Privacy versus Open Science

Simon Dennis[1,6]
Paul Garrett[2]
Hyungwook Yim[1,3]
Jihun Hamm[4]
Adam Osth[1]
Vishnu Sreekumar[5]
Ben Stone[1,6]

University of Melbourne[1]
University of Newcastle[2]
University of Tasmania[3]
Ohio State University[4]
National Institutes of Health[5]
Unforgettable Technologies Pty Ltd[6]

Correspondence concerning this article should be addressed to Simon J. Dennis (School of Psychological Sciences, The University of Melbourne, Australia. E-mail: simon.dennis@gmail.com)

Acknowledgments

This research was supported by the Australian Government through the Australian Research Council's Discovery Projects funding scheme (project DP150100272). The views expressed herein are those of the authors and are not necessarily those of the Australian Government or Australian Research Council.

Conflict of Interest: A critical part of the open science movement involves the open disclosure of both the explicit and potential implicit motivations for scientific work. In that spirit, the reader should be aware that Professor Dennis is the CEO of a startup called Unforgettable Technologies Pty Ltd (UT) that specializes in providing privacy preserving experience sampling collection and analysis services. Dr Stone is the CTO of UT.

Abstract

Pervasive internet and sensor technologies promise to revolutionize psychological science. However, the data collected using these technologies is often very personal - indeed the value of the data is often directly related to how personal it is. At the same time, driven by the replication crisis, there is a sustained push to publish data to open repositories. These movements are in fundamental conflict. One cannot publish private data. In this paper, we propose a way to navigate this issue. We argue that there are significant advantages to be gained by ceding the ownership of data to the participants who generate it. Then we provide desiderata for a privacy preserving platform. In particular, we suggest that researchers should use an API to perform experiments and run analyses rather than observing the stimuli themselves. We argue that this method not only improves privacy, but will also encourage greater compliance with good research practices than is possible with open repositories.

Keywords: privacy, open science, open repositories, differential privacy

Introduction

The scientific community is in the midst of two revolutions which are about to collide. On the one hand, as a consequence of the replication crisis that has occurred across multiple disciplines (Open Science Collaboration, 2015), researchers are being encouraged to share their data in open repositories thus facilitating reanalysis and increasing the transparency of the scientific enterprise. At the same time, the introduction of a range of consumer sensor devices along with the availability of online social data is providing researchers with an unprecedented window into the everyday lives of people. The data collected using these experience sampling methodologies have the potential to transform our understanding of human behavior. However, it is an unfortunate fact that the most valuable data is often the most sensitive. Our ability to fully exploit these new data sources will be directly related to our ability to preserve the privacy of the people who provide it.

The recent Cambridge Analytica scandal, in which tens of millions of Facebook™ profiles were leaked to a political consulting firm, has demonstrated both the enormous power and the enormous danger involved in large scale data collection. As the public take stock and governments react, we can expect research practices to come under increasing scrutiny. There is a very real danger that the public good that could be derived from studying dense data sets will be thwarted.

The Cambridge Analytica case is particularly instructive as there was no “hack”. The data was given to a legitimate researcher, Aleksandr Kogan from Cambridge University, as he was conducting personality research on the platform. He then passed the data to Cambridge Analytica, allegedly in breach of Facebook’s terms and conditions.

The case illustrates a fundamental fact of data - once it is given it is very hard to take back. Cambridge Analytica claim to have deleted all copies and have submitted to audit, but there really is no way to put the genie back in the bottle with complete confidence. If we are to realize the full potential of big data without seriously compromising civil liberties we need to reconceptualize our relationship to data.

In particular, one cannot simply upload participants’ email, GPS, phone call and SMS data to a public repository. Even if obvious personal identifiers such as names and addresses are removed, the scope for reverse engineering identity from diverse sources is surprisingly easy (Narayanan & Shmatikov, 2010). To upload knowing one cannot protect the identity of participants is both unethical and ultimately self defeating, as engaging participants will become very difficult once they understand the implications of being involved in these studies.

