

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:

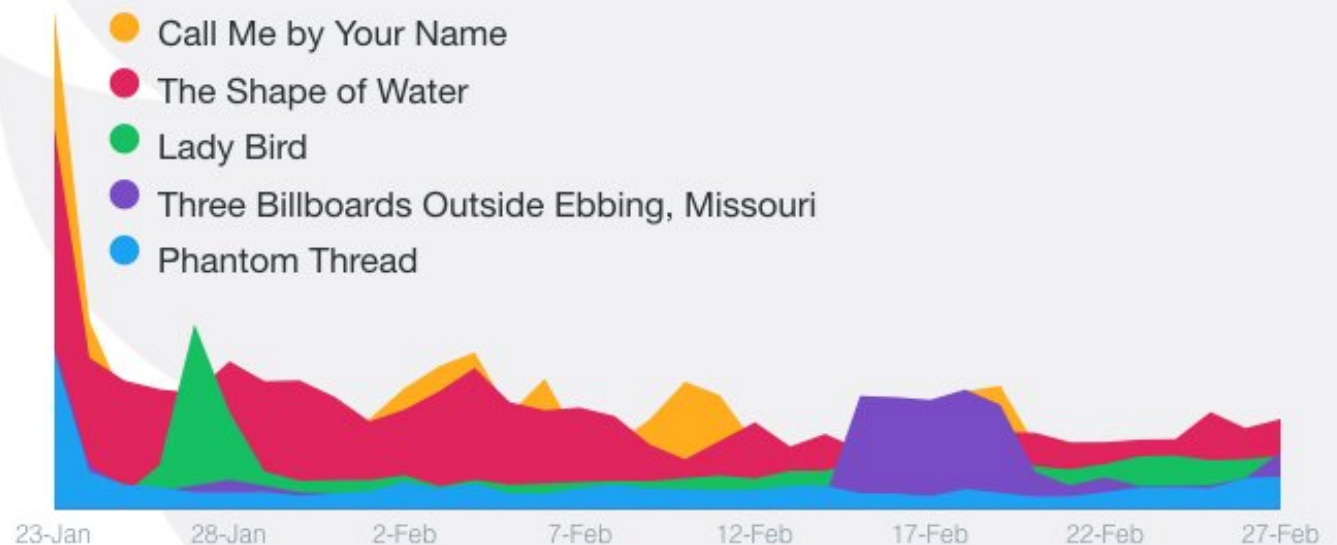
“Shape of water”

