Anonymizing search click-and-query data while maintaining utility

Joseph Jerome, DuckDuckGo

on behalf of

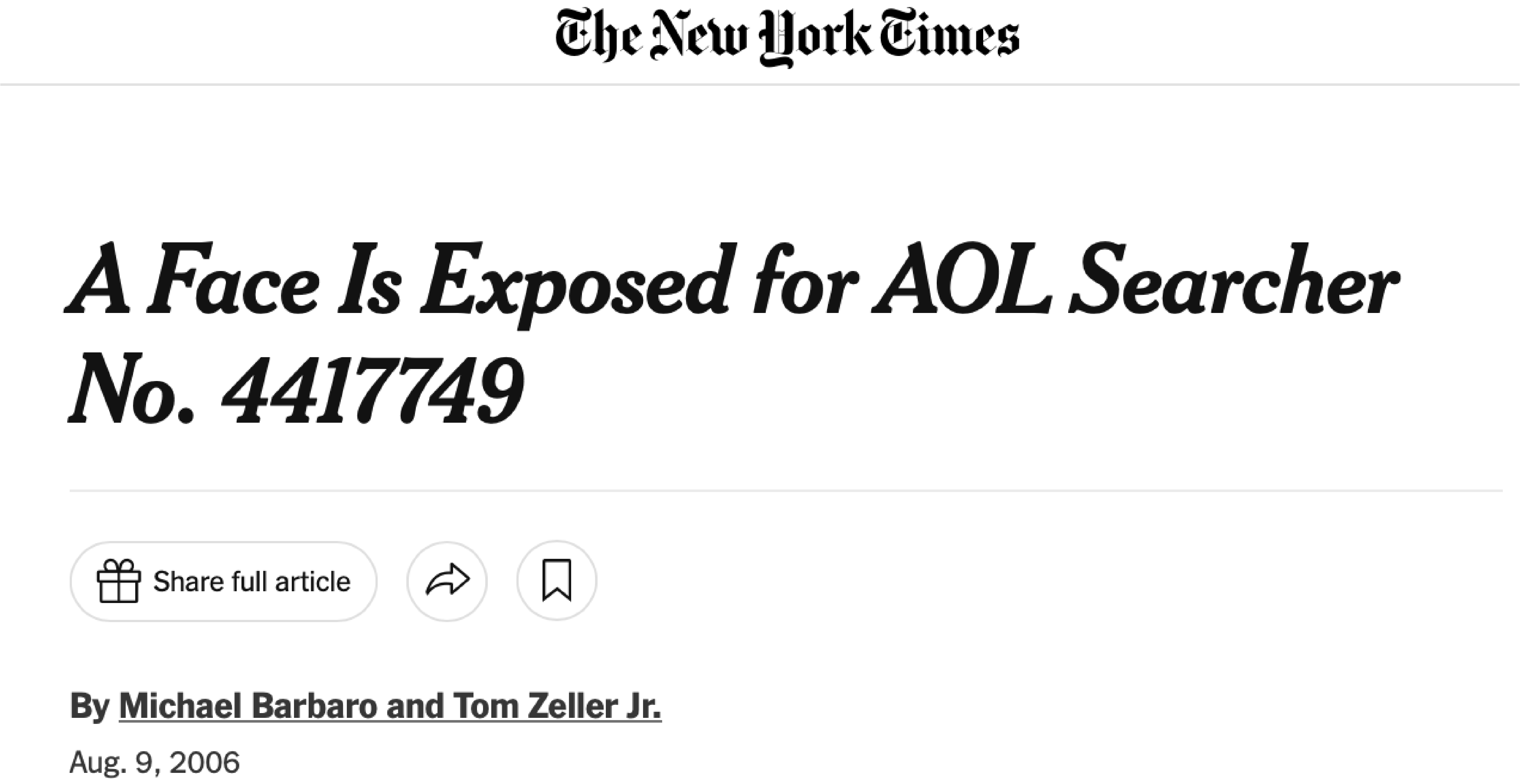
Paul Francis (Director at the Max Planck Institute)

Andreas Dewes (Privacy Engineer, DuckDuckGo)

February 5, 2026

Click and Query / User-Side Data

- The U.S. User-Side Data remedy and the EU Article 6(11) obligation differ in scope, frequency, and duration, but they pose similar **privacy challenges** because they both require sharing **user query text**, **click behavior**, and **result interaction data**.
    - *U.S. v. Google*: "User-side Data" encompasses "all data that can be obtained from users . . . through a search engine's interaction with the user's Device, by automated means. User-side Data includes information Google collects when answering commercial, tail, and local queries."
    - EU DMA: Article 6(11) requires provision of anonymous "ranking, query, click and view data."



