AI and the Changing Landscape of Privacy Notice and Choice

Kassem Fawaz (University of Wisconsin-Madison)

Collaborators:

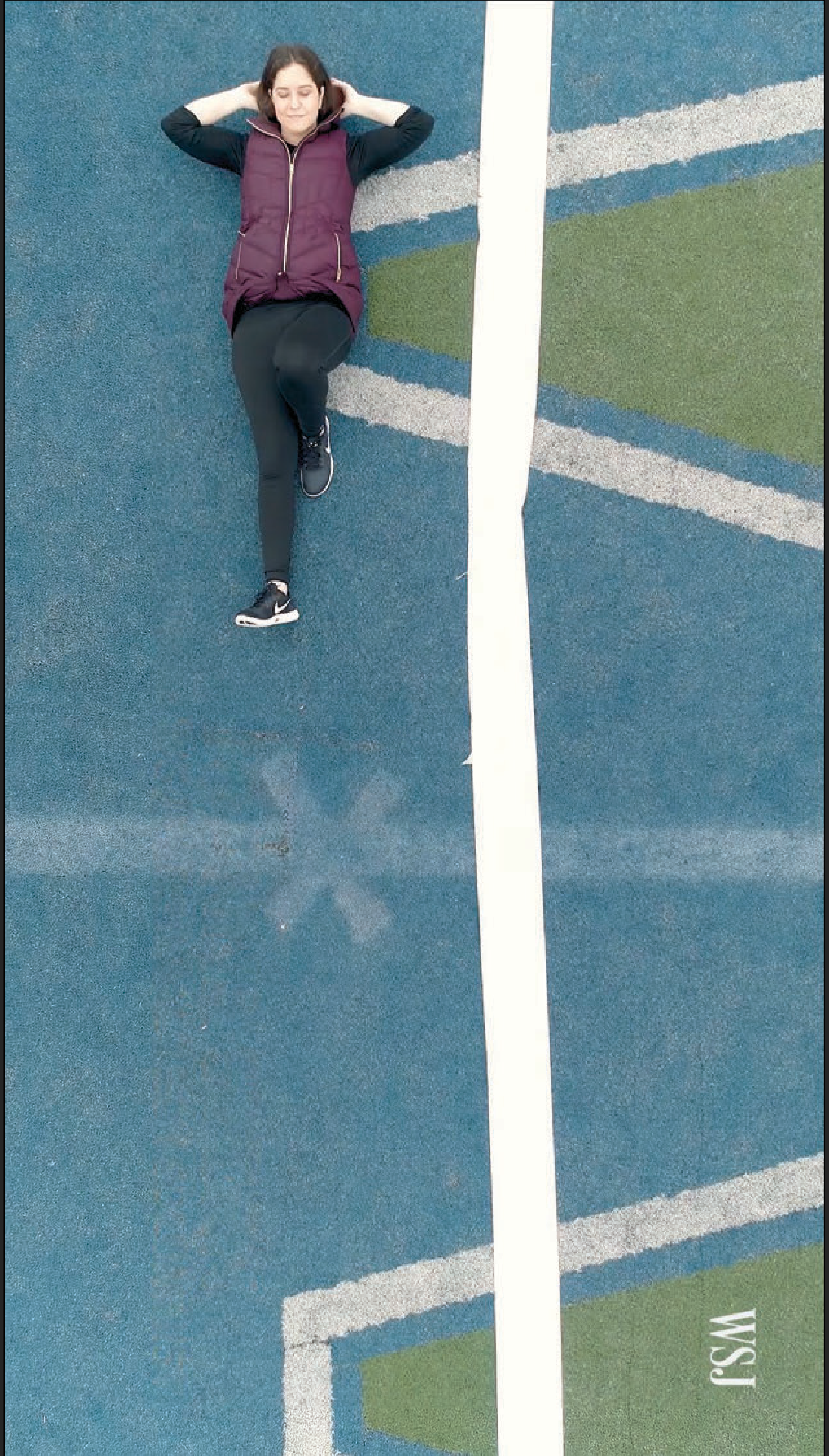
Hamza Harkous (EPFL)

Thomas Linden (University of Wisconsin-Madison)

pribot.org

kassemfawaz.com

Problem?



PRIVACY POLICIES, *AS THE DE FACTO PRIVACY NOTICE MECHANISM*, ARE LONG AND COMPLEX

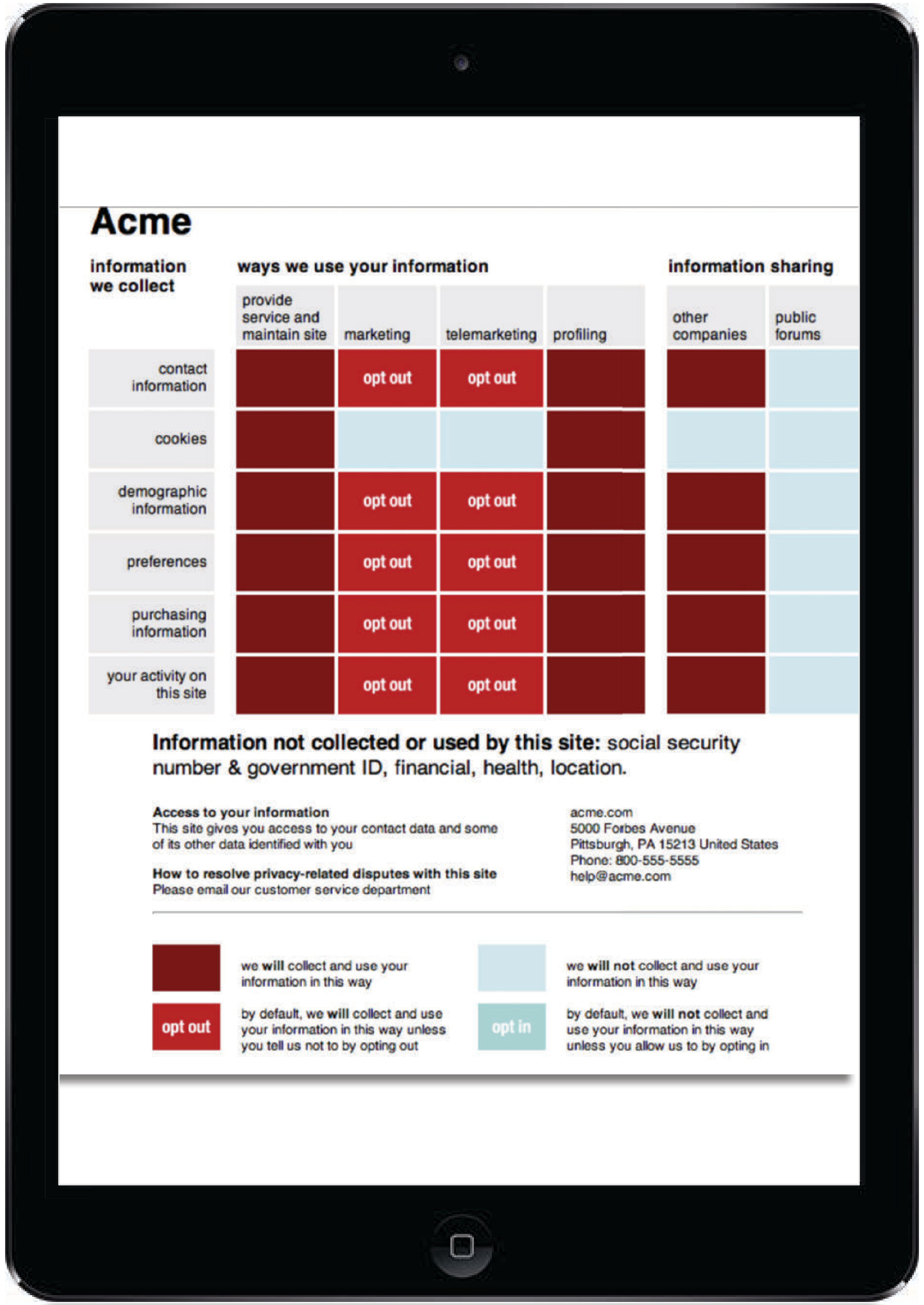
● ● ●

201 hours per year on average to read policies of services we encounter*

*(McDonald and Cranor -2008)

APPROACHES SO FAR?

Put more lawyers on the task.



Acme

information we collect	ways we use your information				information sharing	
	provide service and maintain site	marketing	telemarketing	profiling	other companies	public forums
contact information	🟥	opt out	opt out	🟥	🟥	🟦
cookies	🟥	🟦	🟦	🟥	🟦	🟦
demographic information	🟥	opt out	opt out	🟥	🟥	🟦
preferences	🟥	opt out	opt out	🟥	🟥	🟦
purchasing information	🟥	opt out	opt out	🟥	🟥	🟦
your activity on this site	🟥	opt out	opt out	🟥	🟥	🟦

Information not collected or used by this site: social security number & government ID, financial, health, location.

Access to your information
This site gives you access to your contact data and some of its other data identified with you

How to resolve privacy-related disputes with this site
Please email our customer service department

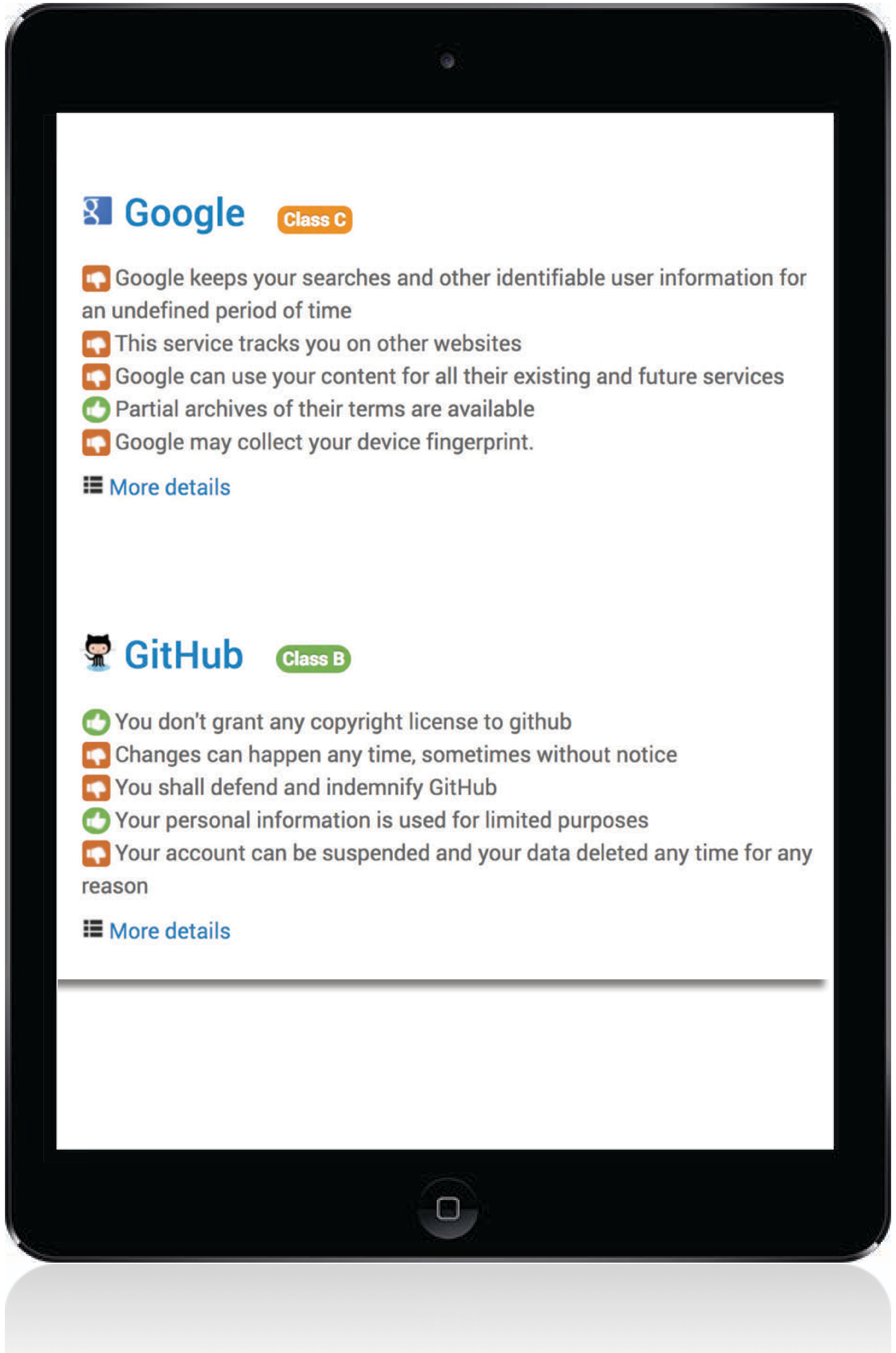
acme.com
5000 Forbes Avenue
Pittsburgh, PA 15213 United States
Phone: 800-555-5555
help@acme.com

🟥	we **will** collect and use your information in this way	🟦	we **will not** collect and use your information in this way
opt out	by default, we **will** collect and use your information in this way unless you tell us not to by opting out	opt in	by default, we **will not** collect and use your information in this way unless you allow us to by opting in

Standardization

- A Nutrition Label for privacy
- Required providers to act
- **Surprise**: They didn’t.

Kelley et al., "A nutrition label for privacy." SOUPS’09



Google Class C

- 👎 Google keeps your searches and other identifiable user information for an undefined period of time
- 👎 This service tracks you on other websites
- 👎 Google can use your content for all their existing and future services
- 👍 Partial archives of their terms are available
- 👎 Google may collect your device fingerprint.

More details

GitHub Class B

- 👍 You don't grant any copyright license to github
- 👎 Changes can happen any time, sometimes without notice
- 👎 You shall defend and indemnify GitHub
- 👍 Your personal information is used for limited purposes
- 👎 Your account can be suspended and your data deleted any time for any reason

More details

Crowdsourcing

- TOSDR.org
- Limited by volunteers' availability
- Available for ~100 policies
- Unstructured → can only be used for limited automated labeling*

*Zimmeck and Bellovin , "Privee: An Architecture for Automatically Analyzing Web Privacy Policies". USENIX Security 2014

Manual work doesn’t scale.