Oscars were held March 4th

Image posted March 1

2018 Best Picture Nominees

Tweets per day



 #Oscars

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

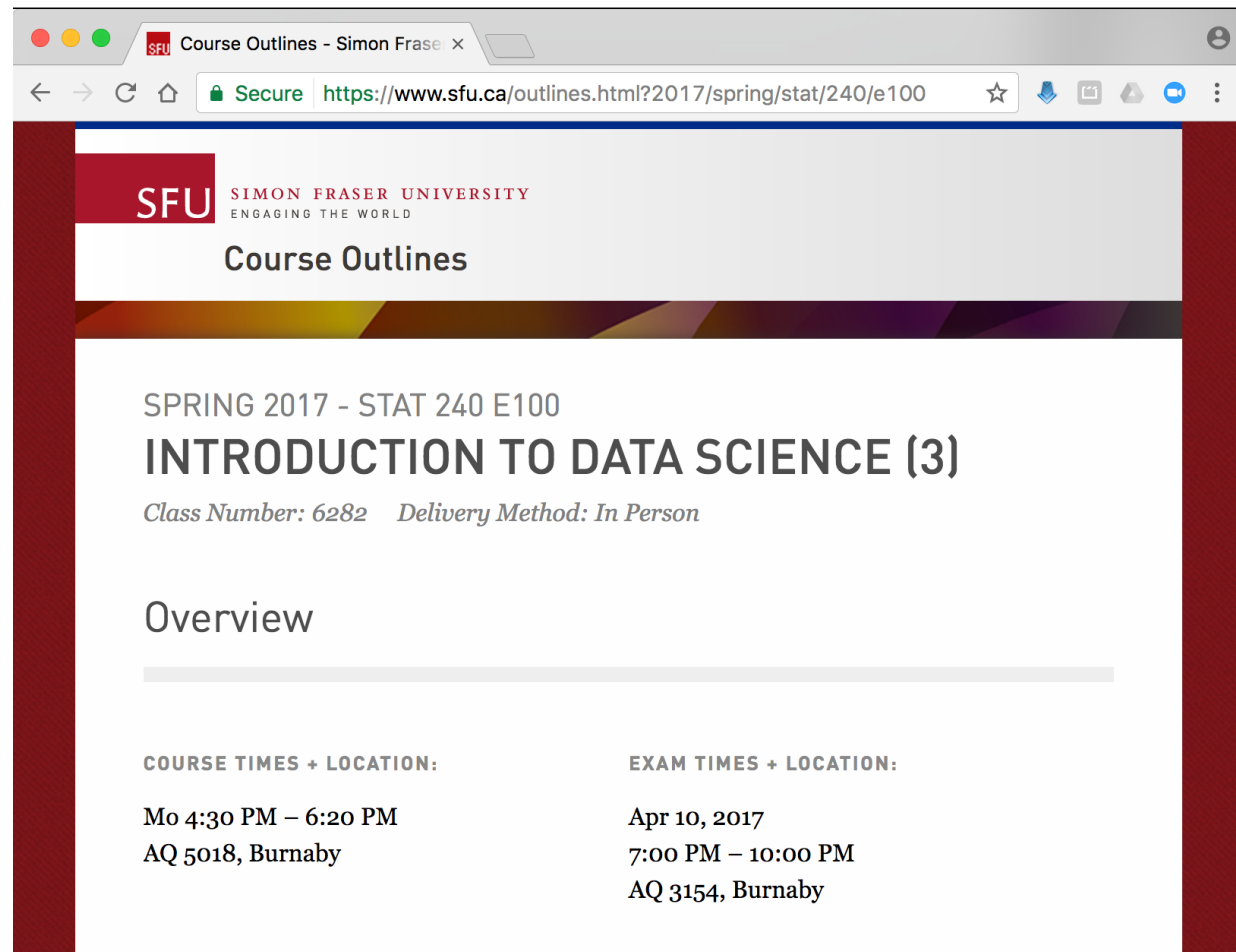
Disallow: <— No pages are disallowed

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 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 Fraser x

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:	EXAM TIMES + LOCATION:
Mo 4:30 PM – 6:20 PM AQ 5018, Burnaby	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>
```

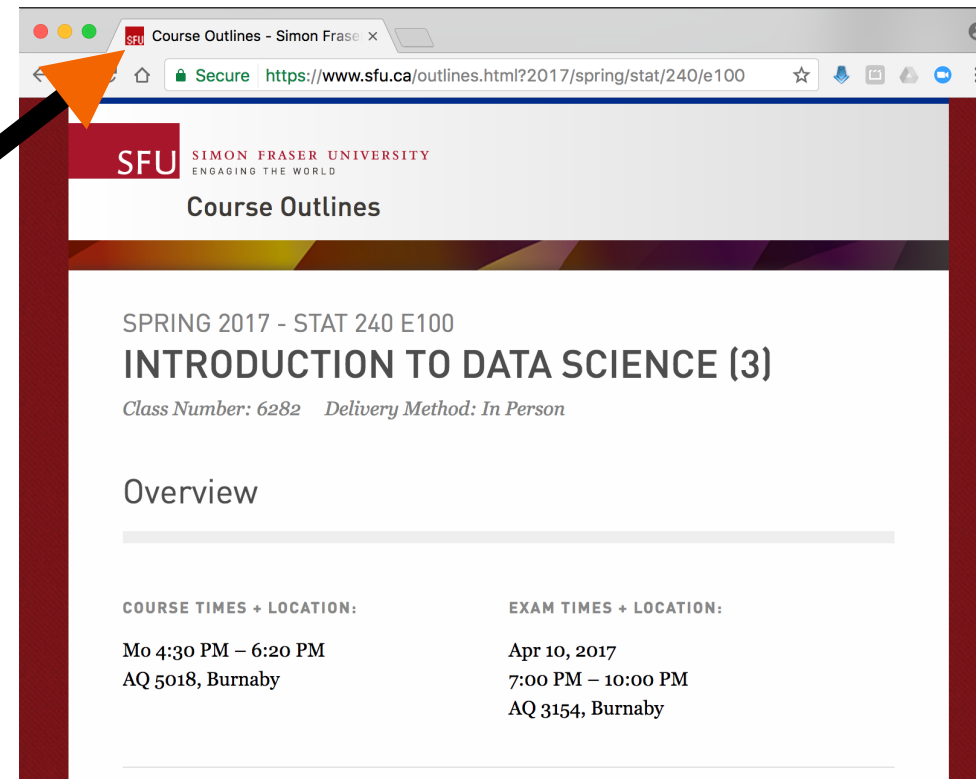
```
<head>
```

```
  <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>
```



Search the source code and the webpage for reference points

The image shows a side-by-side comparison of a webpage's source code and its rendered output. On the left, the source code is displayed in a browser's 'view-source' window. On the right, the rendered webpage is shown. An orange arrow points from the source code to the rendered page, indicating the mapping between the two.

Source Code (Left):

