From Raw Data to Understanding Systemic Risks

Mapping the Data of Online Platforms

Jeff Allen

Data Access Days, 2025-09-25

What is the Integrity Institute?

The Integrity Institute is a growing community of 500+ tech workers who believe in building integrity-first online platforms that help individuals, communities, and democracies thrive.

With years of experience mitigating harms to people and communities within 55+ online platform companies, we bring seasoned, insider knowledge to leaders theorizing, building, and governing online platforms and help them put integrity front and center.

- We build and empower a community of integrity professionals in tech, giving them the tools and research they need to make online platforms safer and healthier for people and societies
- We advise online platforms, policymakers, and academics to put integrity at the heart of company governance, compliance, and tech regulation.
- We educate the public about what an integrity-first future looks like for the social internet.

Outline

Paper in collaboration with EDMO
Matt Motyl, Spencer Gurley, Jeff Allen, Sofia Bonilla
https://edmo.eu/publications/platform-datasets-challenges-insights-and-examples-for-researchers-under-article-40-of-the-digital-services-act/

- What data do platforms collect?
    - And what don't they collect?
- How is data organized internally?
- Case Study: Mastodon
    - Comparisons Between Platforms
- Mapping data to systemic risks

What data do platforms collect?

What data do platforms collect?

A lot!

What data do platforms collect?

- Can break it down into categories:
  - **User provided**
  - **Extracted from user**
  - **Created by user**
  - **Behaviors**
  - **Received from other users**
  - **Inferred**
  - **Purchased**

What data do platforms collect?

- Can break it down into categories:
  - **User provided:** Consciously shared by user. Email, age, relationship status, home location...
  - **Extracted from user**
  - **Created by user**
  - **Behaviors**
  - **Received from other users**
  - **Inferred**
  - **Purchased**

What data do platforms collect?

- Can break it down into categories:
  - **User provided:** Consciously shared by user. Email, age, relationship status, home location...
  - **Extracted from user:** Not consciously shared. Device IDs, GPS location, contacts, IP addresses...
  - **Created by user**
  - **Behaviors**
  - **Received from other users**
  - **Inferred**
  - **Purchased**

What data do platforms collect?

- Can break it down into categories:
  - **User provided:** Consciously shared by user. Email, age, relationship status, home location...
  - **Extracted from user:** Not consciously shared. Device IDs, GPS location, contacts, IP addresses...
  - **Created by user:** On platform content created by user. Posts, photos, videos, text...
  - **Behaviors**
  - **Received from other users**
  - **Inferred**
  - **Purchased**

What data do platforms collect?

- Can break it down into categories:
  - **User provided:** Consciously shared by user. Email, age, relationship status, home location...
  - **Extracted from user:** Not consciously shared. Device IDs, GPS location, contacts, IP addresses...
  - **Created by user:** On platform content created by user. Posts, photos, videos, text...
  - **Behaviors:** Generated by users activity on the platform. Views, engagement, time spent, searches...
  - **Received from other users**
  - **Inferred**
  - **Purchased**

What data do platforms collect?

- Can break it down into categories:
  - **User provided:** Consciously shared by user. Email, age, relationship status, home location...
  - **Extracted from user:** Not consciously shared. Device IDs, GPS location, contacts, IP addresses...
  - **Created by user:** On platform content created by user. Posts, photos, videos, text...
  - **Behaviors:** Generated by users activity on the platform. Views, engagement, time spent, searches...
  - **Received from other users:** Generated by other users about user. Views, blocks, engagement...
  - **Inferred**
  - **Purchased**

What data do platforms collect?

- Can break it down into categories:
  - **User provided:** Consciously shared by user. Email, age, relationship status, home location...
  - **Extracted from user:** Not consciously shared. Device IDs, GPS location, contacts, IP addresses...
  - **Created by user:** On platform content created by user. Posts, photos, videos, text...
  - **Behaviors:** Generated by users activity on the platform. Views, engagement, time spent, searches...
  - **Received from other users:** Generated by other users about user. Views, blocks, engagement...
  - **Inferred:** Predicted properties of the user. Interests, demographics, likelihood to violate policies...
  - **Purchased**

What data do platforms collect?