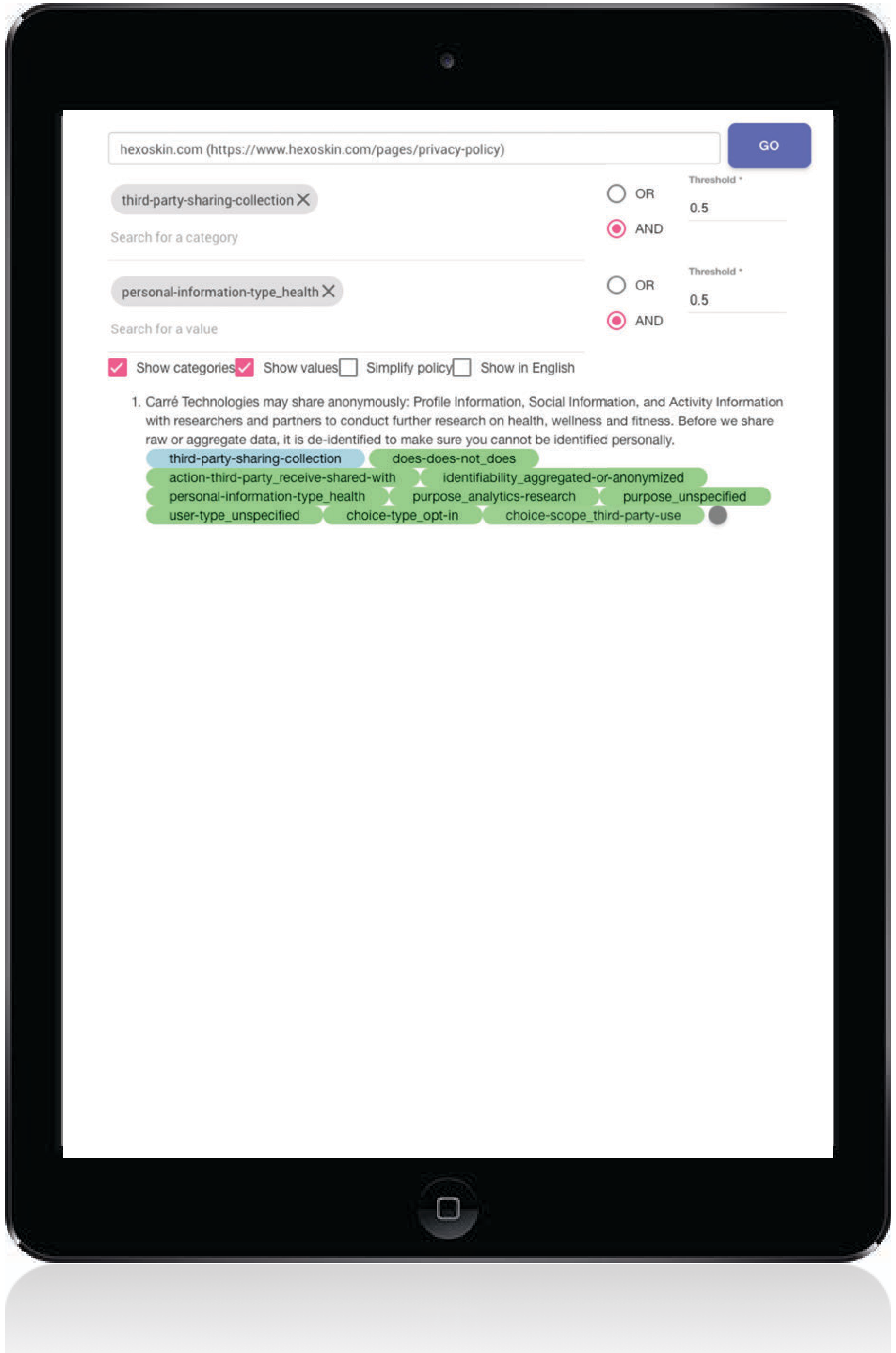
Fails to cope with emerging technologies.

Voice-Activated Devices

- Read the whole policy?
- Show tables on small screens?

Unstructured Query (User Questions)

Regulation Compliance (e.g. GDPR)



hexoskin.com (https://www.hexoskin.com/pages/privacy-policy) GO

third-party-sharing-collection ✕

Search for a category

○ OR ● AND Threshold * 0.5

personal-information-type_health ✕

Search for a value

○ OR ● AND Threshold * 0.5

☑ Show categories ☑ Show values ☐ Simplify policy ☐ Show in English

1. Carré Technologies may share anonymously: Profile Information, Social Information, and Activity Information with researchers and partners to conduct further research on health, wellness and fitness. Before we share raw or aggregate data, it is de-identified to make sure you cannot be identified personally.

third-party-sharing-collection
does-does-not_does
action-third-party_receive-shared-with
identifiability_aggregated-or-anonymized
personal-information-type_health
purpose_analytics-research
purpose_unspecified
user-type_unspecified
choice-type_opt-in
choice-scope_third-party-use

Find Statements About Health Data Sharing

```
Get Segments such that
Category: third party sharing
personal information type: health information
```

Structured Query

Solution

Types of info they collect
Collection reasons
What choices do they give?
Computer Information
Contact
Cookies And Tracking Elements
Demographic
Ip Address And Device Ids
Location
Survey Data
Personalization Customization
Advertising
Additional Service Feature
Other Purposes
Basic Service Feature
Analytics Research
Service Operation And Security
Marketing
Unspecified Choice
Browser Device Privacy Controls
Third Party Privacy Controls
Opt In
Opt Out Via Contacting Company

POLISIS

Unified Framework for Privacy Policies Analysis

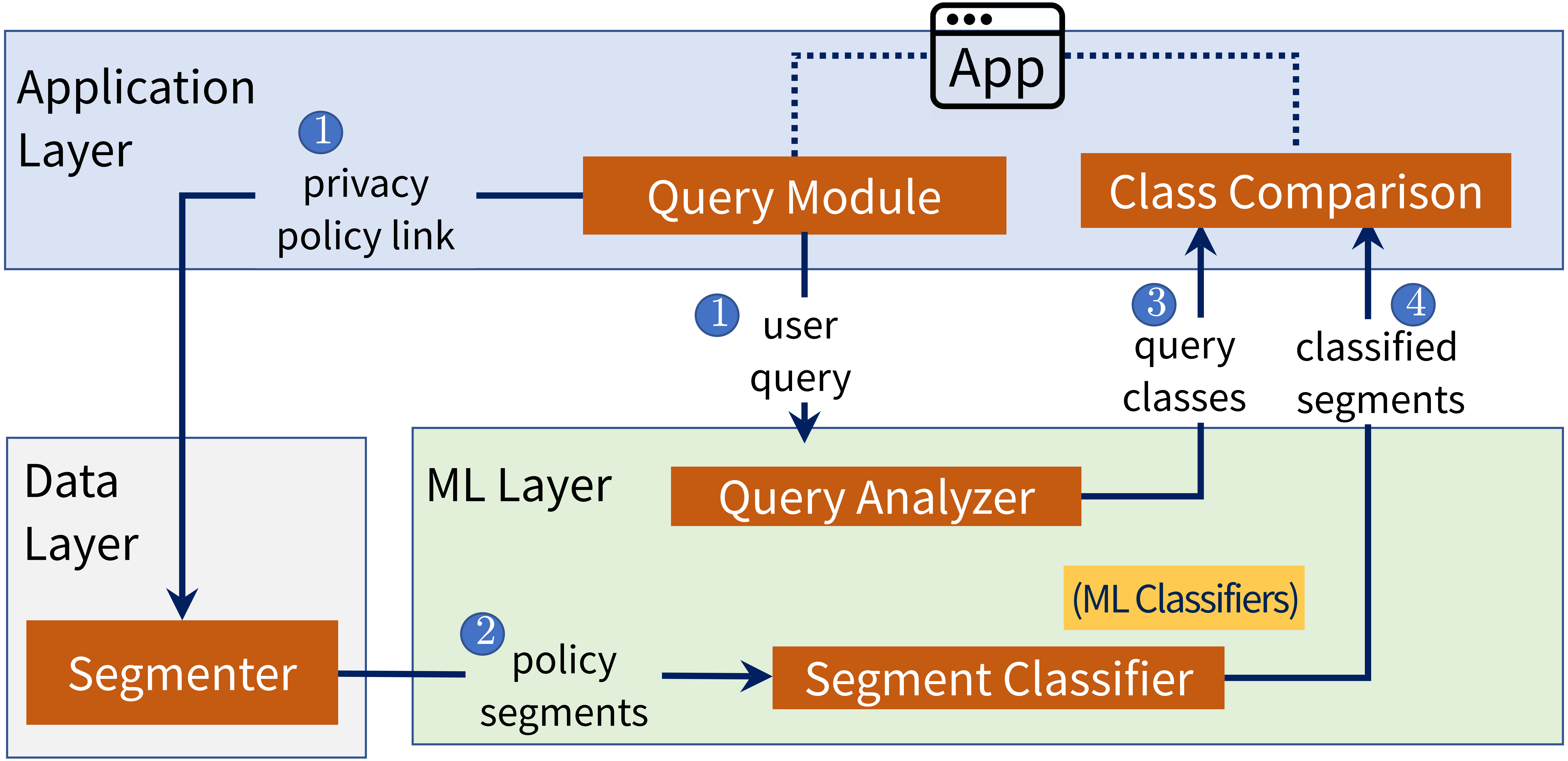
Once we automate policies' analysis, we can create a new interface for millions of policies with a single program.

Structured Queries

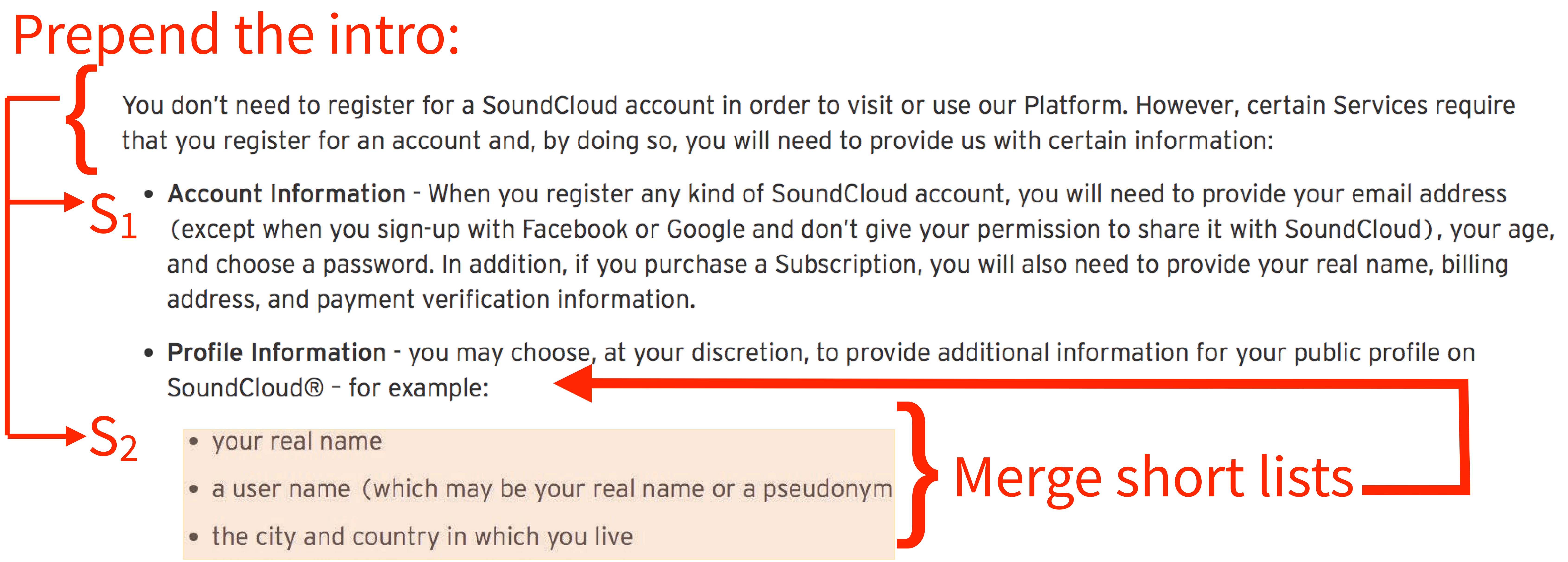
Unstructured Queries

pribot.org

Framework


App
Application Layer
1
privacy policy link
Query Module
Class Comparison
1
user query
3
query classes
4
classified segments
Data Layer
ML Layer
Query Analyzer
(ML Classifiers)
2
policy segments
Segmenter
Segment Classifier

Policy Segmenter

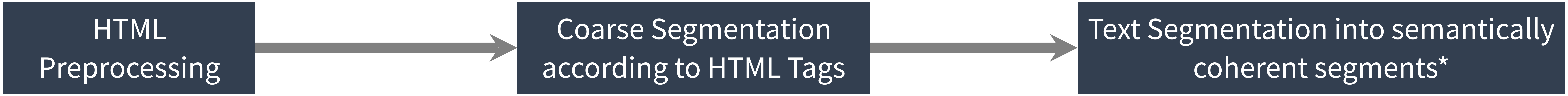


Prepend the intro:

You don't need to register for a SoundCloud account in order to visit or use our Platform. However, certain Services require that you register for an account and, by doing so, you will need to provide us with certain information:

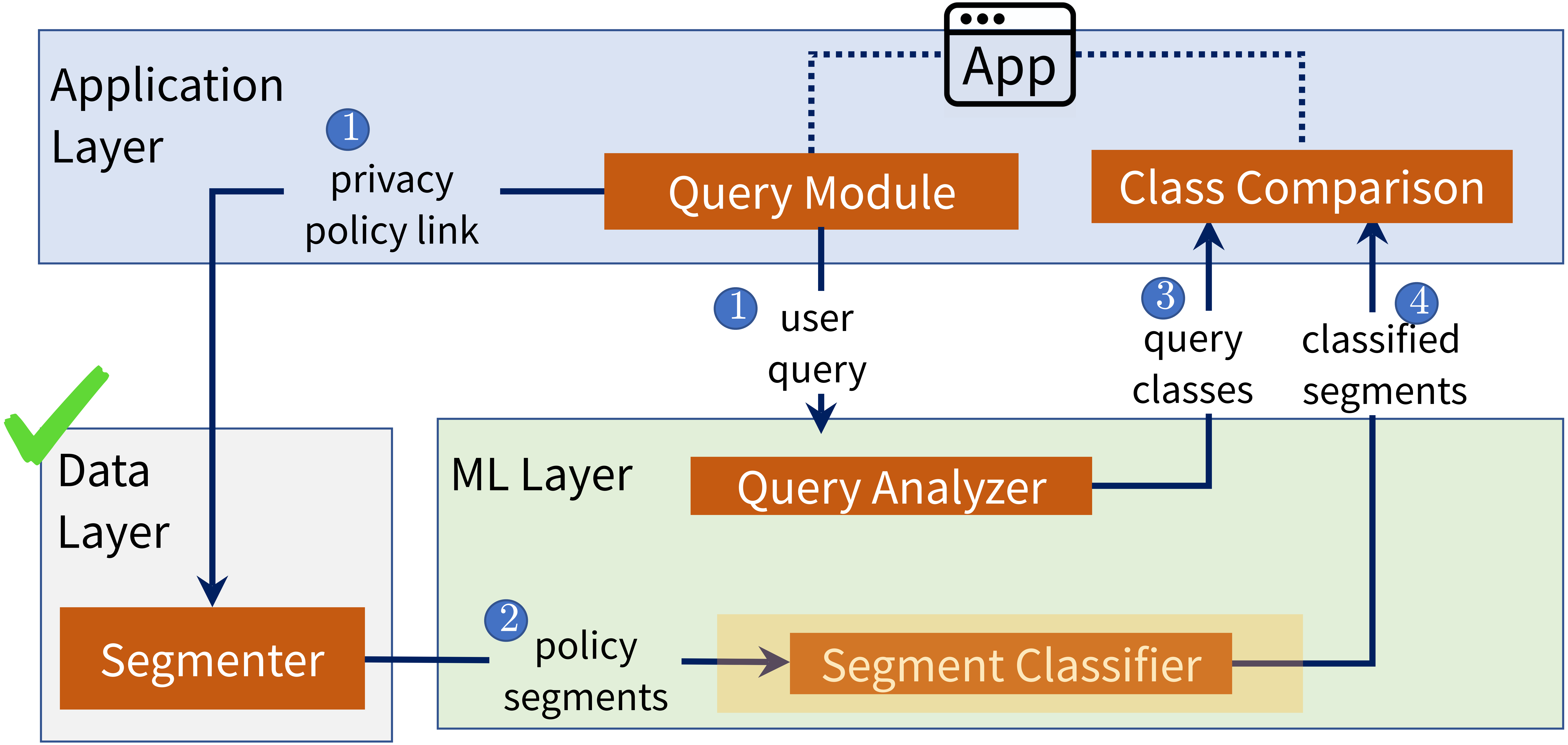
- **Account Information** - When you register any kind of SoundCloud account, you will need to provide your email address (except when you sign-up with Facebook or Google and don't give your permission to share it with SoundCloud), your age, and choose a password. In addition, if you purchase a Subscription, you will also need to provide your real name, billing address, and payment verification information.
- **Profile Information** - you may choose, at your discretion, to provide additional information for your public profile on SoundCloud® – for example:
  - your real name
  - a user name (which may be your real name or a pseudonym
  - the city and country in which you live

