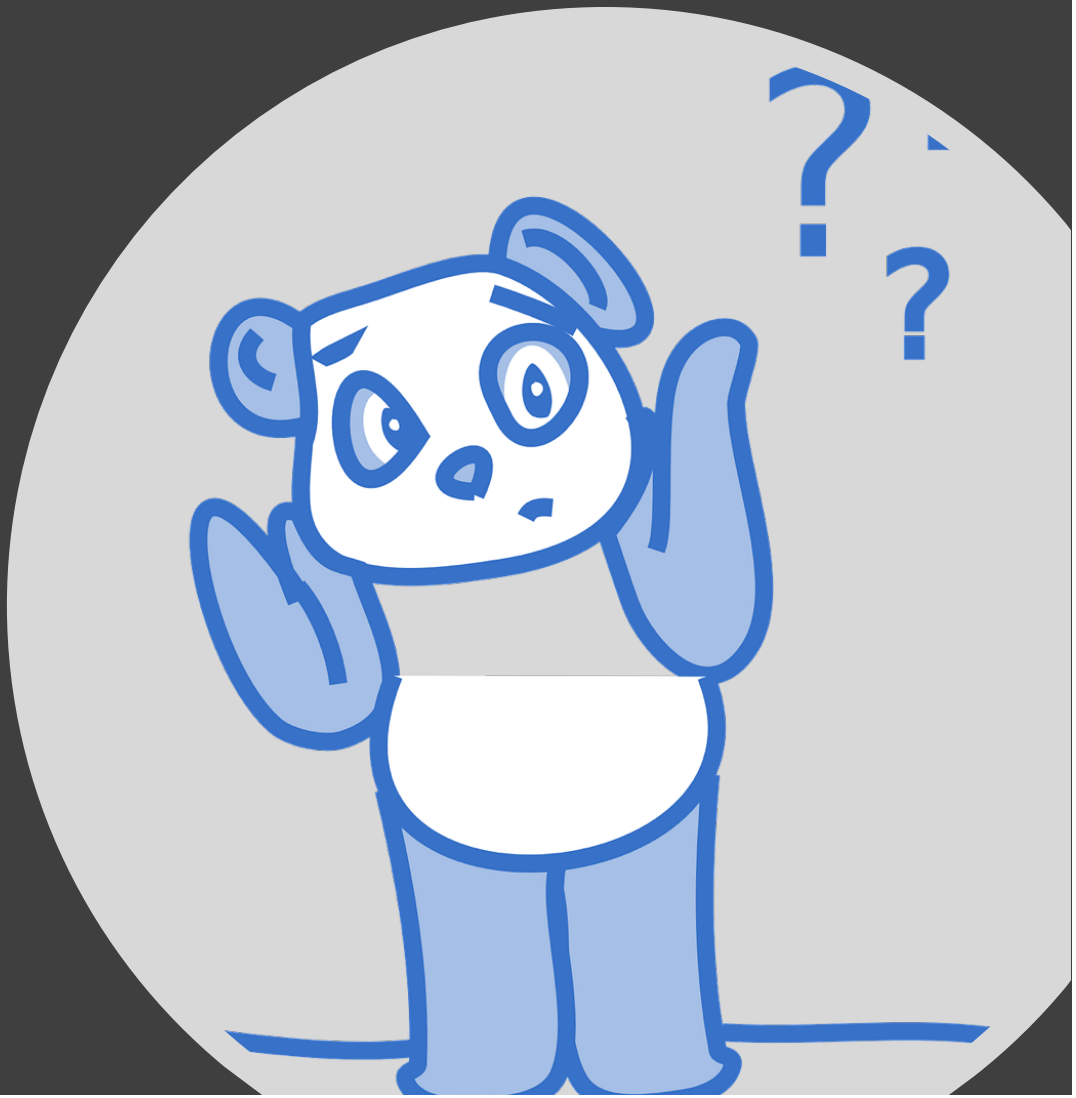
*Sebastian Zimmeck**, Peter Story*, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel Reidenberg, N. Cameron Russell, and Norman Sadeh*

MAPS: Scaling Privacy Compliance Analysis to a Million Apps

The 19th Privacy Enhancing Technologies Symposium
Royal Institute of Technology, Stockholm, Sweden
July 18, 2019

* Corresponding Authors: Sebastian Zimmeck, Department of Mathematics and Computer Science, Wesleyan University, szimmeck@wesleyan.edu, Peter Story, School of Computer Science, Carnegie Mellon University, pstory@andrew.cmu.edu, Norman Sadeh, School of Computer Science, Carnegie Mellon University, sadeh@cs.cmu.edu. The first two authors contributed equally to this study. Previously, Sebastian Zimmeck was a postdoctoral associate at Carnegie Mellon's School of Computer Science.

What does Privacy Compliance Mean?



Privacy Policies

California Online Privacy Protection Act (CalOPPA)

Section 22575 (b) Cal Bus & Prof Code

- Categories of personally identifiable information that the operator collects
- Categories of third-party persons with whom the operator may share
- Whether other parties may collect personally identifiable information

…

Privacy Compliance of Android Apps in the Google Play Store

"Google Play requires developers to provide a valid privacy policy when the app requests or handles sensitive user or device information."