The New York Times

A Face Is Exposed for AOL Searcher No. 4417749

By Michael Barbaro and Tom Zeller Jr.

Aug. 9, 2006

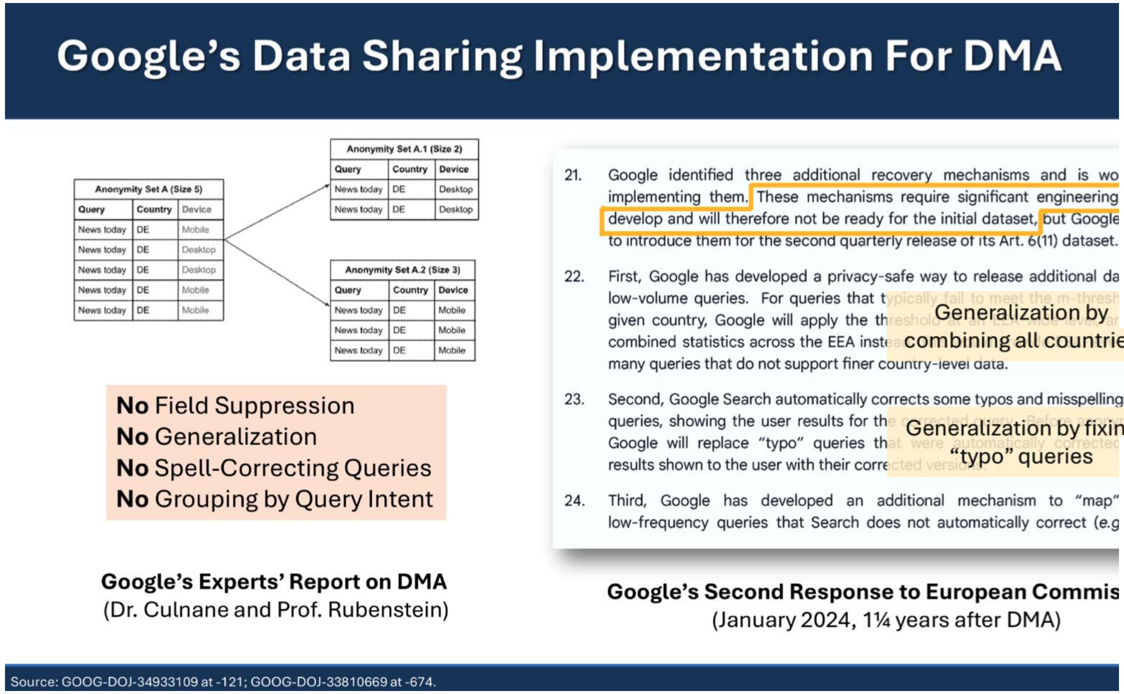
EU Search Data Licensing Program

Google’s current terms for search click and query data under Article 6(11) DMA includes all European Economic Area (EEA) queries where, cumulatively:

- The query has been searched by at least 30 signed in users across the world in the past 13 months (**k-threshold**).
- The query has been searched by at least 5 different, unique signed-in users with the same query / result / device (or device type) / country combination in the relevant quarter (**m-threshold**). If the m-threshold is not met in a given EEA country, Google will provide the data at EEA level if the threshold is met at that level.

This methodology omits **more than 99% of distinct search queries**, excluding 42% of Google’s total volume of queries.

