

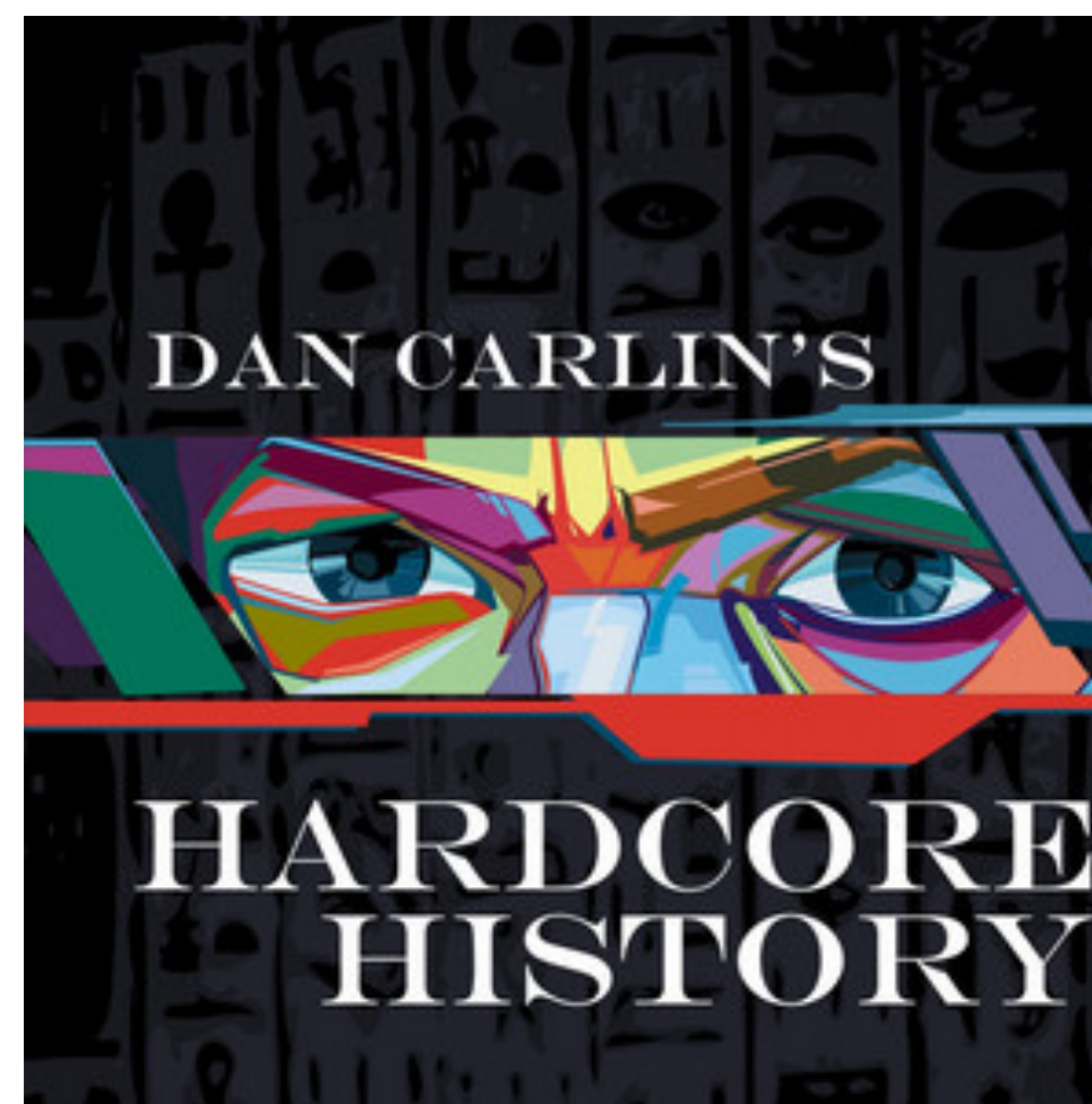
Podcasts as Social Data

End-to-End Pipelines for Large-Scale
Audio, Text, and Network Analysis

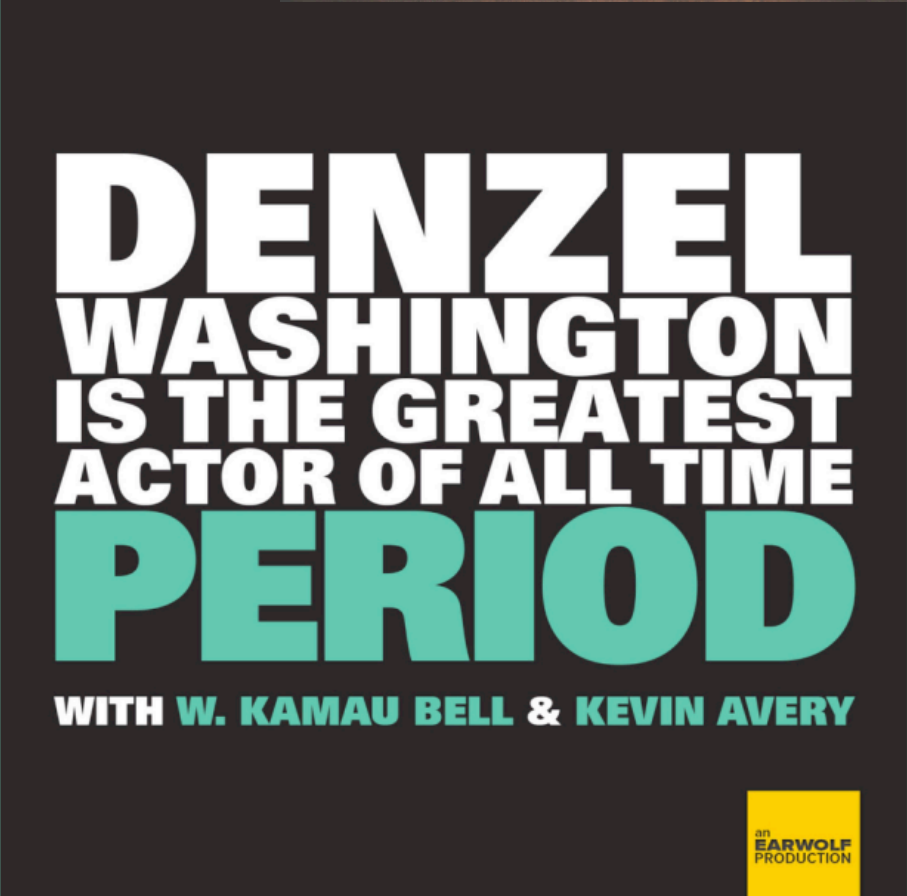
David Jurgens and Dallas Card
July 28, 2026



Who here uses podcasts?



