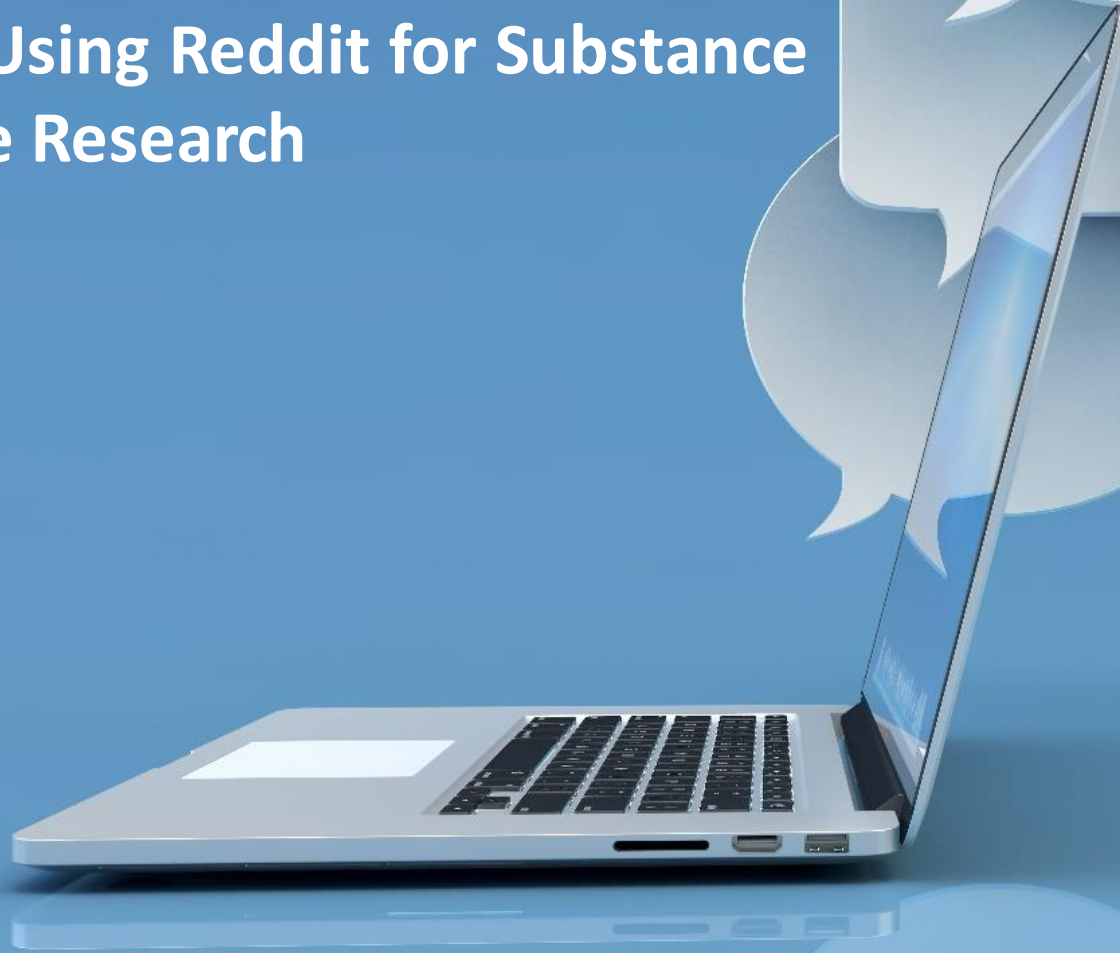



# r/AskAcademia: Special Considerations in the Practice of Using Reddit for Substance Use Research



Savannah Brenneke, MPH  
Meredith Meacham, PhD, MPH  
Amanda Bunting, PhD  
Alexandra DeLucia, MSE  
Nicholas Proferes, PhD

This released version of the presentation has notes included. Notes are linked to relevant slides with a 

College on Problems of Drug Dependence  
Annual Meeting - June 17, 2023

# Disclosures | Funding | Contact

- No financial disclosures or conflicts of interest (MM)
- Research funded by NIH T32 DA007250 and K01 DA046697

- Contact:

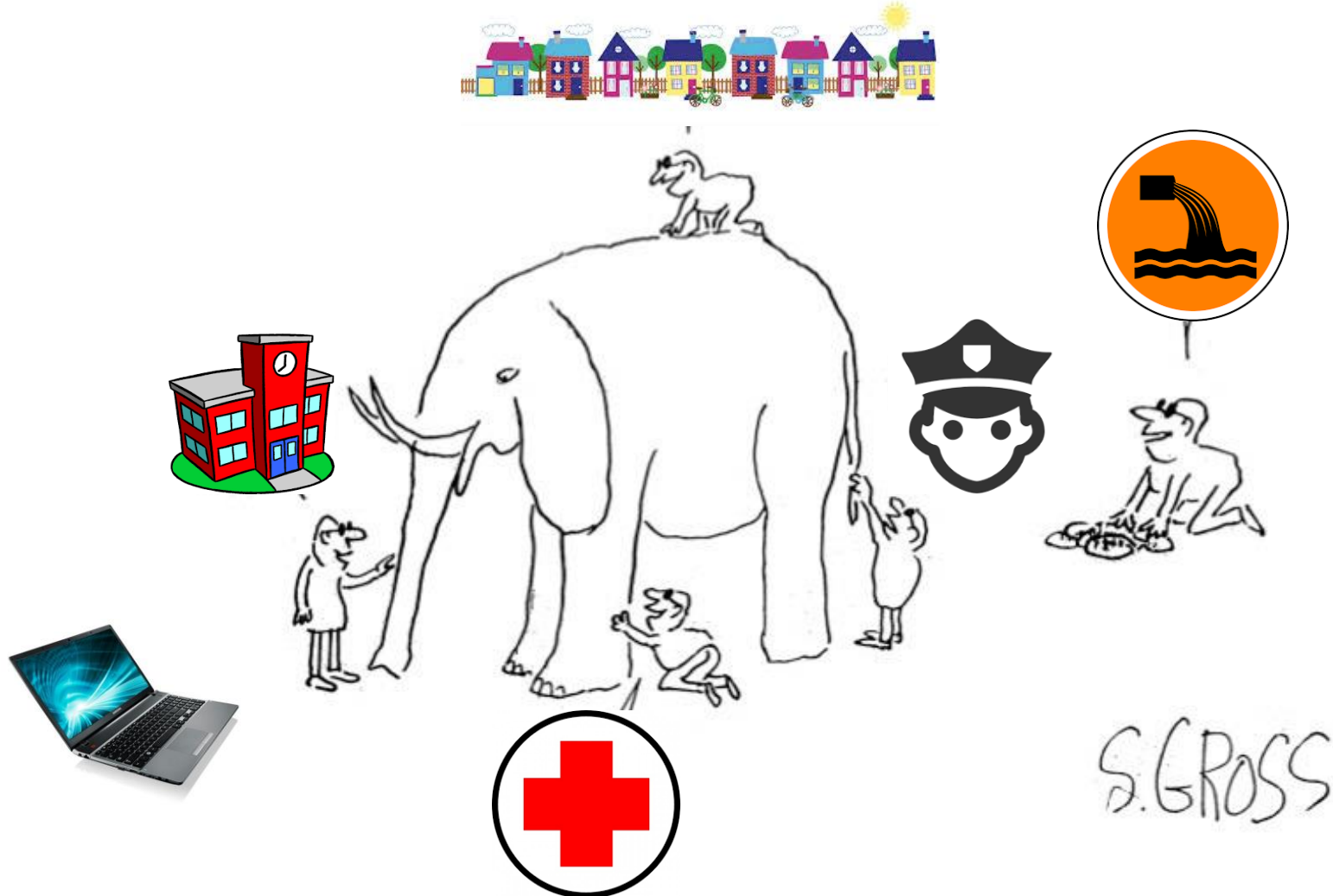
[Meredith.Meacham@ucsf.edu](mailto:Meredith.Meacham@ucsf.edu)

Assistant Professor, Department of Psychiatry & Behavioral Sciences

University of California San Francisco

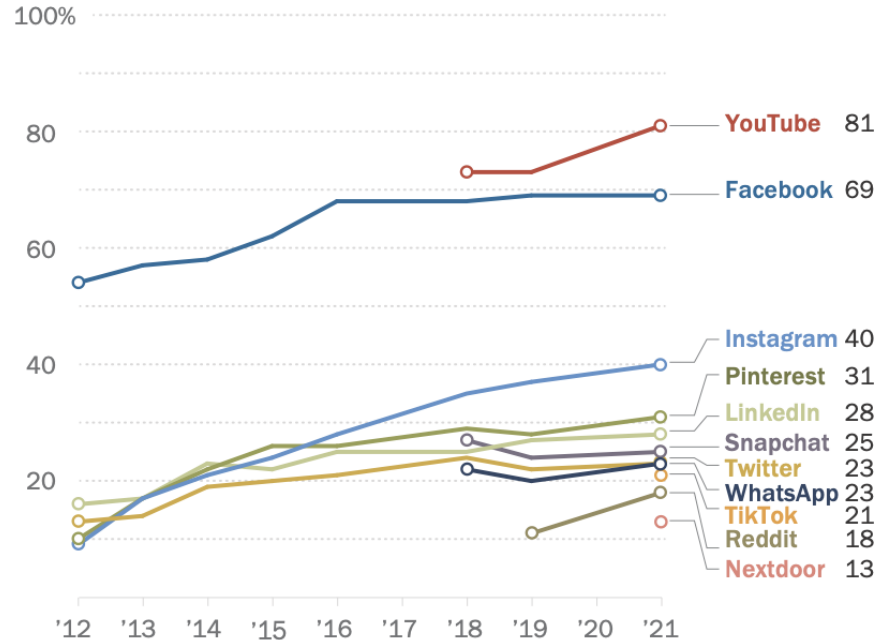
u/ProjectReddibles

# Observational Substance Use Research Data



## Growing share of Americans say they use YouTube; Facebook remains one of the most widely used online platforms among U.S. adults

% of U.S. adults who say they ever use ...



Note: Respondents who did not give an answer are not shown. Pre-2018 telephone poll data is not available for YouTube, Snapchat and WhatsApp; pre-2019 telephone poll data is not available for Reddit. Pre-2021 telephone poll data is not available for TikTok. Trend data is not available for Nextdoor.

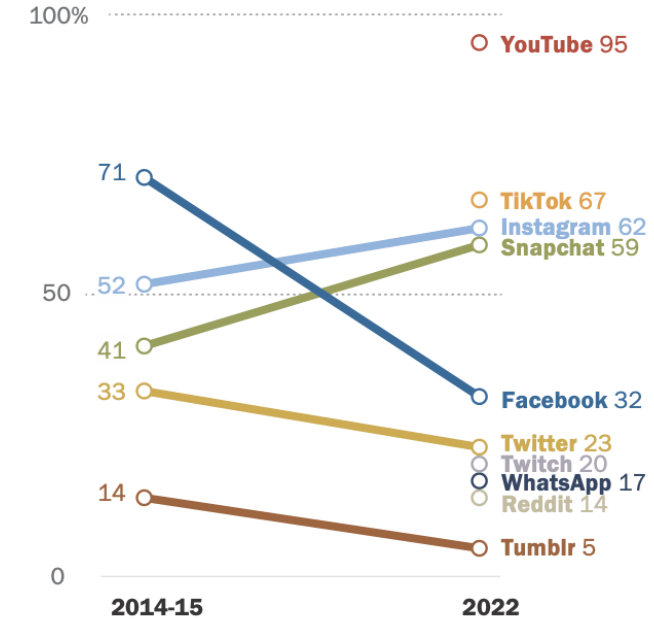
Source: Survey of U.S. adults conducted Jan. 25-Feb. 8, 2021.

"Social Media Use in 2021"

PEW RESEARCH CENTER

## Since 2014-15, TikTok has arisen; Facebook usage has dropped; Instagram, Snapchat have grown

% of U.S. teens who say they ever use any of the following apps or sites



Note: Teens refer to those ages 13 to 17. Those who did not give an answer are not shown. The 2014-15 survey did not ask about YouTube, WhatsApp, Twitch and Reddit. TikTok debuted globally in 2018.

Source: Survey conducted April 14-May 4, 2022.

