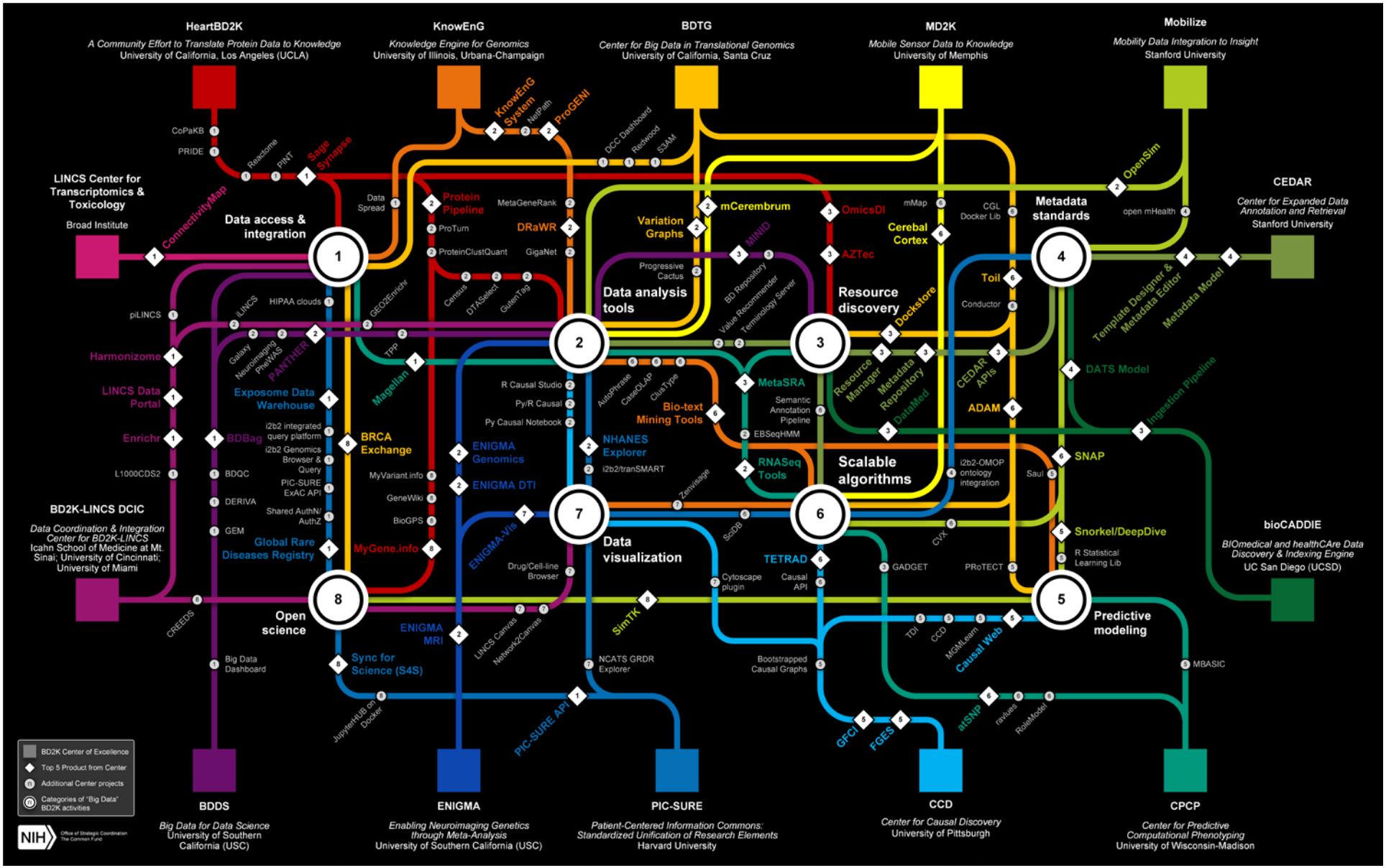
Table of Contents

BD2K Center	Product/Activity Name	Type
BDDS	Minimal Variables Identifier (Minids)	Open exchange format
	BDBag	Software tool
	Big Data Catalog	Software platform
	Data Discovery Dashboard	Software platform
	PANTHER	Software tools
BDTG	Toil	Platform
	Variation Graphs	Platform
	ADAM	Software algorithm/tools
	BRCA Exchange	Portal
	Dockstore	Platform/API standard
BD2K-LINCS DCIC	Harmonizome	Platform
	Enrichr	Platform
	Big Data MOOCs on Coursera	Training/educational materials
	CREEDS	Portal
	LINCS Data Portal and iLINCS	Platform
bioCADDIE	DataMed	Portal/platform
	DatA Tag Suite (DATS) Model	Open exchange format
	Annotated Corpus (Challenge materials)	Dataset
	Ingestion Pipeline	Software platform
	DCIP Pilot	Community engagement
CCD	Fast Greedy Equivalence Search (FGES)	Software algorithm
	Greedy Fast Causal Inference (GFCI)	Software algorithm
	TETRAD	Platform
	Causal Web	Platform
	Causal Modeling and Discovery Educational Materials	Training/educational materials
CEDAR	CEDAR Resource Manager	Platform
	CEDAR Template Designer and Metadata Editor	Software tools
	CEDAR REST APIs	Platform
	CEDAR Metadata Model	Open exchange format
	CEDAR Metadata Repository	Platform
CPCP	MetaSRA Metadata and Pipeline	Portal/software platform
	Single-cell RNASeq Analysis Tools	Software tool
	Magellan Entity Matching Tools	Software tool
	atSNP	Portal/platform
	CPCP Videos	Training/educational materials

BD2K Center	Product/Activity Name	Type
ENIGMA	ENIGMA MRI	Software platform
	ENIGMA DTI	Software tools
	ENIGMA Genomics	Software tools
	ENIGMA-Vis	Portal/platform
	ENIGMA Training	Training/educational materials
HeartBD2K	AZTec	Portal/platform
	OmicsDI	Portal/platform
	MyGene.info	Platform
	Sage Synapse	Portal/platform
	Protein Pipeline	Software tools
KnowEnG	KnowEnG: Knowledge Network Guided Analysis System	Platform
	ProGENI	Software algorithm
	DRaWR	Software algorithm
	ClustENG and TeachEng	Training/educational materials
	Bio-text Mining Suite	Software tools
MD2K	mCerebrum	Software platform/tools
	Cerebral Cortex	Software platform/tools
	MotionSenseHRV & EasySense	Physical device (sensor)
	mHealthHUB	Portal
	mHealth Summer Training Institute (mHTI)	Training
Mobilize Center	OpenSim	Software tools
	Snorkel/DeepDive	Software tools
	Stanford Network Analysis Platform (SNAP)	Software toosl
	SimTK	Platform
	Women in Data Science Conference (WiDS)	Training activity
PIC-SURE	Sync for Science (S4S)	Platform
	PIC-SURE (RESTful) API	Platform
	NHANES Database	Portal
	Exposome Data Warehouse	Portal
	Global Rare Diseases Registry	Portal

Putting the Products into Perspective: BD2K Centers' Software Tools, Platforms, and Standards



1. Minids

Minids provide a lightweight means of identifying distributed data and verifying integrity. They can be applied at any stage of the data lifecycle from raw to archived data. Minids support unambiguous data reference, and require only minimal metadata to be created. Each Minid captures a checksum that ensures that data integrity can be checked. An API supports a full range of Minid management operations (e.g., create, update and delete URIs and titles). To support BDBags, which have different checksums depending on whether or not the fetch file has been materialized, Minids support a content-based checksum. Minids can also be versioned. Here, users can assign a status to Minids (active, tombstoned, obsoleted) and in the case of a new version, they can associate a reference to the new version of the Minid.

Community. The community of users include BD2K investigators; Trans-Proteomic Pipeline (TPP); ENCODE and Panther communities; Peptide Atlas; TReNA; and BDDS pilot project users.

Usability assessment and evaluation. Minids are designed to be both lightweight and easy to use. As such, we have focused on developing a range of interfaces via which they can be used. We offer a web-based service for programmatic access (as used by the ENCODE2BDBag service), CLI access for command line usage, and a Python programming library for integration in external software. We have produced tutorial documentation to support users, and published documentation of the various interfaces. We have worked with collections of pilot users, in particular researchers from Peptide Atlas and also the BDDS and bioCADDIE Centers. In each case, we have met frequently with the users to derive requirements from their use cases, we have observed their usage of the service and adapted the service and tooling to meet their needs.

Discoverability. We have presented Minid in several conferences and the BD2K California meeting. We have also presented hands-on tutorials, with worked examples and exercises for users to explore the system. We have published a white paper and a conference paper (IEEE Big Data) that describe the Minid system and highlights it benefits to researchers. The BDDS website (bd2k.ini.usc.edu/tools/minid/) provides information about the tool as well as several domain-specific use cases that highlight its usage. The Minid software, clients, and APIs are documented in a publicly accessible GitHub repository.

Dissemination. Minids are an integral part of the BDDS platform and have been integrated with many of the BDDS services (Galaxy, LONI pipeline, BDDS catalog, etc.). Minids are now used for every release of the Peptide Atlas project. Minids and BDBags are also supported for new raw datasets from Peptide Atlas. Users can obtain the Minid software from the BDDS website and the public BDDS repository. It is also operated as a hosted service (operated by BDDS) via which any user can access it without requiring any software installation. Finally, we developed documentation and web-based training material to support the use of Minids. In the last year, 20 users have created 665 Minids. Minids have also been used to integrate BDDS analysis services with ENCODE. Here, the ENCODE service creates a BDBag with associated Minid that captures the exact provenance of an ENCODE query.

Future relationship to Commons. Minids provide an unambiguous reference to data that may be stored, analyzed, or shared in different contexts. By associating such an identifier with data, when it is first created, Commons services can more easily track data throughout complex lifecycles and between services.

2. BDBag

The BDBag software (bd2k.ini.usc.edu/tools/bdbag) allows researchers to address a significant challenge of assembling, identifying, and providing access to subsets of big data in a large and complex data collection workflow such as from a catalog search to an analysis pipeline and to a publication service. This collection of utilities simplify usage of BDBags and ensure that created BDBags adhere to the BDDS Bagit/RO (link to https://github.com/ResearchObject/bagit-ro) profiles. A unique aspect of this work is that the data that is aggregated need not be collocated: instead, data collections can be uniquely identified where large elements may be located in cloud or enterprise storage. This is critical for big data elements where the cost of transfer of the data can be prohibitive. Another important feature is the use of JSON-LD to provide a standard way for linking metadata with existing ontologies and vocabularies. As the first example use of JSON-LD metadata, a model has been developed for representing ontology-based file types.

Community. The community of users include BD2K

investigators; Trans-Proteomic Pipeline (TPP); ENCODE and Panther communities; Peptide Atlas; TReNA; and BDDS pilot project users.

Usability assessment and evaluation. BDBag was developed to address a need in the BDDS platform: the scalable and standard representation and exchange of complex datasets comprised of potentially large file-based data and metadata. As such, we developed the system in collaboration with computer scientists developing large scale analysis and cataloging infrastructure and also the scientists creating and analyzing large biomedical data. Our pilot usage period enabled us to work with these communities (including the KnowEng Center) to ensure that the model was flexible enough for their needs. We have since released the software publicly and have used it in production to exchange data between BDDS services. It has also been used for Peptide Atlas data releases and the representation of ENCODE queries for verifiable, point-in-time, analysis and reproducibility.

Discoverability. We have presented BDBag at multiple conferences, workshops, and hands-on tutorials. The latter has resulted in rich tutorial materials that can be used by others to learn about the system and also provide examples of how BDBags can be used in practice. We have published a paper in the IEEE Big Data conference that describes the BDBag software and highlights how it has been used in real application examples. Finally, the BDDS website includes information about BDBag, links to download BDBag tools, descriptions of integrations in other services and software as well as domain-specific use cases that highlight their value in real-world scenarios.

Dissemination. BDBags are used by the BDDS platform to enable exchange of data. It has also been used by the Peptide Atlas project for all data releases and as a model for retrieving data from ENCODE. Users can obtain the BDBag software from the BDDS website and the public BDDS repository. We have created regular releases of the software for public usage of the CLI, Python library, and GUI. To simplify usage of the software we have developed multiple interfaces that address different users' needs. BDBags can be created, managed, used via an integrated command line client that can be installed on a PC or server, they can be used programmatically via a Python client library, and they can be used in an Windows/MacOS native GUI client.

Future relationship to Commons. BDBags provide a standard model for exchanging data between services. They encode not only the data but also metadata in a structured way that can be easily understood by different services. They therefore support interoperability, provenance capture, and reproducibility. We have optimized BDBags to support very large amounts of data (for example by using Globus transfer) thereby enabling researchers to share what are essentially descriptors of their datasets from which others can materialize only the data needed. BDBags are likely to be an integral model for the Commons as it allows researchers and services to exchange data in a reliable and scalable manner throughout the data lifecycle (from creation to publication), while also ensuring that integrity of data can be maintained.

3. Big Data Catalog

The Big Data Catalog (DERIVA) provides a digital asset management system for scientific data that streamlines the acquisition, modeling, management and sharing of complex, big data. It also provides interfaces so that these data can be delivered to diverse external analytic tools.

Community. Users include BD2K investigators; the ENCODE communities; BDDS pilot project users, and the ABIDE Catalog (bd2kcat1.ini.usc.edu/abide/search).

Usability assessment and evaluation. DERIVA is composed of a suite of tools and services that are designed to reduce the overhead and complexity of creating and managing complex, big data sets. DERIVA is an infrastructure platform, and as such is not intended to work "out of the box" but rather can be easily configured to adapt

Discoverability. Details about DERIVA have been published in e-Science, 2016 IEEE 12th International Conference (doi: 10.1109/eScience.2016.7870883). The source code for DERIVA is available on GitHub and it is indexed via the BDDS web site. The GitHub site includes extensive documentation including well-defined web service APIs that promote interoperability.

Dissemination. DERIVA is a platform comprised of multiple modules that can be downloaded and integrated in-part or in-total from GitHub:

1. UI (Chaise, github.com/informatics-isi-edu/chaise).
2. Database (ERMrest, github.com/informatics-isi-edu/ermrest).
3. Object store (Hatrac, github.com/informatics-isi-

edu/hatrac).
4. Data transfer (IOBox, github.com/informatics-isi-edu/iobox; IOBox-Win32, github.com/informatics-isi-edu/iobox-win32; ioboxd, github.com/informatics-isi-edu/ioboxd).

DERIVA has been integrated into projects from numerous other scientific communities including: FaceBase www.facebase.org); GUDMAP (www.gudmap.org); and Rebuilding a Kidney (www.rebuildingakidney.org).

Future relationship to Commons. The DERIVA platform is part of the Software Services & Tools set of Commons components, specifically: user interfaces, APIs, and containers.

4. Discovery Data Dashboard

The Discovery Data Dashboard encompasses:

- A simulation algorithm for on-the-fly data sampling, which maintains data privacy without the need for encryption. The simulated population pool consists of approximately 300,000 individual "samples," with each sample representing 1,000 individuals. To integrate a new variable, a probability distribution is generated for each possible value, and each sample is assigned a value using random sampling based on the calculated probability distribution (e.g., a county with 3,000 males and 5,000 females would be represented as 8 samples, of which on average 3 will be male, and 5 will be female),
- Use of a non-relational database (MongoDB) provides horizontal scalability, making the app just as proficient at manipulating traditional datasets as it is at true "big data" datasets.

Data simulation and fusion reduces the overall data footprint of the web app, resulting in an application that is small enough to deploy onto a smartphone or other portable device.

Community. The Discovery Data Dashboard service has two types of users:

1. Non-technical users that are interested in conducting exploratory data analytics on predefined and aggregated datasets; and
2. Clinical, biomedical, and biosocial investigators that are engaged in data harmonization, fusion of multi-source data and tool development.

Usability assessment and evaluation. The Data Dashboard provides a mechanism for: 1) integrating dispersed multi-source data; and 2) servicing the mashed information via human and machine interfaces in a secure, scalable manner. The Dashboard enables the exploration of subtle associations between variables, population strata, or clusters of data elements, which may be opaque to standard independent inspection of the individual sources. This service is a device agnostic service for graphical querying, navigating and exploring the multivariate associations in complex heterogeneous datasets.

Discoverability. The Discovery Data Dashboard has been published (dx.doi.org/10.1186/s40537-015-00-18-z); is available as a web service (socr.umich.-edu/HTML5/Dashboard); and has an online tutorial (bd2k.ini.usc.edu/tools/big-data-dashboard). This software has been presented at a number of conferences and meeting presentations.

Dissemination. The software is available through GitHub (github.com/ini-bdds/Data-Dashboard), with 5-forks on Git. There are dozens of users employing the service daily.

Future relationship to Commons. The Dashboard has a couple of complementary relations to the NIH Commons Framework:

- Datasets. The Dashboard can ingest data from Commons and contribute datasets to Commons.
- Services. The Dashboard can provide data visualization services for display and interrogation of multi-source archives.

5. PANTHER

The PANTHER (Protein ANnotation Through Evolutionary Relationship) classification system comprehensive system that combines gene function, ontology, pathways and statistical analysis tools that enable biologists to analyze large-scale, genome-wide data from sequencing, proteomics or gene expression experiments (bd2k.ini.usc.edu/tools/panther).

