User-Centric Personal Data Analytics on the Edge

Hamed Haddadi

Queen Mary University of London
--> Imperial College London

The Data Ecosystem

Data about us:

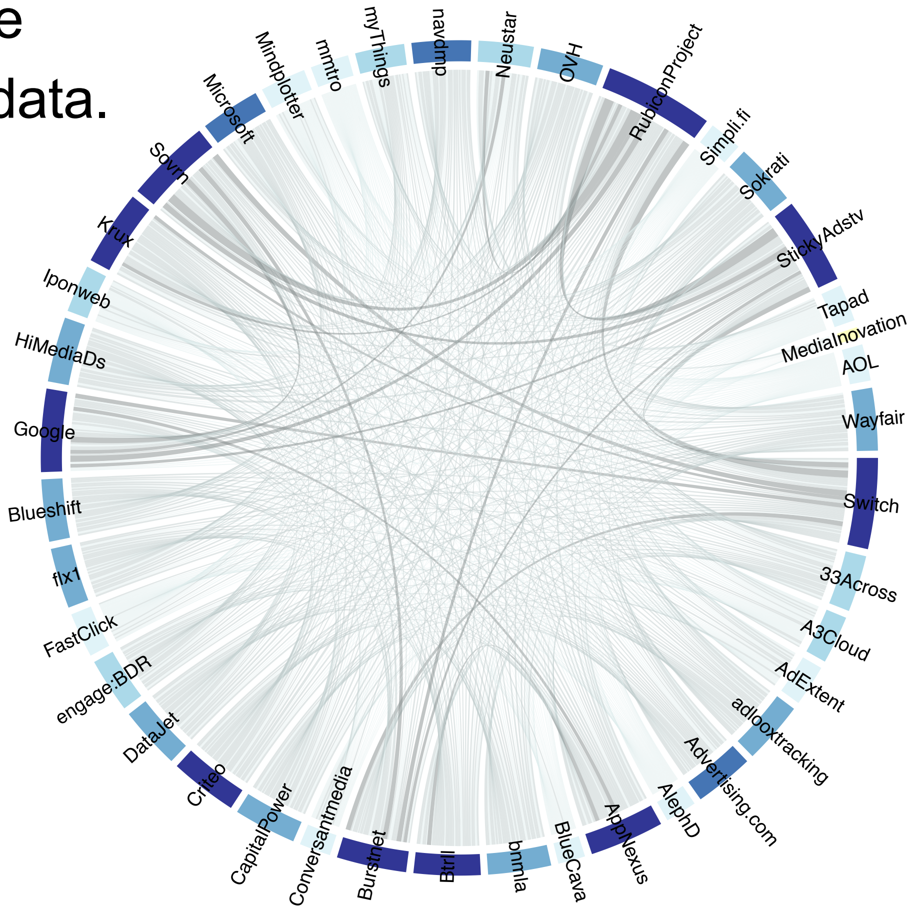
Data generated by us:

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 include data we are not even aware of. Hence Provenance is a major issue.



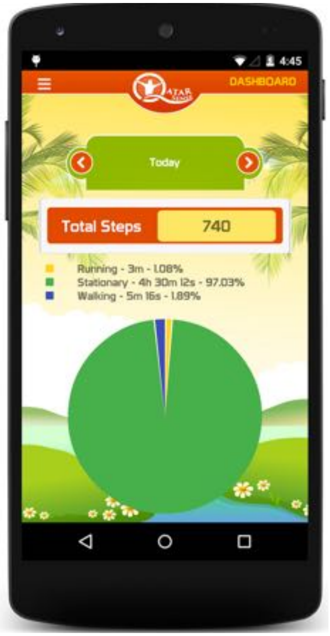
TMA 2014, PAM 2016 and “Anatomy of the Third-Party Web Tracking Ecosystem” on MIT TR 2014.

- 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

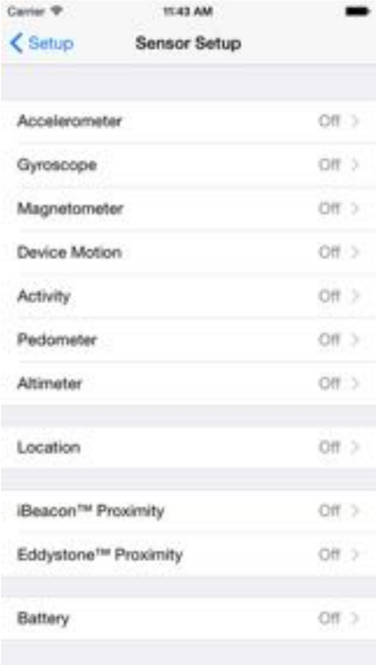
- Online Social media
- Wearable devices
  - Signals indicative of physical & mental health
  - Largely suffering from data isolation and poor user interaction (see publications: qmwearables.eecs.qmul.ac.uk)

Emotional S... 8:15
?
very happy



Total Steps 740

🟨 Running - 3m - 1.08%
🟩 Stationary - 4h 30m 12s - 97.03%
🟦 Walking - 5m 16s - 1.89%



Sensor Setup

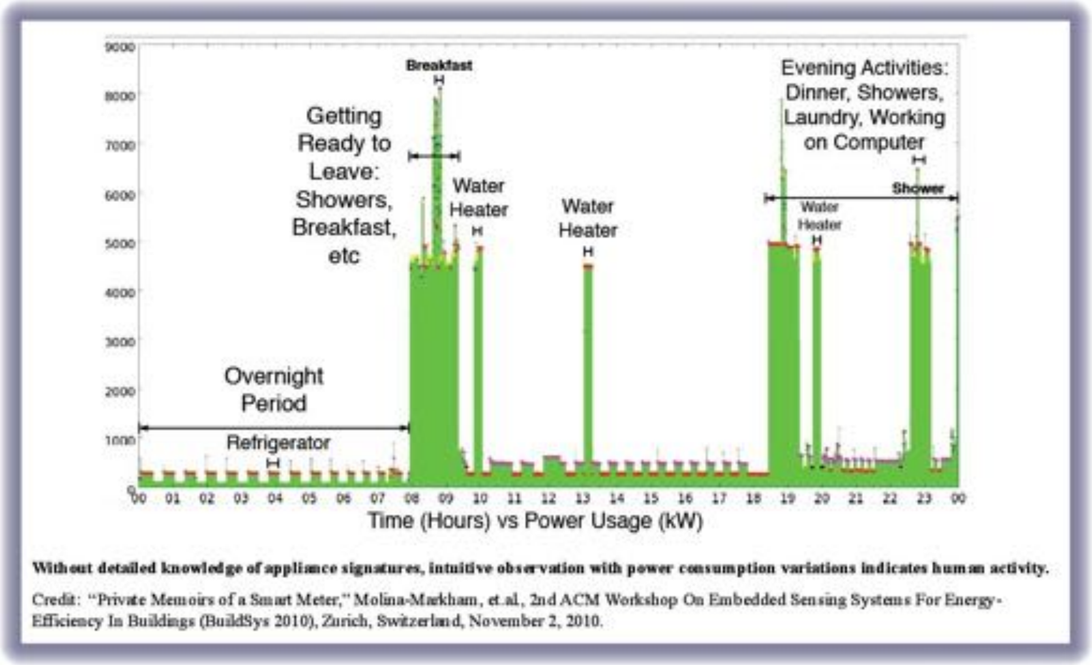
Accelerometer	Off
Gyroscope	Off
Magnetometer	Off
Device Motion	Off
Activity	Off
Pedometer	Off
Altimeter	Off
Location	Off
iBeacon™ Proximity	Off
Eddystone™ Proximity	Off
Battery	Off

SensingKit

CrowdSense
for iOS / Android

Data around us

- IoT devices
- Cyber Physical Systems



Without detailed knowledge of appliance signatures, intuitive observation with power consumption variations indicates human activity.

