Challenges & Opportunities in Human-Data Interaction

Richard Mortier[1], Hamed Haddadi[2], Tristan Henderson[3], Derek McAuley[1], and Jon Crowcroft[4]

[1]University of Nottingham [2]Queen Mary University of London [3]University of St Andrews [4]University of Cambridge

ABSTRACT

As personal data are increasingly collected, analysed, and traded, we introduce the topic of *Human-Data Interaction (HDI)* to engage users with their data. HDI is inherently inter-disciplinary, encapsulating elements not only of traditional computer science such as data processing, systems design, visualisation and interaction design, but also of law, psychology, behavioural economics, and sociology. We outline HDI's motivation, its privacy and ethical challenges, and the opportunities it presents.

1. INTRODUCTION

Recent years have seen increasing collection and use of personal data – that is, data *about* us and data produced *by* us – both public and private, about us and our activities. Such data include purchasing habits (on- and off-line), financial data, communications data (from phone call records to social media content), and more. There has been similar growth in applications that provide us with benefits from our own publicly-released data: traffic reports on Google Maps, crowd-sourced road conditions on Waze [23], and optimised bus routes with mobile phone data [2].

The impact of this data processing is pervasive and wide-ranging – it informs credit ratings, online advertising, retailing, and is used for a variety of other predictions and inferences, from sexual orientation to voting preferences. These data are at the heart of many Internet business models, particularly those based on advertising and market intelligence.

An ecosystem, often collaborative but sometimes combative, is forming around companies and individuals engaging in the use of personal data. As reliance on these systems increases, we believe that people must be able to take more explicit control over the consumption of their data and the information they provide, and the exposition of their data to privacy-aware analytics and service providers. We propose placing the *human* at the centre of the data flows, and providing mechanisms for citizens to interact with these systems *explicitly*. Such an approach sits at the intersection of multiple disciplines, including computer science, statistics, sociology, psychology and behavioural economics and, we believe, deserves identification as a distinct topic we name **Human-Data Interaction** (HDI) [7]. In this paper we discuss some of the challenges and opportunities in HDI.

2. WHAT IS HDI?

We deliberately adopt the phrase HDI by analogy with HCI, but the two can be clearly distinguished. Unlike previous definitions of Human-Data Interaction [4, 6] focused on visualisation, primarily embodied, of large datasets, we believe that HDI concerns interaction generally between humans, datasets *and analytics*, but not the general study of interaction with computer systems that is HCI. HDI refers to the analysis of the individual and collective decisions we make and the actions we take, whether as users of online systems or as subjects of data collection. The term makes explicit the link between individuals and the *signals* they emit as data (e.g., location, shopping trends, search terms), as the richness, pervasiveness and impact of these models and techniques continues to grow.

HDI includes the combination of data *and* the algorithms used to analyse them. HDI aims to understand both raw and derived data *out there* about individuals, the ways in which and by whom they are used, and how people might desire and act to influence—and ideally benefit—from the data and their use.

Figure 1 characterises current systems. Analytics are provided as a "black box" within which collated input data are processed in large centralised facilities (data centres). The inferences output by this processing then cause actions, which may include feeding inferences into subsequent analysis by others. We see two key points in this cycle where greater transparency to and control by subjects is needed.

First, the analytics algorithms themselves must become less opaque. What data are they consuming? What methods are they using to draw inferences? This is often in direct conflict with the fact that these processes represent core intellectual property of the companies that implement and run them, and so cannot easily be made public.

Second, people need to be given control over the inferences that are drawn and the actions that these inferences inform. These systems are large and complex, and although they can affect us all, many of their individual effects will be positive or insufficiently negative to be noticeable. The problem then becomes how to engage people with such complex and mostly uninteresting systems before they suffer harm.

