Free for All! Assessing User Data Exposure to Advertising Libraries on Android

Soteris Demetriou, Whitney Merrill, Wei Yang, Aston Zhang, Carl Gunter
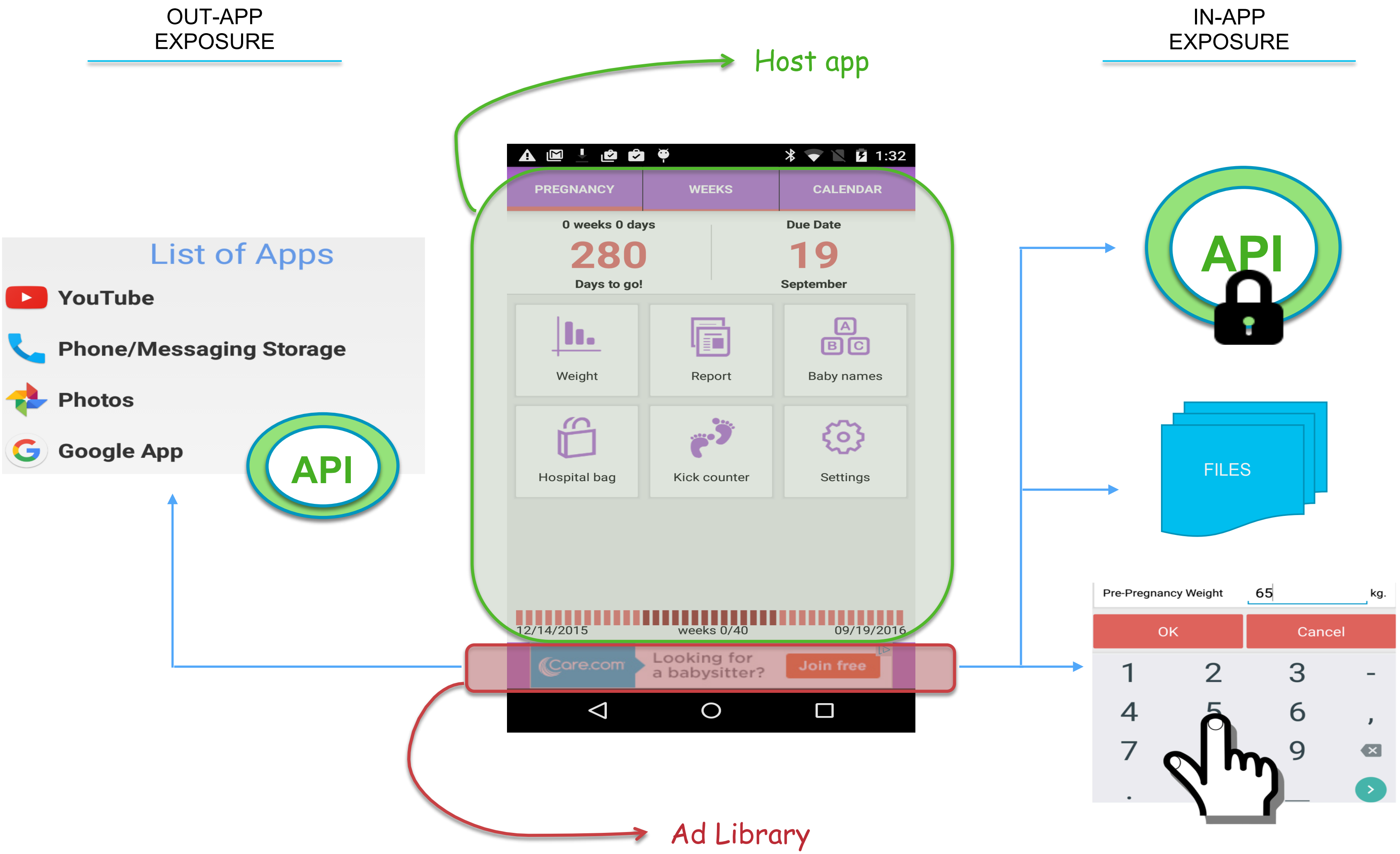
University of Illinois at Urbana - Champaign

Approach

Approach

- GOAL: Assess the **RISK** of integrating advertising libraries in Android apps

- RISK: Potential compromise of an *asset* as a result of an exploit of a *vulnerability* by a *threat*.

All the different ways an **ad library** can access **private user data**

Private User Data

Ad Library

Approach

OUT-APP
EXPOSURE

IN-APP
EXPOSURE

Host app

List of Apps

YouTube

Phone/Messaging Storage

Photos

Google App

API

API

FILES

Ad Library

PREGNANCY   WEEKS   CALENDAR

0 weeks 0 days
280
Days to go!

Due Date
19
September

Weight   Report   Baby names

Hospital bag   Kick counter   Settings

12/14/2015   weeks 0/40   09/19/2016

Care.com Looking for a babysitter?   Join free

Pre-Pregnancy Weight   65   kg.

OK   Cancel

1   2   3   -
4   5   6   ,
7   8   9

Is there any interesting information in local files?

