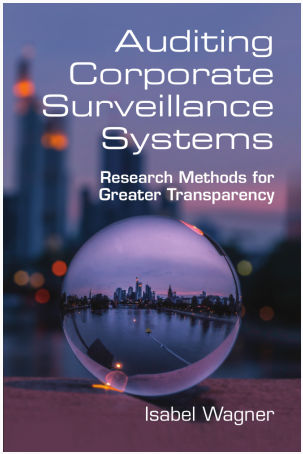
# DATA COLLECTION

Isabel Wagner

De Montfort University



# OVERVIEW

- Aim of data collection: execute the designed experiment and record response variables
- Ethical considerations: how to identify and minimize harmful impacts of data collection?
- Techniques for collecting data from different sources:
  - Archival sources
  - Passive data collection
  - Active data collection
  - Data collection from mobile apps
  - Data collection via crowdsourcing
- Methods to influence input variables
- Data storage

# Ethical considerations

# RESEARCH ACTIVITY INVOLVING HUMANS

- Institutional Review Board (IRB): university process to consider/approve research designs
- When is IRB approval needed?
  - Gathering information from or about individuals and organizations.
  - Using archived data in which individuals are identifiable
  - Research into criminal activities
- Guidelines for considering ethics in different situations
  - Codes of ethics, e.g., ACM Code of Ethics, IEEE Code of Ethics
  - Privacy by Design[1]
  - Ethical research guidelines[2]

[1] A. Cavoukian, "Privacy by Design," Information and Privacy Commissioner of Ontario, Tech. Rep., 2013.

[2] E. Kenneally and D. Dittrich, "The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research," U.S. Department of Homeland Security, Tech. Rep., Aug. 2012.

# ADNAUSEAM: ETHICAL JUSTIFICATION FOR A CONTROVERSIAL STUDY

- AdNauseam: browser extension hides ads, clicks on ads in the background[3]
- Three-part ethical and moral justification
- Is the aim laudable?
  - Ubiquitous surveillance violates tenets of liberal democracy, is against user's wishes
  - Privacy as societal value should be infused in systems
  - Protest against tracking by disrupting corporate surveillance business model (random clicking on ads pollutes user profiles, making them less valuable)
- Do alternatives exist?
  - Best alternative is regulation, which can seem unlikely to be realized
  - Cost induced by AdNauseam: server resources and bandwidth small compared to resources spent on ad serving
- Is it different from click fraud?
  - Fraud normally benefits the perpetrator (financially), not the case here

[3] D. C. Howe and H. Nissenbaum, "Engineering Privacy and Protest: A Case Study of AdNauseam," in *Proceedings of the 3rd International Workshop on Privacy Engineering*, vol. 1873, San Jose, CA, USA: IEEE, 2017, pp. 57–64.

# HARMFUL EFFECTS ON INDIVIDUALS

- Which research methods may cause harmful effects on individuals?
- Data collection from researcher-created app or browser extension
  - For example, to study cloud usage of mobile apps; price discrimination; mobile tracking; or ad pricing
- Studies with human participants
  - For example, studies that use surveys; crowdsourcing
- Data collection from websites/apps that display user data
  - For example, scraping from freelance marketplaces or résumé search engines
- Data collection through residential Internet connections
  - For example, crawling via Luminati to study vantage points in many countries
- Researcher-created synthetic data is visible to users
  - For example, researcher-run ad campaigns; measurement clients for Uber

# MINIMIZING HARMFUL EFFECTS ON INDIVIDUALS

- Informed consent
  - Inform participants about extent and purpose of data collection
  - Participants free to decide whether to participate
  - May not be possible if participants' awareness that they are participating biases the results
  - Additional duty to minimize harm; inform participants as soon as possible
- Data minimization
  - Only collect data that is necessary to achieve study objectives
  - For example, avoid collecting personally identifiable information
- Allow users to turn off data collection
- Process data on user device, only transmit summary data to researchers
- Limit sharing of datasets with user data
- Ensure no negative effects, e.g., no booking of tasks on freelance marketplaces, no requesting of rides on Uber

# HARMFUL EFFECTS ON ORGANIZATIONS

- User accounts
  - Limit account creation, minimize number of virtual personas
- Server resources and bandwidth
  - Rate limit requests to web servers
  - Respect robots.txt
- Responsible disclosure
  - Notify services before making vulnerabilities public, allow time to fix
- Ad impressions/clicks
  - Limit cost incurred by automated scraping of ads/clicking on ads
  - Estimate cost to advertisers, minimize