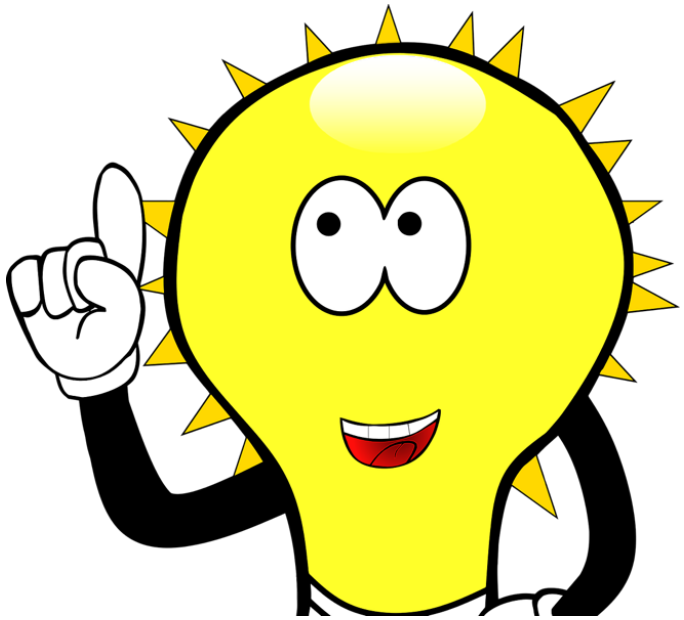
- Google Play Developer E-Mail, Feb '17 (emphasis added)

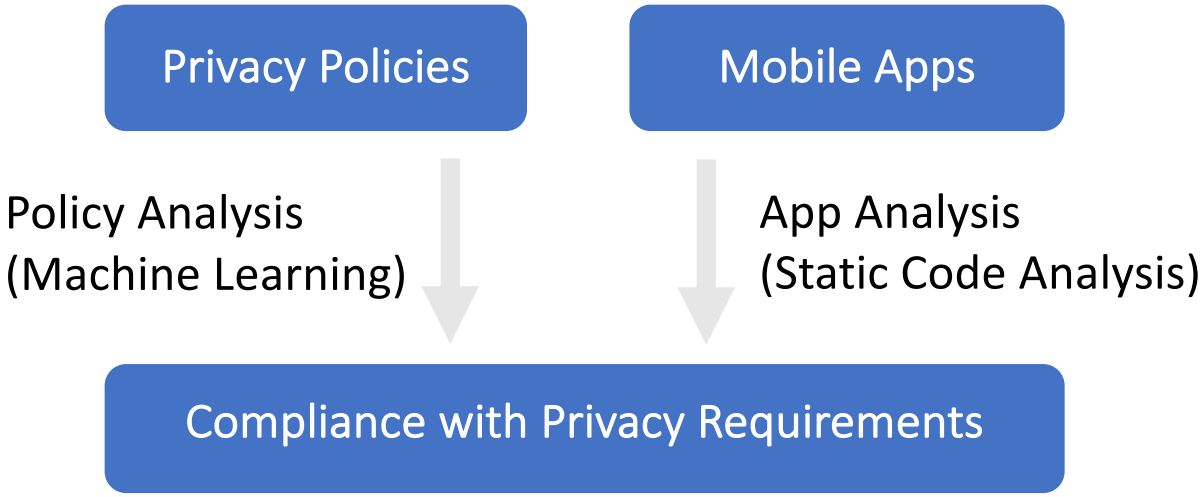
Snapchat's Location Data Disclosures

- "We do not ask for, track, or access any location-specific information [...]."
- Snapchat Android app transmitted Wi-Fi- and cell-based location data from users' devices to analytics service providers
- **Accidental discovery** by researcher who examined Snapchat's data deletion mechanism

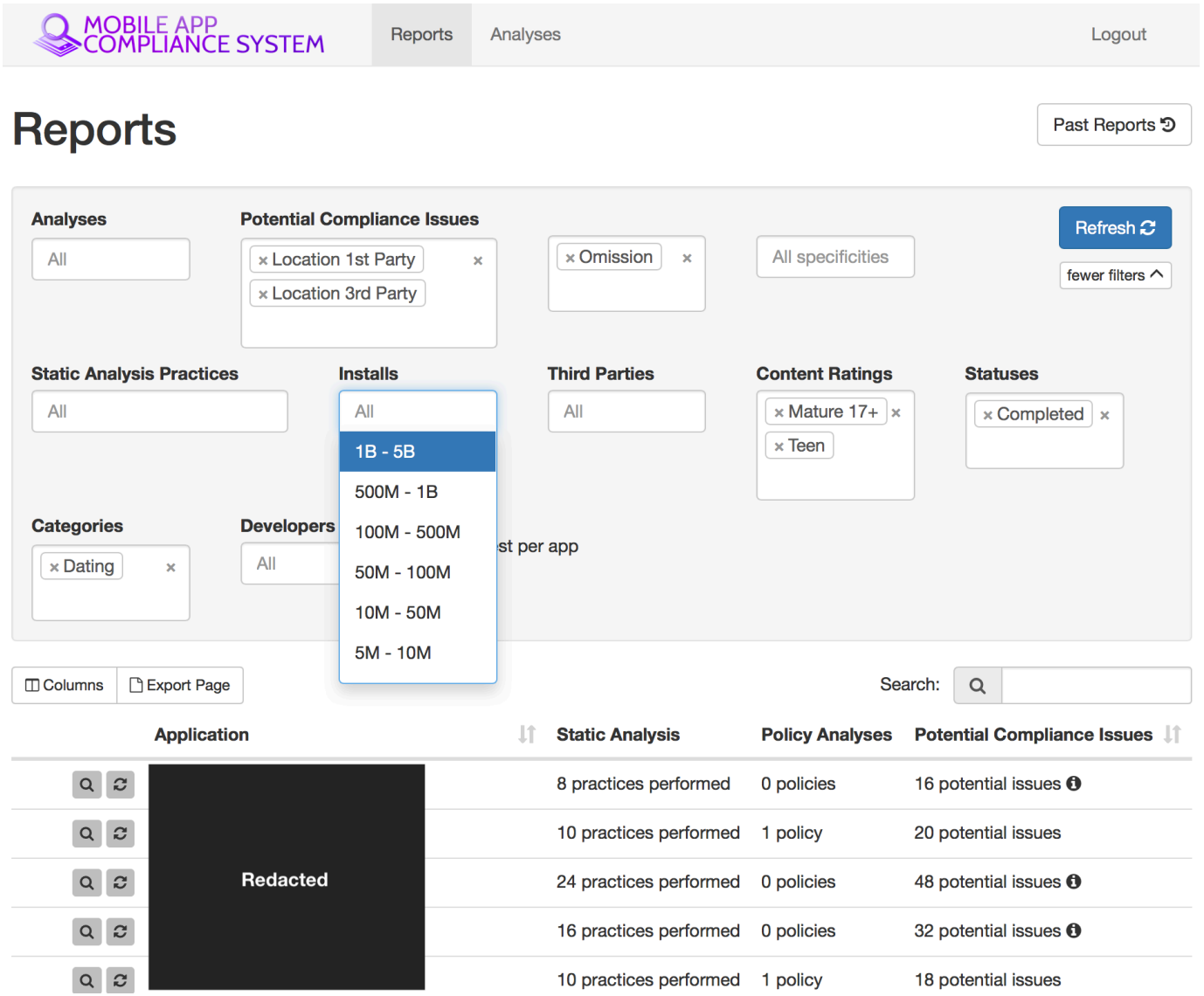
FTC, Complaint In the Matter of Snapchat, Inc. (December 31, 2014) https://www.ftc.gov/enforcement/cases-proceedings/132-3078/snapchat-inc-matter.

