Digital Trace Data

Part 2

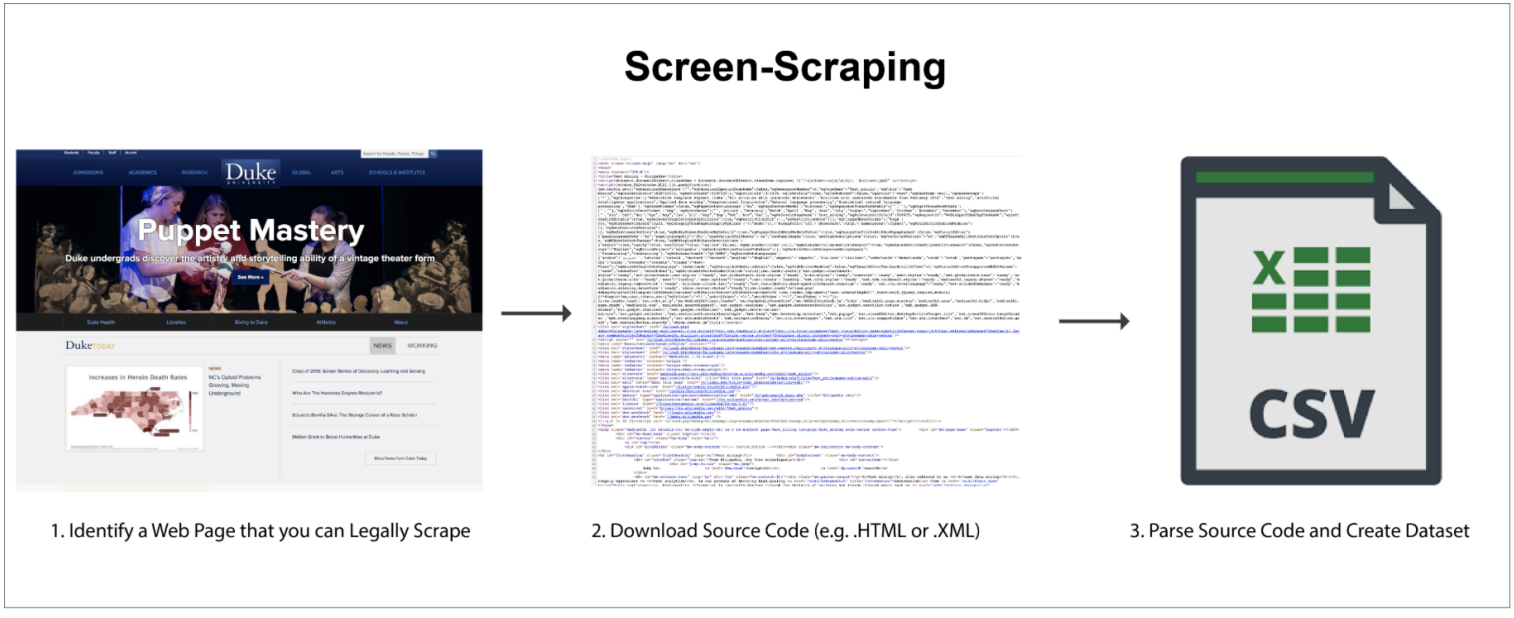
Ridhi Kashyap

University of Oxford

SICSS-Oxford
June 18, 2019

Web Scraping

- The process of automatically extracting data from web pages



Material on screen-scraping in these slides draw on
Chris Bail's tutorial on screen-scraping

Legal Issues in Web Scraping

- Website Terms of Service: are you allowed to do this?
- Websites tend to have a "robots.txt" policy that specifies rules about automated data collection on the site.
- Larger websites like Facebook, Instagram, NY Times do now allow these practices – but some tend to have Application Programming Interfaces (more later) to facilitate access to their public data.

Robot.txt

```
# robots.txt for http://www.wikipedia.org/ and friends
#
# Please note: There are a lot of pages on this site, and there are
# some misbehaved spiders out there that go _way_ too fast. If you're
# irresponsible, your access to the site may be blocked.
#

# Observed spamming large amounts of https://en.wikipedia.org/?curid=NNNNNN
# and ignoring 429 ratelimit responses, claims to respect robots:
# http://mj12bot.com/
User-agent: MJ12bot
Disallow: /

# advertising-related bots:
User-agent: Mediapartners-Google*
Disallow: /

# Wikipedia work bots:
User-agent: IsraBot
Disallow:
```

Wikipedia Page

Create account Log in

Article Talk Read Edit View history Search

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here

World Health Organization ranking of health systems in 2000

From Wikipedia, the free encyclopedia

The **World Health Organization (WHO) ranked the health systems** of its 191 member states in its World Health Report[1] 2000. It provided a framework and measurement approach to examine and compare aspects of health systems around the world.[2] It developed a series of performance indicators to assess the overall level and distribution of health in the populations, and the responsiveness and financing of health care services. It was the organization's first ever analysis of the world's health systems.[3]

Contents [hide]