"Teens, Social Media and Technology 2022"

PEW RESEARCH CENTER

# What's Reddit?



- Social news aggregation, content rating, discussion website founded in 2005
- 6<sup>th</sup> most visited website in the United States, 10<sup>th</sup> most visited worldwide (2023)
- Content-specific communities or forums called “subreddits”
- Pseudo-anonymous contributors and volunteer moderators
- Customizable interaction and engagement
  - subscribing to subreddits
  - commenting (in nested threads)
  - upvotes/downvotes (“karma points”) → visibility



# How I use Reddit



**Australia**

r/australia



**bicycling**

r/bicycling



**Cats**

r/cats



**DataIsBeautiful**

r/dataisbeautiful



ELSEVIER

Drug and Alcohol Dependence

Volume 188, 1 July 2018, Pages 364-369



Full length article

Understanding emerging forms of cannabis use through an online cannabis community: An analysis of relative post volume and subjective highness ratings

Meredith C. Meacham <sup>a</sup>  , Michael J. Paul <sup>b</sup>, Danielle E. Ramo <sup>a</sup>

## PLOS ONE

 OPEN ACCESS  PEER-REVIEWED

RESEARCH ARTICLE

*"I got a bunch of weed to help me through the withdrawals":*  
Naturalistic cannabis use reported in online opioid and opioid recovery community discussion forums

Meredith C. Meacham , Alicia L. Nobles, D. Andrew Tompkins, Johannes Thru

Published: February 8, 2022 • <https://doi.org/10.1371/journal.pone.0263583>

## Benefits of using Reddit



Content specific communities



Open discussion & rich language



Timeliness

## Challenges of using Reddit



Large amounts of data



Ephemeral data



Lack of consistent demographic  
& geographic information



# Who uses Reddit in the United States?

(compared to other social media)

	YouTube	Facebook	Instagram	Pinterest	LinkedIn	Snapchat	Twitter	WhatsApp	TikTok	Reddit	Nextdoor
Total	81	69	40	31	28	25	23	23	21	18	13
Men	82	61	36	16	31	22	25	26	17	23	10
Women	80	77	44	46	26	28	22	21	24	12	16
White	79	67	35	34	29	23	22	16	18	17	15
Black	84	74	49	35	27	26	29	23	30	17	10
Hispanic	85	72	52	18	19	31	23	46	31	14	8
Ages 18-29	95	70	71	32	30	65	42	24	48	36	5
30-49	91	77	48	34	36	24	27	30	22	22	17
50-64	83	73	29	38	33	12	18	23	14	10	16
65+	49	50	13	18	11	2	7	10	4	3	8

Pew Research Center 2021



# Changes and Uncertainties

TECHNOLOGY

## Thousands of Reddit communities 'go dark' in protest of new developer fees

June 12, 2023 · 5:00 AM ET

By Bobby Allyn, Tilda Wilson



Thousands of communities on Reddit are "going dark" in response to changes the company announced that would charge third-party developers for access to its site data.

## COALITION *for* INDEPENDENT TECHNOLOGY RESEARCH


### Signatures & Mutual Aid - Restricting Reddit Data Access Threatens Online Safety & Public-Interest Research

Last week, soon after Reddit announced plans to [restrict free access](#) to the Reddit API, the company [cut off access to Pushshift](#), a data resource widely used by communities, journalists, and thousands of academics worldwide. The impacts of losing access to Reddit data are twofold:

- It will disrupt critical projects from thousands of journalists, academics, and civil society actors who study some of the most important issues impacting our societies today, and
- It will hinder Reddit's volunteer moderators from the vital positions they hold keeping their communities and Reddit's user base safe from harm.



# Workshop Overview

 Activity 1: Exploring Reddit



## Rapid Use of Reddit to Understand Emerging Drug Trends

Amanda Bunting, PhD  
New York University



Activity 2: Reddit  
Research Questions



## r/AskAComputerScientist: Processing Reddit Data for the Social Sciences

Alexandra DeLucia, MSE  
Johns Hopkins University



## Harnessing Reddit: The Methods and Their Limitations in Analyzing Unstructured Data from Social Media

Savannah Brenneke, MPH  
Johns Hopkins University



Activity 3:  
Working with  
Reddit Data



## Ethical Responsibility and Reddit Research: How Contextual Integrity Can Help Guide Practice

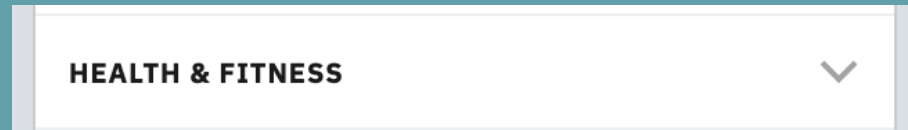
Nicholas Proferes, PhD  
Arizona State University



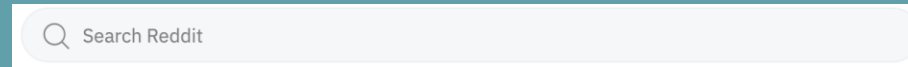
Q & A | Discussion | Further Resources

# Activity 1: Exploring Reddit

Go to Reddit.com or download the Reddit App. Search for a topic or community related to your research interests:



Or explore communities by topic:



Pick a subreddit community and note:

- *What stands out?*
- *How is the community described? When did it start?*
- *How many subscribers does it have?*
- *What kind of posts are there (text, links, image, video)?*
- *What are the subreddit rules?*

# Activity 1 Example: *r/askdrugs*

## About Community

Ask questions about drugs.

 Created Dec 20, 2010

nsfw **Adult content**

**81.9k** **158**  
Members Online

## r/askdrugs Rules

1. Ensure information is accurate and reduces harm
2. Don't discuss places to buy or sell (sourcing)
3. Don't be judgemental, be positive and constructive
4. No drug identification
5. Drug test questions are allowed

**Strictly no requesting, mentioning or giving sources of drugs or paraphernalia**, whether legal or illegal. If in doubt - if your post, or a reply to it would make it easier for someone to get drugs, it's not permitted. This includes sourcing conducted in private messages.

This does not include harm reduction related paraphernalia\*\* such as [testing kits](#), syringes, micron filters and so on.

# **RAPID USE OF REDDIT TO UNDERSTAND EMERGING DRUG TRENDS**

Amanda Bunting, PhD

@ABuntingPhD



# Acknowledgments

- No conflicts of interest to report
- Funding: K01DA053435, R25DA037190
- Collaborators
  - Noa Krawczyk
  - David Frank
  - Marie Bragg & Lab
  - Sam Friedman
  - Yuanqi Gu
  - Thomas Lippincott
  - Meredith Meacham
  - Simran Arya
  - Suhas Nagappala







# Use of Reddit to examine opioid use during COVID-19 lockdowns



Drug and Alcohol Dependence

Volume 222, 1 May 2021, 108672



Socially-supportive norms and mutual aid of people who use opioids: An analysis of Reddit during the initial COVID-19 pandemic

[Amanda M. Bunting](#)<sup>a</sup> , [David Frank](#)<sup>b</sup>, [Joshua Arshonsky](#)<sup>c</sup>, [Marie A. Bragg](#)<sup>c,d</sup>, [Samuel R. Friedman](#)<sup>e</sup>, [Noa Krawczyk](#)<sup>e</sup>

## Informal Coping Strategies Among People Who Use Opioids During COVID-19: Thematic Analysis of Reddit Forums

[Josh Arshonsky](#)<sup>1</sup> ; [Noa Krawczyk](#)<sup>1</sup> ; [Amanda M Bunting](#)<sup>1</sup> ; [David Frank](#)<sup>2</sup> ; [Samuel R Friedman](#)<sup>1</sup> ; [Marie A Bragg](#)<sup>1</sup>



International Journal of Drug Policy

Volume 92, June 2021, 103140



Research paper

*“How will I get my next week's script?”*  
Reactions of Reddit opioid forum users to changes in treatment access in the early months of the coronavirus pandemic

[Noa Krawczyk](#)<sup>a</sup> , [Amanda M. Bunting](#)<sup>b</sup>, [David Frank](#)<sup>c</sup>, [Joshua Arshonsky](#)<sup>d</sup>, [Yuanqi Gu](#)<sup>e</sup>, [Samuel R. Friedman](#)<sup>a</sup>, [Marie A. Bragg](#)<sup>d,e</sup>

## COVID-19–Related Changes to Drug-Selling Networks and Their Effects on People Who Use Illicit Opioids

*Journal of Studies on Alcohol and Drugs*, 84(2), 222–229 (2023).

Article Tools

[David Frank](#), PH.D.,<sup>a,\*</sup> [Noa Krawczyk](#), PH.D.,<sup>b</sup> [Joshua Arshonsky](#), M.A.,<sup>c</sup> [Marie A. Bragg](#), PH.D.,<sup>c,d</sup> [Sam R. Friedman](#), PH.D.,<sup>c</sup> & [Show All...](#)  
[+ Affiliations](#)

## Fentanyl and other synthetic opioid overdose involved fatalities have been rising since 2013


FAKE PRESCRIPTION PILLS • WIDELY AVAILABLE • INCREASINGLY LETHAL

DEA LAB TESTING REVEALS THAT

**4** OUT OF EVERY **10** PILLS

WITH FENTANYL CONTAIN A POTENTIALLY

**LETHAL DOSE**



Counterfeit pills often contain fentanyl and are more lethal than ever before

**FENTANYL IS CAUSING OVERDOSE DEATHS**

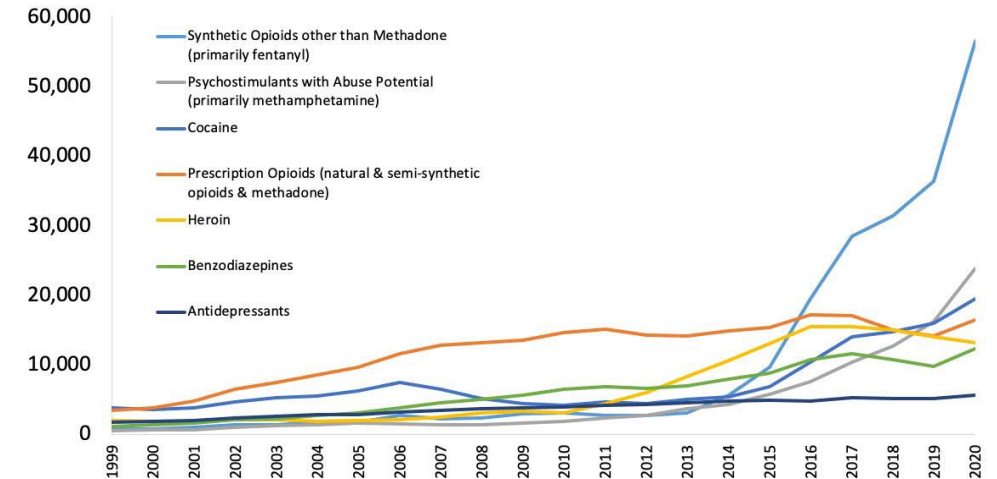
**ANYONE USING HEROIN, COCAINE, OR CRACK, EVEN OCCASIONALLY, IS AT RISK.**

Prevent opioid overdose:

- Carry naloxone. Naloxone can reverse an opioid overdose.
- Avoid mixing drugs. Mixing opioids with alcohol and Xanax, Valium and Klonopin increases the risk of overdose.
- Avoid using alone. If you do, have someone check on you.

The best way to prevent an overdose is to not use drugs.

**Figure 2. National Drug-Involved Overdose Deaths\*, Number Among All Ages, 1999-2020**






\*Includes deaths with underlying causes of unintentional drug poisoning (X40–X44), suicide drug poisoning (X60–X64), homicide drug poisoning (X85), or drug poisoning of undetermined intent (Y10–Y14), as coded in the International Classification of Diseases, 10th Revision. Source: Centers for Disease Control and Prevention, National Center for Health Statistics. Multiple Cause of Death 1999-2020 on CDC WONDER Online Database, released 12/2021.

# Use of Reddit to understand how persons who drugs were experiencing the rise of fentanyl in the drug supply

Original Articles

## Fentanyl in Pressed Oxycodone Pills: A Qualitative Analysis of Online Community Experiences with an Emerging Drug Trend

Simran Arya, Suhas Nagappala, Noa Krawczyk , Yuanqi Gu, Meredith C. Meacham & Amanda M. Bunting  

Pages 1940-1945 | Published online: 15 Sep 2022

[Trends in fentanyl content on Reddit substance use forums, 2013-2021](#)

Accepted

Journal of General Internal Medicine  Springer

# Is your research question appropriate for Reddit data?

- Accelerated data collection
  - Overcomes some of the limitations of the scientific process, data collection
- Real time
  - Data can be collected historically or current
- Subpopulation behaviors that are organic
  - Less likely to have social desirability effects, stigma reduced
  - Explore phenomenon that we are ethically unable to explore
- Resources to investigate are otherwise limited
  - Only if all the above are still important

# Example research questions

- What are the motives of substance use? (e.g., [Pestana et al., 2021](#); [Sharma et al., 2017](#))
- How has substance use (and characteristics of use) changed over time? (e.g., Bowen et al., 2019; [Bunting et al., 2023](#); [Meacham et al., 2018](#))
- How do persons who drugs resonate with televised portrayals of substance use? (e.g., [Kaufman et al., 2020](#))
- How do persons who drugs support each other? (e.g., [D'Agostino et al., 2017](#); [Bunting et al., 2021](#); [Rhidenour et al., 2021](#))



# National Drug Early Warning System

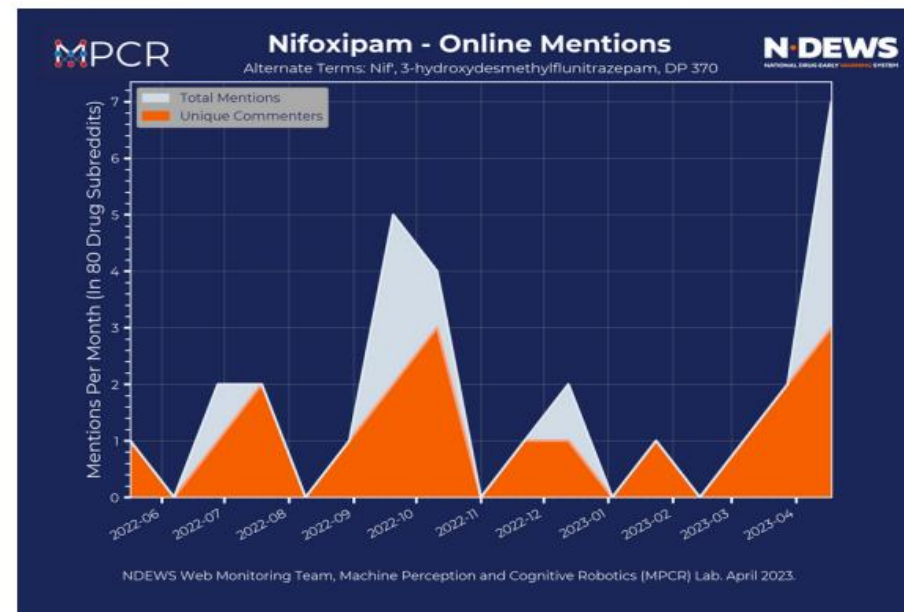
## Web Surveillance

**Social media platforms** provide a unique indicator of activity in the rapidly changing market for new psychoactive substances (NPS). On the popular Reddit website, drug-related discussion occurs in user-generated subreddits dedicated to a drug, a class of drugs, or more general topics related to drug use or experimentation. The contents of drug subreddit discussions can be useful for estimating temporal trends in NPS use, including early and real-time identification of emerging drugs.