In this paper, we will discuss the value of experience sampling paradigms, the objectives of the open science movement and provide a pathway to resolve the conflict by discussing the architecture of a privacy preserving experience sampling platform.

The Promise of Experience Sampling

Most empirical research in psychology either involves administering surveys across a cohort or occurs in the laboratory. While a great deal has been learned about psychological processes using these methods, they provide very sparse samples of human behaviour under quite unnatural conditions. Experience sampling methods (Csikszentmihalyi & Larson, 1992) use smartphones, wearable sensors, social media and the internet of things to collect much denser data over longer time periods as people engage in the activities of daily living. For instance, in an ongoing project with bipolar patients, we are collecting accelerometry data 8-15 times a second, 24-7, for a year. That accumulates to about ½ billion data points per subject providing a fairly complete record of their movement during the collection period.

Experience sampling methods (ESM) come in both passive and active forms. Passive ESM involves the use of data sources that people generate automatically as a consequence of their activity such as accelerometry and GPS. Active ESM (also known as Ecological Momentary Assessment, EMA, Shiffman, Stone & Hufford, 2008) interrupts participants throughout their day and requires them to provide a response. While active ESM is less naturalistic than passive ESM, it allows the sampling of internal mood and other cognitive states that cannot be reliably ascertained based on passive data alone. We are using active ESM to measure mood states in the bipolar study mentioned above.

With the appropriate analysis methods one can ask questions using ESM data that cannot be addressed using traditional methodologies. For instance, we have been able to characterize the dimensionality of people's visual experience (Sreekumar, Dennis, Doxas, Zhuang & Belkin, 2014), to examine the neural representations of time and space over time scales up to a month (Nielson, Smith, Sreekumar, Dennis & Sederberg, 2015) and to model the processes involved in real world episodic memory (Dennis, Yim, Sreekumar, Evans, Garrett & Sederberg, 2017).

The last of these studies illustrates the promise that experience sampling methods hold. In this study, participants wore a smartphone in a pouch around their necks for two weeks. The phone collected audio, image, movement and spatial information. A week after data collection, the participants were presented with a selection of their images and asked on which day they were taken. Because of the specificity of the data, predictions could be made on an item by item and person by person basis and the relative contributions of audio, visual, movement and spatial information to memory performance could be assessed.

In the laboratory, it is common for experimenters to administer lists of random words for participants to study, deliberately striping away the structure that participants might exploit in their everyday memory to better elucidate the underlying processes. However, when presented with an unfamiliar structure, the participants may adopt strategies that they do not normally employ. The introduction of control may perversely complicate the problem that researchers are

trying to solve, as the results of such experiments may not reflect the type of retrieval that participants normally engage in.

Furthermore, people's lives are dominated by repeating experiences (Sreekumar et al., 2014; Sreekumar, Dennis, & Doxas, 2017) and such recurrence structures are thought to influence performance (Dennis & Humphreys, 2001; Osth & Dennis, 2015). ESM provides us a way to quantify recurrence structures in people's lives and how they relate to cognition - something that has not been possible until now.

Across a range of areas, experience sampling approaches promise to provide a more comprehensive, ecologically valid and translationally relevant psychological science. However, the data being collected is very sensitive and issues of ownership and access become much more acute. In the next section, we discuss the open science movement and highlight the fundamental conflict between it and experience sampling research.

