Personal Data Ecosystem: Privacy, Utility, and Efficiency Challenges

Hamed Haddadi

Oxford Internet Institute

June 2018

Imperial College
London

The Data Ecosystem

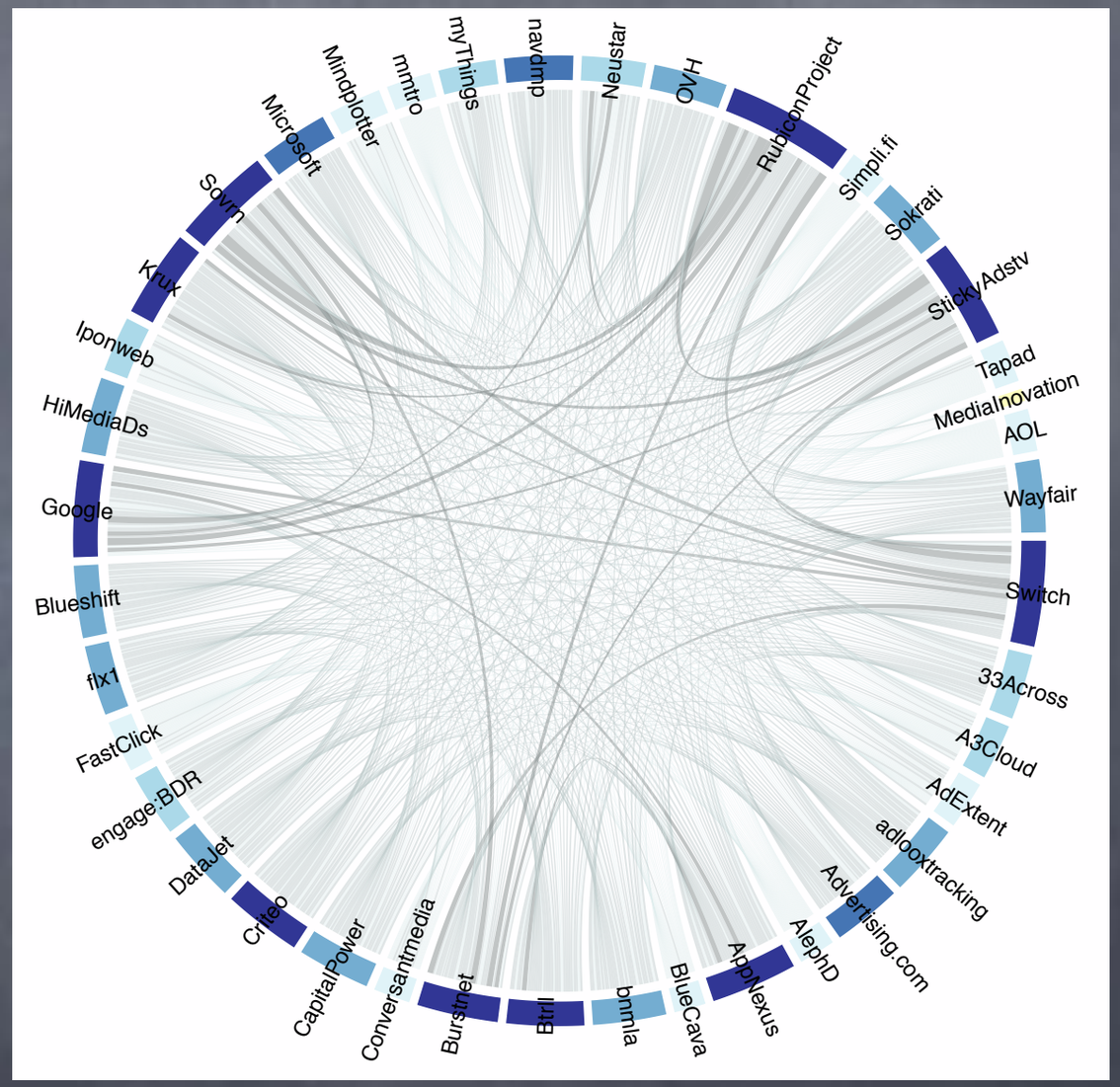
Data about us:

Data generated by us:

Data around us:

Data About Us

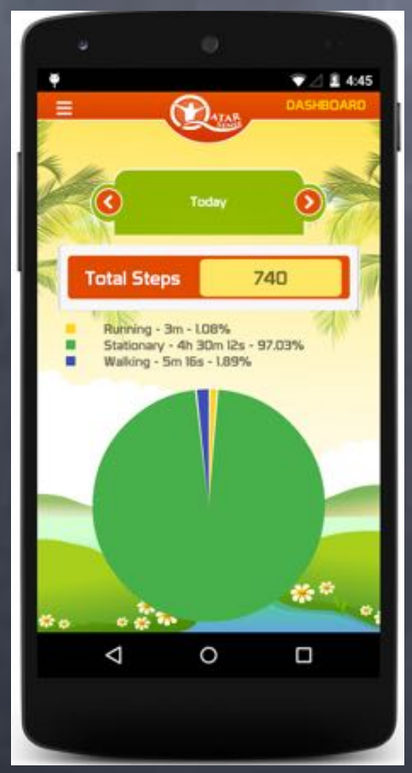
- We found thousands of trackers across the world who follow our clicks and trade our data.
- Our digital footprint
- we are not even aware of.
- Provenance is a major issue.

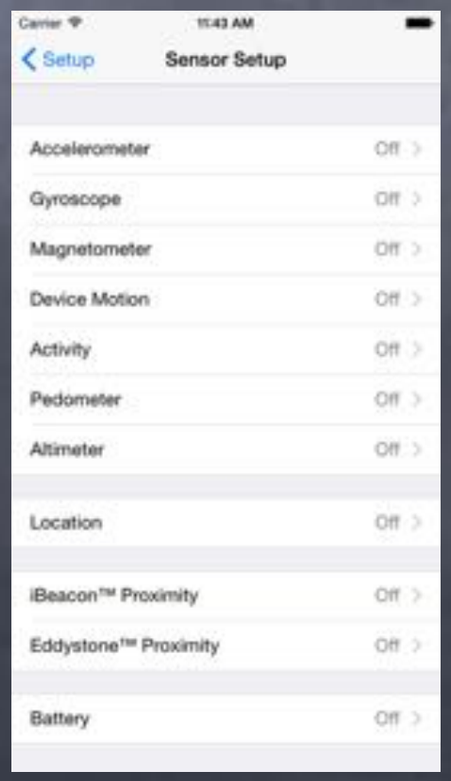


- TMA 2014, PAM 2016, and "Anatomy of the Third-Party Web Tracking Ecosystem" on MIT TR 2014.
  - Ad Blocking is not the long-term solution, see: "Ad-Blocking and Counter Blocking: A Slice of the Arms Race", USENIX 2016.

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)



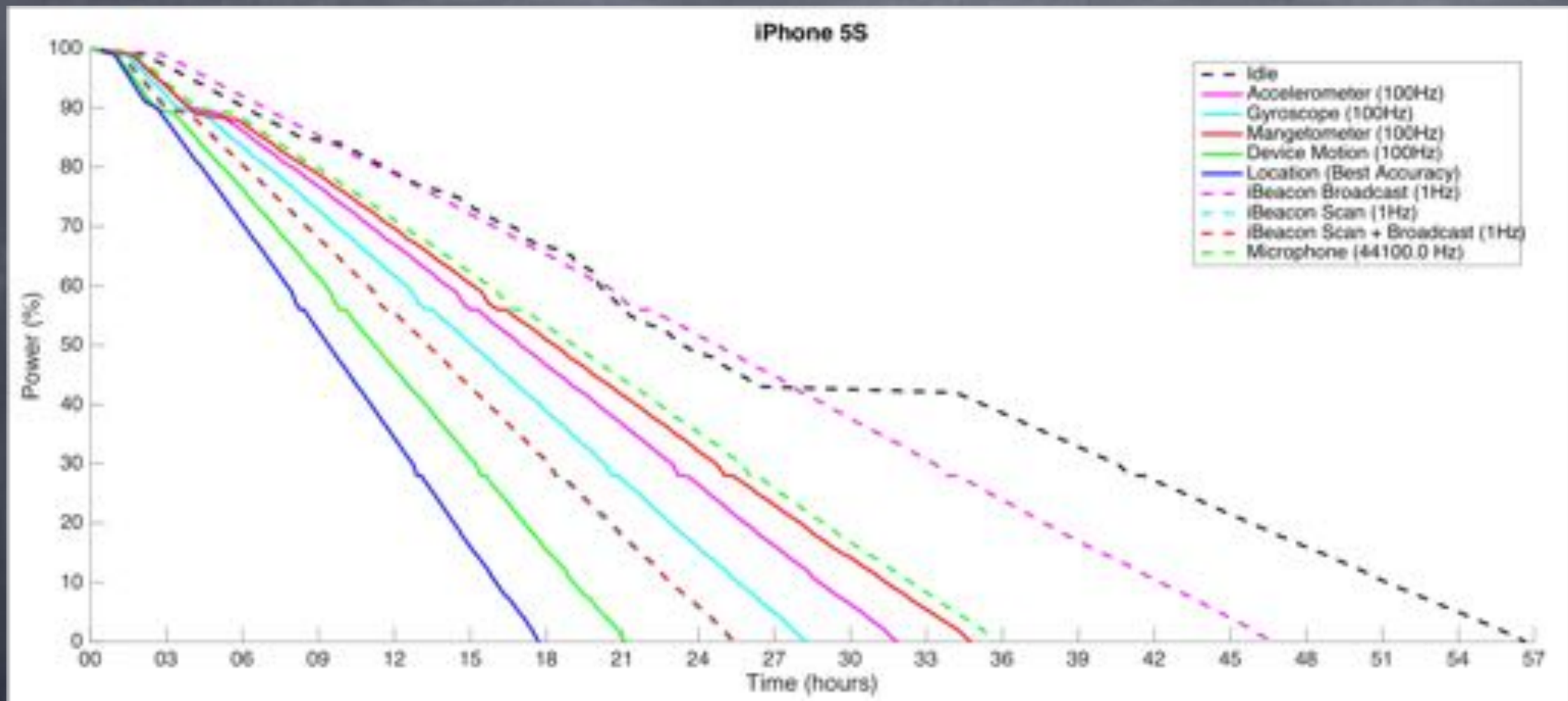


SensingKit

CrowdSense
for iOS / Android

Sensingkit.org
(ACM MobiSys 2017 demos)

Sensing costs energy too!

