Meet the Coordination Centers

BD2K All-hands Grantee Meeting

Friday, November 13

9:00-10:00 am

Meet the BD2K Coordination Centers

Data Discovery Index Coordination Consortium

(DDICC)

Lucila Ohno-Machado (UCSD)

Centers Coordination Center

(CCC)

Peipei Ping (UCLA)

Training Coordination Center

(TCC)

Jack Van Horn (USC)

Standards Coordination Center

(SCC)

Anand Basu (ESA Inc)

coordination activities

a data discovery index coordinating consortium:
interconnecting indexing efforts
for the NIH Commons

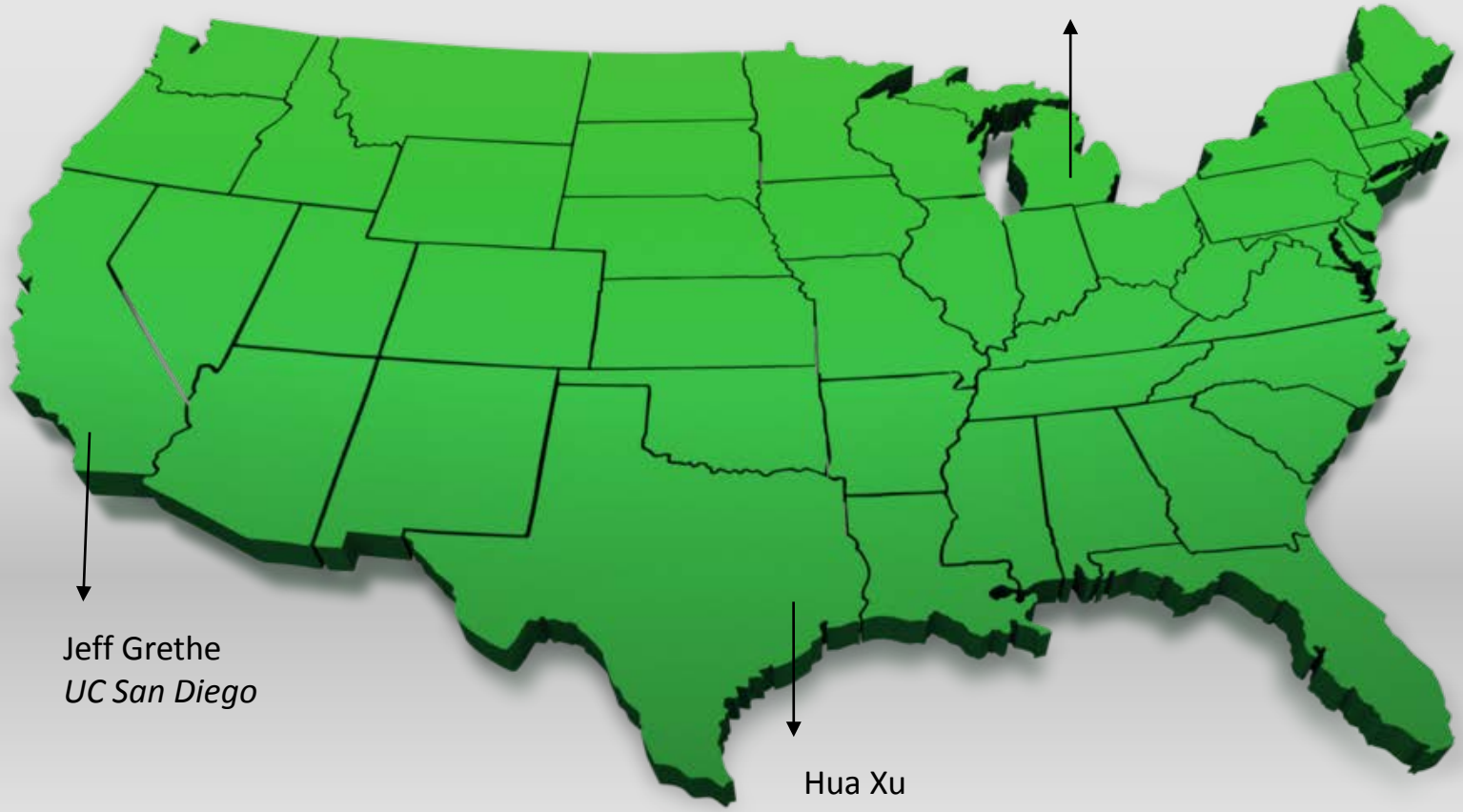
2nd BD2K all-hands meeting

Bethesda, 11/13/15

Consortium Hubs

Susanna Sansone
Oxford University

George Alter
U Michigan



Jeff Grethe
UC San Diego

Hua Xu
UT Houston

Goals

Engage the community to

- Define minimal set of metadata and identifiers
- Develop search engine (functionality, testing)
- Engage data producers & aggregators in linking their assets

through

- Working groups
- Supplements to bioCADDIE or Centers of Excellence
- Small pilot projects funded by bioCADDIE
- Webinars

bio CADDIE

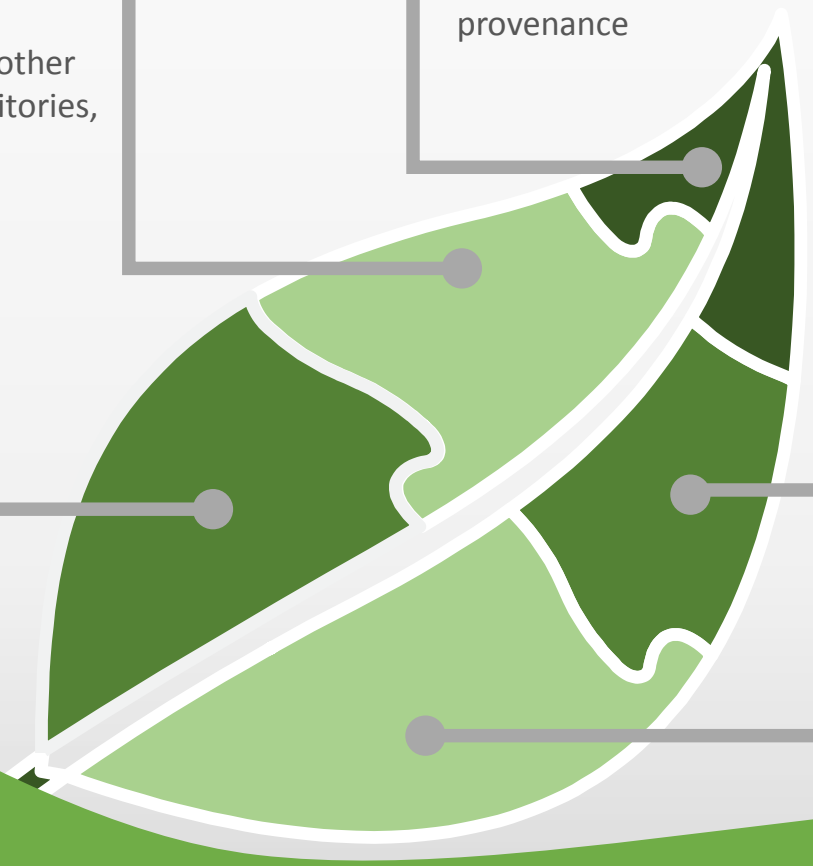
Metadata
minimal set
guidelines for mapping,
accessibility information,
provenance

