Too Big to Mail: On the Way to Publish Large-scale Mobile Analytics Data

Ella Peltonen, Eemil Lagerspetz, Petteri Nurmi, Sasu Tarkoma
University of Helsinki, Finland

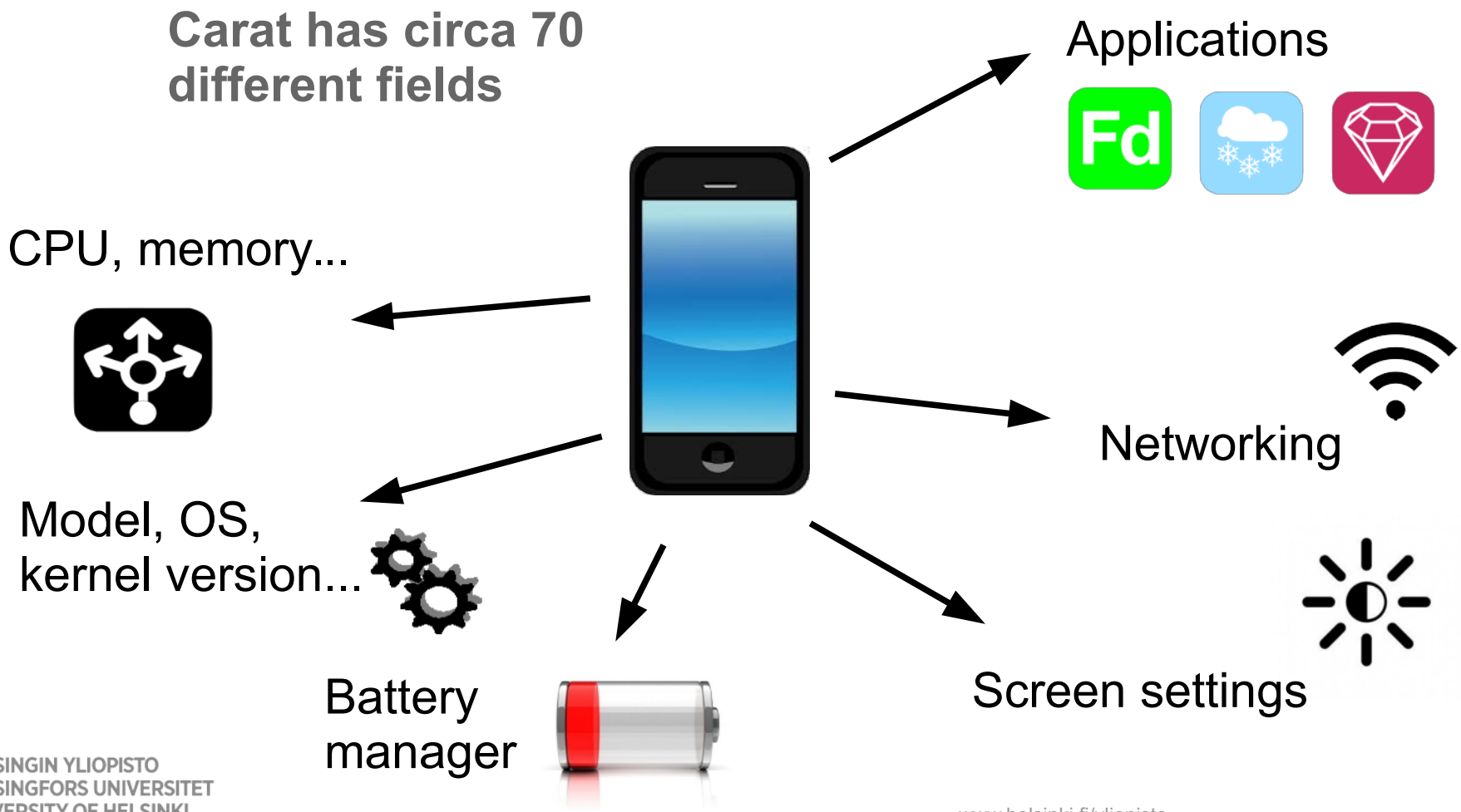
http://carat.cs.helsinki.fi/research
first.last@cs.helsinki.fi

Carat Mobile Data Set

- Since 2012 we have collected mobile data from over 850,000 unique devices:
  - Android and iOS, available in app stores
  - Worldwide, circa 200 countries
  - 250 million measurements (and more is coming)
  - 1.5 TB in binary objects
  - Runnable by Apache Spark in e.g. 10 x ~60GB RAM, 8 cores Amazon EC2 VMs
  - Coarse-grained measurements: based on 1% battery change

Mobile Device as a Data Source

Carat has circa 70 different fields



Multipurpose Data Set

- *How is my app's battery consumption?*
- *Where to find reference data?*

- System optimization
- Long-term mobile usage
- Combinations of energy measurements, system settings, applications, and different features

