Introduction to Data Science

Data Collection Methods & Ethics (Lecture 1)

Dr. Oul Han

Special thanks to:
Juhi Kulshrestha (http://www.juhikulshrestha.com/)
Aniko Hannak (http://ancsahannak.me)

Institute for Web Science and Technologies · University of Koblenz-Landau, Germany

What is data? One example of MANY

Qualitative and small

- Observation of individual cases
- Explaining, characterizing
- *Example*:
  - How many friends do overseas students have?
  - How to they interact with them in the holidays?

What is data? Another example of MANY

Quantitative and large

- Observation of categories
- Counting, sorting by features
- *Example*:
  - How many people in Koblenz are registered as self-employed?
  - How many of them are merchants, manufacturers, or service providers?

Data from online platforms

- Data is everywhere!
- The only limitation is your imagination
  - *And the terms of service*

(iconimage / stock.adobe.com)

Data from online platforms - Pros & Cons

- Larger & cheaper than surveys or field experiments
- Examine human interactions in their natural environments
- Immediate feedback after external events

Data from online platforms - Pros & Cons

- Larger & cheaper than surveys or field experiments
- Examine human interactions in their natural environments
- Immediate feedback after external events

- Big data isn’t more representative or of better quality
- Data-driven analysis, or over-simplified representation?
- Your conclusions may be wrong about the world (external validity)

Example: WeST Facebook group

Representative?

- “Does the WeST Facebook group represent the opinion of all WeST students?”

Wrong about the world?

- „A high number of likes in the WeST Facebook group means that all WeST students are highly satisfied”

In this class

Collecting data:
Ethical & legal
considerations

Overview of data
collection tools:
APIs
Web scraping
Browser automation

Sharing data:
Ethical & legal
considerations

Image source: https://www.flaticon.com/

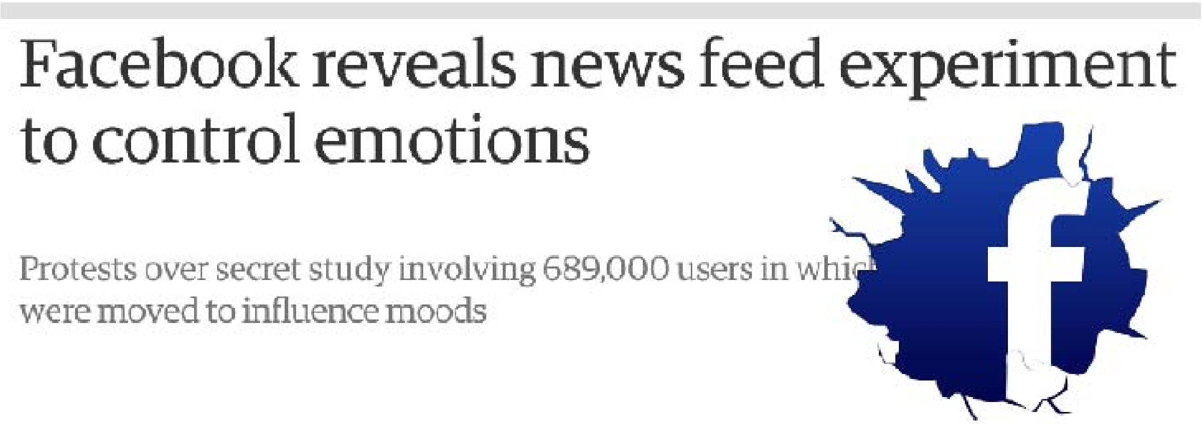
Am I harming the users?

Am I harming the platform?

Collecting Data: Ethical and Legal

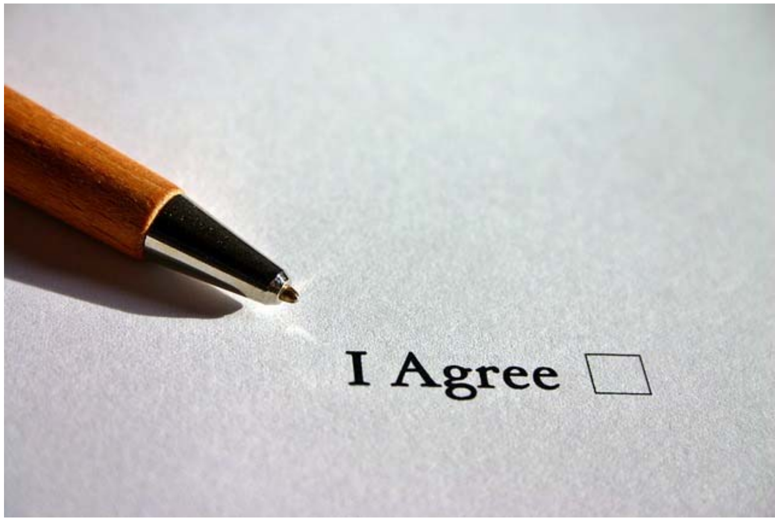
Am I harming the users?

- Interference through experiment
- Manipulating the user’s behavior without consent



Facebook reveals news feed experiment to control emotions

Protests over secret study involving 689,000 users in whic
were moved to influence moods



Am I harming the users?

- Collecting personal or sensitive information

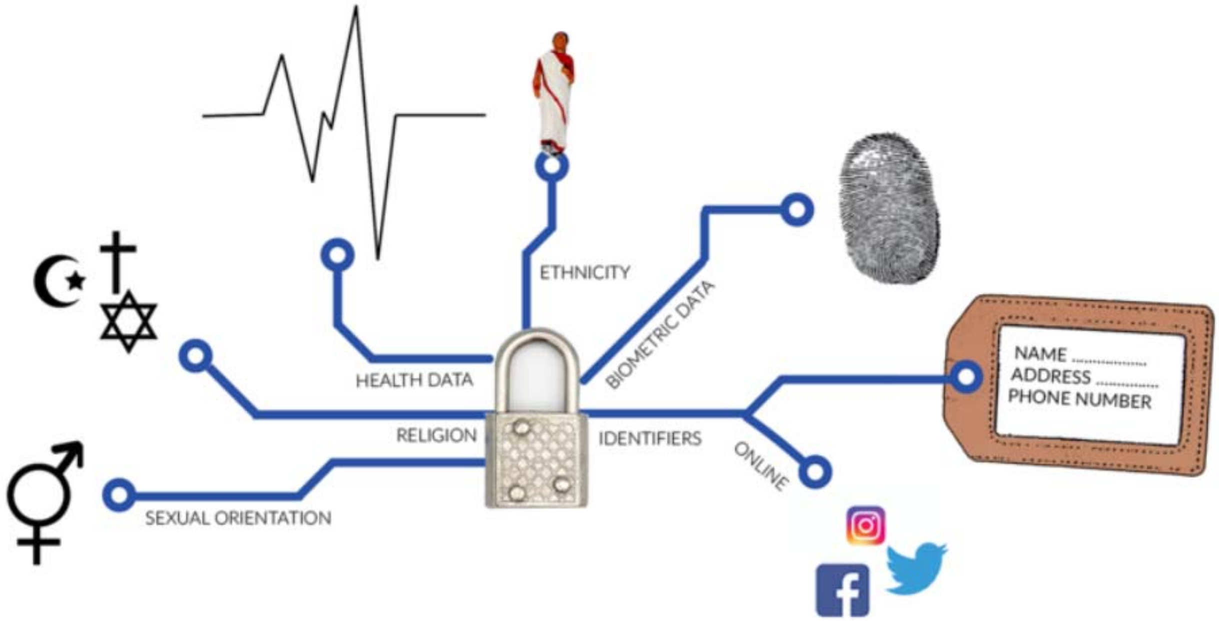


Image reused from CESSDA ERIC Expert Tour Guide on Data Management which is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

Am I harming the users?

- Personal information
  - Any data that can be used to identify living (or deceased) individuals
- Sensitive information
  - name or date of birth
  - person’s origin,
  - political opinion,
  - religious beliefs,
  - health,
  - trade union membership,
  - sexual orientation …
- If sample size is small
  - number of children a person has
  - shoe size …

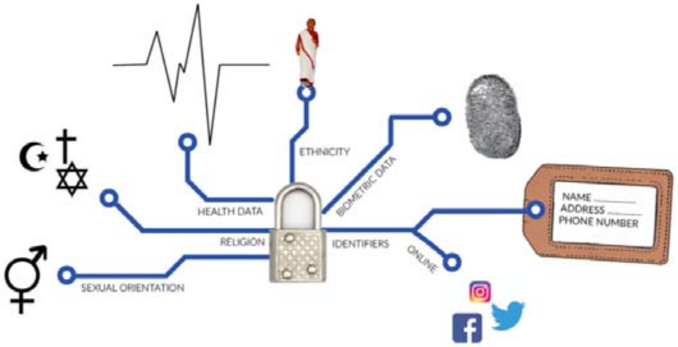


Image reused from CESSDA ERIC Expert Tour Guide on Data Management which is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License

Am I harming the users?

