1 Reviewer 1

1. Revise the text to improve its readability. Correct grammatical and typing errors. Maybe proof-read by native speaker? Maybe revise structure? *We have revised and improved a substantive part of the text, as can be seen in the track-changes document. In addition, the revised version of the manuscript has been proof-read by a native speaker. At last, we have revised the structure in such a way we focus more on related work, the structure of DDPs in general and how Instagram DDPs fit within that profile.*

2. Moreover, I think that the sections "Evaluation" and "Results" are in need of further work: some results are presented in the "evaluation" section, and I found the "result" to be lacking of some quantitative results and statistics. *We agree that these sections have become messy due to the fact that after the first round of results, we further improved the algorithm and we present new results. We believe that it is transparent to show that we took this step before obtaining the final results, but we understand that it might lead to some confusion for the reader. Therefore, we have now rewritten this into a single section where subheadings make the different steps that we take more clear. In addition, we included a separate section where we describe the dataset and also provide the descriptive statistics of that dataset in that section. Furthermore, the evaluation procedure is now described in a separate subsection under 'methods'. Overall, this new structure better corresponds to the procedure that we used to develop the script and makes the steps that we took in this procedure more striaghtforward to understand. In addition, we have supplemented many of our conclusions with new tables and provide new descriptives regarding the evaluation set. We refer to the track-change document to see the adjusted structure and new tables.*

2 Reviewer 2

1. (automatically) redacting documents/data is not a new problem. No related work on this is provided. In a slightly different context, solutions for detecting/removing sensitive info about to be sent on a network were proposed in the past. *We agree with the reviewer that there are many related studies on automatic de-identification of documents and media. We have expanded our related work section where we now include these studies. In this section we also give a more elaborate explanation on why the automatic de-identification of DDPs is a challenge and why existing approaches may not be applicable.*

2.1 Related work

~~To remove direct personal data from DDPs, the software should be able to adhere to the five key characteristics of DDPs introduced in the previous subsection. A first step is to investigate to what extent existing software and literature is able to remove direct personal data from DDPs. A well-known approach is $k-$ anonymity (1) which requires that each record in a data-set is similar to at least $k-1$ other records on the potentially identifying variables (2). However, parts of the DDPs are highly unstructured and thereby unique per DDP and reaching $k-$ anonymity is therefore not feasible. Much research has focused on the de-identification of electronic health records, for example to enable their use in multi-center research studies (3). Scientific open source de-identification tools are available such as DEDUCE (4) as well as commercial tools, such as Amazon Comprehend (5) and CliniDeID (6) (7). Similar initiatives have taken place to de-identify personal data in other types of data, such as for human resource purposes (8). However, textual content generated from structured data-bases such as for electronic health records or human resources typically have a higher level of structure compared to DDPs and does not handle key identifying information in DDPs, such as usernames or visual content and therefore existing software was not sufficient for our purpose. Alternatively, software has been developed focusing on the removal of usernames, for example for Twitter data (9). Furthermore, many different types of both open source and commercial software are available to identify and blur faces on images and videos, such Microsoft Azure (10), and Facenet-PyTorch (11). However, none of the investigated software was able to handle both textual and visual content and both structured and unstructured data within one procedure.~~

~~To summarize, a de-identification procedure is required that works appropriately when file structures change rapidly over time, while there are substantive differences in the level of structure within the~~

~~files, that is able to handle different file formats, that is able to handle both visual and textual content and that recognizes the username as the primary identifier for natural persons, while other types of person identifying information (PII) should also be accounted for, such as first names, phone numbers and e-mail addresses. The developed software aims for such a level of protection that the privacy of the DDP owners (the participants) is always preserved. Importantly, the goal is not to prepare the DDPs for public sharing, however, in the unlikely event of a data breach, the individual research participants should not be directly identifiable. Therefore, the de-identification procedure introduced here should always be supplemented with other security measures such as using a shielded (cloud)environment to store the data and using privacy-preserving algorithms when analyzing the data.~~

De-identification of data in the medical domain has extensively been researched. Medical patient data, like electronic health records and clinical notes, are increasingly used for clinical research. As imposed by privacy legislations such as the US Health Insurance Portability and Accountability Act (HIPAA) (12) and the GDPR, the privacy of patients includede in these data has to be protected. Medical data are therefore de-identified by removing all categories of protected health information (PHI) that are defined by the HIPAA. PHI types typically found in medical data are person names and initials, names of institutions, social security numbers and dates (3; 4; 13; 14). Automatic de-identification approaches in the literature are either rule-based, machine learning based or a combination of both, where machine-learning approaches show the best performance (3; 13; 14). Scientific open-source de-identification tools are available such as DEDUCE (4) and Amnesia (15) as well as commercial tools, such as Amazon Comprehend (5) and CliniDeID (6) (7). Most automatic de-identification approaches are constrained to English medical documents and little is known about their generalizability across languages or domains. Although neural networks have shown good generalization performance compared to rule-based and feature based approaches, a substantial decrease of performance has to be expected when applying these out of the box to new languages or domains (13).

User privacy in social media is an emerging research area and has attracted increasing attention recently. To avoid privacy attacks, like identity disclosure and attribute disclosure, publishers of social media data are obliged to protect users' privacy by anonimizing these data before they are published publicly (16). Anonymizing social media data is a challenging task due to their heterogeneous, highly unstructured and noisy nature (16). Commonly used statistical disclosure control approaches (17; 1; 2; 15; 18) are designed for relational and tabular data and cannot be directly applied to social media data. In addition, PHI types that are common in medical data are unlikely to be found in textual social media data. These data rather contain person names, usernames or IDs, email addresses and locations (19; 20), but in fact there is limited work on the types of person identifying information (PII) that may be present in textual social media data and how these should be removed (20; 21). Yet, removing such information has been shown to be far from sufficient in preserving privacy since users' identity or attributes may be inferred from the public data available on social media platforms (22; 23; 24; 16). Finally, social media data may also consist of visual content. Many different types of both open source and commercial software are available to identify and blur faces on images and videos, such as Microsoft Azure (10), and Facenet-PyTorch (11). However, modern image recognition methods based on deep learning have demonstrated that hidden information in blurred images can be recovered (25).

