

# On the Edge of Human-Data Interaction with the

DATABOX

01000100 01100001 01110100 01100001 01100010 01111000

Richard Mortier



# Living in a Big Data World



- Challenges and Opportunities
  - Who's tracking us, to what end?
  - Personalisation, Internet of Things
- Digital Footprints
  - Intimate information in large, rich data silos
  - Never forgets or forgives



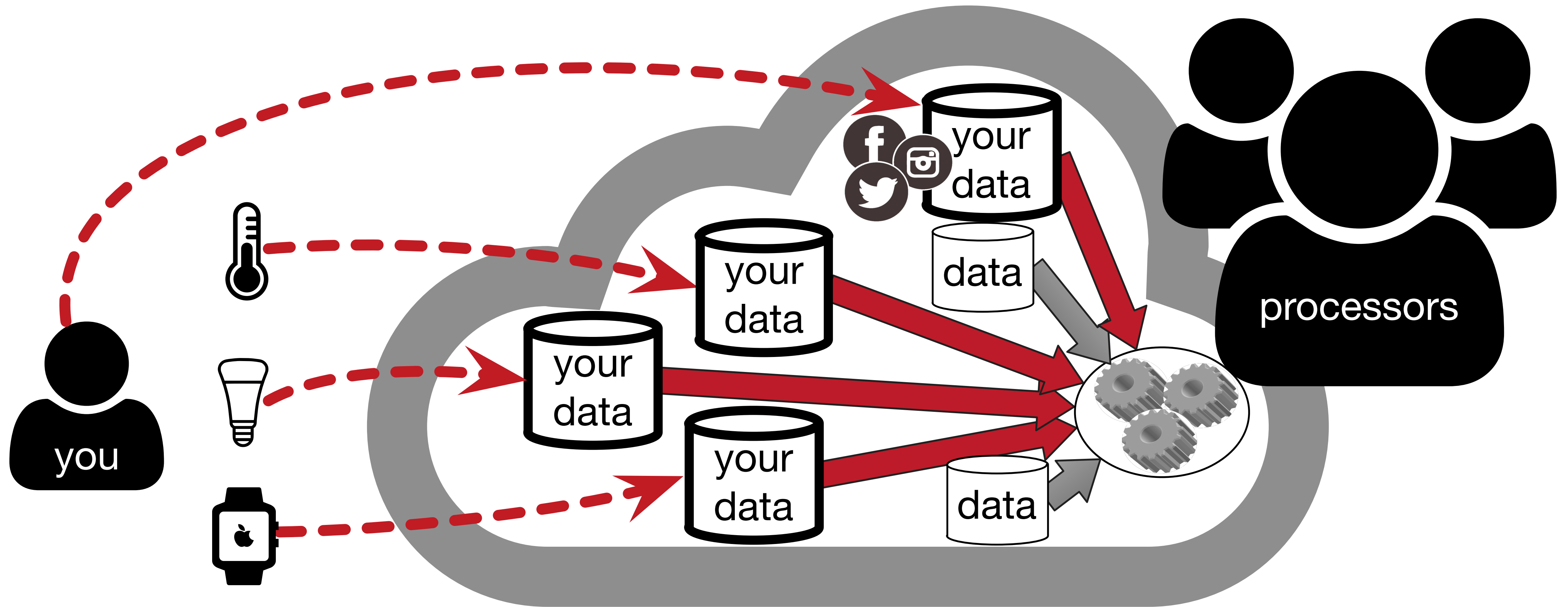
Key Challenge:

**How do we enable data subjects to control collection and exploitation of both *their data* and *data about them*?**

<http://bigdatapix.tumblr.com/> "Big Data is visualized in so many ways... all of them blue and with numbers and lens flare."

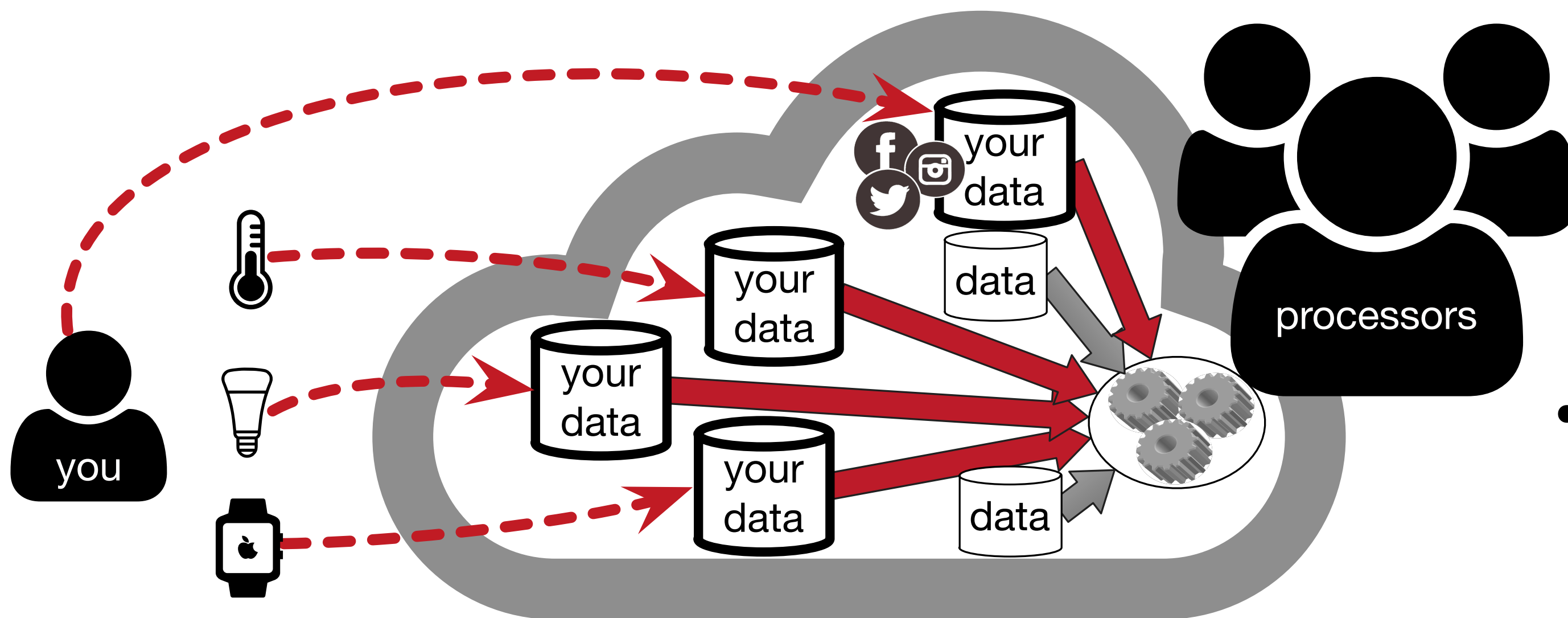
<http://weputachipinit.tumblr.com/> "It was just a dumb thing. Then we put a chip in it. Now it's a smart thing."

# Existing Ecosystem: Move Data



# A Structural Problem?

- The Internet is fragmented, distributed systems are difficult
  - Centralising simplifies things
  - With the cloud, we can, so we do!



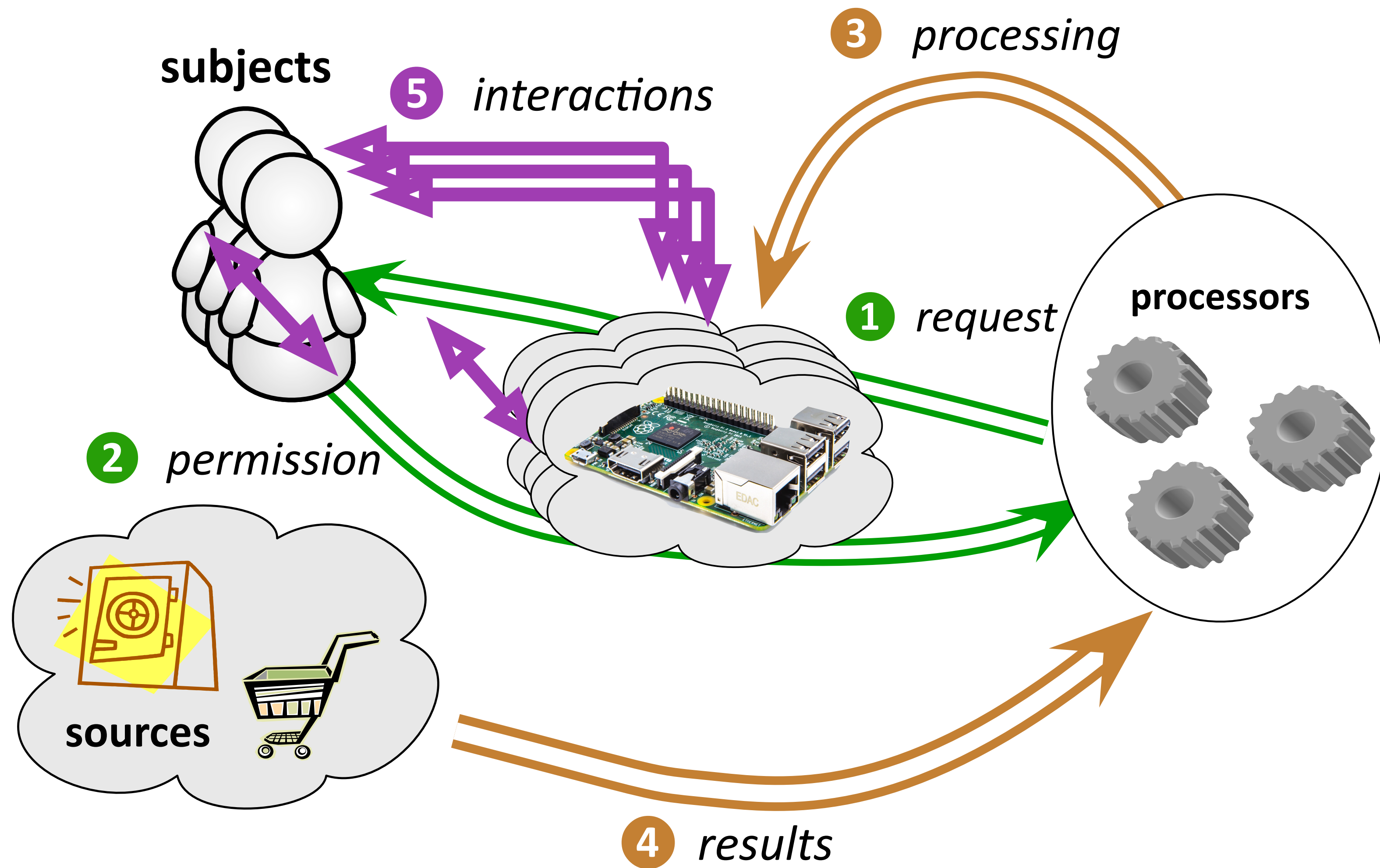
<https://www.stickermule.com/marketplace/3442-there-is-no-cloud>

- Ease of cloud computing means, by default, **we move data to the cloud** for processing

# Restructuring the Problem

- Horizon Digital Economy Research, Nottingham, UK ~2009
  - [*Them*] Build us a Magic Context Service! [*Me*] WTF even is that?!
  - No-one could explain, but it definitely involved using personal data
- I'm a lazy computer scientist so I punted on the hard problems
  - I don't know what you want when you say you want context
  - But if you give me some program that encodes what you want, I'll run it for you
- **Dataware** — effectively a service-oriented architecture for personal data processing
  - Data Processor writes some code to process the Data Subject's data
  - Subject provides the platform on which to run that code
  - Processor gets the result
- Key: **Move code to data, not data to code**

