

Day 5 - The Web, Computational Infrastructure, and Innovative Datasets

UMN CSS Workshop 2025

Instructor: Alvin Zhou

Group

- Group 1
 - Gretchen Corcoran
 - Jikai Sun
 - Shreepriya Dogra
 - Jialu Fan
- Group 2
 - Jiacheng Huang
 - Paulina Vergara Buitrago
 - Jong Won Lee
 - Eun Sun Kyoung
- Group 3
 - Sijin Chen
 - Michael Ofori
 - Jinny Zhang
 - Dongwook Kim
- Group 4
 - Raj Wahlquist
 - Nicole Marie Klevanskaya
 - Wenhui Cheng
 - Rita Rongwei Tang
- Mixed background and coding skills
- Group members who are confident in their coding skills, please help other members during the afternoon coding labs

Learning Goals

- Understand behavioral trace data
- Overview of prior-decade data collection methods (API)
- Post-API era: Scraping, data donation, collaboration, audit, etc.
- Discuss data ethics

A Paradigm Shift in Social Science

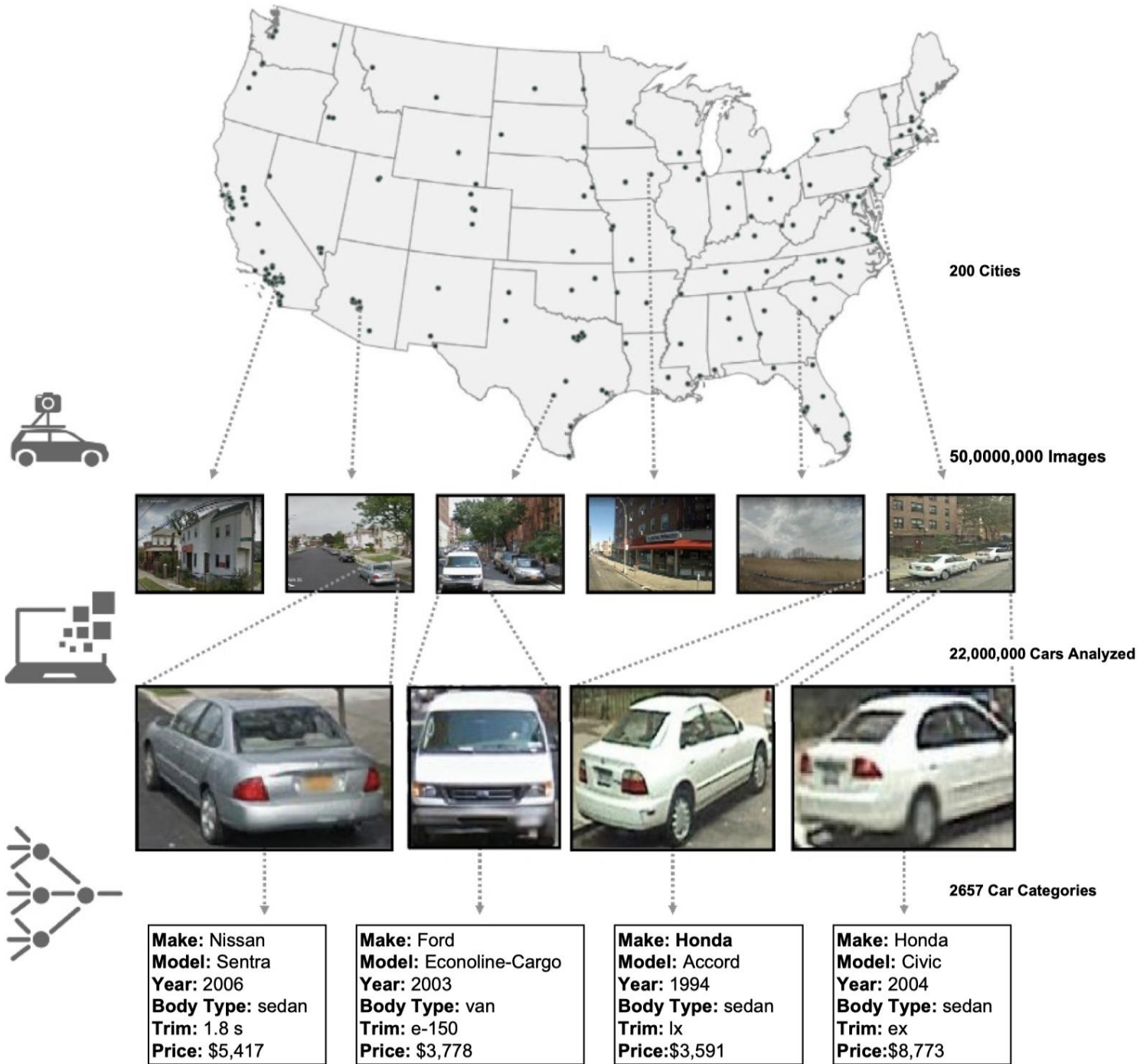
- From surveys and self-reports → behavioral traces and digital exhaust
- Surveys: What people say they do
- Traces: What people actually do
- **Both are biased**, but differently
- Key advantage: Less filtered, less recall bias
- Key risk: Less context, harder to interpret meaningfully

