Developing a Framework for Digital Objects in the Big Data to Knowledge (BD2K) Commons: Report from the Commons Framework Pilots Workshop

Kathleen M. Jagodnik[1], Simon Koplev[1], Sherry Jenkins[1], Lucila Ohno-Machado[2,3], Benedict Paten[4], Stephan C. Schurer[5], Michel Dumontier[6], Ruben Verborgh[7], Alex Bui[8,9], Peipei Ping[10], Neil J. McKenna[11], Ravi Madduri[12], Ajay Pillai[13], Avi Ma'ayan[1,*]

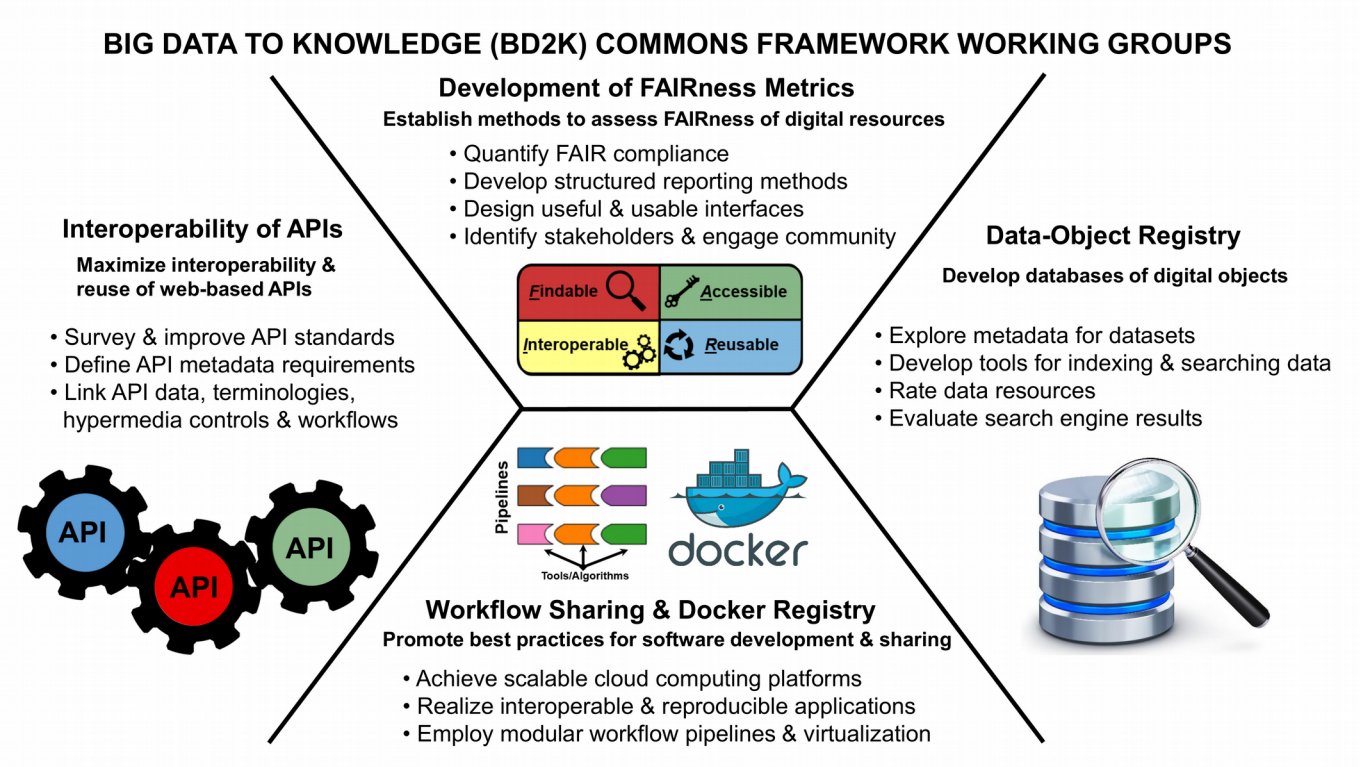
[1]Department of Pharmacological Sciences, BD2K-LINCS Data Coordination and Integration Center, Mount Sinai Center for Bioinformatics, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, Box 1215, New York, NY 10029, USA
[2]Health System Department of Biomedical Informatics, University of California San Diego, 9500 Gilman Dr., La Jolla, CA 92083, USA
[3]Health Services Research, San Diego Veterans Administration Health System, San Diego, CA 92083, USA
[4]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, 1156 High St., Santa Cruz, CA 95060, USA
[5]Department of Molecular and Cellular Pharmacology, University of Miami, 331461120 NW 14th Street, CRB 650 (M-857), Miami, FL 33136, USA
[6]Institute for Data Science, Universiteit Maastricht, Minderbroedersberg 4-6, 6211 LK Maastricht, Netherlands
[7]Ghent University – iMinds Research Foundation Flanders, St. Pietersnieuwstraat 33, 9000 Gent, Belgium
[8]Department of Radiological Sciences, UCLA School of Medicine, Los Angeles, CA 90095, USA
[9]Department of Bioengineering, UCLA Henri Samueli School of Engineering, Los Angeles, CA 90095, USA
[10]Departments of Physiology, Medicine, and Bioinformatics, UCLA School of Medicine, Los Angeles, CA 90095, USA
[11]Department of Molecular and Cellular Biology, Baylor College of Medicine, 1 Baylor Plaza, Houston, TX 77030, USA
[12]Department of Mathematics and Computer Science, Argonne National Laboratory, 9700 S. Cass Avenue, Argonne, IL 60439, USA
[13]Division of Genome Sciences, National Human Genome Research Institute, National Institutes of Health, 31 Center Drive, MSC 2152, 9000 Rockville Pike, Bethesda, MD 20892, USA

*To whom correspondence should be addressed: avi.maayan@mssm.edu

Abstract

The volume and diversity of data in biomedical research has been rapidly increasing in recent years. While such data hold significant promise for accelerating discovery, their use entails many challenges including: the need for adequate computational infrastructure, secure processes for data sharing and access, tools that allow researchers to find and integrate diverse datasets, and standardized methods of analysis. These are just some elements of a complex ecosystem that needs to be built to support the rapid accumulation of these data. The NIH Big Data to Knowledge (BD2K) initiative aims to facilitate digitally enabled biomedical research. Within the BD2K framework, the Commons initiative is intended to establish a virtual environment that will facilitate the use, interoperability, and discoverability of shared digital objects used for research. The BD2K Commons Framework Pilots Working Group (CFPWG) was established to clarify goals and work on pilot projects that would address existing gaps toward realizing the vision of the BD2K Commons. This report reviews highlights from a two-day meeting involving the BD2K CFPWG to provide insights on trends and considerations in advancing Big Data science for biomedical research in the United States.

Graphical Abstract



Highlights

• The NIH BD2K initiative facilitates digitally enabled biomedical research.

• The BD2K Commons Framework Pilots Working Group met to clarify goals.

• This report summarizes key topics of discussion during this March 2016 meeting.

• Four BD2K working groups facilitate and advance use of digital objects for research.

• Methods and tools are being developed to promote interoperable exchange of data.

Keywords: Accessibility, Big Data, FAIR Principles, Findability, Interoperability, Reusability

Introduction

New emerging and diverse technologies that profile biological samples, including cells and tissues, are increasingly producing large quantities of data. The accumulation of such "Big Data" presents an unprecedented opportunity to discover new knowledge that would likely lead to rapid development of novel therapeutics projected to revolutionize health care. Approaches that utilized those new technologies that produce masses of data are transforming disciplines, including pharmacology [1], neuroscience [2], and genomics [3]. However, despite rapid progress, harnessing the full potential of Big Data has many challenges. For example, there is a need to develop effective and more elaborate computational infrastructure, improve methods for data sharing and access, and establish the ability for researchers to integrate diverse datasets, as well as standardize analytical methods. A complex infrastructure must be developed in order to permit the effective use of the digital resources to keep pace with their swift growth in volume and diversity.

