Datathon and Machine Learning Competition on Antisemitism
Workshop 2 – Creating a Discourse Dataset from X

Presentation available at:
https://github.com/damieh1/datathon_2025/blob/main/Datathon_Workshop-Session-2.pdf

INDIANA UNIVERSITY

Dr. Daniel Miehling

18.07.2025

What happened so far?

Workshop 1 Recap:

- Prof. Jikeli introduced patterns of antisemitism online
- Focus on dynamics before and after October 7

Online Antisemitism Before and After October 7

Workshop 1
Datathon & Machine Learning Competition

Günther Jikeli, Daniel Miehling, and Damir Cavar
Research Lab *Social Media & Hate*

Conceptual and Theoretical Background:

- Reach out to him with questions about content/context

Missed it? His slides are available in the GitHub repo

Team formation:

- Most groups should be in place by now
- Team Communication & Responsibilities

Today’s Session

1. Bright Data Introduction

2. Hands-on Session for Challenge #1
   - Work in your teams to collect, annotate, and prepare data

3. Q&A → Conceptual: Prof. Jikeli | Content: Dr. Miehling | ML: Prof. Cavar

Meet the Instructors

Workshop #1

Prof Günther Jikeli

gjikeli@iu.edu

Workshop #2

Dr Daniel Miehling

damieh@iu.edu

Workshop #3

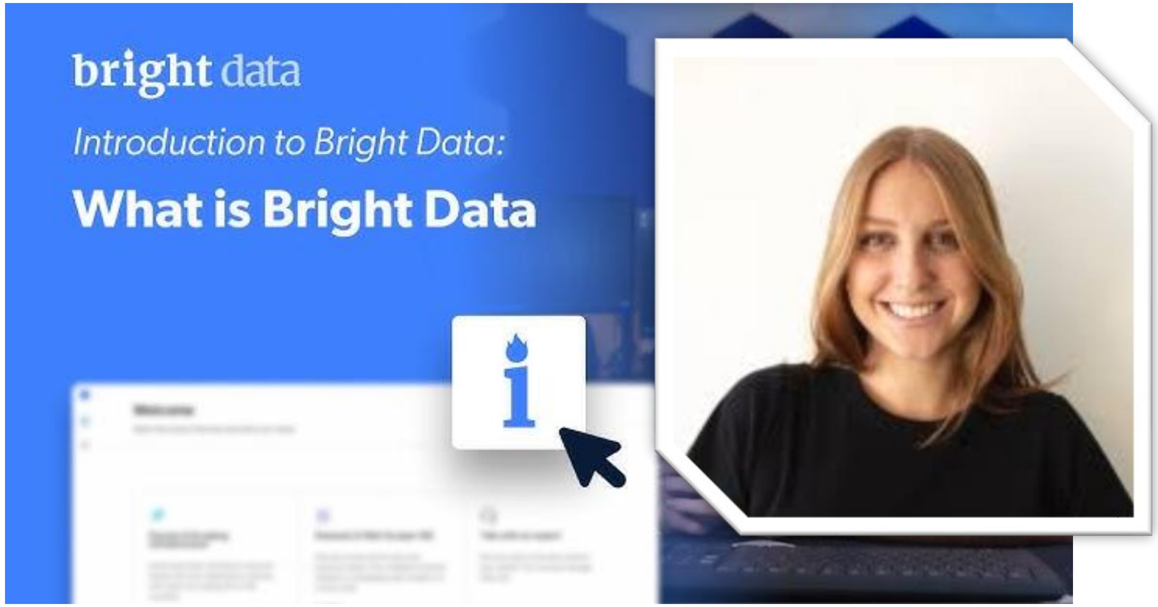
Prof Damir Cavar

dcavar@iu.edu

Workshop Schedule Overview

Date	Focus	Description
July 13	Workshop 1 – Kickoff ☑	Input on antisemitism online & team formation
July 20	Workshop 2 – Challenge #1	Scraping, annotating, exporting discourse data (Hands-On Session)
July 27	Workshop 3 – Challenge #2	Automated Content Detection: The Basics
August 5	Final Submission Deadline	Submit both challenge deliverables and documentation

Bright Data



https://www.youtube.com/watch?v=AGaiVApKfmc

Tutorial Challenge #1

Team Setup, Roles & Prerequisites

Each team should have:

- **Data Manager** → uses Bright Data for scraping
- **Portal Manager** → sets up the annotation project
- **Annotation Team** → annotates and reviews the content
- Discuss roles & workflow internally
- Don’t forget to choose a **team name or tag**!