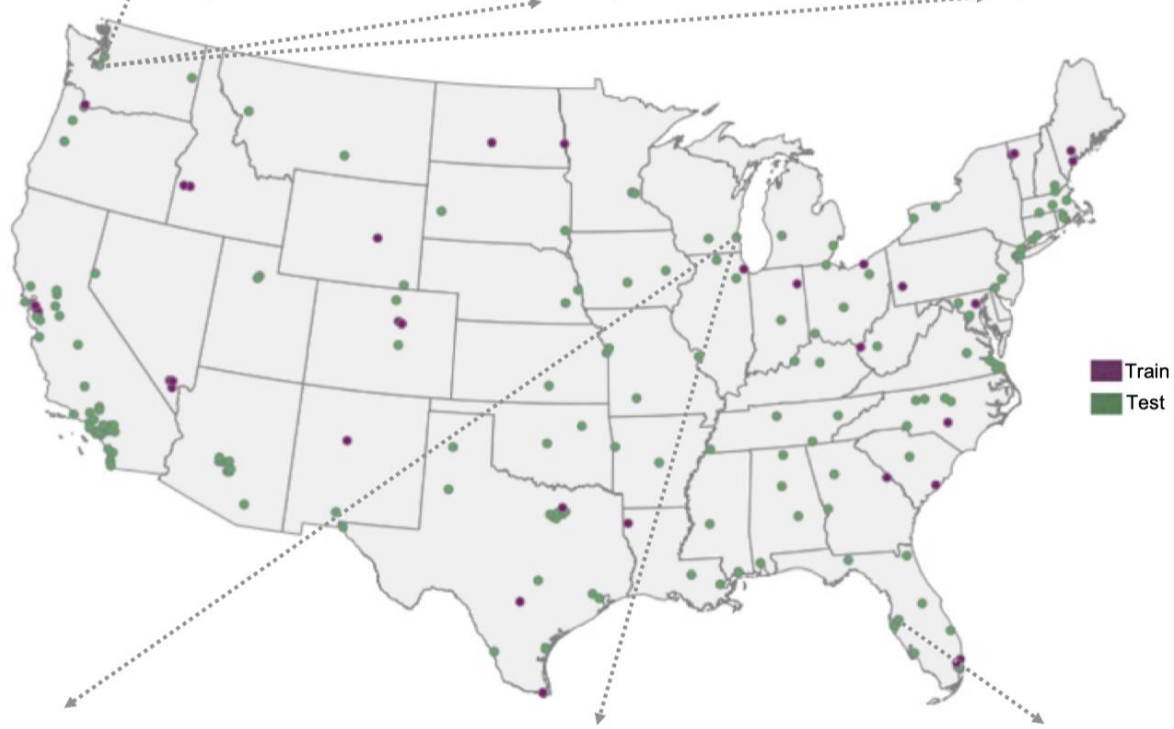
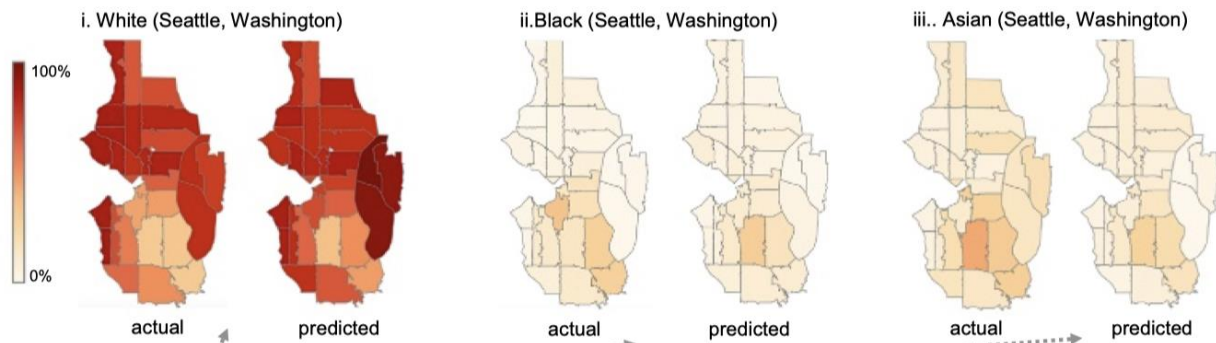
Behavioral Trace Data: Beyond Social Media

- First of all, we should not equate **behavioral trace** to **online behavioral trace** or **online social media behavioral trace**
- Online social media activity \neq entire digital life \neq entire life
- Browsing logs
- Search histories
- Location & mobility data (e.g., Google Maps)
- Environmental imagery (e.g., Street View, satellite)
- Wearables, app logs, e-commerce trails

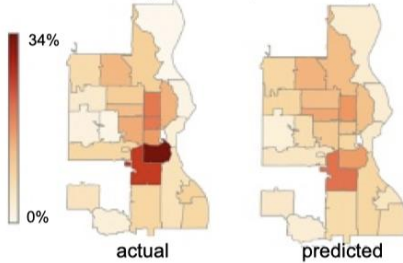
Gebru et al. (2017): Google Street View

- RQ: Can we infer socioeconomic patterns from the environment people live in?
- Analyzed 50 million Google Street View images across 200 U.S. cities.
- Used deep learning to identify 22 million vehicles by make, model, and year.
- Found strong correlations between vehicle types and:
 - Income levels
 - Educational attainment
 - Racial composition
 - Voting patterns:
 - More sedans → likely Democratic; More pickups → likely Republican

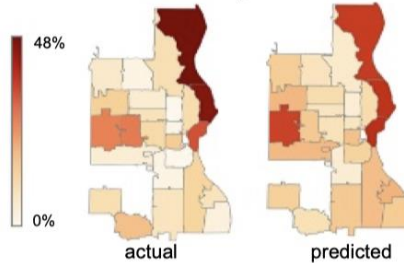




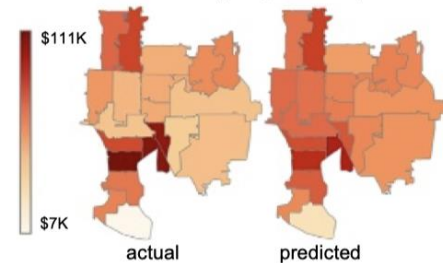
iv. Less than High school (Milwaukee, Wisconsin)