Community. Bench biologists, researchers.

Usability assessment and evaluation. PANTHER is being integrated with the BDDS Platform Tools to support genotype data as inputs. A number of software modules are now available to take data from BDDS data repository or other BDDS tools, update the genome coordinates to those in the current genome, and update the file format. The data is then sent to PANTHER tools through the PANTHER web service. Because a standard file format (VCF) is used, the process

is seamless. The output of the analysis can be sent back to other LONI software or directly displayed on the PANTHER website.

Discoverability. The website contains detailed information about the web services. These services can easily be incorporated into new or existing workflows to perform statistical analysis.

Information about PANTHER is published in Nucleic Acids Research with each major release of the PANTHER data, tools, or features. Site usage statistics increase with each new publication. Top search results from Google for “overrepresentation test” or “gene enrichment analysis” brings up links to PANTHER.

Dissemination. PANTHER is part of the Gene Ontology (GO) Consortium. PANTHER web services are invoked from the GO website and users are forwarded to the PANTHER website for viewing results. PANTHER is also involved with the Model Organism Databases to provide support for annotating genes. It is relatively easy for researchers to build, download or purchase a statistics package. However, it is non-trivial to combine the latest gene set with the latest expert curated annotations and combine with a statistics library. PANTHER has an extensive workflow for mapping the Reference Proteome gene set (released yearly) to the GO annotations (updated monthly). Within the scientific community, the PANTHER tools and data collectively are becoming the de facto standard for statistical analysis. This is reflected in the website usage statistics (~2,000 sessions daily) collected via Google Analytics. New functionality and updates are done based on user feedback and usage statistics.

Future relationship to Commons. PANTHER web services can be easily incorporated into workflows from the Commons.

1. Toil

Toil is an open-source pure-Python workflow engine (github.com/BD2KGenomics/toil) that lets people write better pipelines. You can:

- Write your workflows in Common Workflow Language (CWL);
- Run workflows on your laptop or on huge commercial clouds such as Amazon Web Services (including the spot market), Microsoft Azure, OpenStack, and Google Compute Engine;
- Take advantage of high-performance computing environments with batch systems like GridEngine , Apache Mesos, and Parasol;
- Run workflows concurrently at scale using hundreds of nodes and thousands of cores;
- Execute workflows efficiently with caching and resource requirement specifications; and
- Easily link databases and services.

Community. Toil is an open source project, downloaded thousands of times, with more than 300 GitHub stars and averaging more than 500 individual visitors to the site every two weeks; and >30 individual contributors from more than a dozen organizations.

Usability assessment and evaluation. Toil is easily installed and runnable across cloud and HPC environments, and being written in pure Python is accessible to a broad range of developers. At time of inception no open source project existed to run modern, open source scientific workflow standards, such as CWL, at scale within cloud environments. Toil fulfills this need.

Discoverability. The project is open source on GitHub, the documentation is also open source and freely available at toil.readthedocs.io. The documentation pages are hit thousands of times per month. We have also had a publication accepted in Nature Biotechnology describing Toil and coming out in April 2017. This manuscript shows how Toil can be used to analyze massive biomedical datasets on the cloud – in the paper 20,000 omics samples in 4 days on AWS using an elastic cluster of 32,000 cores.

Dissemination. Through publication, through conference presentations, through social media and by demonstration projects we have raised awareness of the project. We are also working with Global Alliance for Genomics and Health (GA4GH) to ensure Toil conforms to emerging standards in the domain, making it more useful to a broader ecosystem.

Future relationship to Commons. Enabling scalable, portable, reproducible scientific analyses is a core aim of the Commons. Toil facilitates this by providing software and tools to author and run these workflows across different clouds and HPC environments in a manner that is efficient and precisely reproducible while being fault tolerant and robust. By conforming to GA4GH container and workflow standards Toil will interoperate with other services in the commons.

2. Variation Graphs

Reference genomes provide a prior to guide our interpretation of DNA sequence data. However, conventional linear references are fundamentally limited in that they represent only one version of each locus, whereas the population may contain multiple variants. If the reference represents an individual's genome poorly, it can impact read mapping and introduce bias. Genome graphs are DNA sequence graphs that compactly represent local genetic variation. Variation Graphs (vg; github.com/vgteam/vg) is a software toolkit of computational methods for creating, manipulating, and utilizing these structures as references at the scale of the human genome.

vg provides an efficient approach to mapping reads onto variation graphs using generalized compressed suffix arrays, with improved accuracy over alignment to a linear reference, and provides data structures to support downstream variant calling and genotyping. These capabilities make using variation graphs as reference structures for DNA sequencing practical at the scale of vertebrate genomes.

Community. vg is an open source project, which has been downloaded thousands of times, receiving more than 200 GitHub stars and with >20 individual contributors from multiple organizations.

Usability assessment and evaluation. The software provides a comprehensive software toolkit for developing and testing ideas using genome graphs. Increasingly the *de facto* standard in the space. A small, focused, community of experts have developed the project together.

Discoverability. A preprint (biorxiv.org/content/early/-2017/01/18/101378) has been made, with the publication undergoing review at Nature Biotechnology. Despite being released for only a month, the preprint is in the top 120 preprints of all time on bioxriv (of 9000 – 98th percentile). A review article we wrote describing

progress in the field, and heavily citing vg, has been accepted in a special issue of Genome Research. Numerous other publications building on vg have been published.

Dissemination. All code and documentation is open source and available through GitHub. We have organized workshops on Pangenomics, in particular one last year at ECCB that drew considerable interest to the project.

Future relationship to Commons. As genome graphs become the standard way to model sequence variation, we believe vg will become widely used within genomics, and thus of utility within the Commons.

3. ADAM

In the past year, we have extended the ADAM software (github.com/bigdatagenomics/adam) to support new analyses, while pushing the software towards production readiness. A new area of emphasis for us this year has been machine learning and statistical analysis of genomic data. We added the Gnocchi project, which uses the ADAM APIs to analyze genotype data and to perform genotype-phenotype association tests, and the Endive project, which uses the ADAM APIs to perform machine learning on functional genomics datasets. Additionally, we extended ADAM's APIs for storing variant and genotype data, to bring our API more in line with the GA4GH schemas and APIs. Using these APIs, we submitted ADAM-based queries to the Mayo Institute's VariantDB challenge.

In the past year, we broadened the ADAM developer community by increasing the number of developers who had contributed to ADAM by 20% from 50 to 62, published three new releases of ADAM, and added support for the GA4GH reads API, which ADAM exposes as an endpoint in the Mango project. In addition to the new work highlighted above, we have begun new collaborations with the National Marrow Donor Program (NMDP)/Be The Match and Janssen Pharmaceuticals. The NMDP collaboration is using ADAM to evaluate host-versus-graft disease data from a set of 250 matched donor-recipient WGS pairs. In the Janssen collaboration, ADAM is being used as an exploratory data analysis tool for annotated variant and genotype calls.

Community. Genomics programmers and tool developers, researchers performing variant analysis.

Usability assessment and evaluation. ADAM provides a series of APIs, file formats, and utilities for processing common genomic data types. It leverages the SPARK compute environment for high performance. Success can be measured by the active GitHub user community, ADAM has been forked over 200 times and starred by over 600 users.

Discoverability. GitHub provides the primary interface for developers to use and extend the ADAM framework. It is findable software and accessible through it's open source nature and publication history.

Dissemination. The project has been engaged with collaborations offering proof of concept utilization of ADAM in real-world scenarios. For example, a collaboration to use ADAM with the National Marrow Donor Program (NMDP)/Be the Match and Janssen Pharmaceuticals. Over time we expect the rich tool chain of ADAM to offer increased performance on single-machine tool implementations such as GATK 3.x, Samtools, and Picard.

Future relationship to Commons. ADAM defines several useful file formats for storing genomic data more efficiently and this can be leveraged in cloud environments. Furthermore, integration with Toil shows that ADAM can be used in high-performance, cloud-based workflows leveraged by Commons efforts.

4. BRCA Exchange

The BRCA Exchange (brcaexchange.org) aims to advance our understanding of the genetic basis of breast cancer, ovarian cancer and other diseases by pooling data on BRCA1/2 genetic variants and corresponding clinical data from around the world. It is the world's single largest public repository of BRCA variants and hosts gold standard community curations by the ENIGMA consortium. The project was developed using BD2K funds and is now being expanded to provide a generic platform for disease gene communities to share data.

Community. Front end and technical products of the global BRCA Challenge project (genomicsand-health.-org/work-products-demonstration-projects/-brca-challenge-0). BRCA Challenge is a global community of individuals interested in understanding BRCA variation and its relationship to cancer. BRCA Exchange has a community forum for these members with more than 200 members, see: brcaexch-ange.org/community.

Usability assessment and evaluation. The website

provides two portals, an end-user portal aimed at clinicians, genetic counsellors and individuals to provide them up-to-date, gold standard information about more than 18,000 genetic variants. This site was developed in consultation with experts in the community to be approachable. In addition, a research portal provides a deep view of the variant data for the purposes of enabling variant curation.

Discoverability. The site was announced at the World Congress in Genetics at Kyoto in April 2016 by world experts in BRCA Biology. We have subsequently presented the project at numerous scientific meetings.

Dissemination. The site is fully public, all the underlying data is provided for download and analysis. All the underlying code for the site is open source. BRCA Exchange is now the world's largest public, single source of curated BRCA variants, we are now working to export these variants back into other genetic variant databases, including ClinVar and LOVD.

Future relationship to Commons. BRCA Exchange is a web-portal that, being fully open source, can be developed for the sharing of genetic variants within other disease genes. For example, the Canadian Open Genetics Repository (COGR) (opengenetics.ca) is developing the platform for sharing variants within its network. In this way it could provide a platform for many individual communities in a distributed commons.

5. Dockstore

The Dockstore project (dockstore.org) was started in 2015 as a response to the lack of standards present in the community for exchanging Docker-based tools and workflows. It was created at OICR and currently is a joint project between OICR and the UCSC Genomics Institute. The major innovation of Dockstore is its use of WDL or CWL to describe, in a standardized, machine- and human-readable way, how to invoke Docker-based tools – and therefore provide the vital metadata that must be present for scientific usage. Dockstore acts as a public repository for these tools as well as workflows that link them together. For researchers releasing their tools and workflows on Dockstore, we provide convenient methods to link to source repositories on GitHub and Bitbucket an Docker-build services on DockerHub and Quay.io. By using highly popular and well regarded services, Dockstore ensures the site fits in well with existing developer flows. Likewise, for end users, Dockstore provides a compelling interface and consistent documentation, making it much easier to find and use portable scientific tools.

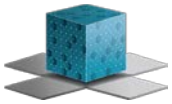
Community. Any researcher making genomics tools and workflows and wanting to share them with the community. The collection of users in the community that want to use high-quality, portable genomics tools.

Usability assessment and evaluation. This product includes a friendly web-based user interface. It walks developers through sharing their Docker-based tools and workflows. It also walks users through finding and using workflows. We evaluate its functionality through the uptake in the community both for developers and users. We also run training workshops and get feedback on usability in this manner.

Discoverability. The Dockstore makes all tools and workflows findable, all items are publically accessible, tools and workflows interoperate on different platforms, and the tools are reusable by others in the community. For these reasons we feel it supports the FAIR principles.

Dissemination. The Dockstore uses, and supports, the development of GA4GH tool registry standards. The source for Dockstore is also available and we encourage others to setup their own instances of Dockstore. We are using Dockstore for HCA and other large scale projects. It was also the platform used for the GA4GH Tool Execution challenge which showed that 30 individuals and organizations could run the same tool and get the same result.

Future relationship to Commons. Dockstore, and the underlying GA4GH Tool Registry Service API standard, are key components of the Commons. They represent and implementation and a standard respectively of how tools and workflows can be shared in the community. Dockstore directly implements the FAIR principles.



1. Harmonizome

The Harmonizome is a collection of processed datasets gathered to serve and mine knowledge about genes and proteins from over 70 major online resources. We extracted, abstracted and organized data into ~72 million functional associations between genes/proteins and their attributes. Such attributes could be physical relationships with other biomolecules, expression in cell lines and tissues, genetic associations with knockout mouse or human phenotypes, or changes in expression after drug treatment. We stored these associations in a relational database along with rich metadata for the genes/proteins, their attributes and the original resources. The freely available Harmonizome web portal provides a graphical user interface, a web service and a mobile app for querying, browsing and downloading all of the collected data.

Community. The community user base is diverse, including biomedical researchers, physicians, high school, undergraduate and graduate students, patients; as well as individuals from biotech and pharma.

Usability assessment and evaluation. As this product has a large user base, we receive constant feedback from users who ask for new features and report bugs. We also analyze user input to see how users interact with the system. We save user queries and perform analysis on such queries to identify trends and heavy users.

Discoverability. We published an article describing the system in the journal Database in 2016. We use social media to report any software updates. We describe the project in two of the MOOCs we teach on Coursera. Most users are acquired by Google search engine searches. Users that are searching Google for terms that describe cell lines, genes, small molecules, gene pages are routed to the Harmonizome resource through returned suggested Harmonizome pages as top search results. API and metadata are well documented.

Dissemination. The product is delivered as a standalone web server and a database. The source code, processing scripts, and processed datasets are available on GitHub. The entire system is dockerized. So far the resource acquired: 145,525 sessions; 104,532 unique users; 435,880 page views; based on Google Analytics as of 3/27/2017.

Future relationship to Commons. The Harmonizome epitomizes the concept of the Commons as it integrates many of the data produced by the NIH Common Fund projects: LINCS, GTEx, ENCODE, Roadmap Epigenomics, CTD2, IDG and more.

2. Enrichr

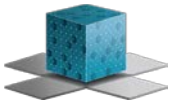
Enrichr is a gene set enrichment analysis tool that currently contains a large collection of diverse gene set libraries available for analysis and download. In total, Enrichr currently contains 180,184 annotated gene sets from 102 gene set libraries. Features of Enrichr include the ability to submit fuzzy sets, upload BED files, API and visualization of the results as clustergrams. Overall, Enrichr is a comprehensive resource for curated gene sets and a search engine that accumulates biological knowledge for further biological discoveries.

Community. The community user base is diverse, including biomedical researchers, physicians, high school, undergraduate and graduate students, patients; as well as individuals from biotech and pharma.

Usability assessment and evaluation. As Enrichr has a large user base, we receive constant feedback from users who ask for new features and report bugs. We also analyze user input to see how users interact with the system. We save user queries and perform analysis on such queries to identify trends and heavy users.

Discoverability. We published two articles describing the system including a recent update article in the journal Nucleic Acid Research in 2016. We use social media to report any software updates. We describe the project in two of the MOOCs we teach on Coursera. Most users are acquired by word of mouth and most users are returning users (75%). The Enrichr software is now well-known and considered one of the leading enrichment analysis tools. So far the articles describing Enrichr were cited more than 400 times with citation doubling every year. API and metadata are well documented.

Dissemination. The product is delivered as a standalone web server and a database. The source code is available on GitHub and gene set libraries can be downloaded from the Enrichr site. The entire system is Dockerized. So far, the resource acquired: 220,076 sessions; 57,510 unique users; 965,275 page views; and 5,259,762 submitted queries. These statistics are based on Google Analytics as of 3/27/2017.



Future relationship to Commons. Enrichr is an important tool for the Commons because it is a search engine for biologists who conduct genomics, transcriptomics and proteomics studies. It enables querying results from new experiments in the context of prior knowledge.

3. Big Data MOOCs on Coursera

Two MOOCs, called Network Analysis in Systems Biology and Big Data Science with the BD2K-LINCS Data Coordination and Integration Center, provide an introduction to big data analysis in biology, including statistical methods used to identify differentially expressed genes, performing various types of enrichment analyses, and applying clustering algorithms. Students learn how to construct, analyze and visualize functional association networks that can be created from many resources, including gene regulatory networks connecting transcription factors to their target genes, protein-protein interaction networks, cell signaling pathways and networks, drug-target and drug-drug similarity networks and other functional association networks. Methods to process raw data from genome-wide mRNA expression (microarrays and RNASeq) are presented. Processed data is clustered, and gene-set enrichment analyses methods are covered.

Community. Coursera students are coming from all over the world and from diverse backgrounds. The range of students that took our courses are faculty members, postdocs, grad students, undergrads, and professionals.

Usability assessment and evaluation. Coursera provides detailed analytics for each course. These analytics include time spent on each lecture, grades, retention, and user feedback. This feedback has helped us to adjust the syllabi and increase interactivity.

Discoverability and dissemination. The courses are advertised by Coursera, who attracted so far millions of students from all over the world. Our courses are widely attended. We estimate that so far over 45,000 students took at least one of the two courses, while over 100,000 visited the courses at least once. All lectures slides and videos are available for download and the videos are distributed through YouTube besides their availability on Coursera.

