

Ryerson  
University

TED  
ROGERS  
SCHOOL  
OF MANAGEMENT



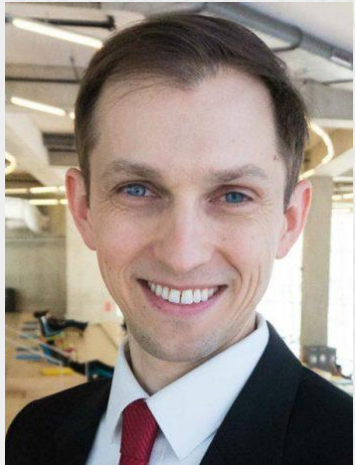
# Social Media Lab's Computational Social Science Bootcamp Summer 2021

## Session 1:

Getting Started with  
Commanalytic:  
Data Collection from Reddit

May 13, 2021

# Bootcamp Instructors



**Dr. Anatoliy Gruzd**  
Canada Research Chair,  
Associate Professor, Director  
of Research at the Ryerson  
University Social Media Lab



**Philip Mai M.A., J.D.**  
Co-Director and Senior  
Researcher at the Ryerson  
University Social Media Lab

## Social Media Lab's Computational Social Science (CSS) Bootcamp

Summer 2021

# CSS Bootcamp Schedule Summer 2021

**Session #1** Getting Started with Communalytic: Data Collection from Reddit

May 13, 2021,  
10:00-11:30am  
(EDT)

**Session #2** Toxicity Analysis with Reddit Data using Perspective API

May 27, 2021,  
10:00-11:30am  
(EDT)

**Session #3** Getting Started with Communalytic: Data Collection from Twitter (Twitter Thread via API v2.0 and Twitter Academic Track)

June 10, 2021,  
10:00-11:30am  
(EDT)

**Session #4** Toxicity Analysis of Twitter Threads using Perspective API

June 24, 2021,  
10:00-11:30am  
(EDT)

**Session #5** Social Network Analysis of Signed Networks with Reddit and Twitter data

July 8, 2021,  
10:00-11:30am  
(EDT)

**Session #6** Getting Started with Communalytic: Data Collection from Facebook & Instagram (via CrowdTangle API) + Social Network Analysis of Two-mode Semantic Networks with CrowdTangle data

July 22, 2021,  
10:00-11:30am  
(EDT)





# We're Hiring a Postdoctoral Researcher to Study Dis/Mis-Information Campaigns at Scale.



- Must have expertise in developing & evaluating distributed computing solutions for large-scale network visualization and analysis
- Ideal for candidates with a doctorate in Computational Social Science, Information Systems, Computer Science, Network Science, Complex Systems, Computer Engineering, or a related field
- **More Info: [SocialMediaLab.ca](https://socialmedialab.ca)**





# The Lab & The People

“Making Sense of a Networked World”

## Research Foci



Academia

Social Media for: Teaching | Learning | Research | Learning Analytics | Alt-metrics



Analytics

Development of analytical tools and methods for social media analysis



Health

Health, Well-being and Social Media



Information Privacy

Social Media Data Use, Reuse & Misuse | Privacy & Trust | Privacy Paradox



Misinformation

The Spread of Dis- and Misinformation via Social Media Sites



Online Communities

Social Networks | Online Crowds & Communities



Organizational Social Media Use

Adoption & Usage Trends | Factors Predicting Engagement | ROI | Data Analytics & Visualizations



Politics

Political Engagement, Polarization, Social Movements and E-Government

# Public Reports

Our various public reports provide a snapshot of social media usage trends in Canada and attitudes of Canadians towards issues such as misinformation, politics and privacy in the context of social media.

The cover features a red background with a white border. At the top left is the SocialMediaLab.ca logo. The date 'May 11, 2020' is centered. The title 'INOCULATING AGAINST AN INFODEMIC:' is in large, bold, white letters on a red background. Below it, the subtitle 'A Canada-Wide COVID-19 News, Social Media, and Misinformation Survey' is in smaller white text. The authors 'By Anatoly Gruzd and Philip Mai' are listed at the bottom. Logos for Ryerson University and Ted Rogers School of Management are at the bottom left. Contact information for media inquiries is at the bottom right. A recommended citation is at the very bottom.

SocialMediaLab.ca  
May 11, 2020

## INOCULATING AGAINST AN INFODEMIC:

A Canada-Wide COVID-19 News, Social Media, and Misinformation Survey

By Anatoly Gruzd and Philip Mai

FOR MEDIA OR OTHER INQUIRIES:  
Email: [info@socialmediablab.ca](mailto:info@socialmediablab.ca)

Ryerson University | TED ROGERS SCHOOL OF MANAGEMENT

RECOMMENDED CITATION: Gruzd & Mai. (2020). Inoculating against an Infodemic: A Canada-wide COVID-19 News, Social Media, and Misinformation Survey. Ryerson University Social Media Lab. DOI: [10.5663/SP2/JLULYA](https://doi.org/10.5663/SP2/JLULYA)

The cover has a white background with a red header bar containing the date 'July 2020'. The SocialMediaLab.ca logo is at the top right. The title 'THE STATE OF SOCIAL MEDIA IN CANADA 2020' is in bold black letters. Below it, the authors 'By Anatoly Gruzd and Philip Mai' are listed. The central image shows a group of stylized blue figures holding up various social media devices (phones, tablets) with icons for likes, shares, and comments, set against a blue background with a red maple leaf. The subtitle 'A Canada-Wide Survey About Social Media Adoption and Use in Canada' is at the bottom of the image. Contact information for media inquiries is at the bottom right. Logos for Ryerson University and Ted Rogers School of Management are at the bottom left. A recommended citation and Creative Commons license are at the very bottom.

July 2020

SocialMediaLab.ca

## THE STATE OF SOCIAL MEDIA IN CANADA 2020

By Anatoly Gruzd and Philip Mai

A Canada-Wide Survey About Social Media Adoption and Use in Canada

FOR MEDIA OR OTHER INQUIRES:  
Email: [info@socialmediablab.ca](mailto:info@socialmediablab.ca)  
Twitter: @SMLabTO

Ryerson University | TED ROGERS SCHOOL OF MANAGEMENT

RECOMMENDED CITATION: Gruzd & Mai. (2020). The State of Social Media in Canada 2020. Ryerson University Social Media Lab. DOI: [10.5663/SP2/XIWB8EW](https://doi.org/10.5663/SP2/XIWB8EW)

CC BY 4.0 <https://creativecommons.org/licenses/by/4.0/>

The cover features a dark background with a grid of white code snippets. At the top left is the date 'November 2020' and the SocialMediaLab.ca logo at the top right. The title 'SOCIAL MEDIA PRIVACY IN CANADA 2020' is in large, bold, white letters. Below it, the authors 'By Anatoly Gruzd and Philip Mai' are listed. The central image shows a hand holding a smartphone with a large eye icon on the screen, surrounded by icons for a padlock, a Wi-Fi signal, and a USB plug. The subtitle 'A Canada-wide survey on privacy perceptions and expectations of Canadian social media users' is at the bottom of the image. Contact information for media inquiries is at the bottom right. Logos for Ryerson University and Ted Rogers School of Management are at the bottom left. A recommended citation and Creative Commons license are at the very bottom.

November 2020

SocialMediaLab.ca

## SOCIAL MEDIA PRIVACY IN CANADA 2020

By Anatoly Gruzd and Philip Mai

A Canada-wide survey on privacy perceptions and expectations of Canadian social media users

FOR MEDIA OR OTHER INQUIRES:  
Email: [info@socialmediablab.ca](mailto:info@socialmediablab.ca)  
Twitter: @SMLabTO

Ryerson University | TED ROGERS SCHOOL OF MANAGEMENT

RECOMMENDED CITATION: Gruzd & Mai. (2020). Social Media Privacy in Canada 2020. Version: 1.0. Ryerson University Social Media Lab. DOI: [10.5663/SP2/XIWB8EW](https://doi.org/10.5663/SP2/XIWB8EW)

CC BY 4.0 <https://creativecommons.org/licenses/by/4.0/>


# Data Visualization Dashboards



The Twitter Learning Analytics Dashboard is an information dashboard designed for instructors who use Twitter for their class discussions.



PoliDashboard is a visualization dashboard designed to help voters, members of the media, and campaign staffer to detect the presence and influence of bots in a political discussion on Twitter and to gauge the health of that conversation.




The COVID19MisInfo.org Portal is a rapid response project of the Ryerson Social Media Lab at Ted Rogers School of Management. The aim of this project is two-fold: (1) put a spotlight on COVID-19 related misinformation and (2) to provide Canadians with timely and actionable information.

# Social Media Data Collection & Analysis Tools



Netlytic is a text and social networks analyzer that can automatically summarize public conversations and discover communication networks from social media posts. It uses various APIs to collect public posts from Twitter, YouTube and RSS feed. It can also support the analysis of your own datasets via .text/.csv. files.



Commanalytic is a research tool for studying online communities and online discourse. Commanalytic can collect and analyze public data from various social media platforms including Reddit, Twitter, and Facebook/Instagram (via CrowdTangle). It uses advanced text and social network analysis techniques to automatically pinpoint toxic and anti-social interactions, identify influencers, map shared interests and the spread of misinformation, and detect signs of possible coordination among seemingly disparate actors.

## Additional Social Media Research Resources



A project focusing on studying practices behind and attitudes towards the collection, storage, use, reuse, analysis, publishing and preservation of social media data from the perspective of social media users.



A list of 70+ social media research tools curated by researchers at the Social Media Lab. Only tools that have been used in peer-reviewed academic studies are included.

# Research Tools & Methods

Data Analytics and Data Visualizations Software



## Commalytic is a research tool for studying online communities and online discourse.

Commalytic can collect and analyze public data from social media platforms. It uses advanced text and social network analysis techniques to automatically pinpoint toxic and anti-social interactions, identify influencers, map shared interests and the spread of misinformation, and detect signs of possible coordination among seemingly disparate actors.



### collect

Collect public data from Reddit, Twitter and CrowdTangle (Facebook/Instagram) or import your own dataset



### extract

Extract and visualize emojis, keywords and engagements as emoji/word clouds, reaction box plots and other summary stats about your dataset



### analyze

Analyze text and patterns of relations to detect toxic and anti-social interactions, identify influencers, map shared interests and the spread of misinformation, and detect signs of possible coordination among seemingly disparate actors



### visualize

Visualize two-mode semantic networks and who-replies-to-whom signed communication networks with each tie storing toxicity-related scores to uncover hidden patterns and correlations

## Communalytic Edu

- **Account Type:** Free
- **Account Caps:** ≤ 30K records shared across 3 datasets
- **Built-in Network Visualizer**
- **Collaboration Features**
- **CSV Import**
- **Reddit:** Live data collection for ≤ 7 consecutive days
- **Twitter Threads:** Public replies (≤ 10K) to any public tweet posted within the previous 7 days ([req. Twitter's Developers Account](#))
- **Twitter Academic Research Track:** Not supported
- **Crowdtangle Facebook/Instagram URL Search:** Public Facebook/Instagram posts (≤ 10k ) that shared the same URL ([req. CrowdTangle's Account](#))



[communalytic.com](https://www.communalytic.com)

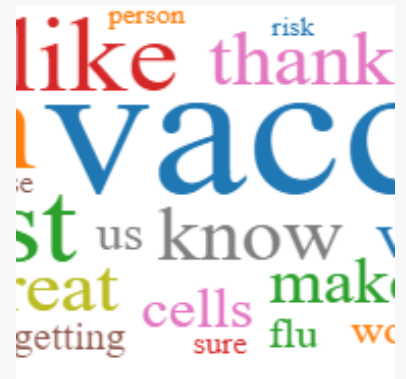
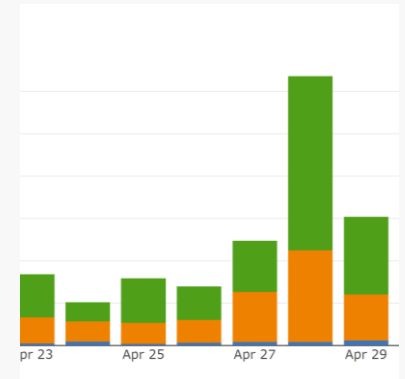
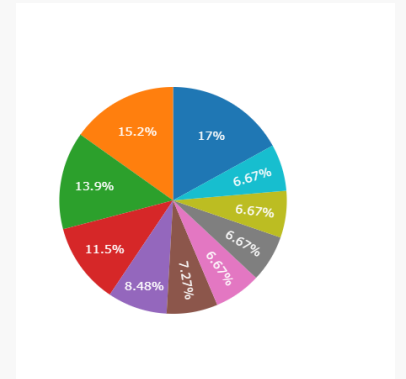
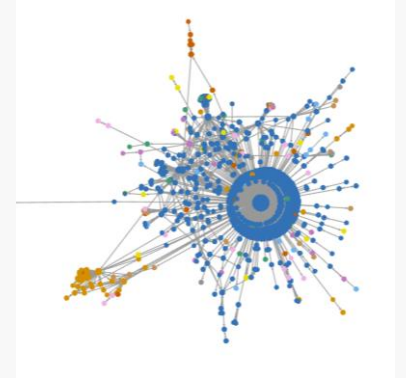
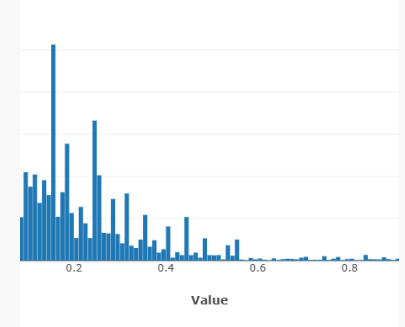
## Communalytic Pro

- **Account Type:** Community-supported
- **Account Caps:** ≤ 10M records shared across 50 datasets
- **Built-in Network Visualizer**
- **Collaboration Features**
- **CSV Import**
- **Reddit:** Live & historical data collection for ≤ 31 consecutive days
- **Twitter Threads:** Public replies (≤ 500K/mo.) to any public tweet posted within the previous 7 days ([req. Twitter's Developers Account.](#))
- **Twitter Academic Research Product Track:** ≤ 10M Tweets/month via 'full-archive (historical) search' ([req. Twitter's approval](#))
- **CrowdTangle Facebook/Instagram URL Search:** Public Facebook/Instagram posts that shared the same URL ([req. CrowdTangle Account](#))

# Examples of Research Inquiries enabled by Commanalytic :

1. What is the prevalence and the types of toxic and anti-social interactions observed in online discourse?
2. How does the presence of toxic interactions change the network structure of an online community over time?
3. And what is their potential impact on the conversation, the community, and its members?
4. How effective is a social media platform at combatting online toxicity and harassment?
5. How closely are platforms' content moderators (Human & AI) adhering to their stated public content moderation policy?
6. What types of accounts tend to share misinformation (e.g., links to questionable sources)?
7. Are these actions likely to be coordinated in some way? And what narratives are these accounts trying to inject into an online discussion?
8. Can/how social media platforms facilitate informal learning and knowledge exchange among community members?
9. Who are the opinion leaders or influential voices in an online discussion?
- 10.....

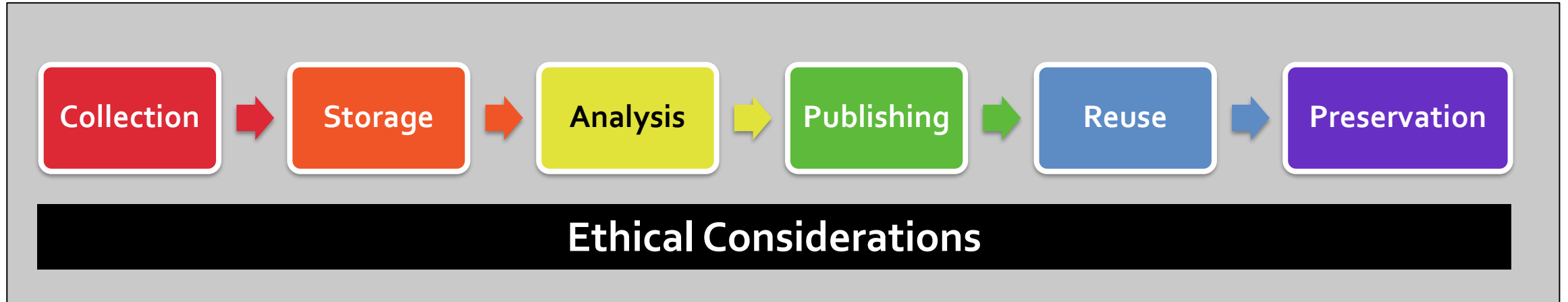
Toxicity Values Distribution Chart





# Research with Social Media Data: Ethical Considerations

Common stages of data-driven social media research:



- In Canada research that involves human participants is governed by the Tri-Council Policy Statement (**TCPS**) on Ethical Conduct for Research Involving Humans (<http://pre.ethics.gc.ca/>)
- Check with your university Research Ethics Board and funders to find out what the regulation are for your region

# Research with Social Media Data: Ethical Considerations

- Research that is **non-intrusive**, does not involve direct interaction between the researcher and individuals **through the Internet**, and where there is **no expectation of privacy does not require REB** (Research Ethics Boards) **review**.
- **REB** review is **not required** for **cyber-material**, such as documents, records, performances, online archival materials, or published third party interviews to which the public is given uncontrolled access on the Internet and for which there is no expectation of privacy.

(TCPS 2018 Ed,p.16)

# Research with Social Media Data: Ethical Considerations

- There are situations where REB (Research Ethics Board) review is required. There are digital sites in the public domain where there is a **reasonable expectation of privacy**. Privacy expectations may be outlined in the sites' terms of use.
- When accessing identifiable information in digital sites, such as online groups with restricted membership, the privacy expectation of contributors of these sites is much higher. Research involving information from these types of sources shall be submitted for REB review (Article 10.3).

**BUT!**

- (TCPS 2018 Ed, p.16)















# Social Media Privacy Survey 2020 (n=1500)

How comfortable would you be if one of the following **entities** accessed information about you or posted by you publicly on social media?

## Percentage of Respondents who are Uncomfortable with Third-Party's Social Media Data Access

(n=1500)

Third Party	Uncomfortable
Marketer	 56%
Political party	 52%
Journalist	 50%
Insurance company	 49%
Financial institution	 48%
Person you used to date	 48%
Government	 47%
Customs/border protection officer	 46%
Legal professional	 46%
Potential employer	 42%
Your current employer	 40%
Academic researcher	 36%

Source: Gruzd & Mai (2020b)

Note: All percentages are rounded to the nearest 1%. "Uncomfortable" includes both "Somewhat uncomfortable" and "Extremely uncomfortable" responses.

# Key Takeaways

- Social Media Research Ethics → not a “one size fits all” framework
- Privacy concerns varies across uses, users, and data types

(Gruzd & Hernández-García, 2018; Dubois, Gruzd, & Jacobson, 2018  
Jacobson, Gruzd & Hernandez-Garcia, 2019; Gruzd & Mai 2020)



## Further Reading

<https://aoir.org/ethics/>

## Ethics

AoIR has an ongoing commitment to ensuring that research on and about the Internet is conducted in an ethical and professional manner. The Ethics Working Committee, as composed of ethicists and researchers from various regions and countries, has produced three major reports to assist researchers in making ethical decisions in their research:

- 2019: *Internet Research: Ethical Guidelines 3.0* [PDF]
- 2012: *Ethical decision-making and Internet research 2.0: Recommendations from the AoIR ethics working committee* [PDF]
- 2012: *This chart provides a useful starting point for internet researchers to consider ethics.*
- 2002: *Ethical decision-making and Internet research: Recommendations from the AoIR ethics working committee* [PDF]

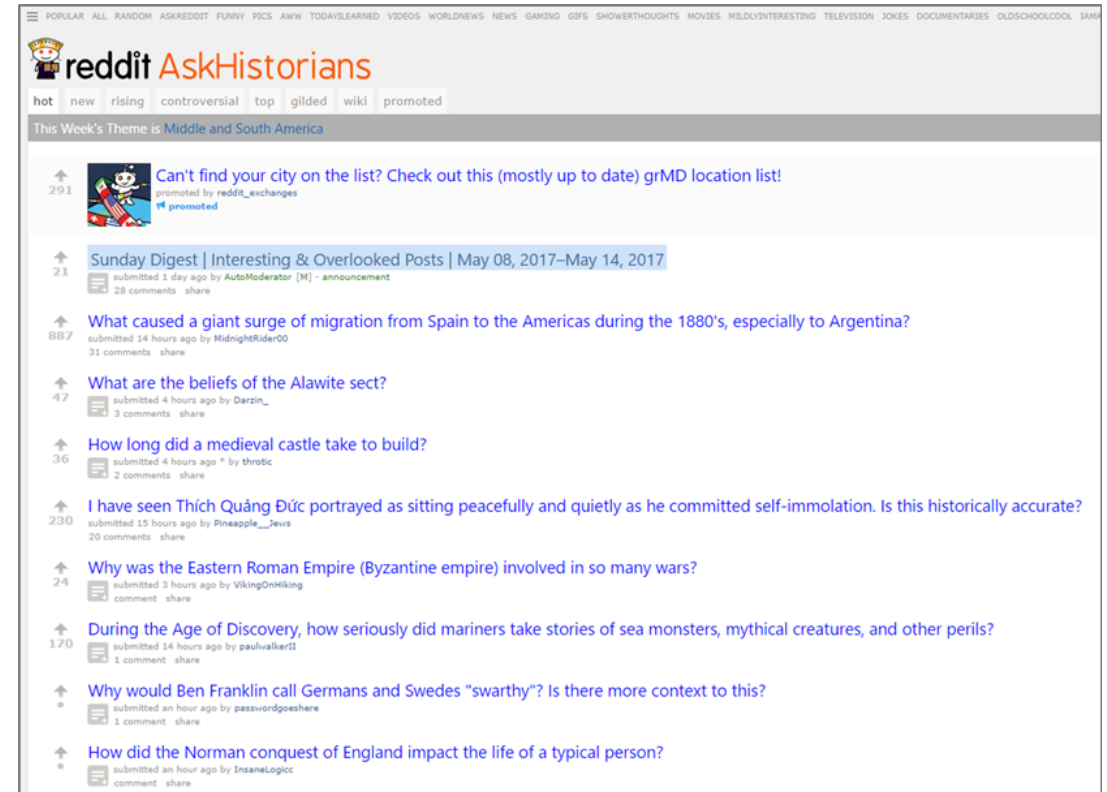
Researchers, students, ethicists, and related institutional bodies and academic organizations in the domain of Internet research may turn to these ethics document as a starting point for their inquiries and reflection. Just as these documents were immeasurably enriched by comments and contributions from AoIR members, we hope that readers will continue to call attention to issues and resources in Internet research ethics for debate and deliberation by the ethics working committee.



