LOCAL ENDS

BEYOND DATA SETS

Understanding data requires more than access to a spreadsheet. All data are entangled with places, institutions, processes, and people that fundamentally shape their significance and use. If we haven't understood the data's setting, we haven't understood the data. Over the course of four cases and six chapters, I have sought to impart this deceptively simple message along with its implications for scholarship and practice.

The case of the Arnold Arboretum illustrates the complex local attachments that data hold, suggesting that place should be an important consideration in data presentation. The DPLA reveals the challenges that can arise when data from different settings are brought into dialogue. Understanding data infrastructures means taking a comparative approach. Looking at the news as a source demonstrates that data can scarcely be separated from the algorithms used to process them. Instead of trying to differentiate the substance of data from their activation through algorithms, we should learn to see the two as part of a data system. Unpacking data requires intimate knowledge of the algorithms that shape them, and vice versa. Finally, Zillow illuminates the broader implications of interfaces: those visual, discursive, and procedural settings that shape our use of data.

Building on these lessons, the previous chapter introduces a variety of localized models intended to put the book's principles into practice through existing and accessible design examples. Such examples model the means of working with data locally, but not necessarily the broader goals in doing so.

In seeking to understand data locally, what kinds of outcomes might we hope for? Thus far, I have only hinted at the answers. Now I would like to make those aims more explicit. What does it mean to successfully put data to use in the service of local ends? As I mentioned at the outset of the book, data seem useful in the first instance because they hold the promise of insight at a distance. Yet being mindful of the local contingencies of data—both where data are made and where they are used—can reveal other benefits. Below I reflect on four commonplace goals for data: orientation, access, analysis, and optimization. Then I consider how such ambitions might be reconsidered to account for additional local ends: place making, restraint, reflexivity, and contestation.

Orientation and Place Making

Data can be important tools for orientation in complex environments. At the Arnold Arboretum, where more than seventy thousand trees, vines, and shrubs have lived since its establishment in 1872, visitors cannot hope to know the extent of the place through experience alone. Data-embossed tags on each plant specimen turn the arboretum

into an inhabitable catalog for species that are not otherwise known in the surrounding region. But these same data have another role. Seeing which specifications tags hold, how they are organized, and even which plants are not tagged at all tells visitors much about the institution itself: a place in which data are deeply enmeshed in its history, materiality, and culture. Data are not just representative of the collected specimens of the arboretum; they are an integral part of the way the place works.

As another pressing example, consider the increasingly widespread notion of the smart city: a place in which all the elements of civic life are potentially mediated through data. Michael Totty reports in the *Wall Street Journal* that, "Whether it's making it easier for residents to find parking places, or guiding health inspectors to high-risk restaurants or giving smoke alarms to the households that are most likely to suffer fatal fires, big-data technologies are beginning to transform the way cities work."[1]

The smart city is heralded as a potentially seamless experience in which place and data converge to meet the needs of city residents as well as reduce administrative costs through preventive maintenance. But early efforts to develop smart city infrastructures are only beginning to grapple with the question of how the phenomenon might materialize in different local ways. We must put aside the rhetoric of digital universalism to ask, How will divergent practices of data collection and use come to characterize different manifestations of the smart city? Some places might encourage grassroots organizing and activism around data, as in the case of the Anti-Eviction Mapping Project introduced in chapter 6. Others may follow more centralized models of surveillance and control. The smart city is not one place.

Access and Restraint

Data act as bridges to large collections of digital resources that are held remotely or in distributed locations. The DPLA, for instance, brings together resources from contributing collections across the United States. Data provide access, but not in a homogeneous way. A local perspective can help users of composite collections to better understand the limitations of knowledge gathered through data.

