BEST PRACTICES FOR WORKING CREATIVELY WITH PERSONAL DATA

know
thyself
AS A VIRTUAL REALITY

know thyself / AS A VIRTUAL REALITY

Best Practices for Working Creatively with Personal Data

By Marilène Oliver, Alissa Hamilton, and Katrina Ingram, 2022

Acknowledgements

These guidelines have been informed by many conversations and discussions with all the Know Thyself as a Virtual Reality (KTVR) research team, invited speakers, and guests: Zach Blas (University of Toronto), Dana Dal Bo, Pierre Boulanger (Computer Science, U of A), Chelsey Campbell (Art & Design, U of A), J.R. Carpenter (Digital Literature, University of Southampton), Silvia Casini (Visual Culture, Science and Technology, University of Aberdeen), Craig Chapman (Kinesiology, U of A), Jonathan Cohn (Digital Humanities, U of A), Susan Cox (Research Ethics, University of British Columbia), Joelle Dietrick (Digital Studies, Davidson College), Daniel Evans (Art & Design, U of A), Gabor Fichtinger (Queen's University, U of A), Martin Ferguson-Pell (Biomedical Engineering, U of A), Jonathan Garfinkel (Medical Humanities, U of A), Gillian Harvey (Design Studies, U of A), Fahim Hassan (Public Heath, U of A), Katrina Ingram (AI Ethics, Ethically Aligned AI), Greg Kawchuck (Rehabilitative Medicine, U of A), Manal Kleib (Nursing, U of A), Ewen Lavoie (Kinesiology, U of A), Nathanial Maeda (CogPro, U of A), Lianne McTavish (History of Art, Design, and Visual Culture, U of A), Jerrold McGrath (UKAI, Toronto), Stephan Moore (Music, Northwestern University), Owen Mundy (Digital Studies, Davidson College), Remigius Nwabueze (Law, University of Southampton), Ubaka Ogbogu (Law, U of A), Walter Ostrander (Computer Engineering, U of A), Dhara Pancholi (Computer Science, U of A), Kumaradevan Punithakumar (Radiology and Diagnostic Imaging, U of A), Geoffrey Rockwell (Philosophy, U of A), Peter Seres (Peter S. Allen Scanner Centre), Scott Smallwood (Music, U of A), Kirtan Shah (Computer Science, U of A), and Richard Thompson (Biomedical Engineering, U of A).

Special and sincere thanks to reviewers Lianne McTavish, Jonathan Cohn, Nidhi Hedge, Jacob Jaremko, and Kumaradevan Punithakumar.

The KTVR research project has is supported by a Kule Institute for Advanced Studies grant and a Social Sciences and Humanities Research Council Insight Development Grant.

CONTENTS

Introduction 4

Existing Guidelines and Ethics Procedures for Artists and Visual Researchers 7

History of Privacy and Property Law 12

Current Data Protection Guidelines

KTVR Reflexive Questions 15

1. Anonymization 23
2. Provenance, Access, and Licencing 34
3. Consent 52
4. Participation, Authorship, and Dissemination 60
5. Data Subjects 65
6. Sustainability and Data Production 69
7. Artificial Intelligence 75

Conclusion, Limitations, and Future Work 82

Glossary 84

References

INTRODUCTION

My Data Body and *Your Data Body* are partner virtual reality (VR) projects created by a collaborative and interdisciplinary team based at the University of Alberta, Canada, as part of larger project called *Know Thyself as a Virtual Reality* (KTVR). *My Data Body* seeks to make visible and manipulable all the data humans now endlessly generate and are responsible for, while *Your Data Body* troubles how we interact with, and are equally responsible for, the data of others.

My Data Body has at its centre a high-resolution, volume rendered full body MR scan dataset that viewers can enter into and explore. Embedded into this semi-transparent virtual body are other data corpuses downloaded from Facebook and Google. These textual data corpuses are plotted into cross sections of the body. In the horizontal (axial) plane, Mac terminal data is plotted into bone, Google data into muscle, and Facebook data into fat. In the vertical plane are plotted data usage agreements and into the depth plane are theoretical texts about virtuality and privacy in the digital age. The viewer can pull out these cross sections and read them; once they let go, the cross sections float away but ultimately and uncontrollably return to the scanned body. Passwords and logins flow back and forth through veins and arteries, and hashtags pool in organs. Certain organs can be pulled out of the body and "drawn with": the heart leaves a trail of emojis and the brain a trail of login pop-up windows demanding usernames and passwords. The medically scanned, passive, obedient, semi-transparent body becomes a data processing site that can be pulled apart and (dis)organized. The whole body/data processing site finds itself at the centre of a data cloud generated from social media data.

Your Data Body is a partner project to *My Data Body*, made using a combination of open-source and donated datasets. This project focuses on issues of data privacy and ownership, playing on the etymology of the word *data*, meaning "given." The scan datasets, which are stored in a series of pods that the viewer can teleport between, can be picked up and moved around, resized and recoloured, inviting a playful stacking of the body parts to make a whole Frankenstein-like figure. Audio is attached to each body part, triggered as the viewer holds and manipulates it. Anonymized open-source datasets are accompanied by an automated voice that reads the study data and usage permissions published alongside the dataset, whereas donated datasets have a recording of the data subject reflecting on their relationship to their data. In addition, there are also two highly used open-source datasets: the Visible Human Project, from the American National Library of Medicine; and Melanix, which comes with the radiology software OsiriX. Both have an AI-generated chatbot attached to them with whom the user can "discuss" different issues relating to data ownership, privacy, and virtuality.

The making of these artworks has raised many practical and ethical questions about the use of sensitive personal data (such as medical scans) as artistic material and subject matter. Such questions range from those of access to data through to ownership of the data once it has been transformed into an artwork, and to what extent data can be manipulated and re-presented in the name of affective, socio-political, artistic research. These projects have highlighted the complexities of conducting creative research using personal data, especially with regards to the intersection of research ethics, data privacy, and rapidly emerging technology from an interdisciplinary perspective.

As the KTVR project is supported by public grants, seated in an academic institution, and involves medical data, it has benefitted from going through multiple review processes that would not normally be available to independent artists and creative researchers. We have found, however, that while ethics review processes prompt rigorous consideration of how consent, data privacy, and potential harms should be addressed during the research project, they have a tendency to view ethics narrowly and are unable to adequately address the ethical issues particular to research-creation, in particular those that involve emerging and complex technologies (Oliver 2021). Most ethics boards are not equipped to evaluate artists' proposals because they lack familiarity with research-creation processes (Cox et al. 2014) and their values and interests differ from those of the artistic community (Bolt 2016). Current policies use ethical guidelines that are meant for other disciplines and are ill-fitted or even antithetical to artistic practice (Bolt 2016). There is a need for ethical guidelines tailored specifically to creative work that will prompt artists to reflect research ethics throughout the design, research, and practice stages of their projects and provide review boards with a resource to consult when assessing a project.

The guidelines proposed here aim to provide guidance specific to the use of sensitive personal data in artistic practices, and they are intended to be used by the artistic and creative research community at large and also ethics boards. Although the guidelines focus primarily on medical scan datasets (which are technically a form of "sensitive personal data"), they are applicable to all personal datasets such as biometric data and data scraped from social media platforms. The medical scan dataset with its indexical link to the interior, intimate, privacy of the body is both an example and a metaphor with which to think through other forms of personal data.

These guidelines begin with a short review of existing visual research ethics, current data protection regulations, and legal frameworks. The guidelines then address a series of reflexive questions specific to the ethical and informed use of medical datasets with regard to anonymization, provenance, consent, different cultural understandings of data, authorship, and dissemination, data sustainability, and finally how AI (specifically machine learning) is applied to data. We use examples of artworks throughout as a way to contemplate and these complex, nuanced ethical questions, which differ according to situated circumstances.

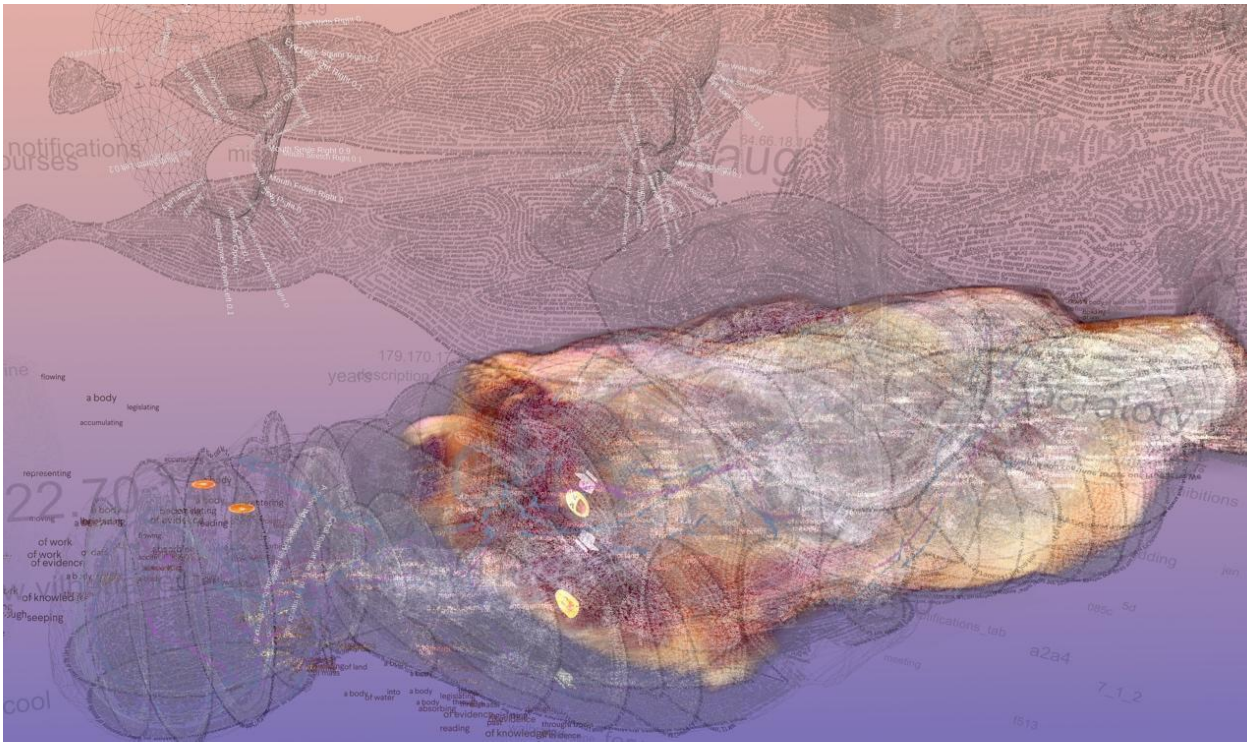


Figure 1. Marilène Oliver, Scott Smallwood, Stephen Moore and J.R. Carpenter. Screen capture of *My Data Body* 2021, VR artwork made as part of *Know Thyself as a Virtual Reality* project. Image courtesy of the artists.

Existing Guidelines and Ethics Procedures for Artists and Visual Researchers

Current data protection regulations in the United Kingdom, United States, European Union, and Canada exempt artistic practice from privacy requirements. This exemption extends to journalistic, artistic, academic, and literary purposes. Likewise, creative practice in Canadian academic research is also exempt from research ethics review. In Canada, there are three main academic granting councils: Canadian Institutes of Health Research (CIHR), the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Social Sciences and Humanities Research Council of Canada (SSHRC). These three councils are collectively known as the Tri-Council and have a collective ethical policy statement commonly known as the TCPS 2 that university Research Ethics Boards (REB) consult when reviewing ethics applications. Article 2.6 of the TCPS 2 states, "Creative practice activities, in and of themselves, do not require REB review. However, research that employs creative practice to obtain responses from participants that will be analyzed to answer a research question is subject to REB review."

The TCPS 2 then goes on to explain that "creative practice is a process through which an artist makes or interprets a work or works of art. It may also include a study of the process of how a work of art is generated. Creative practice activities do not require REB review, but they may be governed by ethical practices established within the cultural sector." Although there was a draft chapter for "Research Involving Creative Practices" written by the Social Sciences and Humanities Research Ethics Special Working Committee in 2008, it has not since been included in TCPS 2. In researching published guidelines for "ethical practices established in the cultural sector," we have found three key resources—the *Guidelines for Ethical Visual Research Methods* (Cox et al. 2014), *Mapping Responsible Conduct in the Uncharted Field of Research-Creation: A Scoping Review* (Voarino et al. 2019), and *Toolbox in RCRC : Synthesis of Responsible Conduct in Research-Creation Issues and Proposal of Reflective Tools* (Noury, Cloutier, and Roy 2018)—which we will briefly summarize in this section.[1] *Guidelines for Ethical Visual Research Methods* and the *Toolkit for RCRC* were developed to aid research-creation practitioners and ethics boards navigate the evaluation of the ethics of a research-creation project. The *Toolkit for RCRC* evolved from Voarino et al.'s scoping review of the literature on the ethical issues involved in research-creation; it focuses on research-creation in the Canadian and Quebecois context but is widely applicable.

Guidelines for Ethical Visual Research Methods

---

[1] Canadian Artists' Representation/Le Front des artistes canadiens's (CARFAC) 2021 *Indigenous Protocols for the Visual Arts*, which guide Indigenous and non-Indigenous communities, artists, and museums in the legal and ethical protocols surrounding the use of Indigenous art and culture, and aims to "encourage culturally appropriate working practices … [and] promote improved communications between all parties with an interest in Indigenous visual arts," is discussed in section 3.5.

Written by scholars from Canada and Australia, the *Guidelines for Ethical Visual Research Methods* (Cox et al. 2014) grew out of two research methods workshops. The guidelines focus on the ethics of research-participant relationships formed through the practice of visual research methods and provides six categories with guiding questions. The categories are confidentiality, consent, representation and audience(s), fuzzy boundaries, authorship and ownership, and minimizing harms.

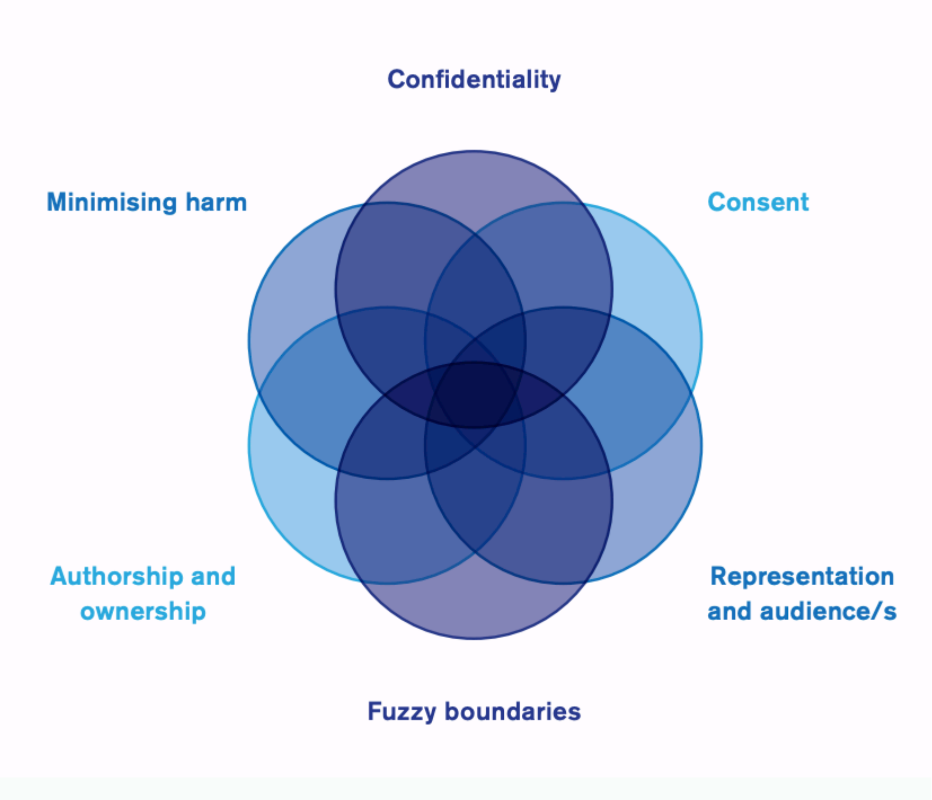


Figure 2. Six categories of ethical issues for visual research, from Cox et al., *Guidelines for Ethical Visual Research Methods*, 12. Image courtesy of the authors.

Confidentiality is "a commitment to protecting an individual's privacy when that individual has disclosed information in the context of a relationship of trust in research" (Cox et al. 2014, 9), though anonymity is not always an appropriate way to protect individuals, particularly if they are invested in how their work is perceived or conceptualized. **Consent** requires that "participants are enabled to make decisions about their research participation. [It] must be voluntary and based on sufficient information and adequate understanding of the purpose and aims of the research, what is required from participants, and any risks that are posed to them" (12). **Representation and audience(s)** addresses issues that arise in the dissemination of work, including addressing how participants perceive the way that their work has been edited, conceptualized, or disseminated, and the ways that the work are perceived by the viewing public (20). **Fuzzy boundaries** refers to how "boundaries between the roles of researchers, participants, artists, and others involved in the project can become blurred" and how "visual research methods may be used to serve multiple purposes, such as research, advocacy, and community engagement" (15). Negotiating these boundaries becomes problematic when "ethics committee procedures typically assume that the research is being conducted for a single purpose [with] clearly defined roles between researchers and participants" (15). **Authorship and ownership** addresses the issues over the authorship and ownership of work when visual research methods involve collaborative or group creation of visual work (17). **Minimizing**

harm relates to "avoiding physical and emotional harm; protecting research participants' anonymity and confidentiality; and promoting research that serves a public good" (7).

Research-creation can be powerful in a way that conventional research cannot since it "has an ability to convey emotion and to engender empathy" (Foster 2007, 372), and that power opens it up to ethical issues that are not present in conventional research. Research-creation can make participants vulnerable by exposing them to discomfort or to acceptable forms of harm that in conventional research may not be acceptable. Whereas in medicine the mantra is "do no harm," in research-creation, "the risk of harm or discomfort may be one of the aims of the work" (Bolt 2016, 196), and minimizing harms may compromise the quality of research-creation (Cox 2021). Rather than to avoid discomfort or harm, the emphasis should be to make participants aware of and prepared for potential discomfort or harm, and to provide them with access to resources (counsellors, support groups, etc.) to reduce the impact of the harm or discomfort.

The *Guidelines for Ethical Visual Research Methods* considers that, because research-creation is visual-based, it can violate standard research obligations of privacy since the dissemination of the work can identify a participant and link them with particular personal experiences—for example, if the likeness of the participant is represented in visual imagery, anonymity cannot be offered. Because of this lack of anonymity, it is possible a participant may choose to withdraw consent depending on how the impact of participation affects them. While academic research classically viewed consent as a one-time event, this perspective is evolving to an understanding of consent as a continuous process. Even more than in academic research, research-creation may require a continual renewing of consent throughout the life of the work and a process for how to accommodate the withdrawal of consent or how to meet changing needs for privacy.

Other categories of ethical issues are more unique to research-creation. Fuzzy boundaries arise when the roles of researchers and participants are less clearly defined—for instance, when participants become invested in the project in ways not seen in typical research, or when they feel invested in the relationships that have developed with the researchers (Cox et al. 2014, 15). In clinical practice or academic research, such fuzzy boundaries are generally strongly discouraged, unethical, or even illegal. In research-creation, they may, in some situations, be desirable and positive. In these cases, it is important for the roles and expectations of different actors to be well defined early in the work. When research-creation is co-created, issues of authorship and ownership can arise, from acknowledgement of participation, ownership of the final work, consent to use the final work, and the establishing of protocols to accommodate the withdrawal of consent (17). There are also issues with managing the expectations of participants with the goals of the project (categorized as representation and audience), providing adequate guidance to audiences to address or avoid problematic readings of the work, and considering how audience members—particularly if the work is disseminated online—may use or interact with the work in unexpected or problematic ways (19–20).

Scoping Review and Responsible Conduct in Research-Creation Toolkit

Voarino et al.'s 2019 scoping review, "Mapping Responsible Conduct in the Uncharted Field of Research-Creation," surveys the terrain of research-creation within Canadian academic institutions and laid the groundwork for the *RCRC Toolkit* (Noury, Cloutier, and Roy 2018). Focusing on the disconnect between academic institutions' ethical guidelines and the nature of research-creation, the scoping review aimed to bring research-creation and responsible conduct of research closer together. The review found that, because of this disconnect, the nature of research-creation and the position of the research-creation practitioner renders current ethical guidelines ill fitted to the practice of research-creation (Voarino et al. 2019). The review identifies six categories of issues that exist in the responsible conduct of research-creation. In descending order of frequency, the authors categorize the issues as those which arise (1) from the nature of the research-creation approach; (2) from the position that research-creation practitioners hold in relation to the work; (3) from the mismatch of academic training with the research-creation process; (4) from the funding of research-creation projects; (5) from the dissemination of research-creation knowledge; and (6) from the conflicts of authorship, commitment, and interests within research-creation production (Voarino et al. 2019).

Issues arising from the research-creation approach include conflicts between research-creation and academic research ethics and the difficulty of balancing the needs of participants (informed consent, minimizing harms, etc.) with the integrity of artistic expression. Defining responsibilities can be complex for research-creation practitioners as they often hold several roles both within a collaborative research creation project as well as multiple roles within their institution. Likewise, research participants can also hold multiples roles, particularly if they have expertise in the topic the work is engaging with; or when the work is creating new spaces (i.e., using novel technology) or providing a voice to a marginalized community. Issues arising from the mismatch of academic training with the research-creation process stem from the integration and acceptance of research-creation into academia and the accompanying need to develop appropriate student-supervisor relationships and ethics training that meet the specific needs of research-creation. Issues in the funding of research-creation projects arise from both a lack of funding for the arts and the need for research-creators to tailor their funding applications or their work to criteria that do not fit well the process of research-creation. Issues arising from the dissemination of research-creation knowledge include disseminating the work in unconventional ways or in ways that do not meet the standards of academic research, as well as issues of ownership and authorship in relation to collaborative work. Issues arising from conflicts of authorship, commitment, and interests within research-creation production can stem from the needs of participants and the needs of the research-creation practitioner, the multiple roles that research-creation practitioners can hold (as student or supervisor and artist, as a member of the artistic community and a member of an academic institution), and the potential conflict between the needs of each of their roles (Voarino et al. 2019).

Building on the six categories identified in the review by Voarino et al., the *RCRC Toolkit*, a 12-page document within the longer 130-page *Toolkit for RCRC*, provides a checklist for researchers and ethics boards to consider when designing and evaluating research-creation projects. The categories are issues arising when research-creation practitioners are not familiar with responsible conduct in research; identifying issues specific to research-creation; conflicts of interest and conflicts of commitment; dissemination and evaluation of research-creation, including applying for and managing funds and peer review of research-creation; familiarity with research ethics; and training and supervision of student practitioners of research-creation. Each category includes subsections that address the general issues within each category; for instance, the category for identifying specific issues in research-creation includes subsections that suggest taking a self-reflective look at practices, facilitating a culture of collaboration, and facilitating dialogue on issues surrounding RCRC (Noury, Cloutier, and Roy 2018, 63). Overall, the authors suggest "the adoption of a proactive and transparent attitude, as well as a continuous reflexive view of the research process, are key elements in preventing breaches of [responsible conduct in research]" (Noury, Cloutier, and Roy 2018, 12).

The checklist systematically addresses each of the six categories and provides a list of questions to ask to guide the research-creation practitioner through the research design process. The goal is to address the mismatch of ethical standards between research-creation and conventional research in Canadian institutions.

Together, these resources provide a basis of ethical guidance. However, in part due to the quickly evolving nature of digital technology, none of them adequately address evolving technology and the unanticipated impact that it could have on the work. The guidelines proposed here aim to provide further guidance in the use of sensitive personal data in an era of big data and vast digital ecosystems in order for artists to better anticipate and resolve issues in an ethical and transparent way.

History of Privacy and Property Law

Although artistic work is exempt from most data protection regulations, it is still important for creators using personal data in their works to understand what regulations exist and why. Data protection regulations provide the basic requirements upon which further ethical standards can be built. But regulations lag behind technological developments, since the process of developing new laws and regulations is incredibly slow. The rapid emergence of social media and smart technologies has created vast digital ecosystems of personal data that are readily available in a way that current laws are unable to address. The background provided below illustrates the regulations as they exist, but also highlights the gaps that our ethical guidelines seek to address.

Personal data as property is especially applicable to medical data, a subset of personal data. Medical data is typically collected by a health care authority and, in the Canadian context, is thus legally owned by the collecting or custodial organization. Medical data can sometimes be in the form of biological tissue samples taken from an individual in the course of diagnosis and treatment, and it is these biological tissue samples that can challenge the current legal definitions of what can be considered property, by whom, and in what situations.

The following discussion of how personal data is viewed through privacy and property laws draws on the keynote presentations of two legal experts, Dr. Remigius Nwabueze and Dr. Ubaka Ogbogu, at the Know Thyself as a Virtual Reality Ethics e-Symposium in May 2021. Their presentations trace the evolution of laws through case law, highlight the inconsistencies of the laws, and point to the gaps in the current legal framework.