# Introducing Reddit

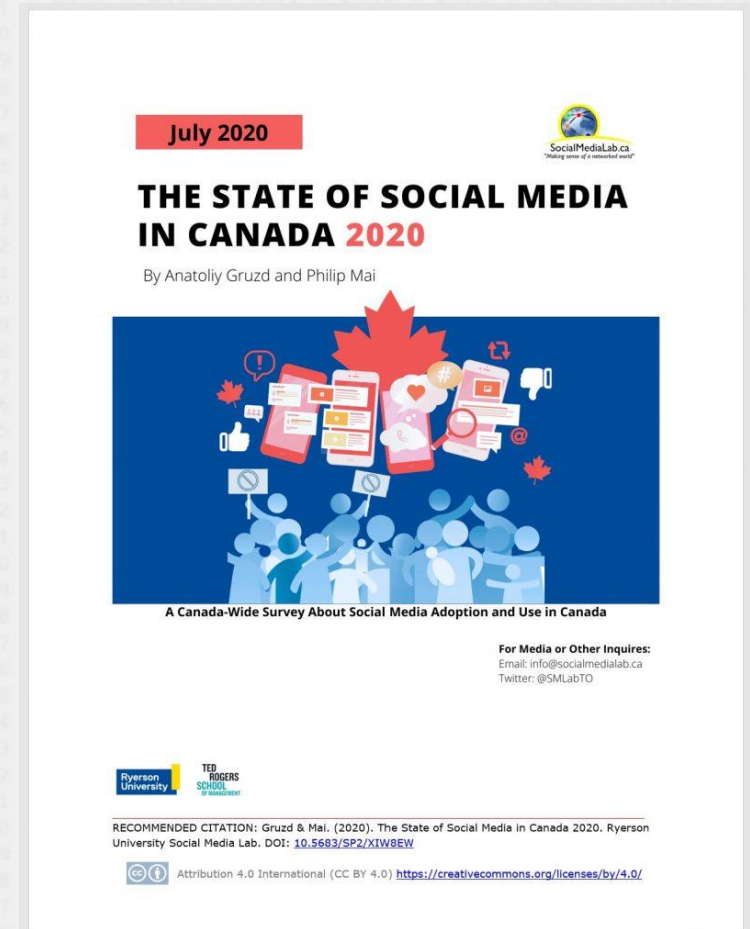
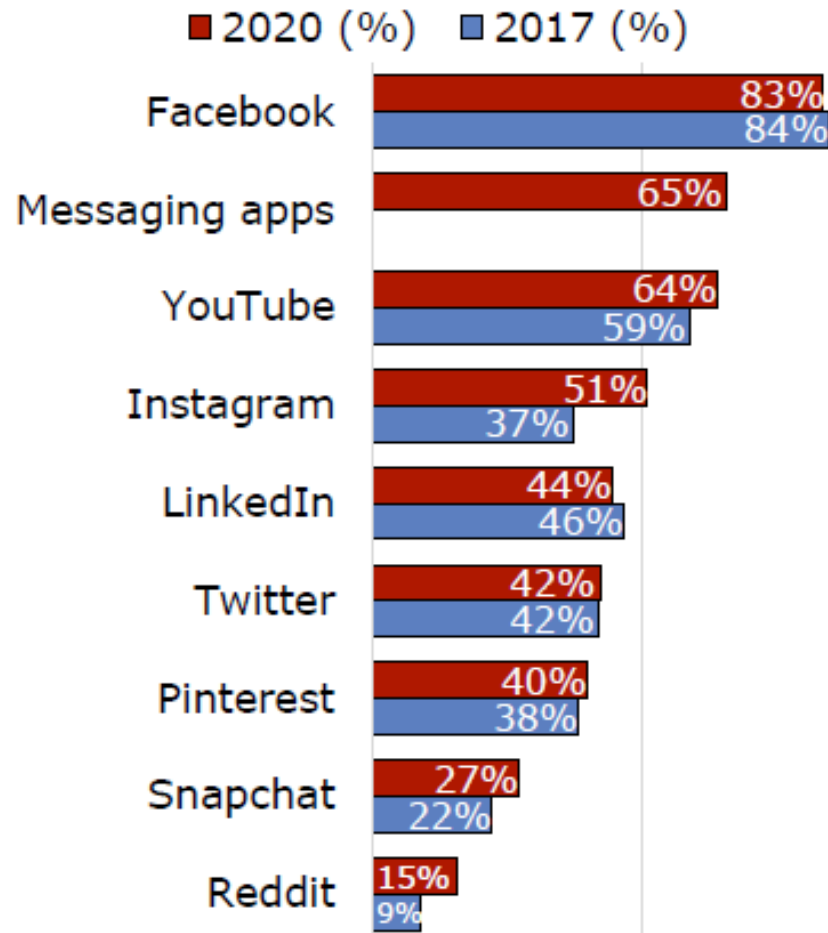


- “The front page of the Internet”
- Founded: 2005
- Reddit ranks 19<sup>th</sup> in terms of global traffic (most from the US, UK, India) – based on Alexa.com
- Diversity of niche communities called ‘subreddits’ maintain distinct subcultures, thematic interests and subject focus (>100k subreddits)
- Subreddit communities are user-generated and moderated by Reddit members
- Content curation (Upvoting/downvoting)



Sample Subreddit - ‘AskHistorians’

## Online Canadian adults with a Social Media Account (%, n=1500)




<https://doi.org/10.5683/SP2/XIW8EW>

# 15% of online Canadian adults have a Reddit account

	2017		2020	
	Total	MAU	Total	MAU
<b>All Online Adults</b>	<b>9%</b>	<b>8%</b>	<b>15%</b>	<b>14%</b>
Female	8	7	13	11
Male	10	9	18	17
18-24	28	26	41	38
25-34	16	15	30	28
35-44	8	8	18	15
45-55	7	6	9	8
55+	2	2	3	3






**July 2020**



## THE STATE OF SOCIAL MEDIA IN CANADA 2020

By Anatolij Gruzd and Philip Mai



A Canada-Wide Survey About Social Media Adoption and Use in Canada

For Media or Other Inquiries:  
Email: [info@socialmedialab.ca](mailto:info@socialmedialab.ca)  
Twitter: [@SMLabTO](https://twitter.com/SMLabTO)

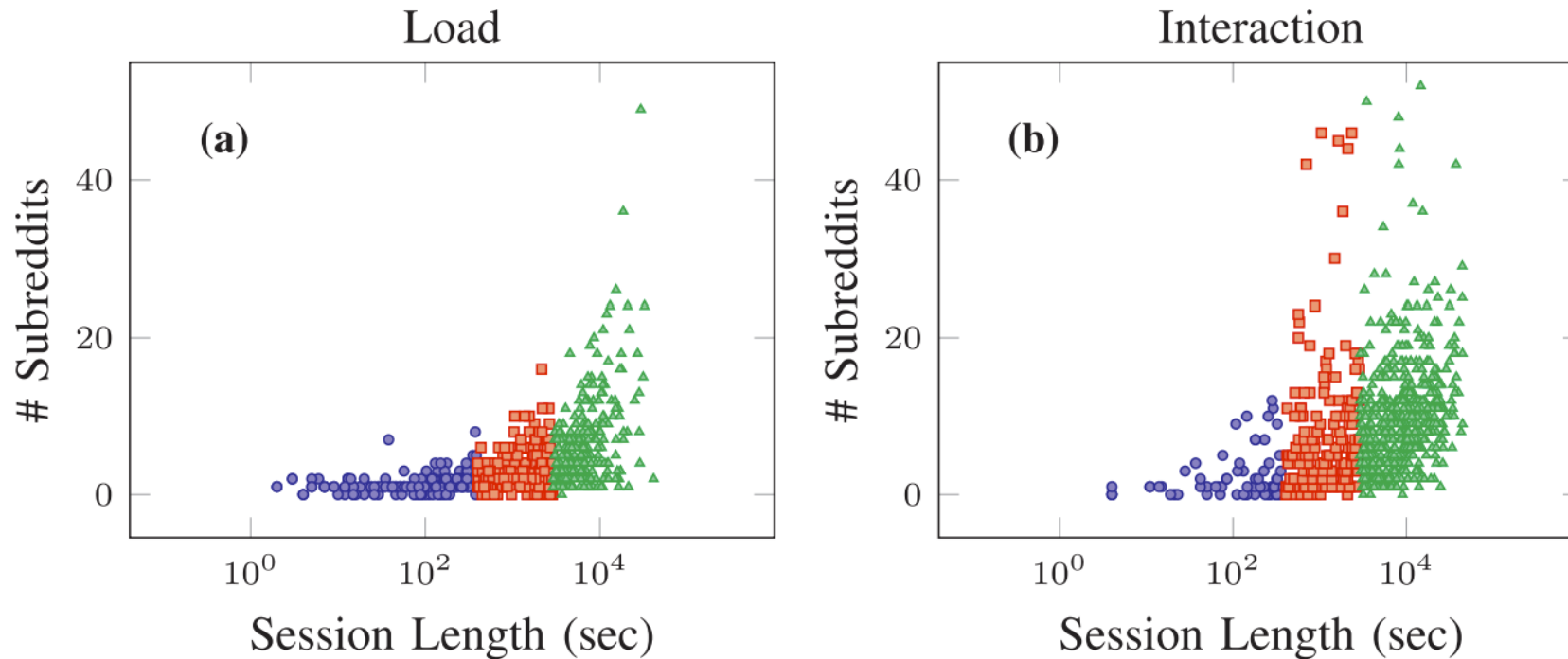



RECOMMENDED CITATION: Gruzd & Mai. (2020). The State of Social Media in Canada 2020. Ryerson University Social Media Lab. DOI: [10.5683/SP2/XIW8EW](https://doi.org/10.5683/SP2/XIW8EW)

Attribution 4.0 International (CC BY 4.0) <https://creativecommons.org/licenses/by/4.0/>

<https://doi.org/10.5683/SP2/XIW8EW>

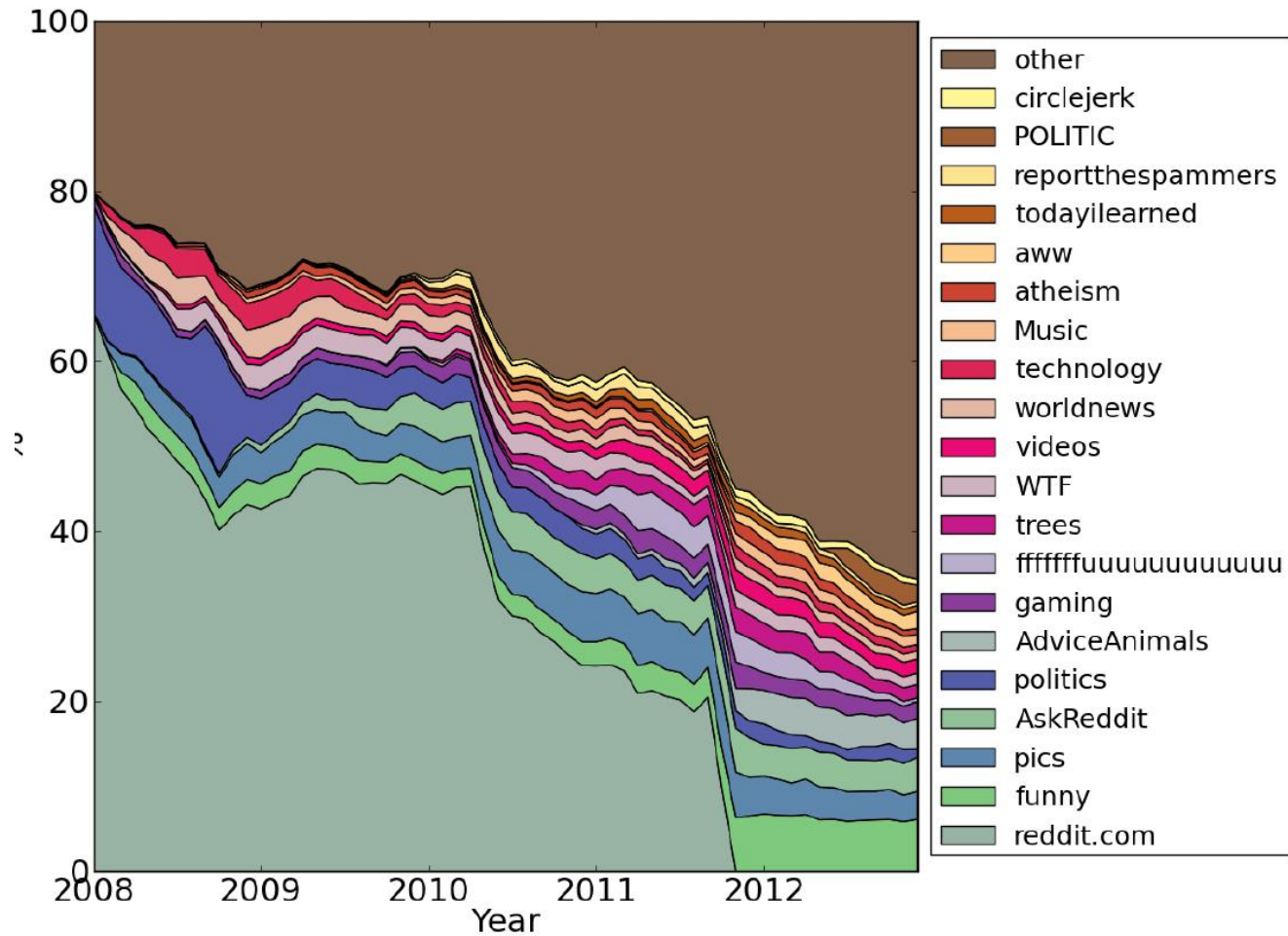
# Reddit users are highly engaged across multiple groups



**Fig. 9.** Content variety. Distinct subreddits (a) loaded and (b) interacted with during a session by session length. Despite being highly skewed, the mean-average number of subreddits loaded is 1.4, 2.5, and 4.8 for short-, medium-, and long-browsing users, respectively, and even lower 1.0, 1.3, and 1.5, respectively, for subreddits users actively interacted with.

Glenski, M., Pennycuff, C., & Weninger, T. (2017). Consumers and Curators: Browsing and Voting Patterns on Reddit. *IEEE Transactions on Computational Social Systems*, 4(4), 196–206. <https://doi.org/10.1109/TCSS.2017.2742242>

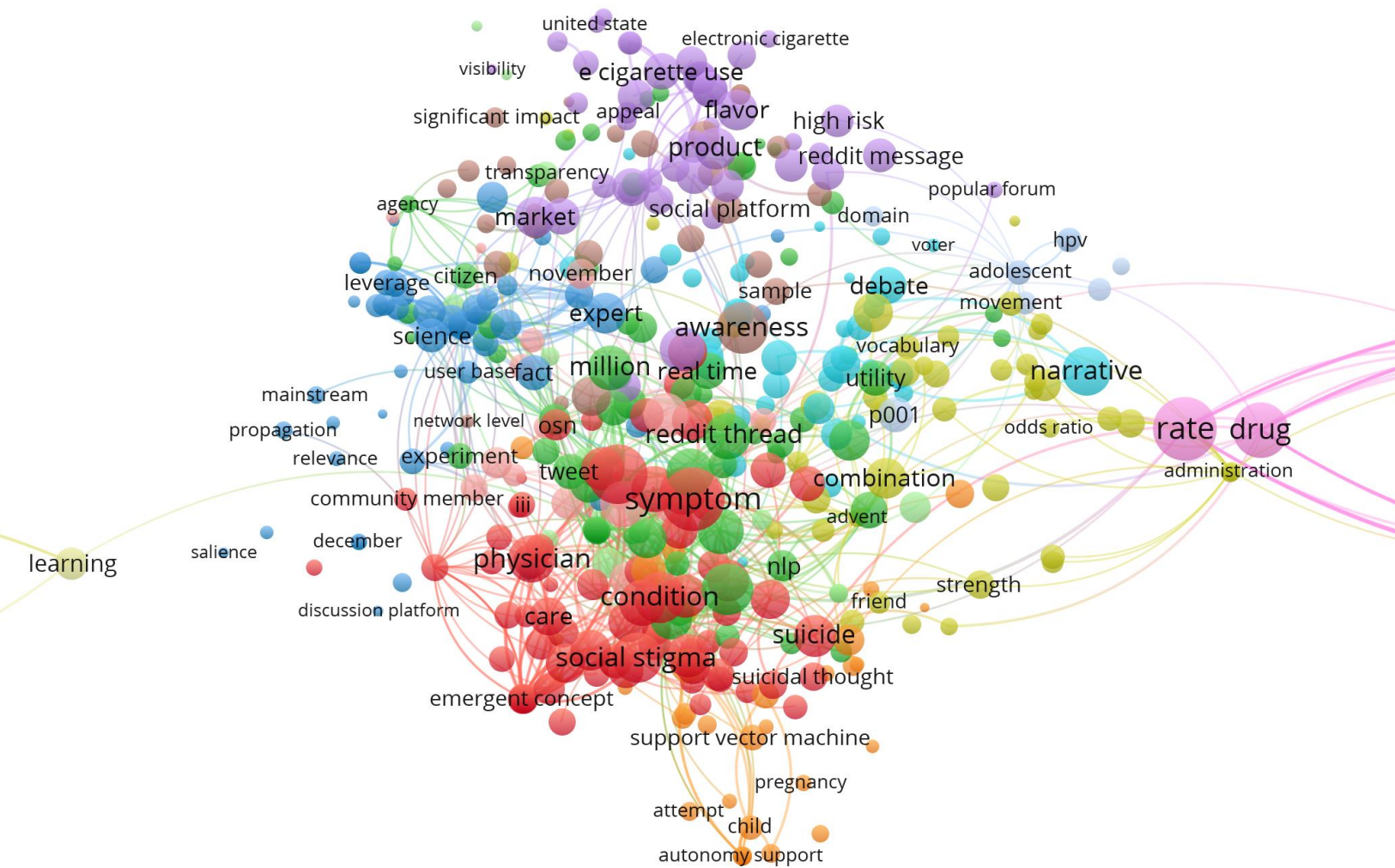
# Changing Topical Focus of Subreddits Over Time (2008-2012)



(a) Subreddits

Singer, P., Flöck, F., Meinhart, C., Zeitfogel, E., & Strohmaier, M. (2014). Evolution of reddit: From the front page of the internet to a self-referential community? *Proceedings of the 23rd International Conference on World Wide Web*, 517–522.  
<https://doi.org/10.1145/2567948.2576943>





# Research with Reddit data

*Frequently used words in abstracts  
of Reddit-related academic papers  
(n=194)*

*Source: Web of Science*

# Sample Study with Reddit Data



ELSEVIER

## Drug and Alcohol Dependence

Volume 194, 1 January 2019, Pages 358-362



Full length article

## A content analysis of JUUL discussions on social media: Using Reddit to understand patterns and perceptions of JUUL use

Emma I. Brett <sup>a</sup>, Elise M. Stevens <sup>b</sup>, Theodore L. Wagener <sup>b, c</sup>, Eleanor L.S. Leavens <sup>a, b</sup>, Taylor L. Morgan <sup>d</sup>, Whitney D. Cotton <sup>b</sup>, Emily T. Hébert <sup>b</sup>  

<https://doi.org/10.1016/j.drugalcdep.2018.10.014>







### Highlights

- The study used the **manual content analysis** of public posts on Reddit.
- Posts relevant to both **JUUL** and **youth** were included in the analysis (364 total posts and comments).
- Most frequent reasons for using JUUL included: the **popularity of JUUL** (34.2%), followed by using it to **quit smoking** (23.3%), and to **feel a buzz** (20.2%).
- Age restrictions were the most common barrier to use.

# Sample Study with Reddit Data



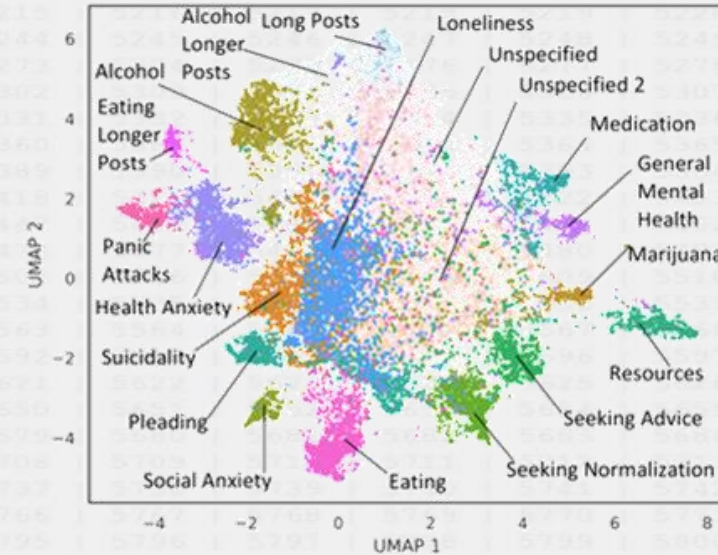
## Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19: Observational Study

Daniel M Low <sup>1, 2, 3</sup> ; Laurie Rumker <sup>4, 5, 6, 7</sup> ; Tanya Talkar <sup>1, 8</sup> ; John Torous <sup>9</sup> ; Guillermo Cecchi <sup>10</sup> ; Satrajit S Ghosh <sup>1, 2, 11</sup> 

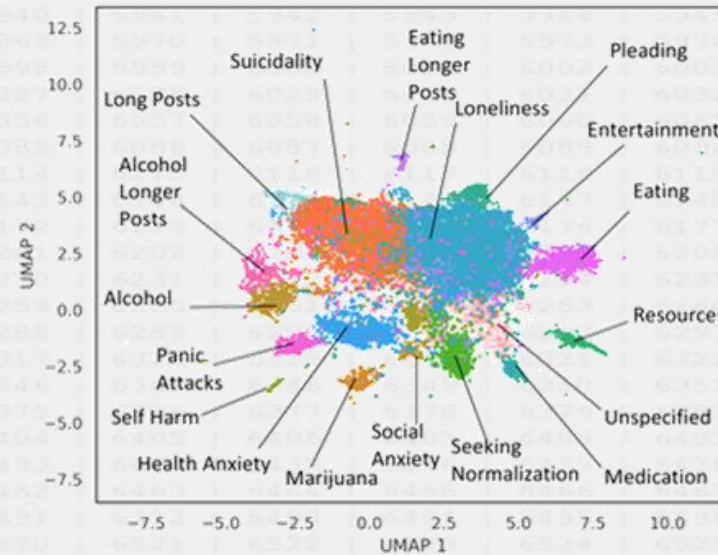
<https://doi.org/10.2196/22635>

## Highlights

- The study applied different NLP techniques to analyze a large sample of reddit posts before and during the COVID-19 pandemic.
- The authors demonstrated “how specific mental health problems manifest in language, identified at-risk users, and revealed the distribution of concerns across Reddit, which could help provide better resources to its millions of users.”



Unsupervised clustering of pre-pandemic (year 2019) posts from 15 mental health subreddits



Unsupervised clustering of mid-pandemic posts



# Sample Study with Reddit Data

RESEARCH-ARTICLE

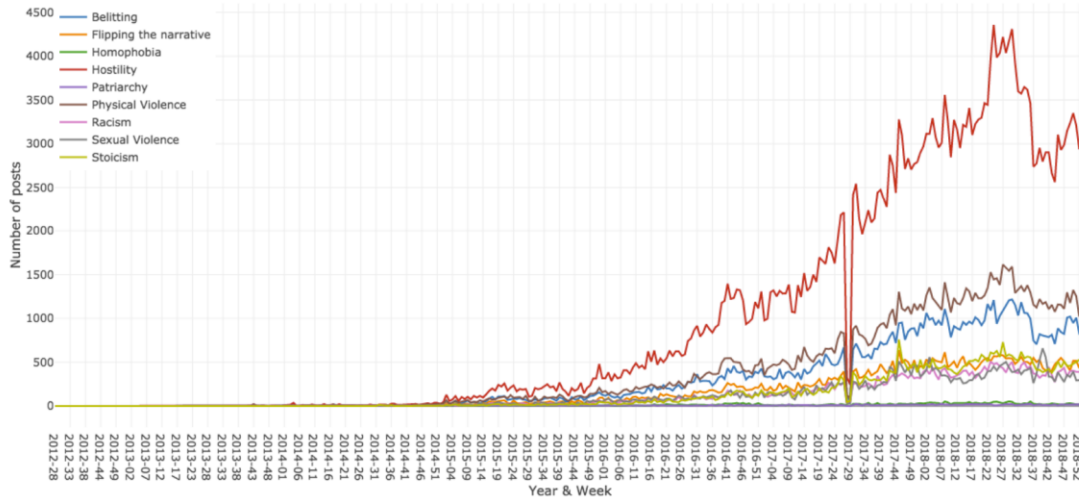
## Exploring Misogyny across the Manosphere in Reddit



Authors: Tracie Farrell, Miriam Fernandez, Jakub Novotny, Harith Alani [Authors Info & Affiliations](#)

Publication: WebSci '19: Proceedings of the 10th ACM Conference on Web Science • June 2019 • Pages 87–

Community MGTOW

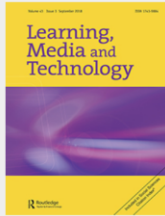


<https://doi.org/10.1145/3292522.3326045>

## Highlights

- The study examined the flow of extreme language across seven subreddits with openly misogynistic discourse, such as MGTOW (“Men Going Their Own Way”), r/badwomensanatomy, and r/IncelsWithoutHate, to study if and how misogynistic ideas spread within and across these communities.
- The study was based on 9 lexicons, including: Harassment Corpus, Violence verbs, Hatebase, Profanity words, Incel specific words.
- The results showed increasing patterns on misogynistic content and users as well as violent attitudes from 2011 to 2018.

# Sample Study with Reddit Data



Learning, Media and Technology >

Volume 43, 2018 - Issue 3

## Learning in the wild: coding for learning and practice on Reddit

Caroline Haythornthwaite, Priya Kumar , Anatoliy Gruzd , Sarah Gilbert,

Marc Esteve del Valle & Drew Paulin

Pages 219-235 | Received 19 Mar 2018, Accepted 05 Jul 2018, Published online: 19 Jul 2018

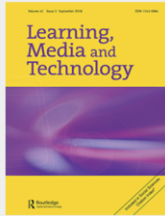
<https://doi.org/10.1080/17439884.2018.1498356>

### Highlights

- The paper studied learning practices found in 'Ask' subreddits **AskScience**, **Ask\_Politics**, **AskAcademia**, and **AskHistorians** to develop a coding schema for **informal learning**.

