divider -->

        <div id="page-content" class="two-column no-navigation">




            <section class="main">
                                    <div class="inherited-parsys above-main">              </div>

                    <div class="parsys main_content"><div class="course-outline-finder sectio
<link rel="stylesheet" href="/etc/clientlibs/granite/jquery-ui.min.css" type="text/css">
<link rel="stylesheet" href="/etc/clientlibs/foundation/jquery-ui/themes/default.min.cs   type="text/css">

<script type="text/javascript" src="/etc/designs/sfu/clientlibs/courseoutlines.mi   s"></script>

<link rel="stylesheet" href="/etc/designs/sfu/clientlibs/courseoutlines.mi   s" type="text/css">


<link rel="stylesheet" type="text/css" href="//code.jquery.com/ui   0.3/themes/smoothness/jquery-ui.css"/>

    <div id="course-outline-details">
    <div class="custom-header">
        <h1 id="name">Spring 2017 - STAT 240 <span>E1   /span></h1>
                <h2 id="title">
            Introduction to Data Science
                                    (3)
                        </h2>

                <h3 id="class-number">Class Number: 6282</h3>

                <h3 id="delivery-method">Delivery Method: In Person</h3>
            </div>

            <h2 class="overview"><a name="overview">Overview</a></h2>
```

Ideally, the information we want on multiple pages can be found based on a consistent location or id tag within a page.



view-source:https://www.sfu.c ×
view-source:https://www.sfu.c ×
Secure
view-source:https://www.sfu.ca/outlines.html?2017/spring/sta...
on to Statistical Computing 1 of

```
        <div id="page-content" class="two-column no-navig



            <section class="main">
                                      <div class="inherited-parsys above-main">
</div>

                     <div class="parsys main_content"><div class="course-outline-f
<link rel="stylesheet" href="/etc/clientlibs/granite/jquery-ui.min.css" type="text
<link rel="stylesheet" href="/etc/clientlibs/foundation/jquery-ui/themes/default.m
type="text/css">

<script type="text/javascript" src="/etc/designs/sfu/clientlibs/courseoutlines.min

<link rel="stylesheet" href="/etc/designs/sfu/clientlibs/courseoutlines.min.css" t


<link rel="stylesheet" type="text/css"
href="//code.jquery.com/ui/1.10.3/themes/smoothness/jquery-ui.css"/>

    <div id="course-outline-details">
    <div class="custom-header">
        <h1 id="name">Spring 2017 - STAT 341 <span>D100</span></h1>
                    <h2 id="title">
                Introduction to Statistical Computing and Exploratory Data Analysis - R
                                    (2)
                            </h2>

                    <h3 id="class-number">Class Number: 6196</h3>
```

Course Outlines - Simon Frase ×
Course Outlines - Simon Frase ×
Secure
https://www.sfu.ca/outlines.html?20...
on to Statistical Computing 1 of 1
SFU
SIMON FRASER UNIVERSITY
ENGAGING THE WORLD
Course Outlines

SPRING 2017 - STAT 341 D100

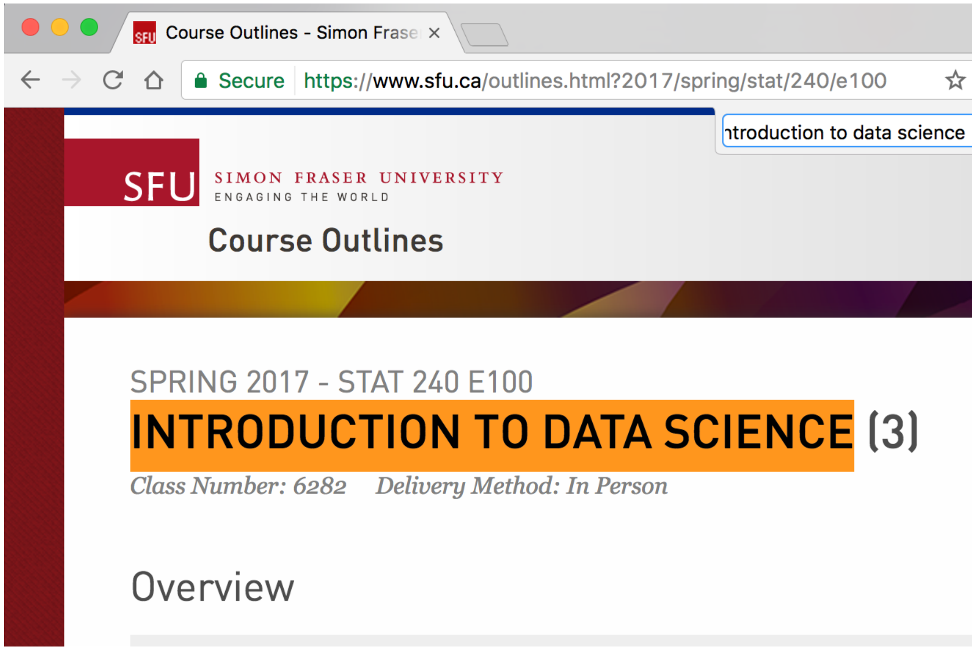
INTRODUCTION TO STATISTICAL COMPUTING AND EXPLORATORY DA ANALYSIS - R (2)

Class Number: 6196 *Delivery Method: In Person*

Overview

Ideally, the information we want on multiple pages can be found based on a consistent location or id tag within a page.