```
186 <!-- end global-links -->
187 </section>
188 <!-- end .actionable-header -->
189 </header>
190
191 <div class="header-divder">
192 </div> <!-- end header-divder -->
193
194 <div id="page-content" class="two-column no-navigation">
195
196
197
198
199 <section class="main">
200
201 <div class="inherited-parsys above-main">
202 </div>
203
204 <div class="parsys main_content"><div class="course-outline-finder section">
205 <link rel="stylesheet" href="/etc/clientlibs/granite/jquery-ui.min.css" type="text/css">
206 <link rel="stylesheet" href="/etc/clientlibs/foundation/jquery-ui/themes/default.min.css" type="text/css">
207 <script type="text/javascript" src="/etc/designs/sfu/clientlibs/courseoutlines.min.js"></script>
208 <link rel="stylesheet" href="/etc/designs/sfu/clientlibs/courseoutlines.min.css" type="text/css">
209
210 <link rel="stylesheet" type="text/css" href="//code.jquery.com/ui/1.10.3/themes/smoothness/jquery-ui.css"/>
211
212 <div id="course-outline-details">
213 <div class="custom-header">
214 <h1 id="name">Spring 2017 - STAT 240 <span>E100</span></h1>
215 <h2 id="title">
216 Introduction to Data Science
217 (3)
218 </h2>
219
220 <h3 id="class-number">Class Number: 6282</h3>
221
222 <h3 id="delivery-method">Delivery Method: In Person</h3>
223 </div>
224
225 <h2 class="overview"><a name="overview">Overview</a></h2>
226 </div></div></div></div>
```

Rendered Page (Right):

Course Outlines - Simon Fraser University

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

Introduction to data science

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

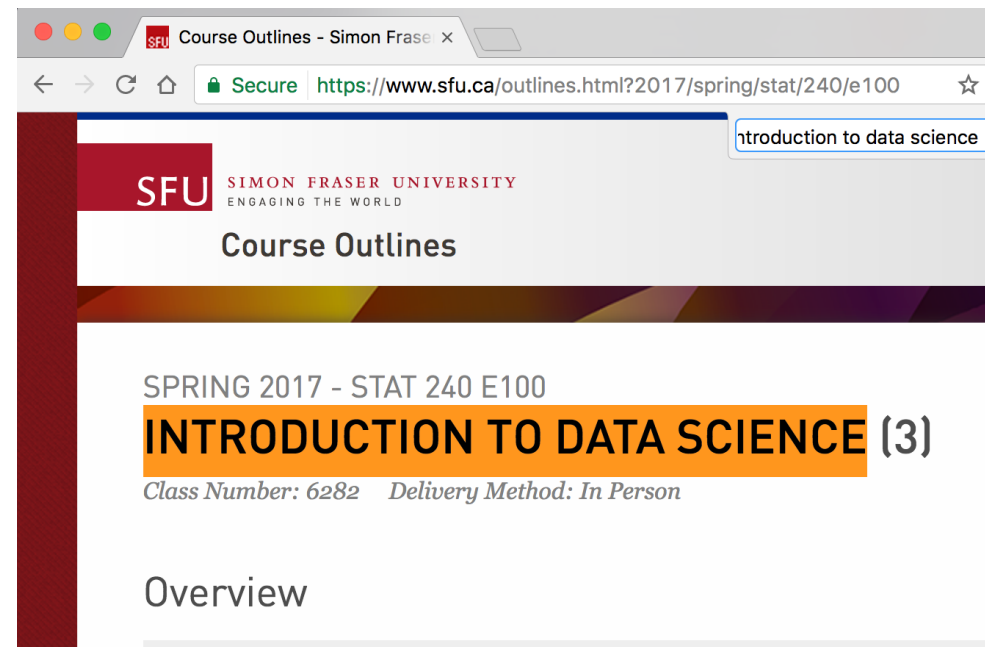
```
<h1 id="name">Spring 2017 - STAT  
240 <span>E100</span></h1>
```

```
<h2 id="title">
```

Introduction to Data Science

(3)

```
</h2>
```



tag start

tag end

```
<h1>
```

```
</h1>
```

everything between is treated the same

Some tags are generic and define a text style
<h1>, <h2>, <p>, , ,...

Some tags are generic

Some tags define
specific content

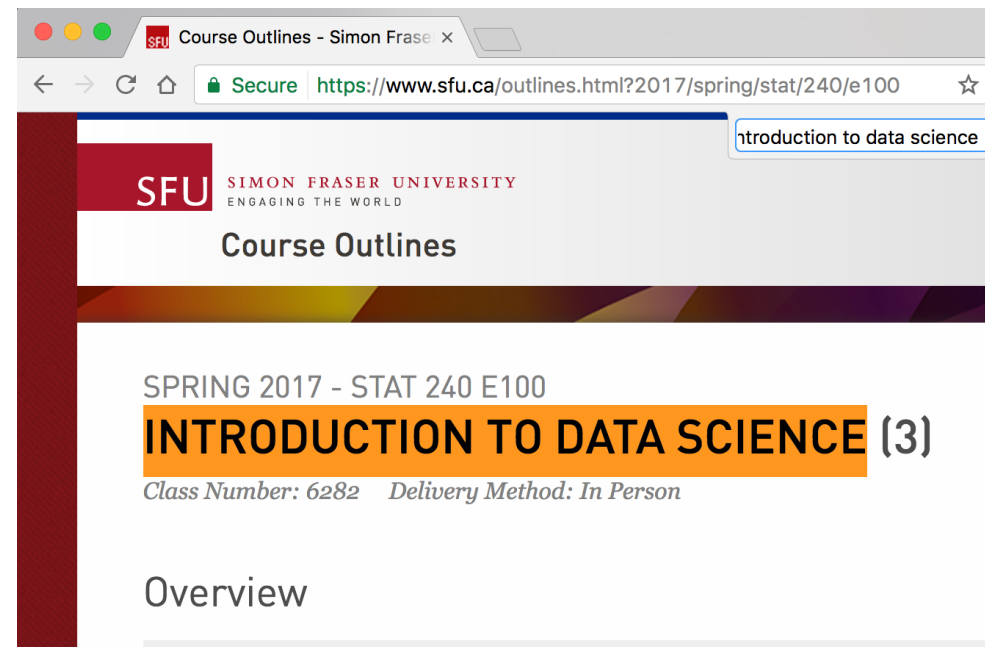
```
<h1 id="name">Spring 2017 - STAT  
240 <span>E100</span></h1>
```

