



# COLLECTING WEB DATA FROM WEBSITES AND APIS FOR ACADEMIC RESEARCH

**HANNES DATTA**

*BASED ON JOINT WORK WITH JOHANNES BOEGERSHAUSEN, ABHISHEK BORAH &  
ANDREW STEPHEN*

WU VIENNA  
1 APRIL 2022



## ▶ Introductory disclaimer

- By any means, we are really not the first (marketing) scholars to gather web data via scraping, APIs, etc.,
  - but we have used this in our own work + reviewed such research (extensively)
  - we currently have a methodological paper about this in the review process  
(Boegershausen, Datta, Borah, and Stephen 2022; <https://tiu.nu/scraping>)

## ▶ Introductory disclaimer

- By any means, we are really not the first (marketing) scholars to gather web data via scraping, APIs, etc.,
  - but we have used this in our own work + reviewed such research (extensively)
  - we currently have a methodological paper about this in the review process  
(Boegershausen, Datta, Borah, and Stephen 2022; <https://tiu.nu/scraping>)
- **There is no boilerplate template for gathering web data for marketing research.**

## ▶ Introductory disclaimer

- By any means, we are really not the first (marketing) scholars to gather web data via scraping, APIs, etc.,
  - but we have used this in our own work + reviewed such research (extensively)
  - we currently have a methodological paper about this in the review process  
(Boegershausen, Datta, Borah, and Stephen 2021; <https://tiu.nu/scraping>)
- There is no boilerplate template for gathering web data for marketing research.
- **This workshop won't be sufficient to teach you how to scrape or use APIs, given time constraints.** Consider it as a starter. And follow <https://odcm.hannedatta.com> if I've sparked your interest.

## ▶ Introductory disclaimer

- By any means, we are really not the first (marketing) scholars to gather web data via scraping, APIs, etc.,
  - but we have used this in our own work + reviewed such research (extensively)
  - we currently have a methodological paper about this in the review process  
(Boegershausen, Datta, Borah, and Stephen 2021; <https://tiu.nu/scraping>)
- There is no boilerplate template for gathering web data for consumer research.
- This workshop won't be sufficient to teach you how to scrape or use APIs, given time constraints. Consider it as a starter. And follow <https://odcm.hannedatta.com> if I've sparked your interest.
- When you feel that I am going to fast, please slow me down.
- This is **designed to be an interactive session**, so we might not get through all materials, but I will share extended slides and supporting docs (see also [hannedatta.com](https://hannedatta.com) + <https://tiu.nu/scraping>)

## ▶ Inventorizing your needs

- 2x2
    - CB vs. quant
    - No experience – a lot of experience/used in papers
  - Done the tutorial?
  - For those that use web data...
    - Experience w/ web scraping vs. APIs
  - Technical vs. conceptual requirement
    - Want to learn scraping now? (tutorial...), vs. develop paper based on method. framework? (source selection, design)
  - Questions on current research projects?
- + Name tags...?



# ▶ Agenda

- Motivation & what's in it for YOU
- Web scraping & APIs for dummies
- Methodological framework & design decisions
  - Data Source Selection
  - Extraction Design
- Future Research Opportunities
- Your (remaining) questions

Thinking more deeply about web scraping & APIs...

# **MOTIVATION & WHAT'S IN IT FOR YOU**



## ▶ The Internet is ubiquitous

**7:11**  
hours

time spent online per day by the average American consumer

**85%**

proportion of US consumers that use the Internet every single day

### Number of active users in November 2021 (global)



>2.9b



2.3b



Instagram

1.4b



TikTok

1b



463m

## ▶ Generation of massive digital traces



~ **224m** reviews



~ **988m** reviews & opinions

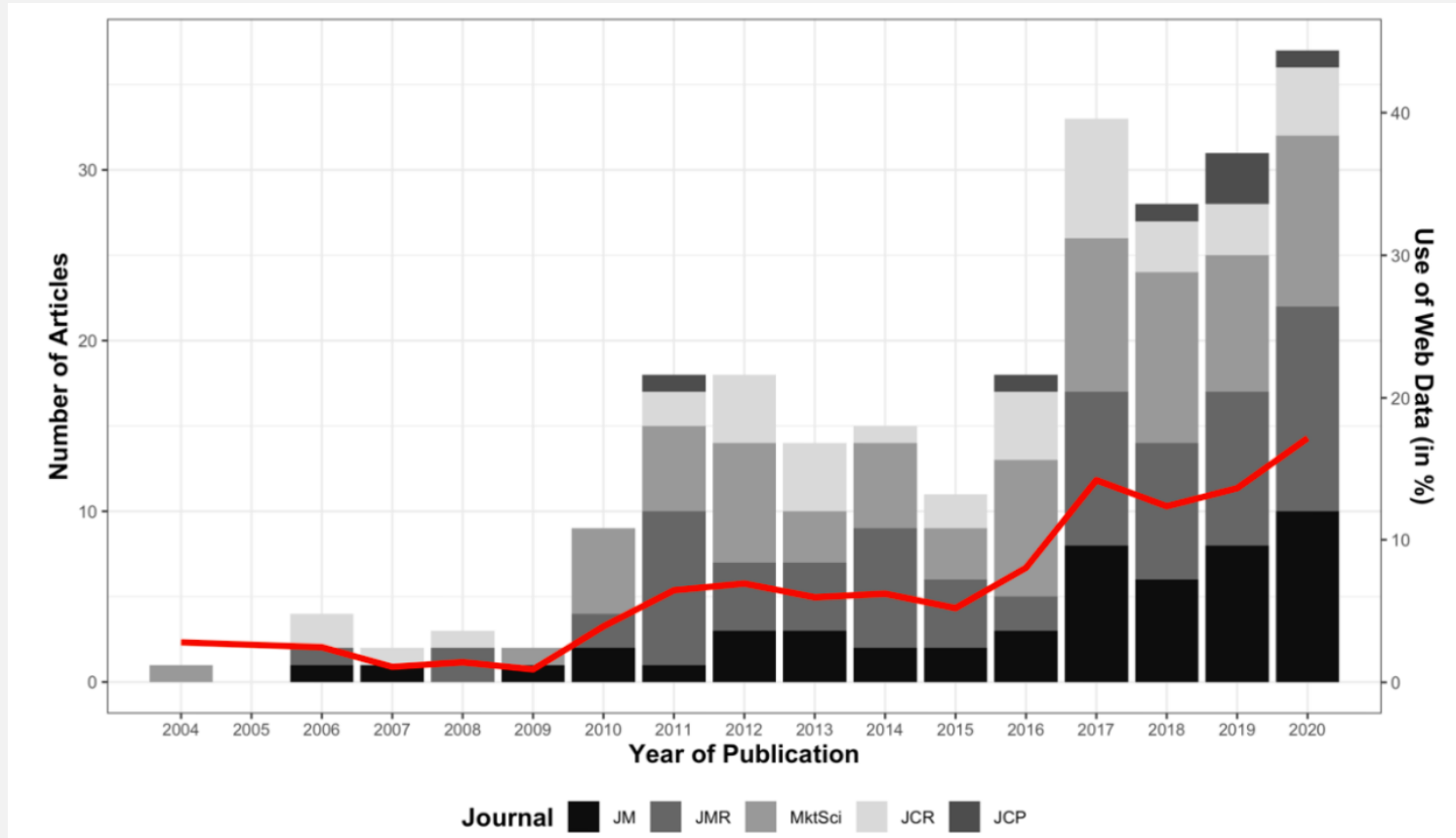


**500m/day** \* <https://www.internetlivestats.com/one-second/#tweets-band>



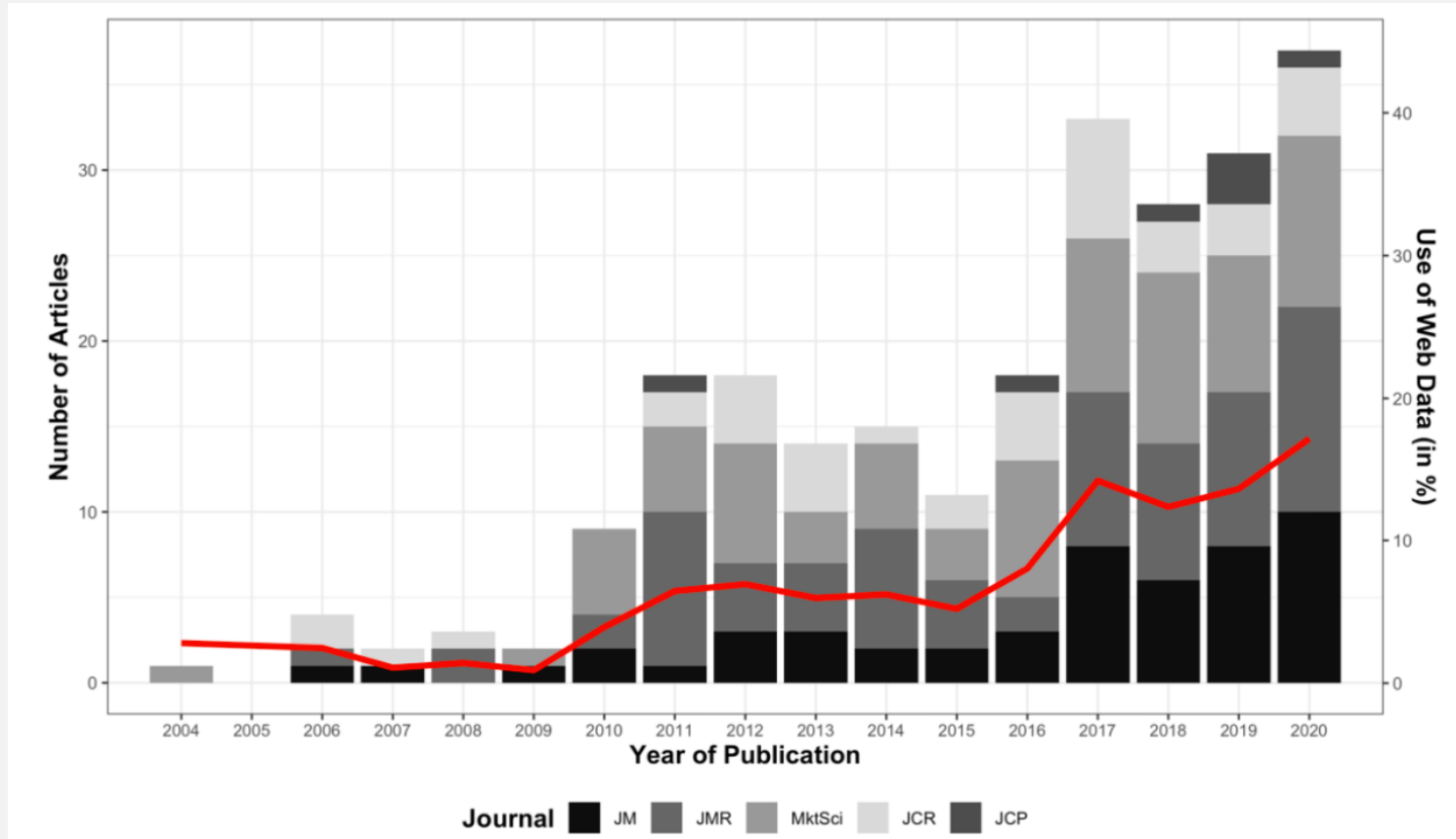
**544K** projects

## ▶ Increasing usage of web data in marketing research



Source: Boegershausen, Datta, Borah, and Stephen (2022)

