Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Structural biology meets data science: does anything change?

Cameron Mura[1,3], Eli J Draizen[1,3] and Philip E Bourne[1,2]

Data science has emerged from the proliferation of digital data, coupled with advances in algorithms, software and hardware (e.g., GPU computing). Innovations in structural biology have been driven by similar factors, spurring us to ask: can these two fields impact one another in deep and hitherto unforeseen ways? We posit that the answer is yes. New biological knowledge lies in the relationships between sequence, structure, function and disease, all of which play out on the stage of evolution, and data science enables us to elucidate these relationships at scale. Here, we consider the above question from the five key pillars of data science: acquisition, engineering, analytics, visualization and policy, with an emphasis on machine learning as the premier analytics approach.

Addresses
[1] Department of Biomedical Engineering, University of Virginia, Charlottesville, VA 22908, USA
[2] Data Science Institute, University of Virginia, Charlottesville, VA 22904, USA

Corresponding author: Bourne, Philip E (peb6a@virginia.edu)
[3] CM and EJD contributed equally to this work.

Current Opinion in Structural Biology 2018, **52**:95–102

This review comes from a themed issue on **Biophysical and computational methods**

Edited by **Gregory Voth** and **Mark Yeager**

https://doi.org/10.1016/j.sbi.2018.09.003

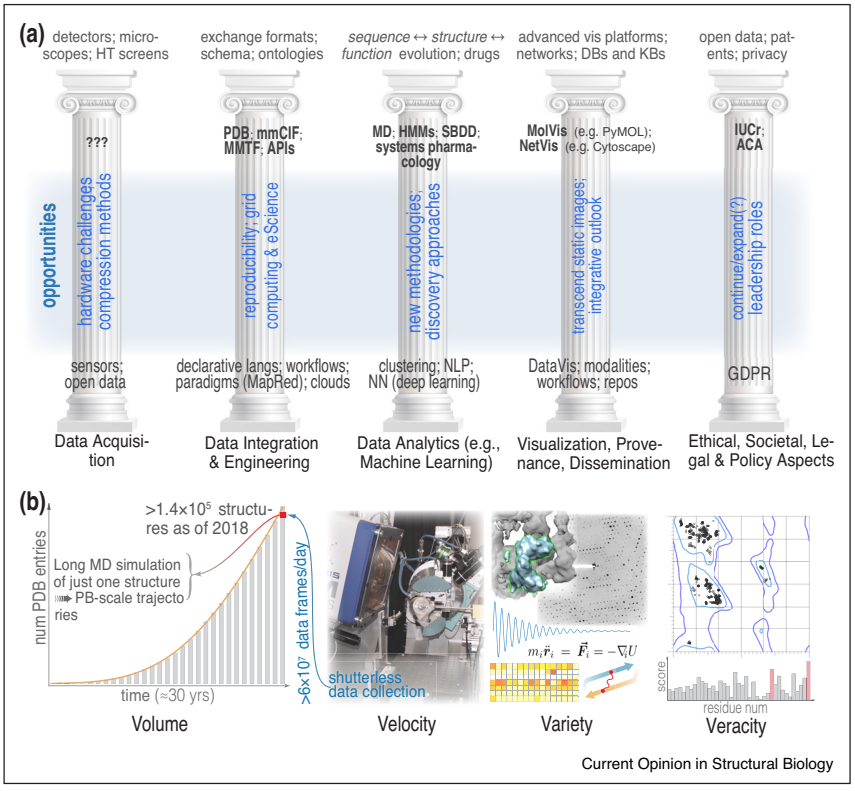
Introduction

The term *Structural Biology* (SB) can be defined rather precisely as a scientific field, but *Data Science* (DS) is more enigmatic, at least currently. The intrinsic difference is two-fold. First, DS is a young field, so its precise *meaning*-based on what we practice and how we educate its practitioners — has had less time than SB [1,2] to coalesce into a consensus definition. Second, and more fundamental, DS is interdisciplinary to an extreme; indeed, DS is not so much a field in itself as it is a way of *doing* science, given large amounts of diverse and complex data, suitable algorithms and sufficient computing resources. Such is the breadth and depth of DS that it has been described as a fourth paradigm of science, alongside the theoretical, experimental and computational [3,4]. Because it is so vast and sprawling, a helpful organizational scheme is to consider four *V*'s and five *P*'s that characterize data and DS (Figure 1).