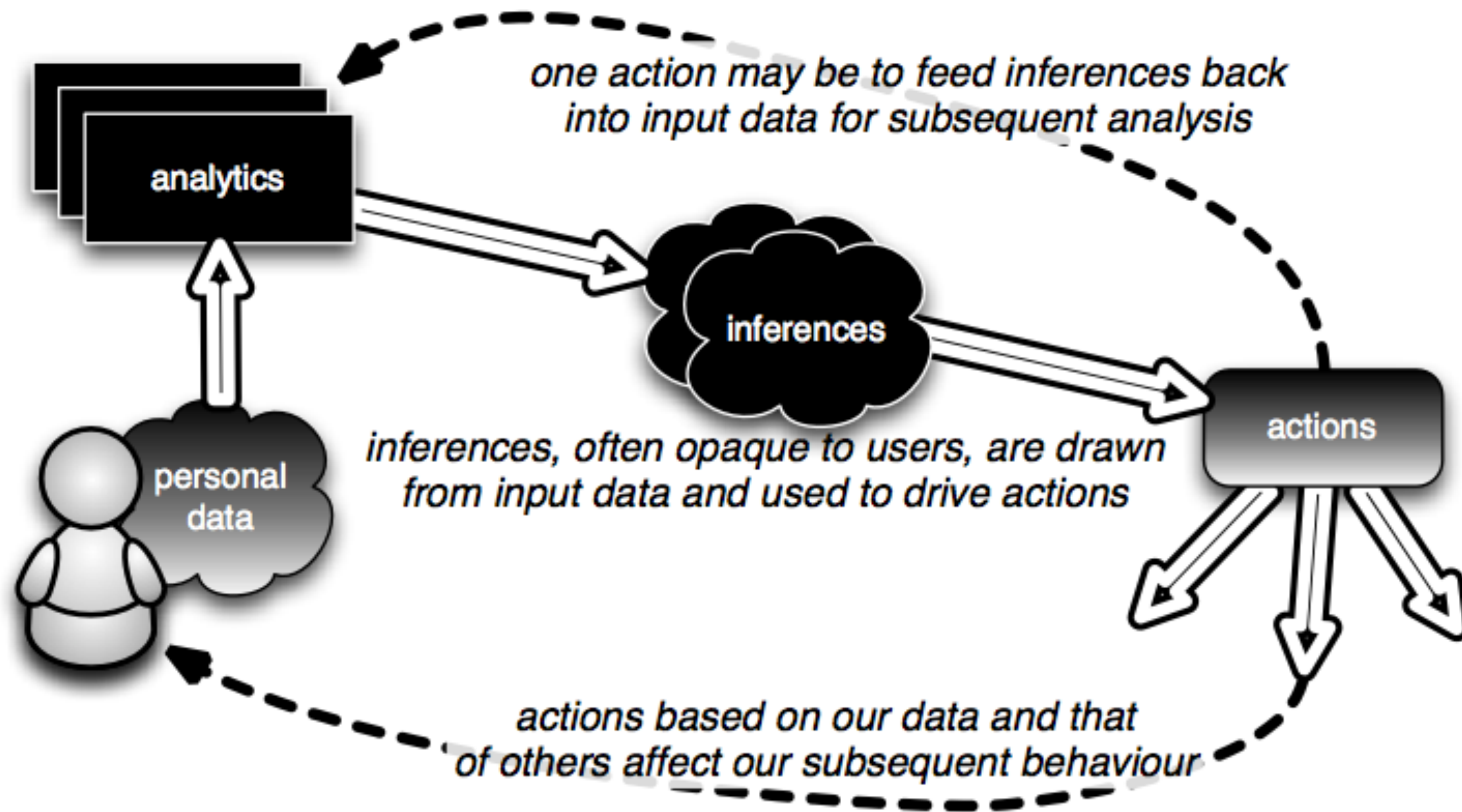
# Dataware



# Constructing Interaction

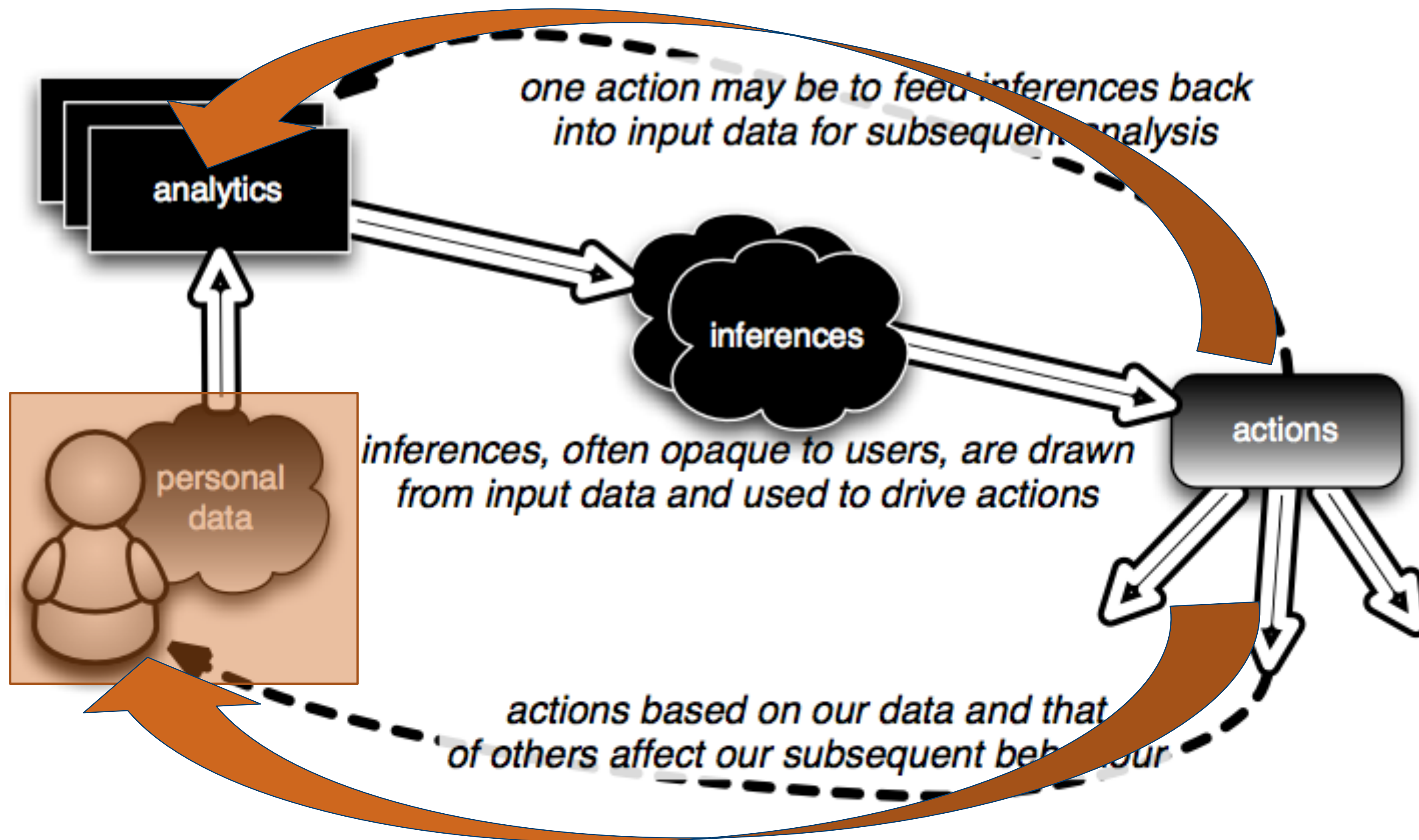
- Many proposed interaction models
  - E.g., pay-per-use
- Little about how to actually provide for it
  - E.g., Exactly what am I being paid for?
- Dataware was a technical proposal supporting some forms of interaction
  - Accountable transaction between parties in terms of request, permission, audit
- But there's a lot more to consider here...

# Human-Data Interaction





# Human-Data Interaction

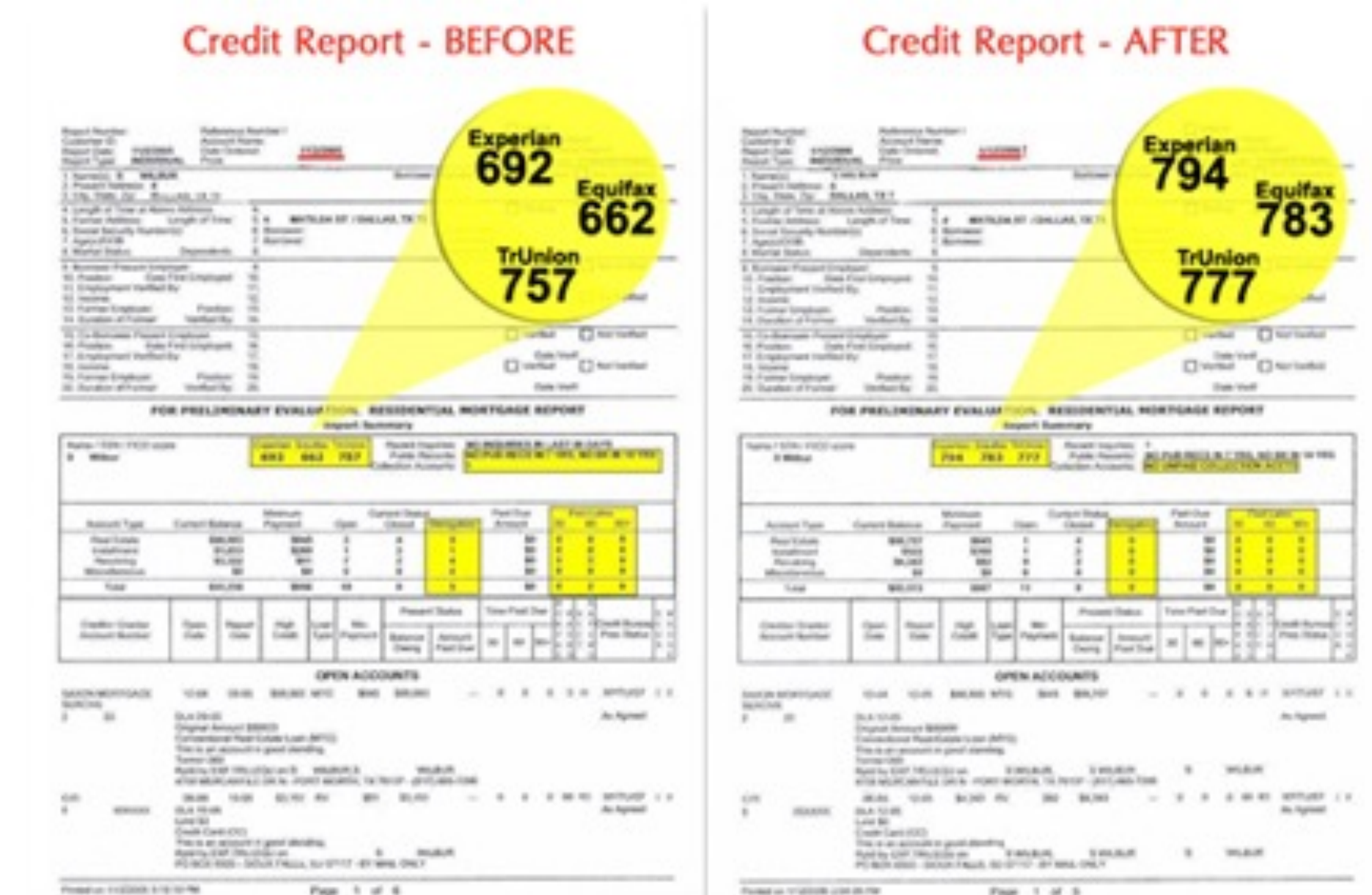


- Data is collected
- Analytics to process data
- Inferences are drawn
- Actions taken as a result

# Lack of Legibility

## *Visualisation & comprehension*

- We are generally unaware of
  - the many **sources of data** collected about us,
  - the **analyses performed** on this data, and
  - the **implications** of these analyses



<https://flic.kr/p/6thmfN>

E.g., Computation of credit scores

# Lack of Agency

## *Capacity to act*

- We are generally unaware of
  - the means we have to **affect data collection**,
  - the means we have to **affect data analysis**,
  - if they even exist, and we know enough to want to employ them

<http://appadvice.com/appnn/2012/04/facebooks-acquisition-of-instagram-just-another-question-mark-for-internet-privacy>

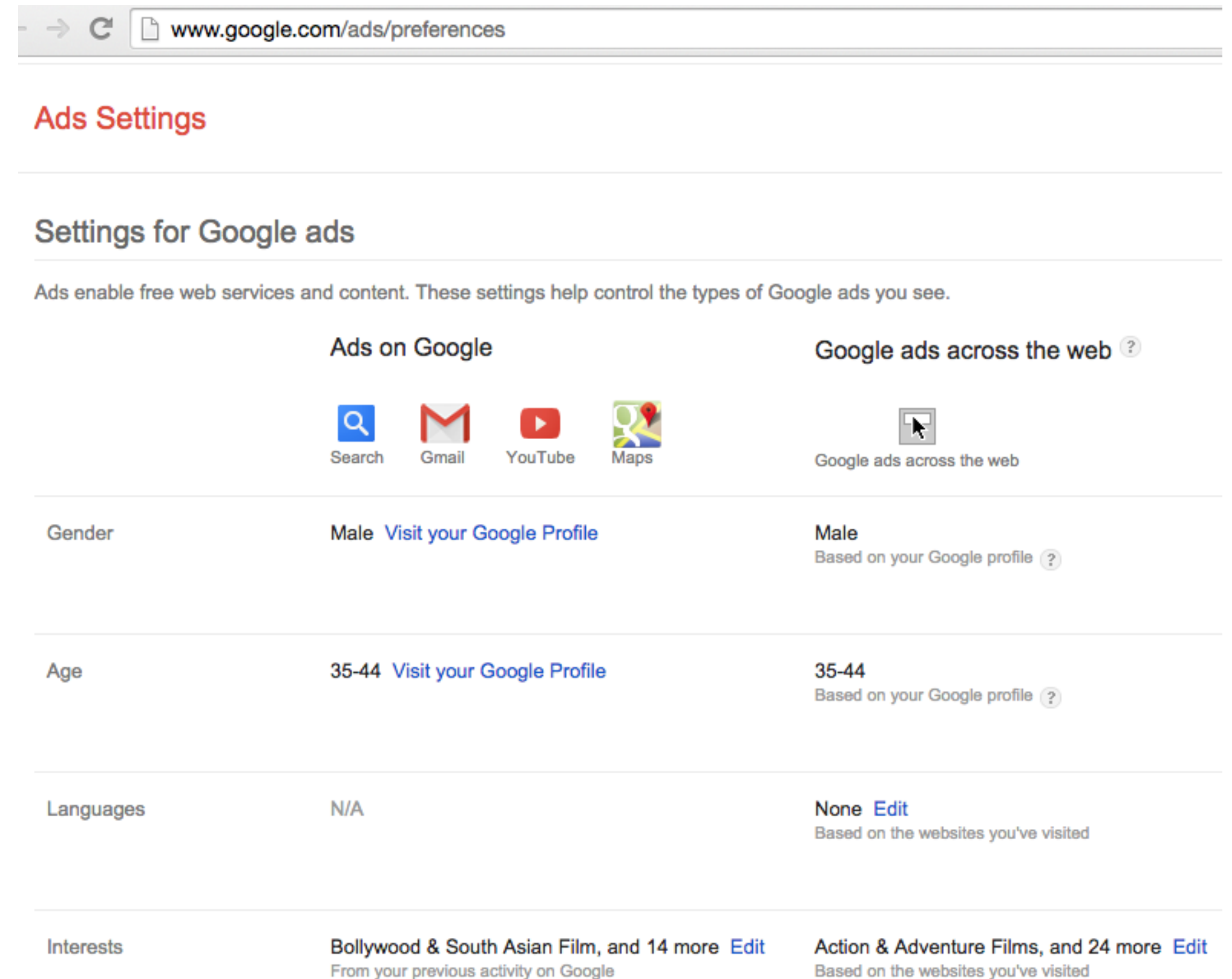


E.g., Use of retail data to profile your propensity to risk for sale to an insurance agency