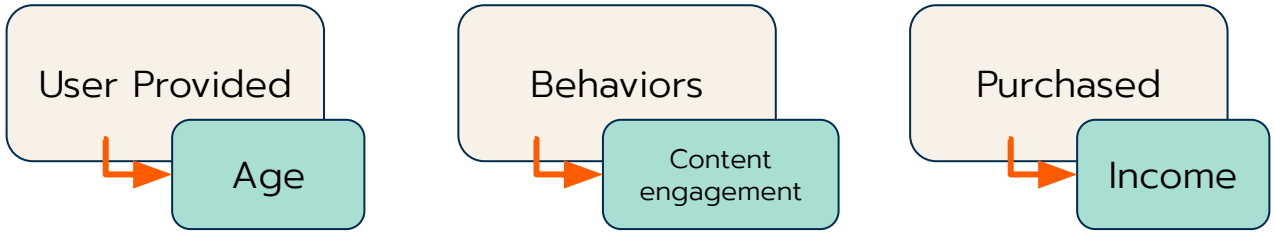
- Can break it down into categories:
  - **User provided:** Consciously shared by user. Email, age, relationship status, home location...
  - **Extracted from user:** Not consciously shared. Device IDs, GPS location, contacts, IP addresses...
  - **Created by user:** On platform content created by user. Posts, photos, videos, text...
  - **Behaviors:** Generated by users activity on the platform. Views, engagement, time spent, searches...
  - **Received from other users:** Generated by other users about user. Views, blocks, engagement...
  - **Inferred:** Predicted properties of the user. Interests, demographics, likelihood to violate policies...
  - **Purchased:** Data purchased from third parties. Income, consumer behavior, credit score...

Why do they collect all this data?

- It is in their business interests to!
- Broadly speaking, three big reasons
  - Inherently required for the platform experience
  - Advertising revenue!
    - More data -> more and better targeting options for advertisers
    - Leads to higher revenue per impression on ads
  - Tailoring the user experience
    - More data -> more customization and personalization for platform experience
    - Can lead to better overall experiences for users
    - Also leads to more time spent on platform, and thus more impressions on ads

Why do they collect all this data?

- **Example: Advertisers want to target wealthy new parents**
  - Age of user relevant
    - Mid 20's to early 30's age bucket most likely to be relevant
  - Content the user engages with is relevant
    - Can identify interest in "baby" related or "parenting" related content
  - External data sources
    - Can combine with external data around income or net worth, tied to email



What don't they collect?

Perfect knowledge is hard to come by

- Not all platforms gather all details from users
  - Many platforms don't ask users age
    - And almost none verify ages outside of new regulation
  - "Age" might be a predicted thing
    - Predicted age
    - Probability of being in age range
  - So asking for "Examples of content engaged with by users between 13 and 30"
    - If platform doesn't collect age, how do you map that to predicted ages?
- Some data platforms might explicitly avoid
  - Changing platform experience by political party affiliation might cause backlash
  - Platform may explicitly not collect or predict things around political party
  - So, "Examples of content engaged with by people in X political party" might not be meaningful

What don't they collect?

Perfect knowledge is hard to come by

- Content topics might not be explicitly known, or might only be predictions
    - "Examples of most popular civic content": There may not be a topic classifier for that.
    - How to map that question to the predicted scores from the classifier?
- Inauthentic accounts or FIMI accounts
    - Inauthenticity can be a complicated assessment to make
    - Sometimes they use a known IP/device
    - But sometimes it's looser affiliations

April 24, 2025

Busting myths: do mobile devices listen to us all the time?

by Sergey Korol

How is data organized internally?

