Personal Data: Thinking Inside the Box

Hamed Haddadi
Queen Mary
University of London
hamed@eecs.qmul.ac.uk

Amir Chaudhry, Jon Crowcroft, Heidi Howard, Anil Madhavapeddy, Richard Mortier
University of Cambridge
first.last@cl.cam.ac.uk

Derek McAuley
University of Nottingham
first.last@nottingham.ac.uk

ABSTRACT

We are in a 'personal data gold rush' driven by advertising being the primary revenue source for most online companies. These companies accumulate extensive personal data about individuals with minimal concern for us, the subjects of this process. This can cause many harms: privacy infringement, personal and professional embarrassment, restricted access to labour markets, restricted access to highest value pricing, and many others. There is a critical need to provide technologies that enable alternative practices, so that individuals can participate in the collection, management and consumption of their personal data. In this paper we discuss the Databox, a personal networked device (and associated services) that collates and mediates access to personal data, allowing us to recover control of our online lives. We hope the Databox is a first step to re-balancing power between us, the data subjects, and the corporations that collect and use our data.

Author Keywords

Personal Data; Computer Systems; Privacy

ACM Classification Keywords

Information systems applications; Computing platforms

THE PERSONAL DATA ECOSYSTEM

Many Internet businesses rely on extensive, rich data collected about their users, whether to target advertising effectively or as a product for sale to other parties. The powerful network externalities that exist in a rich dataset collected about a large set of users make it difficult for truly competitive markets to form. A concrete example can be seen in the increasing range and reach of the information collected about us by third-party websites, a space dominated by just two or three players [7]. This dominance has a detrimental effect on the wider ecosystem: online service vendors find themselves at the whim of large platform and API providers, hampering innovation and distorting markets.

Personal data management is considered an intensely personal matter however. Dourish argues that individual attitudes towards personal data and privacy are very complex and context dependent [5]. A recent three-year study showed that the more people disclosed on social media, the more privacy they said they desired (*We Want Privacy, but Can't Stop Sharing*, Kate Murphy, New York Times, 2014-10-05). This paradox implies dissatisfaction about what participants received in return for exposing so much about themselves online and yet, "*they continued to participate because they were afraid of being left out or judged by others as unplugged and unengaged losers*". This example also indicates the inherently social nature of much "personal" data: it is impractical to withdraw from all online activity just to protect one's privacy [3].

Context sensitivity, opacity of data collection and drawn inferences, trade of personal data between third parties and data aggregators, and recent data leaks and privacy infringements all motivate means to engage with and control our personal data portfolios. However, technical constraints that ignore the interests of advertisers and analytics providers and so remove or diminish revenues supporting our "free" services and applications, will fail [12, 25].

A host of other motivations and uses for such a Databox have been presented elsewhere [20, 17, 9]. These include privacy-conscious advertising, market research, personal archives, and analytical approaches to mental and physical health by combining data from different sources such as wearable devices, Electronic Health Records, and Online Social Networks. All these examples point to a need for individuals to have tools that allow them to take more explicit control over the collection and usage of their data and the information inferred from their online activities.

DATABOX: A USER-CENTRIC ALTERNATIVE

To address this need we propose the Databox, enabling individuals to coordinate the collection of their personal data, and to selectively and transiently make those data available for specific purposes. Following the Human-Data Interaction model [18], a Databox assists in provision of:

- **Legibility**: means to inspect and reflect on "our" data, to understand what is being collected and how it is processed.
- **Agency**: means to manage "our" data and access to it, enabling us to act effectively in these systems as we see fit.
- **Negotiability**: means to navigate data's social aspects, by interacting with other data subjects and their policies.

We do not envisage Databoxes entirely replacing dedicated, application-specific services such as Facebook and Gmail. Such sites that provide value will continue receiving personal data to process, in exchange for the services they offer.

Databox simply provides its user with means to understand, control and negotiate access by others to their data across a range of sites, including such currently dominant players. As a physical object it offers a range of affordances that purely virtual approaches cannot, such as located, physical interactions based on its position and the user's proximity.

