This forum provides a space to engage with the challenges of designing for intelligent algorithmic experiences. We invite articles that tackle the tensions between research and practice when integrating AI and UX design. We welcome interdisciplinary debate, artful critique, forward-looking research, case studies of AI in practice, and speculative design explorations. **— Juho Kim and Henriette Cramer, Editors**

UX of Data: Making Data Available Doesn't Make It Usable

Laura Koesten, University of Vienna, **Elena Simperl,** King's College London

Data plays an important part in our daily lives. It shapes how we view the world and, for better or worse, informs the decisions we make. Despite controversies around who collects and does what with data, data-centric work is seen as critical to solving the most complex problems of today, from climate change and security to health crises and inequality. It spans many types of tasks in various professions, undertaken by multiple stakeholders, who may or may not share a common understanding of the domain or the task. It also raises questions of data literacy, inclusion, and fairness in ensuring that the value the data creates is shared as widely as possible.

In our research over the past five years, we have asked what makes interaction and working with data unique for different types of audiences.

Data is increasingly available online, for instance, in science and government. Organizations invest in infrastructure to share data internally and with their partners. Machine learning is heavily dependent on the availability of datasets to train algorithms, but creating these from scratch comes with costs. Overall, it means that more and more data is used outside the context in which or for which it was produced—reuse is often the best way to add value to the data and perhaps to recover some of the investment that went into publishing it and policy support.

But how comfortable are we with reading, interpreting, and working with other people's data? Using a dataset beyond the context in which it originated remains challenging. Simply making data available, even when following existing guidlines and best practices, does not mean it can be easily used by others. Further, there is little evidence that such guidance, including technical standards and community support, do indeed lead to more data engagement and generate value.

This is not surprising if we consider the complex environment around data. Data-centric work is often carried out in diverse teams, drawing upon different skills, including domain knowledge, statistics, data science, engineering, and project management. While some data tasks are heavily automated, others depend on people's judgments and their decisions on how and what to record, document, and analyze, and what to leave out. Similar to other fields, there is a tension between the resources required to document each step in detail and the ability at a later point in time to reproduce what happened. This is especially the case as deep, exploratory or data-curation work is often seen as secondary to model design or machine learning.

Insights

- → We need better ways to record and communicate that data represents only a perspective of reality.
- → There is no typical data user, just typical data tasks. Data tools need to provide better support for each task in the data life cycle.
- → Interaction challenges for making data usable include making data understandable, supporting collaboration, managing changes, and facilitating different modes of access and user engagement.

In our research, we learned a great deal about how different audiences go about finding data online, how they choose what data to work with, and how they make sense of data that others have created. We explored scenarios where these tasks were carried out by individuals as well as in teams of data professionals. Using both qualitative and quantitative methods, we carried out several studies drawing upon a diverse range of data sources, including interviews, diaries, surveys, and digital traces of user engagement on data platforms such as Kaggle, GitHub, and open-data portals.

Our initial exploratory studies were framed in theories about information seeking: We asked people what their data tasks were—what they did with data in their jobs, how they searched for it, how they evaluated and selected datasets as search results, and how they explored and understood data with which they were not familiar.

We also used different forms of summarization (verbal and textual) to document the cognitive processes involved in data sensemaking. We asked people to describe and discuss data they knew or data that was new to them, and identified common activity patterns [1]. We could distinguish among three clusters: *inspecting, engaging with content,* and *placing data in contex*t. Our research suggests that these relate to certain characteristics of a dataset that are conducive to sensemaking.

DATA CAPTURES A PERSPECTIVE, BUT HOW DO WE COMMUNICATE THAT BETTER?

We want to highlight *placing* as an activity in its own right, which relates to the less tangible aspects of data understanding. When engaging with a dataset, people try to place data in different contexts. First, we'll look at the immediate environment of the dataset's creation—for example, a study setup or the conditions surrounding data collection, with timeframes, geospatial boundaries, or configurations of collection devices. Next, more conceptually, we'll note the norms of the discipline in which the data was collected, including methods of analysis and validation, as well as limitations (e.g., common margins of error). Finally, we also place data within the world, gauging how representative it is and reflecting on assumptions about how much it mirrors reality. This includes assessments of the data itself, but also of what might be missing from the data (e.g., if the dataset contains countries, are they complete? Which source of reference are we using and can we trust it?).

