Advances in Data Modeling and Text Mining for Covid-19 Research

Amy Neustein

Nathaniel Christen

Approx 250 pages, 15 chapters
Manuscript Submission Date: August 31, 2020

1. PROJECT SUMMARY

1.1. Purpose/Summary of the Project

The Covid-19 pandemic has inspired an unprecedented convergence of scientific research, driven in part by publishers choosing to allow open public access to many research papers and data sets relevant to Covid-19 (the disease) and SARS-CoV-2 (the viral agent). The sheer volume of this data presents both a practical challenge — how should scientists find the information most relevant to them — and a valuable case-study in the difficulties and possibilities for curating multi-disciplinary information spaces unified around a common scientific theme: in this case, determining how best to approach Covid-19 and to mitigate the global SARS-CoV-2 pandemic.

In a matter of mere weeks, the scientific and publishing community has essentially engineered the origination of an entirely new data ecosystem, centered around the challenges of studying and predicting properties of Covid-19 and SARS-CoV-2 at the molecular, genomic, epidemiological, clinical, and public-health levels. In many respects, this ecosystem has emerged haphazardly. There appears to be a rush to publicly share text and data which has taken priority over rigorous curation and accuracy. In this context, we are proposing a volume for scientists, computer scientists, and policymakers who may benefit from a volume which provides a critical and analytic overview of the Covid-19 data ecosystem: the different genres of data which are marshaled toward scientific investigation of Covid-19 and SARS-CoV-2; how this data is obtained, consumed, analyzed, and interpreted; how research data supports scientific claims pertaining to Covid-19's biological and epidemiological mechanisms and trajectory; and how these scientific claims should translate into public policy.

At one level, we intend to examine the operational logistics of Covid-19 data: its structures, protocols, analytic methodology, and empirical significance. That is, we intend to identify the distinct scientific disciplines which each concern one facet of Covid-19 research — molecular biology, genomics, radiology, epidemiology, clinical informatics, and their various subfields — and, for each of these disciplines, review their distinct paradigms for data acquisition, analysis, and modeling. The point of these expositions is to bring the reader from conceiving "data" as something abstract and amorphous, to understanding data as the building-blocks of scientific research and biomedical claims. One way to supply this backstory is to examine Covid-19 data from the viewpoint of software engineering: to demonstrate the methodology for data acquisition and management from the perspective of programmers implementing software which manipulates Covid-19 data. This explication would therefore examine the data structures, file formats, **API** protocols, and other technical details intrinsic to writing code which works with Covid-19 data as a digital artifact. Such an exposition might be of interest to programmers who actually are writing Covid-19-related code, but the primary goal of these discussions will be to help scientists (who may be well-versed in data structures relevant to their specialization but less so vis-à-vis other disciplines), policymakers, and the general public understand the technical chains which transform Covid-19 information from the realm of laboratories and experiments to the realm of public health and policy.

Apart from that topical focus — the analysis of Covid-19 data as a concrete ecosystem manifest in standardized formats, global identifiers, and other concrete information artifacts — we also intend to write for a more theoretical audience for whom the pandemic is a unique case-study in data curation and integration. From this perspective, the Covid-19 data ecosystem is a concrete example through which theories of data integration can be presented and exercised.

There are two distinct phenomena which render inter-disciplinary data integration significant for Covid-19 in particular, and clinical/biomedical practices in general. First, certain forms of analysis explicitly combine information or statistical parameters from distinct subject areas. For example, in addition to epidemiological models of SARS-Cov-2 within an entire population, it is important to study the present or projected spread of the disease among different social groups, identified by age, gender, race, economic status, and so forth. This form of analysis will therefore merge epidemiological and sociodemographic data and methods; as such it is an instance of analyses wherein it is explicitly necessary to pool data that is typically represented via different schemas, and accessed via different protocols, into a single algorithmic or computational environment. This volume will therefore examine cross-disciplinary analysis along these lines as case studies of data integration on a procedural level: how computer code can obtain and marshal heterogeneous data into a common form suitable for qualitative and quantitative analysis.

