"Digital Natives": How Medical and Indigenous Histories Matter for Big Data

*by Joanna Radin**

ABSTRACT

This case considers the politics of reuse in the realm of "Big Data." It focuses on the history of a particular collection of data, extracted and digitized from patient records made in the course of a longitudinal epidemiological study involving Indigenous members of the Gila River Indian Community Reservation in the American Southwest. The creation and circulation of the Pima Indian Diabetes Dataset (PIDD) demonstrates the value of medical and Indigenous histories to the study of Big Data. By adapting the concept of the "digital native" itself for reuse, I argue that the history of the PIDD reveals how data becomes alienated from persons even as it reproduces complex social realities of the circumstances of its origin. In doing so, this history highlights otherwise obscured matters of ethics and politics that are relevant to communities who identify as Indigenous as well as those who do not.

WHAT'S IN A NAME?

Several years ago I found myself in conversation with a mathematician. She was an expert in a field of problem solving called machine learning. As she explained it to me, applications of her work served to do things like optimize Google search rank orders, to make sure that people found what they were looking for or, perhaps, what they did not even know they were looking for. At the time of our conversation, she was using her expertise to help the electricity provider Con Edison to predict fires sparking in the underground power grid in New York City. Such fires, in addition to disrupting service, could lead to dangerous explosions of manhole covers.[1] To address the chal-

* Section of the History of Medicine, Sterling Hall of Medicine, Yale University, 333 Cedar Street, L132, New Haven, CT 06520; joanna.radin@yale.edu.

I would like to thank Cynthia Rudin and David Aha for conversations about their machine learning work and Jennifer Brown, Laurel Waycott, Laura Stark, and participants in the "Big Data and Invisible Labor" symposia at the Max Planck Institute for the History of Science.

[1] Cynthia Rudin, "21st Century Data Miners Meet 19th Century Electrical Cables," *Computer* 44 (2011): 103–5; Rudin, "Machine Learning for the New York City Power Grid," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34 (2012): 328–45. These circular, cast-iron caps cover holes constructed to provide repair workers access to the underground power grid when maintenance is required. For some data on the frequency and severity of fires beginning in manholes, see https://users.cs.duke.edu/~cynthia/docs/RudinEtAl2011ComputerMagazine.pdf (accessed 26 September 2014).

lenge of trying to anticipate and prevent such critical disruptions to the electricity supply, she was testing and optimizing the predictive algorithms that were the coin of her professional realm. These tests required access to sufficiently complex and validated data sets to feed to the algorithm. I asked her whether she collected the data to test these algorithms herself. "No," she answered. Then where did the data come from? She recited the names of several preexisting, freely available data sets. One struck a chord of recognition: Pima.

Having studied the history of efforts to enroll Indigenous peoples in biomedical knowledge projects, I knew that "Pima" could refer to members of an Indigenous community who live in the southwestern region of the United States.[2] The community has come to experience extremely high rates of diabetes and obesity, which has made its members the focus of extensive research by epidemiologists, geneticists, and medical anthropologists.[3] In 1990, those living at the Gila River Indian Community Reservation, outside of Phoenix, Arizona, were defined as having "the highest recorded prevalence and incidence of non-insulin-dependent diabetes of any geographically-defined population."[4] Members of this community, known to science as "Pima," refer to themselves as Akimel O'odham, which has been translated as River People.[5]

The political boundaries of the Gila River Indian Community, which was created in the mid-nineteenth century, made it possible for public health officials to conceptualize the reservation, one of the oldest in the United States, as a laboratory for observing the epidemiology of diabetes. Since the 1960s, before the rise of genomics, medical information collected from Akimel O'odham bodies was regarded as a valuable resource for improving and even defining general knowledge about the disease. In the process, this medical information was translated into digital form that would enable it to be reused for knowledge projects unrelated to diabetes or even biomedicine.

The short answer, then, to the question of whether or not there was a connection between the "Pima" I had encountered in my studies and the "Pima" this mathematician was using in hers was "yes." A longer answer is provided in this essay, where I describe historical circumstances that have led data about Akimel O'odham people to

[2] I describe the history of efforts to collect and freeze blood from members of Indigenous communities around the globe in Joanna Radin, *Life on Ice: A History of New Uses for Cold Blood* (Chicago, 2017).

[3] Mariana Leal Ferreira and Gretchen Chelsey Lang, eds., *Indigenous Peoples and Diabetes: Community Empowerment and Wellness* (Durham, N.C., 2006); Michael Montoya, *Making the Mexican Diabetic: Race, Science and the Genetics of Inequality* (Berkeley and Los Angeles, 2011); Carolyn Smith-Morris, *Diabetes among the Pima: Stories of Survival* (Tuscon, Ariz., 2006). Others may have read Malcolm Gladwell's article about the struggle of the Pima with diabetes: Gladwell, "The Pima Paradox," *New Yorker*, 2 February 1998.

[4] William C. Knowler, David J. Pettit, Mohammed F. Saad, and Peter H. Bennett, "Diabetes Mellitus in the Pima Indians: Incidence, Risk Factors and Pathogenesis," *Diabetes/Metabol. Rev.* 6 (1990): 1–27. For more recent statistics, see Centers for Disease Control and Prevention, "Diabetes Prevalence among American Indians and Alaska Natives and the Overall Population—United States, 1994–2002," *Morbid. Mortal. Week. Rep.* 52 (2003): 702–4.

[5] In the early 1600s, the first Spanish settlers questioned members of the community about their identity. Akimel O'odham, unfamiliar with the language of their inquisitors, are said to have responded with the phrase "pi-nyi-matchi," translated as "I don't know." The Spanish colonizers assigned them the name Pima, which is how they heard the response. Carl Waldman, *Encyclopedia of Native American Tribes* (New York, 2014), 4. When I use the word "Pima," it is to point to how the community was described historically, by those who studied it, as well as certain social scientists who have since written about its members. Tracking the label "Pima" matters for me as a historian for reasons that I will make clear.

be produced and mobilized beyond the reservation where they live and where they have long participated in medical research. I argue that the broader enterprise of data collection, use, and reuse has reproduced certain patterns of settler colonialism. This insight is not merely consequential for those who identify as Indigenous but also provides inspiration for intervening in the creation and management of new digital technologies of representation and knowledge that perpetuate—often needlessly—exploitative ideas about property, innovation, and self-determination.[6]

With the title "Digital Nativity," I am referencing a phrase coined by education consultant Mark Prensky, who in 2001 observed of the most recent information age that "today's students are no longer the people our education system was designed to teach."[7] I engage in my own practice of reuse by resignifying the term to extend questions of sovereignty and justice that have focused on land and land rights to the digital territories of our current information age, which is so often inappropriately cast as a virtual Wild West or an uninhabited frontier. The practice of reproducing frontier narratives short-circuits the potential to unsettle ideas about how innovation works, including the potential to learn from Indigenous peoples who have been at the vanguard in encouraging innovative and decolonial approaches to research.[8] Today's research subjects are no longer the objects Euro-American knowledge systems designed to make invisible.

At the core of this history are questions about the origins, ownership, and reuse of the personal and bodily data that fuels information economies. The story of how Indigenous participants in the National Institutes of Health's longitudinal research on diabetes at Gila River became understood as donors of data used to study diabetes and later, how that data was used to refine algorithms that had nothing to do with diabetes or even to do with bodies, is exemplary of the history of Big Data writ large. What makes data "big" is not so much its size—though that is relevant too—but its ability to radically transcend the circumstances and locality of its production.[9] Computers and algorithms make that possible, but understanding the politics of Big Data also requires attention to the creation and processing of the data itself, including the recognition that it often

[6] In other words, I am arguing for the relevance of indigenous studies theory to the critical study of Big Data. E.g., Kim TallBear, "Narratives of Race and Indigeneity in the Genographic Project," *J. Law. Med. & Ethics* 35 (2007): 412–24; Audra Simpson, "The Ruse of Consent and the Anatomy of 'Refusal': Cases from Indigenous North America and Australia," *Postcolon. Stud.* published online 6 June 2017, http://dx.doi.org/10.1080/13688790.2017.1334283 (accessed 20 June 2017); Linda Tuhiwai Smith, *Decolonizing Methodologies: Research and Indigenous Peoples* (1999; rep., London, 2012); Jodi A. Byrd, *The Transit of Empire: Indigenous Critiques of Colonialism* (Minneapolis, 2011).

[7] Marc Prensky, "Digital Natives, Digital Immigrants: Part 1," *On the Horizon* 9 (2001): 1–6.

[8] For instance, what would it mean to understand these spatially distributed territories as "data country," with reference to "Indian country," broadly defined as any of the self-governing Indigenous communities in the United States? See Vine Deloria Jr. and Clifford M. Lytle, "Indian Country," in *American Indians, American Justice* (Austin, Tex., 1983), 58–79. A recent example of an Indigenous community's efforts to produce protocols that are animated by their values is described by Linda Nordling, "San People of Africa Draft Code of Ethics for Researchers," *Science*, 17 March 2017, http://www.sciencemag.org/news/2017/03/san-people-africa-draft-code-ethics-researchers (accessed 21 March 2017). Notably, the San refuse to grant broad consent for other researchers to reuse data for purposes not specified in their original agreements.