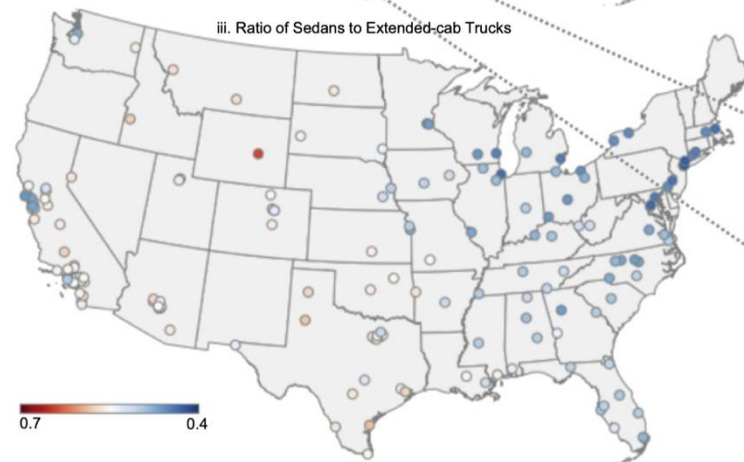
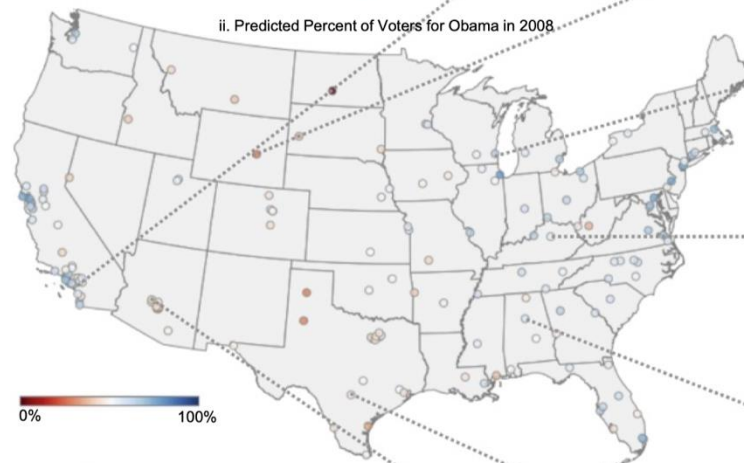
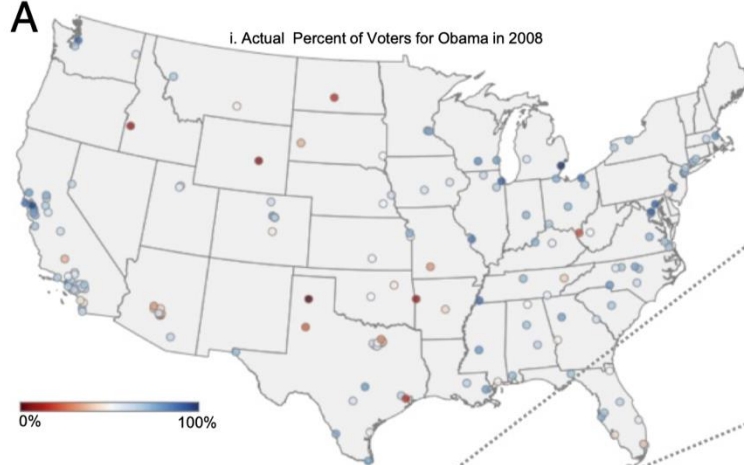
v. Graduate school (Milwaukee, Wisconsin)



vi. Income (Tampa, Florida)