The Open Science Movement

Driven by the specter of the replication crisis (Open Science Collaboration, 2015), a deepening concern for the integrity of scientific evidence has led to the open science movement and the construction of best practice guidelines for research (Nosek et al., 2015). Open science is a term that incorporates a set of distinct practices and perspectives (Mehler & Weiner, 2018; see in particular Whitaker quote). Open can refer to:

- the transparency of research practices and analysis methods (with the rise of preregistration as an emerging standard for confirmatory research)
- the degree of access that people have to research (particularly taxpayer funded research) - open access
- the transparency of the commercial and other motivations for the conduct of the research
- the ability of people to examine, contribute to and use software - open source
- the ability of non-academics to engage in the scientific process - citizen science
- the ability of all people to enter academia regardless of race, gender, nationality etc
- the ability to access the data from which conclusions have been draw - open data

Working towards open science in all of these senses is laudable. In this paper, however, we are concerned with the problems associated with open access to data. An increasing number of repositories and sharing standards have been created to facilitate this aim. These include Dataverse (King, 2007), Dryad (White, Carrier, Thompson, Greenberg & Scherle, 2008), the Interuniversity Consortium for Political and Social Research (ICPSR, Taylor, 1985), the Open Science Framework (Foster & Deardorff, 2017), and the Qualitative Data Repository (QDR; Kirilova & Karcher, 2017). Other commentators have advocated the immediate and automatic uploading of data to stores such as github - the so called ‘born open approach’ (Rouder, 2016).

In many laboratory paradigms, sharing in this fashion is unproblematic. When data is more sensitive, open sharing of this form is not possible.

The problems associated with sharing data are particularly acute in the case of qualitative data and so the issue has been a long standing point of discussion (Kirilova & Karcher, 2017). Qualitative researchers tend to foster much closer relationships with their participants and even before the advent of sophisticated computational techniques for reverse engineering identity become feasible, the data that they collected was difficult to de-identify. The nature of qualitative analysis requires researchers to personally engage with the raw materials. Many qualitative researchers have come to the conclusion that data sharing is simply not possible. Others, however, have sought to elucidate policies and processes to allow sharing (Kirilova & Karcher, 2017).

A key issue raised by this work is the nature of consent. Kirilova and Karcher (2017) advocate researchers engaging in extended conversations and providing participants with multiple options in terms of the way their data might be shared. Such a policy though places a great deal of responsibility on the researcher to interpret the subjects' wishes. The willingness to share may depend on many factors - most notably the purpose of the research - that are not available at the time that the data is collected. The question arises - should data ever be shared without the consent of the participant in that particular case? Underlying this question is an even more fundamental one - to whom should the data belong?

Whose data is it?

A critical issue in discussions of privacy is who should own the data. In current practice, the ownership of the data that psychological researchers collect typically transfers to their host institution. Restrictions to this ownership enshrined in ethical protocols and privacy law usually afford participants the right to amend and/or delete data that they have provided (e.g. the Australian Privacy Act 1988, www.oaic.gov.au/privacy-law/privacy-act/). In practice, logistical barriers tend to mean that few participants exercise their right to modify their data and researchers typically treat data as if they own it - for instance, feeling little compunction about taking a dataset from one institution to the next when they move and/or publishing to open platforms without seeking institutional approval.

While participants may show little concern about giving away an hour's worth of laboratory data, they may feel rather differently about giving away experience sampling data that may have been collected over several years and contain much more sensitive information. The public outcry about Facebook™ and the Cambridge Analytica scandal demonstrates that individuals are becoming increasingly protective and concerned about their online data and their privacy.

Any such discussion necessarily sits within the context of efforts by government to institute national data banks. In the health sphere, these efforts are gathering pace. In 2016, the

Australian government launched “My Health Record”, a permanent electronic record of interactions with healthcare providers across the nation. On February 25th 2018, 5.5 million people were registered with My Health Record (23% of the Australian population) and some 10,754 healthcare providers were connected (Australian Digital Health Agency, 2018). The objective is to reach 98% coverage by the end of 2018. Similar efforts have been under way for some time across multiple countries (Ludwick & Doucette, 2009). While the potential advantages are substantial, the government ownership of data remains controversial (Gagnon et. al, 2016; Anderson, 2007). Similarly, the corporate ownership of data is coming under increasing scrutiny due to events like the Cambridge Analytica breach discussed earlier.