Data Engineering 101

- Data internally can be very chaotic
- Will vary wildly between companies
- There will be entire company functions around data
    - Typically Data Engineering, Analytics, and Science teams
- But there are some useful principles in how it works
- **Common framework: "Fact Tables" and "Dimension Tables"**

Dimension Tables

- "Dimension" tables are data tables that
  - Collect a large amount of information
  - About entities or objects on the platform
- Each row will represent a single instance of an object
  - Represent "nouns" in the "language of data"
- For example, a user dimension table will collect
  - User id
  - Name
  - Email
  - Date created/joined
  - Any other user provided information

Dimension table data dictionary example

Variable	Description	Data Type	Retention	Restricted access?
user_id	Assigned numeric ID for each user	INTEGER	90	N
name	User's name	STRING	90	Y
email	User's email address	STRING	90	Y
date_joined	Date user created account	DATE	90	N
country	User's specified country	STRING	90	N
age	User's specified age	INTEGER	90	N

Dimension table data example

user_id	name	email	date_joined	country	age
12	John Doe	johndoe@example.com	2013-01-01	UK	45
53	Jane Doe	janedoe@example.com	2014-01-01	US	63
891	Fulano	fulano@example.com	2020-01-01	MX	32
156	Mengano	mengano@example.com	2010-01-01	CO	22
932	Zutano	zutano@example.com	2025-01-01	ES	15

Fact Tables

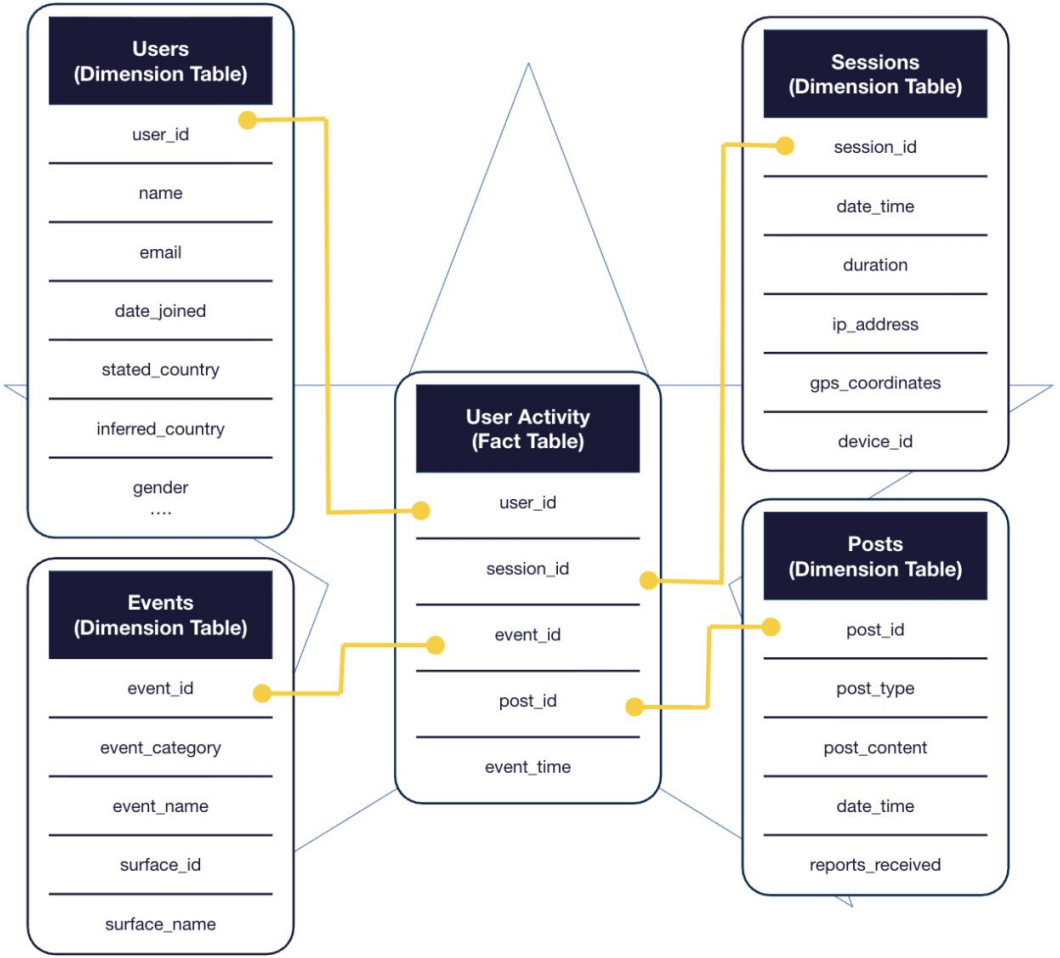
- "Fact" tables are data tables that
  - Collect facts about events on the platform
- Each row will represent an occurrence of something
  - Represent "verbs" in the "language of data"
- They will often take the form of "event logs"
  - Entity taking an action
  - Action taken
  - The entity that was acted upon
- So a content feed fact table will include
  - User ids of content feed users
  - An action they took (Viewed, engaged)
  - Content that was viewed/engaged with