```
<h2 id="title">
```

Introduction to Data Science

(3)

```
</h2>
```



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

The image shows a web browser window with the address bar displaying `https://www.sfu.ca/outlines.html?2017/spring/stat/240/e100`. The browser window is split into two panes. The left pane shows the source code of the page, and the right pane shows the rendered page.

Source Code (Left Pane):

```
186 <!-- end global-links -->
187 </section>
188 <!-- end .actionable-header -->
189 </header>
190
191 <div class="header-divider">
192 </div> <!-- end header-divider -->
193
194 <div id="page-content" class="two-column no-navigation">
195
196
197
198
199 <section class="main">
200
201 <div class="inherited-parsys above-main">
202
203 <div class="parsys main_content"><div class="course-outline-finder section">
204 <link rel="stylesheet" href="/etc/clientlibs/granite/jquery-ui.min.css" type="text/css">
205 <link rel="stylesheet" href="/etc/clientlibs/foundation/jquery-ui/themes/default.min.css" type="text/css">
206 <script type="text/javascript" src="/etc/designs/sfu/clientlibs/courseoutlines.min.js"></script>
207 <link rel="stylesheet" href="/etc/designs/sfu/clientlibs/courseoutlines.min.css" type="text/css">
208 <link rel="stylesheet" href="/etc/designs/sfu/clientlibs/courseoutlines.min.css" type="text/css">
209
210 <link rel="stylesheet" type="text/css" href="//code.jquery.com/ui/1.10.3/themes/smoothness/jquery-ui.css"/>
211
212 <div id="course-outline-details">
213 <div class="custom-header">
214 <h1 id="name">Spring 2017 - STAT 240 <span>E100</span></h1>
215 <h2 id="title">
216 Introduction to Data Science
217 (3)
218 </h2>
219
220 <h3 id="class-number">Class Number: 6282</h3>
221
222 <h3 id="delivery-method">Delivery Method: In Person</h3>
223 </div>
224
225 <h2 class="overview"><a name="overview">Overview</a></h2>
226 </div>
227 </div>
228 </div>
```

Rendered Page (Right Pane):

SFU SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

Course Outlines

Introduction to data science

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.

```
view-source:https://www.sfu.ca/outlines.html?2017/spring/sta...
<div id="page-content" class="two-column no-navi...
on to Statistical Computing 1 of 1

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

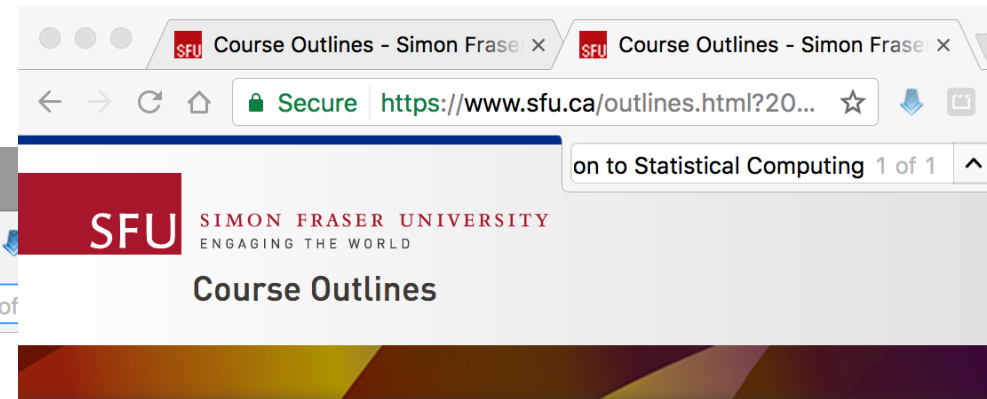
    <div class="parsys main_content"><div class="course-outline-s...
    <link rel="stylesheet" href="/etc/clientlibs/granite/jquery-ui.min.css" type="text
    <link rel="stylesheet" href="/etc/clientlibs/foundation/jquery-ui/themes/default.m
    type="text/css">

