Design Enquiry Through Data

Appropriating a Data Science Workflow for the Design Process

Kun, Peter; Mulder, Ingrid; Kortuem, Gerd

Published in:
Proceedings of British HCI Conference 2018

Publication date:
2018

Document Version
Accepted author manuscript, peer reviewed version

Design Enquiry Through Data: Appropriating a Data Science Workflow for the Design Process

Peter Kun
Delft University of Technology
Landbergstraat 15, 2628CE Delft
The Netherlands
p.kun@tudelft.nl

Ingrid Mulder
Delft University of Technology
Landbergstraat 15, 2628CE Delft
The Netherlands
i.j.mulder@tudelft.nl

Gerd Kortuem
Delft University of Technology
Landbergstraat 15, 2628CE Delft
The Netherlands
g.w.kortuem@tudelft.nl

The recent developments in data science and end-user data tools indicate an opportunity for designers to adapt new data tools for design enquiry. Data has an unquestionable role in the future of the design practice creating new digital products and services. Today's data deluge also opens up new ways of enquiring about the world through data. The current work explores how designers could appropriate a data science workflow in their design research process. Two studies are conducted to explore how a data science workflow could be adapted into a design research process. We present how the participants appropriated data techniques for creative uses and how they synthesized a data-centric enquiry into their research process. We found that designers appropriate data using their creative capacities in hypothesis forming for data collection and exploratory data analysis, and we highlight some implications what this would result. Our findings can inform the design space of the creativity support of future data tools and future data-centric design methods.

Design enquiry, data exploration, data science workflow, end-user data tools

INTRODUCTION

We live in an ever-increasing abundance of digital data, and HCI designers need to take into account this data deluge at an increasing pace. With the growing ubiquity of data and trends of datafication (Lycett, 2013), the designs of third-wave HCI artefacts (Bødker, 2006) (i.e., context-aware or connected artefacts) and the designs of digital services have been informed by more-and-more digital data (Mortier *et al.*, 2014). Under digital data, without aiming for a comprehensive list, we refer to quantitative data, data in databases, sensor data, open data, and so forth. While HCI has a long tradition of discussing various aspects of digital data (such as information visualization (Card, Mackinlay and Shneiderman, 1999), knowledge discovery and information retrieval (e.g., (Fayyad, Piatetsky-Shapiro and Smyth, 1996; Marchionini, 2006; Dörk, Carpendale and Williamson, 2011)), or users' engagement with personal data (Mortier *et al.*, 2014)), using digital data in the HCI design process to enquire about different phenomena is a more recent, and still emerging phenomenon (e.g., (Speed and Oberlander, 2016; Bogers *et al.*, 2016; Giaccardi *et al.*, 2016; Feinberg, 2017)).

HCI designers were called out for more informed uses of data (of any size and scale) to enquire about the world and the design spaces of potential contemporary artefacts and services (Churchill, 2012). More precisely, Churchill argued for the potentials of a data-aware design process, one where design thinking is combined with data analysis. In a more recent article, Churchill (2017) expanded on the notion of using digital data in the design research process, intertwining an ethnographic lens with data analysis to enquire about the users and contexts of connected artefacts and digital services. To conclude, the use cases and the potentials of utilizing digital data in making sense of different phenomena in the design process have been established.

However, for more informed ways of utilizing digital data in the HCI design process, we need more established data practices compatible with the design process. With the growth of data science as a field (Cao, 2017), new techniques, methods and tools can be used by designers to use and make sense of various of digital data in their research process. In this paper, we explore how a data science approach (with its constituting mindset, techniques and tools) is appropriated for a design research process. We present two studies we

conducted with master-level design students to investigate whether with existing end-user data tools, could designers inexperienced with data appropriate a design research enquiry with digital data, and what consequences such an approach would mean for the design process. Our results indicate, that existing end-user data tools can be incorporated into design enquiry, and designers can use their creative capacities in hypothesis forming of data collection and data exploration of digital data. Based on our discussion we distil guidelines for using data in design enquiry.

RELATED WORK

There are two main strands of research relating directly to our exploratory studies; one strand focuses on how data science has opened up for the masses, and the other on how digital data has been explored in the design process.

Data science for the masses

Even though, data science has matured to be a field on its own (Cao, 2017), there is no formal definition of what steps constitutes a data science workflow. It is generally agreed that the typical stages are: starting from formulating a question, acquiring data to answer the question, and then further steps of inference happen (e.g., through visualisation or modeling), and at last communicating the results (Kandel *et al.*, 2012).

The traditional ways of teaching data techniques were based on statistics and computer science theory. Contemporary approaches follow a more applied and holistic way. For example, Baumer (2015) describes their holistic approach in undergraduate education, teaching a full spectrum of tools to prepare students working with data from real environments. In this way, students learn how to think with data, from asking a question to analyse the data and communicate (i.e., visualise) the findings. This encapsulating process aimed at enquiring the world and then inferring learnings is similar to how designers would use research techniques in their design process. It is unlikely that designers inexperienced with data take on such a formal academic training, so it is worthwhile to look at alternative approaches that been established for learning data science workflows. Hill and colleagues (2017) reflect on their experiences of *'democratizing data science'* through community workshops. In their work, they teach the basics of programming for the aim to empower novices to be able to formulate a question from data, and to be able to collect data and get to an answer. This programmatic way, using common data science tools as Python and data libraries, provides the most flexible skill and toolset for working with data, but with the price of the steepest learning curve.

As the design practice is often conducted through methods and tools, corresponding data approaches (in this case data tools targeted at end-users) are potentially effective to utilize in the design process. D'Ignazio and Bhargava (2016) present DataBasic, a set of learning tools for data literacy, that is intentionally not programming-driven, but targeting the acquisition of data skills through single-focused learning tools. Their explicit focus is on learners, but the DataBasic tools can address actual (but narrow) use-cases.