Here is the idea … MAPS*





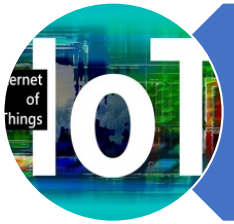
* Mobile App Privacy System



Two Case Studies

Federal Trade Commission

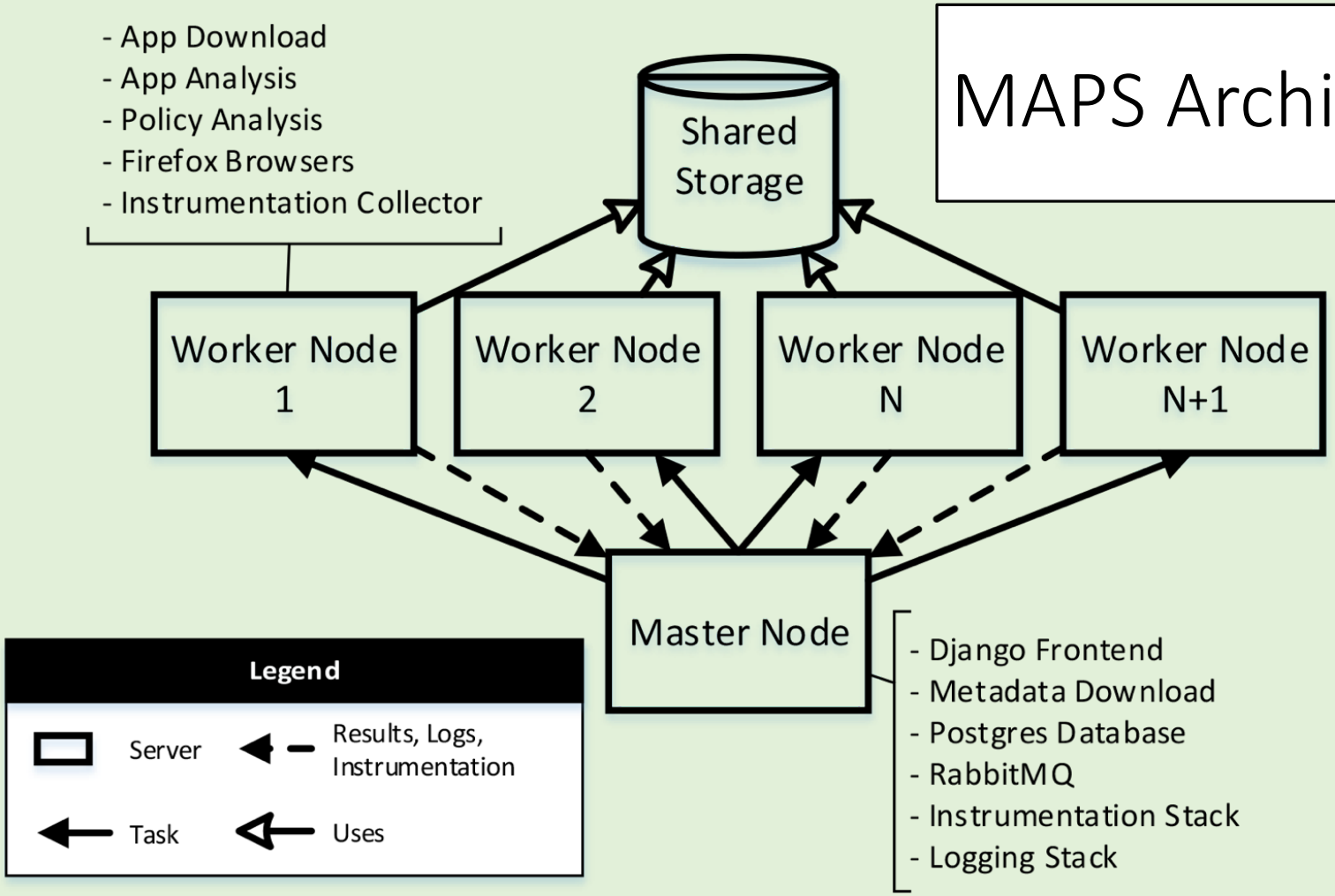
- Nine apps from Google's Designed for Families Program



