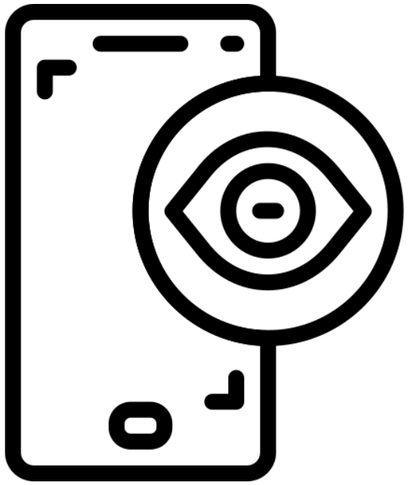
Inferring the Purposes of Network Traffic in Mobile Apps



Who (which app) sends the data?

Where the data is being sent to?

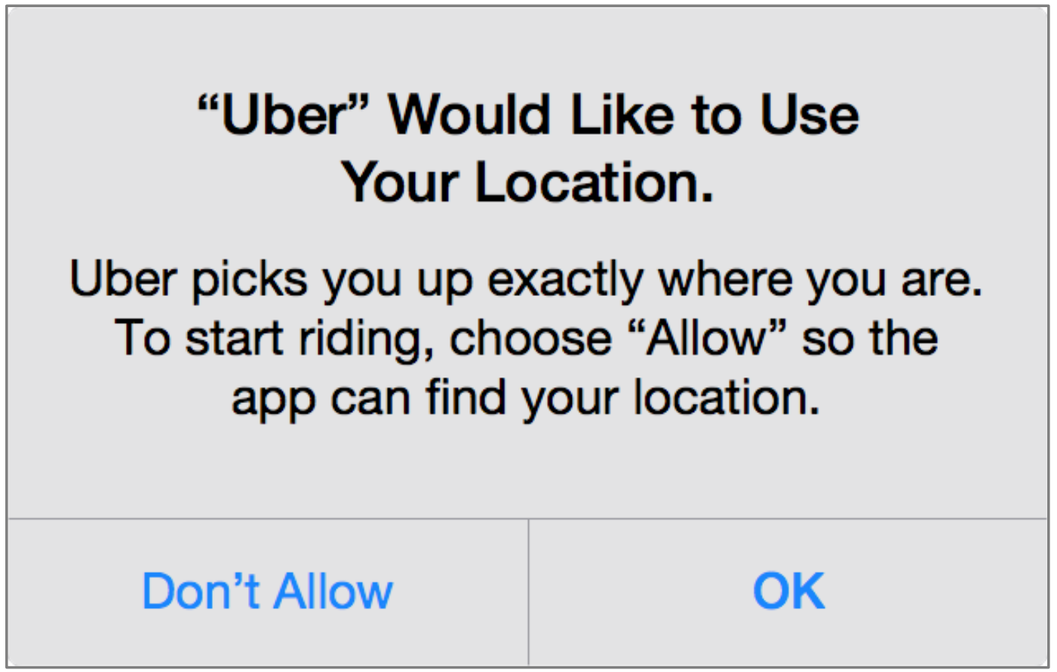
What data is being collected?

Why the data is being collected?

Haojian Jin, Minyi Liu, Yuanchun Li, Gaurav Srivastava,
Matthew Fredrikson, Yuvraj Agarwal, Jason Hong

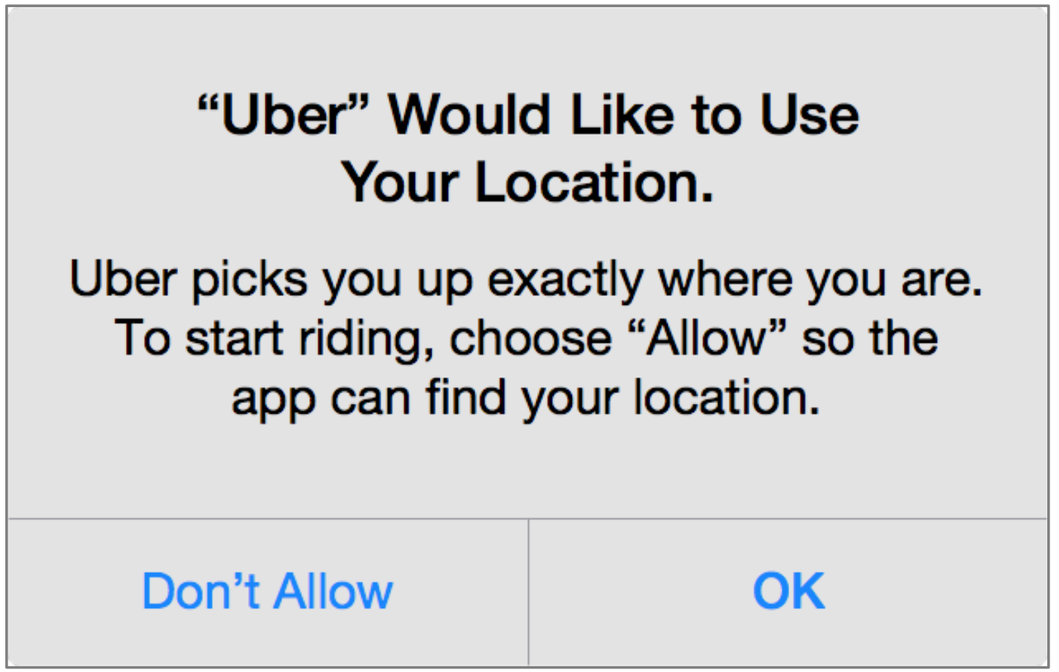
Carnegie Mellon University

who: Camera app
what: location
why: to tag photos



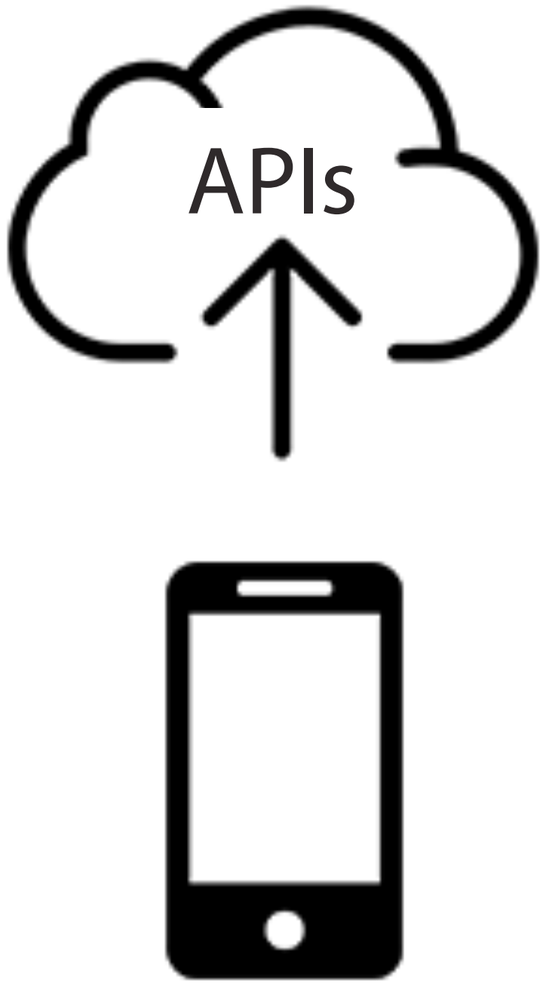
who: Uber
what: location
why: to locate pickup location

who request what data, and why



These descriptions are only shown at the user interface layer and can be arbitrary text.

No way to verify and not yet widely adopted.

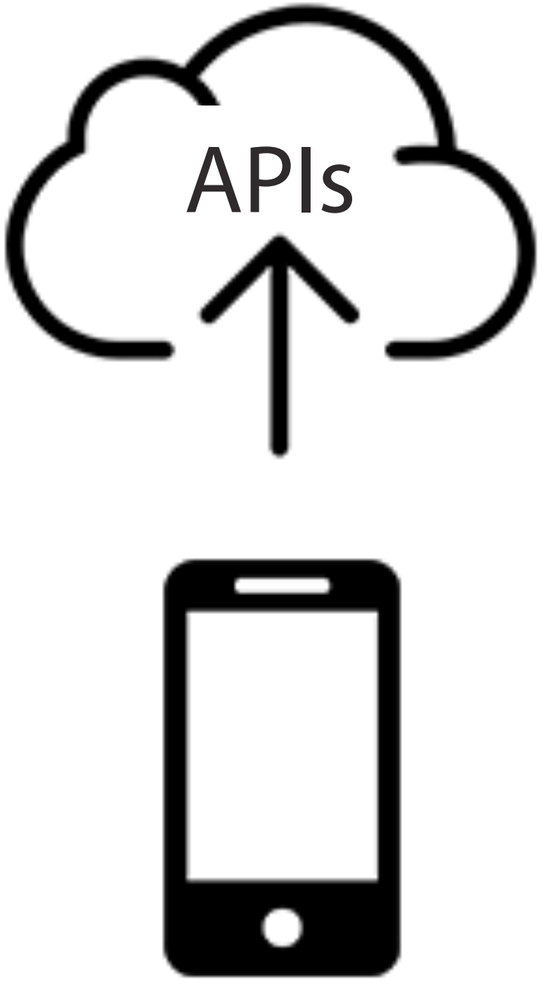


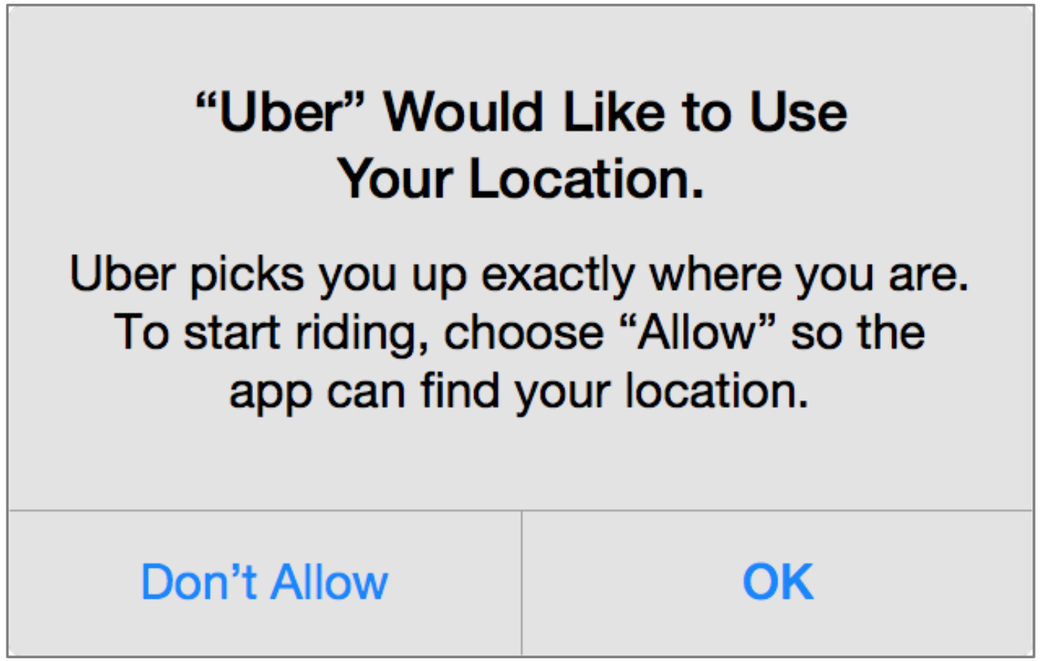
~~user interface layer, arbitrary text~~
~~no way to verify~~

when an app calls an API and post data to **remote servers** over the network.

