Computational Social Sciences

Sorin Adam Matei
Nicolas Jullien
Sean P. Goggins *Editors*

Big Data Factories

Collaborative Approaches

Springer

Computational Social Sciences

Computational Social Sciences

A series of authored and edited monographs that utilize quantitative and computational methods to model, analyze and interpret large-scale social phenomena. Titles within the series contain methods and practices that test and develop theories of complex social processes through bottom-up modeling of social interactions. Of particular interest is the study of the co-evolution of modern communication technology and social behavior and norms, in connection with emerging issues such as trust, risk, security and privacy in novel socio-technical environments.

Computational Social Sciences is explicitly transdisciplinary: quantitative methods from fields such as dynamical systems, artificial intelligence, network theory, agent based modeling, and statistical mechanics are invoked and combined with state-of the-art mining and analysis of large data sets to help us understand social agents, their interactions on and offline, and the effect of these interactions at the macro level. Topics include, but are not limited to social networks and media, dynamics of opinions, cultures and conflicts, socio-technical co-evolution and social psychology. Computational Social Sciences will also publish monographs and selected edited contributions from specialized conferences and workshops specifically aimed at communicating new findings to a large transdisciplinary audience. A fundamental goal of the series is to provide a single forum within which commonalities and differences in the workings of this field may be discerned, hence leading to deeper insight and understanding.

Series Editors

Elisa Bertino
Purdue University, West Lafayette, IN, USA

Claudio Cioffi-Revilla
George Mason University, Fairfax, VA, USA

Jacob Foster
University of California, Los Angeles, CA, USA

Nigel Gilbert
University of Surrey, Guildford, UK

Jennifer Golbeck
University of Maryland, College Park, MD, USA

Bruno Goncalves
New York University, New York, NY, USA

James A. Kitts
Columbia University, Amherst, MA, USA

Larry Liebovitch
Queens College, City University of New York, Flushing, NY, USA

Sorin A. Matei
Purdue University, West Lafayette, IN, USA

Anton Nijholt
University of Twente, Enschede, The Netherlands

Andrzej Nowak
University of Warsaw, Warsaw, Poland

Robert Savit
University of Michigan, Ann Arbor, MI, USA

Flaminio Squazzoni
University of Brescia, Brescia, Italy

Alessandro Vinciarelli
University of Glasgow, Glasgow, Scotland, UK

More information about this series at http://www.springer.com/series/11784

Sorin Adam Matei • Nicolas Jullien
Sean P. Goggins
Editors

Big Data Factories

Collaborative Approaches

Springer

Editors
Sorin Adam Matei
Purdue University
West Lafayette
IN, USA

Nicolas Jullien
Technopôle Brest-Iroise
IMT Atlantique (Telecom Bretagne)
Brest Cedex 3, France

Sean P. Goggins
Computer Science
University of Missouri
Columbia, MO, USA

ISSN 2509-9574 ISSN 2509-9582 (electronic)
Computational Social Sciences
ISBN 978-3-319-59185-8 ISBN 978-3-319-59186-5 (eBook)
https://doi.org/10.1007/978-3-319-59186-5

Library of Congress Control Number: 2017958439

© Springer International Publishing AG 2017
Open Access Chapter 9 is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/). For further details see license information in the chapter.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Chapter 1
Introduction

Nicolas Jullien, Sorin Adam Matei, and Sean P. Goggins

Human interactions facilitated by social media, collaborative platforms, and the blogosphere generate an unprecedented volume of electronic trace data every day. These traces of human behavior online are a unique source for understanding contemporary life behaviors, beliefs, interactions, and knowledge flows. The social connections we make online, which reveal multiple types of human connection, are also recorded on a scale and to a level of granularity previously unimaginable, except possibly by science fiction writers. To many in the data analytics world, these traces are a gold mine. New sub-domains of inquiry have emerged as a consequence of this revolution: computational social science, big data, data science, open innovation data analytics, network science, and undoubtedly new ones yet to appear in the near future. Massive amounts of data, each counting millions of data records and behaviors, are now available to the academic, governmental, or industry research and teaching communities. They promise faster access to real-time social behavior and better understanding of how people behave and interact. Such "social" data include complete records of Wikipedia edits, interactions on social coding platforms like GitHub, and the expression of affiliations and engagement of participation on social media (Twitter, Facebook, YouTube, etc.).

Working with data of this kind and of this magnitude requires cleaning up and preprocessing prodigious amounts information, which is nontrivial and costly.

N. Jullien (✉)
Technopôle Brest-Iroise, IMT Atlantique (Telecom Bretagne), Brest Cedex 3, France
e-mail: nicolas.jullien@imt-atlantique.fr

S.A. Matei
Purdue University, West Lafayette, IN, USA
e-mail: smatei@purdue.edu

S.P. Goggins
Computer Science, University of Missouri, Columbia, MO, USA
e-mail: gogginss@missouri.edu

© Springer International Publishing AG 2017

S.A. Matei et al. (eds.), *Big Data Factories*, Computational Social Sciences,
https://doi.org/10.1007/978-3-319-59186-5_1

Providing documentation and descriptors for the data is also costly. In addition to defining and cleaning, documentation is developed separately for each dataset, as the variables and the procedures are created for each individual dataset. Furthermore, at the end of the process, datasets end up in locked box repositories, not easily accessible to the research community. As the storage and bandwidth necessary for saving and disseminating the data tends to be costly, reaching out across projects is a difficult and onerous operation.

In essence, big social data has created a research landscape of isolated projects. One of the costs of working in isolation is redundancy. Each time a research group aims to analyze a dataset, even if it is relatively well known and central (e.g., Wikipedia editorial history or open source software repositories), work starts anew. Furthermore, as the research products are delivered as papers and findings, the steps the data moved through, from raw, source data through intermediate data products and analysis products, are often lost to the idiosyncrasies of each lab's process. This makes cross-checking, secondary data analysis and methodological validation difficult at best to realize. Concerns go beyond research because the systematic limitations identified by big social data research mirror challenges faced in general data governance, civic action, education, and even business intelligence.

A number of specific solutions might address the issues commonly experienced by data-centric researchers and practitioners. For example, common data ontologies, social scientific analysis protocols, documentation standards, and dissemination workflows could generate repeatable processes. As new approaches emerge, training and teaching materials need to be created from the common store of previously accomplished work. Yet, for this, data professionals need to be trained in data extraction, curation, and analysis with an eye to integrating data procedures, analysis, and dissemination techniques.

The cacophony of current processes and the ideal of a "data factory" or an "open collaboration data exchange" are at two ends of a spectrum. The first is the state of practice; the second, aspirational. This volume aspires to support the second goal.

The "data factory approach" presented in this volume expresses several systematic approaches to tackle the challenges of data-centric research and practice. Our strategic goal is to open and consolidate the conversation on how to vertically integrate the process of data collection, analysis, and result dissemination by standardizing and unifying data workflows and scientific collaboration. One of our goals is to support those who work on projects to create repositories and documentation procedures for large datasets. At the same time, the ideal of a data factory needs to advance core methodologies for preprocessing, documenting, and storing data that can connect information sets across domains and research contexts. Successful implementation of data factory methodologies promises to improve collaborative research, validate methodologies, and widen the dissemination of data, procedures, and results.

Our conceptualization of "data factories" straddles many scholarly communities, including information studies, communication, sociology, computer science, data science, sociology, and political science. Industry practitioners focused on marketing, customer relations, business analytics, and business intelligence governmental

and policy analysts might also benefit from this vision. Scholars are piloting the assembly line in our conceptualization of data factories. We do not expect a "genesis moment" where a particular factory might emerge into the world due to the genius of one scholar or her research group. Instead, we expect domain- and question-specific paths to develop from each "data factory assembly line." From these initial assembly lines, practitioners, students, collaborators, and scholars in related disciplines will have a more solid starting place for their work or discourse. At the same time, we also hope that as new practices emerge, these will converge rather than diverge through the common methods discussed in this volume. The "data factory" vision aims to cross disciplinary boundaries. We hope that social computing researchers, computer scientists, social scientists, organizational scientists, and other scholars will in the end develop a common language.

The data factory approach can cover a variety of activities, but in a more tangible way and as an overture to our volume, its core activities should at the very least include:

1. Creating standard workflows for data processing and documentation and formatting inspired by a variety of projects and which cover several core dimensions: actors, behaviors, levels of analysis, artifacts, outcomes
2. Determining standard ontologies for categorizing in a standard manner records (observable units) and variables (fields)
3. Creating tools for easily processing existing and future datasets for standardized processing and documentation
4. Creating online storage and discovery tools that can easily identify records and variables across datasets, disciplines, and scientific observation domains
5. Facilitating data recombining by matching on various variables of heterogeneous datasets
6. Creating methods for documenting and sharing statistical tools and procedures for analyzing recombined datasets
7. Creating platforms and methods for research collaboration that rely on expertise, intellectual interest, skills, data ownership, and research goals for connecting individuals
8. Creating courses to teach these methods and to train graduate, undergraduate, and mid-career professional

The chapters of this volume address all of these issues, proposing, we hope, an integrated strategy for data factoring. Although some of the chapters can be read as "use case," "how to," or "guideline" contributions, their value should be seen in the context of the overall goal, which is to propose an integrated vision to what a "data factory" research and methodological program should be.

The volume is divided into three large sections. The first is dedicated to the theoretical principles of big data analysis and the needs associated with a data factory approach. The second proposes some theoretical principles and ideas for designing and deploying data factory approaches. The third presents these approaches in action through case studies of data-based research, best practice scenarios, or educational briefs.

The first two chapters are a theoretical overture to the rest of the volume.

The chapter "Accessibility and Flexibility: Two Organizing Principles for Big Data Collaboration" by Hemphill and Jackson argues that *accessibility and flexibility* are the two principles and practices that can bring big data projects the closest to a data factory ideal. The chapter elaborates on the necessity of these two principles, offering a reasoned explanation for their value in context. Using two big data social scientific research projects as a springboard for conversation, the chapter highlights both the advantages and the practical limits within which accessibility and flexibility principles move. The authors consciously avoid both utopian and dystopian tropes about big data approaches. In addition, they offer a critical feminist discussion of big data collaboration. Of particular interest are also the manners in which specific characteristics of big data projects, especially volume and velocity, may affect multidisciplinary collaborations.

The chapter "The Open Community Data Exchange: Advancing Data Sharing and Discovery in Open Online Community Science" by Sean P. Goggins and collaborators argues that while online behavior creates an enormous amount of digital data that can be the basis for a new level and kind of social science research, possibilities are hampered by many shortcomings. Scientists lack the tools, methods, and practices to combine, compare, contrast, and communicate about online behavior across domains of interest or temporal intervals. The chapter presents an effort to (1) specify an Open Community Data Exchange (OCDX) metadata standard to describe datasets, (2) introduce concepts from the data curation lifecycle to social computing research, and (3) describe candidate infrastructure for creating, editing, viewing, sharing, and analyzing manifests.

"Levels of Trace Data for Social and Behavioral Science Research" by Kevin Crowston opens the second part of the book, dedicated to designing strategies for data factories. It highlights another set of theoretical challenges brought about by the big data revolution. Data sources are not "primary" in the traditional sense of the word; they are most of the time "secondary." They are not recorded with the intention to capture human behaviors. Human behaviors are an incidental "capture" of social media data. While social media, which is at the heart of the big data revolution, are in the end tools that support and reflect human behaviors, information is captured incidentally, not purposefully. Check-ins, likes, reposts, and so on reflect a human act, not the meaning or the context of that act. In other words, data carries the mere traces of human behaviors as they are captured after the fact. Adopting a framework adapted from Earth Observation science, the paper proposes an avenue for advancing from partial to more complete understanding of the actions and contexts that generated social media data. The author suggests that the framework may be essential for shaping, sharing, and reusing of big social media data in a data factory context.

In the chapter "The 10 Adoption Drivers of Open Source Software that Enables e-Research in Data Factories for Open Innovations," Kerk Kee inventories factors that lead to the adoption of open source software production platforms, which are major sources for data factories, especially in the field of open innovation.

The inventory goes beyond description. Its goal is to isolate the factors that may predict adoption. By this, the chapter provides a map for identifying the most important prerequisites for developing long-lasting open innovation and potentially data factoring environments. The chapter also raises critical questions community stakeholders should keep in mind when promoting the diffusion and dissemination of software applications that will support data factories for open innovations.

The next chapter, "Aligning online social collaboration data around social order: theoretical considerations and measures," by Matei and Britt proposes that at a higher level of abstraction, datasets generated via data factories need to be comparable on the basis of a common theoretical and methodological ground. The core proposition is to align datasets around the conceptual framework of "social order." Social order is conceptualized as meaningful patterns of interaction that support convergent growth and evolution of online groups. Capturing social order can be accomplished through a series of measures, including social entropy and social network statistics (assortativity and various types of centrality). Theoretical alignment will make datasets not only comparable but the social scientific enterprise in the social media/big data realms more reliable and comprehensive.

Squire and Crowston in "Lessons learned from a decade of FLOSS data collection" open up the third part of the volume, dedicated to practical applications and teaching initiatives. The chapter presents one of the most ambitious data collection and dissemination initiatives, FLOSSmole, which is one of the first projects that embraced a data factory vision. The project is dedicated to understanding how Free/Libre Open Source Software (FLOSS) projects emerge, survive, are successful, or die. Embodying the FLOSS ethos, the project relied on a public-facing repository for data and analyses, encouraging other researchers to use it and contribute to it. The chapter presents the project emergence, design, goals, and, most important, lessons learned. Especially relevant for this book are the conclusions regarding sustainability and relevance of large, data factory-like, data collection, collaboration, and dissemination.

Mahmud, Hogan, Zeffiro, and Hemphill continue the third part of the volume with the chapter "Teaching Students How (NOT) to Lie, Manipulate, and Mislead with Information Visualizations." The authors delve on the intellectual and pedagogical implications of big data visualizations. Representing data visually implies simplifying and essentializing information. However, the selective nature of information visualization can lend itself to lies, manipulations, and misleading information. To avoid these pitfalls, data analysts should focus and embrace specific principles and practices that aim to represent complete, contextualized, comparable, and scalable information, in a way that reveals rather than isolates the viewer and the problem at hand from the problem space it reflects.

The chapter "Democratizing Data Science: The Community Data Science Workshops and Classes" by Hill, Dailey, Guy, Lewis, Matsuzaki, and Morgan introduces the pedagogical concept of "community data science" and the practices associated with it. The chapter reviews several years of experimentation in designing course materials and teaching data science as short workshops and long-form graduate

seminars. The goals of the learning activities were twofold: to teach new methods for scientific inquiry and to democratize access to social scientific methods, especially those applied to big, social media data. The chapter discusses both the philosophy and the lesson learned from course evaluations.

We hope that the collection of chapters gathered within the covers of this volume creates a round, complementary vision of what a data factory perspective can and should be. The ultimate "prize" is to help the next generation of researchers, teachers, and practitioners avoid the mistakes of the previous generations. Of these, the most costly is the temptation to reinvent the wheel. Data factoring should and can help the research and practitioner community root their efforts in a vision of information gathering, analysis, and sharing that is not only more open but also evolutionary. New practices and ideas should build and extend the old ones. This will make data factoring and open social media research more productive and more inclusive.

Part I
Theoretical Principles and Approaches to Data Factories

Chapter 2
Accessibility and Flexibility: Two Organizing Principles for Big Data Collaboration

Libby Hemphill and Susan T. Jackson

Introduction

This chapter's main argument is that in big data collaborations both the data and the collaboration ought to be *accessible* and *flexible*. We offer reflections and recommendations on approaches to big data collaboration through the vehicles of two cases of collaborative big social data research. We avoid both the utopian and dystopian tropes so often found in conversations about big data while still offering a critical feminist discussion of big data collaboration. We focus here on the challenges presented by the volume and velocity of big social data for multidisciplinary collaborations. We address challenges to organizing both the staff and data required by such endeavors and ground our discussion in details from two international collaborations that study political uses of social media.

While we offer general recommendations for approaching and managing big data collaborations, we do not offer specific "best practices," ontologies, or metadata standards for big [social] data. These omissions are purposeful. Instead of offering a set of practices, we propose a set of principles that should guide decisions about and within collaborations. Instead of proposing ontologies, we focus on how human beings, rather than machines, will use data. Making machine-readable ontologies for data that humans can understand in other ways takes resources and time away from the intellectual work those data may support. The work of using open data need not wait for us to make ontologies.

L. Hemphill (✉)
University of Michigan, Ann Arbor, MI, USA
e-mail: libbyh@umich.edu

S.T. Jackson
Stockholm University, Stockholm, Sweden

© Springer International Publishing AG 2017

S.A. Matei et al. (eds.), *Big Data Factories*, Computational Social Sciences,
https://doi.org/10.1007/978-3-319-59186-5_2

A Short Note on Ethics in Big Data Collaboration

The ethical standards governing big data research are in flux, but their lack of fixity does not mean we should ignore the ethical aspects of big data collaborations. At the very least, we suggest that researchers "only use the data that you need, and that doesn't create risk for others" (Metcalf and Crawford 2016, footnote 5). Privacy is certainly an important ethical concern for big data researchers, and we argue that researchers must consider the ethics of propensity, access, and data sharing and combination as well. For more detailed discussions about privacy and big data ethics generally, we recommend a special issue of *First Monday* ("Making data – Big data and beyond" 2013), another in the *International Journal of Communication* (Crawford et al. 2014), and recent discussions about the OKCupid data leak (Markham 2016; Zimmer 2016).

When the goal of analysis is to predict rather than to understand causation (Siegel 2013), big data presents a unique set of ethical challenges in addition to issues of privacy. Predictive policing provides a useful case study here: what is the responsibility of an agency to act on the knowledge that a harmful event is 95 % likely to occur? What about 80 %? Setting intervention thresholds presents ethical dilemmas for those with access to the data. As Zwitter (2014) mentions, the impacts on the people affected by the intervention also must be considered.

By "ethics of access" we mean that researchers must recognize the power of big data collectors and users relative to the individuals whose behavior constitutes that data. Andrejevic (2014) calls this the "big data divide" and argues that data mining's ability to detect unexpected correlations requires our ethical consideration because users are often unaware of how their data is used, thereby reinforcing and potentially exacerbating power imbalances between the users and the data miners. We recognize another big data divide between researchers who collect big data and those who wish to analyze it. The disciplinary posture that enables data mining (where running multiple analyses without making specific predictions about relationships ahead of time is normal) accounts for some of this divide, and relative differences in technical expertise are also at play. One common argument for addressing this second divide is to have data collectors and data analyzers collaborate, but as we discuss below, such collaborations are difficult at best.

Much like combining disciplinary expertise creates problems for researchers, combining datasets and separating them from their contexts also can produce unexpected and even harmful outcomes. One challenge we face in collaboration is that often, even when working with anonymized datasets, combining datasets can reveal the identities of individuals, putting them at risk for a variety of harms. King (2011) offers specific suggestions for social scientists facilitating data sharing while protecting individuals' privacy and serves as a useful overview of the issues data sharing efforts currently face. Existing work on privacy-preserving data mining also is informative here (e.g., Aggarwal and Yu 2008a, b; Hajian et al. 2014; Sánchez and Batet 2016; Xu et al. 2016), arguing that privacy protection in data mining and other big data endeavors is a necessary and promising field of research in itself.

Overview of Projects from Which We're Drawing

To ground our discussion of approaches to big social data collaboration, we use stories and experiences from our work studying politicians and militarism in two separate international collaborations.

Politicians and Social Media (POSM): This research project conducted from 2011 to 2016 examined the use of social media by politicians in the USA, the European Union, and the Republic of Korea. Team members on the project were located in the USA, Cyprus, and Korea. We were interested primarily in Twitter use by elected officials and its impacts on political news (Shapiro and Hemphill in press) and constituent communication (Hemphill et al. 2013; Hemphill and Roback 2014). We leverage data from social media (mainly Twitter but also blogs and websites), *The New York Times*, and the US census to address these issues.

Militarization 2.0 (Mil2.0): "Militarization's social media footprint through a gendered lens" is a 5-year international collaboration funded by the Swedish Science Research Council. It is part of the Digitized Society—Past, Present, and Future framework grant series and is meant to establish a foundation for studying militarization on/in/through social media. The research covers three industries that are generally overlooked in the International Relations (IR) literature: conventional arms production, military videogames, and private and military security. The project looks at industry actors through a gendered lens to see whether and how we can understand the persistence of militarism more broadly by focusing on a particular set of representations on social media. The principal investigator is based in Sweden and the grant includes teams at universities in the UK and Germany. Across the project, we focus on YouTube, Facebook, and Twitter.

Challenges in Big Data Collaboration

Big data collaborations face many of the same challenges as other scientific collaborations, and we recommend *Scientific Collaboration on the Internet* (Olson et al. 2008), an edited volume from MIT Press, and a large body of work on collaboration in science and engineering research (e.g., Bietz and Birnholtz 2003; Bozeman and Corley 2004; Coleman and Rippin 2000; Corley et al. 2006; Cummings and Kiesler 2003; Olson and Olson 2000) for more information on those fundamental difficulties (e.g., infrastructure and incentives). A central challenge for scholars to utilizing big data is a lack of technical skills to gather and access the sheer volume and velocity of digital data[1] and to link that data with conceptual development that moves data

[1]The designs of social media and their affordances for data scientists also impact scholars' ability to work with big social data, for instance, APIs and terms of service change, affecting what data is available and under what conditions. Tools modify the core functions and impact the behaviors users are able to engage in – e.g., Twitter is removing usernames and media attachments from

analysis from standard causation claims (as is usual in conventional large-N studies) to the correlation claims for which big data scholars call. One area where this type of hesitation becomes apparent is in sampling. Typically, quantitative IR scholars rely on large-N datasets that are seen as "complete" in part because the variables are conceptualized and then populated, whereas big data datasets are more "fluid" because the data is captured and then can be used to facilitate pattern finding. Big data can be especially helpful in generating new questions since the "sample" does not need to be decided in advance. It is possible through identifying patterns to locate areas for further, deeper investigation (both patterns that show presence and patterns that show absence) (Cukier and Mayer-Schoenberger 2013).

Differing Technical and Theoretical Skills

This gap between approaches to data would suggest that scholars need to branch out and collaborate more closely with those who have the technical and methodological/method skills for generating, processing, and analyzing big data. However, collaboration across disciplinary lines presents its own challenges. In addition to communication challenges across the project and across disciplines, general software issues also have been barriers. The Mil2.0 project's computer science masters' students collected data, but still some team members have no access to the data they captured because they do not have the necessary computer science (CS) skills, e.g., the ability to query in SQL. The students are now learning how to format the data in ways it can be exported that will be useful, but another difficulty has been relaying to them in non-CS terms what technical issues they as developers face in formatting the data.

Conversations about big data often ignore the other side of this methodological coin—that people with the technical skills to collect and manage big data also lack the training necessary to responsibly and appropriately analyze that data. Social theories help us distinguish signal from noise (González-Bailón 2013) and help us understand how the particulars of the data (e.g., the technology used to collect it, the vantage point it uses) shape it and the insights we can draw from it (Kitchin 2014; Ribes and Jackson 2013). Also, the idea of letting data tell researchers what to attend to (as big data evangelists suggest when they call for an end to theory (Anderson 2008)) is hardly a new concept in social sciences (see, e.g., Charmaz 2006; Glaser and Strauss 1967). As much as big data presents technical challenges, it presents theory challenges as well, and we turn shortly to the epistemological

the 140-character limit; Facebook does not treat all crisis and find your friend functionality the same way. The legal milieu about rights to be forgotten differs between the USA and European Union. This obviously is not an exhaustive list but rather an illustrative one that makes clear that the changing technological landscape impacts the research we can do.

debates that big data collaborations must address. First, a brief discussion of why learning new things—whether technical skills or social theories—is actually quite difficult, even for professional researchers.

Expertise and Identity

Acquiring new methodological capabilities and theoretical orientations presents challenges to egos, research agendas, budgets, and time. In general, researchers face two main issues when considering methodology: determining the appropriate means for discovering or producing knowledge and determining the validity of the knowledge produced by different methods (see, e.g., Bird 2012; Jackson 2010). Scholars are uncomfortable relying on others or on a mid-level (rather than expert level) of knowledge in order to make methodological determinations.

Along these lines, Bleiker (2015) recently prompted IR scholars to think outside their respective comfort zones and approach social media and other digital research using mixed methods—to learn new skills and to work more collaboratively across methodologies and disciplines. He stated that IR scholars, and we would agree, generally are resistant to using a variety of methods because of the difficulty in becoming an expert in multiple, potentially unrelated, methods and because of the hesitation to work with something in which one is not an expert. As Bleiker points out, there is pressure in academia to appear to have unquestioned authority and therefore can limit one's toolbox to one main methodological approach and one or two key methods—a limitation that can impact the depth and breadth of big social data research.

Competing Epistemologies

One of the main challenges present in big data collaborations that include both technical and social scientists is that people trained in technical fields often possess fundamentally different research values and subscribe to epistemologies that conflict with the social scientists' (or humanists' in the POSM case) values and epistemologies.[2] What it means to know something in computer science is different from what it means to know something in communications or in IR, for instance; similarly, what constitutes a valid or reasonable claim also differs, sometimes irreconcilably. Referring to the privileging of computational skills (as compared to other skills such as small-N qualitative methods), Boyd and Crawford

[2]This statement is not meant to conflate the epistemological differences or conflicts within the social sciences but rather to point to the broad stroke differences between the logics in computational or machine-based sciences and those found in the sciences that are people centered.

(2012) discuss the gendering of skills, and they note (citing Harding and others) that who asks the questions impacts what questions get asked and what data gets pulled. McCarthy (2013) reminds us that while technological determinism is not all-pervasive, there is a hierarchy in terms of who has input in how our systems work and thus in how we use these systems. These types of issues/hurdles need much further investigation by social scientists in collaboration with computational experts inside of and outside of the social sciences.

Both within and across disciplines, the hesitation to consider new ways of conceptualizing and operationalizing what in the past have been contested constructed variables can pose high barriers to collaboration in big social data projects. For example, in issues of gender, race, class, sexuality, ability, and so on, there has been a separation between the positivists and post-positivists that leaves each end of the spectrum talking past each other or not talking at all. Scholars who do gender analysis tend toward small-N qualitative studies and generally are resistant to ideas about incorporating a broader range of empirical indicators into discussions on "measuring" or tracking data. Quantitative scholars who do include a "gender" indicator often rely on a sex variable as an indication of gendered behavior (see below for more discussion).

Because of the predictive quality of big social data, it might be possible to use these data for intersectional analyses in which the data can transcend the typical binaries that conventional large-N data assigns to various social categories such as gender, race, class, sexuality, and so on. That is, we might be able to understand digital behavior as performance (see, e.g., Boyd 2014; Hogan 2010; Marwick and Boyd 2014) and in this way capture the intersection of these various social characteristics and how people perform online. In that way, instead of relying on the sex variable to make claims about how women and men behave on social media, we could experiment with various ways to "measure" behavior in terms of feminine and masculine and how these behaviors intersect with other aspects of people's identities. This type of approach is seen, for instance, in the work on gendered language in which both women and men used more masculine language in social media spaces considered to be masculine and feminine language on feminine spaces (Bamman et al. 2014). This example of gender is also useful for illustrating the next challenge we discuss: concept and variable construction.

Concept and Variable Construction

Another challenge can be explaining concepts and eventual data use across disciplines, e.g., how the database the technicians are creating does not actually contain the indicators researchers eventually need conceptually and theoretically. For example, the Mil2.0 project relies on critical perspectives including gender as a central aspect of militarization (Jackson 2016). That said, it took several weeks to explain how the militarism database the Mil2.0 CS students were building would not have a gender indicator per se since gender should be conceptualized as more than

a binary sex variable. This is true not just because we are arguing to move beyond the gender binary, but also because these types of user-indicated variables do not account for missing data or for when users intentionally choose a different indicator, e.g., when females designate themselves males in order to avoid online harassment. Yet, the students asked on several occasions to point out which field/column in the database would house the gender variable. The two "sides" of this conversation had very different ways of conceptualizing and therefore operationalizing gender, difference that stem from the epistemological choices being made.

As stated earlier, because of its sheer volume, big data might make it possible to operationalize very complicated concepts in new ways. For feminist scholars, gender is by and large recognized as socially constructed, though a tension between positivists and post-positivists on how to capture this construction remains. Sarantakos (2012, p. 67) broadly outlines feminist research claiming that at its foundation feminist research is "based on the assumption that the world is socially constructed, displays a relative aversion to empirical positivist methodology, and rejects the value-free nature of research." Gender and sex often are used interchangeably (Caprioli 2004); however by now it should be generally commonly accepted that the terms are socially constructed. Through big data analysis, we can understand people's online behavior and potentially use data to "measure" gender as a social construct by bypassing the binary sex variable. Within the Mil2.0 project, we are looking at whether, how, and to what degree empirical data can inform our understanding of gender in relation to militarization and social media. Both the quantitative feminist literatures on social media and IR challenge conventional scholarship to nuance its understanding of gender to more than a binary sex variable and to rely on theory to make better measures and proxies (e.g., Bamman et al. 2014). But is this possible given the constraints of measuring gender as a social construct? There have been some inroads into this debate and the use of nuanced gender variables, for example, Caprioli's (2009) cluster variables.

However, there are times when sex can be an indicator of gender, e.g., in measuring gender inequality by mapping where women are and are not in order to analyze why. Because with Mil2.0 we have an IR project (albeit infused with multidisciplinary perspectives) that was proposed as a response to the many deficiencies in conventional IR, we have to consider the empirical/analytical distinction made in both conventional and feminist scholarship. A workshop on Feminism and Social Media Research at the ACM Conference on Computer-Supported Cooperative Work explored the complications the gender field on Facebook (FB) presents (Hemphill et al. 2014). The workshop notes point out that while FB has expanded its gender field to include many more categories for gender representation, FB still imposes its own notions of what gender is and who gets to select a more complex self-identity. The workshop participants also formulated a number of questions that reflect a wide variety of interests and concerns that among other things point to how the exposure to expanded categories can impact how the wider audience thinks, whether/how youth might be impacted by having a wider presentation of gender choices, if there will be an impact on automated advertising, and what role FB and other Internet giants play in society more generally.

The questions/issues posed by the CSCW workshop are the same type of questions we face in the Mil2.0 project because from the social media literature, we can call into question how performing gender also needs to account for the impact of external constraints on self-identification and what UGC results through this performance. To quote the workshop website in response to the question how can we research gender in social media: "Qualitatively. Quantitatively. Using mixed methods. With situated data, contextualized data, and thick description. Acknowledging historical context and noting power structures. Respecting our subjects by situating social media users as people" (Hemphill et al. 2014). We would add that we need to keep in mind what impact we want to have and who our audience is. As this discussion of a "gender" variable illustrates, big data collaborations must wrestle with how to reconcile incompatible concepts and how to construct relevant variables that meet the needs of multiple disciplinary perspectives.

We have introduced a number of challenges that face big data collaborations, especially those that involve researchers from multiple disciplines: technical and theoretical skill differences, egos and demands of expertise, epistemological incompatibilities, and variable construction. In the next section, we propose the principles of *accessibility* and *flexibility* as tools for avoiding or addressing these issues.

Organizing Principles for Productive Big Data Collaboration

We present the principles of *accessibility* and *flexibility* alphabetically because their relative salience will depend on the specifics of each collaboration and the data it examines. We have chosen to present these organizing principles instead of step-by-step or more specific advice because of big data's velocity and variability. These principles hold for all big data collaboration regardless of the data's origins or the investigators' disciplinary trainings.

Accessibility

Accessibility is the principle designed to address the problems of differing skill levels and competing disciplinary values. We encourage you to think of your big data collaboration as an "environment" and strive to make it accessible. We borrow the concept of accessibility from disability and rehabilitation studies (see, e.g., Iwarsson and Stahl 2003). Even there, it has a number of complementary definitions including "capable of being entered or approached; easy of access; readily reached or got hold of" ("accessible, adj." n.d.), and "an accessible environment must match the abilities of an individual or a group" (Iwarsson and Stahl 2003, p. 58). We are deliberately invoking a physical space metaphor here—what would it mean for your collaboration to match the abilities of the individuals involved? How would your data need to be formatted in order for it to be readily reached? A Heideggerian

sense of "ready" is useful here: for Heidegger tools are "ready-to-hand" when we can use them unreflexively (Heidegger and Stambaugh 1996). That means that we can use the data without having to think about how it came to be. This does not mean, of course, that we should not be able to access information about how it came to be but rather than the details of equipment, in this case data, should be transparent (Dreyfus 1991; Koschmann et al. 1998). An accessible collaboration environment, then, is one where the tools are transparent and where the individuals' abilities are matched.

In practical terms, this means that data and resources necessary for the project will be available in formats and procedures that match the abilities of all of the members of the team. That also means that code will be usable even by people who did not write it and that theoretical arguments will be presented in ways that non-experts can understand.

Flexibility

The principle of flexibility helps collaborations address the problems of data velocity, competing disciplinary values, and general challenges to large-scale collaborative work. We recommend a sort of "agile management" (Anderson 2003; Morien 2005) approach for researchers to ensure flexibility without sacrificing rigor or productivity. While agile management will help address challenges in the collection and analysis of data and the drafting of publications and presentations, special attention is required in respect to time.

We borrow "agile" from software development where agile management is characterized by uncertainty, frequent deliverables, and recognizable constraints (Morien 2005). While generally thought of as something for the IT industry, collaborative big social data research would benefit from research that has some set parameters but that is "agile" enough to make adjustments along the way. It requires transparent, open, and ongoing channels of communication among the project stakeholders. This approach is useful for bringing together technology and non-technology staff as well as any different perspectives from within the social science and/or humanist elements of the team, such that project requirements become more defined and developed over time as the different parts of the team learn each other's "language." While agile management is a crucial organizational tool, it is just as important to spell out basic expectations from the start. What is the final deadline? How will the team work toward meeting that point, e.g., what intermediate timing goals are appropriate? What kind of style will the output have? How much time does each participant have to offer for the output? Are there potential hazards that might need to be accommodated, e.g., unexpected childcare, external work responsibilities such as advising, or conference organizing? How will the team communicate when it is time to reorganize?

To illustrate the utility of an agile approach, we use the example of writing a publication. As with academics feeling uncomfortable outside their own zones of

expertise, we often do not want to share drafts that are not near actual completion and looking polished. Nor do we excel at speaking openly and honestly about expectations of the working process during the collaboration. It is almost as if writing happens magically—that somehow publications just materialize when necessary, but obviously researchers know that not to be the case. We recommend that researchers establish general communication guidelines at the front of the collaboration, guidelines that include basic respect as well as vocalization about preferred means for feedback. For instance, some people need to have verbal communication and would prefer to stop by for a coffee or to Skype if at a distance; others need feedback in writing. Whatever style, within reason each researcher should be accommodated along the way.

It is important that team members remain flexible during the collaboration, whether within the team about conflicting ideas on the output and handling the iterative process or because of external factors that come up along the way that might change who is available when for the project. Another potential area that might require flexibility is with data gathering and issues involving data host's decisions on the malleability of data properties. Given the velocity at which big data moves and accumulates, it is tempting to assume that work must also happen constantly and at a fever pitch. However, that is just not true. Good scholarship takes time to think, to analyze, and to write (Mountz et al. 2015), and flexible collaborations are able to make time for those activities while respecting the whole persons and whole lives of all the members of the team.

Flexibility with regard to the data itself is also paramount. Rather than propose rigid ontologies or controlled vocabularies or specific metadata standards, we recommend big data collaborations use rigorous documentation and flexible data structures to manage their data. It is helpful here to think of each collaboration as a data sharing effort. Doing so foregrounds the challenges of documentation, data context, tacit knowledge, data quality, and misaligned incentives that plague broader data sharing efforts (Bietz and Birnholtz 2003). These challenges are exacerbated when researchers disagree on the types of problems to be addressed and the specific methods to be employed (as we have shown in common in big data collaborations). Now that storage costs, even for big data, are declining, it is becoming possible to afford to store and share data in multiple formats, rendering it accessible throughout the team.