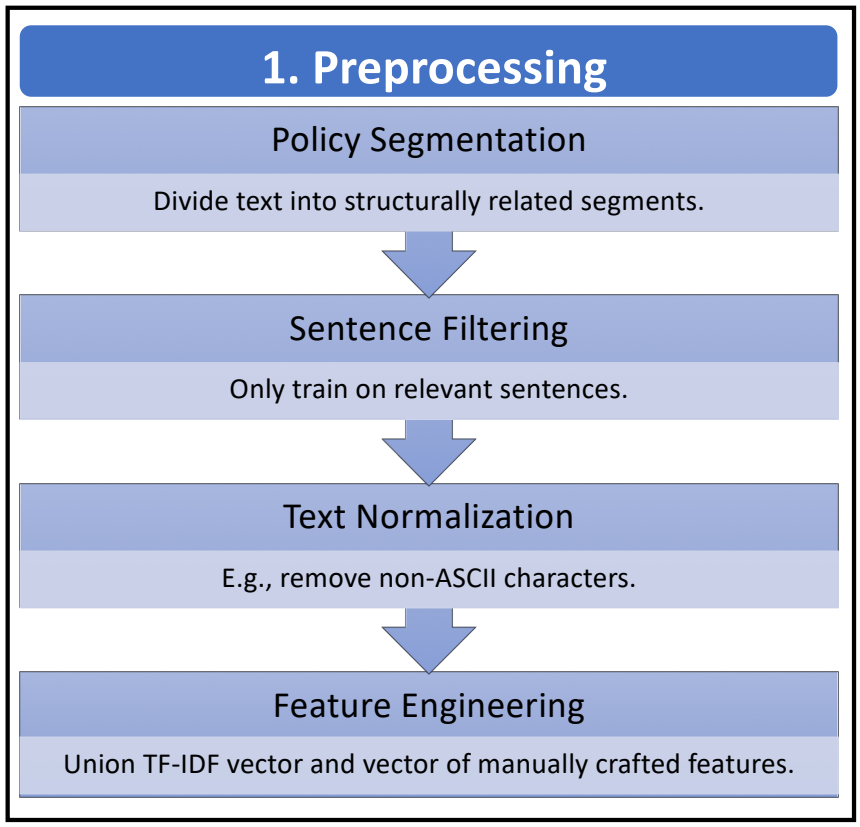
European Device Manufacturer

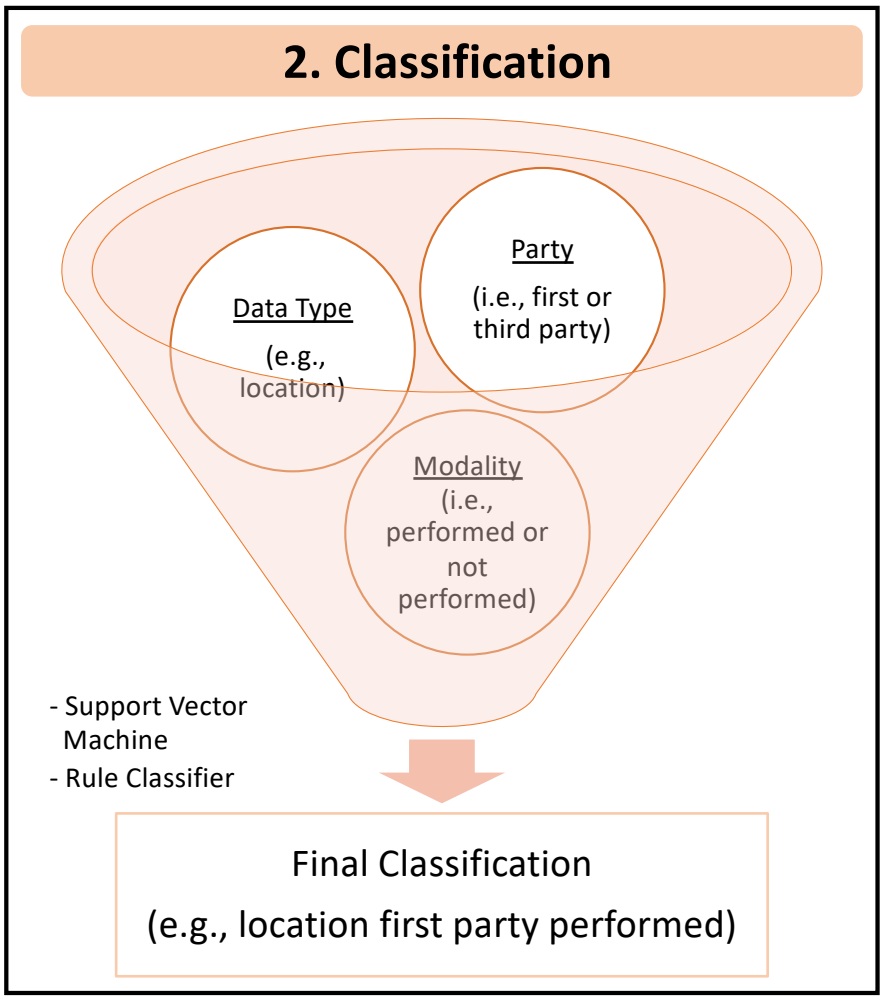
- IoT (legacy) apps

MAPS Architecture



Policy Analysis





The APP-350 Corpus

Supervised learning requires ground truth

- 350 manual expert-annotated policies
- 250 policies for training/validation
- 100 policies for testing
- 35 policies double-annotated by three experts
- Average agreement per practice: Krippendorff's Alpha = 0.78

The dataset is available at https://data.usableprivacy.org.