Credit: "Private Memoirs of a Smart Meter," Molina-Markham, et.al., 2nd ACM Workshop On Embedded Sensing Systems For Energy-Efficiency In Buildings (BuildSys 2010), Zurich, Switzerland, November 2, 2010.

www.connectedseeds.org/about/sensors

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,...
  - Same with IoT/mobile data: see "Privacy Leakage in Mobile Computing: Tools, Methods, and Characteristics" 2014.

Can we do detailed, user-centric, contextual analytics <u>without</u> privacy disasters and legal challenges?

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

https://www.stickermule.com/marketplace/3442-there-is-no-cloud

Outline

- Introduction & Motivations
- The Databox platform
- Privacy-preserving sensing & analytics

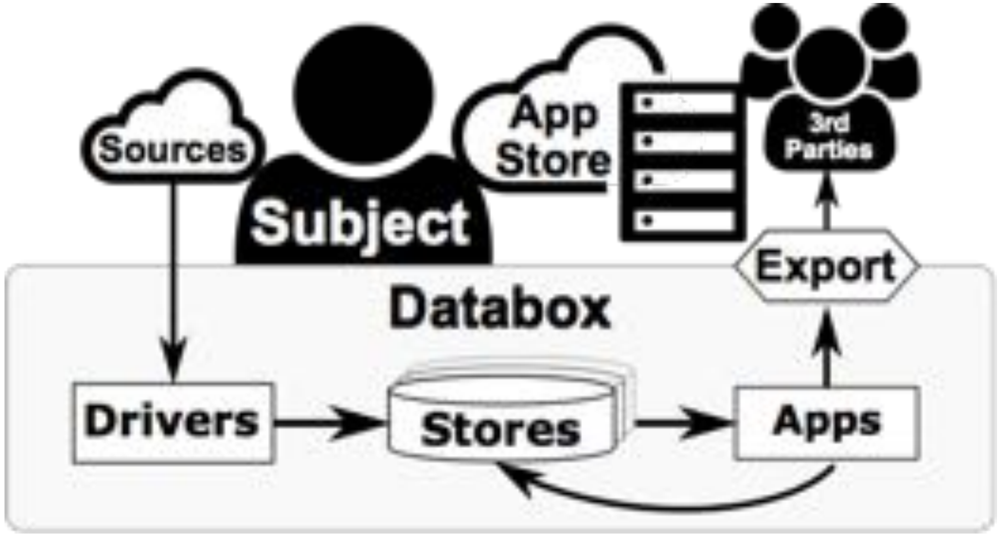
Databox vision

- An open-source personal networked system:
  - collates, curates, and mediates access to our personal data.
  - Enables interaction, sense-making, and privacy-preserving analytics on personal data, with potential wider societal benefits (Haddadi et al., CCR 2013)

- **Not** yet another data silo:
  - cooperative design approach, involving engagement with **all** stakeholders (sources, collectors, processors, organisations, and subjects)

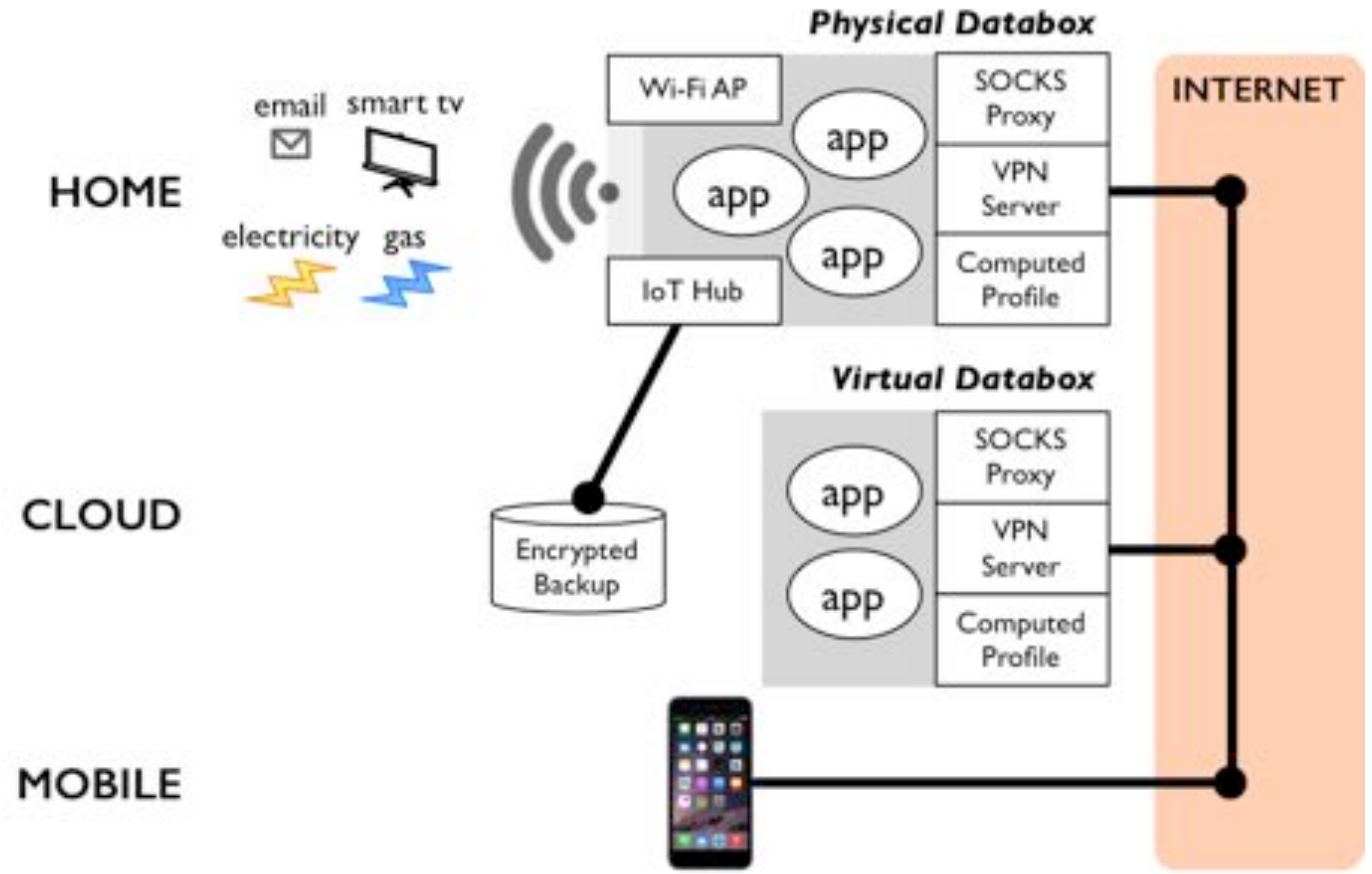
See Haddadi et al., "Personal Data: Thinking Inside the Box", (MIT-TR, Aarhus 2015)