Although a programmatic approach would provide the necessary flexibility to appropriate a data science workflow for the design process, simplified tools to learn data competencies fall short on addressing the use-cases and needs designers would have. Interestingly, more and more professional communities, such as data journalists or digital humanities scholars, have emerged harnessing the ubiquity of data in new ways. These communities often share their know-how publicly as methods or tool libraries (e.g., (Gray, Chambers and Bounegru, 2012; School of Data, 2016)). These libraries curate end-user tools targeting professional workflows with data, and go beyond ubiquitous spreadsheet software (e.g., Excel).

Data in the design process

The end-user perspective was also explored by Grammel and colleagues (2010), more specifically, how novices make sense of data and construct visualisations. Their work provides insights into the mental models how novices approach a dataset to explore it through visualisations. Bigelow and colleagues (2014) particularly studied designers as non-data experts in their enquiry, investigating how *designers* approach working with data to create *visualisations*. They elaborate on the patterns they noticed with designers' sensemaking of data, and discuss themes how the processing of a dataset could be better supported. These examples show how visualisations are fundamental elements in making sense of data, and illustrate the thinking process for the creation of visualisations. The potential of appropriating a data science workflow goes beyond creating *visualisations*; therefore, the current work emphasises an entire data science workflow in its use from defining the right question that informs the data collection with designerly inference as an outcome, and in this way leading to design enquiries targeted a better understand of phenomena in the world.

Going beyond visualization and looking at design as a creative process, D'Ignazio's work (2017) on creative data literacy provides a set of tactics to support developing data competencies geared towards creative work. She argues to support novices' learning of data using datasets they care about and that are from the real-world, instead of

unreal datasets used decades ago to teach statistics.

order to inform design concepts. To be able to better navigate through the data science workflow, we

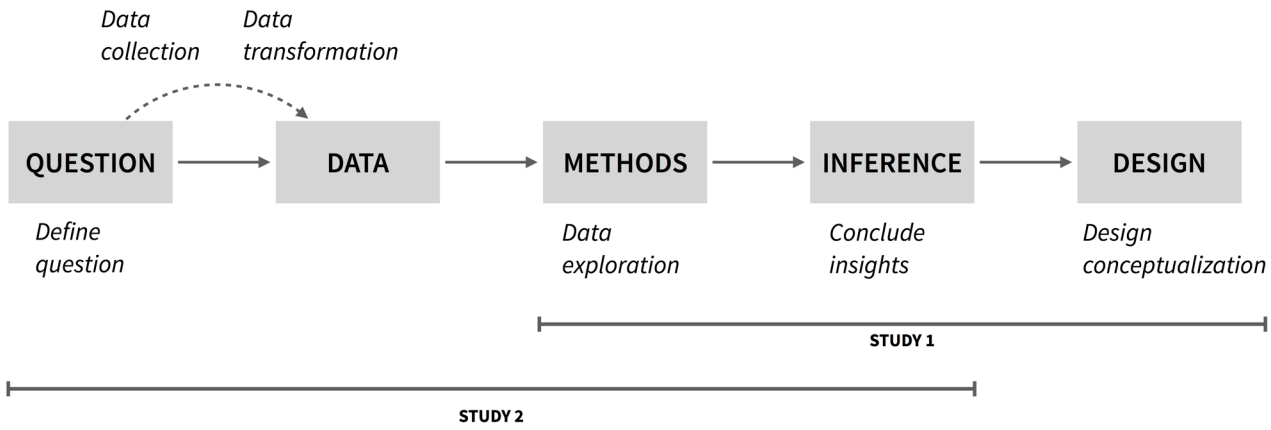

Figure 1. Schematic of a generic data and design process, which also served as the basis of the conducted studies. The scope of the two studies are indicated in relation to the process.

APPROACH

The review of related work has indicated two contexts where a data science workflow could be appropriated for the design process. These contexts bear similarities with each other, but also assume different conditions.

The first context reflects on the ubiquity of existing datasets (such as accessing datasets from open data portals), and that it still remains unclear how could designers work from a provided dataset in finding the right problem to solve. This addresses analytical work, e.g., extract additional knowledge out of a dataset. In design research, this usually refers to data exploration of a public or received dataset, to extract value from it. The second context hints to the increasingly easier ways to collect and store data. Here, designers use a data science workflow to augment their research process in capturing and analysing data to answer enquiries. In this case, designers can use the capturing and analysis of digital data to complement qualitative techniques (such as interview studies or ethnography) to gain additional insights from the data.

These two contexts have informed the set up two exploratory studies:

- Designers analysing a provided dataset to identify a problem space for design concepts (Study 1 - master thesis records);
- Designers with identified problem space capturing and analysing datasets (Study 2 - tourism).

The scope of studies is limited to the conceptual phase (i.e., *'fuzzy front end'*) of the design process (Sanders and Stappers, 2008), focusing on the using data in gaining understanding of the world in

adapt the process described by Baumer (2015) to match a generic design process, which is shown on Figure 1. The workflow starts with defining a question to investigate. In order to answer this question, data is collected. It is most likely necessary to transform and clean this data to prepare it for exploration. When a dataset is available for exploration, different analytical methods are applied on it (statistical analysis, visual analysis, etc.). The exploration generates an inference, such as insights. These insights then can contribute to the designer's understanding of the problem, or can be further communicated as visualizations, design concepts and so on.

Both exploratory studies aim to answer the following research questions:

- What are the conditions that enable a data science workflow to be integrated into a design process?
- Can end-user data tools support designerly work?
- How does the design process and the design reasoning change when using digital data?

