# Mining Apps to Learn Normal Behavior

Andreas Zeller
Saarland University, Saarbrücken, Germany

Joint work with Alessandra Gorla, Ilaria Tavecchia, Vitalii Avdiienko, Konstantin Kuznetsov, and Florian Gross


Saarbrücken

# Saarbrücken

**1700**
BSc + MSc students

**375**
PhD students

**200**
Researchers (post PhD)

**8**
New buildings since 2001

**7**
ERC Grant holders

**6**
Leibniz Awardees

**4**
ACM Fellows

**1**
Software Engineer

UNIVERSITÄT DES SAARLANDES | max planck institut informatik | Max Planck Institute for Software Systems | DFKI | intel Visual Computing Institute | CISPA Center for Information Security, Privacy and Accountability

# Specifications

$$
\begin{array}{l}
\textbf{Schema } \mathit{removeChild} \\
\hline
\Delta \mathit{XMLElement} \\
\mathit{child?} : \mathit{XML\_ELEMENT} \\
\hline
\mathit{child?} \in \mathit{enumerateChildren} \\
\mathit{child?} \neq \mathit{null} \\
\mathit{enumerateChildren}' = \mathit{enumerateChildren} \setminus \mathit{child?} \\
\mathit{getChildrenCount}' = \mathit{getChildrenCount} - 1 \\
\hline
\end{array}
$$

# Specifications

*removeChild*

$\Delta XMLElement$
$child? : XML\_ELEMENT$

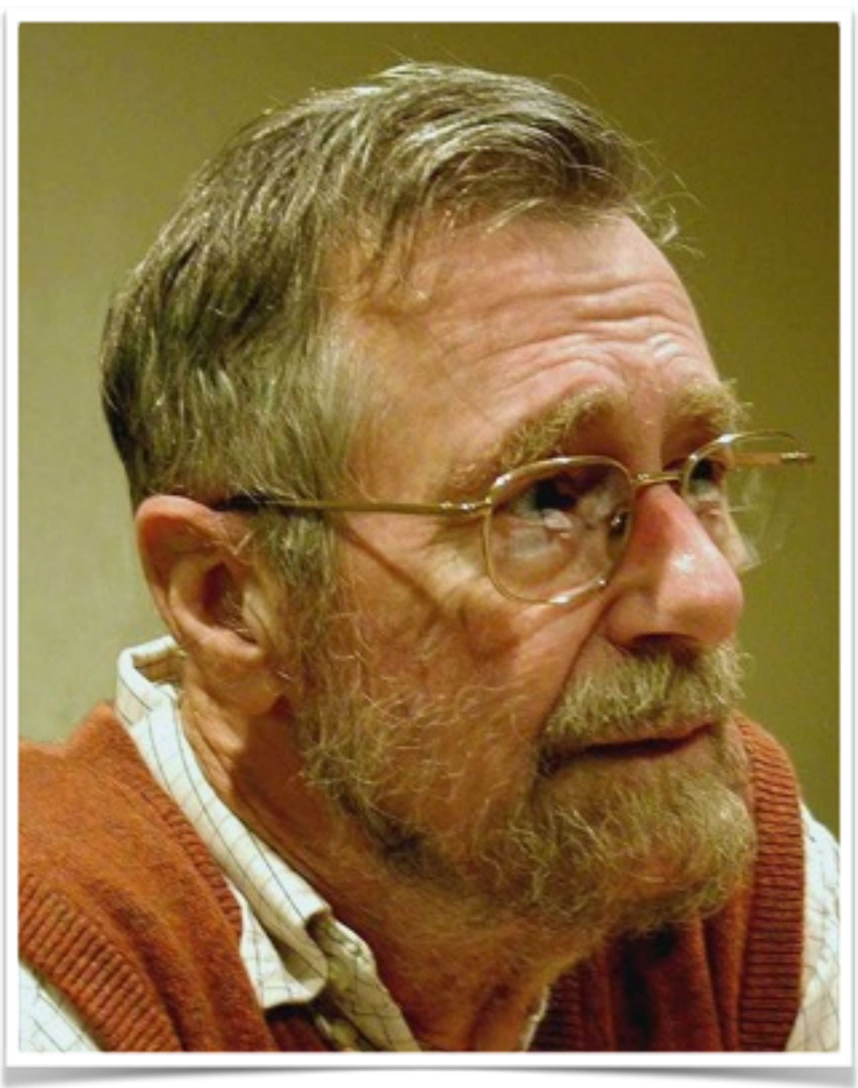
$child? \in enumerateChildren$
$child? \neq null$
$enumerateChildren' = enumerateChildren \setminus child?$
$getChildrenCount' = getChildrenCount - 1$

**fully automated testing**

**fully automated debugging**

**widely automated verification**

# Specifying Correctness

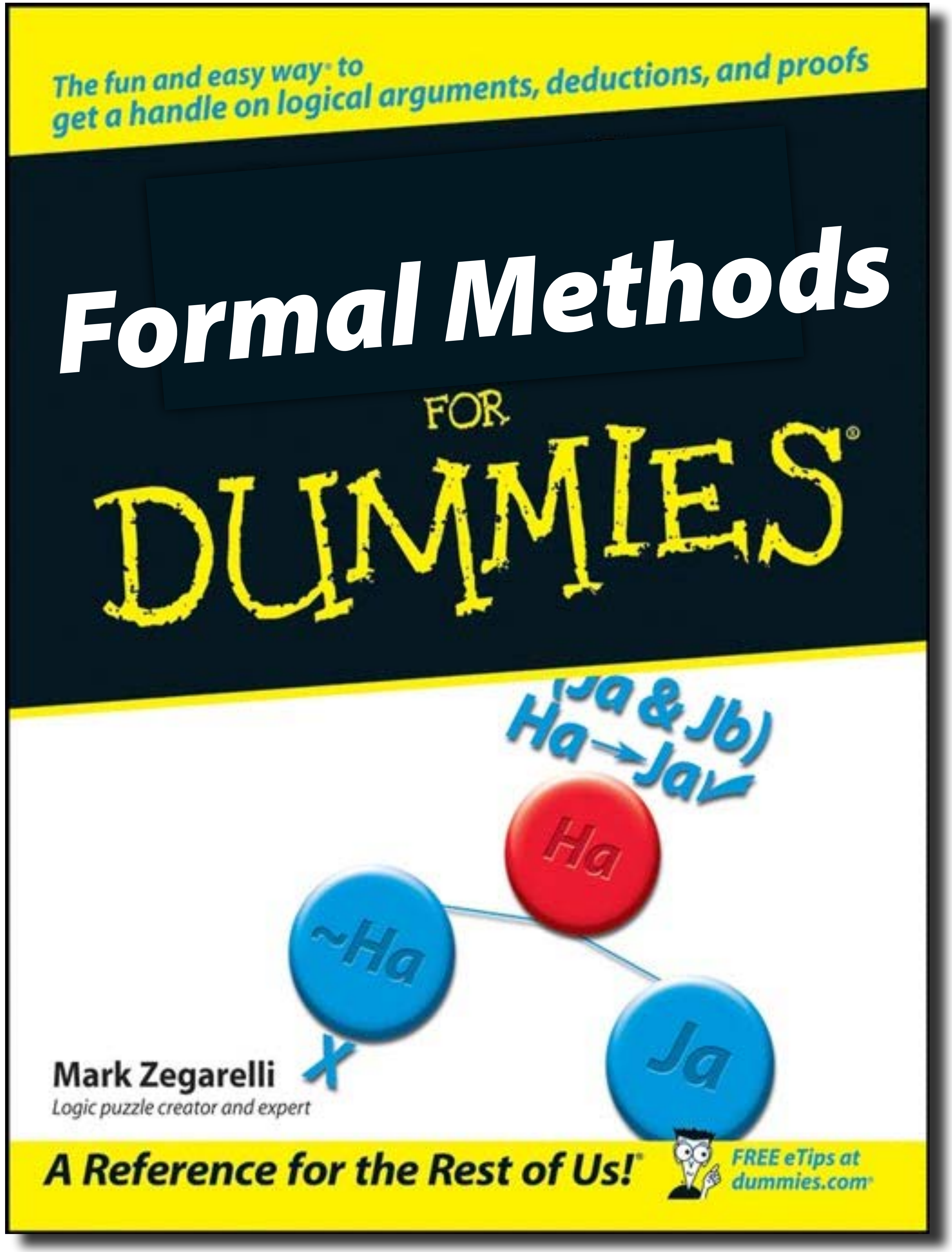


*removeChild*

$\Delta XMLElement$
$child? : XML\_ELEMENT$

---

$child? \in enumerateChildren$
$child? \neq null$
$enumerateChildren' = enumerateChildren \setminus child?$
$getChildrenCount' = getChildrenCount - 1$



The fun and easy way to get a handle on logical arguments, deductions, and proofs
Formal Methods
FOR
DUMMIES
(Ja & Jb)
Ha→Ja
Ha
~Ha
Ja
Mark Zegarelli
Logic puzzle creator and expert
A Reference for the Rest of Us!
FREE eTips at dummies.com


Microsoft Outlook
Unknown error
OK
Was this information helpful?

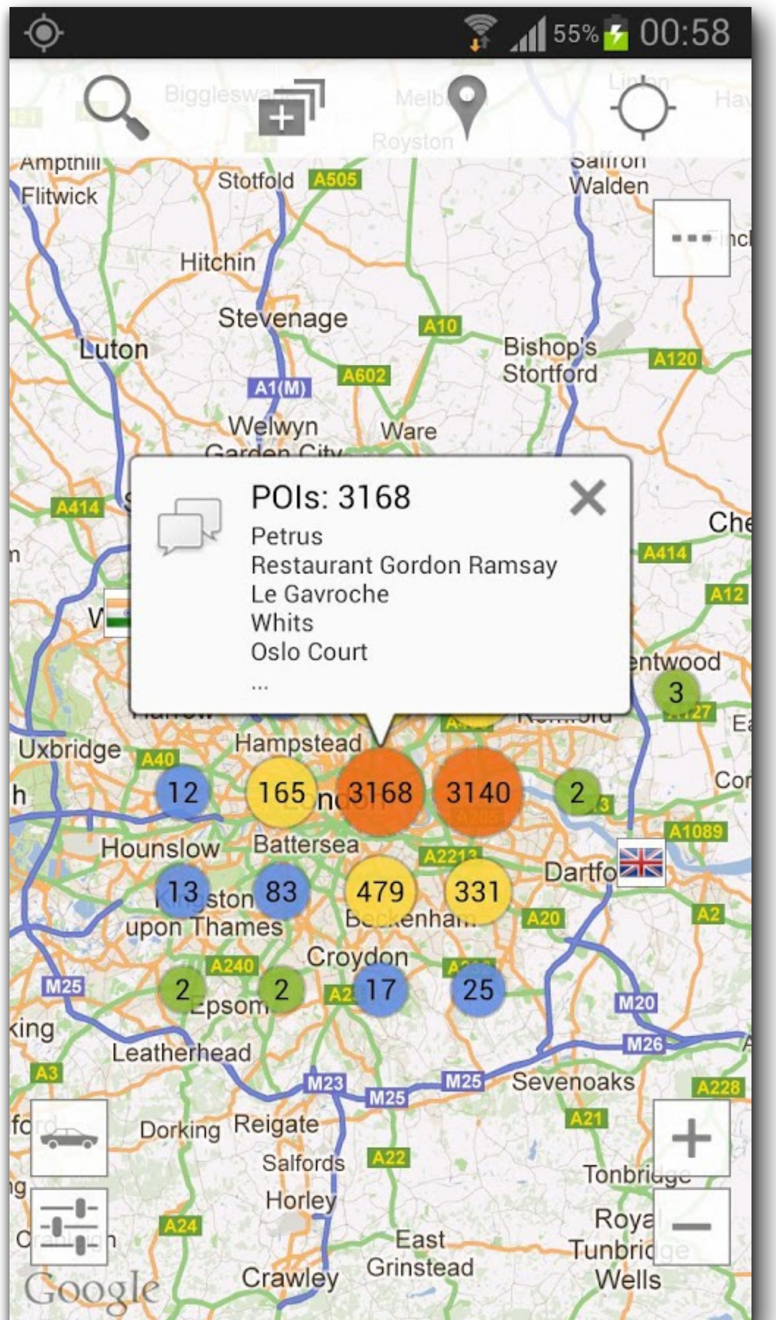
# Normality

# Mining Normality


$$
\begin{array}{l}
\textit{removeChild} \\
\hline
\Delta \textit{XMLElement} \\
\textit{child?} : \textit{XML\_ELEMENT} \\
\hline
\textit{child?} \in \textit{enumerateChildren} \\
\textit{child?} \neq \textit{null} \\
\textit{enumerateChildren}' = \textit{enumerateChildren} \setminus \textit{child?} \\
\textit{getChildrenCount}' = \textit{getChildrenCount} - 1 \\
\hline
\end{array}
$$

# Outliers


# London Restaurants



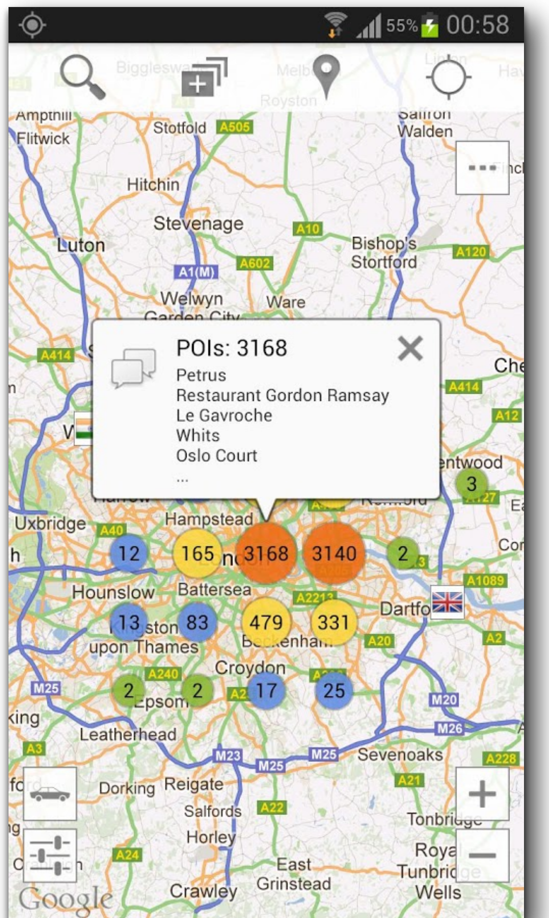

Looking for a restaurant, a bar, a pub or just to have fun in London? Search no more! This application has all the information you need:

• You can search for every type of food you want: french, british, chinese, indian etc.
• You can use it if you are in a car, on a bicycle or walking
• You can view all objectives on the map
• You can search objectives
• You can view objectives near you
• You can view directions (visual route, distance and duration)
• You can use it with Street View
• You can use it with Navigation

Keywords: london, restaurants, bars, pubs, food, breakfast, lunch, dinner, meal, eat, supper, street view, navigation