[9] This is a feature that emerges from but does not contradict the tripartite definition of Big Data— a function of technology, analysis, and mythology—provided by danah boyd and Kate Crawford, "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon," *Inform. Comm. & Soc.* 15 (2012): 662–79, on 663.

comes from living, breathing people.[10] Not unlike recent conceptualizations of diabetes as a problem of food justice, wherein metabolic problems arise from decades of alienation from traditional food ways and land use, the history of a particular data set known as the Pima Indian Diabetes Dataset (often referred to as PIDD) makes political and economic subjectivity visible in an algorithmic age sustained by a steady diet of repurposed data. In doing so, it provides an approach—grounded in Indigenous practices of refusal as well as self-governance—for resisting or differently engaging with research in an age of Big Data.

INVISIBLE LABOR

A medical and Indigenous history of Big Data is, by necessity, one of invisible labor, in which the freedom or autonomy of participation of persons from whom data is generated is too often taken for granted or even sometimes celebrated as a form of "citizen science."[11] In the realm of computing, scholars have had to develop new methods to even identify let alone understand the vast and hidden human labor force that has become responsible for all kinds of tasks that enable our information infrastructures.[12] They are trying to access a realm that Ivan Illich called "shadow work": that which is functionally necessary to maintain institutions but is either not compensated or undercompensated.[13]

Sociologists Susan Leigh Star and Anselm Strauss demonstrated the utility of Illich's ideas in the realm of the hospital, where nurses—a traditionally feminized form of labor—struggled to "change work previously embedded under a general rubric of 'care' and usually taken-for-granted to work that is legitimate, individuated and traceable across settings."[14] The point of their research agenda was to make clear the often unseen—and uncompensated—social energy that was nonetheless necessary to sustain institutions.

More recently, these ideas have been applied to the role of individuals who engage with health care as patients and participants in biomedical research. Bioeconomic theorists Melinda Cooper and Cathy Waldby consider these individuals to be performing "clinical labor."[15] Cooper and Waldby begin with the premise that forms of "*in vivo* labor (either through the production of experimental data or the transfer of tissues) are

[10] See Malte Ziewitz, "Governing Algorithms: Myth, Mess, and Methods," *Sci. Tech. Hum. Val.* 41 (2016): 3–16, and other papers in that issue.

[11] The ambiguous valences and forms of recognition in recent citizen science efforts are described by Etienne Benson, "A Centrifuge of Calculation: Managing Data and Enthusiasm in Early Twentieth-Century Bird Banding," in this volume.

[12] Kavita Philip, Lilly Irani, and Paul Dourish, "Postcolonial Computing: A Tactical Survey," *Sci. Tech. Hum. Val.* 37 (2010), 3–29; Lilly Irani and M. Six Silberman, "Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk," in *CHI '13: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, 2013), 611–20.

[13] Illich wrote, in 1981, "While for wage labor you apply and qualify, to shadow work you are born or are diagnosed for." See Ivan Illich, *Shadow Work* (Boston, 1981), 100. My interest in Illich's concept of shadow work is directly inspired by Laura Stark's history of studies involving "normal" human subjects at the NIH; Stark, *The Normals: A People's History of the Human Experiment* (Chicago, forthcoming).

[14] Susan Leigh Star and Anselm Strauss, "Layers of Silence, Arenas of Voice: The Ecology of Visible and Invisible Work," *Comput. Support. Cooperat. Work* 8 (1999): 9–30, on 15.

[15] Melinda Cooper and Catherine Waldby, *Clinical Labor: Tissue Donors and Research Subjects in the Global Bioeconomy* (Durham, N.C., 2014).

increasingly central" to late capitalist economies. They further argue that professional bioethical practices have played an ironic role in placing research on human subjects "under an exceptional regime of labor" that allows them to have been understood as exempt from the standard protections of twentieth-century labor law. The resulting practice of viewing participants in biomedical research as "volunteers"—part of the bioethical insistence that biological labor should not be waged—has, in their view, only served to facilitate inadequate forms of compensation. Irrespective of the forms of compensation participants receive—such as health care itself—they argue that "services should be considered as labor when the activity is intrinsic to the process of valorization of a particular bioeconomic sector, and when the therapeutic benefits to the participants and their communities are . . . incidental."[16]

Whether it is through participation in longitudinal diabetes research such as that which created the PIDD or more recent direct-to-consumer genomic services that offer information about genetic risk and ancestry to individuals in exchange for access to their genome, many people—Indigenous and otherwise—also participate in unintentional shadow work when they use Google, Facebook, and other ostensibly "free" services. It is shadow work because this freely given "digital exhaust"—the mundane evidence of how people live and breathe on the web—becomes part of what allows these platforms to be valued in the billions of dollars on the stock market.[17] "It is on . . . 'invisibilities' that the collective delusions and collusions of the modern economy run," historian of science Rebecca Lemov has argued, "particularly as that economy merges with the virtual realm."[18] Tracing the itineraries of Big Data derived from medical information about Indigenous bodies—and, crucially, not the knowledge of Indigenous persons—reveals commonalities and settler colonial pathologies that exist across genres of invisible labor, in biomedical research as well as machine learning. Those who make the investments in repositories of data, including their creation and maintenance, are not always those who help conceptualize their design or benefit from their use.[19]

If, as information historian Geoff Bowker has argued, "Data is always already cooked," the circumstances of the initial preparation of Pima diabetes data have made it a newly valuable ingredient to be used for different kinds of meals, even as Akimel O'odham persons have been effaced in the process. By mixing medical and Indigenous history with the history of Big Data, in which metaphors of consumption abound, I reckon with the following questions: What kind of embodied or clinical labor is involved in generating the ingredients? What kind of effort is recognized or is legible as having been required to cook them? Are these ingredients ones that may ultimately lead to serious health problems? Is there a point at which data has been processed so much that they become inedible or unhealthful? And, perhaps most important of all, who gets to be at the table when these meals are served?

[16] Ibid., 4–6.

[17] Jaron Lanier, *Who Owns the Future* (New York, 2014); Christian Rudder, *Dataclysm: Who We Are When We Think No One's Looking* (New York, 2014); Chris Anderson, *The Long Tail: Why the Future of Business Is Selling Less of More* (New York, 2006). See also Dan Bouk, "The History and Political Economy of Personal Data over the Last Two Centuries in Three Acts," in this volume.

[18] Rebecca Lemov, "On Not Being There: The Data Driven Body at Work and at Play," *Hedgehog Rev.* 17 (2015), 44–55. See also Lemov, "Anthropology's Most Documented Man, Ca. 1947: A Prefiguration of Big Data from the Big Social Science Era," in this volume.

[19] Christine L. Borgman, *Big Data, Little Data, No Data* (Cambridge, Mass., 2015), 229. See also Benson, "Centrifuge" (cit. n. 11).

Through this historical case study of data about diabetes and a discussion of its limitations, I am arguing for an alternative approach to "data ethics"—different from one that might be modeled on that of the normative bioethics critiqued by scholars like Waldby and Cooper. I ask instead for critical engagement with the metabolism of Big Data grounded in awareness of settler colonial history and lived experience that can unsettle approaches to research and forms of compensation that challenge the capitalist and colonialist logics that have animated biomedicine and Big Data alike. Before it is possible to determine best practices for the appropriate use, reuse, and maintenance of data, it is imperative to understand how data comes into being.

THE GILA RIVER COMMUNITY RESERVATION: AN UNNATURAL LABORATORY

Akimel O'odham refer to their ancestors as HuHuKam, known for their engineering prowess in supporting an agricultural lifestyle. For centuries they built and maintained miles of irrigation infrastructure (canals that were 10 feet deep and as much as 30 feet wide) drawing on the Gila River.[20] The annexation of the Arizona Territory to the United States following the Mexican-American War in the mid-nineteenth century coincided with the California gold rush, which brought thousands of prospectors through the region, greatly disrupting local life. The Pima Gila River Indian Community was established during this time, in 1859. Encompassing 372,000 acres along the Gila River, it was the first reservation in what is now Arizona (which did not achieve statehood until 1912). The Pima Gila River Indian Community was also one of four federally recognized tribes at the time (today there are 566).[21] The 1870s and 1880s marked a downturn in the agricultural prosperity of the tribe when the construction of upstream diversion structures and dams by non-Native farmers cut Pima off from the water of the Gila River with devastating consequences.[22]

Anthropologist Frank Russell came to Gila River in 1901 as part of a study authorized by the Bureau of American Ethnology (BAE). The reservation had already been in place for several decades and offered Russell the convenience of an apparently bounded field of ethnographic study. However, as museum curator Joshua Roffler has argued, Russell's book, *The Pima*, published posthumously in 1908 (Russell died of tuberculosis in 1903 at the age of 35), "elided the reality that the lives of the Gila River Pima were deeply intertwined with the non-Indian residents of central Arizona."[23] Even though missionaries and BAE officials urged assimilation (to the extent of granting a horse wagon to each man who cut his hair and built an adobe house), modern references were cropped from the already deliberately posed photos in *The Pima*. It was a work of salvage ethnography that relied upon making certain aspects of lived experience invisible.[24]

[20] David H. DeJong, *Stealing the Gila: The Pima Agricultural Economy and Water Deprivation, 1848–1921* (Tucson, Ariz., 2009).

[21] The consequential differences between tribes that are recognized by the federal government and those that are not are described in M. E. Miller, *Forgotten Tribes: Unrecognized Indians and the Federal Acknowledgement Process* (Lincoln, Neb., 2004).

[22] Ibid.; Cary Walter Meister, *Historical Demography of the Pima and Maricopa Indians of Arizona, 1846–1974* (New York, 1989); David H. DeJong, *Forced to Abandon Our Fields: The 1914 Clay Southworth Gila River Pima Interviews* (Salt Lake City, 2011).