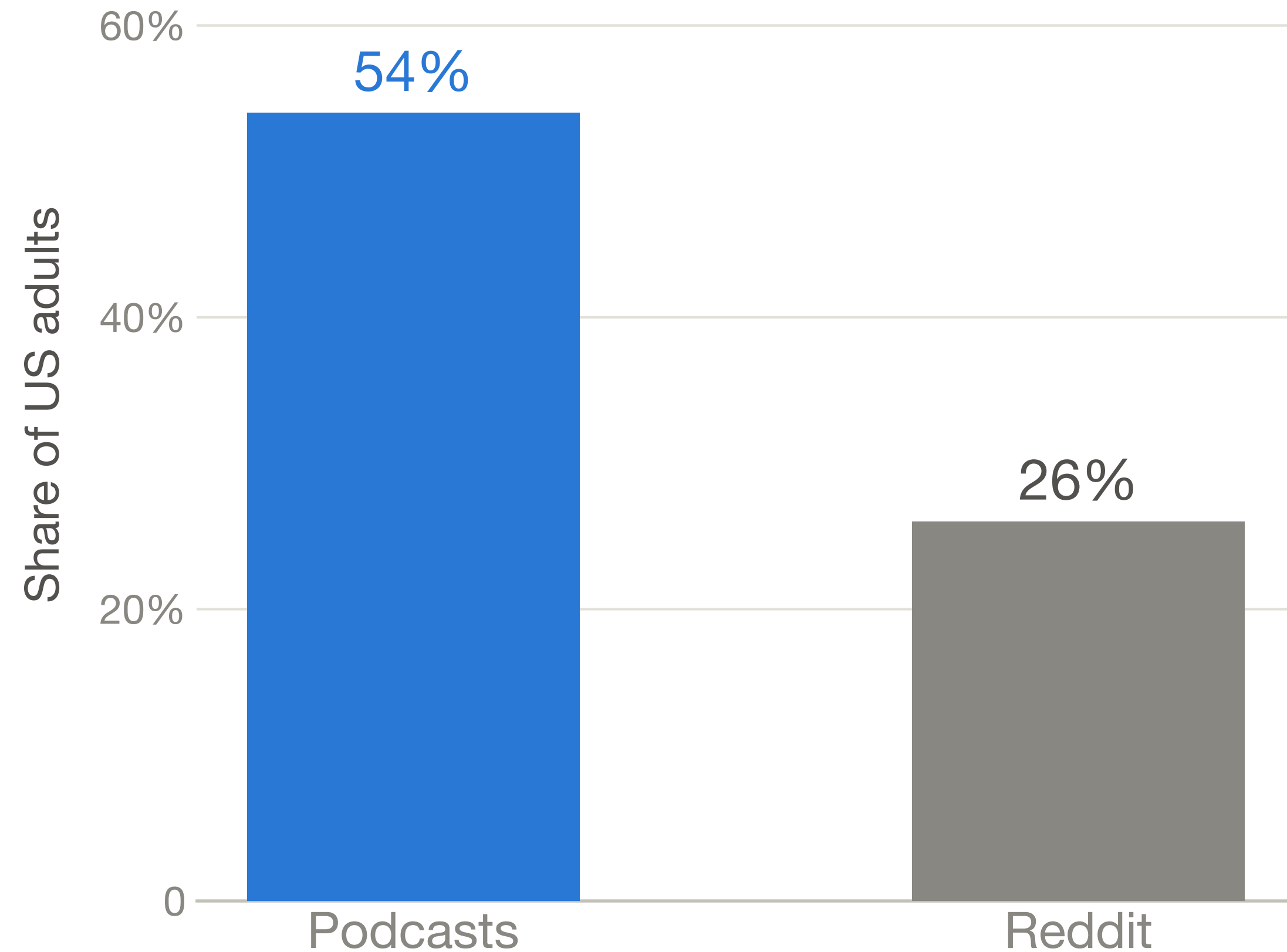
Number of
listeners



Podcast rank

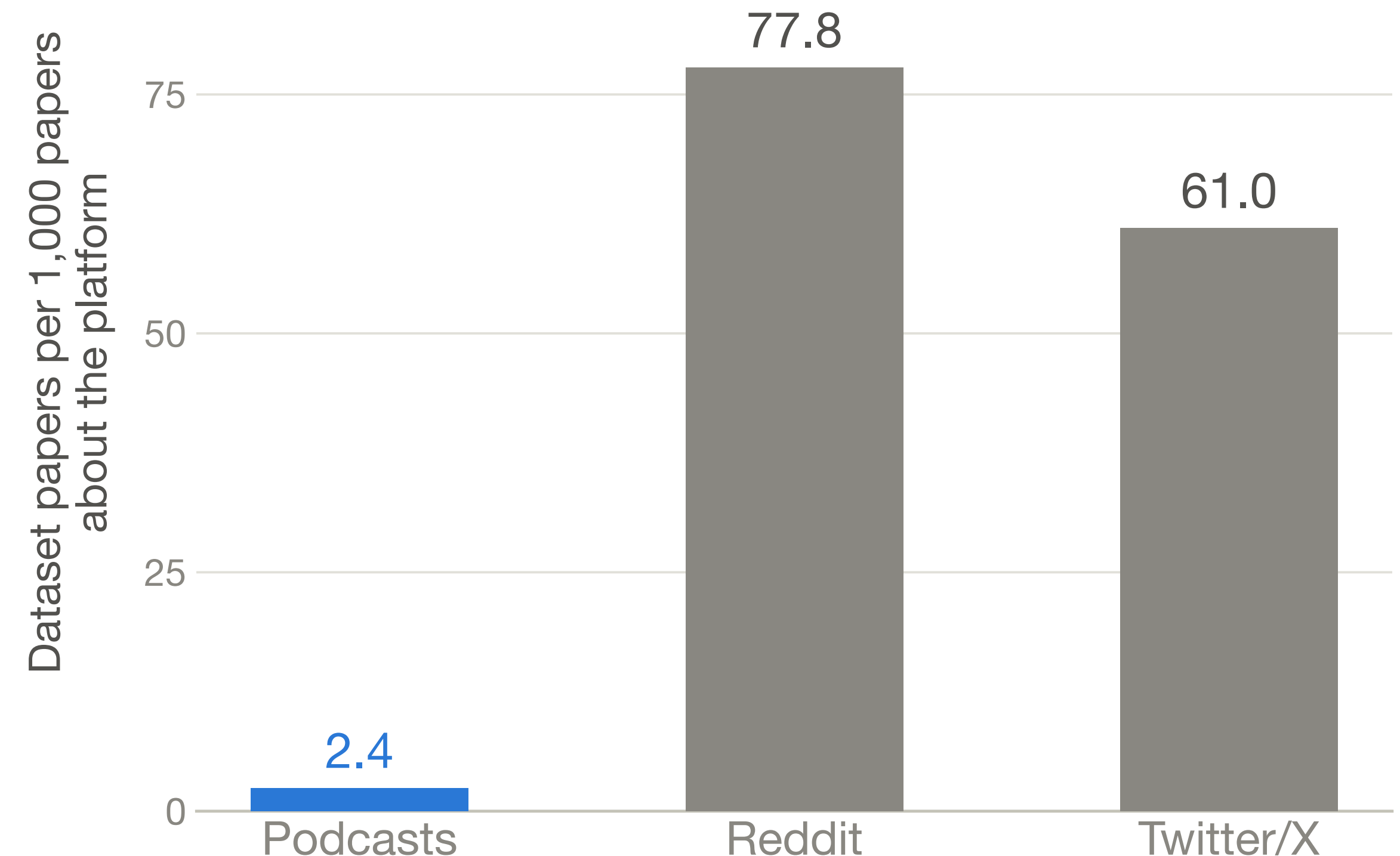
Most Americans Listen to Podcasts

- **54%** of US adults listened to a podcast in the past 12 months
- vs. **26%** who use Reddit



Relative Few Researchers *Study* Podcasts

- Research *about* podcasts is **~13–30× less likely to be built on a corpus** than research about Reddit or Twitter
- Podcasts are studied plenty as a *medium*, but rarely used as *data*



Why the gap in attention?

- **Podcasts are initially opaque:** ~hour long audio files with no speaker attribution.
- **Why study podcasts *now*?**
 - The bad...
 - Feb 2023: Twitter/X ends free API access; the Academic Research track is dismantled over 2023; enterprise tier priced at \$42,000/month
 - Aug 2024: Meta shuts down CrowdTangle
 - Pushshift goes offline
 - The good...
 - ASR got cheap and good; diarization is a solved-enough problem;
 - Released corpora now exist (more on that later)
 - Podcasts remain openly syndicated over RSS—crawlable, unmetered, no gatekeeper

Podcasts got *more* accessible while most social media else got less.

Murfeldt, Ryan, Sejin Paik, Naomi Alterman, Ihsan Kahveci, and Jevin D. West. "RIP Twitter API: A eulogy to its vast research contributions." arXiv:2404.07340, 10 April 2024. <https://arxiv.org/abs/2404.07340>

Blakey, Elizabeth. "The Day Data Transparency Died: How Twitter/X Cut Off Access for Social Research." *Contexts* 23(2), May 2024, pp. 30–35. DOI: [10.1177/15365042241252125](https://doi.org/10.1177/15365042241252125)

Gotfredsen, Sarah Grevy. "Q&A: What happened to academic research on Twitter?" Tow Center for Digital Journalism, *Columbia Journalism Review*, 6 December 2023 https://www.cjr.org/tow_center/qa-what-happened-to-academic-research-on-twitter.php

Meta Transparency Center. "CrowdTangle." "As of August 14, 2024, CrowdTangle is no longer available." <https://transparency.meta.com/researchtools/other-datasets/crowdtangle/>

Baumgartner, Jason, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. "The Pushshift Reddit Dataset." *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)* 14 (2020), pp. 830–839. arXiv:2001.08435 <https://arxiv.org/abs/2001.08435>

Today's tutorial aims to close this gap in attention!

Start	End	Session
09:00	09:20	Introduction to podcasts
09:20	09:35	Data sourcing and collection
09:35	09:55	Processing podcast data
09:55	10:10	Evaluating data quality
10:10	10:20	Annotating podcast data
