UNIVERSITY OF CAMBRIDGE

Big Data Analytics, Human Data Interaction, and the Databox

Richard Mortier

Cambridge University Computer Laboratory

Networks & Operating Systems
SRG, Computer Laboratory

Outline

Part I

- We are all data subjects, and increasingly so
- How can we operate? Human-Data Interaction!
- Move the computation, not the data?

Part II

- Moving computation, Becoming Dataware
- Open challenges of interaction
- A physical realisation, the Databox

Outline

Part I

- We are all data subjects, and increasingly so
- How can we operate? Human-Data Interaction!
- Move the computation, not the data?

Part II

- Moving computation, Becoming Dataware
- Open challenges of interaction
- A physical realisation, the Databox

Our Digital Footprints

Digital footprints pose **major societal challenges**…

https://flic.kr/p/ppMdY1

BE DARING
GO
TRANSPARENT
OWN YOUR NAME.
MODEL A POSITIVE DIGITAL FOOTPRINT.
SHARE SHAMELESSLY.
Flickr CC Photo by Kristina Alexanderson
@GwynethJones - TheDaringLibrarian.com

In the future, your "digital footprint" will carry far more weight than anything you might include on a resume. -Chris Betcher
f
in
Implantable glucose sensor 0.5 x 0.5 x 5 mm
Regular 18-gauge hypodermal needle utilized for sensor implantation
Continuous monitoring and recording of glucose levels
Insta
g+

https://flic.kr/p/6sdrZB

…as the same time as opportunities for **economic growth**

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

Living in a Big Data World

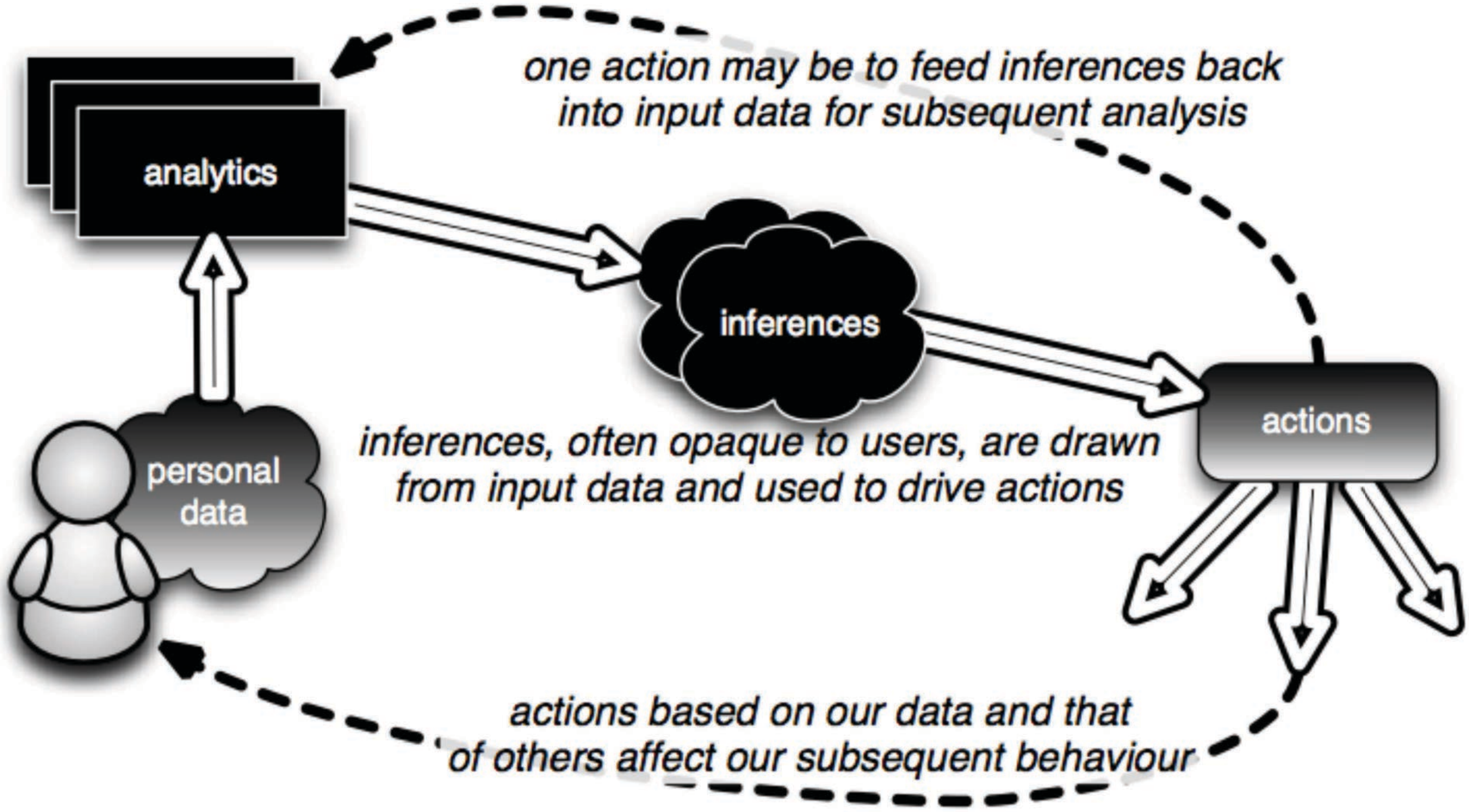
- Intimate information about us is collected and used
- It augments already large, rich data silos
- Never forgetting or forgiving

Key Challenge:

*How do we enable individuals 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."

Human-Data Interaction

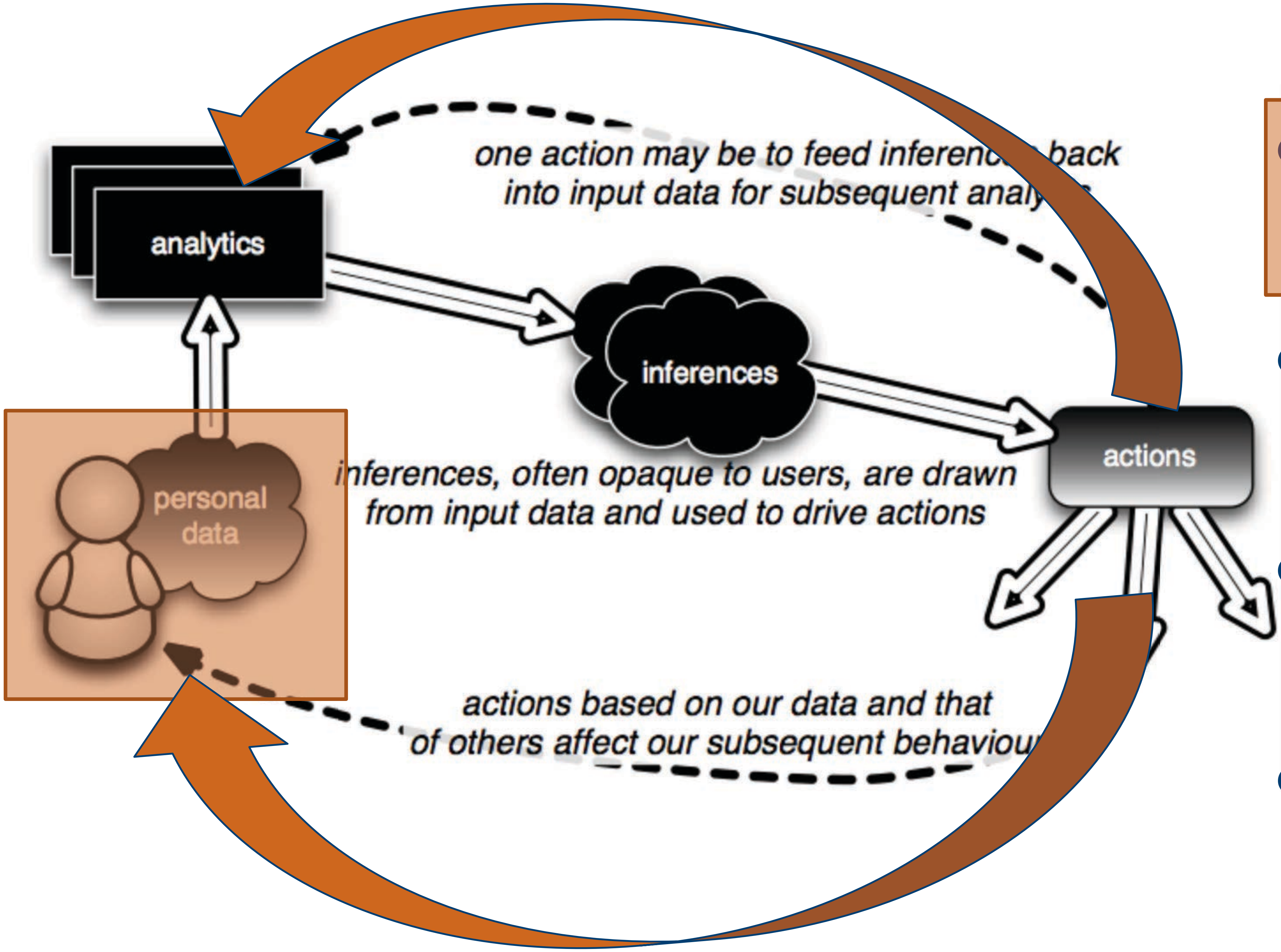


one action may be to feed inferences back into input data for subsequent analysis

inferences, often opaque to users, are drawn from input data and used to drive actions

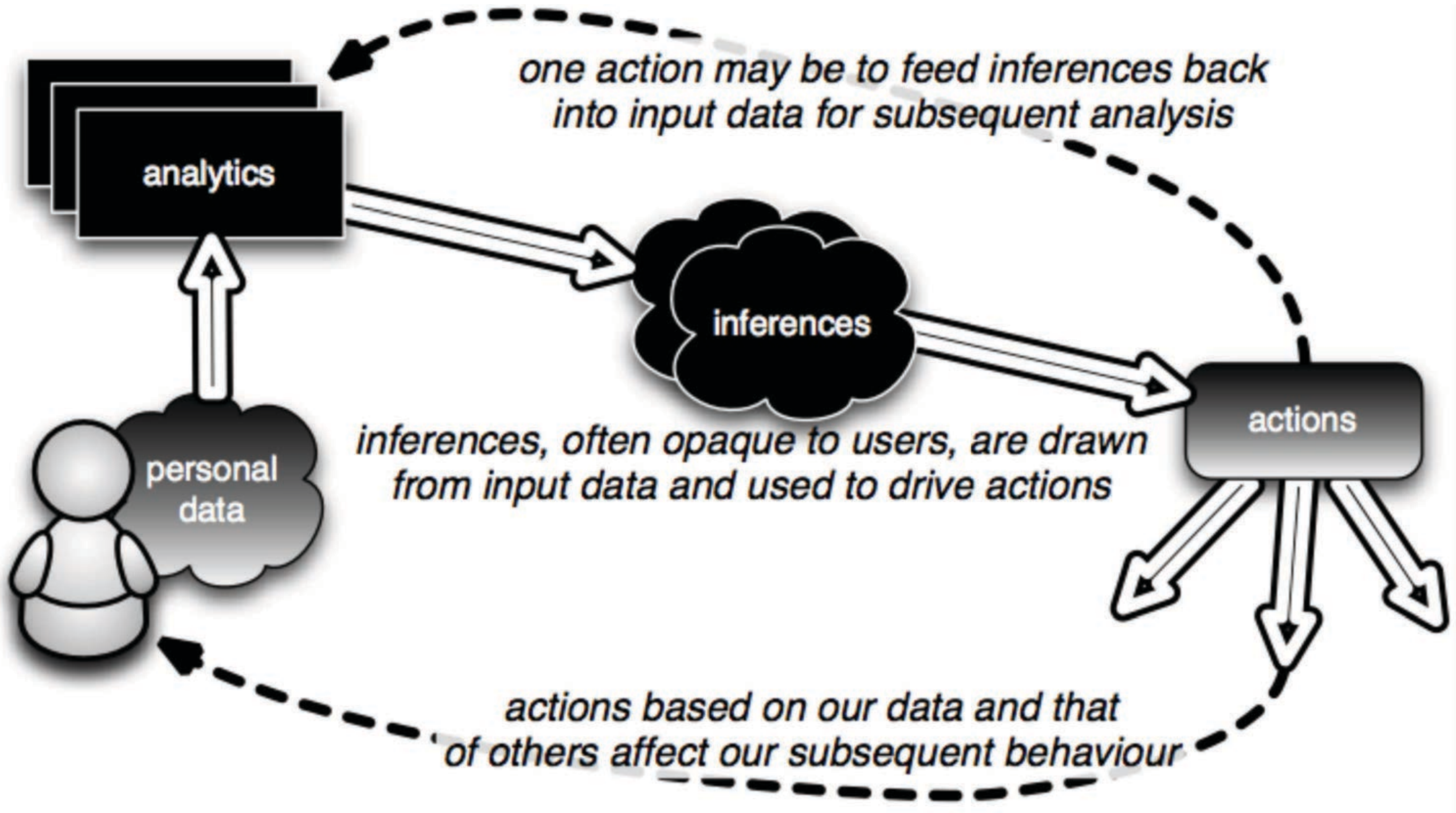
actions based on our data and that of others affect our subsequent behaviour

