# Data Science and Mass Media: Seeking a Hermeneutic Ethics of Information

Christine A. James

Valdosta State University

Valdosta, Georgia, USA

Proceedings of the Society for Phenomenology and Media, vol. 15, 2014, pages 49-58

*"The inborn human instinct to imitate — that and man's commonest weakness, his aversion to being unpleasantly conspicuous, pointed at, shunned, as being on the unpopular side." – Mark Twain, 1901*

In recent years, the growing academic field called "Data Science" has made many promises. On closer inspection, relatively few of these promises have come to fruition. A critique of Data Science from the phenomenological tradition can take many forms. This paper addresses the promise of "participation" in Data Science, taking inspiration from Paul Majkut's 2000 work in *Glimpse*, "Empathy's Impostor: Interactivity and Intersubjectivity," and some insights from Heidegger's "The Question Concerning Technology." The description of Data Science provided in the scholarly literature includes "the study of the generalizable extraction of knowledge from data" (Dhar 2013, 64), "data stewardship and data sharing…access to data at higher volumes and more quickly, and the potential for replication and augmentation of existing research" (Hartter et al., 2013, 1), and "personal information, health status, daily activities and shopping preferences that are recorded and used to give us instant feedback and recommendations based on previous online behavior." (Shin 2013) United States universities have begun to offer graduate programs in "data science", anticipating the growth of this field for marketing, national security, and health industries. These universities include New York University, Columbia University, Stanford,

Northwestern, and Syracuse. (Shin 2013) Data science graduate programs typically provide analysis of the academic credentials of their graduate students, including star chart diagramming.[i]

Those publishing in the Data Science literature emphasize that their work opens new access to knowledge and data sharing, but in reality, there is often a distorted relation of self and other. This is explained clearly in work by Paul Majkut (2000 and 2010). One example comes from Data Science's spatially located data systems, which can be used to classify and describe survey takers reducing them each to specific attributes: Black, Male, over 50 years old, working within 5 miles of home address.[ii] The products of data science keep the subjectivity of the Other (the one with these attributes) at a distance, and allows for "handling" of the object. Majkut describes how this results in alienation of subject and object, and a distortion of the Other. This alienation and distortion result in a false "empathy", the mere appearance of understanding and caring for the Other. "New media distances the participant to a degree that a qualitative dissociation of empathy…is unavoidable." (Majkut 2010, 2013) Data Science is an example of a media that provides the means to distance oneself from, but still enact a false "empathy" for, the subjects one studies.

Another way to understand this distancing of subjects and objects can be clarified by delineating the three major groups of "stakeholders" addressed regularly in the Data Science literature. The first major group is biologists and geneticists. In the early days of molecular genetics, a great deal of time was spent in decoding genomes (the order of amino acids involved in individual strands of DNA, deoxyribonucleic acid). One of the first well-publicized uses of Data Science as a tool involved mapping the human genome, analyzing a great deal of genetic information quickly. Notably for the medical industry, there was always an attached use value in

connecting this data with particular medical treatments. (Katayama et al. 2014, 24) In the process of seeking such medical treatment, the Other who will be treated is alienated from their own data. They are simultaneously made to feel empowered and participating in their own healing, but others control, categorize and quantify their means to become well. The speed of data collection, and the flood of new knowledge, gives a counterfeit sense of reciprocity and shared experience between patient and provider.