Subreddit Community	# Moderators	# Subscribers	# Posts in 2016	Coded Sample (1%)
Askscience	433	14M	223K	2,235
Ask_politics	8	26K	46K	464
AskHistorians	40	600K	122K	1,227
AskAcademia	3	32K	26.9K	269

# Sample Study with Reddit Data



Learning, Media and Technology >  
Volume 43, 2018 - Issue 3

## Learning in the wild: coding for learning and practice on Reddit

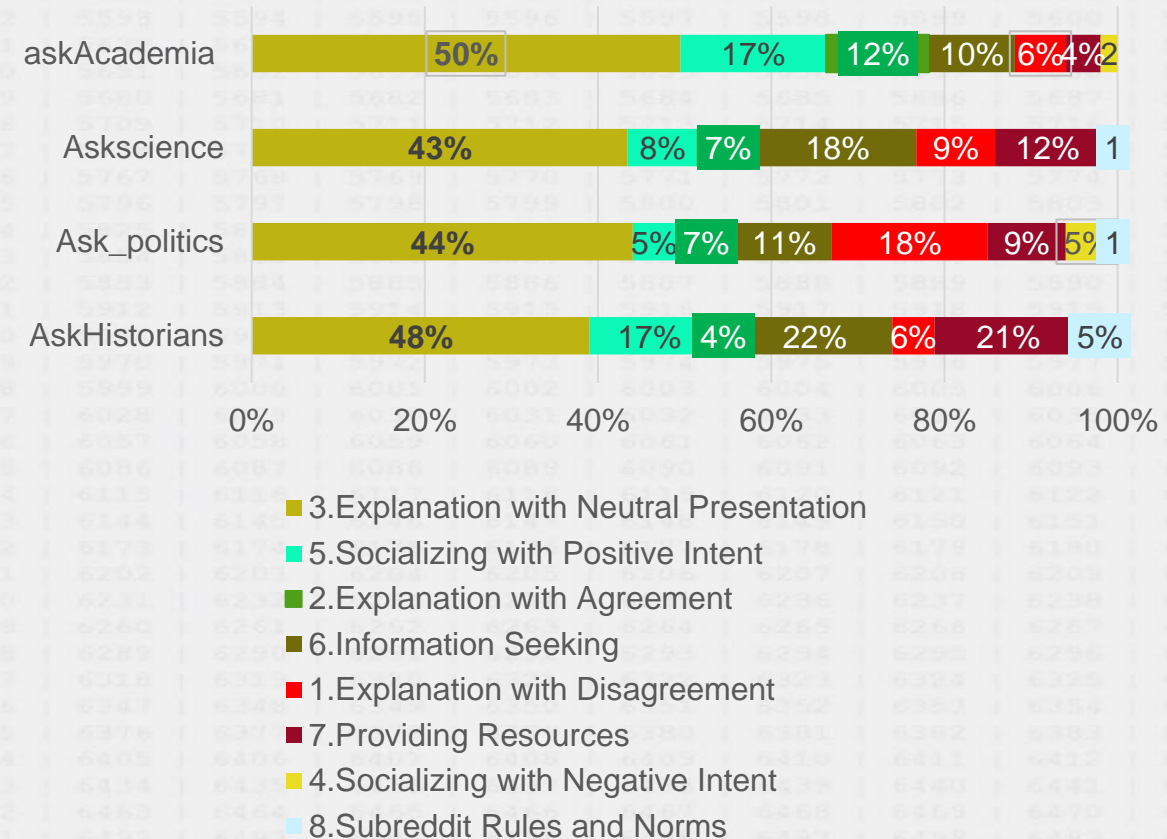
Caroline Haythornthwaite, Priya Kumar , Anatoliy Gruzd , Sarah Gilbert, Marc Esteve del Valle & Drew Paulin

Pages 219-235 | Received 19 Mar 2018, Accepted 05 Jul 2018, Published online: 19 Jul 2018

<https://doi.org/10.1080/17439884.2018.1498356>

### Highlights

- The paper studied learning practices found in 'Ask' subreddits **AskScience**, **Ask\_Politics**, **AskAcademia**, and **AskHistorians** to develop a coding schema for **informal learning**.





# Sample Study with Reddit Data

Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality

## Detecting Anxiety through Reddit

Judy Hanwen Shen | Frank Rudzicz

<u>anxiety subreddits</u>	<u>control subreddits</u>	
r/anxiety	r/askscience	r/relationships
r/healthanxiety	r/writingprompts	r/teaching
r/socialanxiety	r/writing	r/parenting
r/panicparty	r/atheism	r/christianity
	r/showerthoughts	r/jokes
	r/lifeprotips	r/writing
	r/personalfinance	r/talesfromretail
	r/theoryofreddit	r/talesfromtechsupport
	r/randomkindness	r/talesfromcallcenters
	r/books	r/fitness
	r/askdocs	r/frugal
	r/legaladvice	r/youshouldknow
	r/nostupidquestions	

Table 1: Subreddits used for data collection.

<https://www.aclweb.org/anthology/W17-3107/>

## Highlights

- The study applied N-gram language modeling, vector embeddings, topic analysis, and emotional norms to generate features that accurately classify posts related to binary levels of anxiety (accuracy of 98%).

Category	most frequent n-grams
<b>Anxiety unigrams</b>	<i>anxiety, myself, anyone, social, panic, friends, feeling, having, anxious, else, talk, bad, thought, better, felt</i>
<b>Control unigrams</b>	<i>our; call, us, edit, old, tell, phone, use, give, same, customer, post, money, let, reddit</i>
<b>Anxiety bigrams</b>	<i>(my anxiety), (social anxiety), (my life), (anxiety and), (anxiety i), (anyone else), (talk to), (panic attacks), (panic attack), (where i)</i>
<b>Control bigrams</b>	<i>(we are), (from the), (we have), (she was), (he was), (thank you), (that the), (and he), (and she), (what is)</i>

Table 5: Frequent  $N$ -grams ( $N=1, 2$ ) in each class.

# Sample Study with Reddit Data

SHORT-PAPER

## Identifying social roles in reddit using network structure



Authors:  [Cody Buntain](#),  [Jennifer Golbeck](#) [Authors Info & Affiliations](#)

### Publication:

WWW '14 Companion: Proceedings of the 23rd International Conference on World Wide Web • April 2014 • Pages 615–620 • <https://doi.org/10.1145/2567948.2579231>

<https://doi.org/10.1145/2567948.2579231>

## Highlights

- The study explored user posting behavior on reddit (based on the top 100 submissions in July 2013 + the top ~200 comments per submission).
- The authors showed how users taking the “answer-person” role can be identified automatically based solely on user interactions (SNA).

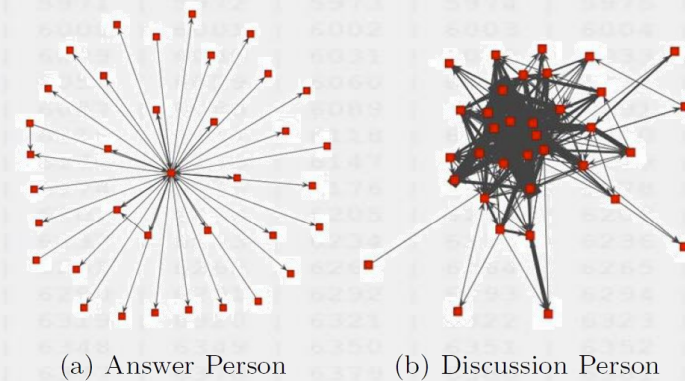
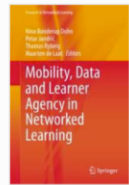


Figure 1: Exemplary Role Signatures

# Sample Study with Reddit Data



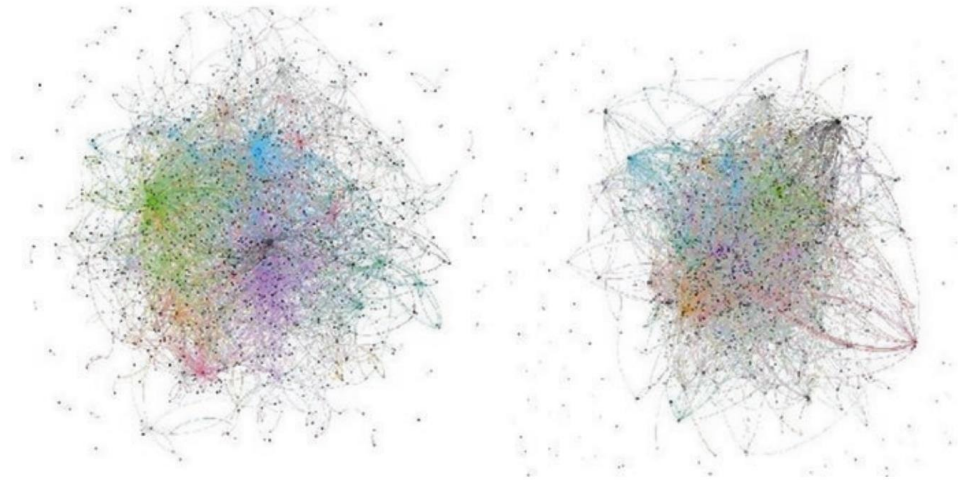
[Mobility, Data and Learner Agency in Networked Learning](#) pp 51-68 | [Cite as](#)

## Learning in the Wild: Understanding Networked Ties in Reddit

Authors

[Authors and affiliations](#)

Marc Esteve Del Valle , Anatoliy Gruzd, Priya Kumar, Sarah Gilbert

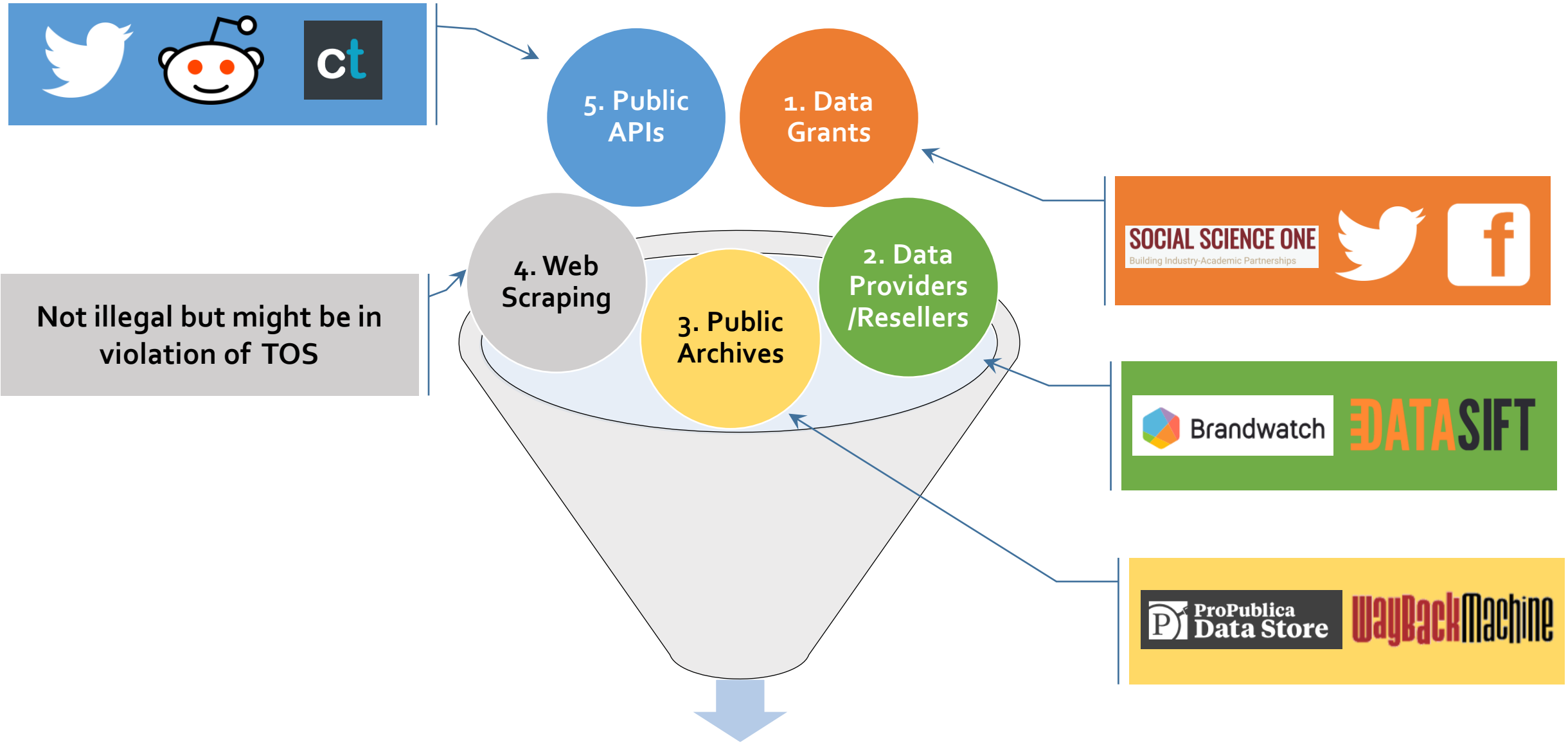


**Fig. 4.1** Communication networks among Redditors in AskStatistics (on the left) and AskSocialScience (on the right)

[https://doi.org/10.1007/978-3-030-36911-8\\_4](https://doi.org/10.1007/978-3-030-36911-8_4)

## Highlights

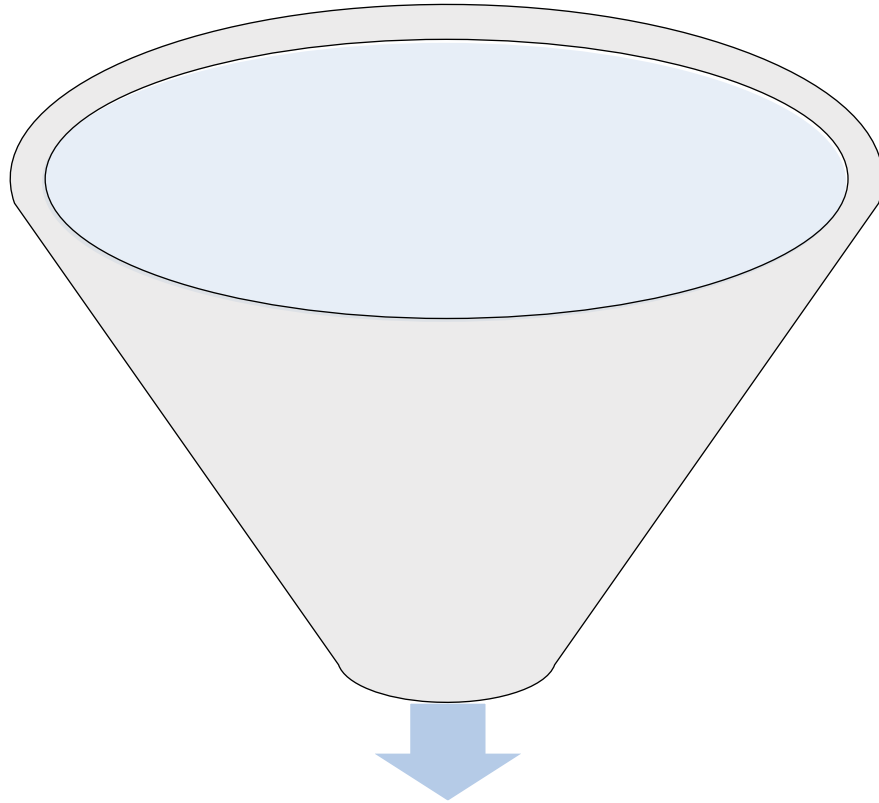
- The study used the Social Network Analysis (SNA) to examine two subreddits: **AskStatistics** and **AskSocialScience**.
- The results show that **informal learning** processes occurring in these two communities are determined by the **reciprocal** and **transitive** nature of communicative ties among their members.
- The paper also found that **moderators** play a key role in fostering interactions in these two communities.



# Social Media Data Sources



5. Public APIs



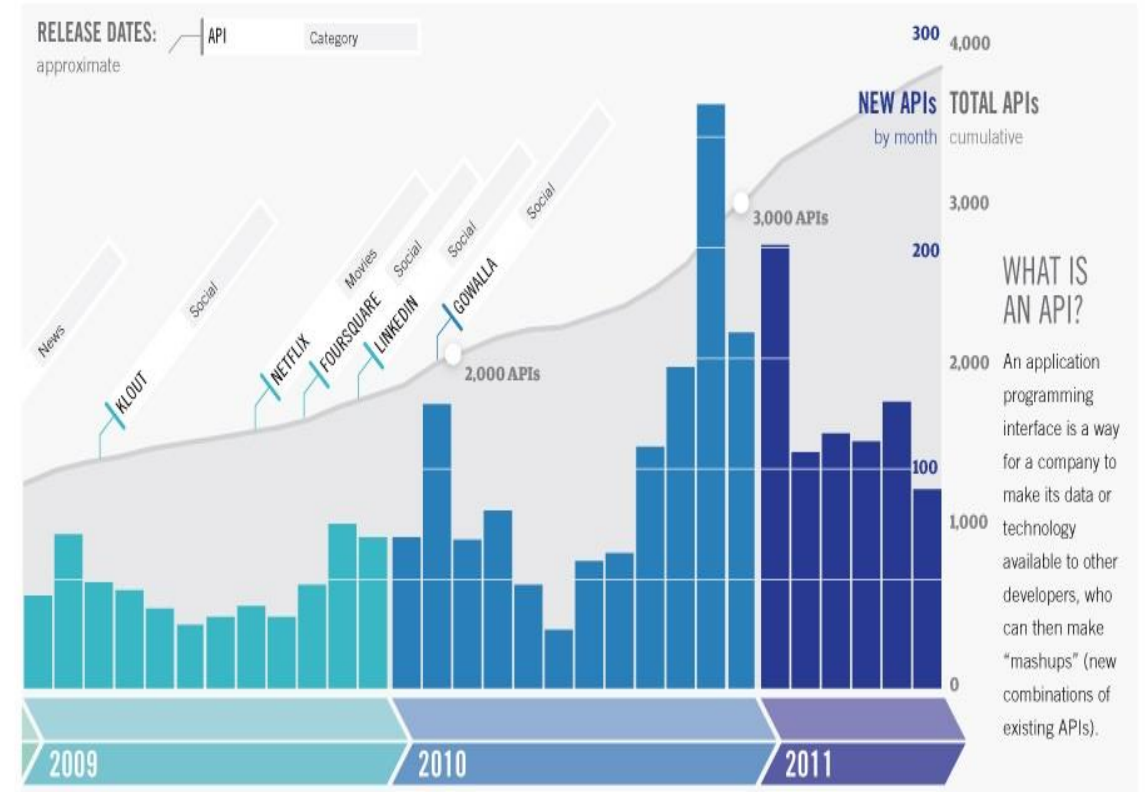
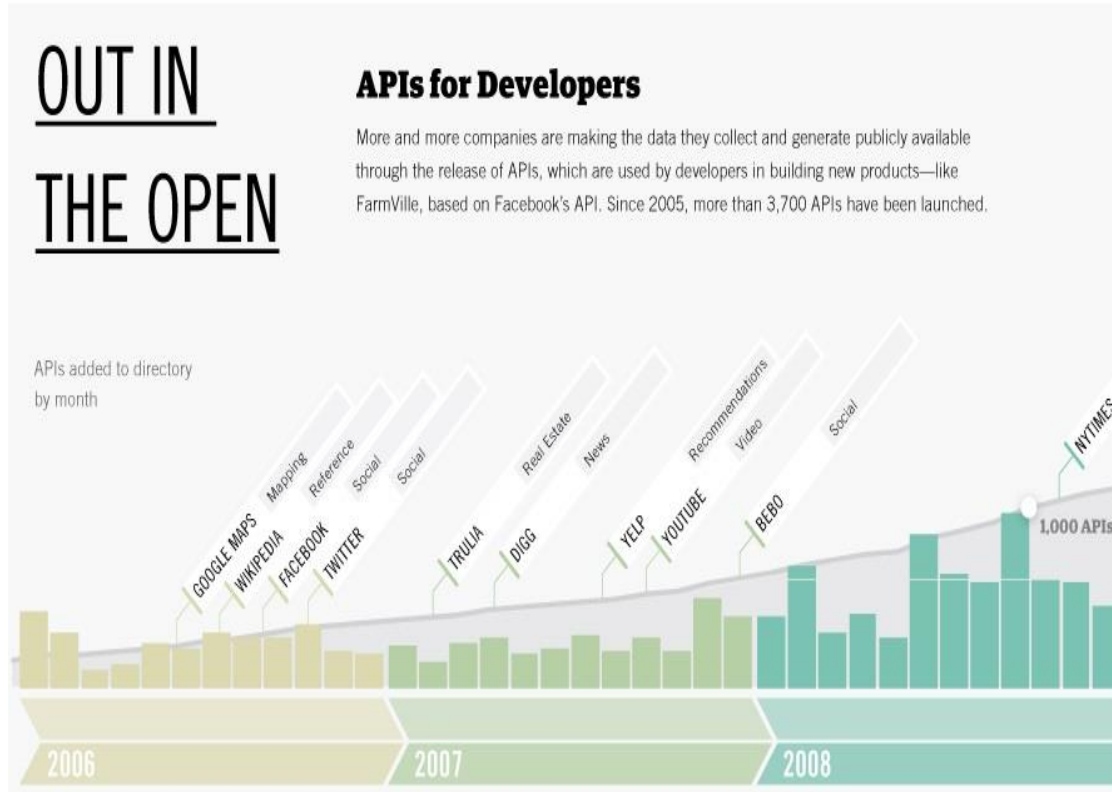
**Social Media Data Sources**

## **Application Programming/ Programmable Interfaces (APIs)**

– a data exchange protocol that platforms (like Reddit and Twitter) creates and supports to make their data/services available to third-party developers who can then combine it with existing APIs to create new “mashup” solutions.

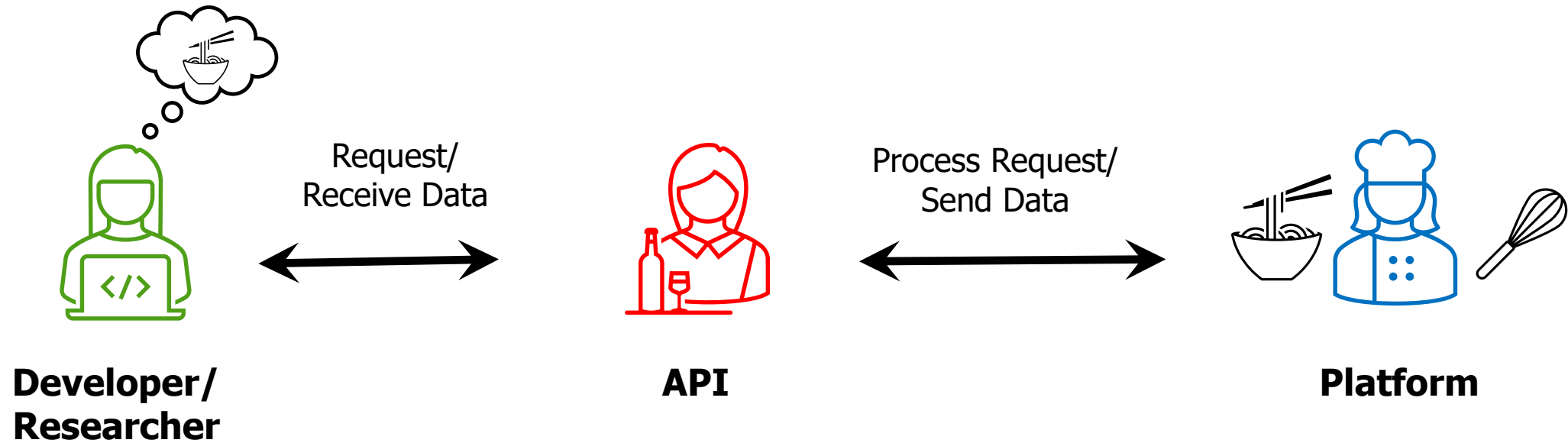


# Application Programming/Programmable Interfaces (APIs)



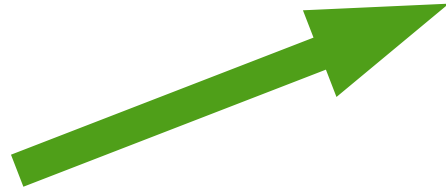
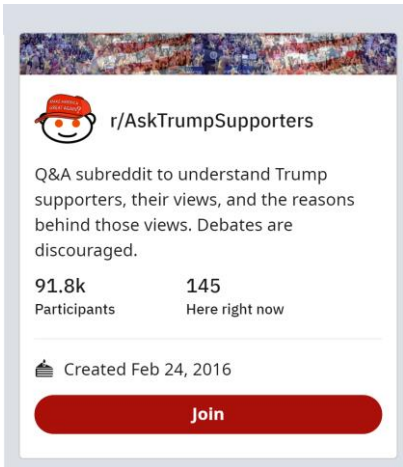
Source: <https://www.programmableweb.com/api-university>

# Application Programming/Programmable Interfaces (APIs)



# Reddit API: Machine-readable representation of a submission

<https://www.reddit.com/dev/api/>



## Flair

**Trump Supporters** are those who support President Trump and either voted for him, or would vote for him if the election were held today. Trump Supporters are the primary draw for this subreddit. As such, they are the only ones who can make top level comments with no requirements to ask [clarifying questions](#).

**Nonsupporters** are those who do not support President Trump or actively oppose his presidency. Non-Supporters are not permitted to make top level comments and their comments must be clarifying in nature, [with some exceptions](#).

**Undecided** are those who do not actively support President Trump, but do not actively oppose him. They may agree with him on certain policy topics. They might consider voting for him if the election were held today. Undecided are not permitted to make top level comments and their comments must also be clarifying in nature, [with some exceptions](#).

Posted by u/HonestlyKidding **Nonsupporter** 1 year ago

40  
What do you think of the Republican Party?

Other

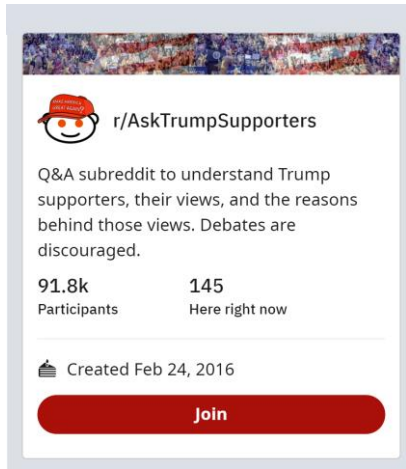
I get the impression from a lot of comments here that not all Trump supporters are GOP supporters.

