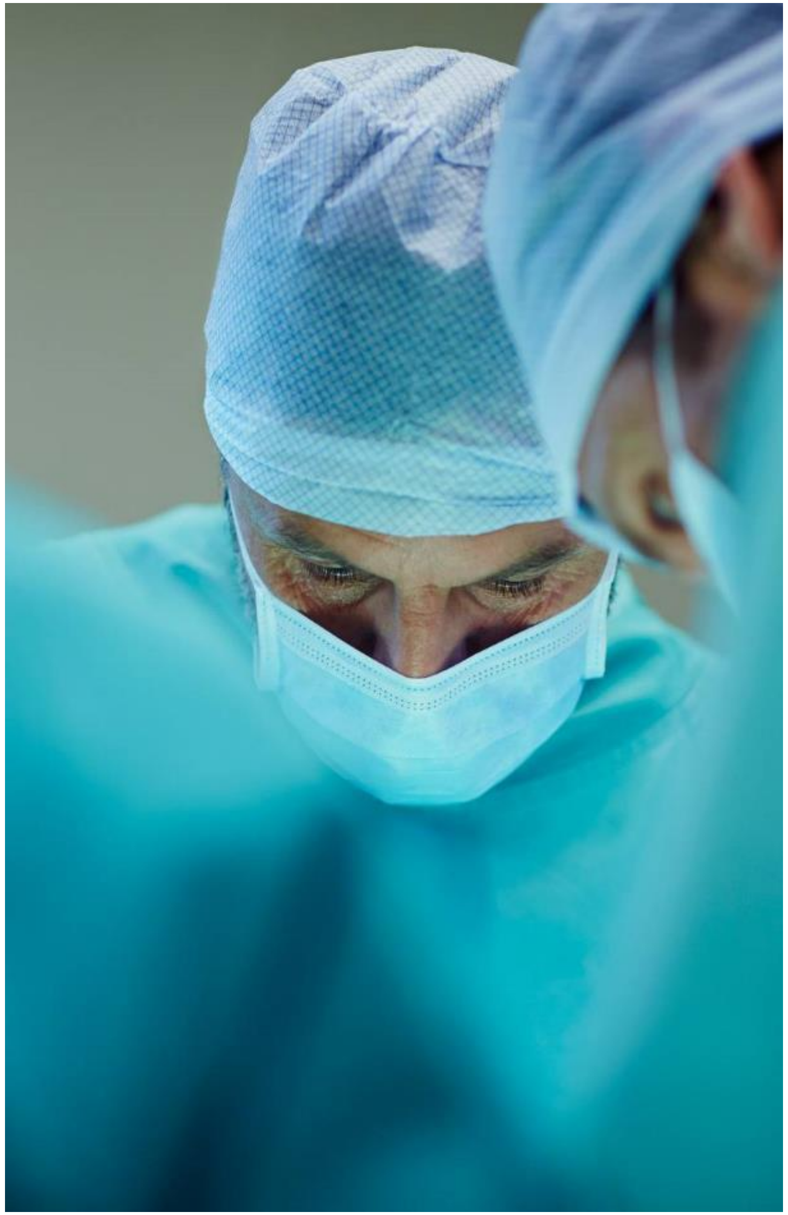
TERM PROJECT – MIGRAINE DATASET

Group 11 - Robert Sliwa, Adam Gallegos, Huiling Yang

AGENDA

- ❑ Introduction
- ❑ Project plan and goal
- ❑ Data gathering and processing (Data sources, Feature Extraction Workflow, Feature Extraction Examples)
- ❑ Dataset creation
- ❑ Conclusions



MIGRAINES

- Lots of people affected
- Not well understood
- Experienced differently by individuals
- We know family and friends who suffer from migraine

