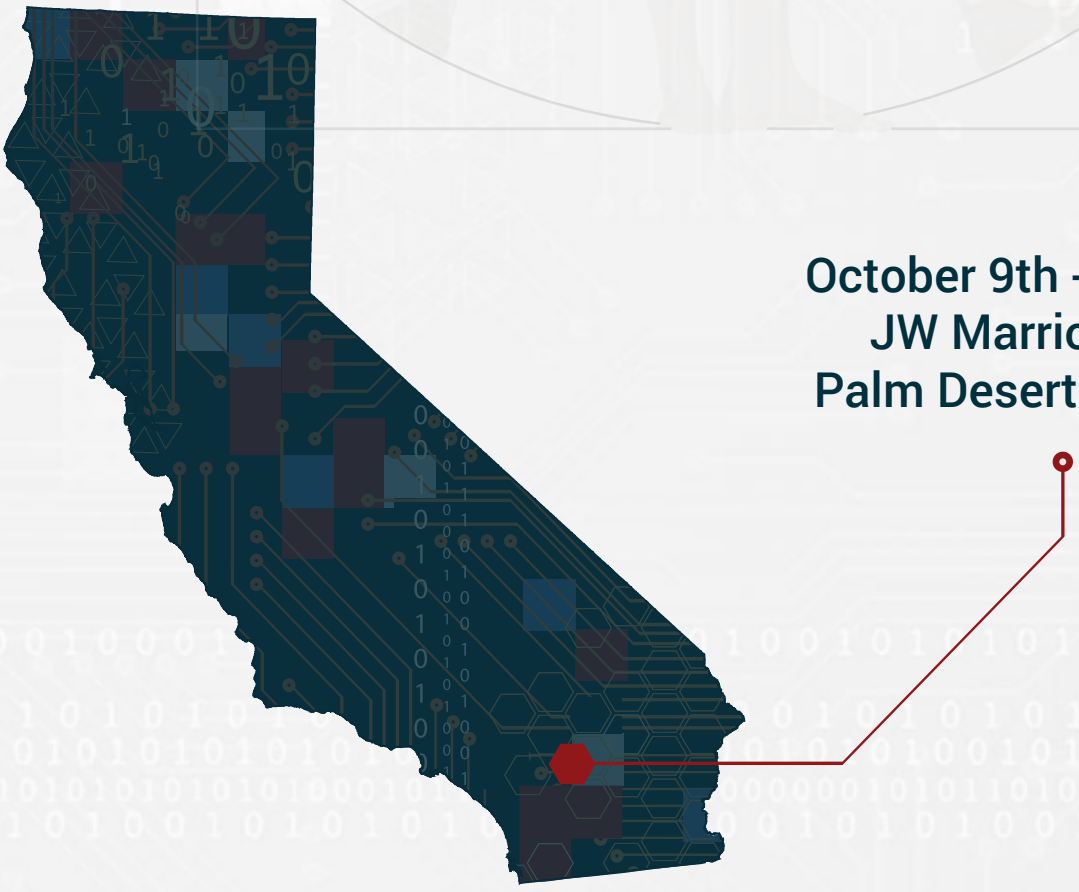
CALIFORNIA

BIG DATA

BIOMEDICAL WORKSHOP



October 9th -10th, 2015
JW Marriott Hotel
Palm Desert, California

HOSTED BY

BDDS
BIG DATA FOR DISCOVERY SCIENCE

ENIGMA

WELCOME

Welcome and thank you for your participation at the California Big Data Biomedicine Workshop held at the JW Marriott Hotel, here in Palm Desert, California! We are excited to have you join us.

California has been the beneficiary of a number of the NIH's BD2K grants – from U54 centers of excellence, coordinating centers, to training project grants. Given that the NIH expects us all to cultivate "consortium activities," we felt that a get together to share ideas and plan for such interactivities would be in order.

In our discussions over the next day and a half we seek to introduce ourselves, our projects, and to specifically identify points of synergy which we can utilize in joint activities and training opportunities. In so doing, we will conduct our interactions around working groups. We are eager to work with you to find answers to these and other questions.

Sincerely yours,

Arthur W. Toga, Ph.D.
Carl Kesselman, Ph.D.
Paul Thompson, Ph.D.
John Darrell Van Horn, Ph.D.

AGENDA

FRIDAY, OCTOBER 9TH	
7:30-8:30	Breakfast
8:30-8:40	Welcome and Opening Remarks
8:40-9:10	BD2K: The View from the NIH
9:10-9:25	Project Overview: MOBILIZE
9:25-9:40	Project Overview: ENIGMA
9:40-9:55	Project Overview: bioCADDIE
9:55-10:10	Project Overview: BDDS
10:10-10:20	Coffee Break
10:20-10:35	Project Overview: HEART BD2K
10:35-10:50	Project Overview: CEDAR
10:50-11:00	Breakout Sessions - Charge and Organization
11:00-12:30	Breakout Sessions
12:30-2:00	Lunch
2:00-3:00	Breakout Sessions Report Out/Discussion
3:00-3:30	Preparation for Saturday – Synergies & Actionable Commonalities
3:30-5:00	Software and Tool Demos & Coffee Break
SATURDAY, OCTOBER 10TH	
7:30-8:30	Breakfast
8:30-8:45	Review Synergies & Actionable Commonalities Organizing Breakout Sessions
8:45-10:15	Breakout Sessions
10:15-10:30	Coffee Break
10:30-11:15	Report Outs from Synergies & Actionable Commonalities Breakout Sessions
11:15-12:00	Next Steps, Assignments, Milestones
12:00	Adjourn

WORKING GROUPS

REVIEWING SCIENTIFIC TOOLS & WORKFLOW TECHNOLOGIES

Discussion Leaders: Carl Kesselman and Ivo Dinov

Charge is to determine:

- What software tools are being used and/or are being newly developed?
- What development platforms have been chosen and why?
- Is there a catalogue of these tools or can one be created?
- What Big Data workflow technologies are being brought to bear?