Human-Data Interaction


one action may be to feed inferences back into input data for subsequent analysis
analytics
inferences
actions
personal data
inferences, often opaque to users, are drawn from input data and used to drive actions
actions based on our data and that of others affect our subsequent behaviour

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

Human-Data Interaction


analytics
one action may be to feed inferences back into input data for subsequent analysis
inferences
actions
personal data
inferences, often opaque to users, are drawn from input data and used to drive actions
actions based on our data and that of others affect our subsequent behaviour

We believe current systems lack

Legibility, **Agency**, **Negotiability**

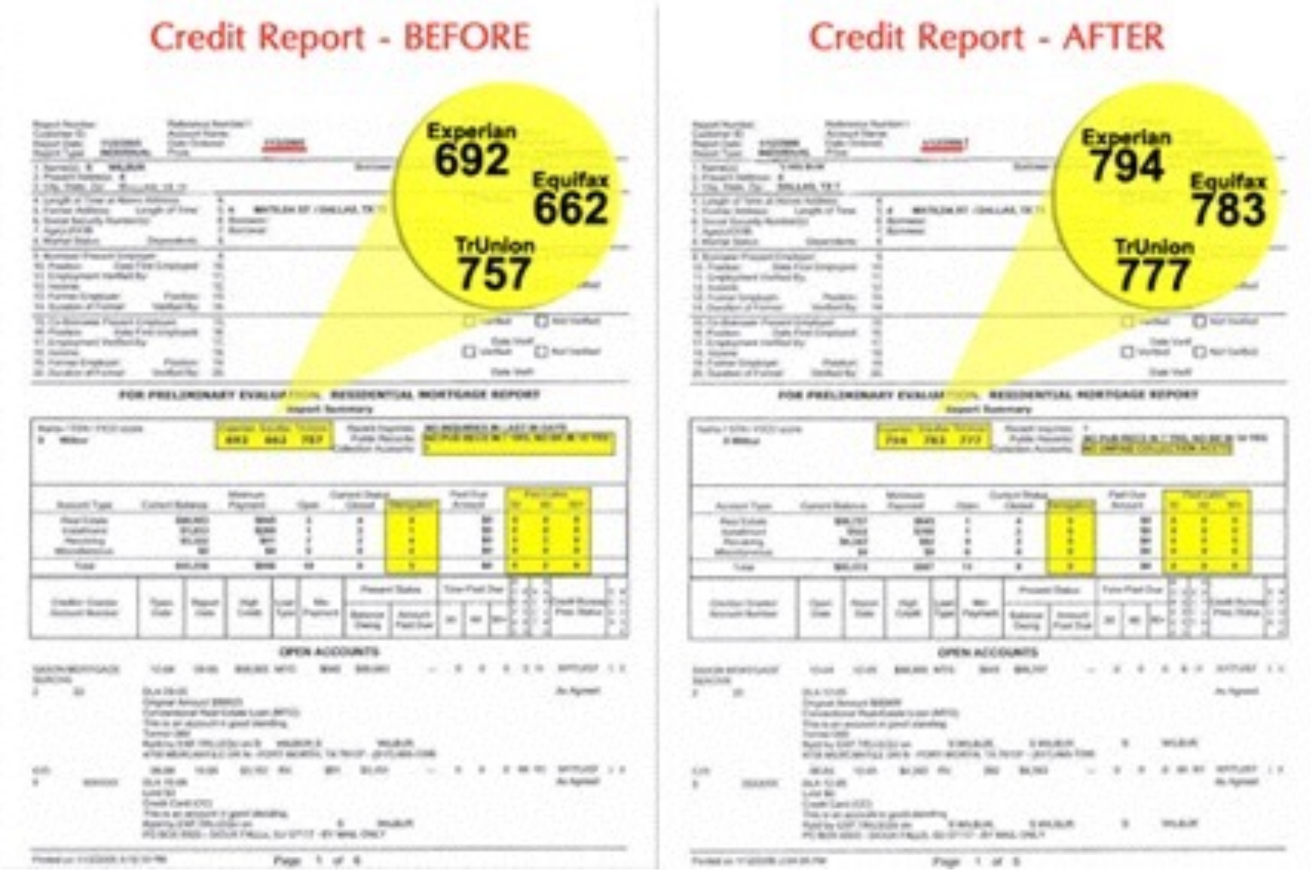
Legibility

Visualisation & comprehension

- E.g., Nest thermostat
  - Simple information display
  - Supports many interaction modalities
  - Hides details of internal processes

https://flic.kr/p/azwi7q

Lack of Legibility



Credit Report - BEFORE

Experian 692

Equifax 662

TrUnion 757

Credit Report - AFTER

Experian 794

Equifax 783

TrUnion 777

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

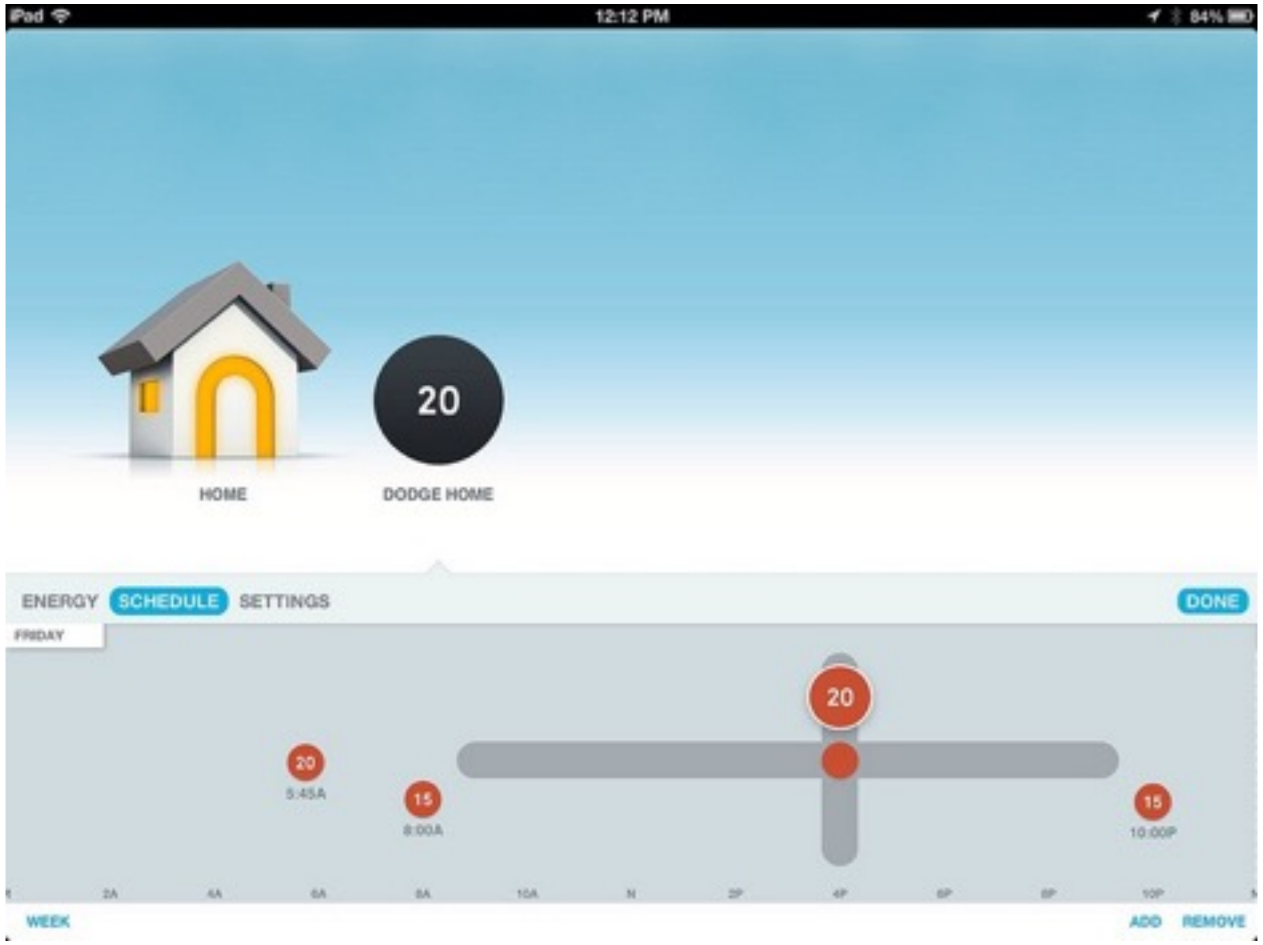
- We are unaware of
  - the many **sources of data** collected about us,
  - the **analyses performed** on this data, and
  - the **implications** of these analyses

E.g., Computation of credit scores

Agency

Capacity to act

- E.g., Nest Thermostat
  - Learns a schedule, but
  - Supports user override, by
  - Setting desired temperature on-demand



https://flic.kr/p/e3oH3k

Lack of Agency

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

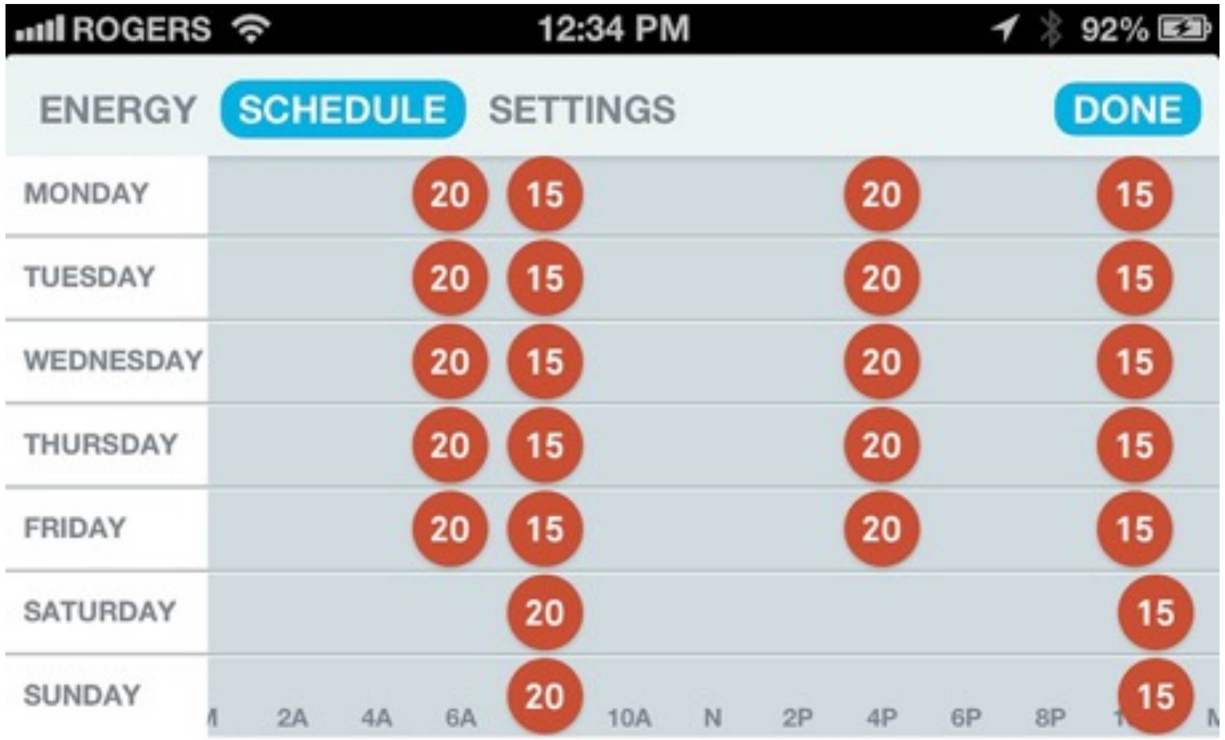
E.g., Use of purchase details to profile your propensity to risk and sell this to an insurance agency