More specifically, for the first study, our research objective is to see how novice designers inexperienced with data work from an unknown dataset towards a specific design goal, using end-user data tools without prior tutorial. For the second study, our research objective is to see how novice designers inexperienced with data appropriate a data science workflow in design enquiry, using end-user data tools after trying them through a homework prior to the study. Table 1 shows the setup of the studies, which are elaborated below.

Both studies are promoted as learning workshops to teach designers data competencies and tools, by hands-on working on a design problem with data.

Table 1. *The setup and methodology overview of Study 1 and Study 2.*

	Study 1 (master theses records)	***Study 2 (tourism)***
Research questions	How is a data science workflow appropriated for the design process when the starting point is a design brief and a dataset?	How is a data science workflow appropriated when used as a complementary method for designerly enquiry?
Setting	One-day elective class.	Three consecutive days workshop, part of a semester-long project.
Participants	First year master design students (n=20, 13 females, 7 males) from Service Design, Interaction Design and Product Design. Participants worked in pairs.	First year master design students (n=26, 20 females, 6 males) from Service Design. Participants worked in groups of 4-5.
Apparatus		
Dataset	1884 master theses records with complete metadata from the participants' university repository [redacted for anonymity]. Scraping and moderate cleaning (removing spelling and capitalization errors) was done by first author.	No provided dataset (the participants captured data as part of the study).
Software tools	Microsoft Excel, Google Sheets, RAWGraphs, OpenRefine, Google Fusion Tables	WebScraper, Microsoft Excel, Google Sheets, RAWGraphs(Mauri *et al.*, 2017), OpenRefine, Carto
Materials	Worksheets for Activity 1: dataset column titles, process reflection sheet. Worksheets for Activity 2: Data design canvas (Data, Model, Experience, Problem, Added value), design reflection sheet.	No additional materials provided.
Procedure		
Pre-study task	No pre-study task.	Homework a week before the study on scraping a page (with WebScraper), and to extract one insight from the Titanic dataset with RAWGraphs.
Study	Basic introduction to data processing and tools. *Activity 1 (Data exploration)*: Processing the provided dataset and analysing it towards concluding 3 insights and make a presentation. *Activity 2 (Conceptualization)*: Based on one insight from Activity 1, generate a data-inspired design concept and make a presentation.	Basic introduction to data processing and tools and debriefing the pre-study task. *Activity 1 (Question definition)*: Related to the semester project, defining three research questions to be answered with data. *Activity 2 (Data collection)*: Capture data (by scraping or downloading) for the questions from Activity 1. *Activity 3 (Data transformation)*: clean, prepare, transform the captured data from Activity 2. *Activity 4 (Data exploration)*: Make sense of the dataset from Activity 3 by analysis or visualization. Conclude on three main insights gained. Iterate from Activity 1, if necessary. Prepare a presentation about the process and the insights.
Follow up	Post-study questionnaire (fill rate: 75%) about learning goals, individual reflections and impact of the learning on participants' future work.	Post-study questionnaire (fill rate: 50%) about learning goals, individual reflections and impact of the learning on participants' future work.
Research data	Content analysis of participants' worksheets and presentations from Activity 1 and 2, post-study survey and observations.	Content analysis of presentations from Activity 4, ethnographic field notes throughout the study, post-study survey and observations.

The studies are similar in several aspects. Following the guidelines by D'Ignazio (2017) on selecting familiar datasets for the participants, Study 1 features a dataset relatable for the participants' personal experience, whilst the participants of Study 2 collect data in a problem space they immersed in prior the study. We provide open-source or freely available tools established in other end-user communities (e.g., data journalists), as shown in Table 1. Throughout the studies, the participants work towards tangible outcomes (as insights and concepts) captured during mid-term and final presentations. In the following section, we present the two studies and their respective results in detail.

STUDY 1 (MASTER THESIS RECORDS)

With Study 1, our aim was to see how a data science workflow is appropriated for the design process, with the conditions of novice designers (master-level design students) facing an unknown dataset without prior experience in data. Prerequisite or part of their education, the design students had previous

coursework on basic statistics, programming and design research methods. Based on this background, we assumed, that design students (of bachelor degrees from technical universities) will have some tacit data knowledge that can inform their approaches. Furthermore, we assumed a basic level of familiarity with spreadsheets software (e.g., Excel), and familiarity with typical visualization techniques (e.g., charts, graphs).

Participants and setup: Twenty students (13 females, 7 males) participated in the current study, as a one-day elective class. The students were first year master students in different orientation of design (product, interaction and service design) of a large, European industrial design faculty. During this study, participants worked in pairs.

Apparatus: The participant pairs were provided with a dataset, several software tools and worksheets to use. The dataset was a complete database of the participants' university's [withheld for anonymity] internal repository for master's thesis at the time of the study, containing 1884 rows and 28 columns of various metadata, including the theses' *Title*, *Abstract*, *Mentors*, *Keywords*, etc.

The participants' process was supported by additional worksheets; these worksheets were used in collecting research data, but also to guide the process for the participants. *Activity 1 (Exploration)* was supported with a printout of the column titles of the dataset, if the participants wanted to take notes about it. *Activity 2 (Conceptualization)* was supported with a data design canvas worksheet, that had fields with guiding questions for the process: Data (*"What are the available data?"*), Model (*"How will it work?"*), Experience (*"How will it look like?; What will it do?"*), Problem, Added value. The reflection sheets contained an empty timeline to visualize and describe the process of the participant pairs (an example of the filled reflection is shown on Figure 3).