Databox



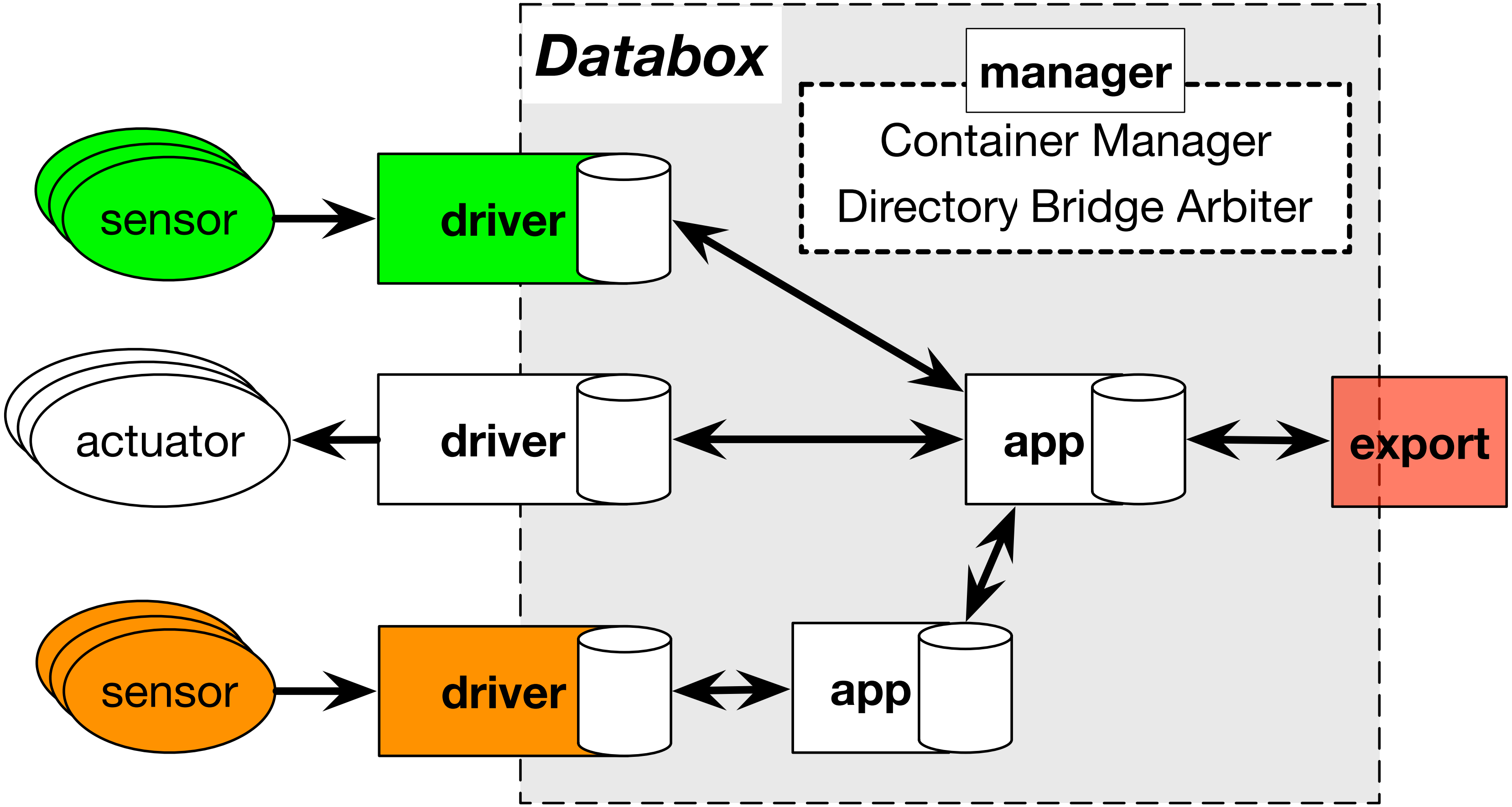
- Mediates access to data, stored locally as appropriate
- Computations (*apps*) move to data, not data to compute
- Maintain control over internal comms and export
- All operations logged for users to inspect, control

