



Clean Insights

We started with consent

David Oliver
david@guardianproject.info

John Hess
john@jthess.com



tl;dr

Clean Insights is
privacy-preserving measurement
(PPM) that empowers rather than
alienates users

- Match measurement need with support for the privacy & anonymity of those measured
- Consent at the forefront
- Time-bound measurement
- Client-side aggregation
 - measure now; send later
- Minimize need for side-channel measurement
- Tools and Best Practices



Origin Story

<https://www.berkmankleinassembly.org/fellowship-2017projects>

<https://cleaninsights.org>

- 2017
 - Twelve-week hackathon project hosted by the Berkman Klein Center for the Internet and Society (Harvard University), and the Media Lab (MIT)
 - Participants from the United Nations, Square, Apple and Google, Guardian Project
- 2020-2022
 - Internews supports the Clean Insights Symposium
 - Concept clarification, iterative implementation, trials
 - cleaninsights.org is born!



Origin Story

The original questions

- How can funders better understand the impact of ideas they fund without putting users at risk?
- How can companies strike the right balance between preserving user privacy and driving successful product development?
- Is it possible to enable measurement of digital interactions in a safe, secure, and sustainable way?
- Can privacy precepts be upheld, even for small open source projects with tiny user bases?



Why Clean Insights?

Outcomes of the Clean Insights
Symposium (May 2020)

- Developers want to understand the impact or usage patterns but not in a way that endangers or alienates their users
- Developers seek a secure and private platform but also guidance in designing measurement campaigns in a manner that is not harmful to individuals
- No measurement is not an option
- Many invasive options for measurement exist
- Build it ('consentful' measurement) and they will come
 - Lack of an option is what pushes developers to invasive alternatives
- Improvement can be incremental



Focus / Refocus

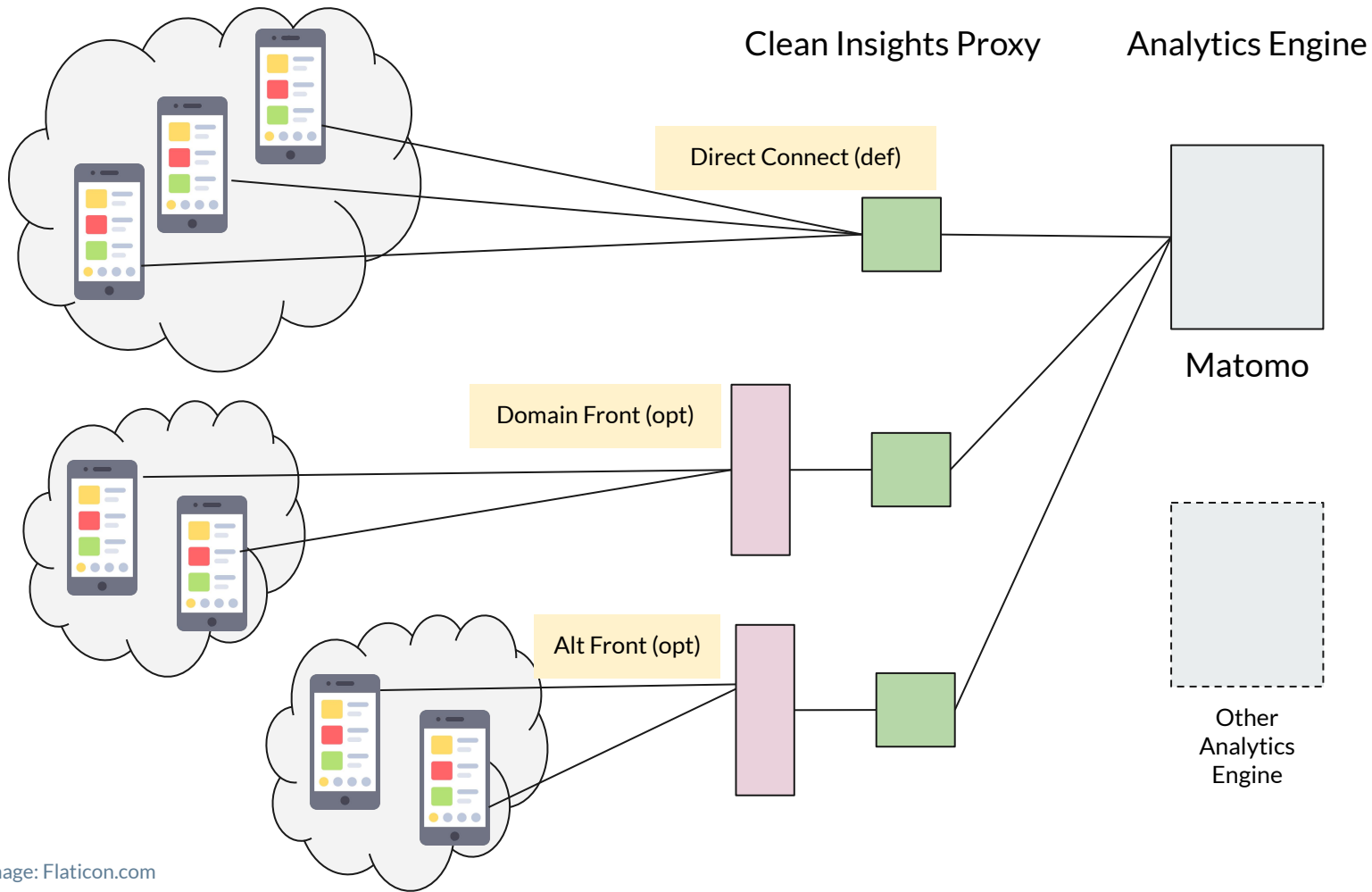
Can measurement be brought in line with respect for the user?