All relationships require some kind of trust, both in the other people involved and in yourself, to be effective. Trust is important as researchers that we know what we can and cannot do and are willing to learn, that we offer to others what is reasonable but challenging and fun at the same time, and that we trust in our partners to be open and honest as well. Trust affords flexibility by ensuring reliability (Das and Teng 1998)—it is possible to be flexible in accommodating external responsibilities when I know I can rely on my colleague to do her part.

Trust also affords flexibility through face-saving. One important lesson learned from the projects referred to here is the importance of knowing when to admit you do not know something, when to admit that you could use some help, and when to be willing to learn something new (even if it will not make you an expert in

that something new). As mentioned above, threats to one's identity as an expert are especially challenging for researchers, and collaborative partners must be able to trust both one another's expertise and its limits. The learning curve from trying to stay within your respective discipline and use the same old toolbox is really steep. It might take time to learn a bit of something new, but it seems to take a lot longer to reinvent the wheel. We encourage researchers at the very least to try a new method and a new way of looking at things and to work with others who not only challenge your notions but also can offer support in developing them further. We, the authors, did just that together and have had a fruitful and fun collaboration that has led to co-convened workshops in New Orleans (USA) and Tübingen (Germany), joint authorship, and the plan to look for ways to hold data trainings to bring together technical and nontechnical academics to do big social data research.

Conclusion

We proposed accessibility and flexibility as guiding principles for successful big data collaborations. Together, these principles prepare researchers to address the challenges posed by interdisciplinary collaboration, the volume and velocity of big data and its associated systems, and the constraints created by researchers having other responsibilities. Using the environment metaphor when talking about accessibility encourages us to think about how skills and abilities may differ across the team. The agile approach to project management popular in software development provides a useful model for managing flexible research collaborations. These principles provide a common frame for big data collaborations that remains adaptable to the specific disciplinary and infrastructural positions collaborations face, and we hope you find them as useful as we have in our own work.

References

accessible, adj. (n.d.). *OED online*. Oxford University Press. Retrieved from http://www.oed.com/view/Entry/1034.

Aggarwal, C. C., & Yu, P. S. (2008a). A general survey of privacy-preserving data mining models and algorithms. In C. C. Aggarwal & P. S. Yu (Eds.), *Privacy-preserving data mining* (pp. 11–52). New York: Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-0-387-70992-5_2.

Aggarwal, C. C., & Yu, P. S. (Eds.). (2008b). *Privacy-preserving data mining: Models and algorithms*. New York: Springer.

Anderson, D. J. (2003). *Agile management for software engineering: Applying the theory of constraints for business results*. Upper Saddle River: Prentice Hall Professional.

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. Retrieved June 13, 2016, from http://www.wired.com/2008/06/pb-theory/.

Andrejevic, M. (2014). Big data, big questions| the big data divide. *International Journal of Communication, 8*(0), 17.

Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of SocioLinguistics, 18*(2), 135–160.

Bietz, M. J., & Birnholtz, J. P. (2003). *Data at work: Supporting sharing in science and engineering* (pp. 339–348). New York: ACM. Retrieved from http://doi.acm.org/10.1145/958160.958215.

Bird, S. (2012). *Feminist methods of research*. Iowa State University. Retrieved from http://www.slideshare.net/miryammastrella/presentation-on-feminist-methods-of-research.

Bleiker, R. (2015). Pluralist methods for visual global politics. *Millennium – Journal of International Studies, 43*(3), 872–890. https://doi.org/10.1177/0305829815583084.

Boyd, D. (2014). *It's complicated: The social lives of networked teens*. New Haven: Yale University Press.

Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society, 15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878.

Bozeman, B., & Corley, E. (2004). Scientists' collaboration strategies: Implications for scientific and technical human capital. *Research Policy, 33*(4), 599–616. https://doi.org/10.1016/j.respol.2004.01.008.

Caprioli, M. (2004). Feminist IR theory and quantitative methodology: A critical analysis1. *International Studies Review, 6*(2), 253–269.

Caprioli, M. (2009). Making choices. *Politics & Gender, 5*(03), 426–431.

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis* (2nd ed.). Los Angeles: Sage Publications.

Coleman, G., & Rippin, A. (2000). Putting feminist theory to work: Collaboration as a means towards organizational change. *Organization, 7*(4), 573–587. https://doi.org/10.1177/135050840074004.

Corley, E. A., Boardman, P. C., & Bozeman, B. (2006). Design and the management of multi-institutional research collaborations: Theoretical implications from two case studies. *Research Policy, 35*(7), 975–993. https://doi.org/10.1016/j.respol.2006.05.003.

Crawford, K., Gray, M., & Miltner, K. (Eds.). (2014). Big data|critiquing big data: Politics, ethics, epistemology. *International Journal of Communication, 8*, 1663–1672.

Cukier, K. N., & Mayer-Schoenberger, V. (2013). The rise of big data: How it's changing the way we think about the world. *Foreign Affairs*. Retrieved from https://www.foreignaffairs.com/articles/2013-04-03/rise-big-data.

Cummings, J. N., & Kiesler, S. (2003). *KDI initiative: Multidisciplinary scientific collaborations*. Washington, DC: National Science Foundation. Retrieved from http://netvis.fuqua.duke.edu/papers/NSF_KDI_report.pdf.

Das, T. K., & Teng, B.-S. (1998). Between trust and control: Developing confidence in partner cooperation in alliances. *Academy of Management Review, 23*(3), 491–512. https://doi.org/10.5465/AMR.1998.926623.

Dreyfus, H. L. (1991). *Being-in-the-world: A commentary on Heidegger's being and time, division I*. Cambridge, MA: MIT Press.

Glaser, B., & Strauss, A. (1967). *Discovery of grounded theory: Strategies for qualitative research*. New Brunswick: Aldine.

González-Bailón, S. (2013). Social science in the era of big data. *Policy & Internet, 5*(2), 147–160. https://doi.org/10.1002/1944-2866.POI328.

Hajian, S., Domingo-Ferrer, J., Monreale, A., Pedreschi, D., & Giannotti, F. (2014). Discrimination- and privacy-aware patterns. *Data Mining and Knowledge Discovery, 29*(6), 1733–1782. https://doi.org/10.1007/s10618-014-0393-7.

Heidegger, M., & Stambaugh, J. (1996). *Being and time: A translation of Sein und Zeit*. Albany: State University of New York Press.

Hemphill, L., Roback, A. J. (2014). Tweet acts: How constituents lobby congress via twitter. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 1200–1210). New York: ACM. https://doi.org/10.1145/2531602.2531735.

Hemphill, L., Otterbacher, J., Shapiro, M. (2013). What's congress doing on twitter? In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (pp. 877–886). New York: ACM. https://doi.org/10.1145/2441776.2441876.

Hemphill, L., Erickson, I., Ribes, D., Mergel, I. (2014). Feminism and social media research. In *Proceedings of the Companion Publication of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 319–322). New York: ACM. https://doi.org/10.1145/2556420.2558864.

Hogan, B. (2010). The presentation of self in the age of social media: Distinguishing performances and exhibitions online. Bulletin of Science, Technology & Society, 0270467610385893. 30 377–386. https://doi.org/10.1177/0270467610385893.

Iwarsson, S., & Stahl, A. (2003). Accessibility, usability and universal design – positioning and definition of concepts describing person-environment relationships. *Disability and Rehabilitation, 25*(2), 57.

Jackson, P. T. (2010). *The conduct of inquiry in international relations: Philosophy of science and its implications for the study of world politics*. London: Routledge.

Jackson, S. T. (2016). Marketing militarism in the digital age: Arms production, YouTube, and selling "national security." In Hamilton & Shepherd, *Understanding popular culture and world politics in the digital age*. Oxon: Routledge.

King, G. (2011). Ensuring the data-rich future of the social sciences. *Science (New York, N.Y.), 331*(6018), 719–721. https://doi.org/10.1126/science.1197872.

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1), 2053951714528481. https://doi.org/10.1177/2053951714528481.

Koschmann, T., Kuutti, K., & Hickman, L. (1998). The concept of breakdown in Heidegger, Leont'ev, and Dewey and its implications for education. *Mind, Culture, and Activity, 5*(1), 25–41. https://doi.org/10.1207/s15327884mca0501_3.

Making data – Big data and beyond. (2013). *First Monday*, *18*(10).

Markham, A. (2016). OKCupid data release fiasco: It's time to rethink ethics education. Retrieved May 18, 2016, from https://medium.com/@amarkham/okcupid-data-release-fiasco-ba0388348cd#.3irk2qca8.

Marwick, A. E., & Boyd, D. (2014). Networked privacy: How teenagers negotiate context in social media. *New Media & Society, 16*(7), 1051–1067. https://doi.org/10.1177/1461444814543995.

McCarthy, D. R. (2013). Technology and "the International" or: How I learned to stop worrying and love determinism. *Millennium – Journal of International Studies*, 0305829813484636. https://doi.org/10.1177/0305829813484636.

Metcalf, J., & Crawford, K. (2016). Where are human subjects in big data research? The emerging ethics divide. *Big Data & Society*. Retrieved from http://papers.ssrn.com/abstract=2779647.

Morien, R. (2005). Agile management and the Toyota way for software project management. In *Industrial informatics, 2005*. INDIN'05. 2005 3rd IEEE International Conference on (pp. 516–522). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1560430.

Mountz, A., Bonds, A., Mansfield, B., Loyd, J., Hyndman, J., Walton-Roberts, M., … Curran, W. (2015). For slow scholarship: A feminist politics of resistance through collective action in the neoliberal university. *ACME: An International E-Journal for Critical Geographies*, *14*(4), 1235–1259.

Olson, G. M., & Olson, J. S. (2000). Distance matters. *Human Computer Interaction, 15*(2 & 3), 139–178.

Olson, G. M., Zimmerman, A., & Bos, N. (2008). *Scientific collaboration on the internet*. Cambridge: MIT Press.

Ribes, D., & Jackson, S. L. (2013). Data bite man: The work of sustaining a long-term study. In L. Gitelman (Ed.), *"Raw data" is an oxymoron* (pp. 147–166). Cambridge, MA: The MIT Press.

Sánchez, D., & Batet, M. (2016). C-sanitized: A privacy model for document redaction and sanitization. *Journal of the Association for Information Science and Technology, 67*(1), 148–163. https://doi.org/10.1002/asi.23363.

Sarantakos, S. (2012). *Social research* (4th ed.). South Melbourne: Palgrave Macmillan.

Shapiro, M. A., & Hemphill, L. (in press). Agenda building & indexing: Does the U.S. congress direct New York Times content through Twitter? *Policy & Internet*.

Siegel, E. (2013). *Predictive analytics: The power to predict who will click, buy, lie, or die*. Hoboken: Wiley.

Xu, L., Jiang, C., Chen, Y., Wang, J., & Ren, Y. (2016). A framework for categorizing and applying privacy-preservation techniques in big data mining. *Computer, 49*(2), 54–62. https://doi.org/10.1109/MC.2016.43.

Zimmer, M. (2016). OkCupid study reveals the perils of big-data science. Retrieved May 18, 2016, from https://www.wired.com/2016/05/okcupid-study-reveals-perils-big-data-science/.

Zwitter, A. (2014). Big data ethics. *Big Data & Society*, *1*(2). https://doi.org/10.1177/2053951714559253.

Chapter 3
The Open Community Data Exchange: Advancing Data Sharing and Discovery in Open Online Community Science

Sean P. Goggins, A. J. Million, Georg J. P. Link, Matt Germonprez, and Kristen Schuster

Introduction

In social computing research today, it is unlikely that any two papers from different labs examining Wikipedia, GitHub, eBird, Facebook, Twitter, or any other space where open online communities emerge will have (a) clear descriptions of the provenance of their data, (b) open access to scripts and anonymized samples of the data, (c) complete methods descriptions, or (d) consistency between them, even if the research questions are similar. As scientists, we should recognize that this state of affairs represents a hole in our process; not a hole to acknowledge and accept, but a scientific hole we should actively seek to close. The development of the open collaboration data exchange (OCDX) is one attempt to close this whole.

Open online communities (OOCs) are fundamentally distinct phenomena that facilitate the collective construction of flexible, distributed, and nonhierarchical forms of organization. The emergence of widely available, highly flexible, interactive information infrastructure technologies significantly altered the universe of feasible organization structures and strategies. OOCs represent a new class of organizing solutions, in which individuals self-organize in order to collaboratively

S.P. Goggins (✉) • A.J. Million
University of Missouri, Columbia, MO, USA
e-mail: gogginss@missouri.edu; ajmillion@gmail.com

G.J.P. Link • M. Germonprez
University of Nebraska at Omaha, Omaha, NE, USA
e-mail: glink@unomaha.edu; mgermonprez@unomaha.edu

K. Schuster
Kings College—London, London, UK
e-mail: kristen.schuster@kcl.ac.uk

© Springer International Publishing AG 2017

S.A. Matei et al. (eds.), *Big Data Factories*, Computational Social Sciences,
https://doi.org/10.1007/978-3-319-59186-5_3

produce any number of artifacts and experiences. OOCs differ from other popular online structures, such as crowdsourcing platforms or online social networks in significant ways. In crowdsourcing, the firm or client proposing the project typically controls the decision-making process. In online social networks, organized collective production is not usually a goal for participants.

Online behavior creates an enormous amount of digital data that can be the basis for social science research. Such behavioral data has been used for research in diverse online contexts, such as scientific advances (Irwin 1995), online learning outcomes (Bishop and Verleger 2013), political use of social media (Nahon and Hemsley 2014), citizen engagement (Tandoc 2014), group identity formation (Ren et al. 2007), and valued health benefits (Moorhead et al. 2013). To date, however, this science has been conducted piecemeal, one Internet address at a time, often without social or scholarly impact beyond the site's own stakeholders. Thus, there is an urgent scientific need to make sense of human behavior across technologies and an urgent human need to better understand how to apply online technologies for social benefit. To address these scientific and human needs, we propose a cyber-infrastructure that will enable researchers to effectively look across online contexts to explain, in more general terms, (1) how online interactions affect participants, groups, and society as a whole and (2) how to design online communities and platforms to maximize their positive effects.

Addressing these issues requires the systematic sharing and analysis of datasets that are currently fragmented and unavailable to most researchers. Scientists lack the tools, methods, and practices to combine, compare, contrast, and communicate about online behavior across location and over time. This is not because the differences across sites are poorly understood. Goggins et al. (2013), for example, provide a coherent ontological framework for classifying online human interactions as principally between people and each other (i.e., online health forums) or people and artifacts (i.e., ebird.org). To advance science beyond a deluge of studies focused on singular sites for online human interaction, we develop an infrastructure where scientists can systematically share, annotate, analyze, and integrate data from multiple online sources.

In biology, GenBank enables scientists to share, describe, and leverage data from hundreds of labs, accelerating the development of knowledge about the human genome. Like GenBank, we are building the infrastructure for social scientists, computational social scientists, and citizens to make corresponding advances in our understanding of online human interactions.

Specifically, the large volume of online behavioral data, combined with its poor description to date, creates a number of persistent research challenges that (1) limit the discovery and reuse of large datasets built from these traces, (2) hinder researchers in combining or comparing datasets, (3) fail to provide proper attribution for those creating the datasets, and (4) make the study of how scientists are creating and using datasets in scientific inquiry difficult. In short, scientists lack the tools, methods, and practices to combine, compare, contrast, and communicate about online behavior over time and across online locations.

The Particular Challenge of Multidisciplinary Research

Unlike GenBank, social computing researchers do not have a common ground that is connected to fundamental, life sciences like biology or genetics. Social computing and "big data" research draw scholars and practitioners from a myriad of different disciplines (e.g., computer science, sociology, mathematics, economics, physics, anthropology, organization science, communications). Each discipline engages in research about OOCs from its own traditions and points of view. For example, management scholars in free and open source software (FOSS) focus on developing theories of collaboration on these projects drawn from rich, qualitative methods (Howison and Crowston 2014), while software engineering scholars address developer coordination tools (Blincoe et al. 2012) and specific issues of how to make sense of electronic trace data through software repository mining (Bird et al. 2009). Human computer interaction (HCI) scholars in FOSS are particularly focused on how tools might be designed to support different modes of collaboration (Dabbish et al. 2012). The research contexts are identical, but differences in data and method prevent the development of coherent understandings across these disciplines. For these reasons, then, the work of building a data exchange will need to work across disciplines.

Spanning intellectual disciplines is potentially risky. Developing new interdisciplinary practices and methodological approaches could conflict with a discipline's current discourse and findings. However, the potential benefits of interdisciplinary OOC work are significant for both science, which gains leverage from integrated research models and the corresponding advancement in knowledge, and society, which is growing increasingly reliant upon OOCs.

How one goes about trying to span disciplines involves a) surrender of a singular, disciplinary view of the process of science and b) *actively* working across disciplines. Unlike established journals within a discipline, where editors and associate editors help to reify practice, top-down mechanisms for more systematic OOC research across disciplines are not likely to succeed. Within this chapter, then, we present a set of flexible and adaptive methods, tools, and data structures for building multidisciplinary social computing and "big social data" practice.

Open Community Data Exchange

Online, behavioral datasets must be described consistently in order to be discoverable by others, compared with each other, and studied in aggregate. Core to this proposal is advancing the Open Community Data Exchange (OCDX), a metadata specification and robust infrastructure for long-term sustainability. This project specifically builds on the prototyped capability of the OCDX, including a *bill of materials* for datasets (OCDX manifest) as derived from the OCDX metadata specification (Fig. 3.1).

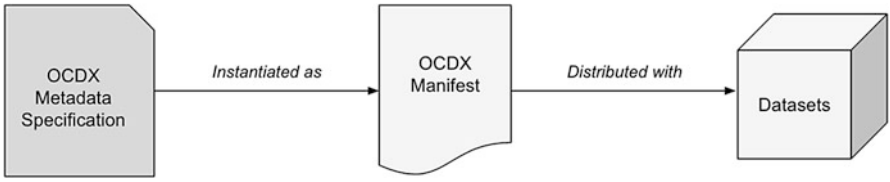


Fig. 3.1 The relationship between the OCDX metadata specification, the OCDX manifest, and datasets. The relationship is similar to the relationship between the W3C specification used to define HTML5 and the actual use of HTML5 in practice. The OCDX metadata specification contains details about metadata fields including acceptable formats and cardinality. The OCDX manifest is the instantiation of those details in practice

The precise metadata describing fundamental dataset information and recommended analytical practices are included in the OCDX manifest. The OCDX metadata standard, related OCDX manifest, and supporting OCDX cyber-infrastructure and tooling (collectively referred to as the OCDX Initiative) have been initially designed and tested by members of several scientific communities, including social science, computer science, and information systems. To date, the OCDX initiative has been evaluated and advanced through academic and practitioner workshops in Vancouver (Morgan et al. 2015), Copenhagen (May 2015), Omaha (January 2016), San Francisco (CSCW, February 2016), Chicago (May 2016), and Portland (February 2017).

Research Approach

Technology alone will not bridge the gaps identified at the outset. Advancing scientific practices, which require people, is both more complex and more critical for success. To meet this challenge, our project will use engaged scholarship as a dominant methodological approach within which more localized methods are applied (Chiasson et al. 2009), as illustrated in Fig. 3.2.

The pluralist approach provides context for our project and frames the setting within which we manage our project. It enables refined analyses and theoretical representation of community development, standards creation, and scientific practices that emerge as part of the OCDX initiative (Weick 1989). For instance, within our workshops, we may conduct surveys before and after the event and interviews at the event. But, since the tools we are deploying and using in the workshops are themselves trace data collectors, we can use that data in conjunction with the surveys and interviews to create a holistic view of the experiences and events that occurred during the workshop.

To advance the OCDX initiative, a new open online community science cyber-infrastructure is designed, deployed, and managed through four integrated tracks within we will participate in engaged scholarship and our localized methods. Each

Fig. 3.2 Engaged scholarship as the research approach within which localized research methods are applied in the proposed project

track focuses on specific work in support of this goal. Additionally, research questions for each track are geared toward helping us understanding how the OCDX initiative can both improve and learn from scientific practice in the ongoing refinement of our infrastructure.

Infrastructure Implementation Track 1 is aimed at creating a robust and sustainable infrastructure that supports OCDX manifest creation, governance, sharing, and access. In this effort, Track 1 advances analytic systems for the aggregation, visualization, and analysis of OCDX manifests and their use in scientific activities. We accomplish this through fostering our relationships and scaling our development efforts with the Wikimedia Foundation for robust information system platforms. Further, we will populate the information systems with an initial corpus of manifests by partnering with FLOSSmole (Morgan et al. 2015) to annotate their archives and with GitHub for continuous open online community data sourcing. Connecting the OCDX initiative with information organizations (Wikimedia), communal engagements (GitHub), and scientific endeavors (FLOSSmole) strengthens ties with our foundational, corporate, and academic partners, fostering diverse support for the OCDX initiative. Track 1 addresses the following questions:

(a) *How is massive online community data infrastructure understood, advanced, and fostered?*
(b) *What are the impacts of infrastructure design decisions on the sharing and analysis of online community data?*

Deep Dives Track 2 advances the integration of the ODCX infrastructure into scientific practices associated with dataset development, management, and discovery. We accomplish this by engaging with several ongoing research projects as deep-dive cases that will use the OCDX infrastructure as part of their research workflow. We will explore the ways research teams use the OCDX infrastructure in the creation of ODCX manifests. These partners include Syracuse University (political election campaigns on social media), the University of Missouri (focusing on online health support), and projects at the University of Maryland (relating online behavior to

offline actions). In addition to creating a corpus of OCDX manifests generated from different types of ongoing open online community research, efforts in Track 2 will provide feedback to improve the OCDX metadata specification and supporting infrastructure. Track 2 addresses the following questions:

(a) How do formalized architectures for online community data fit within the research workflow impact the practice of science?
(b) How can individual use cases be studied in order to gain insight to affect the development of the sharing and analysis infrastructure?

Outreach and Sustainability Track 3 is aimed at the outreach and sustainability of the OCDX initiative, requiring ongoing efforts to engage and grow the community. In Track 3, we actively connect with academic and practitioner participants through two types of OCDX-sponsored workshops recurring a total of ten times over the course of the project. The first type of workshop includes hands-on engagement with the OCDX manifest and infrastructure as participants come to understand and advance the OCDX initiative. In this workshop, participants will integrate existing datasets with OCDX manifests and infrastructure to highlight successes and concerns. The second type of workshop will include relationship building between participants through presentations of how the OCDX initiative is currently being designed, developed, and deployed. The aim of the second workshop is to highlight real-world implementations, stimulating points of common interest between participants. Both workshops are constructed with the goal of building outreach and improving sustainability of the OCDX initiative through regular and engaged community building activities. Track 3 addresses the following questions:

(a) What are key motivators for people to share their online community data and analyses?
(b) What is the impact of outreach and sustainability efforts on promoting the sharing of such data?

Science of Science Research Track 4 is primarily aimed at advancing the science of science, with a focus on data-intensive open online communities. In the fourth track, we study the scientific enterprise using OCDX manifests and infrastructure created from Tracks 1 to 3. In the science of science track, we think of the corpus of OCDX manifests as a kind of human trace data that we can study in similar ways that researchers study open online communities. We will develop analytical techniques, as well as empirical and theoretical models that leverage the OCDX manifests to help reveal the ways data-intensive open online community science takes place. We will also link our findings with other published scientific data (e.g., citations) to identify factors related to scientific productivity and impact. We will demonstrate ways that the OCDX initiative will be useful in informing scientific policy associated with the systematic sharing and analysis of datasets.

Feedback from Track 4 will be used to improve the OCDX metadata specification and infrastructure in ways to specifically support the science directly associated with the OCDX initiative. This is a sharp contrast to GenBank, which was designed to support sharing and discovery of data, but not to directly support the study of the scientific endeavor itself. Track 4 addresses the following questions:

(a) How does analysis of such data sharing initiatives reveal new scientific practice and inform science policy?
(b) What is the impact of science of science findings on online community data sharing?

The research questions in each track help us understand why and how participants engage the OCDX initiative, ways in which the OCDX metadata standard, tooling, and infrastructure are engaged and ways that scientific metadata reveals how data-intensive research takes place and becomes part of scientific practice. Figure 3.3 illustrates the four interrelated tracks.

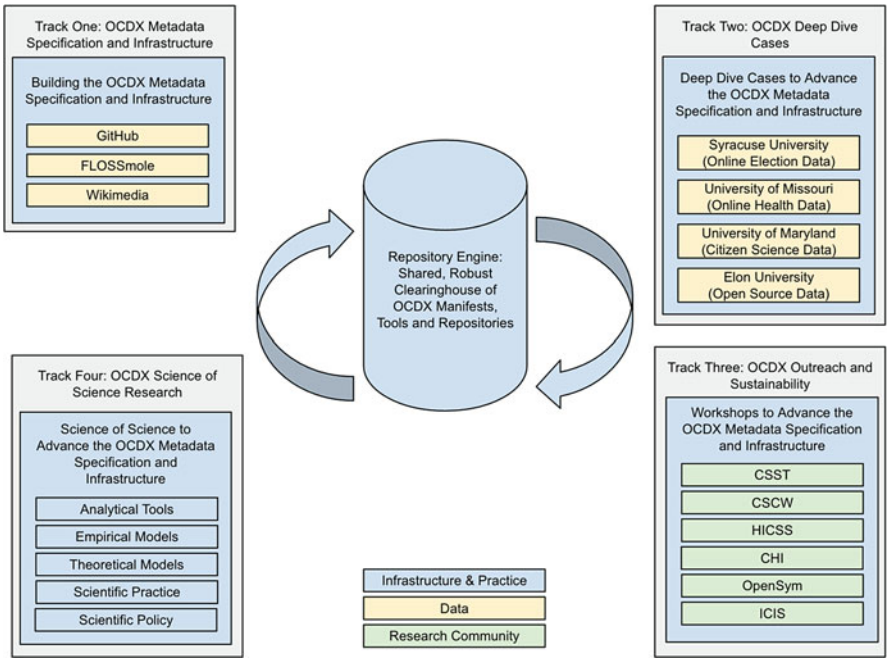


Fig. 3.3 Project tracks for facilitating the description and use of open online community data across scientific practice. Note that infrastructure and practice include all of the tools, metadata, repositories, hardware, and scientific practice that is reflexively constructed across the four tracks

Expected Outcomes

The OCDX initiative provides metadata for researchers' intent on sharing and discovering open online community data and studying the enterprise of open online science in hopes of informing scientific practice and policy. We are working on three primary artifacts: a metadata specification, tooling, and infrastructure. Each is introduced here.

A Manifest for OOC Data

The OCDX's interest in inclusion and collaboration supports the integration of multiple technical and theoretical models for dataset evaluation and documentation. Specifically, interests in balancing the development of infrastructures and technologies with the evaluation and discussion of ethical and concerns framed the need for a metadata schema capable of capturing and reflecting the value and complexity of OOC datasets. The OCDX metadata specification builds on an existing data curation lifecycle model and prior efforts to standardize open source project metadata, including the Linux Foundation's Software Package Data Exchange (SPDX).

Building on the Data Curation Lifecycle Model

The OCDX workflow uses the DCC Curation Lifecycle Model (Fig. 3.4) to guide a series of iterative tasks that support the identification of actors, actions, and technologies that contribute to the collection, creation, and maintenance of records. Based on these iterative tasks, it has been possible to foster discussion and collaboration about characteristics of OOC datasets while simultaneously evaluating methods and technologies for replicating scholarship that uses them.

The lifecycle model contains six total rings, while each requires a series of action and contributions. Establishing iterative tasks maximizes participant input on the quality and accuracy of specific metadata fields and/or entire metadata records. Using a data curation lifecycle that promotes flexible and ongoing data management has made possible to integrate multiple points of view into the standards for metadata creation and the interface users interact with while creating records for datasets. To augment the coherence of the curation practices expressed in the lifecycle, an additional metadata workflow model was adopted (Fig. 3.5).

Identifying areas of overlap in metadata creation and/or revision practices created additional space for conversations among researchers, which has provided opportunities for different areas of expertise and interest to take priority, but not at the cost of overarching interdisciplinary needs. In general, there are two different goals of establishing a workflow: first, establishing where to collect information about datasets from and, second, outlining best practice guidelines for metadata creation, revision, and maintenance.

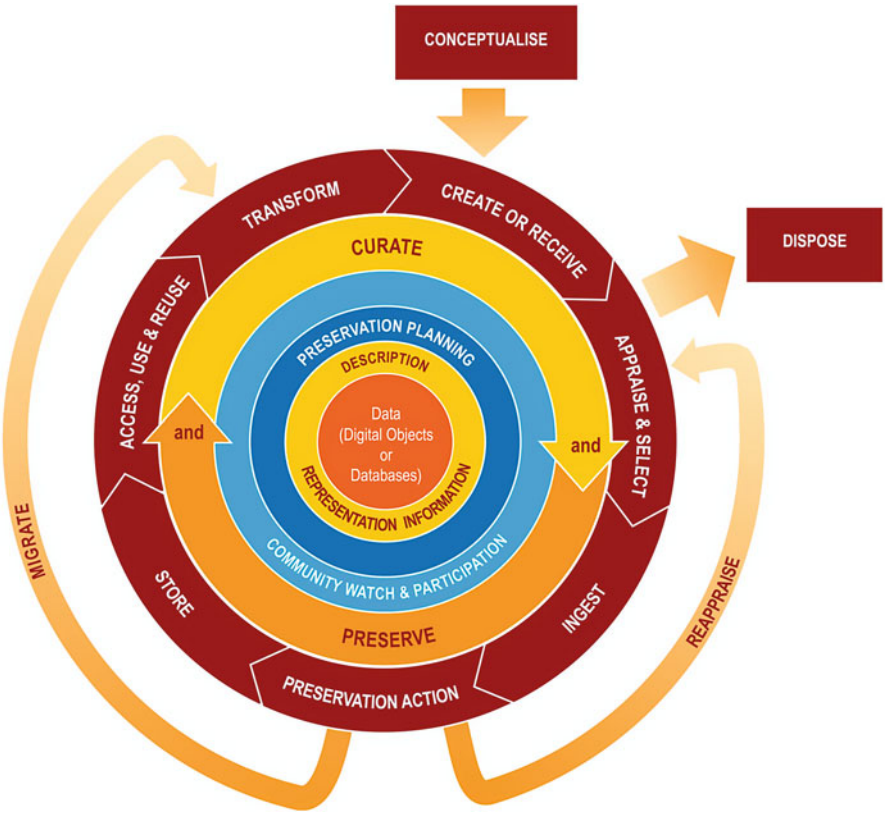


Fig. 3.4 DCC Curation Lifecycle Model (Jisc n.d)

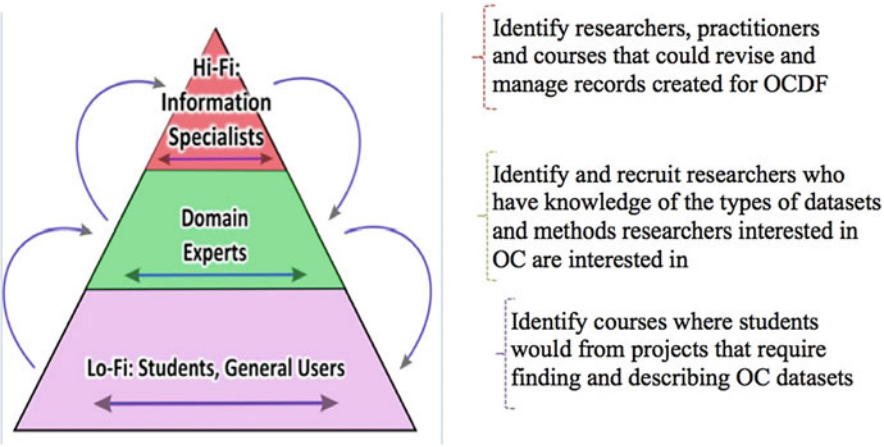


Fig. 3.5 Metadata evaluation and creation workflow (Maron et al. 2015)

Taken together the DCC Curation Lifecycle Model and metadata workflow established standards for metadata entry and revision that synthesize research interests and needs while simultaneously recognizing the need for flexibility and diversity in some fields based on disciplinary and technological practices. Developing metadata standards that reflect the interests of researchers from a variety of academic disciplines will enhance the variety and quality of contributions to the OCDX. Keeping these interests and goals in mind led to the development of a manifest, which consists of two documents: a metadata schema and documentation outlining how to use the schema.

The manifest consists of two documents: a schema and documentation outlining how to use the schema. The schema contains four general descriptive categories (agent, description of dataset, description of data source, metadata creation) and within these four general descriptive areas a series of refinements that facilitate more specific descriptions of documentation, processing, and accessibility of a dataset. Documentation outlining how to use the metadata schema provides structure and guidance on the function of each field. Offering guidelines on implementation supports consistent description of OOC datasets, which furthers the OCDX's goal to support interdisciplinary scholarship.

Establishing the structure and purpose of the manifest will support further discussions of interests, needs (technical and practical), and resources needed to perform interdisciplinary research on OOCs. Additionally, building a manifest that reflects interests and needs of scholars participating in the OCDX will make it possible to describe how metadata can enhance projects undertaken by other working groups contributing to OCDX and, thus, provide opportunities for enhancing the infrastructure and tools available for recording information about datasets. Also, it will promote further opportunities for courses and research relating to the exploration and analysis of open online communities.

Building on Open Source Metadata Specifications

The OCDX metadata specification is used to define metadata manifests to accompany partner datasets. Metadata specifications have proven valuable in bridging and connecting community members aiming to share information in the overall advancement of community health and sustainability. The Software Package Data Exchange (SPDX) community is a Linux Foundation initiative aimed at explicating license and vulnerability metadata for software packages as exchanged throughout software supply chains (Germonprez et al. 2014).

The OCDX metadata specification represents a key artifact from which tooling and infrastructure are derived. It is expected that through these relationships, the OCDX metadata specification will be better understood in practice, leading to its refinement to potentially include such fields as author annotations, dataset dependencies, and dataset lifecycles. A condensed form of the current OCDX metadata specification is shown in Fig. 3.6.

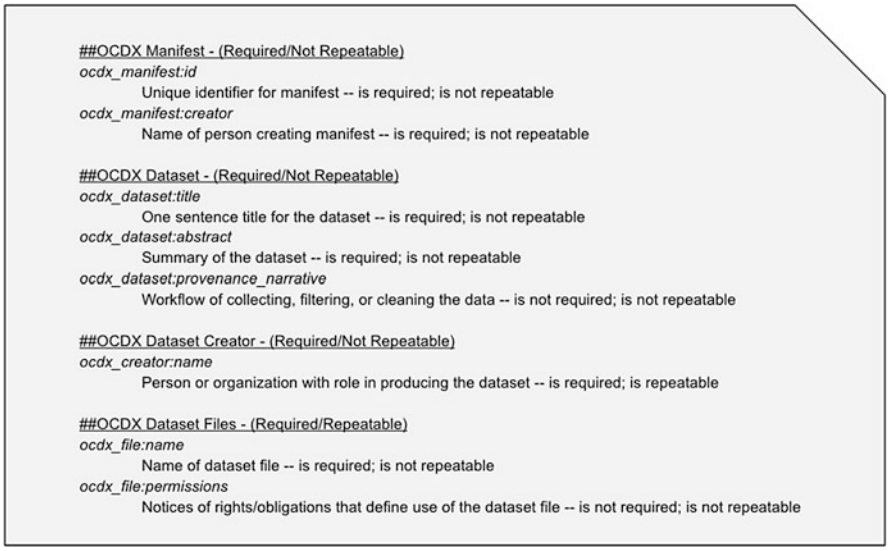


Fig. 3.6 The OCDX metadata specification in condensed form

The advancement of the OCDX metadata specification alone will move us a considerable way toward the goal of making data more reusable by a larger group of scientists. Making sure that a large percentage of open online community datasets have explicit OCDX manifests attached to them that describe what the dataset is, how it was collected, and what permissions are provided for reuse of the dataset will make it much easier for scientists to identify datasets of interest to them, to understand datasets that were used in other contexts, and to use those datasets in their own work. Moreover, this would create labeled and related datasets demonstrating community activity, and such a set of related datasets becomes an object of study in its own right. Technology that enables the easy use of related OCDX manifests will make this work much more powerful, which is what we will describe in the next section.

