HHS Public Access

Author manuscript

Per Med. Author manuscript; available in PMC 2020 October 09.

Published in final edited form as:

Per Med. 2019 May 01; 16(3): 247–257. doi:10.2217/pme-2018-0145.

Preparing next-generation scientists for biomedical big data: Artificial intelligence approaches

Jason H. Moore, **Mary Regina Boland**, **Pablo G. Camara**, **Hannah Chervitz**, **Graciela Gonzalez**, **Blanca E. Himes**, **Dokyoon Kim**, **Danielle L. Mowery**, **Marylyn D. Ritchie**, **Li Shen**, **Ryan J. Urbanowicz**, **John H. Holmes**

Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104

Abstract

Personalized medicine is being realized by our ability to measure biological and environmental information about patients. Much of these data is being stored in electronic health records yielding big data that presents challenges for its management and analysis. We review here several areas of knowledge that are necessary for next-generation scientists to fully realize the potential of biomedical big data. We begin with an overview of big data and its storage and management. We then review statistics and data science as foundational topics followed by a core curriculum of artificial intelligence, machine learning, and natural language processing that are needed to develop predictive models for clinical decision making. We end with some specific training recommendations for preparing next-generation scientists for biomedical big data.

Personalized and Precision Medicine

Personalized medicine and precision medicine are used interchangeably to describe the process of tailoring medical treatment to the individual characteristics of a given patient and the creation of new targeted pharmaceuticals [1–3]. In order to tailor medical treatments to individual patient characteristics, researchers have focused on designing ‘meaningful patient subgroups’. These ‘meaningful’ subgroups would all be similar with respect to a particular patient characteristic. For example, this could be a particular ethnicity, gender or sexual identity, socioeconomic group (low or high income), or even disease status (e.g., asthma). Patients can also be similar to each other with respect to their allergy status. For example, all patients with egg allergies is a meaningful patient subgroup that requires tailored treatment regimens that are important for clinicians to be aware of – for instance they must receive their flu vaccination in a split dose [4] rather than a single dose.

Some patient subgroups are already known, such as the previously described food allergies. However, other patient subgroups are not known. Informatics methods have been designed to identify subgroups of patients based on physiological signals in the electronic health

Corresponding author details: Jason H. Moore.

Author contributions: All authors participated in the design and writing of the content

Ethical disclosure: NA

record (EHR) [5], temporal changes in laboratory value states (e.g., controlled vs uncontrolled diabetes) [6], adverse drug reactions due to genetic factors such as CYP mutations [7], and cancer types [8]. Further informatics methods are required to identify and stratify patient populations into meaningful subpopulations to enable precision medicine. This remains a challenging area, as each individual patient contains a complex constellation of disease phenotypes and symptomology [9].

Biomedical Big Data

The best definition of 'big data' is data that is of a size that challenges your current computational workflow thus limiting your ability to perform analysis and/or interpret results. Computational challenges could come in the form of data storage capacity, network bandwidth for moving data from one storage device to another, or the compute cycles necessary to process and analyze the data. In addition, it is often the case that the computational results generated might exceed the volume of the data itself. More formally, big data is often described according to the *four* Vs that were originally defined by IBM, Inc. in a widely available infographic. The first V is the *volume* of the data. This, of course, refers to storage capacity needed to manage the data and is the first challenge anyone thinks about when discussing big data. The second V is *velocity*. The data might be arriving from some measurement system such as a wearable device faster than your computational infrastructure can cope with it. The third V is *variety*. Big data is not always uniform and is often a heterogeneous mix of data types from different measurement sources. Electronic health record data is a perfect example of this. The fourth V is *veracity*. How clean is the data? Can you trust the data? Big data is often messy, biased, and plagued with missing values. This creates enormous challenges for trying to get a data set ready for analysis and could impact reproducibility. These are the most common Vs that are discussed. However, there are at least two others could be included in this list. The first is *vexedness* or complexity. For example, data that is hierarchical, high dimensional, and/or longitudinal adds complex dimensions to data that can be especially problematic when combined with all the other Vs. This is especially true in the era of new technologies such as smart watches and other mobile devices that can monitor measures of health in real time. Not only does data like this have high volume and velocity but it can be quite complex with discontinuous measures over time. This kind of data is being integrated into electronic health records along with data from social media, for example. Social media data from Facebook, Twitter, and other sources bring natural language into the picture that adds its own layer of complexity. Finally, some might be concerned with the *value* of the data. Is the data being collected worth all the trouble to put in place massive data storage and high-performance computing? Big data is not always the answer and some have argued that more focused small data approaches might make more sense for some scientific questions [10].

The EHR is a modern example of an important source of big data and has been implemented widely across academic medical centers worldwide as their value for tracking patient data and information along with improved billing has been realized. The adoption of EHRs has been enabled by computing technology such as inexpensive data storage and databases that can handle big data efficiently. These EHR databases integrate an impressive range of different patient data including demographics, laboratory tests, imaging, medication history

and use, and clinical notes that include comments from both the clinician and patient in free text. Additional data from genomics and wearable or smart devices are increasingly being captured and integrated into the EHR. These new big data sources create challenges around data management and their use for clinical decision making. An important question is how to process these big data streams and turn them into actionable information that a busy clinician can use. An additional issue that needs to be addressed is that patients are increasingly generating their own big data through direct-to-consumer marketing. It is conceivable in the near future that each patient will have many terabytes or even petabytes of data and information that will need to be stored and processed as part of patient care. These data serve as the raw materials for both personalized and precision medicine that will be enabled by the concepts and methods outlined below.