Search engine
connection to other
engines, repositories,
data sets

Identifiers
existing or new ID
issuing services

Standards
metadata
data

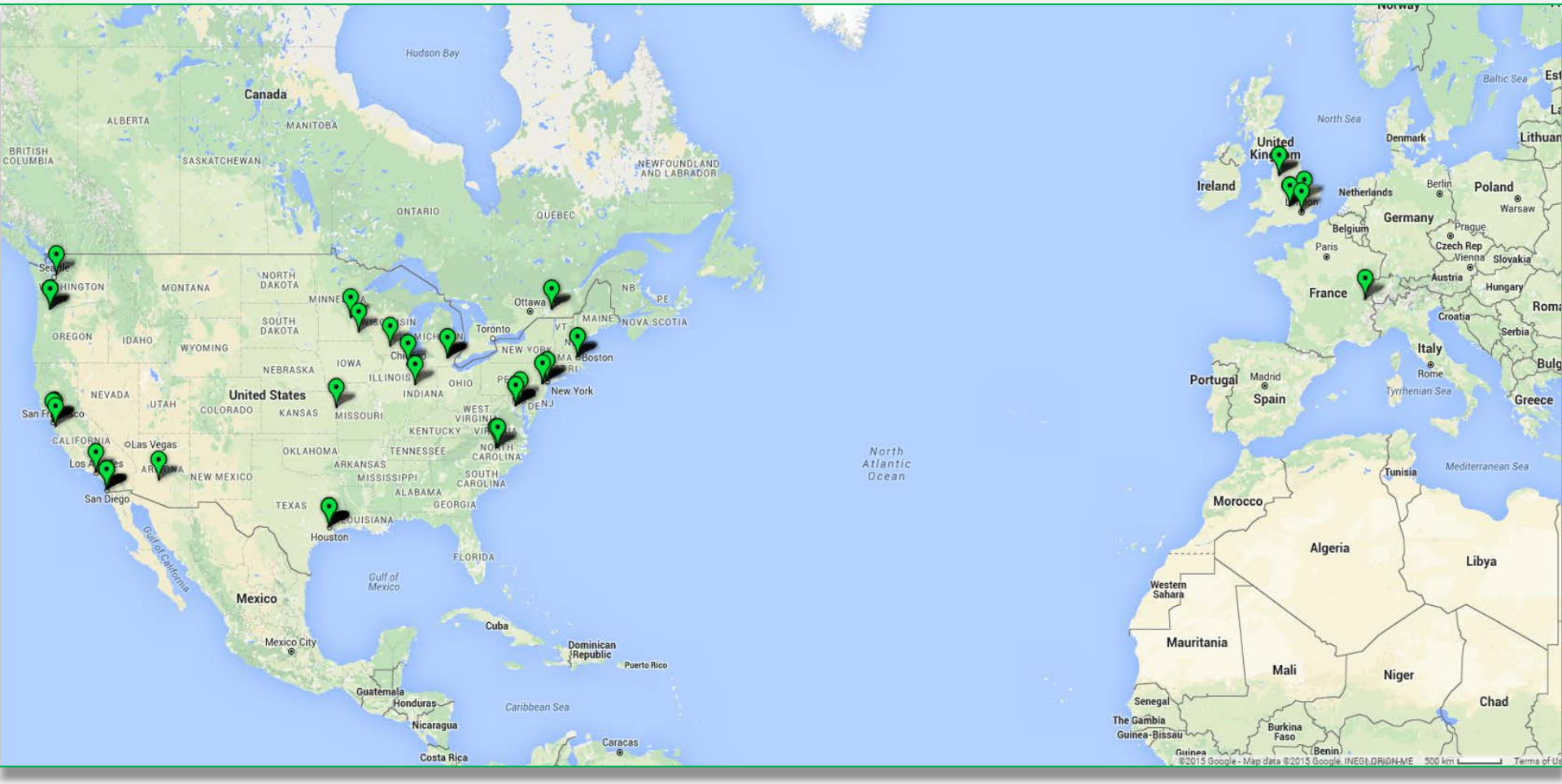
Policies
criteria for prioritization
sustainability

