Privacy-Preserving Sensing & Analytics on the Edge

Hamed Haddadi

@realhamed

Imperial College
London

The Data Ecosystem

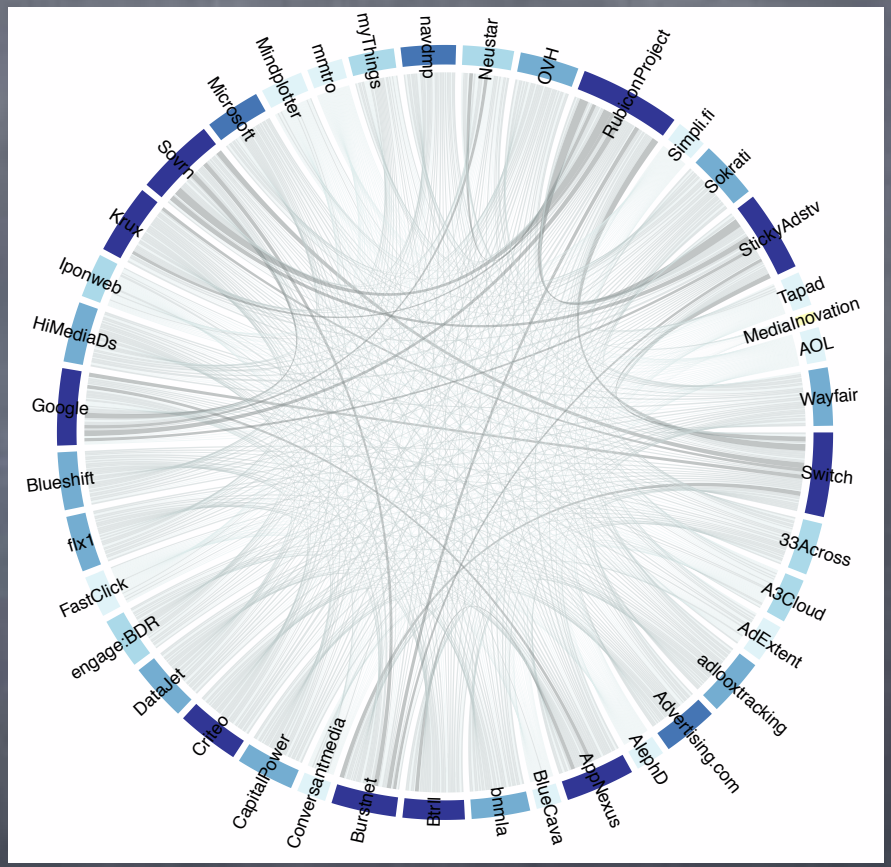
Data about us:

Data generated by us:

Data around us:

Data About Us

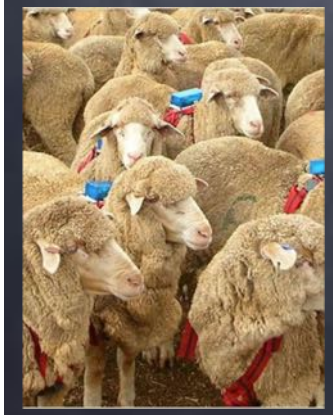
- We found thousands of trackers across the world who follow our clicks and trade our data.
- Our digital footprint
- we are not even aware of.
- Provenance is a major issue.

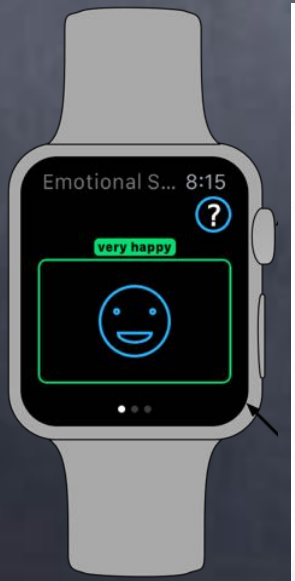


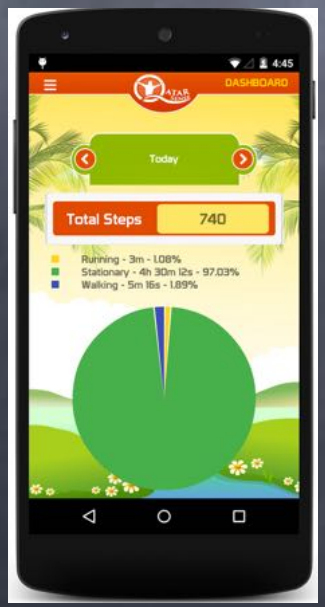
- TMA 2014, PAM 2016, and "Anatomy of the Third-Party Web Tracking Ecosystem" on MIT TR 2014.
  - Ad Blocking is not the long-term solution, see: "Ad-Blocking and Counter Blocking: A Slice of the Arms Race", USENIX 2016.

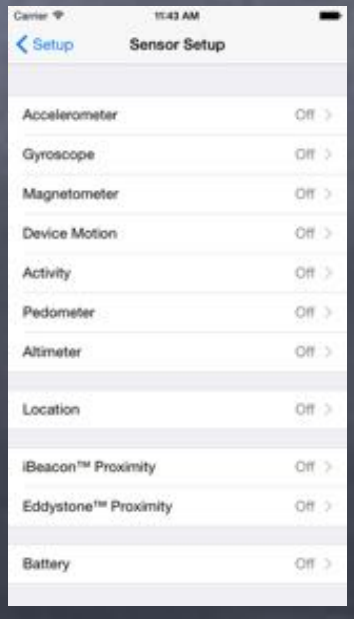
Data Generated by Us

- Online Social media (Tweets, Instagram images, FB posts..)
- Wearable devices
  - Signals indicative of physical & mental health (Current Biology, CHI'2018, UbiComp 2016 MentalHealth)







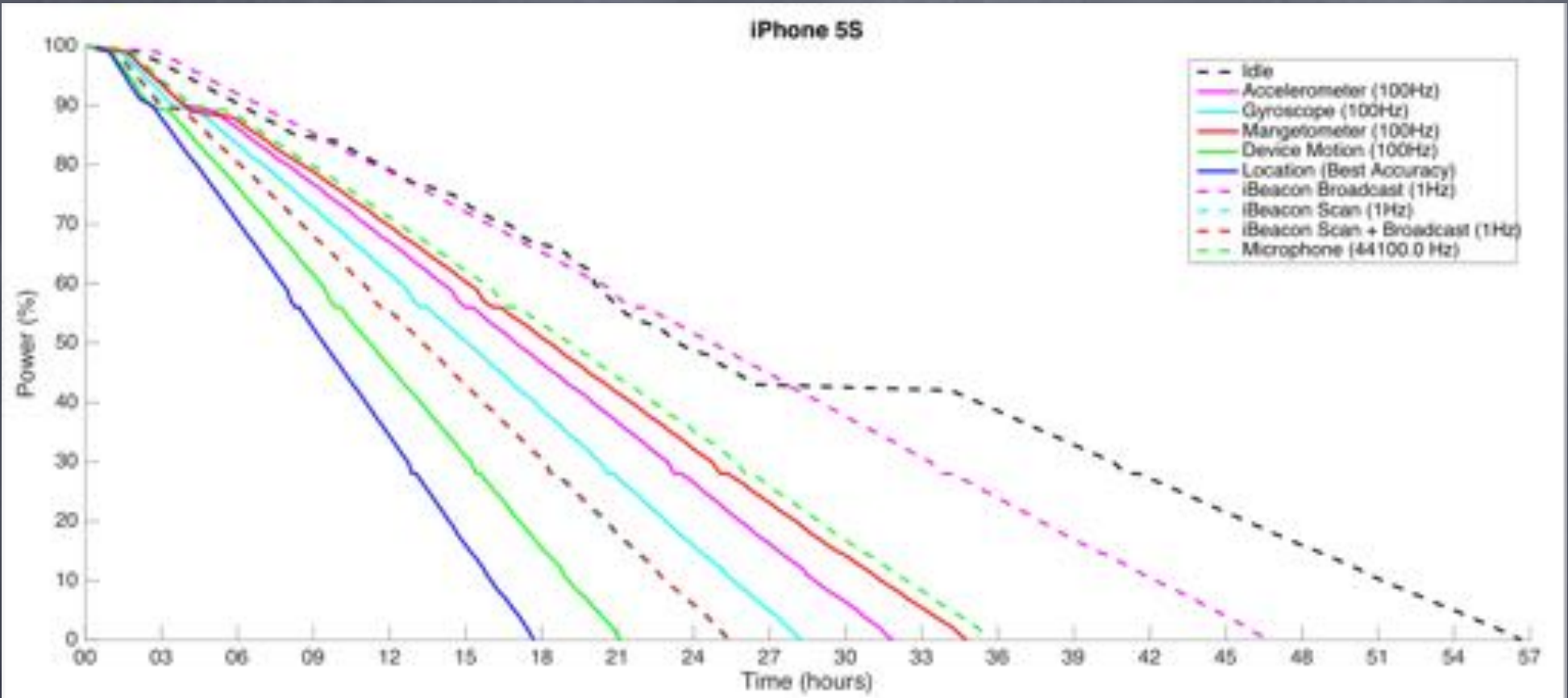


SensingKit

CrowdSense
for iOS / Android

Sensingkit.org
(ACM MobiSys 2017 demos)

Sensing costs energy too!