COMPUTING, DATA & STORAGE INFRASTRUCTURES

Discussion Leader: Ian Foster

Charge is to determine:

- Is there a CA-based network and computing infrastructure in place to facilitate Big Data exchange?
- What data sources are being used?
- What data types do these encompass?
- How does one gain access?

ESTABLISHING EDUCATIONAL & TRAINING EXPERIENCES

Discussion Leader: Jack Van Horn

Charge is to determine:

- Can we develop CA BD2K educational opportunities?
- Can we have an invited speaker series exchange between centers?
- Can we share student exchange 'experiences' across one or more CA BD2K centers?
- Can a shared stipend model be established?

MATHEMATICS OF BIG DATA ALGORITHMS

Discussion Leaders: Paul Thompson and Peipei Ping

Charge is to determine:

- What are the mathematical approaches being used to tackle high dimensional data across CA-BD2K centers, sites, and initiatives?
- What are the common mathematical themes?
- What have we learned in terms of what works and what fails?

FRIDAY GROUPS

Petros Petrosyan
Peter Rose
Sam Hobel
Ian Bowman
Ravi Madduri
Mike D'Arcy
Arthur Toga
Eric Deutsch
Vivien Bonazzi
Vinay Pai
Judy Pa
Chunlei Wu

Jeff Grethe
Magali Champion
Mu Zhou
William Hsu
Nova Smedley
Darvin Yi
Houri Hintiryan
Edward Lau
Jennifer Couch
Ben Heavner
Dan Moyer
Nathan Price

Kyle Chard
Naveen Ashish
Michel Dumontier
Hongwei Dong
Kristi Clark
Jennifer Larkin
Tevfik Umut Dincer
Christina Liu
Clio Gonzalez-Zacarias
Andrew Su
Jennifer Hicks
Sonynka Ngosso

Neda Jahanshad
Daijiang Zhu
Gautam Prasad
Olivier Gevaert
Michelle Dunn
Erica Rosemond
Howard Choi
Maggie Pui Yu Lam
Stacia Friedman-Hill
Purvesh Khatri

BIG DATA FOR DISCOVERY SCIENCE

http://bd2k.ini.usc.edu/

The University of Southern California
Contact PI: Arthur W. Toga
Grant Number: 1U54EB020406-01

Researchers at the Big Data for Discovery Science Center will focus on proteomics, genomics, and images of cells and brain collected from patients and subjects across the globe. They will enable detection of patterns, trends and relationships among these data with user-focused data management, sophisticated computational methodologies, and leading-edge software tools for the efficient large-scale analysis of biomedical data. Interactive visualization tools created at this center will stimulate fresh insights and encourage the development of modern treatments and new cures for disease.

Modern biomedical data acquisition, from genes to cells to systems, is producing exponentially more data due to increases in the speed and resolution of data acquisition methods. Yet, "big data" is a moving target. What is considered big data today will be relatively "small data" tomorrow. Moreover, singularly large data sets arise from the efforts of single laboratories or are accumulated from a collection of more modest studies across common or heterogeneous study protocols. Simply having large-scale biomedical data and making it available online, however, is not a means to an end but only the next step in turning data into actionable knowledge. Our Big Data for Discovery Science (BDDS) Center, has the following aims: 1) create a user-focused graphical system to dynamically create, modify, manage and manipulate multiple collections of big datasets, 2) enrich next generation "Big Data" workflow technologies coupled to modern computation and communication strategies specifically designed for large-scale biomedical datasets, 3) develop a knowledge discovery interface to enable modeling, visualizing, and the interactive exploration of Big Data. In addition to these overarching aims, the goals of this BDDS Center include training and consortium activities. Here we will create university-level degree programs in big data informatics, develop annual workshops on strategies for big data best practices, and contribute to national BD2K consortium efforts. The innovations of our BDDS Center include: 1) providing a novel data science framework for characterizing and big data as a shared resource either singularly or collectively, 2) deriving novel computer algorithms for the joint processing o multi-modal data with an emphasis on the challenges that big data present for computation, 3) designing and deploying a unique data management system focused on the user experience which is ontology agnostic, easy to use, and puts the data first, 4) providing enhanced technologies for remote data access, scientific workflow construction, and cloud-based computation on big data sets, 5) providing compelling means for big data set visualization, interaction, and hypothesis generation. Building on these technologies, we will construct and validate tools so that they may be translated to any biological system or biomedical research domain. Our team is comprised of leading neuroscience, biology, and computer science researchers, with expertise in large-scale biomedical data, experience with the present challenges and promise of big data, and a demonstrable history of delivering unique computational resources, thereby insuring big data solutions which promote a "science of discovery". PUBLIC HEALTH RELEVANCE: The overarching goal of our BDDS Center is to ease the management and organization of biomedical big data and accelerate data-driven discovery by eliminating or reducing three distinct barriers to effective discovery science: complexity with respect to physical distribution and heterogeneity, scalability of analysis, and ease of access and interaction with big-data and associated analytic methods. These issues are fundamental to discovery science and transcend the specifics of the research question as we span levels of scale from cells to organs to systems, and integrate data from imaging, genetics, "omics," and phenotypes.