121 Comments Share Save Hide Report

87% Upvoted

# Reddit API: Machine-readable representation of a submission

<https://www.reddit.com/dev/api/>



Posted by u/HonestlyKidding **Nonsupporter** 1 year ago

## What do you think of the Republican Party?

Other

I get the impression from a lot of comments here that not all Trump supporters are GOP supporters.

121 Comments    Share    Save    Hide    Report

87% Upvoted

```
{
  "_id" : "dx19pr",
  "created_utc" : 1573873250,
  "author" : "HonestlyKidding",
  "body" : "I get the impression from a lot of comm",
  "comment_karma" : 12630,
  "count" : 770,
  "downs" : 0,
  "flair" : "Nonsupporter",
  "likes" : null,
  "link_karma" : 6239,
  "name" : "t3_dx19pr",
  "post_flair" : "Other",
  "praw" : true,
  "score" : 40,
  "selftext" : "I get the impression from a lot of",
  "strings" : "What do you think of the Republican",
  "subreddit" : "AskTrumpSupporters",
  "subreddit_id" : "t5_3cr3o",
  "title" : "What do you think of the Republican Pa",
  "type" : "Submission",
  "ups" : 40,
  "toxicity" : "DONE"
}
```



# Reddit API: Machine-readable representation of a reply



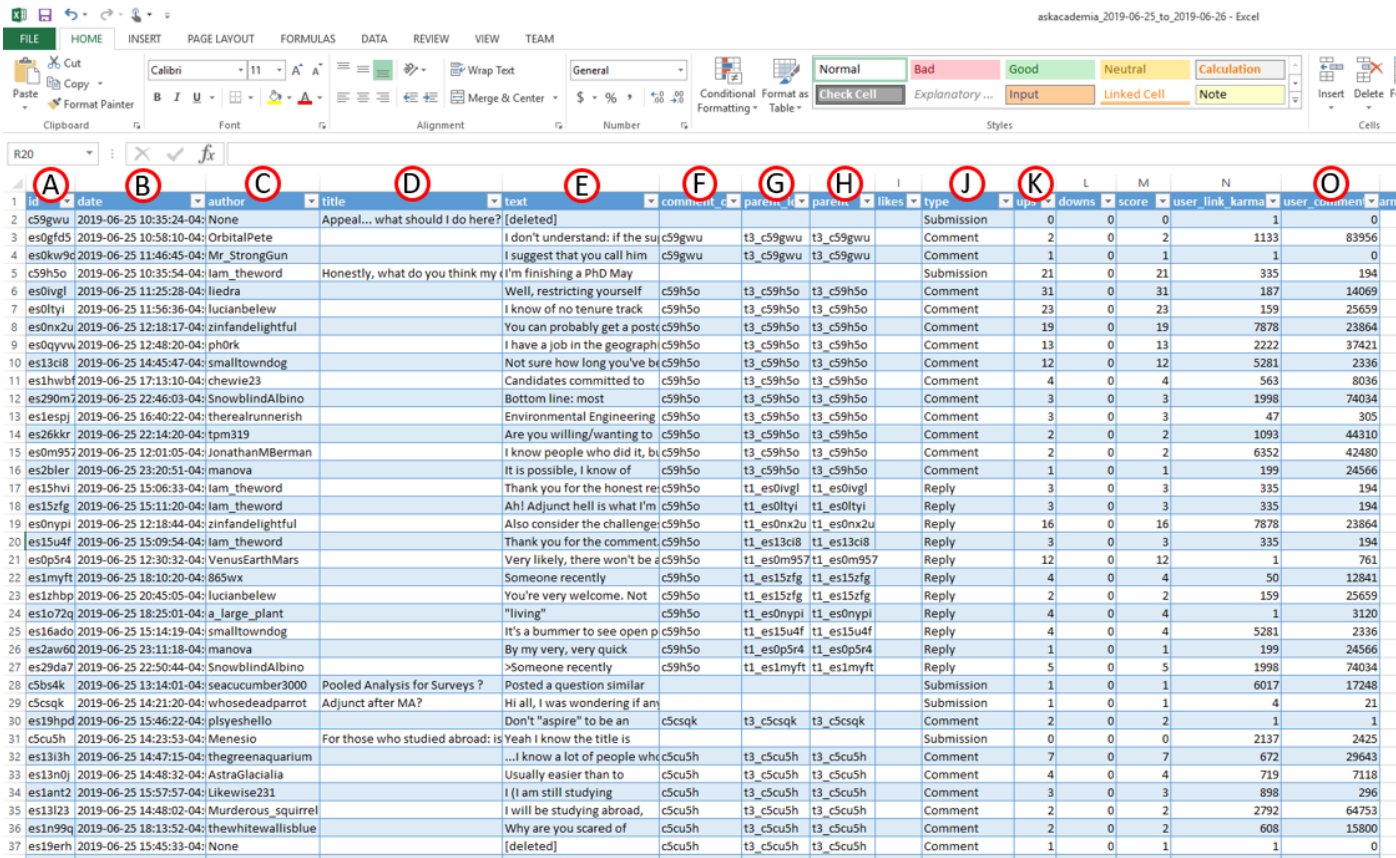
ConsciousFractals 2 months ago

Some people use valid talking points to justify hate. It's sad, because it shuts down any real discussion and just leads to animosity all around.

↑ -4 ↓ Reply Give Award Share Report Save

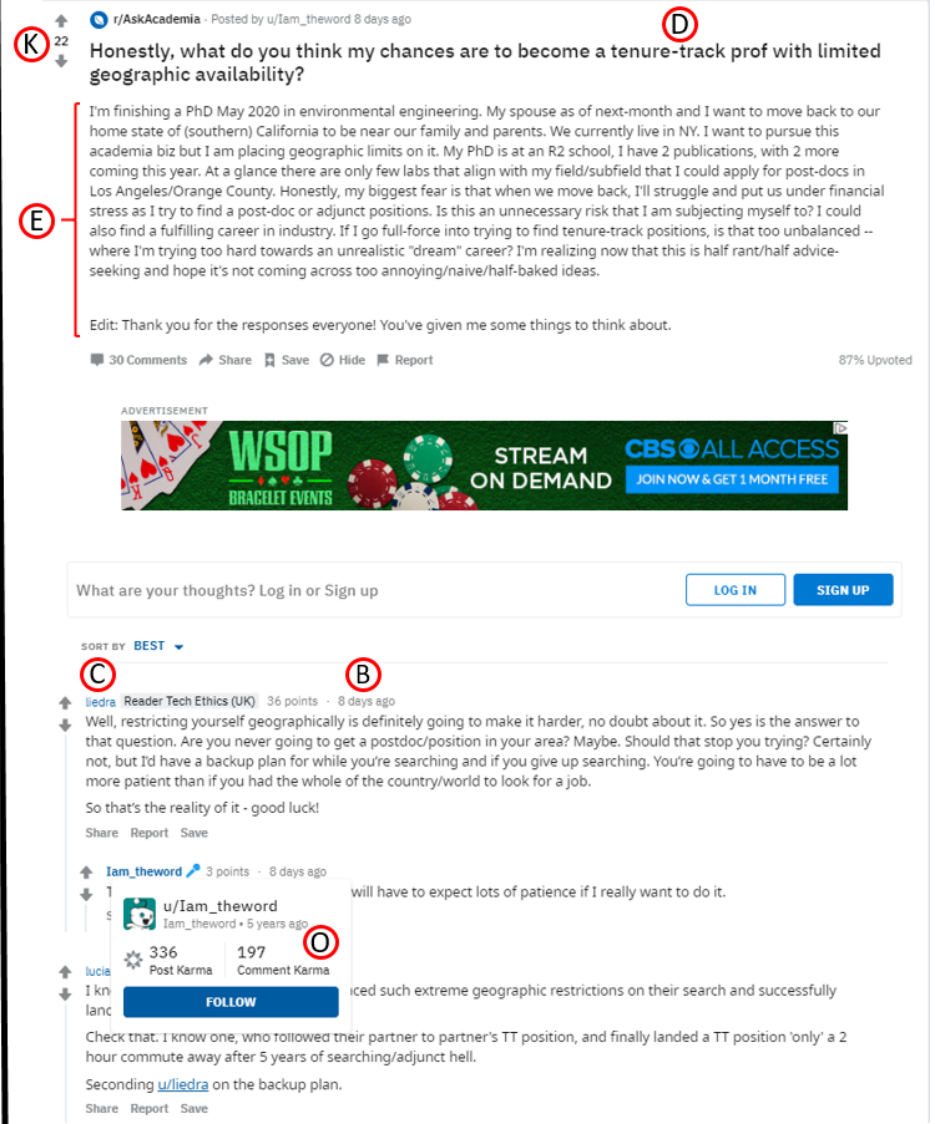
```
{
  "_id" : "gnhusnj",
  "praw" : true,
  "subreddit_id" : "t5_2qh4r",
  "author" : "ConsciousFractals",
  "subreddit" : "conspiracy",
  "created_utc" : 1613358180,
  "ups" : -2,
  "name" : "t1_gnhusnj",
  "body" : "Some people use valid talking points to justify hate. It's sad, because it shuts down any real discussion and just leads to animosity all around.",
  "score" : -2,
  "downs" : 0,
  "likes" : null,
  "parent" : "t1_gnhcebf",
  "link_id" : "t1_gnhcebf",
  "url" : "https://www.reddit.com/r/conspiracy/comments/ljye8m/i_pissed_off_the_ministry_of_truth_guys_reddit/gnhusnj/",
  "permalink" : "/r/conspiracy/comments/ljye8m/i_pissed_off_the_ministry_of_truth_guys_reddit/gnhusnj/",
  "upvote_ratio" : null,
  "type" : "Reply",
  "count" : 10085,
  "flair" : "None",
  "submission_id" : "ljye8m".
```

# Exported Dataset



- A:** The id of the submission/Comment/reply
- B:** Date of when the message was posted
- C:** Author of the post
- D:** title of the submission/subject line
- E:** The body of the post
- F, G & H:** The id of the parent post if it's a Comment/reply
- J:** Post type
- K:** Number of upvotes minus the number of downvotes on a post
- O:** Karma score for the post

# Subreddit

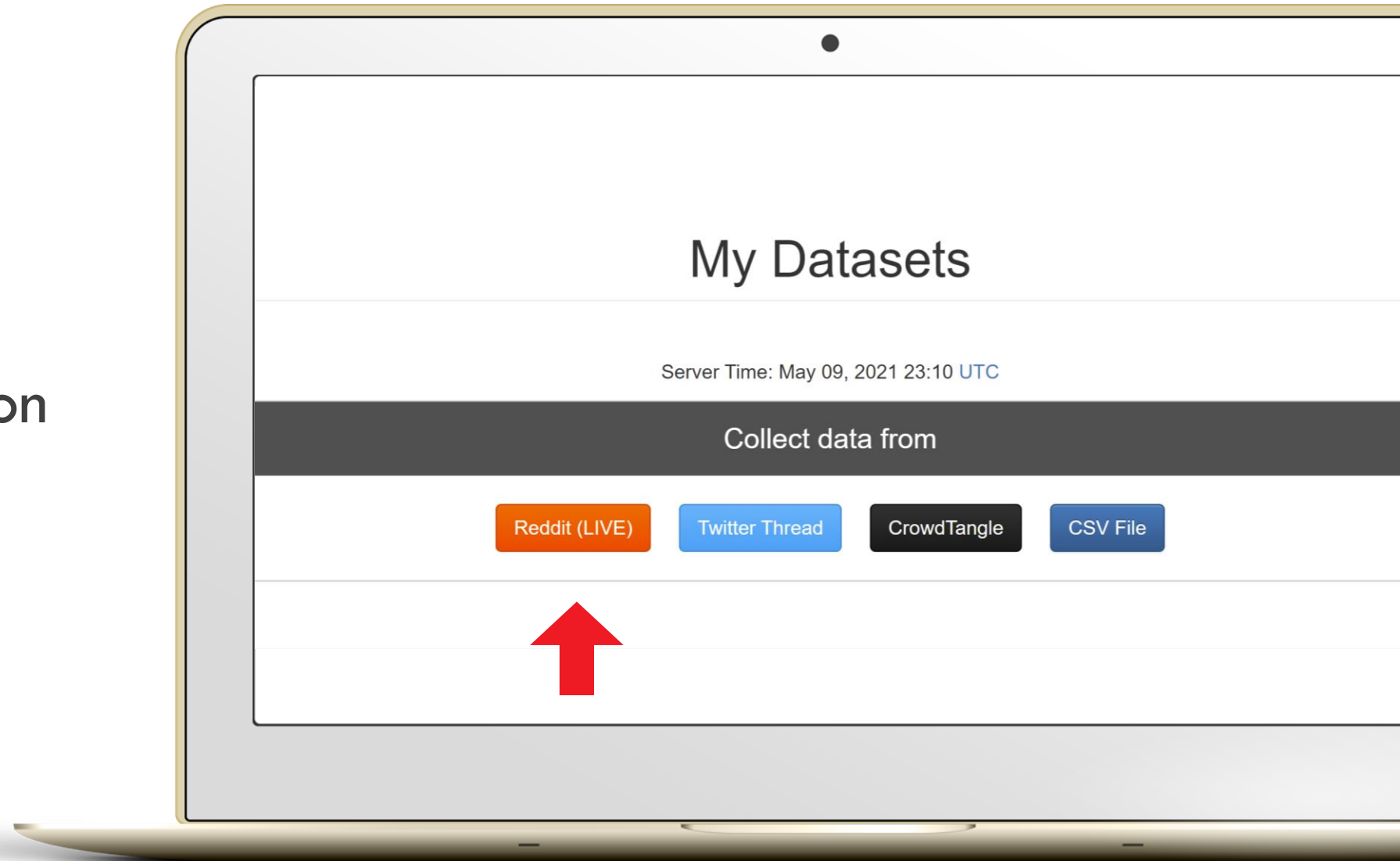


**(K)** **(D)** **(E)**

**(C)** **(B)**

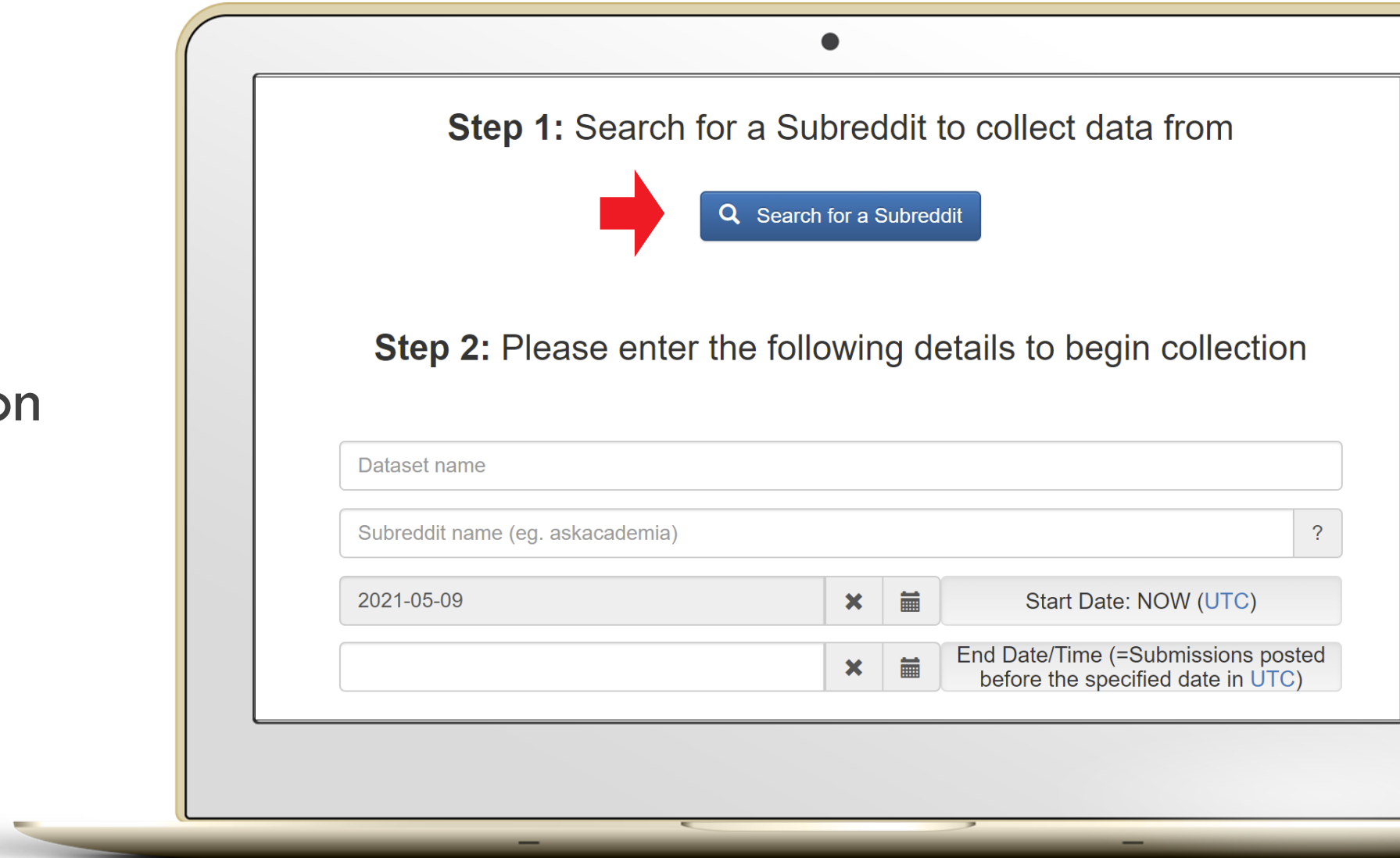
## Hands-On Part 1:

- 1) Find a relevant subreddit, and
- 2) Start data collection




## Hands-On Part 1:

- 1) Find a relevant **subreddit**, and
- 2) Start data collection



**Step 1:** Search for a Subreddit to collect data from



**Step 2:** Please enter the following details to begin collection

Dataset name

Subreddit name (eg. askacademia)  ?

2021-05-09   Start Date: NOW (UTC)

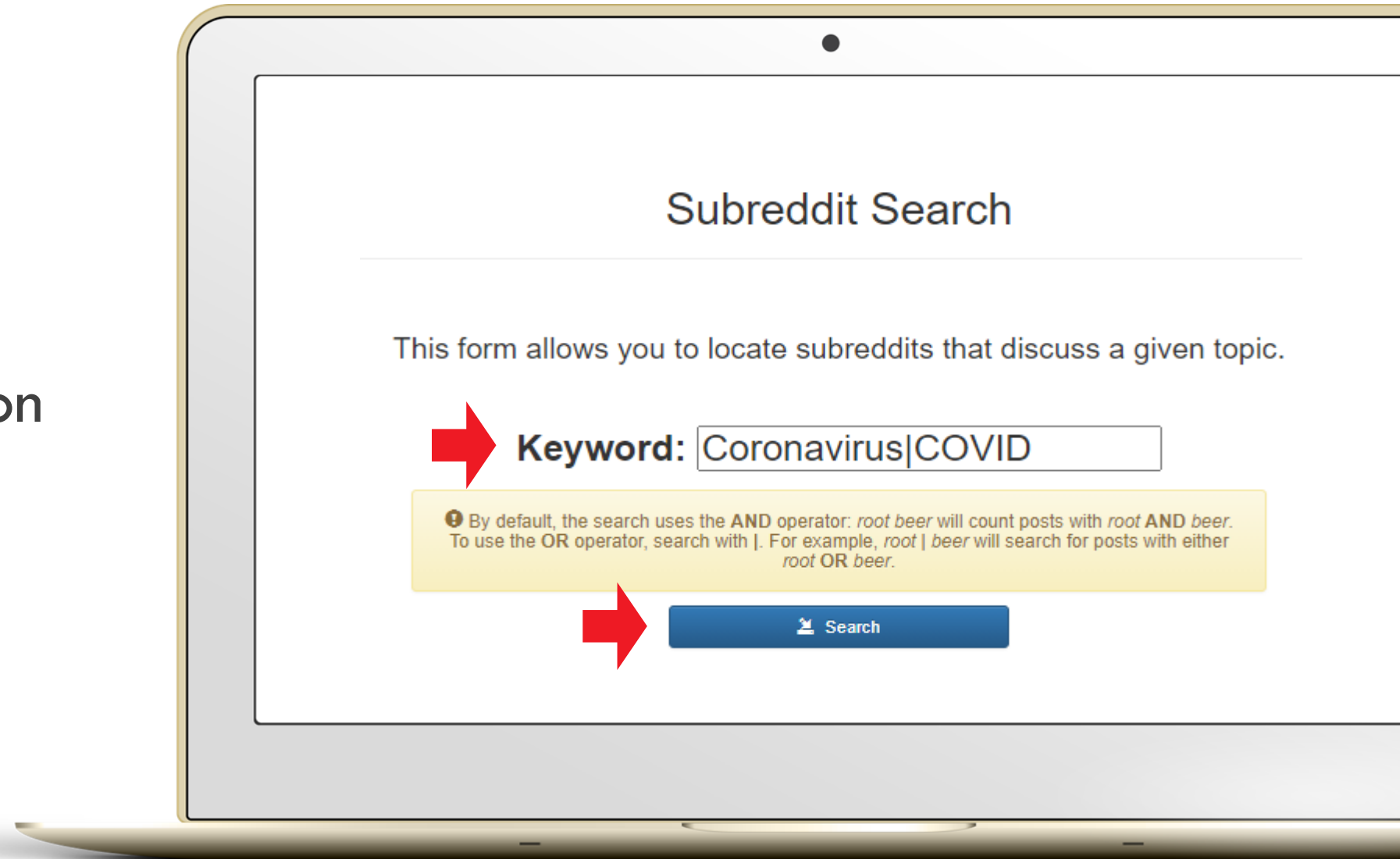
End Date/Time (=Submissions posted before the specified date in UTC)



## Hands-On Part 1:

- 1) Find a relevant **subreddit**, and
- 2) Start data collection

Enter one or more search keywords to find relevant subreddits




## Hands-On Part 1:

1) Find a relevant subreddit, and

2) Start data collection

Review a sample of subreddits with at least 100 comments made within the last 7 days corresponding to your search criteria.



The screenshot shows the 'Subreddit Search Results' page. It includes a title, a descriptive paragraph, two informational yellow boxes, and a table with three columns: 'Start collection', 'Subreddit', and 'Sample comments corresponding to the search criteria'. The table lists two subreddits: r/ADHD\_partners and r/IRS. A red arrow points to the bottom of the laptop screen.

### Subreddit Search Results

Below is a sample of subreddits with at least 100 comments made within the last 7 days

Communalystic Edu version does not collect posts from subreddits with 10 million or more subscribers like r/askReddit. If you need to collect data from subreddits with 10 million or more subscribers, please use [Communalystic Pro](#).

To find out how many members there are in a given subreddit, click on a subreddit name and look under "About Community".

