

Big data for humans or humans for big data?: a human-data interaction perspective

Shin'ichi Konomi¹

¹HDI Lab, Faculty of Arts and Science, Kyushu University, 744, Motooka, Nishi-ku, Fukuoka 819-0395, JAPAN

Abstract

Designing "big data for humans" would require so-called *human-data interaction*. In this paper, we discuss key dimensions of human-data interaction to enable a look at the field from a broader perspective and facilitate developments of "big data for humans". Our discussion is based on the relevant research projects in our group at the intersections of human-data interaction and recommendation and search, pervasive computing, civic computing and learning analytics.

Keywords

Human-data interaction, human-centered big data, calm technology, data science

1. Introduction

Bell and Gray (1997) predicted that all information about physical objects, humans, buildings, processes, and organizations will be online by 2047 [1]. Twenty five years have passed since their prediction, and there are only 25 years left before the possible dawn of the fully datafied world according to their prediction. By 2025, it's estimated that 463 exabytes of data will be generated each day globally [2].

The sheer volume, variety and velocity of the ever-increasing data can easily create the situations of information overload. Quick fixes for the information overload problem often rely on straightforward automation, which may fail to fit human needs in different contexts. Going beyond such myopic approaches would require smartness at a different level to embed right opportunities for people to interact with and intervene big-data systems at the right time and in the right way. This can be a key step towards the design of *calm technology* [3].

Having people involved in big-data environments requires *human-data interaction*. Human-data interaction (HDI) is an emerging field of interdisciplinary inquiry that is concerned with understanding and developing technologies for supporting human interactions with digital data. Such interactions may occur in the contexts of data collection, data wrangling, algorithm design, analytics, visualization, recommendation, classification, prediction, interpretation, and so on.

Proceedings of CoPDA2022 - Sixth International Workshop on Cultures of Participation in the Digital Age: AI for Humans or Humans for AI? June 7, 2022, Frascati (RM), Italy

✉ konomi@artsci.kyushu-u.ac.jp (S. Konomi)

🌐 <http://hdi.ait.kyushu-u.ac.jp/> (S. Konomi)

🆔 0000-0001-5831-2152 (S. Konomi)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

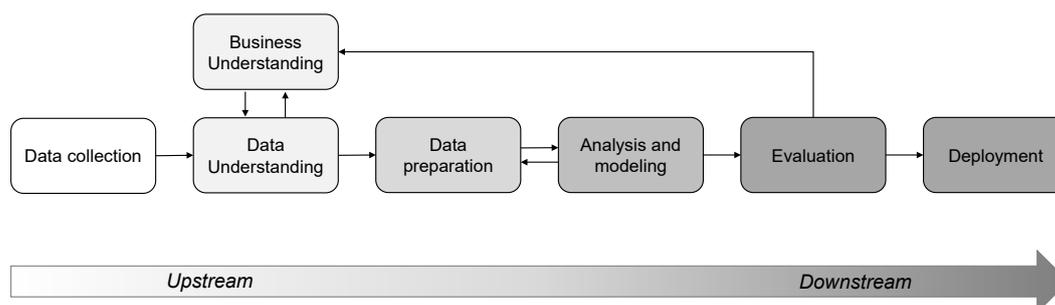


Figure 1: A process for using big data inspired by CRISP-DM [19]

Human-data interaction emphasizes the human-centered approach and existing works in this field focuses on its different facets [4]. Mortier, Haddadi, Henderson, McAuley and Crowcroft discuss human-data interaction with their proposal to place humans at the center of the flows of data, and provision of the mechanisms for citizens to interact with these systems and data explicitly [5]. They also propose and elaborate on the three core themes relevant to human-data interaction, namely, *legibility*, *agency*, and *negotiability*. Crabtree and Mortier discuss human-data interaction from social and interactional perspectives, and look at the need to develop social models and mechanisms of data sharing that *enable users to play an active role in the process* [6]. Mashhadi, Kawsar and Acer draw our attention to the importance of human-data interaction in Internet of Things environments with ubiquitous devices [7]. Cabitza and Locoro discuss healthcare data through the lens of human-data interaction [8]. Other studies look into embodied interactions for exploring large data sets [9], and a media service that exploits personal data to provide content recommendations [10].

