Digital Society Initiative

What hinders researchers from making more frequent use of unstructured, digital data for research?

Results from a survey among UZH scientists

Jana Sedlakova, Viktor von Wyl, Thomas Friemel, Nico Pfiffner, Markus Wolf, Andrea Horn, Wintsch, Kaspar Staub, Gerold Schneider, Tilia Ellendorff, Oliver Grübner, Fabio Rinaldi, Giovanni Spitale, Dominik Alois Ettlin, Jürgen Bernard, Paola Daniore, Christina Haag, Chloé Sieber, Mina Stanikic, Sonja Schläpfer

Executive Summary

Unstructured digital data, defined as information in non-tabular formats such as texts, audio files, images, or complex sensor data, offer exciting potential for research. However, the utilization of unstructured data carries numerous challenges.

We conducted an online survey among scientists working at the University of Zurich (UZH). Our study aimed to understand the type and frequency of unstructured data use, as well as to identify skill- and infrastructure needs among the UZH-researchers.

Our findings suggest that textual data – such as patient-reported data, data from social media or administrative data – are by far the most frequently used unstructured data type. Researchers experienced multiple barriers when working with unstructured data, particularly regarding data quality and data harmonization. Other challenges pertained to the alignment of the available unstructured data and analysis methods with study questions.

Our findings underscore the need for further education and capacity building on methods and the use of unstructured data, particularly for early career researchers. Furthermore, there is also a growing need for interprofessional collaboration together with fostered interdisciplinary exchange. In response to these findings, the Health Community of the Digital Society Initiative is creating an environment for interprofessional and interdisciplinary exchange and is developing teaching modules on textual data analysis and on study question development. These teaching modules will be openly accessible to all UZH researchers as the textual data were most frequently used by the UZH researchers.

Overview

1 Background

Unstructured data, defined as digital information in a non-tabular format, create novel inroads towards gaining more comprehensive descriptions, for example, of health, well-being, or disease phenotypes. Examples of unstructured data include textual or audio data from electronic health records, physicians' notes, social media posts, diaries, or data from sensors and wearables.

Not only in the health domain, but also in other domains, the combination of unstructured data with structured research databases harbors great research potential. However, multiple challenges inhibit the more frequent exploitation of unstructured information for research purposes.

In order to gain a deeper understanding of specific hurdles for unstructured data use by researchers, the DSI-Health Community has conducted a survey among UZH researchers to elicit researchers' opinions and needs regarding the use of unstructured data in research. This survey was aligned with an additionally performed literature review that aimed at answering the following question:

"*How to reach and ensure proper (systematic, reliable, valid, effective, ethical) integration of unstructured data from different sources into the health research to gain new knowledge and enhance the research quality?*"

The survey and the upcoming literature review aimed to provide an overview of 1. researchers' motivation of and experiences with utilizing unstructured data in their past, current or future research projects, 2. methodological approaches and frameworks to define a research questions and design for projects with unstructured data, 3. researchers' needs and obstacles when utilizing unstructured data, and 4. Reflective questions about the possible changes in scientific activity when using unstructured data in research. The survey was also intended to shed light on potential unmet training and infrastructure needs at the UZH.

2 Survey methods

Based on the conducted literature review, we defined unstructured data as raw data that are not in a pre-defined, recognizable structure (e.g., tables) or that are in a loosely defined structure, which need further formalization, pre-processing, or feature extraction (e.g., from real-time data streams or small-interval measurements such as electrocardiograms, mobility measurements). Typical examples of unstructured data include textual data, sensor data from wearable devices, audio, and image data.

The structure and contents of the survey were discussed with the Executive Committee of the Joint Project in the DSI-Health Community and afterwards, we received feedback from the DSI Health-Community. Thomas Friemel and Nico Pfiffner from the DSI Communication Community and DIZH Data Donation Lab have provided additional input for the survey design.