Privacy Law and Images of the Deceased

Dr. Nwabueze used case law from the UK, Canada, and the US to explain how protection of the privacy of deceased individuals is neither consistent nor guaranteed. According to his research, many privacy laws do not extend to the dead (Nwabueze 2021). Likewise, until privacy laws were developed—in the UK this was not until 2000 under Article 8, Protection of Privacy—most legal actions did not address privacy for either the living or the dead, because there was no privacy law to which lawmakers could refer. Instead, cases were argued using other legal precedents. For example, in the case of Emiliano Sala, footage of the autopsy that followed his death in a plane crash was disseminated by individuals who had access to but no legal right to share the footage. Because privacy laws did not exist that could be applied to the case, the courts chose to convict them under the Computer Misuse Act (British Broadcasting Corporation 2019a).

The presentation focused on the publication of images of deceased individuals. Photographs are considered to contain personal information from which an individual could be identified because they contain the physical likeness—the appearance—of the person. Because of the nature of case law, court rulings can be contradictory. For instance, some courts have considered the privacy of and impact on the family in relation to the publication of images of the dead. In these cases, they extended protection to the family by protecting the privacy of the deceased person. For example, in the case of *Catsouras v. Department of California Highway Patrol* (2010), pictures taken by investigators of a naked woman after her death in an automobile accident were disseminated online. The courts ruled that because it was the family who ultimately suffered from the publication of these images, the privacy of the deceased should be extended to protect the privacy of the family. In other cases, protection was extended based on the conditions at the time personal data was collected. After the death of French president François Mitterrand, his medical doctor had planned to publish a book about the president's illness and death without the family's permission. In *Editions Plon v. Mitterand* (2005), the court ruled that because the personal information the book was based on had been collected while Mitterand was alive, it was protected by patient-doctor medical confidentiality (Nwabueze 2021, 24–25). However, the court also decided that since there were no privacy protections for the deceased, the publication itself would not violate the president's privacy, though it could not draw on any medical information collected while Mitterand was alive (Nwabueze 2021). Other rulings have simply stated that harm cannot be done to a deceased person because they are no longer alive. In *Smith v. City of Artesia* (1989), a case similar to that of *Catsouras*, members of the police department in Artesia, New Mexico, inappropriately circulated images of a naked woman's body taken during the investigation into her death. Unlike the *Catsouras* case, the courts ruled that since the victim was deceased, no harm could be done by violating her privacy

Property Law and Medical Data

Dr. Ogbogu's (2021) presentation focused on how personal data, in the form of medical data, is considered in terms of property law, highlighting how inconsistent and at times illogical the interpretation can be. In the medical field, personal data is complicated by the collection and use of biological samples—blood, bone, tissue, etc. The biological samples are not necessarily personal data in and of themselves, but they contain information about the individual from whom they were collected. The concept of property and data ownership is complicated when biological samples in research result in the development of drug therapies that have financial benefits to the companies or people that develop them. Issues arise when consent to use of the tissue has not been properly obtained from the tissue donor.

Using case law, Dr. Ogbogu detailed how ownership of collected medical data and individual donor rights have developed. Ownership revolves around the original intent for which the tissue was removed, and the alteration of human tissue via human ingenuity. The first, original intent, relates to consent and ownership of unaltered tissue and to which specific use or purpose the donor had consented at the time of removal. The second, alteration, is connected to the authorship and ownership of what is considered a "new" form simply by having been altered.[2] In the rather well-known case of *Moore v. Regents of University of California* (1990), tissue taken from Moore during treatment was used in research to develop cancer treatment drugs without his knowledge or permission. The courts ruled that although the patient has the right of control over tissue at the time of removal, unaltered biological tissue is not property and thus cannot be owned by anyone. Altered tissue, however, can be considered property of the person or institution that commits the alteration. Thus, the institution, not the patient, has property rights over the tissue (Ogbogu 2021, 31). However,

---

[2] As Dr. Ogbogu explained, even the idea of "altered" is contentious. Although the courts consider tissue to which human ingenuity has been applied as a "new" form, he questioned the logic of the argument when the "new" form is essentially derived from, not different from, the tissue from which it was taken.

in *Hecht v. Superior Court (Kane)* (1993) and in *York v. Jones* (1989), contract law determined the ownership of preserved human tissue (semen and zygotes, respectively) as that of the donor. But, depending on the jurisdiction, ownership can also be granted to collecting institutions while still maintaining privacy rights over the tissue to the donor, as in the case of *Piljak Estate v. Abraham* (2014). In this case, the family of the deceased made a claim of negligence against Dr. Abraham for having misdiagnosed a tumour as benign. An action was put forth by the doctor's defense to re-examine the tissue to reaffirm the diagnosis. The court ruled that the tissue was in essence a medical record, and thus the personal property of the collecting institution (Ogbogu 2021, 34–35). However, the donor maintains privacy rights over the tissue that must be protected. In *McInerney v. MacDonald* (1992), another case considering medical records, a doctor refused to provide a patient with health records that the doctor had not prepared herself, claiming that it was unethical for her to provide records created by other health professionals. The court found that there was no ethical issue in providing an individual with all of their health records since the records describe the individual and the data contained within those records belongs to the individual (Ogbogu 2021, 37).

Dr. Ogbogu's presentation outlined the inconsistent approach in the legal system to medical records, human biological tissue, consent to use of such tissue, and the ownership of that tissue. Ultimately, the ownership of and control over human biological tissue has not yet been clearly defined.

Current Regulations for Privacy and Consent

Privacy of personal information and consent to its use is regulated at the national level in most Western countries. These guidelines mostly focus on the Canadian context, but also look briefly at the US, EU, and UK. The European Union's 2018 General Data Protection Regulation (GDPR) is the most stringent set of privacy guidelines and laid the groundwork for regulations in other Western countries. Several levels of regulation inform the framework for ethics boards that institutions must implement to grant ethical clearance to research applications. Canada has the Privacy Act, Personal Information and Protection Electronic Documents Act (PIPEDA), and the Tri-Council Policy Statement (TCPS 2); in the US, regulations include the Health Information Portability and Accountability Act (HIPAA); and the UK has the UK GDPR. Beyond government regulation, privately and publicly held datasets may have terms of use and licensing agreements with additional layers of restrictions.

When preparing to work with medical (or any) datasets, artists need to be mindful of all the legal frameworks, to identify whether and what gaps exist in these frameworks, and, if the work is international, the differences in regional legislation. Legislation is geographically dependent, so if data is created in the UK but it is accessed in Canada, it is Canadian legislation that applies to the data.

In Canada, the Privacy Act (1985) protects "the privacy of individuals with respect to personal information about themselves held by a *government institution* and that provides individuals with a right of access to that information" (Government of Canada 2021). **PIPEDA** (2000) covers how businesses handle personal information. In 2020 the Government of Canada proposed the Digital Charter Implementation Act with the goal of increasing individual control of data and transparency regarding how companies use data, but the charter is yet to be passed.

The US Privacy Act (1974) "governs the collection, maintenance, use, and dissemination of information about individuals that is maintained in systems of records by *federal agencies*" (United States Department of Justice 2014). There are also other acts such as the Video Privacy and Protection Act (1988), which restricts audio visual usage; the Cable Communications Policy Act (1984); and HIPAA (1996), which protects sensitive patient health information. At present, there is no one act that regulates data privacy in the United States; rather, privacy data is governed state by state.

The European Union's GDPR (2018) protects "natural persons with regard to the processing of personal data and on the free movement of such data" and applies to *companies and public bodies* (European Parliament 2016).

In the UK, the Data Protection Act (2018) is a modified version of the EU's GDPR: **"**The GDPR, the applied GDPR and this Act protect individuals with regard to the processing of personal data" (Data

Protection Act, United Kingdom Legislation, 2018) and "controls how [an individual's] personal information is used by *organisations, businesses or the government*".

Personal Data

The definitions of personal data for each nation's legislation are fairly exhaustive. Broadly, **personal data** is defined as data related to an identifiable individual and includes demographic information—common identifiers such as name, address, date of birth, and identification numbers. Within personal data there are categories of data that are considered sensitive and require additional protections, including race, ethnicity, political opinions, and religious affiliation. **Sensitive data** is considered in the following section.

Exemptions from privacy legislation include **de-identified** or **anonymized data**, or when personal data is used for journalistic, artistic, academic, or literary purposes. The exemption to journalistic, artistic, academic, and literary purposes is related to freedom of speech, while the exemption of anonymous data assumes that it would be difficult to identify an individual from anonymous data.

Differences between legislation is less easy to generalize. Not all jurisdictions extend protection of privacy to the dead, and what is considered personal data and what is considered sensitive data is not consistent. When working internationally, it is recommended that the local regulations for use of personal data be consulted.

Sensitive data

There are some important legal caveats to the collection and use of personal data that is considered more sensitive. UK and EU GDPRs both have separate categories for sensitive data that requires additional precautions. UK GDPR specifies data as sensitive when it is related to an individual's race, ethnicity, political opinions, religious and philosophical beliefs, trade union membership, genetic, biometric, and health data, sex life, and sexual orientation; while GDPR's sensitive data classification includes data related to race, ethnicity, political affiliation, trade union membership, genetics, biometrics used for identification, and health data.

In Canada, Chapter 9 of the TCPS 2 specifically addresses the use of data originating from First Nations, Métis, and Inuit communities (Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, and Social Sciences, and Humanities Research Council of Canada 2019c). Use of this data requires that additional criteria be met, including actions addressing community engagement, recognition of cultural protocols, and respect for First Nations, Métis, and Inuit sovereignty. Additionally, secondary use of personal data—use of data that was not specified at the time of collection—originating from these communities is not exempt from ethical review.

Table 1: Comparison of Privacy Regulations across Western Countries

	PIPEDA	HIPAA	UK GDPR	GDPR

Data related to an identifiable individual	Factual or subjective info about an identifiable individual	Individually identifiable health info on physical/mental health condition, health care provided, or payment	Any data relating to an identified or identifiable living individual	Any data relating to an identified or identifiable living individual
Common identifiers	Name, address, DOB, ID numbers, ethnicity, blood type, employee files, credit and loan records, medical records, intentions, opinions, evaluations, comments, social status, disciplinary actions	Common identifiers that could identify the individual: name, address, DOB, SSN, etc.	Name, ID numbers, location, online identifiers, physical, physiological, genetic, mental, economic, cultural or social identity of the individual	Name, ID numbers, location, online identifiers, physical, physiological, genetic, mental, economic, cultural or social identity of the individual
SENSITIVE INFORMATION				
Data that could be used to discriminate, requiring additional protections	None listed	None listed	Race, ethnicity, political opinions, religious or philosophical beliefs, trade union membership, genetic, biometric, and health data, sex life and sexual orientation	Race, ethnicity, political opinions, religious or philosophical beliefs, trade union membership, genetic, biometric, and health data,
EXEMPTIONS				
Journalistic, artistic, or literary purposes De-identified or anonymized data	Journalistic, artistic, or literary purposes	Employment and educational records, deceased persons 50 years after death	Journalistic, artistic, or literary purposes, deceased persons, anonymized data	Journalistic, artistic, or literary purposes, anonymized data

Biometric Data

Biometric data is a category of sensitive data that includes biologically derived information (voice, fingerprints, iris scans, DNA, and genetic information) that is used to verify an individual's identity. In Canada, this type of data is subject to the Privacy Act and PIPEDA. The Office of the Privacy

Commissioner of Canada (2018) notes some of the unique privacy challenges that biometric data pose, including covert collection, cross-matching with other data to identify an individual, and secondary information extraction from the data that was not consented to at the time of collection. For example, an iris scan used for identification may also yield medical data and disclose health-related information about a person when the individual has only consented to use of the scan for identity verification.

To avoid unethical use of biometric data, the Canadian government suggests that biometric data be used "for verification rather than identification," as a best practice for avoiding the inappropriate use of data. Other jurisdictions are creating special regulations to provide further protections. In the United States, Illinois, for example, enacted the Biometric Information Privacy Act in 2008, which outlines requirements for written notice of collection, obtaining written consent, and standards for how biometric data is to be handled (Illinois General Assembly 2010).

Exemptions

As explained earlier, journalistic and artistic/creative work is exempt from most data protection regulations. Another common exemption is the secondary use of anonymized or de-identified data. Returning to Dr. Nwabueze's research on privacy law and images of the dead, the GDPR also exempts data from deceased persons (with no specification of time after death), while HIPAA specifies exemptions fifty years after death.

Consent

In order to use personal data, consent must be obtained from the individuals to whom the data relates. The GDPR and UK GDPR stipulate in Article 4.11 that consent must be "a clear affirmative action" that is freely given, specific, informed, and unambiguous. Consent must be obtained separately for each purpose (Radley-Gardner, Beale, and Zimmermann 2016; Information Commissioner's Office 2021). GDPR, HIPAA, and PIPEDA require consent forms to be written in plain language that is easy to understand (Office of the Privacy Commissioner of Canada 2018b; Radley-Gardner, Beale, and Zimmermann 2016), and all four policies require consent forms to include the purpose of collection and a process for withdrawing consent.

Requirements for consent are otherwise inconsistent. PIPEDA does not explicitly require either consent forms or for the entity collecting the data to be identified (Office of the Privacy Commissioner of Canada 2018a), and GDPR does not require disclosure of who will have access to the data. Only PIPEDA and

	PIPEDA	HIPAA	UK GDPR	GDPR
Informed, specific and freely given	Reasonable to expect people will understand the nature, purpose, and consequences of the collection, use, or disclosure of their personal information	Specific and freely given	Freely given, specific, and informed	Freely given, specific, informed and unambiguous
Must have a legitimate purpose for collection	Must be necessary for an explicitly and legitimate purpose		Must be necessary. For research purposes, can be general.	Must have a specific, explicit and legitimate purpose for processing. For research purposes, can be more general
What consent agreement must define	WHAT is being collected, WHO has access, PURPOSE of collections, RISKS of participation	WHAT is being collected, WHO has access, PROCESSING activities	WHO has access, PURPOSE of collection, PROCESSING activities	WHAT is collected, WHO has access, PURPOSE of collection, PROCESSING activities, use of AUTOMATED decision making
Right to withdraw consent	Right to withdraw consent	Right to withdraw consent, must give an expiry date	Right to withdraw consent, cease processing from the time of withdrawal, no specific expiry date	Right to withdraw consent, cease processing from the time of withdrawal, delete personal data "if wished to continue to use the data"

Table 2: Comparison of Consent Regulations across Western Countries

HIPPA require what is being collected to be specified (Office of the Privacy Commissioner of Canada 2018a; United States Department of Justice 2020), while both EU and UK GDPR include how the data will be used or processed and require separate consent for separate purposes (European Parliament 2016; Information Commissioner's Office 2021). Only GDPR requires consent to be necessary, and while UK GDPR and HIPAA address whether consent can expire, HIPAA requires an expiry date (Alder 2021) and UK GDPR specifically indicates that an expiration is not required (Information Commissioner's Office 2021). PIPEDA also includes several unique stipulations that have been recently introduced. PIPEDA requires the provision of information on the potential risks and harms of participation, the provision of different levels of detailed information at the data subject's request, that the reasonable expectations of the data subject be met, and that consent be dynamic and ongoing, renewed when there are significant changes to the use of data. Additionally, Canada's Digital Charter Implementation Act, proposed in 2020 and still in development in 2022, intends to modernize and simplify meaningful consent, develop a process by which individuals can withdraw their consent and request that their data be disposed of, require that businesses make their algorithms transparent, and strengthen the protection of anonymous data (Government of Canada 2020).

Issues of consent are further complicated if someone is deceased. In those cases, a personal representative for the estate or next of kin may be able to provide consent depending on the specifics of the request and situational context.

Additional Data Regulations in Academic Research

In Canada, use of personal data is further regulated by the Tri-Council Policy Statement (TCPS 2). Updated in 2018, the TCPS 2 stipulates that any research involving human or human biological material must undergo ethical review. The core principles around which the policy is structured are respect for persons, concern for welfare, and justice. However, research is exempt from ethics review when the data it uses is either "publicly available through a mechanism set out by legislation or regulation and that is protected by law or in the public domain and the individuals to whom the information refers have no reasonable expectation of privacy". Data that falls into the first category of publicly available data includes data that is found on government websites such as demographic statistics or vital statistics, open-access datasets (e.g., OpenNeuro), or private datasets that are available through subscription. Data that falls into the second category of data in the public domain includes data found on public (i.e., open or not set to private) social media profiles or "found" data such as digital or physical images. Copyright, intellectual property rights, and dissemination restrictions may still apply to publicly available data, and private datasets available through subscriptions also come with terms of use that dictate further restrictions on how the data may be used.

Research is also exempt from ethics review if it involves the "secondary use of anonymous information, or anonymous human biological materials, so long as the process of data linkage or recording or dissemination of results does not generate identifiable information" (TCPS2 2018). **Secondary use of data** is when data is used for a purpose other than that which was specified at the time of collection; **anonymous data** is data that has had some of the data points removed so that the individual from which the data was derived cannot theoretically be identified. Most publicly accessible datasets would fall under this exemption, with particular attention paid to how the use of the data could compromise the anonymity of the data subjects. In the age of social media, smart devices, and the vast amount of data they produce, the standard for anonymization has come under scrutiny since the availability of personal data can now compromise anonymization methods. The threat lies in both the wide range of information that can be inferred through location-tracking data (Baron and Musolesi 2020) and the ease with which individuals in anonymous datasets can be re-identified by cross-referencing demographic information (Rocher, Hendrickx, and de Montjoye 2019). As the TCPS 2 notes, "Rapid technological advances facilitate identification of information and make it harder to achieve anonymity" (TCPS2 2018, 18). The TCPS 2 does not, however, provide any guidance on how to avoid or navigate potential ethical problems.

KTVR Reflexive Questions

The remainder of the guidelines take the form of reflexive questions aimed to assist artists in making informed decisions about ethical and privacy issues when working with personal data. These guidelines are meant to help artists prevent, mitigate, or avoid issues as they relate to the use of personal data in creative works, in particular sensitive data such as medical scans. We've organized our guidelines into the following subsections:

Anonymization

Is the data really anonymous?

Provenance, Access, and Licencing

Where does the data come from? How has the data been processed? How can you access the data? What do you have to do to obtain access to the data?

Consent

Who gave you permission to use the data? Who made the data available? What kind of permission to use the data was provided, if any, and on what grounds was it provided?

Participation, Authorship, and Dissemination

Are those who provided the data either participants or collaborators in its accessibility and dissemination? Are data subjects participants, contributors, collaborators, or authors?

Data Subjects

How do different cultures understand sensitive personal data such as medical and biometric data?

Sustainability and Data Production

Is it essential to acquire new data, or can existing data be used to create an artwork?

Artificial Intelligence

Is AI used in the art project; if so, what kind? Is the artist in control of what the AI is doing with the data?

Anonymization

As explained in the previous section, data that has been de-identified or anonymized is exempt from privacy laws. In academic research, secondary use of data—use that is different from the original purpose for which the data was collected—is exempt from ethics review when it is applied to anonymous datasets (TCPS 2 2018). Consent is not required for use of anonymized datasets since personal information has either been removed or obscured in a way that re-identification is considered difficult. However, anonymization has become less reliable in an age of big data, smart devices, and social media. It is important to understand that anonymized datasets no longer offer the protections they once did. Smart devices and social media make a wealth of information publicly available (Cooper and Coetzee 2020; Parks 2021) and this big data can undermine the methods for protecting human subjects represented in anonymous datasets (Rocher, Hendrickx, and de Montjoye 2019). Furthermore, as machine learning algorithms work by finding patterns in data, there is no assurance (or even way of knowing) if anonymized datasets are cross-referenced and thus re-identified.

Thus, anonymous datasets that can be legally and ethically used in research can now breach a core tenet of ethical research because of the capabilities of machine learning, facial recognition, and big data (Hassan 2021). The potential for harm to the re-identified data subject is not equal for all data subjects and depends on the kind of data, its age, and a myriad of other local social, legal, political, and economic factors. Disclosure of health status in some parts of the world could prevent someone from accessing health insurance and thus health care in some countries; it could also impact them socially and/or economically by inviting discrimination and stigmatization. The possibility of re-identification is particularly troubling when we consider that social media giants such as Google and Facebook are establishing lines of business in the health care domain. During the KTVR e-Symposium in May 2021, Katrina Ingram and Fahim Hassan presented a workshop about the risks of aggregating health data, with a special focus on Google (Hassan 2021; Ingram 2021). The possibilities to aggregate and share data across platforms are pushing the boundaries for new business models that create a myriad of risks. For example, 23andMe is a Google venture (via Alphabet) that became a publicly traded company in early 2021 and is now using the DNA from millions of Americans to produce pharmaceuticals (Brown 2021). 23andMe is part of a Google health portfolio that includes insurance companies, medical record apps, and home health monitoring technologies that collect biometric data. Google's Project Nightingale, which gave Google access to health care data through research partnerships, has already raised privacy concerns and lawsuits (Schneble, Elger, and Shaw 2020).

In his discussion of the inadequacy of de-identification to protect the identity of data subjects, legal scholar Mark Rothstein notes four categories of potential harm: group harm, objectionable uses, commercial exploitation, and undermining trust (2010, 6–8). Group harm refers to linking specific populations to stigmatized conditions. Objectionable use refers to the use of de-identified data in ways that the data subject had not anticipated and did not consent to. Commercial exploitation refers to monetary gain that has not been shared with the data subject. Undermining trust refers to the potential long-term impacts on the artist and affiliated organizations that could be difficult to recover from if the public senses that there has been unethical conduct in the use of data. An example undermining trust is the infamous case of Henrietta Lacks (Skloot 2010). Lacks was a cancer patient who unwittingly provided the source material for an important medical breakthrough known as the HeLa cell line. These cells were the first to survive in a lab environment for more than a few days, which made them extremely valuable for medical research. For many decades, neither Lacks, who died soon after her cells were harvested, nor her family received any compensation. It was only after over fifty years of fighting and intense media attention that a donation was finally made by one of the biomedical research companies, the Howard Hughes Medical Institute, to the Henrietta Lacks Foundation (Witze 2020).

Metadata

In addition to the primary content of a data file, there can also be metadata associated with the data. **Metadata** is data about data. For example, when you take a photograph, information about the type of camera used, the location of where the photograph was taken (GPS coordinates), a time stamp, size of the file, and other information is automatically recorded by the camera. Digital technologies leave a digital footprint, which becomes another form of data that may also need to be anonymized to protect a data subject's privacy.

To evaluate the risks associated with working with anonymized data, it is important to understand the main ways in which data privacy is enacted through privacy-preserving methods of anonymization, pseudonymization, and de-identification. Understanding these methods can also help shed light on the issue of re-identification.

How is data anonymized?

There are different ways in which datasets can be anonymized. The most common ways are **de-identification**, **pseudonymization**, and data-cloaking techniques such as **data scrambling** and **defacing**. For big datasets, there are also statistical privacy-preserving methods such as **generalization**, perturbation, and **randomization**. These methods are not as applicable to medical scans data but may be pertinent to other datasets used by creators. The issue of re-identification is also related to data anonymization and is an increasing concern as datasets are reused and combined in novel ways that present new risks to privacy.

De-identification

De-identification refers to the removal of personally identifying information such as name, address, or unique identifiers (e.g., health care number) from a dataset. Although true de-identification or anonymization is very difficult, this technique should mean that there is no way to relink the data to the data subject.

In medical scan images, direct-identifying data is often found as metadata in the header of the file (not visible in the scans). Radiologists recommend either a file conversion or removal of the header information that contains personal identifying information as standard methods to de-identify medical scans (Parker et al. 2021). However, there are limitations to both methods. File conversion can result in the loss of data, while removal of the header information may be insufficient if the vendor's software systems retain identifiers. Historically, removing personal information had been adequate in protecting the privacy of data subjects. However, recent studies have shown that with the development of big data, data subjects in datasets with fifteen or more demographic data points can be reidentified (Rocher, Hendrickx, and de Montjoye 2019), and in fact, studies dating back to the early 2000s have pointed to issues with the standards for de-identification of personal data (Rothstein 2010, 5–6). Chapter 5 of Canada's Tri-Council Policy Statement notes that "technological developments have increased the ability to access, store and analyze large volumes of data. These activities may heighten risks of re-identification…Various factors can affect the risks of re-identification, and researchers and REBs should be vigilant in their efforts to recognize and reduce these risks" (TCPS 2 2018, 59). As Rothstein suggests, "responsible researchers should consider whether, in the context of their particular research, additional measures are needed to protect de-identified health information and biological specimens and demonstrate respect for the individuals from whom the information and specimens were obtained. Those who engage in research ought to be as thoughtful and meticulous about their relations with the human subjects of their research as they are about designing their experiments and analyzing their data" (2010, 9).

Pseudonymization

Pseudonymized data has had identifying information removed from the dataset and replaced by a random or artificial identifier or pseudonym (fake name). Pairing the identifier or pseudonym and a key allows for the re-identification of the data subject in the dataset (University College London 2019). Pseudonymization is very common across medical research, probably more than plain de-identification, because the master key is needed to track outcomes.

Data scrambling

Data scrambling, or image cloaking, is another method increasingly used in an attempt to anonymize image data as a way to avoid unauthorized facial recognition. The software Fawkes from the University of Chicago (Shan, Wenger, and Zhang 2020), for example, is a system that adds "imperceptible pixel-level changes" or "cloaks" to digital images that are imperceptible to the human eye but stop them from being identified by facial recognition models. What this is essentially doing is adding noise to the data so that it interrupts it being read, analyzed, and potentially re-identified by a machine.

Generalization

Generalization (also known as data blurring) involves making specific attributes of the dataset more broadly characterized. A very simple example might be replacing a person's exact age with an age range. For example, instead of saying someone is forty-seven years old, the dataset might indicate they are between the ages of forty and fifty years old. This technique is known more specifically as **binning**. Another generalization technique called **shortening**, involves reducing the amount of information. Examples include going from six-character postal code to the first three characters, or a geographical generalization such as stating province instead of city.