# Lack of Negotiability

## *Support for dynamics of interaction*

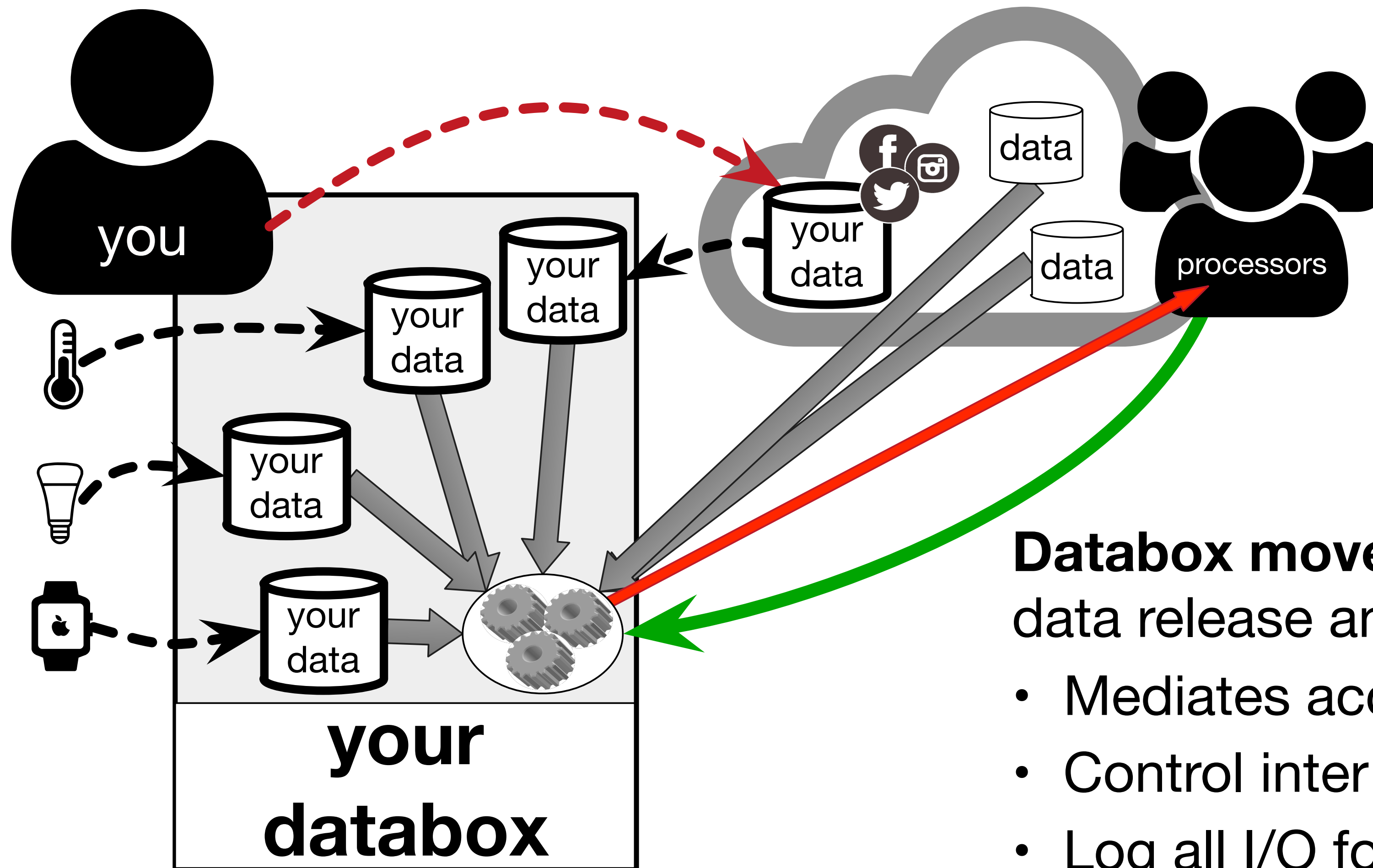
- Even if we know the data collected and analysed about us, and understand how to enact choices over these
- We're **still trapped** by current systems and services
  - Binary accept/reject of terms
  - Cannot subsequently modify or refine our decisions



The screenshot shows the Google Ads Preferences page at [www.google.com/ads/preferences](http://www.google.com/ads/preferences). The page is titled "Ads Settings" and "Settings for Google ads". It contains a table of settings for Google ads, including "Ads on Google", "Google ads across the web", "Gender", "Age", "Languages", and "Interests".

Setting	Value	Source
Ads on Google	Search, Gmail, YouTube, Maps	Based on your Google profile
Google ads across the web	Google ads across the web	Based on your Google profile
Gender	Male	Based on your Google profile
Age	35-44	Based on your Google profile
Languages	N/A	Based on the websites you've visited
Interests	Bollywood & South Asian Film, and 14 more	Based on the websites you've visited

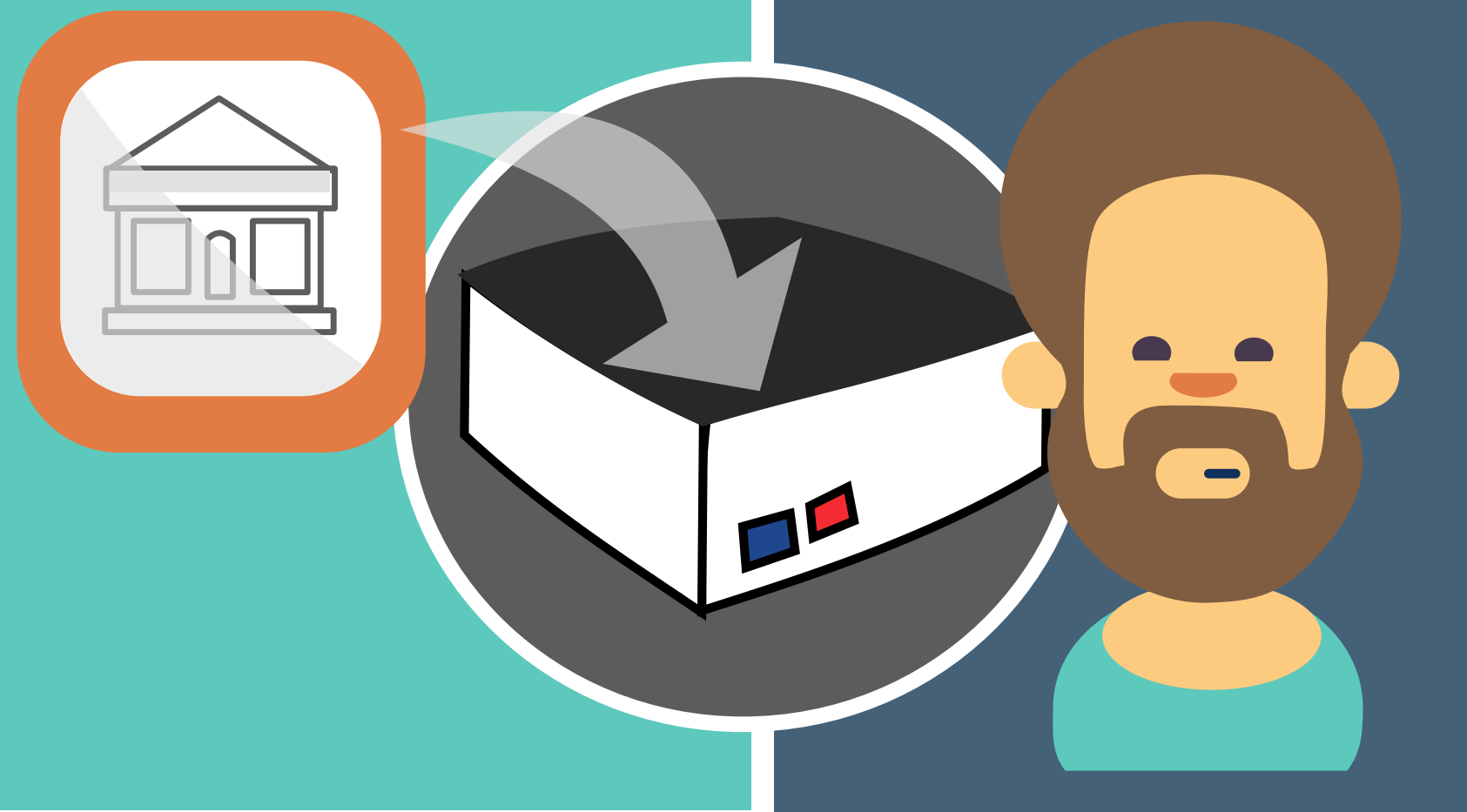
# Databox: Dataware v2



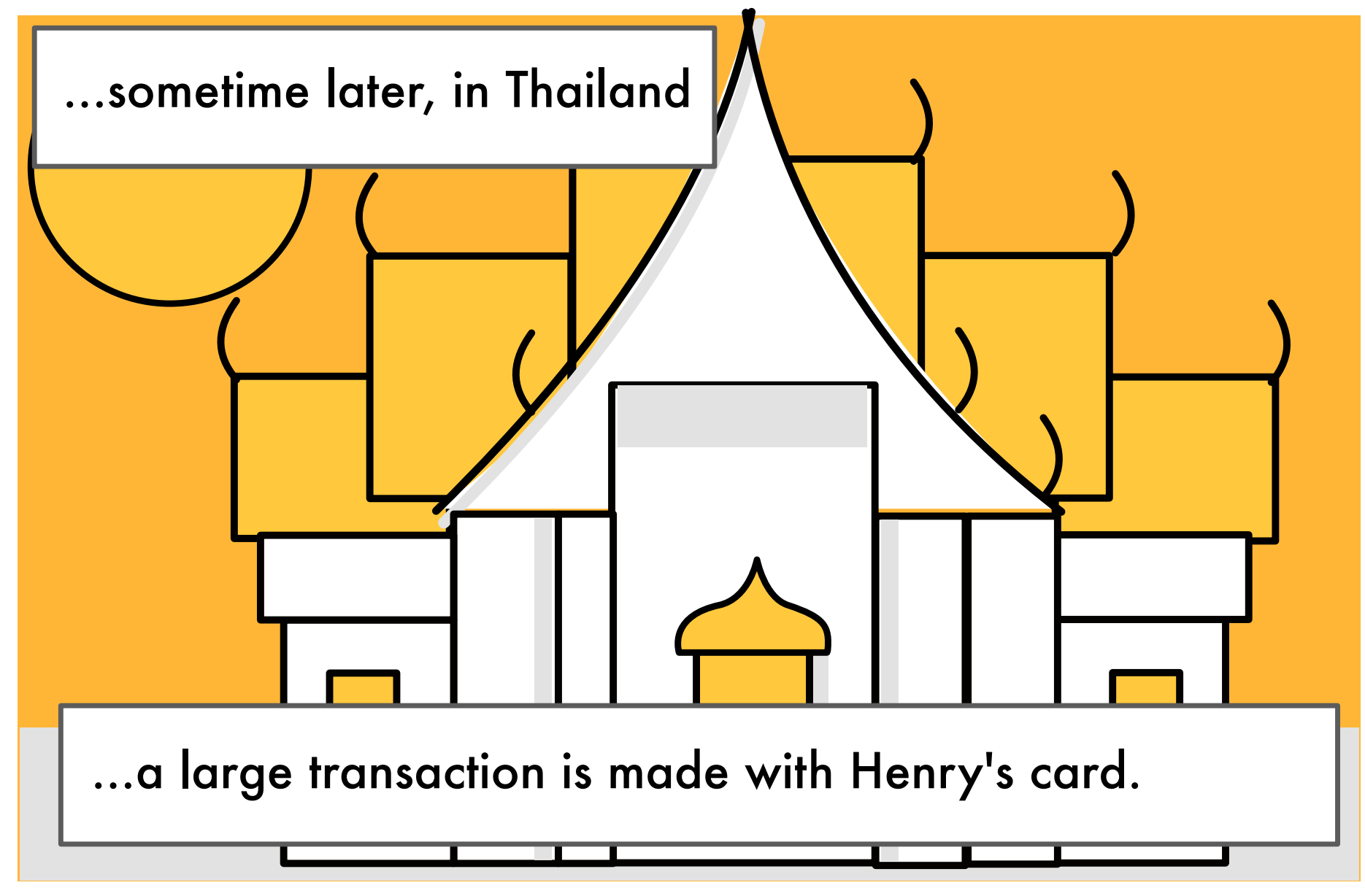
**Databox moves code to the data**, minimising data release and retaining control over processing

- Mediates access to data, local or remote
- Control internal and external communications
- Log all I/O for users to inspect, control

Henry downloads his bank's app onto his databox.

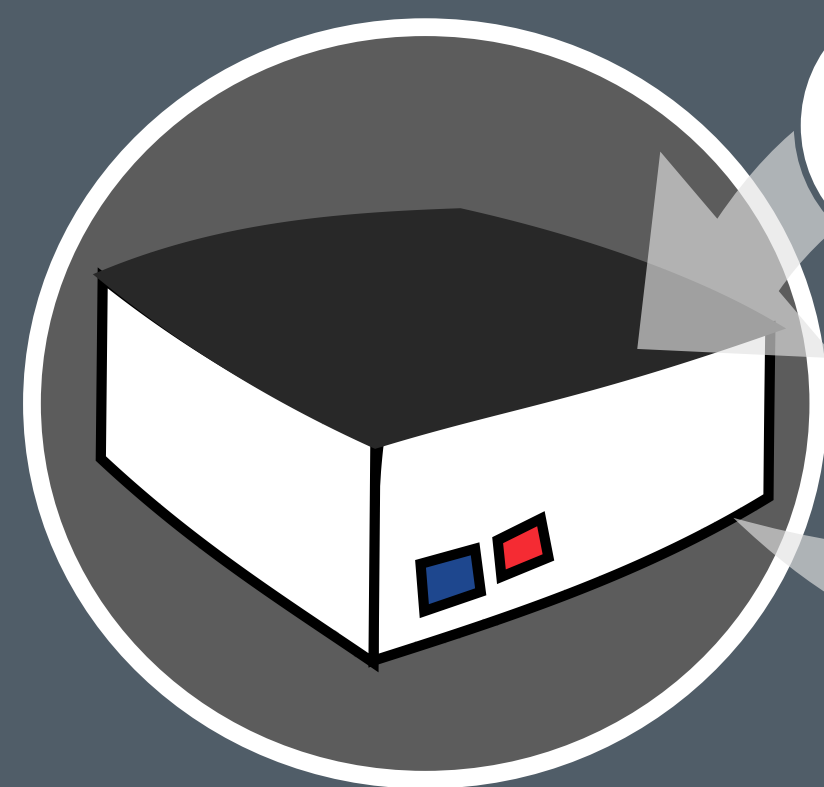


...sometime later, in Thailand



# DATABOX FRAUD DETECTION

Henry's banking app checks his location.



is Henry in Thailand?