Tooling and Infrastructure

Stemming from the metadata specification, we are advancing robust tooling through participant engaged design, development, and deployment activities. These activities involve our foundational, academic, and corporate partners. Foundationally, we are partnered with the Wikimedia Foundation to integrate OCDX tooling with existing toolkits including JupyterHub and Wikibase. Academically, we are partnered with open online community researchers to provide OCDX tooling aimed

at advancing and understanding scientific practice. Corporately, we are partnered with GitHub to integrate OCDX tooling with continuously sourced community metric data. OCDX tooling includes support for the generation, management, and consumption of OCDX metadata standard-derived manifests.

We propose to design and build an infrastructure and toolset that enables the sharing of electronic trace data from a wide range of systems, including open online community systems, in such a way that the content, structure, and associated analysis tooling for each dataset are explicitly noted in an instance of the OCDX manifest. *The proposed manifest will advance the present one by describing the entire research ecosystem around an online behavioral dataset.* Advancing this technical goal makes the analysis of similar online environments and the identification of similar analytical strategies practical and possible for the first time.

OCDX infrastructure is aimed at supporting services by which OCDX metadata standard-based tooling is made publically available for scientific communities. The OCDX infrastructure will support public instances of all OCDX tools by which OCDX manifests are produced, managed, and discovered. Finally, the OCDX infrastructure will be available for local deployments via full source, install scripts, and documentation provided through our GitHub repository.

Conclusion and Future Work

Through participant engaged design, development, and deployment, we consider the OCDX initiative as an evolving endeavor where points of interest are identified in ways that the metadata standard, tooling, and infrastructure are used, adapted, and validated. In this chapter, we outlined the background and goal of this OCDX project and described our methods and outcomes. We believe big social data research will benefit from this work, but the path will surely not be linear.

One source of nonlinearity in this work is the substantial diversity in scholarly approaches. We covered that at the outset. Another challenge emerged through our work. As it turns out, there is a continuum of structured data generated and required by different OOC's and researchers who approach them. That continuum goes from "Must have structured, specified data" to "Hey, our data is conversations and there is little structure around those." Most of the lessons and candidate approaches we present work across both the structured and unstructured OOC datasets.

Some of our work extends from work on building manifest descriptions for the Linux Kernel and GenBank. The Linux Kernel's OSS development is most similar to our work. The intrinsic need for stability is a characteristic it shares with GenBank. In the case of GenBank, what is negotiated are representations of experimentally and computationally defined abstractions of biology and genetics. In the case of the Linux Kernel, we continue to refine the specification of the OCDX metadata standard as well as the tooling and infrastructure required for open

online community scientists to find, understand, create, maintain, and share dataset metadata. We invite scientists who have datasets to compile an OCDX manifest and provide feedback.

References

Bird, C., Rigby, P.C., Barr, E.T., Hamilton, D.J., German, D.M., Devanbu, P. (2009). The promises and perils of mining git. Proceedings from Mining Software Repositories, 2009. MSR'09. 6th IEEE International Working Conference on Mining Software Repositories.

Bishop, J.L. Verleger, M.A. (2013). The flipped classroom: A survey of the research. ASEE National Conference Proceedings, Atlanta.

Blincoe, K., Valetto, G., Goggins, S. (2012). Leveraging task contexts for managing developers' coordination. Proceedings from ACM Conference on Computer Supported Cooperative Work, 2012, Seattle.

Chiasson, M., Germonprez, M., & Mathiassen, L. (2009). Pluralist action research: A review of the information systems literature. *Information Systems Journal, 19*(1. (Jan. 2009), 31–54.

Dabbish, L., Stuart, C., Tsay, J., Herbsleb, J. (2012). Social coding in Github: Transparency and collaboration in an open software repository. Proceedings from CSCW'12, Seattle, Washington.

Germonprez, M., Kendall, J. E., Kendall, K. E., & Young, B. (2014). Collectivism, creativity, competition, and control in open source software development: Reflections on the emergent governance of the SPDXtextregistered working group. *International Journal of Information Systems and Management, 1*(1/2. (2014), 125–145.

Goggins, S. P., Mascaro, C., & Valetto, G. (2013). Group informatics: A methodological approach and ontology for sociotechnical group research. *Journal of the American Society for Information Science and Technology, 64*(3. (Mar. 2013), 516–539.

Howison, J., & Crowston, K. (2014). Collaboration through open superposition: A theory of the open source way. *MIS Quarterly, 38*(1).

Irwin, A. (1995). *Citizen science: A study of people, expertise and sustainable development*. New York: Psychology Press.

Jisc. (n.d.). DCC curation lifecycle model. Retrieved from: http://www.dcc.ac.uk/resources/curation-lifecycle-model.

Maron, D., Missen, C., McNeirney, K., Elnora, K.T. (2015). Lo-fi to hi-fi crowd cataloging: Increasing e-resource records and promoting metadata literacy within WiderNet. Poster presented at the iConference.

Moorhead, S. A., Hazlett, D. E., Harrison, L., Carroll, J. K., Irwin, A., & Hoving, C. (2013). A new dimension of health care: Systematic review of the uses, benefits, and limitations of social media for health communication. *Journal of Medical Internet Research, 15*(4. (Apr. 2013), e85.

Morgan, J.T., Halfaker, A., Taraborelli, D., Goggins, S., Hwang, T., Computing, S. (2015). Bridging the data divide. (2015).

Nahon, K., & Hemsley, J. (2014). Homophily in the guise of cross-linking: Political blogs and content. *American Behavioral Scientist, 58*(10. (Sep. 2014), 1294–1313.

Ren, Y., Kraut, R., & Kiesler, S. (2007). Applying common identity and bond theory to design of online communities. *Organization Studies, 28*(3. (Mar. 2007), 377–408.

Tandoc, E.C. (2014). Journalism is twerking? How web analytics is changing the process of gatekeeping. *New Media & Society*. (Apr. 2014), 1–17.

Weick, K. E. (1989). Theory construction as disciplined imagination. *The Academy of Management Review, 14*(4. (1989), 516–531.

Part II
Theoretical Principles and Ideas for Designing and Deploying Data Factory Approaches

Chapter 4
Levels of Trace Data for Social and Behavioural Science Research

Kevin Crowston

Introduction

The social and behavioural sciences are said to be on the verge of a data-driven revolution. There is great interest in the scientific inferences that can be drawn from digitally captured records of human activity, such as in an online community, user-generated content systems, search engine searches, cellular phones or digital badges (Lazer et al. 2009; Manovich 2012), what Howison et al. (2011) call trace data. As Agarwal et al. (2008) stated: "Most transactions and conversations in these online groups leave a digital trace ... this research data makes visible social processes that are much more difficult to study in conventional organizational settings". For example, researchers have noted that social media data show great potential to address long-standing research questions about human behaviour (Edwards et al. 2013). Chang et al. (2014) go as far as to suggest that the rise of big data is leading to a "paradigm shift in scientific research methods", what Watts (2007) called a "21st century science".

However, these claims about the transformative capacity of big data for the social and behavioural sciences need to be viewed with caution. Records of online behaviour certainly amount to terabytes of data, but these data are of a very different sort than social and behavioural scientists would obtain from more traditional research approaches such as surveys or experiments and so require different research approaches. The most closely related commonly used data are events data in international relations (e.g. McClelland 1967), and consideration of the issues in using these data provides some insights.

K. Crowston (✉)
Syracuse University School of Information Studies, Hinds Hall 348, Syracuse, NY 13244, USA
e-mail: crowston@syr.edu

© Springer International Publishing AG 2017

S.A. Matei et al. (eds.), *Big Data Factories*, Computational Social Sciences,
https://doi.org/10.1007/978-3-319-59186-5_4

The goal of this chapter is to discuss differences between trace data and traditional social and behavioural science data and the implications of these differences for using trace data for social and behavioural science research. The main contribution of the paper is a more precise vocabulary for talking about the processes of using trace data and the products of these processes that clarify different levels of processing. The framework also highlights issues involved in sharing and reusing trace data.

Framework: From Trace to Variable

Howison et al. (2011) identify three differences between long-used sorts of social and behavioural research data and trace data: trace data are event-based, longitudinal and, most importantly, found, rather than created to support research. These features are found in other settings, e.g. political scientists have built databases of events data (e.g. the World Event/Interaction Survey, WEIS (McClelland 1967)), and longitudinal data are common across many fields.

The difference that is key for our argument is the final point. Data from scientific sources such as surveys or experimental measurements are most often purposefully collected to measure constructs of theoretical interest. Rigorous quantitative research employs carefully refined instruments with known psychometric properties to ensure that the instrument reliably measures what it should. (Poorly designed research might be sloppier, but is hard to argue as a model for future research.) In contrast, social media and other trace data are records of human activity without inherent theoretical import. As Howison et al. (2011) say, "Wikipedia was not designed to test theories about knowledge production, nor are corporate email systems designed to collect research data". Rather, these data need to be interpreted to be useful for social and behavioural scientists.

In some ways, the interpretive flexibility of trace data is an advantage. They reflect actual behaviour rather than opinion, belief or attitude and can be used for different kinds of studies, unlike data from most surveys or experiments that measure specific constructs. The implication though is that trace data require considerable additional processing to be useful for research. Unfortunately, the term "data" is overloaded and does not distinguish between different kinds of data, processed or not, leading to potential confusion and unwarranted optimism about the utility of found data. A framework is needed to sort out the different kinds of data. The main contribution of this chapter is to develop such a framework.

Levels of Data in the Earth Sciences

This situation—having multiple kinds of data with different levels of scientific interpretation—is by no means unique to the social and behavioural sciences or

Table 4.1 Levels of Earth observation data

Level	Definition
Level 0	Reconstructed, unprocessed instrument/payload data at full resolution; any and all communications artefacts, e.g. synchronization frames, communications headers, duplicate data removed
Level 1A	Reconstructed, unprocessed instrument data at full resolution, time referenced, and annotated with ancillary information, including radiometric and geometric calibration coefficients and georeferencing parameters, e.g. platform ephemeris, computed and appended but not applied to the Level 0 data
Level 1B	Level 1A data that have been processed to sensor units
Level 2	Derived geophysical variables at the same resolution and location as the Level 1 source data
Level 3	Variables mapped on uniform space-time grids, usually with some completeness and consistency
Level 4	Model output or results from analyses of lower-level data, e.g. variables derived from multiple measurements

From raw data as collected to processed and synthesized data Parkinson et al. 2006

to trace data. It is thus instructive to examine how the distinctions among data with different kinds of processing are addressed in other disciplines. The earth sciences provide a particularly helpful framework, as the kinds of data created by processing satellite observations have been given different labels with clear definitions in this research community.

The NASA Earth observation program distinguishes data at six levels of processing, as shown in Table 4.1 (from Parkinson et al. 2006). Data at each level is derived from the data at the lower level through defined data-processing steps. For example, consider a satellite collecting data about the Earth using a sensor that receives some signal from the Earth (e.g. light or radar reflections) that can be interpreted as evidence for a geophysical variable (e.g. temperature or sea wave heights). To move from Level 1 to Level 2 data in the framework, for example, data from the sensor are interpreted to reveal geophysical variables, e.g. certain wavelengths of light indicate particular kinds of vegetation; particular scattering of radar indicates wave heights. In the earth sciences, Level 0 and 1 data are generally not useful for research, other than for studies of the properties of the satellite and its sensors. Instead, earth scientists want Level 2 or 3 data, data about a geophysical process, plotted on a map. That is, rather than a time series of voltages from a sensor, scientists want a map showing what vegetation is where (for example).

Example: From Tweet to Variable

We can apply the Earth observation data framework to the case of trace data. We use as an example data from the social media platform, Twitter. By analogy to Table 4.1, we define different levels for Twitter data, as shown in Table 4.2. Level 0 are the raw

Table 4.2 Levels of twitter data

Level	Definition
Level 0	Raw tweets
Level 1	Raw tweets annotated with ancillary information, e.g. sender information
Level 2	Derived social and behavioural science variables at the same resolution as Level 1 (i.e. coded tweets)
Level 3	Derived social and behavioural science variables at unit of analysis of interest (e.g. data about individuals)
Level 4	Model output or results from analysis that merges multiple sources of data

tweets, e.g. collected from a Twitter API. Level 1 adds metadata about the tweets as they were collected (e.g. time, date, sender). Level 2 interprets the tweet content as indicating some social and behavioural science variable of interest (e.g. political discourse, topic or sentiment). Level 3 aggregates evidence from multiple tweets to develop data about the unit of analysis of interest for the study: an individual, a political figure, a topic, etc. Note that our interpretation of this level for trace data differs somewhat from the definition of Level 3 in the original Earth observation framework, which refers to mapping data to a uniform space-time grid. Here we generalize that concept to mapping data to other conceptual spaces. Finally, Level 4 is created by linking data from the tweet corpus to data from other datasets or to a model.

The same distinctions can be made for other kinds of trace data. For example, a study about leadership in an open source project (Crowston et al. 2010) might draw on developer emails (Level 0) (see Table 4.3), annotated with information about the sender (e.g. the role in the project, Level 1), coded for evidence of leadership behaviours (Level 2), aggregated to suggest which members of the project exhibit signs of being project leaders (Level 3) and linked to other data about contributions or project outcomes (Level 4).

As with satellite data, for social and behavioural science research, Level 0 or 1 social media data are unlikely to be of much interest for research: raw tweets or email messages by themselves and "as is" are not that useful for research. However, it is at this level that we see the explosion in available data. To test theory, social science researchers need data at Level 3, which corresponds to the kind of data a researcher would get from a survey. Unfortunately, such data are much less readily available. An implication for development of data archives is that it would likely be more useful to focus these on higher levels of data.

Discussion: Moving Up the Levels

The issue then is how to process data to move from Level 0 to Level 3 or 4. For geospatial data, scientists have developed data-processing algorithms based on their knowledge of the physical properties of the satellites and sensors and the

Table 4.3 Levels of open source development data

Level	Definition
Level 0	Raw email messages
Level 1	Raw email messages annotated with ancillary information, e.g. sender information
Level 2	Derived social and behavioural science variables at the same resolution as Level 1 (i.e. coded email messages)
Level 3	Derived social and behavioural science variables at unit of analysis of interest (e.g. data about individuals)
Level 4	Model output or results from analysis that merges multiple sources of data

geophysical properties of the systems being observed: e.g. known performance of instruments converting radiation to a sensor signal, mathematical models for translating between satellite position and orientation to the observed location on the ground or models of what different vegetations look like to support inference from an observed intensity of light at a particular wavelength to geophysical data about ground cover. Even with this level of theoretical development and knowledge of the geophysical processes, automated algorithms are not always sufficient by themselves. For example, for best precision, images might have to be adjusted by hand by manually matching known benchmarks on the base map. Predicted geophysical variables (e.g. vegetation) might need to be ground-truthed to verify the reliability of the interpretation.

Interpretation of data is also a common analysis approach in social research. Qualitative researchers frequently employ the technique of content analysis (Krippendorff 2004) to code textual documents for theoretical constructs of interest. In the framework above, content analysis is a technique to move data from Level 0 or 1 to Level 2. The political science databases described above take newspaper or wire series press reports as Level 0 data and code them against an event coding scheme that identifies actors and actions of theoretical interest (Veen 2008). For example, WEIS's (McClelland 1967) coding scheme codes events reported by *The New York Times* into 61 categories of action. Researchers employing observational techniques develop coding schemes that identify which observed behaviours are of interest, essentially skipping Levels 0 and 1 and collecting data directly at Level 2. Considering social media again, tweets might be interpreted as indicating support for or opposition to a political candidate.

Unfortunately, moving up levels of social media and other social and behavioural trace data is less routinized and predictable than for Earth observation data and even for international relations. Some of these problems are inherent in the nature of the social and behavioural sciences. The processes by which the social and behavioural constructs of interest (e.g. leadership) get reflected in recorded behaviours (e.g. emails) are much less regular than the corresponding geophysical processes (e.g. vegetation reflecting light). But there are also differences that reflect the rigour and reproducibility of the data processing in research practice. At present, social and behavioural researchers typically derive variables from observed data in their own idiosyncratic ways. As with satellite data, processing may require

manual intervention and validation, making the process hard to replicate or even to completely describe. And unfortunately, provenance of data is often not well recorded, so how these steps were carried out may be unclear to those reading the research. For example, Liang and Fu (2015) found that they could not reproduce the results of six out of ten studies of Twitter they examined using a random sample of tweets, which they attributed to "variations of data collection, analytic strategies employed, and inconsistent measurements".

We next discuss the specific issues involved in each step of the chain from event to Level 4 data to further explore the issues involved in using trace data for social and behavioural research.

Collecting Level 0 Data

Level 0 is the lowest level in the framework, but it is worth noting that even Level 0 data has had some processing. As noted in Table 4.1 above, satellite data is processed to remove communications artefacts. For trace data or social media data, there is a comparable process of removing artefacts from the data collection that needs to be documented (e.g. removing spam emails from an email archive before analysis). However, additional problems can arise. Howison et al. (2011) point out that collecting trace data from an information system raises a number of validity issues. They focus on validity issues for social network analysis, but a number of their issues are more general. Two relate in particular to the collection of trace data from an information system, that is, the creation of what we are labelling Level 0 data: "system and practice issues" and "reliability issues".

The first issue refers to the need to understand actual system use in order to be able to interpret the data created. An example given by Howison et al. (2011) is a group-support system that requires individuals be team "members" to access team documents, leading to many people being listed as members mostly to enable document access. The point is that the system definition of a team member in this case is different than the offline definition, posing challenges for interpreting the system data about membership.

The second issue refers to the need to assure that the data have been collected reliably by the system itself. As system databases are maintained for the operation of the system rather than for scientific purposes, decisions about data collection are usually made for operational reasons rather than to preserve the scientific integrity of the data (e.g. system databases might be periodically purged of old data for performance reasons). However, those decisions and their consequences are unlikely to be visible to an external researcher. In political science, similar concerns are raised about biases in news sources' selection of events to report and, indeed, whether certain relevant events are reported at all (McClelland 1983). Boyd and Crawford (2012) note that most Twitter APIs yield a subset of tweets, but it is not clear how that subset is selected, making the generalizability of the sample questionable.

Data Processing from Level 0 to 1

To move from Level 0 to Level 1, data are annotated with additional information about the observations that were made. The issues here for trace data parallel those for collecting Level 0 data, namely, ensuring the completeness and reliability of the data collected. As an example, email messages include a timestamp (a kind of metadata for the email observation), but may omit the time zone, making the interpretation of the timing of messages problematic (Howison et al. 2011).

Data Processing from Level 1 to 2

For Earth observation satellite data, data at Level 2 are the results of interpreting satellite sensor data as geophysical variables. Such an interpretation is inherently theoretically based. For example, to interpret light reflected from the Earth as evidence of vegetation requires a good model (possibly empirical rather than strictly theoretical) of how different kinds of vegetation reflect light under varying conditions.

In the world of trace data, traces need to be interpreted to serve as evidence for social and behavioural concepts of interest. For example, political science events databases arise from human or machine coding of events reported in news stories. As Veen (2008) notes, each "event scheme is informed by theoretical assumptions about the international system and the interaction of political actors". However, Venn notes that researchers often want to analyse variables such as the level of cooperation or conflict between two countries, which requires further interpreting the events as evidence for these constructs.

Returning to social media data, to create Level 2 Twitter data, raw tweets can be content analysed for any number of social and behavioural science concepts, e.g. for what topic a tweet addresses or what speech act the tweet represents (Hemphill and Roback 2014). Again, such interpretation relies on a theory about the concept in question and how it affects or is reflected in the observed behaviour. As Manovich (2012) notes, online behaviour is "not a transparent window into peoples' imaginations, intentions, motifs, opinions, and ideas" and thus needs to be carefully and thoughtfully interpreted.

In some cases, interpretation is sufficiently well understood and mechanical that it can be done automatically. For example, natural language processing techniques have been developed to determine the sentiment of a text, albeit with some imprecision. Recent political science events databases are generated by automatic coding of wire service articles (Veen 2008). In other cases, human judgement might be needed, which can pose a significant bottleneck for processing as well as potentially adding individual human errors or biases to the judgements. These issues have led to the development of citizen science projects that have multiple human volunteers assess images or other data. In many cases though, this processing

is more akin to processing Level 1B in the original Earth observation framework. For example, annotating an image of a galaxy for its shape (as in the original Galaxy Zoo) or an image of an animal with the species (as in Snapshot Serengeti) provides useful information, but the data are still about the image with limited theoretical import. (This argument might also be made about political science events.)

Unfortunately, in many cases, analyses of trace data essentially skip this processing step: data instead remain at the level of the original phenomenon. For example, a social network can be constructed from email messages by interpreting replies to a message as creating links. While this process does yield a network, the theoretical import of such a network is unclear. At best, a reply suggests that the person replying read and was interested in the message, but many others likely also read the message without feeling a need to reply. Similarly, data from digital badges can identify how people move through space or who they have been close to, but without some theory about movement or propinquity, it is hard to interpret the data as evidence for research. Even when an interpretation is made, it may not be theoretically justified. In a study of published communications and social computing studies of hyperlinks, Twitter followers and retweets (three kinds of trace data), Freelon (2014) found that "substantial proportions of articles from both disciplines failed to justify the social implications they imputed to trace data", "more extensively in the latter" discipline (social computing).

Data Processing from Level 2 to 3

Level 3 data require aggregating data from Level 2. To aggregate the data requires picking a unit of analysis and linking related observations. An obvious unit for aggregation for trace data of behaviours is the person involved in the recorded activity. For example, political science events are coded for the actor and recipient of an event to permit such aggregation. For social media data, one might link submitted tweets by the user ID. However, just as an event may not have an identified actor (e.g. an anonymous terrorist attack), in some settings users may have an option to work anonymously (Panciera et al. 2010), which means that a user ID might not capture all work done by a person. In particular, it may omit work done while lurking in the early stages of involvement with a group (e.g. reading others' posts), creating problems for studies of new members in particular. Data might also be aggregated to a population, e.g. to determine the average properties of particularly kinds of contributors. Interpreting such aggregated data requires more attention to the nature of the sample. As a specific example, Boyd and Crawford (2012) note that "it is an error to assume 'people' and 'Twitter users' are synonymous: they are a very particular sub-set".

Data Processing from Level 3 to 4

Level 4 data are derived from the composition of different datasets. Unfortunately, such composition is difficult for trace data and for social and behavioural science data more generally. The problem is that to connect datasets, there needs to be a way to link the data. In database terms, there needs to be common field on which to join the data tables. More simply, the different data need to be about the same thing.

For geospatial data from a satellite, data are typically tied to a particular spot on the Earth. There are difficulties in working out which spot a sensor has measured and aligning data collected in different patterns or at different resolutions, but once these issues are addressed, then collected data can be connected to other data about that spot, no matter how it was collected. The same principle also applies in astronomy: data about the same spot in the sky can be connected.

Alternately, data may be about a specific entity with a stable identity, allowing linking. For example, astronomical data can be thought of as being about particular celestial objects (stars or galaxies) that can be linked from dataset to dataset. Finally, data may be about an identifiable class of object. For example, ecological data might be about particular species and so of interest to others who study the same or similar species. Astronomical data can be about a particular type of star.

In the social sciences, datasets may sometimes be about identifiable entities, allowing linking of datasets. In particular, economic data are often about countries or companies, which makes it possible to link data about the same countries or companies (though even here there can be issues in making connections). This situation may also describe data collected about entire online communities: different perspectives on the various language Wikipedia projects can be compared.

For behavioural research though, data are likely to be about people. As with aggregating data to Level 3, it is possible to link data from a system for a particular user by using the user's system ID. However, such an ID likely has little meaning beyond the system. We might therefore be able to link a user across multiple Twitter databases, but not Twitter and anything else, meaning that we might not know anything more about users of a social media site than what they post. For the specific case of free/libre open source software developers, Crowston et al. (2006) argued that developers are often attached to the user IDs and so attempt to use them on different sites, but it is not clear that this phenomenon generalizes. Without knowing the identity of the specific respondents, it is not possible to link individual responses to other data. At most, data can be cumulated with other data to increase the sample size, as in a meta-analysis, deepening the analysis but not broadening it.

Conclusion: Recommendations for Future Research

The framework presented here reinforces several recommendations that have already been made about social and behavioural research. First, there are clear implications for reporting research. Specifically, research using trace or social media

data needs to provide more detail on the processing that took data from level to level. It would also be valuable to share techniques for moving between levels to promote reproducibility of research and to allow researchers to leverage each other's findings.

There are further implications for sharing data. Researchers sometimes face limitations on sharing Level 0 or Level 1 data. For example, the terms of service of some social media sites limit sharing such raw data. Data from proprietary services may simply be unavailable outside the organizations that run them (Lazer et al. 2009). It is worth noting that there are serious problems for the reproducibility of science if the datasets underlying studies can't be shared, meaning that other researchers are unable to check or reproduce findings. On the other hand, Level 2 or 3 data may not be so encumbered, and these are the levels that are likely to be of the most interest to other researchers. Coupled with a sufficiently detailed description of the data processing used to create the data, a Level 2 or 3 dataset may be sufficient, at least for checking results.

However, the discussion of creating Level 4 data suggests that even Level 2 or 3 data may be difficult for others to link to their own data. To be useful, the data set needs some ID on which to link the data. But if researchers know the identity of users, it is likely that they will not be able tell anyone else in order to maintain the privacy of participants (Daries et al. 2014), i.e., the ID will not be available. A possible direction for research is to apply the notion of a species as an entity for data collection. If researchers using trace data could agree on clearly defined classes of users of interest, then data might be shareable and reusable when aggregated at that level.

In summary, it is unarguable that the increased penetration of information technology across the spectrum of life activities is creating a vast trove of trace data and that such data can be of great interest to social and behavioural scientists. However, such trace data are different in kind from data more traditionally used in social and behavioural research. Applying a framework from Earth observation studies, we have shown how raw trace data must be processed to create data useful for advancing social and behavioural studies and identified the issues that arise. A particularly problematic issue is identifying what data are about in order to be able to link across datasets.

References

Agarwal, R., Gupta, A. K., & Kraut, R. (2008). Editorial overview: The interplay between digital and social networks. *Information Systems Research, 19*(3), 243–252. https://doi.org/10.1287/isre.1080.0200.

Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society, 15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878.

Chang, R. M., Kauffman, R. J., & Kwon, Y. (2014). Understanding the paradigm shift to computational social science in the presence of big data. *Decision Support Systems, 63*, 67–80. https://doi.org/10.1016/j.dss.2013.08.008.

Crowston, K., Wei, K., Li, Q., Howison, J. (2006). Core and periphery in free/libre and open source software team communications. In *Proceedings of Hawai'i International Conference on System System (HICSS-39)*. Kaua'i.

Crowston, K., Wiggins, A., Howison, J. (2010). Analyzing leadership dynamics in distributed group communication. In *Proceedings of Hawaii International Conference on System Sciences (HICSS-43)*. Lihue. doi:https://doi.org/10.1109/HICSS.2010.62.

Daries, J. P., Reich, J., Waldo, J., Young, E. M., Whittinghill, J., Ho, A. D., Seaton, D. T., & Chuang, I. (2014). Privacy, anonymity, and big data in the social sciences. *Communications of the ACM, 57*(9), 56–63. https://doi.org/10.1145/2643132.

Edwards, A., Housley, W., Williams, M., Sloan, L., & Williams, M. (2013). Digital social research, social media and the sociological imagination: Surrogacy, augmentation and re-orientation. *International Journal of Social Research Methodology, 16*(3), 245–260. https://doi.org/10.1080/13645579.2013.774185.

Freelon, D. (2014). On the interpretation of digital trace data in communication and social computing research. *Journal of Broadcasting & Electronic Media, 58*(1), 59–75. https://doi.org/10.1080/08838151.2013.875018.

Hemphill, L., & Roback, A. J. (2014). Tweet acts: How constituents lobby congress via Twitter. In *Proceedings of ACM conference on computer supported cooperative work & social computing* (pp. 1200–1210). Baltimore.

Howison, J., Wiggins, A., & Crowston, K. (2011). Validity issues in the use of social network analysis for the study of online communities. *Journal of the Association for Information Systems, 12*(12), 323–346.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Newbury Park: Sage.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Van Alstyne, M. (2009). Life in the network: The coming age of computational social science. *Science, 323*(5915), 721–723. https://doi.org/10.1126/science.1167742.

Liang, H., & Fu, K.-W. (2015). Testing propositions derived from twitter studies: Generalization and replication in computational social science. *PloS One, 10*(8), e0134270. https://doi.org/10.1371/ journal.pone.0134270.

Manovich, L. (2012). Trending: The promises and the challenges of big social data. In M. K. Gold (Ed.), *Debates in the digital humanities* (pp. 460–475). Minneapolis: University of Minnesota Press.

McClelland, C. A. (1967). *Event interaction analysis in the setting of quantitative international relations research*. Los Angeles: Department of Political Science, University of Southern California.

McClelland, C. A. (1983). Let the user beware. *International Studies Quarterly, 27*(2), 169–177. https://doi.org/10.2307/2600544.

Panciera, K., Priedhorsky, R., Erickson, T., Terveen, L. (2010). Lurking? Cyclopaths? A quantitative lifecyle analysis of user behavior in a geowiki. In *Proceedings of ACM conference on Computer-Human Interaction (CHI)*. Atlanta.

Parkinson, C. L., Ward, A., & King, M. D. (Eds.). (2006). *Earth science reference handbook: A guide to NASA's earth science program and earth observing satellite missions*. Washington, DC: National Aeronautics and Space Administration. Available from: http://eospso.gsfc.nasa.gov/sites/default/files/publications/2006ReferenceHandbook.pdf.

Veen, T. (2008). Event data: A method for analysing political behaviour in the EU. In *Proceedings of prepared for the fourth Pan-European conference on EU Politics, Riga, Latvia*. Available from: http://www.jhubc.it/ecpr-riga/virtualpaperroom/002.pdf.

Watts, D. J. (2007). A twenty-first century science. *Nature, 445*(7127), 489–489. https://doi.org/10.1038/445489a.

Chapter 5
The Ten Adoption Drivers of Open Source Software That Enables e-Research in Data Factories for Open Innovations

Kerk F. Kee

Introduction

According to the Oxford dictionary online, a factory is "[a] building or group of buildings where goods are manufactured or assembled chiefly by machine." To use the word "factory" in conjunction with "data," one can interpret the idea of "data factory" as a virtual arrangement or group of arrangements where big data sets are produced, aggregated, recombined, and/or repurposed mainly by cyberinfrastructure. The meta-platform of cyberinfrastructure includes open source software, visualization systems, remote instruments, distributed sensors, high-speed networks, supercomputers, communication technologies, and the multidisciplinary experts involved in the aggregation of big data and the production of knowledge based on the data (Atkin et al. 2003; Kee et al. 2011). Towns et al. (2014) refer to these also as advanced digital services for research and education with big data.

In the metaphor of a factory, an important point is that raw materials get turned into useful products through material manipulations and industrial treatments. Similarly, in a data factory, raw digital data get turned into meaningful insights through computational processing and data analysis. A critical component in data factory is the software that preprocesses and analyzes raw digital data. In fact, many results from the analysis of big data depend on and/or are tied to specific software applications. Therefore, insights drawn from big data are software dependent; without good and appropriate software, the hidden insights in big data cannot be fully tapped.

K.F. Kee (✉)
School of Communication, Chapman University, Orange, CA, 92866, USA
e-mail: kerk.kee@gmail.com

© Springer International Publishing AG 2017

S.A. Matei et al. (eds.), *Big Data Factories*, Computational Social Sciences,
https://doi.org/10.1007/978-3-319-59186-5_5

Additionally, the metaphor of a factory also conjures up the notion of "standardization," a practice made commonplace during the industrial revolution with the introduction of Taylor's scientific management and time and motion studies (Miller 2008). Standardization is important to the idea of data factories, as the standardization of data format and the interoperability of data make data aggregation, recombination, and repurposing for even larger-scale analysis possible. Therefore, a piece of good software should be designed to be easily adopted and widely diffused in order to facilitate the standardization and interoperability of data for data factories.

While much attention has been given to big data as the raw materials that have hidden insights, limited attention has been given to the open source software that turn raw materials into powerful insights. However, without successful design, development, adoption, and implementation of useful software, raw materials will remain raw materials with hidden insights. Metaphorically, a factory full of raw materials without machines to process them is just that – a factory full of raw materials. A factory full of raw materials processed and assembled in a meaningful way can turn the rawness into usefulness. Therefore, intentional and strategic efforts should be carried out to promote wider adoption of good software applications.

The purpose of this chapter is to explore what drives the adoption and diffusion of open source software that can usher in the vision of data factories. With the adoption of good software applications across the community, researchers can begin moving individual data sets developed by independent projects across geographic locations and disciplinary domains into a broader data ecosystem sustainable over the long term. The data ecosystem should also be easily accessible and used by present and future researchers not directly involved with data collection and documentation of the individual data sets.

In order to achieve the stated goal for this chapter, it is organized with the following sections. First, the concepts of data, big data, and e-research are defined. Second, the largest National Science Foundation's (NSF) supercomputing consortium, XSEDE (Extreme Science and Engineering Discovery Environment), is discussed as a specific case of a data factory. Third, based on interviews conducted with community stakeholders of XSEDE, ten drivers of open source software adoption are discussed along with associated critical questions to promote intentional design of software for successful diffusion in the larger research community. Finally, a conclusion with implications wraps up the chapter.

Data, Big Data, and e-Research

Schroeder (2014) defines *data* as the materials that belong to the object(s) or phenomenon(a) of investigation and that data are the most useful unit of analysis for the investigation, which involves data collection before the interpretation. To take it further, Meyer and Schroeder (2015) argue that when a data set is a magnitude larger than any other existing data sets in size and scope within a given domain, the

data set is qualified as *big data*. Furthermore, they suggest that big data represents a new form of collaborative interaction with and around materials for research. The idea is that big data do not exist simply as materials; they require multidisciplinary experts to collaborate in order to harness big data for important insights.

Besides the scholarly definition of big data offered by Schroeder and Meyer, big data is more commonly defined in the industries by several keywords that begin with the letter V. More specifically, the concept of big data was defined by what was first known as the three Vs of big data: volume, variety, and velocity (Laney 2001). The first V of *volume* refers to the size of the data, and it is often measured in terabyte and petabytes. This characteristic is almost intuitive, as the volume is what makes a data set big or bigger than other existing data sets in a given domain. The second V of *variety* indicates that big data have a range of data formats, often referred to as structured and/or unstructured data. If a big data set is made up of simply structured data, its aggregation, recombination, and analysis are relatively straightforward. If a big data set consists of mainly unstructured data, computational analysis will require a lot of data cleaning and conversion, in order to create format consistency (which is also known as the interoperability of data). This is critical for the need of recombining and repurposing of previously isolated data sets from independent projects. The third V of big data is *velocity*, which refers to the speed at which data are produced and processed. The production and processing of big data are usually in real time or near real time. It is also this characteristic that gives big data the currency and dynamic advantage over traditional dated and static data.

Recently, Gandomi and Haider (2015) further argue that big data possess three additional Vs of variability, veracity, and value. *Variability* describes the flow rates of big data as fluctuating, unpredictable, and erratic. The fluctuation of big data's flow rates is due to the fact that big data sets usually are the aggregated results of data coming from various sources. Therefore, big data usually show periodic and sporadic ups and downs in flow rates. The next V of *veracity* implies that despite big data's inexactitude, imprecision, and uncertainty, they hold significant and hidden insights. The insights require strategic harnessing by humans and machines. Finally, the last V of *value* signifies that there is important worth that can be drawn from big data's large volume. As previously mentioned, the large volume of big data is the obvious defining characteristic of big data. Although the large volume of big data, often measured in terabytes and petabytes today, is commonly used as the primary definition of big data, Gandomi and Haider argue that the notion of volume is relative – what is regarded as big at the present time may be small in the future.