At a practical level, these multiple contexts need to be recorded and communicated to support data sensemaking. This includes, among others, protocols, tools, documentation, and other research objects that were involved in creating the dataset.

The assumptions and categorization activities during data creation influence and shape the resulting data (e.g., [2]). Categories are created, for instance, when we decide how to measure something, to what level of detail, and what to exclude. Omitting this information can have unintended effects in data reuse. Historically, some domains and disciplines have been more sensitive to these aspects than others. In AI, there is a growing awareness that tools and scientific methodologies need to do more to understand, collect, and consider data-related processes, to support users in interacting with datasets beyond simply releasing and describing them.

Actively supporting people in understanding and working with datasets is critical to a good user experience.

THERE IS NO TYPICAL DATA USER, JUST A SET OF TYPICAL DATA TASKS

Our studies have shown repeatedly that everyone, including data professionals, can be experts in one data task and novices in the next. At the same time, information needs on datasets are heavily task dependent.

There is risk and uncertainty attached to using someone else's data. We found that the most tedious task for many is the process of assessing a dataset. To be fit for a task, a dataset needs to be *usable*, *relevant*, and have *good enough quality*. Our studies suggested that these notions are different for data than for other types of information such as text or multimedia.

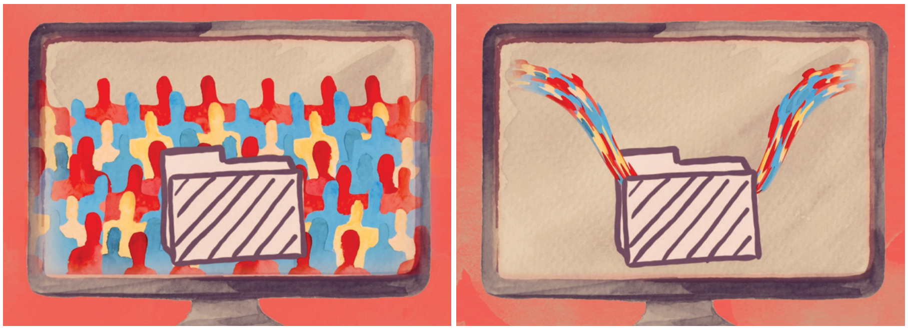
To be *usable*, a dataset might need to be of a certain size, have the right format and license, and not contain sensitive information. To be *relevant*, a dataset needs to cover the topic of interest at the right level of detail. Finally, data *quality* is a multidimensional concept. The literature offers many frameworks to choose from. In our studies, we noticed that people's perception of what constitutes good-quality data changed as they engaged with the data. This makes sense, considering that quality is inherently task dependent. To use Christine Borgman's words, "one person's signal is another's noise" [3]. In our research, we were nevertheless able to identify some recurring themes pertaining to quality. These include indicators such as completeness, provenance, accuracy, cleanliness, and consistency of formatting. Not all facets of quality are straightforward to assess, for instance, understandability and the methodology used to create the dataset. Data quality might need to be fuzzy as a concept to have meaning, but we can think about how it translates to certain tasks and how to capture, measure, and communicate it.

MAKING DATA USABLE MEANS DESIGNING FOR SPECIFIC INTERACTION CHALLENGES

In the following sections, we highlight some of the key challenges for data reuse: making data understandable, accessing data via different access modes to cater to different tasks, the management of changes in data, as well as the collaborative nature of data work processes.

Making data understandable. One of the key challenges to data reuse is formats and capabilities that make it useful in as many contexts as possible. Actively supporting people in understanding and working with datasets is critical to a good user experience. Our results confirm that structured documentation (e.g., ReadMe files) has a positive impact on dataset reuse. This is evident both from asking people and from analyzing the activity logs of data platforms. We found that textual descriptions of data are of high importance, as they often constitute the first points of interaction between a user and a dataset. We studied hundreds of dataset summaries and identified common information categories that people want to know to make sense of new datasets. This includes, for instance, information about the time frames or location that the data refers to, or about the types and ranges of values in a dataset.

There is plenty of advice on how to make data easier to reuse, including technical standards, legal frameworks, and guidelines. However, their focus has been mostly on operational problems such as interoperability and machine readability. More recently, researchers have proposed formats that document context to datasets beyond what is typically stored as metadata, for example, data nutrition labels (https://datanutrition.org/), datasheets for machine learning data [4], or minimal information standards (https://fairsharing.org/standards/) in life sciences. Such proposals provide important reference points for thinking about data reuse. However, most of them



Data-collection and categorization activities.

result in high documentation efforts, and some offer little guidance about what exactly should be documented. While there are efforts to auto-summarize data, it is unclear how useful the results will be in specific contexts.