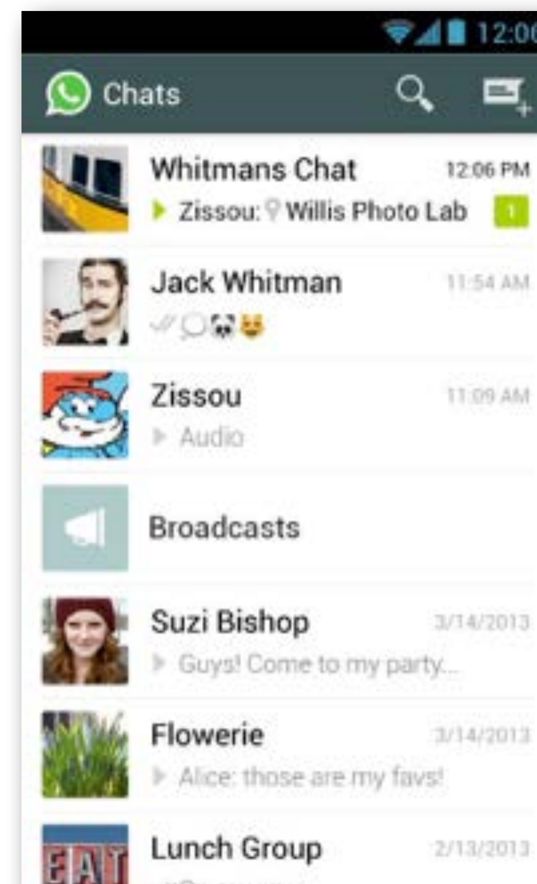
**Also sends out *account info***
**Also sends out *mobile phone number***
**Also sends out *your device ID***

# What is malicious?




London Restaurants

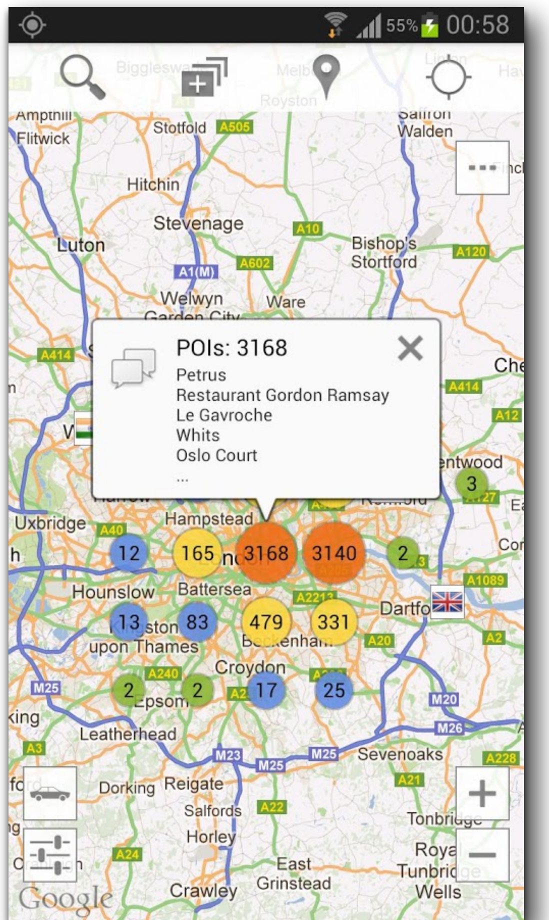
Also sends out *account info*
Also sends out *mobile phone number*
Also sends out *your device ID*




Also sends out account info
Also sends out mobile phone number
Also sends out your device ID

WhatsApp messenger

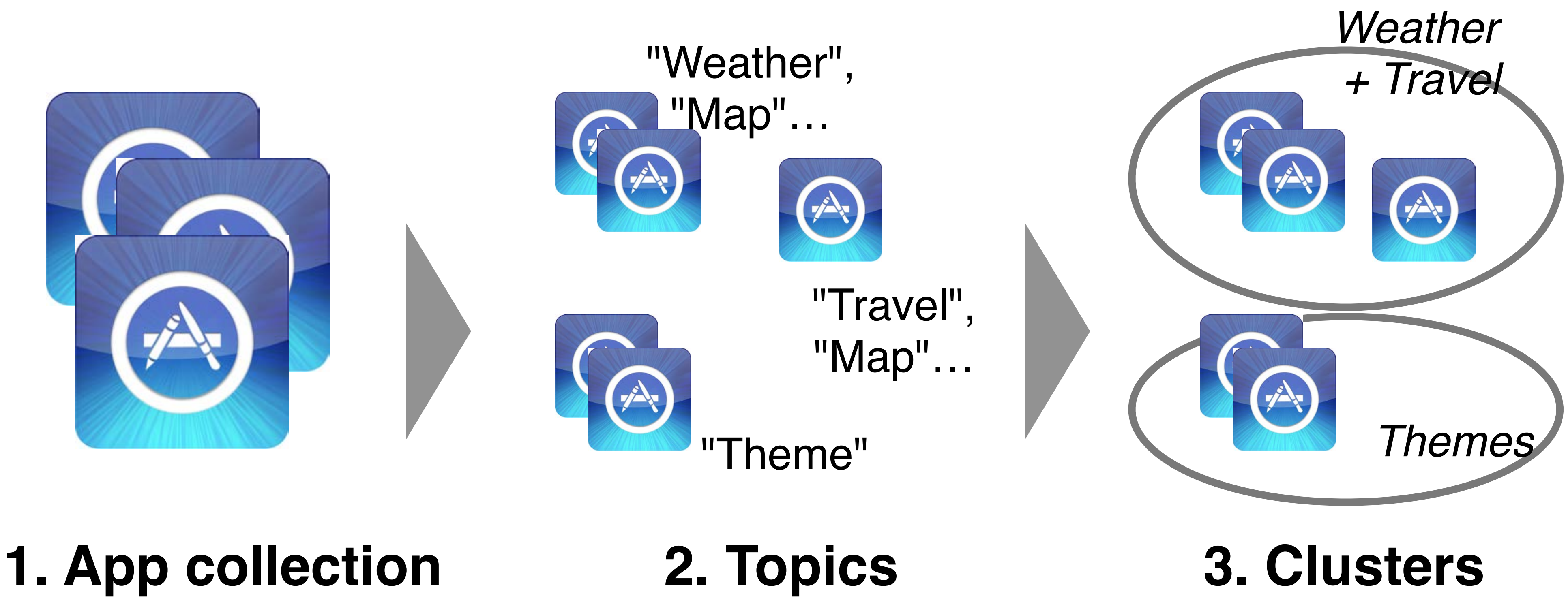
# What is normal?



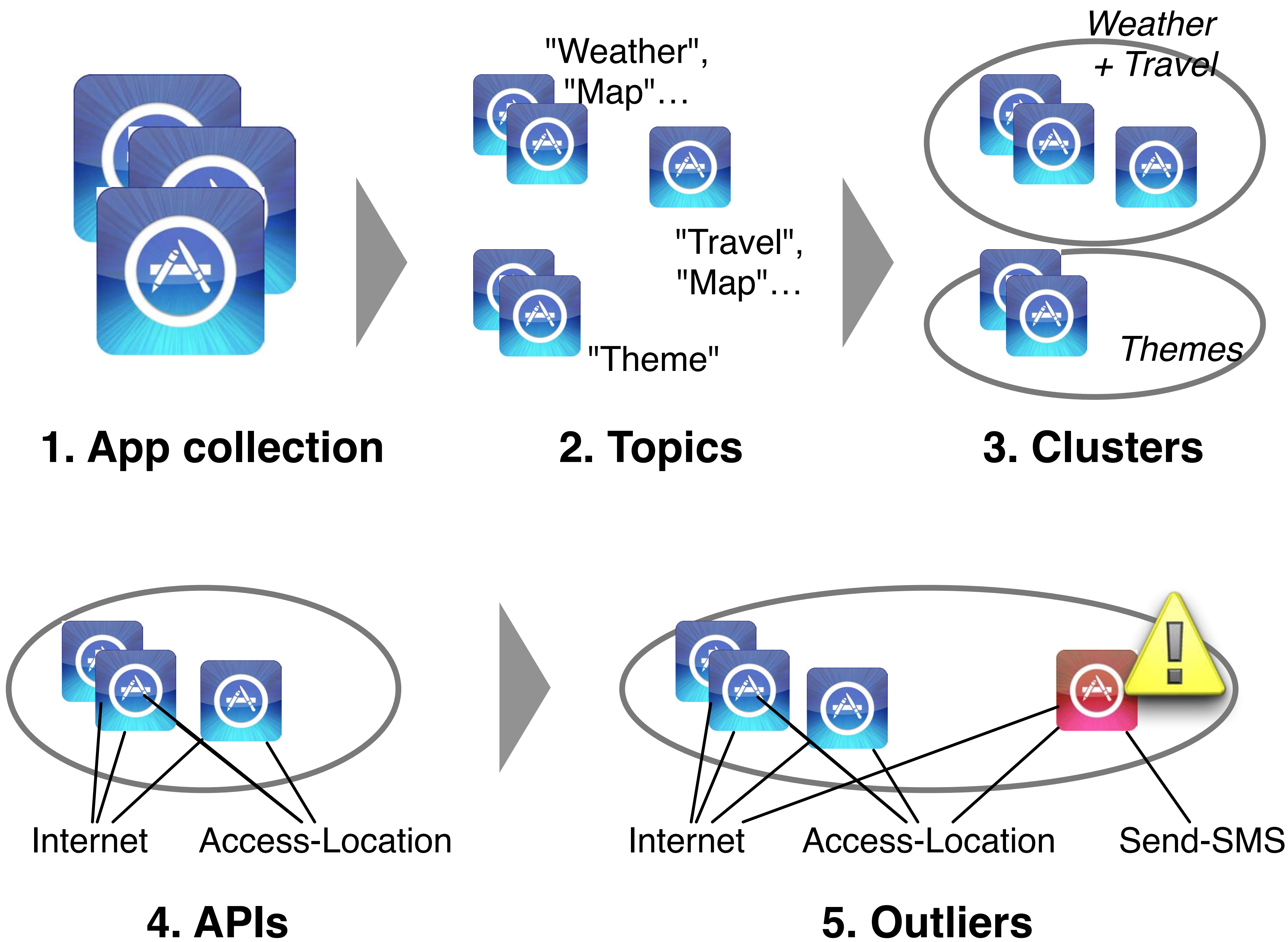
London Restaurants

- “London Restaurants” is a “travel” app
- For “travel” apps, sending account infos is *abnormal*
- For “messaging” apps, this is far more likely

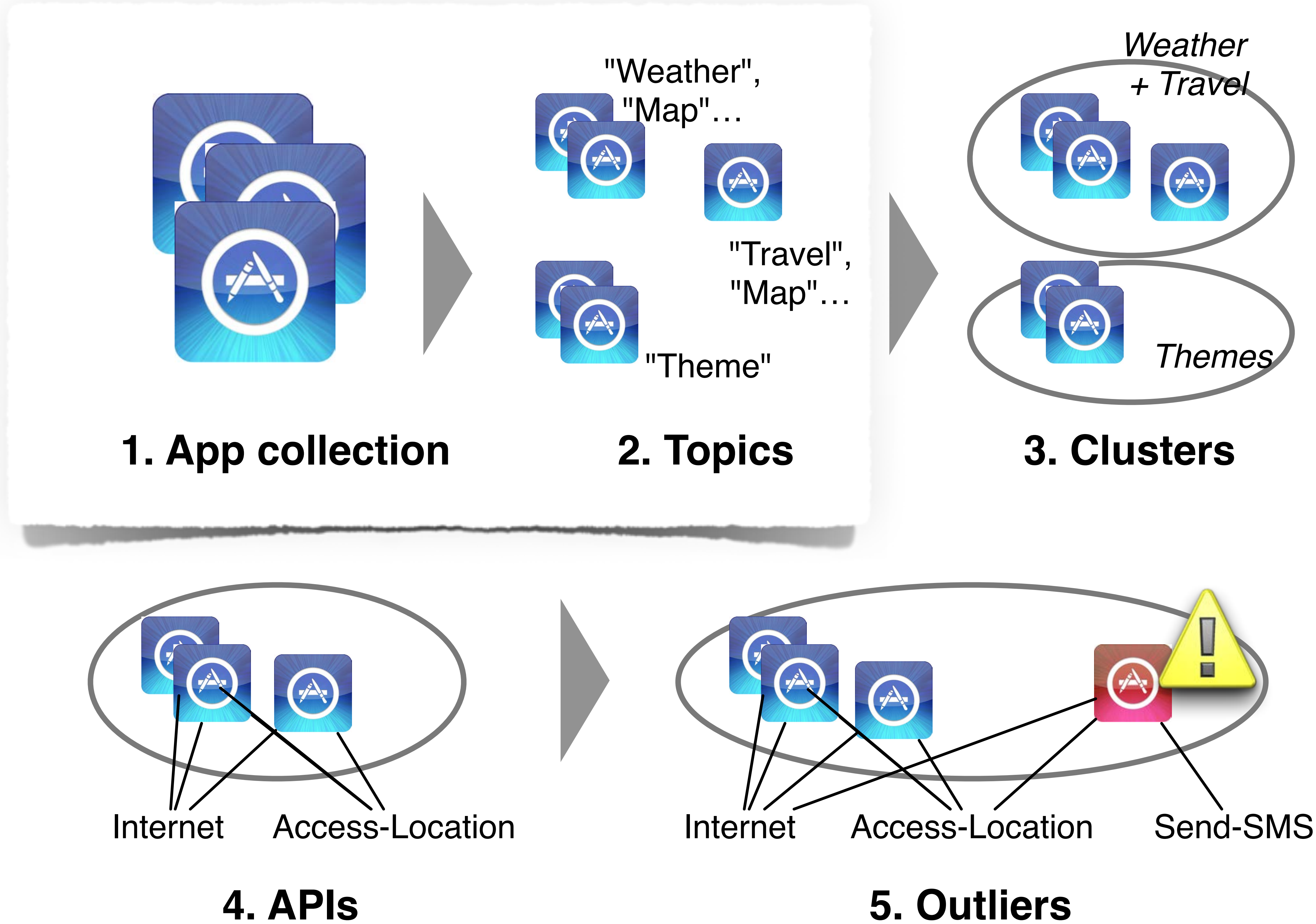
# CHABADA


"Weather",
"Map"…
"Travel",
"Map"…
"Theme"
Weather
+ Travel
Themes
1. App collection
2. Topics
3. Clusters

# CHABADA


"Weather",
"Map"…
"Travel",
"Map"…
"Theme"
Weather
+ Travel
Themes
1. App collection
2. Topics
3. Clusters
Internet
Access-Location
Internet
Access-Location
Send-SMS
4. APIs
5. Outliers