Merge short lists



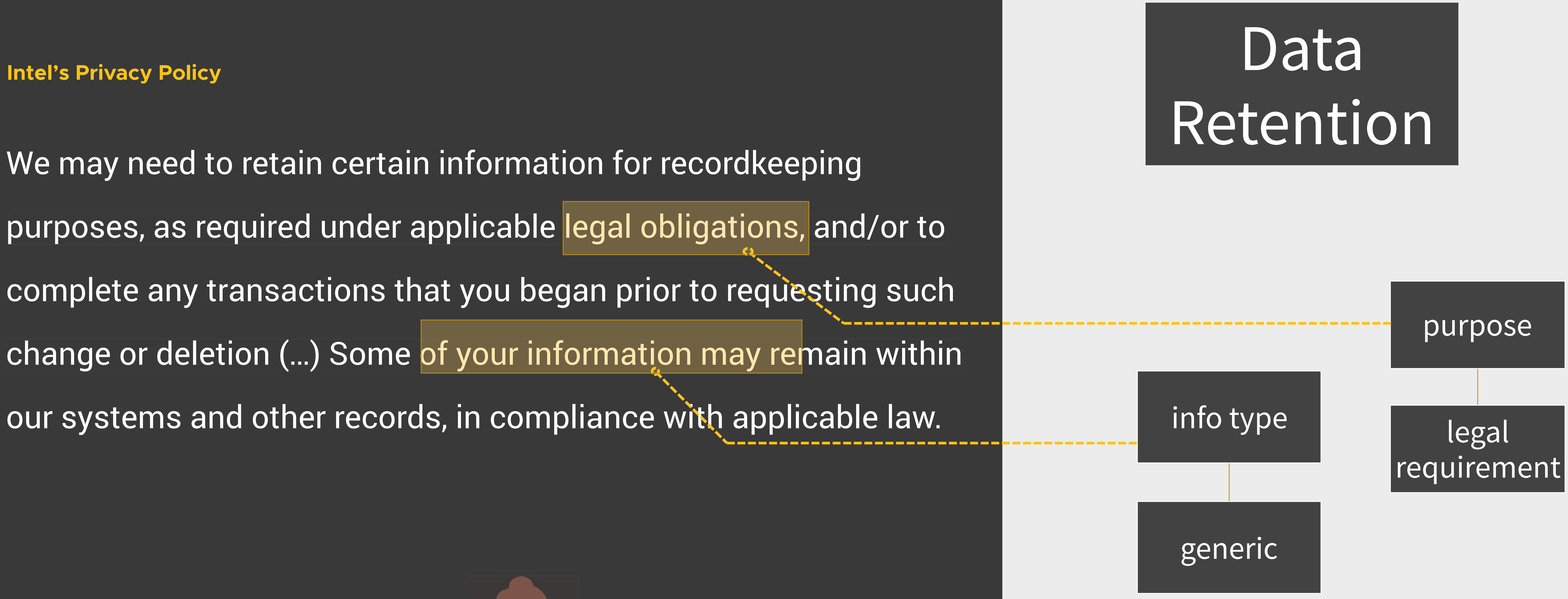
HTML Preprocessing
Coarse Segmentation according to HTML Tags
Text Segmentation into semantically coherent segments*

*Glavas et al., "Unsupervised Text Segmentation Using Semantic Relatedness Graphs", ACL 2016


App
Application Layer
1 privacy policy link
Query Module
Class Comparison
1 user query
3 query classes
4 classified segments
Data Layer
ML Layer
Query Analyzer
Segmenter
2 policy segments
Segment Classifier

EXAMPLE

Intel’s Privacy Policy



We may need to retain certain information for recordkeeping purposes, as required under applicable legal obligations, and/or to complete any transactions that you began prior to requesting such change or deletion (…) Some of your information may remain within our systems and other records, in compliance with applicable law.

EXPERT ANNOTATIONS

*OPP-115 dataset by Wilson et al., ACL 2016

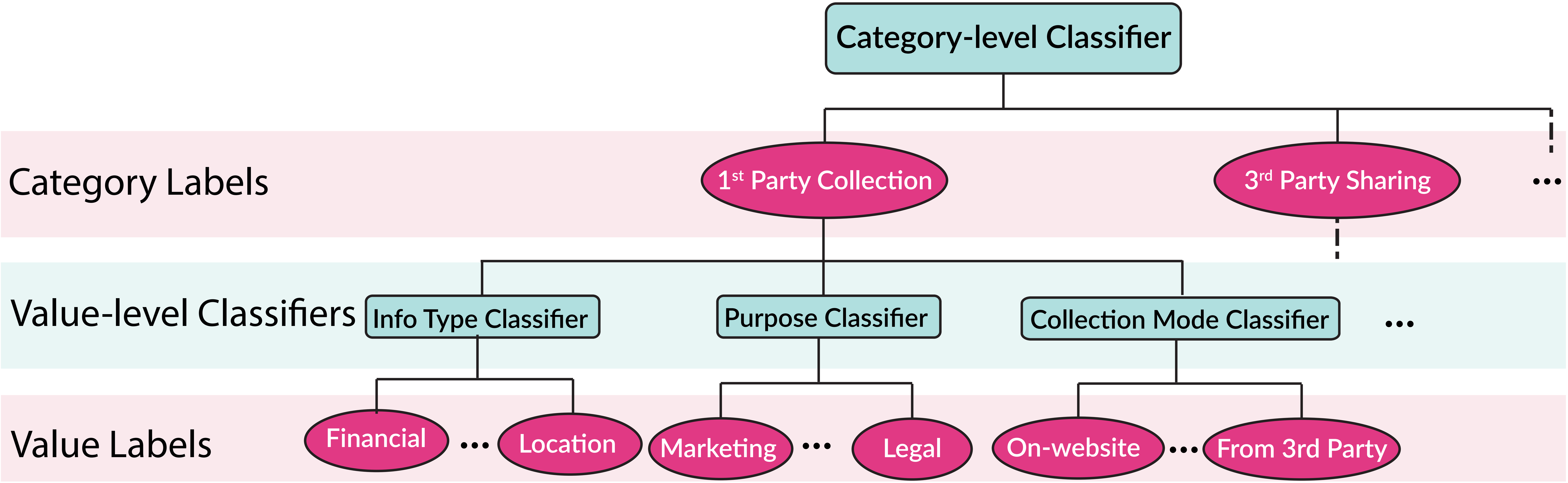
Online Privacy Policies Dataset

- 115 annotated policies
- 23K annotations

Information Type
• financial
• health
• contact
• location
• …
Purpose
• advertising
• marketing
• analytics
• legal requirement
• …
1st Party Collection
Collection Mode
Information Type
Purpose
3rd Party Sharing
Action
Information Type
Purpose
Access, Edit, Delete
Access Scope
Access Rights
Data Retention
Retention Period
Retention Purpose
Information Type
Retention Period
• stated period
• limited
• indefinitely
• unspecified
• other
Data Security
Security Measure
Specific Audiences
Audience group
Do Not Track
Do Not Track Policy
Policy Change
Change Type
User Choice
Notification Type
Other
Introductory
Contact Information
Practice not covered
Choice, Control
Choice Type
Choice Scope
Choice Type
• opt-in
• opt-out
• opt-out-link
• …

Hierarchical Data → Hierarchical Architecture

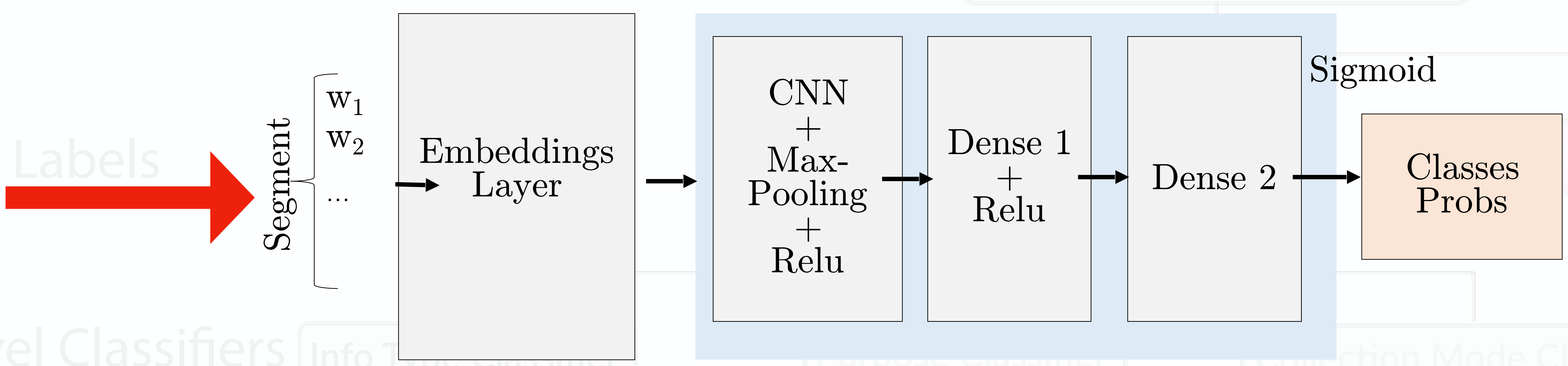
Hierarchical Architecture



Hierarchical Architecture

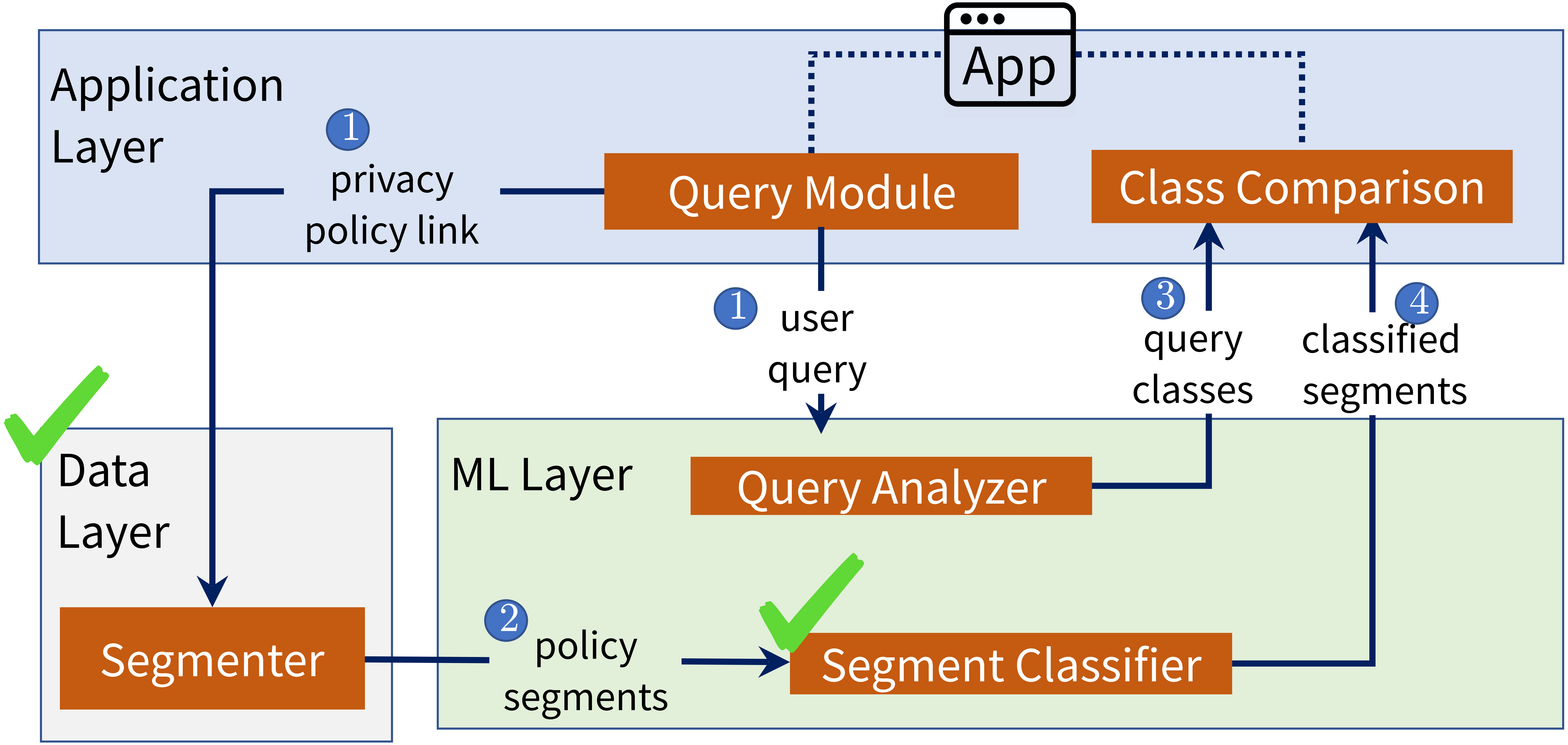
Similar architecture for the 21 classifiers

130,000 privacy policies from Play Store to train our custom word embeddings



Loss function: multi-label cross entropy

Embeddings size: 300, Number of filters: 200, Filter Size: 3, Dense Layer Size: 100, Batch Size: 40



Structured Querying

Unstructured Querying

Structured Querying

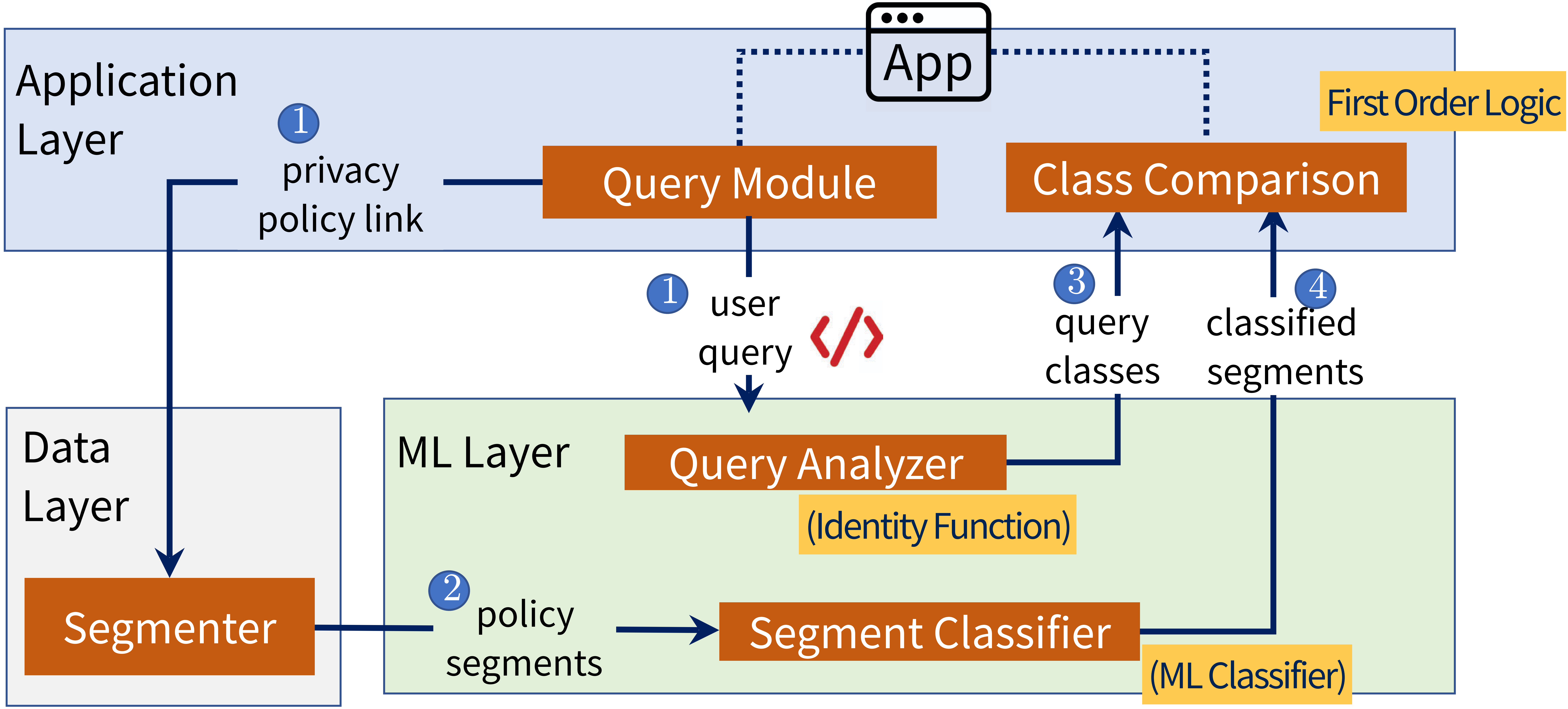
GDPR Compliance as a Case Study

Structured Query

```
Get Segments such that

Category: third party sharing

purpose:  advertising
```



What is the GDPR?