1 Ranking
2 Methodology
3 Criticism
4 See also
5 References

Wikipedia Page HTML

```
<!DOCTYPE html>
<html class="client-nojs" lang="en" dir="ltr">
<head>
<meta charset="UTF-8"/>
<title>World Health Organization ranking of health systems in 2000 - Wikipedia</title>
<script>document.documentElement.className = document.documentElement.className.replace( /(^|\s)client-nojs(\s|$)/, "$1client-js$2" );</script>
<script>(window.RLQ=window.RLQ||[]).push(function()
{mw.config.set({"wgCanonicalNamespace":"","wgCanonicalSpecialPageName":false,"wgNamespaceNumber":0,"wgPageName":"World_Health_Organization_ranking_of_health_systems_in_2000","wgTitle":"Worl
d Health Organization ranking of health systems in
2000","wgCurRevisionId":809846878,"wgRevisionId":809846878,"wgArticleId":24088175,"wgIsArticle":true,"wgIsRedirect":false,"wgAction":"view","wgUserName":null,"wgUserGroups":
["*"],"wgCategories":["Health by country","World Health Organization"],"wgBreakFrames":false,"wgPageContentLanguage":"en","wgPageContentModel":"wikitext","wgSeparatorTransformTable":
["",""],"wgDigitTransformTable":["",""],"wgDefaultDateFormat":"dmy","wgMonthNames":
["","January","February","March","April","May","June","July","August","September","October","November","December"],"wgMonthNamesShort":
["","Jan","Feb","Mar","Apr","May","Jun","Jul","Aug","Sep","Oct","Nov","Dec"],"wgRelevantPageName":"World_Health_Organization_ranking_of_health_systems_in_2000","wgRelevantArticleId":2408817
5,"wgRequestId":"Wv0axwpAAD4AAC6g9PsAAAAL","wgIsProbablyEditable":true,"wgRelevantPageIsProbablyEditable":true,"wgRestrictionEdit":[],"wgRestrictionMove":[],"wgFlaggedRevsParams":{"tags":
{}},"wgStableRevisionId":null,"wgCategoryTreePageCategoryOptions":"{\"mode\":0,\"hideprefix\":20,\"showcount\":true,\"namespaces\":false}","wgWikiEditorEnabledModules":
[],"wgBetaFeaturesFeatures":
[],"wgMediaViewerOnClick":true,"wgMediaViewerEnabledByDefault":true,"wgPopupsShouldSendModuleToUser":true,"wgPopupsConflictsWithNavPopupGadget":false,"wgVisualEditor":
{"pageLanguageCode":"en","pageLanguageDir":"ltr","pageVariantFallbacks":"en","usePageImages":true,"usePageDescriptions":true},"wgPreferredVariant":"en","wgMFExpandAllSectionsUserOption":tru
e,"wgMFEnableFontChanger":true,"wgMFDisplayWikibaseDescriptions":
{"search":true,"nearby":true,"watchlist":true,"tagline":false},"wgRelatedArticles":null,"wgRelatedArticlesUseCirrusSearch":true,"wgRelatedArticlesOnlyUseCirrusSearch":false,"wgULSCurrentAut
onym":"English","wgNoticeProject":"wikipedia","wgCentralNoticeCookiesToDelete":[],"wgCentralNoticeCategoriesUsingLegacy":
["Fundraising","fundraising"],"wgWikibaseItemId":"Q8035804","wgScoreNoteLanguages":
{"arabic":"العربية","catalan":"català","deutsch":"Deutsch","english":"English","espanol":"español","italiano":"italiano","nederlands":"Nederlands","norsk":"norsk","portugues":"português","su
omi":"suomi","svenska":"svenska","vlaams":"West-
Vlams"},"wgScoreDefaultNoteLanguage":"nederlands","wgCentralAuthMobileDomain":false,"wgCodeMirrorEnabled":false,"wgVisualEditorToolbarScrollOffset":0,"wgVisualEditorUnsupportedEditParams":
["undo","undoafter","veswitched"],"wgEditSubmitButtonLabelPublish":true});mw.loader.state({"ext.gadget.charinsert-
styles":"ready","ext.globalCssJs.user.styles":"ready","ext.globalCssJs.site.styles":"ready","site.styles":"ready","noscript":"ready","user.styles":"ready","ext.globalCssJs.user":"ready","ex
t.globalCssJs.site":"ready","user":"loading","user.options":"ready","user.tokens":"loading","ext.cite.styles":"ready","mediawiki.legacy.shared":"ready","mediawiki.legacy.commonPrint":"ready
","jquery.tablesorter.styles":"ready","wikibase.client.init":"ready","ext.visualEditor.desktopArticleTarget.noscript":"ready","ext.uls.interlanguage":"ready","ext.wikimediaBadges":"ready","
mediawiki.skinning.interface":"ready","skins.vector.styles":"ready"});mw.loader.load("/w/load.php?
debug=false\u0026lang=en\u0026modules=user\u0026skin=vector\u0026version=0um5hfh");mw.loader.implement("user.tokens@ldqfd7l",function($,jQuery,require,module)
{/*@nomin*/mw.user.tokens.set({"editToken":"+\\","patrolToken":"+\\","watchToken":"+\\","csrfToken":"+\\"});
});mw.loader.load(["ext.cite.a11y","site","mediawiki.page.startup","mediawiki.user","mediawiki.hidpi","mediawiki.page.ready","jquery.tablesorter","mediawiki.toc","mediawiki.searchSuggest","
ext.gadget.teahouse","ext.gadget.ReferenceTooltips","ext.gadget.watchlist-notice","ext.gadget.DRN-wizard","ext.gadget.charinsert","ext.gadget.refToolbar","ext.gadget.extra-toolbar-
buttons","ext.gadget.switcher","ext.centralauth.centralautologin","mmv.head","mmv.bootstrap.autostart","ext.popups","ext.visualEditor.desktopArticleTarget.init","ext.visualEditor.targetLoad
er","ext.eventLogging.subscriber","ext.wikimediaEvents","ext.navigationTiming","ext.uls.eventlogger","ext.uls.init","ext.uls.compactlinks","ext.uls.interface","ext.3d","ext.centralNotice.ge
oIP","ext.centralNotice.startUp","skins.vector.js"]);});</script>
<link rel="stylesheet" href="/w/load.php?
debug=false&amp;lang=en&amp;modules=ext.cite.styles%7Cext.uls.interlanguage%7Cext.visualEditor.desktopArticleTarget.noscript%7Cext.wikimediaBadges%7Cjquery.tablesorter.styles%7Cmediawiki.le
gacy.commonPrint%2Cshared%7Cmediawiki.skinning.interface%7Cskins.vector.styles%7Cwikibase.client.init&amp;only=styles&amp;skin=vector"/>
<script async="" src="/w/load.php?debug=false&amp;lang=en&amp;modules=startup&amp;only=scripts&amp;skin=vector"></script>
<meta name="ResourceLoaderDynamicStyles" content=""/>
<link rel="stylesheet" href="/w/load.php?debug=false&amp;lang=en&amp;modules=ext.gadget.charinsert-styles&amp;only=styles&amp;skin=vector"/>
<link rel="stylesheet" href="/w/load.php?debug=false&amp;lang=en&amp;modules=site.styles&amp;only=styles&amp;skin=vector"/>
<meta name="generator" content="MediaWiki 1.32.0-wmf.3"/>
<meta name="referrer" content="origin"/>
<meta name="referrer" content="origin-when-crossorigin"/>
<meta name="referrer" content="origin-when-cross-origin"/>
<link rel="alternate" href="android-app://org.wikipedia/http/en.m.wikipedia.org/wiki/World_Health_Organization_ranking_of_health_systems_in_2000"/>
<link rel="alternate" type="application/x-wiki" title="Edit this page" href="/w/index.php?title=World_Health_Organization_ranking_of_health_systems_in_2000&amp;action=edit"/>
<link rel="edit" title="Edit this page" href="/w/index.php?title=World_Health_Organization_ranking_of_health_systems_in_2000&amp;action=edit"/>
<link rel="apple-touch-icon" href="/static/apple-touch/wikipedia.png"/>
<link rel="shortcut icon" href="/static/favicon/wikipedia.ico"/>
<link rel="search" type="application/opensearchdescription+xml" href="/w/opensearch_desc.php" title="Wikipedia (en)"/>
<link rel="EditURI" type="application/rsd+xml" href="//en.wikipedia.org/w/api.php?action=rsd"/>
<link rel="license" href="//creativecommons.org/licenses/by-sa/3.0/"/>
<link rel="canonical" href="https://en.wikipedia.org/wiki/World_Health_Organization_ranking_of_health_systems_in_2000"/>
<link rel="dns-prefetch" href="//login.wikimedia.org"/>
<link rel="dns-prefetch" href="//meta.wikimedia.org" />
<!--[if lt IE 9]><script src="/w/load.php?debug=false&amp;lang=en&amp;modules=html5shiv&amp;only=scripts&amp;skin=vector&amp;sync=1"></script><![endif]-->
</head>
<body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-subject page-World_Health_Organization_ranking_of_health_systems_in_2000 rootpage-
World_Health_Organization_ranking_of_health_systems_in_2000 skin-vector action-view">       <div id="mw-page-base" class="noprint"></div>
        <div id="mw-head-base" class="noprint"></div>
        <div id="content" class="mw-body" role="main">
            <a id="top"></a>
            <div id="siteNotice" class="mw-body-content"><!-- CentralNotice --></div><div class="mw-indicators mw-body-content">
</div>
<h1 id="firstHeading" class="firstHeading" lang="en">World Health Organization ranking of health systems in 2000</h1>          <div id="bodyContent" class="mw-body-content">
                <div id="siteSub" class="noprint">From Wikipedia, the free encyclopedia</div>              <div id="contentSub"></div>
                            <div id="jump-to-nav" class="mw-jump">
                    Jump to:                    <a href="#mw-head">navigation</a>,                 <a href="#p-search">search</a>
                </div>
                <div id="mw-content-text" lang="en" dir="ltr" class="mw-content-ltr"><div class="mw-parser-output"><p>The <b><a href="/wiki/World_Health_Organization" title="World Health
```