An alternative to institutional ownership of data is to have participants retain ownership. Under this model, data becomes an asset that participants allow researchers to license - either in the interests of the public good or for compensation. Participants would build a personal data warehouse that might include generic data that could be used for multiple purposes such as GPS coordinates as well as surveys or experimental responses requested by researchers. As time progresses, the value of the data asset would grow with its extent. Researchers might then offer compensation for a given type of data and participants would consent on a case by case basis. The researchers would be purchasing the right to analyze data, not the data itself, so the subject is then free to participate in other studies (including replications) and to earn additional compensation from other researchers for the same data.

While this proposal requires a shift in the way in which researchers understand their relationship to data, it has a number of advantages:

(a) participants are making decisions about the use of their data on a case by case basis. Researchers would provide a statement about the use to which the data would be put in their request and participants would provide their consent with this use in mind. An increase in transparency would make privacy easier to regulate.

(b) participants would be incentivised to curate their data to ensure it is as complete as possible as this would make it more likely to be requested. Missing data is a much more significant problem in experience sampling paradigms than is typically the case in laboratory work, so any dynamic that engages participants is desirable.

(c) currently, people’s understanding of the relative value of data and the privacy implications of allowing others to access it is rudimentary. Global information technology companies like Google™ and Facebook™ collect large amounts of data in exchange for allowing people to access their systems, but do not provide financial compensation. If subjects retain ownership of their data and participate in a data marketplace, they will come to understand which kinds of data are most valuable both to them collectively and to researchers. The promise then is that a more nuanced understanding of privacy will emerge.

(d) in many cases the data that is most valuable to researchers belongs to members of special populations who are commonly financially disadvantaged. Ensuring they are able to retain ownership of their data could provide a supplemental income to people with financial needs. If this mechanism is to work to a substantial degree, the data should be seen as capital for rent not as labour as is the case with internet work providers such as Amazon Mechanical Turk or Prolific Academic, currently.

(e) the weak link with current open repositories is the time between the publication of the paper and the posting of the data. Well meaning researchers struggle to format, document and post their data. Publication standards aimed at sharing will certainly affect this tendency, however, this requires surveillance and enforcement. By contrast, in the approach advocated herein the data would be submitted directly to the repository by the participant. Using a key published with the paper, a would be replicator can immediately access the set (with the permission of the participants) without additional processes, thus removing a key impediment to sharing (c.f. born open, Rouder, 2016).

(f) participant ownership of data may lead to increased engagement in and understanding of the scientific process - common objectives in citizen science projects (Bonney, Shirk, Phillips, Wiggins, Ballard, Miller-Rushing & Parrish, 2014).

In order to realize this kind of privacy preserving platform, several technical challenges must be addressed. We outline these in the next section.

A Privacy Preserving Platform

There are several critical aspects to consider when designing a privacy preserving experience sampling platform. These include the collection mechanisms, the search and visualization interfaces, the data analysis platform, the experimental platform and the legal framework in which the service operates.

Collection Mechanisms

The first question when trying to maintain privacy is what data is collected in the first instance. Too often, current apps and services lack transparency in what they collect. Even when it is possible to select what forms of data are collected, the interface is often obscure. There are a few design principles that apps and services can implement to improve the privacy interface:

a) There should be a prominent all stop button to allow users to cease all recording when they wish and to easily resume collecting when they wish. If users must disable each data stream (e.g. GPS, audio, accelerometry) individually there is a greater probability that they will miss one and continue collecting data when they did not intend to.

b) Conversely, users should be able to turn on and off individual data streams so that they can decide what they are comfortable sharing and change this at any time. Each stream comes with

a different tradeoff in terms of the privacy that is relinquished and the degree it is necessary for the purposes of collection. Excessive bundling can be used to coerce users to share data they would not otherwise choose to share.