## ▶ Increasing usage of web data in marketing research

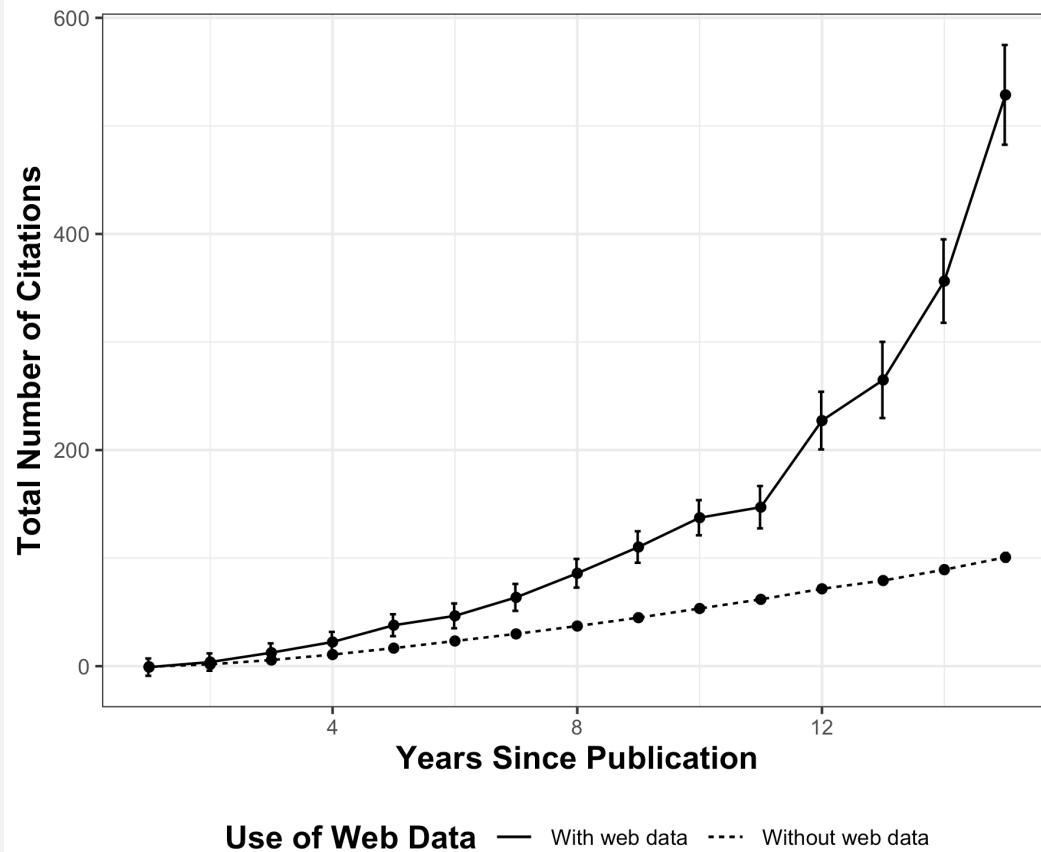


2021 thus far:

**78**

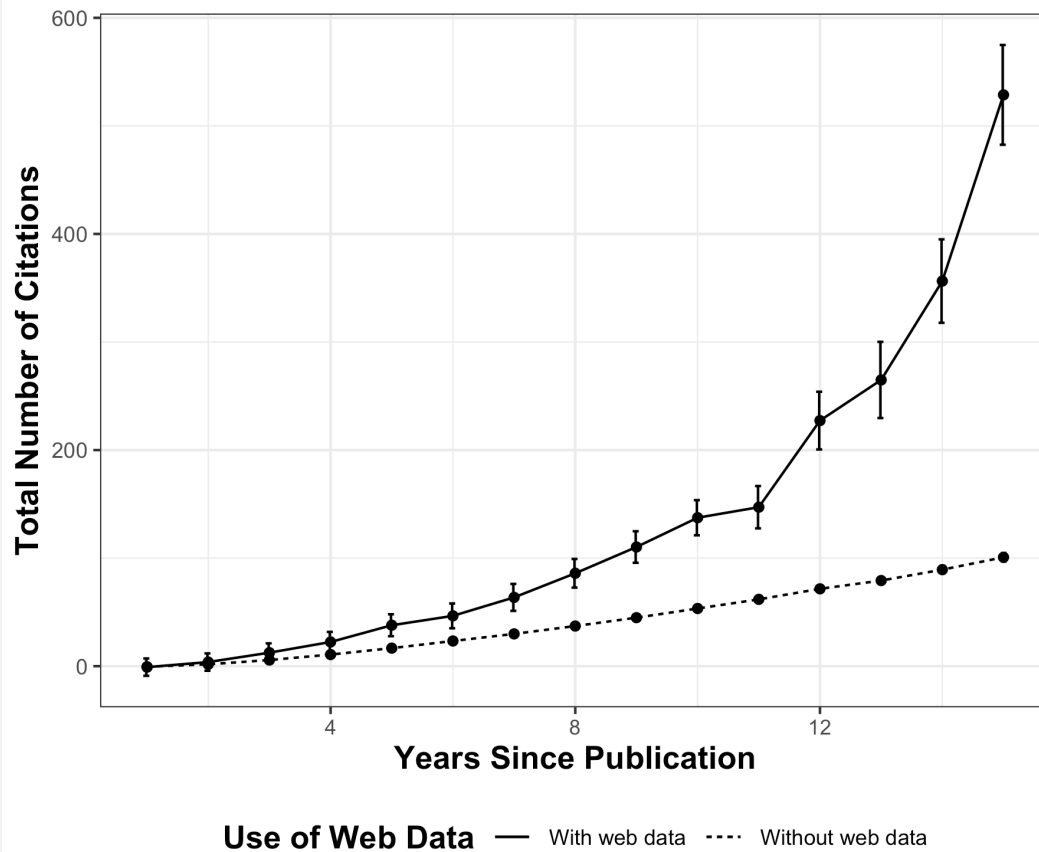
accepted/published  
papers that use web data

## ► Impactful research...



Notes: Regression yearly cites on journal-issue fixed effects, and dummy variables indicating the years since publication – for about 244 webdata-based papers and 4k non-web-data-based papers in the five top marketing journals. See <https://github.com/hannesdatta/webdata-in-marketing>.

## ► Impactful research...



- that is somewhat concentrated (based on author-provided keywords):
  - #1 WORD OF MOUTH 14%
  - #2 SOCIAL MEDIA 14%
  - #3 USER-GENERATED CONTENT 10%
  - #4 ADVERTISING 8%
  - #5 ONLINE REVIEWS 7%
- Uses mostly websites, NOT APIs (**85%**)
- Uses mostly single web sources (**61%**)

Notes: Regression yearly cites on journal-issue fixed effects, and dummy variables indicating the years since publication – for about 244 webdata-based papers and 4k non-web-data-based papers in the five top marketing journals. See <https://github.com/hannesdatta/webdata-in-marketing>.

▶ ... but web data is also SO MUCH MORE!



Google Trends

Google Cloud  
Vision API

Spotify for Developers

amazon  
Product Advertising API

wu  
WEATHER UNDERGROUND

## ▶ **What are you interested in...**

+ search on Google for “your interest + API”

**What pops up?**



▶ ... but web data is also SO MUCH MORE!



Google Trends

Google Cloud  
Vision API

Spotify for Developers

amazon  
Product Advertising API

wu  
WEATHER UNDERGROUND

ENDPOINTS

- Albums >
- Artists >
- Shows >
- Episodes >
- Tracks ▾
  - Get Track GET
  - Get Several Tracks GET
  - Get User's Saved Tracks GET
  - Save Tracks for Current User PUT
  - Remove Tracks for Current User DELETE
  - Check User's Saved Tracks GET
  - Get Tracks' Audio Features GET
  - Get Track's Audio Features GET
  - Get Track's Audio Analysis GET
  - Get Recommendations GET
- Search >
- Users >
- Playlists >
- Categories >
- Genres >
- Player >

# Get Track's Audio Features

OAuth 2.0

Get audio feature information for a single track identified by its unique Spotify ID.

## Request

GET /audio-features/{id}

**id** string required  
 The Spotify ID for the track.  
 Example value: "1dFghVXANMIKmjXsNCbNI"

## Responses 200 401 403 429

Audio features for one track

Body application/json

**acousticness** number<float>  
 A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.  
 >= 0 <= 1

**analysis\_url** string

Request Sample: Shell / cURL ▾

```
curl --request GET \
  --url https://api.spotify.com/v1/audio-features/id \
  --header 'Authorization: ' \
  --header 'Content-Type: application/json'
```

### Response Example

```
1 {
2   "acousticness": 0.00242,
3   "analysis_url": "https://api.spotify.com/v1/audio-analysis.
4   "danceability": 0.585,
5   "duration_ms": 237040,
6   "energy": 0.842,
7   "id": "2takw0aAZwiXQijPHix7B",
8   "instrumentalness": 0.00686,
9   "key": 9,
10  "liveness": 0.0866,
11  "loudness": -5.883,
12  "mode": 0,
13  "speechiness": 0.0556,
14  "tempo": 118.211,
15  "time_signature": 4,
16  "track_href": "https://api.spotify.com/v1/tracks/2takw0aA
17  "type": "audio_features",
18  "uri": "spotify:track:2takw0aAZwiXQijPHix7B",
19  "valence": 0.428
20 }
```

## ► Websites vs. APIs

**Table W1: Commonalities and Differences Between Web Scraping and APIs<sup>1</sup>**

	<b>Web scraping</b>	<b>Application Programming Interfaces (APIs)</b>
Use cases	Extract data from websites	Extract data from APIs
Scope	Any content from publicly available websites	Any content or algorithm made available by the API provider
Data extraction	Programmatically browse on the website and capture information as it becomes visible or available in the website's HTML (HyperText Markup Language) source code	Capture information directly from programming interfaces designed for content extraction and enrichment at scale, typically in JSON (JavaScript Object Notation) or XML (Extensible Markup Language) format
Costs	Free	Usually on a subscription (e.g., free, freemium, paid)
Scalability	Moderate	High
Data documentation	Mostly undocumented	Mostly documented for app developers
Legal and ethical risks	Low-high	Low-moderate <sup>2</sup>
Illustrative data sources	E-commerce ( <a href="https://amazon.com">https://amazon.com</a> ), Online review ( <a href="https://yelp.com">https://yelp.com</a> )	Discussion forum (Reddit API), Social media (Twitter API)
Illustrative articles	Chevalier and Mayzlin (2006); Ludwig et al. (2013)	Tellis et al. (2019); Toubia and Stephen (2013)

<sup>1</sup> Table compares data extraction using web scraping with data extracting using APIs.