- Focus on asking the right questions and collecting just enough data to answer them
- Aggregate data at the source
- Server discards needlessly-toxic PII (e.g. IP address)
- Make the measurement experience legible to, and engaging for, the user
- Generalize data collected, use deresolution to reduce identifiability



Clean Insights: Software + Experience

- Client SDKs
 - iOS
 - Android
 - Javascript
 - Rust
 - Python
- Anonymizing Proxy
 - OoB support for Matomo analytics application (open source)
 - API compatible with general analytics packages
- Best Practices Guide
 - Grounded in user experience research and implementations





Support the Basics

What people tell us they want:

- To collect “normal” analytics, but make them anonymous
 - Counting users, installs, app actions (from a small, enumerable set) etc.
 - Cross-tabulation by location or other user-level metrics
 - Time spent on activities
- Crash reports
- Surveys

If the Clean Insights solution does not support the basics, people will find other ways to get what they need



Improving Anonymity

- Batch reports to be sent (e.g. every Sunday)
 - Hides timestamps of measured activities
- Generalization and deresolution as appropriate on the client
- Domain fronting
- Tension between respecting consent and padding the anonymity set
 - Particularly for small projects
 - Our point of view is that people understand their own risks, therefore err on the side of consent



Borrow from Law

- Time-bound Contracts
 - Use “Campaigns”
- Consideration
 - Often elided in shrinkwrap, EULA, TOS tradition.
 - Meaningful consent is made more difficult by user fatigue
 - One option: sharing the aggregate data with the subjects
- Avoid “Contracts of Adhesion”
 - Recognition of disparate power between parties
 - Refusing consent shouldn’t needlessly deprive a user



Consent

Consider who is using your app, what for, and in what situations

Consider the data you're collecting

- Start with a specific question
- Collect data for a set time period
- Handle that data carefully
- Get rid of the source data once you've arrived at the insight

Common application design patterns and information needs yield CI's Best Practices guidelines



Consent Principles

Digestible

Digestible: The user understands what they're agreeing or not agreeing to.

Transparent

Transparent: Nothing is hidden.

**Affords
Variance**

Affords Variance: The user has options.