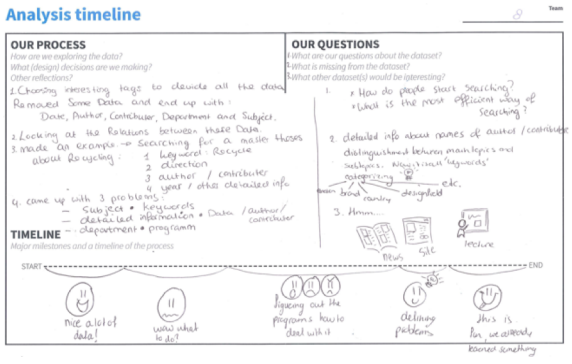

Figure 3. Worksheet from Study 1, a participant pair reflecting on their data analysis process.

Procedure: The study started with a basic introduction to data processing and the provided tools (the participants worked on their own computers) and the different worksheets. The first half of the study was *Activity 1 (Analysis)*; the participant pairs received a task to explore the provided dataset and extract three main insights that they found as design problems to solve. The participant pairs could use the provided additional worksheet (dataset column titles), but it was not compulsory. They received minimal guidance how to open the dataset in spreadsheet tools, and to do basic data cleaning and transformations in OpenRefine. For the visual analysis of the dataset, participants were provided basic guidance to use RAWGraphs (Mauri *et al.*, 2017), and Google Fusion Tables. At the end of the activity, participants needed to fill up the process reflection sheet. The second half of the study was *Activity 2 (Conceptualization)*; participant pairs received the task to develop a design concept based on one selected insight from their output of *Activity 1 (Exploration)*. The procedure was based on the process from Figure 1, and Table 1 provides an overview of how it was operationalized during the study.

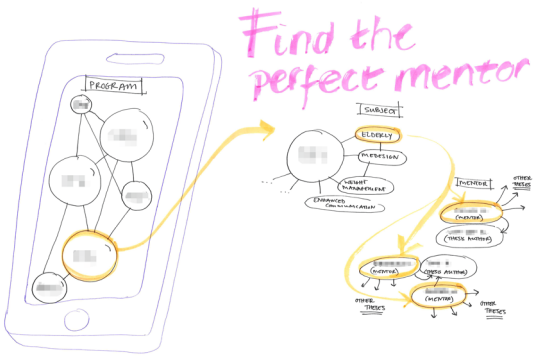

Figure 2. A concept on finding the right mentor for a certain master thesis (Study 1).

Data collection and analysis: Throughout the study, observations, notes and photos were captured. For both activities, we provided worksheets to capture the participants' self-reflections on their process (see Apparatus). Both activities were concluded with the participants preparing a short visual presentation with 3 insights and a design concept, respectively. Following the study, we analysed the presentation materials, the self-reflection worksheets, and the observations to identify patterns, similarities and differences. The study was also followed by a questionnaire sent to the participants to collect immediate data about the learning goals and reflection on the impact of the workshop on their future work.

Results of Study 1

Example project: We first present one concept generated by one participant pair throughout the study, to illustrate the kind of complexity and novelty

Table 2. The design concepts generated during Study 1, and the used data attributes informing the concepts.

Group	Concept description	Data properties used							
		Keywords	Mentors	Departments	Programme	Title	Author	Date	Abstract
M1	Finding the right mentor for your graduation	x				x	x		x
M2	Network visualization of finding the right topic for your research	x	x						
M3	Tinder for finding the right thesis subject	x				x			
M4	Finding the right subject for your graduation	x	x	x					
M5	Personalized search based on user data								
M6	Finding the right mentor for your graduation		x	x	x				
M7	Finding the right mentor for your graduation	x	x	x					
M8	Finding the right subject for your graduation	x	x		x		x		
M9	Connect people around the same interests	x		x	x		x		
M10	Showing trends in graduations			x	x	x		x	

achieved by a one-day study. The dataset contained 1884 records of different master thesis entries. All of these thesis entries had multiple keywords (such as: “design”, “sustainability”, “Internet of Things”, etc.). The average keyword count per thesis was 4.50 (SD=2.34, min=1, max=29). Similar to the keywords, all thesis entries had multiple mentors (faculty members, in rare cases also company mentors). The average mentor count per thesis was 2.32 (SD=0.67, min=1, max=6). Participant pair **M6** argued, that based on the characteristics of the keywords and mentors, it is possible to explore the most common keywords for a given mentor, and vice-versa, which mentors are most common for given keywords (i.e., keywords and mentors formed a bipartite graph). Following this insight, this participant pair presented a concept to find the perfect mentor based on keyword interests (see Figure 2).

Process: We observations participant pairs during *Activity 1 (Data exploration)* to be daunted by the initial task of taking a previously unknown dataset and extract valuable meanings out it. They performed this task without formal training in working with datasets, following a hands-on learning process. An example reporting of this process from **M4**:

> 1. Start with repository and identify users and use-cases.
> 2. Looking at the dataset, trying to understand.
> 3. Trying out the tools: without any questions behind, just exploring.
> 4. Visualizing random columns [with RAWGraphs]
> 5. Seeing some patterns? [pointing back to point 2.]
> 6. We looked back at the dataset and started to ask ourselves some questions
> 7. Trying to simplify the dataset to our needs using OpenRefine

The participant pairs generally followed a similar process: an unstructured, ad-hoc process of data analysis, where they continuously gained a better understanding of the dataset, learnt the usage of the tools, and got familiar with the techniques of working with data. The main data techniques used were cleaning the dataset to remove inconsistencies, such as character capitalizations or spelling errors, and the transforming of the dataset in various ways so it can be inputted into the used tools. In the end, 9 out of 10 participant pairs succeeded in presenting three insights based on the dataset, one pair misunderstood the task.