Google's approach "does not seem like the approach one would use if you wanted to release high utility data." – Prof. David Evans



Google's Data Sharing Implementation For DMA

Anonymity Set A (Size 5)		
Query	Country	Device
News today	DE	Mobile
News today	DE	Desktop
News today	DE	Desktop
News today	DE	Mobile
News today	DE	Mobile

Anonymity Set A.1 (Size 2)		
Query	Country	Device
News today	DE	Desktop
News today	DE	Desktop

Anonymity Set A.2 (Size 3)		
Query	Country	Device
News today	DE	Mobile
News today	DE	Mobile
News today	DE	Mobile

No Field Suppression
No Generalization
No Spell-Correcting Queries
No Grouping by Query Intent

Google's Experts' Report on DMA
(Dr. Culnane and Prof. Rubenstein)

21. Google identified three additional recovery mechanisms and is wo implementing them. These mechanisms require significant engineering develop and will therefore not be ready for the initial dataset, but Google to introduce them for the second quarterly release of its Art. 6(11) dataset.

22. First, Google has developed a privacy-safe way to release additional da low-volume queries. For queries that typically fail to meet the m-thresh given country, Google will apply the threshold at an EEA-wide level, ar combined statistics across the EEA inste[illegible]ta for many queries that do not support finer country-level data.

Generalization by combining all countries

23. Second, Google Search automatically corrects some typos and misspellings in user queries, showing the user results for the corrected query. Before anonymization, Google will replace "typo" queries that were automatically corrected results shown to the user with their corrected versions.

Generalization by fixing "typo" queries

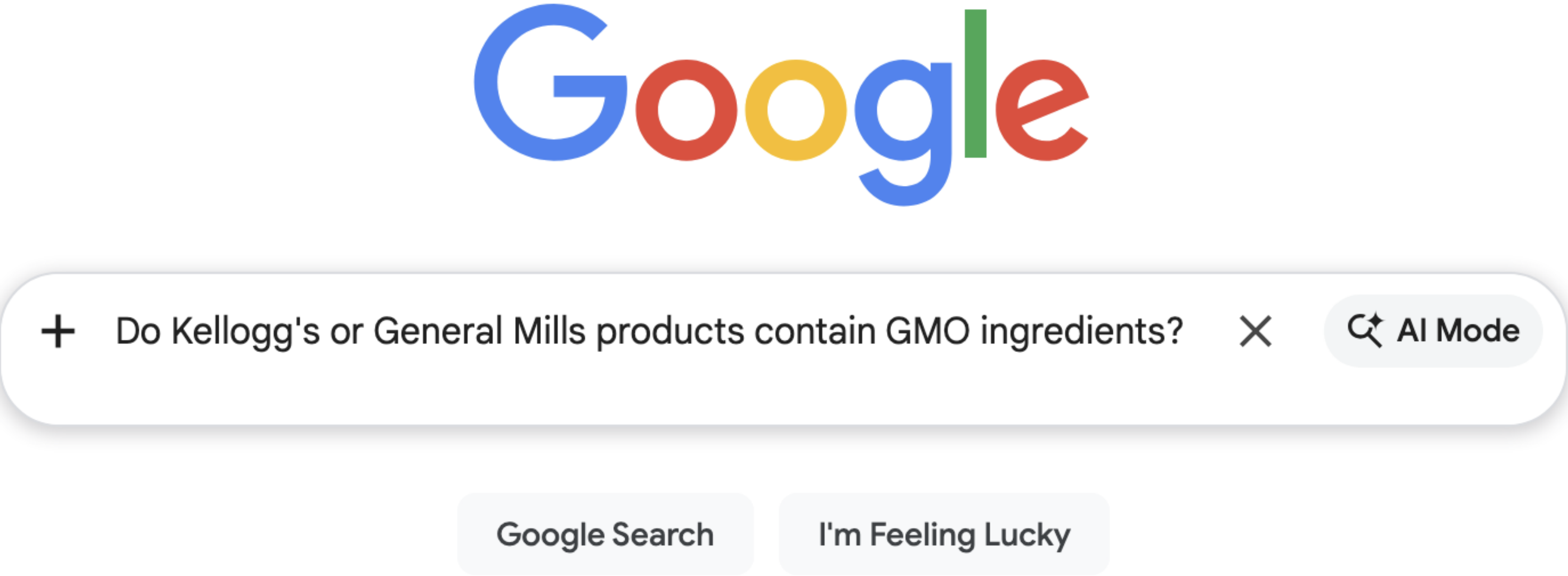
24. Third, Google has developed an additional mechanism to "map" low-frequency queries that Search does not automatically correct (e.g

Google's Second Response to European Commis
(January 2024, 1¼ years after DMA)

Source: GOOG-DOJ-34933109 at -121; GOOG-DOJ-33810669 at -674.