Data Management and Integration

If data are the bricks of personalized medicine research and practice, then databases are the scaffold that ensures the integrity of the science of personalized medicine. Databases provide the structure within which data are maintained and made available for future use. Database systems provide the computational mechanisms one needs to store, manipulate, and retrieve data, typically through graphical interfaces and query languages. The predominant architecture of biomedical databases is relational, in that data are stored in tables that represent a specific domain, such as demographics. More recently, there has been growing interest in graph databases, where rather than tables, data are represented as nodes in an undirected graph, and relationships between the nodes are represented as links [11,12]. More specifically, nodes contain *properties*, or attribute-value pairs, and each node is labeled to indicate its identity. The links also contain properties, which express the semantics between two or more nodes. For example, links indicate the direction of a relationship between nodes, and this directionality, in concert with such properties as possession (such as an “Is-A” relationship) or the strength of a relationship. In either database architecture, one can manipulate the data through queries using a language specific to the architecture. Regardless of the database architecture, all databases are designed and implemented using a *data model*, which defines the characteristics of the database and provides a map for database personnel.

There is an embarrassment of riches when it comes to biomedical data and the databases that store them. One now has the capability of integrating or *linking*, seemingly disparate databases to provide a more complete landscape of a clinical problem. For example, in assessing the reason why a patient’s HbA1C is out of control, data obtained by persistent monitoring of physical activity by means of a personal fitness device such as a Fitbit, could be added to the patient’s clinical data. This could help the clinician to personalize a patient’s physical activity to enhance his or her total diabetes control program. With the burgeoning of new types of data resources, there has been increasing interest in developing new methods of record linkage while recognizing the challenges associated with linkage [13,14].

While there is an increasing diversity of data resources, such as one finds in EHRs, environmental monitoring, and administrative claims data, there is a corresponding heterogeneity of data representation, even within the same biomedical domain. For example, units of measurement may differ from one database to another, where one uses metric

weight and another avoirdupois. The diversity of such data representation constitutes a major challenge to effective and accurate data linkage and integration. In order to address this challenge, one must turn to the epistemological dimensions of biomedical data, as represented in syntax and semantics, both of which are needed to achieve data harmonization, which is in turn essential for data integration. Increasingly, informatics professionals are turning to *ontologies* to effect the data harmonization process, in order to map concepts and the relationships between them in a graphical format. As such, ontologies can be considered a type of data model. An example of a graphical approach to the integration of multi-omics data with existing biological knowledge found in a study by Kim et al. [15]. In this study, the authors proposed and evaluated an "intermediate integration" method that incorporated genomic, epigenomic, and transcriptomic data that included pathway, motif and gene ontology knowledge sources in order to predict an ovarian cancer phenome. This phenome was characterized on three dimensions: survival, tumor stage, and grade.

Another pressing issue in data management and integration is assuring the quality of data. A major reason for poor data quality is that it is frequently missing. Laboratory reports might not find their way into the medical record, or a component of a physical examination might not be completed. In order to make effective inferences from data, there needs to be a regime for handling missing data, which might include various methods of imputation such as hot deck or multiple imputation as well as deep learning [16]. Kim et al. applied a novel integrative framework for predicting censored survival time, itself a form of missing data [17].

Large-scale clinical data warehouses provide a possible avenue to creating and maintaining data resources that serve a wide variety of users and stakeholders in the personalized medicine domain. During the extract-transfer-load process that is the hallmark of the data warehouse paradigm, procedures can be implemented to ensure data quality, to integrate data from a variety of sources, and to provide a secure platform for analysis of de-identified data. These warehouses could provide researchers with the capability of intelligent cohort identification and extraction, and for precision medicine practitioners with the means to address a given patient's phenotype at the point of care.

Statistical Analysis

Statistical analysis is the area of mathematics that uses models for the data to summarize and draw conclusions. The boundary between statistical analysis and machine learning (the field of artificial intelligence that enables computer systems to learn from data) is therefore blurry and a common subject of debate. Each of these two disciplines emphasizes a different aspect of the task of extracting conclusions from the data: machine learning focuses on making accurate predictions from the data, while statistical analysis can be utilized to assess the validity of a model for the data and for making inferences. Thus, statistical analysis and machine learning are complementary in many aspects. Many of us strongly believe that, as the communities of applied mathematics and computer science continue to interact, the two disciplines will be merged into a single field, sometimes called statistical learning [18]. It is our responsibility to train and prepare the next generation of scientists to lead this

transformation, presenting them with the tools of statistical analysis in ways that foster this transformation.

The importance of statistical analysis for biomedical applications is notable. By putting the focus of study in the model for the data, and making the modeling assumptions explicit, statistical analysis allows for interpretable and justifiable statements. This becomes critical in the clinical setting, where decisions that affect the health of patients need to be justified in precise, rational, and arguable ways to fulfill the common ethical and legal requirements. Probability and statistical inference, the two components of statistical analysis, precisely allow biomedical researchers to quantify statements about the data in terms of likelihoods or frequencies of occurrence relative to a model. Using the tools of statistical inference, researchers can test multiple models or hypotheses and determine those that better explain the data. By framing the inference process in terms of probability distributions, they can quantify the uncertainty in their conclusions as a consequence of stochastic factors, such as measurement errors, and missing data. In the era of big data, these aspects become of paramount importance, as the difficulties inherent in testing large numbers of hypotheses or optimizing complex models can lead to misinterpretation of the data. Statistical analysis thus complements machine learning and artificial intelligence approaches to biomedical data analysis in essential ways, and it is critical that the next generation of scientists is equipped with a solid and modern background in statistical analysis to be able to produce meaningful and defensible predictions from large biomedical datasets.