The trans-NIH Big Data to Knowledge (BD2K) initiative was established in 2012 to facilitate digital research in biomedical science for the purpose of enabling further scientific discovery and promoting engagement of the scientific community [4]. BD2K encompasses four main aims intended to improve the utility of Big Data employed in biomedical research. First, the initiative seeks to promote the widespread use of biomedical digital resources by ensuring that they are Findable, Accessible, Interoperable, and Reusable (FAIR). Second, BD2K is mandated to develop software tools and methods that will facilitate a more organized approached for the analysis of big biomedical data. Third, BD2K aims to enhance training to engage more students to enter the field, and to disseminate methods and tools useful for biomedical discovery using Big Data; Finally, the BD2K initiative aims to facilitate a data ecosystem that will promote new discoveries through data reuse and data integration.

The BD2K Commons initiative was established within the BD2K framework [5]. One idea behind developing the Commons environment was to make digital objects accessible by a diverse community of researchers through the biomedical and healthCAre Data Discovery Ecosystem (bioCADDIE) data discovery index. The Commons idea was defined at a high conceptual level with a need to begin putting parts of it into practice. The BD2K Commons

Framework Pilots Working Group (CFPWG) met in March 2016, in Bethesda, Maryland, to plan the group's activities, clarify goals of the Year 1 Commons pilot projects, and identify gaps within the existing Commons framework. Participants included representatives from the BD2K Centers of Excellence, the Human Microbiome Project, the Model Organism Database, the Anonymization and Sharing groups, the Commons Credit Pilot initiative, the BioCADDIE project, the BD2K Interoperability projects, and NIH staff. Moderated discussions were held on topics including computational infrastructure, data indices, the development of an NIH Cloud Credits model, and metrics of success for software development projects. This report summarizes some of the key points from this two-day meeting. The report touches general considerations in Big Data biomedical science and presents some innovative solutions.

Developing an Ecosystem for Finding, Accessing, Interoperating, and Reusing Biomedical Data Digital Objects

The Implementation of the FAIR Principles

The digital objects shared among biomedical researchers on the BD2K Commons platform are expected to be Findable, Accessible, Interoperable, and Reusable (FAIR) [6] (Figure 1). The FAIR principles were developed by FORCE11, a growing online community of stakeholders who aim to accelerate and facilitate the sharing of scientific research output through information technologies. FORCE11 motto is that we should not be communicating science results and ideas primarily through print, when there are more advanced digital option now available. Distinct from other guidelines, the FAIR principles are not tied to any specific technology, but instead focused on essential features of data and the metadata that enable the maximization of data reuse. FAIR-compatible datasets require deep metadata elements. The metadata that should be associated with each type of data can be defined in a guideline, and Biosharing.org is a leading resource for the curated collection of data standards that include metadata reporting guidelines [7]. However, while it is agreed upon that having more and better metadata is desired and even required, obtaining it is challenging.

Several BD2K initiatives have adopted the FAIR principles as a core strategic component. These are the NIH BD2K bioCADDIE project, the BD2K Center for Expanded Data Annotation and Retrieval (CEDAR), the BD2K Library of Integrated Network-based Cellular Signatures

(LINCS) Data Coordination and Integration Center (DCIC), and the Big Data for Discovery Science Center (BDDS). bioCADDIE seeks to develop a search engine for biomedical data objects, namely DataMed (Table 1). The search engine is expected to improve through learning by engagement with the biomedical research community. It is an extramural effort modeled after the popular and successful search engine for biomedical publications, PubMed. To achieve the goal of producing a useful search engine for digital objects, bioCADDIE is promoting rich metadata collection and adherence to a shared high-level schema that was recently defined (http://biorxiv.org/content/early/2017/01/25/103143). CEDAR directly complements bioCADDIE by developing infrastructure to help data submitters craft rich, standards-compliant, and repository-mandated metadata [4]. The CEDAR technology aims to facilitate the capture of standardized metadata via reusable templates and template elements that can be linked to terminologies by integration with the National Center for Biomedical Ontology (NCBO) BioPortal, a registry for biomedical ontologies [8]. This is expected to improve data discovery and reuse of biomedical research digital objects. Toward aiding data providers and data consumers to understand the degree to which digital resources adhere to the FAIR principles, a Commons working group has been formed to explore the feasibility and utility of FAIRness metrics with the goal of developing a rating system that could be used to grade resources. This project is still at a conceptual level of discussion that is expected to lead to the development of a prototype tool that would begin to realize these concepts into practice. Via working groups, and in alignment with other BD2K and international efforts, the bioCADDIE project also develops recommendations for data identifiers, data citation, and search result ranking. The BD2K-LINCS DCIC is ensuring that data produced by the LINCS data generation centers maximally adhere to the FAIR principles. The BDDS center is developing tools, services, and standards for organizing, naming, and describing large biomedical datasets for interoperability [9]. In collaboration with bioCADDIE, the BDDS has extended the BagIT specification [10] to enable the exchange of big biomedical datasets. BagIt is a general purpose hierarchical file packaging format.

Lessons from Related Approaches

The working group discussions included a comparison of differing data management strategies employed by industry leaders, including a manually curated directory vs. an automated indexing

strategy. Although the scope of the search universe is remarkably different when comparing general web page search engines, with a search system designed specifically for biomedical digital objects, the successes and failures of early versions of general search engines such as Yahoo and Google led the group to discuss these examples.

The original strategy of Yahoo was to create a directory that required content producers to submit and classify their websites [11]. Such activity can be considered a bottom-up approach. It requires manual curation and manual updating to ensure that entries are classified correctly, and the directory is balanced and was free of spam. Historically, the Yahoo strategy was not scalable but provided a valuable lesson to subsequent development. In contrast, Google indexed websites with a web-bot that crawled pages and ranked them for search without the need for manual curation, a top-down approach. Google's innovative PageRank algorithm rated websites in a mechanical and objective manner, gauging the level of human interest associated with each page by the number of other pages pointing at it [12]. Google improved its search engine by learning from users' queries and from their clicks on results pages, as well as by implementing other enhancements to the PageRank algorithm; for example, personalizing PageRank vectors using URL features including internet domains [13] and generating query-specific importance scores for webpages [14]. The CFPWG discussed the advantages and drawbacks of different approaches, as well as the differences in scope and resources. The consensus was that a hybrid strategy that combines automated and manual (bottom-up and top-down) curation of digital biomedical objects would likely fit best the diverse nature of biomedical datasets and tools. High-quality metadata by manual curation was viewed as particularly necessary for the discovery of biomedical resources, while some automation would be required for scalability. The bioCADDIE metadata specifications, which are aligned with schema.org [15], represent one way for data producers to expose their datasets for passive retrieval by bioCADDIE for indexing into DataMed.