10:20	10:30	Ethical and legal considerations
10:30	10:45	<i>Coffee break</i>
10:45	10:55	The sporc package
10:55	11:15	Working with SPoRC for research
11:15	11:40	Guided exercises
11:40	11:50	Results sharing
11:50	12:00	Closing and future directions

Our learning goals for this tutorial

- Familiarity with podcasts and podcast data
- Knowledge of the processing pipeline for podcasts
- Ability to access, process, annotate, and use podcast data
- Skill in evaluating podcast quality and knowledge of common errors
- Recognition of the ethical and legal considerations

tutorial website



This is a **hands-on** tutorial where you'll work with data in Part 2

A short history of Podcasts

Podcasting began as a file attachment

Nobody designed podcasting. `<enclosure>` is a file-attachment field, and podcasting is what happened when people put audio in it.

- **Dec 2000** — RSS 0.92 adds `<enclosure>`: a generic way to attach *any* file to a feed item. Dave Winer implements it after Adam Curry argues for distributing MP3s
- **11 Jan 2001** — Winer demonstrates it by attaching a Grateful Dead track to his weblog
- **12 Feb 2004** — Ben Hammersley, *The Guardian*, "Audible revolution," asks: "*But what to call it? Audioblogging? Podcasting? GuerillaMedia?*" — the word's first use
- **28 Jun 2005** — Apple ships iTunes 4.9 with a directory of 3,000+ free podcasts.
 - One million subscriptions in the first two days

RSS Advisory Board. "RSS 2.0 Specification." `<enclosure>` was introduced in RSS 0.92, December 2000, by Dave Winer. <https://www.rssboard.org/rss-specification>

Hammersley, Ben. "Audible revolution." *The Guardian*, 12 February 2004 <https://www.theguardian.com/media/2004/feb/12/broadcasting.digitalmedia>

Apple. "Apple Takes Podcasting Mainstream." Press release, 28 June 2005 <https://www.apple.com/newsroom/2005/06/28Apple-Takes-Podcasting-Mainstream/>

Apple. "iTunes Podcast Subscriptions Top One Million in First Two Days." <https://www.apple.com/newsroom/2005/06/30iTunes-Podcast-Subscriptions-Top-One-Million-in-First-Two-Days/>

Serial and the first mainstream moment

- Oct–Dec 2014: *Serial*, a *This American Life* spinoff, becomes the **fastest podcast to reach 5 million downloads and streams** in Apple's history; ~1.5M downloads per episode
- Establishes the narrative-documentary format and the idea that podcasts can carry investigative journalism
- Marks the shift from hobbyist audioblogging to a professionalized production industry

