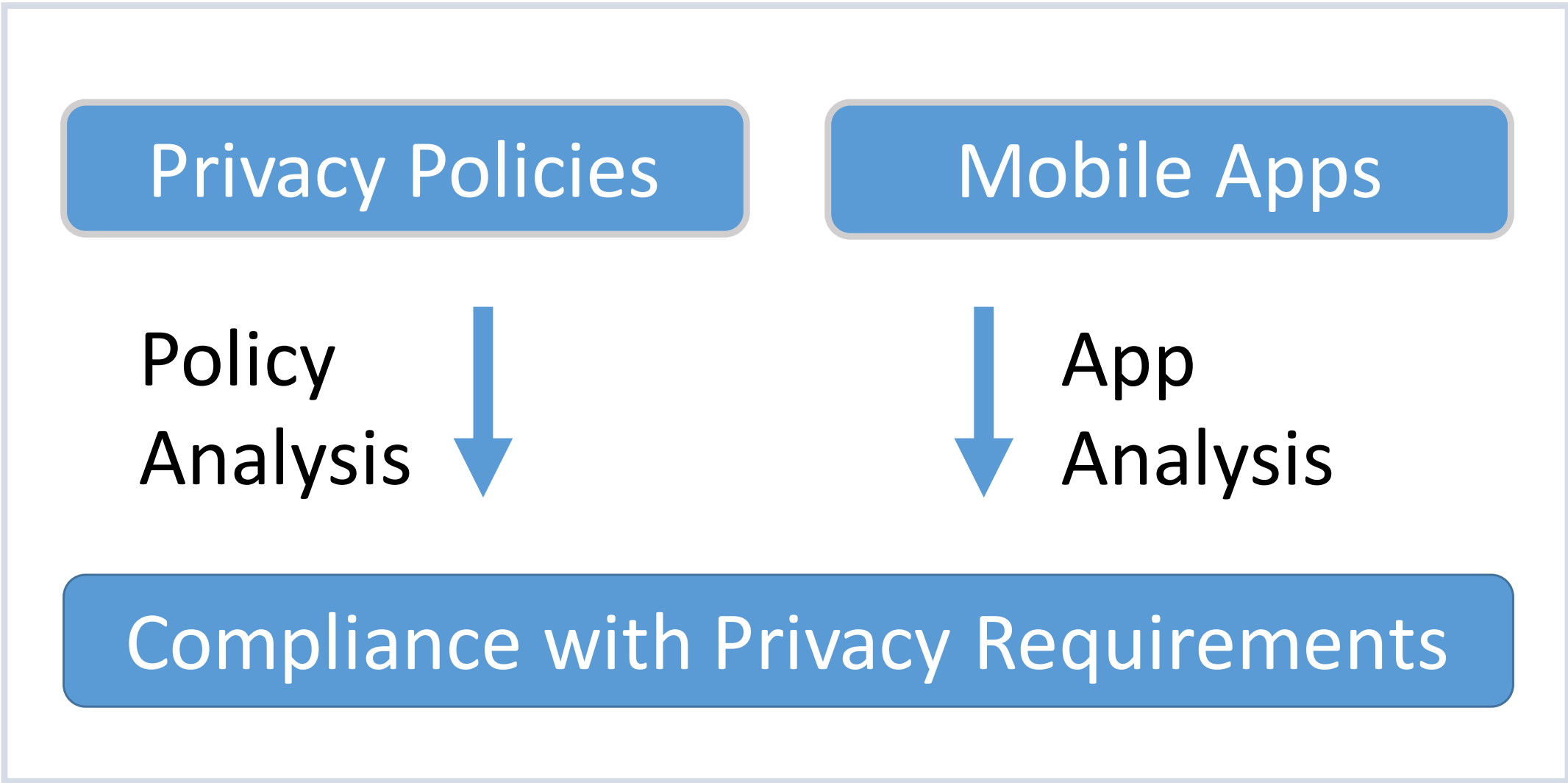
Automated Analysis of Privacy Requirements for Mobile Apps