Indexing of Data, Metadata, and Other Digital Objects

Data Discovery

The bioCADDIE initiative has engaged several working groups involving stakeholders from different countries to plan the implementation of the DataMed platform. Processes for ingesting

data from existing data repositories, designing and evaluating user interfaces, and developing benchmarks for the information retrieval task were determined by the various working groups. Different ranking algorithms that have been implemented, or are intended to be employed include: a) Salton's vector space model [16] using Term Frequency (TF) and Inverse Document Frequency (IDF), a widely used approach employed by Elasticsearch [17]; b) Citation count, which is an alternative metric for certain repositories including the Gene Expression Omnibus (GEO); c) Ranking in reverse chronological order, as employed by PubMed; and d) Result relevance after terminology-based query expansion. The DataMed platform does not host the actual data and relies solely on indexing, searching and ranking metadata.

The Importance of Metadata Capture at Publication Submission

Capturing sufficient metadata at the stage of data generation or publication is much more cost- and time-efficient than undertaking subsequent metadata curation. However, this is not current practice because of various reasons, including the lack of advanced annotation tools that make this task easier for experimentalists. When accepted manuscripts report new data, for example transcriptomic gene expression data, or a solved three-dimensional structure of a protein, scientific journals often require that the authors deposit these data into an appropriate repository, with corresponding metadata, and provide an accession number to mention in the publication. Provisions for data deposition help promote data reuse and facilitate reproducibility of results. However, currently not all journals require this type of deposition, and for some data types there is not a clear choice for the repository. Additionally, metadata quality in some repositories does not always conform to standards, making computation across datasets difficult. High-quality metadata is important for data integration, but currently there are few incentives for data producers to annotate their data for proper reuse. While bioCADDIE currently focuses on ingesting data from many repositories and mapping metadata into a global metadata specification schema, other efforts within the BD2K consortium, in particular CEDAR and BD2K-LINCS DCIC, are developing tools, specifications, and best practices to better capture deep and standardized metadata. This can be achieved with auto-complete web forms; machine learning methods that suggest metadata; as well as methods to incentivize the submission of high-quality metadata. Proper metadata annotation involves the mapping of

named entities to qualified standard identifiers. These identifiers are subsequently mapped to higher-order relationship models such as ontologies. If structured correctly, these knowledge models can enable sophisticated semantic search and seamless data integration that can facilitate new biological discovery. While most of the discussions at the meeting stayed at an abstract level, some specific technical recommendations were made. For example, a practical solution for improved sharing of data objects on the web is the JavaScript Object Notation for Linked Data (JSON-LD), a standard format that makes dataset files interpretable by machines [18].

Educational and Other Efforts to Involve the Community

Crowdsourcing

Crowdsourcing in biomedical research involves the distributed effort of numerous individuals to solve substantial and complex problems. In biomedical research, this strategy can be divided into two principal types: microtasks and megatasks [19]. Microtasks are useful to achieve many simple tasks that together produce a quality resource, for example, genome annotation [20, 21], drug indication curation [22], extraction of gene expression signatures [23], and human gene-disease annotation [24], as well as many other examples in recent years [25]. Megatasks address more challenging problems and are set as a competition between teams or individual experts, for example, the reconstruction of the topology of biological networks, or the imputation of missing data by the development of novel algorithms [26]. Challenges related to the use of crowdsourcing include task completion, efficient assessment, and allocation of resources. The BD2K Commons is interested in further promoting the participation of citizen scientists and further engaging the biomedical research community through crowdsourcing opportunities. Participation of individuals with varying levels of scientific experience could be facilitated by tutorials, courses, webinars, and discussion forums within the BD2K Commons initiative.

Expanding Public Use of Big Data and Promoting Associated Education

Among the aims of the BD2K initiative is to enhance training activities related to the methods essential to advance biomedical research involving Big Data. The NIH offers its Commons Data Science training events, Data Science Distinguished Seminar Series, and Frontiers in Data

Science Lecture Series to contribute toward this part of the BD2K initiative. Massive open online courses (MOOCs), including the Big Data Science with the BD2K-LINCS Data Coordination and Integration Center Coursera offering (https://www.coursera.org/learn/bd2k-lincs), provide instructions on how to get started with LINCS and other related Omics resources as well as general instruction about mainstream methods such as clustering, supervised learning and gene set enrichment analysis. Additionally, BD2K established a training and education center that coordinates training activities across the BD2K Centers and other BD2K components. The Big Data for Patients (BD4P) initiative is a data science training program that provides patient advocates with a basic understanding of this platform in order to facilitate their active participation in Big Data research. Several modes of patient engagement have been reported, including crowdsourcing, dynamic consenting, and the use of social networking platforms. The BD2K initiative will benefit from embracing these diverse forms of patient involvement with Big Data, and involving patients to actively participate in the Big Data analysis community. However, it was noted at the meeting that such community involvement also presents risks. As health care becomes more personalized and participatory [27], there is a risk that patients will more likely make uninformed decisions about their own health choices, and due to their lack of proper training, jeopardize their own health.

Software and Systems

The CFPWG established four working subgroups with the aim of bringing some of the high level concepts established by the BD2K Commons into practice. These four working subgroups include: 1) Development of FAIR-ness Metrics; 2) Interoperability of APIs; 3) Data-Object Indexing; and 4) Workflow Sharing & Docker Registries. These working subgroups are open to all interested participants. The working groups are summarized in Table 1 , and to join them, the group chairs may be contacted.

The division of the working groups into four segments is aligned with various existing standards and software development efforts. In Table 2 we list some relevant efforts divided into the following categories: API, computational platforms, initiatives, metrics, searching and indexing projects, and standards.

Software Repositories, APIs, Docker Containers, and Interactive Notebooks

Tools and workflows operate on raw experimental data to generate new knowledge by abstracting, visualizing, summarizing, and integrating it with other data. Datasets are processed in many different ways, and new datasets can result from the processing of the original data (Figure 2). It is thus critical that all the tools, algorithms, pipelines, and workflows are considered as digital objects, and are also catalogued and annotated in a similar way that DataMed (Table 2, searching and indexing category) is indexing datasets for search. Besides improved data handling with enhanced metadata for tools and pipelines, there is also a need to develop better standards, including metadata, for organizing and indexing tools and workflows. For example, one effort carried out by the HeartBD2K center, named Aztec (Table 2, searching and indexing category) is developing a directory of bioinformatics tools with their corresponding metadata. Aztec provides the ability to automatically create pipelines of tools by relating the upstream/downstream or input/output relationships of these tools. This feature is also being developed by other tool repositories such as OMICtools [62]. There are other efforts to build directories of bioinformatics tools, including the Online Bioinformatics Resources Collection (OBRC) [28] and ExPASy:SIB [29]. A complementary effort led by CEDAR in collaboration with the HeartBD2K center aims to develop smartAPI (Table 2, API category), a coordinated facility for the intelligent annotation of web-based APIs. smartAPI aims to improve finding and reusing APIs developed for accessing and operating on biomedical research data. The smartAPI initiative is built on the code base of the Swagger editor [30]. The Swagger editor is a standards-compliant API metadata authoring tool. The API Interoperability Commons working subgroup is examining the usability and utility of smartAPI and other API interoperability technologies.