- Collect anonymously
- Use consent forms
- Anonymize data
- Securely store, access, transfer

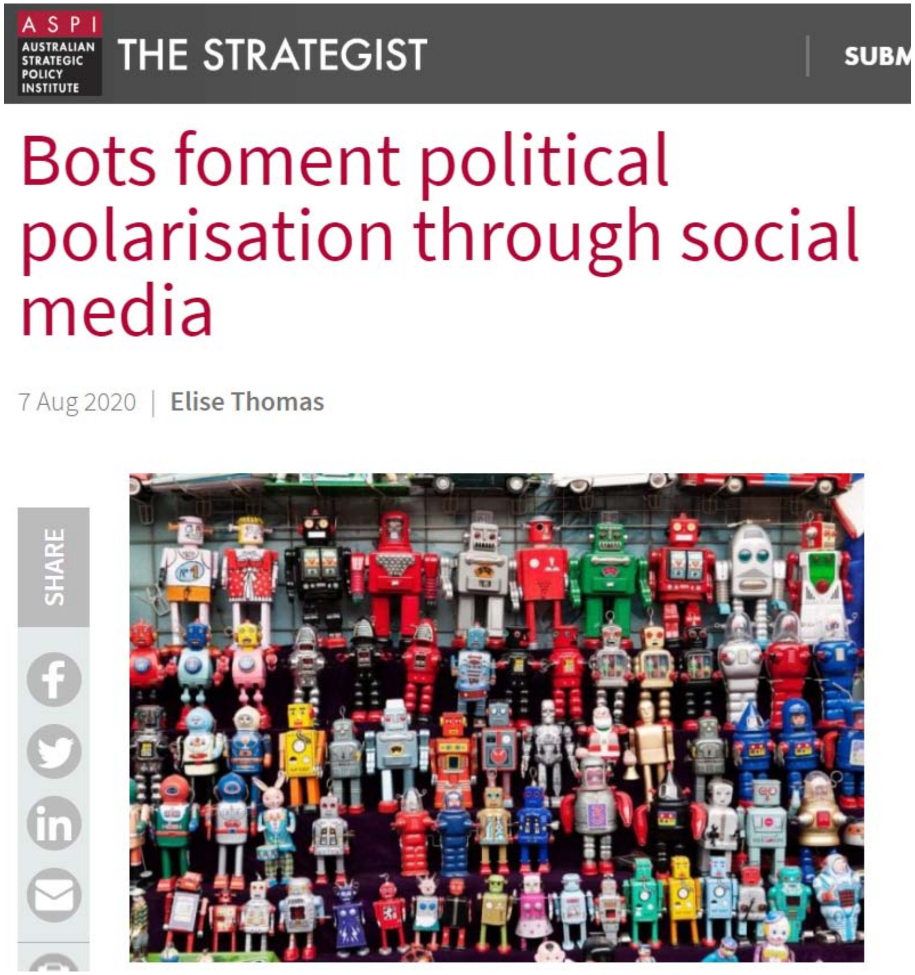
EU General Data Protection Regulation (GDPR)[1]

- Transparency
  - processing personal data “lawfully, fairly and in a transparent manner”
- Data Minimization
  - data use shall be limited to the purpose of the respective research
- Accuracy
  - inaccurate data must be “erased or rectified without delay”
- Integrity and Confidentiality
  - data must be protected by appropriate security measures (technical and organizational)

[1] https://www.fosteropenscience.eu/learning/data-protection-and-ethics/#/id/5ace27ca8ee5d6920ab94c13

Am I harming the platform?

- Interference with algorithms on the site by clicking or searching
  - Bots
  - Search words hijacking



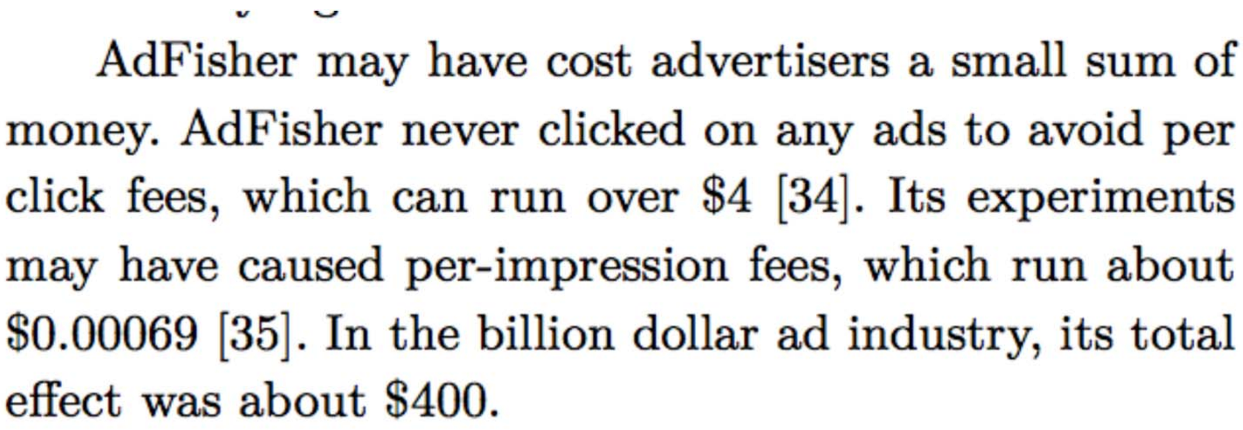
A S P I AUSTRALIAN STRATEGIC POLICY INSTITUTE THE STRATEGIST | SUBM

Bots foment political polarisation through social media

7 Aug 2020 | Elise Thomas

Am I harming the platform?

- Click fraud
  - Advertisers pay for every click or impression[1]



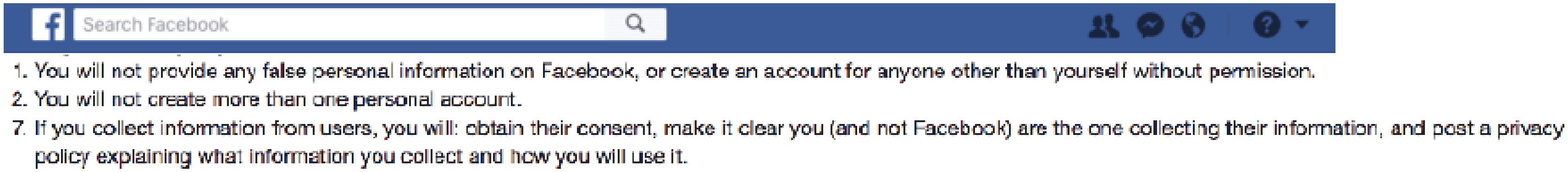
AdFisher may have cost advertisers a small sum of money. AdFisher never clicked on any ads to avoid per click fees, which can run over $4 [34]. Its experiments may have caused per-impression fees, which run about $0.00069 [35]. In the billion dollar ad industry, its total effect was about $400.

[1] Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *Proceedings on privacy enhancing technologies*, *2015*(1), 92-112.

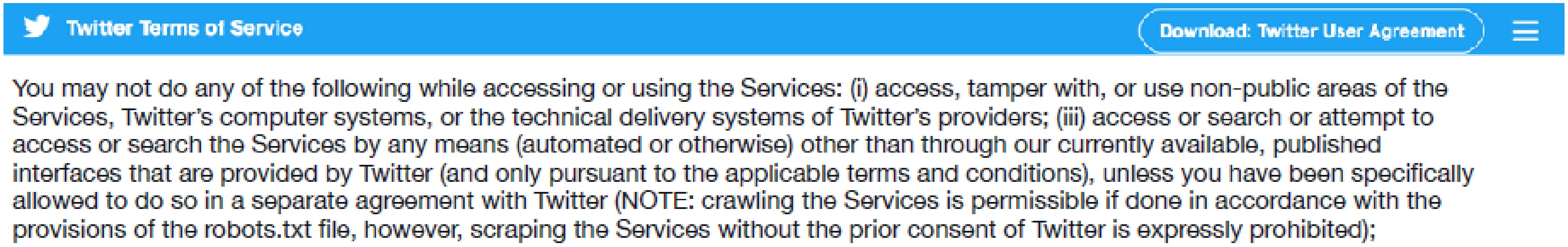
Am I harming the platform?

- Load balance
  - Check if an API exists or if data are available for download
  - Do not overload the servers such that their service is affected!

Terms of Service

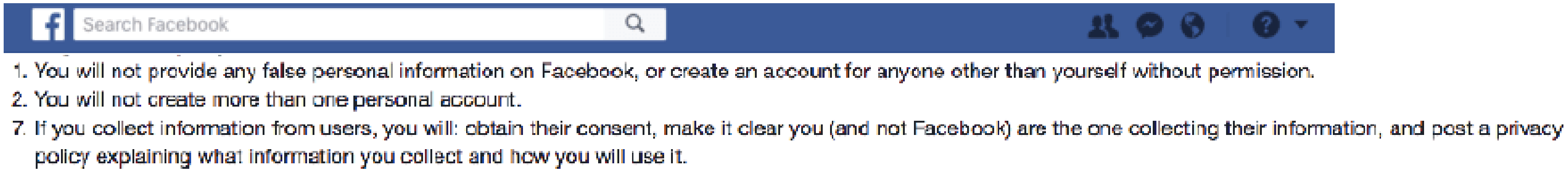


1. You will not provide any false personal information on Facebook, or create an account for anyone other than yourself without permission.
2. You will not create more than one personal account.
7. If you collect information from users, you will: obtain their consent, make it clear you (and not Facebook) are the one collecting their information, and post a privacy policy explaining what information you collect and how you will use it.

