CHANGING THE NATURE OF QUANTITATIVE BIOLOGY EDUCATION: DATA SCIENCE AS A DRIVER

RAINA S. ROBEVA, JOHN R. JUNGCK, AND LOUIS J. GROSS

Abstract. We live in a data-rich world with rapidly growing databases with zettabytes of data. Innovation, computation, and technological advances have now tremendously accelerated the pace of discovery, providing driverless cars, robotic devices, expert healthcare systems, precision medicine, and automated discovery to mention a few. Even though the definition of the term data science continues to evolve, the sweeping impact it has already produced on society is undeniable. We are at a point when new discoveries through data science have enormous potential to advance progress but also to be used maliciously, with harmful ethical and social consequences. Perhaps nowhere is this more clearly exemplified than in the biological and medical sciences. The confluence of 1) machine learning, 2) mathematical modeling, 3) computation/simulation, and 4) big data, have moved us from the sequencing of genomes to gene editing and individualized medicine; yet, unsettled policies regarding data privacy and ethical norms could potentially open doors for serious negative repercussions. The data science revolution has amplified the urgent need for a paradigm shift in undergraduate biology education. It has reaffirmed that data science education interacts and enhances mathematical education in advancing quantitative conceptual and skill development for the new generation of biologists. These connections encourage us to strive to cultivate a broadly skilled workforce of technologically savvy problem-solvers, skilled at handling the unique challenges pertaining to biological data, and capable of collaborating across various disciplines in the sciences, the humanities, and the social sciences. To accomplish this, we suggest development of open curricula that extend beyond the job certification rhetoric and combine data acumen with modeling, experimental, and computational methods through engaging projects, while also providing awareness and deep exploration of their societal implications. This process would benefit from embracing the pedagogy of experiential learning and involve students in open-ended explorations derived from authentic inquiries and ongoing research. On this foundation, we encourage development of flexible data science initiatives for the education of life science undergraduates within and across existing models.

1. Introduction

Significant progress has been made in mathematical biology education since the calls for change at the Cullowhee Conference on Training in Biomathematics held in 1961 [1] and particularly after the report *BIO2010: Transforming undergraduate education for future research biologists* was published in 2003 [2]. The *BIO 2010* report was an impetus to four subsequent major initiatives that attempted to implement some of its recommendations at the national policy level: The AAAS *Vision and Change in Undergraduate Biology Education* [3]; MAA's *Math and Bio 2010: Linking*

2010 *Mathematics Subject Classification.* 97m60; 92B05; 68t09; 62r07.

Key words and phrases. Mathematical biology education, data science education, education reform, big data.

The first author was partially supported by the Karl Peace Fellowship in Mathematics, Randolph-Macon College, VA. The third author was supported by NSF Award DBI-1300426 to the University of Tennessee .

undergraduate disciplines [4], the joint report *Scientific Foundations for Future Physicians* developed jointly by the American Association of Medical Colleges (AAMC) and the Howard Hughes Medical Institute (HHMI) [5], and the report *Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics* from the Executive Office of the President [6].

Concurrently, modern biology has emerged as a field offering significant professional opportunities for applied mathematicians, thus fueling the debate on how to best introduce biology into the undergraduate mathematics curricula (see, e.g., the program report on mathematical biology in the 2015 CUPM Curricular Guide [7]). Those changes have bolstered government initiatives [8], faculty development opportunities by professional groups and organizations [9], [10], [11], [12], and multiple specialized sessions at the annual Joint Mathematics Meetings of the American Mathematical Society (AMS) and the Mathematical Association of America (MAA), as well as the annual MAA MathFest meetings – all aimed at transforming existing curricula in mathematics and biology.[1] As a result of these efforts, the need for convergence of mathematics and biology at the undergraduate level is now widely recognized, and there is a general agreement that biology programs should educate students to think quantitatively and in terms of models.

Substantive changes to biology and mathematics programs however have remained slow. One reason for this may be the ever-present inertia in higher education, where new pedagogies are easy to promote and difficult to implement (see, e.g., [15],[16] [17]). Programmatic and pedagogical challenges have been magnified by numerous existing administrative hurdles – problems with establishing teaching loads for interdisciplinary and team-taught courses, narrow disciplinary criteria for tenure and promotion, and general disconnects between mathematics and biology departments are often mentioned as factors which slow reform. Additional constraints in recent years have been imposed by the tremendous growth in numbers of non-tenure-track faculty in mathematics (see e.g., [18], [19]) without institutional support or incentives to innovate and collaborate with colleagues from biology. Progress has been uneven across institutions, reflecting differences in institutional cultures and readiness to embrace change.

While the debate regarding how to effectively foster mathematical biology education continues, technological advances have brought big data to biology and medicine research, significantly accelerating the pace of scientific discovery. Examples that demonstrate the role of big data approaches across the breadth of biology abound: novel data collection methods including aerial remote sensing and satellite imagery as well as intensive ground-based data-collection in projects such as NEON and LTER are now used widely in ecology, conservation biology, and natural resource management, to mention a few [20], [21], [22],[23], [24], [25]; new data methods are applied in many evolutionary contexts including evolutionary morphology [26], [27]; genome-wide association study algorithms are used to identify possible links between DNA variants and specific deceases [28]; next generation sequencing methods have brought about faster and cheaper techniques that have revolutionized molecular biology and provided insights and actionable results for precision medicine [29]. In the medical field, big data approaches already drive high-throughput technologies [30], imaging (e.g., X-Ray, CT scan, MRI, ultrasound [31], [32], [33]), remote sensing from personalized devices (e.g., heart rate, blood glucose [34], [35]), including their use for potential contact tracing in public health

[1]Two articles in this special issue discuss in more detail societies, communities, and organizations whose main focus is to support mathematical biology research and education [13], [14].

settings, time-course data (e.g., EKG, EEG [36], [37]), integrative brain modeling [38], [39], [40], and identifying anti-aging compounds for humans [41], [42].

Thus, in the era of big data, biology has become more interdisciplinary and more quantitative than ever: discovering new knowledge hidden in petabytes of data depends on using mathematics, statistics, computer science and technological innovation – all attributes that define the emerging field of data science. This new reality has compounded many of the existing challenges in quantitative biology education and raised new ones.

Our view is that we need a paradigm shift toward biology education that evolves in tandem with changes and research advances in the discipline – a reform that combines essential data science approaches with biology content, modeling skills, and societal awareness. We present possible paths forward and advocate that a change of such magnitude could succeed only if it reflects the combined will and expertise of faculty, administration, and professional communities and only when all parties are prepared to enact significant changes in existing curricular models.

The rest of the paper is organized as follows: We begin with a discussion of major trends in biology arising from the explosion of data and its impact on education; proceed with an analysis of some parallels between biology and data science education; summarize the challenges faced by biology education arising from the rapidly-expanding set of quantitative topics appropriate for inclusion in the curriculum; and provide suggestions for how the math biology education community might address these challenges at various levels in educational hierarchy (courses, departments, institutional, and professional levels).

2. Big Data Trends in Biology and Impact on Education

2.1. **Big data – a catalyst for change.** The term *big data* refers to datasets of enormous sizes – data so large that specialized methods may be necessary to carry out analyses. Such data sets are often described by the following five properties (termed the five V's of big data): velocity, volume, value, variety, and veracity. Those refer respectively to the fast pace of data generation, its volume that requires new methods for storing and analyzing data (e.g., distributed systems), the value of data to answer questions of specific interest, its diverse origins and formats (data coming from a variety of sources have different, and often incompatible formats), and the trustworthiness of the data. The last property is essential – if the accuracy of data is questionable, results from analyses will likely be questionable too. More recently, some authors have suggested adding four more V's to the list: variability, visualization, volatility, and validity. The first two refer to how data is captured (dynamic formats of data) and how it is presented (novel ways of big data visualization are often the only way for viewing and understanding the analyses and results). Volatility refers to the rate of change of data over its lifetime, while validity refers to ensuring its integrity. [43].

Big data have already transformed our ability to generate and collect data: in biology, databases of genomes and proteins are growing at an unprecedented pace; in health monitoring, apps and mobile devices beam user measurements into massive storage facilities, and in social science, millions of posts and tweets are added to the public records every day. In 2017, an IBM study found that 90% of all data existing on public and private storage devices had been collected during the two prior years [44]. Stunning as this fact was in 2017, the rate at which data are currently being generated worldwide [45], makes an understatement in 2020.

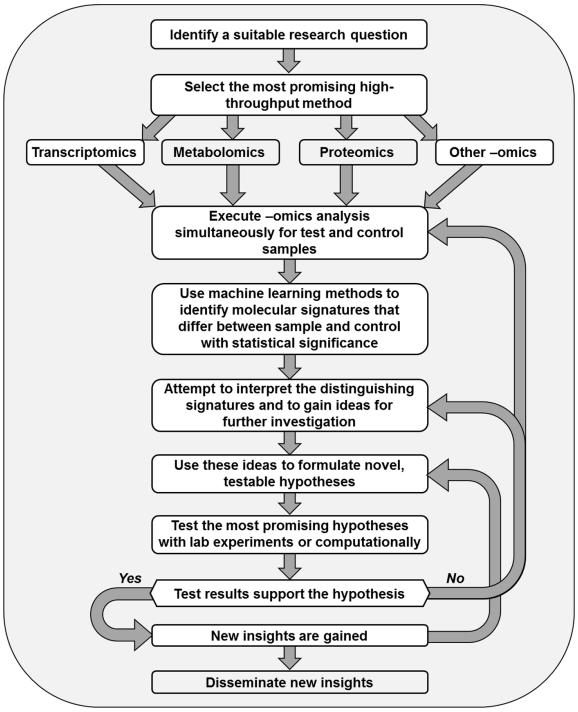

FIGURE 1. Dimension of data-mining–inspired induction. Data-driven research begins with an untargeted exploration, in which the data speak for themselves. Machine learning extracts patterns from the data, which suggest hypotheses that are to be tested in the lab or computationally. From: Voit, E (2019). Perspective: Dimensions of the scientific method. *PLoS Comput Biol*, 15(9). CC License [47].

The discipline of data science has emerged from this reality as an inherently multidisciplinary and transformative field. The fusion of big data, mathematical and statistical modeling, and computational technology have brought about powerful data mining techniques and a new generation of artificial intelligence, machine learning, and deep learning approaches, assisting progress in science, business, sociology, medicine, commerce, and education among others [46]. (See Figure 1 for a schematic of data mining-inspired induction at the omics level). These developments have demonstrated that once we break free from traditional disciplinary silos, there is virtually limitless potential for strengthening the power of data-driven discovery.