The survey was structured in a demographic part and four main blocks. The focus in the first block was to learn about researchers' motivation to use unstructured data, types of unstructured data they used in their previous research or would like to use in future research and whether they had previously integrated the data. The second block aimed at understanding how researchers framed the use of unstructured data and how it aligned with their research questions and hypotheses. The third block focused on obstacles faced by researchers and their requirements when working with unstructured data. Finally, the fourth block was theoretical in nature and focused on questions regarding the possible changes in the scientific practice that might be caused by increased used of unstructured data.

The survey was implemented in Qualtrics and distributed among UZH researchers from all scientific disciplines since the topic of unstructured data is not restricted to health research only. We used the

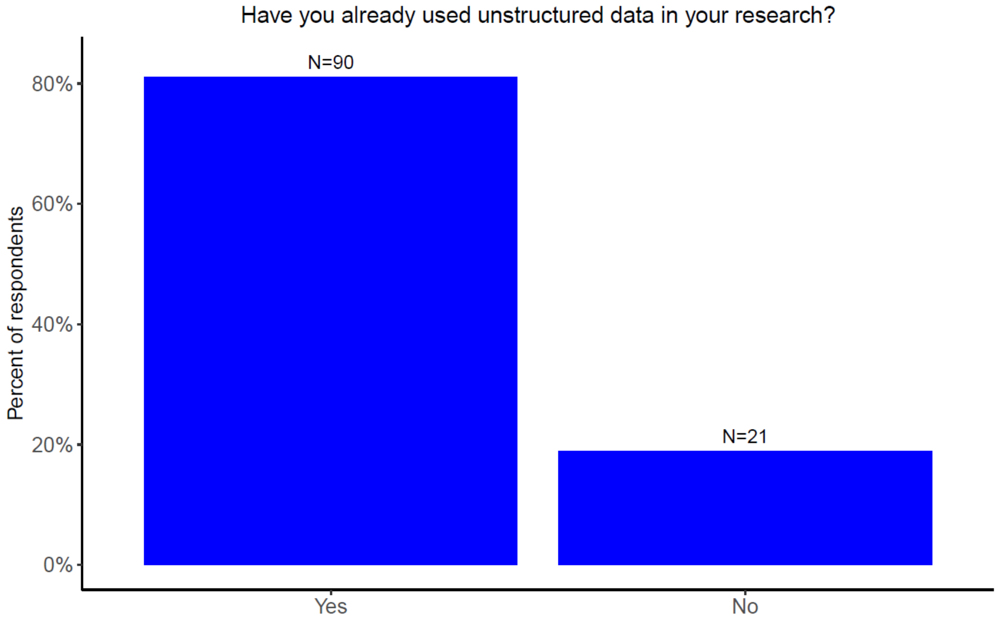
UZH mailing list for distribution and included masters' students, PhD candidates and all researchers with a PhD or higher degree. Furthermore, we used DSI and Participatory Science Academy mailing lists to distribute the survey. In total, the survey was sent to approximately 20 500 researchers. The survey was online from 26. July 2021 to 26. September 2021. For the survey distribution, the central UZH mailing list from President's service and DSI communication channels were used.

Results

Our survey was answered by 177 researchers across all UZH faculties and from diverse research domains such as computer science, linguistics, health research and psychology, law and economics, chemistry, literature and many others. Approximately 30% of all participants answered all survey questions. Almost half (49%) of the researchers were junior researchers.[1] 69 of the responders were involved in health research.

To follow, we report the main results of the survey and provide recommendations based on the upcoming literature review and survey results. In the results, we compare answers from junior and senior researchers.