Access to Key Links

- Challenge Description (PDF)
- Annotation Portal
- GitHub Overview
- Colab Script for Preprocessing

Agenda & Objectives

Agenda

1. Working with X’s Advanced Search
2. Scraping with Bright Data
3. Pre-Processing the Data
4. Annotations Portal Walkthrough

Objectives: What You’ll Learn Today:

- How the Annotation Portal works
- How to approach the ML challenge
- Where to find tools and datasets
- How to succeed as a team

Prerequisites & Setup

Before You Start:

A computer with internet access

A Gmail account

Twitter

A X account

Access to Google Colab (https://colab.research.google.com/)

Stop & Do Now:

Register on the Annotation Portal: https://annotate.osome.iu.edu

Check Github: https://github.com/AnnotationPortal/DatathonandHackathon.github.io/blob/main/README.md

Read Challenge Description: https://github.com/damieh1/datathon_2025/blob/main/Datathon_Challenge.pdf

Challenge #1

📍 **Subtasks include:**

1. Define your scraping focus (hashtags, user groups, topics) and document your rationale and potential biases.

2. Use the Bright Data interface to scrape at least 100 relevant user-generated posts from X.com.

3. Annotate your data using a structured definition of antisemitism and hate speech.

4. Prepare a X/Twitter dataset, and include a dataset report with label definitions, distribution information, and annotation rationale.

Earn Bonus Points

📍 **Deliverables for Challenge #1**

- Adapting and implementing an existing definition of antisemitism
- Reporting how the data was scraped and which guidelines were used to classify and annotate the data in a standardized way

Gain **+10 bonus points** by evaluating the consistency of your team's annotations using an inter-annotator agreement (IAA) metric.

This means:

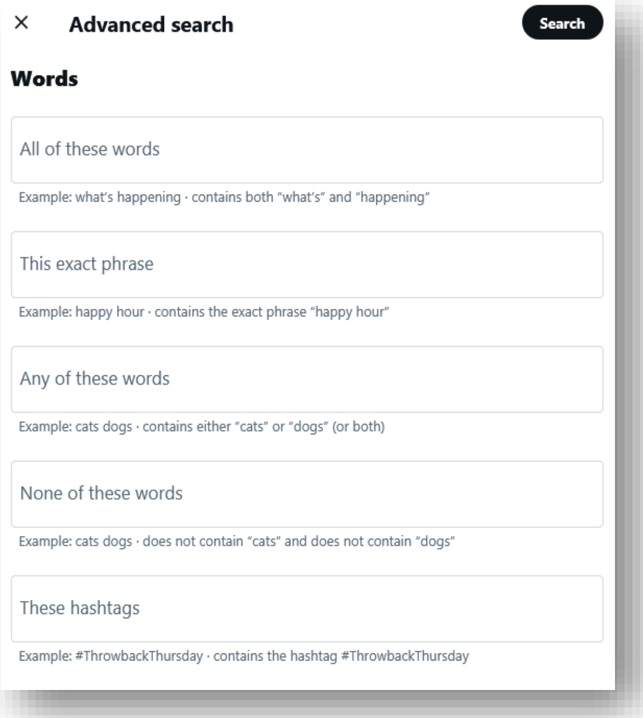
- Having at least two annotators label a shared subset of the data
- Calculating a formal agreement score, such as:
    - **Cohen's Kappa** (for binary or pairwise categorical annotation)
    - **Krippendorff's Alpha** (especially for multi-class or missing data)

Clearly report:

- Which subset was double-annotated
- Your score and a brief interpretation (e.g., "moderate agreement," "high agreement")

Working With X’s Advanced Search Function

Top of the pop-up menu



× Advanced search Search

Words

All of these words

Example: what’s happening · contains both “what’s” and “happening”

This exact phrase

Example: happy hour · contains the exact phrase “happy hour”