```
 <h1 id="name">Spring 2017 - STAT
240 <span>E100</span></h1>

                    <h2 id="title">

                    Introduction to Data
Science

                                  (3)

                              </h2>
```


Course Outlines - Simon Frase ×
Secure
https://www.sfu.ca/outlines.html?2017/spring/stat/240/e100
ntroduction to data science

SFU SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

Course Outlines

SPRING 2017 - STAT 240 E100

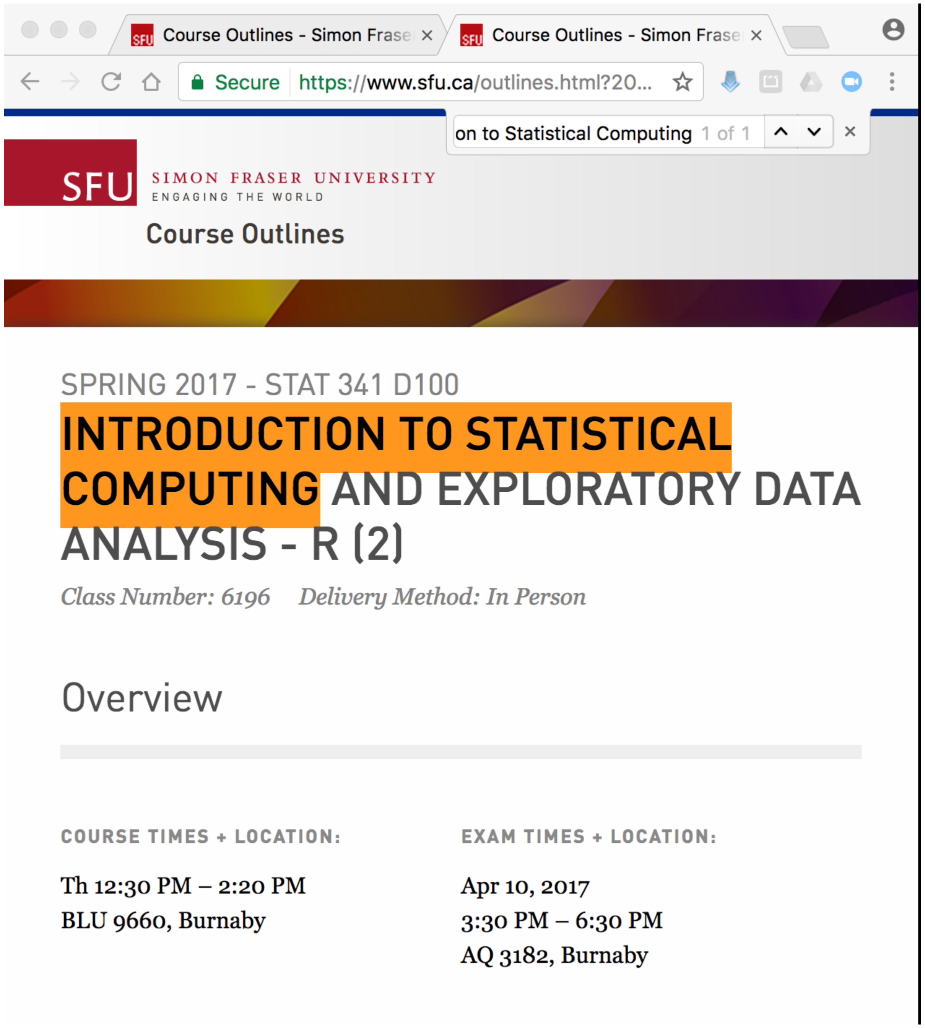
INTRODUCTION TO DATA SCIENCE (3)

Class Number: 6282   *Delivery Method: In Person*

Overview

Ideally, the information we want on multiple pages can be found based on a consistent location or id tag within a page.

