INTRODUCTION

FROM DATA SETS TO DATA SETTINGS

While reading through the accessions data of Harvard University's Arnold Arboretum, one of the largest and most well-documented living collections of trees, vines, and shrubs in the world, I came across the record for a cherry tree, *Prunus sargentii*, named for its collector, the botanist Charles Sprague Sargent. The data suggest that this specimen was retrieved by Sargent on an expedition to Japan in 1940. Yet Sargent died thirteen years earlier, in 1927. How might we decipher the convoluted origins of this tree: uprooted from Japan and planted in US soil on a timeline that makes little sense to an outsider?

In the collections data of the New York Public Library, I discovered 1,719 different conventions for writing the date (i.e., _ _-_ _-_ _ _ _ is just one). Some of these date formats are strange, some are approximate, and some are in languages other than English. Taken together, they reveal the unexpected diversity of cataloging practices that one institution can contain. Recently, the institution contributed its data to a broad initiative called the Digital Public Library of America (DPLA). Containing data from libraries, museums, and archives across the country, this "mega meta collection" manages a confounding number of conflicting formats.[1] How can we expect to make sense of such heterogeneous sources and draw connections among them?

Querying NewsScape, a real-time television news archive hosted by the University of California at Los Angeles, offers access to more than three hundred thousand broadcasts dating back to the Watergate era—so much data that it cannot be seen independent of the algorithms used to search it. How should we differentiate the substance of news data from the computational procedures, such as natural language processing, necessary to access and analyze them?

The website Zillow, an interface to real estate data, purportedly, on all the homes in the United States, seems to facilitate a new level of transparency for the housing market. I can use the site to track the fluctuating market value of my own house or any one of the more than one hundred million properties listed, most of which are not even for sale or rent. But from within the consumer-centered context that Zillow has created, the effects of the inflated housing market on low-income communities across the country remain invisible. How are we to learn about the hidden impacts of our own uses of data?

These four examples introduce a number of challenges that can arise when trying to make sense of unfamiliar data: contradictions, conflicts, and opacities as well as the unintentional effects of both data collection and use. Yet they reinforce a single point, expressed in the title of this book: all data are local. Indeed, data are cultural artifacts

created by people, and their dutiful machines, at a time, in a place, and with the instruments at hand for audiences that are conditioned to receive them.[2]

When I made this observation in 2013 to a roomful of colleagues at the University of California at Berkley in the course of an irreverently titled symposium, The Data Made Me Do It, my words were met with a level of incredulity.[3] But in the ensuing years, we have all become more suspicious of the apparent biases and skewing effects in data. Even major news outlets have begun to report on the dark side of the data revolution, including accusations that Google has inadvertently trained its search algorithms on racist data, that strict measures of scholastic achievement can compel schools to "teach to the tests" or even attempt to cheat them, and that manufactured evidence in so-called fake news might have greatly influenced the 2016 US presidential election.[4] Even academics in the social sciences, who are expected to treat sources with more nuance, are embroiled in debates about how their own data might be unethically skewed by p-hacking: a technique by which researchers artfully manipulate the variables and scope of their analyses in order to produce results that might be considered statistically significant.[5]

A broad range of data-driven professions, which have become accustomed to using evidence collected in distant places and times, are publicly raising questions about how to best handle their most valued sources of information. It is not sufficient to identify and eliminate the most evident biases in data. We must learn to work differently, to uncover the inherent limitations in all such sources, before they lead us further astray.

Today, it is too easy to acquire data sets online without knowing much about their locality—where are they produced and used elsewhere—and how that may matter. We have come to rely on the availability of data as generic resources for reasoning not only in scholarship but in education, politics, industry, and even our personal lives. It is now commonplace for researchers, government institutions, and businesses alike to make their data available online, although often without enough accompanying guidance on how to put those data to good use. The problem starts with our language: the widely used term *data set* implies something discrete, complete, and readily portable. But this is not the case. I contend that we must rethink our terms and habits around public data by learning to analyze data *settings* rather than data *sets*.