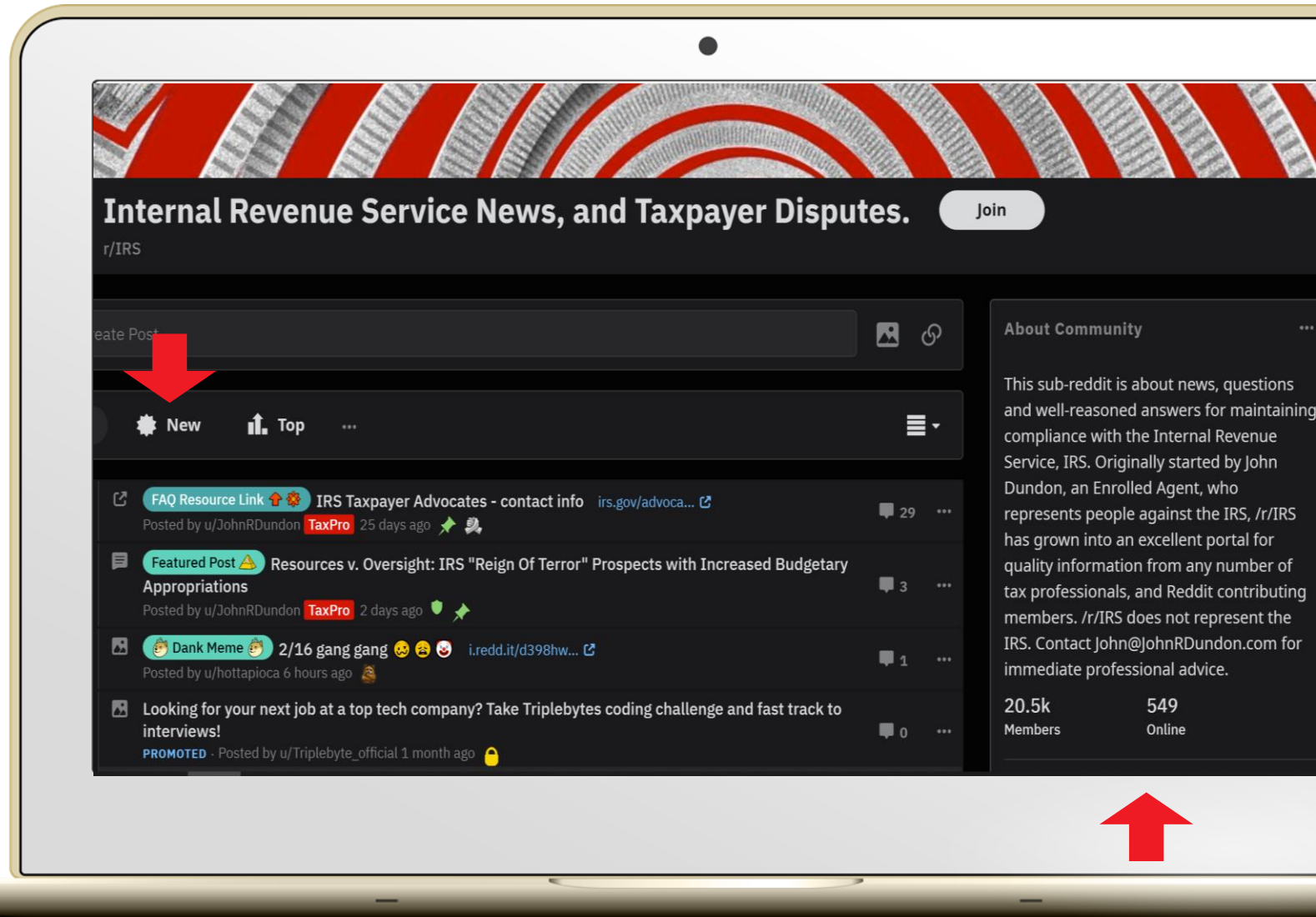
Start collection	Subreddit	Sample comments corresponding to the search criteria
<a href="#">↓ Start Collection on r/ADHD_partners</a>	<a href="#">r/ADHD_partners</a>	I do have great friends and they are one of the top joys in my life but due to the pandemic, I've lost a lot of connection with them. We text often but everyone is feeling "zoom fatigue". My major hobby was training mixed martial arts but also due to <b>covid</b> my gym has only been able to do virtual classes, and it's been very difficult finding the energy to keep up with training because I don't have a person to drill or spar with. So there are major factors at play that's making this situation especially hard for me to cope with right now. I definitely look forward to being able to meet with friends. Where I live, we're not even allowed to meet socially distanced outside. Thank you for the reminder to take care of myself. It's the sad truth I push aside a lot of my needs to accommodate for others and it's something I need to work on.
<a href="#">↓ Start Collection on r/IRS</a>	<a href="#">r/IRS</a>	Accepted 3/10. We've called 3 times, the first time we were told it's just a <b>covid</b> backup, second our return was being held for manual review but no errors noted, today...lo and behold...were in the errors dept and while there aren't any errors noted

## Hands-On Part 1:

1) Find a relevant subreddit, and

2) Start data collection

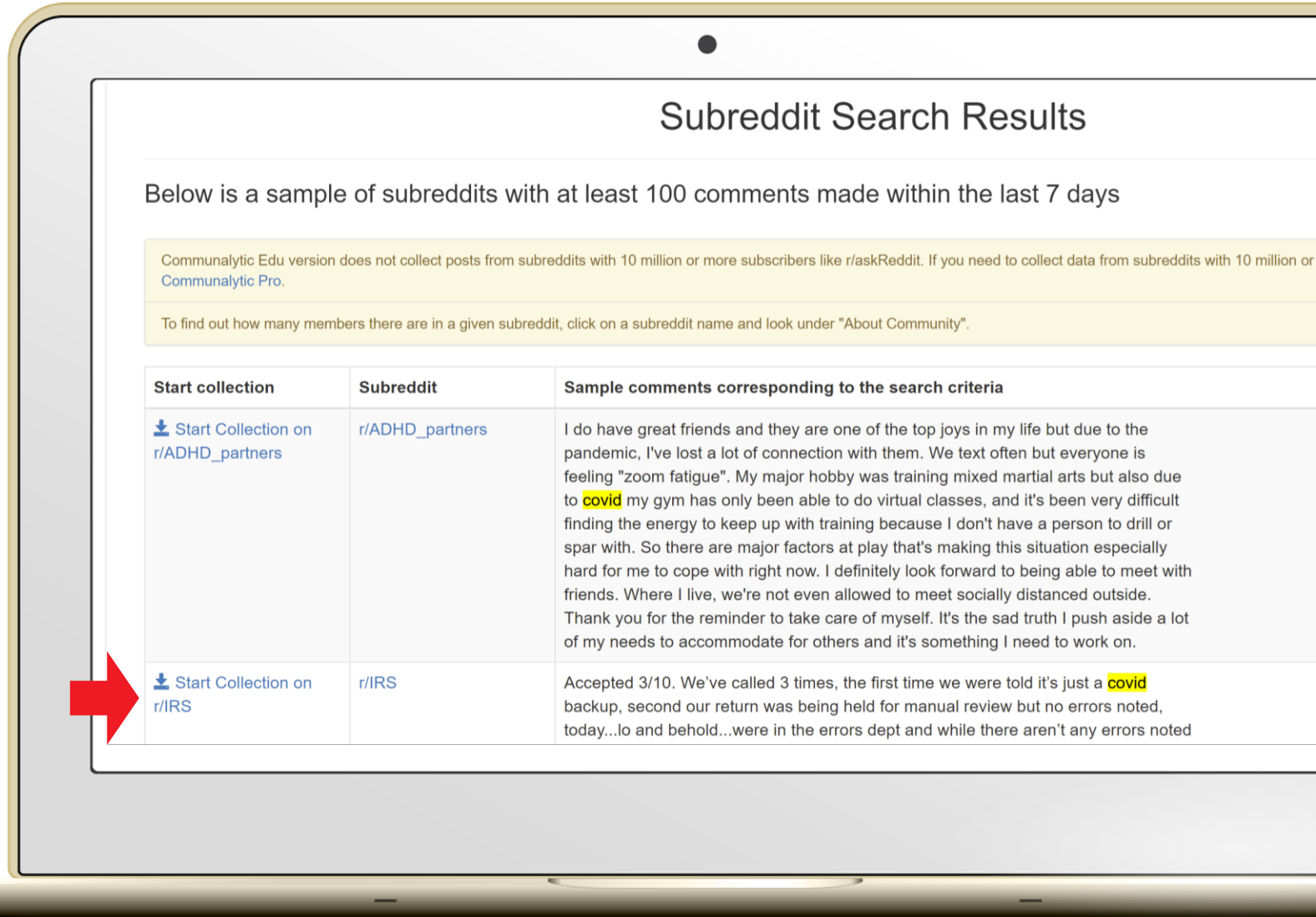
To find out how many members in a given subreddit, click on a subreddit name and look under **"About Community"**.



## Hands-On Part 1:

- 1) Find a relevant subreddit, and
- 2) Start data collection

Click on “Start Collection on ...”



### Subreddit Search Results

Below is a sample of subreddits with at least 100 comments made within the last 7 days

Communalytic Edu version does not collect posts from subreddits with 10 million or more subscribers like r/askReddit. If you need to collect data from subreddits with 10 million or more subscribers, please use [Communalytic Pro](#).

To find out how many members there are in a given subreddit, click on a subreddit name and look under "About Community".

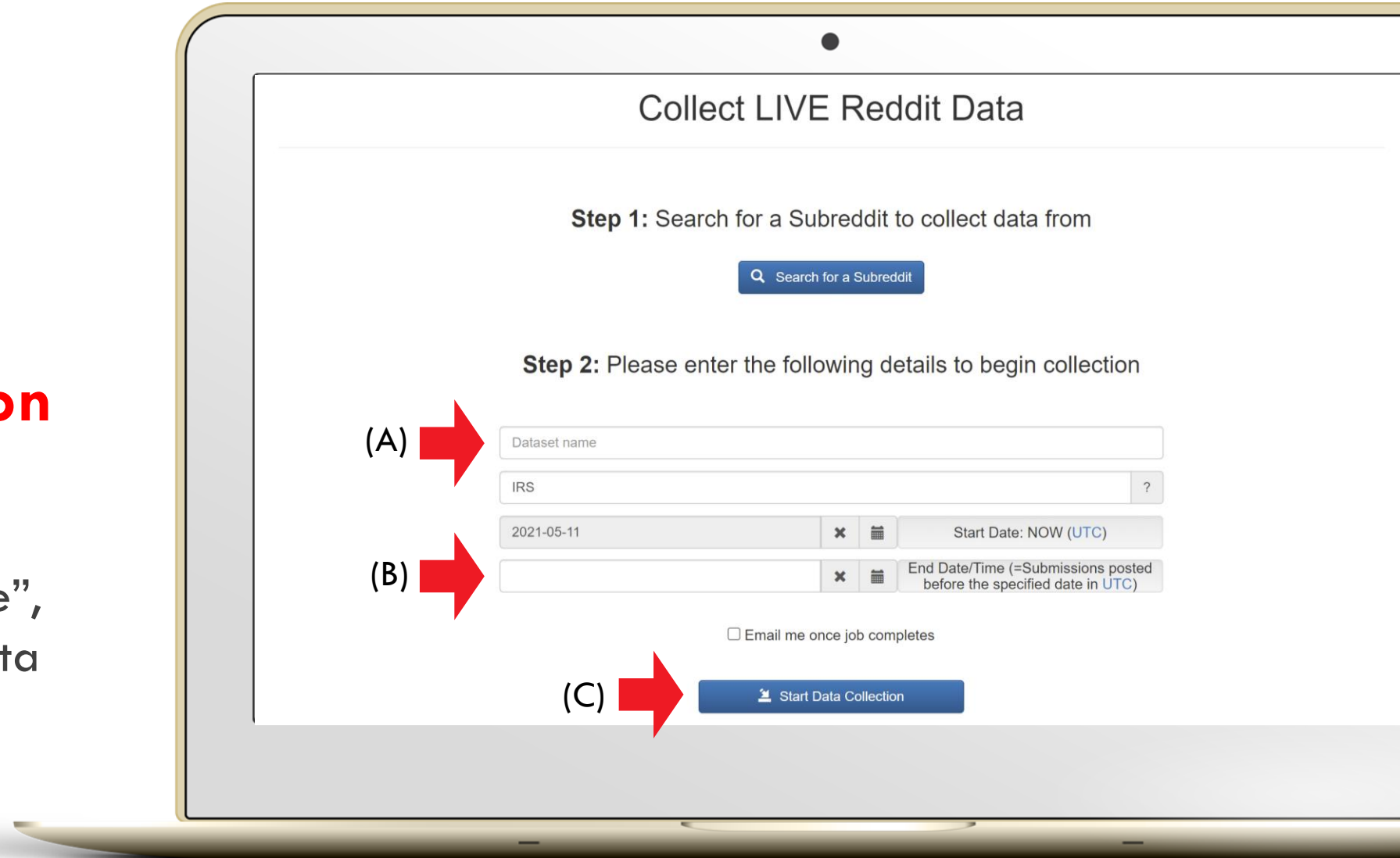
Start collection	Subreddit	Sample comments corresponding to the search criteria
<a href="#">↓ Start Collection on r/ADHD_partners</a>	<a href="#">r/ADHD_partners</a>	I do have great friends and they are one of the top joys in my life but due to the pandemic, I've lost a lot of connection with them. We text often but everyone is feeling "zoom fatigue". My major hobby was training mixed martial arts but also due to <b>covid</b> my gym has only been able to do virtual classes, and it's been very difficult finding the energy to keep up with training because I don't have a person to drill or spar with. So there are major factors at play that's making this situation especially hard for me to cope with right now. I definitely look forward to being able to meet with friends. Where I live, we're not even allowed to meet socially distanced outside. Thank you for the reminder to take care of myself. It's the sad truth I push aside a lot of my needs to accommodate for others and it's something I need to work on.
<a href="#">↓ Start Collection on r/IRS</a>	<a href="#">r/IRS</a>	Accepted 3/10. We've called 3 times, the first time we were told it's just a <b>covid</b> backup, second our return was being held for manual review but no errors noted, today...lo and behold...were in the errors dept and while there aren't any errors noted

## Hands-On Part 1:

1) Find a relevant subreddit, and

**2) Start data collection**

Enter **(A)** a name for the dataset, **(B)** the “end date”, and click on **(C)** “Start Data Collection”.



Collect LIVE Reddit Data

**Step 1:** Search for a Subreddit to collect data from

**Step 2:** Please enter the following details to begin collection

**(A)**

**(B)**

Email me once job completes

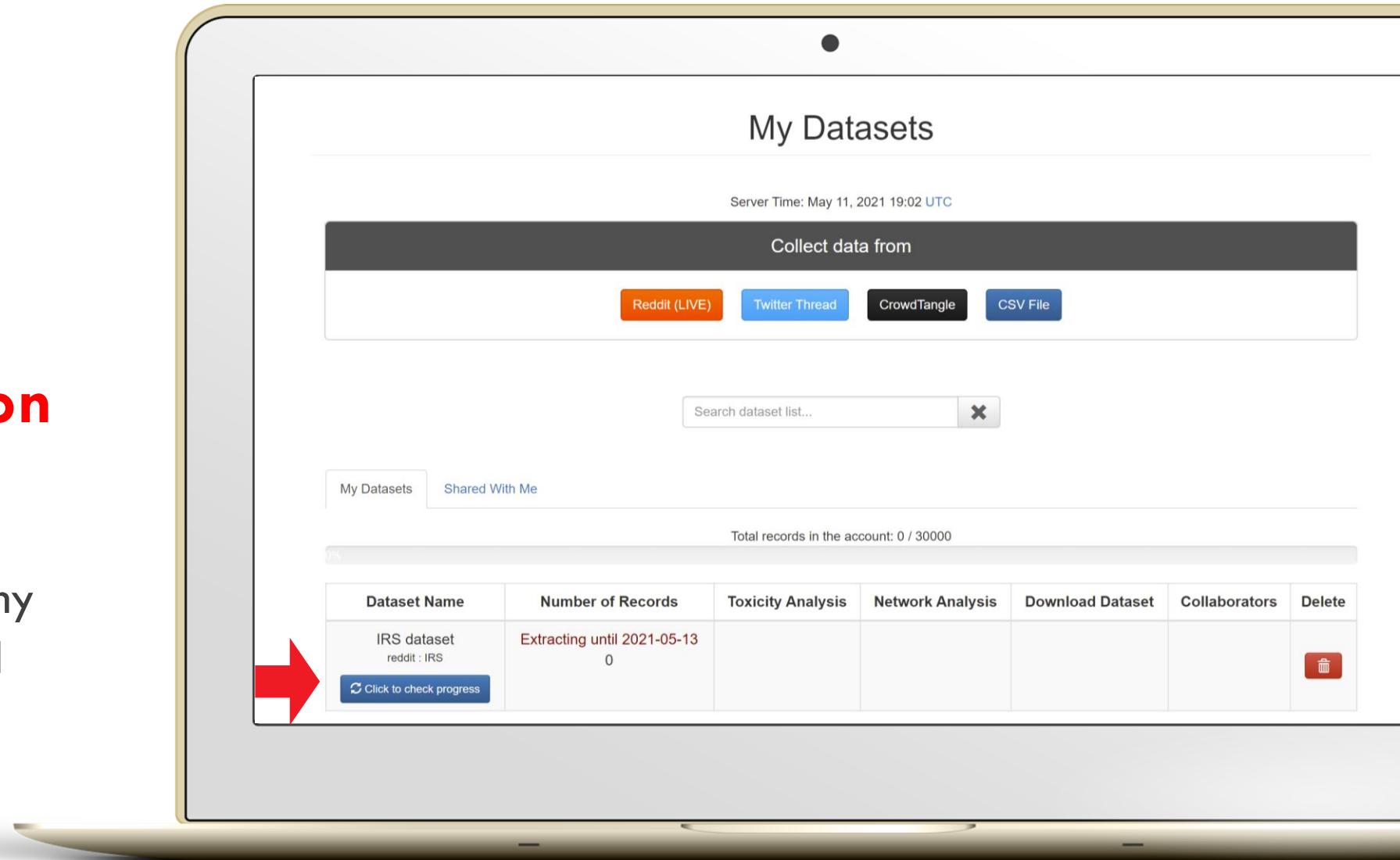
**(C)**



## Hands-On Part 1:

- 1) Find a relevant subreddit, and
- 2) Start data collection**

Click on “**Click to check progress**” to see how many posts have been collected thus far.

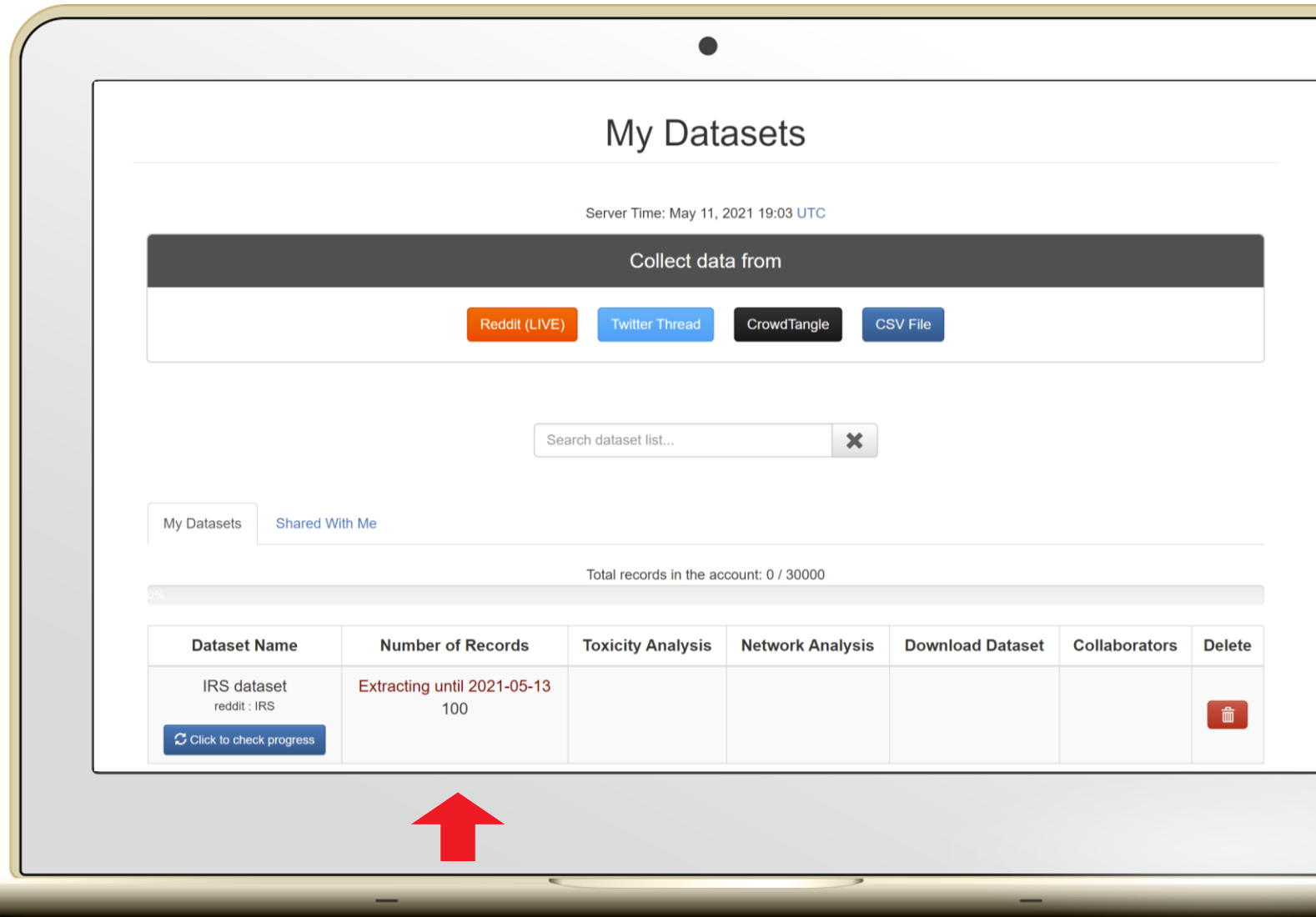


## Hands-On Part 1:

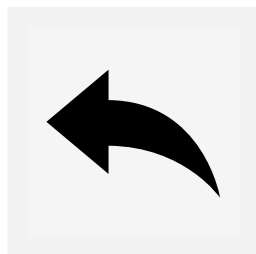
1) Find a relevant subreddit, and

**2) Start data collection**

Communalytic starts by retrieving 100 most recent submissions (=thread starting posts), even if they are posted before the current date.



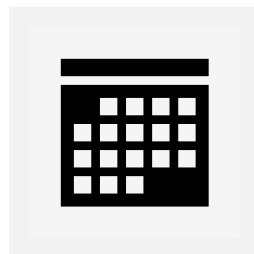
## Three Stages of Live Data Collection from Reddit



### Stage 1

**Collect 100 most recent submissions**

Communalytic starts by retrieving 100 most recent submissions (=thread starting posts), even if they are posted before the current date.



### Stage 2

**Collect new submissions until the end date at 00:01 UTC**

The collection continues until the beginning of the end date at 00:01 am (UTC). For example, if you started collecting data on May 29 and you set the end date to May 31, the collection will stop at the beginning of May 31 at 00:01 UTC.



### Stage 3

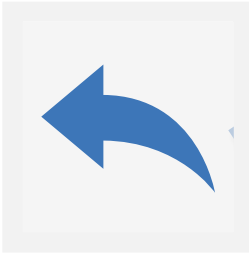
**Collect comments and replies**

During the final stage, Communalytic attempts to retrieve all comments and replies to comments corresponding to the submissions that have been collected during Stage 1 and 2.

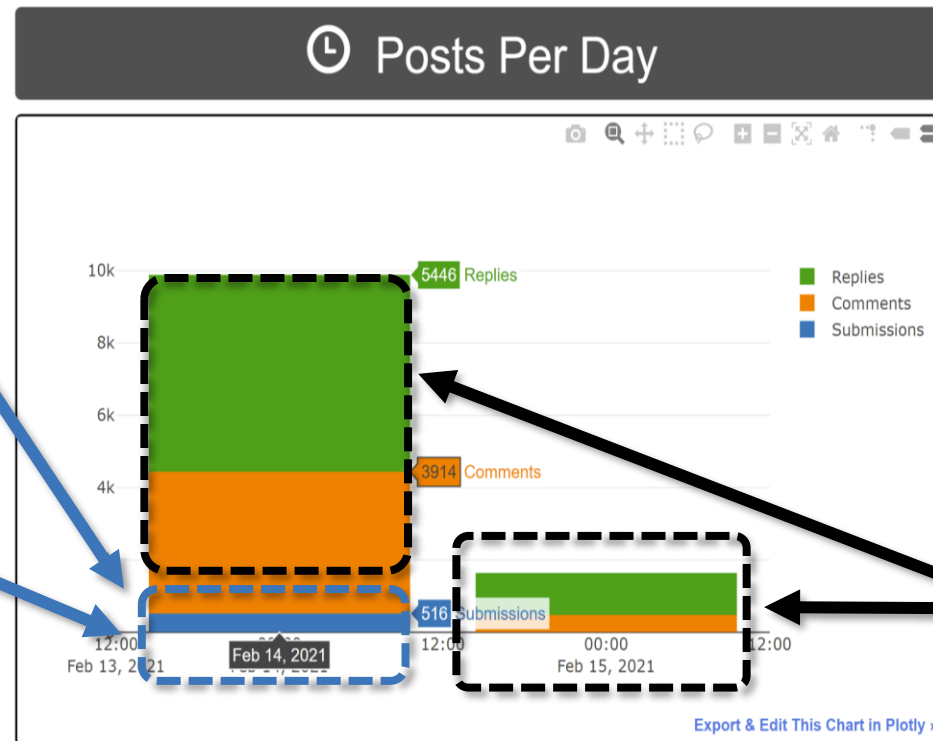
# Three Stages of Live Data Collection from Reddit:

Example: Sample Dataset of 11,512 posts from r/conspiracy

Stage 1  
Collect 100 most recent submissions



Stage 2  
Collect new submissions until the end date at 00:01 UTC



Overview	
Dataset Name:	conspiracy test
Platform:	reddit
Subreddit:	conspiracy
Collection started:	02/14/2021
Collection before:	02/15/2021
Records:	11,512

Stage 3  
Collect comments and replies





# Hands-On Part 2: Practice with a sample dataset

Sample subreddit:  
[r/CoronavirusCanada](https://www.reddit.com/r/CoronavirusCanada)

The screenshot shows the Reddit interface for the subreddit r/CoronavirusCanada. At the top, there's a search bar and navigation links for 'Log In' and 'Sign Up'. The subreddit header includes the name 'CoronavirusCanada', a 'Join' button, and the URL 'r/CoronavirusCanada'. Below the header, there are sorting options: 'Posts', 'Filter By Provinces', and a menu with 'Hot', 'New', 'Top', and a dropdown arrow. The main content area displays three posts. The first post is titled 'COVID-19: Recent deaths in Peterborough had first vaccine dose, health unit says' with a link to 'globalnews.ca/news/7...' and a thumbnail image of a hand in a blue glove. The second post is titled 'anyone else have this reaction?' and describes a reaction to the AZ vaccine. The third post is titled 'All sizes of household portable medical oxygen generator oxygen machine with atomization function O2 concentrator plant for home use'. On the right side, there's an 'About Community' section with a description, member count (8.7k), online count (54), creation date (Feb 12, 2020), and a list of topics including 'Coronavirus' and 'Place'. Below that is a 'Filter by flair' section with buttons for 'ON - Ontario', 'HCoV - Transmission / Safety', 'HCoV - Vaccine', and 'Question'.