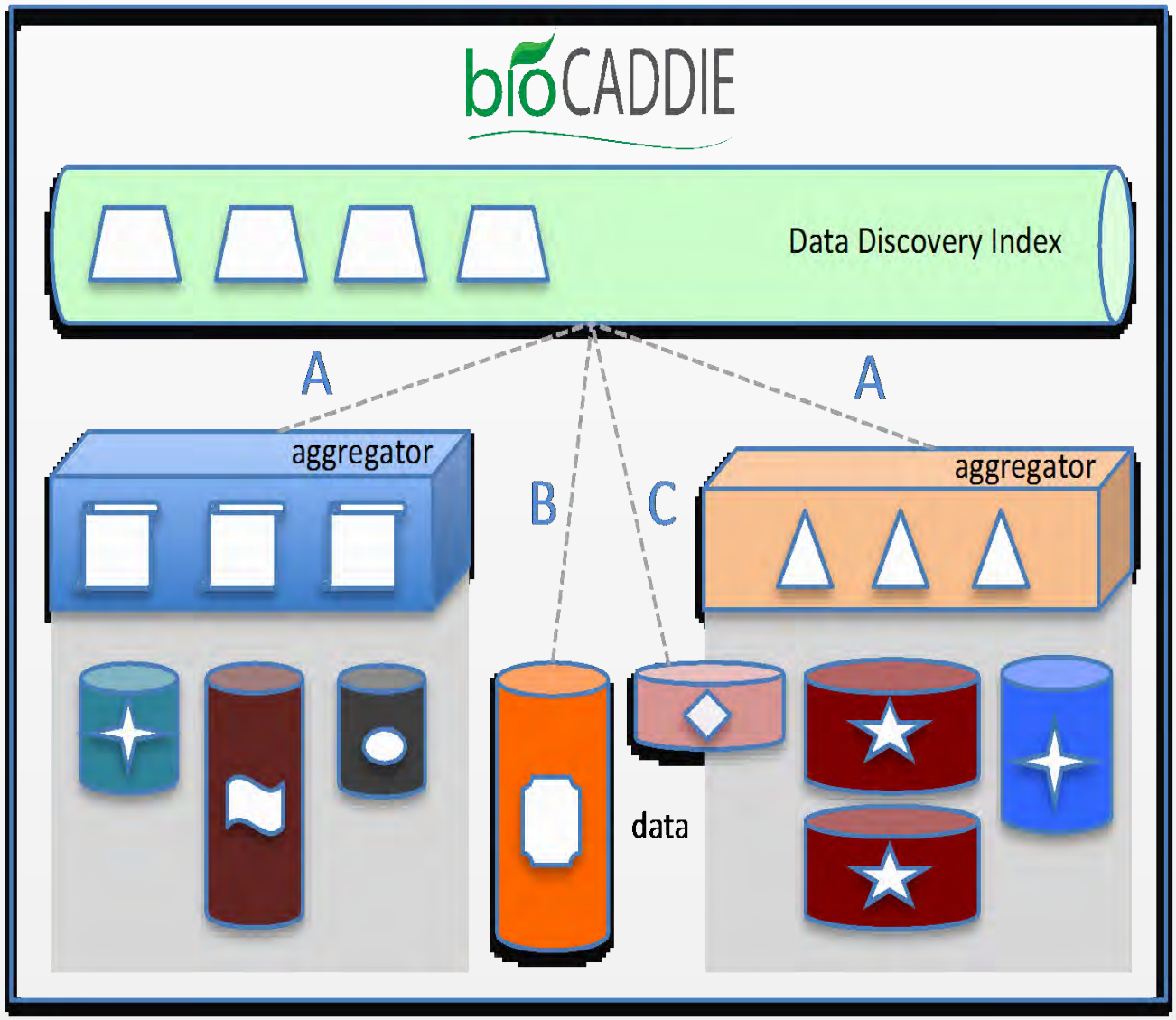


ecosystem components for finding, accessing, interoperating with repositories/indices, and (re)using data

bio CADDIE

Working Group Engagement





collaborations

data citation (FORCE11)

mapping of metadata, links to aggregators, APIs (Count Everything)

aggregators
various indices whose metadata are or can be mapped into Commons metadata (eg, Omics DI, EKG DI, Aztec, iDASH)

data (Athena)

working groups

pilots & supplements

bioCADDIE

Working Group: Data Identifiers

Check document at biocaddie.org

Tim Clark, Harvard Medical School
Merce Crosas, Harvard University
Michel Dumontier, Stanford University
Claudiu Farcas, UCSD
Ian Fore, NIH
Carole Goble, University of Manchester
Alejandra Gonzalez-Beltran, University of Oxford
Jeffrey Grethe, UCSD
John Kunze, California Digital Library
Ron Margolis, NIH
Jo McEntyre, EMBL-EBI
Julie McMurry, Oregon Health and Science University
Philippe Rocca-Serra, University of Oxford
Susanna-Assunta Sansone, University of Oxford
Heidi Sofia, NIH
Stian Soiland-Reyes, University of Manchester
Joan Starr, California Digital Library
Hua Xu, University of Texas Houston

Attend breakout session on Identifiers

Omics Discovery Index

Eric Deutsch
Institute for Systems Biology
HeartBD2K
BDDS

dataset discovery across a heterogeneous, distributed group of

- genomics,
- proteomics and
- metabolomics data resources

spanning

- 8 repositories in
- 2 continents and
- 6 organizations,

including both

- open and
- controlled access

data resources

Peipei Ping
UCLA
Heart BD2K

Henning Hermjakob
EBI
Heart BD2K

Susanna Sansone
Oxford University
bioCADDIE
CEDAR

Metadata

Check documents at biocaddie.org

George Alter, ICPSR, University of Michigan
Muhammad Amith, University of Texas Houston
Tanya Barrett, NIH
Helen Berman, Protein Data Bank
Michael Braxenthaler, Roche
Matthew Brush, Oregon Health and Science University
Allen Dearry, NIH
Michel Dumontier, Stanford University
Ian Fore, NIH
Carole Goble, University of Manchester
Alejandra Gonzalez-Beltran, University of Oxford
Jeffrey Grethe, UCSD
Melissa Haendel, Oregon Health and Science University
Marcelline Harris, University of Michigan
Michael Huerta, NLM
Hyeon-Eui Kim, UCSD
Steven Kleinstein, Yale School of Medicine
Jared Lyle, ICPSR, University of Michigan
Ron Margolis, NIH
Matthew McAuliffe, NIH
Kevin Read, NYU
Philippe Rocca-Serra,
University of Oxford
Susanna-Assunta Sansone, University of Oxford
Heidi Sofia, NIH
Joan Starr, California Digital Library
Cui Tao, University of Texas Houston
Shraddha Thakkar, NCTR FDA
Weida Tong, NCTR FDA
Mary Vardigan, ICPSR, University of Michigan
John Westbrook, Protein Data Bank
Rai Winslow, John Hopkins University
Hua Xu, University of Texas Houston

Attend breakout session on metadata

Count Everything

David Haussler
UC Santa Cruz
Translational Genomics

How many individuals in your database have property X

X is a property that can be expressed using a formal data schema from a broadly adopted API

How many individuals

- are > 65yo?
- have congestive heart disease?
- have the allele 'T' at position 3,283,491 of chromosome 3?
- are taking warfarin?
- are taking < 1000 steps a day?

