Clean Insights

A Privacy Preserving Mobile Measurement Platform
V0.3 July 2017

Nathanial Freitas

THE GUARDIAN PROJECT
https://guardianproject.info

A Berkman-Klein Assembly Project

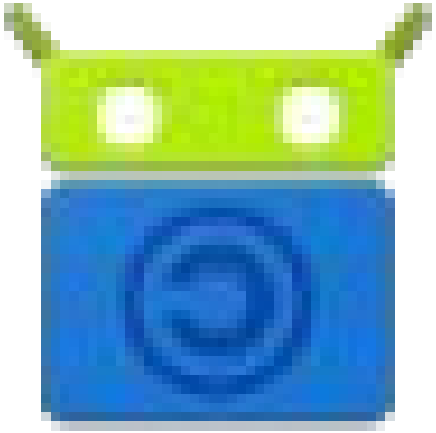
Guardian Project creates easy to use secure apps, open-source software libraries, and customized mobile devices that can be used around the world by any person looking to protect their communications and personal data from unjust intrusion, interception and monitoring.

THE GUARDIAN PROJECT

https://guardianproject.info

Orbot

Orfox



F-Droid

Network Security, Circumvention, Onion Routing and Anonymity

Obscuracam

CameraV
(InformaCam)

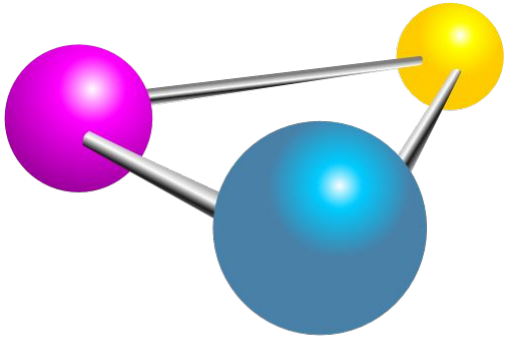
Proof Mode

ChatSecure

Conversations

Zom

Messaging: Encrypted, Resilient and Decentralized



CommonsWare

Empowering Developers… including Facebook and WeChat!

Success?

Not Sure. We have no good way to measure!

Assembly, at the Berkman Klein Center & MIT Media Lab, gathers developers, managers, and tech industry professionals for a rigorous spring term course on internet policy and a twelve-week collaborative development period to explore hard problems with running code.

@ https://bkmla.org/

Assembly Assembled!

How do we move beyond a world where virtually every computing device and network is insecure?

Decision makers, developers, and data scientists need to understand their products' effectiveness and their users' happiness.

This must not come at the cost of privacy, security and trust.

Doing it Wrong

Meitu details why it needs all those permissions.

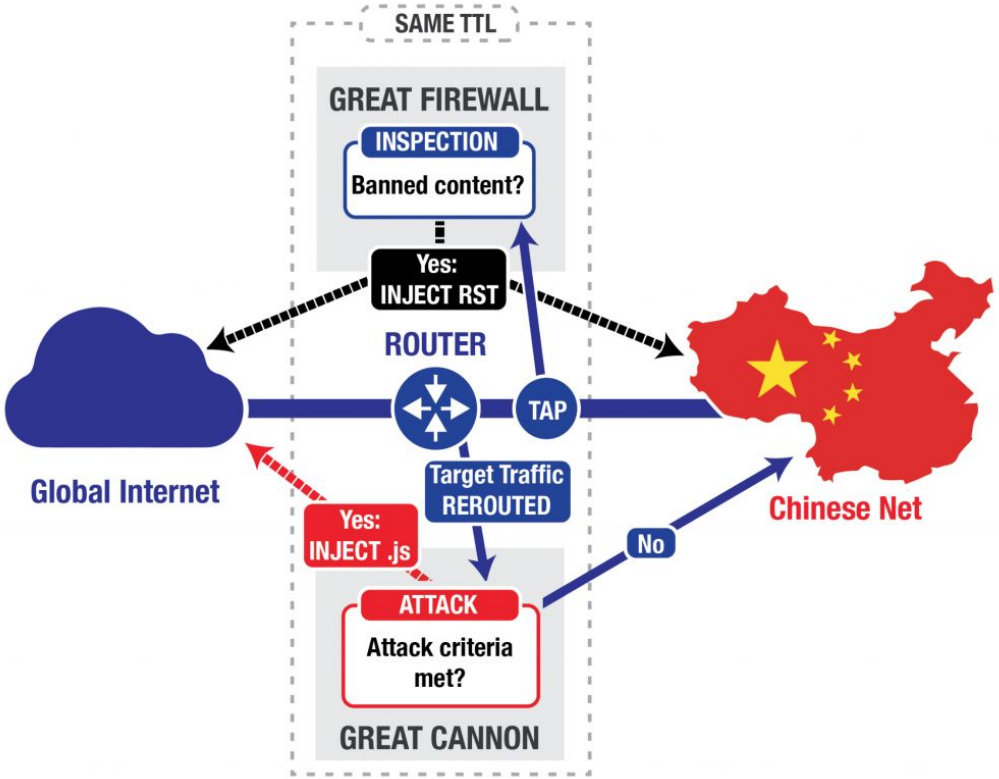
Chinese photo editing app Meitu made landfall in the U.S. recently, with the free app shooting up the Play Store rankings over the course of the week. The app adds anime-style filters to photos, an the final results end up being equal parts wonderful and weird.

Meitu also went into detail over the permissions it requires:

- **MAC address/IMEI number:** In some cases, Meitu cannot get both info at the same time and in some cases different devices even have the same IMEI number, so we combine these two details into one unique ID to track user devices.
- **LAN IP address** is used to prevent business fraud.
- **SIM card country code** is used for a rough location detection.
- **GPS and network location** are used for detecting countries and regions for Geo-based operation and advertisement placement.
- **Phone carrier info** is used as a standard tracking channel for analytics, just like the other third-party analytics tools(e.g., Flurry).
- **RUN_AT_START:** because the Google service (including GCM) is not available in mainland China, Meitu uses a third-party push notification service called Getui (www.getui.com).

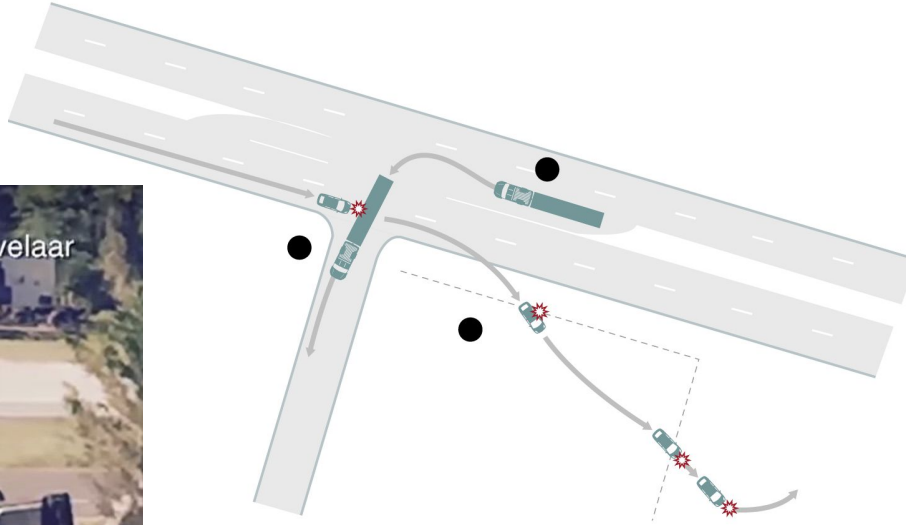
That's certainly a lot to put up with for a photo filter app. If you're satisfied with Meitu's explanation, the app is available for free from the Play Store.

All your permissions and data belong to us!



In the attack on GitHub and GreatFire.org, the GC intercepted traffic sent to Baidu infrastructure servers that host commonly used analytics, social, or advertising scripts.

Weaponized users though insecure analytics



REPORT: TESLA'S FATAL CRASH CAN'T BE BLAMED ON SOFTWARE ERRORS

“Blackbox” exonerates corporation

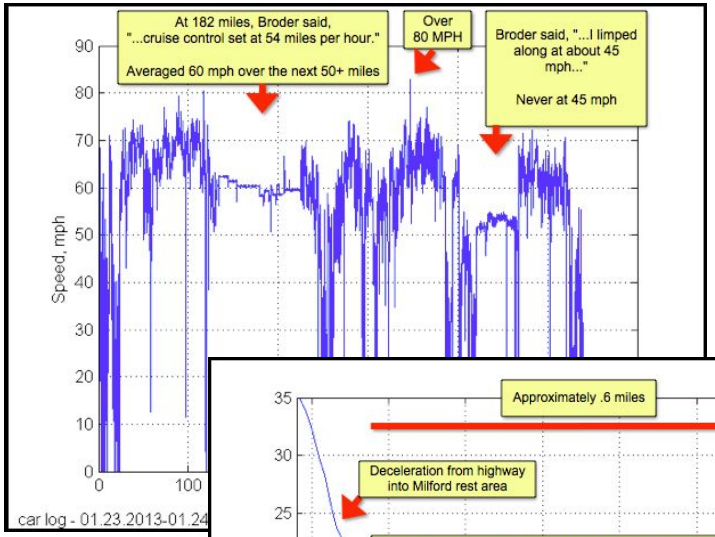
Tesla publishes Model S driving logs that show The New York Times' blatant lies

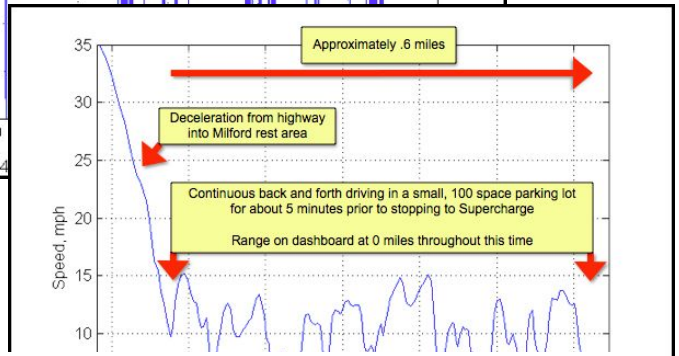
By Sebastian Anthony on February 14, 2013 at 8:17 am | 143 Comments

0 shares

Following Elon Musk's initial denouncement of The New York Times for publishing a fake review of the Tesla Model S electric car, he has now published the actual logs recorded by the car — and boy are they damning. In short, the NYT's John Broder lied through his teeth to smear electric vehicles in general, and the Model S in specific.