```
 <h1 id="name">Spring 2017 - STAT
341 <span>D100</span></h1>

                <h2 id="title">

                    Introduction to Statistical
Computing and Exploratory Data
Analysis - R

                                    (2)

                    </h2>
```


Course Outlines - Simon Frase ×
Course Outlines - Simon Frase ×
Secure
https://www.sfu.ca/outlines.html?20...
on to Statistical Computing 1 of 1
SFU
SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

Course Outlines

SPRING 2017 - STAT 341 D100

INTRODUCTION TO STATISTICAL COMPUTING AND EXPLORATORY DATA ANALYSIS - R (2)

Class Number: 6196 *Delivery Method: In Person*

Overview

COURSE TIMES + LOCATION:

Th 12:30 PM – 2:20 PM
BLU 9660, Burnaby

EXAM TIMES + LOCATION:

Apr 10, 2017
3:30 PM – 6:30 PM
AQ 3182, Burnaby

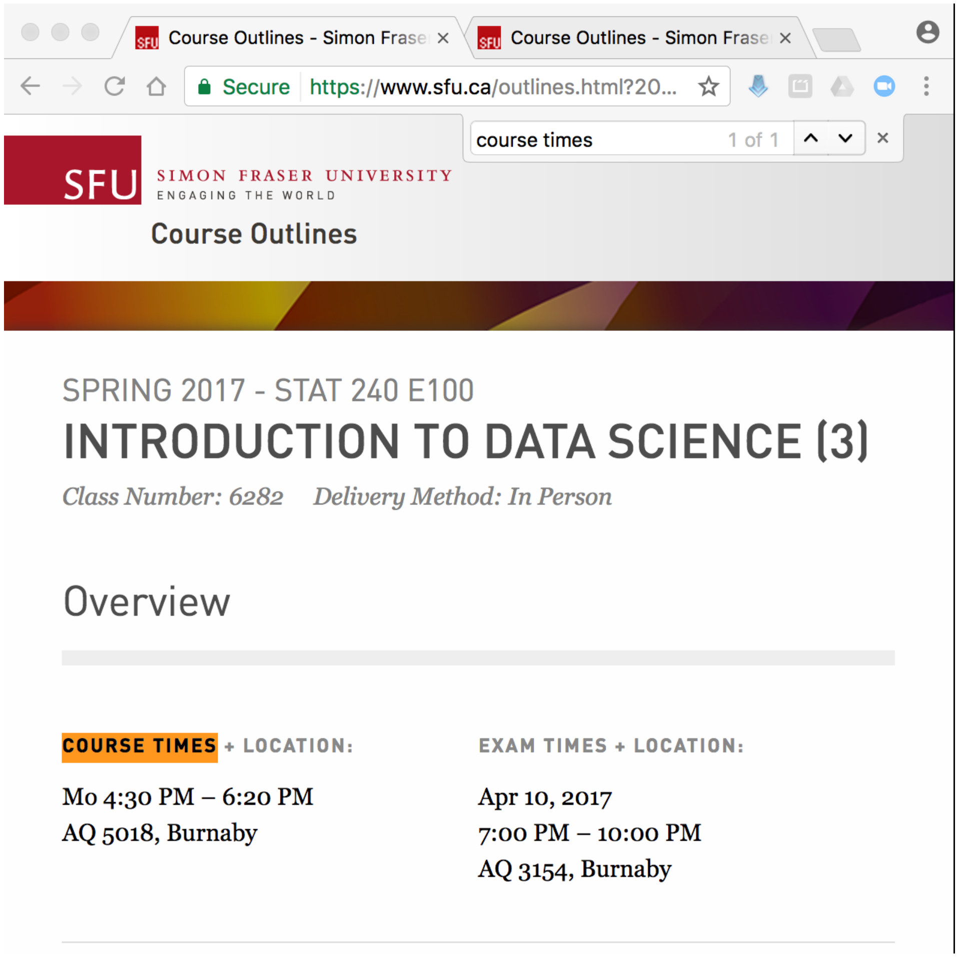
Get html source code