Another significant reminder of the limits of aggregate data was the 2016 US presidential election. As sociologist of science and technology Anne Pollock and I explain in a recent essay, waves of polling data were brought together from every state in the country.[2] The flood of data seemed to indicate, overwhelmingly, that Hillary Clinton would triumph over Donald Trump. Instead, voting on November 8 produced the opposite outcome. This event fostered increased skepticism over the explanatory power of big data. And yet in the immediate wake of the election, we observed widespread efforts to redeem data. Pollsters and journalists, the producers and disseminators of election data, have responded to the apparent limitations of their polls by reworking their approaches to data aggregation and restraining their projections as opposed to rejecting polling's relevance altogether.

Analysis and Reflexivity

Data are useful not only to access large collections of media but to analyze them as well. Archives like NewsScape, described in chapter 4, can be used in conjunction with algorithms for NLP to help analysts "extract" information from the news and explore it for salient patterns. But those algorithms are not simply neutral tools to be applied to any source. They are historically and materially local, because of the data on which they were tested and even trained. Acknowledging the local conditions of such tools can help us work more reflexively, with a concrete understanding of the processes and even people—like those invisibly at work behind algorithms—that make analysis possible.

Another domain where reflexivity would benefit the application of NLP is "search." Once an experimental research project in artificial intelligence, search is now a commonplace form of automation that we use every day on Google's home page. Journalists are on the forefront of this issue, and calling public attention to how search can work to "replicate and deepen the biases" in society.[3] Carole Cadwalladr of the *Guardian*, for instance, has investigated Google's autocomplete function, which attempts to finish your search query for you, using data on common searches related to yours—and even preemptively displaying the results. She reports typing "a-r-e" and "j-e-w-s," and seeing Google complete her entry as "white," "Christian," or "evil." Furthermore, she found that typing "d-o" and "b-l-a-c-k-s" can autocomplete in Google as "commit more crimes?" When questioned by Cadwalladr about these obviously racist results, Google responded by deferring responsibility:

> Our search results are a reflection of the content across the web. This means that sometimes unpleasant portrayals of sensitive subject matter online can affect what search results appear for a given query. These results don't reflect Google's own opinions or beliefs—as a company, we strongly value a diversity of perspectives, ideas and cultures.[4]

In seeking to "organize the world's information," though, Google is anything but neutral.[5] In a recent book, information studies scholar Safiya Noble calls the company to task for creating "algorithms of oppression" that reinforce racism and sexism, particularly in search results for black girls and women.

> We have to ask what is lost, who is harmed, and what should be forgotten with the embrace of artificial intelligence in decision making. It is of no collective social benefit to organize information resources on the web through processes that solidify inequality and marginalization.[6]

Moreover, the purveyors of fake news and other forms of clickbait are explicitly working to game existing algorithms by discovering the "tricks that will move them up Google's PageRank system."[7] Learning how algorithms work in coordination with

existing data can help us reflexivity attend to and take responsibility for (rather than grudgingly accept) the problems inherent in pervasive data-driven services.[8]

Optimization and Contestation

Finally, data can enable the optimization of systems that are otherwise too large or complex to configure by hand. Market-based platforms like Zillow promise to bring increased transparency and thus efficiency to the housing market by allowing buyers, sellers, and realtors to not only access real estate data but add or update data about individual homes too. Optimization, however, is only possible when the context is fixed and outcomes are agreed on. In housing, as in many areas of public life, context is always contested. Understanding how interfaces like Zillow's work to establish an operational context for data can prompt us to question the status quo implicit in all optimization efforts.[9]

Another example of the use of optimization that might be similarly contested is the "gig" economy, where employers in fast-food services, among others, use data-driven management practices to make the most efficient use of their workforce. Utilizing such techniques, businesses seek to avoid overstaffing. But the resulting work conditions for "on-demand" employees—unpredictable hours, few or no benefits, and little job security—only make sense if you understand them within the context of profit maximization.[10] How might the data currently used to optimize gig work be harnessed instead to contest unsustainable labor conditions and perhaps enhance collective bargaining between employers and labor unions?