Answer these questions while preserving patient privacy

Isaac Kohane
Harvard University
PIC-SURE

Santosh Kumar
mD2K

Ida Sim
UCSF
mD2K

Claudiu Farcas
UC San Diego
bioCADDIE

Dataset Citation Metrics

Check documents at biocaddie.org

Trisha Cruse, California Digital Library

Sunje Dallmeier-Tiessen, CERN

Martin Fenner, DataCite

Alejandra Gonzalez-Beltran, University of Oxford

Jeffrey Grethe, UCSD

Kristi Holmes, Northwestern University

Rebecca Lawrence, F1000

Jennifer Lin, PLOS

Jared Lyle, ICPSR, University of Michigan

Ron Margolis, NI

Jo McEntyre, EMBL-EBI

Daniel Mietchen, NIH

Philippe Rocca-Serra, University of Oxford

Susanna- Assunta Sansone, University of Oxford

Michael Taylor, Elsevier

Future of Research Communications and e-Scholarship - FORCE11

test out the joint FORCE11-RDA data citation implementation recommendations using biomedical data resources like the PDB

- stakeholders are brought together, including the data repositories and publishers
- authors of submitted manuscripts are asked to adopt the new form of data citation that is machine-actionable and uniform across publishers
- Launch FORCE 11 Software Citation Group

Tim Clark
Harvard University

Maryan Martonne
UC San Diego
FORCE11

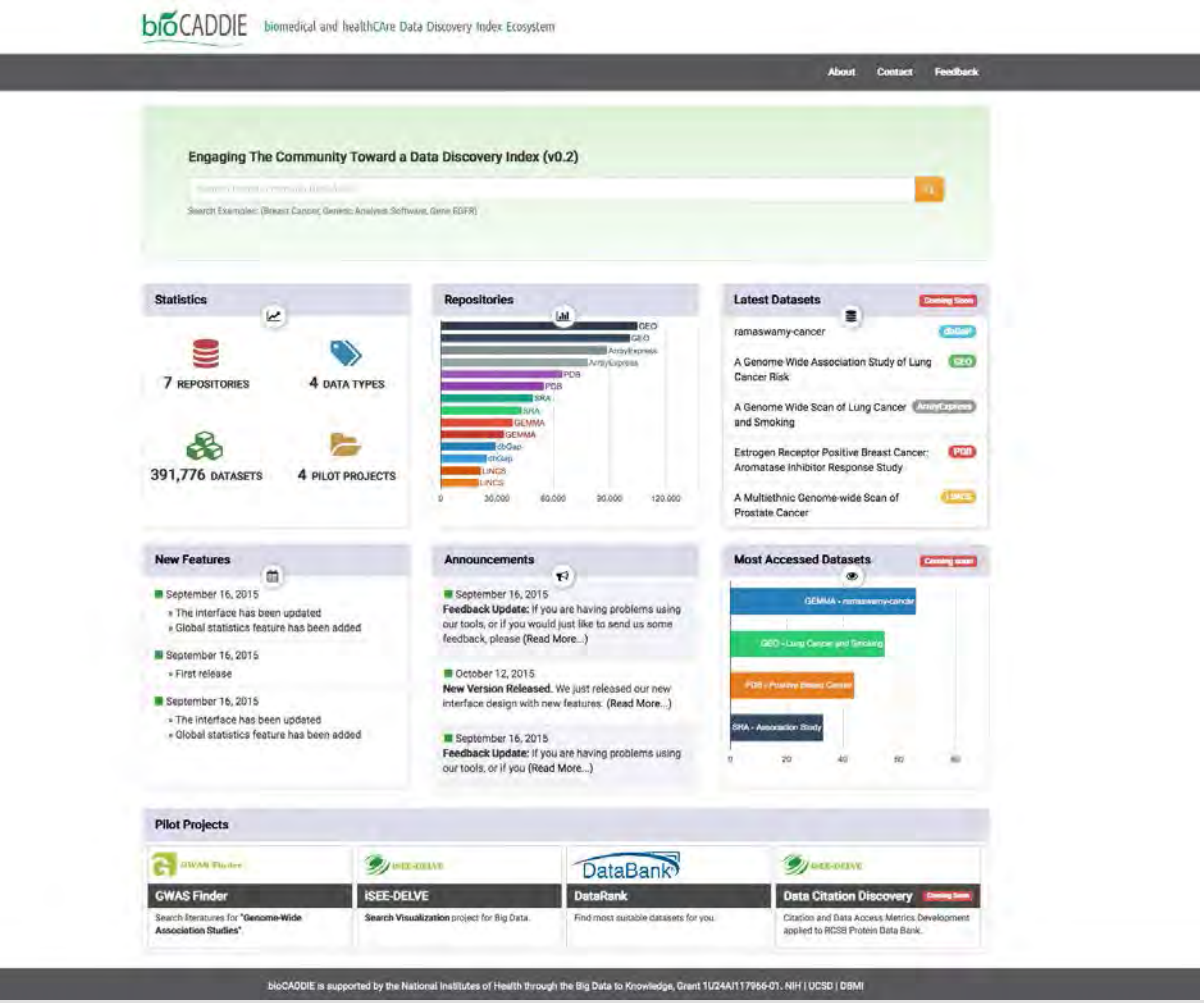
Use Cases and Testing Benchmarks

Check documents at biocaddie.org

George Alter, ICPSR, University of Michigan
Trevor Cohen, University of Texas Houston
Dina Demner-Fushman, NIH
Todd Johnson, University of Texas Houston
Dave Kaufman, Arizona State University
Hyeon-Eui Kim, UCSD
Steven Kleinstein, Yale School of Medicine
Zhiyong Lu, NIH
Thomas Radman, NIH
Susanna-Assunta Sansone, University of Oxford
Cui Tao, University of Texas Houston
Hua Xu, University of Texas Houston

