

Interactive Machine Learning: Managing Information Richness in Highly Anonymized Conversation Data

Ari Alamäki, Lili Aunimo, Harri Ketamo, Lasse Parvinen

▶ **To cite this version:**

Ari Alamäki, Lili Aunimo, Harri Ketamo, Lasse Parvinen. Interactive Machine Learning: Managing Information Richness in Highly Anonymized Conversation Data. 20th Working Conference on Virtual Enterprises (PRO-VE), Sep 2019, Turin, Italy. pp.173-184, 10.1007/978-3-030-28464-0_16 . hal-02478744

HAL Id: hal-02478744

https://inria.hal.science/hal-02478744

Submitted on 14 Feb 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution| 4.0 International License

Interactive Machine Learning: Managing Information Richness in Highly Anonymized Conversation Data

Ari Alamäki[1*], Lili Aunimo[1], Harri Ketamo[2] and Lasse Parvinen[3]

[1] Haaga-Helia University of Applied Sciences, Helsinki, Finland
{ari.alamaki, lili.aunimo}@haaga-helia.fi,
[2] HeadAI Oy, Pori, Finland
harri.ketamo@headai.com
[3]Terveystalo Oy, Helsinki, Finland
lasse.parvinen@terveystalo.com

Abstract. This case study focuses on an experiment analysing textual conversation data using machine learning algorithms and shows that sharing data across organisational boundaries requires anonymisation that decreases that data's information richness. Additionally, sharing data between organisations, conducting data analytics and collaborating to create new business insight requires inter-organisational collaboration. This study shows that analysing highly anonymised and professional conversation data challenges the capabilities of artificial intelligence. Machine learning algorithms alone cannot learn the internal connections and meanings of information cues. This experiment is therefore in line with prior research in interactive machine learning where data scientists, specialists and computational agents interact. This study reveals that, alongside humans, computational agents will be important actors in collaborative networks. Thus, humans are needed in several phases of the machine learning process for facilitating and training. This calls for collaborative working in multi-disciplinary teams of data scientists and substance experts interacting with computational agents.

Keywords: interactive machine learning, unstructured text, big data, anonymisation, information richness, collaboration, privacy

1 Introduction

Sharing data between organisations is becoming an important process for the co-creation of value in collaborative networks [1,2]. But in most business or scientific research cases, sharing data over organisational boundaries requires anonymisation. For example, the European General Data Protection Regulation (GDPR) allows the secondary use of data, but pseudonymisation or anonymisation is required depending on the intended use for that information (EU GDPR, Article 89). There is, however, little research on how anonymisation affects the way in which data are used or how

machine learning algorithms are able to process highly anonymised, unstructured textual data.

Sharing data between multiple organisations is becoming more and more common, with organisations looking to enrich their own data analysis by combining it with third party data. Gaining deeper business insight and ensuring the reliability of results both require the processing of several data sets and a collaborative network of organisations. Many organisations are also seeking new value by combining inter-organisational data and advanced data analytics methods such as machine learning [1]. Sharing data will become a crucial process in the digital transformation of business processes in the value chains of ecosystems. Additionally, machine learning is becoming a vital part of the process of digitising data that has been analysed.

There are numerous benefits to sharing data sets for collaborative networks of organisations. But there are also challenges. One is data privacy, anonymisation and the loss of information richness. Preserving data privacy is very important from the points of view of both lawful use and trust. Very often, data that have monetary value include confidential information. Organisations protect their customers' privacy by anonymising data before they use them in their own business processes, but they also anonymise data that are shared with their partner companies. However, anonymisation typically reduces the information richness of the data and thus also the value that can be created from the data.

Prior research on the relationship between anonymisation and information richness in sharing data between organisations is scant. This is even more true in the realm of natural language data. Organisations are increasingly sharing natural language text data, for example, data from voice messages, chat conversations and customer communications. The techniques for anonymising natural language data differ significantly from the many techniques used for anonymising sensitive structured data, such as k-anonymity [3], a traditional anonymisation model that has been used when sharing aggregated health records. In the field of electronic health record anonymisation, the trade-off between privacy and loss of accuracy is a well-studied problem [4,5].