The basic premise of John Broder's story for The Times was that the car lied about its self-reported estimated remaining range; when it said there was 79 miles left in the batteries, there was in actual fact only 60. Eventually, after a few such cases of the on





A Most Peculiar Test Drive

Elon Musk, Chairman, Product Architect & CEO February 13, 2013

"Blackbox" incriminates the user

NETWORKWORLD
FROM IDG

PRIVACY AND SECURITY FANATIC
By Ms. Smith | Follow

About
Ms. Smith (not her real name) is a freelance writer and programmer with a special and somewhat personal interest in IT privacy and security issues.

Cops use pacemaker data to charge homeowner with arson, insurance fraud

Police called pacemaker data an 'excellent investigative tool' that provided 'key pieces of evidence' to charge a man with arson and insurance fraud

Network World | JAN 30, 2017 7:08 AM PT

RELATED

Cool Yu… disrupti…

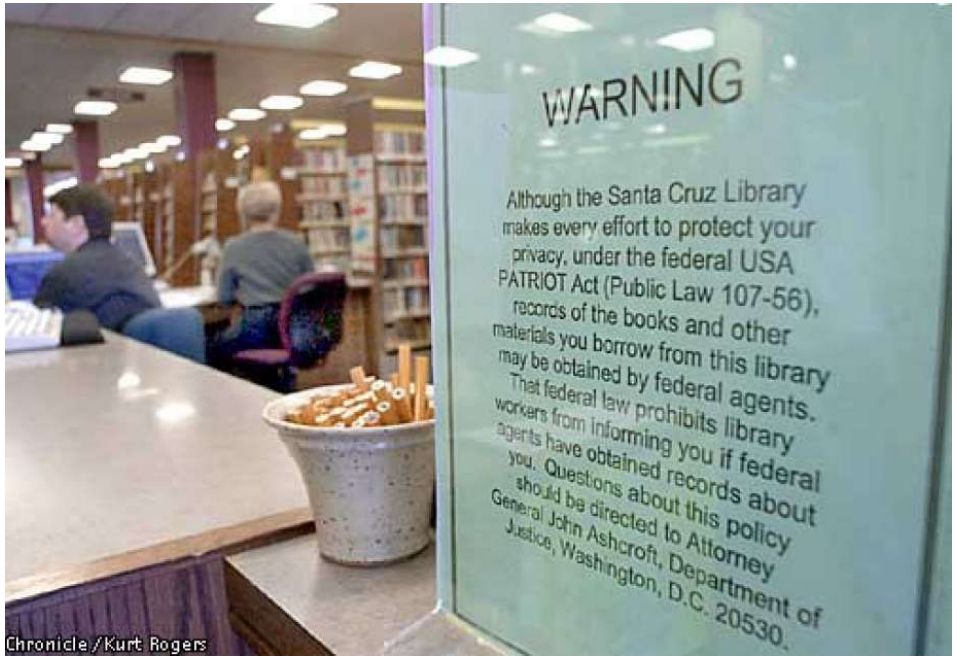
3 replac… phones… week

Pastor: … wife's n… sent pic…

VIDEO

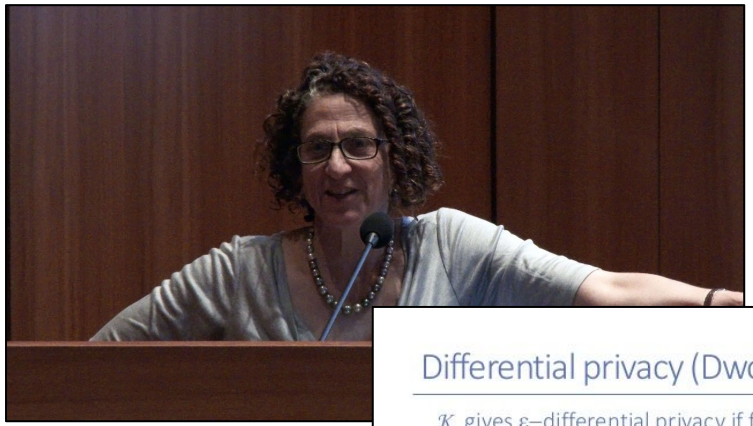
Middletown Police said this was the first time it had used data from a heart device to make an arrest, but the pacemaker data proved to be an “excellent investigative tool;” the data from the pacemaker didn’t correspond with Compton’s version of what happened. The retrieved data helped to indict Compton.

Your body will be used against you

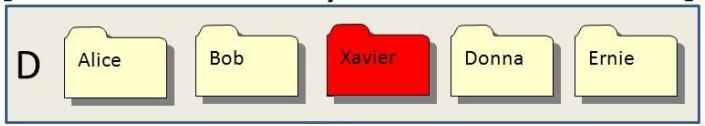


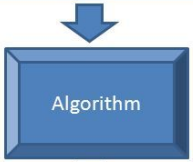
Be careful what you read

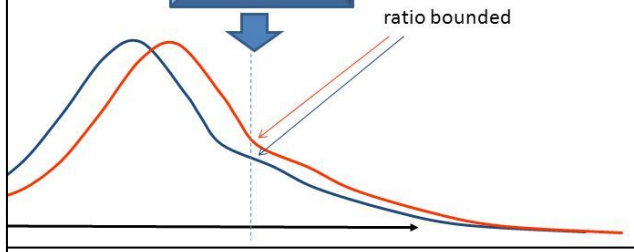
Existing Work



Differential Privacy [Dwork-McSherry-Nissim-Smith 06]



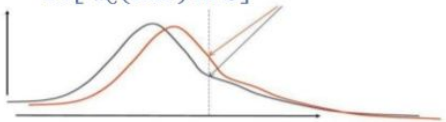




Differential privacy (Dwork 2006)

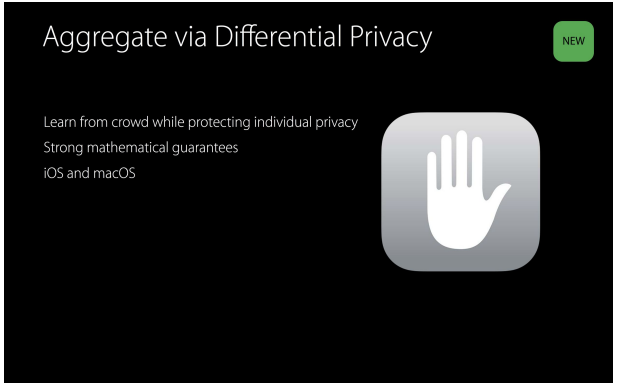
$\mathcal{K}$ gives ε–differential privacy if for all values of DB, DB' differing in a single element, and all S in $\mathrm{Range}(\mathcal{K})$

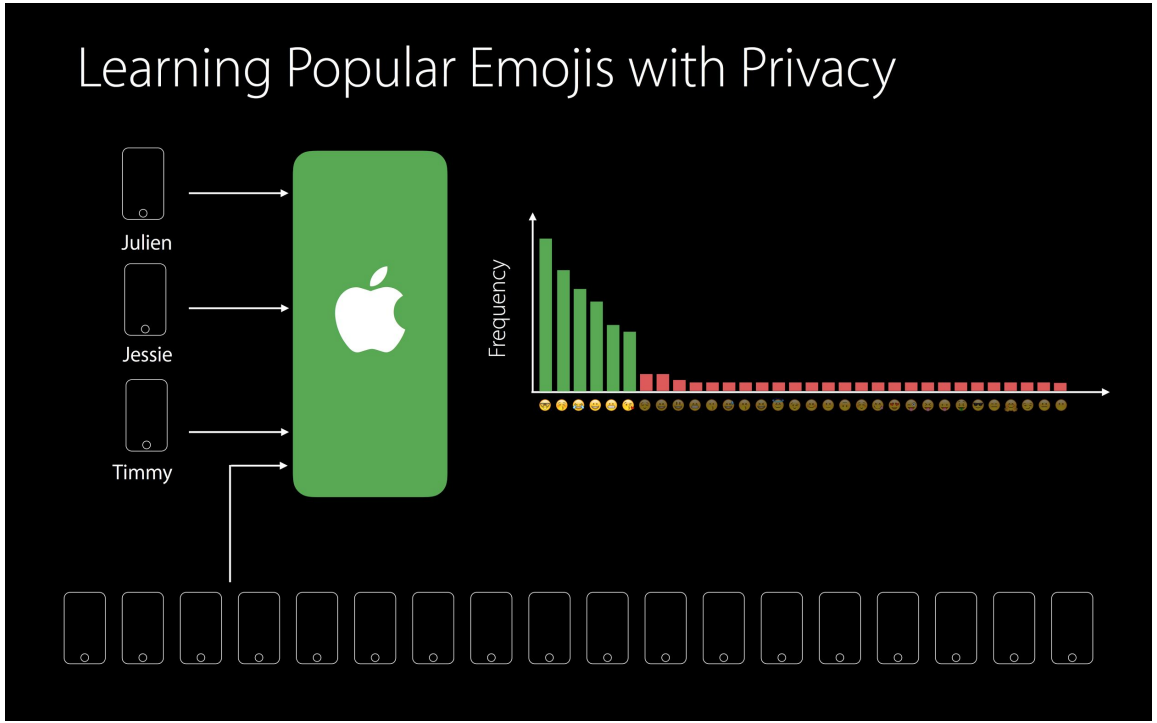
$$\frac{\Pr[\mathcal{K}(\mathrm{DB}) \text{ in } S]}{\Pr[\mathcal{K}(\mathrm{DB}') \text{ in } S]} \leq e^{\varepsilon} \sim (1+\varepsilon)$$