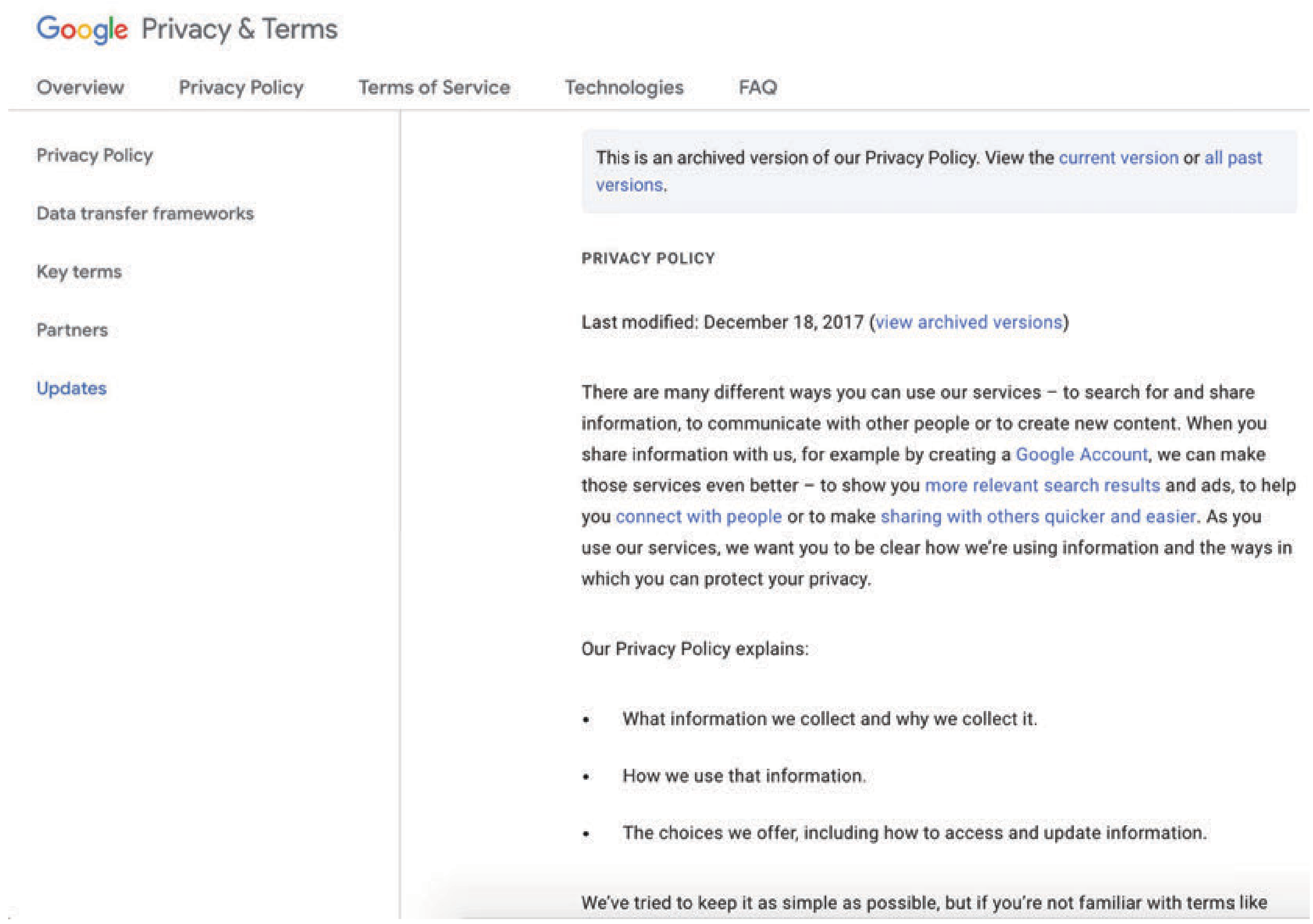
The General Data Protection Regulation

- **Approved:** 14 April 2016
- **Enforced:** 25 May 2018
- **Enforcement:** *max*(€20 million, 4% of the worldwide annual revenue):
  - The basic principles for processing, and consent, under Articles 5, 6, 7, and 9
  - The data subjects’ rights under Articles 12-22

Typically described in the privacy policy

How Did Privacy Policies Change?

1. Presentation
2. Readability
3. Coverage
4. Ambiguity
5. Compliance



This is an archived version of our Privacy Policy. View the current version or all past versions.

PRIVACY POLICY

Last modified: December 18, 2017 (view archived versions)

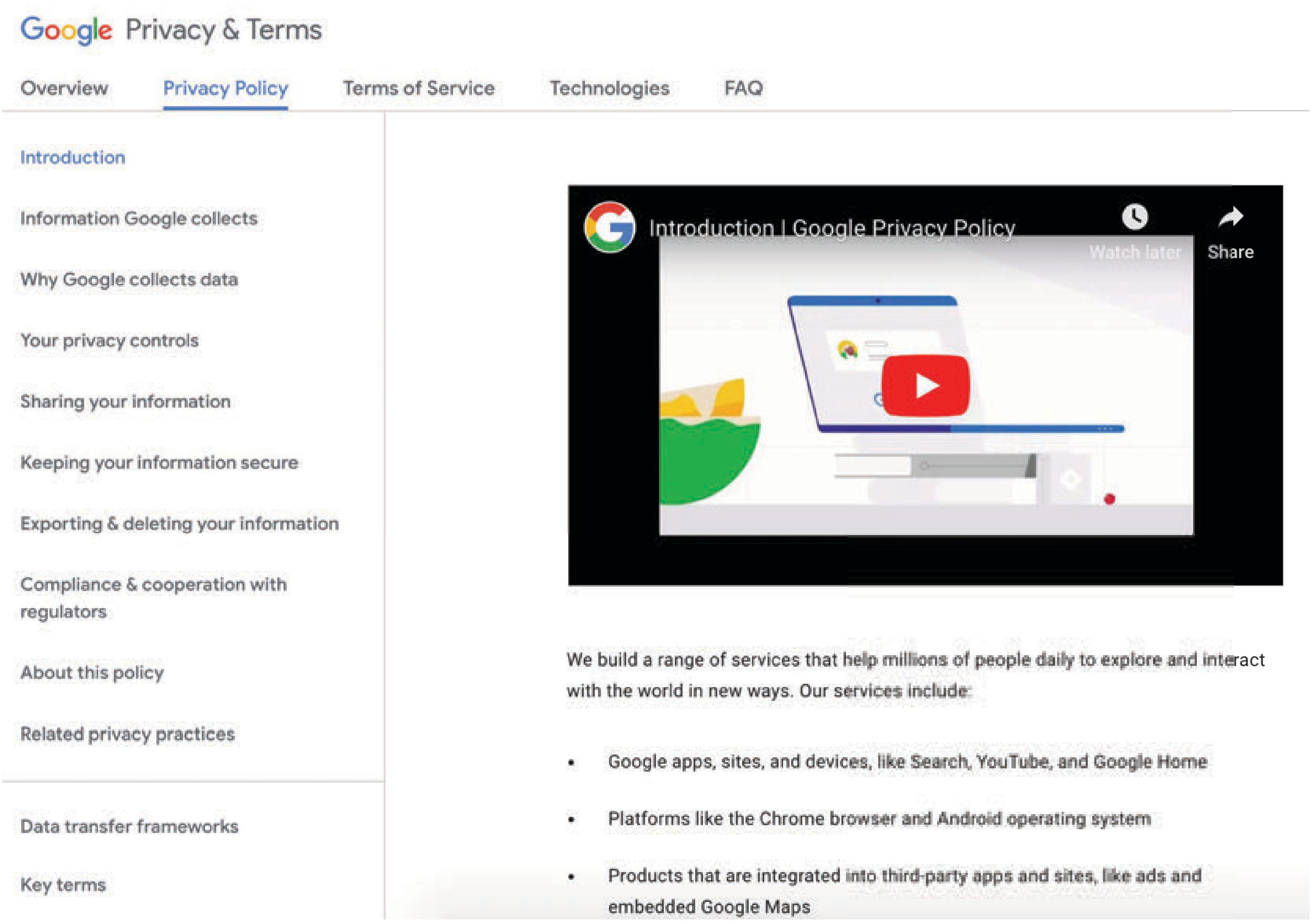
There are many different ways you can use our services – to search for and share information, to communicate with other people or to create new content. When you share information with us, for example by creating a Google Account, we can make those services even better – to show you more relevant search results and ads, to help you connect with people or to make sharing with others quicker and easier. As you use our services, we want you to be clear how we're using information and the ways in which you can protect your privacy.

Our Privacy Policy explains:

- What information we collect and why we collect it.
- How we use that information.
- The choices we offer, including how to access and update information.

We've tried to keep it as simple as possible, but if you're not familiar with terms like

Pre-GDPR Google Privacy Policy

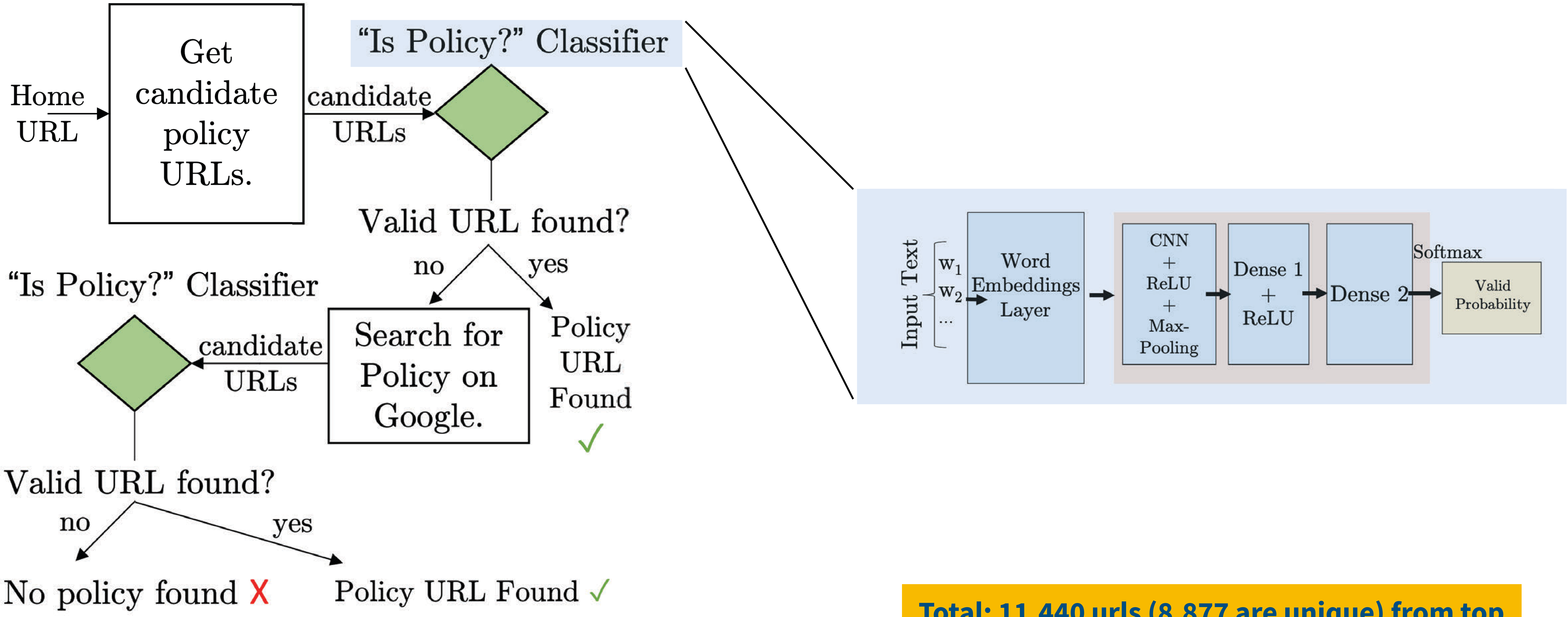


We build a range of services that help millions of people daily to explore and interact with the world in new ways. Our services include:

- Google apps, sites, and devices, like Search, YouTube, and Google Home
- Platforms like the Chrome browser and Android operating system
- Products that are integrated into third-party apps and sites, like ads and embedded Google Maps

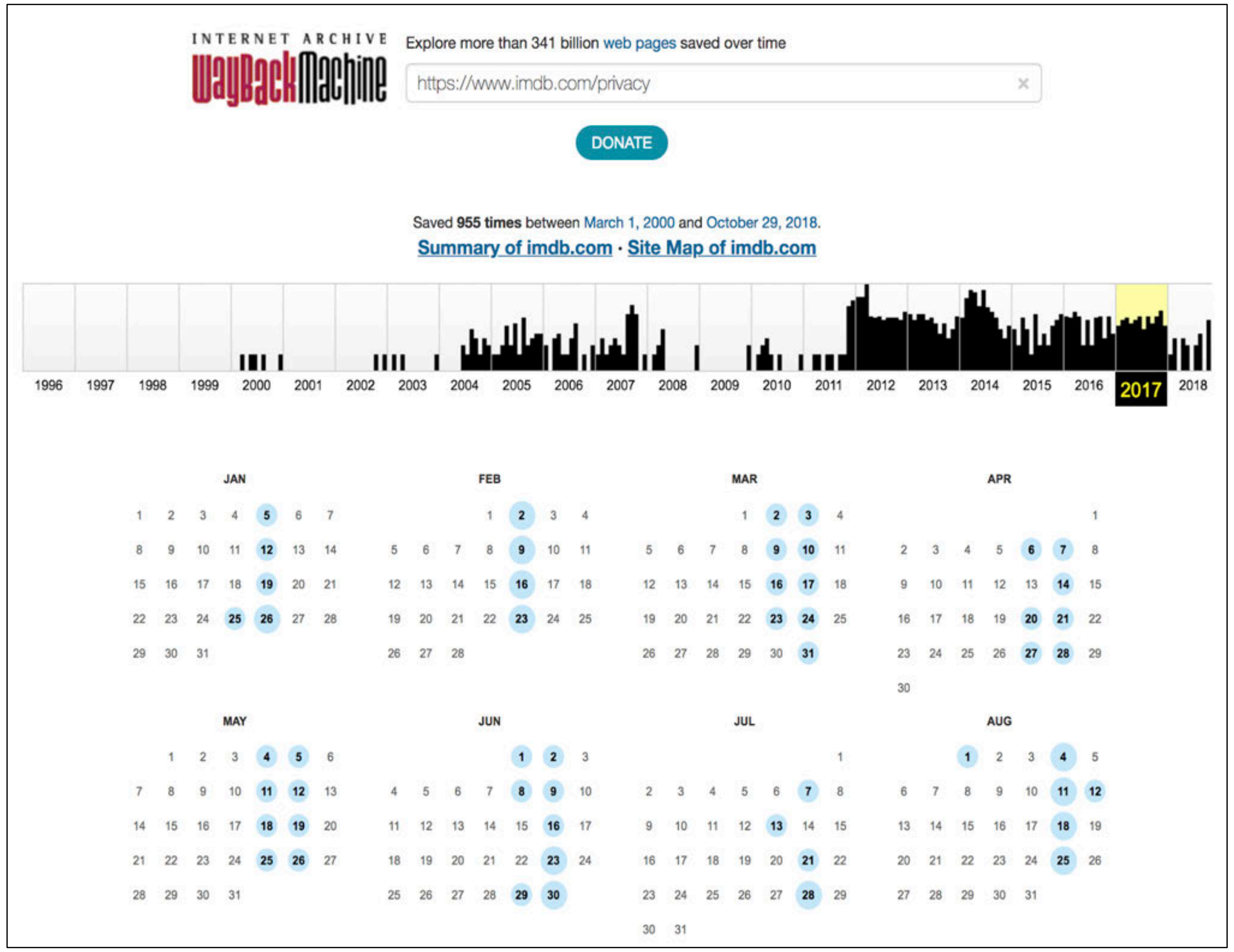
Post-GDPR Google Privacy Policy

Dataset Creation



Total: 11,440 urls (8,877 are unique) from top 500 websites in 25 Alexa categories