This book is an exploration of nuances in data practice long debated in scientific laboratories, libraries, newsrooms, and activist communities, but more recently set aside in the contemporary rush to capitalize on the increasing availability of data.[6] I have found that experienced scientists, librarians, journalists, and activists implicitly know that looking for the local conditions in data can help them to work more effectively, and counter biases when necessary. We rarely need to discard data simply because they are strange. After all, data are useful precisely because they provide unfamiliar perspectives, from other times, places, or standpoints that we would not be able to access otherwise. The strangeness of data is its strength.

One ready example of how to use data locally is already at your fingertips. It is the way you might use this book's index. An index, like a data set, is a collection of related yet discrete expressions (the key terms in the book) gathered into a condensed, accessible reference. If the reader were to flip to the index now, they would find that it is most useful in conjunction with the corresponding source (this text) to which each independent entry refers. On its own, an index serves as little more than a teasingly abstract trace of what could be learned by reading the entire book. Nevertheless, the index is useful, provided the book is also at hand. Too often we attempt to use a given data set as a complete work, such as a book, rather than an index to something greater.

Instead of treating data as independent sources, we should be asking, *Where* do data direct us, and *who* might help us understand their origins as well as their sites of potential impact? The implications of these questions are threefold. For practitioners who want to work with data, understanding local conditions can dispel the dangerous illusion that any data offer what science and technology studies scholar Donna Haraway calls "the view from nowhere."[7] For students and scholars, attention to the local offers an opportunity to compare diverse cultures through the data that they make or use. Finally, local perspectives on data can awaken new forms of social advocacy. For wherever data are used, local communities of producers, users, and even nonusers are affected.

COLLECTIONS AS CASES

This book demonstrates how to understand data settings, not simply data sets, by taking the reader through six principles over an equal number of chapters. Chapter 1 takes on the first principle and the title claim that *all data are local*. The next four principles are illustrated by the concrete cases first introduced at the start of this chapter. They exemplify areas of utmost importance for creating an informed public: science communication, cultural history, journalism, and the housing market.

- The accessions data of the Arnold Arboretum can help us understand, first and foremost, that *data have complex attachments to place*, which invisibly structure their form and interpretation.
- The DPLA can help us see that *data are collected from heterogeneous sources*, each with their own local attachments.
- NewsScape offers an opportunity to learn how *data and algorithms are entangled*, with far-reaching implications for what it may mean to be informed in the future.
- Finally, the case of Zillow shows how *interfaces recontextualize data*, with striking consequences for the value that we place on our homes and those of others.

These cases reflect the challenges of working with publicly available data—challenges that are often overlooked in the abundant and pressing conversations on personal

data and privacy.[8] The first two cases explain the local contingencies of data, and how discontinuities among data can lead to conflicts. The next two look at the implications of data's locality for how we might understand higher-level computational structures: first algorithms, and then interfaces.

I use the term *local*, further explained in the next chapter, as a relative designation. Over the course of this book, each case offers an opportunity to incrementally explore and elaborate on what local can mean in relationship to data: from a form of place attachment, exemplified by the accessions data of the Arnold Arboretum, to the traces of such attachments found in the accumulated sources of data infrastructures, such as the DPLA or NewsScape. In the final case, on Zillow, the local is primarily identifiable in negative terms; local details are stripped away from data in order to create the "frictionless" interfaces desired by today's harried users. For their fickle audiences, companies in today's "interface economy" seek to make data accessible and actionable anywhere.[9] In doing so, they both obscure and then supplant the traditional meaning-making power of the local.

Toward its end, this book shifts from theoretical principles to strategies for practice. Chapter 6 leaves the reader with a culminating principle only hinted at above—*data are indexes to local knowledge*—and a set of practical guidelines that build on the preceding cases:

- Look at the data setting, not just the data set
- Make place a part of data presentation
- Take a comparative approach to data analysis
- Challenge normative algorithms using counterdata
- Create interfaces that cause friction
- Use data to build relationships

The book concludes with a question: How can we rework open data initiatives to make data settings versus data sets both accessible and actionable? Keeping this long-term ambition in mind, the reader might approach each case in the book by considering what it takes, beyond simply access, to make data usable effectively and ethically.