# CHABADA

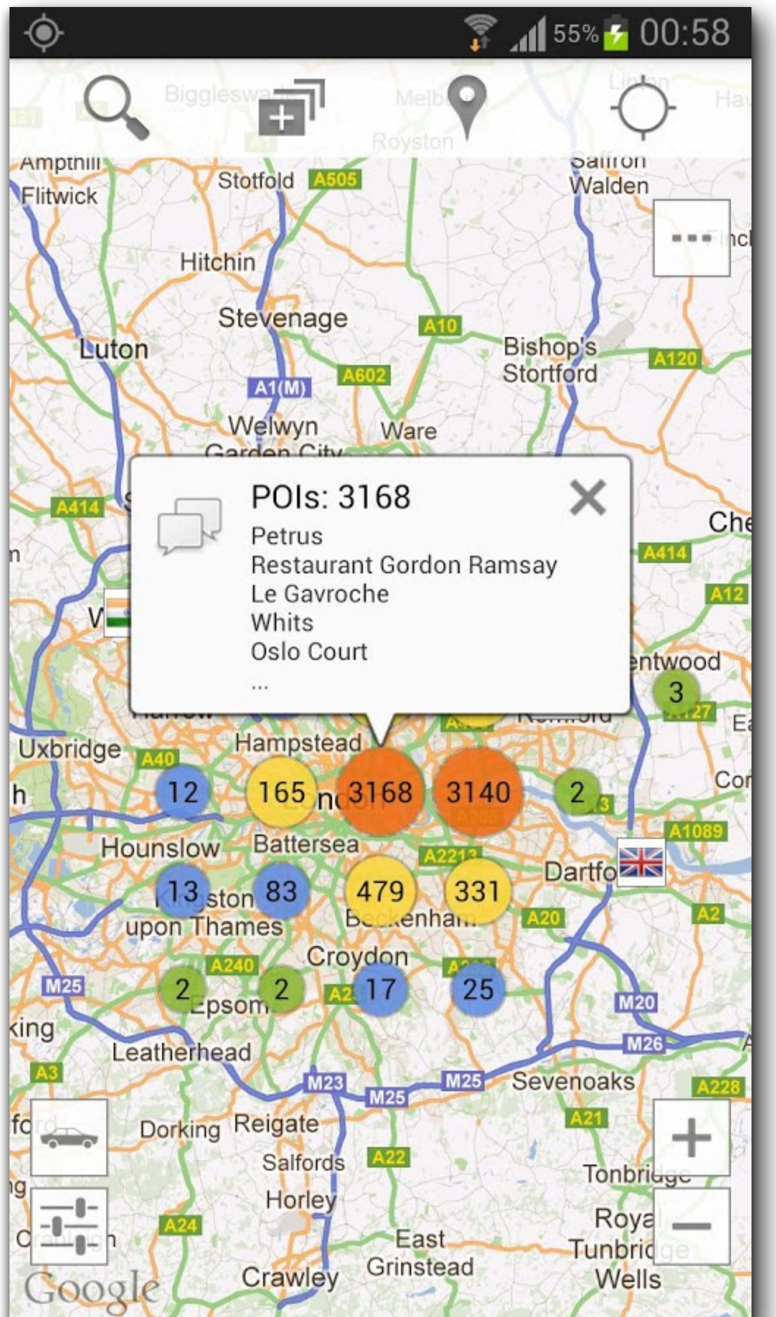


"Weather",
"Map"…
"Travel",
"Map"…
"Theme"
Weather
+ Travel
Themes
1. App collection
2. Topics
3. Clusters
Internet
Access-Location
Internet
Access-Location
Send-SMS
4. APIs
5. Outliers

# App Collection

- Source: Google Play Store
- Downloaded top 150 apps + metadata from each of the 30 categories
- Time frame: Winter to Spring 2013
- Total: 32,136 apps
- Data package available on Web site

# Stemming

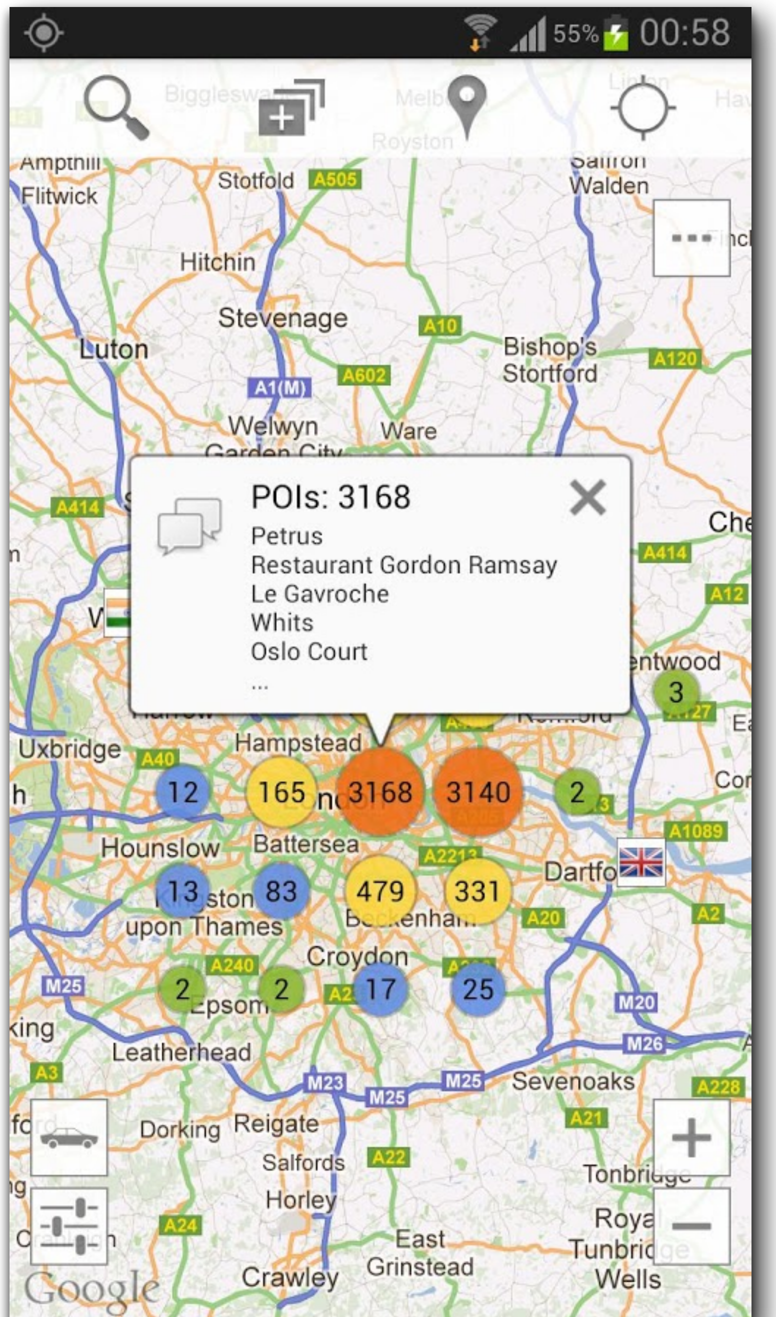



looking for a restaurant, a bar, a pub or just to have fun in london? search no more! this application has all the information you need:

- you can search for every type of food you want: french, british, chinese, indian etc.
- you can use it if you are in a car, on a bicycle or walking
- you can view all objectives on the map
- you can search objectives
- you can view objectives near you
- you can view directions (visual route, distance and duration)
- you can use it with street view
- you can use it with navigation

keywords: london, restaurants, bars, pubs, food, breakfast, lunch, dinner, meal, eat, supper, street view, navigation

# Stemming

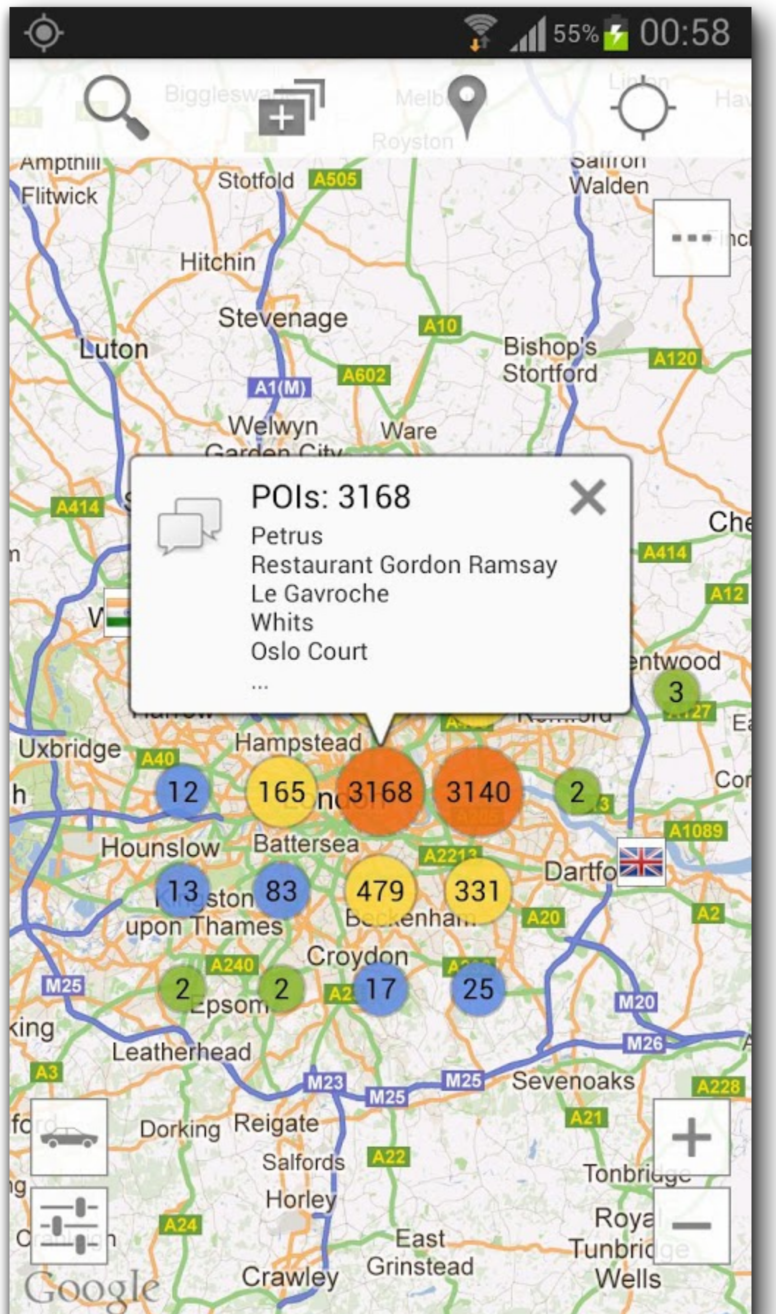



looking for a restaurant, a bar, a pub or just to have fun in london? search no more! this application has all the information you need:

- you can search for everi type of food you want: french, british, chinese, indian etc.
- you can use it if you are in a car, on a bicycle or walking
- you can view all objectives on the map
- you can search objectives
- you can view objectives near you
- you can view directions (visual route, distance and duration)
- you can use it with street view
- you can use it with navigation

keywords: london, restaurants, bars, pubs, food, breakfast, lunch, dinner, meal, eat, supper, street view, navigation

# Stemming




look london restaur search bar pub just applic fun

inform can search need everi type food want french

british chines indian etc car bicycl walk

can us can view object map visual rout

can search object search can view distanc

durat can view direct object near

can us street view can us navig

keyword london restaur bar pub food view

breakfast lunch dinner meal eat supper street navig

# Topic Analysis

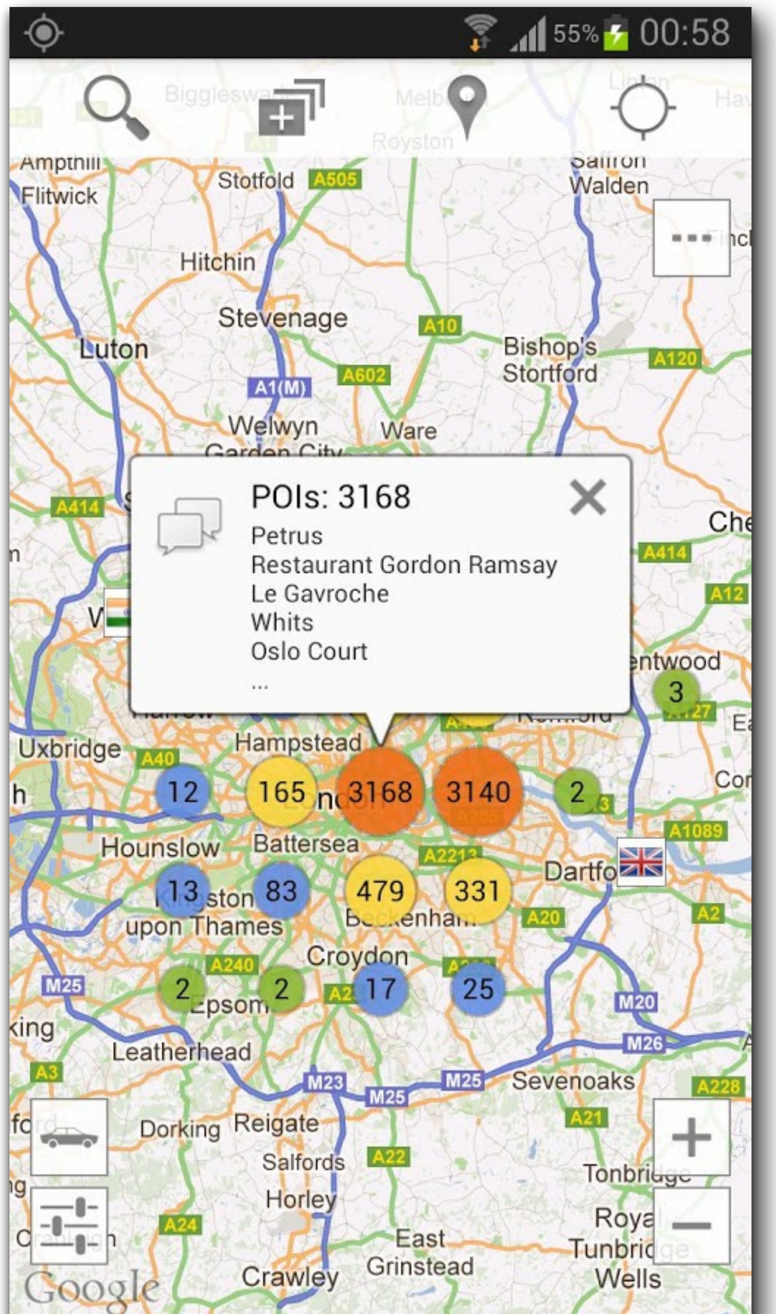
- Eliminated all apps with $\leq$10 words,
  now 22,521 apps
- Want to discover the *topics* that occur in a collection of unlabeled text
- A *topic* consists of a cluster of words that frequently occur together
- Used *Latent Dirichlet Allocation* (LDA)
  to identify 30 topics

# Topics

| Id | Assigned Name | Most Representative Words (stemmed) |
|---|---|---|
| 0 | “personalize” | galaxi, nexu, device, screen, effect, instal, customis |
| 1 | “game and cheat sheets” | game, video, page, cheat, link, tip, trick |
| 2 | “money” | slot, machine, money, poker, currenc, market, trade, stock, casino coin, finance |
| 3 | “tv” | tv, channel, countri, live, watch, germani, nation, bbc, newspap |
| 4 | “music” | music, song, radio, play, player, listen |
| 5 | “holidays” and religion | christmas, halloween, santa, year, holiday, islam, god |
| 6 | “navigation and travel” | map, inform, track, gps, navig, travel |
| 7 | “language” | language, word, english, learn, german, translat |
| 8 | “share” | email, ad, support, facebook, share, twitter, rate, suggest |
| 9 | “weather and stars” | weather, forecast, locate, temperatur, map, city, light |
| 10 | “files and video” | file, download, video, media, support, manage, share, view, search |

| | | |
|---|---|---|
| 13 | "design and art" | life, peopl, natur, form, feel, learn, art, design, uniqu, effect, modern |
| 14 | "food and recipes" | recip, cake, chicken, cook, food |
| 15 | "personalize" | theme, launcher, download, install, icon, menu |
| 16 | "health" | weight, bodi, exercise, diet, workout, medic |
| 17 | "travel" | citi, guid, map, travel, flag, countri, attract |
| 18 | "kids and bodies" | kid, anim, color, girl, babi, pictur, fun, draw, design, learn |
| 19 | "ringtones and sound" | sound, rington, alarm, notif, music |
| 20 | "game" | game, plai, graphic, fun, jump, level, ball, 3d, score |
| 21 | "search and browse" | search, icon, delet, bookmark, link, homepag, shortcut, browser |
| 22 | "battle games" | story, game, monster, zombi, war, battle |
| 23 | "settings and utils" | screen, set, widget, phone, batteri |
| 24 | "sports" | team, football, leagu, player, sport, basketbal |
| 25 | "wallpapers" | wallpap, live, home, screen, background, menu |
| 26 | "connection" | device, connect, network, wifi, blootooth, internet, remot, server |
| 27 | "policies and ads" | live, ad, home, applovin, notif, data, polici, privacy, share, airpush, advertis |
| 28 | "popular media" | seri, video, film, album, movi, music, award, star, fan, show, gangnam, top, bieber |
| 29 | "puzzle and card games" | game, plai, level, puzzl, player, score, challeng, card |