Wikipedia Page

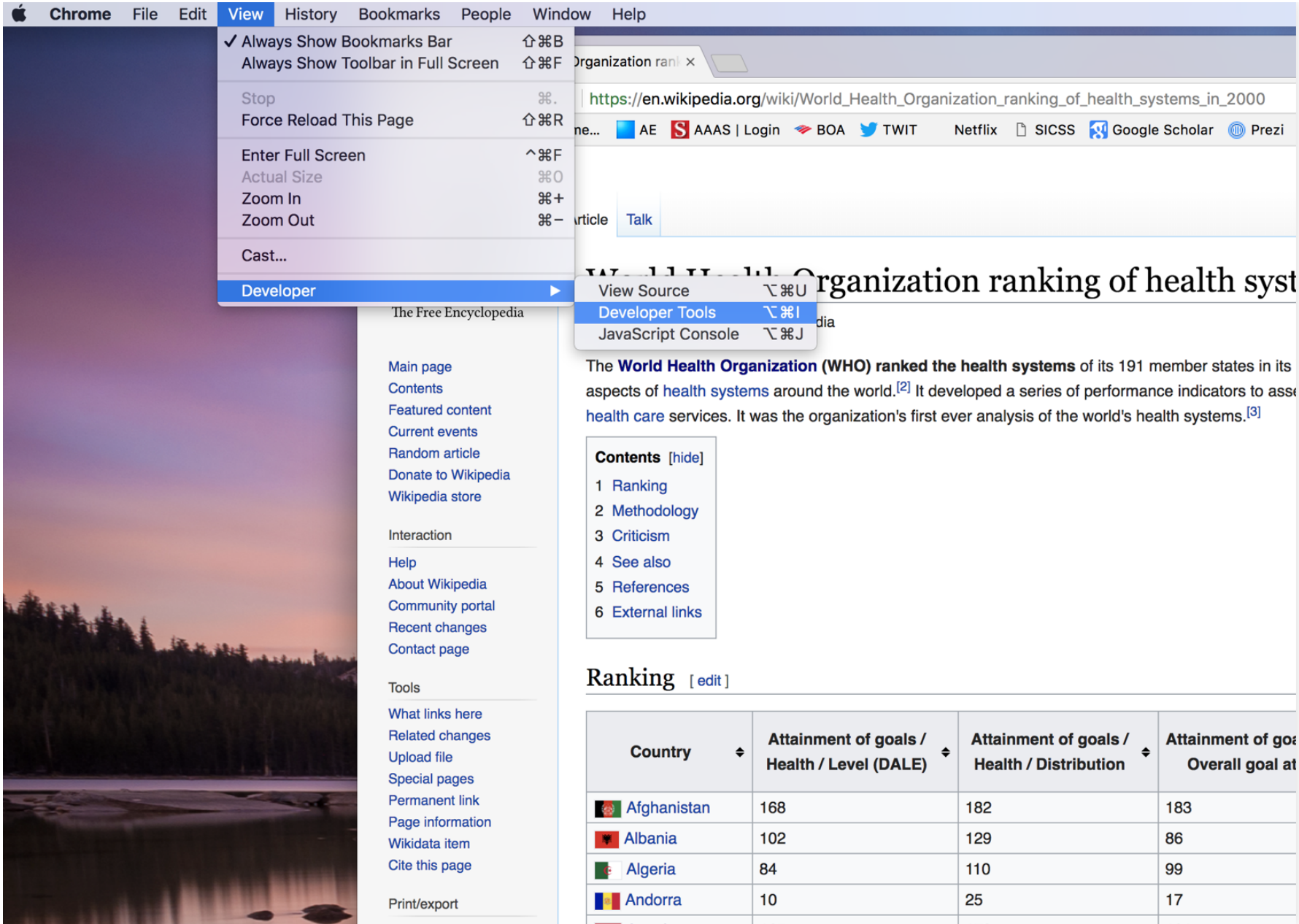
```r
#install package
install.packages("rvest")
library(rvest)

wikipedia_page<-read_html("https://en.wikipedia.org/wiki/World_Health_Organization_ranking_of_health_systems_in_2000")
```