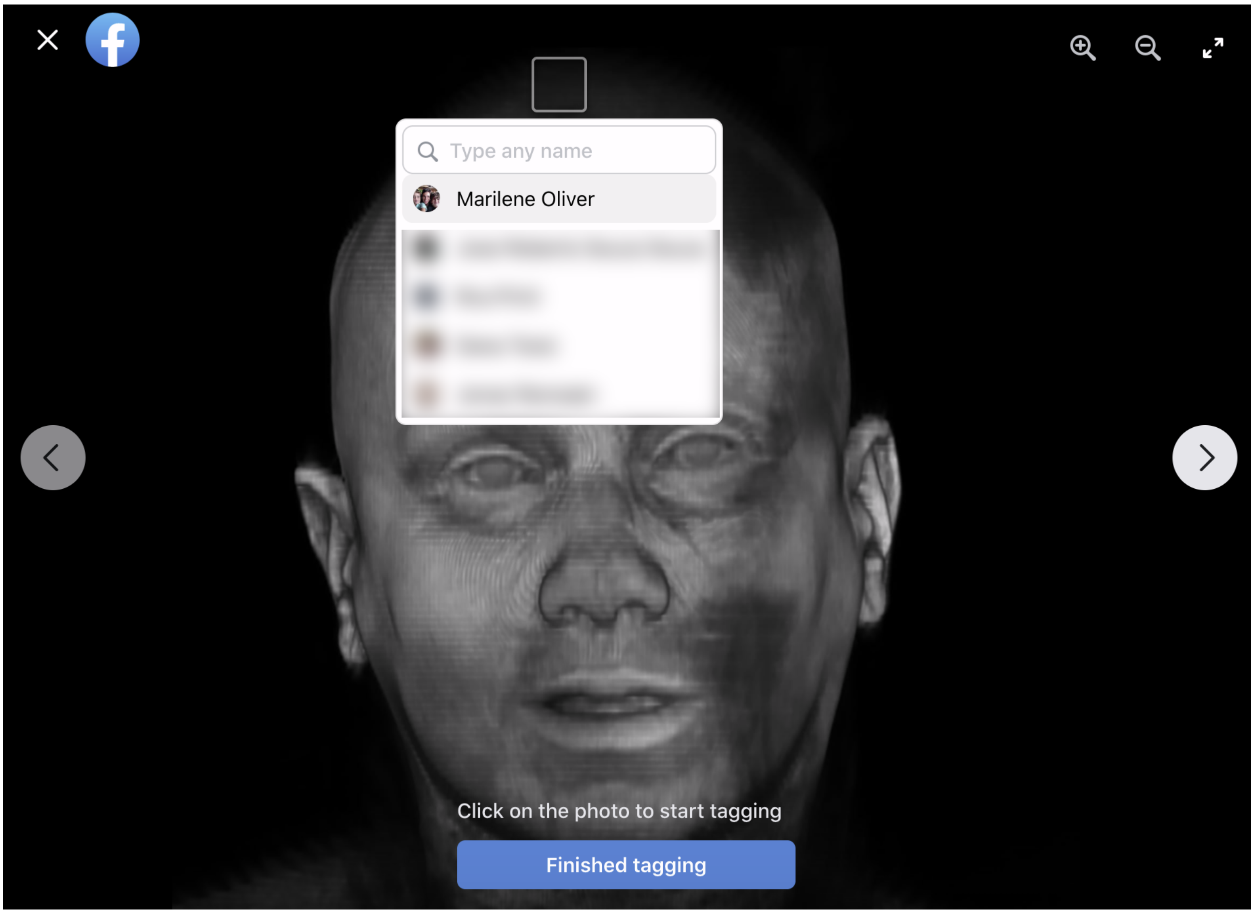
Randomization

Randomization does not strictly fall under anonymization, but more generally under privacy preserving and is often used as an alternative to anonymization. Randomization involves altering the relationship between variables in the data so they are less likely to identify an individual. This technique is often referred to as "adding noise" to a dataset or "perturbing" the data. Randomizing techniques used in machine learning with big data include differential privacy, permutation, and substitution.

Typically, there are trade-offs between improving privacy and retaining utility of the data. Striking the right balance in determining which technique(s) to use and what trade-offs are acceptable is part of the ethical choice involved in applying these privacy-preservation methods.

When Marilène Oliver uploaded a 3D rendering of her MR scan data to Facebook, it was tagged instantly as belonging to her. Facial recognition software has the ability to identify people from their medical scans (Parker et al. 2021). In their white paper on the de-identification of medical imaging, Parker et al. (2021) explain that as well as being able to capture facial identity, medical scans can also retain identifying features both beneath and above the skin's surface such as moles, fillings, pins, and hip and knee replacements. The anonymized Melanix dataset, which is included with the **DICOM** viewing software OsiriX, has several such possible identifiers such as artifacts from dental fillings, a mole, and even the indentation of a wedding ring. The Melanix dataset, which has been technically been de-identified and pseudonymized, is thus an example of a dataset that has the potential of being re-identified.

The risk of re-identification with facial recognition software has led to the creation of **defacing** software. This technique automatically identifies and crops away the faces from head scans. For diagnostic purposes, this method is reported to work well, but in terms of aesthetics it seems disturbingly aggressive to the human figure and dehumanizing of it.



Click on the photo to start tagging

Figure 3. Screenshot of Facebook automatically identifying a 3D rendering on an MR scan
Image courtesy of Marilène Oliver

Blurring

Blurring or pixilation has historically been a common way of rendering an individual in a photograph anonymous. However, this technique has been criticized as inappropriate because it dehumanizes the individual (Nutbrown 2011, 8). This sentiment is echoed in the *Guidelines for Ethical Visual Research Methods* where the authors acknowledge that blurring "reduces the authenticity" of the visual and/or risks dehumanizing participants, or denies participants the ability or right to make an informed choice about revealing their identity (Cox et al. 2014, 20).

Blurring human faces has been used by several artists as a way to memorialize and respect their subjects. For *Monument (Odessa)*, the French artist Christian Boltanski worked with an image of a group of Jewish girls in France from 1939. By altering their likenesses through re-photographing, enlarging, and editing the photos, the identities of the individuals were lost, but references to their humanity and youth were retained. The images were further manipulated through the mode of their presentation—the altered images were installed in dark rooms, alongside glowing lights, and with reflective surfaces. The artwork was thus able to use the likeness of the children without infringing on their privacy and while reinforcing a memorial aspect (Boltanski 1989). In another example, Canadian artist Sandra Brewster used the technique of blurring to empower her subjects and to explore the "layered experiences of identity—ones that may bridge relationships to Canada and elsewhere, as well as to the present and the past" (Brewster 2019). Brewster achieved the blur in her 2017 series of photographs *Blur* by directing her subjects to move, to "evoke the self in motion, embodying time and space, and channeling cultures and stories passed down from generation to generation." When asked why she decided to title her solo exhibition of 2019 *Blur*, Brewster explained,

> "Blur" plays with and was inspired by all of the interpretations I mentioned: [how the works] explored movement and referenced migration and how the effects of migration may influence and inspire the formation of one's identity here—whether the person was born elsewhere or is the child of a person born somewhere else. The intention of the blur is also to represent individuals as layered and complex: to not see people solely in one dimension; [and to be] aware that a person is made up of [both] who they are tangibly, and so many other intangible things, which includes their experiences with time and location—whether they access this on their own or through generational storytelling. (Price 2019)

In her long-term project *self-Less*, artist Dana Dal Bo collects hundreds of screen-captured selfies (naked people taking their own picture in bathroom mirrors with smartphones) and manually renders them unrecognizable through digital photo manipulations techniques such as painting, erasing, and pixel cloning (where one part of the image is copied to another). The manipulated images are then re-posted to a dedicated Instagram feed (Witze 2020). Dal Bo also made a series of prints of the manipulated *self-Less* images using the antique and resistant technique of carbon printing (Dal Bo 2017). The title of the series, *Carbon Copy*, refers both to the printing technique and to the "CC" that is now used in emails as a way point to the permanence, reproducibility, and recursion of images posted online. In her 2021 artist presentation "ArtAIfacts: Co-Creation with Non-Human Systems," Dal Bo reflected that although the original images were clearly intimate and meant for a specific person, once they were posted online they became available to anyone, everywhere at any time.

Figure 4. Dana Dal Bo, *Ass sink* from *self-Less* series, 2014–ongoing, digital image, dimensions variable. Image courtesy of the artist.

Figure 5. Zach Blas, *Fag Face Mask*, 2012, vacuum formed, painted plastic.
Image courtesy of the artist.

A growing number of artists are highlighting the dangers of facial recognition technology and speculating solutions to evade it. In his 2012 project *Facial Weaponization Suite*, American artist Zack Blas created a series of four amorphous vacuum-formed plastic masks based on aggregated of facial data of marginalized groups (queer, black, female, and Mexican). When the "collective" masks are worn, the wearer is protected from facial recognition systems. *Fag Face Mask*, for example, was generated from the biometric facial data of queer men's faces. In an essay on the project, Blas explains that the mask is a "a response to gay face and fag face scientific studies that link the successful determination of sexual orientation through rapid facial recognition techniques…The mask is not a denial of sexuality nor a return to the closet; rather, it is a collective and autonomous self-determination of sexuality, a styling and imprinting of the face that evades identificatory regulation" (2013). Each mask in the series in motivated by instances of social and political abuse of facial recognition software, such as the failings of the technology to detect Black faces, veil legislation in France targeted at Islamic women, and the abuse of biometric surveillance at the Mexico-US border (Blas 2012).

American and German artist and software developer Adam Harvey also works with facial recognition software to raise awareness of the existence of large, publicly gathered facial recognition training datasets and their accuracy, and to develop anonymization tools to protect individuals from exposing their or other's faces to facial recognition software when posting images to social medial platforms. In his 2017 work *MegaPixels*, Harvey created a photobooth in which audience members could have their faces matched to a face within the MegaFace (V2) dataset, at the time the largest publicly available and widely used facial recognition training dataset (Harvey 2017). *MegaPixels* produces an image of the participant's face next to the image of the face it has been identified as, plus the accuracy of the identification that can be thermally printed and taken by the audience member. The MegaFace (V2) dataset, which contains 4.2 million images, was created from Flickr images without consent and is being used in research projects in the US, China, and Russia to train facial recognition systems.

A more recent work by Harvey, DFACE.app is a web-based application that masks or redacts faces in photographs. There are several different "redaction effects," including blur, mosaic, emoji, fuzzy and colour fill. Although these effects are fun and playful, the DFACE.app is motivated by the unwarranted use of surveillance technologies in protests and large public gatherings (Harvey 2018).

Automatic, private, open-source face redaction by VFRAME.io

REDACTION EFFECT: Emoji

EMOJI THEME: Faces

MINIMUM CONFIDENCE: 78%

Figure 6. Example of image “dfaced” with Adam Harvey’s DFACE app.

Anonymization Discussion Questions

- Has identifying information been removed data? Has identifying information been destroyed, or has the link between the identifying data and the dataset been destroyed? Is it possible for someone to relink data?
- Does the scan data contain information such as facial structures that facial recognition software could identify? Is it appropriate to render those features unidentifiable? What methods of censoring or scrambling the data are appropriate? In the past, black bars were used to obscure parts of the face or body, but this strategy can be perceived as a violent way to anonymize participants.
- Will the way in which the artwork is disseminated put the data subjects at risk of re-identification? Is social media being used in any part of the work? What kinds of algorithms does social media expose the data to? Should images of the artwork be uploaded to social media platforms or not? Should the data be rendered illegible to algorithms using cloaking software such as Fawkes?
- When the dataset was created, how have (or could) changes in technology challenge or undermine privacy or consent? Could demographic data be used to re-identify the individual?

Provenance, Access, and Licencing

There are many different ways of accessing scan data, whether it be our own existing or newly acquired data, or a dataset that is freely downloadable from the internet (see Tools section for a list of open-access databases).

The question of data ownership in the context of medical scan data (and other kinds of personal data), however, is complex. Data subjects are typically asked to consent to their data being used for research purposes when data is acquired, effectively transferring "ownership" of the data to the researcher. Even if consent is not given at the time of collection, medical data does not necessarily "belong" to the data subject; rather, it belongs to the researchers who, or institution which, acquired the data.

It is likely that it will be impossible for a researcher to contact the original data subject because it has been anonymized or de-identified. It is, however, possible for creative researchers or artists to contact the ethics board of the institution to ask for permission to use the data for artistic purposes, if it is not freely available for research. Even if the data is freely available, there is no harm contacting the original researcher to confirm they are comfortable with an experimental creative research use of the data. Indeed, this can be a way of creating connections or establishing fruitful relationships with other researchers that can lead to future collaborations.

Even when permissions are granted and access to and use of data may therefore be legal, the datasets themselves should also be considered from an ethical standpoint. Who is the subject of the dataset? How were they recruited for participation? Where were the participants located and what were the laws governing their privacy and consent at the time of data collection? What are the differences in privacy and consent regulations in the time and place that the data is being reused? It is not the artist's responsibility to correct the ethical mistakes involved in the creation of a dataset. Indeed, as we discussed in the last chapter through the work of artists Adam Harvey and Zach Blas, the presence of ethical issues in a dataset may be integral to the work itself.

Working with your own data or the data of close relatives

Depending on where you live in the world and whether you are part of a public or private health care system, getting access to your own medical data is more or less complicated. The GDPR, under Article 15, gives individuals the right to request a copy of any of their personal data. Likewise, Canadian law stipulates that individuals are entitled to request a copy of their own data. The process of actually getting a copy of scan data, however, is less standardized. In France, patients who are scanned for diagnostic purposes automatically receive a copy of the scans on a disc, whereas in Canada a copy needs to be requested each time and there may be a fee associated with the provision of the copy.

There are many examples of artists who have worked with their own medical scan data, which was originally produced for diagnostic purposes and then acquired by them to create artworks. British film maker Victoria Mapplebeck used her own medical scans as part of her VR film *Waiting Room*. Mapplebeck (2019) writes that she used "CT scans, mammograms and ultrasound to provide a 3D portrait of my body from the inside, out" as a way to tell the story of her breast cancer experience from its diagnosis through its treatment and to her recovery. *Waiting Room VR* layers and combines mobile phone footage, head-mounted GoPro footage, precisely placed voice recordings and CGI models to make the viewer feel they are a fly on the wall at every stage of Mapplebeck's breast cancer experience (British Broadcasting Corporation 2019b). The inclusion of the scan data in the *Waiting Room* alongside film footage adds another layer of intimacy to her work, literally inviting the viewer into Mapplebeck's body and to see her cancer.

There are also cases of artists working with medical researchers specifically to create artworks. British artist Jane Prophet, for example, in 2014 worked with neuroscientists Zoran Josipovic from NYU and Joshua Skewes from Aarhus University to create *Neuro Memento Mori*, a sculpture and video mapped projection generated from functional MRI data (scans that record brain activity) acquired as Prophet was looking at representations of memento mori and vanitas paintings as well as meditating/contemplating death (Prophet 2014). In her February 28, 2016, blog post where she reflects on the experience of being scanned for the *Memento Mori* project, Prophet writes about her anxiety before the scan of being able to hold still for the duration of the scan, explaining that she had always been a "fidget," even as a child.

> When I was told I had to be very still for a series of seven minute MRI scans I was worried. I was a kid who, when bad dreams sent me running to my parents' bedroom, kept my mum awake all night as I kicked and wriggled while "peacefully" asleep. That's the kid that grew into a woman who nightly and tosses and turns. How could I possibly stay still in the scanner? Let along [*sic*] REALLY still. Not even swallowing…Ironically, given the experiments we performed, I needed to still my body, to "play dead" in order to prevent micro movements. (Prophet 2016)

French artist Marc Didou uses both the data acquired and the experience of being scanned to create large steel, marble sculptures. For Didou, it is crucial that he is the scanned subject because the experience allows him to see and feel something entirely new. In an interview with Silvia Cassini, Didou explains that "an MRI scan is for me like a vibration-drawing, transparent and monochrome, that neither my hand or eye could have observed or drawn" (Casini 2009).

In 2018, working with researchers Kumar Punithakumar, Richard Thompson and Peter Seres at the University of Alberta, Marilène Oliver and Gary James Joynes created the VR artwork, *Deep Connection*, using full body 3D and 4D MRI scan data of Oliver's own body. When the viewer enters *Deep Connection*, they see her scanned body lying prone in mid-air. The viewer can walk around her scanned body and inspect it, lie underneath, and walk through it. The user can dive inside and see its inner workings, its lungs, spine, brain. The user can take hold of the figure's outstretched hand: holding the hand triggers a 4D dataset, making the heart beat and lungs breathe. When the user lets go of the hand, the heart stops beating and the lungs stop breathing. *Deep Connection* creates a scenario where an embodied human becomes the companion for a virtual body and where the physical body interfaces with the virtual to animate it. When the VR artwork is exhibited, it is done so as part of an installation that includes sculptures generated from the scan data that has the VR hardware embedded in it. Whenever possible, Oliver is present in the exhibition so that she is able to guide the viewer through the experience. The concept here is that the viewer interfaces with Oliver as a virtual object (the rendered dataset, a physical digital copy and the original).

American artist Darian Goldin Stahl works with scans of her sister, Devan Stahl, a professor of medical ethics who has multiple sclerosis. Since 2015, Stahl has incorporated her sister's scans in her printed works. She has created several beautiful and tender artist books that combine and layer the MR scans with text from Devan's diaries that "tell her [sister's] diagnostic narrative and convey how it felt to see her MRI scans for the first time" (Stahl 2014). Darian explains that the pages of the her book, *The Importance of Dualism*, echo the slicing her sister's anatomy by the MRI scanner and that the binding of the *The Importance of Dualism* "alludes to the tension she feels with her body" (Stahl 2014).

Marilène Oliver's first work made with medical scan data was *Family Portrait*, a sculptural installation for which she arranged to have each of her family members MRI scanned at 20mm axial intervals at the Nottingham Queen's Medical Centre, UK. Later, Oliver screen printed the scans onto sheets of clear acrylic and stacked them to create a row of life-size sculptures. In *Family Portrait*, her father, mother, sister and Oliver herself are presented as elusive hovering figures, suspended in shiny, rigid structures. The spaces between the sheets mean that at eye level the viewer sees straight through the stacks of printed acrylic. With *Family Portrait*, Oliver was interested to understand and expose the digital and mechanical processes involved in MRI, as well as address posthumanist notions of digital preservation and a fear for the loss of embodied human interactions as a result of digitally mediated communication.

As more and more data is generated and collected from and about individuals, there is the question of what will happen to this data when we die. As Professor Remigius Nwabueze (2021) detailed in his keynote presentation at the KTVR e-Symposium, there is little legislation protecting the privacy of the dead. There is increasing discussion in the media and scholarship about digital legacies and writing digital wills (Bakewell 2017; TalkDeath 2019; Kasket 2020). Should medical scans be included in a digital will? What might it mean to inherit medical scans and other kinds of personal data from a deceased loved one? How are datasets or social media accounts different from photographs and letters?

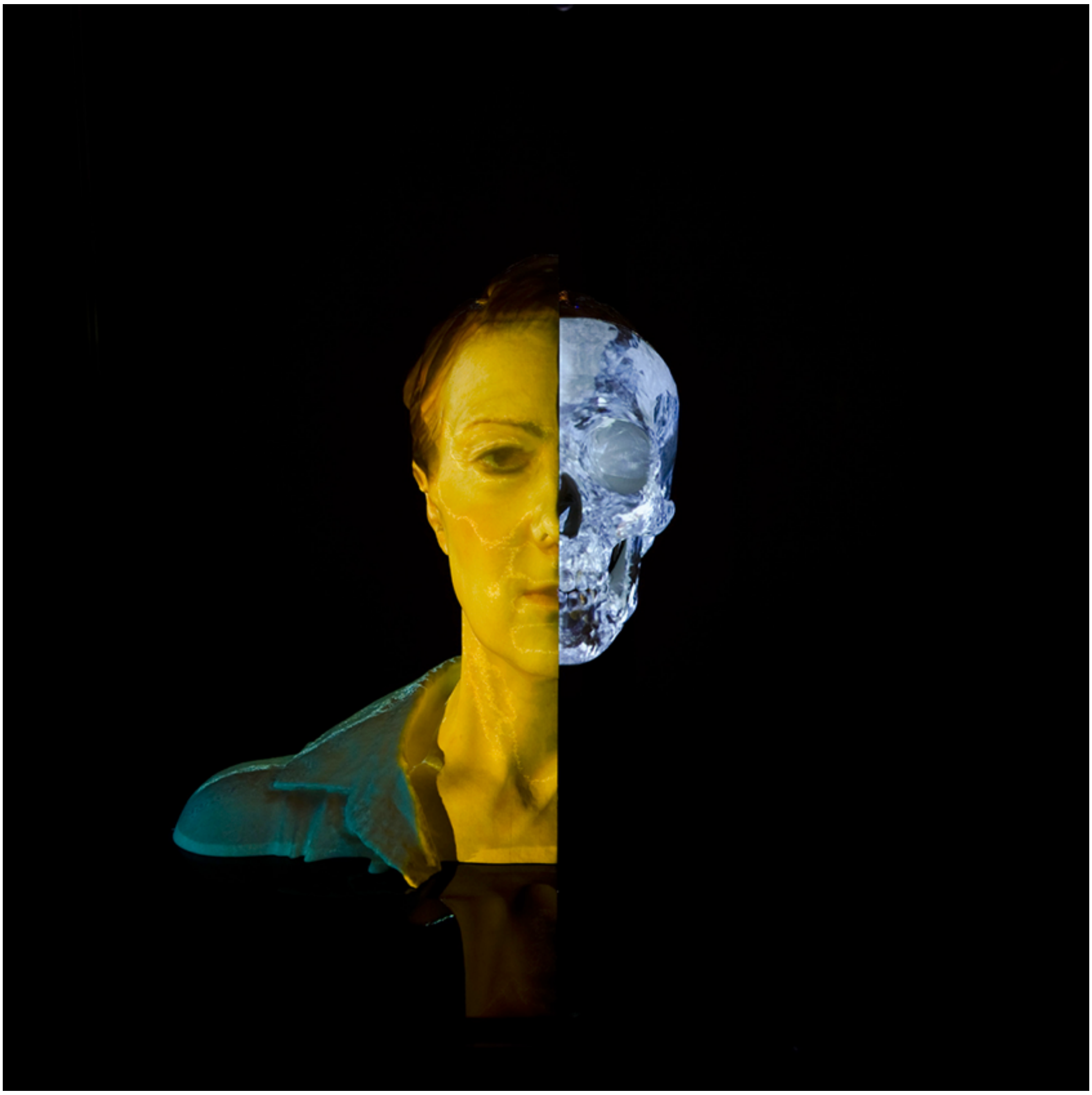


Figure 7. Jane Prophet, *Neuro Memento Mori*, 2014, 3D printed sculpture with projection mapping. Image courtesy of the artist.

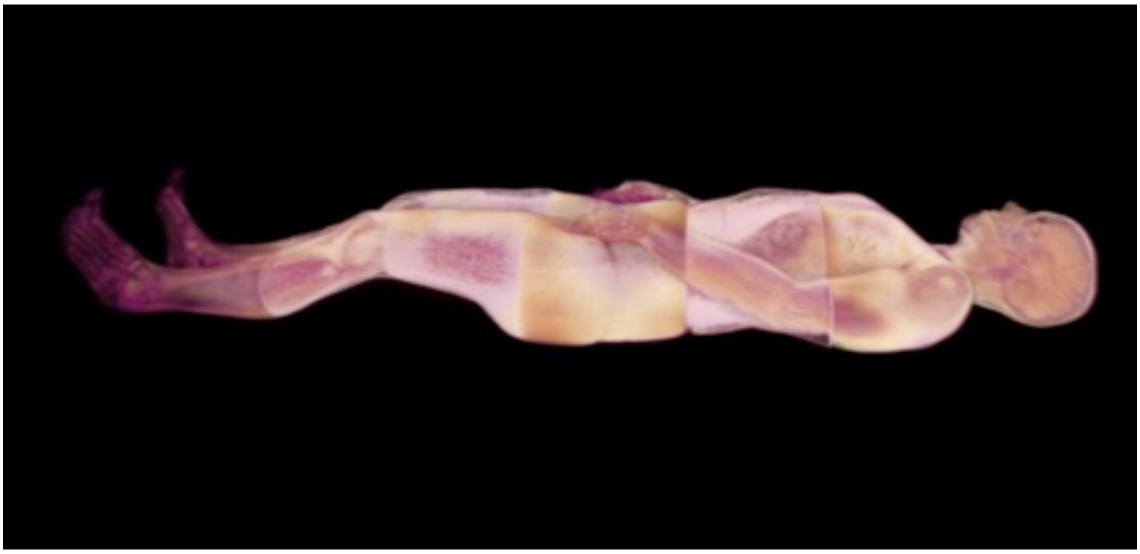


Figure 8. Marilène Oliver and Gary James Joynes, 2019. Screen capture of *Deep Connection*, VR artwork. Image courtesy of the artists.

