



Data Sources

Tuğrulcan Elmas / Tj



THE UNIVERSITY of EDINBURGH
informatics



How To Collect Data

- Primary data collection (create it)
- Secondary data collection (download it)



How To “Create” Data

- Surveys, Experiments
- Sensors to Log Behavior
- Simulations
- Synthetic Data (a.k.a. stealing from ChatGPT)

... too much work for the summer school but maybe interesting as a supplement



How to “Download” Data?

1. Scrape

- **Public but the source does not share explicitly**

2. API

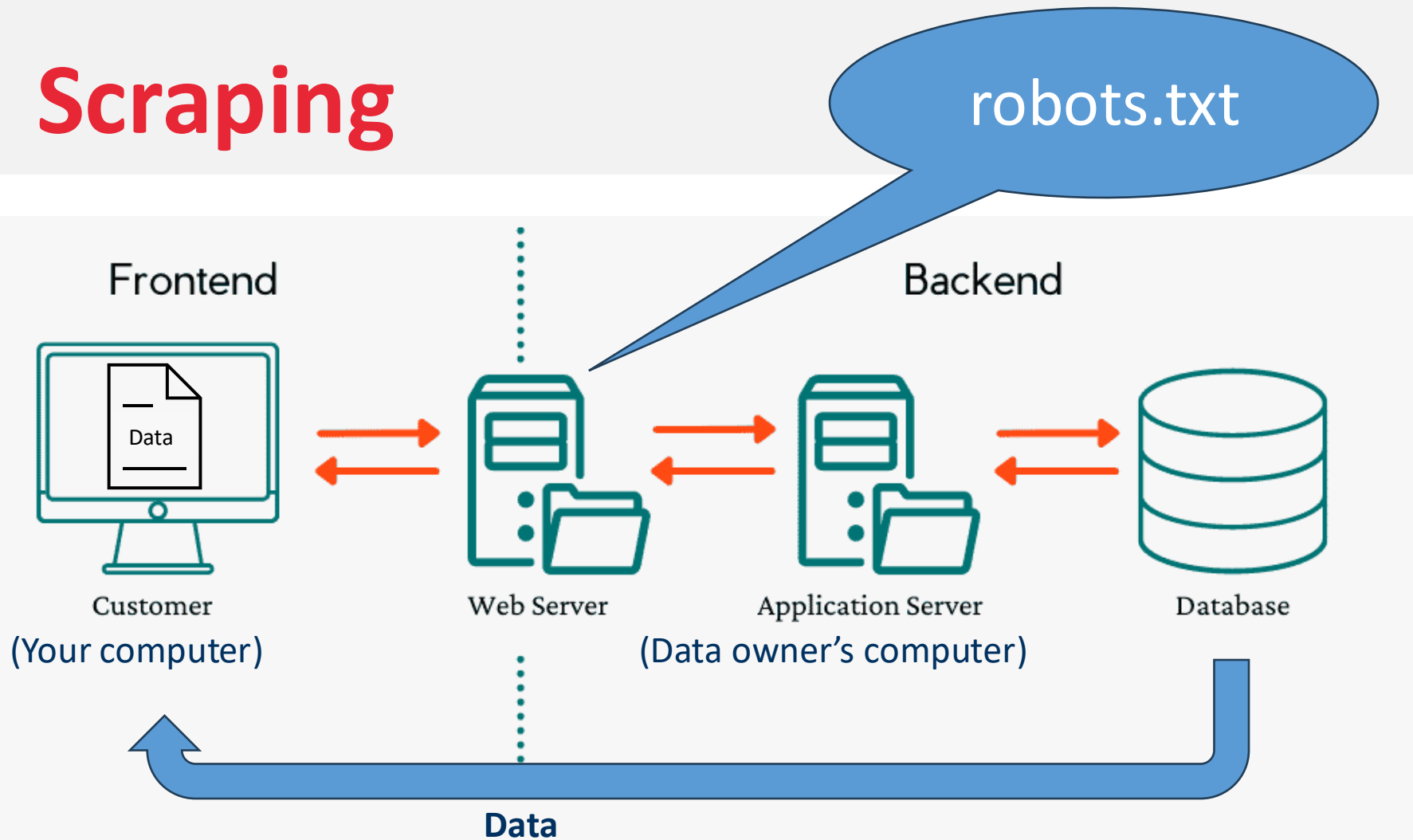
- **(may be) Public and the source shares explicitly often under conditions**

3. Download Existing Datasets

- **Somebody else scraped or API'ed for you**



Scraping



- Data goes from their computer to yours
- Cleaning the data + asking for more = scraping
- Don't ask for too much!

