Challenges for Human-Data Interaction – A Semiotic Perspective

Heiko Hornung[1(✉)], Roberto Pereira[1,2], M. Cecilia C. Baranauskas[1], and Kecheng Liu[2]

[1] Institute of Computing, University of Campinas (UNICAMP), Campinas, Brazil
{heiko, rpereira, cecilia}@ic.unicamp.br
[2] Informatics Research Centre, University of Reading, Reading, UK
k.liu@henley.ac.uk

Abstract. Data has become ubiquitous and pervasive influencing our perceptions and actions in ever more areas of individual and social life. Data production, collection and editing are complex actions motivated by data use. In this paper we present and characterize the field of study of Human-Data Interaction by discussing the challenges of how to enable understanding of data and information in this complex context, and how to facilitate acting on this understanding considering the social impact. By understanding interaction with data as a sign process, and identifying the goal of designing human-data interaction as enabling stakeholders to promote desired and to avoid undesired consequences of data use, we employ a semiotic perspective and define research challenges for the field.

Keywords: Human-data interaction · Semiotics · Digital display

1 Introduction

Due to informatization and computerization, we have today an unprecedented access to data about all aspects of life. This data includes data about individuals, groups of people or even societies, as well as data about things, events, and so on. Examples include personal diet or exercise data, metadata related to email or web site traffic, data about economic or environmental indicators, or real-time traffic data.

The ecosystem within which data is produced, collected, edited (e.g. analyzed and synthesized) and used is complex and ranges from simple scenarios where data producer, collector, editor and consumer are the same person (e.g. nutrition diaries in paper form) to complex scenarios where many stakeholders are involved (e.g. a population census). Furthermore, the methods and purposes of how and why data is produced, collected, edited, and used vary greatly.

Data production, collection and editing are actions motivated by data use. While the abstract goal of data use commonly is to "extract" information or even "gain" knowledge, concrete goals vary and might be related to several domains e.g. entertainment, commerce, security, politics, arts, or science. These actions have given rise to

M. Kurosu (Ed.): Human-Computer Interaction, Part I, HCII 2015, LNCS 9169, pp. 37–48, 2015.
DOI: 10.1007/978-3-319-20901-2_4

questions such as: how to facilitate understanding of data and information, and how to facilitate acting on this understanding considering its potential social impact.

Scientific fields of study related to the interaction of people with data include but are not limited to the multi-disciplinary fields of data visualization or sonification, information visualization, and big data. While these areas are concerned with facilitating understanding and reasoning, big data focuses on the amount of data, visualization and sonification on representation.

More recently, different authors have begun to investigate challenges under the perspective of "Human-Data Interaction" (HDI; [3, 5, 14]), proposing to investigate how people interact with data in analogy to how HCI investigates the relation between people and computers. The common ground of these works seems to be the assumption that, in order to gain information or knowledge from data, people (or "end users of data") need to interact with data instead of only passively consuming them. This interaction goes beyond data analysis and includes interactive exploration of data sets. HDI should investigate interaction with data and questions related to understanding, but also technical, social and ethical issues such as privacy, transparency, commerce, as well as the question of how knowledge gained from data might benefit society.

We agree with Elmquist, Cafaro and Mortier et al. that HDI poses relevant and scientifically interesting challenges. We believe that these challenges can and should be addressed by the HCI community and also involve other disciplines. Furthermore, we believe that the challenges are more ample than presented by the cited authors. We argue that data production, collection, editing and use need to be investigated systematically, focusing on social impact they might provoke. This means an approach is required that goes beyond the challenges of representation and sense-making, and that is capable of conceptualizing pragmatic and social issues, i.e. issues related to meaning in context, intentions, negotiations and the effects and impact of decisions and other manifestations of data or information use.

