INFORMATION, COMMUNICATION & SOCIETY, 2016
http://dx.doi.org/10.1080/1369118X.2016.1211722

Taking Big Data apart: local readings of composite media collections

Yanni Alexander Loukissas

Program in Digital Media, School of Literature, Media and Communication, Georgia Institute of Technology, Atlanta, GA, USA

ABSTRACT

If we are to think critically about Big Data initiatives, we must learn to take them apart. This paper explains how to interrogate Big Data, not as large homogenous resources, but as heterogeneous collections with varied and discordant local ties. My argument focuses on the Big Data of media collections: composite digital repositories of texts, images, and video created in different contexts, but brought together online. The primary example used in this paper is the Digital Public Library of America (DPLA), a collection composed of digitized library, museum and archive records from institutions across the United States. I demonstrate how local readings of DPLA data can uncover schemata, errors, infrastructures, classifications, absences, and rituals that have important origins. Moreover, I explain how identifying these local features can support new forms of scholarship, pedagogy, and advocacy in the face of Big Data. For this last point, I use two additional cases: NewsScape, a real-time archive of broadcast news, and Zillow, a marketplace for real estate listings. The range of examples demonstrates how the stakes change from one Big Data initiative to the next. The paper concludes with a set of speculative guidelines for attending to the local conditions in Big Data: get dirty, take a comparative approach, show context, use data to connect people, and create opportunities for the collection of counter-data. When working with Big Data, I argue that thinking locally is thinking critically.

ARTICLE HISTORY
Received 12 January 2016
Accepted 5 July 2016

KEYWORDS
Big Data; local data; critical data studies; media collections; close reading

Introduction

Data are local: made by people and their dutiful machines, at a time, in a place, with the instruments at hand, in existing organizations, with limited resources, for disciplined audiences. In recent years, scholars of information have sought to understand the local contexts in which data are created (Bowker & Star, 1999), reused (Zimmerman, 2008), aggregated (Edwards, 2010), and exchanged (Vertesi & Dourish, 2011). This work has challenged the narrow instrumental conception of data as inscriptions: mobile and immutable representations of observed phenomena applicable primarily to the production of scientific arguments (Latour, 1990).

However, the Big Data phenomenon has largely been debated independently of local sites of production and use. Scholars of Big Data have sought to define their subjects in

CONTACT Yanni Alexander Loukissas ✉ yanni.loukissas@lmc.gatech.edu Program in Digital Media, School of Literature, Media and Communication, Georgia Institute of Technology, TSRB 85 5th Street NW, Rm 318A, Atlanta, GA 30308, USA

more general terms: technically, in terms of their 'volume, velocity and variety' (Kitchin & McArdle, 2016); historically, as a 'step change' from past practices (Gantz & Reinsel, 2011); practically, focusing on the difficulty of managing them (Schneiderman, 2014); and ideologically, by illuminating their 'mythology' (boyd & Crawford, 2012). Moreover, popular articles about Big Data depict them as ubiquitous tools for research and decision making across domains (Anderson, 2008; Lohr, 2012; Mayer-Schoenberger & Cukier, 2013). Such definitions keep discussions of Big Data separate from questions about the importance of their origins.

In this paper, I argue that attention to the local is necessary for developing critical discourses on Big Data. Scholarship in the emergent field of critical data studies - focused on 'the nature of data, how they are being produced, organized, analysed and employed, and how best to make sense of them and the work they do' (Kitchin & Lauriault, 2014, p. 1) – should reframe Big Data as a agglomerations of local data. I use the word agglomeration to focus attention on the distributed origins of Big Data. Indeed, Big Data are heterogeneous collections, created in varied sites of production and shaped by their conflicting values and norms. We should be asking: where do Big Data come from and how do the local conditions of their creation shape subsequent research and practice?

The implications of addressing this question are three-fold. For scholars, attention to the local offers an unprecedented opportunity to learn about varied cultures of data collection brought together in Big Data. For students, learning about the heterogeneity of Big Data can dispel the illusion that any data can offer what Haraway calls, 'the view from nowhere' (1988). Finally, an awareness of data's enduring local ties suggests new forms of social advocacy around Big Data. For wherever data travel, local communities of producers, users, and non-users are affected.