- We are 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

Negotiability

Support the dynamics of interaction

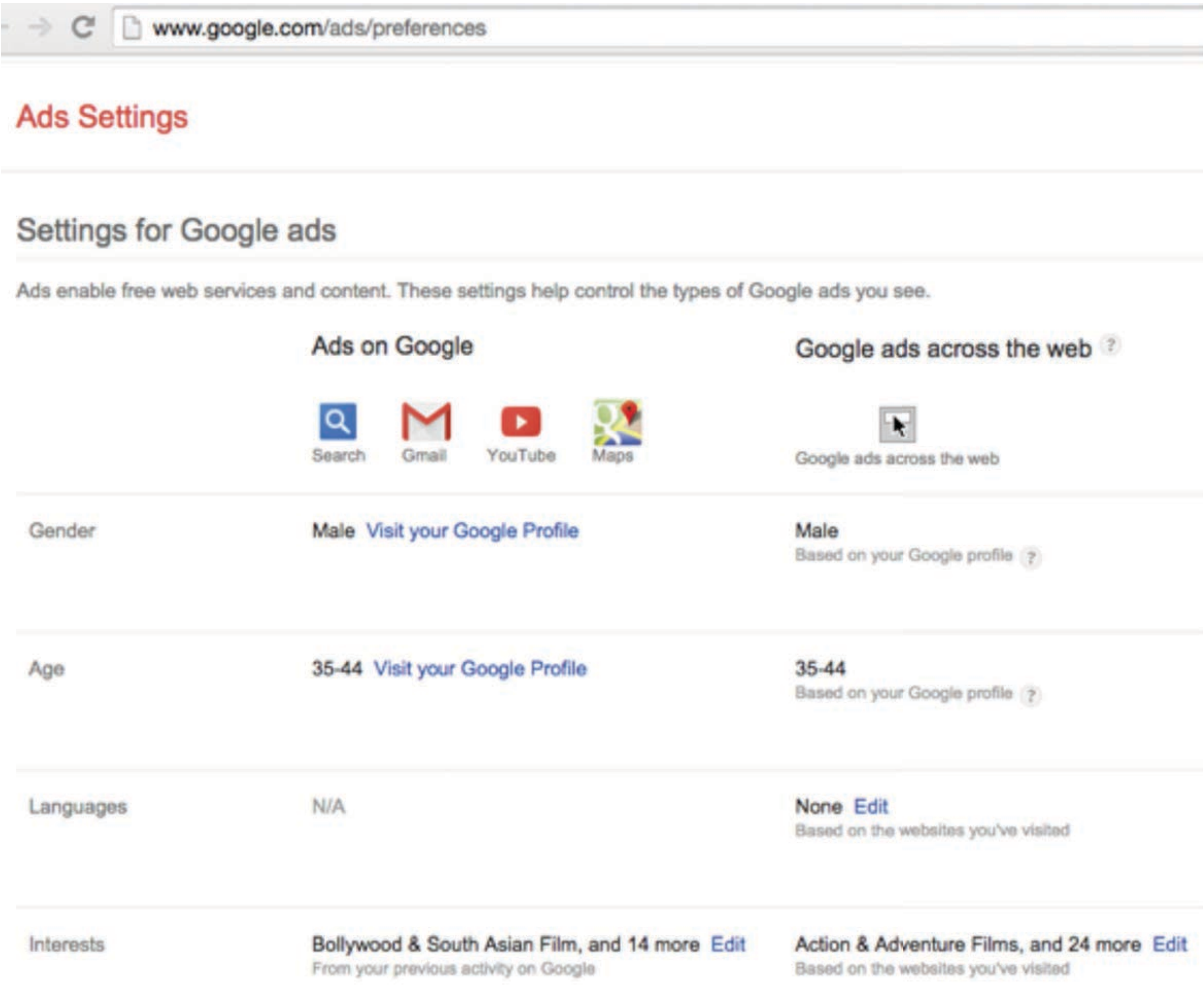
- E.g., Nest Thermostat
  - Provides means to inspect and edit the schedule it has learnt
  - Continually updates learnt behaviour to adapt to changes in context
  - Based on context-dependent patterns of past user interaction



MONDAY	20	15	20	15
TUESDAY	20	15	20	15
WEDNESDAY	20	15	20	15
THURSDAY	20	15	20	15
FRIDAY	20	15	20	15
SATURDAY	20	15		
SUNDAY	20	15		

https://flic.kr/p/i8cHvi

Lack of Negotiability



www.google.com/ads/preferences

Ads Settings

Settings for Google ads

Ads enable free web services and content. These settings help control the types of Google ads you see.

	Ads on Google	Google ads across the web ?
	Search Gmail YouTube Maps	Google ads across the web
Gender	Male Visit your Google Profile	Male Based on your Google profile ?
Age	35-44 Visit your Google Profile	35-44 Based on your Google profile ?
Languages	N/A	None Edit Based on the websites you've visited
Interests	Bollywood & South Asian Film, and 14 more Edit From your previous activity on Google	Action & Adventure Films, and 24 more Edit Based on the websites you've visited

Even given

- we know the data collected and analyzed about us, and
- we 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
- Cannot easily correct data or inferences held about us

An Underlying Structural Problem

- The Internet is fragmented, distributed systems are difficult
  - Everything is much easier if you centralise
  - With the cloud, we can!

- Ease of cloud computing has led to two poor defaults:
  1. Move the data …
  2. … to a centralised location

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

Implications

http://cliparts.co/honey-pot-clip-art

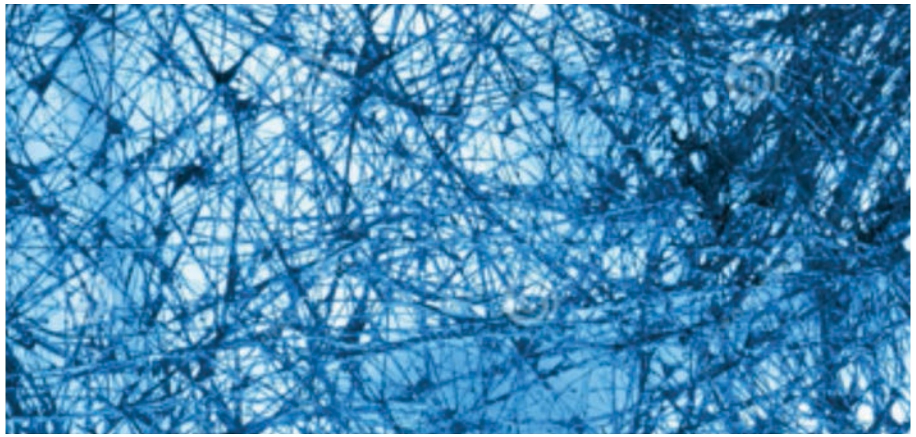
Security

- Creation of a honey-pot
- Highly desirable to attackers

Performance

- Creation of a performance challenge
- Require enormous, reliable, connected resource

http://autoguide.com.vsassets.com/blog/wp-content/uploads/2014/05/traffic-jam.jpg

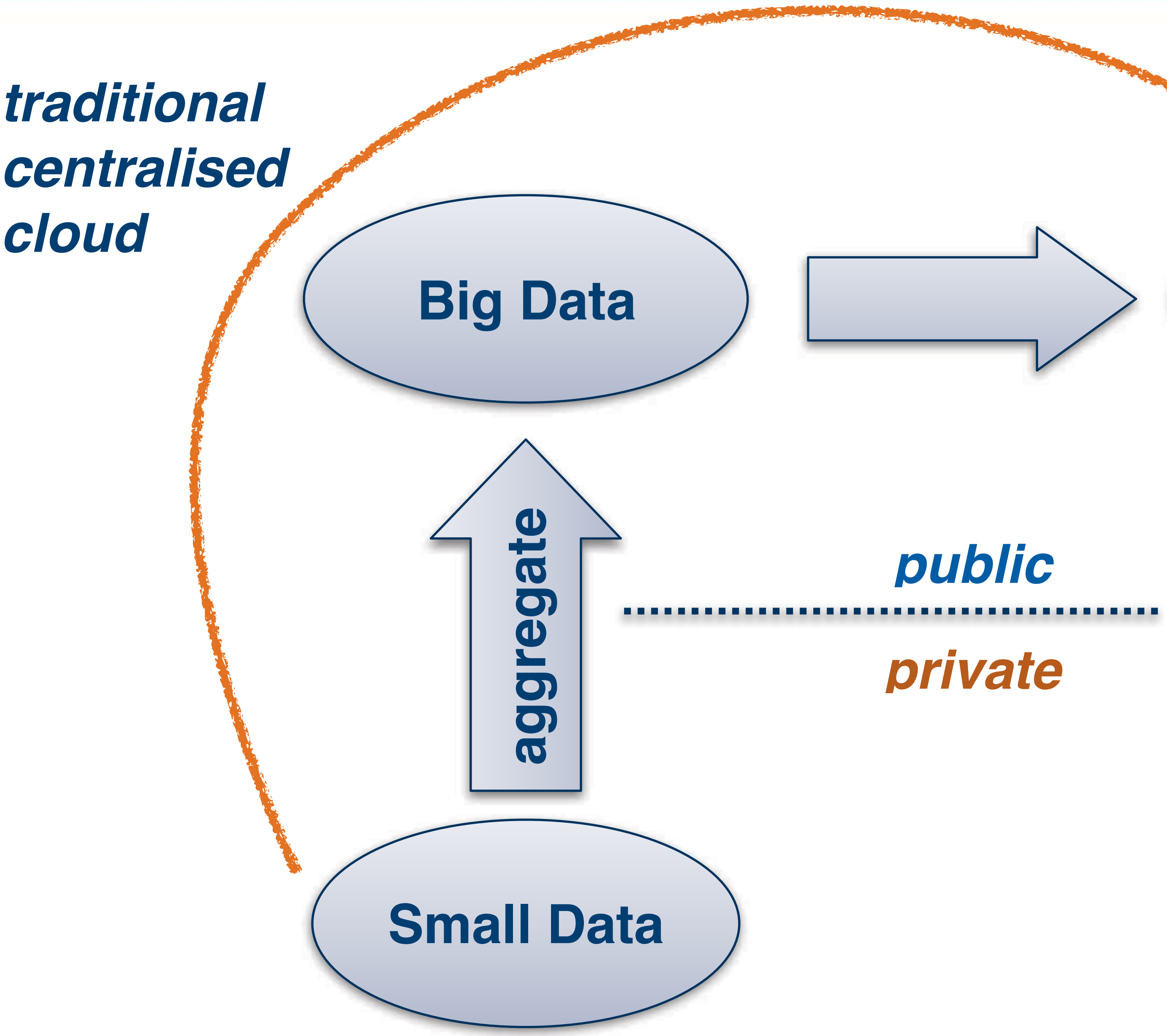


https://www.dreamstime.com/royalty-free-stock-photography-complex-abstract-communication-image18615337