In this paper, we define our view of Human-Data Interaction and adopt a perspective informed by (Organizational) Semiotics [12] in order to systematically describe challenges and outline a research agenda for Human-Data Interaction that considers the complex processes of data production, collection, editing and use. This paper extends previous HDI work by presenting a more systematic and comprising view of HDI research challenges, and builds upon Liu's [13] semiotic perspective on digital visualization by investigating HCI-related issues of data production, collection and editing, besides use.

The paper is structured as follows: Sect. 2 provides a brief motivation for investigating HDI, and characterizes HDI considering HCI and previous work; Sect. 3 presents a semiotic perspective on HDI; Sect. 4 outlines research challenges for HDI; Sect. 5 concludes.

2 Characterizing the Human-Data Interaction Problem

People using digital artifacts come into contact with data on many occasions. They produce data both intentionally, e.g., when using fitness trackers, and unintentionally, e.g., when leaving digital traces in browser search histories or being tracked by cookies.

People analyze data using methods of varying complexity, e.g., when reading an infographic or when trying to find an affordable flight from London to Tokyo during two weeks of May. They use data with different intentions, e.g. when sharing fitness data it might be to elicit encouraging feedback from peers ("I ran the mile faster than last week") or brag about their lifestyle ("I ran the mile at Rio de Janeiro's Ipanema Beach"). Accordingly, data use might have different consequences.

The way data is captured and analyzed influences data use and its consequences. People are often not aware of how the digital traces they leave are used by third parties, e.g. advertisers, credit-scoring companies, or data brokers (e.g. [1]). This has given rise to questions about privacy and other ethics-related issues, and how Interaction Design might or should address these (e.g. [8]).

People with no statistical background are often not able to analyze certain data. This might lead to them drawing false conclusions based on flawed analysis or to being dependent on data analysts. In complex data sets it is often not straightforward to extract relevant and significant information, and different analysts might thus focus on different aspects of the data and present different results. People using these results might not be aware that in order to get a more complete picture, one needs to look at data from different perspectives and might need to consider additional data sets.

Data, and specifically seemingly quantitative data, are often equated with objective facts. However, data is often subjective since methods for data collection carry subjectivity. Examples include medical, political or economical data, e.g. patients' self-reports or data collected according to ideologically charged collection methods (e.g. regarding crime statistics, is an assault on a person of different skin color an act of random violence, politically motivated or an act of racism or terrorism?). Data analysts bring their own beliefs and values into the analysis and act within complex contexts of organizational norms and values. Different analysts might interpret data and present analyses differently, influencing people's behaviors differently.

Without further context, the term "Human-Data Interaction" is not very specific. "Humans" have "interacted" with "data" for thousands of years, e.g. using astronomic or calendar data for planning hunting or agricultural activities. In scientific literature, particularly in the area of Computing, but later also in other areas, the expression is known at least since the 1990s (e.g. [10]).

Regarding the context of HCI, we found three independent lines of inquiry about Human-Data Interaction [3, 5, 14, 15]. While Cafaro [3] and Elmqvist [5] seem to focus on embodied aspects of data analysis (building on the work of Johnson [9], Dourish [4], and Lakoff and Johnson [11]), Mortier et al. seem to focus less on HCI-related aspects and more on the interaction between humans, datasets and analytics (e.g. algorithms). Furthermore, they seem to concentrate on personal data [14, 15].

Elmqvist [5] defines HDI as "*the human manipulation, analysis, and sensemaking of large, unstructured, and complex datasets*". Cafaro [3] defines HDI as "*the problem of delivering personalized, context-aware, and understandable data from big datasets*" and HDI systems "*as technologies that use embodied interaction to facilitate the users' exploration of rich datasets*". Mortier et al. [14] state that "*HDI concerns interaction generally between humans, datasets and analytics […]. HDI refers to the analysis of the individual and collective decisions we make and the actions we take […]. The term makes explicit the link between individuals and the signals they emit as data […]*".