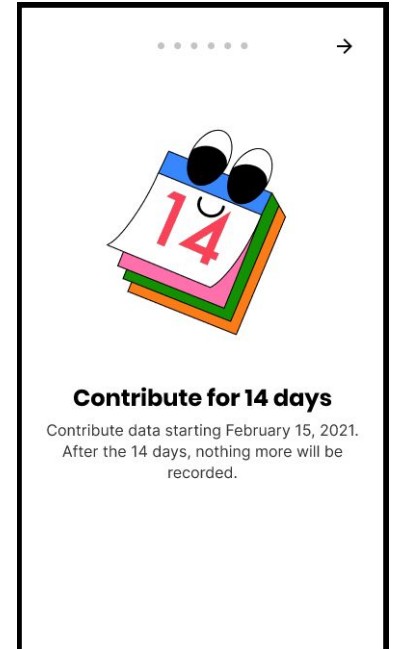
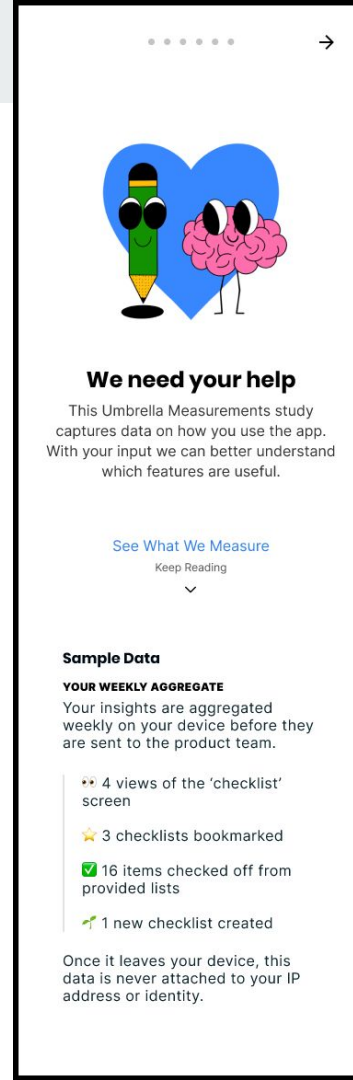
**People
Centric**

People-Centric: The contribution has an impact that benefits the user and their community.

A Sample Consent Model

When you want to measure usage patterns across multiple features

Framed as a collective campaign experience

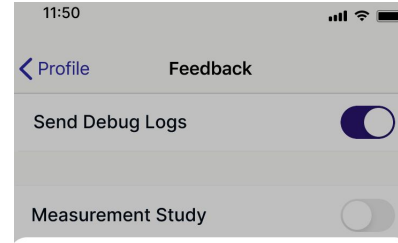


Umbrella



Another Consent Model

When you're using metrics as a focus group study



Want to contribute?

Help us understand how Círculo is used.
The study measures 5 things for 4 weeks.

- ✓ Response time
- ✓ How long a status is active
- ✓ Use of conditions
- ✓ Frequency of location sharing
- ✓ If messages are sent offline

Your contribution is **anonymous**. Your location is never exposed.

Yes

No

CONFIRM

Círculo



Roll Your Own Consent Model

Defining Your Consent Experience

When making decisions about the consent experience:

- Ask for consent in relation to the engaged view/task
- Ask subtly so workflow isn't interrupted
- Give the user the option to decide later
- Consider letting the user choose what to share
- Delay the ask until the user has experience
- Support temporary opt-out



Implementations

Impact reports available!

Mailvelope

- Email privacy

Save by Open Archive


- Secure archiving and sharing

Tella

- Securely document events

WeClock

- Self-tracking (work and wellbeing)



Become a “Conscientious Collector”

Determine a specific decision to be made from data collection

Consider your audience and their situations

Take consent to the next level
(legible and clear terms)

Collect only for a short, specified period of time

Strive to keep a database you would make public (from a privacy POV)



Other PPM Approaches

Commonalities and differences
from other IETF PPM activities

Common Affinities

- Proxying to protect identity
- Start with the question/decision; don't collect it all and figure it out later
- Neutralize the dataset such that leaks are nontoxic

Clean Insights Sacrifices:

- Assumes some trust in the measuring entity (“collector”)
- Can fail on one-time visits
- Counts on implementers to know what measurements would (and would not) be inherently revealing



More Information

<https://cleaninsights.org/>

- Consent guide: <https://okthanks.com/blog/2021/5/14/clean-consent-ux>
- Gitlab repo: <https://gitlab.com/cleaninsights>
- Impact reports: <https://cleaninsights.org/impact>

<https://guardianproject.info/>

<https://matomo.org/>



END