THE WITNESS: Right. If these were implemented, they would be able to release more data instead of this small, only releasing 1 percent of the queries that they were doing up to this point. And these are things they were saying they were considered doing but would require a significant engineering effort to be able to do that for the DMA.

The second one of correcting spelling typos, this is something that I have students do in the introductory computer science course that I teach for nonmajors. So they're very well known, well established, and it's, of course, something Google already does on every search query.



Do Kellogg's or General Mills products contain GMO ingredients?

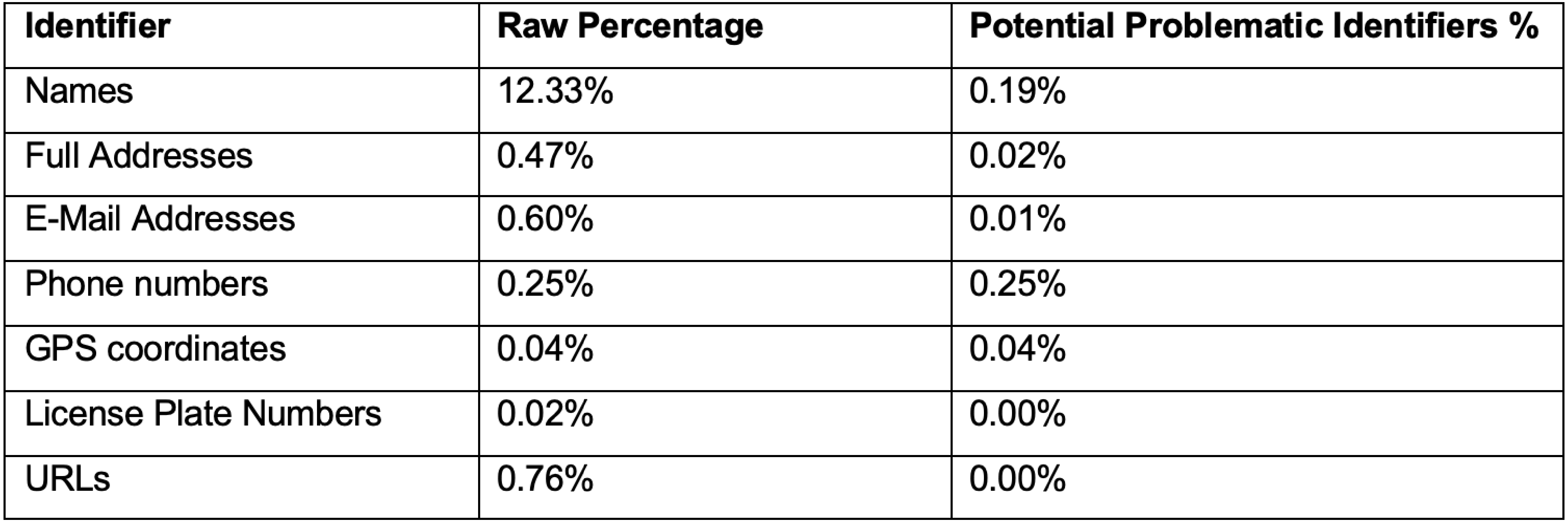
European Commission/European Data Protection Board’s **Joint Guidelines on the Interplay between the GDPR and DMA**

- The Joint Guidelines clarify that only the **personal data of the end user generating the search data needs to be anonymized**, not personal information about other individuals that might appear in queries (e.g., a person mentioned in a search).
- The Joint Guidelines acknowledge that anonymization should be achieved through a combination of **technical measures** complemented by **organizational**, **administrative** and **contractual measures**.