FILES

Motivation: in-app

FILES

I'm Pregnant / Pregnancy App

Installs
1,000,000 - 5,000,000

4.4
★★★★⯪
👤 15,605 total

★ 5	9,969	
★ 4	3,342	
★ 3	1,231	
★ 2	470	
★ 1	593	
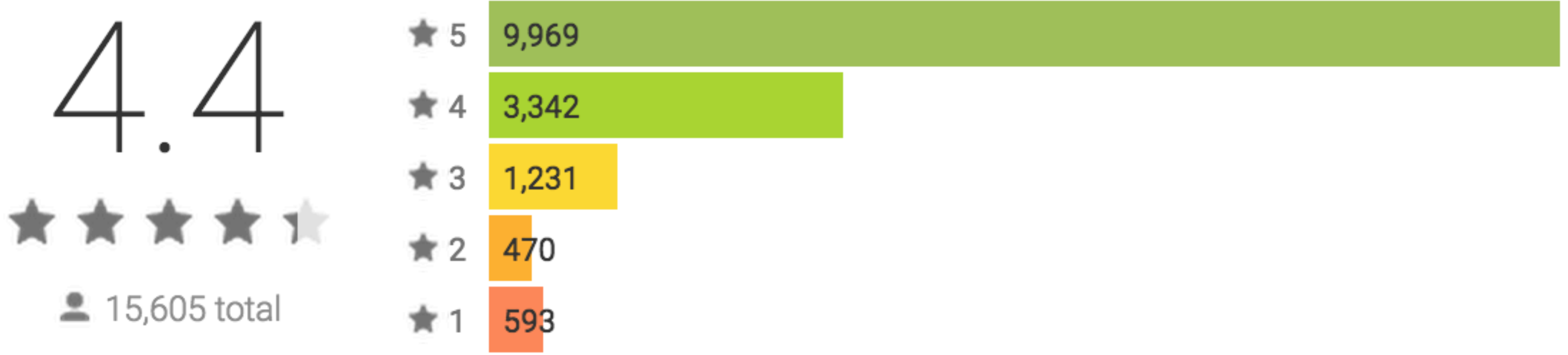
- Weight
- Height
- Pregnancy month and day

- Symptoms (headaches, backache, constipation)
- Events (date of intercourse)
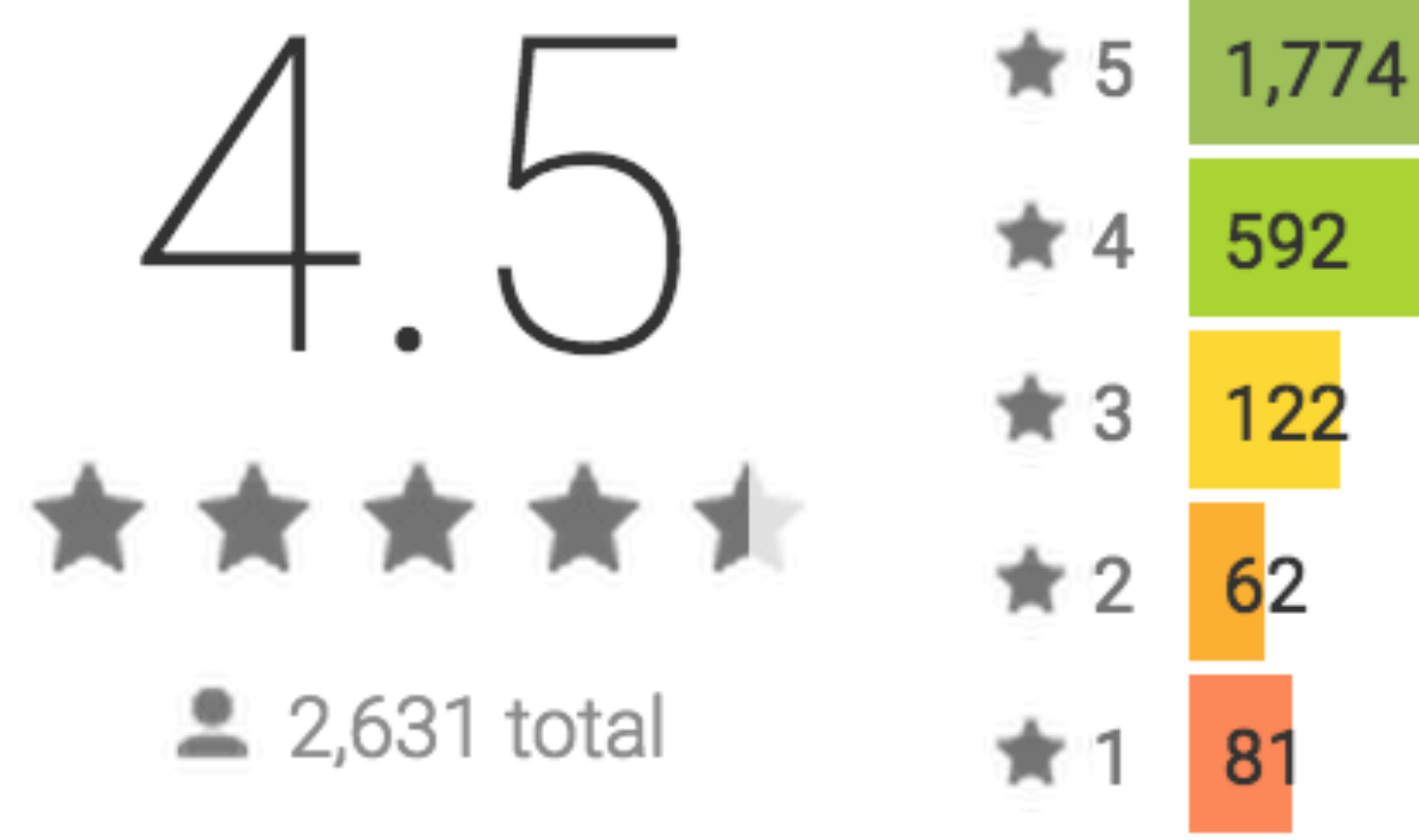- Outcomes (miscarriage, birth date)
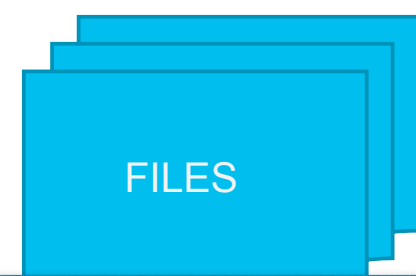
Motivation: in-app