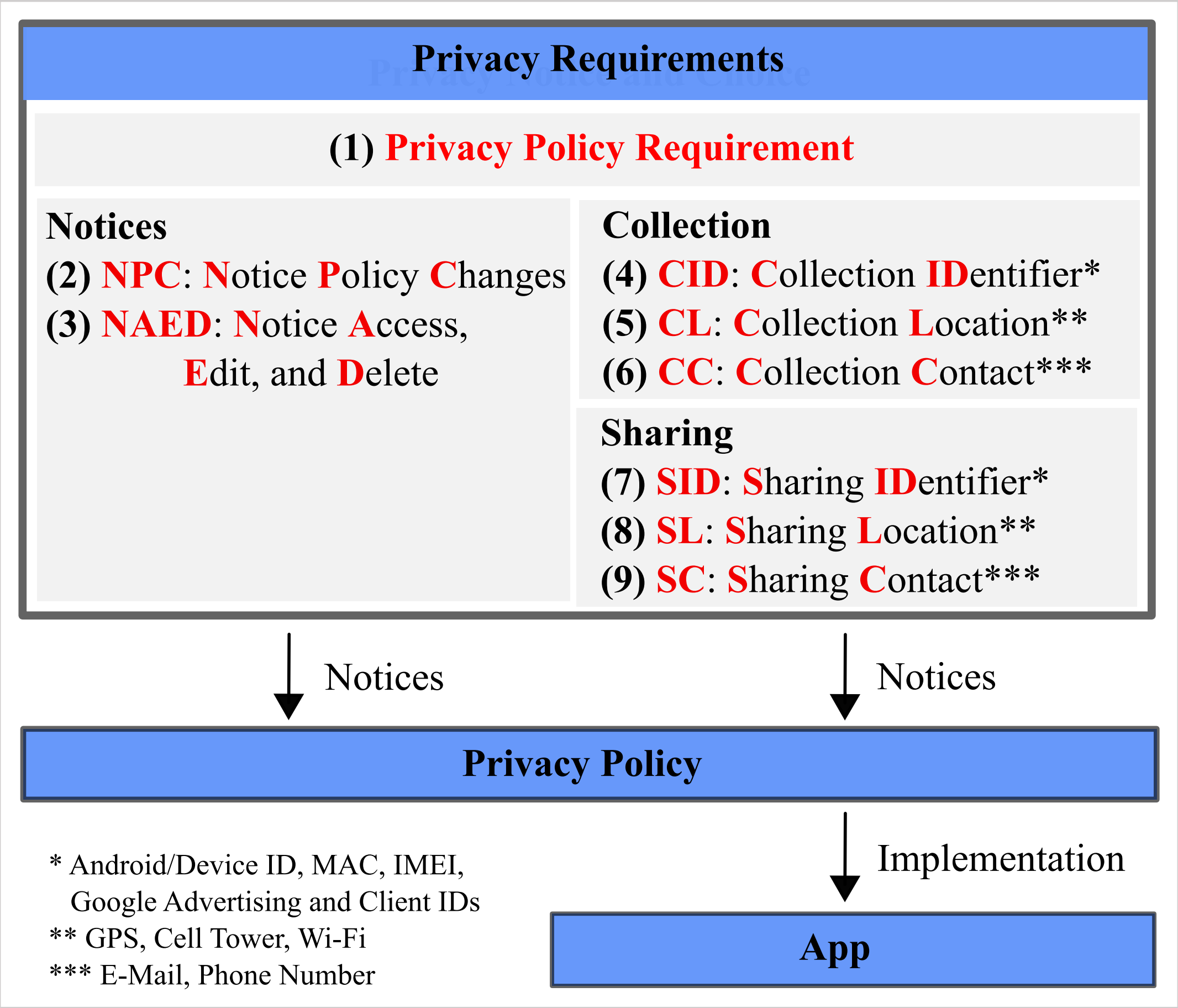
Sebastian Zimmeck, Ziqi Wang, Lieyong Zou, Roger Iyengar, Bin Liu, Florian Schaub, Shomir Wilson, Norman Sadeh, Steven M. Bellovin, Joel Reidenberg

Background



We introduce a <u>system to analyze Android apps' compliance with privacy requirements</u>

- We define privacy requirement compliance to mean that <u>apps need a privacy policy and must behave according to it</u>
- In addition, the <u>policy by itself is required to follow requirements</u> (e.g., on notifying a user on access, edit, and deletion rights)



Privacy Requirements

(1) Privacy Policy Requirement

Notices
(2) NPC: Notice Policy Changes
(3) NAED: Notice Access, Edit, and Delete

Collection
(4) CID: Collection IDentifier*
(5) CL: Collection Location**
(6) CC: Collection Contact***

Sharing
(7) SID: Sharing IDentifier*
(8) SL: Sharing Location**
(9) SC: Sharing Contact***