This rapid rise of data science approaches to multitudes of questions has also left some with a sense that the historical precedence of hypothetico-deductive approaches to science – those that construct and rely upon general theories – may no longer be a dominant mode. With availability of extensive data sets, data-driven discovery approaches for pattern detection (and associated prediction) devoid of a general, abstract framework have even led to comments noting that this is the "end of theory" [48].

Our view however is that, in addition to the great potential for data-mining and machine learning methods to provide novel insights from extensive data sets, there are many reasons to preserve

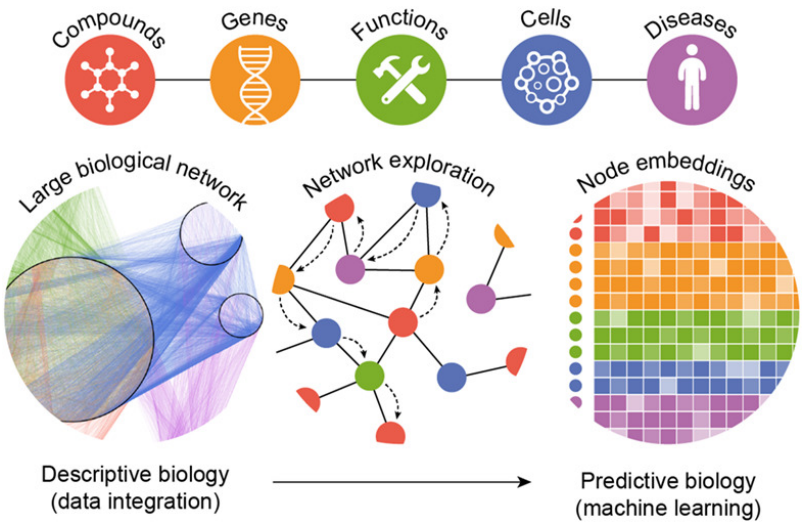

FIGURE 2. From Descriptive biology to Predictive biology. Reprinted with permission from Duran-Frigola, Miquel, Adrià Fernández-Torras, Martino Bertoni, and Patrick Aloy (2018). Formatting biological big data for modern machine learning in drug discovery. *Wiley Interdisciplinary Reviews: CMS*: e1408 [54].

educational efforts that utilize traditional investigative methods. This is particularly appropriate in biological sub-disciplines for which the massive amounts of data needed to develop effective machine learning methods are not readily available, or for which there are ethical considerations associated with obtaining and maintaining the data.

2.2. **Big data impact on research in the life sciences.** Big data in biology and medicine have allowed us to better understand how individual components at all levels of biological organization form systems interacting with one another, among themselves, and with the organisms' ecosystems. These interactions form the informational pathways in living organisms and their environment, and various stresses and diseases perturb these signaling networks in different ways. With higher speed and decreased cost of systems and omics approaches, it is increasingly possible to screen populations for genes and molecules of interest and use deep learning to discover patterns in data, classifying patients accordingly, and developing targeted therapies (see, e.g., [49], [50], [51], [52]). Today "Data have become a resource, rather than a result . . . "[53], and we have transitioned from "descriptive biology to predictive biology" [54] (see Figure 2), leading to promises for future medical breakthroughs, more potent medications, novel treatments, and disease-management strategies. Quantitative approaches have become essential to many sub-disciplines of biology in which research advances at the intersection of the life sciences, mathematics, statistics, and data science have demonstrated the importance of modeling, computations, and data driven approaches. Moreover, for those fields, it appears increasingly difficult to draw a line between the terms "biology" and "quantitative biology."

Still, finding ways to fully utilize the potential of big data for biology and medicine remains challenging, as the field is too broad and heterogeneous to rely on "one size fits all" approaches. There

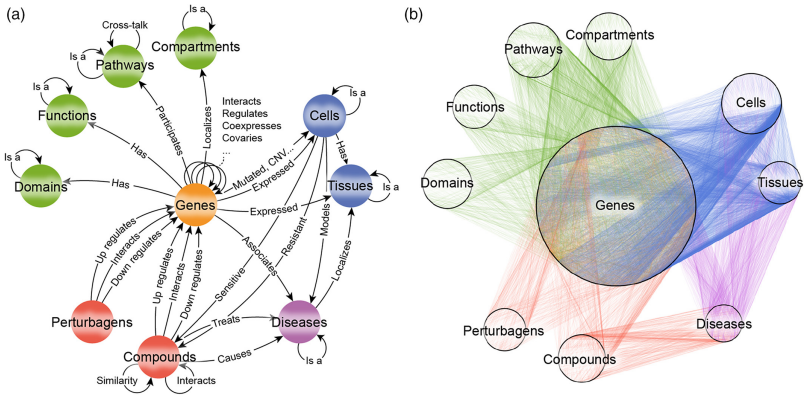

FIGURE 3. Heterogeneous network of biology. Reprinted with permission from Duran-Frigola, Miquel, Adrià Fernández-Torras, Martino Bertoni, and Patrick Aloy (2018). Formatting biological big data for modern machine learning in drug discovery. *Wiley Interdisciplinary Reviews: CMS*: e1408 [54].

are real differences in data types and structure across biology – e.g., data needed for environmental science and ecology are conceptually different in scale and form from data in genetics or proteomics or many forms of human health data. Biomedical data are heterogeneous – assembled by different means, from various samples, at different levels of biological organization and granularity, and reflecting different time/space attributes (see Figure 3). It is also generally unstructured (that is, the data are not organized in a predefined manner) and this problem is expected to only intensify with time – IDC anticipates that by 2025, most data will be unstructured [55]. Further, biological data are multi-dimensional – containing information from various scales of biological organization: from subcellular to geospatial (see Table 1). It is spread over thousands of databases, and millions of published articles; despite efforts to create open data repositories, a large fraction of biological data are only available by request from the authors. More often than not, big data in biology are incomplete and imprecise – for example, Manzoni et al. note that information in existing interactomics databases may overlap by as little as 50% [56]. Biomedical big data is "diverse, complex, disorganized, massive, and multimodal..., being generated by researchers, hospitals, and mobile devices around the world" [57]. These challenging aspects of biological data are certainly not limited to biomedical applications but occur across the range of biology.

This situation adds further challenges to the many "Vs" of big data and causes significant difficulties related to organization, storage, retrieval, verification, and processing of big data. Gaining meaningful insights often requires a detailed understanding of the origin and format of the data, and calls for *ad hoc* means of collecting, wrangling, and structuring. Frequently, this also necessitates *ad hoc* approaches to modeling, hypervisualization, algorithmic optimization and machine learning methods to uncover valuable knowledge "lost in [the] literature and data landslide" (see e.g., [58] [50], [54]). Large annotated datasets, from which algorithms learn to make correct classifications are still hard to find. The FAIR principles for responsible data management and stewardship of scientific data (Findability, Accessibility, Intraoperativity, and Reusability), created by a diverse international set of stakeholders – representing academia, industry, funding agencies, and scholarly publishers – are designed to guide and guard the veracity of data. These principles are expected

to enhance the ability of machines to automatically find and analyze data, support its reuse and increase our ability to discover and use data for creating new knowledge [59], [60]. A wider adoption of the FAIR principles, together with the emergence of novel algorithms that require smaller data sets and are less sensitive to experimental variation, is expected to broaden the use of deep learning approaches. Finally, questions concerning biomedical data processing, analysis, privacy, confidentiality, ethical use, and data sharing are far from settled, and debates regarding the ownership of data, proper approaches to medical forecasting, and various uses of data for commercial enterprises are ongoing. Those pressing questions require a mindful approach toward the interplay between societal needs, priorities, ethics, and policies, and an acute awareness of the possible consequences, should data be used for malicious purposes.

Progress is still needed on the theoretical front: while there are many deep learning algorithms that appear to work well, there is no solid theory that provides a set of conditions ensuring that the methods will always lead to a desired outcome under those conditions (even though we know that those algorithms performed well with the data sets they have been tested on). This may not be a serious problem when such methods are used to train systems to play chess or Jeopardy but may lead to severe implications if algorithms malfunction while diagnosing patients, designing treatment strategies, or making health management decisions. Thus, we will benefit from methods to close existing theoretical gaps and facilitate progress in aspects of theoretical machine learning that have so far been intractable [61]. There is also significant need for better coupling of mathematical models and big data. Hierarchy in biology leads to highly connected systems and big models have to work across scales and focus on linking data and patterns of different types – e.g., integrating multiple heterogeneous data by multimodels to incorporate different scales of biological organization and different quantitative approaches and scales of detail; better coupling of complex layered tools like GIS, with modeling and integration of non-spatial-temporal data; and systems level approaches based on model-based machine learning.

2.3. **Biology inspired data science.** Historically, there has been considerable association and feedback between advancement in the mathematical and biological fields. Mathematical analyses applied to models of biological phenomena at many scales have provided novel biological insights, suggested new experiments/observations, and served as effective alternative "microscopes" for biology [62], [63]. The reverse has also occurred: questions from biology have driven novel mathematics, such as the growth of studies on reaction-diffusion equations (see, e.g., [64]), analysis of spatial control problems (e.g., [65]), and virtually the entire field of evolutionary game theory (see, e.g., [66], [67]). Examples of novel mathematics in algebra, geometry and combinatorics can be found in Bernd Sturmfels's "Can biology lead to new theorems" [68], as well as in the article by Macauley and Youngs in this special issue [69].

The fields of biology and data science have similarly influenced one another. However, while the benefits to biology from using big data and data science approaches are widely acknowledged, the benefits to data science from its intersection with biology still need to be highlighted. In fact, many methods and algorithms empowering artificial intelligence have been influenced by neuroscience, ecology, evolutionary biology, and genetics, emulating the behavior of living organisms. Artificial neural networks, for example, come in multiple types and topologies and are designed to mimic the work of neurons in the brain. They are designed to learn, improve, and grow as more data are used for their training (see, e.g., [70]).

Organismal swarming (e.g., birds, bees, fish) has given rise to Swarming Intelligence models as well as optimization methods in machine learning such as Particle Swarming Optimization and Ant Colony Optimization. Similarly, Genetic and Evolutionary Algorithms are based on mechanisms inspired by genetics (e.g., mutation, selection, reproduction) and concepts from evolution (fitness, evolutionary landscapes) (see, e.g., [71]).