In collaboration with the [Machine Perception and Cognitive Robotics \(MPCR\) Lab](#), under the direction of Dr. Daniel Barenholtz, PhD candidate Paul Morris, and PhD student Daniel Van Zant the NDEWS Coordinating Center developed a web monitoring platform for early detection of NPS in drug subreddits. Trends in drug discussion are quantified through anonymized, aggregate keyword counts derived from algorithmic monitoring of ~80 drug-oriented subreddits. Keyword metrics count mentions over time of keywords that refer to a drug. Machine learning models are employed for automated detection and aggregation of drug keywords.

In combination, these methods detect drug-related activity that is anomalous and potentially indicative of an emerging trend in the development or use of novel substances. [Validation on historical trends](#) reveals that the detection of an NPS in drug subreddit discussion is predictive of its subsequent emergence in toxicology and other real-world signals. Early detection of NPS trends by web monitoring serves as a source for further investigation and collaboration with NDEWS partners.

<https://ndews.org/novel-surveillance/web-surveillance/>



### Alert from the NDEWS Web Monitoring Team: Online mentions of Nifoxipam

**What was found?** The drug nifoxipam has had very sparse activity in drug-related subreddits, with the largest increase occurring over the past two months. Before this point, the drug had consistent low activity online. Based on online activity, this potentially indicates an early signal for a substance that could replace etizolam or clonazepam as these drugs are beginning to come under tighter control.

**What is Nifoxipam?** Nifoxipam is a metabolite of the benzodiazepine flunitrazepam. The drug produces strong tranquilizing and sleep-prolonging effects. It is one of many "designer benzodiazepines" currently available.

**How is it being discussed?** Most commenters on Reddit discuss using the drug as a replacement for etizolam or clonazepam. Reported motivations for use have included easier availability, lower perceived legal risk, lower perceived toxicity, and a weaker effect.

**Drug Terms:** Nifoxipam, Nif, 3-hydroxydesmethylflunitrazepam, DP 370

[amanda.bunting@nyulangone.org](mailto:amanda.bunting@nyulangone.org)

[amandabunting.com](http://amandabunting.com)

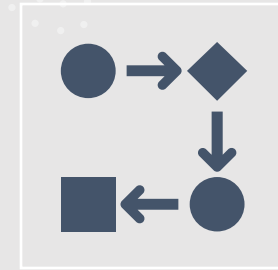
[@ABuntingPhD](#)



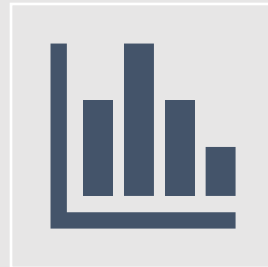
# Activity 2: Reddit Research Questions



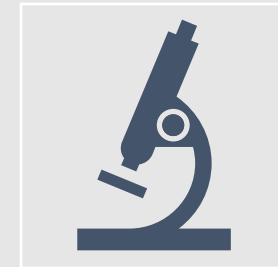
Do you want to focus on a single community or many?



What time periods are you interested in?



Are your interests qualitative, quantitative, or both?



What has been published on this topic (biomedical, social sciences, computer sciences)?

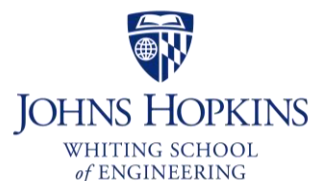
# r/AskAComputerScientist: Processing Reddit Data for the Social Sciences

Alexandra DeLucia, MSE

PhD Candidate

Center for Language and Speech Processing

Department of Computer Science, Johns Hopkins University



Please direct all questions or comments to [✉ aadelucia@jhu.edu](mailto:aadelucia@jhu.edu) or [🐦 @alexir563](https://twitter.com/alexir563)

# Overview

↑ 1 ↓ Tenets of Data Collection

↑ 2 ↓ Anatomy of a Reddit Thread

↑ 3 ↓ Processing Reddit Data

↑ 3.1 ↓ Obtaining + Future of APIs

↑ 3.2 ↓ Storing

↑ 3.3 ↓ Cleaning

↑ 4 ↓ TL;DR + AMA

# Disclosures

**No conflicts of interest to report**

## Scale

How many posts, users, subreddits, etc, are needed?

# Tenets of Data Collection

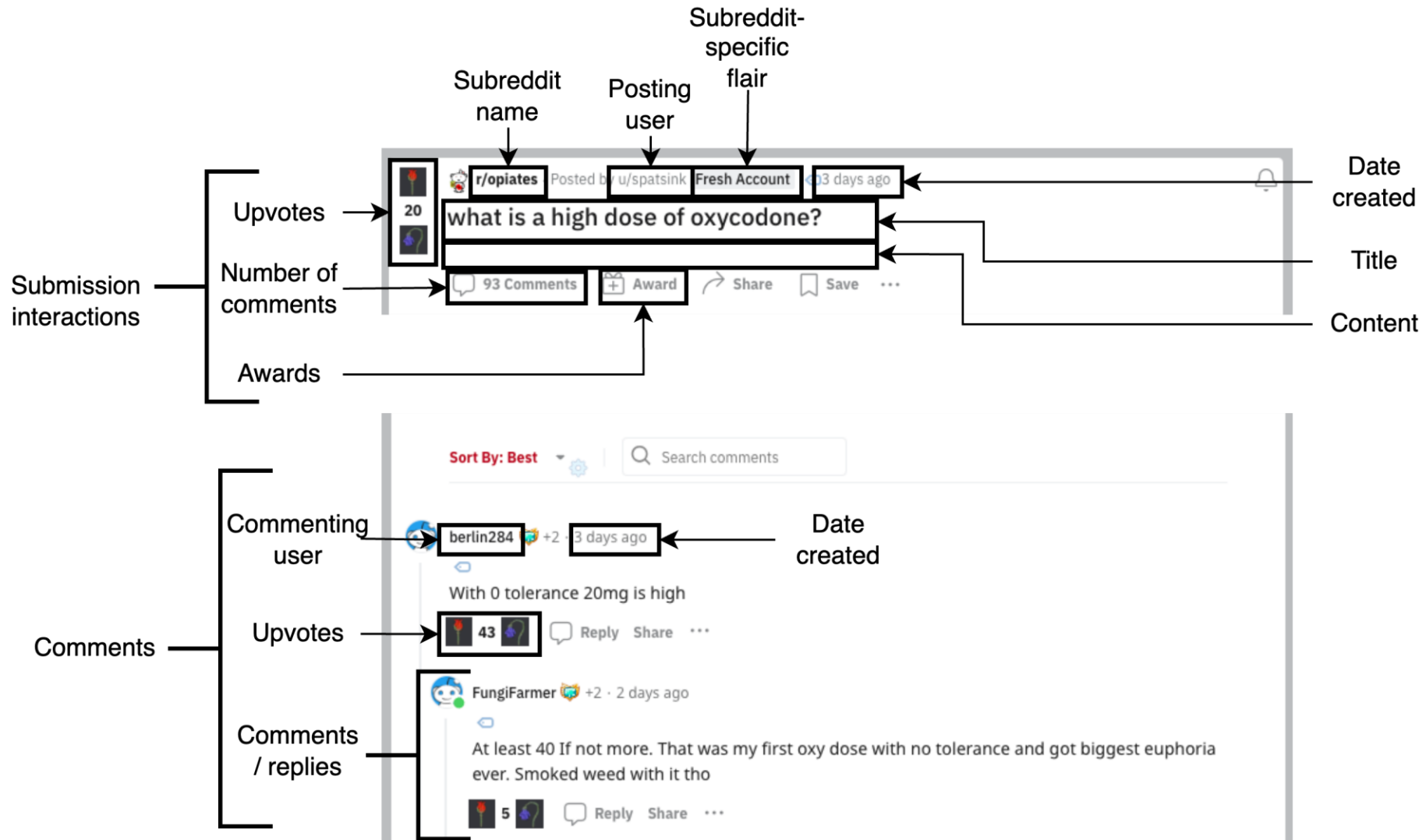
## Time

How much time is allotted for the collection?

## Resources

Which compute resources are available, for processing and storing?

# Anatomy of a Reddit Thread



# Anatomy of a Reddit Thread...in code\*

```
{
  'id': '141udhx',
  'permalink': '/r/opiates/comments/141udhx/what_is_a_high_dose_of_oxycodone/',
  Author: {
    'author': Redditor(name='spatsink'),
    'author_flair_text': 'Fresh Account',
  },
  Submission content: {
    'title': 'what is a high dose of oxycodone?',
    'selftext': '',
    'is_self': True,
    'url': 'https://www.reddit.com/r/opiates/comments/141udhx/what_is_a_high_dose_of_oxycodone',
  },
  Submission interactions: {
    'score': 21,
    'ups': 21,
    'upvote_ratio': 0.89,
    'num_comments': 93,
    'num_crossposts': 0,
    'num_duplicates': 0a,
    'num_reports': None,
  },
  Comments: {
    '_comments': ...,
  },
  Other: {
    'over_18': False,
    'edited': False,
    'gilded': 0,
    'gildings': {},
    ..
    ..
  }
}
```

\*This is a JavaScript Object Notation (JSON) object, it is a way to store structured data. Similar to comma-separated values (CSV).



Obtain

Store

Clean

Processing Reddit Data

# Obtaining Reddit Data

Two main methods for obtaining Reddit data

## 1. Application Programming Interfaces (APIs)\*

- a. Pushshift.io
- b. Official Reddit API

## 2. Manually (via Reddit Search bar)

- a. Screenshots
- b. Copy/paste

```
import praw

reddit = praw.Reddit(
    client_id="my client id",
    client_secret="my client secret",
    user_agent="my user agent",
)
```

\*API access is currently undergoing changes due to Reddit policy

# Obtaining Reddit Data

APIs

## Reddit API

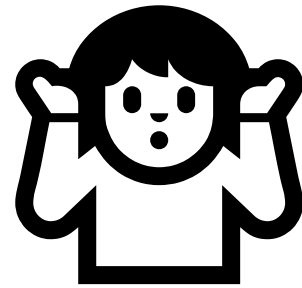
- Allows for search, collecting specific posts and comments, collecting subreddit-specific posts
- Contains up-to-date post attributes (i.e., upvotes)

## Pushshift.io

- Third-party (i.e., unaffiliated) collection and host of Reddit data
- Large "dumps" available containing all comments/submissions for all subreddits per month (roughly 30GB per compressed file)
- Search for subreddits, keywords, and specific date ranges, and other archived posts



# Future of Reddit API



# Future of Reddit API

- Official Reddit API Access for researchers and third-party apps is changing this year. User access remains the same at no cost.
- Effective July 1, the free tier will allow 100 queries (requests) per minute
- Pushshift.io is currently down will potentially allow access for moderators and researchers
- Adult-only (NSFW) subreddits and posts could be restricted, which includes some substance-use communities



# Storing the Data

How to Store?