Let me now make a caveat: despite the provocatively broad claim on the cover of this book, I do not address *all* types of data. Most of the examples that I use throughout the text can be characterized as collections data. These are data that help people to manage distributed work with large quantities of objects, organisms, texts, images, and more. I focus on collections data for three reasons that I hope will make my argument more accessible to readers.

My choice, first of all, has to do with the concreteness of collections data. They refer to actual subjects in the world: plants, books, broadcasts, homes, and even people. Second, collections data are likely to be familiar to many readers. Social media have turned our lives into vividly documented collections of "friends," "favorites," and

"shares." Likewise, e-commerce sites like Amazon are collections. This is partially because, in recent years, standards for collections data have converged with object-oriented approaches to programming—a strategy for defining computational systems in terms of classes of objects and their attributes—in order to produce a powerful model for a broad array of online interactions.[10] Third, collections data have historically been used to do the work of curation (from the Latin *cura*, meaning "care")—a practice that for reasons I will get into later in this chapter, necessitates a local perspective. But when the data that describe large, complex collections are aggregated, without regard to their localities, we can be blinded to important distinctions within data. If unacknowledged, these distinctions can sometimes become structural fissures and even lead to a collapse.

Consider, as a stark example readily available in US public consciousness, the role of data in creating and, at first, obscuring the mortgage crisis of 2007. Several years before the market collapsed, in 2004, an eccentric financial manager named Mike Burry with a knack for identifying unique investment opportunities pored over reams of documents describing home loans that comprised a financial product known as a mortgage bond. At the time, private home mortgages were deemed the most stable kind of investment. Beyond ensuring the American dream of homeownership, the resulting mortgage-backed bond market served as the bedrock of the US economy.

As Burry slowly uncovered, the dream would become a nightmare for many homeowners. These bonds weren't based on uniform home loans with fixed terms. Rather, they were comprised of claims on returns from a heterogeneous reserve, including thousands of independent mortgages with varying risks. Many of them turned out to be "subprime": loans made at alarmingly high, variable interest rates and with a high risk of foreclosure. In order to tease apart the risks that each bond contained and understand the chance that the entire bond could fail, Burry had to work through a lengthy legal and financial prospectus. Back then, he might have been the only person to have done so, apart from the attorneys responsible for its assembly.

Michael Lewis recounts this tale in *The Big Short: Inside the Doomsday Machine.* Lewis's book, later adapted into the Oscar-winning film of the same name, tells of the creation and collapse of the mortgage bond market. At the time, all subprime mortgage bonds were considered equivalent, with their value set and secured by the unimpeachable ratings agencies, Moody's and Standard & Poor's.[11] Each mortgage bond represented innumerable pieces of loans that remained largely unexamined by Wall Street. Bonds based entirely on mortgages, explains Lewis, "extended Wall Street into a place it had never before been: the debts of ordinary Americans."[12]

Based on interviews with the few eccentric investors who saw it coming, Lewis's book introduces us to the backroom world of Wall Street where the housing crisis of 2007 began. "The people at Moody's and S&P," notes Lewis, "didn't actually evaluate the individual home loans, or so much as look at them. All they and their models saw,

and evaluated, were the general characteristics of loan pools."[13] Meanwhile, the banks presumed that they were passing off any potential liability by repackaging the risk. Also, they strongly suspected that even if the liability did catch up to them, the federal government would bail them out, which ultimately it did, but not before hundreds of thousands of people lost their homes to foreclosure.

By reading between the lines of the mortgage bonds, Burry discovered the contingent nature of each mortgage: its size, interest rate, payment structure, and inherent risk. Moreover, he learned that the number of interest-only, riskier mortgages contained within these bonds was increasing over time. This meant defaults were immanent. Burry leveraged this insight to bet against the housing market so as to "short" the mortgage bonds.

While others dealt blindly with the bonds as aggregates, Burry's research allowed him to see the housing crisis several years before it hit. Unfortunately, rather than using this knowledge to help those most imperiled by these practices, he chose to profit from their effects. *The Big Short* works as a cautionary tale about financial bubbles, but also as a lesson about the locality of data: data have heterogeneous sources, and there are severe implications for those who don't know how to read them with a discerning eye.