Biology and data science advances here go hand in hand. Our desire to build intelligent systems is deeply connected with growing knowledge of biological systems, and the powerful combination of big data and increased computing capacity is now bringing this goal within reach. Even though artificial systems that teach themselves and learn in ways humans do have not entered the mainstream yet, the progress of biology-inspired AI clearly demonstrates that there are still no better models to follow in data science than those evolved in living systems. In this, biology informs progress in data science more than any other discipline. It will be interesting to monitor whether specific biology domains would matter more (or in a different way) than others in facilitating progress in data science and if, in the long-run, future progress in data science would assist with reintegrating biology or, instead, enhance the existing sub-disciplinary silos.

2.4. **Data science and the ongoing math-bio education debate.** The definition for what data science is as a discipline is still emerging, even though the fundamental understanding of its nature has been developing at least since the early 1960s. In his highly-influential and controversial 1962 paper in the Annals of Mathematical Statistics [72], John Tukey discusses the evolution of mathematical statistics toward a new discipline, which strongly resembles what we call today data science (when Tukey talks of "data analysis," his meaning is very close to what we now refer to as data analytics). He describes it as being comprised of "procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data." He further argues that this definition transcends the fields of mathematics and statistics and should therefore be considered a separate branch of science. It is notable that nearly sixty years later, some still see data science "as a broader, task-driven, and computationally oriented version of statistics" [73], encapsulating a widespread perspective that data science is nothing new, just a modern repackaging of well-known statistical techniques. Others would argue convincingly that data science is not just glorified statistics [74].

In our view, even though data science has its roots at the intersection of mathematics, statistics, and computation, we see its province extending far beyond any of the three disciplines. The use of big data to address questions from medicine, biology, environmental science and politics could lead to findings involving ethics, public policy, and international affairs. Thus, instead of looking for an exact definition, it would be more practical to identify a set of skills needed for learning from big data and the conceptual underpinnings that empower students to develop these skills.

The NAS report "Data Science for Undergraduates: Opportunities and Options" published in 2018 [75] provides a broad outline of what this skillset may look like and makes a call to encourage academic institutions to create curricula that would introduce students to the fundamentals of data science early in the college cycle. The set of essential skills includes associative thinking, the ability to use computation, data, statistics, mathematical methods and models, collaborative attitude, and the capability to present results to multidisciplinary audiences. How to best cultivate such skills raises some challenging questions, many of which the mathematical biology community has grappled with in the last few decades: How can we develop a workforce of biologists, capable of

navigating the interdisciplinary landscapes of mathematics and data science? How do we bridge the gap between research, industry needs, and education? What approach to teaching would be most beneficial for a rapidly evolving discipline? How can we measure progress and evaluate effectiveness of diverse modalities of learning? And, in the context of unprecedented pace of biological and medical discoveries and advancement in technology: How do we create engaged problem-solvers and life-long learners?

Sadly, the transformational changes and challenges brought to biology by data science have remained largely invisible on the education front. Except for student research projects, biostatistics courses with focus on applications in certain biology sub-disciplines, and unconventional isolated courses at institutions where dedicated faculty have taken on designing and teaching such courses, the existing biology curricula appear to have remained oblivious to the ongoing data revolution in biology. The gap between research and education is growing and there is pressing need for comprehensive reform: we need to ensure that biology students are aware of the capabilities data science brings to modern biology and of a range of methods across quantitative science that can be used to address important questions in biology using big data. Batut et al. [76] provide the following assessment of the current state of the biomedical workforce: "The primary problem with the explosion of biomedical datasets is not the data itself, not computational resources, and not the required storage space, but the general lack of trained and skilled researchers to manipulate and analyze these data." To train those skilled researchers, educators might begin by making sure students at the undergraduate level are exposed to important examples of how mathematical models and data benefit biology, understand the fundamentals of data wrangling across variable data to utilize them effectively, and are generally aware of the existing challenges of working with big data. The ongoing dialog for advancing data science education further accelerates the need for substantive changes in biology programs toward problem solving, modeling, and quantitative approaches using big data.

3. Data Science and Biology Education

As educators, we strive to create classrooms where students learn through immersion in the methods, culture, and practices of their disciplines, and employ the most-promising pedagogies toward achieving the learning outcomes we consider essential. The education research literature offers abundant evidence that engagement and learning attitudes improve significantly when students are guided through projects, independent discovery, teamwork, and small-group discussions [77], [78]. Yet, despite numerous calls for a shift toward experiential-learning methods and project based-pedagogy, STEM classes are often still dominated by lectures [79]. This is particularly disconcerting for biology education: as content and technology become obsolete rather quickly, teaching facts and skills that are specific only to a certain software, platform, or method is a poor fit for a rapidly evolving landscape. We should instead strive to find ways to help students develop their own ability and capacity for independent learning. Involving students as much as possible in supervised research and authentic projects would teach them to generate and test ideas, encourage them to check out various sources, and provide an environment where they could "learn on the job."

The desired progression toward educating students who take charge of their own education has led to the evolution from pedagogy and andragogy to heutagogy – a form of self-determined learning with "emphasis placed on development of learner capacity and capability with the goal of producing learners who are well prepared for the complexities of today's workplace" [80] – see Figure 4. In our view, the principles and practices of heutagogy may be an effective vehicle for continuing

Biological Level	**Example of Big Data**	**URL**
Continent	National Ecological Observatory Network	https://www.neonscience.org/
Ecosystem	Long Term Biological Research	https://lternet.edu/
Phylogenetic Biodiversity	Tree of Life	http://tolweb.org/tree/ https://cyverse.org
Population	Haplotype Map Gene Haplotype Alleles	https://www.broadinstitute.org/ international-haplotype-map-project/ haplotype/haplotype-map https://genome.ucsc.edu/l goldenPath/help/haplotypes.htm
Organism	Physionet Flybase Wormbase iDigBio	https://physionet.org http://flybase.org/ http://www.wormbase.org http://www.iDigBio.org
Tissue	Allen Brain Map	https://portal.brain-map.org/
Cell	American Type Culture Collection	https://www.atcc.org/en/Products/ Cells_and_Microorganisms/Cell_Lines.aspx
Genome	GenBank	https://www.ncbi.nlm.nih.gov/genbank/
Metabolisms	KEGG EcoCyc	https://www.genome.jp/kegg/pathway.html https://ecocyc.org/
Transcriptomics	Mammalian Transcriptomic Database	http://mtd.cbi.ac.cn/
Proteome	PDB Protopedia	https://www.rcsb.org/ http://proteopedia.org/
Small Molecules	Human Metabolome Database	http://www.hmdb.ca/
Atomics Level	Biological “Periodic Table”	https://i-biology.net/ibdpbio/ 01-cells-and-energy/ periodic-table-for-biologists/

TABLE 1. **Some existing databases at various levels of biological organization.**

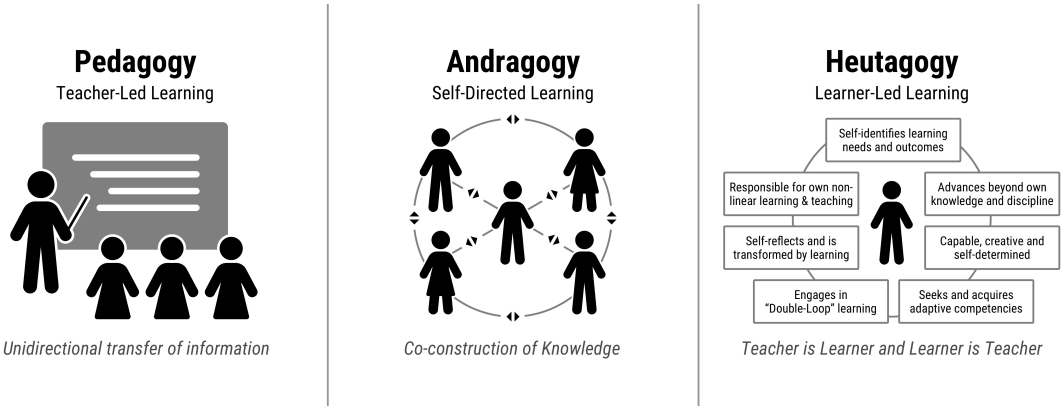

FIGURE 4. Over the past century and a half, models for education have gone through major shifts: from a broadcast transmission model (pedagogy) promoted in the 19th century ([81], [82]), to a more learner-centered model (andragogy [83]) developed by John Dewey in the 1930's and elaborated in the 1960's by Joseph Schwab [84] and Jerome Bruner [85], to a model influenced by the massive potential of the World Wide Web (heutagogy) that focuses more on students' own motives for learning ([86], [80], [87]).

education of students in rapidly advancing areas such as data visualization, and generally the use of quantitative methods and models in biology.

3.1. **Parallels between biology and data science education.** The two fields share some common traits and challenges.

A need to educate a qualified workforce. In many ways the NAS report "Data Science for Undergraduates: Opportunities and Options" [75] may develop to be for data science what Bio 2010 was for biology. It makes a clarion call for change, and its broad recommendations for inclusivity, diversity, good practices, and interdisciplinarity parallel those made in the Bio 2010 report for creating modern programs in biology. Both reports recognize the need for faculty development, the significant challenges that creating or overhauling existing programs present for administrative and organizational units, and the need for encouraging cooperation and collaboration between units and institutions.

The recommendations of the "Data Science for Undergraduates" report are not specific to any disciplinary domain of application and as such are not immediately applicable to the biology curriculum. Students in the physical sciences will likely find the integration of data science in their coursework quite natural, while biology students, who may have less affinity to the quantitative or technical fields, could see it as taking an already existing quantitative challenge to a higher and insurmountable level (unfortunately, the latter may also be true for many biology faculty). In this non-specificity however, one could also find a parallel with BIO 2010 and subsequent reports related to biology education – they emphasize a need for cultivating quantitative and modeling skills while mostly avoiding prescriptive lists of topics from mathematics and statistics to be included in the curriculum.

The reason is twofold: First, the various sub-disciplines of biology require different mathematical, statistical, and modeling approaches. Similarly, the toolkit of data science is broad and diverse – the types of data and methods that work for one discipline may not work for or be optimal for another. Biology students may need to learn specific data science concepts essential to biology, due to the importance of non-linearity and hierarchical levels of organization, that may not benefit physics students in the same way. Simply put, there is too much biology, and too much data science, making a comprehensive curriculum impossible. Second, because of this wide diversity of approaches, different institutions, institutional units, and individual faculty would need to decide how to develop a curriculum that serves their interests and institutional goals best.