3

Dwork 06





Apple Thinks Differentially

Google Security Blog

The latest news and insights from Google on security and safety on the Internet

Learning statistics with privacy, aided by the flip of a coin

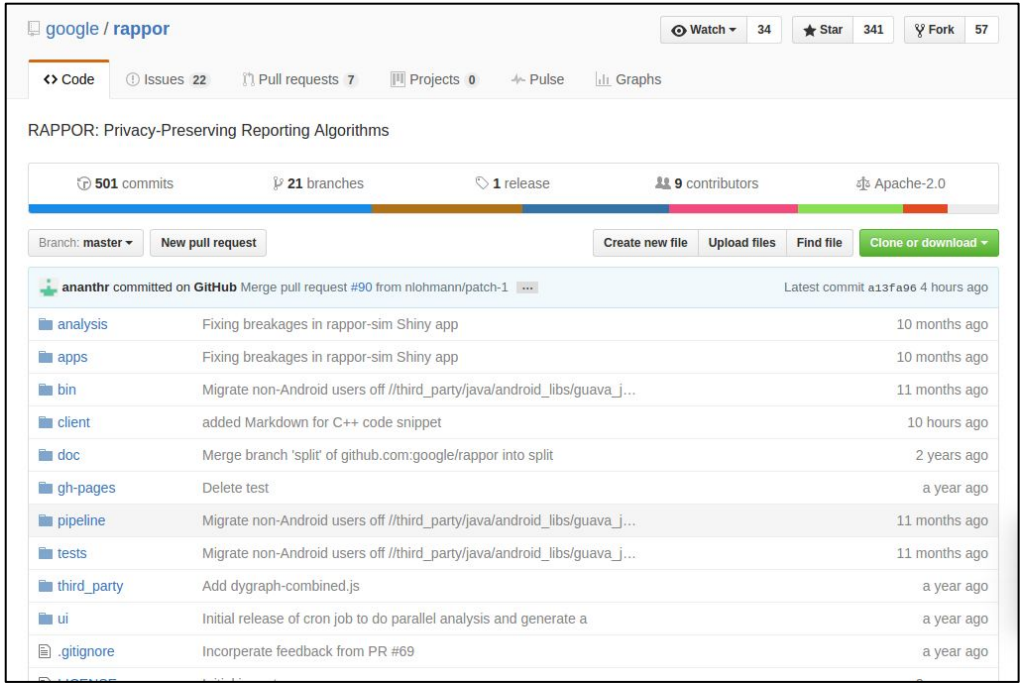
October 30, 2014

Cross-posted on the Research Blog and the Chromium Blog

At Google, we are constantly trying to improve the techniques we use to protect our users' security and privacy. One such project, RAPPOR (Randomized Aggregatable Privacy-Preserving Ordinal Response), provides a new state-of-the-art, privacy-preserving way to learn software statistics that we can use to better safeguard our users' security, find bugs, and improve the overall user experience.

Building on the concept of randomized response, RAPPOR enables learning statistics about the behavior of users' software while guaranteeing client privacy. The guarantees of differential privacy, which are widely accepted as being the strongest form of privacy, have almost never been used in practice despite intense research in academia. RAPPOR introduces a practical method to achieve those guarantees.

To understand RAPPOR, consider the following example. Let's say you wanted to count how many of your online friends were dogs, while respecting the maxim that, on the Internet, nobody should know you're a dog. To do this, you could ask each friend to answer the question "Are you a dog?" in the following way. Each friend should flip a



Google Open-Source Rappor

Solutions do exist...

RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response

Úlfar Erlingsson
Google, Inc.
ulfar@google.com

Vasyl Pihur
Google, Inc.
vpihur@google.com

Aleksandra Korolova
University of Southern California
korolova@usc.edu

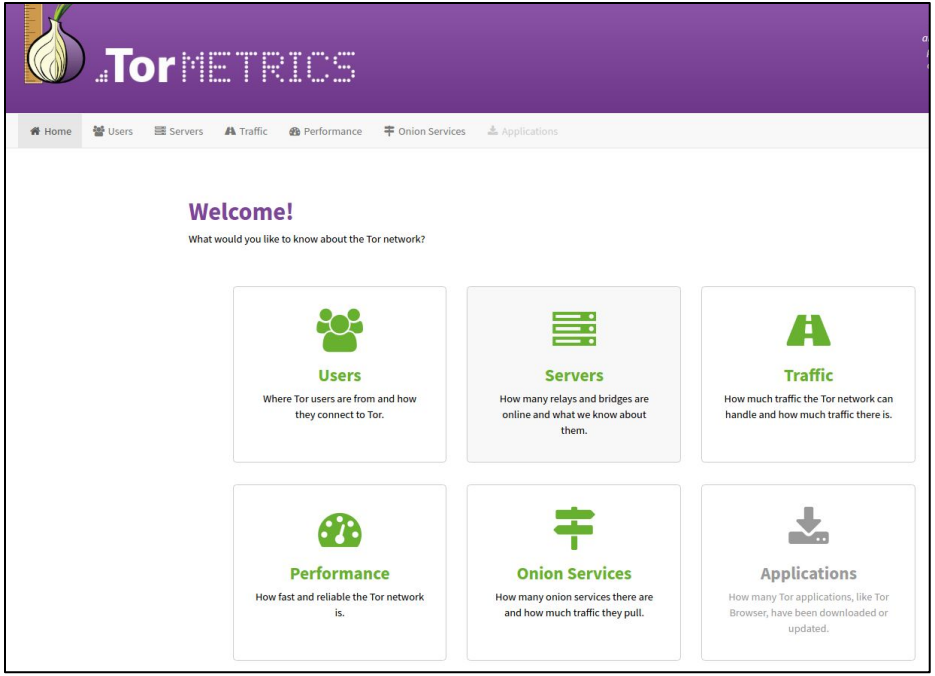
Google

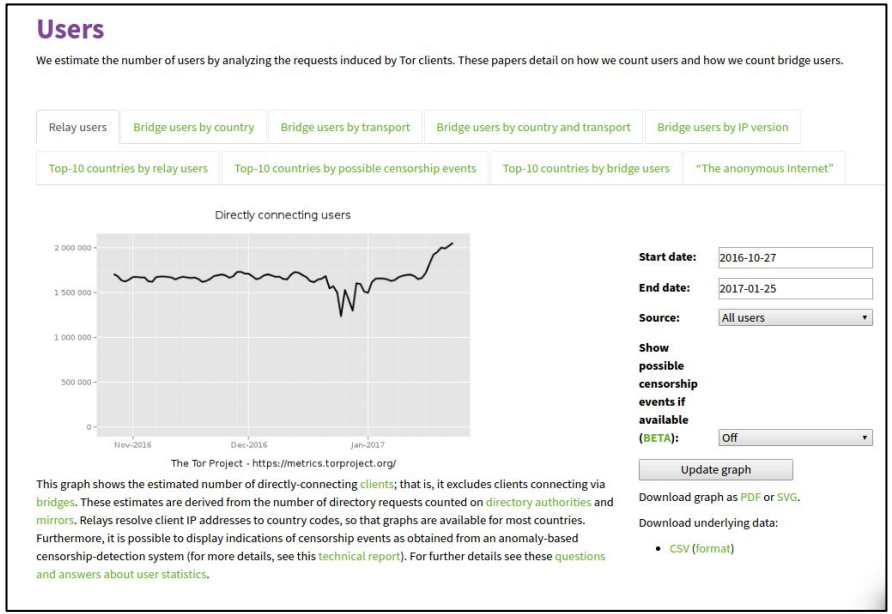
WIRED

ANDY GREENBERG SECURITY 06.13.16 7:02 PM

APPLE'S 'DIFFERENTIAL PRIVACY' IS ABOUT COLLECTING YOUR DATA—BUT NOT *YOUR* DATA

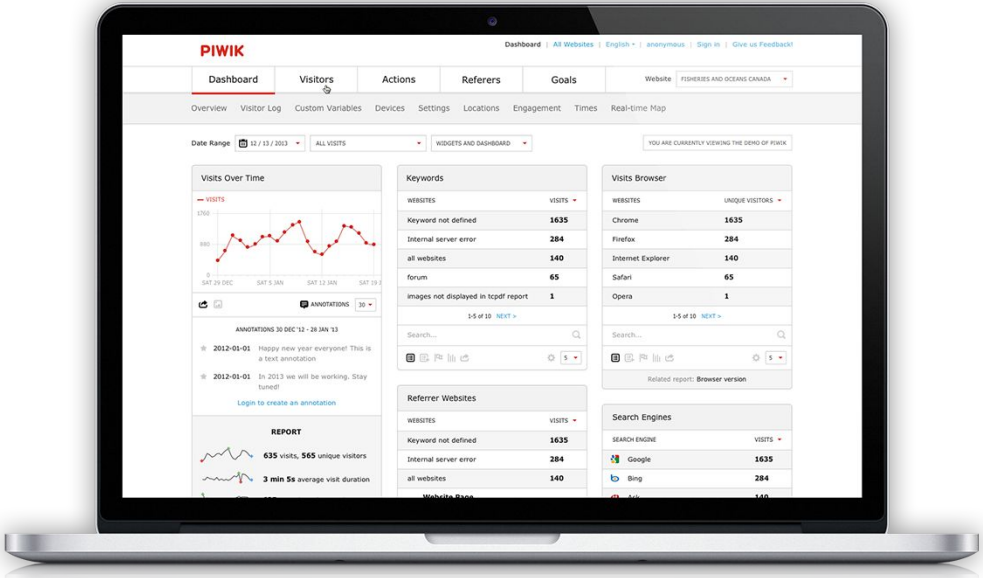
...but are not readily available to most





Tor's Anonymous Metrics

Piwik is leading open-source analytics systems, with 100% data ownership, user privacy protections, and extensibility.