Given Gandomi and Haider's point above, perhaps the definition offered by Mayer-Schönberger and Cukier (2013) can be added to the list of defining characteristics. They argue that a data set is considered big data when the size of a sample drawn from a population is equal to the size of the entire population (i.e., $N =$ all). Their argument stems from big data analytics' departure from the traditional practice of sampling and inferential statistics when it was impossible to obtain and/or analyze population data of an entire organization, community, country, or social system. Due to previous limitations in terms of data collection, researchers carefully drew a sample for analysis and then appropriately inferred from the sample

certain insights about the population. This inference was determined by statistical calculations and probability. However, since population data can be obtained today, sometimes through passive data recording, there is no longer a need to simply draw a sample. Moreover, data analysis was previously limited to what a single computer can process. Given today's network capability, big data set can be processed by a network of supercomputers, such as in the case of the Extreme Science and Engineering Discovery Environment (XSEDE).

In summary, big data can be defined by volume, variety, velocity, variability, veracity, and value. These six Vs have also been reduced to simply the five Vs (volume, velocity, variety, value, and veracity) of big data. Today, the five Vs are widely used to define big data, such as in the call for papers by the 2016 IEEE Big Data conference in Washington DC. The main characteristic of volume can be understood also as when the size of the sample is equal to the size of the population or when the volume is at least one magnitude bigger than the size and scope of other existing data sets within a given domain. Finally, big data present the need for multidisciplinary collaborations with and around the data.

What is the purpose of big data then? Meyer and Schroeder (2015) offer the answer that big data are being used for e-research (Borgman 2010; Dutton and Jeffreys 2010). They define *e-research* as "the use of shared and distributed digital software and data for the collaborative production of knowledge." They use the term e-research to be inclusive of e-science, computational social science, digital humanities, and any other computational analyses of big data for advancing knowledge by collaborative researchers. Interestingly, their definition of e-research has an emphasis on the collaborative nature of knowledge production. In other words, if a researcher simply digitalizes the data (e.g., scanning images of historical manuscripts for computational analysis) for personal use, and the researcher does not share the digitized manuscripts with a wider community of researchers, this researcher's work does not fully qualify as e-research. The emphasis of the collaborative nature of e-research is critical for the notion of data factories, as these factories are set up to support open innovations.

The challenge of volume can be addressed by high-performance computing (HPC) and/or high-throughput computing (HTC). When a data set is too big and a single desktop computer cannot process the data (i.e., choked and frozen when the "process" button is pushed), a researcher can apply for an allocation to access HPC and/or HTC at national resources, such as XSEDE. Therefore, the major challenge addressed by the data factory metaphor is that of variety. Metaphorically, a producer of goods made the goods from start to finish before the industrial revolution. Because the process was done by a holistic approach, each product was unique. While the uniqueness may be celebrated by some, the variety can be a problem when there is a need to aggregate, recombine, and repurpose them.

Taking a pro-innovation and innovation diffusion stance, this chapter presents the purpose of data factories as threefold. First, it is about standardization and interoperability to reduce the challenges that come with big data's variety and variability. Second, it is about having centralized data repositories and computational resources to process big data, supporting big data's volume and velocity. Finally, it is

about creating and maintaining a thriving and collaborative community around open innovations, so big data's veracity and value can be fully realized. Given the purpose discussed, the next section presents XSEDE as a specific case of data factory.

XSEDE as a Data Factory

The Extreme Science and Engineering Discovery Environment (XSEDE, www.xsede.org) is the largest supercomputing consortium that provides computational resources and expertise for data-intensive research and education in science, engineering, social sciences, and humanities in the USA. XSEDE consists of more than 20 supercomputers and resources for advanced visualization and analysis of big data. The consortium is led by the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign, and it includes partner centers such as the Texas Advanced Computing Center at the University of Texas at Austin, the San Diego Supercomputer Center at the University of California at San Diego, and universities such as Purdue University and University of Southern California, to name a few (for a full list, please visit https://www.xsede.org/leaders). These partner institutions each contribute one or more allocatable services to the consortium.

XSEDE is funded by the Office of Advanced Cyberinfrastructure (OAC) of the NSF's Computer and Information Science and Engineering (CISE) Directorate to continue advancing NSF's efforts in providing a national infrastructure to support the e-research community and cyberinfrastructure ecosystem started by the TeraGrid (2001–2006) and TeraGrid2 (2006–2011) projects (for information on TeraGrid and TeraGrid2, please see Lawrence and Zimmerman 2007, Towns 2011, and Zimmerman and Finholt 2006). Launched in July 2011 and funded at about $125 million for 5 years, XSEDE transitioned into XSEDE2 in September 2016 for another 5 years with a new round of funding at about $92 million. Similar to TeraGrid and TeraGrid2, XSEDE provides all the resources and support at no cost to the e-research community. XSEDE2 will continue in this approach to support big data and open innovations in the USA.

The goal of XSEDE is not simply to provide supercomputing power; the goal also includes the goal to provide a comprehensive and cohesive set of distributed infrastructure, digital services, support services, and technical expertise to enable e-research and cyberlearning (Towns et al. 2014). Broadly, XSEDE has supported researchers in computational finance, genomics, epidemiology, digital humanities, and social network analysis. Notable examples of groundbreaking research supported by XSEDE include a study of high-frequency trading in the US stock market (see O'Hara et al. 2014) and the hydrogen sorption in a metal organic framework (see Pham et al. 2013), to name a couple. More importantly, based on theoretical assumptions from classical and quantum mechanics, the use of XSEDE's predecessor was utilized for performing the simulation and prediction of the behavior of biomolecules, a study that led to a discovery that awarded Martin

Karplus, Michael Levitt, and Arieh Warshel the 2013 Nobel Prize in Chemistry. Their computational simulation was innovative in taking chemistry research outside of the traditional laboratory (Towns et al. 2014).

XSEDE stores tens to hundreds of petabytes of data, supports a few hundreds of software packages, as well as provides training and services to more than 10,000 researchers and 2500 projects across all 50 states to harness big data for research discovery and knowledge production. XSEDE also supports international researchers from over 100 universities in more than 35 countries who collaborate with the US researchers XSEDE directly supports. XSEDE is an exemplar of a data factory, as Towns et al. (2014) explain – the purpose of XSEDE is for:

> Making codes run faster and more easily allows researchers to get more science done in a fixed amount of time. Lowering the barrier for access to and use of digital services enables additional research in established communities and in new communities who haven't harnessed these services to date. Such productivity increases can be the difference between an infeasible project and a feasible one, reducing the time to publishing scientific findings.

The notions of efficiency and productivity, two characteristics of the machine metaphor of industrial revolution (Miller 2008), are prominent in XSEDE.

As previously stated, the idea of "data factory" can be interpreted as a virtual arrangement or group of arrangements where big data sets are produced and aggregated mainly by cyberinfrastructure. A key feature of cyberinfrastructure is the open source software applications necessary for processing big data. Understanding the adoption drivers that promote diffusion of these software applications is important because existing efforts should not be wasted and new users do not need to reinvent the wheel. Furthermore, wider diffusion of good software will also help create standardization and interoperability of data, further promoting the vision of data factories. Standardization can reduce idiosyncratic measures, and data formats, instead, move data from isolated projects and locked box repositories more easily into a longitudinal data warehouse associated with certain data factories. Finally, with wider adoption, more data can be aggregated, recombined, and integrated to perform analysis at unprecedented scale, to tackle big problems previously limited by the volume and variety of data and the limitation of existing software and supercomputing resources. The ultimate outcome of a pro-innovation diffusion effort in this sense can lead to more innovations and breakthroughs that benefit societies and humanity worldwide. In order to promote diffusion, the next section explores the ten drivers that promote the adoption of open source software for data factories and open innovations within the XSEDE community.

The Ten Adoption Drivers of Open Source Software in XSEDE

The ten drivers discussed in this section were identified in an analysis based on 135 in-depth interviews with domain researchers (as technology users), computational technologists (as software developers), and center administrators (as data center leaders) who consider themselves stakeholders of the XSEDE community (for more

details, see Kee et al. 2016). The interviews were systematically analyzed using the grounded theory approach (Glaser and Strauss 1967; Kee and Thompson-Hayes 2012; Strauss and Corbin 1998). The ten drivers are also discussed with critical questions from the perspective of potential new users. These questions represent the kind of issues that stakeholders should keep in mind while designing and promoting their software within the larger research community to support data factories and open innovations.

Driven by Needs

The first adoption driver is the software's ability to meet users' existing needs. While research to date is still inconclusive about if users' needs drive innovation (see von Hippel 2005 on how lead users created innovations to meet their own needs) or an innovation creates a market for an unknown need (see Daly 2011 on how iPod created a completely new market), the development and adoption of open source software for big data are usually driven by known needs in the research community. This is because big data usually exist before the software to process them is available, and the software is designed to harness existing data. The segment that makes up the potential user market are busy professionals who do not have time to adopt a piece of software simply for personal enjoyment, but for a compelling reason, such as a pressing problem that represents a dire need for a solution.

Furthermore, the design and development of open source software can be very time consuming and financially expensive. This is why many software applications are developed by federally funded projects for 3–5 years (Kee and Browning 2010), such as those supported by NSF's OAC. In these projects, if the inception teams are not able to articulate a compelling rationale with clear reasons for the need to develop a piece of software for research, the projects would not be funded by NSF and other federal agencies (such as the Department of Energy, National Institutes of Health, National Oceanic and Atmospheric Administration). The rationales are often based on grand challenges and critical problems well-documented in the research literature. Therefore, in order for a piece of open source software to widely diffuse, it needs to clearly meet the needs of potential users and the community/funders behind their work. In fact, in their discussion of XSEDE, Towns et al. (2014) open the article by stating that the establishment of XSEDE itself was "[d]riven by community needs." Therefore, a critical question stakeholders should keep in mind that a potential user may ask is "Does this software meet my needs?"

Organized Access

Once there is a compelling need, potential users require organized access to find the open source software they may adopt. The notion of organized access is not

simply having an online link to download a piece of software; the notion includes having a systematically designed location (usually a website, such as HUBzero at https://hubzero.org/ and Galaxy Tool Shed at https://toolshed.g2.bx.psu.edu/) where inception teams post their software, active users rate, review, and comment on the software and potential adopters read about the software online easily. The website should be designed to facilitate a vibrant community where the interactions among different groups of stakeholders (inception teams, active users, and potential adopters) come together to carry a piece of open source software forward.

Having organized access to an online marketplace where the marketplace is well known is important for diffusion. This driver is important for data factories as the community of users need to participate in the marketplace in order to generate open innovations collectively. A piece of diffusing software has to have a strong web presence, and it can be located at a known marketplace that is open and organized for a community of users. Therefore, the critical questions stakeholders should keep in mind that a potential user may ask are "Is the software easily available?" and "Can I find the software at a known location?"

Trialability

The third adoption driver of a piece of diffusible open source software is that it allows potential adopter to try it out before full adoption. Many open source software applications in e-research to date have a high degree of trialability because they are open sourced. These software applications are different from their propriety counterparts in that all the source codes are freely open, so interested developers and savvy users can add to the software and extend the software features based on their existing needs. In other words, being open sourced allows for open innovations and ecologically driven evolution of software. This is an important point for trialability because it is often during the open trials that potential adopters cultivate an understanding of the software and how it works, what it means for them in their particular contexts.

The notion of trialability is important for data factories because the notion of open innovation is eventually driven by open free trials and organic contributions. Within the community of e-research and open innovations, members subscribe to the open sharing philosophy. Once a piece of software is aligned with the potential users' philosophical orientation, the software should also be easily implemented for a trial without too much learning time. A steep learning curve will discourage adoption. Therefore, the critical question stakeholders should keep in mind that a potential user may ask is "Can I try this software without much time investment?"

Well-Documented

Documentation refers to having a complete record of how the software was developed, instructions on integrating and using the software, the decisions that went into the design, the updates, and exemplars of how the software has been successfully used to solve different big data problems. A piece of software that is well documented not only offers potential adopters simply basic information to download the software, it offers a learning environment that is akin to a fully developed course on a piece of software. In other words, the documentation cannot be outdated and/or skeletal. Otherwise, another software with better documentation will likely attract more active users and potential adopters.

Being well documented is also an important characteristic for data factories, because the community members for open innovations are diverse, and the vision is to maintain long-term data and technologies that allow longitudinal analyses. Even when the pioneering stakeholders are no longer alive 100 years from now, their well-documented software applications can continue being updated and used by future researchers. Therefore, the critical question stakeholders should keep in mind that a potential user may ask is "Is the software well documented with a complete track record and robust user guides?"

Community Driven

Building on the adoption drivers of trialability and being well documented as discussed above, the more people can try out a piece of software with helpful documentations, the more active a community will develop around the software. The open sourced nature continues to manifest in the adoption driver of being community driven. The open source philosophy does not only give innovations freely to a marketplace, it empowers a community of stakeholders to rally around the software. The source codes are open online; this allows many savvy users, potential adopters, and interested developers to participate in trying out the software, integrating the software, fixing the bugs, updating the codes, improving its functionality, and extending its usage to new problems and contexts previously not considered by the inception team.

Shirky (2009) beautifully elaborates on Eric Raymond's notion of "a plausible promise" – the promise that the original developer will not take advantage of community contribution for personal financial gain. A plausible promise is what gives community members the reason to join and contribute to the community. The

driver of community driven is fundamental to data factories for open innovations. It is also this driver that gives future adopters the confidence that the software will continue to thrive with the support of the community. Therefore, the critical question stakeholders should keep in mind that a potential user may ask is "Is there a thriving community that will carry this piece of software forward for the long term?"

Observability

The next adoption driver is observability. The notion of observability manifests in terms of how often near peers talk about a piece of software (i.e., word of mouth), how frequently the software is showcased in research presentations and/or demonstrations at a conference (i.e., community visibility), and its success in enabling good research and producing useful results (i.e., citation index). The notion of observability based on the three dimensions of word of mouth, community visibility, and citation index allows a piece of software to create the impression that the software has a strong potential to be useful for potential adopters.

The driver of observability is also important for data factories and open innovations because the contribution to and access to repositories depends on whether community members are aware of the software and related data archives. The more observable the software is, the more likely it will attract a group of stakeholders around it. Therefore, the critical question stakeholders should keep in mind that a potential user may ask is "What software are my peers using, and how are they using it?"

Relative Advantage

The adoption driver of relative advantage refers to a piece of software's ability to outperform an existing software in multifold. It is important to note that potential adopters are often entrenched in their existing technologies. Therefore, it is difficult or painful for them to transition. The new open source software has to offer a multifold advantage for potential adopters to overcome their resistance to avoid pain during a software transition.

In the case of open source software, a large segment of potential adopters are not existing users of other open source software, but potential adopters of the computational approach to gain insights from big data. In other words, these individuals have to be convinced not simply that the software is going to help them do their work better, but that the computational approach and big data will help them solve problems that are bigger in scale and more complex in scope or to solve a problem that otherwise cannot be solved with their existing technologies and approaches based on sampling techniques and samples drawn from larger populations of interest.

The driver of relative advantage is also important for data factories because the idea of open sharing an open innovation is still relatively new for the traditional research community grounded in individual credits for hiring, tenure, and promotion. The bundle of software, big data, and computational approach need to appear a lot more beneficial than the traditional way of doing research. Therefore, the critical question stakeholders should keep in mind that a potential user may ask is "Is this software a lot better than what I have right now?"

Simplicity

Simplicity is key to successful software adoption. Very few people will take the time and effort to adopt a piece of complex software that is difficult to learn. There are always some die-heart users who believe that to fully do computational data processing, one needs to know the nitty gritty of programming and supercomputers. However, these individuals make up a small segment of the market place, possibly only those who are referred to as "innovators" (2.5% of total population) in Rogers' (2003) original diffusion model.

Instead of the need to learn how to program like those previously referred to as active and savvy users, there is now a steady effort in creating science gateways to lower the barrier of entry (Wilkins-Diehr et al. 2008). Science gateways are essentially open source software designed with a user-friendly interface. According to the XSEDE website:

> A Science Gateway is a community-developed set of software, applications, and data that are integrated via a portal or a suite of applications, usually in a graphical user interface, that is further customized to meet the needs of a specific community. Gateways enable entire communities of users associated with a common discipline to use national resources through a common interface that is configured for optimal use. Researchers can focus on their scientific goals and less on assembling the cyberinfrastructure they require. Gateways can also foster collaborations and the exchange of ideas among researchers.

As described above, with a science gateway, users can simply use the point-and-click method to navigate and use the software to process big data. According to Towns et al. (2014), more than 40% of XSEDE users in 2013 were users of one of more than 35 science gateways associated with XSEDE in the same year. This portion of users is expected to continue growing over time. Therefore, the critical question stakeholders should keep in mind that a potential user may ask is "Is this software simple to learn and easy to use?"

Compatibility

The adoption driver of compatibility refers to a piece of software's fit with a potential adopter's technological repertoire, behavioral practices, and ideological

orientation toward data-driven research. If the innovation is disruptive (technologically, behaviorally, and ideologically), both for the potential adopters and/or their collaborators, the innovation will suffer greatly in terms of compatibility. As today's researchers are heavily dependent on their technologies, the further a new piece of software departs from their existing routine and/or the norms in their disciplines, the more difficult it is for the software to be adopted.

This driver is also important for data factories because in order for a community of open innovations to thrive, it needs to attract many members. A potential member may compare and contrast if his/her data format is compatible with the format chosen by a data factory of interests. Without data interoperability, the aggregation of data sets into a big data set is difficult. The software and the data format go hand in hand for the adoption decision by potential users. Therefore, the critical question stakeholders should keep in mind that a potential user may ask is "Can I easily integrate this software into my existing routine and collaborations?"

Adaptability

Traditionally, adoption with a deviation from the original purpose of a piece of software is considered as "noise" in diffusion research. This bias is understandable because a deviation does not count as a full adoption if a researcher or manufacturer is interested in tracking "successful adoption" of a new technology as originally designed. However, in the Web 2.0 era, a deviation from the original purpose (such as in terms of adaptability, repurposing, and reinvention) may aid in a piece of open source software's ability to diffuse. In other words, a piece of software's ability to adapt and be repurposed for a new problem and/or a new context may promote its wider adoption ultimately.

The adoption driver of adaptability should not be left as simply a happy accident. In fact, it can be an intentional diffusion strategy – a piece of software is designed to repurpose across problems, contexts, fields, and domains. Therefore, the critical question stakeholders should keep in mind that a potential user may ask is "Can I take this piece of software from that domain and bring it into my domain?" Table 5.1 below summarizes the ten adoption drivers and associated critical questions as discussed above.

Conclusion, Discussion, and Implications

This chapter set out to explore the definitions of data, big data, and e-research in the context of data factories for open innovations. The metaphor of a factory for data is compelling as it implies key characteristics for data such as standardization and interoperability and for open innovations such as efficiency and productivity. The chapter presents the case of XSEDE as an exemplar of a data factory in the

Table 5.1 The ten adoption drivers of open source software in the e-research community for data factories and open innovations

Adoption drivers	Critical questions
Driven by needs	*"Does this software meet my needs?"*
Organized access	*"Is this software easily available?" and "Can I find the software at a known location?"*
Trialability	*"Can I try this software without much time investment?"*
Well documented	*"Is the software well documented with a complete track record and robust user guides?"*
Community driven	*"Is there a thriving community that will carry this piece of software forward for the long term?"*
Observability	*"What software are my peers using, and how are they using it?"*
Relative advantage	*"Is this software a lot better than what I have right now?"*
Simplicity	*"Is this software simple to learn and easy to use?"*
Compatibility	*"Can I easily integrate this software into my existing routine and collaborations?""*
Adaptability	*"Can I take this piece of software from that domain and bring it into my domain?"*

USA. Most importantly, this chapter laid out ten drivers that promote the adoption and diffusion of open source software in the e-research community to usher in the vision of data factories and open innovations. It is important to note that the ten drivers make up a need-based diffusion model, a broader technology adoption framework. Although the ten drivers were presented in a linear and sequential way, it is important to keep in mind that they are interconnected and they influence each other in a complex way at any given time.

The topic of software adoption is not simply a theoretical question; it is also an important practical question. Instead of providing direct recommendations for practice, the ten drivers were presented with associated critical questions (see Table 5.1 for a summary) to prompt the stakeholders to ponder upon and discussed the ten different drivers at any given point in time. In a fast-changing world of technologies, a specific recommendation is likely to be outdated in the foreseeable future. Furthermore, a recommendation that works well in one particular disciplinary domain may not work in another domain. However, by engaging with the critical questions, stakeholders can come up with the best answers for themselves in their given contextual and historical contexts. Therefore, the critical questions are useful for facilitating stakeholders' regular reflections on the challenges and opportunities to promote their software applications for data factories and open innovations.

While the focal point was on the adoption of open source software as a technology, an important insight stemmed from the discussion above is that the adoption decisions are multidimensional. Kee (2017) uses the adoption of green technologies within the workplace as an example to make this point. More specifically, the adoption of the green technologies also involves the adoption of the recycling and/or conservation behaviors and the belief and mindset that environmental sustainability is of critical urgency and importance within the workplace. If the push to adopt a

green technology only focuses on the technology itself, the stakeholders are missing the critical fact that the adoption of the technology is not complete without the adoption of the associated behavioral practices and philosophical ideologies.

Similarly, the argument can be extended to the adoption of open source software for data factories and open innovations. The potential adopters need to be willing to modify existing practices to make the software fit into existing routines and collaborations. The potential adopters also need to strongly believe that open source software, data factories, and open innovations are the ways of the future of research and knowledge production. The adoption decision is multidimensional, as it involves the adoption of the material objects (i.e., open source software, big data), the behavioral practices (i.e., large-scale scientific collaborations, open sharing of data and documentation), and philosophical ideologies (i.e., data factories, open innovations). The adoption of one dimension without the others would be considered incomplete. The case of XSEDE presents an interesting context to study the diffusion of multidimensional innovations for adoption.

Acknowledgment The author thanks Mona Sleiman, Rion Dooley, Nancy Wilkins-Diehr, and John Towns for their support of this project. This research was funded by NSF ACI 1322305.

References

Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Molina, H., Klein, M. L., Messerschmitt, D. G., et al. (2003). Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation Blue-ribbon Advisory Panel on Cyberinfrastructure. Washington, DC: National Science Foundation. Retrieved from 19 Dec 2006. http://www.communitytechnology.org/nsf_ci_report/.

Borgman, C. L. (2010). *Scholarship in the digital age: Information, infrastructure, and the internet*. Cambridge, MA: MIT press.

Daly, J. A. (2011). *Advocacy: Championing ideas and influencing others*. New Haven: Yale University Press.

Dutton, W. H., & Jeffreys, P. W. (2010). *Worldwide research: An introduction*. Cambridge, MA: MIT Press.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*, 137–144.

Glaser, B. G., & Strauss, A. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Piscataway: Aldine Transaction.

Kee, K. F. (2017). Adoption and diffusion. In C. Scott & L. Lewis (Eds.), *International encyclopedia of organizational communication, 1* (pp. 41–54). Chichester: Wiley-Blackwell.

Kee, K. F., & Browning, L. D. (2010). The dialectical tensions in the funding infrastructure of cyberinfrastructure. *Computer Supported Cooperative Work, 19*, 283–308.

Kee, K. F., Cradduck, L., Blodgett, B., & Olwan, R. (2011). Cyberinfrastructure inside out: Definitions and influencing forces shaping its emergence, development, and implementation. In D. Araya, Y. Breindl, & T. Houghton (Eds.), *Nexus: New intersections in Internet research* (pp. 157–189). New York: Peter Lang.

Kee, K. F., & Thompson-Hayes, M. (2012). Conducting effective interviews about virtual work: Gathering and analyzing data using a grounded theory approach. In S. D. Long (Ed.), *Virtual work and human interaction research* (pp. 192–212). Hershey: IGI Global.

Kee, K. F., Sleiman, M., Williams, M., & Stewart, D. (2016). The 10 attributes that drive adoption and diffusion of computational tools in e-science. In P. Navrátil, M. Dahan, D. Hart, A. Romanella, & N. Sukhija (Eds.), *Proceedings of the 2016 XSEDE Conference: Diversity, big data, & science at scale*. New York: ACM.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META Group Research Note, 6*, 70.

Lawrence, K. A., Zimmerman, A. (2007). TeraGrid planning process report: August 2007 user workshops. *Collaboratory for Research on Electronic Work, School of Information, University of Michigan.* Retrieved January 10, 2010, from http://deepblue.lib.umich.edu/handle/2027.42/61842

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. New York: Houghton Mifflin Harcourt.

Meyer, E. T., & Schroeder, R. (2015). *Knowledge machines: Digital transformations of the sciences and humanities*. Cambridge, MA: MIT Press.

Miller, K. (2008). *Organizational communication: Approaches and processes* (5th ed.). Belmont: Wadsworth Publishing.

O'Hara, M., Yao, C., & Ye, M. (2014). What's not there: Odd lots and market data. *The Journal of Finance, 69*, 2199–2236. https://doi.org/10.1111/jofi.12185.

Pham, T., Forrest, K. A., Nugent, P., Belmabkhout, Y., Luebke, R., Eddaoudi, M.,. .. Space, B. (2013). Understanding hydrogen sorption in a metal–organic framework with open-metal sites and amide functional groups. *The Journal of Physical Chemistry C, 117*, 9340–9354. doi: https://doi.org/10.1021/jp402304a.

Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). New York: Free Press.

Schroeder, R. (2014). Big data and the brave new world of social media research. *Big Data & Society, 1*, 2053951714563194.

Shirky, C. (2009). *Here comes everybody: The power of organizing without organizations*. New York: Penguim.

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (2nd ed.). Thousand Oaks: Sage.

Towns, J. (2011). *The sunset of TeraGrid and the dawn of XSEDE*. Paper presented at the The 10th Annual Meeting on High Performance Computing and Infrastructure in Norway (NOTUR2011), Oslo. http://www.notur.no/notur2011/material/TG-to-XSEDE-for-NOTUR11_Towns.pdf.

Towns, J., Cockerill, T., Dahan, M. F., Ian, Gaither, K., Grimshaw, A., Hazlewood, V., ... Wilkins-Diehr, N. (2014). XSEDE: Accelerating scientific discovery. *Computing in Science & Engineering, 16*, 62–74.

von Hippel, E. (2005). *Democratizing innovation*. Cambridge, MA: The MIT Press.

Wilkins-Diehr, N., Gannon, D., Klimeck, G., Oster, S., & Pamidighantam, S. (2008). TeraGrid science gateways and their impact on science. *Computer, 41*(11), 32–41.

Zimmerman, A., & Finholt, T. A. (2006). TeraGrid user workshop final report. *Collaboratory for Research on Electronic Work, School of Information, University of Michigan.* Retrieved January 10, 2010, from http://deepblue.lib.umich.edu/handle/2027.42/61841.

Chapter 6
Aligning Online Social Collaboration Data Around Social Order: Theoretical Considerations and Measures

Sorin Adam Matei and Brian C. Britt

Introduction

The development of online collaboratories, which enable users to engage in the production and creation of content, has proliferated to such an extent that it has become routine (Wang et al. 2007). Rather than being monopolized by a few individuals and institutions and tethered to a specific place and time as their site for work, as it was in the past, collaborative content produced on the Internet is created and distributed worldwide by millions of individuals, groups, and organizations that meet online (Harley and Blismas 2010). Nonetheless, while some collaborative efforts, such as that of Wikipedia, are able to attract and retain contributors, countless others are unable to fulfill the core requirement for successful, sustained collaboration (Koschmann 2016; Marshak 2005) and either stagnate or collapse entirely (Harley and Blismas 2010; Ingram and Hathorn n.d.; Koschmann 2016).

Understanding the evolution and configuration of collaborative online organizations demands a flexible, yet comprehensive framework of investigation, which sets at its core the issue of social order. This should take into account the fact that, while user interactions and behaviors are easy to collect, conceptualizing a model of social order that works well across contexts and interaction environments is substantially more difficult. Data heterogeneity and model complexity are some of the recurrent challenges that need to be overcome. Most importantly, when it comes to interpreting the social and technological patterns identified via automated

S.A. Matei (✉)
Purdue University, West Lafayette, IN, USA
e-mail: smatei@purdue.edu

B.C. Britt
South Dakota State University, Brookings, SD, USA
e-mail: brian.britt@sdstate.edu

© Springer International Publishing AG 2017

S.A. Matei et al. (eds.), *Big Data Factories*, Computational Social Sciences,
https://doi.org/10.1007/978-3-319-59186-5_6

harvesting, interactional complexity is often one or two orders of magnitude greater than that encountered when examining traditional organizations (Capocci et al. 2005, 2006).

We argue that some of the obstacles to handling collaborative social media analytic tasks can be mitigated or eliminated by early and rigorous definition of the research space, which should include reliable and flexible instruments for defining social order. Preliminary steps should, however, be taken even before engaging these issues. These should include defining the actors and their behaviors, their domain of interaction, the quality and quantity of their output, and the manner in which they are seen as contributors to social order. Furthermore, collaborative data analysis should propose, from the beginning, a sound theoretical framework for understanding social interaction and social order as a "social fact" (Greenwood 2003), that is, as a reality that is greater than the sum of its parts. Social groups should be seen not as derivatives, but as emergent realities that transcend the members, even though they may be characterized as "plural subjectivities" (Gilbert 1992) or "intersubjective realities" (Schutz 1970). Order should be observable as patterns that exist long after their initiators ceased being active in a group, on a website, or in a project. Collaboration, likewise, should be viewed not as a simple aggregation of local, individual rules applied in isolation by otherwise uncoordinated individuals, but as an emergent order directed by common goals, norms, values, implicit ideologies, and a need for coordination and control. Such order must be defined by transindividual rules and behavioral patterns, which emerge from group-defined and group-enforced norms, values, principles, and goals.

A well-defined set of measures and indicators should complement the theoretical approach. Regularities and patterns that transcend individuals need to be considered as observable facts that may be measured with reliable and accurate instruments. Of these, measures of social structure and social structuration are the most important. There is no more vital task when observing collaborative spaces than being able to tell whether they stem from a social structure and an emerging social order. As such, a substantial portion of this chapter is devoted to directly addressing the question of structure. Along the way we will suggest ways to translate the newly rediscovered concept of social order into a cross-group comparison tool and articulate an argument for platforms that can leverage this concept through visualization and interaction affordances.

Social Order as (Neg)entropy

Online collaboration occurs in many forms. Researchers have proposed numerous definitions for what constitutes "online collaboration" (Faraj et al. 2011) and how such efforts should be undertaken, but no clear consensus exists (Faraj et al. 2011; Harley and Blismas 2010). More importantly, much research should be conducted to clearly understand online collaborative processes. For example, despite recurring claims that online collaboration is innately egalitarian and potentially superior due

to some form of "collective intelligence" that spontaneously emerges with minimal coordination (Kelly 1994), there is mounting evidence that online interaction instead follows traditional patterns of collaboration (Matei and Bruno 2015).

Additionally—and quite interestingly—individual effort, inputs, and outputs are regularly observed to be unevenly distributed (Huberman 2001). Emergent coordination and/or power hierarchies accompany these uneven distributions (Shaw and Hill 2014), and online groups that are rooted in these uneven distributions are consequently more likely to be productive.

These complex and uneven social arrangements require theoretical models and derived tools that can explain how social encounters take shape online. They should emphasize social structural approaches. The models should give an account for the emergence of social structures, roles and reputations, and the tools derived from them, in turn, may be used to explain, with maximum efficiency, individual and group effectiveness. To this end, we propose an approach to measuring and defining social structure via entropy as a negentropy phenomenon (Stepanić et al. 2005), as detailed below.

A significant amount of empirical evidence indicates that collaborative effort in online environments, such as those of Wikipedia, the open source software (OSS) and Linux movements, and other social interactions such as online bulletin boards or discussion groups are typically distributed in the shape of a highly skewed curve (Barabasi and Frangos 2002; Huberman 2001). These findings suggest that online social environments tend to naturally lead to social aggregates that are dominated by a few sources, voices, or actors. However, this might or might not lead to highly hierarchical and strictly compartmentalized groups. Power structures that are tightly scripted and organized do not perform very well online. They often run into problems of their own, including inefficient utilization of resources and poor allocation of effort, and they manifest an inability to fully capture and circulate local and tacit knowledge throughout the organization. These problems contribute to increasing transaction costs, which limit the scale of any organization, even if well structured (Coase 1937).

With this in mind, online social processes need and often find an equilibrium point between extreme decentralization and total egalitarianism, which instantiates a certain form of social order. This is, presumably, the point where a sufficient but not excessive number of roles and methods of work, authority, and reward allocation have emerged (Blau 1959; Coase 1937). More importantly, this is a point of social structuration, around which patterns of authority, work investment, and role allocation take place (Blau 1975). Finding this optimal point on the curve of social structuration which marks the presence of social order should be a central concern when trying to understand data collection, interpretation, and cross-project comparisons for online collaboration (Sydow et al. 2017).

Previous work (Matei and Bruno 2015) has proposed that Shannon's theory of communication and its companion measure, social entropy, can be applied as an index for user participation and collaboration in online and/or technological systems (Shannon and Weaver 1949). The point at which groups become structured can be identified by plotting the evolution of a group's entropy over time.

Although the complete demonstration of what entropy is and why it can be used as a proxy for social order is long and cannot be made here (Matei and Britt 2017), suffice it to say that entropy is a relatively concise and elegant solution for handling order because it is based in a simple idea. Systems are made of elements, which can take random or nonrandom (organized) states. Random states, statistically speaking, imply that no element should be expected to be in a certain state more than what chance alone predicts. They are, by definition, "egalitarian." In other words, simpler, disorganized systems have a maximum level of entropy because their constitutive elements are equally likely to be in a particular state at a given moment in time. This also means that the elements are equal in other respects. Taking the example of a socio-collaborative system, when the participants in such a system have no clear role delineations and spontaneously contribute an equal amount of work, they are likely to work in isolation and perform redundant tasks. In contrast, when some members contribute much more than others, the group is more likely to be organized because some members lead by example and shape the project, being in more than one place (collaboratively speaking) at the same time.

Social entropy, in this context, is an indicator of several dimensions of social collaboration. On the one hand, it measures diversity and evenness of collaboration in raw terms. It can also be used as a quantitative indicator of how strongly or weakly skewed the collaborative effort is in terms of inputs to the collaborative process. Seen through this lens, we consider social entropy as a higher-level indicator of group structuration. Groups that are dominated by one or more members are also more likely to have a command, power, and communication structure. In short, we can use social entropy to measure a group's level of coalescence. This can be a first step toward characterizing it as an entity that denotes institutional characteristics. In previous research, we have shown how this process works on Wikipedia (Matei and Britt 2017). We have also explicated an evolutionary model inspired by the community of practice paradigm which shows that entropy measurements can be used to track and describe organizational behaviors and patterns over time (Matei and Britt 2017).

At this point, a brief note about using entropy to measure social structuration is warranted. As a measurement tool, entropy increases as system chaos increases. It is, in fact, a measure of social organization in the negative. Social order increases (at least to a point) as entropy *decreases*. Due to this measurement subtlety, it is useful to think about social order as a manifestation of what Schrodinger called "negative entropy," or in short, negentropy (Schrödinger 1948). This is also similar to the amount of "free energy," a concept that has been explored in social sciences as well. Negentropy, in this context, is a term that covers the same territory as entropy, with the difference that we consider it as a measure of social structuration. We do not use (neg)entropy in the more technical sense, as differential entropy, as we instead measure it in a very traditional and simple manner via the formula $H = -\Sigma p\log(p)$. To mark this distinction, we will put the negative particle (neg) in parentheses.

This theoretical work, conducted in collaboration with physicists, computer, and social scientists, has resulted in the Visible Effort wiki visualization tool (http://veffort.us). The main goal of this site is to make social order visible on every page

via (neg)entropy measurements. Social order is represented as unevenness of gross and net contributions. Visible Effort compares page versions on a word-by-word basis, and users are credited with any deleted, modified, or inserted words. Once words are counted and allocated to each user, the entropy value is generated and stored for each edit and each iteration of editing. (Neg)entropy values reflect how even or uneven collaboration was, with the recognition that, in this case, evenness also indicates the degree to which the process of collaboration was structured or unstructured.