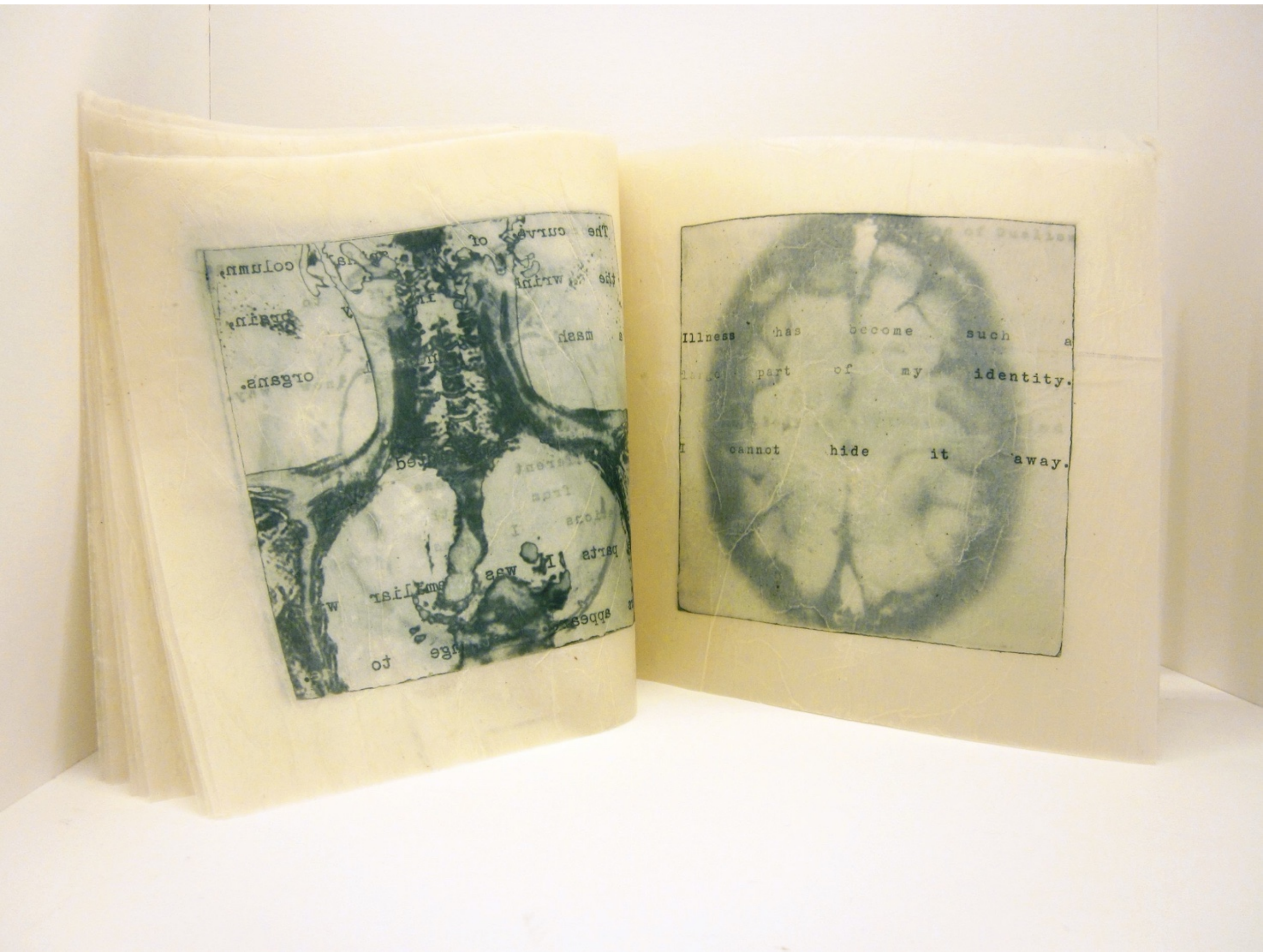


Figure 9. Darian Goldin Stahl, *The Importance of Dualism* 2014, photo intaglio and encaustics, 23 x 25cm. Image courtesy of the artist.

Figure 10. Marilène Oliver, *Family Portrait* 2003, screen prints on acrylic, each sculpture 50 x 70 x 192 cm. Image courtesy of the artist.

Open-source datasets

In many cases and especially in the exploratory stages of a project, it may not be necessary or relevant to work with personal scans/data. The acquisition of new data is a complex, lengthy, expensive process that requires access to research collaborators and institutions. It also requires a lot of energy to run the large and complex machines such as MRI and CT scanners. Furthermore, it takes time away from medical research. At the end of this section there is a list of open-access scan data repositories from around the world. These are almost all free to use for "research purposes." Most databases and repositories will include information about the study for which the data was originally acquired. The OpenNeuro database, for example, is an open-access repository of scan datasets that researchers have uploaded for other researchers to use. Researchers who upload data to OpenNeuro have agreed that they have ethics permissions to share the data publicly, that the data has been de-identified and defaced (using pydeface), and that it is publicly available under a **Creative Commons license**. Each dataset has its own digital object identifier (DOI), making it easily traceable to the original study. Each dataset has its own README file with information about the original study, which typically includes demographic information such as when the data was acquired, how many subjects were scanned, and what other data was collected as part of the study. Often, the title of the dataset will also give explanatory information about why it was acquired.

For example, the dataset titled "Emotion Regulation in the Ageing Brain, University of Reading" (Lloyd et al. 2021), accessible on the OpenNeuro platform, quickly indicates the original reason for acquiring the data, and suggests the scan subjects will be older and that there will be comparative functional scans that compare different emotions. It would be important in this case to read published papers related to the study to understand what the data represents. It is understandable that the original researcher and data subjects would object to an "incorrect" use of the data (such as if the aforementioned datasets where used as a part of an artwork about children's emotions). This is a good example of when it would be wise to reach out to the original researchers both to confirm they are agreeable to the data being used for creative research and to invite a conversation about the data and its findings. It is so easy to download data, but someone somewhere worked hard to create it and will probably be very happy to hear from another researcher or artist who is interested in working with the data in a surprising and experimental way.

Other data repositories, such as Open Access Series of Imaging Studies (OASIS), are available through institutions via an application process, which often includes an ethics review of the proposed work and may include restrictions on how the data can be used. The American National Institute of Mental Health's brain mapping project, the Human Connectome Project, also has a huge repository of brain datasets. Accessing the repository requires the completion and approval of a time-limited Data Use Certificate (DUC) (National Institute of Mental Health 2021). The DUC terms and conditions include clauses such non-transferability of the data, research-only use, no redistribution, deletion of data after the study has ended, and agreeing to share publications of "other public disclosure" with the NIMH. It also includes a useful clause (no. 5) that there will be no attempted re-identification of subjects or their relatives, and that in the case that "identifying information is discovered," users will notify the NIMH (2021).

In addition to open-access research datasets collected and made available by research institutions, radiology software often comes with libraries of datasets. The software OsiriX-Viewer, for example, which is available via subscription, comes with an image library of high-resolution scan datasets. OsiriX-Viewer stipulates that the datasets are for research and teaching only and cannot be redistributed, sold, or used for commercial purposes. Another open-source research software platform, 3D Slicer also provides scan datasets to be used with the software. A wiki page about the "Sample Data" invites new data to be added to the Sample Data module of the software. There is no restriction on the use of 3D Slicer or the datasets, but the developers make clear that it is the user's responsibility to ensure compliance with any applicable rules and regulations: "Slicer is **NOT** approved for clinical use and the distributed application is intended for research use. Permissions and compliance with applicable rules are the responsibility of the user" (3D Slicer 2020).

Through making several artworks with scan data over more than two decades, Marilène Oliver has found that most researchers and institutions respond to requests to use data to explain either why it isn't possible to share the data, or the conditions under which it can be used. Indeed, connecting with researchers with requests to use scan data has led to both a positive exchange of ideas and a broadening of audience. Since 2007, Oliver has made numerous artworks using one of the datasets available in the OsiriX-Viewer Image Library called Melanix. When she first started working with Melanix, Oliver contacted the creators of OsiriX-Viewer to tell them about her creative work with Melanix and to confirm that her creative use didn't contravene their data permissions. This contact later led to both an exhibition of her works in the Geneva University Hospital gallery, where the software was developed, and the inclusion of her works in *Le corps et son image*, a book by one of the key developers of OsiriX-Viewer (Ratib 2011). Likewise, Oliver's request to work with the CT scans of the infamous mummy Otzi the Iceman led to the sculpture she created with the scans, *Iceman: Frozen, Scanned and Plotted,* being exhibited at the Südtiroler Archäologiemuseum in Italy in 2011.

Key Datasets

In the KTVR research project, a number of scan datasets have been particularly helpful to think about the ethical use of secondary data—notably, the Visible Human Project, *Ben Body*, and BrainWeb.

The Visible Human Project

In 1994 the National Library of Medicine released the **Visible Human,** a dataset that includes CT, MRI, and cryosections of a male cadaver. Until 2019 a license was required to work with the data, but it is now publicly available with no permissions. The male cadaver was Joseph Paul Jernigan, a Texan convicted murderer who was killed by lethal injection in 1993. Before he was executed, he was reportedly convinced by the prison chaplain to donate his body to medical science. Despite only having one testicle and a missing tooth, Jernigan was selected from thousands of possible specimens to become the Visible Human (Waldby 2000). The Visible Human is still one the few full body scan datasets available and has been used by thousands of research teams. The Visible Human has also

Figure 11. Lisa Nilsson, *Shoulders* 2013, mulberry paper, 58 x 32 x 4 cm.
Image courtesy of the artist.

been written about in countless newspapers articles, and has been the subject of several documentaries and books. It has also been used by numerous artists to create artworks. Lisa Nilsson, for instance, has worked extensively with the VHP dataset to create complex and captivating quilled (rolled and shaped paper) works. Likewise, Miltos Manetas and Aaron Russ Clinger's web-based work *Man in the Dark* (Moss 2009) transformed the VHP images into a poetic floating body that dangles from the cursor.

The Visible Human is now part of a larger Visible Human Project (VHP) which has grown to include multiple datasets from different subjects. A later VHP dataset created from the body of Susan Potter, is quite a different case from Jernigan. After learning of the early Visible Human Projects, Potter, who was a cancer survivor and disability rights activist, convinced Dr. Victor Spitzer, the lead VHP researcher, to make her into a "visible human" when she died (*Becoming Immortal* 2018). At first, Spitzer resisted because the VHP was meant only to focus on digitizing healthy, "normal" bodies. Fifteen years passed between the time Spitzer agreed that Potter would be the next VHP subject and when she died in 2015. Over that time, they worked closely together and became friends. A National Geographic documentary made about their relationship shows how Potter was committed to making sure that "abnormal" bodies were also digitized for the purposes of medical education. When Potter first approached Spitzer, she had had twenty-six surgeries related to a car accident and was later diagnosed with melanoma, breast cancer, and diabetes. Spitzer insisted on digitizing Potter's life as well as her death by gathering her personal experiences and thoughts about her body, including her descriptions of her pain and her desire to donate her body to medical science. The Susan Potter VHP dataset is the most fully digitized human body, extending to these video and audio recordings. Yet the Visible Human Female dataset, available for free download on the National Library of Medicine website, is not that of Susan Potter but of a "Maryland Housewife" whose body was explicitly donated to the VHP by her husband after she died of a heart attack in 1995 (Waldby 2000, 2). It is not clear why the data that Susan Potter authorized for public release is not yet freely available.

Paywall datasets that have been expertly segmented and cleaned

Echoing the history of anatomy where beautifully illustrated hand-printed anatomical encyclopedias sold for a high price, there are also commercially available high-resolution datasets that include organs, the lymphatic, vascular, and nervous systems expertly mapped into them. *Ben Body*, by the Swedish company Interspectral in collaboration with Benjamin Moreno, is an "exhibition" based on a full body CT scan that can be rented by museums or institutions and explored via a touchscreen table. Renting the exhibition has a high price tag of €3,000, but given that the data took over a year to segment, this is a model of data usage that recognizes the work and expense of acquiring and processing data. The full Interspectral catalogue (Interspectral 2020) advertises a virtual ark of datasets for sale, including a scan of a woman who died in a traffic accident, a stroke patient, a full body scan of a man in his fifties in good health whose scan demonstrates aging, a chimpanzee, a golden eagle, a grey seal, a lion, a moose, several sharks, two stingrays, a fly, a beetle, a spider, and an ant.

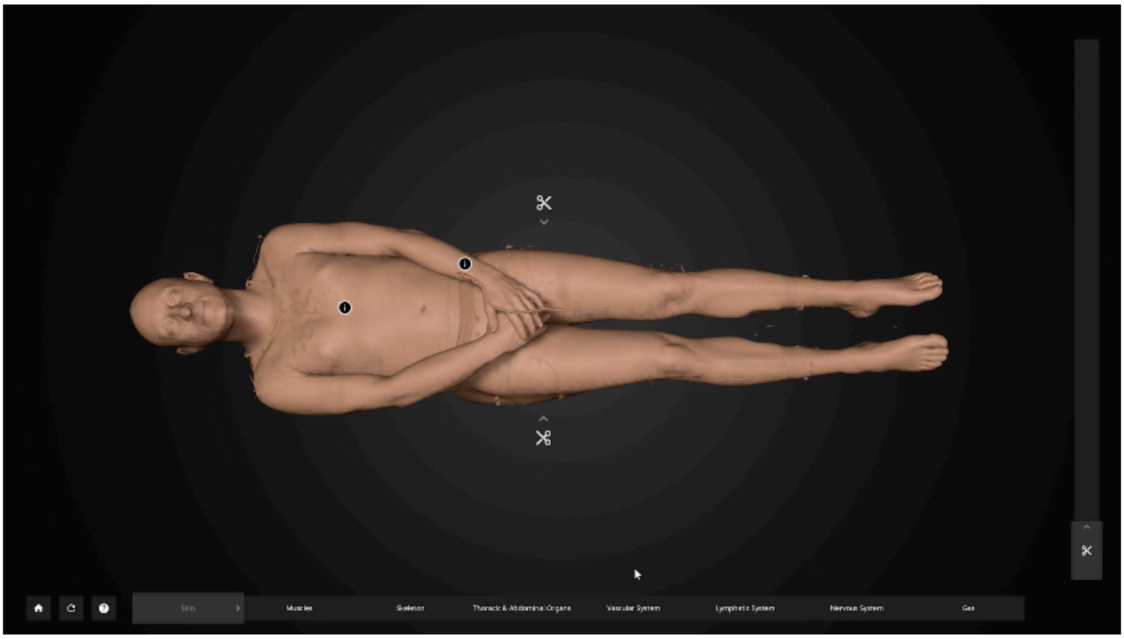


Figure 12. Screenshot of *Ben Body* exhibition, Interspectral website 2022

Simulated and Synthetic Data

Already in the late 1990s, researchers at McGill University had created datasets of a simulated, phantom brain. The project, called BrainWeb, includes both a "normal brain" and an "MS brain" (Montreal Neurological Institute and McConnell Brain Imaging Centre, n.d.). These datasets were created by averaging multiple datasets to create a single standardized set of scans that other scans can be registered or aligned to (Collins et al. 1998). It is important in medical research when comparing multiple brain scans first to ensure they are all orientated in the same way as this is difficult to ensure when the subject is scanned. It would be important when using the BrainWeb datasets as creative researchers to be thoughtful about working with an averaged MS brain rather than the averaged "normal" brain and vice versa—not only because mis-using a dataset would no doubt invite objection by the researchers, but because it would be disrespectful to the original data subjects whose data contributed to the creation of the BrainWeb datasets.

Machine learning is also being used to generate high-resolution synthetic datasets. Researchers are working with generative adversarial networks (GANs) trained on large datasets of real scans to learn patterns and rules within scan datasets from which to generate synthetic or "fake" data (Mirsky 2019). Again, here the original subject of the scan does not exist—the machine learning model is generating new scans based on what it has observed in the dataset of scans it has been fed.

In some cases, synthetic data may be an alternative to conventional data, and, depending on the proposed use, synthetic data could provide a dataset relatively free of privacy issues. According to a 2020 Benaim et al.) study, when based on a large enough population, synthetic data resulted in similar findings to conventional data. Because the data subjects in synthetic data do not exist, there is no possibility of disclosure. Depending on the type of data used, it may still be important to consider who the dataset was modelled on, including the make-up and size of the sample population.

While this emerging area promises to alleviate certain privacy issues, it is not without its own challenges. Decisions made in the construction of the dataset can introduce bias into the results, and the value of the data itself can be compromised if the underlying dataset is processed multiple times. How much "reality" is left and does that matter? Or could be a conceptual element of the work that the data visualized is an aggregation of thousands of bodies or brains? Media artist Refik Anadol's 2021 work *Sense of Space* was created from synthetic data generated from the Human Connectome Project (HCP), which is built from 70 terabytes of multimodal MRI scans from over 4,500 subjects (Anadol and Hotamisligil 2021). The project, which was made in collaboration with the coordinator of the HCP, Dr. Taylor Kuhn, presents the architecture of the brain as an immersive experience into which the viewer can enter. In *Sense of Space*, the fact that the brain space that viewers enter is generated from thousands of brains "from birth to nonagenarians" is conceptually central to the work, speaking to the 2021 Venice Architecture Biennale's theme "how will we live together?." Anadol and his collaborators present a utopian, technologically enabled, corporately sponsored (Siemens, NVIDIA, Epson, Arup, AiBuild) architectural structure that is an "*every* brain." The HCP's massive database includes scans of healthy brains from subjects of various ages, as well as brain scans acquired as part of studies for epilepsy, anxiety and depression, aging and dementia, early psychosis, and anxious misery. Only "healthy brains," however, are used for *Sense of Space* (Anadol and Hotamisligil 2021), most likely because of the complexity of amalgamating so many datasets and the scientific need to control the data when making the connectome model. From an artistic perspective, however, excluding neuro diverse datasets when making an artwork that proposes a future way of 'living together' is important to be conscious of and carefully considered.

Missing datasets

Despite the abundance of data that exists, there are times when data for a particular situation does not exist. Since data is collected for a purpose, a lack of data can speak to power imbalances about who is determining what data should be collected. Dataset acquisition is expensive, and most research occurs in economically and technologically developed parts of the world, reflecting the social and economic disparities in access to health care both within a given society and across the globe. If diverse populations are not accessing health care equally, there will be less data (scans, etc.) from those populations. Considering medical datasets and AI, there is a good deal of literature examining how homogenous (usually white) datasets have caused significant problems when applied to a diverse population. Certain groups, such as older minority communities, can be systematically excluded from medical research, resulting in institutional bias or racism (Bécares, Kapadia, and Nazroo 2020). Women have also historically been underrepresented in clinical research (Criado-Perez 2019). Mimi Ọnụọha is an artist whose work highlights the social relationships and power dynamics behind data collection. In her bright white installation *The Library of Missing Datasets* (2016), Ọnụọha presented a filing cabinet full of empty files representing uncollected data as "things that have been excluded." Each empty file is labelled to identify the missing data. Examples of empty files in the cabinet include people excluded from public housing because of criminal records, public lists of citizens undergoing domestic surveillance, white children adopted by People of Colour, and the quantifiable effect of corruption in lean economies. Ọnụọha added another volume to the work in 2018, which focuses on Blackness. As well as the physical installation, there is also a Github repository for the project, which is empty save a README file explaining the project and an essay about information networks and power dynamics (Ọnụọha 2016b; Bossewitch and Sinnreich 2013).

Figure 13. Refik Anadal, *Sense of Space* 2021, AI projection mapping and 3D printing.
Image courtesy of the artist.

Figure 14. Mimi Ọnụọha, *The Library of Missing Datasets* 2016, mixed-media installation.
Image courtesy of the artist.

Provenance, Access, and Licencing Discussion Questions

- Who are the data subjects? Is there demographic information about who is in the dataset? (This could be data within the dataset or it could be a generalized summary of the data subjects.) From what geographical location were the participants recruited?
- How were the data participants recruited? Was it an open call to the public? Was a particular group of people targeted (e.g., students, employees of an institution, etc.)? Does the method of recruitment affect the artwork that uses the data?
- For older datasets or for datasets where the donors are known to be deceased, is the consent that was given while alive appropriate in order to respect the privacy for the dead or their families? Does using the data impact those who knew the data subjects when they were alive?
- Is there a Creative Commons license for the data? If so, what kind?
- Should you contact the original creator of the data to find out more about the original study and confirm permission for creative research?

Synthetic Data Discussion Questions

- Does the work examine issues related to a specific community or group of people? Does the synthetic data accurately reflect that group? Does a dataset exist that can be used without compromising the wishes or the privacy of the data subjects?
- Does the work depend on the lived experiences of individuals captured in the data, or will the data be used in a much more general, non-specific way? Does synthetic data exist that can be used instead?

List of selected open-access medical image repositories

Alzheimer’s Disease Neuroimaging Initiative

http://adni.loni.usc.edu/data-samples/access-data/

Cancer Imaging Archive

https://www.cancerimagingarchive.net/collections/

Embodi3D

https://www.embodi3d.com/

Interspectral

https://interspectral.com/

MedPix

https://medpix.nlm.nih.gov/home

OpenNeuro

https://openneuro.org/

Oasis

https://www.oasis-brains.org/

OsiriX

https://www.osirix-viewer.com/resources/dicom-image-library/

Further links

https://www.aylward.org/notes/open-access-medical-image-repositories

https://radrounds.com/radiology-news/list-of-open-access-medical-imaging-datasets/

Consent

Consent is the way that parties agree to participate in a proposed activity, event, or arrangement. In the context of research, consent is a particularly important topic as it respects the autonomy of the data subject to voluntarily engage in the research process. If we breakdown the word *consent*, it is comprised of the following root terms: *con* (together) + *sentire* (feel) = Feel together, a mutual feeling (Peña and Varon 2019, 24). Thus, the etymology of the word *consent* speaks to the idea of mutuality of feeling. There are legal requirements for consent, which are outlined in more detail in the *Current Data Protection Guidelines* section. Here, the aim is to move beyond legalities and examine the social context or contract that informs the notion of consent. This broader understanding of consent is especially important when it comes to both the original collection of data online as well as the reuse or secondary uses of data.

Click to Consent

We are all familiar with checking boxes and clicking buttons as a way to consent to our data being collected by social media and commercial websites in exchange for access to their services and platforms. The Edward Snowden files and the Cambridge Analytica scandal revealed the negative political, social, and economic impacts of this practice and how little we understand what our data is, its value to others, and how it can be used against us for political marketing.

Although there are far more protections around medical scan data than data we provide to social media platforms, it is important to keep in mind broader discussions around the problematics of "consent" in the digital age and ask whether they can steer creative researchers towards considering consent issues more sensitively and mindfully as artists working with personal data. In their 2019 publication "Consent to Our Data Bodies: Lessons from Feminist Theories to Enforce Data Protection," Paz Peña and Joana Varon bring a feminist lens to thinking through consent to data usage. They remind us that consent is at the heart of social movements against sexual abuse and harassment movements such as #MeToo, and that its meaning, when discussing justice as it relates to the violation of the physical (typically female) body, is contested or avoided. They explain the problematics of binary consent options and the illusion that consent can be a free, rational, and individual choice. When a simple click can give access to a website, or, in the case of some medical research projects, to a potentially life-saving study, is there really a choice? In the same way that terms and conditions for social media sites are unreadable, is there a risk that research ethics information sheets and consent forms (and, dare we say, artist statements and project proposals) are equally unintelligible, prompting us to click through rather than meticulously scroll through the entire text?

In his paper "Biometric Privacy Harms," American legal scholar Matthew Kugler (2019) points to a series of studies showing that people's opinions about the use of biometric data are far from binary. Rather than just opting in or out from data-sensitive technologies, people want more nuanced, granular, and informed choices. Kugler cites a study by Research Now/SSI (which has since become Dynata) that found that most people were comfortable using their thumbprint to unlock a smartphone or banking app but much less comfortable with allowing department stores to use facial recognition software to send them targeted ads. Studies also showed that younger people generally have "greater privacy expectations" than older people. Kugler and Peña and Varon all point to evidence that people are not generally willing to pay for more privacy, and that the more choices offered by organizations using biometric identity systems, the better. These studies show that it is necessary to review and reassess boundaries and expectations around privacy.

Incidental Findings

When medical scan data is acquired for research, there is a requirement that scans are reviewed for **incidental findings** by a medical doctor. Any incidental findings are confidential and only shared with the data subject and their physician (if this is requested in the consent form by the data subject). At this stage, the data subject may decide to withdraw their consent for their data to be used in the study.

The nature of artistic practice is that artworks evolve and shift as they are developed, making consent withdrawal or ongoing consent a best practice when seeking permission to use personal data in the production of an artwork. Ideally, the person who has given permission for their data to be used is given the opportunity to see the artwork as it is being made and before it is exhibited publicly so they understand how their data is being presented and contextualized. There may be instances where subjects are uncomfortable with how their data is being represented or if something is revealed, much like an incidental finding. When making the VR artwork *My Data Body*, which uses the artist Marilène Oliver's Facebook data, Oliver discovered that Facebook stores all searches that have ever been made by the user. Oliver was uncomfortable with including her Facebook search terms in the work, so this data was deleted from the project.

A framework that could be adopted to protect all parties when an incidental finding is discovered in personal data while creating an artwork is one used in contracts for artist commissions where the artist provides preliminary sketches and final sketches to the client that both parties agree upon at agreed dates in the project. CARFAC Ontario provides examples of contracts for private and public commissions via their website for a small fee (CARFAC Ontario 2016).This process would protect both the artist (who is investing considerable amount of time and resources in the creation of the artwork) and the person giving permission to use their data (who can't fully understand the impact of the artwork until they see/experience it).