NO



and tells the Bank Henry is NOT in Thailand.

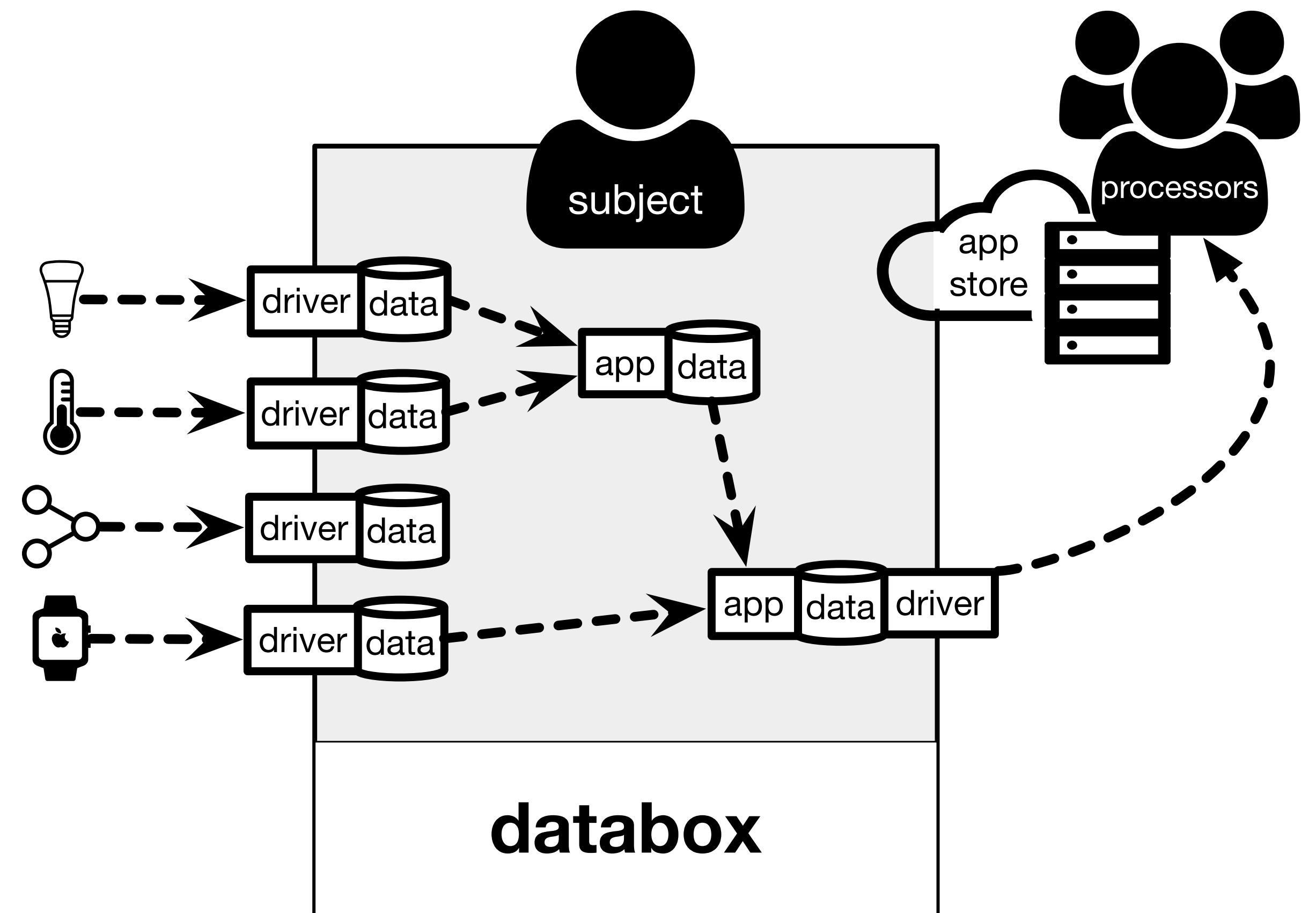
The transaction is refused.



Henry is happy. So is his bank manager.

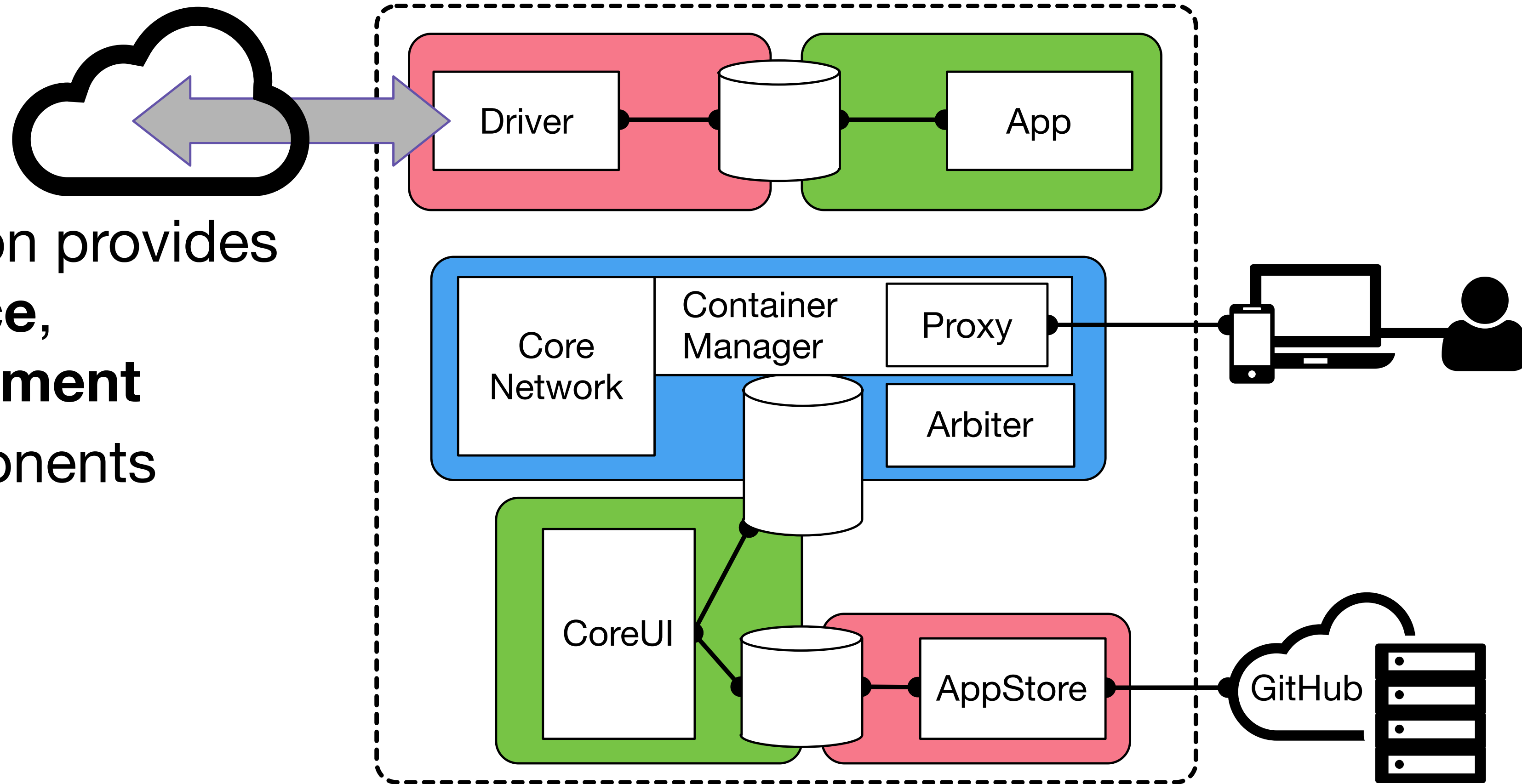
# Databox: Move Contained Code!

- Install **apps** to process data locally
- Ingest/release data via **drivers**
- App **manifests** describe data they will access,
  - ...when made into concrete **SLAs** on installation



# Databox Platform

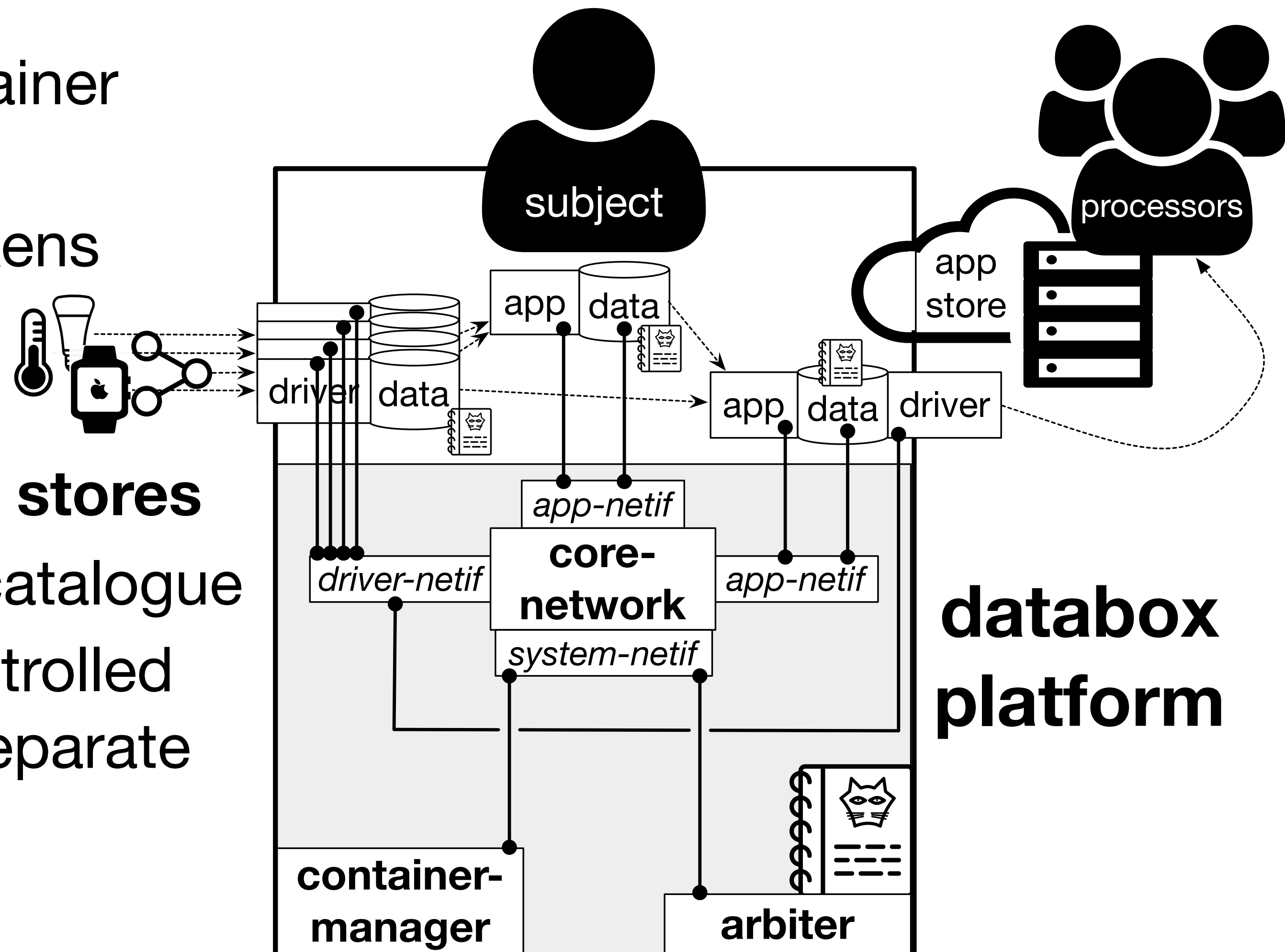
- All components are **Docker containers**
  - Lightweight virtualisation provides **platform independence, isolation, and management**
- Four core platform components
  - Container Manager
  - Arbiter
  - Core Network
  - Data store(s)





# Databox Platform

- **Container Manager** manages container lifecycle
- **Arbiter** manages access control tokens
- Persistent storage and 0MQ-based **middleware** layer via provided **data stores**
- Data stores registered in **hypercat** catalogue
- Inter-container communications controlled by **core-network** interconnecting separate virtual interfaces