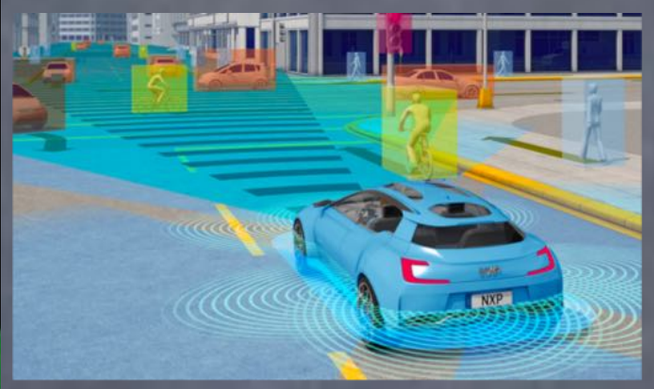


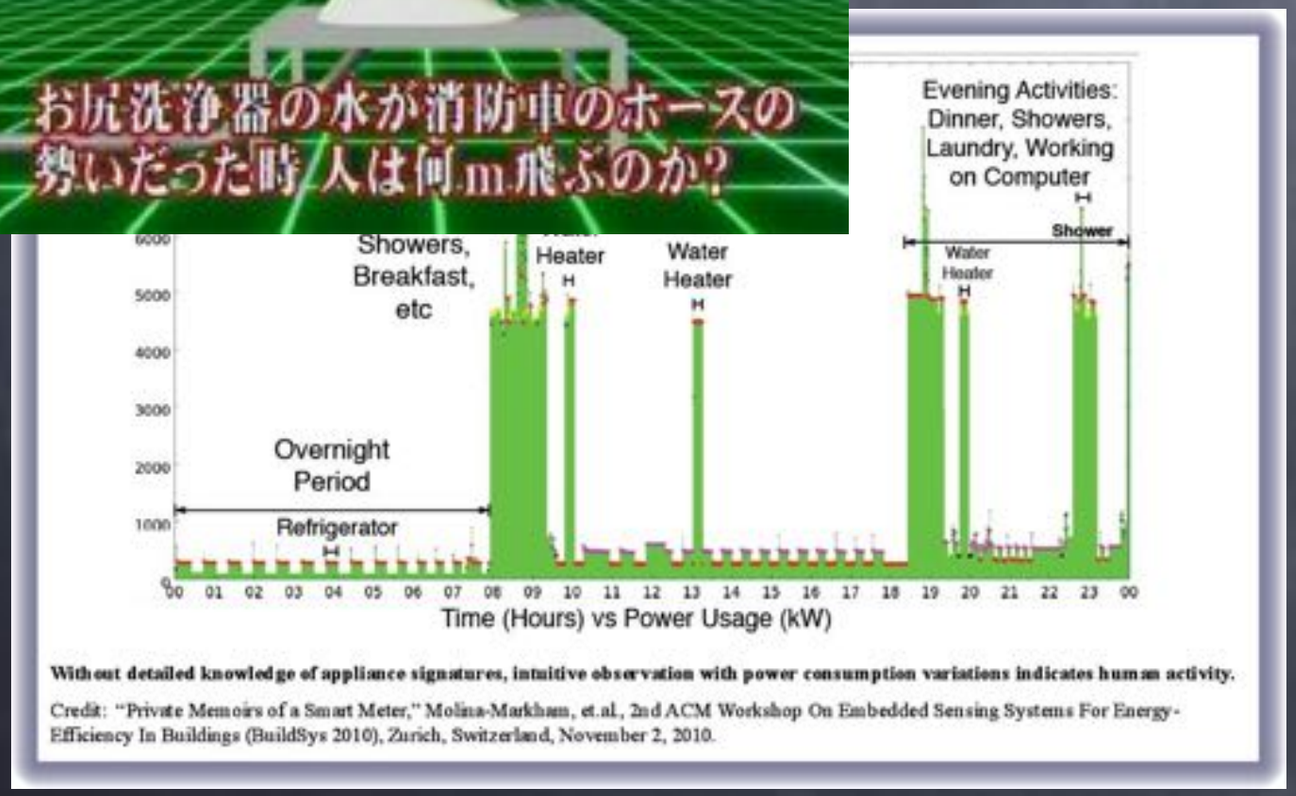
Power consumption of SensingKit running on an iPhone 5S smartphone.

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

Data around us

- IoT devices
- Cyber Physical Systems





www.connectedseeds.org/about/sensors

IoT Traffic

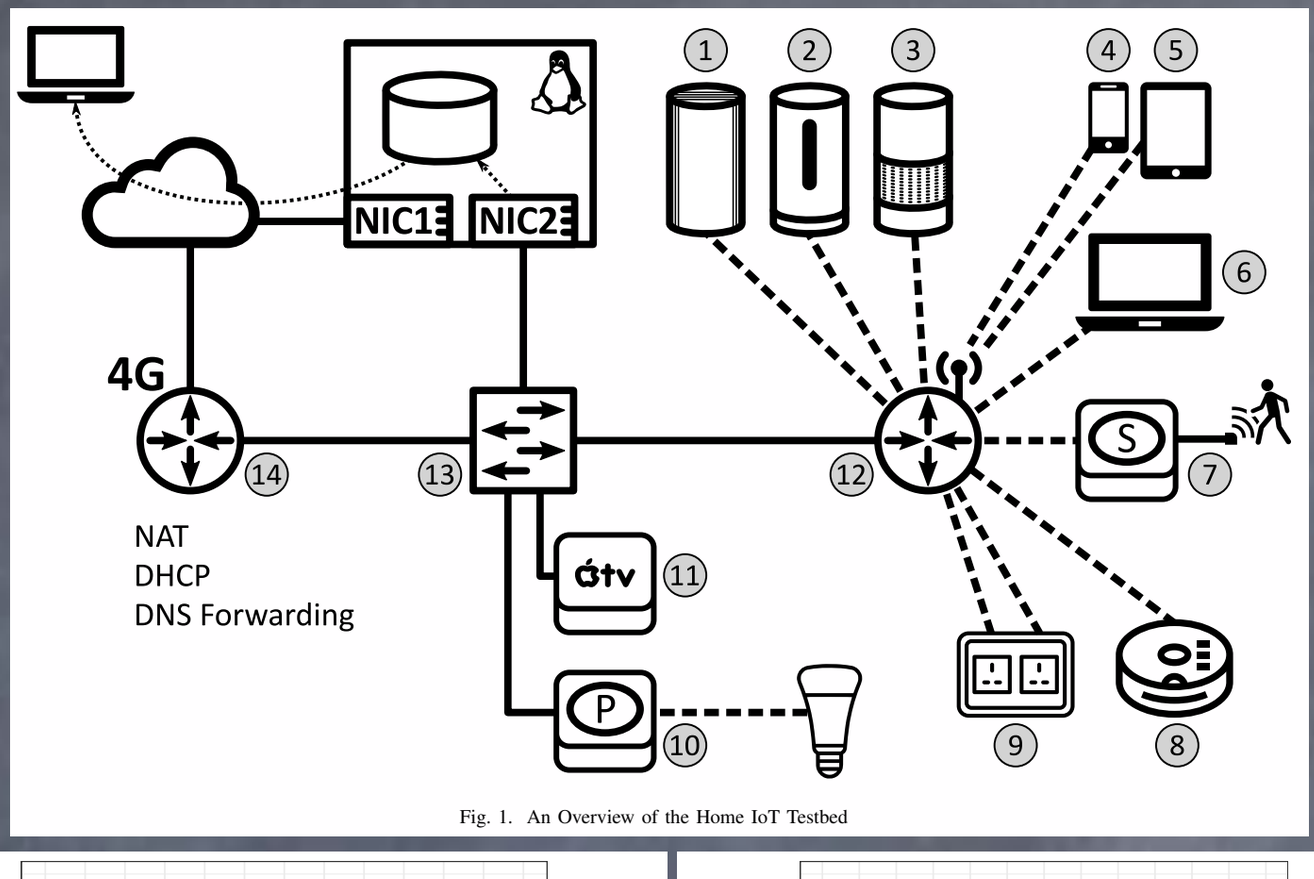
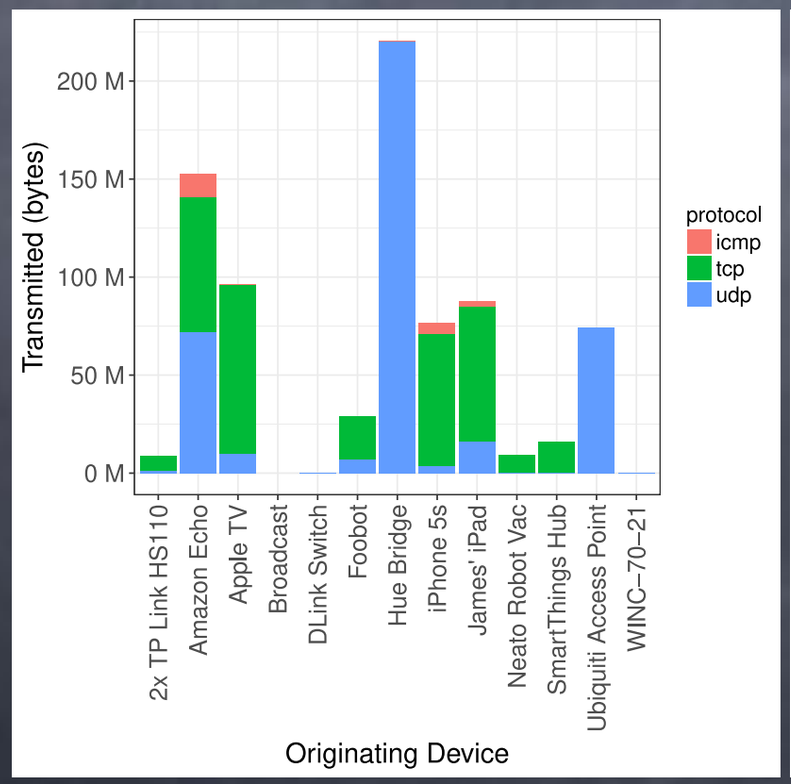
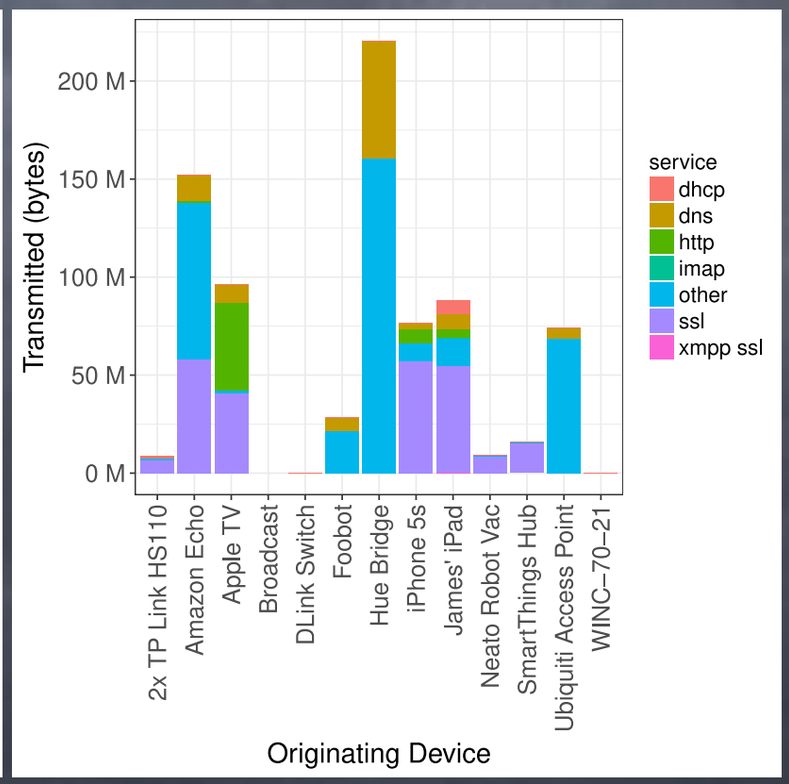


