Register Login

Current Archives About Search

Home / Archives / Vol. 24 No. 2 (2021): dark / Articles

Darkness, Datafication, and Provenance as an Illuminating Methodology

Suneel Jethani
University of Technology Sydney

Robbie Fordyce
Monash University

DOI: https://doi.org/10.5204/mcj.2758

Vol. 24 No. 2 (2021): dark
Articles

> Data are generated and employed for many ends, including governing societies, managing organisations, leveraging profit, and regulating places. In all these cases, data are key inputs into systems that paradoxically are implemented in the name of making societies more secure, safe, competitive, productive, efficient, transparent and accountable, yet do so through processes that monitor, discipline, repress, coerce, and exploit people. (Kitchin, 165)

Introduction

Provenance refers to the place of origin or earliest known history of a thing. It refers to the custodial history of objects. It is a term that is commonly used in the art-world but also has come into the language of other disciplines such as computer science. It has also been applied in reference to the transactional nature of objects in supply chains and circular economies. In an interview with Scotland's Institute for Public Policy Research, Adam Greenfield suggests that provenance has a role to play in the "establishment of reliability" given that a "transaction or artifact has a specified provenance, then that assertion can be tested and verified to the satisfaction of all parities" (Lawrence).

Recent debates on the unrecognised effects of digital media have convincingly argued that data is fully embroiled within capitalism, but it is necessary to remember that data is more than just a transactable commodity. One challenge in bringing processes of datafication into critical light is how we understand what happens to data from its point of acquisition to the point where it becomes instrumental in the production of outcomes that are of ethical concern. All data gather their meaning through relationality; whether acting as a representation of an exterior world or representing relations between other data points. Data *objectifies* relations, and despite any higher-order complexities, at its core, data is involved in *factualising* a relation into a binary. Assumptions like these about data shape reasoning, decision-making and evidence-based practice in private, personal and economic contexts.

If processes of datafication are to be better understood, then we need to seek out conceptual frameworks that are adequate to the way that data is used and understood by its users. Deborah Lupton suggests that often we give data "other vital capacities because they are about human life itself, have implications for human life opportunities and livelihoods, [and] can have recursive effects on human lives (shaping action and concepts of embodiment ... selfhood [and subjectivity]) and generate economic value".

But when data are afforded such capacities, the analysis of its politics also calls for us to "consider context" and "making the labour [of datafication] visible" (D'Ignazio and Klein). For Jenny L. Davis, getting beyond simply thinking about *what* data affords involves bringing to light *how* continually and dynamically to *requests*, *demands*, *encourages*, *discourages*, and *refuses* certain operations and interpretations. It is in this re-orientation of the question from *what* to *how* where "practical analytical tool[s]" (Davis) can be found. Davis writes:

> requests and demands are bids placed by technological objects, on user-subjects. Encourage, discourage and refuse are the ways technologies respond to bids user-subjects place upon them. Allow pertains equally to bids from technological objects and the object's response to user-subjects. (Davis)

Building on Lupton, Davis, and D'Ignazio and Klein, we see three principles that we consider crucial for work on data, darkness and light:

1. data is not simply a technological object that exists within sociotechnical systems without having undergone any priming or processing, so as a consequence the data collecting entity imposes standards and way of imagining data before it comes into contact with user-subjects;
2. data is not neutral and does not possess qualities that make it equivalent to the things that it comes to represent;
3. data is partial, situated, and contingent on technical processes, but the outcomes of its use afford it properties beyond those that are purely informational.

This article builds from these principles and traces a framework for investigating the complications arising when data moves from one context to another. We draw from the "data provenance" as it is applied in the computing and informational sciences where it is used to query the location and accuracy of data in databases. In developing "data provenance", we adapt provenance from an approach that solely focuses on technical infrastructures and material processes that move data from one place to another and turn to sociotechnical, institutional, and discursive forces that bring about data acquisition, sharing, interpretation, and re-use. As data passes through open, opaque, and darkened spaces within sociotechnical systems, we argue that provenance can shed light on gaps and overlaps in technical, legal, ethical, and ideological forms of data governance. Whether data becomes exclusive by moving from light to dark (as has happened with the removal of many pages and links from Facebook around the Australian news revenue-sharing bill), or is publicised by shifting from dark to light (such as the Australian government releasing investigative journalist Andie Fox's welfare history to the press), or even recontextualised from one dark space to another (as with genetic data shifting from medical to legal contexts, or the theft of personal financial data), there is still a process of transmission here that we can assess and critique through provenance. These different modalities, which guide data acquisition, sharing, interpretation, and re-use, cascade and influence different elements and apparatuses within data-driven sociotechnical systems to different extents depending on context. Attempts to illuminate and make sense of these complex forces, we argue, exposes data-driven practices as inherently political in terms of whose interests they serve.