FILES

Diabetes Journal

Installs
100,000 - 500,000

4.5

★ ★ ★ ★ ½

👤 2,631 total

★ 5	1,774	
★ 4	592	
★ 3	122	
★ 2	62	
★ 1	81	

- Birth date
- Gender
- First name
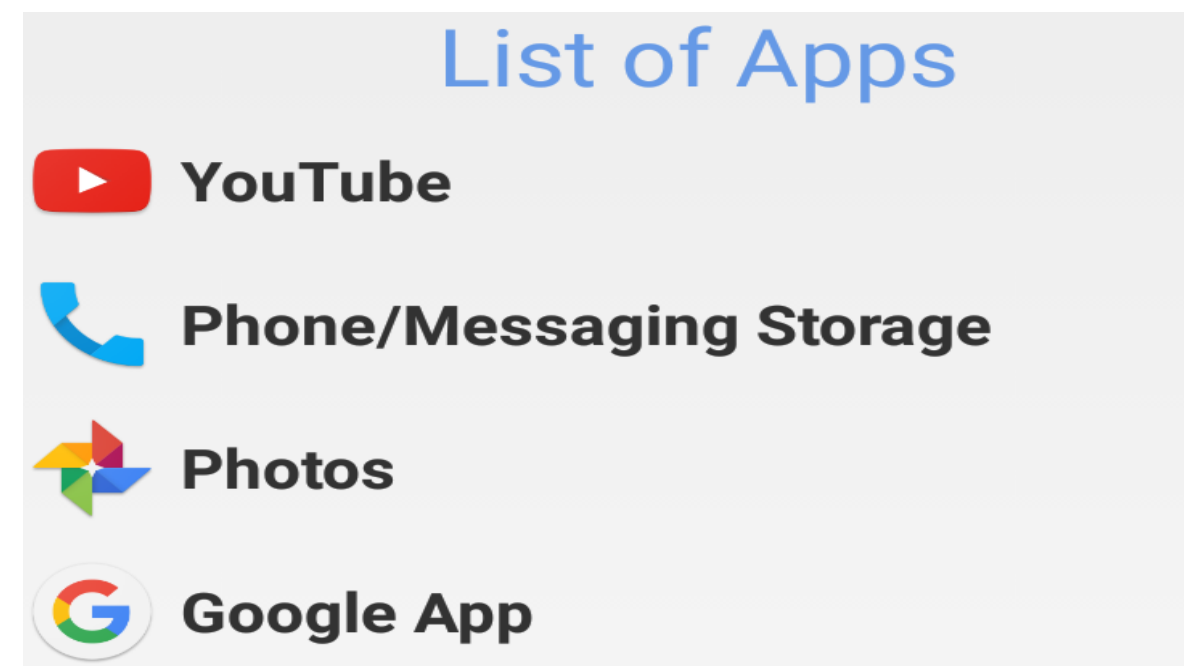- Last name

- Weight
- Height
- Blood glucose levels
- Workout activities

FILES

- There is a plethora of private user information in app local files.

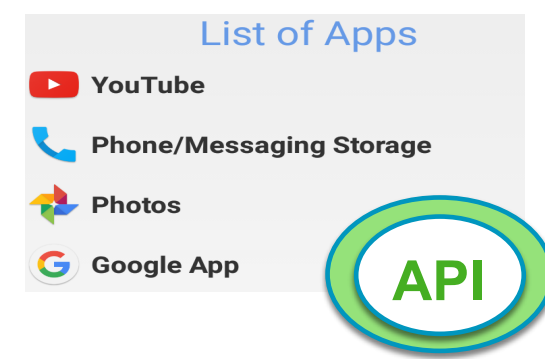- It is trivial for ad libraries to access such information.

- Last name

- Workout activities

Are ad libraries interested in app bundles?

List of Apps

- YouTube
- Phone/Messaging Storage
- Photos
- Google App

Motivation: out-app

METHODOLOGY

- Call graphs on 2700 Google Play apps

  - getInstalledPackages (gIP)

  - getInstalledApplications (gIA)
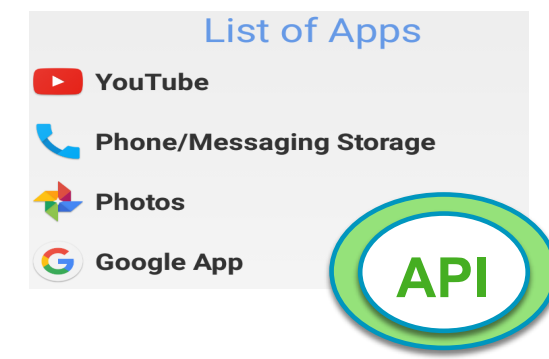
- Manual analysis of packages containing gIP and gIA