network perspective

Can we **index** the **privacy attributes** of each network request **similarly** as the permission dialog?





who, where, what, why

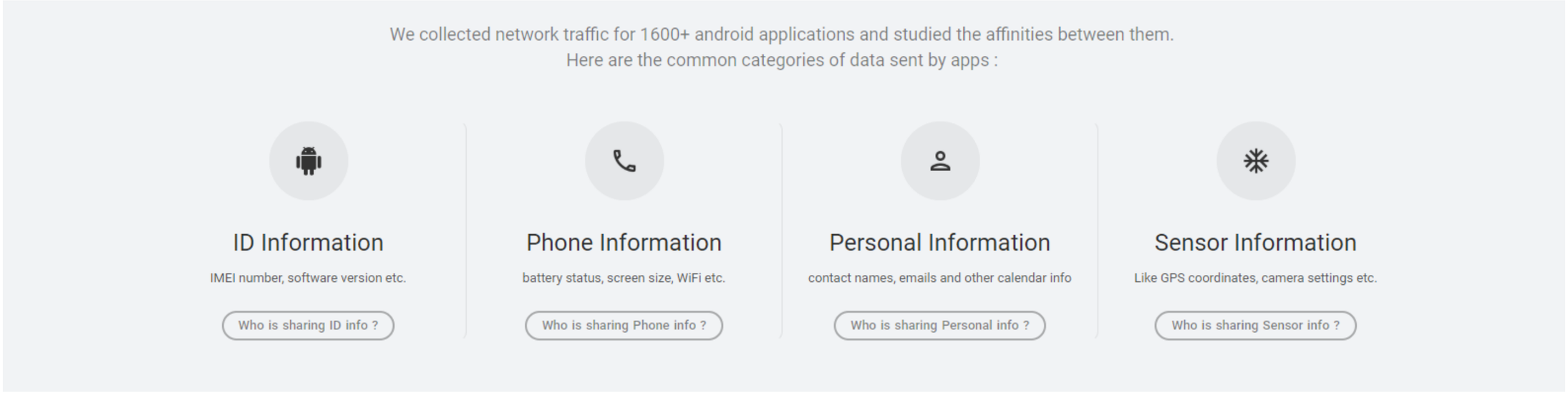
myLat: 40.4435877
myLon: -79.9452883

→ https://maps.google.com

Who (which app) sends the data?	Uber
Where the data is being sent to?	Google
What data is being collected?	Location
Why the data is being collected?	Map/navigation

Towards a public, large scale privacy database

to improve the transparency of mobile data collection



Related work

myLat: 40.4435877
myLon: -79.9452883

https://maps.google.com

State of the art [1, 2]

Who (which app) sends the data?	Uber
Where the data is being sent to?	Google
What data is being collected?	Location
Why the data is being collected?	Map/navigation

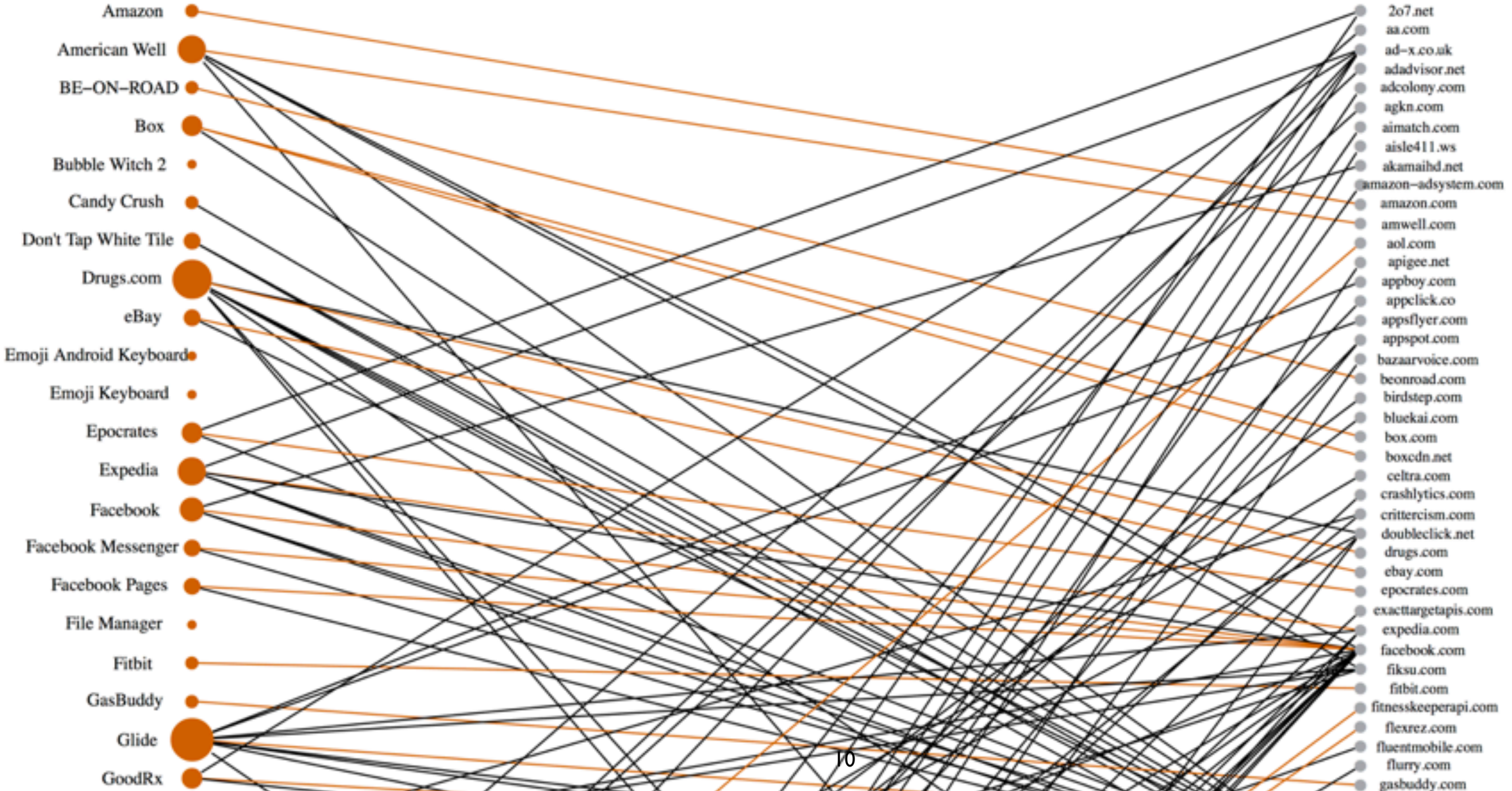
[1] Who Knows What About Me? A Survey of Behind the Scenes Personal Data Sharing to Third Parties by Mobile Apps

[2] ReCon: Revealing and Controlling PII Leaks in Mobile Network Traffic

Related work

Who Knows What About Me?

Zang et al. https://techscience.org/a/2015103001/



Related work

Who Knows What About Me?

https://techscience.org/a/2015103001/

Recruit ~10 participants
Install VPN on their phones
Test 10~100 apps
Raw data type (e.g., email addr.)

myLat: 40.4435877
myLon: -79.9452883

https://maps.google.com

State of the art

Who (which app) sends the data?	Uber
Where the data is being sent to?	Google
What data is being collected?	Location
Why the data is being collected?	Map/navigation

less explored.

Related work

b) Permissions + Purposes

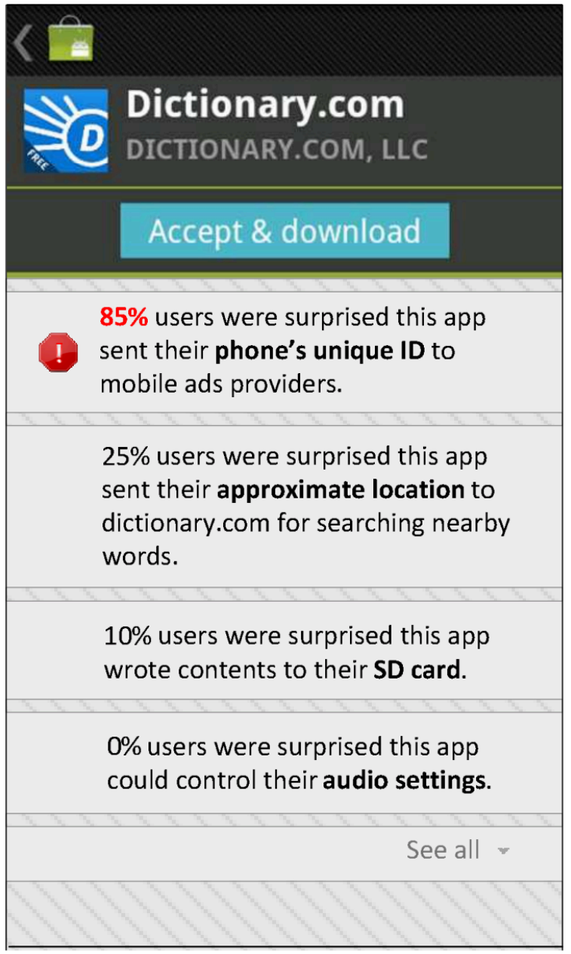
bbc iplayer
may request access to:

Location — Functionality, Marketing
- Access your approximate location (City/Town)
- Access your precise location via GPS and WiFi.