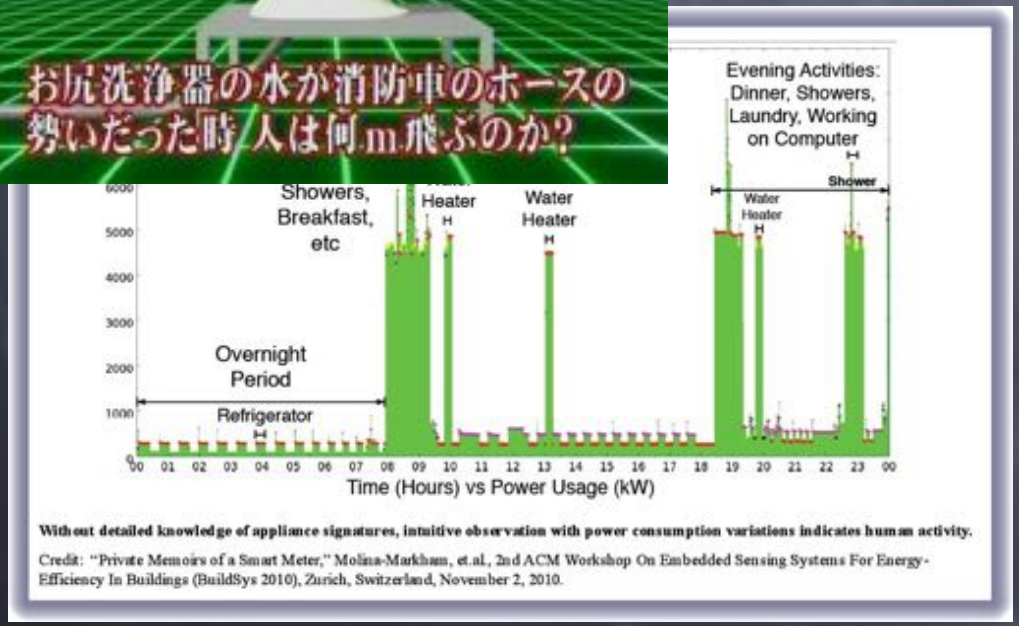


Power consumption of SensingKit running on an iPhone 5S smartphone.

ACM MobiSys 2017 Demo, IE 2016, ACM SIGCOMM IMC 2013

Data around us

- IoT devices
- Cyber Physical Systems



www.connectedseeds.org/about/sensors

An Underlying Structural Problem

The Internet is fragmented, distributed systems are difficult

- Centralising simplifies things
- With the cloud, we can, so we do!

Ease of cloud computing has led to two suboptimal defaults:

1. Move the data … (by copying)
2. … to a centralised location

Applications and Challenges

Opportunities

- Infrastructure monitoring
- Understanding individuals' wellbeing & public health
- Enabling personalised services

Challenges

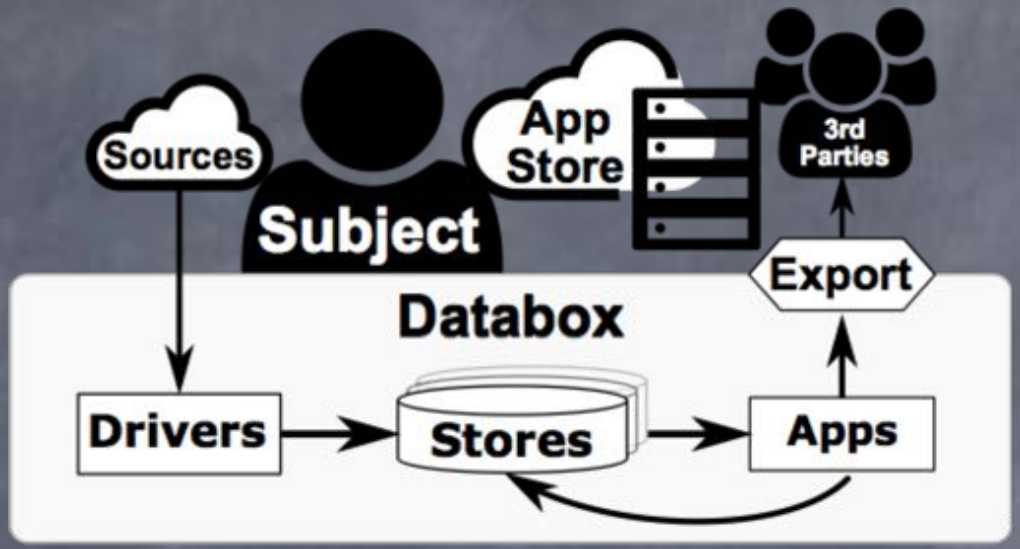
- Real-time control & adaptation, scalability
- Accountability & liability
- Algorithmic bias, privacy, security,...

Can we do detailed, user-centric, contextual analytics without some of the inefficiencies, privacy disasters, and legal challenges?

Efforts in this space

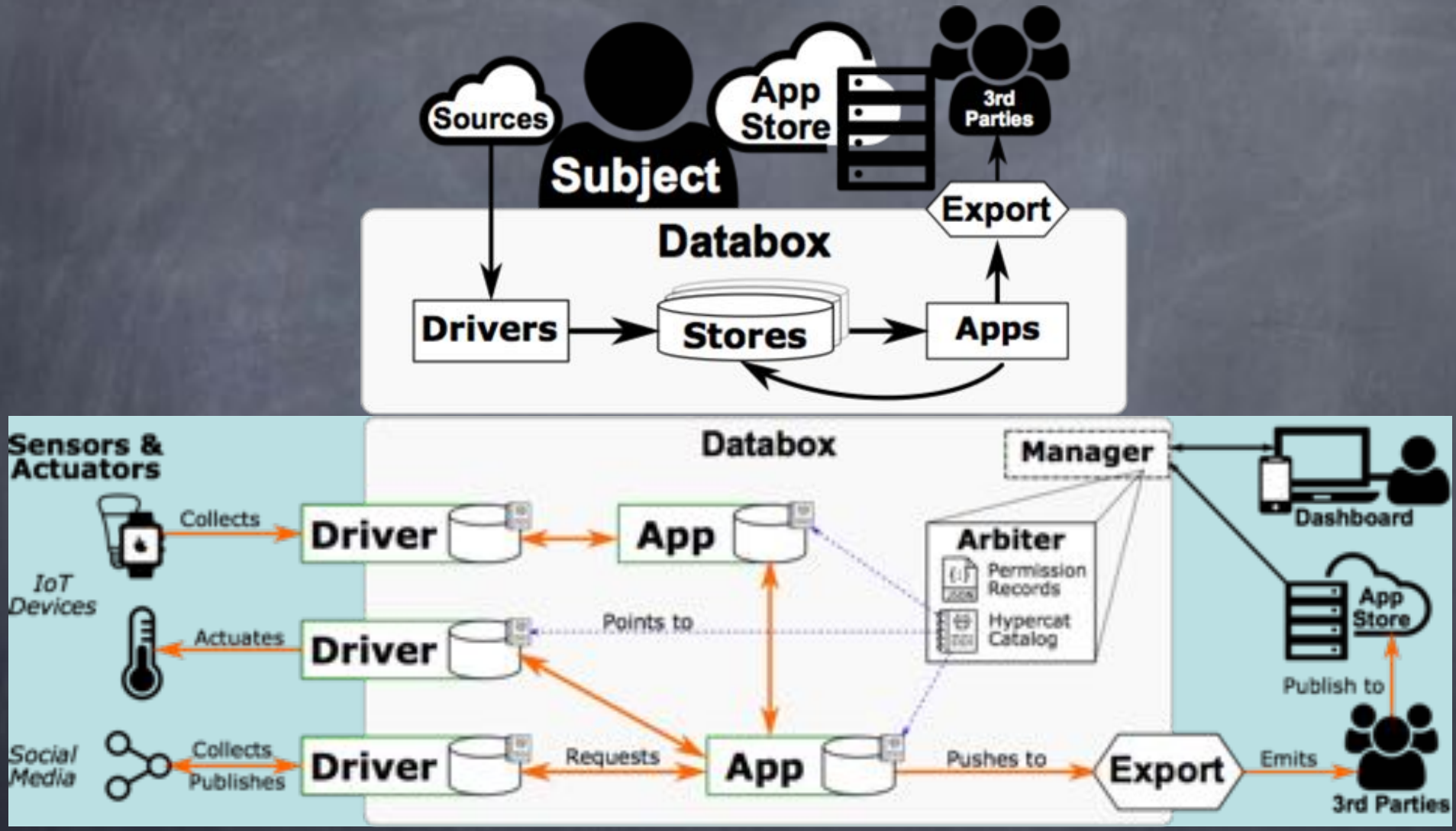
- NyMote
- OpenPDS
- dowse.eu
- Hub of All Things
- Name your favorite data silo...
- ..... Databox