Fig. 1. An Overview of the Home IoT Testbed





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

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

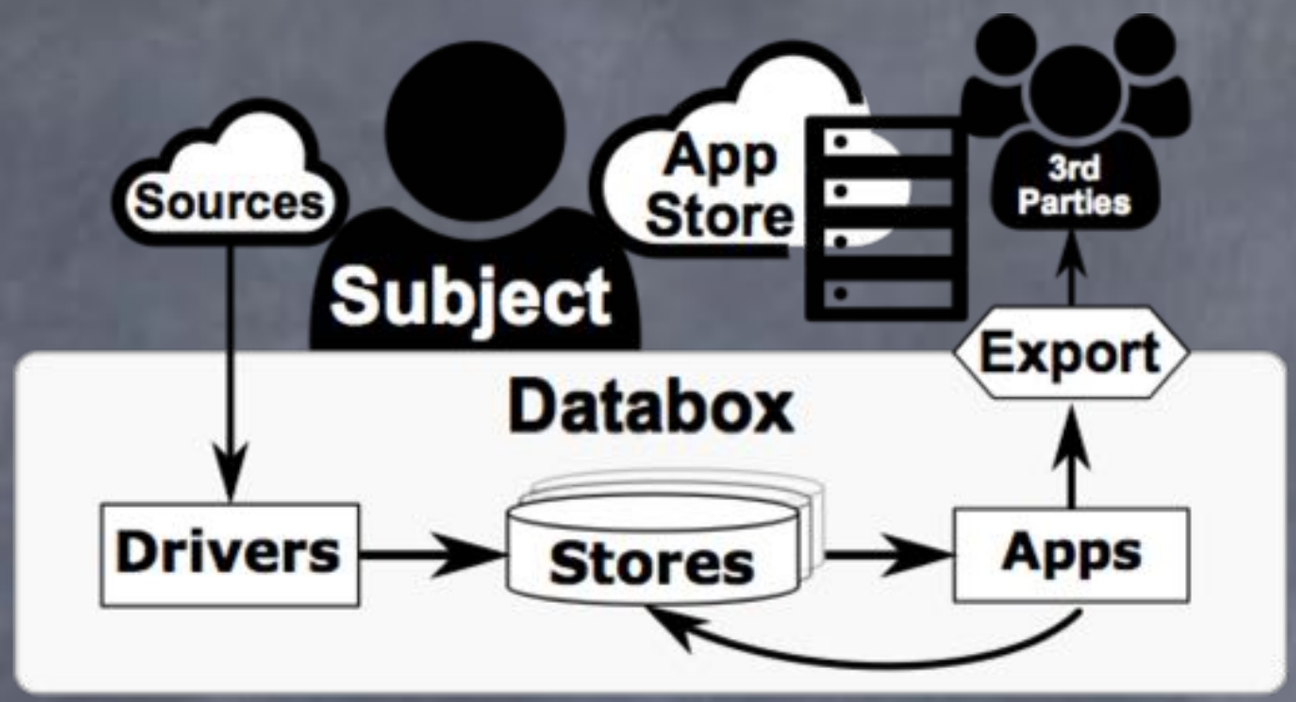
- Real-time control & adaptation, scalability
- Accountability & liability
- Algorithmic bias, privacy, security,...

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

Efforts in this space

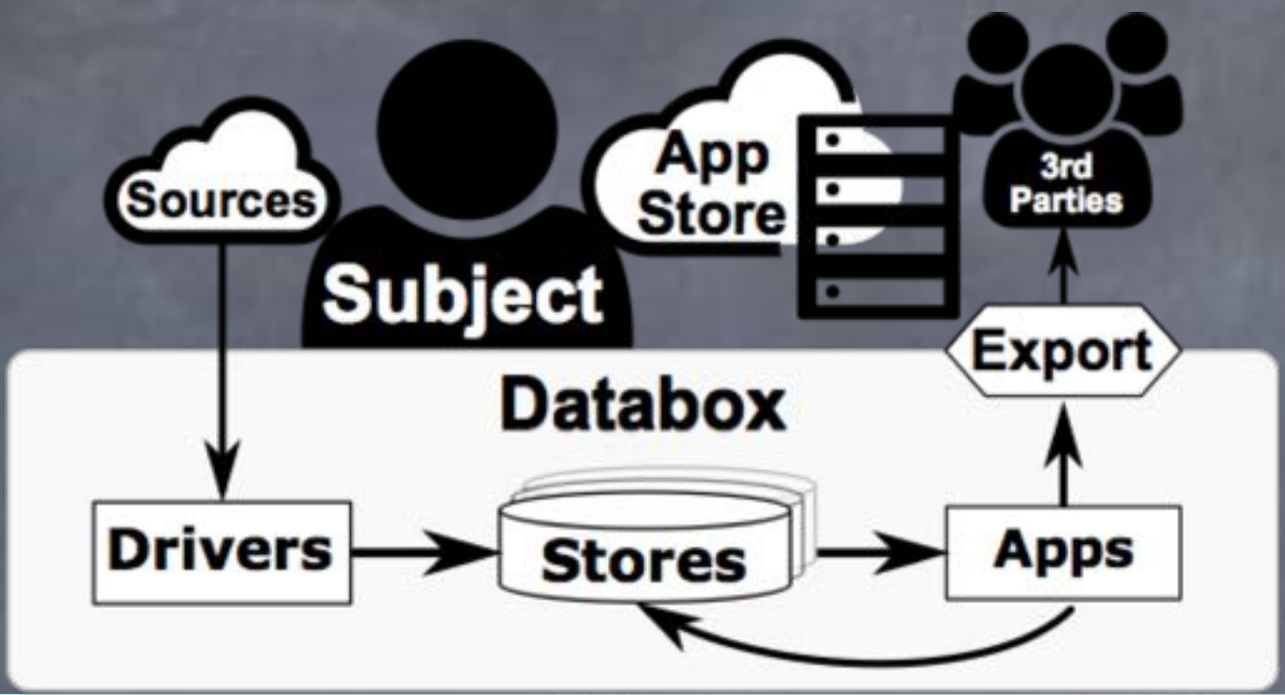
- NyMote
- OpenPDS
- dowse.eu
- Hub of All Things
- Name your favorite data silo…
- ….. Databox