Variable	Description	Data Type	Retention	Restricted access?
user_id	Assigned numeric ID for each user	INTEGER	30	N
session_id	Assigned numeric ID for each user session	INTEGER	30	N
event_id	Label for each event type	STRING	30	N
post_id	Assigned numeric ID for each post	INTEGER	30	N
event_time	The time when event occurred	TIMESTAMP	30	N

user_id	session_id	event_id	post_id	event_time
12	5021	LIKE	90	2025-01-01T00:10:10Z
12	5021	VIEW	93	2025-01-01T00:20:15Z
12	5022	LIKE	93	2025-01-01T00:20:20Z
53	5023	REPLY	143	2025-01-01T00:30:23Z
53	5023	SHARE	143	2025-01-01T00:30:24Z

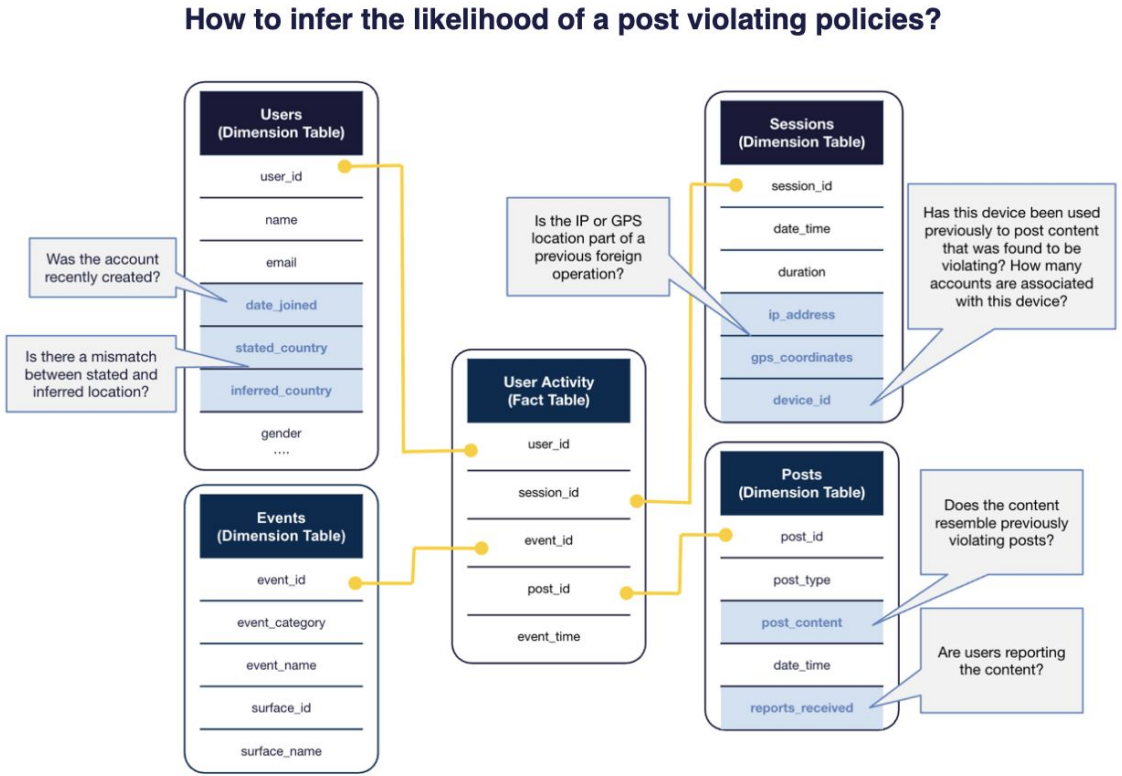
“Star Schema”