Another important development in this area has been the introduction of Docker containers [31] (Table 2, Computing Platform category). This entails the ability to package software tools developed using different technologies as a relatively lightweight executable and installable package that can run on any server that supports Docker. Dockerizing apps makes software applications more reusable and accessible [32]. It also provides the opportunity to chain tools for developing workflows and pipelines. The Dockstore.org project (Table 2, Searching and

Indexing category) is jointly developed by the Ontario Institute for Cancer Research (OICR), the BD2K Center for Big Data in Translational Genomics (BD2K Genomics Center), and the Global Alliance for Genomics and Health (GA4GH) [33, 34]. It is similar to Aztec.bio in that it is developing a curated repository of tools and workflows with searchable metadata. Moreover, Dockstore provides all tools in Docker containers ready to be added to workflows by using the Common Workflow Language (CWL) [35] and the Workflow Description Language (WDL) (Table 2, Standards category). Docker containers that are coded in CWL and WDL facilitate scalable, efficient, and reproducible deployment of tools across platforms including cloud environments. In addition, the BD2K Genomics Center has developed Toil [36] (Table 2, Standards category), and the BDDS center has developed Globus Genomics [37] (Table 2, Computing Platform category). Similar to Cromwell, Nextflow, and Arvados (Table 2, Computing Platform category), the aim of Toil and Globus Genomics is to make it easier for users to run large-scale analyses. For Toil, this was recently demonstrated in a single combined workflow facilitating the successful analysis of 20,000 next-generation sequencing (NGS) samples on the Amazon Web Services (AWS) platform in under four days across 32,000 processing cores. The pilot project Reproducibility by Design complements these efforts by providing the iDASH [38], a HIPAA-compliant compute environment in which Docker containers can be used to analyze protected health information, including human genome sequences and corresponding phenotypes derived from electronic medical health records.

Yet another relevant development is the emergence of online interactive notebook [39]. Systems such as Jupyter/IPython or R Markdown provide a web-based platform where users can interactively execute open source scripts online [40], together with embedded markup text, and interactive animated figures. Such systems can make publication of data processing pipelines transparent, shareable, and modifiable for reuse. The ability of interactive notebooks to provide an easy way to document code by incorporating text and figures within a notebook, can potentially become a new mode of publishing biomedical research results. It was suggested at the meeting that scientific journals should better support this form of publication.

Usability of Software Tools

One aspect of software development that is not rewarded by current funding mechanisms is investment in the improvement and sustainability of existing useful tools and databases. Most tools that are developed and published in the area of bioinformatics do not always consider the user perspective and requirements first [41]. One recommendation from the meeting is to start thinking about how to incentivize more user-centered design principles [42]. While initially slowing the development process, the implementation of these principles can accelerate the development phase and ultimately yield tools that are more suitable for their intended use. It was recommended at the meeting that the inclusion of user-centered design in proposed projects to develop computational resources should be considered as a criteria for evaluating grant applications by funding agencies. Key usability metrics include effectiveness, efficiency, and the perceived satisfaction of bioinformatics experts and bench researchers. Usability metrics can be categorized according to aspects such as: time to complete a task, layout complexity, error frequency, and task effectiveness. Techniques for studying and improving usability might include user testing sessions, user surveys and focus groups, design workshops, and the provision of user guides and training resources [43]. The social context influencing the use of bioinformatics tools requires further consideration [44]. At the meeting, the presence of social media and community message boards was brought up. Sites such as ResearchGate [45], Biostars [46], and StackOverflow [47] have been highly successful, suggesting that community-building platforms, which are living ecosystems that benefit their users, should be considered by BD2K as key resources for accomplishing a variety of goals, including better implementation of software. In general, proven practices in usable design and web engineering could inform the development of effective bioinformatics tools [48]. This is a new endeavor for extramural NIH-funded projects, but new policies and approaches are expected to eventually penetrate.

A related concern is the lack of incentives for academic investigators to maintain widely used tools. If funding expires, there is a risk that successful tools will disappear due to insufficient support and upgrades. Current mechanisms for NIH grant support, and guidelines for the review process, require innovation and discovery, so grant proposals that request funds to maintain and incrementally enhance an existing valuable tool are at risk of not receiving funds. It was recommended at the meeting that funding agencies consider establishing new mechanisms that would support existing digital resources that are valuable to the research

community in addition to maintaining existing mechanisms that promote innovation. Dockerizing tools and databases opens the opportunity to host such applications on a public server so they can continue to serve their users, even after the projects expire. It was noted at the meeting that there are too many dead links to previously published tools and resources that can potentially benefit existing users.

The Commons Cloud Credits Business Model

A central and timely consideration for BD2K is the transition of software and data to the cloud. In the 1970s and 1980s, the cost of personal computers was high, and hence most scientific computation was done through a client-server environment. In the 1990s and the 2000s, there was a shift, and most bioinformatics tools were designed to run locally on a desktop. However, the past 10 years have seen a shift back to client-server computing. Most of the newest and popular bioinformatics tools and databases are web-based, and increasingly also cloud-based, due to lower storage costs, faster speed of communication networks, and increased size of data files, particularly for genomic sequencing and imaging data. The cloud provides several additional advantages over local computing. Cloud providers manage the hardware and software resources such that storage and computing are done remotely without the need of the user, or the tool developer, to know exactly where and how this is accomplished. Managing the cost of cloud computing services for biomedical research was a central topic of the meeting. Practical questions arose such as: Who should pay for cloud services? If the NIH covers cloud computing costs, should a principal investigator be required to submit a proposal for using such services? How much cloud computing is needed to enable scientific progress? Can usage of tools by laboratories help track demand for better allocation of resources? The BD2K team at NIH has developed a cloud credits pilot model to begin addressing some these questions and needs.

Benchmarking to Employ the Best Available Tools

Disconcertingly, highly cited bioinformatics tools often draw the most users, whereby the adoption of newer, potentially superior solutions can be overlooked. For example, a popular method to process DNA or RNA sequencing, or mass spectrometry data, may miss important

results that could be detected by a better but less widely adopted pipeline. This situation is due to a lack of established objective benchmarks that can be used to compare and evaluate tools (Figure 2). In addition, improved provenance of tools and data processing pipelines is required to ensure the reproducibility of results. There are many factors that influence users' choice of bioinformatics tools [49]. These factors can be grouped into system-related factors such as: platform, interface, and cost; considerations of functionality include: customizability, scalability, and speed; overall quality of the tool; and personal factors such as: usability and availability of documentation. The development of a system to filter and rank bioinformatics tools according to their objective performance to extract more knowledge from the data is an important goal of BD2K.

Benchmarking Pipelines

Benchmarking involves the comparison of algorithms and related tools for processing data at different stages of analysis (Figure 2). Best practices for benchmarking must consider that: 1) Evaluation metrics can vary widely and influence the rankings of tools and the algorithm implemented within them; 2) The data used to compare tools may be critical in affecting the rankings; 3) Evaluation metrics have different aspects such as speed, scalability, usability, accuracy, precision, and sensitivity, which have different levels of importance for different types of data and research projects; and 4) The use of synthetic data vs. real data can influence the results, with each type of data having advantages and disadvantages. Benchmarking is directly related to megatask crowdsourcing challenges such as those run by the Dialogue on Reverse-Engineering Assessment and Methods (DREAM) [50] or Kaggle [51]. For example, the bioCADDIE team developed an information retrieval crowdsourcing challenge and generated benchmark indexing data to evaluate the submissions of teams (https://biocaddie.org/biocaddie-2016-dataset-retrieval-challenge-registration). The discussions at the meeting emphasized the importance of such crowdsourcing projects to benchmark biomedical informatics practices.

Case Studies

There are several pilot projects and new tools and databases that have been already developed by BD2K awardees to address some of the challenges discussed. For example, the BD2K LINCS-DCIC developed a system that aggregates knowledge about genes and proteins called the Harmonizome [52]. The Harmonizome resource was brought online in September 2015, and since then, as of April 30th 2017, the Harmonizome website and API attracted over 112,585 unique users based on Google Analytics. The BD2K LINCS-DCIC also conducted a successful crowdsourcing project in which participants used a Chrome extension developed by the center [53] to extract gene expression signatures from the Gene Expression Omnibus (GEO) for single drug, gene, and disease perturbations. The outcome of the project was the identification of many novel associations between genes, drugs, and diseases [23]. As mentioned previously, the bioCADDIE project, which started two years ago, developed a prototype search engine for datasets based on indexing of 63 highly utilized repositories. The DataMed search engine backend already indexed more than 1.3 million datasets. Community engagement was achieved through working groups and funded pilot projects on various topics, including data citation, result ranking, and automated metadata extraction.