[23] Joshua Roffler, "Frank Russell at Gila River: Constructing an Ethnographic Description," *Kiva* 71 (2006): 373–95.

[24] Jacob Gruber, "Ethnographic Salvage and the Shaping of Anthropology," *Amer. Anthropol.* 72 (1970): 1289–99.

While many community members participated as research assistants to Russell, to the point of effectively conducting ethnography on themselves, they did not have control over the final cultural description that Russell produced. He was aware of this, writing to his boss in 1902, "These people are starving so that I am getting specimens and information that would not otherwise be obtainable."[25] These specimens included farming implements rendered obsolete by the drying of the Gila River.

Around this time, Aleš Hrdlička, considered to be one of the founding leaders of American physical anthropology, reported only one case of diabetes in the population.[26] But over the next forty years, those who lived on the reservation experienced mass famine and starvation, to which the American government responded with canned and processed food assistance. Between 1908 and 1955, the Bureau of Indian Affairs ran a health program, and, when a forty-two-bed hospital opened in 1953, the Indian Health Service (initially the Division of Indian Health, a branch of the U.S. Public Health Service) became responsible for providing comprehensive health care for the eligible residents of the Gila River Reservation.[27]

Tribal members have since argued that the loss of their ability to productively farm the land, combined with the change of diet introduced by government assistance, spurred the rise of obesity and Type 2 diabetes.[28] Such explanations have intermingled with genetic theories, such as James Neel's controversial 1962 "thrifty genotype" hypothesis, which posited that such communities developed genetic adaptations to help them survive during periods of famine.[29] In fact, Neel visited the region as a consultant for the NIH at the beginning of their longitudinal study in 1965. He claimed that "certain problems relevant to Indian health can be much more easily approached on the reservation than in, e.g., Phoenix or Tucson," including questions of genetics.[30] Critics have more recently suggested that biomedical researchers' emphasis on genetically determinist explanations have led to a sense of "fatalism" among some members of the community, who report feeling that diabetes is an unavoidable component of their identity.[31]

What has come to be defined by biomedical scientists as "cooperative" research between community members at Gila River and the NIH began formally in 1963. This

[25] Quoted in Roffler, "Frank Russell" (cit. n. 23), 388.

[26] Aleš Hrdlička, *Physiological and Medical Observations among the Indians of the Southwestern United States and Northern Mexico*, Bureau of American Ethnology Bulletin 24 (Washington, D.C., 1908).

[27] Stephen J. Kunitz, "The History and Politics of US Health Care Policy for American Indians and Alaskan Natives," *Amer. J. Public Health* 86 (1996): 1464–73; Abraham B. Bergman, David C. Grossman, Angela M. Erdrich, John G. Todd, and Ralph Forquera, "A Political History of the Indian Health Service," *Millbank Memorial Fund Quart.* 77 (1999): 571–604.

[28] Smith-Morris, *Diabetes* (cit. n. 3); Melanie Rock, "Classifying Diabetes; or, Commensurating Bodies of Unequal Experience," *Public Cult.* 17 (2005): 467–86; Emoke J. E. Szathmary, "Non-Insulin Dependent Diabetes Mellitus among Aboriginal North Americans," *Annu. Rev. Anthropol.* 23 (1994): 457–82.

[29] Puneet Chawla Sahota, "Genetic Histories: Native Americans' Accounts of Being at Risk for Diabetes," *Soc. Stud. Sci.* 42 (2012): 821–42; Margery Fee, "Racializing Narratives: Obesity, Diabetes, and the 'Aboriginal' Thrifty Genotype," *Soc. Sci. & Med.* 62 (2006): 2988–97. Indeed, Neel visited Gila River to see how his hypothesis might apply to the community.

[30] Neel, memo to National Cancer Institute, 8 November 1965, MS Coll. 96, Box 92, Folder "Pima—Tohono O'odham—1965–1968," James V. Neel Papers, American Philosophical Society, Philadelphia.

[31] Diane Weiner, "Interpreting Ideas about Diabetes, Genetics, and Inheritance," in *Medicine Ways: Disease, Health, and Survival among Native Americans*, ed. Clifford E. Trafzer and Diane Weiner (Walnut Creek, Calif., 2001), 108–33.

was when the National Institute of Arthritis, Diabetes and Digestive and Kidney Diseases (now known as the National Institute of Diabetes and Digestive and Kidney Diseases [NIDDK]) made a survey of rheumatoid arthritis among the groups they referred to as "Pima" in Arizona and "Blackfeet" in Montana. Researchers were surprised to find an extremely high rate of diabetes. In 1965, the Epidemiology and Field Studies Branch (EFSB) of the Institute—in partnership with the Indian Health Service—sent a team to begin an observational study of the Pima community at Gila River. The research was supposed to last for ten years. It continued for more than forty.[32]

Beginning in 1965, every resident of the Pima study area (which refers to the Gila River Indian Community) of at least five years of age was asked to participate in a "comprehensive longitudinal study of diabetes," for which they were examined approximately every two years. In 1984, the EFSB became the current Phoenix Epidemiology and Clinical Research Branch (PECRB), which oversaw the study. By the 1990s, the PECRB had come to oversee prevention programs, incorporated in response to criticism that the observational nature of the epidemiological study had not yielded any findings that directly benefited its participants. At the turn of the twenty-first century, despite great advances in knowledge about diabetes more generally, there were no discernible decreases in obesity or diabetes rates among community members.

Writing in the early 1990s, two leaders of the project, Clifton Bogardus and Stephen Lillioja, argued that "the Pima Indian model of this disease [diabetes] affords . . . major advantages," not least of all because "the population is genetically homogenous compared to Caucasian populations, and therefore the causes of NIDDM [non-insulin-dependent diabetes mellitus] are less heterogeneous, simplifying linkage studies."[33] In other words, the legal and political boundaries of the reservation functioned to enclose the Pima, making them appear as a natural and perhaps captive laboratory in which to study diabetes.[34]

A government pamphlet from 1996, "The Pima Indians: Pathfinders for Health," quoted Bogardus, who reported that "NIDDK scientists . . . have studied well over 90 percent of the people on the reservation at least once."[35] This same document cast Akimel O'odham as magnanimous donors of biomedical knowledge:

[32] National Institute of Diabetes and Digestive and Kidney Diseases, "Prospective Studies of the Natural History of Diabetes Mellitus and Its Complications in the Gila River Indian Community," ClinicalTrials.gov, National Institutes of Health, https://clinicaltrials.gov/ct2/show/NCT00339482 (accessed 21 March 2017).

[33] Clifton Bogardus and Stephen Lillioja, "Pima Indians as a Model to Study the Genetics of NIDDM," *J. Cell. Biochem.* 48 (1992): 337–43.

[34] Historian Matthew Klingle is undertaking an environmental history of diabetes, including work at Gila River. This research, like my own, is indebted to a large body of literature on efforts to transform indigenous communities into living laboratories. See, e.g., Helen Tilley, *Africa as a Living Laboratory: Empire, Development, and the Problem of Scientific Knowledge, 1870–1950* (Chicago, 2011); Warwick Anderson, "The Colonial Medicine of Settler States: Comparing Histories of Indigenous Health," *Health Hist.* 9 (2007): 144–54; Roffler, "Frank Russell" (cit. n. 23); David S. Jones, "The Health Care Experiments at Many Farms: The Navajo, Tuberculosis, and the Limits of Modern Medicine, 1952–1962," *Bull. Hist. Med.* 76 (2002): 749–90; Christian W. McMillen, *Discovering Tuberculosis: A Global History 1900 to the Present* (New Haven, Conn., 2015). On the racialized history of diabetes more broadly, see Arleen Tuchman, "Diabetes and Race: A Historical Perspective," *Amer. J. Public Health* 10 (2011): 24–33; Tuchman, "Diabetes and 'Defective' Genes in the Twentieth Century United States," *J. Hist. Med. Allied Sci.* 70 (2015): 1–33.

[35] National Institute of Diabetes and Digestive and Kidney Diseases, *The Pima Indians: Pathfinders for Health* (Bethesda, Md., 1996), 7.

> Once trusted scouts for the US Cavalry, the Pima Indians are pathfinders for health. . . . The Pima Indians are giving a great gift to the world by continuing to volunteer for research studies. Their generosity contributes to better health for all people, and we are all in their debt. . . . The Pima Indians' help is so important . . . because of the uniqueness of the community. There are few like it in the world.[36]

As William Knowler, an NIH researcher at the reservation since 1975 who is also recognized as one of the world's 250 most highly cited researchers in clinical medicine, biology, and biochemistry, testified before Congress, "This study has contributed much to the world's current understanding of the causes and consequences of Type 2 diabetes and its complications, for which we are indebted to this community."[37] The year that "Pathfinders for Health" was published was the first one in which the NIH funded a large-scale prevention program involving Akimel O'odham participants, raising concern that the publication may have been produced to help assuage local feelings of exploitation, that they had been guinea pigs for the benefit of others.[38]

In the 1990s, the issue of compensation was addressed by ensuring that those who participated in the study continued to receive medical care and also $50, free transportation, and a meal each time they received diagnostic testing.[39] Before considering the significance of this form of compensation, I want to pause this epidemiological history to shift attention to a quite different area of expertise that was emerging in parallel, that of machine learning.