c) When data is collected on a smartphone or similar device there ought to be a delay between when the data is collected and when it is uploaded during which the user can delete it - somewhat like the mechanisms live television programs implement to avoid the broadcasting of inappropriate content. Once the data leaves the device it is more difficult to control and so there should be an opportunity for users to prevent it from uploading.

d) Consideration should be given to the format of the data that is being uploaded to assess whether a more private form would serve similar purposes. For instance, when recording audio it may not be necessary (or legal) to retain raw audio in a form that can be replayed. Sometimes, however, it is enough to sample sporadically and to retain only frequency information. Machine learning algorithms can be used to determine ambient qualities of the audio such as whether it contains voices or traffic, without being able to replay the stream. As another example, when recording phone calls or SMSs, for many research purposes it is sufficient to retain the time of calls and perhaps a unique identifier for the caller that is not their name or number. In this way, the temporal characteristics and the diversity of callers can be ascertained without retaining more sensitive information like the content of the SMSs or the identity of the callers. It is a common maxim in experimental research that one should record everything as you can never be sure what future analyses you might want to conduct. With private data, however, that edict must be balanced against the need to protect participants from future analyses they might not want conducted.

Search and Visualization Interfaces

In order for participants to actively maintain their privacy, it is critical that the system have a usable mechanism to allow participants to understand what data they they are allowing researchers to access. That interface will also be critical to provide participants with the ability to delete portions of the data that they do not wish to be available. It is typically the case in experimental protocols that participants are afforded the right to have their data deleted should they wish. In practice, however, the mechanisms to allow people to sift through their data are rudimentary and so very few participants ever make a deletion request.

Making data available to participants in a usable form is the most difficult and perhaps the most under appreciated component of a privacy preserving platform. For analysis purposes, we often store data in cryptic files or relational databases. Participants are more likely, however, to be familiar with search engines and should be provided with such a mechanism to access their data. However, search engines are only as good as the tags that are indexed to recover the data. For instance, you may wish to collect GPS data, and the coordinates alone may be sufficient for your purposes. However, they will only be usable by participants if they can be referenced by address, and so additional effort is required. Similarly, participants need straightforward interfaces to be able to select data by date and to visualize the data that is

stored in the form of maps and calendars, so they truly understand what they are allowing researchers to access.

Beyond their usefulness when completing transactions with researchers, search and visualization interfaces can be intrinsically motivating - thus encouraging engagement of participants with their data. A search tool provides a form of memory prosthesis that people can use to recall what they were doing at any given time. A visualization tool can allow users to discover patterns and relationships in their lives about which they may not have been conscious. These kinds of facilities are critical if we are to transform into a more data aware populace.

Analyzing Data

“The emergence of powerful re-identification algorithms demonstrates not just a flaw in a specific anonymization technique(s), but the fundamental inadequacy of the entire privacy protection paradigm based on “de-identifying” the data. De-identification provides only a weak form of privacy. It may prevent “peeping” by insiders and keep honest people honest. Unfortunately, advances in the art and science of re-identification, increasing economic incentives for potential attackers, and ready availability of personal information about millions of people (for example, in online social networks) are rapidly rendering it obsolete.” (Narayanan & Shmatikov, 2010)

If the “release and forget” approach employed by open repositories is not a viable solution to providing greater transparency in science then what can be done? Narayanan and Shmatikov (2010) argue that a better (though not foolproof) solution is to define privacy in terms of computation rather than in terms of data. Rather than provide access to the data directly, researchers would be given an application programming interface with which to interact with the data. Although the researcher is unable to access the raw data, they would be able to run analyses that do. Code could be written that runs on the raw data and tests hypotheses, but that provides only the inferential statistics and groupwise descriptive statistics to the researcher. These measures are derived from many participants and could be constructed to bound the probability that individual data could be reconstructed from the composite measures. This notion has been formally encapsulated in the concept of differential privacy that we will discuss next.