The least common denominator of these three definitions is that people need to make sense of large and complex data sets. Furthermore, according to the three authors, the area overlaps multiple disciplines including HCI. Apart from that, the proposals of Elmqvist and Cafaro seem to be individual research endeavors that focus on the embodiment of interaction and try to prescribe a method to solve the respective research challenges: Elmqvist proposes tangible interaction, Cafaro gestural interaction as possible solutions. Mortier et al. frame the problem as a more general research challenge. Acknowledging the works of Elmqvist and Cafaro, we build upon the work of Mortier et al. [14, 15] to define our perspective on HDI and the related challenges.

"Data" plays an important role in many disciplines. In order to define a research challenge that is distinct from existing ones and that does not simply hijack topics from other areas, we need to investigate and try to define important core concepts such as "data", "interaction" and "humans". Furthermore, we need some conceptual framework that supports the organization of research challenges and a research agenda. Without this framework, we run the risk to simply create an unordered bag of topics and to omit important aspects.

Mortier et al. [15] started from the dictionary definition of "data", which they did not find very helpful, and then compounded "data" with seemingly arbitrary nouns, some of which appear in scientific literature, arriving at "data trail", "data smog", "big data", "small data" (potentially complex data sets about a single person [6]), participatory personal data ("any representation recorded by an individual, about an individual, using a mediating technology" [19]), or "open data"[1].

In the next step, Mortier et al. [15] look at "interacting" and at the "humans", i.e. at who is interacting with data. They contrast their proposal to the ones of Elmqvist and Cafaro and state, among others, that HDI does not only refer to embodied interaction, but to all kinds of interaction, especially not only to data exploration but also to other activities. Furthermore, they emphasize that data is "under constant revision and extension", and that "data" does not only concern the individual who provided the data or about who the data is, but also other stakeholders that might have different interpretations of the data. They then describe a data flow in which personal data (data by or about a person) is subject to analytics, which results in inferences, which in turn results in actions influencing people's behavior or in feedback to further analytics.

We think "personal data" as described by Mortier et al. [15] needs clarification. "Data about a person" could refer to personally identifiable information in the sense of privacy law or information security, but it could also refer to personally unidentifiable information such as anonymized or aggregated data, e.g. the unemployment rate or average creditworthiness of the neighborhood a person is living in. "Data by a person" might refer to data recorded or authored by a person or simply to data provided by a person, e.g. after this person conducted some data processing or acquired the data from a third person. To clarify the subject we propose that HDI should investigate "data that affects people". This includes stakeholders in the data lifecycle as well as data provenance, i.e. investigating previous stages in the data lifecycle such as data generation, collection, processing, etc. Moreover, the data lifecycle has not necessarily a simple

[1] https://okfn.org/opendata/.

sequential or circular form but might take the form of an arbitrary graph, e.g. when data sets are split or merged and used in other HDI processes.

The aforementioned authors tried to characterize "data" by citing a range of properties (e.g. "unstructured", "complex" or "large scale"). In fact, there exist numerous definitions of the term "data" that all have their strengths and weaknesses.

Merriam-Webster's definition of "data"[2] seems unnecessarily limited. "Factual information" for example excludes information that is uncertain, "numerical form" excludes qualitative data. "Information output by a sensing device or organ" seems to indicate that there is an agent actively providing data, which is questionable. Despite these limitations, these definitions show that "data" is something that has to be used, and indicate what can be done with data: data can be measured, collected, given meaning, analyzed or used for reasoning, discussion, interpretation, etc. The definitions also show that the term is somewhat problematic. The general Wikipedia definition[3] first conflates data and information ("pieces of data are individual pieces of information"), only to put them into a hierarchical relationship one sentence later ("Data as an abstract concept can be viewed as the lowest level of abstraction, from which information and then knowledge are derived."). This hierarchical relationship seems to be compatible with the "knowledge pyramid" [18], a model used in Information Science that, however, is controversial [7].

