Big Data and the Study of Dementia: Epistemological Promises and Pitfalls

D. Grant Campbell
Faculty of Information and Media Studies
University of Western Ontario
London, Ontario N6A 5B7
gcampbel@uwo.ca

ABSTRACT
This paper examines the potential contribution of big data analytical methods to our understanding of dementia and the best means of caring for those afflicted with dementia. Using Ian Hacking’s history of statistics as a basis, the paper examines the three fundamental premises of big data: large datasets, a high tolerance of error and a focus on correlation rather than causation. Big data both evokes and departs from the epistemological assumptions that gave rise to statistical analysis, and its innovations merit both optimism and caution. While big data offers important means of reframing questions of dementia and dementia care along lines more conducive to the needs of patients and caregivers, the enthusiastic adoption of big data in business circles threatens to distort the practices into cost-saving measures that reflect false efficiencies rather than genuine improvements.

Keywords
Big data, predictive analytics, dementia, epistemology

INTRODUCTION
In the past few years, big data has emerged as a means of solving just about every problem that besets the human condition, from recommending entertainment to circumventing environmental disaster. As dementia increases in our aging population, and as the cure for its various forms shows no sign of appearing soon, it is only natural to speculate on the possible applications of big data analysis to the discovery of patterns and correlations that could lead to significant and useful insights, both in the diagnosis and treatment of dementia, and in the compassionate and effective treatment of those afflicted. It is equally natural for information scientists to speculate on the potential role of information studies in these applications. This paper uses a philosophy of information framework to assess the claims of big data enthusiasts Viktor Mayer-Schonberger and Thomas Cukier against two separate backdrops: current research in the care of dementia patients, and the history of data analysis as an instance of specific epistemological assumptions. The result of the analysis is a cautious and wary optimism: big data methods, I argue, might well lead to enhancements in the care of dementia patients, but only if its assumptions are held in vigilant check.

BIG DATA ASSUMPTIONS
Mayer-Schonberger and Cukier (2013) argue that the revolutionary significance of big data methods lies in their adoption of three presuppositions, each of which flies in the face of standard empirical analysis: large data collections, messy and inconsistent data, and a focus on correlation over causation.

First, big data is premised on very large datasets: datasets that would have been either impossible or unfeasible to create before modern technology’s new affordances for data collection. This enhanced volume shifts data analysis from methods grounded in sampling and inference to methods grounded in census data collection and analysis:

Reaching for a random sample in the age of big data is like clutching at a horse whip in the era of the motor car. We can still use sampling in certain contexts, but it need not—and will not—be the predominant way we analyze large datasets. Increasingly, we will aim to go for it all. (Mayer-Schonberger & Cukier 2013, 31)

Second, the very size of the datasets, and the diversity of their origins and structures, implies a much higher tolerance for errors within the data. Since measurement is presumably being done upon a population, the researcher need not be as concerned about the inferential power of the
data, and therefore need neither preserve nor display for scrutiny the data’s integrity as a representative sample. Furthermore, if the data points are numerous, the consequences of error or failure in the collection of one piece of data are significantly less. For Mayer-Schonberger and Cukier, rigorous concerns with data quality are a hallmark of sampling in “an information-deprived analog era” (2013, 40), and blind us to the possibilities of exploring large, messy datasets. Using Twitter user tags as an example, they argue:

in return for messiness in the way we organize our photo collections, we gain a much richer universe of labels, and by extension, a deeper, broader access to our pictures. (2013, 43)

Only our own fears, it seems, keep us from the joys of being richly inconsistent.

Finally, big data is concerned less with causation than with correlation. As such, this analysis does not extract variables from their context in order to examine their interaction and determine the effects of one variable upon the other. Instead of causation, big data works on the principles of the predictive model: “a mechanism that predicts the behavior of an individual … It takes characteristics of the individual as input and provides a predictive score as output” (Siegel 2013, 26). Big data analysis involves the search for patterns within the dataset: patterns that reflect how variables behave within their contexts:

Correlations exist; we can show them mathematically. We can’t easily do the same for causal links. So we would do well to hold off from trying to explain the reason behind the correlations: the why instead of the what. (Myer-Schonberger & Cukier 201, 67).