Should You Learn Scraping?

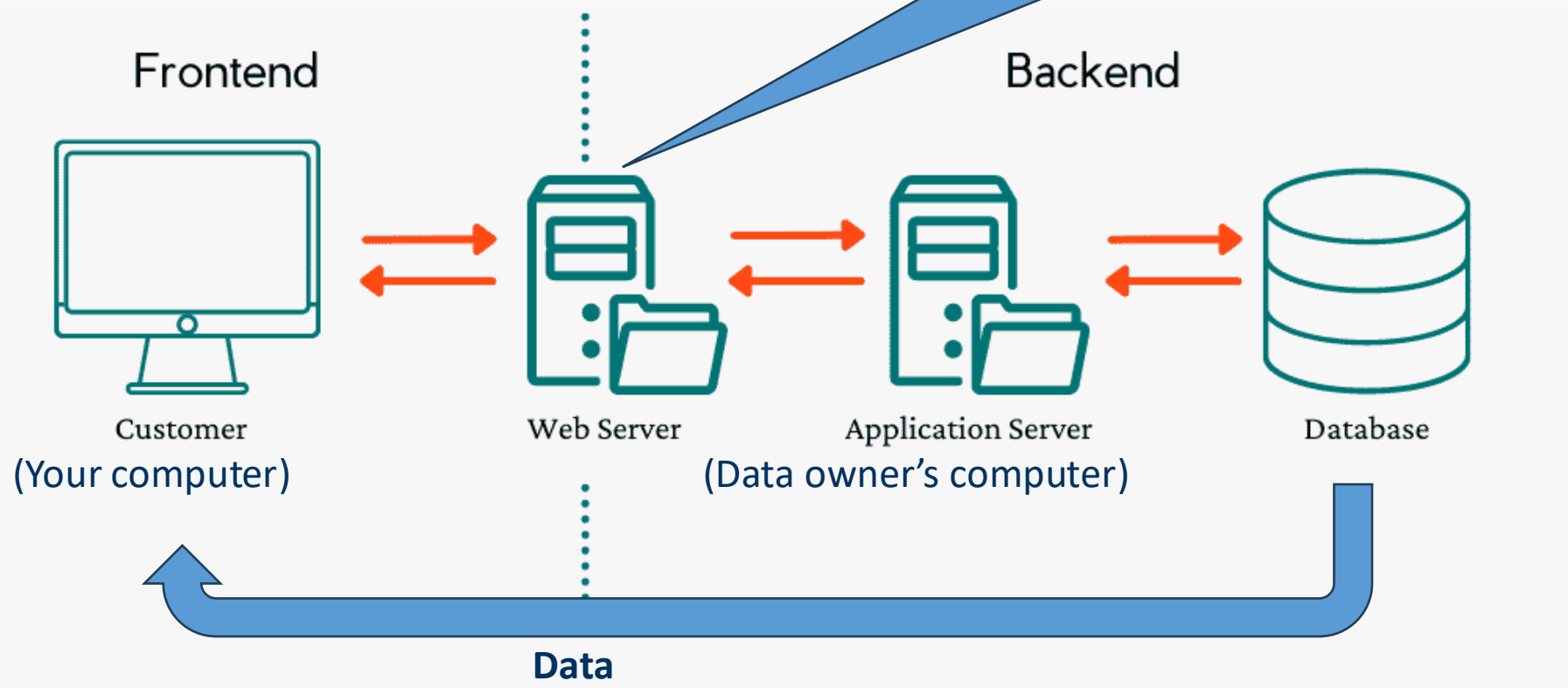
No

- Risky: Legal & Ethical issues
- Easy to Delegate
 - AI, Fiverr, UG Students, Professionals
- Hard to Maintain
 - Website Updates
- Time Better Spent Elsewhere
 - No scientific insight
 - Plenty of Public Datasets & APIs



API

RATE LIMITS!