# London Restaurant Topics




look london restaur search bar pub just applic fun

inform can search need everi type food want french

british chines indian etc car bicycl walk

can us can view object map visual rout

can search object search can view distanc

durat can view direct object near

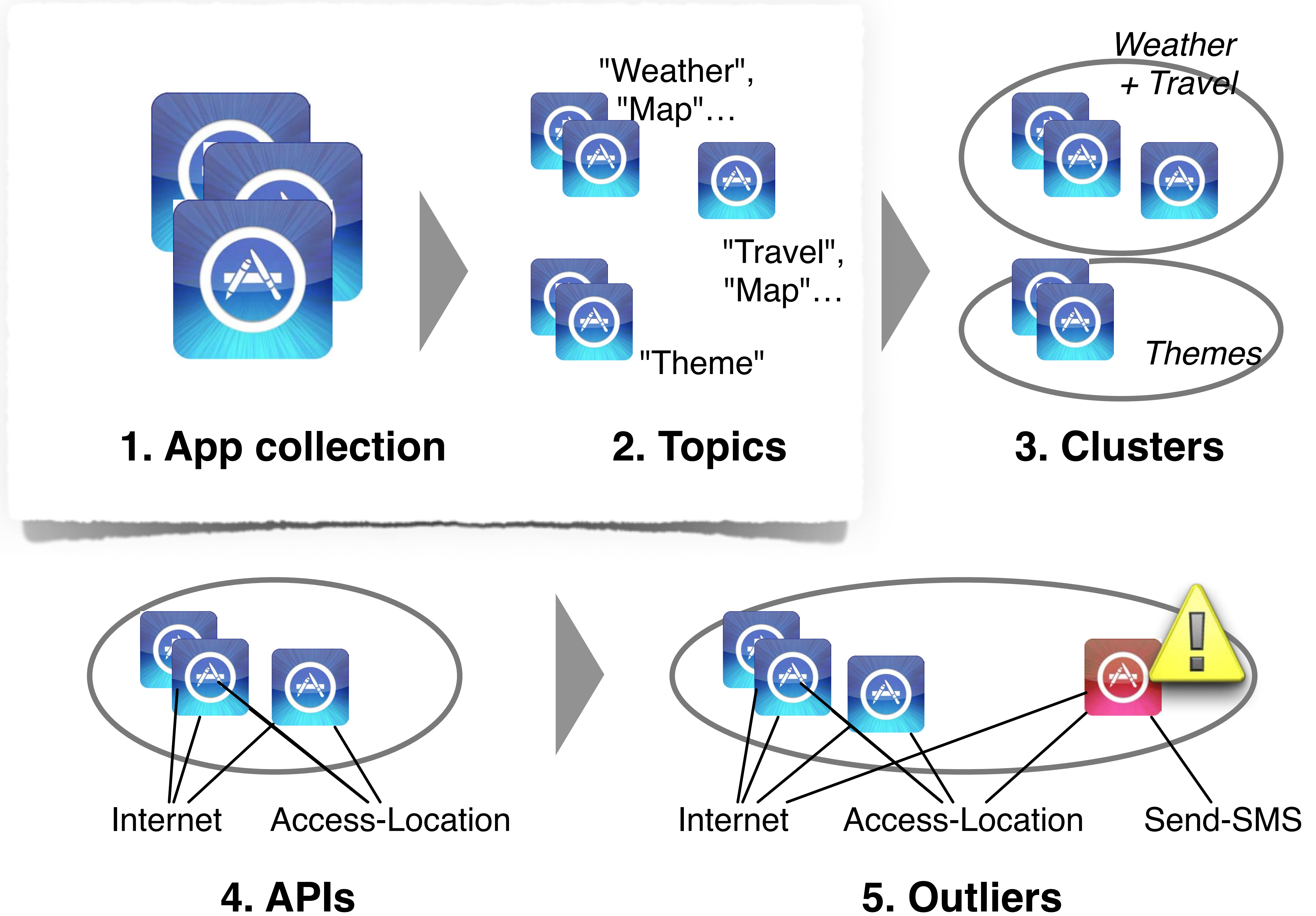
can us street view can us navig

keyword london restaur bar pub food view

breakfast lunch dinner meal eat supper street navig

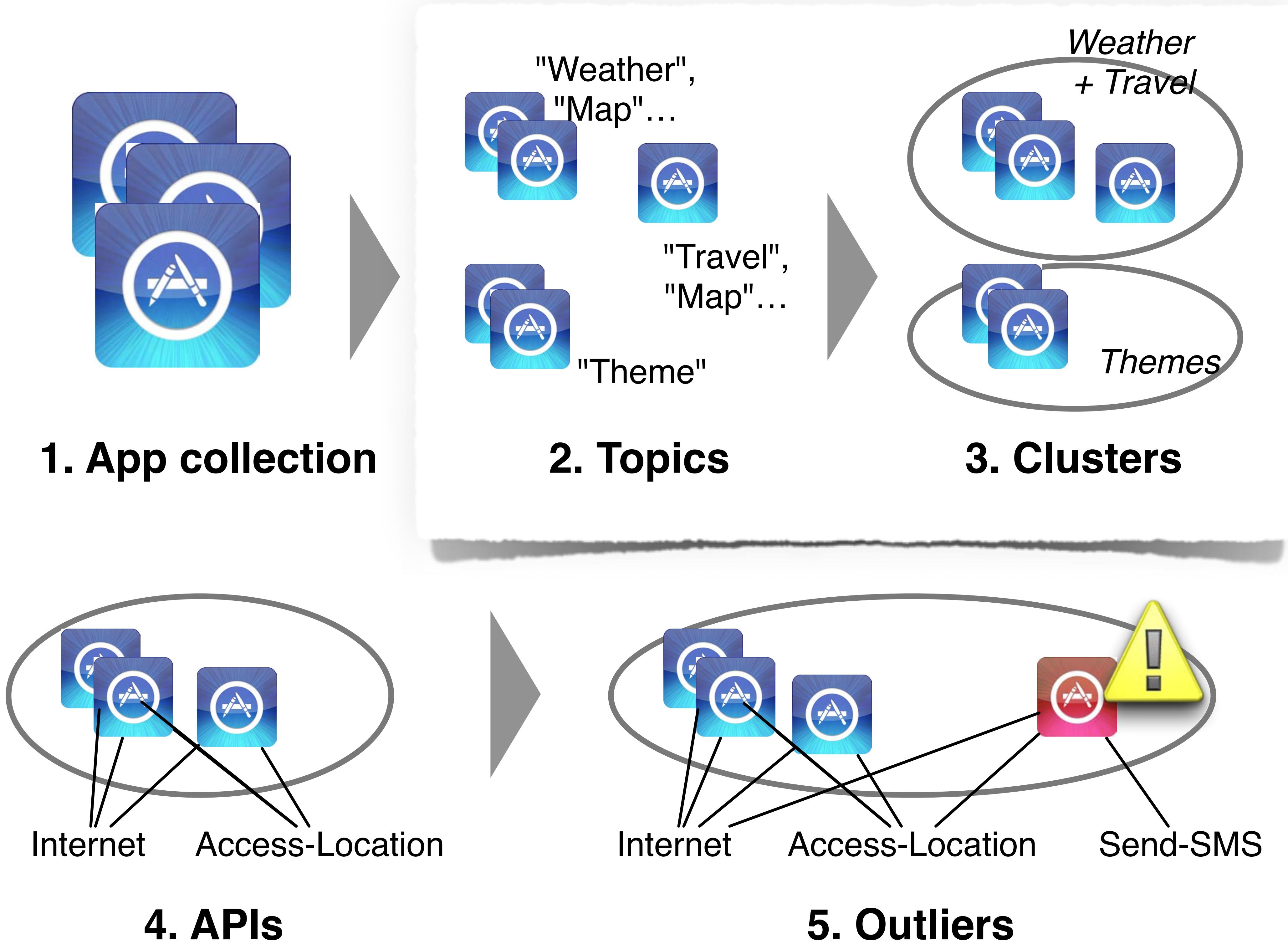
"navigation and travel" (59.8%)
"food and recipes" (19.9%)
"travel" (14.0%)

# CHABADA



"Weather",
"Map"…
"Travel",
"Map"…
"Theme"
Weather
+ Travel
Themes
1. App collection
2. Topics
3. Clusters
Internet
Access-Location
Internet
Access-Location
Send-SMS
4. APIs
5. Outliers

# CHABADA


"Weather",
"Map"…
"Travel",
"Map"…
"Theme"
Weather
+ Travel
Themes
1. App collection
2. Topics
3. Clusters
Internet
Access-Location
Internet
Access-Location
Send-SMS
4. APIs
5. Outliers

# Clustering

- Want to identify *groups of applications* that are similar according to their descriptions.
- Used *K-Means* to identify such clusters
- Used *elements silhouette*
  to identify best number K of clusters

# Clusters

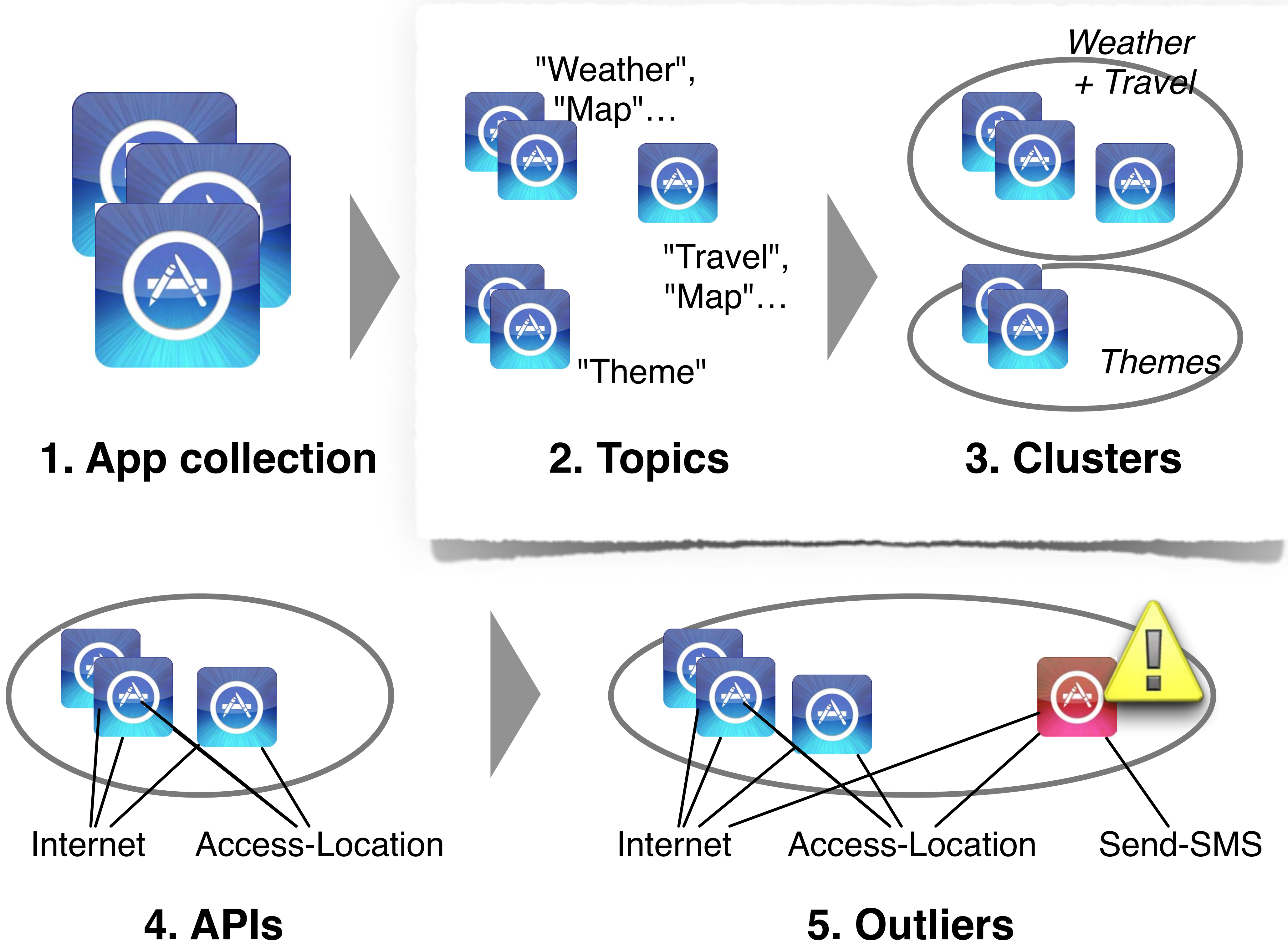
| Id | Assigned Name | Size | Most Important Topics |
|---|---|---|---|
| 1 | “sharing” | 1,453 | **share** (53%), settings and utils, navigation and travel |
| 2 | “puzzle and card games” | 953 | **puzzle and card games** (78%), share, game |
| 3 | “memory puzzles” | 1,069 | **puzzle and card games** (40%), game (12%), share |
| 4 | “music” | 714 | **music** (58%), share, settings and utils |
| 5 | “music videos” | 773 | **popular media** (44%), **holidays and religion** (20%), share |
| 6 | “religious wallpapers” | 367 | **holidays and religion** (56%), design and art, wallpapers |
| 7 | “language” | 602 | **language** (67%), share, settings and utils |
| 8 | “cheat sheets” | 785 | **game and cheat sheets** (76%), share, popular media |
| 9 | “utils” | 1,300 | **settings and utils** (62%), share, connection |
| 10 | “sports game” | 1,306 | **game** (63%), battle games, puzzle and card games |
| 11 | “battle games” | 953 | **battle games** (60%), **game** |

| | | | |
|---|---|---|---|
| | | | and travel |
| 19 | “sports” | 580 | **sports** (62%), share, popular media |
| 20 | “files and videos” | 679 | **files and videos** (63%), share, settings and utils |
| 21 | “search and browse” | 363 | **search and browse** (64%), game, puzzle and card games |
| 22 | “advertisements” | 380 | **policies and ads** (97%) |
| 23 | “design and art” | 978 | **design and art** (48%), share, game |
| 24 | “car games” | 449 | **cars** (51%), game, puzzle and card games |
| 25 | “tv live” | 500 | **tv** (57%), share, navigation and travel |
| 26 | “adult photo” | 828 | **photo and social** (59%), share, settings and utils |
| 27 | “adult wallpapers” | 543 | **wallpapers** (51%), share, kids and bodies |
| 28 | “ad wallpapers” | 180 | **policies and ads** (46%), wallpapers, settings and utils |
| 29 | “ringtones and sound” | 662 | **ringtones and sound** (68%), share, settings and utils |
| 30 | “theme wallpapers” | 593 | **wallpapers** (90%), holidays and religion, share |
| 31 | “personalize” | 402 | **personalize** (86%), share, settings and utils |
| 32 | “settings and wallpapers” | 251 | **settings and utils** (37%), **wallpapers** (37%), personalize |