1. The majority of researchers have already worked with unstructured data

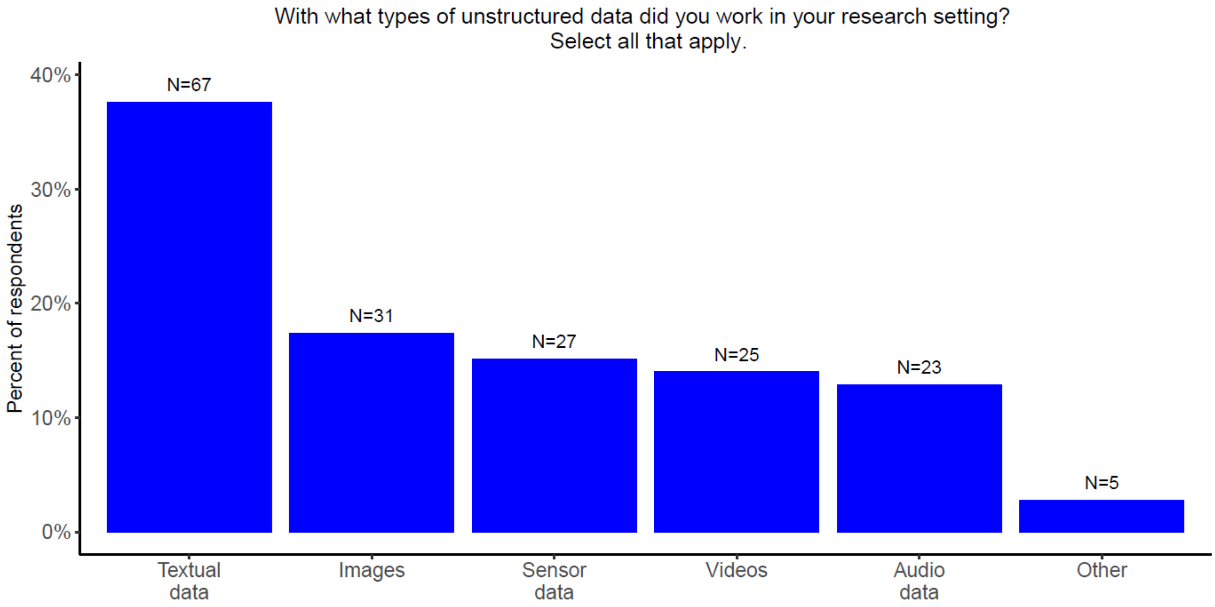


The majority of researchers (80%) across all faculties and independent of their level of seniority have already used unstructured data in their research.

The result reflects the importance of unstructured data sources generated, processed or analyzed by digital technologies that can enrich structured datasets as well as research in general.

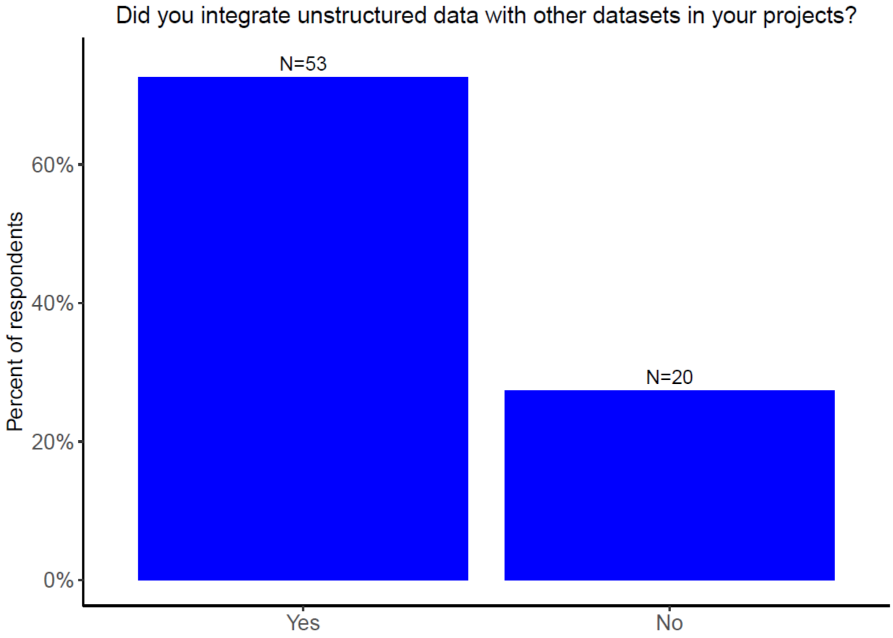
[1] We refer to junior researchers as researchers with a Bachelors' or Masters' degree; and to senior researchers as researchers with a PhD or a higher degree.