I draw upon examples from the Big Data of media collections: texts, images, and video brought together from distributed sources. Three cases support my argument. The Digital Public Library of America (DPLA), a non-profit organization, has linked millions of digitized resources from libraries, archives, and museums; NewsScape, an academic initiative based at the University of California, Los Angeles (UCLA), has a growing collection of video for more than 300,000 broadcast news programs, extending back to Watergate; Zillow, an online real estate marketplace, has drawn together sales data on more than 100 million homes in the United States. By including a range of cases, I hope to provide a broader sense of how Big Data are assembled and what work they can do.

My examples do not always conform to the strict definition of Big Data emerging from critical data studies: high in volume, variety, and velocity (Kitchin & McArdle, 2016). The DPLA, for example, is not a high-velocity initiative. Nevertheless, these examples are important references for examining the broader cultural phenomenon of Big Data, which I believe is better understood as a movement towards datasets that are comprehensive. In technical language, these are characterized as datasets in which n (the number of elements in the set) = all. (Mayer-Schoenberger & Cukier, 2013). This definition speaks to the universalizing ambitions of Big Data and prompts us to think about when and why they fall short.

I approach media collections data through a method I call *local reading*. This is a technique that I have adapted from close reading - originally developed for literary and cultural studies (Culler, 2010) - to examine how isolated features in Big Data produce meaning. In order to explain what is significant about local reading, it is useful to compare the practice to another common technique for analysing data, visualization. Data visualization is the practice of graphically presenting abstract patterns in data (Drucker, 2014). Current visualization techniques require a homogeneous dataset, which is achieved through reductive practices of parsing and filtering (Fry, 2007). Indeed, visualization is a method for examining the global, not the local features in data (Jänicke, Franzini, Faisal, & Scheuermann, 2015). In recent years, the popularity of data visualization has normalized both data and ways of looking at them. Local readings of data operate as a critical counterpoint to data visualization. They reveal what visualization does not: the heterogeneity of data before they are processed. Although the techniques of local reading and data visualization can be used in complementary ways, this paper relies exclusively on the former.

Local readings of data rely heavily on access to local knowledge. This necessary interpretive context can come from interviews with those who either make or make use of data. However, local does not mean exclusive. There may be several local contexts in which data have significance (Star & Griesemer, 1989). In preparation for this paper, I conducted semi-structured interviews around each of the three Big Data initiatives mentioned above, the DPLA, NewsScape, and Zillow: thirty in total.[1] I spoke with technologists and researchers directly involved in these projects as well as specialist librarians, journalists, and real estate agents who manage, contribute to, or work closely with data that have been 'ingested' or agglomerated in Big Data. I share only the most salient findings here. However, these findings are enough to illuminate a broader pattern that I have witnessed: Big Data are assembled from local conditions that are important for understanding the whole. In fact, by looking at Big Data as agglomerations of local data, we can learn about the heterogeneity of data in general and the importance of data's origins.

In summary, if we are to develop critical perspectives on Big Data, we must learn to take Big Data apart. This paper explains how to look for the local conditions in Big Data. Moreover, it illustrates new modes of scholarship, pedagogy, and advocacy that engage Big Data, not as large homogenous sources of information, but as sites of controversy (Latour, 1987) where varied conceptions of data come into conflict. The paper concludes with a set of speculative guidelines that can help researchers, students, and social advocates cultivate a new sensibility toward Big Data: get dirty, take a comparative approach, show context, use data to connect people, and create opportunities for the collection of counter-data.

Looking for the local

In order to understand Big Data as local, we must first clarify what 'local' means. Geertz explains local as a relative term.

> In the solar system, the earth is local; in the galaxy, the solar system is local; and in the universe, the galaxy is local. To a high energy physicist, the particle world – or zoo – is, well, the world. It's the particle, a thread of vapour in a cloud of droplets, that's local. (Geertz, 1992, p. 129)

In common use, the term gains relevance in relation to national or global contexts (Gieryn, 2000). In computing, local indicates the relative placement of a digital file: a folder is local on your hard drive; your hard drive is local in a network; your network is local on the Internet. Geertz offers that one local condition might be most productively understood – not in relation to some imagined universal – but relative to another local condition. In fact, he argues that all understanding is local. ‘No one knows everything’, writes Geertz, ‘because there is no everything to know’ (Geertz, 1992, p. 129). It is from this perspective, that we can see Big Data as assembled from local conditions.

In media collections data, there are many ways in which the simple act of reading can reveal how the local is manifest. Below, I demonstrate local readings of data using examples drawn from contributing collections to the DPLA. The primary interface to the initiative – a standard search bar with the heading ‘A Wealth of Knowledge: explore 11,578,169 items from libraries, archives, and museums’ – promises equal access to each contributing repository (Digital Public Library of America, 2016). But a basic search of the DPLA’s unified collections conceals the striking heterogeneity of the underlying data. The examples below variously reveal schemata, errors, terminologies, constraints, classifications, and rituals, all of which are rooted in local conditions.