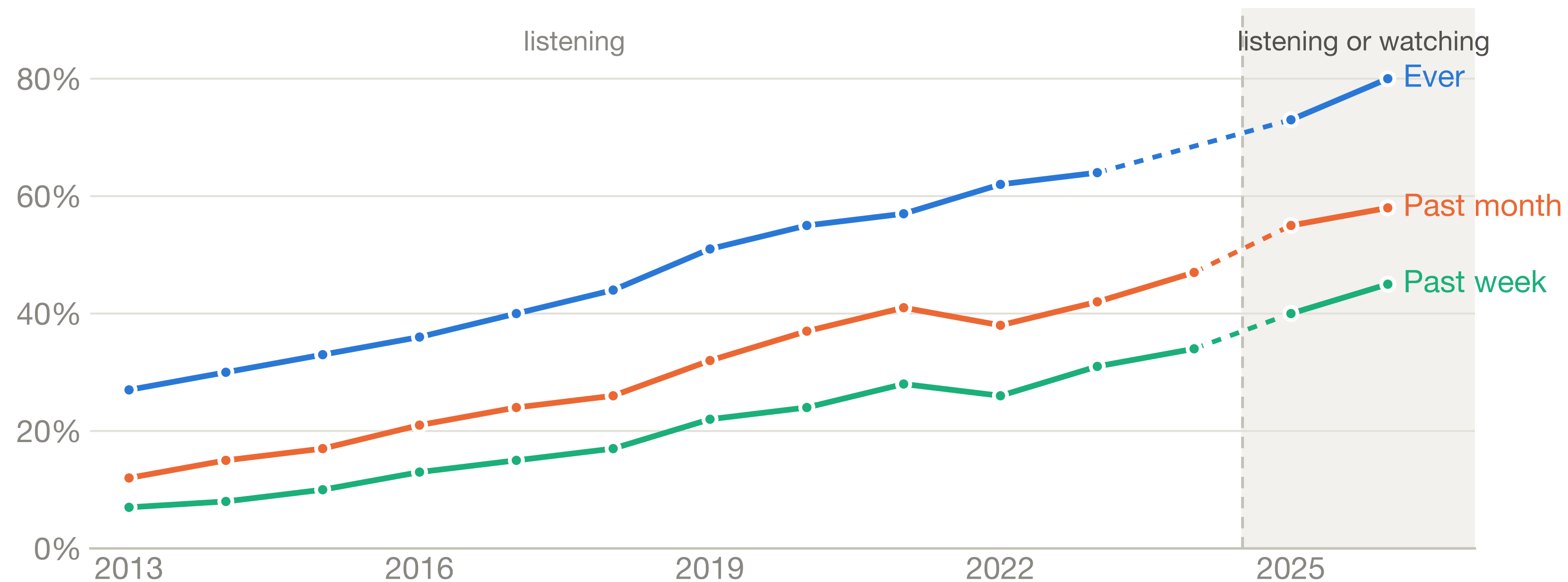
CNN. "The *Serial* podcast: By the numbers." 18 December 2014. <https://www.cnn.com/2014/12/18/showbiz/feat-serial-podcast-btn/index.html>

CNN Money. "The *Serial* effect." 18 December 2014. <https://money.cnn.com/2014/12/18/media/serial-podcast/>

The Irish Times. "November 2014." <https://www.irishtimes.com/business/technology/serial-is-the-fastest-downloaded-podcast-in-itunes-history-1.2040889>

The pandemic surge

- Monthly US listening: **12% (2013) → 42% (2023) → 58% (2026)**
- The ruler changed in 2025. Through 2024, Edison asked about *listening*. From 2025 it reports “consumption” which counts listening *or watching*
- Edison published both for 2025: 70% ever listened vs. 73% ever consumed.



Pew Research Center. "Audio and Podcasting Fact Sheet." 15 June 2023 <https://www.pewresearch.org/journalism/fact-sheet/audio-and-podcasting/>

Edison Research. *The Infinite Dial 2024*. <https://www.edisonresearch.com/the-infinite-dial-2024/>

Edison Research. *The Infinite Dial 2025*, <https://podnews.net/press-release/edison-research-infinite-dial-2025>

Edison Research. *The Infinite Dial 2026*, <https://www.edisonresearch.com/the-infinite-dial-2026/> · <https://podnews.net/press-release/infinite-dial-us-2026>

Consolidation: 2019–2021

- **2019** — Spotify acquires Gimlet and Anchor (€300M combined, Feb) and Parcast (€49M, Apr)
- **2020** — The Ringer; Megaphone (\$235M, Nov 2020); Joe Rogan exclusive (reported ~\$200M), exclusive on Spotify from Dec 2020
- **2022** — Podsights + Chartable (€83M, measurement), Findaway (€117M, audiobooks), Sonantic (€91M, synthetic voice)
- **2024** — the Rogan renewal (reported up to \$250M) drops exclusivity; the show returns to other platforms

Spotify Technology S.A. Form 6-K, FY2019 (US Securities and Exchange Commission), April 2019. https://www.sec.gov/Archives/edgar/data/1639920/000156459019027494/ck0001639920-6k_20190630.htm

Graham Holdings Company. Form 8-K, Exhibit 99.1, 10 November 2020. <https://www.sec.gov/Archives/edgar/data/104889/000010488920000068/a2020-11x10form8xkex991.htm>

Spotify SEC filing following the 27 July 2022 earnings report, as reported by TechCrunch, 28 July 2022 <https://techcrunch.com/2022/07/28/spotify-forks-out-295m-for-findaway-podsights-chartable-and-sonantic-filing-reveals/>

Spitz, Todd, and Cynthia Littleton. "Joe Rogan's Spotify Deal Renewal Worth Up to \$250 Million, Podcast Will No Longer Be Exclusive to the Platform." *Variety*, 2 February 2024. <https://variety.com/2024/digital/news/joe-rogan-renews-spotify-deal-not-exclusive-1235895424/>

The correction: 2022–2023

- **Oct 2022** — Spotify cancels 11 original shows, cuts <5% of staff
- **Jan 2023** — Spotify lays off ~600 company-wide
- **Mar 2023** — NPR lays off ~10% of staff and ends four podcasts (*Invisibilia*, *Louder Than A Riot*, *Rough Translation*, *Everyone & Their Mom*) against a projected **\$30M** ad shortfall
- **Jun 2023** — Spotify cuts ~200 podcast roles; Gimlet and Parcast merged into one unit
- Also hit: Amazon Music, New York Public Radio, The Athletic, Sony Music, Pushkin

Spotify, October 2022: TechCrunch, 6 October 2022. <https://techcrunch.com/2022/10/06/spotify-cancels-11-original-podcasts-lays-off-under-5-of-staff/>

NPR, March 2023: ~10% of staff laid off and four podcasts ended — *Invisibilia*, *Louder Than A Riot*, *Rough Translation*, *Everyone & Their Mom* Podnews. <https://podnews.net/update/npr-layoffs-23>

Sherman, Justin. "The Podcast Bubble Burst." *Slate*, 12 December 2023 <https://slate.com/business/2023/12/podcasts-layoffs-spotify-heavyweight-stolen-amazon.html>

Podcasting is More Than Audio

- Platform used most often by US weekly podcast consumers in 2026:
 - YouTube 32%
 - Spotify 25%
 - Apple Podcasts 15%
- 57% of Americans 12+ have both *listened to* and *watched* a podcast
- Edison's redefined what a “listen” means in light of video podcasts
- **The collection consequence:** a YouTube-first show *may* have no useful RSS feed at all.