<sup>2</sup> Before 2010, companies focused on the “social” aspects of their business model and actively offered APIs (usually in XML or JSON feeds) to grow their ecosystem. As companies matured and grew their userbase, many of these firms became increasingly restrictive to monetize their data (e.g., Amazon, Twitter).

▶ **Many opportunities:** *discovery orientation*



**“scout out”  
novel consumer  
phenomena**

streaming (Datta et al. 2018)  
mobile devices (Melumad et al. 2019)



**different levels  
of analysis +  
effects over time**

brand public (Arvidsson & Caliandro 2016)  
psychological distances (Huang et al. 2016)



**explore  
geographic  
variation**

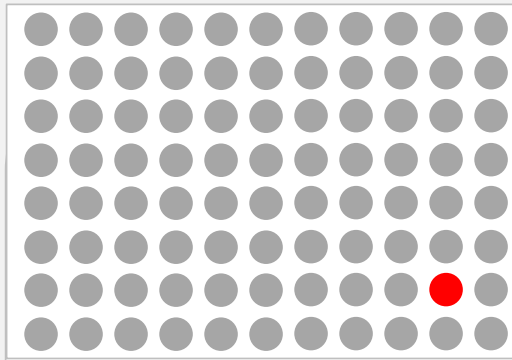
Sensitivity to prices and ratings  
across the globe (Kübler et al. 2018)

▶ **Many opportunities:** *phenomena*



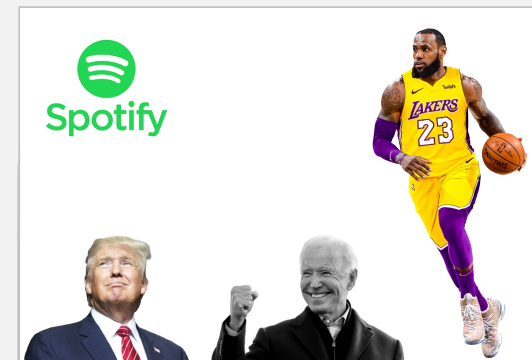
**socially sensitive  
phenomena**

controversy (Chen & Berger 2013)  
violent protests (Mooijman et al. 2018)



**rare events**

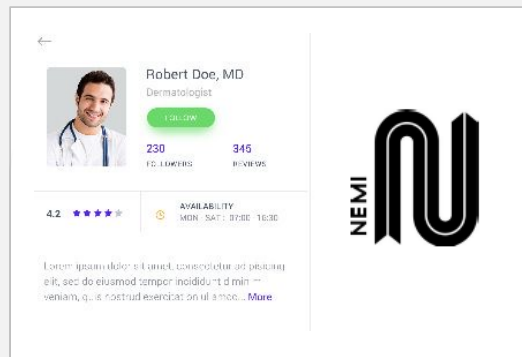
Bright (2017)



**hard-to-reach  
populations**

political elites (Brady et al. 2019)  
professional athletes (Grijalva et al. 2020)  
early Spotify adopters (Datta et al. 2018)

► **Many opportunities:** *creative applications*



**stimuli  
generation**

provider profiles (Howe & Monin 2017)  
brand logos (Luffarelli et al. 2019)



**data  
enrichment**

Govind et al. (2020)

## ► Highly versatile data collection technique

**Table 1: A taxonomy of the pathways to knowledge creation via web data**

Effect on (with typical outcome variables in parentheses)	Primary knowledge pathway resulting from the use of web data			
	Pathway 1: Studying new phenomena	Pathway 2: Boosting ecological value	Pathway 3: Facilitating method- logical advancement	Pathway 4: Enhancing Inferences
<b>Consumers</b> (e.g., learning and social media use of consumers; sentiments of customers; interaction in a network)	Toubia and Stephen (2013): testing the motivations of users to contribute content to social media.	Sridhar and Srinivasan (2012): understanding peer effects in evaluating online product reviews.	Huang (2019): understanding how picture quality improves due to consumer learning.	Huang et al. (2016): exploiting within-user variation to measure how psychological distances interact.
<b>Organizations</b> (e.g., sales and profits of firms, donations to non- profits)	Chevalier and Mayzlin (2006): demonstrating the impact of online reviews on book sales.	Wu and Cosguner (2020): demonstrating the prevalence and profit implications of decoy effects.	Netzer et al. (2012): generating marketing insights from user- generated content to predict market share.	Datta et al. (2021): gathering national holidays across 14 countries to capture seasonality.
<b>Other marketing stakeholders</b> (e.g., market reaction of investors, public health outcomes)	Hemosilla et al. (2018): examining how consumers' aesthetic preferences create biases in firms' hiring decisions.	Blaseg et al. (2020): examining whether consumers are protected against false price advertising claims on Kickstarter.	Tirunillai and Tellis (2012): developing novel online metrics based on user- generated content to predict stock returns.	Kim and KC (2020): exploring the effect of ads for erectile dysfunction drugs on birth rates.

*Notes:* The table cross-tabulates the pathways through which web data has advanced marketing thought (the columns), with three of the most studied actors in marketing research (the rows).

Source: Boegershausen, Datta, Borah, and Stephen (2022)





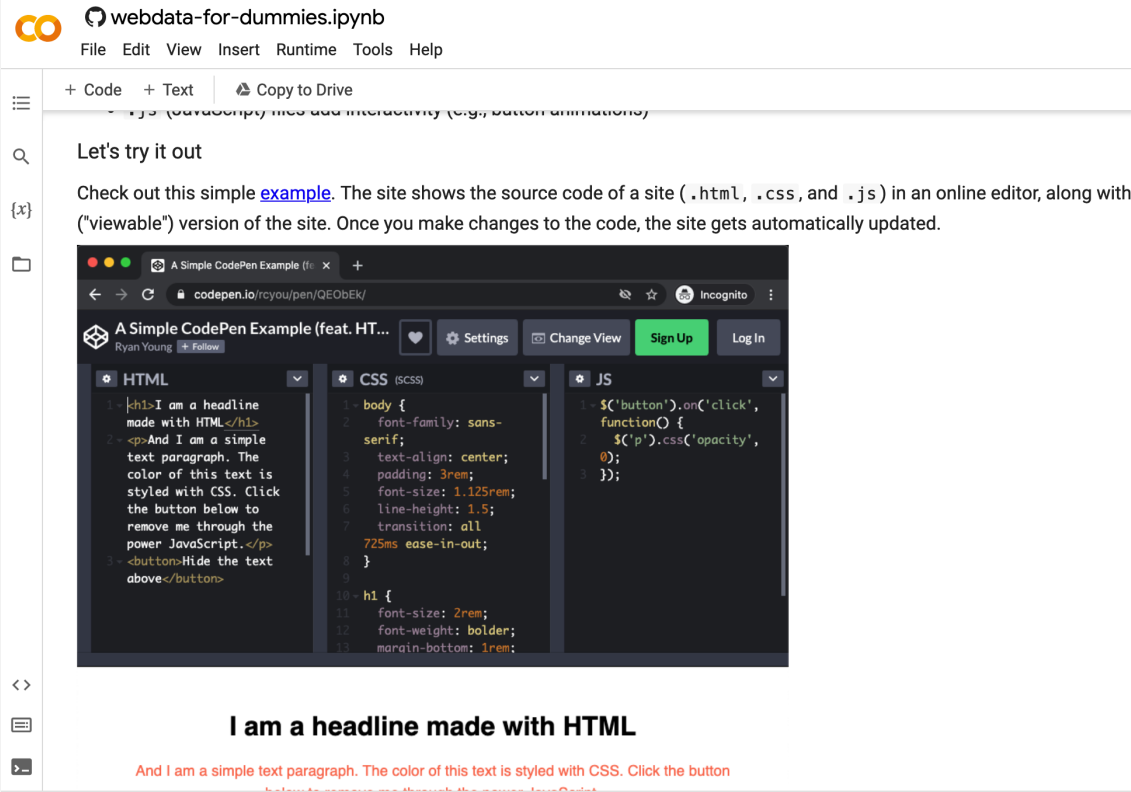
# **WEBSCRAPING & APIS FOR DUMMIES**

# WEB SCRAPING & APIS FOR DUMMIES

- Tutorial(s) available here

<https://odcm.hannedatta.com/docs/tutorials/webdata-for-dummies/>

- Discussion points
  - Any experience with scraping/APIs?
  - Which software tools do you use?



The screenshot shows a CodePen editor interface for a file named 'webdata-for-dummies.ipynb'. The editor has a menu bar with 'File', 'Edit', 'View', 'Insert', 'Runtime', 'Tools', and 'Help'. Below the menu bar are tabs for '+ Code', '+ Text', and 'Copy to Drive'. The main content area is titled 'Let's try it out' and contains the following text: 'Check out this simple [example](#). The site shows the source code of a site (.html, .css, and .js) in an online editor, along with ("viewable") version of the site. Once you make changes to the code, the site gets automatically updated.'