Test a pre-release version of the prototype, give feedback

Prototype



- User needs analysis
- User feedback during development
  - Issue reporting at GitHub
  - Brief survey at the BD2K meeting
- User-centric evaluation, system-centric evaluation
- Formal usability study
  - User analysis
  - Function analysis
  - Representation analysis
  - Task analysis

bio CADDIE

New Working Groups

WG6: Criteria for Repository Inclusion (Standards, Interoperability, Sustainability)

George Alter, ICPSR, University of Michigan
Tim Clark, Harvard Medical School
Ian Fore, NIH
(Still recruiting)
Jeffrey Grethe, UCSD
Lucila Ohno-Machado, UCSD
Susanna- Assunta Sansone, University of Oxford

WG7: Accessibility Metadata for Datasets

George Alter, ICPSR, University of Michigan
Damon Davis, HealthData.gov
Alex Kanous, University of Michigan
Hyeon-Eui Kim, UCSD
Jared Lyle, ICPSR, University of Michigan
Frank Manion, University of Michigan
Reagan Moore, University of North Carolina
Mark Phillips, McGill University
Kendall Roark, Purdue University
Jessica Scott, GlaxoSmithKline
Anne-Marie Tassé, McGill University

WG8: Ranking Search Results

Elmer Bernstam, University of Texas Houston
Trevor Cohen, University of Texas Houston
Anupama Gururaj, University of Texas Houston
Christina Kendziorski, University of Wisconsin-Madison
Hyeon-Eui Kim, UCSD
Hongfang Liu, Mayo Clinic
Hua Xu, University of Texas Houston
Jim Zheng, University of Texas Houston

Pilot$

Newly awarded – Metadata Discover

- Distributed data discovery using gym: github, yaml and markdown

  Chris Mungall, Lawrence Berkeley National Laboratory
- Feasibility study of indexing clinical research data using HL7 FHIR

  Guoqian Jiang, Mayo Clinic College of Medicine
- Metadata discovery and integration to support repurposing of heterogeneous data using the Openfurther platform

  Ram Gouripeddi and Julio Facelli, University of Utah

Results of Ranking (1: top proposal)									
Project	Reviewer 1	Reviewer 2	Reviewer 3	Reviewer 4	Reviewer 5	Reviewer 6	Reviewer 7	Reviewer 8	Sum
Project 1	2	NR	NR	1	NR	1	1	1	6
Project 2	3	2	NR	2	1	2	NR	NR	10
Project 3	4	1	NR	NR	2	NR	2	3	12
Project 4	1	4	1	NR	4	NR	4	NR	14
Project 5	NR	3	2	NR	3	3	3	NR	14
Project 6	NR	NR	4	3	5	4	NR	2	18
Project 7	5	NR	3	4	NR	5	NR	4	21

Planned Working Groups

- Innovation in metadata generation
- Interoperability with repositories
- Standards-driven curation

Pilots

- Experimental projects funded by bioCADDIE
- The goal is to integrate with the prototype
- Peer-review process with objective criteria
- RFA for *Machine Actionable Licenses*

Who do I talk to?

- I have digital objects that need indexing
- I built an index and want to disseminate it
- I want to contribute my domain knowledge to an index activity
- I want to contribute my data science knowledge to build this ecosystem

CCC

centers' objects

SCC

standards

biosharing

TCC

training materials

bioCADDIE

data

What do you want to index?

What is BD2K CCC About

Service & Support

Supporting the Daily Operations of BD2K Initiatives.

Facilitating the Dissemination of BD2K Accomplishments.

NIH BD2K CCC Focus Areas & Deliverables

NIH Office of the ADDS,
NIH BD2K Program Directors & Staff

- **12 BD2K Centers**
- **BD2K Working Groups** (e.g., metadata standards)
- **Scientific Communities**

BD2K Centers Coordination Center Leadership Committee

NIH Program Representatives;
Peipei Ping, Alex Bui, Wei Wang,
Henning Hermjakob, and Andrew Su.

- **bioCADDIE**
- **BD2K Training Coordination Center**
- **Others (TBA)**

Focus Area 1

Administration and Coordination Support

- ❑ Perform administrative duties following the instructions of the NIH Office of the ADDS and NIH BD2K Program Officers;
- ❑ Organize PI Meetings, Advisory Panel Meetings, Working Group Meetings, Teleconferences, and Consortium Activities (e.g., Hackathon);
- ❑ Create and maintain a web portal of the Coordination Center; update publications, press releases, meetings, and other relevant information, etc.

Focus Area 2

Tool and Knowledge Management

- ❑ Create a Directory to document, index, and register all tools, APIs, and digital resources from the 12 BD2K Centers;
- ❑ Collaborate with bioCADDIE and NIH BD2K metadata standard Working Group to adopt and comply with NIH metadata standards;
- ❑ Work with Training Coordination Center to index training materials.

BD2K CCC | NIH National Institutes of Health

BD2K CCC Service & Resources

- Introduction to BD2K Services & Support (Peipei)
- Introduction to BD2K Hackathon
  (led by Andrew Su).
- Introduction to Project AZTEC, a BD2K Resource Directory
  (led by Wei Wang; collaborations with BioCADDIE & KnowEng)
- Integration with NIH Commons Framework
  (led by Henning Hermjakob; collaborations with BDDS & Elixir).
- Introduction to BD2K Working Group (WG) coordination
  (led by Alex Bui).

BD2K CCC Service & Resources

Highlight the achievement and success of your center

- BD2KCCC Website, twitter, blog, etc.
- 5 min videos to visualize scientific content of your center.
- One page white paper to highlight your center.
- Journals, special issues on BD2K.
- Workshops to assist you synergizing with and to disseminate BD2K accomplishments in your primary scientific organization(s).

BD2K CCC Hackathon & Code-Fest Call for Applications

Goal:

Promote BD2K & data science synergies, both within the BD2K consortium and in the broader scientific community; facilitating education and training of data science workforce.

Mechanism:

Sponsor up to six Hackathons annually ($3,000 each).