- You ask for data explicitly & they give
- Less cleaning + ask for more under a quota
- Pay to get more

API

- Application Programming Interface
 - an interface for applications to communicate
- Requires authentication
 - Get an “API” key
- Still requires a bit of cleaning
 - Use “wrapper” libraries
- 3rd Party - Unofficial “API”s
 - API to somebody else’s data
 - If official API is not available or too expensive



Existing Datasets

- Many were there before you
- Data Repositories:
 - Crowdsourced: Kaggle, Github
 - Professional: Harvard Dataverse, Zenodo, OpenICPSR
- Dataset / Resource Papers
 - ICWSM, CIKM, Webconf
 - May be to milk citations
 - Providers are often responsive and open to collaborate



Main Sources

Administrative Data

- Census, public statistics etc.

Social Science Products

- Surveys, experiments etc.

“Digital Trace Data”

- Wikipedia edits, News comments, Google reviews, Online Gaming
- **Social Media**



Social Media Data Sources Overview



THE UNIVERSITY of EDINBURGH
informatics

Criteria on Social Media

- Can I get the data?
- Can I get the data legally and ethically?
- Is the data I get useful?
 - Relevant
 - Clean
 - Big
 - Diverse

Can I just download it?

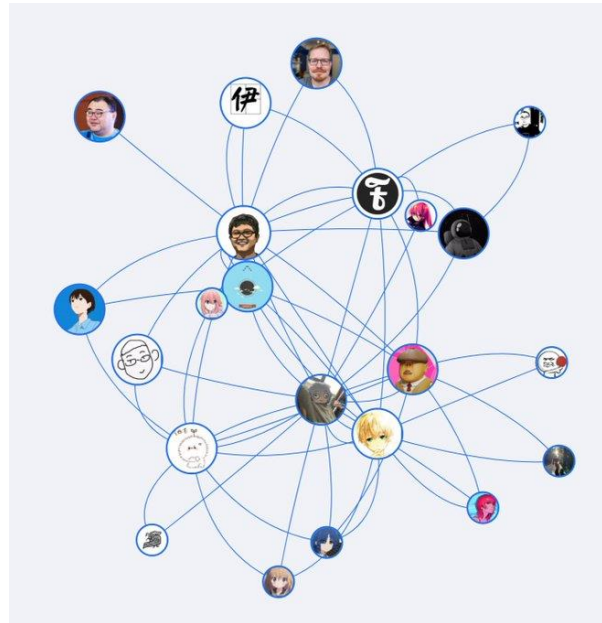


Every Social Media Platform Be Like



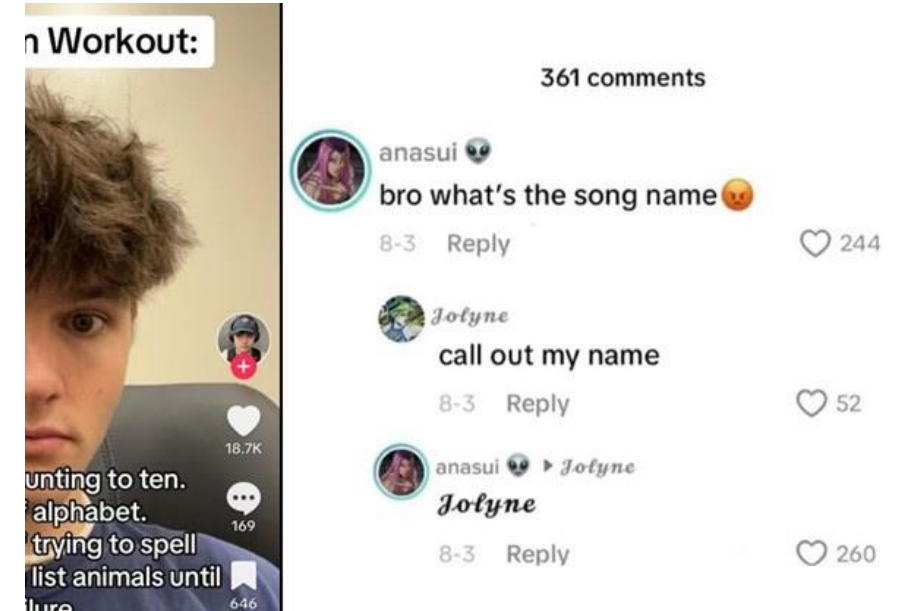
Profile

attributes, all content
by the user



Relationships

Follows, retweets



Content

Engagement metrics,
comments



THE UNIVERSITY of EDINBURGH
informatics

Relevance

- Is the data relevant to the topic of study or population of focus?
- You need to infer by some assumptions
 - Relevant subreddit -> relevant data
 - No subreddits on Twitter