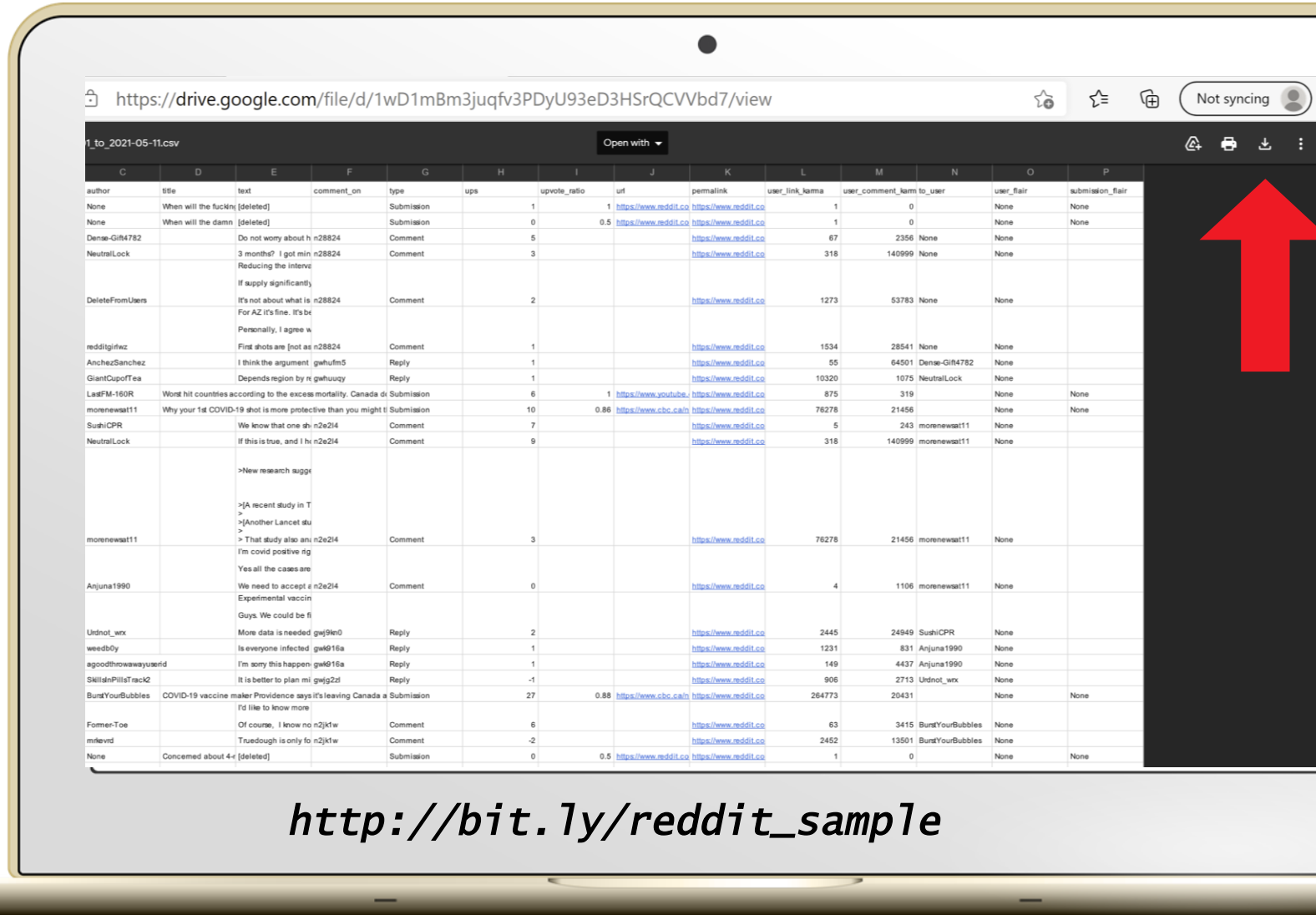
## Hands-On Part 2:

1) Download a sample dataset from [http://bit.ly/reddit\\_sample](http://bit.ly/reddit_sample)

2) Open and review the downloaded dataset in Excel

3) Import the CSV file into Communalytic

4) Review the Overview Charts in Communalytic



https://drive.google.com/file/d/1wD1mBm3juqfv3PDyU93eD3HSrQCVbd7/view

1\_to\_2021-05-11.csv

C	D	E	F	G	H	I	J	K	L	M	N	O	P
author	title	text	comment_on	type	ups	upvote_ratio	url	permalink	user_link_karma	user_comment_karma	user_flair	submission_flair	
None	When will the fuckin	[deleted]		Submission	1	1	<a href="https://www.reddit.co">https://www.reddit.co</a>	<a href="https://www.reddit.co">https://www.reddit.co</a>	1	0	None	None	
None	When will the damn	[deleted]		Submission	0	0.5	<a href="https://www.reddit.co">https://www.reddit.co</a>	<a href="https://www.reddit.co">https://www.reddit.co</a>	1	0	None	None	
Dense-GiR4782		Do not worry about h	n28824	Comment	5			<a href="https://www.reddit.co">https://www.reddit.co</a>	67	2356	None	None	
NeutralLock		3 months? I got min	n28824	Comment	3			<a href="https://www.reddit.co">https://www.reddit.co</a>	318	140999	None	None	
		Reducing the interv											
		If supply significant											
DeleteFromUsers		It's not about what is	n28824	Comment	2			<a href="https://www.reddit.co">https://www.reddit.co</a>	1273	53783	None	None	
		For AZ it's fine. It's be											
		Personally, I agree w											
reddigifwz		First shots are [not as	n28824	Comment	1			<a href="https://www.reddit.co">https://www.reddit.co</a>	1534	28541	None	None	
AnchezSanchez		I think the argument	gahufm5	Reply	1			<a href="https://www.reddit.co">https://www.reddit.co</a>	55	64501	Dense-GiR4782	None	
GiantCupofTea		Depends region by r	gahuuzy	Reply	1			<a href="https://www.reddit.co">https://www.reddit.co</a>	10320	1075	NeutralLock	None	
LastFM-160R		Worst hit countries according to the excess mortality, Canada di		Submission	6	1	<a href="https://www.youtube.c">https://www.youtube.c</a>	<a href="https://www.reddit.co">https://www.reddit.co</a>	875	319	None	None	
morenewat11		Why your 1st COVID-19 shot is more protective than you might t		Submission	10	0.86	<a href="https://www.cbc.ca/h">https://www.cbc.ca/h</a>	<a href="https://www.reddit.co">https://www.reddit.co</a>	76278	21456	None	None	
SushiCPR		We know that one shi	n2e214	Comment	7			<a href="https://www.reddit.co">https://www.reddit.co</a>	5	243	morenewat11	None	
NeutralLock		If this is true, and I h	n2e214	Comment	9			<a href="https://www.reddit.co">https://www.reddit.co</a>	318	140999	morenewat11	None	
		>New research sugg											
		>A recent study in T											
		>Another Lancet stu											
		> That study also an	n2e214	Comment	3			<a href="https://www.reddit.co">https://www.reddit.co</a>	76278	21456	morenewat11	None	
		I'm covid positive rig											
		Yes all the cases are											
Anjuna1990		We need to accept e	n2e214	Comment	0			<a href="https://www.reddit.co">https://www.reddit.co</a>	4	1106	morenewat11	None	
		Experimental vaccin											
		Guys. We could be fi											
Urdnot_wrx		More data is needed	gwj9ln0	Reply	2			<a href="https://www.reddit.co">https://www.reddit.co</a>	2445	24949	SushiCPR	None	
weedboy		Is everyone infected	gwj916a	Reply	1			<a href="https://www.reddit.co">https://www.reddit.co</a>	1231	831	Anjuna1990	None	
agoodthrowawayuserid		I'm sorry this happen	gwj916a	Reply	1			<a href="https://www.reddit.co">https://www.reddit.co</a>	149	4437	Anjuna1990	None	
SkillsnPills4Track2		It is better to plan mi	gwjg2d1	Reply	-1			<a href="https://www.reddit.co">https://www.reddit.co</a>	906	2713	Urdnot_wrx	None	
BurstYourBubbles		COVID-19 vaccine mater Providence says it's leaving Canada a		Submission	27	0.88	<a href="https://www.cbc.ca/h">https://www.cbc.ca/h</a>	<a href="https://www.reddit.co">https://www.reddit.co</a>	264773	20431	None	None	
		I'd like to know more											
Former-Toe		Of course, I know no	n2jkt1w	Comment	6			<a href="https://www.reddit.co">https://www.reddit.co</a>	63	3415	BurstYourBubbles	None	
mifevrd		Tuedough is only fo	n2jkt1w	Comment	-2			<a href="https://www.reddit.co">https://www.reddit.co</a>	2452	13501	BurstYourBubbles	None	
None		Concerned about 4+	[deleted]	Submission	0	0.5	<a href="https://www.reddit.co">https://www.reddit.co</a>	<a href="https://www.reddit.co">https://www.reddit.co</a>	1	0	None	None	

[http://bit.ly/reddit\\_sample](http://bit.ly/reddit_sample)

## Hands-On Part 2:

1) Download a sample dataset from [http://bit.ly/reddit\\_sample](http://bit.ly/reddit_sample)

2) Open and review the downloaded dataset in Excel

3) Import the CSV file into Communalytic

4) Review the Overview Charts in Communalytic

text	comment_on	type	ups	upvote_ratio	url	perma	user_link_karma	user_comment
cking [deleted]		Submission	1		1 https://ww	https://w		1
arn in [deleted]		Submission	0		0.5 https://ww	https://w		1
Do not worry about how efficacious a vaccine booster is after 2	n28824	Comment	5			https://w		67
3 months? I got mine two weeks ago and I'm booked for August	n28824	Comment	3			https://w		318
Reducing the interval means we stay in this pandemic longer,	n28824	Comment	2			https://w		1273
For AZ it's fine. It's been studied and it actually seems to be	n28824	Comment	1			https://w		1534
I think the argument is its better to get as many 1st doses into	gwhufm5	Reply	1			https://w		55
Depends region by region I guess. Let's just say I live in a city th	gwhuuqy	Reply	1			https://w		10320
es according to the excess mortality. Canada doing well with almost		Submission	6		1 https://ww	https://w		875
VID-19 shot is more protective than you might think		Submission	10	0.86	https://ww	https://w		76278
We know that one shot is effective after 4-5 weeks. We don't kno	n2e2l4	Comment	7			https://w		5
If this is true, and I hope it is, it would be nice if they started es	n2e2l4	Comment	9			https://w		318
	n2e2l4	Comment	3			https://w		76278
I'm covid positive right now and received Pfizer in early	n2e2l4	Comment	0			https://w		4
Experimental vaccine, emergency use authorized...AND THEN	gwj9kn0	Reply	2			https://w		2445
Is everyone infected the same?	gwk916a	Reply	1			https://w		1231
I'm sorry this happened to you! If you don't mind me asking, h	gwk916a	Reply	1			https://w		149
It is better to plan mistakes and learn by ourselves than succee	gwjg2zl	Reply	-1			https://w		906
e maker Providence says it's leaving Canada after calls for more feder		Submission	27	0.88	https://ww	https://w		264773
I'd like to know more about this company and their product	n2jk1w	Comment	6			https://w		63
Truedough is only for big pharma	n2jk1w	Comment	-2			https://w		2452
t 4-mc [deleted]		Submission	0		0.5 https://ww	https://w		1
What help are you actually after?	n2lxay	Comment	4			https://w		110
I don't think it will be 4 months for many by the time June hits	n2lxay	Comment	4			https://w		1
I'm covid positive right now (I received a first dose in early	n2lxay	Comment	1			https://w		4
yea but my mom got her vaccine last week and now here secor	gwlbx98	Reply	1			https://w		10320
So basically, you got COVID, but you were asymptomatic becau	gwk8f47	Reply	3			https://w		10320
So basically the vaccine allowed you and others to have almost	gwk8f47	Reply	2			https://w		1
I completely agree but the concern is now are silently spreadi	gwkaxq3	Reply	2			https://w		4
I don't disagree if the chance of death with one shot is close to	gwk8m8	Reply	2			https://w		1
Asymptomatic spread has always been a major driver of	gwk8m8	Reply	1			https://w		110
UK has much more of it's population vaccinated. That's why the	gwlf32s	Reply	1			https://w		1111

## Hands-On Part 2:

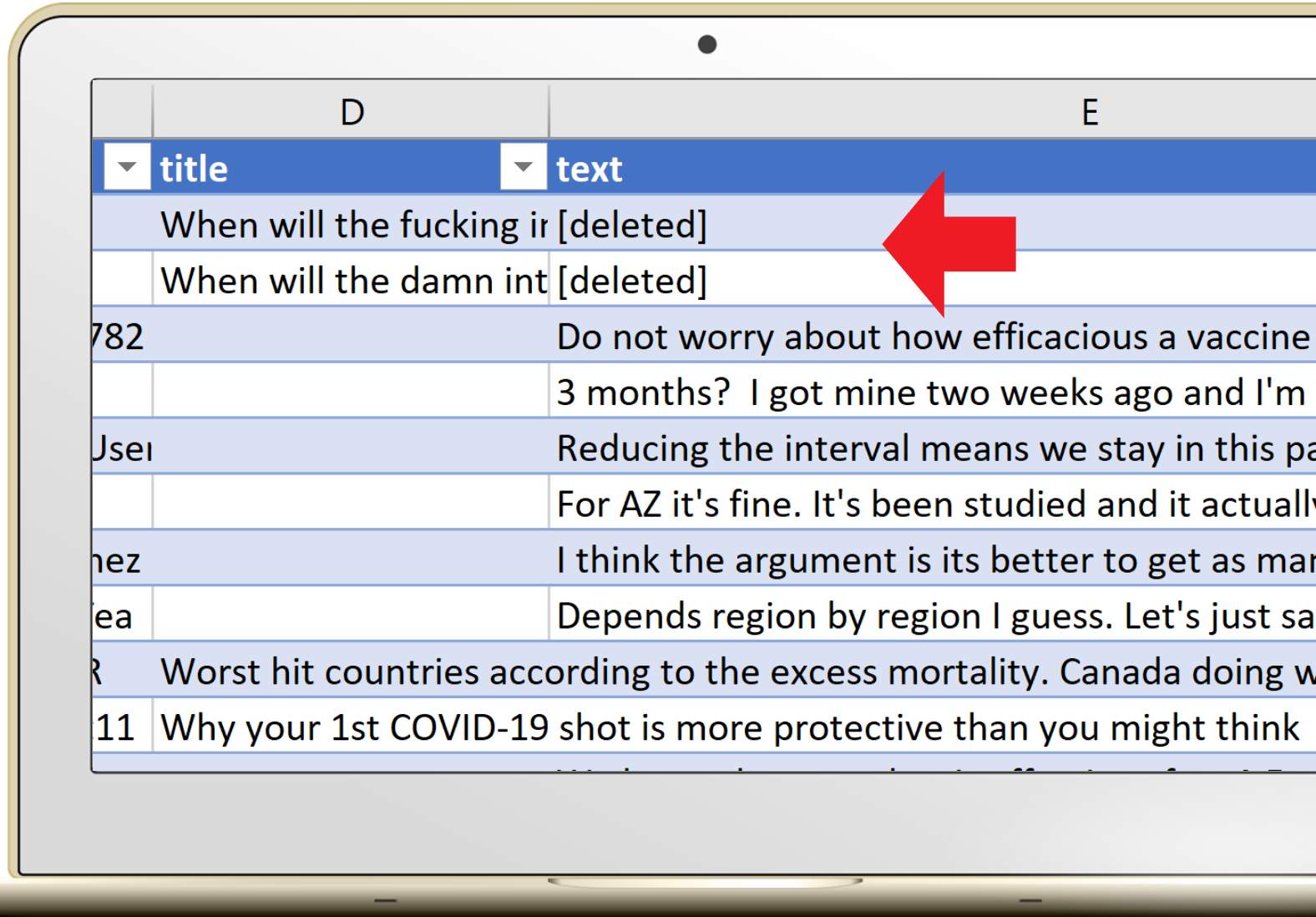
1) Download a sample dataset from [http://bit.ly/reddit\\_sample](http://bit.ly/reddit_sample)

### 2) Open and review the downloaded dataset in Excel

Posts deleted either by the moderator or the poster will be shown as “[deleted]”

3) Import the CSV file into Communalystic

4) Review the Overview Charts in Communalystic



	D	E
	▼ title	▼ text
		When will the fucking ir [deleted]
		When will the damn int [deleted]
782		Do not worry about how efficacious a vaccine
		3 months? I got mine two weeks ago and I'm
Jsei		Reducing the interval means we stay in this pa
		For AZ it's fine. It's been studied and it actually
hez		I think the argument is its better to get as mar
ea		Depends region by region I guess. Let's just sa
R		Worst hit countries according to the excess mortality. Canada doing w
11		Why your 1st COVID-19 shot is more protective than you might think



## Hands-On Part 2:

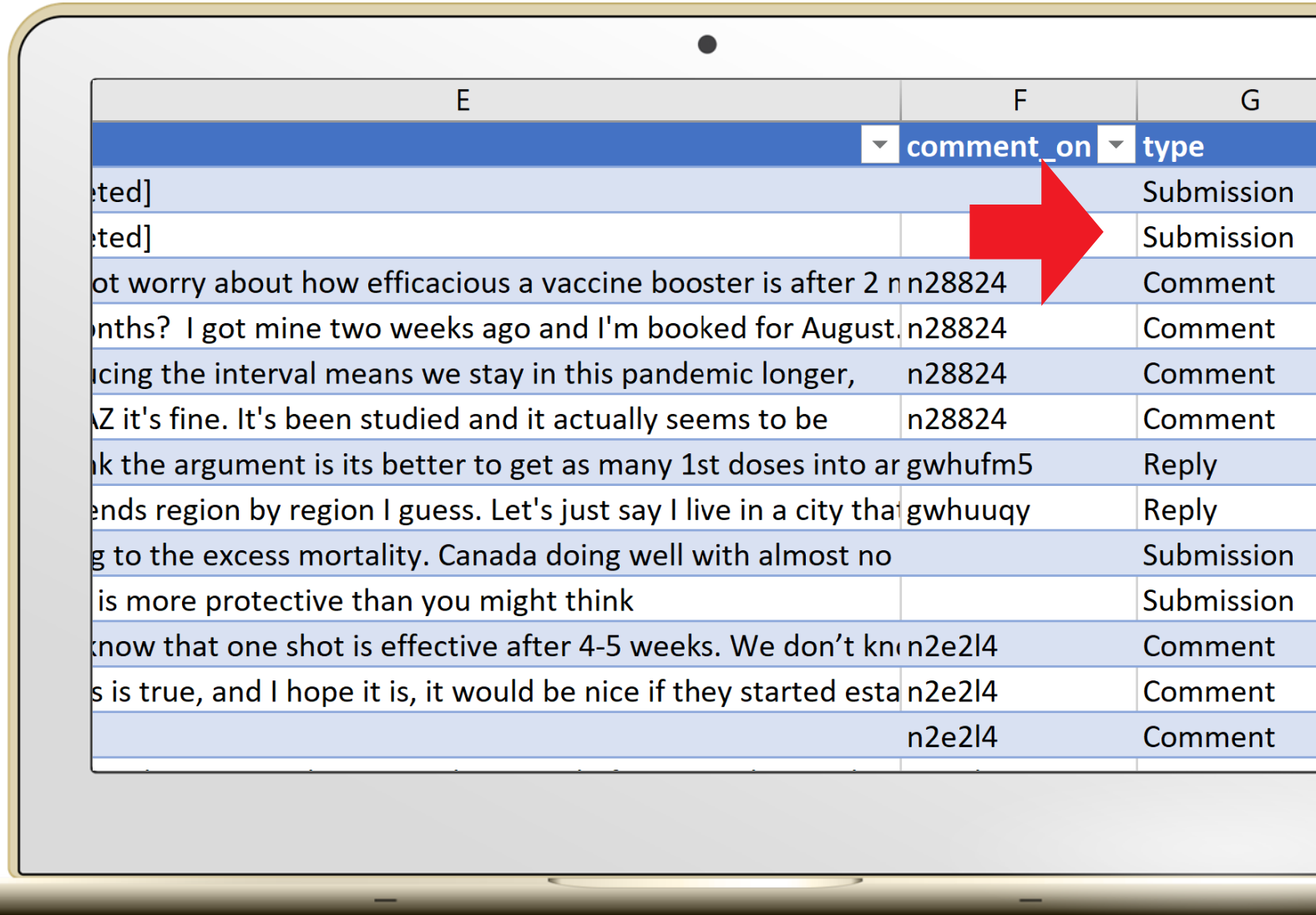
1) Download a sample dataset from [http://bit.ly/reddit\\_sample](http://bit.ly/reddit_sample)

**2) Open and review the downloaded dataset in Excel**

Three types of Reddit posts:  
Submission, Comment & Reply

3) Import the CSV file into Communalystic

4) Review the Overview Charts in Communalystic



E	F	G
	comment_on	type
ted]		Submission
ted]		Submission
ot worry about how efficacious a vaccine booster is after 2 n	n28824	Comment
onths? I got mine two weeks ago and I'm booked for August.	n28824	Comment
ancing the interval means we stay in this pandemic longer,	n28824	Comment
WZ it's fine. It's been studied and it actually seems to be	n28824	Comment
nk the argument is its better to get as many 1st doses into ar	gwhufm5	Reply
ends region by region I guess. Let's just say I live in a city tha	gwhuuqy	Reply
g to the excess mortality. Canada doing well with almost no		Submission
is more protective than you might think		Submission
know that one shot is effective after 4-5 weeks. We don't kn	n2e2l4	Comment
s is true, and I hope it is, it would be nice if they started esta	n2e2l4	Comment
	n2e2l4	Comment

## Hands-On Part 2:

1) Download a sample dataset from [http://bit.ly/reddit\\_sample](http://bit.ly/reddit_sample)

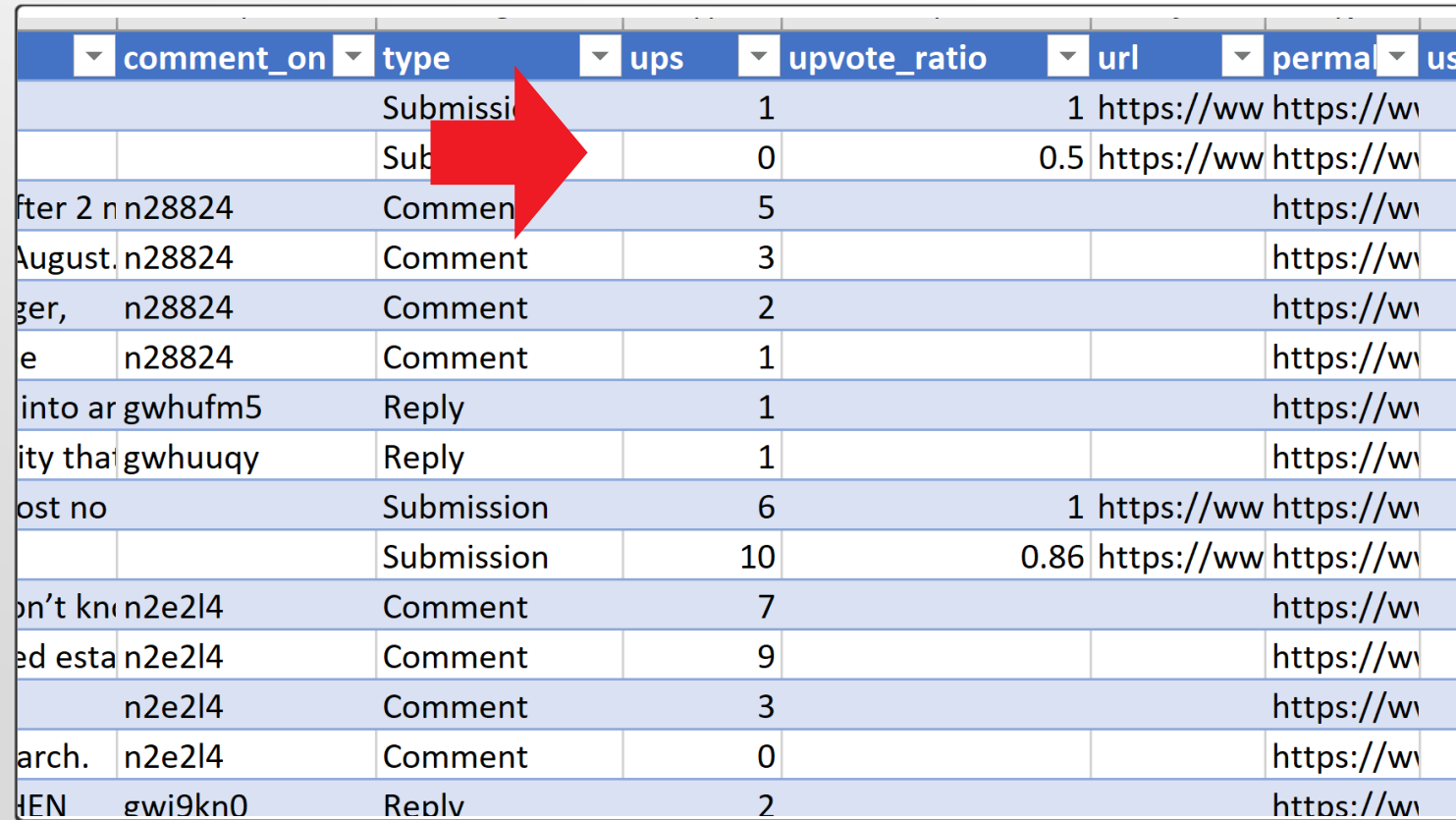
2) Open and review the downloaded dataset in Excel