## In a Database

- MySQL, Apache Solr
- Re-running analysis with various settings is easier (e.g., specific date ranges, excluding users)
- Learning curve for storing and querying

## In Files

- CSVs, JSON, etc
- Requires programming knowledge for efficient reading
- Iterate through all files for each analysis

# Storing the Data

Tips

- Store *compressed* files (e.g., .gz)
  - Text is significantly smaller than images and videos
- Only save the needed information
  - Not all attributes are needed (e.g., "author flair")
  - Thorough cleaning can reduce storage needs
- Save files grouped in ways they will be frequently accessed
  - e.g., by subreddit, by date, by post (grouped with comments)



# Cleaning the Data

## 1 Filtering unwanted data



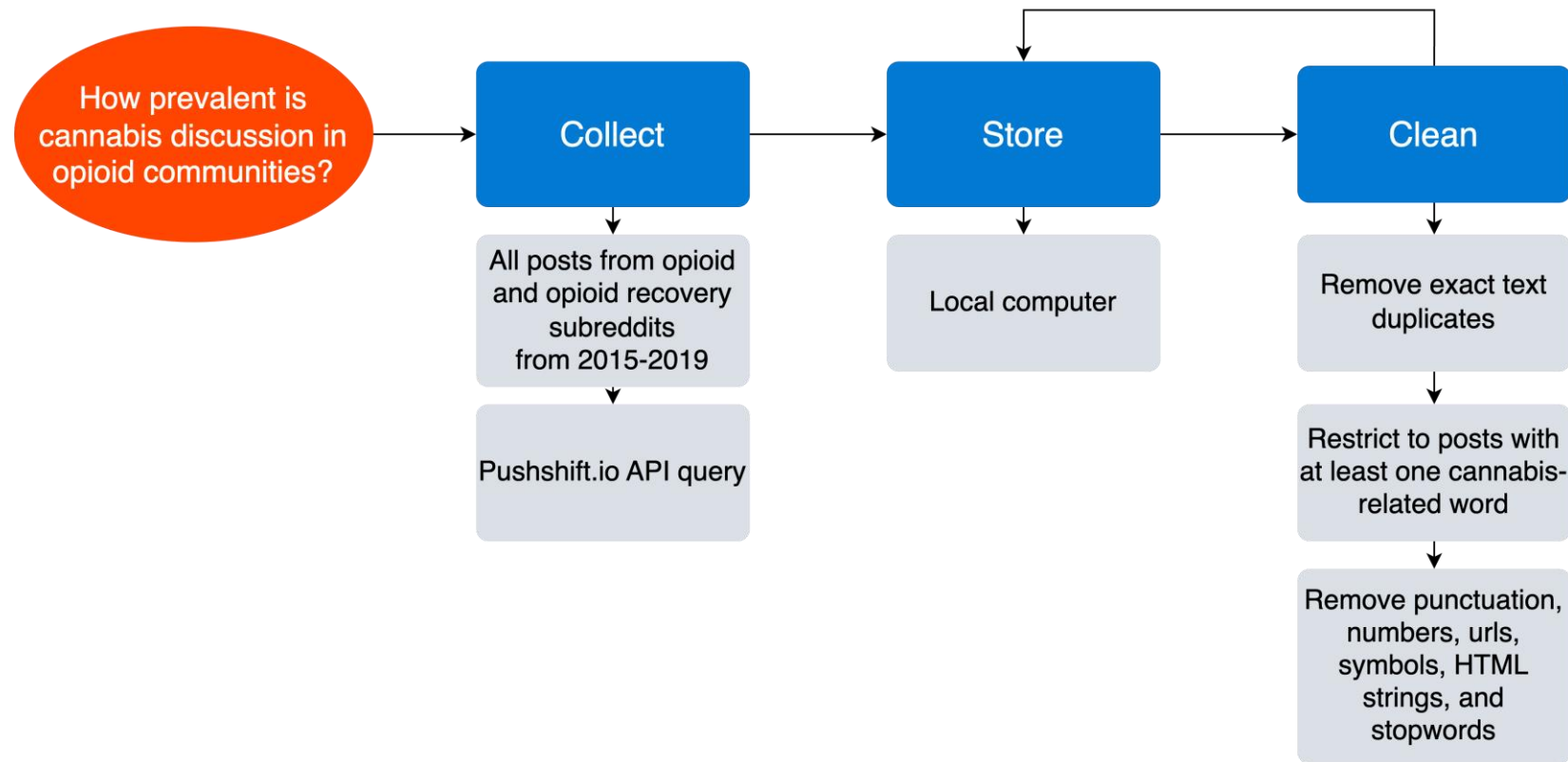
- Spam, duplicates, bots, ads
- Restricting to posts with specific keywords
- Removing posts with images, crossposts

## 2 Post-level processing



- Removing HTML and Markdown formatting code
- Stem words for frequency stats
  - e.g., {smoking, smoked, smoke} -> smok

# Example Process

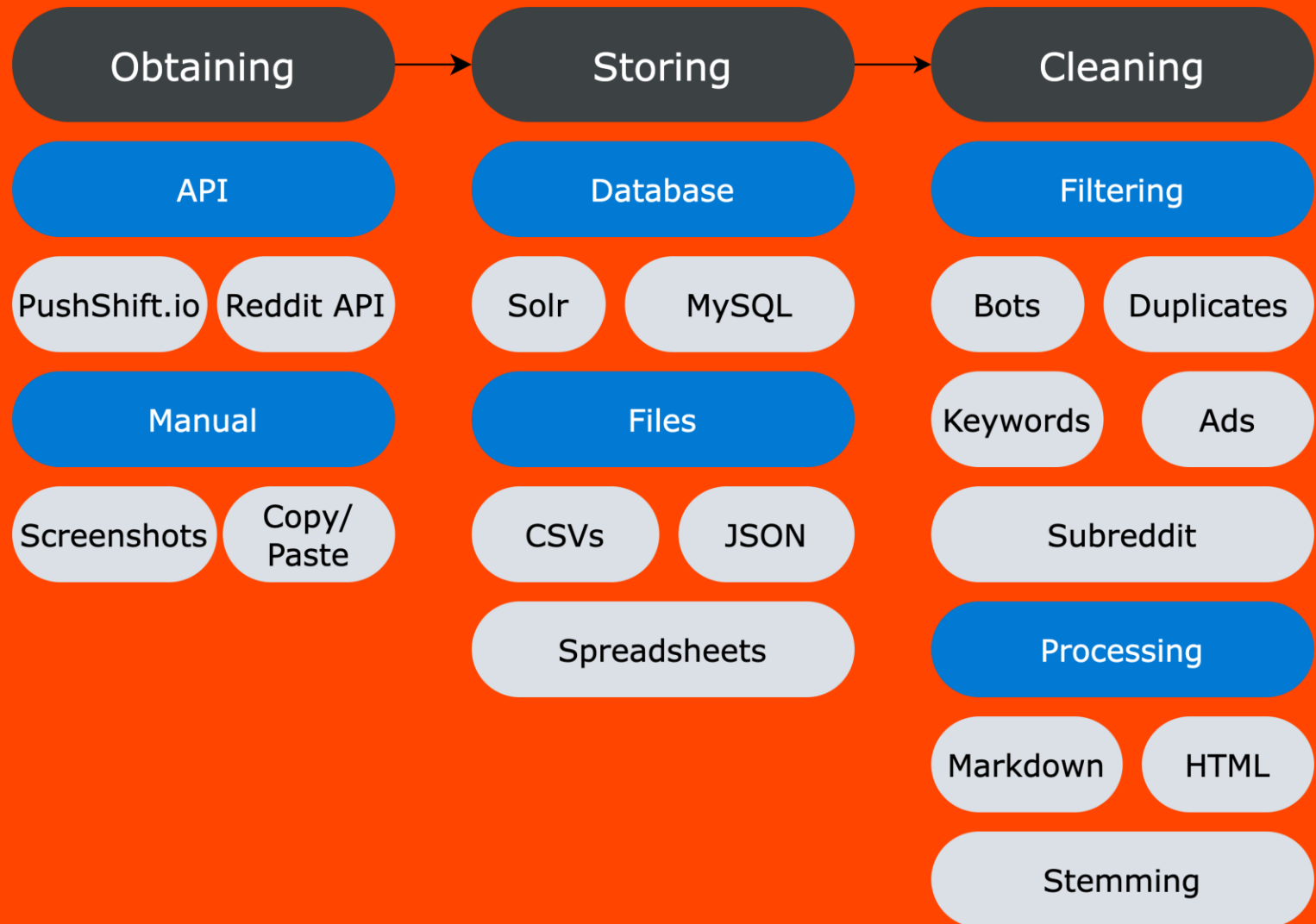


Meacham MC, Nobles AL, Tompkins DA, Thrul J (2022)

**“I got a bunch of weed to help me through the withdrawals”: Naturalistic cannabis use reported in online opioid and opioid recovery community discussion forums.**

PLoS ONE 17(2): e0263583. <https://doi.org/10.1371/journal.pone.0263583>

# TL;DR





JOHNS HOPKINS  
BLOOMBERG SCHOOL  
of PUBLIC HEALTH

# Harnessing Reddit: The Methods and Their Limitations in Analyzing Unstructured Data from Social Media

Savannah G. Brenneke, MPH

✉ sbrenne7@jh.edu | 🐦 @sbrennek  
Bloomberg School of Public Health  
Johns Hopkins University



## Disclosures

I have NO financial disclosure  
or conflicts of interest with  
the presented material

## Funding

NIDA T-32 Drug Dependence  
Epidemiology Training Program  
*T32DA007292*

The material in this video is subject to the copyright of the owners of the material and is being provided for educational purposes under rules of fair use for registered students in this course only. No additional copies of the copyrighted work may be made or distributed.

# Overview

- Assumptions in Reddit Data:
  - Representativeness
  - Generalizability
- Quantitative Methods
  - Computational
  - Infodemiologic
- Qualitative Methods
  - Focus groups & interviews
  - Observation
- Example – *Meacham, M.C., Nobles, A.L., Tompkins, D.A., & Thrul, J. (2022)*



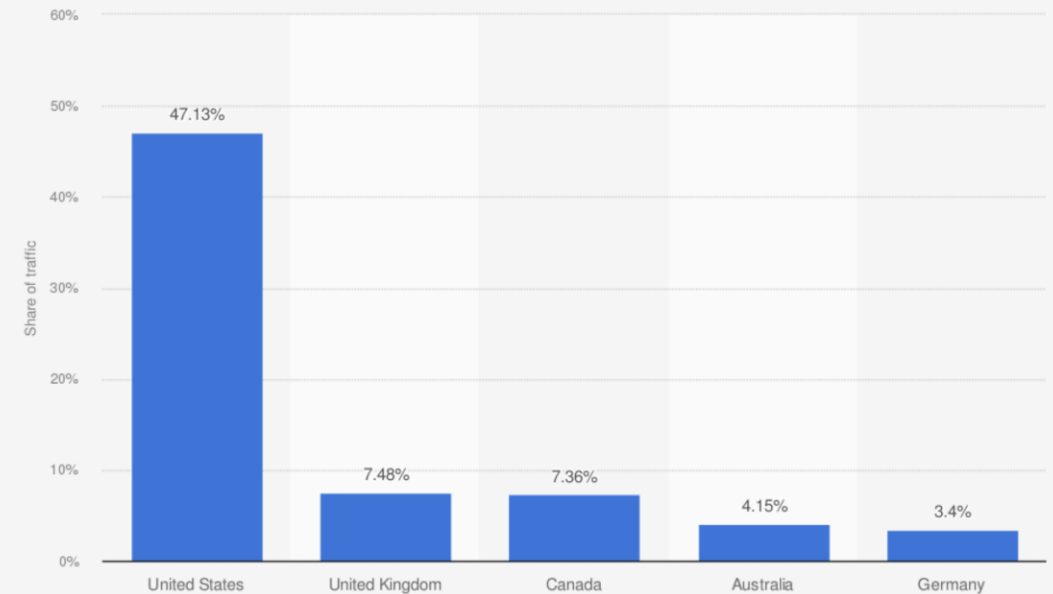
# Assumptions within Reddit



# WHO is using Reddit

- ▶ More than 430 million users MONTHLY
- ▶ >60% are between 18-29 years of age
- ▶ Majority of users are:
  - ▶ MALE
  - ▶ LEFT-LEANING
  - ▶ COLLEGE EDUCATED
- ▶ Over 90% of posts are in ENGLISH \*including non-native speakers

Regional distribution of desktop traffic to Reddit.com as of May 2022 by country



Source  
SimilarWeb  
© Statista 2023

Additional Information:  
Worldwide; 6 months ending May 2022; all traffic

statista

Assumptions

Quantitative

Qualitative

Example



# Estimating Demographic Information

- ▶ Individuals OCCASSIONALLY disclose demographic details
- ▶ Nature of disclosures vary, with disclosures serving certain purposes
  - Narration
  - Establish credibility
  - Relate to topic information
- ▶ Researchers, e.g. Chew et al (2021), have attempted to use computational methods to estimate user demographics

r/saplings • 11 days ago  
by coralushuttkneke

Join ...

## Wish I could actually get STONED

M 40's started trying weed ~4 years ago, and at this point, I've tried everything and different strains and different methods. it does relax me when I smoke or vape and loosen up a bit but that's the limit— my brain/body must have a governor on it that only lets it go past a certain point because I've never just been fkin stoned. I've smoked flower oil shatter budder, consumed gummies and tinctures, low dose high dose you name it. High dose just ends up with me falling asleep. I hear stories about people getting so high I'd love to just space out watching tv and have a good laugh but it just doesn't do that for me. Theory— maybe years of being on daily Vicodin prescription did this? Haven't had an opiate in years but I feel like my brain is still trained for it. If that makes sense, that might be messing up my experience

r/saplings • 18 days ago  
by [redacted]

Join ...