Personal Details — Functionality, Marketing
- Personal details about you, such as: age, sex, weight, height, or date of birth.

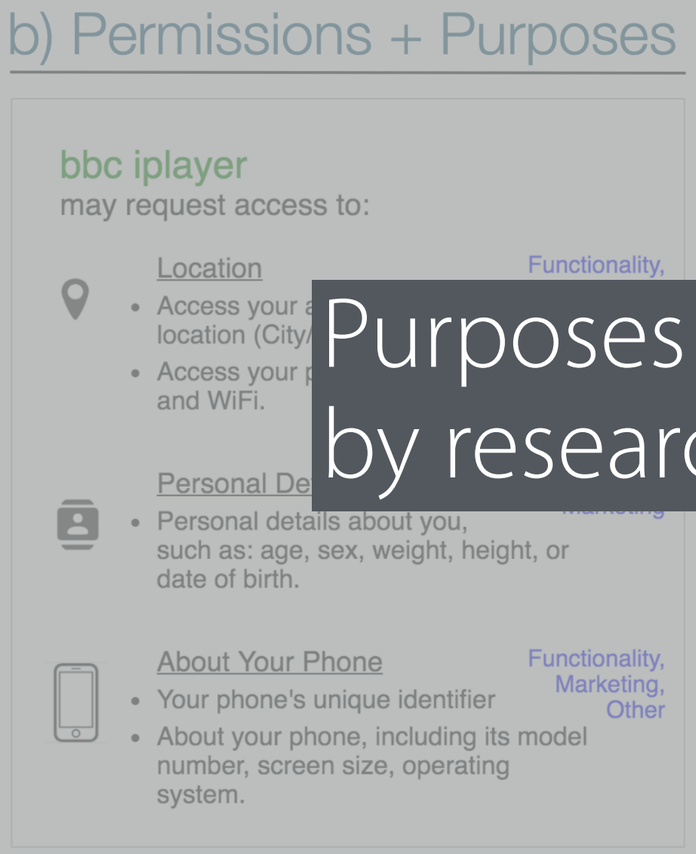
About Your Phone — Functionality, Marketing, Other
- Your phone's unique identifier
- About your phone, including its model number, screen size, operating system.

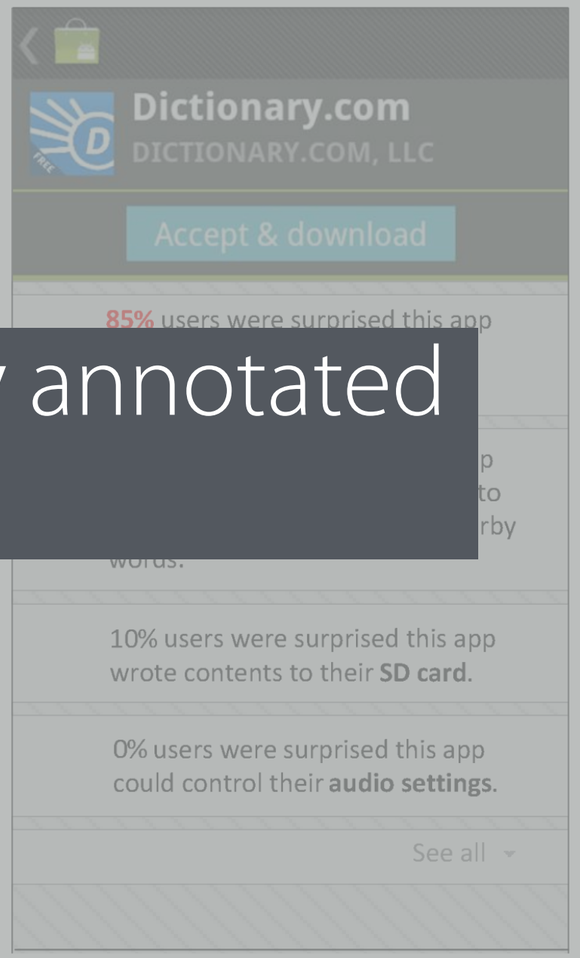
Exposing the Data Sharing Practices of Smartphone Apps [CHI' 17]



Expectation and Purpose [Ubicomp'12]

Related work

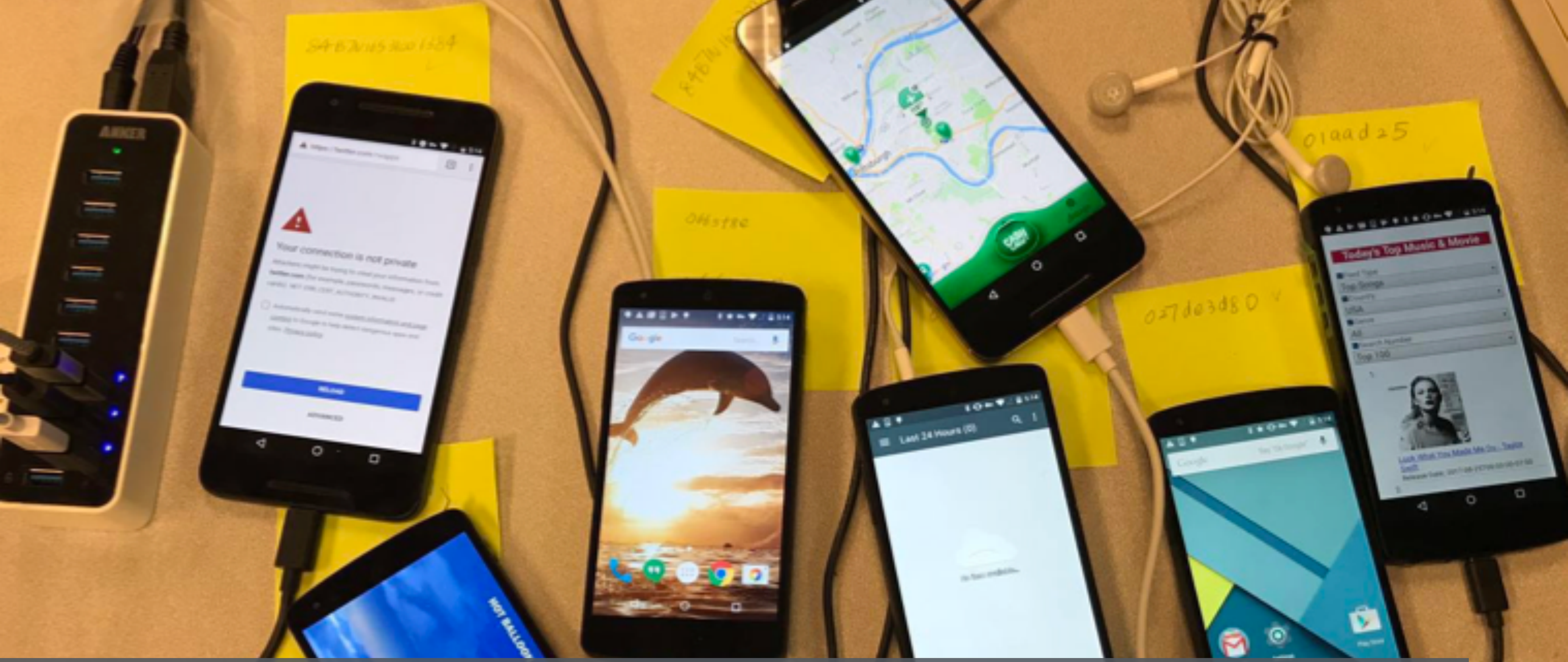




Purposes are **manually** annotated by researchers.

Exposing the Data Sharing Practices of Smartphone Apps [CHI' 17]

Expectation and Purpose [Ubicomp'12]



MobiPurpose is a **scalable in-lab** solution that can index **fine-grained** privacy attributes (**who**, **where**, **what**, **why**) of **outgoing** network requests.