    <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>
```



SPRING 2017 - STAT 341 D100

INTRODUCTION TO STATISTICAL COMPUTING AND EXPLORATORY DATA 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.

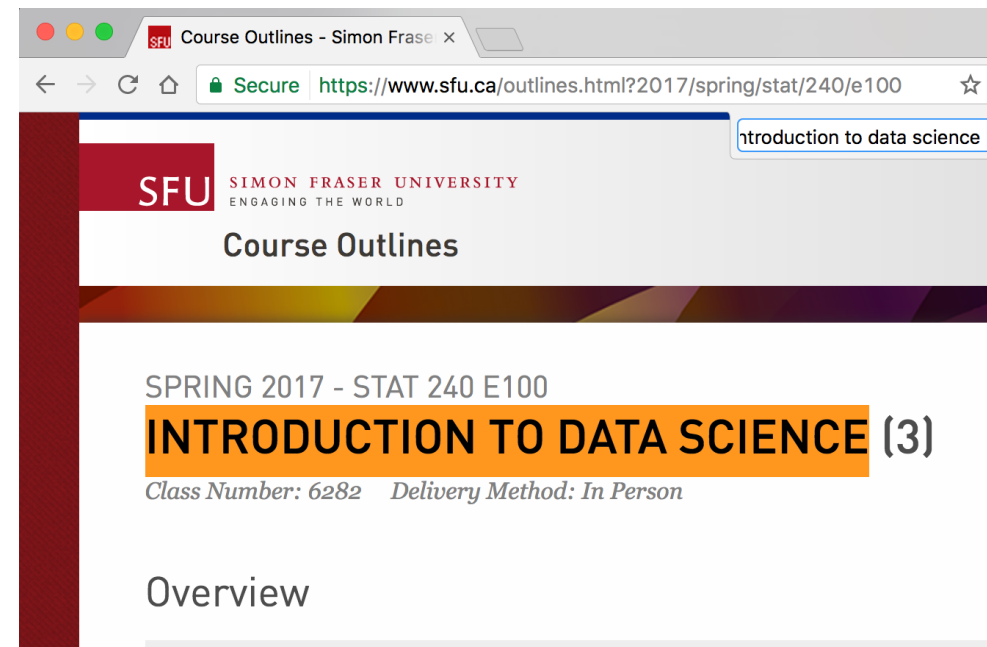
```
<h1 id="name">Spring 2017 - STAT  
240 <span>E100</span></h1>
```

```
<h2 id="title">
```

Introduction to Data
Science

(3)

```
</h2>
```



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

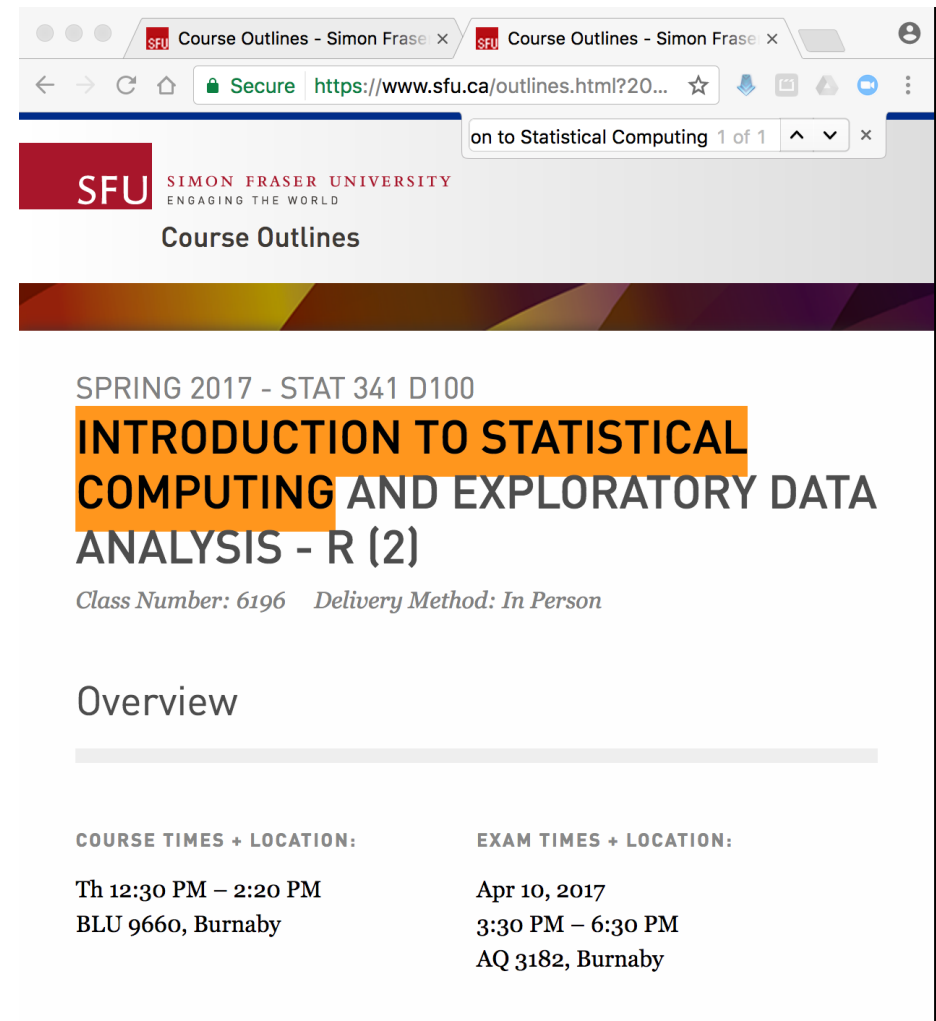
```
<h1 id="name">Spring 2017 - STAT  
341 <span>D100</span></h1>
```