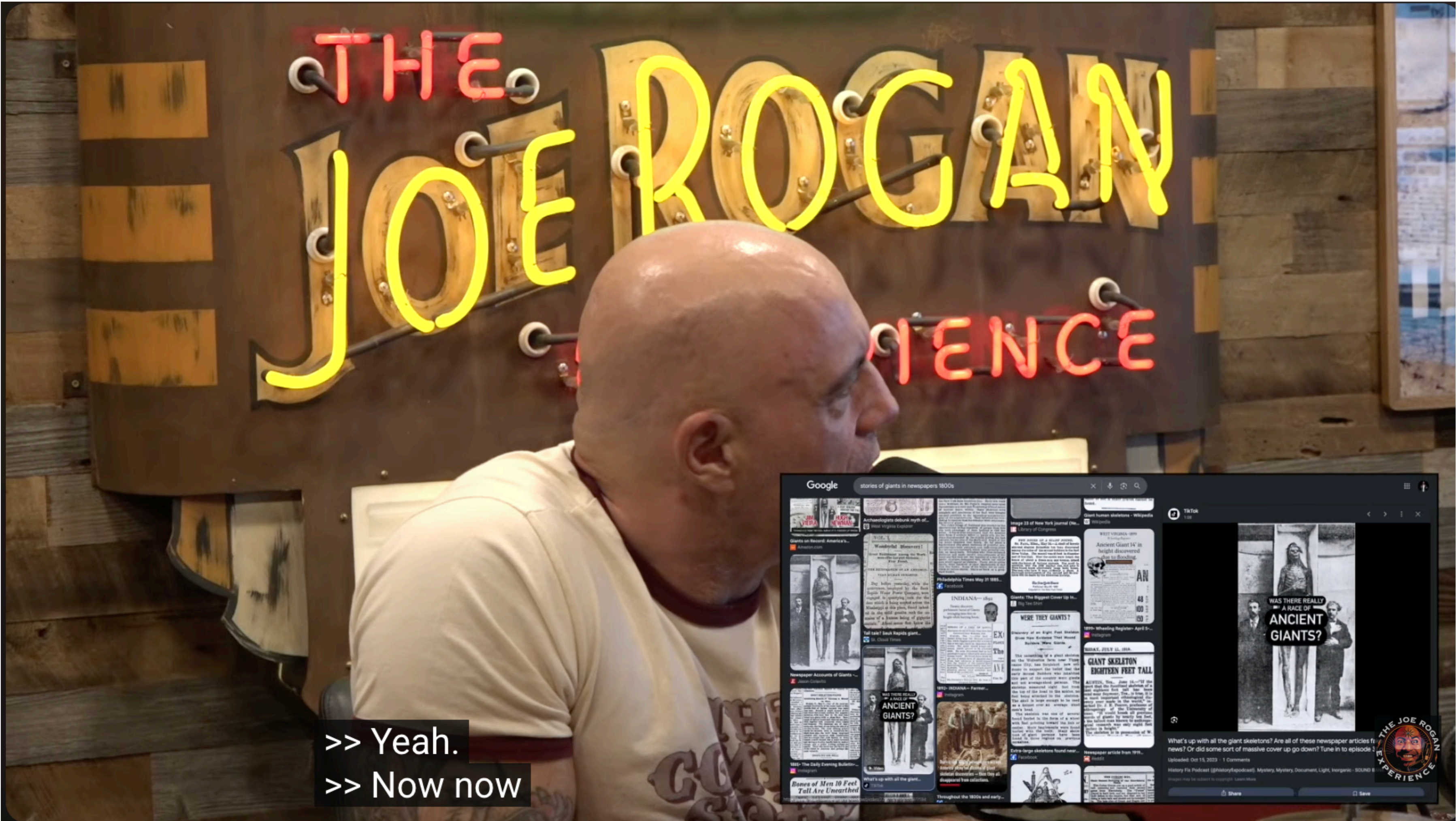
The most-used podcast platform in America is not a podcast platform.

Edison Research. *The Infinite Dial 2026*, released 13 March 2026. *n* = 2,050, US age 12+, fielded January 2026. <https://www.edisonresearch.com/the-infinite-dial-2026/>

Reuters Institute for the Study of Journalism. *Digital News Report 2026*, 16 June 2026; 48 markets, ≈97,520 interviews. <https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2026>

Example: The Joe Rogan Experience

Episode #2530 - Timothy Alberino (July 23, 2026)



>> Yeah.
>> Now now

A Podcast is an Edited Artifact

- Podcast content may include cuts, retakes, or re-recorded lines
 - Podcast files can even be edited post release
- Plenty of non-conversation segments: music, scripted segments, inserted ads
- Podcasts vary widely in format and editing
 - A narrative-documentary show may contain almost no spontaneous speech
- Pauses, overlaps, and turn boundaries are partly an editor's decisions