```
course_url = "https://www.sfu.ca/outlines.html?2018/spring/stat/240/d100"
```

```
(course_page = readLines(course_url))
```

Then use regular expressions!

Tips!
tags define boundaries


Course Outlines - Simon Frase
Course Outlines - Simon Frase
Secure
https://www.sfu.ca/outlines.html?20...
course times
1 of 1
SFU
SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

Course Outlines

SPRING 2017 - STAT 240 E100

INTRODUCTION TO DATA SCIENCE (3)

Class Number: 6282 *Delivery Method: In Person*

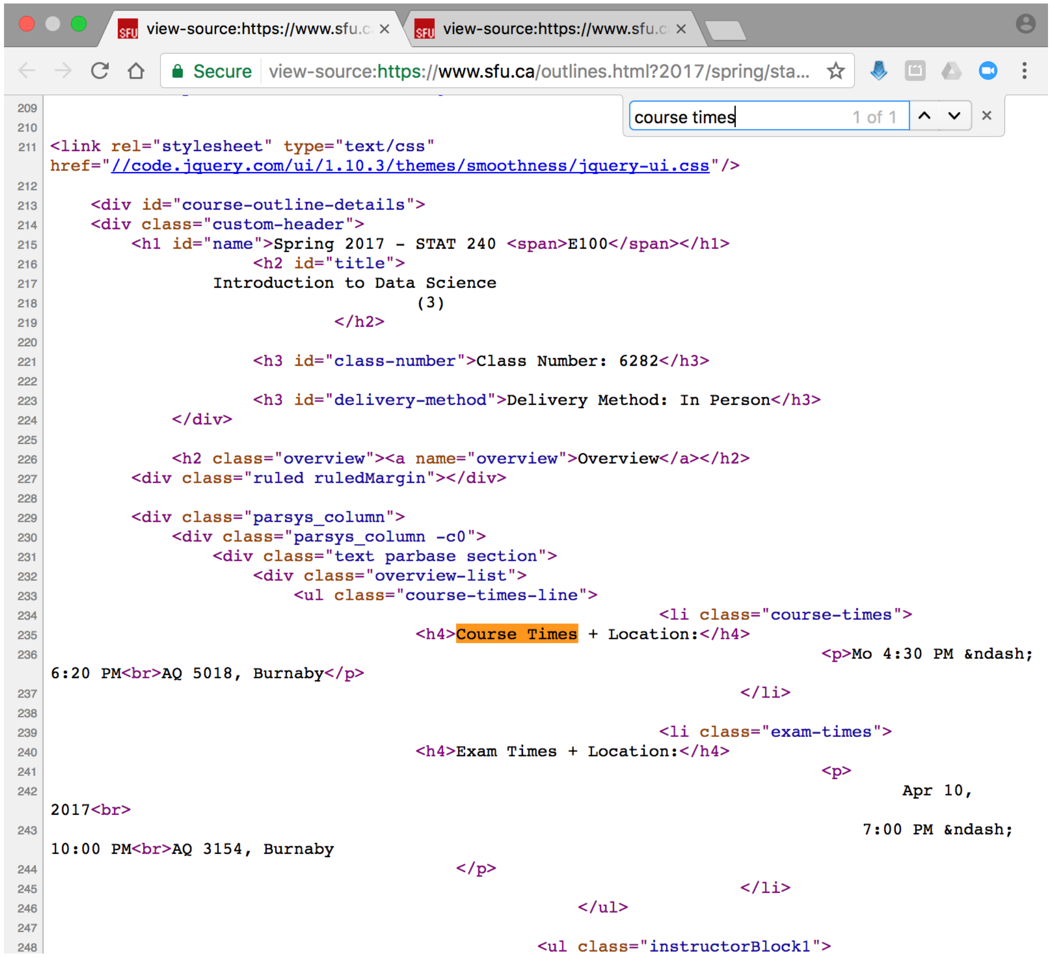
Overview

COURSE TIMES + LOCATION:

Mo 4:30 PM – 6:20 PM
AQ 5018, Burnaby

EXAM TIMES + LOCATION:

Apr 10, 2017
7:00 PM – 10:00 PM
AQ 3154, Burnaby


view-source:https://www.sfu.c
view-source:https://www.sfu.c
Secure
view-source:https://www.sfu.ca/outlines.html?2017/spring/sta...
course times
1 of 1

