

# Social Media as an Early Data Source for Emerging Substance Use Trends

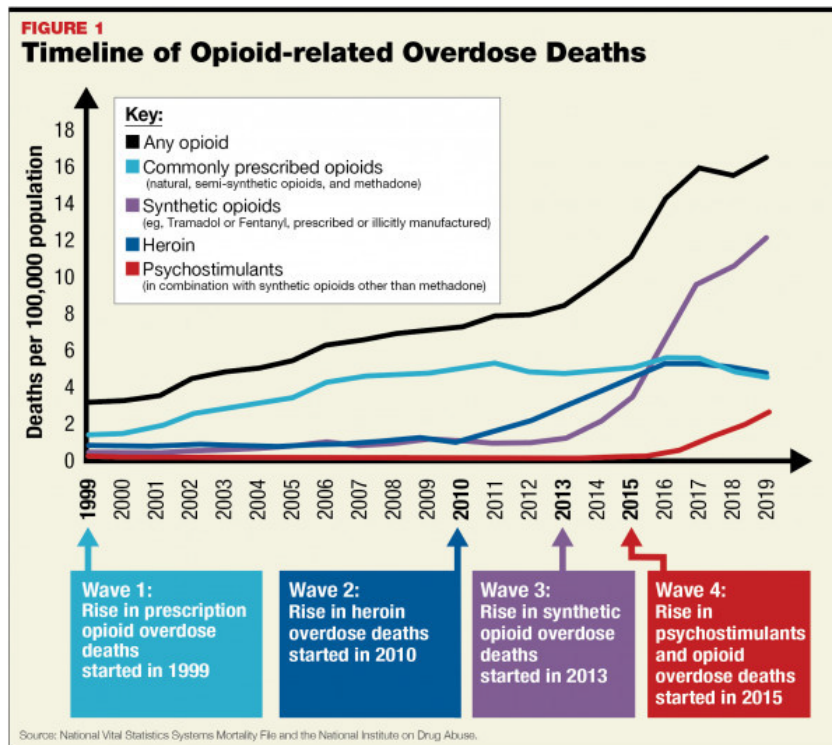
Sandy Preiss (RTI International)

Mark Edlund (RTI International)

Peter Baumgartner (Explosion AI)

Georgiy Bobashev (RTI International)

# Motivation



- Rapidly changing substances and use practices
- Prevention and treatment efforts need real-time info
- **Can we measure substance use trends in real time?**

# Social media can fill gaps in traditional surveillance

## Surveys and traditional surveillance

👍 Representative

🗨️ Slow

🗨️ Rigid

## Social Media

👍 Real-time

👍 Flexible

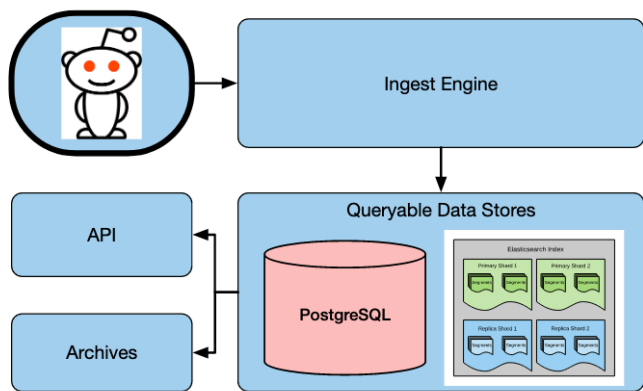
🗨️ Convenience sample



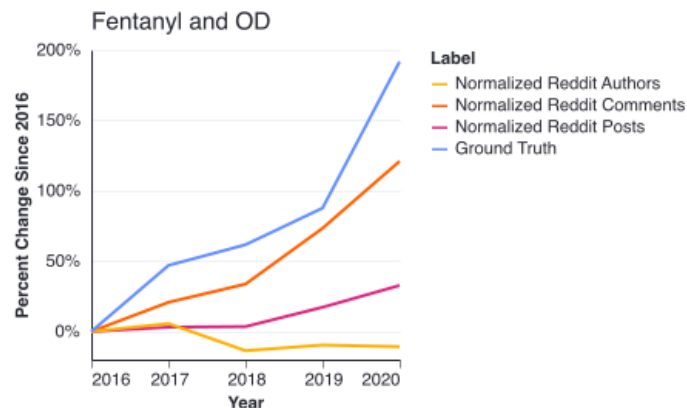
# Data science can unlock the wealth of info in social media

Free, public APIs provide access to massive quantities of Reddit data.

Natural language processing (NLP) extracts structured info from large Reddit corpuses.



**Figure 1:** Pushshift's Reddit data collection platform.



# Agenda

1. **Getting the data**
2. **Structuring the data**
3. **Comparison to “ground truth”**
4. **Next steps**

# 1. Getting the data

↑ 75 [Opiate Withdrawal Tips that have helped me through kicking over 20 times.](#) (self.opiates)  
submitted 2 years ago \* by 88Knuckles88

Alright guys, I'm gonna preface this routine I use by telling you straight up, I am not a doctor- taking medical advice from me might be stupid I couldn't tell you- but one thing I certainly am is a hardcore ig junkie. So when I tell you this provides relief, I'm not some doc telling you this from the sidelines. I have lived this routine too many times, and this is how I have managed to reduce 70-90% of withdrawals each time I have kicked the habit, making it bearable enough to get through by allowing me to eat and get some sleep in that dreaded first week. I dont know if all of the items are necessary or if it would be just as effective without some of them, but each time I've kicked (at least 20 times, sadly) this is the routine I've used. It is expensive- overall around \$300-400 for everything- but I'll mention some alternatives to reduce the cost at the bottom. Let's get started with a list of the things you'll need...

- Kratom (preferably an extract, specifically OPMS Gold Liquid Vials or The original UEI Kratom or tincture made from it are the three most effective I have come across for opiate relief)
- Loperamide (Immodium)
- Xanax
- Cannabis
- Restless legs tablets
- Clonidine
- Unisom
- Vitamin C/Ascorbic Acid
  - Mucinex
  - Food that is easy on the stomach (yogurt, apple sauce, peabut butter, bread, fruit)

↑ [-] [deleted] 8 points 2 years ago

Thanks for this! I'm sure a lot of people will find it helpful, I've honestly wanted to try Kratom quite abit but I never know of a place to get any. What is Kratom concentrate like? How does one ingest it?