The Commons Ecosystem

Risks of Decisions Driven by Big Data

The proper use of Big Data has great promise in informing medical and scientific decisions; with sound analysis, more comprehensive collection of data is expected to reveal better biomarkers for diseases [9] and to improve health care in numerous other ways [54]. However, the use of Big Data also raises challenges associated with incompleteness and inaccuracy of data collection, processing, and analysis [55]. In addition, if multiple datasets are integrated, challenges related to differences in formatting and nomenclature among datasets often arise. Caution is needed when interpreting reports generated by Big Data methodologies because large sample size can lead to inferential errors. Biases associated with errors that stem from poor study design or biased sampling can be magnified [56]. It should be considered that at least currently data-driven decisions are not always more correct that decision made by expert knowledge, and hence, caution is advised when advancing toward data-driven solutions.

Metrics and Evaluation

One area in which Big Data is already making strides and demonstrating impact is improved metrics and evaluation of researchers' output and resources' impact. The opportunity is to make evaluation more objective and transparent so that resources are allocated to efforts that are more productive. Not long ago, biomedical research relied on few sources for such evaluation, mostly through peer review and journal citation. With digitization of research output, and tracking analytics tools such as Google Analytics [57] and Altmetric [58] there is an increase in the ability for instant assessment of the popularity and usage of research output. These metrics and analytics tools can provide more objective assessment of impact and productivity. However, relying on algorithms alone to objectively assess impact and productivity of scientific research can be dangerous because elements of scientific quality are complex [59]. Impact and productivity have a temporal aspect; it may take time for a method to become adopted, or for a researcher's work to be fully appreciated. The numbers do not always tell the whole story. Caution should be used when comparing reported statistics for websites and tools because different web analytics providers use different methods to compute web access statistics. Reported statistics currently have few methods for systematic verification.

Future Vision

The attendees of the meeting selected several objectives on which to focus in near-future activities; a working subgroup has been formed for each of these objectives. The digital object registry subgroup has specified that this registry should be open source in nature, standardized, customizable, scalable, extensible, redeployable, decentralized, collaborative and semantic, using ontologies to describe content. The working subgroup addressing API-related planning has been developing API metadata requirements based on a survey performed to assess the properties of metadata elements, and to specify necessary attributes (smartAPI, Table 2). The intention is that these subgroups will begin translating the BD2K Commons principles into practice with the hope that the BD2K Commons will gradually emerge. The Commons is expected to consist of many interacting components where some efforts will succeed while

others may fail. It is possible that major impact may only be realized in the long term, so measuring it now could be challenging.

It should be mentioned that there are domestic and international initiatives similar to BD2K. The National Science Foundation (NSF) established the eXtreme Science and Engineering Discovery Environment (XSEDE) program with the aim to serve diverse and integrated digital resources and advanced cyberinfrastructure supporting a wide range of scientific endeavors, from biological and geological sciences, to social, economic, and behavioral research, and electrical and structural engineering [60]. XSEDE has a strong Training, Education, and Outreach Services (TEOS) program that seeks to diversify the STEM workforce by offering training classes and online training resources, and working in collaboration with higher education institutions to develop certificate and degree programs in STEM fields. The XSEDE initiative complements BD2K, with the former having a broad scientific focus across numerous theoretical, experimental, and engineering disciplines, and the latter having a more strictly biomedical research focus to improve human health and cure disease. Additionally, ELIXIR is a European initiative that is similar to BD2K in its focus on biomedical research, Big Data solutions, and training activities [61]. ELIXIR aims to improve the interoperability and accessibility of bioinformatics resources for academia and industry in Europe.

In the future, the BD2K Commons initiative seeks to extend biomedical discovery through the development of a computing environment that supports the access, use, and storage of biomedical research digital objects; to support the transition of publicly available datasets to be more compliant with the FAIR principles; and to facilitate software tools and services that are scalable, shareable, and interoperable with other registries, repositories, and resources. Input from the research community and the public will be essential to realize these goals, ensuring that an accessible and useful organization of resources is developed.

Table 1: NIH Commons Pilot Projects Working Groups

Working Group	Chairs	Brief Description	Topics
Development of FAIR-ness Metrics	Neil McKenna (Baylor) & Michel Dumontier (U. Maastricht)	Identify and prototype methods to assess the FAIRness of a digital resource.	Identification of stakeholders, structured reporting methods, quantifying FAIRness, community engagement strategies, utility and usability of capture, and reporting interfaces
Data-Object Registry	Lucila Ohno-Machado (UCSD) & Michel Dumontier (U. Maastricht)	Promote integration of activities related to the development of easy-to-use, broad-scope "catalogs" of data objects.	Metadata for datasets, rating of data resources, tools to facilitate indexing and search for data objects, and evaluation of search engine results
Workflow Sharing and Docker Registry	Umberto Ravaioli (U. Illinois) & Brian O'Connor (UCSC)	Promote best practices for software development, deployment, and sharing, through the use of modular workflow pipelines and virtualization based on Docker containers.	Efficient scalability of cloud computing platforms, simplifying the realization of interoperable and reproducible software applications
Interoperability of APIs	Chunlei Wu (Scripps Research Institute) & Michel Dumontier (U. Maastrict)	Develop a strategy for maximizing interoperability and reuse of web-based biomedical APIs.	Topics of interest include API standards, API metadata requirements, Linking API Data, terminologies, hypermedia controls, matchmaking, and workflows

Table 2: Online Resources for Digital Object Sharing in Biomedical Research

Resource Name	Description	URL	Category	BD2K?
Big Data Genomics: ADAM	API proxy to harmonize genomics-oriented APIs	http://bdgenomics.org/projects/adam/	API	Yes
smartAPI	Enables API publishers to annotate their services	https://github.com/Network-of-BioThings/smartAPI	API	Yes
Arvados	Platform for data science employing very large datasets	https://arvados.org/	Computing Platform	No
Docker	Bundle software into packages	https://www.docker.com/	Computing Platform	No
Globus Genomics	Galaxy-based platform for NGS analysis	http://www.globus.org/genomics	Computing Platform	Yes
Nextflow	Data-driven computational pipelines	https://www.nextflow.io/	Computing Platform	No
Project Jupyter	Interactive, web-based computational environment	https://github.com/jupyter/	Computing Platform	No
ELIXIR	European effort similar	http://www.elixir-	Initiative	No

	to BD2K; a distributed infrastructure that coordinates, integrates, and maintains bioinformatics resources	europe.org/		
FORCE11	International consortium focused on the future of scientific publications	https://www.force11.org/	Initiative	No
NIH Commons	A shared environment for the use, interoperability, and discoverability of digital research objects	https://datascience.nih.gov/commons/	Initiative	Yes
XSEDE	NSF Big Data initiative	https://www.xsede.org/	Initiative	No
Altmetric	Efficient platform to measure the impact of research	http://www.altmetric.com/	Metrics	No
Orbitera	Tracks expenses from cloud computing providers	http://www.orbitera.com/	Metrics	Yes
Aztec	Indexing for software	http://aztec.bio/	Search and Indexing	Yes
bioCADDIE	Discovery index search engine	https://biocaddie.org/	Search and Indexing	Yes
DataMed	Prototype biomedical data search engine	https://datamed.org/	Search and Indexing	Yes
DockerStore	Depository for Docker containers for tools and workflows from science	https://dockstore.org/	Search and Indexing	Yes
Harmonizome	Repository for aggregating data collected from genes and proteins	http;//amp.pharm.mssm.edu/Harmonizome	Search and Indexing	Yes
Common Workflow Language	A language for computational pipelines	https://github.com/common-workflow-language/	Standards	Yes
Cromwell	Workflow execution engine using Workflow Description Language	https://github.com/broadinstitute/cromwell/	Standards	No
Toil	Specification for pipelines	https://github.com/BD2KGenomics/toil/	Standards	Yes
Workflow Description Language	Language used to implement Docker containers	https://github.com/broadinstitute/wdl/	Standards	No
Biosharing.org	A repository of biomedical standards, policies, and databases	https://biosharing.org/	Standards	Yes

