Comprehensive TikTok Data Collection for Computational Social Science

Gayoung Jeon, Cameron Moy, Deen Freelon

University of Pennsylvania
Annenberg School for Communication
{gjeon, moycam, dfreelon}@upenn.edu

Abstract

This 4 hour hands-on tutorial provides researchers with practical tools and frameworks for TikTok data collection for Computational Social Science. Recent work systematically testing three TikTok data collection techniques (currently under review at the 2026 ICWSM Conference) reveals TikTok data collection method decisions dramatically alters research results. Participants in our tutorial will learn how to use web-scraping data collection methods (Pyktok and Apify) as well as the official TikTok Research API. This tutorial will explore best practices for data collection from three endpoints—*Users*, *Hashtags*, and *Comments*—using strategies identified through stress testing that: 1) Reduce algorithmic selection bias in data collection; 2) Substitute or fill missing data by combining multiple tools for a more complete dataset; and 3) Improve collection efficiency by balancing resources and dataset size (including strategies to minimize resource waste). Lastly, we introduce a checklist for reporting data collection procedures and results to increase the transparency, replicability, and generalizability of TikTok research. By engaging in this tutorial, researchers will be equipped with actionable methods to obtain high-quality TikTok datasets and decision-making criteria for optimizing collection parameters to answer empirical TikTok research questions.

Organizer Details and Qualifications

Gayoung Jeon is a doctoral student at the University of Pennsylvania Annenberg School for Communication. Her research combines computational psycholinguistics and artificial intelligence (AI) to examine how AI technologies influence cognitive processes and scientific research. She currently studies the factors driving LLM misalignment that result in the generation of anti-democratic content. Her work has been published in IJOC, JITP, and Visual Communication (VisCom).

Cameron Moy is a doctoral student at the University of Pennsylvania Annenberg School for Communication. His research interests include social media, marginalized communities, and data access. Currently, he employs web scraping techniques to monitor algorithmic changes on TikTok amid changing US ownership. His work has appeared at top venues including CHI, FAccT, and DIS.

Deen Freelon (PhD) is the Allan Randall Freelon Sr. Professor and a Presidential Professor at the Annenberg School for Communication, where he directs the Politics, Identities, and Communication Lab (PICL). A widely recognized expert on digital politics and computational social science, he has authored or coauthored over 60 book chapters, funded reports, and articles in journals such as Nature, Science, and the Proceedings of the National Academy of Sciences. He was one of the first communication researchers to apply computational methods to social media data and has developed eight open-source research software packages.

Freelon is the main author of Pyktok, an open-source Python module for collecting video, text, and metadata from TikTok. Pyktok is designed primarily to serve scientific research, enabling data collection from hashtags, user profiles, comments, and "You may like" videos (or so-called related videos). The tool has been ported to R as "traktok," demonstrating its broad impact across the computational social science community. He also developed ReCal, a free online intercoder reliability service, that has been running continuously since 2008 and used by tens of thousands of researchers worldwide.

He has been awarded over 6 million USD in research funding from grantmakers including the Knight Foundation, the Hewlett Foundation, the Spencer Foundation, and the US Institute for Peace. He was a founding member and remains Senior Researcher at the Center for Information, Technology, and Public Life at the University of North Carolina at Chapel Hill, one of five academic research centers in the Knight Research Network (est. 2019) to receive its highest level of funding. His research and commentary have been featured in press outlets including the Washington Post, NPR, The Atlantic, Buzzfeed, Vox, USA Today, the BBC, PBS NewsHour, CBS News, NBC News, and many others. Unlike many computational social scientists, he centers questions of identity and power in his work, paying particular attention to race, gender, and ideology.

Topic and Relevance

Introduction & Overview

The recent growth of TikTok, especially among younger generations, has sparked substantial scientific interest, making it a central platform for data-driven behavioral and social

science research. Yet, TikTok research remains in its preliminary stage. Little is known about the quality, validity, and reliability of data retrieved from the platform's newly released *official Research API Wrapper*, or how it compares to open-source alternatives such as *Pyktok* and commercial third-party platforms like *Apify*. These three tools vary in accessibility, cost, and technical requirements, presenting challenges for researchers looking to build comprehensive and/or randomly sampled datasets of general user behavior or platform phenomena.

This 4 hour tutorial, based on ongoing empirical stress-testing and comparative analyses by the authors (Jeon et al. 2026), introduces practical strategies to increase the rigor of empirical TikTok datasets. It addresses challenges related to algorithmic bias, including but not limited to popularity, recency, and geolocation biases. These biases have been noted in recent TikTok research, where scholars have noted discrepancies between API documentation and the data they receive(Pearson et al. 2025), underscoring the necessity of transparently reporting the limitations of TikTok data collection and methodologies (Corso, Pierri, and De Francisci Morales 2024). We aim to equip researchers across all levels of computational expertise—from beginners to advanced programmers—with practical tools and methodological guidance for collecting and preprocessing TikTok data using both the official Research API and web-scraping approaches (Pyktok and Apify).