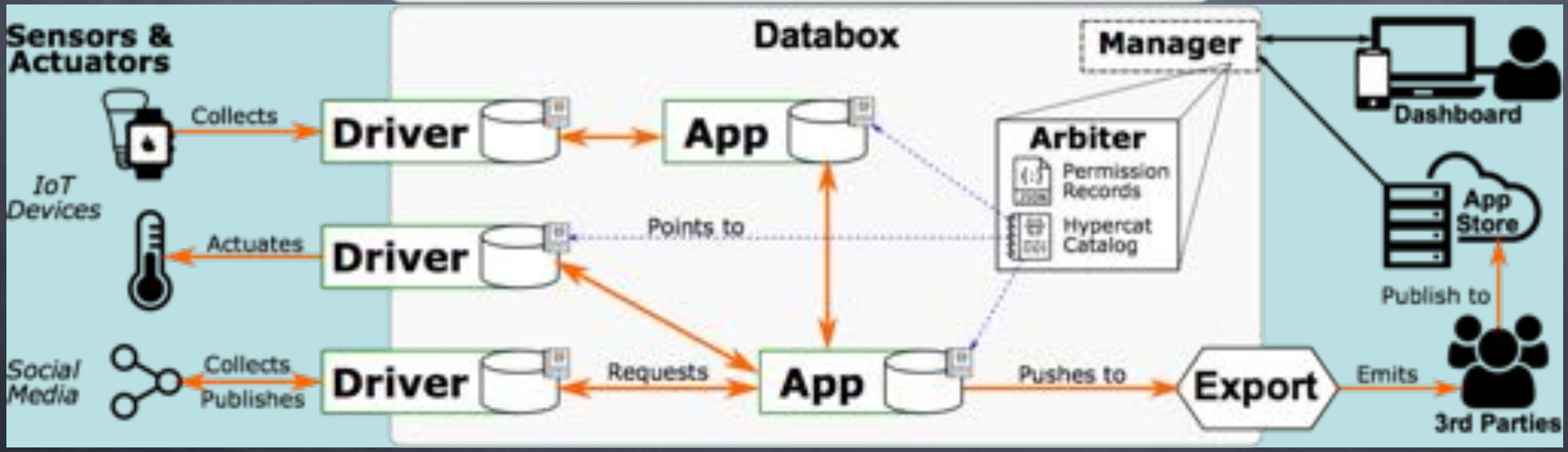
Databox



- 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

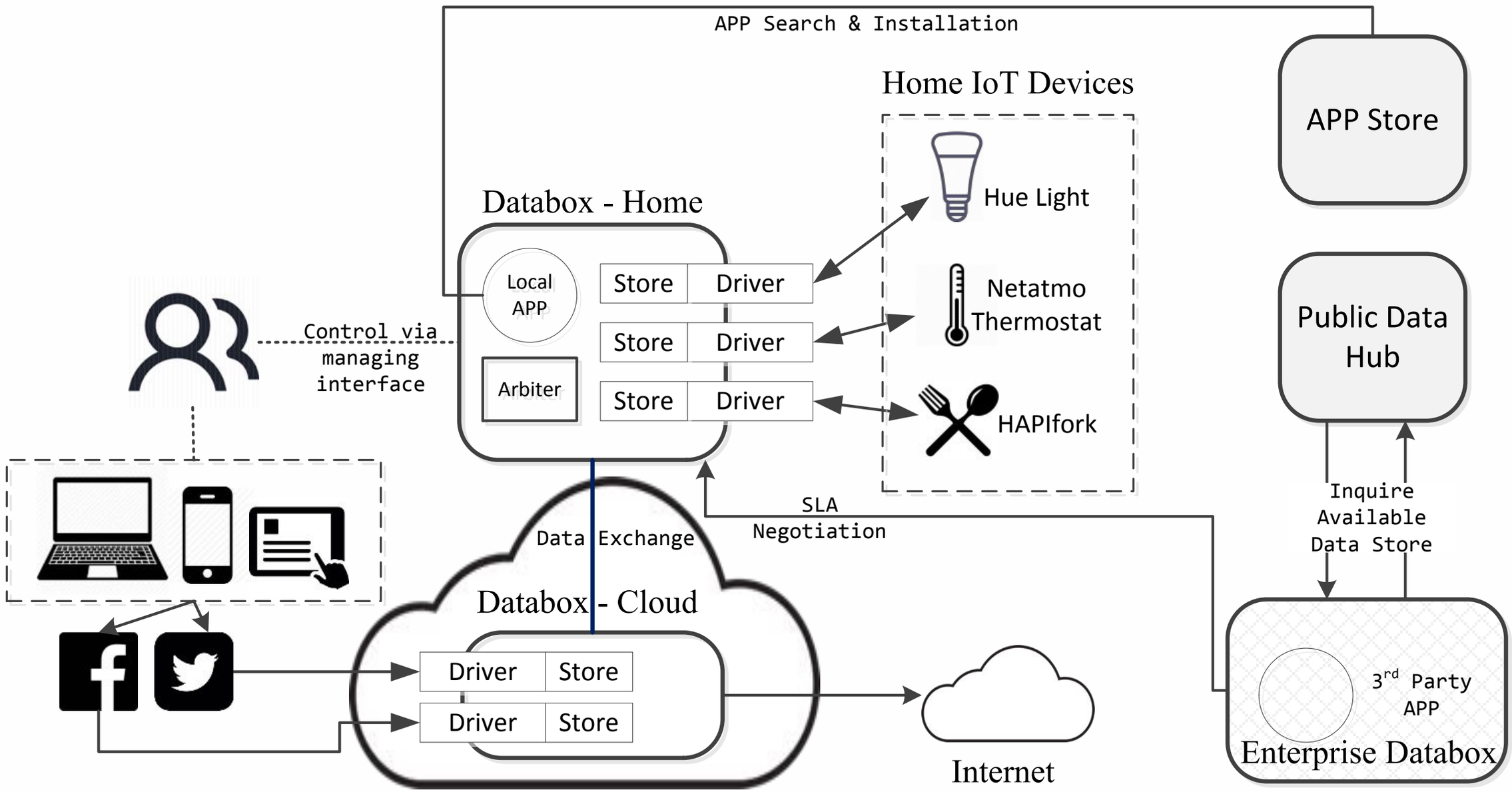
Databox Platform





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

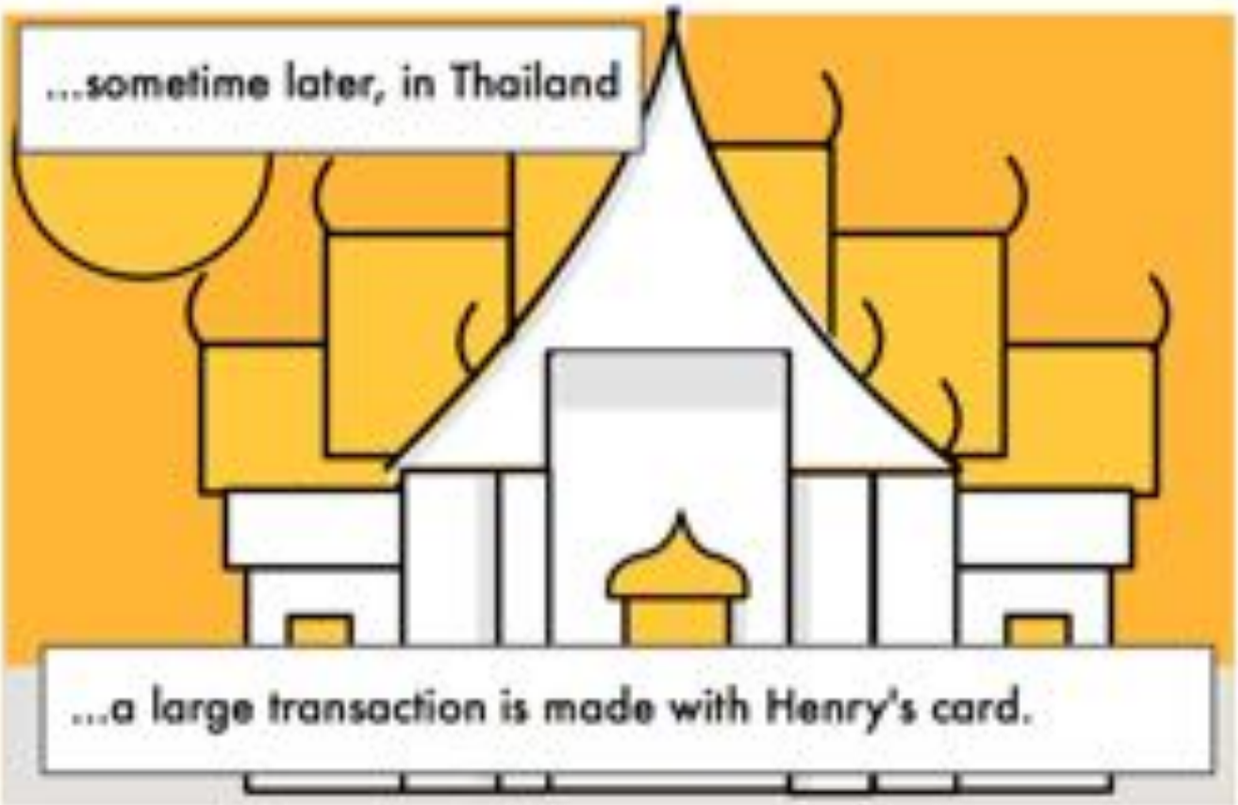
Databox and apps ecosystem



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.

Insure
compare

TESCO

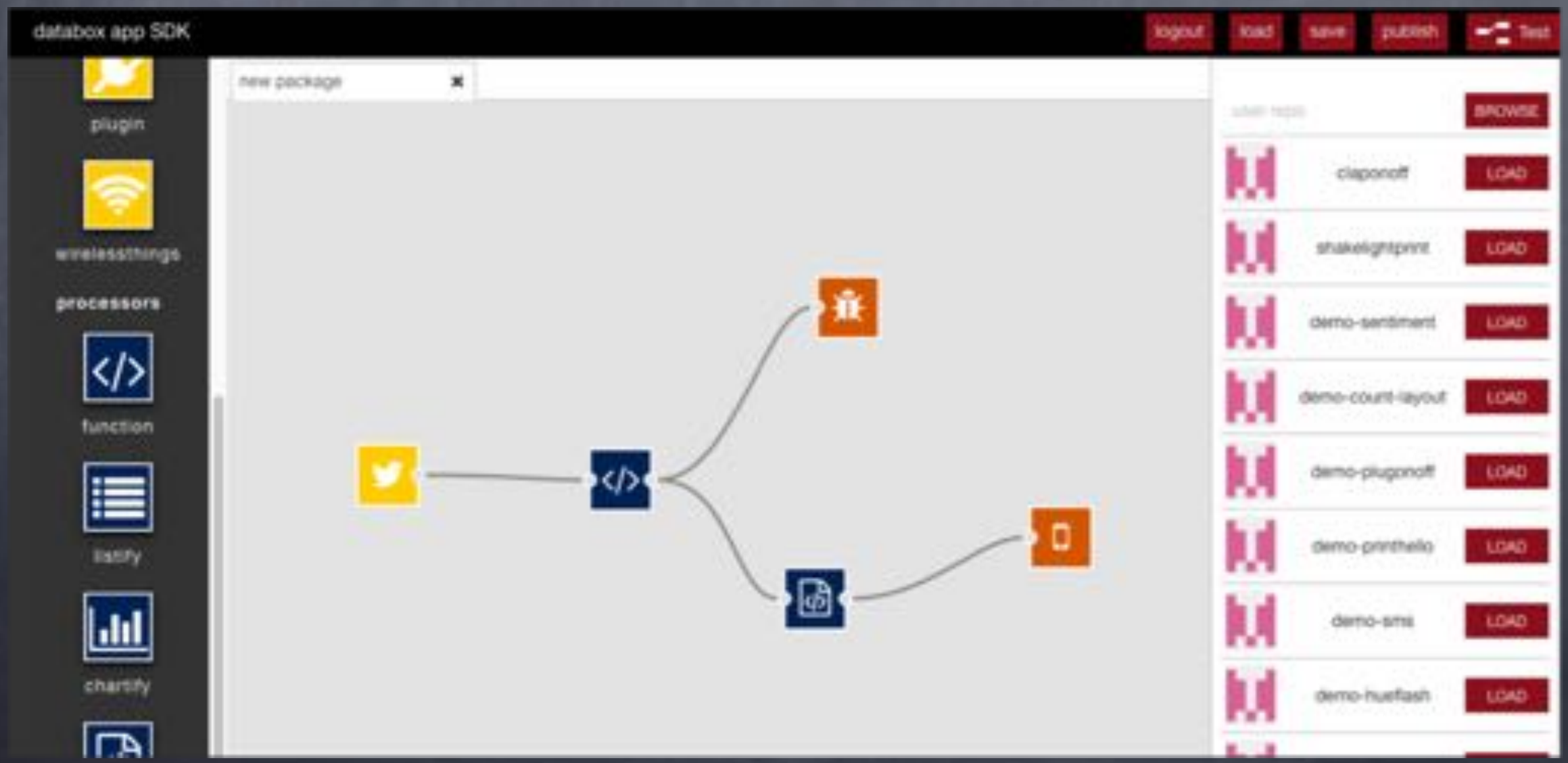
The app analyses her home, mobile, location and grocery shopping data.