- Common jargon for this framework
- Involves core event logs as fact tables
- And unique ids to link to dimension tables



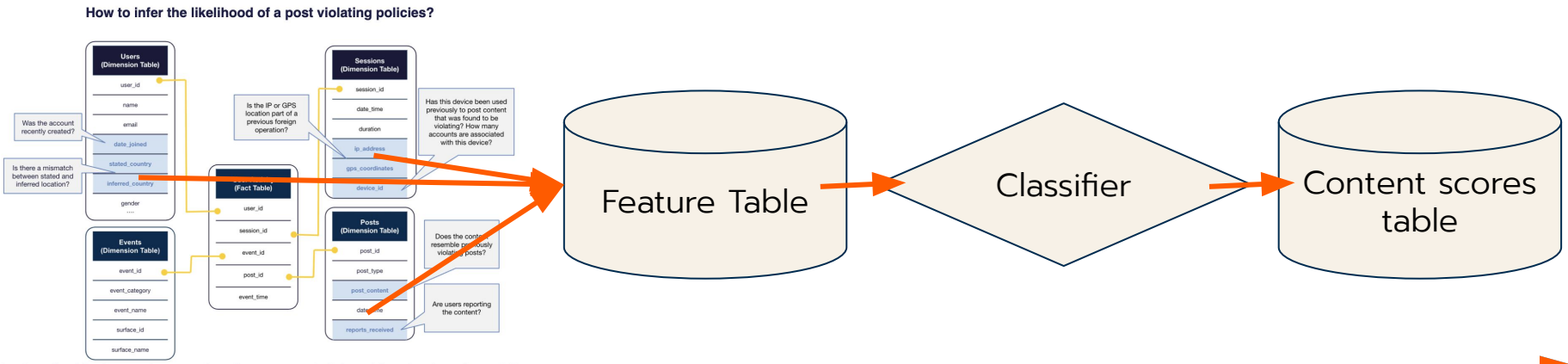
Example: Predicting Violating Posts

- Predicting is a post violates relies on lots of data
  - The content itself
    - Dimension table
  - Account that posted it
    - Dimension table
  - User responses
    - Fact table aggregations



Example: Predicting Violating Posts

- Data will be joined to feed the features of the classifier
- Classifier output will go to a dimension table