Seeing schemata

In the collections data of the New York Public Library (NYPL), a major public research library and an early contributor to the DPLA, one can find at least 1719 unique date schemata – ways of recording the moment a book, image, or other library artefact came into the world. Below is a sample of abstracted date schemata. These are not actual dates. Rather, each represents one way of documenting a date of publication.

Printed by Thomas; Badger, Jun (1)

pref _ _ _ _] (1)

_ _ March, _ _ _ _ (3)

probably before _ _ _ _ (7)

[c_ _ _ _]/ _ _ _ _ (130)

_ _ _ _ - _ _ _ _ , _ _ _ _ - _ _ _ _ (209)

_ _ _ _ -_ _ _ _ , re- issued through _ _ _ _ (240)

_ _ _ _ -_ _ -_ _ /_ _ _ _ -_ _ -_ _ (438)

ca. _ _ _ _ ’s (640)

The ‘_’ in each schema is a variable standing in for a variety of possible integers. The number in parenthesis indicates the total times the format appears in the NYPL catalogue. Thus, the common schema ca. _ _ _ _’s, used 640 times, might be applied as ca. 1950s. The less common formats are typically extreme: either highly ambiguous or strangely specific. In one case cited here, the format includes the name of the printer. It appears only once. Although we cannot understand these obscure date schemata without further inquiry, we can imagine that each has its own local history.

Decoding errors (dirt)

Every contributing collection to the DPLA contains what appear to be errors. One does not need much instruction to notice misspelled words or misplaced punctuation. But it takes a degree of local knowledge to understand that such errors are not random. In fact, we can see them as evidence of localized cataloguing practices (Battles & Loukissas, 2013). Indeed, they stem from situated processes of data production.

Badly scanned text, blurred images, and moiré effects are a result of specific technologies and ways of making use of them within a local setting. Typographic errors sometimes originate in the use of type from a particular historical moment. They may be brought on by optical character misrecognition of unusual typefaces, ligatures, or unexpected characters. For example, the standing 's' in early-modern English typography is routinely mistaken for an 'f' by character recognition systems. Understanding this has value beyond the occasional amusement for contemporary readers. Alternatively, errors can be brought about when content is mistaken for code. Brackets, dollar signs, and semicolons can be interpreted as instructions to be carried out by a computer program.

We often hear about the arduous but imperative necessity to rid datasets of such flagrant errors through acts of cleaning or filtering. But as anthropologist Mary Douglas tells us, dirt is nothing more than matter out of place (1978). Indeed, dirty bits in collections data are local signifiers taken out of their interpretive context. We should learn to read such dirt as important traces of their own local production.

Attending to infrastructures

Ways of inscribing data are always constrained by infrastructural conditions. One well-known example of the technical limits on data comes from the turn of the millennium. Leading up to the year 2000, digitized date codes had to switch from two to four digits. Otherwise '00, 01, 02' might be mistaken as 1900, 1901, 1902 rather than 2000, 2001, 2002. Across contributing collections to the DPLA, databases had to be updated at great cost. And still, fears persisted that some unseen conflict in formats might cause whole systems to fail. Two-digit date formats are local markers of a specific technological moment. In previous eras, when storage space was much more expensive, programmers used two-digit date codes to save space. We should read such legacy infrastructure as evidence of the way that data are situated in time. Without the software and hardware of their era, as well as operating knowledge thereof, data would not be accessible at all.

Questioning classifications

Data are also shaped by local classifications. Audiences typically do not take notice of classifications unless their origins are unfamiliar. Jeffrey Licht, a technologist working for the DPLA, calls attention to a record that would appear unfamiliar outside of South Carolina.[2] The state's digital library, another early contributor to the DPLA, contains a group portrait from Clemson University with a field titled 'coverage'. The field contains a single string: 'upstate', presumably referring to a place in South Carolina. However, the string is meaningless to Licht and to me. Neither of us are from the region. But such language should not be presented as anomalous. Rather, it should compel us to think about the situated nature of all place names. Local classifications are a product of geographies as well as other social boundaries, such as those that separate disciplines.