ACRA

Acra catches exceptions, retrieves **lots of context data** and send them to the **backend of your choice**.

Best of all, it is **FREE** and **OPEN SOURCE**.

Expressed Needs

—

Is the network latency reduced since last week?

Do users like the change in the user interface?

How many people are typically in a group chat?

Is the battery usage better or worse with the new version?

How many average conversations do you users have open?

What Developers of Secure Messaging Apps Want to Know

—

A first step would be to not have identifiable information on well-behaving users.

I would love if we could have something self hosted, secure and useful that we could use to make our app a better tool but we couldn't find anything that would meet those requirements.

I feel very uncomfortable, but the mis-use over our system is a problem that poses a threat to the whole project.

We don't know which features are popular. Therefore, it is hard to expand the service.

The biggest thing we generally use analytics for is just tracking the health of our service in terms of overall user growth and location.

We (a funder) are frequently asked for additional metrics demonstrating the impact of the projects we support.

Feedback from our broad user survey

Half of Smartphone Owners Don't Want Their Locations Tracked Study suggests mobile marketers should cool it By Lauren Johnson

July 23, 2014, 10:12 AM EDT | Technology

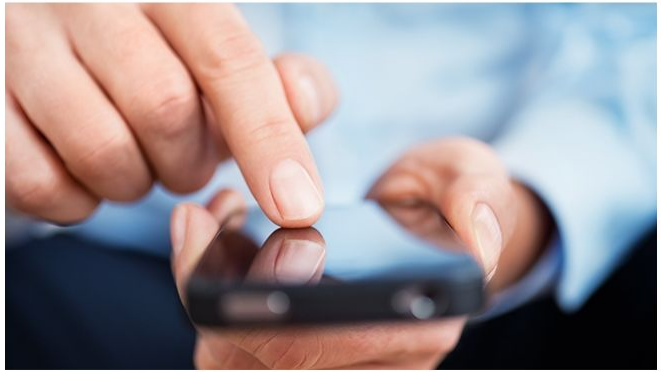
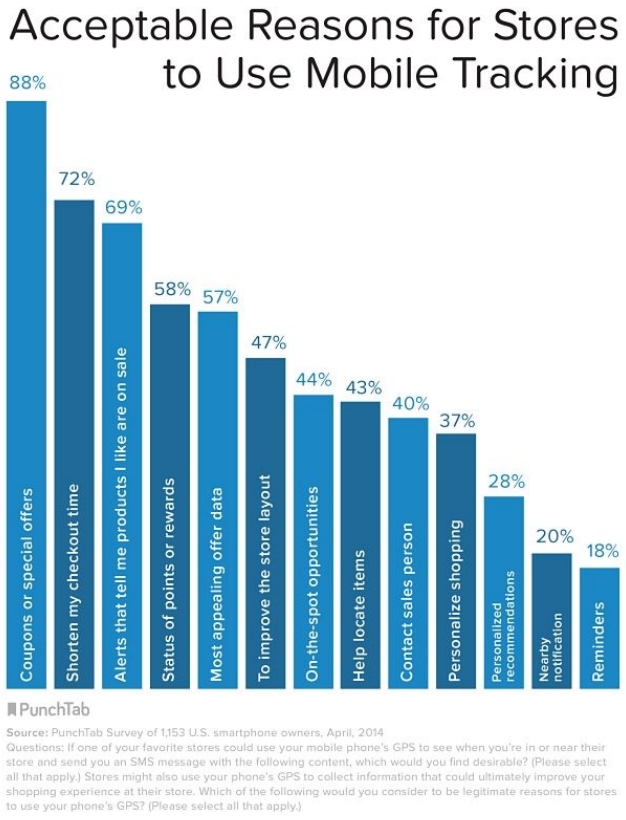


Photo: Getty Images



Consumers Want Coupons Without Being Geo-Located

We want developers to have a means to understand how to improve, but to do so in a way that respects privacy and security

Research Design Sprint

Phase 1: Research

- Identify privacy problems of existing app analytic tools
- Threat Modeling
- Identify privacy requirements of target apps
- Study users' privacy expectation for these apps
- Identify key metrics needed by target apps
- Study Differential Privacy
- **Deliverable: Build up knowledge in the problem domain and possible solutions**

Threat Modeling

Vulnerable Assets

- Unique user / hardware identifiers
- Internet addresses
- Biometric data
- Geolocation data
- Social graphs
- Behavioral logs
- Political preferences
- And on, and on....

Attack Vectors

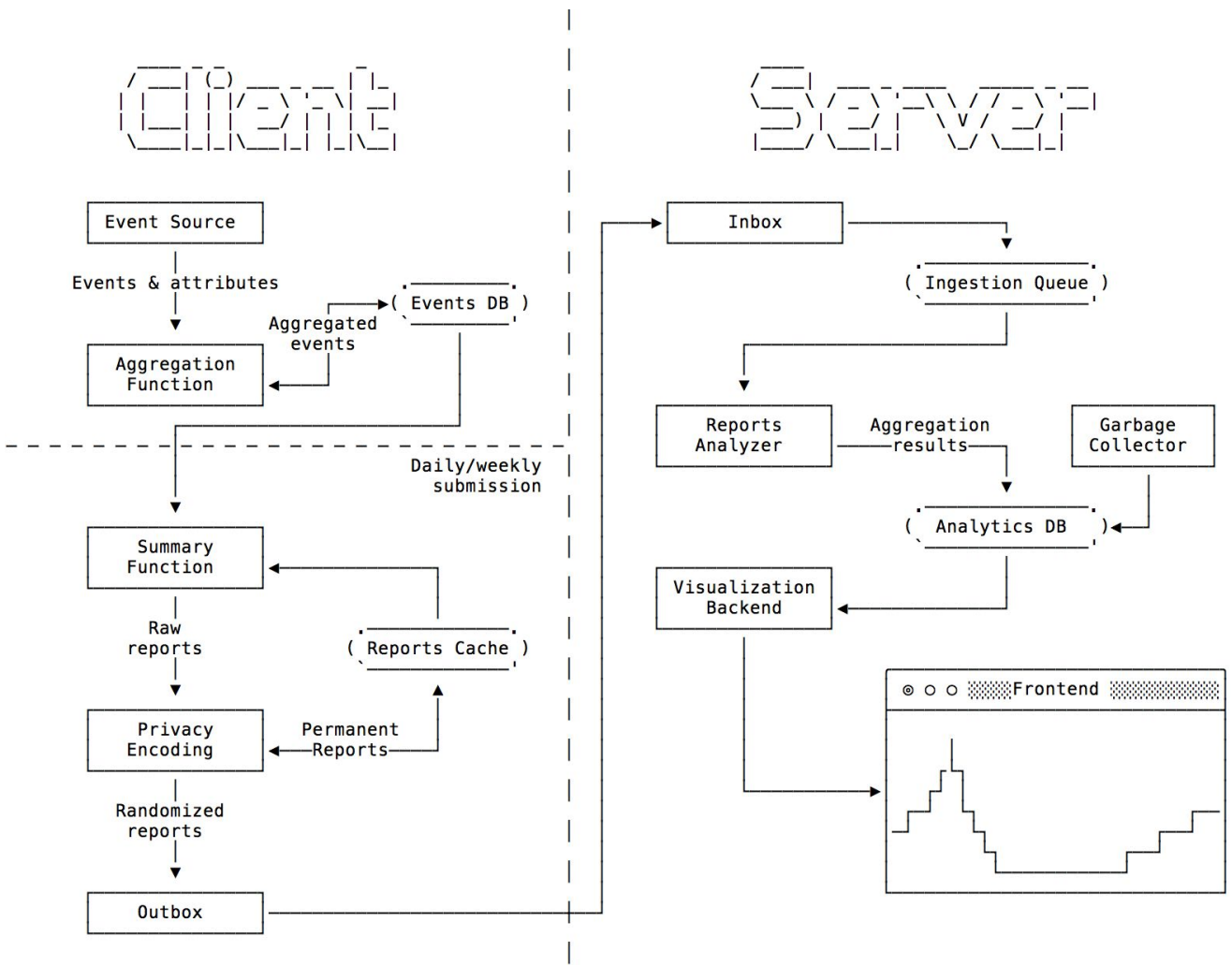
- Client/device temporary cache
- Server storage database
- Developer code libraries
- Network transports
- The Algorithm

Potential Mitigations

- FLOSS code
- Hardened network transports
- Data minimization
- Client-side processing
- Differential Privacy, Randomized Response, Randomized Controlled Trials and more...