Unresolved issues in TikTok Data Collection

The authors and colleagues conducted a 3-wave panel data collection, querying the TikTok API at three time points (10-day intervals) to systematically compare the coverage and overlap of three data collection techniques for hashtags, keywords, comments, user accounts, and related videos over time. Holding queries constant across collection tools, we revealed substantial variations in data recency, engagement metrics, and API hyperparameters (notably one that is supposed to generate random samples). In doing so, we corroborate concerns raised by recent studies (Corso, Pierri, and De Francisci Morales 2024; Jeon et al. 2026; Pearson et al. 2025), across the two dominant modes of TikTok data collection:

- **Web-Scraping (Front-End Data Collection)**— Captures data directly from the web interface, closely emulating the user browsing experience. While this method provides high-fidelity snapshots of what users actually see, it is highly influenced by TikTok's opaque algorithmic content curation, introducing biases in the visibility and selection of videos.
- **Official TikTok Research API**— Retrieves data from the platform's back-end servers through RESTful HTTP/HTTPS protocols with structured documentation. This method allows for direct access to TikTok's databases, but video selection processes are obscured behind TikTok's internal API decision scaffolding.

Findings from our related paper show that data collected through the API and via web-scraping are often distinct—sometimes exhibiting no overlap (Jaccard similarity ≈ 0)—even when retrieved simultaneously using identical queries and hyperparameters. These discrepancies limit *replicability* and *generalizability* of empirical TikTok research. Factors such as quota limits, undocumented data collection infrastructures, differing supported endpoints, and inconsistent data formats further hinder scientific reproducibility. We also identify undocumented behaviors of the three tools—such as unanticipated data loss, quota waste, and unclear limits on resource usage—that make it difficult to assess data completeness or collection efficiency. This lack of transparency and standardization poses major challenges to advancing reliable, large-scale TikTok data analysis.

Content overview & Detailed Schedule

In the 4 hour tutorial, we will provide in-depth examples of how researchers might scaffold data collection pipelines using each data collection tool–the official research API, Pyktok, and Apify (Table 1, sessions 2-4). Additionally, to address the challenges mentioned above, we introduce a data collection framework that combines web-scraping and the official API as complementary methods, using the strengths of each, to systematically reduce bias and improve the completeness of the data set (Table 1, session 5). The session will conclude with guidelines for methods and limitations reporting and a brief reflection on how the web and social media research community can and should collect TikTok (and other social media data) in the future (Table 1, sessions 6-7)

Session	Presenter	Duration
1. Introduction & Overview	Cameron Moy	30 mins
2. Official Research API Data Collection	Gayoung Jeon	30 mins
3. Pyktok Data Collection	Cameron Moy	30 mins
4. Apify Data Collection	Gayoung Jeon	30 mins
======= Break =======		15 mins
5. Combining Tools	Gayoung Jeon	45 mins
6. Methods & Limitations Reporting	Cameron Moy	30 mins
7. Closing Thoughts and Reflections	Cameron Moy	30 mins

Table 1: Tutorial Schedule

Transparency checklist for replicability

We argue that reflecting on the scope of datasets—and therefore the scope of valid claims—is critical. One approach is to systematically consider factors that influence the data collection process. For example, resource constraints such as hardware limitations or budget considerations may explain methodological choices (e.g., why Pyktok was selected over Apify). Additionally, researchers should clearly define constructs to reduce ambiguity. For instance, "influential posts" could refer to content with high engagement metrics (e.g., $>$1,000 comments or $>$50,000 likes) or content with extended visibility duration (e.g., remaining visible for $>$30 days)—these require fundamentally different data collection settings. Given this, explicitly reporting the thresholds defining conceptual boundaries and the specific metrics used to operationalize these constructs will enhance scientific rigor and facilitate cross-study communication. Additionally, not every study will necessarily make causal claims or require data that represents the broader platform

environment. If research interests address a specific niche rather than platform-wide patterns, this scope should be explicitly acknowledged. Transparency about these boundaries strengthens rather than weakens empirical TikTok research.

To facilitate such a practice within the web and social media research community, we propose a checklist for reporting data collection procedures for academic papers and beyond. The checklist will document tool selection, queries, quotas, data access points (e.g., back-end API vs. front-end user interface), collection dates, collection modes (automated via API or manual through user interface), sampling strategies and their rationale, collected data attributes, and validation checks to confirm that returned data is available on the platform (via web or mobile). We propose this checklist as a starting point, not a set of requirements, for thinking about methodological choices in web and social media research.