Revealing absences

More subtly than in the previous examples, data are defined by what they leave out. Smithsonian historian Marya Mcquirter recounts having searched her catalogue, one of the largest contributors to the DPLA, for the terms 'black' and 'white'.[3] The first brings up various examples of African-American artists, for the museum diligently documents work created by artists who identify racially as 'black'. But the search term 'white' brings up little about race, other than the occasional piece linked to white supremacy. White is not a racial identity that the Smithsonian typically tracks. Instead, the category exists as an absence that reveals a bias. White, the racial identity of the vast majority of artists whose work is shown at the Smithsonian, is not critically examined by the institution. As this example demonstrates, absences are part of deeply rooted systems of representation reified in data.

Observing rituals

Finally, and less overtly visible, are the *rituals* that shape local data. I use the term ritual here to identify cultural practices with data that have their own significance as symbolic expressions or community-making activities (Gusterson, 1998). Thomas Ma, a cataloguer at the Harvard Library – another large contributor to the DPLA – reflects on the way cataloguing has changed over the course of his career.

> I remember when I first started at the law school, I was told by the person in charge of technical services that 'you weren't worth anything if you didn't have a backlog.' And nowadays it's like if you have a backlog, you have the cooties. So the backlog [used to be] evidence of a certain kind of care and quality and attention in the catalogue processing. And now the backlog is a distinct liability.[4]

In this example, practices with data are closely tied to professional identity and status. Moreover, the change in cataloguing practices has significant practical implications. For when backlogs pile up – sometimes consisting of tens of thousands of accumulated books that need to be processed – cataloguing is outsourced. Ma explains, 'We used a company in Arkansas. I guess they find cheap labour … they just grab people off the street and say, here, slap a record together and move it on.'[5] His words are laden with an implicit argument for preserving the social milieu in which he was trained. Indeed, without the proper rituals, argues Ma, the quality of library data is in danger. Ma forecasts a dark future for collections: they contain more artefacts than ever before, but their data – the maps to those collections – are increasingly thin. From this perspective, the movement toward Big Data can paradoxically make information less accessible. Ma's rituals are evidence of a local social order in which data are enmeshed.

As the examples above illustrate, there is no such thing as universal data. Data are situated within the means of their production, the infrastructure required to maintain them, their systems of representation, and the social order they reproduce. These insights build on social studies of information that focus on the specificity of data at the level of the institution, such as the museum (Star & Griesemer, 1989) or the laboratory (Latour & Woolgar, 1979). However, such often assume that standards are the primary forces that shape data. Gitelman writes 'every discipline and disciplinary institution has its own norms and standards for the imagination of data' (2013, p. 3). But as local readings of DPLA data suggest, variations in data can be the result of a number of historical, technological, and environmental contingencies, many of which are not merely institutional. Too often, these differences are passed off as anomalies, to be resolved by ‘normalizing’ data, a process by which data are made to conform to an expected range of categories and values. I argue that we should consider the utility of differences in data as markers of an otherwise invisible local context that is important for meaningful analysis.

The six features examined above – schemata, errors, infrastructures, classifications, absences, and rituals – do not constitute a fixed typology for local readings of data. Rather, they convey the contingent character of data through examples that cover a range of possible local ties. What appears to be local in data depends on emergent differences among data agglomerated in one place. As Geertz explains, the local is only intelligible when seen through a comparative lens. When data are drawn together from disparate origins, conflicting practices of data production are made apparent.

Local markers are particularly relevant when exploring Big Data. For although the term suggests a departure from the local, the rise of the Big Data phenomenon has ironically made the local qualities of data more significant. Under Big Data, distributed records with discordant local ties are increasingly estranged from their developers and presented to audiences other than those first intended. In the remainder of this paper, I explain how renewed attention to the local can help us critically engage Big Data for purposes of scholarship as well as pedagogy and advocacy. I will rely on local readings of the DPLA, but also NewsScape and Zillow. Showing a diversity of Big Data initiatives is important for conveying what can be at stake when local data are agglomerated on a massive scale.

Local conditions in Big Data

Today, there are a proliferation of initiatives that assemble Big Data from local, distributed sources like libraries, museums, and archives. Each Big Data push promises to reveal new patterns across previously independent datasets. However, these initiatives are not all the same in their motivations and goals. The DPLA, referenced throughout this paper, is an example from the non-profit world. Created with a mission ‘to educate, inform, and empower everyone in current and future generations’ (dp.la), the DPLA is supported by a combination of private foundations, donors, and United States federal research agencies. Meanwhile, academia and industry have their own models. NewsScape, which is supported entirely by public funds, is used for research in disciplines ranging from communication to computer science. Meanwhile, Zillow is a for-profit platform motivated by the potential for surplus value gained through the agglomeration of data.