Databox



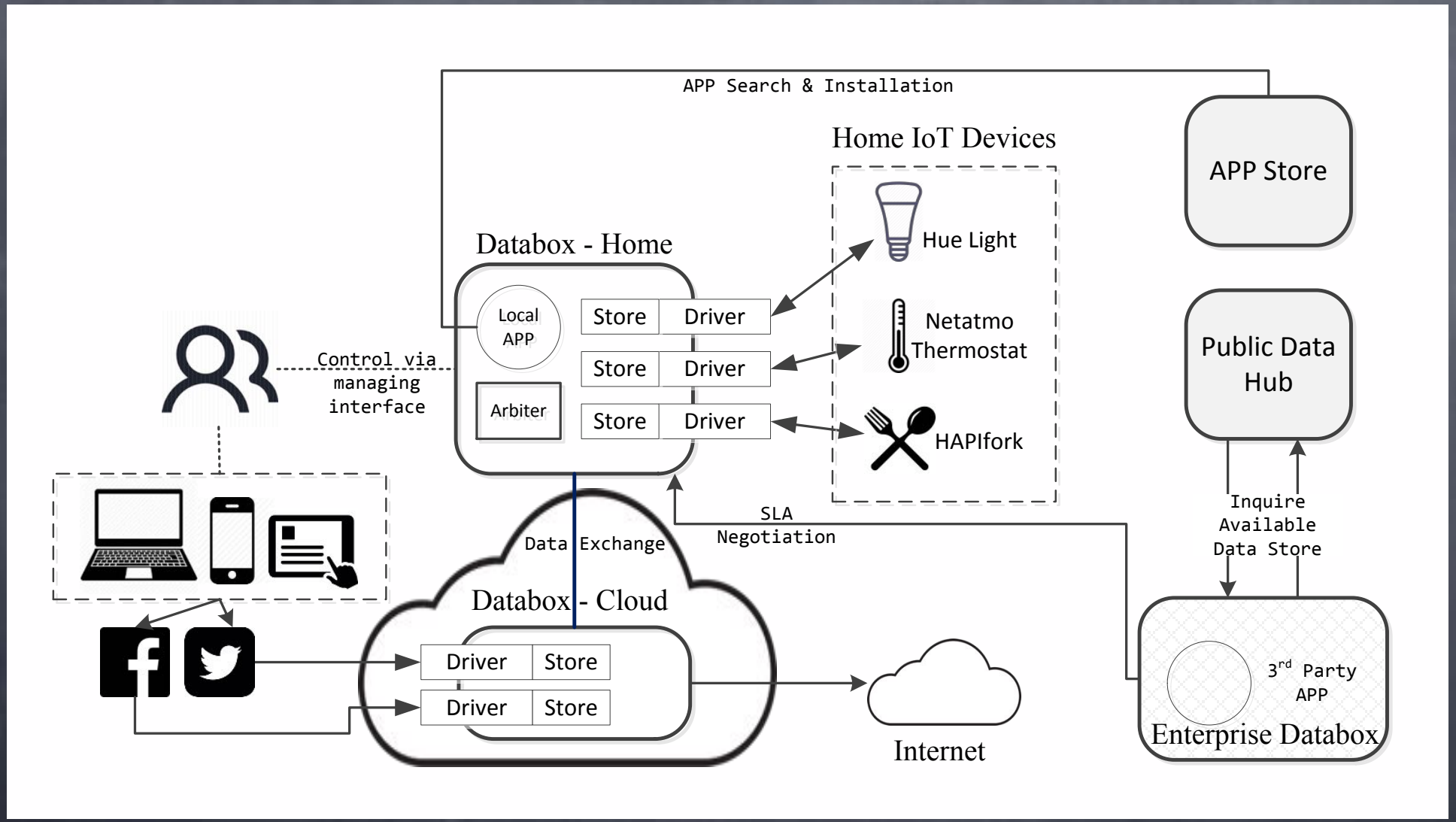
- Mediates access to data, stored locally as appropriate
- Computations (*apps*) move to data, not data to compute
- Maintain control over internal comms and export
- All operations logged for users to inspect, control

Databox Platform



EPSRC Databox: Privacy-Aware Infrastructure for Managing Personal Data
www.databoxproject.uk

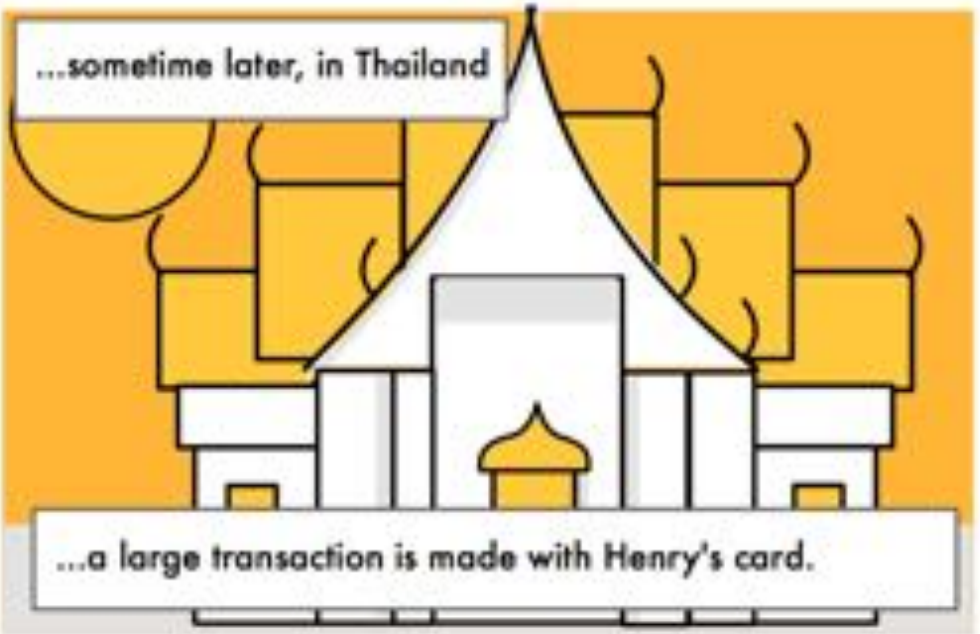
Databox and apps ecosystem



Code available on https://github.com/me-box/

Henry downloads his bank's app onto his databox.

...sometime later, in Thailand



...a large transaction is made with Henry's card.

DATABOX FRAUD DETECTION

Henry's banking app checks his location.

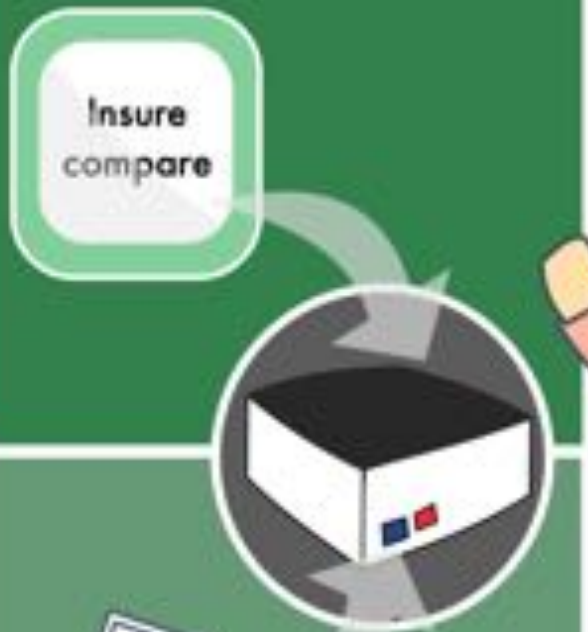
and tells the Bank Henry is NOT in Thailand.