As with structured data, anonymisation the identity of any text data user may significantly decrease those data's information richness. However, the problem of how to preserve information richness while conserving data privacy has not been widely studied in relation to textual natural language data. This has resulted in two practical consequences: 1) Sensitive textual data are not shared outside the organisation at all, in the fear of breaking privacy regulations; or 2) if the data are shared, they are anonymised using unnecessary obfuscating anonymisation methods, resulting in a dramatic loss in data richness.

The goal of this experiment was to study machine learning and highly anonymized conversation data in the inter-organizational setting. This study also seeks to define information richness in textual conversation data, study the relationship between data anonymisation and information richness, and, lastly, describe the machine learning experiments related to using highly anonymised conversation data. Additionally, it suggests a model for how machine learning algorithms could create useful insight from highly anonymised textual conversation data.

The study contributes to discussions related to using big data, information management and artificial intelligence to automate business processes. The paper is

organised as follows. After this introduction, Section 2 reviews the essence of anonymisation and information richness in conversation data. Section 3 describes the research method and data used in the case study. In Section 4, we then present our findings. In conclusion, Sections 5 and 6 discuss the contributions this study makes to the field.

2 Anonymisation and Information Richness

Information richness and conversation data

In-person conversations between customers and service providers use a wide range of symbolic systems and therefore contain semantically rich information. Conveying verbal and non-verbal messages in these conversations facilitates mutual understanding. When conversation takes place via the online chat, information is not semantically as rich as in face-to-face situations. However, using video as a part of a chat conversation does enable non-verbal cues to be transmitted and interpreted, allowing for a less ambiguous understanding than with simple textual information. Textual interaction does not create logical connections between various symbolic systems and cannot convey the meanings of conditional events or causes as well as multimedia or face-to-face interaction [6,7]. This means that analysing chat conversations that include only textual information results in a decrease in information richness as compared to analysing multimedia or physical conversations. In addition, the anonymisation of textual conversation data decreases semantic information.

The concept of information richness [8,9] provides a theoretical framework for discussing the richness of conversational data before and after anonymisation. Information richness refers to the data mediums, such as text, audio or voice, that deliver informational or emotional cues. The structure of text includes the logical connections and cues that form stories and meanings. Information richness, also referred to as media richness, is an objective property of media that indicates the extent to which a medium can facilitate understanding or interpretation within a specific amount of time [10,11]. Information richness does not have a causal connection to the actual performance of communication [12]. Thus, richness of information does not directly correlate to the richness of data being used in data analytics. One level of information richness may also result in different levels of understanding for a particular piece of communication [13,14]. Similarly, results may change over time [15], as information is also context-dependent.

Anonymisation and data analytics

The European Union (EU) [16] defines personal data as information concerning an identified or identifiable natural person. A person can be identified directly or indirectly by these data, which could include an "*identification number, location data, an online*

identifier or […] one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person" [16].

Data anonymisation refers to the process of obfuscating data so that they cannot be used to identify any individual. Personal identifiers can be categorised into two classes: 1) identifying attributes, such as social security numbers, names, driver license IDs, etc.; and 2) quasi-identifiers, which are a combination of key attributes that can be used to narrow down identity to a certain individual. The terms de-identification and pseydonymisation are also used in this area. De-identification is sometimes used interchangeably with anonymisation [5], while anonymisation is sometimes used in a stronger sense, denoting an irreversible form of de-identification [17]. In this article, we use the latter interpretation of anonymisation. Pseudonymisation refers to the process of personal identifiers being replaced with artificial identifiers. Pseudonymised data can be re-identified if additional information is available. Encrypting is a commonly used pseudonymisation technique. An important difference between anonymisation and pseudonymisation is that pseudonymised data still fall under the scope of privacy legislation, while anonymised data do not [18].

In medical fields, structured data are typically shared between organisations in integrated health care and public health studies. Based on consent and ethical approval, patient records can be used for secondary purposes that include clinical trials and studies, even if the data in those records are not anonymised. However, as clinical data are increasingly shared between organisations, the techniques used to anonymise those data need more attention. When different datasets are linked with each other and when intelligent algorithms are used to mine the data, even anonymised data records can contain attributes that allow for an individual's identity to be narrowed down.