TL;DR:

Reddit
X/Twitter

Good

YouTube
Telegram
TikTok
Facebook Ads

OK

Instagram
Facebook

Danger Zone

BlueSky
TruthSocial
Mastodon

what data?



Reddit

- Easy To Use API – PRAW
- Relevance Guaranteed
 - Comments labelled by post, posts labelled by Subreddit
 - On Topic
- Moderated
 - No spam
 - Inauthenticity still exists – but this is the normal for social media
- **Rich Data:** Content: Yes, Profile & Relationship: Limited
- Massive Archival Data: **Pushshift**
- Massive Literature
 - Too many related work already!



PushShift Reddit Dataset

- 2005-2023
- 1 billion submissions, 11 billion comments
- Very high coverage (but not 100%)
- Deleted content
- Accessible through the Internet Archive or Academic Torrents
- TJ has this in a database for 2022

The Pushshift Reddit Dataset

Jason Baumgartner^{1,*}, Savvas Zannettou^{2,☺}, Brian Keegan³, Megan Squire⁴, Jeremy Blackburn^{5,☺}

¹Pushshift.io, ²Max Plank Institute, ³ University of Colorado Boulder, ⁴Elon University, ⁵Binghamton University

^{*}Network Contagion Research Institute, [☺]iDRAMA Lab

jason@pushshift.io, szannett@mpi-inf.mpg.de, brian.keegan@colorado.edu, msquire@elon.edu, blackburn@cs.binghamton.edu