App Analysis

Permissions

- e.g., `ACCESS_FINE_LOCATION`

API Calls

- e.g., `android.location.LocationManager.getLastKnownLocation`

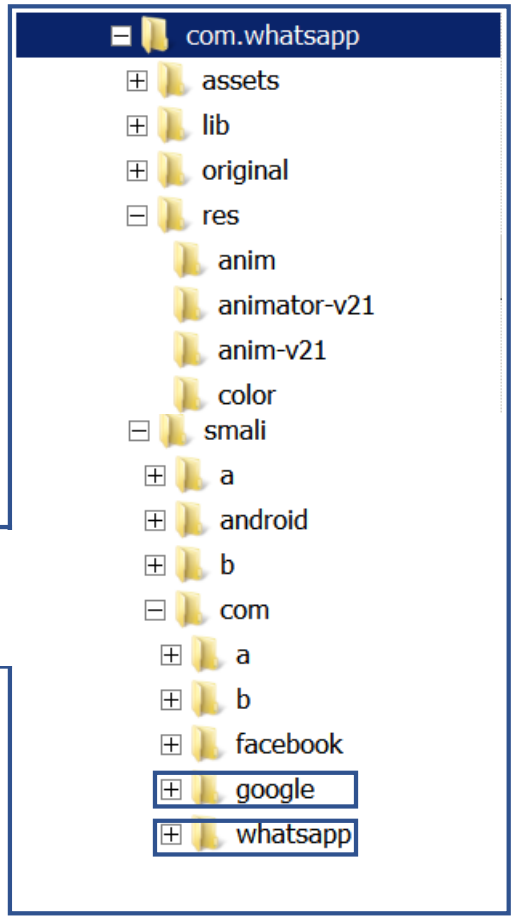
Parameter String Call Graph Analysis

- e.g., `GPS_PROVIDER`

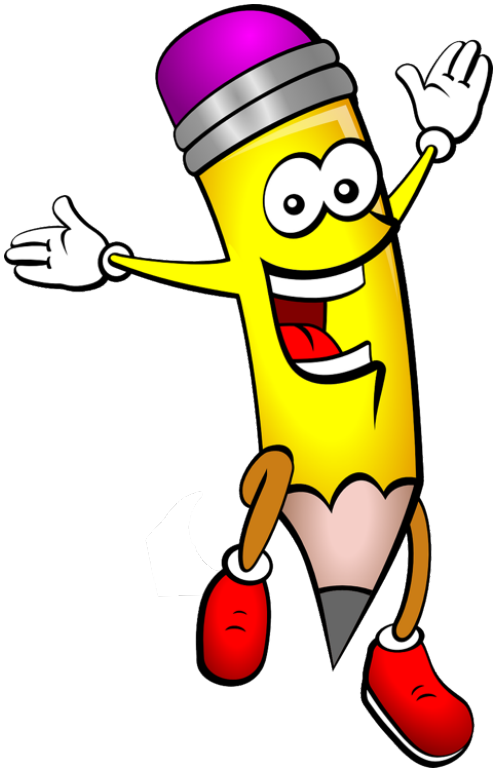
Class Structure (leverage reverse package naming convention)

- e.g., `com.whatsapp` is a first party and `com.google` is a third party

Mapping to Privacy Practice (e.g., Location GPS First Party)

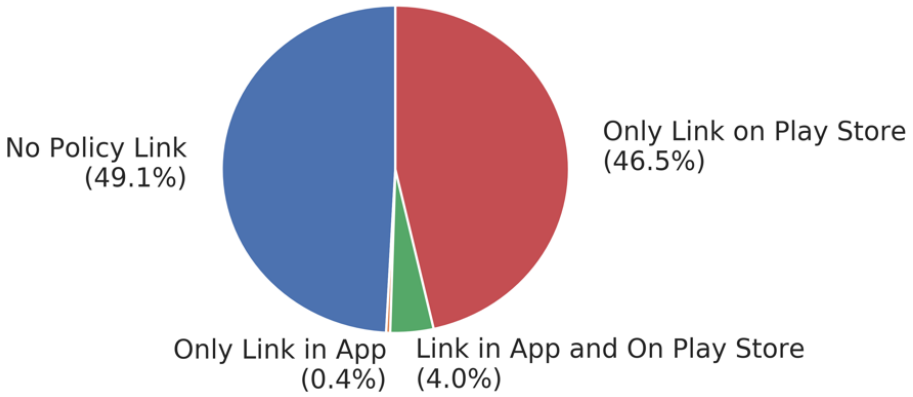


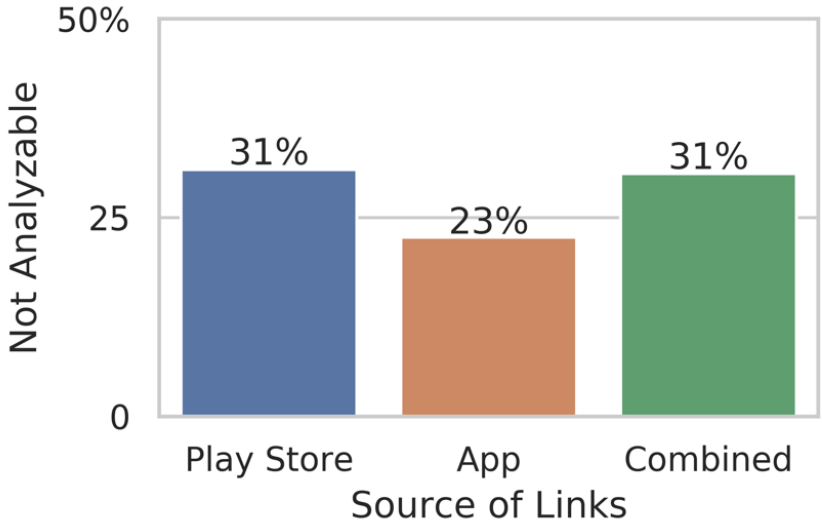
Performance Results*



- Policy Analysis
  - Negative F1 scores ranging from **78% to 100%**
  - Compared against ground truth from expert-annotated policies
- App Analysis
  - F1 scores ranging from **62% to 99%**
  - Manual dynamic analysis with custom Xposed module
  - If a practice could not be verified, we counted it against us
- Compliance Analysis
  - F1 scores ranging from **40% to 100% (mean: 71%)**
  - Statistical analysis seems to suggest we underestimate the number of potential compliance issues

* All for app/policy test set (n = 100). For the app and compliance analysis 17 apps could not be considered due to forced automatic app updates, apps' refusal to run on a rooted phone, or failures in API logging.





- *n* = 1,035,853 Android apps
- A lot of details: policy link identifier, policy crawler, policy classifier, dealing with JS, pdfs, …

How Many Apps Have Privacy Policies?

Which Practices are Disclosed in Policies?