Each of these Big Data projects manifests traces of the local with important implications for data-driven scholarship, pedagogy, and advocacy. However, I have chosen the most evocative case for each of the implications I wish to explain. Thus, I will use the DPLA to explain how attention to the local might serve scholarship in the area of critical data studies. I will rely on NewsScape to illustrate the local perspective on data pedagogy. Finally, I will show how Zillow necessitates new forms of advocacy around data.

The DPLA: a case for data studies

Big Data offer untapped opportunities for scholarly inquiry into the differing local conditions in which data are made. A straightforward resource for carrying out such a research program is the DPLA, for it can support comparative studies of cataloguing at institutions across the United States.

Nowhere are the differences across DPLA data more accessible to inquiry than in efforts to normalize them. One such effort, referred to as 'deduping' within the organization, was introduced at the DPLA's first 'hack-a-thon'– an event that brought together technologists, designers, librarians, and academics for three days of collaborative experimentation around their data. The term dedupe is shorthand for an automated process that will rid the DPLA of redundant entries. However, the process of identifying seemingly identical digital versions of books, newspapers, and other collection objects also reveals key differences in the data that have the potential to illuminate what each object means in its originating context.

For example, there are innumerable copies of Adam Smith's book *The Wealth of Nations* amidst the assembled collections of the DPLA. One might ask: why do users need access to all of these? Deduping might lead to a more streamlined experience, but it also erases historical context, which might be crucial to understanding how *Nations* and other popular texts were taken up over time (metaLAB (at) Harvard, 2016).

Duplicates are key to learning about the heterogeneity of the DPLA. Instead of riding the collection of redundancies, we might study them. When seen in this way, the DPLA exemplifies an opportunity to use Big Data to study the production of data – raising important questions about the local histories of heterogeneous cultures of collecting. Such work would serve the basic goals of critical data studies.

Newsscape: a case for data pedagogy

For pedagogy, as for scholarship, Big Data initiatives offer new tools with which to think about how data are made in varied local contexts. However, students at the undergraduate level sometimes have difficulty seeing the accumulated historical collections of the DPLA as directly relevant for their lives. As an alternative site in which to learn about the heterogeneity of data, consider NewsScape.[6] NewsScape addresses contemporary events of broad interest. Drawing together data from CNN, BBC, Fox News, MSNBC, Comedy Central, and Al Jazeera, among other news outlets, NewsScape can offer students an opportunity to reflect on how such events are framed in different ways by the social and technological conditions of individual news makers.

NewsScape has assembled video recordings as well as associated data including closed captioning, on-screen text, and other identifying information for more than 300,000 news programs into patchwork portrait of the news. Students examining this patchwork must grapple with gaps in local coverage, a lack of non-broadcast sources (such as social media and print), as well discordant audio, visual, and text data that present different analytic challenges and are not easily mined for corresponding themes. Finally, NewsScape demonstrates that when decoupled from regional and temporal contexts of interpretation, news content is not easily understood. For instance, students would have a hard time parsing stories on the 2012 hurricane Sandy from New York City without local knowledge about the yearly marathon it disrupted and the associated political fallout for Mayor Bloomberg.

NewsScape offers a resource for students to learn that data about the news, like the news itself, do not present 'the view from nowhere' (Haraway, 1988). Rather, students will find that the news does not submit to easy flattening as data in order to trace issues across time and news outlets. Thus, Big Data can be a component of a new critical data pedagogy that builds off of the lessons of critical data studies, but makes those lessons more concrete and subject to hands-on engagement by students.

Zillow: a case for data advocacy

Finally, Big Data raise new challenges for advocacy – to support communities, specifically, the most vulnerable, in voicing their opinions on issues that concern them. One example that makes the need for advocacy around Big Data apparent is Zillow.com, a commercial initiative that amasses data on house sales across the United States, as well as data on demographics, crime, schools, and other related and quantifiable concerns. Zillow's data come from public records as well as private real estate agents and even homeowners themselves. The company uses these data drawn from local conditions to generate values for more than 100 million homes, not just those on the market.

The company's 'Zestimates' (a term coined by Zillow) are generated by a proprietary set of algorithms that are opaque to its users. Zestimates are not simply representations of the housing market. They help to drive the market: by normalizing the definition of value to what can be measured universally and quantitatively (i.e., floor area, crime rates and school test scores but not historical significance or community cohesion); by enabling financial speculators to see housing as a homogeneous set of investment opportunities; by encouraging homeowners to anxiously check in on Zillow's evaluations of their property and contribute additional data; and by excluding those who are not already homeowners from participating in the determination of value in housing. When we understand data as an operational part of an economic system, it becomes important to ask: who creates the data? who are the data created for? who benefits and who does not?