Interaction

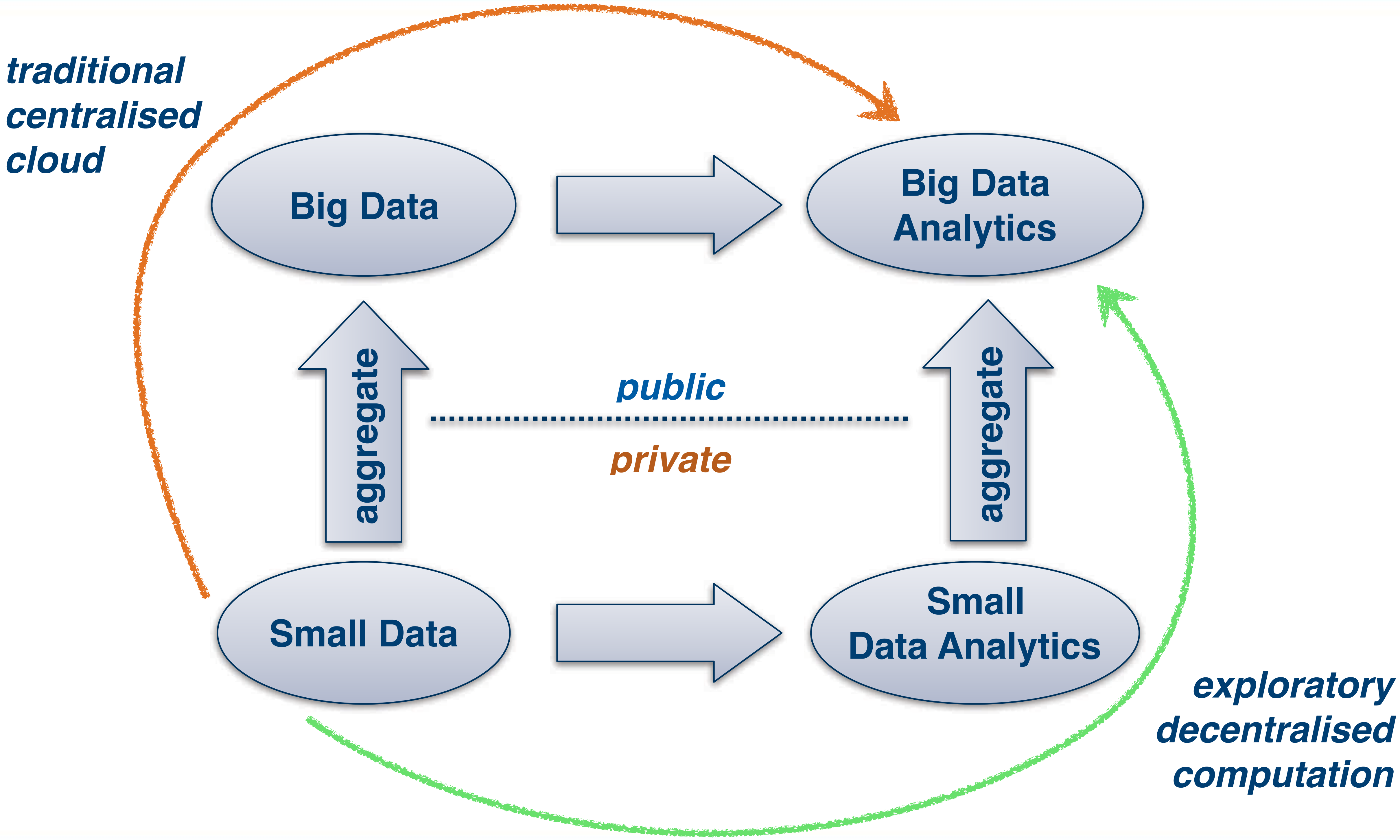
- Creation of an abstraction
- It’s all “out there somewhere”

Big Data Analytics?

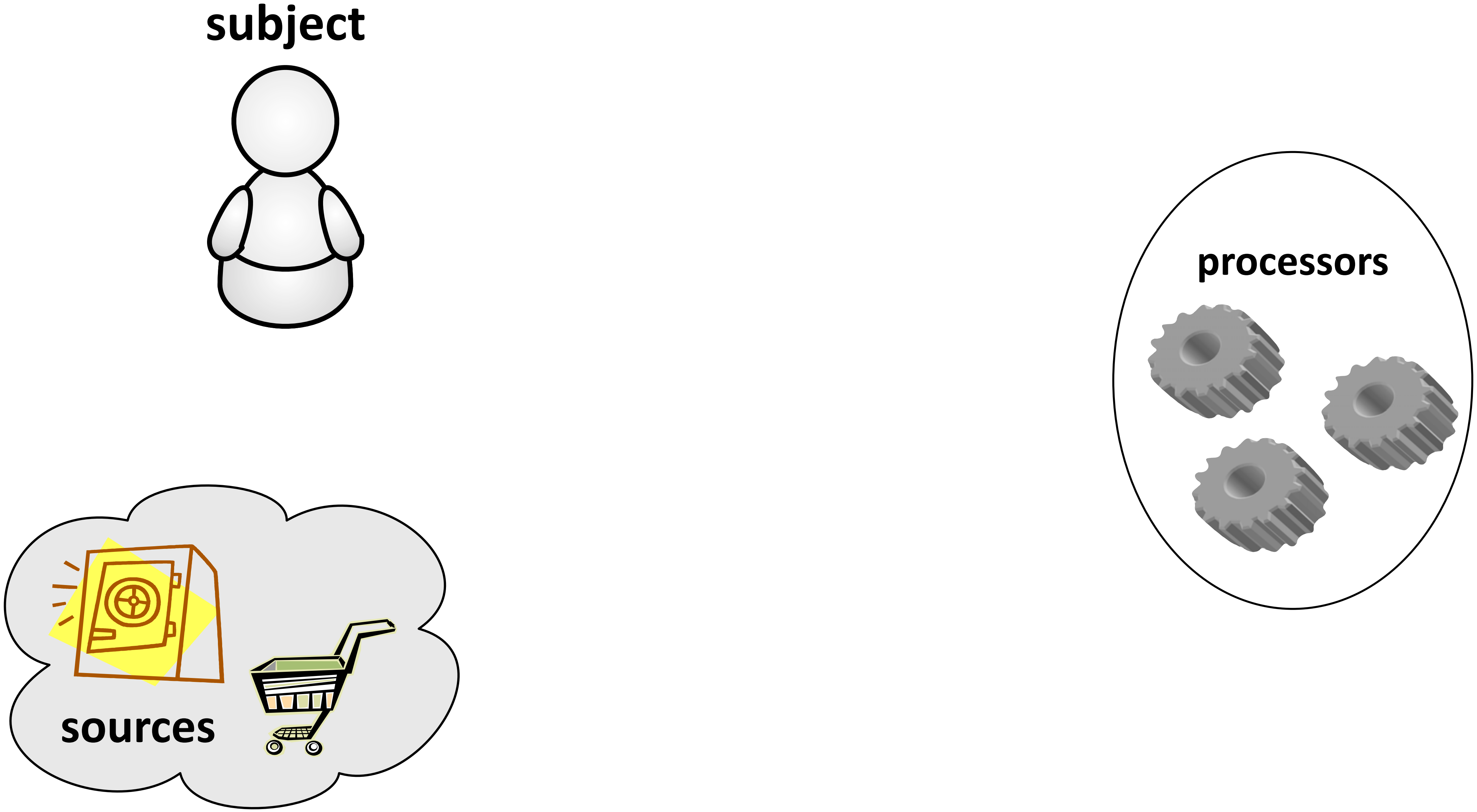


- Loss of contextual information
- Ethical and legal issues arise
- Platform technology challenges

Big Data Analytics? Small Data Analytics!



Dataware: The Actors

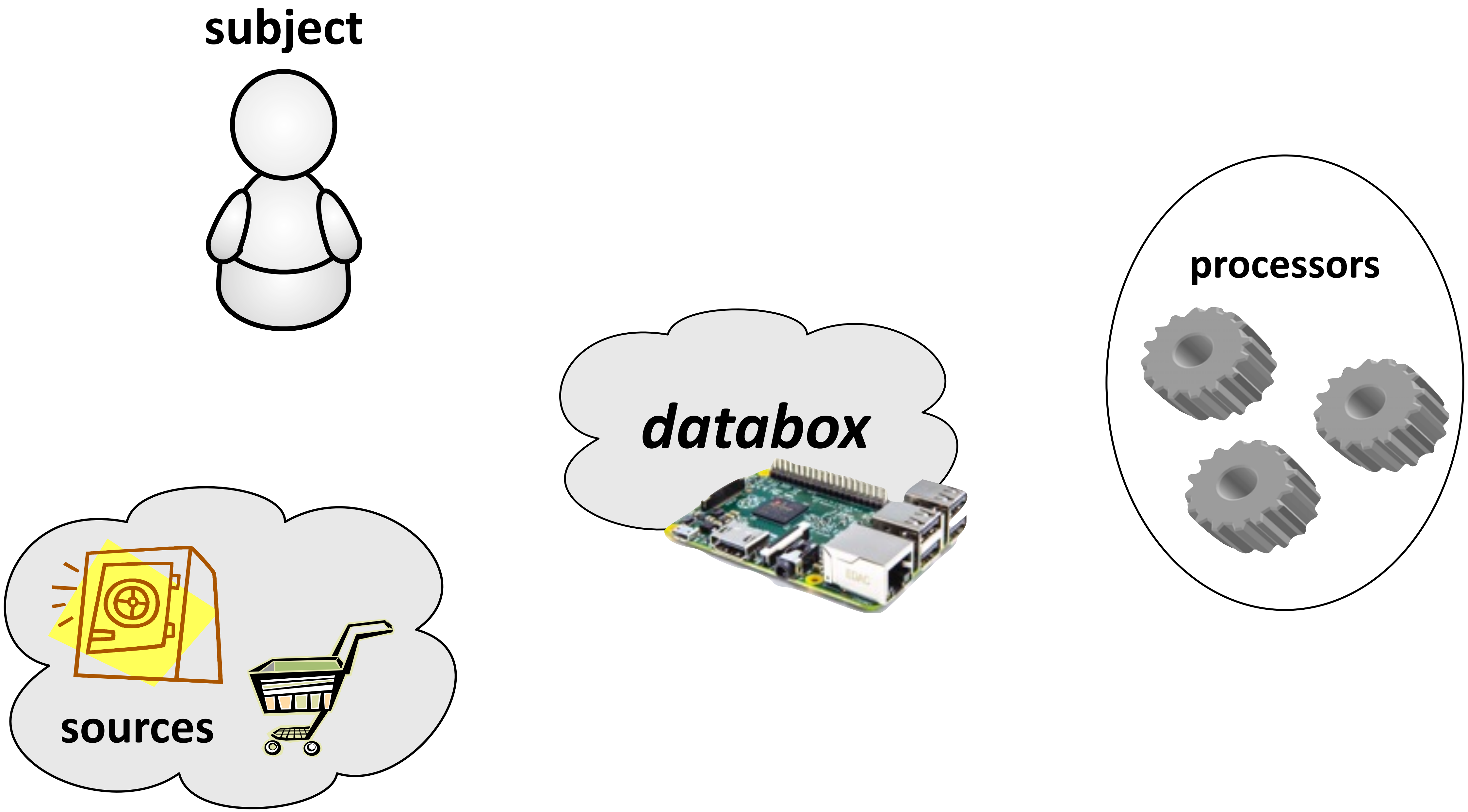


subject

processors

sources

Dataware: Implementing HDI



subject

processors

databox

sources

End Part I! Questions?

http://mort.io/
richard.mortier@cl.cam.ac.uk

http://hdiresearch.org/
http://homenetworks.ac.uk/
https://mirage.io/
https://forum.databoxproject.uk/

Mortier et al, SSRN'14

Angelopoulos et al, ICIS'16

Mortier et al, HCI Encyclopedia (2016)