Wayback Machine for Pre-GDPR Sites

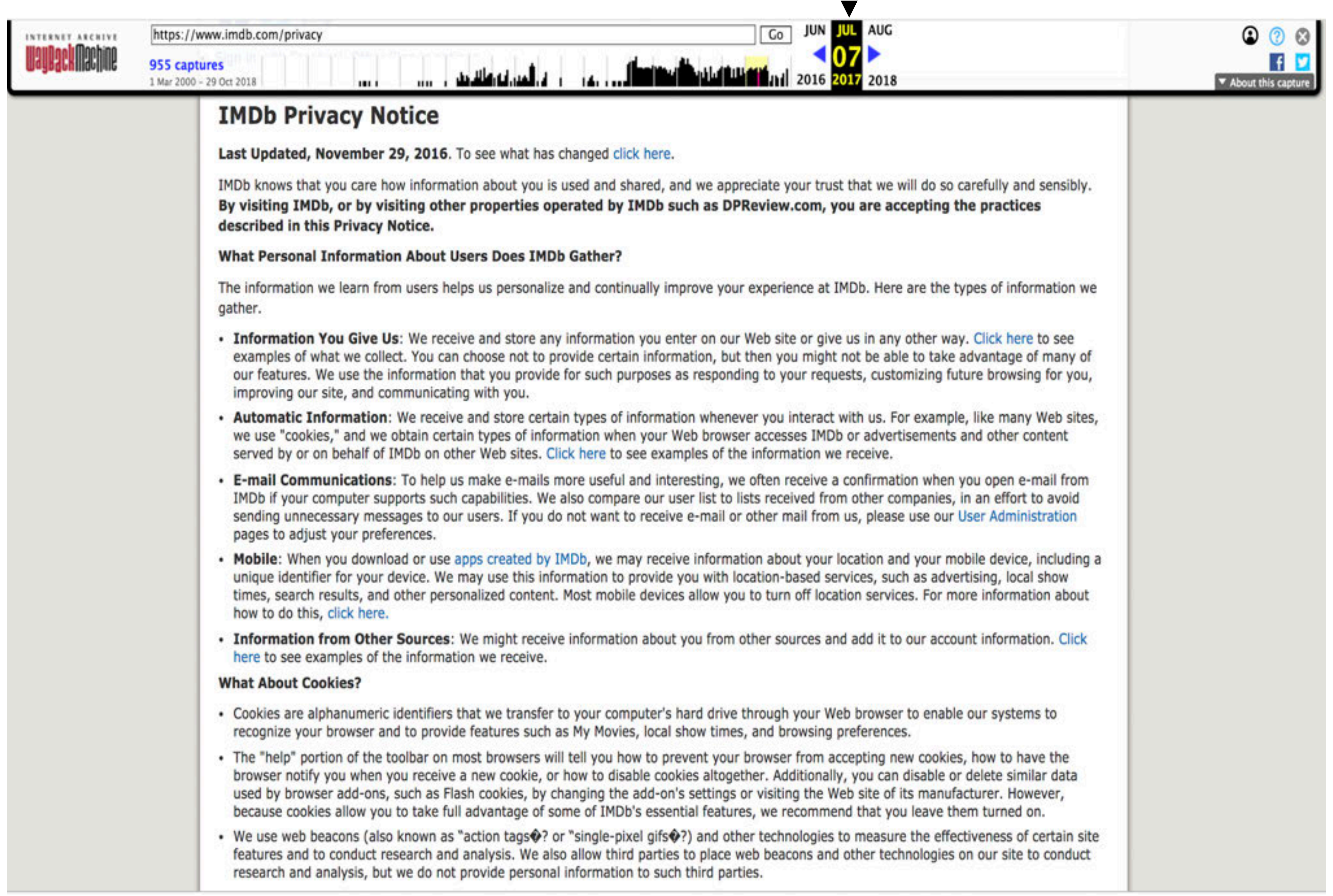


Explore more than 341 billion web pages saved over time

Saved 955 times between March 1, 2000 and October 29, 2018.

Summary of imdb.com · Site Map of imdb.com

Focus on July 2017



IMDb Privacy Notice

Last Updated, November 29, 2016. To see what has changed click here.

IMDb knows that you care how information about you is used and shared, and we appreciate your trust that we will do so carefully and sensibly. **By visiting IMDb, or by visiting other properties operated by IMDb such as DPReview.com, you are accepting the practices described in this Privacy Notice.**

What Personal Information About Users Does IMDb Gather?

The information we learn from users helps us personalize and continually improve your experience at IMDb. Here are the types of information we gather.

- **Information You Give Us**: We receive and store any information you enter on our Web site or give us in any other way. Click here to see examples of what we collect. You can choose not to provide certain information, but then you might not be able to take advantage of many of our features. We use the information that you provide for such purposes as responding to your requests, customizing future browsing for you, improving our site, and communicating with you.
- **Automatic Information**: We receive and store certain types of information whenever you interact with us. For example, like many Web sites, we use "cookies," and we obtain certain types of information when your Web browser accesses IMDb or advertisements and other content served by or on behalf of IMDb on other Web sites. Click here to see examples of the information we receive.
- **E-mail Communications**: To help us make e-mails more useful and interesting, we often receive a confirmation when you open e-mail from IMDb if your computer supports such capabilities. We also compare our user list to lists received from other companies, in an effort to avoid sending unnecessary messages to our users. If you do not want to receive e-mail or other mail from us, please use our User Administration pages to adjust your preferences.
- **Mobile**: When you download or use apps created by IMDb, we may receive information about your location and your mobile device, including a unique identifier for your device. We may use this information to provide you with location-based services, such as advertising, local show times, search results, and other personalized content. Most mobile devices allow you to turn off location services. For more information about how to do this, click here.
- **Information from Other Sources**: We might receive information about you from other sources and add it to our account information. Click here to see examples of the information we receive.

What About Cookies?

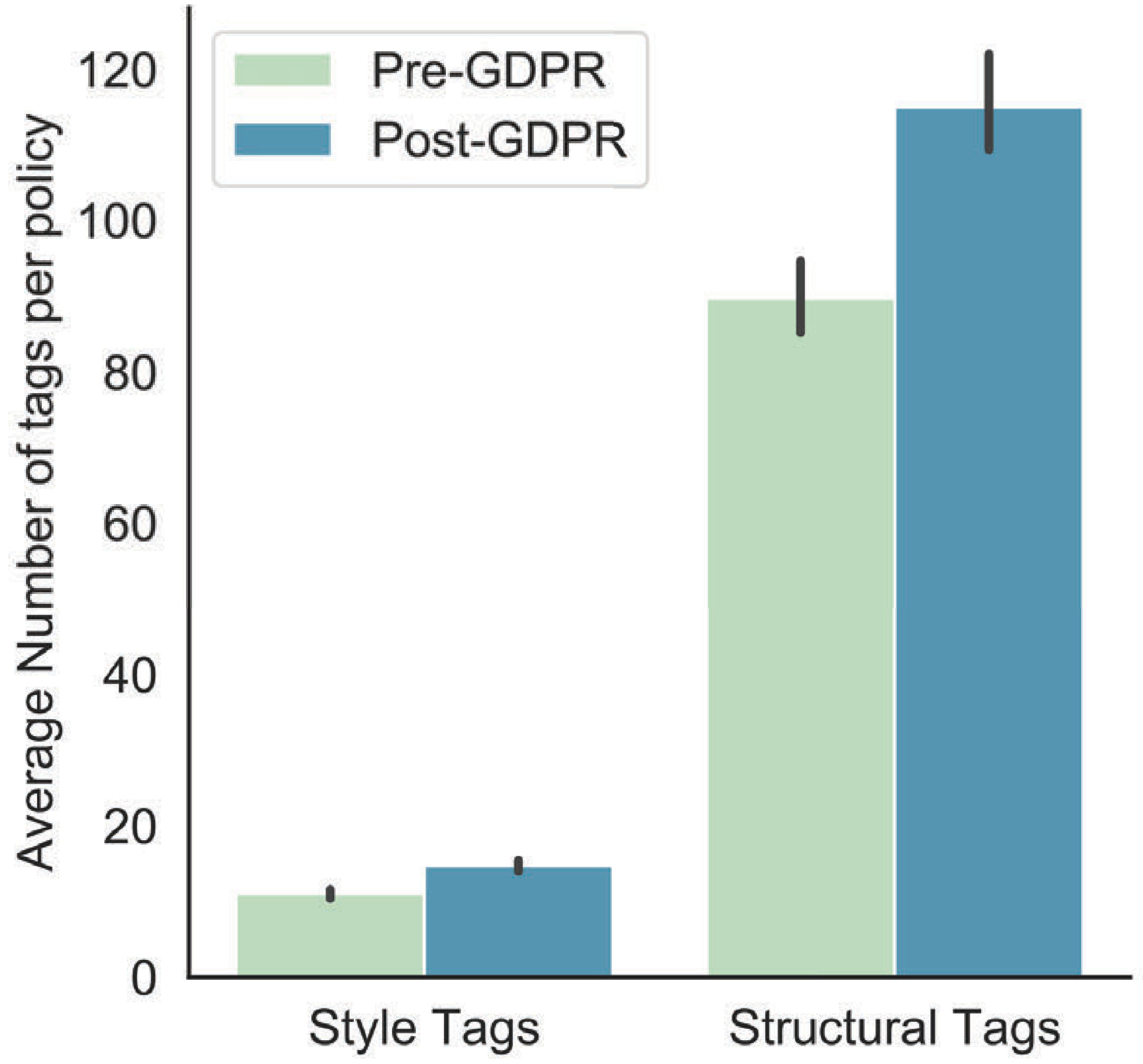
- Cookies are alphanumeric identifiers that we transfer to your computer's hard drive through your Web browser to enable our systems to recognize your browser and to provide features such as My Movies, local show times, and browsing preferences.
- The "help" portion of the toolbar on most browsers will tell you how to prevent your browser from accepting new cookies, how to have the browser notify you when you receive a new cookie, or how to disable cookies altogether. Additionally, you can disable or delete similar data used by browser add-ons, such as Flash cookies, by changing the add-on's settings or visiting the Web site of its manufacturer. However, because cookies allow you to take full advantage of some of IMDb's essential features, we recommend that you leave them turned on.
- We use web beacons (also known as "action tags�? or "single-pixel gifs�?) and other technologies to measure the effectiveness of certain site features and to conduct research and analysis. We also allow third parties to place web beacons and other technologies on our site to conduct research and analysis, but we do not provide personal information to such third parties.

https://web.archive.org/web/20170707132059/https://www.imdb.com/privacy

Analysis of 3086 privacy policies with valid pre- and post- GDPR versions

Presentation

- Increase of usage of style and structural HTML tags between pre- and post- GDPR policies
- User study (*N*=530) did not reveal significant change in the users' perception of the policies.



Structural tags: br, li, div, dl, table, a, optgroup, select, p, span, center, blockquote, dd, details, ul, ol, h1, h2, h3, h4, h5, h6.

Style tags: strong, style, sp, b, basefont, big, em, font,button, i, strike, u, img.

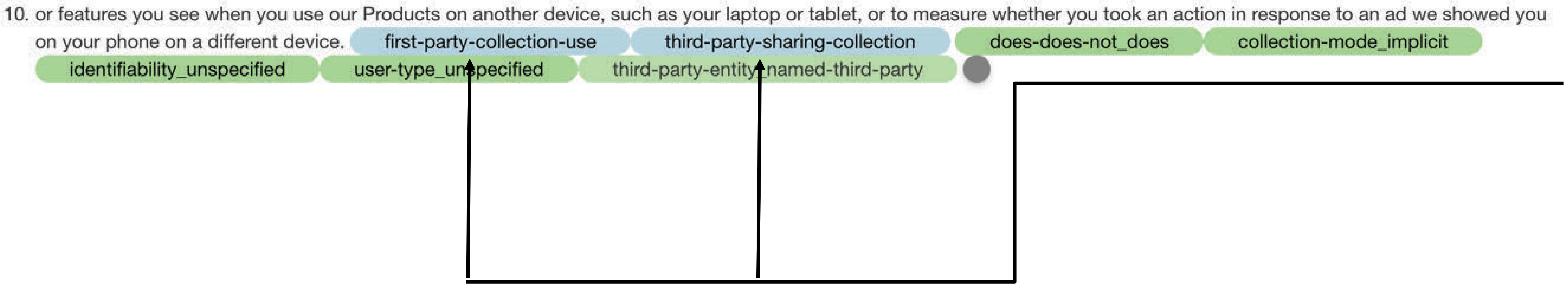
Readability

- Policies have become significantly longer
- Policies relying on distinct vocabulary

Metric	Pre-GDPR	Post-GDPR	*p*
#Words	2073.02 (1621)	2655.17 (7500)	0
#Sentences	88.62 (68.15)	118.28 (598.75)	0.006
Dale-Chall	7.42 (0.7)	7.40 (0.7)	0.26

NLP-based Analysis

Coverage: How many privacy categories where described?



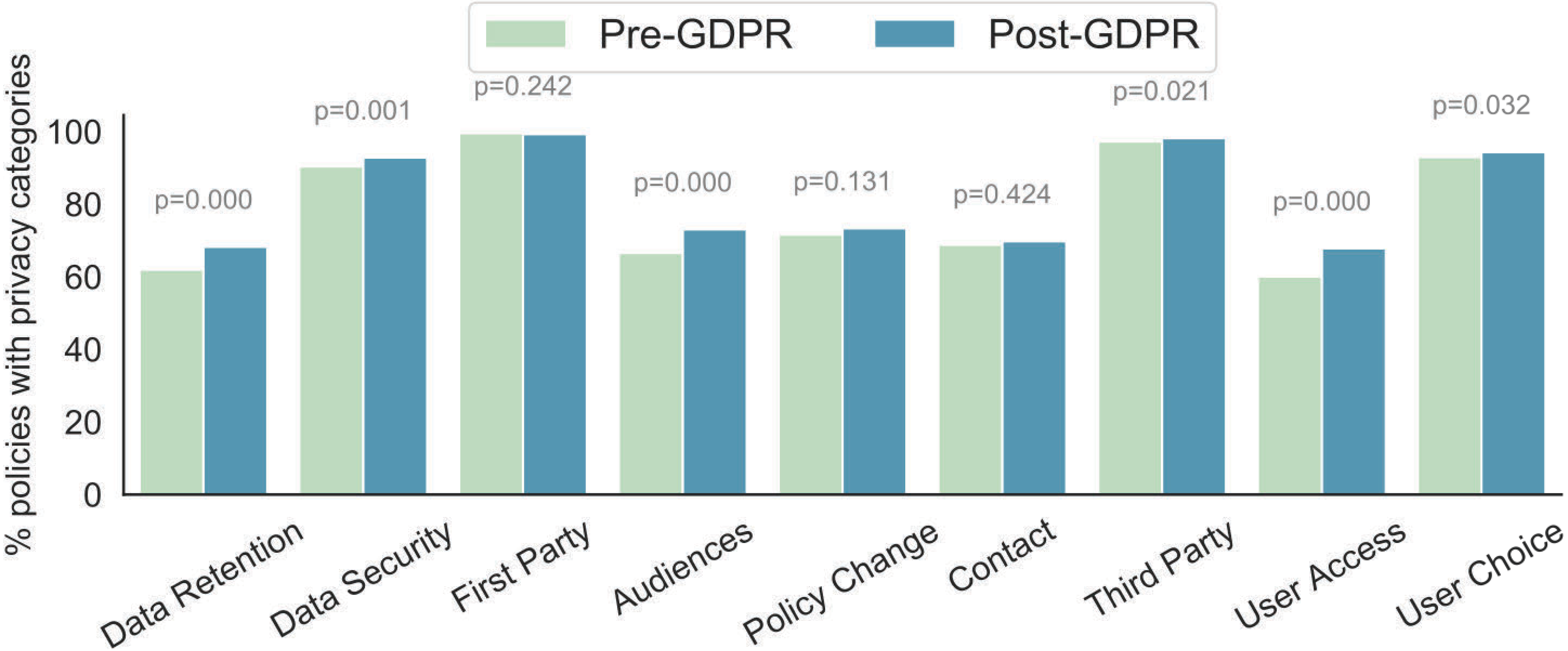
10. or features you see when you use our Products on another device, such as your laptop or tablet, or to measure whether you took an action in response to an ad we showed you on your phone on a different device. first-party-collection-use third-party-sharing-collection does-does-not_does collection-mode_implicit identifiability_unspecified user-type_unspecified third-party-entity-named-third-party

Coverage Checklist

☑ 1st party…
☑ 3rd party…
Choice…
Specific Audience…
Data Security
Access, Edit, Delete…
Data Retention…
Policy Change…
Other

A Structured Query on Polisis Output

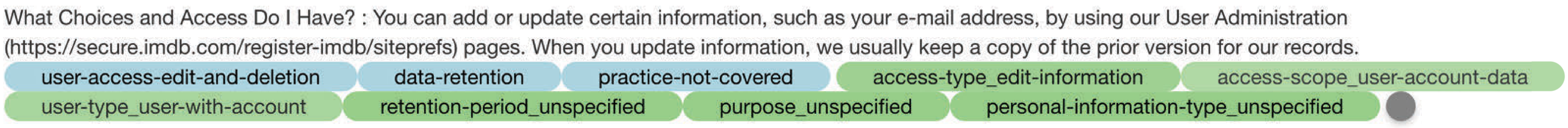
Privacy Coverage



Higher coverage of privacy practices from pre- to post- GDPR

NLP-based Analysis

Ambiguity: How specific is the policy in describing privacy practices?