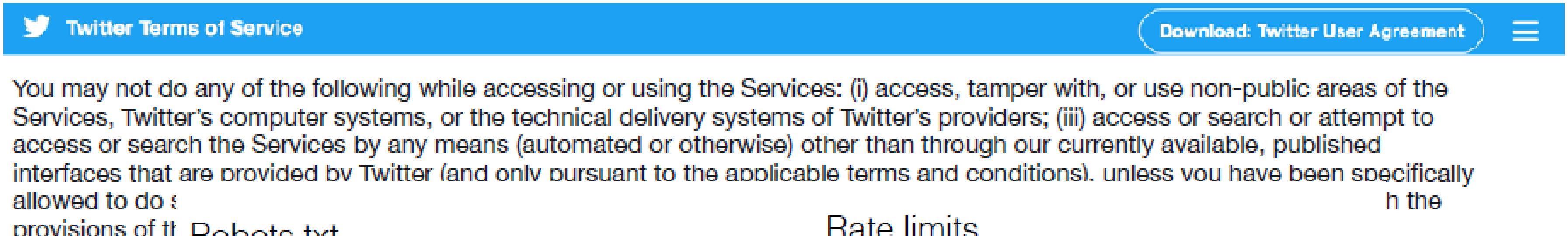


You may not do any of the following while accessing or using the Services: (i) access, tamper with, or use non-public areas of the Services, Twitter's computer systems, or the technical delivery systems of Twitter's providers; (iii) access or search or attempt to access or search the Services by any means (automated or otherwise) other than through our currently available, published interfaces that are provided by Twitter (and only pursuant to the applicable terms and conditions), unless you have been specifically allowed to do so in a separate agreement with Twitter (NOTE: crawling the Services is permissible if done in accordance with the provisions of the robots.txt file, however, scraping the Services without the prior consent of Twitter is expressly prohibited);

Terms of Service

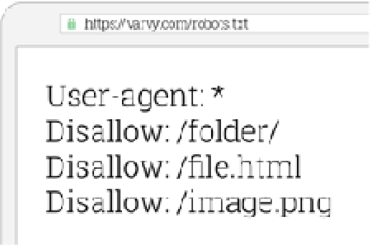


1. You will not provide any false personal information on Facebook, or create an account for anyone other than yourself without permission.
2. You will not create more than one personal account.
7. If you collect information from users, you will: obtain their consent, make it clear you (and not Facebook) are the one collecting their information, and post a privacy policy explaining what information you collect and how you will use it.



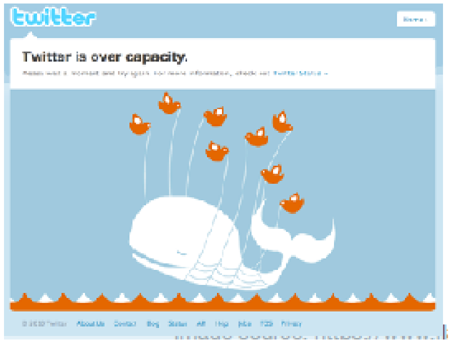
You may not do any of the following while accessing or using the Services: (i) access, tamper with, or use non-public areas of the Services, Twitter's computer systems, or the technical delivery systems of Twitter's providers; (iii) access or search or attempt to access or search the Services by any means (automated or otherwise) other than through our currently available, published interfaces that are provided by Twitter (and only pursuant to the applicable terms and conditions), unless you have been specifically allowed to do s h the provisions of tl

Robots.txt



```
User-agent: *
Disallow: /folder/
Disallow: /file.html
Disallow: /image.png
```

Rate limits



Exceptions for research

- Computer Fraud and Abuse Act (CFAA) (US)
  - Crawling is considered illegal for business practices such as discrimination

[1] https://www.fosteropenscience.eu/learning/data-protection-and-ethics/#/id/5ace27ca8ee5d6920ab94c13

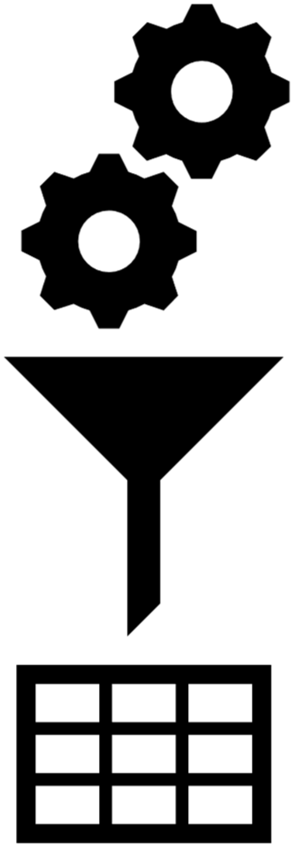
Exceptions for research

- Computer Fraud and Abuse Act (CFAA) (US)
  - Crawling is considered illegal for business practices such as discrimination
- GDPR research exemptions (EU)[1]
  - If for “the public interest, scientific or historical research purposes or statistical purposes” (Art. 5.1 2016/679/EU)
  - If “the data subject has given consent to the processing of his or her personal data for one or more specific purposes“ (Art. 6.1 2016/679/EU)

[1] https://www.fosteropenscience.eu/learning/data-protection-and-ethics/#/id/5ace27ca8ee5d6920ab94c13

Data collection methods

- APIs
- Web scraping
- Browser automation
- Personalized data collection

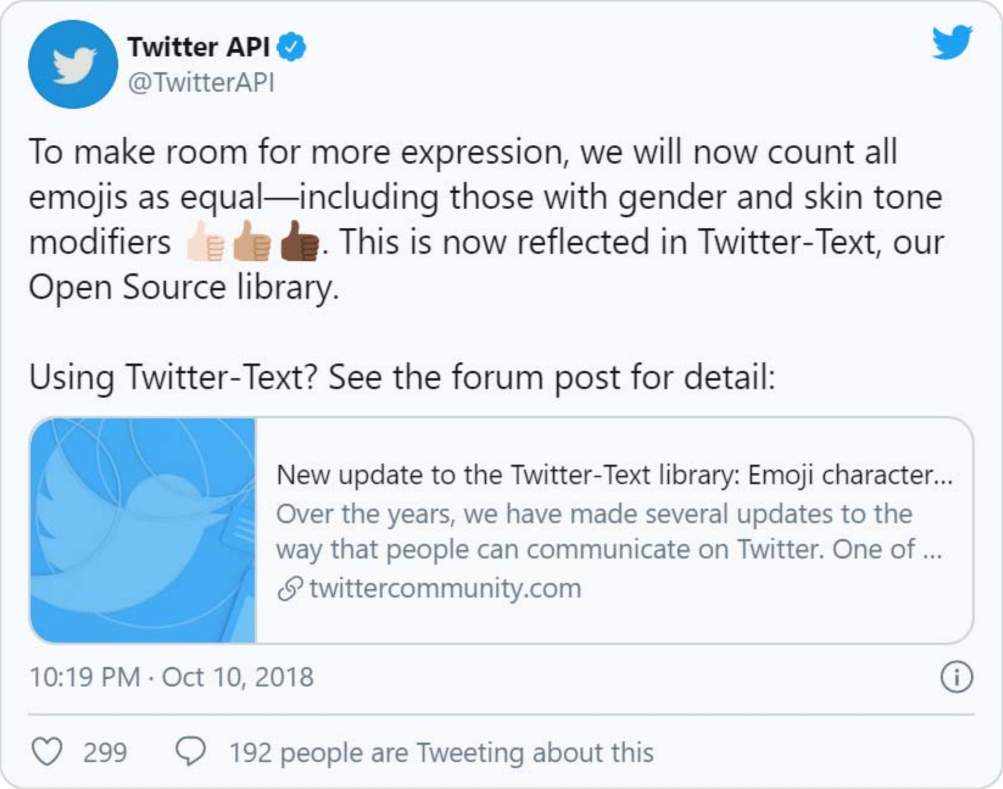


APIs

APIs

- API - Application Programming Interface
  - set of http requests that returns structured data (JSON, XML)
- Two types
  - Restful APIs
    - queries for static information at current moment
      - user profiles, posts, …
  - Streaming APIs
    - changes in users’ data in real time
      - new tweets, weather alerts, …

Twitter API example



Twitter API
@TwitterAPI

To make room for more expression, we will now count all emojis as equal—including those with gender and skin tone modifiers 👍🏻👍🏽👍🏿. This is now reflected in Twitter-Text, our Open Source library.

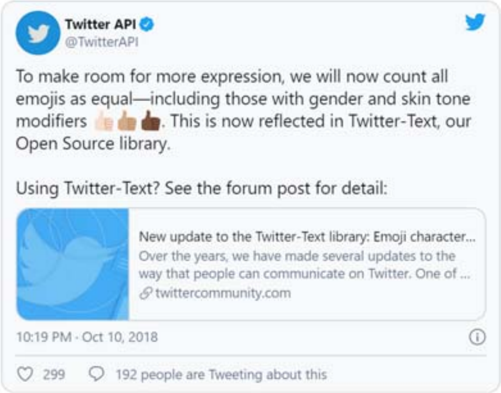
Using Twitter-Text? See the forum post for detail:

New update to the Twitter-Text library: Emoji character...
Over the years, we have made several updates to the way that people can communicate on Twitter. One of ...
twittercommunity.com

10:19 PM · Oct 10, 2018

299 192 people are Tweeting about this

Twitter API example



Twitter API
@TwitterAPI

To make room for more expression, we will now count all emojis as equal—including those with gender and skin tone modifiers 👍🏻👍🏽👍🏿. This is now reflected in Twitter-Text, our Open Source library.

Using Twitter-Text? See the forum post for detail:

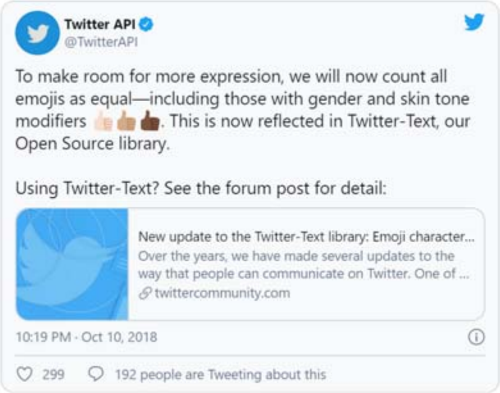
New update to the Twitter-Text library: Emoji character...
Over the years, we have made several updates to the way that people can communicate on Twitter. One of ...
twittercommunity.com

10:19 PM · Oct 10, 2018

299 192 people are Tweeting about this

```
{
 "created_at": "Wed Oct 10 20:19:24 +0000 2018",
 "id": 1050118621198921728,
 "id_str": "1050118621198921728",
 "text": "To make room for more expression, we will
 now count all emojis as equal—including those with
 gender and skin t… https://t.co/MkGjXf9aXm",
 "user": {}, "entities": {}
}
```

Twitter API example



Twitter API
@TwitterAPI

To make room for more expression, we will now count all emojis as equal—including those with gender and skin tone modifiers 👍🏻👍🏽👍🏿. This is now reflected in Twitter-Text, our Open Source library.