A Counterproposal

- Aligned with the prescriptions in the European Commission/European Data Protection Board's **Joint Guidelines on the Interplay between the GDPR and DMA**.
- Using DuckDuckGo's datasets, this analysis shows that less than **5% of distinct queries need to be removed** due to privacy risks (in contrast to Google removing more than 99%).
- Propose to remove queries using:
  - Targeted PII filtering (identifiers, addresses, etc.),
  - Metadata generalization, especially location,
  - Frequency-based word filtering, to eliminate potential unidentified PII

Potential identifiers are uncommon in search queries



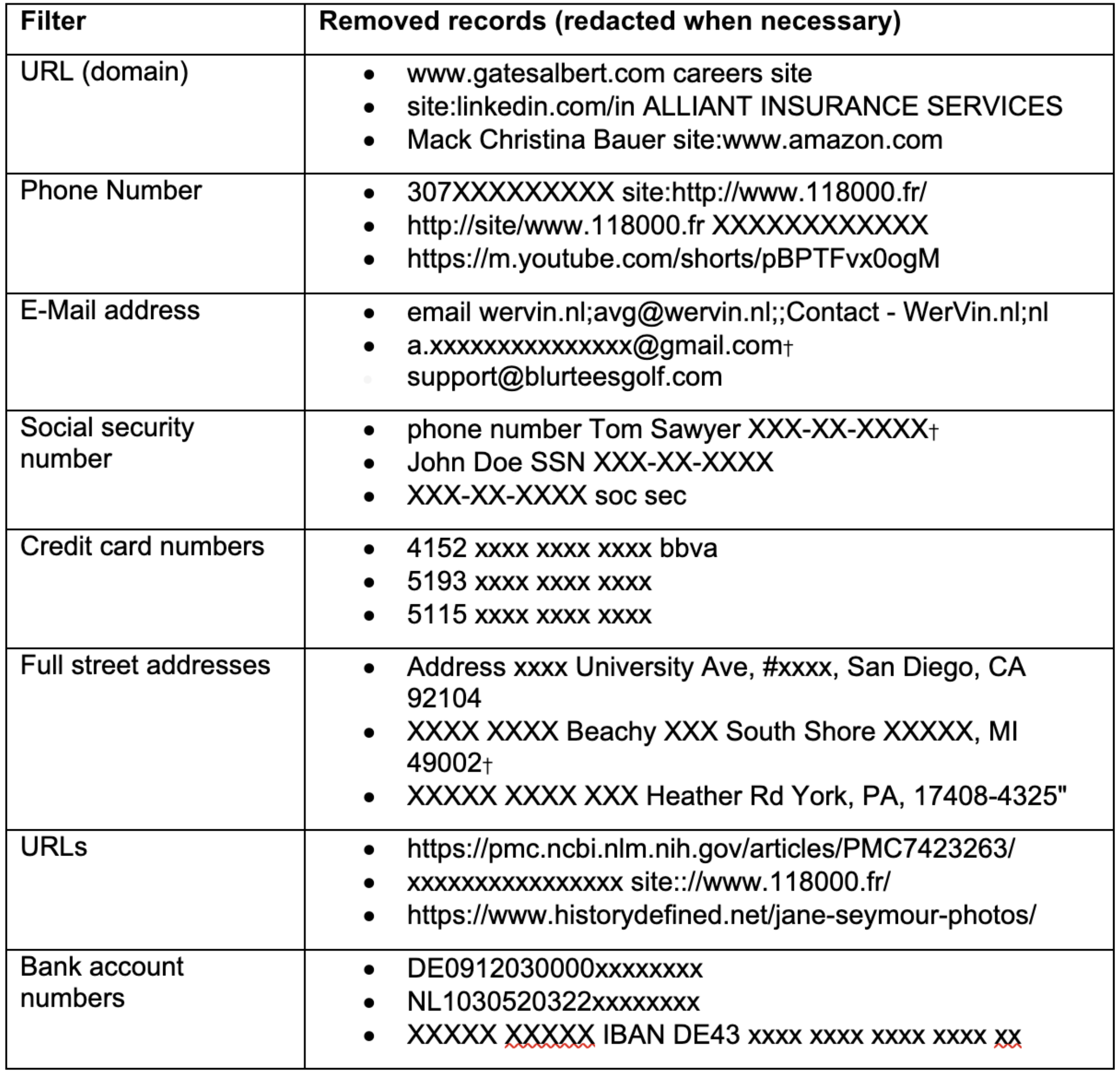
Identifier	Raw Percentage	Potential Problematic Identifiers %
Names	12.33%	0.19%
Full Addresses	0.47%	0.02%
E-Mail Addresses	0.60%	0.01%
Phone numbers	0.25%	0.25%
GPS coordinates	0.04%	0.04%
License Plate Numbers	0.02%	0.00%
URLs	0.76%	0.00%

Step 1: Filtering known identifiers

For testing purposes, filtered:

- **Email addresses**
- **Phone numbers**
- **Social Security numbers**
- **Credit card numbers**
- **URLs**
- **Bank account numbers**
- **Full street addresses**

The queries filtered by these techniques include both problematic queries requiring removal and false positives from a strict anonymization perspective.