Future Deadlines:

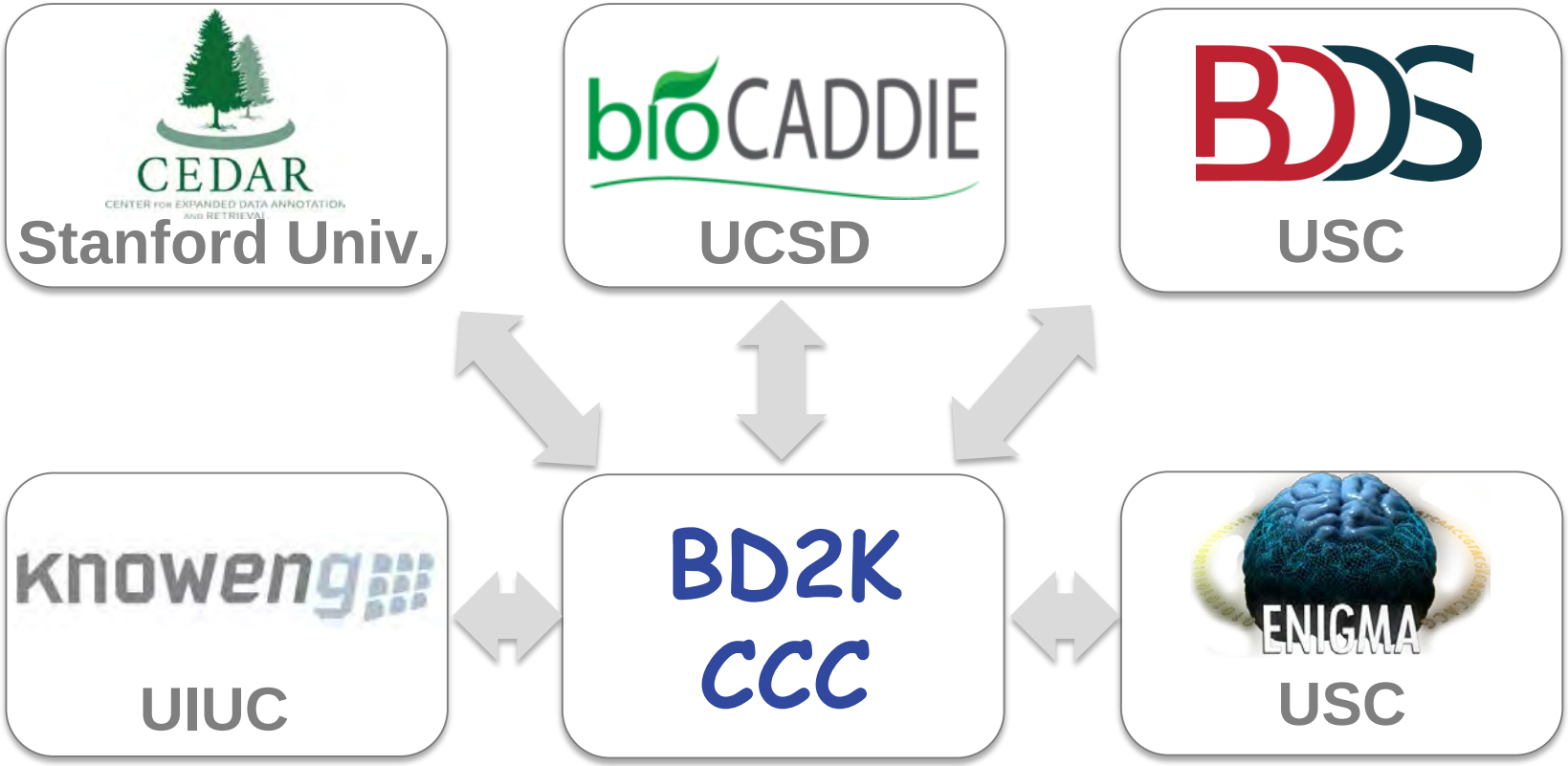
Feb 29th, 2016; Aug 31st, 2016; Feb 28th, 2017; Aug 31st, 2017.

Upcoming Events:

CEDAR COE Nov 2015	MyGene / MyVariant U01 Jan 2016	LINCS DDIC Feb 2016	Cancer Genomics Cloud, NCI Apr 2016	Heart BD2K COE Apr 2016

BD2K CCC | NIH National Institutes of Health

Leveraging Existing Resources and Collaborations with BD2K Centers



BD2K Metadata WG

- **Working group leaders:** Mark Musen (Stanford, CEDAR), George Alter (NIH, BioCADDIE)
- **Working group contact:** John Graybeal (Stanford, CEDAR)
- **Meetings:** The Metadata WG holds online meetings for members every 3rd Friday of the month at 11 AM PST.

https://workinggroups.bd2kccc.org/community/metadataworkinggroup

The Metadata WG focuses on topics related to the development of standards, annotation frameworks, and tools for dealing with meta-information about biomedical data, software, literature, and systems.

A large part of the NIH BD2K initiative focuses on developing frameworks that enable researchers to easily discover information - be it datasets, software, models, research reports, or other type of data. Key to this ability will be the creation of consistent, easily searchable metadata, which will enable sharing. The Metadata Working Group looks at issues related to representations and interface for metadata creation; and the use of ontologies, controlled vocabularies, and other knowledge resources for consistency in terminology. The Working Group also has interest in methods that help to automate metadata creation and validation; as well as techniques for allowing associated metadata to change over time (and for such provenance to be tracked).

BD2K API WG

- **Working group leaders:** Wei Wang (UCLA, HeartBD2K), Omar Sobh (U Illinois, KnowEng)
- **Working group contact:** TBD
- **Meetings:** Regular time TBD; initial WG formation kicked off on Nov 3rd, 2015.

The API WG broadly focuses on topics surrounding the definition, standards, and interoperability of a software ecosystem driven by the products of the BD2K Centers of Excellence and Targeted Software Development projects.

It is anticipated that the BD2K community will develop a wide array of software, packages in a range of ways (web services, Docker modules, etc.). Efforts to make the developments of each site a part of a larger, integrated whole requires coordination of standards to ensure interoperability at multiple levels (representational, contextual, semantic, etc.).