Like social media data, DDPs are heterogeneous and unstructured and are likely to contain the same types of sensitive information. Yet, the limited de-identification approaches that are available for social media data focus either on textual or visual content and the presence of both types of information within one DDP poses a major de-identification challenge (26). An important difference is that on social media platforms information on large groups of users is widely available, whereas DDPs are only available for a single individual. The goal of this research is not to prepare the DDPs for public sharing. DDPs will either be stored on the owner's device or in a shielded (cloud)environment and analyzed using privacy-preserving algorithms. In that sense, handling DDP's is comparable to handling medical data and we therefore assume that the risk of privacy attacks is very low. However, for ethical reasons and in the unlikely event of a data breach, DDPs should still be de-identified.

To summarize, we need a de-identification procedure that is able to handle unstructured and heterogeneous data, and can de-identify both visual and textual content within one procedure. It should be able to recognize usernames as the primary identifier for natural persons, while other types of PII, such as person names, phone numbers and e-mail addresses, should also be accounted for.

2. the utility aspects are not covered. Are the resulting DDP still useful. To assess this, the authors should explain what data scientist typically need from DDP and why they collect them in the first place. Also, the general pipeline for collecting DDP is not clear (users do it and then pass it to the researchers I guess; it should

be explained that the de-identification would be done on the user side). Also, it is not clear how the utility is preserved for studies that focus on the *interaction* between users. It should be made clearer how interaction/relation information is preserved. *We already discussed some of these issues in the discussion section, but in addition we now explain the workflow and the usability of the resulting de-identified DDPs in the introduction as well.*

> However, the data present in DDPs can be deeply private and potentially sensitive. This poses a major challenge to using DDPs for scientific research. Participants might not be willing to share this sensitive data. However, researchers are often only interested in a part of the DDP and do not need the sensitive data. Although an interesting solution is to extract relevant features locally on the device of the participant (27), this workflow is not suitable for all research purposes. When, for example, an exploratory approach is of interest, or when the aim is to develop or improve the performance of an extraction algorithm, local extraction would limit the analytic possibilities. In such situations, collection of the *complete* DDP is desired, which requires challenges caused by the sensitivity of the data to be overcome. An example of such a research project is Project AWeSome (28), which collects complete Instagram DDPs from research participants. The participants' DDPs are stored in a secured environment where they are de-identified using the de-identification algorithm proposed in this manuscript. Only after the sensitive information is adequately masked, can the DDPs be shared with the applied researchers for substantive analyses.

3. the privacy goals are not clear. whose privacy should be protected? that of the DDP owner or that of the mentioned individuals? If it's the owner, I'm afraid a simple Google (image) search would still re-identify the original data (if it's public) from the redacted data. For instance, a google image search fed with the picture with blurred face would probably return the original image. *The goal is to preserve the privacy of participants in a way that is common for sensitive research data. Decisions here are aimed at maintaining a balance between the reduction of identifying information on the one hand and usability of the data for research purposes on the other hand. Our goal is therefore to strive for de-identification at such a level that when the data is used in a secured environment and without linkage to other data-sources, identification of individual research participants is not possible. In addition, the same holds for the identification of other individuals who's images or messages appear in the DDP of the research participant as a by-product. Originally, we discussed this in the related work section but we have removed this part and make the privacy goals more explicit in the introduction of the manuscript.*

We argue that in situations where complete DDPs are collected for research, the DDPs should be treated in a similar fashion as any other sensitive data that is collected for research purposes. We therefore follow the guidelines for sensitive research data[1], which were established by Utrecht University for handling sensitive data from official statistical agencies and governmental bodies like Statistics Netherlands (29) and the European Commission(30). From these guidelines it can be concluded that two important measures should be taken. First, security measures such as using shielded (cloud) environments for data storage should be used. Second, the privacy of participants should be preserved while their data is analysed by researchers.

4. Sections 2.1 and 2.2 lack references to (official) documentation. *We have now restructured these two sections into one separate section on DDPs. In this section, we now provide a table that illustrates how the Instagram DDP relates to DDPs from other data controllers. In this table, we also provide the links to official documentation from these data controllers.*

3 ~~Background~~Data download packages

~~The aim of the software introduced in this paper is to enable researchers to use DDPs for scientific research while preserving the privacy of participants. In this section, we explain in more detail the specific type of data that can be found in DDPs, define our aims in terms of data protection in more detail and discuss relevant existing literature and software.~~

3.1 ~~Data Download Packages~~

~~Most large data controllers currently comply with the right of data access by providing users with the option to retrieve an electronic "Data Download Package" (DDP). This DDP typically comes as a .zip-file containing .json, .html, .csv, .txt, .JPEG and/or .MP4 files in which all the digital traces left behind by the data subject with respect to the data controller are stored. The structure and content of a DDP varies per data controller, and even within data controllers there are differences among data subjects. Data subjects may use different features provided by the data controller and this is reflected by their DDP, for example, if a data subject does not share photos on Facebook, there will be no data folder with .JPEG files in the corresponding DDP.~~

[1] https://www.uu.nl/en/research/research-data-management/faq

~~One particular characteristic of DDPs is that their content and structure is often subject to change. For example, if a data subject downloads the DDP at a data controller, and repeats this a month later, differences may be found in the structure of the DDP. This can have several causes. The most straightforward cause is that the data subject generated additional data throughout this month. However, other important factors also play a role. First, data controllers can develop new features by which new types of data regarding the data subject are collected. Second, other features are phased out. Third, some data (for example search history) is only saved for a limited amount of time and is destroyed by the data controller after that period. In that case, it will also not be present in the DDP anymore. At last, the GDPR is still relatively new and data controllers continue to optimize the processes used to transfer the relevant data to its subjects, leading to changes in the structure of DDPs.~~

3.2 ~~Instagram DDPs~~

~~As the software in this research project was initially developed to de-identify Instagram DDPs, the structure of these DDPs has been thoroughly investigated. Instagram DDPs come as one or multiple zipfiles (depending on the amount of data available on the data subject). The .zip-file contains a number of folders in which all the visual content is stored, namely "photos", "videos", "profile" and "stories". The different folders refer to the different Instagram features used by the data subject to generate the visual content. For example, in the folder "profile", a subject's profile picture can be found, while in the folder "stories", visual content can be found generated using the "stories" feature in Instagram, a form of ephemeral sharing. All textual information is collected in a number of .json files. Some of these files have a simple list structure. For example the file "likes.json" lists all the 'likes' given by the subject, supplemented with a timestamp and the username of the Instagram account to which the 'like' was given. Files such as 'connections.json', 'searches.json' and 'seen_content.json' have similar structures. Other files, such as 'profile.json' are typically shorter in size but have a more complex structure, as they typically contain different auxiliary characteristics. Other files with such a structure are for example 'account_history.json', 'devices.json' and 'settings.json'. However, a substantial number of files contains data that is less structured. Examples of such files are 'comments.json', 'media.json', 'messages.json' and 'stories_activities.json'. Furthermore, data subjects at Instagram are not necessarily natural persons. Data subjects at Instagram can be identified by a single and unique Username. Typically, natural persons have individual~~