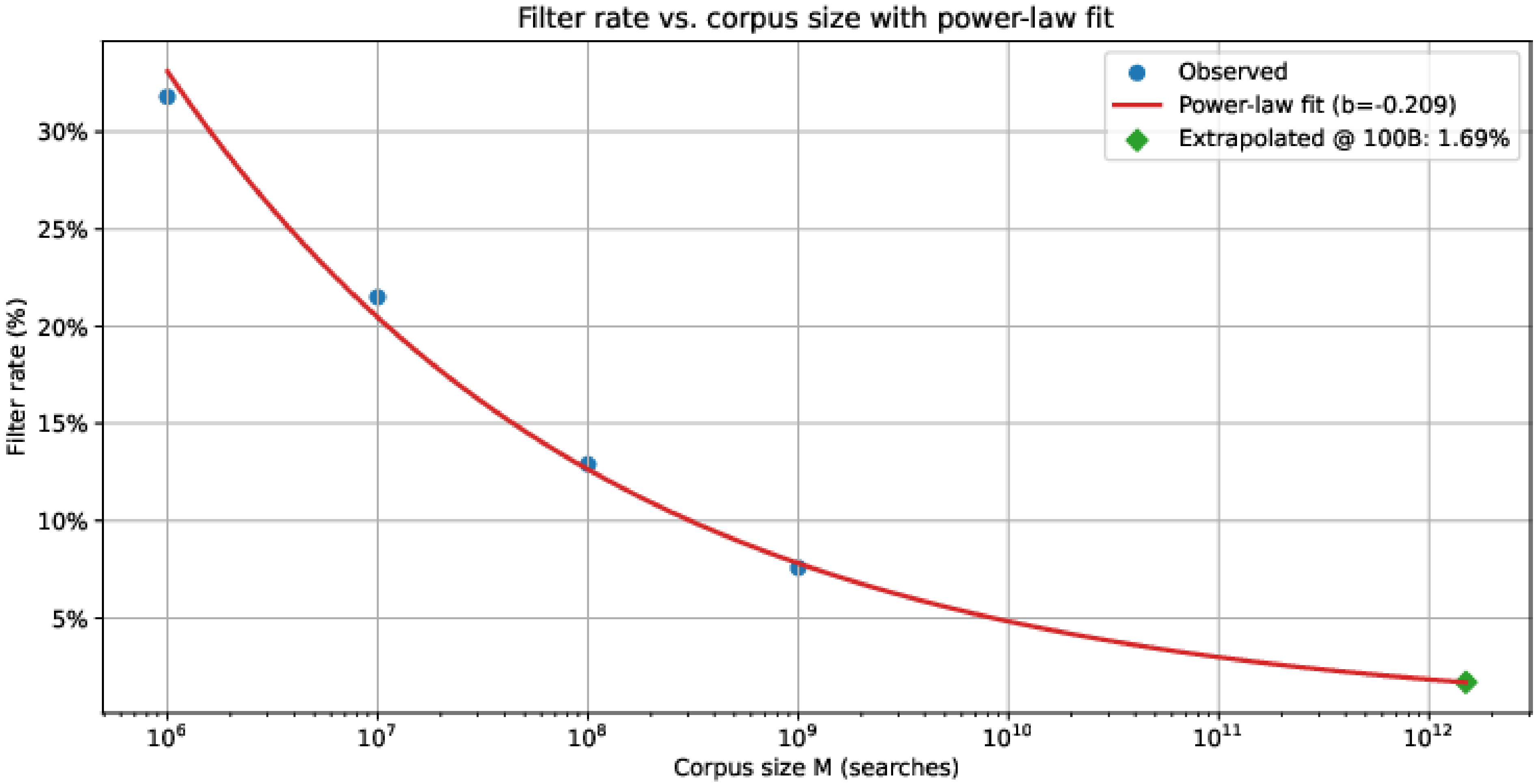


Filter	Removed records (redacted when necessary)
URL (domain)	• www.gatesalbert.com careers site • site:linkedin.com/in ALLIANT INSURANCE SERVICES • Mack Christina Bauer site:www.amazon.com
Phone Number	• 307XXXXXXXXX site:http://www.118000.fr/ • http://site/www.118000.fr XXXXXXXXXXX • https://m.youtube.com/shorts/pBPTFvx0ogM
E-Mail address	• email wervin.nl;avg@wervin.nl;;Contact - WerVin.nl;nl • a.xxxxxxxxxxxxxxx@gmail.com† support@blurteesgolf.com
Social security number	• phone number Tom Sawyer XXX-XX-XXXX† • John Doe SSN XXX-XX-XXXX • XXX-XX-XXXX soc sec
Credit card numbers	• 4152 xxxx xxxx xxxx bbva • 5193 xxxx xxxx xxxx • 5115 xxxx xxxx xxxx
Full street addresses	• Address xxxx University Ave, #xxxx, San Diego, CA 92104 • XXXX XXXX Beachy XXX South Shore XXXXX, MI 49002† • XXXXX XXXX XXX Heather Rd York, PA, 17408-4325"
URLs	• https://pmc.ncbi.nlm.nih.gov/articles/PMC7423263/ • xxxxxxxxxxxxxxxx site:://www.118000.fr/ • https://www.historydefined.net/jane-seymour-photos/
Bank account numbers	• DE0912030000xxxxxxxx • NL1030520322xxxxxxxx • XXXXX XXXXX IBAN DE43 xxxx xxxx xxxx xxxx xx

Step 2: Filtering unknown & uncommon words

- Apply a threshold to individual words in a query rather than the entire query text (as Google proposes):
  - Tokenize search queries into “words” by splitting queries on whitespace.
  - Use this word list to construct a dictionary of known words over a given time window and for a given language.
  - Filter out all queries that contain words which are below a given threshold, e.g., 10 occurrences in our reference data.
- An example:
  - Query: **“is ct3EZ944f in REI hacked passwords”**
  - Tokenization: **[‘is’, 'ct3EZ944f ', ‘in’, 'REI', 'hacked', 'passwords’]**
  - Dictionary Mismatch: **[‘ct3EZ944f’, 'REI’]**
    - REI is the name of a U.S.-based retailer and thus likely to be included in more than 10 queries.