## Any idea what could be making my house (allegedly) smell?

I've (23) been smoking for about 8 years and have lived with my grandparents for (almost) the entire time. I recently moved back in with them and my grandma has started getting up in arms about how I'm stinking the whole house up, but the only difference between now and when I lived here previously when it comes to my smoking habits is that I openly smoke on the porch. When I was younger I was dumb and would smoke in the house, but nothing was ever said about it until one of them saw one of my pieces and said that they didn't care but wanted me to smoke outside. I eventually started taking my stuff out to my car to smoke there and would bring it inside with me after, but nothing was ever said to me about bringing any smell in.

# What people are DOING on Reddit

**About Community**

A community for images, videos, discussions, artwork, and everything dachshund related. Feel free to share your doxie!

Created Dec 6, 2008


341k Members   716 Online   Top 1% Ranked by Size

**r/Dachshund Rules**

1. All content must be related to dachshunds
2. Keep the community wholesome
3. No Impersonation
4. Follow reddiquette, don't ask for upvotes
5. Self-promotion is only allowed for active community members
6. Help by reporting spam

Posted by 18 days ago

2.7k **Rest in Peace** This was my blind doxie, Shadow. He lived for 16.5 years. Please ask me questions about blind doxies and I'll answer what I can.



140 Comments   Share   Save   ...

· 18 days ago

Special needs dogs seem even more loving than usual. I adopted a blind from birth puppy mill dachshund and Chloe was amazing...she started pacing everyday at 4:15 waiting for her 4:30 meal. And she would steal the end of the toilet paper roll and run it out to the living room and chew it up...my other ween got blamed until she was caught

27   Reply   Share   ...

1 more reply

· 18 days ago

No questions but he looks like a beautiful sweet boy

29   Reply   Share   ...

· 18 days ago

Aww thank you. He was definitely a beautiful boy.

14   Reply   Share   ...

· 18 days ago

Ours is fully blind and has lost half hearing now too.. did yours adjust okay to not seeing/hearing?

18   Reply   Share   ...

· 18 days ago

He did fine. We had to pay a bit more attention to him. When he was losing his hearing, we had to stomp a lot so he could feel the vibration of where we were in the house.

21   Reply   Share   ...

· 18 days ago

He is so precious. I took my Doxie to the vet because his eyes are looking a little Smokey and they said there was nothing wrong but I've noticed him acting extremely clingy now that he's 12. I don't know if he's being clingy because his eyesight is not that good. I've seen this change in Behavior where sudden noises like a door creaking or him walking on the wood floor he catches himself like he's going to fall and I see him jumping and freezing up when he's walking on the dark floor. I put down light rugs and it has made an improvement I am wondering if it's his vision

18   Reply   Share   ...

Assumptions

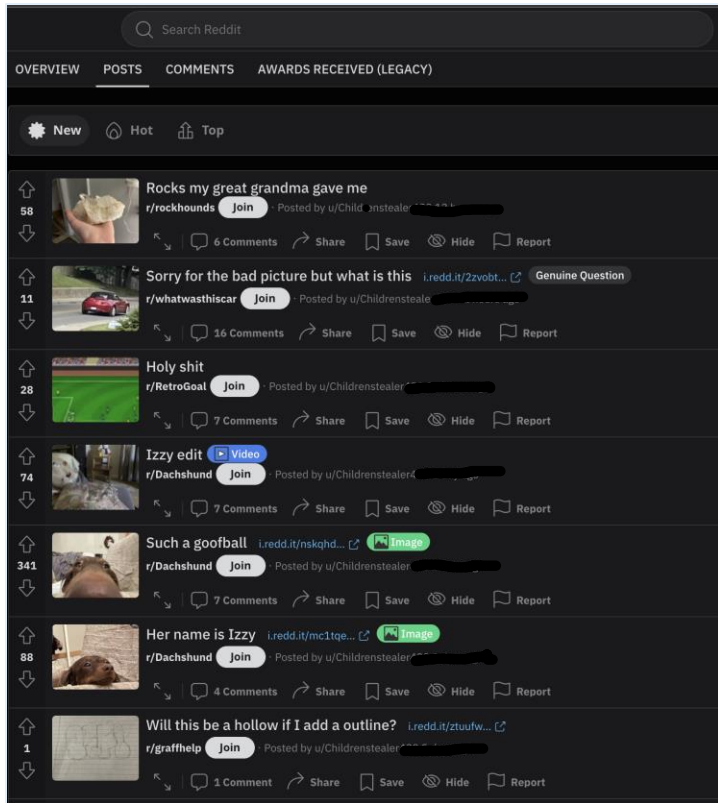
Quantitative

Qualitative

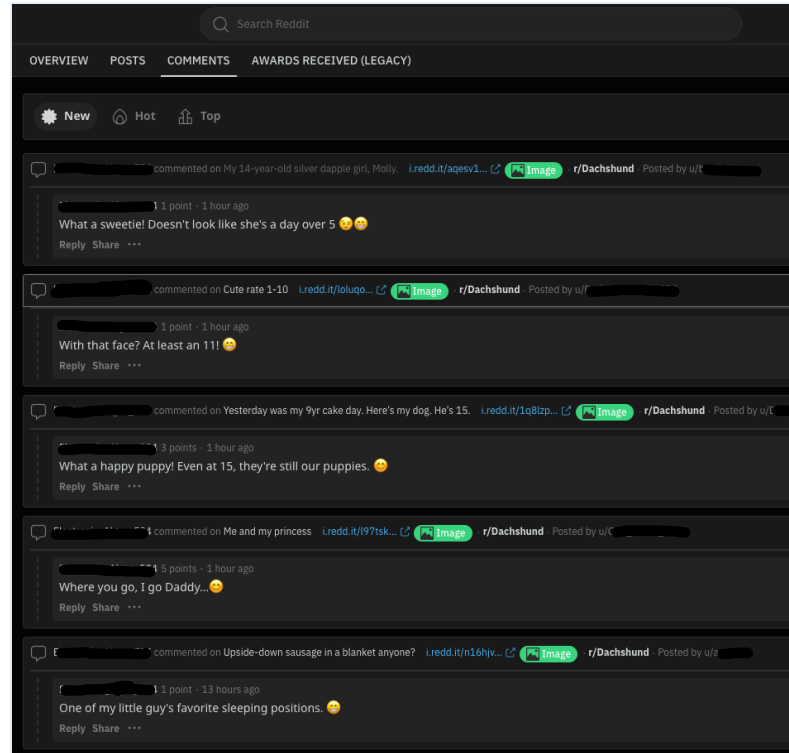
Example

# Redditor Engagement

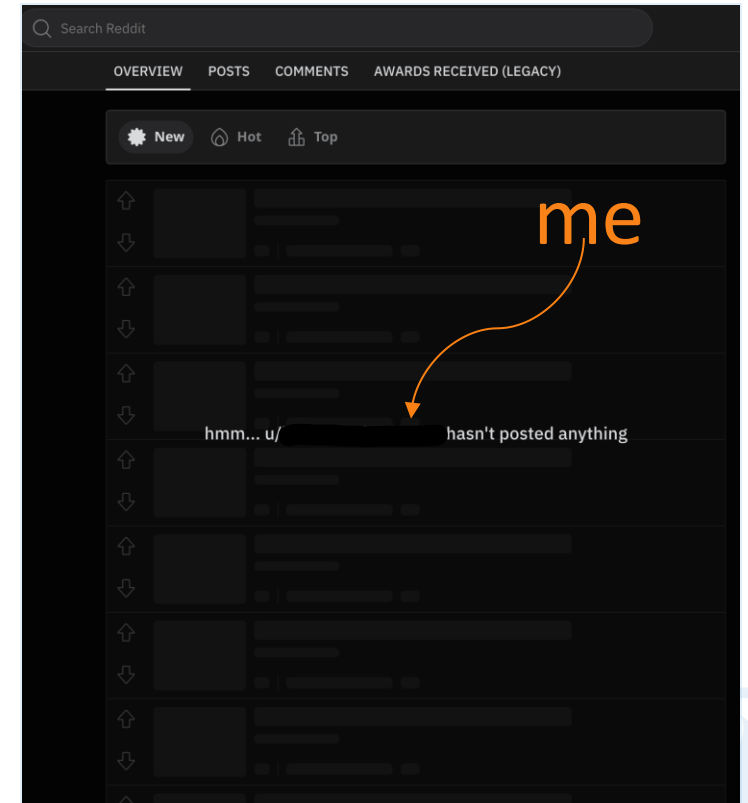
## POSTER



## COMMENTER



## LURKER



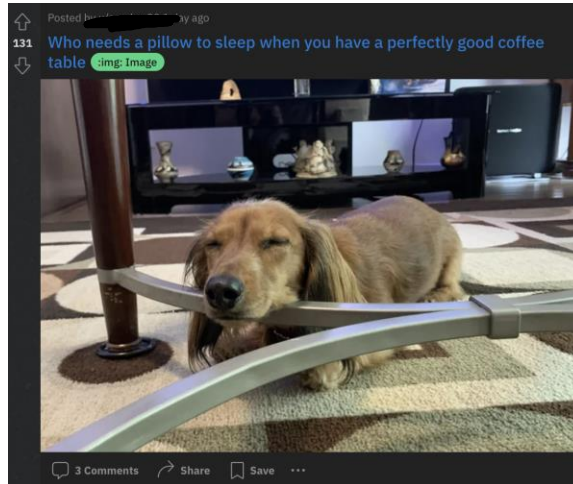
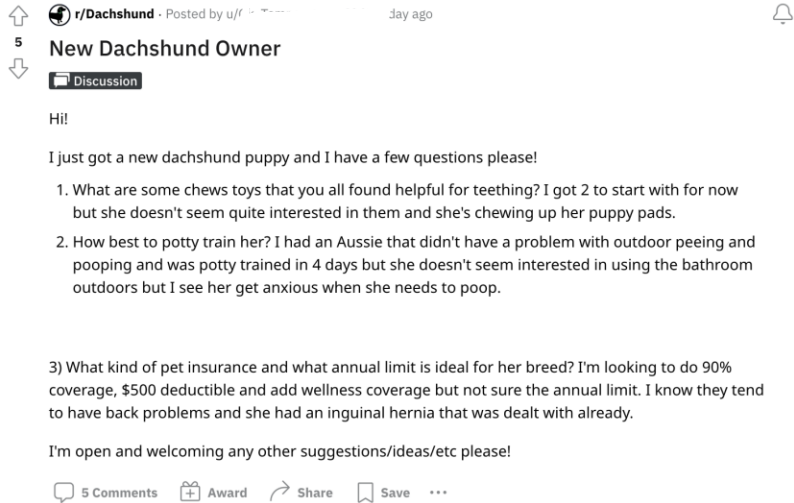
Assumptions

Quantitative

Qualitative

Example

# User Level Active-Engagement: SUBMISSIONS



- ▶ Redditor is initiator and creator of original content shared in community
- ▶ Initiation can be high-stakes for some
- ▶ Submissions as data points may be informative for:
  - ▶ Story-telling / Narration
  - ▶ Advice-seeking
  - ▶ General disclosure
  - ▶ Entertainment/Trend Sharing

Assumptions

Quantitative

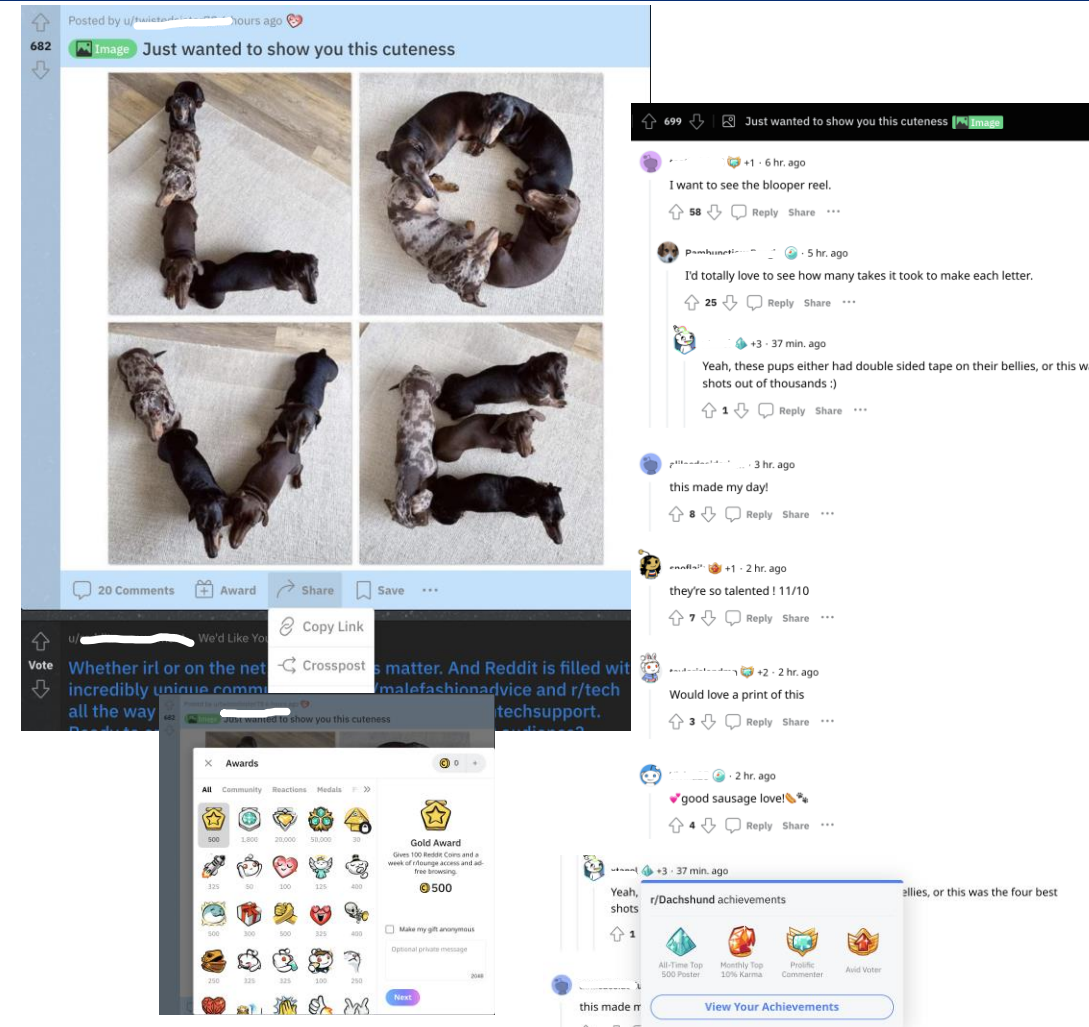
Qualitative

Example



# User Level Active-Engagement: COMMENTING & INTERACTING

- ▶ Multiple levels of engaging without being an "initiator":
  - ▶ Commenting
  - ▶ Up & Down Voting
  - ▶ Gifting
  - ▶ Cross posting
- ▶ Analyses may examine submissions + comments, comments only, other engagement like Karma and Voting
  - ▶ See metadata
  - ▶ Analysis of submissions + comments inform discourse, community, support networks, etc.
- ▶ Special considerations for topic drift\*