Not surprisingly then, and just like existing mathematical and computational biology programs, undergraduate programs in data science come in many shapes and sizes. The 2020 "Discover Data Science" list provides details on over sixty undergraduate programs, showing a wide variety of requirements, names, and department/school affiliations [88]. In many cases, a common core of foundational courses in computer science (e.g., Computer programming, Data structures), mathematics (e.g., Discrete mathematics, Linear algebra), and statistics (e.g., Statistical Reasoning, Statistical modeling and regression analysis) is complemented by advanced course electives (e.g., algorithm design, machine learning, Bayesian statistics, data mining and visualization), requirements related to a discipline of application, and a capstone. For the latter, students are encouraged to work in a research group or with a local business or organization to gain experience in solving authentic problems.

A need for appropriate curricular materials. Many new introductory and intermediate-level data science textbooks have been published in the last few years to assist faculty in developing curricular materials at the junction of biology and data science. New texts on data science for biology and the life sciences complement those on bioinformatics and biostatistics. The titles below are listed to assist faculty as starting points for new courses they might wish to develop. The resources in this area are rapidly expanding and constantly in flux, so this list simply represents a small set of possible options.

Selected recent introductory texts in data science:

- Brian Godsey (2017). *Think Like a Data Scientist: Tackle the data science process step-by-step* [89];
- Jeffrey Saltz, Jeffrey Stanton (2017). *An introduction to Data Science* [90];
- Oliver Theobald (2017). *Machine Learning for Absolute Beginners* [91];
- Mark Fenner (2019). *Machine Learning with Python for Everyone* [92];
- Jon Krohn et al. (2019). *Deep Learning Illustrated* [93].

Selected recent data science books at the intermediate and advanced levels

- Joel Grus (2019). *Data Science from Scratch* [94];
- Max Kuhn, Kjell Johnson (2013). *Applied Predictive Modeling* [95];
- Aurélien Géron (2017). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* [96];
- Sebastian Raschka and Vahid Mirjalili (2017). *Python Machine Learning* [97].

Selected data science texts for biology and medicine

- Ramsundar et al. (2019). *Deep Learning for the Life Sciences* [98];
- Daniel Zelterman (2015). *Applied Multivariate Statistics with R," Statistics for Biology and Health* [99];
- Alan Moses (2016). *Statistical Modeling and Machine Learning for Molecular Biology* [100].

Recent new texts in bioinformatics and statistics for biology:

- Compeau and Pevzner (2018). *Bioinformatics Algorithms: An Active Learning Approach* [101];
- Dan MacLean (2019). *R Biolinformatics Cookbook* [102];
- Arthur Lesk (2019). *Introduction to Bioinformatics* [103].

Many additional resources are available, although scattered and catalogued under different disciplinary labels – statistics, computer science, mathematics, bioinformatics, data science, biology, life sciences, environmental science, and others. As the number and types of methods, approaches, tools, and textbooks for data science grows rapidly, even the most ambitious curricula won't be able to cover them comprehensively. Thus, effective prioritization will be critical.

A need to create flexible programs. With the rapid pace of discovery in both disciplines, "content" coverage has been growing, not leaving much room in the curriculum for flexibility. However, as different questions may require completely different algorithmic, computational, and mathematical approaches, both data science and biology programs will need to provide different degree pathways for students. For data science the pathways would depend on the content domain; for biology – on the biology sub-discipline. It is thus baffling that many biology programs around the country still require only a semester or two of calculus, which in many instances is taught without any connection to the life sciences. Some institutions may instead require statistics but, again, in many cases completely disconnected from biology. There are also programs that require both courses and those that, unfortunately, require neither. Even in cases when institutions have developed specialized calculus or statistics courses for the life sciences, it may not be prudent to require them to the exclusion of other mathematical topics. Calculus and statistics, for example, do not align well with questions regarding gene regulation and trophic networks – for those, having background in discrete mathematics and graph theory would be more helpful. Linear algebra and probability also provide methods of fundamental importance to certain questions in biology and data science; yet, there may be no room in the curriculum to require them.

Thus, biology and data science need to build flexible programs of study that would permit student exposure to several mathematical topics. Such programs would need multiple entry points, breaking away from the current norms of linearly structured biology curricula. Similarly, data science curricula would need to be flexible in allowing the disciplinary domain of application to drive the data science focus. Through the successful initiatives that have created specialized mathematics courses for the life sciences, student projects and interdisciplinary modules, and student engagement in research, biology could provide examples and offer educational models that data science faculty might find useful. Many articles in this special issue describe such successful programs, and they all have a shared goal of teaching biology students to work with data, models, and quantitative techniques that would require a diverse mathematical toolbox. The main challenge now is in finding ways to translate these successful approaches to foster adoption at other institutions and accelerate the pace of reform.

A need to instil ethical norms. Both biology and data science face questions regarding privacy and ethical use of data. When biology students work with data and models that have the potential to affect public health (epidemiology, vaccination, sanitation), medicine (diagnosis, treatment, prognosis), agriculture (food safety, land management, water utilization), and pharmacology (efficacy, toxicity, allometric scaling, chronotherapy), it is important that they learn about data privacy and ethical use of data and comply with all established protocols. Those include, e.g., de-identification, data scrubbing, ensuring privacy, and limiting access, as well as the use of animals in medical research. In a similar way, we should make sure to include the ethical underpinnings of data and models when designing new data science curricula. Students not only need to be able to analyze big data but they also need to be concerned with what data should be collected, how it is collected, stored, analyzed, and used, and with any policy ramifications of the models that they develop. It would be beneficial for students in either degree program to be aware of the Institutional Review Board policies in place at their college and include the training for these at some point in their curriculum.

Cathy O'Neil's book "Weapons of Math Destruction" [104] lays out a variety of issues that may arise because of "the power and risk of mathematical models." Our students will potentially face many of the same issues of data privacy and use when seeking employment, obtaining insurance, applying for funding, and seeking further education. As educators, we should help them understand the potential abuses of big data and prepare them to be professionals who reverse O'Neil's concerns about current practices that "increase inequality and threaten democracy."

Here again, biology curricula could inform the general data science education debate: biology programs have grappled with these questions for a long time, and institutions that have incorporated these elements into their existing biology programs may provide roadmaps or supply 'adopt and adapt' building blocks for new programs in data science that also align with the FAIR principles for data management and stewardship. Because big data often comes from populations, it is not surprising that public health researchers have faced many issues involved with big data. Figure 5 maps the ethical issues of digital disease detection – the use of internet, media sources, and technology to indicate outbreaks and spread of diseases. Modern survey technologies for disease outbreaks are prevalent in the news, online, and on social media, so data on the spread of infectious diseases can be made publicly available faster than through traditional surveillance systems that rely on data coming from physicians' offices or hospitals. Good intentions for early detection and control based on such digital disease detection systems, however, may lead to (intentional or unintentional) undermining of established ethical norms, and data science students need to develop understanding and awareness for possible ethical breaches.

Rapp and Tirassa [106] identify four vital elements in the handling of big data: (1) Privacy; (2) Accuracy; (3) Property; and, (4) Access. These are the nuts and bolts of ethical practices in big data analysis, but they leave out some social justice and equity considerations. Richards and King note [107] that the ethical issues of big data are so revolutionary that we might perceive we are on the cusp of a massive qualitative social change: "The potential for social change means that we are now at a critical moment; big data uses today will be sticky and will settle both default norms and public notions of what is 'no big deal' regarding big data predictions for years to come. Individuals have little idea concerning what data is being collected, let alone shared with third parties. Existing privacy protections focused on managing personally identifying information are not enough when secondary uses of big data sets can reverse engineer past, present, and even future breaches of privacy, confidentiality, and identity." And, as Vayena et al. [105] stipulate, "At the crux of the

Categories	Ethical Challenges	Specific Examples	Values
Context sensitivity	Differentiating between commercial versus public health uses of data	Is identification permitted? Is consent required for DDD uses? If so, has consent been obtained? Can it be revoked?	Privacy and contextual integrity
	User agreements, terms of service, participatory epidemiology	Are users protected in all contexts irrespective of privacy laws that differ according to jurisdiction?	Transparency
	Global health issues	Are privately collected data open to global public health uses?	Global justice
Nexus of ethics and methodology	Robust methodology: algorithm validation, algorithm recalibration, noise filtering, and feedback mechanisms	False identification of outbreaks and inaccurate predictions of outbreak trajectory	Risk of harm
		Pressure to mobilize public health resources in light of rapidly spreading unvalidated predictions	Fair use of resources
	Data provenance	Awareness about public health uses of personal data (in aggregated form)	Trust, transparency, accountability
Legitimacy requirements	Best practice standards	Is there a shared code of practice amongst all those working on DDD?	Trustworthiness
	Monitoring bodies (policies for ongoing monitoring and action plans for correction of false results)	Is there a mechanism for quick response to inaccuracies about outbreaks?	Trust, transparency, accountability
	Paced integration of DDD to standard surveillance systems	Are there mechanisms for redressing harms caused by DDD activities?	Justice
	Communication to the public (prevent hype)	Management of expectations	Common good

doi:10.1371/journal.pcbi.1003904.t001

FIGURE 5. Mapping the ethical issues in digital disease detection. From: Vayena E, Salathé M, Madoff LC, Brownstein JS (2015) Ethical Challenges of Big Data in Public Health. *PLoS Comput Biol* 11(2). CC License [105].

debate on the ethics of big data lies a familiar, but formidably complex, question: how can big data be utilized for the common good whilst respecting individual rights and liberties, such as the right to privacy? What are the acceptable trade-offs between individual rights and the common good, and how do we determine the thresholds for such trade-offs?"

As educators and mentors, we will need to find good answers to these questions and assist students develop a strong moral compass and grow sensitive to possible unintended consequences of their actions.

A need for assessment and evaluation practices. Two scales at which the effectiveness of novel and adapted programs at the interface of data science and biology can occur are that of assessment of particular modules and teaching methods, and that of evaluation at the broader level of courses, curricula and programs. The scientific teaching movement [108] and the discipline-based education research program [109] emphasize the benefits of applying education research methods to determine the effectiveness of teaching, which is also appropriate for initiatives at the interface between disciplines. The NAS report [75] notes the importance of evaluation for the new flexible programs in data science and suggests that data science can provide novel methods to coordinate cross-institutional data to compare the effectiveness of alternative implementations. Examples of efforts to compare alternative teaching methodologies at the course level include the development of instruments for quantitative reasoning in biology [110] and for biocalculus [111]. These examples are steps towards educational research to enhance the efficacy of course-level designs that incorporate quantitative methods in a biological context. As new curricula and programs are developed, a more continuous, regular evaluation approach will be beneficial as these programs evolve through the rapidly changing connections of biology to data science.