3 modules

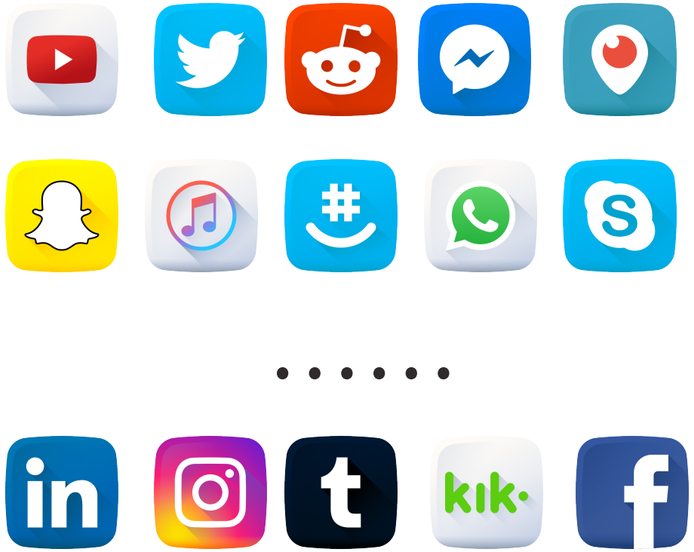
1 Scalable network tracing

2 Data types & purposes taxonomy

3 Automated Inference

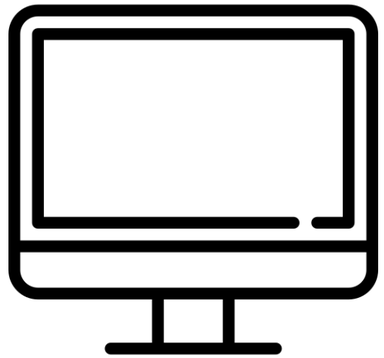
1 Network tracing

large scale network requests
at a low cost

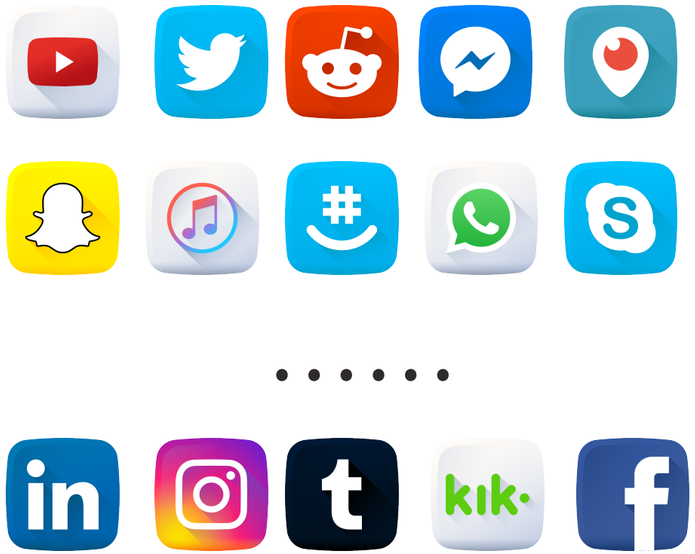


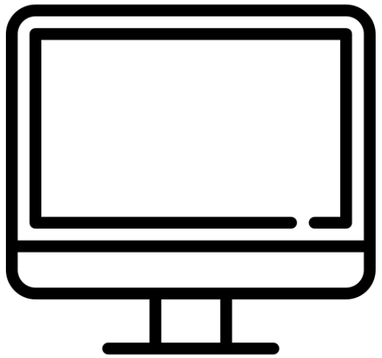
Google play

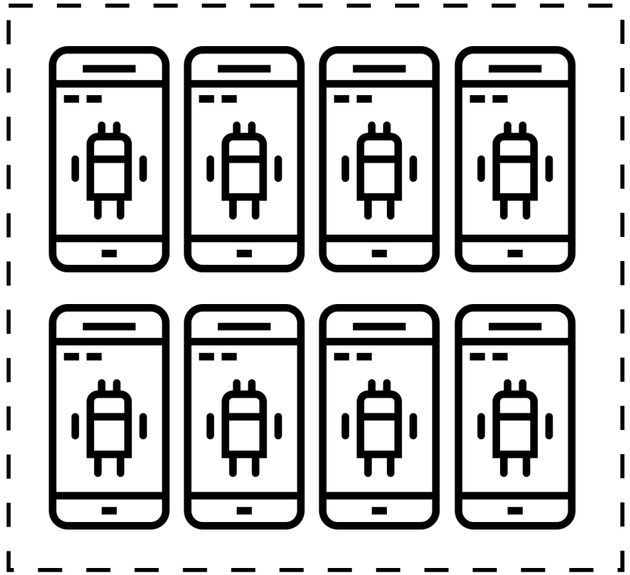
downloaded 185, 173 apps



Hardware & software setup (1)



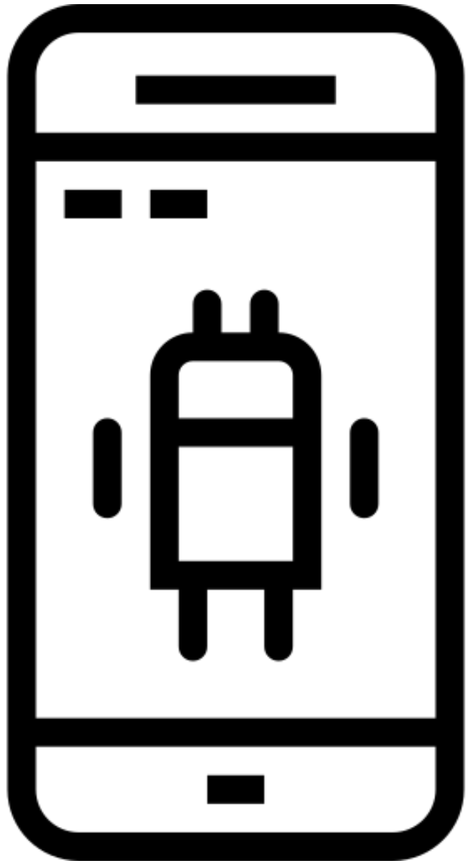




installed 30,075 apps

(due to OS compatibility, etc)

Hardware & software setup (2)



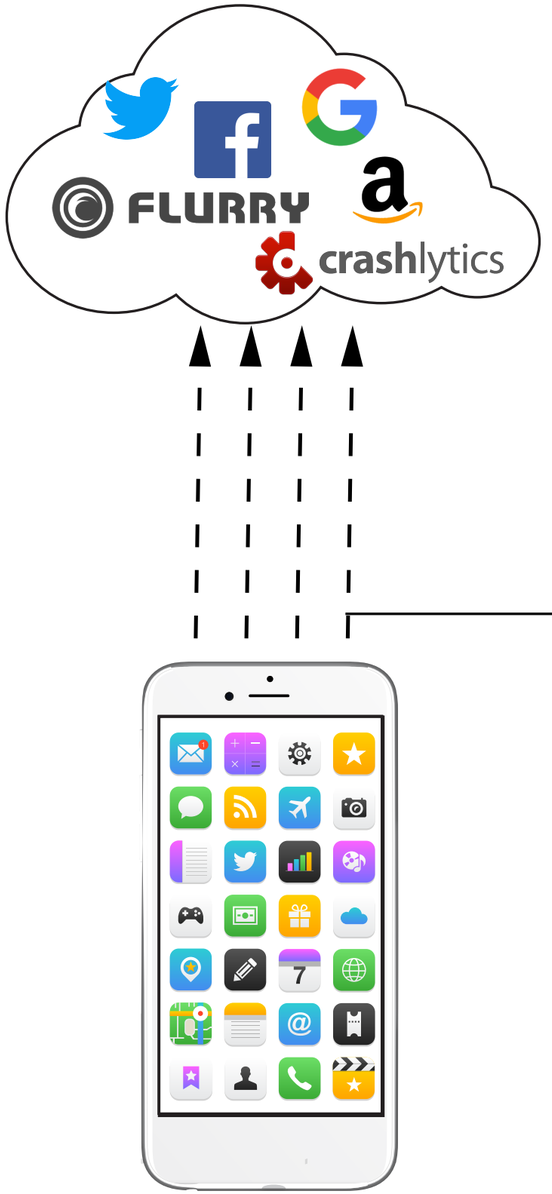
a men-in-the-middle VPN proxy app

3 minutes UI automation for each

running for 50 days

We open source the tools at:
http://bit.ly/mobipurpose

Hardware & software setup (3)



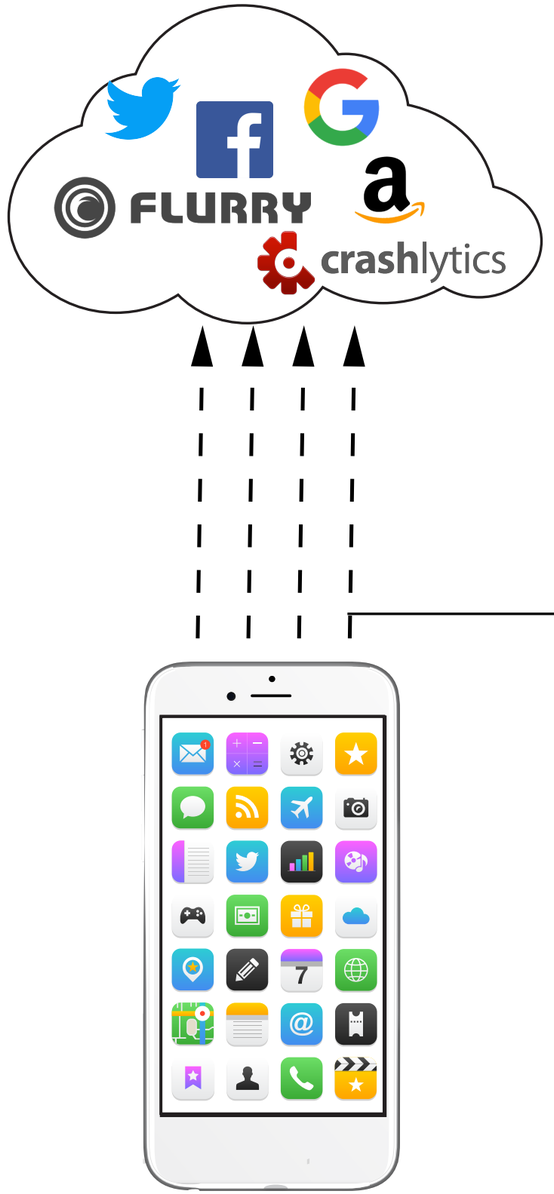
Traffic request snapshot

source app:
com.inkcreature.predatorfree
connect to host:
inkcreature.com
server path:
/_predatorServer/

key-value pairs in request body:
myLat: 40.4435877
myLon: -79.9452883
....

Raw Traffic Data

Traffic request snapshot



source app:
com.inkcreature.predatorfree
connect to host:
inkcreature.com
server path:
/_predatorServer/

key-value pairs in request body:
myLat: 40.4435877
myLon: -79.9452883
....

Who?

Where?

Key-value pairs

Raw Traffic Data

Traffic request snapshot

source app:
com.inkcreature.predatorfree
connect to host:
inkcreature.com
server path:
/_predatorServer/

key-value pairs in request body:
myLat: 40.4435877
myLon: -79.9452883
....