- **Problem 1: Silence**

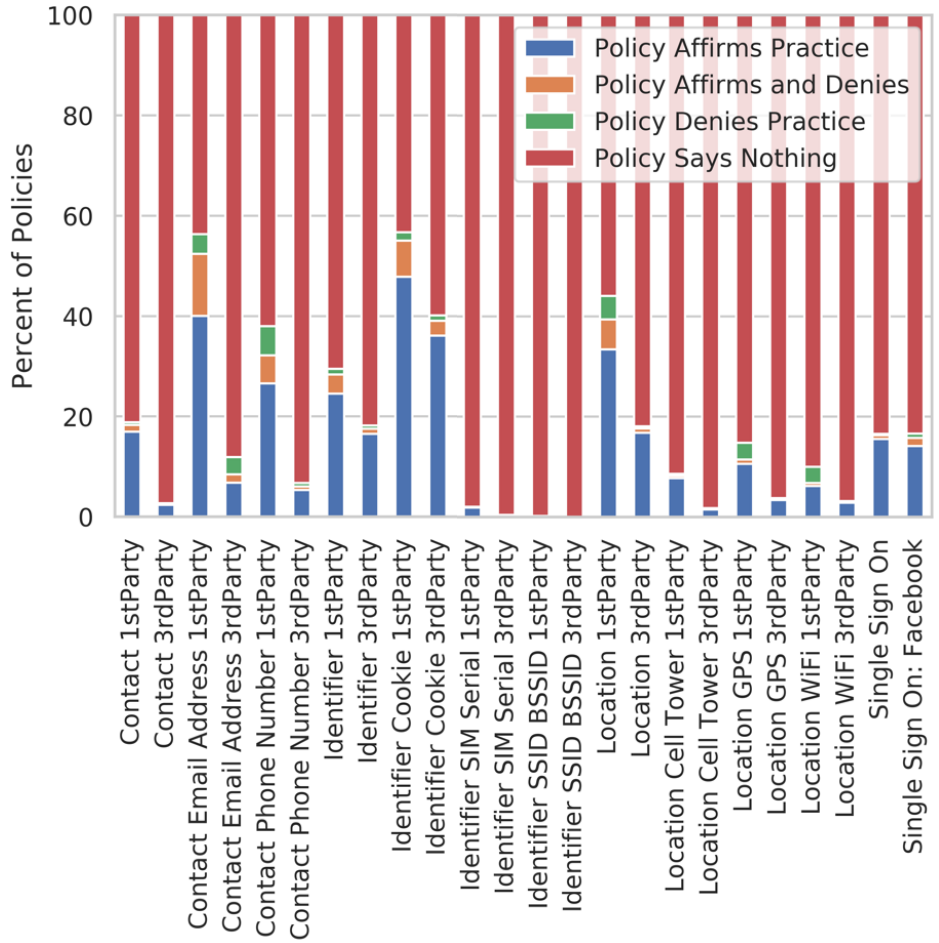
  What does silence mean? Can a service perform a practice that it simply does not mention in its privacy policy?

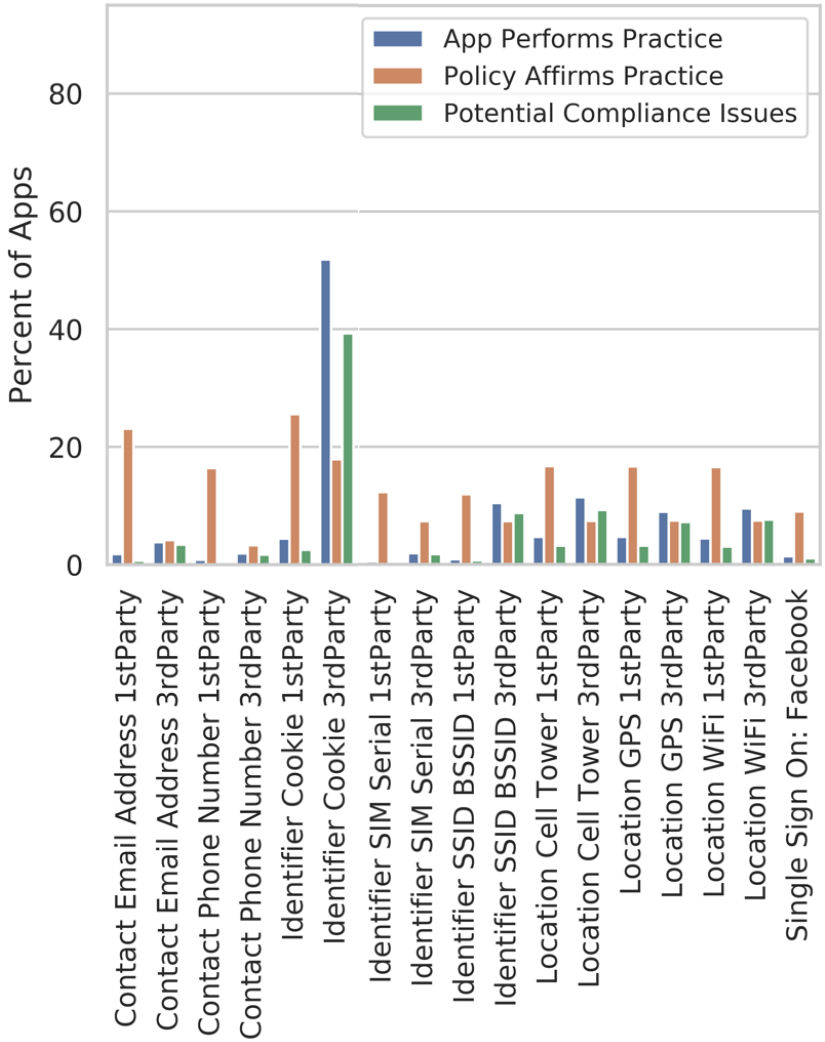
- **Problem 2: Few Negative Statements**

  We created synthetic data in our corpus

- **Problem 3: Ambiguity & Vagueness**

  Practices can be disclosed in general or specific terms ("We collect your location data." vs "We collect your GPS data.")