(Neg)entropy values are used dynamically to shape the page layout using easily comprehensible conventions. Key visual elements of the collaborative space (page), especially its frame, darken or condense as the level of (neg)entropy increases. This communicates, at a glance, to the system administrators and to the users how even (or structured) the collaborative process currently is. When the color is the lightest, entropy is 0 and (neg)entropy at a maximum. Credit for the collaborative effort should be assigned to only one member of the team. When the color is the darkest, there is perfect equality (evenness, high entropy, low (neg)entropy). In addition, a chart visually reflects the distribution of effort for each collaborator as well as tabular information such as the number of words or characters contributed by each individual. In the administrative space of the page, there is also a line chart that tracks the entropy level of each page as it has evolved over time.

The theoretical and practical approach presented so far has several advantages. Since it relies on measuring relative proportions of effort, (neg)entropy is easy to conceptualize in any collaborative system. (Neg)entropy measurements can be repeated over time, which allows the longitudinal development of social order to be tracked. Finally, (neg)entropy can be examined in tandem with other measurements to assess its covariance with, for instance, group effectiveness, content quality, and the generation of trusted content. In this form, (neg)entropy becomes a very meaningful measure for calibrating the effect of social order on variables that measure outcome and effectiveness.

Overall, (neg)entropy represents, again, a simple and clear means to measure social order that is easily convertible and transferable across social spaces and collaborative situations. With it, wiki groups or similar online collaboration collectives can be simply and efficiently characterized. (Neg)entropy enables at-a-glance assessment and measurement of social order, on which we can build other more meaningful social scientific research questions.

Social Order as Social Embedding: Some Common Measures

(Neg)entropy is a probabilistic measure that synthesizes system states. It works across and, to a certain extent, smooths out the complexity of social roles, connections, and interactions. The study of complex collaborative spaces demands specific tools to examine the extent to which various social influence processes operate in the collaborative online environment. There are clear strengths and weaknesses in

emerging forms of collaboration and knowledge creation stemming from the manner in which participants are embedded in particular interactional configurations. Given that the level of diversity in the knowledge creation process is one of the key indicators of optimal collaboration (standing alongside the knowledge creation outcomes themselves), the extent of social embeddedness represents a crucial determinant for collaboration. In large-scale and complex systems of knowledge creation, embeddedness needs to be considered from the perspective of coeditorial work. Moreover, the concept of embeddedness is in fact coextensive, at least in part, with a core attribute of social order, patterned interactions.

Collaborative coeditorial networks offer unique opportunities for operationalizing a variety of research questions about embeddedness and, ultimately, social order. First, they can be used for mapping complete interaction graphs. For instance, on Wikispaces, a widely used wiki service, every action is recorded and sequentially ordered, just as on Wikipedia. By mapping interactions that occur as multiple users coedit the same page, we gain the opportunity to explore social graphs of interaction in their entirety and across time. Further, the social structure of Wikispaces incorporates node-level attributes that provide longitudinal information regarding typical and atypical node behaviors at various stages of the network's development. Most importantly, coeditorial graphs speak about the degree of social embeddedness, which may be conceptualized as social work with peers or with community leaders that leads to social order.

The notion of social embeddedness (Granovetter 2005; Uzzi 1996) in network theory has highlighted the fact that network formation is necessarily intertwined with various types of social relations including personal connections, trust, social capital, and co-work. It acknowledges the dynamic nature of the systems of social relations within which individual actions are embedded (Granovetter 1983), and it can thus be used to examine how social structures, technological systems, and individual actors interact with each other. It is extremely productive, then, to visualize the structure of collaboration at the level of the work performed and content produced and to measure the attributes of and relationships between coeditors (i.e., who interacts with whom).

Early work on Wikispaces focused on visualizing histories of edits made to Wikipedia articles over time in order to operationalize collaboration patterns (Wattenberg et al. 2007). More recently (Leskovec and Sosič 2016), much more complex approaches have been proposed (Brandes et al. 2009; Halatchliyski and Cress 2014; Kenis and Lerner 2014). Of course, social embeddedness research is not without costs on pseudonymous sites like Wikipedia. While this approach holds value, it remains limited in its ability to process large numbers of edits and does not consider the identities of authors/collaborators involved in a particular instance of editing (who edits what). Still, this is a temporary and small problem, as a diverse array of studies has expanded the vitality of the approach from those conducted in the early days (Brandes et al. 2009; Capocci et al. 2005) to the more contemporary (Kenis and Lerner 2014), building upon the analysis of online interactive content pioneered two decades ago (Barabasi and Frangos 2002).

Any social structural characterization of collaborative spaces via graph approaches should include properly chosen system-level measures. We propose the following subset which, much like entropy, can be used across projects and domains as standard methods of group characterization:

1. Degree assortativity (Newman 2010), which represents the tendency (or lack thereof) for well-connected users to interact with similarly well-connected users rather than those with more or fewer network connections
2. Degree and betweenness centrality (Freeman 1978), which represent the relative importance of an individual within a community based on his or her connections and position within the network
3. Reciprocity (Shi et al. 2007), which on Wikipedia represents the tendency for users to edit the text of others who have already edited their own textual contributions
4. Transitivity (Newman 2010), the tendency toward editing an individual's text if he or she has already served as a coeditor for one of a user's own coeditors
5. Multiplexity (Haythornthwaite 2001), the tendency toward connecting with other users in multiple distinct networks (such as the collection of articles and "Discussion" pages or articles housed in different categories representing distinct knowledge bases)

Each of these graph characteristics is more than a measurement procedure; it is a yardstick and a point of articulation for connecting multiple datasets and data processes around a theoretically grounded idea of social structure. Anchoring heterogenous data alignment strategies around these measurements allows researchers to see how group configurations and, at the same time, individual behaviors (assessed via degree and betweenness centrality) emerge and ultimately stabilize across collaborative communities. Furthermore, such measures can be correlated with output-related activity. This provides the capability to track the impact of structural features on the collaborative process overall and on outcomes in particular. Comparisons of structural features of assortativity and betweenness with longitudinal changes in overall activity levels may suggest fine-grained structural impacts on effective collaboration.

The application of common network analytic measurements across collaborative online spaces is more than an issue of convenience. It may, in fact, presage socially intelligent data factory processes. As high-level indicators, network measurements sort and define contexts of interaction. Their differential effect can suggest ready-made explanations for differences between groups and collaborative spaces. Findings derived from this type of research may allow us to identify sociostructural determinants that influence participation in the knowledge creation process and yield optimal collaborative outcomes. They may also enhance our understanding of the factors contributing to sustainable, socially intelligent systems of collaboration.

Final Considerations

The development and utilization of standardized metrics for characterizing collaborative big data sets is not only of theoretical import; it also has practical significance, as such metrics can be used in the development of a next-generation collaborative platform for knowledge production. They provide a venue for dynamically analyzing the processes of collaboration, and they may support compelling visualization and self-moderation techniques, as well. New collaborative platforms may be strategically built around these metrics, as they provide invaluable avenues of assessment for the very functionality of the collaborative space. At the same time, as users engage with the platforms, the effects that visible visualizations and metrics have in terms of inspiring or facilitating self-moderation or collaboration can be further mined for additional theoretical insights.

Seen from an integrative perspective, the deployment of specific metrics that are theoretically grounded across problem spaces provides multiple advantages. Our vision of integrating and aligning data via metrics referring to social structure, social collaboration, and social embeddedness may contribute to the science of collaborative systems in these ways:

1. They extend and make more tangible the concept of integrated data factoring within collaborative knowledge production.
2. They provide guidance in developing visualization tools for collaborative populations and especially of their efforts, review, reputation, and relationships.
3. They offer integrated real-time interactive tools to facilitate processes of collaboration.
4. They yield longitudinal data that can subsequently be used to further refine existing models of collaborative processes and devise increasingly socially intelligent data factoring tools.
5. They facilitate indexing and searching of new collaborative opportunities via code collaboration, sharing, and reutilization.

To conclude, the vision we offer provides theoretical and metric-based guidelines for modeling and visualizing the emergence of effective social structuration discovery in collaborative spaces. Theoretical models based upon entropy measurement and social network modeling can also be further refined as a consequence. We hope that these suggestions may spur the further development of integrated approaches to studying collaboration online.

References

Barabasi, A.-L., & Frangos, J. (2002). *Linked: The new science of networks*. Cambridge, MA: Perseus.

Benkler, Y. (2006). *The wealth of networks: How social production transforms markets and freedom*. New Haven: Yale University Press.

Blau, P. M. (1959). Social integration, social rank, and processes of interaction. *Human Organization, 18*(4), 152–157.
Blau, P. M. (1975). *Approaches to the study of social structure*. New York: Free Press.
Brafman, O., & Beckstrom, R. A. (2006). *The starfish and the spider: The unstoppable power of leaderless organizations*. New York: Penguin.
Brandes, U., Kenis, P., Lerner, J., van Raaij, D. (2009). Network analysis of collaboration structure in Wikipedia. In *Proceedings of the 18th international conference on World wide web – WWW '09* (pp. 731–740). https://doi.org/10.1145/1526709.1526808
Capocci, A., Servedio, V. D. P., Caldarelli, G., & Colaiori, F. (2005). Detecting communities in large networks. *Physica A, 352*(2–4), 669–676.
Capocci, A., Servedio, V. D. P., Colaiori, F., Buriol, L. S., Donato, D., Leonardi, S., & Caldarelli, G. (2006). Preferential attachment in the growth of social networks: The internet encyclopedia Wikipedia. *Physical Review E, 74*(3), 036116.
Coase, R. H. (1937). The nature of the firm. *Economica, 4*(16), 386–405.
Faraj, S., Jarvenpaa, S. L., & Majchrzak, A. (2011). Knowledge collaboration in online communities. *Organization Science, 22*(5), 1224–1239.
Freeman, L. C. (1978). Centrality in social networks: Conceptual clarification. *Social Networks, 1*(3), 215–239.
Gilbert, M. (1992). *On social facts*. Princeton: Princeton University Press.
Granovetter, M. (1983). The strength of weak ties: A network theory revisited. *Sociological Theory, 1*, 201–233.
Granovetter, M. (2005). The impact of social structure on economic outcomes. *The Journal of Economic Perspectives: A Journal of the American Economic Association, 19*(1), 33–50.
Greenwood, J. D. (2003). *The disappearance of the social in American social psychology*. Cambridge, UK: Cambridge University Press.
Halatchliyski, I., & Cress, U. (2014). How structure shapes dynamics: Knowledge development in Wikipedia – a network multilevel modeling approach. *PloS One, 9*(11), e111958.
Harley, J., & Blismas, N. (2010). An anatomy of collaboration within the online environment. In M. Anandarajan & A. Anandarajan (Eds.), *e-Research collaboration* (pp. 15–34). Berlin: Springer.
Haythornthwaite, C. (2001). Exploring multiplexity: Social network structures in a computer-supported distance learning class. *The Information Society, 17*(3), 211–226.
Huberman, B. A. (2001). *The laws of the web: Patterns in the ecology of information*. Cambridge, MA: MIT Press.
Ingram, A. L., & Hathorn, L. G. (n.d.). Methods for analyzing collaboration in online communications. In T. S. Roberts (Ed.), *Online collaborative learning: Theory and practice* (pp. 215–241). Hershey: Information Science Publishing.
Kelly, K. (1994). *Out of control: The new biology of machines, social systems and the economic world*. New York: Basic Books.
Kenis, P., & Lerner, J. (2014). Wikipedia collaborative networks. In R. Alhajj & J. Rokne (Eds.), *Encyclopedia of social network analysis and mining* (pp. 2406–2410). New York: Springer.
Koschmann, M. A. (2016). The communicative accomplishment of collaboration failure. *Journal of Communication, 66*(3), 409–432.
Leskovec, J., & Sosič, R. (2016). SNAP. *ACM Transactions on Intelligent Systems and Technology, 8*(1), 1–20.
Marshak, D. (2005). *Evaluating online meeting platforms for collaboration: Conclusion*. https://doi.org/10.1571/ca1-6-05cc.
Matei, S. A., & Britt, B. C. (2017). *The 1% effect: Structural differentiation and entropy in social media groups*. New York: Springer Nature.
Matei, S. A., & Bruno, R. J. (2015). Pareto's 80/20 law and social differentiation: A social entropy perspective. *Public Relations Review, 41*(2), 178–186.
Newman, M. (2010). *Networks: An introduction*. New York: Oxford University Press.
Schrödinger, E. (1948). *What is life? The physical aspect of the living cell*. Cambridge, UK: Cambridge University Press.

Schutz, A. (1970). *Alfred Schutz on phenomenology and social relations*. Chicago: University of Chicago Press.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana: University of Illinois Press.

Shaw, A., & Hill, B. M. (2014). Laboratories of oligarchy? How the iron law extends to peer production. *Journal of Communication, 64*(2), 215–238.

Shi, X., Adamic, L. A., & Strauss, M. J. (2007). Networks of strong ties. *Physica A: Statistical Mechanics and its Applications, 378*(1), 33–47.

Stepanić, J., Sabol, G., & Žebec, M. S. (2005). Describing social systems using social free energy and social entropy. *Kybernetes: The International Journal of Cybernetics, Systems and Management Sciences, 34*(6), 857–868.

Sydow, M., Katarzyna, B., Paweł, T. (2017). Diversity of editors and teams versus quality of cooperative work: Experiments on wikipedia. *Journal of Intelligent Information Systems, 48*(3), 601–632. https://doi.org/10.1007/s10844-016-0428-1.

Uzzi, B. (1996). The sources and consequences of embeddedness for the economic performance of organizations: The network effect. *American Sociological Review, 61*, 674–698.

Wang, F.-Y., Carley, K. M., Zeng, D., & Mao, W. (2007). Social computing: From social informatics to social intelligence. *IEEE Intelligent Systems, 22*(2), 79–83.

Wattenberg, M., Viégas, F. B., & Hollenbach, K. (2007). Visualizing activity on Wikipedia with chromograms. In *Human-computer interaction–INTERACT 2007* (pp. 272–287). Berlin: Springer.

Part III
Approaches in Action Through Case Studies of Data Based Research, Best Practice Scenarios, or Educational Briefs

Chapter 7
Lessons Learned from a Decade of FLOSS Data Collection

Kevin Crowston and Megan Squire

Introduction

The FLOSSmole project began in 2004 as a multi-institution, multidisciplinary effort to gather, share, and store data and analyses about free/libre open source software (FLOSS) development for academic research (Howison et al. 2006). The original goals of the project were to coordinate among the ongoing collection and analysis efforts of different research groups studying FLOSS development, thus "reducing duplication and promoting compatibility" between both the data sources themselves and the findings from different research groups.

The initial data sources for FLOSSmole included easy-to-collect "low-hanging fruit" (Conklin 2006) such as metadata from web-based project repositories like SourceForge and Google Code. Later, FLOSSmole data grew and became more varied, comprising dozens of disparate sources and data types. Over the years, hundreds of research papers have been written using FLOSSmole data by researchers at every level, from undergraduate students to multi-institution teams.

Along the way, FLOSS development practices and FLOSS itself have changed and are no longer considered as peculiar a phenomenon as it was initially. The organizing principles and development practices of FLOSS have now been thoroughly integrated into mainstream business, media, and scientific research. Accordingly, the data sources within FLOSSmole have also changed in important ways, reflecting the changes we see in the contemporary FLOSS ecosystem. Some of these changes

K. Crowston
Syracuse University School of Information Studies, Hinds Hall 348, Syracuse, NY 13244, USA
e-mail: crowston@syr.edu

M. Squire (✉)
Department of Computing Sciences, Elon University, Elon, NC 27244, USA
e-mail: msquire@elon.edu

© Springer International Publishing AG 2017

S.A. Matei et al. (eds.), *Big Data Factories*, Computational Social Sciences,
https://doi.org/10.1007/978-3-319-59186-5_7

include the decline of the project-based software forge as an organizational entity and the concurrent increase in transparent, social development practices in non-open "walled gardens," many times without a traditional FLOSS license.

As FLOSSmole continues to serve the research community into the future, we will continue to face many challenges in collecting and analyzing data for such a constantly evolving, dynamic community of practice. This means we will need to collect new data sources, support new and expanded analysis techniques, and pursue intersections with complementary research fields such as open data and open communities.

This chapter will serve as an important documentation of the history of FLOSSmole, starting with an overview of the data sources and data types in our repository, and describing the kinds of findings researchers have learned from analyzing this data. Next, we outline the challenges we have faced along the way, including data collection, data validation, and data integration. Our hope is that these insights can illuminate possible warning signs and courses of action for other research groups hoping to build similar systems. Finally, we outline our many goals for future growth and expansion of the system.

Data Sources and Data Types

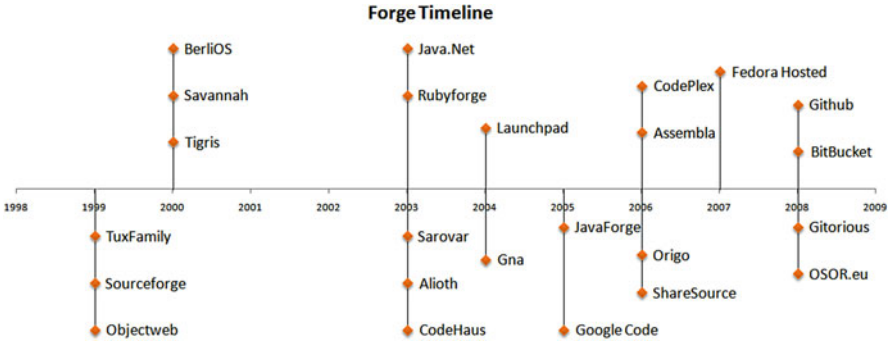
FLOSSmole collects, cleans, stores, and analyzes data about FLOSS and FLOSS development from a variety of online sources, including open source software forges, directories of open source software resources, individual project communication channels, and other project- and developer-related sources. Our data collection methodology consists of a spidering step where we automatically download data from the target online archive, a cleaning and parsing step where we extract the interesting bits of information from the downloaded data, and a storage step where we add metadata so that we can preserve the data for the long term. Our artifact collections are distinct, each with their own attributes and characteristics, but are also part of a common schema that can be federated as necessary. Some examples of each type of collected artifact are given below, along with ideas for the types of questions that can be answered with each.

Forge Metadata

Traditionally, a software forge is a collection of centralized, online tools designed to support collaboration between software developers working on a team to produce a product. The idea of a forge grew out of the earlier concept of a CDE, or collaborative development environment. CDEs were first positioned in the literature as web-enabled and virtualized extensions of the traditional developer desktop IDE (integrated development environment) (Booch and Brown 2003). Well-known IDEs

include Eclipse, ActiveState Komodo, or Microsoft Visual Studio. An IDE typically provides features such as a text editor, shell, file uploads, compiler integration, interactive debugger, integration with bug tracking systems, integration with version control systems, etc. The CDE was described by Booch and Brown as "a virtual space wherein all the stakeholders of a project...labor together to ...create an executable deliverable and its supporting artifacts." A software CDE is, then, a set of tools that facilitates the same tasks as a software IDE (writing code, writing documentation, finding and fixing bugs, distributing releases) but does so in a way that meets the needs of distributed (over both time and space) groups of developers.

With the commercialization of the Internet in the mid- to late 1990s, software development teams continued to become more geographically dispersed and dependent upon Internet-based tools for collaboration. FLOSS teams in particular tended to collaborate in this decentralized way, and many early software forges were specifically created to be used by FLOSS teams. During the late 1990s and early 2000s, many FLOSS-oriented software forges sprung up, offering some kind of combination of hosted services, including file downloads, email mailing lists, wikis, source code control, chat or forums, documentation hosting, and so on. Examples of early FLOSS software forges include SourceForge, RubyForge, Launchpad, GNU Savannah, Google Code, and Microsoft CodePlex. Each of these sites hosts (or, in the case of Google Code, hosted) between several thousand and several hundred thousand projects. Today, while not strictly a FLOSS forge, the very popular Github site hosts over 35 million projects, providing source code control, bug tracking, comments, and other development tools.



Collecting Data from Forges Because of the large number of projects available on a forge and the rich possibilities for studying so many project artifacts (e.g., for cross-project studies), these FLOSS forges were some of the first places that we began collecting data when the FLOSSmole project began in 2004. SourceForge, as the largest and easiest to use forge, was a very rich source of data about projects, developers, and teams.

The type of interesting data that is typically available on a forge includes both metadata and project artifacts. Metadata includes information about the projects, teams, and developers. Artifacts could include source code, bug reports, mailing list

messages, documentation, and the like. Which artifacts are available on which forge depends solely on the forge itself, what data it stores about a user or project, and what the forge chooses to make available to others to see.

In a 2011 paper (Squire and Williams), we investigated which metadata elements and which artifacts were made available at each of the forges in existence at that time, so that we could determine the most fruitful avenues for data collection. The full results of the 2011 forge study are available as an interactive web page on FLOSSmole.org (link: http://flossmole.org/content/everything-you-ever-wanted-know-about-software-forges-code-forges-june-2011). Below we show the lists of the 60 metadata elements and artifacts that we tracked.

Forge Hosting Features
- Bug/issue tracking system
- Database management system
- Designated spot for documentation
- Discussion forums for teams
- Mailing list service for teams
- News/announcement feature
- Project management software
- Survey module for team developers
- Task management software
- Trac for development teams
- Web space for project
- Wiki for development team

Forge Policies
- Is itself ad-free
- Allows use of anonymous ftp by projects
- Provides public, remote API into project data
- Requires approval for hosting
- Provides a directory or list of all projects
- Provides DOAP descriptions of each project
- Does not require payment for FLOSS projects
- Allows only FLOSS projects to host there
- Allows use of gravatar for user profiles
- Provides OpenID for login
- Has a commercial or paid option
- Provides shell access

Project Artifacts Available at Forge
- Bug tracker archives are publicly available
- Forum archives are publicly available
- Mailing list archives are publicly available
- Project-level news stream or rss activity
- Project reviews are publicly available
- To-do list or task manager list is publicly available

- Wiki change history pages are publicly available
- Wiki pages are publicly available

Project Metadata Available at Forge
- How active this project is (ranking)
- Administrators on this project
- Project's development status
- Environments project is designed for
- One or more external URLs for project
- Project can take donations via the forge
- Project's intended audience
- Persistent internal URL for each project
- Date of last release
- License(s) the project is released under
- List of members on this project
- Operating system the project is designed for
- Programming languages used to write project
- Textual description of the project
- When project was registered on this forge
- Project activity statistics
- Tags describing this project
- Topic of this project
- Translations for this project

Revision Control Available at Forge
- Arch access for teams
- Bazaar access
- Software to support code reviews
- CVS access
- DARCS revision control system
- Git access
- Mercurial access
- Microsoft Team Foundation Server
- Subversion access

These lists of the features and artifacts that can be collected have provided some of the basis for what FLOSSmole has collected over time and from what forges. For example, we found that RubyForge provided some of the same data as SourceForge, so we began to collect data from there in 2006. When RubyForge closed down and was succeeded by RubyGems, we began to collect data from this new forge starting in 2015. Subsequently, we have been able to use entity matching techniques to connect approximately 5000 RubyForge projects to their new RubyGems identities (Squire 2016). Other forges we have collected from include Google Code, Tigris, ObjectWeb, Alioth, Launchpad, Github, and GNU Savannah.

While forges are attractive as a quick source of data, we quickly learned that collecting data from SourceForge was fraught with peril (Howison and Crowston

2004). For example, automated scripts required a rewrite each time the page layout of the site changed. The calculation of data item shown on a forge project page could evolve over time, complicating longitudinal studies. These issues reinforced the value of sharing the effort of collecting data. Another change during this time frame was when another research team [SRDA] (Van Antwerp and Madey 2008) made arrangements with the company running SourceForge to receive periodic data dumps, making it unnecessary to collect this data via spidering. Later, some forges created search APIs that made it much easier to collect data about specific projects. For example, currently the GHTorrent service [Gousios 2013] samples the Github project API and provides a query interface to it.

Directory Metadata

A second source of data has been directories of FLOSS projects. In the early days of FLOSS, it was important to some users to be able to find software packages that had been developed using free and open principles and which were distributed with free or open source licenses. To help keep track of what FLOSS projects existed and where to find them, several groups created web portals or directories. Examples include Freshmeat.net (later called Freecode) and the Free Software Foundation (FSF) Directory. The necessity of these directories has declined in recent years due to the ubiquity of FLOSS projects and development methodologies and better integration of project search capabilities in software package managers. Freshmeat.net itself was sold several times during the 2000s, was rebranded as Freecode in 2011, and then became read-only in June of 2014. Black Duck Software manages a directory called Ohloh, which has rebranded as Open Hub, and claims to index 672,000 projects.

Since the primary mission of these FLOSS directories is to help users find software that matches their needs, the directories are typically organized as one page per project, with each project page listing metadata such as license type and version information, where to find the project (a URL), a contact name for the project, and sometimes basic download statistics (or in the case of Freshmeat.net and Open Hub, a type of popularity metric). The FSF Directory is similar, except that it limits listings to only projects that meet the FSF's definition of free software and which are distributed with a free software license.

Open Hub is an interesting case, as it represents more of a meta-directory with facts aggregated from other sources. Specifically, Open Hub creates a page for each project, but uses data from other forges and directories to populate the facts about the project. For example, the FLOSSmole page on Open Hub seems to have most of its information pulled from an older SourceForge listing for our project. (FLOSSmole moved off of SourceForge to Google Code in 2009, and then in 2013 we moved to Github. Presumably Open Hub intends for this updated information to be manually corrected on Open Hub by a person who claims the title of project manager.) Open Hub is also different in that it attempts to suggest similar projects based on tags

or topics, and it attempts to determine other facts automatically about the project based on an analysis of the source code. For example, Open Hub tries to estimate the COCOMO development effort that went into the project or the comment density based on source lines of code. It also tries to identify the top programming languages and how recently the project has been edited and by whom. This automated analysis occasionally yields odd or amusing results, for example, the FLOSSmole Open Hub listing states that the project was written between 2004 and 2009 (apparently based on our SourceForge hosting years), was written in Ruby (it was not), and has an COCOMO model project effort of 106 years.

Collecting Data from Project Directories FLOSSmole began collecting from Freshmeat.net in June of 2005 and continued until it shut down in 2014. During this time, we collected data from Freshmeat.net 65 times. Rather than scraping the text of each project page, we used an RDF data dump provided by Freshmeat/Freecode. We collected data from the FSF Directory 36 times starting in 2007 and going through 2012. For FSF, we had to scrape the project metadata from the FSF Directory project web pages themselves, until a site redesign in 2012 caused us to stop collecting from there.

Details we collected about Freshmeat.net projects include dates (date project was added, date the project was last updated), URLs (home page, mirrors, locations of various packages), project ratings (popularity, vitality, number of subscriptions), project descriptions, screenshots, software license, release information (version numbers and dates), project authors, project dependencies, and tags to describe the project. Tags include topics for the project, user level, operating system it was designed for, and so on.

Metadata we collected from the FSF Directory varied as the site was redesigned several times, but generally we were able to collect project names and URLs, dates (release date and registration date on the site), descriptions of the project, what kind of user it is intended for, developers on the project (who and how many), categories for the project, license, related projects, and requirements (dependencies).

Because we collected from these directories so many times over so many years, we have the ability to see changes in the projects over this time period. For example, the Linux project was first listed on Freshmeat.net in 2005 with a C++ tag, but that tag was removed in 2012. The Bitcoin project was listed with several language translations in 2009, including Chinese, which was added in 2012. When the project metadata was updated on the directory, we have a record of the update and can learn about the evolution of projects in this small way.

Individual Project Website Metadata

Another important source for data about FLOSS projects is the individual project web pages themselves. Many FLOSS projects use a forge as a central location for their development activity and downloads, but many others do not. The Apache

family of projects, for example, is hosted on the Apache.org domain. Each of the 300 or so Apache projects may choose to use a non-Apache forge for organizing or mirroring code releases (Apache Spark, for example, mirrors its code releases on Github), and some FLOSS directories chose to list Apache projects. For example, Apache Mina was listed on Freecode starting in 2005. However, many of the metadata fields are blank, so if we want to find out substantive details about these projects, we need to use their actual project pages.

Collecting Data from Project Websites The task of collecting data from individual project websites means developing a separate collector for each site. Apache projects may all live on one central apache.org domain, but each project has a distinctive web presence with information about the project stored in a unique format and page structure, complicating data collection. Nonetheless, in 2013 FLOSSmole began attempting to extract several pieces of data from Apache projects, including names of developers and other contributors who have signed a contributor license agreement, developer roles on the project (Squire 2013a), and Twitter handles (Squire 2013b). In 2014 FLOSSmole was given a donation of DOAP (description of a project) data for all Apache projects by Davide Galletti. These DOAP files are optionally listed by each Apache project on its apache.org-hosted site and include structured data about each person working on the project, their roles, and the versions of the project.

Communication Archives and Social Media

Finally, over time, FLOSSmole has been collecting more and more communication artifacts from FLOSS projects. Owing to the decentralized and distributed nature of FLOSS development, most FLOSS teams prefer to communicate online, and many communities prioritize the transparency and availability of these communications to enable broad participation in the work of the team and as a form of institutional memory. Apache's own guideline for project management committees (PMCs) states "Virtually all PMC communication should happen on the dev@ list or any other appropriate public mailing list" (link: http://www.apache.org/foundation/governance/pmcs.html). The Apache Code of Conduct states "We preferably use public methods of communication for project-related messages, unless discussing something sensitive" (link: http://www.apache.org/foundation/policies/conduct.html). The level of commitment to email is similar on the Linux project, where the Linux Kernel Mailing List (LKML) receives hundreds of messages each day. Message content includes hammering out changes to procedures, sharing and critiquing code fixes, discussing priorities for future development, and so on.

In addition to email, many FLOSS projects use IRC to support users and developers. The Ubuntu distribution of Linux has 360 different discussion channels

on the freenode IRC network. Its channels are organized by topic (Xubuntu, website, news, bugs), demographic (youth, women), and geography (us, za, cn) or are specifically designated for meetings. A disadvantage of the synchronous communication in IRC is that it is unlikely that all interested parties will be online at the same time. Making logs available partly mitigates this problem. For example, the logs for the Ubuntu IRC channels are posted each day for anyone to read. Other FLOSS projects that use IRC and publicly post the logs include Django, Perl6, Bitcoin, Openstack, Puppet, and some Apache projects. Still, IRC may be more suitable for uses such as user support or quick answers to questions, rather than discussions that should involve multiple points of view.

Another communication-oriented data source we have collected is the channel topics for freenode IRC channels. Freenode is a public IRC server where anyone can create a channel about any topic. Each channel operator can write a brief description ("topic") describing the purpose of the channel. For example, the Bitcoin IRC channel on Freenode listed the following topic on April 8, 2015:

```
v0.10.0 | Bitcoin http://bitcoin.org/ | https://en.bitcoin.it/
wiki/Faq | No: pricetalk (#bitcoin-pricetalk), ads, trading
(#bitcoin-otc), begging, NOR altcoins | Web wallets can steal
your money | URLs are often SCAMS or MALWARE | All keys gener-
ated with brainwallet.org should be considered compromised
```

This description tells what software is being discussed in the channel, gives additional URLs for where to find more information, sets a few social norms ("No: pricetalk..., ads, trading..., begging..."), and gives some general admonitions about frequent problems on the channel ("URLs are often SCAMS or MALWARE"). These channel topics have been important for learning about the social expectations of a channel. For example, in many IRC chat channels, it is considered bad behavior to paste in large amounts of text or source code. Instead, IRC users prefer to trade links that point to the source code (e.g., links to a pastebin copy of the text or a JSFiddle copy of the text). In Squire and Smith (2015) we examined the proliferation of pastebin links in email mailing lists, but we find similar suggestions in IRC channel topics, as well as in the IRC channel chat texts themselves.

FLOSS researchers have studied FLOSS developer email for a variety of reasons such as to understand team culture, communication styles, decision-making practices, efficacy of bug fixing and leadership structures, and so on. To study these topics, researchers have applied a variety of techniques to the emails including social network analysis, text mining strategies, qualitative analysis of content, and quantitative analysis of the email headers. A review of email mining techniques used in FLOSS research is found in Squire (2012).

Collecting FLOSS communication archives like IRC and email can benefit other disciplines beyond software engineering and FLOSS studies. Applications where large amounts of email and IRC data will be helpful include natural language tasks, for example, learning to separate source code from human speech or performing sentiment analysis. Another area where IRC has been suggested as an interesting

application area is in text summarization, for example, in Zhu and Hovy (2005) and Sood et al. (2013). In both these papers, human-generated summaries of the IRC chat channel for the FLOSS GNUe project were used to train a computerized IRC chat summarizer. However, over time, the gold standard human-generated summaries that these papers were based on were lost, as were the original GNUe IRC messages. We rebuilt both of these data sets and made them available on FLOSSmole (link: http://flossmole.org/content/software-archaeology-gnue-irc-data-summaries).

Collecting Communication Archives When evaluating candidates for data collection, we chose to only collect IRC and email logs from projects which have publicly posted the archives and which have a tradition or expectation of public consumption of the logs. We do *not* run "log bots" that join an IRC channel and log the chats, and we do *not* join mailing lists simply to create archives of the content. However, if a project makes its logs available, we may choose to work with them, as was the case with Django, Bitcoin, Ubuntu, and the other projects mentioned earlier. For email, we have collected messages from projects hosted on the Tigris forge, from the LKML, and from Apache.

Collecting data from communication archives has all the same problems as collecting data from other sources, including changes in the data formats and page layouts. However, in collecting communication archives, we have noticed a few new problems unique to these data sources. First, the data is much messier. Text archives for software projects are replete with source code, unicode, html text, attachments (in the case of email), and other challenging data cleaning issues. Second, because communication is so central to the workings of the team, teams tend to move servers, change software, and upgrade platforms frequently. These changes lead to lapses in archives and page layout changes and occasionally mean that we can no longer collect data from a source, leaving our collection incomplete.

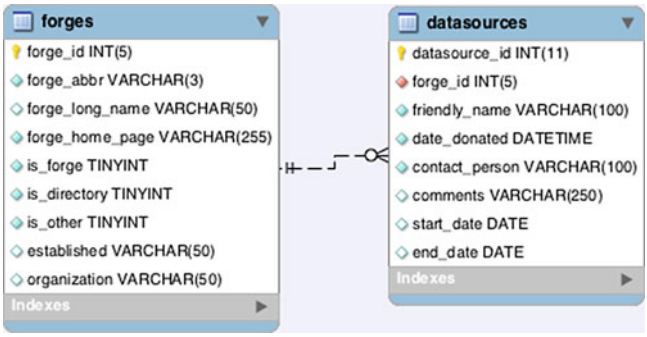
To illustrate this last issue, we describe the case of WordPress, the open source publishing platform, and its IRC channels. Beginning in 2009, WordPress created and made publicly available chat logs for several of its channels, including a general chat channel, numerous developer channels, and channels for subprojects like bbPress and BuddyPress. However, in 2014, WordPress announced that their developer chat would be moved to Slack, a privately owned real-time communication platform (link: https://make.wordpress.org/core/2014/10/29/modernising-real-time-communication/). Since this time, the chat logs for WordPress are no longer available for download or study. Other FLOSS projects that have moved to Slack as of this writing include Babel, Bootstrap, Apache Cordova (still maintains an email list), Mido.net, Socket.io, Ghost, and Bitcoin Core.

Nonetheless, many FLOSS projects still host message archives for their team communication, and FLOSSmole does collect some of these, for example, for Apache projects. There are also third-party archives for some email lists; MarkMail and Marc.info are general collections of thousands of mailing list archives. Many universities also host email archives. For example, to produce our cleaned LKML collection, we started with emails collected from the Indiana University archive.

Data Model and Data Availability

Given the array of FLOSS data artifacts that FLOSSmole has collected, it is important to understand how they are organized and stored. In this section we will explain the data model that underlies the FLOSSmole system and some of the historical reasons that it is designed in this way.