2. The majority of researchers have worked with textual data



The type of unstructured data which was most commonly used was textual data, originating mainly from text documents such as electronic health records, journals or pdf files. Images were also a frequently used type of unstructured data. Finally, many researchers have also worked with sensor data predominantly from wearable sensors. The observed distribution of researchers' answers regarding types of unstructured data was almost identical for junior and senior researchers.

3. The majority of researchers have integrated unstructured data with other data sources.

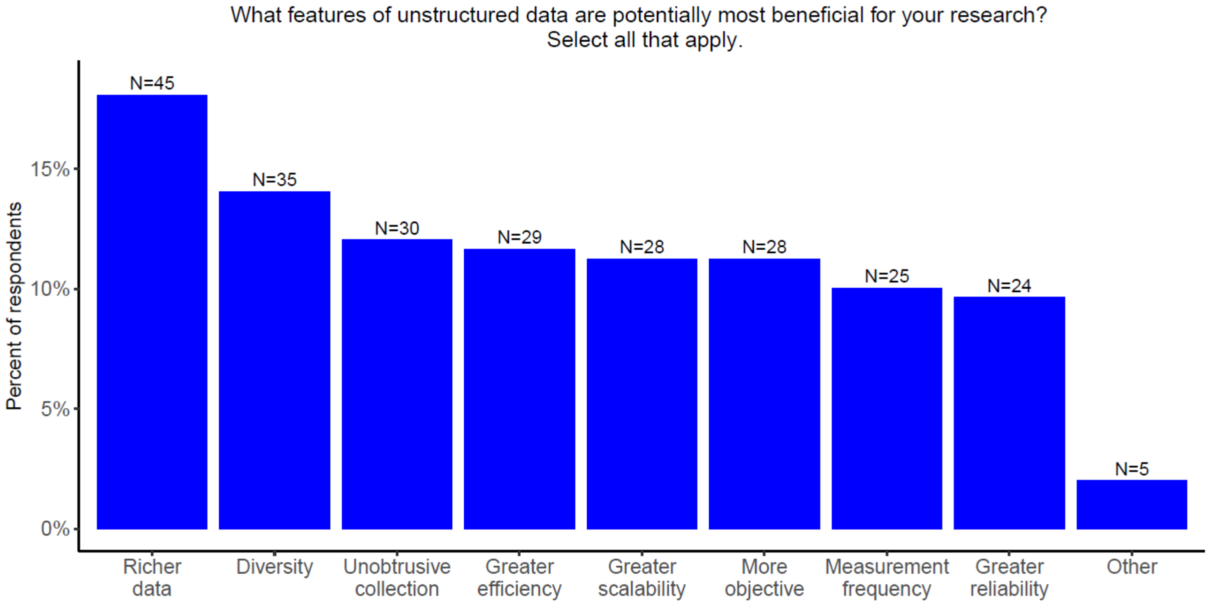


The majority of researchers has experience with unstructured data integration in their research. Here, 77% of the senior researchers and 68% of the junior researchers have integrated unstructured data with other data sources.

Researchers integrated unstructured data usually with tabular data, metadata, other types of unstructured data, survey data and clinical data such as patient reported outcomes (PROMS), diagnostic and measurement data or data from databanks.

This result reflects the importance of data integration in research. Unstructured data can contribute to the enrichment of other, structured and established datasets.

4. Features



The features of unstructured data found to be potentially the most beneficial for research projects were their richness[2], unobtrusive collection[3], greater level of objectivity[4], and greater efficiency regarding their collection and the possibility of the reuse of the data.

Most senior researchers considered the richness of data (58%) as the most beneficial features. Greater reliability and diversity were mentioned least often (34%). Junior researchers also chose richness (69%) and diversity (66%) as the most important reasons for using unstructured data. The least mentioned feature was reliability (31%) and greater measurement frequency (34%).