Since we are proposing HDI as a research challenge for HCI, we think it is unwise to subscribe to a definition from a specific scientific area since this might restrict approaches and methods to tackle the challenge and consequently restrict the number of interested researchers. For now, we preliminarily define "data" in the context of HDI as artifact-mediated representations of phenomena that need to be given meaning by people and that serve some purpose.

After characterizing the term "data", Mortier et al. [15] subsequently arrive at three themes (legibility, agency, and negotiability) which they use to structure further discussion. Legibility is concerned with processes of understanding, which requires making transparent data processing. Agency is concerned with the power of acting upon data and within systems that process data. Negotiability seems to be related to legibility and agency. It stresses contextual and dynamic factors such as changes over time or different social, legal contexts. As Mortier et al. [15] state, organizations of the HDI landscape other than legibility-agency-negotiability are possible. Legibility, agency and negotiability can be interpreted as requirements for interacting meaningfully with data or based on data. There might be other requirements, and from the point of view of HCI research or design, it is also pertinent to consider contexts where these requirements have not been met. We thus propose to employ the more general "understanding data" and the consequences thereof as top-level concerns of HDI. A main goal of HDI then should be to design human-data interaction that enables stakeholders to promote desired and avoid undesired consequences of data use.

"The consequences of understanding data" are actions that are or are not taken based on the actual understanding. For example, "data" might refer to private pictures

[2] http://www.m-w.com/dictionary/data.

[3] http://en.wikipedia.org/wiki/Data.

a person publishes in online social network services, "understanding" might refer to the interpretation of these pictures by an employee in a human-resources department, and "consequence" might refer to influencing a hiring decision by the interpretation of said pictures. Another simple example is credit scoring based on generalized data of a person (e.g. gender, race, or neighborhood). "Understanding" and "consequences" then cover the "interaction" part of HDI, and the two examples above show that this also encompasses legibility, agency and negotiability according to Mortier et al.

On the other hand, if we open the context from Mortier et al.'s "personal data" to "data that affects people", then legibility, agency and negotiability are not sufficient. Even if we completely understood issues related to legibility, agency and negotiability, we would still not know much about consequences of data interpretation in the social world.

Consider for example the 2014/2015 water crisis in the Brazilian state of São Paulo, which at the time of writing this paper is still ongoing[4]. News media started printing or posting water levels of the principal reservoirs that supply water to the region. This data is legible to a certain degree: it is known where the data comes from and people understand a part of its meaning ("drinking water supply is on a critical level" and "if the current data trend continues we will run out of water soon"). People have some agency and there exists some degree of negotiability, e.g. exerting public pressure or suggesting to news outlets to collect data more frequently and to also include secondary water reservoirs, weather forecasts, or other prognoses.

Based on the water level data, only a small part of the problem is understood, and this limited understanding might lead to short- or mid-term consequences such as reducing water consumption, denunciating water squanderers, complaining against the government, or moving to another region of the country. In order to understand a larger part of the problem and be able to take different actions, one has to ask questions such as what is the ratio of water consumption between private households and industry; has the water supply infrastructure kept up with population and industry growth; is the draught a singular phenomenon, or is it related to climate change or deforestation in other regions, etc. Answering these questions might result in a broader understanding and different consequences.

To give an example closer to Mortier et al.'s personal data [15], consider a person keeping a food log and monitoring body weight and related figures using a smartphone app. Even if legibility, agency and negotiability regarding collected data were guaranteed, we would not know whether this self-monitoring would lead to a higher degree of well-being or to an eating disorder, and how the latter behavior might be avoided or mitigated.

In order to be able to understand consequences or even design "data interaction" that promotes or inhibits certain consequences, we need to consider complex contextual factors including the systems of beliefs, values and norms of the involved people. Instead of proposing additional topics to legibility, agency and negotiability, that might need to be amended as our knowledge of the problem grows, we thus propose to stay at the more general level of understanding data and the consequences thereof.