RESULTS

- 2535 unique apps

- 27.5% contain at least one invocation of gIP or gIA

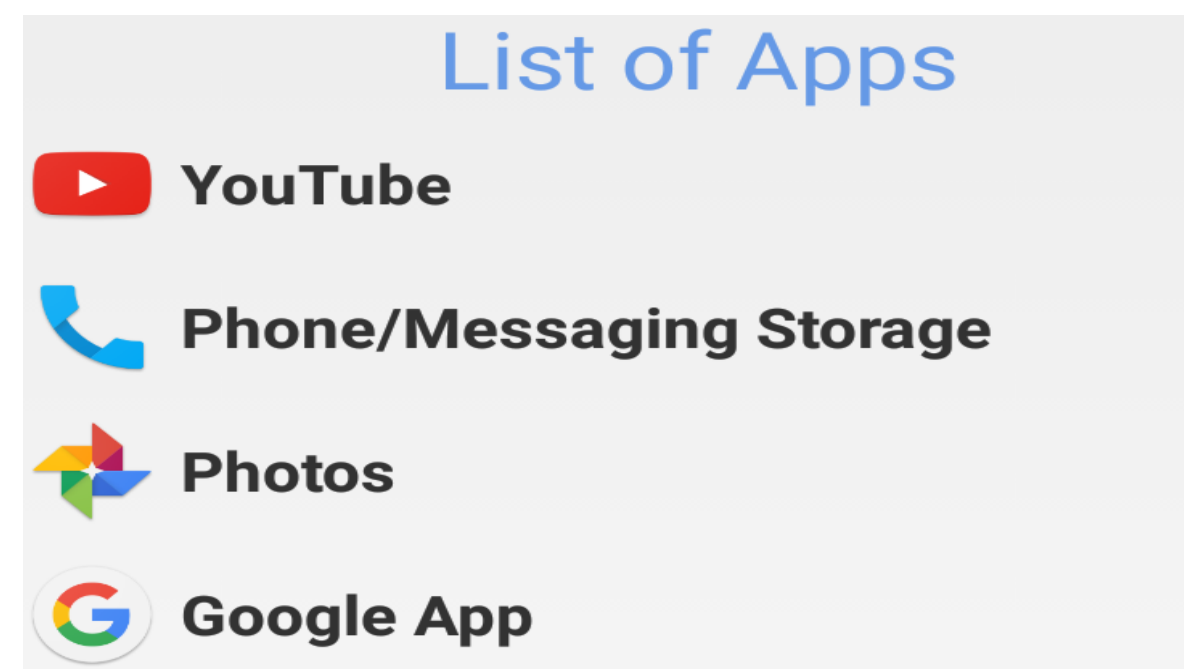- 12.54% contain an ad library that invokes gIP or gIA

- 28 unique ad libraries

Motivation: out-app

List of Apps

▶ YouTube
📞 Phone/Messaging Storage
📷 Photos
G Google App   API

Ad Libraries are increasingly collecting app bundles from user devices.
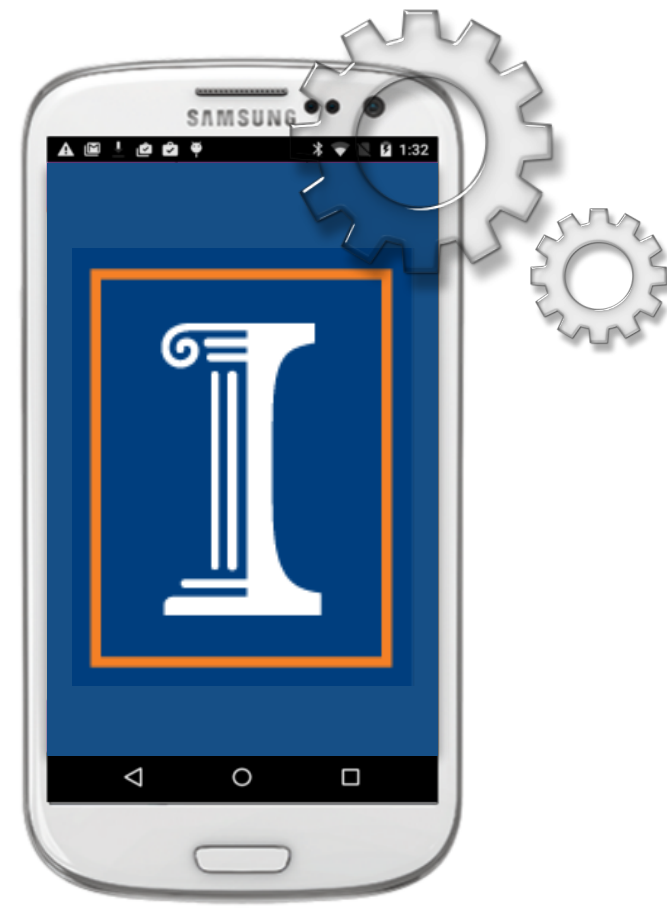
What can ad libraries learn from app bundles?

Motivation: out-app

Ground Truth collection: Private User Data

List of Apps
- YouTube
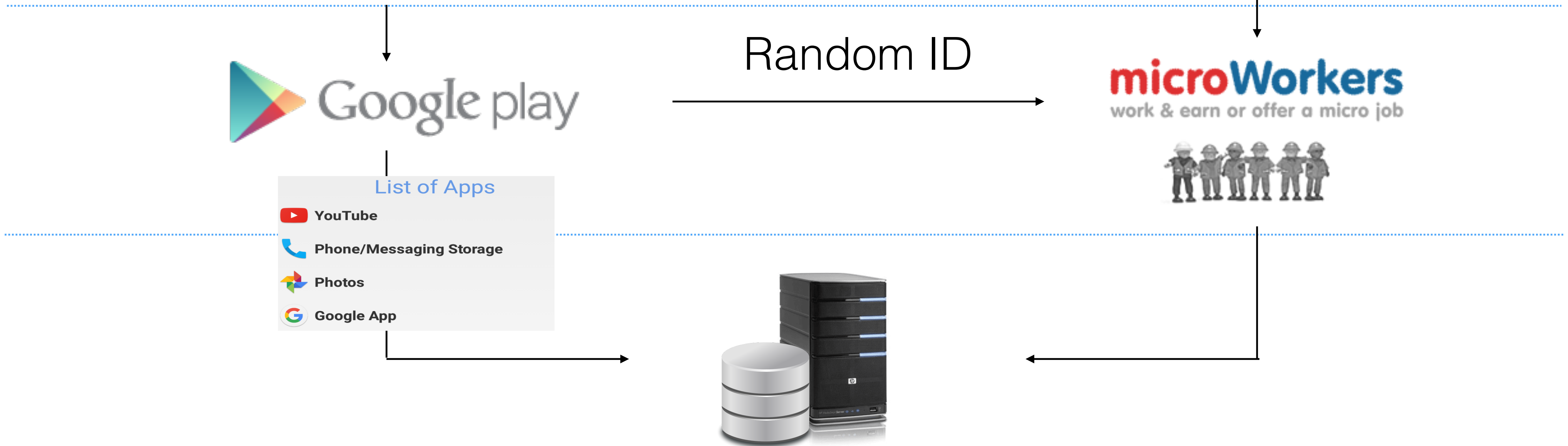- Phone/Messaging Storage
- Photos
- Google App
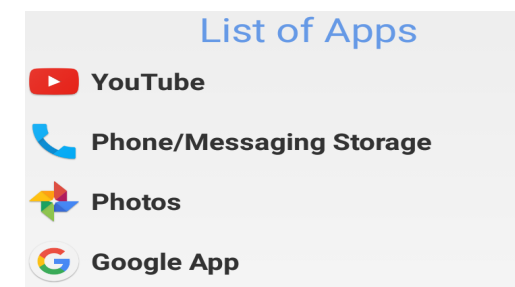
FINANCIAL TIMES