The implication of this statement is startling and contains a curious irony. Statisticians have struggled all along to distinguish causation and correlation, offering up instances of inference that are rendered absurd by confusing the two. Scientists have always known that causal links are difficult to find, which explains the rigour of both statistical inference and experimental design. But none of it particularly requires information science. If, however, we look at big data from the perspective of the philosophy of information, and in particular through the prism of Ian Hacking’s The Taming of Chance (1990), we find two things. First, the epistemological assumptions of big data are less dramatic than they might at first seem; second, that far from breaking from previous epistemologies of data, big data analysis repositions those paradigms in ways that justify responses of both optimism and caution. The remainder of this paper attempts to articulate the reasons for both.

**THE PROMISE FOR DEMENTIA STUDIES**

The significance of big data analysis for the study of dementia has yet to be fully understood, but early signs are promising. Many manifestations of dementia involve some form of aphasia which renders the patient unable to communicate with others. Advanced dementia frequently causes what used to be called “problem behaviours”: wandering, exit seeking, shouting, grabbing, biting and cursing. These behaviours are often bewildering or disturbing to family and caregivers, and while they are generally considered to be responses to stimuli, the patient’s inability to articulate these responses makes them difficult to predict or control. However, collecting large amounts of data about wandering patterns, repetitive behavior, movements and verbalizations, and correlating them with other data such as time and the occurrences of regular or unusual events, may enable researchers to detect the hidden reasons behind a patient’s seemingly cryptic behavior, and make it easier for caregivers to respond appropriately.

Big data assumptions have figured implicitly for some time in research on smart home design, in which analysis of patient behaviour enables the computer to prompt the resident with suggestions (Orpwood, et al. 2005, 205; Orpwood, et al. 2010, 156). Conceivably, such analysis could be expanded by combining data from multiple sources and looking for patterns of correlation.

In addition, big data analysis might help to separate the analysis of patient behavior from the caregiving narratives that explain such behavior. Nursing and caregiving communities have developed narrative structures to explain what are now called “responsive behaviours,” and to create compassionate procedures for coping with them. These narratives proceed from a sound and humane assumption: that every action by a dementia patient, however bewildering it may seem, has a reason, and is a response to something (University of Waterloo 2005, 5). However, the strength of these narratives lies less in their accuracy than in the emotional, professional and procedural guidance they provide for the caregiver dealing with the situation (Campbell 2014). Supplementing such structures and narratives with data that analyzes correlations may leave the care community open to surprises and insights that might tangibly relieve certain situations, and offer solutions to problems long assumed to be permanently opaque.

This could be very exciting. But none of it particularly requires information science. If, however, we look at big data from the perspective of the philosophy of information, and in particular through the prism of Ian Hacking’s The Taming of Chance (1990), we find two things. First, the epistemological assumptions of big data are less dramatic than they might at first seem; second, that far from breaking from previous epistemologies of data, big data analysis repositions those paradigms in ways that justify responses of both optimism and caution. The remainder of this paper attempts to articulate the reasons for both.

**PHILOSOPHY OF INFORMATION**

For Hacking in The Taming of Chance, our modern conceptions of knowing about the world date back to a transformation that took place at the beginning of the nineteenth century: a transformation that involved two related but distinct changes. First, philosophy in the early nineteenth century began to break from the determinist frameworks of eighteenth-century thought: “Causality, long
of metaphysics, was toppled or at least tilted: the past does not determine exactly what happens next (1). Determinism, as a philosophical approach, rests upon an assumption of the primacy of causation:

Determinism [is] the doctrine that every event has a cause. The usual explanation of this is that for every event, there is some antecedent state, related in such a way that it would break a law of nature for this antecedent state to exist yet the event not to happen. (Blackburn 2008, 97)

If all phenomena have an original cause, and rational thought pursues and detects the underlying causal relations between phenomena, then chance signifies chaos: something to be rejected. As the primacy of causation waned, a greater interest sprang up in chance, and the concept of a “normal” population began to gain traction. Determinism began to yield to a way of thinking that sought to analyze the patterns of chance, particularly the normal distribution (Hacking 1990, 163).