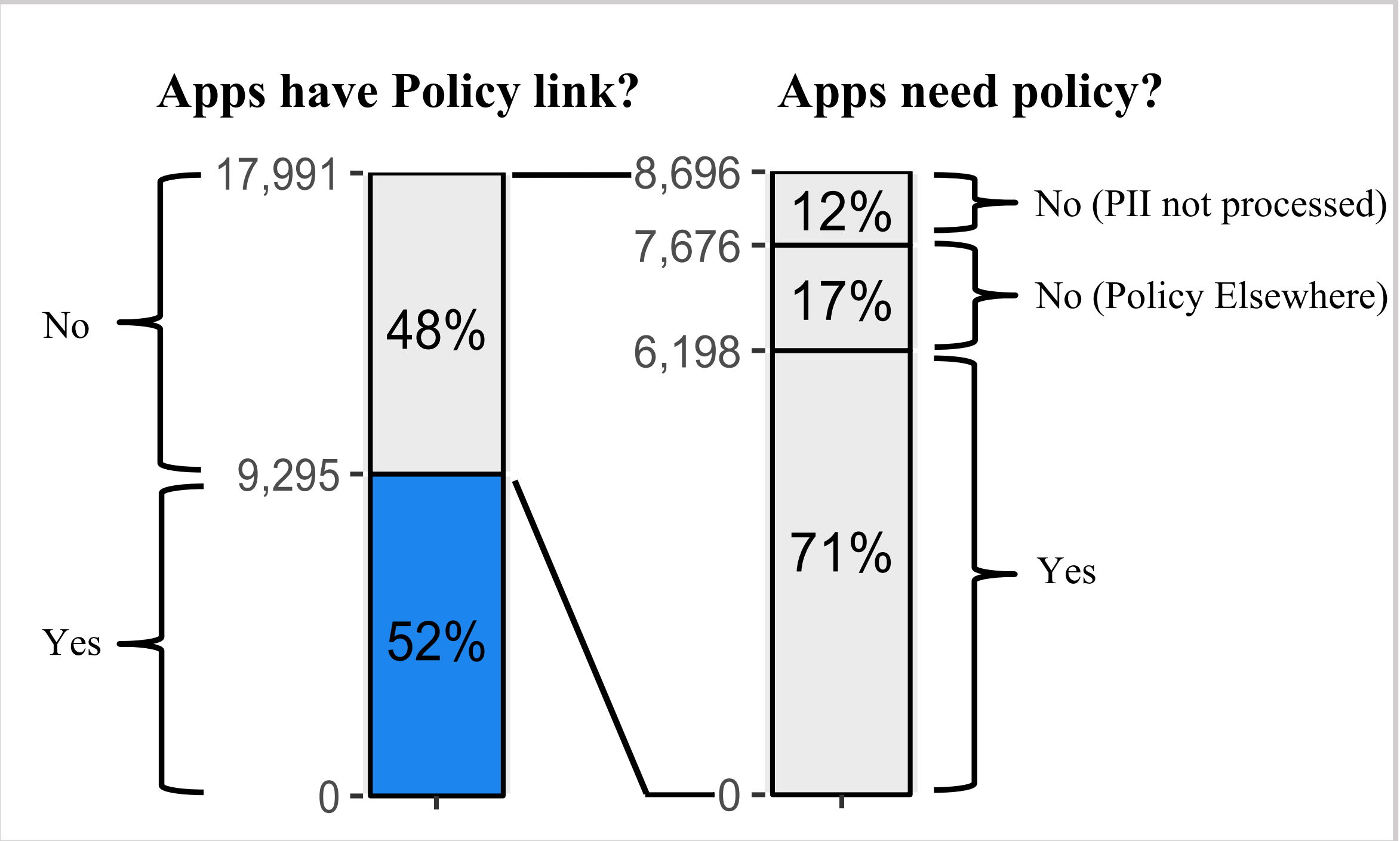
* Android/Device ID, MAC, IMEI, Google Advertising and Client IDs
** GPS, Cell Tower, Wi-Fi
*** E-Mail, Phone Number

- In detail, apps that process Personally Identifiable Information (PII) are generally required to:

  (1) have a privacy policy (either on its Google Play page or inside the app);
  (2) include notices about policy changes and access, edit, and deletion rights;
  (3) notify users of data collection practices; and
  (4) disclose how data is shared with third parties

Policy Analysis

<u>71% (6,198/8,696) of apps appear to have no privacy policy despite processing PII</u>