Provenance in Darkness and in Light

When processes of data capture, sharing, interpretation, and re-use are obscured, it impacts on the extent to which we might retrospectively examine cases where malpractice in responsible data custodianship and stewardship has occurred, because it makes it difficult to see how things have been rendered real and knowable, changed over time, had causality ascribed to them, and to what degree of confidence a decision has been made based on a given dataset. To borrow from this issue's concerns, the paradigm of dark spaces covers a range of different kinds of valences on the idea of private, secret, or exclusive contexts. We can parallel it with the idea of 'light' spaces, which equally holds a range of different concepts about what is open, public, or accessible. For instance, in the use of social data garnered from online platforms, the practices of academic researchers and analysts working in the private sector often fall within a grey zone when it comes to consent and transparency. Here the binary notion of public and private is complicated by the passage of data from light to dark (and back to light). Writing in a different context, Michael Warner complicates the notion of publicness. He observes that the idea of something being public is in and of itself always sectioned off, divorced from being fully generalisable, and it is "just whatever people in a given context think it is" (11). Michael Hardt and Antonio Negri argue that publicness is already shadowed by an idea of state ownership, leaving us in a situation where public and private already both sit on the same side of the propertied/commons divide as if the "only alternative to the private is the public, that is, what is managed and regulated by states and other governmental authorities" (vii). The same can be said about the way data is conceived as a public good or common asset.

These ideas of light and dark are useful categorisations for deliberately moving past the tensions that arise when trying to qualify different subspecies of privacy and openness. The problem with specific linguistic dyads of private vs. public, or open vs. closed, and so on, is that they are embedded within legal, moral, technical, economic, or rhetorical distinctions that already involve normative judgements on whether such categories are appropriate or valid. Data may be located in a dark space for legal reasons that fall under the legal domain of 'private' or it may be dark because it has been stolen. It may simply be inaccessible, encrypted away behind a lost password on a forgotten external drive. Equally, there are distinctions around lightness that can be glossed – the openness of Open Data (see: theodi.org) is of an entirely separate category to the AACS encryption key, which was illegally but enthusiastically shared across the internet in 2007 to the point where it is now accessible on Wikipedia. The language of light and dark spaces allows us to cut across these distinctions and discuss in deliberately loose terms the degree to which something is accessed, with any normative judgments reserved for the cases themselves.

Data provenance, in this sense, can be used as a methodology to critique the way that data is recontextualised from light to dark, dark to light, and even within these distinctions. Data provenance critiques the way that data is presented as if it were "there for the taking". This also suggests that when data is used for some or another secondary purpose – generally for value creation – some form of closure or darkening is to be expected. Data in the public domain is more than simply a specific informational thing: there is always context, and this contextual specificity, we argue, extends far beyond anything that can be captured in a metadata schema or a licensing model. Even the transfer of data from one open, public, or light context to another will evoke new degrees of openness and luminosity that should not be assumed to be straightforward. And with this a new set of relations between data-user-subjects and stewards emerges.

The movement of data between public and private contexts by virtue of the growing amount of personal information that is generated through the traces left behind as people make use of increasingly digitised services going about their everyday lives means that data-motile processes are constantly occurring behind the scenes – in darkness – where it comes into the view, or possession, of third parties without obvious mechanisms of consent, disclosure, or justification. Given that there are "many hands" (D'Iganzio and Klein) involved in making data portable between light and dark spaces, equally there can be diversity in the approaches taken to generate critical literacies of these relations. There are two complexities that we argue are important for considering the ethics of data motility from light to dark, and this differs from the concerns that we might have when we think about other illuminating tactics such as open data publishing, freedom-of-information requests, or when data is anonymously leaked in the public interest. The first is that the terms of ethics must be communicable to individuals and groups whose data literacy may be low, effectively non-existent, or not oriented around the objective of upholding or generating data-luminosity as an element of a wider, more general form of responsible data stewardship.