MACHINE LEARNING

Machine learning is a subdiscipline of artificial intelligence that practitioners date back to the late 1950s. It focuses on algorithms capable of learning and/or adapting their parameters based on a set of observed data without having been programmed to do so. An algorithm is, in its simplest form, a set of instructions or a code. It is a form of "software" that organizes data to generate meaningful information. It is what allows computers to do computational work.[40] As media theorist Tarleton Gillespie has explained,

> Algorithms are inert, meaningless machines until paired with databases upon which to function. For us, algorithms and databases are conceptually conjoined . . . but before results can be algorithmically provided, information must be collected and readied for the algorithm and sometimes excluded or demoted.[41]

[36] Ibid., 5.

[37] Knowler testimony before Committee on Indian Affairs, U.S. Senate, 8 February 2007, https://olpa.od.nih.gov/hearings/110/session1/testimonies/diabetesAI.asp (accessed 21 March 2017).

[38] Rachel Winer, "Diabetes, the National Institutes of Health, and the Arizona Pima Indians: A Study of Ethics and Experimentation in American Medicine, 1965 to the Present" (senior thesis, Yale Univ., 2006). In her thesis, Winer also analyzed a three-part exposé on the diabetes research undertaken at Gila River; Graciela Sevilla, "A People in Peril: Pimas on the Front Lines of an Epidemic," *Arizona Republic*, 31 October–2 November 1999.

[39] As observed by Winer, "Diabetes" (cit. n. 38).

[40] David Berlinski, *The Advent of the Algorithm: The 300-Year Journey from an Idea to the Computer* (New York, 2000).

[41] Tarleton Gillespie, "The Relevance of Algorithms," in *Media Technologies*, ed. Tarleton Gillespie, Pablo J. Boczkowski, and Kirsten Foote (Cambridge, Mass., 2013), 167–94, on 169. A more cynical view is offered in Cathy O'Neill, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (New York, 2016).

Machine learning theorists are concerned with issues such as computational complexity, computability, and generalization. Algorithms are the coin of their realm; data is used to refine them. The field, in ways that are described by Hallam Stevens in this volume, is a marriage of applied math and computer science.[42]

Machine learning and the related field of statistical pattern recognition have been the subject of increasing interest to the biomedical community because they offer the promise of improving the sensitivity and specificity of detection and diagnosis of disease, while at the same time purportedly increasing the objectivity of the decision-making process. According to an early editorial published in the *Journal of Machine Learning*, "unlike psychology, machine learning is fortunate in that it can experimentally study the relative effects of 'nature' versus 'nurture.'"[43] The author, Pat Langley, based at the University of California, Irvine (UCI), believed that by looking at unprecedentedly large amounts of supposedly raw data, machine learning could avoid human biases.[44] He looked forward to the day when standardized databases would facilitate such studies. And it was at UCI that an important resource was established to help achieve this goal.

The UCI Machine Learning Repository is, in the words of its stewards, "a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms."[45] It is an archive of data sets, created in 1987 by David Aha and several other graduate students at UCI, including Jeff Schlimmer and Doug Fisher. According to Aha, it began after he had heard several calls for such a repository at machine learning conferences, but nothing suitable materialized. "I became convinced it would not exist without some reasonable number of datasets," he recalled. "I requested all those I had read about in publications at that time and from other sources, and I believe I announced it only after it had 25 (a number I only vaguely recall; the actual number may have differed) mostly well-known datasets. . . . It was pre-web, available by ftp [File Transfer Protocol]. I was gratified by the community's interest, and Pat Langley (a mentor) [and the author of the editorial referenced above] was particularly encouraging and no doubt suggested others to contact me proactively, which many folks did."[46]

Among the major issues initially raised by the repository were those of attribution and ownership. Aha recalled that he "realized that UCI needed credit for this effort, and I broadly requested and publicized a reference format to cite the repository. . . . I reviewed many papers over the years where I requested this to be inserted, or contacted folks after they had published their paper to remind them to include its citation in their future papers, when appropriate."[47] Yet, as information theorist Christine Borgman has recently argued, such citation practices "largely presume that objects

[42] Paul Sajda, "Machine Learning for Detection and Diagnosis of Disease," *Annu. Rev. Biomed. Eng.* 8 (2006): 537–65. See also Hallam Stevens, "A Feeling for the Algorithm: Working Knowledge and Big Data in Biology," in this volume.

[43] Pat Langley, "Machine Learning as an Experimental Science," *J. Machine Learning* 3 (1988): 5–8, on 6.

[44] A form of "mechanical objectivity." See Lorraine Daston and Peter Louis Galison, *Objectivity* (New York, 2007).

[45] UCI Machine Learning Repository, "About," http://archive.ics.uci.edu/ml/about.html (accessed 21 March 2017).

[46] David Aha, UCI Repository History, as told to Padhraic Murphy, 21 June 2009, transcript of e-mail correspondence from the personal files of David Aha (used with his permission).

[47] Ibid.

are fixed, stable, and complete units. None of these conditions can be assumed with data."[48] The radical instability and alienability of this kind of data, as we will see, would become central to its power and also its critique.

Issues of accuracy in citing data sets were also of concern, especially for data sets that dealt with medical issues. "Some of the databases were donated only on the condition that their use was cited accurately," Aha explained, "as was especially the case for medical datasets I obtained. . . . Most of the others had no such condition."[49] One such medical resource is the "Heart Disease Data Set" from the Cleveland Clinic, which became part of the repository in 1988. In this case, Aha noted that those researchers who donated the data were primarily concerned with attribution—making certain that their work and funders were credited properly—not about compensating patients, whose data had been made anonymous and was therefore regarded as protected.

By the first decade of the twenty-first century, the UCI Machine Learning Repository had been cited over 1,000 times, making it on its own one of the top 100 most cited "papers" in all of computer science.[50] It is worth noting that Aha did not list it as a publication on his CV, making his own connection to the repository invisible, at least in a formal capacity. He did, however, acknowledge that the labor of encouraging people to donate their data sets—itself a kind of shadow work—actually contributed to raising his profile in the field of machine learning.[51]

Still known as the UCI Machine Learning Repository, it is now maintained by the University of Massachusetts, Amherst, with financial support from the National Science Foundation. The home page for the repository thanks "the donors and creators of the databases and data generators," which refers to data scientists, not research subjects.[52] Aha's sense is that, in recent years, the field has shifted in ways that have allowed other kinds of repositories to become more important, but the "discretized" data in the UCI repository make it appropriate for certain, specific research and also a resource for those new to the field of machine learning.[53]

One of the oldest data sets on file is the "Pima Indians Diabetes Data Set," which was donated in 1990. The data set is often referred to by its initials, PIDD, and "has become a standard for testing data mining algorithms to see their accuracy in predicting diabetic status from the 8 variables given."[54] These variables, or "attribute information," were extracted from paper patient records, not any kind of DNA-based data, and included (1) number of pregnancies, (2) plasma glucose concentration after two hours in an oral glucose tolerance test, (3) diastolic blood pressure, (4) triceps skin fold thickness, (5) two-hour serum insulin level, (6) body mass index, (7) diabetes pedigree function, and (8) age.

In a pregenomic era, the diabetes pedigree function (DPF) provided a synthesis of the diabetes history of relatives (including parents, grandparents, full and half siblings,

[48] Borgman, *Big Data* (cit. n. 19), 242.
[49] Aha, UCI Repository History (cit. n. 46).
[50] Google Scholar reports a much higher number, closer to 6,000, but many of those are redundancies.
[51] David Aha, telephone interview by Joanna Radin, 30 June 2015.
[52] UCI Machine Learning Repository, "About" (cit. n. 45).
[53] See, e.g., machinelearnringmaster.com/case-study-predicting-the-onset-of-diabetes-within-five-years-part-1-of-3/, 29 March 2014 (accessed 15 July 2015).
[54] Joseph L. Breault, "Data Mining Diabetic Databases: Are Rough Sets a Useful Addition?," in *Proceedings of the 33rd Symposium on the Interface*, ed. Edward J. Wegman, D. T. Gantz, and J. J. Miller (Fairfax Station, Va., 2002), 597–606.

full and half aunts and uncles, and first cousins) and the genetic relationship of those relatives to the subject. It has been described in academic publications as providing "a measure of the expected genetic influence of affected and unaffected relatives on the subject's eventual diabetes risk."[55] Effectively, the DPF was a technology that functioned to devalue Akimel O'ohdam understandings of kinship—which include familial relationships across the community as well as with nonhuman animals and landscapes—folding biogenetic information into the data set, enabling researchers to extrapolate—in the absence of DNA—patterns of heredity.[56] It also folded in assumptions about life course and gender in that "all patients [768 in total] were females at least 21 years old of Pima Indian heritage."

Experts in the field of machine learning pay little attention to these particular attributes but have observed that "diabetes is a particularly opportune disease for data mining technology. . . . First, because the mountain of data is there. Second, diabetes is a common disease that costs a great deal of money, and so has attracted managers and payers in the never ending quest for saving money and cost efficiency. Third, diabetes is a disease that can produce terrible complications . . . so physicians and regulators would like to know how to improve outcomes as much as possible. Data mining might prove an ideal match in these circumstances."[57]

Let us take a closer look at the PIDD as it appears on the UCI Machine Learning Repository (fig. 1). The donor of the data is listed as Vincent Sigillito, who was part of the Applied Physics Laboratory at Johns Hopkins. The Applied Physics Laboratory (APL) is a not-for-profit engineering research and development center, founded in 1942 to assist the military with ballistics detonation. According to the APL's website, its mission is to "provide solutions to national security and scientific challenges with systems engineering and integration, research and development, and analysis."[58] The APL is located not far from the Johns Hopkins School of Public Health, which was crucial for facilitating interactions between researchers in epidemiology and machine learning. Aha, I learned, did a postdoc at Johns Hopkins University in the early 1990s, and Sigillito was his sponsor. This is how the PIDD wound up in the UCI Machine Learning Repository. The data set was constructed from a larger database by the NIDDK. As of March 2017, it had been viewed nearly 260,000 times.[59]

From the information available on the UCI Machine Learning database it is possible to discern that data in the PIDD was based on research first reported by epidemiologists in 1981.[60] The 1981 paper, published in the *American Journal of Epidemiology*, indicated that diabetes incidence was computed for 3,137 subjects, each of whom had been examined at least twice. Estimates of age- and sex-adjusted incidence and prevalence rates with 95 percent confidence intervals were computed comparing Pima

[55] Jack W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus," *Johns Hopkins APL Tech. Digest* 10 (1988): 262–6.