~~accounts with an accompanying username, but other institutions, such as for example retail shops or bands can also have an individual account with an accompanying username.~~

Most large data controllers currently comply with the right of data access by providing users with the option to retrieve an electronic DDP. This DDP typically comes as a compressed folder containing text and/or media files in which all the digital traces left behind by the data subject with respect to the data controller are stored. Table 4 shows that the content and structure of DDPs differs among data controllers. Differences between DDPs from the same data controller can also occur among data subjects and over time. These differences may be caused by data subjects using different features provided by the data controller or by the fact that the DDP is a snapshot of the data collected by the data subject up to that point. However, other important factors also play a role. First, data controllers can develop new features through which new types of data of the data subject are collected. Second, other features may be phased out. Third, some data (for example search history) is only saved for a limited amount of time and is destroyed by the data controller after that period. In that case, it will also not be present in the DDP anymore. Finally, the GDPR is still relatively new and data controllers continue to optimize the processes used to transfer the relevant data to its subjects, leading to changes in the structure of DDPs.

From Table 4 it can be concluded that the Instagram DDP contains many features that can also be found in DDPs of other data controllers. Common features are the presence of both text and/or media files, the presence of both structured and unstructured text and the presence of specific types of person identifying information (PII). Therefore, an algorithm that is able to de-identify Instagram DDPs also contains the features needed to de-identify many of the DDPs of other data controllers. To summarize, the developed algorithm is able to handle: ~~To summarize, software to de-identify Instagram DDPs should be able to handle:~~

- An ever changing file structure,
- both visual and textual content,
- different file formats,
- ~~Files in highly structured and highly unstructured format and different variants in between~~files ranging from highly structured to highly unstructured formats,
- ~~Natural persons and other users which are identified by their unique username.~~the masking of usernames of natural persons or other users.

		FACEBOOK[a]	**WHATSAPP**[b]	**TWITTER**[c]	**SNAPCHAT**[d]	**INSTAGRAM**[e]
DDP INFO	DDP name	facebook-¡profile_name¿	My account information.zip WhatsApp chat-¡group or contactname¿.zip	Archive	mydata¡hashed_code¿	username_¡date_of_download¿
	DDP format	.zip	.zip	.zip	.zip	.zip
	Type of files	media, text	media, text	media, text	text	media, text
	Structure	Content folders ¿ content files	Separate DDP per conversation	Content folders ¡ content files	Index file & Format (i.e., json and html) folders ¿ content files	Content text files and Content folders ¿ content media files
MEDIA FILES	Format of media files	.PNG, .JPG, .MP4	.JPG, .MP4, .HTML	.PNG	-	.JPG, .MP4
	Folder structure	All images, videos, stickers are categorized and stored in corresponding folders. There are no loose files.	All images, videos, stickers in single folder	-	All images and videos are categorized and stored in corresponding folders. There are no loose files.	
PII IN MEDIA	Faces	-/+	-/+	-/+	-	-/+
	Written text	-/+	-/+	-/+	-	-/+
	(user)name tags	-	-	-	-	-/+
TEXT FILES	Format of text files	.JSON or .HTML	.TXT, .OPUS, .HTML	.JS or .HTML	.JSON and .HTML	.JSON
	Folder structure	All text files are categorized and stored in corresponding folders. There are no loose files.	All text in single file per conversation	There is one text file per month.	Text files are not categorized and stored in (sub) folders. They are displayed as loose files.	Text files are not categorized and stored in (sub) folders. They are displayed as loose files.
	Structured data	+	+	+	+	+
	Unstructured data (i.e., containing free text)	+	+	+	-/+	+
PII IN TEXT	Usernames	-/+	-/+	+	+	+
	(first) Names	+	+	-/+	-/+	+
	Email addresses	+	+	+	-/+	+
	Phone numbers	-/+	+	-/+	-/+	-/+
	Locations	-/+	-/+	-/+	-/+	-/+

Table 1: Overview of content and structure of DDPs of five data controllers. Note that if a certain object is present in DDPs, this is indicated with +. If it often occurs within the DDP, a -/+ is used. Finally, if said object is not present, a - is used.

[a] https://www.facebook.com/help/1701730696756992

[b] https://faq.whatsapp.com/general/account-and-profile/how-to-request-your-account-information/

[c] https://help.twitter.com/nl/managing-your-account/how-to-download-your-twitter-archive

[d] https://support.snapchat.com/en-US/a/download-my-data

[e] https://help.instagram.com/181231772500920?helpref

	Information	Instagram DDP
Overall	Main language	Dutch; English
Text	Structure	Unstructured; Loose text files
	Number of files	20
	File names	account_history; autofill; comments; connections; devices; events; fundraisers; guides; information_about_you; likes; media; messages; profile; saved; searches; seen_content; settings; shopping; stories_activities; uploaded_contacts;
	File format	.JSON
Media	Structure	Structured: Folder ¿ subfolder ¿ media files
	Folders	photos; profile; stories; videos
	Subfolders	Date (format: YYYYMM)
	File format	.JPG/.MP4

Table 2: The content of a typical Instagram DDP of a Dutch user

5. The dataset is relatively small. Also, except for some data, the annotation was done by a single annotator; this is not very robust. *We agree that the dataset is relatively small in terms of number of DDPs. However, as the data were generated for research purposes and during the generation process, we accounted for the variety of Instagram features. We explain this procedure and highlight the variety that remained now in more detail in the Data section. However, we also comment on this issue in the discussion section.*

4 Data

4.1 Development set

For the development of this new de-dentification procedure, the researchers initially used two DDPs of their own personal Instagram accounts. The functionality of the algorithm was based on the typical Instagram DDP file structure (see Table 2). To ensure that the developed algorithm can adequately handle possible varieties in DDP structures (over different Instagram accounts), a validation data corpus was created. Using this corpus, the de-identification procedure could be tested and improved, maximizing its effectiveness.