There are several methods for anonymising structured data. These include directory replacement, masking, scrambling and blurring. There are also several measures designed to tackle the problem of indirect identifiers, a problem which arises when datasets are joined (in the medical field, this is called record linkage) so that a new dataset includes records that identify an individual even if all direct identifiers have been removed. For example, if information about patients' native languages is added to a set of records that have been linked, a single patient with a unique combination of nationality and native language may be identifiable in that combined dataset. The problem with establishing sound methods for removing all indirect identifiers from a single dataset is that it is not always possible to predict which sensitive datasets might be combined in the future. Changes to datasets over time also affect the indirect identifiers. Methods based on k-anonymity are a well-known solution for de-identification of combined datasets [3]. K-anonymity requires that every individual must be indistinguishable from at least *k* other individuals within that dataset, where a greater *k* value correlates with stronger de-identification levels in the data.

Joining two datasets, i.e. record linkage, often occurs when data originating from different organisations is combined for research purposes [19]. The main principle behind secure record linkage is that information identifying a patient is separated from actual health-related information. This has resulted in the creation of independent linkage centres that ensure this principle is adhered to [19].

As stated before, research on the effects of data anonymisation on the information richness of textual data is scarce. As several authors have stated, de-identifying free text is more complex that de-identifying structured data [see, for example, 20]. Cardinal

[21] presents a method for anonymising psychiatric patients' textual records which is based on fuzzy matching of recognizable phrases. This method uses cryptography to map patient identifiers to research identifiers (also called pseudonyms).

3 Methods

The case experiment

In this study, we analyse a case in which anonymised natural language chat data were shared between two organisations. These data were analysed using machine learning and neural network methodologies. The organisations represent different roles in the value chain of an ecosystem, making this case particularly relevant when considering the challenges of anonymisation and information richness. The study also provides a case experiment for using machine learning to analyse highly anonymised conversation data.

We used the Headai-artificial intelligence platform (https://www.headai.com/), which utilises machine learning methods. The platform combines semantic neurocomputing and learning algorithms to create semantic neural networks and deep insight based on unstructured or structured data. The data used in the study consist of medical information, an example of sensitive information with a high requirement for anonymisation. The data included 57,000 dialogue-loops and more than 800,000 trigger-response pairs. The data were anonymised using strict standards, removing all indications of personal information. After that, machine learning methods were applied to the data.

The case study method

Since the aim of this research is to develop a new understanding of the relationship between anonymisation and information richness, the method we adopted is the case study approach [22]. We extended our research approach to include abductive qualitative research methods, [23] since our goal is to build a new model that assists companies in managing anonymisation without losing information richness. The abductive research method enabled the researchers to build explanations from the findings and elaborate on a conceptual model that combines a literature review and the study's empirical findings. This method also enabled researchers to simultaneously process prior literature on anonymisation and information richness and the analysis of the data gathered in the experiment [22]. Using an iterative research process allowed for a deeper understanding of the experiment results while also contributing to the model of anonymisation.

The machine learning experiment procedure

This scientific research experiment grew out of the needs of a specific healthcare service provider. This organisation wanted to learn about patterns in chat conversations between patients and healthcare professionals. These chat conversations followed similar question-and-answer formats. The goal was to find patterns that could later be used to automate or improve customer experience, or to streamline business processes by improving information management practices related to the chat conversations. The conversation data were highly anonymised before being provided to data scientists.