2,008,912 unique traffic requests

from 14,910 apps

contacting

12,046 unique domains

302,893 unique URLs

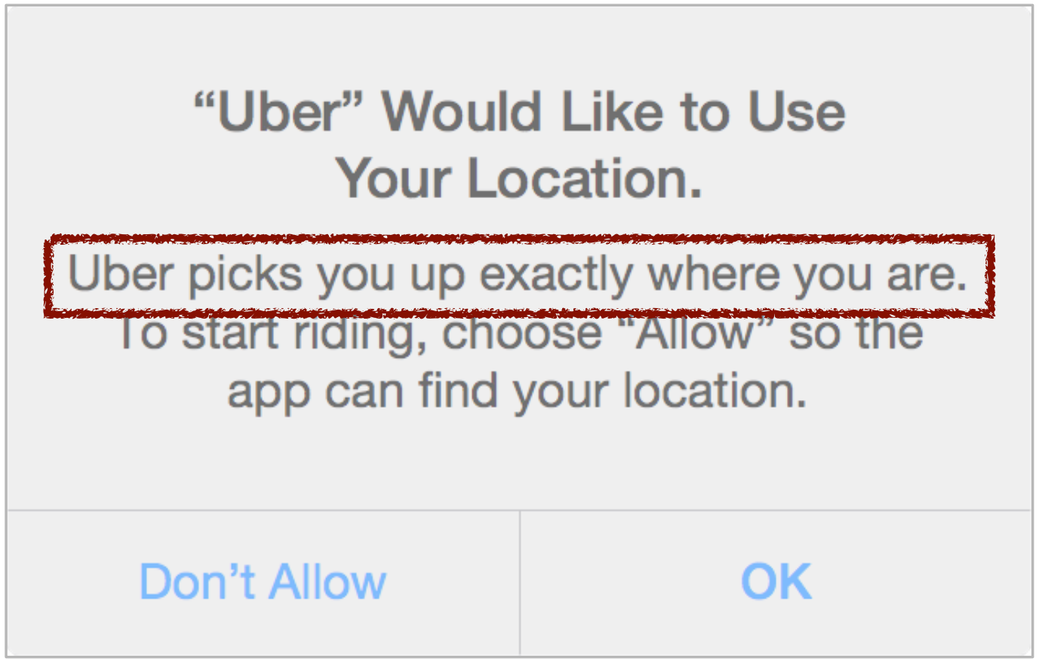
We publish the dataset at:
http://bit.ly/purposedata

Traffic Data stats

2 Taxonomy

define and categorize purposes

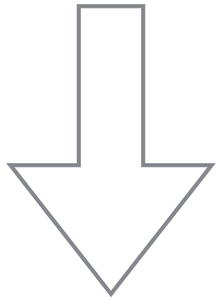
“**usage strings**” in iOS/Android



Arbitrary texts are hard to **aggregate**, **analyze** and **verify.**

- Many apps collect users' data for **similar** purposes.
- There are **enumerable** purposes.
  10-50 depends on the granularity.

~~generate text describing the purpose~~



build a taxonomy and classify the purpose

1 Comprehensive and extendable

covers the majority of use cases

2 Meaningful granularity

not too narrow nor too broad

3 Understandable

minimal explanation for dev and users

10 CS graduate students

categorizing 1000+ network requests

and 300+ permission usages

3 independent sessions

affinity diagram

Purpose at App level

why a user downloads the app (e.g., app categories - Game)

Purpose at Network level

why an app sends the request the app (e.g., library categories - Ad)

purpose granularity

Purpose at App level

why a user downloads the app (e.g., app categories - Game)

Purpose at Network level

why an app sends the request the app (e.g., library categories - Ad)

Purpose at Data level

why a developer collects the data (e.g., nearby search)

purpose granularity

Purpose at App level

why a user downloads the app (e.g., app categories - Game)

Purpose at Network level

why a app sends the request the app (e.g., library categories - Ad)

Purpose at Data level

why a developer collects the data (e.g., usage descriptions)

contains most privacy details, consistent with usage strings

purpose granularity

data types

location

data types

data purposes

examples

nearby search

location

data types

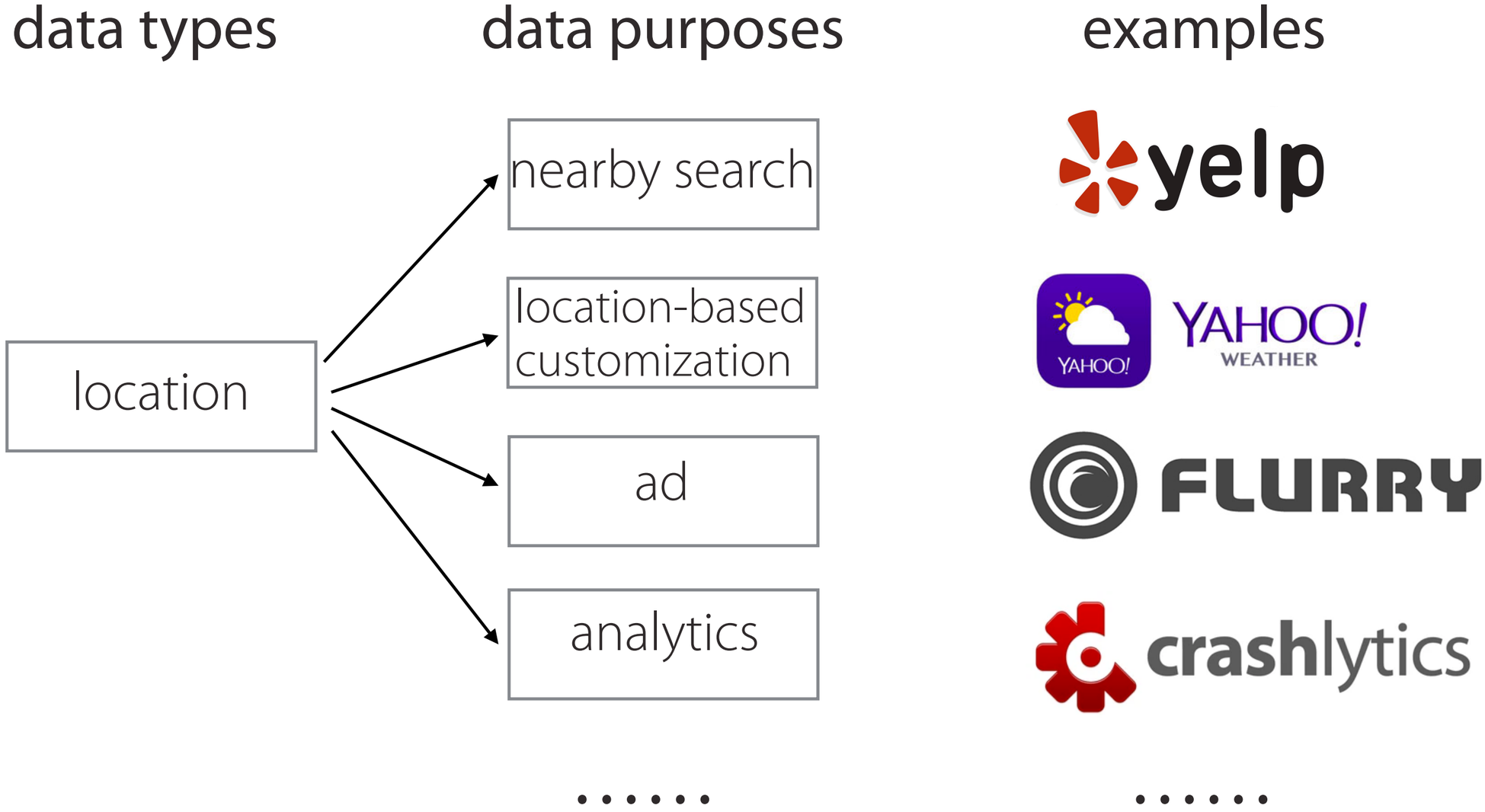
data purposes

examples

location

nearby search

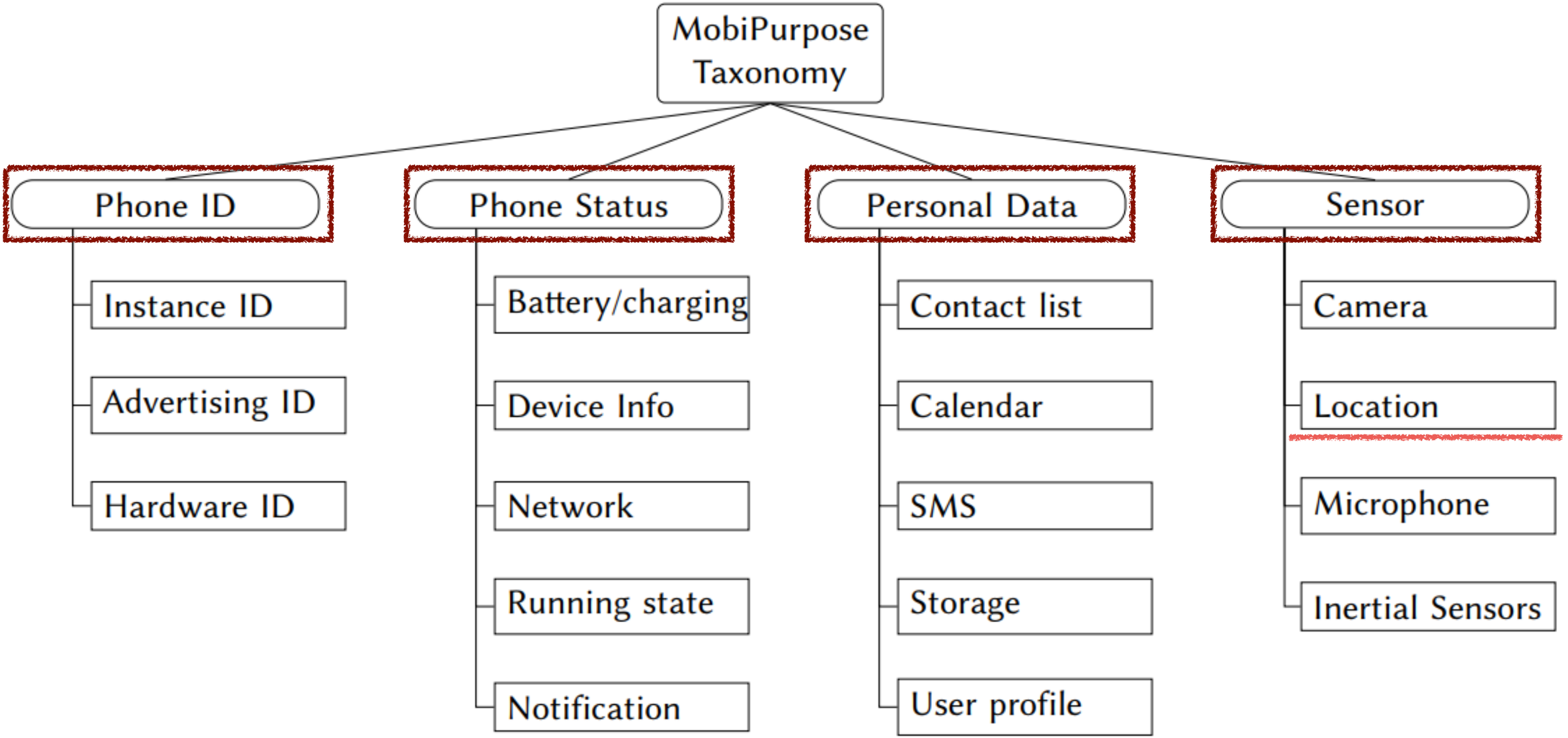
location-based customization


data types
data purposes
examples
location
nearby search
location-based customization
ad
analytics
.....
yelp
YAHOO!
YAHOO! WEATHER
FLURRY
crashlytics
.....