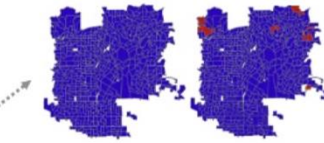
A



B

Republican
Democrat

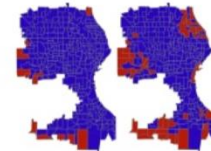
Los Angeles, California



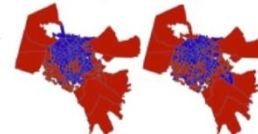
Casper, Wyoming



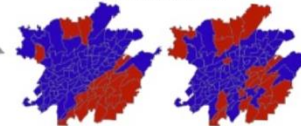
Milwaukee, Wisconsin



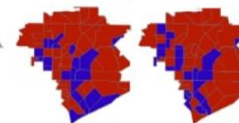
Lexington, Kentucky



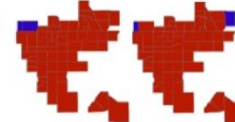
Birmingham, Alabama



Garland, Texas

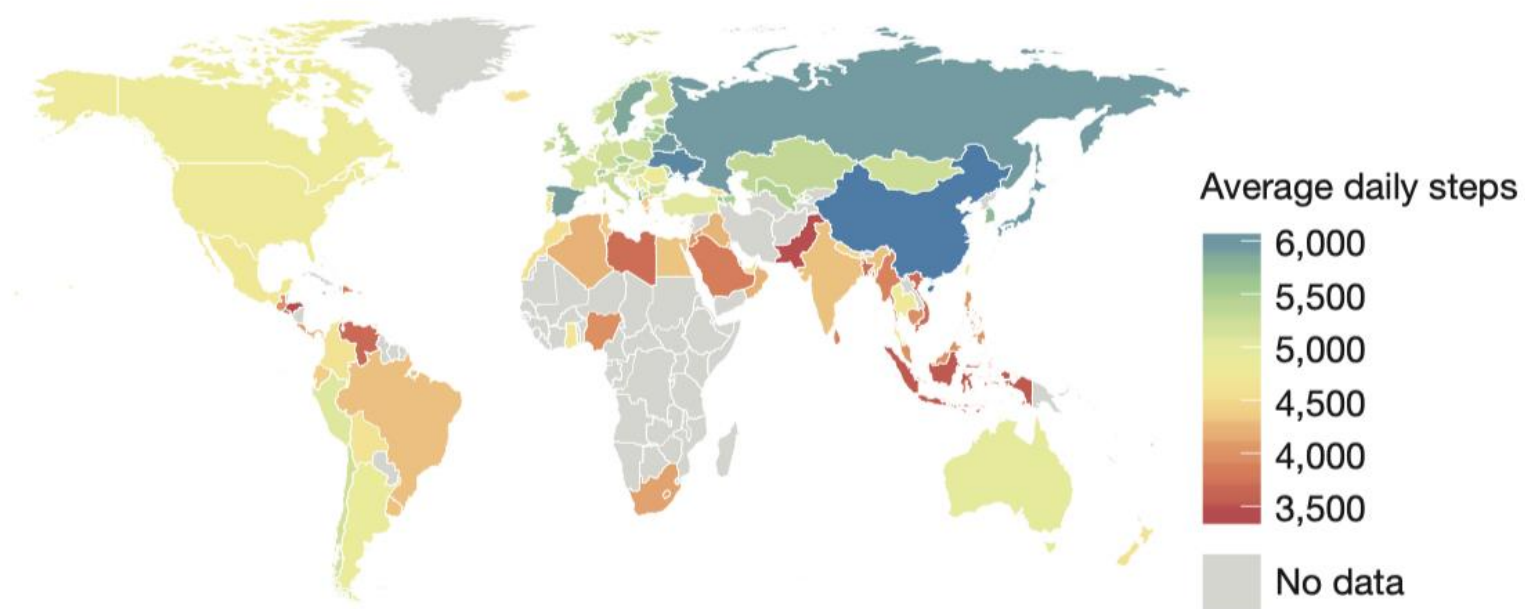


Gilbert, Arizona

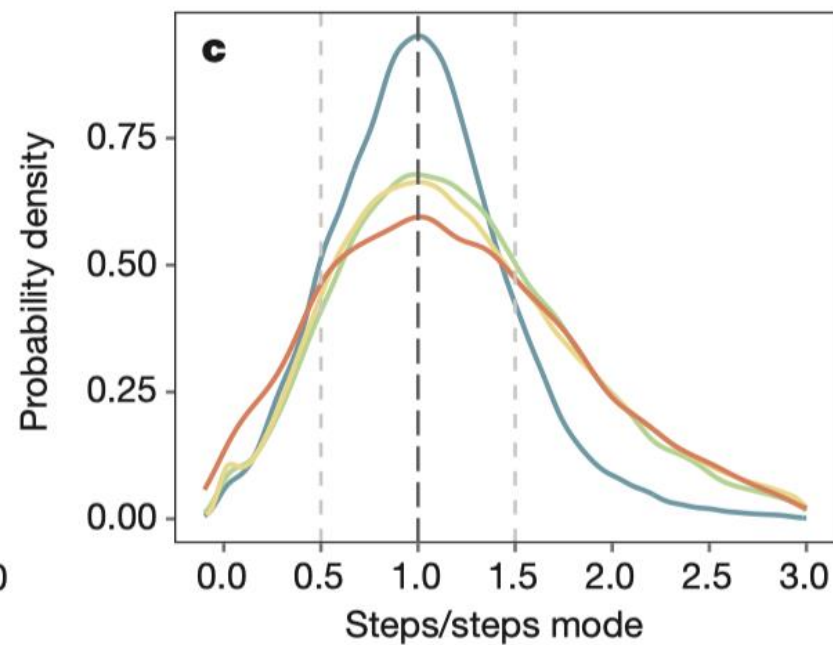
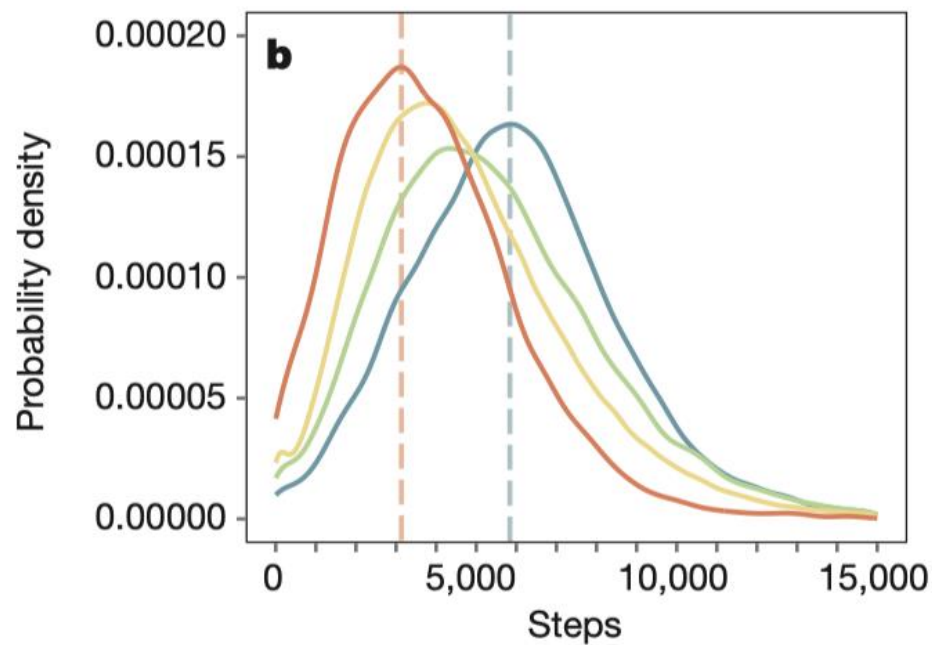