Program Officer: Vinay Pai
Science Officers: Stacia Friedman-Hill, Christina Liu, Keyvan Farahani, Thomas Radman

CENTER FOR BIG DATA IN TRANSLATIONAL GENOMICS

https://genomics.soe.ucsc.edu/bd2k

The University California Santa Cruz
PIs: David H. Haussler, David Patterson, and Laura Van't Veer
Grant Number: 1-U54HG007990-01

The Center for Big Data in Translational Genomics is a multinational collaboration between academia and industry that will create data models and analysis tools to analyze massive datasets of genomic information. Such tools can be used for analysis of the genomes and the gene expression data from thousands of individuals to uncover the contribution of gene variants to disease, with an initial focus on cancer. This knowledge will be instrumental in the development of precision diagnostic and treatment methods.

The Center for Big Data in Translational Genomics is a multi-institution partnership coordinated by the University of California at Santa Cruz to create scalable infrastructure for the broad application of genomics in biomedicine. Our U.S. partners include UC San Francisco, UC Berkeley, Oregon Health Science University, Caltech, and several major big data companies. International partners include the European Bioinformatics Institute, the Sanger Centre, the Ontario Institute for Cancer Research and a computer systems provider. The Center will make software solutions interoperable through the development of standard Application Programming Interfaces (APIs) and tools at multiple levels, from raw sequence data to genetic variation and functional data, through to systems, pathways and phenotypes. The overriding goal is to create implementations capable of handling genomics datasets that are orders of magnitude larger than those that can now be handled. The APIs and all academic reference implementations will be open source, while several major corporate partners not funded by the project will provide proprietary implementations, creating a competitive ecosystem of interoperable big data genomics software. All-comers extensive benchmarking will be performed on all implementations within and external to our center to identify best-of-breed and results made broadly available. Design will be in part driven by the needs of a diverse set of separately funded specific biomedical projects that will serve as pilots. These include the Pan-Cancer whole genome analysis project of the International Cancer Genomics Consortium to analyze 2,000 cancer genomes, the UK10K project to analyze 10,000 personal genomes, the UCSF-led I-SPY2 adaptive breast cancer trial, and the omics-guided leukemia therapy project BeatAML at Oregon Health Sciences University. PUBLIC HEALTH RELEVANCE: At least half of all diseases have a substantial genomic component, often including contributions from the millions of individually rare but collectively common genetic variations. Only by studying the genomes and transcriptomes of very large numbers of individuals will scientists have the statistical power to discover and understand this vital aspect of the genomic contribution to disease. For this it is essential that genomics is brought into the big data era, so that analyses of huge datasets is possible and precision diagnosis and treatment based on genomic information is widely deployed.

Program Officer: Lisa Brooks
Science Officers: Jerry Li, Weiniu Gan, Dawei Lin, Jane Ye, Heidi Sofia

CENTER FOR EXPANDED DATA ANNOTATION AND RETRIEVAL (CEDAR)

http://metadatacenter.org/

Stanford University
PIs: Mark A. Musen
Grant Number: 1U54AI117925-01

The ability to locate, analyze, and integrate Big Data depends on the metadata that describe the content of data sets. The Center for Expanded Data Annotation and Retrieval (CEDAR) will facilitate automated annotation of data with high quality metadata by generating community-based metadata standards and a metadata repository for training learning algorithms to develop metadata templates. These templates will initially be evaluated, validated, and adapted with the NIAID ImmPort multi-assay data repository and other data repositories.

The Big Data revolution requires that biomedical scientists be able to locate, analyze, and integrate the large datasets that now pervade biomedicine. Such work is possible only when experimental datasets are made available online and when they are annotated with metadata that explain how the data are organized, what the data represent, and how the data were collected. The Center for Expanded Data Annotation and Retrieval (CEDAR) will take advantage of the recent growth in community-driven metadata standards to develop innovative computational methods to ease the authoring and use of metadata annotations. Our specific aims focus on working with communities of investigators to standardize descriptions of the data generated through biomedical studies; creating a computational collective for development, evaluation, use, and refinement of metadata templates for describing laboratory studies; developing a comprehensive and open repository of metadata that will inform the learning algorithms that will drive much of our Center's technology; training the biomedical community in the use of metadata and in CEDAR's resources; and evaluating our work in the context of ImmPort, an NIAID-supported multi-assay data repository that will offer end-to-end opportunities to demonstrate and validate our ideas. We anticipate a growing community of users, starting with the Human Immunology Project Consortium, then the BD2K Center Consortium, then the Stanford Digital Repository, growing until we have developed a wide user base leading to measurable changes in the quality of the metadata used to annotate online datasets. The Overall description of our project provides a synopsis of CEDAR's activities and overall specific aims. PUBLIC HEALTH RELEVANCE: The ability to locate, analyze, and integrate Big Data depends on the metadata that describe data sets and the experiments that have been performed. This project will facilitate annotation of data with high quality metadata. The results of our work will lead to better data and, thus, to better science. Ultimately, such results will lead to better health.

Program Officer: Maria Giovanni
Science Officers: Allen Dearry, Valerie Florance, Quan Chen, Ashley Xia, Punam Mathur

A COMMUNITY EFFORT TO TRANSLATE PROTEIN DATA TO KNOWLEDGE: AN INTEGRATED PLATFORM

http://www.heartbd2k.org/

The University of California Los Angeles
PIs: Peipei Ping, Merry Lindsey, Andrew Su, and Karol Watson
Grant Number: 1U54GM114833-01

The UCLA Center of Excellence for Big Data will embark on the project, A Community Effort to Translate Protein Data to Knowledge: An Integrated Platform, in order to fundamentally alter biomedical research culture to enable full employment of technological modeling innovations, such as crowdsourcing to biomedical Big Data analysis. The goal of this center is to democratize data research to include non-computational scientists and individuals and to apply innovative global community-driven data integration and modeling methods to address challenges involved in the study of protein structure, function, and networks with a focus on cardiovascular research.