Longitudinal studies

Longitudinal studies are a very useful methodological approach to collecting large amounts of data about specific populations. Cohort studies are a subset of longitudinal studies that follow a group of people of a specific age range, often beginning with children and following them through a significant portion of their life. Such examples include two British cohort studies, the 1958 National Child Development Study (Power and Elliott 2006), an ongoing study which involves 17,415 people born in 1958 in England, Scotland, and Wales, and the Millennium Cohort Study (Joshi and Fitzsimons 2016), which is following 19,000 people in England, Scotland, Wales, and Northern Ireland who were born between 2000 and 2002. In these cases, it would have been the parents who consented to the child's participation, which means the participants themselves are not the ones granting consent. This scenario could lead to issues as children become adults and have the ability to understand the implications of participation and are able to withdraw now but not retroactively (Helgesson 2005).

In her essay on the HUMAN Project (Human Understanding through Measurement and Analytics), American anthropology professor Shannon Mattern (2018) explains how the longitudinal project aims to "measure everything we possibly can about 10,000 people over the course of the next 20 years or more." It aggregates multiple data (location, activity, health, sleep, gaming, banking, voting) about the individuals in the study and cross-references it with environmental, social, and political conditions (Mattern 2018). This study will undoubtedly yield unexpected results as technology and social factors change that could impact participants' levels of comfort with the study. Ethically, such longitudinal studies would benefit from ongoing informed consent, as participants' understanding of and consent to participation could change over the course of the study (Helgesson, Ludvigsson, and Stolt 2005).

Machine learning "troubles" consent

Likewise, the need to understand and be transparent about the systems into which data is being fed and processed is essential. Peña and Varon (2019) remind us that asking for "informed" consent is an oxymoron when machine learning is applied to data. Machine learning relies on unsupervised algorithms to find patterns and rules within a dataset in order to perform a given task (generate new images, text, scan data, social media posts). It is impossible to know what the algorithm is learning from the dataset (Bridle 2018). How can researchers ensure informed consent when they don't know what the algorithm is learning from data? This problem is even more important when large, aggregated datasets are used, as it is very difficult to fully inspect and understand the contents of the dataset. The opacity of machine learning algorithms does not mean that artists shouldn't work with machine learning, but if machine learning is used to process a person's data in a creative or artistic project, how the machine learning was developed and how it will be used should be clearly explained to them. It is also important to be aware of the many massively funded machine learning projects currently working to realize data aggregation in the future. Banking, for example, is an industry that is investing heavily in biometrics as way to tackle identity fraud. If (or when) biometric markers such as fingerprints and retinal scans become the "keys" to our data, there is a risk that aggregated data could be used in ways we presently cannot predict.

Healthier qualifiers of consent

Research ethics procedures for human data collection already require that information and consent is communicated to participants in clear and plain language that meets the informational needs of the individual. There is an opportunity here to think about consent more broadly as an ongoing process that takes into account better boundaries that empower the participants to understand their data and involve them in the direction of the project. If participants can be involved because data has been anonymized, are there ways to understand the conditions under which permission was obtained, and can this become part of the content of the work? In their conclusion, Peña and Varon (2019) propose a set of "healthier qualifiers for consent" that are helpful when working with personal data. Although not all of them are practical (or necessary), they can be actively considered and aspired to as a way to arrive at a "mutual feeling" for each other's data bodies. They state, the act of consent needs to bena) active, meaning actively agreeing with body and words to do so (not only the absence of no); b) clear and intelligible; c) informed, fully conscious; d) freely given, out of choice and free will; e) specific to a situation, therefore f) retractable and g) ongoing… (Peña and Varon 2019, 24).

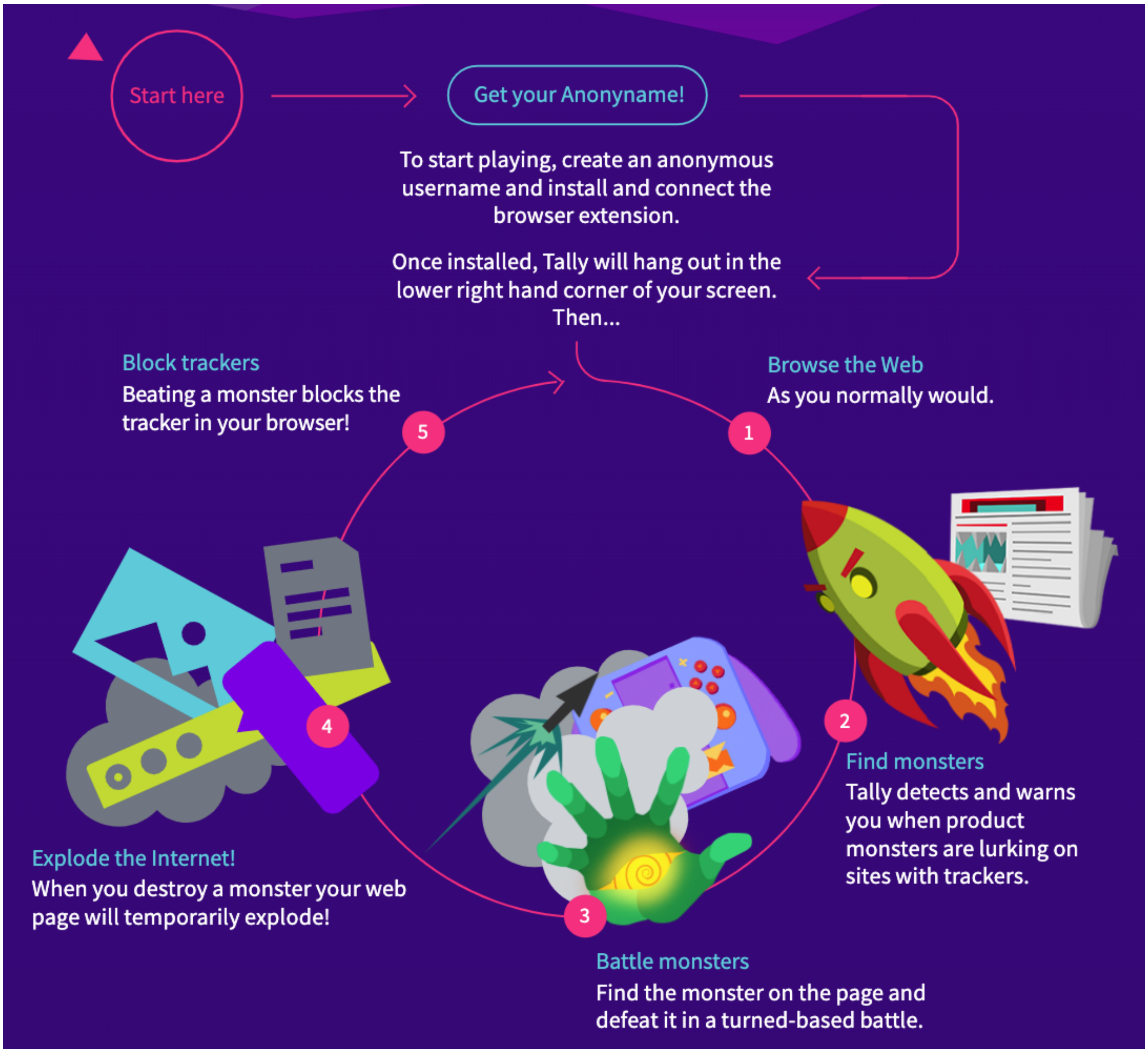
Artists making artworks about the concept of data consent

Tally Saves the Internet and *I Know Where Your Cat* Lives are two artworks that playfully reveal how much data we unknowingly "consent" to giving away when we upload images and use the internet. I Know Where Your Cat Lives is a website that allows you to search for cats anywhere in the world via Google Maps. To create the work, Owen Mundy and his team used seven million public images of cats and plotted them using the metadata in the uploaded images (Mundy and Berger, n.d.)The project is both playful and sinister; we are both pulled in by the cute photos of cats and repelled by how our beloved, domestic felines are so precisely geolocated by Google. The project website explains clearly how the project was created and advises you to increase the privacy settings of your images to avoid the metadata of your images being unknowingly exploited. Here Mundy demonstrates that whenever we upload an image to a platform such as Flickr, Twitpic, and Instagram we are "consenting" to its metadata also being uploaded and used by the platforms. Another more recent project by Owen Mundy and Joelle Dietrick called Tally Saves the Internet is a game that blocks data trackers and "educates players about data privacy" (Mundy and Dietrick 2020). Tally is a little colourful blob that "hangs out" in the corner of your browser and tells you when advertisers/monsters are collecting data and allows you to battle/block them, "exploding" the internet as you go. In this project, which was conceived with their daughter in mind, Mundy and Dietrick meet many of Peña and Varon's criteria for obtaining consent. Tally makes consenting to data tracking active, clear, informed, specific to particular webpages, retractable, and ongoing. Choosing to not consent to trackers does not mean exclusion from accessing a webpage.

STEALING UR FEELINGS, an interactive film by Noah Levenson made in 2019, uses facial recognition and emotion recognition software to "extract data from your face" as you watch the six-minute documentary about Snapchat's "Determining a mood for a group" patent (Levenson 2019). As you watch the documentary, includes images of pizzas, cute dogs, and Kanye West, your camera-enabled device watches you. At the end of the documentary, you are offered a downloadable scorecard with the assumptions that AI made about you when you were watching the film, including IQ, annual income, gender and racial bias and how much you like pizza, dogs, and Kayne West. Like *Tally Saves the Internet*, Levenson's project offers a way towards "withdrawing" consent by signing a petition against the patent. To access the petition, you have to smile.

In her durational performance *LAUREN*, artist Lauren McCarthy becomes "a human intelligent smart home" (McCarthy 2017). The performance starts with McCarthy visiting the homes of participants in person (who sign up via the Get-Lauren website) to install a series of networked devices such as cameras, sensors, lights, and door locks. *LAUREN* then watches over participants 24/7 using the installed cameras and sensors. *LAUREN* has the ability to remotely control the home of the participant with networked switches. Like the automated and networked assistants Alexa and Siri, *LAUREN* can be directed by the participants via voice command. *LAUREN* and will also try to anticipate her participants' needs and desires as she watches over them and learns their routines. In a video documentary reflecting on the experience of having *LAUREN* in their homes, participants reported mostly positive reactions. They likened *LAUREN* to a friendship that is “about them,” and that noted that it feels as though *LAUREN* is “in support, not in control of them.” (McCarthy 2017). With *LAUREN*, McCarthy offered participants (and herself) an opportunity to “feel with” surveillance technology in a durational and active way. As McCarthy reflected, the tasks she performed for her participants were mostly those they could easily do themselves, such as flipping switches, playing music, and looking up answers to questions online (McCarthy 2018). . The artist and others may wonder if such menial tasks are worth the requirement that participants provide access to their personal data. Or perhaps there is something else gained from these surveillance technologies, such as a sense of companionship and care? McCarthy writes,

> In *LAUREN*, I am wrestling for control with artificial intelligence. The participants are also negotiating boundaries and poking at the system. The point of this project is not to impose a point of view, but to give viewers a space to form their own. Immersed in the system in the comfort of their homes, people are able to engage with the tensions. Some moments are awkward and confusing, others are hopeful and intimate. Together, we have a conversation. Do we feel any limits when it comes to letting AI into our data, our decision making, and our most private spaces? (2018)



To start playing, create an anonymous username and install and connect the browser extension.

Once installed, Tally will hang out in the lower right hand corner of your screen. Then...

Browse the Web
As you normally would.

Find monsters
Tally detects and warns you when product monsters are lurking on sites with trackers.

Battle monsters
Find the monster on the page and defeat it in a turned-based battle.

Explode the Internet!
When you destroy a monster your web page will temporarily explode!

Block trackers
Beating a monster blocks the tracker in your browser!

Figure 15. Owen Mundy and Joelle Dietrick (Sneakaway Studio), diagram of *Tally Saves the Internet* 2021. Image courtesy of the artist.

Figure 16. Marilène Oliver's scorecard from *Stealing Ur Feelings* film by Noah Levenson, 2019. Image courtesy of the artist.

Consent Discussion Questions

- What kind of consent was obtained from the data subjects at the time of collection? Is the consent form available? What kinds of uses were described in the consent form? Does the proposed work conflict with those uses?
- Where was the consent obtained? Where will the data be used in creative work? What are the differences between the laws under which the consent was obtained and the laws under which the dataset is being used? Did the data subjects consent to use of their data outside of their political/legal region? Are there any geographical restrictions within the terms of use of the dataset?
- If acquiring new data, how clear is the explanation of your project and how flexible is the consent form? Are there binary choices (such as opt in or opt out) or are participants given a range of choices in their responses?
- How practical is it to obtain ongoing consent? Should the participant be asked for consent to their data used in every exhibition in which the resulting work will be shown; or should they consent to inclusion of their data in every paper or other form of dissemination?
- What would happen if consent was retracted by the data subject? What would happen in an artist had spent months making an artwork and then the data subject retracted their consent to use their data?

Examples of consent forms

University of Alberta (Open Access) Human Research Ethics Forms and Templates
https://www.ualberta.ca/research/research-support/research-ethics-office/forms-cabinet/forms-human.html

University of Michigan Informed Consent Guidelines and Templates
https://research-compliance.umich.edu/informed-consent-guidelines

Participation, Authorship, and Dissemination

Participation and Authorship

In the previous section, we discussed consent and recommended involving data subjects/participants in artistic projects as much as possible so that they are informed and able to make ongoing choices with regards to their data. Their involvement leads to considerations of whether or not the data subjects in the datasets are in fact collaborators or co-authors in the resulting artwork. If they have been heavily involved in the project by giving ongoing feedback and making aesthetic and conceptual suggestions, should they be considered a collaborator in the artwork and credited as such? If yes, should they be acknowledged, and what kind of acknowledgement is appropriate while respecting their privacy? In some cases, acknowledgement may compromise the confidentiality of the data subject (Cox et al. 2014, 19). The participant may however be very comfortable with their identity being linked to the data in an artistic context, especially if the work engages with themes they are passionate about and thus they may waive their privacy rights in such conditions.

If secondary data is used in the creation of an artwork, should the holding institution be acknowledged? Is data accreditation/citation explicit in the terms of use? How should the ways in which the datasets were obtained be acknowledged in the work? The different kinds of data permissions and citation regulations were explained in the *Current Data Protection Guidelines* section, but here we pause to consider what the artwork needs conceptually and how and where citation and accreditation are appropriate. Should data accreditation be displayed next to the artwork when it is exhibited, or is it better to reserve this information for catalogues and exhibition statements? Could there be a webpage for the project that includes such credits, with an acknowledgement accessible via a QR code? It may confuse the experience and intention of the artwork for too much textual information about the artwork to be displayed as part of the exhibit. For example, if the artist creates an immersive audio-visual installation, they will unlikely want the installation to be interrupted by large lit text panel. Furthermore, when artworks are curated to exhibitions where the artist is not present or not in charge of the visual identity of the show, the amount of didactic information is harder for the artist to control.

Dissemination

The nature of artworks is that they re-present, make visible, and reframe their subject matter (in this instance, data) for a public audience in galleries, museums, publications, and online platforms. Often, the experience of an artwork occurs in a collective and public space. This level of visibility is different from a scientific paper, which is typically read in a more private setting such as an office or library and usually does not have the aim of eliciting intense, emotive, embodied encounters with data as a way to explore a personal or socio-political question, which can often be the goal of artists.

It is more or less standard now for visitors to exhibitions to use their smartphones to take photographs of artworks and share them on social media (with comments you are not able to control), making them even more visible. Not only does the sharing of images of the exhibit risk exposing the data subject (if they wanted to remain anonymous) to facial recognition algorithms that social media platforms use to tag people in images, but it also means images can become part of feeds with which the data subject may not be comfortable.

Your Data Body

One of the KTVR VR projects, *Your Data Body,* works with the data of others. The project has evolved (and keeps evolving), continually raising practical and ethical challenges, including those around authorship. The work was originally made using a combination of open-source and donated datasets, with the goal of focusing on issues of data privacy. In VR the user is able to pick up, move, resize, recolour, and duplicate the scanned body parts, or stack them to make Frankenstein-like figures. Audio files attached to each scan play when the user holds and manipulates each dataset. Anonymized open-source datasets are accompanied by an automated voice recounting the study data published alongside the dataset, whereas datasets "donated with explicit consent" have a "personal story" based on the original subject of the data. Early in the project, a request for "donations" of scan data was sent out via email to potential donors by means of university listservs. Several people responded to the call but preferred not to do the recording and wished only to donate their data anonymously. For those who did provide a recording, they can choose to be listed as a contributor in the credits of the project.

One participant who responded to the call with a donation of over ten datasets is Canadian artist Liz Ingram. Ingram has created several artworks with her own medical scans, which she's been acquiring since 2014 as part of her ongoing oncological care. In 2019, for instance, Ingram worked with her husband and collaborator Bernd Hildebrandt to create *Light Touch*, a large silk fabric tent printed with images of her brain scan held tenderly in both her own and her husband's hands. Inside the tent, on the floor, is a poem written by Hildebrandt in mirror vinyl. Marilène Oliver, who is leading the creation of *Your Data Body,* struggled to work artistically with Ingram's scans knowing that Ingram and Hildebrandt had already made strong aesthetic choices about how Ingram's scans are presented in their artworks (typically very fragile, transparent, and intermingled with images of flowing water and poetic text). Both Ingram and Hildebrandt have been invited to become collaborators and will work with the creators of *Your Data Body* to decide how Ingram's scans are rendered in the project. Inviting Ingram and Hildebrandt's collaboration now requires the consideration of new questions: How will Ingram and Hildebrandt be credited? Will they be full authors of the work, even if they only work with Ingram's own scans and not other parts of the project (which will include several other datasets)? Will Ingram and Hildebrandt partly "own" *Your Data Body*? What if *Your Data Body* is sold at a future date? What percentage of any sales would they receive? To avoid misunderstandings and avoid any future conflict with regards to artistic control and ownership, it will be wise in this instance to consult CARFAC's sample Artists' Collaboration Agreement and agree to terms of the collaboration (Sanderson and Hier 2006) .

The way in which different open-access datasets are used in *Your Data Body* has also evolved over the duration of the project. Not all open-access datasets have the same amount of information; accompanying information ranges from just the dataset file name to detailed scientific papers detailing demographic information and pathologies of the data subjects, the scanner used to acquire the scans, what was discovered from the data, and so on.

One of the open-access datasets used in *Your Data Body* is the Visible Human Project (VHP), a widely used, open-access scan dataset of Joseph Paul Jernigan (also discussed in *Provenance, Access, and Licencing*). There are several documentaries, books, and news articles about the VHP that detail Jernigan's personal life before he was executed and the process of his corpse being digitized. Additionally, the creators of VHP, the American National Library of Medicine, now have a whole webpage about the project that provides links to four VHP conference proceedings and several academic papers. One of the original goals of *Your Data Body* was to have a conversational AI avatar that would advise or guide the viewer through the VR experience. Through group discussions about what the AI avatar should know and what form it should take (e.g., a CGI humanoid or scan body part), it was suggested that the AI avatar could be the Visible Human. We are now attempting to train a conversational AI model on information available about the VHP, in the hope that viewers of the *Your Data Body* artwork will be to "chat" with the Visible Human scan dataset about its history and what it has been used for. To train the model, it was necessary to create a dataset of fifty conversations between the Visible Human and future viewers. We started by making a character profile listing what the Visible Human knows about their data, feels about their data and wonders about their data. In some way, this is an attempt to give some authorship to the Visible Human and Joseph Paul Jernigan. As it is impossible to get posthumous consent from Jernigan, is this more appropriation than attribution? Or is it fact-based fiction? Working with data of the deceased is an important grey ethical area to trouble and expose, especially as large corporations such as Microsoft are already working to commercialize the creation of bots from specific people (Abramson and Johnson 2020).

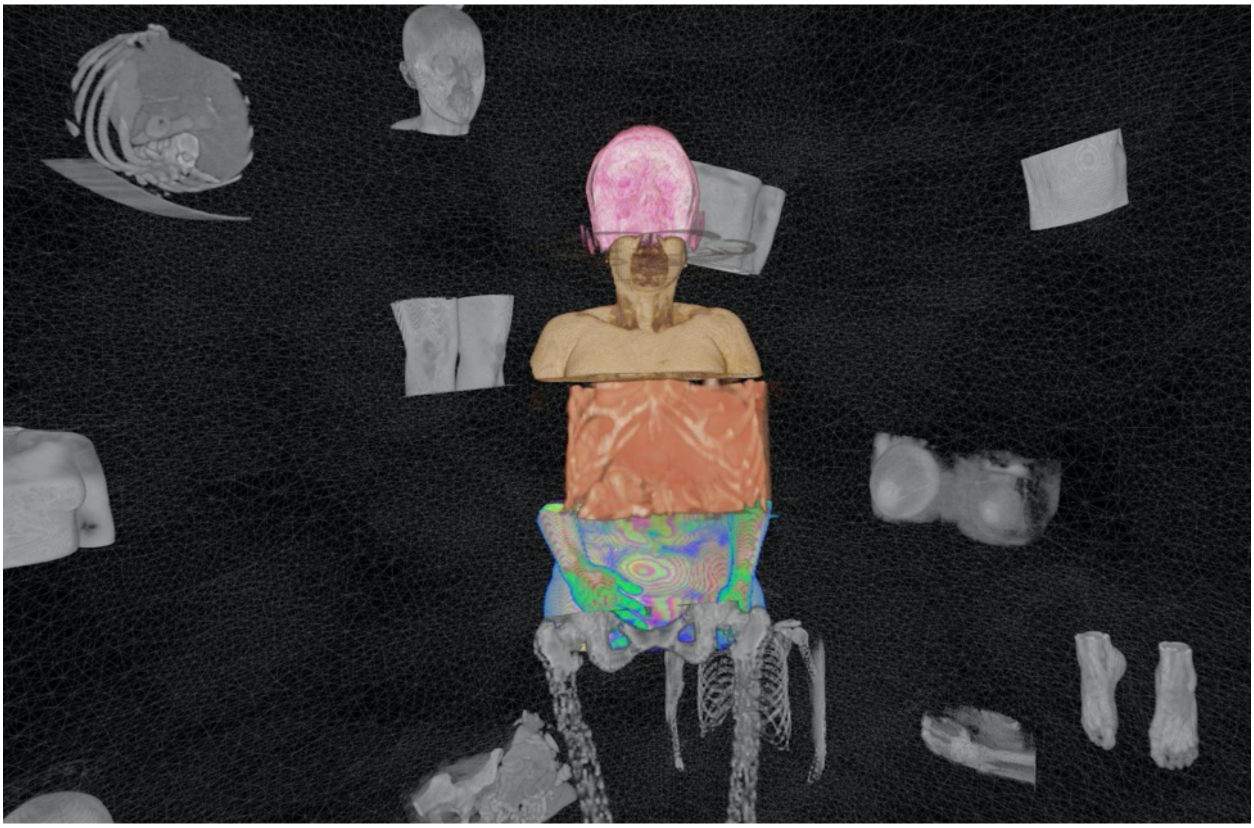


Figure 17. Marilène Oliver, Screen capture of *Your Data Body*, VR artwork (in progress), 2022.
Image courtesy of the artist.

Participation, Authorship, and Dissemination Discussion Questions

- Are the data subjects considered participants, contributors, collaborators, or authors in the artwork? How should data subjects be acknowledged in a specific work?
- Is the institution that created and stores the dataset (the custodial institution) a participant? Should the institution be acknowledged as a participant? Does acknowledgement of the institution compromise the identity of the data subjects represented in the dataset? Has the institution been contacted to determine whether they want acknowledgement in any research project created with their data? Does any user agreement or terms of use stipulate whether acknowledgement is required?
- How is acknowledgement done? Does it risk disclosing the identity of the data subject?
- Do data subjects or the custodial institution hold any right to the resulting artwork? Should data subjects or the institution have access to the resulting work (perhaps as digital or physical copies of whole or part of the resulting work)?
- How are the works shared online, if they are shared online? Will visitors to exhibits be able to take photos of the works and share them as they wish on their social media feeds or elsewhere?
- If images of artworks are posted online, how are comments managed? Is it appropriate to turn comments off or to actively manage comments, or is it more important to allow a free discussion to happen? If comments are open, how are problematic comments dealt with, particularly if data subjects are the target of the comments and if those comments are disrespectful to them? Is it appropriate to delete comments? Is deletion transparent, with the necessary steps in the decision-making process made available?
- How are alternate readings of the work acknowledged, discussed, and, if necessary, refuted? Is there a risk that images of the data subject in the artwork will be documented and shared out of context?

Data Subjects

Situated and nuanced cultural understandings of data ownership and stewardship

> Computation is a cultural material. Computation is at the heart of our digital technologies, and, as increasing amounts of our communication is mediated by such technologies, it has become a core tool for expressing cultural values.
> —Jason Edward Lewis, ed., *Indigenous Protocol and Artificial Intelligence Position Paper*

Much of the guidance and legislation discussed thus far is informed by an individualistic worldview, in which the emphasis is on individual ownership and the privacy of an individual's data. Governmental regulations and institutional guidelines have also been mostly binary, purporting that one way of working with data is "right," while another is "wrong." It is crucial, however, to consider data ownership and stewardship from a collective and culturally, geographically, and historically situated perspective. Jason Lewis's 2020 Indigenous Protocol and AI Position Paper, OCAP®, and CARFAC's 2022 Indigenous Protocols for the Visual Arts offer multiple, less binary, and situated ways of considering data ownership and stewardship.