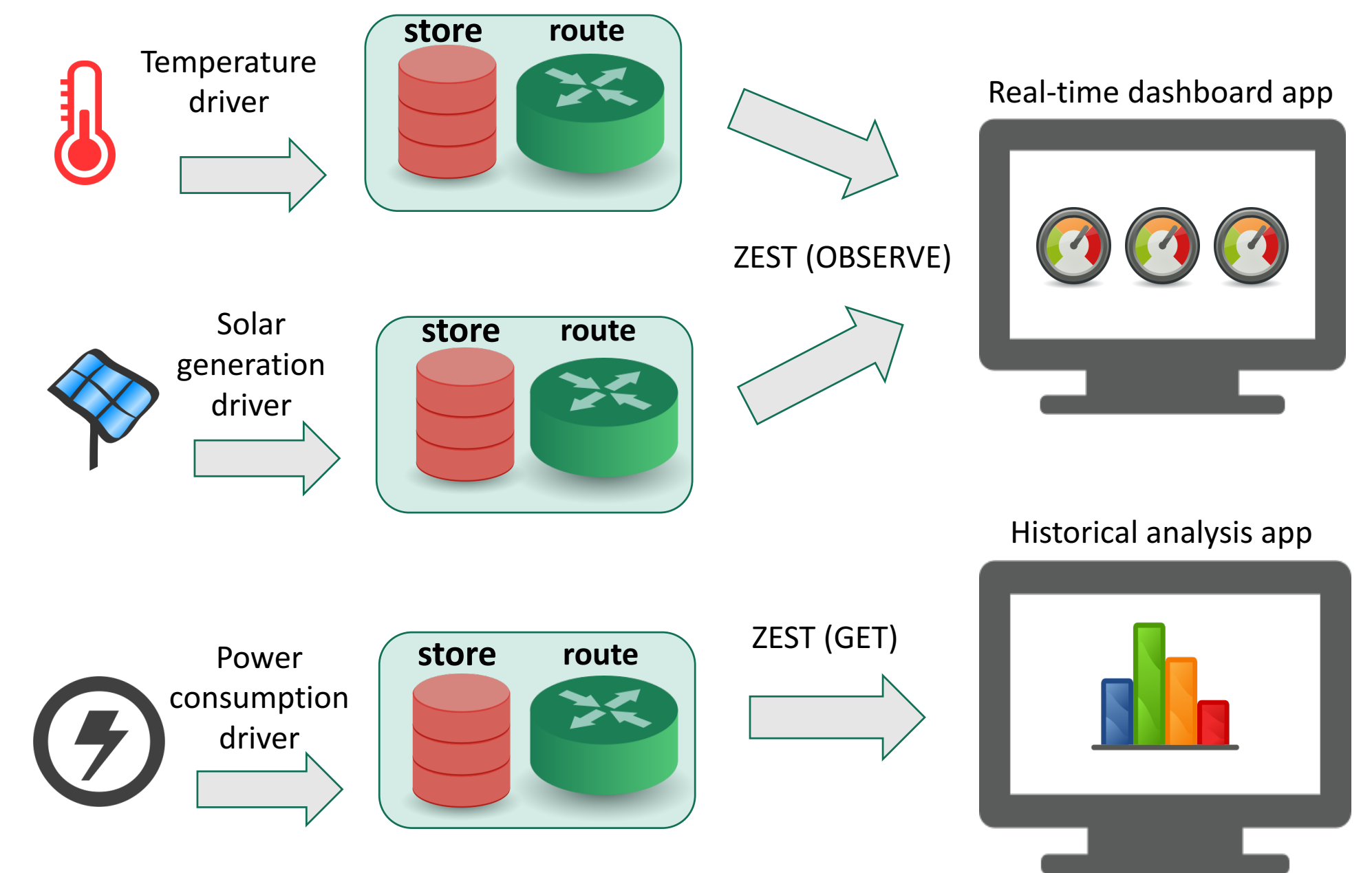


# Container Lifecycle

- Apps and drivers come with a **Manifest**, covering
  - origination metadata,
  - data access and storage requirements,
  - remote access requirements
- **Installation**
  - user input realises manifest as a **Service Level Agreement**,
  - obtains access tokens (*macaroons*) from the **Arbiter**,
  - creates a per-app bridge and configures connectivity via **Core Network**,
  - starts the app/driver's containers, including a **Store**

# Accessing Data Stores with Zest

- Originally simple HTTP/REST API
  - Unsuitable to high-frequency sensor data
  - Memory footprint unsuitable to rPi
- Zest: CoAP over 0MQ
  - RESTful-like, key-value and timeseries retrieval controlled by *macaroon*s
  - Irmin (git-like) backend supporting JSON, text, binary data
  - Encryption via CurveZMQ, integration with HyperCat
  - About half the CPU load and memory footprint of HTTP solution
- Audit logging



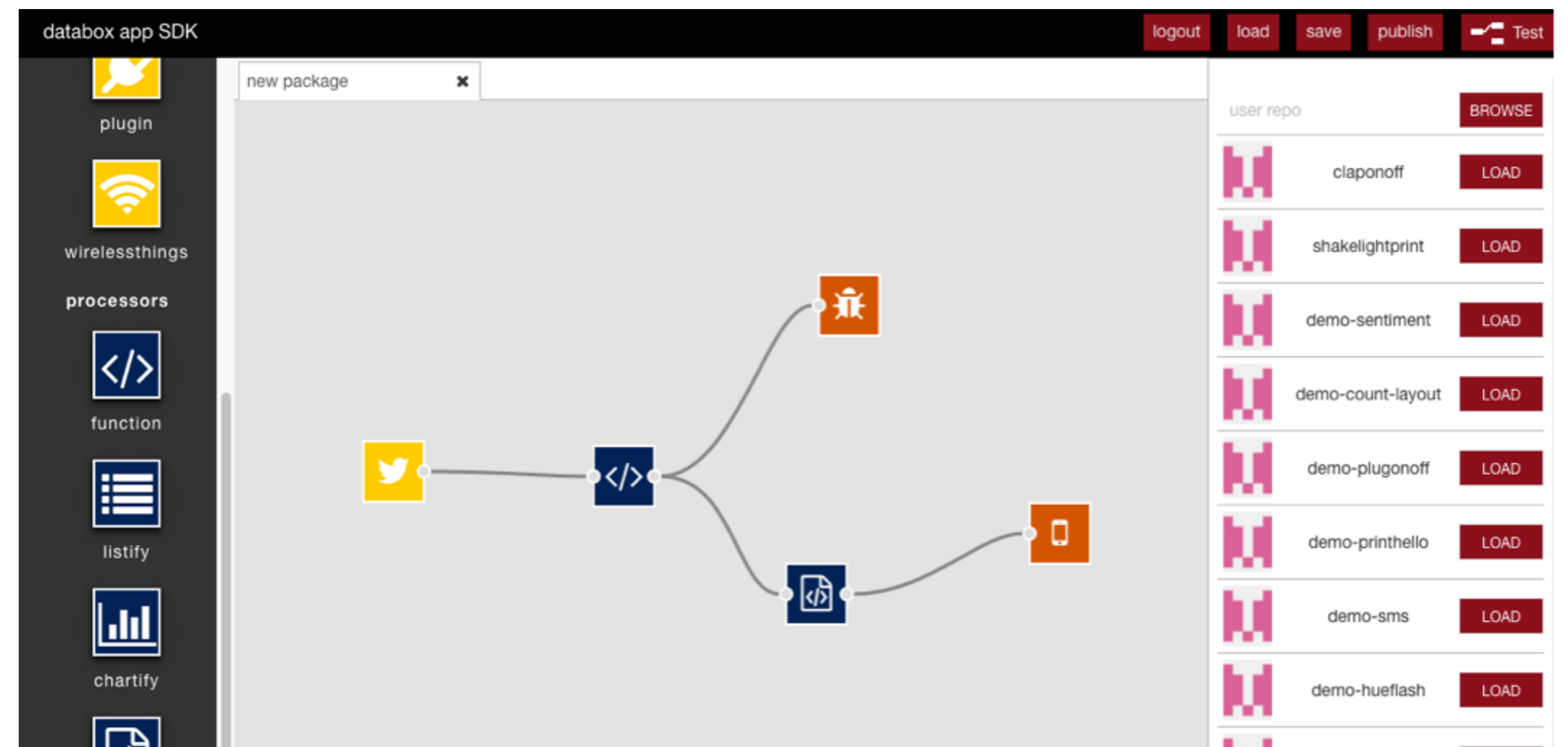
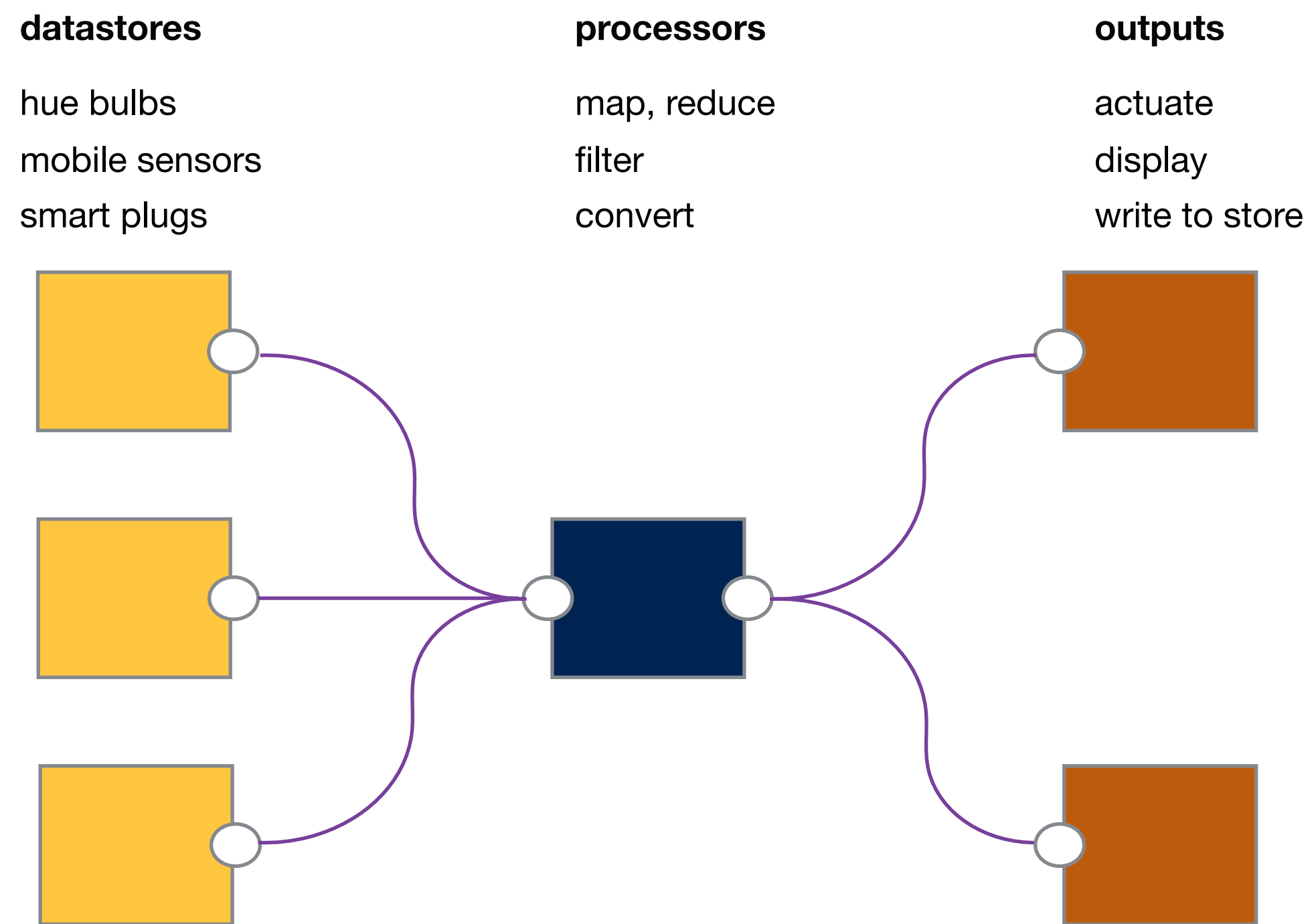
# Enabling Physical Interactivity

- Physical devices often easier to reason about
  - Visible; Located; Proximate; Portable
  - Physical access control (“bag of keys”) is widely understood
- For example,
  - “access to our smart meter data allowed only if a green tag is in my Databox and in my partner’s Databox, or when the green tag is in one Databox and we’re both in the house”
- Alternatively, **physical interactions** providing for **virtual connectivity**



# Democratising App Development

- Install and connect existing apps
- Plug together apps and components to customise **your** apps

