Chapter I: Introduction

1 Prelude

On May 11, 2021, Dr. Rochelle Walensky, director of the US Centers for Disease Control (C.D.C.), provided testimony to a hearing of the Senate "HELP" (Health, Education, Labor and Pensions) Committee which proved, after the fact, to be surprisingly controversial. At issue was how Dr. Walensky defended the C.D.C.'s guidance on outdoor activities by indirectly citing a study which was a statistical outlier, with her testimony later challenged by the very authors whose work she had relied upon as a source of up-to-date information. The senators criticizing Dr. Walensky focused on the matter of SARS-CoV-2 outdoors transmissibility, particularly in the context of when it was safe to reopen schools and summer camps. Outside of the committee itself, however, it was not so much the C.D.C. guidelines which garnered controversy but the fact that Dr. Walensky's testimony misconstrued the results of a recent publication, giving improper weight to one of multiple studies considered as part of a "systematic review."[1]

Hours after Dr. Walensky's testimony, one of the authors of the systematic review she referenced used twitter to dispute the C.D.C. directors' interpretation of their results. The specific issue in contention was an estimation of the percentage of Covid-19 cases that can be traced to outdoors rather than indoors transmission. This actual number is likely to be below 1%.[2] Dr. Walensky, however, cited the Systematic Review paper as claiming a transmission rate of *less than 10%*, a maximum bound derived from the statistical extremum among all the reviewed articles. In an interview with the New York Times, the author who criticized Dr. Walensky's testimony argued that she and her co-authors "were very clear we were not making a summary number" with the 10% upper bound, and that their paper was technically a "systematic review" and not (as Dr. Walensky described it) a "meta-analysis." Scientifically, as the Times puts it, "A meta-analysis often includes a precise estimate — a best guess, based on the data" whereas "A systematic review is more general." In short, the author implied that Dr. Walensky was misreading the analytic methods of their paper and as a consequence had presented a quantitative summary which was significantly different from their actual findings.

Of all the points of contention surrounding Covid-19, this particular controversy stands as little more than a footnote in the annals of US Government response to the pandemic (although confusing statements about outdoor transmissions continue to be cited as a contributing factor in the public's mistrust of the C.D.C.). However minor it may be, though, this episode raises interesting questions about how scientists' and policymakers' utilization of scientific research. After all, we certainly believe that policies (especially in contexts related to medicine and health care) should be informed by empirical data. Yet how should policymakers interpret research which actually produces empirical data sets? Insofar as many areas of large-scale public concern (certainly Covid-19 is a prime example) engender numerous different research projects, often yielding inconsistent results, how should such diverse data be consolidated into a single model for data-driven policy? Or, is there some threshold of divergence beyond which government officials should simply acknowledge that the science is inconclusive, and defer to scientists to produce more consistent findings before attempting to legitimize policies on empirical terms?

Insofar as multiple research projects often yield conflicting results, citing one specific research work can give a misleading overview of the relevant field as a whole — even if that work was conducted professionally and responsibly. In other words, our criteria for assessing the merits of scientific work inevitably shifts based on the relation of a given research endeavor to others on similar topics. Individual scientists, of course, can do no more than conduct their own research according to disciplined protocols, yielding results which are as conclusive as possible. In this sense well-executed research can be deemed definitive in the specific manner that it was conducted responsibly and in accord with protocols designed to minimize randomness and error. Findings which are *methodologically* sound, however, may still be inaccurate. Whenever multiple research projects address similar themes, to the degree that their findings can be contrasted, the totality of results among multiple studies should presumably be considered as a context for assessing the claims of any one study individually.

These points may seem obvious, even trivial, but they raise interesting questions under the general rubric of a philosophy of science — research methods are supposed to adhere to rigorous protocols *so that* the corresponding work is empirically persuasive. Empirical science is driven by the belief that we can arrive at factually decisive results by conducting research which is optimized to yield statistically significant results. It is therefore at least philosophically uncomfortable when multiple studies — all done correctly and professionally — yield substantially incommensurate conclusions. Unless specific details in trial design or sample populations (whatever these may be in the research context) are identified which could explain scientific discrepancies, how can we be confident in the accuracy of individual (even well-implemented) scientific projects, when history shows us many cases of equally meritorious research work yielding inconsistent results?