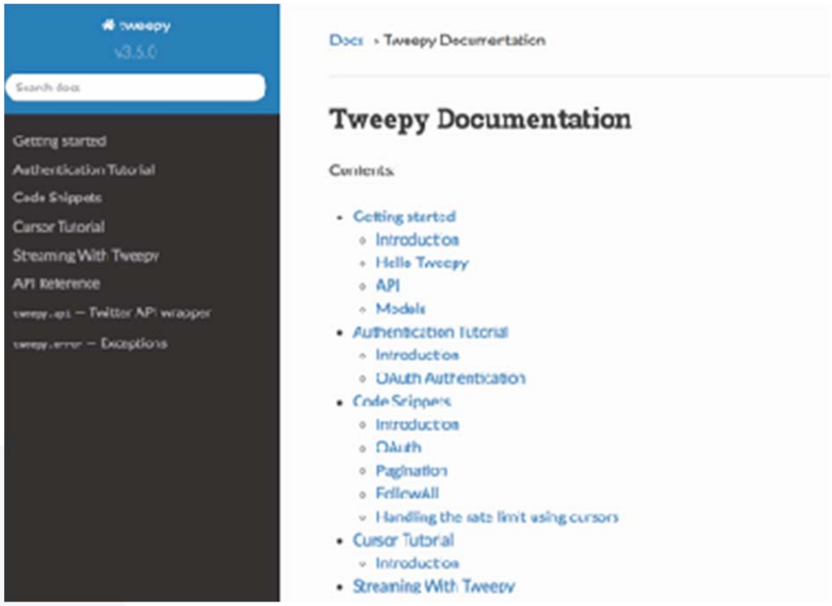
Using Twitter-Text? See the forum post for detail:

New update to the Twitter-Text library: Emoji character...
Over the years, we have made several updates to the way that people can communicate on Twitter. One of ...
twittercommunity.com

10:19 PM - Oct 10, 2018

```
{
 "created_at": "Wed Oct 10 20:19:24 +0000 2018",
 "id": 1050118621198921728,
 "id_str": "1050118621198921728",
 "text": "To make room for more expression, we will
 now count all emojis as equal—including those with
 gender and skin t… https://t.co/MkGjXf9aXm",
 "user": {}, "entities": {}
}
```

GET followers/ids



Tweepy Documentation

Contents:

- Getting started
  - Introduction
  - Hello Tweepy
  - API
  - Models
- Authentication Tutorial
  - Introduction
  - OAuth Authentication
- Code Snippets
  - Introduction
  - OAuth
  - Pagination
  - FollowAll
  - Handling the rate limit using cursors
- Cursor Tutorial
  - Introduction
- Streaming With Tweepy

APIs - Pros

- ToS compliant
- Easy to use, replicate
- Well-formatted data

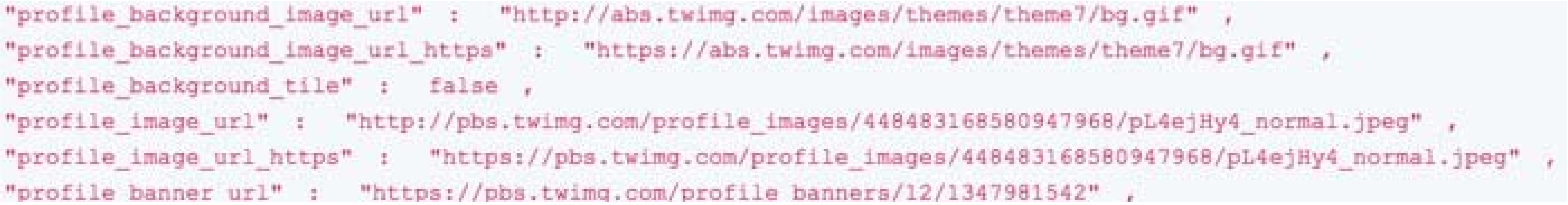
APIs - Cons (I)

- ToS compliant
- Easy to use, replicate
- Well-formatted data

- Authentication
  - IP based
  - User tokens based
- Rate limits
  - Restricted number of API call per user/IP address
  - x% of data

APIs - Cons (II)

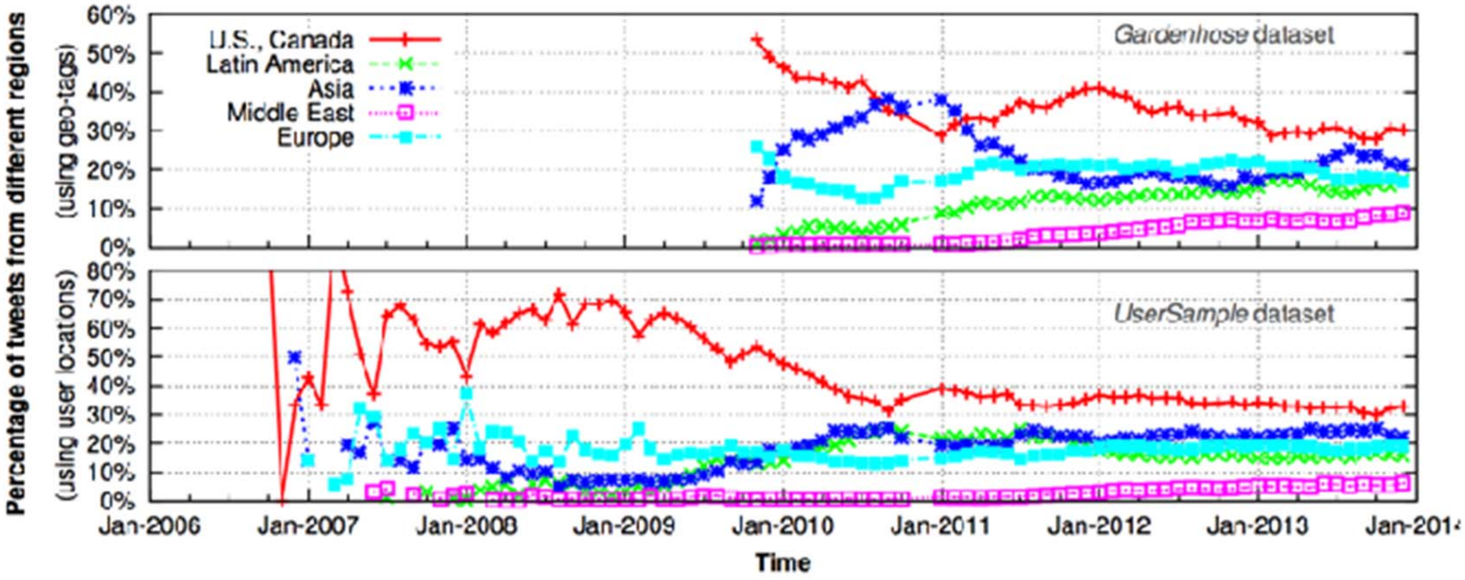
- Possible incompleteness of data
  - Missing info, such as images



```
"profile_background_image_url"  :   "http://abs.twimg.com/images/themes/theme7/bg.gif"  ,
"profile_background_image_url_https"  :   "https://abs.twimg.com/images/themes/theme7/bg.gif"  ,
"profile_background_tile"  :   false  ,
"profile_image_url"  :   "http://pbs.twimg.com/profile_images/448483168580947968/pL4ejHy4_normal.jpeg"  ,
"profile_image_url_https"  :   "https://pbs.twimg.com/profile_images/448483168580947968/pL4ejHy4_normal.jpeg"  ,
"profile_banner_url"  :   "https://pbs.twimg.com/profile_banners/12/1347981542"  ,
```

APIs - Cons (III)

- Possible unknown biases in data
  - Unclear what the platform provider may be giving



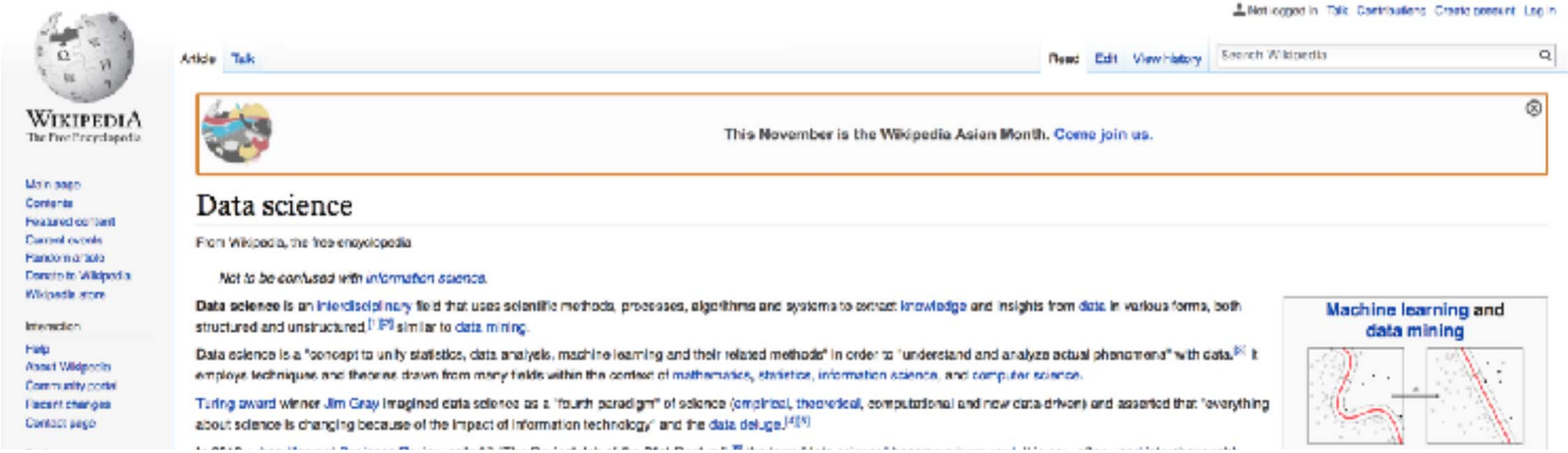
Morstatter et al, 2013, ICWSM

Web scraping

Web scraping

- Extracting data from source code of website

```
BASH: curl https://en.wikipedia.org/wiki/Data_science > wiki_ds.html
Python Requests:
requests.get(“https://en.wikipedia.org/wiki/Data_science”)
```



Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search Wikipedia

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

This November is the Wikipedia Asian Month. Come join us.