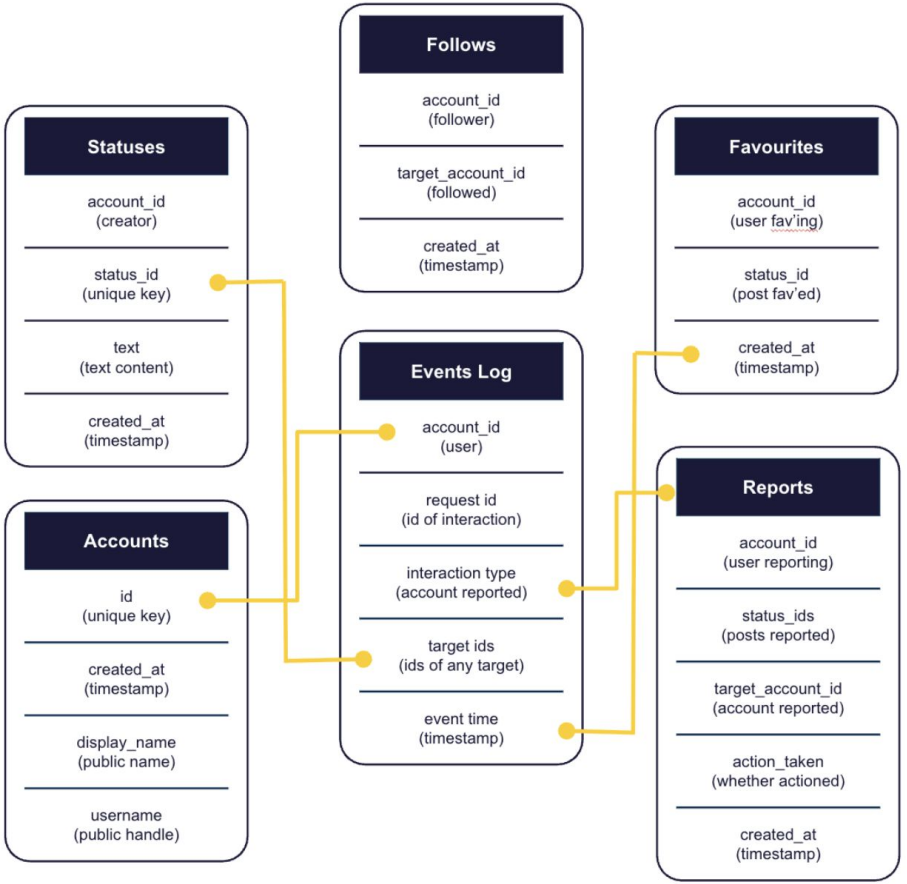
Case Study: Mastodon

Data Tables in Mastodon

- Fun fact: There are open source platforms!
- And they generate data tables!
- Mastodon is a really great example
- Can act as a "minimalist" stand in for social media platforms
    - If Mastodon has the data, VLOPs almost certainly do
- Mastodon has "dimension" like tables and "fact" like tables

Data Tables in Mastodon

- Mastodon uses a "star schema" type framework



Comparisons Between Platforms

APIs vs. Raw Data

- Not all platform APIs stack up against Mastodon or what we need

	Facebook	X	TikTok	Mastodon
FIMI				
Text	post_media_text	text	video_description hashtag_names region_code	statuses media_attachments
Engagement	post_reactions post_comments post_shares	organic_metrics nonpublic_metrics promoted_metrics	like_count comment_count share_count view_count favourites_count	favourites_count replies_count reblogs_count views (from event log)
Reports				account_warnings reports report_notes
Classifier		possibly_sensitive	video_label	follow_recommendation_suppressions global_follow_recommendations
Survey Data Manual review				account_moderation_notes account_notes admin_action_logs appeals
User Controls				blocks conversation_mutes custom_filters follow_recommendation_mutes
Creator Information			region_code	Location from IP of creator