The inception of the BD2K Initiative is a testament to the foresight of NIH and our community. Clearly, the future of biomedicine rests on our collective ability to transform Big Data into intelligible scientific facts. In line with the BD2K objectives, our goal is to revolutionize how we address the universal challenge to discern meaning from unruly data. Capitalizing on our investigators' complementary strengths in computational biology and cardiovascular medicine, we will present a fusion of cutting-edge innovations that are grounded in a cardiovascular research focus, encompassing: (i) on-the-cloud data processing, (ii) crowdsourcing and text-mining data annotation, (iii) protein spatiotemporal dynamics, (iv) multi-omic integration, and (v) multi-scale clinical data modeling. Drawing from our decade of experience in creating and refining bioinformatics tools, we propose to amalgamate established Big Data resources into a generalizable model for data annotation and collaborative research, through a new query system and cloud infrastructure for accessing multiple omics repositories, and through computational-supported crowdsourcing initiatives for mining the biomedical literature. We propose to interweave diverse data types for revealing biological networks that coalesce from molecular entities at multiple scales, through machine learning methods for structuring molecular data and defining relationships with drugs and diseases, and through novel algorithms for on-the-cloud integration and pathway visualization of multi-dimensional molecular data. Moreover, we propose to innovate advanced modeling tools to resolve protein dynamics and spatiotemporal molecular mechanisms, through mechanistic modeling of protein properties and 3D protein expression maps, and through Bayesian algorithms that correlate patient phenotypes, health histories, and multi-scale molecular profiles. The utility and customizability of our tools to the broader research population is clearly demonstrated using three archetypical workflows that enable annotations of large lists of genes, transcripts, proteins, or metabolites; powerful analysis of complex protein datasets acquired over time; and seamless aggregation of diverse molecular, textual and literature data. These workflows will be rigorously validated using data from two significant clinical cohorts, the Jackson Heart Study and the Healthy Elderly Longevity (Wellderly). In parallel, a multifaceted strategy will be implemented to educate and train biomedical investigators, and to engage the public for promoting the overall BD2K initiative. We are convinced that a community-driven BD2K initiative will best realize its scientific potential and transform the research culture in a sustainable manner, exhibiting lasting success beyond the current funding period.

Program Officer: Susan Gregurick
Science Officers: Pothur Srinivas, Weiniu Gan, Sal Sechi

ENIGMA CENTER FOR WORLDWIDE MEDICINE, IMAGING, AND GENOMICS

http://enigma.ini.usc.edu/

The University of Southern California
PIs: Paul M. Thompson
Grant Number: 1U54EB020403-01

The ENIGMA Center for Worldwide Medicine, Imaging and Genomics will incorporate the scientific acumen of more than 300 scientists worldwide, and their biomedical datasets, in a global effort to combat human brain diseases. This center will develop computational methods for integration, clustering, and learning from complex biodata types. This center's projects will help identify factors that either resist or promote brain disease, and those that help diagnosis and prognosis, and will also help identify new mechanisms and drug targets for mental health care.

The ENIGMA Center for Worldwide Medicine, Imaging and Genomics is an unprecedented global effort bringing together 287 scientists and all their vast biomedical datasets, to work on 9 major human brain diseases: schizophrenia, bipolar disorder, major depression, ADHD, OCD, autism, 22q deletion syndrome, HIV/AIDS and addictions. ENIGMA integrates images, genomes, connectomes and biomarkers on an unprecedented scale, with new kinds of computation for integration, clustering, and learning from complex biodata types. ENIGMA, founded in 2009, performed the largest brain imaging studies in history (N>26,000 subjects; Stein +207 authors, Nature Genetics, 2012) screening genomes and images at 125 institutions in 20 countries. Responding to the BD2K RFA, ENIGMA'S Working Groups target key programmatic goals of BD2K funders across the NIH, including NIMH, NIBIB, NICHD, NIA, NINDS, NIDA, NIAAA, NHGRI and FIC. ENIGMA creates novel computational algorithms and a new model for Consortium Science to revolutionize the way Big Data is handled, shared and optimized. We unleash the power of sparse machine learning, and high dimensional combinatorics, to cluster and inter-relate genomes, connectomes, and multimodal brain images to discover diagnostic and prognostic markers. The sheer computational power and unprecedented collaboration advances distributed computation on Big Data leveraging US and non-US infrastructure, talents and data. Our projects will better identify factors that resist and promote brain disease, that help diagnosis and prognosis, and identify new mechanisms and drug targets. Our Data Science Research Cores create new algorithms to handle Big Data from (1) Imaging Genomics, (2) Connectomics, and (3) Machine Learning & Clinical Prediction. Led by world leaders in the field who developed major software packages (e.g., Jieping Ye/SLEP), we prioritize trillions of computations for gene-image clustering, distributed multi-task machine learning, and new approaches to screen brain connections based on the Partition Problem in mathematics. Our ENIGMA Training Program offers a world class Summer School coordinated with other BD2K Centers, worldwide scientific exchanges. Challenge-based Workshops and hackathons to stimulate innovation, and Web Portals to disseminate tools and engage scientists in Big Data science. PUBLIC HEALTH RELEVANCE: The ENIGMA Center for Worldwide Medicine, Imaging and Genomics is an unprecedented global effort uniting 287 scientists from 125 institutions and all their vast biomedical data, to work on 9 major human brain diseases: schizophrenia, bipolar disorder, major depression, ADHD, OCD, autism, 22q deletion syndrome, HIV/AIDS and addictions. ENIGMA integrates images from multiple modalities, genomes, connectomes and biomarkers on an unimaginable scale, with new computations to integrate, cluster, and learn from complex biodata types.