Below the text is a preview window showing a simple web page. The preview window has a title bar 'A Simple CodePen Example (feat. HT...' and a user profile 'Ryan Young'. It has buttons for 'Settings', 'Change View', 'Sign Up', and 'Log In'. The preview window shows three panels: HTML, CSS (SCSS), and JS. The HTML panel contains the following code:

```
<h1>I am a headline made with HTML</h1>
<p>And I am a simple text paragraph. The color of this text is styled with CSS. Click the button below to remove me through the power JavaScript.</p>
<button>Hide the text above</button>
```

The CSS panel contains the following code:

```
body {
  font-family: sans-serif;
  text-align: center;
  padding: 3rem;
  font-size: 1.125rem;
  line-height: 1.5;
  transition: all 725ms ease-in-out;
}

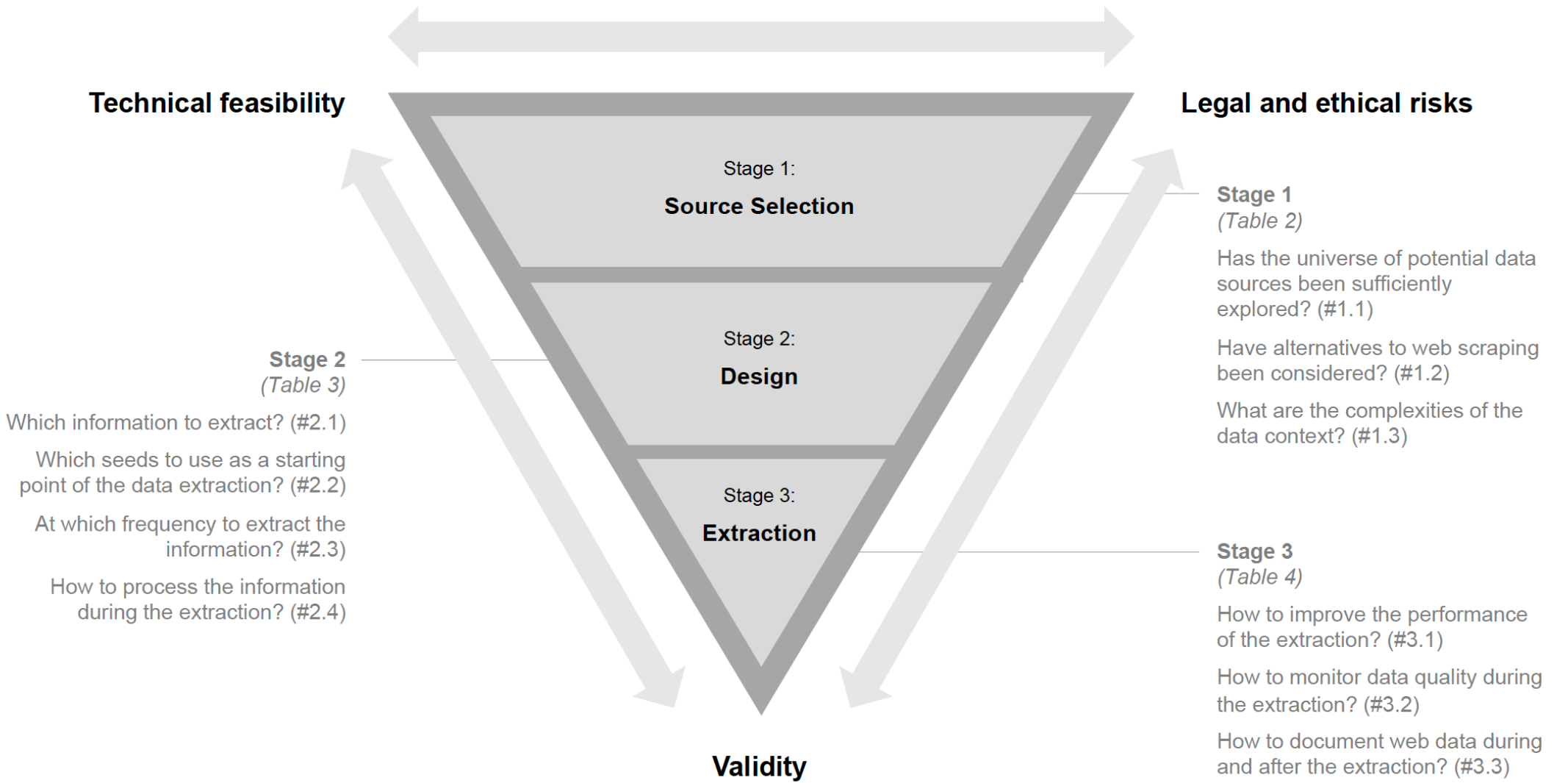
h1 {
  font-size: 2rem;
  font-weight: bolder;
  margin-bottom: 1rem;
}
```

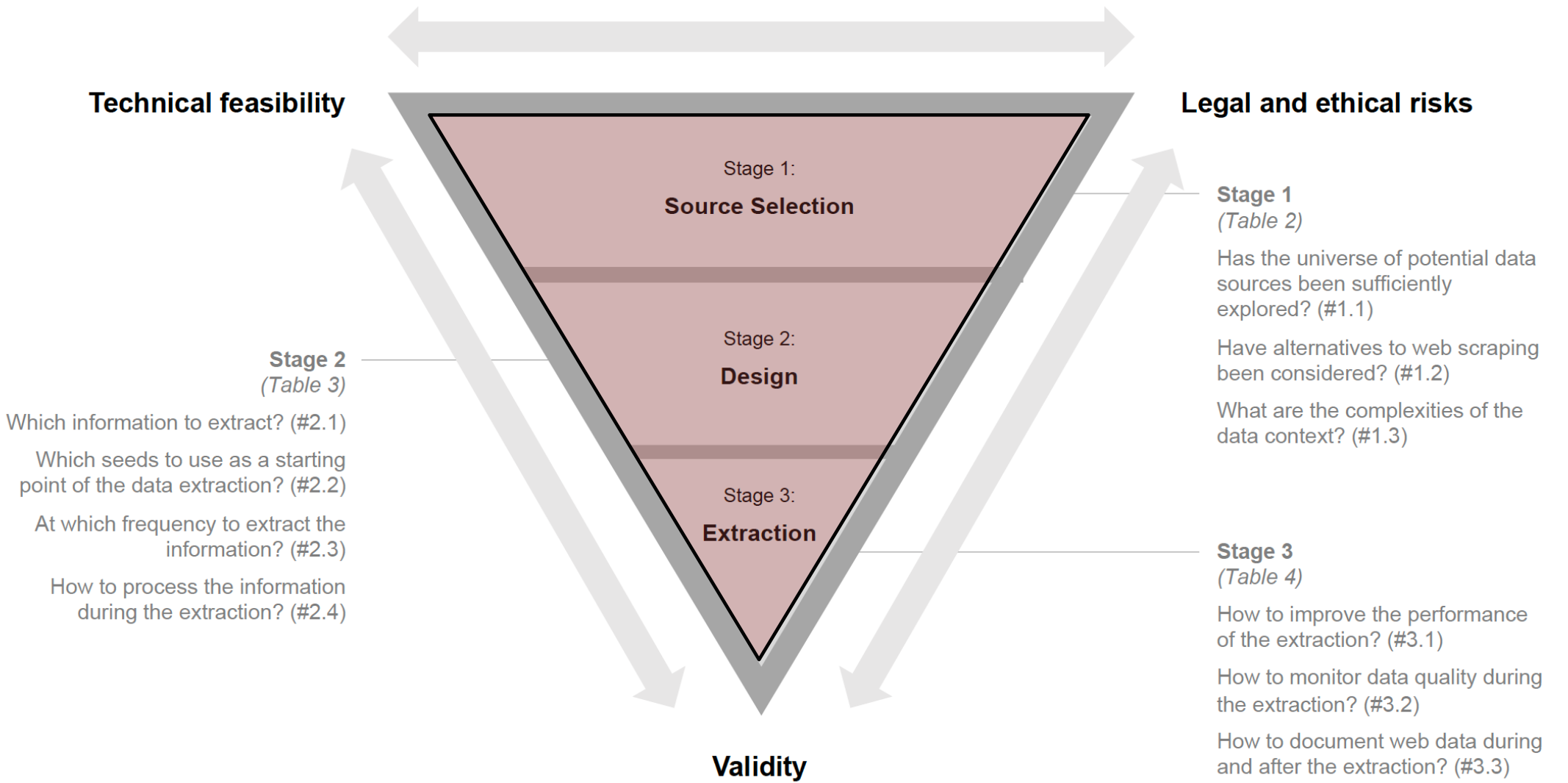
The JS panel contains the following code:

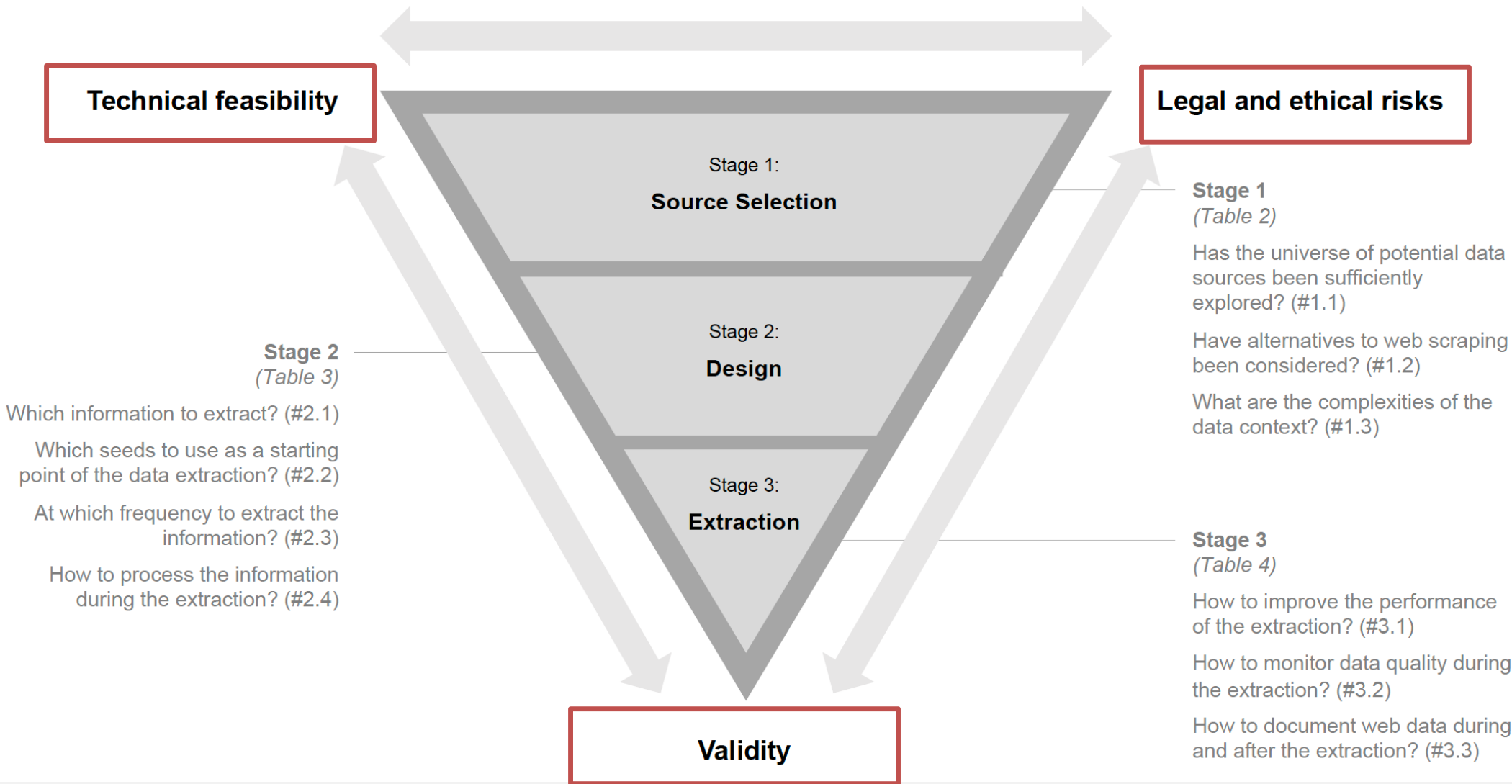
```
$('#button').on('click', function() {
  $('p').css('opacity', 0);
});
```