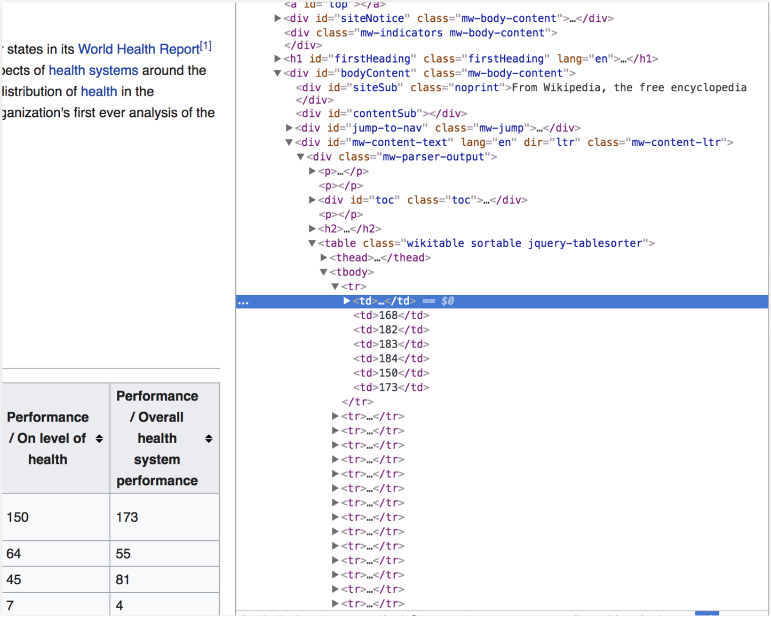
Wikipedia Page

```
{xml_document}
<html class="client-nojs" lang="en" dir="ltr">
[1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset=UTF-8">\n<meta charset="UTF-8">\n<title>World Health O ...
[2] <body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-subject mw-editable page-World_Health_Organization_rankin ...
```