Since we are primarily a data repository that began by periodically collecting data from FLOSS forges, FLOSSmole is still organized around the concept of a "forge" and "datasource_id". Each place we collect data from is still called a "forge" (although some are directories and some are individual projects, as described above). In addition, each time we collect data from one of these places (or forges), we give that collection a new, unique number, which we call a "datasource_id". Across the entire FLOSSmole system, the one thing all the tables have in common is the idea of a datasource_id. Every piece of data is stamped with its datasource_id, indicating where the data came from, when it was collected, and who collected it. The figure below illustrates the relationship between a forge and a datasource_id.



Below are examples of the forges table and data sources table, each with one sample record shown.

Forges

Column	Sample data
forge_id	71
forge_abbr	RG
forge_long_name	RubyGems
forge_home_page	http://rubygems.org
is_forge	0
is_directory	1
is_other	0
established	2009
organization	RubyGems

Data Sources

Baudry	Baudry
datasource_id	61240
forge_id	71
friendly_name	RubyGems Nov 2015
date_donated	2015-11-06
contact_person	msquire@elon.edu
Comments	RubyGems Nov 2015
start_date	2015-11-06
end_date	2015-12-04

Storing communication raises its own issues. We have experimented with several ways to store email data, including in a standard RDBMS where one row equals one email, with the most relevant email headers parsed into columns (sender, date, message body). We have also tried creating document-oriented databases of email, where one JSON record equals one email and headers are unlimited (Squire 2013c). Storing IRC data is more straightforward. For IRC, we have created relational tables where each table is a channel and within the table one row equals one chat line. Depending on what is available in the archive, we create columns for the timestamp, the message sender, the message line, and whether the message is a system message, an action message, or a standard chat message.

In the FLOSSmole system, each collection of data is organized into databases with other, similar data. For example, our data from Freshmeat.net/Freecode has its own database with seven tables to hold the data just for these projects. The LKML email has its own database, the Apache IRC channels are grouped together in their own database, and so on. With 31 databases and hundreds of tables, we are unable to show a complete entity relationship diagram (ERD) in this paper, but all the databases and tables are described in a color-coded set of ERDs on the FLOSSmole website. The database can be queried from any MySQL database client, including from Jupyter notebooks such as those hosted by the Wikimedia Foundation. Database access instructions can be found on the flossmole.org site.

In addition to the queryable database, some of the data has been made available in raw text format as well as compressed text dumps of the MySQL database. These files are also available through the FLOSSmole website and require no sign up or special access permissions. We should point out that not everything in the MySQL database is available in the flat files or database dumps and vice versa. For instance, IRC logs started out as text files that we cleaned and imported into the database. There is thus no need to re-create text archives for IRC chat or emails that came to us as text archives. Conversely, a few of the directories on the FLOSSmole server

include files that are not in the database. One example is the files of links to IRC and email examples of insults, jokes, and profanity described in (Squire and Gazda 2015).

What Researchers Have Learned from This Data

The data collected by the FLOSSmole project has been used in several hundred published research papers on a diversity of topics and with many different research approaches. Many of these papers have been collected on the FLOSShub site (http://flosshub.org/). In this section, we describe a few of these papers to give a sense of how FLOSSmole data have been used.

A number of researchers have used FLOSSmole data as a convenient source of basic metrics about FLOSS development, e.g., to determine the number of projects on Github (Biazzini and Baudry 2014), to create a list of projects on SourceForge (Mockus 2009), or to obtain a list of developers to survey to Kina et al. (2016). Others have used FLOSSmole for general project metadata to augment other data collection, e.g., project descriptions to determine the application domain of a project whose source code is being analyzed (Zhang et al. 2013).

Other researchers have used the data more intensively to examine relations between project features. For example, Samoladas et al. (2007) used machine learning techniques, including classification rules and decision trees, on data from 1 month to predict the metrics for the following month. Often, the independent variable in these studies is some measure of project success. Crowston et al. (2006) proposed a framework for FLOSS success measures and used data from FLOSSmole data to operationalize three example measures. Wasserman and Ashutosh (2007) used FLOSSmole data to assess a project readiness for business use. Rossi et al. (2010) examined how download rates (a measure of user uptake) were affected by new releases. They found that different projects showed different impacts of a new release, suggesting different usage patterns for the software. Other researchers have examined the state of development of a project. For example, Schweik and English (2012) proposed a way to use the data to classify projects as successful or abandoned. Piggot and Amrit (2013) had good success using eight project metrics to predict the state of a project.

Researchers have also used FLOSSmole data to look in more detail at the software development processes within projects. For example, one theme in this work is the nature of leadership in FLOSS projects. Valverde et al. (2006) compared FLOSS development to the self-organizational processes of wasp colonies and noted that a small number of developers seemed to stand out from the rest. Taylor et al. (2008) analyzed authorship patterns proposing the concept of author entropy. Crowston et al. (2010) used social network analysis to identify project leaders, but

noted that the technique also seemed useful for identifying interesting periods in the project history, e.g., when leadership was changing. Corona and Rossi (2013) extended the analysis to look at "linchpin" developers, those who connect across multiple projects.

Finally, a few projects have looked at the issues in connecting data across multiple repositories. Howison (2008) and Iqbal et al. (2012) suggested that Semantic Web technologies could be useful in making the linkage between projects hosted in different places. Rezende et al. (2010) suggested the use of data mining techniques on these repositories to identify clusters of similar projects.

Challenges

In this section we describe some of the ongoing challenges we have with maintaining and growing FLOSSmole as a data repository. Rather than just detailing the frustrations of collecting and storing large-scale data from a group of evolving online resources, our focus in this section will be to outline some of the new developments within the FLOSS ecosystem itself that challenge our ideas of what data elements are important and how they should be used.

Challenge: Availability and Integration of the Data

As we stated earlier, the goal of FLOSSmole is to collect and store data about free, libre, and open source software and its development. We began collecting data from software forges because they were an easy place to get a lot of data in a relatively organized format. However, the forge landscape itself has changed enormously since our project began in 2004, and this has impacted where we get data and how much we can get. Our initial vision was very much driven by the visible success of SourceForge. SourceForge, launched in 1999, has always tried to be a one-stop place where developers could collaborate to create FLOSS and users could download the software. The model of SourceForge is designed around the idea of a project, each of which has an owner, contributing developers, and users. SourceForge hosts mailing lists, bug reporting software, databases, wikis, downloadable content, documentation, and discussion boards and provides source code revision control services. As of this writing, the site claims to host 430,000 FLOSS projects and 3.7 million registered users. The rate of new project registrations on SourceForge from 1999 through 2014 is shown in Fig. 7.1. Projects that were created without any textual description are likely throwaway test projects or spam projects, and these are shown in blue. The recent decline in the volume of registrations is clearly visible.

A more recently successful site is Github. Launched in 2008, Github currently has over 14 million users and 35 million public repositories (similar to projects), completely eclipsing the size and scale of SourceForge in a relatively short amount

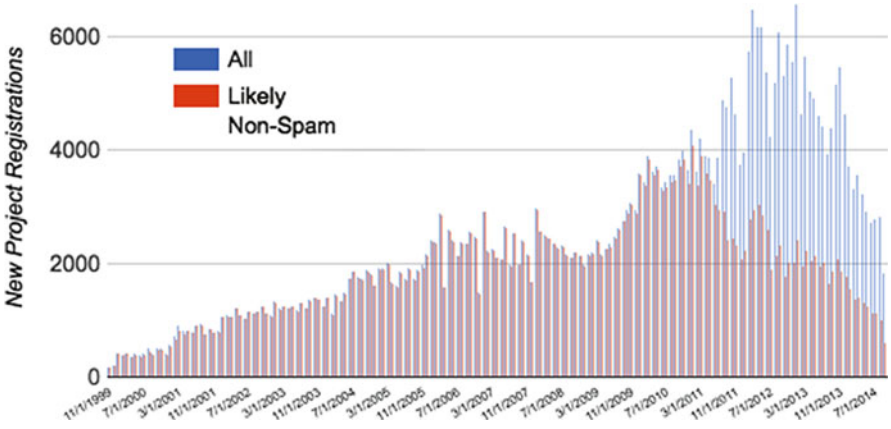


Fig. 7.1 New project registrations on SF (1999–2014)

of time. Rather than project, Github is organized around the idea of a user who can create public or private repositories/projects and who may be a member of teams. The services provided to users and teams are limited to revision control, bug reporting, and a few basic documentation services. Github is growing at such a rate that makes it difficult to create a timestamped "static" mirror of its data. Instead, Github provides an API to allow searching of its projects and users. Numerous projects have been started to provide a queryable interface to Github data, for example, the very popular GHTorrent.org and GithubArchive.org. Because Github is now the most popular repository for *any* project – whether or not it is explicitly FLOSS – this has caused a mass exodus from other, formerly important FLOSS forges. These older forges have languished or disappeared entirely (as was the case with Google Code, Berlios, and JavaForge), as projects abandon them and move to Github instead.

At the same time as the mass exodus away from the FLOSS forges and onto Github, we notice a shift in the communication strategies for projects as well. Most projects used to have an email mailing list and an IRC channel where most development discussion happened. In many cases, these were hosted on the same forge as other artifacts, enabling one-stop data collection. Currently we see instead a multiplication of communication media, spread across multiple services, e.g., an increase in the use of Slack for discussions (as noted above for WordPress), Pastebin, and other paste-type sites to share code snippets and an increase in the use of Stack Exchange sites for developer support (Squire 2015). Indeed, Github has completely opted out of providing communication channels, with the exception of comment threads on bugs. Many projects use Github only for source code hosting, preferring to glue together a group of services to manage the rest of their communication and coordination tasks. For example, they may use Github for posting code, Bugzilla for bug reporting, Slack for real-time communication, an email list hosted at a university or private corporation for asynchronous communication, and so on.

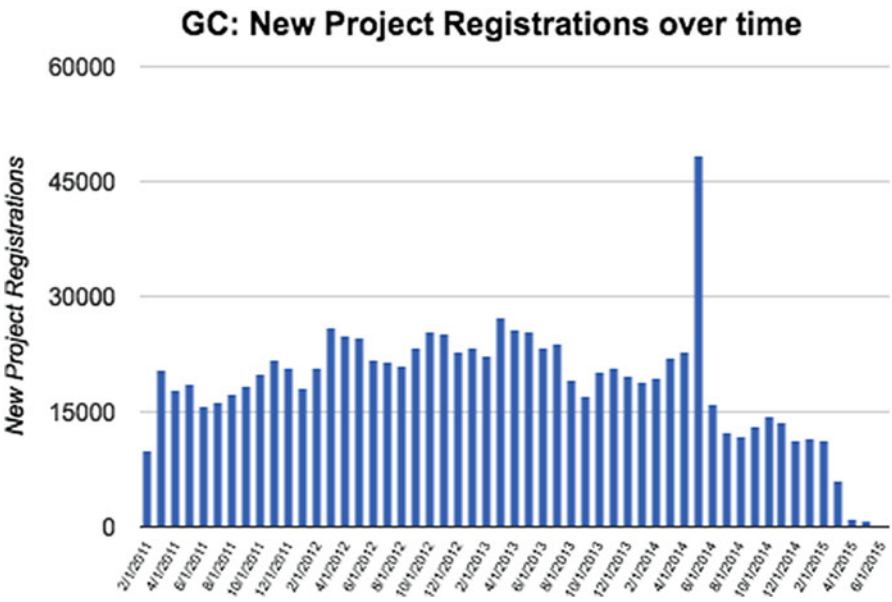
The FLOSSmole model of writing spidering scripts to collect all the metadata and communication artifacts for a number of projects all hosted on the same FLOSS forge has thus become more difficult to apply. Multiple spiders are needed as well as techniques to identify connections across services. Slack is particularly problematic for us since it affects the visibility of developer communication to researchers (and other interested outsiders). Although there is a Slack-created archive of chat messages, it is limited for our use in three important ways. First, the archives are not available to anyone who is not invited to use the Slack channel, which stretches the definition of a "public" archive. (For example, the address for the WordPress archives is https://wordpress.slack.com/archives, but it is only available to registered, logged in Slack users.) Second, the Slack message archive is created inside a JavaScript front end, and while it is mostly browsable and searchable, it is not easily scrapable or downloadable unless you are the team owner. Third, if you are not the channel owner, attempting to create an archive of messages is likely a violation of Slack's Terms of Service.

In summary, the constant shift of projects from one forge to another and a piecemeal approach to choosing services means that understanding the evolution of a project over its lifetime is much more difficult. Compiling a story about a given software project will necessarily involve integrating data from multiple sources, including deciding which of many options is the canonical location for a project as it forks and changes over time.

As an example, consider the Bitcoin project. As the highest profile and arguably the most innovative cryptocurrency ever invented, Bitcoin started as a private code base in 2008 with email messages sent to the cryptography mailing list. It then moved as an open source project to SourceForge in January of 2009, then in October of 2009 moved development to Github while keeping a mailing list on SourceForge. It was entered into the Freecode directory in 2010 and archives of its IRC channel were created. In 2011 leaders started a new mailing list for developers at LinuxFoundation.org, and the Bitcoin Stack Exchange site was started in 2011 as well. To build a data-oriented history of this, one project will require, at a minimum, integrating data from these nine disparate, unconnected sources. We have not even considered the various forks of the project or the connection libraries (for example, Ruby connectors, Java connectors). When we consider that there are hundreds of thousands of FLOSS projects, the job of collecting "all" the data seems indeed daunting.

Challenge: Validity of the Data

We mentioned spam projects in passing earlier, but weeding out real projects from spam projects continues to be a challenge. The figure below shows new project registrations on the now-defunct Google Code FLOSS hosting site.



The sharp spike in May 2014 represents an enormous influx of new projects, most of which were created as empty pages showing no details except for links to advertising. One year earlier, in May 2013, Google had announced that projects hosted on its services could no longer host downloadable files since so many of the files were malware and copyrighted files. Google Code closed down entirely the following year, in 2015. As we saw with the SourceForge graph shown earlier, the problem of spam projects, fake projects, and spam is not limited to Google Code alone. The SourceForge graph presented earlier with non-spam projects shown in red was our attempt to describe accurately the state of new project registration on that site, in light of a high number of spam projects. The challenge for researchers using any data from software forges is to disambiguate the good projects from the bad ones, and this is no different with FLOSSmole data, Github data, Stack Exchange data, or any other data that comes from a site that allows open contributions.

A further complication with Github is that projects on Github are not guaranteed to be FLOSS, unlike some earlier FLOSS-only forges. Indeed, only about 20% of the projects on Github are explicitly released with a free or open source license (many have no license at all). This lack of license information makes it difficult to figure out which projects to collect and which to ignore if we are interested in studying FLOSS development specifically.

Challenge: Providing Analyses of Data

The discussion above has focused on collection of raw data about FLOSS development. Part of the FLOSSmole original mission statement was to produce and distribute derived datasets and analyses. FLOSSmole collects data, and it also provides some cleaned, augmented, or processed versions of that raw data. For example, the Linux Kernel Mailing List email messages are provided in raw format and in several cleaned formats as well: with headers removed, with source code removed, with extraneous signature lines and whitespace removed, and so on. The RubyGems data includes processed analyses for what the first release date was for each gem in the collection. The Apache project data has been parsed to extract the roles and corporate affiliations of each contributor on each team.

Still, the amount of processed analyses stored in FLOSSmole could be vastly increased. To do so will require a larger number of donated data sets or a larger core team of developers. Furthermore, the level of analysis could be increased, i.e., moving from observational trace data to measures of concepts of theoretical interest (e.g., derived measures of leadership as in Crowston et al. 2010). As it stands, people who use FLOSSmole data are encouraged to share it, and FLOSSmole offers to create the data model and storage infrastructure for FLOSS-related, processed data sets. However, few researchers have taken up this offer. An ongoing challenge is convincing people to take advantage of the infrastructure to store and publicize the results of their own work. The increased interest in data sharing to promote replicability and to meet funding agency requirements may provide a further impetus.

Challenge: Usability of the Data

A further challenge facing FLOSSmole is making the collected data more useful to and usable by researchers. A first step is simply describing what is there so researchers can use it. The website does this at a basic level of indicating what datasets exist, what they contain, and their provenance. A complication with the latter is in describing how data have evolved over time. However, there is a more challenging issue of explaining what the data mean and how they can be connected to interesting research questions. To fully address this challenge will require better documentation and instructions provided by a strong user community using the datasets.

Challenge: Sustainability of the Project

A final challenge is the sustainability of the project: the effort needed to continue to collect data and make it available. The project currently has no financial support

and instead relies on donated time and facilities. Data collection and processing is largely done by one project leader and her students. The current system of file and database storage and access relies on donated facilities at Syracuse University. These facilities fit the size of the community using the data, at least for the moment, but a new CIO at Syracuse could start charging for computer/storage/network time. To address the former challenge requires building more of a community around data collection, so the project is not overly reliant on one person. (Ironically, FLOSS projects face the same challenge to their sustainability.) To address the latter, data could be moved to another host, perhaps one with a mission that is more specifically aligned with the open source or open data movement.

Possibilities for the Future

The use of FLOSS development methodologies, licensing strategies, and business models continues to grow. While many early FLOSS research projects sought to define and explore this perplexing new phenomenon, today studying FLOSS is an accepted – and in many cases an expected – part of many diverse research programs. As such, FLOSSmole has gradually moved away from its early "mile wide, inch deep" approach to data collection. Where we once tried to collect as much (admittedly shallow) data from a software forge as possible, this is no longer feasible or interesting in the age of Github's exponential growth through 35 million forked repositories. Therefore, several years ago, FLOSSmole began taking a depth over breadth approach to some projects, for example, building in-depth studies of particular FLOSS communities of interest. Our initial collection of the RubyForge repository, and its follow-on successor the RubyGems repository, has morphed into a 12-year long evolutionary history of thousands of projects as they grew and then moved across the two sites. Our collection of data on the Apache family of projects includes diverse artifacts such as board meeting minutes, email, IRC chat, Twitter handles, and roles of people on each project as well as, in some cases, the corporations for whom they work.

As we mentioned earlier, FLOSSmole has also begun to focus more on text artifacts, particularly communication media such as IRC chat and email discussions. One challenge for FLOSSmole in the future is that these communication channels generate an enormous amount of traffic, especially on very popular projects. Many FLOSS projects have mailing lists with hundreds or thousands of messages per day. The Linux Kernel Mailing List, for example, includes 2.4 million emails starting at its inception in 1995 through today. The Openstack developer IRC chat channel has 390,000 lines of chat dialogue over 5 years, and the Openstack user channel has created twice that many messages. The Ubuntu community created its general IRC channel in 2004, and it has logged 30 million lines of dialogue in that time. Daily traffic on the Ubuntu chat channel is typically in the range of 100–800 messages per hour, with many days topping 1000 messages per hour. This creates minor storage

questions for FLOSSmole, but more importantly it creates an attention issue. Do we have enough time and attention to collect, clean, and store everything? Should we attempt to collect the communication artifacts of many projects, or just for a few? Should we only collect the developer channels or the user channels as well? Even if we create space to collect the data, do we have the time and attention to analyze it all? Right now we are collecting a small number channels and list archives, from a variety of different sources, and using these to learn what is important and what data other researchers find most useful. Projecting which projects will be important or interesting in the future is certainly a challenge.

As we collect these richer data sources, our hope is that they will enable researchers to tackle more difficult questions about software evolution, community structure, and so on. Example questions that we hope our data will address include: What issues are important to the developer and user communities? How does decision-making happen? How do the leaders emerge? What is the organizational culture of the group? How do all of these things change over time? How do various communities compare to each other? These questions probably cannot be answered with empirical, artifact-based research alone, but richer data sources can certainly illuminate some aspects of each line of inquiry.

References

Biazzini, M., & Baudry, B.. (2014) "May the fork be with you": Novel metrics to analyze collaboration on GitHub. Proceedings of the 5th International Workshop on Emerging Trends in Software Metrics – WETSoM 2014, New York, pp. 37–43.

Booch, G., & Brown, A. W. (2003). Collaborative development environments. *Advances in Computers, 59*, 1–27.

Conklin, M. (2006). Beyond low-hanging fruit: Seeking the next generation in FLOSS data mining. In *Proceedings of the 2nd IFIP WG 2.13 International Conference on Open Source Systems*. Como: IFIP, Elsevier. June 8–10. pp. 47–57.

Corona, E. I. M., & Rossi, B. (2013). Linchpin developers in open source software projects. In *Proceedings of The IASTED International Conference on Software Engineering* (pp. 8). Innsbruck: ACTA Press. February 11–13.

Crowston, K., Howison, J., & Hala, A. (2006). Information systems success in free and open source software development: Theory and measures. *Software Process–Improvement and Practice, 11*(2), 123–148.

Crowston, K., Wiggins, A., Howison, J. (2010). Analyzing leadership dynamics in distributed group communication, 43rd Hawaii International Conference on System Sciences (HICSS 2010), Honolulu, Hawaii, USA, pp. 1–10.

Gousios, G. (2013). The GHTorent dataset and tool suite. In *Proceedings of the 10th Working Conference on Mining Software Repositories* (pp. 233–236). IEEE Press. May 18.

Howison, J. (2008) Cross-repository data linking with RDF and OWL. 3rd Workshop on Public Data about Software Development (WoPDaSD 2008), pp. 15–22.

Howison, J., & Crowston, K. (2004). The perils and pitfalls of mining SourceForge. In *Proceedings of the International Workshop on Mining Software Repositories (MSR 2004)* (pp. 7–11).

Howison, J., Conklin, M., & Crowston, K. (2006). FLOSSmole: A collaborative repository for FLOSS research data and analyses. *International Journal of Information Technology and Web Engineering, 1*(3), 17–26.

Iqbal, A., Cyganiak, R., Hausenblas, M. (2012). Integrating FLOSS repositories on the Web, Technical Report #2012-12-10 of the Digital Enterprise Research Institute (DERI) at the National University of Ireland, Galway.

Kina, K., Tsunoda, M., Tamada, H., & Igaki, H. (2016). Analyzing the decision criteria of software developers based on prospect theory. 23rd IEEE International Conference on Software Analysis, Evolution, and Reengineering (SANER 2016) at Osaka, 03/2016.

Mockus, A. (2009). Amassing and indexing a large sample of version control systems: towards the census of public source code history. 6th IEEE Working Conference on Mining Software Repositories, May 16–17.

Piggot, J., & Amrit, C. (2013). How healthy is my project? Open source project attributes as indicators of success. IFIP Advances in Information and Communication Technology Open Source Software: Quality Verification, Volume 404, Berlin, Heidelberg, pp. 30–44.

Rezende, H. R., & Esmin, A. A. A. (2010). Proposed application of data mining techniques for clustering software projects. INFOCOMP Special Edition (pp. 43–48).

Rossi, B., Russo, B., & Succi, G. (2010). Download patterns and releases in open source software projects: A perfect symbiosis? *Open Source Software: New Horizons, 319*, 252–267.

Samoladas, I., Bibi, S., Stamelos, I., Sowe Sulayman, K., Deligiannis, I. (2007). A preliminary analysis of publicly available FLOSS measurements: Towards discovering maintainability trends. 2nd Workshop on Public Data about Software Development (WoPDaSD 2007).

Schweik, C. M., & English, R. (2012). *Internet success: A study of open source software commons*. Cambridge, MA: MIT Press.

Sood, A., Mohamed, T. P., Varma, V. (2013). Topic-focused summarization of chat conversations. Advances in Information Retrieval, Volume 7814 of the series Lecture Notes in Computer Science. 35th European Conference on IR Research, ECIR 2013, Moscow, Russia, March 24–27. Springer. pp. 800–803.

Squire, M. (2012). How the FLOSS research community uses email archives. *International Journal of Open Source Software and Processes, 4*(1), 37–59.

Squire, M. (2013a). Project roles in the apache Software Foundation: A dataset. In *Proceedings 10th Working Conference on Mining Software Repositories (MSR2013)* (pp. 301–304). San Francisco: IEEE. May 18–19.

Squire, M. (2013b). Apache-affiliated Twitter screen names: A dataset. In *Proceedings 10th Working Conference on Mining Software Repositories (MSR2013)* (pp. 305–308). San Francisco: IEEE. May 18–19.

Squire, M. (2013c). A replicable infrastructure for empirical studies of email archives. In *Proceedings 3rd International Workshop on Replication in Empirical Software Engineering (RESER 2013*) (pp. 43–50). Baltimore: IEEE. October 9.

Squire, M. (2015). "Should we move to Stack Overflow?" Measuring the utility of social media for developer support. In *Proceedings of 37th International Conference on Software Engineering (ICSE-2015)* vol. 2 (pp. 219–228). Florence: IEEE. May 20–22.

Squire, M. (2016). Data sets: The circle of life in Ruby hosting, 2003-2015. In *Proceedings of the 13th International Conference on Mining Software Repositories (MSR2016)* (pp. 452–455). Austin: ACM. May 15.

Squire, M. & Gazda, R. (2015). FLOSS as a source for profanity and insults: Collecting the data. In *Proceedings of 48th Hawai'i International Conference on System Sciences (HICSS-48)* (pp. 5290–5298). Hawaii: IEEE. January 6–8.

Squire, M., & Smith, A. (2015). The diffusion of pastebin tools to enhance communication in FLOSS mailing lists. In *Proceedings of the 11th International Conference on Open Source Systems (OSS2015)* (pp. 45–57). Florence: IFIP, Elsevier. May 16.

Taylor, Q. C., Stevenson James, E.., Delorey Daniel, P., Knutson Charles, D. (2008). Author entropy: A metric for characterization of software authorship patterns, 3rd Workshop on Public Data about Software Development (WoPDaSD 2008), pp. 42–47.

Valverde, S., Theraulaz, G., Gautrais, J., Fourcassie, V., & Sole, R. V. (2006). Self-organization patterns in wasp and open source communities. *IEEE Intelligent Systems*., 03/2006*, 21*(2), 36–40.

Van Antwerp, M., & Madey, G. (2008). Advances in the Sourceforge Research Data Archive. In *Workshop on Public Data about Software Development (WoPDaSD) at The 4th International Conference on Open Source Systems*. Milan.

Wasserman, A., & Das, A.. (2007). Using FLOSSmole data in determining business readiness ratings. 2nd Workshop on Public Data about Software Development (WopDaSD 2007).

Zhang, F., Mockus, A., Zou, Y., Khomh, F., and Hassan Ahmed, E. (2013). How does context affect the distribution of software maintainability metrics? Proceedings of the 29th IEEE International Conference on Software Maintainability.

Zhu, L. & Hovy, E. (2005). Digesting virtual "geek" culture: The summarization of technical internet relay chats. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL '05)* (pp. 298–305). Stroudsburg: Association for Computational Linguistics..

Kevin Crowston is a Distinguished Professor of Information Science in the School of Information Studies at Syracuse University. He received his Ph.D. (1991) in Information Technologies from the Sloan School of Management, Massachusetts Institute of Technology (MIT). His research examines new ways of organizing made possible by the extensive use of information and communications technology. Specific research topics include the development practices of free/libre open source software (FLOSS) teams and work practices and technology support for citizen science research projects, both with US National Science Foundation support.

Megan Squire is Professor of Computing Sciences at Elon University. Her primary research focus is in the collection and analysis of software development artifacts, specifically those from free, libre, and open source software (FLOSS) projects. She co-founded FLOSSmole to collect and analyze this FLOSS data and to provide the results back freely to the FLOSS research community. The data and results from this work are readily applicable to research in a variety of fields, including software engineering and the social sciences.

Chapter 8
Teaching Students How (Not) to Lie, Manipulate, and Mislead with Information Visualization

Athir Mahmud, Mél Hogan, Andrea Zeffiro, and Libby Hemphill

Introduction

Information visualization is an appropriate method for displaying big data. However, the selective nature of information visualization can lend itself to portraying information that lies, manipulates, and misleads. Most people are unlikely to question these deceiving practices, so the ethical burden is placed on the designer of the visualizations. The purpose of this chapter is to help students avoid some of these inadvertent deceptions and make purposeful choices about the claims their visualizations make. This would entail gaining a deep understanding of the appearance of deceitful visualizations. There are debates in the visualization field regarding data and the goal of visualization and even about the data involved and data discussions mirror those about big data broadly. This is not surprising considering that even data can be inaccurate or deceitful, so why would we expect the visualizations that rely on this data to always be accurate?

There are numerous perspectives as to the purpose of visualization. One of these is the use of visualizations to represent a concept or idea that is already found in the physical world, such as maps (MacEachren 1995). Electronic, satellite-based maps have taken map visualization to an entirely unchartered territory, as they can communicate where a user is in real time, even during motion. A major

A. Mahmud (✉) • L. Hemphill
Illinois Institute of Technology, Chicago, IL, USA
e-mail: athir.mahmud@gmail.com

M. Hogan
University of Calgary, Calgary, AB, Canada
e-mail: mhogan@ucalgary.ca

A. Zeffiro
McMaster University, Hamilton, ON, Canada

© Springer International Publishing AG 2017

S.A. Matei et al. (eds.), *Big Data Factories*, Computational Social Sciences,
https://doi.org/10.1007/978-3-319-59186-5_8

perspective for the use of information visualization is also to communicate or explain information that provides a narrative (Tufte 1997). Unfortunately, a great deal of information that is being communicated is done so by those who have little or no expertise in the information they are conveying. The rise of the data scientist job title to its near rockstar-like reputation comes burdened with the risk that individuals who understand the numbers lack the knowledge to translate the significance of those numbers and how they connect with one another. Information related to medical conditions is not entirely similar to data about educational testing or social media use. The ability to translate this information and make sense of the information found within big data ultimately results in a narrative. This brings us to the last purpose of information visualization. That is, to tell a story based on the available data (Yau 2011). Without a story, visualizations are nothing more than pretty pictures. But those stories can sometimes also lack a compelling one. We conclude this chapter by providing a framework of exercises to encourage students to reflect on the practice of designing good visualizations and avoiding some of the common traps of poor visualization design.

Since the publication of J.C.B Grant's *An Atlas of Anatomy* in 1943, medical illustrations have served to simplify complex information systems for medical practitioners and for surgeons in particular. Medical illustrations simplified information for those who needed to see quickly and precisely the details of an operation without being overwhelmed by the totality of information available through photography, text descriptions, or live petri dish comparison. Illustrations were drawn by artists, often women who only had rudimentary access to the field of medicine to which they were greatly contributing. And yet despite the lack of official medical training, it was illustrators who were best suited to convey the necessary data that was then instrumentalized by (Western medicine trained) surgeons. The illustrator's' ability to convey only what was important anatomically, and from a purely visual point of view, was an essential (and still largely under documented) part to the evolution of medical practices. And this is because of their *expertise in communicating information* rather than in medicine per se.

We use this anecdote to make the point that while the tools for visualization have evolved – from pen to paper to big scale data sets – the kind of visual thinking and translating required to make sense of data, or make data make sense, remains an expertise in its own right. Taking "raw" data and "cooking" it (Bowker 2006; Gitelman 2013) means creating a narrative based on analysis. But of course this is more difficult than it sounds because data is never "raw," objective, or neutral and the methods and systems for data analysis aren't either. Machines don't pump out more accurate data than humans; humans are their programmers. Machines can, however, handle much more data than a brain on its own can. However, whether more is better is also a matter of the questions you pose of the machine, the deployment of its algorithms and the logics of code and dispositions that underlie it (Bivens 2015; Easterling 2014; Gitelman 2013). While some approaches are more rigorous than others, we make the case here that in designing informational visualizations, it's important to know what methods you are using, what data is, and its cultural and political significance at different junctures.

What is data? When thinking about the collection, application, interpretation, and manipulation of data, you might want to first consider the source of your data. You want to ask yourself: Where does my data come from? In what context was the data collected, and by which mechanisms? Data is always incomplete because it's collected in a particular way and often within fixed parameters. Especially with large data sets, data may feel total or objective, but no data exists outside of the realm of a (human) logic which generates the conditions of its existence. Over 40 years ago, John Berger made a case for the inherent power of images. In *Ways of Seeing* (1972), Berger described the ways in which visual representations of the world are inherently ideological. We argue that the same holds true for representations of data. Charts are not facts. Graphs are not truths.

If we return briefly to the question of medical illustrations, we might ponder whether illustrations are less "truthful" than photographs? Perhaps the question itself becomes problematic if it suggests that only one "truth" is possible or that attempts at objective and accurate depictions are always inherently neutral, rather than a product of specific processes and world views. What can we learn about data representation from this example that can be applied to newer, bigger data sets? Why is it important to recognize data as never raw and as always cooked? This kind of recognition of data's "loadedness" becomes the critical gaze necessary for designers working on data visualizations and representations because "data visualizations wield a tremendous amount of rhetorical power" as Catherine D'Ignazio (2015) argues in her blog post, "What would feminist data visualization look like?". Expanding on D'Ignazio's (2015) post, we ask: What shapes our current visual landscape and its politics of representation? What's at stake? Why is developing a critical gaze on data visualization so important, now more than ever?

Many media scholars are currently addressing the invisible labor driven by and driving online communications. Subjects include humans that function as code or algorithms, such as those behind commercial content moderation (Irani 2015; Roberts 2016), social media filtering mechanisms (Gillespie 2016; Seaver 2013), and gendered software programming (Bivens 2015; D'Ignazio 2016). These are only a few examples among a growing body of research that reflects on the blurring of machine and human work or, more specifically, workers who labor in big data (Gillespie 2016). According to these critics and others, what is of concern is the pervasive invisibility that shapes the contours of what users believe possible of mediated communications. In simpler terms, invisibility is intentional and serves to make users complicit by surrendering their agency in two ways.

One way is that users come to understand algorithms that process large data sets as too complex for them, the layperson, as well as offering up an objective calculation from which to make decisions (Kitchin 2014). The second way is that users believe big data necessarily drive algorithms. This means, for example, surrendering to prompts that suggest, for instance, a new book or movie based on our browsing histories, precisely the kind of counsel we would likely resist from a random stranger making the same suggestion! In this way, we relinquish part of the decision-making process because we believe in big data, as networked and collective data, to be fundamentally informative and unbiased. Or perhaps

we want it to be so, to alleviate some of our indecisions or to assure us that we are making a good decision based on more than our own whims and desires. No less, this is important because it also informs the context in which we perceive data and its visual representations. And because we are largely led to believe that data is neutral – big data as a source ripe with expansive insights if only the right questions are asked of it – we could critique the results while forgetting to ask more fundamental questions about its origins.

Good data visualizations become even harder to assess as good design tends to erase its process and render invisible its guiding structures. As Anna Munster (2009) explains, structures that dictate the organization and aggregation of data remain largely invisible and are thought of as the imperceptible objects of data visualization. Similarly, Stalbaum (2004) observes that "same representations might exist in different terminal states (as either data or information) on a larger conveyor belt of ubiquitous digital processing" (par. 11). Further yet, Alexander Galloway (2011) argues that by developing a critical gaze onto data visualization, we come to see that what is represented are the underlying structures and rules that produce the data. But whose rules? While we all inadvertently generate data by daily uses of mobile devices and web-based applications, fewer of us have the capacity to collect and store data, and fewer still have the infrastructures and tools to interpret it (Manovich 2012).

In short, we are not all equal in our relationship to big data or what can be made from it. And why this matters is because those who make the rules determine the game and come to control its players (Harding 1996). Data is never simply inert material to be converted into knowledge – it embodies culturally specific principles of organization and representation – it both informs and is informed by the knowledge of a society. These tensions between organization, representation, and information mirror debates within the broader visualization literature. We turn now to three popular perspectives on the purpose of visualizations: to represent (MacEachren 1995), to communicate or explain (Tufte 1997), and to tell stories (Yau 2011).