There's a headshop here that sells just straight up Kratom, but I've heard that the headshop quality stuff isn't usually that good. I've always really wanted to try it but I'm quite skeptical on whether or not it'll actually work. I'm your experience, does it provide analgesia? Anything close to what say, Morphine might do ? Curious - look forward to hearing back. Thank you .

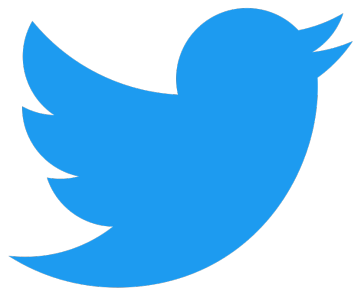
[permalink](#) [embed](#) [save](#)

## Full Data:

- Comments & Posts
- *r/opiates* & *r/OpiatesRecovery*
- 2015-2020

# Why Reddit?

- Topic-specific subreddits
- Long texts
- ***Easy access***





## API Documentation

### List of Endpoints

There are three main endpoints for the API to get information on comments, submissions and subreddits. The main endpoints are:

- /reddit/comment/search
- /reddit/submission/search
- /reddit/subreddit/search

### Full List of Pushshift Reddit Specific Parameters

Show  entries

Search:

Parameter	Type	Endpoint	Description
sort	Filter	All Endpoints	Sort direction of results ("asc" or "desc")
sort_type	Filter	All Endpoints	Parameter used for sort
after	Integer	All Endpoints	Restrict results to those made after this epoch time

# Data Acquisition

Google Cloud Platform

NIDA-OPIATES-WD

Search products and resources

BigQuery

FEATURES & INFO

SHORTCUT

Query history

Saved queries

Job history

Transfers

Scheduled queries

Reservations

BI Engine

Resources

+ ADD DATA

Search for your tables and datasets

hackernews

leadahq\_challenge

liquor

meetup\_extracts

mexico

mta\_nyc\_si

n

o

o

popular\_names

public\_dump

pypi

python\_extracts

reddit

reddit\_comments

reddit\_extracts

reddit\_posts

Get Opiate reddit comments Edited

LINK SHARING

COMPOSE NEW QUERY

HIDE EDITOR

FULL SCREEN

1 SELECT

2 body,

3 author,

4 created\_utc,

5 link\_id,

6 parent\_id,

7 score,

8 retrieved\_on,

9 id,

10 subreddit

11 FROM (TABLE\_QUERY([fh-bigquery:reddit\_comments], "table\_id BETWEEN '2007' AND '2014' OR table\_id CONTAINS '2015\_' OR table\_id CONTAINS '2016\_' OR table\_id CONTAINS '2017\_' OR table\_id CONTAINS '2018\_' OR table\_id CONTAINS '2019\_'"))

12 WHERE

13 LOWER(subreddit) IN ('opiates', 'opiatesrecovery')

14 LIMIT 100

Legacy SQL dialect

Run

Save query

Save view

Schedule query

More

This query will process 1.5 TB when run.

Query results

SAVE RESULTS

EXPLORE DATA

Query complete (5.5 sec elapsed, 1.5 TB processed)

Job information

Results

JSON

Execution details

Some cell values are very long and the display is truncated to the first 1024 characters to improve browser performance. If full values are necessary, try lowering the number of rows per page before clicking "Show full values".

11 I'm sure but I owe so much money to so many people what's the point anymore. I'll just file for bankruptcy then live paycheck to paycheck. Isn't that what everyone does

12 Thent it probably was laced, a test kit is the only way you're going to find out. Be safe!

13 I didn't ask a question, maybe you're referring to something interpreted as a question, on your behalf. I'm no ruling out I had a snidey tone & came across that way. I asked a question last weekend, on this sub. Check it out, it's the submission I refer to a

&#x200B;

Msg me direct if u like w ever maman x

Rows per page: 100 1 - 100 of 100 First page < > > Last page

## Query Pushshift on Google BigQuery

# Data Acquisition 2

The screenshot shows the GitHub repository page for `RTIInternational / PushshiftRedditDistiller`. The repository has 6 stars and 0 forks. The main branch is `main` with 10 branches and 1 tag. The repository was last updated 13 hours ago by `pmbaumgartner` with commit `c98ecf4`. The repository contains the following files:

File	Commit	Time
<code>.github/workflows</code>	first commit	16 hours ago
<code>src</code>	cleanup	14 hours ago
<code>test</code>	add fields test	14 hours ago
<code>.gitignore</code>	first commit	16 hours ago
<code>.travis.yml</code>	update travis & badges	16 hours ago
<code>LICENSE</code>	first commit	16 hours ago
<code>Project.toml</code>	add compats	13 hours ago
<code>README.md</code>	remove unused compression utils	15 hours ago

The `README.md` file contains the following content:

## PushshiftRedditDistiller

DOI: 10.5281/zenodo.4157726 | build: passing | codecov: 94%

This package is intended to assist with downloading, extracting, and distilling the monthly reddit data dumps made available through [pushshift.io](https://pushshift.io).

### Install

```
pkg> add https://github.com/RTIInternational/PushshiftRedditDistiller
```

### Example Use

Preexisting File

```
julia> using PushshiftRedditDistiller
```

**About**

This package is intended to assist with downloading, extracting, and distilling the monthly reddit data dumps made available through [pushshift.io](https://pushshift.io).

**Releases**

Initial Release (Latest) 16 hours ago

**Packages**

No packages published. [Publish your first package](#)

**Languages**

Julia 100.0%

<https://github.com/RTIInternational/PushshiftRedditDistiller>

# Data Acquisition 3

☰ README.md

## PMAW: Pushshift Multithread API Wrapper

circleci passing codecov 87% pypi v2.1.3 python 3.5 | 3.6 | 3.7 | 3.8 | 3.9 License MIT