Wikipedia Page

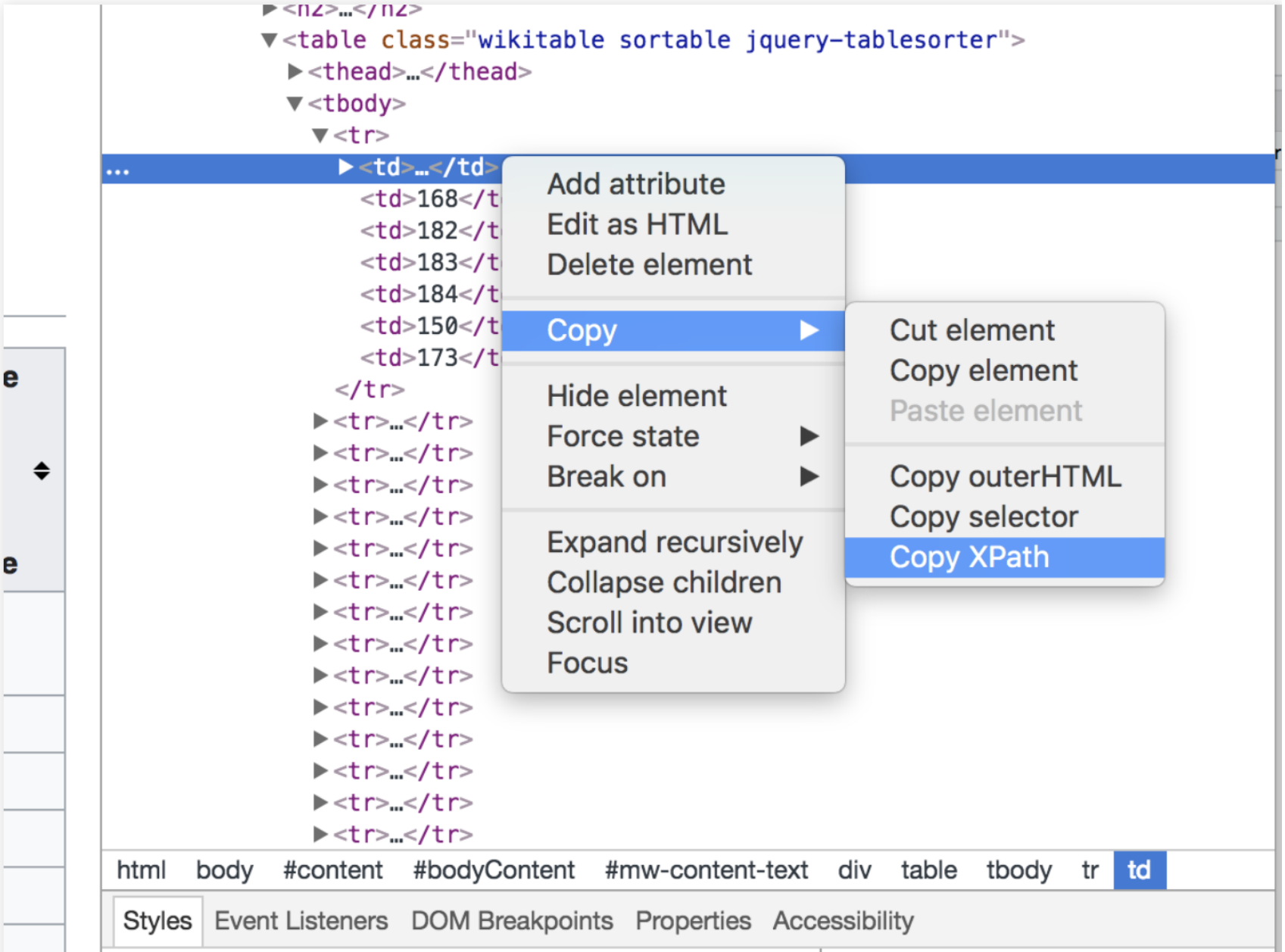


Inspect



states in its World Health Report[1]
pects of health systems around the
listribution of health in the
ganization's first ever analysis of the
Performance / On level of health
Performance / Overall health system performance
150
173
64
55
45
81
7
4
<div id="siteNotice" class="mw-body-content">…</div>
<div class="mw-indicators mw-body-content">
</div>
<h1 id="firstHeading" class="firstHeading" lang="en">…</h1>
<div id="bodyContent" class="mw-body-content">
<div id="siteSub" class="noprint">From Wikipedia, the free encyclopedia
</div>
<div id="contentSub"></div>
<div id="jump-to-nav" class="mw-jump">…</div>
<div id="mw-content-text" lang="en" dir="ltr" class="mw-content-ltr">
<div class="mw-parser-output">
<p>…</p>
<p></p>
<div id="toc" class="toc">…</div>
<p></p>
<h2>…</h2>
<table class="wikitable sortable jquery-tablesorter">
<thead>…</thead>
<tbody>
<tr>
<td>…</td> == $0
<td>168</td>
<td>182</td>
<td>183</td>
<td>184</td>
<td>150</td>
<td>173</td>
</tr>
<tr>…</tr>

Copy Xpath


▶<h2>…</h2>
▼<table class="wikitable sortable jquery-tablesorter">
▶<thead>…</thead>
▼<tbody>
▼<tr>
▶<td>…</td>
<td>168</t
<td>182</t
<td>183</t
<td>184</t
<td>150</t
<td>173</t
</tr>
▶<tr>…</tr>
▶<tr>…</tr>
▶<tr>…</tr>
▶<tr>…</tr>
▶<tr>…</tr>
▶<tr>…</tr>
▶<tr>…</tr>
▶<tr>…</tr>
▶<tr>…</tr>
▶<tr>…</tr>
▶<tr>…</tr>
▶<tr>…</tr>
▶<tr>…</tr>
▶<tr>…</tr>
Add attribute
Edit as HTML
Delete element
Copy
Hide element
Force state
Break on
Expand recursively
Collapse children
Scroll into view
Focus
Cut element
Copy element
Paste element
Copy outerHTML
Copy selector
Copy XPath
html body #content #bodyContent #mw-content-text div table tbody tr td
Styles Event Listeners DOM Breakpoints Properties Accessibility

HTML node

```
section_of_wikipedia <- html_node(wikipedia_page, xpath = '//*[@id="mw-content-text"]/div/table[2]')
head(section_of_wikipedia)
```

```
$node
<pointer: 0x7feb444a56d0>

$doc
<pointer: 0x7feb4449e0a0>
```

Extracting HTML Table

```
> health_rankings<-html_table(section_of_wikipedia)
> head(health_rankings[,(1:2)])
              Country Attainment of goals / Health / Level (DALE)
1         Afghanistan                                          168
2             Albania                                          102
3             Algeria                                           84
4             Andorra                                           10
5              Angola                                          165
6 Antigua and Barbuda                                           48
```

CSS Selector Gadget

- For more complex webpages, particularly those with CSS elements, other approaches might be needed.
- `http://selectorgadget.com/`
- Select css elements with Selector.

Selector Gadget

Selector Gadget

```
duke_page<-read_html("https://www.duke.edu")
duke_events<-html_nodes(duke_page, css="li:nth-child(1) .epsilon")
html_text(duke_events)
```

```
## [1] "Is There a Limit to Human Endurance? Science Says Yes\n\n\t\t\t\t\t\t\t\t"
## [2] "University Worship - Rev. Dr. Luke A. Powery"
## [3] "Duke Fuqua Insights: What Your Digital Footprint Says About Credit Worthiness"
```

Some Limits

- In practice, this can often yield messy data depending on web page complexity
- Web scraping can be frustrating, unfeasible or illegal. What then?
  - For complex web pages, crowd workers (Mechanical Turk) could be an option. Talk to Roberto at dinner.
  - Some websites make their content available through APIs.

What is an API?

- Application programming interfaces or APIs are a software intermediary that allows two applications to talk to each other.
- Web APIs allow one computer (a client) to ask another computer (a server) for some resource over the internet.

Application Programming Interface

- Modern APIs adhere to standards, that make data exchange programmatically accessible, safe and structured
- APIs generally require some kind of authentication and credentialling process that gives them more legitimacy.
- Contrast with web scraping.