Data science

From Wikipedia, the free encyclopedia

Not to be confused with information science.

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured,[1][2] similar to data mining.

Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data.[3] It employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science, and computer science.

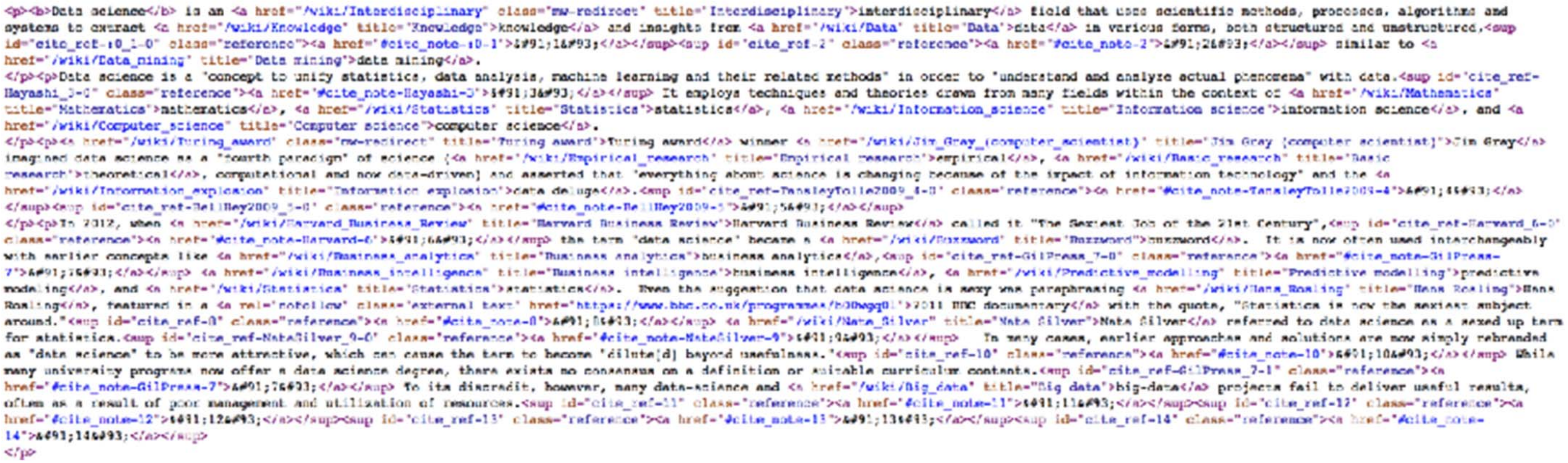
Turing award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge.[4][5]

Machine learning and data mining

Web scraping

- Extracting data from source code of website

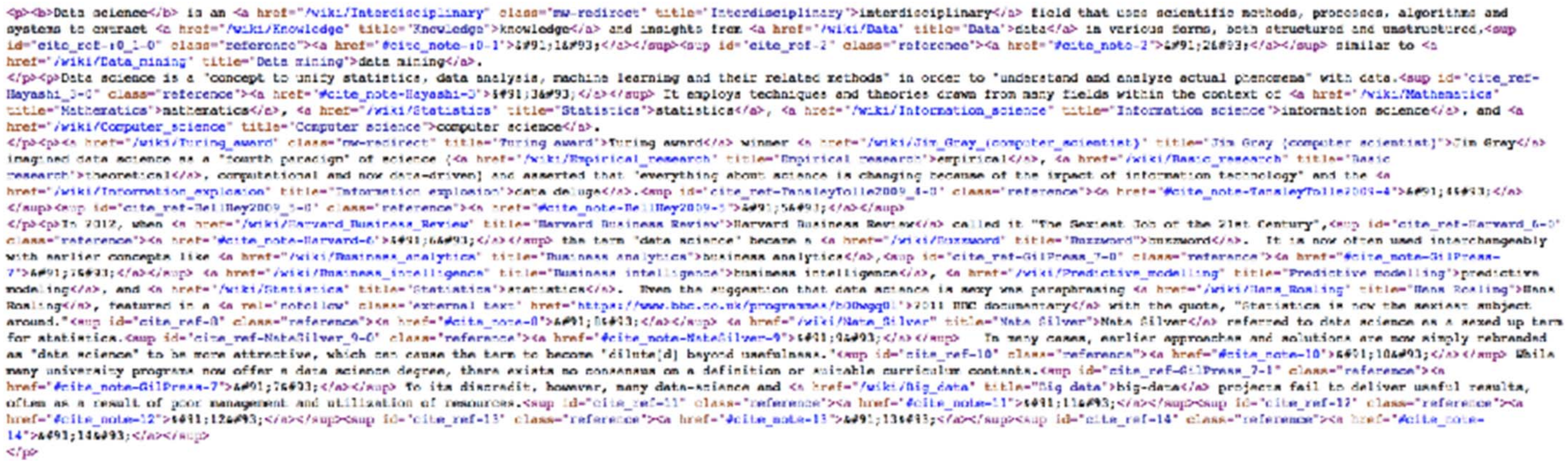
```
BASH: curl https://en.wikipedia.org/wiki/Data_science > wiki_ds.html
Python Requests:
requests.get(“https://en.wikipedia.org/wiki/Data_science”)
```



```
<p><b>Data science</b> is an <a href="/wiki/Interdisciplinary" class="mw-redirect" title="Interdisciplinary">interdisciplinary</a> field that uses scientific methods, processes, algorithms and
systems to extract <a href="/wiki/Knowledge" title="Knowledge">knowledge</a> and insights from <a href="/wiki/Data" title="Data">data</a> in various forms, both structured and unstructured,<sup
id="cite_ref-:0_1-0" class="reference"><a href="#cite_note-:0-1">&#91;1&#93;</a></sup><sup id="cite_ref-2" class="reference"><a href="#cite_note-2">&#91;2&#93;</a></sup> similar to <a
href="/wiki/Data_mining" title="Data mining">data mining</a>.
</p><p>Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data.<sup id="cite_ref-
Hayashi_3-0" class="reference"><a href="#cite_note-Hayashi-3">&#91;3&#93;</a></sup> It employs techniques and theories drawn from many fields within the context of <a href="/wiki/Mathematics"
title="Mathematics">mathematics</a>, <a href="/wiki/Statistics" title="Statistics">statistics</a>, <a href="/wiki/Information_science" title="Information science">information science</a>, and <a
href="/wiki/Computer_science" title="Computer science">computer science</a>.
</p><p><a href="/wiki/Turing_award" class="mw-redirect" title="Turing award">Turing award</a> winner <a href="/wiki/Jim_Gray_(computer_scientist)" title="Jim Gray (computer scientist)">Jim Gray</a>
imagined data science as a "fourth paradigm" of science (<a href="/wiki/Empirical_research" title="Empirical research">empirical</a>, <a href="/wiki/Basic_research" title="Basic
research">theoretical</a>, computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the <a
href="/wiki/Information_explosion" title="Information explosion">data deluge</a>.<sup id="cite_ref-TansleyTolle2009_4-0" class="reference"><a href="#cite_note-TansleyTolle2009-4">&#91;4&#93;</a>
</sup><sup id="cite_ref-BellHey2009_5-0" class="reference"><a href="#cite_note-BellHey2009-5">&#91;5&#93;</a></sup>
</p><p>In 2012, when <a href="/wiki/Harvard_Business_Review" title="Harvard Business Review">Harvard Business Review</a> called it "The Sexiest Job of the 21st Century",<sup id="cite_ref-Harvard_6-0"
class="reference"><a href="#cite_note-Harvard-6">&#91;6&#93;</a></sup> the term "data science" became a <a href="/wiki/Buzzword" title="Buzzword">buzzword</a>.  It is now often used interchangeably
with earlier concepts like <a href="/wiki/Business_analytics" title="Business analytics">business analytics</a>,<sup id="cite_ref-GilPress_7-0" class="reference"><a href="#cite_note-GilPress-
7">&#91;7&#93;</a></sup> <a href="/wiki/Business_intelligence" title="Business intelligence">business intelligence</a>, <a href="/wiki/Predictive_modelling" title="Predictive modelling">predictive
modeling</a>, and <a href="/wiki/Statistics" title="Statistics">statistics</a>.  Even the suggestion that data science is sexy was paraphrasing <a href="/wiki/Hans_Rosling" title="Hans Rosling">Hans
Rosling</a>, featured in a <a rel="nofollow" class="external text" href="https://www.bbc.co.uk/programmes/b00wgq0l">2011 BBC documentary</a> with the quote, "Statistics is now the sexiest subject
around."<sup id="cite_ref-8" class="reference"><a href="#cite_note-8">&#91;8&#93;</a></sup> <a href="/wiki/Nate_Silver" title="Nate Silver">Nate Silver</a> referred to data science as a sexed up term
for statistics.<sup id="cite_ref-NateSilver_9-0" class="reference"><a href="#cite_note-NateSilver-9">&#91;9&#93;</a></sup>  In many cases, earlier approaches and solutions are now simply rebranded
as "data science" to be more attractive, which can cause the term to become "dilute[d] beyond usefulness."<sup id="cite_ref-10" class="reference"><a href="#cite_note-10">&#91;10&#93;</a></sup> While
many university programs now offer a data science degree, there exists no consensus on a definition or suitable curriculum contents.<sup id="cite_ref-GilPress_7-1" class="reference"><a
href="#cite_note-GilPress-7">&#91;7&#93;</a></sup> To its discredit, however, many data-science and <a href="/wiki/Big_data" title="Big data">big-data</a> projects fail to deliver useful results,
often as a result of poor management and utilization of resources.<sup id="cite_ref-11" class="reference"><a href="#cite_note-11">&#91;11&#93;</a></sup><sup id="cite_ref-12" class="reference"><a
href="#cite_note-12">&#91;12&#93;</a></sup><sup id="cite_ref-13" class="reference"><a href="#cite_note-13">&#91;13&#93;</a></sup><sup id="cite_ref-14" class="reference"><a href="#cite_note-
14">&#91;14&#93;</a></sup>
</p>
```