Each of these local ends—place making, restraint, reflexivity, and contestation—are efforts to rethink the now-commonplace roles for data as universal tools for orientation, access, analysis, and optimization. Local ends require a new ethics of data work tied to local viewpoints. Using data effectively cannot simply be about securing personal benefits. Understanding data is increasingly a social and civic project.

GUIDING FUTURE RESEARCH

Accepting the message of this book that all data are local leaves us with a significant challenge: How can we make data broadly accessible while acknowledging their attachments? A general ethos of "open data" has pervaded government, academic, and industry approaches in recent years.[11] But what good are open data if they cannot be understood by outside audiences? Data made in civic and cultural institutions, exemplified by the cases in this book, are not often made in the first instance for public use. Rather, those data are part of systems for curation and management originally defined in the nineteenth century; they are expert tools, designed for use within complex knowledge systems. In order to embrace the locality of data while also making their home institutions relevant today, we must reconsider what we mean by *open data*.

One way of doing so, and the departing recommendation of this book, is to establish a new genre of open data guides, which can introduce potential users to data settings and not just data sets. What might local guides for open data consist of, who will make them, and how will they affect data use? This is a pressing research question for all of us.[12]

We might think of a local guide for open data as related to, yet more reflexive than, a traditional data user guide or manual. Such materials sometimes accompany data sets created by government institutions or organizations that hold themselves to high expectations for transparency.[13] A user guide is typically created by those who make data as a means of helping others put those data to good use.[14] It might offer a codebook, which explains each of the fields and values in the data, as well as information about their purposes, processes, and potential contexts of use. For example, the Western Pennsylvania Regional Data Center has produced a sample template for creating a data user guide with eight distinct sections: original purpose and application; history, standards, and formats; organizational context; workflow; things to know about the data (including limitations); current applications; field values; and sources and acknowledgments.[15]

This template is a productive starting point for data-creating institutions that want to make their data more accessible. But such user guides can fall short in a number of ways. First, user guides frequently do not address and might even intentionally obscure the answers to difficult ethical questions around data in order to protect their collecting institutions. Second, user guides are often fixed, while the data and regulations around their use change over time. Third, understanding data must be immersive. It requires local reading in dialogue with data experts and subjects as well as hands-on work such as data visualization. Fourth, different users have an additional burden of trying to understand their own standpoints and those of their particular audiences.

As I have explored the data featured in this book, I have also developed an awareness of my own specific subject position: a technically adept user with the time, resources, and status necessary to access complex as well as sometimes-shrouded, if ostensibly open, data settings. I do not presume to know what other, less privileged users might need to facilitate access for their own ends. For all the reasons above, I imagine that local guides for open data would need to diverge from existing templates in a number of important ways.

Beyond establishing the contents of such guides, we must consider their social context. Making local guides means moving beyond the user data dyad; there is more than one subject position for a data user. A number of users and even nonusers might need to be accounted for. Moreover, local guides need not be made by insiders: those who make the data. Outsiders, especially students, can learn from the practice of creating such local guides as a means of coming to terms with what, beyond a spreadsheet,

comma-separated values file, database, or application programming interface, is necessary for understanding data.

One of the additional challenges in developing local guides for open data will be establishing a local ethics of data use. When it comes to working with data, our ends are only just if they are arrived at justly. A number of authors have written about the ethics of data involving issues of persuasion, privacy, security, and even exploitation.[16] While these issues are of utmost significance, I would ask, What unanswered ethical questions do the cases in this book suggest? For example, I expect local ethics to learn from the ethics of care. Given that data are never raw, Geoffrey Bowker tells us, they need to be cooked "with care."[17] Unfortunately, there is no formula for doing so.