PROJECT PLAN AND GOALS

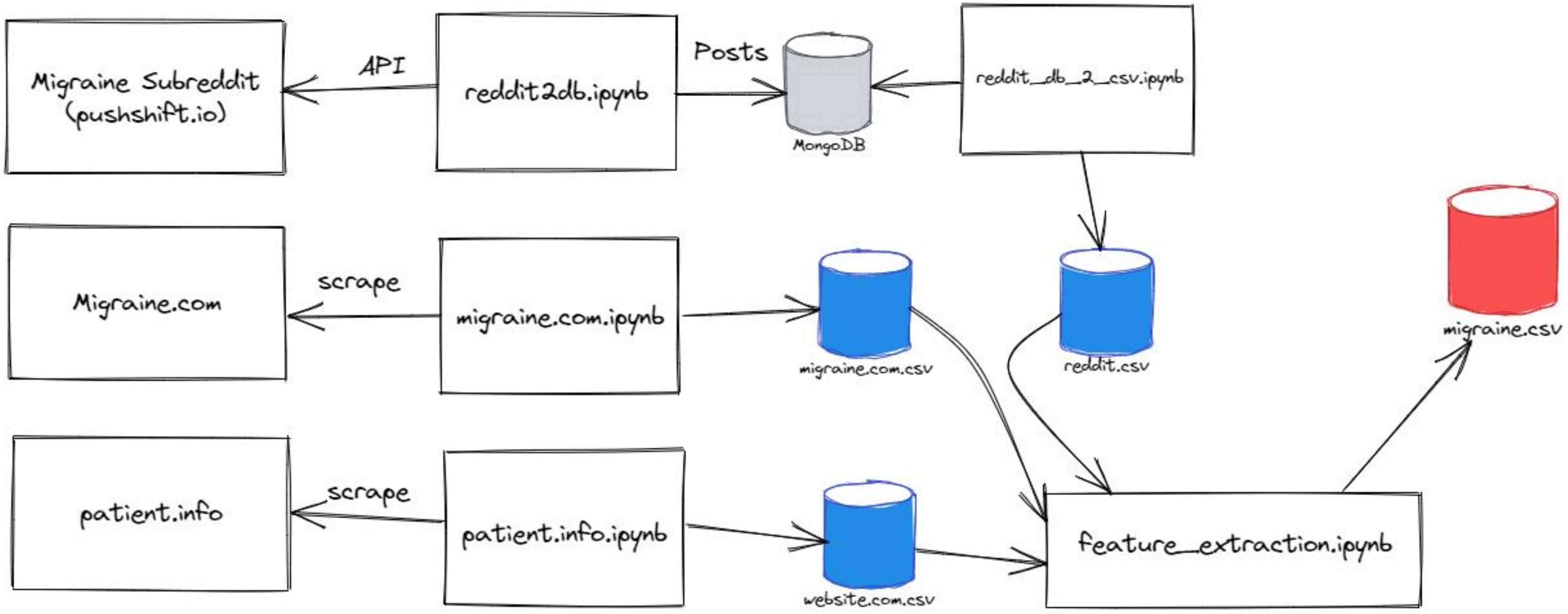
Create a useful dataset, which can lead to insights on causes and treatments of migraines.

Methods: scraping various websites that discuss migraine headaches and pulling the data together.

WHO MIGHT BE INTERESTED IN OUR DATASET

- Sufferers: get more knowledge about migraine, and get access to other people’s insights, tips, and experiences with migraine
- Family member, friends or employers of migraine suffers get to know migraine experience and better understand or take good care of migraine sufferers
- Doctor or health provider: connect to a huge number of migraine sufferers
- Medical school student: get to know more about migraine and people’s experiences with migraine
- Insurance company: get more information of migraine experience related to prevention, medication and treatment
- Data scientist: help them build their own dataset
- Researchers: find interesting topics (problems) and work on it (or try to solve the problems)
- Medical app developers: build their own app

DATA GATHERING



DATA SOURCES: MIGRAINE SUBREDDIT

Migraine Subreddit is a large collection of posts about variety of migraine experiences like medicines used, techniques to deal with headaches, descriptions of auras, rants about people not understanding migraine sufferers and implications of suffering from migraines on work and family.