The use of information visualization for the purposes of representation can be seen in various forms of visuals. The use of maps is one well-known method of representing geographic locations and their relation to one another. Time, an abstract concept, is typically represented by clocks in the physical world, but can be represented by a visualization as simple as line graphs or as complex as a streamgraph. Different visualization techniques can be combined to offer other forms of data. For example, a weather radar map combines both geographic maps and time, with visual indicators as to the direction of wind or the movement of precipitation. These are forms of visualization that most people know and trust. The trouble with unethical representation usually stems from using visualization techniques that are not appropriate for the data. Many of these can be seen in every infographics where something like the outline of a cow is used to determine the

percentage of people who eat beef every day, five times a week, four times a week, three times a week, two times a week, once a week, or never. While the creative aspect is apparent, the use of an organically shaped image makes it very difficult to determine the correct percentages.

Another perspective proposes the goal of information visualization is to produce an aesthetically pleasing, appropriate manner in which to communicate and explain different kinds of data, particularly big data (Keim et al. 2013). Big data is large, unruly, and difficult to understand. When left in its raw form, big data cannot provide an explanation of what the information really means. Edward Tufte (1997) justified the need to transform quantities evidence into visual designs in *Visual Explanations: Images and Quantitates, Evidence and Narrative*. He explains that creating visualizations that deceive is actually "disinformation design," (p. 55) where the display is nothing more than a form of magic. If one purpose of visualization is to communicate, then misinformation essentially accomplishes the opposite.

Another goal of information visualization is to tell a story about data that is otherwise difficult to understand, because it is originally in the form of raw data and numbers. The storytelling perspective may be fraught with the greatest amount of ethical concern. This is because visualizations can easily be manipulated or modified to tell the story that most suits the data analyst or designer. Furthermore, this is possible even if the data is accurate or unflawed. If a conservative news program is interested in portraying liberal politicians in a negative light (see Fox News examples), simply truncating axes could change the entire narrative of the data presented.

Information that is presented visually is perceived to have greater reliability than if the information was presented alone, regardless of the integrity of the data source. However, just as information visualization can be used to make this considerable amount of data clearer, more manageable, or more interesting, it can also be used to manipulate viewers. This can be accomplished, whether advertently or inadvertently, in a variety of ways. Unintentional visualization manipulation may more likely be due to a lack of visual or data literacy. There is a necessary level of skill required to understand that a streamgraph is more appropriate than a pie chart, for example.

Manipulation of users can also be achieved simply as a result of information visualization's nature to only portray some information, sometimes at the expense or exclusion of other, often critical information. Information visualization can easily be used to manipulate the message of that data, even critically examined and describe data, because the very nature of information visualization is that it must exclude and occlude some of the information available. In the sections that follow, we propose activities you can use with students (or on your own) to improve their abilities to ethically and attentively design visualizations that accomplish a variety of representation, communication, and storytelling goals.

Suggested Activities

So what is a good instructor to do? This section first provides a list of questions students and instructors can use as prompts for developing data visualizations, to account ethically and with political intent the way they communicate information. We encourage students to reflect on these questions as a means of making their design work more transparent and also as a tool to critically assess data visualization more broadly. Following the questions, we address three common dimensions in visualization and suggest activities for teaching about them: physical space, time, and comparisons. We suggest that you use these questions to guide a discussion of each of the exercises outlined below.

Maps

Together, the recent pushes to open government data and to make mapping software more accessible have contributed to the rise in popularity of mapping visualizations. Maps have an advantage of being intuitive to many people because of their experiences driving, taking public transportation, and watching or reading national and international news. We are used to seeing scaled down representations of physical space, and we're familiar with common boundaries (e.g., states) and conventions (e.g., political party colors, "you are here" markers).

In this section, we use the Centers for Disease Control's Community Health Status Indicators (CHSI) to combat obesity, heart disease, and cancer[1] to illustrate how decisions about data and semiotics impact what functions map visualizations serve. We have chosen this data because it highlights questions about narratives and political uses of visualizations. According to the CDC, the data set is explicitly designed for public health professionals and the public broadly, and its curation was supported by federal funds. Each of the maps below was created in Tableau Public.[2] We use the same variable – "Prim_Care_Phys_Rate," a county-level measure of the number of primary care physicians per 100,000 people – to determine the colors to display in each map. The PCP rate ranges from 44 to 123 with a median of 84. Primary care helps prevent disease, and access to primary care is associated with more equitable heath distributions (Starfield et al. 2005). States have different policies around health insurance and primary care, especially under the Affordable Care Act. When looking at the three maps below, think about how health policy makers and activists might use these maps to argue for changes in health policy. At whom are these maps aimed? How might they be misunderstood? What narratives do they create?

[1] https://catalog.data.gov/dataset/community-health-status-indicators-chsi-to-combat-obesity-heart-disease-and-cancer.

[2] https://public.tableau.com/s/.

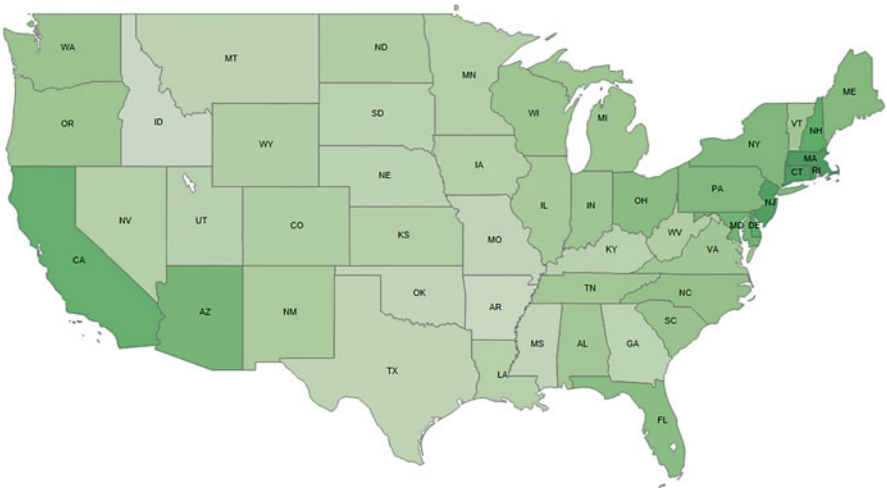


Fig. 8.1 Primary care physician rates in the USA – *darker colors* indicate more PCPs per 100,000 residents

In Fig. 8.1, we use a standard map of the 48 contiguous states and use a single-color gradient to indicate the primary care physician rate in each state. As the rate of PCPs increases, the value of the green color becomes darker. MacEachren (1995) refers to maps like this one as "designative" (p. 247). The colors designate explicit properties of the underlying shapes – here sequential greens indicate the rate of PCPs in a state. Use the questions in Table 8.1 to guide a discussion that compares these two maps.

Figure 8.2 presents the same data on the same map but using a diverging color scheme. In this case, the color ranges from deep red to deep green, and the gradient is centered around the median PCP rate of 84 PCPs per 100,000 residents. By using a two-color gradient, especially red and green, the map here is an "appraisive" (MacEachren 1995, p. 248). The red-green gradient evokes a metaphor of relative safety where red means "danger" and green means "safe." In doing so, the gradient imbues this map with judgments about the states. Red implies that states are dangerous while green implies relative safety.

Given the familiarity of maps, we are also able to adjust the ways physical space is represented without losing the affordances of map metaphors. Regional comparisons are still accessible, for instance, and we can still recognize how a particular point on map corresponds to a place in space. In Fig. 8.3 we use a tile map instead of a standard projection. In this map, each state, regardless of its geographic size, is represented by a tile of the same size. The tiles are colored according to the same appraisive gradient used in Fig. 8.2. The tile map also makes it easier to include Alaska and Hawaii because the map is divorced from some of the space constraints of a projection map. Projection views crowd out geographically small areas such as Rhode Island and the District of Columbia. How do the functions of

Table 8.1 Prompts for designing data visualizations

1. Data collection
(a) What kind of data are you working with?
(i) That is, numbers, correlations, timelines, equations, quotations, stories
(b) What were the motives for collecting the data?
(i) That is, research, play, software testing
(c) What is the origin or source of your data? Who or what collected the data?
(i) That is, researchers, NGOs, artists, corporations, governments, law enforcement
(d) Through what mechanism and by what method was the data initially collected?
(i) That is, insurance purposes, academic research, advertising revenue, API
2. Data analysis
(a) What are the contours of the data? What is included and what is left out? What are their limits?
(i) That is, geography, time, gender
(b) How does a particular tool, technique, or method inform how data is collected?
(i) That is, biometric data, surveys, user-generated content, crowdsourced
(c) What are the tool, technique, or method's orientations? How does it position, frame, or encode the data within a particular logic or world view?
(d) How is information about the process of data collection made visible? How is it hidden? How might absences be accounted for?
(e) At what point in the data collection process are you interjecting your analysis?
(f) How are the data analyzed? What questions are guiding the analysis? What assumptions are embedded in those questions?
3. Data visualizations
(a) What tools (software, program, or practices) are used to generate the data visualizations? What are the aesthetics that underlie or guide the tools? How might these shape interpretations of the visualizations?
(b) What's kind of narrative does the visualization create? Who is it aimed at? Is it a widely accessible narrative?
(c) How is the data visualization framed?
(i) That is, explanatory text, link, contact information
(d) How might the visualization be understood? How might it be misunderstood?
(e) What are the potential social and political implications of the visualization?
(i) That is, surveil, represent, call to action
(f) Where will the visualization be featured? How will it be featured?
(g) Who stands to benefit from the way the data is represented?
(h) What stage(s) of a project is the visualization supporting?
4. Self-reflexive analysis
(a) Who is your design and research team comprised of? How do these factors inform the choice of data sets, data collection, interpretation and storage?
(i) That is, ethnicity, authority, age, rank, gender
(b) Was anyone funded to collect the data? Is anyone profiting, financially or otherwise, from the data?
(c) What assumptions and biases are embedded in the data? How might these impact the data's use?
(d) Who or what is subsumed by the data? Who or what is privileged by the data? What are the possible social and political effects of these data entitlements?
(e) What can other data collectors, analysts, publishers, and users learn from your work?
(f) Does the work contribute to data ethics, equity, sovereignty, and access

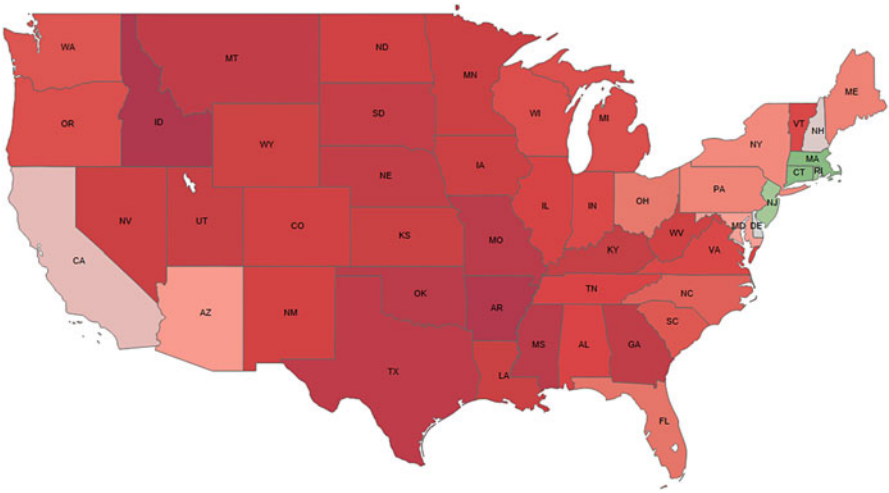


Fig. 8.2 Primary care physician rates in the USA

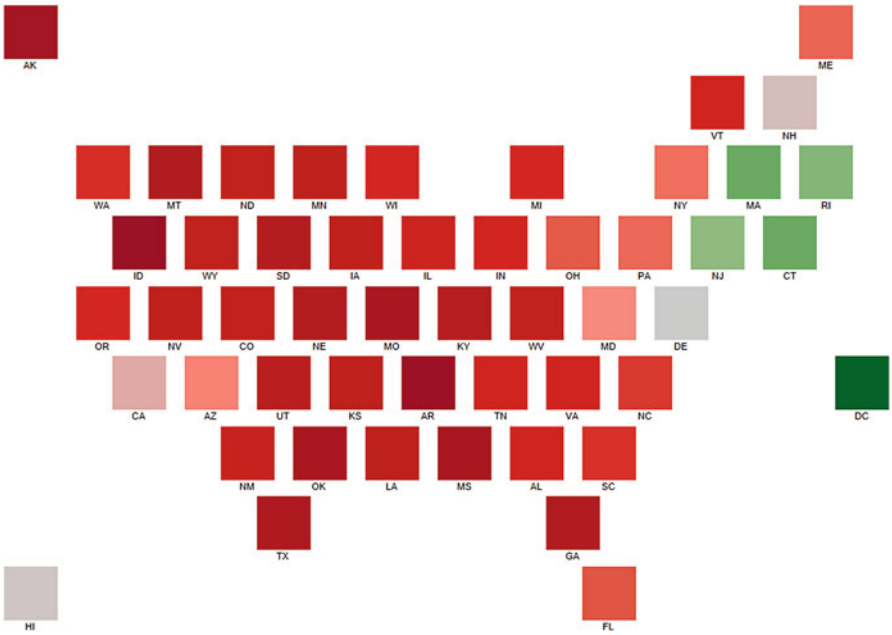


Fig. 8.3 Primary care physician rates in the USA using a tile map

the tile map differ from those of the projection map? What kinds of narratives are easier to see in a tile map? What differences does a tile map occlude?

This example of mapping PCP rates across the USA just scratches the surface of issues visualization designers face when mapping data. Here are some additional resources for designing mapping visualizations.

Comparisons

There are a number of situations that can benefit from or can make better sense when there is some level of visual comparison. This type of data is generally categorical, with very clear distinctions between those categories, offering room to compare and contrast. Comparisons can be made in a variety of ways, but need a technique in which to differentiate the items that are being compared. This can include the use of one or any combination of color, symbols, patterns, and text.

There are a number of types of visualizations that are used to compare data. In a sense, one of the purposes of information visualization is to compare data, so nearly all forms of visualization compare *something*. A bar chart can be used to compare the individual test scores of a group of students in a particular class. A line graph may be used to compare a company's monthly spending over a year's time. Even a simple pie chart can be used to compare how much of available funds and individual spends on different expenses within his or her budget.

Take, for example, the concept of the "red state" versus the "blue state," in terms of designated political symbolism in the USA. Ordinary citizens can usually recognize that this refers to a state's tendency to vote for either Republican or Democratic candidates, particularly in presidential elections. However, it should be noted that colors are not universal symbols, so they can even be problematic for use in comparison when used internationally. For instance, in the United Kingdom, the color blue is associated with the Conservative Party, so the connection between color and ideology breaks down.

The use of symbols represents visually differences in information. Comparisons between males and females are also accomplished through the use of the color, but also through the use of male and female symbols. There is also a symbol used to indicate transgender individuals. The individual shapes of traffic signs, sometimes without their corresponding colors or patterns, hold symbolic meaning that many experienced drivers would easily recognize. There are even symbols that are used internationally, such as methods of transport, including airplanes, taxi cabs, buses, escalators, and stairs. When colors and symbols cannot be used, patterns, such as dashed or dotted lines, stripes, polka dots, or zigzags, can be assigned to represent data.

However, comparisons can be made using a visualization that is as simple as a column chart and such as Fig. 8.4, a Fox News visual where comparisons are being

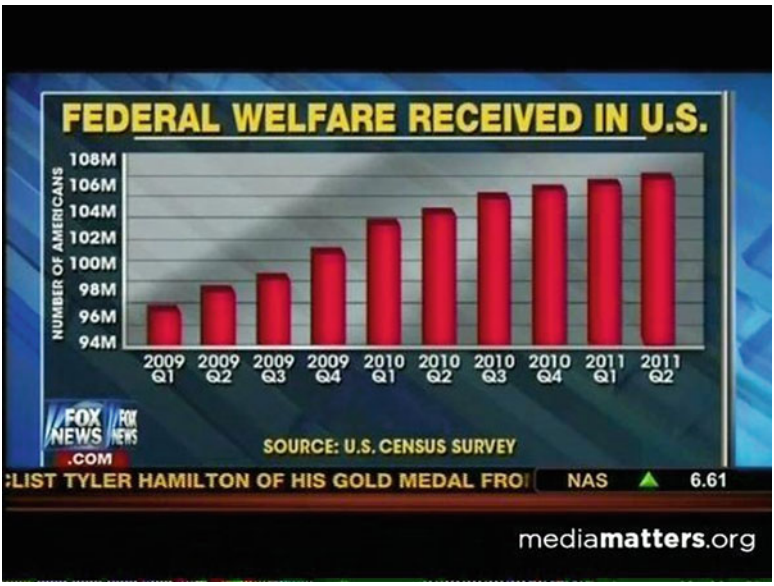


Fig. 8.4 Fox News truncated y-axis

made between yearly quarters for federal welfare dollars that were distributed in the USA. On first glance, the visualization appears harmless, if not informative, and indicating an increase in spending on welfare. However, upon further inspection, it becomes clear that the y-axis was truncated to result in a more exaggerated perception of welfare increases.

But assuming that the data is correct, if the visual was modified where the y-axis was not truncated or, at minimum, expanded to include more numbers, the graph would look closer to the graph found in Fig. 8.5. The variations become markedly less significant and appear to have a much smaller increase in the number of dollars spent.

Another highly effective tactic is to choose not to label one or more of the axes presented in a visualization. This can be problematic for many reasons, including allowing different variables or data types to be compared at the same. In Fig. 8.6, the lack of a y-axis provided the creator of the visual with nearly free range to determine how it appeared. Just glancing at the visual would have a viewer believe that that the baseball pitcher's speed has drastically fallen between 2012 and 2013 – nearly by half. However, closer inspection of the numbers above each column reveals that difference in speed is only 2 miles per hour on average. Given that the pitch described is a knuckleball, the visualization may be trying to highlight the slowness of the 2013 velocity – slower knuckleballs are better – and so exaggerate the difference to make Dickey look like he's improved significantly.

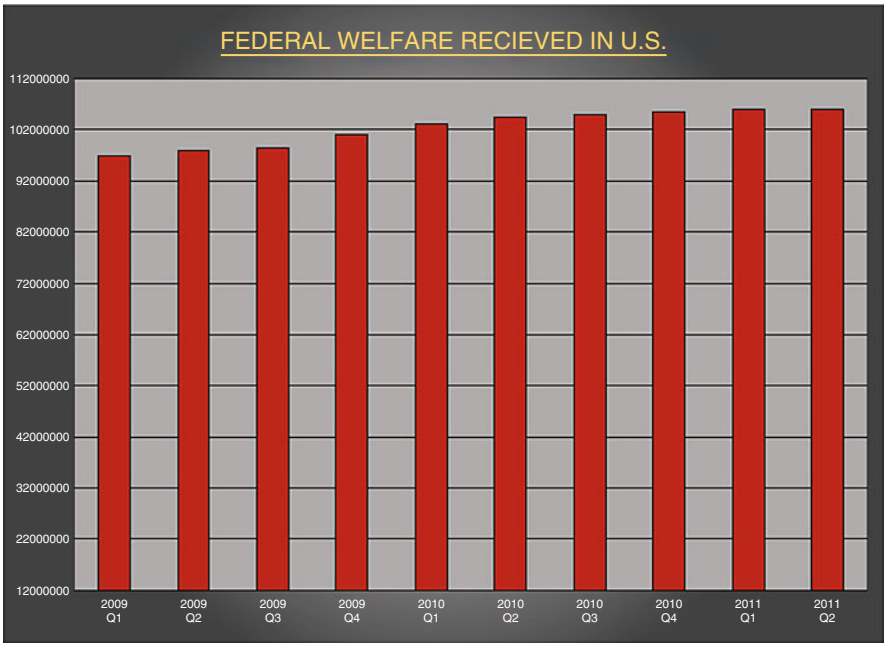


Fig. 8.5 Modified visualization of welfare data

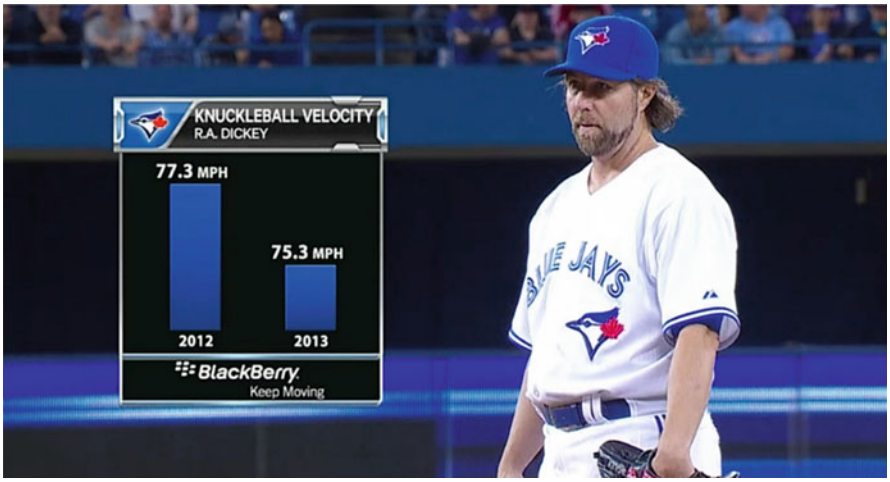


Fig. 8.6 Visualization lacking y-axis

Conclusion and Additional Resources

Information visualization is a powerful approach to portraying data, particularly big data. Without a visual story, massive quantities of data are nearly impossible to understand, but visualization can help people form a greater understanding of the numbers behind the pictures. However, with this powerful tool comes a great deal of responsibility. The responsibility to visually display data is one that rests on the shoulders of the visualization designer. However, because data scientists and designers can assume that people generally have little data or visual literacy, some designers have used this to their own advantage. Perhaps they have a story that needs to be told, they have been paid to create the false story, or perhaps they are lacking in the very form of literacy they have been tasked with demonstrating.

The types of visualizations available for use span a wide variety of information and come in many forms. We have only been able to discuss a very small number of these, including maps, comparisons, and temporal visualizations.

Using information visualization to lie and mislead is not a novel endeavor. Therefore, there are many resources available to demonstrate good, as well as bad, visualizations. The list below is a good, albeit not exhaustive, list of resources.

- MacEachren, A. M. (1995). *How Maps Work: Representation, Visualization, and Design*. Guilford Press.
- Yau, N. (2011). *Visualize This: The FlowingData Guide to Design, Visualization, and Statistics*. John Wiley & Sons
- http://www.informationisbeautiful.net/visualizations/what-makes-a-good-data-visualization
- http://www.gooddata.com/blog/5-data-visualization-best-practices
- http://viz.wtf
- http://data.heapanalytics.com/how-to-lie-with-data-visualization
- http://www.visualisingdata.com/2014/04/the-fine-line-between-confusion-and-deception

Ultimately, the goal of information visualization is to represent, communicate, and to tell stories. Large data sets, or big data, contain meaningful information that is not apparent to the everyday individual. The challenge, however, is to create visualizations that do not give dishonest, misleading information, regardless of the purpose of the visualization. Making certain that students are aware of these pitfalls will make them better prepared to avoid making the mistakes and decisions that result in misleading or dishonest visualizations.

Parts of this paper were written in consultation with Antonia Hernández and Corina MacDonald of mat3rial.com. This material is based upon work supported by the National Science Foundation under Grant No. IIS-1525662.

References

Berger, J. (1972). *Ways of seeing*. London: Penguin.

Bivens, R. (2015). The gender binary will not be deprogrammed: Ten years of coding gender on Facebook. *New Media & Society*, 1461444815621527. https://doi.org/10.1177/1461444815621527.

Bowker, G. C. (2006). *Memory practices in the sciences*. Cambridge, Mass: MIT Press.

D'Ignazio, C. (2015). What would feminist data visualization look like? Retrieved June 16, 2016, from https://civic.mit.edu/feminist-data-visualization.

D'Ignazio, C. (2016). A primer on non-binary gender and big data. Retrieved June 16, 2016, from https://civic.mit.edu/blog/kanarinka/a-primer-on-non-binary-gender-and-big-data

Easterling, K. (2014). *Extrastatecraft: The Power of Infrastructure Space*. Verso.

Galloway, A. (2011). Are some things unrepresentable? *Theory, Culture & Society, 28*(7–8), 85–102. https://doi.org/10.1177/0263276411423038.

Gillespie, T. (2016). Facebook trending: It's made of people!! (but we should have already known that). Retrieved from http://culturedigitally.org/2016/05/facebook-trending-its-made-of-people-but-we-should-have-already-known-that/.

Gitelman, L. (2013). Raw data is an oxymoron. Retrieved June 16, 2016, from https://mitpress.mit.edu/books/raw-data-oxymoron.

Harding, S. (1996). Feminism, science and the anti-enlightenment critiques. In A. Garry & M. Pearsall (Eds.), *Women, knowledge, and reality: Explorations in feminist philosophy* (2nd ed., pp. 298–320). Boston: Unwin Hyman.

Irani, L. (2015). Difference and dependence among digital workers: The case of Amazon mechanical Turk. *South Atlantic Quarterly, 114*(1), 225–234. https://doi.org/10.1215/00382876-2831665.

Keim, D., Qu, H., & Ma, K. L. (2013). Big-data visualization. *IEEE Computer Graphics and Applications, 33*(4), 20–21. https://doi.org/10.1109/MCG.2013.54.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. Thousand OAks: SAGE Publications.

MacEachren, A. M. (1995). *How maps work: Representation, visualization, and design*. Guilford Press.

Manovich, L. (2012). Trending: The promises and the challenges of big social data. In M. K. Gold (Ed.), *Debates in the digital humanities* (pp. 460–475). Minneapolis: University of Minnesota Press.

Munster, A. (2009). Data undermining: The work of networked art in an age of imperceptibility. Retrieved from http://turbulence.org/project/data-undermining/#.

Roberts, S. (2016). Digital refuse: Canadian garbage, commercial content moderation and the global circulation of social media's waste. *Wi: Journal of Mobile Media*. Retrieved from http://ir.lib.uwo.ca/commpub/14.

Seaver, N. (2013). Knowing algorithms. Presented at the Media in Transition 8, Cambridge, MA. Retrieved from http://nickseaver.net/s/seaverMiT8.pdf.

Stalbaum, B. (2004). An interpretive framework for contemporary database practice in the arts. Presented at the College Art Association 94th annual conference, Boston, MA. Retrieved from http://paintersflat.net/database_interpret.html.

Starfield, B., Shi, L., & Macinko, J. (2005). Contribution of primary care to health systems and health. *Milbank Quarterly, 83*(3), 457–502. https://doi.org/10.1111/j.1468-0009.2005.00409.x.

Tufte, E. R. (1997). *Visual explanations: Images and quantities, evidence and narrative*. Cheshire: Graphics Press.

Yau, N. (2011). *Visualize this: The flowing data guide to design, visualization, and statistics*. Chichester: Wiley.

Chapter 9
Democratizing Data Science: The Community Data Science Workshops and Classes

Benjamin Mako Hill, Dharma Dailey, Richard T. Guy, Ben Lewis, Mika Matsuzaki, and Jonathan T. Morgan

Nearly every published discussion of data science education begins with a reflection on an acute shortage in labor markets of professional data scientists with the skills necessary to extract business value from burgeoning datasets created by online communities like Facebook, Twitter, and LinkedIn. This model of data science—professional data scientists mining online communities for the benefit of their employers—is only one possible vision for the future of the field. What if everybody learned the basic tools of data science? What if the users of online communities—instead of being ignored completely or relegated to the passive roles of data producers to be shaped and nudged—collected and analyzed data about themselves? What if, instead, they used data to understand themselves and communicate with

Dharma Dailey, Richard T. Guy, Ben Lewis, Mika Matsuzaki, and Jonathan T. Morgan contributed equally to this work.

B.M. Hill (✉)
Department of Communication, University of Washington, Seattle, WA, USA
e-mail: makohill@uw.edu

D. Dailey
Department of Human Centered Design and Engineering, University of Washington, Seattle, WA, USA
e-mail: ddailey@uw.edu

R.T. Guy • B. Lewis
Microsoft, Redmond, WA, USA
e-mail: richardtguy84@gmail.com; benjf5@outlook.com

M. Matsuzaki
Department of Biostatistics, University of Washington, Seattle, WA, USA
e-mail: mika@post.harvard.edu

J.T. Morgan
Wikimedia Foundation, San Francisco, CA, USA
e-mail: jmo25@uw.edu

© The Author(s) 2017

S.A. Matei et al. (eds.), *Big Data Factories*, Computational Social Sciences,
https://doi.org/10.1007/978-3-319-59186-5_9

each other? What if data science was treated not as a highly specialized set of skills but as a basic literacy in an increasingly data-driven world?

In this chapter, we describe three years of work and experimentation around a vision of *community data science* that attempts to explore one set of answers to these "what if?" questions. This work has primarily involved designing curriculum for, and then running, five series of 4-day workshops, plus three traditional university courses taught to masters students. We have used these workshops and classes to explore the potential of, and challenges around, this vision of democratized data science. We begin by framing our goals and approach in terms of similar and analogous efforts. Next, we present our philosophy and design goals. With this background, we describe the structure of the curriculum we have developed. Finally, we use data from several pre-session, within-session, and post-session surveys to discuss some of the promises and limitations of our approach.

Background

Data Science Education

There is little doubt that, driven by surging interest in the power and potential of "big data" for business, data scientists have found themselves in high demand (Manyika et al., 2011). *Harvard Business Review* has called "data scientist" the "sexiest job in the twenty-first century" (Davenport and Patil, 2012), and several reports have pointed to massive shortages of data scientists in labor markets. For example, in their widely cited report published by the McKinsey Global Institute, Manyika et al. (2011) suggested that the United States is already facing a massive shortfall of skilled data scientists that will only be aggravated in the coming years. In 2014, Dwoskin (2014) suggested that there were 36,000 advertised, but unfilled, positions for data scientists in more than 6,000 firms.

In response, a whole series of education programs have been created, or rebranded, in what West and Portenoy (2016) have described as a "data science gold rush in higher education." Using a dataset of more than 100 degree-granting programs in related spaces collected by North Carolina State University's Institute for Advanced Analytics,[1] West and Portenoy point to dozens of new programs created in a matter of years in the United States alone.

Although there is no consensus—either in popular accounts or among data scientists educators—on exactly what such programs should cover (Davenport and Patil, 2012; Miller, 2013; Gellman, 2014), there is some agreement that data scientists should be able to collect and integrate datasets and conduct analyses using some combination of programming, statistical modeling, and data mining

[1] http://analytics.ncsu.edu/?page_id=4184 (https://perma.cc/6MKH-7KVY)

techniques. Similarly, there is consensus that a critical skill for professional data scientists is the ability to ask and answer questions of substantive interest and to be able to clearly communicate their results (Davenport and Patil, 2012).

End User and Conversational Programmers

Although not all descriptions of data science involve social media, many of the most widely cited accounts of the rise of data science focus on the massive growth of datasets of online behavior from sites like Facebook, LinkedIn, Google, and Etsy (Manyika et al., 2011; Dwoskin, 2014). The absence of any mention of users of these websites from these discussions of data science is striking. Although left largely implicit, the role of end users in these accounts is to produce data and, ultimately, have their behavior shaped by the output of algorithms. Of course, as evidenced by the quantified self-movement (Wolf, 2010; Nafus, 2016; Neff and Nafus, 2016), at least some users of these systems are likely interested in the data created and stored by these systems.

Data analysis is often pointed to as a classic example of *end user programming*—commonly defined as the authoring of code by nonprofessional programmers (Nardi, 1993; Jones, 1995). Intriguingly, as data science has grown into an established professional practice itself, the potential emerges for *end user data science*. Through web application programming interfaces (APIs) created to facilitate user access to personal data from online communities, the infrastructure already exists to provide users with structured data about themselves and their friends from many of the most widely used social computing systems. That said, this access is almost only ever taken advantage of through apps with preset interfaces and dashboards. What remains missing is widespread access to the knowledge and skills to facilitate *end user data science* using currently available data.

Recent research has suggested that learning to program can be understood as a valuable tool even among users who never engage in programming. A study by Chilana et al. (2015) showed that students from majors like management with no intention to engage in programming of any sort expressed a strong interest in learning to program so that they could speak effectively with programmers they might work with. In a follow-up survey of non-programmers in a large multinational software company, Chilana et al. (2016) found that nearly half of their respondents (42.6%) had invested time in learning to program and that over half of these individuals were what they called "conversational programmers" who were interested only in improving technical conversations and their own marketability. To the extent that it is increasingly common for nonprofessional data scientists to encounter data scientific analyses, being exposed to the basic tools of data science may be seen as useful for these conversational data scientists with no intent to engage in analysis themselves.

To extend the metaphor to programming one final time, it is worth considering how, over the last several decades, computer science educators have explored what curriculum might best serve the goals of teaching nonprofessional programmers. To cite one example, Mark Guzdial and Andrea Forte have published a series of papers that reported on, in various ways, an attempt to develop, deploy, and evaluate curriculum teaching programming to noncomputer science majors (Guzdial, 2003; Guzdial and Forte, 2005; Forte and Guzdial, 2005). The degree to which this type of curriculum might differ from attempts to teach conversational programmers has been described as an open issue by Chilana et al. (2016). We know of no attempts to develop curriculum or explore pedagogical approaches around end user and conversational data science.

Democratizing Data Science

To the extent that data science is powerful and provides its practitioners with the ability to understand and affect behavior, it can be understood as politically important to make access to these tools more widespread. Although statistics are much less solid than they are in more established fields, there is evidence that data scientists are overwhelmingly white and overwhelmingly male. Though women, minorities, people with disabilities, and veterans are underrepresented in STEM fields generally, they remain most underrepresented in the fields that data science draws upon most strongly: computer science, math, and statistics.[2]

One important approach to reducing inequality in participation used in feminist critiques of computer science is to attempt to remove systematic barriers to participation. Margolis and Fisher (2001) famously use the metaphor of unlocking clubhouses to describe the goal of breaking down these systematic barriers to interested women in computing communities. A second approach involves designing new forms of participation that appeal to wider audiences. Buechley and Hill (2010) use the metaphor of building new clubhouses to evoke the idea that computing can be reimagined to appeal to women uninterested by computing as it is typically framed. Buechley and Hill argue that this approach can broaden participation in computing. Although there are almost certainly many systematic barriers to participation in data science that affect members of underrepresented groups, imagining data science as practiced by the large majority of people uninterested in careers as professional data scientists is the first step on the path of "democratizing" data science in the ways suggested by Buechley and Hill.

There have been a series of efforts to involve users of online communities in data science. The most famous and common techniques are citizen science projects. The citizen science model, made famous by Galaxy Zoo (Raddick et al., 2007),

[2] http://www.nsf.gov/statistics/2015/nsf15311/digest/nsf15311-digest.pdf (https://perma.cc/74E5-T4YJ)

Zooniverse (Smith et al., 2013; Simpson et al., 2014), and eBird (Sullivan et al., 2009; Wood et al., 2011), is similar to "crowdsourcing" where participants' role is active and intentional but also limited to a handful of typically low-level and repetitive tasks. In citizen science, participants act as sources of distributed labor and human computation (Howe, 2006). Like crowdsourcing, task execution is distributed, but the tasks of posing questions and performing analyses remain the exclusive domain of the platform operators and the "real" scientists (Benkler, 2016).

A smaller body of work has explored the potential of involving online communities in participatory data analysis where both task selection and execution are distributed. There are a number of attempts to support data analysis through participatory and social data visualization on the web (e.g., Heer et al., 2007; Viegas et al., 2007; Wattenberg and Kriss, 2006; Luther et al., 2009). Although powerful, these systems are often restricted to particular datasets provided by researchers or to a set of predefined types of visualizations or analyses. For example, users of these systems are often unable to create new variables in ways that are a basic part of most data scientists' work. Another interesting approach occurred on the Reddit online community through an experimental research process used by Matias (2016). In his study of a large social mobilization in Reddit, Matias discussed initial results and worked with participants to refine models and hypotheses. Although users were deeply involved in the process of hypothesis construction, they still relied on an academic researcher with access to programming and statistical knowledge and skills to carry out tests. Both social visualization systems and Matias's work are limited by their desire to involve users without also asking them to learn new technical skills.