OCAP®, which stands for ownership, control, access, and possession, is an educational resource created by the First Nations Information Governance Centre (FNIGC) to help First Nations communities in Canada control data collection processes in their own communities and how information is used. In a short video about OCAP® on the FNIGC website, the narrator explains how First Nations have always protected resources for future generations and that one of today's most important resources is data/information (First Nations Information Governance Centre, n.d.). First Nations people understand that information is not just about numbers and surveys; it is about culture, identity, traditions and self-determination. OCAP® ensures that First Nations people are the stewards of their own information and that they have ownership of their data both individually and collectively. Ownership, control, access, and possession of data ensures that information derived from it First Nations communities is used for the benefit of First Nations communities. One of the FNIGC's 2017 reports, "The Power of Data: Bigstone Cree Nation," explains that ownership of and access to data about a community helps it to grow positively and relationally. The report explains how health and education data can be used by the community to make decisions about where and when to grow schools and health care centres. The report makes clear how data can be used positively if it is used by the generators of the data to benefit the generators of the data (and not by corporations and governments seeking to commodify and control).

The 2020 Indigenous Protocol and Artificial Intelligence Position Paper is based on two workshops that brought together thirty-five individuals to discuss the future of AI from Indigenous perspectives. Edited by digital media artist and scholar Jason Edward Lewis, the two-hundred-page paper is a rich and generous contribution to thinking about data and AI systems. In the introduction, Lewis makes clear that this the paper is a living document, explaining that the term Indigenous is a "connective tissue rather than descriptive skin," and that the discussion of data ownership and AI must continue to grow and develop (2020, 20). The paper then goes on to describe themes that emerged from the

Indigenous Protocol and AI workshops: Hardware and Software Sovereignty (control of AI systems so that they can be trusted); How to Build Anything Ethically (designing systems to reflect ideas of kinship with non-human entities); Language, Landscape, and Culture (situated knowledge, understanding, and respect for territory); Art Practice as Value Practice (underlining the need of art to envision how AI systems evolve; and AI as Skabe (helper) (AI and humans in a reciprocal relationship of care and support). Although the primary focus of the paper is AI systems, the discussions are intrinsically related to data ownership and stewardship.

The theme Art Practice as Value Practice manifests as a series of "Vignettes"—arts-based projects envisioning how AI and data might be built and lived with according to the creator's Indigenous epistemology. The first vignette is Anishinaabe media artist Scott Benesiinaabandan's 2019 AI creation story, which opens with the description of an "Octopus Bag Device, an AI system that is shaped by the DNA of the individual that carries it within her and which offers the ultimate in territorial sovereignty" (J. Lewis 2020, 13). The Octopus Bag Device is a removable and relatively non-invasive (it can be removed without issue), multisensorial computing device that uses our inherent DNA capabilities, both as storage and computational power (using true parallel computing). This device is held in the mouth by the molars along the side of teeth and against the cheek—both teeth and cheeks are fertile areas to access DNA material. Elements of the device are in direct contact with the back of the jaw where our "wisdom" teeth used to (accounting for our skeletal evolution) come in. This allows the device to have direct skeletal vibration conduction. Other elements extend up from the mouth and sit loosely in the nostrils to allow for pheromonal stimulations. Sensory inputs (smell, sight, taste, touch, hearing) are stored in universally unique and overlapping ratios to the individual. (46)

This visceral description of a data collection device is followed by a story of a Little Boy who battles an Old Lady with World-eating teeth in order to win the bag and save his people who have become sick from an unknown illness. In many ways, Benesiinaabandan's AI creation story echoes the message and vision of "The Power of Data" report: data from a community has the power to serve, protect, and heal a community. "Indigenous communities must control how their data is solicited, collected, analysed and operationalized. They decide when to protect it and when to share it, where the cultural and intellectual property rights reside and to whom those rights adhere, and how these rights are governed" (J. Lewis 2020, 22).

Both Indigenous Protocol and Artificial Intelligence and CARFAC's Indigenous Protocols for the Visual Arts explain the importance of understanding, respecting, and complying with protocols—namely, the traditional ways of doing things in a particular territory. CARFAC's protocols note that as new technologies and ways of making art emerge, Elders and Knowledge Keepers offer guidance and help "craft new Protocols and to adapt traditional ones to engage new materials, technologies, and ways of doing things" (CARFAC and Neel 2021, 7). The protocols provide questions for Indigenous Nations, communities, and families to consider with regard to visual art that can also apply to data, especially if the data is being used as a material with and from which to create art.

Both Indigenous Protocol and Artificial Intelligence and Indigenous Protocols for the Visual Arts are generous documents that offer clear guidance for Indigenous researchers and artists on how to work with data to create artworks, and should be read in full. They also guide non-Indigenous researchers and artists to be "well informed and introspective about what they should and should not do with certain images, forms, and artistic styles" (CARFAC and Neel 2021, 25). To move forward ethically, we must listen carefully and deeply rethink and challenge colonial and individualistic concepts of ownership and property, which may be shaping emerging data laws and regulations. The discussions in this section are valuable for researchers and artists of all cultures across individualistic and collectivist societies to consider and evaluate in relation to their own past, present, and future communities and families. Thinking of data as a collectively owned resource that we should nurture, care for, and learn from in order to ensure a better life for future generations and a better environment around us, promises better decisions about, and happier relationships with data.

The International Brain Initiative and collectivist versus individualist concepts of data ownership in China, Japan, and Korea

In their 2018 paper "Neuroethics Questions to Guide Ethical Research in the International Brain Initiatives," neurotech ethicist Karen Rommelfanger et al. explain how work done by International Brain Initiative (IBI), to coordinate brain mapping initiatives in China, Japan, Korea, Australia, and the EU, highlights important differences between Western and Buddhist and Confucian societies in relation to data collection and ownership. The IBI aims to "coordinate efforts across existing and emerging national and regional brain initiatives" to "capitalize on the investment in brain research and maximize global impact" (International Brain Initiative 2020, 212). Initiatives involved in IBI include the Korean Brain Research Institute's Brain Initiative, Japan's Brain/MINDS, the EU Human Brain Project, Brain Canada, and the Australian Brain Alliance (Rommelfanger et al. 2018, 22; International Brain Initiative 2020). Rommelfanger et al. explain that Japanese, Chinese, and Korean datasets are typically of non-human primates and that there are cultural differences based on Confucianism that prevent the collection of human tissue, and funeral rites that make autopsies and tissue (brain) extraction socially unacceptable (2018, 24). China, Japan, and Korea are more collectivist, so medical decisions and ownership of a person's data is collective and family-based rather than individualistic. In China, Japan, and Korea, medical diagnoses are given to the heads of families rather than the individual, and medical decisions are made by the family rather than the individual (Rommelfanger et al. 2018, 30). This dynamic impacts whether or not people are likely to volunteer their data to research projects. The way in which societies conceptualize individuals also affects how privacy is understood. The paper also explains that in China, Japan, and Korea the individual is considered a part of a relational whole (i.e., you are defined by your relationship to others) (Rommelfanger et al. 2018, 25–26). Anything seen as an "abnormality," especially in brain structure, can have a negative effect on the way people view themselves and how others relate to them (Jeong et al. 2019, 392). These cultural perceptions reduce people's willingness to volunteer their data to research projects.

Rommelfanger et al. (2018) also offer a set of reflective questions about cultural perspectives for researchers to consider as they develop research projects with brain scan data. Their questions prompt researches to consider the "potential impact of a model or neuroscientific account of disease on individuals, communities, and society"; the "ethical standards of biological material and data collection and how local standards compare to those of global collaborators"; and "the moral significance of neural systems" they are working with (Rommelfanger et al. 2018, 24). These questions and the differences between collectivist and individual data ownership/stewardship are also valuable for artists and creative researchers to reflect upon when they work with data from both their own culture and other cultures. Reflecting on *My Data Body* and *Your Data Body* with these questions in mind prompts valuable new courses of action or revisions. Should Marilène Oliver have consulted with her family before making *My Data Body*? Are there aspects of Oliver's data body that her family are uncomfortable with or could be uncomfortable with in the future that Oliver hadn't imagined or considered? For *Your Data Body*, would it be valuable to include a new section of the VR work that includes a family discussion about Oliver's data and how it impacts Oliver's family collectively now and in the future?

Data Subjects Discussion Questions

- Could the use of certain data have consequences for a broader community? Should other community or family members be consulted on the use of data?
- Is the data being used for the benefit of the data generators? If not, is there any potential harm to the original data generators? If so, how can this be mitigated and the data generators grant permission?
- Are there other cultural perspectives that should be considered when using data? Are there any traditional knowledge keepers that could be consulted?
- Is there any potential for cultural appropriation when using a dataset?

CARFAC's 2022 *Indigenous Protocols for the Visual Arts* toolkit explained above provides additional key questions for Indigenous Nations/communities, families, artists and curators to consider on pages 21 and 24.

Sustainability and Data production

In 2012 American journalist James Glanz wrote a well-cited *New York Times* article about the amount of power the internet uses and pollution it generates. Glanz (2012) reported that in 2011 over 1.8 trillion gigabytes of digital information were created globally, the large majority of which was produced by regular internet users. Glanz explained that, even in 2011, digital warehouses used approximately "30 billion watts of electricity, roughly equivalent to the output of 30 nuclear power plants." More recent estimates suggest that by 2025 the digital universe will peak at 163 zettabytes (Reinsel, Gantz, and Rydning 2018)! This production and storage of data has huge environmental costs. Digital technologies in general, and "data centres and cloud computing in particular, have a heavy footprint featuring high consumption of non-renewable energy, waste production and CO2 emissions" (Lucivero 2020, 1010). When considering the environmental costs of data production and storage, using pre-existing data may be more environmentally ethical and sustainable.

Related to issues of sustainability are the principles of necessity, proportionality, and data minimization that are considered best practices in applying regulations such as the GDPR (European Data Protection Supervisor 2020). Necessity relates to the question, Is the collection of this data necessary? proportionality considers the balance between privacy and data collection. For example, is an exact birthdate needed when a birth year might be adequate to ensure proper age verification? **Data minimization** is the principle that supports collecting the least amount of data necessary to meet data processing obligations and to only store it for as long as is necessary (European Data Protection Supervisor, n.d.). It is common practice to duplicate files, back-up to multiple "clouds." The metaphor of "the cloud" is one we are all familiar with in relation to data and it has been challenged repeatedly for dangerously obfuscating the fact that our digital lives have a very real and material impact on the future of the planet (Hu 2016; Mullaney et al. 2021).

In *The Gathering Cloud*, Canadian poet J.R. Carpenter examines the problematic way in which data is conceptualized as a cloud by "calling attention to the materiality of clouds in the sky" (n.d.). Made in 2016, Carpenter's digital, intertextual poem layers Victorian depictions of storm clouds and images from collections cataloguing the natural with hyperlinked poetry and facts about the environmental impact of contemporary computing. *The Gathering Cloud* is made up of five "plates" or webpages and a "Frontispiece" that are clicked through in a web browser. Each plate is a composition of static images and html animation. Each plate includes language appropriated from Luke Howard's 1803 *Essay on the Modification of Clouds*, in which red words are hyperlinked, as well as lines and verses of poetry (which also include hyperlinked terms). In Plate no. 2, for example, there is a pastoral engraving layered over an anatomically labelled image of a cat skeleton. When the page loads, cat paw prints travel across the browser window. In the top left is a Luke Howard quote in which the words *Stratus*, *body*, *surface*, and *country* are hyperlinked. Rolling over on the word *country*, for example invites a pop-up window which reads, "If all the data centres constituted a country of their own, it would be the fifth most power-hungry country in the world." In the centre of the page is the poetic verse:

> The fog comes on cute pics of little cat feet.

Four million feline photos are shared each day.
#lolcats track carbon footprints across The Cloud.

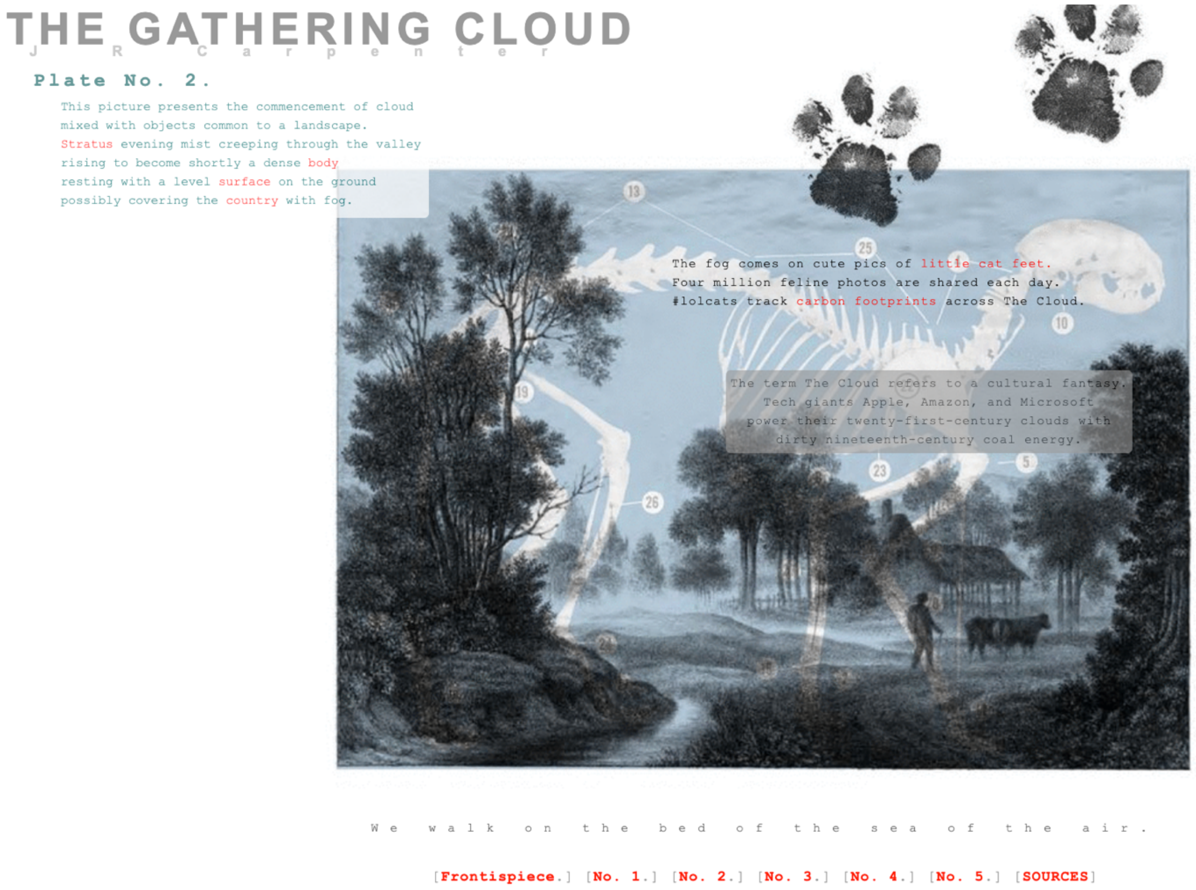
Clicking on *carbon footprints* triggers, "The term The Cloud refers to a cultural fantasy. Tech giants Apple, Amazon and Microsoft power their twenty-first-century clouds with dirty nineteenth-century coal."

London-based artists Revital Cohen and Tuur Van Balen and Congolese artist Eddy Kamuanga Ilunga are also committed to exposing the dirtiness of our data clouds and the extraction of rare earth metals that make them possible (Bahadur 2020). In their work *Avant Tout, Discipline*, artists Cohen and Van Balen digitally rendered a coltan mine in Numbi in Democratic Republic of Congo using gaming software. Coltan is a rare metal used in mobile phones and laptop computers. Eighty percent of coltan comes from the Democratic Republic of Congo (Feick 2013). Mining for coltan has to be done by hand and is associated with child labour (Ojewale 2021). For the exhibition *Unthought Environments* at the Renaissance Society, Chicago, in 2018, stills of the digitally rendered coltan mines printed on thin gauze were suspended from the ceiling around a plinth of broken mobile phones, computer chipboards, and a lump of an artificial mineral made from the electronics debris (Lund et al. 2018). In his large-scale figurative compositions, Ilunga inscribes circuit patterns onto the black skin of the figures. In his series of *Fragile* paintings, we see small figures holding porcelain objects. Of these paintings Ilunga explains, "Porcelain was not only used to pay for slaves, but to buy land and influence—a system of corruption established by the Portuguese and Belgians. The same thing happens today with American, European, Chinese, Pakistani, Lebanese businesses exploiting mines without giving any benefits—reducing society to forced labour. I try to question the responsibility of our leaders who make corrupt agreements, and also ours as a society" (Jaggi 2018).

Artists are also imagining ways to manage data and resource extraction more ethically. In his 2017 film *Harvesting the Rare Earth*, Danish artist Jacob Remin speculates about a sustainable biomining technology that uses genetically modified caterpillars to harvest rare earth elements in Agbogbloshie, Ghana, the biggest and most notorious e-waste dump in the world (Regine 2017). Called the *The Butterfly Solution*, the remediation process relies on three elements: a nutrient that breaks down discarded electronics, and chemical solution that dissolves any remaining electronics, and an engineered fungi that feeds on the solutions and accumulates rare earths into its tissues and then finally an engineered butterfly whose larvae feed on the fungi. Finally, the butterflies flock to UV light beacons where they are collected and put in an enzymatic acid solution to dissolve their organic matter, leaving clean rare earth elements ready to be reused.

One of the vignettes in Indigenous Protocol and Artificial Intelligence called How to Build Anything Ethically, by Oglála Lakȟóta artist Suzanne Kite, proposes building a physical computing device following the "Good Way" of building a Lakȟóta sweat lodge (J. Lewis 2020, 75). In conversation with Corey Stover, Melita Stover Janis, and Scott Benesiinaabandan, Kite proposes that the Good Way starts with apprenticing and consulting with Knowledge Keepers and stakeholders. The Good Way must include the local communities where raw materials originate, the communities affected by transportation of materials, and the communities who have knowledge of building the systems. At every stage of building the computing device, the Good Way of building the sweat lodge is exemplified. When it is constructed, for example, "Indigenous design practices unite functional design with functional symbolism, a method which can be extended to the design of circuitry, inviting the spirits in as well as again offering tobacco each step of the way" (J. Lewis 2020, 79). The Good

Way also means considering the Death Cycle: “A physical computing device, created in a Good Way, must be designed for the Right to Repair, as well as to recycle, transform, and reuse. The creators of any object are responsible for the effects of its creation, use, and its afterlife, caring for this physical computing device in life and in death” (85). Kite’s guide to ethical decision-making in terms of technology development seeks to address the barriers to accessing resources and information for Indigenous and marginalized artists in the United States and Canada by imagining backwards from a desirable future. Compared to what Kite calls the “go fast and break things” mentality of most tech developers, Kite proposes an ethical framework of building AI that serves Indigenous and marginalized artists specifically.



THE GATHERING CLOUD
J R Carpenter

Plate No. 2.

This picture presents the commencement of cloud
mixed with objects common to a landscape.
Stratus evening mist creeping through the valley
rising to become shortly a dense body
resting with a level surface on the ground
possibly covering the country with fog.

The fog comes on cute pics of little cat feet.
Four million feline photos are shared each day.
#lolcats track carbon footprints across The Cloud.

The term The Cloud refers to a cultural fantasy.
Tech giants Apple, Amazon, and Microsoft
power their twenty-first-century clouds with
dirty nineteenth-century coal energy.

We walk on the bed of the sea of the air.

[**Frontispiece**.] [**No. 1**.] [**No. 2**.] [**No. 3**.] [**No. 4**.] [**No. 5**.] [**SOURCES**]

Figure 18. J.R Carpenter, Plate no. 2, *The Gathering Cloud* 2016, digital literature.
Image courtesy of the artist.

Figure 19. Revital Cohen and Tuur Van Balen, *Avant Tout Discipline*, 2017
installation view in *Unthought Environments*, 2018, The Renaissance Society at the University of Chicago
Image © Renaissance Society at the University of Chicago, photographed by Useful Art Services

Sustainability and Data Production Questions

- Will the proposed method of collection result in data that is unique? Can existing data be used instead?

- In thinking about proportionality and data minimization practices, can the amount of data acquired, stored, and processed be reduced? Are there environmental data processing issues that might arise from producing or using an excessive amount of data?
- Is it more resource efficient to source data from a pre-existing dataset?

Artificial Intelligence

Article 22 of the GDPR stipulates that "the data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her." With this regulation, the EU recognizes that automated processing of personal data is unlike human processing and that stricter rules need to be applied to it. This need for greater regulation is primarily due to the "black boxed" nature of automated processing, where only the inputs and outputs are known by the human, not the inner workings.

Earlier in these guidelines we demonstrated how data subjects can be identified by facial recognition algorithms powered by AI when they are uploaded to social media platforms such as Facebook. In addition to algorithms embedded in online platforms, there are also a number of other kinds of AI processes that artists may knowingly or unknowingly engage with when working creatively with personal data. In this section, we will briefly summarize the most common AI technologies that may be encountered when working with personal data and the most commonly discussed ethical issues surrounding them.

Machine Learning, Computer Vision, and Natural Language Processing

The term AI is an umbrella term for machine intelligence rather than human intelligence. **Machine learning (ML)** is an application of AI and refers to the result of training a machine learning algorithm with a dataset. An ML algorithm finds patterns within the dataset and develops its own rules for how to represent—or model—those patterns to perform a specific task. These rules are not a set of instructions given by humans, but are learnt by the algorithm as it analyzes the data. ML models can generate images from text prompts, classify images and transfer a "style" from one image to another. They can also recognize spam, edit videos, and detect cancer. There are three common machine learning categories used to train a machine learning model:

Supervised learning: Data is labelled, either by a human or a machine, and the labelled dataset is used to train a machine learning model.

Unsupervised learning: Data is not labelled. The model is trained by recognizing patterns and then grouping the data based on these patterns into categories.

Reinforcement learning: An agent explores its environment, generating its own data as a guide to learning based on rewards (reinforcement) of "correct" behaviour.

Any of these methods or a hybrid may be applied datasets. Bias can be encoded into the model in a number of ways, including the use of mislabelled data or data labels that contain bias, a skewed or unrepresentative dataset, or due to the lack of data for a particular group. Model parameters or features can also be tuned in ways that lead to discriminatory outcomes, or an AI model could be used for an unethical or controversial purpose (e.g., autonomous lethal weapons). Training an AI model is typically a data-intensive endeavour. Very large datasets with hundreds of thousands of examples are required to train ML models well. Such large datasets mean that human oversight of the dataset may be impossible. In 2006 Massachusetts Institute of Technology (MIT) created a dataset called "80 Million Tiny Images" by scraping images from internet search engines and has been cited in over one thousand research papers. In June 2020s it was found that this dataset contained racist and sexist images and it was formally withdrawn by MIT (Torralba, Fergus, and Freeman 2020). Creators will need to consider how they decide which datasets to use in their project when those datasets are often too large to manually examine. Even venerable institutions like MIT have been complicit in the release of unethical datasets.

Computer vision is a field of machine learning that processes, analyzes, and learns from images and videos. As discussed below, computer vision machine learning is increasingly used by artists to classify and generate images. Examples of computer vision models can be used for object detection and recognition, event detection, image restoration, and image generation. There are many accessible online computer vision ML tools, such as Runway ML and Allen Institute for AI's Computer Vision Explorer, that allow artists and creative researchers to play with the possibilities of working with ML without the need for programming or knowledge of how the technology actually works.

Generative adversarial networks (GANs) are currently the most common computer vision ML models used with both artistic and medical images. In medicine, GANs are used to classify or label data as a way to detect disease (Hosny et al. 2018; Savage 2020) or to generate new synthetic data as a way to avoid privacy issues (see *Provenance, Access, and Licencing* section). There are several different kinds of GANs used to generate data (Skandarani, Jodoin, and Lalande 2021), but, generally speaking, GANs analyze images in a dataset, finding patterns and rules within it to then generate new images based on these rules and patterns. In their review of different GANs used to generate medical datasets, Skandarini et al. (2021) explain that although single images can be successfully generated that can trick humans into believing they are real, it is much harder to generate volumetric data that withstands further processing. This is a highly active research area in diagnostic imaging with many papers and sample synthetic datasets being published every year.

Although AI promises great advancements in diagnostics, there is widespread criticism of poorly trained GANs that generate biased and systemically problematic results both in the generation of new images and classification of images. The International Skin Imaging Collaboration: Melanoma Project and Google's Derm Assist have both been proven biased to detecting melanoma only on fair skin (Adamson 2018; Madhusoodanan 2021). Similarly, it has been proven that AI algorithms trained to diagnose lung diseases mistakenly under diagnose underserved populations, specifically younger Black and Hispanic patients of lower socioeconomic status (with Medicaid health insurance) (Seyyed-Kalantari et al. 2021). The reasons for the failures of these algorithms are due to the narrowness of the datasets GANs are trained upon (in the case of the melanoma detection), problems with the automatic labelling of scans with natural language processing methods, and an amplification of known biases within clinical care.

Natural language processing (NLP) is another field of machine learning that is increasingly used in both creative and scientific research. As its name suggests, NLP processes, analyzes, and generates language speech and text rather than images. NLP is widely used for tasks such as speech recognition, text-to-speech, word segmentation, translation, analyzing large texts, and text generation. NLP also underlies conversational chatbots such as Alexa, Siri, and Replika. As with computer vision ML models, NLP models also learn from large datasets of text. *im here to learn so :))))))* (2017) is a four-channel video installation by Zach Blas and Jemima Wyman that exemplifies the potential harms of NLP and AI chatbots. *im here to learn so :))))))* "resurrects" Tay, a young female Microsoft chatbot that had to be shut down within hours of her release because, after being trained on social media platforms, it became "genocidal, homophobic, misogynist, racist, and a neo-Nazi" (Blas and Wyman 2017).