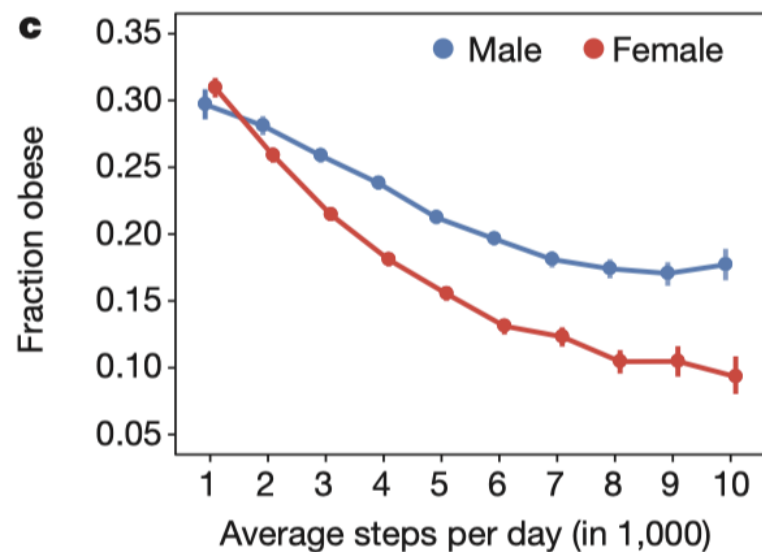
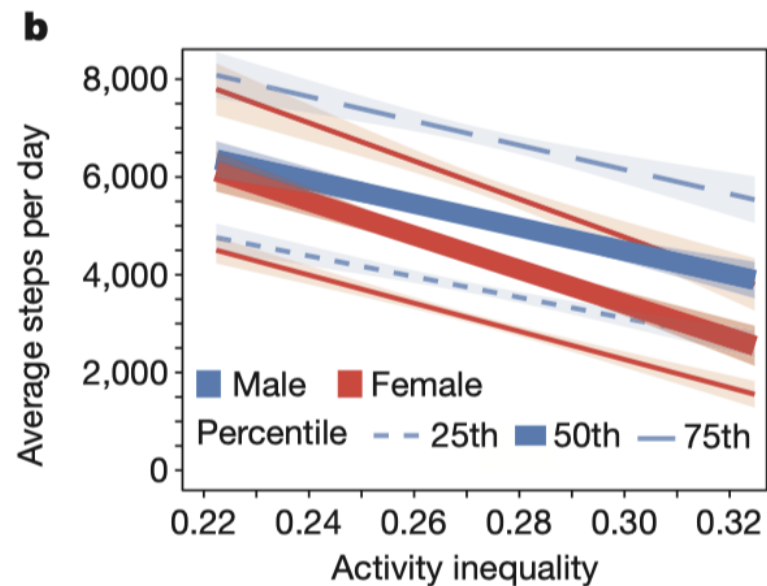
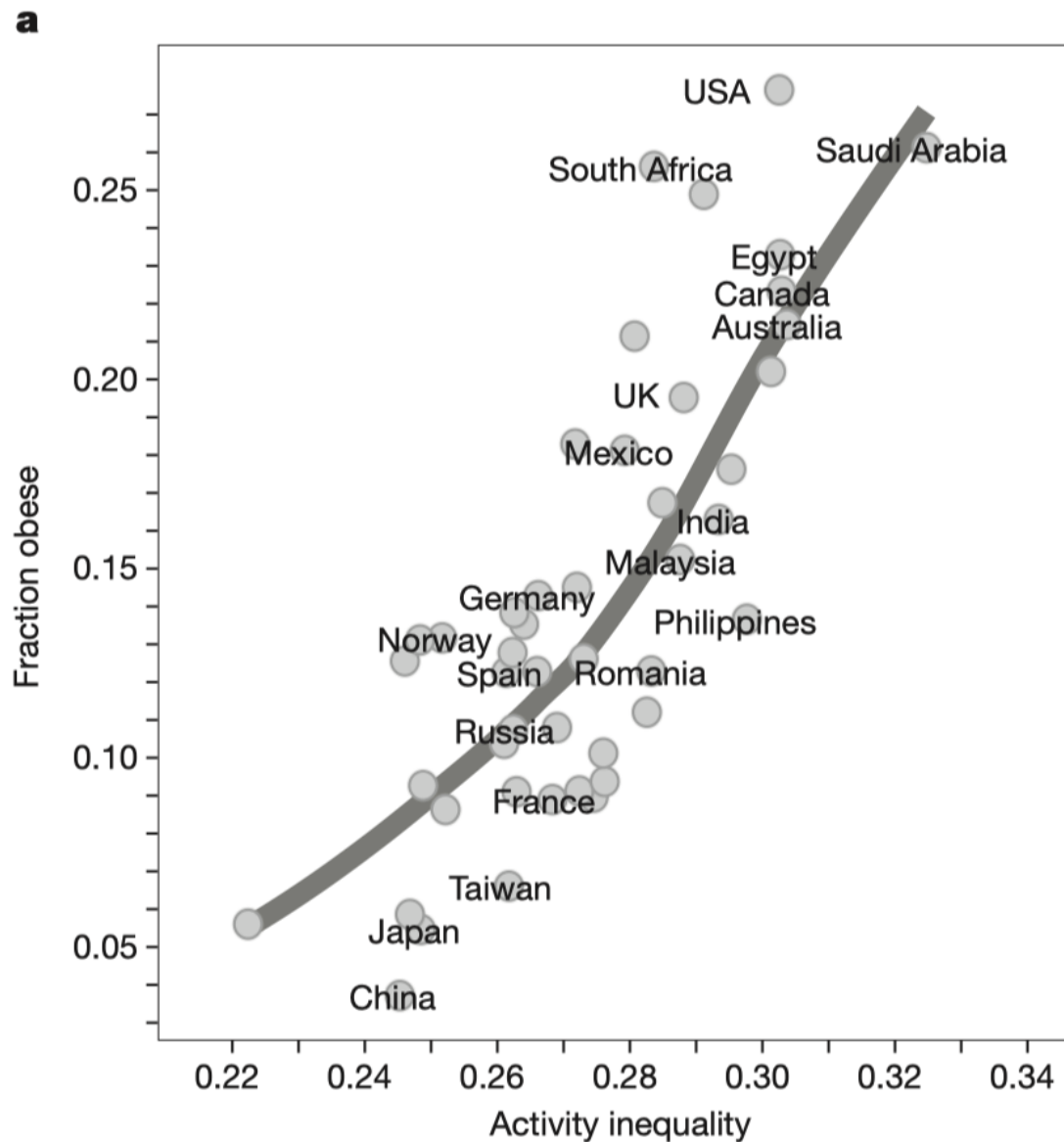
Althoff et al. (2017): Fitness Tracking