4. CREEDS

CREEDS is a citizen-science crowdsourcing project. Through a massive open online course on Coursera, over 70 participants from over 25 countries identify and annotate 2,460 single-gene perturbation signatures, 839 disease versus normal signatures, and 906 drug perturbation signatures. All these signatures are unique and are manually validated for quality. Global analysis of these signatures confirms known associations and identifies novel associations between genes, diseases and drugs. The manually curated signatures are used as a training set to develop classifiers for extracting similar signatures from the entire GEO repository.

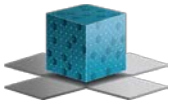
Community. The project engaged ~100 students from the Coursera courses. These students who were from over 25 countries, worked together on a crowdsourcing project that produced the CREEDS resource. The CREEDS resource is used by the entire biomedical research community.

Usability assessment and evaluation. Some of the students that participated in the project were recruited to our Center. Some reported that they mentioned the project in their interviews for entering a new job or an academic program. It is too early to assess usability of the CREEDS resource as it was published in the summer of 2016.

Discoverability and dissemination. The project was published in the journal Nature Communications in 2016. There was a press release about the paper, and it was picked up by several news outlets. The recruitment of the students was done through mass communication via Coursera. The CREEDS site is delivered as a standalone web server and a database. The source code is available on GitHub and gene set libraries can be downloaded from the CREEDS site with their metadata. The entire system is Dockerized. Using machine learning and text mining, we extracted automatically an additional 14,000 signatures that we made available.

5. LINCS Data Portal and iLINCS

Integrative LINCS (iLINCS) data and signatures analysis portal facilitates online user analytics of LINCS data and signatures. iLINCS brings together data, computational tools, and user interfaces to be used by biomedical scientists with only conceptual understanding of analysis protocols. iLINCS is meant to utilize and complement tools designed for specific tasks by organizing them into easily executable workflows. The



LINCS Data Portal (LDP) is a multi-tier, web-based application intended to present a unified interface to access LINCS datasets and metadata with mappings to several external resources. LDP provides various options to explore, query, and download LINCS data that have been described using the LINCS metadata standards. LDP enables download of data packages, consisting of datasets and associated metadata.

Community. The community user base is diverse, including biomedical researchers, physicians, high school, undergraduate and graduate students, patients; as well as individuals from biotech and pharma.

Usability assessment and evaluation. These resources are developed in tight collaboration with the LINCS Data and Signature Generation Centers (DSGCs). For the LINCS Data Portal, the DSGCs provide metadata and SOPs and ensure that the data is well presented. For the iLINCS portal the DSGCs provide and share data processing scripts and feedback on workflows.

Discoverability. These tools and resources are presented in webinars, workshops, symposiums and invited talks. We use social media to update the community about the release of new features. Publications of these resources are in preparation. All data in the LINCS Data Portal is served through a web interface and an API. iLINCS also have APIs that are well documented.

Dissemination. iLINCS is provided through R and R packages. All code is made available on GitHub. The tools usage is tracked through Google Analytics. Tracking was enabled only one year ago. The iLINCS portal has accumulated 8,578 unique users; 19,115 sessions, and 172,730 page views as of 3/28/2017 based on Google Analytics. The LINCS Data Portal attracted a similar level of traffic. LDP has been used as the template for developing the CEDAR metadata forms.

Future relationship to Commons. LDP and iLINCS provide an example of how to serve diverse datasets with deep metadata and readily available analysis tools coupled with the data.

1. DataMed

DataMed is a data discovery index (DDI) prototype developed by bioCADDIE to provide a platform for users to find data distributed in multiple repositories and other sources. DataMed incorporates the DATS metadata model to unify and map metadata for indexing that is utilized by the search engine. Various enhancement modules such as publication, terminology and an NLP-server are employed to provide an efficient search engine and intuitive user interface (datamed.org). The search by CDE is a relatively new collaboration with NIH.

Community. Biomedical researchers, granting agencies (NIH) via feedback mechanism in DataMed, OMICS DI engagement. The BD2K-LINCS team has downloaded and make extensions to the code as well.

Usability assessment and evaluation. bioCADDIE involved a specific Usability working group (WG) (WG9) for summative and formative evaluation. Formal user needs analysis were conducted prior to initiation of the development. User testing has been done after v0.5 and is currently ongoing for v2.0 to inform future development. We also collect feedback via email, online "contact us" forms, and GitHub.

Discoverability. A recent addition to the UCSD team is helping to market the tool via social media. Searchathons (FORCE11, NIH) have also been held. Additional routes for making the scientific community aware of DataMed include:

- A publication in Nature Genetics (in press; preprint available in bioRxiv); and
- Schema.org markup for discoverability by search engines and search engine optimizations have been done by adding keywords, creating a sitemap, etc. (to increase traffic to DataMed).

Data discovery will be available to all users of the NLM CDE Repository and other CDE registries via DataMed being built into those services.

Dissemination. Efforts have included conference presentations (Health Datapalooza 2016, FORCE16, Elixir All Hands, AMIA 2016, AMIA 2017 (submitted)); engagement with the NIH/NLM CDE repository; and an initial presentation aimed to NIH CDE taskforce.

Future relationship to Commons. DataMed has indexed the following datasets which are intended to be part of the NIH Commons: Human Microbiome Project; Rat Genome Database; and Genomic Data Commons. Linking data in the Commons (via DataMED) to CDEs in the NLM Repository will bridge two NIH initiatives.

2. DatA Tag Suite (DATS) Model

We have designed and implemented the DAta Tag Suite (DATS) model to support the DataMed data discovery index. Akin to the Journal Article Tag Suite (JATS) used in PubMed, the DATS model enables submission of metadata on datasets to DataMed. DATS has a core set of elements, which are generic and applicable to any type of dataset, and an extended set that can accommodate more specialized data types. DATS is a platform-independent model also available as an annotated serialization in schema.org, which in turn is widely used by major search engines like Google, Microsoft, Yahoo, and Yandex.

Community. DATS' intended audience is threefold: 1) the DataMed development team that will implement and test the model; 2) prospective data sources that wish to be indexed in the DataMed prototype; and 3) developers of data harvesting and other metadata tools and catalogs. Especially notable are our successful collaborations with other data aggregators and service providers that are working to implement DATS. These include, but are not limited to:

- The Inter-university Consortium for Political and Social Research (ICPSR), the world's largest archive of digital social science data and one of the NIH-supported repositories (www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html);
- The NIH BD2K OmicsDI15, a data discovery index for proteomics, genomics and metabolomics datasets;
- The NIH Federal Interagency Traumatic Brain Injury Research (FITBIR; fitbir.nih.gov), an informatics system to share data across the entire TBI research field;
- ImmPort (www.immport.org), an informatics system supporting the NIH mission to share data, focused on the immunology data; and
- DataCite (www.datacite.org), a global non-profit organization that supports the creation and allocation of digital object identifiers (DOIs) for research data and accompanying metadata.

Another example of use is provided by the NIH BD2K CEDAR metadata authoring tool that works to provide a DATS-compliant template to help researchers to describe and expose their datasets, which are not yet in public repositories, to indexing in DataMed.

Usability assessment and evaluation. DATS was developed collaboratively by a WG of experts, and based on the use cases collected by NIH, Force11, NIH researchers and a large scientific community (biocaddie.org/group/working-group/-working-group-3-descriptive-metadata-datasets). DATS is currently implemented by the bioCADDIE team in DataMed.

Discoverability. Efforts to make the broader community aware of DATS have included: 1) a publication in Nature Scientific Data (in press; preprint at: bioRxiv 103143; doi: doi.org/10.1101/103143); specification development, tracked in GitHub (github.com/biocaddie/WG3-MetadataSpecifications); making specification citable via Zenodo (zenodo.org/communities/-dats/?page=1&size=20); and annotation with schema.org and bioschemas.org to enhance discoverability.

Dissemination. DATS is being shared via the ELIXIR-endorsed BioSharing portal (biosharing.org/collection/bioCADDIE), through various workshops, and through direct contact with repositories.

3. Annotated Corpus

The benchmark dataset for information retrieval of biomedical datasets was developed and used for the 2016 bioCADDIE Dataset Retrieval Challenge. This benchmark includes both a corpus (biomedical datasets), a set of queries, and relevance judgments relating these queries to elements of the corpus. We used a collection of metadata (structured and unstructured) from biomedical datasets generated from a set of 20 individual repositories. A total of 794,992 datasets were made available for use from the set of indices that was frozen from the DataMed backend on 3/24/2016. This set of records is available as both XML and JSON files. The resulting benchmark set has been made publicly available to advance research in biomedical dataset retrieval.

Community. This dataset is intended for informaticists focusing on information retrieval (IR) research. Thus far, 10 Challenge respondents were able to use this dataset successfully.

Discoverability and dissemination. A paper describing the annotated corpus is being published as a part of a special issue in Database (Journal). The dataset is also published online at biocaddie.org/benchmark-data. The dataset is freely available for community use.

4. Ingestion Pipeline

The bioCADDIE prototype (datamed.biocaddie.org) is backed by a scalable dataset indexing infrastructure. This infrastructure maps the disparate metadata from the diverse data sources into a unified specification provided by various working groups organized by bioCADDIE and related communities. This pipeline involves an automated component that provides controlled translation and curation of metadata using special tools such as a transformation language and JSON-Path. The overall infrastructure consists of the following components:

1. A data and metadata extraction system that is able to connect to various repositories and data aggregators. All metadata information is converted to JSON documents for each dataset being described and stored in MongoDB.
2. A messaging infrastructure, utilizing Apache ActiveMQ, distributes dataset description documents from MongoDB, and depending on their status value, dispatches them to persistent point-to-point queues.
3. A collection of multiple concurrent consumers retrieve the documents from MongoDB, process it, update the job status and save it back to the MongoDB. Consumers can be written using the STOMP protocol. Documents are transformed, to align with the bioCADDIE metadata model, and processed by a collection of modules that enhance the metadata records.
4. Fully processed documents are then exported to an ElasticSearch endpoint that serves the dataset indices via standard ElasticSearch RESTful services.

Community. Informaticists involved with DataMed development.

Usability assessment and evaluation. Testing of deployment and usability has been conducted by a second group from the bioCADDIE CDT to get the pipeline up and running. This activity has been extended with modules developed by multiple CDT groups.

Discoverability and dissemination. A publication describing this framework is in preparation. A number of recent conference presentations with DataMed have also covered this framework (Health Datapalooza 2016, FORCE16, ELIXIR All Hands) and individually (AMIA, BD2K).

Future relationship to Commons. The ingestion

pipeline is used in the indexing of content for DataMed and towards the Commons (see above).

5. DCIP Pilot

bioCADDIE, through a collaboration with FORCE11, jointly established work on implementing data citation as a community effort through the Data Citation Implementation Pilot (DCIP), to drive adoption of data citation standards and to standardize citation metadata and literature connections. A key roadblock to data citation was identified in the traditional use of locally assigned database accession numbers. Globally resolvable identifiers, as outlined by bioCADDIE's working group on identifiers, is a requirement for development of upper-level tools in the ecosystem, such as DataMed, which require the use of software agents operating across the Web. The DCIP is structured as a series of "Expert Groups" developing detailed roadmaps, specifications and training materials for publishers, repositories, and identifier and metadata service providers. It is also exciting to note that Elsevier has announced implementation of data citation across 1,800 of its journals (www.elsevier.com/about/-press-releases/science-and-technology/elsevier-implements-data-citation-standards-to-encourage-authors-to-share-research-data).

Community. The working and expert groups involved in this effort have included representatives from many communities: publishers, libraries, informaticists, biomedical researchers, data repositories, and software vendors. Outreach has also been to the broader community.

Usability assessment and evaluation. Elsevier and Nature have committed to adopting the Data Citation principles. They are in the process of adopting them.

Discoverability and dissemination. Publisher and Repository Roadmaps to Data Citation, along with a draft specification for converging compact identifier resolvers on a common namespace prefix registry, have been released as preprints (Fenner et al. 2016, bioRxiv 097196; doi: doi.org/10.1101/097196; Cousijn et al. 2017 bioRxiv 100784; doi: doi.org/10.1101/100784; Wimalaratne et al. 2017 bioRxiv 101279; doi: doi.org/-10.1101/101279). Broad stakeholder and community involvement in the bioCADDIE working group has occurred with the DCIP pilot.

Future relationship to Commons. The DCIP Pilot builds on work from bioCADDIE WG2, which recommended: "*The identifier recommendations contained within this document are focused on the need for the DDI to uniquely identify data sets within biomedical data repositories and the NIH Commons.*"

1. Fast Greedy Equivalence Search

Fast Greedy Equivalence Search (FGES) is an algorithm that takes as input a biomedical dataset of samples on a given set of variables, as well as optional prior biological knowledge. It then searches over a large number of possible causal Bayesian network (CBN) structures (models) and outputs the most probable model that it finds according to a Bayesian scoring measure. The model it returns is a data-supported hypothesis about causal relationships that exist among the variables in the dataset. The model serves to help biomedical scientists form hypotheses and guide their design of controlled experiments to investigate those hypotheses.

Community. Users of this algorithm include biomedical scientists and data scientists.

Usability assessment and evaluation. We have extensively optimized and parallelized FGES to enable the algorithm to handle datasets with a large number of variables, which now commonly occur in biomedical research. We tested it extensively using both simulated datasets and real biomedical datasets. As an example, in simulation experiments involving datasets with 100,000 continuous variables and 200,000 causal relationships, FGES learned a model in 17.4 minutes using 120 processors; it found 87% of the causal relationships that existed (recall) and 99% of the relationships that it labeled as causal were correct (precision). To our knowledge, the speed of FGES exceeds that of any other available causal Bayesian network (CBN) learning algorithm. We are continually refining the usability of FGES and other CCD software based on feedback obtained from users during our Causal Discovery Summer Short Course (including a survey with questions on software usability) and via feedback from our Helpdesk call-in service.

Discoverability. We announce its availability and updates via Reddit, Google+, and other email lists. We also describe and demonstrate its use at our annual week-long Causal Discovery Summer Short Course (which has ~70 faculty, postdoc, student, and industry participants each year) and our day-long Causal Discovery Datathon, both of which are held at Carnegie Mellon University. Attendees of both events can use FGES with their own data or supplied datasets for hands-on, real-world learning in a supportive environment. Moreover, we discuss its availability at conferences at which we make research presentations and at workshops in which we give tutorials on causal modeling and discovery (e.g., the 2016 American Medical Informatics Association (AMIA) Joints Summits on Translational Science meeting).

Dissemination. We make FGES available as a well-documented API on several platforms, including Java, Python, and R. It also is available within the Tetrad desktop and the Causal Web application (see these separate products). The source code is available for free download via GitHub (github.com/cmu-phil/tetrad) and by way of our Center website (www.ccd.pitt.edu/tools), and the resources page at the BD2K-CCC website (bd2kccc.org/bd2k-resources/). User documentation is available via our CCD Wiki (www.ccd.pitt.edu/wiki).

Future relationship to Commons. We plan for FGES to be available via the Commons and discoverable as software to analyze data for causal relationships. In our ongoing Commons pilot supplement project, we have developed a proof-of-principle integrated data ecosystem that serves as a prototype for sharing and analyzing biomedical big data in a secure and scalable manner in a cloud environment. FGES is an algorithm that is available in that prototype. The prototype currently runs on the Amazon cloud and supports authen-tication using institutional single sign-on. We have assigned its digital objects unique identifiers, implemented software objects within Docker containers, and used the Commons Credit model to support computation.

2. Greedy Fast Causal Inference

The Greedy Fast Causal Inference (GFCI) algorithm is similar in scope and purpose to the FGES algorithm (see Product 1, above), with one important difference. Unlike FGES, the GFCI algorithm models the possibility of unmeasured (hidden or latent) confounder variables of the measured variables. As biomedical data often contain unmeasured confounders, the ability to model and detect their presence and absence is often important.

Community. Users of this algorithm include biomedical scientists and data scientists.

Usability assessment and evaluation. We have optimized and parallelized GFCI to enable the algorithm

to handle datasets with a large number of variables, which now commonly occur in biomedical research. We tested it extensively using both simulated datasets and real biomedical datasets. As an example, in simulation experiments involving datasets with 1,000 continuous variables and 2,000 causal relationships, GFCI learned a model in 15 seconds using only 1 processor; it found 93% of the causal relationships that existed (recall), and 92% of the relationships that it labeled as causal were correct (precision). We are continually refining the usability of GFCI and other CCD software based on feedback obtained from users during our Causal Discovery Summer Short Course (including a survey with questions on software usability) and via feedback from our Helpdesk call-in service.

Discoverability. We announce its availability and updates via Reddit, Google+, and other email lists. We also describe and demonstrate its use at our annual week-long Causal Discovery Summer Short Course (which have ~70 faculty, postdoc, student, and industry participants each year) and our day-long Causal Discovery Datathon, both of which are held at Carnegie Mellon University. Attendees of both events can use GFCI with their own data or supplied datasets for hands-on, real-world learning in a supportive environment. Moreover, we discuss its availability at conferences at which we make research presentations and at workshops in which we give tutorials on causal mod-eling and discovery (e.g., the 2016 American Medical Informatics Association (AMIA) Joints Summits on Translational Science meeting).