# "Personalize" Cluster

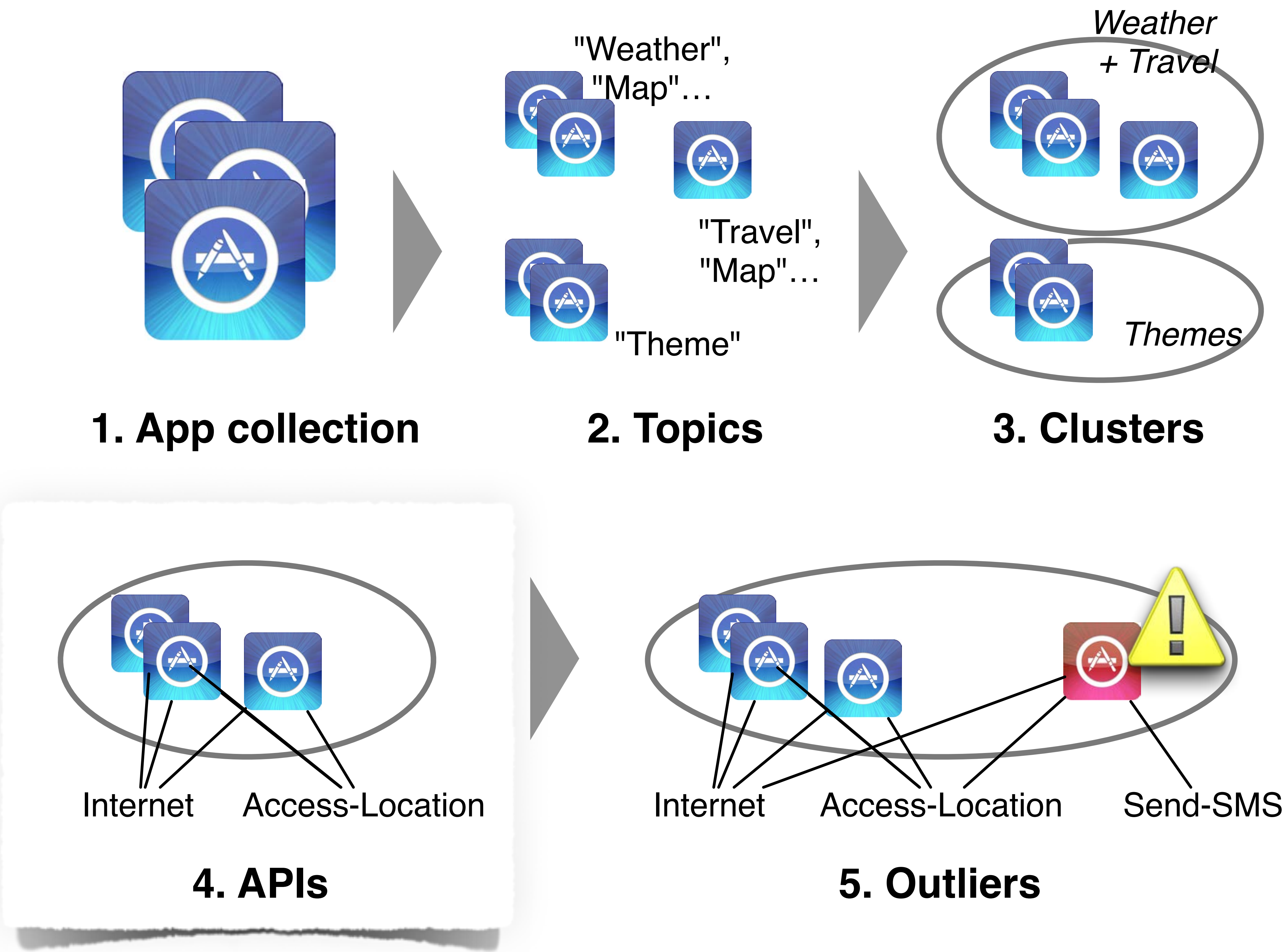
# “Travel” Cluster

# CHABADA



"Weather",
"Map"…
"Travel",
"Map"…
"Theme"
Weather
+ Travel
Themes
1. App collection
2. Topics
3. Clusters
Internet
Access-Location
Internet
Access-Location
Send-SMS
4. APIs
5. Outliers

# CHABADA

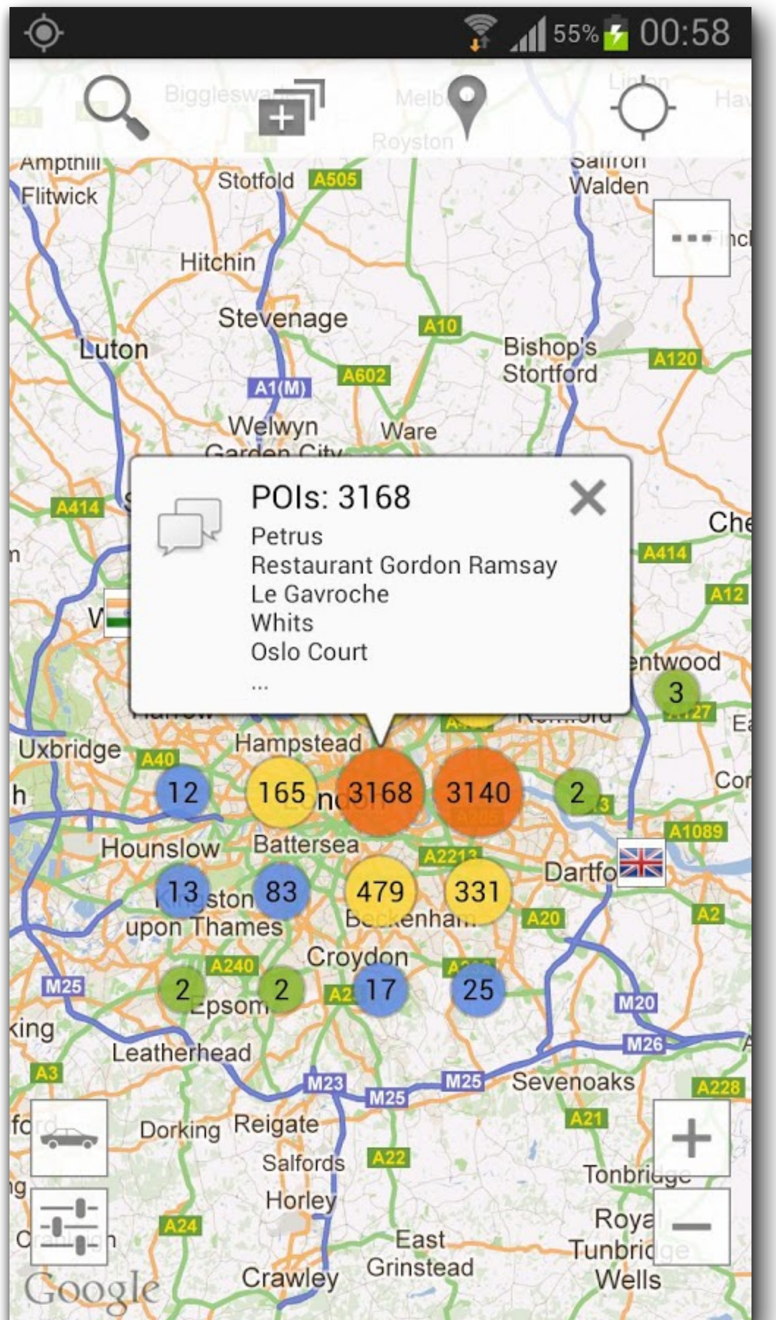


"Weather",
"Map"…
"Travel",
"Map"…
"Theme"
Weather
+ Travel
Themes
1. App collection
2. Topics
3. Clusters
Internet
Access-Location
4. APIs
Internet
Access-Location
Send-SMS
5. Outliers

# API Analysis

- For each APK, we identified the APIs used
- Used simple static analysis
- Only considered *sensitive APIs* which would be governed by *permissions*

# London Restaurants




android.net.ConnectivityManager.getActiveNetworkInfo()
android.webkit.WebView()

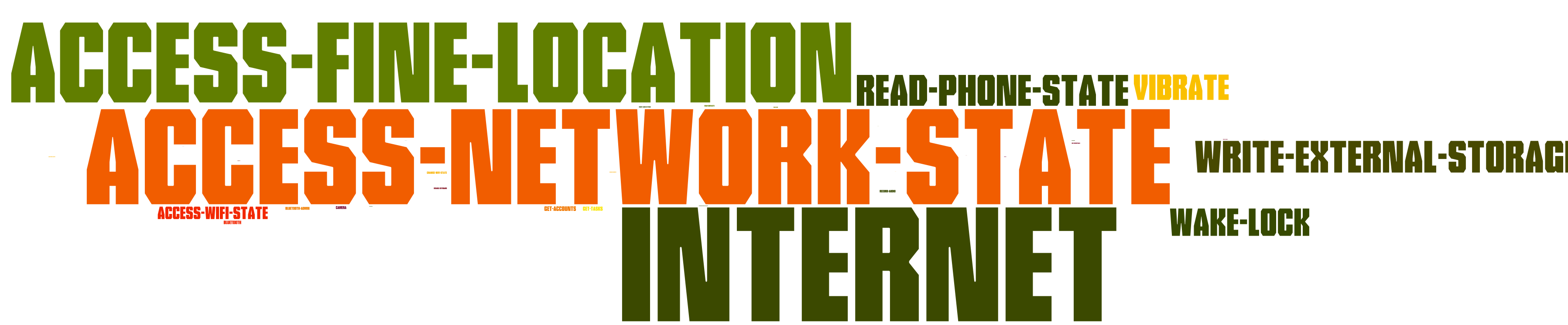
INTERNET
GET-ACCOUNTS
ACCESS-WIFI-STATE
ACCESS-NETWORK-STATE
ACCESS-FINE-LOCATION
READ-PHONE-STATE
VIBRATE

android.net.NetworkInfo.isConnectedOrConnecting()
android.net.ConnectivityManager.getAllNetworkInfo()