4.2 Validation corpus sampling

A group of 11 participants generated Instagram DDPs by actively using a new Instagram account for approximately a week. The participants were instructed not to share any of their own personal information via the Instagram accounts. Instead, participants were instructed to share either fake or publicly available information by, for example, sharing URLs of news websites, posting images of celebrities, or liking and following verified Instagram accounts. As the final data corpus does not contain any personal information it is publicly available at `http://doi.org/10.5281/zenodo.4472606`.

6. Not sure Section 4.4 is needed for the audience of a journal on Data Science. *We have made this section shorter, but we think some information should remain in place, to make clear what evaluation criteria we focus on and the argumentation for this.*

4.3 Evaluation criteria

~~For each category of PII in each filetype in the set of DDPs regarding textual content, we count the number of TP, FP and FN. For the visual content, we calculate the TP and FN.~~ We use scikit learn to further evaluate the performance of the procedure on the different aspects (31). First, we calculate the recall, or the sensitivity, as

$$Recall = \frac{TP}{TP + FN}. \tag{1}$$

Here, we measure the ratio of the correctly de-identified cases to all the cases that were supposed to be de-identified (i.e. ground truth). Each false negative potentially results in not preserving the privacy of a research participant and therefore a high value for the recall is particularly important. The precision is calculated as

$$Precision = \frac{TP}{TP + FP}. \tag{2}$$

Precision shows the ratio of correctly de-identified observations to the total of de-identified observations and a high precision illustrates that the amount of additional information lost due to unnecessary de-identification is limited. Given that DDPs are typically collected to analyze aspects such as the free text or the images, losing a lot of this information by the de-identification process challenges the intended research goal. At last, we calculate the F1 score

$$F1 - score = 2 \times \frac{precision \times recall}{precision + recall}, \tag{3}$$

which combined the precision and recall and considered both false positives and false negatives. Note that we do not calculate the accuracy as the number of true negatives cannot be determined appropriately in our data-set.

	PII	**File**	**N**	**Count**	**Proportion**
Textual					
	Username	comments.json	10	261	0.03
		connections.json	10	1222	0.14
		likes.json	10	883	0.10
		media.json	10	43	0.00
		messages.json	10	2947	0.33
		profile.json	10	10	0.00
		saved.json	11	6	0.00
		searches.json	11	314	0.04
		seen_content.json	11	3144	0.35
		shopping.json	11	1	0.00
		stories_activities.json	11	35	0.00
		Total	**115**	**8866**	**1.00**
	Name	comments.json	10	105	0.18
		media.json	10	54	0.09
		messages.json	10	427	0.72
		profile.json	10	10	0.02
		Total	**40**	**596**	**1.00**
	Email	comments.json	10	28	0.13
		media.json	10	28	0.13
		messages.json	10	152	0.70
		profile.json	10	10	0.05
		Total	**40**	**218**	**1.00**
	Phone	comments.json	10	29	0.16
		media.json	10	9	0.05
		messages.json	10	140	0.79
		Total	**30**	**178**	**1.00**
	URL	comments.json	10	1	0.00
		messages.json	10	267	0.96
		profile.json	10	10	0.04
		Total	**30**	**278**	**1.00**
Visual					
	PII	**Folder**	**.JPG**	**.MP4**	**Proportion**
	Username	photos	49	-	0.11
		stories	255	105	0.84
		videos	-	21	0.05
		Total	**304**	**126**	**1.00**
	Face	direct	20	-	0.01
		photos	1046	-	0.67
		stories	290	163	0.29
		videos	-	36	0.02
		Total	**1356**	**199**	**1.00**

Table 3: Descriptive statistics of visual and textual content in the generated Instagram DDP validation corpus

5 Reviewer 3

1. On the downside, it is unclear how the submitted manuscript relates to data science and if the proposed de-identification method relates to the journal's aims and scope. *According to our understanding, one of the aims and scope of the journal includes the processing of data, and the journal has an interest in specific tools. One of the goals of the journal is to unleash the power of scientific data to deepen our understanding of digital systems and to gain insight into human social and economic behavior. Our de-identification algorithm is a tool that processes DDPs in such a way that it preserves the privacy of participants. By using our algorithm, the privacy of research participants can be preserved in a similar way that is done with any type of research data, and thereby we open up the possibility to use DDPs for scientific research, which will help gain insight into human social and economic behavior, as all digital traces left behind by participants on certain platforms can be collected via DDPs and then analysed.*

2. Second, the authors missed several highly related work/software that propose anonymization tools for research data: `https://arx.deidentifier.org`, `https://amnesia.openaire.eu` **and** `https://cran.r-project.org/web/packages/sdcMicro/index.html`. **I encourage the authors to cite them and position their work with respect to them.** *We agree with the reviewer that there are many more related studies on automatic de-identification of documents and media. We have included the references mentioned by the reviewer and additional studies in the related work section. In this section we also give a more elaborate explanation on why the automatic de-identification of DDPs is a challenge and why existing approaches may not be applicable.*

5.1 Related work

~~To remove direct personal data from DDPs, the software should be able to adhere to the five key characteristics of DDPs introduced in the previous subsection. A first step is to investigate to what extent existing software and literature is able to remove direct personal data from DDPs. A well-known approach is $k-$ anonymity (1) which requires that each record in a data-set is similar to at least $k-1$ other records on the potentially identifying variables (2). However, parts of the DDPs are highly unstructured and thereby unique per DDP and reaching $k-$ anonymity is therefore not feasible. Much research has focused on the de-identification of electronic health records, for example to enable their use in multi-center research studies (3). Scientific open source de-identification tools are available such as DEDUCE (4) as well as commercial tools, such as Amazon Comprehend (5) and CliniDeID (6) (7). Similar initiatives have~~

~~taken place to de-identify personal data in other types of data, such as for human resource purposes (8). However, textual content generated from structured data-bases such as for electronic health records or human resources typically have a higher level of structure compared to DDPs and does not handle key identifying information in DDPs, such as usernames or visual content and therefore existing software was not sufficient for our purpose. Alternatively, software has been developed focusing on the removal of usernames, for example for Twitter data (9). Furthermore, many different types of both open source and commercial software are available to identify and blur faces on images and videos, such Microsoft Azure (10), and Facenet-PyTorch (11). However, none of the investigated software was able to handle both textual and visual content and both structured and unstructured data within one procedure.~~