- Question 1
- Question 2
- …

Random ID

Google play

microWorkers
work & earn or offer a micro job

List of Apps
- YouTube
- Phone/Messaging Storage
- Photos
- Google App

Motivation: out-app

Ground Truth collection: Private User Data

FINANCIAL TIMES

243 approved users

1985 distinct apps

Evaluation: out-app

	AGE		MARITAL STATUS		SEX	
	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)
Random Forest	88.6	88.6	95.0	93.8	93.8	92.9
SVM	44.8	35.4	66.9	50.5	80.9	70.1
KNN	85.7	83.6	92.5	91.2	91.6	89.9

P: Precision
R: Recall

Pluto Risk Assessment Framework

Pluto Design

PURPOSE: "offline" estimation of the private user data a target app can expose to an embedded ad library that utilizes:

- in-app attack channels

- out-app attack channels [please see the paper for details]

Pluto Design: in-app exposure discovery
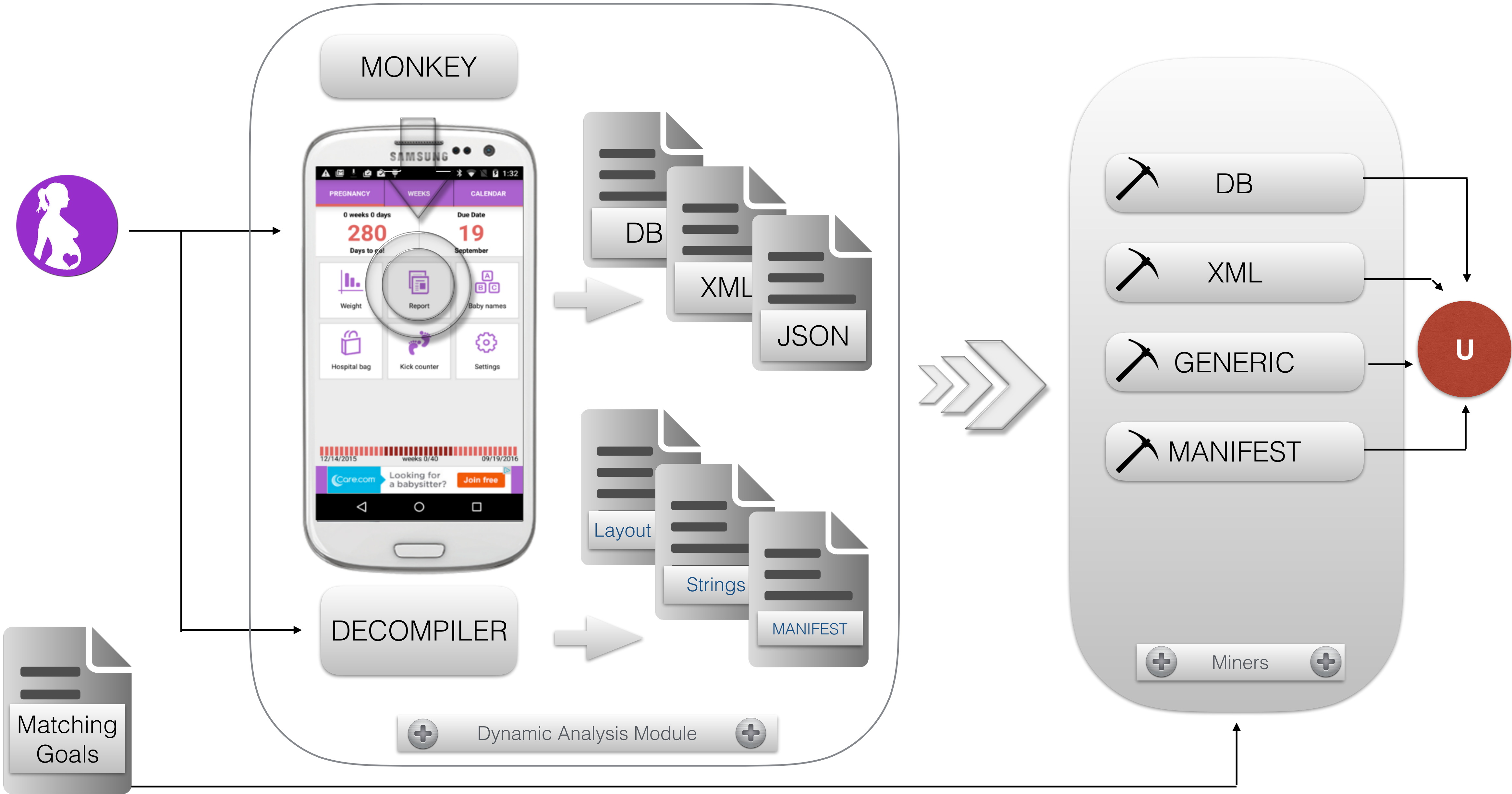
Evaluation