# “Travel” Cluster

Description

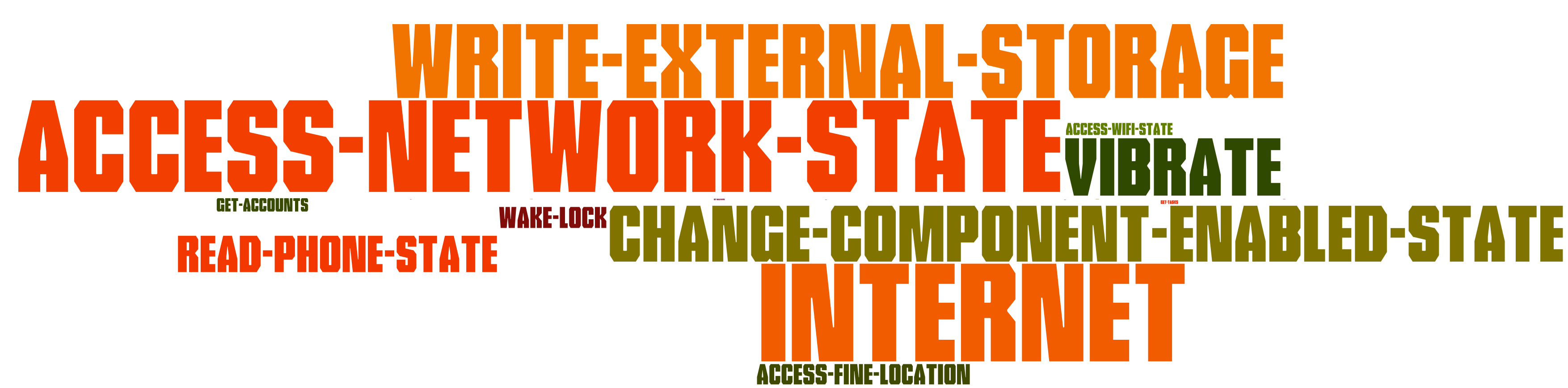

Permissions of APIs used



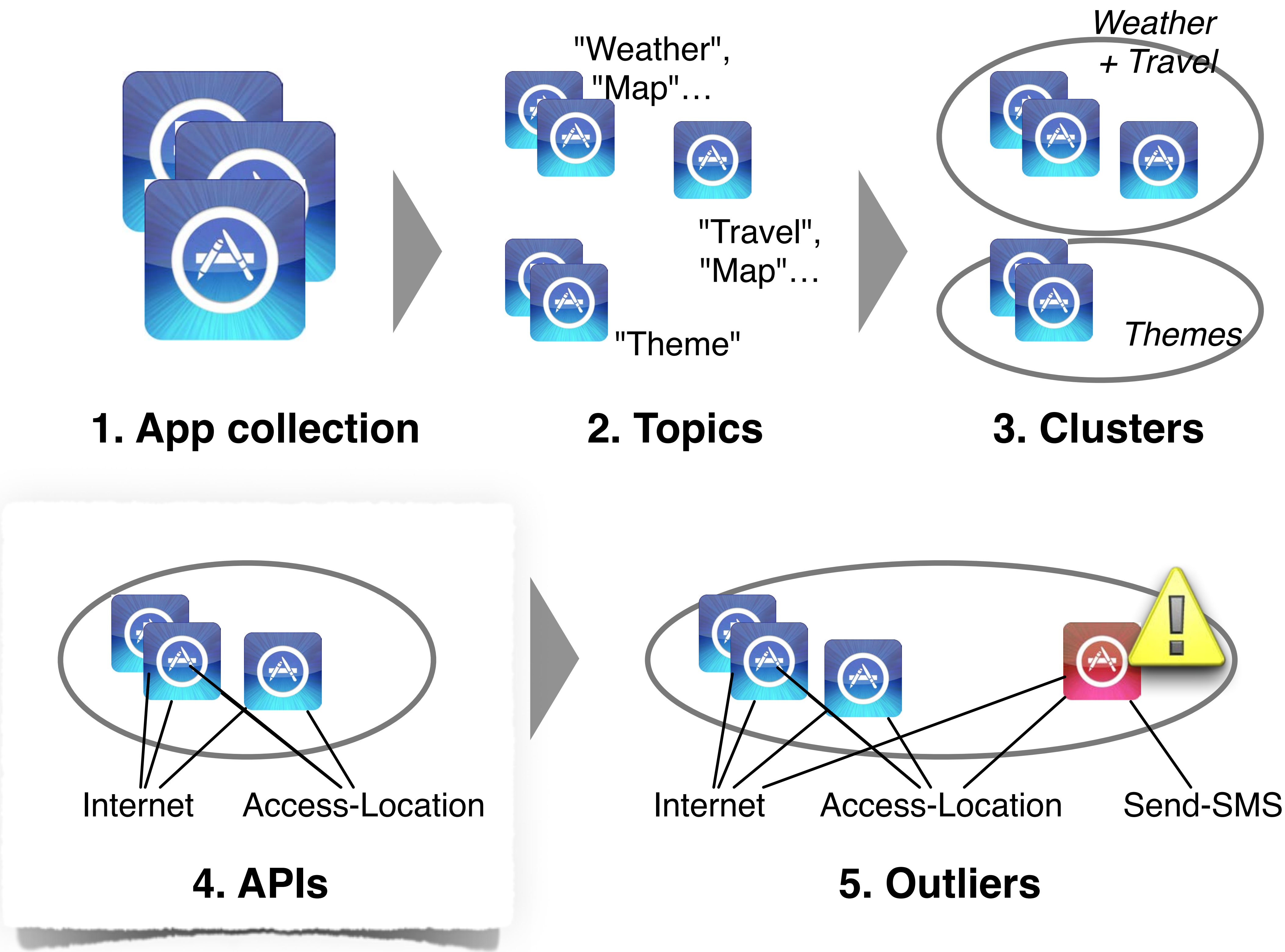
# “Personalize” Cluster

Description

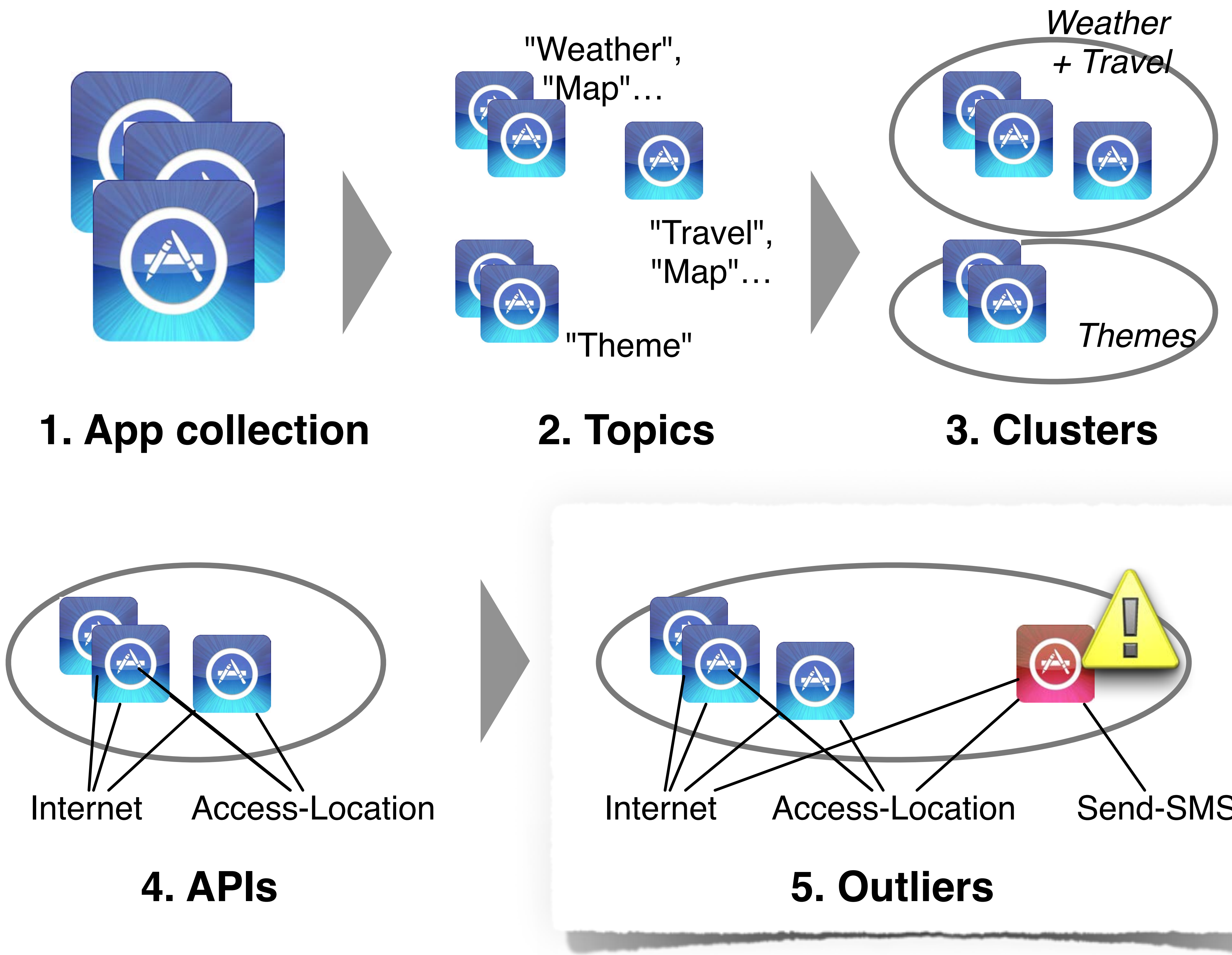

Permissions of APIs used



# CHABADA


"Weather",
"Map"…
"Travel",
"Map"…
"Theme"
Weather
+ Travel
Themes
1. App collection
2. Topics
3. Clusters
Internet
Access-Location
4. APIs
Internet
Access-Location
Send-SMS
5. Outliers

# CHABADA

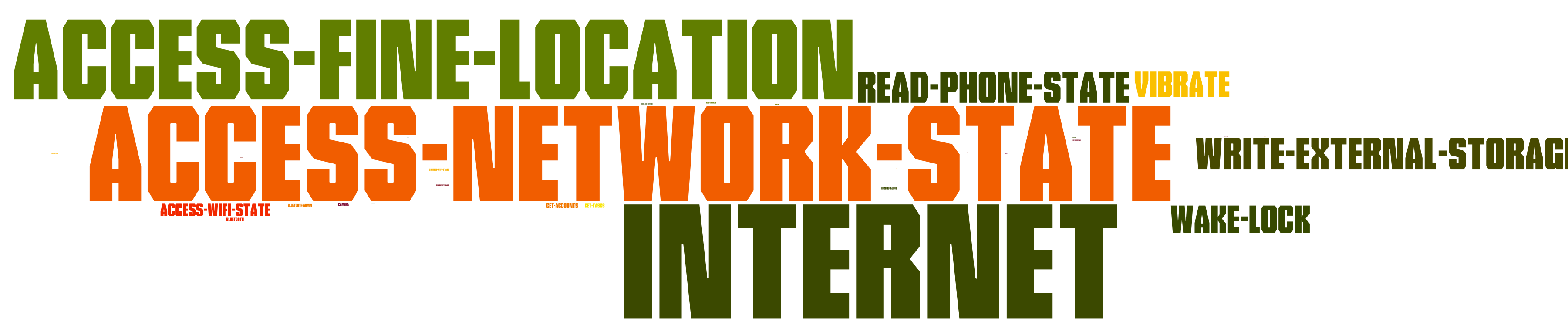


"Weather",
"Map"…
"Travel",
"Map"…
"Theme"
Weather
+ Travel
Themes
1. App collection
2. Topics
3. Clusters
Internet
Access-Location
Internet
Access-Location
Send-SMS
4. APIs
5. Outliers

# "Travel" Cluster

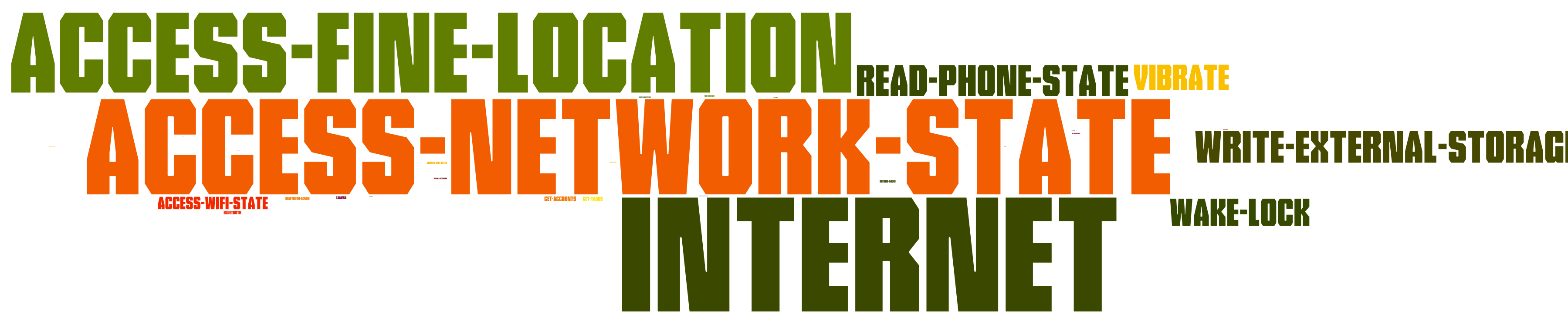
Description


Permissions of APIs used



# “Travel” Cluster

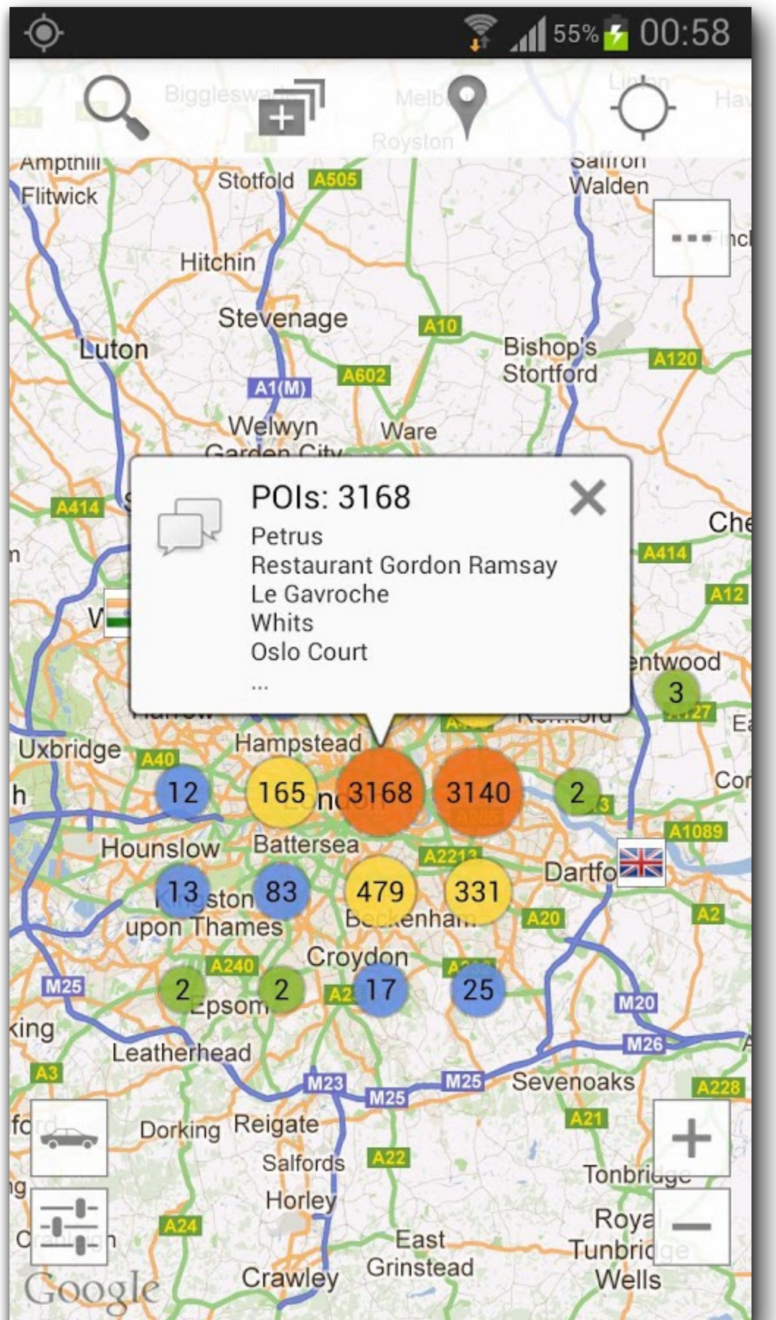
Permissions of APIs used




# London Restaurants

Permissions of APIs used

# London Restaurants

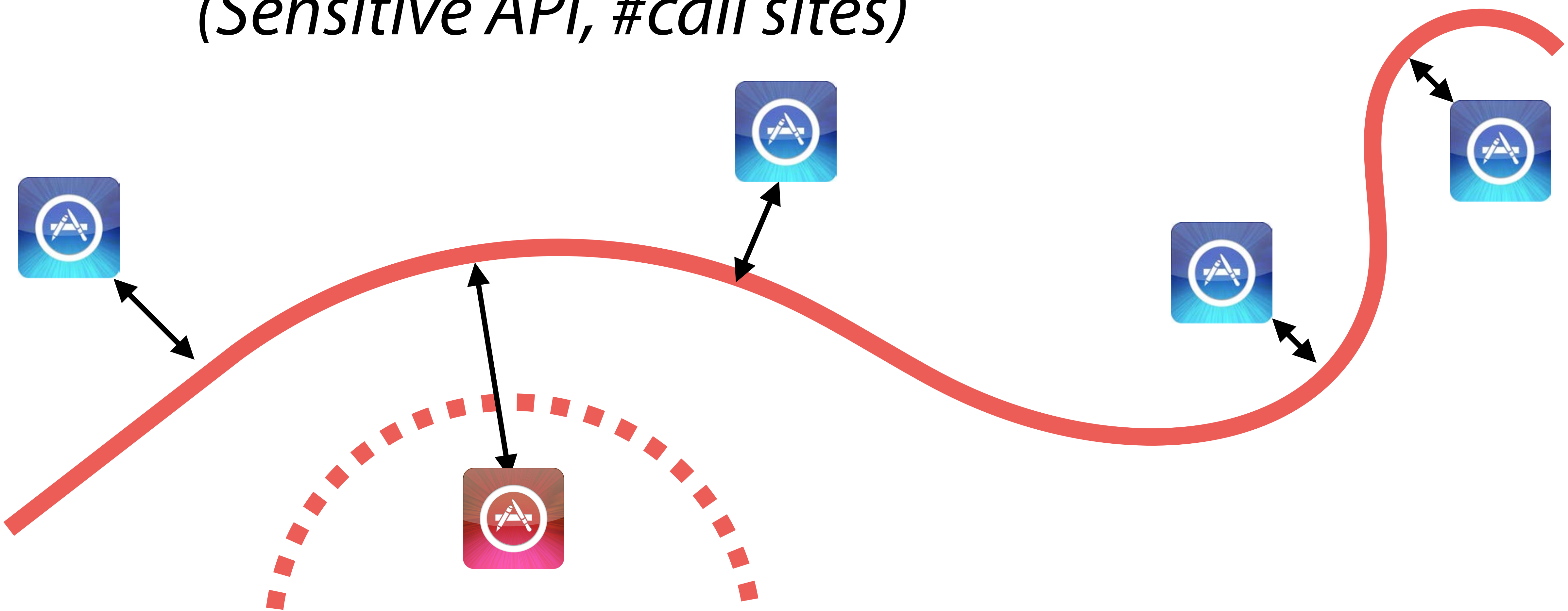



android.net.ConnectivityManager.getActiveNetworkInfo()
android.webkit.WebView()
java.net.HttpURLConnection.connect()
android.app.NotificationManager.notify()
java.net.URL.openConnection()
**android.telephony.TelephonyManager.getDeviceId()**
org.apache.http.impl.client.DefaultHttpClient()
org.apache.http.impl.client.DefaultHttpClient.execute()
android.location.LocationManager.getBestProvider()
**android.telephony.TelephonyManager.getLine1Number()**
android.net.wifi.WifiManager.isWifiEnabled()
**android.accounts.AccountManager.getAccountsByType()**
android.net.wifi.WifiManager.getConnectionInfo()
android.location.LocationManager.getLastKnownLocation()
android.location.LocationManager.isProviderEnabled()
android.location.LocationManager.requestLocationUpdates()
android.net.NetworkInfo.isConnectedOrConnecting()
android.net.ConnectivityManager.getAllNetworkInfo()