References

1. Xie L, Draizen EJ, Bourne PE: **Harnessing big data for systems pharmacology**. *Annual Review of Pharmacology and Toxicology* 2017, **57**:245-262.
2. Landhuis E: **Neuroscience: Big brain, big data**. *Nature* 2017, **541**(7638):559-561.
3. Schmidt B, Hildebrandt A: **Next-generation sequencing: big data meets high performance computing**. *Drug Discovery Today* 2017.
4. Margolis R, Derr L, Dunn M, Huerta M, Larkin J, Sheehan J, Guyer M, Green ED: **The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data**. *Journal of the American Medical Informatics Association* 2014, **21**(6):957-958.
5. Bourne PE, Bonazzi V, Dunn M, Green ED, Guyer M, Komatsoulis G, Larkin J, Russell B: **The NIH Big Data to Knowledge (BD2K) initiative**. *Journal of the American Medical Informatics Association* 2015, **22**(6):1114-1114.
6. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE: **The FAIR Guiding Principles for scientific data management and stewardship**. *Scientific data* 2016, **3**.
7. Sansone S-A: **Omics data sharing–BioSharing: On data policies's plans and reporting standards**. 2010.
8. Fergerson RW, Alexander PR, Dorf M, Gonçalves RS, Salvadores M, Skrenchuk A, Vendetti J, Musen MA: **NCBO BioPortal Version 4**. 2015.
9. Toga AW, Foster I, Kesselman C, Madduri R, Chard K, Deutsch EW, Price ND, Glusman G, Heavner BD, Dinov ID: **Big biomedical data as the key resource for discovery science**. *Journal of the American Medical Informatics Association* 2015:ocv077.
10. Kunze J, Littman J, Madden L, Summers E, Boyko A, Vargas B: **The BagIt File Packaging Format**. *https://toolsietforg/html/draft-kunze-bagit-13* 2016.
11. Labrou Y, Finin T: **Yahoo! as an ontology: using Yahoo! categories to describe documents**. In: *Proceedings of the eighth international conference on Information and knowledge management: 1999*. ACM: 180-187.
12. Page L, Brin S, Motwani R, Winograd T: **The PageRank citation ranking: bringing order to the web**. 1999.
13. Aktas MS, Nacar MA, Menczer F: **Personalizing pagerank based on domain profiles**. In: *Proc of WebKDD: 2004*. 22-25.
14. Haveliwala TH: **Topic-sensitive pagerank**. In: *Proceedings of the 11th international conference on World Wide Web: 2002*. ACM: 517-526.
15. Ronallo J: **HTML5 Microdata and Schema. org**. *Code4Lib Journal* 2012, **16**.
16. Salton G, Wong A, Yang C-S: **A vector space model for automatic indexing**. *Communications of the ACM* 1975, **18**(11):613-620.
17. Kononenko O, Baysal O, Holmes R, Godfrey MW: **Mining modern repositories with elasticsearch**. In: *Proceedings of the 11th Working Conference on Mining Software Repositories: 2014*. ACM: 328-331.
18. Sporny M, Kellogg G, Lanthaler M, Group WCRW: **JSON-LD 1.0: a JSON-based serialization for linked data**. *W3C Recommendation* 2014, **16**.
19. Good BM, Su AI: **Crowdsourcing for bioinformatics**. *Bioinformatics* 2013:btt333.
20. Hingamp P, Brochier C, Talla E, Gautheret D, Thieffry D, Herrmann C: **Metagenome annotation using a distributed grid of undergraduate students**. *PLoS Biol* 2008, **6**(11):e296.
21. Brister JR, Le Mercier P, Hu JC: **Microbial virus genome annotation—Mustering the troops to fight the sequence onslaught**. *Virology* 2012, **434**(2):175-180.
22. Khare R, Burger JD, Aberdeen JS, Tresner-Kirsch DW, Corrales TJ, Hirchman L, Lu Z: **Scaling drug indication curation through crowdsourcing**. *Database* 2015, **2015**:bav016.

23. Wang Z, Monteiro CD, Jagodnik KM, Fernandez NF, Gundersen GW, Rouillard AD, Jenkins SL, Feldmann AS, Hu KS, McDermott MG: **Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd**. *Nature communications* 2016, **7**.
24. Loguercio S, Good BM, Su AI: **Dizeez: an online game for human gene-disease annotation**. *PloS one* 2013, **8**(8):e71171.
25. Khare R, Good BM, Leaman R, Su AI, Lu Z: **Crowdsourcing in biomedicine: challenges and opportunities**. *Briefings in bioinformatics* 2015:bbv021.
26. Prill RJ, Saez-Rodriguez J, Alexopoulos LG, Sorger PK, Stolovitzky G: **Crowdsourcing network inference: the DREAM predictive signaling network challenge**. *Science signaling* 2011, **4**(189):mr7.
27. Hood L, Friend SH: **Predictive, personalized, preventive, participatory (P4) cancer medicine**. *Nature Reviews Clinical Oncology* 2011, **8**(3):184-187.
28. Chen Y-B, Chattopadhyay A, Bergen P, Gadd C, Tannery N: **The Online Bioinformatics Resources Collection at the University of Pittsburgh Health Sciences Library System—a one-stop gateway to online bioinformatics databases and software tools**. *Nucleic acids research* 2007, **35**(suppl 1):D780-D785.
29. Artimo P, Jonnalagedda M, Arnold K, Baratin D, Csardi G, De Castro E, Duvaud S, Flegel V, Fortier A, Gasteiger E: **ExPASy: SIB bioinformatics resource portal**. *Nucleic acids research* 2012:gks400.
30. Lucky MN, Cremaschi M, Lodigiani B, Menolascina A, De Paoli F: **Enriching API Descriptions by Adding API Profiles Through Semantic Annotation**. In: *International Conference on Service-Oriented Computing: 2016*. Springer: 780-794.
31. Merkel D: **Docker: lightweight linux containers for consistent development and deployment**. *Linux Journal* 2014, **2014**(239):2.
32. Chamberlain R, Schommer J: **Using Docker to support reproducible research**. *DOI: http://dx doi org/106084/m9 figshare* 2014, **1101910**.
33. Terry SF: **The global alliance for genomics & health**. *Genetic testing and molecular biomarkers* 2014, **18**(6):375-376.
34. Paten B, Diekhans M, Druker BJ, Friend S, Guinney J, Gassner N, Guttman M, Kent WJ, Mantey P, Margolin AA: **The NIH BD2K Center for Big Data in Translational Genomics**. *Journal of the American Medical Informatics Association* 2015, **22**(6):1143-1147.
35. Leipzig J: **A review of bioinformatic pipeline frameworks**. *Briefings in bioinformatics* 2016:bbw020.
36. Vivian J, Rao A, Nothaft FA, Ketchum C, Armstrong J, Novak A, Pfeil J, Narkizian J, Deran AD, Musselman-Brown A: **Rapid and efficient analysis of 20,000 RNA-seq samples with Toil**. *bioRxiv* 2016:062497.
37. Madduri RK, Sulakhe D, Lacinski L, Liu B, Rodriguez A, Chard K, Dave UJ, Foster IT: **Experiences building Globus Genomics: a next-generation sequencing analysis service using Galaxy, Globus, and Amazon Web Services**. *Concurrency and Computation: Practice and Experience* 2014, **26**(13):2266-2279.
38. Ohno-Machado L, Bafna V, Boxwala AA, Chapman BE, Chapman WW, Chaudhuri K, Day ME, Farcas C, Heintzman ND, Jiang X: **iDASH: integrating data for analysis, anonymization, and sharing**. *Journal of the American Medical Informatics Association* 2012, **19**(2):196-201.
39. Shen H: **Interactive notebooks: Sharing the code**. *Nature* 2014, **515**(7525):151-152.
40. Ragan-Kelley M, Perez F, Granger B, Kluyver T, Ivanov P, Frederic J, Bussonier M: **The Jupyter/IPython architecture: a unified view of computational research, from interactive exploration to communication and publication**. In: *AGU Fall Meeting Abstracts: 2014*. 07.
41. Al-Ageel N, Al-Wabil A, Badr G, AlOmar N: **Human factors in the design and evaluation of bioinformatics tools**. *Procedia Manufacturing* 2015, **3**:2003-2010.
42. Pavelin K, Cham JA, de Matos P, Brooksbank C, Cameron G, Steinbeck C: **Bioinformatics meets user-centred design: a perspective**. *PLoS Comput Biol* 2012, **8**(7):e1002554.

