iDASH

integrating Data for Analysis, Anonymization and Sharing
A National Center for Biomedical Computing
http://idash.ucsd.edu

iDASH is the newest National Center for Biomedical Computing, funded in late 2010. Its goal is to develop infrastructure, services, and tools to allow privacy-preserving data sharing. The Working Group on Data and Informatics has recently made recommendations to the NIH Advisory Committee to the NIH Director to accelerate research: Data and meta-data should be shared, incentives should be offered to those who share data, and investments in user training and infrastructure need to be coordinated to ensure efficient utilization of resources. On the training side, the number of informatics professionals and researchers needs to increase. On the infrastructure side, a backbone for data and software sharing needs to be implemented through a network of biomedical computing centers. iDASH addresses both challenges. We are exploring how biomedical researchers and healthcare providers can remain focused on their activities and outsource data storage, de-identification, annotation, curation, some analysis, and distribution to reliable third parties/processes.

Research Impact

Despite its young age, iDASH has already developed different models, tools and infrastructure for data sharing that allows it to broker the relationship between data owners and data users. The infrastructure, service and tools developed by iDASH protect the privacy of individuals and of institutions, and provide meaningful information for patients to make informed decisions about sharing their data and specimens. We have developed a HIPAA-compliant hardware and software infrastructure at the San Diego Supercomputer Center that combines over 300 terabytes of cloud storage with high performance computing to allow computation on sensitive data such as human genomes and clinical records. Our infrastructure is supporting advanced research in cloud computing and privacy technology that involves commercial and private HIPAA-compliant clouds.

EXAMPLE 1: We have deployed cloud storage computing and associated policy infrastructure for researchers to share data. We are hosting several data sets including different data modalities (whole genomes, transcriptome data, images, specialty reports, clinical trial data, structured and unstructured clinical data) in our annotated data repository, including many related to Kawasaki Disease, a relatively rare disease of unknown etiology for which we have one of the world's largest data collections, which is annotated and mapped to public ontologies using tools from NCBO and other tools that we have developed. **EXAMPLE 2:** Our data sharing models also include facilitating access to federated databases. We host the hub for five University of California health systems, a collection of 11 million patients. Our tools complement our implementation of i2B2 software for count queries with analytical software for privacy-preserving predictive model building. We have also enabled policy-based data exchanges by developing a legal framework of data-use agreements (DUA) between both (a) data providers and iDASH as data custodian (i.e., honest broker similar to an escrow service), and (b) data recipients and iDASH. These DUAs allow the provider to precisely specify what is shared and when (e.g., embargo prior to article publication), the sensitivity of the data (e.g., identified vs. de-identified), and restrictions on who can access the data with a fine control. We have executed over 15 DUAs and this number is increasing fast since our deployment in March 2012. We also developed an electronic informed consent tool to allow patients to express their preferences towards the use of their data and institutions to automate solutions.

Community Resources/Software/Collaborations

We provided letters of support for 14 grant applications to NIH, NSF, PCORI, and private foundations. We participated in 4 linked R01 proposals, and had one trainee receive a K99/R00 award for his work on privacy technology. We have over 22 data sets from different studies and 11 software tools for privacy protection, data analysis, annotation, and genome query. Our new web site (http://idash.ucsd.edu) containing the data sets and related tools has been up since April 2012 and has been visited by over 3,000 unique users, with over 6,000 views. We are collaborating directly on data access sharing with federal (Tennessee VA), public (UC system) and private institutions.

Dissemination and Training

We were invited by the OSTP to present iDASH at a White House announcement for Big Data, and received several invitations to speak about iDASH internationally. We sponsored 15 free iDASH webinars from speakers in academia, industry, and government, which were all attended by over 30 individuals. We provided webinar support for 15 journal clubs featuring *J Amer Med Inform Assoc* (JAMIA) editor's choice freely accessible articles. Attendance ranged from 40 to 130 remote attendees per session. We organized 5 workshops (Imaging Informatics, Natural Language Processing, High Performance Computing, Privacy Technology, and Ethical, legal and policy perspectives of data sharing), with attendances averaging about 40 participants.

During the past 2 years, we have trained 65 individuals: 6 postdocs, 10 graduate students, and several short-term trainees, including 3 who were under-represented minorities in science. Approximately half of our trainee pool is female. Our internship program involves high school students (paid from another source), undergrads, graduate students from five different states, and a multitude of public and private universities. The trainees come primarily from computer science/engineering and biomedical backgrounds. They have produced over 30 posters and 27 journal publications and their scientific presentations are available at our web site. During the first and second years of iDASH, we have published 14 and 22 journal papers, respectively.

Supported by grant U54HL108460