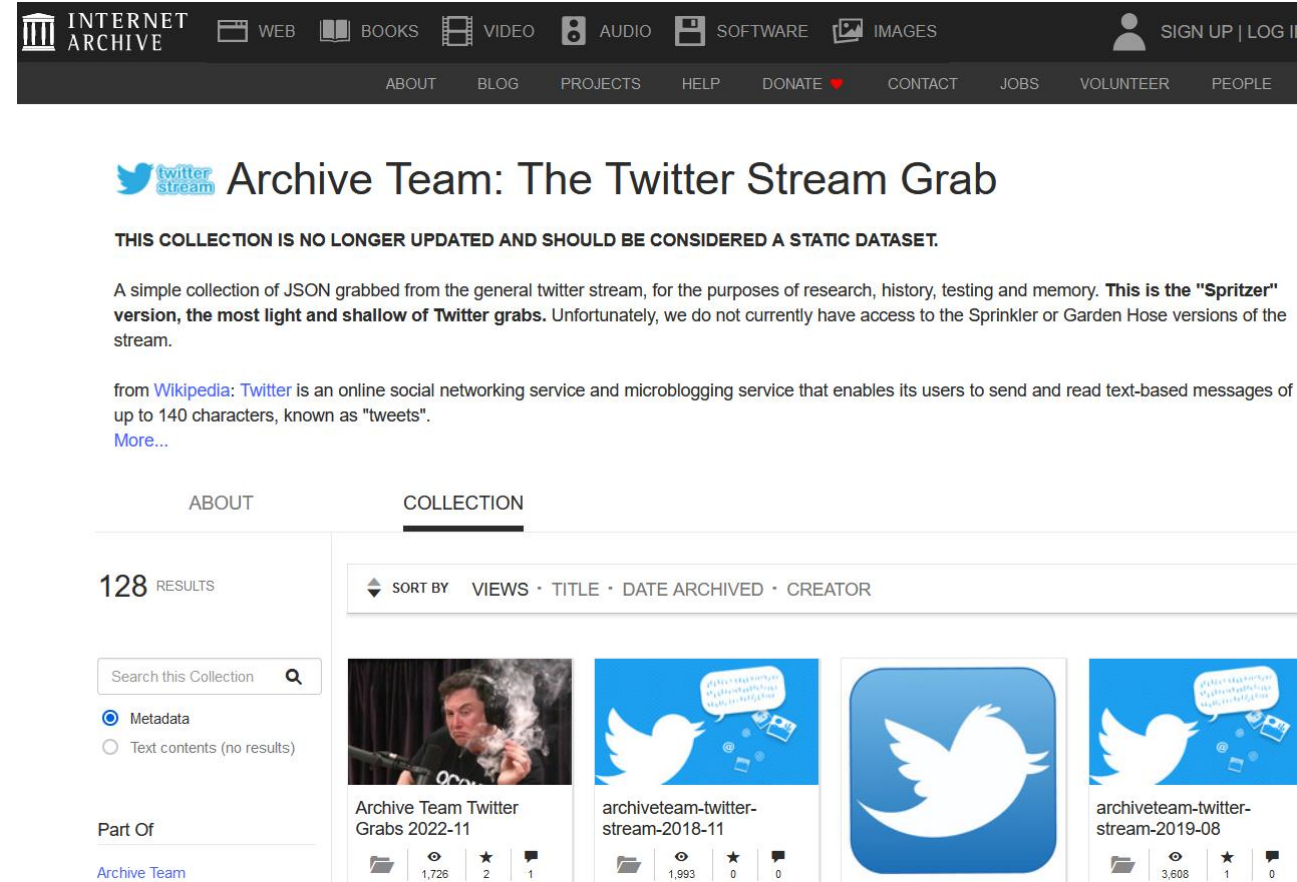
X / Twitter

- ~~Easy To Use API~~ — EXPENSIVE!
 - Scrapers available
- Relevance **NOT** Guaranteed
 - Keyword & time-based collection – recall not guaranteed
 - Maybe Off Topic
- **Weakly** Moderated
 - A lot of spammers, bots and trolls
- **Rich Data:** Profiles, Relationships, Content
- Massive Archival Data: ~~Twitter Stream Grab~~ **Gone AWOL**
- Massive Literature
 - Too many related work already!



Twitter Stream Grab

- 1% of all tweets 2011-2023
- 7TB compressed
- Download tweets from any day
- Recently made inaccessible
 - No idea why
- Tj has a copy
- Tweets are collected real-time
 - Deleted and suspended users



The screenshot shows the Internet Archive website interface. At the top, there's a navigation bar with icons for WEB, BOOKS, VIDEO, AUDIO, SOFTWARE, and IMAGES, along with links for ABOUT, BLOG, PROJECTS, HELP, DONATE, CONTACT, JOBS, VOLUNTEER, and PEOPLE. The main heading is 'Archive Team: The Twitter Stream Grab'. Below it, a notice states: 'THIS COLLECTION IS NO LONGER UPDATED AND SHOULD BE CONSIDERED A STATIC DATASET.' A paragraph explains that it's a simple collection of JSON grabbed from the general twitter stream for research, history, testing, and memory, identifying it as the 'Spritzer' version, the most light and shallow of Twitter grabs. It mentions that they don't have access to the Sprinkler or Garden Hose versions. A link to Wikipedia is provided for more information. The 'COLLECTION' tab is selected, showing 128 results. A search bar is present with 'Search this Collection' and a magnifying glass icon. Below the search bar, there are radio buttons for 'Metadata' (selected) and 'Text contents (no results)'. The 'Part Of' section shows 'Archive Team'. The collection list displays four items: 'Archive Team Twitter Grabs 2022-11' (1,726 views, 2 stars, 1 comment), 'archiveteam-twitter-stream-2018-11' (1,993 views, 0 stars, 0 comments), a large Twitter logo icon, and 'archiveteam-twitter-stream-2019-08' (3,608 views, 1 star, 0 comments).



Twitter Information Operations (Trolls!)

- “State-affiliated” coordinated accounts
- Russia, China, Iran, UAE, Saudi Arabia, Serbia, Armenia, Turkey
- No non-state related data (e.g., no Qanon)
- Ground Truth for Trolls
- Trolls are fun 😊
- Collect original data or Indiana University Vetted Version
- [“Labeled Datasets for Research on Information Operations”](#)



YouTube

- Easy To Use API
 - "YouTube Research Program" for scaled access
- Relevance Guaranteed
 - Comments labelled by video, videos labelled by channel
- Moderated
 - By channels, so inauthenticity is still a problem
- **NOT** Rich Data: Only content, comments and channel info
- Massive Archival Data: **YouNiverse**
- NOT a Massive Literature
 - Less work to inspire from
 - Many opportunities!