Any of these words

Example: cats dogs · contains either “cats” or “dogs” (or both)

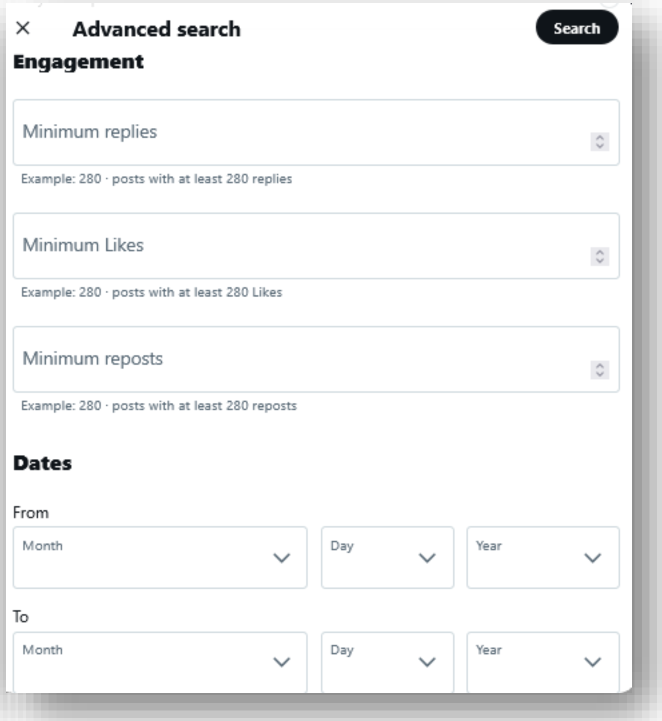
None of these words

Example: cats dogs · does not contain “cats” and does not contain “dogs”

These hashtags

Example: #ThrowbackThursday · contains the hashtag #ThrowbackThursday

Bottom of the pop-up menu



× Advanced search Search

Engagement

Minimum replies

Example: 280 · posts with at least 280 replies

Minimum Likes

Example: 280 · posts with at least 280 Likes

Minimum reposts

Example: 280 · posts with at least 280 reposts

Dates

From

Month Day Year

To

Month Day Year

1. Go to: → X (Twitter) → https://x.com/search-advanced

2. Specify dates, e.g., May 8, 2024.

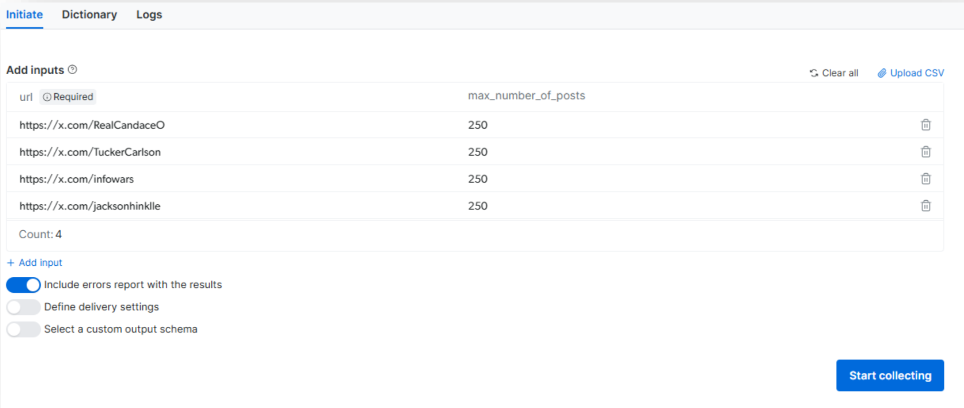
3. Click "Search.“

4. Select posts with a minimum of 200 views.

5. Go to user profiles and copy URLs to a spreadsheet.

6. Goal: Find a wide range of users who engage in online discourse.

Working With Bright Data



Initiate | Dictionary | Logs

Add inputs ⓘ

Clear all | Upload CSV

url ⓘ Required	max_number_of_posts
https://x.com/RealCandaceO	250
https://x.com/TuckerCarlson	250
https://x.com/infowars	250
https://x.com/jacksonhinklle	250