API Directory

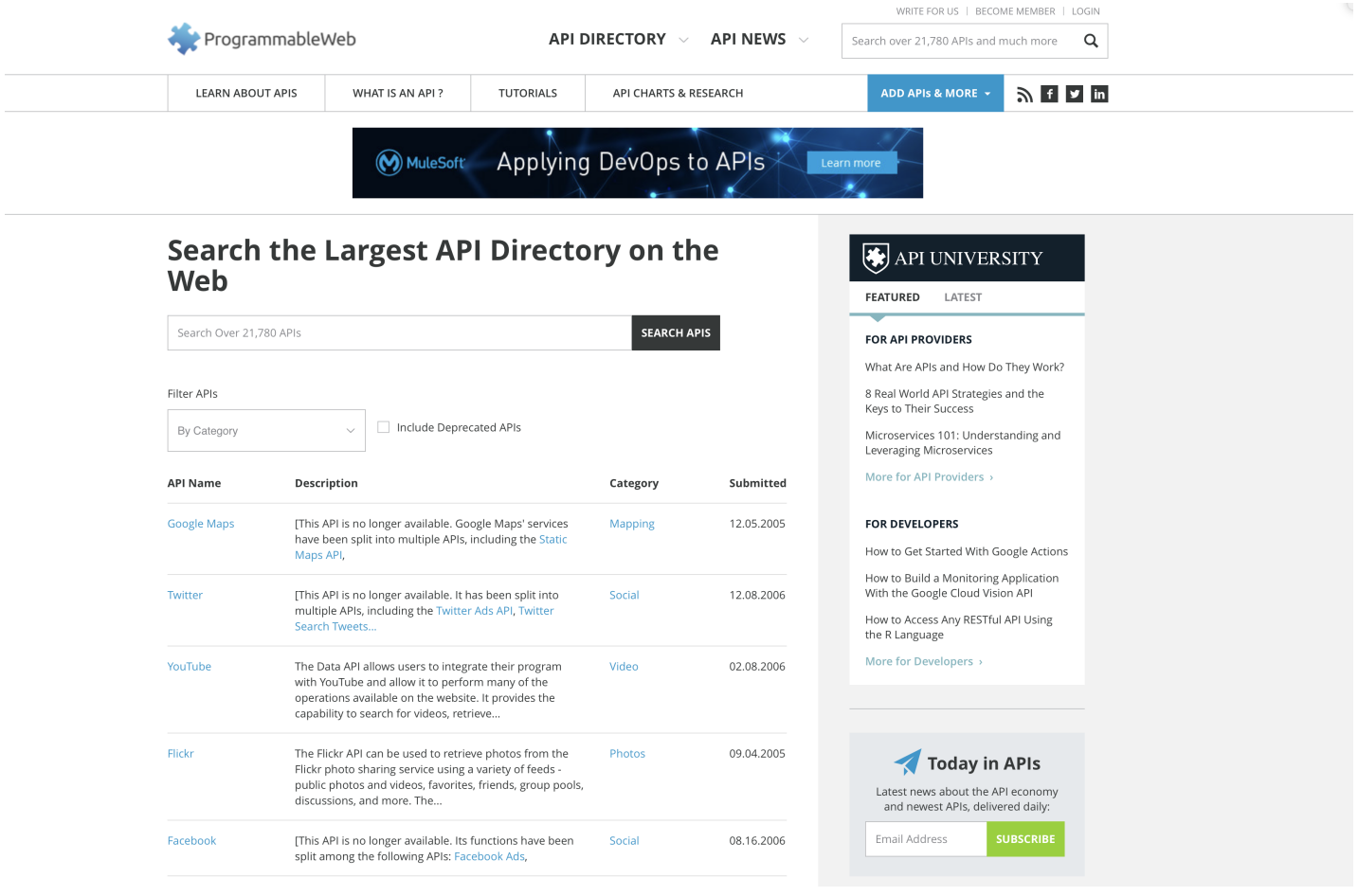


Figure: `https://www.programmableweb.com/apis/directory`

Extracting Data from APIs

- Web APIs tend to use HTTP methods.
- HTTP is the network protocol used to deliver virtually all files and other data on the World Wide Web, such as HTML files, image files, query results, etc
- These are verbs: GET(), POST(), DELETE(), and so on.
- GET() and POST()
  - GET() is used by your browser when requesting a page
  - POST() is (usually) used when submitting a form to a server.