This eclipse of determinism was made possible largely by the enormous rise in the quantity of printed data in the early nineteenth century, as more and more governments began to accumulate more and more census reports. The opportunity to collect larger data sets and to analyze them in hitherto neglected ways gave rise to the belief, entrenched in empirical research methods, that data can tell stories that humans cannot tell, and that these stories, in their impersonal patterns, have a credibility that counterbalances the biases, delusions and false impressions that mark human accounts.

Long before we had big data, then, Western thought developed a skepticism towards causality as a means of knowing the truth of things. Certainly, the break was not firm or complete; empirical experiments continue to isolate variables and observe their interaction in an effort to discover causal relationships that have significant and useful predictive power in real-life situations. But the groundwork for big data was prepared in a gap made possible by this epistemological shift: the link between causation and understanding, while still prevalent, was no longer seen as necessarily inevitable.

In the same way, the shift described by Hacking offered conceptual room both to investigate the patterns of data distributions and to admire their beauty and their capacity to surprise. While, to be sure, data collection methods throughout the nineteenth century never approached the levels of modern affordances, and therefore required elaborate and sophisticated sampling methods, the rise in data analysis laid the basis for big data analysis in its fascination with data patterns that emerge as local eccentricities merge into greater, more stable verities brought about, not by causality’s defeat of chance, but by a yielding to the mysteries of chance and probability.

Hacking, however, notes two effects of the rise of data analysis: an emphasis on control, and an exaltation of the principle of “normality.” As the power of data analysis became more obvious, the mechanisms for authorizing, collecting, analyzing, and measuring data became entrenched in national bureaucracies as a means of monitoring and controlling populations: “There is a seeming paradox: the more the indeterminism, the more the control. … Behind it lay new technologies for classifying and enumerating, and new bureaucracies with the authority and continuity to deploy the technology” (Hacking 1990, 2,3).

As Foucault notes in much of his work, the power to control lies not merely in the power to suppress or to exclude, but in the power to classify, measure, analyze and interpret. In The History of Sexuality, for instance, he argues that an discursive interaction was established between activities and behaviours perceived to be “unnatural” and the entities that sought to exert power by drawing out the very unnatural practices that it monitored (1990, 45).

The power of the normal distribution created a self-perpetuating tendency, particularly in the social sciences, to exact and encourage conformity to social definitions of “the normal”:

People are normal if they conform to the central tendency of such laws, while those at the extremes are pathological. Few of us fancy being pathological, so ‘most of us’ try to make ourselves normal, which in turn affects what is normal. (Hacking 1990, 2)

Accordingly, this led to rough and sometimes cruel distinctions between normal and abnormal existence and behavior.

**DATA ANALYSIS AND DEMENTIA**

The new data environment described by Hacking has had an important effect on the shaping of conventional understandings of dementia and how it should be treated. For one thing, dementia is typically seen as a form of madness: that state of mind that, according to Foucault, occupies a simultaneously marginalized and exalted position in contemporary discourse, in which the mad person is seen both as someone whose utterance possesses a special truth-telling power and at the same time as someone whose discourse lies outside our conventional frames of reference (2010). Even other marginalized communities can often see dementia as the last straw. “Don’t go crazy on me, girlfriend,” warns the kind and wise caregiver in Tony Kushner’s AIDS epic, Angels in America: “I already got enough crazy queens for one lifetime. For two. I can’t be bothering with dementia” (61).