Nor is the Databox oriented solely to *privacy* and prevention of activities involving personal data. It enables *new* applications that combine data from many silos to draw inferences presently unavailable. By redressing the extreme asymmetries in power relationships in the current personal data ecosystem, the Databox opens up a range of market and social approaches to how we conceive of, manage, cross-correlate and exploit "our" data to improve "our" lives.

FEATURES OF A DATABOX

What features must a Databox provide to achieve these aims? We answer in four parts: it must be a *trusted platform* providing facilities for *data management* of data at rest for the data subjects as well as *controlled access* by other parties wishing to use their data, and *supporting incentives* for all parties.

Trusted Platform. Your Databox coordinates, indexes, secures and manages data about you and generated by you. These data can remain in many locations, but it is the Databox that holds the index and delegates the means to access that data. It must thus be highly trusted: the range of data at its disposal is potentially far more intrusive – as well as more useful – when compared to data available to traditional data silos. Thus, although privacy is not the primary goal of the Databox, there are clear requirements on the implementation of the Databox to protect privacy [11]. Trust in the platform requires strong security, reliable behaviour and consistent availability. All of the Databox's actions and behaviours must be supported by pervasive logging, with associated tools, so that users and (potentially) third-party auditors can build trust that the system is operating as expected and, should something unforeseen happen, the results can at least be tracked. We envisage such a platform as having a physical component, perhaps in the form-factor of an augmented home broadband router, under the direct physical control of the individual. Thus, while making use of and collating data from remote cloud services, it would also manage data that the individual would not consider releasing to any remote cloud platform.

Data Management. A Databox must provide means for users to reflect upon the data it contains, enabling informed decision-making about their behaviours, and particularly whether to delegate access to others. As part of these interactions, and to support trust in the platform, users must be able to edit and delete data via their Databox as a way to handle the inevitable cases where bad data is discovered to have been inferred and distributed. Similarly, it may be appropriate for some data to *not* exhibit the usual digital tendency of perfect record. Means to enable the Databox automatically to forget data that are no longer relevant or have become untrue may increase trust in the platform by users [15]. Even if data has previously been used, it may still need to be "put beyond use" [2]. Concepts such as the European Union's *Right to be Forgotten* require adherence to agreed protocols and other forms of cooperation, by third-party services and data aggregators. The Databox can be used as a central point for negotiating such data access and release rights.

Controlled Access. Users must have fine-grained control over the data made available to third parties. At the very least, the Databox must be selectively queryable, though more complex possibilities include supporting privacy-preserving data analytics techniques, such as differential privacy [6] and homomorphic encryption [21]. A key feature of the Databox is its support for revocation of previously granted access. In systems where grant of access means that data can be copied elsewhere, it is effectively impossible to revoke access to the data accessed. In contrast, a Databox can grant access to process data locally *without* allowing copies to be taken of raw data unless that is explicitly part of the request. Subsequent access can thus easily be revoked [16]. A challenge is then to enable users to make informed decisions concerning the impact of releasing a given datum as this requires an understanding of the possible future information-states of all third parties that might access the newly released datum. One way to simplify this is to release data only after careful and irreversible aggregation of results to a degree that de-anonymisation becomes impossible. More complex decisions will require an on-going dialogue between the user and their Databox, to assist in understanding the impact of their decisions.

Supporting Incentives. A consequence of the controlled access envisioned above is that users may deny third-party services access to data. The Databox thus must enable services alternate means to charge the user: those who wish to pay through access to their data may do so, while those who do not may pay through more traditional financial means. One possible expression of this would be to enable the Databox to make payments, tracing them alongside data flows to and from different third-party services made available via some form of app store. Commercial incentives include having the Databox act as a gateway to personal data currently in other silos, and as an exposure reduction mechanism for commercial organisation. This removes their need to be directly responsible for personal data, with all the legal costs and constraints that entail, instead giving control over to the data subject. This is particularly relevant for international organisations that must be aware of many legal frameworks. A simple analogy is online stores' use of payment services (e.g., PayPal, Google Wallet) to avoid the overhead of Payment Card Infrastructure compliance.