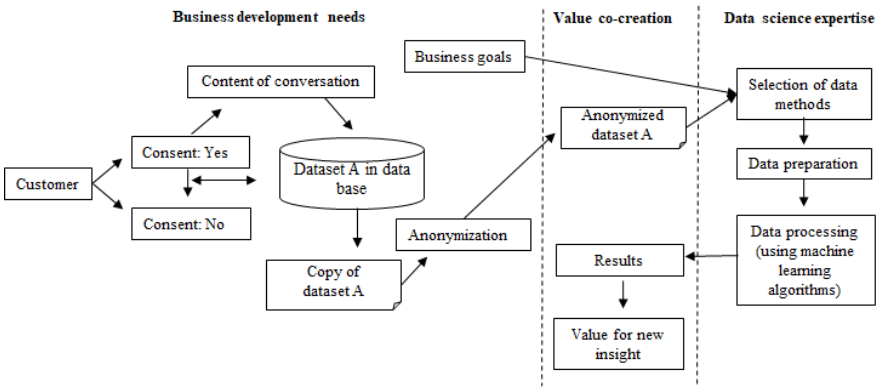


Fig. 1. Experiment procedure

The anonymised data were tested using analytic methods including Bayesian machine learning, support vector machines and feedforward networks. They were then prepared for deeper analyses that would increase the understanding of conversation dialogues. Data were used to train machine learning algorithms, but due to the high levels of anonymisation, special vocabulary and open conversation flows, several human-machine loops were needed to identify the dialogue used to create patterns.

4 Results

Our experiment showed that highly anonymised sensitive data included significantly less information than original datasets. In fact, the anonymised dataset was another dataset compared to the original dataset, in practice. The anonymised dataset excluded the job titles or roles of patients and healthcare experts, since this might have included personal information such as names, titles or organisations. Although machine learning did recognise question-answer pairs, it did not recognise which conversation partner was asking a question and which was answering and artificial intelligence could not conclude who was a patient and who a healthcare (although a human could easily make this determination based on the conversation content). The following question-answer pair illustrates the challenge that machines without the necessary algorithms have in

concluding which conversation partner represents which role, as compared to human listeners: "*My children had fevers and coughs yesterday" / "Would you like to make an appointment with a paediatrician?"*

The experiment showed that, although this conversation was a professional discussion between a patient and healthcare expert, it nevertheless included lots of conversation with "blank substance," which we also call noise. This type of content includes words such as "*hello*", "*hi*", "*how can I help you?*", "*what is your address?*" and "*video does work*". This meant that data modelling using existing machine learning methods, such as Bayesian machine learning, support vector machines and feedforward networks, did not immediately detect meaningful patterns of conversation. These machine learning methods work well in analyses of question-answer data or news data, which include sentences with straightforward meanings and textual data with a lot of different kinds of content. The preliminary analyses revealed that the kind of professional conversation in this case study includes a lot of noise. This needed to be taken into account in order for the machine learning algorithms to be able to successfully detect which repeated question-answer pairs were useful for creating new business insight as opposed to being only noise.

Unlike conversations in public discussion forums, the data in this study followed a linear path. Noise was removed from the analyses, which helped create larger entities and more meaningful question-answer pairs. In this phase, human analysis was needed to differentiate noise from substantive conversation. After this, the basic Bayesian machine learning algorithms began to work properly and were improved by the use of reinforcement learning principles. Additionally, an analysis that recognised the meanings of words also helped to find the essential topics across all conversations. These findings could help service providers incorporate additional useful information into chat conversations, for example, related articles or instructions.

5 Discussion

Our experiment showed that applying machine learning to highly anonymised data requires several human-machine loops to aid in training the artificial intelligence software that is being used. This finding supports Holzinger's [24] research, which showed that using machine learning to generate meaningful results from sensitive and complex medical information requires several rounds of expert training. In interactive machine learning, human agents (experts) interact with computational agents in order to train those computational agents to create meaningful and correct analyses. The human agent or expert also needs to perform a final check to ensure the meaningfulness of analytic results. This kind of machine learning is especially relevant in new scenarios where neither data scientists nor experts have prior experience. A new approach is needed for developing and training machine learning algorithms so that they compute in a meaningful way. For example, analysing of professional conversation data or abstract and domain-specific data require typically several human-machine interactions.