Relevance to ICWSM

Given TikTok's ubiquity, influence, and controversy, the platform has become a popular site of internet and social science research, with over 72,000 Google Scholar entries mentioning "TikTok" in 2024 alone. TikTok's popularity has garnered significant scholarly attention at ICWSM spanning topics of global political issues, misinformation and hate speech, and algorithmic bias (Ji et al. 2025; Luik, Setiawan, and Sitindjak 2025; Rejeb et al. 2024; Sharevski and Zeidieh 2025; Galdeman and Aiello 2025; Ungless, Markl, and Ross 2025; Yang et al. 2025; Pinto et al. 2024). Yet, as research interest in TikTok data initially accelerated in the absence of established data collection methods, we know less about the quality of collected data from the different available tools. In a full paper concurrently under review at ICWSM, we find that different TikTok data collection methods return drastically different data samples (Jeon et al. 2026). As TikTok research remains in a preliminary stage, transparent and reproducible data collection practices are especially critical for making research accessible and comparable across studies. Given that ICWSM is at the forefront of internet and TikTok research, our hands-on tutorial aims to give researchers the tools they need to conduct rigorous TikTok data collection, as well as the reflexive skills to critically examine the validity of their research methods, results, and limitations.

Tutorial Type

This will be a 4 hour hands-on tutorial, consisting of demonstrations, guided coding sessions, and supplementary documentation. Participants will perform live data collection using their own queries (e.g., hashtags, usernames, or events). All materials will be accessible on-site via a QR code linking to our GitHub repository. Participants should bring a laptop with VS Code (or a different code editor) pre-installed. Although not required for participation, to make the most of the workshop, we recommend participants 1) apply for TikTok's Official Research API (https://developers.tiktok.com/products/research-api/) ***at least one month in advance***, and 2) sign up for an Apify account (https://apify.com/), which offers $5 worth of free credits.

Audience Prerequisites

The tutorial welcomes participants of all backgrounds. Programming experience is not required for participation, but familiarity with Python is recommended. Attendees will gain hands-on experience with:

- Accessing TikTok data using both the Research API and web-scraping tools.
- Identifying overlapping data attributes and merging multi-source datasets.
- Recognizing and reducing TikTok sampling biases.
- Applying transparent documentation practices for replicability.

Tutorial Materials

Participants will receive Python scripts, documentation, and example datasets. All materials are open-access and free of copyright restrictions.

Previous Tutorials

This tutorial has not previously been presented.

Acknowledgments

We give special thanks to members of the Politics, Identities, and Communication Lab (PICL) for internal reviews and especially to Cristina Monzer and Silvia Téliz for their valuable feedback in developing technical recommendations to increase data collection transparency and methodological guidance.

References

Corso, F.; Pierri, F.; and De Francisci Morales, G. 2024. What We Can Learn from TikTok through Its Research API. In *Companion Proceedings of the 16th ACM Web Science Conference*, 110–114. ACM. ISBN 979-8-4007-0453-6.

Galdeman, A.; and Aiello, L. M. 2025. Mapping the Climate Change Landscape on TikTok. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, 2614–2621.

Jeon, G.; Moy, C.; Téliz, S.; Monzer, C.; Alayón, N.; and Freelon, D. 2026. WhichTok? Comparing Three TikTok Data Acquisition Tools. ICWSM'26 Proceedings. Association for the Advancement of Artificial Intelligence (AAAI).

Ji, J.; Xu, X.; Tam, V.; and Zhang, Y. 2025. Revealing public attitudes toward 'substituting plastic with bamboo' in China: Sentiment and topic analyses using social media data. *Forest Policy and Economics*, 176.

Luik, J.; Setiawan, D.; and Sitindjak, R. 2025. Media logic and educational micro-content: Presentational themes and approaches on TikTok. *Communication Review*, 28(2): 170–196.

Pearson, G. D.; , S., Nathan A.; , R., Jessica Y.; , A., Mona; , S., Barbara A.; ; and Kreslake, J. M. 2025. Beyond the margin of error: a systematic and replicable audit of the TikTok research API. *Information, Communication & Society*, 28(3): 452–470. Publisher: Routledge _eprint: https://doi.org/10.1080/1369118X.2024.2420032.

Pinto, G.; Burghardt, K.; Lerman, K.; and Ferrara, E. 2024. Fighting for Democracy: The Attempted Coup in Peru through the Lens of TikTok. In *Proceedings of the ICWSM Workshop on Data for the Wellbeing of Most Vulnerable*.

Rejeb, A.; Rejeb, K.; Appolloni, A.; Treiblmaier, H.; and Iranmanesh, M. 2024. Mapping the scholarly landscape of TikTok (Douyin): A bibliometric exploration of research topics and trends. *Digital Business*, 4(1): 100075.

Sharevski, F.; and Zeidieh, A. 2025. "I Don't Think TikTok Really Cares About the Truth": Experiences of Users Who Are Low Vision or Blind with Misinformation on TikTok. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, 1786–1797.

Ungless, E. L.; Markl, N.; and Ross, B. 2025. Experiences of Censorship on TikTok Across Marginalised Identities. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, 1952–1965.

Yang, C.; Mousavi, S.; Dash, A.; Gummadi, K. P.; and Weber, I. 2025. Studying Behavioral Addiction by Combining Surveys and Digital Traces: A Case Study of TikTok. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, 2106–2123.