YouNiverse

- 2005-2019
- 137k (popular, English) channels
- 73m videos
- 8.6b comments on 20.5m videos
- Interestingly, an understudied dataset



Why Not Facebook & Instagram?

- Mostly Private Data
- Company's Stance Against Research
- Meta Content Library
 - Hard to Get Access
 - Does not work well

TECHNOLOGY

NYU Researchers Were Studying Disinformation On Facebook. The Company Cut Them Off

AUGUST 4, 2021 · 6:45 PM ET



Shannon Bond



informatics

BlueSky, TruthSocial, Mastodon, Parler, Gab

- Small scale
- Future unsure
- Echo chambers
- Twitter Clones – same research different names
- May be fun
- Easy to get full data
- Tom works with TruthSocial API



Truth Social Dataset

- Most original name ever
- 800k posts
- 454k users
- 4 million follow relationship

Truth Social Dataset

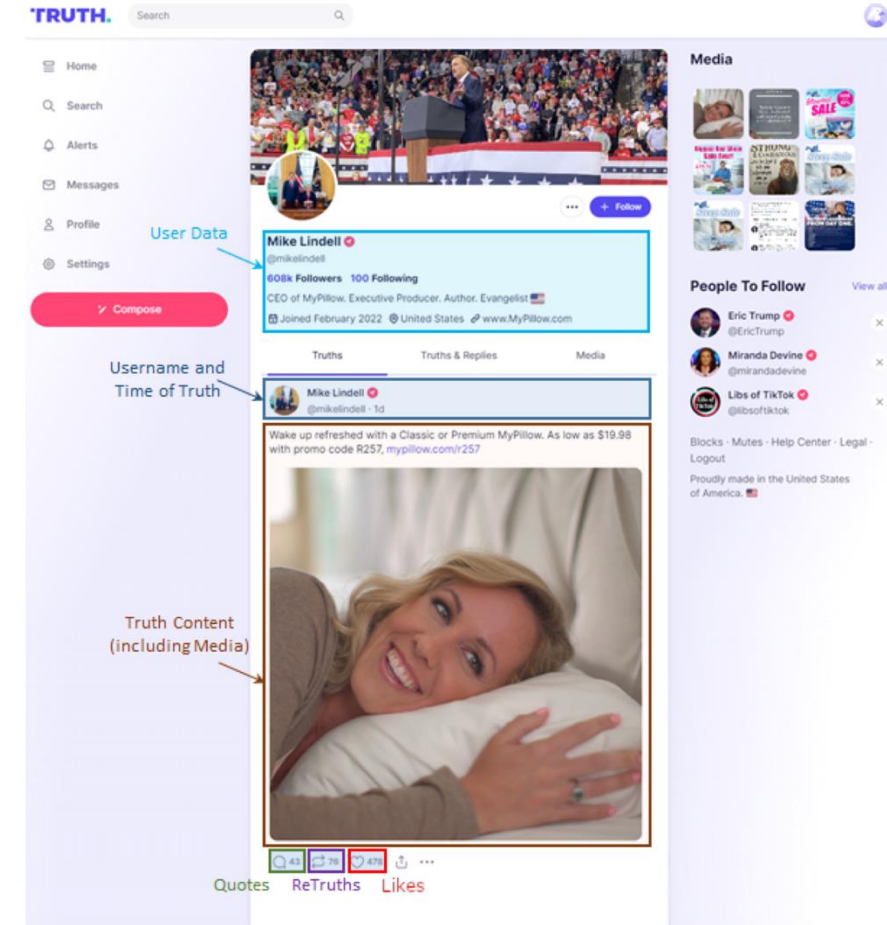
Patrick Gerard, Nicholas Botzer, Tim Weninger

Department of Computer Science and Engineering

University of Notre Dame

Notre Dame, Indiana, USA

{pgerard2, nbotzer, tweninger}@nd.edu





THE UNIVERSITY of EDINBURGH
informatics