The speed and usefulness of data collection becomes an end in itself for the second major group of stakeholders, the computer scientists who develop new and more "elegant" forms of aggregated data analysis, and seek ways to make profitable applications for businesses and corporations. According to computer science professionals and administrators who rely on their work, "predictive modeling and machine learning are increasingly central to the business models of Internet-based data-driven businesses." (Dhar 2013, 73) One example of the development and refinement of Data Science is the "BioHackathon," a regular event dedicated to problem-solving and refinement of specific issues in data storage and retrieval.[iii] A great deal of effort is spent creating a variety of different kinds of data, including "linked data" cross-referencing between multiple aggregated reports through hypertext markups within documents and pdf files. (Katayama et al. 2014, 24)

The previously mentioned biological and genetic data relates directly to the third group of stakeholders in Data Science. These are the individuals whose knowledge (of their own health, habits, genetic predispositions, and home energy use) is portrayed as accessible and empowering through Data Science. Their biological information is useful for health care providers and insurance actuaries, and those with access to such data hold a great deal of power and decision making capacity based on that data. But this is inherently problematic given the resulting

alienation of subject and object, and the false sense of empathy in those who rely on Data Science as a way to define and categorize the other. This is made all the more problematic when Data Science is extolled as a way to allow the individual to "participate" actively in their own choices about health care and energy uses. A misleading narrative of empowerment and self-determination is predominant in the literature on Data Science.[iv]

I argue that beside Majkut's impostor empathy in those who use the technology, there is also an impostor "participation" in the objects (in the Other(s) studied through Data Science.) Individuals can increasingly collect data about their habits, routines, and environment technology that is always at the ready, such as smartphones. The data collected by these devices are both personal (identifying of an individual) and participatory (accessible by that individual for aggregation, analysis, and sharing). (Shilton 2012, 1905) In "The Question Concerning Technology," Heidegger noted that our real relationship to the essence of technology is out of our grasp, precisely because we are so routinely connected to it: "Everywhere we remain unfree and chained to technology, whether we passionately affirm or deny it." The crux of the problem is that "participatory" Data Science makes us feel empowered even as we are connected or chained to the technology itself, and in a broader political sense, we do not have control over the data we are participating in creating.

Data Science literature describes two types of data that are portrayed as empowering for the individual: participatory and personal data. Participatory data are accessible to the individual. Examples might be apps that monitor personal fitness, such as MyFitnessPal. In contrast, personal data are authored by an individual and describe an individual, but need not be accessible or usable by the individual (Kang, 1998). For example, one's home may have instrumentation constantly reporting energy use to a utility company, but not to the homeowner. (Shilton 2012,

1906) Much of the scholarship on Data Science holds a positive view of the potential for both participatory and personal data to be used for political ends that will benefit the individual, such as enabling individuals in marginalized social positions to access and analyze data to confront people in power (a member of a village in India using data on their water consumption to challenge water charges imposed by a corporate entity such as Nestle or Veolia, for example.)[v] Because our understanding of technology is always informed by our instrumental conception of what it does for us (its instrumental definition), one can see how the "what it does for us" is easily manipulated and packaged as a benefit to us.

While participatory data might sound helpful to the individual and even "democratic", it is important to note that legal and ethical discussions on Data Science are still in flux and less optimistic in tone. For example, many European countries are interested in keeping biobanks of DNA information but such projects have often failed because of a lack of clarity on intellectual property rights and individual skepticism. (Rose 2006, 184)[vi] Individuals often disagree on which data should be saved in public record, or how publicly accessible data should be. Not all institutions follow "Institutional Review Board" concepts about privacy. U.S. law does not interpret personal data to be owned by the subject of those data. Instead, legal regimes give control of, and responsibility for, personal data to the institution that collected the data. (Waldo et al., 2007 as cited in Shilton 2012, 1909)[vii]

While the Data Science community is beginning to discuss these political and ethical issues, even including ethics training in graduate program curricula, in general such concerns tend to fall behind concerns of elegance in programming.[viii] Those who participate in studies that use Data Science as an instrumental tool may have data collected from friends and loved ones via social networking, which skips over a variety of concerns such as Institutional Review Board (IRB)

approval.[ix] Out of concern for good stewardship of data, the US National Science Foundation (NSF) "prescribed that a two-page data management plan must accompany all research proposals." (Hartter et al. 2013, 1)