### Description

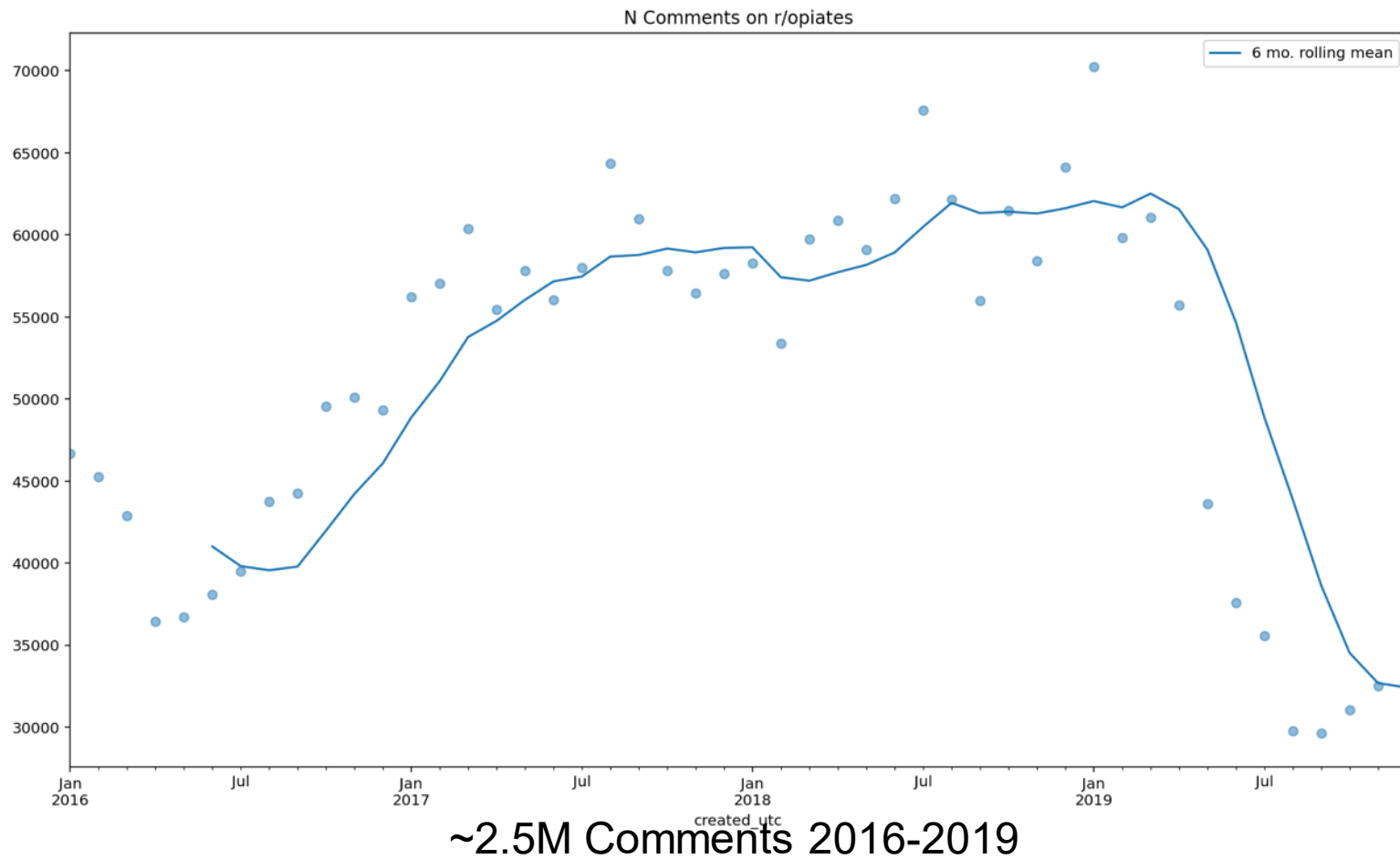
**PMAW** is a wrapper for the Pushshift API which uses multithreading to retrieve Reddit comments and submissions. General usage is through the `PushshiftAPI` class which provides methods for interacting with different `Pushshift` endpoints, please view the [Pushshift Docs](#) for more details on the endpoints and accepted parameters. Parameters are provided through keyword arguments when calling the method, some methods will have required parameters. When using a method **PMAW** will complete all the required API calls to complete the query before returning a `Response` generator object.

The following three methods are currently supported:

- Searching Comments: `search_comments`
  - [Details](#)
- Search Submissions: `search_submissions`
  - [Details](#)
- Search Submission Comment IDs: `search_submission_comment_ids`
  - [Details](#)

<https://github.com/mattpodolak/pmaw>

# How much data?



## 2. Structuring the data

# Named Entity Recognition

**Effect:** a therapeutic or adverse effect mentioned as a result or rationale for consuming a substance.

**Substance:** a drug, remedy, supplement, or other consumable item used to treat an effect or induce a desired effect.

EFFECT 1

SUBSTANCE 2

Yeah I know **subs** SUBSTANCE . Just never used them. **Mdone** SUBSTANCE is much cheaper to get and you dont have to wait untill you can use them. But I think if you really want to get off your DOC then **subs** SUBSTANCE are more powerfull cause they block the receptors. I myself have taken once 120mg of **mdone** SUBSTANCE and craved **H** SUBSTANCE so much that I used it to and nearly OD'd but everytime I detox I use **Diazepam** SUBSTANCE too. It really helps me with the **insomnia** EFFECT and the **RLS** EFFECT . But I prefer **Xanax** SUBSTANCE more. They are stronger and most of the detox time I sleep cause of them but I take high doses like 10-15mg per day plus 100mg of **Diazepam** SUBSTANCE through out the day and take 3 times a day the 5mg **red devil Xans** SUBSTANCE . It works wonders for me. Then I also use **Loperamid** SUBSTANCE against the **diarrhea** EFFECT . And yeah that makes wds easy for me. **Benzos** SUBSTANCE are really a godsend for wds. If you have enough then detox is super easy. The last time I detoxed I had nothing because of my poor planing amd it was super hard but it was bearable . Just the **puking** EFFECT was the hell and the **RLS** EFFECT and **insomnia** EFFECT . But after 5 days everything was over and I managed