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/

Related efforts

OPAL: Bringing the code to the data

- Secured question-and-answer system (API)
  - To be installed in Senegal and Colombia by the end of 2018
- All open-source software & published research

- Developed by:

mit media lab

orange

DATA POP ALLIANCE

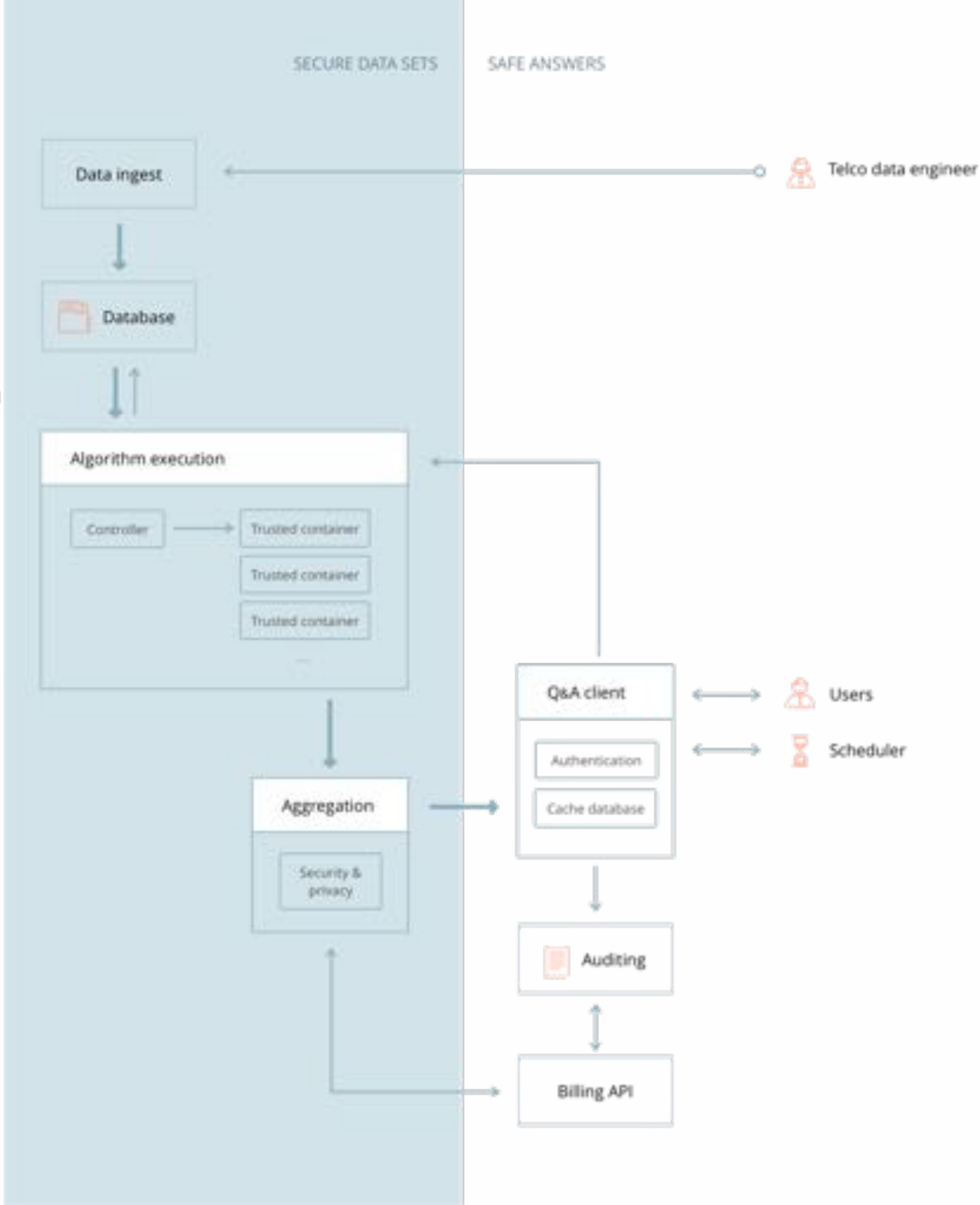
WORLD ECONOMIC FORUM

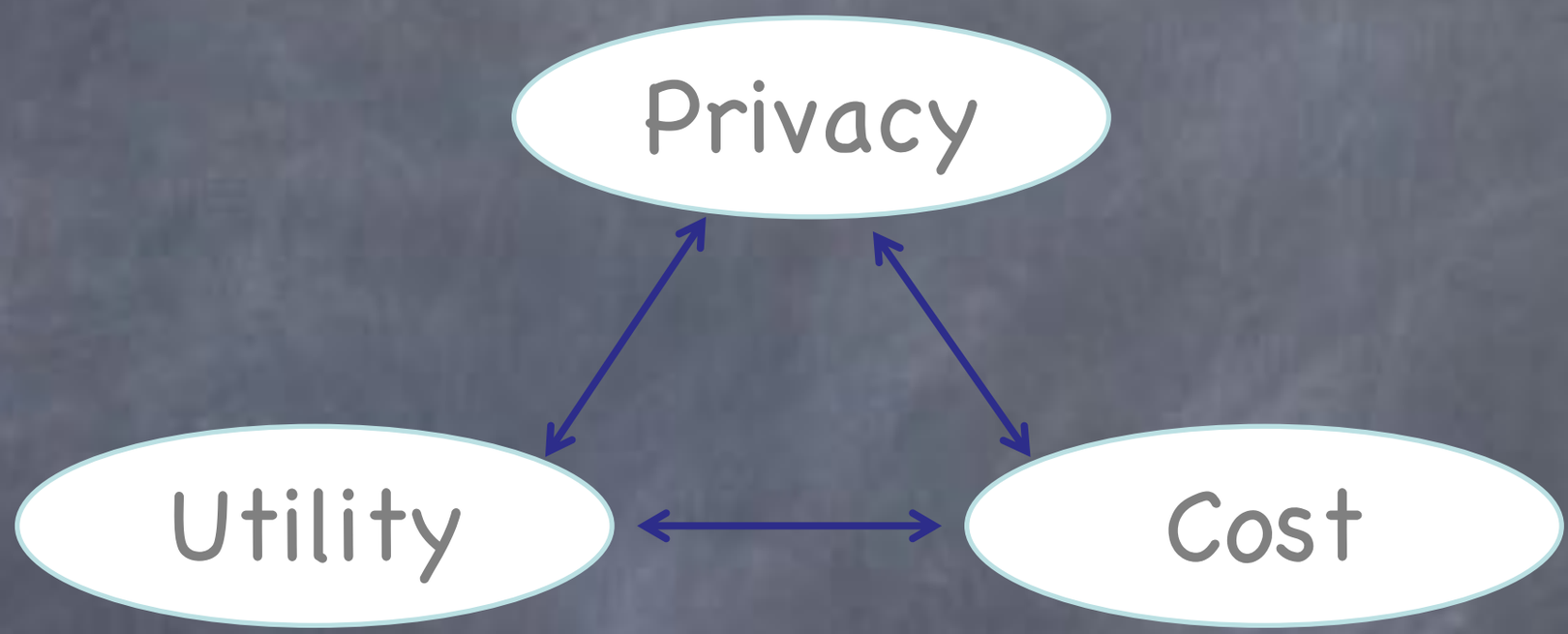
Imperial College London

- With support from:

afd AGENCE FRANÇAISE DE DÉVELOPPEMENT

WORLD BANK GROUP

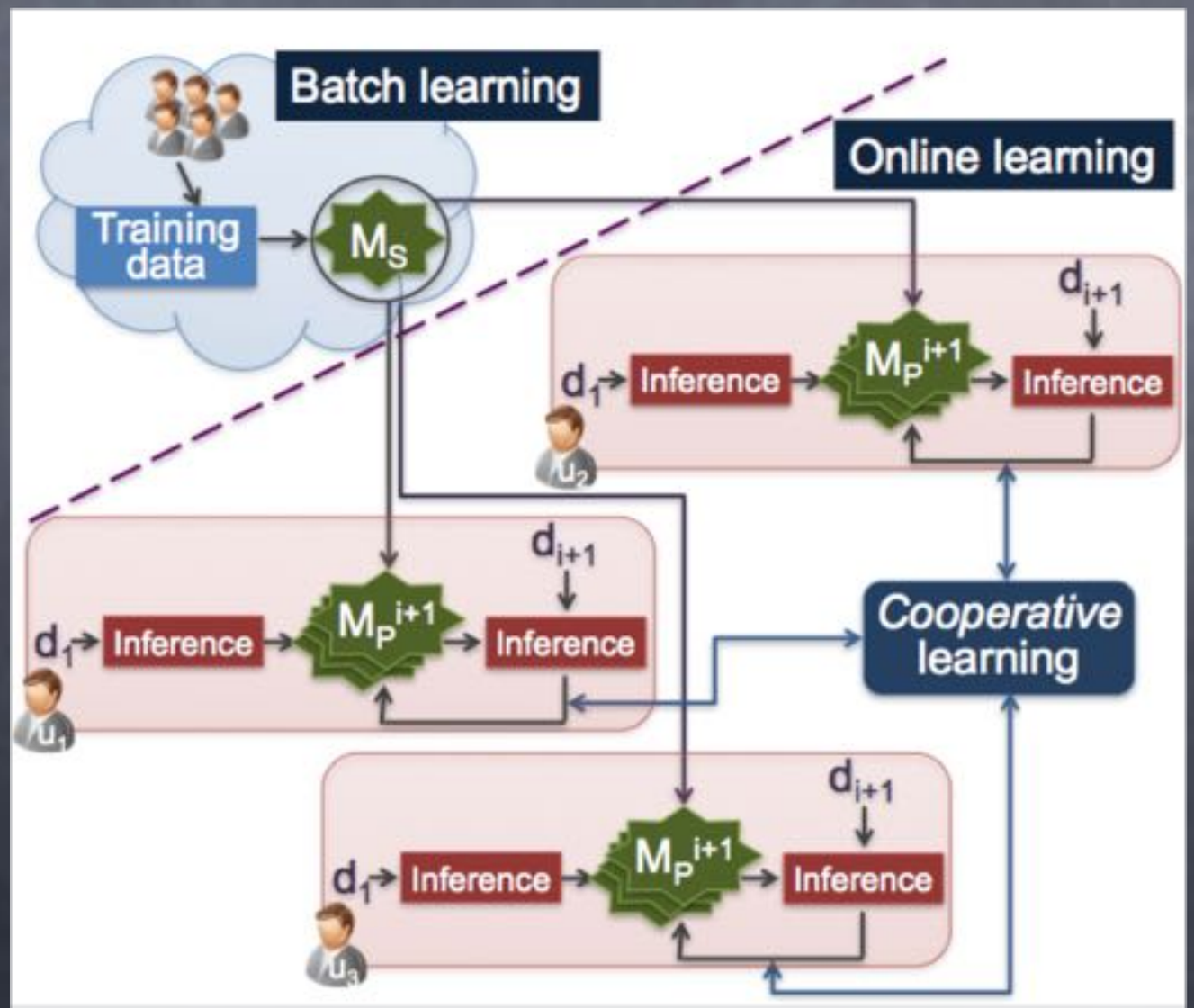



Privacy
Utility
Cost

Distributed Analytics

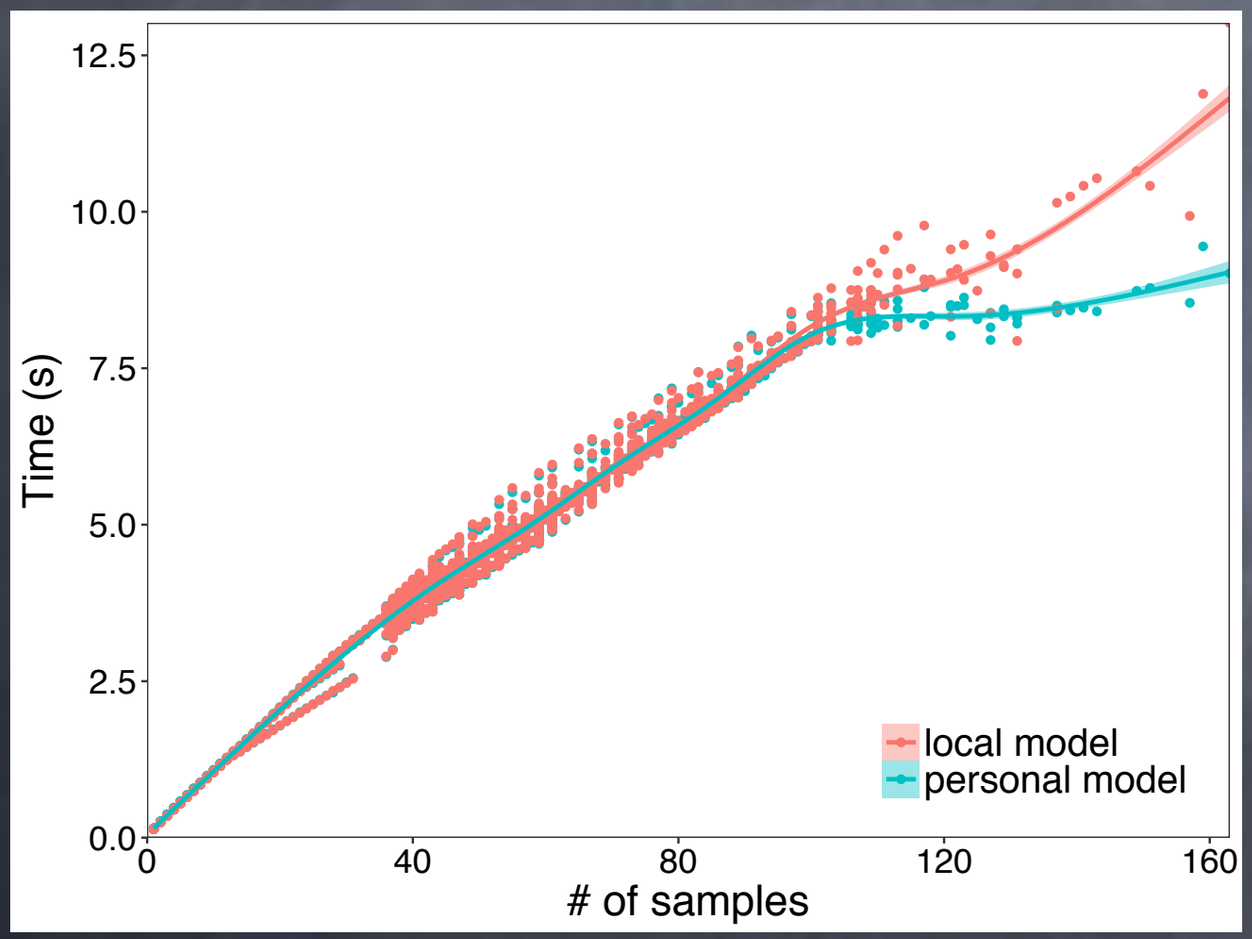
How to handle scale, heterogeneity, dynamics?

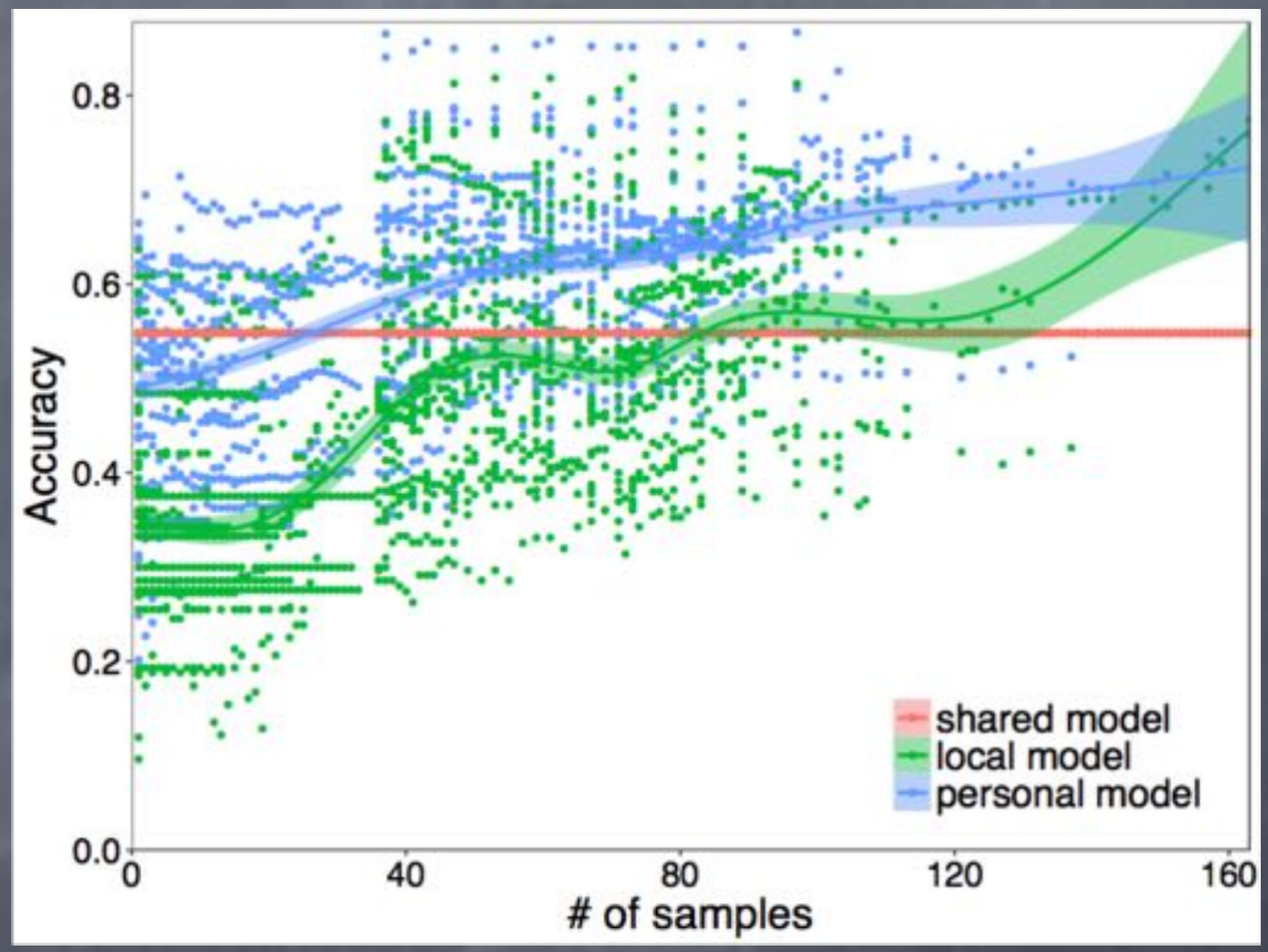
- Cohort vs individual processing
- Distributed model building
- Personalised local analytics