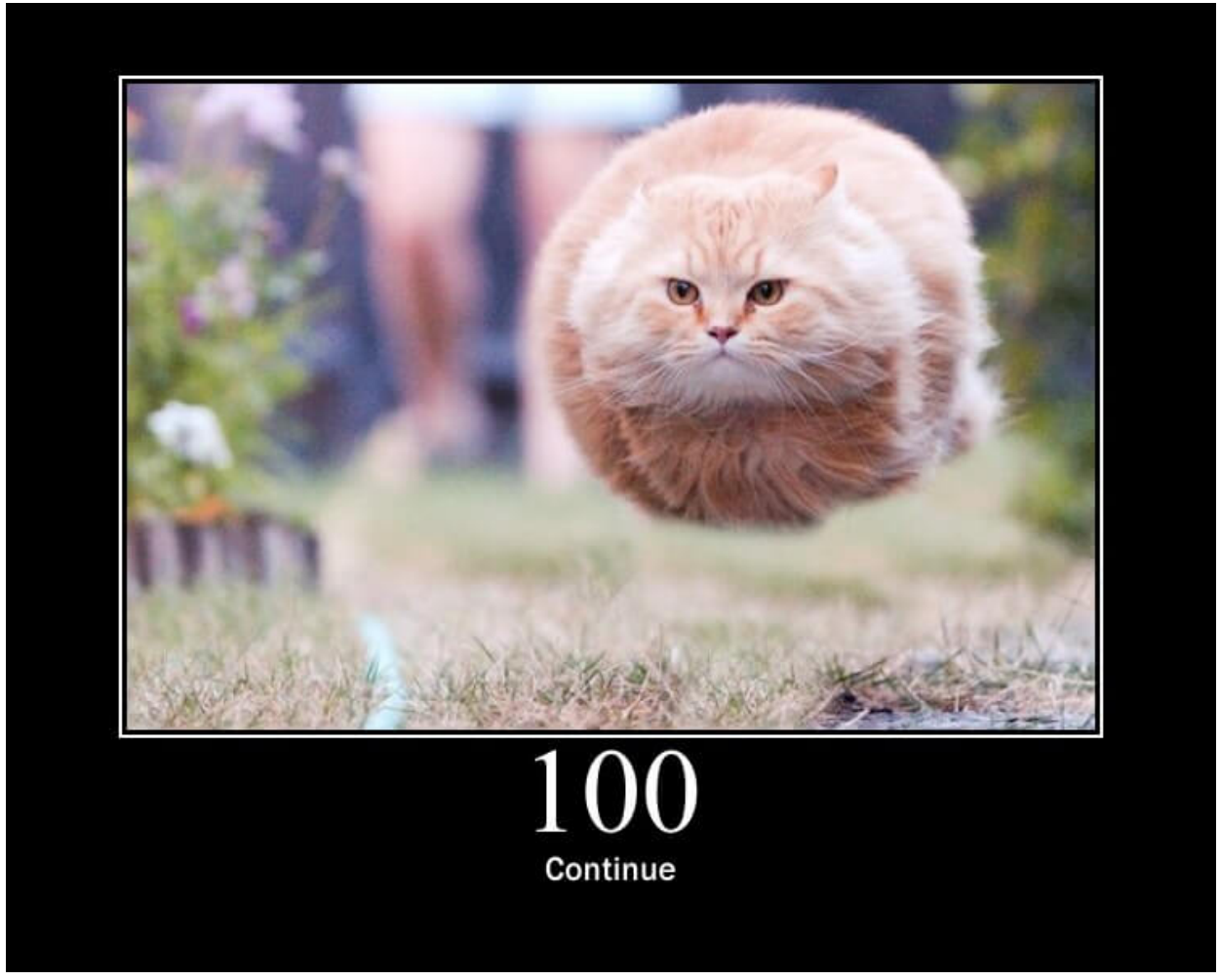
httr in R

- The `httr` package in R is useful to work with APIs
- You can make a request (to a url) and get a response.
- Response contains a status, header and body.

httr in R

```
> library(httr)
>
> trial <- GET("https://http.cat/100")
> status_code(trial)
[1] 200
>
> trial_content <- content(trial, as = 'raw')
> head(trial_content)
[1] ff d8 ff e0 00 10
>
> write_file(trial_content, "100.jpg")
```

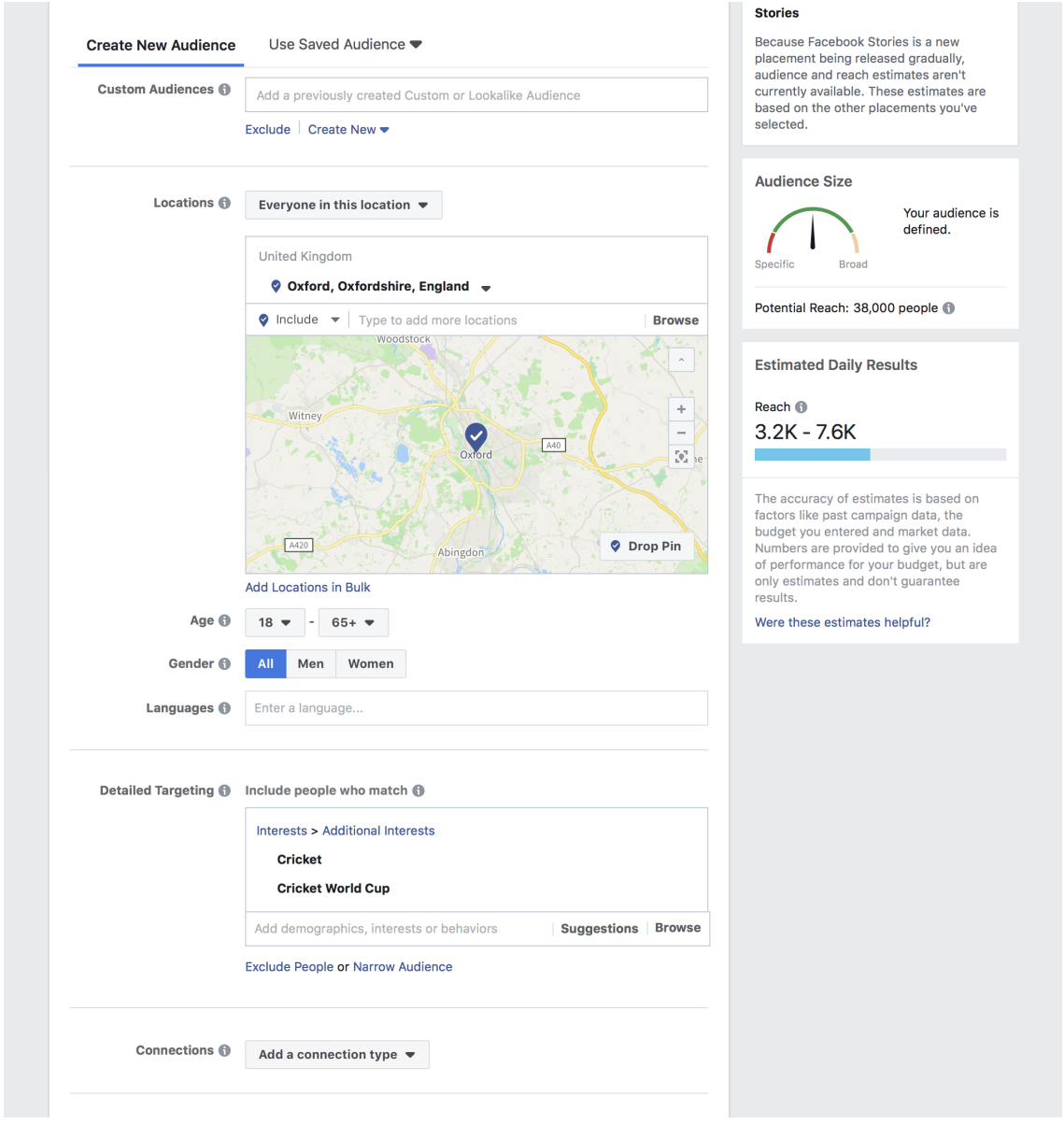
httr in R



Facebook Marketing API

- You need:
  - Facebook account
  - Marketing app with token and an ad account number ("act")
  - See this helpful tutorial by Sofia Gil-Clavel from MPIDR on how to obtain these credentials – `https://github.com/SofiaG1l/Using_Facebook_API/blob/master/First_Step.pdf`