Program Officer: Vinay Pai
Science Officers: Patrick Bellgowan, Yantian Zhang, Harold Gordon

THE NATIONAL CENTER FOR MOBILITY DATA INTEGRATION TO INSIGHT (THE MOBILIZE CENTER)

http://mobilize.stanford.edu/

Stanford University
PIs: Scott L. Delp
Grant Number: 1U54EB020405-01

The Mobilize Center is poised to provide access to mobility data for over ten million people. The center will develop and disseminate a range of novel data science tools, including modeling and analysis methods to predict and improve the outcomes of surgeries in children with cerebral palsy and gait pathology; to identify new approaches to optimize mobility in individuals with osteoarthritis, running injuries, and other movement impairments; and to discover methods that motivate overweight and obese individuals to exercise more and in ways that promote joint health.

Mobility is essential for human health. Regular physical activity helps prevent heart disease and stroke, relieves symptoms of depression, and promotes weight loss. Unfortunately, many conditions, such as cerebral palsy, osteoarthritis, and obesity, limit mobility at an enormous personal and societal cost. While vast amounts of data are available from hundreds of research labs and millions of smartphones, there is a dearth of methods for analyzing this massive, heterogeneous dataset. We propose to establish the National Center for Mobility Data Integration to Insight (the Mobilize Center) to overcome the data science challenges facing mobility big data and biomedical big data in general. Our preliminary work identified four bottlenecks in data science, which drive four Data Science Research Cores. The Cores include Biomechanical Modeling, Statistical Learning, Behavioral and Social Modeling, and Integrative Modeling and Prediction. Our Cores will produce novel methods to integrate diverse modeling modalities and gain insight from noisy, sparse, heterogeneous, and time-varying big data. Our data-sharing consortia, with clinical, research, and industry partners, will provide mobility data for over ten million people. Three Driving Biomedical Problems will focus and validate our data science research. The Mobilize Center will disseminate our novel data science tools to thousands of researchers and create a sustainable data-sharing consortium. We will train tens of thousands of scientists to use data science methods in biomedicine through our in-person and online educational programs. We will establish a cohesive, vibrant, and sustainable National Center through the leadership of an experienced executive team and will help unify the BD2K consortia through our Biomedical Computation Review publication and the Simtk.org resource portal. The Mobilize Center will lay the groundwork for the next generation of data science systems and revolutionize diagnosis and treatment for millions of people affected by limited mobility. PUBLIC HEALTH RELEVANCE: Regular physical activity is essential for human health, yet a broad range of conditions impair mobility. This project will transform human movement research by developing tools for data analysis and creating software that will advance research to prevent, diagnose, and reduce impairments that limit human movement.

Program Officer: Grace Peng
Science Officers: Theresa Cruz, Daofen Chen

Award: Biological and HealthCare Data Discovery and Indexing Ecosystem (bioCADDIE)

bioCADDIE seeks to develop a prototype DDI that will enable finding, accessing and citing biomedical big data. bioCADDIE has a Community Engagement mandate that seeks to work with the broader biomedical community to better identify data, and other digital objects, so that they may find shared data in ways that allow for extracting maximal knowledge.

Biomedical and healthcare data sharing efforts are currently impaired by lack of (1) proper incentives and sharing tools for data producers, (2) practical frameworks for data standardization and indexing of data, and (3) effective data discovery mechanisms. BioCADDIE is a consortium of data producers, curators, publishers, and consumers who will work together to develop practical, sustainable solutions to the problem of biomedical and healthcare data discovery. Through task forces and corresponding pilot projects addressing the barriers enumerated above, we will promote open discussion of why millions of dollars are currently spent in the generation of data that remain captive at their origin or are shared in a sub-optimal way just to comply with mandates from funding agencies and scientific journals. We will promote the development of incentives, policies, and tools for data sharing and data discovery. We will engage researchers, clinicians, patients, and the community in general in an open dialogue focused on pros and cons of biomedical and clinical data sharing. BioCADDIE's specific aims are to: (1) Organize task forces with representatives from communities who have interest in data production, dissemination, and utilization. We will organize an annual symposium, workshops, and internet-based discussions among biomedical and clinical researchers, professional societies, journal publishers, funding agencies, clinicians, patients, and information scientists on best, sustainable practices for making data easily discoverable by different types of users. (2) Promote the development of realistic, minimal, friendly meta-data specifications and annotations for biomedical and healthcare data collections, and corresponding tools for automated indexing so that users will be able to locate data that are relevant to their specific free text searches. (3) Incubate new technologies by funding highly innovative, high-risk pilot research projects that enable the development of novel data discovery and indexing engines and have them tested by our diverse community of stakeholders. We only describe a small number of seed pilot projects in this proposal because BioCADDIE will solicit proposals for new pilot projects every year and select them through a review process involving the various stakeholder communities. PUBLIC HEALTH RELEVANCE: Biomedical research and healthcare data are not fully utilized in part due to lack of incentives and tools to share these data in a way that makes it possible to reproduce results and make new discoveries. We will develop a consortium involving data producers, data disseminators, and data consumers (including patients) to develop tools and processes for easy discovery and access to data.