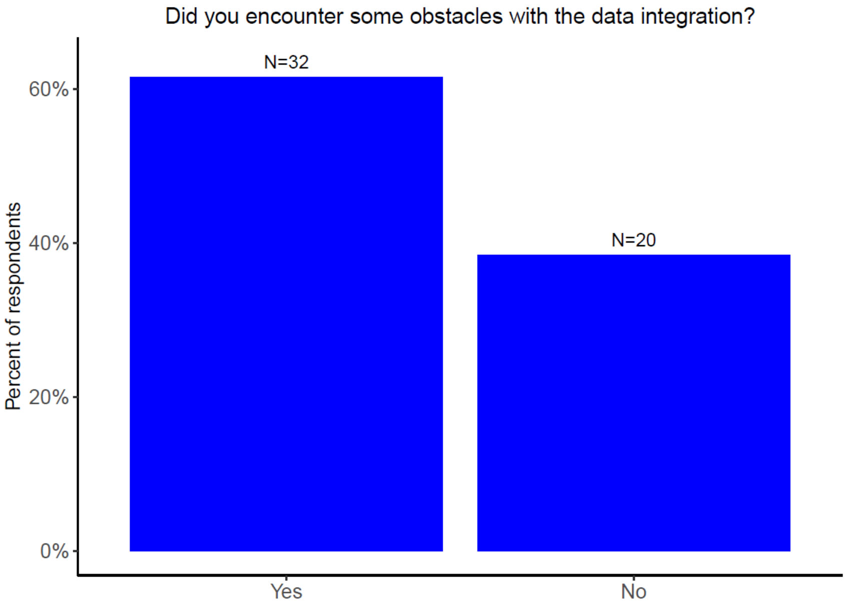
These data shed light on why unstructured data are increasingly popular and used in research. The wealth and richness of unstructured data can be seen as a double-edged sword. On the one hand, richness of these novel data provides opportunities in terms of enriching information and knowledge in a specific research field. On the other hand, this richness is also inevitably connected with challenges regarding data processing, analysis and interpretation. These common challenges were identified both in the survey as well as in the ongoing literature review.

[2] Data provides more details and deeper insights or covers additional dimensions.

[3] Data can be collected with limited involvement of participants or tasks required from them.

[4] Data is collected unobtrusively, in a natural environment and without observer's paradox – when collected data is influenced by the fact that the subject is observed.

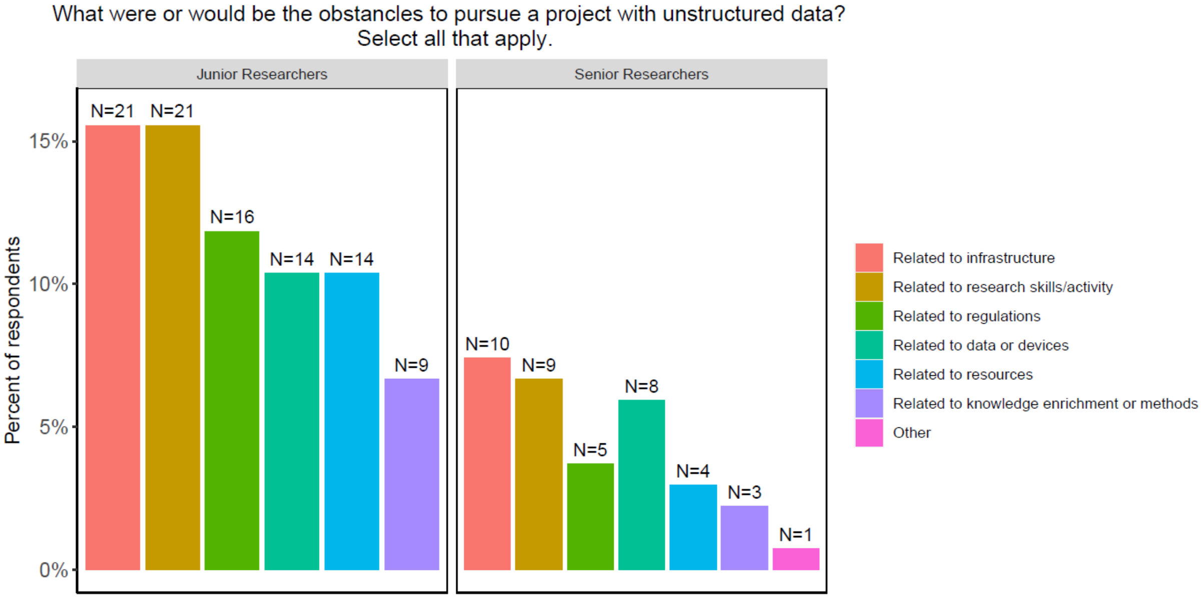
5. The majority of researchers have encountered obstacles when working with unstructured data