In principle, aggregative compilations of related research work — whether these be called "systematic reviews," "meta-analyses," or something else — can try to resolve the discrepancies between multiple related publications. Philosophically, however, we should consider whether this simply translates the epistemological uncertainties engendered by contradictory findings onto a different scale. After all, if multiple (well-executed) papers yield divergent results, why should we expect a simply numeric average across all such papers to be more empirically accurate, if we have no explanatory mechanism for

[1] See https://www.nytimes.com/2021/05/26/briefing/CDC-outdoor-covid-risks-guidelines.html for an overview

[2] See the specific publication, "Outdoor Transmission of SARS-CoV-2 and Other Respiratory Viruses: A Systematic Review," for details (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7798940).

what caused the discrepancies in the first place?

In the case of Dr. Walensky's testimony, the manner in which she drew conclusions from the Systematic Review inadvertently placed undue weight on one or two specific papers included within that review (according to the Times, the 10% upper bound on outdoor versus indoor Covid-19 transmission may have been skewed by one study among construction workers, whose working conditions are not representative of outdoor activities in general). This is an indirect version of the fallacy in citing one single paper as a conclusive source when its findings conflict with other similar papers. And yet, data-driven policy has to draw facts from *somewhere* to be empirically grounded in the first place. It is not philosophically evident that some sort of quantitative synthesis of multiple papers should be deemed authoritatively more accurate than the results of one single paper. Is not a meta-analysis itself one form of research, which can coexist with other meta-analyses potentially yielding different results in turn, and so on, *ad infinitum*?

Perhaps the best way out of such an impasse is to be more rigorous in how multiple research projects are integrated. To the degree that similar research work yields a spectrum of distinct results, scientists should actually try to explain those differences, rather than relying on some sort of statistical averaging effect. If findings from one project prove difficult to reproduce, scientists should try to identify the source of this difficulty in the original work, the attempted replications, or some combination. If one (well-executed) paper appears to be a statistical outlier, scientists should attempt to explain its divergent results.

Of course, some of this interpretive work doubtless occurs, informally and colloquially as scientists conduct literature reviews or discuss others' work amongst themselves. However, one could argue that more granular research integration and meta-analysis should be promoted as a formal mechanism within publishing platforms and scientific technology. On this theory, publishers should curate the tools allowing meta-analysis to be performed (more rigorously than today) and the software used to generate or analyze scientific data-sets should more aggressively implement procedures to integrate data spanning multiple research sources.

In recent years, the scientific community has indeed increasingly embraced open-access paradigms such as FAIR (Findable, Accessible, Interoperable, Reusable) and the Research Object protocol, which we will cite more explicitly in Chapter 2 and elsewhere. These initiatives are intended to make scientific data more transparent and open-access, promoting study replication and/or multi-project syntheses. These same paradigms can also apply to meta-analyses themselves, of course. Even when merely aggregating prior research work, there are methodological options which can affect the outcome of a meta-analysis, and the tactics and limitations of aggregative efforts can potentially be modeled systematically. Antecedent research in a meta-analysis (or systematic review) does not necessarily neatly align; some interpretive effort and data-marshaling may be needed to finesse disparate publications and data sets into straightforward comparable. A study of Covid-19 transmissibility *among construction workers*, for example, is not precisely comparable to studies of outdoors SARS-CoV-D infectiousness in general. In order to have sufficient data in the first place (the authors of the controversy-inspiring Systematic Review report having to exclude many papers from their analysis to begin with) it may be necessary to rely on research which has certain reciprocal methodological anomalies along these lines, but perhaps there is some formal mechanism to annotate and systematize the complications which arise from unifying multiple not-fully-compatible research environments into one integrated picture.

2 Data Integration, Hypergraphs, and Type Theory

There is, of course, no magic wand which anyone could wave over the worlds of science and publishing to re-engineer meta-analytic paradigms in a single historical inflection-point. Paradigm-shifts (notwithstanding Kuhnian narratives) tend to happen gradually, and in a decentralized manner. Anyone writing a single chapter or a single book can only hope to predict potential paradigm-shifts or articulate the disciplinary dynamics which could motivate them. Indeed, in this book we are trying to take such an evaluative perspective applied to the subject of biomedical research data and data-integration practices, in both research and clinical settings.