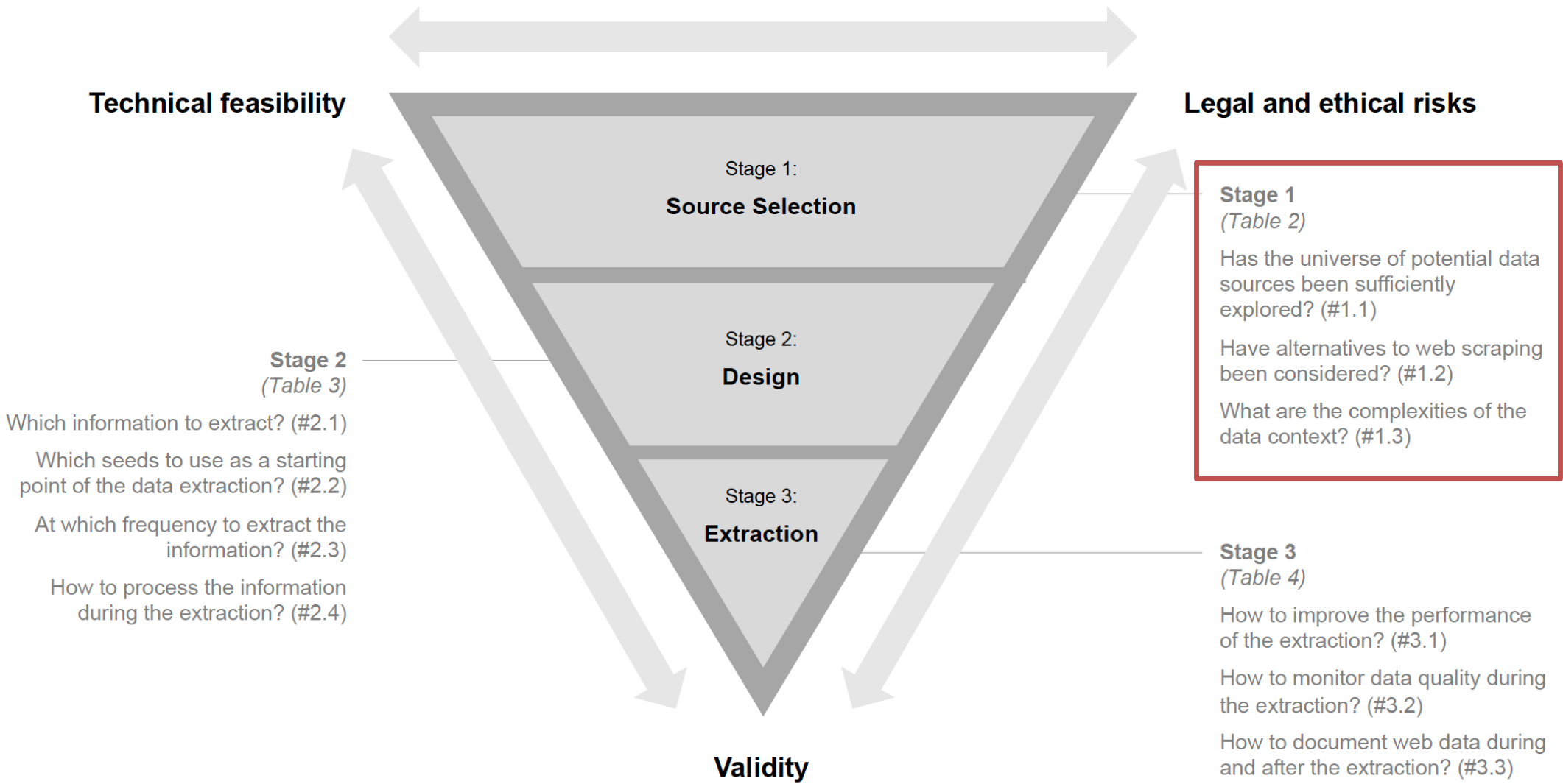
The rendered output of the code is shown below the preview window. It displays a large bold heading 'I am a headline made with HTML' and a paragraph of text 'And I am a simple text paragraph. The color of this text is styled with CSS. Click the button below to remove me through the power JavaScript.' The text is styled with a serif font, centered, and has a larger font size. A button is visible below the text.

# **METHODOLOGICAL FRAMEWORK & DESIGN DECISIONS**









## ▶ **Challenge #1.1: exploring the dataverse**

- Access to near-to infinite number of potential sources without traditional gatekeepers.
  - But sources vary vastly in terms of quality, stability, and retrievability.
- **Might prompt researchers to only consider dominant or familiar platforms only.**





## ▶ **Challenge #1.1: exploring the dataverse**

- Access to near-to infinite number of potential sources without traditional gatekeepers.
  - But sources vary vastly in terms of quality, stability, and retrievability.
- Might prompt researchers to only consider dominant or familiar platforms only.

BUT **thorough exploration** of the vast data universe allows for

- more compelling theory-testing
- identifying novel, emerging marketing phenomena



## ▶ **Challenge #1.1: solutions**

- Search from different angles (e.g., consumers, analysts, managers)
- Broaden geographical search criteria
- Identify related data sources using cross-searches on Google Trends
- Expand search to non-primary data providers (e.g., aggregators)
- Expand search by including terms such as “API” or “data set”
- Understand popularity and legitimacy of data sources
- Sign up to the service (e.g., to gain experience using the data source)
- Explore all pages available at a source
- Explore the website's source code

▶ **Looking beyond the usual suspects**



## ▶ Looking beyond the usual suspects

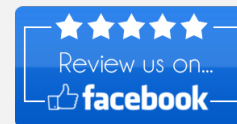


tripadvisor

amazon



How America finds a doctor.\*



## ▶ Looking beyond the usual suspects



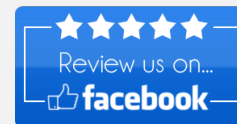
tripadvisor

amazon

SKYTRAX



How America finds a doctor.\*



**bol.com**

de winkel van ons allemaal

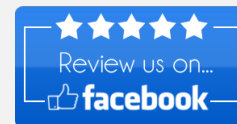
wayfair.co.uk

+ entity/page mapping

▶ Looking beyond the usual suspects



tripadvisor



## ▶ Challenge #1.1: reflection

- Search from different angles (e.g., consumers, analysts, managers)
- Broaden geographical search criteria
- Identify related data sources using cross-searches on Google Trends
- Expand search to non-primary data providers (e.g., aggregators)
- Expand search by including terms such as “API” or “data set”
- Understand popularity and legitimacy of data sources
- Sign up to the service (e.g., to gain experience using the data source)
- Explore all entities available at a source
- Explore the website’s source code

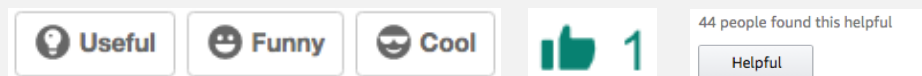
Is the suggested data source superior to existing or potentially collectable (non-) web data?

## ▶ **Challenge #1.1:** reflection | food for thought

- Selecting a source from the vast dataverse is challenging, yet critical

→ Remedy: Present a clear rationale to motivate the sampling choice; some useful approaches below:

- identify *idiosyncratic feature(s)* (e.g., Yelp funny votes; McGraw et al. 2015)



- particular *type* of webpage (e.g., discussion forum; Chen & Berger 2013)



## ▶ **Challenge #1.1:** reflection | food for thought

- Selecting a source from the vast dataverse is challenging, yet critical

→ Remedy: Present a clear rationale to motivate the sampling choice; some useful approaches below:

- identify *idiosyncratic feature(s)* (e.g., Yelp funny votes; McGraw et al. 2015)



- particular *type* of webpage (e.g., discussion forum; Chen & Berger 2013)

More compelling argumentation is facilitated:

- by **pretests** that demonstrate that certain instances (e.g., brands, Henkel et al. 2018; industries, Umashankar et al. 2017) map onto the focal construct(s) of interest
- picking a **representative example** for the focal construct  
[e.g., Paharia et al. (2014): Peet's Coffee = an underdog brand with a strong rival (i.e., Starbucks)]

## ▶ **Challenge #1.1:** reflection | food for thought

- Selecting a source from the vast dataverse is challenging, yet critical
- When agnostic about the source, sampling multiple websites can increase confidence about effect generalizability (e.g., Ordenes et al. 2019; Melumad et al. 2019)

## ▶ **Challenge #1.2:** defaulting to web scraping

- **Web scraping = most popular data retrieval method**

BUT might not always be optimal for many sources:

- APIs provide a structured and legit way to obtain web data
- data dumps are also widely available



## ▶ **Challenge #1.2:** defaulting to web scraping

- **Web scraping = most popular data retrieval method**

BUT might not always be optimal for many sources:

- APIs provide a structured and legit way to obtain web data
- data dumps are also widely available

Using APIs and data dumps may lead to...

- swifter data collection with better documentation
- novel research opportunities
- minimization of exposure to legal risk (*more on this later*)



## ▶ Challenge #1.2: solutions

There are many readily downloadable datasets that can be used in lieu of collecting your own novel dataset.



6,685,900 reviews, 192,609 businesses,  
200,000 pictures, 10 metropolitan areas

Other sources:

Recommender Systems Database, Kaggle, Webrobots.io, IMDb

→ Schoenmueller et al. (2020, JMR): <https://osf.io/6n2kt/>

## ▶ Selection of readily available datasets

- Recommender Systems Database:  
<https://cseweb.ucsd.edu/~jmcauley/datasets.html>
- <https://www.kaggle.com/>
- Webrobots.io's Kickstarter datasets:  
<https://webrobots.io/kickstarter-datasets/>
- IMDb: <https://www.imdb.com/interfaces/>
- Various datasets cf. <https://osf.io/6n2kt/>  
(see the web appendix of <https://doi.org/10.1177/0022243720941832>)

## ▶ **Challenge #1.2:** reflection

- ☑ Formulate an explicit rationale for extraction method (e.g., scraping vs. API vs. other). ☑

## ▶ **Challenge #1.3:** capturing contextual complexity

- Despite its richness, web data on its own **may not include sufficient information** about the complex context it is generated in.
  - lack of documentation





## ▶ **Challenge #1.3:** capturing contextual complexity

- Despite its richness, web data on its own **may not include sufficient information** about the complex context it is generated in.
  - lack of documentation

Researchers need to **proactively identify and capture relevant meta data** (i.e., data about the focal web data).