THE BD2K CONCEPT NETWORK: OPEN SHARING OF ACTIVE-LEARNING AND TOOLS ONLINE

http://ceils.ucla.edu/index.php/education-projects

The University of California Los Angeles
PIs: Christopher Lee
Grant Number: 1R25GM114822-01

This project will provide instructors with a platform for collaborative, peer- reviewed sharing and remixing of active-learning materials, which are known to greatly increase conceptual understanding and real-world problem-solving ability. In addition, the project will create freely available, online courselets; each will be centered on one key Big Data concept. Courselets will consist of brief videos tightly integrated with concept tests and active-learning exercises.

The proposed project will create online services, teaching materials sharing, and training for instructors and students to 1) expand and tailor Big Data To Knowledge (BD2K) learning for new audiences in bioinformatics, medical informatics and biomedical applications; 2) use active- learning to greatly increase conceptual understanding and real-world problem-solving ability; 3) directly measure learning effectiveness; and 4) boost the number of students that successfully complete BD2K courses. Tailoring the core concepts for BD2K success to teach diverse biomedical audiences is crucial both because these interdisciplinary concepts are a key barrier to entry, and because they are vital for real-world BD2K problem-solving ability. The UCLA/UCSD project team will: 1) provide an open, online repository where BD2K instructors worldwide can find, author, and share peer-reviewed active-learning exercises such as concept tests (already over 600), and immediately use them in class (with students answering with their smartphones or laptops); 2) catalyze the development, usage and validation of candidate BD2K concept inventories for rigorously measuring learning gains, via an accelerated approach of open- response concept testing and online data collection; 3) provide BD2K instructors a collaborative, peer-reviewed sharing and remixing platform for active-learning materials such as algorithm projects, hands-on data mining projects (via convenient “cloud projects”), exercises and problems, as well as “courselet” recording tools that automatically record video and audio on the instructor’s laptop while they teach; 4) provide students anywhere free online courselets each about one key BD2K concept, consisting of brief videos tightly integrated with concept tests and all the active- learning exercises described above, and designed as an online persistent-learning community unified by concepts, in which students learn from the community’s consolidated error models (common errors for a specific BD2K concept), effective remediations and counter-examples for each error model. Testing of this instructional approach for 3 years has doubled successful student completions of a BD2K methods course at UCLA, by reducing attrition, while simultaneously increasing conceptual understanding (mean exam scores). This approach will also be disseminated by: 1) pilot projects with BD2K instructors at UCLA and partner institutions, with detailed evaluation studies to identify critical success factors; 2) workshops (both online and onsite) for training instructors how to teach effectively with these tools in their BD2K courses; 3 online services and courselets. PUBLIC HEALTH RELEVANCE: Big Data to Knowledge (BD2K) education means bringing sophisticated data mining skills and thinking to researchers and clinicians throughout the biomedical enterprise, a most challenging interdisciplinary learning curve. This cannot succeed without the kinds of hands-on learning exercises that are hard to find in BD2K textbooks, but that students need, such as data-mining projects with real datasets and real computational powertools, concept tests and concept inventories that rigorously teach and measure conceptual understanding, and algorithm projects where students prove their understanding of a challenge problem, by writing code that can correctly solve any test case thrown at it. We will provide BD2K instructors a collaborative, peer-reviewed sharing platform for immediately using all of these kinds of active-learning materials in class (currently containing over 2000 BD2K exercises and related materials), and BD2K students free online courselets each about one key BD2K concept, consisting of brief videos tightly integrated with concept tests and all the active-learning exercises described above.

INTEGRATED ACTIVE LEARNING FRAMEWORK FOR BIOMEDICAL BD2K

The University of California San Diego
PIs: Pavel A. Pevzner
Grant Number: 1R25GM114819-01

This project will create active and adaptive open online resources for students and educators by developing two massive open online courses (MOOCs) for biomedical Big Data aimed at bioinformaticians and at biologists. It will also develop two new problem tracks on the online Rosalind platform, which facilitates independent learning of bioinformatics through automatically tested challenges. Finally, it will form an open community of educators with the goal of developing a vast set of open learning modules for biomedical Big Data.

The proposed project will create active and adaptive open online resources for students and educators. We propose the development of two massive open online courses (MOOCs) for Biomedical Big Data (BBD). BBD for Bioinformaticians will be aimed at bioinformatics students who know some introductory programming and need specialized tutorials focusing on BBD analysis. BBD for Biologists will provide biologists having no previous exposure to programming with the skills required to effectively apply existing software tools in BBD. These MOOCs will have three different adaptive learning tracks that will help guide readers through the courses based on their computational experience. Creating such an adaptive environment would not be possible without our substantial experience in offering the first bioinformatics MOOC, Bioinformatics Algorithms, in fall 2013 on Coursera. By making our learning materials open for use by individual learners and professors, we hope to bring down resource barriers that have prevented BBD courses from growing at individual universities. We will also develop two new problem tracks on our online Rosalind platform that facilitates independent learning of bioinformatics through automatically tested challenges. One of these problem sets will focus on implementing the algorithms required for BBD analysis; the second problem set will focus on applying existing online tools to analyze BBD. By creating a comprehensive set of assessments, we will eliminate the need for BBD professors to ever again think about automating their own homework assignments. Combined with the efforts of our open, adaptive learning environment, these problem sets will help reduce the barriers to creation of new BBD courses at universities. We will foster an open community of BBD educators by forming the BBD Education Alliance. This network will be founded at the RECOMB Conference on Bioinformatics Education at UCSD in 2015, which will focus on BBD education. Members in the alliance will create open learning modules to supplement our content as well as provide feedback to other members of the alliance on their modules. These educators will also work to design BBD courses at their own universities. Finally, PUBLIC HEALTH RELEVANCE: Big Data to Knowledge initiatives at NIH cannot succeed without educating students, researchers, and clinicians in how to analyze Biomedical Big Data (BBD). This project will create active and adaptive open online resources for students and educators by developing two massive open online courses (MOOCs) for BBD. BBD for Bioinformaticians will be aimed at bioinformatics students who know some introductory programming and need specialized tutorials focusing on BBD analysis. BBD for Biologists will provide biologists having no previous exposure to programming with the skills required to effectively apply existing software tools in BBD. We will also develop two new problem tracks on our online Rosalind platform that facilitates independent learning of bioinformatics through automatically tested challenges. Finally, we will foster an open community of BBD educators by forming the BBD Education Alliance, with the goal of developing a vast set of open learning modules for BBD.