Outline

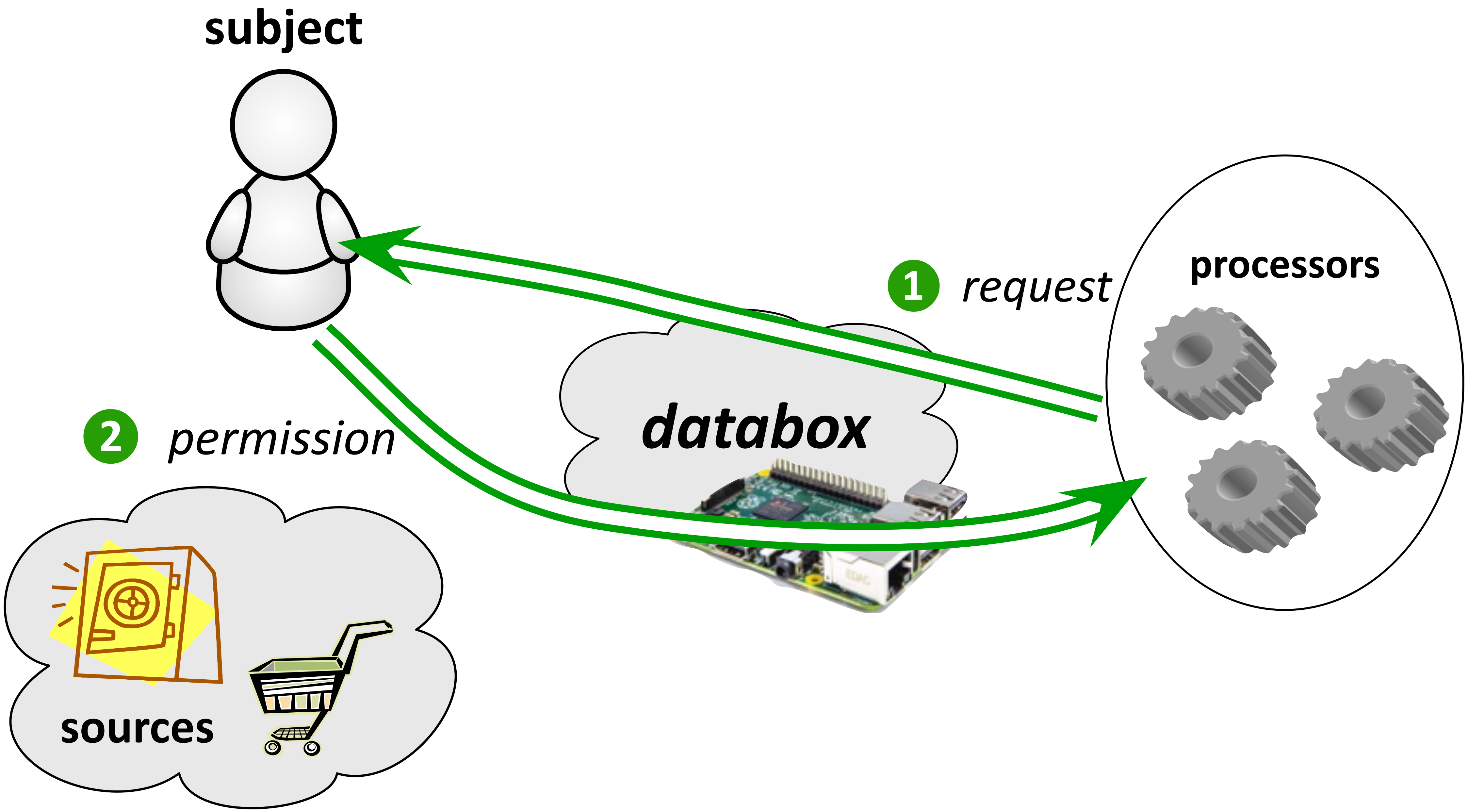
Part I

- We are all data subjects, and increasingly so
- How can we operate? Human-Data Interaction!
- Move the computation, not the data?

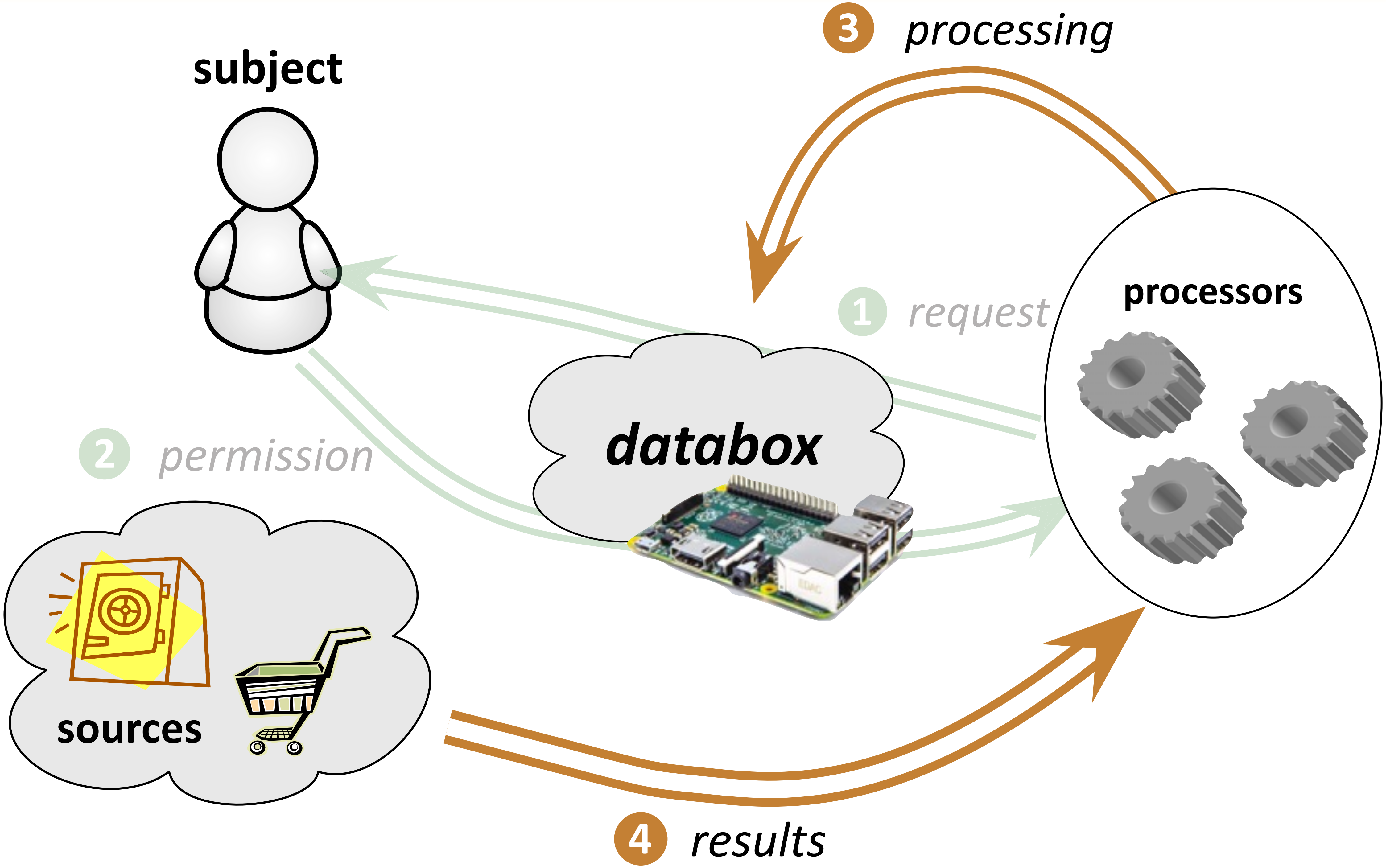
Part II

- Moving computation: Becoming Dataware
- A physical realisation: the Databox
- Some open challenges of interaction