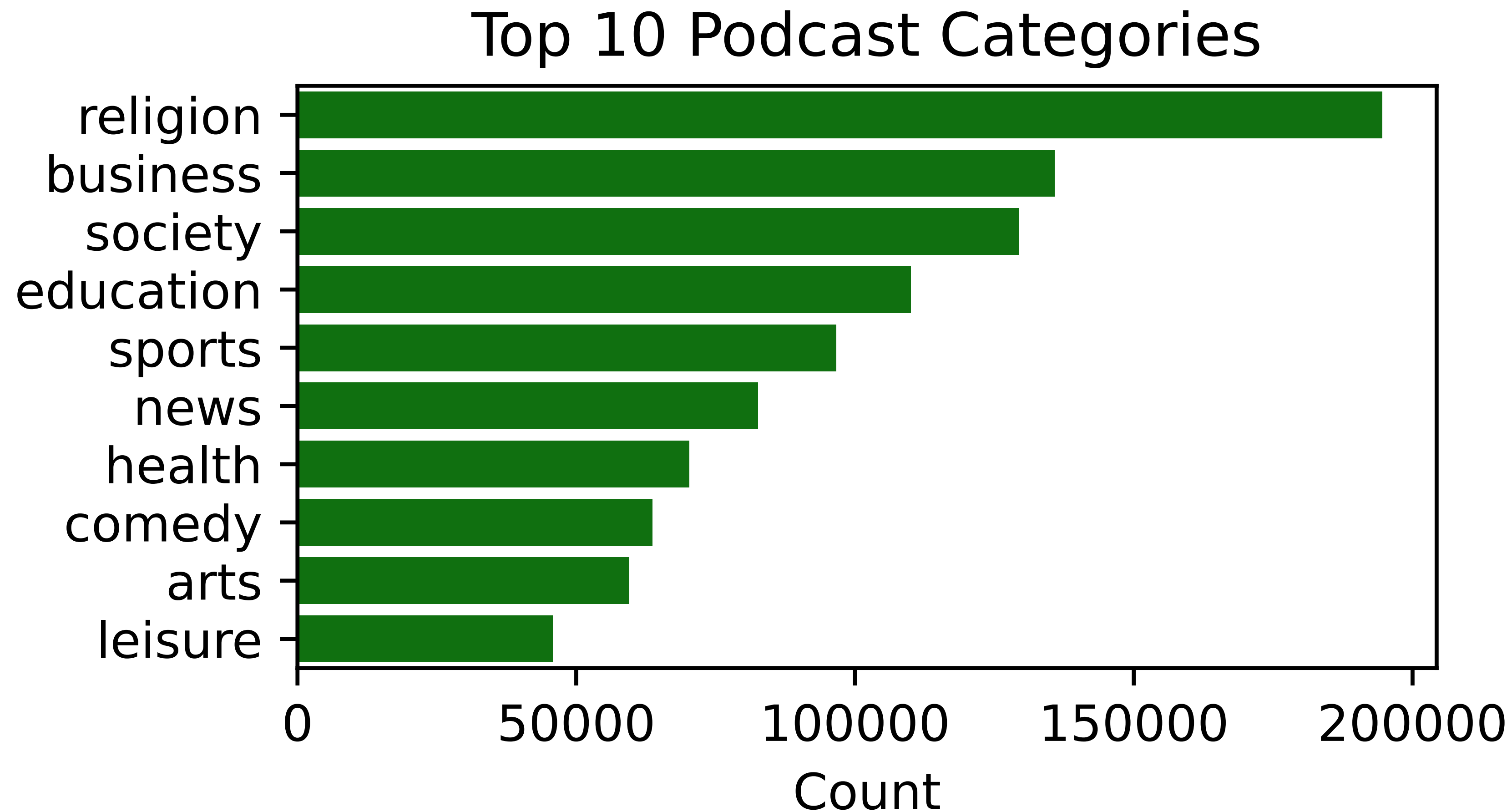
If you measure interruption, pause length, or turn-taking on a produced show, you are partly measuring post-production.

The Many Formats of Podcasts

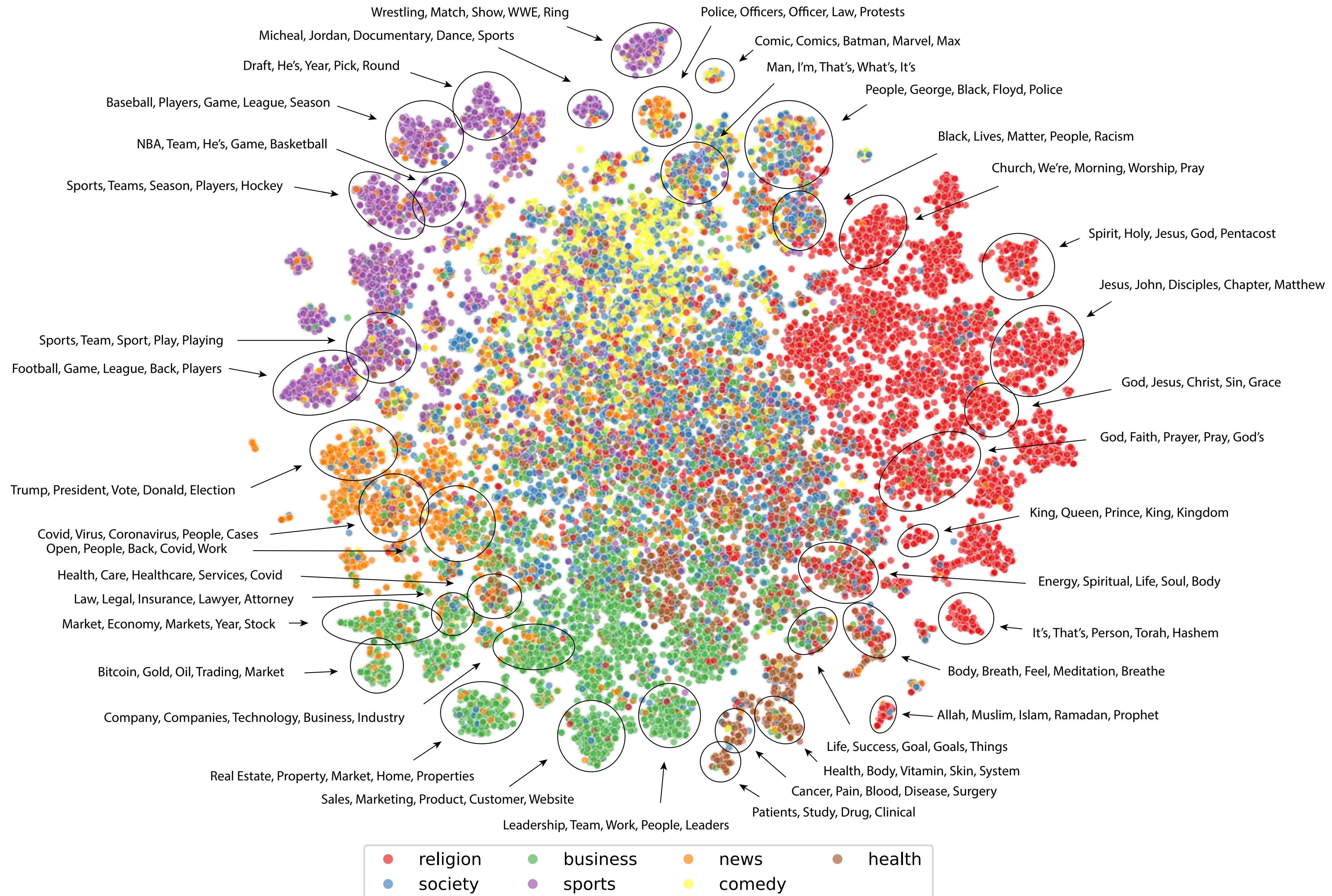
Format	Example shape	Dialog Consequences
Interview	host + guest, 1:1	Cleanest turn-taking
Co-host chat	2–4 regulars	Rich interaction; heavy overlap
Monologue	one speaker	No turn-taking to measure at all
Panel / roundtable	4+ speakers	Diarization degrades fast
Narrative documentary	scripted + tape	Speech is written; prosody is performed
Fiction / drama	actors	Not natural speech in any sense