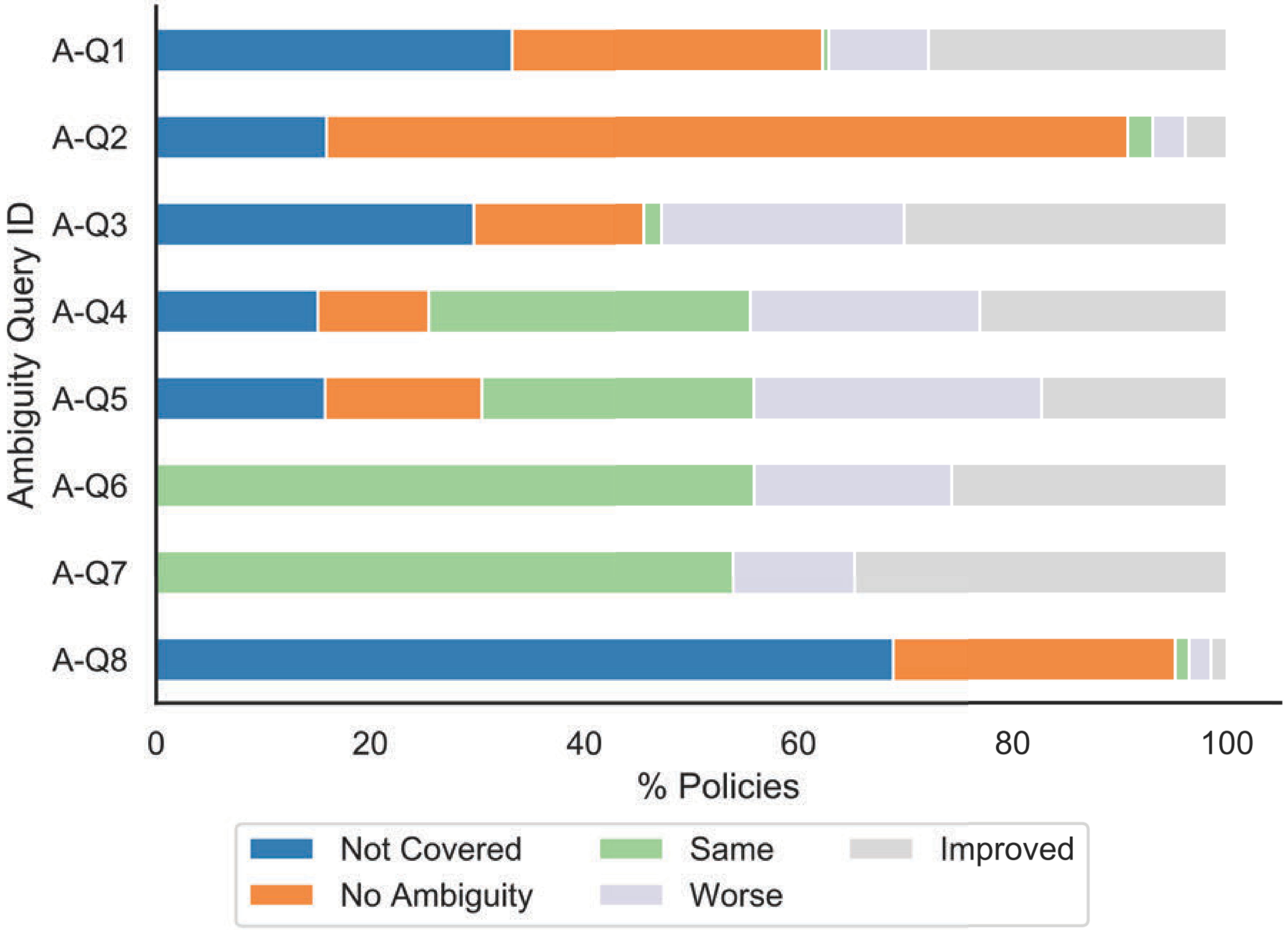
What Choices and Access Do I Have? : You can add or update certain information, such as your e-mail address, by using our User Administration (https://secure.imdb.com/register-imdb/siteprefs) pages. When you update information, we usually keep a copy of the prior version for our records.

user-access-edit-and-deletion
data-retention
practice-not-covered
access-type_edit-information
access-scope_user-account-data
user-type_user-with-account
retention-period_unspecified
purpose_unspecified
personal-information-type_unspecified

What is the purpose for keeping prior data?

A Structured Query on Polisis Output as FOL clause

Ambiguity



A-Q1: how the first party is obtaining user data.

A-Q2: how the third party is collecting user data.

A-Q3: the type of information accessed by the first party.

A-Q4: the type of information shared with the third party

A-Q5: the third party receiving user information.

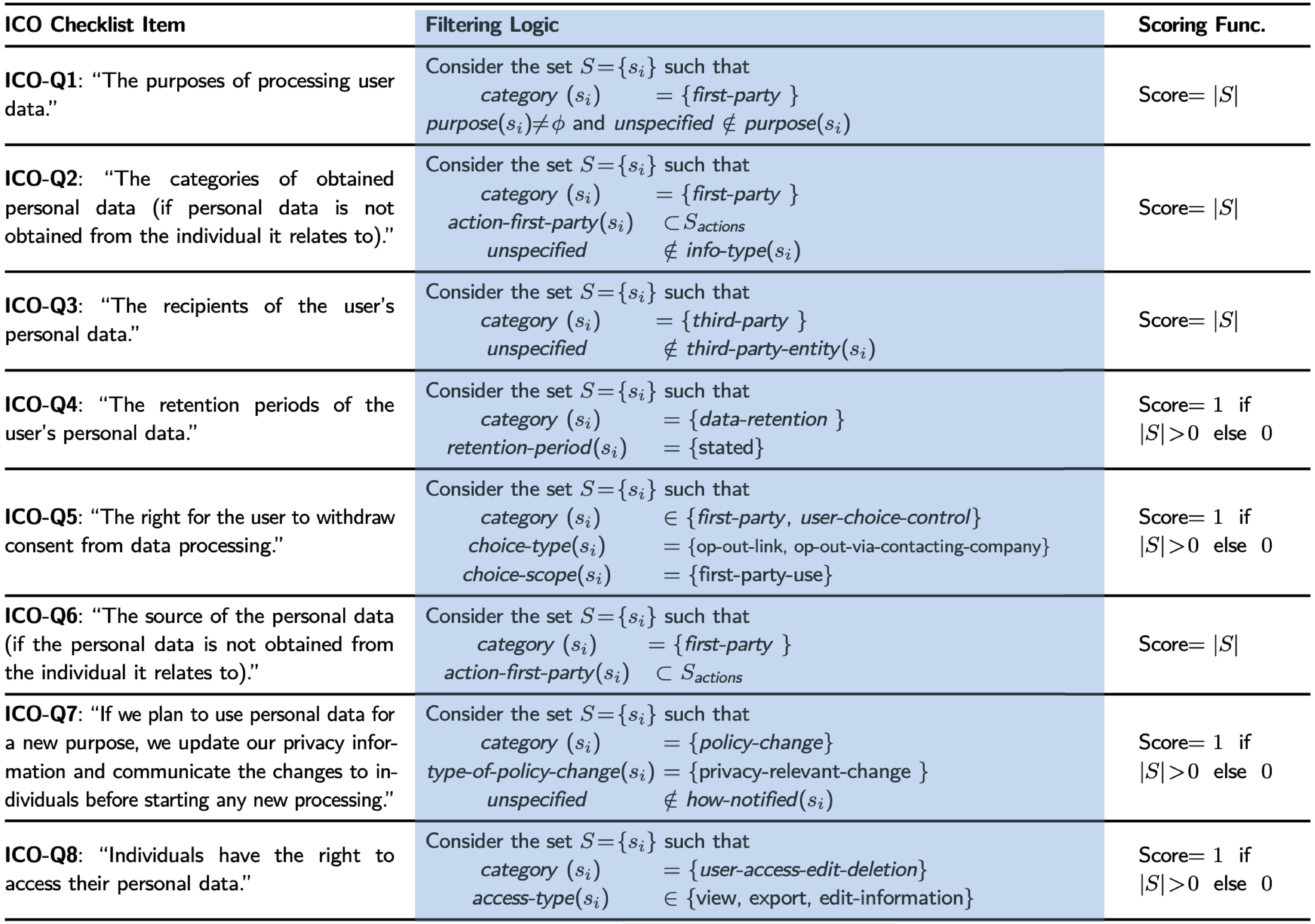
A-Q6: coverage of first party collection purposes relative to all possible purposes in our taxonomy.

A-Q7: coverage of third party sharing purposes relative to all possible purposes in our taxonomy.

A-Q8: the purpose for the data retention.

NLP-based Analysis

Compliance: How much is the privacy policy compliant with clauses from the GDPR?



ICO Checklist Item	Filtering Logic	Scoring Func.
ICO-Q1: "The purposes of processing user data."	Consider the set $S=\{s_i\}$ such that *category* (s_i) $= \{$*first-party* $\}$ *purpose*$(s_i) \neq \phi$ and *unspecified* $\notin$ *purpose*(s_i)	Score$= \lvert S \rvert$
ICO-Q2: "The categories of obtained personal data (if personal data is not obtained from the individual it relates to)."	Consider the set $S=\{s_i\}$ such that *category* (s_i) $= \{$*first-party* $\}$ *action-first-party*(s_i) $\subset S_{actions}$ *unspecified* $\notin$ *info-type*(s_i)	Score$= \lvert S \rvert$
ICO-Q3: "The recipients of the user's personal data."	Consider the set $S=\{s_i\}$ such that *category* (s_i) $= \{$*third-party* $\}$ *unspecified* $\notin$ *third-party-entity*(s_i)	Score$= \lvert S \rvert$
ICO-Q4: "The retention periods of the user's personal data."	Consider the set $S=\{s_i\}$ such that *category* (s_i) $= \{$*data-retention* $\}$ *retention-period*(s_i) $= \{$stated$\}$	Score$=1$ if $\lvert S \rvert > 0$ else 0
ICO-Q5: "The right for the user to withdraw consent from data processing."	Consider the set $S=\{s_i\}$ such that *category* (s_i) $\in \{$*first-party*, *user-choice-control*$\}$ *choice-type*(s_i) $= \{$op-out-link, op-out-via-contacting-company$\}$ *choice-scope*(s_i) $= \{$first-party-use$\}$	Score$=1$ if $\lvert S \rvert > 0$ else 0
ICO-Q6: "The source of the personal data (if the personal data is not obtained from the individual it relates to)."	Consider the set $S=\{s_i\}$ such that *category* (s_i) $= \{$*first-party* $\}$ *action-first-party*(s_i) $\subset S_{actions}$	Score$= \lvert S \rvert$
ICO-Q7: "If we plan to use personal data for a new purpose, we update our privacy information and communicate the changes to individuals before starting any new processing."	Consider the set $S=\{s_i\}$ such that *category* (s_i) $= \{$*policy-change*$\}$ *type-of-policy-change*(s_i) $= \{$privacy-relevant-change $\}$ *unspecified* $\notin$ *how-notified*(s_i)	Score$=1$ if $\lvert S \rvert > 0$ else 0
ICO-Q8: "Individuals have the right to access their personal data."	Consider the set $S=\{s_i\}$ such that *category* (s_i) $= \{$*user-access-edit-deletion*$\}$ *access-type*(s_i) $\in \{$view, export, edit-information$\}$	Score$=1$ if $\lvert S \rvert > 0$ else 0

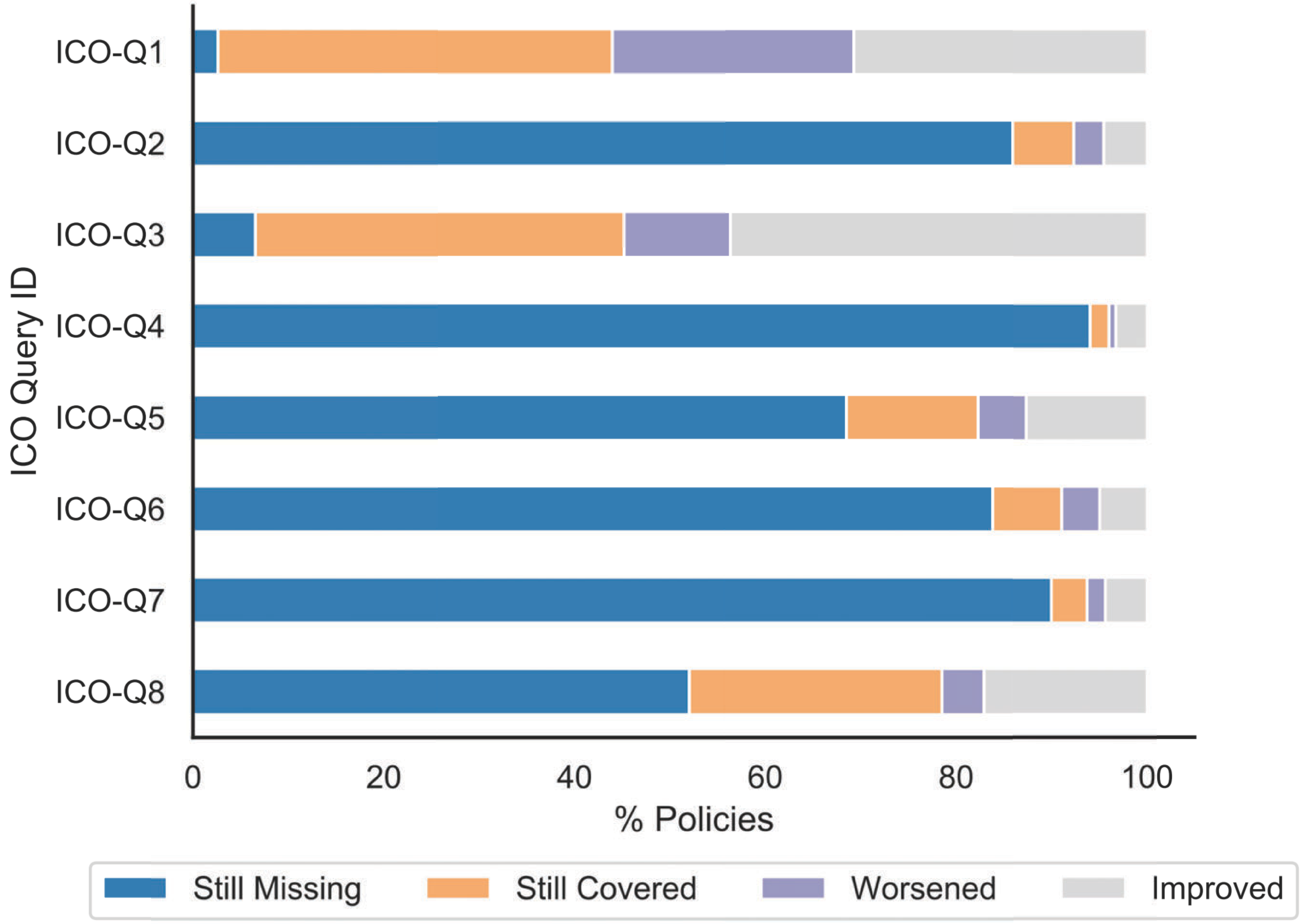
FOL clauses

Compliance

Requirements still missing:

trulia.com (data retention - Q4):
Pre-GDPR: *"We will retain your Personal Information for as long as you have an active account, as needed to provide you with the Services, to comply with our legal, financial reporting, or compliance obligations, and to enforce this privacy policy."*