An ethics of care does not presuppose a set of universal rules for treating data and associated humans (and nonhumans) ethically but rather implies maintaining relationships with them. Maria Puig de la Bellacasa writes that "we need to ask, 'how to care' in each situation."[18] Furthermore, we need to understand who is empowered by acts of caring. As the Grassroots Mapping project introduced in chapter 6 illuminates, centralized institutions, such as the World Bank, often collect data in ways that reproduce colonialist patterns of external representation and control, veiled behind the discourse of care.

A strict ethics template might lead us to another form of digital universalism or one-size-fits-all solution. We are beginning to see this in domains like the design of so-called smart cities where ethics are an obvious requirement.[19] Consider, as mentioned previously, that few city officials are thinking about what *smart* should mean in their particular locality. How will smartness address their city's local social, economic, and geographic advantages as well as challenges in ethical ways? Chapter 5 illustrates the problems of simply making civic data open—one of the defining steps toward creating the generic smart city. Publicly accessible data, as I explain in the case of Zillow, can be placed in visual, discursive, and algorithmic contexts that undermine their original role as resources for the public good. What are the local effects of Zillow on neighborhoods in Atlanta, where rampant economic speculation is displacing historically black communities at a rapid pace? What ethical obligations does Zillow have to those communities, and can those obligations be simply generalized for the entire country? Using data ethically is a local problem, which requires attention to the differences among data settings and how they might change over time, necessitating continued maintenance and adjustment. Thus a local ethic also implies evolving roles for those who act as intermediaries and stewards for data. Understanding those identities and their future potential is a necessary component of such research too.

Below I propose first steps toward creating local guides to open data in five parts, which build on the findings in this book. This sequence might be particularly useful for students, scholars, and practitioners who frequently encounter data sets that are new to them. That said, it is a provocation to further research rather than another template to be followed exactly. For we do not yet know what kind of guidance will work, where, and for whom.

Step 1: Read

Acquire a human-readable version of the data set. Try to identify at least two kinds of data: one that seems typical, and one that is surprising or confusing. Consider the data format(s). What can we learn about their history? Where else are they used? What other formats are used for similar data? Question why the data looks the way that it does.

Step 2: Inquire

Establish a rapport with a diverse group of informants who are local experts on the data set: data collectors, data analysts, or data subjects (someone who the data represents). Ask them about the provenance and purpose(s) of the data. Ask them about known patterns. Also ask them to identify problems: errors, absences, and limitations that can illuminate the context(s) of collection. With their help, create an accessible codebook, explaining the various data fields and their structure, along with your own code of ethics for working with the data set.

Step 3: Represent

Use a simple visualization technique, such as a scatterplot, line graph, timeline, map, tree map, or network diagram, to confirm or contest a known pattern in the data, first revealed by the informants. Return to the informants with any new questions prompted by the visualization.

Step 4: Unfold

Work with the informants to create a diagram of the collection, normalization, maintenance, and distribution processes used with the data set. Learn about any specialized algorithms used on or developed along with the data.

Step 5: Contextualize

Identify and analyze several contexts of use for the data. Who is using these data and what claims are they making? What visual, discursive, or algorithmic tools are they employing? What ethical issues do these uses present? What friction do they "kill" or potentially kindle?

Learning to venture into local knowledge systems has long been the role of an ethnographer: one who documents the practices of different cultures. But today we encounter new and confounding knowledge systems in each new data setting. While not all of us will become ethnographers in the traditional sense, we can learn to take an "ethnographic stance" when encountering unfamiliar sources by making a commitment to contextualization.[20] Where are data made and by whom? What assumptions and values do data carry? How are data entangled with otherwise-invisible processes? What

are the contexts in which we confront data? Such questions can help us establish new expectations of data as well as our relationships with both their keepers and subjects.

We increasingly live in societies driven by data and the accompanying promise of connectivity. But as this book has sought to demonstrate, connections based solely on data are relatively weak. If I can offer one takeaway message to the reader, it is this: treat data as a point of contact, a landing, an opportunity to get closer, to learn to care about a subject, or the people and places beyond data. Do not mistake the availability of data as permission to remain at a distance.