Obstacles regarding data integration were encountered by the majority of researchers. Here, 67% of senior researchers and 56 % of junior researchers have encountered obstacles with data integration. The most frequent obstacles mentioned in the open-ended questions were problems with 1) data quality, 2) methodological obstacles, 3) alignment with research design, 4) standardization issues, 5) regulatory and ethical issues. These categories can be described as follows.

1) Problems regarding data quality included missing, noisy or inaccurate data and lack of contextual information and metadata. These problems tend to translate to broader issues such as validity and replicability problems.
2) Methodological problems were firstly linked to the heterogeneity of data, which might limit its interpretation and consequently replicability of studies. Secondly, methodological issues included lack of skills and knowledge on methods needed for data pre-processing and cleaning. Finally, methodological problems were also connected with infrastructure problems when data access and sharing was difficult and insufficient resources were available for data management, analysis, and processing.
3) Problems with the alignment of data with research design were expressed in the context of research question and goals. The large quantity of data that it is often general might decrease the data relevancy for a research question or make it difficult to find a proper research question for the data to be analyzed. Another problem was the choice of the right methods for analysis and interpretation of unstructured data.
4) The non-standardized data makes it difficult to merge datasets or link data to an individual patient. The lack of standardization makes data sharing, access and storage difficult.
5) Regulatory and ethical issues were mentioned in the context of anonymization problems, privacy, lack of informed consent and insufficient overview over regulatory standards regarding data collection and use.

6. Types of Obstacles

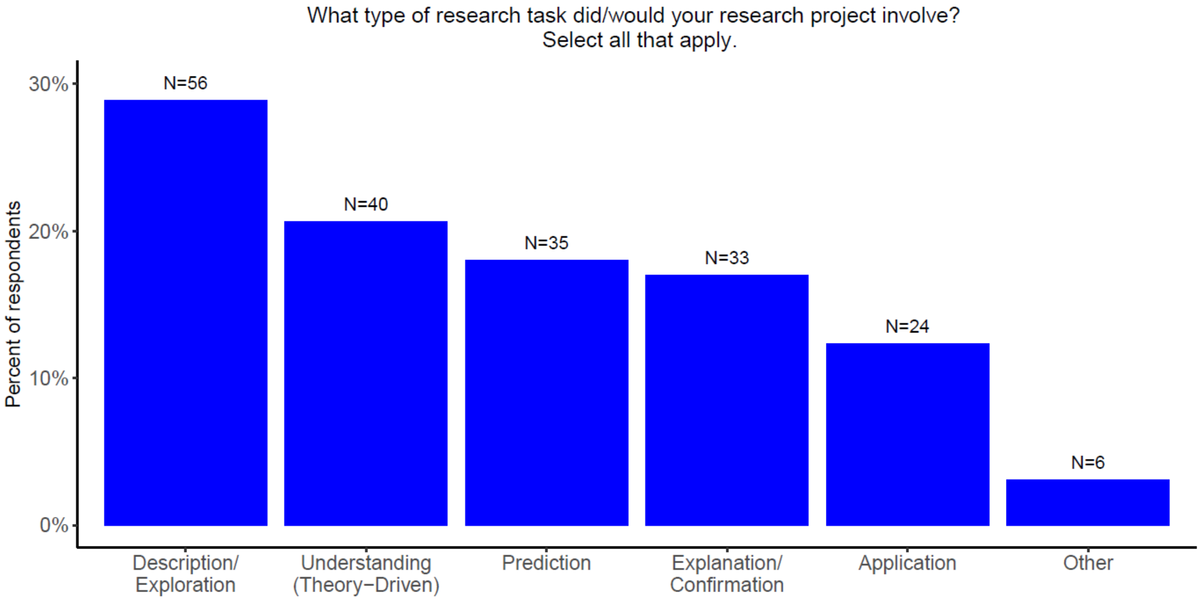


The most frequently mentioned obstacle categories were related to infrastructure (e.g., access to data, tools for analysis), unavailable research skills and legal or ethical regulations.