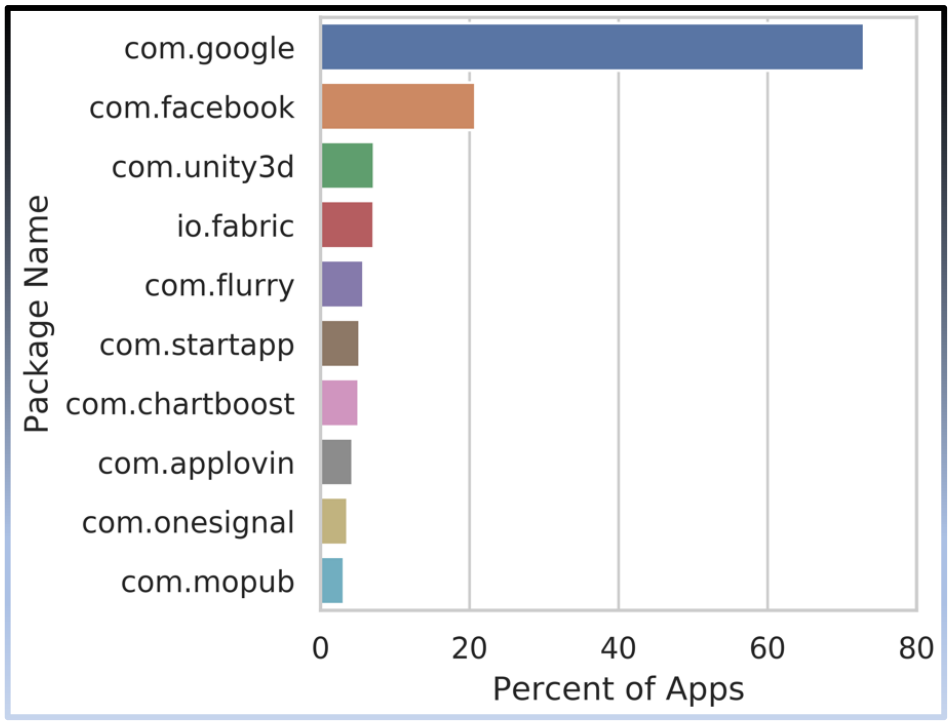




What are Common Potential Compliance Issues?

- **Identifier-related issues** are the most frequent ones, but **location-related issues** are present in a substantial number of apps as well
- If an app performs a practice, there is a good chance that a compliance issue exists
- **Third party compliance issues** are more frequent than first party compliance issues

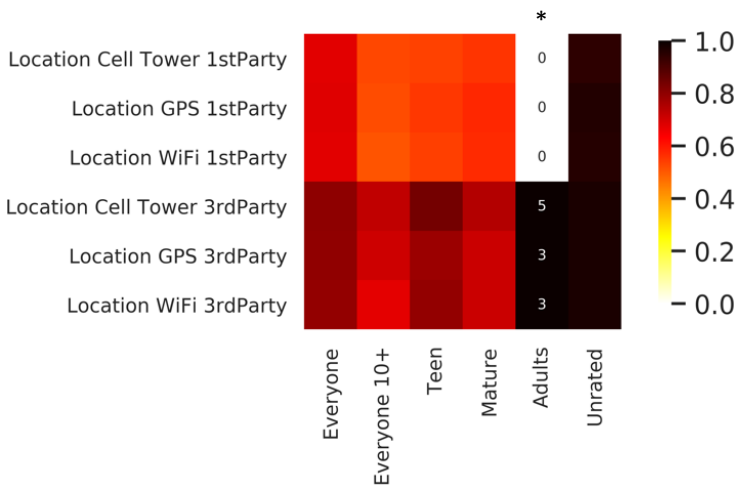
Who Are the Third Parties?



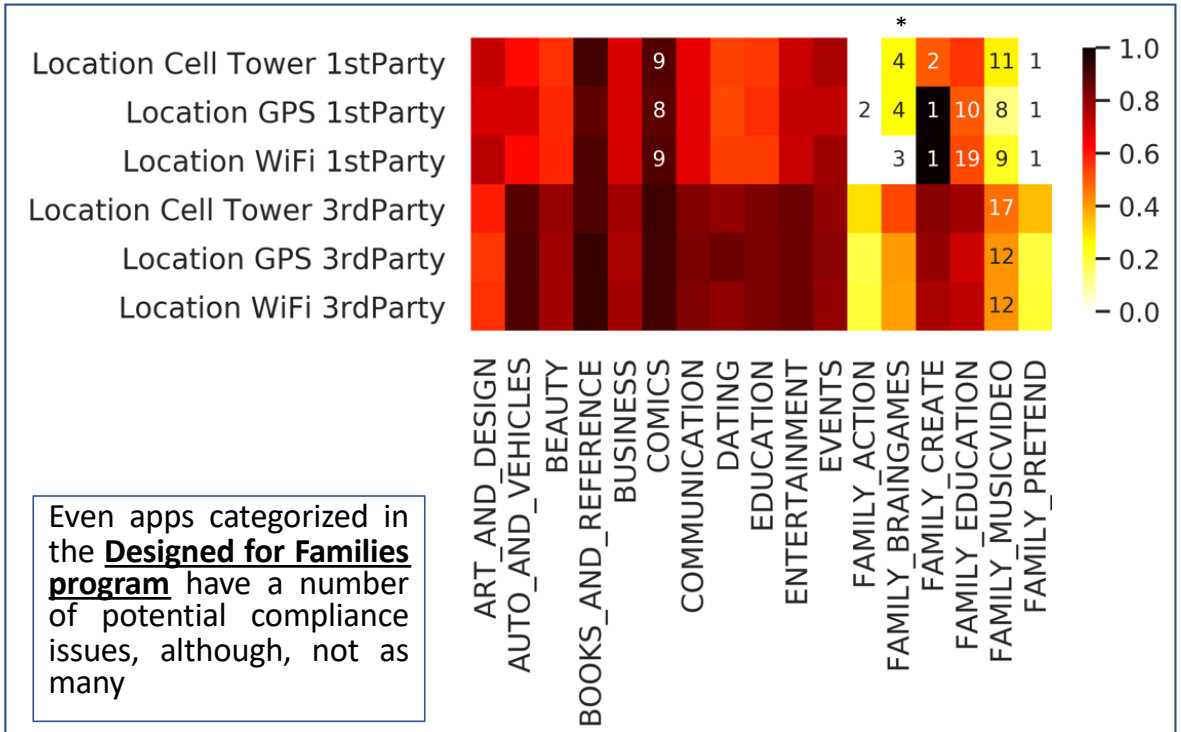
Different types of third parties:

- Advertisers ← most common ad packages
- Analytics
- Developer Tools
- Compatibility Libraries
- Authentication Packages

…



Apps without **Entertainment Software Rating Board (ESRB) Content rating** tend to have a higher number of potential compliance issues

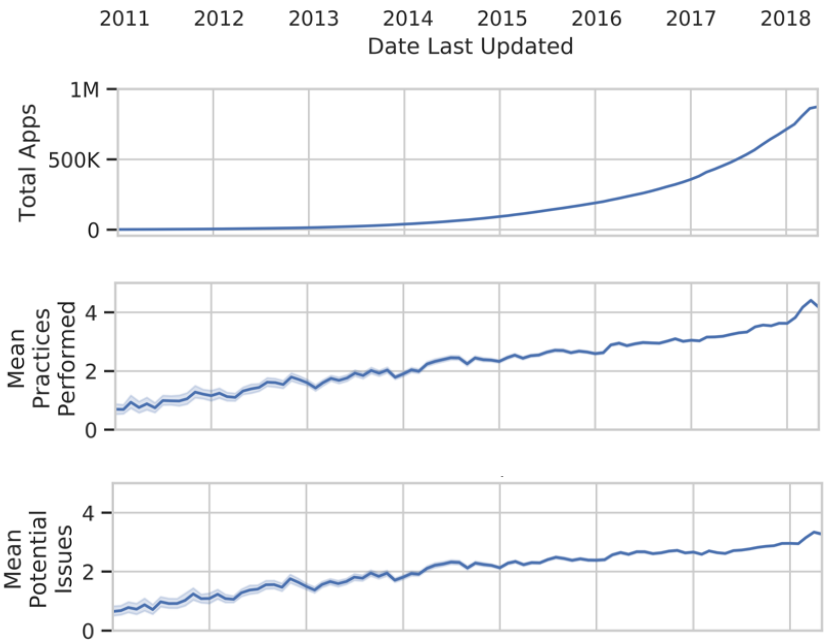


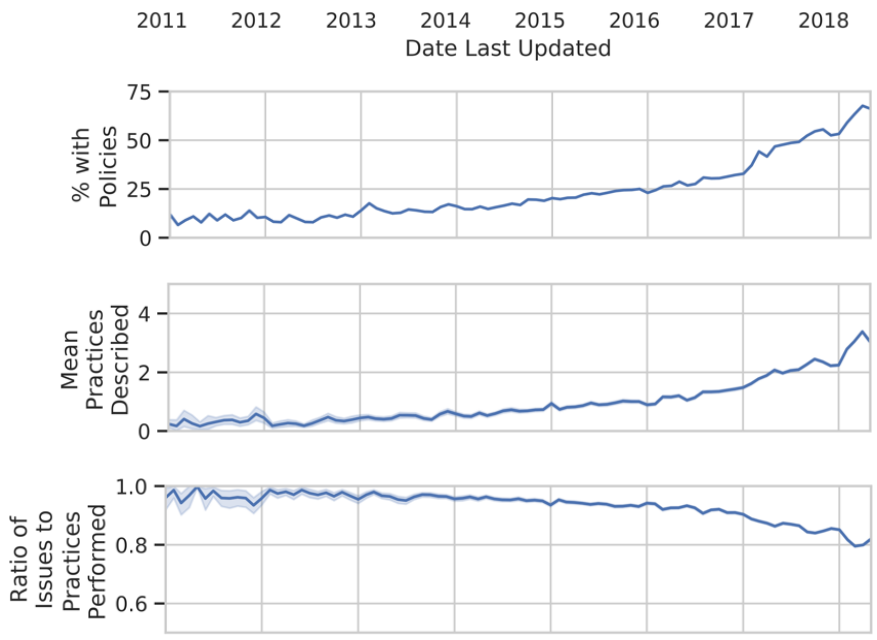
Even apps categorized in the **Designed for Families program** have a number of potential compliance issues, although, not as many

COPPA- and ESRB-related Potential Compliance Issues

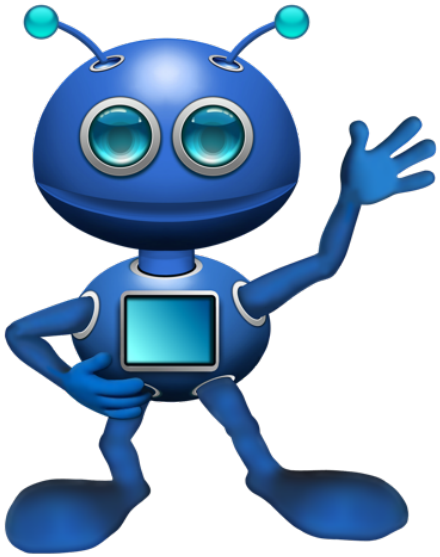
* Cells with fewer than 25 apps performing the practice are annotated with the respective number of apps.

Potential Compliance Issues: The Big Picture





What's the story? What should we do next?



- Many privacy compliance issues are due to policies' silence and opaque third party libraries
- Scale vs depth, especially, challenging for taint analysis
- Automation only supplements manual analysis but does not replace it
- That does it for today … **questions**?

This study was supported in part by the NSF Frontier grant on Usable Privacy Policies (CNS-1330596, CNS-1330141, and CNS-1330214) and a DARPA Brandeis grant on Personalized Privacy Assistants (FA8750-15-2-0277). The US Government is authorized to reproduce and distribute reprints for Governmental purposes not withstanding any copyright notation. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the NSF, DARPA, or the US Government.
This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing high performance computing resources that have contributed to the research results reported within this paper.