In this paper, we discuss key dimensions of human-data interaction to enable a look at the field from a broader perspective and facilitate developments of "big data for humans". Our discussion is based on the relevant research projects in our group at the intersections of human-data interaction and recommendation and search, pervasive computing, civic computing and learning analytics.

2. Three key dimensions of human-data interaction

In this section, we introduce the three dimensions for classifying human-data interaction environments. We identified these dimensions based on a survey of related works [4, 5, 6, 7, 8, 9, 10], our own experiences with relevant projects [11, 12, 13, 14, 15, 16, 17, 18] as well as an existing process model for data science [19]. Table 1 shows these three dimensions in a tabular format.

The first dimension concerns with the process for using big data (see Figure 1). The process starts with data collection, followed by data understanding, data preparation through data wrangling, analysis and modeling via visualization and/or machine learning algorithms, evaluation, and deployment of the resulting model or actions based on the gained insights. For

Table 1
The three key dimensions of HDI.

	Personal data		Public data	
	Upstream	Downstream	Upstream	Downstream
Synchronous	<i>Real-time</i> interaction with <i>personal</i> data at <i>upstream</i> steps (e.g., Collecting personal health data interactively)	<i>Real-time</i> interaction with <i>personal</i> data at <i>downstream</i> steps (e.g., Interactive analysis of data in personal informatics)	<i>Real-time</i> interaction with <i>public</i> data at <i>upstream</i> steps (e.g., Collecting urban public data interactively)	<i>Real-time</i> interaction with <i>public</i> data at <i>downstream</i> steps (e.g., Interactive analysis of urban public data sets, possibly using an embodied interaction interface)
Asynchronous	<i>Long-term</i> interaction with <i>personal</i> data at <i>upstream</i> steps (e.g., Collecting personal health data automatically and use it at a later point in time. Improving data collection to address privacy issues.)	<i>Long-term</i> interaction with <i>personal</i> data at <i>downstream</i> steps (e.g., Personalized news recommendation based on an incrementally improved machine-learning model)	<i>Long-term</i> interaction with <i>public</i> data at <i>upstream</i> steps (e.g., Collecting urban public data automatically and use it at a later point in time. Improving data collection to address ethical issues.)	<i>Long-term</i> interaction with <i>public</i> data at <i>downstream</i> steps (e.g., Non-personalized recommendation of popular news based on an incrementally improved machine-learning model)

example, human-data interaction can take place *downstream* in this process during the analysis and modeling phase by using interactive visualization tools. In other cases, it can take place *upstream* during data collection phase by turning on and off GPS tracking on one's smart phone. This *upstream-downstream* dimension captures the point of human-data interaction in this process, and allows us to consider the differences of human-data interaction accordingly.

The second is the *personal-public* dimension that concerns with the characteristics of data with which people interact. For example, embodied interaction with *public* data sets in a VR environment is *public* in this dimension, whereas *personal* news recommendation systems may use *personal* data about people. This dimension allows us to consider different concerns around the interaction with *public* and *personal* data.

The third is the *synchronous-asynchronous* dimension that concerns with the time aspects of human-data interaction. This dimension distinguishes the different modes of human-data interaction in a similar way as the *synchronous-asynchronous* classification of computer-supported cooperative work environments. For example, interactive analysis of data sets using an em-

Table 2

Classification of existing systems for *recommendation and search*, *pervasive computing*, *civic computing*, and *learning analytics*.

	Personal data		Public data	
	Upstream	Downstream	Upstream	Downstream
Synchronous	- Deai Explorer[12]	- Deai Explorer[12]	- Askus[13] - Vacant House[17] - Community Reminder[14]	- CourseQ[11] - Deai Explorer[12]
Asynchronous	- Learning Analytics for All[16]	- e-Book Reading Analytics[15]	- Vacant House[17] - Community Reminder[14]	- Co-location networks [18]

bodied interaction interface falls into the *synchronous* category. When people improve the behaviors of a recommendation system by changing some preference settings or by replacing its algorithm with a more privacy-preserving and less biased one, such interactions can be considered as *asynchronous*.