BARRIERS TO ADOPTION

Many past systems provide some or all of these features, but none have really been successful due, we claim, to fundamental barriers that have yet to be coherently addressed.

Trust. The growth of third party cloud services means users must trust, not only the service they are directly using, but also any infrastructure providers involved, as well as other parties such as local law enforcement. If the Databox is to take such a central place in our online lives, it must be trusted. Two key aspects stand out here for the Databox: (*i*) the need to trust that it will protect the user against breach of data due to attacks such as inference across different datasets; and

(*ii*) the need to trust that the software running on it is trustworthy and not acting maliciously. Two features of the design support this. First, all keys remain with the user themselves such that not even a Databox provider can gain access without permission. Second, by making a physical artefact a key component in a user's Databox, e.g., a low-energy computing device hosted in their home, physical access controls – including turning it off or completely disconnecting it from all networks – can be applied to ensure data cannot leak. While this minimises the need to trust third parties, it increases trust placed in the software: we mitigate this by using open-source software built using modern languages (OCaml) on platforms (Xen) that limit the Trusted Computing Base, mitigating several classes of attack to which existing software is vulnerable.

Usability. Personal data is so complex and rich that treating it homogeneously is almost always a mistake and, as noted above, user preferences in this space are complex: socially derived and context dependent. A very broad range of intents and requirements must be captured and expressed in machine-actionable form. We will build on techniques developed in the Homework platform [19] which prototyped and deployed a range of novel task-specific interfaces that assisted users in the complex business of managing their home networks. Deciding which devices should be able to share in and access the digital footprint, even before considering sharing with other people, makes it even harder. Issues such as mixed, sometimes proprietary data formats, high variability in datum sizes, the multiplicity of standards for authentication to different systems to access data, lack of standard data processing pipelines and tools, and myriad other reasons make this job complex and time consuming. In addition, most data is inherently shared in that it implicates more than one individual and thus ownership and the concomitant right to grant access is not always clear. E.g., Use of cloud email services like Gmail: even if a user opts out by not using Gmail, there is a high chancethat a recipient of their email *is* using Gmail.

Cost. There are a range of incentives that must align for the success of such a platform. The day-to-day costs of running a Databox have to be acceptable to users. Similarly, costs that third-parties incur when accessing the system will have to be recouped, including perhaps recompensing for access to data that previously they would have simply gathered themselves. It remains to be seen how this can be done in practice: Are users willing and able to pay in practice? What will be the response of users when offered pay-for versions of previously free-to-use services? There is some evidence that some users will be willing to make this trade-off [1], but studies also show that the situation is complex [24].

THE WIDER ENVIRONMENT

In response to growing awareness about how our data is processed, many startups have formed in recent years aiming to put users explicitly in control of their personal data or metadata. They typically provide platforms through which users can permit advertisers and content providers to enjoy metered access to valuable personal data, e.g., OpenPDS [4]. In exchange, users may benefit by receiving a portion of the monetary value generated from their data as it is traded in an increasingly complex ecosystem [7]. Considering the churn experienced in the personal data startup space, with so many new but typically short-lived entrants, it seems that few truly viable models have yet been discovered. Our belief is that the power of personal data can only be realised when proper consideration is given to its social character, and it can be legibly combined with data from external sources. In this case, we might anticipate many potential business models [22].

Unfortunately, these approaches typically entail lodging *all* personal data into cloud-hosted services giving rise to concerns about privacy and future exploitation or mistaken release of data. In contrast, your Databox retains data – particularly *control over* that data – locally under your sole control. From a technology point of view, the general approach of Databox is that of "privacy by design", though it remains to be seen if it can be successful in a space such as this, where policy and technology need to co-evolve. In order to sell personal data, there needs to be a method for determining the marginal rate of substitution (the rate at which the consumer is willing to substitute one good for another) for personal data. The sale of personal data and insights derived from it is the key utility in this ecosystem, and individuals' preferences are the fundamental descriptors and success indicators.