```
<h2 id="title">
```

Introduction to Statistical
Computing and Exploratory Data
Analysis - R

```
(2)
```

```
</h2>
```



Course Outlines - Simon Fraser x

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

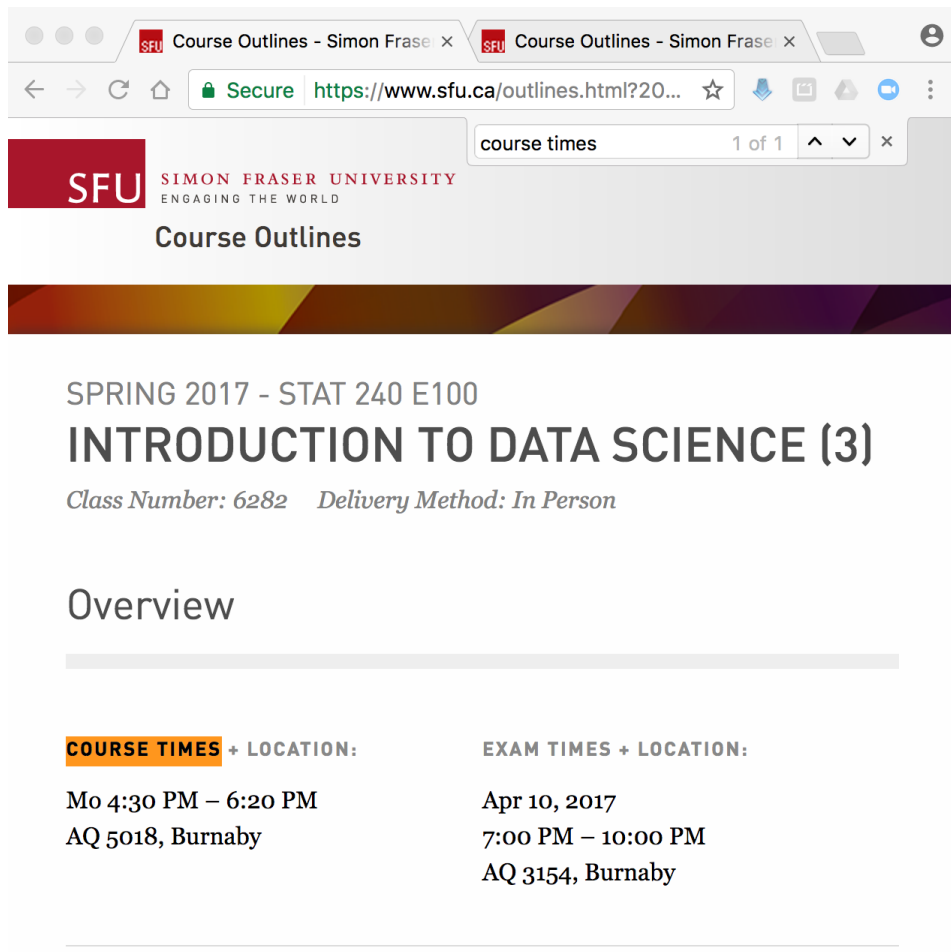
```
course_url = "https://www.sfu.ca/outlines.html?  
2018/spring/stat/240/d100"
```

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

Then use regular expressions!

Tips!

tags define boundaries



Course Outlines

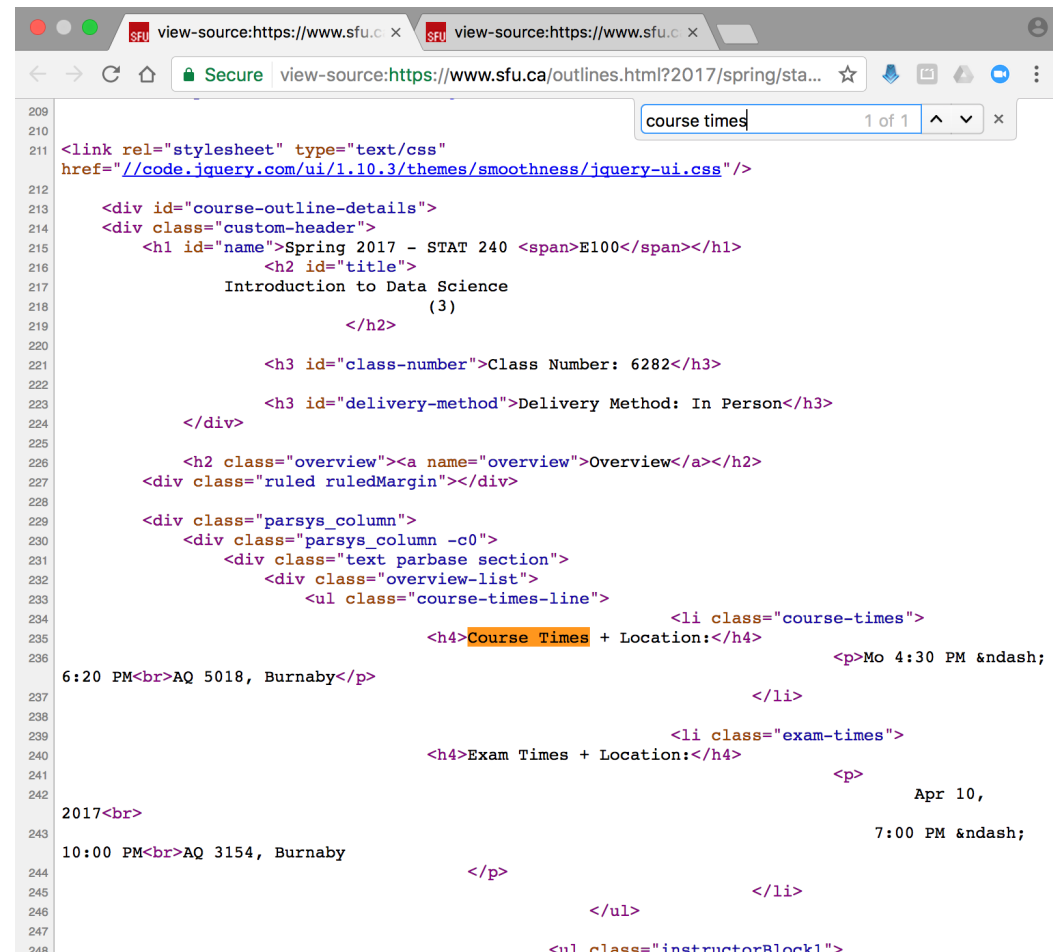
SPRING 2017 - STAT 240 E100

INTRODUCTION TO DATA SCIENCE (3)

Class Number: 6282 Delivery Method: In Person

Overview

COURSE TIMES + LOCATION:	EXAM TIMES + LOCATION:
Mo 4:30 PM – 6:20 PM AQ 5018, Burnaby	Apr 10, 2017 7:00 PM – 10:00 PM AQ 3154, Burnaby