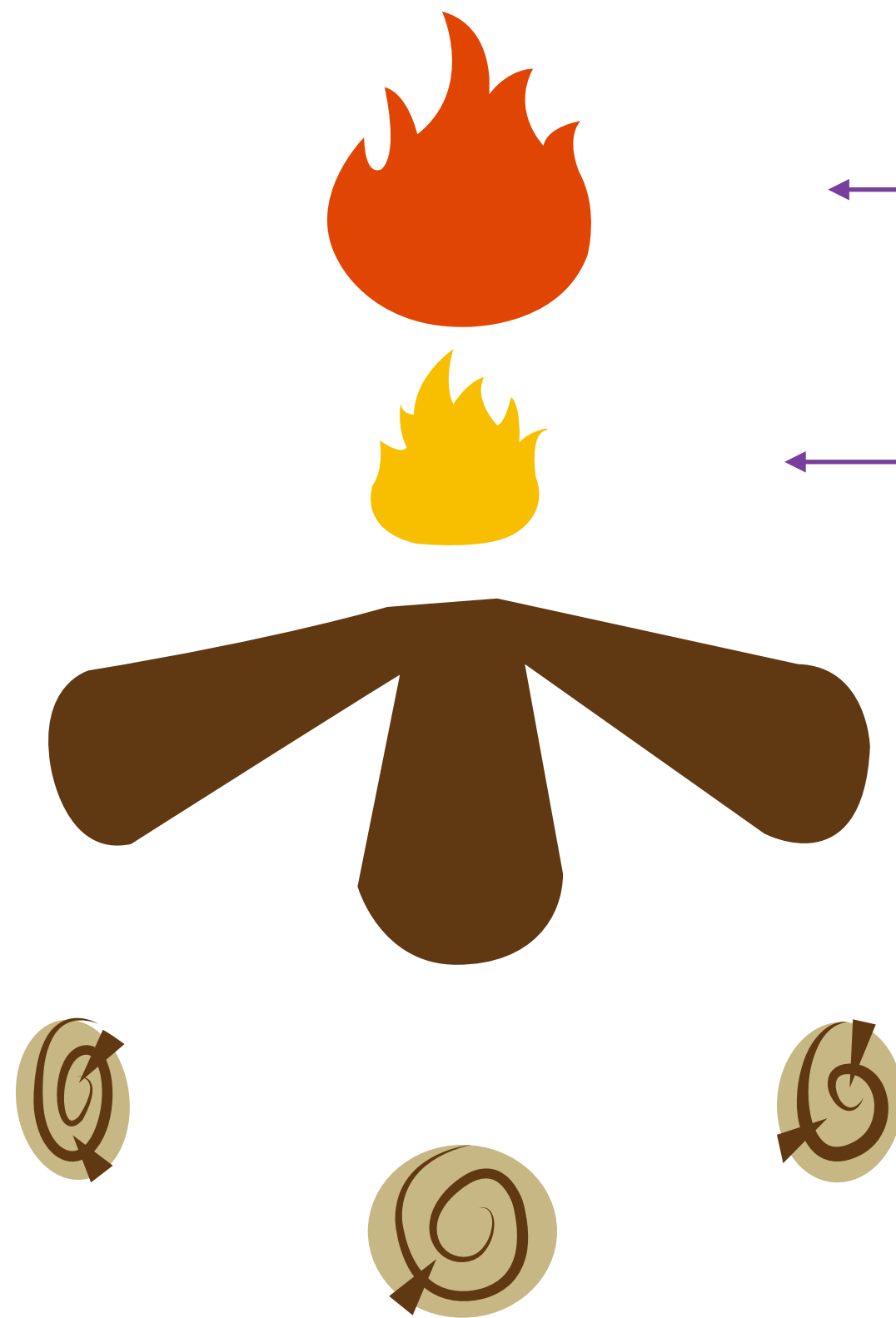


# Rich Visualisations of Rich Data

svg image



image parts



transform

rotate  $x$  degrees

scale by  $y/2$

fill with colour  $z$

translate to  $(i,j)$

data

$x$

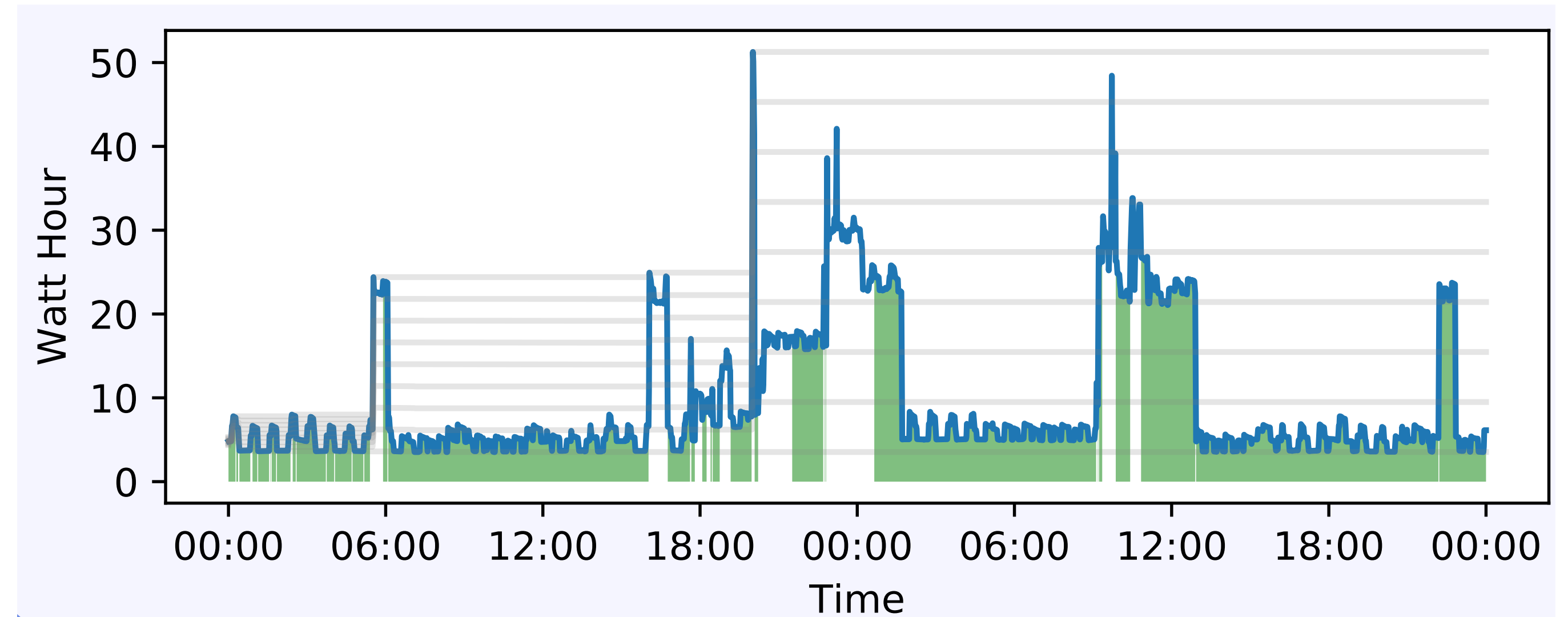
$y$

$z$

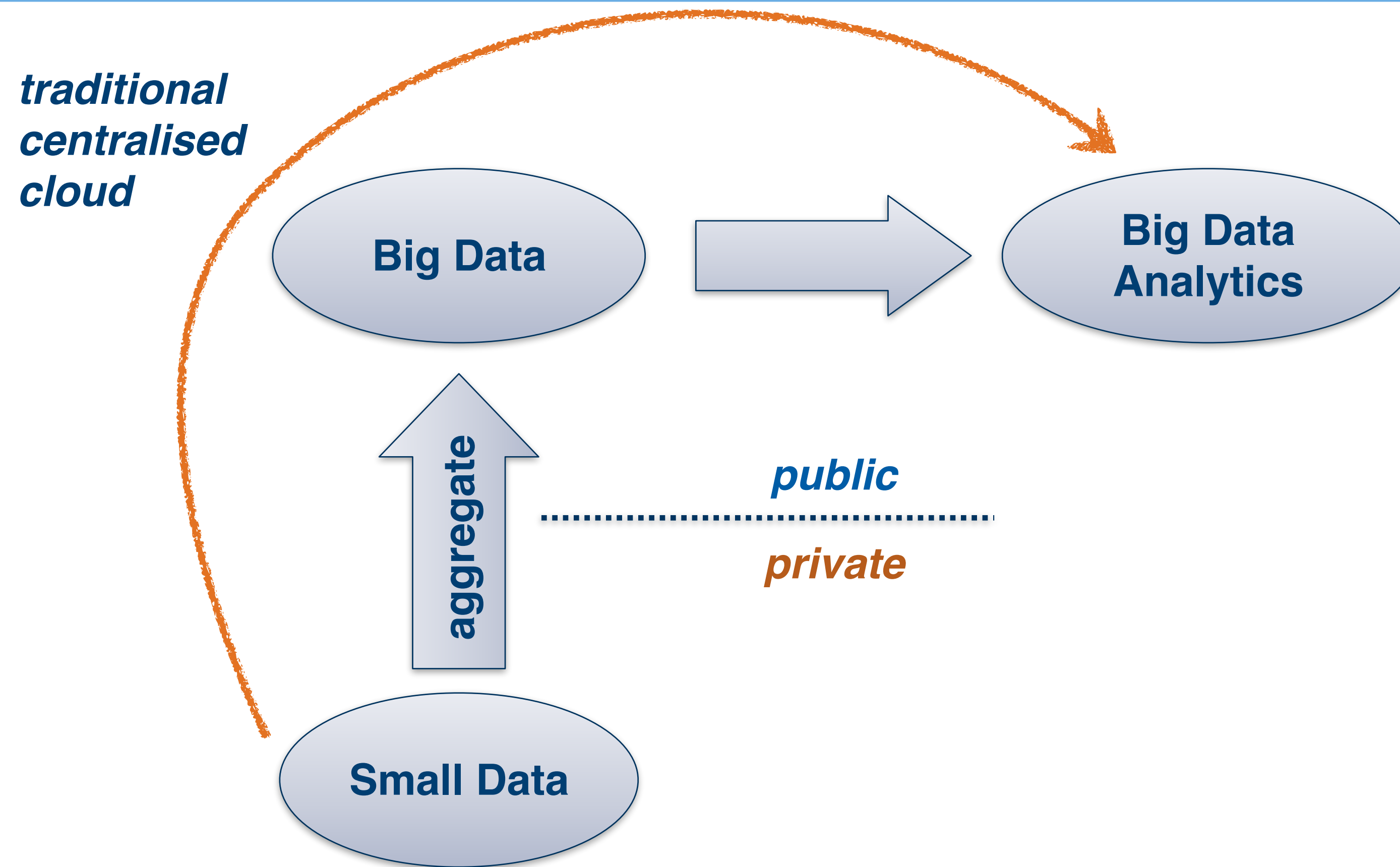
$(i,j)$

# Privacy-Informed Access Control

- Access control through tokens (*macaroons*)
  - minted by the Arbiter,
  - verified by a Store,
  - apply to URI paths
- Exploring generic measures of privacy risk, e.g.,
  - (entropic) surprisal,
  - (statistical) autocorrelation,
  - (similarity)  $k$ -anonymity,  $l$ -diversity,  $t$ -closeness
  - Dynamic determination of risky access
  - Static analysis of overall configuration risk