## ▶ Challenge #1.3: example

★★★★★ Well worth its cost.

October 5, 2017

Style: W/ CR123A Batteries | Package Type: Plastic Clamshell Pa

Without a doubt, a top notch light instrument for everyday carry, never leaves my possession. I've kept it clipped into a back pocket. Furthermore, the lumen power is plenty powerful enough to more. I've exposed it to free flowing water... to extended day and overnight shifts. You won't be disappointed... especially if you also purchase 18650 Button Top AC Li-Ion 120V which is also found here on Amazon.

11 people found this helpful

|  |



3 people found this helpful



Helpful |  |



35 people found this helpful



Helpful |  |

## ▶ **Challenge #1.3: solutions**

- Understand changes to the data generating mechanism (e.g., mapping out changes over time using archive.org (for websites), or versions of the API)
- Search for blogs, sites with press releases, a firm's "changelogs" on software about important (technical) firm developments
- Explore forms/user forums to see how users talk about the service
- Use reverse-search on Google
- Identify market and usage statistics (e.g., using Statista, news media, social media, firm reports & stock filings)
- Identify competitive landscape and dependencies with other firms/services
- Assess whether and how entities can be linked to external data

## ▶ Challenge #1.3: reflection

- Understand changes to the data generating mechanism (e.g., mapping out changes over time using archive.org (for websites), or versions of the API)
- Search for blogs, sites with press releases, a firm's "changelogs" on software about important (technical) firm developments
- Explore forms/user forums to see how users talk about the service
- Use reverse-search on Google
- Identify market and usage statistics (e.g., using Statista, news media, social media, firm reports & stock filings)
- Identify competitive landscape and dependencies with other firms/services
- Assess whether and how entities can be linked to external data

☑ Which statistics would you consider informative about the data, which cannot be captured on the targeted source itself?  
Which sources could you potentially gather such data from? ☑

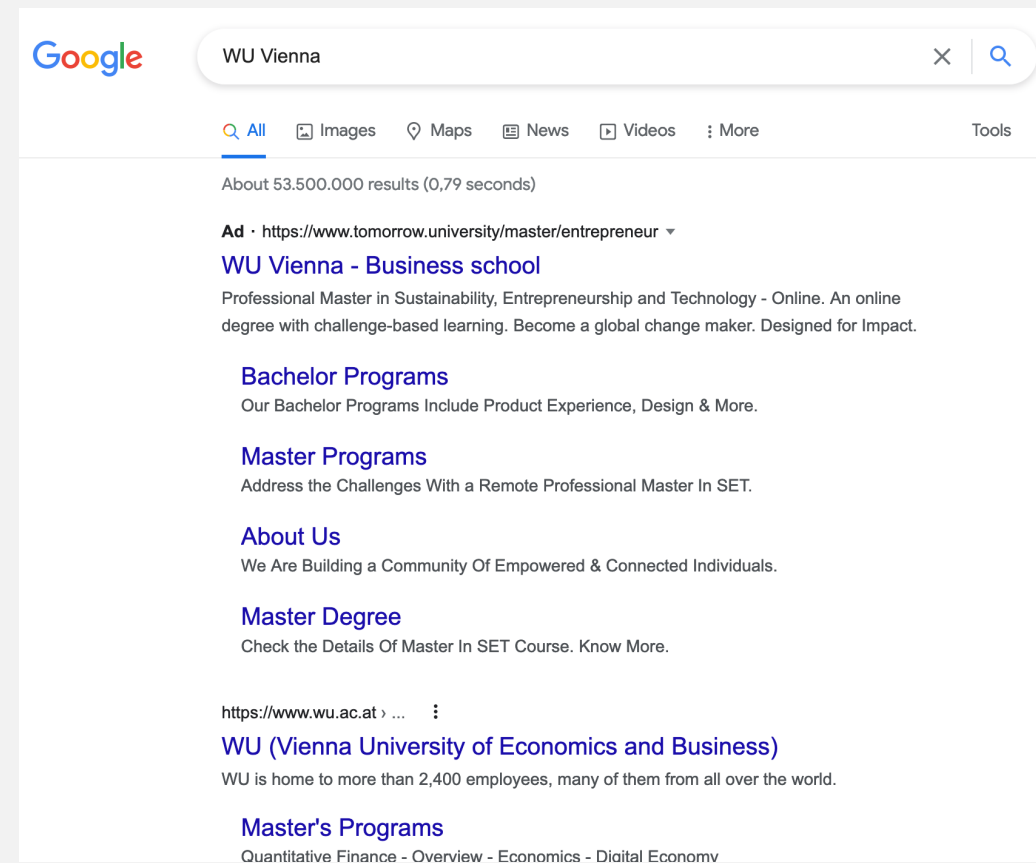
## ▶ **Your projects**

- Let's take YOUR (early?) projects and apply some of our proposed solutions.



## ▶ Alexander's project

- Input: List of keywords
- Scraper: Put keywords in Google search, save organic and paid ads
- Requirement: Many keywords, US-based, rotating IPs
- Affects...
  - Validity: e.g., generalizability
  - Tech. feasibility: rotating IP addresses, being blocked
  - Legal issues: Google doesn't want to be scraped...!



Google search results for "WU Vienna". The search bar shows "WU Vienna" and the results page displays "About 53.500.000 results (0,79 seconds)". The top result is an advertisement for "WU Vienna - Business school" with a link to <https://www.tomorrow.university/master/entrepreneur>. Below the ad are several program links: "Bachelor Programs", "Master Programs", "About Us", "Master Degree", and "Master's Programs".

Google

WU Vienna

All Images Maps News Videos More Tools

About 53.500.000 results (0,79 seconds)

Ad · <https://www.tomorrow.university/master/entrepreneur>

**WU Vienna - Business school**

Professional Master in Sustainability, Entrepreneurship and Technology - Online. An online degree with challenge-based learning. Become a global change maker. Designed for Impact.

**Bachelor Programs**

Our Bachelor Programs Include Product Experience, Design & More.

**Master Programs**

Address the Challenges With a Remote Professional Master In SET.

**About Us**

We Are Building a Community Of Empowered & Connected Individuals.

**Master Degree**

Check the Details Of Master In SET Course. Know More.

<https://www.wu.ac.at> › ...

**WU (Vienna University of Economics and Business)**

WU is home to more than 2,400 employees, many of them from all over the world.

**Master's Programs**

Quantitative Finance - Overview - Economics - Digital Economy

## Technical feasibility

## Legal and ethical risks

Stage 1:  
**Source Selection**

Stage 2:  
**Design**

Stage 3:  
**Extraction**

**Validity**

### Stage 1 (Table 2)

Has the universe of potential data sources been sufficiently explored? (#1.1)

Have alternatives to web scraping been considered? (#1.2)

What are the complexities of the data context? (#1.3)

### Stage 3 (Table 4)

How to improve the performance of the extraction? (#3.1)

How to monitor data quality during the extraction? (#3.2)

How to document web data during and after the extraction? (#3.3)

### Stage 2 (Table 3)

Which information to extract? (#2.1)

Which seeds to use as a starting point of the data extraction? (#2.2)

At which frequency to extract the information? (#2.3)

How to process the information during the extraction? (#2.4)

## ▶ Challenge #2.1: Which information to extract, from which page?

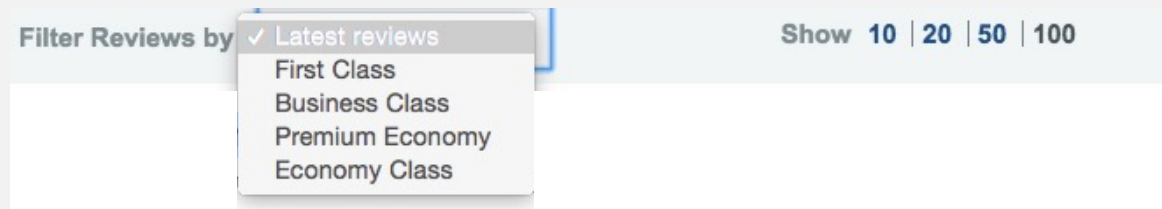
- Go to Amazon.com and find out on which pages you can find reviewer information (e.g., valence, text, demographics, ...)
- Make a list of URLs & variables you could capture





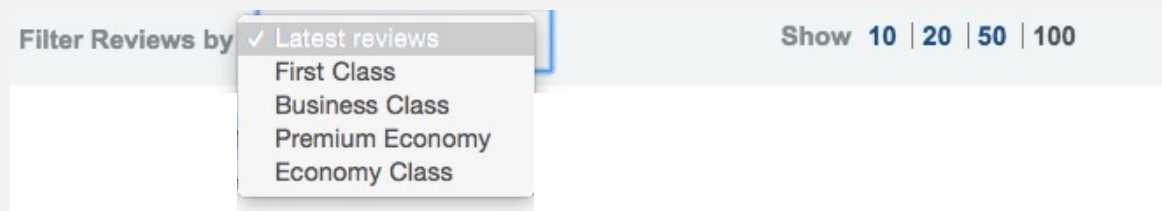
## ▶ Preparation for effective extraction

- Understand & leverage the **structure** of the target website
  - Are there different ways to display the data (e.g., cases per page)?
  - Are there different ways to sort the data (e.g., date, rating, recommended)?
  - Are there other relevant exclusion filters (e.g., language)?



## ▶ Preparation for effective extraction

- Understand & leverage the **structure** of the target website
  - Are there different ways to display the data (e.g., cases per page)?
  - Are there different ways to sort the data (e.g., date, rating, recommended)?
  - Are there other relevant exclusion filters (e.g., language)?



- How consistent is the structure across different pages? Are there “unusual” incidences that deviate from the basic structure (e.g., updated reviews)?
- ▶ **Goal:** minimize the burden on the server + maximize data quality & reproducibility

Influence, New and Expanded: The Psychology of Persuasion Customer reviews

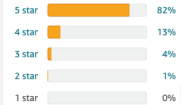
**Customer reviews** **Influence, New and Expanded...** by Robert B Cialdini PhD

★★★★★ 4.8 out of 5

748 global ratings

See All Buying Options

Add to Wish List



Write a review

How are ratings calculated?