Web scraping

- Tool to parse HTML code: beautifulsoup
- More in exercise hour!



```
<p><b>Data science</b> is an <a href="/wiki/Interdisciplinary" class="mw-redirect" title="Interdisciplinary">interdisciplinary</a> field that uses scientific methods, processes, algorithms and
systems to extract <a href="/wiki/Knowledge" title="Knowledge">knowledge</a> and insights from <a href="/wiki/Data" title="Data">data</a> in various forms, both structured and unstructured,<sup
id="cite_ref-:0_1-0" class="reference"><a href="#cite_note-:0-1">&#91;1&#93;</a></sup><sup id="cite_ref-2" class="reference"><a href="#cite_note-2">&#91;2&#93;</a></sup> similar to <a
href="/wiki/Data_mining" title="Data mining">data mining</a>.
</p><p>Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data.<sup id="cite_ref-
Hayashi_3-0" class="reference"><a href="#cite_note-Hayashi-3">&#91;3&#93;</a></sup> It employs techniques and theories drawn from many fields within the context of <a href="/wiki/Mathematics"
title="Mathematics">mathematics</a>, <a href="/wiki/Statistics" title="Statistics">statistics</a>, <a href="/wiki/Information_science" title="Information science">information science</a>, and <a
href="/wiki/Computer_science" title="Computer science">computer science</a>.
</p><p><a href="/wiki/Turing_award" class="mw-redirect" title="Turing award">Turing award</a> winner <a href="/wiki/Jim_Gray_(computer_scientist)" title="Jim Gray (computer scientist)">Jim Gray</a>
imagined data science as a "fourth paradigm" of science (<a href="/wiki/Empirical_research" title="Empirical research">empirical</a>, <a href="/wiki/Basic_research" title="Basic
research">theoretical</a>, computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the <a
href="/wiki/Information_explosion" title="Information explosion">data deluge</a>.<sup id="cite_ref-TansleyTolle2009_4-0" class="reference"><a href="#cite_note-TansleyTolle2009-4">&#91;4&#93;</a>
</sup><sup id="cite_ref-BellHey2009_5-0" class="reference"><a href="#cite_note-BellHey2009-5">&#91;5&#93;</a></sup>
</p><p>In 2012, when <a href="/wiki/Harvard_Business_Review" title="Harvard Business Review">Harvard Business Review</a> called it "The Sexiest Job of the 21st Century",<sup id="cite_ref-Harvard_6-0"
class="reference"><a href="#cite_note-Harvard-6">&#91;6&#93;</a></sup> the term "data science" became a <a href="/wiki/Buzzword" title="Buzzword">buzzword</a>.  It is now often used interchangeably
with earlier concepts like <a href="/wiki/Business_analytics" title="Business analytics">business analytics</a>,<sup id="cite_ref-GilPress_7-0" class="reference"><a href="#cite_note-GilPress-
7">&#91;7&#93;</a></sup> <a href="/wiki/Business_intelligence" title="Business intelligence">business intelligence</a>, <a href="/wiki/Predictive_modelling" title="Predictive modelling">predictive
modeling</a>, and <a href="/wiki/Statistics" title="Statistics">statistics</a>.  Even the suggestion that data science is sexy was paraphrasing <a href="/wiki/Hans_Rosling" title="Hans Rosling">Hans
Rosling</a>, featured in a <a rel="nofollow" class="external text" href="https://www.bbc.co.uk/programmes/b00wgq0l">2011 BBC documentary</a> with the quote, "Statistics is now the sexiest subject
around."<sup id="cite_ref-8" class="reference"><a href="#cite_note-8">&#91;8&#93;</a></sup> <a href="/wiki/Nate_Silver" title="Nate Silver">Nate Silver</a> referred to data science as a sexed up term
for statistics.<sup id="cite_ref-NateSilver_9-0" class="reference"><a href="#cite_note-NateSilver-9">&#91;9&#93;</a></sup>   In many cases, earlier approaches and solutions are now simply rebranded
as "data science" to be more attractive, which can cause the term to become "dilute[d] beyond usefulness."<sup id="cite_ref-10" class="reference"><a href="#cite_note-10">&#91;10&#93;</a></sup> While
many university programs now offer a data science degree, there exists no consensus on a definition or suitable curriculum contents.<sup id="cite_ref-GilPress_7-1" class="reference"><a
href="#cite_note-GilPress-7">&#91;7&#93;</a></sup> To its discredit, however, many data-science and <a href="/wiki/Big_data" title="Big data">big-data</a> projects fail to deliver useful results,
often as a result of poor management and utilization of resources.<sup id="cite_ref-11" class="reference"><a href="#cite_note-11">&#91;11&#93;</a></sup><sup id="cite_ref-12" class="reference"><a
href="#cite_note-12">&#91;12&#93;</a></sup><sup id="cite_ref-13" class="reference"><a href="#cite_note-13">&#91;13&#93;</a></sup><sup id="cite_ref-14" class="reference"><a href="#cite_note-
14">&#91;14&#93;</a></sup>
</p>
```

Web scraping - Pros & cons

- Easy to set up
- Easy to parallelize

Web scraping - Pros & cons

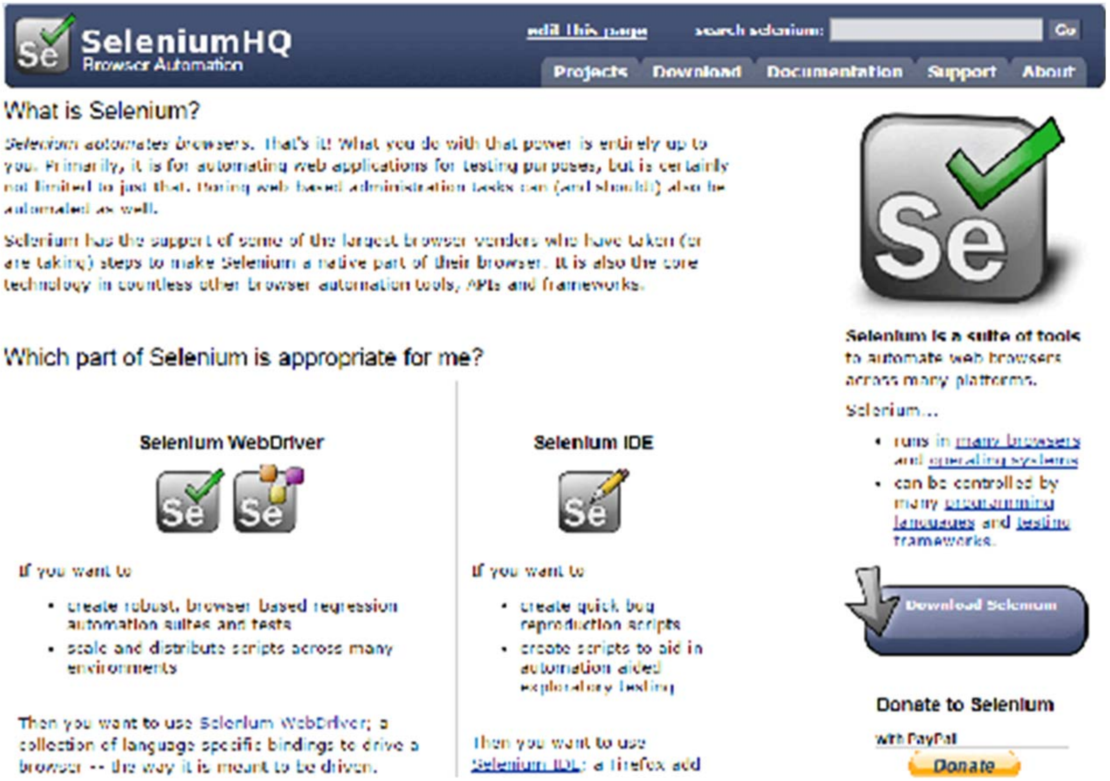
- Easy to set up
- Easy to parallelize

- Not ToS compliant
- No ajax, no javascript, no images
- Parsing needs to be updated every time the platform makes a change

Browser automation

Browser automation

- Automate browser to scrape dynamically rendered webpages
- Could be used to:
  - fill forms
  - enter text
  - scroll
  - click on buttons



What is Selenium?

Selenium automates browsers. That's it! What you do with that power is entirely up to you. Primarily, it is for automating web applications for testing purposes, but is certainly not limited to just that. Boring web-based administration tasks can (and should!) also be automated as well.

Selenium has the support of some of the largest browser vendors who have taken (or are taking) steps to make Selenium a native part of their browser. It is also the core technology in countless other browser automation tools, APIs and frameworks.

Which part of Selenium is appropriate for me?

Selenium WebDriver

If you want to

- create robust, browser-based regression automation suites and tests
- scale and distribute scripts across many environments

Then you want to use Selenium WebDriver; a collection of language specific bindings to drive a browser -- the way it is meant to be driven.