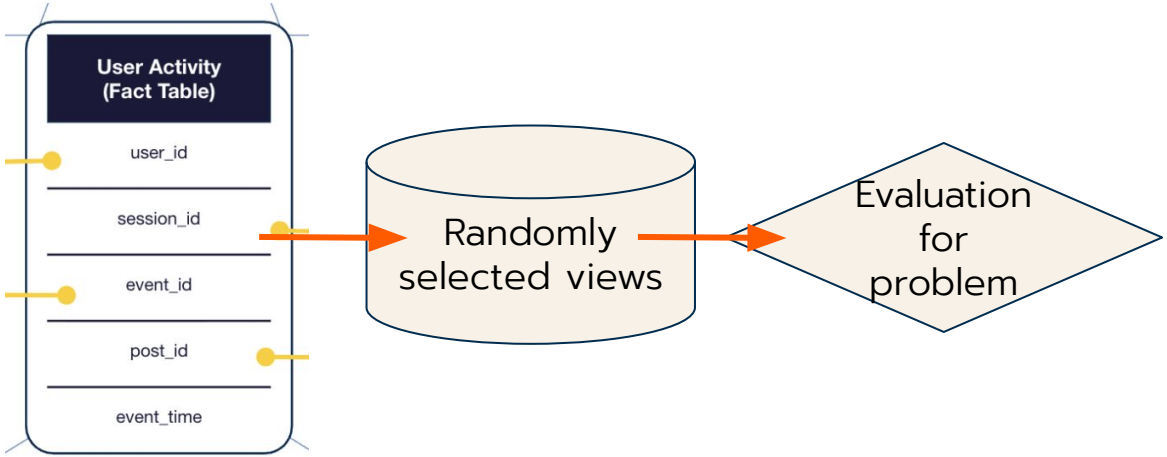
From Data to Systemic Risks

Tables useful for systemic risks

- How do we go from the core dimension and fact tables to a dataset that can answer questions around systemic risks?
- **Key example: Can we produce a dataset that allows researchers to validate prevalence metrics for violating content? Or estimates prevalence for novel problems?**

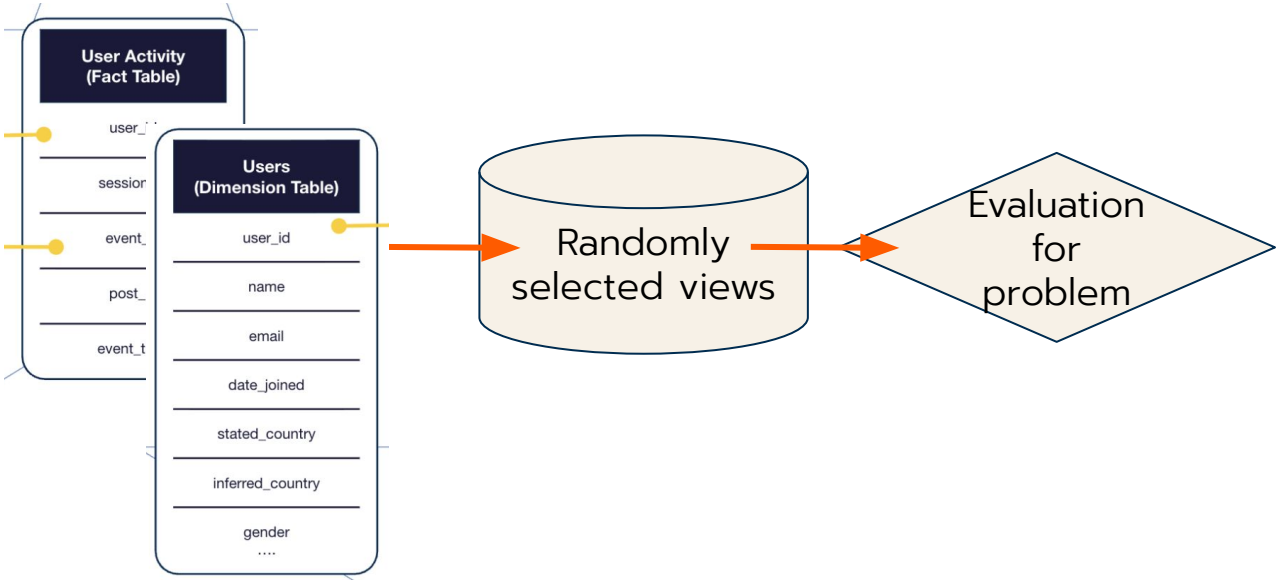
Tables useful for systemic risks

- **Key example: Can we produce a dataset that allows researchers to validate prevalence metrics for violating content? Or estimates prevalence for novel problems?**