There is plenty of on-the-ground evidence indicating that scientists are taking issues of transparent data sharing and research-replication seriously. We will consider the notion of "replication crises" in greater detail in Chapter 4. In a nutshell, though, science has been rocked by a recurring failure of reproducing research results in follow-up studies, even when the original research appears to be well-executed. Separate and apart from the merits of a single research project, then, a string of similar projects which yield convergent findings have greater scientific weight. Consequently, researchers have an incentive to spur replication studies to solidify the work they themselves publish. It logically follows, moreover, that an emergent criterion for newly published research is *how well it facilitates* potential reproductions.

This situation does indeed portend a paradigm-shift which has ramifications in numerous scientific and technological domains. The value of having research consolidated through replication can in many contexts outweigh the value of keeping some data or methodological details private (as potential Intellectual Property, say), which has helped expand the scope of the open-access data publishing ecosystem. Many publishers have likewise embraced an open-access publishing model, often charging authors fees to have their work appear in peer-reviewed contexts. One can debate the ethics of such arrangements (since many skilled scientists may hail from countries, or may be at an early stage in their career, where exorbitant publishing fees are a real burden). However, leaving evaluations

aside and simply taking open-access publishing as a given phenomenon, the momentum behind open-access models surely points to publishers' assessment of their market. There is plenty of good research which slips outside the gates of *de jure* paywalls, even when it is not technically open-access. A business model which is based on restricting access to certain academic content to paying subscribers will inevitably be eroded by readers' possibility of finding their desired materials in a free version elsewhere, or at least finding comparable freely-available work, of similar quality, from other sources.

We should quickly clarify that not all "open-access" publications or data sets are actually driven by "author fees" models; there are plenty of resources where quality research can be found involving no expense to either readers *or* authors, aside from the time it takes for authors to curate their publications. The contrast between different open-access models is not especially important to our arguments. More to the point is that open-access publication is one prominent example of a series of paradigm-shifts which have pushed the scientific community to embrace openness and transparency in sharing research data and methods. Scientists have a reputational stake in exposing their protocols to communal assessment, by analogy to how computer programmers profit more from open-access code than from closed-source alternatives, because open-source projects get vetted and refined by a large community of users and fellow developers. Commercially licensed versions of open-source projects therefore have a trust factor which closed-source projects often lack, making open-source code (at least in many commercial domains) more competitive on the open market.

Against this background we can therefore see a certain narrative emerging in how contemporary scientific research is assessed: researchers have an interest in producing work which is (to the degree possible) easily replicable, which is transparent about methods and open vis-à-vis data access, and which is poised to be integrated with other projects on similar topics. Sometimes these paradigms are explicitly enforced by publishers or funding sources. The Bill and Melinda Gates foundation, for example, in their "guidelines for authors" (themselves published as one specification with the FAIRSharing Initiative) essentially requires depositing open-access data sets on popular hosting platforms such as Dryad, Open Science Foundation, or GitHub as a precondition for fieldwork grants.[3] Similarly, academic publishers strongly encourage authors to submit data sets to open-access platforms and to reference those platforms via "supplemental materials" and/or "data availability" sections of their articles.[4] In particular, the onus is on authors to justify a *failure* to publicly share research data, such that pairing academic papers with open-access data is intended to be the norm, rather than the exception.

Aside from merely observing the industry trends toward freely-shared research data and open-access research publications, we can also draw lessons from *why* these trends are becoming entrenched. What do scientists actually hope to accomplish by making research data publicly available? One facet of this trend involves double-checking research (and also statistical/computational) methods. The text of an article may condense or summarize findings into a single table or illustration, or even just a single number (say, a 1% estimate for the proportion of Covid-19 transmissions which occur outdoors). However, researchers may find that these summaries are more convincing when they can demonstrate their provenance (by publishing data sets in their entirety rather than just statistical overviews, as well as computer code via which summarial calculations are attained, data-acquisition protocol descriptions, and so forth). Such transparency, among other benefits, can help prevent (or at least retroactively clarify) misinterpretations such as Dr. Walensky's precis of outdoor Covid-19 infectiousness.

Aside from building trust in their research, data-sharing helps research projects prepare for potential replications, as we intimated earlier, as well as for integration with comparable projects. These factors also augment the value of any one individual project.

