

# Cross-Disciplinary Data Integration Models for the New Covid-19 Data Ecosystem

Amy Neustein

Nathaniel Christen

Approx 250 pages: 15 chapters

Manuscript Submission Date: October 15, 2020

**Key Words:** *Covid-19; SARS-Cov-2; Data Modeling; Text Mining; Computational Epidemiology; Software Language Engineering; Hypergraphs; Research Objects; Electronic Health Records; Public Health*

## 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 present in the curation of 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 weeks, the scientific and publishing community has essentially engineered the formation 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 somewhat 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, biostatisticians, computer scientists, and policy makers who may benefit from a book 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, each of which is aimed at one particular 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 that works with Covid-19 data as a digital artifact. 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), policy makers, and a discerning general public to better understand the technical chains which transform Covid-19 information from the realm of laboratories and experiments to the realm of public policy and public health.

Aside from the empirical focus geared toward scientists and policy makers, as explained above, this book will also satisfy a more theoretical and IT-focused 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 cross-disciplinary data integration can be presented.

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, this is an example 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 analyses.

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 that is 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. It is understood that 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 heterogeneous 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 presumably accommodate a diversity of data models, and this diversity certainly presents technological challenges. And in fact, 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 presents a case-study in the challenges of implementing large-scale heterogeneous data repositories. *The purpose of this volume, in addition to reviewing Covid-19 data models, is therefore to offer theoretical analyses and practical recommendations to improve such technology for the present and future.*

The volume will accordingly be organized in a format 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 above, this volume will examine Covid-19 research on two levels: one that is empirically focused, and one that is more theoretical. As such, this volume would be of interest to two distinct audiences. On an empirical level, scientists and policy makers 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 theoretical level, the book would 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 theories and methods explicated in this book have practical applications to fields such as database implementation and Software Language Engineering. Such applications will be documented via the supporting code available within publicly-accessible data sets, helping to further the goals of the Research Object Protocol.

## 2. Table of Contents

### Foreword (Invited)

### Authors' Introduction

## Part I: Architecting Data Models for Scientific Disciplines Associated with Covid-19

- Chapter 1: Data structures for molecular biology and virology

This chapter will consider data pertaining to scientists' investigation of SARS-Cov-2's viral mechanisms. It will consider how data modeling the pathogen's proteins, physical structure, and interactions with human cells is collected and utilized. Emphasis will be placed on cheminformatic pipelines and protocols for working with molecular-biological information.

- Chapter 2: How Genomic Data is Stored and Analyzed in the Coronavirus Context

This chapter will examine the genomic data ecosystem with SARS-Cov-2 as a case-study. It will focus on the challenges presented by the large scale of genomic data, and how these challenges are addressed through data acquisition protocols and distributed software networks.

- Chapter 3: Structuring of Radiographic and Diagnostic Data in the Context of Covid-19

This chapter will focus on information germane to identifying Covid-19 infections. It will be centered on diagnostic imaging, but will also consider the structure of data generated by Covid-19 tests or biometric indicators of possible SARS-Cov-2 infection.

- Chapter 4: Reviewing Epidemiological Structures and Methodology for SARS-Cov-2 Research

This chapter will consider epidemiological modeling both *a posteriori* (via empirical clinical data) and via simulations (which try to predict different possible trajectories for the Covid-19 pandemic). The chapter will review fundamental epidemiological measures such as infection and transmission rate, mortality rate, and  $R_0$  — documenting how these magnitudes are assessed both empirically and theoretically.

- 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 **APIs**, 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 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 symptomatology 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 Record (**EHR**) system.

### 3. About the Market

#### 3.1. Audience

Epidemiologists; Computational Epidemiologists; Statisticians; Biostatisticians; Software Engineers; Linguists; Policy Makers; Government Health Officials; and Healthcare Administrators

#### 3.2. Secondary Markets

Epidemiology; Computational Epidemiology; Biostatistics in Medicine; Health Policy; Software Engineering; Quantitative Analytics/Statistics; Data Visualization; and Text Mining

#### 3.3. 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 of 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*). Although many publications, including Telea's, are paired with Open-Access data sets, they fail to provide software to help researches make optimal use of this data (this book will be unique in implementing customized software to work with the associated data sets). An additional shortcoming of Telea's textbook is that it is narrowly focused on data visualization, and does not rigorously examine data models. By contrast, the present volume will examine data modeling in detail at both an empirical and theoretical level. Moreover, the present volume uses Covid-19 research as a case study for data-modeling examples, and therefore is especially important at the current moment in time.