7. Hypothesis and Research Question

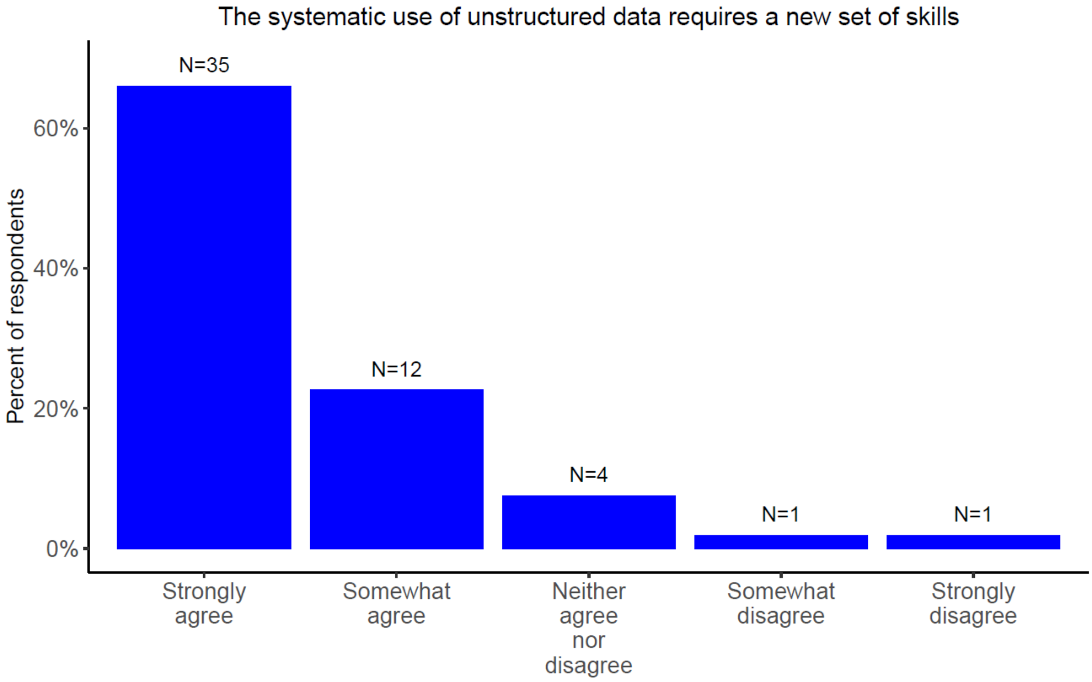
Researchers were asked to provide a research question and hypothesis for their current research projects involving unstructured data. Even though the majority of researchers described the hypothesis and research question, those answers were often formulated in a broad and general manner. Many of them used unstructured data with an explanatory approach to generate a hypothesis for future research. A notable difference between junior and senior researchers was observed. Junior researchers answered more often that they were not sure or clear about the hypothesis and research question.