In short, among the factors driving current data-curation and publishing paradigms is the hope that new research will be attentively reviewed, synthesized with other research, and potentially reproduced/replicated in whole or in part, all of which can make new research more valuable to the relevant scientific community which is in position to judge the work. In particular, *replication* and *integration* are key goals of new research: the potential for replication (by follow-up studies) and integration (with prior or future projects) are among the criteria by which new research is evaluated.

To the degree that these are indeed compelling motivations — researchers strive to model their text, data, and protocols to promote replication and integration — we can anticipate that research methods and tools (including software and computational tools) will be preferred that amplify these aspects of research's dissemination. In short, this can plausibly be deemed a driving factor in how scientific software, data modeling paradigms, and publication technologies will evolve in the immediate future: technologies will come to the fore to the degree that they promote research data integration and replication.

For a concrete example of these issues, consider clinical data-sharing networks such as OMOP (Observational Medical Outcomes Partnership), PCOR (Patient-Centered Outcomes Research Network) or CDISC (the Clinical Data Interchange Standards Consortium). These initiatives promote data-integration largely through conventional relational-database mergers, relying on Controlled Vocabularies or table-schema to enforce data field names and table columns. Conversely, Semantic-Web based data-integration projects such as the OBO (Open Biological and Biomedical Ontology) Foundry pursue data-integration via more free-form graph structures, regulated by Ontologies or "shape constraints" rather than by static table layouts (we discuss such indirect constraint logics

[3] See https://gatesopenresearch.org/for-authors/data-guidelines

[4] See https://www.elsevier.com/authors/tools-and-resources/research-data/data-guidelines, for example.

in Chapter 6). A variation on graph-based integration techniques, based on property-graphs rather than Semantic Web networks, is evident in the University of Pennsylvania "Carnival" project, which we summarize in Chapter 3. Other model-sharing networks emphasize data models prioritizing computational simulations and Object-Oriented representations (we will cite specific examples in Chapter 4). Moreover, individual disciplines within biomedicine and bioinformatics have their own data-sharing protocols, often based on special-purpose file types and domain-specific software, a few of which we will review in Chapter 2.

In short, researchers working in a biomedical context have multiple options for preparing publicly-shared data in preparation for potential integration with other studies (or with replication efforts seeking to reproduce their own specific work). Should data be modeled as relational tables, Semantic Web style graphs, property graphs, "objects" backed by Object-Oriented computer code, or as instances of domain-specific data formats? All of these data-representations coexist in contemporary technology with more or less comparable equilibrium, in the sense that no one paradigm dominates the others. How will data-representation strategies evolve in the future?

Insofar as paradigm-shifts in science and publishing are being driven, as we claim, by priorities of data-integration and research-replication, we can anticipate that technical details such as data meta-models will evolve under pressures from these scientific considerations. Data representations which are conducive to replication and integration will likely become more widely used; less flexible models, such as relational-database technologies, may become deemphasized.

As large-scale initiatives such as OMOP, CDISC, and PCOR evince, the conventional Relational Database model remains influential. Nonetheless, over the past two generation, the Semantic Web — and graph database technology in general — was predicted to substantially displace SQL-style technologies. That has not really happened (even if Semantic Web formats such as RDF have indeed been widely used). One explanation for why predictions centered on the Semantic Web have fallen flat is that Semantic Web models, intended to be flexible and conceptually realistic, arguably fail by their own standards — this has engendered a trickle of computer scientists criticizing Semantic Web implementations more than motives, and presenting alternative models (such as Conceptual Spaces) which we will analyze further in chapters 6 and 9. Meanwhile, graph database technology itself is more general than the Semantic Web alone, and non-RDF formats (more expressive or structurally detailed than labeled graphs as such) have emerged as popular NoSQL database models, including hypergraphs and property-graphs. These technologies, too, may become increasingly influential in structuring how data is modeled for public sharing and publishing in the future.

In short, we can predict that paradigms such as Conceptual Spaces, hypergraphs, and property-graphs will potentially become more substantial foundations for future data-sharing protocols which deviate from both Relational-Database and Semantic Web precedents. Work which synthesizes several of these developments — such as hypergraphs and Conceptual Space theory — is therefore especially interesting. One version of a unified hypergraph/Conceptual Space model has emerged in the context of Quantum Natural Language Processing, and our evaluation of that model's assumptions, potential, and limitations will be an important focus of Chapters 6 and 9.