43. Macaulay C, Sloan D, Jiang X, Forbes P, Loynton S, Swedlow JR, Gregor P: **Usability and user-centered design in scientific software development**. *IEEE Software* 2009, **26**(1):96.
44. Douglas C, Goulding R, Farris L, Atkinson-Grosjean J: **Socio-Cultural characteristics of usability of bioinformatics databases and tools**. *Interdisciplinary Science Reviews* 2011, **36**(1):55-71.
45. Thelwall M, Kousha K: **ResearchGate: Disseminating, communicating, and measuring scholarship?** *Journal of the Association for Information Science and Technology* 2015, **66**(5):876-889.
46. Parnell LD, Lindenbaum P, Shameer K, Dall'Olio GM, Swan DC, Jensen LJ, Cockell SJ, Pedersen BS, Mangan ME, Miller CA: **BioStar: an online question & answer resource for the bioinformatics community**. *PLoS Comput Biol* 2011, **7**(10):e1002216.
47. Hanrahan BV, Convertino G, Nelson L: **Modeling problem difficulty and expertise in stackoverflow**. In: *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion: 2012*. ACM: 91-94.
48. Bolchini D, Finkelstein A, Perrone V, Nagl S: **Better bioinformatics through usability analysis**. *Bioinformatics* 2009, **25**(3):406-412.
49. Bartlett J, Ishimura Y, Kloda L: **Why choose this one?: Factors in scientists' selection of bioinformatics tools**. *Information Research* 2011, **16**(1):15.
50. Stolovitzky G, Monroe D, Califano A: **Dialogue on Reverse-Engineering Assessment and Methods**. *Annals of the New York Academy of Sciences* 2007, **1115**(1):1-22.
51. Carpenter J: **May the best analyst win**. *Science* 2011, **331**(6018):698-699.
52. Rouillard AD, Gundersen GW, Fernandez NF, Wang Z, Monteiro CD, McDermott MG, Ma'ayan A: **The Harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins**. *Database* 2016, **2016**:baw100.
53. Gundersen GW, Jones MR, Rouillard AD, Kou Y, Monteiro CD, Fledmann AS, Hu KS, Ma'ayan A: **GEO2Enrichr: browser extension and server app to extract gene sets from GEO and analyze them for biological functions**. *Bioinformatics* 2015:btv297.
54. Nambiar R, Bhardwaj R, Sethi A, Vargheese R: **A look at challenges and opportunities of big data analytics in healthcare**. In: *Big Data, 2013 IEEE International Conference on: 2013*. IEEE: 17-22.
55. Liu J, Li J, Li W, Wu J: **Rethinking big data: A review on the data quality and usage issues**. *ISPRS Journal of Photogrammetry and Remote Sensing* 2016, **115**:134-142.
56. Kaplan RM, Chambers DA, Glasgow RE: **Big data and large sample size: a cautionary note on the potential for bias**. *Clinical and translational science* 2014, **7**(4):342-346.
57. Clifton B: **Advanced web metrics with Google Analytics**: John Wiley & Sons; 2012.
58. Adie E, Roe W: **Altmetric: enriching scholarly content with article-level discussion and metrics**. *Learned Publishing* 2013, **26**(1):11-17.
59. Costas R, Zahedi Z, Wouters P: **Do "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective**. *Journal of the Association for Information Science and Technology* 2015, **66**(10):2003-2019.
60. Towns J, Cockerill T, Dahan M, Foster I, Gaither K, Grimshaw A, Hazlewood V, Lathrop S, Lifka D, Peterson GD: **XSEDE: accelerating scientific discovery**. *Computing in Science & Engineering* 2014, **16**(5):62-74.
61. Crosswell LC, Thornton JM: **ELIXIR: a distributed infrastructure for European biological data**. *Trends in biotechnology* 2012, **30**(5):241.
62. Henry VJ, Bandrowski AE, Pepin AS, Gonzalez BJ, Desfeux A: **OMICtools: an informative directory for multi-omic data analysis**. *Database* 2014, bau069.