Facebook Marketing API



Facebook Marketing API

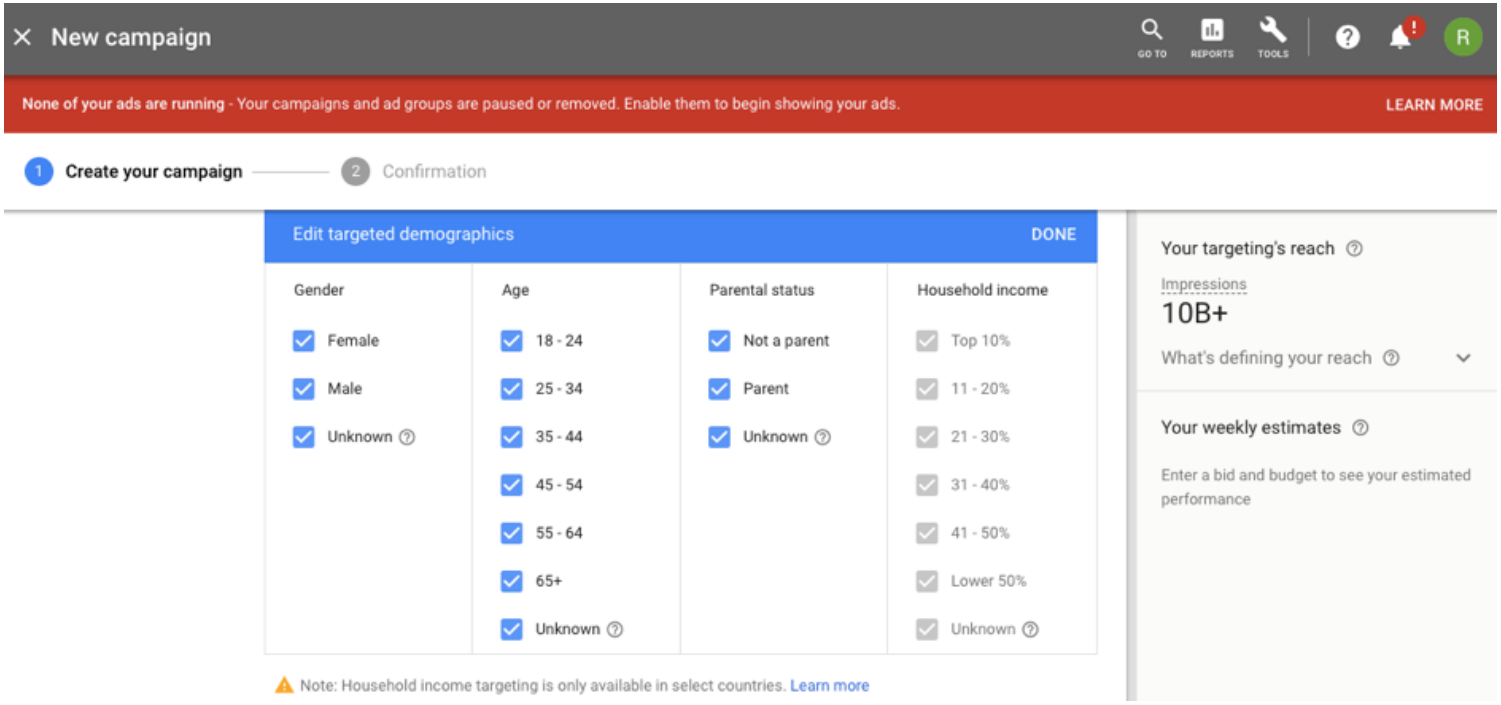
- We can programmatically make a query to the API to retrieve these ad audience estimates.
- For more information on how targeting specifications work – see `https://developers.facebook.com/docs/marketing-api/buying-api/targeting/`
- To search for available targeting options – see `https://developers.facebook.com/docs/marketing-api/targeting-search`

Facebook Marketing API

```
credentials_2 <- paste0('https://graph.facebook.com/',version,'/act_',act,'/delivery_estimate?access_token=', token)
```

- `version` refers to the version of the API (v 3.3. is current)
- `act` is the ad account number
- `token` is the access token
- Important to remember rate limiting when working with APIs

AdWords


New campaign
GO TO
REPORTS
TOOLS
R
None of your ads are running - Your campaigns and ad groups are paused or removed. Enable them to begin showing your ads.
LEARN MORE
1 Create your campaign
2 Confirmation
Edit targeted demographics
DONE
Gender
Female
Male
Unknown
Age
18 - 24
25 - 34
35 - 44
45 - 54
55 - 64
65+
Unknown
Parental status
Not a parent
Parent
Unknown
Household income
Top 10%
11 - 20%
21 - 30%
31 - 40%
41 - 50%
Lower 50%
Unknown
Note: Household income targeting is only available in select countries. Learn more
Your targeting's reach
Impressions
10B+
What's defining your reach
Your weekly estimates
Enter a bid and budget to see your estimated performance

Exercise

- Familiarise yourself (and obtain credentials) to work with one of the following three data sources:
  1. Facebook marketing API
  2. Twitter – use Chris Bail's helpful tutorial using the `rtweet` package.
  3. ...Or another API of your choosing.

Group Exercise

- Divide yourselves into groups of four by counting off in order around the room.
- Work together to identify a research question that you believe could be answered using one of the following three sources of digital trace data:
  1. Facebook marketing API
  2. Twitter
  3. Google Trends
- Think of a research design using digital trace data either on its own, or in combination with existing observational or survey data.
- Collect preliminary data to test the feasibility of your approach.
- What are the strengths and limitations of this approach?
- How could it be improved with a hybrid design?