There is little existing literature on the role of anonymisation and machine learning. This study has filled that gap by showing that anonymisation causes problems for machine learning. Unlike prior research, the present study examined highly anonymised data and its information richness using real-life data. Service providers need to anonymise highly sensitive conversation data—typically natural language data (text, voice, audio, image) that includes personal information or other identifying information not located in separated columns in the database. These data differ from structured textual data that locates in database cells according to the data model of software system. Locating direct identifiers that need to be anonymised is significantly easier in structured data sets, because the identifying data are located in specific columns and cells. For example, in unstructured conversation data, it is difficult to locate all personally identifying information because names, addresses or other physical, physiological, genetic, mental, economic, cultural, social or ethnic data are a part of the conversation. In addition to recognizing direct identifiers such as proper names, the data should also preserve at least some of the content of the information, such as different actor types (patient, physician, possibly a patient's relatives, etc.). If these were recognized in the data, the anonymisation method could consist of replacing personal identifiers with general terms, for example, all patient names may read "patient", all home addresses "home address" and so on. However, if simplistic anonymisation methods are used, this creates an unintended loss of information cues and their logical connections, which are needed for a full understanding of the meanings of sentences. One advantage of unstructured conversation data is that they provide richer conversation samples, unlike data that is entered into pre-defined information categories.

The findings of this study pointed out that anonymisation resulted in some loss of significant informational cues and that this, in turn, destroyed sentence logic and decreased information richness. This is not necessarily a problem for humans, who can easily determine which question-answer pairs belong together, however, machine learning interprets words mechanically and needs informational cues in order to create patterns. For example, machine learning needs to conclude who is a patient in a conversation. If informational cues do not have a clear logical connection to meaning, then machine learning algorithms cannot determine their internal connections. An issue may be, for example, stated as suggestion, example, conclusion, diagnostic or warning in the conversation. Thus, human agents or experts are needed in several phases of the machine learning process in order to facilitate and train machine learning. This consists of an iterative process of selecting the proper methods by running several trial-and-error loops that can map concepts logically and locate meaningful patterns. This calls for collaborative and multi-disciplinary teams of data scientists and substance experts interacting with computational agents. For example, substance experts can explain to data scientists why conversation goes forward with certain logic or why some experts ask questions before they present their recommendations.

Collaborative networks are data-rich environments that have started to adopt new technologies such as artificial intelligence and data management [25, 26, 27]. These findings contribute to discussions around the ways in which data are often used across collaborative networks. Healthcare sector is one of industries that are adopting data management and machine learning within collaborative networks [28]. Healthcare and social welfare data, for example, can be shared and used in these networks, but it must

be anonymised. This study demonstrates that the anonymisation of unstructured clinical or patient textual data often results in a significant loss of information richness. Thus, the secondary use of highly anonymised data creates potential reliability and validity problems for the interpretation of results, if these issues are not taken into account when conducting machine learning analytics, especially given that artificial intelligence software cannot automatically manage these challenges. Healthcare companies can, in addition, request their customers' consent to use their data for the company's own purposes, which allows for the use of richer information. The kind of anonymisation required when sharing personal data across organisational boundaries should pay particular attention to linguistic challenges and differences. For example, some languages have several word endings that pose a challenge when trying to locate all personal identifiers in a set of unstructured textual data. Data are anonymised, and their processing falls outside the scope of GDPR when it no longer includes any identifiable personal data [18].

The proposed method for anonymising unstructured textual data uses natural language processing techniques to identify all personal identifiers and to classify them according to a domain-specific ontology. Once this has been done, all personal identifiers are replaced with terms from that ontology. The first phase of the identification of personal identifiers is achieved using an off-the-shelf named entity recognition software. Named entity recognition software uses lemmatization or the stemming of words, as well as syntactic sentence analysis, to classify entities into categories such as: names of persons, organisations, locations, expressions of time, currency and other numerical expressions (see, for example, the GATE software created by the University of Sheffield). When these named entities have been identified, a human specialist is brought in to produce training data for the domain-specific personal identifier recognizer. This involves marking in the processed text all entities listed in the domain-specific ontology, which contains a hierarchy of terms. Based on this training corpus, patterns for identifying ontology-specific personal identifiers are created. Anonymisation is achieved by replacing the specific entities recognised in the text with the corresponding term from the ontology.

In many practical cases, researchers end up facing an overly anonymised dataset. This is typically because organisations wish to stay on the safe side of privacy regulations. While this is understandable from the organisation's point of view, for a researcher, this often strips the data of most of its utility. To tackle this challenge, we propose using an interactive machine learning method with a human specialist in the loop to assist the algorithm. This method uses machine learning in an iterative manner. After the first iteration with initial training data, a human specialist inspects the results and makes one systematic improvement to the training data. The goal of this improvement is to add to the data's information richness. For example, in our case study, the first iteration consisted of a human agent classifying the parts of the conversation as either belonging to the physician or the patient. After this, the algorithm is run again and the iteration starts from the beginning. This iterative process ends either when the results of the machine learning algorithm are satisfactory with regard to the task at hand or when the human expert cannot do any systematic improvements to the training data.