[4] E.g. http://www.bbc.com/news/world-latin-america-29947965, last access on Feb 25th, 2015.

As a consequence of the need to consider people's norms, values and beliefs, regarding the "human" part of HDI, we subscribe to Bannon's "more human-centered perspective", i.e. we give "primacy to human actors, their values, and their activities" and understand HCI as "human activities mediated by computing" [2]. This human-centered perspective should not be confounded with flavors of human-centered design that are too narrowly focused on users and atomistic interactions with artifacts [16]. Furthermore, the "humans" in HDI do not only include those who directly access and use data but also those who affect and are affected by the consequences of this use.

3 A Semiotic Perspective on HDI

Semiotics is the doctrine of signs, and in Peircean Semiotics, "a sign is something […] which denotes some fact or object […] to some interpretant thought" ([17], vol. 1, par. 346), or paraphrased, "a sign is something which stands to somebody for something in some respect or capacity" [20]. Framing data as "artifact-mediated representations of phenomena that need to be given meaning by people and that serve some purpose" means we can understand data as signs that are subject to processes of interpretation. The three main branches of Semiotics—Syntax, Semantics and Pragmatics—are concerned with the structure, meaning and use of sign.

To give some examples of areas related to HDI, Data Visualization is mostly focused on Syntax, trying to reveal the structure of data (e.g. local or global maxima or minima, trends, distribution) by choosing adequate visualizations (e.g. bar charts, box plots, scatter plots). Information Visualization is focused on Semantics, trying to reveal the meaning of data (e.g. how the gross domestic product per person is related to life expectancy in different countries). Pragmatics is always present since any visualization or other representation of data or information is created with a purpose or intention. In fact, Pragmatics is already relevant when choosing data sources, deciding how to select which data from these sources, how to cleanse and process data, etc. From an HCI and HDI perspective, these are all design decisions that need to be taken deliberately and consciously, considering users and other relevant stakeholders.

Apart from Syntax, Semantics and Pragmatics, additional aspects are relevant for HDI as conceptualized in this paper, in particular the effects that intentional and purposeful actions have in the social world. Examples include privacy, trust, health or personal well-being. Organizational Semiotics [12] understands organizations of people as complex systems of sign processing and extends the traditional semiotic framework by including the Physical World, Empirics, and the Social World [20]. The physical layer is concerned with the media in which signs appear and the hardware with which they are transmitted or processed. The empirical layer is concerned with statistical properties and the coding of signs across different media. The social layer is concerned with the effects of the use of signs in the social world, i.e. with beliefs, expectations, or commitments. HDI-related topics on this layer include privacy, trust or security.

We can use the six layers of the extended semiotic framework and cross them with the different stages of the data lifecycle, e.g. data origin, selection, cleansing, mapping, display or interaction. As a third dimension we can introduce the different stakeholders

that appear at each stage in the data lifecycle. For each triple [layer in the Semiotic Framework, stage in the data lifecycle, stakeholder] we can now map design issues or questions (Fig. 1) that make the data lifecycle more explicit, clarify how data might be understood, and help to understand possible consequences of data use.

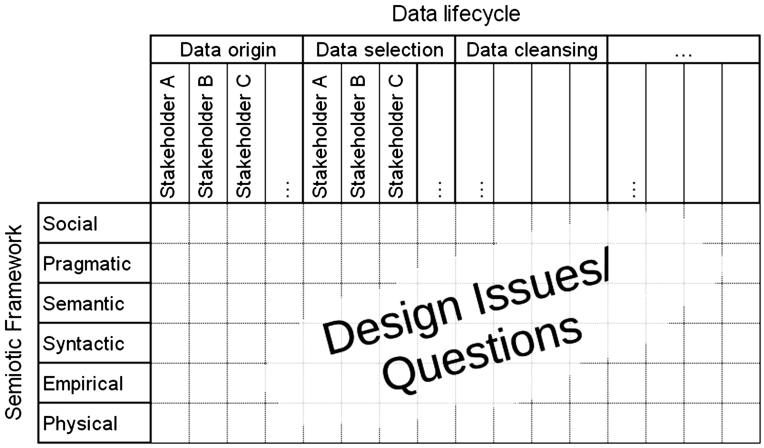

Fig. 1. Mapping design issues in the data lifecycle using the extended semiotic framework