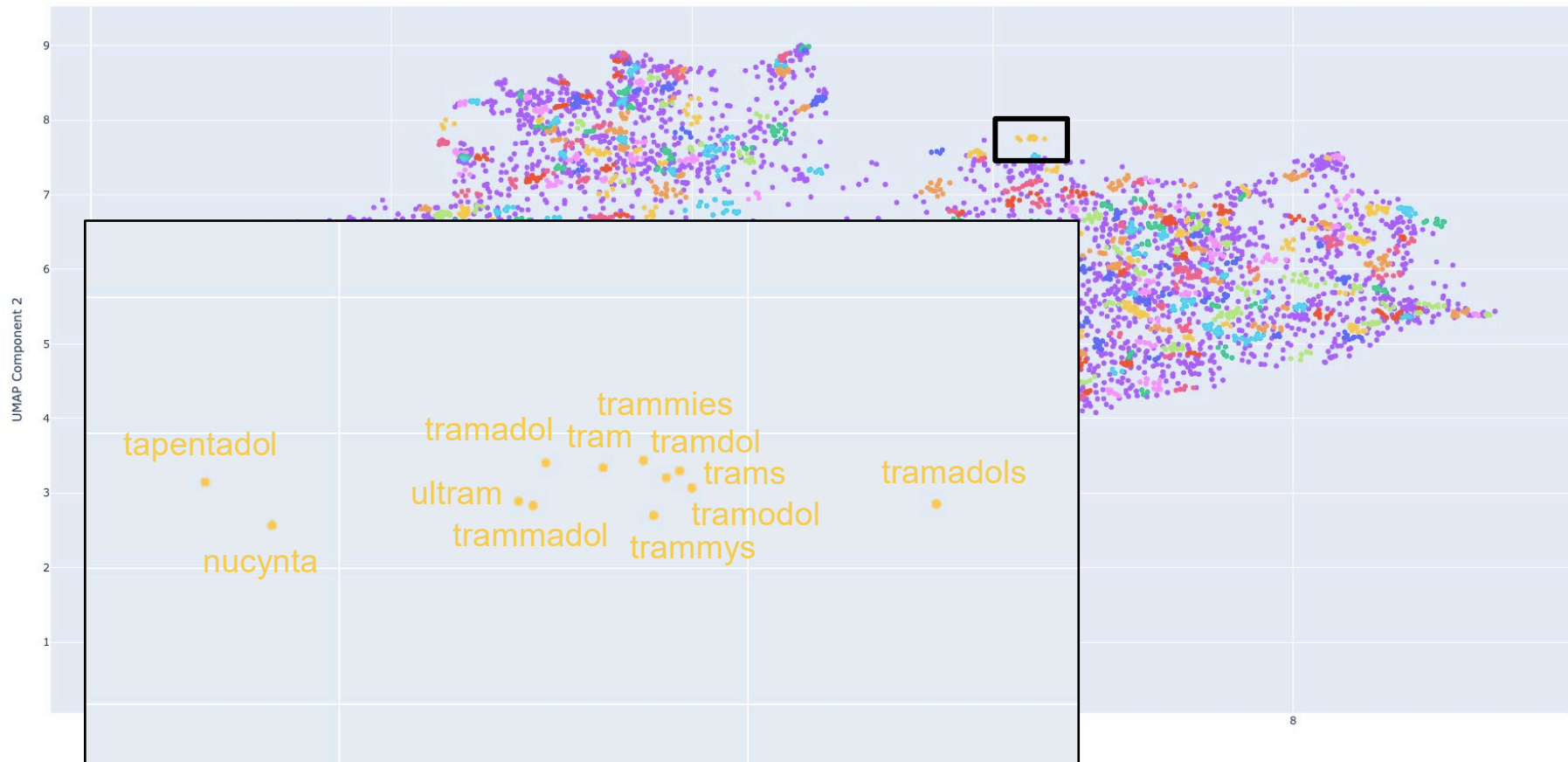
# Problem

Entity	Count
lope	16853
lope imodium	2
lope lope	3
lopeamide	2
lopedium	6
lopeermide	2
lopemide	13
loperadime	3
loperaide	2
loperamade	2
loperamaide	4
loperamid	37
loperamide	7424
loperamide hci	2
loperamide hcl	49
loperamide hydrochloride	18

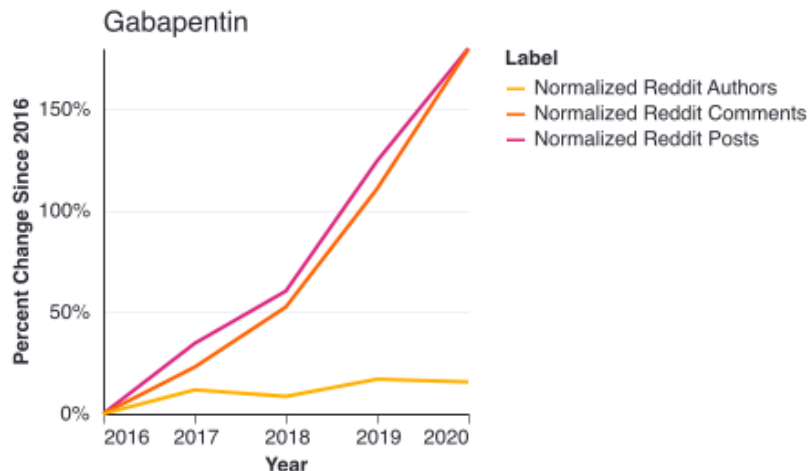
Entity	Count
immodium	2
immodium	3598
imodiom	3
imodium	3093



# Clustering deduplication



# Measure trends in entity mentions over time



- Posts, comments, and unique authors
- Normalized as proportion of all posts which mention topic