6 Conclusion

This study's results contribute to debates related to information processing and management and artificial intelligence in several ways. First, by reviewing the relationship between anonymisation and information richness in unstructured conversation data and secondly, by demonstrating that an interactive machine learning method in which humans and computational agents collaborate is the best mode of analysing highly anonymised conversation data.

These findings highlight that the digital transformation of business intelligence processes is not linear, but that it instead requires multi-disciplinary teamwork in inter-organisational settings. Incorporating artificial intelligence into business processes requires an understanding the role anonymisation plays in information richness and machine learning methods. The insight generated by artificial intelligence is directly dependent on the ways in which data and human-machine interaction takes place.

This study's most basic limitation is that its reliance on a specific case study limits the results' transferability. Nonetheless, the findings provide a basic understanding of anonymisation, information richness and interactive machine learning. The present research raised questions concerning interactive machine learning which merit further examination. This study's results should also encourage researchers to conduct empirical research into how best to involve topic experts in the process of interactive machine learning, in particular when data scientists are not able to solve all domain-related conceptual and procedural problems.

In future research, the method proposed for anonymising unstructured textual data should be augmented with the capability to measure the level of anonymisation achieved with the data at hand. It should also be possible to define the desired level of anonymity beforehand. These two steps could be achieved by forming a structured database record based on each dialogue loop. The columns would consist of named entity classes and the rows would consist of the corresponding ontology-specific named entities extracted from a dialogue loop. The k-anonymity level of the anonymised data set could then be measured. If that level is lower than desired, the anonymisation method could replace the currently used ontology terms in the data with the original term. For example, the expression of location "City name: Kauniainen: role: home address" would be replaced by the "District name: Uusimaa, role: home address". In the original data, the location was the exact home address.

Acknowledgments. The authors would like to thank the Big data big business-project, all the parties behind the project, and Business Finland – for its support for this study**.**

References

1. Alamäki, A., Rantala, T., Valkokari, K., Palomäki, K..: Business Roles in Creating Value from Data in Collaborative Networks. In: Collaborative Networks of Cognitive Systems. PRO-VE 2018. IFIP Advances in Information and Communication Technology, vol 534, pp. 612—622. Springer, Cham (2018)
2. Lindquist, J.: Data Science Under GDPR with Pseudonymisation in the Data Pipeline. Dativa, https://www.dativa.com/data-science-gdpr-pseudonymisation-data-pipeline (2018)
3. Sweeney, L.: k-anonymity: A Model for Protecting Privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(05), pp. 557--570 (2002)
4. Lu, Y., Sinnott, R. O., Verspoor, K., Parampalli, U.: Privacy-Preserving Access Control in Electronic Health Record Linkage. In: 17th IEEE International Conference on Trust, Security And Privacy In Computing And Communications (TrustCom)/12th IEEE International Conference on Big Data Science And Engineering (BigDataSE), pp. 1079--1090 (2018)
5. Board on Health Sciences Policy, Institute of Medicine, https://www.ncbi.nlm.nih.gov/books/NBK285994/ (2015)
6. Lim, K. H., Benbasat, I.: The Effect of Multimedia on Perceived Equivocality and Perceived Usefulness of Information Systems. MIS Quarterly, 449--471 (2000)
7. Salomon, G.: Interaction of Media, Cognition, and Learning: An Exploration of How Symbolic Forms Cultivate Mental Skills and Affect Knowledge Acquisition. Jossey-Bass, San Francisco (1979)
8. Daft, R. L., Lengel, R. H.: Information Richness: A New Approach to Managerial Behavior and Organisation Design (No. TR-ONR-DG-02). College of Business Administration, Texas A&M University, College Station, TX (1983)
9. Daft, R. L., & Lengel, R. H.: Organisational Information Requirements, Media Richness and Structural Design. Management Science 32(5), 554--571 (1986)
10. Sun, P. C., Cheng, H. K.: The Design of Instructional Multimedia in E-Learning: A Media Richness Theory-based Approach. Computers & Education 49(3), 662--676 (2007)
11. Alamäki, A., Pesonen, J., Dirin, A.: Triggering Effects of Mobile Video Marketing in Nature Tourism: Media Richness Perspective. Information Processing & Management 56 (3), 756--770 (2019)
12. Dennis, A. R., Kinney, S. T.: Testing Media Richness Theory in the New Media: The Effects of Cues, Feedback, and Task Equivocality. Information Systems Research 9.3, 256--274 (1998)
13. Mayer, R. E.: Multimedia Learning (2nd ed.). Cambridge University Press, New York, NY (2009)
14. Fiorella, L., Mayer, R. E.: Effects of Observing the Instructor Draw Diagrams on Learning from Multimedia Messages. Journal of Educational Psychology 108(4), 528 (2016)
15. Tan, W. K., Tan, C. H., Teo, H. H.: Conveying Information Effectively in a Virtual World: Insights from Synthesized Task Closure and Media Richness. Journal of the American Society for Information Science and Technology 63(6), 1198--1212 (2012)