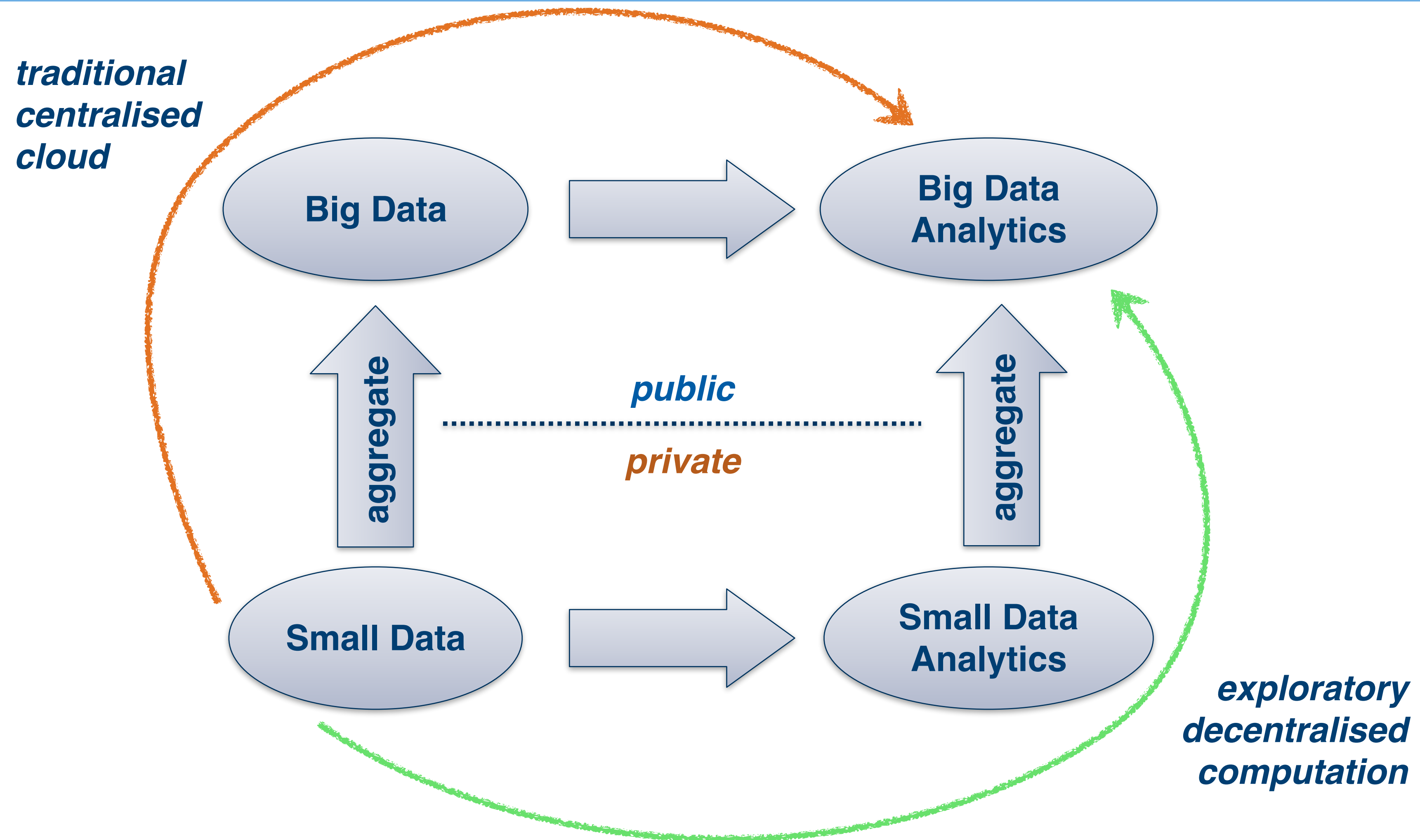


# Big Data Analytics?



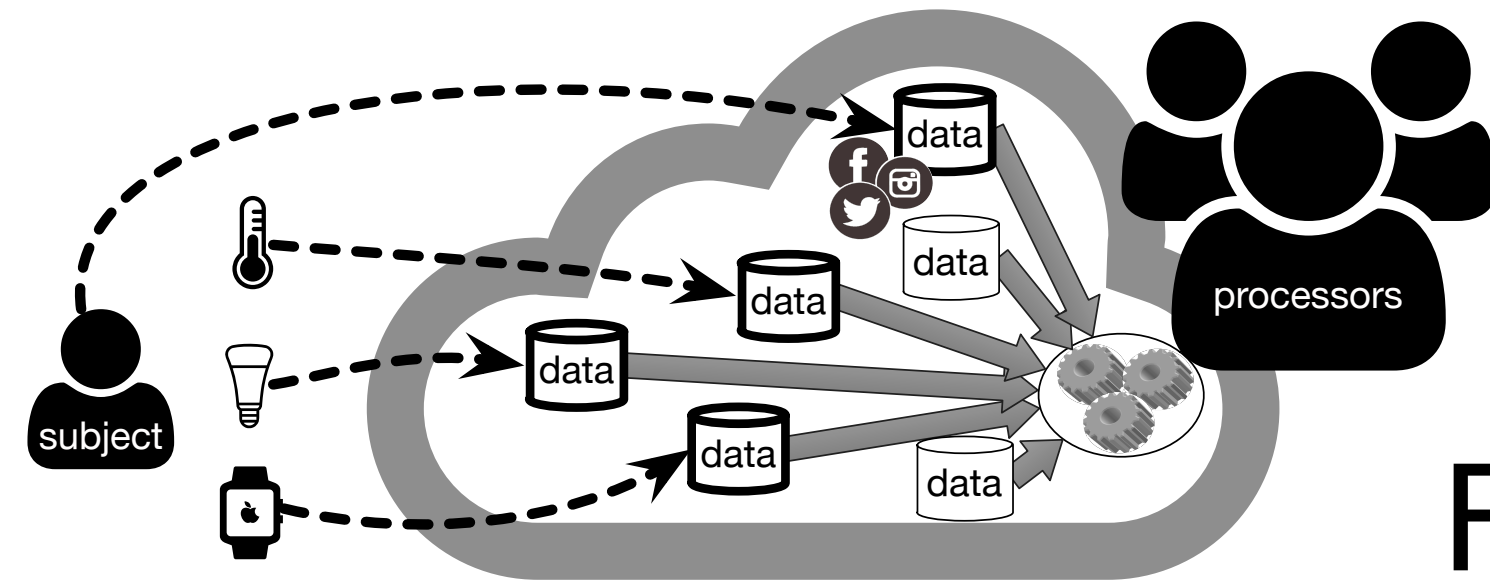


# Big Data Analytics? Small Data Analytics!

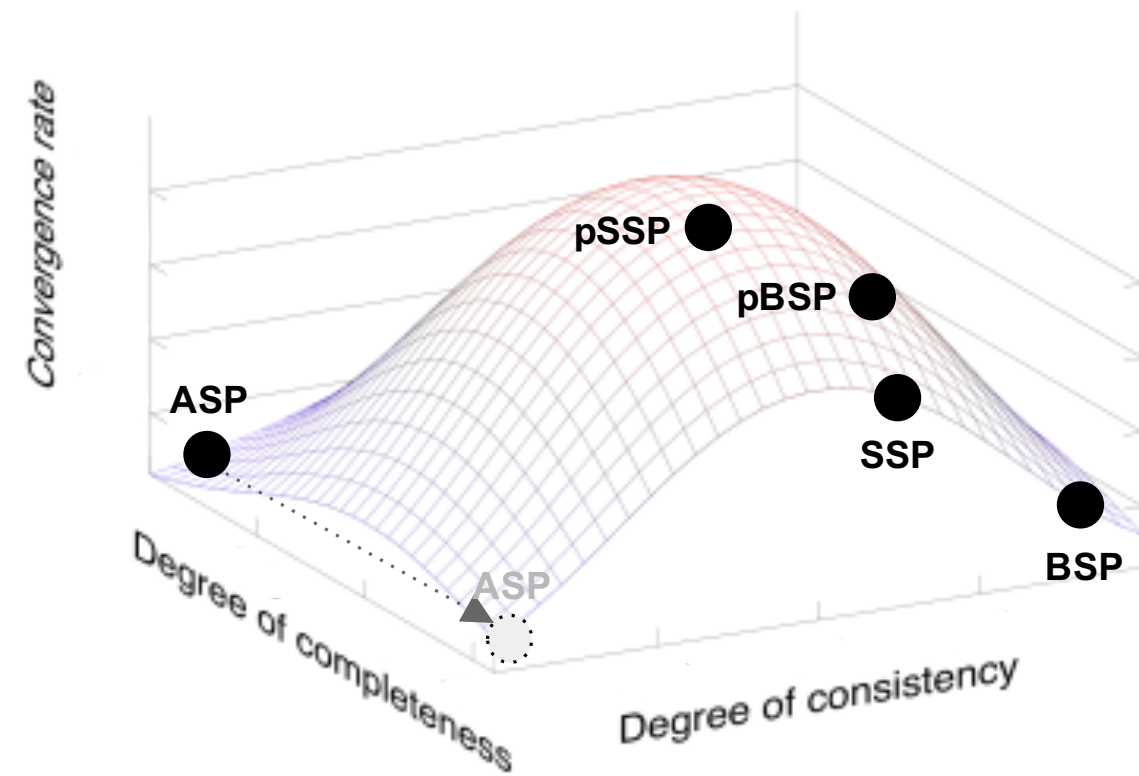
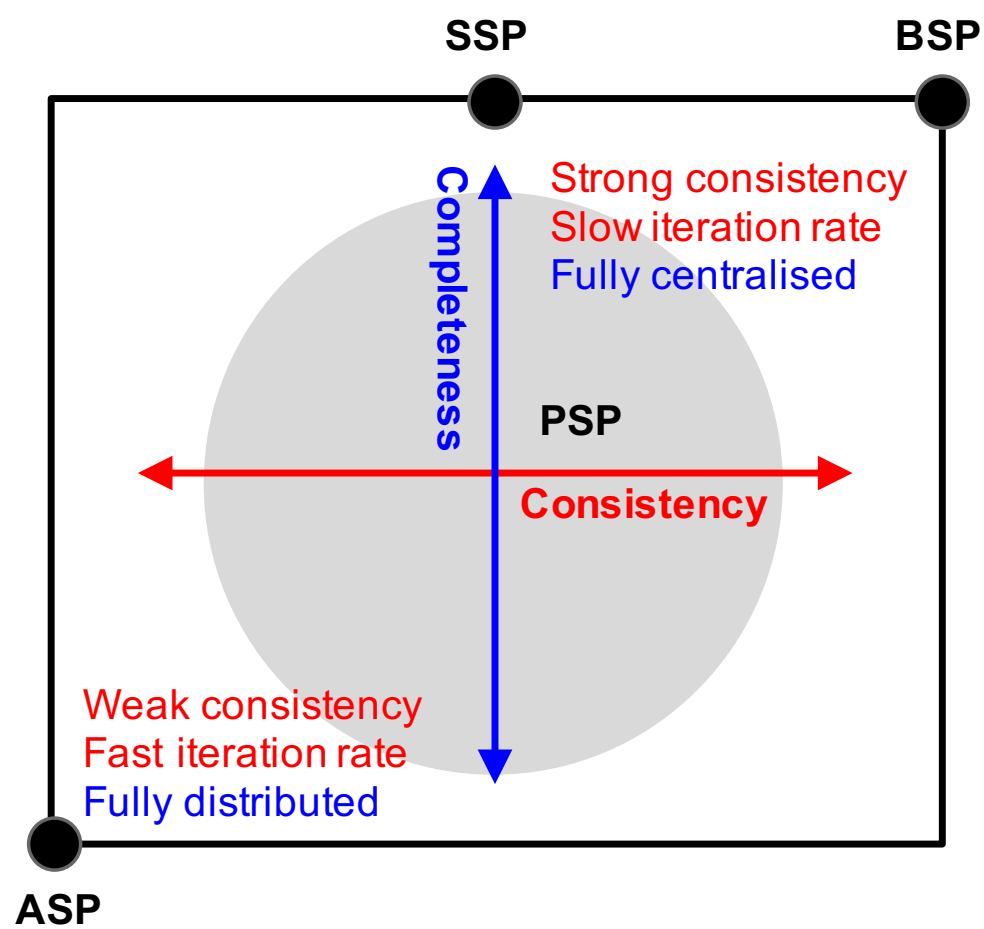
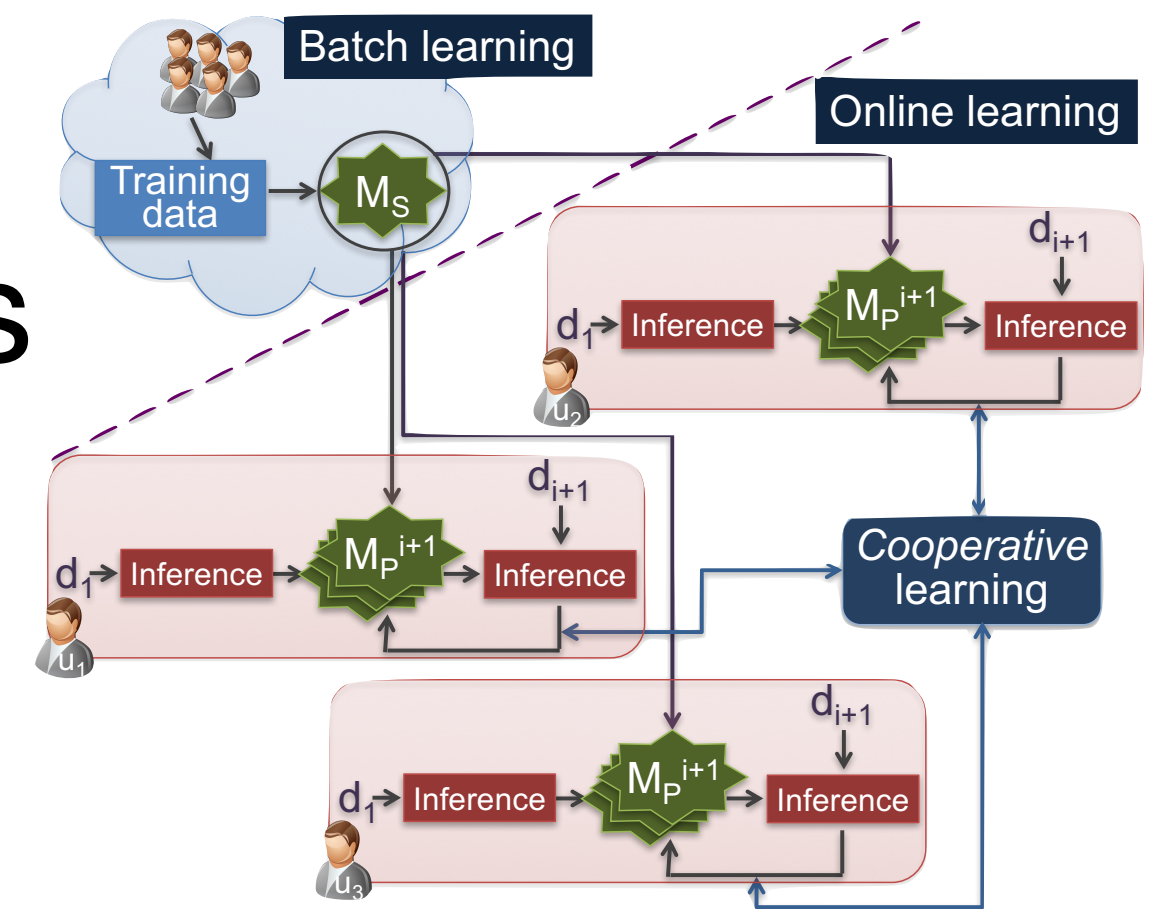


# Wide-Area Distributed Analytics

Current: centralise data so it can be processed, usually in big datacenters



First attempt: distribute models and then refine locally



Goal?

**Fully distributed inference and learning at scale**

# Welcome to BBC Box



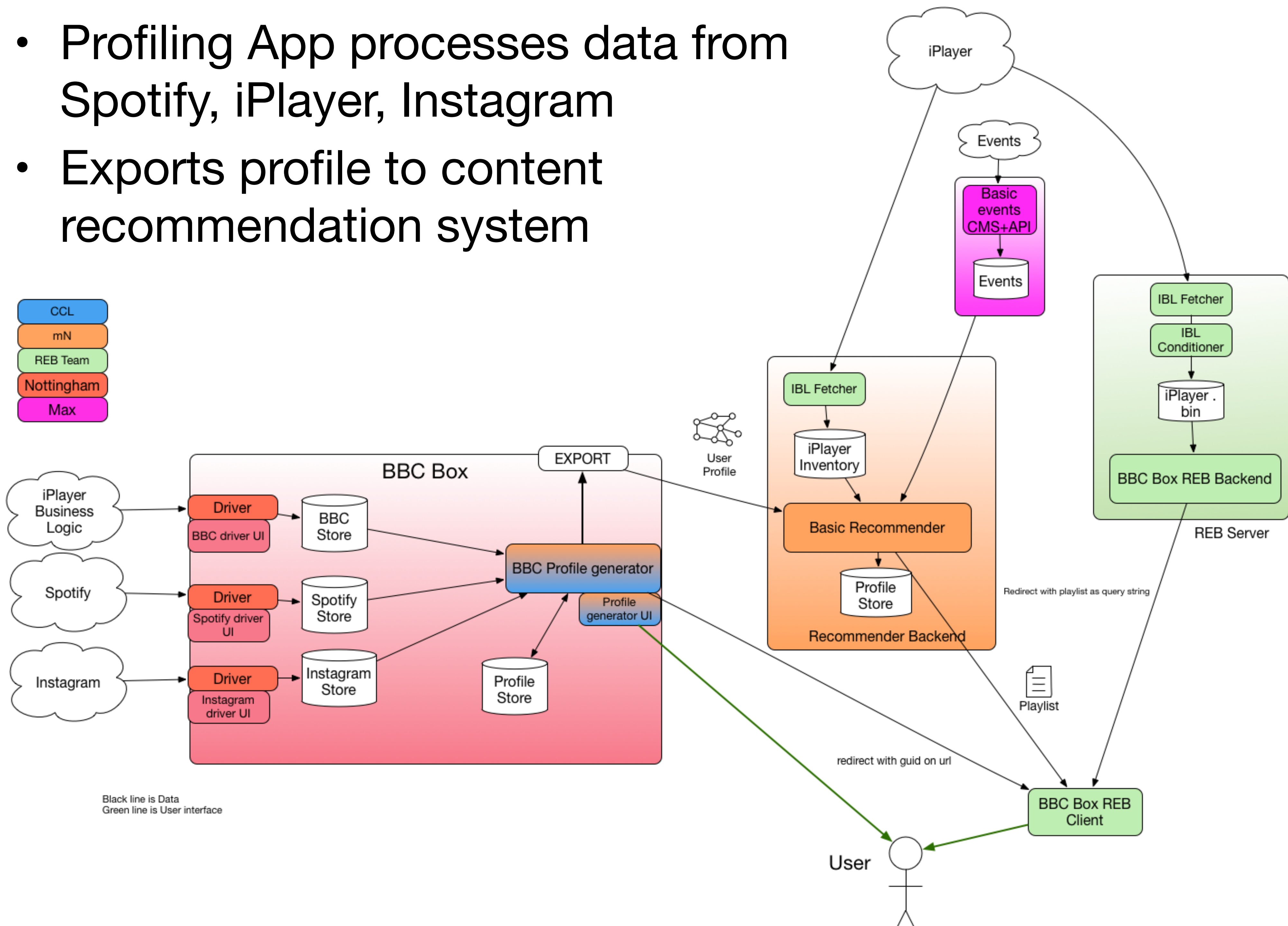
Did you know, under new data laws you have the right to access data about you and move it from one place to another?

Box is here to help!