**Top positive review**  
All positive reviews

Peter Clarke, Ph.D.  
★★★★★ **Authoritative and practical**  
Reviewed in the United States on May 4, 2021  
Like other seasoned readers of books about human behavior, I've grown justifiably skeptical about "new editions" of these works, suspecting a light update or even just a new forward. But Robert Cialdini's Influence, New and Expanded shatters these expectations.

**Top critical review**  
All critical reviews

Dennis  
★★★★☆ **Very insightful but too repetitive**  
Reviewed in the United States on August 11, 2021  
The book has a lot of good information and insights, but it becomes repetitive. The chapters on Commitment and Unity especially feel like they drag on forever.

**DARK PSYCHOLOGY 6 BOOKS IN 1: Int...**  
\$35.07 or Prime **Shop now**

★★★★☆ 454

Sponsored

Read more  
74 people found this helpful

Need customer service? [Click here](#)

Search customer reviews

**SORT BY** **FILTER BY**  
Most r... All reviewers All stars Text, image, ...

748 global ratings | 78 global reviews

**From the United States**

Frank Z.  
★★★★★ **Love the book!**  
Reviewed in the United States on December 13, 2021  
**Verified Purchase**  
I got a friend from an event that he recommended me this book. It enlightens me about marketing. I'm happy to read this book during pandemic. Gave me a lot of ideas. Highly recommended.

Helpful Report abuse

Jose S  
★★★★★ **Magnificent**  
Reviewed in the United States on December 6, 2021  
**Verified Purchase**  
This book is very good. Quite illustrating. The approach with which it was created really captured my attention.  
Highly recommend  
One person found this helpful

Helpful Report abuse

Jay  
★★★★☆ **Intimidated by the length of the book, but well worth reading**  
Reviewed in the United States on November 29, 2021  
**Verified Purchase**  
I loved this book and it had so many eye opening stories to share. I am glad that I am now aware of what social researchers and compliance professionals are capable of. As stated in the book, hopefully it is more often done for good. If it is not done for good, at least the reader can be better prepared to combat against it. I would have given it five stars but you'll have to read the book to find out why 4 is as high as I could go. I will definitely come back to this book in an effort to retain the information.  
One person found this helpful

## ▶ **Challenge #2.1: Impact on validity, tech. feasibility & legal/ethical risks**

- Validity implications
  - Is information subject to algorithmic biases or missing data?
  - Are there significant changes to the data-generating process?
  - Meta data required to make sense of variables?
- Legal/ethical risks
  - Publicly accessible vs. login? Consent to ToU? Implicit or explicit?
  - Personal or sensitive information?
  - Overlap original intent of posting & research question / scientific justification
- Technical feasibility?
  - All information extractable?
  - Limits to iterating through pages?
  - Does the extraction software obtain information reliably?

## ▶ Challenge #2.1: Impact on validity, tech. feasibility & legal/ethical risks

- Validity implications
  - Is information subject to algorithmic biases or missing data? *Delete cookies & check?*
  - Are there significant changes to the data-generating process? *Archive.org*
  - Meta data required to make sense of variables? *Save timestamps/IP addresses*
- Legal/ethical risks
  - Publicly accessible vs. login? Consent to ToU? Implicit or explicit? *Find public pages*
  - Personal or sensitive information? *Anonymize while collecting*
  - Overlap original intent of posting & research question / scientific justification *Tweak RQ; be critical!*
- Technical feasibility?
  - All information extractable? *Build prototype*
  - Limits to iterating through pages? *Check last page, try a few inbetween*
  - Does the extraction software obtain information reliably? *Run it for longer amount of time*

## ▶ **Challenge #2.2: Which seeds to use?**

- With scraping, you don't have access to a firm's "entire" database,
- but... you only see snapshots of the data
  
- **How do you go from a snapshot to a sample?**
  
- Collect so-called seeds
  - On Amazon.com – a list of books (e.g., from a product category)
  - On Etsy.com – a list of projects, e.g., using a search term
  - On social networks – a list of "seeding" users

## ▶ **Challenge #2.2: Which seeds to use?**

- Check out [trakt.tv](https://trakt.tv) (a site that monitors movie streaming services)
- **Come up with ways to sample users from the site!**



## ▶ Challenge #2.2: Which seeds to use?

- So far, we have talked about “internal seeds”
- Sometimes, “external” seeds are also useful
  - **Necessity**: *ability to link data (e.g., via identifiers)*



## ▶ **Challenge #2.2: Impact on validity, tech. feasibility & legal/ethical risks**

- Validity implications
  - Sample size sufficient to effectively inform the research question?
  - To which population does the sample generalize? Is it random?
  - How prevalent is panel attrition of seeds?
- Legal/ethical risks
  - Excessive portion relative to all data available?
  - Similar data available elsewhere? RQ only answerable with this data?
  - Are there any (potentially) vulnerable seeds?
- Technical feasibility?
  - Is the required sample size technically feasible?
  - Can external seeds be consistently matched to the web data?

## ▶ Challenge #2.2: Impact on validity, tech. feasibility & legal/ethical risks

- Validity implications
  - Sample size sufficient to effectively inform the research question? Explore alternative click paths
  - To which population does the sample generalize? Is it random? Internal vs. external seeds
  - How prevalent is panel attrition of seeds? Collect extra meta data
- Legal/ethical risks
  - Excessive portion relative to all data available? Narrow down collection
  - Similar data available elsewhere? RQ only answerable with this data? Reduce dependency
  - Are there any (potentially) vulnerable seeds? Exclude by design, anonymize
- Technical feasibility?
  - Is the required sample size technically feasible? Prototype
  - Can external seeds be consistently matched to the web data? E.g., search by ID

## ▶ **Challenge #2.3: At which frequency to extract data?**

- When to extract the data?
  - Cross-sectional, one-time captures
  - vs.
  - over-time captures
- How to schedule the collection?
  - E.g., hourly, daily, weekly
- **Can you come up with ideas of when extraction at higher frequency makes sense?**



## ▶ Frequency directly affects *technically feasible* sample size

- **Project how long a data collection will take**
- Basics
  - How many seeds?
  - How many pages to visit, for each seed?
  - How long does it take to extract data for one page?
  - How many computers do you use?
  - How often to run the data collection (frequency)?
- Build an Excel model of sample size / frequency & time it takes to collect data



## ▶ Challenge #2.3: Impact on validity, tech. feasibility & legal/ethical risks

- Validity implications
  - Extraction frequency in sync w/ studied phenomena? Any gains from a live collection?
  - Refresh rate of source sufficient? Check over time
  - Data really archival? Consistently available over time? Data source theory
- Legal/ethical risks
  - Excessive server load due to extraction frequency? Inspect robots.txt, adhere to defaults
  - Does sensitivity increase because of higher frequencies?
- Technical feasibility?
  - Technical hurdles (e.g., blocking)? Consider costs for storage
  - How to guarantee stability over time? Monitoring
  - How to distinguish batches of data? Meta data enrichment

## ▶ **Challenge #2.4: How to process data *during* the collection**

- Web data is "messy"
- Many directly clean data in scraping/API scripts, and then save them in CSV/Excel files
- **What could be a potential problem in cleaning data on-the-fly?**

## ▶ **Challenge #2.4: Impact on validity, tech. feasibility & legal/ethical risks**

- Validity implications
  - Could erroneous processing lead to data loss?
  - Could there be scientific value in retaining raw data?
- Legal/ethical risks
  - Collected data in conflict w/ GDPR?
  - Secured from unauthorized access?
  - Anonymization required?
- Technical feasibility?
  - Which storage facility to use?
  - Normalization necessary?

## ▶ Challenge #2.4: Impact on validity, tech. feasibility & legal/ethical risks

- Validity implications
  - Could erroneous processing lead to data loss?
  - Could there be scientific value in retaining raw data?

But: always parse some minimal amount for monitoring  
Ensure proper encoding, retain in original format
- Legal/ethical risks
  - Collected data in conflict w/ GDPR?
  - Secured from unauthorized access?
  - Anonymization required?

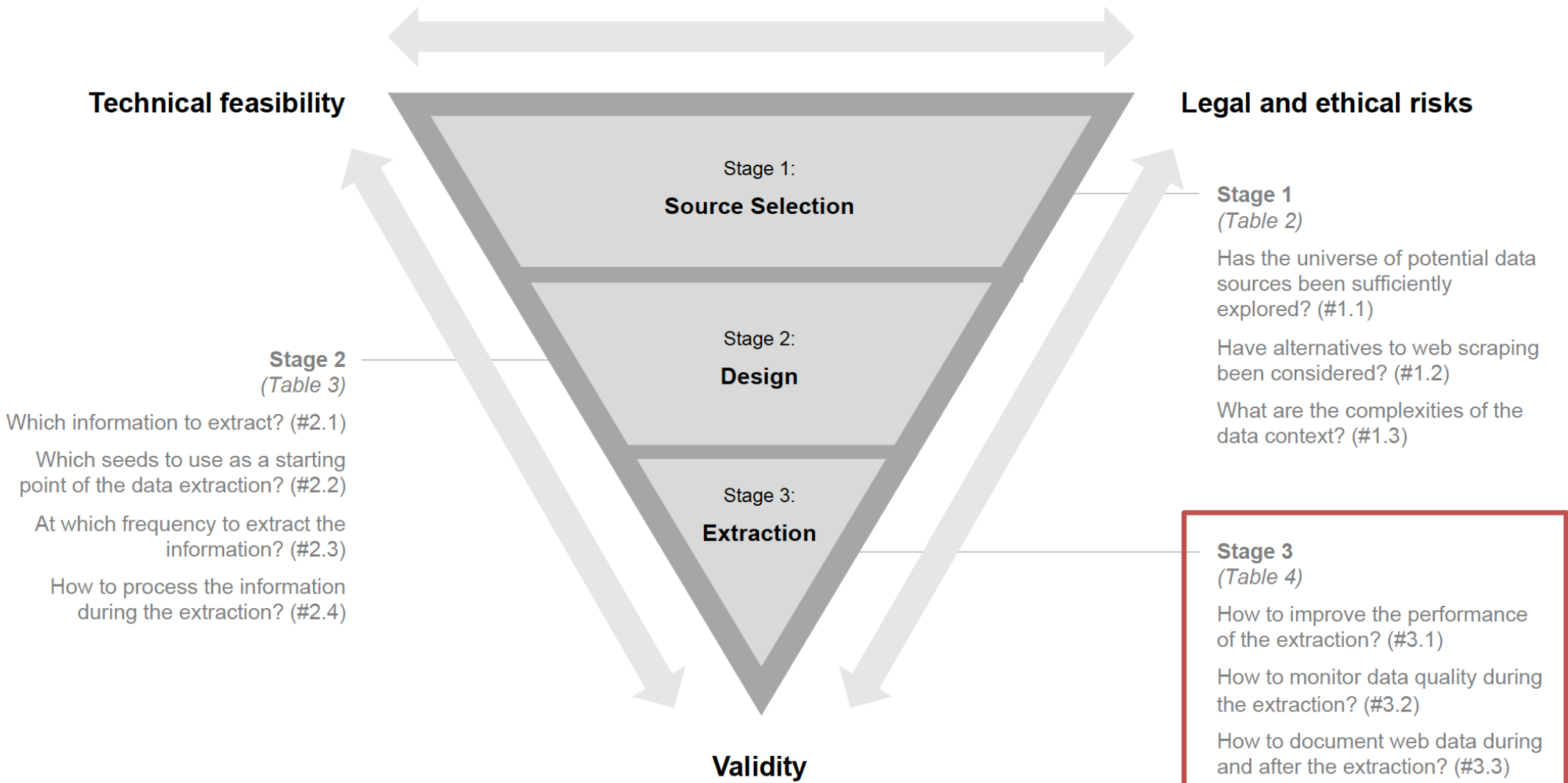
Use third parties in case of highly sensitive information
- Technical feasibility?
  - Which storage facility to use?
  - Normalization necessary?