Historically, a productive approach to data literacy has been finding appropriate metaphors from adjacent fields that can help add depth – by way of analogy – to understanding data motility. Here we return to our earlier assertion that data is more than simply a transactable commodity. Consider the notion of "giving" and "taking" in the context of darkness and light. The analogy of giving and taking is deeply embedded into the notion of data acquisition and sharing by virtue of the etymology of the word data itself: in Latin, "things having been given", whereby in French *données*, a natural gift, perhaps one that is given to those that attempt capture for the purposes of empiricism – representation in quantitative form is a *quality* that is given to phenomena being brought *into* the light. However, in the contemporary parlance of "analytics" data is "taken" in the form of recording, measuring, and tracking. Data is considered to be something valuable enough to give or take because of its capacity to stand in for real things. The empiricist's preferred method is to take rather than to accept what is given (Kitchin, 2); the data-capitalist's is to incentivise the act of giving or to take what is already given (or yet to be taken). Because data-motile processes are not simply passive forms of reading what is contained within a dataset, the materiality and subjectivity of data extraction and interpretation is something that should not be ignored. These processes represent the recontextualisation of data from one space to another and are expressed in the landmark case of Cambridge Analytica, where a private research company extracted data from Facebook and used it to engage in psychometric analysis of unknowing users.

Data Capture Mechanism Characteristics and Approach to Data Stewardship	
Historical	Information created, recorded, or gathered about people of things directly from the source or a delegate but accessed for secondary purposes.
Observational	Represents patterns and realities of everyday life, collected by subjects by their own choice and with some degree of discretion over the methods. Third parties access this data through reciprocal arrangement with the subject (e.g., in exchange for providing a digital service such as online shopping, banking, healthcare, or social networking).
Purposeful	Data gathered with a specific purpose in mind and collected with the objective to manipulate its analysis to achieve certain ends.
Integrative	Places less emphasis on specific data types but rather looks towards social and cultural factors that afford access to and facilitate the integration and linkage of disparate datasets

Table 1: Mechanisms of Data Capture

There are ethical challenges associated with data that has been sourced from pre-existing sets or that has been extracted from websites and online platforms through scraping data and then enriching it through cleaning, annotation, de-identification, aggregation, or linking to other data sources (tab. 1). As a way to address this challenge, our suggestion of "data provenance" can be defined as *where* a data point comes from, *how* it came into being, and how it became valuable for some or another purpose. In developing this idea, we borrow from both the computational and biological sciences (Buneman et al.) where provenance, as a form of qualitative inquiry into data-motile processes, centres around understanding the origin of a data point as part of a broader almost forensic analysis of quality and error-potential in datasets. Provenance is an evaluation of *a priori* computational inputs and outputs from the results of database queries and audits.

Provenance can also be applied to other contexts where data passes through sociotechnical systems, such as behavioural analytics, targeted advertising, machine learning, and algorithmic decision-making. Conventionally, data provenance is based on understanding *where* data has come from and *why* it was collected. Both these questions are concerned with the evaluation of the *nature* of a data point within the wider context of a database that is itself situated within a larger sociotechnical system where the data is made available for use. In its conventional sense, provenance is a means of ensuring that a data point is maintained as a single source of truth (Buneman, 89), and by way of a reproducible mechanism which allows for its path through a set of technical processes, it affords the assessment of a how reliable a system's output might be by sheer virtue of the ability for one to retrace the steps from point A to B. "Where" and "why" questions are illuminating because they offer an ends-and-means view of the relation between the origins and ultimate uses of a given data point or set. Provenance is interesting when studying data luminosity because means and ends have much to tell us about the origins and uses of data in ways that gesture towards a more accurate and structured research agenda for data ethics that takes the emphasis away from individual moral patients and reorients it towards practices that occur within information management environments. Provenance offers researchers seeking to study data-driven practices a similar heuristic to a journalist's line of questioning *who, what, when, where,* why, and *how?*

This last question of *how* is something that can be incorporated into conventional models of provenance that make it useful in data ethics. The question of *how* data comes into being extends questions of power, legality, literacy, permission-seeking, and harm in an entangled way and notes how these factors shape the nature of personal data as it moves between contexts. Forms of provenance accumulate from transaction to transaction, cascading along, as a dataset 'picks up' the types of provenance that have led to its creation. This may involve multiple forms of overlapping provenance – methodological and epistemological, legal and illegal – which modulate different elements and apparatuses. Provenance, we argue is an important methodological consideration for workers in the humanities and social sciences.

Provenance provides a set of shared questions on which models of transparency, accountability, and trust may be established. It points us towards tactics that might help data-subjects understand privacy in a contextual manner (Nissenbaum) and even establish practices of obfuscation and "informational self-defence" against regimes of datafication (Brunton and Nissenbaum). Here provenance is not just a declaration of *what* means and ends of data capture, sharing, linkage, and analysis are. We sketch the outlines of a provenance model in table 2 below.