All of this points to the illusion of a "participatory" claim on the part of Data Science. Use of technology to understand one's own health and utility use on a daily basis invokes a "reiterated deception," and especially a deception about the individual's own empowerment and self-understanding. It involves quantifying our habits and attributes, giving us a false sense of certainty, self-definition, and comfort from the numbers. Those who do not participate in the gathering of data, in the aggregation, are left to feel conspicuous. From the view of the data scientist, the participatory individual, the participant, is not rendered empowered, but rendered zombie-like, as a mere locus of information rather than a self-motivated individual. (Majkut 2010, 203)[x] The motivation might seem to be empowering participation of the individual and taking control of one's own data, but sadly, this motivation soon becomes secondary to participation in the gathering of data in itself.[xi]

As in the epigraph from Mark Twain's "The United States of Lyncherdom," Data Science brings a pressure to participate in the gathering of data. Twain anticipated much of the later criticism of media from phenomenology, while using a different vocabulary. Twain was concerned with the interplay of "democratic control" (read as participation, empathy for the Other) over "technological change." He noted in writings like "The Tragedy of Pudd'nhead Wilson" that data storage and data mining and data aggregation (in Wilson's case via fingerprinting everyone in his town) would bring about a simultaneous dissociation between individuals, and an illusory self-empowerment for individuals who believe themselves to be masters of their own data. In the literature on Twain, scholars find him concerned to "ensure that

local citizens and groups have continuous access to scientific and technological expertise so that they can educate themselves regarding new developments…not restricting access to technological expertise to the elites." (Smith 2000)[xii] One can compare this to Heidegger's goal in The Question Concerning Technology: opening up or "democratizing" the conversation on technology, bringing it to those who are not experts.[xiii] Perhaps the first step in such a goal is to acknowledge when empowerment and participation are in fact mere illusions.

Bibliography

Cargo, M., & Mercer, S.L. "The value and challenges of participatory research: Strengthening its practice." *Annual Review of Public Health* 29 (2008): 325–350. Print.

Coupland, R. et al. "Protecting Everybody's Genetic Data." *Lancet* May 21, 365.9473, (2005): 1754-1756. Print.

Dhar, Vasant. "Data Science and Prediction." *Communications of the ACM* 56.12 (2013): 64-73. Academic Search Complete. Web. 6 March 2014.

Dreyfus, Hubert and Stuart. "Why Computers May Never Think Like People." *Technology Review MIT Technology*, 1986. Rpt. in Readings in the Philosophy of Technology. Ed. David M. Kaplan. New York: Rowman and Littlefield, 2004. 375-390. Print.

Gale, Robert L. "Twain's Pudd'nhead Wilson." *Explicator*, Fall 1979, Vol 38 Issue 1: 4-6. Print.

Haraway, Donna J. "A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century," in Simians, Cyborgs and Women: The Reinvention of Nature, New York: Routledge, 1991: 149-181. Print.

Hartter, Joel, et al. "Spatially Explicit Data: Stewardship and Ethical Challenges in Science." *Plos Biology* 11.9 (2013): 1-5. Academic Search Complete. Web. 6 March 2014.

Heidegger, Martin. "The Question Concerning Technology." in The Question Concerning Technology and Other Essays, Trans. and Ed. William Lovitt, New York: Garland Publishing/Harper and Row, 1977: 3-35. Print.

Katayama, Toshiaki, et al. 2014. “BioHackathon Series in 2011 and 2012: Penetration of Ontology and Linked Data in Life Science Domains.” *Journal of Biomedical Semantics* 5.1 (2014): 1-33. Academic Search Complete. Web. 6 March 2014.

Majkut, Paul. "Empathy’s Imposter: Intersubjectivity versus Interactivity." *Glimpse: Proceedings of the Society for Phenomenology and Media* 2 (2000): 59-65. Print.

Majkut, Paul. “Media.” in *Handbook of Phenomenological Aesthetics, Contributions to Phenomenology* 59, (2010): 201-205. Web. 1 March 2014.