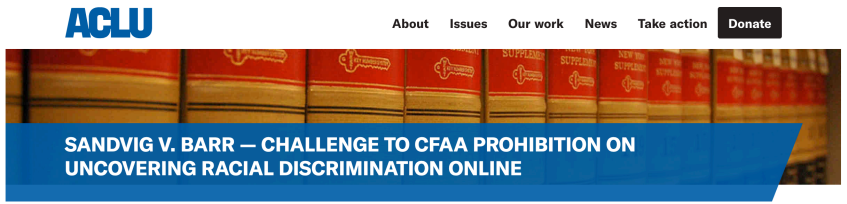
# KEEPING IT LEGAL

- Compliance with laws, regulations, policies
  - Different rules depending on location and research!
  - Computer Fraud and Abuse Act (CFAA)
  - Digital Millennium Copyright Act (DMCA)
  - General Data Protection Regulation (GDPR)
  - Terms of service[4] (may forbid automated queries, reverse engineering, scraping)
  - "non-commercial research for the public good that deals with issues of societal importance must be able to access public Web resources for research purposes as long as automated processes do not produce an unreasonable load.[5]"
- Compliance with ethical codes, e.g. ACM code of ethics
  - Need to comply with terms of service, "unless there is a compelling ethical justification to do otherwise"

[4] K. Vaccaro, K. Karahalios, C. Sandvig, *et al.*, "Agree or cancel? research and terms of service compliance," in *ACM CSCW Ethics Workshop: Ethics for Studying Sociotechnical Systems in a Big Data World*, Vancouver, Canada: ACM, 2015.

[5] G. Soeller, K. Karahalios, C. Sandvig, *et al.*, "MapWatch: Detecting and Monitoring International Border Personalization on Online Maps," in *Proceedings of the 25th International Conference on World Wide Web*, ser. WWW '16, Montréal, Québec, Canada: International World Wide Web Conferences Steering Committee, 2016, pp. 867–878. DOI: 10.1145/2872427.2883016.

# SANDVIG V. BARR[7]




- Lawsuit filed 2016
  - Does First Amendment grant researchers the right to provide false information to websites in the course of testing for discrimination on the basis of race, gender, and other characteristics protected under civil rights laws?
- 2020: federal court rules that research aimed at uncovering whether online algorithms result in racial, gender, or other discrimination does not violate the Computer Fraud and Abuse Act (CFAA) [6]

[6] *https://www.aclu.org/press-releases/federal-court-rules-big-data-discrimination-studies-do-not-violate-federal-anti*

[7] *https://www.aclu.org/cases/sandvig-v-barr-challenge-cfaa-prohibition-uncovering-racial-discrimination-online*

# Data sources

# ARCHIVAL DATA SOURCES: WEBSITES

- Internet archive's Wayback Machine
  - https://web.archive.org
  - Archive of website snapshots since 1996
  - APIs available to find available snapshots
  - Some issues: not all sites archived; some embedded resources not archived; resources may be archived at different times[8]
- OpenWPM: Monthly crawls of top 1million websites, 2015-2018[9]
- Open Observatory of Network Interference (OONI): Blocking and censorship measurements from many vantage points[10]
- CommonCrawl: regular snapshots of 3 billion websites since 2013, HTML pages without embedded resources[11]

[8] A. Lerner, A. K. Simpson, T. Kohno, *et al.*, "Internet Jones and the Raiders of the Lost Trackers: An Archaeological Study of Web Tracking from 1996 to 2016," in *25th USENIX Security Symposium (USENIX Security 16)*, Austin, TX, USA: USENIX, 2016.

[9] `https://webtransparency.cs.princeton.edu/webcensus/data-release/`

[10] `https://ooni.org/`

[11] `https://commoncrawl.org/`

# ARCHIVAL DATA SOURCES: ADVERTISING AND TRACKING ECOSYSTEM

- WhoTracks.me
  - Dataset of tracker domains and who owns them
  - Monthly data starting 2017[12]
- Ad block filter lists: EasyList[13]
- Tracker filter lists: EasyPrivacy
- Anti-ad block filter lists: Anti-Adblock Killer list[14], ad block warning removal list[15]

[12] *https://github.com/cliqz-oss/whotracks.me*
[13] *https://easylist.to/*
[14] *https://github.com/reek/anti-adblock-killer*
[15] *https://easylist-downloads.adblockplus.org/antiadblockfilters.txt*

# PASSIVE TRAFFIC CAPTURE

- Record real live traffic that was not influenced by researchers
  - netflows or full traffic at ISP
  - application-layer traffic (search, email) at search/email provider
- Advantages:
  - Service use by real users, no artificially induced effects
- Challenges:
  - Cannot vary experimental factor => analysis is more limited
  - Data can be difficult to obtain
  - Typically not possible to decrypt HTTPS traffic

# ACTIVE TRAFFIC CAPTURE

- Recording traffic that was initiated by researchers' measurement
- Desktop
  - Instrument browser to record cleartext HTTP, HTML, JavaScript, screenshots
- Mobile
  - Depending on platform, can be hard to instrument apps or OS
  - Can record outgoing network traffic, use VPN permissions/mitmproxy to decrypt HTTPS traffic
- Advantages:
  - Easy to influence input variables, define control/experimental groups
  - Record plaintext traffic at application layer
  - Easier control of noise factors
- Challenges:
  - Traffic may not resemble real user traffic (e.g., crawlers see more third-party requests than humans[16])

[16] D. Zeber, S. Bird, C. Oliveira, *et al.*, "The Representativeness of Automated Web Crawls as a Surrogate for Human Browsing," in *Proceedings of The Web Conference 2020*, ser. WWW '20, Taipei, Taiwan: ACM, Apr. 2020, pp. 167–178. DOI: *10.1145/3366423.3380104*.