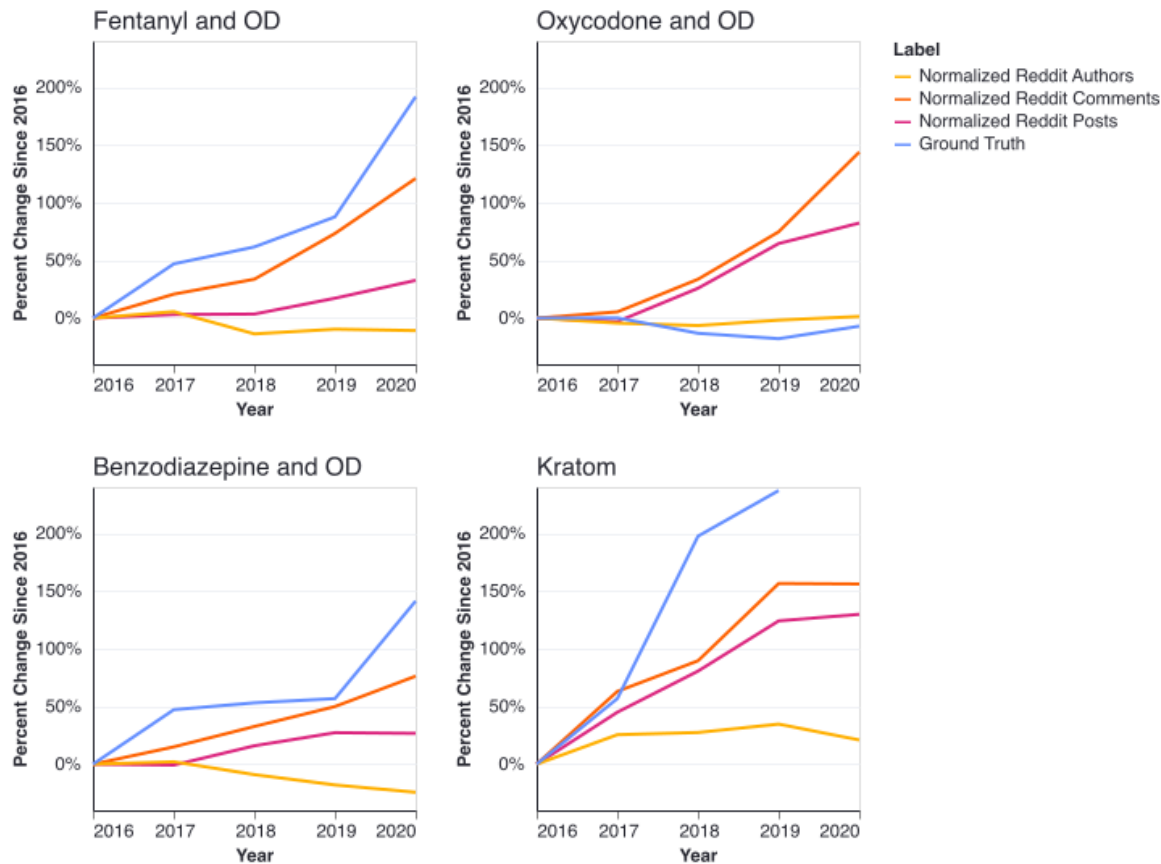
### **3. Comparison to “ground truth”**

# Comparison Methods

- Select topics where a “ground truth” comparison is possible (e.g., fentanyl overdose)
- Extract Reddit discussion trends and comparison data over same time frame
- Visualize and estimate correlation



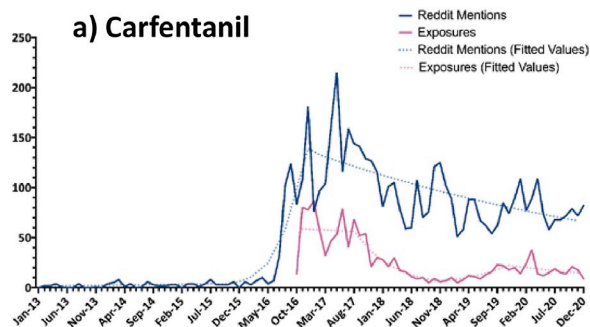
# Comparison Results



## 4. Next steps

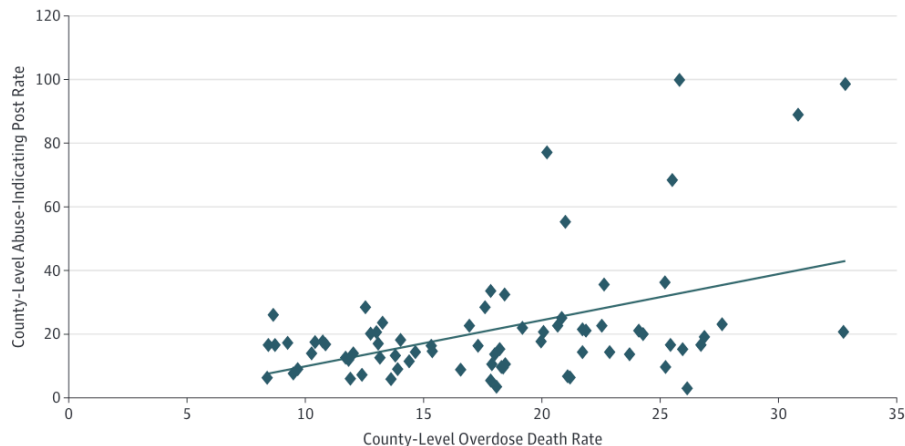
# Many proofs of this concept

Barenholtz, E., et al. (2021). Online surveillance of novel psychoactive substances (NPS): Monitoring Reddit discussions as a predictor of increased NPS-related exposures. *International Journal of Drug Policy*, 98, 103393. <https://doi.org/10.1016/j.drugpo.2021.103393>

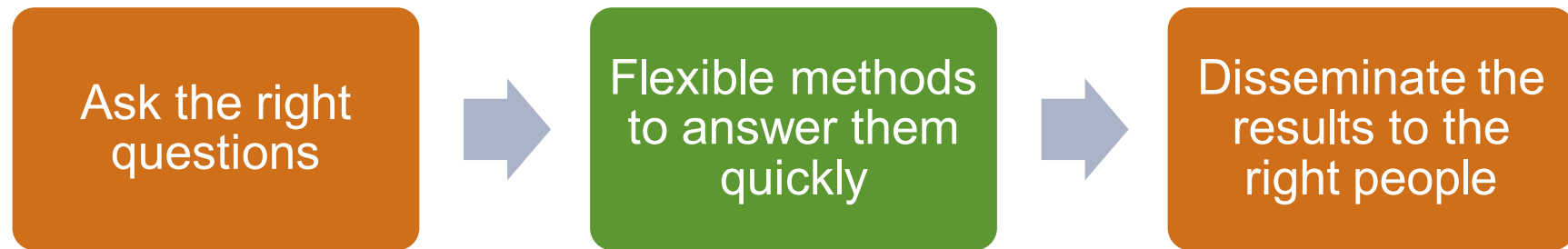


Sarker, A., Gonzalez-Hernandez, G., Ruan, Y., & Perrone, J. (2019). Machine Learning and Natural Language Processing for Geolocation-Centric Monitoring and Characterization of Opioid-Related Social Media Chatter. *JAMA Network Open*, 2(11), e1914672. <https://doi.org/10.1001/jamanetworkopen.2019.14672>