Access via Reddit API

- Reddit' User Agreement has provisions for web scrapping, but it allows to access data via API
- No need for authentication to access Subreddit posts
- Don't provide long history
- Throttling required

Access via Pushshift.io API

- Copy of Reddit for purposes of Big Data and Social Media ingest
- Simple API with no authentication
- Long history of posts maintained
- Throttling required
- Total posts: 42878, with comments: 421030
- Unique authors: 48845

WORKING WITH PUSHSHIFT.IO APIS

- Pushshift.io maintains copy of Reddit posts for purpose of Big Data and Social Media Ingest
- Simple APIs that do not require authentication
- Documentation doesn’t specify throttling policies, but we empirically discovered them
- We waited for 1 second between requests and 5 seconds on retries
- Server returns 429 Too Many Requests error when it is overwhelmed
- You must request list of posts that can be paginated based on creation time expressed in Unix epoch
- For each post you must take post id and use it to retrieve list of comments for associated using different API
- Too gather 10000 posts is takes about 8 hours.

DATA SOURCES: MIGRAINE.COM

Migraine.com is an excellent source on people sharing their experiences with migraine, the challenges they face, and how they deal with them. The advantage of Migraine.com over migraine subreddit is that it is more organized and focused on the topics within the available categories.

How to access data:

- No APIs, needs Web scraping
- Terms of Use require permission for scraping articles but not forums
- Required use of Selenium module to open each forum page
  - Time consuming, but effective
  - A function was required to open-up hidden posts

Data:

- 326 pages of listed discussion posts
- 3215 discussion topics
- 23267 unique comments

DATA SOURCES: PATIENT.INFO

Patient.info is a more general medical forum site. It does have an active section dedicated to migraine headaches. This was scraped in a similar manner to migraine.com. Less active than the other site, but still added significant data.

How to access data

- No APIs, needs Web scraping
- Also used selenium module to scrape each page
- A different format than migraine.com, responses were instead on multiple pages
  - Looped through all available pages for each topic