Based on an insight from Activity 1 (Exploration), during Activity 2 (Conceptualization) each participant pair developed one design concept. Table 2 provides a detailed overview of the developed concepts. Most participant pairs focused on a few data attributes from the dataset, namely the thesis title, abstract, graduation mentors and their departments and keywords. Three concepts focused on finding the right mentor and four concepts focused on finding the right graduation subject. One concept targeted improving the overall search experience, one concept aimed at

connecting people with similar interests based on the subjects and one concept focused on showing trends in graduation projects.

Participant reflections: In the post-study questionnaire, the majority of participants primarily valued learning about the tooling to work with data. In detail, they found learning about the generic workflow of working with data as something new. Furthermore, they found the provided end-user tools approachable to integrate data into their design process.

The participants also reflected on the shift in the thinking process necessary to utilize data. As one participant phrased his main learning: *"Asking the right questions at the beginning of the data, what do want to know, helps to understand what to look for."* (participant from **M4**). Another participant phrased it differently: *"The importance of a research question or hypothesis for structuring and processing the data"* (participant from **M1**).

STUDY 2 (TOURISM)

With Study 2, our aim was to see how would the appropriation of a data science workflow using end-user tools could complement the design research process. We assumed, that novice designers inexperienced with data will need to do multiple iterations of the activities to get comfortable with data capturing and analysis for designerly insights. Similarly to Study 1, the design students had previous coursework on programming and quantitative and qualitative research methods for design. Based on this, we expected that the design students (with backgrounds from technical universities) have basic familiarity with spreadsheets software, and familiarity with typical visualization techniques (e.g., charts, graphs).

Participants and setup: 26 students (20 females, 6 male) participated in the study, which ran for three consecutive days as a part of the participants' semester project. All students were first year master students in service design from a European design faculty. During this study, participants worked in groups of 4-5.

Apparatus: The provided software tools are summarized in Table 1.

Procedure: Prior to the study, participants received a *Pre-study task* to get familiar with data capturing and data visualization for analysis. The participants were instructed to scrape a specified webpage (their university library's search results page), and to visually explore a sample dataset from RAWGraphs (Mauri *et al.*, 2017) and extract three insights from it. The study started with a basic introduction to the data workflow and the provided tools (the participants worked on their own computers) and a debriefing of the pre-study task. The beginning of the study was *Activity 1 (Question definition)*; the participant groups needed to define research questions based on their semester brief (that was focused on tourism in a Nordic capital city). *Activity 2 (Data collection)* continued with the research questions from *Activity 1 (Question definition)*, with the task to capture data in relation to the research questions. The task of *Activity 3 (Data transformation)* was to clean, prepare and transform the captured dataset from *Activity 2 (Data collection)*. The end of the study was *Activity 4 (Data exploration)*, during which the participant groups needed to make sense of the dataset by analysis and visualization and prepare a presentation about it and their process. The participant groups could iterate from *Activity 1 (Question definition)* to *Activity 4 (Data exploration)*, if necessary. The procedure was based on the process from Figure 1 and an overview of how it was operationalized can be found in Table 1.

Data collection and analysis: Throughout the study, observations and photos were captured by the researchers, and ethnographic field notes were taken by an independent observer throughout the three days of the study. Following the study, we processed the presentation materials, the observations and the field notes to identify patterns, similarities and differences. The study was followed by a questionnaire sent to the participants to collect immediate data about the learning goals and self-reflection on the impact of the workshop on their future work.

Results of Study 2

Example project: The problem space of this study was centred around tourism in a Nordic capital city (the participants' semester project brief). For the current study, the participants were told to utilize a data science workflow to further their research about the problem space. In order to illustrate the kind of problems and what complexity the participant groups operated on, we first present the work of participant group **T2**. This group focused on a specific neighbourhood from the lenses of tourism. Their leading research questions were:

- Which places are recommended in [certain neighbourhood]?
- Where do locals and visitors spend their time in [certain neighbourhood]?
- What do people search about [certain city] abroad on Google?

For example, in their approach, **T2** analysed social media hashtags for a specific neighbourhood, and especially looked into the less common hashtags from slang and subcultures.

Process: Prior to the study, the participants received two *pre-study tasks* as homework. The task to visually explore a dataset (to be done individually)

Table 3. The problem areas under investigation during Study 2, and an overview of the data acquired by participant groups and their tool usage.

Group	Participants (# of female, male)	Problem area	Data sources	Tools used
T1	2 F / 2 M	What are the places locals visit and how to provide local experiences to visitors?	Crowdsourced review sites (2), curated travel sites (1), social hospitality site (1)	WebScraper, OpenRefine, Google Sheets, RAWGraphs
T2	4F / 1 M	Focused on a specific neighbourhood, what are the recommended places and places of interest for locals and visitors?	Crowdsourced review sites (2), curated travel sites (2), social media (1), qualitative interviews	WebScraper, OpenRefine, Google Sheets, RAWGraphs
T3	4 F	What places are recommended by locals? How far visitors go from the hot spots?	Crowdsourced review sites (1), curated travel sites (1)	WebScraper, OpenRefine, Excel, Carto
T4	5 F	In detail comparing the different neigbourhoods.	Crowdsourced review sites (1)	WebScraper, OpenRefine, RAWGraphs, Google Mapmaker, Carto
T5	2 F / 2 M	Can data-driven technologies support providing visitors the experience of locals?	Social media (2)	Twitter API, WebScraper, RAWGraphs, Carto
T6	3 F / 1 M	How can the visits of business travellers be extended?	Crowdsourced review sites (1), Open weather data (1)	WebScraper, OpenRefine, Excel

was done by all participants, whilst the task of scraping a webpage (to be done as a group) was done by half of the groups. During the debriefing, the participants reported difficulty in extracting interesting findings from the sample dataset without background knowledge and knowing what would be interesting to know about this dataset.