Online Learning

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



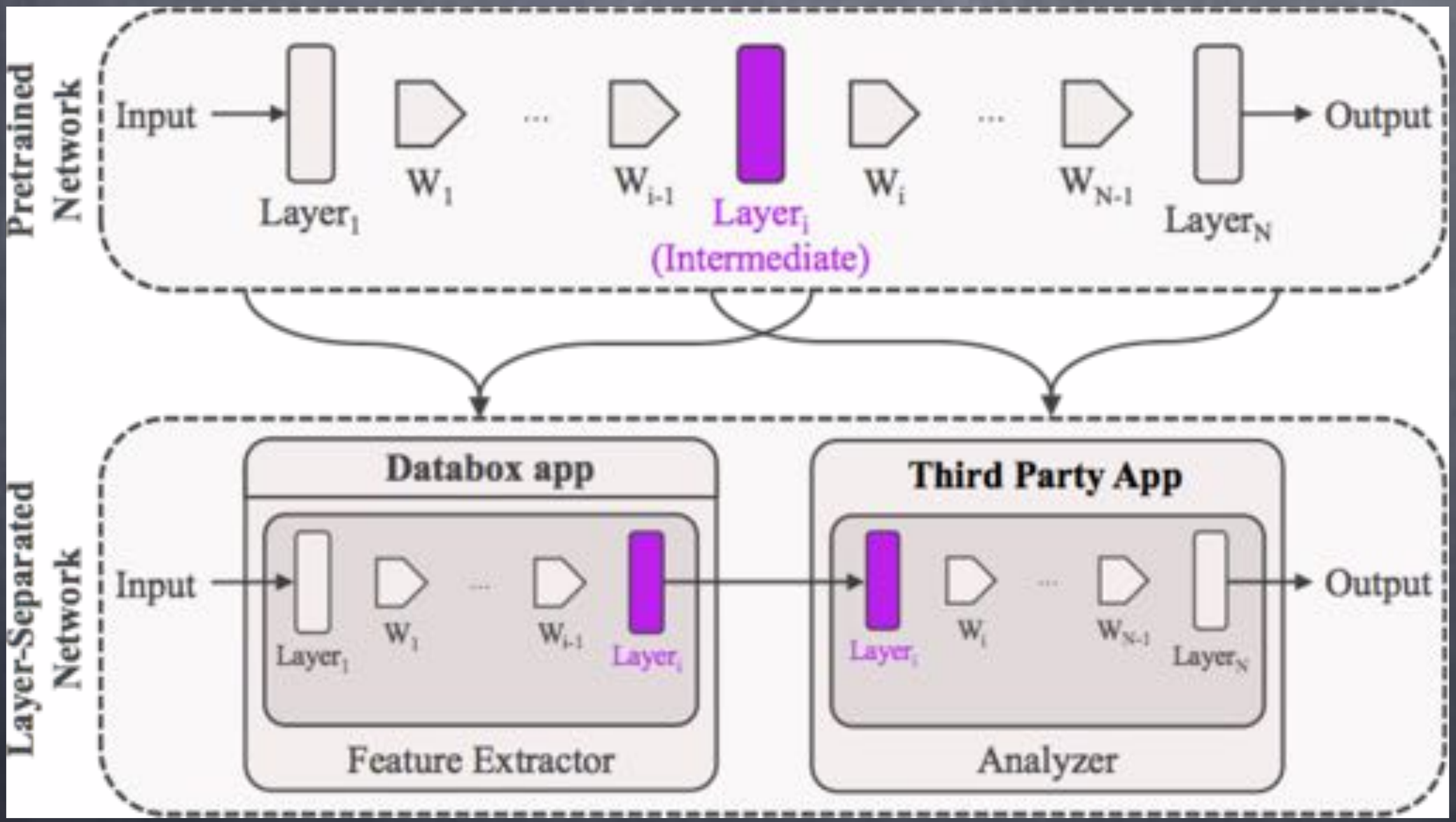


Sandra Servia-Rodriguez, Liang Wang, Jianxin R. Zhao, Richard Mortier, Hamed Haddadi, “Privacy-Preserving Personal Model Training”, The 3rd ACM/IEEE International Conference on Internet-of-Things Design and Implementation, April 2018, Orlando, Florida.