    - ‘ct3EZ944f’ likely not be to be found in either a dictionary or searched more than 10 times, causing this query to be excluded.

Filtering Results



Step 3: Generalize explicit metadata with k-anonymity

- Useful metadata:
  - Approximate location
  - Device type ("mobile" or “desktop”)
  - Derived language (e.g., “en-US”)
  - Approximate timestamp (rounded to one day)
- Implementation:
  - If a metadata combination is observed in searches from fewer than k=1,000 users , generalize its attributes until it meets the threshold.
  - Group search query data records by device type, derived language, country, and approximate timestamp. (If any results below the k-threshold, drop the record.)
  - Partition records in each group by their associated exact location, forming groups based on a 100x100 meter grid based on WGS84. Merge any partition below the k-threshold with adjacent partitions necessary to form a new square partition.
  - Set the location of each record to the bounding box of the associated partition of the record.
- This process ensures that location information and other metadata is only shared to a level of detail that will not allow re-identification or singling out of users.

Thank you!

- Despite Google's claims to the contrary, **search click and query data can be safely anonymized in a way that maintains utility** for search business users.
  - These three steps result in the exclusion of 7.25% of DuckDuckGo English searches, which scales to ~2% at Google's volume. It addresses all potentially problematic queries in our manually-reviewed sample.
- The **European Commission's Article 6(11) enforcement proceeding** and the 6-month sprint for the ***U.S. v. Google*** **Technical Committee to recommend privacy safeguards** mean more attention is warranted.
- Any questions or follow-up? Please reach out to me at jjerome@duckduckgo.com.