```


<link rel="stylesheet" type="text/css"
href="//code.jquery.com/ui/1.10.3/themes/smoothness/jquery-ui.css"/>

  <div id="course-outline-details">
  <div class="custom-header">
    <h1 id="name">Spring 2017 - STAT 240 <span>E100</span></h1>
          <h2 id="title">
        Introduction to Data Science
                  (3)
              </h2>

          <h3 id="class-number">Class Number: 6282</h3>

          <h3 id="delivery-method">Delivery Method: In Person</h3>
      </div>

      <h2 class="overview"><a name="overview">Overview</a></h2>
    <div class="ruled ruledMargin"></div>

    <div class="parsys_column">
      <div class="parsys_column -c0">
        <div class="text parbase section">
          <div class="overview-list">
            <ul class="course-times-line">
                              <li class="course-times">
                  <h4>Course Times + Location:</h4>
                                      <p>Mo 4:30 PM &ndash;
6:20 PM<br>AQ 5018, Burnaby</p>
                                  </li>

                              <li class="exam-times">
                  <h4>Exam Times + Location:</h4>
                                      <p>
                                          Apr 10,
2017<br>
                                        7:00 PM &ndash;
10:00 PM<br>AQ 3154, Burnaby
                    </p>
                                  </li>
                        </ul>

                          <ul class="instructorBlock1">
```

Tips!
define start and ends to your info

```
<h4>Course Times +
Location:</h4>


<p>Mo 4:30 PM &ndash; 6:20
PM<br>AQ 5018, Burnaby</p>


</li>

   <li class="exam-times">
```

```
(startindex =
grep("<h4>Course
Times \\+
Location:",course_page)
)

(endindex = grep("<li
class="exam-
times">",course_page))

course_page[(startindex
+1):(endindex-1)]
```

Note the minor code fail and the fix

```
paste(course_page[(startindex+1):(endindex-1)],collapse="")   #collapse all elements into a single string
```

Then remove html formatting (see twitter client assignment question).

Get a lot of course info
run a for loop over course numbers and sections

```
baseurl = "https://www.sfu.ca/outlines.html?"

year = "2017"

term = "spring"

dept="stat"

courseNo = "240"

section = "e100"

course_url1 = paste(baseurl,year,sep="")

(course_url = paste(course_url1,term,dept, courseNo,
section,sep="/"))
```

HTML

```
<p>Directly parsing Canada's <a href="https://
en.wikipedia.org/wiki/
List_of_National_Parks_of_Canada"
title="National Parks of Canada">National
Parks</a> table should be much easier than
what I did in class.  But that’s step 2</p>
```

WebScraping version 2