Mortgage data, by the way, are collections data too: records on individual entities used to identify and organize them as part of a larger composite. Nevertheless, the principles espoused in *All Data Are Local* can be quite broadly applied, beyond data that deal exclusively with collections. Other types of data, not addressed in this book, are also local and dependent on knowledge about their settings for responsible use. My own varied experiences with data have impressed this on me. In a study of human and machine interactions from the first lunar landing, I learned how Apollo 11 astronauts Neil Armstrong and Buzz Aldrin, aboard the lunar module (nicknamed "Eagle"), were distracted by unexpected and ultimately inconsequential *feedback data* from their guidance computer.[14] The astronauts could not decipher a series of outputs—the values "1201" and "1202"—on their display/keyboard interface. Recognizing them as alarm signals, the astronauts wasted critical seconds reaching out to ground control for help in deciphering these data. In the time that elapsed, the Eagle overshot its landing site and nearly crashed into the surface of the moon. In another study of human-machine communication, this time in a hospital operating room, I witnessed a surgeon, overly reliant on *sensor data* from an electrocardiogram, overlook a pool of blood slowly forming around his sneakers.[15] Another observer in the room warned the surgeon in time to save the patient from bleeding out. In both cases, astronauts and medical professionals were focused on data, and not the broader setting or context.

Time and again, I have encountered such signs of the insistent locality of data across data types. My discussion of collections data, however, is not meant to be comprehensive in scope. I have selected examples that illustrate the limits of universalizing ambitions for data and prompt us to think about how they might be used more

conscientiously. The reader might notice that my focus is predominantly on US data. In fact, the cases were chosen specifically because of their proximity and interrelationships. Together, they characterize a particular data-driven society. Although this is a significant limitation to my work, it also presents strategic opportunities. These cases can be used to challenge the unwarranted dominance on the internet of data created in the United States.[16] Seeing how data are local, I argue, can help us put data in their place, materially as well as politically.

LOCAL METHODS AND GOALS

All Data Are Local is assembled from a combination of qualitative findings on data cultures and exploratory data visualizations. Both are informed by extended ethnographic fieldwork, including interviews, workshops, and hands-on engagements with data, conducted over the course of seven years. My use of the term *ethnographic* echoes anthropologist Sherry Ortner's explanation of the method as an "attempt to understand another life world using the self—as much of it as possible—as the instrument of knowing."[17] Indeed, this is a book based largely on my own experiences as an observer and participant in data settings, guided by a desire to understand data through the perspectives and practices of both their keepers and subjects.

My approach is unconventional, but it builds on substantial research in data studies—an area of scholarship that has emerged recently in response to the increasing importance of data in everyday life. Data studies, which seeks to make sense of data from a social and humanistic perspective, has been a significant area of scholarship ever since information scholars Geoffrey Bowker and Susan Leigh Star published *Sorting Things Out: Classification and Its Consequences* almost two decades ago. Their book established the terminology and stakes for thinking about the social lives of data. But the worlds of data look strikingly different today, in 2019, than they did at the end of the twentieth century when *Sorting Things Out* was published.[18] We need new ways of thinking about and looking at the role of data in the public realm.

As explained above, my empirical focus is on four different collections of data. Each chapter documents my efforts to understand one of those collections within its spatial as well as social and technological contexts.[19] These cases might have been a means of reinforcing similar points by tracking one or more themes across many examples. Instead, I take each collection as an opportunity to open up new territory, to ask what each data setting can reveal that is distinct about the locality of data. Moreover, I try to engage these collections reflexively, considering my own position and relationship to the data and their subjects. Each collection is local for me, the investigator, in a different way.