[56] Deborah House, "'Know Who You Are and Where You Come From': Ties of Kin, Clan, and Homeland in Southwestern Indian Identity," *Rev. Anthropol.* 33 (2004): 371–91.

[57] Breault, "Data Mining" (cit. n. 54).

[58] Johns Hopkins Applied Physics Laboratory, "About APL," http://www.jhuapl.edu/aboutapl/default.asp (accessed 21 March 2017).

[59] This is up from about 90,000 times, when I began this research in 2015. The exponential uptick in views may track with the explosion of machine learning applications.

[60] William C. Knowler, David J. Pettitt, Peter J. Savage, and Peter H. Bennett, "Diabetes Incidence in Pima Indians: Contributions of Obesity and Parental Diabetes," *Amer. J. Epidemiol.* 113 (1981): 144–56.

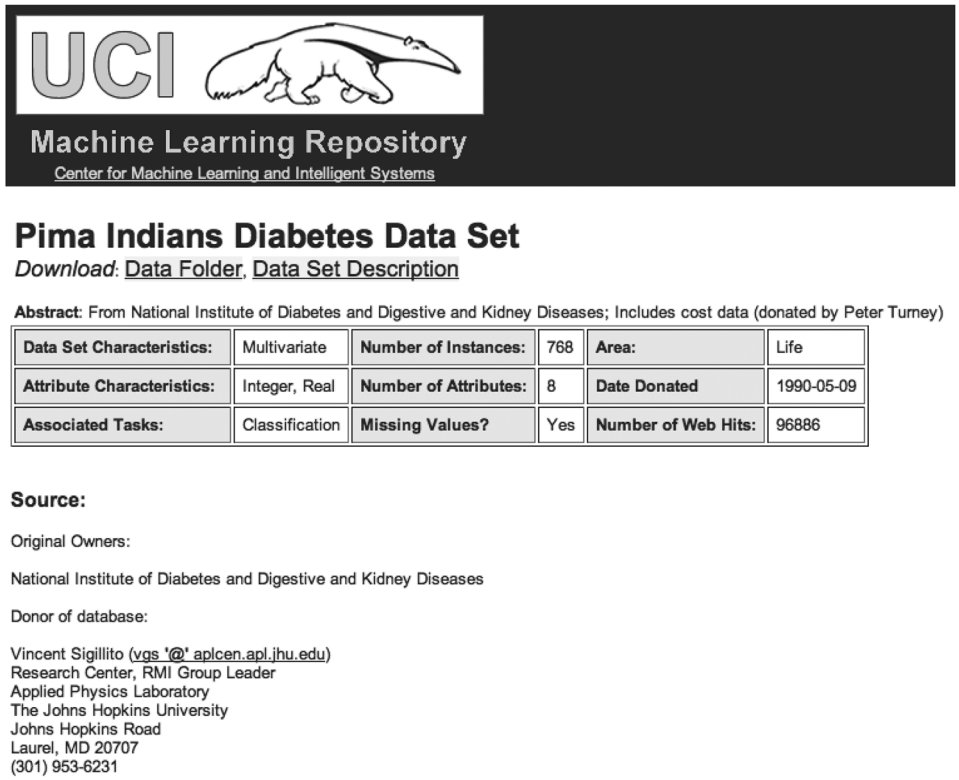

Figure 1. *Screenshot of Pima Indians Diabetes Data Set on UCI Machine Learning Repository,* "*About*" (cit. n. 45).

data to the 1970 U.S. Census of the white population, including armed forces. Incidence rates were also stratified by body mass index. The authors concluded that their findings were "consistent with Neel's hypothesis that diabetes results from the introduction of a steady food supply to people who have evolved a 'thrifty genotype.'"[61]

In 1988, a group of researchers at the NIH and Johns Hopkins, with connections to the Applied Physics Laboratory, sought to further extract and digitize data from this resource and apply it to the machine-learning context.[62] The goal of that paper, published in a very different forum—the *Johns Hopkins APL Technical Digest*—was to test the ability of an early neural network model algorithm called ADAP (proclaimed to be "loosely" modeled after computation in the brain), to forecast the onset of diabetes in the high-risk population of Pima Indians. This required transforming the original, manually recorded data set into one that could be entered into a computational machine (fig. 2).[63]

[61] Ibid., 145. For an analysis of the remarkable endurance of the "thrifty genotype hypothesis," see Isabel Beshar, "A Tale of a Mutating Theory: The Evolution of the Thrifty Genotype Hypothesis from 1962–2007" (senior thesis, Yale Univ., 2014).

[62] Smith et al., "Using the ADAP Learning Algorithm" (cit. n. 55).

[63] The digitization step was essentially punch card technology. See Martin Campbell-Kelly and William Aspray, *Computer: A History of the Information Machine*, 1st ed. (New York, 1996). The early history of efforts to use computers to model living systems has been told in Timothy Lenoir, "Shaping

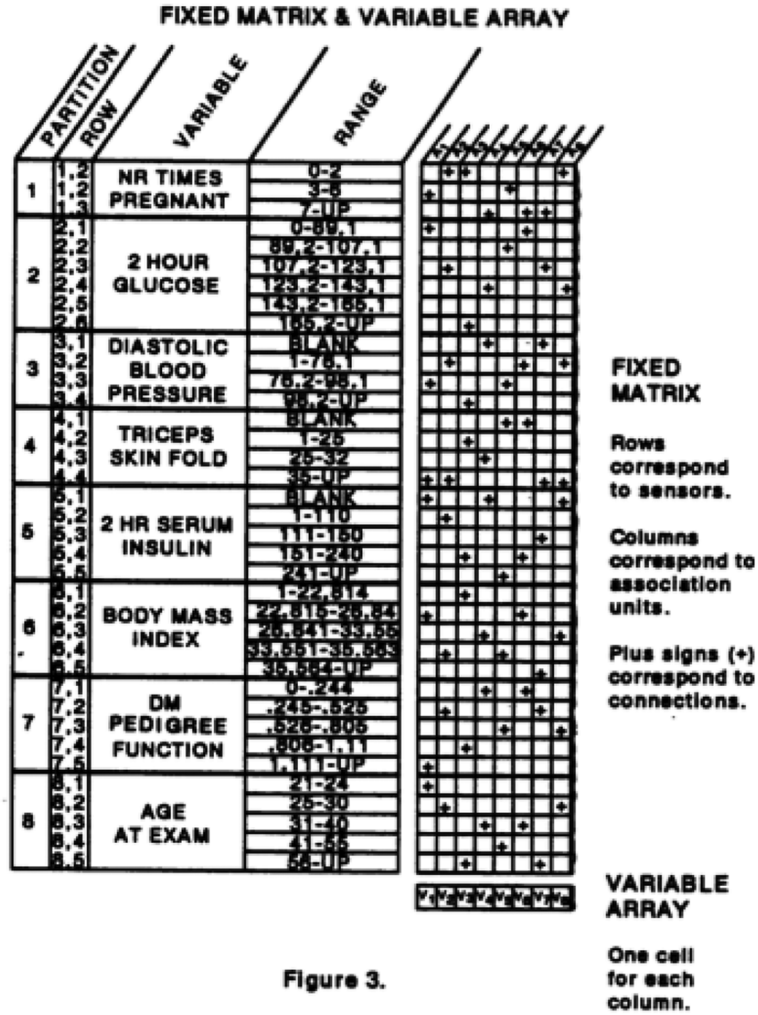

Figure 2. *Example of data reformatted for machine-based computation from Smith et al.,* "*Using the ADAP Learning Algorithm*" *(cit. n. 55).*

The authors explained that neural network models worked by using a "training" data set to discover patterns in data. Once the algorithm had been trained using 576 cases, it was used to forecast whether another 192 test cases would develop diabetes within five years. In the materials and methods section of the paper, the authors presented a four-part justification for privileging the study population. First and foremost, they argued that the fact that the Pima population could be "recognized as such due to its location on the reservation"—its ability to function as a natural laboratory—was crucial to the validity of the data set. Related was the matter of the longitudinal

Biomedicine as an Information Science," in *Proceedings of the 1998 Conference on the History and Heritage of Science Informaton Systems*, ed. Mary Ellen Bowden, Trudi Bellardo Hahn, and Robert V. Williams (Medford, N.J., 1999); Joseph November, *Biomedical Computing: Digitizing Life in the United States* (Baltimore, 2012).

aspects of the research, that Pima had been the subject of "continuous study since 1965 by NIHDDK." The third justification focused on the WHO's standardization of criteria for diagnosing diabetes (based on a Technical Report of a Study Group, TRS 727, 1985; a prior report was published in 1980, TRS 646).[64] The final source of legitimation was the prior ubiquity of the data set itself, which was already known to the investigators and served as a "well-validated" resource.[65]

REUSING AND REFUSING

Through its availability on the UCI Machine Learning Repository, the PIDD became a standard resource for testing algorithms of all kinds. The data set, by virtue of the fact that it was and continues to be freely available, has also been used to refine algorithms intended for "knowledge discoveries" that have nothing do with diabetes, including the prediction of manhole fires.[66] To David Aha, what made PIDD such a valuable data set in the world of machine learning beyond the realm of diabetes research was that it was data on a topic that seemed important, but, perhaps more relevant to its long life in the machine learning community, it was not overly large: "people could work with it. It was all tables and this made it easier to be used with the algorithms that were being created at the time." Furthermore, the attributes were straightforward, which made it "amenable from a number of perspectives." Even though there were some missing values, this became part of its appeal or a "feature" of the data set.[67] It was not too big or too small. Even what it did not include came to be exemplary, appearing to replicate the contingency and complexity of real-world situations.