The participant groups started with *Activity 1 (Question definition)*: the groups first considered their project and defined initial research questions to be answered with data. Moving forward to *Activity 2 (Data collection)*, the groups captured data from online resources, primarily by scraping and downloading existing datasets. Scraping was mainly daunting for participants without extensive programming skills, nevertheless, by the end most participants managed to develop non-trivial scrapers, tackling pagination and similarly complex problems. All scraping was done using browser extensions. The groups ended up capturing data about tourism, primarily by scraping publicly accessible data from social media (e.g., Twitter and Instagram) and tourism websites (TripAdvisor, etc.), as shown in Table 3. As next step, the participant groups worked on *Activity 3 (Data transformation)*. The main needs of data cleaning were to eliminate inconsistencies, hidden characters and similar string operations. As a significant portion of the captured data was location-specific (e.g., addresses), some groups used OpenRefine to enrich their datasets with GPS coordinates. This was accomplished by following an OpenRefine recipe that called an external API with the address input to enrich the data with GPS coordinates. The participant groups finished the study with *Activity 4 (Data exploration)*. The groups explored their dataset through visualisations in RAWGraphs and Carto. Throughout the three days of the study, all groups went through several iterations of *Activity 1 (Question definition)* to *Activity 4 (Data exploration)*. Table 3 shows each participant groups' main research direction, the data sources and the tools used. In the end, all participant groups managed to find valuable insights for their semester project. To better illustrate the kind of research questions the teams attempted to answer, an example: One team focused on approaching how seasons influence tourism and when found that the correlation of seasonality and tourism is probably low for their target group, focused on comparing the target city with similar cities, based on weather and other predictors.

Participant reflections: In the post-study questionnaire, the majority of participants' reflections were unanimous: all responses noted data acquisition as primary learning, followed by visualization of data and an increased general understanding of data, its processes and its potentials for the design process. Besides three respondents with more technical background, the participants were also unanimous to report how challenging it was to scrape data.

Participants emphasized the transition from Data collection to Data exploration: *"[...] the moment we visualized the data using the tools provided to us. Finally all those lines of data were converted into a visual representation of the three days of hard work."* (participant from **T1**). Some responses further reflected on the necessity for visualization to see the data in context: *"[...] visualizing the data. For me it first really makes sense and is useful, when I can see it visually, since this makes the data more concrete. Finding out that there were many different*

ways and different tools to visualize it, was nice." (participant from **T3**).

There were also various other points raised in the responses. A participant with a technical background reflected on demystifying working with data: *"[the study] helped me to understand that there is no need of any deep technical knowledge, to start playing with data and applying it [in the design process]"* (participant from **T2**). More participants noted the study helping them better understanding the phenomenon around big data, and increasing their awareness of the online data traces: *"[the study] also made me more aware of the digital footprints I leave online, everyday. Many people are warning about this, but I had not quite understood it until now."* (participant from **T6**).

DISCUSSION

The current study explored the appropriation of a data science workflow by two groups of master-level design students into a design process. In this section, results from the two studies are positioned in HCI and design literature, highlighting further research opportunities.

Gaining domain-knowledge

The two studies differed in working from provided data (Study 1) and capturing data (Study 2), and the participants familiarised with the datasets differently. For Study 1 (master thesis records), we followed guidelines by D'Ignazio (2017) to work with familiar datasets and messy data. Being master-level design students, the participants were familiar with the general domain of the dataset (as writing a master thesis to finish their studies), however, several data properties were unclear for them (having one more year before starting their master thesis). The dataset was not entirely clean (Wickham, 2014), thus the participants needed to do some data cleaning on them. This *'friction'* work turned out to contribute to gaining a more detailed understanding of the dataset. For Study 2, as the participants worked on their ongoing semester project and had done research prior to the study, gaining domain-knowledge was less pronounced. The importance of domain-knowledge has long been acknowledged and researched in data mining (Anand, Bell and Hughes, 1995) and later data science (Waller and Fawcett, 2013). Gaining domain knowledge needs to be considered when pursuing a data science workflow in the design process; access to a dataset (such as stumbled upon open data or a design process at a hackathon) still requires building up the understanding what is inside the dataset. This understanding can be fostered by additional description of the dataset (sometimes called *data dictionary*) to describe the different properties in the dataset. Designers can also use other, qualitative data enquiries for gaining domain knowledge, or can collaborate with a domain expert too.

End-user data tools as an assemblage

The steps of the data science workflow – such as capturing online data or cleaning a dataset – were followed through various end-user data tools that we selected appropriately for the needs and skill levels of the participants. This approach was appreciated, as learning about these different data tools was highlighted as a a major take-away from the studies. But, this approach leads to end-user data tools forming a system assemblage (Kling and Scacchi, 1982), where the different tools enable different actions to be taken on the dataset. Following through the multiple steps of such a data workflow happen by using non-programmatic tools. The system assemblage has positive and negative consequences. The assemblage enables designers to optimize their workflow using different tools for different tasks, choosing more appropriate tools for certain jobs. Furthermore, whilst some tooling is generic, such as a text editor that can perform basic string operations on a dataset (e.g., find and replace), other tools are data type specific (e.g., geo-located data is typically inspected through map-based visualizations, whilst data with numbers and categories are plotted on graphs). However, different tools can require certain formats and data transformations to prepare the input. This complicates the learning curve of different end-user data tools and the assemblage's overall usability.