Phase 2: Design

- Scope the MVP
- One mobile platform and web
- Wireframe/visual design
- Software architecture design
- Data processing pipeline
- **Deliverables: A vision about the product; a scope that can fit into the schedule.**

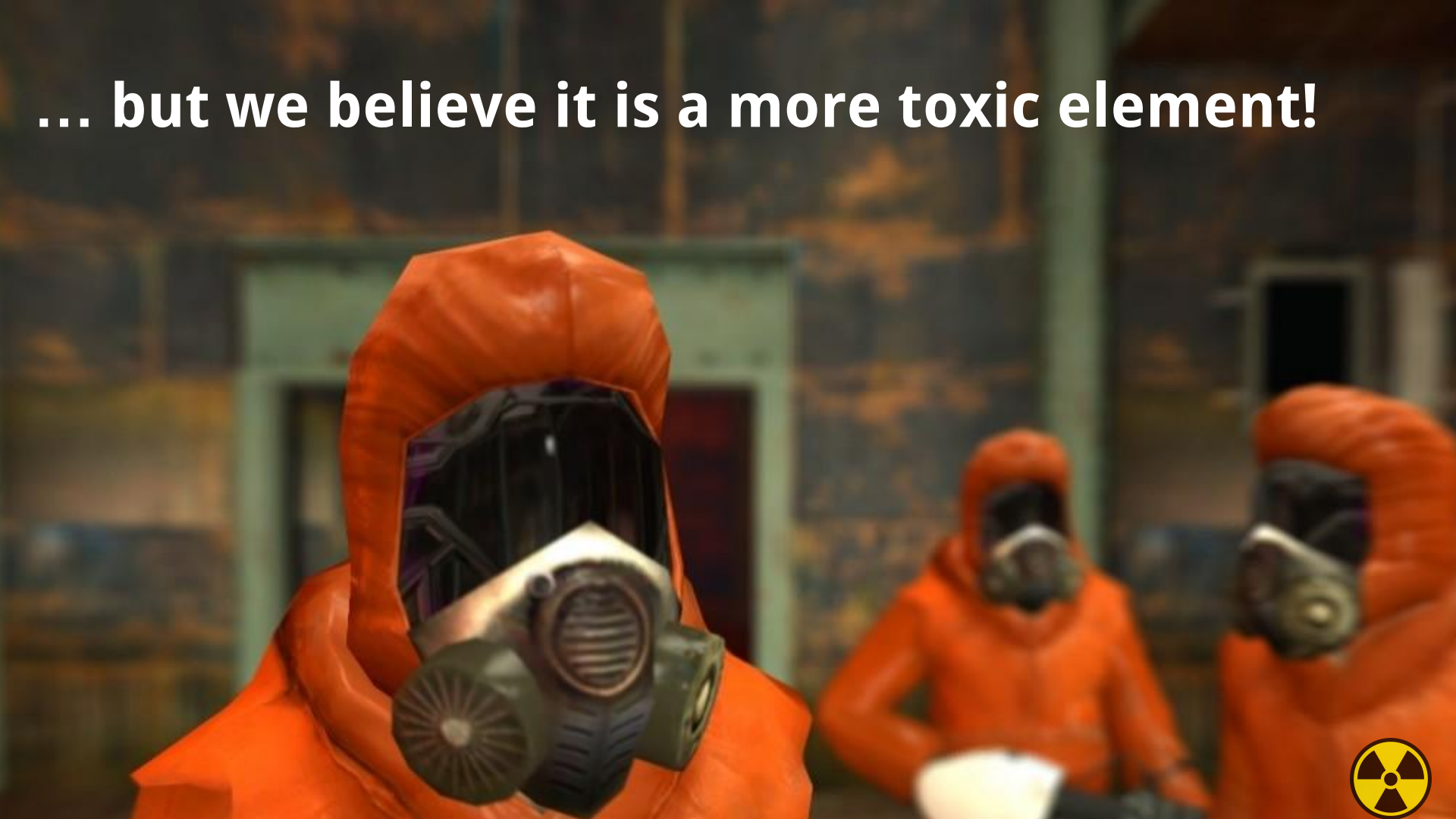

Client
Server
Event Source
Events & attributes
Aggregation Function
Aggregated events
Events DB
Daily/weekly submission
Summary Function
Raw reports
Privacy Encoding
Randomized reports
Outbox
Reports Cache
Permanent Reports
Inbox
Ingestion Queue
Reports Analyzer
Aggregation results
Garbage Collector
Analytics DB
Visualization Backend
Frontend

Phase 3: Implement / Sprint

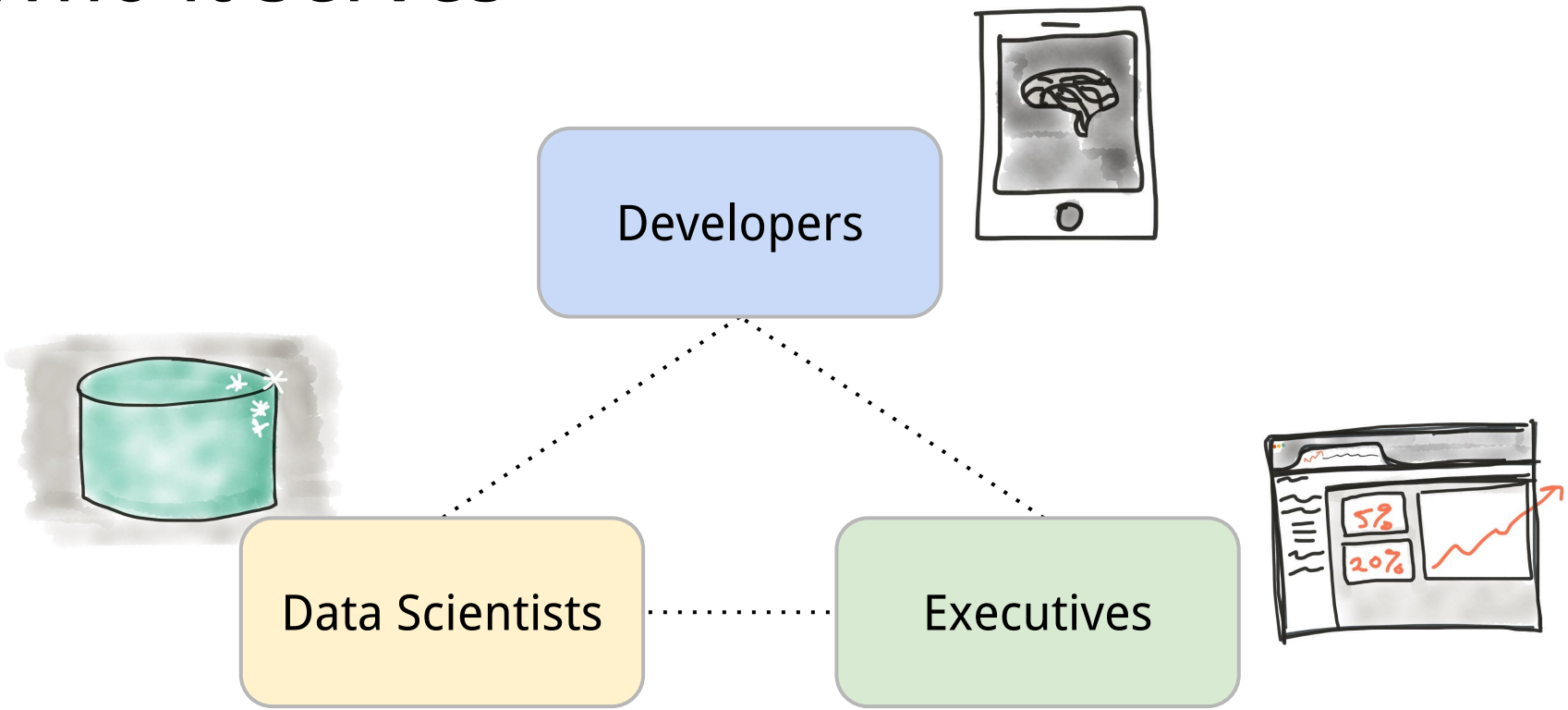
- Data processing pipeline: DP algorithms
- Frontend: Analytic panels, Account and authorization, Data visualization
- Backend:API, Persistence
- Collector: Run by user apps to collect and preprocess data
- Privacy policy
- **Deliverables: A working MVP.**