The second context where multi-disciplinary integration becomes relevant operates at a higher level: the development of heterogeneous information spaces which can absorb data from many environments, evincing a variety of disciplinary orientations. The rationale for such heterogeneous repositories is often practical and logistical; institutions have operational reasons for curating a single, comprehensive data ecosystem shared by multiple information producers and consumers, such as a "Semantic Data Lake". In these situations, one large central repository will take the place of numerous narrower, domain-specific databases. A central repository may be subdivided into smaller components implementing narrower protocols — e.g., a clinical software network may provide diagnostic images via a **PACS** (Picture Archiving and Communication System) service, and treatment/outcome data via an **EMR** (Electronic Medical Record) architecture. The structure and use of data in these two environments (**PACS** and **EMR**) is very different. Nevertheless, institutions will often unify these systems into a single data platform, for logistical reasons: it is more convenient for doctors and researchers to have a single access point, a single login account, a single query framework, etc., which accesses the totality of information used across the organization's activities.

These institutional repositories present challenges which are different from granular syntheses of heterogeneous data into a single procedural/algorithmic context. Disparate data structures in a heterogenous archive, such as a "Data Lake", may never be directly combined in a single computation. Nevertheless, Data Lakes and their kin seek to provide a single software, query, and accession infrastructure which can accommodate a diversity of data models, and this diversity presents technological challenges. It appears that Covid-19 demonstrates the problems engendered by these complexities in a tangible way, insofar as health and governmental officials have criticized the lack of integrated data across disciplinary and jurisdictional boundaries — poor coordination between city, state, and federal governments in the US, for example, as well as between medical and governmental institutions. Covid-19 therefore offers a case-study in the challenges of implementing large-scale heterogeneous data repositories, and we intend to offer theoretical analyses and practical recommendations which could potentially improve such technology in the future.

The volume will therefore be organized in a pattern which progresses from domain-specific to integrative styles of analysis: the first part will focus on data models and protocols within individual disciplines, while the second part will discuss cross-disciplinary integration at both a theoretical and practical level. The third part will then delve deeper into integrative paradigms in several areas, particularly text mining and software development.

1.2. Coverage and Approach

As just outlined, the book will examine Covid-19 research on two levels — one more empirical and one more theoretical. As such, we believe the volume will be of interest to two distinct audiences. On the one hand, scientists and policymakers could benefit from a broad overview of Covid-19 research, one which uses empirical case-studies to illustrate how Covid-19 data is accessed and analyzed, within the disparate disciplines that collectively contribute to our knowledge about the disease. On a more technical level, the book may be of interest to computer scientists and software engineers who will find new theoretical models and type systems with which to investigate data-integration problems. The new theory developed here has practical applications to fields such as database implementation and Software Language Engineering, which we will document via supporting code.

2. Table of Contents

- Chapter 5: Modeling Clinical Data in the Covid-19 Patient Population

This chapter will examine medical records, in the context of Covid-19 as a case-study: formats for representing diagnostic, treatment, and outcome data in a clinical setting, and the construction of patient cohorts. Focus will be placed on one specific technology — the Clinical Looking Glass software application and Object Model — as a representative example of how patient-centered data is structured and queried.

Part II: Creating a Cross-Disciplinary Ecosystem for Covid-19

- Chapter 6: Approaches for Merging Heterogeneous Data Sets: Ontologies and Hypergraphs

This chapter will review the use of ontologies to model specific information domains and, via ontology integration, to construct unified knowledge systems that encompass multiple domains. The chapter will present varieties of hypergraphs as generic data containers applicable to heterogeneous domains, with the goal of identifying sufficiently generic data structures appropriate for cross-disciplinary integrated ontologies.

- Chapter 7: Scientific Workflows and Inter-Application Networking: Reviewing data pipelines commonly used in Covid-19 research

This chapter will focus on **API**s, command-line interfaces, and related technologies through which software components communicate with one another. The purpose of this review is to direct attention not at specific procedures which are implemented within a given software component, but rather to show how components expose functionality to the "outside world." This review sets the stage for a more internal focus (implementation details within a software component) in the following chapter.