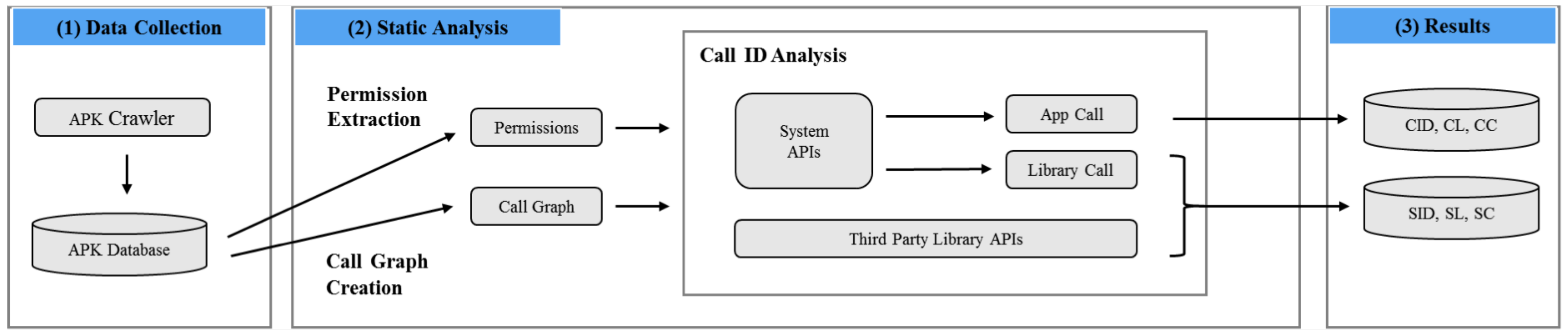
The system <u>classifies descriptions of practices in privacy policies based on machine learning</u>

(1) Keyword sets are used to identify practices: data type keywords and action keywords
(2) Sentences in policies are extracted based on data type keywords (e.g., all sentences that contain the term "location")
(3) Using action keywords unigram and bigram feature vectors are constructed from the extracted sentences (e.g., "share location")
(4) The unigram and bigram features are leveraged by Support Vector Machine (SVM) and Logistic Regression (Log. Reg.) classifiers

Practice	*Classifier*	*Base (n=40)*	*Acc_{pol} (n=40)*	*95% CI (n=40)*	*$Prec_{neg}$ (n=40)*	*Rec_{neg} (n=40)*	*F-1_{neg} (n=40)*	*F-1_{pos} (n=40)*	*Pos (n=9,050)*
NPC	SVM	0.7	0.9	0.76–0.97	0.79	0.92	**0.85**	0.93	**46%**
NAED	SVM	0.58	0.75	0.59–0.87	0.71	0.71	**0.71**	0.78	**36%**
CID	Log. Reg.	0.65	0.83	0.67–0.93	0.77	0.71	**0.74**	0.87	46%
CL	SVM	0.53	0.88	0.73–0.96	0.83	0.95	**0.89**	0.86	34%
CC	Log. Reg.	0.8	0.88	0.73–0.96	0.71	0.63	**0.67**	0.92	56%
SID	Log. Reg.	0.88	0.88	0.73–0.96	0.94	0.91	**0.93**	**0.55**	**10%**
SL	SVM	0.95	0.93	0.8–0.98	0.97	0.95	**0.96**	-	**12%**
SC	SVM	0.73	0.78	0.62–0.89	0.79	0.93	**0.86**	**0.47**	**6%**