3.2. **Challenges specific to biology education in the era of big data.** Despite similarities in the challenges that both biology and data science education are facing, there are many others that are specific to the life sciences. Some of them are not new, while others have formed more recently with the emergence of data science. Many are rooted in the reality that biology students are less quantitatively prepared than those in other STEM disciplines. Virtually all would require broad multidisciplinary thinking, investments in faculty development, and a major shift in curricular organization and requirements. Below we presents a list of what we would consider major challenges in biology education. We then follow up with our thoughts on potential solutions.

1. ***Educational practice is not keeping up with either research in biology or education research***. The era of big data has brought fundamental changes to the life sciences. However, faculty and student training is happening much slower than the increase in demand, and the inertia of our current academic system that is designed to enforce disciplinary silos, not tear them down, is impeding the process.
2. ***Rote approaches still prevail in teaching introductory statistics***. Communication of statistical ideas is a first step toward success in data science, thus statistics courses should include doing as much statistics as possible, tackling projects using real and relevant data. Coding, data visualization, and ample practical problem-solving should be included. However, statistics education has faced its own problems and setbacks in the last decade, and recommended curricular shifts toward computation [112] have not yet been widely implemented. Many introductory courses in statistics (at the high school and college level) still rely on rote approaches, and students who take such classes rarely develop solid conceptual understanding of the material. There is however a large movement in the statistics education community towards driving concepts through careful use of data examples [113].
3. ***Data visualization training is lacking***. Data visualization and talking intelligently about data are of great importance in society, and especially in biology and medicine. However, visualization of big data requires specialized software and technologies that may not be among the tools covered in traditional biology courses or even in advanced undergraduate biology courses.
4. ***Algorithmic thinking and use of quantitative computational packages are not covered in the biology curriculum***. Computing, basic coding structures, and use of technology are now just as essential for biology as are quantitative thinking, problem solving, and modeling. The importance of data science in biology will continue to grow as datasets grow ever larger. However, biology students, as well as many biology teachers and college faculty, generally consider those out of their comfort zone.
5. ***Emphasis on modeling is insufficient***. Various types of models (mathematical, graphical, statistical, simulation, conceptual, etc.) provide ways for describing and understanding complex biological systems as well as for discovering appropriate controlling mechanisms. Making students aware of the value of models and working with them to help them understand and create/modify models using data is essential for modern biology education. However, mathematical modeling courses in the US generally do not expose students to the connections between models and data, and many institutions do not even offer courses in mathematical modeling (or only offer advanced level mathematics courses that biology students could hardly take because of their mathematics prerequisites).
6. ***Data acumen is necessary for success in biology***. Introductory data science concepts of great importance to biology and medicine include data wrangling with heterogeneous, complex

data sets to address a problem: organizing, combining, aggregating, disaggregating, and transforming data. However, students in general are not exposed to this in traditional biology and introductory statistics courses, and biology students are not required to take data analytics or computer science courses (some of which may cover those topics). There seems to be no place in the curriculum for that.

7. ***Biology and mathematics curricula are rigid and difficult to change***. There is the perception that there is too much content to teach, and the typical linear progression through the curriculum, often involving multiple prerequisites for courses, makes updates difficult. However, students need to be exposed to new concepts and practices brought about by emerging topics related to big data, including mathematical models, ethical uses of models and data, and veracity of open-source data. The large number of credit hours required by some programs as well as caps on the credit hours for majors that some institutions impose, limit the options for including a diversity of quantitative topics in the curriculum.
8. ***Calculus is not sufficient for modern education in quantitative biology***. Data science approaches and uses of data in the life sciences are not generally built on calculus-related methods, but require concepts from other branches of mathematics including, e.g., discrete mathematics, linear algebra, and geometry in addition to statistics. Biology programs that have mathematics requirements for the degree usually require calculus. Some degree programs do not require any mathematics or even, in some cases, any statistics.
9. ***Ability to account for the multiple scales of processes and interactions are critical for biology***. Multiple scales are inherent in many biological questions, which may require approaches different from those in the physical sciences. This requires perhaps different ways of dealing with heterogeneous data in biology (at least at the undergraduate level) than working with data from physical systems. This question is not even on the radar in many cases, as even the general goal of incorporating any authentic data-driven projects into biology courses is challenging.
10. ***New assessment and evaluation methods are necessary.*** With the ongoing growth of data science, there is need for assessment instruments and evaluation methodologies that measure how well colleges and universities are preparing their graduates to join the professional workforce in view of the challenges above. Such metrics for biology programs, particularly regarding quantitative connections, are not yet widely available. The mathematical biology community would benefit from collaboration with the education research community to develop new assessment tools and foster cross-institutional methods to broadly evaluate the effectiveness of alternative methods at the interface of biology and data science.

4. Moving Forward

These challenges may appear insurmountable in the fractured disciplinary setting of academia, but a solution should be possible if the professional math-bio education community embraces and facilitates a much-needed paradigm shift. Faculty, institutions, and the professional community at large should not think of targeting those challenges independently but instead think of designing comprehensive strategies for reform that target them in parallel. We encourage the community to approach teaching biology in a way that truly reflects the nature of the discipline in the 21st century – a vibrant multidisciplinary field in which scientific discovery relies on advancements in technology, mathematics, data science, statistics, and computer science, and a discipline with profound impact on our well-being and our environment. The impact of such discoveries (current and potential) on

society is huge, making it necessary to position our revised programs on solid foundations of ethical practices and social policy.

Below, we give a list of suggestions for consideration and action grouped by organizational levels where, in our view, enacting the change would be done most effectively. We do not pretend to know how all of them could be accomplished or by what specific means. However, by formulating these thoughts, we hope to initiate conversations within the professional community and rely on our collective wisdom to find a path forward. We consider a professional community comprised of societies (e.g., SMB, SIAM, ABLE, NABT, ACUBE), interest groups and organizations (e.g., Bio SIGMAA, IBA), government agencies (NSF, NIH), centers and institutes (e.g., NIMBioS, SESYNC), initiatives (e.g., QUBES, BEER, BioQUEST), academic institutions, departments, and programs, as well as individual scholars and educators. Some of the articles in this special issue are devoted to describing the impact some of these societies, organizations, and initiatives play in shaping a vision for the future of biology education. Finally, even the best ideas may fizzle and be forgotten if there is no financial support for their development and implementation. In that, government support is critical and new initiatives such as NSF's Reintegrating Biology, could have transformative effect for the field of biology education.

At the Course Level:

1. ***Introduce simple basic ideas early***. Finding ways to convey important statistical ideas in a simple way may well be the most powerful and beneficial quantitative learning at the lower level. Emphasis on data literacy, understanding data, and using data to answer questions should be embedded in the core quantitative literacy requirements that occur in the core curriculum of all undergraduate programs. It may also be time to consider separate introductory-level statistics courses for biology students. Addresses Challenge 2.
2. ***Use different courses to gradually develop data acumen***. This could be done initially on a small scale in a variety of courses (e.g., encouraging students to do small data collection projects and develop hypotheses, learning a bit of R or Python along the way) or by organizing tutorials with group projects (requiring, e.g., multivariate statistics). Addresses Challenges 3, 4.
3. ***Use real and relevant data in all courses***. With the open data movement gaining momentum, there are many repositories and databases that contain reliable and current information from research institutions, health departments and organizations, municipal records, government databases, community resources (e.g., Open Data Repositories [114], and those combining open data resources with code and results from contributors [115]). Short, intensive initiatives without multilayered prerequisites could work best. Addresses Challenge 2.
4. ***Expose students to data from complex systems***. Hierarchy in biology leads to highly connected systems. We might build a place in the curriculum where linking, aggregation and aggregation vs. linking can be discussed and demonstrated. Students should be able to see data and models of different types and scales and understand how time/space issues are reflected in data. Developing ways to expose them to some basic systems ideas might encourage them to make a start on the capability to derive links between the underlying system's structure and behavior. Readily available tools for agent-based modeling, such as NetLogo, provide one means to consider how properties at one-scale of biological understanding (cell, or individual organism) scale to some of the properties at tissue or population level. Addresses Challenges 8, 9.
5. ***Introduce students to technology tools used by practicing data scientists***. There is no need to use "training wheels" in classes. If students are taught to use uncommon software,

they will need to unlearn as soon as they decide to join the professional community. This brings more harm than benefit. Students should be exposed to computing and technology in as many traditional biology courses as possible. Addresses Challenges 3, 4.

6. ***Emphasize data visualization across multiple courses***. Address visualization of data in as many introductory statistics and biology courses as possible. Using projects and data relevant to your students' interests, allowing them to choose, visualize, and interpret data sets of their choice to increase their learning engagement. Adapt and adopt can work well, given the multitude of existing visualization tools, inspiring TED talks, and teaching resources offered (e.g., Gapminder, Many Eyes, Tableau, Worldmapper). There are also many great data visualization texts (e.g., Atlas of Knowledge: Anyone can map [116], Data Visualization: A Practical Introduction [117], Fundamentals of Data Visualization [118], and others). Addresses Challenge 3.
7. ***Develop problem-solving acumen***. Help students develop awareness for the specificity of mathematical tools needed to answer various questions in biology. When is calculus really important and when is it likely to be the main tool for dealing with certain types of questions? What type of statistical models would be appropriate to target the underlying question from biology? When are methods from discrete mathematics, linear algebra, geometry, and modern algebra important? Think about how to integrate content and pedagogy to best expose students to those problem-solving concepts. Addresses Challenges 2, 5, 8.

At the Curriculum/Department Level:

8. ***Create flexibility in the biology and mathematics curricula***. This would allow capability for different units/groups of faculty or individual faculty to prioritize opportunities for their students. A way to do this would be to create different pathways for biology majors by broadening existing requirements for the degree. One way to do this would be to break away from the largely linear structure of most biology curricula and replace some of their requirements with electives from an expanded lists of options from mathematics, computer science, data science, and statistics, as well as biology. Similar to the recommendations of the *Data Science for Undergraduates* report, this process may take many forms - see [75, Finding 3.1]. Specific curricular revisions would depend on the needs and resources of each institution, while the extent of changes would depend on the institution's goals and the learning objectives it sets forth for its students. Addresses Challenge 5 and 7.
9. ***Revisit the lists of mathematics courses required for biology students***. The life sciences are different from the other STEM disciplines as, at least for some areas and programs, an entire quantitative curriculum could be rationally developed without a calculus focus at all, but with an emphasis on data science, discrete mathematics, and modern algebra instead. Combining approaches and perspectives from different sub-fields of biology can be used to provide a way for prioritizing topics, methods, and data science approaches. Addresses Challenges 7, 8.
10. ***Emphasize the ethical implications by working with biomedical data***. Inspire students to be attentive and take initiative to develop ethical worldviews that are just as important as their science education. This would be done best early in the curriculum and in as many courses as possible and become a component of any philosophy of science course. Find ways to engage in meta-cognitive reasoning about unintended consequences and subtle second and third order ramifications of the use and misuse of big data. Addresses Challenge 10.
11. ***Encourage students to understand the benefits and hazards of open data and open resources***. The open data, open source journals, and publications under the Creative Commons

license have removed many "paywalls" and catalyzed progress toward democratizing science. However, a complement would teach students how to detect questionable content and verify the veracity of data and integrate these discussions with those regarding ethical use of data. As Rapp and Tirassa point out "Just because it's accessible, doesn't make it ethical." [106]. Addresses Challenge 7.