Advocates for housing justice in local contexts should be asking, does Zillow stimulate gentrification – abrupt changes in the character, culture, and demographics of a neighbourhood – and if so, how can its impact be countered? In many communities across the United States, gentrification does not simply mean an influx of affluent residents; it means that low-income residents who cannot afford higher rents or property taxes are forced to leave. Calling attention to the role of Big Data in the housing market can be a starting point for developing new forms of advocacy to support the rights of those for whom existing data do not do justice.

Despite their differences, the DPLA, NewsScape, and Zillow are unified by a consistent motivation: to assemble collections that are complete. Their creators aspire to build comprehensive perspectives on the world, for example offering vistas across all the books, all the news, and all the real estate opportunities. We might think of Big Data as an instance of what Nye calls the 'technological sublime', for they test the limits of human perception and imagination (1994). But as the cases above reveal, Big Data only appear to be comprehensive from a distance. They have important limits. Seeing Big Data as local can help us understand their limits and challenge the processes of agglomeration that shape the phenomenon of Big Data as it exists today.

Cultivating a local sensibility

Our conceptions about data have not kept pace with changes in the size of data sets, as well as their diversity, distribution, and use. In a 1954 article in Atlantic Magazine entitled 'As We May Think' – written before the adoption of the digital computer – Vannevar Bush imagined how researchers would handle large datasets (Bush, 1945). Conceived of in response to the rapid increase in scientific publication during and immediately following the Second World War, his solution was the memex, short for 'memory expansion'. Today, we might think of the memex as an early attempt to leverage Big Data, though the term itself would not come into common use for another half century or more. Contemporary digitalized collections act as Bush imagined, but only in a superficial sense. They help scientists, but also educators, professionals, and an increasingly broad public, manage streams of data that would otherwise overwhelm an individual. However, Bush did not predict some of the most important social changes in practices with data: changes that would make data simultaneously smaller and bigger than he could have imagined.

First, knowledge practices have rapidly diverged, leading to a variety of data cultures, each of which manages data in its own way. Second, data are widely distributed. The notion of a personal database has been replaced by a fascination with the potential of the web as a platform for access to data from almost anywhere. Third, data have become big business. As mentioned in the example of Zillow, the potential surplus value of data has stimulated agglomeration and, thereby, severed important ties between data and the local communities in which they are made.

The three cases presented in this paper, the DPLA, NewsScape, and Zillow, offer good indications of the range of challenges that accompany the proliferation of Big Data. In order to develop a robust critique of Big Data, I argue that we must see the phenomenon as existing alongside local data. Indeed, the bigger the dataset, the more important it is to acknowledge local sources of heterogeneity. It has been more than 70 years since Vannevar Bush's article in *The Atlantic*. We need to reconsider our expectations of data. Below, I offer a set of speculative guidelines for approaching Big Data as agglomerations of the local:

(1) *Get dirty*. Ask what might be lost by filtering out local features in data? Data that initially appear out of place might offer insights into past audiences or lost provenances.
(2) *Be comparative*. Ask how are data created otherwise? The local dimensions of data are most evident in the context of Big Data, when different datasets are brought into productive dialog.
(3) *Attend to context*. Ask what are the environments in which data evolve? Illuminate the infrastructure of data collection and interpretation.
(4) *Stay connected*. Ask who lives with data and understands their importance? Data should facilitate communication among their authors and users, not serve as proxies for local knowledge.

If we revisit the DPLA with these principles in mind, we discover many opportunities for deeper inquiry. Each guideline invites the inquirer to look more closely at the DPLA's agglomerated data, but also to see them in relation to the whole. The first invites one to consider the DPLA catalogue not as a single dataset, but rather as a varied history of data collection practices. The second suggests that the data from one contributor to the DPLA, such as the NYPL, might be productively understood by seeing them in relation to those of another. The third opens the inquiry to the inclusion of other representations of each collection; for instance, the architecture of contributing libraries might aid in making sense of variations among datasets. The fourth reminds the inquirer that the data are an excellent resource in direct interactions with library staff at contributing intuitions. Librarians are themselves living repositories of institutional knowledge. Examining the data of the DPLA through one or more of these guidelines can be a powerful exercise in learning to look critically at Big Data.

