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.

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

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

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

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