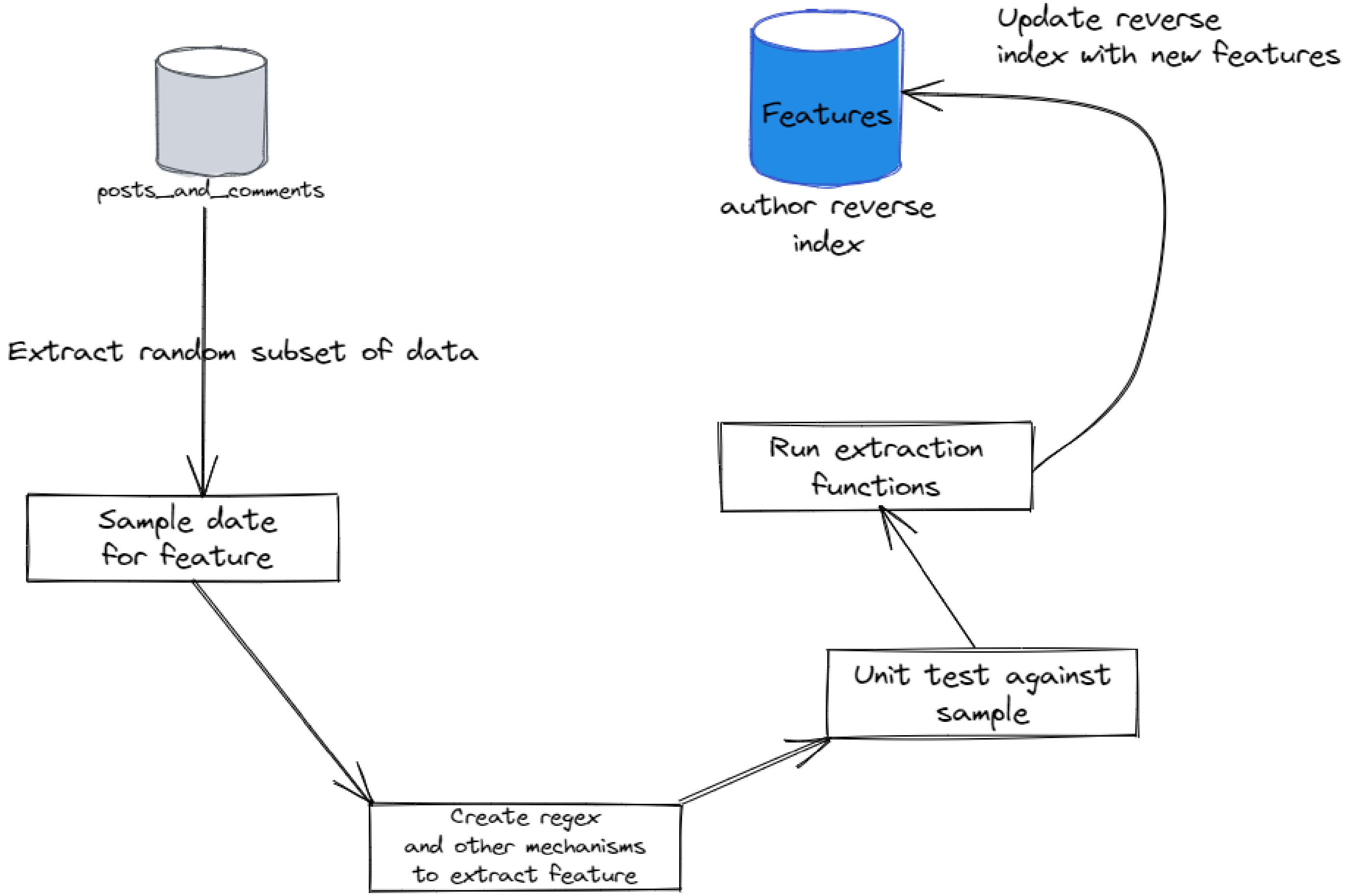
Data:

- 23 pages of listed discussion posts
- 965 discussion topics
- 10698 unique comments

FEATURE EXTRACTION WORKFLOW

Features:

- Age
- Gender
- Suicidal thoughts
- Presence of aura
- Migraine triggers
- ADHD
- Treatments/drugs and dosage
- Effectiveness of treatments



posts_and_comments

Extract random subset of data

Sample date for feature

Create regex and other mechanisms to extract feature

Unit test against sample

Run extraction functions

Update reverse index with new features

Features

author reverse index

FEATURE EXTRACTION TECHNIQUES

- There is general workflow that we used for working on all the features
- Different features needed somewhat different approach to retrieving them from the posts and comments
- The nature of the data is that authors write multiple posts and comments
- To capture all the features, we create reversed index where author is the key and value is a dictionary of the features