The Many Categories of Podcasts

From a sample in 2020



Even within categories, podcasts are topically diverse



Podcasts for News—US and Internationally

- **32% of US adults** get news from podcasts often/sometimes (2020: 22%)
 - Under-50s: 39% · 50+: 24%
- **Globally, weekly news podcast reach is ~11%** — US 14%; Australia and Peru rising fastest
 - Nordic charts are dominated by public broadcasters; US charts by personality-led shows

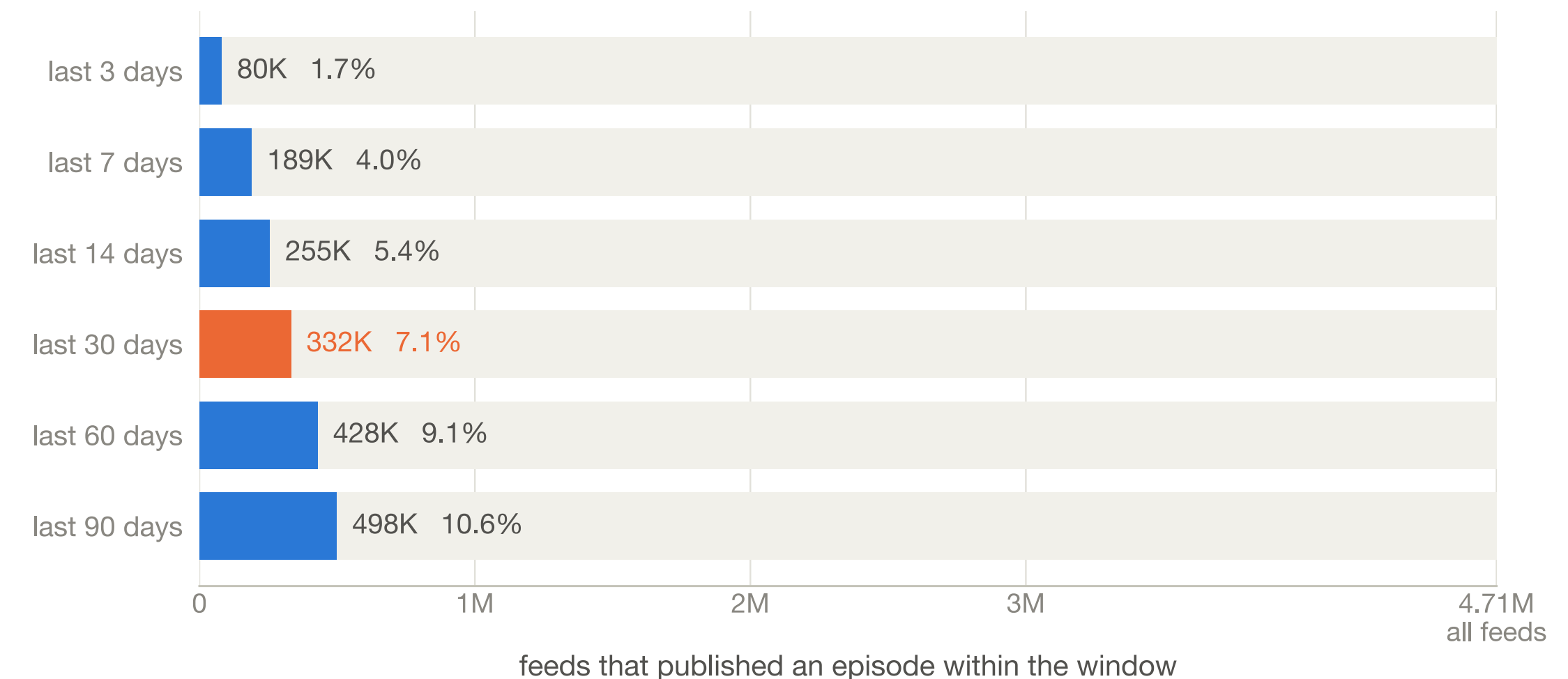
Pew Research Center. "Podcasts and News Fact Sheet." 25 September 2025 <https://www.pewresearch.org/journalism/fact-sheet/podcasts-and-news-fact-sheet/>

Reuters Institute for the Study of Journalism. *Digital News Report 2026*, 16 June 2026; <https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2026>

Newman, Nic. *Digital News Report 2025*, executive summary. Reuters Institute, 17 June 2025. <https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2025/dnr-executive-summary>

Concentration: Most Podcasts Have Almost No Audience

- While most Americans listen to podcasts, and most podcasts have no listeners
- Active production is concentrated: only 332K of 4.71M feeds published in the last 30 days (7.1%)
- "How many podcasts are there" has no single answer
 - The activity threshold *is* the sampling frame
- Listening concentration is more concentrated, and we (researchers) cannot easily measure it.
 - Per-show audience data is proprietary (Edison Podcast Metrics, subscription-only).
 - There is no large public download distribution for podcasting
 - but there is the Open Podcast Prefix Project (OP3)!



Podcast Index. Daily counts, retrieved 26 July 2026 https://stats.podcastindex.org/daily_counts.json

Edison Research. *Edison Podcast Metrics* <https://www.edisonresearch.com/edison-podcast-metrics/>

Open Podcast Prefix Project (OP3) — <https://op3.dev/>

Current Research on Podcasts

Major Themes in Current Podcast Research

- Content & discourse — what is said, framed, claimed
- Speaker behavior — how people talk, interact, take turns
- Ecosystem structure — who appears with whom; how shows connect
- Audience & reception — who listens and what it does to them
- Many different methods: quantitative, qualitative, mixed, design/position

Very few studies using “large” data to draw conclusions.