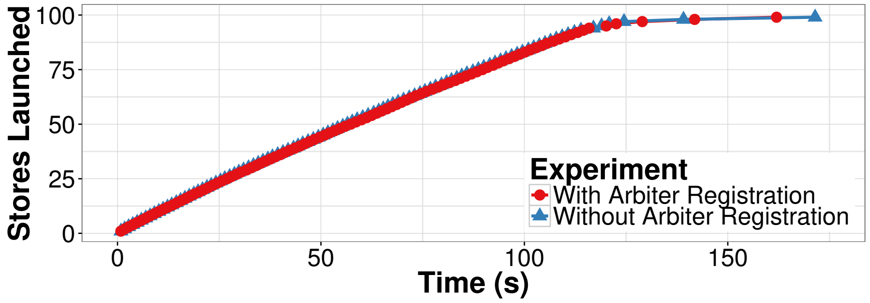
Privacy-Aware Personal Data Platform



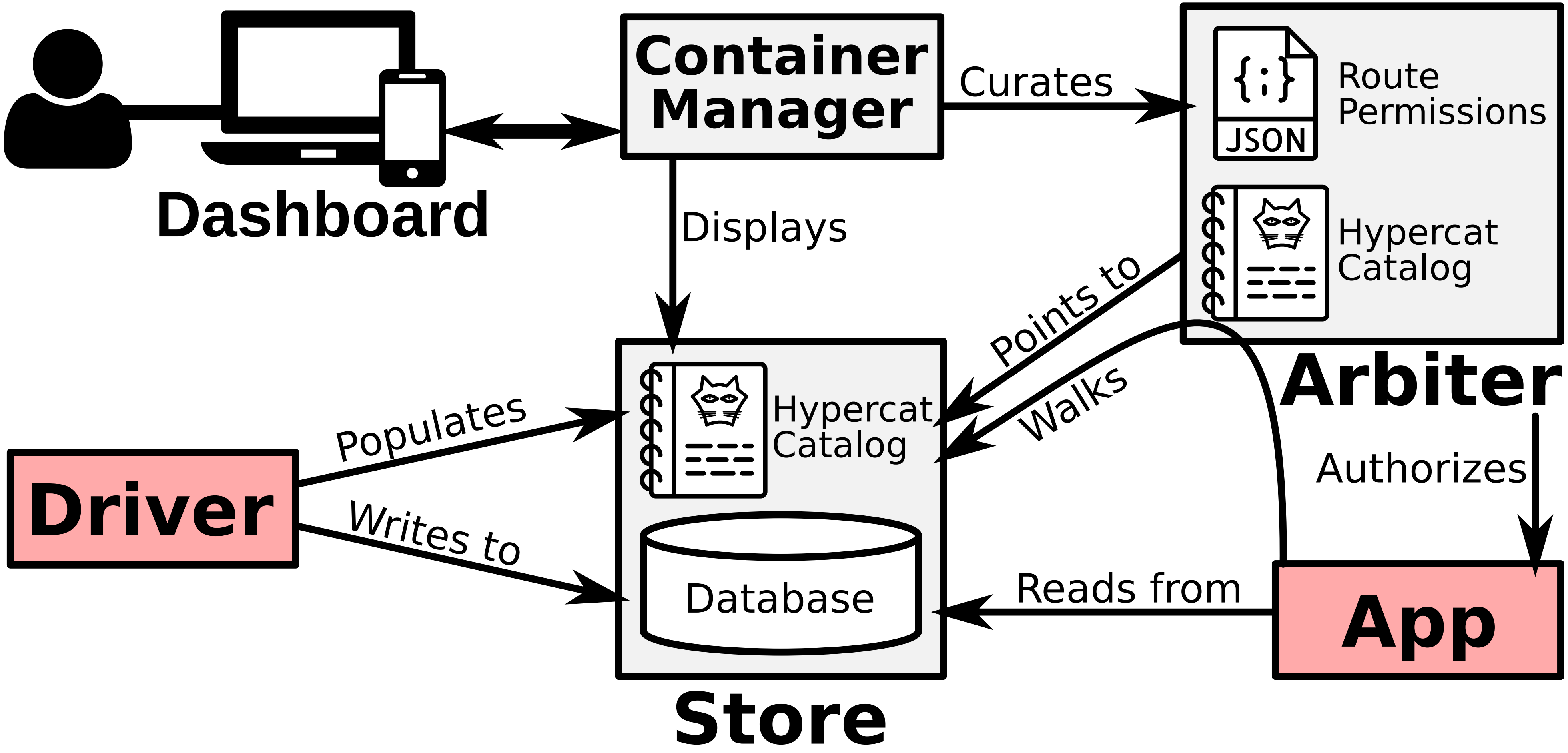
EPSRC Databox: Privacy-Aware Infrastructure for Managing Personal Data
3-years, started October 2016: www.databoxproject.uk

System architecture

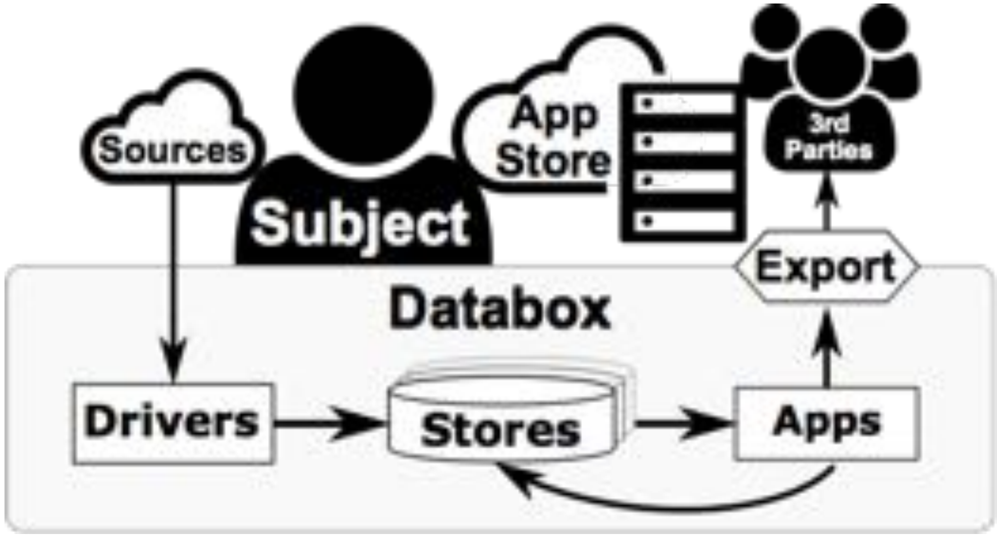


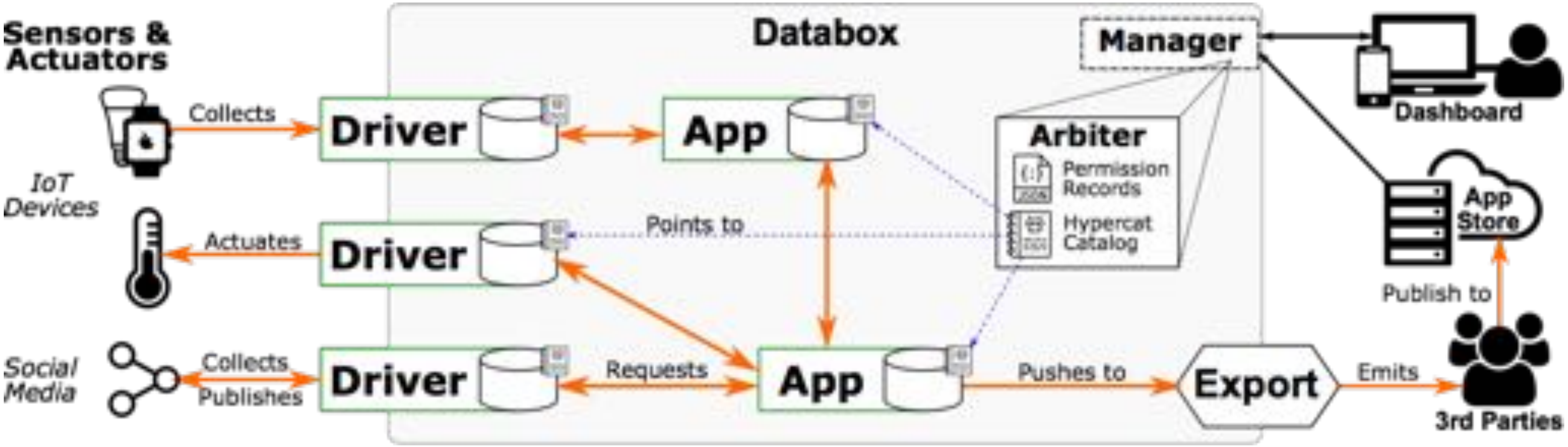


Interaction between the components



Databox Platform

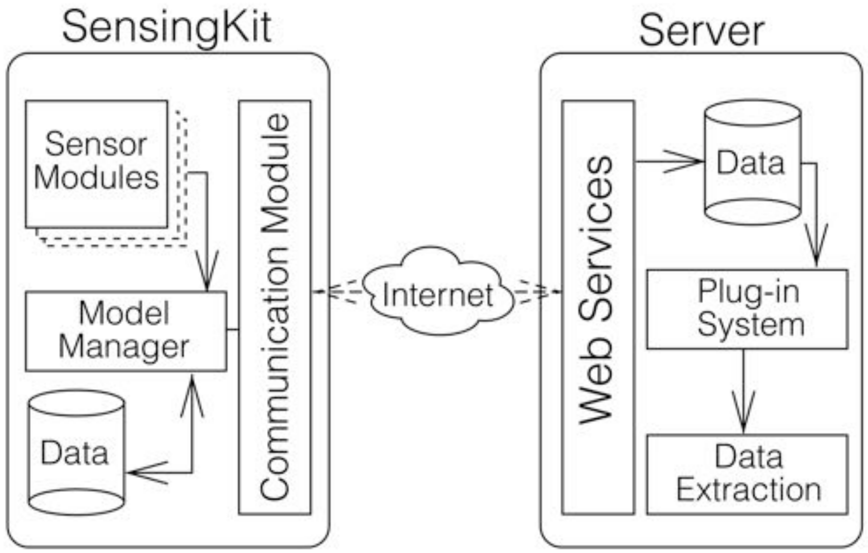




https://github.com/me-box/databox

Integrating mobile sensing

- Smartphone sensors an invaluable source of external information.
- Energy efficiency and privacy are major challenges in this space (see www.sensingkit.org).
  - Potential dual approach to separate data and processing stages