Question-driven enquiry

Following through a data science workflow in Study 2, the participants struggled initially with the computational thinking required by data acquisition through scraping, and to understand what kind of questions could they possible answer by capturing and analysing data. Participants' understanding increased through an iterative process in defining better questions, and as a consequence capturing more targeted data (approximately half the time of Study 2 was spent on doing multiple iterations). This iterative process of refining the research question and collecting data to extract insights applies the data science workflow of the co-evolution of problem and solution space (Dorst and Cross, 2001).

Designers are exposed to thinking about wicked problems (Rittel and Webber, 1973; Buchanan, 1992) and formulate design questions that generate design spaces (such as 'How might we…?' questions). However, during the study, questions towards falsifiable/provable hypotheses (resembling the *'scientific method')* turned out to be more productive. Throughout iterations, participants both continuously learnt more-and-more about the domain, and also improved the imposed questions that can be addressed via data enquiry. In our

observation, working with digital data for design enquiry requires a more precise question formulation by designers. While a qualitative enquiry such as field observations can be 'forgiving' while conducted, enquiry through digital data collection requires precision in instructing a software tool. In this way, the creativity of designers is channelled into hypothesis and research question formulation.

Creative uses of data exploration

A common data science terminology for the early step of exploring data is *'Exploratory Data Analysis'* (EDA). EDA was originally introduced for the exploration of numerical datasets using a statistical toolbox (Tukey, 1977). Commonly during EDA, various statistical techniques are applied to better understand the data, generate various hypotheses and test those against the data. Yu (1994), following Pierce's pragmaticism explains how deduction, induction and abduction plays a role in EDA: abduction is used to generate a hypothesis, deduction to evaluate the hypothesis and induction to justify the hypothesis with empirical data. Most commonly, data and visual analytics is targeted at using deduction to analyse data (Wong and Thomas, 2004).

Interestingly, the early phase of design is largely influenced by abduction (Kolko, 2009; Dorst, 2011). We observed the use of data as a source of inspiration, following an abductive sensemaking. The approach of **T2** highlights this: they visualised social media hashtags to find subcultural and slang hashtags, looking for knowledge that would have been hard to gain from user interviews or field studies. They used their findings not to prove a hypothesis, but used a creative thought process to explore a phenomenon otherwise they would hardly access. This is a creative way of using data – using data as a generative design tool –, and one where the human abductive sensemaking is necessary to create the right connections.

Designers are trained in making sense of the world following patterns of thought where what is being designed is being informed by a constantly reframed problem space (Dorst, 2011). Our observations indicate that this skillset can be transferred for exploratory data analysis, using an abductive hypothesis generation as a creative process. Further studies in understanding the creative process throughout the data science workflow could help informing new data uses, and in generating future design methods with data.

Limitations

Our study contributes an overview of incorporating a data science workflow in the design process, it was nonetheless conducted with master-level design students. Future research with expert designers and in design practice would support generalizing our findings for designers on all expertise level and designers working in a range of non-educational settings. Furthermore, the current study was limited to working with data collected from online resources. For future research, it would be especially interesting to explore designers working with sensor data, product log-files and so forth. Ultimately we hope this study inspires future research on improving data methods and tools tailored for the particularities of the design process.

GUIDELINES FOR USING DATA IN DESIGN ENQUIRY

The discussion enables to distil the following guidelines how to integrate a data science workflow for designer enquiry:

1) **Access to data is not enough**; although data can be acquired virtually about any phenomenon, it still needs to be made sense of and insights to be concluded to be used in the design process.
2) **Explore data creatively**; designerly sensemaking of abductive synthesis can be applied in data exploration to seek insights and inspirations. In this case, using digital data can make an unseen phenomenon visible.
3) **Navigate the whole process**; the power of using a data science workflow lies in the various steps from formulating a question, acquiring data and then making sense of data for actionable insights.
4) **Use data exploration for its strengths**; qualitative methods to gain knowledge on contexts have their merits, and digital data at best can augment and be used in concert with them.

CONCLUSIONS

The current work presented two exploratory studies elaborating how a data science workflow could be adapted in a design research process. The findings demonstrate that designers transfer their creative capacity to hypothesis forming for data collection and use their designer sensemaking to synthesise data exploration of digital data in design enquiry. Currently existing end-user data tools can help in combining a data science workflow in the design process to harness the ever-increasing abundance of digital data describing different phenomena in the world. In our future work, we aim to continue to contribute to the current debate on method development of data-aware design methods.

ACKNOWLEDGEMENTS

[The acknowledgements have been redacted for anonymity]

REFERENCES

Anand, S. S., Bell, D. A. and Hughes, J. G. (1995) 'The role of domain knowledge in data mining'. Proceedings of the Fourth International Conference on Information and Knowledge Management, pp. 37–43. ACM, New York, NY, USA.

Baumer, B. (2015) 'A Data Science Course for Undergraduates: Thinking With Data', *The American Statistician*. Taylor & Francis, 69(4), pp. 334–342.

Bigelow, A., Drucker, S., Fisher, D. and Meyer, M. (2014) 'Reflections on How Designers Design with Data'. Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces, pp. 17–24. ACM, New York, NY, USA.

Bogers, S., Frens, J., van Kollenburg, J., Deckers, E. and Hummels, C. (2016) 'Connected Baby Bottle: A Design Case Study Towards a Framework for Data-Enabled Design'. Proceedings of the 2016 ACM Conference on Designing Interactive Systems 2016, pp. 301–311. ACM, New York, New York, USA.

Buchanan, R. (1992) 'Wicked Problems in Design Thinking', *Design Issues*, 8(2), p. 5.

Bødker, S. (2006) 'When second wave HCI meets third wave challenges'. Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles 2006, pp. 1–8. ACM, New York, NY, USA.