3. Case studies to explore the dimensions

We next look further into the proposed dimensions of human-data interaction based on several existing systems, which have been developed by our group. The purposes of the systems include *recommendation and search*, *pervasive computing*, *civic computing*, and *learning analytics*. Their HDI features can be classified into different categories as shown in Table 2.

3.1. Recommendation and search

CourseQ [11] is a course recommendation system for university students based on a syllabus data set and a topic modeling-based algorithm. Although many existing course recommendation systems focus on the accuracy of recommendation, they may fail to recommend the courses that the students feel truly relevant. We introduced various interactive features in CourseQ so as to improve user-centric metrics such as user acceptance as well as understandability of recommendation results.

The interactive features of CourseQ include keyword-based search and filtering, interactive visualization of recommended courses, dynamic presentation of relevant auxiliary information and explanation with the recommended results, and a 'like' button. These features mainly support synchronous interactions with the publicly available data set, however, the interactivity does not allow users to change the data sets and other elements in the upstream process. CourseQ thus provides synchronous downstream human-data interaction based on public data.

3.2. Pervasive computing

DeaiExplorer [12] is a social-network display that responds to RFID badges carried by conference participants and displays social connections between colocated conference participants. The system exploits a public data set from a publication database as well as data collected from RFID readers based on participants' agreement. The system visualizes social network structures based on these two types of data in order to facilitate social interactions among conference participants. The interactive feature of DeaiExplorer allows conference participants to access their social network visualizations just by showing their RFID badges to the RFID reader. This feature allows users to interact with public and private data in a synchronous manner. As this interactivity allows users to control the capture of their RFID data by (not) showing badges to the RFID reader, the system allows synchronous human-data interaction in both upstream and downstream processes.

Askus [13] is a type of so-called *participatory sensing systems*, which allows users to collect data manually by using mobile phones. It thus concerns with the upstream process, and mainly supports synchronous human-data interaction with public environmental data, etc.

Co-location networks [18] analyze urban mobility data sets based on network analysis techniques. This analysis was performed multiple times based on an iterative improvements of network analysis techniques. It thus concerns with asynchronous interaction with public data in the downstream process.

3.3. Civic computing

Our WiFi-based sensing tool to predict vacant houses [17] is a type of so-called *opportunistic sensing systems*, which allows local community members to collect data automatically by just walking around in their community with their mobile phones in their backpacks. Users' choices of walking routes can control data collection in a synchronous manner. Data collection can be controlled asynchronously by changing the setting of WiFi sensing software. It thus concerns with the upstream process, and mainly supports synchronous and asynchronous human-data interaction with WiFi signals in public spaces.

Community Reminder [14] is also a type of *participatory sensing systems*, which allows local community members to collect information about the safety in their communities using mobile phones. Local community members can also participate in the design of the data collection mechanisms of this system using an intuitive tangible user interface. This system concerns with both synchronous and asynchronous aspects of public data collection.

3.4. Learning analytics

Our research projects in this area include analysis of e-book reading patterns [15] as well as an effort to provide learning analytics for all age groups and in developing communities without reliable internet access [16]. The former analysis can be performed in an iterative manner based on different research questions and techniques. It thus concerns with asynchronous interaction with personal data in the downstream process. The interactive features of the latter includes delayed data transmission of learning log data using mobile phones[16]. It thus concerns with asynchronous interaction with personal data in the upstream process.

4. Discussion and conclusion

We introduced the three key dimensions for classifying human-data interaction environments, i.e., the *upstream-downstream*, *personal-public*, and *synchronous-asynchronous* dimensions. Existing research projects tend to focus on one or more areas with respect to these dimensions.

The discussions of the several existing systems for recommendation and search, pervasive computing, civic computing, and learning analytics provided an opportunity for a further look into the proposed HDI dimensions, and demonstrated how they can highlight commonalities and differences of various human-data interaction systems.

Interaction is everywhere in the space defined by the proposed dimensions. Thinking about big data systems from the perspectives enabled by these dimensions would help us analyze and/or design a broader range of big data and AI systems with humans at the center and their data interactions in mind. It reminds the designers and the users of such systems that they are embedded in larger contexts, and the importance of providing the right opportunities for humans to play active roles in the context. One of the major advantages of emphasizing interactions in big data and AI systems can be, as our experiences with CourseQ [11] suggests, people's increased trust with data-centric smart mechanisms such as recommender systems, which could in turn lead to people's improved satisfaction with such systems.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP20H00622.