The four *V*'s describe the properties of data: *volume*, *velocity*, *variety* and *veracity*. The *P*'s are the five disciplinary pillars (P-i through P-v) of DS (Figure 1): (i) *data acquisition*, (ii) *data reduction*, *integration and engineering*, (iii) *data analysis* (often via machine learning), (iv) *data visualization*, *provenance and dissemination*, and (v) *ethical*, *legal*, *social and policy-related matters*. The *P*'s are interrelated, as are the *V*'s. For example, the fifth pillar leans into each of the other four: a host of privacy matters surround data acquisition, aggregation can have unforeseen security concerns, analytics algorithms can introduce unintended bias, and dissemination policies raise licensing and intellectual property issues. Similarly, many modes of data analysis (P-iii) rely on advanced visualization approaches (P-iv). The *P*'s also closely link to the four *V*'s. For example, P-i, the *data acquisition* pillar, clearly relates to *volume* and *velocity*. More subtle linkages also exist, e.g., between *data analysis* and *variety*: in structural biology, hybrid approaches [5–7,8•] involve joint integration/analysis of heterogeneous varieties of data (e.g., cryo-EM, mass spectrometry, cross-linking), for instance via a Bayesian statistical formulation of the structure determination process [9,10]. The philosophy and epistemology of DS is an entire field unto itself, and helpful starting points can be found in recent texts [11••].

The rest of this review focuses on the junction of data science and structural biology. We consider DS approaches that have been applied in SB recently, including examples from crystallography and protein interactions. We focus mostly on pillar P-iii (Figure 1), and specifically machine learning. In so doing, we largely ignore traditional disciplinary labels. For example, the junction of DS and SB could be viewed as simply expanding the field of structural bioinformatics [12]; but, such disciplinary labels and boundaries matter less than the actual scientific impact. Analogously, definitions of '*the internet*' vary greatly, yet the impact of the internet on science is unmistakable. For convenience, we use the term 'SB' as including structural bioinformatics, simply to distinguish what has gone before versus what may lie on the horizon. We suspect much lies on the horizon: akin to the rapid growth [13] of databases such as the Protein Data Bank (PDB; [14]), our assessment of bibliometric data (Figure 2) suggests that data science will profoundly impact the biosciences, including structural biology. (The best-fit curve in Figure 2 is supra-exponential, with no inflection point in sight.) Conversely, can SB impact the broader field of DS? This has yet to occur in a definitive way, but, given the maturity of SB as a discipline, much can be learnt from it and its history; thus, we start with a short review of how SB might influence DS.

Figure 1



SB mapped onto the five pillars of DS, and in relationship to the four V's of big data. DS rests upon five central pillars, schematized in **(a)** as (i) data acquisition; (ii) data integration & engineering; (iii) data analytics (e.g., machine learning); (iv) visualization, provenance and dissemination; and the (v) ethical, societal, legal and policy aspects. General concepts and keywords from the data sciences are near the bottom of each column (e.g., MapReduce, a distributed computing paradigm), while more domain-specific examples rest atop each column (e.g., structure-based drug design [SBDD], middle column). A band of opportunity arises as SB meets the data sciences. Realizing these potential opportunities requires big data, which enables a question or system to be addressed via DS approaches like deep learning. The four V's of big data — *volume*, *velocity*, *variety* and *veracity* — are shown in **(b)**, illustrated by vignettes from SB. As indicated, the *volume* and *velocity* characteristics are intertwined; for instance, modern X-ray diffraction technologies enable shutterless data collection, with upwards of many millions of diffraction patterns acquired per day (a concomitant increase in the rate of structure determination means growth in the volume of the PDB). Fits of the data in the PDB histogram (b) to different functional forms — (i) a simple power law, (ii) a pure exponential, (iii) a stretched exponential and (iv) the product of an exponential and a power law — reveal form (iv) to be the best fit (orange trace). The *Variety* panel illustrates the challenge addressed by 'hybrid methods': data arise from cryo-EM, X-ray diffraction, NMR spectroscopy, molecular simulations, chemical cross-linking/mass spectrometry, phylogenetic analyses and a host of other potential approaches. DS provides a framework for integrating such data in an optimal manner (in an information theoretic sense) so as to create 3D structural models.

What structural biology has to offer data science

Open science

SB has pioneered open science through the provision of the PDB and many derivative data sources. The complete corpus of structural information in the PDB is free of copyright and is available for unfettered use, non-commercial or otherwise (P-v). Moreover, community practices—such as virtually no journal publishing an article without its data deposited in the PDB [15]—is a precedent that, if broadly adopted in other disciplines, would deepen the amount and diversity of data available for DS-like approaches in those other scientific and technical domains. The creation and free distribution of software (SW) tools has echoed this trend, as epitomized by the *Collaborative Computational Project 4* (*CCP4*); developed and meticulously maintained since 1979 [16], the *CCP4* suite has been a mainstay of the crystallographic structure-determination process. *CCP4* and kindred projects, alongside myriad other SW tools and attendant data, have fostered an open discipline. DS draws upon data and ideas from a wide range of disciplinary areas, but some of these areas have been less open than SB, at least historically. To succeed, we believe that any DS must abide by the 'FAIR' principles, enabling researchers to *Find*, *Access*, *Interoperate* and *Reuse* data and analytics [17•]. SB has exercised this for decades, and is thus positioned to lead the way.

Reproducibility

In principle, reproducibility is the bedrock of the scientific enterprise. And, as a byproduct of open science, reproducibility has been central in SB, though often less so in other realms of DS. Cultural differences across various disciplines, often driven by (perceived) competitive pressures,