Dissemination. We make GFCI available as a well-documented API on several platforms, including Java, Python, and R. It also is available within the Tetrad desktop and the Causal Web application (see these separate products). The source code is available for free download via GitHub (github.com/cmu-phil/tetrad) and by way of our Center website (www.ccd.pitt.edu/tools), and the resources page at the BD2K-CCC website (bd2kccc.org/bd2k-resources/). User documentation is available via our CCD Wiki (www.ccd.pitt.edu/wiki).

Future relationship to Commons. We plan for GFCI to be available via the Commons and discoverable as software to analyze data for causal relationships that may include unmeasured confounders. In our Commons pilot supplement project, we have developed a proof-of-principle integrated data ecosystem that serves as a prototype for sharing and analyzing biomedical big data in a secure and scalable manner in a cloud environment. GFCI is an algorithm that is available in that prototype.

3. TETRAD

TETRAD is a desktop system with a graphical user interface that provides an extensive set of causal modeling and discovery operations. A biomedical researcher can define prior biological knowledge, read in a large biomedical dataset, apply one of over 20 causal discovery algorithms, and display the resulting causal Bayesian network (CBN) on the desktop monitor. The user can also browse the CBN graphically and append notes that describe current insights and interpretations. The algorithms available in TETRAD include FGES (see Product 1, above) and GFCI (see Product 2, above), as well as algorithms that model and discover cyclic CBNs (representing feedback relationships) and causal relationships among latent variables.

Community. Users of this platform include biomedical scientists and data scientists.

Usability assessment and evaluation. TETRAD is a desktop system with an easy-to-use point-and-click graphical user interface and built-in user documentation and help features. We are continually refining the usability of TETRAD and other CCD software based on feedback obtained from users during our Causal Discovery Summer Short Course (including a survey with questions on software usability) and via feedback from our Helpdesk call-in service.

Discoverability. We announce its availability and updates via Reddit, Google+, and other email lists. We also describe and demonstrate its use at our annual week-long Causal Discovery Summer Short Course and our day-long Causal Discovery Datathon, both of which are held at Carnegie Mellon University. Attendees of both events can use TETRAD with their own data or supplied datasets for hands-on, real-world learning in a supportive environment. Moreover, we discuss its availability at conferences at which we make research presentations and

at workshops in which we give tutorials on causal mod-eling and discovery.

Dissemination. TETRAD is written in Java and therefore can run on most desktop machines, including PCs, Macs, and Linux computers. The source code is available for free download via Github (github.com/cmu-phil/tetrad) and by way of our Center website (www.ccd.pitt.edu/tools), and the resources page at the BD2K-CCC website (bd2kccc.org/bd2k-resources/). We estimate that the TETRAD system is being applied over 500 times per month by over 200 unique users.

Future relationship to Commons. We plan for TETRAD to be available via the Commons and discoverable as software to analyze data for causal relationships.

4. Causal Web

Causal Web is a web application that we developed so that biomedical researchers can easily perform causal discovery analyses on large biomedical datasets using just a web browser. Otherwise, such analyses would typically require direct access to a high performance computing (HPC) cluster with a large number of CPUs and large memory, as well as the skill to execute jobs via a command line interface. The Causal Web application avoids such requirements and removes significant barriers to applying causal discovery tools to large datasets. It has a wizard-style user interface that walks the user through the process of selecting an algorithm, choosing datasets from a local desktop machine, and submitting a causal discovery analysis job to an HPC machine via a web browser. The Causal Web application supports file uploads for very large datasets, file annotation to maintain provenance, a built-in visualizer for causal graphs, and a method for comparing causal graphs. Currently, we use the Bridges supercomputer at the Pittsburgh Supercomputing Center as the backend HPC machine, although the architecture can support using any HPC cluster. We make a large number of CPU hours freely available to all users who complete a simple one-minute sign-up process.

Community. Users of this platform include biomedical scientists and data scientists.

Usability assessment and evaluation. *Please see the description above*.

Discoverability. We announce its availability and updates via Reddit, Google+, and other email lists. We also describe and demonstrate its use at our annual week-long Causal Discovery Summer Short Course and our day-long Causal Discovery Datathon, both of which are held at Carnegie Mellon University. Attendees of both events can use Causal Web with their own data or supplied datasets for hands-on, real-world learning in a supportive environment. Moreover, we discuss its availability at conferences at which we make research presentations and at workshops in which we give tutorials on causal modeling and discovery.

Dissemination. Information and documentation for using Causal Web is available on our Center website (www.ccd.pitt.edu/tools/) and via a site at the PSC (ccd2.vm.bridges.psc.edu/ccd), which is pointed to by the resources page at the BD2K-CCC website (bd2kccc.org/bd2k-resources/). Since deploying Causal Web, we have registered over 50 users of the application, and they have used over 50,000 CPU hours to perform causal discovery analyses of their data.

Future relationship to Commons. We plan to make the availability of Causal Web discoverable to users of the Commons as a web application to analyze data for causal relationships. At the point that the Commons is supporting web services, the Causal Web backend server can be run from there.

5. Causal Modeling & Discovery Educational Materials

The CCD website (www.ccd.pitt.edu) has pointers to a rich set of educational materials on causal modeling and discovery that are easily accessible and freely available. The materials include:

- All of the lectures in our Summer Short Course on causal modeling and discovery (www.ccd.pitt.edu/2016-short-course-datathon/),
- Slides and videotaped presentations of Distinguished Lectures that discuss state-of-the-art work in causal modeling and discovery (www.ccd.pitt.edu/distinguished-lecture-series).
- An introductory, online, computer-based course on causal and statistical reasoning

(www.ccd.pitt.edu/open-learning-initiative-courses/).

Community. Users of this platform include biomedical scientists and data scientists at all levels of training and experience, from undergraduates to faculty biomedical investigators.

Discoverability and dissemination. The educational materials on the CCD website are indexed by the BD2K-TCC website (bigdatau.org/about_erudite). The Summer Short Course and Distinguished Lecture videos are also available on our CCD channel on YouTube (www.youtube.com/channel/UCBWDYSwP-bUMmRWZHuGJcJ5g), where they are also indexed. At our annual week-long Causal Discovery Summer Short Course, we make sure that the attendees are aware of these materials. We also discuss their availability at workshops at which we give tutorials on causal modeling and discovery. In the past year, on the CCD website the educational material web pages have received more than 6,000 views. On YouTube, the Summer Short Course videos have received over 3000 views for Lecture 1 alone, with over 790 views for the Distinguished Lecture videos.

Future relationship to Commons. We plan to make the availability of these educational materials discoverable to users of the Commons.

CEDAR Top Five Products

In biomedicine, metadata are essential for finding experimental datasets, for understanding how experiments have been performed, and for reproducing those experiments. The Center for Expanded Data Annotation and Retrieval (CEDAR) is developing an array of software tools to help create, manage, and submit high-quality biomedical metadata for use in online data repositories. Our approach centers on the use of editable metadata templates, which define collections of data elements needed to describe particular types of biomedical experiments. CEDAR tools provide a series of features that lower the barrier to the creation, management, and population of metadata templates, and that allow users to submit these metadata to public repositories. The overall goal is to provide the ability for investigators to easily create and submit metadata that are comprehensive and standardized, and to ensure that the corresponding data sets satisfy the FAIR principles in support of open science.

CEDAR provides a set of Web-based tools and REST APIs to manage the creation, storage, search, reuse, and submission of metadata. The system provides easy-to-use services that help scientists to make sense of the fragmented landscape of metadata formats and metadata submission tools.

A key goal has been the creation of a common language for metadata templates that serve as the underpinning of the system. To this end, we first defined a lightweight standards-based metadata model that provides principled interoperation with Linked Open Data. Unlike other approaches, the metadata model in CEDAR provides the foundation for a comprehensive, user-managed library of metadata templates that are themselves first-class representations that are fully editable and customizable.

Architecturally, the system can be divided into five core components (Figure 1): (1) the CEDAR Workbench, a set of highly-interactive Web-based tools that enable the collaborative development, sharing and publication of metadata resources; (2) the CEDAR Metadata Management APIs, which support the creation, modification, and archiving of metadata; (3) the CEDAR Metadata Model, which provides a standards-based representation and exchange format for metadata; (4) the CEDAR Metadata Repository, which provides a queryable storage system for metadata and templates; and (5) the CEDAR Submission Service, which offers a collection of tools to manage the transmission of metadata to various public repositories.

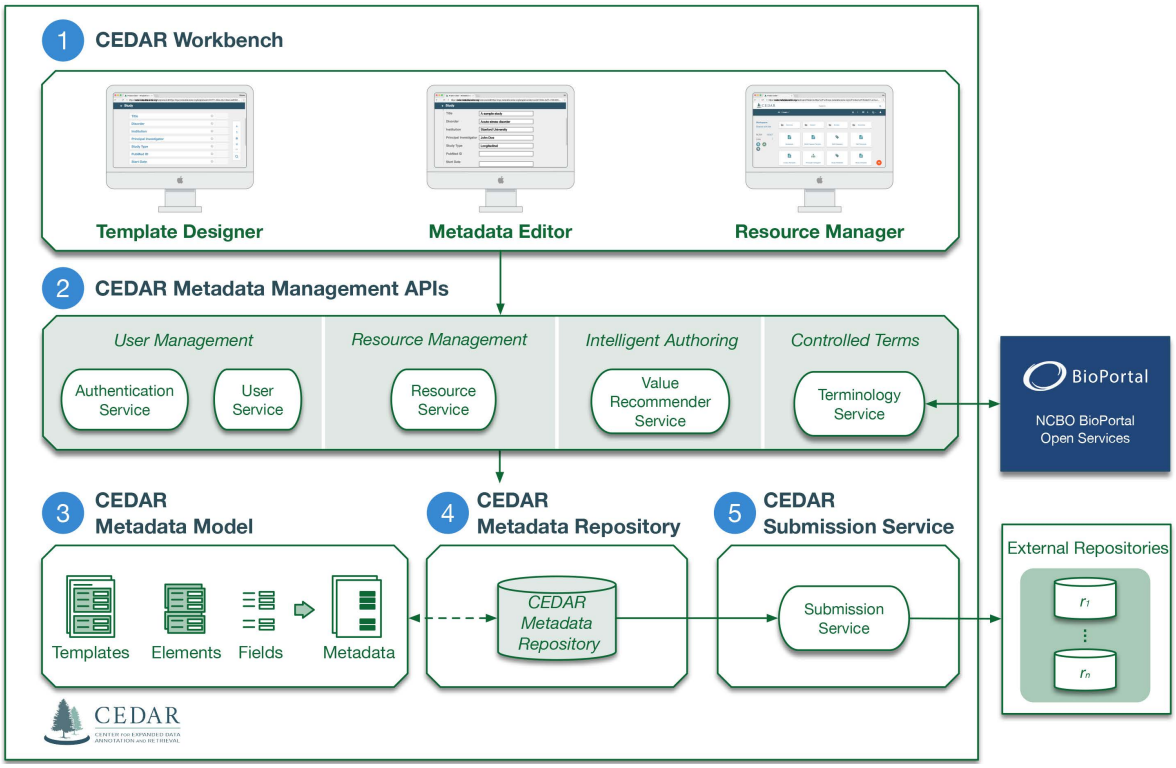


Figure 1. Overview of the architecture of the CEDAR Workbench

1. CEDAR Workbench

The CEDAR Workbench provides a unified platform for developing and publishing metadata resources. The primary goal is to support the collaborative discovery, exchange, and management of metadata templates and metadata instances. The Workbench comprises a set of highly-interactive Web-based tools. Users first create metadata templates with the Template Designer, interactively specifying the templates' structure and content, and defining semantic relations using interactive, user-friendly look-up services linked to the NCBO BioPortal ontology repository. The Metadata Editor automatically generates a metadata entry interface for each template, prompting users with drop-down lists, auto-completion suggestions, and hints to significantly reduce errors and to increase the speed of metadata entry. Users can then manage and share template resources created with these tools using the Resource Manager.

Community. Several groups use the CEDAR Workbench to develop metadata templates and instances. Representative groups include: (1) the Adaptive Immune Receptor Repertoire (AIRR) community, creating metadata templates to characterize high-throughput repertoire sequencing (Rep-Seq); (2) the Library of Network-Based Cellular Signatures (LINCS) Consortium, developing metadata templates to describe cell lines and a variety of instrument-based experiments; (3) the Data Commons Credit Pilot, an NIH-funded effort to support the creation and management of FAIR data in the Cloud; and (4) LD4P, an initiative of a consortium of University Research Libraries to ease the authoring of metadata for their collections in a manner that is consistent with linked open data principles.

Usability assessment and evaluation. The Workbench allows non-specialist users to collaboratively develop and publish metadata resources, so usability is key. We modeled Workbench capabilities on the Google Drive platform, presenting a gentle learning curve to users. Several cycles of user testing and over 16 months of post-release field tests confirmed the tool's usability.

Discoverability. Our primary CEDAR website, metadatacenter.org, contains documentation for the Workbench and outreach products, and it points to an online training guide. The CEDAR Web presence is complemented by our GitHub portal and wiki, and social media channels have been activated for Twitter (@metadatacenter) and Facebook. We have driven discoverability of this resource via academic publications, attendance at various conferences, social media outreach, and by running training workshops for with interested communities.

Dissemination. The Workbench can be accessed at cedar.metadatacenter.org. The product has over 600 registered users, with over 1000 templates and 34000 instances in CEDAR's repositories. These repositories serve biomedical projects across NIH. The product has been incorporated into several research workflows, including two workflows that have reached production status and several more that are targeting production use in the next few months.

Future applicability to shared data commons. We anticipate the CEDAR Workbench will be a capable tool for submitting metadata and content to any shared data commons, for managing the shared metadata (including updates), and for supporting complete life cycle operations (including versioning of information artifacts). CEDAR enables project-level curation of metadata resources and submission of data and metadata to public repositories.

2. CEDAR Metadata Management APIs

Using the metadata model as a foundation, CEDAR provides an array of REST APIs for managing metadata. The APIs support rich metadata resource management to create, update, and share metadata resources. Unique services provide metadata value suggestions and terminology services from BioPortal.

Community. Interacting groups include: (1) ImmPort, (2) the Adaptive Immune Receptor Repertoire (AIRR) community, (3) the Library of Network-Based Cellular Signatures (LINCS) Consortium, (4) the Human Immunology Project

Consortium (HIPC), (5) the Commons Credits Model Pilot, and (6) individual researchers developing their own metadata pipelines.

Usability assessment and evaluation. We performed field-testing with collaborating groups before and after the public release of the APIs in September 2016. The REST APIs have exceeded our usability goals, with many user groups adopting them with no in-person training required, thanks to an elegant specification built around the data artifacts and to detailed online documentation with example code.

Discoverability. Our primary CEDAR website, metadatacenter.org, contains documentation for the APIs and outreach products. The CEDAR Web presence and GitHub documentation describe the APIs, and a complete Swagger documentation set is at resource.metadatacenter.org/api. Because the APIs are centered on our principled and well-specified metadata model, their application is generally straightforward.

Dissemination. The REST APIs are provided as part of the CEDAR Workbench and are available at resource.metadatacenter.org. The APIs are extensively incorporated into broader platforms and workflows, being designed for adoption (in their entirety, or as subcomponents) in a system-of-systems framework. All CEDAR's users have access to the APIs, and we've explicitly coordinated API use with the collaborating projects listed above.

Future applicability to shared data commons. The API services are fundamental to capabilities required by any shared data commons. Providing fully authenticated access to all metadata services, the APIs enable metadata workflows for any external software system that handles metadata following the CEDAR general-purpose metadata model. We envision robust access via the APIs to all the metadata provided to shared data commons through CEDAR systems.

3. CEDAR Metadata Model

The standards-based CEDAR metadata defines a common format for describing metadata templates and instances, promoting FAIR principles and interoperating with Linked Open Data. It offers a means to support template composition, semantic markup and linkage, and value constraint and validation capabilities.

Community. The Metadata Model is used implicitly by all users of the CEDAR Workbench, and explicitly by many groups, including: (1) ImmPort, a repository of immunology-related datasets; (2) the Adaptive Immune Receptor Repertoire (AIRR) community; (3) the Library of Network-Based Cellular Signatures (LINCS) Consortium; (4) the Human Immunology Project Consortium (HIPC); and (5) FAIRSharing, which applies and analyzes this model in multiple contexts.

Usability assessment and evaluation. We performed significant field-testing with these external research groups to ensure that the model was both expressive and usable for real production workflows. We assessed the ability of our model to represent real-world metadata from these collaborating groups. We also evaluated the model's ability to comprehensively link metadata elements and value constraints to controlled vocabularies and ontologies, as well as other essential criteria. Finally, we developed a validation library to ensure model conformance.

Discoverability. The CEDAR Metadata Model is comprehensively documented at metadatacenter.org/tools-training/outreach/cedar-template-model and in a scientific publication (see metadatacenter.org/open-repository-model). Many presentations showed the model's value proposition and its organization. Extensive work with our collaborating organizations and with BioCADDIE provided forums for discussing the model in more detail.

Dissemination. The Metadata Model is used as the foundation of the CEDAR Workbench and is being considered for additional interoperability applications. Many tools for working with this model exist on GitHub, and CEDAR collaborators often use these tools, or work directly with the model.