~~To summarize, a de-identification procedure is required that works appropriately when file structures change rapidly over time, while there are substantive differences in the level of structure within the files, that is able to handle different file formats, that is able to handle both visual and textual content and that recognizes the username as the primary identifier for natural persons, while other types of person identifying information (PII) should also be accounted for, such as first names, phone numbers and e-mail addresses. The developed software aims for such a level of protection that the privacy of the DDP owners (the participants) is always preserved. Importantly, the goal is not to prepare the DDPs for public sharing, however, in the unlikely event of a data breach, the individual research participants should not be directly identifiable. Therefore, the de-identification procedure introduced here should always be supplemented with other security measures such as using a shielded (cloud)environment to store the data and using privacy-preserving algorithms when analyzing the data.~~

De-identification of data in the medical domain has extensively been researched. Medical patient data, like electronic health records and clinical notes, are increasingly used for clinical research. As imposed by privacy legislations such as the US Health Insurance Portability and Accountability Act (HIPAA) (12) and the GDPR, the privacy of patients includede in these data has to be protected. Medical data are therefore de-identified by removing all categories of protected health information (PHI) that are defined by the HIPAA. PHI types typically found in medical data are person names and initials, names of institutions, social security numbers and dates (3; 4; 13; 14). Automatic de-identification approaches in the literature are either rule-based, machine learning based or a combination of both, where machine-learning approaches show the best performance (3; 13; 14). Scientific open-source de-identification tools are available such as DEDUCE (4) and Amnesia (15) as well as commercial tools, such as Amazon Comprehend (5) and CliniDeID (6) (7). Most automatic de-identification approaches are constrained to English medical documents and little is known about their generalizability across languages or domains. Although neural networks have shown good generalization performance compared to rule-based and feature based approaches, a substantial decrease of performance has to be expected when applying these out of the box to new languages or domains (13).

User privacy in social media is an emerging research area and has attracted increasing attention recently. To avoid privacy attacks, like identity disclosure and attribute disclosure, publishers of social media data are obliged to protect users' privacy by anonimizing these data before they are published publicly (16). Anonymizing social media data is a challenging task due to their heterogeneous, highly unstructured and noisy nature (16). Commonly used statistical disclosure control approaches (17; 1; 2; 15; 18) are designed for relational and tabular data and cannot be directly applied to social media data. In addition, PHI types that are common in medical data are unlikely to be found in textual social media data. These data rather contain person names, usernames or IDs, email addresses and locations (19; 20), but in fact there is limited work on the types of person identifying information (PII) that may be present in textual social media data and how these should be removed (20; 21). Yet, removing such information has been shown to be far from sufficient in preserving privacy since users' identity or attributes may be inferred from the public data available on social media platforms (22; 23; 24; 16). Finally, social media data may also consist of visual content. Many different types of both open source and commercial software are available to identify and blur faces on images and videos, such as Microsoft Azure (10), and Facenet-PyTorch (11). However, modern image recognition methods based on deep learning have demonstrated that hidden information in blurred images can be recovered (25).

Like social media data, DDPs are heterogeneous and unstructured and are likely to contain the same types of sensitive information. Yet, the limited de-identification approaches that are available for social media data focus either on textual or visual content and the presence of both types of information within one DDP poses a major de-identification challenge (26). An important difference is that on social media platforms information on large groups of users is widely available, whereas DDPs are only available for a single individual. The goal of this research is not to prepare the DDPs for public sharing. DDPs will either be stored on the owner's device or in a shielded (cloud)environment and analyzed using privacy-preserving algorithms. In that sense, handling DDP's is comparable to handling medical data and we therefore assume that the risk of privacy attacks is very low. However, for ethical reasons and in the unlikely event of a data breach, DDPs should still be de-identified.

To summarize, we need a de-identification procedure that is able to handle unstructured and heterogeneous data, and can de-identify both visual and textual content within one procedure. It should be able to recognize usernames as the primary identifier for natural persons, while other types of PII, such as person names, phone numbers and e-mail addresses, should also be accounted for.

3. Third, as mentioned already, the tool, in particular the improved script, is tailored to a specific application, which limits its scope and impact. It would be good to discuss how the proposed improvements could also apply in other contexts, such as other social networks or applications. *We now discuss in more detail why we are convinced that the Instagram DDP is very representative to many other DDPs, as they contain a wide scope of typical characteristics of a very diverse group of DDPs. Therefore, many of the features included in our algorithm can be used to de-identify DDPs of other platforms as well. We have restructured this section in such a way that we now provide a table that illustrates how features from DDPs of different platforms are also present in the DDP of Instagram. In addition, we provide links to where and how the DDPs of the investigated data controllers can be downloaded and make more clear which information is based on the inspection of DDPs.*

6 ~~Background~~Data download packages

~~The aim of the software introduced in this paper is to enable researchers to use DDPs for scientific research while preserving the privacy of participants. In this section, we explain in more detail the specific type of data that can be found in DDPs, define our aims in terms of data protection in more detail and discuss relevant existing literature and software.~~

6.1 ~~Data Download Packages~~

~~Most large data controllers currently comply with the right of data access by providing users with the option to retrieve an electronic "Data Download Package" (DDP). This DDP typically comes as a .zip-file containing .json, .html, .csv, .txt, .JPEG and/or .MP4 files in which all the digital traces left behind by the data subject with respect to the data controller are stored. The structure and content~~

~~of a DDP varies per data controller, and even within data controllers there are differences among data subjects. Data subjects may use different features provided by the data controller and this is reflected by their DDP, for example, if a data subject does not share photos on Facebook, there will be no data folder with .JPEG files in the corresponding DDP.~~

~~One particular characteristic of DDPs is that their content and structure is often subject to change. For example, if a data subject downloads the DDP at a data controller, and repeats this a month later, differences may be found in the structure of the DDP. This can have several causes. The most straightforward cause is that the data subject generated additional data throughout this month. However, other important factors also play a role. First, data controllers can develop new features by which new types of data regarding the data subject are collected. Second, other features are phased out. Third, some data (for example search history) is only saved for a limited amount of time and is destroyed by the data controller after that period. In that case, it will also not be present in the DDP anymore. At last, the GDPR is still relatively new and data controllers continue to optimize the processes used to transfer the relevant data to its subjects, leading to changes in the structure of DDPs.~~