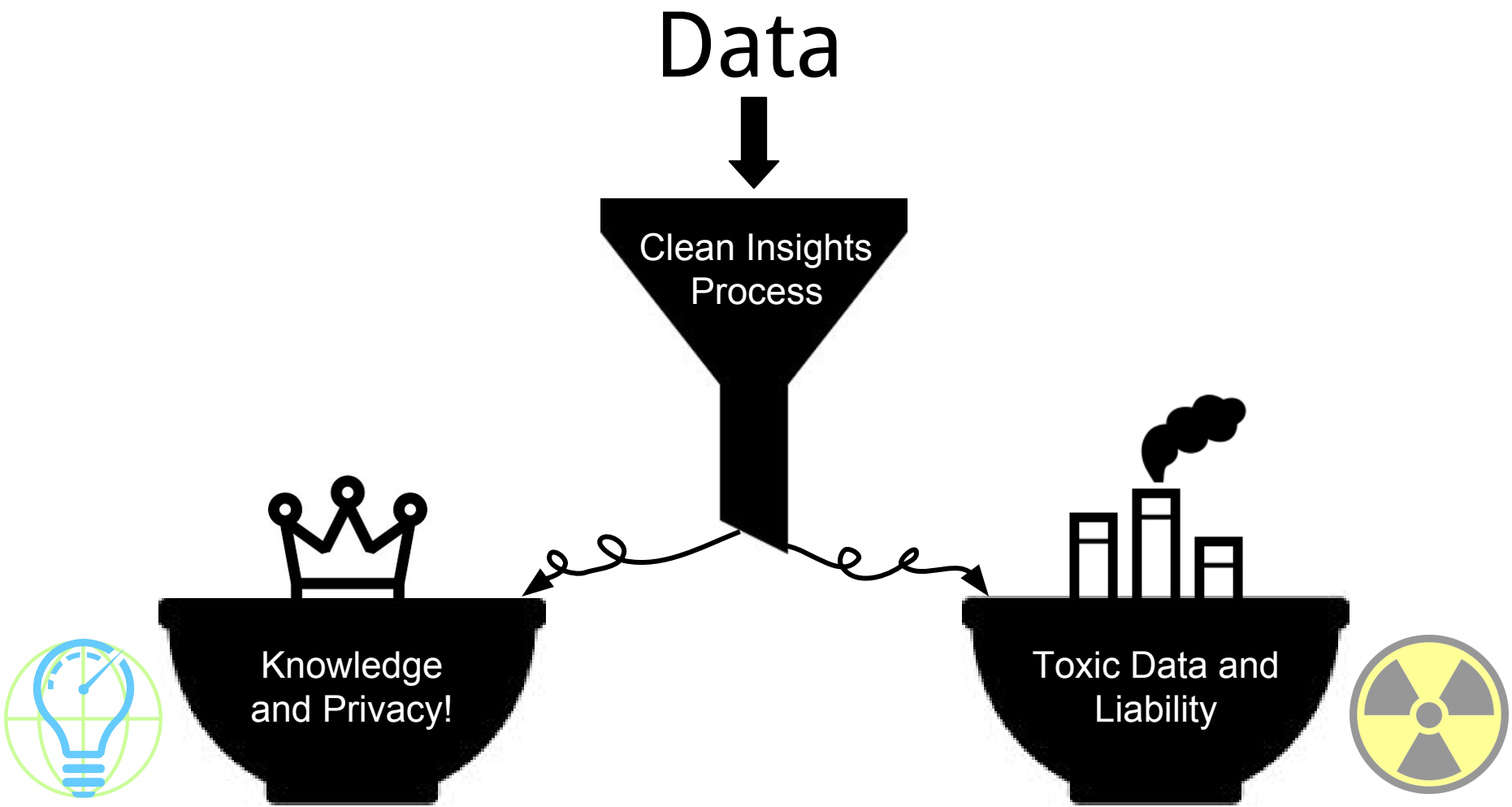
The “Clean Insights” Pitch

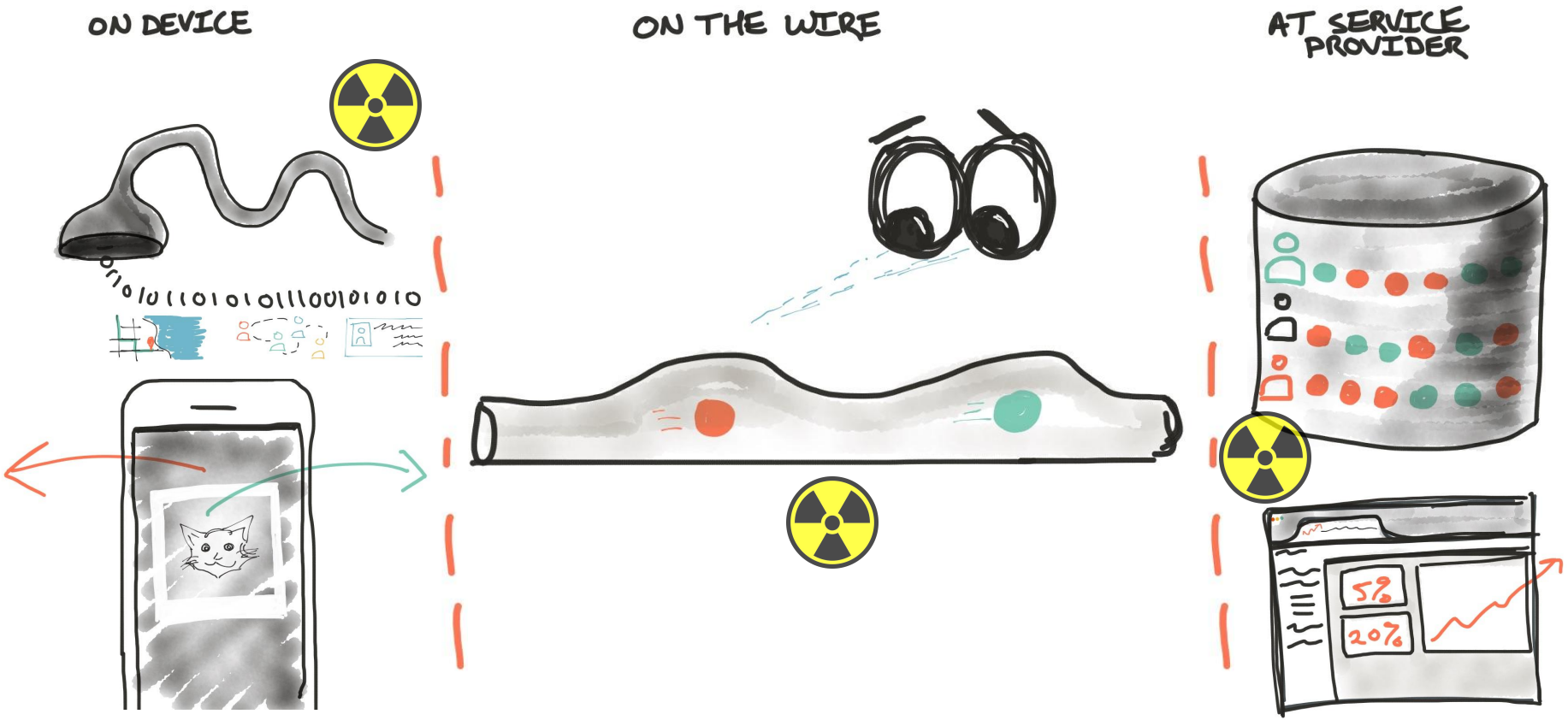
Most companies treat data like gold...


… but we believe it is a more toxic element!

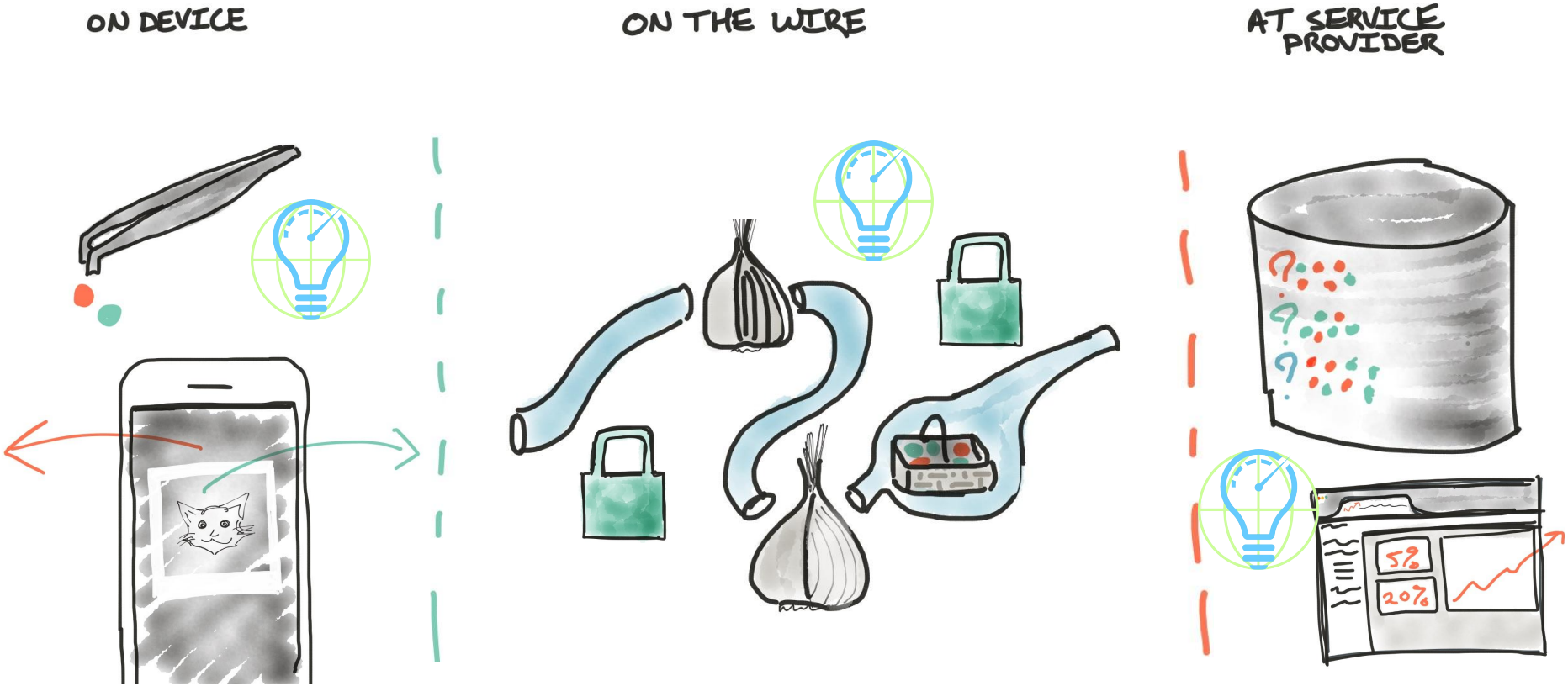
Who it serves




Data
Clean Insights Process
Knowledge and Privacy!
Toxic Data and Liability



Typical analytics services vacuum up every interaction, transmitting the raw data to a centralized cloud data warehouse, often in an insecure manner.