**B** Association between abuse-related social media posts and opioid-related death rates

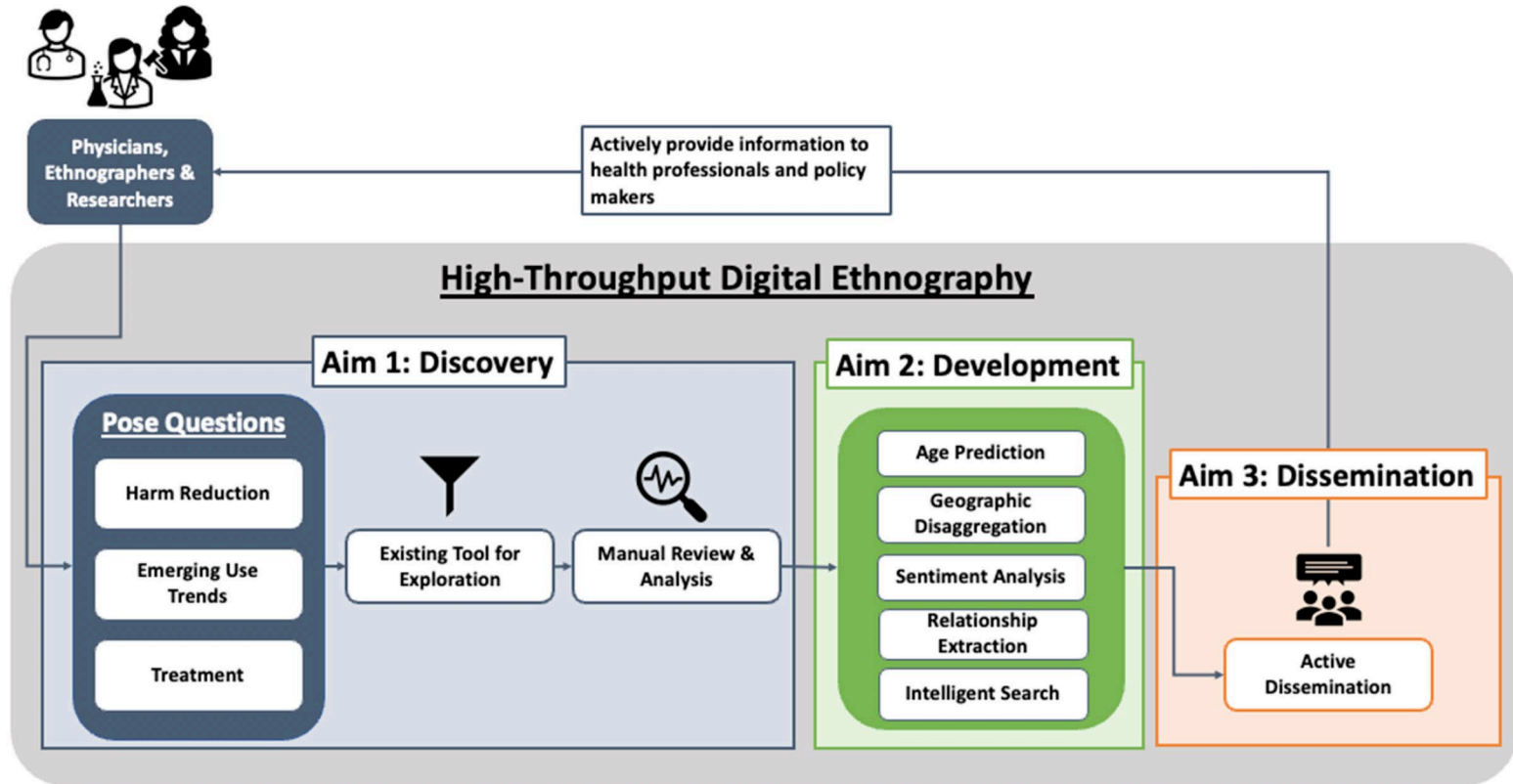


# On-the-ground implementation remains a goal





# NIDA R01: High Throughput Digital Ethnography for Substance Use



Sandy Preiss  
Research Data Scientist  
RTI International  
**[apreiss@rti.org](mailto:apreiss@rti.org)**

# Appendix

# Dealing With Large Data in Natural Language Processing

## 1. Solve a smaller & simpler problem

- a. **Identify** relevant documents with a keyword list derived from word vectors
- b. **Build** labeled data as fast as possible with binary decision-making on token-based entities

## 2. Iterate with the data

- a. **Train** *disposable* models as soon as feasible ( $n < 250$ )
- b. **Apply** the disposable model to remaining data, aggregate results, do error analysis on results, identify documents with errors for training

# Where to start?

1. Bootstrap identifying relevant documents with a terminology list derived from word vectors
  1. Start with obvious ones: “opiates”, “oxycontin”, “tiredness”, “yawning”, etc.
2. Start generating labels as fast as possible with binary decision-making on token-based entities
3. Make corrections on simple binary annotations rather than start from the blank page of no annotations

**In sum:** identify likely relevant comments and solve simpler problems first

# How to improve the model

1. Train a model with limited training data ( $n < 1000$ ) and run it on the full dataset.
2. Aggregate predicted entity occurrences.
3. Perform error analysis on aggregated entity counts.
  1. i.e. “lack of” as an effect, “wal-mart” or “adhd” as a substance
4. Identify documents not in training dataset with erroneous entities, create a new batch of training data with those documents
5. Annotate again on identified erroneous examples.
6. Repeat from top until you have an acceptable error rate within the *top n* most common entities.

**In sum:** Iterate on your data

# Results: Most Common Entities

Substance			Effect	
Rank	Entity	Count	Entity	Count
1	dope	184420	pain	70171
2	opiates	162949	anxiety	31013
3	heroin	152477	depression	23567
4	oxy	91823	sleep	22294
5	subs	83456	cravings	19991
6	sub	80744	depressed	11141
7	opiate	79997	rls	8794
8	methadone	73054	nausea	7454
9	kratom	65298	craving	6355
10	fent	64020	insomnia	5594
11	suboxone	62785	puke	5484
12	h	51548	mood	5366
13	weed	49659	seizures	5251
14	morphine	45814	anxious	5190
15	benzos	36115	puking	5063