“ups” = #ups - #downs (as provided by Reddit API)

“upvote\_ratio” is only available for “Submissions” (as provided by Reddit API)

3) Import the CSV file into Communalytic

4) Review the Overview Charts in Communalytic



comment_on	type	ups	upvote_ratio	url	perma	us
	Submissi		1	1 https://ww	https://w	
	Sub		0	0.5 https://ww	https://w	
fter 2 n n28824	Commen		5		https://w	
August. n28824	Comment		3		https://w	
ger, n28824	Comment		2		https://w	
e n28824	Comment		1		https://w	
into ar gwhufm5	Reply		1		https://w	
ity tha gwhuuqy	Reply		1		https://w	
ost no	Submission		6	1 https://ww	https://w	
	Submission		10	0.86 https://ww	https://w	
on't kn n2e2l4	Comment		7		https://w	
ed esta n2e2l4	Comment		9		https://w	
n2e2l4	Comment		3		https://w	
arch. n2e2l4	Comment		0		https://w	
HEN gwi9kn0	Replv		2		https://w	

## Hands-On Part 2:

1) Download a sample dataset from [http://bit.ly/reddit\\_sample](http://bit.ly/reddit_sample)

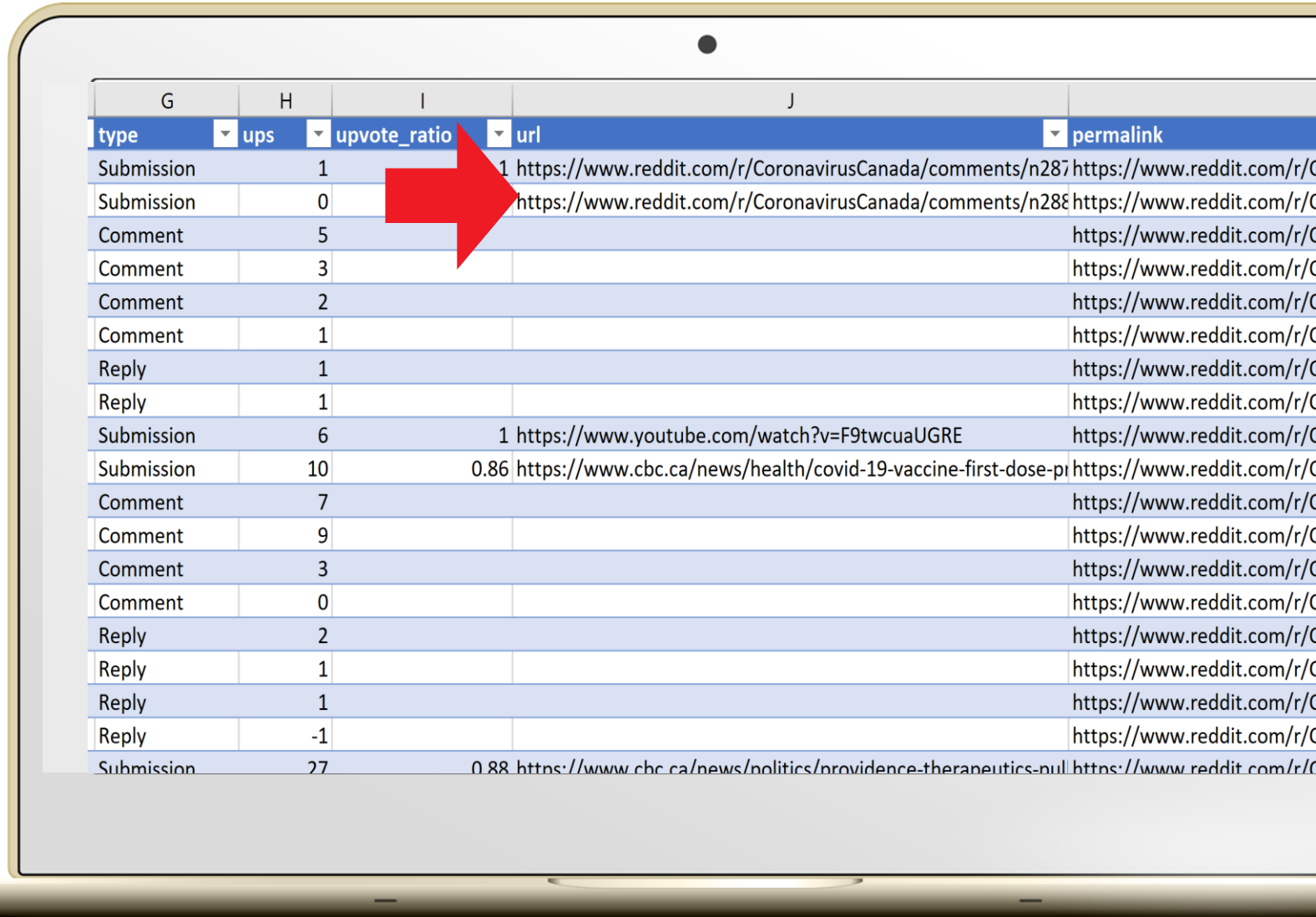
2) Open and review the downloaded dataset in Excel

“permalink” = a direct link to the post on Reddit

“url” = a link to a web resource shared (for “submissions” only)

3) Import the CSV file into Communalytic

4) Review the Overview Charts in Communalytic



	G	H	I	J	
	type	ups	upvote_ratio	url	permalink
	Submission	1	1	https://www.reddit.com/r/CoronavirusCanada/comments/n287	https://www.reddit.com/r/C
	Submission	0		https://www.reddit.com/r/CoronavirusCanada/comments/n288	https://www.reddit.com/r/C
	Comment	5			https://www.reddit.com/r/C
	Comment	3			https://www.reddit.com/r/C
	Comment	2			https://www.reddit.com/r/C
	Comment	1			https://www.reddit.com/r/C
	Reply	1			https://www.reddit.com/r/C
	Reply	1			https://www.reddit.com/r/C
	Submission	6	1	https://www.youtube.com/watch?v=F9twcuaUGRE	https://www.reddit.com/r/C
	Submission	10	0.86	https://www.cbc.ca/news/health/covid-19-vaccine-first-dose-p	https://www.reddit.com/r/C
	Comment	7			https://www.reddit.com/r/C
	Comment	9			https://www.reddit.com/r/C
	Comment	3			https://www.reddit.com/r/C
	Comment	0			https://www.reddit.com/r/C
	Reply	2			https://www.reddit.com/r/C
	Reply	1			https://www.reddit.com/r/C
	Reply	1			https://www.reddit.com/r/C
	Reply	-1			https://www.reddit.com/r/C
	Submission	27	0.88	https://www.cbc.ca/news/politics/providence-therapeutics-null	https://www.reddit.com/r/C

## Hands-On Part 2:

1) Download a sample dataset from [http://bit.ly/reddit\\_sample](http://bit.ly/reddit_sample)

### 2) Open and review the downloaded dataset in Excel

There are some additional metadata elements for link/comment “karma” & subreddit-specific user/submission “flairs”.

3) Import the CSV file into Communalystic

4) Review the Overview Charts in Communalystic

user_link_karma	user_comment_karma	to_user	user_flair	submission_flair
1	0		None	Question
5	652		None	Question
614	45822		None	HCoV - Vaccine
614	45822		None	HCoV - Vaccine
68	3586		None	Question
18560	5620		Moderator	HCoV - Vaccine
55	21		None	Opinion
156	348		None	Question
1	14		None	Question
18560	5620		Moderator	HCoV - Vaccine
3521	5		None	Canada 🍁
14611	129973		None	ON - Ontario
222	56		None	HCoV - Vaccine
220	-16		None	Question
34906	567		None	HCoV - Vaccine
19	347		None	HCoV - Vaccine
7535	1331		None	HCoV - Vaccine
20875	19692		None	MB - Manitoba

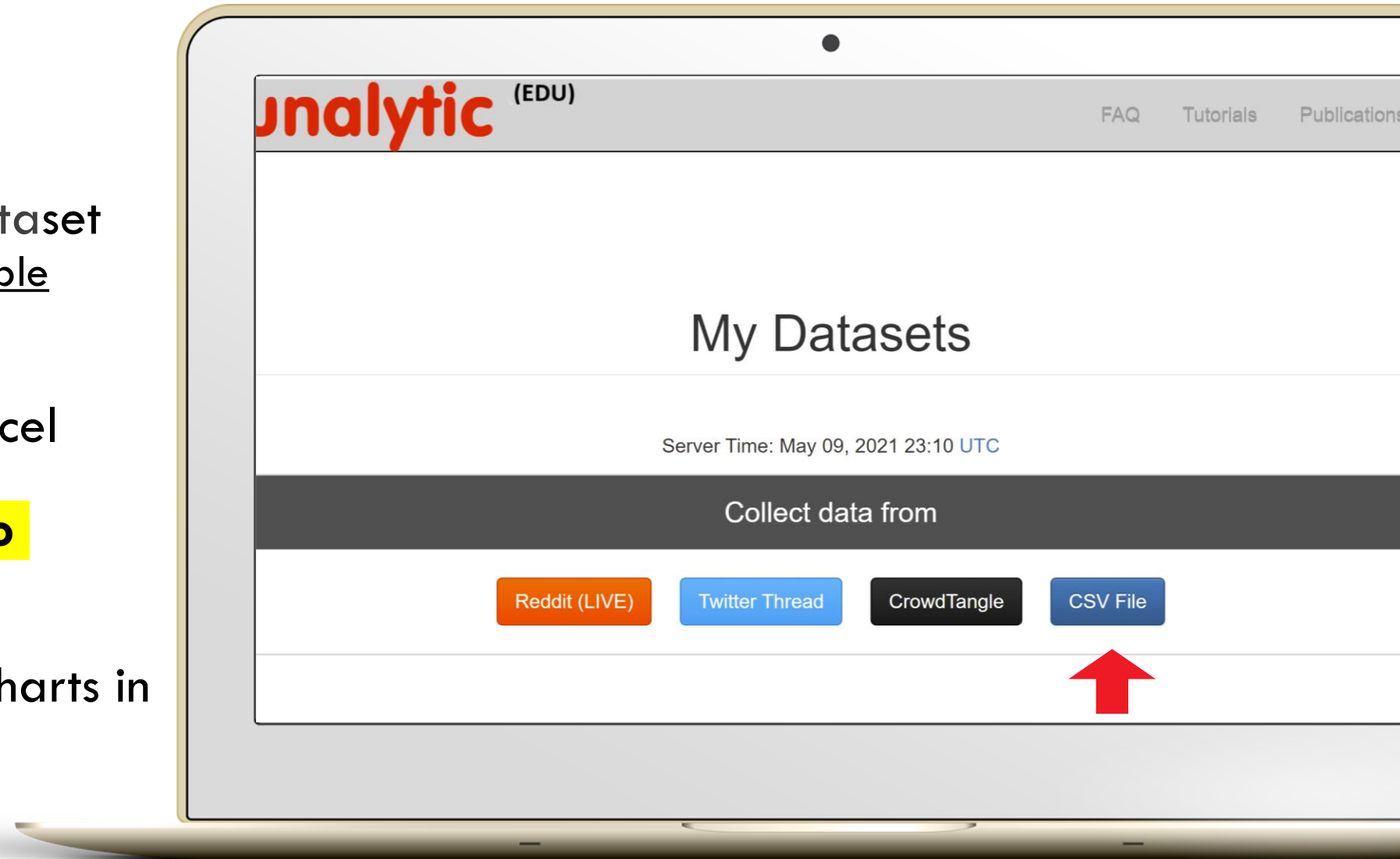
## Hands-On Part 2:

1) Download a sample dataset from [http://bit.ly/reddit\\_sample](http://bit.ly/reddit_sample)

2) Open and review the downloaded dataset in Excel

**3) Import the CSV file into Communalytic**

4) Review the Overview Charts in Communalytic





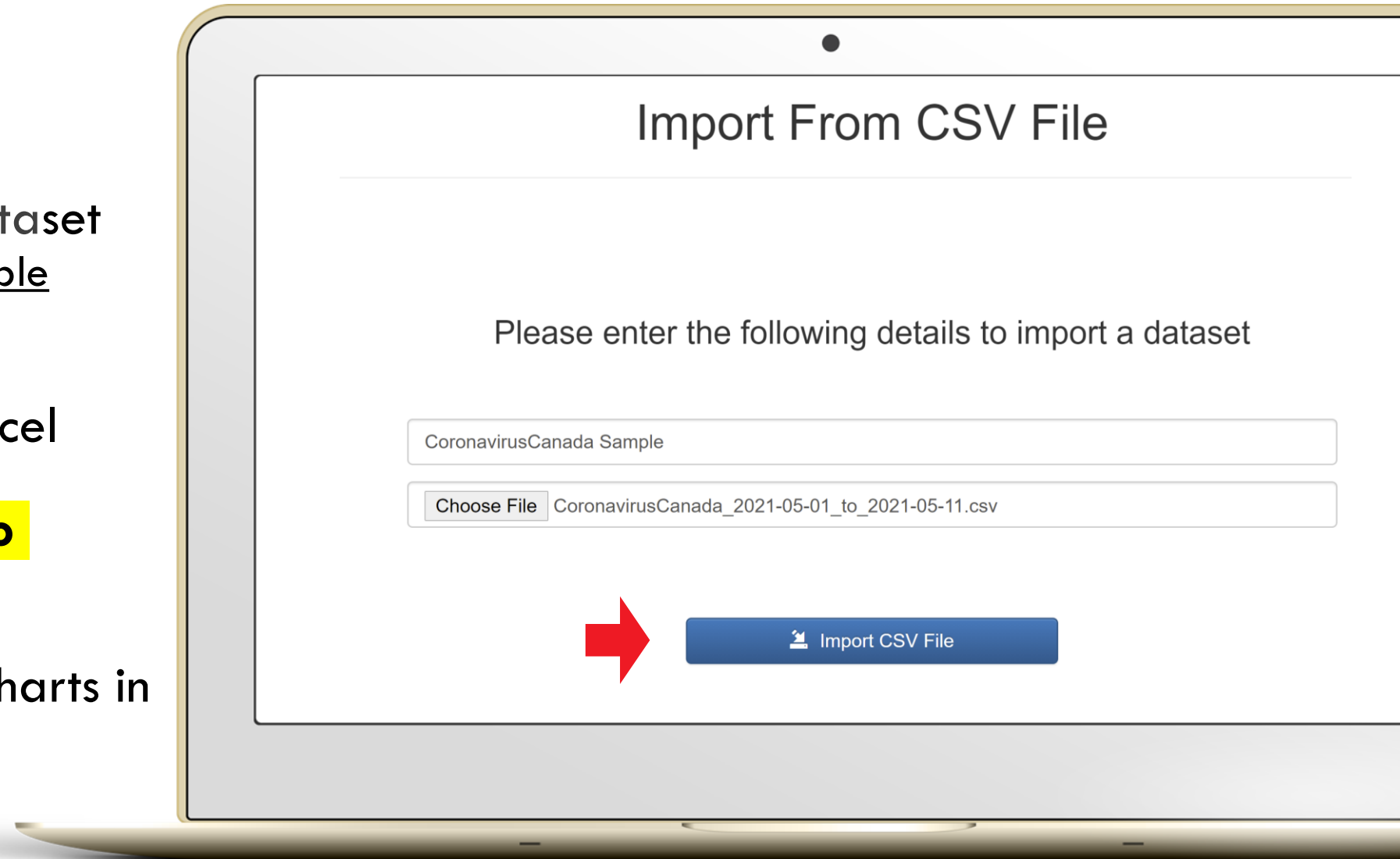
## Hands-On Part 2:

1) Download a sample dataset from [http://bit.ly/reddit\\_sample](http://bit.ly/reddit_sample)

2) Open and review the downloaded dataset in Excel

**3) Import the CSV file into Communalytic**

4) Review the Overview Charts in Communalytic



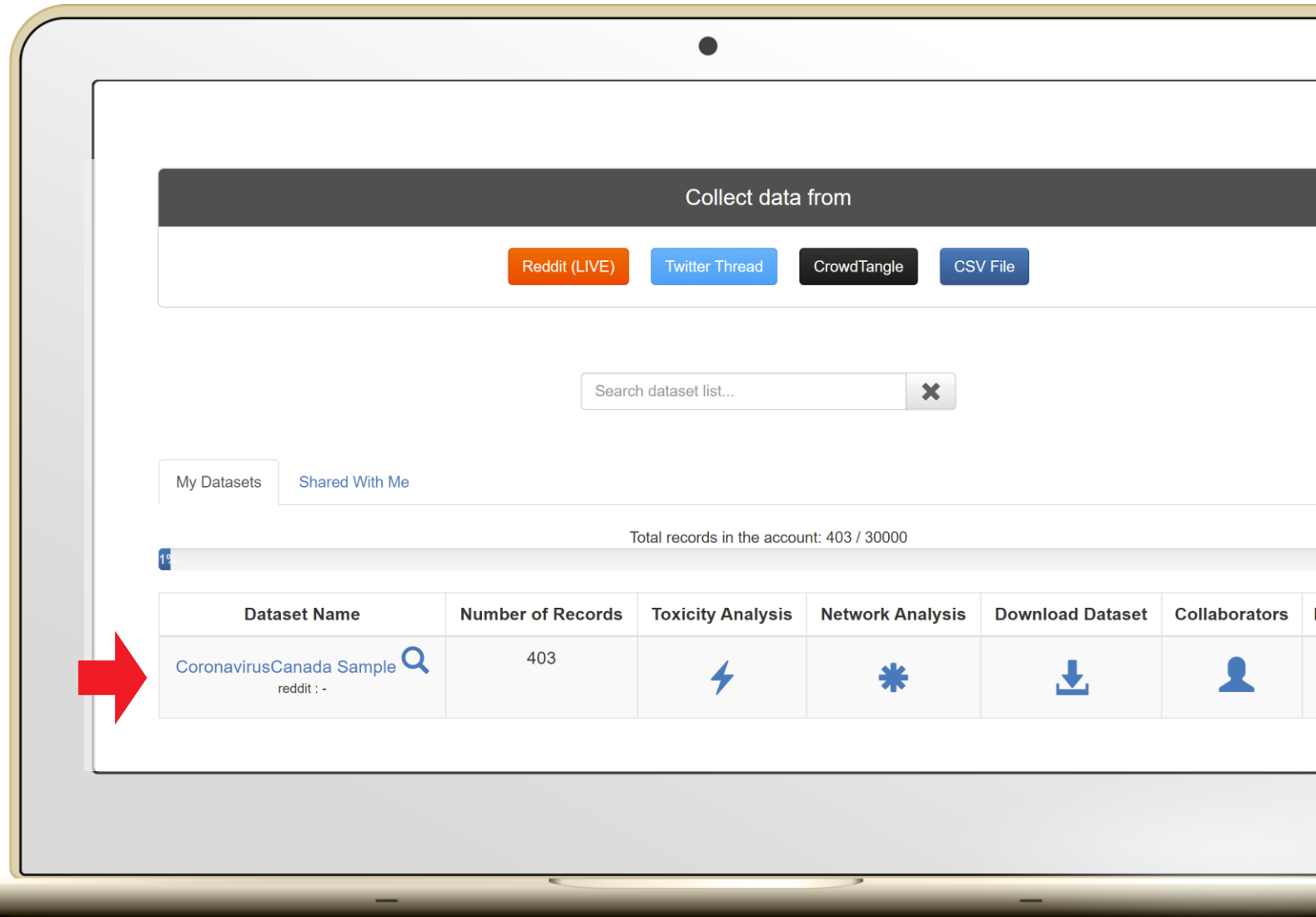
## Hands-On Part 2:

1) Download a sample dataset from [http://bit.ly/reddit\\_sample](http://bit.ly/reddit_sample)

2) Open and review the downloaded dataset in Excel

3) Import the CSV file into Communalytic

4) Review the Overview Charts in Communalytic



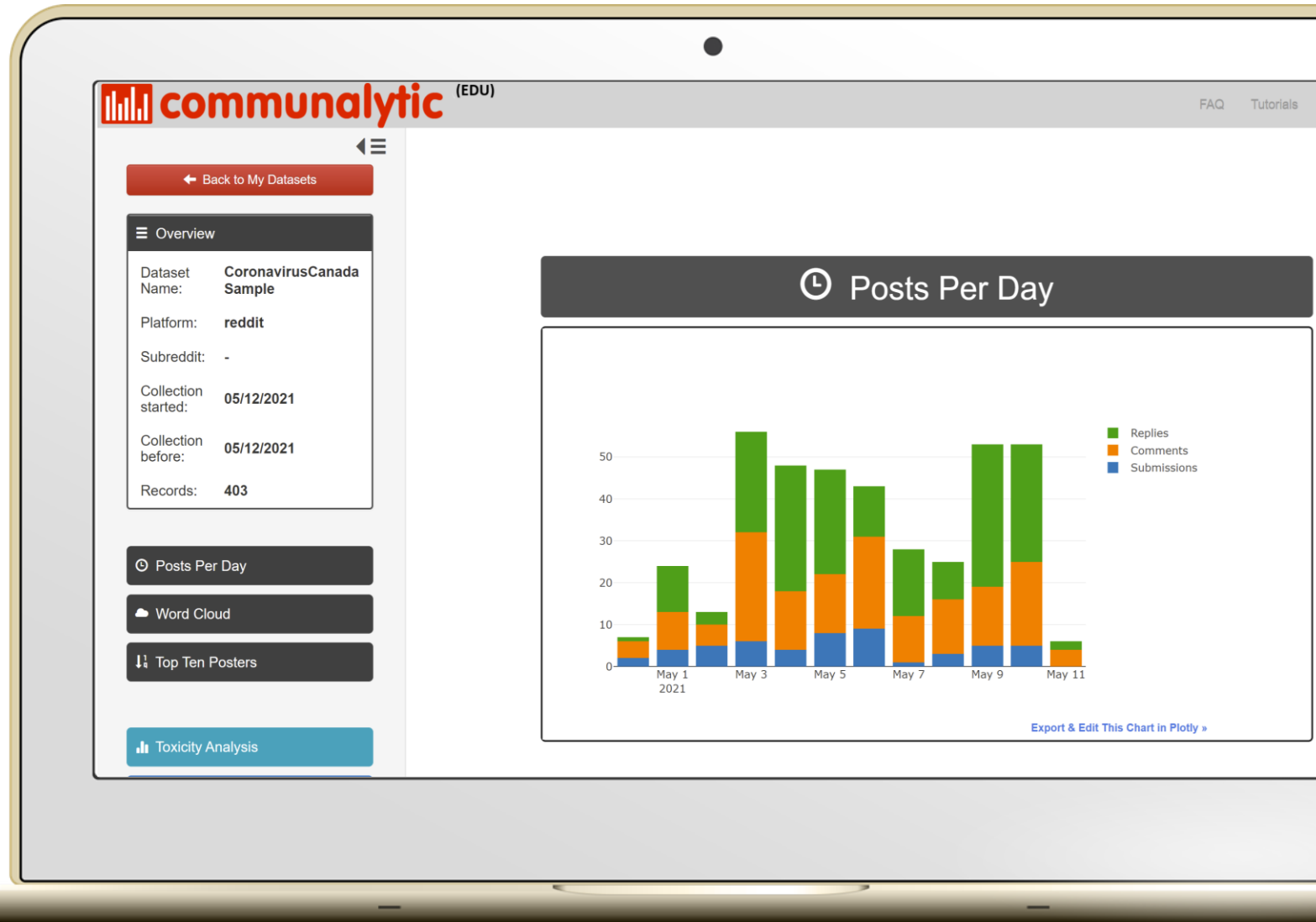
## Hands-On Part 2:

1) Download a sample dataset from [http://bit.ly/reddit\\_sample](http://bit.ly/reddit_sample)