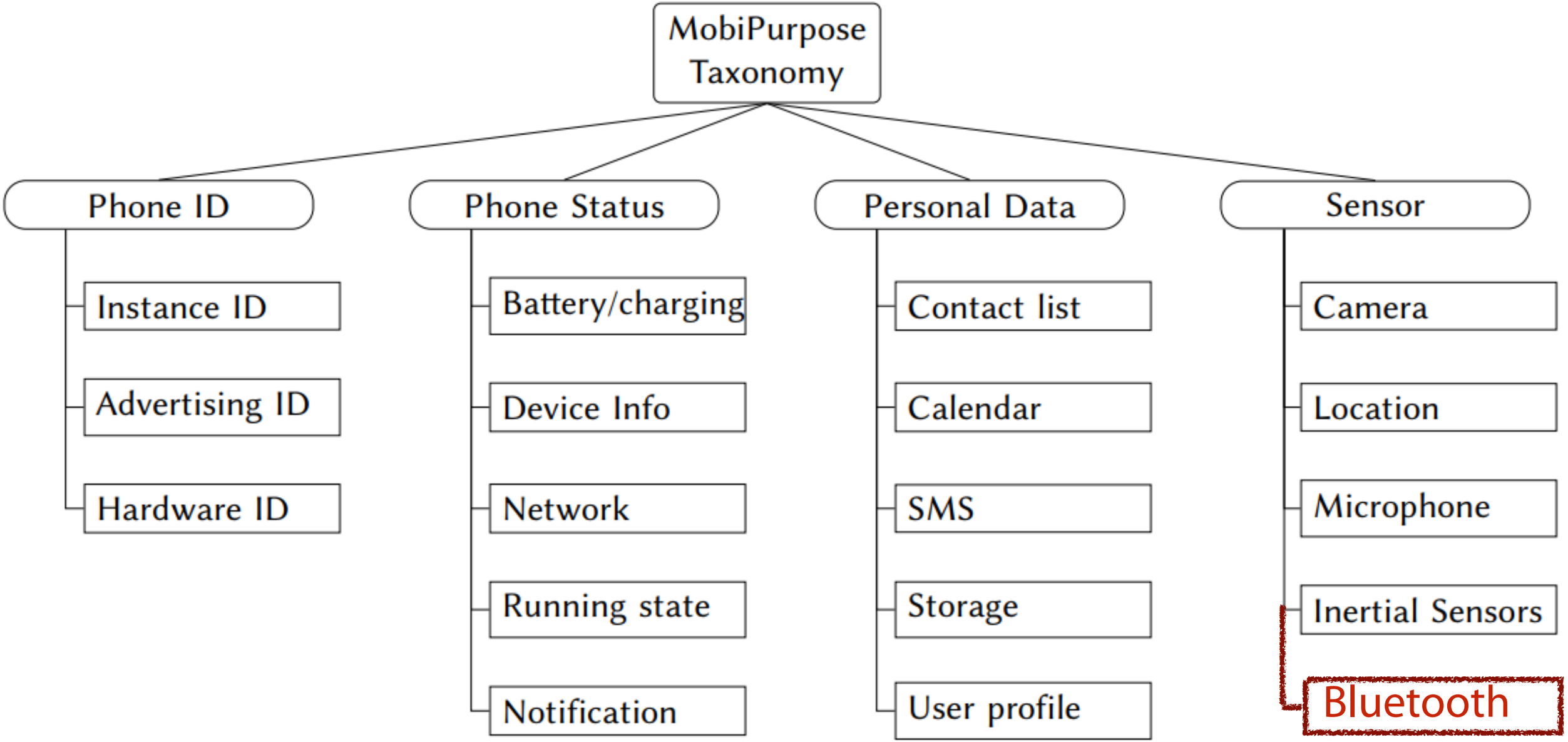
Data purposes for location data

Location[7]	Nearby Search	Search nearby POIs/real estates
	Location-based Customization	Fetch local weather/radio information
	Query Transportation Information	Estimate the trip time through Uber API
	Recording	Track the running velocity
	Map and Navigation	Find the user location in Map apps
	Geosocial Networking	Find nearby users in the social network
	Geotagging	Tag photos with locations
	Location Spoofing	Set up fake GPS locations
	Alert and Remind	Remind location-based tasks
	Location-based game	Play games require users' physical location
	Reverse geocoding	Use the GPS coords to find the real world address.
	Data collection for analytics	Collect data for marketing analysis
	Data collection for ad	Collect data for ad personalization

See the complete taxonomy at:
http://bit.ly/mobitaxonomy


MobiPurpose Taxonomy
Phone ID
Instance ID
Advertising ID
Hardware ID
Phone Status
Battery/charging
Device Info
Network
Running state
Notification
Personal Data
Contact list
Calendar
SMS
Storage
User profile
Sensor
Camera
Location
Microphone
Inertial Sensors

data types



extensibility

3 Automated inference

Traffic request snapshot

source app:
com.inkcreature.predatorfree
connect to host:
inkcreature.com
server path:
/_predatorServer/

key-value pairs in request body:
myLat: 40.4435877
myLon: -79.9452883
....

input

What data is being collected?

Why the data is being collected?

output

1 Self-explainable patterns

userAdvertisingId : 901e3310-3a26-487e-83c7-2fa26ac2786c

advertising, Id

machine generated UUID

http://reports.crashlytics.com

report, crash, analytics

1 Self-explainable patterns

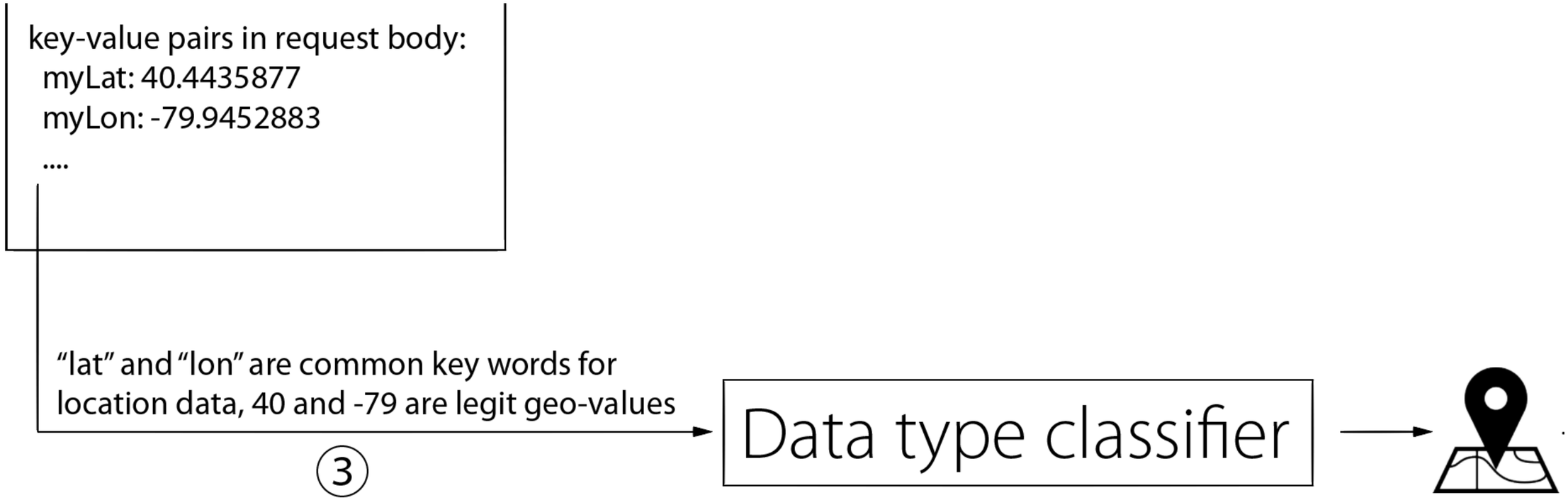
2 External knowledge (app type, server domain)

a game app sends location data to http://admob.com

A mobile ad company

intuitions

a **bootstrapping** method to predict the data type



data type inference

taxonomy lookup to get
the purpose candidates

purposes candidates

search nearby
location-based customization
transportation information
recording
map/navigation
geosocial networking
geotagging
location spoofing
alert and remind
location-based game
reverse geocoding
advertising
analytics

taxonomy lookup

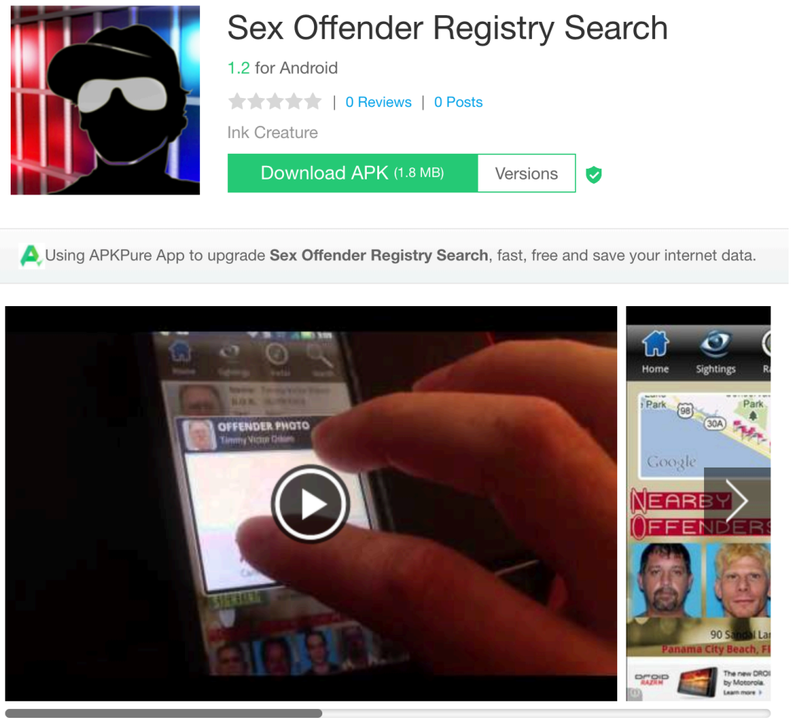
Traffic request snapshot

source app:
com.inkcreature.predatorfree
connect to host:
inkcreature.com
server path:
/_predatorServer/

key-value pairs in request body:
myLat: 40.4435877
myLon: -79.9452883
....