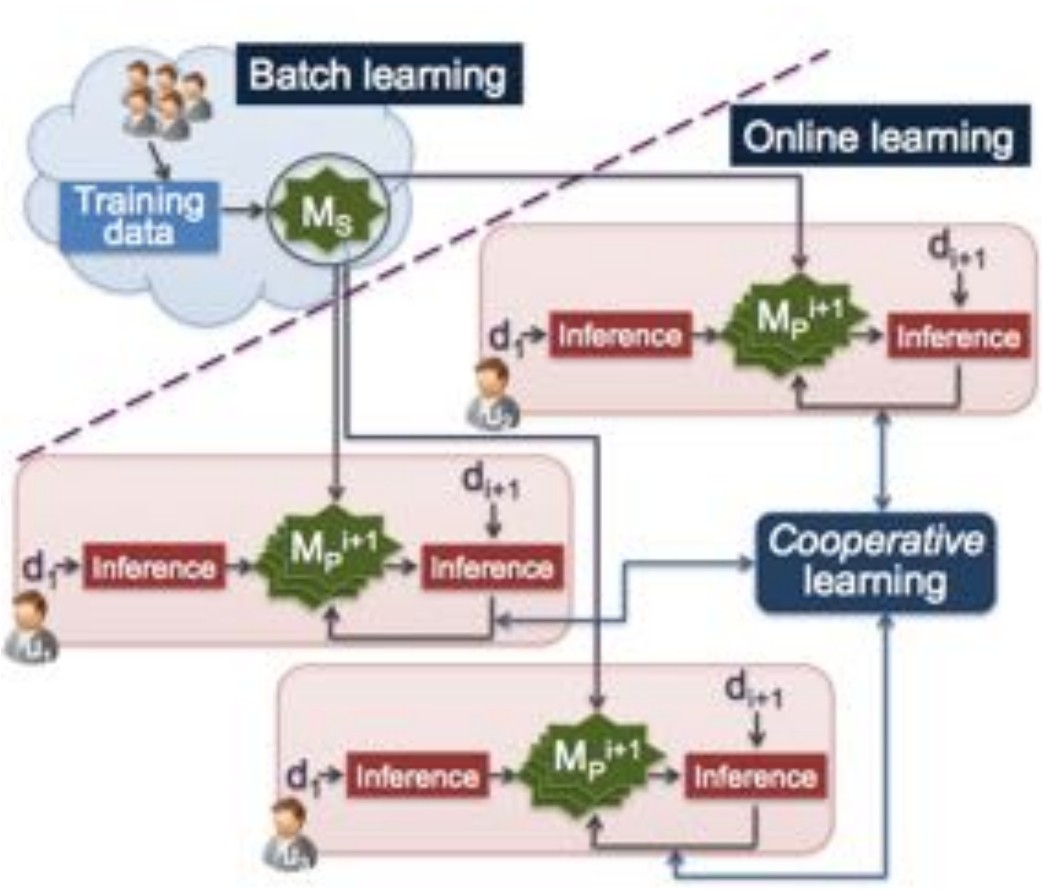
SensingKit System Architecture

Outline

- Introduction & Motivations
- The Databox platform
- Privacy-preserving sensing & analytics