As an example of how to use Fig. 1, consider data in the context of Quantified Self applications. In this case, the origin of data are sensor outputs of different devices (e.g. for tracking hear rate, insulin level, weight), as well as the actual self-quantifying person (e.g. keeping a manual food log). Stakeholders include the self-quantifying person, the manufacturers of hardware devices, software developers, as well as possibly a spouse, family and other relatives or friends, physicians, nutritionists, etc.

A design issue on the physical level that is of interest to varying degrees to the self-quantifier, the hardware manufacturer, and people in the self-quantifiers social environment is whether the used hardware devices can be fit into the physical context of the daily routine. Related issues on this layer include the hardware specification, e.g. memory capacity, processing power, or battery life. These issues might have ramifications in other layers of the semiotic framework or other stages of the data lifecycle.

An example issue on the empirical level related to "noise on the data channel" is whether the used devices allow accurate and precise observation of data, for instance when keeping a food log to track calorie intake "one apple" is not very precise since the weight of an apple might vary. A related issue is the choice of data encodings and formats, e.g. paper annotation or digital, standard or proprietary format. Again—and this can be generalized to all semiotic layers and the complete data lifecycle—these issues might affect other semiotic layers and later stages in the data lifecycle.

The syntactic layer is concerned with structure, e.g. in the case of self-quantification and data origin the procedures required for making data observable. Issues here are related to the correct use of capturing devices, as well as to questions such as whether it

is possible to cross-reference food intake with insulin levels or heart rate with type of physical activity.

Regarding semantics, an example question is whether it is meaningful to register food intake without registering physical activity, or even mental activity or psychological states. For a nutritionist it might be sufficient to simply monitor body weight in order to assess the success of a diet plan, while for a person who has a low self-awareness of food intake it might be more meaningful to observe more detailed data.

Issues in the pragmatic layer are related to people's intentions. The intentions of a self-quantifying person might include self-improvement, curious exploration, staying in control, increasing performance or mitigating health problems. The intentions of device manufacturers might be of a commercial nature with sub-goals such as providing reliable and pleasant means for observing data, or be related to altruistic motives such as improving people's well-being.

In the social layer, i.e. regarding effects or consequences, the self-quantifier might gain a higher self-awareness when recording data. Friends or family members might become amused, annoyed, interested or inspired to also start self-quantifying.

We can make similar investigations for the data lifecycle stages posterior to "data origin". Data use by the self-quantifying person or other stakeholders depends on the tools provided by software manufactures. On the syntactic level, if these tools provide limited functionality (e.g. only provide static graphs or tables of aggregated data) or have poor usability, some aspects of the data might remain unexplored. If the tools use poorly implemented gamification mechanisms ("new personal record!" after an insignificant change) data might be misinterpreted. This might influence stakeholders' understanding on the semantic level (e.g. a change of body weight within a short time span might not be significant while a trend during a longer period might be). The understanding shapes how stakeholder intentions are materialized, e.g. if weight loss or gain is interpreted as significant, data might be shared with friends or a nutritionist. On the social level, understanding of data and materialization of intentions might lead to a change of the self-quantifiers attitudes and behaviors. Device manufacturers on the other hand might be perceived as facilitators of personal well-being or exhibitionism, and public expectation regarding the quality of their products might change.

This superficial and very incomplete investigation of HDI-related issues of self-quantification, an arbitrarily chosen topic, demonstrates that HDI encompasses more than the interaction of an end-user with the product of some data processing. It also shows that for understanding this complex context, not only semantics, but all layers of the extended Semiotic Framework are required. Furthermore, consequences of data understanding and use appear in the social layer, and are shaped by all stages of the data lifecycle.