Supplemental materials for this book will be deposited on the Open Science Foundation, Dataverse, and GitHub; for archive locations please visit https://github.com/scignscape/DataIntegration-ConceptualSpaceModeling. These materials include small bioimaging and clinical data sets which are encoded and annotated using techniques we discuss in Chapters 6-9. These data sets are not intended for research purposes themselves, but rather to illustrate ideas about how data sets may be structured, to the degree that scientists really do prioritize expressive data models, microcitations (see below), integration with machine-readable text, and other data-publishing features we mention in this introduction and elsewhere in the book. In addition, the repository contains machine-readable text of the book's individual chapters, encoded as separate documents using a special text-representation system designed to facilitate cross-references between publications and data sets.

The book's supplemental materials also include code libraries targeted at the (sample) data sets. This is in keeping with design patterns we will discuss in later chapters, which assume that data sets will in general be publishing alongside code libraries providing functionality to read and manipulate the associated data, including parsers for the data's serialization format. By sharing code alongside data — optimized as necessary for the specific information comprising a data set — programmers can shift the burden of annotation and documentation to the computer code rather than the data itself. As we will discuss in Chapter 6 and elsewhere, source code presents a richer foundation (as compared to static data models) for pre- and post-condition annotations, requirements engineering, and other technical details which can express scientific theories and research protocols. Equipping data sets with custom-designed code libraries allows the data types uniquely instantiated via those libraries' implementations to serve as models and documentations for the data set's specific profiles; such a type-theoretic foundation, in particular, allows us to examine data set organization in a systematic fashion. We will consider type theory as a date-modeling paradigm in Chapters 5 and 6.

3 Philosophy and the Semantic Web

Data-exchange formats might seem like a pretty mundane scientific topic, part of the minutia of research practice which academics attend to as a matter of professional competence, like curating biobliographic references, but hardly interesting outside its backstage role. It is curious therefore to consider that the Semantic Web has a colorful philosophical backstory and has had a relative center-stage position in the theater of Artificial Intelligence and debates over the AIs potentials and lim-

its; over whether human intelligence is mechanical enough to be digitally simulated. The concept of Semantic Web *Ontologies* has become rather conventionalized, such that so-called Ontologies tend to serve essentially as Controlled Vocabularies or Taxonomies, constraining the classifications of data within structured information spaces, and the labels used to connect disparate data points.

Ontologies are formally rooted in graph database technologies (or information spaces which emulate them, such as the Semantic Web itself). In this context, the principal elements of Ontologies are enumerations of labels which can be attached to graph nodes (providing a classification of the data set embodied by a graph) and graph edges (classifying the kinds of interrelationships that may be asserted between nodes). Ontologies are *controlled vocabularies* in the sense that conformant graphs may only utilize node or edge labels proscribed by Ontologies applied to the graph. They are *taxonomies* in that labeled terms may be sorted *hierarchically*: labels can name classification elements which are super- or subkinds, relative to other elements in their respective Ontology.

Philosophers understand the term "ontology" to name something of much greater metaphysical weight than just taxonomies on graph data-stores. Notwithstanding that background, the term "Ontology" was not chosen by accident; every *domain-specific* Ontology, essentially a data-model applied to empirical data sets, is understood by Semantic Web practitioners to be potentially unified into more general "upper" Ontologies, which have broader (and more philosophical) scope. Upper Ontologies aspire a global classification of "objects" in general, both abstract and concrete, so that in addition to domain-specific concepts and definitions (*carcinoma is a kind of cancer*, say) one has broader metaphysical annotations (cancers are "disease processes," tissues are "spatially extended regions," and so forth). The rationale for this metaphysical superstructure is rooted in AI — Ontologies seek to endow scientific and technical data (particularly biomedical data, where Ontologies are especially popular) with a conceptual scaffolding analogous to human intelligence, insofar as we instinctively conceptualize objects through the lens of spatiotemporal extension, of stasis and change, events and processes, and so forth.