Data Science

Data science refers broadly to integrating statistical and computational techniques with domain knowledge to gain insights from big data. As a *data-driven* discipline, data science is able to address pre-specified questions, as well as discover novel hypotheses in an unbiased fashion. In the case of biomedical data, data science can be applied to gain novel insights and uncover biologically actionable knowledge for transforming the way we diagnose, treat, and prevent disease.

Statistics, as described in a previous section, is an essential foundation of data science. Knowledge of statistical theory alone, however, is not sufficient for the analysis of large and real-world datasets. Computing and informatics skills are crucial to data science, as the size of datasets has increased, requiring the use of computers to store, query, and analyze datasets effectively. The computing skills necessary for effective data science are applied, however, rather than theoretical in nature. A background in computer science can be helpful to a data scientist, but only if an individual is able to apply programming skills to wrangle with and analyze data. The most commonly used programs by data scientists for analysis of data are Python and R. For large datasets, use of various other programs written in languages such as C is necessary, while storing of large datasets requires the use of databases such as SQL.

Domain expertise is indispensable in data science to ensure questions posed of data are reasonable and to guide the interpretation of results. Data scientists do not passively analyze data. Rather, critical choices are made in the selection and/or transformation of variables, appropriateness of methods to answer specific questions, and subsequently, how to best

communicate and interpret findings using effective visualization techniques. Indeed, many data scientists collect their own data to answer questions of their choice, driving scientific areas, rather than serving as analysts in service of others' questions.

The fusion of expertise in statistics, computing, informatics, and domain knowledge yields practical skills necessary for data science, including the ability to retrieve and clean data, perform exploratory analyses, build models to answer scientific questions, and present informative and visually appealing results. The process of data analysis is not linear; cycles are often required before "final" results are obtained. That is, after performing exploratory analyses or building initial models, features of the data and/or models used may need to be adjusted in accordance with the topic matter at hand. In various cases, data scientists develop their own methods and tools, inspired by needs encountered during analyses of real datasets.

To ensure the results of the research enterprise yield maximum benefit, many data scientists have led reproducible research efforts. This includes the creation of open-source software packages that are widely distributed, sharing specific steps followed to obtain results in publications, and the sharing of data necessary to reproduce findings [19–23]. A variety of technologies are available to promote the transparency and reproducibility of methods applied to big datasets. RStudio is an integrated development environment that greatly enhanced R's usability and popularity among data scientists, for reasons that include improving workflow and facilitating the creation of R Markdown documents that can be converted into a variety of formats (e.g., HTML, PDF) for reporting of results [22]. More broadly, laboratory notebooks such as Jupyter and Apache Zepplin provide interactive, web-based computing environments and support the use of open-source software and computer programming languages, including Python, R, Scala, Groovy, and SQL. Data scientists can leverage these notebooks for data wrangling, analytics, visualization, and collaboration. For example, a data scientist using Python can clean and pre-process data with pandas, analyze data with scikit-learn, and visualize data with Altair. These notebooks also support big data processing and computing technologies such as Hadoop, Spark, and Hive. Version control systems such as Git provide effective means of tracking large projects over time. GitHub, a web-based platform that hosts projects using Git for version control, has become a widely-used repository where code, small datasets, and results of analyses are shared. More recently, containers such as Docker and Singularity provide a user-friendly means of sharing code with pre-installed software dependencies and user-restricted processes that can also help with reproducibility [24]. Container orchestration software (e.g., Kubernetes, OpenShift) supports scalable application build, management, and deployment at an enterprise level. The rise of accessible cloud computing has enabled many of these tools and approaches to be leveraged for big data [25]. Academia and industry alike are leveraging these powerful technologies to support data science efforts that seek to improve health.

Artificial Intelligence

Fundamentals

The term 'artificial intelligence' (AI) is one that has evolved to have a meaning that is more general, interdisciplinary, and encompassing, than when it was first coined. As a subfield of computer science, AI is often used interchangeably with the term 'machine learning', which

itself is more accurately a subfield of AI dealing with the broader concept of *inductive reasoning*. However, a wealth of key prerequisite topics that focus on *deductive reasoning* align with the bulk of biomedical informatics applications being actively utilized today.

These founding principles of AI and their intersection with biomedical informatics applications [26] are essential for those hoping to fully exploit big data for personalized medicine and other applications in healthcare [27]. These principles also serve to inform a deeper understanding of the popular topic of machine learning and the future of AI research. In summary, AI fundamentals focus on how biomedical data can be organized, represented, interpreted, searched, and applied in order to derive knowledge, make decisions, and ultimately how to make predictions.

According to a popular AI textbook, training should begin with a historical overview of the development of artificial intelligence as a field, surveying definitions, key advancements, applications, and ethical considerations [28]. This should be followed by topics in logic (i.e. propositional and first order logic) describing the common formal language for data and knowledge allowing for an interface between person and machine. Next are frameworks for data representation including frames, rules, trees, ontologies, and semantic networks. Representation is an essential topic connecting both deductive and inductive reasoning.