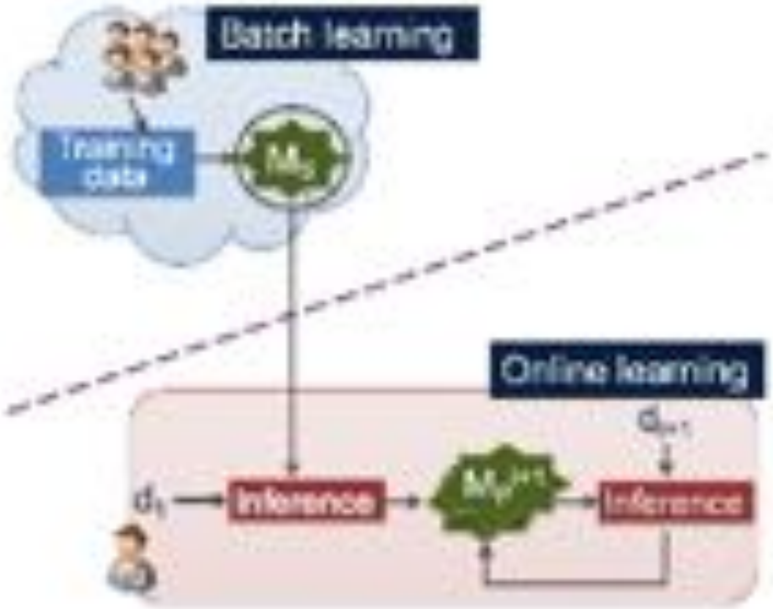
Distributed Analytics

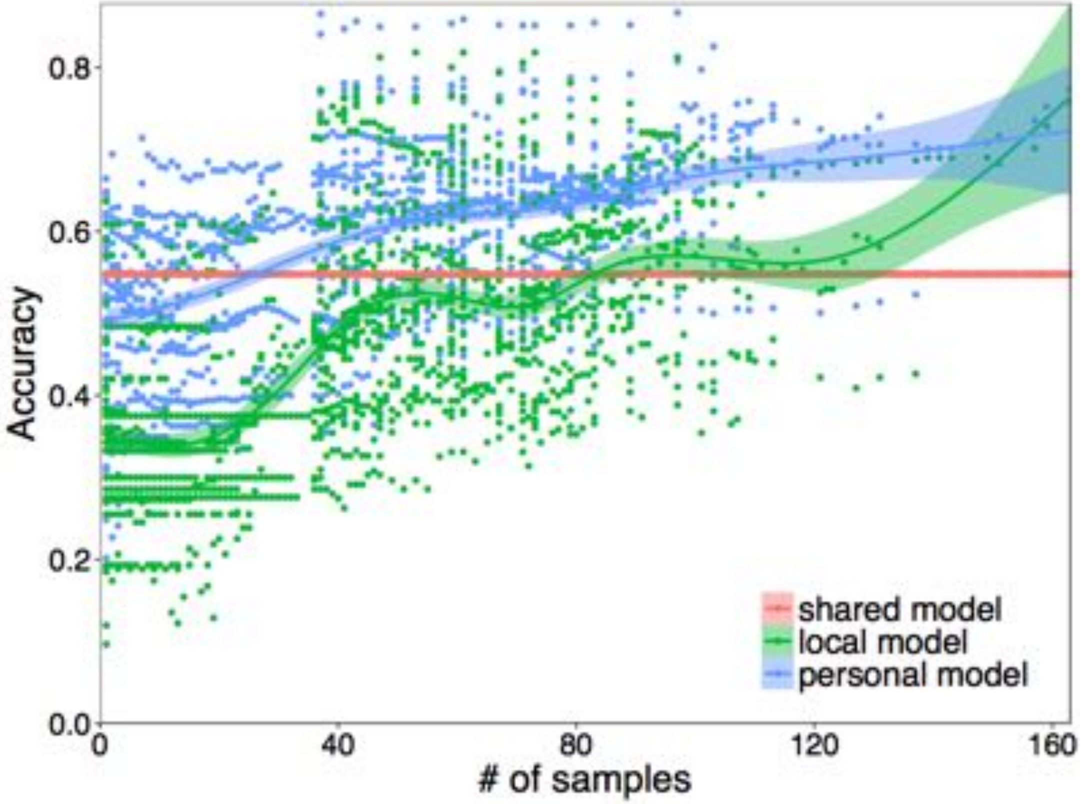
- How to handle scale, heterogeneity, dynamics?
- Subject vs processor driven
  - App stores vs cohort discovery
- Cohort vs individual processing
  - Distributed model building
  - Personal local visualisation



Online Learning

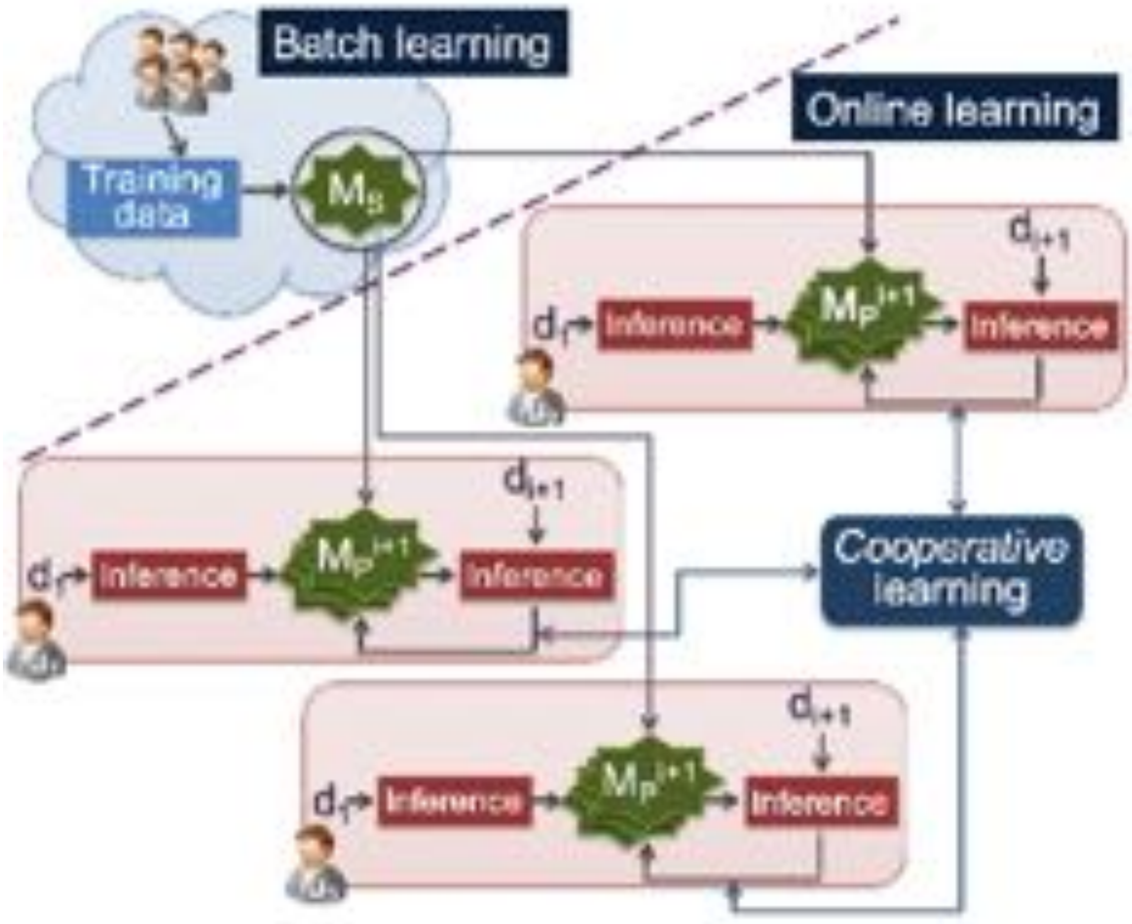
Can we use personal data to improve public, pre-trained ML models?



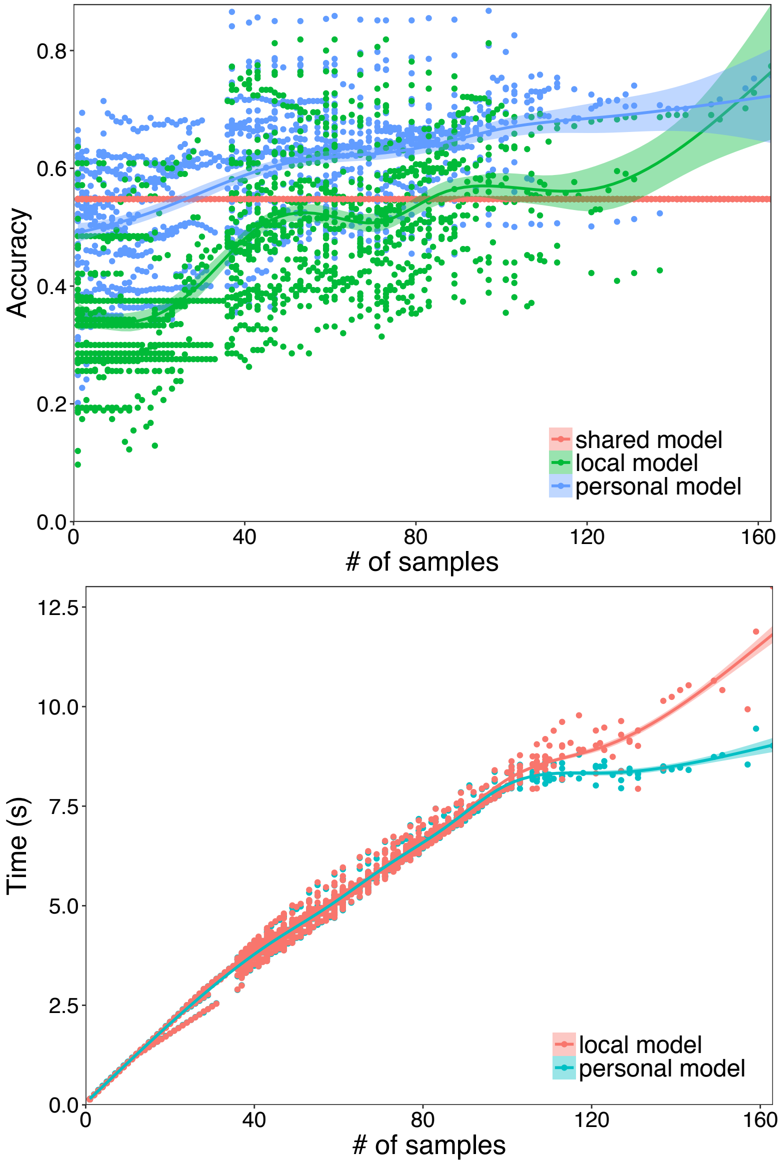


Cooperative Learning

Or train our models cooperatively over distributed users?