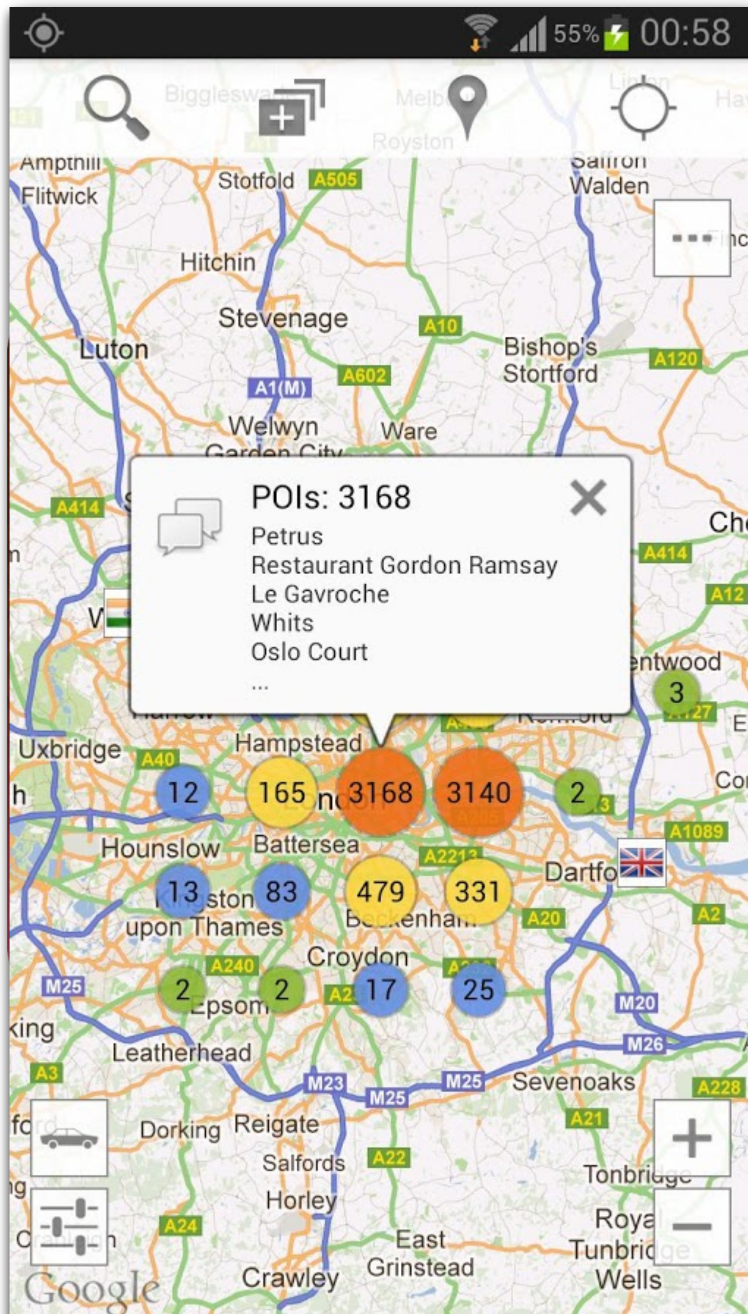
**→ An *Outlier* in the “Travel” Cluster**

# Outlier Analysis

- In each cluster, identified outliers through *one-class support vector machine* (OC-SVM)
- Features of each APK: a vector of *(Sensitive API, #call sites)*



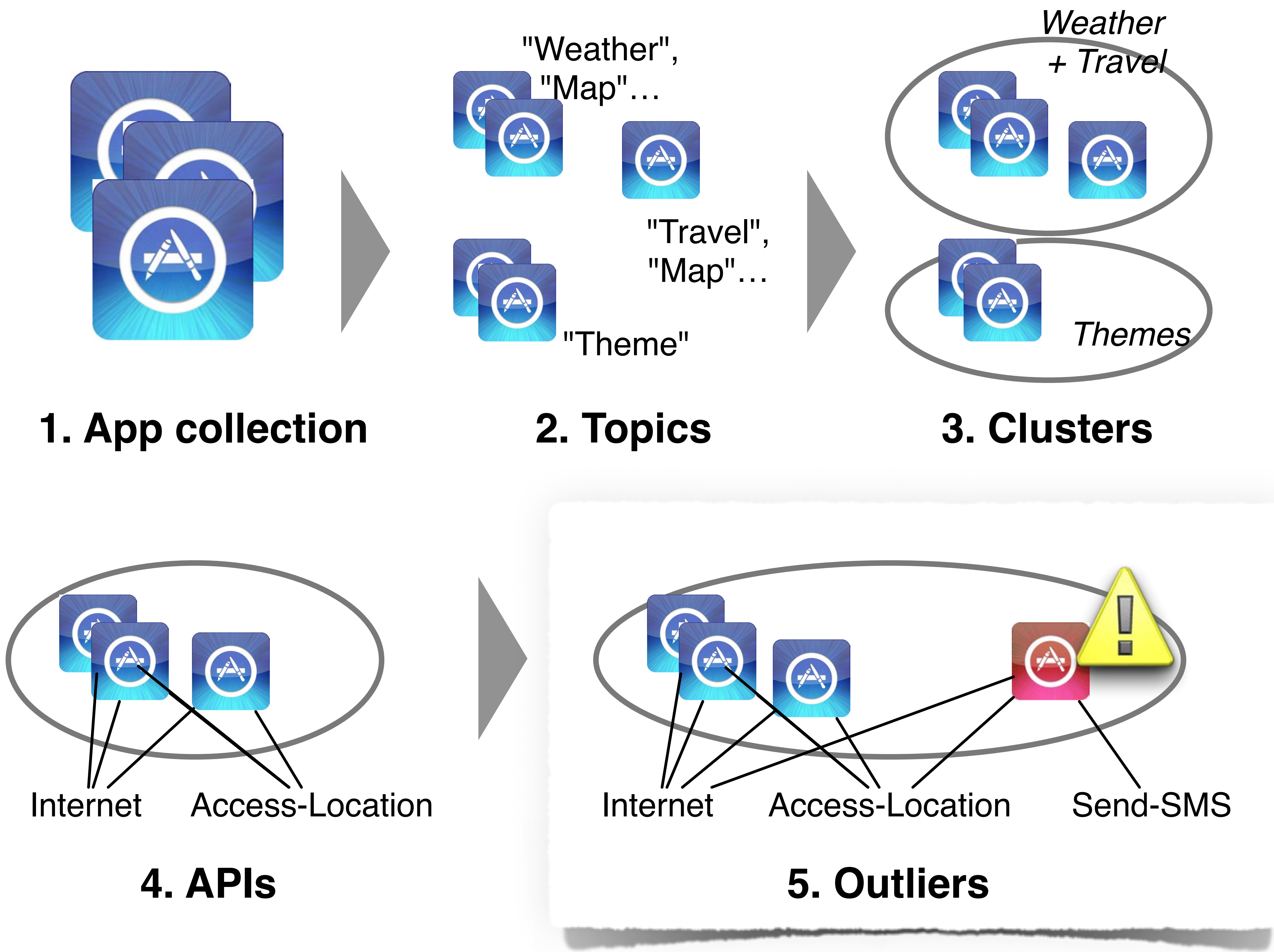
# London Restaurants




android.net.ConnectivityManager.getActiveNetworkInfo()
android.webkit.WebView()
java.net.HttpURLConnection.connect()
android.app.NotificationManager.notify()
java.net.URL.openConnection()
**android.telephony.TelephonyManager.getDeviceId()**
org.apache.http.impl.client.DefaultHttpClient()
org.apache.http.impl.client.DefaultHttpClient.execute()
android.location.LocationManager.getBestProvider()
**android.telephony.TelephonyManager.getLine1Number()**
android.net.wifi.WifiManager.isWifiEnabled()
**android.accounts.AccountManager.getAccountsByType()**
android.net.wifi.WifiManager.getConnectionInfo()
android.location.LocationManager.getLastKnownLocation()
android.location.LocationManager.isProviderEnabled()
android.location.LocationManager.requestLocationUpdates()
android.net.NetworkInfo.isConnectedOrConnecting()
android.net.ConnectivityManager.getAllNetworkInfo()

→ **Identified as Outlier**

# CHABADA


"Weather",
"Map"…
"Travel",
"Map"…
"Theme"
Weather
+ Travel
Themes
1. App collection
2. Topics
3. Clusters
Internet
Access-Location
Internet
Access-Location
Send-SMS
4. APIs
5. Outliers

# Evaluation: Outliers

- *Can our technique effectively identify anomalies (i.e., mismatches between description and behavior) in Android apps?*
- Manually checked
  top 5 outliers in each cluster (160 total)
- 26% showed *covert behavior using sensitive APIs that acts against the interest of its users.*

# What makes an outlier?

- Ad frameworks (apploving, airpush)
- Dubious behavior (UNO, WICKED, Yahoo!)
- Uncommon behavior (SoundCloud)
- Benign outliers (Mr. Will's Stud Poker)

# Evaluation: Malware

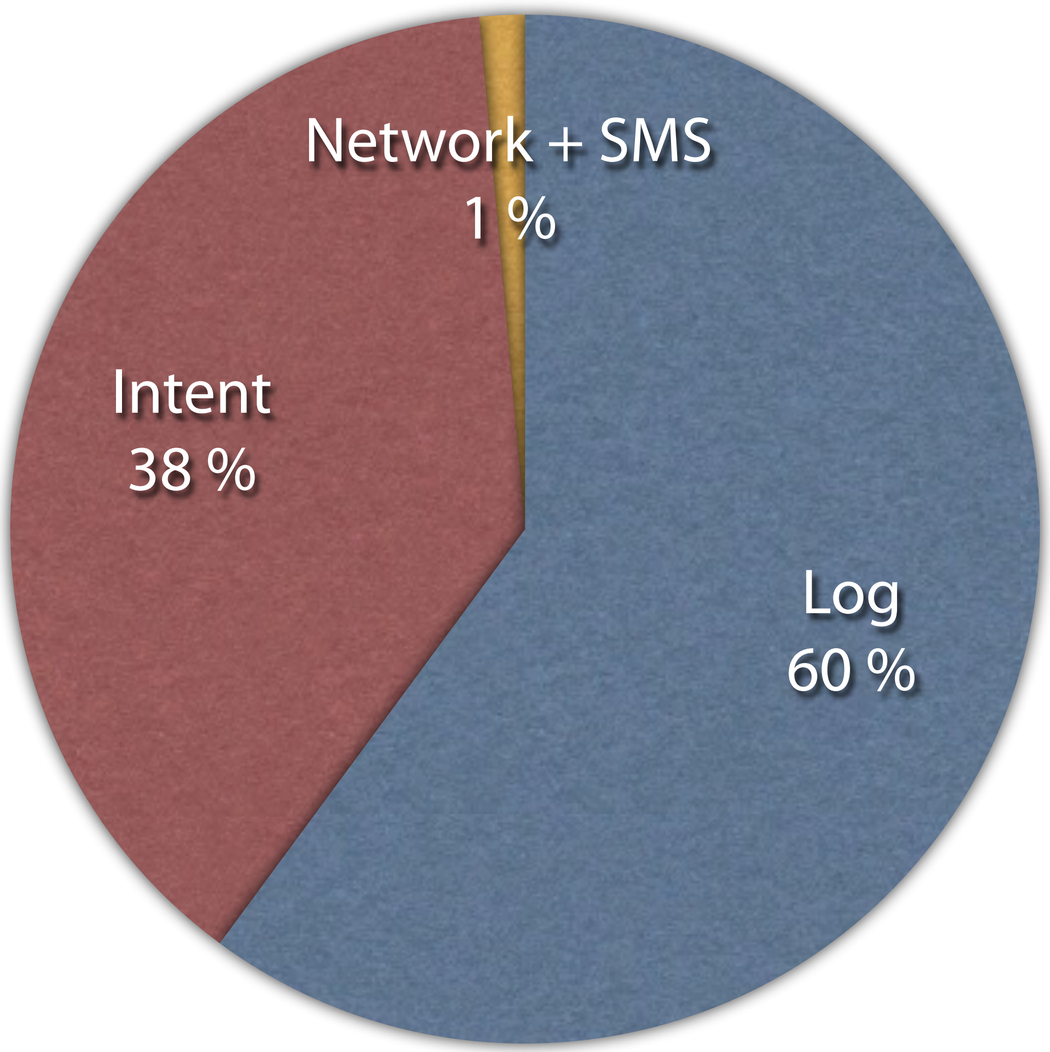
- *Can our technique be used to identify malicious Android applications?*
- In each cluster, trained OC-SVM on 90% of "benign" apps
- Used TF-IDF as classifier on sets with remaining "benign" apps and 173 known malware apps

**Malware recognition rate >80%**

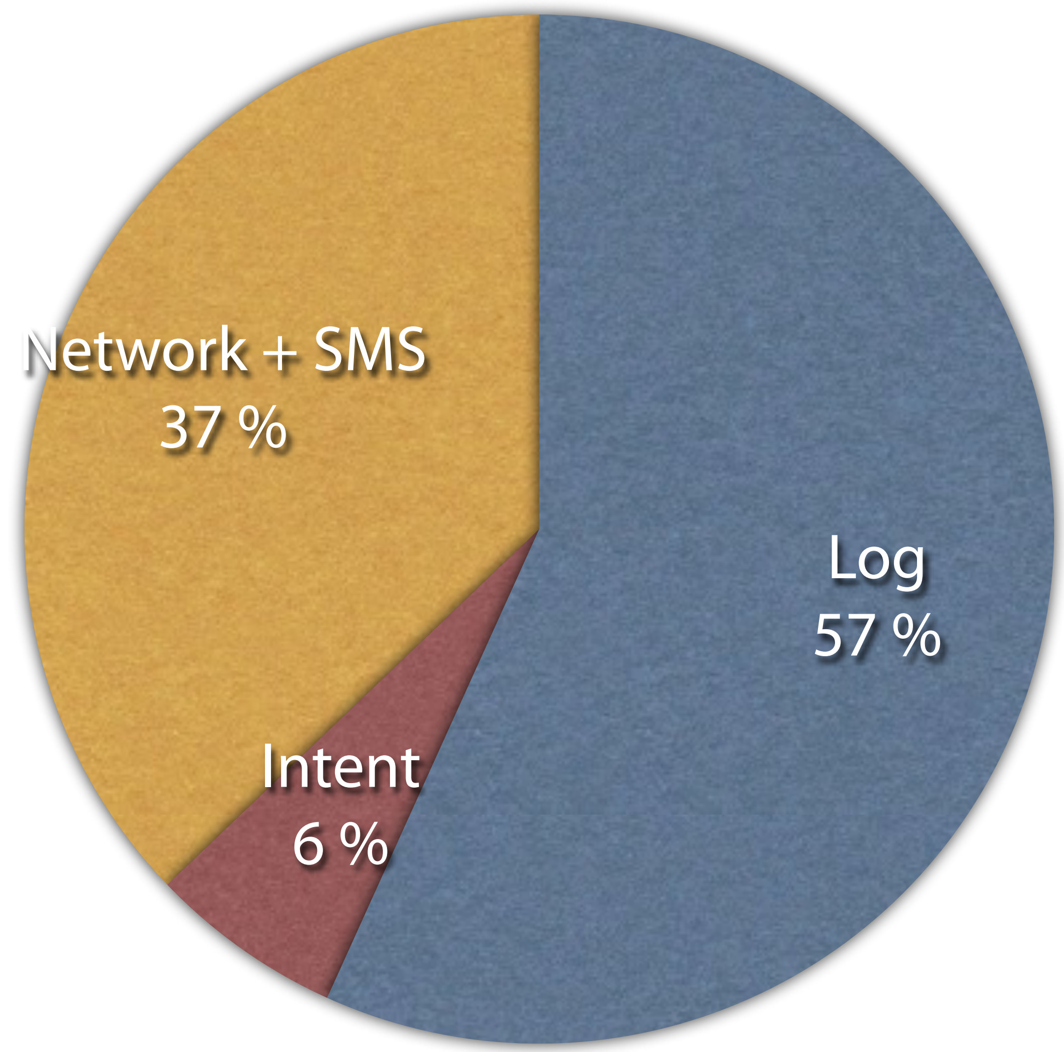
# Information Flow

- Which sensitive APIs does the *device ID* flow to?