```
<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,
                7:00 PM &ndash;
                10:00 PM<br>AQ 3154, Burnaby
              </p>
            </li>
          </ul>
        </div>
      </div>
    </div>
  </div>
</div>
```

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 \\\n+  
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 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:

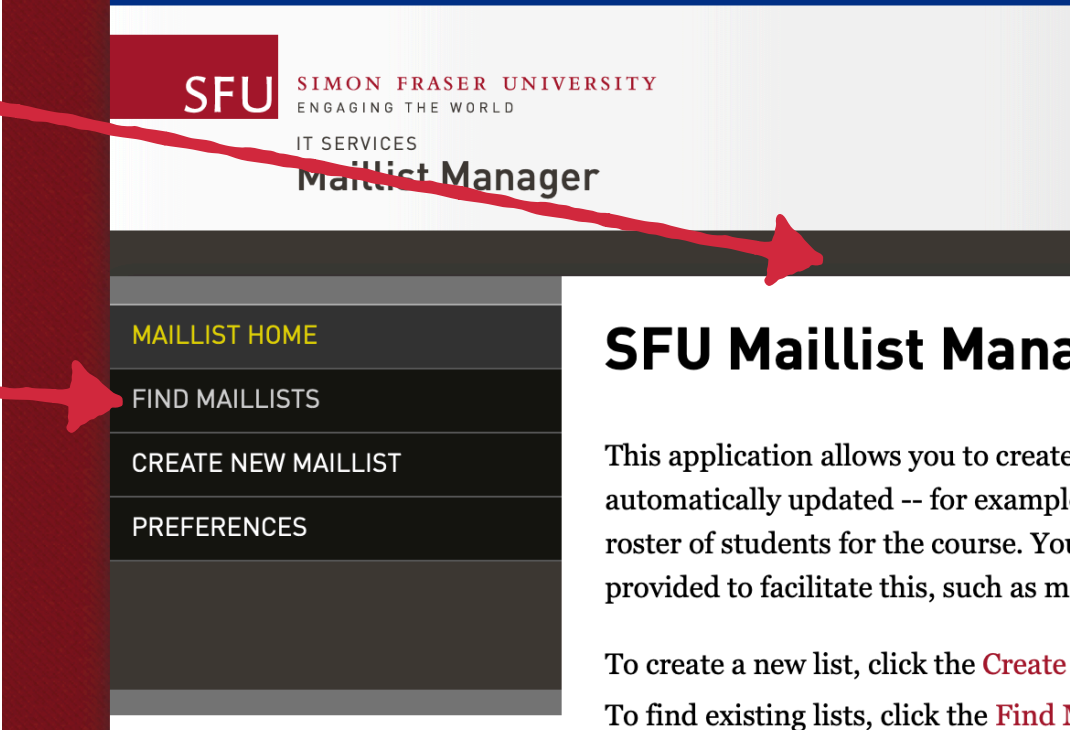
https://shiny.rcg.sfu.ca/u/your_SFU_username_goes_here/myapp

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



The screenshot shows the SFU Maillist Manager web application. At the top, the SFU logo and text "SIMON FRASER UNIVERSITY ENGAGING THE WORLD" are visible, followed by "IT SERVICES" and "Maillist Manager". A vertical red bar on the left contains a menu with the following items: "MAILLIST HOME", "FIND MAILLISTS", "CREATE NEW MAILLIST", and "PREFERENCES". A red arrow points from the text "Log into https://amaint.sfu.ca/cgi-bin/WebObjects/Maillist.woa" to the top of this menu bar. Another red arrow points from the text "go to Manage your Maillists" to the "FIND MAILLISTS" link. To the right of the menu, the heading "SFU Maillist Manager" is displayed, followed by a paragraph: "This application allows you to create automatically updated -- for example, a roster of students for the course. You are provided to facilitate this, such as m...". Below this, two instructions are given: "To create a new list, click the Create" and "To find existing lists, click the Find".