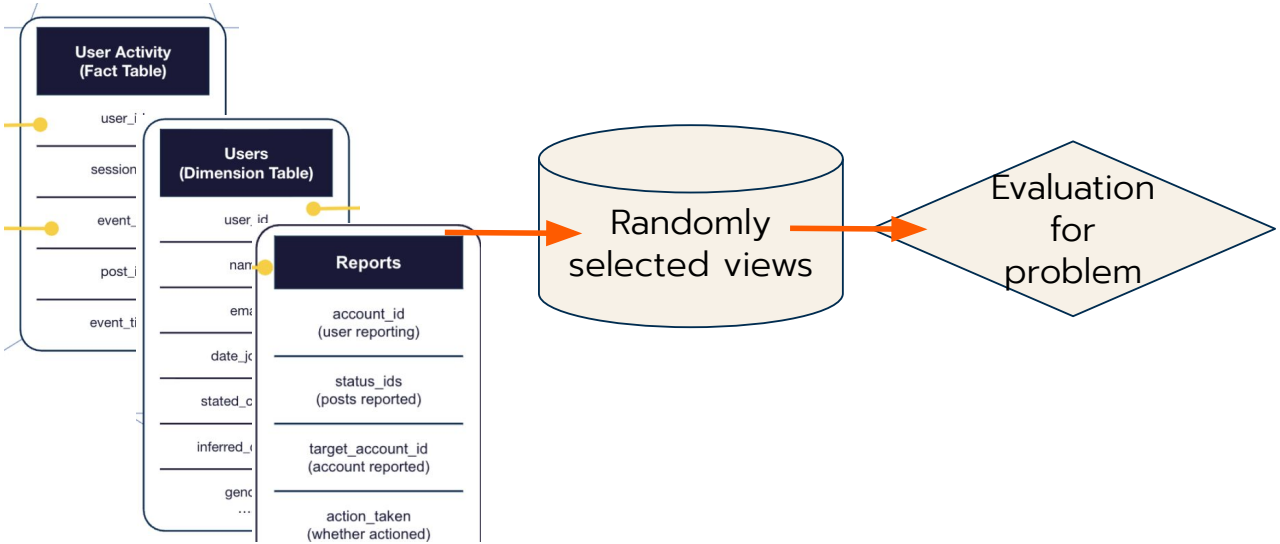
Tables useful for systemic risks

- **Key example: Can we produce a dataset that allows researchers to validate prevalence metrics for violating content? Or estimates prevalence for novel problems?**
- Can also combine with user attributes to filter to prevalence by country or age!



Tables useful for systemic risks

- **Key example: Can we produce a dataset that allows researchers to validate prevalence metrics for violating content? Or estimates prevalence for novel problems?**
- Can also combine with user attributes to filter to prevalence by country or age!
- Can also combine with either predicted violation scores or reports to get stratified random sample!



Conclusion

Key Takeaways

- Platforms collect a lot of data! (But likely won't have perfect data for everything you want)
- Understanding how platform data is stored internally can help you
    - Determine the most likely datasets that will be useful
    - Help you navigate any conversation with platforms about data access
- There are open source example to learn from and point to!
    - If Mastodon has a data you need, every VLOP will
    - Other platforms to look to, like ElasticSearch for ranking systems
- Hopefully we get real access that lets us meaningfully answer questions about systemic risks
    - There are platform experts who can help! (Integrity Institute is one!)
- Check out our paper for more details!
    - Matt Motyl, Spencer Gurley, Jeff Allen, Sofia Bonilla
    - https://edmo.eu/publications/platform-datasets-challenges-insights-and-examples-for-researchers-under-article-40-of-the-digital-services-act/