8. Research Tasks



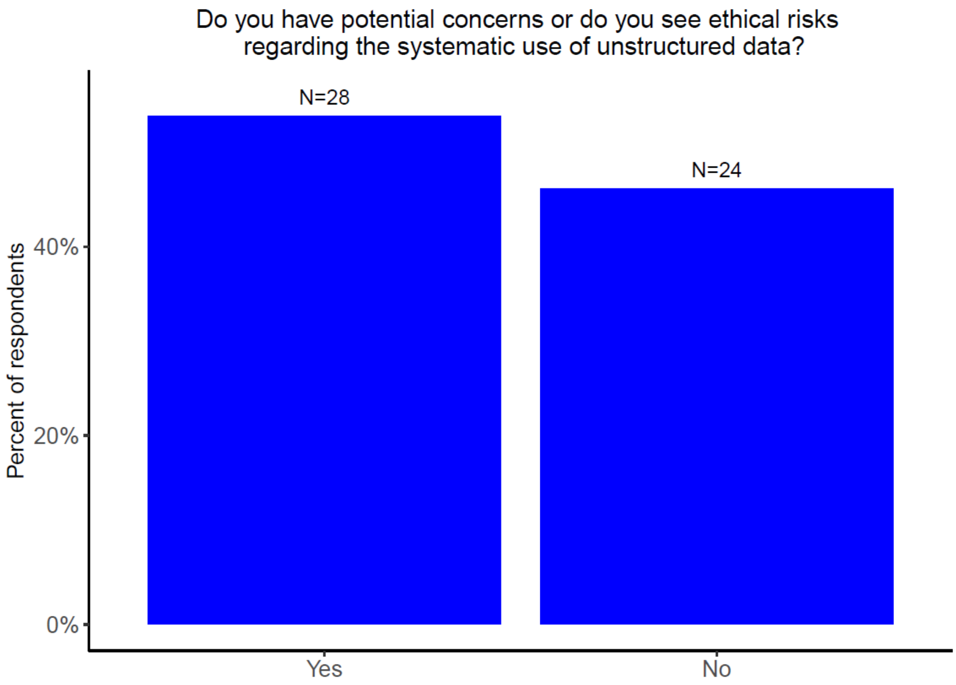
The unstructured data were used in the research mainly with the purpose of exploration. Researchers aimed at finding new patterns, correlation or generating a hypothesis. There was no observed difference between senior and junior researchers. Among researchers from the health domain, junior researchers more often aimed at descriptive goals and goals connected with patient-monitoring as well as improvement of patient-doctor communication, while senior researchers mainly aimed at the goals of clinical diagnosis and prediction.

9. Skills



The majority of researchers agreed that the use of unstructured data requires a new set of skills in research and facilitation of new interdisciplinary collaborations. This was also reflected in the "obstacles" section of the survey where researchers expressed the need for trainings and capacity building regarding methods and tools to deal with unstructured data. No difference between senior and junior researchers was observed.

10. Ethical issues



More than half of the researcher acknowledged that the use of unstructured data is connected with ethical risks. The most often described ethical challenges of unstructured data use are privacy, data protection, data sharing and the risk of biased data that might lead to biased conclusions. Several researchers expressed the need of having a better guidance in ethical issues. Ethical and legal issues pertaining unstructured data use can be part of the capacity building efforts.

Summary and Conclusion

Results from both the survey and upcoming literature review have shown that there is a need for capacity building regarding unstructured data integration to make full use of the data in the research. Furthermore, research efforts to include unstructured data require interdisciplinary, collaborative teams which need new interdisciplinary and integrative skills, research tools and methods to facilitate the exchange. This intensifies the need for new education tools and curricula. Among the results, a need for better guidance and clarity about how to place the unstructured data use in the context of research question and research design was reflected. Furthermore, the infrastructure challenges were a reoccurring topic.

Therefore, we recommend development and inclusion of new courses in the curricula for PhD researchers and other researchers working with unstructured data. Another recommendation is to develop an infrastructure that is suitable for these novel types of data. In response to these findings, the Health Community of the Digital Society Initiative is undertaking several efforts to accelerate capacity building in natural language processing (e.g., through workshops in collaboration with the Text Crunching Center, led by Gerold Schneider), as well as through the creation of a novel online teaching module on how to develop sound research questions when working with unstructured data. These teaching offerings are and will continue to be broadly accessible to researchers within the University of Zurich.

Appendix

Table 1: Overview of the Results (n=177 surveys)

Result	Provided Answers	Missing Answers
1	111	66
2	85	92
3	73	104
4	59	118
5	52	125
6	50	127
7	69	108
8	53	124
9	53	124
10	53	124