Evaluation

Evaluation: in-app

Ground Truth collection: Manual construction of L1 and L2

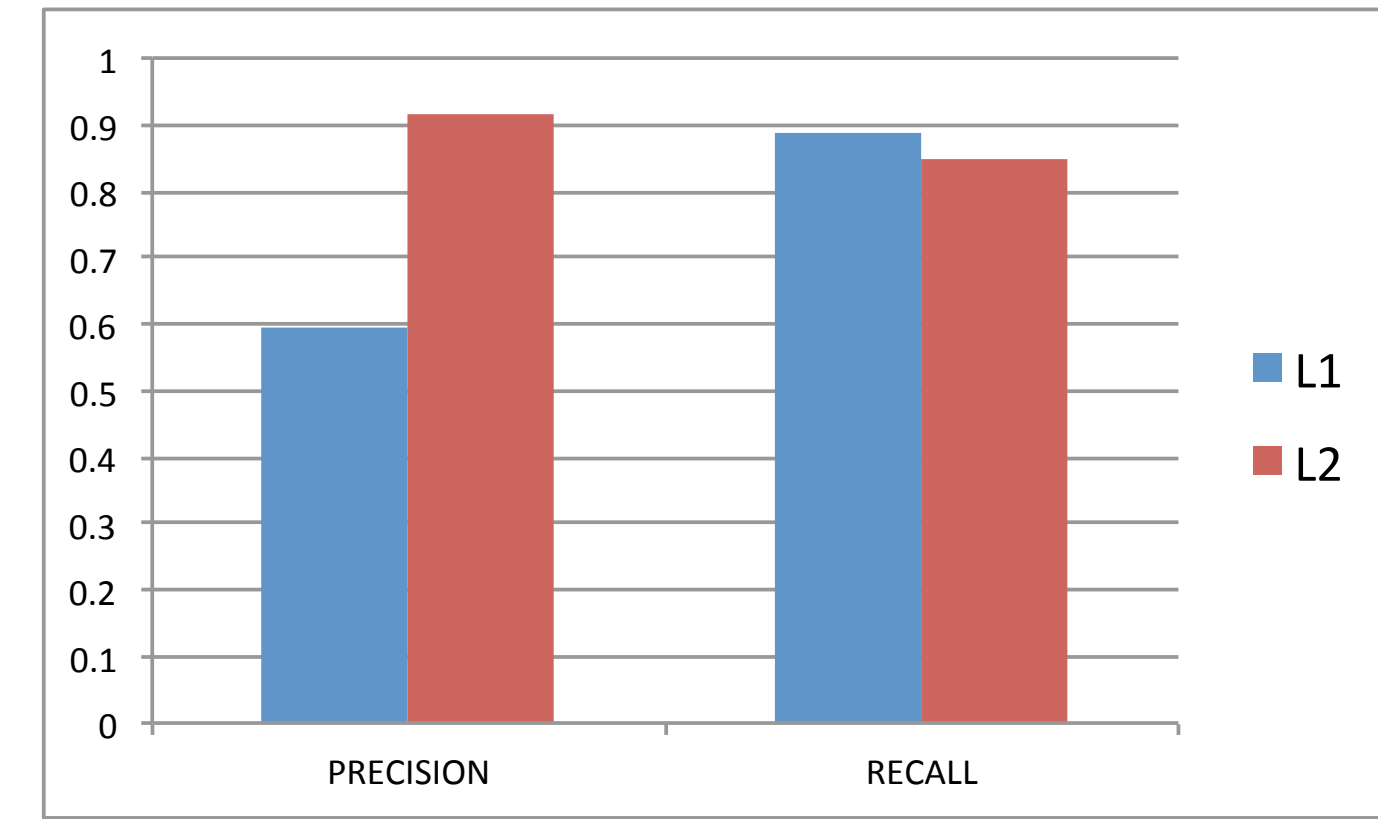| Name	Number	Description
Full Dataset (FD)	2535	Unique apps collected from the 27 Google Play categories
Level 1 Dataset (L1)	262	Apps randomly selected from FD
Level 2 Dataset (L2)	35	Apps purposively selected from L1