Governments and regulatory bodies have attempted to impose regulatory frameworks that force the market to recognise certain individual rights. Unfortunately, legal systems are not sufficiently agile to keep up with the rapid pace of change in this area. Attempts at self-regulation such as the *Do Not Track* headers[1] are ineffective, with only an insignificant fraction of services in compliance [7]. It is even possible that there may be a shift towards consumer protection legislation, as opposed to current prevalence of informed consent [13].

SUMMARY

We are in an era where aggregation and analysis of personal data fuels large, centralised, online services. The many capabilities offered by these services have come at significant cost: lost of privacy and numerous other detrimental effects on the well-being of individuals and society. Many have commented that people simple do not see the need for technologies like this until they suffer some kind of harm from the exploitation of their data. On the other hand, it has been argued that privacy is negotiated through collective dynamics, and hence society reacts to the systems that are developed and released [10]. We speculate that data management may become a mundane activity in which we all engage, directly or through some representative, to a greater or lesser extent.

We have proposed the Databox as a technical platform that would provide means to redress this imbalance, placing the user in a position where they can understand, act and negotiate in this socio-technical system. By acting as a co-ordination point for your data, your Databox will provide means for you to *reflect* on your online presence, restore to you *agency* over your data, and enable a process of *negotiation* with other parties concerning your data. Even if the Databox as currently conceived is not a perfect solution, only

[1] http://donottrack.us/

by taking initial, practical steps can we elicit the necessary knowledge to improve the state-of-the-art. We do not believe further progress can be made without focused effort on the practical development and deployment of the technologies involved. E.g., Before addressing the complex problems of co-managing data [3], a Databox that enables personal data to be collated and reflected upon will allow individuals to explore workflows managing both their own data and, through ad hoc social interaction, data involving other stakeholders.

Thus we have begun development of underlying technologies for Databox: Nymote (http://nymote.org) and its constituent components of MirageOS [14], Irmin [8] and Signpost [23]. In addition, the community is developing methodologies for indexing and tracking the personal data held about us by third parties. However, the successful widespread deployment of such a platform will require that we tackle many significant issues of trust, usability, complexity and cost in ways that are transparent and scalable. Resolving questions such as those above requires that we develop and study Databoxes in-the-wild, in partnership with individuals, consumer rights groups, privacy advocates, the advertising industry, and regulators.

Acknowledgements. Work supported in part by the EU FP7 UCN project, grant agreement no 611001. Dr Haddadi visited Qatar Computing Research Institute during this work.