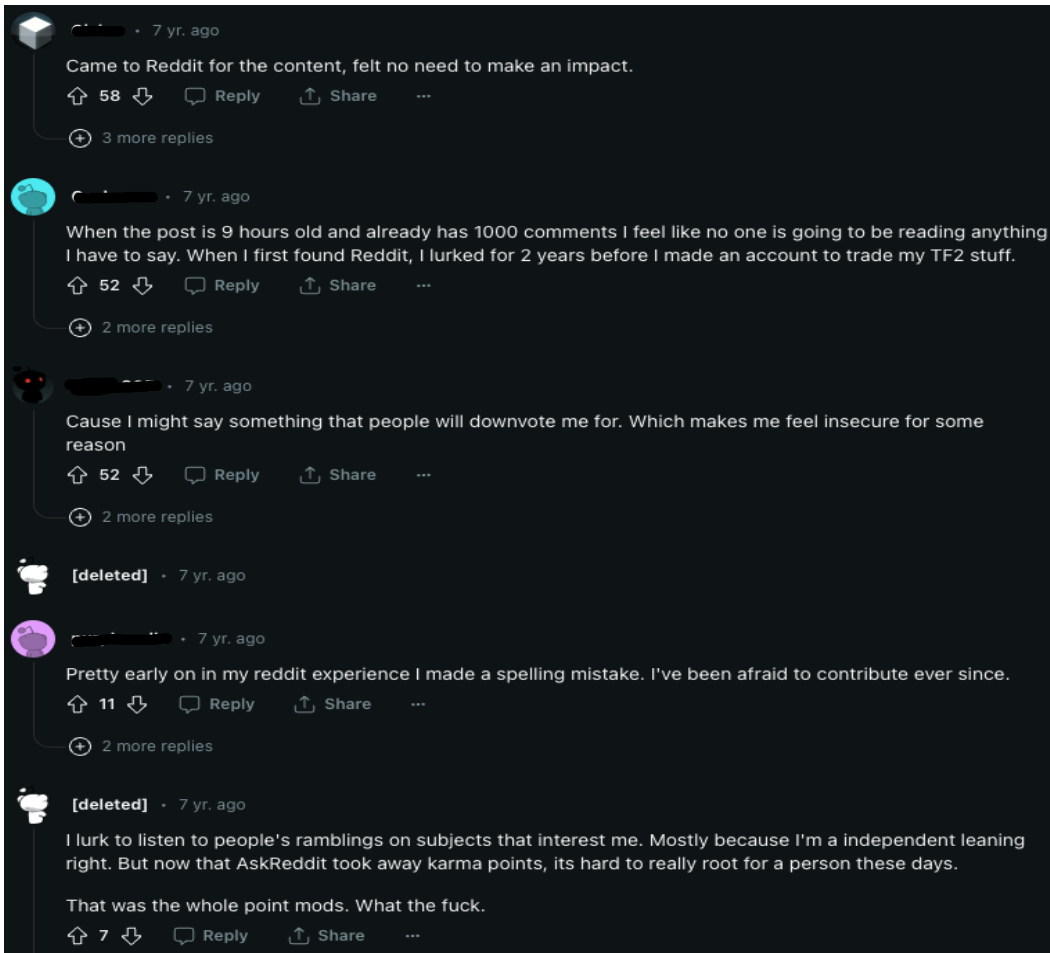
Assumptions

Quantitative

Qualitative

Example

# User Level Passive-Engagement: 'LURKING'



- ▶ Passively consuming information with OR without an account
  - ▶ Browsing
  - ▶ Searching
- ▶ 'lurkers' or passive-consumers cannot be captured in the data
- ▶ Belonging to and content consumption behaviors could be insightful to current or future 'real-life' behaviors

# Why does all this matter?

## Representativeness

- ▶ Reddit user base is unlikely to be representative of populations
  - ▶ Minority
  - ▶ Disconnected
  - ▶ Lower-income
  - ▶ Non-native English speakers
- ▶ Engagement levels have different risks & rewards, i.e. experiences shared/discussed may skew data

## Generalizability

- ▶ Without demographic and geographic information, Reddit data is unlikely to generalize to broader populations
- ▶ Across Reddit generalizability is also limited by
  - ▶ Community rules
  - ▶ Membership
  - ▶ Reward systems



# QUANTITATIVE METHODS





# Computational Methods

## Term frequency – Inverse document frequency (tf-idf)

- Estimation of the most commonly used unique and informative terms within a document and among a collection of documents

$$\text{tf-idf}(t,d) = \text{tf}(t,d) \times \text{idf}(t). \quad \text{idf}(t) = \log \frac{1+n}{1+\text{df}(t)} + 1$$

## Word-embeddings

- Numerical (vector) representation of a word that allows words that occur in similar contexts to have the similar representations.
- E.g., word2vec, GloVe, BERT, ClinicalBERT

## Sentiment analysis

- Context mining of subjective information
- Text classification (e.g., positive, negative, neutral)

# Computational Methods

- ▶ Topic modeling
  - Exploring abstract “topics” within a set of documents
  - Latent Dirichlet allocation (LDA)
- ▶ Network analysis
  - Structural analysis of relationships or processes
- ▶ Predictive modeling
  - Using existing data to forecast future
  - Regression models, decision trees and neural networks



# Infodemiologic Methods

- ▶ Infodemiology coined in early 2000's by Dr. Gunther Eysenbach
- 📎 ▶ “...*the science of distribution and determinants of information in an electronic medium, specifically the Internet, or in a population, with the ultimate aim to inform public health and public policy.*”
- ▶ Supply- and demand-based infodemiology
  - *Supply* = what is published online
  - *Demand* = search and click behaviors
- ▶ Infoveillance as a second arm of infodemiology



# SUPPLY: Information (Concept) Prevalence, Incidence, Co-occurrence



- ▶ Prevalence –
  - Absolute or relative number a term or keyword is present in a pool of information, i.e. Reddit submissions, comments or submissions + comments
- ▶ Incidence –
  - Number of new concept units per unit of time, i.e. incidence of mentions of
- ▶ Co-occurrence –
  - Number of two or more concepts present in the same unit, i.e. two concepts mentioned in a given Reddit submission

# DEMAND: Queries and Clicks

npj | digital medicine

Explore content ▾ About the journal ▾ Publish with us ▾

nature > npj digital medicine > brief communications > article

Brief Communication | [Open Access](#) | [Published: 29 January 2020](#)

## Responses to addiction help-seeking from Alexa, Siri, Google Assistant, Cortana, and Bixby intelligent virtual assistants

[Alicia L. Nobles](#), [Eric C. Leas](#), [Theodore L. Caputi](#), [Shu-Hong Zhu](#), [Steffanie A. Strathdee](#) & [John W. Ayers](#)

[npj Digital Medicine](#) 3, Article number: 11 (2020) | [Cite this article](#)

5553 Accesses

JAMA Network Open

JAMA Network Open

**This Issue** Views 4,157 | Citations 0 | Altmetric 659 | Comments 1

Download PDF

More ▾

Cite This

**Research Letter** | Public Health

June 7, 2023

## Evaluating Artificial Intelligence Responses to Public Health Questions

[John W. Ayers](#), PhD, MA<sup>1,2</sup>; [Zechariah Zhu](#), BS<sup>1</sup>; [Adam Poliak](#), PhD<sup>3</sup>; [et al](#)

[» Author Affiliations](#) | [Article Information](#)

*JAMA Netw Open.* 2023;6(6):e2317517. doi:10.1001/jamanetworkopen.2023.17517

**JMIR Publications**  
Advancing Digital Health & Open Science

Articles ▾ Search articles



Resource Center ▾ Login Register

Journal of Medical Internet Research



Journal Information ▾

Browse Journal ▾

Submit Article

Published on 1.12.2022 in Vol 24 , No 12 (2022) :December

Preprints (earlier versions) of this paper are available at <https://preprints.jmir.org/preprint/41527>, first published July 28, 2022.



## Characterizing Help-Seeking Searches for Substance Use Treatment From Google Trends and Assessing Their Use for Infoveillance: Longitudinal Descriptive and Validation Statistical Analysis

[Patton T](#) <sup>1</sup>; [Daniela Abramovitz](#) <sup>2</sup>; [Derek Johnson](#) <sup>3</sup>; [Eric Leas](#) <sup>4</sup>; [Nobles](#) <sup>1</sup>; [Theodore Caputi](#) <sup>5</sup>; [John Ayers](#) <sup>1</sup>; [Steffanie Strathdee](#) <sup>1</sup>; [Bórquez](#) <sup>1</sup>

### Citation

Please cite as:

Patton T, Abramovitz D, Johnson D, Leas E, Nobles A, Caputi T, Ayers J, Strathdee S, Bórquez A  
Characterizing Help-Seeking Searches for Substance Use Treatment From Google Trends and Assessing Their Use for Infoveillance: Longitudinal Descriptive and Validation Statistical Analysis  
*J Med Internet Res* 2022;24(12):e41527  
doi: [10.2196/41527](https://doi.org/10.2196/41527)  
PMID: [36454620](https://pubmed.ncbi.nlm.nih.gov/36454620/)  
PMCID: [9756118](https://pubmed.ncbi.nlm.nih.gov/9756118/)

JAMA Network

JAMA Internal Medicine

Search All Enter Search Term

**This Issue** Views 2,902 | Citations 8 | Altmetric 147

Download PDF

More ▾

Cite This

Permissions

**Research Letter**

January 14, 2019

## Media Trends for the Substance Abuse and Mental Health Services Administration 800-662-HELP Addiction Treatment Referral Services After a Celebrity Overdose

[John W. Ayers](#), PhD, MA<sup>1</sup>; [Alicia L. Nobles](#), PhD, MS<sup>1</sup>; [Mark Dredze](#), PhD<sup>2</sup>

[» Author Affiliations](#) | [Article Information](#)

*JAMA Intern Med.* 2019;179(3):441-442. doi:10.1001/jamainternmed.2018.6562

THE AMERICAN JOURNAL OF DRUG AND ALCOHOL ABUSE  
2022, VOL. 48, NO. 4, 504-506  
<https://doi.org/10.1080/00952990.2022.2068422>

Taylor & Francis  
Taylor & Francis Group

LETTER TO THE EDITOR

Check for updates

## Substance use-related stigma: an exploratory study of search behavior using Google trends (2004-2021)

[Mike Conway](#) <sup>a,b</sup>, [Cole Citrenbaum](#) <sup>c</sup>, and [Annie T. Chen](#) <sup>d</sup>

<sup>a</sup>School of Computing & Information Systems, University of Melbourne, Parkville, Australia; <sup>b</sup>Centre for Digital Transformation of Health, University of Melbourne, Parkville, Australia; <sup>c</sup>Neuromodulation Division, Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, CA, USA; <sup>d</sup>Department of Biomedical Informatics and Medical Education, University of Washington School of Medicine, Seattle, WA, USA

**ARTICLE HISTORY** Received 6 October 2021; Revised 14 April 2022; Accepted 17 April 2022

**KEYWORDS** Infodemiology [MeSH ID: D000090462]; social stigma [MeSH ID: D057545]; substance-related disorders [MeSH ID: D019966]