It is also important to understand the role of an *agent* as the traditional building blocks of a bottom-up AI system such as a deep learning neural network where agents are nodes. This is in contrast to top-down AI that attempts to build an artificial brain. Top-down AI is premature given we do not fully understand how the human brain works. Another essential topic includes an introduction to the basics of problem solving through *search* algorithms including uniformed search (e.g. breadth or depth-first) and heuristic search (e.g. greedy or A* search). Search is relevant to common challenges in biomedical information access, and is essential to optimization and constraint satisfaction, i.e. problems where constrains on certain variables need to be satisfied in order to result in a solution. Training in AI fundamentals also extends to an understanding of reasoning with uncertainty and how it ties to probabilistic biomedical knowledge. This leads to the topics of conditional probabilities, entropy, Bayesian inference, and knowledge engineering, as well as their integration for knowledge based systems including rule-based inference, expert systems, and modern clinical decision support systems [26]. In other words, how do we take existing knowledge and apply it to making decisions through reasoning? Additional topics to explore here could include state machines, dynamic models, reinforcement learning, adversarial search (e.g. game play), artificial life, and automated discovery. These somewhat advanced AI topics have the potential to intersect with the field of biomedical informatics more frequently in the future.

It is our position that AI is an essential component of any training program where analysis of big biomedical data and/or complex systems are involved. This is particularly true for the goal of personalized medicine. We see evidence of this realization at the governmental level. For example, China has made substantial investments in AI with a stated goal of being a dominant force in this space. Also, Germany has specifically mentioned AI and personalized medicine in its Industry 4.0 initiative that comes with specific funding allocations. Other

countries are likely to follow suit in the coming years as it becomes clear that AI is a necessary part of a comprehensive analytics approach to hard problems in healthcare and other areas of societal importance such as finance, manufacturing, and weather forecasting.

It is of course important to establish realistic expectations for AI in these efforts as has been pointed out previously [29,30]. Artificial intelligence does have important limitations. For example, the use of AI in medicine requires knowledge engineering to encode and make available to the computer what we as humans know. Extracting knowledge from humans is a very difficult and time-consuming process. Further, the black box nature of AI is a concern for those hoping to use the models they generate for developing new drugs or treating patients. We need to be able to understand the model to develop basic science or clinical experiments to validate the finding. Trust is a related concern for medical applications. Clinicians need to be able to trust that AI is generating a result that is both useful and grounded in medical evidence. Finally, there are legal and ethical issues related to treating patients with AI-generated results. Indeed, the threat of lawsuits prevented the MYCIN AI software from being used in the clinic to treat intensive care unit patients with computer-prescribed antibiotics [31]. Students learning about AI should be aware of its limits and even its potential dangers [32,33].

Machine Learning

As mentioned above, machine learning is a subfield of AI dealing with the broader concept of inductive reasoning. In particular, we think of it as a set of methods that can extract patterns from raw data and use these patterns to predict future data or help other types of decision making [26,34,35]. Supervised learning and unsupervised learning are two major categories. In supervised or predictive learning, we learn a function that maps an input object, represented by a set of features, to an output value [28]. This learning task is called *classification* if the output value is *categorical* and called *regression* if the output value is *continuous*. In unsupervised or descriptive learning, we are given just the input data and aim to identify interesting patterns in the data, such as clusters, anomalies, and latent factors [28]. Cluster analysis aims to group similar objects into clusters. Anomaly detection aims to identify outliers in the data. Learning latent factors can help extract compact data representations or informative features. Given that many biomedical problems can be formulated as these tasks, machine learning offers powerful tools for solving data science problems in biomedicine. Existing successful applications include disease diagnosis, biomarker discovery, omics study, drug discovery, clinical outcome prediction and patient monitoring, personalized treatment, smart electronic health records, epidemic outbreak prediction, inferring health status through wearable devices, and image-based decision support in radiology, dermatology, ophthalmology and pathology [27,36]. Python, Java, R, C ++, C, JavaScript, Scala and Julia are among the widely used machine learning languages.

There is a broad range of machine learning methods and algorithms [26,34,35]. For example, one classic supervised learning method is to learn a decision tree [26,34,35], where each internal node describes a test on a feature, each branch corresponds to an outcome, and each leaf node represents an output value. Despite being easy to interpret, decision trees are high variance estimators: slightly different input data can yield very different tree structures.

To overcome this instability, random forest has been proposed by aggregating many decision trees trained on random subsets of the data using random subsets of features. This has been shown much more robust than decision trees. Other examples of classic learning methods [26,34,35] include support vector machines, linear regression, logistic regression, naive Bayes, linear discriminant analysis, and k-nearest neighbor. In these classic methods, features representing an object are user-specified and may not be optimized for the learning task. A new paradigm is to use machine learning to achieve two goals at the same time: (1) to learn the mapping from an object representation to an output; and (2) to discover the object representation itself by automatically identifying the features suitable for the learning task. For example, the deep neural network learning methods [34] belong to this category and have been shown highly successful in many machine learning application domains including biomedical data science [37–41]. It is important to note that no one machine learning method is ideal for all data and choosing the right methods to use can be problematic [42,43].

Given the unprecedented scale and complexity of biomedical big data, machine learning is still facing major computational and methodological challenges. These include (1) the overfitting issue when we fit learning models with many variables to estimate, (2) the model selection issue when we have a number of models with different complexities to choose from, (3) the optimal search strategy when we don’t have a closed-form solution, (4) the hyperparameter optimization issue when we have many parameters to tune, and (5) the biomedical interpretation issue when promising results are predicted by complicated models. To address these challenges and make machine learning more user-friendly to non-expert practitioners, efforts have been made in the field of automated machine learning (AutoML) to automate the process of applying machine learning to real-world problems. The existing AutoML systems (e.g., AutoWeka [44], AutoSklearn [45], TPOT [46] and PennAI [47]) are designed to automate one or more machine learning components such as data preparation, task detection, feature engineering, model selection, hyperparameter optimization, pipeline selection and so on. In sum, given its high promise in effective analysis of biomedical big data, machine learning is an important topic to be included in the curriculum of training next-generation biomedical informaticians and data scientists.