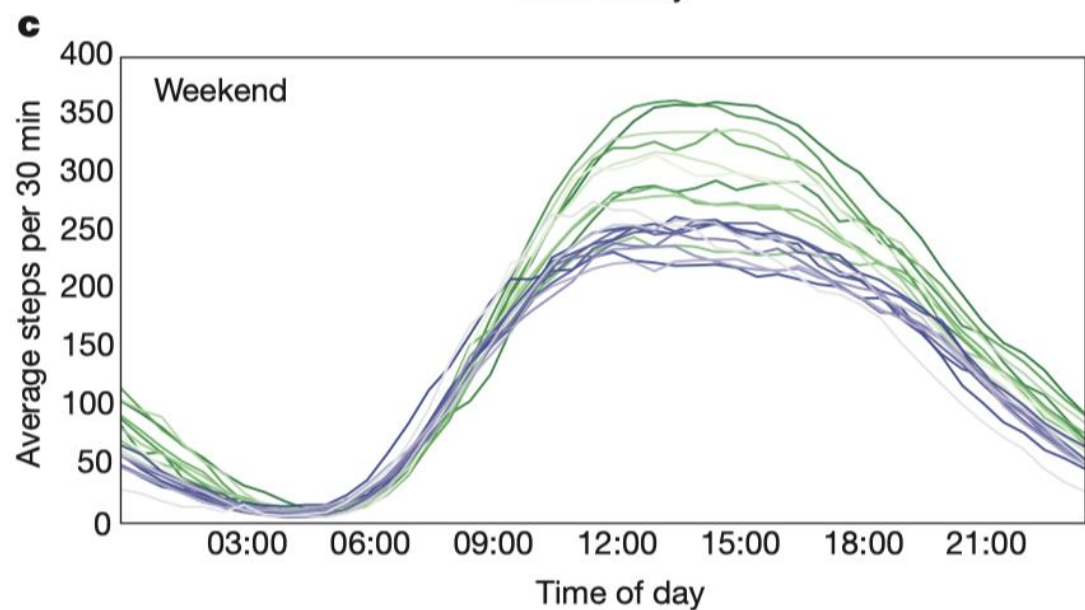
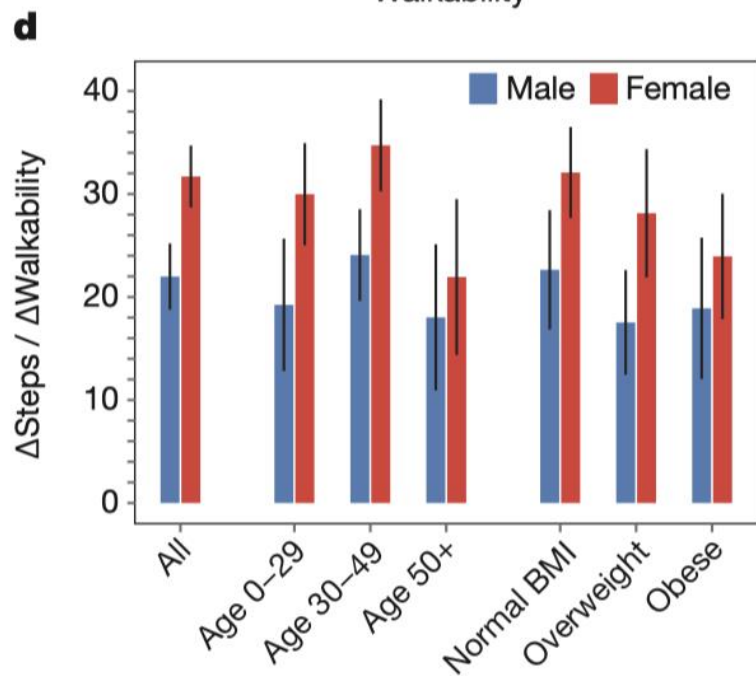
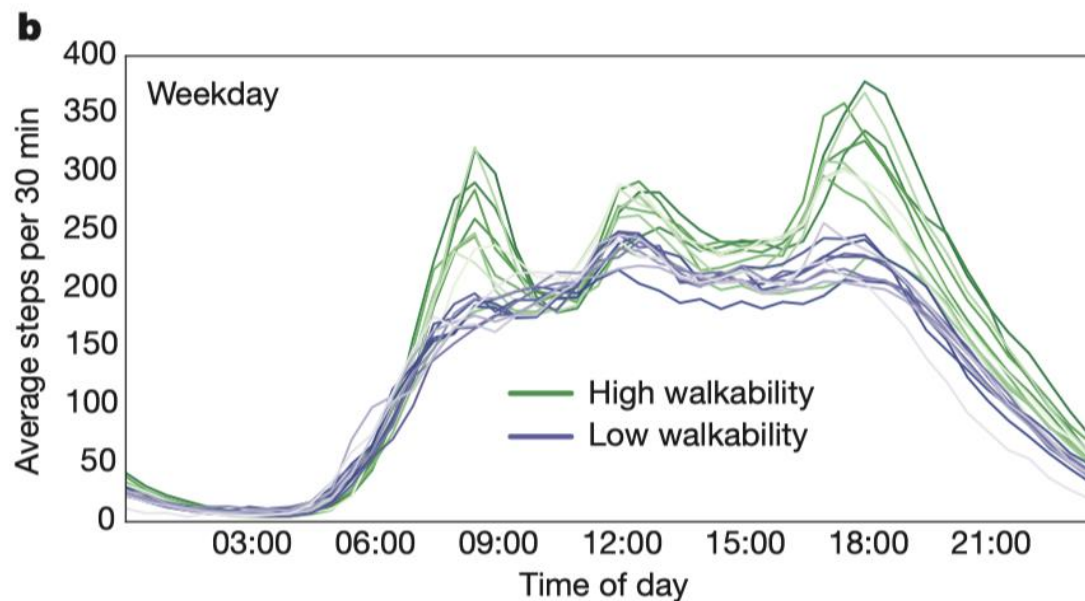
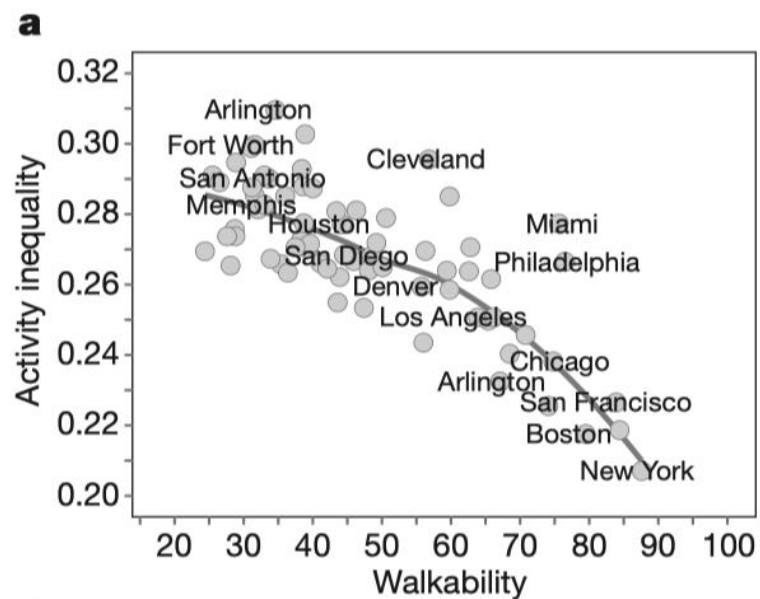
- Traditional social science relied on surveys, diaries, or lab studies to track health behavior.
- Mobile apps and wearable tech (e.g., smartphones, Fitbit) enable real-time, passive, global data collection.
- Collaboration with corporate platforms (e.g., Argus app) allows access to massive behavioral datasets.
- 68 million days of physical activity from 717,000 users across 111 countries.
- Activity inequality as a stronger predictor of obesity.