It is also important to note that working with Big Data can introduce a variety of ethical concerns (boyd & Crawford, 2012). From a local perspective, studying Big Data can reinforce the social importance of the large, well-supported organizations that create them. Hence, there is also a need to acknowledge forms of counter-data or even anti-data: tactical representations that challenge the dominant uses of data to establish cultural hegemony, to reinforce state power or to increase profit. For instance, counter-data might encode alternative perspectives to the cultural histories of the DPLA, the current events of NewsScape, or the developments forecast by Zillow. Although counter-data have their own limits (Dalton & Thatcher, 2014), they are necessary to counteract dominant representations of the past, present and future. With this concern in mind, I suggest one additional guideline for cultivating a local sensibility toward Big Data:

(5) *Collect counter-data.* Convey the incompleteness of data. Moreover, invite and integrate annotation, challenges to interpretation, and alternate representations.

Thinking again about the case of the DPLA, this final guideline encourages us to see library patrons as an equally important resource, with their own means of making do in libraries using their institution's data or even some of their own. These guidelines should support not only local reading, but also other practices with data. For example, critical approaches to visualization might highlight the local features in data, rather than the global ones.

Conclusion

Developing critical perspectives on Big Data means seeing the whole as well as its local, heterogeneous parts. This paper proposes that we engage Big Data with increased attention to the local, in order to acknowledge the importance of Big Data's distributed origins. Local readings of Big Data – informed by interviews with those who make and make use of them – can reveal the variety of local ties they harbour: in schemata, errors, infrastructures, classifications, absences, and rituals. Seeing Big Data as assembled from local conditions opens new opportunities and obligations for scholarship, pedagogy, and advocacy. However, we should not romanticize local ties; they can be lacking in sophistication or even discriminatory. The paper concludes with five speculative guidelines for adopting a local sensibility when working with (or against) Big Data: get dirty, be comparative, attend to context, stay connected, and finally, collect counter-data.

On a tangential note, I have not dealt with a different, but salient, local dimension of Big Data: where and how they are experienced. Indeed, Big Data can be enacted through a variety of local practices – readings and visualizations, mentioned here, as well as other manifestations with social and material characteristics that are not fully determined by the origins of data. Thus, a local sensibility should also be attuned to the specificities of how Big Data are encountered.

In summary, there is an emergent need for alternatives to the universalizing discourses that surround Big Data, infusing the phenomenon with an 'aura of truth, objectivity and accuracy' (boyd & Crawford, 2012, p. 663). Scholars, students, and advocates who wish to critique Big Data initiatives should learn to take them apart. When it comes to Big Data, thinking locally is thinking critically.

Notes

1. These engagements typically began with the question, 'What are data?' Other topics of discussion included data standards and workflows, as well as significant challenges, changes, or absences in those conditions. I also asked, 'Do people outside your field appreciate the complexity of the data you work with? What would you have them know?' Finally, I encouraged subjects to discuss who they share data with and what requirements those exchanges place on data.
2. From an interview with the author, 2013.
3. From a public lecture by Mcquirter at Beautiful Data II. Cambridge, 2015.
4. From an interview by the author, 2013.
5. From an interview by the author, 2013.
6. http://newsscape.library.ucla.edu

Acknowledgements

This paper has benefitted from extensive feedback generously offered by the editors and reviewers of the journal as well as numerous colleagues, including Matthew Battles, Daniel Cardoso, Johanna Drucker, Carl DiSalvo, Lauren Klein, Leeanne Richardson and Greg Zinman.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributor

Yanni Alexander Loukissas is an Assistant Professor of Digital Media in the School of Literature, Media, and Communication at Georgia Tech. His research draws together the fields of information studies, design, and science, technology, and society. Recent projects include an institutional portrait of the Arnold Arboretum using data on 70,000 trees, vines, and shrubs and a visualization of human–machine interactions during the first lunar landing. He is also the author of *Co-designers: Cultures of computer simulation in architecture* (Routledge, 2012). Before coming to Georgia Tech, he was a lecturer at the Harvard Graduate School of Design, where he co-coordinated the Program in Art, Design and the Public Domain and served as a principal at metaLAB, a research project of the Harvard Berkman Center for Internet and Society. Originally trained as an architect at Cornell University, he subsequently received a Master of Science and a Ph.D. in Design and Computation at MIT. He also completed postdoctoral work at the MIT Program in Science, Technology and Society. [email: yanni.loukissas@lmc.gatech.edu]

References

Anderson, C. (2008, June 23). The end of theory: The data deluge makes the scientific method obsolete. *Wired*. Retrieved from http://www.wired.com

Battles, M., & Loukissas, Y. (2013). Data artifacts: Visualizing orders of knowledge in mega-meta collections. In Aida Slavic, Almila Akdag Salah, & Sylvie Davies (Eds.), *Classification and visualization: Interfaces to knowledge: Proceedings of the International UDC Seminar, 24–25 October 2013, The Hague, The Netherlands* (pp. 243–258). Würzburg: Ergon Verlag.