```
library(rvest)

file = read_html("https://en.wikipedia.org/wiki/
List_of_National_Parks_of_Canada")

out = html_table(html_nodes(file, "table")[[1]])



length(out)

out[[1]]

out[[2]]
```

fixing size:

```
head(out)

out[,5]

step1 =  gsub(x=out[,5],pattern =     ,replacement =   )

step2 =    gsub(x=step1,     pattern =      ,replacement=   )

km2 =
as.numeric(gsub(x=step2,pattern=       ,replacement=   )
```

Touch ups

fixing location :

```
out[,3]

prov =                  gsub(out[,3],pattern=“   “,replacement="  “)
```

fixing year:

```
out[,2]

year = as.numeric(gsub(out[,4],pattern=“  “,replacement="  "))
```

Filling in the table

```
NatParks = data.frame(name=out[,1],

              year=year,

              size=km2,

              location = prov)
```

Dashboarding via ShinyApps

Using html

https://twitter.com/ShinyappsRecent

visit a (html) webpage (like: https://istats.shinyapps.io/MultivariateRelationship/ or https://jheppler.shinyapps.io/omaha-bikes/)

use menus to select the data and analysis

R runs on the server and renders analytics to your web browser

Get going (R side)

https://shiny.rstudio.com/articles/shinyapps.html

Get Going (SFU side)

Full instructions:

http://www.rcg.sfu.ca/services/shiny/index.html

Step 1: sign up for the mail list to give you access to our servers

Step 2: Upload your Shiny App

Step 3: tell your friends / show employers / tweet #ShinyApp using your own url:

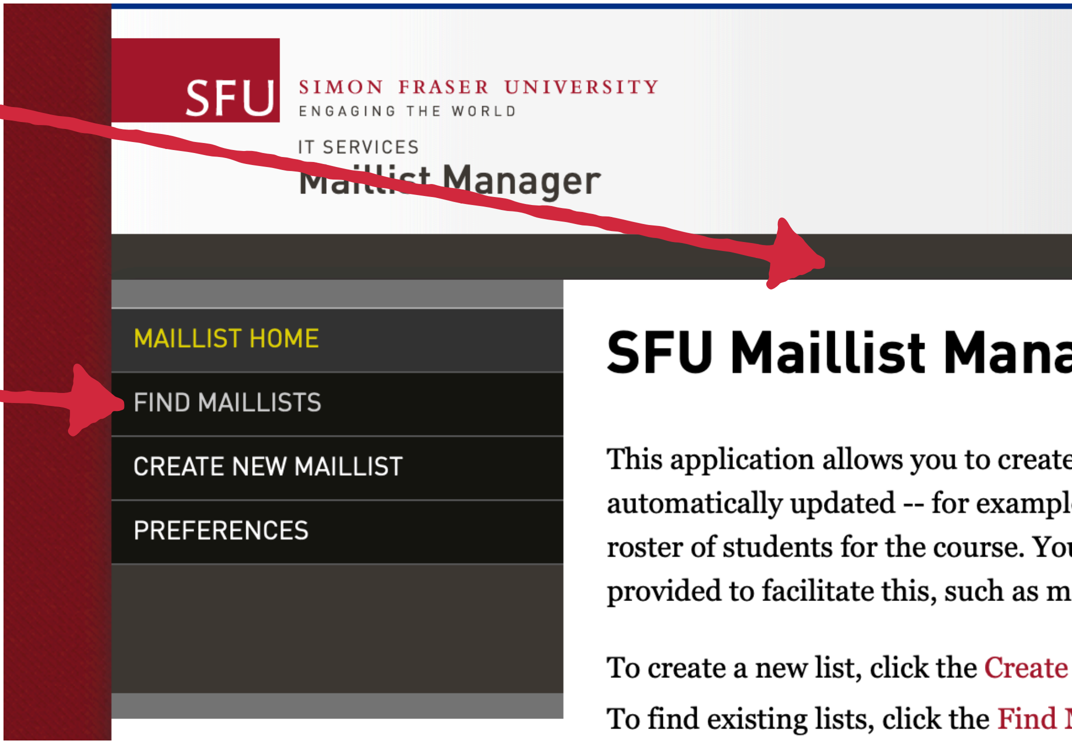
https://shiny.rcg.sfu.ca/u/your_SFU_username_goes_here/myapp

Sign up for that maillist…

(via: https://www.sfu.ca/itservices/sfu_email/user-guide.html)

Log into https://amaint.sfu.ca/cgi-bin/WebObjects/Maillist.woa

go to Manage your Maillists



SFU Maillist Mana

This application allows you to create
automatically updated -- for exampl
roster of students for the course. Yo
provided to facilitate this, such as m

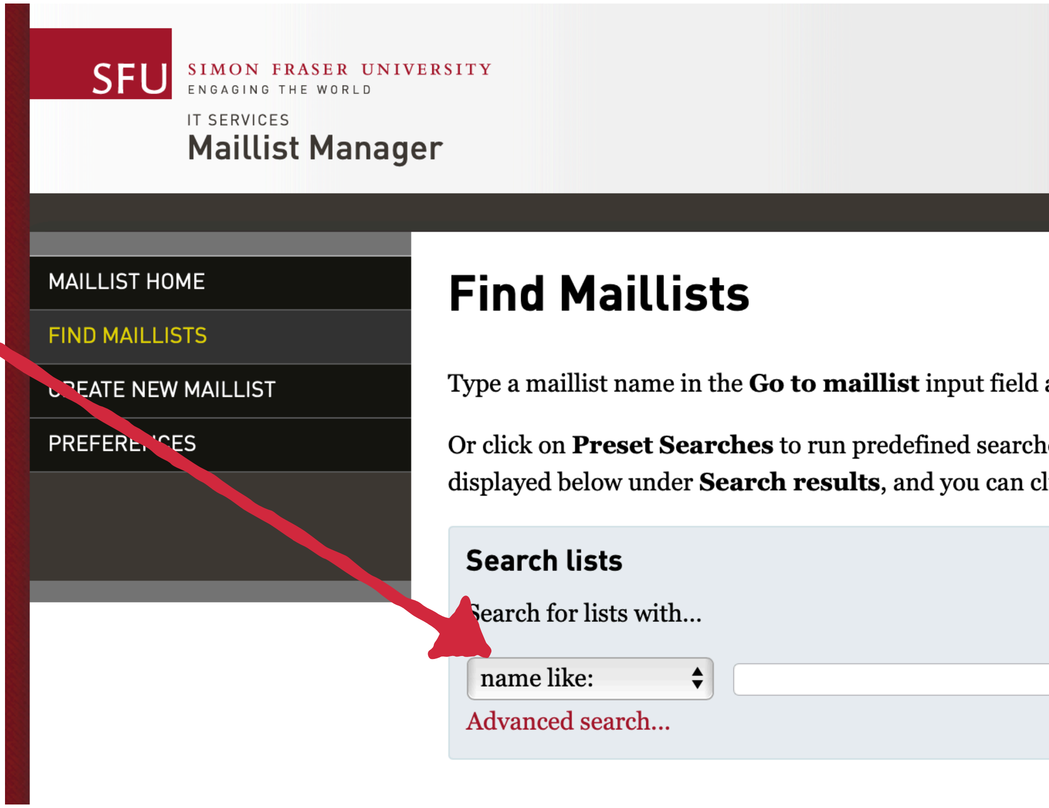
To create a new list, click the Create

To find existing lists, click the Find I

Search for the rcg-shiny-users maillist here

Then tell it you want to subscribe

Procrastinators Beware: (Then wait ~½ hour for the next step)



Find Maillists

Type a maillist name in the **Go to maillist** input field a

Or click on **Preset Searches** to run predefined search
displayed below under **Search results**, and you can cl

Search lists

Search for lists with...

name like:

Advanced search...