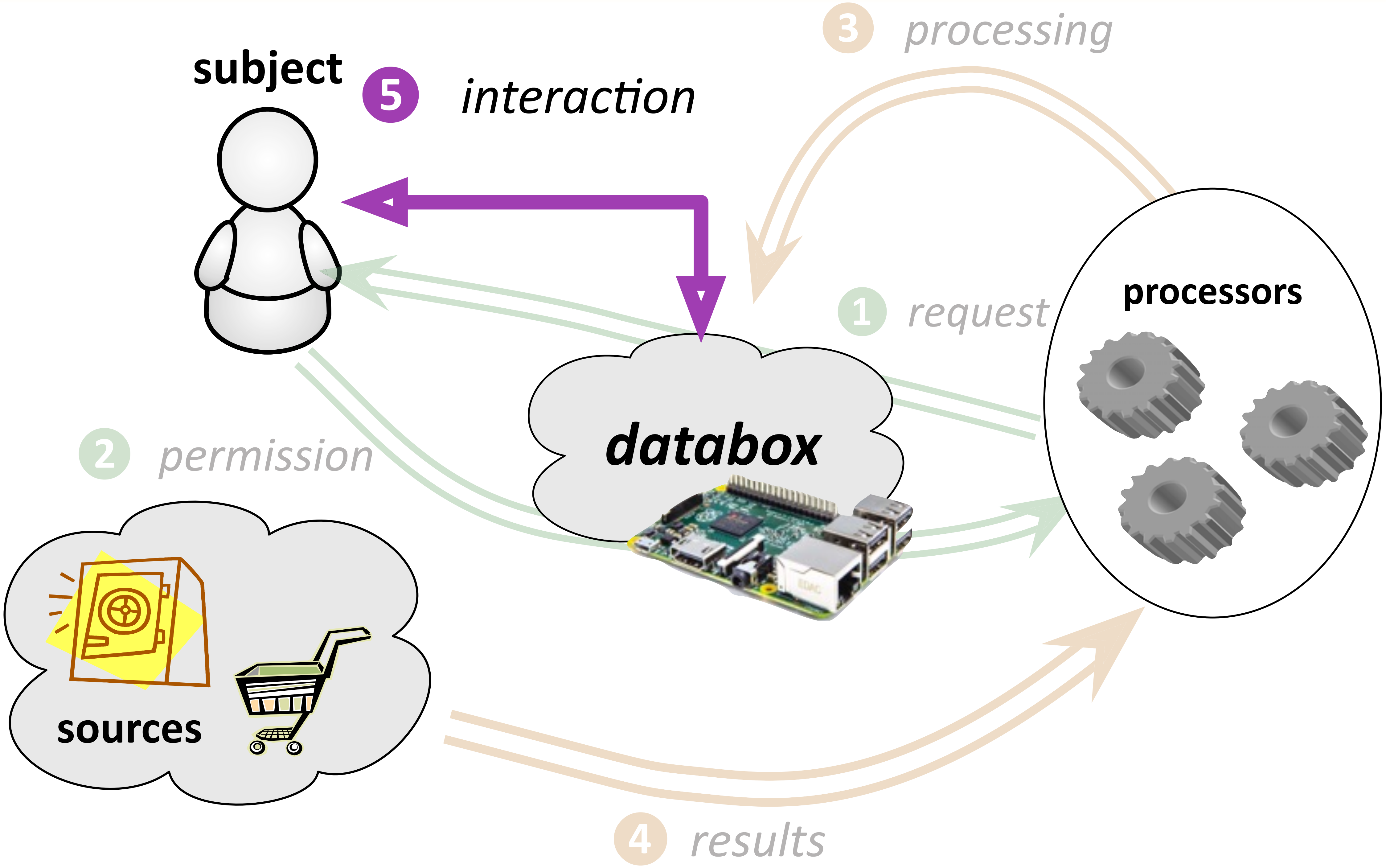
Dataware: Legibility



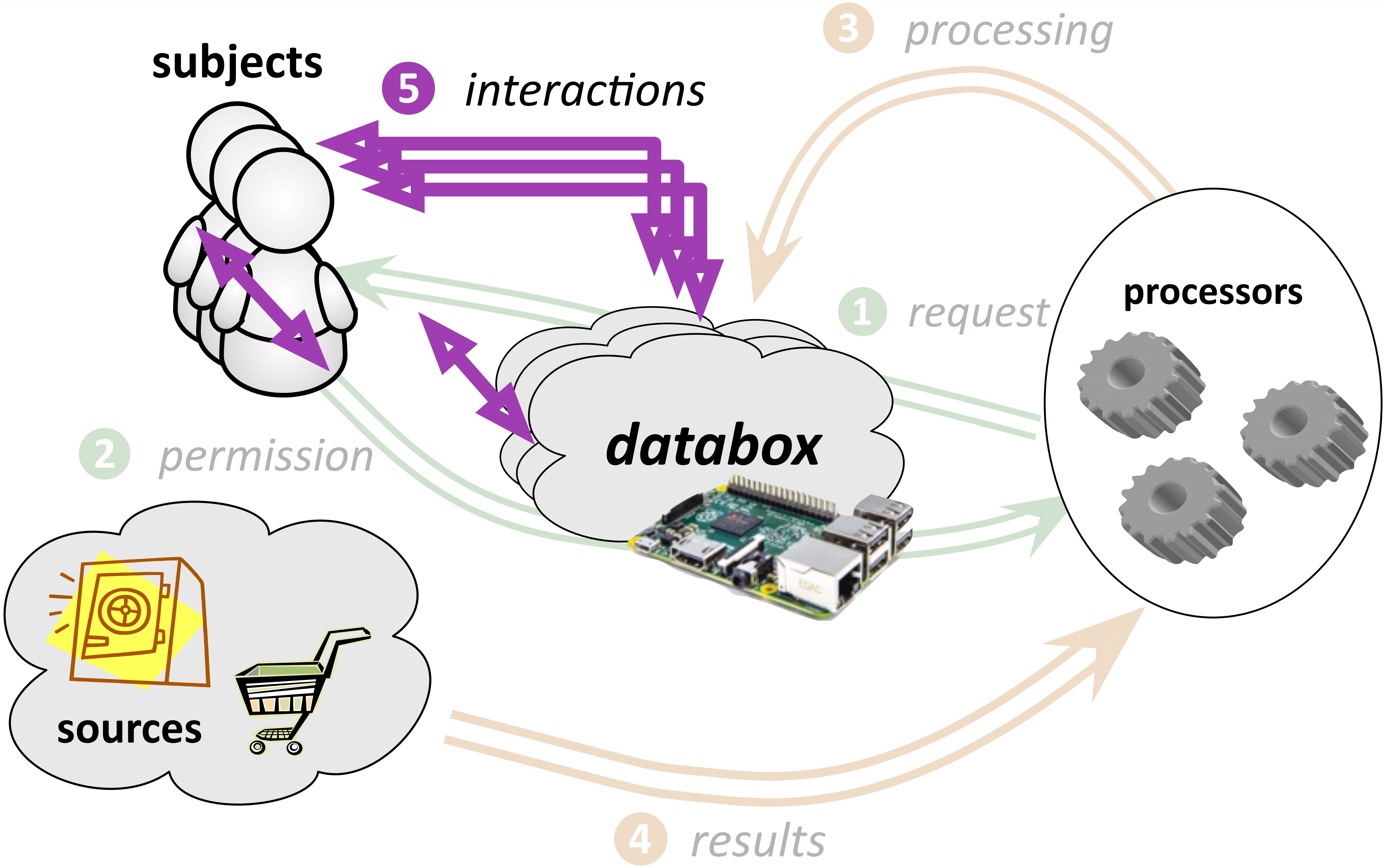
Dataware: Agency



Dataware: Negotiability



Dataware: Constructing Interaction



Dataware: Constructing Interaction

- Numerous proposed interaction models
  - E.g., pay-per-use
- Little about how to actually provide for it
- *Dataware* one such proposal
  - Accountable transaction between parties in terms of request, permission, audit
- But there's a lot more to consider here…

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

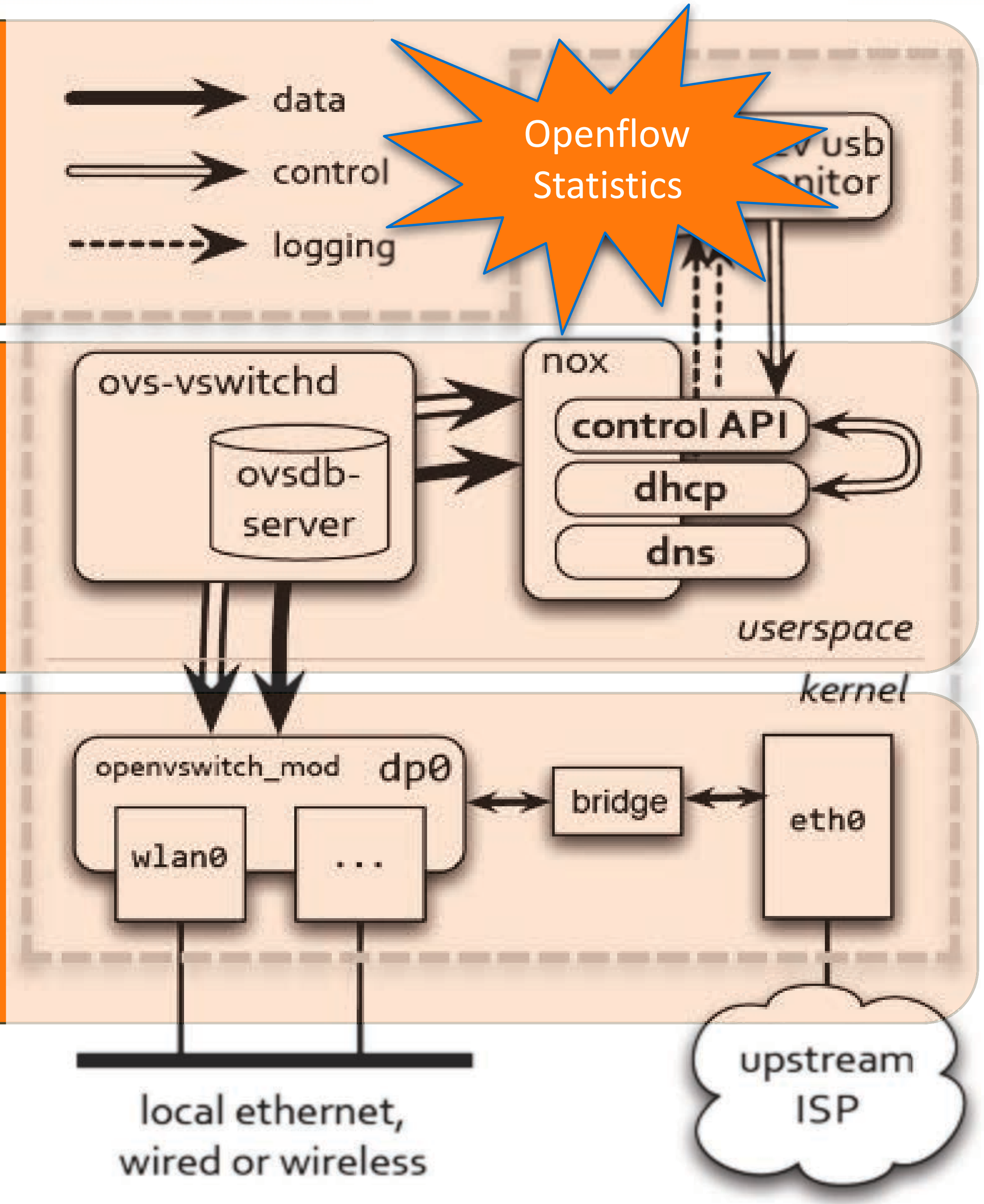
Digression: Home Networking

- Focused at home route
  - Single point of control in the home
  - Avoid ma heterogeneous clients
- Built a hom
  - Used Ope provide c server, DNS interception, and a control API

Monitoring traffic

Controlling traffic

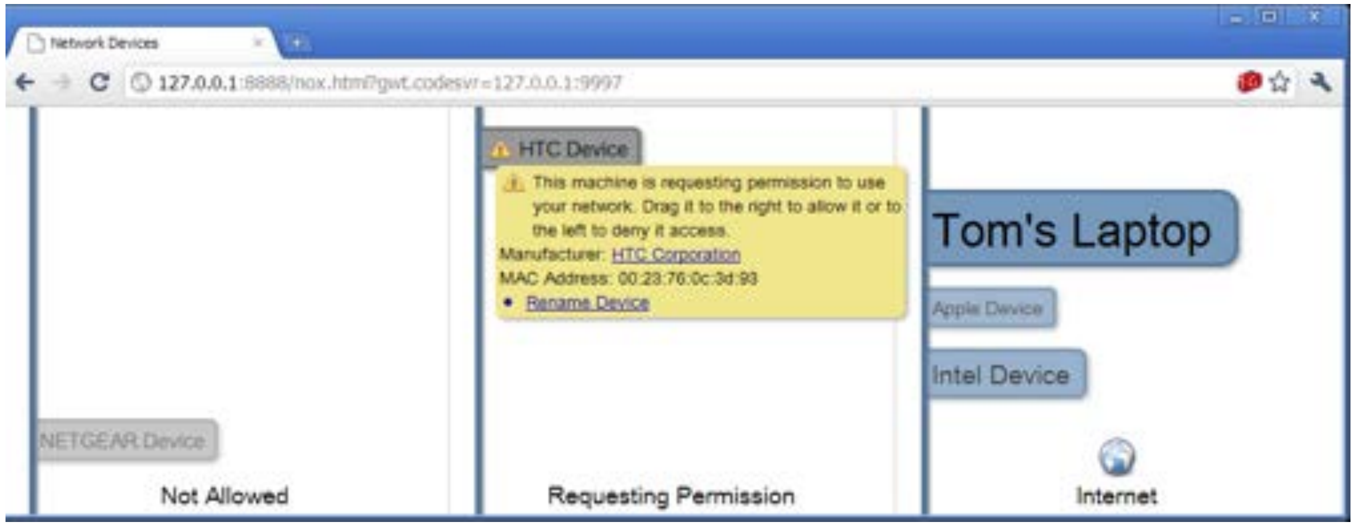
Forwarding traffic



[Mortier et al, ACM UIST'12]

Even More Complex than Home Networking

- Disambiguation can't be delegated to a nominated householder/cohort
  - Too many relational issues wrapped up in this
  - Old, young; Parents, children; Colleagues, friends, lovers
- Not even just about my **vs** our data
  - We may not agree



Network Devices

127.0.0.1:8888/nox.html?gwt.codesvr=127.0.0.1:9997

HTC Device

This machine is requesting permission to use your network. Drag it to the right to allow it or to the left to deny it access.