Future applicability to shared data commons. The CEDAR Metadata Model is designed using an open,

standards-based format for representing metadata templates that conform to FAIR principles. The model is built around the open-science notion of providing a common platform for information exchange, and is being explored (with organizations such as BioCADDIE) as the basis for widespread metadata interoperability. We believe it could serve as an exchange standard for shared data commons metadata.

4. CEDAR Metadata Repository

CEDAR's Metadata Repository provides storage, dissemination and discoverability for all CEDAR metadata resources. This component and the underlying Metadata Model underpin all CEDAR REST services and UI tools. The Metadata Repository also has explicit user-facing functions: (1) permanent storage of metadata for systems that use CEDAR as their end repository; (2) intermediate metadata storage, pending submission to external repositories; (3) searchable repository of public metadata artifacts, via either GUI or API; (4) metadata storage for metadata collections being processed (in a pipeline) or analyzed (e.g., for intelligent metadata suggestions).

Community. The CEDAR Metadata Repository engages all users of the Workbench, as it supports all system functions. We describe it as a stand-alone product because it also serves user needs beyond simple Workbench use (see description above).

Usability assessment and evaluation. We performed significant testing internally and with external research groups to ensure the Metadata Repository met needs of actual production workflows. Extensive use in production releases, including medium-scale storage of >60,000 instances, helped us optimize critical system operations. End-to-end system tests verify proper component behavior.

Discoverability. We have driven discoverability of this resource via the CEDAR Workbench, and by communications with interested communities. We ensured that API documentation and interoperability concerns were thoroughly addressed as we developed other components (viz., REST APIs and Metadata Model). As the Metadata Repository grows to contain more resources, we will advertise its availability to likely user communities.

Dissemination. The Metadata Repository is used as the core metadata storage system in the CEDAR Workbench, the Web-based platform available at cedar.metadatacenter.org. Users may choose to deploy the Metadata Repository (or any CEDAR component) in their own domain for dedicated use.

Future applicability to shared data commons. The CEDAR Metadata Repository is an open, standards-based platform for storing metadata templates and instances that conform to FAIR principles. It serves the open-science goal of a common platform for information exchange, and it has been designed using scalable, configurable, and interoperable principles appropriate for any shared data commons. It offers both central storage of shared data commons metadata and options for distributed deployment for specific projects.

5. CEDAR Submission Service

CEDAR's Submission Service is designed to help authors to submit semantically rich metadata to public repositories. It is designed to provide a uniform API to external metadata repositories. Architecturally, it has been developed to ensure that in can be incorporated into the submission workflows of third party groups in a variety of ways. It works in concert with core CEDAR tools, such as its controlled term recommendation and intelligent authoring services, to ensure that users can quickly create high quality, semantically enriched metadata and to submit these metadata to existing repositories.

Community. Several groups use the Submission Service to submit metadata to public repositories. These groups include: 1) the Adaptive Immune Receptor Repertoire (AIRR) community, which is using CEDAR to submit metadata to NCBI's BioProject, BioSample and SRA repositories;

2) the Library of Network-Based Cellular Signatures (LINCS) Consortium, which is using CEDAR to submit metadata to a public LINCS-based repository; and (3) the Data Commons Pilot, which has developed a submission pipeline to submit metadata to the BioCADDIE repository.

Usability assessment and evaluation. We performed significant testing of the Submission Service during the development of pipelines with collaborating research groups to ensure the service meets the needs of production workflows. The LINCS-based submission pipeline went into production in May 2018, the Commons Credits Model pilot went into production in June 2018, and the AIRR pipeline is going into production in July, 2018. We plan to carry out post-submission analyses of submitted metadata to ensure that the quality goals of the various submitters are met.

Discoverability. We have driven discoverability of this resource via academic publications, attendance at various conferences, social media outreach, and by communications with interested communities. We ensured that the services met core functionality goals in the development of submission pipelines for the three collaborating groups described above. As the functionality of the Submission Service grows, we will advertise its availability to likely user communities.

Dissemination. The Submission Service is used as the core metadata system in CEDAR, the Web-based platform available at cedar.metadatacenter.org. As demonstrated by the current pipelines, the services can be customized to work with a variety of workflows and is designed to be integrated into the submission pipelines of third party groups.

Future application to shared data commons. The CEDAR Submission Service provides the foundation for the development of pipelines for submitting semantically enriched metadata templates to a variety of public repositories. The service aims to meet the open-science goal of providing a common platform for metadata enrichment and submission, and has been designed using scalable, configurable, and interoperable principles appropriate for any shared data commons.

The Submission Service can help build a central submission portal for a shared data commons that can be tuned for the deployment needs of specific projects. In addition to the submission pipelines described above, we plan to provide an array of generic submission pipelines, including pipelines that can be used by researchers to submit to NCBI's BioProject, BioSample, and SRA repositories.

1. MetaSRA Metadata and Pipeline

MetaSRA is a resource providing a mapping of sample-specific metadata in the Sequence Read Archive (SRA) to standard biomedical ontologies. Currently, MetaSRA maps biological samples to biologically relevant terms in the Disease Ontology, Experimental Factor Ontology, Cell Ontology, Uberon, and Cellosaurus. We have also made available the software system for this annotation, which can be repurposed to do similar metadata mappings for other data resources.

Community. The data resource is of value to: 1) scientists who want to do large-scale secondary analysis of data in the SRA; and 2) other groups that provide value-added services for the SRA such as the BD2K-LINCS center. The software system is of value to other projects that want to map unstructured metadata from an arbitrary data resource onto structured ontology-based annotations.

Usability assessment and evaluation. We conducted a rigorous experiment with held-aside data to measure the accuracy of the system in mapping unstructured metadata to ontology-based annotations.

Discoverability. The data resource and software system have been described in a preprint that was deposited in BioRXiv (doi.org/10.1101/090506) and is currently under review for publication in a journal. Additionally, the Resources section of CPCP web site links to both the data resource and the software system.

Dissemination. The product was released just a few months ago, so it is not yet broadly used. However, the BioRXiv abstract has more than 4000 views, the PDF has been downloaded more than 400 times, and there are 87 Tweets referencing the system. The product is being incorporated into a broader scientific workflow for transcriptome-based cellular phenotyping that is under development in CPCP.

Future relationship to Commons. The software is open-source and publicly available. It can be incorporated into a wide range of analysis pipelines on a wide range of platforms. The data resource will significantly enable the SRA to be more widely used in computational studies.

2. Single-cell RNASeq Analysis Tools

We have developed a suite of tools for analyzing and extracting phenotypes from single-cell RNASeq data. This set of tools includes *EBSeqHMM* (auto-regressive hidden Markov model (HMM) for identifying genes and isoforms that have expression changes in ordered RNASeq experiments), *scDD* (a method for characterizing differences in expression in the presence of distinct expression states within and among biological conditions), *Oscope* (a statistical pipeline for identifying oscillatory genes and characterizing one cycle of their oscillation), and *OEFinder* (method to identify genes having the so-called ordering effect in single-cell RNASeq data generated by the Fluidigm C1 platform).

Community. These software tools are being used by a broad range of biomedical scientists conducting single-cell RNASeq experiments.

Usability assessment and evaluation. The methods and software were developed in the context of collaborative projects with biologists that CPCP investigators are involved in. We have evaluated usability by assessing the effectiveness of the methods in these biological studies, as well as by feedback from the large user community (see download statistics below).

Discoverability. The algorithms and software have been described in peer-reviewed articles published in Bioinformatics (EBSeqHMM and OEFinder), Genome Biology (scDD), and Nature Methods (Oscope). The software tools are discoverable and available via BioConductor or GitHub. Additionally, the Publications section and the Resources section of CPCP web site link to the articles and the software systems, respectively.

Dissemination. The products have been disseminated via BioConductor and GitHub. The number of software downloads for the software tools are as follows: EBSeqHMM (> 5700), Oscope (~2,700), scDD (more than 130).

Future relationship to Commons. The software is open-source and publicly available. It can be incorporated into a wide range of analysis pipelines on a wide range of platforms.

3. Magellan Entity Matching Tools

Entity matching (EM) is a fundamental problem in data integration. Magellan is a set of tools and how-to guides for developing entity-matching systems. The Magellan tools are built on the Python data science and big data eco-system, and aim to cover all the tasks involved in developing an entity-matching pipeline.

Community. The potential users of this system are data scientists involved in integrating multiple biological data sets.

Usability assessment and evaluation. The product has been developed in the context of several data-integration tasks that CPCP investigators are involved in. We have evaluated usability by assessing the effectiveness of the software in projects that are mapping unstructured RNASeq metadata to ontologies and aligning three drug ontologies.

Discoverability. Magellan has been described in two publications, including VLDB (Konda et al., 2016). The software and associated tutorials and data are available through a Google site (sites.google.com/site/anhaidgroup/projects/magellan), which the CPCP web site links to.

Dissemination. The product was released just a few months ago, so it is not yet broadly used. It has been incorporated into workflows in two CPCP projects. The code is available through a Google site. It is distributed with a tutorial that walks users step-by-step through the process of assembling an entity-matching pipeline.

Future relationship to Commons. The software is open-source and publicly available. It can be incorporated into a wide range of analysis pipelines on a wide range of platforms.

4. atSNP

atSNP is a method and software tool for statistically assessing the potential change in transcription-factor binding induced by a genetic variant. The atSNP portal (under development) will make available these calculations for 2000 transcription factors at 133 human SNPS (3.1 billion TF-SNP combinations). The portal will support querying, browsing, visualizing and downloading this large set of TF-SNP interactions.

Community. This product is aimed at a broad range of biomedical scientists who have interests in characterizing the potential regulatory roles of genetic variants of interest.

Usability assessment and evaluation. The methods and software were developed in the context of collaborative projects with biologists that CPCP investigators are involved in. We have evaluated usability by assessing the effectiveness of the methods in these biological studies.

Discoverability. The atSNP method was published in an article in Bioinformatics (Zuo et al., 2015). The software is available via GitHub and the Resources section of the CPCP web site. A separate article will be submitted on the web portal which is under development.

Dissemination. The software is available via GitHub. It has also been incorporated in the ENCODE analysis pipeline.

Future relationship to Commons. The software is open-source and publicly available. It can be incorporated into a wide range of analysis pipelines on a wide range of platforms.

5. CPCP Videos

We have produced 23 videos to date that have been publicly released. These fall into several different categories including Small Talks on Big Data, talks and panel discussions from our Symposium Big Privacy: Policy Meets Data Science, presentations from the CPCP annual retreat, and presentations from seminars.

Community. Collectively, these videos are intended for a broad range of audiences including biomedical scientists, clinicians, data scientists, students and postdocs, and the general public.

Discoverability and dissemination. The videos are made available on our Center web site and also on Vimeo where they are hosted. They have been viewed more than 1,100 times. The audiences served, depending on the video, include biomedical scientists, clinicians, data scientists, students and postdocs, and the general public. The videos are indexed and discoverable via search engines such as Google.

1. ENIGMA MRI

ENIGMA's MRI toolkit is used in over 35 countries to analyze brain data, in the largest studies of the brain worldwide.

Community. The ENIGMA Center provides 24/7 support to over 1,000 users, who run our tools in the largest-ever studies of the human brain – analyzing brain MRI scans from >53,000 people across 35 countries. ENIGMA's analysis protocols for brain MRI have been used by over 1,000 labs worldwide; they were used to perform the world's largest neuroimaging studies of *schizophrenia*, *major depression*, *bipolar illness*, and *obsessive compulsive disorder* combining MRI data from >20,000 people.

Usability assessment and evaluation. Using the ENIGMA MRI toolkit, the ENIGMA-ADHD Working Group published the largest-ever neuroimaging study of ADHD (Hoogman et al., Lancet Psychiatry), with over 20,000 views in its first week online. The study was highlighted in 147 news media outlets, including CNN, the Washington Post, and TIME magazine (elsevier.altmetric.com/details/16454649/news).

Discoverability. Highlighted by the New York Times, Science, and The Lancet ("Crowdsourcing meets Neuroscience"), ENIGMA's 33 working groups study 18 brain diseases, uniting data, resources and talents of 800 scientists from 340 institutions. ENIGMA-Shape, a dedicated toolkit for fine-scale anatomical analysis, is also available (enigma.ini.usc.edu/ongoing/enigma-shape-analysis).

Dissemination. ENIGMA's workshops, plenary lectures, and training events at the NIH (2), and across the EU, Russia, Siberia, the Middle East, Mongolia, the Thai Red Cross, Korea (KAIST), and Ecuador, drew thousands of attendees in aggregate, including a keynote ENIGMA lecture to thousands of radiologists at ISMRM (Toronto), the Chinese Congress of Radiology, and Russian Academy of Sciences. Lectures and tutorials are online at the ENIGMA website, with our codebase (github.com/ENIGMA-git/ENIGMA).

Future relationship to Commons. ENIGMA's MRI toolkit can be used to analyze brain imaging data stored in the NIH Data Commons, and is already used to analyze vast datasets from brain MRI databases worldwide.

2. ENIGMA DTI

ENIGMA's Tools for Diffusion Tensor Imaging (DTI) is used to study the brain's microstructural abnormalities in disease; it is used in 9 worldwide DTI studies: 1) ENIGMA-*Bipolar*, whose DTI study is led by researchers in Paris (Houenou et al., OHBM 2016); 2) ENIGMA-*MDD* (Major Depression) whose DTI study is led by researchers in Los Angeles, Boston, and Amsterdam (Kelly et al., OHBM 2016); 3) ENIGMA-*Addiction*, whose DTI study is led from South Africa and Vermont; 4) ENIGMA-OCD's DTI working group, also known as "DOMAIN", led from Rome and Amsterdam; 5) ENIGMA's 22q11 Deletion Syndrome group, which recently published its first DTI analysis (Villalon et al., OHBM 2016); 6) the ENIGMA-*Epilepsy* group, whose DTI project is led by researchers in London and Los Angeles; 7) the ENIGMA-*ADHD* group, led from Nijmegen; 8) the ENIGMA-*PTSD* group, led from USC and Duke; and 9) ENIGMA-*HIV*, led by researchers in South Africa and Los Angeles.

Community. ENIGMA-DTI Core members support the world's largest diffusion MRI studies of *9 major brain diseases*, in coordination with the ENIGMA disease working groups. Trainees from all over the world are using our protocols within the disease working groups, in global initiatives studying white matter microstructure for each disease. These global efforts offer a new opportunity to produce water-tight and convincing findings on white matter abnormalities in each brain disease.

Usability assessment and evaluation. The tools are presently used in the world's largest DTI studies of schizophrenia, bipolar disorder, major depression, 22 deletion syndrome, addiction, obsessive com-pulsive disorder, HIV/AIDS, epilepsy, and post-traumatic stress disorder (PTSD).

Discoverability. As one of the top downloaded tools in the public domain on NITRC (the NeuroInformatics Tools and Resources website, www.nitrc.org/frs/?group_id=981), ENIGMA-DTI has had 1,800+ downloads on the NITRC website, with a minimum of 140 downloads a month. The UK Biobank modeled the release of DTI summary measures from their 100,000 diffusion brain scans on the ENIGMA-DTI protocol, as referenced in their imaging notes.

Dissemination. ENIGMA-DTI is available at the ENIGMA website (enigma.usc.edu). The ENIGMA-Schizophrenia group was the first to use the ENIGMA-DTI protocol, and pooled together a sample of approximately 2,000 patients to compare with 2,000 individu-

als without schizophrenia (paper under revision at Molecular Psychiatry). ENIGMA's 10K-in-1day event in Utrecht was the largest ever supercomputing analysis of brain connectivity, analyzing connectomes from over 14,000 people worldwide.

Future relationship to Commons. ENIGMA DTI can be used to analyze brain imaging data stored in the NIH Data Commons, and is already used to analyze vast datasets from brain MRI databases worldwide.

3. ENIGMA Genomics

ENIGMA Genomics tools are used to perform the largest-ever genetic studies of the human brain – analyzing brain MRI scans from >30,000 people across 35 countries. Tools are included for genetic imputation, quality control, and large-scale analyses, including genome-wide association studies (GWAS).

Community. These tools are used to perform the largest-ever genetic studies of the human brain. One reviewer at Nature Neuroscience noted the sheer scale of this study: "*the unprecedented organizational capabilities and technical prowess required to coordinate, harmonize, and integrate this large scale imaging data into a unified study. At the data acquisition and processing side, the study is truly one-of-a-kind, uniquely remarkable in its size, scope, and synthesis.*" The code is used at over 200 institutions.

Usability assessment and evaluation. In a world-wide project of unprecedented scale, we analyzed brain MRI and GWAS data from 251 institutions (Hibar et al., Nature 2015; Adams et al., Nature Neuroscience 2016; Hibar et al., Nature Communications 2017).

Discoverability. ENIGMA's Genomics toolkit is distributed in a version-controlled format on ENIGMA's GitHub: github.com/ENIGMA-git/ENIGMA/tree/master/Genetics. For all our big data projects in genomics, epigenetics, and CNV analysis of the brain, we detail step-by-step instructions on how to extract and QC imaging phenotypes, and run each step of the genetic associations. We provide tutorials on our protocols; a series of videos also provide instructions and skills for novice users to learn the basics of syntax for Unix, bash, Matlab, and R. These videos can be watched on our YouTube Channel (e.g., www.youtube.com/watch?v=hxixl-22Zks).