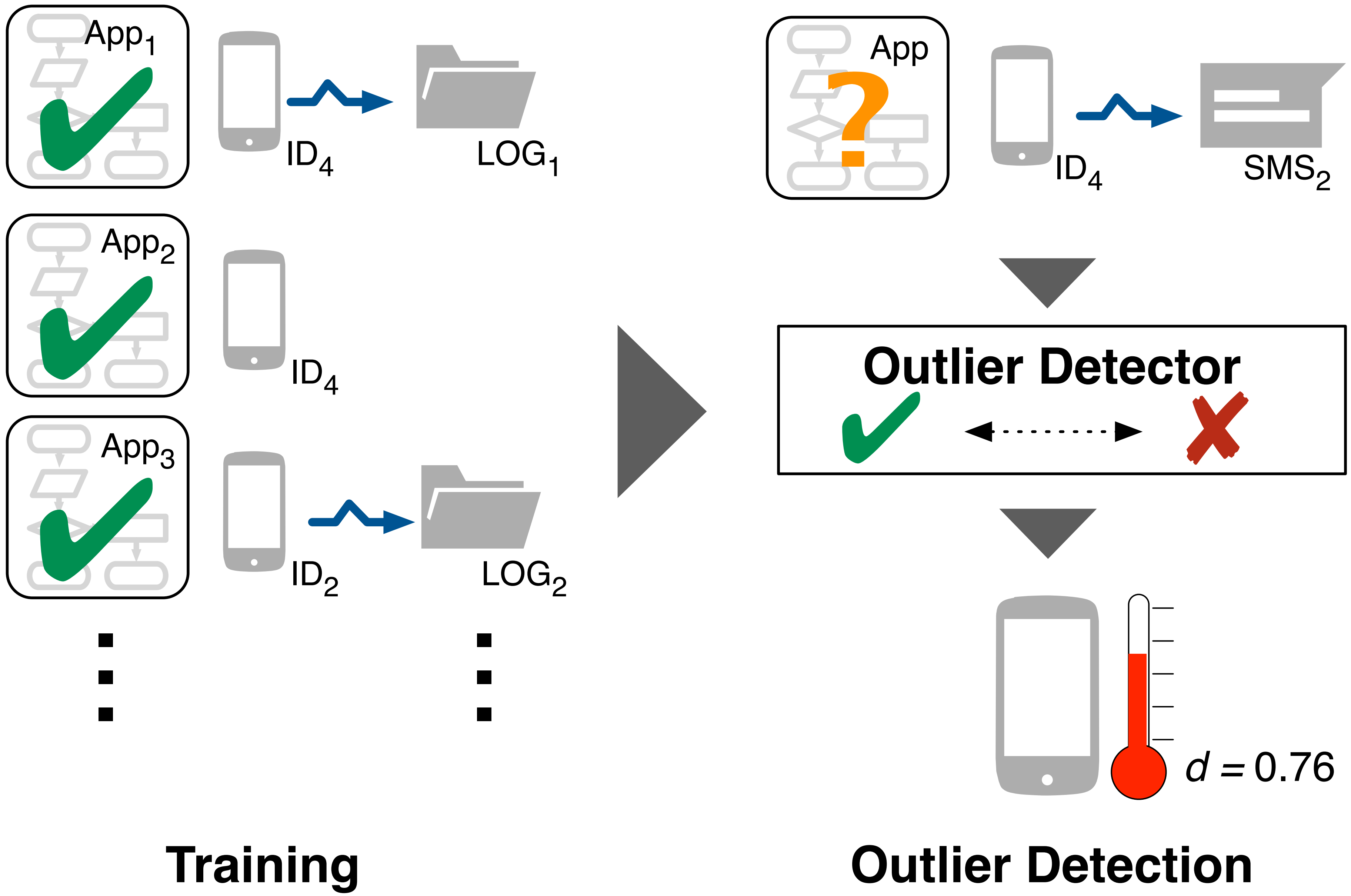
**Benign Apps**




**Malicious Apps**



# MUDFLOW



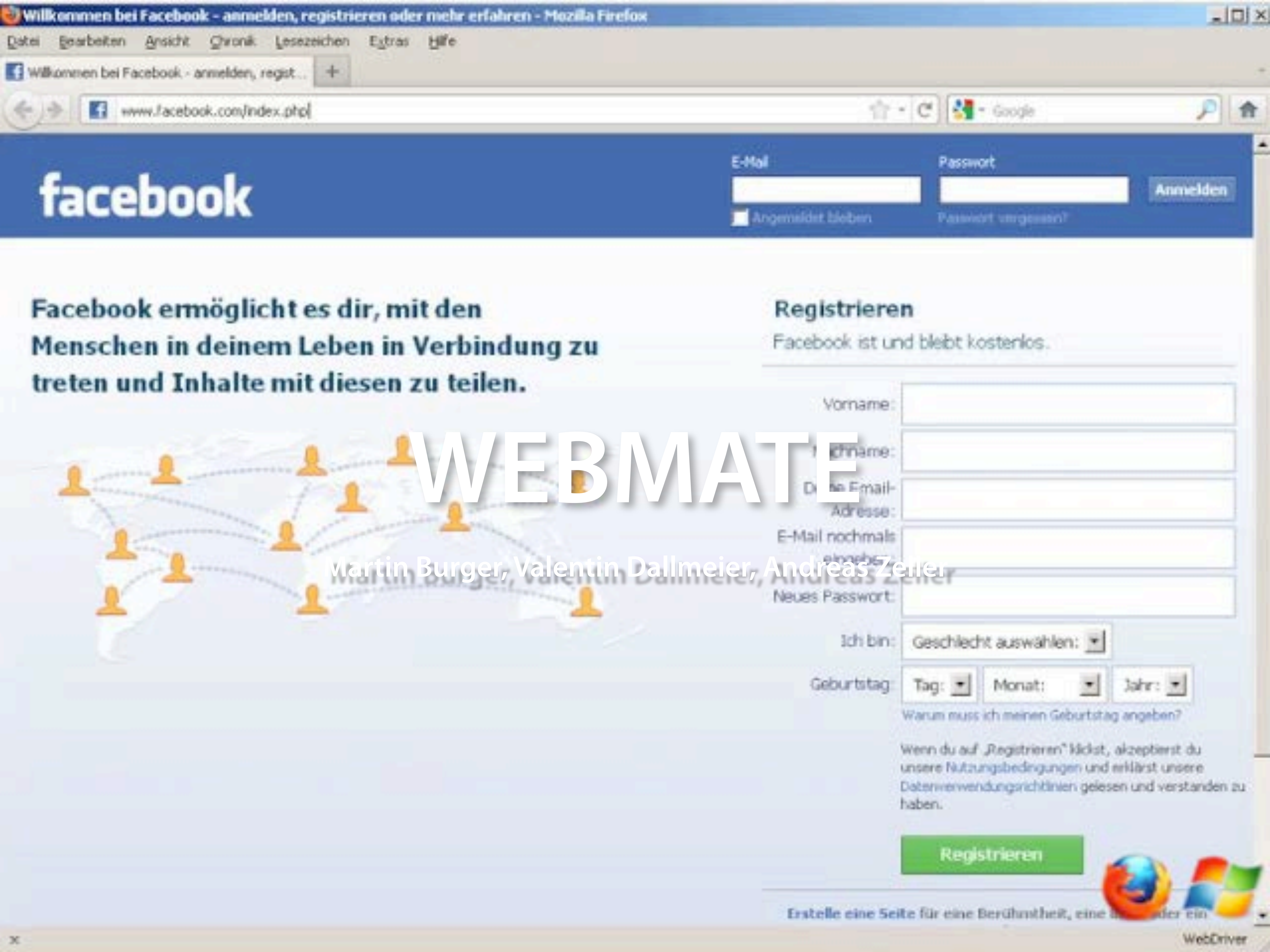

**Malware recognition rate >86%**

# Úlfar Erlingsson

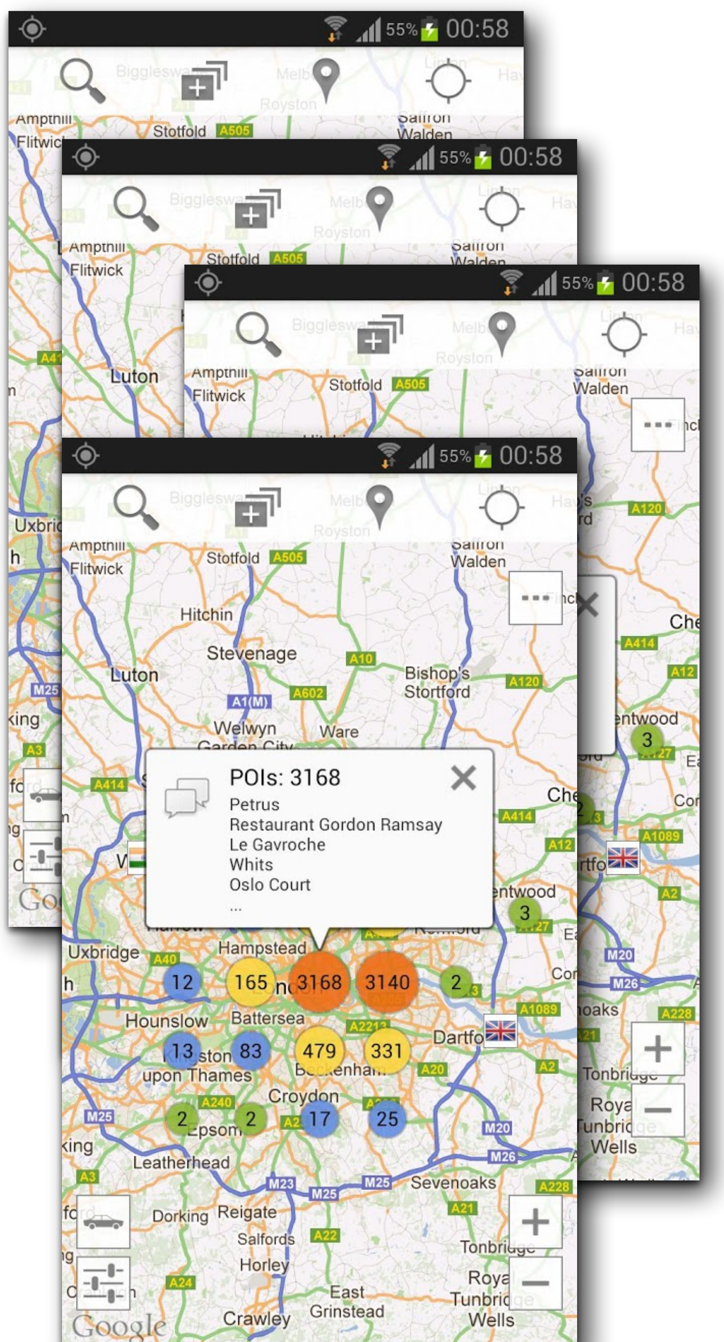

Error.
RAD
SHIFT
DRG
FSE
MR
MS
M+
hyp
sin
cos
tan
ln
log
%
7
9
DEL
AC

# WEBMATE

Martin Burger, Valentin Dallmeier, Andreas Zeller

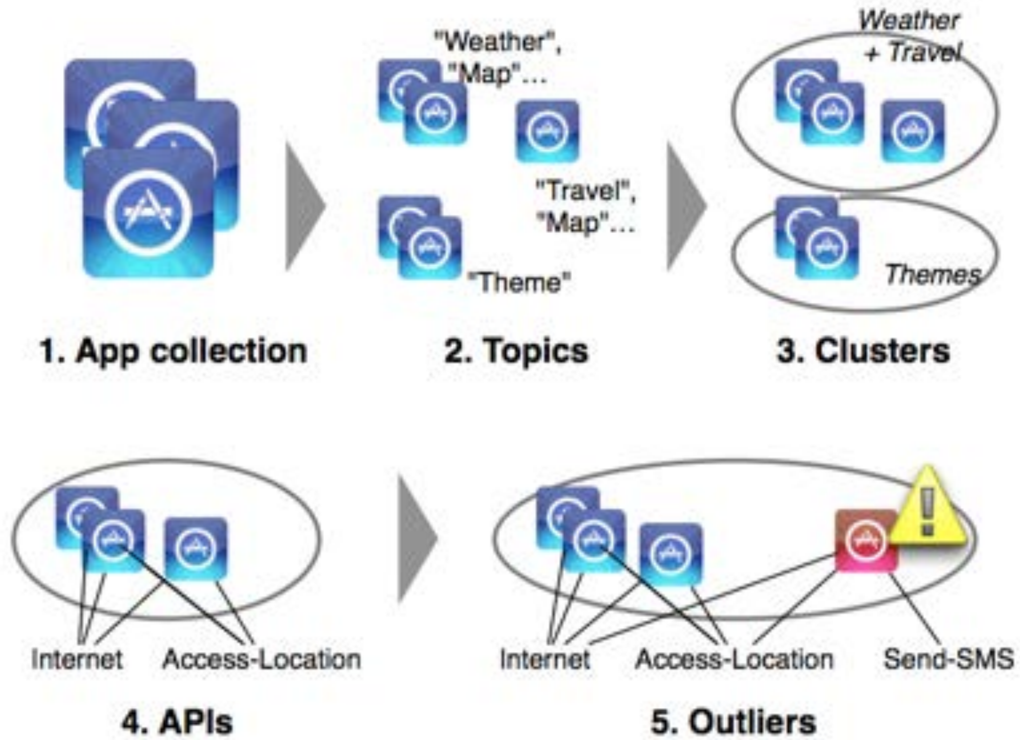


# App Mining




- For 100,000s of apps:
- Gather *descriptions*
- Gather *metadata*
- Gather *execution features*
- Find what is *common* and what is *uncommon*

# CHABADA




# Travel Cluster

Description


APIs used


# London Restaurants

android.net.ConnectivityManager.getActiveNetworkInfo()
android.webkit.WebView()
java.net.HttpURLConnection.connect()
android.app.NotificationManager.notify()
java.net.URL.openConnection()
**android.telephony.TelephonyManager.getDeviceId()**
org.apache.http.impl.client.DefaultHttpClient()
org.apache.http.impl.client.DefaultHttpClient.execute()
android.location.LocationManager.getBestProvider()
**android.telephony.TelephonyManager.getLine1Number()**
android.net.wifi.WifiManager.isWifiEnabled()
**android.accounts.AccountManager.getAccountsByType()**
android.net.wifi.WifiManager.getConnectionInfo()
android.location.LocationManager.getLastKnownLocation()
android.location.LocationManager.isProviderEnabled()
android.location.LocationManager.requestLocationUpdates()
android.net.NetworkInfo.isConnectedOrConnecting()
android.net.ConnectivityManager.getAllNetworkInfo()

**→ An *Outlier* in the "Travel" Cluster**

# Key Findings

- Of the top five API outliers per cluster,
  26% show unadvertised (covert) behavior
  - Typically *ad frameworks* (apploving, airpush)
  - Several anomalies (UNO, WICKED, Yahoo! Mail…)
- Using TF-IDF to classify API outliers per cluster,
  we could flag >80% of novel malware as such
- Current work: Dynamic API usage, information
  flow, user authorization

http://www.st.cs.uni-saarland.de/chabada/

erc Google