Edge Processing on Sensitive Data

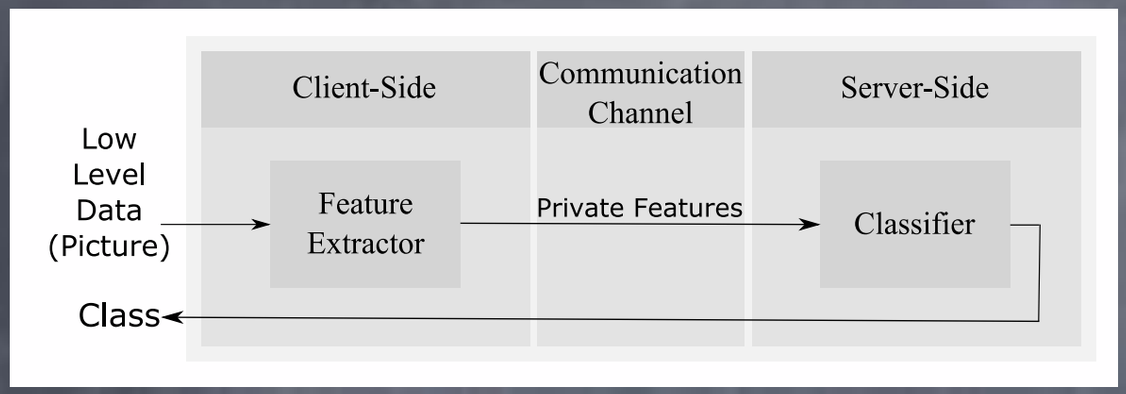
Example: Surveying/Marketing

Privacy-Preserving Analytics

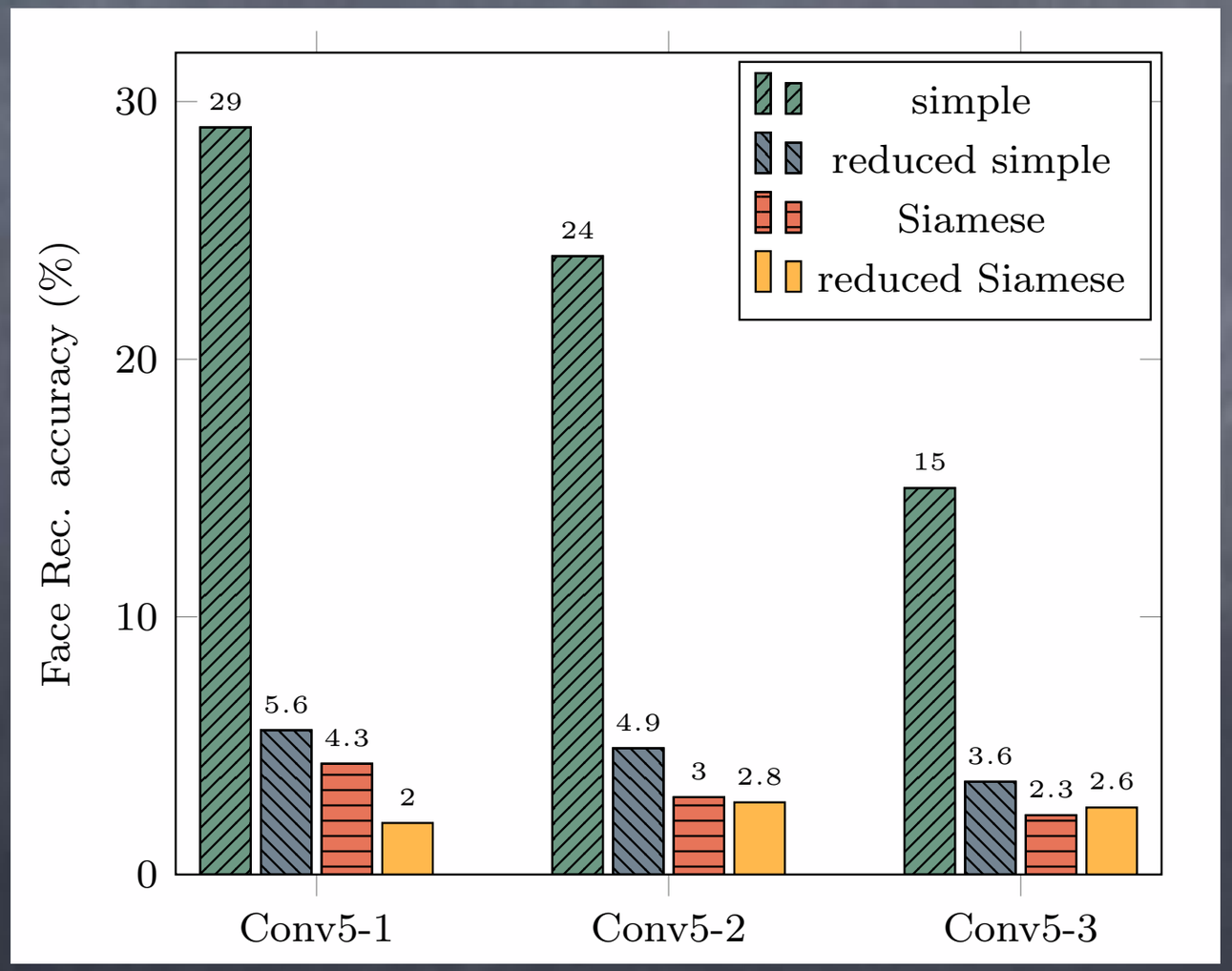


Edge computing paradigm

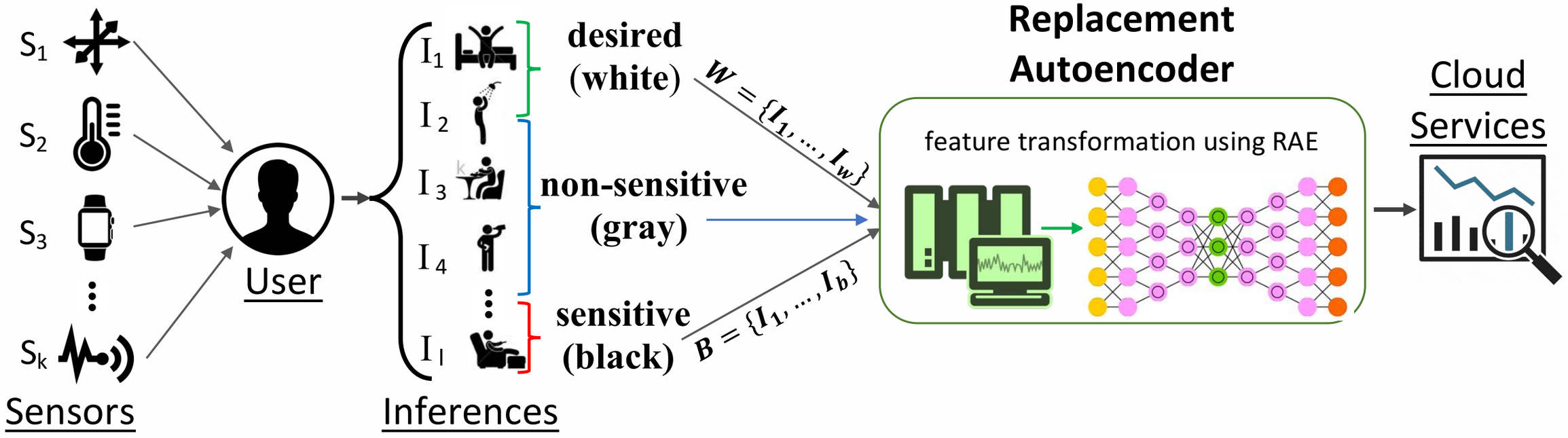
Case study: can we do gender detection without face recognition?

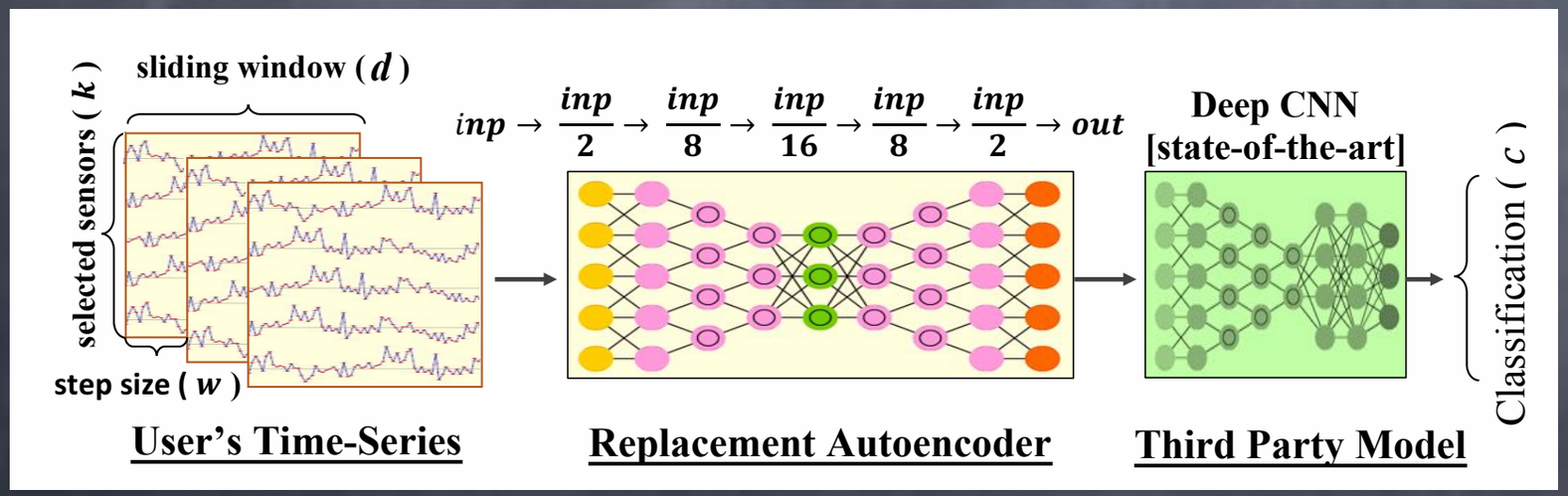


Accuracy on LFW			
	Conv5-1	Conv5-2	Conv5-3
simple	94%	94%	94%
reduced simple	89.7%	87%	94%
Siamese	92.7%	92.7%	93.5%
reduced Siamese	91.3%	92.9%	93.3%



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.

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

threats

Collusion, Generative adversarial Networks, ...

frankmcsherry / blog

Watch 134 Star 726 Fork 66

Code Issues 0 Pull requests 0 Projects 0 Wiki Insights

Branch: master ▾ blog / posts / 2017-10-27.md Find file Copy path

frankmcsherry mortifying typo 8781951 on 30 Oct

1 contributor

200 lines (101 sloc) 28 KB Raw Blame History

Deep learning and differential privacy

Today we'll look at two papers from ACM CCS, one about to be presented at CCS 2017, and one presented at CCS 2015 two years ago. Both are chiefly about deep learning and privacy, but they touch on *differential* privacy, and so I thought it would be good to talk through the two of them. The two papers are somewhat in opposition, which means there is unlikely to be a happy resolution.

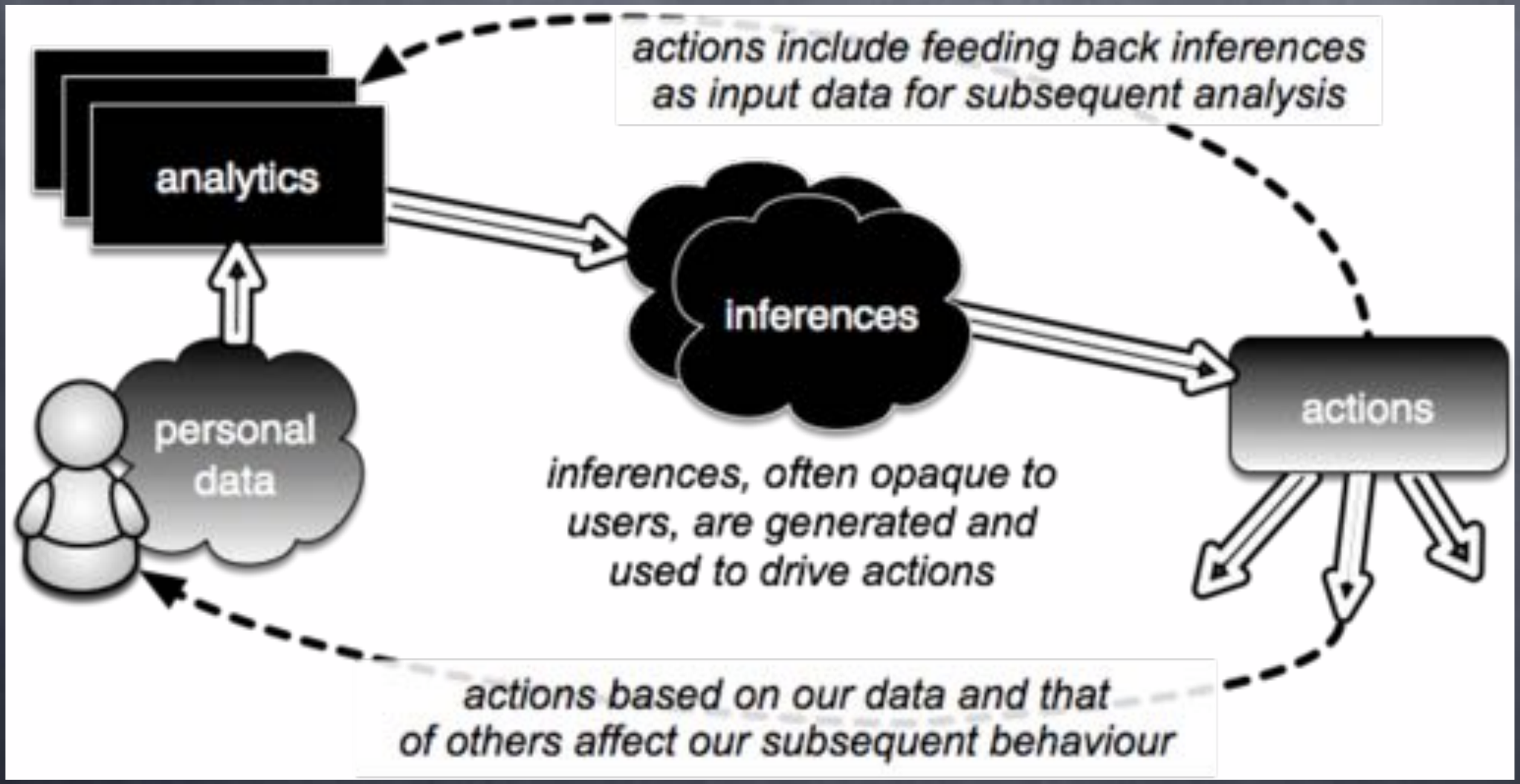
The first paper is Privacy-Preserving Deep Learning, from ACM CCS 2015. The paper describes an approach to deep learning in which participants do not share their raw data, but instead share only partial information about how their raw data affect the training process. From their abstract:

Security, Privacy, Analytics, scalability dichotomy

Scare stories: Mirai IoT Botnet, Smart TVs transmitting conversations & profiling, CIA Hacks, Webcam viewing websites, spamming fridge, Amazon echo ordering dolls, eavesdropping toys...

- IoT device and Network Isolation to limit coordinated attacks
- Crowdsourced or semi-supervised policing & anomaly detection
- Can not rely on constant connectivity
  - Is the "cloud" or your DSL connection always *on*?
  - Remember Amazon AWS outages?

Human-Data Interaction



Legibility, Agency, Negotiability

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