Clean Insights pushes data processing to the edge, selectively collecting and sharing through secure channels, to a self-hosted backend server.

Three Tenets of Clean Insights

Hardened Security

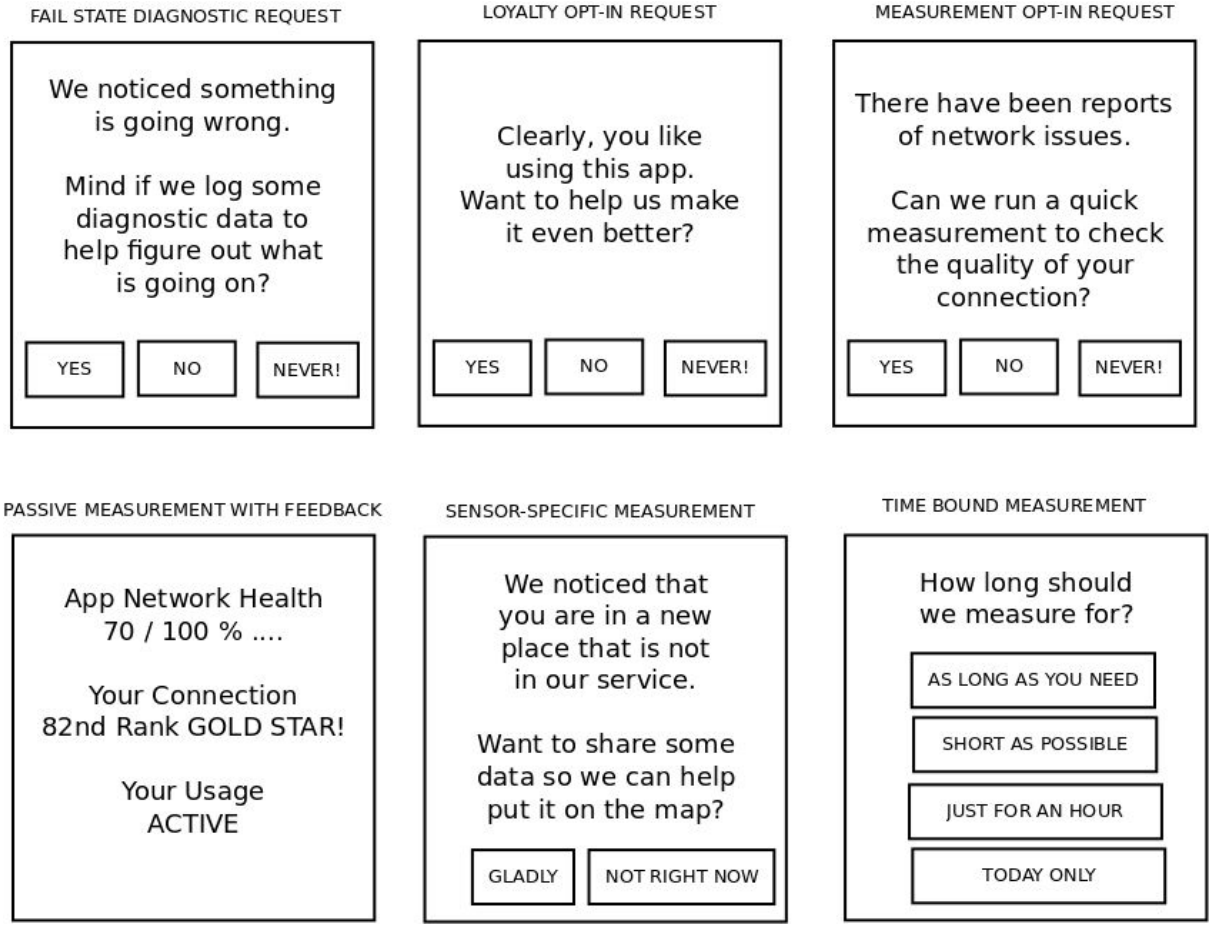
Certificate Pinning
TLS Best Practices
Onion Routing

Powerful Privacy Toolbox

Data Batching
Smart Thresholds
No Perma Cookies

Advanced Anonymity

Differential Privacy
Randomized Response
Machine Learning



Empowering & Engaging User interactions

```
if (CleanInsights.getInstance(mPiwik.getContext()).isTorEnabled())
{
    int proxyPort = CleanInsights.getInstance(mPiwik.getContext()).getTorHttpPort();
    mProxy = new Proxy(Proxy.Type.HTTP, new InetSocketAddress("127.0.0.1",proxyPort));
}
```

```
StrongOkHttpClientBuilder builder= StrongOkHttpClientBuilder.forMaxSecurity(mPiwik.getContext());

OkHttpClient client = null;

if (mCertPin != null) {
    CertificatePinner certificatePinner = new CertificatePinner.Builder()
            .add(packet.getTargetURL().getHost(), mCertPin)
            .build();

        client = new OkHttpClient.Builder()
                .proxy(mProxy)
            .certificatePinner(certificatePinner)
            .build();
}
```

Onion Routing & Certificate Pinning

```java
@Override
public boolean allowMeasurement() {

    if (checkLocationPermission()) {

        //get the last good current location
        Location locationNow = locationManager.getLastKnownLocation(LocationManager.PASSIVE_PROVIDER);

        //check the distance between this location, and the user provided one
        float distanceNow = locationNow.distanceTo(locationNear);

        //if within distance, then allow measurement
        if (distanceNow <= distanceLimit)
            return true;

    }

    return false;
}
```

```java
//only measure when the user is in the app for longer than 60 seconds
getMeasurer().addThreshold(new SessionLengthThreshold(true, 60));

try {
    //only measure between the specified dates and/or times
    Date startDate = SimpleDateFormat.getDateInstance().parse("4/20/2017");
    Date endDate = SimpleDateFormat.getDateInstance().parse("4/21/2017");

    //measure when between these dates, but DON'T require if another threshold matches
    getMeasurer().addThreshold(new DateThreshold(false, startDate, endDate));
}
catch (ParseException pe){}
```

Code: Thresholds of time and space

```java
@Override
protected void onPause() {
    super.onPause();

    //when the app pauses do a private, randomized-response based tracking of the number of likes
    MeasureHelper.track().privateEvent("Vote", "Like per Session", Integer.valueOf(mLikeCount).floatValue(), getMeasurer())
            .with(getMeasurer());

    //dispatch the current set of events to the server
    ((CleanInsightsApplication)getApplication()).getMeasurer().dispatch();
}
```