Selenium IDE

If you want to

- create quick bug reproduction scripts
- create scripts to aid in automation-aided exploratory testing

Then you want to use Selenium IDE; a Firefox add

Selenium is a suite of tools to automate web browsers across many platforms.

Selenium...

- runs in many browsers and operating systems
- can be controlled by many programming languages and testing frameworks.

Donate to Selenium

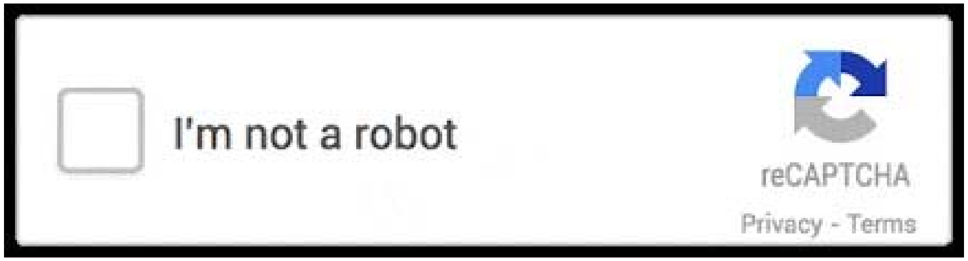
with PayPal

Browser automation - Pros

- Mimics human browsing
- Loads ajax, images, etc.
- Pops open a browser so you can check if it works - easy to debug
- Can design the flow of events : log-in, search, etc.

Browser automation - Cons

- Not ToS compliant
- Need to parse content
- Difficult to scale
- Unpredictable bugs
- Platform may throw a pop up (e.g. a captcha) if you collect too much



Headless browser

- Tool - PhantomJS
- Headless => does not pop open a browser window

Headless browser - Pros

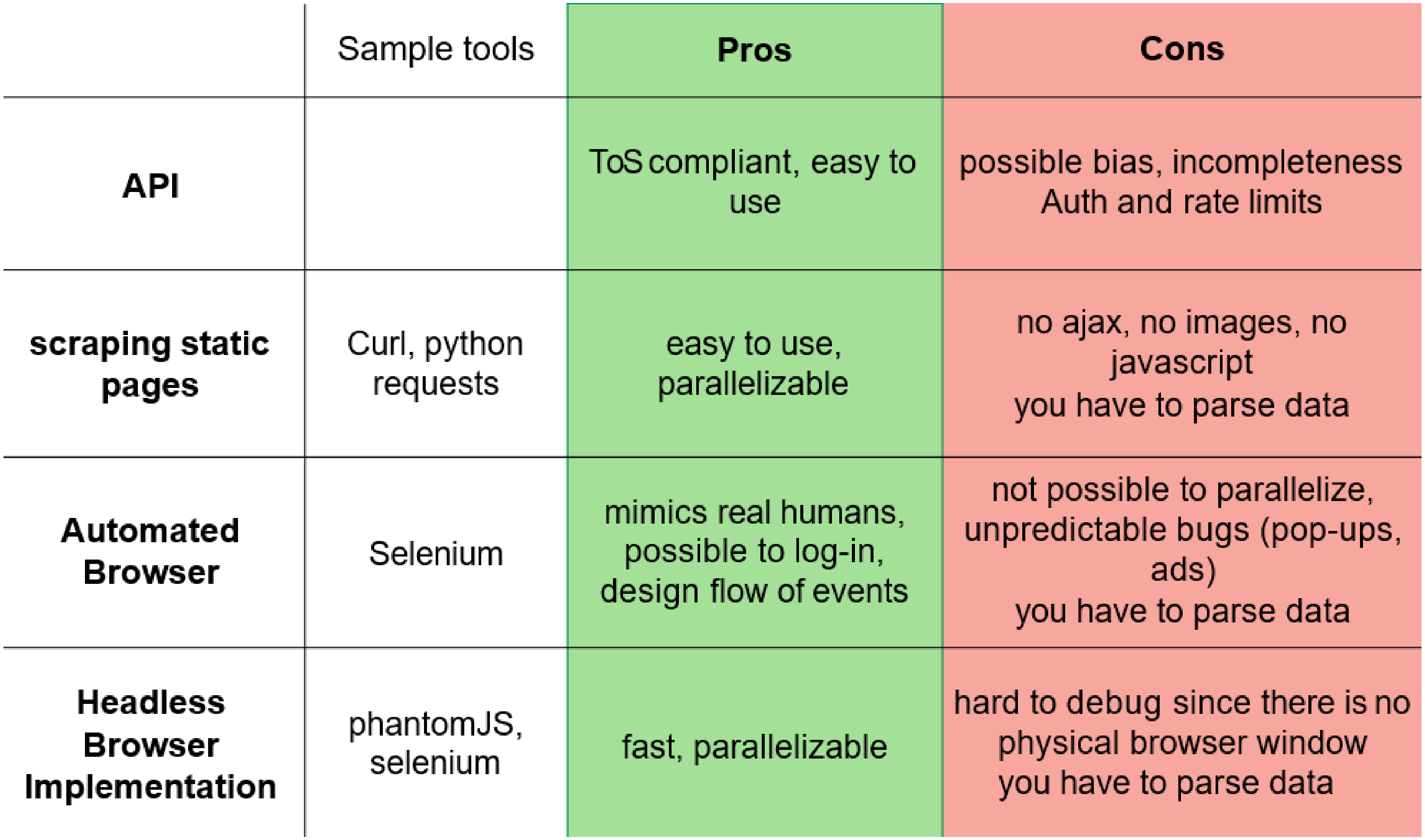
- Mimics human browsing
- Loads ajax, images etc
- Can design the flow of events: log-in, search etc.
- Easy to parallelize
- Less memory intensive

Headless browser - Cons

- Not ToS compliant
- Need to parse content
- Unpredictable bugs
- Messy code
- Hard to debug - no browser window

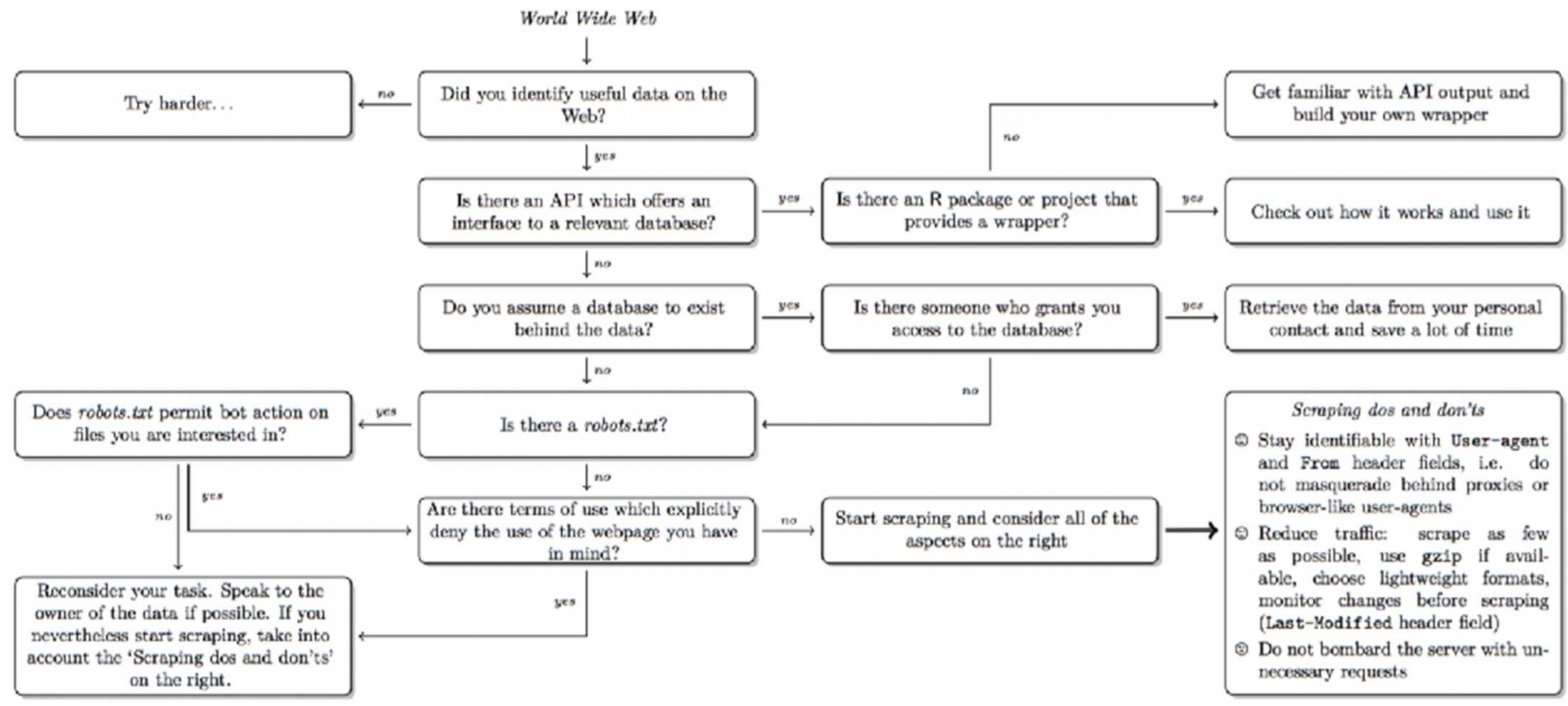
Other options for collecting data?

Summary



	Sample tools	Pros	Cons
API		ToS compliant, easy to use	possible bias, incompleteness Auth and rate limits
scraping static pages	Curl, python requests	easy to use, parallelizable	no ajax, no images, no javascript you have to parse data
Automated Browser	Selenium	mimics real humans, possible to log-in, design flow of events	not possible to parallelize, unpredictable bugs (pop-ups, ads) you have to parse data
Headless Browser Implementation	phantomJS, selenium	fast, parallelizable	hard to debug since there is no physical browser window you have to parse data

How to decide?