6.2 ~~Instagram DDPs~~

~~As the software in this research project was initially developed to de-identify Instagram DDPs, the structure of these DDPs has been thoroughly investigated. Instagram DDPs come as one or multiple zipfiles (depending on the amount of data available on the data subject). The .zip-file contains a number of folders in which all the visual content is stored, namely "photos", "videos", "profile" and "stories". The different folders refer to the different Instagram features used by the data subject to generate the visual content. For example, in the folder "profile", a subject's profile picture can be found, while in the folder "stories", visual content can be found generated using the "stories" feature in Instagram, a form of ephemeral sharing. All textual information is collected in a number of .json files. Some of these files have a simple list structure. For example the file "likes.json" lists all the 'likes' given by the subject, supplemented with a timestamp and the username of the Instagram account to which the 'like' was given. Files such as 'connections.json', 'searches.json' and 'seen_content.json' have similar structures. Other files, such as 'profile.json' are typically shorter in size but have a more complex structure, as they typically contain different auxiliary characteristics. Other files with such a structure are for example 'account_history.json',~~

~~‘devices.json’ and ‘settings.json’. However, a substantial number of files contains data that is less structured. Examples of such files are ‘comments.json’, ‘media.json’, ‘messages.json’ and ‘stories_activities.json’. Furthermore, data subjects at Instagram are not necessarily natural persons. Data subjects at Instagram can be identified by a single and unique Username. Typically, natural persons have individual accounts with an accompanying username, but other institutions, such as for example retail shops or bands can also have an individual account with an accompanying username.~~

Most large data controllers currently comply with the right of data access by providing users with the option to retrieve an electronic DDP. This DDP typically comes as a compressed folder containing text and/or media files in which all the digital traces left behind by the data subject with respect to the data controller are stored. Table 4 shows that the content and structure of DDPs differs among data controllers. Differences between DDPs from the same data controller can also occur among data subjects and over time. These differences may be caused by data subjects using different features provided by the data controller or by the fact that the DDP is a snapshot of the data collected by the data subject up to that point. However, other important factors also play a role. First, data controllers can develop new features through which new types of data of the data subject are collected. Second, other features may be phased out. Third, some data (for example search history) is only saved for a limited amount of time and is destroyed by the data controller after that period. In that case, it will also not be present in the DDP anymore. Finally, the GDPR is still relatively new and data controllers continue to optimize the processes used to transfer the relevant data to its subjects, leading to changes in the structure of DDPs.

From Table 4 it can be concluded that the Instagram DDP contains many features that can also be found in DDPs of other data controllers. Common features are the presence of both text and/or media files, the presence of both structured and unstructured text and the presence of specific types of person identifying information (PII). Therefore, an algorithm that is able to de-identify Instagram DDPs also contains the features needed to de-identify many of the DDPs of other data controllers. To summarize, the developed algorithm is able to handle: ~~To summarize, software to de-identify Instagram DDPs should be able to handle:~~

- An ever changing file structure,
- both visual and textual content,
- different file formats,
- ~~Files in highly structured and highly unstructured format and~~

		FACEBOOK[a]	**WHATSAPP**[b]	**TWITTER**[c]	**SNAPCHAT**[d]	**INSTAGRAM**[e]
DDP INFO	DDP name	facebook-¡profile_name¿	My account information.zip WhatsApp chat-¡group or contactname¿.zip	Archive	mydata¡hashed_code¿	username_¡date_of_download¿
	DDP format	.zip	.zip	.zip	.zip	.zip
	Type of files	media, text	media, text	media, text	text	media, text
	Structure	Content folders ¿ content files	Separate DDP per conversation	Content folders ¡ content files	Index file & Format (i.e., json and html) folders ¿ content files	Content text files and Content folders ¿ content media files
MEDIA FILES	Format of media files	.PNG, .JPG, .MP4	.JPG, .MP4, .HTML	.PNG	-	.JPG, .MP4
	Folder structure	All images, videos, stickers are categorized and stored in corresponding folders. There are no loose files.	All images, videos, stickers in single folder	-	All images and videos are categorized and stored in corresponding folders. There are no loose files.	
PII IN MEDIA	Faces	-/+	-/+	-/+	-	-/+
	Written text	-/+	-/+	-/+	-	-/+
	(user)name tags	-	-	-	-	-/+
TEXT FILES	Format of text files	.JSON or .HTML	.TXT, .OPUS, .HTML	.JS or .HTML	.JSON and .HTML	.JSON
	Folder structure	All text files are categorized and stored in corresponding folders. There are no loose files.	All text in single file per conversation	There is one text file per month.	Text files are not categorized and stored in (sub) folders. They are displayed as loose files.	Text files are not categorized and stored in (sub) folders. They are displayed as loose files.
	Structured data	+	+	+	+	+
	Unstructured data (i.e., containing free text)	+	+	+	-/+	+
PII IN TEXT	Usernames	-/+	-/+	+	+	+
	(first) Names	+	+	-/+	-/+	+
	Email addresses	+	+	+	-/+	+
	Phone numbers	-/+	+	-/+	-/+	-/+
	Locations	-/+	-/+	-/+	-/+	-/+

Table 4: Overview of content and structure of DDPs of five data controllers. Note that if a certain object is present in DDPs, this is indicated with +. If it often occurs within the DDP, a -/+ is used. Finally, if said object is not present, a - is used.

[a] https://www.facebook.com/help/1701730696756992

[b] https://faq.whatsapp.com/general/account-and-profile/how-to-request-your-account-information/

[c] https://help.twitter.com/nl/managing-your-account/how-to-download-your-twitter-archive

[d] https://support.snapchat.com/en-US/a/download-my-data

[e] https://help.instagram.com/181231772500920?helpref

~~different variants in between~~files ranging from highly structured to highly unstructured formats,

- ~~Natural persons and other users which are identified by their unique username.~~the masking of usernames of natural persons or other users.