Challenges of Sharing

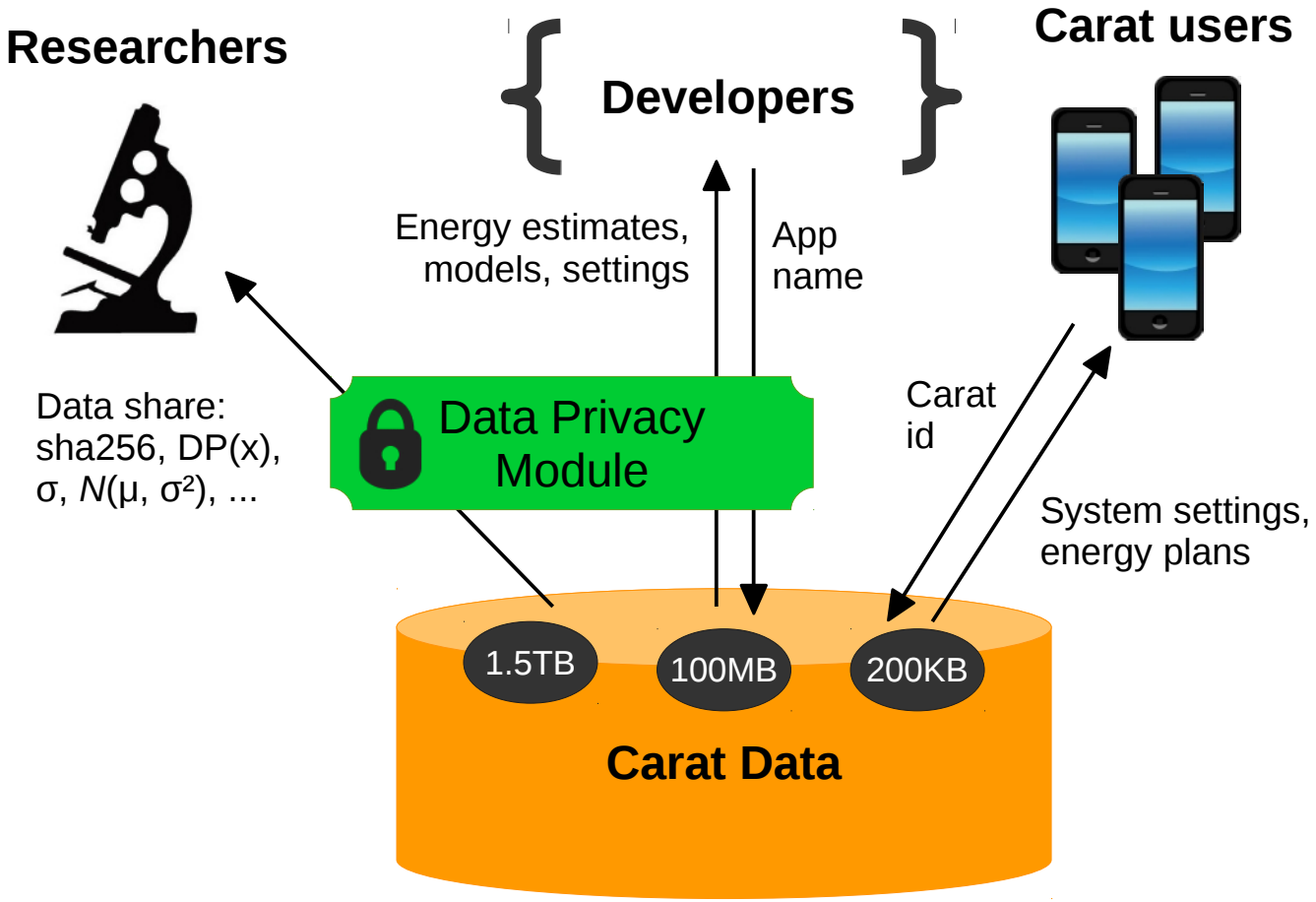
- It's Big – almost 2TB in binary, stored at Amazon S3
  - Requirements of Amazon keys and processing facilities
- It's Private – user's personal data
  - Data from real persons
  - Apps, battery problems, WiFi signals, IDs etc can identify users
  - Privacy implications change all the time
- No commercial use, only for research purposes

Solving “It's Big”

- Sharing in place via remote access
- Requesting only necessary fields / time periods
- API for developers
  - Get access to your own data by using the Android signing key
  - Queries for different feature combinations

Solving “It's Private”

- Differential privacy
  - Statistical method, no user identifiers
  - Must be tailored by use case
- Salted hash
  - Provides privacy for apps and users
  - Can make data unreadable or useless
- Categorization
  - No app/phone/user names, but features


Researchers
Developers
Carat users
Energy estimates,
models, settings
App
name
Data share:
sha256, DP(x),
σ, N(μ, σ²), ...
Data Privacy
Module
Carat
id
System settings,
energy plans
1.5TB
100MB
200KB
Carat Data

Previous Research

- Carat: Collaborative Energy Diagnosis for Mobile Devices, SenSys 2013
- How Carat Affects User Behavior: Implications for Mobile Battery Awareness Applications, CHI 2014
- The Company You Keep: Mobile Malware Infection Rates and Inexpensive Risk Indicators, WWW 2014
- Energy Modeling of System Settings: A Crowdsourced Approach, Percom 2015
- Constella: Recommending System Settings the Crowdsourced Way, Pervasive and Mobile Computing Feb 2016

Takeaways

- There are solutions for scalable data processing and private information retrieval
- It's difficult to combine the solutions addressing all use cases
- We are developing mechanisms for publishing the data and releasing the Carat SDK

- We are open for collaboration
  - Please contact us for details of the data
- We would like to hear your ideas of use cases

We are here this week

Ella Peltonen

ella.peltonen@cs.helsinki.fi

Eemil Lagerspetz

eemil.lagerspetz@cs.helsinki.fi

University of Helsinki

Finland

http://carat.cs.helsinki.fi