AN OPEN RESOURCE FOR COLLABORATIVE BIOMEDICAL BIG DATA TRAINING

The University of California San Diego
PIs: Rommie E. Amaro and Ilkay Altintas de Callafon
Grant Number: 1R25GM114821-01

This award will be used to cultivate a community effort around training and education in biomedical big data research, through the Biomedical Big Data Training Collaborative (BBDTC). The BBDTC will develop a module-based curriculum, MOOCs, and virtual machine environments for hands-on exploration. This and other content will be hosted in an open repository, creating a vibrant, community-based approach to generating high-quality biomedical big data content or training.

Of all the resources required to make gaining insight from big data a success, perhaps the most important is the human one. A major challenge to the big data community generally and especially, the biomedical big data community is training and education of the current and next generation of biomedical scientists. We must work collectively to address this critical challenge. What we seek to do through this proposed project is maximize the impact of biomedical big data training through a large-scale collaborative approach, and to create a training and education framework for other educators (a.k.a., teachers, instructors) and/or learners (a.k.a., students, trainees, researchers) that enables them to construct and deliver customized modules or courses that deliver the highest value to their particular application. Our vision is to cultivate a high-quality, well-informed, freely accessible knowledge and data community effort around training and education in biomedical big data research, through the Biomedical Big Data Training Collaborative (BBDTC). The end-to-end BBDTC open online training framework is a repository allowing faculty, researchers and students access to a state-of-the-art training model over the years to come. To develop such an environment, we employ current best practices and build upon our existing efforts. Initially we focus on building the BBDTC along with example courses, lecture content and hands-on application use cases for biomedical big data training. We will also communicate best practices for developing course content and delivering it to a wide-range of trainees along with associated adaptive learning approaches and assessments. In addition, we will deliver customizable virtual machines (VMs) including the course materials, hands-on tools and example data and additional assessment and make sure that these VMs are portable to a variety of environments. Specific aims in the project include development of: (1) Biomedical Big Data Curriculum; (2) Biomedical Big Data MOOC Framework; (3) Biomedical Big Data Tool Box; and (4) Repository Interfaces to Engage Community Stakeholders. The significance of our approach is that the BBDTC will enable the development of many more courses and training modules (whether they are full-scale MOOCs or much smaller, more targeted units). Although we focus on the present “mission critical” challenges defined by the NIH and biomedical community, we build the BBDTC framework in a way that will allow it to evolve over the years, not just by one person but by a community of biomedical big data researchers as a collective force to handle training challenges of the future. PUBLIC HEALTH RELEVANCE: Our vision is to cultivate a high-quality, well-informed, freely accessible knowledge and data community effort around training and education in biomedical big data research, through the Biomedical Big Data Training Collaborative (BBDTC). The BBDTC open online training framework will enable faculty, researchers and students access to a state-of-the-art biomedical big data training model over the years to come.

BIOMEDICAL BIG DATA TRAINING GRANT

The University of California Los Angeles
PIs: Matteo Pelligreni
Grant Number: 1T32 CA201160-01

Matteo Pellegrini, Ph.D., and co-investigators, Alex Bui, Ph.D. and Alexander Hoffmann, Ph.D., bring together 29 faculty mentors from bioinformatics, clinical/imaging informatics, experimental biology, math, statistics, biostatistics, and computer science to create a program focused on training students in the general problem of analyzing and relating large biomedical data sets. Trainees will complete a tailored set of electives in big data analysis that includes areas of data management and computational algorithms, biostatistics and biomathematical analyses, and informatics. In addition, participation in team-based big data challenges, high performance computing and big data workshops, and an extramural summer internship in NIH BD2K Centers or in the bioinformatics/biotechnology industry are required. Students will have access to several big data resources unique to UCLA that includes data sets in bipolar disorder, depression, autism, lung cancer, and breast cancer. The training program will foster the quantitative thinking of students that are expected to aid the discovery process in biology.