Manufacturer: HTC Corporation

MAC Address: 00:23:76:0c:3d:93

- Rename Device

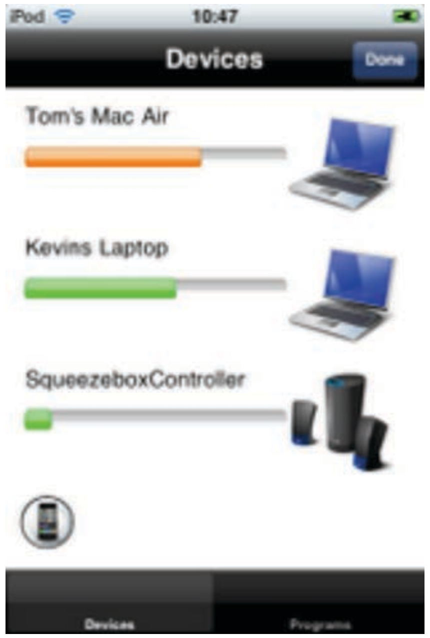
NETGEAR Device

Tom's Laptop

Apple Device

Intel Device

Not Allowed | Requesting Permission | Internet



iPod 10:47

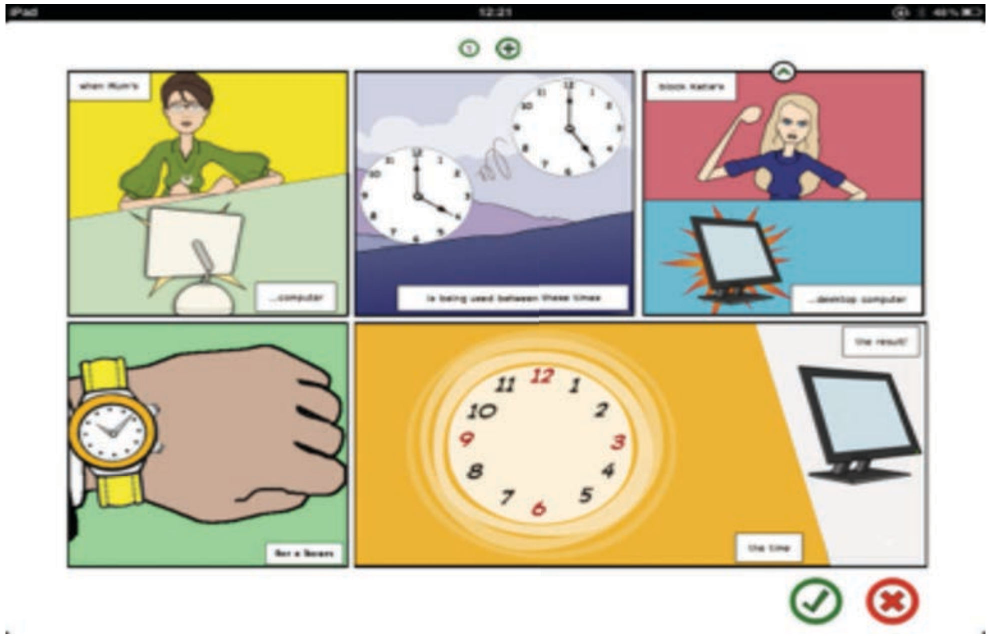
Devices — Done

Tom's Mac Air

Kevins Laptop

SqueezeboxController

Devices | Programs



[Crabtree et al, Springer PUC'15]

Articulation Work

- Dataware subject is engaged in cooperative work
  - There is interdependence between subject, processor, perhaps other subjects
- 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**

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

Interactional Challenges for HDI

User Driven Discovery

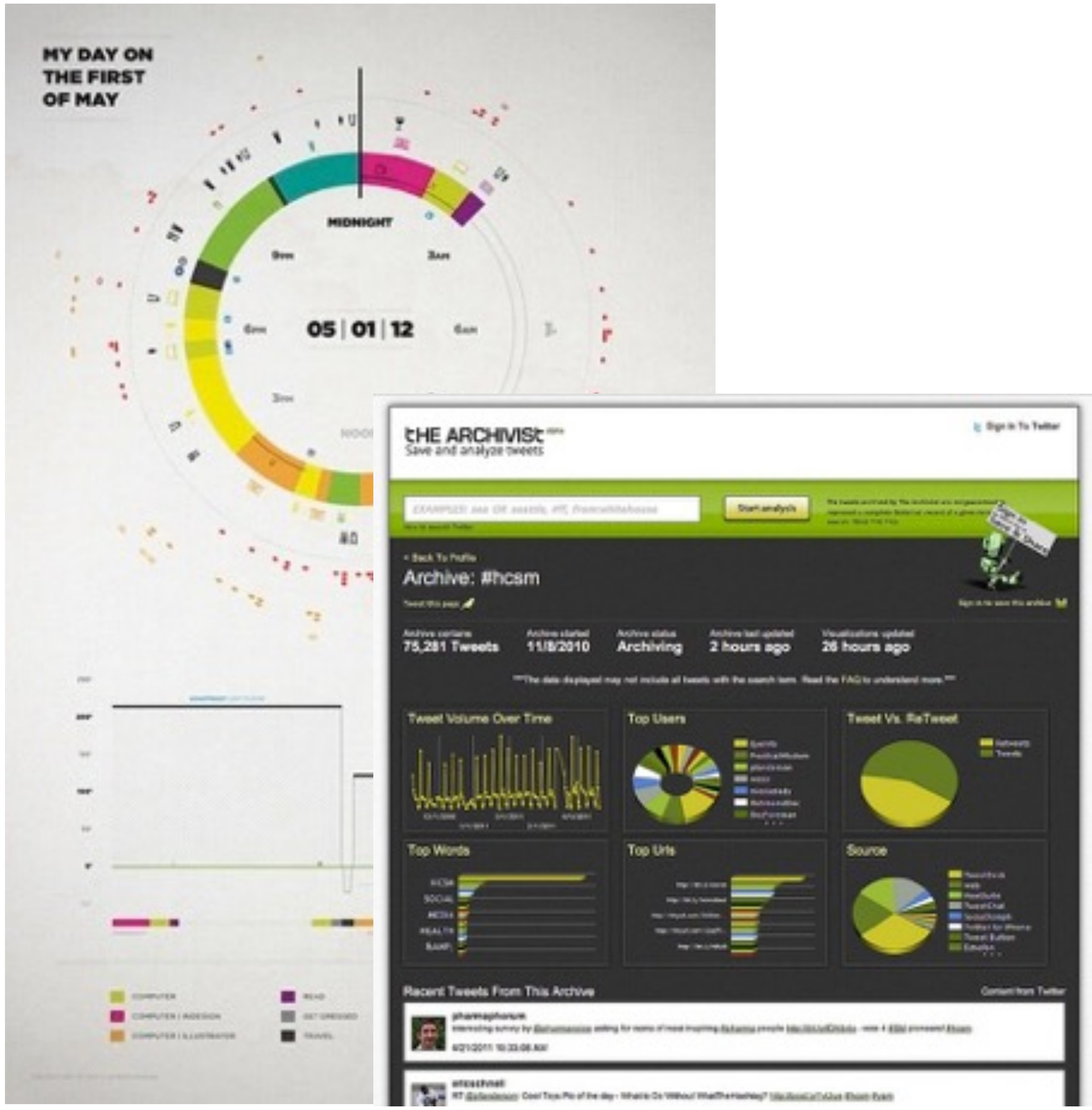
- What is discovered? By whom? Under whose control?
- Need for metadata usage analytics
- Empowering subjects: app stores?
- Permissions, social ratings and exchange

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

UNIVERSITY OF CAMBRIDGE

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
- Present and future public data
- What you have, what others want
- Editing of data; control of presentation to processors — *Recipient design*

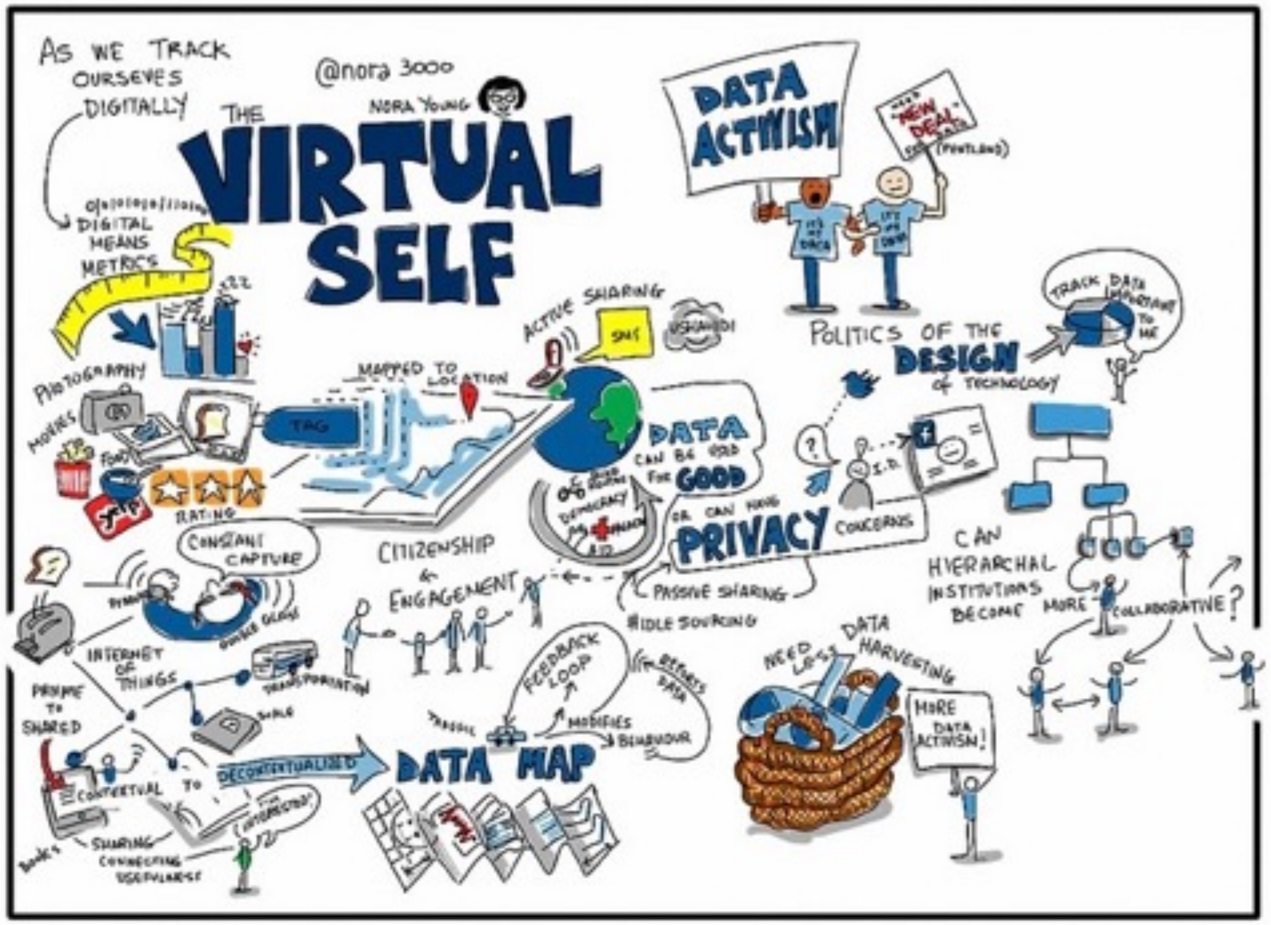
Interactional Challenges for HDI

From My Data to Our Data

- Delegating and revoking control
- Editing, viewing, sharing
- Group management, negotiated collection and control

https://flic.kr/p/drV8zY

Interactional Challenges for HDI



https://flic.kr/p/e57ySb

Salient Dimensions of Collaboration

- Transitivity: to whom is data passed, for what purpose
- Tracking and treatment

Thematic Areas for HDI

Personal data discovery

- Meta-data publication,
- Consumer analytics,
- Discoverability policies,
- Identity mechanisms, and
- App store models supporting discovery of data processers

Thematic Areas for HDI

Personal data ownership and control

- Group management of data sources,
- Negotiation,
- Delegation and transparency/awareness mechanisms, and
- Rights management

Thematic Areas for HDI

Personal data legibility

- Visualisation of what processors would take from data sources,
- Visualisations that help users make sense of data usage, and
- Recipient design to support data editing and data presentation