4 Research Challenges for HDI

Data lifecycles might vary among problem domains, but generally at least four major stages are present: data production and collection; data transformation and filtering; data presentation; and data display and interaction. The example of the previous section

served to argument that if we want to design for purposeful interaction with data, i.e. for data understanding and actions based on this understanding, we need to consider all stages of the data lifecycle and the relevant stakeholders.

The following questions might help to clarify an HDI problem. What does the data lifecycle look like? Who are the respective stakeholders and what are their responsibilities and competences? Who has access to which data at which stage? Who provides the hardware and software tools at each stage, and is it possible or necessary to design or redesign them? How to anticipate social implications of HDI design? These questions are relatively easy to answer—although they might require some effort as in the case of stakeholder identification—and do not depend on the domain and design problem, nor on the adopted conceptual framework.

The chosen conceptual framework enables a systemic view of processes related to the data lifecycle. We proposed to employ Organizational Semiotics, outlining the use of the extended semiotic framework as a tool for mapping design issues in the data lifecycle, but this does not preclude the use of alternative or additional frameworks.

We stated that one of the main goals of HDI should be to design human-data interaction that enables stakeholders to promote desired and avoid undesired consequences of data use. Since it is impossible to know any potentially conflicting consequences of data use beforehand, this is not a goal for design, but a desired outcome of the use of designed artifacts. A goal is to enable stakeholders to understand data on different levels, and in particular on the semantic, pragmatic and social level. Possible actionable objectives include making users aware of and understanding the intentions of different stakeholders involved in the data lifecycle, as well as enabling users to access different data sources and interacting with data representations that match their preferences and capabilities.

We identify two overarching research questions regarding these objectives: (1) how do the different stages of the data lifecycle affect each other; and (2) how do different representations of different data and different mechanisms to use these representations (e.g. explore, manipulate, share) affect people's ability to understand and act upon this understanding?

A first step to answering question (1) would be to investigate existing examples of human-data interaction such as those given in this paper, to reconstruct the data lifecycle as well as involved stakeholders and to try to identify patterns and invariants. A subsequent step might be to conduct different case studies in order to validate and refine knowledge about the impact of different data lifecycle stages and in order to create and test different data representations and interaction mechanisms.

In order to answer both questions, we can build upon already existing work in various areas. Regarding question 1, e.g. Software Engineering, HCI or Organizational Semiotics provide us with methods for identifying stakeholders and their concerns. HCI and Organizational Semiotics also allow investigating pragmatic aspects from different perspectives. Regarding question 2, we can build on concepts and use methods from—to name but a few—HCI, Semiotics, and Data or Information Visualization or Sonification, Theories of Human Perception and Cognition.

Despite this rich base of concepts and methods, many questions remain.

The semiotic concept of abduction, i.e. the process of generating, verifying and refuting hypotheses, allows us to reason about how people gain an understanding of

data. How can we leverage the concept of abduction during design and evaluation, e.g. how can we support users with little or no statistical knowledge to generate sensible hypotheses regarding quantitative data?

Data and Information Visualization and Sonification provide us with knowledge about good representations of data, albeit focused on the syntactic or semantic level. What are good representations that consider the pragmatic and social level? Are visual and auditory displays and interaction mechanisms enough or should we also explore haptic displays or, more general, embodied interaction?

HCI knows various frameworks that consider stakeholder concerns and the social impact of design and use, e.g., Participatory Design or Value Sensitive Design. How can these be employed in HDI? How can we conduct a participatory design of the data lifecycle? Does it make sense and is it possible to conduct a participatory data analysis? How can we apply principles of Value Sensitive Design to HDI?

5 Conclusion

The production, collection, processing and use of data have taken new dimensions in terms of complexity and possible social impacts. In this paper, we have presented problems related to Human-Data Interaction, and identified HDI as a research field for HCI and related areas. We have characterized HDI and outlined research challenges, adopting a semiotic perspective to ground future investigations.