REFERENCES

1. Acquisti, A., John, L. K., and Loewenstein, G. What is privacy worth? *Journal of Legal Studies 42*, 2 (2013), 249–274.
2. Brown, I., and Laurie, B. Security against compelled disclosure. In *Proc. IEEE ACSAC* (Dec 2000), 2–10.
3. Crabtree, A., and Mortier, R. Human data interaction: Historical lessons from social studies and CSCW. In *Proc. ECSCW* (Oslo, Norway, Sept. 19–23 2015).
4. de Montjoye, Y.-A., Shmueli, E., Wang, S. S., and Pentland, A. S. openpds: Protecting the privacy of metadata through safeanswers. *PLoS ONE 9*, 7 (07 2014), e98790.
5. Dourish, P. What we talk about when we talk about context. *PUC 8*, 1 (Feb. 2004), 19–30.
6. Dwork, C. Differential privacy. In *Automata, Languages and Programming*, M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, Eds., vol. 4052 of *LNCS*. Springer, 2006, 1–12.
7. Falahrastegar, M., Haddadi, H., Uhlig, S., and Mortier, R. Anatomy of the third-party web tracking ecosystem. *CoRR abs/1409.1066* (2014).
8. Gazagnaire, T., Chaudhry, A., Crowcroft, J., Madhavapeddy, A., Mortier, R., Scott, D., Sheets, D., and Tsipenyuk, G. Irmin: a branch-consistent distributed library database. In *Proc. ICFP OCaml User and Developer Workshop* (Sept. 2014).
9. Guha, S., Reznichenko, A., Tang, K., Haddadi, H., and Francis, P. Serving ads from localhost for performance, privacy, and profit. In *ACM Workshop on Hot Topics in Networks* (2009).
10. Gürses, S. Can you engineer privacy? *Commun. ACM 57*, 8 (Aug. 2014), 20–23.
11. Haddadi, H., Hui, P., and Brown, I. Mobiad: private and scalable mobile advertising. Proc. ACM MobiArch (2010).
12. Leontiadis, I., Efstratiou, C., Picone, M., and Mascolo, C. Don't kill my ads! balancing privacy in an ad-supported mobile application market. Proc. ACM HotMobile (2012).
13. Luger, E., and Rodden, T. An informed view on consent for UbiComp. In *Proc. ACM UBICOMP* (2013), 529–538.
14. Madhavapeddy, A., Mortier, R., Rotsos, C., Scott, D., Singh, B., Gazagnaire, T., Smith, S., Hand, S., and Crowcroft, J. Unikernels: Library operating systems for the cloud. In *Proc. ACM ASPLOS* (Mar. 16–20 2013).
15. Mayer-Schonberger, V. *Delete: The Virtue of Forgetting in the Digital Age*. Princeton University Press, 2009.
16. McAuley, D., Mortier, R., and Goulding, J. The Dataware Manifesto. In *Proc. IEEE International Conf. on Communication Systems and Networks (COMSNETS)* (January 2011). Invited paper.
17. Mortier, R., Greenhalgh, C., McAuley, D., Spence, A., Madhavapeddy, A., Crowcroft, J., and Hand, S. The personal container, or your life in bits. *Proc. Digital Futures* (2010).
18. Mortier, R., Haddadi, H., Henderson, T., McAuley, D., and Crowcroft, J. Human-data interaction: The human face of the data-driven society. *SSRN* (Oct. 1 2014). http://dx.doi.org/10.2139/ssrn.2508051.
19. Mortier, R., Rodden, T., Tolmie, P., Lodge, T., Spencer, R., Crabtree, A., Sventek, J., and Koliousis, A. Homework: Putting interaction into the infrastructure. In *Proc. ACM UIST* (2012), 197–206.
20. Mun, M., Hao, S., Mishra, N., Shilton, K., Burke, J., Estrin, D., Hansen, M., and Govindan, R. Personal data vaults: A locus of control for personal data streams. In *Proc. ACM CoNEXT* (2010), 1–12.
21. Naehrig, M., Lauter, K., and Vaikuntanathan, V. Can homomorphic encryption be practical? In *Proc. ACM Cloud Computing Security Workshop* (2011), 113–124.
22. Ng, I. C. Engineering a Market for Personal Data: The Hub-of-all-Things (HAT), A Briefing Paper. *WMG Service Systems Working Paper Series* (2014).
23. Rotsos, C., Howard, H., Sheets, D., Mortier, R., Madhavapeddy, A., Chaudhry, A., and Crowcroft, J. Lost in the edge: Finding your way with Signposts. In *Proc. USENIX FOCI* (Aug. 13 2013).
24. Skatova, A., Johal, J., Houghton, R., Mortier, R., Bhandari, N., Lodge, T., Wagner, C., Goulding, J., Crowcroft, J., and Madhavapeddy, A. Perceived risks of personal data sharing. In *Proc. Digital Economy: Open Digital* (Nov. 2013).
25. Vallina-Rodriguez, N., Shah, J., Finamore, A., Grunenberger, Y., Papagiannaki, K., Haddadi, H., and Crowcroft, J. Commercial break: Characterizing mobile advertising. In *Proc. ACM IMC* (2012).