Assumptions

Quantitative

Qualitative

Example

# QUALITATIVE METHODS



# Observational Qualitative Methods

- ▶ **CONTENT ANALYSIS: most commonly used qualitative method** – assuming treated as a qualitative
  - Note: quant + qual approaches to content analysis are common
- ▶ Require development of a **codebook** to guide coding of content
  - *Inductive* – ground-up approach
  - *Deductive* – top-down approach
  - A mix of inductive-deductive
- ▶ Others:
  - Discourse analysis
  - Ethnographic observation
  - Grounded theory
  - Case Studies



## *"I got a bunch of weed to help me through the withdrawals":* Naturalistic cannabis use reported in online opioid and opioid recovery community discussion forums

Meredith C. Meacham , Alicia L. Nobles, D. Andrew Tompkins, Johannes Thru

Published: February 8, 2022 • <https://doi.org/10.1371/journal.pone.0263583>

Article	Authors	Metrics	Comments	Media Coverage	Peer Review
⌵					

### Abstract

- Introduction
- Materials and methods
- Results
- Discussion
- Acknowledgments
- References

### Reader Comments

### Figures

### Abstract

A growing body of research has reported on the potential opioid-sparing effects of cannabis and cannabinoids, but less is known about specific mechanisms. The present research examines cannabis-related posts in two large online communities on the Reddit platform ("subreddits") to compare mentions of naturalistic cannabis use by persons self-identifying as actively using opioids versus persons in recovery. We extracted all posts mentioning cannabis-related keywords (e.g., "weed", "cannabis", "marijuana") from December 2015 through August 2019 from an opioid use subreddit and an opioid recovery subreddit. To investigate how cannabis is discussed at-scale, we identified and compared the most frequent phrases in cannabis-related posts in each subreddit using term-frequency-inverse document frequency (TF-IDF) weighting. To contextualize these findings, we also conducted a qualitative content analysis of 200 random posts (100 from each subreddit). Cannabis-related posts were about twice as prevalent in the recovery subreddit ( $n = 908$ ; 5.4% of 16,791 posts) than in the active opioid use subreddit ( $n = 4,224$ ; 2.6% of 159,994 posts,  $p < .001$ ). The most frequent phrases from the recovery subreddit referred to time without using opioids and the possibility of using cannabis as a "treatment." The most frequent phrases from the opioid subreddit referred to concurrent use of cannabis and opioids. The most common motivations for using cannabis were to manage opioid withdrawal symptoms in the recovery subreddit, often in conjunction with anti-anxiety and GI-distress "comfort meds," and to enhance the "high" when used in combination with opioids in the opioid subreddit. Despite limitations in generalizability from pseudonymous online posts, this examination of reports of naturalistic cannabis use in relation to opioid use identified withdrawal symptom management as a common motivation. Future research is warranted with more structured assessments that examines the role of cannabis and cannabinoids in addressing both somatic and affective symptoms of opioid withdrawal.

 EXAMPLE





# Quantitative Methods: Concept prevalence and tf-idf

- ▶ **Concept prevalence**, reported as “proportion of cannabis-related posts”:
  - Key-term search for - “*weed*”, “*cannabis*”, “*marijuana*”, “*pot*”, “*reefer*”, “*ganja*”, “*thc*”, and “*cbd*”
  - Cannabis-related posts overall by subreddit and estimations overtime (per month)
  
- ▶ **tf-idf**, examining most frequent words and phrases in cannabis-related posts
  - Explored top 10 unigram, bigrams and trigrams, via highest tf-idf weighting

	Recovery Subreddit		Opioid Subreddit	
Total Posts (N)	16,791		159,994	
Cannabis Posts (N, % of Total)	908	5.4%	4,224	2.6%
Total Wordcount (Median, IQR)	101	(32, 224)	24	(8, 88)
Cannabis Post Wordcount (Median, IQR)	279	(151, 435)	173	(80, 347)
Change in proportion of posts mentioning cannabis per month (Slope, p-value)	0.014%	0.29	-0.013%	0.0051

IQR: Inter-quartile range.

Table 1. Comparison of posts containing cannabis terms between active opioid use and opioid recovery subreddits (late 2015-mid 2019).

	Top 10 most frequent terms (tf-idf weighting)
Recovery unigrams	marijuana, cbd, clean, opiate, recovery, cannabis, kratom, day, days, weed
Recovery bigrams	day marijuana, weed recovery, zoloft pot, health benefits, medical marijuana, marijuana treat, opiate substitution, substitution thc, benefits cannabis, weed social
Recovery trigrams	opiate substitution thc, health benefits cannabis, weed social drinking, kratom cbd recovery, medical marijuana treat, marijuana treat addiction, health benefits cbd, benefits cbd oil, facts opiate crisis, opiate crisis cbd
Opioid unigrams	opiates, marijuana, cbd, heroin, weed, like, smoke, oxy, just, get
Opioid bigrams	smoke weed, smoking weed, weed opiates, first time, medical marijuana, hydro weed, weed hydro, feel like, anyone else, right now
Opioid trigrams	pill porn weed, coke h weed, combining tramadol weed, going weed withdrawals, weed valium amazing, xanax ambien weed, growing cannabis mistake, bars weed hash, h alc weed, 10mg hydro weed

Table 2. Top 10 most frequent n-grams in cannabis-related posts

# Quantitative Methods: “Keyness” scoring

Table 3

Top 15 most unique unigrams and bigrams in cannabis-related posts.

Top 15 most unique terms (unweighted)	
Recovery unigrams	recovery, clean, day, days, help, years, withdrawal, support, paws, life, relapse, using, months, feel, na
Recovery bigrams	days clean, feel like, cold turkey, w d [withdrawals], every day, per day, cbd oil, months clean, first days, enough enough, physical symptoms, favorite lyrics, hard time, stay strong, comfort meds
Opioid unigrams	bag, guys, guy, black, white, u, lol, fent, fuck, tolerance, dude, weed, shit, opium, dope
Opioid bigrams	front door, cold cop, gas station, r drugs, high like, get addicted, shit like, opiate tolerance, anyone else, living room, opium poppy, stay safe, happy nods, drug test, shit post

- ▶ **Keyness scoring**, estimating terms or words that are unique and important within a subreddit by comparing to the other
- ▶ Recovery subreddit key terms focus on
  - Time spent "clean"
  - Withdrawal & PAWS
  - Relapse
  - Challenges with recovery
- ▶ Larger opioid subreddit key terms
  - Mentions of people
  - Profanity
  - Tolerance
  - References of amounts or effects of drugs

# Qualitative Methods: Content analysis

	Code Definition	Recovery Subreddit (N = 100)	Opioid Subreddit (N = 100)
<b>False Positive</b>	Post is not about cannabis (e.g., "pulling weeds")	2	0
<b>Opioid use disclosure</b>			
No / Not personal / NA	Refers to someone else's use or media depiction	11	22
Yes	Discloses personal use	89	78
Past	Reflection on initial use or use several years ago	6	9
Current	Recent use, just used, or short-term plans to use	9	40
Reducing	In withdrawal, tapering use, or abstinent	74	29
<b>Cannabis use disclosure</b>			
No / Not personal / NA	Refers to someone else's use or to cannabis legalization news	23	32
Yes	Discloses personal use	77	68
Past	Reflection on initial use or use several years ago	14	14
Current	Recent use, just used, or short-term plans to use	52	49
Reducing	Cutting back or intentionally abstinent	11	5
<b>CBD use disclosure</b>	Specifically mentions CBD use	7	2
<b>Motivations</b>			
Withdrawal management	Cannabis is used to manage opioid withdrawal symptoms (e.g., anxiety, nausea, aches, malaise)	43	12
Helpful on its own	States that cannabis is helpful	9	2
Helpful with other strategies	Describes several strategies, including cannabis	22	10
Unclear	Unclear how helpful cannabis is for withdrawal	10	0
Not helpful	States that cannabis is not helpful	2	0
Questioning compatibility	Questions about effectiveness or social acceptability of cannabis use during opioid recovery	12	3
Polysubstance use	Use of cannabis and an opioid at the same time to achieve desired high	3	18
Pain management	Cannabis use explicitly for pain management	6	5
Helpful with other strategies	Describes several strategies, including cannabis	3	2
Unclear	Unclear how helpful cannabis is for pain	3	3
<b>Healthcare provider interaction</b>	Poster refers to interaction with a healthcare provider or system	26	14
Related to cannabis use	Refers to a provider or program's attitudes towards cannabis	7	3

Motivations	Sample Quotations
Withdrawal symptom management	<p>"Marijuana is AMAZINGLY helpful. I'm going to ramble because it is a nice distraction. I hate marijuana usually. . . I had some pretty bad WDs and I feel better than ever. . . because of weed." (recovery subreddit)</p> <p>"2 days clean. I have no one else to tell. . . I'm taking a few hits of pot every 4 hours to ease the symptoms . . . it's been making it tolerable." (recovery subreddit)</p>
Questioning compatibility	<p>"As far as marijuana goes, I haven't decided if I want to start again. Any feedback from people who have stopped marijuana and then started again would be appreciated. Any regrets starting again? Did it affect your recovery negatively?" (recovery subreddit)</p> <p>"Opiates are usually my first priority, but I'm interested in weed now. My options are either maintenance therapy without weed or to trying to medicate with weed. . . . Should I keep pursuing maintenance? Or really try to get clean and smoke weed instead? Maybe it's just that now that I'm on subs I see maintenance does not mean I am high all the time and I still want to keep doing that somehow." (recovery subreddit)</p>
Polysubstance use	<p>"I'm having a very pleasant night sitting here, vegging out online—I am currently stable on my Suboxone dose, with a nice shot of crystal and some good weed. . . ." (opioid subreddit)</p> <p>"Tonight I smoked up a joint of heady weed with hash in it, then smoked like 500mg heroin 50mg at a time." (opioid subreddit)</p>
Pain management	<p>"Hi chronic pain sufferers—how effective is weed for your pain? . . . does it work for you at all? I find that it is just a distraction for me when I get random back pain." (opioid subreddit)</p> <p>"I came off methadone 9 days ago, and the withdrawal symptoms have sucked, sneezing, watery eyes, restless legs. You know, all that fun stuff. I've been using medicinal cannabis to control the pain, but I'm getting really fed up. Has anyone else done this, and if so, about how long does methadone withdrawal usually last?" (recovery subreddit)</p>

Direct quotations are lightly edited to reduce chances of re-identification.

Assumptions

Quantitative

Qualitative

Example

# Activity 3: Working with Reddit Data

- *What information do you want to access, store, and analyze?*
- *What data might be missing (as it relates your research question)?*
- *What challenges might you encounter?*
- *Who could you collaborate with in computer science, data science, information science for data collection (if needed)?*
- *Could you collaborate with in clinical, health services, policy, civil society, or other fields?*



Ethical Responsibility and  
Reddit Research: How  
Contextual Integrity Can  
Help Guide Practice

---

Nicholas Proferes  
Arizona State University  
nprofere@asu.edu

# DISCLOSURES

No conflicts of interest to report

# Talk Outline

- I. Ethics: Two Ways
- II. Gap between what researchers do and what users expect.
- III. Contextual Integrity



# Ethics as Contemplation

**Ethics:** systems of principles that we use to guide us in making moral evaluations.

- Utilitarianism
- Deontology
- Virtue Ethics, etc.

Using our capacities for **reason** and **judgment** to critically examine our actions and our character.



# “Ethics” as Compliance

Process of ensuring conformity with all relevant laws, policies, and guidelines.

These laws/guidelines/policies often draw on ethics-based principles, such as:

- Respecting human dignity and autonomy
- Maximizing benefit/minimizing harm
- Justice and beneficence

Ensuring researchers do not violate certain baseline conditions for the treatment of others.

# Compliance / Contemplation

“Public” social media data analysis often not considered (by U.S. IRBs) as “research involving human subjects.”

But, projects using publicly available data from social media sites do not lack ethical dimensions.

They require contemplation alongside (often minimal) compliance.

“Do users know their content is being used for academic study and how do they feel about it?”

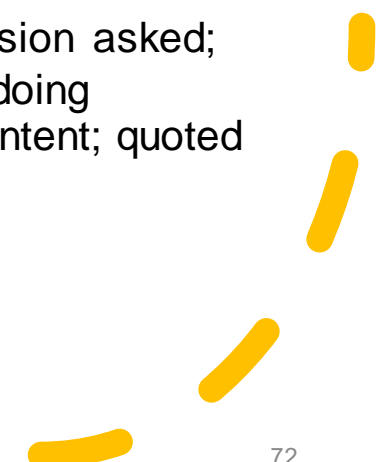
Fiesler, C., & Proferes, N. (2018). [“Participant” Perceptions of Twitter Research Ethics](#). *Social Media+ Society*, 4(1), 2056305118763366.

Surveyed  
Twitter  
users,  
asking:

Whether users think researchers are “allowed” to use their content without re-consent.

Levels of comfort with their tweets being used.

**Contextual factors** (permission asked; study content; size of dataset; who is doing analysis; kinds of content; status of content; quoted in study).