The transaction is refused.

Henry is happy. So is his bank manager.

Elsie's health insurance is due to expire.

Elsie installs an insurance comparison app on her databox.



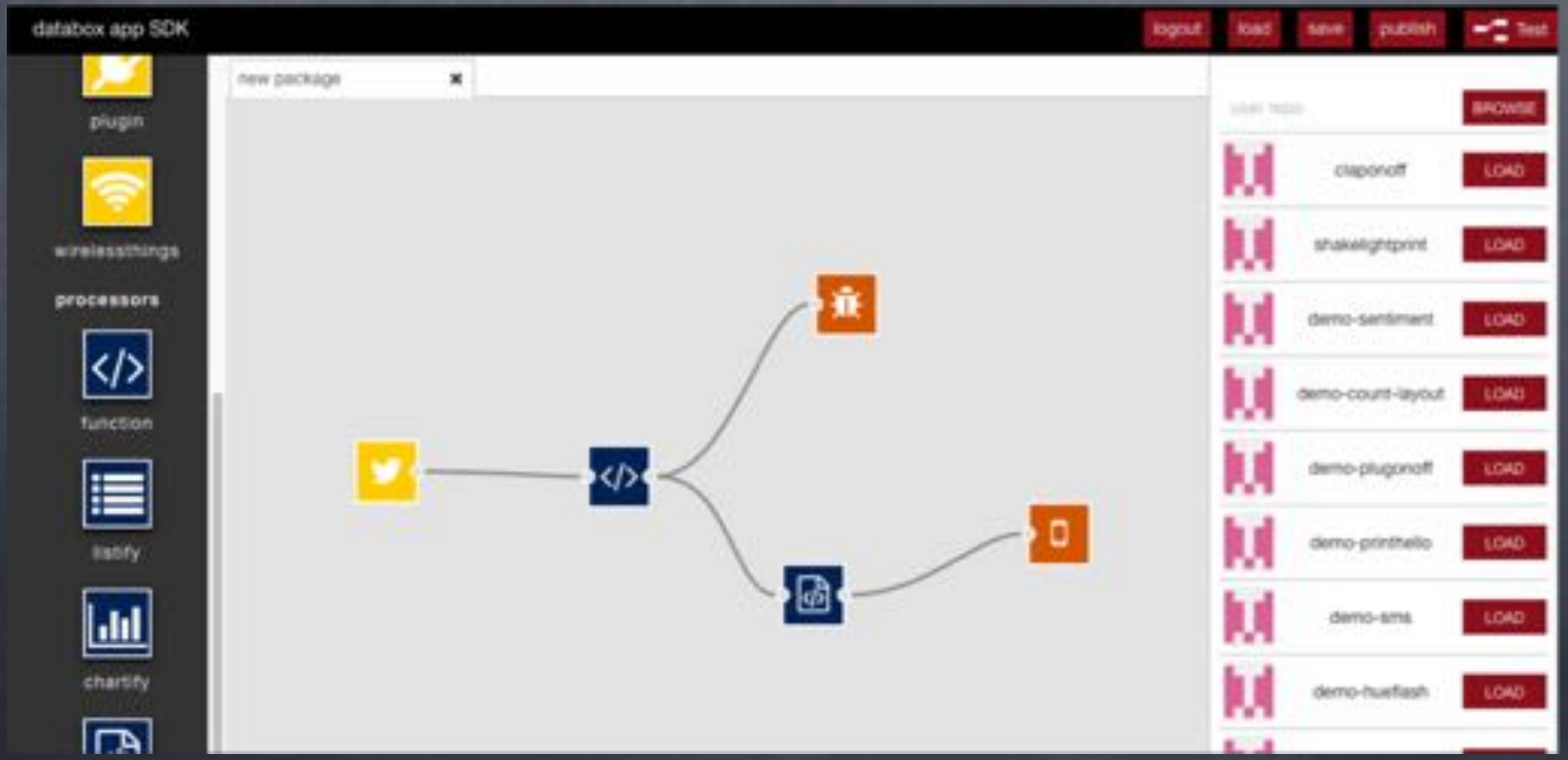
The app analyses her home, mobile, location and grocery shopping data.

The app discovers that Elsie has an active and healthy lifestyle and offers her a big discount.

DATABOX HEALTH INSURANCE

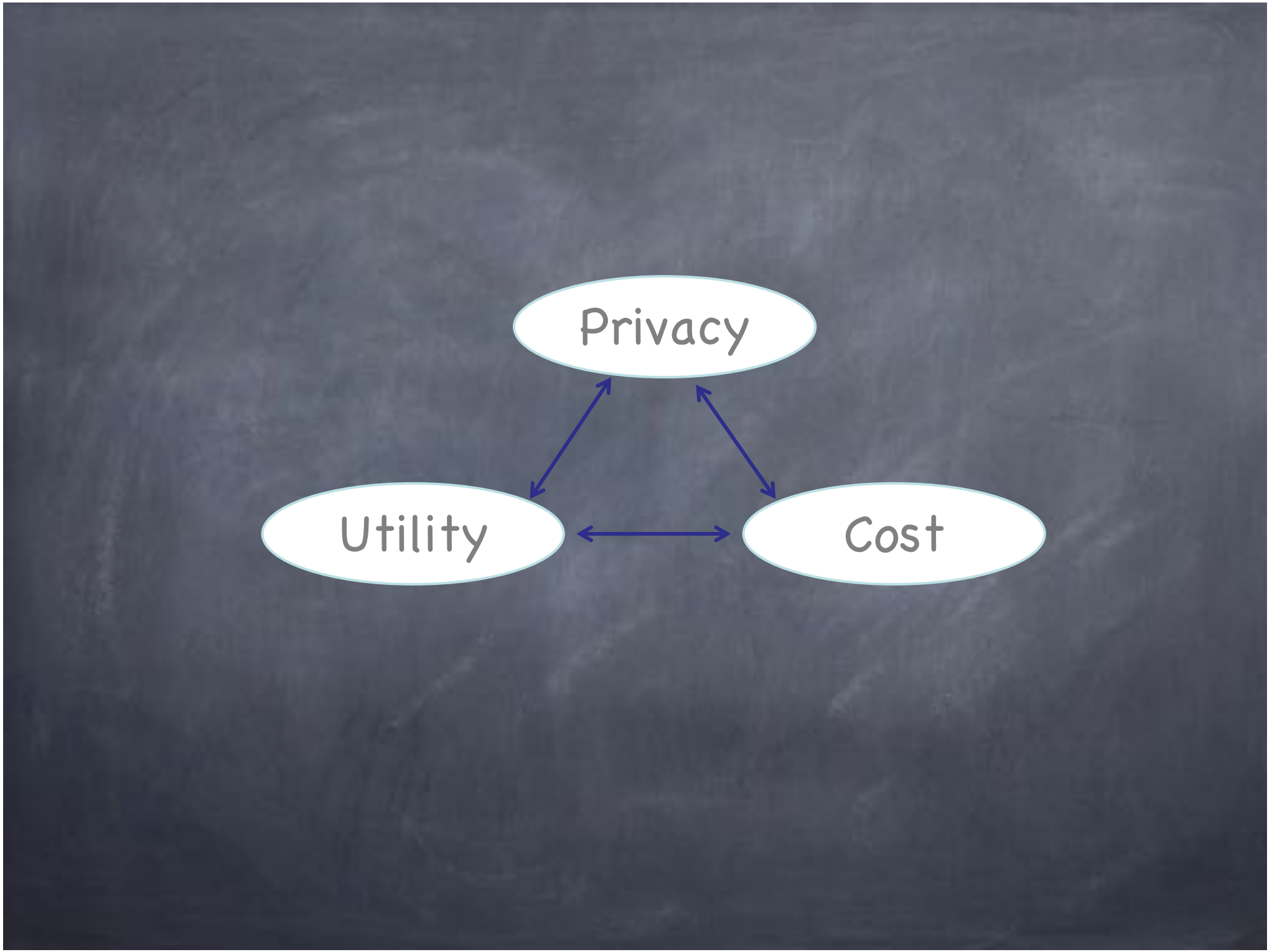
Developing Apps

- Install and connect existing apps
  https://sdk.iotdatabox.com
- Plug together apps & components to customise **your** apps