Research on the prediction of manhole fires—which used algorithms developed with the use of PIDD—featured prominently in Viktor Mayer-Schonberg and Kevin Cukier's popular 2013 exposition, *Big Data*. In their book, they invoke the reuse potential of Big Data as one of its defining and most valuable features. Among the broader lessons they draw: (1) Big Data is being put to new uses to solve difficult real-world problems; (2) to achieve this, however, we need to change how we make knowledge; (3) we have to use all the data, as much as can possibly be collected, not just a small portion; (4) we need to accept messiness rather than treat exactitude as a central priority; and (5) we must put our trust in correlations without fully knowing the causal basis for predictions.[68]

The history that I have provided thus far highlights other equally consequential features of Big Data to which Mayer-Schoenberg and Cukier were not attuned. Chief among them is that today's machine learning scientists have been disincentivized from

[64] On the WHO's technical reports in shaping research agendas, see Joanna Radin, "Latent Life: Concepts and Practices of Human Tissue Preservation in the International Biological Program," *Soc. Stud. Sci.* 43 (2013): 483–508. The standards disseminated by such documents play an important role in the construction of disease categories. See Geoffrey C. Bowker and Susan Leigh Star, *Sorting Things Out: Classification and Its Consequences*, Inside Technology (Cambridge, Mass., 1999).

[65] Smith et al., "Using the ADAP Learning Algorithm" (cit. n. 55).

[66] Krzysztof J. Cios and Witold Pedrycz, *Data Mining Methods for Knowledge Discovery* (Boston, 1998); Jiawei Han and Micheline Kamber, *Data Mining: Concepts and Techniques* (San Francisco, 2001). On the subject of "open access," see Sabina Leonelli, "Why the Current Insistence on Open Access to Scientific Data? Big Data, Knowledge Production, and the Political Economy of Contemporary Biology," *Bull. Sci. Tech. Soc.* 33 (2013): 6–11.

[67] Aha, interview (cit. n. 51).

[68] Viktor Mayer-Schonberger and Kenneth Cukier, *Big Data: A Revolution That Will Transform How We Live, Work and Think* (New York, 2013), 70.

considering the meaning or origins of the data they feed to their algorithms. This may be a function of "algorithmic objectivity," upheld by members of the data mining and machine learning communities as "fundamental to the maintenance of these tools as legitimate brokers of relevant knowledge."[69] The fact that this data is also considered to be naturally occurring—a neutral product of the contingent circumstances of its acquisition—is seen as one of its additional advantageous qualities, even if the data itself is—as in the case of the PIDD—a product of settler colonialism, economic struggle, and biosocial suffering.[70] While it may be tempting to consider the PIDD—the data set itself—as a kind of model organism for machine learning, doing so closes off access to the shadow work of people, Akimel O'odham, who continue to live and die on the reservation and to circulate as disembodied data, stored on the servers of universities and corporations.[71] It also elides the work Akimel O'odham have undertaken to redefine the norms of research encounters with their community that are in keeping with values not grounded in capitalist or settler colonial logics.[72]

As Pima data moved off the reservation, it became available for new and unexpected uses in basic informatics research. As a result, Akimel O'odham lost "direct control over intervention in and treatment of serious diseases affecting their populations," as Richard Narcia, Lieutenant Governor of the Gila River Community, testified before Congress.[73] How do members of the Akimel O'odham community at Gila River, in particular, feel about their loss of control over data derived from studies of the serious diseases affecting them? I have not been able to speak with a representative of the community. The community has recently put into place strict procedures restricting researcher access. Even with a reference from a trusted colleague, the head of the Gila River Indian Community's Committee on Health and Social Standing, which controls IRB permission, did not respond to my inquiries. I also experienced this lack of response from non-Indigenous health workers who have long-term relationships with the community.

A decade ago, Rachel Winer, then an undergraduate at Yale, was able to conduct interviews with individuals from both of these groups for her Yale senior thesis project.[74] In 2015, they chose not to engage with my requests for interviews; it was their right, in accordance with a Health Care Research Ordinance enacted by the Gila River Indian Community in 2009.[75] The ordinance states that "the Community Council has found that Medical and Health Care Research has been conducted in ways that do not respect the human dignity of human subjects and that do not recognize the legitimate interest of the Community in the integrity and preservation of its culture."[76]

[69] Gillespie, "Relevance" (cit. n. 41).

[70] On the politics of biosocial suffering, see Duana Fullwiley, *The Enculturated Gene: Sickle Cell Health Politics and Biological Difference in West Africa* (Princeton, N.J., 2011).

[71] Nathan Ensmenger, "Is Chess the Drosophila of Artificial Intelligence? A Social History of an Algorithm," *Soc. Stud. Sci.* 42 (2011): 5–30.

[72] See, e.g., Kim TallBear, "Beyond the Life/Not Life Binary: A Feminist-Indigenous Reading of Cryopreservation, Interspecies Thinking and the New Materialisms," in *Cryopolitics: Freezing Life in a Melting World*, ed. Joanna Radin and Emma Kowal (Cambridge, 2017), 179–202.

[73] Richard Narcia, Testimony of the Gila River Indian Community before the Senate Committee on Indian Affairs, Washington, D.C., 8 March 2000. Cited in Winer, "Diabetes" (cit. n. 38).

[74] Winer, "Diabetes" (cit. n. 38).

[75] Gila River Indian Community, "Ordinance GR-05-09," http://nptao.arizona.edu/sites/nptao/files/gila_river_indian_ord_gr-05-09_title_17_chapter_9_0.pdf (accessed 21 March 2017).

[76] Ibid. Section 9.107 includes a discussion of the information to be provided to the review board by prospective researchers, including "Who shall own the data from Medical and Health Care Research?"

In addition to specific circumstances leading to the ordinance, in part associated with reuse of tissues of members of other Indigenous communities without consent, research relationships are not stable over time and involve shifting priorities, allegiances, and desires. One facet of this is a phenomenon known as "research fatigue"—a feeling of exhaustion about being subjects of inquiry.[77] This position can grow out of a desire to avoid what has been referred to as "voyeurism," where researchers seek to learn lessons from the study of communities but fail to improve conditions or consider the points of view of community members.[78]

This is a perspective that I respect. However, I am also mindful of the risks of erasure of the role of Indigenous peoples from studies of existing and emerging technoscientific infrastructures.[79] Part of what it means to write the history of Big Data is to be attentive not only to the voices that have been silenced but also to those who have chosen not to speak, or to speak at this particular moment in time. This silence might be understood in terms of "refusal," a position that has been identified and explored by anthropologist Audra Simpson in the context of her research on Mohawk citizenship and sovereignty.[80] Simpson recognized the political dimensions of moments when her desired subjects resisted her inquiries. "Rather than stops, or impediments to knowing," she realized, "those limits may be expansive in what they do not tell us. . . . The refusals speak volumes because they tell us when to stop."[81]

Simpson's decision to embrace her ethnographic subjects' right not to give her the information she initially believed she needed to obtain appears to run counter to the maxim, upheld by digital entrepreneurs like Mark Zuckerberg, that "information wants to be free." The other, less frequently cited coda to that maxim is that information "also wants to be expensive," meaning that it can be hugely valuable to the recipient, though not necessarily to the donor.[82]

Simpson is aware of the apparent perversity of refusal in a Euro-American culture that values openness. However, she notes that the political theory of John Locke, which

---

and "What control will the individual medical and Health Care Research participants have over the use of their own data? What control will the Community or Medical and Health Care Research participants have over the current and future use of the data, and how will the control be exercised?" This does not include oversight of social science research. As documented in Naomi Tom, "Protecting Tribal Nations through Community Controlled Research: An Analysis of Established Research Protocols within Arizona Tribes" (MS thesis, Arizona State Univ., May 2015), https://repository.asu.edu/attachments/150598/content/Tom_asu_0010N_14771.pdf (accessed 1 October 2016).

[77] Tom Clark, "'We're Over-Researched Here!' Exploring Accounts of Research Fatigue within Qualitative Research Engagements," *Sociology* 42 (2008): 953–70. On uneven partnerships in global health research that can contribute to research fatigue, see Johanna Crane, "Unequal 'Partners': AIDS, Academia and the Rise of Global Health," *Behemoth* 3 (2010): 78–97.

[78] Lauri Gilchrist, "Aboriginal Communities and Social Science Research: Voyeurism in Transition," *Native Soc. Work J.* 1 (1997): 69–85, https://zone.biblio.laurentian.ca/bitstream/10219/472/1/NSWJ-V1-art6-p69-85.pdf (accessed 1 October 2016).