12. ***Determine metrics of success***. Define clear learning outcomes that your program or department wants to accomplish. There may be different metrics of success for different groups with different objectives/needs. The flexibility in program development should provide multiple approaches; which ones are covered in the curriculum should be determined by the goals each individual course or program sets for its students. Mapping the progression of students throughout the curriculum and analyzing how this aligns with exposure to the major concepts and skills the program decides to emphasize is one example of an evidence-based method for evaluating success. Addresses Challenge 10.

At the Cross-Department/ Institutional Level:

13. ***Develop partnerships between departments***. Establish connections between biology, mathematics, statistics, and computer science departments. Designing opportunities for connecting certain courses from these disciplines – e.g., by offering joint lectures, labs, and projects – would highlight the value and importance of collaborative efforts and generate student opportunities for interdisciplinary projects. Additionally, connections with the philosophy department could foster ways for students from diverse disciplines to be introduced to the debate on ethical concerns with use of big data. Including bioethics in developing interdisciplinary projects would help students identify and confront unethical use of big data in biology and medicine. Addresses Challenge 3, 4, 7.
14. ***Find synergistic approaches to content delivery***. It may be prudent to look for ways to combine key biology and data science concepts and teach them in parallel. For instance, certain data tools and topics overlay with biology topics – some algorithms have biological cognates (e.g., genetic algorithms, neural networks). This way, biology students would learn how to use those algorithms and DS/CS students benefit from learning genetics, trophic nets, and other elements from biology that have influenced data science. Carefully consider what alternative delivery mechanisms may work in your specific institutional context for removing existing disciplinary boundaries. Addresses Challenges 4, 6, 8.
15. ***Ensure institutional support and faculty development***. Put processes in place to offer opportunities for faculty – from those supporting individuals, to groups, departments, schools, cross-institutional initiatives and those across communities. Encourage faculty to participate in professional development initiatives that bring training, instructional materials, and curricular models to those ready to reform their classrooms. This would enhance the comfort of faculty to introduce novel concepts in their teaching. Develop systems of rewards for those willing to offer new courses or transform their classrooms by employing modern pedagogies aiming for educating life-long, independent learners. Addresses Challenges 1, 3, 4, 5, 9.

At the Professional Level:

16. ***Catalyze education research on modeling the education process***. Develop robust inventories for measuring success of data acumen. Decide what metrics are appropriate for evaluation

at different scales of the education enterprise. Is there a need to expand the existing networks of educational conceptualizations to better assess students in this new data-rich driven environment? Addresses Challenge 10.

17. ***Join the debate on teacher education***. We have responsibility for educating prospective teachers differently. Two articles in this volume by Richard Lehrer and Bob Mayes et al. provide some perspective on ways to do this. Addresses Challenges 2, 3, 4.
18. ***Engage with the professional communities of biology, mathematics, and data science***. Many resources have been developed under the guidance of these entities and are freely available to use in their current form or to adapt and adopt. There are also ongoing initiatives, which interested faculty may join, and discussions where faculty could contribute their ideas for the future of biology education. Finally, participating in the various workshops and seminars organized by NSF, MAA, NIMBioS, MBI, SIAM would allow faculty to stay current in the rapidly-changing world of big data and receive professional support and funding. Addresses Challenge 1.

5. Concluding Remarks

The need to create designated curricula for mathematical biology education has been debated for decades, going back at least to the Cullowhee Conference on Training in Biomathematics held at Western Carolina College, Cullowhee, North Carolina, from August 14-18, 1961 [1]. Since then, mathematical biology education has gone through many attempts to reform but some of the challenges have persisted – some issues raised at the Cullowhee conference have resurfaced, in various forms, during each new attempt for change. Questions about the need to create mathematics courses and training programs for biology students different from those designed for engineering and natural sciences students have sparked multiple debates since the 1960s and appear to be just as pertinent (and in many instances just as controversial) today as they were at the time of the Cullowhee conference.

Concurrently, the emerging importance of computer science and discrete mathematics in the 1980s led Fred Roberts and Tony Ralston to raise similar concerns and ask "Is Calculus Necessary?" [119] and "Will Discrete Mathematics Surpass Calculus in Importance?" [120]. Even though discrete mathematics is yet to become a standard expectation of biologists, progress was made by some biology departments in the 1980s and 1990s by requiring biometrics (especially after the success of textbooks like Rohlf and Sokal [121] and Zar [122]). As much as this was a step in the right direction, at many of the institutions requiring statistics for a biology major, the required courses were only "weakly, if at all, connected to the life science courses in the curriculum."[123].

A major shift in undergraduate equation occurred in 1998 with the publication of the Boyer report [124], [125]. It reported on dramatically increased attention to undergraduate education at research universities and accelerated pace of action toward undergraduate research. In biology departments nationwide, this trend amplified the need for quantitative training of biology students and the debate that ultimately culminated in the major reports on biology education from the 2000s – BIO 2010 [2], HHMI/AAMC [5], and AAAS Vision and Change [3].

Overall, there have been four major concurrent drivers of change toward reform:

1. ***Changes in Technology:*** Available computation tools have taken us from slide rules to mainframe computers with punch cards, from teletypes to calculators and microcomputers, and to

cell phones and tablets with enormous computer power and access to massive information on the web. This, as well as the development of algorithms, heuristics, data structures, and databases, has had an enormous impact on educational expectations;

2. ***Changes in Mathematics Education:*** The post-Sputnik push for major changes in mathematics education has led to many more students choosing STEM disciplines, "new math" wars between Bourbaki advocates, traditionalists, and progressive student-centered movements, the emergence of discrete mathematics as an applied discipline (often in association with the technological changes above), and the professionalization of Discipline Based Educational Research (DBER) which vetted reforms that improved mathematical learning [109];
3. ***Changes in Biology Education:*** Two groups in the sixties and seventies have been of particular importance to this reform: The Biological Sciences Curriculum Study Group (BSCS), placing a trifold focus on the emergence of molecular biology and environmentalism (ecology) as well as sustaining a traditional organismal perspective, and the Commission on Undergraduate Education in the Biological Sciences (CUEBS), active from 1963 through 1972. It produced such books as"The Pre-service Preparation of College Biology Teachers" and "Investigative Labs" [126]. This was followed by the high-profile reports and initiatives mentioned above as well as the development of mathematical biology education materials by many groups described in this volume;
4. ***General Changes in US Higher Education:*** Three initiatives have been essential in the process: inclusion of more historically underrepresented groups of students into STEM careers, the "Boyer Report" [127], advocating for college and university faculty to perceive themselves as scholars investigating how their students were learning, and the Presidential Council of Advisors of Science and Technology (PCAST) Report [128], calling for focusing more on retention of students enrolled in STEM disciplines and less on recruitment into programs.

To add to this list, data science has now emerged as another significant driver of change in biology. It has brought enormous benefits to the discipline and has also highlighted grand challenges that are urgently awaiting solutions. It is our hope that data science will now force the discipline to advance education more rapidly toward quantitative approaches, thus training the next generation of biologists in a way that prepares them to successfully confront those challenges. At the same time, we stress again that the type, pace, and magnitude of change each institution, department, or program chooses to implement should always be derived from the goals and objectives set forth for their graduates. To echo the words from 1961 of J. Z. Hearon, founder and head of the NIH Office of Mathematical Research, that still ring true after nearly six decades: "There does not exist now, and it is unlikely that there ever will exist, a unique answer to the question of the kind and the extent of training that a mathematical biologist should receive" [1].

References

[1] M Turner. Statistics in biology, the cullowhee conference on training in biomathematics, cullowhee, north carolina, august, 1961 (edited by HL Lucas, Jr.). *Typing Service, Raleigh, NC*, pages 259–263, 1962.

[2] National Research Council et al. *BIO2010: Transforming undergraduate education for future research biologists*. National Academies Press, 2003.

[3] *Vision and Change in Undergraduate Biology Education: A Call to Action*. AAAS, Washington, DC, 2009.

[4] Lynn Arthur Steen. *Math and Bio 2010: linking undergraduate disciplines*. MAA, 2005.