Open Source Community Engagement

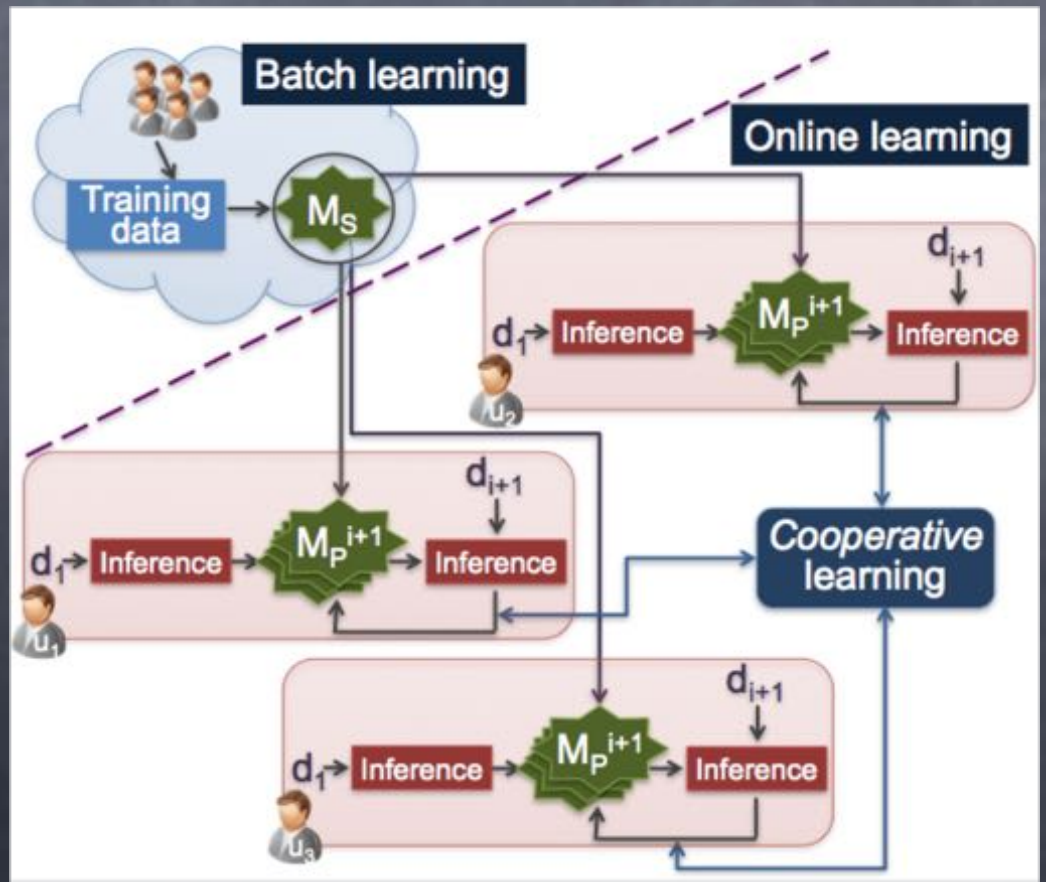
https://forum.databoxproject.uk/
https://github.com/me-box/


Privacy
Utility
Cost

Distributed Analytics

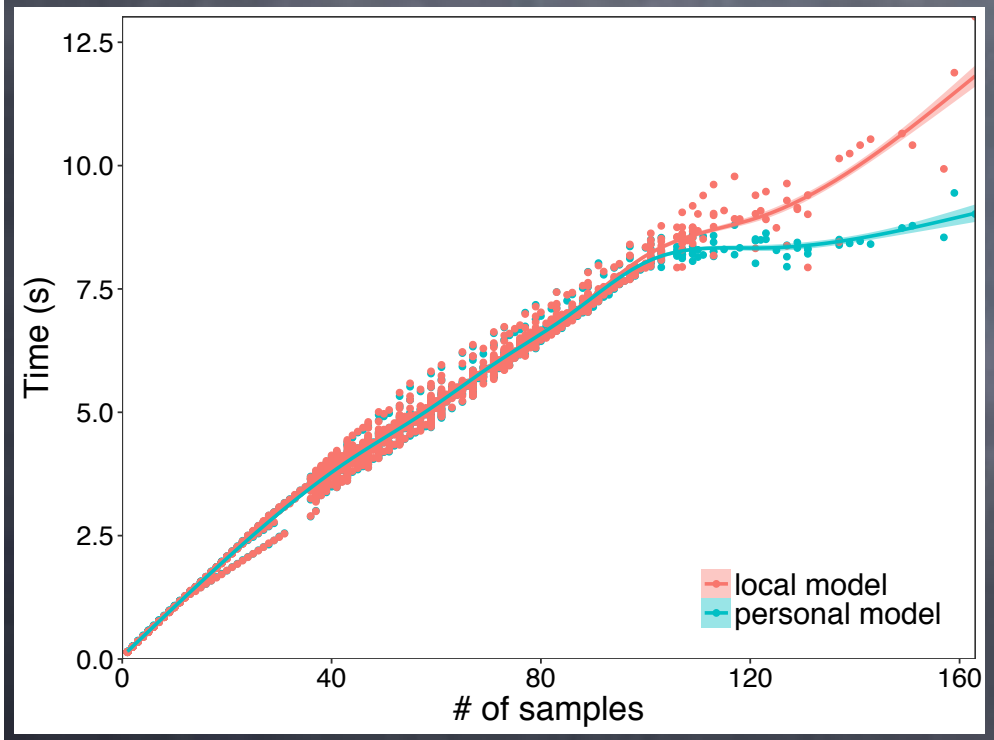
How to handle scale, heterogeneity, dynamics?

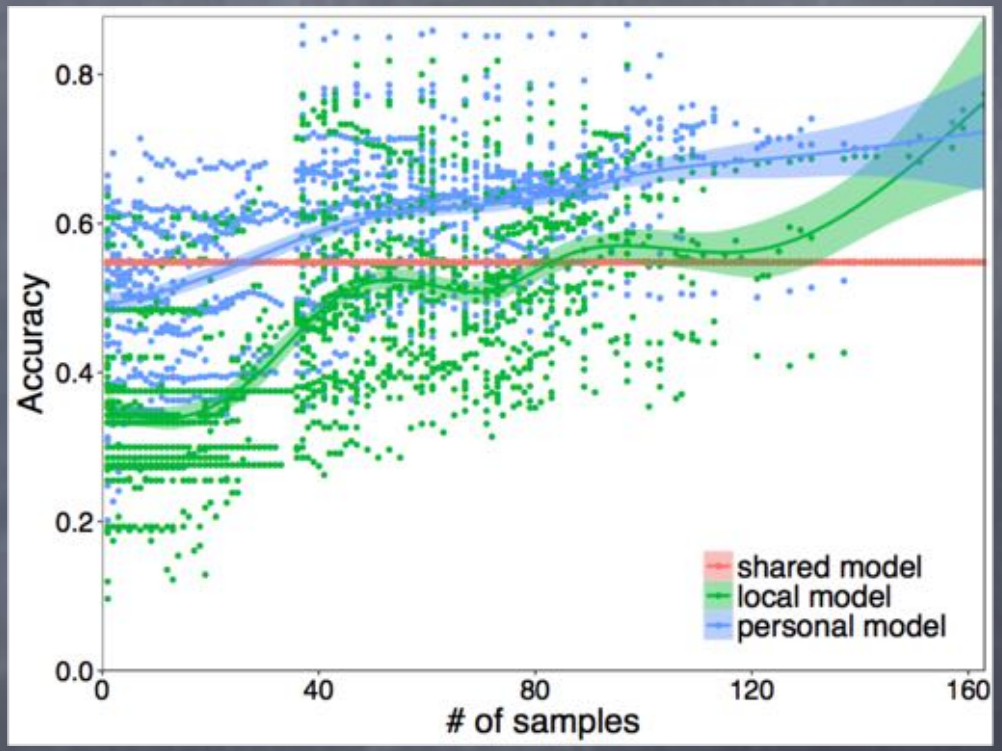
- Cohort vs individual processing
- Distributed model building
- Personalised local analytics



Online Learning

Can we use personal data to improve public, pre-trained ML models?