SFU SIMON FRASER UNIVERSITY
ENGAGING THE WORLD
IT SERVICES
Maillist Manager

MAILLIST HOME
FIND MAILLISTS
CREATE NEW MAILLIST
PREFERENCES

SFU Maillist Manager

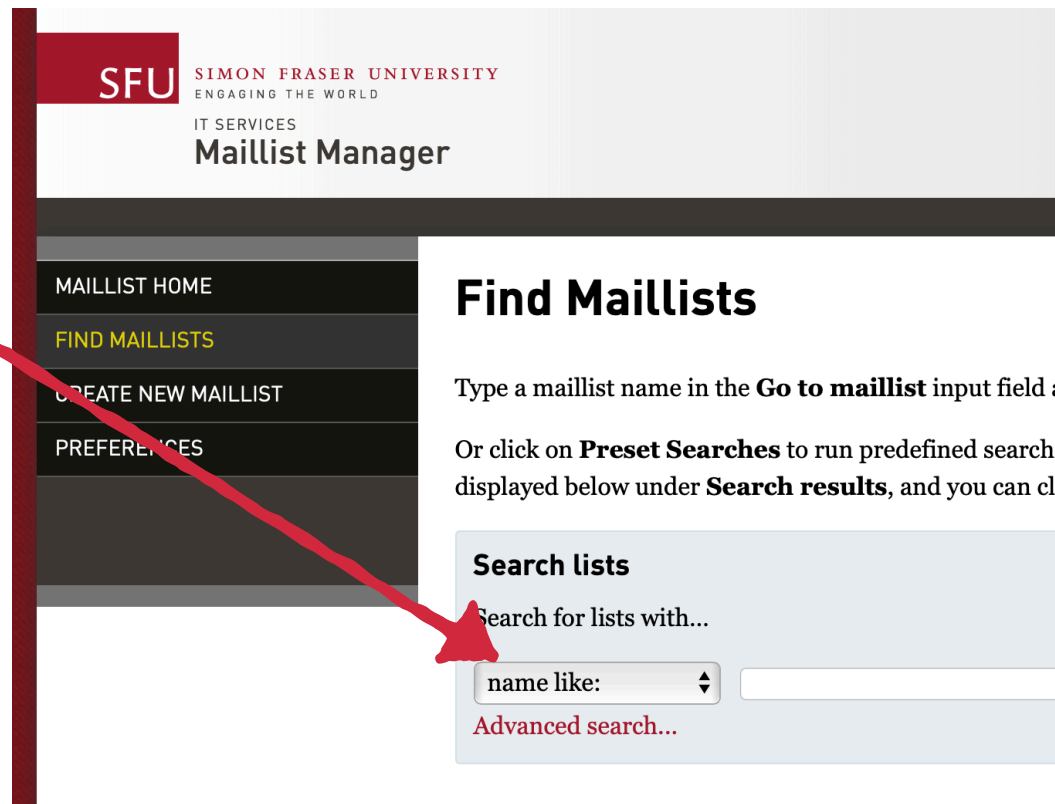
This application allows you to create automatically updated -- for example, a roster of students for the course. You are provided to facilitate this, such as m...

To create a new list, click the **Create**
To find existing lists, click the **Find**

Search for the rcg-shiny-users
maillist here

Then tell it you want to
subscribe

Procrastinators Beware: (Then
wait ~1/2 hour for the next step)



The screenshot shows the SFU Maillist Manager interface. At the top, the SFU logo and 'SIMON FRASER UNIVERSITY ENGAGING THE WORLD' are visible, along with 'IT SERVICES Maillist Manager'. A sidebar on the left contains links: 'MAILLIST HOME', 'FIND MAILLISTS' (highlighted in yellow), 'CREATE NEW MAILLIST', and 'PREFERENCES'. The main content area is titled 'Find Maillists' and includes instructions: 'Type a maillist name in the **Go to maillist** input field' and 'Or click on **Preset Searches** to run predefined searches displayed below under **Search results**, and you can click on the results to subscribe'. Below this is a 'Search lists' section with the text 'Search for lists with...'. It features a dropdown menu currently set to 'name like:' and an adjacent text input field. A red arrow points from the text 'Search for the rcg-shiny-users maillist here' to this input field. Below the input field is a link labeled 'Advanced search...'.