# Techniques for data collection

# BROWSER AUTOMATION

- Most widely used: Selenium
  - Allows to script web browsing sessions
  - Advantage: full browser including JavaScript execution closely imitates real user experience
  - Many academic frameworks built on top, e.g. openWPM, AdFisher
- openWPM[17]
  - Flexible instrumentation, including browser extensions
  - Data storage
- AdFisher[18]
  - Instrumentation for advertising (ad settings, scraping of ads)
  - Permutation tests

[17] S. Englehardt and A. Narayanan, "Online Tracking: A 1-million-site Measurement and Analysis," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '16, Vienna, Austria: ACM, 2016, pp. 1388–1401. DOI: *10.1145/2976749.2978313*.

[18] A. Datta, M. C. Tschantz, and A. Datta, "Automated Experiments on Ad Privacy Settings," *Proceedings on Privacy Enhancing Technologies*, vol. 2015, no. 1, pp. 92–112, Apr. 2015. DOI: *10.1515/popets-2015-0007*.

- "Monkey" tools automate user interaction with apps
  - Android UI/Application Exerciser Monkey: creates random input events
  - DroidBot[19]: creates app-specific state-transition model to trigger sensitive app behaviors at higher rate
  - Both can run in emulator or on real device
- Disadvantages:
  - Additional work needed for apps that require login
  - Random events may not trigger all app behaviors

[19] H. Jin, M. Liu, K. Dodhia, *et al.*, "Why Are They Collecting My Data?: Inferring the Purposes of Network Traffic in Mobile Apps," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 4, 173:1–173:27, Dec. 2018. DOI: 10.1145/3287051.

- Browser extensions, apps
  - Idea: create browser extension/app that offers real benefit to users, hope that users consent to participating in research in exchange (e.g. by contributing some of their data)
  - Examples:
    - FDVT displays value users create for Facebook by viewing/clicking on ads[20]
    - $heriff allows users to compare prices for products from vantage points in other countries[21]
- Crowdsourcing platforms
  - Idea: pay crowd workers small amounts for short tasks
  - Amazon Mechanical Turk, Crowd Flower, Prolific Academic

[20] A. A. Galán, J. G. Cabañas, Á Cuevas, *et al.*, "Large-Scale Analysis of User Exposure to Online Advertising on Facebook," *IEEE Access*, vol. 7, pp. 11 959–11 971, 2019. DOI: 10.1109/ACCESS.2019.2892237.

[21] C. Iordanou, C. Soriente, M. Sirivianos, *et al.*, "Who is Fiddling with Prices?: Building and Deploying a Watchdog Service for E-commerce," in *Proceedings of the Conference of the ACM Special Interest Group on Data Communication*, ser. SIGCOMM '17, Los Angeles, CA, USA: ACM, 2017, pp. 376–389. DOI: 10.1145/3098822.3098850.

- Fair compensation for crowd workers
  - 96% of workers on Amazon Mechanical Turk earn less than US federal minimum wage of $7.25[22]
  - For fair compensation, ensure at least minimum wage for task + required unpaid work (task selection, screening tasks)
- Crowd worker tasks
  - Surveys, labeling images or other data, performing searches or similar to measure personalization based on real user profiles
- Quality control
  - Screening of workers: approval rating, location
  - Task replication: each task is performed by more than one crowd worker, majority vote to filter bad results
  - Controls: subtasks with known results are included in tasks, exclude results from workers who do controls incorrectly

[22] K. Hara, A. Adams, K. Milland, *et al.*, "A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Montréal, Québec, Canada: Association for Computing Machinery, Apr. 2018, pp. 1–14.

# Data storage

# DATA STORAGE

- Collected data may be large
- Goals for data storage:
  - Permanent record
  - Fast to access for data analysis
- US Library of Congress recommends platform-independent, character-based formats based on well-known schemas, such as CSV and SQLite
- Combination of storage systems can help to achieve performance and permanence

# Summary

# SUMMARY: METHODS FOR DATA COLLECTION

- Ethical considerations for data collection
- Techniques to minimize harm to individuals and organizations
- Data sources
  - Archival sources
  - Passive traffic capture
  - Active traffic capture
- Automation of browsers and apps
- Crowdsourcing
- Data storage

# ABOUT THIS SLIDE DECK

- These slides are designed to accompany a lecture based on the textbook “Auditing Corporate Surveillance Systems: Research Methods for Greater Transparency” by Isabel Wagner, published in 2022 by Cambridge University Press.
- 
- 
- Please leave this slide intact, but indicate modifications below.
  - Version 2022-04
  - Improved version for release on book website (Isabel Wagner)
- Updated versions of the original slide deck are available online: corporatesurveillance.org