Significant Differences in Study Scale

BBC World Service hand-audited 15 episodes of one show (of 23 screened, with 4 medical experts), finding an average of 14 harmful health claims per episode

→ **international news coverage**

Moldovan *et al.* 2025 — classified 139,200 episodes for collective-action statements

→ one paper

Moldovan, Theodora, Arianna Pera, Davide Vega, and Luca Maria Aiello. "Podcasts as a Medium for Participation in Collective Action: A Case Study of Black Lives Matter." arXiv:2509.13197, 16 September 2025. <https://arxiv.org/abs/2509.13197>

BBC World Service / BBC Global Misinformation Unit. <https://medicalxpress.com/news/2025-01-influential-podcasts-fuel-health-misinformation.html>

Recent work: Content and Behavior

Study	Question	Data	Method
Reddy et al. 2021 (ACL)	Does linguistic style predict listener engagement?	5,371 episodes (Spotify Podcast Dataset); engagement = <i>stream rate</i>	Quantitative
Rizwan et al. 2025 (ACM MM)	How does toxicity escalate across reply chains in political podcasts?	9,166 episodes, 31 channels , 2022–23; Whisper + pyannote	Quantitative
Moldovan et al. 2025	How is participation in collective action articulated?	155,784 SPoRC episodes → 139,200 with collective-action statements	Mixed
Pathiyan Cherumanal et al. 2024 (ICMI)	Can real-time auditory alerts flag misinformation?	<i>position paper</i>	Design

Reddy, Sravana, Marina Lazarova, Yongze Yu, and Rosie Jones. "Modeling Language Usage and Listener Engagement in Podcasts." *ACL 2021*. arXiv:2106.06605 <https://arxiv.org/abs/2106.06605>

Rizwan, Naquee, Nayandeep Deb, Sarthak Roy, Vishwajeet Singh Solanki, Kiran Garimella, and Animesh Mukherjee. "Toxicity Begets Toxicity: Unraveling Conversational Chains in Political Podcasts." *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)*. DOI: [10.1145/3746027.3754553](https://doi.org/10.1145/3746027.3754553) · arXiv:2501.12640 <https://arxiv.org/abs/2501.12640>

Moldovan, Theodora, Arianna Pera, Davide Vega, and Luca Maria Aiello. "Podcasts as a Medium for Participation in Collective Action: A Case Study of Black Lives Matter." arXiv:2509.13197, 16 September 2025. <https://arxiv.org/abs/2509.13197>

Pathiyan Cherumanal, Sachin, Ujwal Gadiraju, and Damiano Spina. "Everything We Hear: Towards Tackling Misinformation in Podcasts." *ACM ICMI '24*, Blue Sky Papers (Third Place). DOI: [10.1145/3678957.3678959](https://doi.org/10.1145/3678957.3678959) · arXiv:2408.00292 <https://arxiv.org/abs/2408.00292>

Recent work: Audience and Applied Domains

Study	Question	Data	Method
Vilceanu 2025 (<i>Media & Communication</i>)	How does parasocial intimacy relate to nostalgia when a show changes?	12,000+ listener reviews , 2 long-running shows	Mixed
Sharon & John 2024 (<i>New Media & Society</i>)	How do <i>creators</i> imagine and address absent listeners?	Interviews with podcasters	Qualitative
Greeves et al. 2025 (<i>Frontiers in Communication</i>)	What characterizes journal-affiliated health podcasts?	400 episodes, 10 journals	Quantitative content analysis (Gwet's AC $\geq$ 0.80)

Vilceanu, M. Olguta. "Parasocial Intimacy, Change, and Nostalgia in Podcast Listener Reviews." *Media and Communication* 13 (2025), article 9059. DOI: [10.17645/mac.9059](https://doi.org/10.17645/mac.9059) <https://www.cogitatiopress.com/mediaandcommunication/article/view/9059>

Sharon, Tzlil, and Nicholas A. John. "'It's between me and myself': Inverse parasocial relationships in addressing (imagined) podcast listeners." *New Media & Society*, 2024. DOI: [10.1177/14614448241287913](https://doi.org/10.1177/14614448241287913)

Greeves, Scott, Rhessa Ledbetter, Rachel McGovern, and Keith Wiley. "Journal podcasting of health and medical science." *Frontiers in Communication* 10 (2025), article 1589099. DOI: [10.3389/fcomm.2025.1589099](https://doi.org/10.3389/fcomm.2025.1589099) <https://www.frontiersin.org/journals/communication/articles/10.3389/fcomm.2025.1589099/full>

Current Gaps in Podcast Research

- **Non-English podcasts** — the medium is global; the literature is overwhelmingly English
- **Audio-native features** — nearly everyone analyzes transcripts and discards the signal that made it audio
- **Longitudinal work** — most corpora are snapshots (e.g., SPoRC is only two months)
- **Causal designs** — reception is unobservable without extra effort, so measuring effects is hard
 - Some panels do allow collecting listener data, but still limited in sample size
- **The corpus itself** — the Spotify dataset was taken offline in Dec 2023

Preparing for Part 2

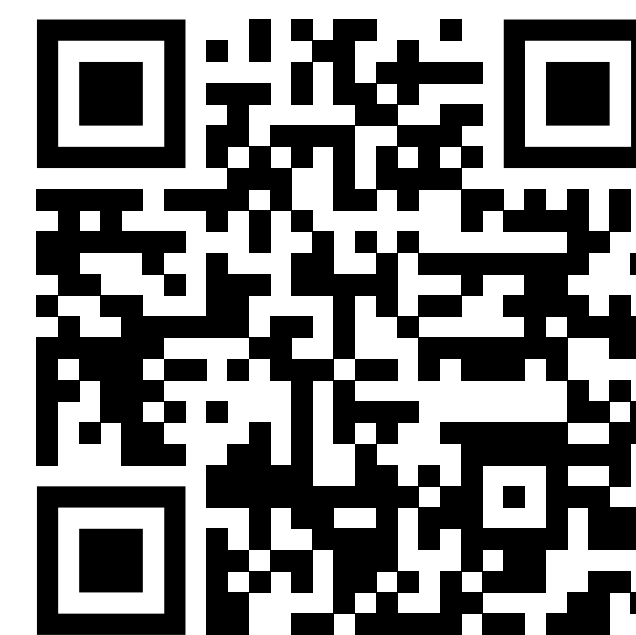
(after the coffee break)

- Get go to the Huggingface page for SPoRC and request access to the data
- Go to the tutorial page and run through the commands to install the packages and start downloading data in **INSTALL.md**:
 - <https://tinyurl.com/ic2s2podcasttutorial>
- You'll only download ~1GB of SPoRC data per genre(s) of choice
 - You can get more if you want!
- Better to start the install now

Website



SPoRC



Code