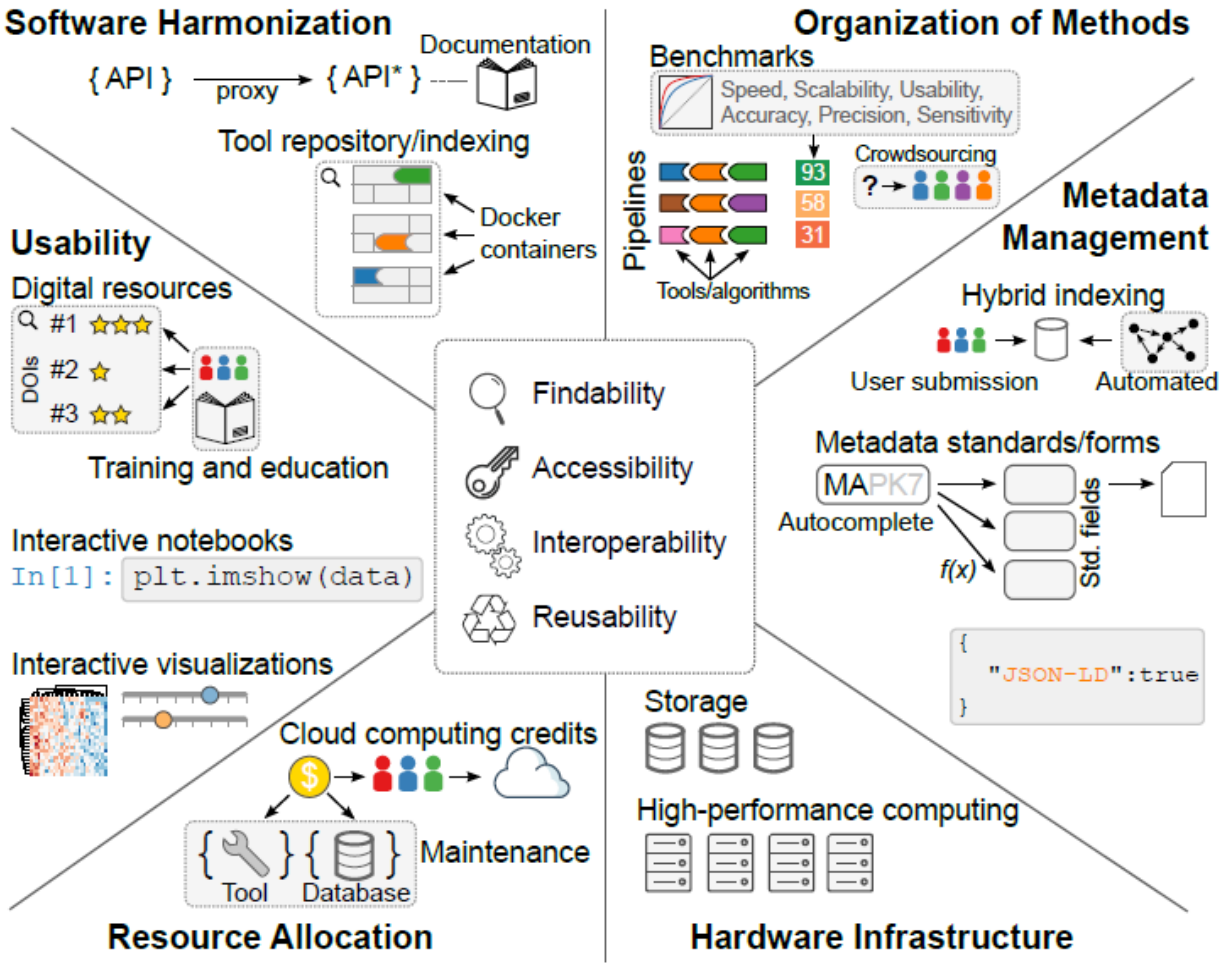


Figure 1. The Findability, Accessibility, Interoperability, and Reusability (FAIR) principles in the context of software harmonization, organization of methods, metadata management, hardware infrastructure, resource allocation, and usability. Organization of Methods illustrates crowdsourcing efforts to establish benchmarks for pipelines and algorithm performance. **Metadata Management** can include hybrid indexing that pairs manual submissions by users with automated analyses (bottom-up and top-down approaches). Metadata standards and forms are employed to implement this concept. **Hardware Infrastructure** includes cloud-based storage and high-performance computing solutions. **Resource Allocation** employs the idea of cloud computing credits model in which funds for computational resources are allocated based on need and cost. **Usability** considerations include training and education related to using digital resources, employing of interactive notebooks to allow reproducible and open analyses, and developing interactive data visualizations that permit dynamic modifications of displays for different data views. **Software Harmonization** facilitates compatibility between application programming interfaces (APIs), and Docker containers can encapsulate implementation detail to facilitate the management, reuse and indexing of tool and data repositories.

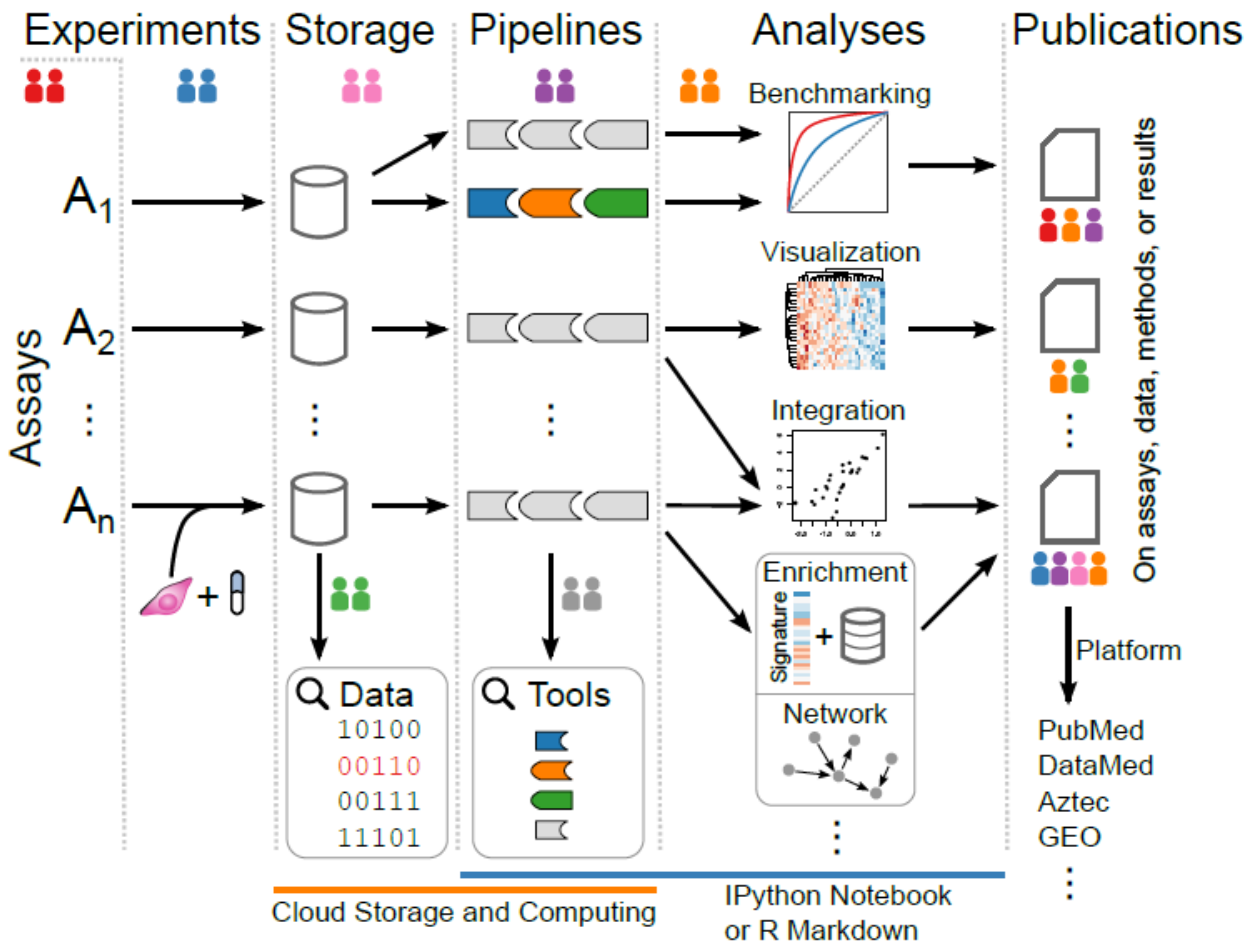


Figure 2. Workflows for biomedical research involving Big Data. Wet bench experiments collect measurements of cellular and tissue variables under different conditions and time points; the resulting data are processed via pipelines that perform data processing in a series of sequential steps. Different analysis steps can be benchmarked to objectively evaluate the quality of a pipeline by comparing pipelines through an objective benchmark. At the final step of the analysis, data is visualized into interactive web-based figures, and integrated with other data using statistical mining approaches such as correlation analyses, enrichment and network analyses. The publications, or other final products that result from the analyses are hosted on platforms that include PubMed, DataMed, and GEO. These repositories facilitate reuse and integration. Data, tools, and pipelines are hosted on the cloud.