References

- [1] G. Bell, J. N. Gray, The revolution yet to happen, in: *Beyond Calculation*, Springer, 1997, pp. 5–32.
- [2] How much data is generated each day? | world economic forum, 2019. URL: <https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/>.
- [3] M. Weiser, J. S. Brown, The coming age of calm technology, *Beyond Calculation* (1997) 75–85. URL: https://link.springer.com/chapter/10.1007/978-1-4612-0685-9_6. doi:10.1007/978-1-4612-0685-9_6.
- [4] E. Z. Victorelli, J. C. Dos Reis, H. Hornung, A. B. Prado, Understanding human-data interaction: Literature review and recommendations for design, *International Journal of Human-Computer Studies* 134 (2020) 13–32.
- [5] R. Mortier, H. Haddadi, T. Henderson, D. McAuley, J. Crowcroft, Human-data interaction: The human face of the data-driven society, Available at SSRN 2508051 (2014).
- [6] A. Crabtree, R. Mortier, Human data interaction: historical lessons from social studies and cscw, in: *ECSCW 2015: Proceedings of the 14th European Conference on Computer Supported Cooperative Work*, 19-23 September 2015, Oslo, Norway, Springer, 2015, pp. 3–21.

- [7] A. Mashhadi, F. Kawsar, U. G. Acer, Human data interaction in iot: The ownership aspect, in: 2014 IEEE world forum on Internet of Things (WF-IoT), IEEE, 2014, pp. 159–162.
- [8] F. Cabitza, A. Locoro, Human-data interaction in healthcare, *Smart Technology Applications in Business Environments* (2017) 184–203. doi:10.4018/978-1-5225-2492-2.CH009.
- [9] M. Trajkova, A. Alhakamy, F. Cafaro, R. Mallappa, S. R. Kankara, Move your body: Engaging museum visitors with human-data interaction, in: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–13.
- [10] N. Sailaja, R. Jones, D. McAuley, Designing for human data interaction in data-driven media experiences, in: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–7.
- [11] B. Ma, M. Lu, Y. Taniguchi, S. Konomi, Courseq: the impact of visual and interactive course recommendation in university environments, *Research and Practice in Technology Enhanced Learning* 16 (2021) 1–24.
- [12] S. Konomi, S. Inoue, T. Kobayashi, M. Tsuchida, M. Kitsuregawa, Supporting colocated interactions using rfid and social network displays, *IEEE Pervasive Computing* 5 (2006) 48–56.
- [13] S. Konomi, N. Thepvilojanapong, R. Suzuki, S. Pirttikangas, K. Sezaki, Y. Tobe, *Askus: Amplifying mobile actions*, in: *International Conference on Pervasive Computing*, Springer, 2009, pp. 202–219.
- [14] T. Sasao, S. Konomi, V. Kostakos, K. Kuribayashi, J. Goncalves, Community reminder: Participatory contextual reminder environments for local communities, *International Journal of Human-Computer Studies* 102 (2017) 41–53.
- [15] B. Ma, M. Lu, Y. Taniguchi, S. Konomi, Exploring jump back behavior patterns and reasons in e-book system, *Smart Learning Environments* 9 (2022) 1–23.
- [16] S. Konomi, L. Gao, D. Mushi, An intelligent platform for offline learners based on model-driven crowdsensing over intermittent networks, in: *International Conference on Human-Computer Interaction*, Springer, 2020, pp. 300–314.
- [17] S. Konomi, T. Sasao, S. Hosio, K. Sezaki, Using ambient WiFi signals to find occupied and vacant houses in local communities, *Journal of Ambient Intelligence and Humanized Computing* 10 (2019) 779–789.
- [18] S. Konomi, T. Sasao, The use of colocation and flow networks in mobile crowdsourcing, in: *UbiComp and ISWC 2015 - Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the Proceedings of the 2015 ACM International Symposium on Wearable Computers*, 2015, pp. 1343–1348. doi:10.1145/2800835.2800967.
- [19] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, R. Wirth, et al., *Crisp-dm 1.0: Step-by-step data mining guide*, *SPSS inc* 9 (2000) 13.