Artificial Intelligence is in fact a key component of Semantic Web architecture, though in practice the role of AI is less on this metaphysical scale and more focused on the use of *axioms* and other Ontology constructions adding structural detail to Semantic data models. Ontologies employ axioms and annotations to assert constraints or patterns on how classifications and relations are used. Ontologies may assert that two relations are inverses (parenthood and childhood, say), or that one relation conceptually implies another — for instance, the relation of two people being *divorced* necessitates that they were previously *married*, such that at some prior point in time the *marriage* relation held. Instead of just sets of graph nodes and edges, then, Ontologies add logical structure of graph-form data: the *divorce* relation, for example, demonstrates how relations cluster into logical networks. Any database which represents a divorce-instance, and which is not fundamentally lacking data, would be expected to provide data about the necessarily antecedent *marriage*. We will return to *divorce* as a case-study in "multi-part" relationships in Chapter 6.

Ontologies, in short, are not just taxonomies or controlled vocabularies for graph databases; they also introduce axioms and logical constraints on graph-structured data. These extra constructions provide graph-based data models with added expressivity and precision. They also allow graph data structures to be targets for "reasoning engines" and other AI-driven analytics. Reasoning over the Semantic Web is analogous to query-evaluation, but uses methods rooted in Symbolic AI rather than the query-engine architectures typical of relational (or even NoSQL) databases. This is one sense in which the Semantic Web as a whole represents a "project" or even ideology closely aligned with AI itself. And the association between AI and the Semantic Web has also been a source of criticism, either from perspectives which are skeptical about the more holistic claims of AI (or "Artificial *General* Intelligence") visionaries, or those which feel that the Semantic Web's fairly modest data-representation paradigms do not fully harness the power of AI (or some combination of the two).

This is some of the milieu in which technical-sounding debates as to optimal data-representation meta-models become invested with unlikely philosophical gravitas. Some of the influential figures in Semantic Web evolution (and criticism) come from the realm of philosophy and humanities/linguistics, not from science (or computer science), such as Barry Smith — a scholar whose earlier work was grounded in Phenomenology and Central European Philosophy, and who pivoted mid-career to information science, spearheading the OBO Foundry — and Peter Gärdenfors, a linguist who originated Conceptual Space theory and has catalyzed certain counter-narratives critiquing the Semantic Web (mentioned earlier).

At some level these are relatively minor episodes in Intellectual History, and of course the philosophical germination of the Semantic Web has relatively little bearing on a researcher who adopts a specific OBO Ontology, for example, as a structuring device for their data. But philosophical controversies around the Semantic Web help alert us to what is at stake in data-sharing protocols. Why is a vision such as the Semantic Web — a globally synthesized network of knowledge subject to common meta-models which can be concretized in domain-specific standards, potentially synthesizing data from myriad sources — what is the intuitive appeal of a vision of standardization and broad-based integration along these lines? The underlying dynamic in debates about (say) Conceptual Spaces versus Ontologies appears to lie in scientists' search for data *representations* which seem to intuitively capture the theoretical commitments and conceptual architecture surrounding scientific data, its research origins, and its technical import. In short, scientists seem to consider research data not only as a digital artifact to be shared, but as an embodiment of a given scientific perspective and research environment. Data sets, on this point of view, should *communicate* something about the

science and research that produce them, as well as serving the practical goals of moving data from one point to another.

Implicitly, then, in the following chapters when we talk about "data sets" we will usually be considering collections of information which are more than just "raw data"; which, specifically, are organized and documented to convey some details about the data set's scientific origins. The data *models* which govern how data sets are encoded thereby need to do more than simply digitize raw data in an unambiguous fashion; instead, data models have to permit data-set curators to use the organizing principles and annotation mechanisms within each data set as communicative tools representing the appropriate scientific and theoretical background. Such requirements point beyond the formats typically used for data encoding in the past (XML, JSON, HDF, ASDF, and so forth), and we will examine criteria and architectures for more expansive and conceptually intuitive formats in later chapters. As mentioned above, we also demonstrate possible data formats and data-set architectures via supplemental materials accompanying this book.

4 Navigating the Proliferation of Research Data

Whatever scientists' motivations in curating research data, a further ineluctable detail of contemporary science and publishing is the large volume of (meritorious) work being produced. It is better to have more good science than less, to be sure, but the sheer scale of research work presents a challenge both to individual scientists (who are charged with making their contributions known to the relevant communities that can leverage them) and to technologies powering science as a whole.