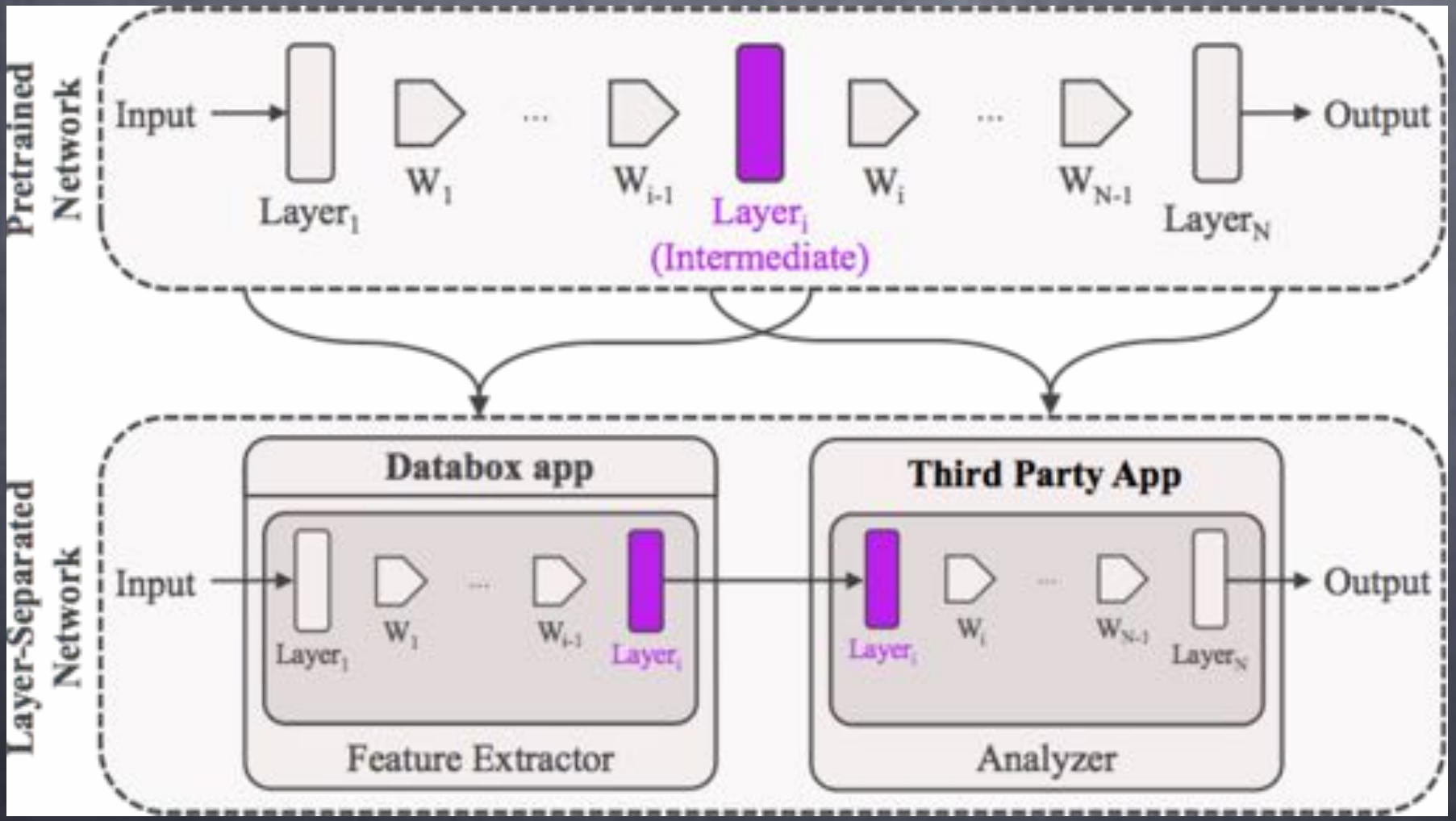


ACM/IEEE IoTDI 2018.

Edge Processing on Sensitive Data

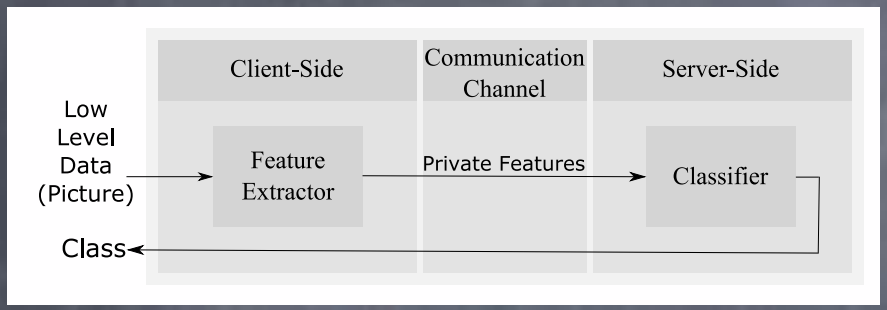
Example: Occupancy-as-a-Service

Privacy-Preserving Analytics

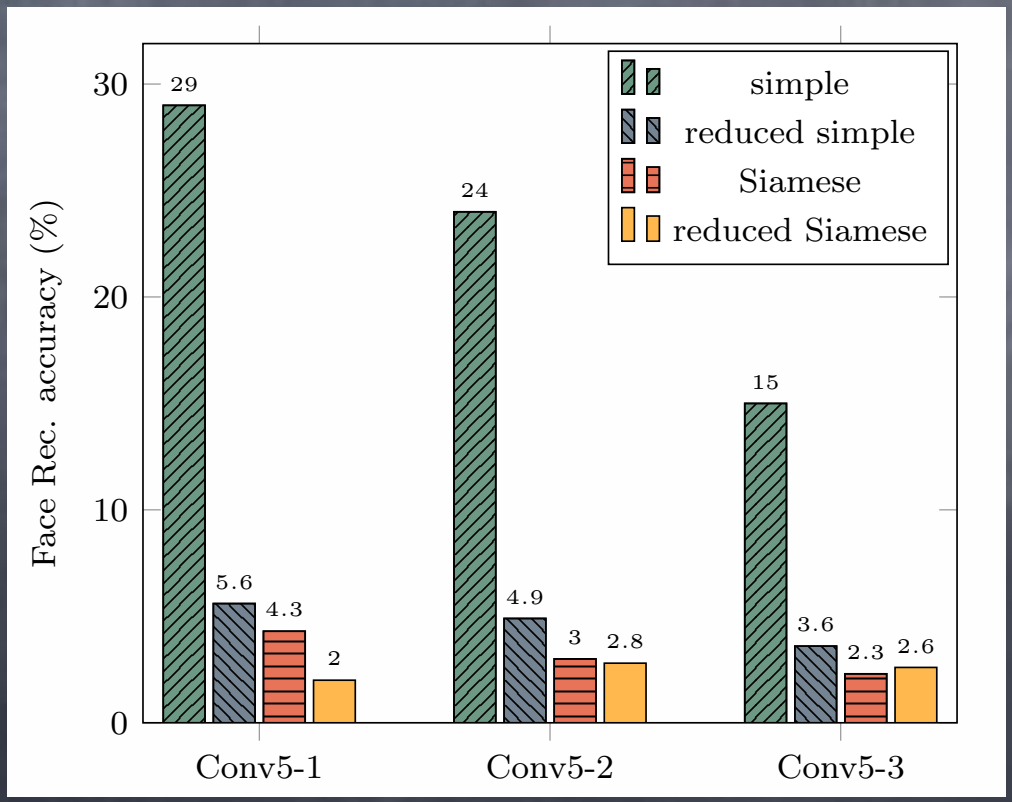


Edge computing paradigm

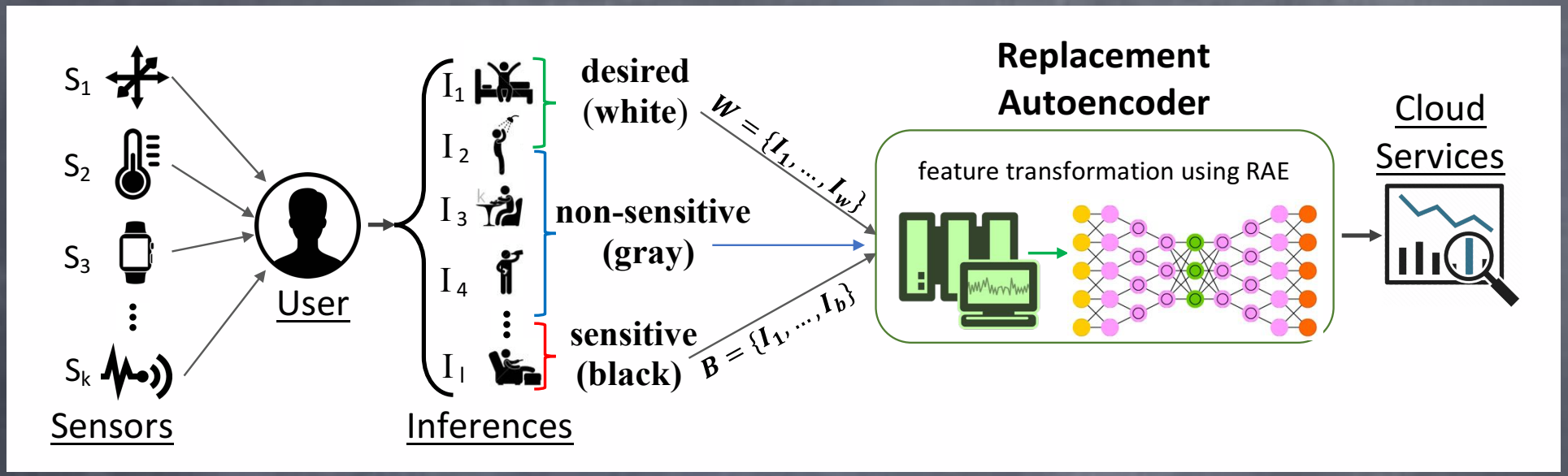
Case study: can we do gender detection without face recognition?