[79] The latent racism that has animated our information infrastructures can be interrupted by recognizing the contributions of Indigenous people. See, e.g., Lisa Nakamura, "Indigenous Circuits: Navajo Women and the Racialization of Early Electronic Manufacture," *Amer. Quart.* 66 (2014): 919–41.

[80] Audra Simpson, "On Ethnographic Refusal: Indigeneity, 'Voice,' and Colonial Citizenship," *Junctures* 9 (2007): 67–80; Simpson, *Mohawk Interruptus: Political Life across the Borders of Settler States* (Durham, N.C., 2014). The concept of refusal was first articulated within the context of anthropology by Sherry B. Ortner, "Resistance and the Problem of Ethnographic Refusal," *Comp. Stud. Soc. Hist.* 37 (1995): 173–93.

[81] Simpson, "On Ethnographic Refusal" (cit. n. 80), 78.

[82] Stewart Brand stated this in 1984 at the first Hackers' Conference. It was printed in the May 1985 *Whole Earth Review*, and again in 1987 in his book, *The Media Lab: Inventing the Future at MIT* (New York, 1987).

promotes shared standards of justice and truth in an intellectual commons, was derived from the violent enclosure of land and alienation of Indigenous peoples from their modes of governance.[83] It is a philosophy of property that has made it such that Indigenous communities' own laws of dominion over their bodies, including knowledge about those bodies, are not legible in settler colonial regimes of power.[84]

In Euro-American history, reaching all the way back to Montaigne and Rousseau, the Indigenous subject has often been situated as someone who is naturalized—taken as being closer to nature and outside of history, even as their existence forms the basis for modern political thought.[85] Naturalness is also a component of what makes Big Data valued by those who consume it; it is supposedly harvested from people going about their daily lives, in the "real world," allowing scientists to learn how people behave and think *in vivo virtual*. The activity of daily life releases exhaust or hidden data into a virtual territory idealized as a commons, yet it is those who have defined the digital commons who are setting the terms upon which it can be valued. Data exhaust is made of the digital traces that are silently, or invisibly, accumulated and given off by the devices people use every day, like smartphones and computers. Discourses of "naturalness" preempt conversation about appropriate forms of compensation and governance.[86] Perhaps more importantly, they often foreclose the question of whether or not it is ever possible to cease being productive if one is making data that is not even visible to oneself.

Today, Akimel O'odham, in part as a result of their long-term involvement in biomedical research, are deeply skeptical about continuing to participate in additional forms of research, including that which is undertaken by historians.[87] Their skepticism or refusal, and the difficult legacies of other Indigenous communities' involvement with biomedicine, anticipate concerns that are beginning to be articulated about the datafication of all kinds of lives.[88] These issues are implied by the idea of "digital natives," a means of suggesting that Indigenous people themselves have been compelled to generate alternative ways of conceiving of what it means to be a citizen in a digital age, including tactics for resisting its embrace.[89]

[83] Simpson, "On Ethnographic Refusal" (cit. n. 80), 74.

[84] Simpson invokes Locke's assertion that "amongst those who are *counted the civilized part of mankind* [emphasis added by Simpson], who have made and multiplied positive laws to determine property . . . is by the labour that removes it out of that common state nature left it in made his property who takes pains about it"; quoted in ibid., 70. "Thus," Simpson concludes, "property could be defined only as that which was mixed with labour and belonged to those who perceived it, in contradistinction to the living histories of Indigenous peoples in those places"; ibid.

[85] Johannes Fabian, *Time and the Other: How Anthropology Makes Its Object* (New York, 1983); Eric Wolf, *Europe and the People without History* (Berkeley and Los Angeles, 1982); Marc Rifkin, *Beyond Settler Time: Temporal Sovereignty and Indigenous Self-Determination* (Durham, N.C., 2017).

[86] See, e.g., William Cronon's arguments about European settlers in colonial America. Cronon, *Changes in the Land: Indians, Colonists, and the Ecology of New England* (New York, 1983).

[87] Anthropologist and STS scholar Puneet Sahota has done important research on Native Americans' perceptions of being at risk for diabetes. Because of concerns about stigmatization, she has anonymized not only the individuals she interviewed but also the tribal community to which they belong. Sahota, "Genetic Histories" (cit. n. 29), 821–42. An alternative approach taken by Kim TallBear "studies up," focusing on the ideas and values of scientists who work with Indigenous peoples. TallBear, *Native American DNA: Tribal Belonging and the False Promise of Genomic Science* (Minneapolis, 2013).

[88] See, e.g., Borgman, *Big Data* (cit. n. 19).

[89] The idea of the "biodefector," one who resists and, in doing so, may even reconfigure expectations of the research encounter, has been advanced in a thoughtful discussion by Ruha Benjamin, "Informed Refusal: Toward a Justice-Based Ethics," *Sci. Tech. Hum. Val.* 41 (2016): 967–90.

I emphasize this because it is not only indigenous studies scholars who have critiqued and sought to redefine the research encounter and the broader political economic regime by which it is bolstered. The implications of the history of the PIDD are intensified and compounded when we consider that even though the UCI Machine Learning Repository was extremely important in the early years of the field of machine learning, it was also sharply criticized within the community. By 1995, Aha recalled, "the problems 'caused' by the repository had become popularly espoused. For example, at . . . [the International Conference on Machine Learning] Lorenza Saitta had [in an invited workshop that Aha co-organized] passionately decried how it allowed researchers to publish dull papers that proposed small variations of existing supervised learning algorithms and reported their small-but-significant incremental performance improvements in comparison studies. But even before this concern became broadly recognized throughout the community (which sadly, implies that the Repository was successful), I recall others could see this coming."[90]

Aha was referring to a paper that would become a widely cited takedown of the "popular practice of exploiting ready-to-use data sets," published by Italian machine learning scientists Lorenza Saitta and Filippo Neri.[91] They interpreted the stakes of relying on what they called "off the shelf data" as impeding a two-way or iterative process of technology transfer. Drawing on their own experiences in developing and applying machine learning systems in fields that included industrial troubleshooting, molecular biology, medicine, industrial robotics, speech recognition, cognitive psychology, and knowledge discovery in large databases, Saitta and Neri invoked experiences "interacting with both domain experts and end-users who displayed different attitudes and different degrees of trust and understanding of the potential of the methodologies we were proposing."[92]

Saitta and Neri explained that "designers of ML [machine learning] systems usually envision a scenario including themselves, data, and possibly, a marginally useful expert, but almost never the user. On the contrary, the user should be a fundamental component of this scenario . . . the user must not only be present, but also actively participate in the development of a 'real-world' application."[93] The conclusions they reached were not radically different from those reached by indigenous studies scholars or historians of science and technology, who make arguments for, respectively, self-determination and the fact that users matter.[94] The case of reuse further complicates these insights, reminding us that what constitutes a "user," let alone a "self," is not stable across time or place, even digital ones.

From the perspective of a historian such as myself, this creates a conundrum: How to tell a history of Big Data in a way that highlights the central role members of Indigenous communities and other generators of data cast as "digital natives" have played

[90] Aha, UCI Repository History (cit. n. 46).
[91] Lorenza Saitta and Filippo Neri, "Learning in the 'Real World,'" *Machine Learning* 30 (1998): 133–63, on 133.
[92] Ibid.
[93] Ibid., 136.
[94] See e.g., Nelly Oudshoorn and T. J. Pinch, *How Users Matter: The Co-Construction of Users and Technologies*, Inside Technology (Cambridge, Mass., 2003); Smith, *Decolonizing Methodologies* (cit. n. 6); Debra Harry, "Acts of Self-Determination and Self-Defense: Indigenous Peoples' Responses to Biocolonialism," in *Rights and Liberties in the Biotech Age: Why We Need a Genetic Bill of Rights*, ed. Sheldon Krimsky and Peter Shorett (Lanham, Md., 2005), 87–97; James Clifford, *Returns: Becoming Indigenous in the Twenty-First Century* (Cambridge, Mass., 2013).

in a way that does not reproduce the very harms that have led to this realm of inquiry? Rather than viewing the purpose of this essay as delivering a statement on what residents of the Gila River Community believe is at stake in the reuse of data based on their bodies, or viewing the nonresponse to my inquiries as a lack of interest, I hope this essay will be read by Akimel O'odham and by machine learning experts alike as an invitation to engage this history in their efforts to collectively redefine what it means to engage or even to refuse research.[95] Such work begins with my acceptance that for some Indigenous people, "research," as Linda Tuhiwai Smith has argued in her important writing on decolonizing methods, has become "a dirty word."[96] The conclusions of this essay can only ever be understood as the partial perspective of a non-Indigenous historian of medicine.[97] I have called for a reconceptualizion of the questions seen as relevant to the study of Big Data, rather than an attempt to seek solutions to problems that do not admit the concerns of all affected groups, especially the people from whom data is made.

CONCLUSION

What kinds of activities are seen as relevant to the regulation of our biomedical and, increasingly, our information infrastructures? Which ones are readily made invisible and with what consequences?[98] How is the project of historicizing Big Data augmented when participation in biomedical research is recognized as a form of labor, or when the imperative to cast new generations of citizens as "digital natives" is considered in terms of Indigenous experience?[99] In particular, I am urging that greater attention be given to alterlives: what happens to bodies that have become alienated from their personhood as well as forms of personhood that have become alienated from bodies.[100] Receiving health care, looking for love, or even ordering takeout food over the Internet have all become activities that can and are being turned into data to ask questions of which many are largely unaware. As authors in this volume have emphasized, Rebecca Lemov and Dan Bouk especially, data are people too.