2) Open and review the downloaded dataset in Excel

3) Import the CSV file into Communalytic

4) Review the Overview Charts in Communalytic



## Hands-On Part 2:

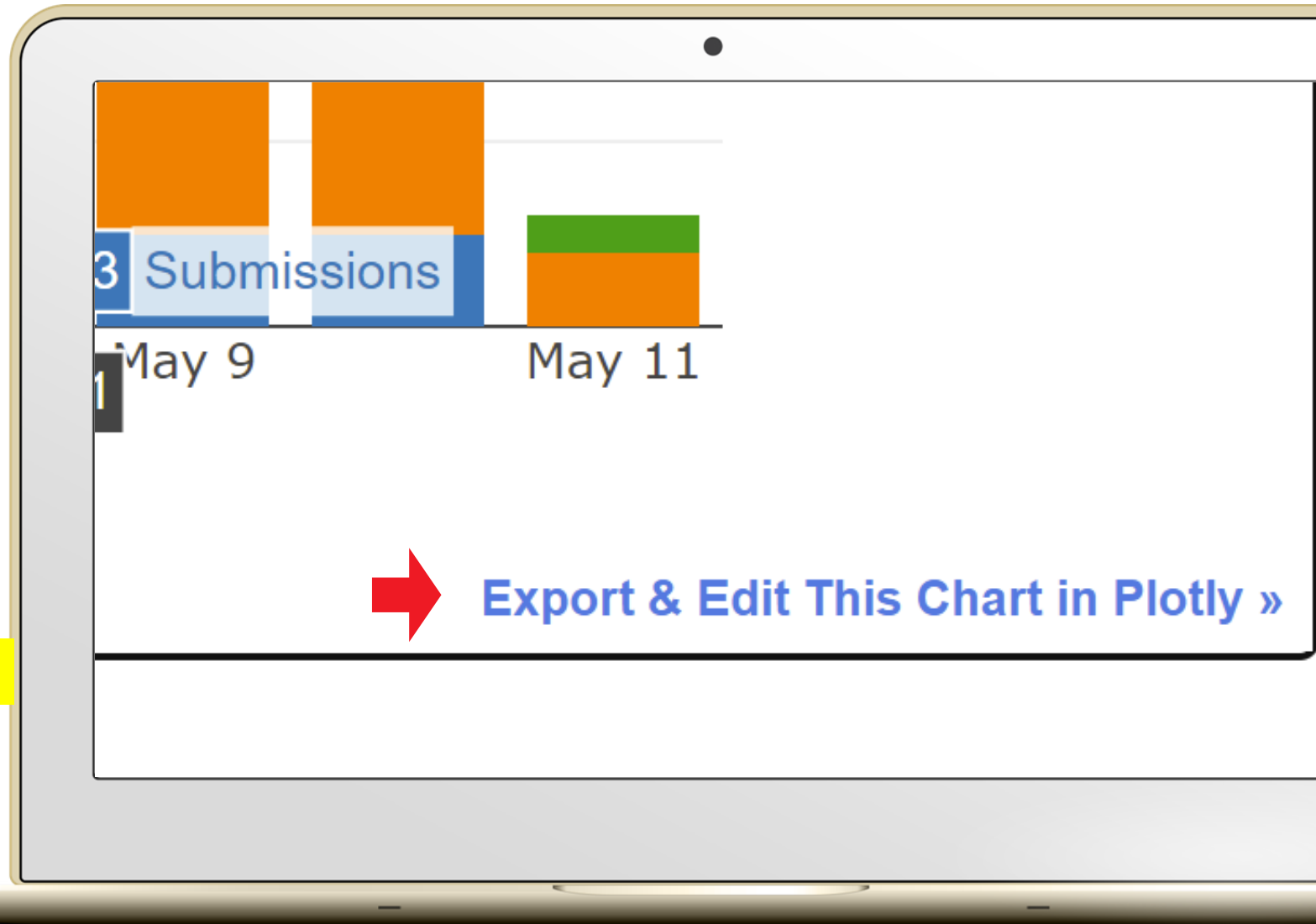
1) Download a sample dataset from [http://bit.ly/reddit\\_sample](http://bit.ly/reddit_sample)

2) Open and review the downloaded dataset in Excel

3) Import the CSV file into Communalystic

**4) Review the Overview Charts in Communalystic**

You can visualize and customize charts using Plot.ly



## Hands-On Part 2:

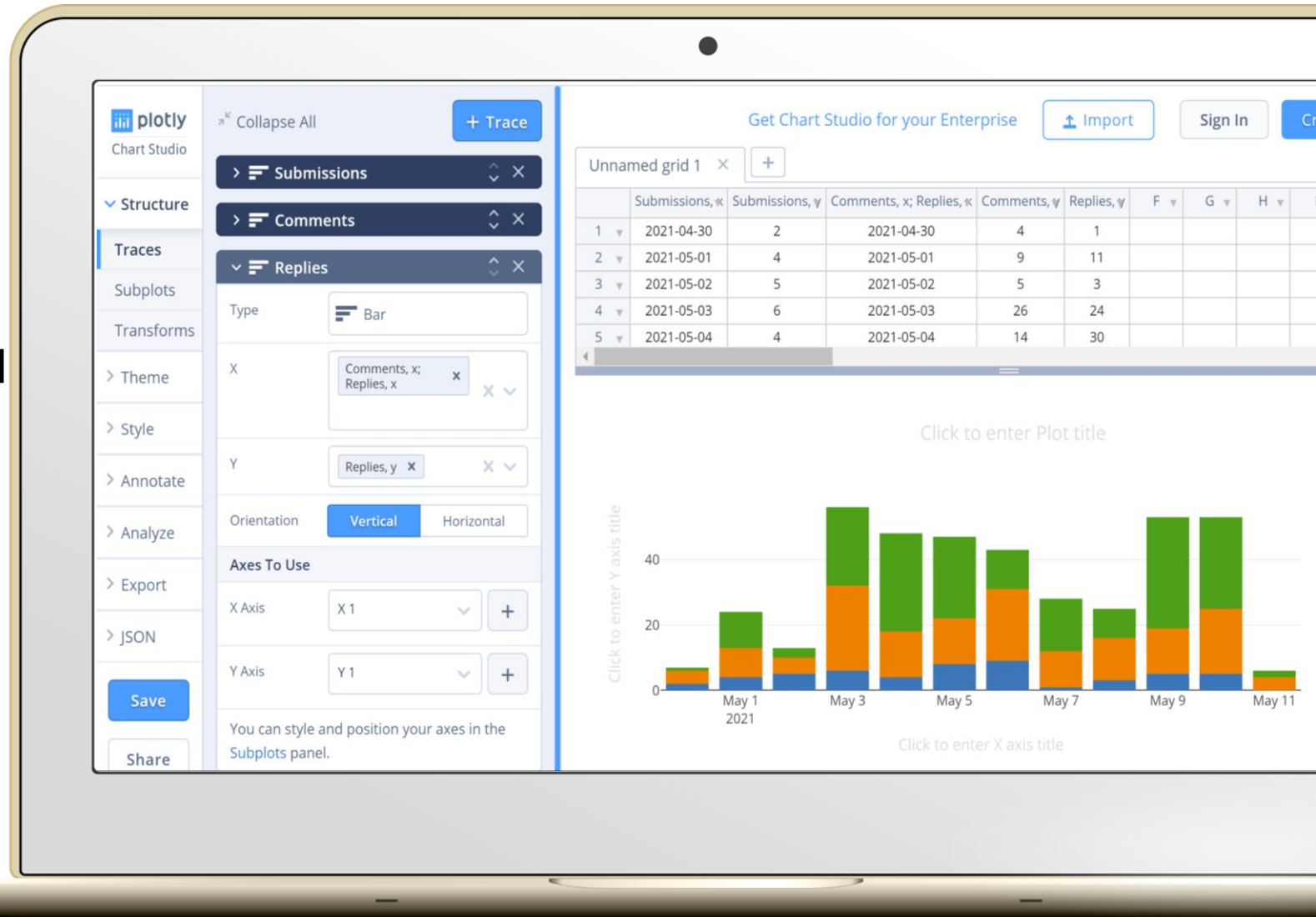
1) Download a sample dataset from [http://bit.ly/reddit\\_sample](http://bit.ly/reddit_sample)

2) Open and review the downloaded dataset in Excel

3) Import the CSV file into Communalytic

4) Review the Overview Charts in Communalytic

You can visualize and customize charts using Plot.ly





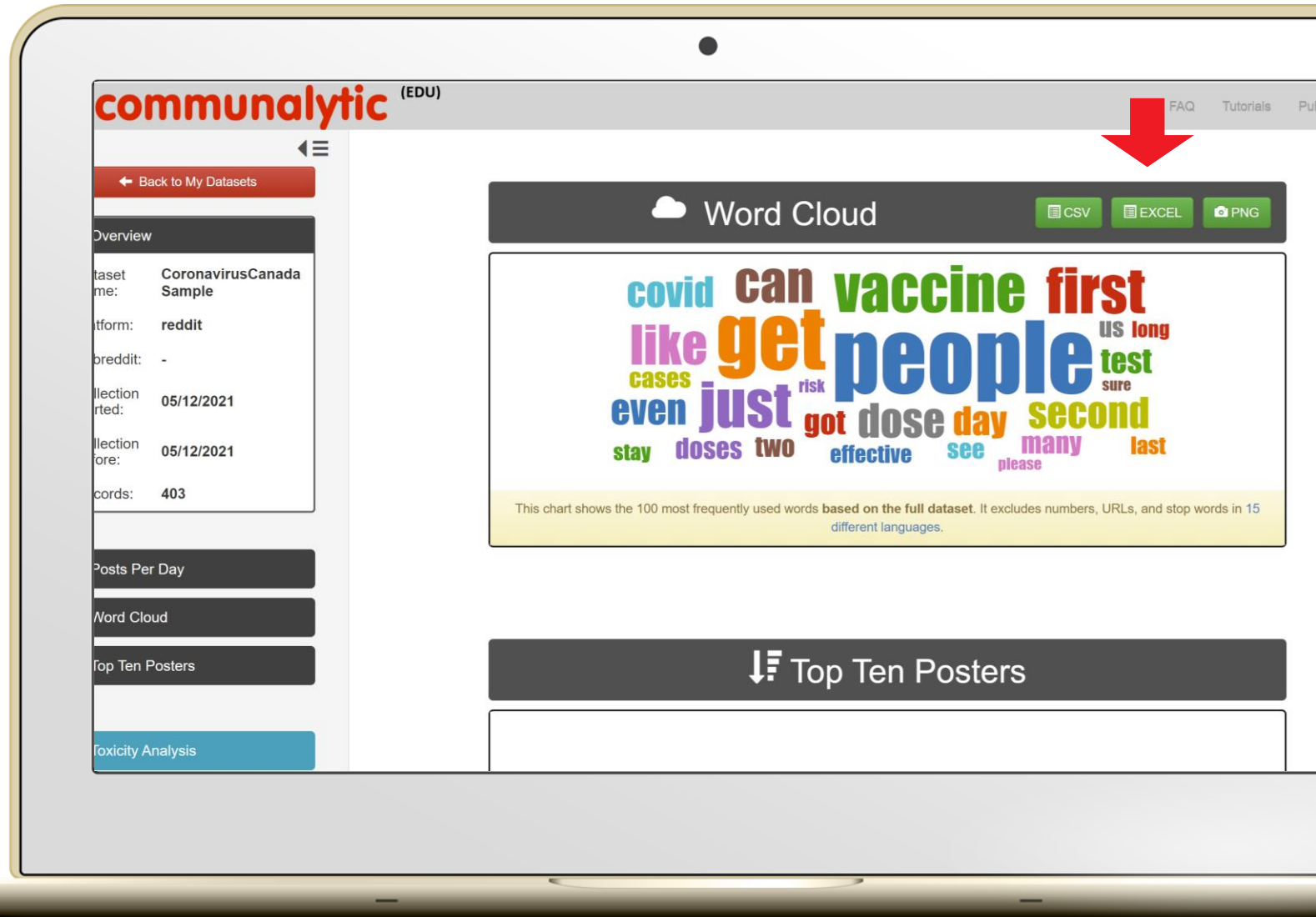
## Hands-On Part 2:

1) Download a sample dataset from [http://bit.ly/reddit\\_sample](http://bit.ly/reddit_sample)

2) Open and review the downloaded dataset in Excel

3) Import the CSV file into Communalytic

4) Review the Overview Charts in Communalytic



## Hands-On Part 2:

- 1) Download a sample dataset from [http://bit.ly/reddit\\_sample](http://bit.ly/reddit_sample)
- 2) Open and review the downloaded dataset in Excel
- 3) Import the CSV file into Communalystic

### 4) Review the Overview Charts in Communalystic

By default, the word cloud is calculated based on the first 1000 posts. To recreate it for the full dataset, click on the “Recreate ...” button (if applicable).



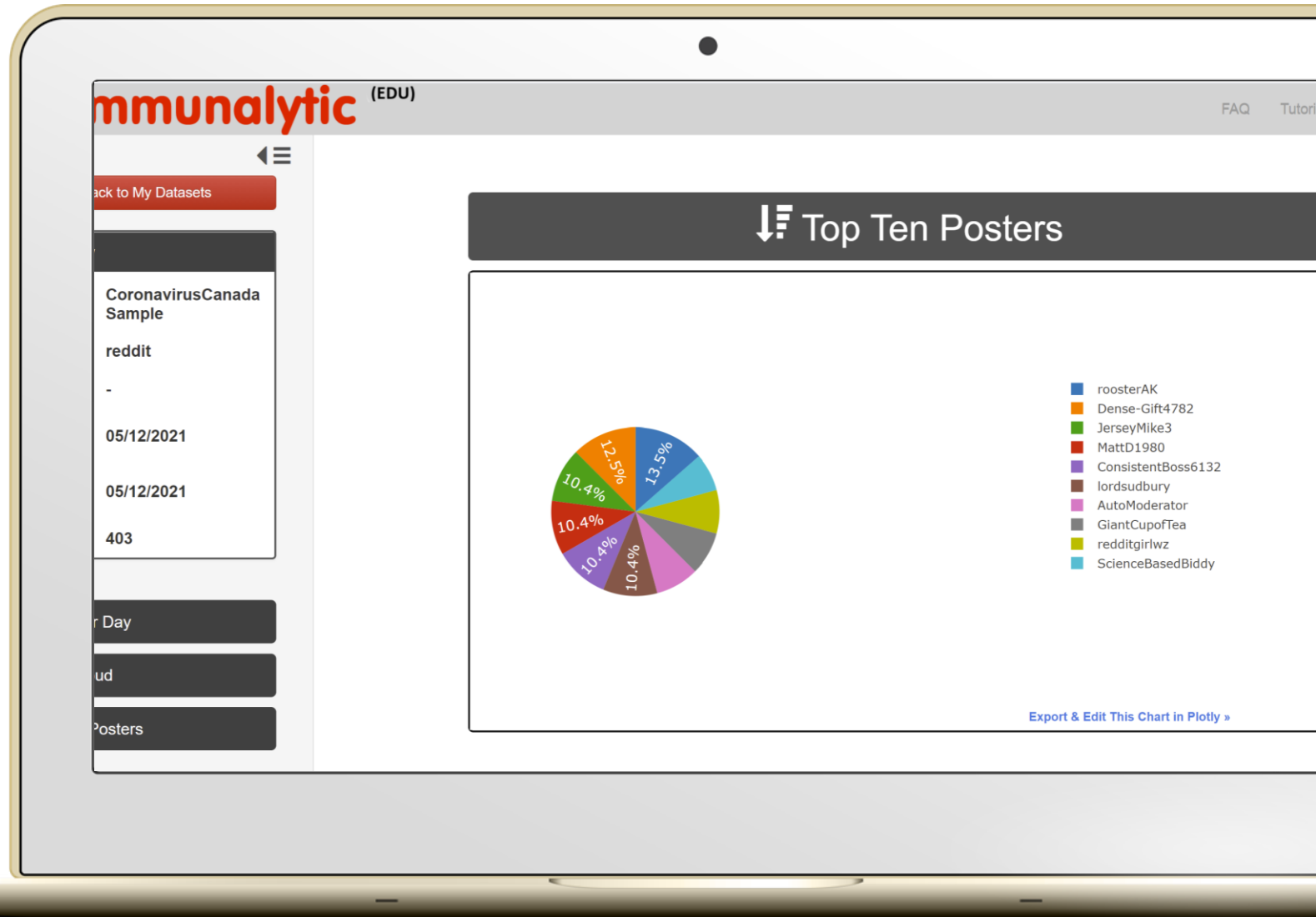
## Hands-On Part 2:

1) Download a sample dataset from [http://bit.ly/reddit\\_sample](http://bit.ly/reddit_sample)

2) Open and review the downloaded dataset in Excel

3) Import the CSV file into Communalytic

4) Review the Overview Charts in Communalytic



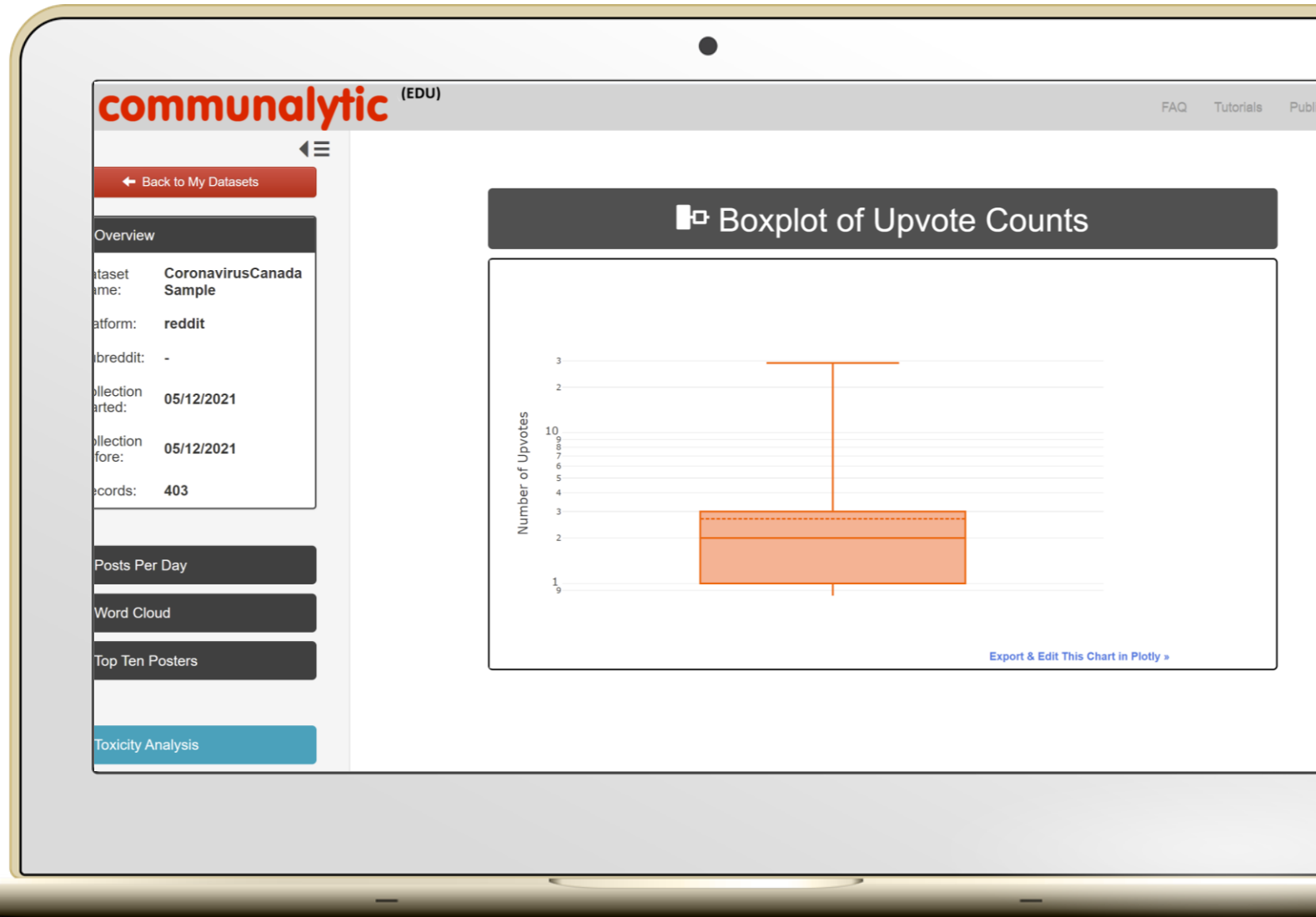
## Hands-On Part 2:

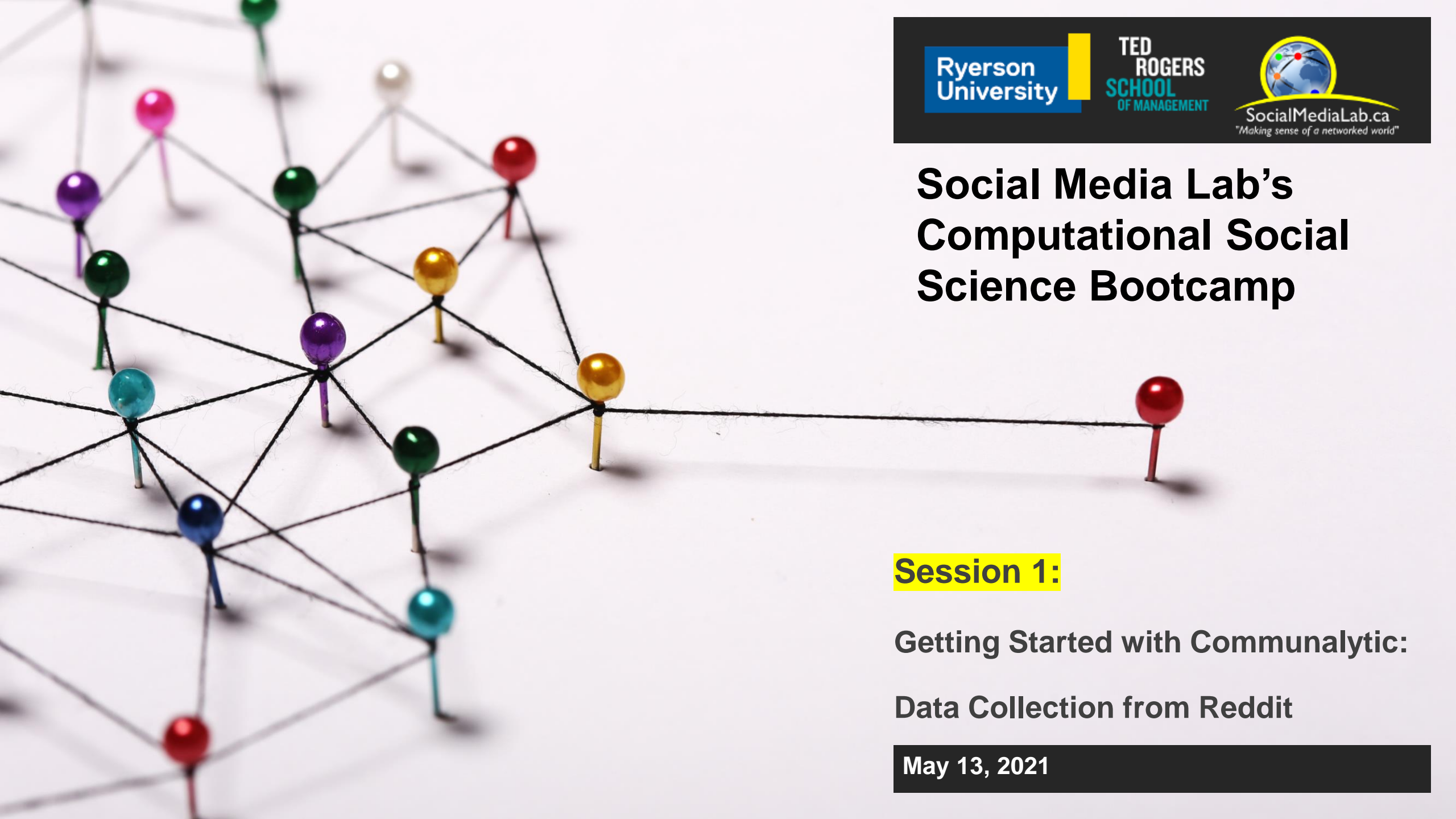
1) Download a sample dataset from [http://bit.ly/reddit\\_sample](http://bit.ly/reddit_sample)

2) Open and review the downloaded dataset in Excel

3) Import the CSV file into Communalytic

4) Review the Overview Charts in Communalytic





Ryerson  
University

TED  
ROGERS  
SCHOOL  
OF MANAGEMENT



# Social Media Lab's Computational Social Science Bootcamp

## Session 1:

Getting Started with CommuAnalytic:

Data Collection from Reddit

May 13, 2021



# References

- Gruzd & Hernández-García. (2018). [Privacy Concerns and Self-Disclosure in Private and Public Uses of Social Media](#). *Cyberpsychology, Behavior, and Social Networking* 21(7), 418–428.
- Dubois, Gruzd, & Jacobson. (2018). [Journalists' Use of Social Media to Infer Public Opinion: The Citizens' Perspective](#). *Social Science Computer Review*.
- Jacobson & Gruzd. (2018). [Employers' Use of Young People's Social Media: Extending Stakeholder Theory to Social Media Data](#). In *Academy of Management Proceedings* (Vol. 2018, p. 18217).
- Jacobson, Gruzd & Hernandez-Garcia. (2019). Who is watching the Watchers? Social Media Marketing in the Post Cambridge Analytica Era. *Journal of Retailing and Consumer Services*.  
<https://doi.org/10.1016/j.jretconser.2019.03.001>
- Gruzd, A., Jacobson, J., Mai, P., & Dubois, E. (2018). [The State of Social Media in Canada 2017](#). Public Report.
- Gruzd, A., Jacobson, J., Mai, P., & Dubois, E. (2018). [Social Media Privacy in Canada](#). Public Report.
- Gruzd & Hernández-García. (2018). Privacy Concerns and Self-Disclosure in Private and Public Uses of Social Media. *Cyberpsychology, Behavior, and Social Networking* 21(7), 418–428.
- Image credit: <https://pixabay.com>