World Wide Web

Did you identify useful data on the Web?

Try harder…

Is there an API which offers an interface to a relevant database?

Is there an R package or project that provides a wrapper?

Get familiar with API output and build your own wrapper

Check out how it works and use it

Do you assume a database to exist behind the data?

Is there someone who grants you access to the database?

Retrieve the data from your personal contact and save a lot of time

Is there a *robots.txt*?

Does *robots.txt* permit bot action on files you are interested in?

Are there terms of use which explicitly deny the use of the webpage you have in mind?

Start scraping and consider all of the aspects on the right

Reconsider your task. Speak to the owner of the data if possible. If you nevertheless start scraping, take into account the 'Scraping dos and don'ts' on the right.

Scraping dos and don'ts

- ☺ Stay identifiable with `User-agent` and `From` header fields, i.e. do not masquerade behind proxies or browser-like user-agents
- ☺ Reduce traffic: scrape as few as possible, use `gzip` if available, choose lightweight formats, monitor changes before scraping (`Last-Modified` header field)
- ☹ Do not bombard the server with unnecessary requests

Image source: http://pablobarbera.com

Case study: Collecting Personalized data

Collecting personalized data (I)

- Volunteered data - crowdsourced
- Browser plugin

Collecting personalized data (I)

- Volunteered data - crowdsourced
- Browser plugin

- Easier to scale
- Leverage real accounts
- Does not break ToS

- Must recruit volunteers
- Sampling bias
- Less control

Collecting personalized data (II)

- Controlled experiments
  - Create test accounts with preferred characteristics
  - Collect the personalized recs or search results for these accounts.

Collecting personalized data (II)

- Controlled experiments
  - Create test accounts with preferred characteristics
  - Collect the personalized recs or search results for these accounts.

- Easier to measure impact of features
- Clean data
- Can mitigate biases

- Synthetic data
- Breaks ToS
- Harder to scale

Can I anonymize the data?

Am I violating Terms of Service?

Sharing data: Ethical & legal considerations

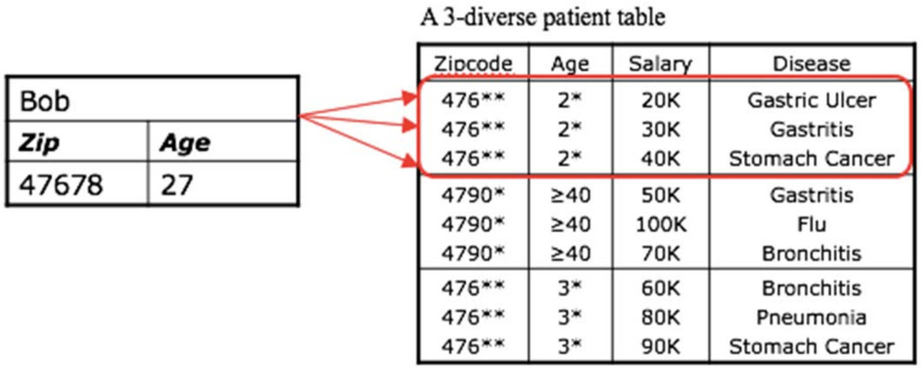
Sharing data publicly

- Anonymize the users
- Even then, users can be fingerprinted if:
  - Sample size is small
  - there are outliers or minorities
  - it can be merged with other available data sets
  - etc.

Sharing data publicly

- Anonymize the users
- Even then, users can be fingerprinted if:
  - Sample size is small
  - there are outliers or minorities
  - it can be merged with other available data sets
  - etc.
- K-anonymization: Data is said to have k-anonymity if the information for each person contained in the dataset cannot be distinguished from at least k-1 individuals whose information also appears in the release

Sharing data publicly



A 3-diverse patient table

Bob	
Zip	***Age***
47678	27

Zipcode	Age	Salary	Disease
476**	2*	20K	Gastric Ulcer
476**	2*	30K	Gastritis
476**	2*	40K	Stomach Cancer
4790*	≥40	50K	Gastritis
4790*	≥40	100K	Flu
4790*	≥40	70K	Bronchitis
476**	3*	60K	Bronchitis
476**	3*	80K	Pneumonia
476**	3*	90K	Stomach Cancer

Source: https://elf11.github.io/2017/04/22/kanonymity.html

- K-anonymization: Data is said to have k-anonymity if the information for each person contained in the dataset cannot be distinguished from at least k-1 individuals whose information also appears in the release

Sharing data publicly

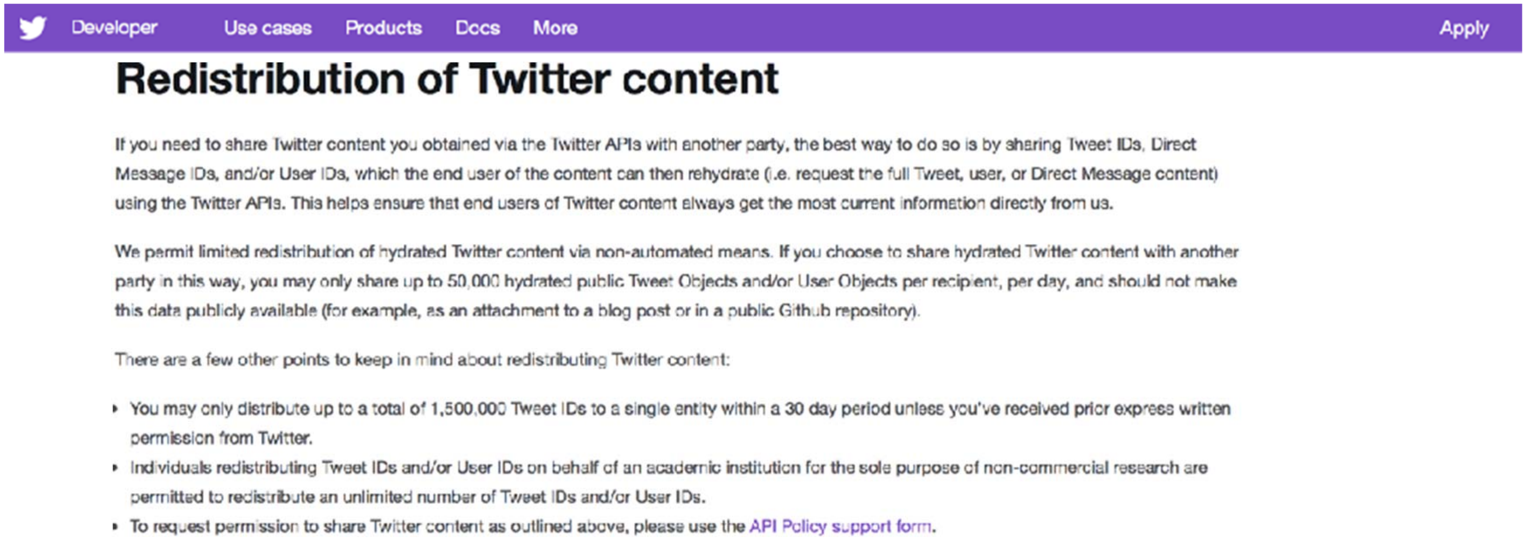
General Data Protection Regulation

- Anonymize the users
- Securely store, control access & transfer your data

General Data Protection Regulation

Sharing data publicly

- Anonymize the users
- Securely store, control access & transfer your data
- Read ToS carefully



Redistribution of Twitter content

If you need to share Twitter content you obtained via the Twitter APIs with another party, the best way to do so is by sharing Tweet IDs, Direct Message IDs, and/or User IDs, which the end user of the content can then rehydrate (i.e. request the full Tweet, user, or Direct Message content) using the Twitter APIs. This helps ensure that end users of Twitter content always get the most current information directly from us.

We permit limited redistribution of hydrated Twitter content via non-automated means. If you choose to share hydrated Twitter content with another party in this way, you may only share up to 50,000 hydrated public Tweet Objects and/or User Objects per recipient, per day, and should not make this data publicly available (for example, as an attachment to a blog post or in a public Github repository).

There are a few other points to keep in mind about redistributing Twitter content:

- You may only distribute up to a total of 1,500,000 Tweet IDs to a single entity within a 30 day period unless you've received prior express written permission from Twitter.
- Individuals redistributing Tweet IDs and/or User IDs on behalf of an academic institution for the sole purpose of non-commercial research are permitted to redistribute an unlimited number of Tweet IDs and/or User IDs.
- To request permission to share Twitter content as outlined above, please use the API Policy support form.

Sharing data publicly

- Anonymize
  - protect against fingerprinting - k-anonymity
- Securely store, control access & transfer your data
- Read TOS carefully
  - allowed to share tweet ids, and a sample of tweet objects for non commercial use
- Do not share copyrighted content
  - Just because you can download it, doesn't mean you can share someone else's intellectual property

In this class… we learnt

Collecting data:
Ethical & legal
considerations

Overview of data
collection tools:
APIs
Web scraping
Browser automation

Sharing data:
Ethical & legal
considerations

Image source: https://www.flaticon.com/

How would you collect data for…

- Which method would you use?
  - – Do you need more information to decide?
- What are the method’s pros and cons for this scenario?
- What are the ethical and legal issues of this method?

How would you collect data for…

- What politicians post on Twitter

How would you collect data for…

- Wikipedia data about scientists from different countries

How would you collect data for…

- News articles from nytimes.com

How would you collect data for…

- Multiple pages of search results on Twitter/ Google

How would you collect data for…

- YouTube recommendations

End of Lecture 1

Questions:

Tutorials 1. TA Office hours (book slot)