McArdle, Jennifer. “Rethinking Property in the Digital Era.” National Interest. 5 May 2014. < http://nationalinterest.org/commentary/rethinking-property-the-digital-era-10388?page=2> Web. 4 May 2014.

Proceedings of the National Academy of Sciences of the United States of America. <http://www.pnas.org/content/102/43/15337/F1.expansion.html> Web. 25 February 2014.

Railton, Stephen. “The Tragedy of Mark Twain, by Pudd'nhead Wilson,” *Nineteenth-Century Literature*, March 2002, Vol. 56 Issue 4: 518-545. Print.

Rose, Hilary. “From Hype to Mothballs in Four Years: Troubles in the Development of Large-Scale DNA Biobanks in Europe.” *Community Genetics*, 2006, Vol. 9, Issue 3: 184-189. Web.

Royal, Derek Parker. 2002. “The Clinician as Enslaver: Pudd'nhead Wilson and the Rationalization of Identity,” *Texas Studies in Literature & Language*, Winter 2002, Vol. 44 Issue 4:414-452. Academic Search Complete. Web. 10 October 2013.

Schutt, Rachel. “The Stars of Data Science.” 2012. < http://columbiadatascience.com/2012/12/08/the-stars-of-data-science/> Web. 13 October 2013.

Shilton, Katie. “Participatory Personal Data: An Emerging Research Challenge for the Information Sciences.” *Journal of the American Society for Information Science and Technology* 63.10 (2012): 1905-1915. Academic Search Complete. Web. 6 March 2014.

Shin, Laura. 2013. “How the New Field of Data Science Is Grappling With Ethics,” September 10, 2013. <http://www.smartplanet.com> Web. 13 October 2013.

Smith, Tony. Technology and Capital in the Age of Lean Production. Albany, New York: SUNY Press, 2000: 135-159. Print.

Spangler, George M. 1970. “Pudd'nhead Wilson: A Parable of Property.” *American Literature*, March 1970, Vol. 42 Issue 1:28-38. Academic Search Complete. Web. 1 October 2013.

Strout, Cushing. 2012. “Crisis in Camelot: Mark Twain and the Idea of Progress.” *Sewanee Review*, 2012, Vol. 120, No. 2: 336-340. Academic Search Complete. Web. 1 October 2013.

Twain, Mark. “Pudd’nhead Wilson.” in The Writings of Mark Twain, New York, XVI, Stormfield Edition, originally published in America on November 28, 1894, as “The Tragedy of Pudd'nhead Wilson and the Comedy of Those Extraordinary Twins.” New York: Harper & Brothers, 1929. Print.

Twain, Mark. “The United States of Lyncherdom.” 1901. Rpt. in *The Complete Essays and Satires of Mark Twain*, New York: e-artnow, Kindle eBook edition 2014. <http://people.virginia.edu/~sfr/enam482e/lyncherdom.html> Web. 20 March 2014.

[i] For an example of a star chart diagram, see Schutt, Rachel. 2012, whose blog The Stars of Data Science, includes one: http://columbiadatascience.com/2012/12/08/the-stars-of-data-science/

[ii] See the Proceedings of the National Academy of Sciences of the United States of America http://www.pnas.org/content/102/43/15337/F1.expansion.html The example shows spatial distribution of residences of hypothetical survey respondents in Washtenaw County, Michigan, with the specific attributes of one respondent highlighted. The primary benefit of capturing locational human subjects’ data (e.g., socioeconomic conditions and demographics) is to support longitudinal research, help avoid over-researched locales, and capture locational effects (e.g., elevated lead levels). The ability to identify and locate these study ‘‘spaces’’ requires even stricter data control to protect confidential information. New methods aggregate social data at larger scales or mask data locations, allowing data interpolation using less distinct spatial patterns… Some IRBs now require that spatially explicit social data be kept confidential or that anyone with data access be made aware of their ethical obligations and added to ethics approval. (Hartter et al. 2013, 3)