Understanding Differential Privacy

Differential Privacy is quantifiable probabilistic definition of what it means to guarantee privacy motivated by cryptography (Dwork et al., 2004; 2006). Suppose a database consists of the body weights of 20 subjects, and we wish to allow a (potentially malicious) analyst to compute the average without providing access to individual data. One may imagine that providing the mean would not be a breach of privacy as any individual’s value cannot be reconstructed uniquely from the mean. However, one must consider what can be learned about a person if the attacker

has access to additional information. For example, if the attacker can issue another query to the system that includes all of the same people as before except for the individual of interest, then the weight of that person can be readily discerned. Differential privacy tries to prevent such inference by reporting a noisy answer to the query. If one scales the noise appropriately one can decrease the probability that the individuals weight can be ascertained, while also preserving the usefulness of the data analysis platform.

Formally, let D and D' be any pair of databases which differ in exactly one entry (called neighbors), and the mechanism M(.) be a random function that takes in a database and outputs a random output such as a real vector. We call the mechanism (ε δ)-differentially private, if for all neighboring databases D and D', and measurable sets S of the range of the output, the inequality $P(M(D) \in S) \leq e^{\varepsilon} P(M(D') \in S) + \delta$ holds.

The interpretation of the inequality is as follows. Suppose D is a database of sensitive measurements from N subjects, and D' is a similar database of the same subjects except one particular subject whose data is included in D' but not in D. The mechanism M(D) is the result of an analysis that the analyst has queried, such as statistics of the data D, plus some random noise to the answer to provide privacy. If the mechanism M is differentially private, then, it is difficult to infer whether the output M(.) is from D or D' (whose probabilities differ only by the multiplicative and additive constants ε and δ). Furthermore, since D and D' differ by only one entry, it is difficult to guess if a particular subject was included (D') or not included (D) from the output, even if the adversary knows the sensitive data of the N-1 subjects.

Differential privacy packages such as diffpriv.R[1] demonstrate how this concept can be implemented. While they are useful for illustrative purposes, they do not provide a solution in themselves as to apply them one must have access to the raw data. Privacy protection is only afforded when such a package is incorporated into a data collection and access control platform.

The Experimental Platform

While some analyses can be completed using only the experience sampling data, it is commonly necessary to also administer experimental paradigms. For instance, in an experiment we are currently running participants are presented with a map and are asked where they were at a given time (see Figure 1). Four alternatives are presented and participants make a response. Rather than have researchers examine the person's data and select the alternatives, these are chosen by the experimental code, which is run within a password protected environment. The participant makes their selections and the data is added to their personal repository. The researcher also has access to these records, but they contain only the event identifiers (random keys) that correspond to the target and distractor coordinates. This approach allows the researchers to run subsequent analyses that incorporate factors such as the distance

[1] cran.r-project.org/web/packages/diffpriv/index.html

between points, without having access to the GPS coordinates themselves. In other experiments, we have created similar algorithms that select images for presentation, automatically excluding those that are too dark, or blurred or that contain too little information.