The API Working Group looks at issues related to how BD2K developments can be linked together at a programmatic level, and ways to assist in the discovery of software tools and algorithms generated by the big data science community. *The scope and mission of the API WG will be defined in the coming months based on discussion with stakeholders.*

BD2K Clinical Informatics WG

- **Working group leaders:** Isaac Kohane (Harvard, PCI Commons), Ida Sim (UCSF, MD2K)
- **Working group contact:** TBD
- **Meetings:** TBD. *An initial meeting of individuals interested in the working group will be held at the November All-Hands Meeting.*

The Clinical Informatics WG focuses on topics related to the use of individual-level clinical information, including in the context of big data computational frameworks, translational considerations, and emergent technical and social issues.

Several of the BD2K Centers of Excellence involve the use of clinical data from a range of information sources (clinical notes, laboratory, imaging, etc.). The Clinical Informatics WG looks at the challenges inherent in using such data not only in the aggregate to drive scientific insights, but at the individual level and its implications for translational efforts, security/privacy, and as new types of personal data (e.g., mHealth, -omic) become prevalent parts of healthcare. *The scope and mission of the Clinical Informatics WG will be defined in the coming months based on discussion with stakeholders.*

BD2K: TRAINING COORDINATION CENTER (TCC)

John Darrell Van Horn, Ph.D.

USC Mark and Mary Stevens Neuroimaging and Informatics Institute

University of Southern California

Overview

- Introducing the BD2K Training Coordination Center (TCC) investigators
- The goals of our Training Coordination Center
- The role of BD2K training working groups
  - What groups exist? Are they open to BD2K affiliates? How do I join?
- TCC web portal for personalizing training resources
  - Mock ups, plans, etc
  - Confined to just BD2K resources? No.
  - Highlighting BD2K resources
  - How does someone get their resource listed
- Coordinating science rotations, workshops, etc
- How to contact us

BD2K TCC Team

USC Laboratory of Neuro Imaging
USC School of Cinematic Arts
GEORGETOWN UNIVERSITY
INI USC Stevens Neuroimaging and Informatics Institute
NIH Big Data to Knowledge (BD2K)
USC Viterbi School of Engineering Information Sciences Institute

Jose-Luis Ambite
Information Sciences In

Rochelle Tractenberg
Georgetown University

Kristina Lerman
Information Sciences In

Michael Taylor
USC School of Cinema

John Berardo
USC Media Institute for Social Change; Shatterproof Films

TCC BD2K Training Coordinating Center

BD2K Training Program Officials

- Michelle Dunn, Ph.D., NIH Office of Data Science
- Erica Rosemond, Ph.D., Division of Clinical Innovation, National Center for Advancing Translational Sciences (NCATS)
- Carol Shreffler, Ph.D., Training & Career Development Program, NIEHS

NIH Data Science at NIH

Introduction

- Training content relevant to Big Data exists in a variety of forms and locations
- Video, printed matter, web content, conference posters, etc.
- Biomedical data hackathons, satellite conferences, special events, etc.
- BD2K Training Grants manage a variety of ongoing training, "summer workshop", and other events.
- Each BD2K Center program has a responsibility for developing and cultivating Big Data training and education
- The BD2K TCC exists to supervise these efforts, cultivate, promote, and leverage them into unique training opportunities
- Moreover, the TCC will help users anywhere locate, curate, and take advantage of these materials for their own personalized training in Big Data biomedicine

BD2K Training Coordination

- The TCC will serve as an outreach resource to all Big Data researchers by
  - Indexing of educational content and materials from BD2K and beyond
  - Hosting a website to facilitate personalized Big Data training ("Big Data U" website)
  - Promote events relevant to Big Data biomedical research
  - Host monthly working group meetings on topics relevant to Big Data training/education
  - Hackathon advertising and organization
  - Lead professional development events
- Coordinate Big Data Science rotations
  - Post-docs and junior faculty
  - Applications due Spring 2016
- Spearhead novel activities relevant to the future of Big Data science

BD2K Center Interactions

- The TCC will work closely with the BD2K Centers Coordinating Center (BD2KCC) to ensure proper coordination of BD2K training activities
- Virtual or in-person visits with each BD2K Training Awardee and COE
- Identify novel training opportunities
- Index training content within BD2K and beyond
  - For instance, work with the BD2KCC and bioCADDIE teams to link instances of data discovery with subsequent educational discovery and vice verse
- Where needed, help to organize, evangelize, and realize these opportunities via the Big Data U web portal

BD2K Training Working Groups

- Monthly conference calls between BD2K training efforts
  - Thank you to Deborah Treu (Morgridge Inst. Res.) for her pioneering efforts!!
  - Convening again in early December 2016 - Watch your email for information!
- BD2K training working groups
  - Plans include but not limited to
    - BD2K Training Standards
    - Hack-a-thon Working Group
    - Training Program Evaluation
  - Initially involving training liaisons but expected to grow over time
- Contact us for additional details on how to get involved