	Very uncomfortable	Somewhat uncomfortable	Neither uncomfortable nor comfortable	Somewhat comfortable	Very comfortable
... you were not informed at all?	35.1%	31.7%	16.4%	13.4%	3.4%
... you were informed about the use after the fact?	21.3%	29.1%	20.5%	22.0%	7.1%
... it was analyzed along with millions of other tweets?	2.6%	18.7%	25.5%	30.0%	23.2%
... it was analyzed along with only a few dozen tweets?	16.5%	30.3%	24.0%	20.2%	9.0%
... it was from your "protected" account?	54.9%	20.5%	13.8%	6.0%	4.9%
... it was a public tweet you had later deleted?	31.3%	32.5%	20.5%	10.4%	5.2%
... no human researchers read it, but it was analyzed by a computer program?	2.6%	14.3%	30.5%	32.3%	20.3%
... the human researchers read your tweet to analyze it?	9.7%	27.6%	25.0%	25.4%	12.3%
... the researchers also analyzed your public profile information, such as location and username?	32.2%	23.2%	21.0%	13.9%	9.7%
... the researchers did not have any of your additional profile information?	4.9%	15.4%	25.1%	34.1%	20.6%
... your tweet was quoted in a published research paper, attributed to your Twitter handle?	34.3%	21.6%	21.6%	13.1%	9.3%
... your tweet was quoted in a published research paper, attributed anonymously?	9.0%	16.8%	26.5%	28.4%	19.4%

User understandings of data-uses are limited.

Important to understand and study phenomena.

Levels of acceptance vary by specifics.

Notifying/consenting users could create anxiety.



 **Nissenbaum's framework of privacy as contextual integrity helps us unpack possible issues.**

Every situation in life has:

- **Norms of Information Appropriateness**

At a health check-up, it's appropriate for a doctor to ask about my health conditions, but probably not my bank balance.

- **Norms of Information Flow**

At a health check-up, it's appropriate for my doctor to share my health information about with my insurance, but probably not my bank.

# Where do these normative expectations come from?

For each contextual situation, we consider:

- Social roles (ours and others)
- Expectations we have for others, and that we think others have of us.
- Likely actions and practices of others, as informed by history, culture, law, and social convention.



# Addressing tensions...

- We must continuously weigh the tension between our values and duties.
- Try to get a sense of the norms of appropriateness and information flow with the subreddit.
  - Size, discoverability, subreddit rules.
- In data-sharing, reporting, find ways to anonymize user-data when possible.
  - Obfuscate quotes, don't include usernames, potentially subreddit names, images.

# Addressing tensions...

- Re-sharing by request only, [with your own user agreement](#).
- Document decision-making and discuss it!
- Ethical obligations to others: co-workers and students.
  - Debriefing as a team
  - Support services

- Association of Internet Researchers.
- Just shy of two decades of tackling ethics questions involving Internet Research.
- Has published two ethical decision-making guides + flowchart: <https://aoir.org/ethics/>
  - 2019 Ethical Guidelines (v 3.0).

# AUDIENCE Q&A



Extra Information

# Anonymization

Reddit is a pseudo-anonymous website but there are further measures that can be taken to protect identity

- Replace each unique username with unique, different ID (e.g., "User1"..."User234")

# Reddit API Resources

- JSON example <https://www.json.org/json-en.html>
- Official API <https://www.reddit.com/dev/api/>
- Python API wrapper <https://praw.readthedocs.io/en/stable/index.html>
- Submission and comment attributes <https://docs.google.com/document/d/1xpOD3dh9CgCrrRPTBsEfObdgYNstHQU6M1DXrLgUrCc/edit?usp=sharing>

# Changes to Reddit API Access

"Though access to Pushshift data for research purposes is not available at this time, we are keen to explore possibilities that might allow us to provide researchers with access to datasets essential for their valuable social media research."

[https://www.reddit.com/r/pushshift/comments/13w6j20/advancing\\_communityled\\_moderation\\_an\\_update\\_on/](https://www.reddit.com/r/pushshift/comments/13w6j20/advancing_communityled_moderation_an_update_on/)

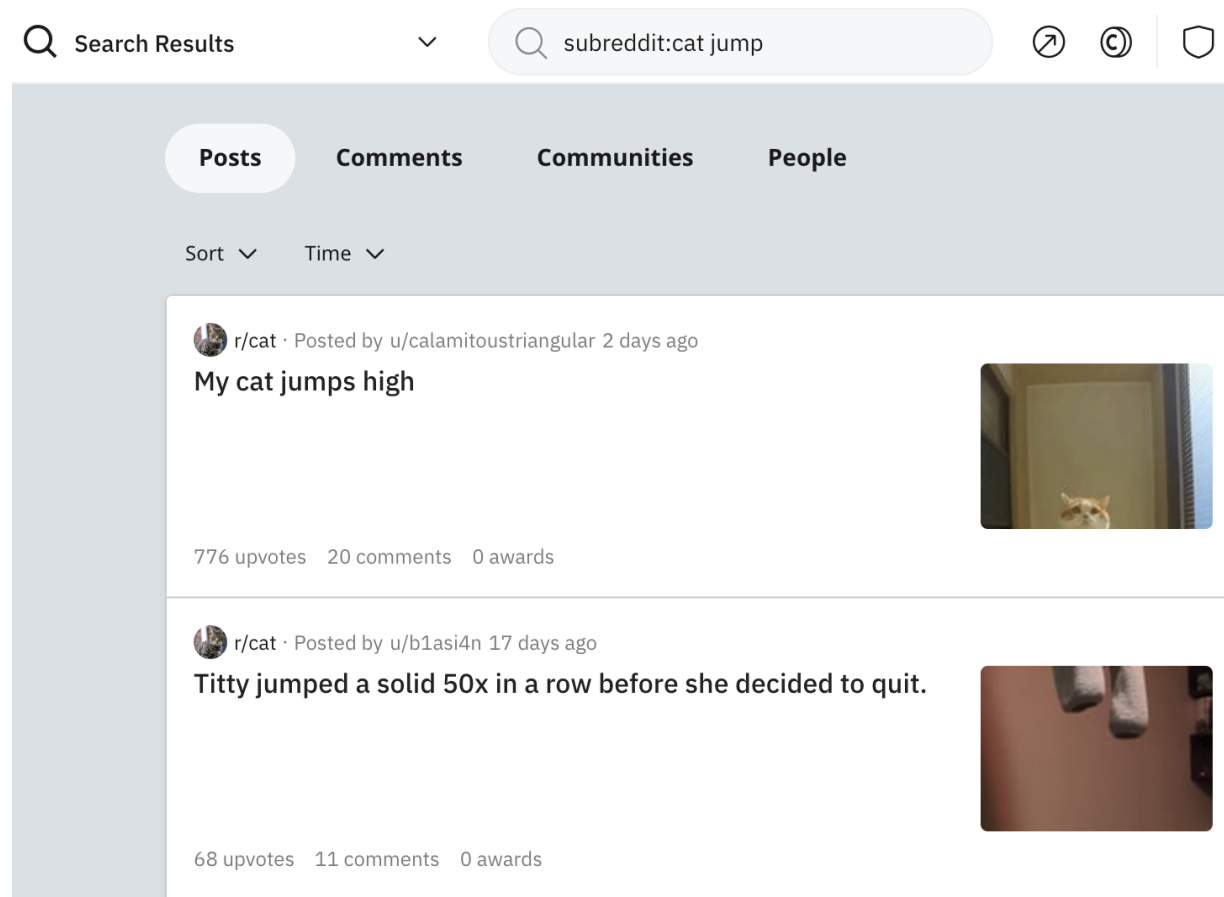
"Effective July 1, 2023, the rate limits to use the Data API free of charge are.. 100 queries per minute per OAuth client id if you are using OAuth authentication and 10 queries per minute if you are not using OAuth authentication."

[https://www.reddit.com/r/reddit/comments/145bram/addressing\\_the\\_community\\_about\\_changes\\_to\\_our\\_api/](https://www.reddit.com/r/reddit/comments/145bram/addressing_the_community_about_changes_to_our_api/)



# Manual Reddit Search

- Search syntax <https://www.reddit.com/wiki/search/>



# Data Size Estimates

Data Point	Size
Single comment or submission	~1KB
All submissions and comments from r/trees (uncompressed)	3GB
Month of all comments (compressed)	30GB

Table only includes text and metadata information, not images or videos

# Apache Solr for Reddit

- Getting started with Solr <https://solr.apache.org/guide/solr/latest/index.html>
- Solr configuration for Reddit <https://github.com/AADeLucia/Solr-Reddit-Utils>

# Post-Level Processing

General great resource for tutorials from The Summer Institutes in Computational Social Science <https://sicss.io/overview>

- Python resources
  - Removing unwanted symbols in text can be done with [Python regular expressions](#) (“regex”, or [pattern-matching](#)) and/or [NLTK](#)
  - Scikit-learn <https://scikit-learn.org/stable/>
  - Loading the files and removing duplicates can be done with [Pandas](#). The [read\\_json](#) method is especially useful.
- R resources
  - LOL learn Python
  - quanteda package for analyzing text data <http://quanteda.io>
  - But please, learn Python.

# Topic Drift

- Jerry Hobbs original overview of “topic drift” in natural human-to-human conversations:
  - <https://www.isi.edu/~hobbs/discourse-references/topic-drift.pdf>
- ResearchGate Article on the quality issues faced in social media date:
  - Nitin Agrawal, January 2010
  - [https://www.researchgate.net/profile/Nitin-Agarwal-14/publication/260337476\\_Information\\_quality\\_challenges\\_in\\_social\\_media/links/515961b0cf2d70ee270abfd/Information-quality-challenges-in-social-media.pdf](https://www.researchgate.net/profile/Nitin-Agarwal-14/publication/260337476_Information_quality_challenges_in_social_media/links/515961b0cf2d70ee270abfd/Information-quality-challenges-in-social-media.pdf)
- JMIR Publication of a study exploring topic drift in online health discussions
  - [Park, A., Hartzler, A. L., Huh, J., Hsieh, G., McDonald, D. W., & Pratt, W. \(2016\). "How Did We Get Here?": Topic Drift in Online Health Discussions. \*Journal of medical Internet research\*, 18\(11\), e284. https://doi.org/10.2196/jmir.6297](https://doi.org/10.2196/jmir.6297)

# Term frequency-inverse Document Frequency (tf-idf)

- Method explanation from scikit-learn [https://scikit-learn.org/stable/modules/feature\\_extraction.html#tfidf-term-weighting](https://scikit-learn.org/stable/modules/feature_extraction.html#tfidf-term-weighting)
- Implementation in Python [https://scikit-learn.org/stable/modules/generated/sklearn.feature\\_extraction.text.TfidfVectorizer.html#sklearn.feature\\_extraction.text.TfidfVectorizer](https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html#sklearn.feature_extraction.text.TfidfVectorizer)
- Article from Int'l Journal of Drug policy deploying tf-idf to identify Novel Psychoactive Substances (NPS) on reddit:
  - [Barenholtz, E., Krotulski, A. J., Morris, P., Fitzgerald, N. D., Le, A., Papsun, D. M., ... & Palamar, J. J. \(2021\). Online surveillance of novel psychoactive substances \(NPS\): monitoring Reddit discussions as a predictor of increased NPS-related exposure](#)

# Word Embeddings

- Word2Vec paper  
<https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf>
- State-of-the-art model, BERT
  - BERT model Python implementation <https://huggingface.co/bert-base-uncased>
  - BERT Paper <https://aclanthology.org/N19-1423>
  - BERT explanation <http://jalammar.github.io/illustrated-bert>
- Article from the British Journal on Social Psych about the application of word embeddings to capture cultural bias
  - [Durrheim K, Schuld M, Mafunda M, Mazibuko S. Using word embeddings to investigate cultural biases. Br J Soc Psychol. 2023 Jan;62\(1\):617-629. doi: 10.1111/bjso.12560. Epub 2022 Jul 23. PMID: 35871272; PMCID: PMC10086990.](#)

# Topic Modeling

- Readings for understanding topic modeling
  - Main paper <https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf>
- Implementations
  - Java implementation (MALLET) <https://mimno.github.io/Mallet/index>
  - Python implementation (uses MALLET)  
[https://radimrehurek.com/gensim 3.8.3/models/wrappers/ldamallet.html](https://radimrehurek.com/gensim/3.8.3/models/wrappers/ldamallet.html)
- JMIR article exploring application of LDA and sentiment analysis to detect discussions on Twitter about the COVID-19 vaccine
  - [Lyu JC, Han EL, Luli GK. COVID-19 Vaccine-Related Discussion on Twitter: Topic Modeling and Sentiment Analysis. J Med Internet Res. 2021 Jun 29;23\(6\):e24435. doi: 10.2196/24435. PMID: 34115608; PMCID: PMC8244724.](#)



# Infodemiology as a Framework

- Journal of Medical Internet Research (JMIR) video & description of infodemiology:
  - <https://infodemiology.jmir.org/>
- Fully article by Gunther Eysenbach on infodemiology in the American Journal of Preventive Medicine:
  - [Eysenbach G. Infodemiology and infoveillance tracking online health information and cyberbehavior for public health. Am J Prev Med. 2011 May;40\(5 Suppl 2\):S154-8. doi: 10.1016/j.amepre.2011.02.006. PMID: 21521589.](#)

# Privacy as Contextual Integrity:

- Link to the original publication of Helen Nissenbaum's 2004 Washington Law Review covering the topics of "Privacy as Contextual Integrity"
  - <https://crypto.stanford.edu/portia/papers/RevnissenbaumDTP31.pdf>