# Performance: Results

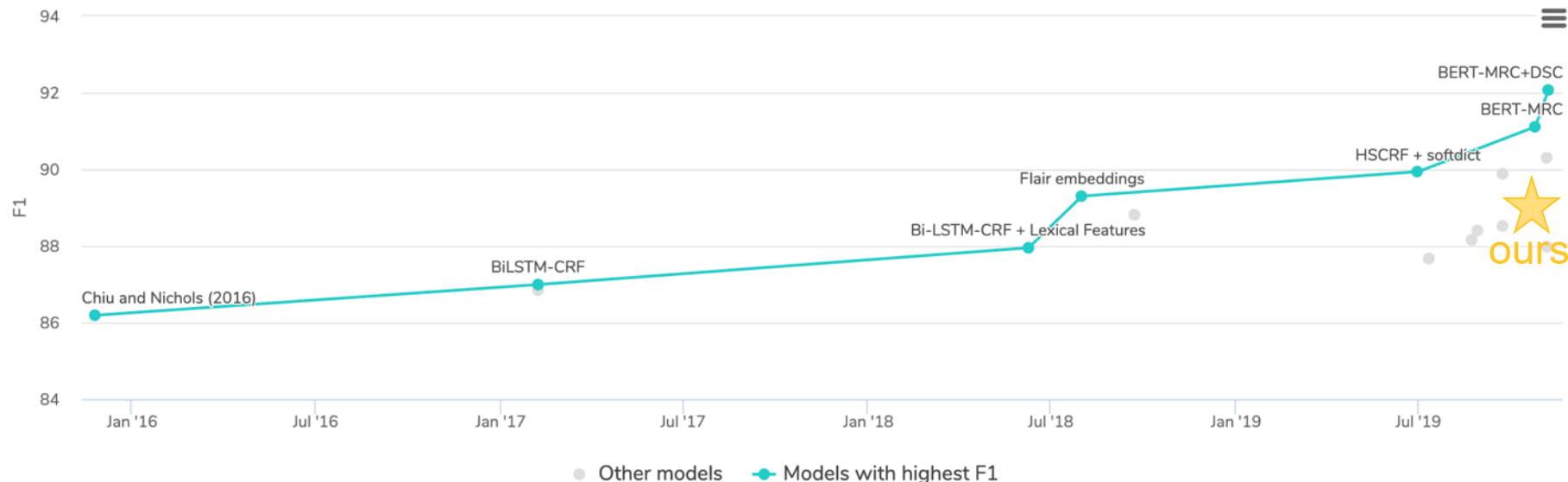
Entity → Measure ↓	EFFECT	SUBSTANCE	OVERALL
Precision	80.90	90.48	88.23
Recall	79.56	91.45	88.63
F1	80.20	90.96	88.42

*5-fold cross validation*  
*Scoring: exact overlap*



# Performance: Is this good?

## Named Entity Recognition on Ontonotes v5 (English)



<https://paperswithcode.com/sota/named-entity-recognition-ner-on-ontonotes-v5>

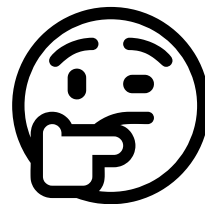
# Orthography and Semantics (Spelling and Meaning)

Fuzzy string matching

m i s p e l d  
/ / / | | | \  
m i s s p e l l e d

Loperamide

Immodium



# Word Embeddings

## Word Embeddings:

Vector representations of word meaning

	living being	feline	human	gender	royalty	verb	plural
<i>cat</i> →	0.6	0.9	0.1	0.4	-0.7	-0.3	-0.2
<i>kitten</i> →	0.5	0.8	-0.1	0.2	-0.6	-0.5	-0.1
<i>dog</i> →	0.7	-0.1	0.4	0.3	-0.4	-0.1	-0.3
<i>houses</i> →	-0.8	-0.4	-0.5	0.1	-0.9	0.3	0.8

**Word Embeddings leverage**

**Distributional Semantics:**

"A word is characterized by the company it keeps." – J.R. Firth

# FastText



# FastText Output

Entity	0	1	...	$N$
lope	0.1	-0.3	0.8	0.5
loperamide	0.2	-0.2	0.8	0.6
immodium	0.2	-0.3	0.5	0.8
imodium	0.1	-0.4	0.5	0.7
heroin	-0.6	0.9	0.1	-0.1