[5] AAMC-HHMI Scientific Foundation for Future Physicians Committee et al. Scientific foundations for future physicians, 2009.
[6] Steve Olson and Donna Gerardi Riordan. Engage to excel: Producing one million additional college graduates with degrees in science, technology, engineering, and mathematics. report to the president. *Executive Office of the President*, 2012.
[7] Carol S. Schumacher and Martha J. Siegel. 2015 cupm curriculum guide to majors in the mathematical sciences. 2015.
[8] National Science Foundation. Interdisciplinary training for undergraduates in biological and mathematical sciences (UBM), 2005.
[9] Michael Pearson. Maa's professional enhancement program (PREP) funded by nsf. *FOCUS*, 2004.
[10] John R Jungck. Constructivism, computer exploratoriums, and collaborative learning: Constructing scientific knowledge. *Teaching Education*, 3(2):151–170, 1991.
[11] Michael Drew LaMar. Qubes: A community supporting teaching and learning in quantitative biology. 2016.
[12] Olcay Akman and Megan Powell. A model for cross-institutional collaboration: how the intercollegiate biomathematics alliance is pioneering a new paradigm in response to diminishing resources in academia. *Letters in Biomathematics*, 5(1):91–97, 2018.
[13] Carrie Diaz Akman, Olcay Eaton, Dan Hrozencik, Kristin Jenkins, and Katerina V. Thompson. Pathways to national reform of interdisciplinary learning across mathematics and biology. *Bulletin of Mathematical Biology*, 2020.
[14] Meredith L Greer, Olcay Akman, Timothy D Comar, Daniel Hrozencik, and Jonathan E Rubin. Paying our dues: The role of professional societies in the evolution of mathematical biology education. *Bulletin of Mathematical Biology*, 82(5):59–59, 2020.
[15] Charlene D'Avanzo. Post–vision and change: Do we know how to change? *CBE Life Sci Educ.*, 12(3):273–382, 2013.
[16] Ruhul H. Kuddus. Who should change biology education: An analysis of the final report on the vision and change in undergraduate biology education conference. *International Journal of Biology Education*, 3(1a):63–83, 2013.
[17] V. Woodin, Terry, Celeste Carter, and Linnea Fletcher. Vision and change in biology undergraduate education, a call for action—initial responses.
[18] Anna Drake, Laura Struve, Sana Ali Meghani, and Beth Bukoski. Invisible labor, visible change: Non-tenure-track faculty agency in a research university. *The Review of Higher Education*, 42(4):1635–1664, 2019.
[19] Lenora M Hayes. Here to stay: An overview of the non-tenure track faculty and their rise to new faculty majority. In *Diversity, Equity, and Inclusivity in Contemporary Higher Education*, pages 160–174. IGI Global, 2019.
[20] Rafael O Wüest, Niklaus E Zimmermann, Damaris Zurell, Jake M Alexander, Susanne A Fritz, Christian Hof, Holger Kreft, Signe Normand, Juliano Sarmento Cabral, Eniko Szekely, et al. Macroecology in the age of big data–where to go from here? *Journal of Biogeography*, 47(1):1–12, 2020.
[21] Lina Zheng, Guoqin Yuan, Yongming Yang, and Haipeng Kuang. Efficient acquisition of geographic big data: Domestic three-line stereo aerial photography system. In *China's e-Science Blue Book 2018*, pages 205–218. Springer, 2020.
[22] Alexander Y Sun and Bridget R Scanlon. How can big data and machine learning benefit environment and water management: a survey of methods, applications, and future directions.

Environmental Research Letters, 14(7):073001, 2019.

[23] Stephanie E Hampton, Carly A Strasser, Joshua J Tewksbury, Wendy K Gram, Amber E Budden, Archer L Batcheller, Clifford S Duke, and John H Porter. Big data and the future of ecology. *Frontiers in Ecology and the Environment*, 11(3):156–162, 2013.

[24] SL LaDeau, BA Han, EJ Rosi-Marshall, and KC Weathers. The next decade of big data in ecosystem science. *Ecosystems*, 20(2):274–283, 2017.

[25] Scott S Farley, Andria Dawson, Simon J Goring, and John W Williams. Situating ecology as a big-data science: Current advances, challenges, and solutions. *BioScience*, 68(8):563–576, 2018.

[26] Brian Sidlauskas, Ganeshkumar Ganapathy, Einat Hazkani-Covo, Kristin P Jenkins, Hilmar Lapp, Lauren W McCall, Samantha Price, Ryan Scherle, Paula A Spaeth, and David M Kidd. Linking big: the continuing promise of evolutionary synthesis. *Evolution*, 64(4):871–880, 2010.

[27] Martha M Munov and Samantha A Price. The future is bright for evolutionary morphology and biomechanics in the era of big data. *Integrative and Comparative Biology*, 59(3):599–603, 2019.

[28] Greg Gibson. Population genetics and gwas: a primer. *PLoS biology*, 16(3):e2005485, 2018.

[29] Yuriy O Alekseyev, Roghayeh Fazeli, Shi Yang, Raveen Basran, Thomas Maher, Nancy S Miller, and Daniel Remick. A next-generation sequencing primer—how does it work and what can it do? *Academic pathology*, 5:2374289518766521, 2018.

[30] Valeria D'Argenio. The high-throughput analyses era: are we ready for the data struggle? *High-throughput*, 7(1):8–20, 2018.

[31] Javier Andreu-Perez, Carmen CY Poon, Robert D Merrifield, Stephen TC Wong, and Guang-Zhong Yang. Big data for health. *IEEE journal of biomedical and health informatics*, 19(4):1193–1208, 2015.

[32] Akash P Kansagra, J Yu John-Paul, Arindam R Chatterjee, Leon Lenchik, Daniel S Chow, Adam B Prater, Jean Yeh, Ankur M Doshi, C Matthew Hawkins, Marta E Heilbrun, et al. Big data and the future of radiology informatics. *Academic radiology*, 23(1):30–42, 2016.

[33] Viktor Wegmayr, Sai Aitharaju, and Joachim Buhmann. Classification of brain mri with big data and deep 3d convolutional neural networks. In *Medical Imaging 2018: Computer-Aided Diagnosis*, volume 10575, page 105751S. International Society for Optics and Photonics, 2018.

[34] Min Chen, Jun Yang, Jiehan Zhou, Yixue Hao, Jing Zhang, and Chan-Hyun Youn. 5g-smart diabetes: Toward personalized diabetes diagnosis with healthcare big data clouds. *IEEE Communications Magazine*, 56(4):16–23, 2018.

[35] Kun Wang, Yun Shao, Lei Shu, Chunsheng Zhu, and Yan Zhang. Mobile big data fault-tolerant processing for ehealth networks. *IEEE Network*, 30(1):36–42, 2016.

[36] Mohammad-Parsa Hosseini, Hamid Soltanian-Zadeh, Kost Elisevich, and Dario Pompili. Cloud-based deep learning of big eeg data for epileptic seizure prediction. In *2016 IEEE global conference on signal and information processing (GlobalSIP)*, pages 1151–1155. IEEE, 2016.

[37] Meysam Golmohammadi, Amir Hossein Harati Nejad Torbati, Silvia Lopez de Diego, Iyad Obeid, and Joseph Picone. Automatic analysis of eegs using big data and hybrid deep learning architectures. *Frontiers in human neuroscience*, 13:76, 2019.

[38] Eren Balevi and Richard D Gitlin. Synergies between cloud-fag-thing and brain-spinal cord-nerve networks. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–9. IEEE, 2018.

[39] Milad Makkie, Heng Huang, Yu Zhao, Athanasios V Vasilakos, and Tianming Liu. Fast and scalable distributed deep convolutional autoencoder for fmri big data analytics. *Neurocomputing*, 325:20–30, 2019.
[40] Moo K Chung. Statistical challenges of big brain network data. *Statistics & probability letters*, 136:78–82, 2018.
[41] Handan Melike Dönertacs, Matías Fuentealba, Linda Partridge, and Janet M Thornton. Identifying potential ageing-modulating drugs in silico. *Trends in Endocrinology & Metabolism*, 30(2):118–131, 2019.
[42] Matías Fuentealba, Handan Melike Dönertacs, Rhianna Williams, Johnathan Labbadia, Janet M Thornton, and Linda Partridge. Using the drug-protein interactome to identify anti-ageing compounds for humans. *PLoS computational biology*, 15(1):e1006639, 2019.
[43] Suhail Sami Owais and Nada Sael Hussein. Extract five categories cpivw from the 9v's characteristics of the big data. *International Journal of Advanced Computer Science and Applications*, 7(3):254–258, 2016.
[44] I Marketing. key marketing trends for 2017. Technical report, Technical report, IBM. URL: https://www-01. ibm. com/common/ssi/cgi-bin/ssialias, 10.
[45] Josh James. What 'data never sleeps 7.0' says—and doesn't say. *Domosphere: https://www. domo. com/learn/data-never-sleeps-7*.
[46] Science/AAAS. *Special Issue: Artificial Intelligence*, volume 349. American Association for the Advancement of Science, 2015.
[47] Eberhard O Voit. Perspective: Dimensions of the scientific method. *PLoS Computational Biology*, 15(9), 2019.
[48] Chris Anderson. The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, 16(7):16–07, 2008.
[49] Sarah Webb. Deep learning for biology. *Nature*, 554(7693), 2018.
[50] Davide Cirillo and Alfonso Valencia. Big data analytics for personalized medicine. *Current opinion in biotechnology*, 58:161–167, 2019.
[51] Wen Y Ding, Michael W Beresford, Moin A Saleem, and Athimalaipet V Ramanan. Big data and stratified medicine: what does it mean for children? *Archives of disease in childhood*, 104(4):389–394, 2019.
[52] Genevieve L Stein-O'Brien, Raman Arora, Aedin C Culhane, Alexander V Favorov, Lana X Garmire, Casey S Greene, Loyal A Goff, Yifeng Li, Aloune Ngom, Michael F Ochs, et al. Enter the matrix: factorization uncovers knowledge from omics. *Trends in Genetics*, 34(10):790–805, 2018.
[53] Hans-Jörg Rheinberger. Infra-experimentality: From traces to data, from data to patterning facts. *History of Science*, 49(3):337–348, 2011.
[54] Miquel Duran-Frigola, Adrià Fernández-Torras, Martino Bertoni, and Patrick Aloy. Formatting biological big data for modern machine learning in drug discovery. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 9(6):e1408, 2019.
[55] Altaf Merchant. Big data: Ushering new vistas in market research. *Projectics/Proyectica/Projectique*, (3):9–12, 2018.
[56] Claudia Manzoni, Paul Denny, Ruth C Lovering, and Patrick A Lewis. Computational analysis of the lrrk2 interactome. *PeerJ*, 3:e778, 2015.
[57] National Institutes of Environmental Health Sciences. Developing a data science competent ehs workforce, 2018.

[58] Teresa K Attwood, Douglas B Kell, Philip McDermott, James Marsh, Steve R Pettifer, and David Thorne. Calling international rescue: knowledge lost in literature and data landslide! *Biochemical Journal*, 424(3):317–333, 2009.

[59] MD Wilkinson, M Dumontier, IJ Aalbersberg, G Appleton, M Axton, A Baak, N Blomberg, JW Boiten, LBD Santos, PE Bourne, et al. Comment: the fair guiding principles for scientific data management and stewardship. scientific data 3: 9, 2016.

[60] Susanna-Assunta Sansone, Peter McQuilton, Philippe Rocca-Serra, Alejandra Gonzalez-Beltran, Massimiliano Izzo, Allyson L Lister, and Milo Thurston. Fairsharing as a community approach to standards, repositories and policies. *Nature biotechnology*, 37(4):358–367, 2019.

[61] Avi Wigderson. Mathematics and computation: Algorithms will rule the earth, but which algorithms? *The institute Letter*, page 4, Fall 2018.

[62] Joel E Cohen. Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better. *PLoS biology*, 2(12), 2004.

[63] Michael C Mackey and Philip K Maini. What has mathematics done for biology? *Bulletin of mathematical biology*, 77(5):735–738, 2015.