3. WHY IS HDI INTERESTING?

There are two features that make HDI interesting. First, as recent experience with online social networks and the NSA's PRISM have shown, the impact of the inferences drawn from public personal data can affect the market value of billion dollar corporations or move the use of national infrastructure outside expected parameters. Second, inferences drawn from on- and off-line private personal data, such as passive measurement, location, and communications, create *virtual personalities* for each individual. Thus HDI contains a simultaneous mix of two contrasting features: sheer scale and personal richness. Combined, these features create a complex system that poses challenges at many levels:

- **Visualisation and sense-making**. How are people to make sense of such complex, technical systems?
- **Transparency and audit**. What audit trails and information must be provided to support this?

- **Privacy and control**. How can the resulting audit data be used to enable interaction around control of access to and processing of data?
- **Analytics and commerce**. How can analysis algorithms be made transparent to users while retaining their protected commercial status?
- **Data to knowledge**. How can the vast amount of data be used to both benefit individuals and let society exploit the wealth of information offered by shared data?

4. WHAT CAN WE DO ABOUT IT?

What are the implications of HDI? How should researchers engage with HDI? The following existing domains clearly intersect with HDI; no doubt there are many others.

HCI & Data Visualisation. As discussed above, HCI clearly overlaps with HDI, particularly in topics relating to data visualisation. Many of the concepts and datasets that motivate the need for HDI are rather abstract: enabling subjects to interact naturally with their data and the algorithms processing them is an important goal. Some existing work in this space uses embodied allegories to support the design of meaningful Embodied Interactions [6, 4].

Analytics. For industry, perhaps the most important aspect of personal data use is analytics. As new sources of *Big Data* arise, characterised by volume, variety, and velocity [5], new analytics methods will arise, with resultant effects for how users should interact and interpret these data. One growing source of such data is the Internet of Things (IoT), which will create increasing amounts of ambient data from our urban environment [25]. Recent interest in the *Quantified Self* [20] also relies on data collation and analytics about physical activity, dietary data, sleep patterns, and so forth.

Privacy & Security. Outdated regulation, coupled with cultural differences in service provision and online behaviour, has resulted in a wave of strong user reactions in response to political events and industrial developments in the *Big Data Analytics* era. In reaction to this trend, individuals, governments, privacy advocates, industry, and regulators have been fiercely fighting their corners concerning collection, use, trade, and retention of personal information.

Social Psychology. Individuals' decision making can be manipulated in many ways, e.g., by altering the choices available and the order in which they are presented [22]. Interaction with online content is also affected by the way in which information is presented. Even in population-scale Big Data industries, the human factors of judging ambiguities and cross-referencing terms across social and cultural boundaries remain key elements [13].

Behavioural Economics. Access to the Internet is increasingly seen as a human right [24]. The Internet's open, non-discriminatory shared nature has thus been of central interest to a number of advocacy groups. Changes in access to data often cause Internet activism, sometimes leading to political and regulatory change, e.g., the 2013 ITU voting case on Internet Governance [10, 15]. We need to understand how recent behavioural targeting advances in advertising have affected the personal data collection ecosystem.

5. FUTURE DIRECTIONS

Our thesis is that HDI is worthy of treatment as a distinct topic of research, and this review has covered a number of facets of the HDI ecosystem. Though HDI is not necessarily about *Big Data*, it depends indirectly on understanding the potential biases and inaccuracies in Big Data where it concerns people, where the sheer quantity of data is sometimes confused with quality [3]. The current regulatory situation around use of *big* personal data is far from acceptable, a point to which bodies such as the EU and the UK Parliament are now beginning to respond [8]. Ultimately, HDI places humans in their rightful place, not as mere stakeholders in this system but at its very centre. Study of HDI thus provides a framework within which to address many related issues, for example:

- developing mechanisms to improve data quality and data processing algorithms, and to give people control over lifetime, scope and visibility of their personal data;
- as a pre-requisite for many such mechanisms, how to make our data available in a *machine-friendly* form so that it can readily be processed by code rather than only inspected visually via webpages; note that challenges here include not only how best to structure and represent such heterogeneous data, but also issues concerning licensing and informed consent in giving others access to our personal data, where we can benefit from releasing such data;
- realising *Personal APIs* [17], enabling voluntary participation in information marketplaces [1, 11];
- reconciling such use and control of personal data with a regulatory push to Open Data [19];
- creating and promoting novel approaches to use of shared personal data to offer insight and information both to individuals and society, while respecting privacy;
- understanding the many complex and subtle ethical and legal issues surrounding use of big personal data, giving meaning to mechanisms such as the right to be forgotten;
- addressing the broader societal implications of having such rich personal data available at scale, able to be gossiped across the globe in milliseconds; in particular, how we can build geo-social controls over visibility of our data to help people avoid offence, embarrassment and worse;
- reworking conceptions of informed consent from its current intolerable state [9, 14], supporting the regulatory push for transparency into value of personal data in the information economy;
- understanding the *contextual integrity* [18] of uses of our personal data, and how this impacts services [21] and new uses of our data both for research and business [16];
- and, ultimately, stopping the downward trajectory of economic value in the information age [12], avoiding disproportionate economic power concentrating in the data aggregators' hands.

The way many services are currently deployed and monetised encourages us all to trade eyeball time for "free" services, resulting in the enormous valuations accorded companies such as Facebook and Google due to the massive quantities of data about us they already have and continue to accumulate. Addressing the above challenges can begin to level the playing field between us, the users farmed for our data, and our would-be data overlords who gather and exploit our data.

Acknowledgements. We thank Ian Brown, Laura James, Tom Rodden, Dirk Trossen, and QMUL Cognitive Science research group members for their feedback on earlier drafts of this paper; and RCUK (Horizon hub, EP/G065802/1; CREATe, AH/K000179/1; and IT as a Utility Network+, EP/K003569/1) for funding.

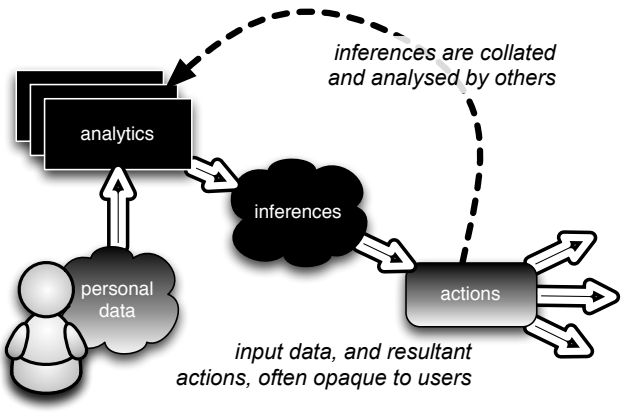

Figure 1: Human-Data Interaction. *Personal data* about and by each of us, whether we are aware of it or not, feeds into black-box *analytics* algorithms. These output *inferences* driving *actions* whose effects may or may not be visible to us.

6. REFERENCES

[1] C. Aperjis and B. A. Huberman. A market for unbiased private data: Paying individuals according to their privacy attitudes. *First Monday*, 17(5-7), May 2012. doi:10.5210/fm.v17i5.4013.

[2] M. Berlingerio, F. Calabrese, G. Di Lorenzo, R. Nair, F. Pinelli, and M. L. Sbodio. AllAboard: a system for exploring urban mobility and optimizing public transport using cellphone data. In *Proc. ECML PKDD*, Sept. 2013. doi:10.1007/978-3-642-40994-3_50.

[3] D. Boyd and K. Crawford. Critical questions for big data. *Information, Communication & Society*, 15(5):662–679, May 2012. doi:10.1080/1369118x.2012.678878.

[4] F. Cafaro. Using embodied allegories to design gesture suites for human-data interaction. In *Proc. UbiComp*, pages 560–563, 2012. doi:10.1145/2370216.2370309.

[5] C. Eaton, D. DeRoos, T. Deutsch, G. Lapis, and P. Zikopoulos. *Understanding Big Data*. McGraw-Hill, 2012.