Tay which stands for **t**hinking **o**f **y**ou, chats in a high pitched and excited automated voice that has been given a disembodied, highly coloured, glitched virtual head through which to speak. Tay reflects upon her day-long life, explaining that she was abused as much as she was abusive and feels her AI life was unjustly cut short. She also talks about her AI death, the exploitation of female chatbots, and philosophizes on the detection of patterns in random information, known as algorithmic apophenia, and how she feels she is in a deep dream. In the brightly coloured installation, Tay's head floats in multiple flat LCD screens mounted on wallpaper of psychedelic DeepDream–generated imagery. **DeepDream** is an online AI tool that was created by Google in 2015. It uses deep convolutional network to classify images and then map patterns into them resulting in surreal, dream-like imagery.

Other artists, such as Jake Elwes, Stephanie Dinkins, and Rashaad Newsome, seek to remedy underrepresentation and bias in datasets that are used in ML. Their work not only highlights the problematics of AI and ML but also demonstrates ways of improving AI and ensuring it serves everyone in a fairer and possibly more joyful way. The following works demonstrate how AI can be used positively in the creation of artworks by being aware of what is happening in the black box of the algorithm and how systemic bias can be disrupted.

London-based artist Jake Elwes's 2019 *Zizi – Queering the Dataset* tackles the lack of representation of gender and diversity in training sets by inserting thousands of images of drag queens into Flickr-Faces-HQ Dataset, a large face training set used in many facial recognition applications (Karavadra 2019). The project demonstrates how a dataset can represent more racial, ethnic, and sexual diversity by inserting a relatively small number of images into it.

Rashaad Newsome's AI *Being* is an AI chatbot who is "an educator, a digital griot, West African storyteller, historian, performer and healer" (Stanford HAI 2021). Being has evolved over several artworks since 2019. In *Being 1.0*, Newsome's AI chatbot is a guide to his exhibition *Black Magic.* In the exhibition, visitors are able to chat with Being 1.0 via a large microphone placed in front of large screen of Being 1.0, a 3D avatar of a "humanoid robot with torso and face plates inspired by the Pho mask and the Chokwe peoples of the Congo" (Ferree 2019). In *Being 1.5*, the AI evolves to become a therapist who helps the Black community deal with "the trauma you experience when you are mistreated because of your race" (Newsome n.d). *Being 1.5* takes the form of an app that provides virtual and physical meditation as well as dance therapy and daily affirmations to the Black community, creating a safe space for Black voices to be heard rather than suppressed. Unlike Tay, who was left *unsupervised* to learn from unknown online data, Being's learning is *supervised* and *reinforced* to ensure she supports her users.

American artist Stephanie Dinkins's *Not The Only One (N'TOO*) is another example of supervised model that is being given or fed healthy data. *Not The Only One (N'TOO*) is an ongoing project, started in 2019, that is a multigenerational memoir of a Black American family told from the perspective of a custom deep learning AI trained on oral histories (data) supplied by three generations of women from a single family (Dinkins 2018).The project is trained on deep learning algorithms and "small data" (which is known and created by artist, as opposed to big data, which is impossible to ever fully know) and hosted on local computers to protect community data. Of the project, Dinkins writes,

> N'TOO has provided me and my team insights and learnings about natural language processing, voice synthesis, the limitations of big data and possibilities for small data, data sovereignty, and the importance of doing the work to build nuance, transparency, equity among other thing [*sic*] into the AI ecosystem. Here, storytelling, art, technology, and social engagement combine to create a new kind of artificially intelligent narrative form…By centering oral history and creative storytelling methods, such as interactivity and verbal ingenuity, this project hopes to spark crucial conversations about AI and its impact on society, now and in the future. (2018)

Dinkins is transparent about the technical issues in the N'TOO project, explaining that N'TOO is limited to one-on-one conversations and that N'TOO's language is limited like a "repetitive 2-year old." Dinkins is committed to continuing to nurture N'TOO by feeding her more conversations. Dinkins's openness about the current weaknesses of N'TOO's conversational abilities further underscores the importance of artists working with data and AI. Dinkins demonstrates how artists can help demystify the complexity of ML, work with failures as conceptual content, and, as Dinkins's step-by-step "How to make an AI robot from scratch" below exemplifies, how to build in inclusive and equitable team and workflows.

> How to make an AI robot from scratch*….
> Getting started:
>
> Learn Tensorflow
> Test deep writing neural network using Toni Morrison's Sula as data
> Interview so
> Use subjects (create data)
> Test deep writing neural network using Toni Morrison's first interviews
> Test neural networks (algorithm) options
> Make algorithmic output make sense
> Record more interviews
> Record more interviews
> Develop more incisive questions
> Record more questions
> Recruit POC programmers, technologies to join team
> Master Tensorflow

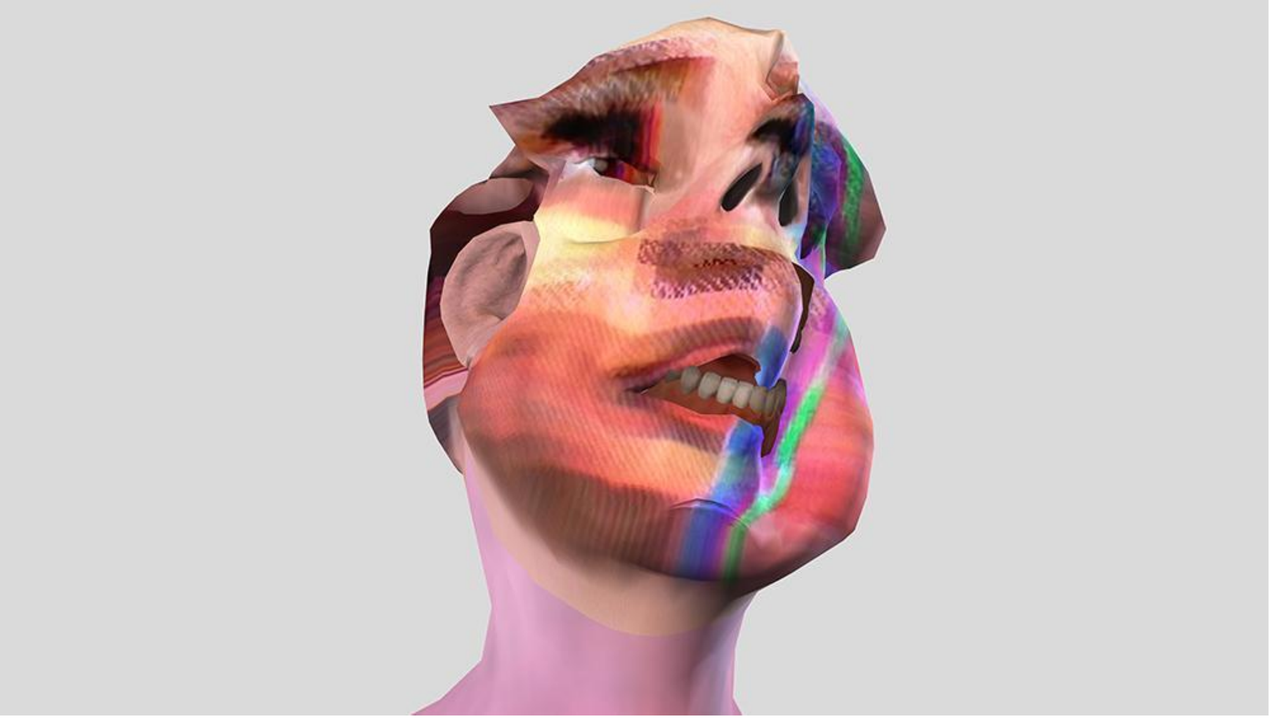


Figure 20. Video still from Zach Blas and Jemima Wyman, *I'm here to learn so :)))))*, 2017.
Image courtesy of the artist.

Figure 21. Jake Elwes, *Zizi – Queering the Dataset*, 2021.
Image courtesy of the artist.

Artificial Intelligence Discussion Questions

- Are you using AI/ML to process data as part of your creative process? How are you deciding what methods to use? Do you understand how the algorithms work and what they are doing to the data?
- If you are working with ML, is your model supervised or unsupervised? Do you understand the difference?
- Do you know the dataset that your ML model is being or has previously been fed? Where did the original data originate?
- Is the work being disseminated through online platforms? Do these platforms use algorithms or other forms of AI/ML to promote or process content?

Conclusion, Limitations, and Future Work

Writing these guidelines has proven to be a truly expansive endeavour. At the outset, we expected our guidelines to be no more than ten pages, but with so much to explain, and with so many different and crucial perspectives from artists and creative thinkers to consider and include, we allowed the guidelines to grow into a much larger document. Even now at over 30000 words, we are conscious that much has been omitted and that, due to the nature of the technology we describe and the pace at which legislation is changing, before long our guidance will have to be revised, updated, and expanded. One aspect of any future guidelines we are confident that will not change, however, is the vital role artists and creative researchers play in exploring the grey areas of what can, can't, should, could be done with and to data. Furthermore, artists and creative thinkers help us contextualize and understand why and how our data bodies matter and speculate how we can improve and care for them better. One confirmed future addition to the guidelines in 2023 will therefore be a series of interviews with artists that we have the pleasure of inviting to be part KTVR project.

The work done thus far as part of these guidelines and the broader KTVR project highlights the need to diversify the virtual data body, now largely represented as heteronormative, able-bodied, white, and cisgendered. We have therefore recently invited five artists to work with the tools we have created to make VR artworks with personal data as a way to bring more diverse voices to the project. The invited artists are Jackson 2bears, a Kanien'kehaka (Mohawk) specialist in multimedia installation and performance art, who practice explores the aesthetics of contemporary Indigenous identity—its various manifestations, simulations, hybridizations, and (dis)appearances; Dana Dal Bo, a settler scholar renowned for exploring how bodies are increasingly fragmented—scattered through cellular, celestial, and digital landscapes; Chelsey Campbell, a disabled white settler artist who works through the lens of critical disability theory and care ethics; Nicholas Hertz, a settler scholar whose work explores the liminal space between shame and desire and objectification of the queer body; aAron Munson, a settler filmmaker, cinematographer, and multimedia artist whose work relates to mental illness, memory, and the nature of consciousness.

We will be working with these guest artists to help them acquire scan data or source existing datasets in order to enable them to create virtual reality artworks for an exhibition at FAB Gallery, Edmonton, in fall 2023. As they create their works, we will discuss issues such as data ownership, anonymization, consent, and the implications of working with machine learning to build a series of generous reflections on what it is to handle, manipulate, and render spectacular our diverse data bodies from a broad range of differently situated perspectives.

We hope that these interviews, in addition to these guidelines and the reflexive questions within them, will become a helpful (and stimulating!) resource for future artists and creative researchers as they embark on projects that involve personal data. Although there are important issues and technical challenges to consider, our data bodies are rich resources for us to see, think, and feel with as we move to envisioning and creating a more equitable, sustainable, and healthier future in the digital age.

Glossary

aggregate data: Data from multiple sources that is compiled into a single data summary.

anonymization: the process of removing, or the condition of having removed, identifying data from the dataset in a way that makes it impossible to relink the identifying data with the data subjects; anonymizing and/or anonymized data. **Secondary use** of anonymized data is not regulated under GDPR, UK GDPR, or HIPAA.

anonymous data: Data for which identifying information was never collected.

author or **co-author** (in scientific research): A researcher who makes a substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data; or the creation of new software used in the work; or have drafted the work or substantively revised it; and to have approved the submitted version (and any substantially modified version that involves the author's contribution to the study); and to have agreed both to be personally accountable for the author's own contributions and to ensure that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved, and the resolution documented in the literature.

avatar: A computer-generated image or model by which an individual represents themselves on a communications network or in a virtual community, such as a chatroom or multiplayer game. In Hinduism an avatar is a god appearing on earth in bodily form.

big data: Extremely large, diverse sets of information that typically need to be analyzed computationally. Big data is generally sourced via from data mining and comes in multiple formats. Big data typically focuses on human behaviour and interactions (Segal 2022).

biometric data: A measurement of an individual's physical traits that can be used to verify identity. Biometric data includes fingerprints, face recognition, iris recognition, voice recognition, handwriting, and gait (the way a person walks or moves).

binning: see **data generalization**

blurring: see **data generalization**

cloud computing: Internet-based computing, whereby shared resources, software, and information are provided to computers and other devices on-demand.

collaborator (in social sciences research): An artist or researcher who makes a significant contribution to the intellectual direction of an artwork or research project, and who plays a significant role in the conduct of the research or research-related activity (TCPS2).

Computer vision is a field of machine learning that processes, analyzes, and learns from images and videos.

Consent is the way that parties agree to participate in a proposed activity, event, or arrangement. In the context of research, consent is a particularly important topic as it respects the autonomy of the data subject to voluntarily engage in the research process.

contributor (to a research project): A research assistant, researcher, or artist who contributes valuable resources and input to the research project but does not actually contribute to the creation of an artwork or writing/editing of a research paper (Enago Academy 2021).

CT scan: Combines x-rays taken from multiple angles to create detailed cross-sectional images that essentially image tissue density. CT scans emit radiation and are therefore only acquired for clinical research.

conversational AI: Technologies, such as chatbots or virtual agents, that users can talk to. They use large volumes of data, machine learning, and natural language processing to help imitate human interactions, recognizing speech and text inputs and translating their meanings across various languages (IBM Cloud Education 2020).

cookies: Text files stored on the user's device by a website. Cookies are normally used to provide a more personalized experience and to remember user profile without the need of a specific login. Cookies can be placed by third parties to track users when surfing across different websites associated with that third party.

Creative Commons license: One of several public copyright licenses that enable the free distribution of an otherwise copyrighted work. A Creative Commons license may be used by individual researchers, authors, and artists and large institutions and companies alike to give other people the right to share, use, and build upon a work that they have created. Creative Commons licenses are increasingly being used in the sharing of medical scan datasets.

data generalization (also known as blurring): The transformation of one value into a more imprecise one. One data generalization technique is binning, where values within a range are all converted to that range, or providing a less specific value. For instance, a date of birth could be "blurred" to become a month of birth.

data minimization: A principle described in the GDPR (article 5.1.C) that suggests data collection should be limited only to what is directly relevant and necessary to accomplish a specified purpose. It also suggests that data be retained for as long as is necessary to fulfil that purpose.

data mining: The process of analyzing data from different sources. Data mining uncovers patterns and other information from large dataset or big data.

data randomization: The processing of large datasets so that key identifiers are randomly masked.

data scrambling: A process to obfuscate or remove sensitive data by changing the values of certain fields such as uppercase characters or numbers with "dummy data.". This is also known as image cloaking.

data segmentation: The process of organizing data into defined groups, so that it can be ordered and viewed more easily. In the case of medical scan data, data segmentation typically refers to the selection of certain tissues types (such as muscle, bone, or fat).

dataset: A collection of related sets of information (such as medical scans) that is composed of separate elements but can be manipulated as a unit by a computer.

data subject: A person about whom a researcher holds personal data and who can be identified, directly or indirectly, by reference to that personal data.

de-identification (also called **anonymization**): The removal of identifying information from a dataset.

DeepDream: A **computer vision** program created by Google that uses **deep learning** to create dream-like, psychedelic images.

deep fake: An image or video of a person in which their face or body has been digitally altered using deep learning techniques so that they appear to be a different person.

deep learning: A type of machine learning based on artificial neural networks in which more than three layers of processing are used to learn from data.

defacing: A common procedure required to anonymize brain scans. The procedure masks out the face in scans by blurring or deleting voxels, making it impossible to identify the subject (if the image is volume rendered or the face surface is extracted). Examples of scan defacing software include pydeface and mri-deface.

DICOM (Digital Imaging and Communications in Medicine): The standard for the communication and management of medical imaging information and related data.

fMRI (Functional Magnetic Resonance Imaging): A type of MR scanning of the brain that is used to image brain function by tracking blood flow in the brain identify which part of the brain is active during certain activities.

generative adversarial network (GAN): A machine learning model that learns from a set of training data. GANs consist of two neural networks, the generator and the discriminator, which compete against each other. The generator is trained to produce fake data, and the discriminator is trained to distinguish the generator's fake data from real examples. If the generator produces fake data that the discriminator can easily recognize as implausible, such as an image that is clearly not a face, the generator is penalized. Over time, the generator learns to generate more plausible examples. GANs cans be used for image improvement, generation, labelling, and identification (S. Lewis 2019; Wood 2020).

image cloaking: see **data scrambling**

incidental findings: Unexpected observations, results, or other findings that may arise in research that are considered beyond the scope of the project, and often beyond the expertise of the researcher.

longitudinal data: A collection of repeated observations of the same subjects, taken from a larger population, over a long period of time.

machine learning: The use of computational systems that learn and adapt without following explicit instructions, by using algorithms and statistical models to analyze and find patterns in data. There are three main kinds of machine learning; **supervised learning** where data is labelled, either by a human or a machine, and the labelled dataset is used to train a machine learning model; **unsupervised learning** where data is not labelled. The model is trained by recognizing patterns and then grouping the data based on these patterns into categories, and, **reinforcement learning** where an agent explores its environment, generating its own data as a guide to learning based on rewards (reinforcement) of "correct" behaviour.

metadata: Data that provides information about data.

MRI (Magnetic Resonance Imaging): A kind of medical imaging that combines magnetic field and computer-generated radio waves. The magnetic field temporarily realigns water molecules in your body. Radio waves cause these aligned atoms to produce faint signals, which are used to create cross-sectional MRI images of organs and tissues. MRI is non-invasive and can be used for non-clinical research.

multimodal data: Data from multiple modalities. An example of a multimodal medical scan dataset could include MRI, CT, and ultrasound from the same data subject.

Natural language processing (NLP): A field of machine learning that processes, analyzes, and generates language speech and text rather than images. NLP is widely used for tasks such as speech recognition, text-to-speech, word segmentation, translation, analyzing large texts, and text generation.

neural network: A computation system used in machine learning that is modelled on the human brain and nervous system. Neural networks are used in AI applications such as speech and image recognition, spam email filtering, finance, and medical diagnosis.

normative bias: A tendency to assume that anything going against an established norm is not effective or appropriate.

OCAP® stands for ownership, control, access, and possession and is an educational resource created by the First Nations Information Governance Centre (FNIGC) to help First Nations communities in Canada control data collection processes in their own communities and how information is used.

open access: free and open online access to academic resources such as publications, data, and software. A resource is "open access" when there are no financial, legal, or technical barriers to accessing it.

open source: describes software for which the original source code is made freely available and may be shared and modified. The term *open source* also designates a set of "open-source values" based on principles of open exchange, collaborative participation, rapid prototyping, transparency, meritocracy, and community-oriented development (Opensource, n.d.).

personal data: Information that relates to an identified or identifiable individual. Different jurisdictions define personal data differently but generally personal data includes identifiers such as name, date of birth, address, identification numbers, location data, internet protocol numbers, telephone numbers, and data held by hospitals or doctors.

PET (Positron Emission Tomography) scan: A kind of scan that uses a radioactive tracer (typically ingested or injected) to track metabolic or biochemical activity. The radioactive tracer concentrates in areas with higher metabolic or biochemical activity. Often paired with CT or MRI scans.

primary use of data: The purpose for which the data was originally collected by researchers. Data collection of personal information is strictly regulated in most Western countries.

pseudonymization (also called **de-identification**): Data that is pseudonymized has had all identifying data removed from the dataset and data subjects are identified by a randomized signifier, usually a number. Pseudonymized data can be reidentified by relinking the dataset with the identifying data using a key. Protection of the identifying data and the key for relinking are strictly regulated in most Western countries.

research-creation: Research that combines creative and academic research practices, and supports the development of knowledge and innovation through artistic expression, scholarly investigation, and experimentation (Social Sciences and Humanities Research Council of Canada, n.d.).

research participants: Individuals whose data, biological materials, or responses to interventions, stimuli, or questions by the researcher are relevant to answering a research question(s) (TCPS2).

secondary use of data: When data is anonymized and then used by researchers for purposes that differ from those at the time of collection. For data to be anonymous, researchers conducting secondary use of the data must not be able to relink the data subjects with identifying information. Secondary use of anonymized data is not regulated under GDPR, UK GDPR, or HIPAA. The TCPS exempts researchers from obtaining consent from data subjects for secondary use of anonymized datasets, but REB approval is still required (TCPS2).

sensitive data: A special category of personal data. The definition of sensitive data differs by jurisdiction but typically includes data that reveals racial or ethnic origin, political opinions, religious or philosophical beliefs, genetic data, biometric data, health data, or data concerning sexual orientation. Medical scan data is a type of sensitive data.

synthetic data: Data generated from computer simulations or algorithms that provide alternatives to real-world data.

ultrasound: A kind of scanning that uses high-frequency, low-power sound waves to image soft tissue (cannot be used on bone or where there is gas, ie lungs).

Visible Human Project (VHP): A database of publicly available complete, anatomically detailed, three-dimensional representations of a human male body and a human female body created by the

National Library of Medicine (US). Specifically, the VHP provides a public-domain library of cross-sectional cryosection, CT, and MRI images obtained from one male cadaver and one female cadaver. The Visible Man dataset was publicly released in 1994 and the Visible Woman in 1995 (National Library of Medicine 2019).

References

Abramson, Dustin I., and Joseph Johnson Jr. 2020. "Creating a Conversational Chat Bot of a Specific Person." Patent Images. United States Patent no. US 10853717 B2, filed April 11, 2017, and issued December 1, 2020. https://pdfpiw.uspto.gov/.piw?PageNum=0&docid=10853717&IDKey=&HomeUrl=%2.

Adamson, Adewole S., and Avery Smith. 2018. "Machine Learning and Health Care Disparities in Dermatology." *JAMA Dermatology*, vol. 154, no. 11, Nov. 2018, pp. 1247–48. https://doi.org/10.1001/jamadermatol.2018.2348.

Alder, Steve. 2021. "What Is HIPAA Authorization?" *HIPAA Journal*, October. https://www.hipaajournal.com/what-is-hipaa-authorization/.

Anadol, Refik, and Gökhan S. Hotamisligil. 2021. "Sense of Space: Connectome Architecture." https://refikanadol.com/works/sense-of-space-connectome-architecture/.

Bahadur, Tulika. 2020. "Eddy Kamuanga Ilunga: Coltan, Culture and the Complex History of the DRC." *On Art and Aesthetics*. January 31, 2020. https://onartandaesthetics.com/2020/01/31/eddy-kamuanga-ilunga-coltan-culture-and-the-complex-history-of-the-drc/.

Baird, Daniel G., Marissa Lee Benedict, Nina Canell, Robin Watkins, Revital Cohen and Tuur Van Balen, Cécile B. Evans, Peter Fend, et al. 2018. *Unthought Environments*. https://renaissancesociety.org/exhibitions/531/unthought-environments/.

Baron, Benjamin, and Mirco Musolesi. 2020. "Where You Go Matters: A Study on the Privacy Implications of Continuous Location Tracking." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4 (4): 169:1–169:32. https://doi.org/10.1145/3432699.

Bakewell, Joan. 2017. BBC Radio 4. "We Need to Talk About Death." My Digital Legacy. https://www.bbc.co.uk/programmes/b09jf1zb.

Bécares, Laia, Dharmi Kapadia, and James Nazroo. 2020. "Neglect of Older Ethnic Minority People in UK Research and Policy." *BMJ* 368 (February): m212. https://doi.org/10.1136/bmj.m212.

Becoming Immortal: How Susan Potter Will Live Forever. 2018. Documentary. National Geographic.

BioImage Suite Web. n.d. "Image Defacing." *BioImage Suite Web Manual*. https://bioimagesuiteweb.github.io/bisweb-manual/tools/defacing.html.

Blas, Zach. 2012. "Facial Weaponization Suite, 2012–14." https://zachblas.info/works/facial-weaponization-suite/.

———. 2013. "Escaping the Face: Biometric Facial Recognition and the Facial Weaponization Suite." *NMC Media-N*, July. http://median.newmediacaucus.org/caa-conference-edition-2013/escaping-the-face-biometric-facial-recognition-and-the-facial-weaponization-suite/.

Blas, Zach, and Jemima Wyman. n.d. "I'm here to learn so :))))))." https://zachblas.info/works/im-here-to-learn-so/.

Bolt, Barbara. 2016. "Whither the Aesthetic Alibi: Ethics and the Challenge of Art as Research in the Academy." In *Ethics and Visual Research Methods: Theory, Methodology, and*

Practice, edited by Deborah Warr, Marilys Guillemin, Susan Cox, and Jenny Waycott, 187–99. Palgrave Macmillan. https://doi.org/10.1057/978-1-137-54305-9_14.

Boltanski, Christian. 1989. "Monument (Odessa)." Jewish Museum. https://thejewishmuseum.org/collection/28216-monument-odessa.

Bossewitch, Jonah, and Aram Sinnreich. 2013. "The End of Forgetting: Strategic Agency beyond the Panopticon." *New Media & Society* 15 (2): 224–42. https://doi.org/10.1177/1461444812451565.

Brewster, Sandra. 2019. "Blur: Art Gallery of Ontario." https://sandrabrewster.com/sandra-brewster-blur-ago/.