This broad-based T32 proposal is focused on supporting graduate students pursuing Biomedical Big Data Analysis research at UCLA. Over the past few years there has been increasing recognition that the biomedical sciences are undergoing a transformation, led by the development of new technologies that have enormously increased the capacity to generate data. These include technologies to sequence DNA and RNA; measure protein and metabolite abundances using mass spectrometry; as well as multiple other high throughput platforms for screening and phenotyping. Coupled with the advances in medical imaging and EHRs, the amount of data is growing faster than ever before. While several other training grants exist at UCLA, none are specifically focused on the general problem of analyzing and relating large biomedical data sets. Thus this new training grant we propose here fills a critical niche that will allow us to support graduate students in fundamental aspects of biomedical “big data” analysis. This effort realizes the novel development of a tailored set of courses in big data analysis, along with specialized team building activities in bi data challenges and extramural internships in big data centers. This training program will position our students for the future growth in big data science, fostering the growth of this critical area at UCLA. We posit that this program will be critical for the growth of our bioinformatics program and big data biomedical science at UCLA. Faculty supporting this training program represent an interdisciplinary group of researchers from across the UCLA campus, including the Schools of Engineering & Applied Sciences; College of Life Sciences and the Medical School. Many of the faculty are nationally recognized leaders in their disciplines; collectively, these individuals provide a comprehensive and complementary set of (funded) research areas and skills that enrich the training experience. PUBLIC HEALTH RELEVANCE: Since 2008, the UCLA Bioinformatics Program has recruited a growing number of outstanding students trained at the boundary of biology and computational sciences. The proposed Biomedical Big data Training program will support a subset of these students with an interest in developing and applying skills to analyze massive scale biomedical data, such as sequence, proteomics, and medical records. The students will take specialized big data courses and work with pairs of mentors with expertise in big data and either experimental of computational biology.

PARTICIPANTS

Joe Ames	joe.ames@ini.usc.edu	BDDS	USC
Naveen Ashish	naveen.ashish@loni.usc.edu	BDDS	USC
Vivien Bonazzi	bonazziv@mail.nih.gov	NHGRI	NIH
Ian Bowman	Ian.Bowman@loni.usc.edu	MCP	USC
Magali Champion	mchampio@stanford.edu	CEDAR	Stanford
Kyle Chard	chard@uchicago.edu	BDDS	U of Chicago
Howard Choi	Cjh9595@gmail.com	Heart BD2K	UCLA
Kristi Clark	kclark@ini.usc.edu	BDDS	USC
Jennifer Couch	Couchj@ctep.nci.nih.gov	NCI	NIH
Eric Deutsch	edeutsch@systemsbiology.org	BDDS	ISB
Tevfik Umut Dincer	Dincer@ucla.edu	Heart BD2K	UCLA
Ivo Dinov	dinov@umich.edu	BDDS	U of Michigan
Hongwei Dong	hongwei.dong@loni.usc.edu	MCP	USC
Michel Dumontier	michel.dumontier@stanford.edu	CEDAR	Stanford
Michelle Dunn	dunnm3@od.nih.gov	OD	NIH
Ian Foster	foster@anl.gov	BDDS	U of Chicago
Stacia Friedman-Hill	friedmans@mail.nih.gov	NIMH	NIH
Olivier Gevaert	ogevaert@stanford.edu	CEDAR	Stanford
Clio Gonzalez-Zacarias	clio.gonzalez-zacarias@loni.usc.edu	BDDS	USC
Jeff Grethe	jgrethe@ncmir.ucsd.edu	BioCADDIE	USCD
Ben Heavner	Bheavner@systemsbiology.org	BDDS	ISB
Jen Hicks	jenhicks@stanford.edu	Mobilize Center	Stanford
Houri Hintiryan	Houri.Hintiryan@loni.usc.edu	MCP	USC
Sam Hobel	samuel.hobel@loni.usc.edu	BDDS	USC
William Hsu	WHsu@mednet.ucla.edu	Heart BD2K	UCLA
Neda Jahanshad	neda.jahanshad@ini.usc.edu	ENIGMA	USC
Carl Kesselman	carl@isi.edu	BDDS	USC
Purvesh Khatri	pkhatri@stanford.edu	CEDAR	Stanford
Maggie Pui Yu Lam	Magelpy@ucla.edu	Heart BD2K	UCLA
Jennie Larkin	larkinj2@od.nih.gov	OD	NIH
Edward Lau	Edward.lau@me.com	Heart BD2K	UCLA
Christina Liu	Liuch2@mail.nih.gov	NIMH	NIH
Ravi Madduri	madduri@mcs.anl.gov	BDDS	U of Chicago
Henrietta Movsessian	Henrietta.Movsessian@loni.usc.edu	BDDS	USC
Dan Moyer	Daniel.Moyer@loni.usc.edu	ENIGMA	USC
Sonynka Ngosso	sonynka.ngosso@nih.gov	OD	NIH
Judy Pa	jpa@ini.usc.edu	BDDS	USC
Vinay Pai	Paiv@mail.nih.gov	NIBIB	NIH
Petros Petrosyan	petros.petrosyan@loni.usc.edu	BDDS	USC
Peipei Ping	pping@mednet.ucla.edu	Heart BD2K	UCLA
Gautam Prasad	gautam.prasad@loni.usc.edu	ENIGMA	USC
Nathan Price	nprice@systemsbiology.org	BDDS	ISB
Peter Rose	Peter.Rose@rcsb.org	PDB	UCSD
Erica Rosemond	rosemonde@mail.nih.gov	NIMH	NIH
Nova Smedley	novasmedley@ucla.edu	Heart BD2K	UCLA
Andrew Su	asu@scripps.edu	Heart BD2K	TSRI
Paul Thompson	thompson@loni.usc.edu	ENIGMA	USC
Arthur Toga	toga@loni.usc.edu	BDDS	USC
John Van Horn	jack.vanhorn@loni.usc.edu	BDDS	USC
Chunlei Wu	cwu@scripps.edu	MyGene.info	Scripps
Darvin Yi	darvinyi@stanford.edu	CEDAR	Stanford
Mu Zhou	muzhou@stanford.edu	CEDAR	Stanford
Daijiang Zhu	Dajiang.Zhu@loni.usc.edu	ENIGMA	USC