Natural Language Processing and Text Mining

Each year, hundreds of thousands of new articles are added to PubMed and other literature repositories. Similarly, an important component of EHRs are the many free-text notes that clinicians write after patient encounters. An important goal of natural language processing (NLP) as a sub-discipline of AI is to automate the curation of documents from the scientific literature and from clinical notes to provide an understanding of their content. More specifically, this entails automatically extracting keywords and phrases from documents and annotating them with meaning. Extracted and annotated content can then be converted to structured data that can be integrated with other data in a relational or graph database, for example.

Automating extraction is not easy and is an active area of investigation. As an example, it might be of interest to extract the drugs that are mentioned from a set of clinical notes. This

is not as easy as it sounds and requires the computer to know all the different drug names along with their abbreviations, acronyms, and human shorthand. Humans are quite good at this task but computers struggle. As Hobbs discusses, a computer can get to 60% of all valid extractions relatively easily [48]. However, getting to 90% requires the computer to be aware of many rarely used terms, abbreviations, etc. This can require an enormous amount of time for programming and algorithm refinement. At the core of the automatic approaches lies what is known as “named entity recognition” (NER): the problem of finding references to entities (*mentions*) such as genes, proteins, diseases, drugs, or organisms in natural language text, and tagging them with their location and type. This is a basic building block for almost all other extraction tasks. Named entity recognition in the biomedical domain is generally considered to be more difficult than other domains because of rapid change and inconsistent acronyms and abbreviations. On the other hand, since entity names in biomedical text are longer on average than names from other domains, it is generally much easier – for both humans and automated systems – to determine whether an entity name is present than to determine its exact boundaries [49,50]. As an example, a common open source tool to tag names of genes in the literature is BANNER [51].

The next step in text extraction complexity is that of extracting *relationships* among two (or more) entities. Various extraction systems have been developed for extracting different kinds of relationships. For example, consider the Pharmspresso tool that was developed to find mentions of genes and drugs and their relationships in full text articles [52]. The goal here is to help the computer understand the gene-drug relationships as they were communicated by the authors of the papers. This extracted knowledge can then be used in biomedical research for tasks such as drug-repurposing that has the goal of identifying new indications for drugs already on the market [53].

Bridging NLP with the rest of the knowledge acquisition pipeline is a final key component for data integration in a biological context. This is called *normalization*. This is the problem of determining unique identifiers for the entities recognized in text [54]. For example, for genes, the name mentioned in text would have to be mapped to its corresponding Entrez Gene identifier. Not only is it key for integration, but automating this process has many potential applications in both information extraction and database curation systems. This problem inherits the difficulties of NER, and is particularly hard-hit by ambiguities [55]. The normalization task includes four basic activities. The first is finding a lexicon to which mentions in text will be mapped. The second is identifying and labelling the mentions of the entities of interest in text. This might include handling of prefixes, suffixes, and lists or ranges of entities. The third is matching mentions in the text against the lexicon. This is usually not a one-to-one match, as several “candidate matches” have to be weighted or assigned “confidence levels”. Finally, post-processing is needed to remove false positives due to ambiguity.

Finally, for both biomedical and clinical texts alike, a critical component in reducing false positives in the knowledge acquisition pipeline is determining the context of an entity’s mention within the text. This is called *assertion*. The entities and their contexts described within biomedical texts differ from those within clinical texts [56,57]. Additionally, there can be general and specific contexts for each entity. For example, in biomedical texts, genes

can be described as on or off; in clinical texts, symptoms can be described as affirmed or negated. The accurate representation of entities, their relationships, and their assertions can have critical implications for understanding how biological pathways explain patient clinical profiles and ultimately the discovery of actionable knowledge from population studies.

Future Perspective

Advances in personalized and precision medicine are dependent on our ability to define the unique characteristics of individuals or small groups of people that require specific disease prevention and treatment strategies. Before we can realize this in clinical practice, we must first determine what the important characteristics are for each disease. This naturally relies on our ability to measure as many different internal and external biological processes and exposures as possible. This creates big biomedical data that requires special technology and informatics methods for its storage, management, analysis, and interpretation. We have reviewed here some of the key disciplines and topical areas that are important for the scientist of the future to be familiar with to have an impact on personalized and precision medicine with big data. We have put an emphasis on AI including machine learning and natural language processing. These computational methods are critical to extracting useful information from complex patterns in big data and require complementary training in data management and integration, statistics, and data science in addition the requisite domain expertise to understand the data and the research questions.

We propose a curriculum for training next-generation scientists for biomedical big data that includes specific coursework or equivalent training opportunities that give students a solid background in 1) data management and integration including database experience, 2) statistical analysis including basic concepts and methods in probability and inference, 3) data science including computer programming and methods for improving reproducibility, and 4) artificial intelligence including courses or modules that AI fundamentals, machine learning, and natural language processing for unstructured data. This would be complemented by a course in personalized and precision medicine that provides the motivation for using big data along with appropriate domain knowledge courses from the biomedical sciences. These could be formal courses offered through an in person or online graduate training program or could be intensive short courses that leave students with a fundamental understanding of the area and some hands-on experience.