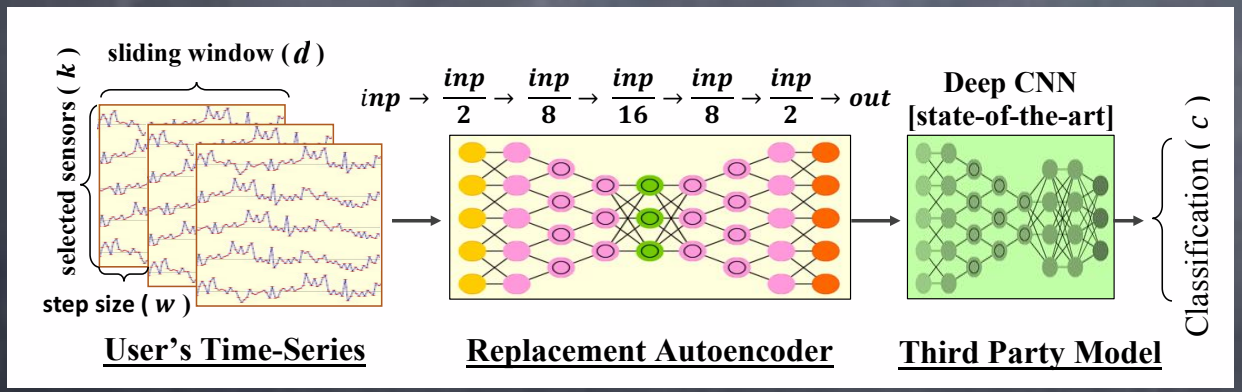


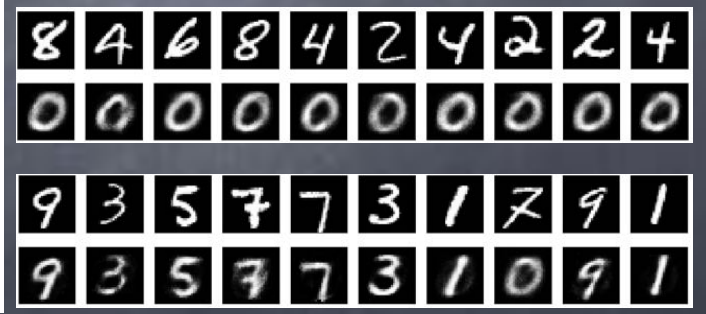
Accuracy on LFW			
	Conv5-1	Conv5-2	Conv5-3
simple	94%	94%	94%
reduced simple	89.7%	87%	94%
Siamese	92.7%	92.7%	93.5%
reduced Siamese	91.3%	92.9%	93.3%



Replacement auto-encoder

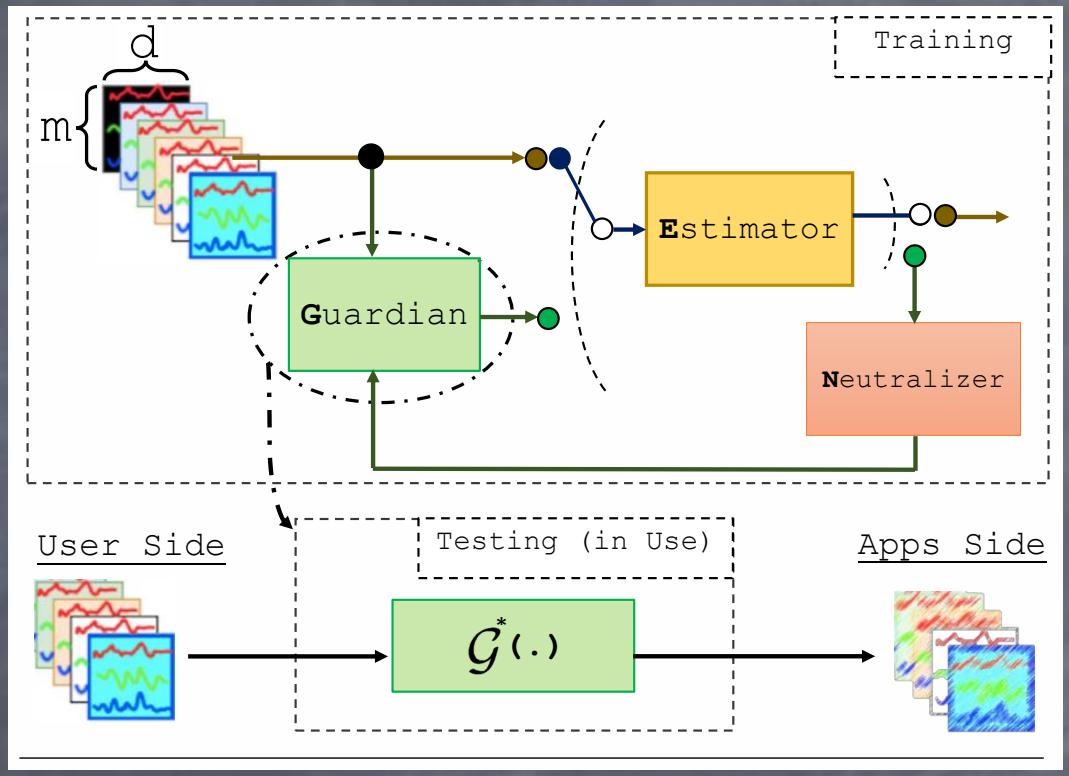




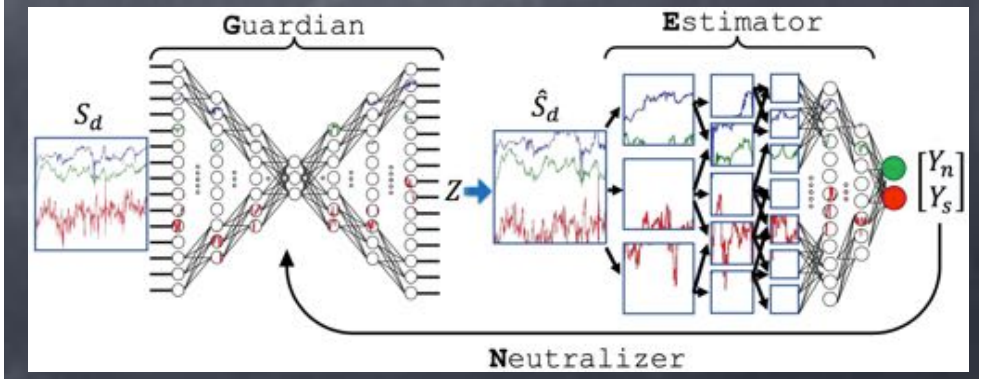


Mohammad Malekzadeh, Richard G. Clegg, Hamed Haddadi, "Replacement AutoEncoder: A Privacy-Preserving Algorithm for Sensory Data Analysis", The 3rd ACM/IEEE International Conference on Internet-of-Things Design and Implementation, April 2018, Orlando, Florida.

