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

Amy Neustein
Linguistic Technology Systems, Fort Lee, NJ.
E-mail: amy.neustein@verizon.net
Nathaniel Christen
Linguistic Technology Systems, Fort Lee, NJ.

1. Introduction

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 published data relevant to Covid-19 (the disease) and SARS-CoV-2 (the viral agent). The sheer volume of the data presents both a practical challenge — to help scientists find the information most relevant to them — and a valuable case-study in the challenges and possibilities for curating multi-disiplinary information spaces unified around a common scientific theme and research project (in this case, determining how best to treat Covid-19 and to mitigate the global SARS-CoV-2 pandemic).

In a matter of mere weeks, the scientific and publication community has essentially engineered the origination of an entirely new data ecosystem, unified 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, with the rush to publicly share text and data taking priority over rigorous curation and accuracy. In this context, scientists and policymakers 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 marshalled 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, taking into consideration both the importance of basing government actions on empirical data and the gaps in scientific knowledge which make such data inexact and provisional.

1.1. Audience

Against this background, our proposed volume will be directed at two distinct audiences. 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 variatious 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 understand data as the building-blocks of scientific research and scientific claims. We want readers to have some awareness, when hearing or assessing claims made by either domain experts or public officials,

the provenance and history of the data behind these 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 put to the test.

1.2. Data Integration in the Covid-19 Context

There are two distinct phenomena which render inter-disciplinary data integration significant for Covid-19. 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. This is an example of analysis where it is explicitly necessary to pool data which is typically represented via different schema, and accessed via different protocols, into a single algorithm or computational environment. The current 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 (often quantitative) analysis.

The second context where multi-disciplinary integration becomes relevant operates at a higher level: the development of heterogenous information spaces which can absorb data from many environments, evincing a variety of disciplinary orientations. The demand for such heterogeneous repositories is often practical and logistical: institutions will curate a single, coprehensive 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; a clinical software network may provide diagnostic images via a **PACS** (Picture Archiving and Communication System) service, and treatment/outcome data via an **PACS** (Electronic Medical Record) architecture. The structure and use of data in these two environments — **PACS** and **PACS** — is very different. Nevertheless, clinical 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 criticised 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 disiplines relevant to Covid-19, while the second part will discuss cross-disiplinary 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.

2. Table of Contents