The rise of FAIR (Findable Accessible Interoperable Reusable; https://www.go-fair.org/fair-principles/) data science showcases how grassroots and policy efforts could lead to more useful, reusable data. FAIR stands for a set of principles, technical and otherwise, to find, access, and integrate data from all branches of science. We believe there is a large design space to create tools and interfaces to (semi-)automatically capture impactful documentation. Rather than designing for a specific type of user, tools should be designed to embrace different levels of expertise, with drill-down capabilities to reach the desired level of detail. This could, for instance, be realized through summary description of datasets, tailored to support known sensemaking activities and accompanied by column descriptions for more information.

Different data-access modes and managing changes. People are collaborating with data in different contexts and therefore need different data formats and representations. Tools should support a range of data-literacy and technical skills. Being able to easily plot data, access subsets of data, or create relations between tables helps build a shared understanding between collaborators. There are increasingly tools available that work with multiple data representations. However, we found that users still perceive data workflows as fragmented and switch between different tools for various tasks, which adds cognitive effort.

The ability to access earlier versions of a data source has long been recognized in tools that allow us to work together with text or code. This includes, for example, being able to revert to a version before particular cleaning methods have been applied, or notification services when someone makes a change in the dataset or when an updated version is published. Version control tailored to datasets needs to track changes at varying levels of granularity, such as rows, columns, and individual cells. Again, this is supported by various tools, but not necessarily easy to use or customized for different data modalities, according to our participants.

Supporting collaboration and social interaction, during data creation and use. Both creating and using data are often done in a team. Our research has shown that the various requirements, expectations, choices, assumptions, and limitations are often informal and mostly undocumented.

As discussed earlier, there is increased awareness in many domains, including AI, of the challenges this creates for potential users of the data. Further, the mere act of reusing someone's data can be seen as a form of collaboration—between the user and the creator.

The literature has shown that social interaction, including the ability to ask questions about the data, supports reuse. However, communication capabilities need to be tailored to data work; for instance, they should facilitate sharing, commenting, and discussions around particular data fragments, as well as the levels of individual cells, rows, and columns [5].

We cannot anticipate documenting everything in a way that serves all types of data tasks and accounts for multiple backgrounds and levels of literacy. Instead, our findings underline the importance of engagement around datasets, including discussions, feedback, reviews, ratings, and means to contact data creators. User communities and peer support can complement documentation efforts and make dataset maintenance sustainable.

The bottom line is that we cannot see datasets as usable end products without telling the story of how they were made. Because the story is complex, the user experience of data relies on tools and environments that try to do exactly that: embedding datasets in the rich context of their creation and use.

Endnotes

1. Koesten, L., Gregory, K., Groth, P., and Simperl, E. Talking datasets – understanding data sensemaking behaviours. *International Journal of Human-Computer Studies 146* (Feb. 2021).
2. Bowker, G.C. and Star, S.L. *Sorting Things Out: Classification and Its Consequences.* MIT Press, 2000.
3. Borgman, C.L. *Big Data, Little Data, No Data: Scholarship in the Networked World.* MIT Press, 2015.
4. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé III, H. and Crawford, K. Datasheets for datasets. arXiv preprint arXiv:1803.09010, 2018.
5. Koesten, L., Kacprzak, E., Tennison, J., and Simperl, E. Collaborative practices with structured data: Do tools support what users need? *Proc. of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, 2019.

Laura Koesten is a postdoctoral researcher at the University of Vienna. She obtained her Ph.D. at the Open Data Institute and the University of Southampton, U.K. In her research, she is looking at ways to improve human-data interaction by studying sensemaking with data, data reuse, and collaboration in data science.
→ laura.koesten@univie.ac.at

Elena Simperl obtained her doctoral degree in computer science from the Free University of Berlin, and her diploma in computer science from the Technical University of Munich. She is a Fellow of the British Computer Society and a former Turing Fellow, and has held positions in Germany, Austria, and the University of Southampton prior to joining King's College London in 2020.
→ elena.simperl@kcl.ac.uk

DOI: 10.1145/3448888