16. GDPR Regulation (EU) 2016/679 of the European parliament and of the council of 27 April 2016. Official Journal of the European Union, https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN (2016)
17. Garfinkel, S.: De-identification of Personal Information (NISTIR 8053). U.S. National Institute of Standards and Technology, http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf (2015)
18. European Medicines Agency Data Anonymisation: A Key Enabler for Clinical Data Sharing, Workshop report, 30 November-1 December, 2017, EMA/796532/2018 (2018)
19. Winkler, W. E.: Overview of Record Linkage and Current Research Directions. Technical Report Statistical Research Report Series RRS2006/02, US Bureau of the Census, Washington, D.C. (2006)
20. Uzuner, O., Luo, Y., Szolovits, P.: Evaluating the State-of-the-art in Automatic Deidentification. J Am Med Inform Assoc (JAMIA) 14:550--63 (2007)
21. Cardinal, R. N.: Clinical records anonymisation and text extraction (CRATE): an open-source software system. BMC medical informatics and decision making, 17(1), 50 (2017).
22. Eisenhardt, K. M., Graebner, M.: Theory Building from Cases: Opportunities and Challenges. Academy of Management Journal 50(1), 25--32 (2007)
23. Dubois, A., Gadde, L. E.: Systematic Combining: An Abductive Approach to Case Research. Journal of Business Research 55(7), 553--560 (2002)
24. Holzinger, A.: Interactive Machine Learning for Health Informatics: When Do We Need the Human-in-the-loop? Brain Informatics 3(2), 119--131 (2016)
25. Camarinha-Matos L.M., Fornasiero R., Afsarmanesh H.: Collaborative Networks as a Core Enabler of Industry 4.0. In: Collaboration in a Data-Rich World. PRO-VE 2017. IFIP Advances in Information and Communication Technology, vol 506, 3—17. Springer (2017)
26. Serrano, D. C., Chavarría-Barrientos, D., Ortega, A., Falcón, B., Mitre, L., Correa, R., & Gutiérrez, A. M. A Framework to Support Industry 4.0: Chemical Company Case Study. In: Collaborative Networks of Cognitive Systems. PRO-VE 2018. IFIP Advances in Information and Communication Technology, vol 534, pp. 387--395. Springer, Cham (2018)
27. Valkokari K., Rantala T., Alamäki A., Palomäki K.: Business Impacts of Technology Disruption - A Design Science Approach to Cognitive Systems' Adoption Within Collaborative Networks. In: Collaborative Networks of Cognitive Systems. PRO-VE 2018. IFIP Advances in Information and Communication Technology, vol 534, pp. 337—348. Springer, Cham (2018)
28. Macedo, P., Madeira, R. N., Camarinha-Matos, L. M.: Cognitive Services for Collaborative mHealth: The OnParkinson Case Study. In: Collaborative Networks of Cognitive Systems. PRO-VE 2018. IFIP Advances in Information and Communication Technology, vol 534, pp. 442-453. Springer, Cham (2018)