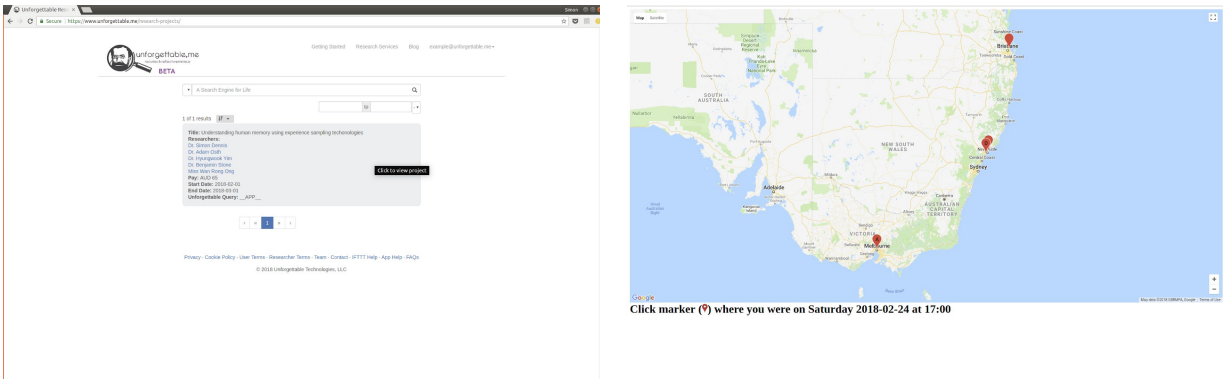

Figure 1: Screenshots showing the implementation of the memory for WHERE experiment. a) the interface that participants see when deciding whether to sign up for an experiment. Note that this happens in a password protected environment. b) The task showing the map with four alternatives. These alternatives are generated by the experimental code from the users' data, but are not available to researchers directly. Users have the ability to pan and zoom the map.

While not being able to examine the stimuli that have been presented to participants can make it more difficult to debug code and to discover regularities, it also introduces a wholesome discipline in the stimuli selection process. The algorithms can and should be published to make the selection process more transparent. The human selection of stimuli can be subject to subtle biases that may influence results in ways that are not communicated and of which the experimenters may not be aware - thus compromising replication and scientific understanding.

Legal Protections

As advocated in the section entitled "Whose data is it?", users should retain ownership of their data and be free to license the data to multiple researchers. Examples of legal agreements that enshine this principle can be found on the unforgettable.me site (the user agreement - www.unforgettable.me/terms, the researcher agreement - www.unforgettable.me/researcher-terms and the privacy policy - www.unforgettable.me/privacy).

There are a couple of implications of the policy that researchers should take into consideration.

Any data that is generated by experiments belong to the users. Therefore they retain the right to license that data to other researchers. In order for another researcher to be able to do that,

however, they must be able to find the data. We encourage researchers to add a unique random key to the data that is posted by experimental code. This key can then be published with the corresponding article thus avoiding scooping while also ensuring that the data will be available for others to replicate analyses.

Another consequence of the policy is that users retain the right to delete data even after analyses have been conducted. In principle, this policy is already in force in most circumstances today. However, the difficulty involved in actually accessing the data means that the right is seldom exercised. If one implements more comprehensive search and visualization interfaces, data deletion will become easier and so it is likely that it will occur more regularly. While this aspect of the system may undermine the ability to reproduce analyses in some circumstances, it is a necessary evil if one is to genuinely implement privacy rights.

Discussion

Privacy and open science are on a collision course. The experience sampling techniques that promise to revolutionize the psychological sciences are also the techniques that are most invasive to privacy. Openly publishing this kind of data is not an option. We have proposed a solution that relies on collection mechanisms that provide the user with precise control over what is collected, search and visualization mechanisms that give users the tools to understand and delete their data, an analysis platform that allows researchers to conduct analyses without seeing the raw data, an experimental platform that allows users to respond to their own data without exposing it to researchers and a legal framework that cedes ownership of data to the users.

The long term impact of using experience sampling data to understand psychological processes could be transformative. In order to reach that goal though we need to reinvent our relationship to data to protect the partnership we share with our participants.

References

Anderson, J. G. (2007). Social, ethical and legal barriers to e-health. *International journal of medical informatics, 76*(5-6), 480-483.

Australian Digital Health Agency (2018). My Health Record Statistics - at 25 February 2018, https://myhealthrecord.gov.au/internet/mhr/publishing.nsf/Content/news-002.

Bonney, R., Shirk, J. L., Phillips, T. B., Wiggins, A., Ballard, H. L., Miller-Rushing, A. J., & Parrish, J. K. (2014). Next steps for citizen science. *Science, 343*(6178), 1436-1437.