Type	**Metaphorical frame**	**Dark**	**Light**
What?	The epistemological structure of a database determines the accuracy of subsequent decisions. Data must be consistent.	What data is asked of a person beyond what is strictly needed for service delivery.	Data that is collected for a specific stated purpose with informed consent from the data-subject. How does the decision about what to collect disrupt existing polities and communities? What demands for conformity does the database make of its subjects?
Where?	The contents of a database is important for making informed decisions. Data must be represented.	The parameters of inclusion/exclusion that create unjust risks or costs to people because of their inclusion or exclusion in a dataset.	The parameters of inclusion or exclusion that afford individuals representation or acknowledgement by being included or excluded from a dataset.

			How are populations recruited into a dataset? What divides exist that systematically exclude individuals?
Who?	Who has access to data, and how privacy is framed is important for the security of data-subjects. Data access is political.	Access to the data by parties not disclosed to the data-subject.	Who has collected the data and who has or will access it? How is the data made available to those beyond the data subjects?
How?	Data is created with a purpose and is never neutral. Data is instrumental.	How the data is used, to what ends, discursively, practically, instrumentally. Is it a private record, a source of value creation, the subject of extortion or blackmail?	How the data was intended to be used at the time that it was collected.
Why?	Data is created by people who are shaped by ideological factors. Data has potential.	The political rationality that shapes data governance with regard to technological innovation.	The trade-offs that are made known to individuals when they contribute data into sociotechnical systems over which they have limited control.

Table 2: Forms of Data Provenance

Conclusion

As an illuminating methodology, provenance offers a specific line of questioning practices that take information through darkness and light. The emphasis that it places on a narrative for data assets themselves (asking what when, who, how, and why) offers a mechanism for traceability and has potential for application across contexts and cases that allows us to see data malpractice as something that can be productively generalised and understood as a series of ideologically driven technical events with social and political consequences without being marred by perceptions of exceptionality of individual, localised cases of data harm or data violence.

References

Brunton, Finn, and Helen Nissenbaum. "Political and Ethical Perspectives on Data Obfuscation." *Privacy, Due Process and the Computational Turn: The Philosophy of Law Meets the Philosophy of Technology*. Eds. Mireille Hildebrandt and Katja de Vries. New York: Routledge, 2013. 171-195.

Buneman, Peter, Sanjeev Khanna, and Wang-Chiew Tan. "Data Provenance: Some Basic Issues." *International Conference on Foundations of Software Technology and Theoretical Computer Science*. Berlin: Springer, 2000.

Davis, Jenny L. *How Artifacts Afford: The Power and Politics of Everyday Things*. Cambridge: MIT Press, 2020.

D'Ignazio, Catherine, and Lauren F. Klein. *Data Feminism*. Cambridge: MIT Press, 2020.

Hardt, Michael, and Antonio Negri. *Commonwealth*. Cambridge: Harvard UP, 2009.

Kitchin, Rob. "Big Data, New Epistemologies and Paradigm Shifts." *Big Data & Society* 1.1 (2014).

Lawrence, Matthew. "Emerging Technology: An Interview with Adam Greenfield. 'God Forbid That Anyone Stopped to Ask What Harm This Might Do to Us'. Institute for Public Policy Research, 13 Oct. 2017. <https://www.ippr.org/juncture-item/emerging-technology-an-interview-with-adam-greenfield-god-forbid-that-anyone-stopped-to-ask-what-harm-this-might-do-us>.

Lupton, Deborah. "Vital Materialism and the Thing-Power of Lively Digital Data." *Social Theory, Health and Education.* Eds. Deana Leahy, Katie Fitzpatrick, and Jan Wright. London: Routledge, 2018.

Nissenbaum, Helen F. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford: Stanford Law Books, 2020.

Warner, Michael. "Publics and Counterpublics." *Public Culture* 14.1 (2002): 49-90.

License

M/C JOURNAL
HOME
CURRENT ISSUE
UPCOMING ISSUES
ARCHIVES
CONTRIBUTORS
ABOUT M/C JOURNAL
USER HOME

JOURNAL CONTENT
SEARCH
BY AUTHOR
BY ISSUE

USER
LOGIN

CURRENT ISSUE
ATOM 1.0
RSS 2.0
RSS 1.0

INFORMATION
FOR READERS
FOR AUTHORS
FOR LIBRARIANS

FONT SIZE

an M/C publication | Supported by QUT creative industries