Since it is impossible for any one person to read all scientific literature produced at any one point in time — or even all literature confined to a specific specialization, such as oncology, cardiac care, or Covid-19 — scientists envision AI tools that could guide researchers toward relevant papers and data sets. More advanced search engines for scientitic documents would help investigators to cut through the mass of literature and hone in on specific studies which are precedents or theoretical foundations for their own work. In short, better search tools would balance the counteract the challenges posed by how scientific work is rapidly expanding. Scientists could then have the best of both worlds: a vibrant scholarly community which produces large quantities of credible science (and increasingly lowers barriers to admission into the circle of professional academia, yielding a more diverse and representative community in terms of race, class, and gender) while, simultaneously, utilize technologies which keep them from being swamped by this very volume.

This is an encouraging idea, but there seems to be little evidence that search tools specifically designed for science perform noticeably better than generic web searches. Truly accurate publication-search capabilities appear to remain a project for the future. A good case-study in the current state and limitations of publishing technology is offered by the CORD-19 corpus, curated by the Allen Institute of Artificial Intelligence, to promote text and data mining targeted at literature related to SARS-CoV-2 and (to the degree that they may benefit Covid-19 research) coronavirus studies in general. The CORD-19 compilation provides machine-readable full-text versions of over 280,000 publications (as of mid-2021). We review the features and architecture of CORD-19 in Chapter 3. For right now we'll simply point out that the data scientists who formulated CORD-19 openly acknowledged limitations in their methodology to obtain full-text representations and to curate them in a searchable manner. They even issued a "call to action" encouraging publishers to develop "distribution formats [for] scientific papers" which are less ambiguous than PDF (i.e., less opaque in text-encoding), to share publication texts in "structured format[s] like JSON, XML, or HTML," and to embrace consistent schemata for article meta-data.[5] If the availability of machine-readable text serves as a qualification distinguishing which papers are included in aggregative corpora — and also which papers, once included, are more prominent in search results due to their being properly annotated (with demarcated keyphrases, findable data links, and so forth) — then publishers have an incentive to adopt the paradigms akin to those which the Allen Institute is advocating in this context.

Similar challenges confront finding and integrating research data across disparate projects. It is difficult to search within data sets because there is no obvious foundation to look for key phrases, for example, the way that search engines can match search terms against the raw text of a publication. There is, in general, no "raw text" within a data set that can be scanned for keywords and phrases. A related problem is that multiple data sets can be hard to aggregate together, even if they use similar methods applied to similar real-world problems. Unless there is a rigorous isomorphism between the statistical parameters and data-types employed between two kindred research projects, there is no automatic process to map one project's parameters onto another's so that they may be analyzed or visualized as a whole, or subjected to integrated statistical processing. Even subtle difference in parameters' variance, distributions, ranges, and data-acquisition methods can complicate attempts at data aggregation spanning two or more research projects.

In light of these difficulties, scientists have proposed numerous formats for describing published data sets, with the hope that common representations would make data sets more searchable and interoperable. One aspect of these standardization projects is the notion of "microcitations," or strategies to demarcate individual parts of a dataset as citable references, by analogy to citations of specific pages within a published document. The process of forming microcitation targets in the context of specific data sets, however, depends on the format through which the data is encoded. Object values in JSON, SQL table rows or columns, XML document nodes, or record-tuples for formats such as **numpy** or CSV may all be feasible microcitation sites. Given that data sets might employ any of these representations (or many others) it is not unproblematic

[5] See Lucy Lu Wang, *et al.*, "CORD-19: The COVID-19 Open Research Dataset" (https://arxiv.org/pdf/2004.10706.pdf), page 8.

to standardize a general-purpose micro-citation format.

These data-sharing and text-mining challenges are significant, but they also give us a lens with which to anticipate what kind of data curation and document-preparation technologies will become popular in the next phase of scientific publishing. It is reasonable to guess that scientists will benefit from standardized, multi-disciplinary data representations that support micro-citations, that allow publication texts to cite specific parts of data sets (much as they cite other articles), and that allow code to be re-used across multiple research projects as a means to achieve data integration. We consider this a hyptohesis as to the general priorities that will shape scientific computing and publishing technologies moving forward. The actual software engineering and data-modeling structures and design patterns that might realize these general goals will be the subject of much of our analyses in several later chapters, particularly Chapter 5, 6, and 9. This will also be a theme (in the specific bioimaging context) of Chapters 7 and 8 as well.