- Chapter 8: Formal Procedural Models: Representing computational procedures applicable to Covid-19

This chapter will concentrate more rigorously on individual procedures implemented within software components. The goal is to use external interface models (discussed in the previous chapter) as entry-points to analyzing components at a procedural level: viz., a good method for analyzing software functionality is to examine how internal procedures provide the capabilities exposed to an external interface. On this basis we will develop a general model of procedural properties which adequately represents computational logic across different programming languages and software-development methodologies.

- Chapter 9: Integrating Procedural and Data Models

This chapter will unify the discussion within the prior three chapters, which will have yielded, first, a general and cross-domain model of data structures and, second, a paradigm-agnostic model of procedural logic. The current chapter will merge both models into an overarching paradigm, exploiting the fact that procedural types and requirements are representable as data structures in their own right.

- Chapter 10: Type Theories for Procedural Data Modeling

This chapter will provide a formal structure supporting the analysis developed in Chapter 9. The priority in this chapter is codifying a hypergraph-based type theory, one which exploits the hypergraph context to concretely anchor a type system: each inhabited type is by definition an attribute of one or more hypernodes. This starting-point, along with the specification of a particular genre of "channelized" hypergraphs (introduced preliminarily in Nathaniel Christen's chapter in Amy Neustein's just-published *Advances in Ubiquitous Computing* volume), permits the construction of a general-purpose type theory applicable to most programming environments.

Part III: Text and Data Mining for Covid-19

- Chapter 11: Applying Data Mining Techniques to Covid-19 Research Corpora

This chapter will take a concrete look at research data published in conjunction with Covid-19 data-sharing initiatives. The machinery of the prior chapters will be leveraged to examine this existing data in a rigorous fashion, using formally described data types and hypergraph meta-models as methodological tools in surveying the Covid-19 data ecosystem.

- Chapter 12: Text Mining of Covid-19 Publication Archives

This chapter will examine Covid-19 corpora with a concrete focus that overlaps with that of Chapter 11, but with an emphasis on text rather than data mining. The goal of the present chapter is to review publishers' efforts to provide open-access document corpora to support Covid-19 research, and to demonstrate how text mining tools can make these corpora more valuable.

- Chapter 13: Human-Computer Interaction Approaches for Covid-19 Software

This chapter will apply hypergraph-based type theory (as developed in Part II) to **HCI** methodology, with Covid-19 Text and Data Mining (**TDM**) as a case-study. The goal of this chapter is to consider **TDM** software targeting Covid-19 corpora as concrete examples illustrating how **HCI** concerns may be analyzed through the lens of the data models and type systems presented earlier in the volume.

- Chapter 14: Using Text-Mining Tools to Extract Medical History from Clinical Narratives

This chapter will examine how text-mining technology considered in Chapter 12 may be applied to clinical and patient narratives for the purpose of extracting relevant patient data, pharmacological data, and epidemiological data, to improve patient care.

- Chapter 15: Annotating Patient Narratives for Emerging Covid-19 Symptomatology

This chapter will introduce techniques for encoding rhetorical structures identified within patient narratives where patients describe their first-hand experiential symptomology associated with Covid-19. These subjective descriptions can be found on social media websites and in the patient histories uploaded to portals which are made part of the overall Electronic Health Records.

3. Competition

The structure and presentation of this book is inspired in part by Alexandru Telea's *Data Visualization — Principles and Practice*. In particular, we incorporate the idea of providing sample data sets associated with individual chapters, and recommending software applications that readers may use to open and explore these data sets. In fact, we intend to provide customized versions of several applications in a downloadable package accompanying the text (reusing code we developed in conjunction with Amy Neustein's just-published *Advances in Ubiquitous Computing*). In contrast to *Data Visualization*, however, the current volume is focused on data modeling rather than data visualization, and of course uses Covid-19 research as a case study for all data-modeling examples.