Cao, L. (2017) 'Data Science: A Comprehensive Overview', *ACM Computing Surveys (CSUR)*. ACM, 50(3), pp. 43–42.

Card, S. K., Mackinlay, J. D. and Shneiderman, B. (1999) *Readings in information visualization: using vision to think*. Morgan Kaufmann.

Churchill, E. F. (2012) 'From data divination to data-aware design', *interactions*, 19(5), pp. 10–13.

Churchill, E. F. (2017) 'Data, design, and ethnography', *interactions*, 25(1), pp. 22–23.

D'Ignazio, C. (2017) 'Creative data literacy: Bridging the gap between the data-haves and data-have nots', *Information Design Journal*, 23(1), pp. 6–18. John Benjamin's Publishing Company.

D'Ignazio, C. and Bhargava, R. (2016) 'DataBasic: Design Principles, Tools and Activities for Data Literacy Learners', *The Journal of Community Informatics*, 12(3).

Dorst, K. (2011) 'The core of "design thinking" and its application', *Design Studies*, 32(6), pp. 521–532.

Dorst, K. and Cross, N. (2001) 'Creativity in the design process: co-evolution of problem–solution', *Design Studies*, 22(5), pp. 425–437.

Dörk, M., Carpendale, S. and Williamson, C. (2011) 'The information flaneur: a fresh look at information seeking'. ACM, pp. 1215–1224. doi: 10.1145/1978942.1979124.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996) 'From Data Mining to Knowledge Discovery in Databases', *AI magazine*, 17(3), p. 37.

Feinberg, M. (2017) 'A Design Perspective on Data'. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems 2017, pp. 2952–2963. ACM, New York, NY, USA.

Giaccardi, E., Cila, N., Speed, C. and Caldwell, M. (2016) 'Thing Ethnography: Doing Design Research with Non-Humans' Proceedings of the 2016 ACM Conference on Designing Interactive Systems 2016, pp. 377–387. ACM, New York, NY, USA.

Grammel, L., Tory, M. and Storey, M. (2010) 'How Information Visualization Novices Construct Visualizations', *IEEE Transactions on Visualization and Computer Graphics*. IEEE, 16(6), pp. 943–952.

Gray, J., Chambers, L. and Bounegru, L. (2012) *Data Journalism handbook*, *Data Journalism handbook*. http://datajournalismhandbook.net/1.0/en/ (Last retrieved: 29 March 2018).

Hill, B. M., Dailey, D., Guy, R. T., Lewis, B., Matsuzaki, M. and Morgan, J. T. (2017) 'Democratizing Data Science: The Community Data Science Workshops and Classes', in Matei, S. A., Jullien, N., and Goggins, S. P. (eds) *Big Data Factories: Collaborative Approaches*. Cham: Springer International Publishing (Big Data Factories: Collaborative Approaches).

Kandel, S., Paepcke, A., Hellerstein, J. M. and Heer, J. (2012) 'Enterprise Data Analysis and Visualization: An Interview Study', *IEEE Transactions on Visualization and Computer Graphics*, 18(12), pp. 2917–2926.

Kling, R. and Scacchi, W. (1982) 'The Web of Computing: Computer Technology as Social Organization', in *Advances in Computers Volume 21*. Elsevier (Advances in Computers), pp. 1–90.

Kolko, J. (2009) 'Abductive Thinking and Sensemaking: The Drivers of Design Synthesis', *Design Issues*, 26(1), pp. 15–28.

Lycett, M. (2013) '"Datafication": making sense of (big) data in a complex world', *European Journal of Information Systems*. 22(4), pp. 381–386. Palgrave Macmillan UK.

Marchionini, G. (2006) 'Exploratory search: from finding to understanding', *Communications of the ACM*. ACM, 49(4), pp. 41–46.

Mauri, M., Elli, T., Caviglia, G., Uboldi, G. and Azzi, M. (2017) 'RAWGraphs'. Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter 2017, p. 28. ACM, New York, NY, USA.

Mortier, R., Haddadi, H., Henderson, T., McAuley, D. and Crowcroft, J. (2014) 'Human-Data Interaction: The Human Face of the Data-Driven Society', *SSRN Electronic Journal*.

Rittel, H. W. J. and Webber, M. M. (1973) 'Dilemmas in a general theory of planning', *Policy Sciences*, 4(2), pp. 155–169. Kluwer Academic Publishers.

Sanders, E. B.-N. and Stappers, P. J. (2008) 'Co-creation and the new landscapes of design', *Co-Design*, 4(1), pp. 5–18. Taylor & Francis.

School of Data (2016) *Online Courses*, *School of Data*. https://schoolofdata.org/courses/ (Last retrieved: 29 March 2018).

Speed, C. and Oberlander, J. (2016) 'Designing from, with and by Data: Introducing the ablative framework', *Proceedings of the the 50th Anniversary Conference of the Design Research Society*.

Tukey, J. W. (1977) *Exploratory data analysis*. Addison-Wesley.

Waller, M. A. and Fawcett, S. E. (2013) 'Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management', *Journal of Business Logistics*, 34(2), pp. 77–84.

Wickham, H. (2014) 'Tidy data', *Journal of Statistical Software*, 59(10), pp. 1–23.

Wong, P. C. and Thomas, J. (2004) 'Guest Editors' Introduction--Visual Analytics', *IEEE Computer Graphics and Applications, 24(5):20-21*, 24(5), pp. 20–21.

Yu, C. H. (1994) 'Abduction? Deduction? Induction? Is There a Logic of Exploratory Data Analysis?'. Annual Meeting of American Educational Research Association. New Orleans, Louisana.