Most of the chapters in this book will be focused on different approaches to data modeling, such as Type Theory, Conceptual Spaces, or Graph Database architectures. Our emphasis from a *theoretical* point of view will be on data-modeling representational paradigms whose goals and criteria are oriented toward software engineering and the interoperation of distinct software components. That is to say, we advocate for data-representations whose rules and conventions prioritize the implementation of software components which produce, share, and consume the modeled data. As an underlying assumption, any database, data set, or information space should be engineered with the expectation that multiple (not fully isomorphic) software components will be interacting with that data, and that parts of such data will be passed and shared between these components, meaning that the data should be structured to facilitate cross-component communication.

From a more practical or "applied" point of view, we will call attention in particular to biomedical research projects which synthesize information with variegated disciplinary provenance and diverse data profiles. Of course, much biomedical research is inherently interdisciplinary. However, new breakthroughs and new research methods and technologies have accelerated the cross-disciplinary insights of research in several specific biomedical disciplines, yielding diagnostic, prognostic, and explanatory models that cut across biophysical scales (molecular, cellular, tissues, organs) and data-acquisition modalities (proteomics, genomics, biopsies, image processing, lab assays — such as for biologic sample analysis — and so forth). Examining literature where these integrative studies are described, it becomes clear that scientists often construct the software ecosystem powering their research in ad-hoc ways, piecing together diverse software components (sometimes standalone applications, sometimes code libraries, or some combination of the two) designed for specific disciplinary contexts.

We contend that the relatively informal and trial-and-error approach often taken to integrating multi-disciplinary biomedical data can act as an impediment to research replication and the systematic evaluation of interdisciplinary research findings. This is one reason for engaging in a detailed review of data profiles, data modeling paradigms, and data integration techniques, so as to lay the foundation for a software ecosystem which can support the emerging paradigm of transparent, replicable research data and digital scientific resources.

We do not claim any special insights into interdisciplinary biomedical methods or data-sharing as such — it is quite well-acknowledged first that data sharing is an increasingly important part of both research and clinical practice, and second that breakthroughs in fields such as oncology and immunotherapy will depend on carefully calibrated multi-disciplinary data integration. However, later chapters in this book will examine aspects of these data-integration, data-sharing, and data-modeling paradigms which we feel have been under-emphasized in existing biomedical computer science.

In particular, while it is obviously true that (given today's highly interconnected digital-health ecosystem) many biomedical and clinical data spaces are utilized by multiple (independent) software components, there are intricate design challenges which confront the engineering of data sources that can allow autonomous components to leverage their data in consistent (but flexible) ways. The impetus for digital health in recent years has, one might say, been rooted in data mining and Artificial Intelligence; less emphasis has been placed on the software engineering side. As a result, our biomedical software ecosystem arguably remains more fragmented and unsystematically designed than would be warranted based on how profoundly different biomedical subdisciplines have been connected together in recent years.

As for Dr. Rochelle Walensky's testimony whose controversy was discussed above, the C.D.C. director might be justly held accountable for misrepresentating a specific Covid-19 study, but she can hardly be faulted for attempting to base her testimony on peer-reviewed scientific literature. The problem is that — although many people believe government policies should be grounded on scientific evidence and should respect scientific consensus wherever possible — all too often there simply *isn't* scientific consensus, even in light of substantial real-world data. Such lack of consensus should not inhibit policymakers from basing government decision on data-driven, empirically-minded deliberation, but it implies that scientists need a more sophisticated model of how to translate scientific findings into public policy insofar as the science itself is sometimes inconclusive and contradictory.

One way to achieve this, we contend, is by formulating more sophisticated presentations of research data, of archives tracking multiple research projects, and of software and algorithms that could be used to integrate disparate data sources (while also modeling the anomalies and structural anisomorphisms which can make data integration inexact). The impetus for such granular and non-oversimplifying (acknowledging integration problems rather than papering them over) methodologies might come from software engineering, as much as from scientific institutions themselves.