Dissemination. ENIGMA's international support group provides around the clock, cross time-zone support for any problems with genetic imputation, cortical phenotype extraction, quality control, and each step of the GWAS along the way. The group consists of ENIGMA Co-investigators Sarah Medland, Neda Jahanshad, Jason Stein, and new trainees from the imaging genetics field across ENIGMA-funded institutions, including Janita Bralten from Radboud UMC, Nijmegen, Netherlands and Penelope Lind from the Queensland Institute for Medical Research in Brisbane, Australia.

Future relationship to Commons. ENIGMA Genomics tools can be used to analyze genomics data stored in the NIH Data Commons, and is already used to analyze vast datasets from genomics databases worldwide.

4. ENIGMA-Vis

ENIGMA-Vis is a tool to query, display, and understand gene effects on a variety of brain measures.

Community. ENIGMA performs the largest-ever genetic studies of the human brain (Hibar et al., Nature 2015; Adams et al., Nature Neuroscience 2016; Hibar et al., Nature Communications, January 2017), discovering new markers in the human genome that boost disease risk or protect us from disease. ENIGMA-Vis provides access to ENIGMA's genomic findings.

Usability assessment and evaluation. ENIGMA's two papers at Nature Communications map gene effects through-out the brain worldwide (Roshchupkin et al., 2016; Hibar et al., 2017). ENIGMA's world-wide map shows how the gene APOE elevates Alzheimer risk throughout life and worldwide in 33,000 people scanned with MRI.

Discoverability. Since April 2016, ENIGMA's genomic data related to the brain's subcortical volumes have been requested by 122 labs. Since November 2016, the GWAS of intracranial volume (a collaboration with the CHARGE Consortium) has been requested by 46 labs worldwide. These requests come from verified researchers across the US and internationally. We have received requests from every continent (except Antarctica), including countries such as Israel, China, Morocco, Australia, Colombia, and Serbia.

Dissemination. The ENIGMA-Vis Genome browser is at enigma.ini.usc.edu/enigma-vis. A user can pick a gene, or a region of the genome, and load up brain measures, using checkboxes, to see which parts of the brain it affects.

Future relationship to Commons. ENIGMA-Vis can

be used to view, browse and analyze data stored in the NIH Data Commons, and is already used to analyze vast datasets from genomics databases worldwide.

5. ENIGMA Training

As a metric of the global impact of ENIGMA Training, ENIGMA has since published the world's largest studies of five psychiatric diseases using brain imaging. These global studies were hailed in the New York Times as, "breaking the logjam in neuroscience," "giving us a power we have not had," and, in The Lancet, as "Crowdsourcing meets Neuroscience" (Mohammadi, 2015). *ENIGMA won the Kent Innovation in Academia Award for breaking down international barriers in science.*

Community. Over 1,000 labs across the world participate. The ENIGMA consortium consists of over 30 working groups (WGs) made up of 800 scientists from over 230 institutions and 35 countries; the WGs have over 150 active projects. The diseases studied include ADHD, autism, addiction, bipolar disorder, epilepsy, HIV, major depressive disorder, OCD, PTSD, and schizophrenia. In the past year, new and active WGs have been created and evolved into worldwide consortia on epilepsy (Whelan 2017), anorexia (King 2017), and anxiety disorders.

Discoverability and dissemination. To help consortium leaders and trainees become involved in ENIGMA, we produce a range of training materials that are also available on our website. These include detailed technical videos on big data analytics in genomics, imaging, and mathematics, to one-on-one interviews with ENIGMA Chairs, with updates on the goals, status, and challenges of running global studies of various brain diseases.

All training videos are available via ENIGMA's website (www.youtube.com/channel/UCT0eih_EoNpXtG9kM-0bjCKg/featured). The American Society for Human Genetics commissioned a 5-minute film on ENIGMA, shown to 8,000 attendees at their annual conference. The film was professionally produced and was shot on location in Sonoma, California, USA, and Nijmegen, in the Netherlands, and features over 20 ENIGMA scientists. The video is available at www.youtube.com/-watch?v=JzQokpBqxIY.

As a very large scale global project in imaging, ENIGMA training events and lectures have been held at a large number of international and national meetings, in China, Russia, Mongolia, the UK, Iran, India, Austria, Hong Kong, Thailand, and Australia, and across the Americas. In the last year, ENIGMA PI Paul Thompson gave over 50 lectures on ENIGMA, at the Russian National Academy of Sciences, the European Molecular Biology Laboratory (EMBL), Cambridge, UK, the University of Maryland, Washington University in St. Louis, the University of Vienna, Austria, and the Broad Institute of Harvard and MIT.

1. AZTec

AZTec is a global biomedical resource discovery index that allows users to simultaneously search a diverse array of tools. It provides the scientific community with a new platform that promotes a long-term, sustainable ecosystem of biomedical research software. AZTec (A to Z technology; aztec.bio), an open-source online software discovery platform that empowers biomedical researchers to find the software tools they need. The resources indexed include web services, standalone software, publications, and large libraries composed of many interrelated functions. AZTec will ensure that software tools remain findable in the long term by issuing persistent DOIs and routinely updating metadata for the entire index. AZTec's established ontologies and robust API support the programmatic query of its entire database, as well as the construction of indexes for specialized subdomains. AZTec currently hosts over 10,000 resources spanning 17 domains, including imaging, gene ontology, text-mining, data visualization, and various omics analyses (i.e., genomics, proteomics, and metabolomics).

Community. Users of AZTec include researchers and tool developers in the biomedical fields. AZTec targets: 1) the broad biomedical basic science community, 2) the translational clinical science community, and 3) funding agencies.

Usability assessment and evaluation. Aztec is currently in its alpha-release phase (version 1.1), in which it is being evaluated and tested by internal users at UCLA, as well as invited external users at Sage Bionetworks, TSRI, and EMBL-EBI. Their feedback and comments have been documented and incorporated into Aztec's next release (version 1.2, anticipated 5/1/17).

Discoverability. AZTec fosters an environment of sustainable resource support and discovery, empowering researchers to overcome the challenges of information science in the 21st century. The web platform of Aztec is hosted on an elastic compute cloud server on Amazon Web Services (AWS). Its current source code is published on GitHub (github.com/BD2K-Aztec).

Dissemination. We disseminate AZTec through publications, routine press releases are made to biomedical communities (through mailing lists and on social media), presentations and software demos at major conferences of targeted user communities. AZTec has been presented in 17 national and international conferences, including BD2K AHM, International Data Week, the California Big Data Workshop, ISMB, Recomb, HUPO, OCC, ISHR, AHA, IEEE, NIPS, etc. AZTec has been advertised on the HeartBD2K website and a video demo on YouTube; in addition, we are producing an animated promotional video on Aztec. Aztec can be accessed at aztec.bio .

Future relationship to Commons. AZTec provides a platform for enabling and ensuring the compliance of software and tools with FAIR principles. Tools and software become findable through the AZTec search engine, supporting multi-faceted queries based on tool function, biological domain, prerequisites, input and output formats, and a host of other vital metadata fields. AZTec's RESTful API enables universal accessibility of tools and software. It enhances interoperability by adopting existing vocabularies to define tool metadata whenever possible. AZTec ensures the reusability of both the platform and the tools indexed by it.

2. OmicsDI

The Omics Discovery Index (OmicsDI) provides dataset discovery across a heterogeneous, distributed group of transcriptomics, genomics, proteomics, and metabolomics data resources including both open and controlled access data resources. Based on metadata, OmicsDI provides extensive search capabilities, as well as identification of related datasets by metadata and data content where possible. In particular, OmicsDI identifies groups of related, multi-omics datasets across repositories by shared identifiers. It also provides metrics on specific dataset impact through referencing data reuse and access counts.

Community. OmicsDI targets: 1) the bioinformatics community to identify reference data sets; 2) the biomedical science community to find relevant data sets; and 3) funding agencies to assess the specific impact of datasets.

Usability assessment and evaluation. OmicsDI builds on 7 years of experience in the international ProteomeXchange Consortium, as well as EBI Search, the major search engine of the European Bioinformatics Institute. Development is done in close collaboration with currently 11 data resource providers, and an active Twitter discussion. Formal user feedback is collected through two separate surveys, one for end users and one for data providers. An intensive user experience testing phase with novice users is budgeted for Q3/2017. Information is then used to enhance the soft-

ware accordingly, providing new and updated functionality that meets users' needs.

Discoverability. The web interface is at www.omicsdi.org, with all web services documented at www.omicsdi.org/ws and source code available on GitHub (github.com/OmicsDI). @OmicsDI is maintaining an active Twitter presence (136 tweets, 83 followers, 110 likes. Key publications providing information on OmicsDI include:

- *Omics Discovery Index – Discovering and Linking Public Omics Datasets*. Yasset Perez-Riverol, et al. doi: doi.org/10.1101/049205; bioRxiv 049205.
- Omics Discovery Index - Discovering and Linking Public Omics Datasets. Yasset Perez-Riverol, et al. Nature Biotechnology 2017, in press.

Dissemination. OmicsDI is disseminated through various channels of communication, including Twitter, a help topic blog and a publication in Nature Biotechnology accepted in 2017.

Future relationship to Commons. Discoverability is a key element of the FAIR principles, and OmicsDI implements a production grade tool for dataset discoverability. The distributed approach of OmicsDI, with a thin centralized metadata layer lends itself well to implementation of a dataset discoverability index for the commons.

3. MyGene.info

MyGene.info provides simple-to-use REST web services to query/retrieve gene annotation data, which are constantly kept up-to-date. It is designed with simplicity and performance emphasized. A typical use case is to use it to power a web application or an analysis pipeline which requires querying genes and obtaining common gene annotations, so that researchers can make more efficient science without the burdens of setting up and maintaining their own local gene annotation databases.

Community. MyGene.info targets the data science and computational biology community.

Usability assessment and evaluation. MyGene.info is currently in production stage with average ~5 million API requests every month (from ~4,000 unique IPs monthly and total 34K unique IPs over last 12 months). It hosts 18 million genes from 19K species with over 200 annotation types. The underlying gene annotation data are updated weekly.

Discoverability. MyGene.info production is hosted as a cluster in AWS, and optimized for 100% reliability. We communicate with our users via our blogs mygene.info, Twitter (@mygeneinfo), and email. The source code is available at github.com/SuLab/mygene.info.

Dissemination. MyGene.info is hosted at mygene.info. Its API doc page is at mygene.info/v3/api and its detailed documentation is at docs.mygene.info. We also provide Python (pypi.python.org/pypi/mygene) and R clients (www.bioconductor.org/packages/release/bioc/html/mygene.html). The mygene R client currently has over 500 monthly downloads, and ranked as one of top 5% downloaded BioConductor packages.

Future relationship to Commons. MyGene.info is the exemplar API used in the smartAPI project developed by the Commons API Interoperability Working Group, which is co-chaired by Dr. Chunlei Wu (PI of MyGene.info project).

4. Sage Synapse

Sage Bionetworks has developed Synapse as an informatics platform dedicated to supporting the large-scale pooling of data, knowledge, and expertise across institutional boundaries to solve some of the most challenge problems in biomedical research.

Community. Sage Synapse targets the computational biology community, as well as the broad biomedical community.

Usability assessment and evaluation. Synapse has both a QA engineer that manages weekly code releases as well as a community manager that handles interactions with researchers who are using the system. We reach out to a sample of our user base annually to assess the direction of the platform and solicit feedback for continuing improvements.

Discoverability. All Synapse source code is open source and available through GitHub (github.com/Sage-Bionetworks); Synapse is primarily deployed through AWS and provided as a Platform as a Service (PaaS); Synapse also allows for minting of DOIs for resources in the system to provide discoverability via external systems.

Dissemination. Synapse is free and accessible to the public through RESTful APIs (rest.synapse.org), via a web portal (www.synapse.org), or via our analytical clients (documentation at docs.synapse.org). Synapse has a total of about 24,000 registered users; approximately 2,400 unique users leverage the system per

month and 900 are "active" users who log in at least 3 times per month.

Future relationship to Commons. Sage Bionetworks has been promoted the concept of the Commons since its inception in 2009 and Synapse was developed to support a "Commons" model. There are a number of ways in which Synapse could be a part of the Commons, but the most straightforward of which is for biomedical researchers to share and describe their digital research assets and their relationships to one another (provenance). These assets can be easily leveraged by third parties who can carry out independent research via any compute infrastructure they like. Synapse also provides a governance system which allows for the sharing of information that is more sensitive.

5. Protein Pipeline

The Protein Pipeline is a complete computational package to calculate protein turnover values. This package encompasses multiple standalone software tools that allows the users to convert raw mass spectrometry files to community standard format (*Raw Converter*, *Raw Extractor*), perform database search for protein identification (*ProLuCID*), and derive protein turnover values using automatic nonlinear model fitting (*ProTurn*).

Community. The Protein Pipeline targets the proteomics community, as well as the broad basic scientific community and the translational clinical science community.

Usability assessment and evaluation. The Protein Pipeline was the first software pipeline available to enable large-scale analysis of protein turnover using deuterium labeling in animal models. Protein turnover studies are increasingly popular in the proteomics community. Prior to ProTurn there was no software tool available to enable deuterium labeling analysis for protein turnover.

Discoverability. *ProTurn* can be discovered via the resource discovery index AZTec and is the top result for the query term "turnover." We have also published an open access "data descriptor" article to promote discoverability and document use cases on Nature Publishing Group's open data journal Scientific Data (www.nature.com/articles/sdata201615). *ProTurn* is highly interoperable with open mass spectrometry data standards and common upstream workflows (search engines). *Raw Extractor* and *Converter* can be downloaded via fields.scripps.edu/yates/wp/?page_id=17.

Dissemination. This pipeline has been published in multiple papers in esteemed journals (Lam et al., J Clin Invest 2014; Lau et al., Sci Data 2016). The software tools in this pipeline are broadly used by the proteomics community. They can be downloaded via fields.scripps.edu/yates/wp/?page_id=17 and www.heartproteome.org/proturn. Additional instructions and test data can be found on ProteomeXchange (PX000561) and Sage Synapse (doi 10.7303/syn2289125).

Future relationship to Commons. We believe the *ProTurn* pipeline will provide essential services for the re-use and re-analysis of any deuterium-labeled protein turnover datasets that may appear in the commons. We will be happy to work with the Commons to ensure interoperability in future use cases.

1. KnowEnG: Knowledge Network Guided Analysis System

We have created a cloud-based infrastructure, called KnowEnG (Knowledge Engine for Genomics), for knowledge-guided analysis of genomics data. The user uploads their data, in the form of a spreadsheet, to the KnowEnG interface and the system performs powerful data mining and machine learning tasks on those data. The unique part of such analysis is that it is carried out while making intelligent use of prior knowledge in the public domain. Such prior knowledge is represented in the form of a massive heterogeneous network called the *Knowledge Network*, which aggregates information from several externally curated databases. The user may choose from several analysis pipelines to deploy on their data. Each pipeline is a complex workflow involving one or more algorithms for data processing and normalization, application of the core machine learning or statistical algorithm, as well as post-processing and visualization.

Community. Biologists and bioinformaticians working with multi-sample genomics data sets (e.g., gene expression, somatic mutation, copy number data sets).

Usability assessment and evaluation. The KnowEnG platform is usable through a web portal. In addition, certain pipelines of the KnowEnG platform are available on Cancer Genomics Cloud (CGC) through CWL descriptors. The functionality of this platform is primarily assessed by utilization of KnowEnG's capabilities in multiple research projects on cancer pharmacogenomics and behavioral neurogenomics. Feedback on the platform is also obtained through two pipelines that are used as part of a lab module in a course on computational genomics taught at UIUC.

Discoverability. This platform is Dockerized, with containers available as Containerized Commons objects in the Docker hub. A paper is presently being prepared for journal publication, describing the overall platform and its components. Additionally, several components of the KnowEnG system are deposited in GitHub. Components of the system have been used in other research projects that are now in the late stages of publication (i.e., in review or preparation). In general, the Knowledge Network (KN) is usable and discoverable in the following ways:

- Containerized pipelines to check, fetch, parse, map, merge, and export data from public biological datasets into Knowledge Network available to developers as GitHub repository.
- Python package to extract gene and gene set mapping and descriptive information from the Knowledge Network Redis database available to developers in GitHub repository.
- This method along with a KN subnetwork fetcher is also available to collaborators as a tool on the Seven Bridges Cancer Genomic Cloud.

Dissemination. A public portal is currently under development, and we are engaging alpha users to perform analysis of their data sets with KnowEnG pipelines. Examples of alpha testers include collaborators at Mayo Clinic, UIUC, and UCLA. Further routes of disseminating the system include:

- Allowing users of TCGA data at the Cancer Genomics Cloud (CGC) to access KnowEnG functionality through CWL and Docker containers.
- Enabling users of Globus Genomics to access KnowEnG functionality through CWL and Docker containers.
- Engaging with other BD2K Centers who have strong dissemination platforms, including LINCS.

