[data project name goes here]

Map to other resources:

1. Brief presentation on project idea (presented to EDMO, EC, CDT-E on May 15, 2024)
2. Spreadsheet containing notes of types of data
3. API Walk-through Workshop (2024-05-30)
4. Wireframe
5. GitHub repo

Some UI examples I like for potential presentation style:

- Focus on Features
- Hadley's R for Data Science (written in Markdown, updated in Quarto, hosted by Netlify)[github repo]
- Something else?

High-level Goal

Help people outside of industry know what they don't know.

Specific Goals and Table of Contents

1) What data do platforms collect
    a) Spreadsheet with high level examples
        i) Explicitly provided by user
        ii) Behavior of user
        iii) Data extracted from a user's devices / networks
            (1) This could include cross-app tracking via cookies, pixel, etc
        iv) Inferences about the user based on what they provide, their behavior, and more complex things like machine learning / AI
        v) Purchased third-party data on users
2) What do platforms generate from those data
    a) Machine learning and/or AI predictions
        i) Covers all sorts of domains like:
            (1) What kinds of ads would that person be likely to click on?
            (2) Whether that person is a bot?
            (3) Whether that person is an abusive account?
            (4) What kinds of content is that person likely to engage with?
3) How do platforms use those data
    a) Serve personalized ads
    b) Make account and content recommendations
    c) Remove harmful content / accounts

4) How is the data stored
    a) Dim tables
        i) Often in a tidy format -- 1 row per user, or group, or page, or whatever
        ii) Columns more likely to be what social scientists are used to (e.g., if column is gender, then the cell for a user might be man/woman/other)
    b) Fact tables
        i) Often in long, semi-/non-structured format
        ii) Lots of JSON / map / array columns of differing lengths
    c) Many hive tables across many namespaces/warehouses
        i) Takes additional steps to join data from tables existing in different namespaces / warehouses
    d) Heavily partitioned (e.g., often by date, but could also be by account type, action type, etc)
    e) Differing sensitivity / security for tables and/or columns within tables
    f) Retention
        i) Dim tables usually have longer retention (e.g., 30-90 days)
        ii) Fact tables typically have shorter retention (e.g., between 1 and 90 days)
        iii) Once max retention duration is reached, data may be moved to “cold storage”
            (1) Cold storage is usually much older hardware, meaning it’s slower to retrieve / work with
            (2) User ids are randomized, so they can no longer (easily) be joined to data in the active storage tables (though there is a decoder ring to figure this out)
        iv) Some exceptions (at least at some platforms) -- there are some tables containing data used to train ML models, and I don’t know that those ever move into cold storage
5) How do you extract the data
    a) Disclaimer: Very carefully -- the data are bigger than most humans can fathom, and if you’re not careful, you could write a query that costs thousands of dollars and/or breaks/freezes and/or breaks/freezes your computer (if trying querying data to your local machine which is unlikely to have petabytes of RAM)
    b) Internally via SQL, NoSQL, Presto, Spark, etc.
        i) Include sample (simulated) data
        ii) Include sample queries working with:
            (1) Dim tables
            (2) Fact tables
            (3) Joins across dim and fact tables
    c) Externally, probably via API
        i) Include list of APIs for for VLOPSEs, with links
            (1) TikTok
                (a) Requirements
                (b) Details
            (2) Facebook

(a) ~~CrowdTangle~~ (deprecated)
(b) Content Library
(i) Details
(c) FORT
(i) Datasets Details
(ii) API Details
(3) Instagram
(a) Same as FB (under META)
(4) Google Search
(a) Requirements
(b) Details
(5) Google YouTube
(a) Requirements
(b) Details
(6) Google Shopping
(a) Requirements
(b) Details
(7) Google Play
(a) Requirements
(b) Details
(8) Google Maps
(a) Requirements
(b) Details
(9) Snapchat
(a) Requirements
(10) LinkedIn
(a) Details
(b) Requirements
(11) Pinterest
(a) Requirements
(12) X (Twitter)
(a) Details
(13) Reddit
(a) Details
(14) Wikipedia
(a) Details
(15) Amazon Store
(a)
(16) Alibaba AliExpress
(a)
(17) Apple AppStore
(a)
(18) Bing
(a) Details (more likely a search api not a research api)

(19) Booking.com
(a)
(20) Zalando

ii) Describe how the platforms probably determined the variables available via the API

iii) Describe how the data being pulled b the API are stored

d) Externally, possibly via a clean room

i) Sometimes researchers need special access to more sensitive data that companies are unwilling to make available off-premises and will invite researcher on campus

ii) When on-campus, they may be given a device to use to query the data locally and work with while on-campus

iii) Known issues with clean room data access:

(1) Data persistence is low; often whatever data were queried or accessed in that session is wiped at the conclusion of the session, even if the researcher isn't done and needs to come back to continue the work

e) Public Interest Scraping (legal in EU because of DSA 40.12)

f) Transparency Centers

i) TikTok

(1) Transparency Center

ii) Facebook

(1) Transparency Center

(2) Widely Viewed Content Report

iii) Instagram

(1) Transparency Center (same as FB)

iv) Google Search

(1) Ads Transparency Center

(2) Transparency Center (All Google under 1 center)

v) Google YouTube

(1) Transparency Center (All Google under 1 center)

vi) Google Shopping

(1) Transparency Center (All Google under 1 center)

vii) Google Play

(1) Transparency Center (All Google under 1 center)

viii) Google Maps

(1) Transparency Center (All Google under 1 center)

ix) Snapchat

(1) Safety and Privacy Hub

x) LinkedIn

(1) Transparency Center

xi) Pinterest

(1) Transparency Report (H1 2023)

xii) X (Twitter)

(1) Transparency Center

xiii) Reddit (not VELOPSE, but probably generally useful

(1) Transparency Center

xiv) Wikipedia

(1) Transparency Report (H2 2023)

xv) Amazon Store

xvi) Alibaba AliExpress

xvii) Apple AppStore

xviii) Bing

xix) Booking.com

xx) Zalando

g) GitHub

i) X (Twitter)

ii) Meta

iii)

h) External Surveys

i) Neely Center Social Media Index

ii) Neely Center Artificial Intelligence Index

iii) Neely Center Mixed Reality Index

iv) Pew Research

v) Eurobarometer

vi) Gallup

vii) Bipartisan Policy Center

viii) Oxford Digital News Report

i) Data disclosed through other means

i) Some (typically academic) publications link to the Open Science Framework, Harvard's Dataverse, UMichigan's ICPSR, etc where platform data used in those studies are publicly available

ii)