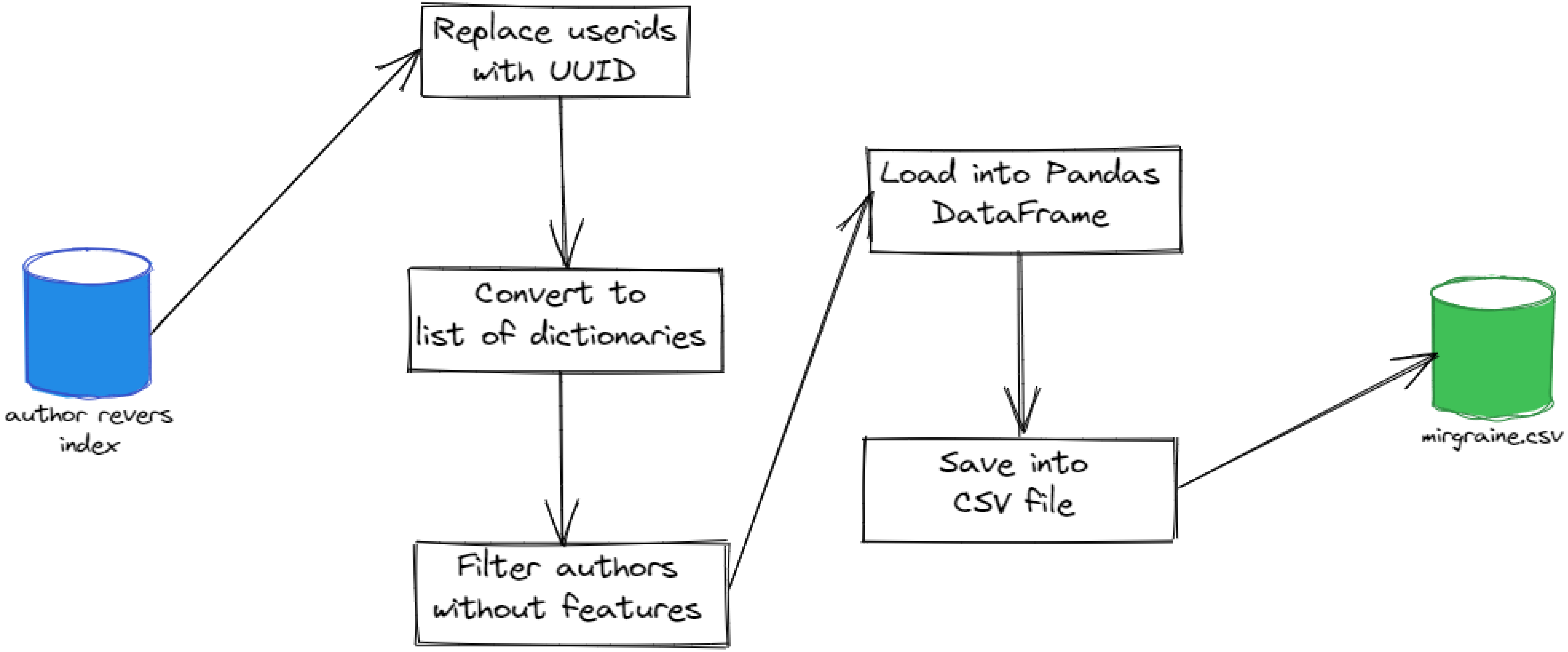
```
gender,medicine,dosage,qty,effectiveness,suicidal,age,triggers,aura,adhd,id

male,triptan,5 mg,1x,yes,no,33,[‘caffeine’],false,false,86aa1692-4bd1-45b9-98d2-fdc3679193ae
```

Samples

- Extract author’s medication, dosage, effectiveness
- Extract author’s triggers

CREATE DATASET



author revers index

Replace userids with UUID

Convert to list of dictionaries

Filter authors without features

Load into Pandas DataFrame

Save into CSV file

mirgraine.csv

CONCLUSION

- The major challenge is parsing through the forum posts and extracting meaningful information

Information extraction from text is hard! but observing people posts over time one can learn a lot about them

- we had to use ensemble of techniques to retrieve information from the articles to finally put all of that together in a simple structured data that can be easily represented and digested.

a.) Vast number of conversations can still lead to limited amount of useful data: ~500000 posts -> ~5000 authors with useful information

b.) Simple text processing techniques can get a lot of information

c.) Simple text processing techniques can only take you so far

THANK YOU

Name: Robert Sliwa

Email: rjs463@drexel.edu

Name: Adam Gallegos

Email: ag3999@drexel.edu

Name: Huiling Yang

Email: hy435@drexel.edu