The most fruitful instruction will occur in graduate courses that can cover the material in some depth in a classroom setting that allows for direct interaction with other students and instructors. This is important because many of the topics presented here will be foreign to students coming from a variety of different backgrounds. Further, learning AI is about more than becoming comfortable with certain computer algorithms, software, and technology. There is an art and a philosophy to working with data that is important and best presented in person in the classroom. Anecdotes and personal experiences are best shared in person and difficult to capture in a condensed setting or in online material. Thus, there is an apprenticeship aspect to learning AI and related topics such as machine learning and natural language processing. One model that might be effective is the flipped classroom where the students study the basic material online on their own time and then come to class to ask

questions and benefit from discussions with the instructor. Indeed, there are a number of studies showing that the flipped classroom mode might be more effective than online or classroom only [58–61].

In addition to covering each of these areas, we recommend that optimal training will occur through a curriculum that is integrated such that each of the courses closely complements all the others through terminology, examples, and perhaps an integrated computational platform where the students can work with big data, carry out analyses, evaluate results, and even design virtual clinical decision support tools that a clinician might use to deliver personalized care. An integrated learning platform that would serve as the basis for each course is possible given the wide range of open-source software tools that are freely available. For example, there are several open-source EHR systems, including OpenMRS [62] that are actively being used in developing countries to manage patient data. These could be adopted locally for free and integrated with a data warehouse and open-source data analysis tools such as R and scikit-learn for data science and machine learning. Designing integrated training programs like this is a challenge but there might be an opportunity for new graduate programs that are in the planning stages for launch within the next few years. Retooling an existing program for this kind of tightly coordinated curriculum is possible but could be prohibitive for some.

Students receiving the kind of training we have outlined here who also have a working knowledge of biomedical science will be in a very strong position to serve as the leaders of scientific projects. This is in contrast to the early days of these quantitative disciplines where computer scientists, data scientists, informaticians, and statisticians were seen as consultants or collaborators who were brought in to address a very specific data management or analysis need with direct involvement in the formulation of the question or the study design. In fact, the next-generation scientist outlined here will be in a strong position to ask more impactful scientific questions because they will have the computational skills to look across disciplines to synthesize information and knowledge in a way that many disciplinarians are not able to. This approach to biomedical research has been called *no-boundary thinking* [10,63].

Executive Summary

- Biomedical big data has arrived and is growing by the day as we measure more and more of our internal and external biological ecosystems.
- The motivation to personalize care is also here and could greatly benefit from scientists receiving the kind of training we have outlined here.
- We propose a curriculum for training next-generation scientists for biomedical big data that includes coursework or equivalent training opportunities that give students a solid background in the areas outlined below.

Data management and integration

- Relational and graph databases are needed to store big data for rapid retrieval. Each type of database has strengths and weaknesses for different types of data and analysis objectives.

- Data integration is important for bringing diverse data types together prior to analysis.

Statistical analysis

- A basic understanding of probability is essential for describing data and is the basis for many AI and machine learning methods.
- Statistical inference including familiarity with basic parametric approaches such as linear regression is necessary to complement AI and machine learning.

Artificial intelligence - Foundations

- A historical overview of the development of artificial intelligence as a field is necessary to provide a foundation for modern developments in AI.
- Equally important are topics in logic describing the common formal language for data and knowledge allowing for an interface between person and machine as well as frameworks for data representation including frames, rules, trees, ontologies, and semantic networks. Representation is an essential topic connecting both deductive and inductive reasoning.

Artificial intelligence – Machine learning

- Machine learning is a subfield of AI dealing with the broader concept of inductive reasoning. In particular, we think of it as a set of methods that can extract patterns from raw data and use these patterns to predict future data or help other types of decision making.
- Supervised learning and unsupervised learning are two major categories. In supervised or predictive learning, we learn a function that maps an input object, represented by a set of features, to an output value. In unsupervised or descriptive learning, we are given just the input data and aim to identify interesting patterns in the data, such as clusters, anomalies, and latent factors.

Artificial intelligence – Natural language processing and text mining

- An important goal of natural language processing (NLP) as a sub-discipline of AI is to automate the curation of documents from the scientific literature and from clinical notes to provide an understanding of their content. More specifically, this entails automatically extracting keywords and phrases from documents and annotating them with meaning.

Acknowledgements:

This work was supported by National Institutes of Health grants TR001878, LM010098, ES013508, and LM012601.

References

1. Collins FS, Varmus H. A new initiative on precision medicine. N. Engl. J. Med 372(9), 793–795 (2015). [PubMed: 25635347]

2. Council NR. Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease [Internet]. Available from: https://www.nap.edu/catalog/13284/toward-precision-medicine-building-a-knowledge-network-for-biomedical-research.

3. Nimmesgern E, Benediktsson I, Norstedt I. Personalized Medicine in Europe. Clin Transl Sci. 10(2), 61–63 (2017). [PubMed: 28083940]

4. Erlewyn-Lajeunesse M, Brathwaite N, Lucas JSA, Warner JO. Recommendations for the administration of influenza vaccine in children allergic to egg. BMJ. 339, b3680 (2009). [PubMed: 19755545]

5. Tran T, Luo W, Phung D, et al. A framework for feature extraction from hospital medical data with applications in risk prediction. BMC Bioinformatics. 15(1), 425 (2014). [PubMed: 25547173]

6. Albers DJ, Elhadad N, Tabak E, Perotte A, Hripcsak G. Dynamical Phenotyping: Using Temporal Analysis of Clinically Collected Physiologic Data to Stratify Populations. PLOS ONE. 9(6), e96443 (2014). [PubMed: 24933368]