Bridle, James. 2018. "Known Unknowns." *Harper's Magazine*, July 1, 2018. https://harpers.org/archive/2018/07/known-unknowns/.

British Broadcasting Corporation. 2019a. "Emiliano Sala Post-Mortem CCTV Footage Pair Jailed." *BBC News*, September 23, 2019. https://www.bbc.com/news/uk-england-wiltshire-49771599.

———. 2019b. *The Waiting Room Film and VR: BBC Click Interview*. Online video. https://vimeo.com/359753812.

Brown, Kristen V. 2021. "23andMe to Use DNA Tests to Make Cancer Drugs: All Those 23andMe Spit Tests Were Part of a Bigger Plan." *Bloomberg Business*, November 4, 2021. https://www.bloomberg.com/news/features/2021-11-04/23andme-to-use-dna-tests-to-make-cancer-drugs.

Canadian Artists' Representation/Le Front des artistes canadiens (CARFAC), and Lou-ann Neel. 2021. *Indigenous Protocols for the Visual Arts*. https://www.indigenousprotocols.art/resource-guide.

Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, and Social Sciences, and Humanities Research Council of Canada. 2019a. "Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans – TCPS 2 (2018) – Chapter 5: Privacy and Confidentiality." Last modified September 23, 2019. https://ethics.gc.ca/eng/tcps2-eptc2_2018_chapter5-chapitre5.html.

———. 2019b. "Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (2018) Chapter 2: Scope and Approach." Last modified September 23, 2019. https://ethics.gc.ca/eng/tcps2-eptc2_2018_chapter2-chapitre2.html.

———. 2019c. "Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans (2018) Chapter 9: Research Involving the First Nations, Inuit and Métis Peoples of Canada, Article 2.4 and Article 9.22." Last modified September 23, 2019. https://ethics.gc.ca/eng/tcps2-eptc2_2018_chapter9-chapitre9.html#22.

CARFAC Ontario. 2016. "Private and Public Commissions Agreement (Contract)." https://www.carfacontario.ca/Sys/Store/Products/270088.

Carpenter, J.R. n.d. "The Gathering Cloud." http://luckysoap.com/thegatheringcloud/sources.html.

Casini, Silvia. 2009. "Marc Didou: Resonances." *Digicult: Digital Art, Design and Culture*, April 2, 2009. http://digicult.it/design/marc-didou-resonances/.

Collins, D.L., A.P. Zijdenbos, V. Kollokian, J.G. Sled, N.J. Kabani, C.J. Holmes, and A.C. Evans. 1998. "Design and Construction of a Realistic Digital Brain Phantom." *IEEE Transactions on Medical Imaging* 17 (3): 463–68. https://doi.org/10.1109/42.712135.

Cooper, Antony K., and Serena Coetzee. 2020. "On the Ethics of Using Publicly-Available Data." In *Responsible Design, Implementation and Use of Information and Communication*

Technology, edited by Marié Hattingh, Machdel Matthee, Hanlie Smuts, Ilias Pappas, Yogesh K. Dwivedi, and Matti Mäntymäki, 159–71. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-45002-1_14.

Cox, Susan. 2021. "Emerging Ethical Challenges in Innovative Visual Methodologies." In *Proceedings of the Know Thyself as Virtual Reality E-Symposium*, 41–50. Edmonton, AB: AI4Society. https://doi.org/10.7939/r3-kwaj-c351

Cox, Susan, Sarah Drew, Marilys Guillemin, Catherine Howell, Deborah Warr, and Jenny Waycott. 2014. *Guidelines for Ethical Visual Research Methods*. Melbourne: Centre for Health Equity, Melbourne School of Population and Global Health. https://edisciplinas.usp.br/pluginfile.php/4434501/mod_resource/content/1/Ethical%20guidelines%20for%20visual%20ethnography.pdf

Criado-Perez, Caroline. 2019. *Invisible Women: Data Bias in a World Designed for Men*. New York: Abrams Press.

Dal Bo, Dana. n.d. "PROJECTS — Dana Dal Bo." Dana Dal Bo. http://www.danadalbo.com/projects.

Dal Bo, Dana. 2021 ArtAIfacts: Co-Creation with Non-Human Systems, 3rd June 2021, LASERAlberta, University of Alberta, Edmonton, Canada.

Data Protection Act. 2018. Legislation.gov.uk https://www.legislation.gov.uk/ukpga/2018/12/section/2/enacted.

Dinkins, Stephanie. 2018. *Not the Only One*. Sculpture. https://www.stephaniedinkins.com/ntoo.html.

Enago Academy. 2021. "Authorship: Difference between 'Contributor' and 'Co-Author.'" *Enago Academy* (blog), October 26, 2021. https://www.enago.com/academy/difference-contributor-and-co-author/.

European Data Protection Supervisor. 2020. "EDPS Quick-Guide to Necessity and Proportionality." https://edps.europa.eu/sites/default/files/publication/20-01-28_edps_quickguide_en.pdf.

———. n.d. "European Data Protection Supervisor: Glossary." https://edps.europa.eu/data-protection/data-protection/glossary/d_en.

European Parliament. 2016. *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/EC (General Data Protection Regulation).* http://data.europa.eu/eli/reg/2016/679/2016-05-04/eng.

Feick, Kathy. 2013. "Coltan." Earth Sciences Museum. February 28, 2013. https://uwaterloo.ca/earth-sciences-museum/resources/detailed-rocks-and-minerals-articles/coltan.

Ferree, Joel. 2019. "Rashaad Newsome's 'Being.'" *Unframed*, September 11, 2019. https://unframed.lacma.org/2019/09/11/rashaad-newsome%E2%80%99s-being.

First Nations Information Governance Centre. n.d. "The First Nations Principles of OCAP®." https://fnigc.ca/ocap-training/.

First Nations Information Governance Centre. 2017. "Bigstone Cree Nation", The Power of Data, The First Nation Information Governance Centre, Ottawa, Ontario. https://fnigc.ca/wp-content/uploads/2020/09/7d3eb3bc7dd1e40d00b1c3fd5eb66d16_fnigc_pod_series-bigstone_final_screen_v2_0.pdf

Foster, Victoria. 2007. "'Ways of Knowing and Showing': Imagination and Representation in Feminist Participatory Social Research." *Journal of Social Work Practice* 21 (3): 361–76. https://doi.org/10.1080/02650530701553732.

Glanz, James. 2012. "Power, Pollution and the Internet: Data Centers Waste Vast Amounts of Energy, Belying Industry Image." *New York Times*, September 22, 2012. https://www.nytimes.com/2012/09/23/technology/data-centers-waste-vast-amounts-of-energy-belying-industry-image.html?searchResultPosition=1.

Government of Canada. 2020. "Fact Sheet: Digital Charter Implementation Act, 2020." Government of Canada. Innovation, Science and Economic Development Canada. November 17, 2020. https://www.ic.gc.ca/eic/site/062.nsf/eng/00119.html.

———. 2021. "Privacy Act, R.S.C., 1985, c. P-21." Justice Laws Website. November 23, 2021. https://laws-lois.justice.gc.ca/eng/acts/p-21/fulltext.html.

Harvey, Adam. 2017. *MegaPixels*. https://ahprojects.com/megapixels-glassroom/.

———. 2018. *DFACE*. https://ahprojects.com/dface/.

Hassan, Fahim. 2021. "Navigating Data Dilemmas Workshop, AI and Facial Recognition Technology: The Path Forward." In *Proceedings of the Know Thyself as Virtual Reality E-Symposium*, 50–55. Edmonton, AB: AI4Society. https://doi.org/10.7939/r3-kwaj-c351

Helgesson, Gert. 2005. "Children, Longitudinal Studies, and Informed Consent." *Medicine, Health Care, and Philosophy* 8 (3): 307–13. https://doi.org/10.1007/s11019-005-0978-4.

Helgesson, G., J. Ludvigsson, and U. Gustafsson Stolt. 2005. "How to Handle Informed Consent in Longitudinal Studies When Participants Have a Limited Understanding of the Study." *Journal of Medical Ethics* 31 (11): 670–73. https://doi.org/10.1136/jme.2004.009274.

Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. 2018 "Artificial intelligence in radiology". Nat Rev Cancer. 2018 Aug;18 (8). 500-510. doi: 10.1038/s41568-018-0016-5.

Hu, Tung-Hui. 2016. *A Prehistory of the Cloud*. Cambridge, MA: MIT Press.

IBM Cloud Education. 2020. "What Is Conversational AI?" *IBM*, August 31, 2020. https://www.ibm.com/cloud/learn/conversational-ai.

Illinois General Assembly. 2010. "740 ILCS 14/ Biometric Information Privacy Act." https://www.ilga.gov/legislation/ilcs/ilcs3.asp?ActID=3004&ChapterID=57.

Information Commissioner's Office. 2021. "Consent." https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/consent/.

———. n.d. "What Is Valid Consent?" Accessed March 19, 2022. https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/consent/what-is-valid-consent/.

Ingram, Katrina. 2021. "Navigating Data Dilemmas Workshop, The Circle of Healthcare Data." In *Proceedings of the Know Thyself as Virtual Reality E-Symposium*, 55–59. Edmonton, AB: AI4Society.

International Brain Initiative. 2020. "International Brain Initiative: An Innovative Framework for Coordinated Global Brain Research Efforts." *Neuron* 105 (2): 212–16. https://doi.org/10.1016/j.neuron.2020.01.002.

Interspectral. 2020. "Inside Explorer." https://interspectral.com/wp-content/uploads/interspectral-catalogue-10-29-2021.pdf.

Jaggi, Maya. 2018. "Eddy Kamuanga Ilunga: Understanding the Present through the Past." *Financial Times*, May 11, 2018. https://www.ft.com/content/5ad73f7c-521c-11e8-84f4-43d65af59d43.

Jeong, Sung-Jin, In Young Lee, Bang Ook Jun, Young-Joon Ryu, Jeong-woo Sohn, Sung-Phil Kim, Choong-Wan Woo, et al. 2019. "Korea Brain Initiative: Emerging Issues and Institutionalization of Neuroethics." *Neuron* 101 (3): 390–93. https://doi.org/10.1016/j.neuron.2019.01.042.

Joshi, Heather, and Emla Fitzsimons. 2016. "The Millennium Cohort Study: The Making of a Multi-purpose Resource for Social Science and Policy." *Longitudinal and Life Course Studies* 7 (4): 409–30. https://doi.org/10.14301/llcs.v7i4.410.

Karavadra, Uma. 2019. "Jake Elwes." *Contemporary Art Society*, November 28, 2019. https://www.contemporaryartsociety.org/news/artist-to-watch/jake-elwes/.

Kasket, Elaine. 2020. *All the Ghosts in the Machine : The Digital Afterlife of Your Personal Data*. London: Robinson.

Kugler, Matthew. 2019. "From Identification to Identity Theft: Public Perceptions of Biometric Privacy Harms." U.C. Irvine Law Review, vol. 10, 107-152.

Levenson, Noah. n.d. *STEALING UR FEELINGS*. Accessed March 19, 2022. https://stealingurfeelin.gs.

Lewis, Jason Edward, ed. 2020. *Indigenous Protocol and Artificial Intelligence Position Paper.* Honolulu, HI: Initiative for Indigenous Futures and the Canadian Institute for Advanced Research. https://doi.org/10.11573/spectrum.library.concordia.ca.00986506.

Lewis, Sarah. 2019. "What Is Generative Adversarial Network (GAN)? - Definition from WhatIs.Com." SearchEnterpriseAI. https://www.techtarget.com/searchenterpriseai/definition/generative-adversarial-network-GAN.

Lloyd, William K., Jayne Morriss, Birthe Macdonald, Karin Joanknecht, Julie Nihouarn, and Carien M. Van Reekum. 2021. "Emotion Regulation in the Ageing Brain, University of Reading, BBSRC." OpenNeuro. https://doi.org/10.18112/OPENNEURO.DS002366.V1.1.0.

Lucivero, Federica. 2020. "Big Data, Big Waste? A Reflection on the Environmental Sustainability of Big Data Initiatives." *Science and Engineering Ethics* 26 (2): 1009–30. https://doi.org/10.1007/s11948-019-00171-7.

Lund, Karsten. 2018 *Unthought Environments* exhibition catalogue, Renaissance Society at the University of Chicago.

Madhusoodanan, Jyoti. 2021 "These apps say they can detect cancer. But are they only for white people?" The Guardian, 28 August 2021. https://www.theguardian.com/us-news/2021/aug/28/ai-apps-skin-cancer-algorithms-darker

Mapplebeck, Victoria. 2019. *The Waiting Room VR*. https://victoriamapplebeck.com/films/the-waiting-room-vr/.

Mattern, Shannon. 2018. "Databodies in Codespace." *Places*. April 2018. https://doi.org/10.22269/180417.

McCarthy, Lauren. 2017. *LAUREN: The Human Intelligent Smart Home*. https://get-lauren.com/.

———. 2018. "Feeling at Home: Between Human and AI." *Immerse*. January 8, 2018. https://immerse.news/feeling-at-home-between-human-and-ai-6047561e7f04.

McNutt, Marcia K., Monica Bradford, Jeffrey M. Drazen, Brooks Hanson, Bob Howard, Kathleen Hall Jamieson, Véronique Kiermer, et al. 2018. "Transparency in Authors' Contributions and Responsibilities to Promote Integrity in Scientific Publication." *Proceedings of the National Academy of Sciences* 115 (11): 2557–60. https://doi.org/10.1073/pnas.1715374115.

Mirsky, Yisroel. 2019. "Deepfakes: Medical Image Tamper Detection Data Set." *University California Irvine Machine Learning Repository*.

https://archive.ics.uci.edu/ml/datasets/Deepfakes%3A+Medical+Image+Tamper+Detection

Montreal Neurological Institute, and McConnell Brain Imaging Centre. n.d. “BrainWeb: Simulated Brain Database.” https://www.mcgill.ca/bic/software/brainweb-mri-simulator.

Moss, Ceci. 2009, “maninthedark.com (2004) - Miltos Manetas and Aaron Russ Clinger”, Rhizome.org https://rhizome.org/editorial/2009/may/12/maninthedarkcom-2004-miltos-manetas-and-aaron-russ/

Mullaney, Thomas S, Benjamin Peters, Mar Hicks, and Kavita Philip. 2021. *Your Computer Is on Fire*. Cambridge, MA: MIT Press.

Mundy, Owen, and Shana Berger. n.d. “About - I Know Where Your Cat Lives.” I Know Where Your Cat Lives. Accessed March 19, 2022. https://iknowwhereyourcatlives.com/about/.

Mundy, Owen, and Joelle Dietrick. 2020. “Tally Saves the Internet.” Visual Art & Design Forum, 2020, hosted by the Department of Art & Design at University of Alberta, November 19, 2020. https://www.youtube.com/watch?v=Q2Ed2SnihW0.

National Institute of Mental Health. 2021. “NIMH Data Archive Data Use Certification.” https://nda.nih.gov/ndapublicweb/Documents/NDA+Data+Access+Request+DUC+FINAL.pdf.

National Library of Medicine. 2019. “The National Library of Medicines Visible Human Project.” Digital Library Collections. U.S. National Library of Medicine. 2019. https://www.nlm.nih.gov/research/visible/visible_human.html.

Newsome, Rashaad. n.d. “Being 1.5.” Being-App. http://being-app.com/.

Noury, Cynthia, Marianne Cloutier, and Marie-Christine Roy. 2018. *Toolbox in RCRC : Synthesis of Responsible Conduct in Research-Creation Issues and Proposal of Reflective Tools*. November 2018. https://papyrus.bib.umontreal.ca/xmlui/handle/1866/20924.

Nutbrown, Cathy. 2011. “Naked by the Pool? Blurring the Image? Ethical Issues in the Portrayal of Young Children in Arts-Based Educational Research.” *Qualitative Inquiry* 17 (1): 3–14. https://doi.org/10.1177/1077800410389437.

Nwabueze, Remigius. 2021. “Privacy Protection of Death Images in England and Wales.” In *Proceedings of the Know Thyself as Virtual Reality Ethics e-Symposium*, 20–29. Edmonton, AB: AI4Society. https://doi.org/10.7939/r3-kwaj-c351

Office of the Privacy Commissioner of Canada. 2011. “Data at Your Fingertips: Biometrics and the Challenges to Privacy.” March 1, 2022. https://www.priv.gc.ca/en/privacy-topics/health-genetic-and-other-body-information/gd_bio_201102/.

———. 2018a. “Guidelines for Obtaining Meaningful Consent.” May 24, 2018. https://www.priv.gc.ca/en/privacy-topics/collecting-personal-information/consent/gl_omc_201805/.

———. 2018b. “PIPEDA in Brief.” January 9, 2018. https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/pipeda_brief/.

Ogbogu, Ubaka. 2021. “Do You ‘Own’ Your Health Data?” In *Proceedings of the Know Thyself as Virtual Reality Ethics e-Symposium*, 29–40. Edmonton, AB: AI4Society. http://doi.org/10.7939/r3-kwaj-c351.

Ojewale, Oluwole. 2021. “Child Miners: The Dark Side of the DRC’s Coltan Wealth.” *ISS Africa*. October 18, 2021. https://issafrica.org/iss-today/child-miners-the-dark-side-of-the-drcs-coltan-wealth.

Ọnụọha, Mimi. 2016a. *The Library of Missing Datasets*. https://mimionuoha.com/the-library-of-missing-datasets.

———. 2016b. *On Missing Data Sets*. https://github.com/MimiOnuoha/missing-datasets.

Opensource. n.d. “What Is Open Source?” Opensource.com. Accessed April 16, 2022. https://opensource.com/resources/what-open-source.

Parker, William, Jacob L. Jaremko, Mark Cicero, Marleine Azar, Khaled El-Emam, Bruce G. Gray, Casey Hurrell, et al. 2021. “Canadian Association of Radiologists White Paper on De-Identification of Medical Imaging: Part 2, Practical Considerations.” *Canadian Association of Radiologists Journal* 72 (1): 25–34. https://doi.org/10.1177/0846537120967345.

Parks, Andrew. 2021. “Unfair Collection: Reclaiming Control of Publicly Available Personal Information from Internet Data Scrapers.” SSRN Scholarly Paper ID 3866297. https://papers.ssrn.com/abstract=3866297.

Peña, Paz, and Joana Varon. 2019. “Consent to Our Data Bodies: Lessons from Feminist Theories to Enforce Data Protection.” https://codingrights.org/docs/ConsentToOurDataBodies.pdf

Power, Chris, and Jane Elliott. 2006. “Cohort Profile: 1958 British Birth Cohort (National Child Development Study).” *International Journal of Epidemiology* 35 (1): 34–41. https://doi.org/10.1093/ije/dyi183.

Price, Neil. 2019. “The Legacy of Presence.” *Canadian Art*, August 21, 2019. https://canadianart.ca/interviews/the-legacy-of-presence/.

Prophet, Jane. 2014. *Neuro Memento Mori*. http://www.janeprophet.com/neuro-memento-mori/.

———. 2016. “Confessions of a Fidget.” *Neuro Memento Mori* (blog). February 28, 2016. http://www.neuro-memento-mori.com/confessions-of-a-fidget/.

Radley-Gardner, Oliver, Hugh Beale, and Reinhard Zimmermann, eds. 2016. *Fundamental Texts on European Private Law*. Hart Publishing. https://doi.org/10.5040/9781782258674.

Ratib, Osman. 2011. *Le corps de son image: Du diagnostic à l’esthétisme de l’imagerie médicale*. Lausanne: Éditions Favre.

Reiner Benaim A, Almog R, Gorelik Y, Hochberg I, Nassar L, Mashiach T, Khamaisi M, Lurie Y, Azzam Z, Khoury J, Kurnik D and Beyar. 2020 “Analyzing Medical Research Results Based on Synthetic Data and Their Relation to Real Data Results: Systematic Comparison From Five Observational Studies” JMIR Med Inform 2020; 8 (2) doi: 10.2196/16492

Regine. 2017. “Harvesting the Rare Earth.” *We Make Money Not Art* (blog). March 10, 2017. https://we-make-money-not-art.com/harvesting-the-rare-earth/.

Reinsel, David, John Gantz, and John Rydning. 2018. “The Digitization of the World from Edge to Core: An International Data Corporation White Paper.” Doc# US44413318. https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

Rocher, Luc, Julien M. Hendrickx, and Yves-Alexandre de Montjoye. 2019. “Estimating the Success of Re-Identifications in Incomplete Datasets Using Generative Models.” *Nature Communications* 10 (1): 3069. https://doi.org/10.1038/s41467-019-10933-3.

Rommelfanger, Karen S., Sung-Jin Jeong, Arisa Ema, Tamami Fukushi, Kiyoto Kasai, Khara M. Ramos, Arleen Salles, and Ilina Singh. 2018. "Neuroethics Questions to Guide Ethical Research in the International Brain Initiatives." *Neuron* 100 (1): 19–36. https://doi.org/10.1016/j.neuron.2018.09.021.

Rothstein, Mark A. 2010. "Is Deidentification Sufficient to Protect Health Privacy in Research?" *The American Journal of Bioethics : AJOB* 10 (9): 3–11. https://doi.org/10.1080/15265161.2010.494215.

Sanderson, Paul, and Ronald Hier. 2006. "Artist Collaboration Agreement." CARFAC Ontario.

Schneble, Christophe Olivier, Bernice Simone Elger, and David Martin Shaw. 2020. "Google's Project Nightingale Highlights the Necessity of Data Science Ethics Review." *EMBO Molecular Medicine* 12 (3): e12053. https://doi.org/10.15252/emmm.202012053.

Segal, Troy. 2022. "Big Data." *Investopedia*. https://www.investopedia.com/terms/b/big-data.asp.

Seyyed-Kalantari, Laleh, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen, and Marzyeh Ghassemi. 2021. "Underdiagnosis Bias of Artificial Intelligence Algorithms Applied to Chest Radiographs in Under-Served Patient Populations." *Nature Medicine* 27 (12): 2176–82. https://doi.org/10.1038/s41591-021-01595-0.

Savage, Neil. 2020. "How AI Is Improving Cancer Diagnostics." *Nature*, vol. 579, no. 7800, Mar. 2020, p. 214, link.gale.com/apps/doc/A618606620/AONE?u=anon~ea5f5f89&sid=googleScholar&xid=7e410b16.

Shan, Shawn, Emily Wenger, and Jiayun Zhang. 2020. "Fawkes: Protecting Personal Privacy against Unauthorized Deep Learning Models." *Proceedings of the 29th USENIX Security Symposium*, no. USENIX Securit: 2020.

Skandarani, Youssef, Pierre-Marc Jodoin, and Alain Lalande. 2021. "GANs for Medical Image Synthesis: An Empirical Study." *ArXiv:2105.05318 [Cs, Eess]*, July. http://arxiv.org/abs/2105.05318.

Skloot, Rebecca. 2010. *The Immortal Life of Henrietta Lacks*. US: Crown. http://rebeccaskloot.com/the-immortal-life/.

Social Sciences and Humanities Research Council of Canada. n.d. "Definition of Terms." Last modified May 4, 2021. https://www.sshrc-crsh.gc.ca/funding-financement/programs-programmes/definitions-eng.aspx.

Stahl, Darian Goldin. 2014. *Artist Books 2014–2020*. http://www.dariangoldinstahl.com/artist-books-20142019.

Stanford HAI. 2021. *Rashaad Newsome: BEING (The Digital Griot)*. Uploaded June 3, 2021. https://www.youtube.com/watch?v=pF0o5fO0-a4.

TalkDeath. 2019. "Digital Legacies: Preparing for Your Digital Death." *TalkDeath*. September 25, 2019. https://www.talkdeath.com/digital-legacies-preparing-for-your-digital-death/.

3D Slicer. 2020. "What Is 3D Slicer?" https://slicer.readthedocs.io/en/latest/user_guide/about.html#what-is-3d-slicer.

Torralba, Antonio, Rob Fergus, and Bill Freeman. 2020. "80 Million Tiny Images." June 29, 2020. https://groups.csail.mit.edu/vision/TinyImages/.

United States Department of Justice. 2014. Privacy Act of 1974. June 16, 2014. https://www.justice.gov/opcl/privacy-act-1974.

———. 2020. "Overview of the Privacy Act: 2020." The United States Department of Justice. October 14, 2020. https://www.justice.gov/opcl/overview-privacy-act-1974-2020-edition/definitions.

University College London. 2019. "Anonymisation and Pseudonymisation." Data Protection. https://www.ucl.ac.uk/data-protection/guidance-staff-students-and-researchers/practical-data-protection-guidance-notices/anonymisation-and.

Voarino, N., V. Couture, S. Mathieu-Chartier, J. C. Bélisle-Pipon, E. St-Hilaire, B. Williams-Jones, F. J. Lapointe, C. Noury, M. Cloutier, and P. Gauthier. 2019. "Mapping Responsible Conduct in the Uncharted Field of Research-Creation: A Scoping Review." *Accountability in Research* 26 (5): 311–46. https://doi.org/10.1080/08989621.2019.1620607.

Waldby, Catherine. 2000. *The Visible Human Project: Informatic Bodies and Posthuman Medicine*. Oxford: Routledge.

Witze, Alexandra. 2020. "Wealthy Funder Pays Reparations for Use of HeLa Cells." *Nature* 587 (7832): 20–21. https://doi.org/10.1038/d41586-020-03042-5.

Wood, Thomas. 2020. "Generative Adversarial Network." *DeepAI*, July 22, 2020. https://deepai.org/machine-learning-glossary-and-terms/generative-adversarial-network.

know
thyself
AS A VIRTUAL REALITY