Presentations at conferences during the past year include: the Rocky Mountain Bioinformatics Conference 2016, SciDataCon 2016, RSG-DREAM 2016, Individualized Medicine Conference 2016, Biological Data Science meeting at CSHL 2016.

Future relationship to Commons. We are working towards integrating the KnowEnG system with major cloud-based data repositories such as TCGA and LINCS, as part of a bigger ecosystem under the Commons umbrella. Pipelines also available as Containerized Commons objects on Docker Hub. We are taking a leadership role in the Commons Working Group on Workflow Sharing and Docker Registry (WSDR), working closely with scientists from the UCSC BD2K Center and the GA4GH consortium. One pipeline has been selected as prototype for implementation of FAIR principles.

2. ProGENI

The ProGENI algorithm employs random walks with restarts (RWR) to rank genes by their association with drug response variation.

Community. Biologists and bioinformaticians interested in analyzing gene expression data sets in conjunction with phenotype measurements to identify the

most phenotype-relevant genes.

Usability assessment and evaluation. The ProGENI tool is usable through a web portal. An Amazon Web Service (AWS)-hosted Hubzero-based front end provides access. It is also available on the Cancer Genomics Cloud through CWL descriptors. The functionality has been assessed by utilization of ProGENI in the comprehensive discovery of genes whose basal expression levels are predictive of drug response.

Discoverability. A research paper based on ProGENI is under review (presently in revision) at a major journal. A draft of this manuscript has also been deposited on bioRxiv. The source code is available through the KnowEnG GitHub repository. Critically, this tool has been selected as prototype for implementation of FAIR principles (e.g., findable and accessible via the Dockstore, interoperable and reusable via CWL).

Dissemination. We presently allow users of TCGA data at the Cancer Genomics Cloud (CGC) to access ProGENI functionality through CWL and Docker containers. Likewise, users of Globus Genomics can access KnowEnG functionality. Lastly, this work has been presented at a number of conferences.

3. DRaWR

DRaWR is a network-based method for ranking genes or properties related to a given gene set. Such related genes or properties are identified from among the nodes of a large, heterogeneous network of biological information. Our method involves a random walk with restarts, performed on an initial network with multiple node and edge types that preserve more of the original, specific property information than current methods that operate on homogeneous networks.

Community. Biologists and bioinformaticians interested in identifying the shared properties and annotations of genes in an experimentally derived gene set.

Usability assessment and evaluation. The DRaWR tool is accessible through a web portal implemented via an AWS-hosted Hubzero-based front-end. The overall functionality of this algorithm has been assessed by utilization of DRaWR for characterization of gene sets obtained from a variety of research domains, such as human cancer studies, fruitfly development, and mouse social behavior. Feedback was also derived from its usage as part of a lab-module in a course on computational genomics taught at UIUC.

Discoverability. A research paper on DRaWR has been published, and the source code is available through KnowEnG's GitHub repository.

Dissemination. Users of TCGA data at the Cancer Genomics Cloud (CGC) have access to DRaWR through CWL and Docker containers. Similarly, users of Globus Genomics can access KnowEnG functionality. Lastly, this work has been presented at a number of scientific conferences.

4. ClusterEnG and TeachEnG

ClusterEnG (Clustering Engine for Genomics) provides an interface for clustering big data and interactive visualizations including 3D views, cluster selection, and zoom features. ClusterEnG also aims at educating the user about the similarities and differences between various clustering algorithms and provides clustering tutorials that demonstrate potential pitfalls of each algorithm. TeachEnG (Teaching Engine for Genomics) is an online educational tool for reinforcing key concepts in sequence alignment and phylogenetic tree reconstruction. Our instructional games allow students to align sequences by hand, fill out the dynamic programming matrix in the Needleman-Wunsch global sequence alignment algorithm, and reconstruct phylogenetic trees via the maximum parsimony, unweighted pair group method with arithmetic mean (UPGMA) and neighbor-joining algorithms.

Community. ClusterEnG is used in the UIUC Coursera course taught by Dr. Jiawei Han, as well as in the UIUC-Mayo Computational Genomics Course. TeachEnG is being used in a bioinformatics course at Fisk University. Our tracking system indicates that TeachEnG is also being frequently accessed by European universities.

Usability assessment and evaluation. Students without much background in bioinformatics algorithms can interact with these resources to perform state-of-the-art clustering of large data sets and to reinforce their understanding of common bioinformatics concepts. We provide real-time feedback and visualization.

Discoverability and dissemination. We have posted our relevant manuscripts on bioRxiv. Twitter estimates that information about TeachEnG has been disseminated to an upper bound of 42,784 followers and that about ClusterEnG to 7,171 followers so far. The websites are also linked to the main KnowEnG website and the BD2K Training Coordination Center website.

5. Bio-Text Mining Suite

This text mining suite of tools utilizes domain-independent phrase-mining, typing, and entity-extraction; and summarization to develop new methodologies, algorithms, and systems to extract information from biomedical text corpora. This suite of tools includes:

- *AutoPhrase* is a phrase mining method that reduces human efforts on anno-tation or labeling and is adaptable in many languages.
- *ClusType* uses entity recognition and domain-specific typing to construct structured networks from unstructured text corpora.
- *CaseOLAP* is a system for ranking a set of genes or proteins for relevance to subtypes of a disease, based on literature mining. We applied it to study how a set of genes shows differential relevance to different subcategories of cardiovascular disease, such as cardiomyopathies and arrhythmias, and a manuscript on the work is in preparation.

Community. Biomedical researchers involved in literature mining.

Usability assessment and evaluation. Traditional knowledge network construction methods rely on heavy training and costly schema and data annotation; thus, they have poor portability to new domains and languages. We developed automated text mining tools that require less human curation and labeling and effectively mine biological text data to construct structured biological networks. The tools are currently used by UCLA Heart BD2K center for literature mining of cardiovascular diseases using PubMed biomedical text corpora.

Discoverability. The tools have been described in several publications and conference proceedings. The software packages are available on GitHub.

Dissemination. The tools will exist as a pipeline in the KnowEnG system and is currently under development.

1. mCerebrum

mCerebrum is a mobile software platform to support data collection from multiple sensors in phones and wearables (e.g., wrist- and chest-worn sensors, smart toothbrushes, as well as weight and blood pressure monitors) to discover and validate new mHealth biomarkers. mCerebrum supports high-frequency raw sensor data collection in excess of 70+ million samples/day, along with their curation, analysis, storage (2GB/day), and secure upload to cloud. Built-in privacy controls allow participants to suspend/resume data collection from specific sensors. Data science research conducted by MD2K has already resulted in ten mHealth biomarkers, including stress, smoking, craving, eating, activity, and drug (cocaine) use. The entire pipeline of mobile sensor big data – collection, curation, feature extraction, biomarker computation, time series pattern mining, and micro-randomization – has been developed and fully-implemented on the phone to support real-time, biomarker-triggered notifications and interventions.

Community. mCerebrum is being used in seven field studies (smoking, eating, oral health, cocaine use, and congestive heart failure) being conducted at seven unique sites throughout the United States. These studies will involve over 2,000 participants who use the software for a total of over 100,000 person days, resulting in 584,640 hours of high-frequency sensor data, consisting of more than 4.3 trillion data points, for a total of at least 300 TB. These studies support projects from NIBIB, NIDA, NCI, NIMHD, and NIDCR from NIH, as well as IARPA and NSF.

Usability assessment and evaluation. mCerebrum consists of 150,000 lines of code that has undergone rigorous testing at the seven sites collecting data with it. They have submitted over 800 requests for new features and updates, all of which has been incorporated in the currently deployed system over the past year.

Discoverability. mCerebrum is an entirely open-source platform that is hosted on GitHub. It spans 55 software repositories and is also discoverable from the MD2K website in addition to GitHub. It has had over 3,000 software commits from 19 contributors. It has had over 3,000 unique page views from 62 countries.

It connects with a variety of commercial sensors and uploads the data to MD2K cloud. It is easily configured for different researcher studies to meet their unique requirements without modifying the underlying source code.

Dissemination. After two years of development, mCerebrum is now in deployment at seven research studies. It is disseminated via GitHub (19 contributors), via MD2K website, in presentations and talks, in media outlets (it was a cover story at MIT Technology Review in December 2016), as well as in scientific articles.

Future relationship to Commons. mCerebrum is the only general-purpose software platform that supports discovery and validation of digital mHealth biomarkers and sensor-triggered interventions. Work is now underway to incorporate provenance in the data stream to enable access and use of the data collected by mCerebrum for third party research. mCerebrum contributes a unique data category to the Common that is rapidly growing in its importance to biomedical research.

2. Cerebral Cortex

Cerebral Cortex is mCerebrum's big data companion, designed to support study-wide data analysis, visualization, model development, and intervention design. Cerebral Cortex supports the computation of 10 mHealth biomarkers for stress, smoking, craving, eating, lung congestion, heart motion, location, activity, driving, and drug (cocaine) use. It also supports scaling (in the thousands) of concurrent mCerebrum instances and geographically distributed studies targeting diverse health conditions. Cerebral Cortex provides the machine learning model development on population-scale data sets and interoperable interfaces for aggregation of diverse data sources.

Community. *The same community of users as mCerebrum (see earlier)*.

Usability assessment and evaluation. *Usability assessment and evaluation has occurred similarly to mCerebrum (see earlier).*

Discoverability. *Efforts to make this software discoverable by the scientific community are analogous to those for mCerebrum (see earlier).*

Dissemination. *Similar to mCerebrum*. In addition, PSSC Labs (a vendor of private cloud servers), is launching a new product line that will have all the necessary compute, network, and storage capacities needed to run Cerebral Cortex and will come with Cerebral Cortex preinstalled. The investigators purchasing these servers will have an option to have the servers installed at their own premises or keep them at PSSC

Labs premises so it can be managed by the vendor itself. Unlike public clouds, the investigators will not need to pay any monthly fees.

Future relationship to Commons. Cerebral Cortex can be a unique contributor to the Commons. In addition to providing analytics of mobile sensor data to aid knowledge discovery, it introduces provenance for high-frequency mobile sensor data and annotation of all the stages of data processing employed to obtain a mHealth biomarker from raw sensor data. As mobile sensor data and mHealth biomarkers become increasingly common in biomedical research, Cerebral Cortex can be a critical and unique component of the Commons.

3. MotionSenseHRV & EasySense

MD2K has developed two sensors that are used in the center's studies. EasySense is a contactless RF sensor for monitoring heart and lung mo-ion and pulmonary edema that was developed by a team at The Ohio State University led by Emre Ertin. It allows sensing of lung fluid and heart/lung motion non-invasively using micro-radar. When the Microsoft Band was withdrawn from the market, MD2K was unable to find a wrist-worn sensor capable of streaming raw sensor data for an entire day on a single battery charge. As a result, Ertin's team developed a new wrist sensor called MotionSense HRV, which has three types of LED sensors (red, infra-red and green) embedded in its underside (as compared to commercial sensors that only use green LEDs). MotionSense HRV has been tested and produced, and is now being used in studies at seven sites.

Community. *The same community of users as mCerebrum (see earlier)*.

Usability assessment and evaluation. These sensors are deployed widely. They have undergone several rounds of testing prior to their deployment. The results of these testing have resulted in significant improvements in their wearability, usability, data yield, and battery life.

Discoverability. Via MD2K website, scientific articles, talks, and news media.

Dissemination. These products are currently limited to use within the projects collaborating with the MD2K Center. But, they will be made available for purchase at low cost starting end of Summer 2017 so it can be used widely. These sensor provide unique capability that is necessary to conduct biomedical research with mobile sensors that will facilitate discovery and validation of new mHealth biomarkers.

Future relationship to Commons. Data collected by these sensors can be incorporated in the Commons, which should have a capability to store these data and allow access and use by third party researchers.

4. mHealthHUB

The mHealthHUB is an online resource for members of the mHealth community that includes, articles, news, training videos, meeting announcements, templates for IRB languages for sharing high-frequency mobile sensor data, and a discussion forum for discussing development and deployment of MD2K software in research studies and data science research.

Community. The mHealtlhHUB is used by members of the mHealth community, which includes investigators and students in medicine, behavioral science, computing, and engineering.

Discoverability and dissemination. The mHealthHUB is discoverable on the Internet and is promoted in talks, articles and the webinars are promoted as they occur and once they are archived via an email list of interested parties, the BD2K CC and Twitter. The product is discoverable/disseminated via the mHTI website, the mHealthHUB, the MD2K website and through promotion on Twitter as they become available. Since its launch in November 2015 (through 4/11/17), the HUB has recorded 30,889 page views in 11,234 sessions with 6,810 users in 125 countries. We use Google analytics to evaluate metrics.

Use of the webinars and of the recorded lectures from the mHealth Training Institutes is tracked via the MD2K YouTube channel. As of 4/11/17, 7,024 views had been reported on the more than 38 hours of archived material. The product is discoverable/disseminated via the mHTI website, the mHealthHUB, the MD2K website and through promotion on Twitter as they become available.

5. mHealth Summer Training Institute

The annual mHealth Training Institute (mHTI) is sponsored by OBSSR/NIDA and co-sponsored by MD2K. mHTI is a week-long immersion boot camp in mHealth where scholars are trained by experienced thought leaders in mHealth. Scholars participate in lectures in all aspects of mHealth covering core educational

grounding in mHealth perspectives and methodologies. They are grouped in small teams from multiple disciplines to work together on a new mHealth grant, which they present at the end of the institute. They also submit their grant for evaluation by NIH program officers to get feedback. They learn both course materials and soft skills on how to work in a multidisciplinary team. The mHTI helps inculcate the intrapersonal and interpersonal skills and attitudes necessary for transdisciplinary collaborations.

Community. The mHTI is an invitation-only training that solicits applications from any-one involved/interested in mHealth research that has a doctoral-level degree. A total of 35 scholars are selected each year to participate in mHTI. The scholars cover all the major disciplines involved in mHealth that includes medicine, engineering, behavioral science, nursing, statistics, engineering, science, informatics, and several others. mHTI has provided training 70 scholars from 50 institutions and 12 disciplines.

Discoverability and dissemination. The mHTI is widely promoted among the BD2K centers and anyone with a doctoral-level degree and interest in mHealth research is welcome to apply. The mHTI is formally evaluated each year by the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) to gauge its effectiveness. These evaluations, shared with NIH, have shown it to be successful and viewed as beneficial by participants. In addition, each year's mHTI draws more than 100 applicants for the 30 slots.

The mHTI has its own website (mhealth.md2k.org/mhealth-training-institute), which is linked on the MD2K website (https://md2k.org) and the mHealthHUB (mhealth.md2k.org). When applications open each year, it is heavily promoted via the websites, a variety of email lists and Twitter.

mHTI lectures are recorded, and following each year's training, are posted on the mHTI website and the MD2K YouTube channel. As noted under Product 4, these videos (46 to date) along with the MD2K webinars, total more than 38 hours of recorded material that has been viewed more than 7,000 times.

1. OpenSim

OpenSim is open-source software that supports importing and integrating data about movement and the musculoskeletal system. The software supports data from a variety of sources, including 3-D motion capture systems, EMGs, force plates, and currently under development, inertial measurement units (IMUs), such as those found in smartphones.

Community. The software has been downloaded by 160,000 individuals and cited in 1,500 publications. Users include biomechanics researchers and clinicians who treat movement disorders. The software is also used extensively for teaching in high school and college courses around the world.

Usability assessment and evaluation. We test the tool for scientific validity by comparing to a wide range of experimental measurements, along with running a regular suite of regression and unit tests that ensure that errors are not introduced. We test for usability by interviewing users at workshops, monitoring our user forum for common bugs and issues, providing a tool for users to file bugs and request new features, and soliciting regular feedback in the form of user surveys. We also face the challenge of a user base with a wide range of backgrounds and computational expertise. To address this challenge, we provide different interfaces to the software, such as a graphical user interface, a Matlab/Python scripting interface, and a C++ API for developers.

Discoverability. We use online repositories, including SimTK (where we post the OpenSim application, supporting scripts, models, experimental data, and simulation results) and GitHub (where we post source code). SimTK includes a search feature and also automatically recommends similar projects based on users' browsing history. Resources hosted on SimTK are also indexed at the Resource Discovery System (biositemaps.org/rds/index.html) and by DataMed (datamed.org). On GitHub, we use tags to help interested users find our software. We also host a project webpage at opensim.stanford.edu that links users to all resources related to the software. OpenSim provides an API that is documented at simtk.-org/api_-docs/opensim/api_docs/.

Dissemination. We disseminate OpenSim using a broad range of avenues. We provide extensive online documentation in the form of a wiki with guides for users and developers, examples and tutorials, and best practices. We host an active user forum (over 3,000 topics) and have a mailing list with 14,000 subscribers. We also run a range of in-person workshops (15+) and conference tutorials (30+). Different events target beginning or advanced users, and computational or clinical users. In addition, we run courses at Stanford that use OpenSim and provide our teaching materials so that instructors at other institutions can also use OpenSim in their teaching.

Future relationship to Commons. OpenSim could potentially be one of the tools within the Commons. Example functions of how OpenSim could be part of the Commons include: enabling the integration of real-time data from IMUs with other experimental data; connecting with other tools within the Commons either through its own API or via the APIs of other tools; generating simulation data that is hosted within the Commons.