Health information sources that describe dementia typically emphasize that the disorder falls outside the category of “normal.” The National Library of Medicine, for instance, states that while memory loss is a normal part of aging,
Alzheimer’s Disease is not a normal aging process (A.D.A.M. 2010). The psychological effects of being branded “abnormal” are all the greater because of the frequent occurrence of aphasia in dementia, which renders communication difficult, even in the earlier stages. As a result, studies in dementia care have noted recurring problems with information communication: both medical professionals and caregivers tend to exchange information among themselves, frequently neglecting to address the patient directly or to allow the patient to participate meaningfully in important health and life decisions (Featherstone et al. 2013, 143).

Further evidence from the dementia literature in information science, nursing and medicine indicates that dementia research continually negotiates an uneasy distinction between efforts to find causes and efforts to meliorate current conditions. Much of the medical literature understandably seeks to find causes for the various symptoms of dementia: the role of the hippocampal complex (Graham & Hodges 1997, 77); the influence of genes (Sarkar 2013, 249); whether semantic memory gives place to other, simpler forms of classification as dementia progresses (Patterson, Graham & Hodges 1994, 57; Au, Chan & Chiu 2003, 737). These causative relations are unquestionably important, and may make significant differences in our ability, not only to alleviate the symptoms of dementia but to isolate causes and find cures.

Nonetheless, the testimony of dementia patients themselves suggests the presence of very different attitudes towards their affliction. Textual analysis of the testimonies of dementia patients, using Farradane’s relational indexing matrix as a template, suggested that for many dementia patients, concern for causes yields to concern for placing themselves conceptually within their contemporary worlds, adapting to the change in their own identities, and planning for the future (Campbell 2011). Big data analysis may perhaps meet some of these non-causative needs, without underestimating the importance of traditional causative inquiries.

THE CASE FOR OPTIMISM

Big data offers two responses to traditional data analysis that bear promise for dementia, one of which supports that traditional paradigm and one which questions it. First, by further rupturing the connection between causality and understanding, big data analysis creates cultural and conceptual room for increasing numbers of data studies that seek to meliorate the effects of dementia without getting caught up in the causes. Information science and computer science offer the main avenues for improvement in this area. Innovations in smart home design have made us aware that algorithmic analysis of resident behaviour can assist in providing patients with safe and comfortable environments (Orpwood 2005; Orpwood et al. 2009). The discursive gap between causation and correlation also gives room for researchers to do justice to the nature of dementia as a disease which requires from the patient constant adaptation and improvisation. Patients repeatedly need to draw together their diminishing resources from multiple sources, including narrative structures, nonverbal gestures and notions of identity, merely to negotiate everyday encounters (Hyden & Orulv 2009). In an environment where the brain is perpetually rewiring itself, patterns of behavior can be more revealing than adherence to formal links of what does what.

Second, while big data analysis continues to use concepts of normality, the paradigm of the “normal distribution” is repositioned to represent, not the population being studied but the data being collected. Because data gathered often involves the entire population rather than a representative sample, there is less concern with inferring from the sample to the larger population. Instead, the normal distribution figures in the assumptions about data collection. Since the data is being collected from a multitude of data points, over considerable lengths of time, the data set can tolerate error to a greater degree. Individual errors in sensor readings or data entry, which might have been disastrous in a small sample, become mere outliers as the data set gets larger and larger. As long as the bulk of the data collected is “normal,” in the sense of being accurate and useful, a certain number of anomalous errors can be absorbed without difficulty.

The effect of this shift in the normal distribution, from the subjects to the data that represents the subjects, is potentially a liberating one, and information science, with its strong traditions of user studies and user-centered systems design, could acquire a stronger discursive scaffolding for presenting its findings. More and more, we are finding examples of information systems designers involving people with dementia in the design process (Astell et al. 2009, 49); they are also working to create more accurate representations of older people (Blythe & Dearden 2009, 21).