Count: 4

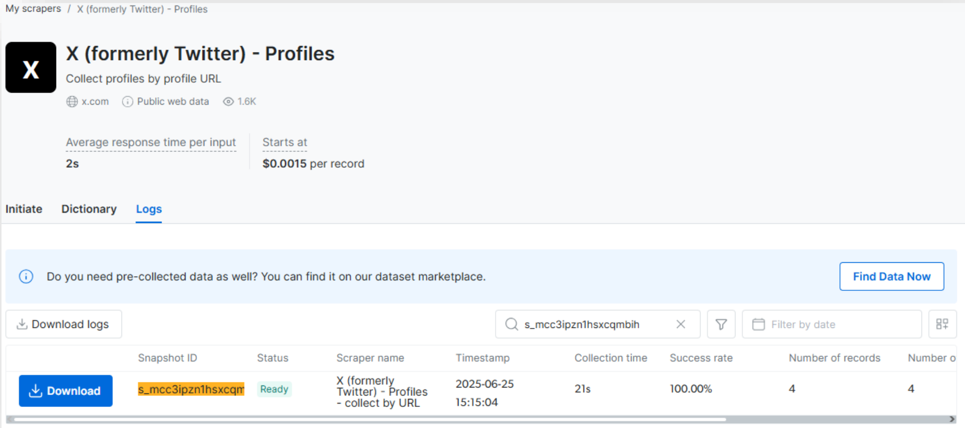
+ Add input

● Include errors report with the results

○ Define delivery settings

○ Select a custom output schema

Start collecting



My scrapers / X (formerly Twitter) - Profiles

X (formerly Twitter) - Profiles

Collect profiles by profile URL

x.com | Public web data | 1.6K

Average response time per input: 2s

Starts at: $0.0015 per record

Initiate | Dictionary | Logs

ⓘ Do you need pre-collected data as well? You can find it on our dataset marketplace. Find Data Now

Download logs | s_mcc3ipzn1hsxcqmbih | Filter by date

	Snapshot ID	Status	Scraper name	Timestamp	Collection time	Success rate	Number of records	Number o
Download	s_mcc3ipzn1hsxcqm	Ready	X (formerly Twitter) - Profiles - collect by URL	2025-06-25 15:15:04	21s	100.00%	4	4

1. Go to: Bright Data → Web Scrapes → X (Twitter) → Posts → Discover by URL → https://brightdata.com/cp/scrapers/no_code

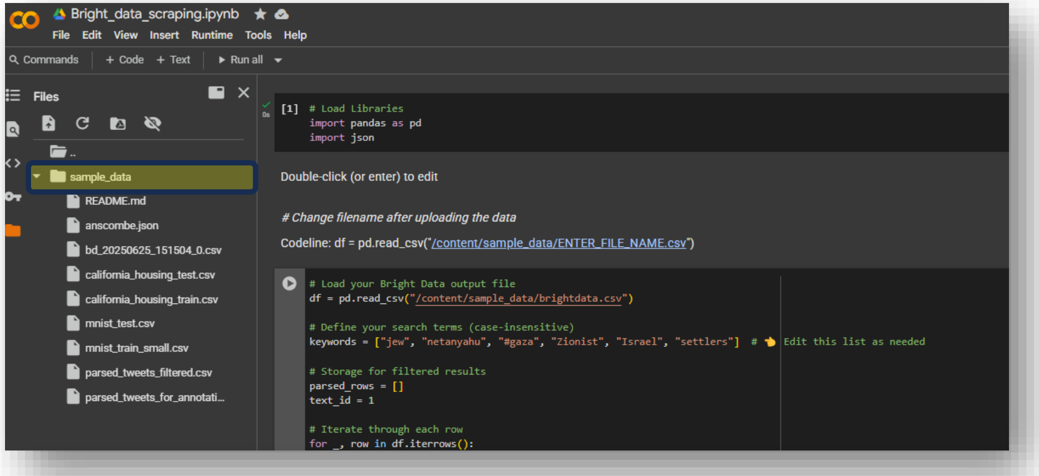
2. Click: Add Inputs → https://x.com/RandomXUser