GEN


Training
d
m
Estimator
Guardian
Neutralizer
User Side
Testing (in Use)
Apps Side
$\mathcal{G}^*(.)$


Original Data
Transformed Data
Normalized Error (%)
100
80
60
40
20
0
Gender
Age
Height
Weight


Guardian
Estimator
S_d
$\hat{S}_d$
Z
$[Y_n \; Y_s]$
Neutralizer

Inter-variable dependencies

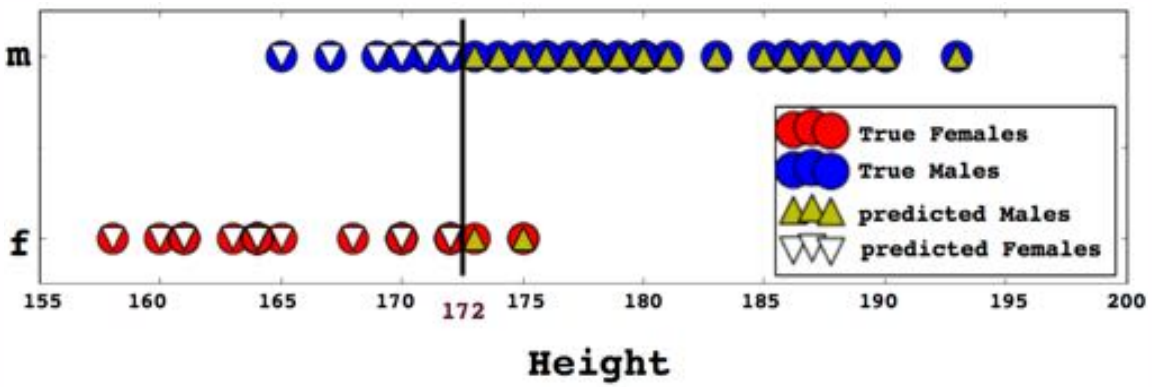
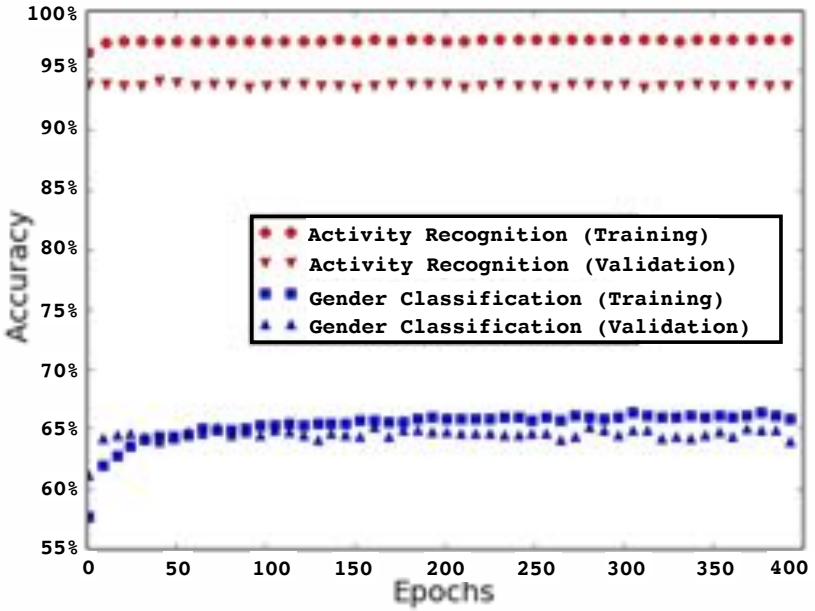


Figure 4. Dependencies between height and gender on the MotionSense and MobiAct datasets. A classification threshold of $172cm$ predicts gender with 84% accuracy.



Mohammad Malekzadeh, Richard G. Clegg, Andrea Cavallaro, Hamed Haddadi, "Protecting Sensory Data against Sensitive Inferences", in proceedings of the 1st Workshop on Privacy by Design in Distributed Systems, co-located with Eurosys 2018, Porto, Portugal, April 2018.

IoT Traffic

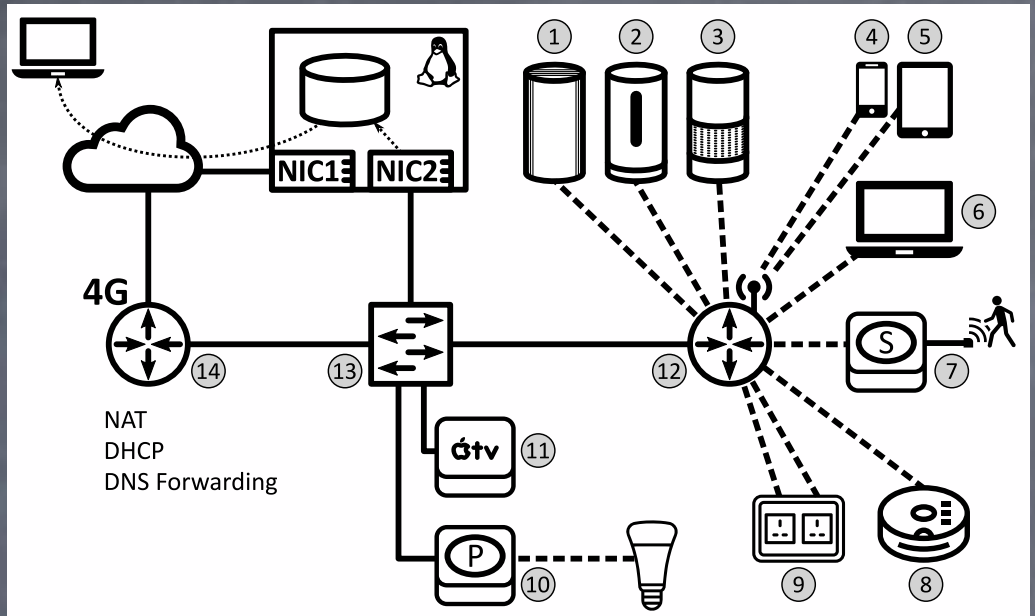
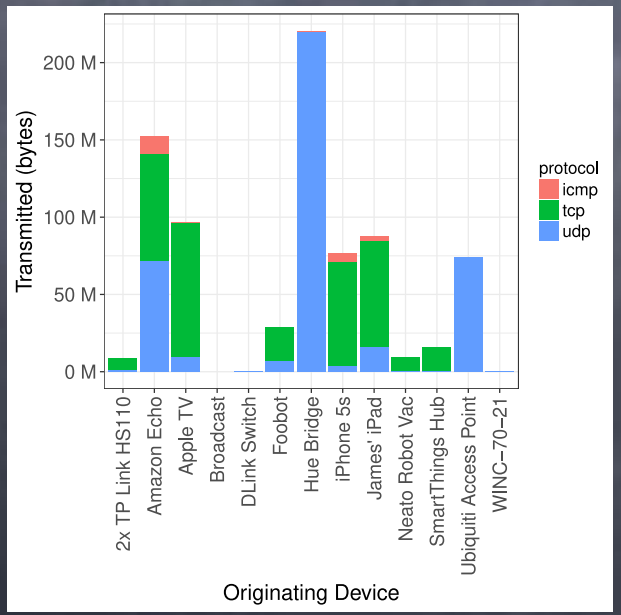
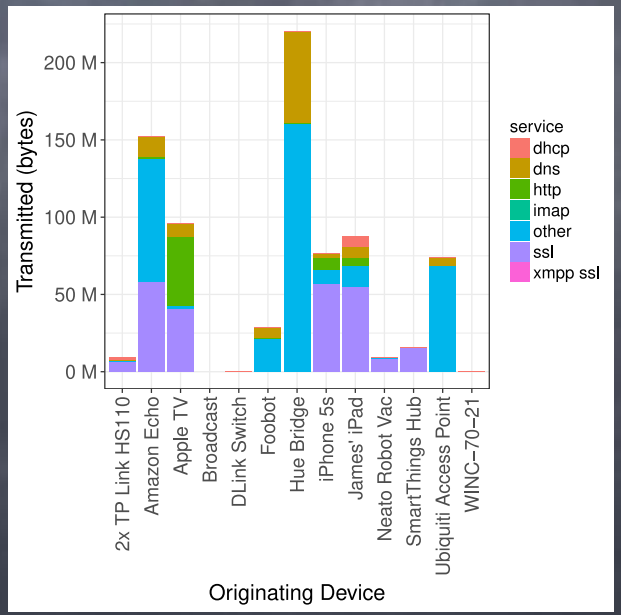


Fig. 1. An Overview of the Home IoT Testbed





Amar et al., , “An Analysis of Home IoT Network Traffic and Behaviour”, Available on ArXiv, March 2018.

Conclusions

- Personal Data analytics face complex challenges and we need new approaches for data utilisation.
- Databox, edge-computing, and user-centric processing methods are timely enablers in this direction
- Interesting new approaches for personal data, ambient sensing, actuation, and HDI

For more information, software, and papers:

haddadi.github.io