While not all future research into dementia will involve big data, then, the effects of big data on our understanding of dementia could be beneficial. Our increased tolerance for large data sets with messy but intriguing data could make it easier to assemble large-scale repositories for the analysis of cryptic behavior such as wandering. By focusing on correlation rather than causation, the analysis of such data could provide insights that we didn’t expect, and lead to new avenues for understanding dementia which, when subsequently analyzed in traditional ways, could yield up its causes and its optimal treatment. And by examining, not why people suffer cognitive impairment, but what they need to survive comfortably despite such impairment, our research, particularly in the areas of information seeking and use, could be significantly better aligned with what dementia patients actually want, and how they actually cope with an increasingly confusing world.
THE CASE FOR CAUTION

Evgeny Morozov, one of the most trenchant critics of the recent developments in big data and predictive analytics, argues in a very different direction. For Morozov, big data is only the latest manifestation of what he calls “solutionism”: the deliberate and large-scale generation of problems which have computational solutions (2013, 6). The fact that correlation is easy to detect mathematically, which attracts the enthusiasm of Mayer-Schoenberger and Cukier, is not, Morozov would argue, a happy coincidence. Rather, problems that call for detecting correlation are gaining higher prominence because they conform to a business, social, cultural and intellectual climate that favours a greater reliance on computational tools, and that aims to reframe as many problems as possible within a programming, and specifically a gaming paradigm (5).

Morozov may well be correct; nonetheless, his objections do not automatically discredit big data in its application to dementia care. In attempting to care for people with diminished communicative ability who are prone to behaviours that appear incomprehensible, any kind of computation of correlations that might lead us past our ham-fisted guesses and makeshift solutions would be helpful.

Of greater concern to this author is the appeal of big data, not to the gaming consciousness, but to the areas of business intelligence and profit seeking. For Eric Siegel, predictive analytics is a key to retaining competitive advantage in business by enabling a company to compete from “the most powerful and unique competitive stronghold,” to enforce rules by “managing fraud,” and satisfying “today’s escalating consumer expectations” (2010).

At first glance, such entrepreneurial rhetoric seems far removed from the thankless, unglamorous and endless task of caring for patients with dementia. However, the rhetoric appeals to a desire to control fraud by enforcing rules, monitoring behavior, managing expectations and maintaining a competitive edge through predictive models that “learn” from an organization’s data. As such, the popular talk around big data is eerily reminiscent of the enthusiasm for enforcing control posited by Hacking: computations upon large groups of numbers for purposes of holding onto power, exercising surveillance and punishing the disobedient.

Again, in a dementia context, this is less alarming than it sounds. No one can work with dementia patients for more than a day without appreciating the necessity for constant surveillance, at a level which would be unacceptable in other cases. However, as the number of dementia patients rises, and as the strain on the public health care system increases, and as for-profit care facilities seek to establish their places in the elder care infrastructure, the ideals of big data might become the pretext for draconian computations which have everything to do with maximizing a surface efficiency without contributing to patient care.

Finally, the theoretical tolerance for messy data, while convincing in the context of big data analysis, could easily become a rationale for ignoring individual narratives of suffering in the face of collective narratives of improvement. The impact of dementia cannot be rendered solely through computation, and optimistic narratives at an intellectual and emotional distance do little to alleviate the close-proximity experience with this wasting affliction.

CONCLUSION

Morozov’s warnings about solutionism are germane and worthwhile, but they should not prevent us from exploring how big data could enhance our understanding of dementia, particularly in the context of behaviours that mark the later stages. Big data offers a shift in perspective and assumptions that provides additional discursive room for approaches and methods that could significantly alleviate some pressing problems in health care and elder care. We should, however, approach the promises and challenges of big data with caution. Many difficult problems can be ameliorated by breaking off easy problems and solving them through computation and correlation. But caring for people with dementia will never be free of deep moral, ethical and spiritual questions. And an approach that evades those deep waters in favour of computational ease may well play a vital and beneficial role, but it will never entirely suffice.

ACKNOWLEDGMENTS

The author wishes to acknowledge the contributions of colleagues Meredith Levine, Jacquelyn Burkell, and Nicole Dalmer at the University of Western Ontario. He also wishes to thank Rogers Communications, which has funded the Rogers Chair which has provided a framework of activities that support this research.

REFERENCES


scenarios. *Universal access to the information society* 8, 21-32.