Perhaps the most clear attempt to democratize data science in the way we have articulated is a system by Sayamindu Dasgupta (Dasgupta, 2016; Dasgupta and Hill, 2016, 2017). Deployed in the Scratch programming community (Resnick et al., 2009), Dasgupta's system provides programmatic access to data about activity in the Scratch community to each member. Dasgupta documented the way that Scratch's young users used the system to enthusiastically analyze their own data in ways that were powerful, unanticipated, and empowering. Dasgupta's system is limited both in the analytic tools it makes available and in the depth and scope of data provided. That said, the level of enthusiasm shown by users of the system, and the creativity these users displayed, is deeply inspiring. Like Dasgupta, our goal is to move one step beyond both citizen science and participatory hypothesis testing to give users of online communities the ability to ask and answer their own questions (*end user data science*) and to build the skills to engage with other analysts and analyses (*conversational data science*).

Toward that end, we designed a series of workshops and courses. In designing, teaching, and evaluating this curriculum, we were motivated by three broad questions. First, what are the essential skills for end user and conversational data scientists? Second, what would a curriculum teaching these skills involve? Finally, how would one evaluate attempts to democratize data science? We describe the work we have done in our workshops to explore potential answers to these questions over the rest of this essay.

Philosophy and Pedagogy

The philosophy informing our pedagogical approach is primarily influenced by Margolis and Fisher's (2001) seminal work on breaking down barriers to the participation of women in computing, Lave and Wenger's (1991) theory of legitimate peripheral participation, and Papert's (1980) concept of constructionism. From Margolis and Fisher, we draw a commitment to broadening participation in data science. From Lave and Wenger, we draw a commitment to the idea of authentic learning environments and the ability to learn through apprenticeship-like relationships. From Papert, we draw the idea that knowledge can be constructed through the creation and manipulation of knowledge in a social environment.

Broadening Participation

The first pillar of our community data science approach is the goal of broadening participation. We seek to broaden participation along several dimensions including not only the kinds of academic fields or professional backgrounds of participants but also demographic characteristics including gender and race. Many other approaches to teaching data science require existing programming or statistical experience. For example, the Software Carpentry and Data Carpentry workshops seek to attract participants with undergraduate-level programming experience (Wilson, 2014). We target absolute beginners. Indeed, one central criterion for making acceptance decisions for our workshops and classes is that applicants have no previous programming experience. This has an additional benefit of ensuring that participants begin with a similar skill level.

Meaningful participation in STEM requires successful negotiation of cultural, social, and symbolic elements of STEM fields (Joshi et al., 2016). Therefore, we strive to create an inclusive environment that considers several factors known to influence inclusiveness in STEM. For example, signifiers of masculine tech culture such as Star Trek posters have been shown to inhibit participation by women. Conversely, more neutral "ambient" signifiers such as nature posters do not inhibit anyone's participation (Cheryan et al., 2009). Toward this end, we have intentionally hosted all of our workshops and classes outside of the engineering buildings at the University of Washington campus. We have made attempts to recruit and encourage women and people of color to act as mentors, lead sessions, and give lectures. Inclusiveness is also influenced by the kinds of examples one uses, and our curriculum emphasizes working with data about people.

Finally, we have sought to offer our workshops at times, and at a cost, that makes participation by diverse groups of people possible. For example, we have scheduled our workshops on evenings and on Saturdays to make it possible for participants with full-time jobs to attend. So far, we have been able to offer all of our workshops at no cost to participants. Similarly, we have built our curriculum entirely around tools, APIs, and datasets that can be installed and used for free.

Project-Based Construction

A second pillar of our approach is a strong emphasis on project-based construction and authenticity. Although we do not entirely eschew more traditional lecture-based pedagogy, the bulk of our workshops and classes involves participants' programming on their own computers. Even during lectures, all participants are encouraged to program using their own computers by repeating the programming constructs being demonstrated by instructors and modifying them in ways that interest them.

The decision to have individuals program on their own computers reflects a strong commitment to creating authentic experiences (Lave and Wenger, 1991). We strongly believe that participants in our workshops and classes should program using the tools that we use in our own work as end user and conversational data scientists. When we teach individuals to use APIs, we have them create API keys and engage directly with real APIs. Although this leads to challenges and unpredictability around the setup related to heterogeneity of participants' devices, it also turns data science into something that happens directly on each participant's computer. When the workshops end, participants leave with all the software necessary to continue engaging in data science.

We ensure that less than half of any session is dedicated to more traditional lecture-based teaching. Instead, participants spend the majority of their time in the sessions writing software and analyzing data. We encourage participants to program and analyze data the way we do—by modifying existing code and by searching sites like Stack Overflow for error messages, recipes, and solutions to problems. This approach encourages people to wrestle with many of the real issues brought up by data analysis in ways that make critical engagement a central part of the process (Ratto, 2011). For example, when we teach about APIs, participants deal with questions about the degree to which APIs are owned or controlled by companies.

Learning Communities

A final pillar of our approach is the idea that learning happens through collaborative construction of knowledge in convivial social environments. In ways that are inspired by both Lave and Wenger's (1991) apprenticeships and Papert's (1980) samba schools, we attempt to maximize one-on-one interactions between beginners and more skilled data scientists. Concretely, this involves recruiting a large number of skilled data scientists to serve as "mentors." We try to keep to a four-to-one student-to-mentor ratio. Over the course of running the workshop series five times, we have observed that the mentors who are most reliably effective at helping learners solve their problems often come from nontraditional engineering backgrounds. Most encouragingly, we have found that many of the most effective mentors were originally introduced to data science through previous iterations of the workshops and classes.

Excellent mentors embody a warm environment by helping participants solve the problems they are facing in ways they will be able to replicate and build upon when they are working on their own rather than trying to teach "their way" or the "right way" to do something. A low student-to-mentor ratio enables opportunities for extensive one-on-one coaching. This is especially helpful for beginners since their ability to troubleshoot a problem can be brittle and because troubleshooting can be stressful and frustrating (Estrada and Atwood, 2012).

A sense of belonging is another factor that has been demonstrated to influence the inclusiveness of STEM participation (e.g., Good et al., 2012). Providing lunch—the workshops biggest expense by far—is a time-honored way to foster informal interactions and an important component of how we help to foster social support for participants. During lunch, participants often debrief with each other over the morning workshop while getting to know each other and mentors. For these reasons, we also encourage and support meet-ups and learning sessions outside of the formal workshops and classes.

Community Data Science Workshops

In early 2014, we designed a set of 4-day workshops in Seattle, Washington, that aimed to answer the three questions we raised in our background section while attempting to adhere to the philosophy and pedagogy laid out above. For the initial set of workshops, we drew both inspiration and some initial curriculum from the Boston Python Workshops (BPW)[3] and Software Carpentry[4]—two curricula with which we had experience. In particular, we leveraged BPW's detailed Python setup instructions and introductory Python programming curriculum. Additionally, the way we structure our daily schedule and our project-based afternoon sessions was drawn directly from BPW. Although several of us teach at the University of Washington, we sought to arrange these workshops as volunteers outside of a formal classroom setting.

The initial workshops were an enormous success with 115 applicants of whom we were able to admit 52. In response to this demand, we ran the workshops again in late 2014, twice again in 2015, and once in early 2016. Additional workshops are planned in Seattle, twice a year, going forward. As we have been able to recruit more mentors, each workshop has been larger than the previous iteration. Our most recent workshop in early 2016 was attended by 97 participants.

Each time we have run them, the workshops were organized over one Friday evening and three Saturdays. A Friday session before the initial Saturday session ensured all participants (and their computers) were prepared for the following morning. The four sessions were numbered from 0 to 3 in reference to about

[3] http://bostonpythonworkshop.com/ (https://perma.cc/5Y36-R9FM)

[4] See Wilson (2014) and http://software-carpentry.org/ (https://perma.cc/23SE-BPHA)

Fig. 9.1 Four photographs from the Community Data Science Workshops held in April and May 2016. The top two panels show mentors working one on one with participants. The bottom left panel shows a breakout afternoon workshop with participants working independently on projects. The bottom right panel shows participants during a morning lecture with mentors standing to the side and ready to help participants when they require assistance

zero-indexing in the Python programming language. We collected feedback from participants after each day and debriefed instructors after each session and again after each series of workshops has concluded. Based on this process, we iterated on the curriculum and design of the workshops each time we ran them.

Each Saturday session begins with a 2-hour interactive lecture in the morning that builds upon the topics presented in previous sessions. Lectures introduce new concepts and show real examples of carrying out tasks through "live coding." A picture of a lecture is shown in the bottom right panel of Fig. 9.1. We encourage participants to participate in the lecture by actively programming on their own computers. The concepts discussed in each lecture introduce participants to a handful of tools and concepts that are then explored in the afternoon challenges. Each afternoon session is organized around open-ended questions designed to foster structured exploration of the morning's concepts to help participants synthesize and use their new skills.

Afternoon sessions involve independent project work. Participants are given an archive of several simple programs written using only concepts that participants were introduced to in the lectures. After a short exposition and explanation of the sample programs by a session leader, participants are encouraged to modify, build upon, or be inspired by these programs to solve problems of their choosing.

Participants work on projects individually, or in groups, with help from more experienced mentors present. This independent project work continues over 3–4 hours. We have experimented with many different projects. In general, we have offered participants two or three choices during each afternoon so that participants can choose projects that align with their interests. All but the bottom right panel in Fig. 9.1 show these project-based sessions. The top two panels both show mentors working one on one with participants. All of our curriculum—including sample projects, code, and recordings of lectures—are made available on our website.[5]

Day 0: Setup

In the first Friday session, participants walk through a checklist for installing Python and installing a programmer's text editor. Next, they work through a brief tutorial on the basics of using the command line. After the participants have completed these setup tasks, they are encouraged to work through some simple Python programming exercises. This makes the next morning lecture easier by pre-introducing the material covered in the Saturday lecture. The evening session is completely self-guided and allows participants to warm up to the concepts presented at their own pace. Mentors are on hand to provide technical assistance, help participants through difficult programming concepts, and verify that each student has completed session goals before they leave.

Day 1: Introduction to Programming

The first Saturday session starts with a reinforcement of how to work in the command line, and then introduces variables, Python's built-in data types including integers, floating point numbers, strings, lists, and dictionaries. As a result, after only one lecture, participants are familiar with all of Python's first-order data structures and all of the data types used in the rest of the workshops. Finally, we introduce conditional logic and loops. As in all of our lectures, we do not use slides. Instead, we demonstrate and discuss concepts while programming example code in an interactive Python interpreter using an iterative trial-and-error method. For example, we demonstrate strings by constructing messages from strings and demonstrate dictionaries by mapping names to ages (`{'Mako': 33, 'Ben': 24}`). Throughout the lecture, mentors are distributed throughout the room to be able to answer participants' questions about issues they are having in their code.

The first afternoon project session aims to support participants in engaging in simple data analysis using Python. For example, one session we have designed

[5] http://wiki.communitydata.cc/CDSW (https://perma.cc/G36T-KLG8)

begins by downloading an archive that includes code and a dataset drawn from the US Social Security Administration on the popularity of different baby names among US children. Projects like this allow participants to start analyzing real-world data to ask and answer questions of their own design almost immediately. For example, participants often begin by answering a question like "How many times does *your* name show up in the dataset?" and proceed to more complicated questions (e.g., "Which names are strict subsets of other names?"). Answering these questions reveals common challenges in data analysis immediately. For example, the exclusion of names given to fewer than five people of one gender leads directly to insights about missing data, while the binary nature of gender in the dataset leads to insights about how data collection decisions can support or suppress specific conclusions.

Day 2: Web APIs

For the second session, we step back from Python to spend time working with web APIs—web services that allow a program to acquire data from online communities and social media sources. One API we rely upon in the lecture is the PlaceKitten API, which takes a request for an image of a specified size and then returns an image of a kitten of that size. Participants are first shown how to make API requests through a web browser. We then show them how to make the same requests in Python.

Next, we demonstrate how to parse more complex API responses. We have often relied on data drawn from Wikipedia about articles related to Harry Potter as an example because there is a very large amount of data and it exhibits interesting patterns (e.g., bursts of edits around the release of each film and book). Afternoon sessions on the second full day involve working through and modifying simple programs that pull data from Twitter's API to build a tweet-gathering tool for use in the third session, from the Yelp API to find out about local restaurants, and from the Wikipedia API to answer questions about editing activity and article metadata.

Day 3: Data Cleanup and Analysis

The final session acts as a capstone highlighting the process of sourcing, cleaning, and using a dataset to ask and answer a question. In the morning lecture, we walk through a program that collects a dataset about every contribution to articles in Wikipedia related to Harry Potter using the Wikipedia API. Using these data, we generate a series of time series plots to answer several questions related to the way that Wikipedia editing on Harry Potter topics has changed over time.

The afternoon projects for this session focus on the process of data analysis and visualization. For example, we have used a pre-collected set of tweets about earthquakes (collected using a code that was crafted, in part, by participants during

an afternoon session on the second day) to generate time series in different resolutions and identify earthquakes around the world as they appear in the dataset. Other sessions have focused on gathering geocoded social media data and visualizing these data on a map. By showing participants different ways of interacting with datasets that they have gathered, we are able to contextualize the act of analyzing data and to provide examples of the process of analyzing social media data from start to finish.

Community Data Science Classes

In response to requests from our university, three of us have developed and taught quarter-length, for-credit, masters level courses based on the Community Data Science Workshops. The classes were taught at two different departments at the University of Washington: three times in the Department of Communication in our Communication Leadership program and once in the Department of Human Centered Design and Engineering. The courses directly incorporate most of the workshop curriculum described above. Unlike most other data science curricula, these classes' central focus is an extended, self-directed project which forms most of each student's grade. Curriculum for these classes are made fully available on our website.[6] Courses were taught to groups of 20 and 30 students with 1 instructor and 1 teaching assistant.

Teaching this material over 10 weeks, instead of 4 days, provided us with more opportunities to iterate on our lesson plans. The practice of sending out anonymous feedback surveys after each class session, carried over from the workshops, helped us adjust the pace and teaching style between sessions. However, other than the addition of more examples of APIs (essentially, the ability to teach more than one of the afternoon session from Day 2), we found that the additional time did not allow us to increase the scope of the material presented. We were challenged to address all core programming concepts thoroughly within the first few weeks of the course so that students would feel confident deploying those concepts in their own work while leaving them with sufficient time to select a dataset, to frame a research question, and to gather, analyze, and report their findings. The nature of the course work changed dramatically at roughly the halfway point: the first half of the quarter provided a crash course in data science programming; the second half focused on supporting students as they applied those lessons to specific datasets and research problems. Students with no previous programming experience needed to absorb a great deal of new knowledge within the first few weeks in order to successfully complete their class project.

The introduction of grades substantially raised the stakes of mastering the material, and it risked conflict with our "low stakes" approach in the workshops.

[6] e.g., https://wiki.communitydata.cc/CDSW#Courses (https://perma.cc/UQ42-ZF9B)

Homework assignments were graded on effort, not code quality. Each course culminated in a final project where success depended more on gathering and synthesizing data to tell a story than on the quality of the code written along the way. As an example, a student would receive full credit for an inefficient program or a program with a few missing edge cases but would lose credit for failing to identify a potential source of error like incomplete data. In one rendition of the class, data visualization was worth 25% of the project grade. Points were awarded if a plot represented the data correctly by using sensible color schemes and axes, not based on the students' choice or mastery of plotting technology (Excel was most commonly used).

Instructors teaching the courses did not always experience the same challenges. One course instructor felt that the move to a traditional classroom setting, which meant dramatically increasing the ratio of students to available mentors, reduced opportunities for ad hoc, one-on-one support. He attempted to compensate for this by building opportunities for peer support into the class and by grouping students with little or no previous programming experience with others who had some familiarity with programming in other contexts and languages. Another instructor found that the shift to more hours in class meant he could spend more time on average with each student.

There was consensus that while it was not possible to cover substantially more material in 10 weeks than in 3 weekends, it was possible to cover it more thoroughly. The higher student-to-mentor ratio made it more difficult to support struggling students, but the addition of assignments, feedback surveys, a more drawn out schedule, and self-directed projects helped assure that students had the opportunity to master the material. Students were also exposed to some new challenges, chiefly the challenge of finding data relevant to their subject of interest.

Outcomes

As we have developed the workshops and classes, we have devoted time to a discussion of our own goals. Although the organizers share a goal of "democratizing data science," this is an amorphous goal understood differently even within the team that developed the curricula. Through discussion, we established that there were several dimensions on which we feel our efforts should be evaluated. First, we believe that our approach should be evaluated in terms of its ability to support *skill development* among participants. In this first sense, we consider our approach effective only if participants are building skills associated with end user or conversational data science.

Second, given our goals of democratization, we believe that it is important that the curriculum be a successful form of *outreach* in that it should attract large numbers of individuals, especially from groups that are underrepresented in more traditional data science communities. Third and finally, we believe a success

criterion for our approach is its ability to support *empowerment*. In this final sense, we believe that it is not enough that learners simply have skills but that they feel able to build on these skills in ways that shift power.

Skill Development

The informal nature of our workshops makes it difficult to systematically ascertain the degree to which participants have learned skills. Some evidence of skill development comes from the opt-in surveys we have run after our sessions. In one typical response to an open-ended question about outcomes, a participant explained that the sessions helped build skills around programming and data analysis:

> It helped me become more comfortable with reading and writing code and taught me how to think more about how to use social media data to answer questions that are not necessarily academic. It also made me more confident to take the lead as the person responsible for writing code in a class project.

Although it is certainly the case that not every participant felt comfortable writing a code at the end of the four sessions, many explained that they felt more comfortable in a role of end user or conversational data scientists. For example, one explained that:

> Before the workshop I had no idea what Python can do, what API is for, or what data visualization is. The workshop basically was my entry point to the world of data analysis.

Another participant's feedback is an example of someone who became a more effective and confident conversational data scientist through their experience in the workshops:

> In my work as a librarian where I help clients navigate various sources of information, I feel more comfortable talking about how they can use programming to find or analyze the data they have access to.

In the classes where students each worked on projects over several weeks, more concrete evidence of skill development included the products they were able to create at the end of the class. For example, one student published a detailed report that attempted to understand the relationship between the release of television shows on Netflix and activity on associated Wikipedia articles. The student collected and compared a dataset of Wikipedia editing activity on articles associated with television shows released on Netflix with a similar dataset about broadcast television shows. Using these data, she provided evidence of a strong correlation between episode release dates and editing activity on Wikipedia.[7] There was also evidence

[7]The student, Nyssa Achtyes, published her analysis on a website titled *Long Term User Engagement of Netflix and Non-Netflix shows*: https://nyssadatascience.wordpress.com/ (https://perma.cc/Z9HK-ZVA3)

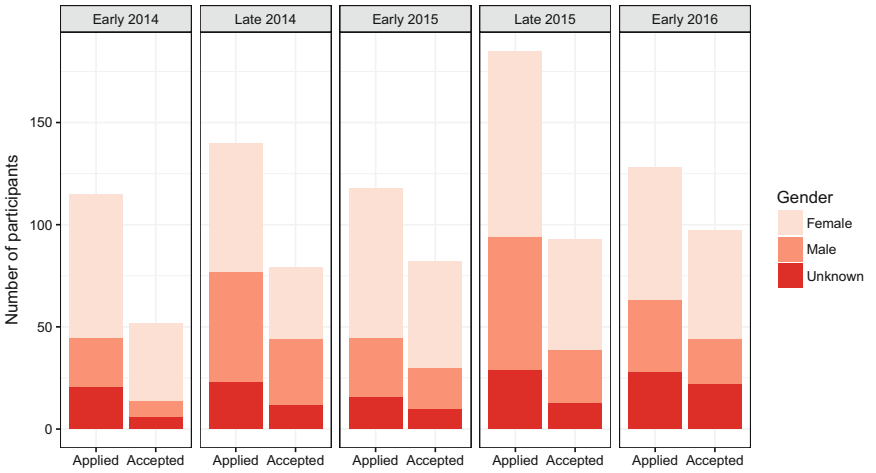


Fig. 9.2 Numbers of admitted participants at each workshop by inferred gender

of skill development among the academics who attended the workshops. At least one participant emailed us to say that they used skills developed in the class to collect and analyze data from the Twitter API that ultimately led to a published paper.

Outreach

The workshops have consistently attracted a large number of participants. Over the 5 series, 686 people applied to the workshops in Seattle, and 403 were accepted (see Fig. 9.2). In each case, we were constrained by the size of the instructional spaces we had access to and the number of mentors we had been able to recruit. Our curriculum has been adapted and taught outside of Seattle as well. For example, a group at the University of Waterloo's Women in Computer Science group has at least twice taught a series of workshops that relies heavily on our curriculum.

One of the most striking aspects of our workshops, so far, has been that our participants seem to come from more diverse backgrounds than in typical data science communities. For example, in every workshop and class, participants have been mostly women. This surprised us since we did not make targeted efforts to include (or exclude) a particular gender. To quantify the gender of participants, we analyzed the first names of the participants using the US Census and Social Security data to assign a probable gender to each name. Results are shown in Fig. 9.2 that show that a majority of both applicants and participants were female for each of the five sessions. There was also a fairly high proportion of women among our mentors—especially in later sessions when most mentors were returning participants.

We saw diversity along other dimensions as well. Because we targeted programming neophytes, a large portion of our attendees came from traditionally less technical departments within our university and from outside the university as well. For example, we attracted participants working for both local government and a large number of local nonprofits. The workshops were also attended by social media users including bloggers and participants in Wikipedia who were interested in building the skills to analyze data from their own communities.

Empowerment

Perhaps the most important—but difficult to measure—determination of whether our curricula have contributed to the democratization of data science is the degree to which participants felt empowered afterward. Although skill development might include the ability to understand or conduct data analysis, we feel that empowerment goes one step further and suggest that skills can affect and change the power structure in which participants find themselves—at least in relation to data and data analysis. Although empowerment is difficult to measure, opt-in post-workshop surveys of participants suggested that at least some participants felt that exposure to data science was empowering. For example, one former student told us:

> It [ultimately] gave me the confidence to accept a job teaching CS at a local CC, which led to me applying to the CS PhD program at [the University of Washington] (and getting in!). So, I guess it contributed to completely changing my life.

Another student reported a similar sense in which the program had led to a shift from a career in administration to one in software engineering:

> Well, I went to Hackbright Academy largely because its curriculum centers on Python. And now I'm a software engineer in San Francisco. So. . . pretty rad, huh?

One thing we encourage participants to do is to return to future workshops as mentors. Many participants, including two of the current organizers, have returned to become new mentors. This is both a good opportunity for the participants to continue engaging in data science and a sign of empowerment. In our most recent workshops, a majority of mentors were former participants.

Participants often did not continue to engage in data science after the workshop when they felt they did not have projects where they could use and improve their knowledge and skills. Participants who continued to engage in data science often had specific projects or pursued resources like Coursera, CodeAcademy, Data Science Dojo, and classes at the University of Washington. In terms of empowerment, assisting participants at this transitional stage—from the workshop to real-world settings—should be considered an integral part of any community data science curriculum and reflects an area we hope to focus on in future curriculum development.

Limitations

We believe that the community data science approach can benefit participants who seek to gain a working knowledge of programming and data science literacy. The first and most fundamental limitation is that we are trying to cover both data literacy and introductory programming simultaneously. Even individuals who are relatively comfortable exploring, aggregating, and describing data using software tools like spreadsheets often struggle to perform familiar, basic data manipulations using Python. Currently, our workshops and courses emphasize programming, but it is unclear that we have the right mix. We could certainly defer more programming concepts, or exclude them altogether, in favor of teaching participants how to use widely available software tools that accomplish the same task.

We could also choose to cover additional programming concepts, such as object orientation, that are useful for working with many common data science libraries. Of course, these decisions—to skip over a basic programming concept or to teach participants a non-programming alternative—would impose new constraints on what we can cover within the workshop as well as what participants will be able to accomplish afterward.

Furthermore, it is not yet clear to us what measures of success we should use to evaluate our approach. Participants seek out our workshops for a variety of reasons, arriving with vastly different types of experience. Some have more practical, immediate, opportunities to continue honing skills than others. Ultimately, success for any individual participant might be best evaluated based on that individual's goals and preparation as well as what they did with what they learned afterward than on direct measures of their performance or engagement during the sessions.

Conclusion

In their highly cited critique around the discourse of big data, danah boyd and Kate Crawford argue that limited access to big data analytic tools is creating new digital divides. The world, they suggest, is divided into the "Big Data rich" and the "Big Data poor" (boyd and Crawford, 2012). The issues boyd and Crawford raise about access to data are formidable and substantive. We see the community data model as one of very few attempts to address these issues directly. However, by framing big data equity as simply an access issue, boyd and Crawford may understate the problem. In ways that Dasgupta and Hill (2017) have shown, nonprofessional data scientists do not ask the same questions that professional data scientists ask. Democratized data science is not only a broader distribution of knowledge, skills, and power, it has the potential to support the development of new *types* of data science.

We believe that what we have developed in our workshops and classes is a proof of concept. That said, we feel confident in our demonstration that there is a broad demand for data science skills outside of traditional engineering circles and among groups, like women, that the fields most closely associated with data science have historically struggled to engage. We hope that we have also provided one vision of what a democratized data science curriculum might look like. A more democratized data science is possible—potentially even with broad societal effects. We encourage you to join us in the process of understanding what it might look like and what it might be able to accomplish.

Acknowledgements The Community Data Science workshops were made possible through the generous actions of dozens mentors who volunteered to spend their weekends teaching strangers data science. Without them, nothing we've described here would have been possible. Our work was also supported by the Department of Communication at the University of Washington which provided physical facilities and other resources. Finally, this work was supported by a Data Science Environments project award from the Gordon and Betty Moore Foundation (Award #2013-10-29) and the Alfred P. Sloan Foundation (Award #3835) to the University of Washington eScience Institute. eScience supports data-driven discovery at the University of Washington in many ways and provided financial and other forms of support for the workshops and for this chapter.

References

Benkler, Y. (2016). Peer production and cooperation. In J. M. Bauer & M. Latzer (Eds.), *Handbook on the economics of the internet*. Cheltenham, UK: Edward Elgar.

boyd, d., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society, 15*(5), 662–679. doi:10.1080/1369118X.2012.678878

Buechley, L., & Hill, B. M. (2010). LilyPad in the wild: How hardware's long tail is supporting new engineering and design communities. In *Proceedings of the 8th ACM Conference on Designing Interactive Systems (DIS'10)* (pp. 199–207). New York, NY: ACM Press. doi:10.1145/1858171.1858206

Cheryan, S., Plaut, V. C., Davies, P. G., & Steele, C. M. (2009). Ambient belonging: How stereotypical cues impact gender participation in computer science. *Journal of Personality and Social Psychology, 97*(6), 1045–1060. doi:10.1037/a0016239

Chilana, P. K., Alcock, C., Dembla, S., Ho, A., Hurst, A., Armstrong, B., & Guo, P. J. (2015). Perceptions of non-CS majors in intro programming: The rise of the conversational programmer. In *Proceedings of the 2015 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)* (pp. 251–259). Piscataway, NJ: IEEE Press. doi:10.1109/VLHCC.2015.7357224

Chilana, P. K., Singh, R., & Guo, P. J. (2016). Understanding conversational programmers: A perspective from the software industry. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI'16)* (pp. 1462–1472). New York, NY: ACM Press. doi:10.1145/2858036.2858323

Dasgupta, S. (2016). Children as data scientists: Explorations in creating, thinking, and learning. Ph.D. Dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Dasgupta, S., & Hill, B. M. (2016). Learning with data: Designing for community introspection and exploration. In *Workshop on Human-Centered Data Science*. Position Paper. Computer supported cooperative work and social computing, San Francisco, CA.

Dasgupta, S., & Hill, B. M. (2017). Scratch community blocks: Supporting children as data scientists. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI'17)*. New York, NY: ACM Press. doi:10.1145/3025453.3025847

Davenport, T. H., & Patil, D. J. (2012). Data scientist: The sexiest job of the 21st century. *Harvard Business Review*. Retrieved from 11 July 2016. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Dwoskin, E. (2014). Big data's high-priests of algorithms; 'Data scientists' meld statistics and software for find lucrative high-tech jobs. *Wall Street Journal (Online): Tech*. Retrieved from 11 July 2016. http://search.proquest.com/newsstand/docview/1552020409/abstract/D70B27FC5DA74D5APQ/1

Estrada, T., & Atwood, S. A. (2012). Factors that affect student frustration level in introductory laboratory experiences. *2012 ASEE Annual Conference & Exposition*. American Society for Engineering Education, 25.629.1–25.629.7. Retrieved from https://peer.asee.org/21386

Forte, A., & Guzdial, M. (2005). Motivation and nonmajors in computer science: Identifying discrete audiences for introductory courses. *IEEE Transactions on Education, 48*(2), 248–253. doi:10.1109/TE.2004.842924

Gellman, L. (2014). Business education: Big data gets master treatment-some business schools offer one-year analytics programs, catering to shift in students' ambitions. *Wall Street Journal*, B.7. Retrieved from 11 July 2016. http://search.proquest.com/newsstand/docview/1620527411/abstract/B21739238EE74F26PQ/1

Good, C., Rattan, A., & Dweck, C. S. (2012). Why do women opt out? Sense of belonging and women's representation in mathematics. *Journal of Personality and Social Psychology, 102*(4), 700–717. doi:10.1037/a0026659

Guzdial, M. (2003). A media computation course for non-majors. In *Proceedings of the 8th Annual Conference on Innovation and Technology in Computer Science Education (ITiCSE '03)* (pp. 104–108). New York, NY: ACM Press. doi:10.1145/961511.961542

Guzdial, M., & Forte, A. (2005). Design process for a non-majors computing course. In *Proceedings of the 36th SIGCSE Technical Symposium on Computer Science Education (SIGCSE '05)* (pp. 361–365). New York, NY: ACM Press. doi:10.1145/1047344.1047468

Heer, J., Viégas, F. B., & Wattenberg, M. (2007). Voyagers and voyeurs: Supporting asynchronous collaborative information visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)* (pp. 1029–1038). New York, NY: ACM Press. doi:10.1145/1240624.1240781

Howe, J. (2006). The rise of crowdsourcing. *Wired Magazine, 14*(6), 1–4.

Jones, C. (1995). End user programming. *Computer, 28*(9), 68–70. doi:10.1109/2.410158

Joshi, K. D., Kvasny, L., Unnikrishnan, P., & Trauth, E. (2016). How do black men succeed in IT careers? The effects of capital. In *Proceedings of the 2016 49th Hawaii International Conference on System Sciences (HICSS)* (pp. 4729–4738). Piscataway, NJ: IEEE Computer Society Press. doi:10.1109/HICSS.2016.586

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge, UK: Cambridge University Press.

Luther, K., Counts, S., Stecher, K. B., Hoff, A., & Johns, P. (2009). Pathfinder: An online collaboration environment for citizen scientists. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)* (pp. 239–248). New York, NY: ACM Press. doi:10.1145/1518701.1518741

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute. Retrieved from 11 July 2016. http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation

Margolis, J., & Fisher, A. (2001). *Unlocking the clubhouse: Women in computing*. Cambridge, MA: The MIT Press.

Matias, J. N. (2016). Going dark: Social factors in collective action against platform operators in the Reddit blackout. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)* (pp. 1138–1151). New York, NY: ACM Press. doi:10.1145/2858036.2858391

Miller, C. C. (2013, April 14). The numbers of our lives. *New York Times: ED*, ED.18. Retrieved from 11 July 2016. http://search.proquest.com/newsstand/docview/1326574891/abstract/88A4A39B52A94D3BPQ/2

Nafus, D. (Ed.) (2016). *Quantified: Biosensing technologies in everyday life*. Cambridge, MA: MIT Press.

Nardi, B. A. (1993). *A small matter of programming: Perspectives on end user computing*. Cambridge, MA: MIT Press.

Neff, G., & Nafus, D. (2016). *Self-tracking*. Cambridge, MA: MIT Press.

Papert, S. (1980). *Mindstorms: Children, computers, and powerful ideas*. New York, NY: Basic Books.

Papert, S. (1987). Computer criticism vs. technocentric thinking. *Educational Researcher, 16*(1), 22–30. doi:10.2307/1174251.JSTOR:1174251

Raddick, J., Lintott, C. J., Schawinski, K., Thomas, D., Nichol, R. C., Andreescu, D., ... Slosar, A., et al. (2007). Galaxy Zoo: An experiment in public science participation. *Bulletin of the American Astronomical Society, 38*, 892.

Ratto, M. (2011). Critical making: Conceptual and material studies in technology and social life. *The Information Society, 27*(4), 252–260. doi:10.1080/01972243.2011.583819

Resnick, M., Silverman, B., Kafai, Y., Maloney, J., Monroy-Hernández, A., Rusk, N., ... Silver, J. (2009). Scratch: Programming for all. *Communications of the ACM, 52*(11), 60. doi:10.1145/1592761.1592779

Simpson, R., Page, K. R., & De Roure, D. (2014). Zooniverse: Observing the world's largest citizen science platform. In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14 Companion)* (pp. 1049–1054). New York, NY: ACM Press. doi:10.1145/2567948.2579215

Smith, A. M., Lynn, S., & Lintott, C. J. (2013). An introduction to the Zooniverse. In *First AAAI Conference on Human Computation and Crowdsourcing (HCOMP '2013)*. Palo Alto, CA: AAAI Press.

Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., & Kelling, S. (2009). eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation, 142*(10), 2282–2292. doi:10.1016/j.biocon.2009.05.006

Viegas, F. B., Wattenberg, M., van Ham, F., Kriss, J., & McKeon, M. (2007). ManyEyes: A site for visualization at Internet scale. *IEEE Transactions on Visualization and Computer Graphics, 13*(6), 1121–1128. doi:10.1109/TVCG.2007.70577

Wattenberg, M., & Kriss, J. (2006). Designing for social data analysis. *IEEE Transactions on Visualization and Computer Graphics, 12*(4), 549–557. doi:10.1109/TVCG.2006.65

West, J., & Portenoy, J. (2016). The data gold rush in higher education. In C. Sugimoto, H. R. Ekbia, & M. Mattioli (Eds.), *Big Data is Not a Monolith*. Information Policy. Cambridge, MA: MIT Press.

Wilson, G. (2014). Software carpentry: Lessons learned. *F1000Research*. doi:10.12688/f1000research.3-62.v1

Wolf, G. (2010). The data-driven life. *The New York Times*. Retrieved 2016, August 12 from http://www.nytimes.com/2010/05/02/magazine/02self-measurement-t.html

Wood, C., Sullivan, B., Iliff, M., Fink, D., & Kelling, S. (2011). eBird: Engaging birders in science and conservation. *PLOS Biology, 9*(12), e1001220. doi:10.1371/journal.pbio.1001220

Benjamin Mako Hill is a data scientist who studies study collective action in online communities. He is an Assistant Professor of Communication at the University of Washington, a Faculty Associate at the Berkman Klein Center for Internet and Society at Harvard University, and a participant in Wikipedia and a number of other peer production communities.

Dharma Dailey studies how people get information during crises. She attended the first Community Data Science Workshop as a student and put what she learned into her research! She found the workshop so helpful, she stuck around to help organize more of them. She is a PhD Candidate in Human-Centered Design and Engineering at the University of Washington.

Richard T. Guy is a Data Scientist at Microsoft where he works on large scale experimentation. He enjoys teaching programming and data science wherever they will let him.

Ben Lewis is a software engineer who advocates for community involvement in decision making, and seeks to expand access to tools for understanding and shaping the world. He is a graduate of McGill University, an occasional contributor to open source projects, and a participant in Wikipedia.

Mika Matsuzaki is an epidemiologist at the University of Washington studying substance use and social support among marginalized populations.

Jonathan T. Morgan is a Senior Design Researcher at the Wikimedia Foundation. He has a PhD from the University of Washington in the Department of Human Centered Design & Engineering.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

Index

© Springer International Publishing AG 2017

S.A. Matei et al. (eds.), *Big Data Factories*, Computational Social Sciences,
https://doi.org/10.1007/978-3-319-59186-5