a

— Japan — UK — USA — Saudi Arabia







The Golden Age of APIs: Then and Now

- APIs once made massive platform data accessible to researchers
- Then: Open endpoints, easy rate limits, little restriction
 - Twitter:
 - Era 1:
 - Firehose: Full access to all public tweets in real time; Only available to a handful of partners or through expensive commercial contracts (e.g., Gnip, now part of Twitter)
 - Garden Hose: A sampled feed, ~10% of public tweets. Researchers request access
 - Spritzer: An even smaller sampled stream, around 1%—what most researchers actually got via the free Streaming API
 - Era 2 (launched in 2021, until Elon takeover in 2023):
 - Academic Research API v2, free, full-archive access, 10M tweets/month cap
- Now:
 - High barriers (e.g., paid access, deprecated tools)
 - Legal gray zones (e.g., scraping = violation of TOS)
 - Platform shutdowns (e.g., Pushshift, Facebook Research)

Twitter API Now

Tier	Monthly Cost	Access Details
Free	\$0	For write-only access You can post 500 posts per month
Basic	\$200	Up to 10,000 tweets read per month. 7-day search history
Pro	\$5,000	Up to 1 million tweets read per month. Search enabled
Enterprise	\$42,000+ (rumor)	Full-archive search access

<https://docs.x.com/x-api/getting-started/about-x-api>

The Golden Age of APIs: Then and Now

- APIs once made massive platform data accessible to researchers
- Then: Open endpoints, easy rate limits, little restriction
 - Reddit:
 - Pushshift: historical and real-time Reddit data access far beyond what Reddit's own API
 - Full archives of comments and posts since Reddit's founding
 - Queryable JSON file and downloadable dumps (still exists it seems like)
 - Reddit cut off Pushshift access in 2023
 - The "PullPush" effort tries to recreate it - we will try it in the lab
- Now:
 - High barriers (e.g., paid access, deprecated tools)
 - Legal gray zones (e.g., scraping = violation of TOS)
 - Platform shutdowns (e.g., Pushshift, Facebook Research)

The Golden Age of APIs: Then and Now

- APIs once made massive platform data accessible to researchers
- Then: Open endpoints, easy rate limits, little restriction
 - YouTube:
 - YouTube Data API v3
 - Originally open and easy to use.
 - Now requires an API key with quota limits.
 - Strict use policy and compliance with YouTube's Terms of Service.
 - Increased risk of API key revocation for TOS violations
- Now:
 - High barriers (e.g., paid access, deprecated tools)
 - Legal gray zones (e.g., scraping = violation of TOS)
 - Platform shutdowns (e.g., Pushshift, Facebook Research)

The Golden Age of APIs: Then and Now

- APIs once made massive platform data accessible to researchers
- Then: Open endpoints, easy rate limits, little restriction
 - Facebook:
 - Graph API (launched 2010)
 - Allowed programmatic access to posts, likes, friends, group memberships, etc
 - Huge for early CSS research (e.g., Facebook friendship networks)
 - Used (and abused) in scandals like Cambridge Analytica
 - After 2018, Facebook locked down its Graph API
 - CrowdTangle (with Instagram access) became the semi-official tool, focusing only on public pages and groups (recently deprecated and now blocked for new access)
- Now:
 - High barriers (e.g., paid access, deprecated tools)
 - Legal gray zones (e.g., scraping = violation of TOS)
 - Platform shutdowns (e.g., Pushshift, Facebook Research)

Freelon (2018): The post-API age

- We used to rely on APIs for social media data. Those days are (mostly) over. Now what?
- Heavy dependence on platform APIs is risky: companies can (and do) change access rules overnight.
- Teaching platform-specific tools is fragile and short-lived.
- We must train students to be adaptable, platform-agnostic, and ethically reflexive.