3. Specify Number of Posts → max. 250 per User

4. Start Collecting → Runes the Query

5. Download Output as .CSV

Working With Google Colab


Bright_data_scraping.ipynb
File Edit View Insert Runtime Tools Help
Commands + Code + Text ▶ Run all
Files
..
sample_data
README.md
anscombe.json
bd_20250625_151504_0.csv
california_housing_test.csv
california_housing_train.csv
mnist_test.csv
mnist_train_small.csv
parsed_tweets_filtered.csv
parsed_tweets_for_annotati...

```
[1] # Load Libraries
    import pandas as pd
    import json
```

Double-click (or enter) to edit

# Change filename after uploading the data

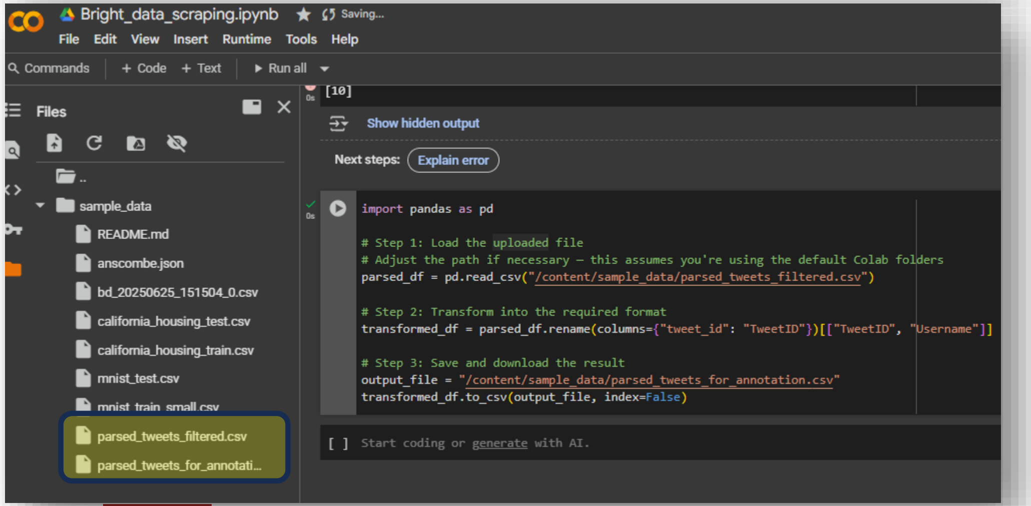
Codeline: df = pd.read_csv("/content/sample_data/ENTER_FILE_NAME.csv")

```
# Load your Bright Data output file
df = pd.read_csv("/content/sample_data/brightdata.csv")

# Define your search terms (case-insensitive)
keywords = ["jew", "netanyahu", "#gaza", "Zionist", "Israel", "settlers"]  # 👈 Edit this list as needed

# Storage for filtered results
parsed_rows = []
text_id = 1

# Iterate through each row
for _, row in df.iterrows():
```


Bright_data_scraping.ipynb
Saving...
File Edit View Insert Runtime Tools Help
Commands + Code + Text ▶ Run all
Files
..
sample_data
README.md
anscombe.json
bd_20250625_151504_0.csv
california_housing_test.csv
california_housing_train.csv
mnist_test.csv
mnist_train_small.csv
parsed_tweets_filtered.csv
parsed_tweets_for_annotati...
Show hidden output
Next steps:
Explain error

```
[10]

import pandas as pd

# Step 1: Load the uploaded file
# Adjust the path if necessary — this assumes you're using the default Colab folders
parsed_df = pd.read_csv("/content/sample_data/parsed_tweets_filtered.csv")

# Step 2: Transform into the required format
transformed_df = parsed_df.rename(columns={"tweet_id": "TweetID"})[["TweetID", "Username"]]

# Step 3: Save and download the result
output_file = "/content/sample_data/parsed_tweets_for_annotation.csv"
transformed_df.to_csv(output_file, index=False)

[ ] Start coding or generate with AI.
```