7. Westphal JF. Macrolide – induced clinically relevant drug interactions with cytochrome P-450A (CYP) 3A4: an update focused on clarithromycin, azithromycin and dirithromycin. British Journal of Clinical Pharmacology. 50(4), 285–295 (2000). [PubMed: 11012550]

8. Phillips HS, Kharbanda S, Chen R, et al. Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. Cancer Cell. 9(3), 157–173 (2006). [PubMed: 16530701]

9. Warner JL, Denny JC, Kreda DA, Alterovitz G. Seeing the forest through the trees: uncovering phenomic complexity through interactive network visualization. J Am Med Inform Assoc. 22(2), 324–329 (2015). [PubMed: 25336590]

10. Huang X, Jennings SF, Bruce B, et al. Big data - a 21st century science Maginot Line? No-boundary thinking: shifting from the big data paradigm. BioData Min. 8, 7 (2015). [PubMed: 25670967]

11. Angles R, Gutierrez C. Survey of Graph Database Models. ACM Comput. Surv 40(1), 1:1–1:39 (2008).

12. Robinson I, Webber J, Eifrem E. Graph Databases. 1 edition O'Reilly Media, Beijing ; Sebastopol, CA.

13. Harron K, Dibben C, Boyd J, et al. Challenges in administrative data linkage for research. Big Data Soc. 4(2), 2053951717745678 (2017). [PubMed: 30381794]

14. Mamun A-A, Aseltine R, Rajasekaran S. Efficient Record Linkage Algorithms Using Complete Linkage Clustering. PLoS ONE. 11(4), e0154446 (2016). [PubMed: 27124604]

15. Kim D, Joung J-G, Sohn K-A, et al. Knowledge boosting: a graph-based integration approach with multi-omics data and genomic knowledge for cancer clinical outcome prediction. J Am Med Inform Assoc. 22(1), 109–120 (2015). [PubMed: 25002459]

16. Beaulieu-Jones BK, Moore JH. MISSING DATA IMPUTATION IN THE ELECTRONIC HEALTH RECORD USING DEEPLY LEARNED AUTOENCODERS. Pac Symp Biocomput. 22, 207–218 (2017). [PubMed: 27896976]

17. Kim D, Li R, Dudek SM, Ritchie MD. Predicting censored survival data based on the interactions between meta-dimensional omics data in breast cancer. J Biomed Inform. 56, 220–228 (2015). [PubMed: 26048077]

18. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition [Internet]. 2nd ed. Springer-Verlag, New York Available from: // www.springer.com/us/book/9780387848570.

19. Greene CS, Garmire LX, Gilbert JA, Ritchie MD, Hunter LE. Celebrating parasites. Nat. Genet 49(4), 483–484 (2017). [PubMed: 28358134]

20. Huber W, Carey VJ, Gentleman R, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat. Methods 12(2), 115–121 (2015). [PubMed: 25633503]

21. Peng RD. Reproducible research in computational science. Science. 334(6060), 1226–1227 (2011). [PubMed: 22144613]

22. Wickham H, Grolemund G. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. 1 edition O'Reilly Media, Sebastopol, CA.

23. Lotterhos KE, Moore JH, Stapleton AE. Analysis validation has been neglected in the Age of Reproducibility. PLoS Biol. 16(12), e3000070 (2018). [PubMed: 30532167]
24. Beaulieu-Jones BK, Greene CS. Reproducibility of computational workflows is automated using continuous analysis. Nat. Biotechnol 35(4), 342–346 (2017). [PubMed: 28288103]
25. Cole BS, Moore JH. Eleven quick tips for architecting biomedical informatics workflows with cloud computing. PLoS Comput. Biol 14(3), e1005994 (2018). [PubMed: 29596416]
26. Russell S, Norvig P. Artificial Intelligence: A Modern Approach. 3rd ed. Prentice Hall Press, Upper Saddle River, NJ, USA.
27. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat. Med 25(1), 44–56 (2019). [PubMed: 30617339]
28. Ertel W Introduction to Artificial Intelligence [Internet]. 2nd ed. Springer International Publishing Available from: //www.springer.com/us/book/9783319584867.
29. Özdemir V, Hekim N. Birth of Industry 5.0: Making Sense of Big Data with Artificial Intelligence, "The Internet of Things" and Next-Generation Technology Policy. OMICS. 22(1), 65–76 (2018). [PubMed: 29293405]
30. Pfeiffer S The Vision of "Industrie 4.0" in the Making—a Case of Future Told, Tamed, and Traded. Nanoethics. 11(1), 107–121 (2017). [PubMed: 28435474]
31. Duda RO, Shortliffe EH. Expert Systems Research. Science. 220(4594), 261–268 (1983). [PubMed: 6340198]
32. Didier C, Duan W, Dupuy J-P, et al. Acknowledging AI's dark side. Science. 349(6252), 1064–1065 (2015).
33. Özdemir V The Dark Side of the Moon: The Internet of Things, Industry 4.0, and The Quantified Planet. OMICS. 22(10), 637–641 (2018). [PubMed: 30260734]
34. Goodfellow I, Bengio Y, Courville A. Deep Learning. The MIT Press.
35. Murphy KP. Machine Learning: A Probabilistic Perspective. The MIT Press.
36. Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nature Biomedical Engineering. 2(10), 719 (2018).
37. Baldi P Deep Learning in Biomedical Data Science. Annu. Rev. Biomed. Data Sci 1(1), 181–205 (2018).
38. Cao C, Liu F, Tan H, et al. Deep Learning and Its Applications in Biomedicine. Genomics Proteomics Bioinformatics. 16(1), 17–32 (2018). [PubMed: 29522900]
39. Ching T, Himmelstein DS, Beaulieu-Jones BK, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface. 15(141) (2018).
40. Ravi D, Wong C, Deligianni F, et al. Deep Learning for Health Informatics. IEEE J Biomed Health Inform. 21(1), 4–21 (2017). [PubMed: 28055930]
41. Wainberg M, Merico D, Delong A, Frey BJ. Deep learning in biomedicine. Nat. Biotechnol 36(9), 829–838 (2018). [PubMed: 30188539]
42. Olson RS, La Cava W, Orzechowski P, Urbanowicz RJ, Moore JH. PMLB: a large benchmark suite for machine learning evaluation and comparison. BioData Min. 10, 36 (2017). [PubMed: 29238404]
43. Olson RS, Cava WL, Mustahsan Z, Varik A, Moore JH. Data-driven advice for applying machine learning to bioinformatics problems. Pac Symp Biocomput. 23, 192–203 (2018). [PubMed: 29218881]
44. Nantasenamat C, Worachartcheewan A, Jamsak S, et al. AutoWeka: toward an automated data mining software for QSAR and QSPR studies. Methods Mol. Biol 1260, 119–147 (2015). [PubMed: 25502379]
45. Feurer M, Klein A, Eggensperger K, Springenberg J, Blum M, Hutter F. Efficient and Robust Automated Machine Learning [Internet] In: Advances in Neural Information Processing Systems 28. Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R (Eds.), Curran Associates, Inc., 2962–2970 (2015) [cited 2018 Mar 2]. Available from: http://papers.nips.cc/paper/5872-efficient-and-robust-automated-machine-learning.pdf.
46. Olson RS, Bartley N, Urbanowicz RJ, Moore JH. Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science [Internet] In: Proceedings of the Genetic and Evolutionary