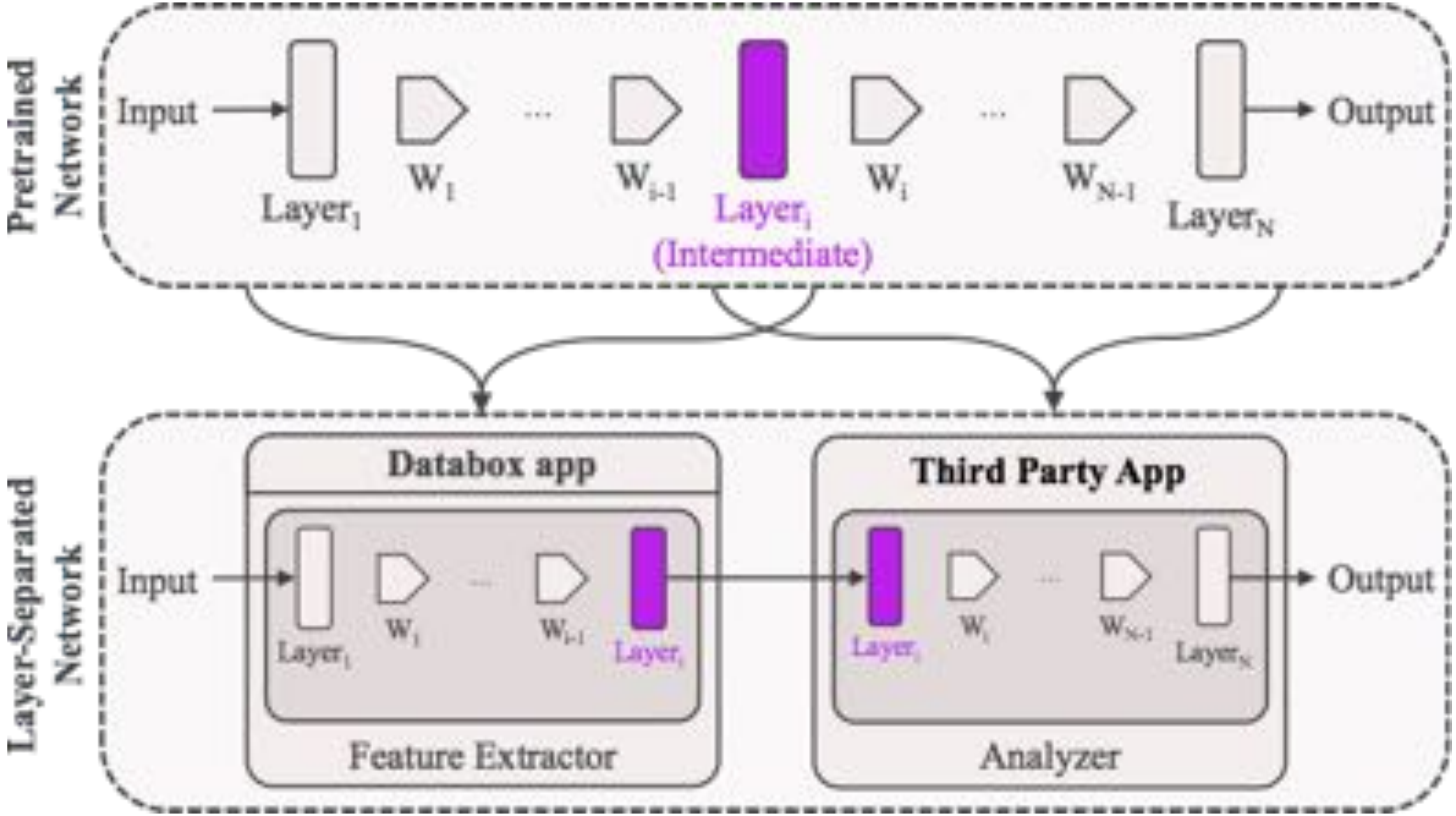
Cooperative learning



“Personal Model Training under Privacy Constraints”, on ArXiv 2017

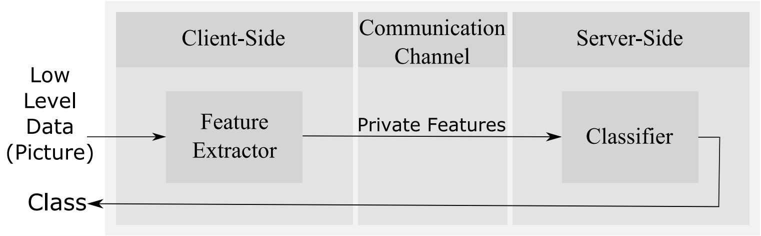
Example: Occupancy-as-a-Service

Privacy-Preserving Analytics

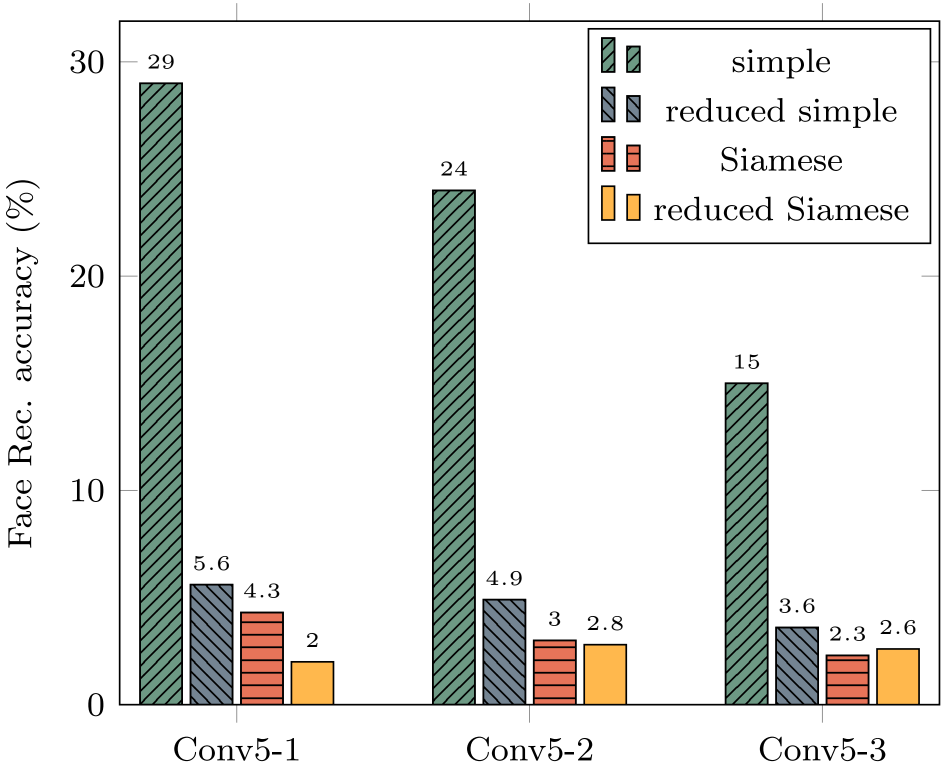


Edge computing paradigm

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

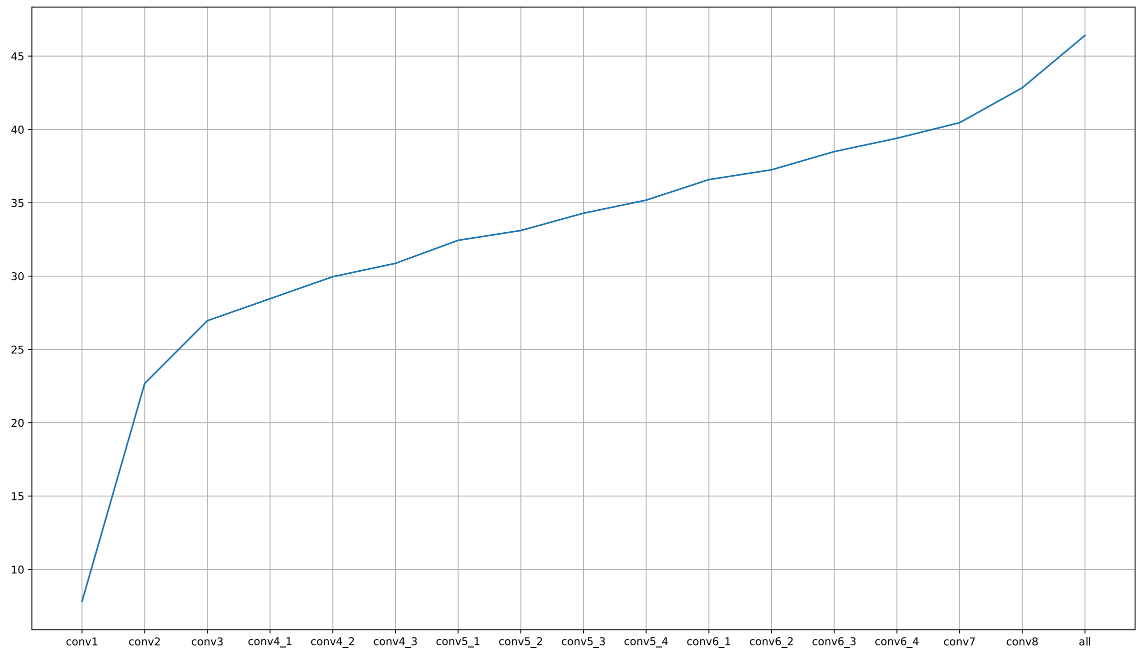


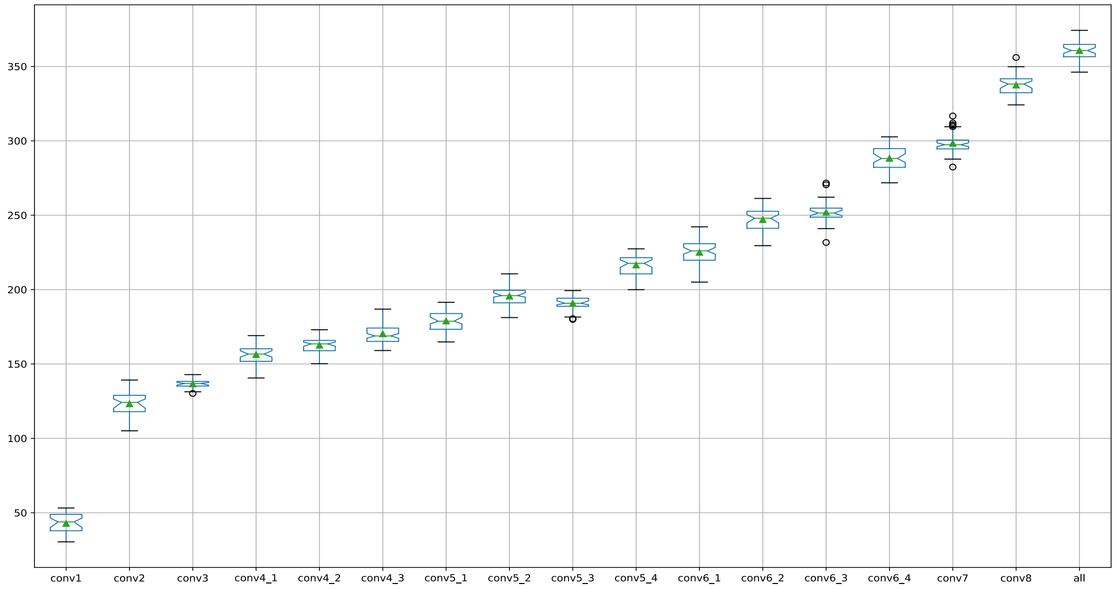
Accuracy on LFW			
	Conv5-1	Conv5-2	Conv5-3
simple	94%	94%	94%
reduced simple	89.7%	87%	94%
Siamese	92.7%	92.7%	93.5%
reduced Siamese	91.3%	92.9%	93.3%



“A Hybrid Deep Learning Architecture for Privacy-Preserving Mobile Analytics” on ArXiv 2017

Mobile Efficiency





Developer Community Engagement

DATABOX

01000100 01100001 01110100 01100001 01100010 01111000

mozilla Festival

www.databoxproject.uk

BBC R&D

Future works

- User-centric sensing and analytics
  - Can a dual approach decrease privacy risk?
    - Large-scale continuous sensing app (multimedia)
    - Understanding contextual requirements
      - See our new paper on ArXiv on this.

- Enabling in-the-wild capabilities for the Databox
  - User and developer Community will be a key part
  - In-house Platform for longitudinal social and experimental studies with real data
  - Providing a home DMZ through the Databox….

CPS Security and Privacy

- Security and Privacy dichotomy
  - Scare stories: Mirai IoT Botnet, Smart TVs transmitting conversations & profiling, CIA Hacks, Webcam viewing websites, spamming fridge, Amazon echo ordering dolls, eavesdropping teddy bears…
- IoT device and Network Isolation
  - limit coordinated attacks
- Crowdsourced or semi-supervised policing & anomaly detection
- Can not rely on constant connectivity
  - Is the “cloud” or your DSL connection always *online*?
  - Remember Amazon AWS outage?

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