Post-GDPR: *"We will retain your information for as long as necessary to fulfill the purposes outlined in this Privacy Policy unless a longer retention period is required or permitted by law."*

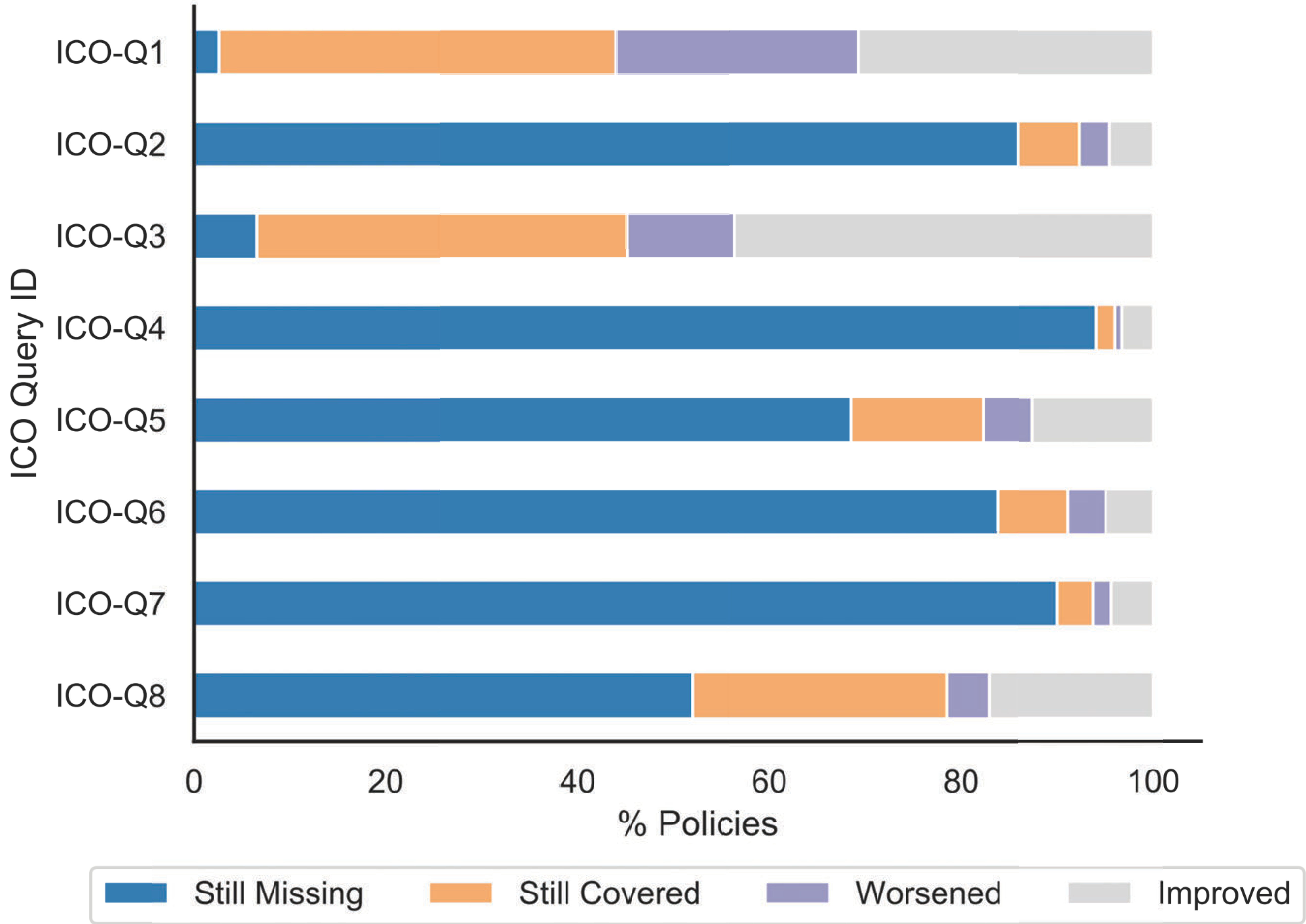


Compliance

Requirements worsened:

mckinsey.com (data retention - Q4):
Pre-GDPR: *"McKinsey would like to know whether a job candidate has previously applied. A minimal amount of personally- identifiable information will be retained for this purpose, and it will be deleted after 5 years."*

Post-GDPR: *"McKinsey retains personal data, as necessary, for the duration of the relevant business relationship. We may also retain personal data for longer than the duration of the business relationship should we need to retain it to protect ourselves against legal claims, use it for analysis or historical record-keeping, or comply with our information management policies and schedules"*



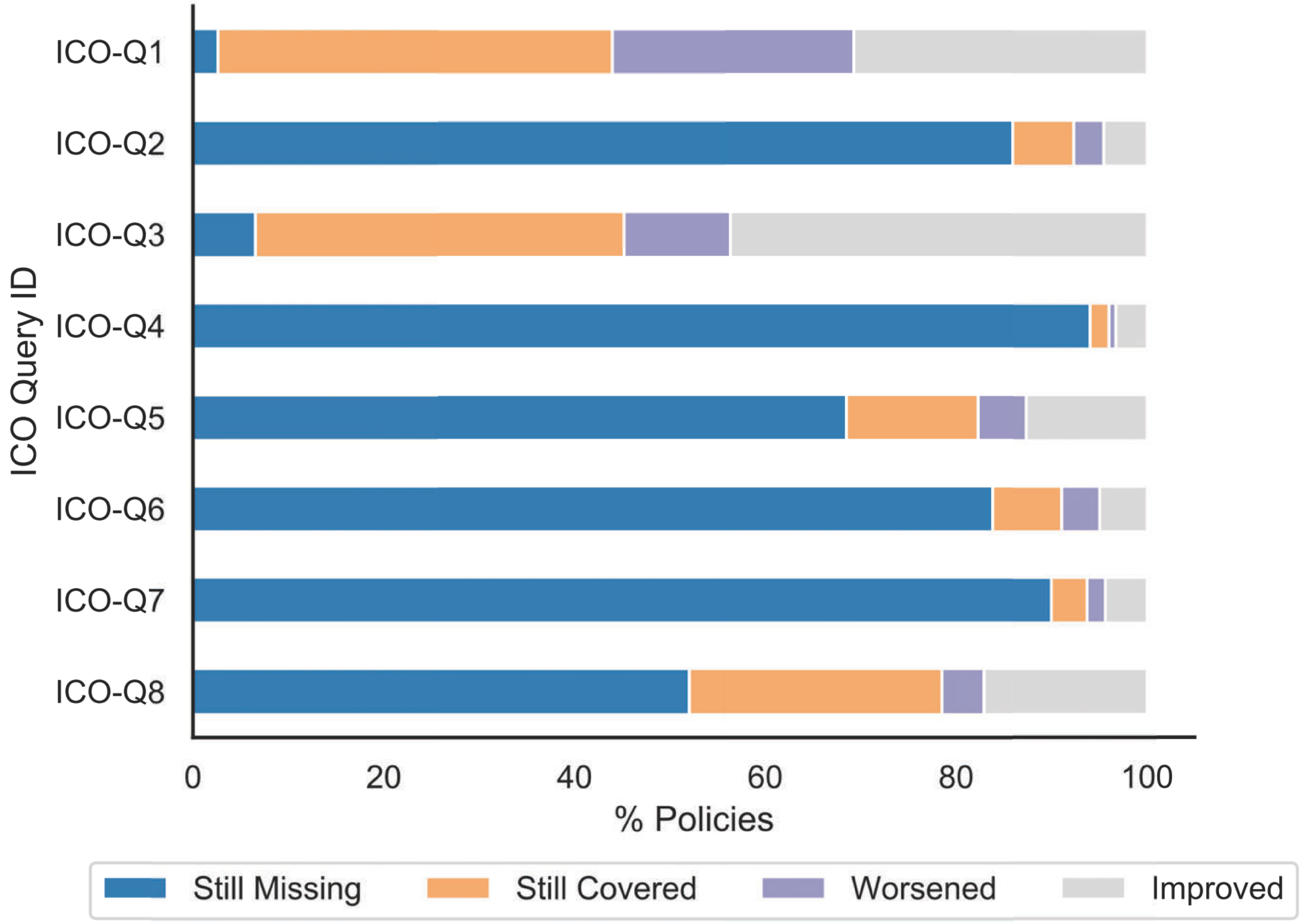
Compliance

Requirements improved:

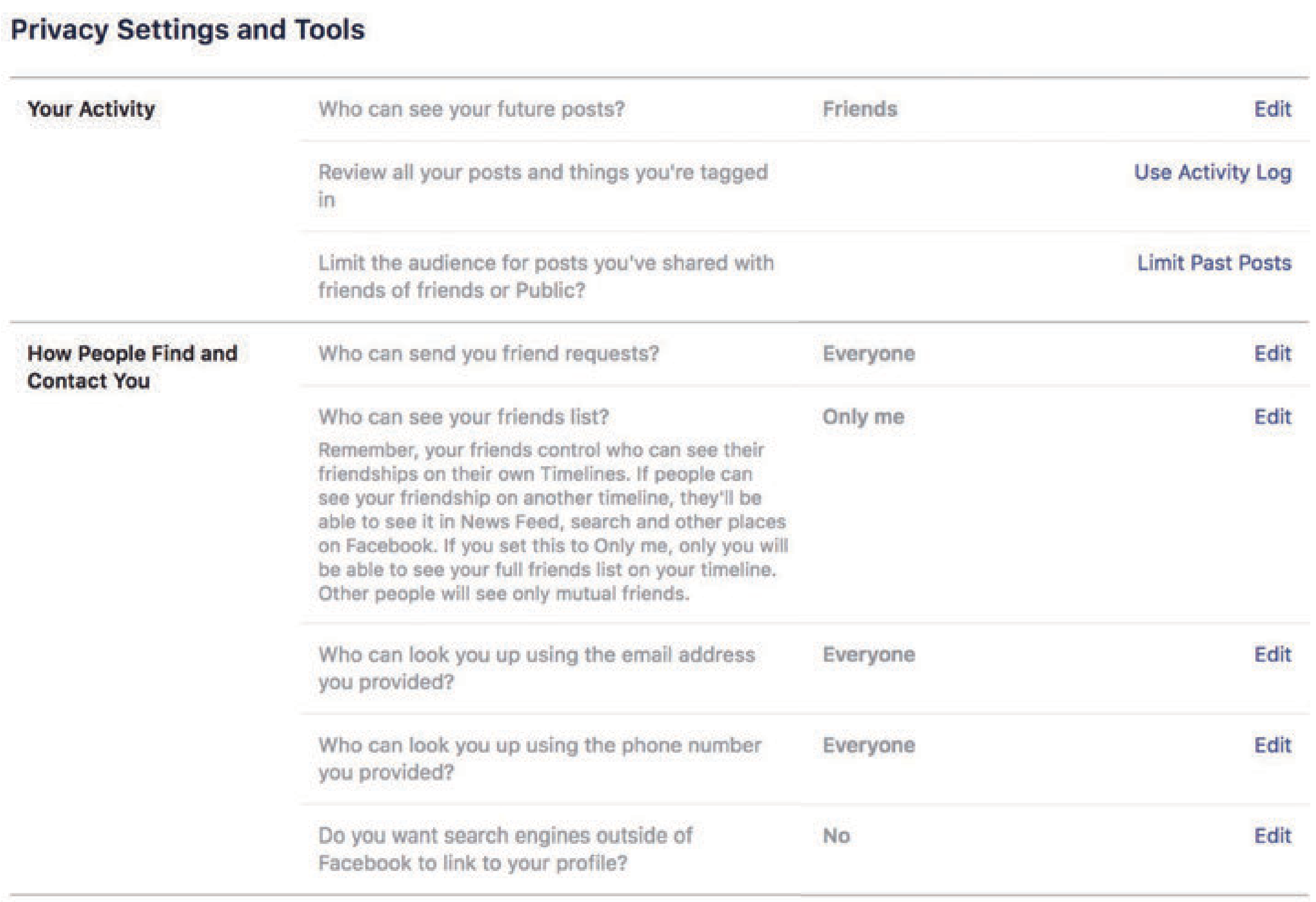
nytimes.com (privacy policy change - Q7)

Pre-GDPR: *“”*

Post-GDPR: *“We evaluate this privacy policy periodically in light of changing business practices, technology and legal requirements. As a result, it is updated from time to time. Any such changes will be posted on this page. If we make a significant or material change in the way we use or share your personal information, you will be notified via email and/or prominent notice within the NYT Services at least 30 days prior to the changes taking effect.”*

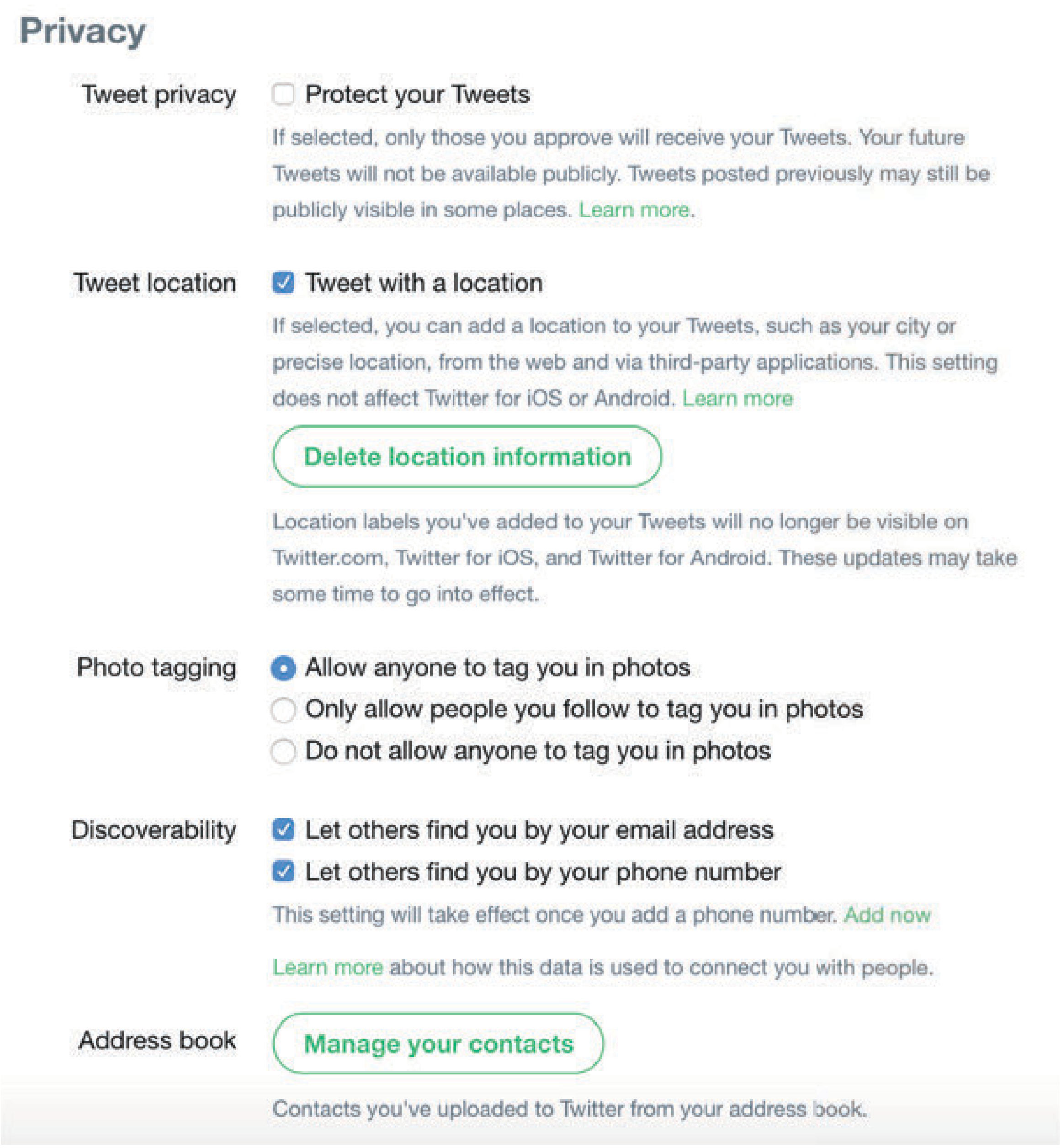


What about privacy choices?