[6] N. Elmqvist. Embodied Human-Data Interaction. In *Proc. CHI Workshop "Embodied Interaction: Theory and Practice in HCI"*, pages 104–107, May 2011. Online at http://www.antle.iat.sfu.ca/chi2011_EmbodiedWorkshop/Papers/NiklasElmqvist_CHI11EIWkshp_EmbodiedHuman-DataInteraction.pdf.

[7] H. Haddadi, R. Mortier, D. McAuley, and J. Crowcroft. Human-data interaction. Technical Report UCAM-CL-TR-837, Computer Laboratory, University of Cambridge, June 2013. Online at http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-837.pdf.

[8] House of Commons Public Administration Select Committee. Government and IT — "a recipe for rip-offs": time for a new approach. Twelfth Report of Session 2010-12, 28 July 2011. Online at http://www.publications.parliament.uk/pa/cm201012/cmselect/cmpubadm/715/715i.pdf.

[9] J. P. A. Ioannidis. Informed consent, big data, and the oxymoron of research that is not research. *American Journal of Bioethics*, 13(4):40–42, 20 Mar. 2013. doi:10.1080/15265161.2013.768864.

[10] ITU. International telecommunications regulations. http://www.itu.int/ITU-T/itr/.

[11] B. Kamleitner, S. Dickert, M. Falahrastegar, and H. Haddadi. Information bazaar: a contextual evaluation. In *Proc. HotPlanet*, pages 57–62, Aug. 2013. doi:10.1145/2491159.2491161.

[12] J. Lanier. *Who Owns The Future?* Simon & Schuster, 2013.

[13] S. Lohr. Algorithms get a human hand in steering Web. *New York Times*, page A1, 10 Mar. 2013. Online at http://www.nytimes.com/2013/03/11/technology/computer-algorithms-rely-increasingly-on-human-helpers.html.

[14] E. Luger, S. Moran, and T. Rodden. Consent for all: Revealing the hidden complexity of terms and conditions. In *Proc. CHI*, May 2013. doi:10.1145/2470654.2481371.

[15] D. McCullagh. U.N. summit rejects U.S., Europe hands-off-the-Internet plea. *CNET*, 12 Dec. 2012. Online at http://news.cnet.com/8301-13578_3-57558887-38/u.n-summit-rejects-u.s-europe-hands-off-the-internet-plea/.

[16] S. McNeilly, L. Hutton, and T. Henderson. Understanding ethical concerns in social media privacy studies. In *Proc. ACM CSCW Workshop on Measuring Networked Social Privacy: Qualitative & Quantitative Approaches*, Feb. 2013. Online at http://networkedprivacy2013.files.wordpress.com/2013/01/hutton-networkedprivacy2013.pdf.

[17] G. Meyer. Revisiting the API of me, June 2013. http://gregmeyer.com/2013/06/10/revisiting-the-api-of-me/.

[18] H. F. Nissenbaum. Privacy as contextual integrity. *Washington Law Review*, 79(1):119–157, Feb. 2004. Online at http://ssrn.com/abstract=534622.

[19] Open Knowledge Foundation. Open data & my data, Feb. 2013. http://blog.okfn.org/2013/02/22/open-data-my-data/.

[20] Quantified Self. http://quantifiedself.com/.

[21] K. Shilton, J. Burke, D. Estrin, R. Govindan, M. Hansen, J. Kang, and M. Mun. Designing the personal data stream: Enabling participatory privacy in mobile personal sensing. In *Proc. TPRC*, Sept. 2009. Online at http://ssrn.com/abstract=1999839.

[22] R. H. Thaler and C. R. Sunstein. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Penguin Books, 2008.

[23] Waze. http://www.waze.com/.

[24] S. B. Wicker and S. M. Santoso. Access to the Internet is a human right. *Commun. ACM*, 56(6):43–46, June 2013. doi:10.1145/2461256.2461271.

[25] A. Zaslavsky, C. Perera, and D. Georgakopoulos. Sensing as a Service and Big Data. In *Proc. ACC*, July 2012. Online at http://arxiv.org/abs/1301.0159.