[Let's get started](#)

All BBC box apps come with trusted certification

- Profiling App processes data from Spotify, iPlayer, Instagram
- Exports profile to content recommendation system



# HDI: So Where's the Interaction?

- Request and processing occur as if in a black-box
  - Can't tell where it's got to, what's going on
  - Status within the arrangement
- Requests, permissions and audit logs
  - Mechanisms of coordination within the field of work
  - Order but do not articulate the field of work
- Real world data sharing is **recipient designed**
  - Shaped by people with respect to the relationship they have with the parties implicated in the act of sharing

# Articulation Work

- Dataware subject is engaged in cooperative work
  - Interdependence between subject, processor, perhaps other subjects
  - E.g., walking down a busy street
- Activities must thus be meshed together, e.g., Schmidt (1994)
  - maintaining reciprocal **awareness of salient activities** within a cooperative ensemble
  - **directing attention towards current state** of cooperative activities
  - **assigning tasks to members** of the ensemble
  - handing over aspects of the work for **others to pick up**

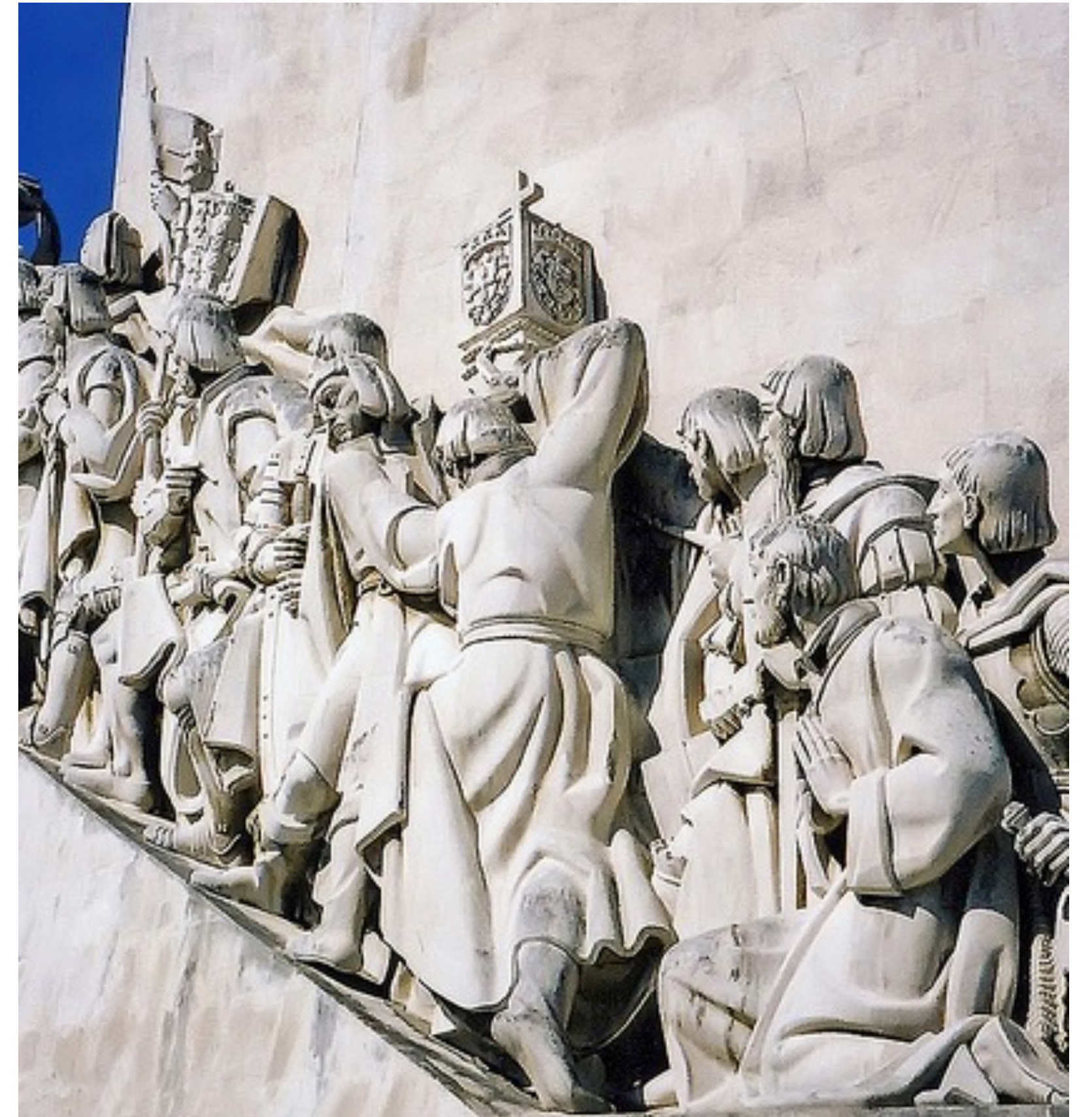
# Data as a Boundary Object

- Contextual nature – plastic adaptation to need
- E.g., Credit card receipt
  - Consumer's proof of **payment**
  - Bank's proof of a **valid transaction**
  - Supermarket's proof that **the bank should pay them**
- Inherently relational and thus social
  - Not so much 'me' or 'you' as 'us'
  - Very little is so private that it involves no-one else

# Interactional Challenges for HDI

## User Driven Discovery

- What is discovered? By whom? Under whose control?
  - Meta-data publication
  - Consumer analytics
- Empowering subjects: app stores?
  - Discoverability policies
  - Identity mechanisms
- Permissions, social ratings and exchange
  - App store models supporting discovery of data processors

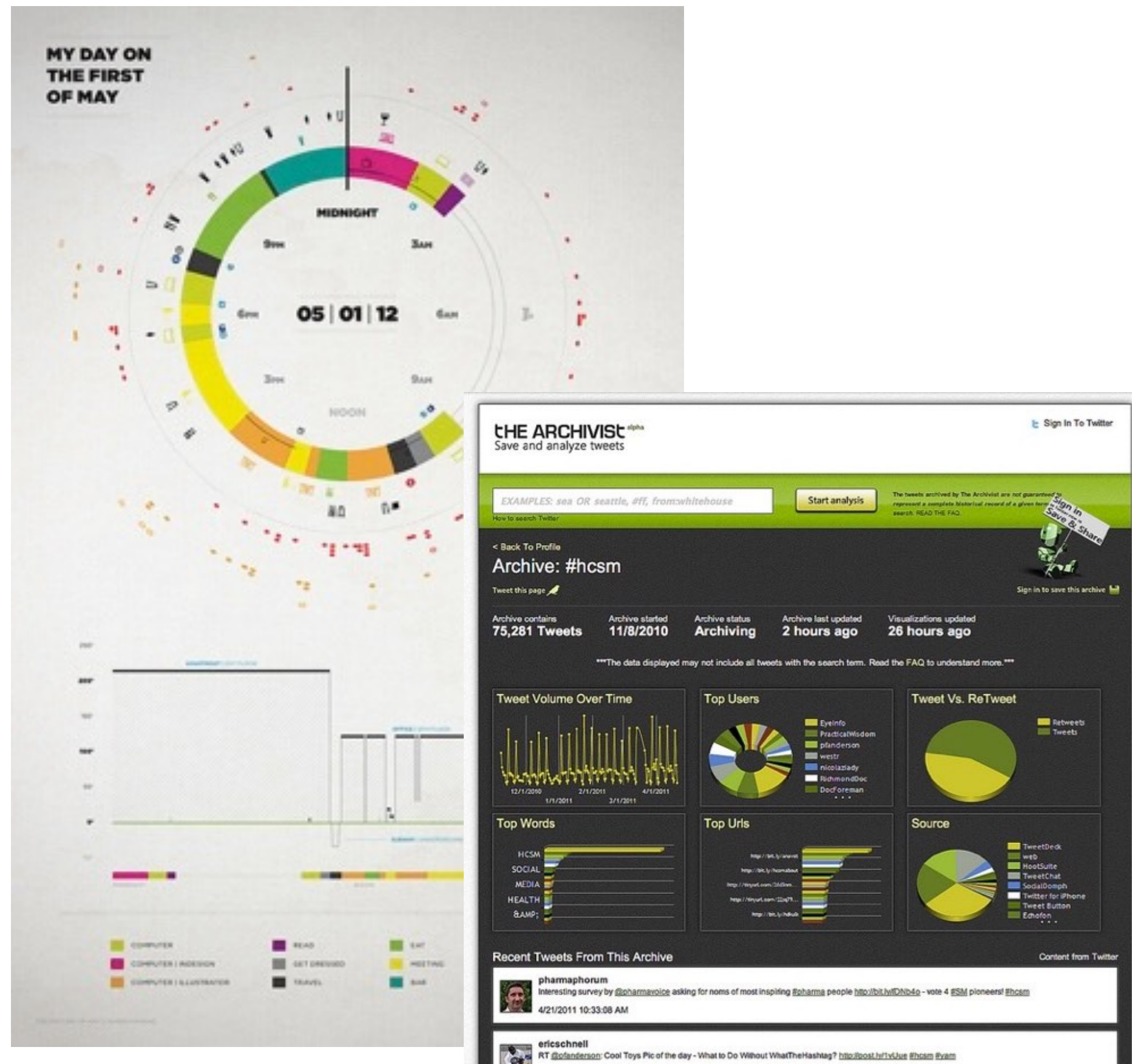


<https://flic.kr/p/4o1wLv>



# Interactional Challenges for HDI

<https://flic.kr/p/c3jJAY>



<https://flic.kr/p/9AwFd3>

## Legibility of Data Sources

- Visualisation of own data, impact of others' data
  - Help users make sense of data usage
  - Both present and future public data
- What you have, what others want
  - What processors would take from data sources
- Editing of data; control of presentation to processors – *Recipient design*
  - Support data editing and data presentation

# Interactional Challenges for HDI

## From My Data to Our Data

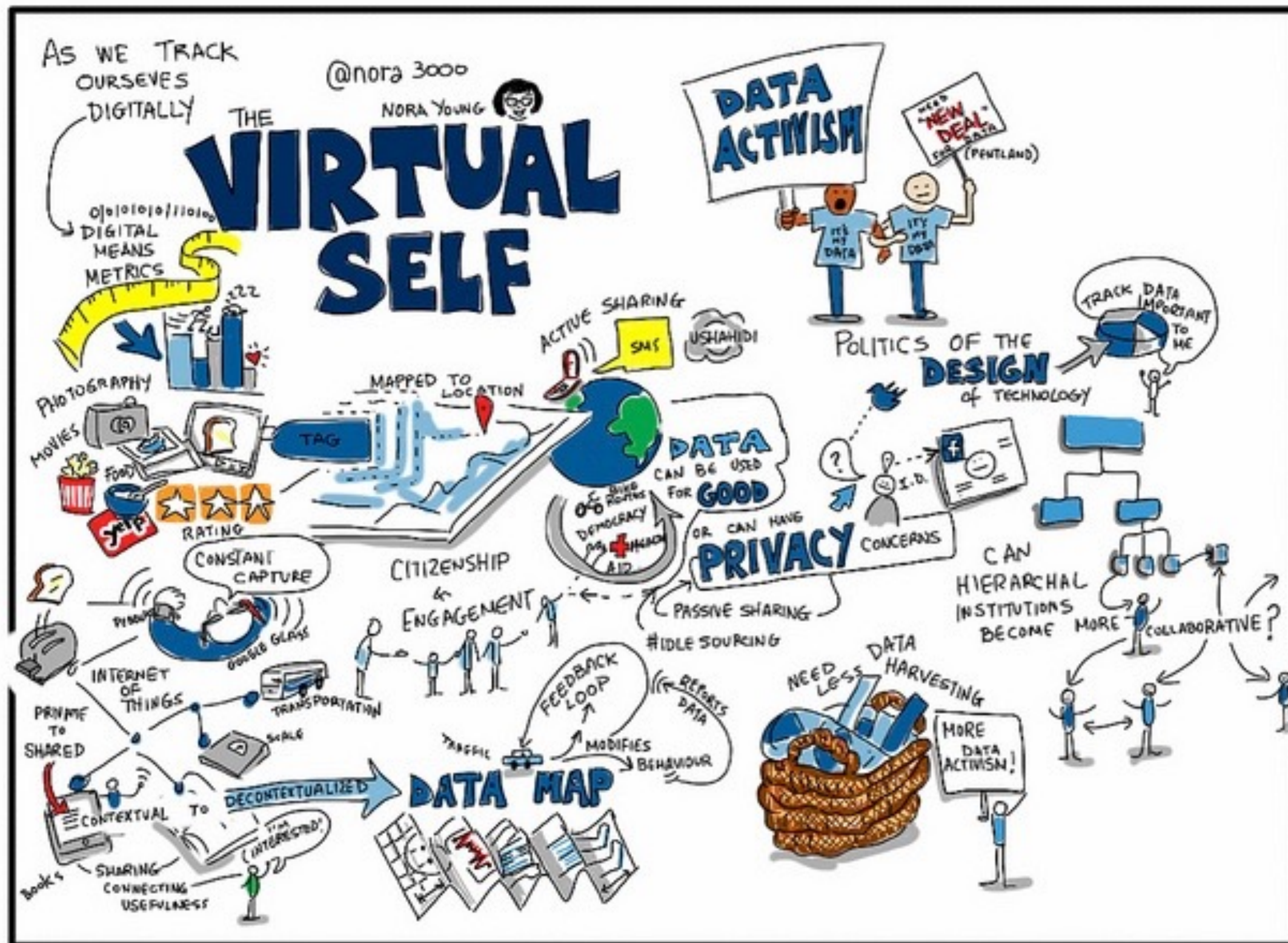
- Delegating and revoking control
  - Transparency/awareness mechanisms
  - Rights management
  - Editing, viewing, sharing
- Negotiation
  - Group management, negotiated collection and control
  - Group management of data sources



<https://flic.kr/p/drV8zY>

# Interactional Challenges for HDI

## Salient Dimensions of Collaboration



<https://flic.kr/p/e57ySb>

- To whom is data passed, for what purpose — *Transitivity*
  - Real time articulation of data sharing processes, e.g., current status reports
- Tracking and treatment
  - Data tracking, e.g., subsequent processing or transfer

# Platform Challenges

- Sharing data
  - Need to support offline data collection from e.g., mobile phones
  - Need a rendezvous and identity service for direct interconnection
- Shared data
  - No current platform is a good fit to social dynamics of a household!
  - Who and how to manage users, groups?
  - Who gets to be root?

# Questions?

<https://bit.ly/encyclopedia-hdi>  
<http://hdiresearch.org/>  
<https://databoxproject.uk/>  
<https://ocaml.xyz/>

<https://mort.io/>  
[richard.mortier@cl.cam.ac.uk](mailto:richard.mortier@cl.cam.ac.uk)