In order to carry out this agenda, I employ a variety of methods for studying data, which I collectively refer to as *local readings*.[20] As the phrase implies, I treat data as texts: cultural expressions subject to interpretative examination. All my readings of

data rely on insights gleaned from their keepers, who use their own local knowledge to explain the contingencies of their data, which are not apparent otherwise. Moreover, local reading necessitates examining data comparatively. As cultural anthropologist Clifford Geertz explains, one local condition is most productively understood not in relation to some imagined universal but instead relative to another locality.[21] Sometimes my local readings are made possible by looking at different collections juxtaposed with one another. In other instances, these local readings involve looking at how data are made differently over time, but within the same institution. More experimentally, reading locally can mean imagining how data might be seen in new ways, using speculative yet nevertheless locally imagined modes of visualization.[22] From my perspective, visualization is simply another way of reading data. Each chapter in the book contains one or more visualizations that extend as well as enrich claims made in the text.[23]

My use of visualization is informed by a long history of design practices that produce informative and expressive experiences of data.[24] Today, most writing in the area of data visualization is pragmatic, offering techniques for hands-on work with data. Edward Tufte's book *The Visual Display of Quantitative Information* first introduced many contemporary scholars and practitioners to the potentials, pitfalls, and pleasures of looking at data graphically. Yet Tufte and more recent authors treat data as given.[25] It is time that we learned how to visualize critical thinking on the subject of data.

The visualizations in this book are meant to be exercises in first-person participation and inquiry within data cultures. As such, the visual results might at first appear odd or atypical to the reader. For example, some are entirely textual as opposed to graphical. These visualizations focus on showing the structure and texture of data, rather than offering clear visual patterns, telling stories, or answering narrowly defined questions, as more conventional instances of data visualization might do. In engaging these visualizations, the reader should be ready (as they must with any evidence) to do some of their own interpretative work. Visualizations are, after all, also texts.

One note about the critical sensibilities of this study: my methods are significantly informed by though distinct from those employed by the cohort of scholars who practice under the banner of *critical data studies*.[26] Geographers Rob Kitchin and Tracy Lauriault explain the purpose of this emergent area of investigation:

> To unpack the complex assemblages that produce, circulate, share/sell and utilize data in diverse ways; to chart the diverse work they do and their consequences for how the world is known, governed and lived-in; and to survey the wider landscape of data assemblages and how they interact to form intersecting data products, services and markets and shape policy and regulation.[27]

The work of critical data studies—to unpack, chart, and survey—is typical of critical approaches to scholarship. Across various areas of information studies, the term *critical* has been used to support projects that challenge the status quo: critical games,

critical literacy, critical making, and critical design.[28] As a mode of engagement that illuminates biases and assumptions we might otherwise unconsciously adopt, a critical approach is imperative for data studies and practice. But critical reflection has its own limits; it can be detached rather than responsible, analytic rather than affective, or conceptual rather than hands-on.

I take a critical stance, but also explore approaches to working with data that are less distant and cerebral than critical reflection implies. In order to do so, my approach integrates lessons from the feminist ethics of care. Unlike critical reflection, care embraces affect, material engagement, and a host of concerns sometimes invisible in conventional work with technology. Care is critical in that it calls attention to neglected things.[29] But it is more than critical reflection; it is a doing practice.[30] In pursuing opportunities not only for critical reflection on data but in support of care too, I hope to bring largely unrecognized and unrewarded local sensibilities into efforts to understand data.

AGAINST DIGITAL UNIVERSALISM

Finally, a note about the stakes of what I am proposing. The increasing availability of data in public life is part of a broader social and technological transformation: the rise of digital media. Since their early manifestations, digital media have been promoted as a means of independence from local constraints. "Being digital," prophesizes tech visionary Nicholas Negroponte in his 1995 text by the same name, means "less and less dependence upon being in a specific place at a specific time." Eventually, he posits, "the transmission of place itself will start to become possible."[31]

Twenty years later, in 2005, Negroponte launched his "One Laptop per Child" (OLPC) project with the aim of producing a rugged, cheap, and low-power computer, complete with its own open-source software, that might help poor, rural schoolchildren worldwide further their own education by "connecting to the world."[32] Much has been written about OLPC that I will not reprise here.[33] In fact, I would like to put aside questions about whether the project has been successful or not, and instead contemplate OLPC as representative of a broader ideology of place agnosticism for digital media. Negroponte's project, after all, is not designed to improve the places where its young intended users live but rather to make them less dependent on those places. Mike Ananny and Niall Winters, critics of the project, explain that "OLPC sees the child as the agent of change and the network as the mechanism of change."[34] We might ask, Why is it so important for digital media to be place agnostic? Who benefits from claims about the boundlessness of being digital?