4. Fourth, the method used for de-identifying faces in media (photos and videos), blurring, has been shown to be prone to re-identification attacks (using deep learning) by McPherson et al.: `https://arxiv.org/abs/1609.00408`**. As a consequence, I would use more robust methods for de-identifying faces in images and videos. This could be a key novel contribution of the paper, which currently lacks strong technical contributions.** *We are aware of this issue. However, we believe that this currently meets the privacy preserving level that we aim for. As we provide the software free and open source, and as it consists of various modules for various types of de-identification, users can adapt certain modules if interested, for example the de-identification method for faces in media. We now make our aims in terms of privacy more explicit in the introduction*

> However, the data present in DDPs can be deeply private and potentially sensitive. This poses a major challenge to using DDPs for scientific research. Participants might not be willing to share this sensitive data. However, researchers are often only interested in a part of the DDP and do not need the sensitive data. Although an interesting solution is to extract relevant features locally on the device of the participant (27), this workflow is not suitable for all research purposes. When, for example, an exploratory approach is of interest, or when the aim is to develop or improve the performance of an extraction algorithm, local extraction would limit the analytic possibilities. In such situations, collection of the *complete* DDP is desired, which requires challenges caused by the sensitivity of the data to be overcome. An example of such a research project is Project AWeSome (28), which collects complete Instagram DDPs from research participants. The participants' DDPs are stored in a secured environment where they are de-identified using the de-identification algorithm proposed in this manuscript. Only after the sensitive information is adequately masked, can the DDPs be shared with the applied researchers for substantive analyses.

and reflect on the potential improvements in the discussion section:

> A last point of discussion considers the safety standards that are currently adhered. We have clearly stated that the algorithm aims to prepare the DDPs in such a way that they can be processed as any other type of sensitive research data, supplemented with other measures such as using shielded (cloud) environments. If the researchers

would like to share the data with others on a more flexible level, for example the currently used blurring algorithm is not sufficient as it can be prone to re-identification (25).

5. Besides, I encouraged the authors to proof-read their paper and correct the numerous typos (incl. in subsections' titles). *We have rewritten and restructured substantive parts of the manuscript, as well have checked it by a native speaker.*

References

[1] L. Sweeney, k-anonymity: A model for protecting privacy, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **10**(05) (2002), 557–570.

[2] K. El Emam and F.K. Dankar, Protecting privacy using k-anonymity, *Journal of the American Medical Informatics Association* **15**(5) (2008), 627–637. https://doi.org/10.1197/jamia.M2716.

[3] C.A. Kushida, D.A. Nichols, R. Jadrnicek, R. Miller, J.K. Walsh and K. Griffin, Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies, *Medical care* **50**(Suppl) (2012), S82. doi:10.1097/MLR.0b013e3182585355.

[4] V. Menger, F. Scheepers, L.M. van Wijk and M. Spruit, DEDUCE: A pattern matching method for automatic de-identification of Dutch medical text, *Telematics and Informatics* **35**(4) (2018), 727–736. https://doi.org/10.1016/j.tele.2017.08.002.

[5] J. Simon, Amazon Comprehend Medical–Natural Language Processing 24 for Healthcare Customers, *Retrieved April* **18** (2018), 2019. https://aws.amazon.com/comprehend/medical/.

[6] L. Liu, O. Perez-Concha, A. Nguyen, V. Bennett and L. Jorm, De-identifying Hospital Discharge Summaries: An End-to-End Framework using Ensemble of De-Identifiers, *arXiv preprint arXiv:2101.00146* (2020).

[7] P.M. Heider, J.S. Obeid and S.M. Meystre, A Comparative Analysis of Speed and Accuracy for Three Off-the-Shelf De-Identification Tools, *AMIA Summits on Translational Science Proceedings* **2020** (2020), 241. doi:PMCID: PMC7233098.

[8] C. van Toledo, F. van Dijk and M. Spruit, Dutch Named Entity Recognition and De-identification Methods for the Human Resource Domain, *arXiv preprint arXiv:2106.02287* (2021).

[9] G. Coppersmith, M. Mitchell, C. Harman, M. Dredze and R. Leary, Deidentify Twitter, 2017. https://github.com/qntfy/deidentify_twitter.

[10] M. Azure, Microsoft Azure cognitive services, 2021. `https://azure.microsoft.com/nl-nl/services/cognitive-services/face/`.

[11] T. Esler, facenet pytorch, 2019. doi:10.34740/KAGGLE/DSV/845275. `https://www.kaggle.com/timesler/facenet-pytorch`.

[12] R. Nosowsky and T.J. Giordano, The Health Insurance Portability and Accountability Act of 1996 (HIPAA) privacy rule: implications for clinical research, *Annu. Rev. Med.* **57** (2006), 575–590.

[13] J. Trienes, D. Trieschnigg, C. Seifert and D. Hiemstra, Comparing rule-based, feature-based and deep neural methods for de-identification of Dutch medical records, *CEUR Workshop Proceedings* **2551** (2020), 3–11.

[14] Ö. Uzuner, Y. Luo and P. Szolovits, Evaluating the State-of-the-Art in Automatic De-identification, *Journal of the American Medical Informatics Association* **14**(5) (2007), 550–563. doi:10.1197/jamia.M2444.

[15] OpenAIRE, amnesia. `https://amnesia.openaire.eu/index.html`.

[16] G. Beigi and H. Liu, *A Survey on Privacy in Social Media*, Vol. 1, 2020, pp. 1–38. ISSN 2691-1922. ISBN ISBN 0001417126. doi:10.1145/3343038.

[17] F. Prasser, F. Kohlmayer, R. Lautenschläger and K.A. Kuhn, Arx-a comprehensive tool for anonymizing biomedical data, in: *AMIA Annual Symposium Proceedings*, Vol. 2014, American Medical Informatics Association, 2014, p. 984.

[18] M. Templ, A. Kowarik and B. Meindl, Statistical disclosure control for micro-data using the R package sdcMicro, *Journal of Statistical Software* **67**(4) (2015). doi:10.18637/jss.v067.i04.

[19] G. Coppersmith, M. Dredze, C. Harman, K. Hollingshead and M. Mitchell, CLPsych 2015 shared task: Depression and PTSD on Twitter, in: *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2015, pp. 31–39.

[20] R. Dorn, A.L. Nobles, M. Rouhizadeh and M. Dredze, Examining the Feasibility of Off-the-Shelf Algorithms for Masking Directly Identifiable Information in Social Media Data **1996** (2020). `http://arxiv.org/abs/2011.08324`.

[21] G. Beigi, K. Shu, R. Guo, S. Wang and H. Liu, Privacy preserving text representation learning, in: *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, 2019, pp. 275–276.