Computation Conference 2016, ACM, New York, NY, USA, 485–492 (2016) [cited 2018 Mar 2]. Available from: http://doi.acm.org/10.1145/2908812.2908918.

47. Olson RS, Sipper M, Cava WL, et al. A System for Accessible Artificial Intelligence In: Genetic Programming Theory and Practice XV. Banzhaf W, Olson RS, Tozier W, Riolo R (Eds.), Springer International Publishing, 121–134 (2018).
48. Hobbs JR, Appelt D, Bear J, et al. FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text. arXiv:cmp-lg/9705013 [Internet]. (1997). Available from: http://arxiv.org/abs/cmp-lg/9705013.
49. Leser U, Hakenberg J. What makes a gene name? Named entity recognition in the biomedical literature. Brief. Bioinformatics 6(4), 357–369 (2005). [PubMed: 16420734]
50. Yeh A, Morgan A, Colosimo M, Hirschman L. BioCreAtIvE task 1A: gene mention finding evaluation. BMC Bioinformatics. 6 Suppl 1, S2 (2005).
51. Leaman R, Gonzalez G. BANNER: an executable survey of advances in biomedical named entity recognition. Pac Symp Biocomput. , 652–663 (2008). [PubMed: 18229723]
52. Garten Y, Altman RB. Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. BMC Bioinformatics. 10 Suppl 2, S6 (2009).
53. Yang H-T, Ju J-H, Wong Y-T, Shmulevich I, Chiang J-H. Literature-based discovery of new candidates for drug repurposing. Brief. Bioinformatics 18(3), 488–497 (2017). [PubMed: 27113728]
54. Crim J, McDonald R, Pereira F. Automatically annotating documents with normalized gene lists. BMC Bioinformatics 6 Suppl 1, S13 (2005).
55. Chen L, Liu H, Friedman C. Gene name ambiguity of eukaryotic nomenclatures. Bioinformatics. 21(2), 248–256 (2005). [PubMed: 15333458]
56. Miwa M, Thompson P, McNaught J, Kell DB, Ananiadou S. Extracting semantically enriched events from biomedical literature. BMC Bioinformatics. 13, 108 (2012). [PubMed: 22621266]
57. Velupillai S, Mowery D, South BR, Kvist M, Dalianis H. Recent Advances in Clinical Natural Language Processing in Support of Semantic Analysis. Yearb Med Inform. 10(1), 183–193 (2015). [PubMed: 26293867]
58. Dehghanzadeh S, Jafaraghaee F. Comparing the effects of traditional lecture and flipped classroom on nursing students' critical thinking disposition: A quasi-experimental study. Nurse Educ Today. 71, 151–156 (2018). [PubMed: 30286373]
59. Lee YH, Kim K-J. Enhancement of student perceptions of learner-centeredness and community of inquiry in flipped classrooms. BMC Med Educ. 18(1), 242 (2018). [PubMed: 30352591]
60. Matthew SM, Schoenfeld-Tacher RM, Danielson JA, Warman SM. Flipped Classroom Use in Veterinary Education: A Multinational Survey of Faculty Experiences. J Vet Med Educ. , 1–11 (2018).
61. Xiao N, Thor D, Zheng M, Baek J, Kim G. Flipped classroom narrows the performance gap between low- and high-performing dental students in physiology. Adv Physiol Educ. 42(4), 586–592 (2018). [PubMed: 30251890]
62. Wolfe BA, Mamlin BW, Biondich PG, et al. The OpenMRS system: collaborating toward an open source EMR for developing countries. AMIA Annu Symp Proc. , 1146 (2006). [PubMed: 17238765]
63. Huang X, Bruce B, Buchan A, et al. No-boundary thinking in bioinformatics research. BioData Min. 6(1), 19 (2013). [PubMed: 24192339]