Evaluation: in-app
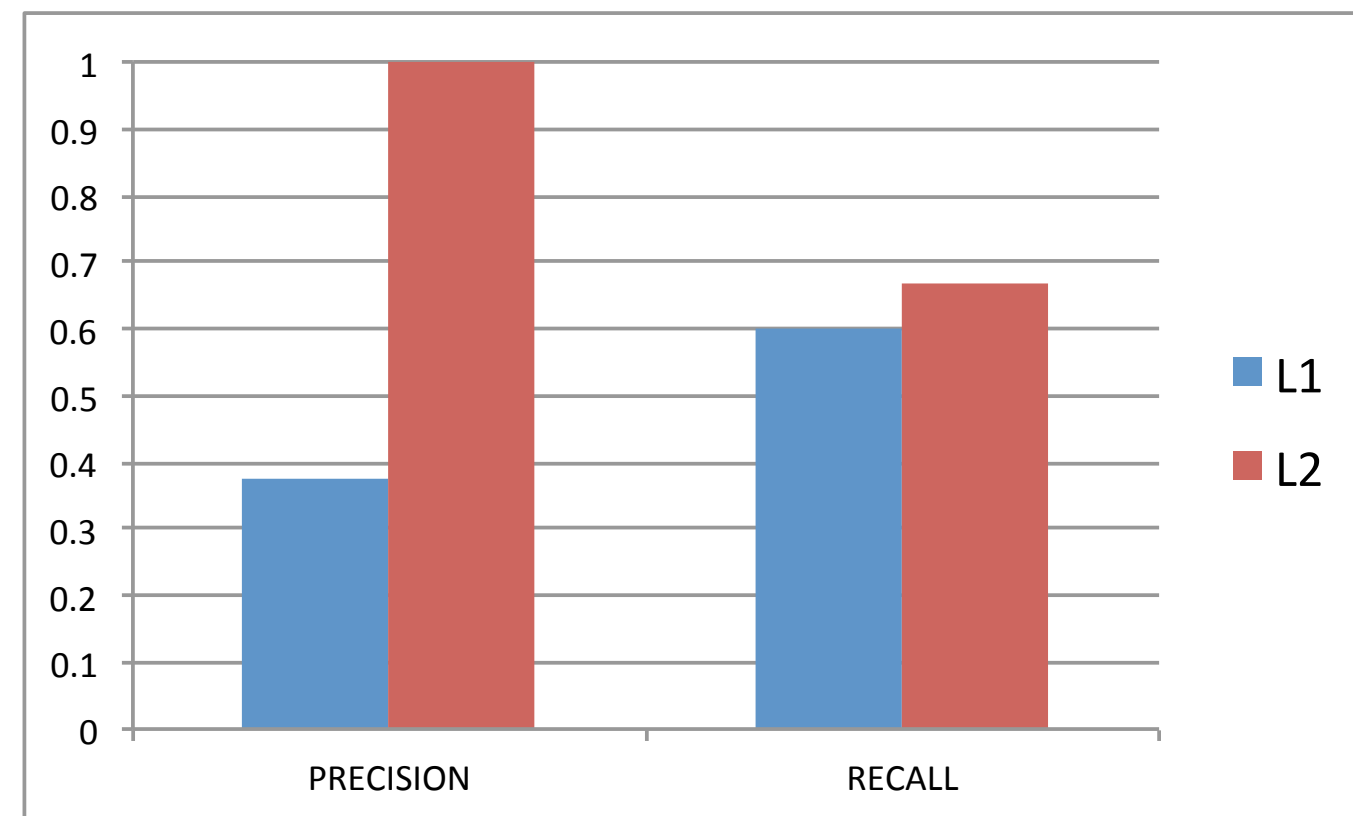

AGE
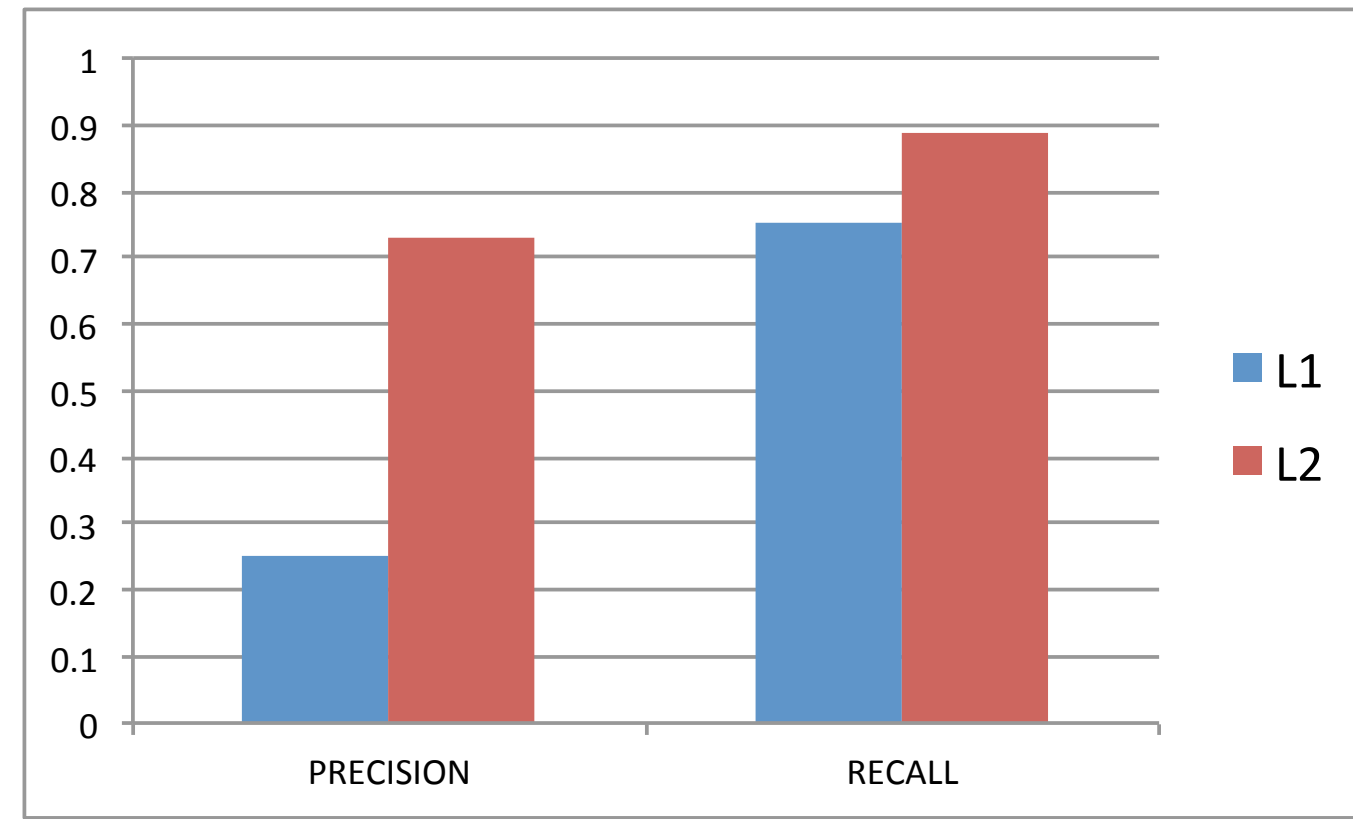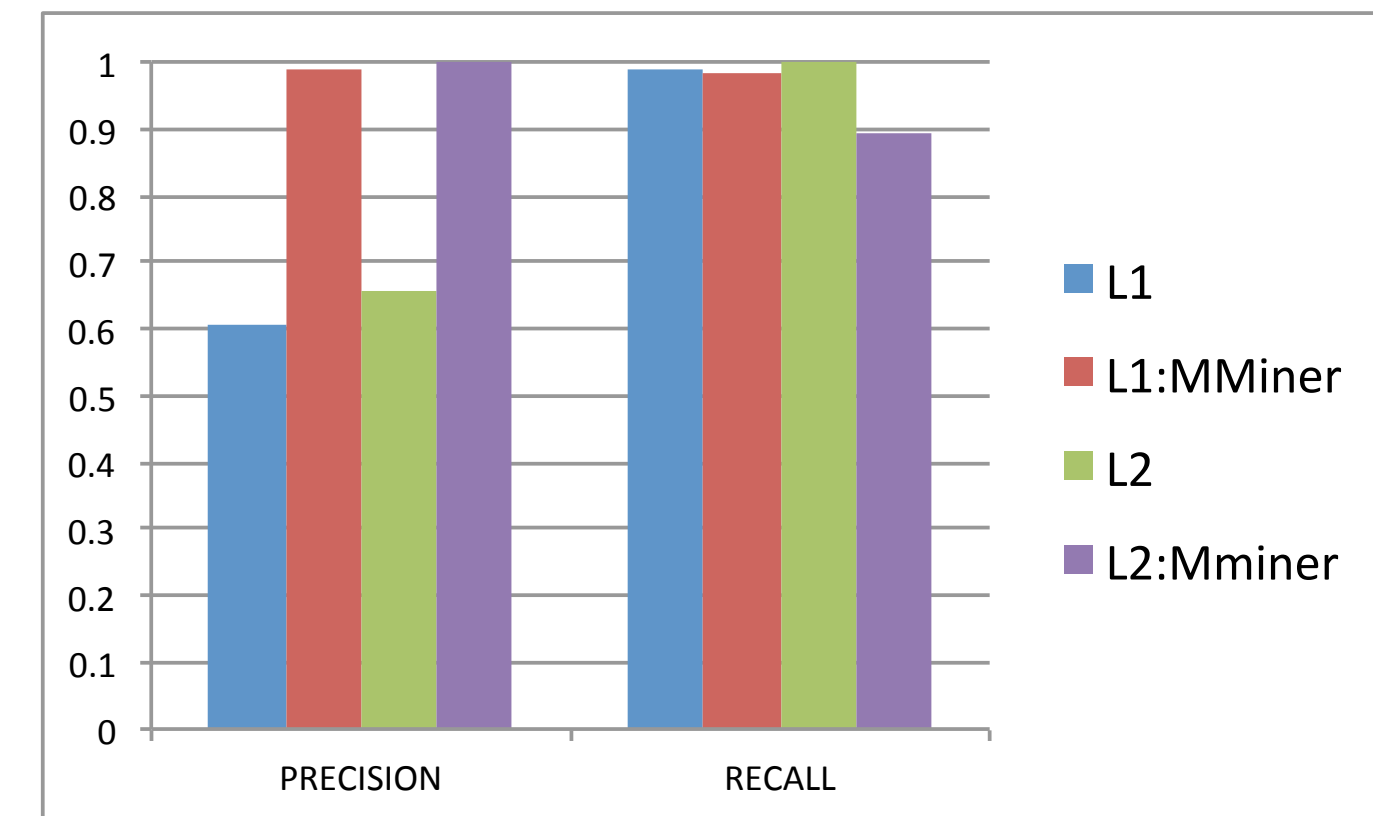

GENDER

WORKOUT

ADDRESS

Privacy Risk App Ranking

Utility: assessing the risk with Pluto

- D: set of data points in cost model (e.g. Financial Times)

- X: set of data point weights in the cost model

- $|D| = |X| = n$

- α: target app

- x_α: sum of all weights of data points exposed by α

risk score: $$z_\alpha = \frac{x_\alpha - min(X)}{\sum\limits_{i=1}^{n} x_i - min(X)}$$

Utility: assessing the risk with Pluto

CATEGORY		APP TITLE	AVG # INSTALLS	RISK SCORE [0 - 10]
MEDICAL		Depression CBT Self-Help Guide	100K - 500K	8.14
MEDICAL		Prognosis: Your Diagnosis	500K - 1M	6.31
HEALTH & FITNESS		Dream Body Workout Plan	100K - 500K	7.33
HEALTH & FITNESS		myCigna	100K - 500K	5.62

Utility: assessing the risk with Pluto

CATEGORY	APP TITLE	AVG # INSTALLS	RISK SCORE [0 - 10]
MEDICAL	Depression CBT Self-Help Guide	100K - 500K	8.14
MEDICAL	Prognosis: Your Diagnosis	500K - 1M	6.31
HEALTH &			

exposes 16 data points

depression, headache, pregnancy,...

Summary

- Apps store an abundance of private user data in local files.

- Revealed a trend of aggressive collection of app bundles.

- New techniques for assessing user sensitive information exposure to libraries. [not covered in this talk]

- Designed a tool (Pluto) to automatically assess the data exposure risk to third-party libraries by apps at scale.

- Pluto is evaluated on real world apps and user data and evidently achieves good prediction performance.

Thank You!

Source code is available online at:

https://github.com/soteris/android-advertising-pluto