[22] L. Backstrom, C. Dwork and J. Kleinberg, Wherefore art thou R3579X? Anonymized social networks, hidden patterns, and structural steganography, in: *Proceedings of the 16th international conference on World Wide Web*, 2007, pp. 181–190.

[23] A. Narayanan and V. Shmatikov, Robust de-anonymization of large sparse datasets, in: *2008 IEEE Symposium on Security and Privacy (sp 2008)*, IEEE, 2008, pp. 111–125.

[24] H. Mao, X. Shuai and A. Kapadia, Loose tweets: an analysis of privacy leaks on twitter, in: *Proceedings of the 10th annual ACM workshop on Privacy in the electronic society*, 2011, pp. 1–12.

[25] R. McPherson, R. Shokri and V. Shmatikov, Defeating image obfuscation with deep learning, *arXiv preprint arXiv:1609.00408* (2016).

[26] S. Ribaric, A. Ariyaeeinia and N. Pavesic, De-identification for privacy protection in multimedia content: A survey, *Signal Processing: Image Communication* **47** (2016), 131–151.

[27] L. Boeschoten, J. Ausloos, J. Moeller, T. Araujo and D.L. Oberski, Digital trace data collection through data donation, *arXiv preprint arXiv:2011.09851* (2020).

[28] I. Beyens, J.L. Pouwels, I.I. van Driel, L. Keijsers and P.M. Valkenburg, The effect of social media on well-being differs from adolescent to adolescent, *Scientific Reports* **10**(1) (2020), 1–11.

[29] A. Hundepool and P.-P. De Wolf, Statistical discosure control, *Method Series* (2012), 1–49. `https://www.cbs.nl/en-gb/our-services/methods/statistical-methods/output/output/statistical-disclosure-control`.

[30] Aticle 29 Data protection working party, Opinion 05/2014 on Anonymisation Techniques, *European Commission* (2014), 1–37. `https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf`.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg et al., Scikit-learn: Machine learning in Python, *the Journal of machine Learning research* **12** (2011), 2825–2830. `https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?source=post_page---------------------------`.

[32] H. Hanke and D. Knees, A phase-field damage model based on evolving microstructure, *Asymptotic Analysis* **101** (2017), 149–180.

[33] E. Lefever, A hybrid approach to domain-independent taxonomy learning, *Applied Ontology* **11**(3) (2016), 255–278.

[34] P.S. Meltzer, A. Kallioniemi and J.M. Trent, Chromosome alterations in human solid tumors, in: *The Genetic Basis of Human Cancer*, B. Vogelstein and K.W. Kinzler, eds, McGraw-Hill, New York, 2002, pp. 93–113.

[35] P.R. Murray, K.S. Rosenthal, G.S. Kobayashi and M.A. Pfaller, *Medical Microbiology*, 4th edn, Mosby, St. Louis, 2002.

[36] E. Wilson, Active vibration analysis of thin-walled beams, PhD thesis, University of Virginia, 1991.

[37] G.D.P. Regulation, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46, *Official Journal of the European Union (OJ)* **59**(1–88) (2016), 294. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L:2016:119:FULL&from=EN.

[38] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks, *IEEE Signal Processing Letters* **23**(10) (2016), 1499–1503. doi:10.1109/LSP.2016.2603342.

[39] G. Bradski, The OpenCV Library, *Dr. Dobb's Journal of Software Tools* (2000). https://opencv.org/.

[40] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He and J. Liang, EAST: An Efficient and Accurate Scene Text Detector, in: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2642–2651. doi:10.1109/CVPR.2017.283.

[41] O. Yadong, frozen east text detection, 2018. https://github.com/oyyd/frozen_east_text_detection.pb.

[42] B. van der Sloot, *The General Data Protection Regulation in Plain Language*, Amsterdam University Press, 2020. ISBN ISBN 978-94-6372-651-1.

[43] B. Zhong, Y. Huang and Q. Liu, Mental health toll from the coronavirus: Social media usage reveals Wuhan residents' depression and secondary trauma in the COVID-19 outbreak, *Computers in human behavior* **114** (2021), 106524. https://doi.org/10.1016/j.chb.2020.106524.

[44] A. Jungherr, *Analyzing Political Communication with Digital Trace Data: The Role of Twitter Messages in Social Science Research*, Contributions to Political Science, Springer International Publishing, 2015. ISBN ISBN 978-3-319-20318-8. doi:10.1007/978-3-319-20319-5. https://www.springer.com/gp/book/9783319203188.

[45] H. Schoen, D. Gayo-Avello, P. Takis Metaxas, E. Mustafaraj, M. Strohmaier and P. Gloor, The power of prediction with social media, *Internet Research* **23**(5) (2013), 528–543. doi:10.1108/IntR-06-2013-0115.

[46] H. van Kemenade, wiredfool, A. Murray, A. Clark, A. Karpinsky, nulano, C. Gohlke, J. Dufresne, B. Crowell, D. Schmidt, A. Houghton, K. Kopachev, S. Mani, S. Landey, vashek, J. Ware, Jason, D. Caro, S. Kossouho, R. Lahd, S. T., A. Lee, E.W. Brown, O. Tonnhofer, M. Bonfill,

P. Rowlands, F. Al-Saidi, M. Górny, M. Korobov and M. Kurczewski, python-pillow/Pillow 8.0.0, Zenodo, 2020. doi:10.5281/zenodo.4088798.

[47] M. Kosinski, D. Stillwell and T. Graepel, Private traits and attributes are predictable from digital records of human behavior, *Proceedings of the National Academy of Sciences* **110**(15) (2013), 5802–5805. doi:10.1073/pnas.1218772110.

[48] G. King, Ensuring the data-rich future of the social sciences, *science* **331**(6018) (2011), 719–721. doi:10.1126/science.1197872.

[49] P. Korshunov and S. Marcel, Deepfakes: a new threat to face recognition? assessment and detection, *arXiv preprint arXiv:1812.08685* (2018).

[50] S.L. Garfinkel et al., De-identification of personal information, *National institute of standards and technology* (2015).

[51] A. Dehghan, A. Kovacevic, G. Karystianis, J.A. Keane and G. Nenadic, Combining knowledge- and data-driven methods for de-identification of clinical narratives, *Journal of Biomedical Informatics* **58** (2015), S53–S59. doi:10.1016/j.jbi.2015.06.029.