[iii] The yearly BioHackathon series of events attempts to provide the environment within which these choices (of data management and interoperability) can be explored, evaluated, and then implemented on a collaborative and community-guided basis. These BioHackathons were hosted by the National Bioscience Database Center (NBDC) and the Database Center for Life Science (DBCLS) as a part of the Integrated Database Project to integrate life science databases in Japan. (Katayama et al. 2014, 14) Metadata activities at the BioHackathon could be grouped into three areas of focus: service quality indicators, database content descriptors, and a broader inclusive discussion of generic metadata that could be used to characterize datasets in a database catalogue for enhanced data discovery, assessment, and access (not limited to but still useful for biodatabases). “The BioHackers coined the phrase “Yummy Data” as a shorthand way of expressing not only data quality, but more importantly, the ability to explicitly determine the quality of a given dataset.” (Katayama et al. 2014, 19)

[iv] “The data capture, sorting, and use performed as part of knowledge discovery can be empowering if it is conducted by the people most affected by the data: research subjects themselves.” (Cargo & Mercer as cited by Shilton 2012, 1908)

[v] The social relations and institutional structures secured by participatory personal data are still forming. The capture and control of participatory personal data are distributed, and people from marginalized social positions may use the power to collect and analyze data to confront the powerful… to hold authorities accountable. (Shilton 2012, 1907-8)

[vi] Attempts to collect DNA samples from indigenous populations have met with similar concerns, as well as post-colonial criticism. (Coupland et al., 2005, 1755)

[vii] For more discussion of property rights and legal approaches to genetic data, see also McArdle, Jennifer. "Rethinking Property in the Digital Era." National Interest. 5 May 2014. < http://nationalinterest.org/commentary/rethinking-property-the-digital-era-10388?page=2> Web. 4 May 2014.

[viii] "Generally speaking, fields such as statistics, computer science and the hard sciences don't teach ethics," says Dr. Rachel Schutt, an adjunct professor at Columbia's Institute for Data Sciences and Engineering. "There are privacy concerns, such as how much corporations and the government should know about individuals…. But software engineers [are taught] about the elegance or the mathematical beauty of the thing that they're building, not how it will affect people's lives." (Shin 2013)

[ix] "If you do these large social network studies, you don't have what they call participant-informed consent. Let's say I have you in one of my Facebook studies, and you're coming to my lab and we are analyzing the strength of the connections between you and your friends. I'm getting information about your friends and their friends without their consent. It's a very, very ethically sensitive area." - Karrie Karahalios, a computer science professor at the University of Illinois at Urbana-Champaign (Shin 2013)

[x] Such zombies, like Haraway's cyborgs, ironically need not the empowerment or participation they are promised. In the ironic myth of the "Cyborg Manifesto", the machine-organism hybrid cyborg has no 'origin story.' "From one perspective, a cyborg world is about the final imposition of a grid of control on the planet, about the final abstraction embodied in a Star Wars apocalypse waged in the name of defence…" (Haraway 1991)

[xi] For a classic work on how computers and persons differ (and how the end-goal of data gathering signifies one such difference between computers and persons) see Dreyfus, Hubert and Stuart. "Why Computers May Never Think Like People." Technology Review MIT Technology, 1986. Rpt. in Readings in the Philosophy of Technology. Ed. David M. Kaplan. New York: Rowman and Littlefield, 2004. 375-390. Print.

[xii] Other scholars who have addressed Twain as a commentator on the issues of technology, information, and industry include Spangler 1970, Gale 1979, Railton 2002, and Royal 2002.

[xiii] Twain invokes a technocratic idea of progress when he claims that "the valuable part" of what we call civilization "had no existence when Queen Victoria emerged upon the planet." (Strout 2012, 337)