2. Snorkel/DeepDive

Snorkel (the successor to DeepDive) is an open-source system that introduces a new approach for rapidly creating, modeling, and managing data for training predictive systems. It is currently focused on accelerating the development of structured or "dark" data extraction applications for domains in which large labeled training sets are not available or easy to obtain. Examples include biomedical literature and clinical notes. Initial results show that Snorkel – with its use of weakly labeled, noisy training data – can achieve the same performance as fully supervised learning approaches with "gold standard" labeled training data.

Community. Snorkel is a general tool with applicability in many domains. Users come from both academia and industry, including many who are using the software for biomedical applications. Some users are data scientists, and others are biomedical researchers with some data science training. Example biomedical domains where Snorkel is being used include the microbiome, joint replacements, and cancer. Snorkel and DeepDive have together achieved over 1,300 stars and nearly 400 forks on GitHub.

Usability assessment and evaluation. To ensure the functionality of Snorkel/DeepDive, we employ industry standard tools for automatic unit testing and documentation generation. To increase usability, we provide regularly updated tutorials for novice and advanced users in the form of Jupyter notebooks. We assess user needs through a number of mechanisms. Users routinely submit bug reports, feature requests, or provide

other feedback through the GitHub issues system. We also maintain weekly "office hours" for local Stanford collaborators who have more complex questions on active projects. One-on-one interactions with external collaborators provide another source of feedback. We also use our events to solicit feedback from our users about needed features and usability bottlenecks.

Discoverability. To enable discovery of Snorkel, we post the source code on GitHub, with relevant tags, and list the package in the Python Package Index. We also host a project webpage at snorkel.stanford.edu and list it in the SimTK repository, which is indexed at the Resource Discovery System (biositemaps.org/rds/index.html) and by DataMed.

Dissemination. The main Snorkel and DeepDive publications have been cited over 150 times. The software websites receive over 1,000 unique visitors per month. Users can obtain the code for Snorkel through GitHub or install it using the pip installer for Python. To further disseminate Snorkel, we host a number of events, including in-person hackathons, workshops, and courses. We are hosting a workshop specifically for biomedical researchers in July 2017 and have also held two events at two conferences for data scientists. Snorkel was also featured in a recent article in the Biomedical Computation Review, to reach the wide biomedical audience who could benefit from the software.

Future relationship to Commons. Snorkel could potentially be one of the tools within the Commons. It could integrate the vast amount of heterogeneous data available within the Commons to generate predictions.

3. Stanford Network Analysis Platform

Stanford Network Analysis Platform (SNAP) is a general purpose, high performance system for analysis and manipulation of large networks. SNAP is optimized for maximum performance and compact graph representation. It easily scales to massive networks with hundreds of millions of nodes, and billions of edges. It efficiently manipulates large graphs, calculates structural properties, generates regular and random graphs, and supports attributes on nodes and edges. Besides scalability to large graphs, an additional strength of SNAP is that nodes, edges and attributes in a graph or a network can be changed dynamically during the computation.

Community. SNAP has been downloaded over 12,500 times in the last 12 months and is used to understand any type of network, e.g., social networks, biological networks, physical networks (such as communications and roads). Users come from academia and industry and have sufficient programming experience to run data analytics in Python. More sophisticated users are experienced C++ programmers. Their interests span the gamut, from studying genes-drugs association at Baylor College of Medicine to examining C. elegans' neuronal network to analyzing the impact of social networks from a smartphone app on physical activity levels.

Usability assessment and evaluation. To ensure the validity of SNAP's output, we have unit tests for all major components in it. We use a mailing list and GitHub issue trackers to collect maintenance-related issues and feature requests from users. Significant new features are normally the results of new research and advanced state-of-the-art methods that address needs for specific applications being run both internally and externally.

Discoverability. SNAP software and its associated datasets are available from snap.stanford.edu. The code itself is open-source and is available at github.com/snap-stanford. SNAP is also listed on the SimTK repository, which is indexed at the Resource Discovery System and by DataMed. Both a C++ and a Python version of SNAP have been made available. Documentation for the classes and functions within SNAP are also available online.

Dissemination. The most recent distribution package for SNAP is available at snap.stanford.edu). Users can download the package at snap.stanford.edu/snap/download.html or obtain the source code via github.com/snap-stanford. SNAP is integrated in the USC ISI WINGS workflow platform under the Cancer Moonshot initiative. To further disseminate SNAP, we have run workshops/tutorials at data science conferences, such as WWW15.

Future relationship to Commons. SNAP could potentially be one of the tools within the Commons, building and analyzing large networks from data available within the Commons.

4. SimTK

SimTK is a web platform for sharing and collaborating on the development of biosimulation software, models,

and data. It hosts more than 930 projects from researchers around the world, and has had more than 400,000 files downloaded from it. Its infrastructure has enabled members to fulfill the data sharing responsibilities in their grants with no added cost, find collaborators, jointly develop simulation tools, build communities around these tools, and experiment with new forms of collaborations, like grand challenges and open-source model development.

Community. SimTK currently has over 58,000 members, representing both industry and academia. Their interests span diverse areas of biosimulation with a large fraction interested in biomechanics or biophysics.

Usability assessment and evaluation. SimTK has an easy-to-use graphical user interface that enables a user to quickly accomplish what they need to do on the site. To ensure the continued usability of the site, we provide an issue tracker so that users can report bugs and request new features on an on-going basis. We also solicit feedback via user surveys and one-on-one conversations with both current and potential users at events where we promote the site. We monitor a number of metrics to assess both the site's performance (e.g., uptime, CPU usage, memory usage) and impact (e.g., number of projects, number of downloads).

Testing of any new features or bug fixes on the site is iterative, beginning with the web developers testing on their own development machines. Major new features may also be tested in small focus groups, providing feedback to the developers during development. Finally, the updated site is tested by both developers and SimTK users on a staging server, a virtual machine that is a clone of the live server.

Discoverability. We index all the publicly available resources on SimTK at the Resource Discovery System (biositemaps.org/rds/index.html) and by Data-Med (datamed.org/). SimTK is often referenced by members in their publications, and our code is made available on GitHub.

Dissemination. Users can access SimTK by visiting simtk.org. The site is designed to run on Chrome, Firefox, IE, Opera, and Safari on the desktop, as well as on mobile devices. The site currently has 58,000+ members whose interests span diverse areas of biosimulation with a large fraction interested in biomechanics or biophysics. The code for the site is also available on GitHub.

Future relationship to Commons. SimTK could potentially be part of the Commons, serving as a repository for individuals to share data, software, and models and also as a component in researcher workflows. We envision developing functionality so that other tools can automatically deposit the data they generate into SimTK, and also automatically pull data from SimTK to use as inputs.

5. Women in Data Science Conference

The Women in Data Science (WiDS) Conference (widsconference.org) is a 1-day technical conference organized by Mobilize faculty member Margot Gerritsen to inspire, educate, and support women in the field – from those just starting out to those who are established leaders across industry, academia, government, and NGOs. The conference features talks from influential female data scientists, a career panel discussion, plenty of opportunities for networking, and breakout sessions for discussing common topics of interest, including one on "Biomedicine." The conference was held for the second time in February 2017, reaching over 50,000 people through the main event at Stanford University, 75+ regional events across 25+ countries, and a live stream. The theme for the regional event at Rutgers University was "Data Science Applications to Healthcare," and the NIH co-hosted the local event in the DC area, along with the American Statistical Association and the Women in Data Science DC meet-up group.

Community. The event seeks to inspire, educate, and support women in the field of data science. This year the conference reached over 50,000 people through the main event at Stanford University, 75+ regional events across 25+ countries, and a live stream. The main event itself was attended by 400+ individuals from 31 universities and 114 companies. The organization of the event has broad support from companies (e.g., Google, Microsoft), universities, professional societies (e.g., American Statistical Association), and local meet-up groups.

Assessment. WiDS clearly fulfills a critical need in the data science community and has already had worldwide impact. To ensure the continued value, or "usability," of the conference, we gather feedback from attendees both in follow-up surveys and during registration. The conference steering committee is also composed of individuals representing diverse organiza-

tions where data science plays a role, ensuring the applicability of the conference for different fields.

Discoverability and dissemination. Discoverability is achieved through a strong online presence, including a website, YouTube videos, Twitter, a Facebook page, and a LinkedIn group. During WiDS 2017, the conference was trending on Twitter all day. All the presentations from WiDS 2015 and WiDS 2017 have been made available on YouTube and continue to be promoted on social media. Our Ambassadors program also plays a critical role in disseminating WiDS, as Ambassadors expand our reach locally and through their own existing social media channels. Press coverage and the event's co-sponsors, such as SAP, Microsoft, Google, and Walmart Labs, provide additional dissemination.

1. Sync for Science (S4S)

S4S offers a streamlined, e-commerce-like user experience for donating data from any electronic health record (EHR) system by making use of HL7 FHIR to represent clinical data and OAuth 2.0 for authorization. S4S leverages the HIPAA requirement that patients must be able to access their own EHR data, and the Meaningful Use Stage 3 and 2015 EHR Certification programs that require patients be able to share a Common Clinical Data Set with software apps of their choice. S4S design incorporates input from the NIH, the ONC, and Office for Civil Rights to be sure it satisfies rigorous technical, operational, ethical, and legal requirements for patients to release their personally identifiable health information to research studies.

Community. Anyone wishing to donate EHR data to research projects, including the Precision Medicine Initiative's All of Us Research Program. Developed in partnership with major EHR vendors.

Usability assessment and evaluation. Currently being tested at seven health provider organizations (HPOs).

Discoverability. Because S4S is designed for use by individual volunteers or health provider organizations and will be released as open source software, anyone wishing to stand up a project requiring participant recruitment will be able to download and use this from GitHub. A full national campaign will be used to present this tool whose first application will be for the PMI All of us Project.

Dissemination. As this workflow has been developed in partnership with the major EHR vendors for use with All of Us, the commercial sector has already endorsed the application and will be using it to provision patient data to the All of Us Coordinating Center which will solicit the first 1M cohort. Presentations planned for AMIA and other professional societies as S4S comes on line.

Future relationship to Commons. To the extent that the Commons imagines itself serving as the repository of PHI for general access by the research community, S4S will provide the gold standard for doing so.

2. PIC-SURE (RESTful) API

Addressing concerns in biomedical research about reproducibility, particularly in "big data" projects, we have engineered the standardization of this RESTful API such that it can be called from a multiplicity of different programming languages and within electronic notebooks (specifically illustrated via open source Jupyter notebook). This standardized web API provides a single programmatic interface to foster the incorporation of, and enabling access to, multiple heterogeneous patient-level clinical, omics and environmental datasets. This system embraces the idea of decentralized datasets of varying types, and the protocols used to access them, while still providing a simple communication layer that can handle querying, joining, and computing on.

Community. Any researcher who wishes to combine and analyze different data modalities without knowing the myriad details of the underlying data structures of each data source.

Usability assessment and evaluation. This platform and API is available both as a freely downloadable application with accompanying instructions and as a resource already implemented on several data sources including an NHANES data set consolidated from the CDC and as a wrapper for the ExAC database.

Discoverability. Multiple public presentations to the NIH, the research community, and shortly through publication; via our PIC-SURE website and GitHub; through our collaborators at CCD (an interoperability project) and CountEverything supplement. We also use it in NCATS GRDR platform. Nature Scientific Data paper, 2016 for the NHANES application.

Dissemination. Documentation is available online at: bd2k-picsure.hms.harvard.edu. Our NHANES instance of this product currently has 1,042 users. The platform is also a core element in a course offered by Harvard University (dbmi.hms.harvard.edu/education/-courses/bmi-705; Course Director, Paul Avillach). Lastly, a PIC-SURE API datathon was held in March 2017 around the PIC-SURE API/platform.

Future relationship to Commons. Directly relevant to the Commons vision of provisioning large and varied data sets to the research community via an easy to use and security compliant interface.

3. NHANES Database

The National Health and Nutrition Examination Survey (NHANES) is a multimodal epidemiological dataset (individual-level environmental, clinical and physiological status) compiled and made publically available by the CDC. A subset of 41K individuals has been extracted and loaded into an i2b2/tranSMART application

(nhanes.hms.harvard.edu) and wrapped with our RESTful API and made freely available (pic-sure.org/-software/nhanes-api-and-dataset-explorer-app/help/-the-picsure-way)

Community. Biomedical researchers.

Usability assessment and evaluation. Previously, elements of the NHANES dataset, while publically available, were not compiled or integrated into a common data structure that could be used by researchers lacking advanced coding skills.

Discoverability. A Nature Scientific Data paper was published in 2016. This product is described through our PIC-SURE and Harvard's DBMI Websites. A number of public presentations/posters have taken place at the BD2K All Hands Meetings (AHMs). The NHANES dataset is used in a number of places, including as core elements of graduate-level courses at Harvard DBMI (dbmi.hms.harvard.edu/education/courses/bmi-704, chiragjpgroup.org/exposome-course, Course Director: Chirag Patel; dbmi.hms.harvard.edu/education/courses/bmi-705, Course Director: Paul Avillach) and a workshop at Emory University).

Dissemination. Currently 1,042 users are registered and using the dataset. The dataset itself was documented in a Nature Scientific Data paper, which has garnered a 59 Altmetric score, placing it in top 5th percentile of all papers tracked by Altmetrics and top 12th percentile of all papers of a similar age in Nature Scientific Data.

Future relationship to Commons. As an exemplar of how multimodal datasets can be made broadly available in a user friendly framework.

4. Exposome Data Warehouse

We are creating and will make publicly available the *Exposome Data Warehouse*, a comprehensive data warehouse containing all data from the EPA Air Data hourly air pollution monitoring program, daily weather data from NOAA (weather.gov and other sources), socio-demographic data from the US Census (going back to 2000s) in a single unified data warehouse. We will also develop an application programming interface to allow access to these data from external resources, such as health records that exist within in a proprietary and sand-boxed clinical warehouse. Second, we will develop a web-browsable application to allow for facile exploration of the data and download. Third, we will provide scripts in python and R programming language to allow investigators to build their own pipelines to use the data in their own clinical data repositories.

Community. Biomedical researchers.

Usability assessment and evaluation. While there are many publicly available data resources that ascertain population-level environmental (e.g., exposure to air pollution), sociodemographic (e.g., income), and ecological (e.g., climate), we lack a centralized repository and software to access this information in clinical data repositories, such as: is weather associated with asthma? Is socioeconomic status associated with hospital readmission? Is air pollution associated with risk for heart attack admission? However, there is no data platform that integrates these environmental data resources because they are heterogeneous: they capture different types of data (e.g., average income at a zip code versus total lead and particulate matter found at a specific latitude and longitude) at different geographical and temporal resolutions.

Discoverability. Efforts to make this data warehouse known are threefold:

1. Through conferences and workshops, such as the NIEHS investigator day (North Carolina, 2016), National Science Foundation Big Data Spoke workshop (New York, NY; Feb 24 2017), National Science Foundation Big Data Investigators meeting (Washington DC; March 17, 2017);
2. We are building a web application for browsing and downloading of the data and plan to publish this as a citable product in Nature Scientific Data or JAMIA; and
3. We will document the APIs in GitHub for public consumption.

Dissemination. Currently, the Exposome Data Warehouse is being evaluated internally and we have three proof-of-concept scientific papers that utilize the data resource.

Future relationship to Commons. This is very much related to other projects in the commons as it contains highly impactful, federally funded, gold standard data resources that are packaged comprehensively into a unified data resource for research use. We anticipate this will be a part of the Commons in the future.

5. Global Rare Diseases Registry

This is a web-based resource that aggregates, secures, and stores de-identified patient information from 10 different registries for rare diseases (5,277 patients

with 178 different rare diseases) all in one place. This platform allows for the wider dissemination of data collected by individual rare disease registries as well as the increased accessibility of data for researchers conducting intra-registry and cross-registry queries in the genesis of hypotheses for potential study. It uses the PIC-SURE RESTful API described elsewhere, international ontologies for semantic interoperability, mappings to UMLS Concept Unique Identifiers, and mappings to 55 different GRDR Common Data Elements.

Community. Researchers interested in looking across several rare disease repositories.

Usability assessment and evaluation. Designed to use the widely adopted i2b2-TranSMART platform for aggregating, securing, and storing patient level data, this registry has been explicitly configured for querying by researchers not proficient in the computer coding skills typically required to access multiple, related data sources originally developed on different platforms. Functionality was confirmed by reproducing published findings.

Discoverability. This registry has been widely presented at national and NIH meetings (grdr.hms.harvard.edu, s3.amazonaws.com/hms-dbmi-docs/GRDR-_Quick_guide.pdf, s3.amazonaws.com/hms-dbmi-docs/GRDR_User_Guide.pdf, vimeo.com/151171529); publications are in process.

Dissemination. The registry is freely accessible at https://grdr.hms.harvard.edu, with a total of 627 registered users and 2,130 sessions since release.

Future relationship to Commons. A perfect example of the Commons vision of organizing and provisioning complex data sets for the scientific community, especially as these original datasets were originally commissioned by the NIH but were largely previously inaccessible due to lack of interoperability.