Our discussion adds to the existing literature by offering an alternative view HDI and its challenges. We argue that the goal of designing Human-Data Interaction should be to enable stakeholders to promote desired and to avoid undesired consequences of data use. Therefore, data production, collection, processing and use need to be investigated systematically, focusing on the social impact they provoke. Understanding interaction with data as a sign process, we draw attention to an approach that goes beyond the challenges of representation and sense-making, also considering pragmatic and social issues related to meaning in context, intentions, negotiations and the effects of data use.

Acknowledgements. This work received financial support by CAPES, CNPq (#308618/2014-9), and FAPESP (#2014/01382-7).

References

1. Anthes, G.: Data brokers are watching you. Commun. ACM **58**(1), 28–30 (2014)
2. Bannon, L.: Reimagining HCI: toward a more human-centered perspective. Interactions **18**(4), 50–57 (2011)
3. Cafaro, F.: Using embodied allegories to design gesture suites for human-data interaction. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, UbiComp 2012, pp. 560–563. ACM, New York, NY, USA (2012)
4. Dourish, P.: Where the Action Is: The Foundations of Embodied Interaction. MIT Press, Cambridge (2001)

5. Elmqvist, N.: Embodied human-data interaction. In: ACM CHI 2011 Workshop "Embodied Interaction: Theory and Practice in HCI", pp. 104–107 (2011)
6. Estrin, D.: Small data, where n = me. Commun. ACM **57**(4), 32–34 (2014)
7. Frické, M.: The knowledge pyramid: a critique of the DIKW hierarchy. J. Inf. Sci. **35**(2), 131–142 (2009)
8. Goodman, E.: Design and ethics in the era of big data. Interactions **21**(3), 22–24 (2014)
9. Johnson, M.: The Body in the Mind. The University of Chicago Press, Chicago (1987)
10. Kennedy, J.B., Mitchell, K.J., Barclay, P.J.: A framework for information visualisation. SIGMOD Rec. **25**(4), 30–34 (1996)
11. Lakoff, G., Johnson, M.: Metaphors We Live By. University of Chicago Press, Chicago (2003)
12. Liu, K.: Semiotics in Information Systems Engineering. Cambridge University Press, New York, NY (2000)
13. Liu, K.: Semiotics in digital visualization. In: Keynote Lecture at the 16th International Conference on Enterprise Information Systems, ICEIS 2014. http://vimeo.com/95737955. Accessed 10 Nov 2014
14. Mortier, R., Haddadi, H., Henderson, T., McAuley, D., Crowcroft, J.: Challenges and opportunities in human-data interaction. In: The Fourth Digital Economy All-hands Meeting: Open Digital (DE), Salford (2013)
15. Mortier, R., Haddadi, H., Henderson, T., McAuley, D., Crowcroft, J.: Human-data interaction: the human face of the data-driven society. Available at SSRN: http://ssrn.com/abstract=2508051 or http://dx.doi.org/10.2139/ssrn.2508051 (2014)
16. Norman, D.A.: Human-centered design considered harmful. Interactions **12**(4), 14–19 (2005)
17. Peirce, C.S.: Collected Papers of Charles Sanders Peirce, vols. 1–6. Harvard University Press, Cambridge (1931–1935)
18. Rowley, J.: The wisdom hierarchy: representations of the DIKW hierarchy. J. Inf. Sci. **33**(2), 163–180 (2007)
19. Shilton, K.: Participatory personal data: an emerging research challenge for the information sciences. J. Am. Soc. Inf. Sci. Technol. **63**(10), 1905–1915 (2012)
20. Stamper, R.K.: A semiotic theory of information and information systems/applied semiotics. In: Invited Papers for the ICL/University of Newcastle Seminar on "Information", 6–10 Sept 1993