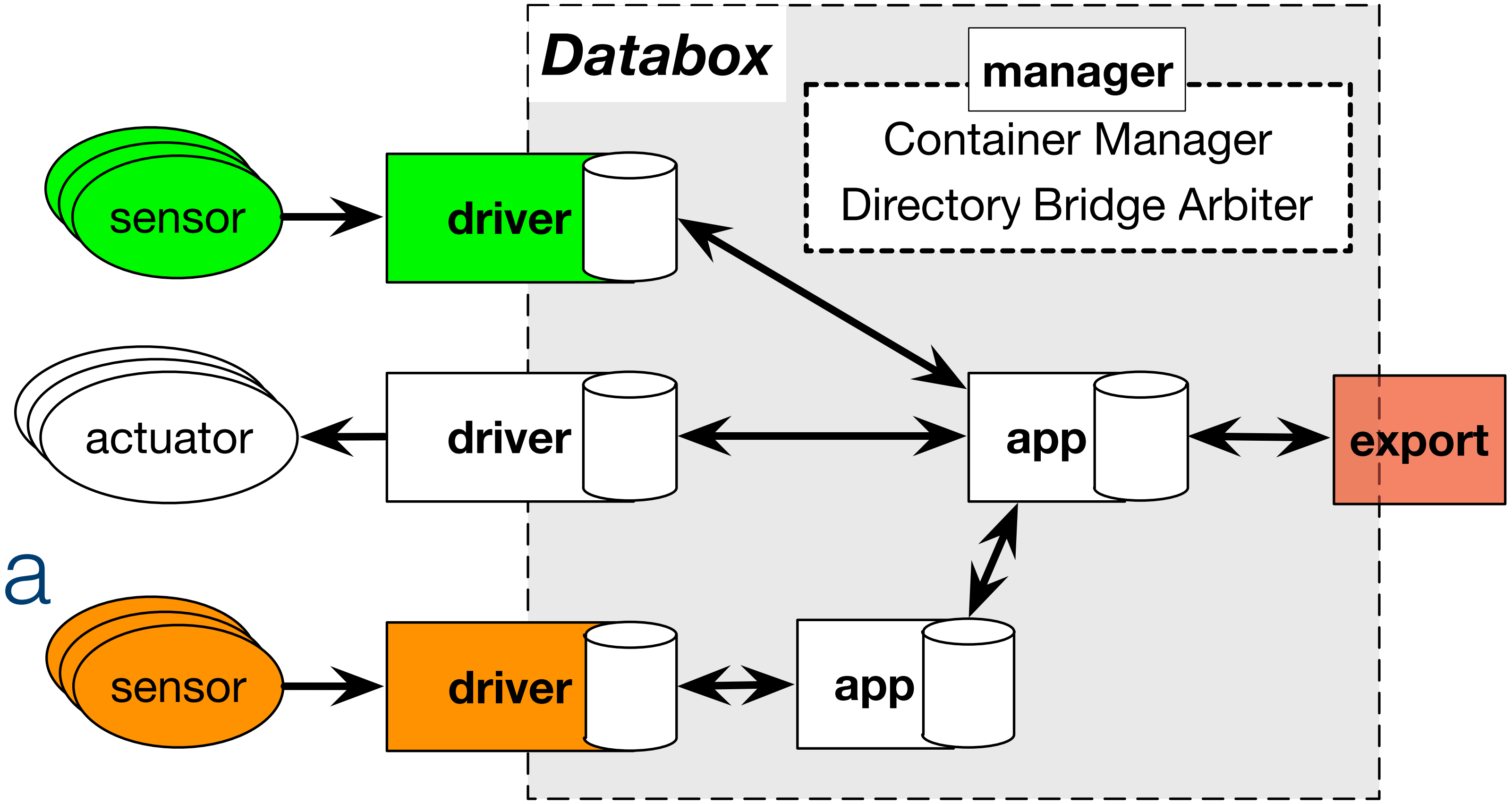
Thematic Areas for HDI

Personal data tracking

- Real time articulation of data sharing processes (e.g., current status reports and aggregated outputs), and
- Data tracking
  (e.g., subsequent consumer processing or data transfer)

Databox: Software Architecture

- Privacy preserving resource discovery
- Support existing development practices
- Control access to cloud originated data
- Network isolation of all datastores and legacy code



Databox: Physical Interactivity

- Physical devices often easier to reason about
  - Visible; Located; Proximate; Portable
- Physical access control is the norm
  - "The bag of keys" is well understood

- For example,
  - "when the grey tag is attached to my iPhone at home, the photos I take are shared with no-one; but when the grey tag is attached to my iPhone away from home, photos I take can be shared with family members"
  - "when the red tag is plugged into my Databox, none of my data may be accessed without direct permission from me"
  - "access to our smart meter data is allowed only when I have the green tag plugged into my Databox, and my wife has the green tag plugged into hers, or when one of our tags is plugged in and we're both in the house"

Databox: Distributed Analytics

- Subject driven *vs* Processor driven
  - App stores *vs* cohort discovery
- Cohort *vs* individual processing
  - Distributed model building
  - Personal local visualisation
- Challenges:
  - Scale, Heterogeneity, Dynamics

User-Centric Infrastructure

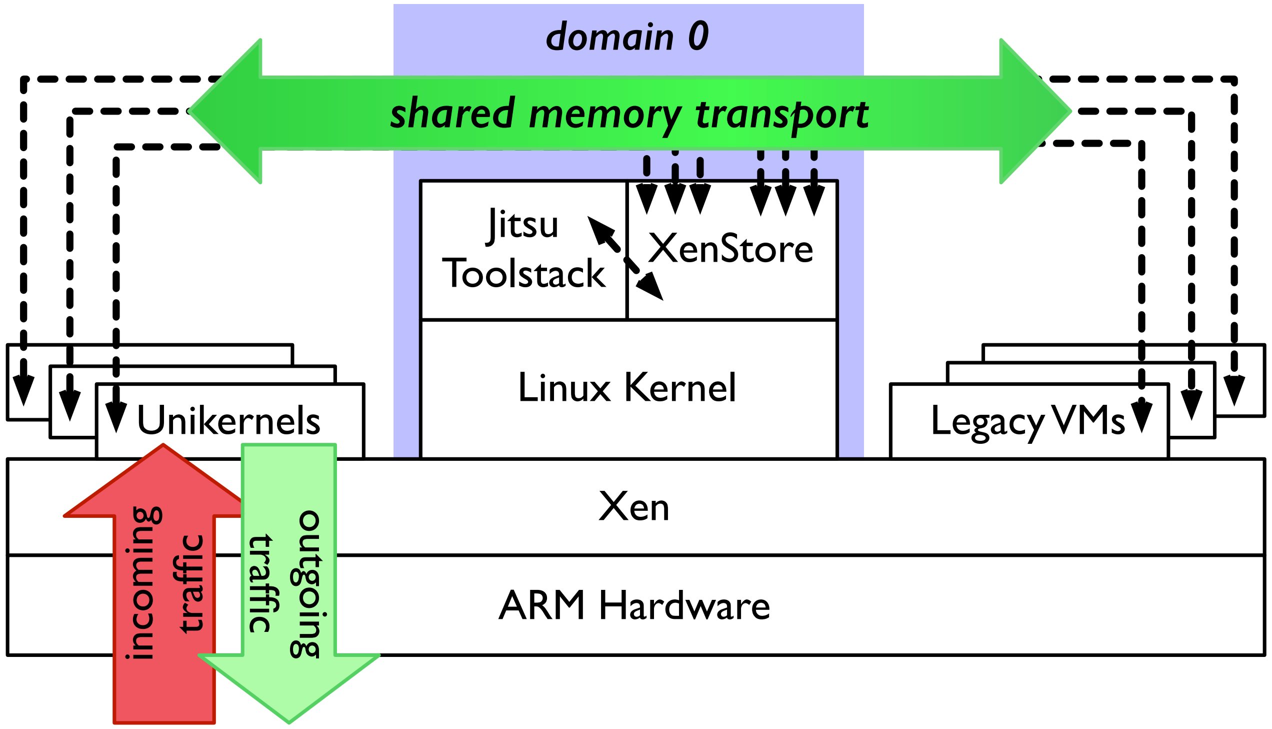
Stable, hidden, shared *vs*

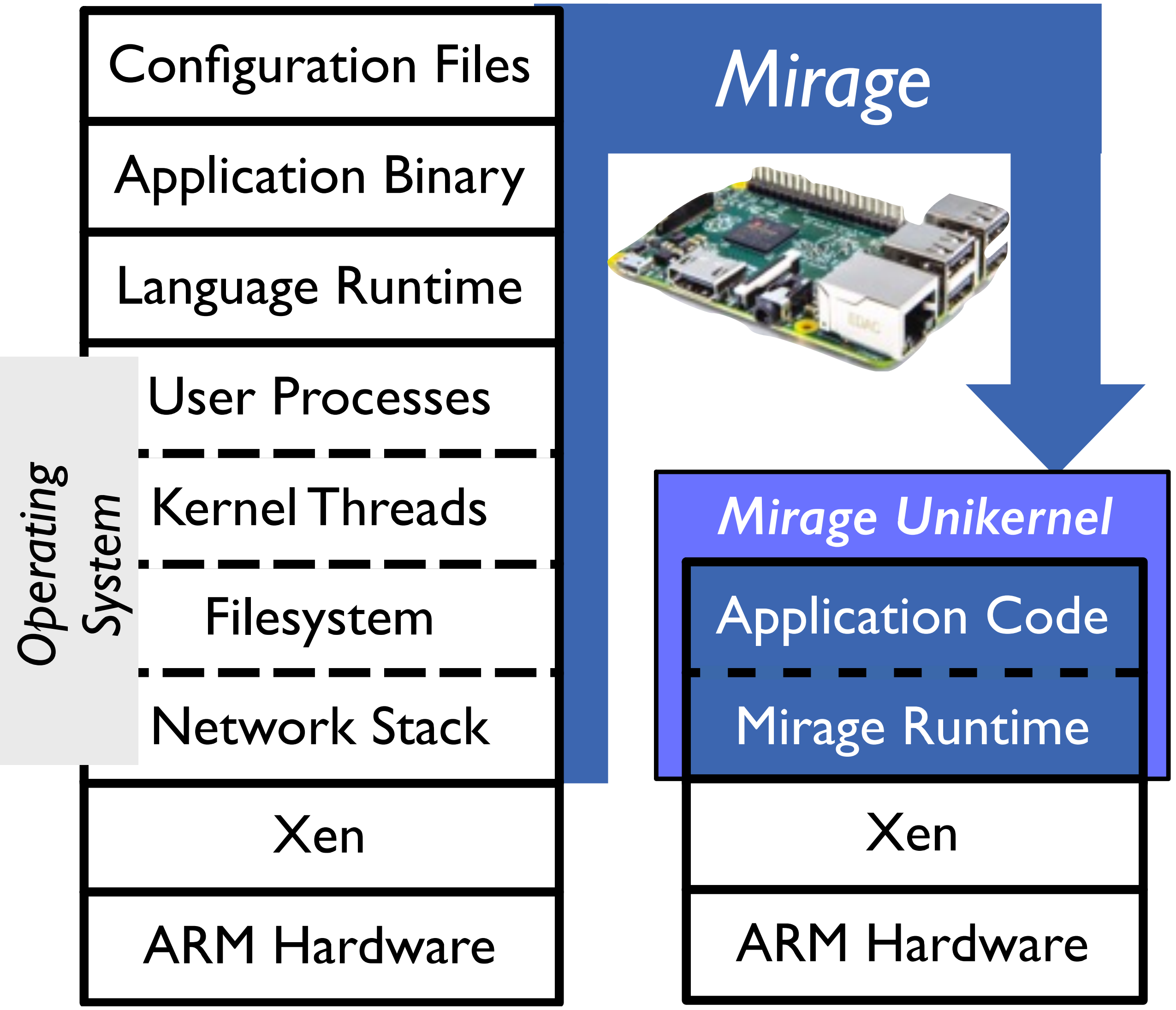
Dynamic, exposed, intimate

Personal Clouds

https://mirage.io/

- We should operate our own infrastructure
  …not abrogate our lives to "the cloud"
- Redesign OS infrastructure for network services to be run by **non-expert admins**





End Part II! Questions?

http://mort.io/

richard.mortier@cl.cam.ac.uk

http://hdiresearch.org/
http://homenetworks.ac.uk/
https://mirage.io/
https://forum.databoxproject.uk/

McAuley et al, COMSNETS'11
Haddadi et al, Aarhus'15
Crabtree & Mortier, ECSCW'15
Mortier et al, CAN'16 (in submission)

UNIVERSITY OF CAMBRIDGE

Particular thanks to Andy Crabtree, Hamed Haddadi, Derek McAuley. Work funded in part by EU FP7 611001, EPSRC EP/N028260/1, EP/N028422/1.