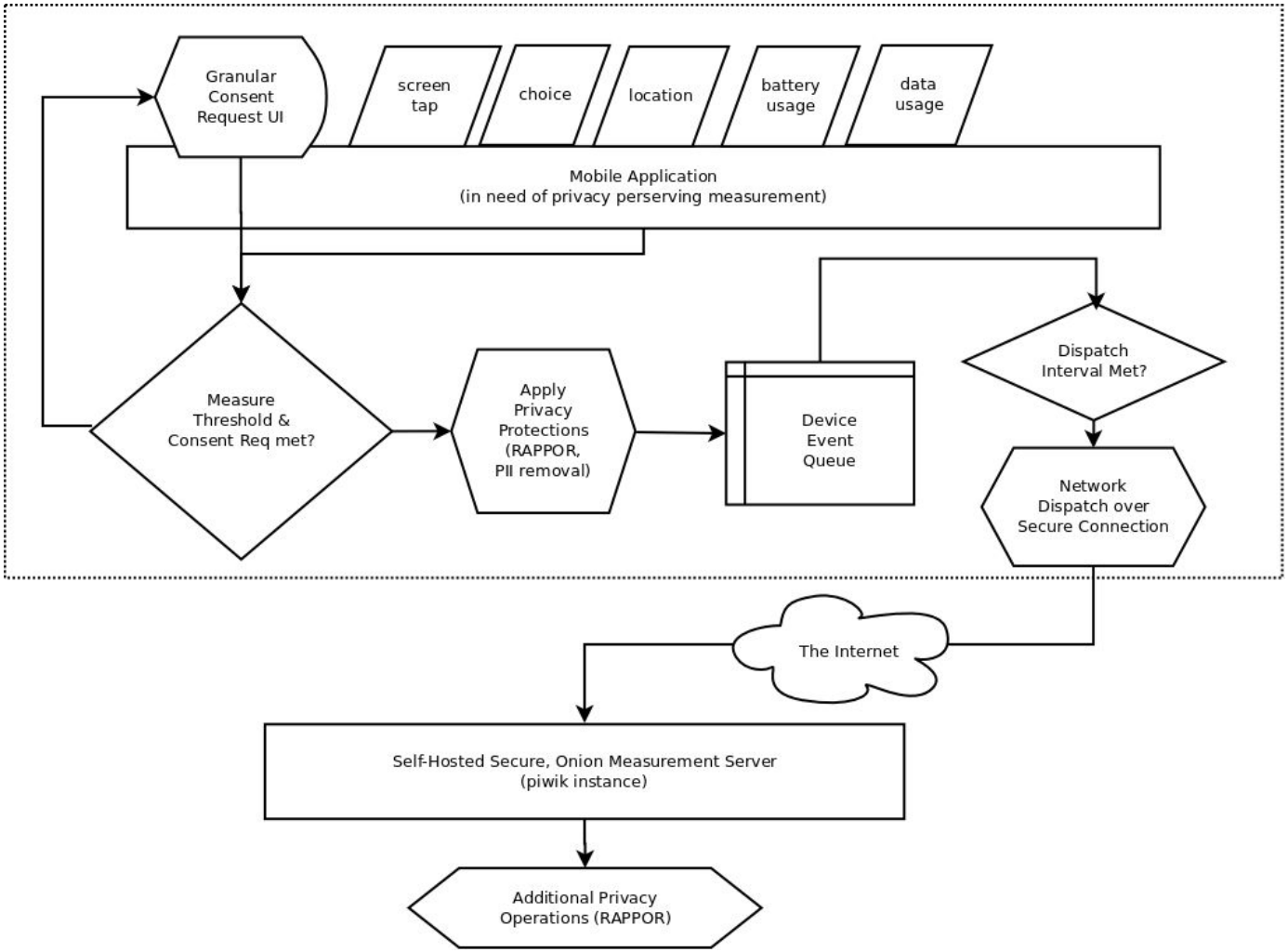
Code: Private Measurement and Dispatch

```
private Encoder createRandomizingEncoder() {
    // TODO: Choose appropriate parameters
    return new Encoder(mMeasurer.getUserSecret(),
            ENCODER_ID,
            4096,
            13.0 / 128.0,
            0.25,
            0.75,
            1,
            2);
}
```

```
  public synchronized MeasureMe set(@NonNull QueryParams key, int value) {
      final String stringValue;
      if (key == QueryParams.EVENT_VALUE) {
          key = QueryParams.EVENT_NAME;
          stringValue = BaseEncoding.base64().encode(createRandomizingEncoder().encodeOrdinal(value));
      } else {
          stringValue = Integer.toString(value);
      }
      set(key, stringValue);
      return this;
  }
```

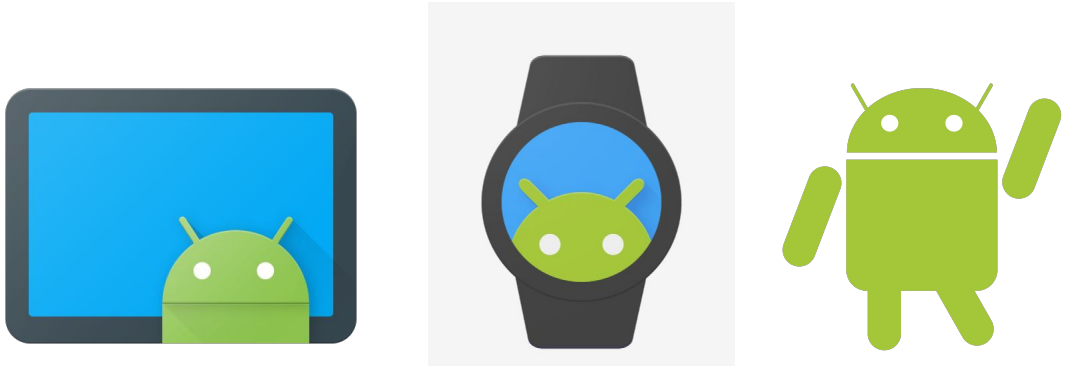
```
public Encoder(byte[] userSecret, String encoderId, int numBits,
               double probabilityF, double probabilityP, double probabilityQ,
               int numCohorts, int numBloomHashes) {
  this(
      null, // random
      null, // md5,
      null, // sha256,
      userSecret,
      encoderId,
      numBits,
      probabilityF,
      probabilityP,
      probabilityQ,
      numCohorts,
      numBloomHashes);
}
```

```
/**
 * Constructs a new RAPPOR message encoder.
 *
 * @param userSecret Stable secret randomly selected for this user.  UserSecret must be at least
 *     MIN_USER_SECRET_BYTES bytes of high-quality entropy.  Changing the user secret clears the
 *     memoized cohort assignments and permanent randomized responses.  Be aware that resetting
 *     these memoizations has significant privacy risks -- consult documentation at go/rappor for
 *     more details.
 * @param encoderId Uniquely identifies this encoder.  Used to differentiate momoized
 *     cohort assignments and permanent randomized responses.
 * @param numBits The number of bits in the RAPPOR-encoded report.
 * @param probabilityF The RAPPOR "f" probability, on the range [0.0, 1.0].  This will be
 *     quantized to the nearest 1/128.
 * @param probabilityP The RAPPOR "p" probability, on the range [0.0, 1.0].
 * @param probabilityQ The RAPPOR "1" probability, on the range [0.0, 1.0].
 * @param numCohorts Number of cohorts into which the user pool is randomly segmented.
 * @param numBloomHashes The number of hash functions used forming the Bloom filter encoding of a
 *     string.
 */
```


Granular Consent Request UI
screen tap
choice
location
battery usage
data usage
Mobile Application
(in need of privacy perserving measurement)
Measure Threshold & Consent Req met?
Apply Privacy Protections (RAPPOR, PII removal)
Device Event Queue
Dispatch Interval Met?
Network Dispatch over Secure Connection
The Internet
Self-Hosted Secure, Onion Measurement Server
(piwik instance)
Additional Privacy Operations (RAPPOR)

Available now -- free and open source

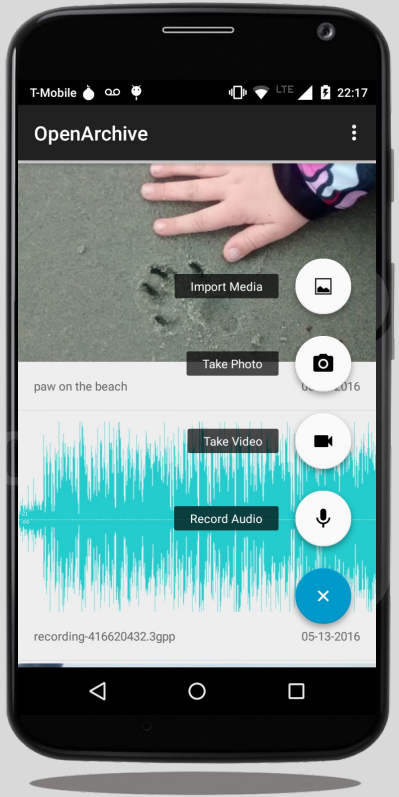
Clean Insights SDK for Android (*Preview Release*!)
https://github.com/cleaninsights



OpenArchive

Preserving Mobilized Culture

Powered by Clean Insights



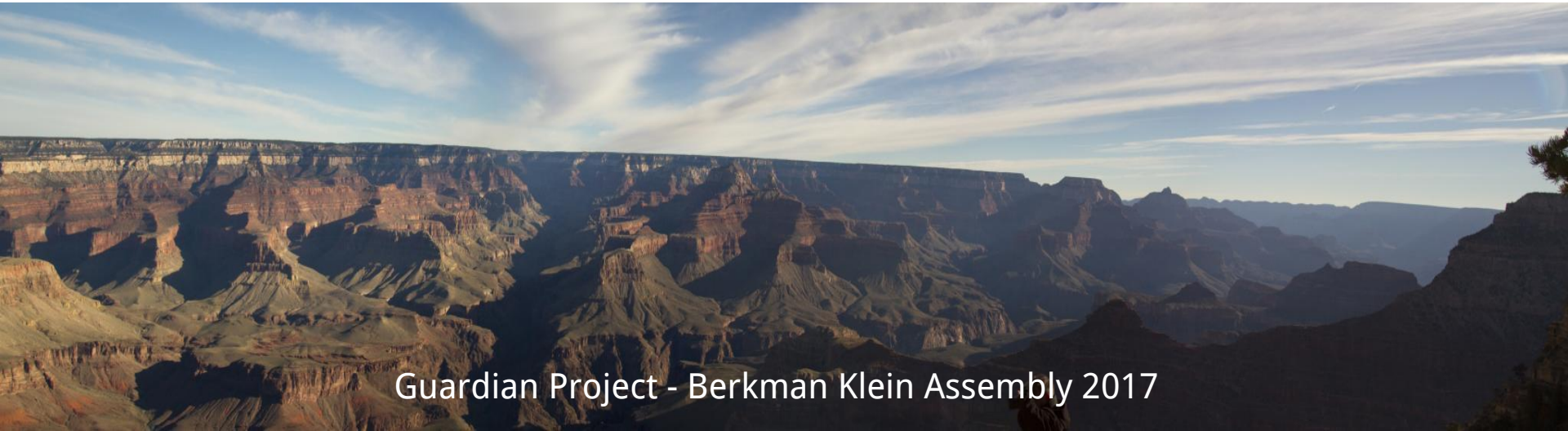
Now measuring media type popularity, offline sharing, network diagnostics, battery usage and more….

GET IT ON Google Play

Clean Insights

cleaninsights.io

@dataistoxic #dataaretoxic



Guardian Project - Berkman Klein Assembly 2017