Project storage costs  
“How will the data be used?”



## ▶ Alexander's project

- Input: List of keywords
- Scraper: Put keywords in Google search, save organic and paid ads
- Requirement: Many keywords, US-based, rotating IPs
- Affects...
  - Validity: e.g., generalizability
  - Tech. feasibility: rotating IP addresses, being blocked
  - Legal issues: Google doesn't want to be scraped...!
- What could be the most efficient way to achieve IP address rotation?
  - Look for APIs that do scraping
  - Look for proxy providers w/ Python packages/R packages to rotate IPs
- Can you suggest any code implementations?
  - <https://www.scrapehero.com/how-to-rotate-proxies-and-ip-addresses-using-python-3/>
- Is there an existing package/framework for R/Python that achieves this?
  - Would rely on commercial scraping API here
  - Duration of project?



## ▶ **Challenge #3.1: Increase performance**

**Quick & dirty code that “just works” may be suboptimal & threaten validity!**

- Use stable selectors (e.g., tags, classes, attributes, styles associated with particular information), and make only selective use of error handling
- When using APIs, choose a stable and supported version
- Check for traces of being banned/blocked/slowed down by the website
- Update the technically feasible retrieval limit
- Verify that computing resources are appropriate (e.g., scale up or down servers, verify that database runs optimally)
- Consider potential benefits from using cloud computing (e.g., for extended, uninterrupted data collection) vs. benefits from local setups (e.g., due to security or privacy concerns)

## ▶ **Challenge #3.2: Monitor data collection**

**Imagine your data collection broke, and you didn't notice it...**

- Log each web request (i.e., URL call), along with response status codes, timestamps of when the collection was started, and when the request was made
- Save raw HTML websites, along with the parsed data, and use them for triangulation
- Verify whether the raw data was correctly parsed (e.g., for a sample of information, compare raw data and parsed data)
- Check file sizes or the number of observations at regular intervals
- Set up monitoring tool (e.g., based on number of files retrieved or requests made, file sizes retrieved, time the collection last ran)
- Automatically generate reports on data quality (e.g., using RMarkdown)
- Record issue(s) in a logbook (e.g., in the documentation); especially if considered critical for data quality
- Extrapolate and monitor costs (e.g., API subscription, storage, and cloud computing)

## ▶ **Challenge #3.3: Document the data**

**Nobody, except you, know how the data was generated!**

- Start from a template (e.g., Datasheets for Datasets, Gebru et al. 2020), and use it during the early stages of the collection
- Maintain a logbook in which to note important events (e.g., when the collection broke down and why)
- Keep and organize copies of relevant files (e.g., screenshots of the website at the time of data extraction, the API documentation, details on variable operationalization with summary statistics, information about the context, etc.)
- Have a plan for long-term, archival storage (e.g., re3data.org), and consider which license to use for the data (e.g., Creative Commons)

▶ **Your questions/submissions**



## ▶ Summary

- Web scraping and APIs have advanced our field (higher citations, about 20% of all publications use web data)
- Web data is largely accessible
- But: there is no “download button” & merely writing a bit of code doesn't guarantee your data is really valid
- Use **methodological framework** to balance validity w/ technical feasibility and legal/ethical concerns

## ▶ **The future is bright... (I)**

### 1. Branch out and reveal the invisible!

1. Draw from underutilized sources
2. Draw from multiple sources
3. Rediscover frequently used sources (e.g., overlooked pages)
4. Altering the extraction frequency

### 2. Boost ecological validity

1. Use scraping for stimuli generation
2. Run self-administered experiments



## ▶ The future is bright... (II)

### 3. New method

1. Can webdata replace traditional marketing metrics? (e.g., advertising & brand equity data is expensive!)
2. Operate API-based microservices

### 4. Enhance inferences through efficiency gains

1. Get data on offline behaviors (e.g., weather, holidays, ...)
2. "Academic's little helper"



# THANK YOU.

BOEGERSHAUSEN@RSM.NL

H.DATTA@TILBURGUNIVERSITY.EDU

ABHISHEK.BORAH@INSEAD.EDU

ANDREW.STEPHEN@SBS.OX.AC.UK

+ <https://tiu.nu/scraping>

+ <https://odcm.hannedatta.com>

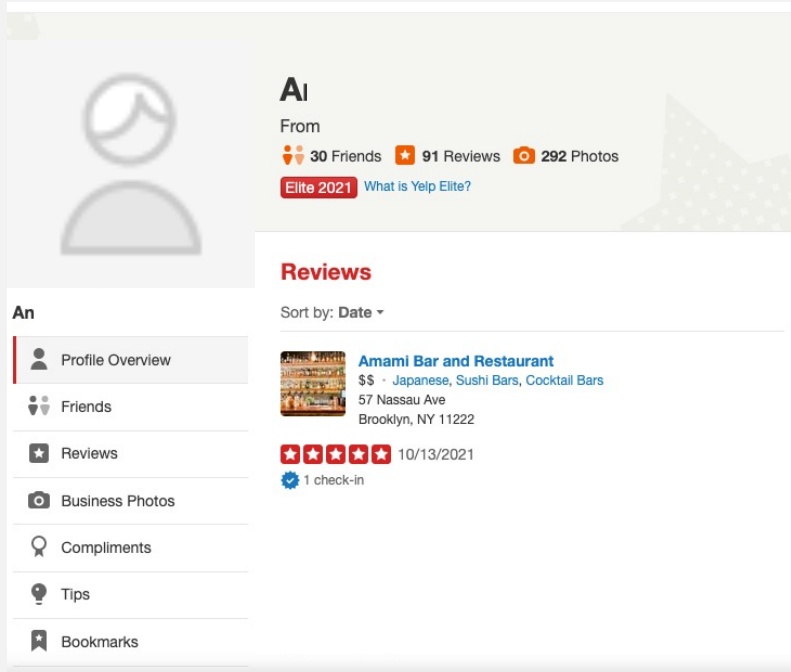
+ <https://tilburgsciencehub.com>



**BACKUP**

## ▶ Design challenges: entities example

- “I want to examine the behavior of reviewers over time”



[https://www.yelp.com/user\\_details?userid=](https://www.yelp.com/user_details?userid=)  
[https://www.yelp.com/user\\_details\\_reviews\\_self  
?userid=XXX&review\\_filter=category  
&category\\_filter=restaurants](https://www.yelp.com/user_details_reviews_self?userid=XXX&review_filter=category&category_filter=restaurants)

## ▶ **Design challenges:** data source theory

- What are **your essential assumptions** about the configuration, data generation process, and characteristics of the data to test predictions?

Recursive process of *formulating a “**data source theory**”* outlining these assumptions, testing, and refining the theory as required (Landers et al. 2016)

## ▶ **Design challenges:** data source theory

- What are **your essential assumptions** about the configuration, data generation process, and characteristics of the data to test predictions?

Recursive process of *formulating a “**data source theory**”* outlining these assumptions, testing, and refining the theory as required (Landers et al. 2016)

- Case study:

Prediction: # friends on Yelp → usage of emotional language in reviews (+)

Sample: all reviews of the 5 most reviewed Japanese restaurants in 5 US cities (NYC, LA, SF, CHI, DC)

## ▶ Design challenges: data source theory

- What are **your essential assumptions** about the configuration, data generation process, and characteristics of the data to test predictions?




Recursive process of formulating a “**data source theory**” outlining these assumptions, testing, and refining the theory as required (Landers et al. 2016)

- Case study:

Prediction: # friends on Yelp → usage of emotional language in reviews (+)

Sample: all reviews of the 5 most reviewed Japanese restaurants in 5 US cities (NYC, LA, SF, CHI, DC)

**User A**  
(scraped today)

 300 friends  
 437 reviews  
 775 photos  
**Elite '2019**

**User A's review in our dataset**  
(scraped today)

 **Sushi House**  
\$\$ · Japanese, Sushi Bars

 1/26/2014

## ▶ Design challenges: data source theory

- What are **your essential assumptions** about the configuration, data generation process, and characteristics of the data to test predictions?




Recursive process of formulating a “**data source theory**” outlining these assumptions, testing, and refining the theory as required (Landers et al. 2016)

- Case study:

Prediction: # friends on Yelp → usage of emotional language in reviews (+)

Sample: all reviews of the 5 most reviewed Japanese restaurants in 5 US cities (NYC, LA, SF, CHI, DC)

**User A**  
(scraped today)

 300 friends  
 437 reviews  
 775 photos  
**Elite '2019**

**User A's review in our dataset**  
(scraped today)

 **Sushi House**  
\$\$ · Japanese, Sushi Bars

 1/26/2014

*Any issues here?*



## ▶ Design challenges: data source theory

- What are **your essential assumptions** about the configuration, data generation process, and characteristics of the data to test predictions?




Recursive process of formulating a “**data source theory**” outlining these assumptions, testing, and refining the theory as required (Landers et al. 2016)

- Case study:

Prediction: # friends on Yelp → usage of emotional language in reviews (+)

Sample: all reviews of the 5 most reviewed Japanese restaurants in 5 US cities (NYC, LA, SF, CHI, DC)

**User A**  
(scraped today)

 300 friends  
 437 reviews  
 775 photos  
**Elite '2019**

**User A's review in our dataset**  
(scraped today)

 **Sushi House**  
\$\$ · Japanese, Sushi Bars

 1/26/2014

**User A**  
joined on  
1/26/2014