The history of our digital media infrastructures is replete with important places. Consider the origins of the World Wide Web, the primary mechanism through which data are made publicly available on the internet today. It began as a local, "home brew" project.[35] The web's tripartite structure—an address system (URI), network protocol (HTTP), and markup language (HTML)—was first developed by Tim Berners-Lee and

his colleagues in a specific setting, CERN (the European Particle Physics Laboratory in Geneva, Switzerland), as a primarily textual scientific communication platform for distributed teams of physicists and engineers. Only years later was it reimagined as an infrastructure for the internet: a model for web page cataloging, delivery, and design that might be expanded indefinitely.[36]

The web's uniquely identifiable pages, connected by unidirectional links, might have solved the local problem of information management at CERN, but it makes for an ungainly way of structuring everything that we do on the web today. Moreover, websites have not become the self-contained places that Negroponte imagined might be "transmitted" anywhere. Rather, they are assemblies of data and algorithms that are composed, maintained, and eventually encountered in a variety of settings that matter to their use.[37]

Why, then, do so many creative people working in digital media today embrace what ethnographer Anita Chan calls "the myth of digital universalism"? This ideology, asserts Chan, leads us to falsely believe that despite our varying local circumstances, "once online, all users could be granted the same agencies on a single network, all differences could dissolve, and everyone could be treated alike."[38] Perhaps, operating from within their cloistered innovation centers, digital elites can become heedless to the contingencies of place. Indeed, it is difficult to predict how other conditions, in other places and times, governed by unfamiliar norms, might clash with their own judiciously designed systems. To use a more contemporary example, despite being mired in scandals over the spread of conspiracy theories and hate speech online, prominent leaders in Silicon Valley are still reluctant to acknowledge that they cannot control what users do with their platforms.[39]

One reason universalist aspirations for digital media have thrived is that they manifest the assumptions of an encompassing and rarely questioned free market ideology.[40] If you are not influenced by your setting, you are a more independent and economically rational individual. If your reach can extend indefinitely, your profits can always grow. If you can participate from anywhere, competition is at its strongest. When digital media normalize human behavior and diminish local effects, the market gains strength.

The market, however, should not be the sole means of evaluating digital media and, by extension, contemporary manifestations of data. The diversity and prosperity of the world's varied and contingent digital practices depend on our acceptance of data's locality. In fact, the stakes for the future of the internet could not be higher. If left unchallenged, digital universalism could become a new kind of colonialism in which practitioners at the "periphery" are made to conform to the expectations of a dominant technological culture.[41]

If digital universalism continues to gain traction, it may yet become a self-fulfilling prophesy by enforcing its own totalizing system of norms. Fortunately, there is still time to halt the march toward placelessness. As I argue in this book, learning to look at the local conditions of data can be a form of resistance to the ideology of digital universalism and threat of erasure that it poses to myriad data cultures.

Resisting digital universalism, and instead seeking out the ways in which data are local requires hard work. But it is work that can begin modestly: by learning about the data-shaping power of settings, such as the Arnold Arboretum; taking account of the inherent discontinuities in practices of data collection, such as those illuminated by the DPLA; acknowledging the limits of algorithms that only recognize normative patterns in the news, as NewsScape demonstrates; and confronting the devastating effects of the data-driven housing market on users and nonusers alike, as exemplified by Zillow. Overlooking the locality of data means being naive to the contradictions, conflicts, opacities, and unintentional impacts produced by efforts to universalize data.

All Data Are Local is meant to stand as an alternative to the disassociated theoretical treatises and practical manuals on contemporary practices with data. The book asks readers to consider how a local perspective can transform practices designed to make sense of data. Readers will learn how to engage with the local conditions of data productively in ways that lead toward just ends for those who make, use, or are themselves the subjects of data. Few of us today do not fall into one or more of these categories.

By the end of this book, I hope the reader will acquire new sensibilities about both data and the local. For I aim to do more than simply proclaim that data are local or muster the evidence necessary to convince readers; indeed, this claim may simply confirm what some data-savvy audiences intuitively know. Rather, this book tackles a pragmatic question: How do local conditions matter for understanding data in everyday practice? In the chapters that follow, I consider that question as it manifests in a range of specific settings.

1