Source app feature

predator is an offender registry search app



purpose features

Traffic request snapshot

source app:
com.inkcreature.predatorfree
connect to host:
inkcreature.com
server path:
/_predatorServer/

key-value pairs in request body:
myLat: 40.4435877
myLon: -79.9452883
....

Source app feature

predator is an offender registry search app

Textual feature

the app sends data to its own server

purpose features

Traffic request snapshot

source app:
com.inkcreature.predatorfree
connect to host:
inkcreature.com
server path:
/_predatorServer/

key-value pairs in request body:
myLat: 40.4435877
myLon: -79.9452883
....

Source app feature

predator is an offender registry search app

Textual feature

the app sends data to its own server

Domain feature

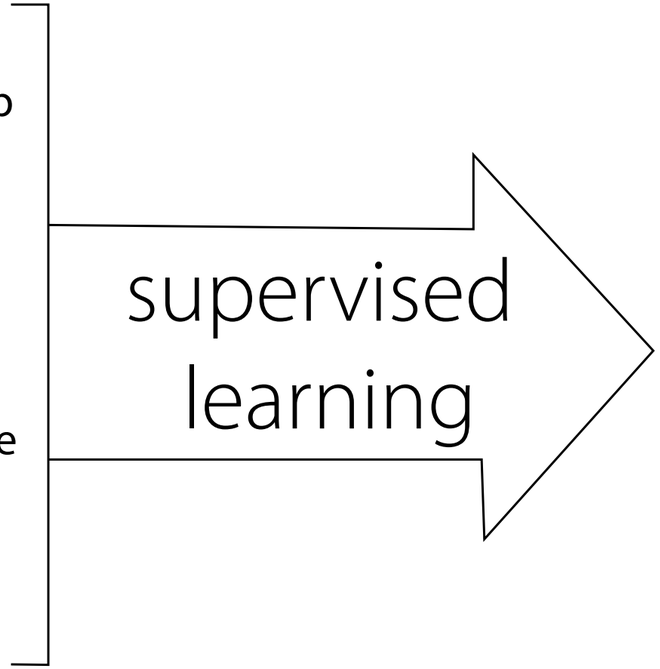
- company business type (Crunchbase)
- decompile app files to mine the domain references

purpose features

source app feature:
predator is an offender registry search app

textual feature:
the app sends data to its own server

domain feature:
- company business type from Crunchbase
- decompile large scale app files to mine the domain references

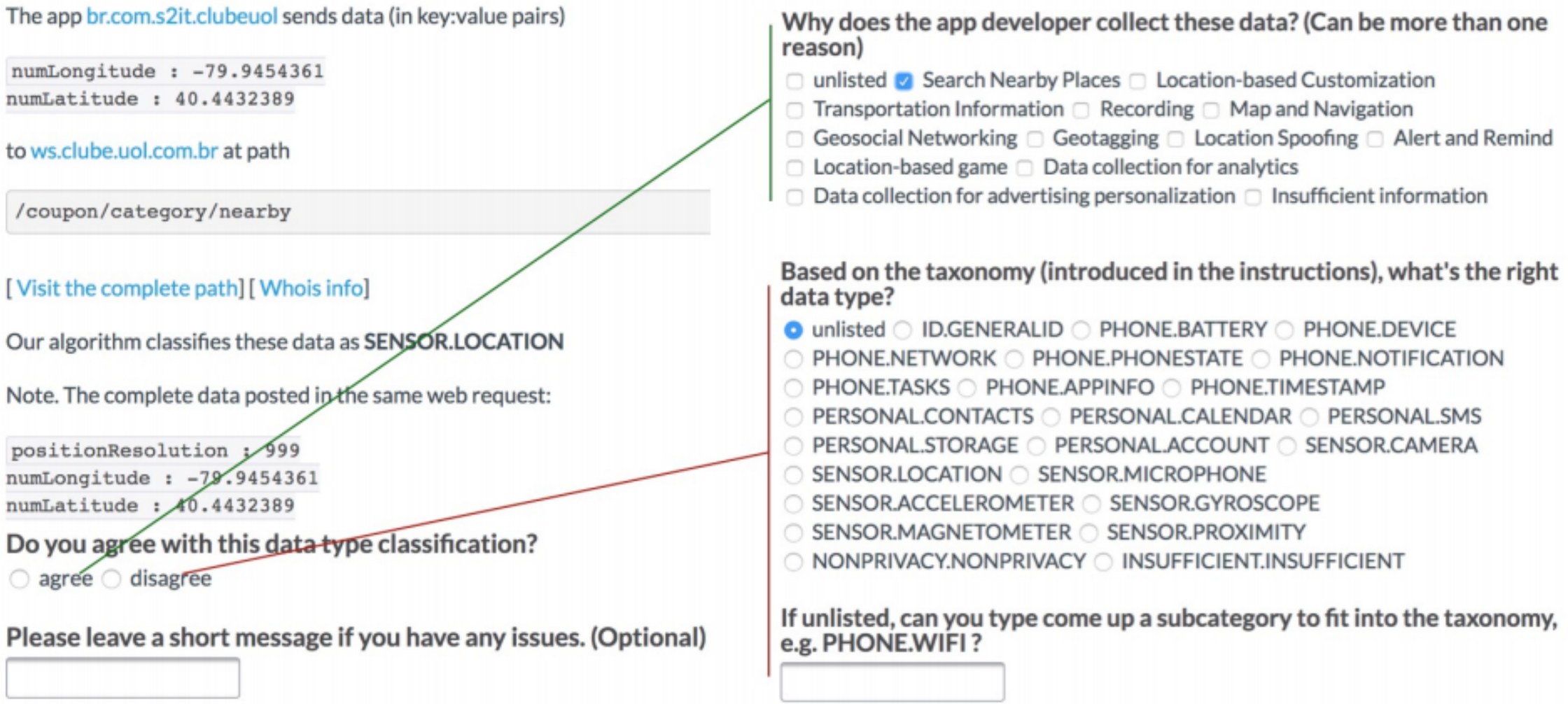


probability	purposes candidates
0.72	**search nearby**
0.2	location-based customization
0.03	transportation information
0.02	recording
0.02	map/navigation
0.01	geosocial networking
0	geotagging
0	location spoofing
0	alert and remind
0	location-based game
0	reverse geocoding
0	advertising
0	analytics

supervised machine learning

Evaluation

accuracy & recall



Labeling ”what” & “why” in each traffic request.
Each request has been labeled by three people.

data set labeling

1059 traffic requests in total
across **7** data categories

consensus on **98%** data type labels,
and **88%** of purpose labels.

method: 10-fold cross validation

Data type inference:

Overall precision of **95.9%**
precision above **93%** for all **7** classes

	ID	Battery	Device	Network	State	Account	Location	**Macro-avg**	**Micro-avg**
Precision	97%	100%	93%	94%	100%	96%	96%	95.6%	95.9%
Recall	86.8%	100%	87.5%	92.1%	100%	92.3%	95.2%	90.8%	89.9%
F-score	0.925	1.00	0.902	0.930	1.00	0.941	0.956	0.931	0.928

results

Data purpose inference:

Overall precision of **84%** for
19 unique categories

		P1	P2	P3	P4	P5	Total
Anti-fraud	P1	26	-	-	1	4	31
Authentication	P2	-	16	1	3	7	27
Personalization	P3	3	1	8	1	11	24
Ad	P4	-	-	-	162	15	177
Analytics	P5	-	-	1	11	114	126

confusion matrix for ID

Data purpose inference:

Overall precision of **84%** for
19 unique categories

		P1	P2	P3	P4	P5	Total
Anti-fraud	P1	26	-	-	1	4	31
Authentication	P2	-	16	1	3	7	27
Personalization	P3	3	1	8	1	11	24
Ad	P4	-	-	-	162	15	177
Analytics	P5	-	-	1	11	114	126

confusion matrix for ID purposes

See more details
in the paper.

Privacy Analytics for Smartphones

We collected network traffic for 1600+ android applications and studied the affinities between them.
Here are the common categories of data sent by apps :

ID Information

IMEI number, software version etc.

Who is sharing ID info ?

Phone Information

battery status, screen size, WiFi etc.

Who is sharing Phone info ?

Personal Information

contact names, emails and other calendar info

Who is sharing Personal info ?

Sensor Information

Like GPS coordinates, camera settings etc.

Who is sharing Sensor info ?

Privacy Analytics for Smartphones

Beta web: http://bit.ly/mobipurposeweb

We collected network traffic for 1600+ android applications and studied the affinities between them.
Here are the common categories of data sent by apps :

ID Information

IMEI number, software version etc.

Who is sharing ID info ?

Phone Information

battery status, screen size, WiFi etc.

Who is sharing Phone info ?

Personal Information

contact names, emails and other calendar info

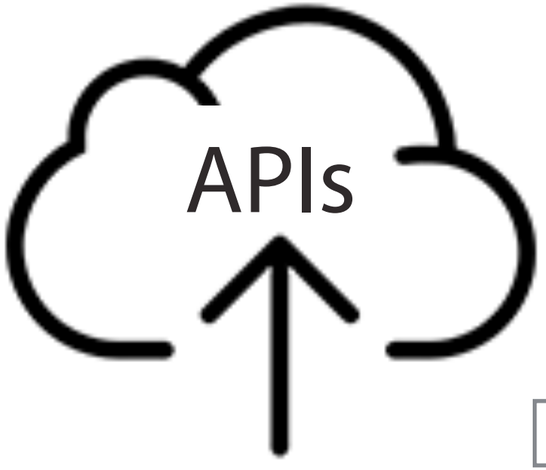
Who is sharing Personal info ?