Information studies scholar Christine Borgman has rearticulated concerns raised in the 1990s with what she calls "enchantment" with the possibilities of the reuse of old data.[101] She calls for a broader conversation about what the reuse of data is intended to accomplish and for greater participation by those who produce the data. Borgman recently coauthored a "Joint Declaration of Data Citation Principles," which is focused primarily on ensuring that data are cited in ways that support the "need to give

[95] See, e.g., Sally M. Reid and Raymond Reid, "Practicing Participatory Research in American Indian Communities," *Amer. J. Clin. Nutr.* 69 (1999): 755S–9S.

[96] Smith, *Decolonizing Methodologies* (cit. n. 6).

[97] Donna Jeanne Haraway, "Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective," *Feminist Stud.* 14 (1988): 575–99.

[98] This is an opportunity to build on the long-standing interests of historians of science in invisible labor. See, e.g., Steven Shapin, "The Invisible Technician," *Amer. Scient.* 77 (1989): 554–63; Naomi Oreskes, "Objectivity or Heroism? On the Invisibility of Women in Science," *Osiris* 11 (1996): 87–113.

[99] For recent high-profile critiques of our existing models of the value derived from data, see Lanier, *Who Owns the Future* (cit. n. 17); Mayer-Schonberger and Cukier, *Big Data* (cit. n. 68).

[100] The concept of alterlives is Michelle Murphy's. See http://www.toxicsymposium.org/conversations-1/2016/3/1/alterlife-in-the-ongoing-aftermath-exposure-entanglement-survivance (accessed 20 September 2016).

[101] Borgman, *Big Data* (cit. n. 19), 222. See also boyd and Crawford, "Critical Questions" (cit. n. 9).

scholarly credit to contributors and the importance of data as evidence."[102] In keeping with long-standing concerns about the importance of credit in expert communities, these principles are intended to ensure that those who produce, maintain, and curate the data and those who use the same data to refine algorithms or to recognize patterns are seen as equal contributors to the knowledge production process. In practice, this means relying on citations to make the "invisible technicians" as visible and as credible as the scientists.[103] Until these principles are adopted, Borgman has stated, and perhaps even then, data provenance will continue to be a "cascading problem" as data sets continue to be reused and recombined. The overarching challenge, as she sees it, is to "understand the many roles associated with data and to reach consensus within communities for which of these roles deserve credit and the best way to assign it."[104]

This is an admirable goal, but it may not extend far enough, not least of all because it is ambiguous about the kinds of credit and compensation that stand to be valued by those whose bodies or behaviors generate data captured by digital infrastructures. Similarly, it may not be able to capture the ways in which the data that supports fields like machine learning can be used to cultivate expertise that serves an array of possible futures. David Aha, who as a graduate student built the freely available UCI Machine Learning Database to support the cultivation of algorithmic expertise, now works for the United States military designing what are known as "adaptive systems." These, he explained in more colloquial terms, include "drones." Drones are vehicles that are piloted without a person on board. They are essentially flying robots, created by humans to do the work of maintaining but also severing connections that have been made onerous or unpalatable. Drones, it is often assumed, are servants of human intention.[105] Yet in the ways that machine learning's drones reorganize the relationships between humans and their machines through artificial intelligence, the drone provides yet another important opportunity to reflect on the deeper histories of alienation that have guided the distribution of labor in data-driven societies.[106]

In the face of research that points to the unexpected and potentially undesirable consequences of data reuse, some scholars have called for a domain of "data ethics," modeled on "bioethics." These scholars, as well as Borgman, are quite right to insist on the importance of wrestling with matters of social and political consequence in the realm of "Big Data." However, rather than reproducing existing models of bioethics, here is an opportunity to rethink and revise institutionalized strategies for guiding the research enterprise. It is a chance to first evaluate the history of bioethics itself—what has worked and what has not.

A good example of a thorny problem for bioethics has been the doctrine of informed consent. Time and again, it has proven inadequate to the task of adequately involving research subjects and potential research subjects in decision making. To cite only a few recognized limitations, the Human Genome Diversity Project of the 1990s struggled to address concerns about indigenous participation in genetic research by devel-

[102] Christine L. Borgman, "An Introduction to the Joint Principles for Data Citation," *Bull. Amer. Soc. Inform. Sci. Tech.* 41 (2015): 43–5.
[103] Shapin, "The Invisible Technician" (cit. n. 98).
[104] Borgman, *Big Data* (cit. n. 19), 242.
[105] Markus Krajewski, "Master and Servant in Technoscience," *Interdiscipl. Sci. Rev.* 37 (2012): 287–98.
[106] Neal Curtis, "The Explication of the Social: Algorithms, Drones and (Counter-) Terror," *J. Sociol.* 52 (2016): 522–36.

oping a model ethical protocol that advocated "group consent." This protocol was meant to recognize that the bioethical principle of autonomy did not enjoy the same status in all societies. Since then, concerns about reuse of human-derived research materials such as blood and DNA have led scholars to consider the ways that it becomes impossible to be informed about future applications.[107] Similar concerns are now being addressed in the realm of Big Data, where the phrase "click here to consent forever" has been invoked to express dissatisfaction with existing models of informed consent.[108]

The implications of the history of the PIDD case have obvious relevance for Indigenous communities, who have been at the vanguard of innovating strategies concerning the unintended uses of bodily extracts, including refusal to provide them in the first place.[109] It is also relevant to the millions of citizens who are increasingly entreated to have their most intimate desires transformed into data. For whom and what are they pathfinders?[110] These are timely opportunities for historians of medicine and Indigenous studies scholars to join forces with those invested in producing kinds of data science that better serves those whose embodied lives make it possible. This approach requires starting with empirically thick descriptions of the conditions of possibility that have given rise to technoscientific infrastructure, rather than with philosophical principles that are grounded in the logic of settler colonialism. The real innovation would be to let go of the historical fantasy of the future frontier and create systems that are accountable to those who live and labor in the present.

[107] For an excellent recent review of the issues, see Klaus Hoeyer and Linda F. Hogle, "Informed Consent: The Politics of Intent and Practice in Medical Research Ethics," *Annu. Rev. Anthropol.* 43 (2014): 347–62. For studies that highlight alternative approaches, largely grounded in Indigenous experience, see Benjamin, "Informed Refusal" (cit. n. 89); Kristin Solum Steinsbekk and Berge Solbgerg, "Biobanks—When Is Re-consent Necessary?" *Public Health Ethics* 4 (2011): 236–50; Joan Cunningham and Terry Dunbar, "Consent for Long-Term Storage of Blood Samples by Indigenous Australian Research Participants: The DRUID Study Experience," *Epidemiol. Perspect. Innovations* 4 (2007), https://epi-perspectives.biomedcentral.com/articles/10.1186/1742-5573-4-7 (accessed 21 March 2017). For research specifically concerned with indigenous participants, see Laura Arbour and Doris Cook, "DNA on Loan: Issues to Consider When Carrying out Genetic Research with Aboriginal Families and Communities," *Community Genet.* 9 (2006): 153–60; Michelle Mello and Leslie E. Wolf, "The Havasupai Indian Tribe Case—Lessons for Research Involving Stored Biological Samples," *New Engl. J. Med.* 363 (2010): 204–7; Constance MacIntosh, "Indigenous Self-Determination and Research on Human Genetic Material: A Consideration of the Relevance of Debate on Patents and Informed Consent, and the Political Demands on Researchers," *Health Law J.* 13 (2005): 213–51; Jenny Reardon, *Race to the Finish: Identity and Governance in an Age of Genomics* (Princeton, N.J., 2005); Annie O. Wu, "Surpassing the Material: The Human Rights Implications of Informed Consent in Bioprospecting Cells Derived from Indigenous People Groups," *Washington Univ. Law Quart.* 78 (2000): 979–1003.

[108] Bart Custers, "Click Here to Consent Forever: Expiry Dates for Informed Consent," *Big Data* 3 (2016): 1–6.

[109] Jenny Reardon and Kim TallBear, "'Your DNA Is Our History': Genomics, Anthropology, and the Construction of Whiteness as Property," *Curr. Anthropol.* 53 (2012): 233–45; Emma Kowal, Joanna Radin, and Jenny Reardon, "Indigenous Body Parts, Mutating Temporalities, and the Half-Lives of Postcolonial Technoscience," *Soc. Stud. Sci.* 43 (2013); Rebecca Tsosie, "Cultural Challenges to Biotechnology: Native American Genetic Resources and the Concept of Cultural Harm," *J. Law Med. Ethics* 35 (2007): 396–411; Debra Harry, "Indigenous Peoples and Gene Disputes," *Chicago Kent Law Rev.* 84 (2009): 147–96; Harry, "Acts of Self-Determination" (cit. n. 94); Emma Kowal, "Orphan DNA: Indigenous Samples, Ethical Biovalue and Postcolonial Science," *Soc. Stud. Sci.* 43 (2013).

[110] For example, recent controversies surrounding Facebook's manipulation of its users' advertising experience, conducted in the name of scientific research. Reed Albergotti, "Furor Erupts over Facebook's Experiment on Users," *Wall Street Journal*, 30 June 2014, http://www.wsj.com/articles/furor-erupts-over-facebook-experiment-on-users-1404085840 (accessed 21 March 2017).