Bowker, G. C., & Star, S. L. (1999). *Sorting things out: Classification and its consequences*. Cambridge: MIT Press.

boyd, d., & Crawford, K. (2012). Critical questions for Big Data. *Information, Communication & Society*, *15*(5), 662–679.

Bush, V. (1945, July). As we may think. *The Atlantic*. Retrieved from http://www.theatlantic.com/

Culler, J. (2010). The closeness of close reading. *ADE Bulletin*, *149*, 20–25.

Dalton, C., & Thatcher, J. (2014). What does a critical data studies look like, and why do we care? Seven points for a critical approach to 'Big Data'. *Society and Space – Environment and Planning D*. Retrieved from https://societyandspace.com

Digital Public Library of America. (2016). (http://dp.la)

Douglas, M. (1978). *Purity and danger: An analysis of the concepts of pollution and taboo*. New York, NY: Routledge & Kegan Paul PLC.

Drucker, J. (2014). *Graphesis: Visual forms of knowledge production*. Cambridge: Harvard University Press.

Edwards, P. N. (2010). *A vast machine: Computer models, climate data, and the politics of global warming*. Cambridge: MIT Press.

Fry, B. (2007). *Visualizing data: Exploring and explaining data with the processing environment*. Sebastopol: O'Reilly Media.

Gantz, J., & Reinsel, D. (2011). *Extracting value from chaos*. Retrieved from http://idcdocserv.com/1142

Geertz, C. (1992). Local knowledge and its limits. *The Yale Journal of Criticism*, *5*(2), 129–135.

Gieryn, T. F. (2000). A space for place in sociology. *Annual Review of Sociology*, *26*(1), 463–496.

Gitelman, L. (2013). *'Raw data' is an oxymoron*. Cambridge: MIT Press.

Gusterson, H. (1998). *Nuclear rites: A weapons laboratory at the end of the cold war*. Berkeley: University of California Press.

Haraway, D. (1988). Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies*, *14*(3), 575–599.

Jänicke, S., Franzini, G., Faisal, C., & Scheuermann, G. (2015). *On close and distant reading in digital humanities: A survey and future challenges. A state-of-the-art (STAR) report*. EuroVis 2015: The EG/VGTC Conference on Visualization, Cagliari.

Kitchin, R., & Lauriault, T. P. (2014). *Towards critical data studies: Charting and unpacking data assemblages and their work*. The Programmable City Working Paper 2. Programmable City, Social Science Research Network.

Kitchin, R., & McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*, *3*(1). Retrieved from http://bds.sagepub.com

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. Cambridge: Harvard University Press.

Latour, B. (1990). Drawing things together. In Michael Lynch and Steve Woolgar (Eds.), *Representation in scientific practice* (pp. 19–68). Cambridge: MIT Press.

Latour, B., & Woolgar, S. (1979). *Laboratory life: The social construction of scientific facts*. Beverly Hills: Sage.

Lohr, S. (2012, February 11). The age of Big Data. *New York Times*.

Mayer-Schoenberger, V., & Cukier, K. N. (2013 May/June). The rise of Big Data. *Foreign Affairs*. Retrieved from https://www.foreignaffairs.com

metaLAB (at) Harvard. (2016). *The book biography machine*. Retrieved from http://metalab.harvard.edu

Nye, D. E. (1994). *American technological sublime*. Cambridge: MIT Press.

Schneiderman, B. (2014). The big picture for Big Data: Visualization. *Science*, *14*(343), 730.

Star, S. L., & Griesemer, J. R. (1989). Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in Berkeley's museum of vertebrate zoology, 1907–39. *Social Studies of Science*, *19*(3), 387–420.

Vertesi, J., & Dourish, P. (2011). *The value of data: Considering the context of production in data economies*. Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work, 533–542.

Zimmerman, A. S. (2008). New knowledge from old data the role of standards in the sharing and reuse of ecological data. *Science, Technology & Human Values*, *33*(5), 631–652.