Sensor Information

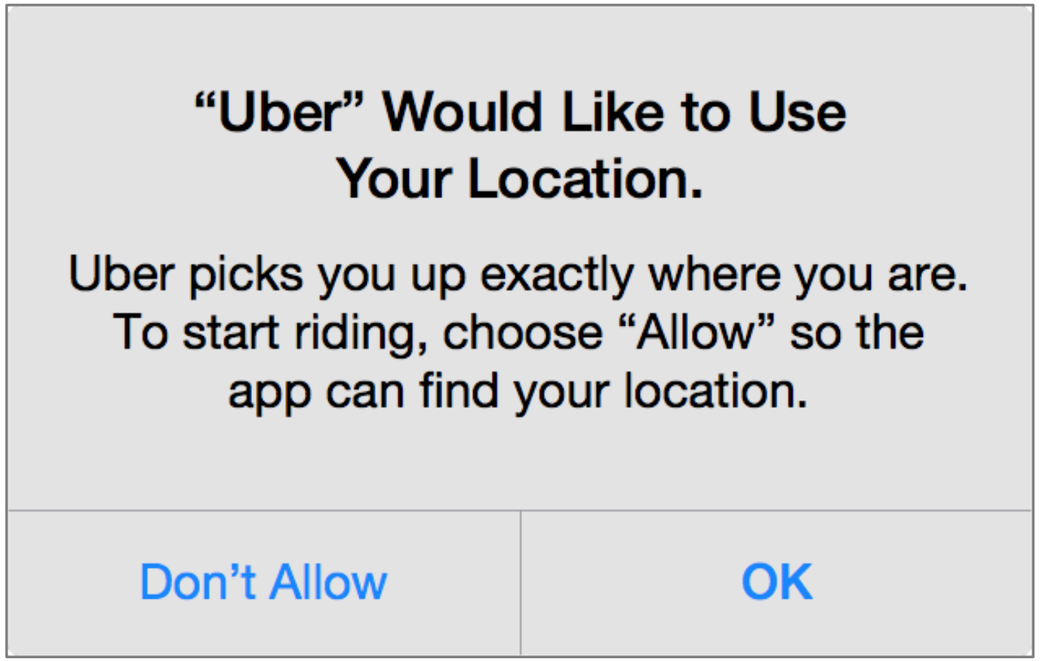
Like GPS coordinates, camera settings etc.

Who is sharing Sensor info ?

Make privacy a **native** feature by inspecting network requests



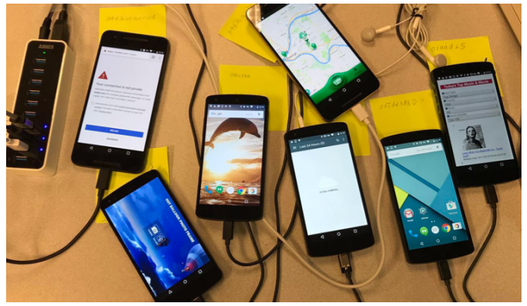
MobiPurpose



who, where, what, why

1 Network tracing tools

http://bit.ly/mobipurposetool



2 Traffic requests data set

http://bit.ly/mobipurposedata

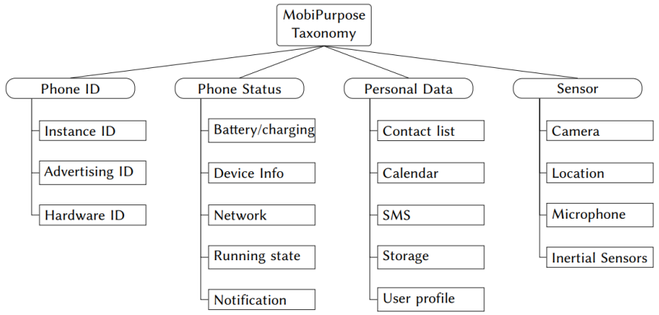
Traffic request snapshot

```
source app:
    com.inkcreature.predatorfree
connect to host:
    inkcreature.com
server path:
    /_predatorServer/

key-value pairs in request body:
  myLat: 40.4435877
  myLon: -79.9452883
  ....
```

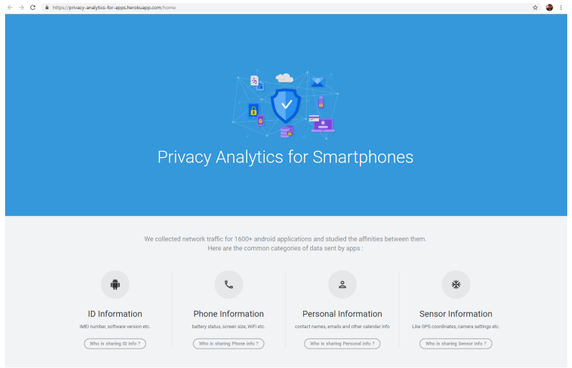
3 Data type & purpose taxonomy

http://bit.ly/mobitaxonomy



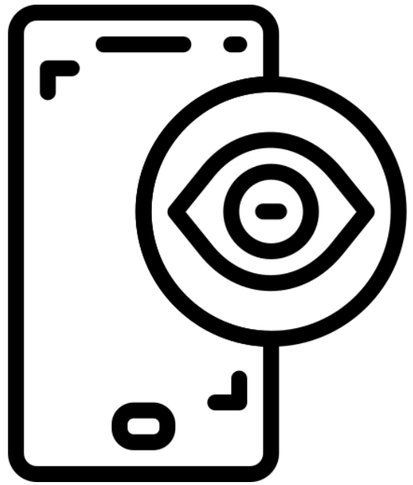
4 Beta web

http://bit.ly/mobipurposeweb



Outreach

Inferring the Purposes of Network Traffic in Mobile Apps



Who (which app) sends the data?

Where the data is being sent to?

What data is being collected?

Why the data is being collected?

Haojian Jin (haojian@cs.cmu.edu)

Carnegie Mellon University

CHIMPS LAB

synerqy
systems, networking and energy efficiency

Backup slides

Approximate Information Flows: Socially-based Modeling of Privacy in Ubiquitous Computing

Xiaodong Jiang, Jason I. Hong, James A. Landay

Group for User Interface Research
Computer Science Division
University of California, Berkeley
Berkeley, CA 94720-1776, USA
{xdjiang, jasonh, landay@cs.berkeley.edu}

Abstract. In this paper, we propose a framework for supporting socially-compatible privacy objectives in ubiquitous computing settings. Drawing on social science research, we have developed a key objective called the *Principle of Minimum Asymmetry*, which seeks to minimize the imbalance between the people about whom data is being collected, and the systems and people that collect and use that data. We have also developed *Approximate Information Flow* (AIF), a model describing the interaction between the various actors and personal data. AIF effectively supports varying degrees of asymmetry for ubicomp systems, suggests new privacy protection mechanisms, and provides a foundation for inspecting privacy-friendliness of ubicomp systems.

Approximate information flow

PRIVACY AS CONTEXTUAL INTEGRITY

Helen Nissenbaum*

Abstract: The practices of public surveillance, which include the monitoring of individuals in public through a variety of media (e.g., video, data, online), are among the least understood and controversial challenges to privacy in an age of information technologies. The fragmentary nature of privacy policy in the United States reflects not only the oppositional pulls of diverse vested interests, but also the ambivalence of unsettled intuitions on mundane phenomena such as shopper cards, closed-circuit television, and biometrics. This Article, which extends earlier work on the problem of privacy in public, explains why some of the prominent theoretical approaches to privacy, which were developed over time to meet traditional privacy challenges, yield unsatisfactory conclusions in the case of public surveillance. It posits a new construct, "contextual integrity," as an alternative benchmark for privacy, to capture the nature of challenges posed by information technologies. Contextual integrity ties adequate protection for privacy to norms of specific contexts, demanding that information gathering and dissemination be appropriate to that context and obey the governing norms of distribution within it. Building on the idea of "spheres of justice," developed by political philosopher Michael Walzer, this Article argues that public surveillance violates a right to privacy because it violates contextual integrity; as such, it constitutes injustice and even tyranny.

Contextual Integrity

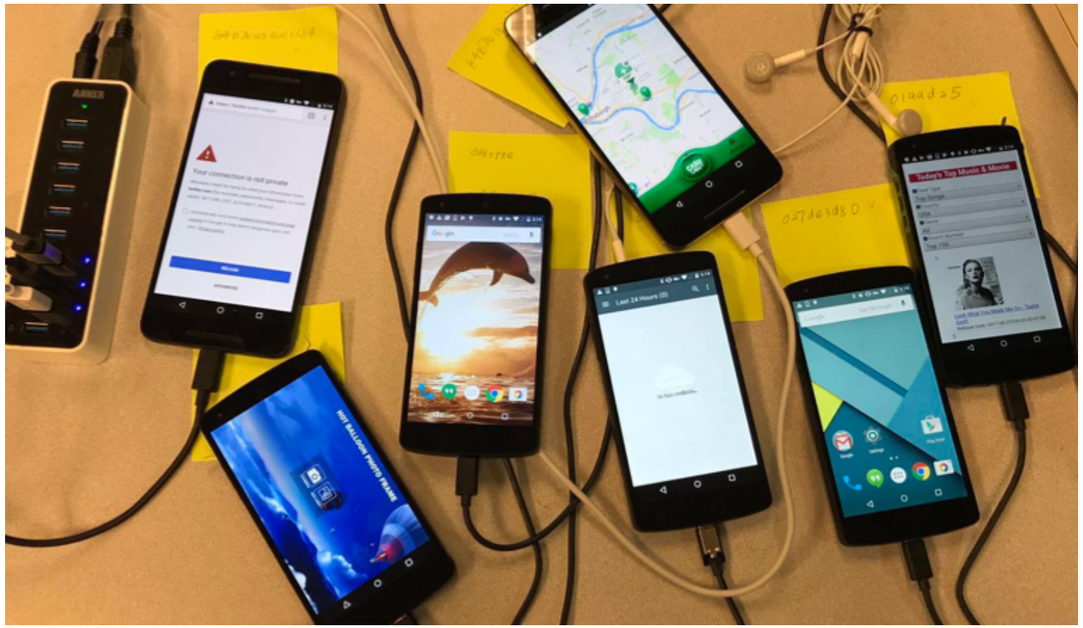
Privacy frameworks

recruiting participants

+ proportional to real usages

- not scalable

- may not be ethical



in-lab devices

- not proportional

+scalable

+ cheap

user study v.s. in-lab devices