[64] Peter Grindrod. *Patterns and waves: The theory and applications of reaction-diffusion equations*. Oxford University Press, USA, 1991.

[65] Guillermo E Herrera and Suzanne Lenhart. Spatial optimal control of renewable resource stocks. *Spatial Ecology*, page 343, 2010.

[66] Karl Sigmund and Martin A Nowak. Evolutionary game theory. *Current Biology*, 9(14):R503–R505, 1999.

[67] Christopher Devers, Christine Lee, Joe Hoffert, Erin Devers, Stephen Burgos, and Jeremy Davis. Followme: A game-based approach to self-regulation. In *Society for Information Technology & Teacher Education International Conference*, pages 754–758. Association for the Advancement of Computing in Education (AACE), 2015.

[68] Bernd Sturmfels. Can biology lead to new theorems? *Annu. Rep. Clay Math. Inst.*, pages 13–26, 2005.

[69] Matthew Macauley and Nora Youngs. The case for algebraic biology: from research to education. *Bull. Math. Biol.*, 2020.

[70] Ivan Nunes Da Silva, Danilo Hernane Spatti, Rogerio Andrade Flauzino, Luisa Helena Bartocci Liboni, and Silas Franco dos Reis Alves. Artificial neural networks. *Cham: Springer International Publishing*, page 39, 2017.

[71] Clinton Sheppard. *Genetic algorithms with python*. Smashwords Edition, 2017.

[72] John W Tukey. The future of data analysis. *The annals of mathematical statistics*, 33(1):1–67, 1962.

[73] Iain Carmichael and JS Marron. Data science vs. statistics: two cultures? *Japanese Journal of Statistics and Data Science*, 1(1):117–138, 2018.

[74] Joe Davison. No, machine learning is not just glorified statistics. *Medium - TowardData-Science.com*, 2018. Accessed: 2020-03-31.

[75] Engineering National Academies of Sciences, Medicine, et al. *Data science for undergraduates: Opportunities and options*. National Academies Press, 2018.

[76] Bérénice Batut, Saskia Hiltemann, Andrea Bagnacani, Dannon Baker, Vivek Bhardwaj, Clemens Blank, Anthony Bretaudeau, Loraine Brillet-Guéguen, Martin Čech, John Chilton, et al. Community-driven data analysis training for biology. *Cell systems*, 6(6):752–758, 2018.

[77] National Research Council et al. *Discipline-based education research: Understanding and improving learning in undergraduate science and engineering*. National Academies Press,

2012.
[78] Scott Freeman, Sarah L Eddy, Miles McDonough, Michelle K Smith, Nnadozie Okoroafor, Hannah Jordt, and Mary Pat Wenderoth. Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23):8410–8415, 2014.
[79] Marilyne Stains, Jordan Harshman, Megan K Barker, Stephanie V Chasteen, Renee Cole, Sue Ellen DeChenne-Peters, MK Eagan, Joan M Esson, Jennifer K Knight, Frank A Laski, et al. Anatomy of stem teaching in north american universities. *Science*, 359(6383):1468–1470, 2018.
[80] Lisa Marie Blaschke. Heutagogy and lifelong learning: A review of heutagogical practice and self-determined learning. *The International Review of Research in Open and Distributed Learning*, 13(1):56–71, 2012.
[81] Johann Friedrich Herbart. *The science of education*. DC Heath & Company, 1896.
[82] Florian Cajori. The teaching and history of mathematics in the united states, washington, 1890. *This book is a gold mine of all sorts of information*, page 94, 1890.
[83] MS Knowles. The modern practice of adult education: Andragogy vs. pedagogy wilton, 1980.
[84] Joseph Jackson Schwab. *The Teaching of Science. The Teaching of Science as Enquiry.[By] JJ Schwab. Elements in a Strategy for Teaching Science in the Elementary School.[By] Paul F. Brandwein*. Harvard University Press, 1962.
[85] Jerome S Bruner. " the process of education" revisited. *The Phi Delta Kappan*, 53(1):18–21, 1971.
[86] S Hase and Chr Kenyon. From andragogy to heutagogy. UltiBASE In-Site, 2000.
[87] Stewart Hase. Heutagogy and e-learning in the workplace: Some challenges and opportunities. *Impact: journal of applied research in workplace e-learning*, 1(1):43–52, 2009.
[88] Discover Date Science. Bachelor degree in data science – guide to choosing a great program. https://www.discoverdatascience.org/programs/bachelors-in-data-science/. Accessed April 2, 2020.
[89] Brian Godsey. *Think Like a Data Scientist: Tackle the data science process step-by-step*. Manning Publications Co., 2017.
[90] Jeffrey S Saltz and Jeffrey M Stanton. *An introduction to data science*. Sage Publications, 2017.
[91] Oliver Theobald. *Machine learning for absolute beginners*. 2017.
[92] Mark Fenner. *Machine learning with Python for everyone*. Addison-Wesley Professional, 2019.
[93] Jon Krohn, Grant Beyleveld, and Aglaé Bassens. *Deep Learning Illustrated: A Visual, Interactive Guide to Artificial Intelligence*. Addison-Wesley Professional, 2019.
[94] Joel Grus. *Data science from scratch: first principles with python*. O'Reilly Media, 2019.
[95] Max Kuhn and Kjell Johnson. *Applied predictive modeling*, volume 26. Springer, 2013.
[96] Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 2019.
[97] Sebastian Raschka. *Python machine learning*. Packt Publishing Ltd, 2015.
[98] Bharath Ramsundar, Peter Eastman, Patrick Walters, and Vijay Pande. *Deep learning for the life sciences: applying deep learning to genomics, microscopy, drug discovery, and more*. " O'Reilly Media, Inc.", 2019.
[99] Daniel Zelterman. *Applied multivariate statistics with R*. Springer, 2015.

[100] Alan Moses. *Statistical Modeling and Machine Learning for Molecular Biology*. CRC Press, 2017.
[101] Phillip Compeau and PA Pevzner. *Bioinformatics Algorithms: An Active Learning Approach. La Jolla*. CA: Active Learning Publishers, 2018.
[102] Dan MacLean. R bioinformatics cookbook: use r and bioconductor to perform rnaseq, genomics, data visualization, and bioinformatic analysis. 2019.
[103] Arthur Lesk. *Introduction to bioinformatics*. Oxford university press, 2019.
[104] Cathy O'neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2016.
[105] Effy Vayena, Marcel Salathé, Lawrence C Madoff, and John S Brownstein. Ethical challenges of big data in public health. *PLoS computational biology*, 11(2), 2015.
[106] Amon Rapp and Maurizio Tirassa. Know thyself: a theory of the self for personal informatics. *Human–Computer Interaction*, 32(5-6):335–380, 2017.
[107] Neil M Richards and Jonathan H King. Big data ethics. *Wake Forest L. Rev.*, 49:393, 2014.
[108] Jo Handelsman, Diane Ebert-May, Robert Beichner, Peter Bruns, Amy Chang, Robert DeHaan, Jim Gentile, Sarah Lauffer, James Stewart, Shirley M. Tilghman, and William B. Wood. Scientific teaching. *Science*, 304(5670):521–522, 2004.
[109] Susan Singer and Karl A Smith. Discipline-based education research: Understanding and improving learning in undergraduate science and engineering. *Journal of Engineering Education*, 102(4):468–471, 2013.
[110] Liz Stanhope, Laura Ziegler, Tabassum Haque, Laura Le, Marcelo Vinces, Gregory K. Davis, Andrew Zieffler, Peter Brodfuehrer, Marion Preest, Jason M. Belitsky, Charles Umbanhowar, and Paul J. Overvoorde. Development of a biological science quantitative reasoning exam (biosquare). 16(4), 2018.
[111] Robin T. Taylor, Pamela R. Bishop, Suzanne Lenhart, Louis J. Gross, Sturner, and Kelly. Development of the biocalculus assessment (bca). 19(1), 2020.
[112] Deborah Nolan and Duncan Temple Lang. Computing in the statistics curricula. *The American Statistician*, 64(2):97–107, 2010.
[113] American Statistical Association et al. Guidelines for assessment and instruction in statistics education (gaise): College report 2016. alexandria, va: Author, 2016.
[114] Scientific Data. Recommended data repositories. *Scientific Data. https://www. nature. com/sdata/policies/repositories*. (accessed April 24, 2020).
[115] Kaggle: Code and data. https://www.kaggle.com/.
[116] Brian Reffin Smith. Atlas of knowledge: Anyone can map, 2016.
[117] Kieran Healy. *Data visualization: a practical introduction*. Princeton University Press, 2018.
[118] Claus O Wilke. *Fundamentals of data visualization: a primer on making informative and compelling figures*. O'Reilly Media, 2019.
[119] Fred S Roberts. Is calculus necessary? In *Proceedings of the Fourth International Congress on Mathematical Education*, pages 50–53, 1980.
[120] Anthony Ralston. Will discrete mathematics surpass calculus in importance? *The Two-Year College Mathematics Journal*, 15(5):371–373, 1984.
[121] F James Rohlf and Robert R Sokal. *Biometry: the principles and practice of statistics in biological research*. Freeman New York, 1961, 1981, 1995, and many subsequent editions.
[122] Jerrold H Zar. *Biostatistical analysis*. Pearson Education India, 1999.
[123] Louis J Gross. Quantitative training for life-science students. *BioScience*, 44(2):59, 1994.

[124] Ernest L Boyer. The boyer commission on educating undergraduates in the research university, reinventing undergraduate education: A blueprint for america's research universities. *Stony Brook, NY*, 46, 1998.
[125] Wendy Katkin. The boyer commission report and its impact on undergraduate research. *New Directions for Teaching and Learning*, 93:19–38, 2003.
[126] Commission on Undergraduate Education in the Biological Sciences and Edward J Kormondy. *CUEBS, 1963 to 1972: its history and final report*. 1972.
[127] John W Moore. The boyer report. *Journal of Chemical Education*, 75(8):935, 1998.
[128] President's Council of Advisors on Science and Technology (US). *Prepare and Inspire: K-12 Education in Science, Technology, Engineering, and Math (STEM) for America's Future: Executive Report*. Executive Office of the President, 2010.

Department of Mathematics, Randolph-Macon College, Ashland, VA 23005

Email address: RainaRobeva@rmc.edu

Center for Bioinformatics and Computational Biology; DENIN Delaware Environmental Institute, University of Delaware, Newark, DE 19716

Email address: jungck@udel.edu

Department of Ecology and Evolutionary Biology; Department of Mathematics; NIMBioS; University of Tennessee Knoxville, Knoxville, TN 37996

Email address: lgross@utk.edu