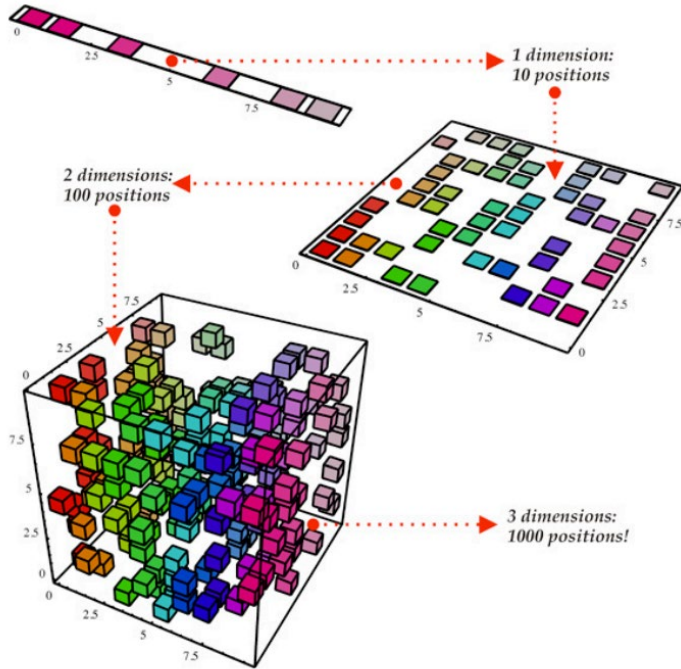
*Next Step:*

*Cluster like with like!*

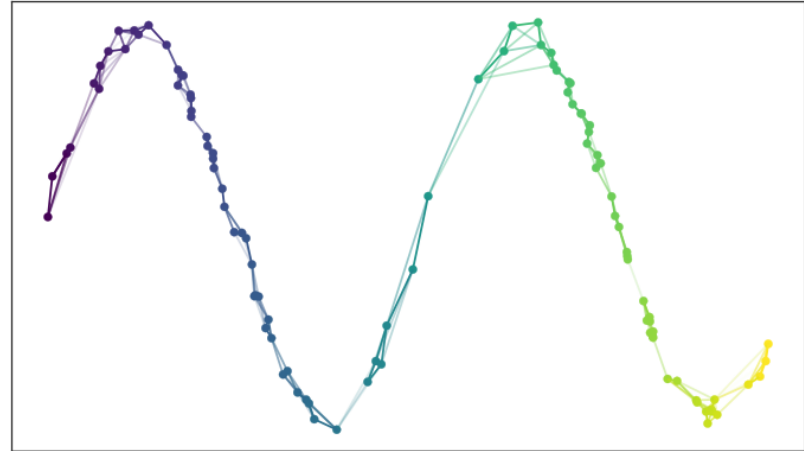
*But first...*

# Dimension Reduction: UMAP

## The curse of dimensionality



## UMAP: Uniform Manifold Approximation



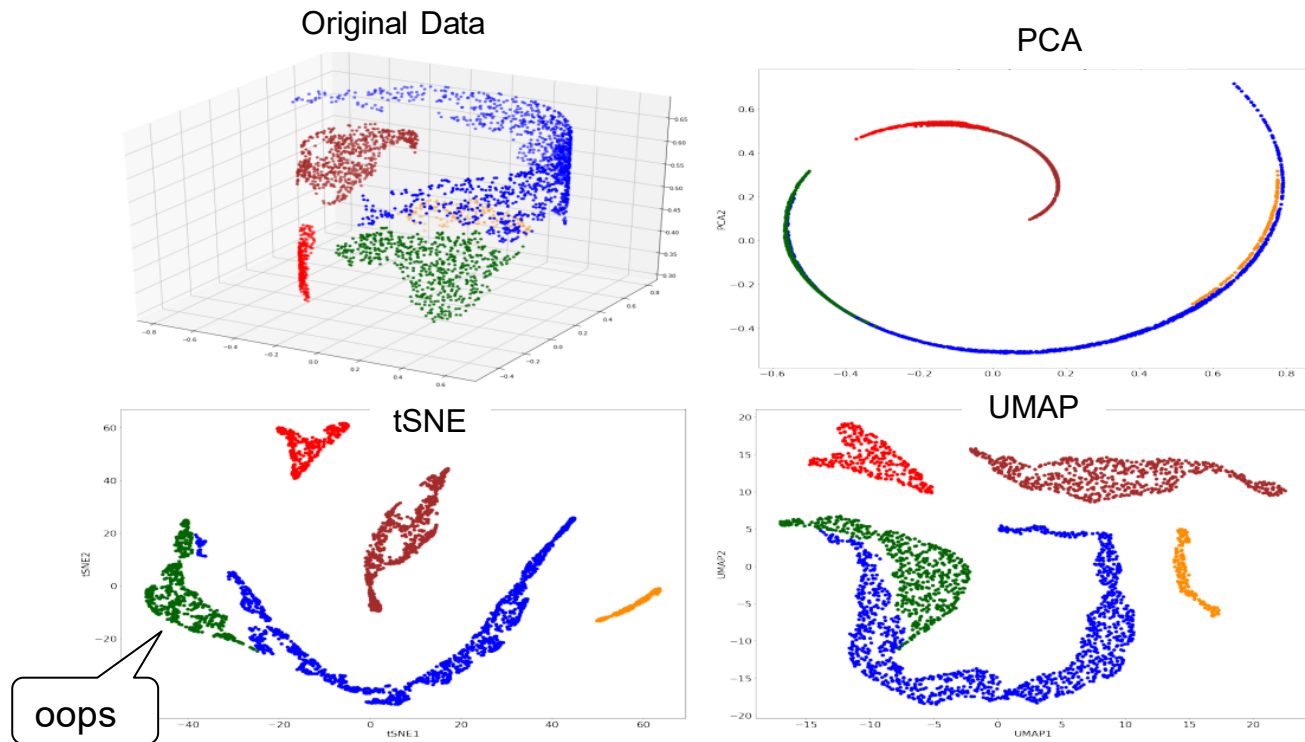
# Why FastText?

- Generates embeddings for character n-grams as well as whole words
- Character n-grams capture infrequent misspellings
- Whole word embeddings capture synonyms (e.g. brand name & generic drug)
- Works on words not in training data



# Why UMAP?

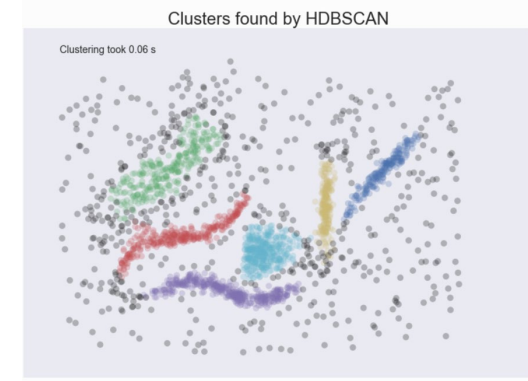
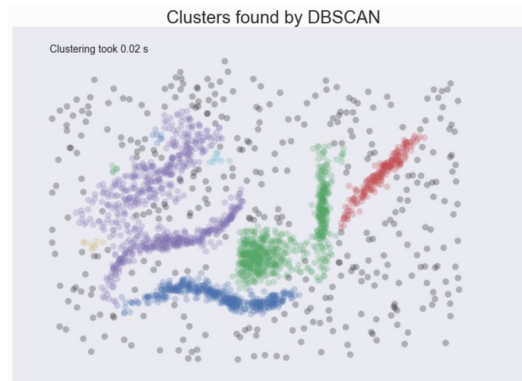
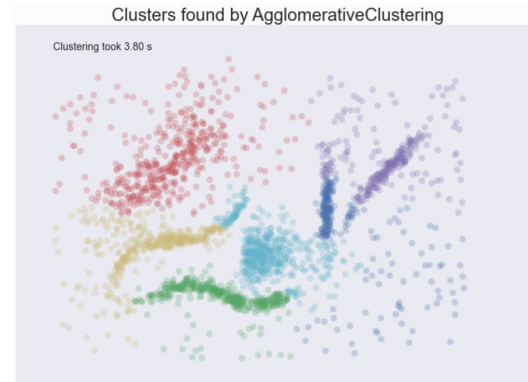
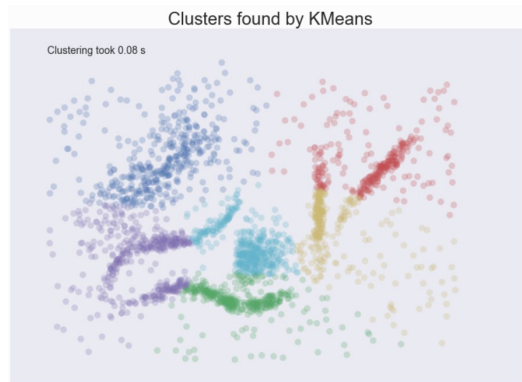
- Retains heterogeneity (groups neighboring points)
- Preserves global structure
- Relatively fast





# Why HDBSCAN?

- Dense clusters with sparse background noise
- Flexible cluster shapes
- Intuitive parameter tuning (vs DBSCAN)
- Fast



# Clustering Output

Cluster ID	Most Common Entity	Number of Entities	Other Entities
120	lope	17	immodium, imodium, lope, lopemide, loperamid, loperamide, loperamide hcl, loperamideimmodium, loperamine, loperimide, lopermade, lopermaide, lopermide, lopermine, loperomide, lopes, nimmodium
221	fentanyl	8	fentanyl, fentanyl 30s, fentanyl 80s, fentanyl blues, fentanyl heroin, fentanyl ods, fentanyl oxy, fentanyl xanax
271	tylenol	14	aceataminophen, aceta, acetametaphine, acetaminophen, acetaminophen 10325, acetaminophen 5325, aceteminophen, acetimophen, acetomenaphin, acetominophen, apap, aspirin acetaminophen, paracetamol, tylenol