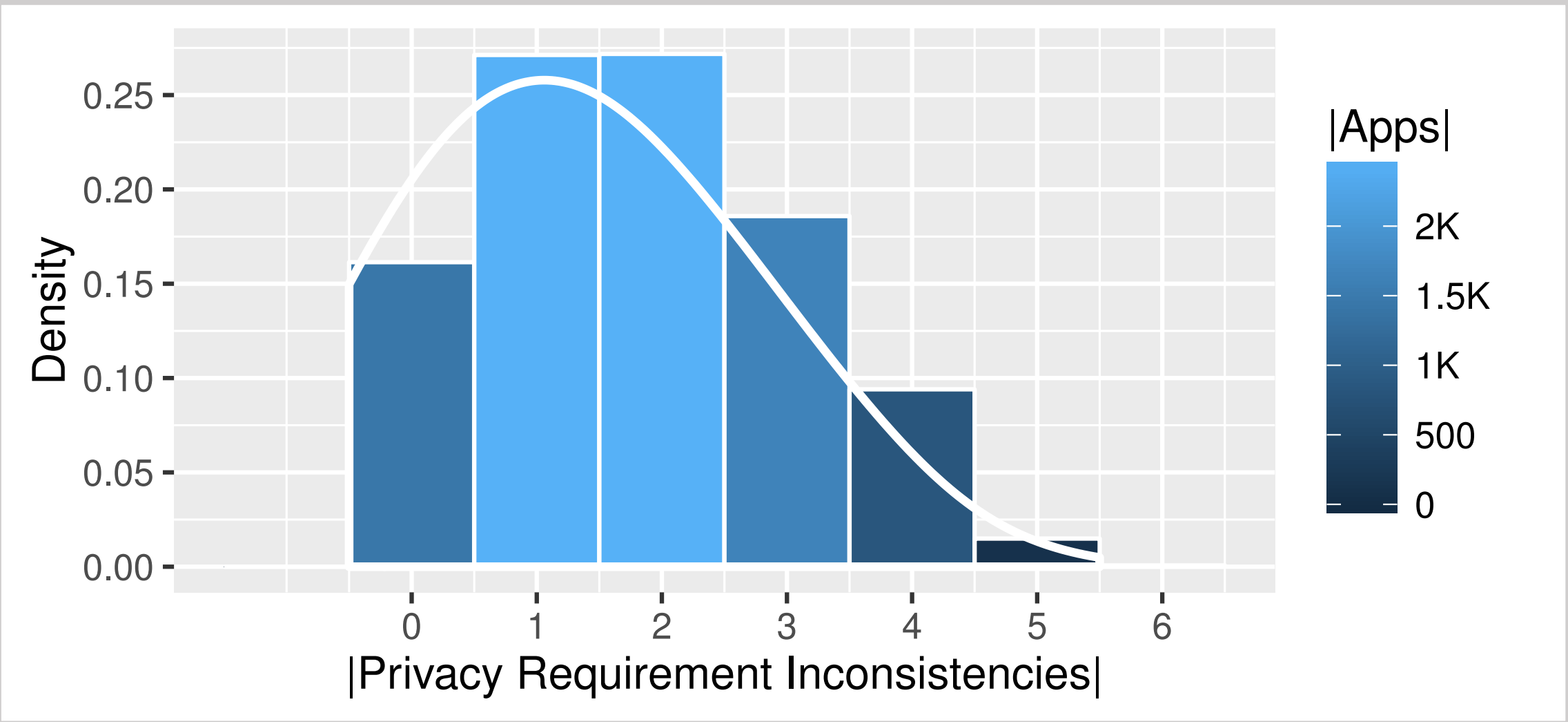
Classification results for a policy test set (n=40) and the occurrence of positive classifications (Pos) in a set of n=9,050 policies

App Analysis



(1) The system first crawls the US Google Play store for free apps
(2) It then performs static analysis on the app code (consisting of permission extraction, call graph creation, and call ID analysis)
(3) The resulting collection and sharing practices of the app are stored in a database

Compliance with Privacy Requirements



- 2,455 apps have one potential privacy requirement non-compliance, 2,460 have two, and only 1,461 adhere completely to their policy (out of n = 9,050 apps)
- **<u>Each app exhibits a mean of 1.83 instances of potential privacy requirement non-compliance</u>**
- Non-compliance does not necessarily mean that a law is violated

Practice	*Acc (n=40)*	*$Acc_{pol} \cdot Acc_{app}$ (n=40)*	*95% CI (n=40)*	*$Prec_{pos}$ (n=40)*	*Rec_{pos} (n=40)*	*F-1_{pos} (n=40)*	*F-1_{neg} (n=40)*	*MCC (n=40)*	*TP, FP, TN, FN (n=40)*	*Inconsistency (n=9,050)*
CID	**0.95**	0.74	0.83–0.99	0.75	1	**0.86**	0.97	0.84	6, 2, 32, 0	**50%**
CL	**0.83**	0.7	0.67–0.93	**0.54**	1	**0.7**	0.88	0.65	8, 7, 25, 0	**41%**
CC	**1**	0.88	0.91–1	-	-	-	1	-	0, 0, 40, 0	**9%**
SID	**0.85**	0.84	0.7–0.94	0.93	0.74	**0.82**	0.87	0.71	14, 1, 20, 5	**63%**
SL	**1**	0.93	0.91–1	1	1	**1s**	1	1	3, 0, 37, 0	**17%**
SC	**1**	0.78	0.91–1	1	1	**1**	1	1	1, 0, 39, 0	**2%**

Identifying privacy requirement non-compliance for a test set of app/policy pairs (n=40) and the percentages of potential non-compliance (Inconsistency) for n=9,050 app/policy pairs

NSF SaTC
Frontier project
Lead PI:
Norman Sadeh
sadeh@cs.cmu.edu

This material is based upon work supported in part by the National Science Foundation under grants CNS-1330596, CNS-1330214, and SBE-1513957, as well as by DARPA and the Air Force Research Laboratory, under agreement number FA8750-15-2-0277. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes not withstanding any copyright notation thereon. 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 DARPA, the Air Force Research Laboratory, the National Science Foundation, or the U.S. Government.