Csikszentmihalyi, M., & Larson, R. (1992). Validity and reliability of the experience sampling method. In Marten W. de Vries (Ed.) *The experience of psychopathology: Investigating mental disorders in their natural settings*, (pp. 43-57). Cambridge, UK: Cambridge University Press.

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. Psychological review, 108(2), 452.

Dennis, S. J., Yim, H., Sreekumar, V., Evans, N. J., Garrett, P., & Sederberg, P. (2017). A hierarchical Bayesian model of “memory for when” based on experience sampling data. In G. Bunzelmann, A. Howes, T. Tenbrink, E. Davelaar (Eds.) *Proceedings of the 39th Annual Conference of the Cognitive Science Society*, (pp. 295-300). Austin, TX: Cognitive Science Society.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. *In Theory of cryptography*, pp. 265-284. Springer.

Dwork, C. and Nissim, K. (2004). Privacy-preserving datamining on vertically partitioned databases. In *Advances in Cryptology-CRYPTO 2004*, pp.528-544. Springer

Ebbinghaus, H. (1913). Memory (HA Ruger & CE Bussenius, Trans.). New York: Teachers College.(Original work published 1885), 39.

Foster, E. D., & Deardorff, A. (2017). Open science framework (OSF). *Journal of the Medical Library Association: JMLA, 105*(2), 203.

Gagnon, M. P., Payne-Gagnon, J., Breton, E., Fortin, J. P., Khoury, L., Dolovich, L., ... & Archer, N. (2016). Adoption of electronic personal health records in Canada: perceptions of stakeholders. *International journal of health policy and management, 5*(7), 425.

Kirilova, D. & Karcher, S., (2017). Rethinking Data Sharing and Human Participant Protection in Social Science Research: Applications from the Qualitative Realm. *Data Science Journal. 16*, p.43. DOI: http://doi.org/10.5334/dsj-2017-043

King, G. (2007). An introduction to the dataverse network as an infrastructure for data sharing. *Sociological Methods and Research 36*(2): 173-199.

Ludwick, D. A., & Doucette, J. (2009). Adopting electronic medical records in primary care: lessons learned from health information systems implementation experience in seven countries. *International journal of medical informatics, 78*(1), 22-31.

Narayanan, A., & Shmatikov, V. (2010). Myths and fallacies of personally identifiable information. *Communications of the ACM, 53*(6), 24-26.

Nielson, D. M., Smith, T. A., Sreekumar, V., Dennis, S., & Sederberg, P. B. (2015). Human hippocampus represents space and time during retrieval of real-world memories. *Proceedings of the National Academy of Sciences, 112*(35), 11078-11083.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... & Contestabile, M. (2015). Promoting an open research culture. *Science, 348*(6242), 1422-1425.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349* (6251).

Osth, A. F., & Dennis, S. (2015). Sources of interference in item and associative recognition memory. Psychological Review, 122(2), 260.

Rouder, J. N. (2016). The what, why, and how of born-open data. *Behavior research methods, 48*(3), 1062-1069.

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *The Annual Review of Clinical Psychology, 4*, 1-32.

Sreekumar, V., Dennis, S., Doxas, I., Zhuang, Y., & Belkin, M. (2014). The geometry and dynamics of lifelogs: discovering the organizational principles of human experience. *PloS one, 9*(5), e97166.

Sreekumar, V., Dennis, S., & Doxas, I. (2017). The episodic nature of experience: a dynamical systems analysis. Cognitive Science, 41(5), 1377-1393.

Taylor, C. L. (1985).The World Handbook Tradition: Producing Data for Cross-National Quantitative Analysis. Inter-university Consortium for Political and Social Research Bulletin.

White, H., Carrier, S., Thompson, A., Greenberg, J., & Scherle, R. (2008, September). The Dryad Data Repository: A Singapore Framework Metadata Architecture in a DSpace Environment. In *Dublin Core Conference* (pp. 157-162).