Freelon (2018): The post-API age

- We used to rely on APIs for social media data. Those days are (mostly) over. Now what?
- Learn Web Scraping
 - More flexible than APIs, works on most sites
 - But: harder to learn, fragile, may violate TOS
- Understand TOS and Legal Risks
 - Do not confuse TOS compliance with basic data privacy and ethics
 - Violating TOS may result in revoked access, lawsuits, or worse

Freelon (2018): The post-API age

- We used to rely on APIs for social media data. Those days are (mostly) over. Now what?

	APIs	Scraping
Access	Provided by platform	Extracted from public web pages
Limits	Rate-limited, selective data	Depends on HTML structure & your own limit
Reliability	Stable but platform-controlled	Fragile; breaks with site redesign
Legal/Ethical	Often within TOS	Often violates TOS or copyright
Use Case	Ideal for structured, historical data	Better for current or hard-to-access content

Freelon (2018): The post-API age

- We used to rely on APIs for social media data. Those days are (mostly) over. Now what?
- New skillsets researchers need:
 - Technical: scraping, automation, browser instrumentation
 - Legal: understanding TOS, DMCA, GDPR, platform policy
 - Ethical: balancing public benefit, user rights, and reproducibility
- Methodological future:
 - Less plug-and-play, more hand-crafted, bespoke pipelines
 - More emphasis on replication packages and ethical transparency

Freelon (2018): The post-API age

- We used to rely on APIs for social media data. Those days are (mostly) over. Now what?
- Should researchers ever violate TOS to obtain important social data?
- How do we balance public interest, user privacy, and corporate control?
- What ethical guidelines should we follow when APIs disappear?

Several Methodological Pivots

Platform Auditing: Studying the Black Box

- What is auditing?
 - Systematic evaluation of platforms' inner workings—what content is recommended, censored, or boosted? Who sees what and why?
- Why audit?
 - Because platforms don't give us the full picture. APIs are limited. Internal data is off-limits. Platform behavior is opaque by design.
- Audit studies often involve data collection methods outside official APIs—scraping, browser instrumentation, or custom logging pipelines are common technical foundations.

Presentation: Haroon et al. (2023)

Platform Auditing: Audit Design as a Genre

- Auditing is a genre, not a single method. Sock puppets are one tool, but many others exist.
- Examples of audit techniques:
 - Bots / Sock Puppet
 - User-side instrumentation (browser extensions, screen recording)
 - Crowdsourced data donation (especially prevalent in Europe)
 - Differential exposure experimental tests: Instruct users to do something first (search terms, video watches), then observe algorithmic outputs
- Design Challenges:
 - Keeping everything constant (e.g., geolocation, cookies)
 - Scaling up without violating terms of service
 - Minimizing bias from volunteer samples

The Hunt for Innovative Data

- Natural experiments from platform changes
 - Jaidka, K., Zhou, A., & Lelkes, Y. (2019). Brevity is the soul of Twitter: The constraint affordance and political discussion. *Journal of Communication*, 69(4), 345–372. <https://doi.org/10.1093/joc/jqz023>
 - Guo, Y., Li, Y., & Yang, T. (2023). Civilizing social media: The effect of geolocation on the incivility of news comments. *New Media & Society*, 14614448231218989. <https://doi.org/10.1177/14614448231218989>
- User-generated data (reviews, images, social features)
 - Yelp Reviews, LinkedIn Job Posting, Google Images, etc.
 - Yu, C., & Margolin, D. (2024). Sharing inequalities: Racial discrimination in review acquisition on Airbnb. *New Media & Society*, 26(3), 1627–1647. <https://doi.org/10.1177/14614448221075774>
- Government or nonprofit data
- Open datasets from unexpected places (e.g., real estate, transportation, Google Street View)

Presentation: Park et al. (2023)

User-Centric Behavioral Tracking

- What if users are your best data collectors?
- Shift focus from platform-provided data to *user-contributed* data:
“If platforms won’t give you the data, work with users who already have it”
 - Data Donation
 - Participants manually export and donate their platform activity data (e.g., Facebook “Download Your Data” or Google Takeout)
 - Screen Tracking
 - Software logs user interaction in real time (e.g., apps, screenomics)
 - User-Centric Behavioral Tracking
 - Usually involves recruiting users to install a new piece of software
- Issues/disadvantages?

Presentation: Robertson et al. (2023)

Lab Preview

- **Web Scraping**

- Basic
- Advanced
- Reddit API – try out
- YouTube API – try out
- TikTok API – Apply and hope for the best

- **Preparing for next week:**

- Sign up for Face++ API: <https://www.faceplusplus.com>
 - Just sign up for the free API; you should not be required to pay
 - Once signed up, go to <https://console.faceplusplus.com/dashboard>
 - Click “Apps” and then click “API Keys”
- Sign up for OpenAI (GPT) API: <https://openai.com/api/>
 - Go to Dashboard (top-right) and then click “API Keys”
 - The lab script may cost you \$1 to run, but
 - Jiacheng will share her API with the class

- **Day 7’s Presentation**

- Gretchen Corcoran
- Paulina Vergara
- Jinny Zhang