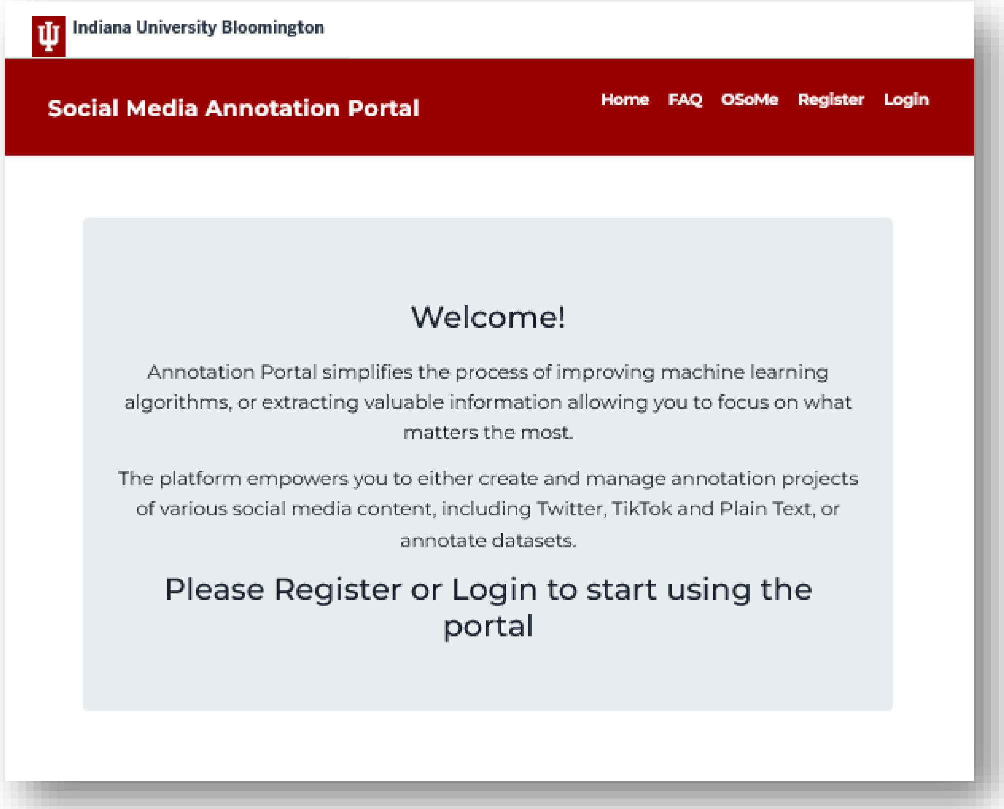
1. Go to: Google Colab → https://colab.research.google.com
2. Upload: Bright Data Output .CSV
3. Parse Data
   → Run Code on Colab
4. Prepare Data for Annotation Portal
   → Run Code on Colab
5. Download compatible .CSV Output

Click this → Open in Colab

Annotation Portal Walkthrough

Use the Annotation Portal:

1. Create New Project
2. Create a Sample
3. Create Annotation Scheme and Questions
4. Important! Do not start annotating before the schema has been fully created.
5. Export when done



Indiana University Bloomington
Social Media Annotation Portal
Home
FAQ
OSoMe
Register
Login

Welcome!

Annotation Portal simplifies the process of improving machine learning algorithms, or extracting valuable information allowing you to focus on what matters the most.

The platform empowers you to either create and manage annotation projects of various social media content, including Twitter, TikTok and Plain Text, or annotate datasets.

Please Register or Login to start using the portal

https://annotate.osome.iu.edu/

What’s next?

📝 **Work with your Team on Challenge #1**

📅 **Important Dates:**

- Workshop 1 – July 13: Kick-Off & Introduction (Team Assignment & Communication) ✅
- Workshop 2 – July 20: Hands-On Session (Data collection, preprocessing & annotation) ✅
- Workshop 3 – July 27: Introduction to Automated Detection (ML modeling & evaluation)
- Final submission deadline: August 5.

📁 **All materials available at:**

- https://github.com/damieh1/datathon_2025

💬 **Q&A – We’ll now open the floor for questions!**

Thanks for your attention

Dr. Daniel Miehling

Visiting Assistant Professor
Borns Jewish Studies Program

Samerian Foundation Research Fellow
Institute for the Study of
Contemporary Antisemitism (ISCA)

Indiana University Bloomington
damieh@iu.edu