Privacy Settings and Tools

Your Activity	Who can see your future posts?	Friends	Edit
	Review all your posts and things you're tagged in		Use Activity Log
	Limit the audience for posts you've shared with friends of friends or Public?		Limit Past Posts
How People Find and Contact You	Who can send you friend requests?	Everyone	Edit
	Who can see your friends list? Remember, your friends control who can see their friendships on their own Timelines. If people can see your friendship on another timeline, they'll be able to see it in News Feed, search and other places on Facebook. If you set this to Only me, only you will be able to see your full friends list on your timeline. Other people will see only mutual friends.	Only me	Edit
	Who can look you up using the email address you provided?	Everyone	Edit
	Who can look you up using the phone number you provided?	Everyone	Edit
	Do you want search engines outside of Facebook to link to your profile?	No	Edit



Privacy

Tweet privacy
□ Protect your Tweets
If selected, only those you approve will receive your Tweets. Your future Tweets will not be available publicly. Tweets posted previously may still be publicly visible in some places. Learn more.

Tweet location
☑ Tweet with a location
If selected, you can add a location to your Tweets, such as your city or precise location, from the web and via third-party applications. This setting does not affect Twitter for iOS or Android. Learn more

Delete location information

Location labels you've added to your Tweets will no longer be visible on Twitter.com, Twitter for iOS, and Twitter for Android. These updates may take some time to go into effect.

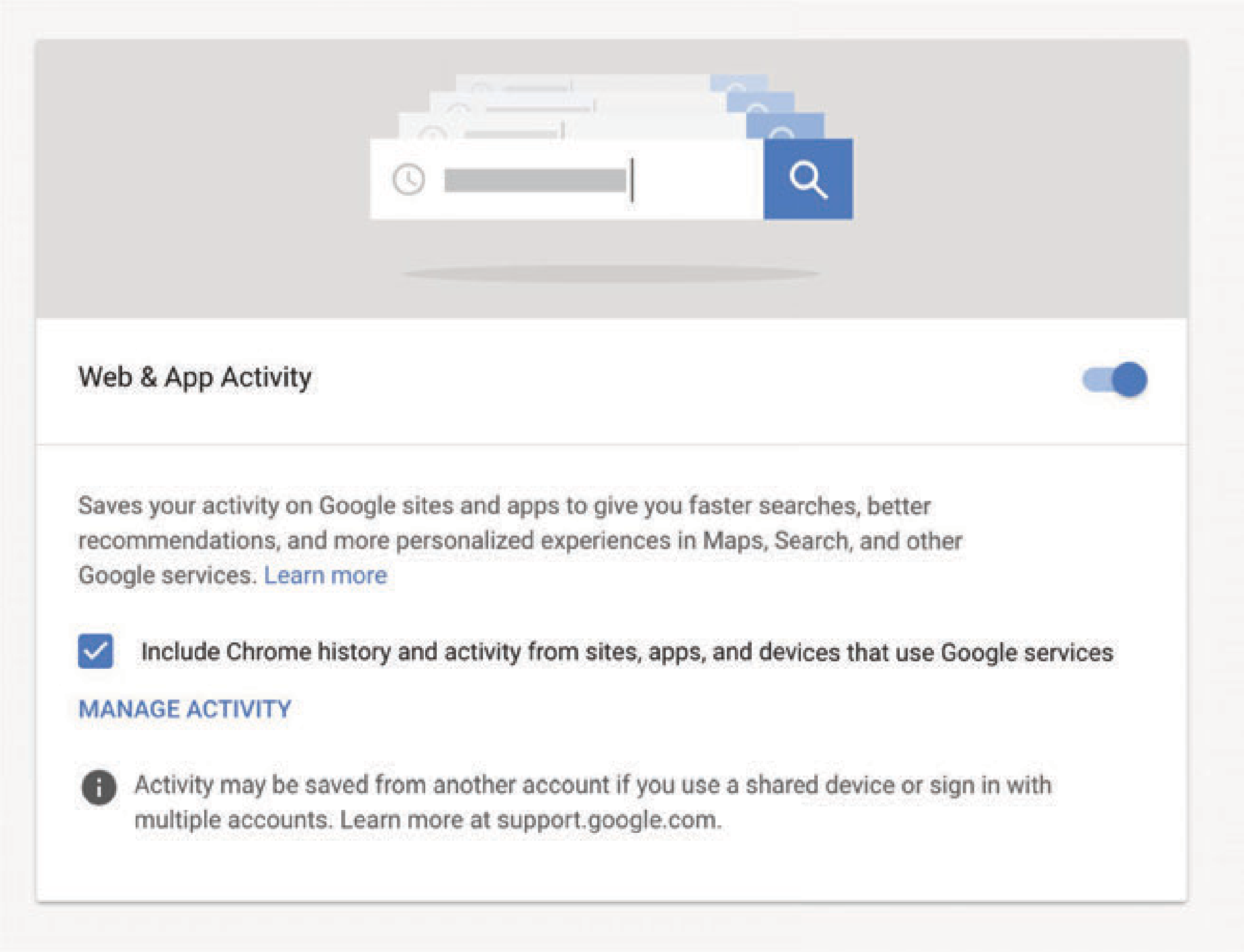
Photo tagging
● Allow anyone to tag you in photos
○ Only allow people you follow to tag you in photos
○ Do not allow anyone to tag you in photos

Discoverability
☑ Let others find you by your email address
☑ Let others find you by your phone number
This setting will take effect once you add a phone number. Add now

Learn more about how this data is used to connect you with people.

Address book
Manage your contacts

Contacts you've uploaded to Twitter from your address book.



Web & App Activity

Saves your activity on Google sites and apps to give you faster searches, better recommendations, and more personalized experiences in Maps, Search, and other Google services. Learn more

☑ Include Chrome history and activity from sites, apps, and devices that use Google services

MANAGE ACTIVITY

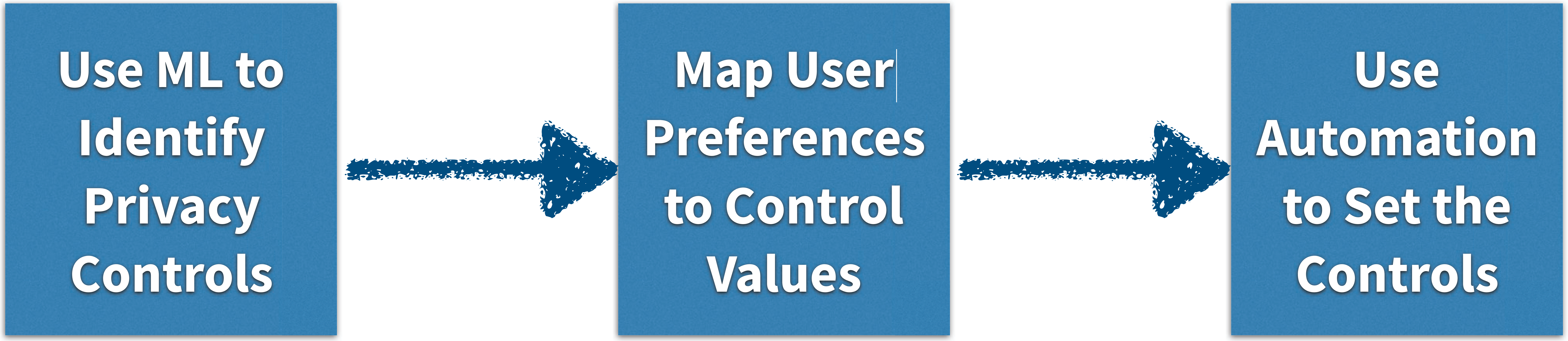
Activity may be saved from another account if you use a shared device or sign in with multiple accounts. Learn more at support.google.com.

Facebook

Twitter

Google

Users have to navigate complex and heterogenous privacy settings menus



Use ML to Identify Privacy Controls

Map User Preferences to Control Values

Use Automation to Set the Controls

Next Step: automatically understand and apply user privacy preferences

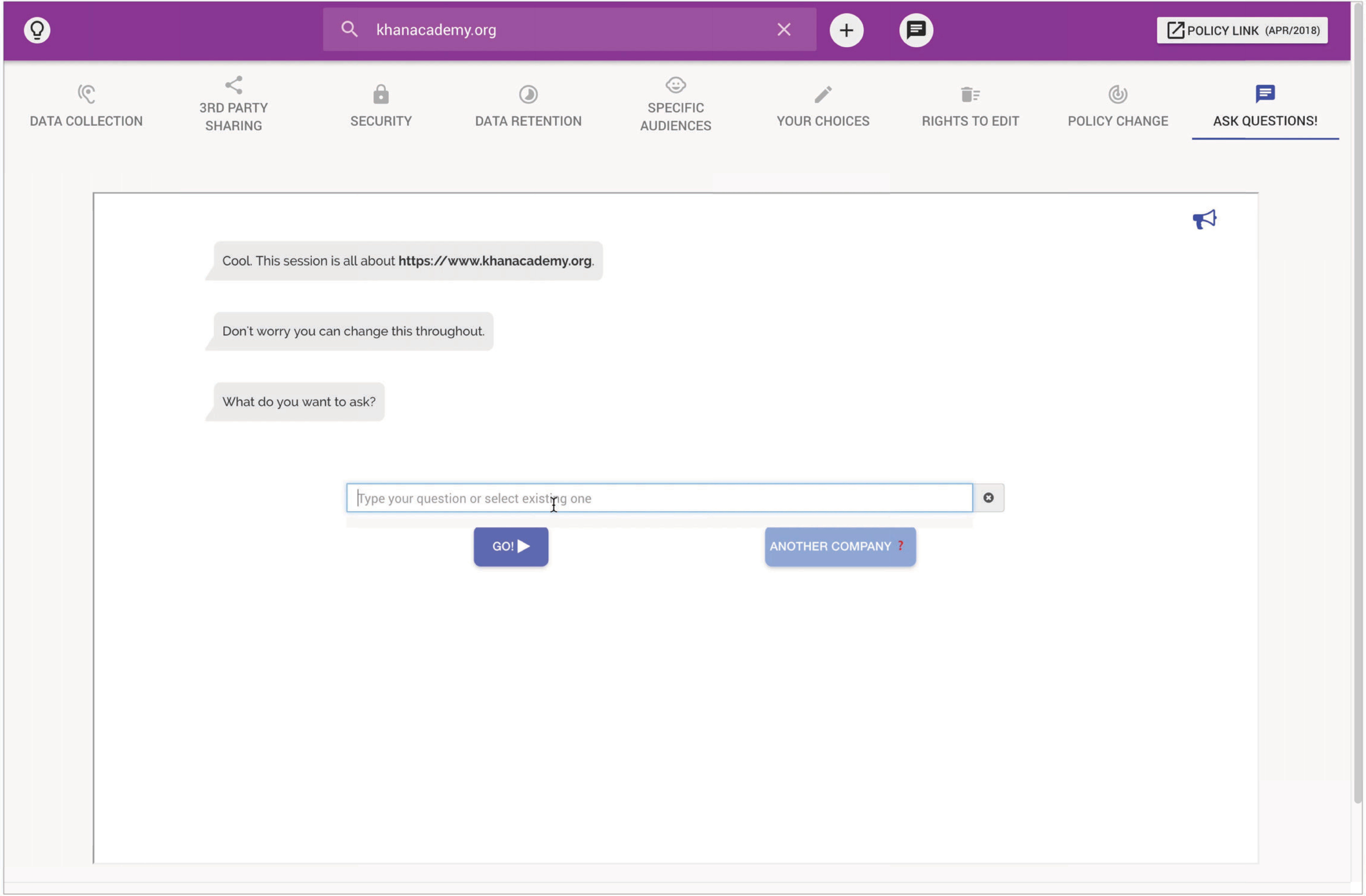
Other Applications of Polisis

tunein.com
POLICY LINK (DOWNLOADED: 7/AUG/2018)
GOOD/BAD ASPECTS
DATA COLLECTION
3RD PARTY SHARING
SECURITY
DATA RETENTION
SPECIFIC AUDIENCES
YOUR CHOICES
RIGHTS TO EDIT
POLICY CHANGE
ASK QUESTIONS!

What data does the company gather for its own use? ?

Types of info they collect
Collection reasons
What options do they give?
Computer Information
Contact
Demographic
Ip Address And Device Ids
Location
Social Media Data
User Profile
Generic Personal Information
Other Data
Other Purposes
Basic Service Feature
Marketing
Additional Service Feature
Personalization Customization
Analytics Research
Advertising
Unspecified Choice

Choice Links (0)



Cool. This session is all about **https://www.khanacademy.org**.

Don't worry you can change this throughout.

What do you want to ask?

Type your question or select existing one

medium.com
POLICY LINK (DOWNLOADED: 16/AUG/2018)

What are the positive and the negative aspects of the policy?

Confidence Threshold 0.2

Good

- You can request access and deletion of personal data
- In certain conditions, data is not shared.
- Data is not shared with third parties for advertising purposes.
- The policy states that third parties do not receive personal information.
- The policy offers you clear links to control your data
- Some of the collected data is anonymized or aggregated.

Bad

- Some data might be retained indefinitely.
- Data might be shared in the case of a merger or acquisition.

COMING SOON!

Impact

Users of the app

>35,000

Minutes on our apps

>88,000

Websites analyzed

>21,000

NVIDIA · FAST COMPANY · Product Hunt · WSJ · WIRED · WORLD ECONOMIC FORUM

Take-aways

- **Polisis**:
  - Unified framework for querying privacy policies
  - Assisting users, regulators, and researchers
  - Example application: Structured querying for GDPR compliance
    - Privacy policies have become longer on average
    - Some ambiguity in describing the practices is still evident

  - **Read more at:**
    - Our polisis paper and GDPR paper
    - **WIRED**: Polisis AI Reads Privacy Policies So You Don't Have To
    - **Fast Company**: This Data Viz Tool Explains Privacy Policies You're Too Lazy to Read
    - **WSJ**: Those Privacy Policies Flooding Your Inbox? Print Them Out and They Span a Football Field

pribot.org **kassemfawaz.com**