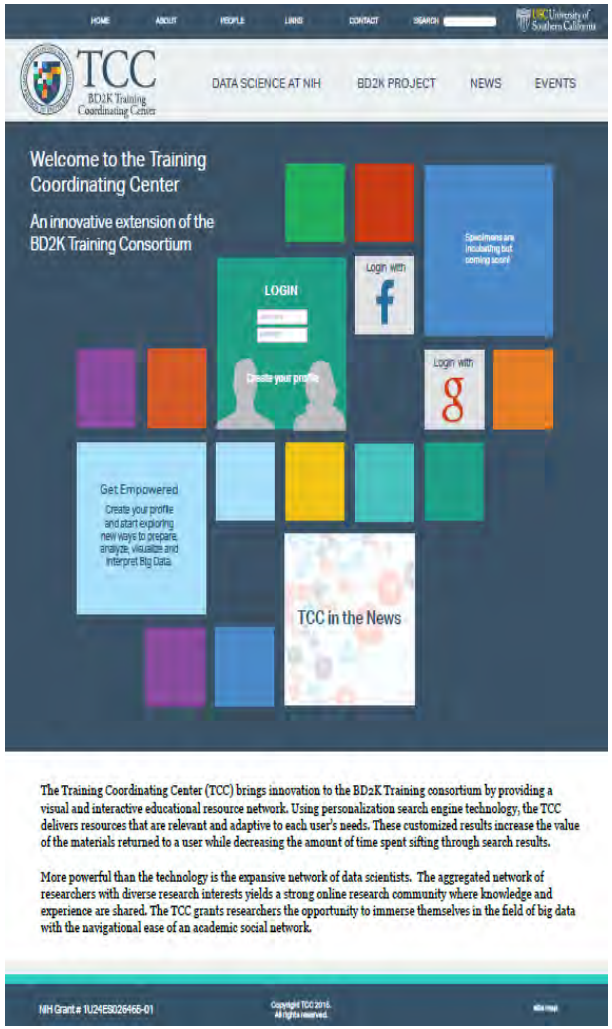
BD2K TCC "Big Data U" Website

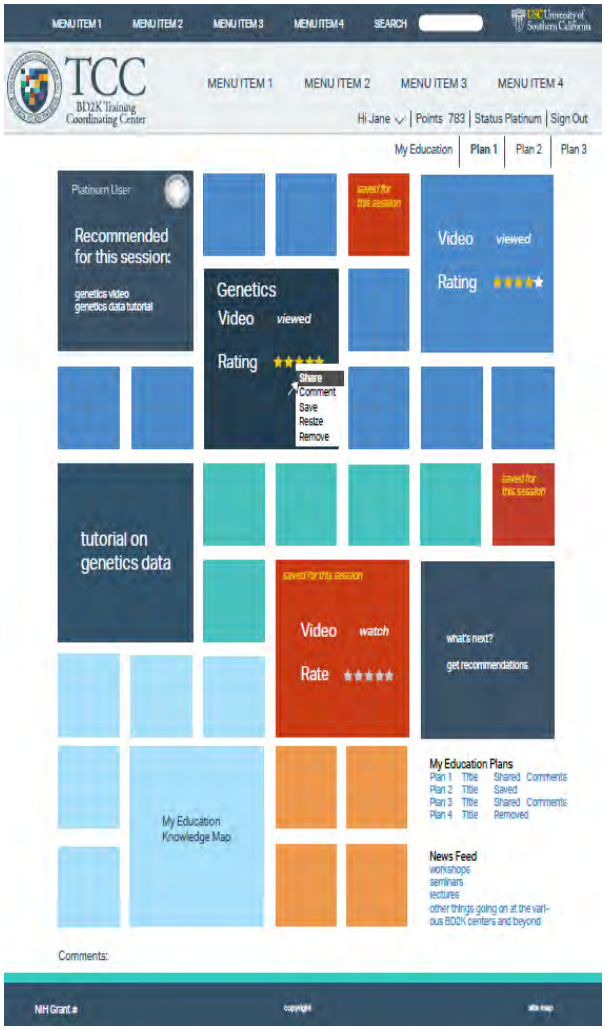
COMING SOON!

www.bigdatau.org

bigdatau.ini.usc.edu

- Your personalized "university" for Big Data biomedicine
- Graphical, sortable, dynamic, searchable, scalable
- Build your own training curricula
- Instantiation of the BD2K ERUDITE Knowledge Map
- Content automatically identified
- Easily link your own content
- BD2K program-specific





BD2K TCC Staffing

- Project Manager
- Communications Specialist
- Educational Outreach Specialist
- Science Education Specialist
- Website Developer
- Web Services Programmer
- Ontology/Database Expert

Contact the BD2K TCC

USC Mark and Mary Stevens Neuroimaging and Informatics Institute

University of Southern California

2001 North Soto Street, SSB1-102

Los Angeles, CA 90049

URL: http://bigdatau.ini.usc.edu, http://www.bigdatau.org

Phone: 323-44-BRAIN

Email: jvanhorn@usc.edu

TCC
BD2K Training Coordinating Center

BD2K Standards Coordination Center

BD2K All Hands Meeting

November 13, 2015

Anand Basu

ESAC
www.esacinc.com

BD2K Standards Coordinating Center (SCC)

- Overview:
  - Contract awarded: Sept 30, 2015
  - Period of Performance: Oct 1, 2015 – Sept 30, 2017
  - Program Manager: Anand Basu (anand.basu@esacinc.com)
  - Project Manager: Laura Sheahan (laura.sheahan@esacinc.com)
  - Our offices are located in Rockville, MD

SCC - Goals

- Serve as a standards resource for BD2K and the biomedical research community to:
  - Collaborate
  - Communicate
  - Collect and Share
- Collaborate
  - Work with the BD2K Centers and research community
- Communicate
  - Standards resources, guidance documents, emerging standards, best practices
- Collect and Share
  - Disseminate using a web portal, webinars, newsletters, etc.

SCC – Program Plan

- Initial outreach to BD2K Centers
- Establish contacts
  - Understand needs and identify use cases
- Establish working groups of interested participants
  - Domain based working groups
  - Cross-domain participation
- How can I participate?
  - Join our listservs, Google groups
  - Register on our portal (coming soon) for a working group
  - Contact us directly

BD2K Standards Coordination Center
Collaborate

- Collaborate with existing stakeholder groups and researchers to address standards needs
- Create a robust, interactive, collaborative resource for the biomedical research community

Organization	Collaboration
NIH BD2K CoEs	Standards needs assessment
NIH BD2K Programs	Reference standards programs
International Standards Consortia	Standards references
Other Stakeholders	Needs and standards references

BD2K Standards Coordination Center
Communicate, Collect, and Share

- Monthly ‘newsletter’ – focusing on a particular domain or standard
- Conduct monthly webinars to highlight use of standards, as well as the SCC resource
- Develop a use case driven standards web portal to facilitate and provide means for:
  - Discovery and browsing
  - Links to existing resources
  - Q&A forums
  - Discussion forums
  - Standards guidance documents
  - Other resources

BD2K Standards Coordination Center

Our Organization

NIH BD2K
Program Office/COR

Program Manager
Anand Basu

Project Manager
Dr. Laura Sheahan

SCC Advisory Committee

- Karen Ketchum, PhD (ESAC)
- Jessica Tenenbaum, PhD (Duke)
- Cheryl Kraft, MS, PMP (biomedical IFX consultant)
- Mark Musen, MD, PhD (Stanford)
- Chris Chute, MD, PhD (JHU)
- Melissa Haendel, PhD (OHSU)

Portal Development Team

Lead: Michael Holck

Standards Curation Team

Lead: Sweta Ladwa

Community Outreach Team

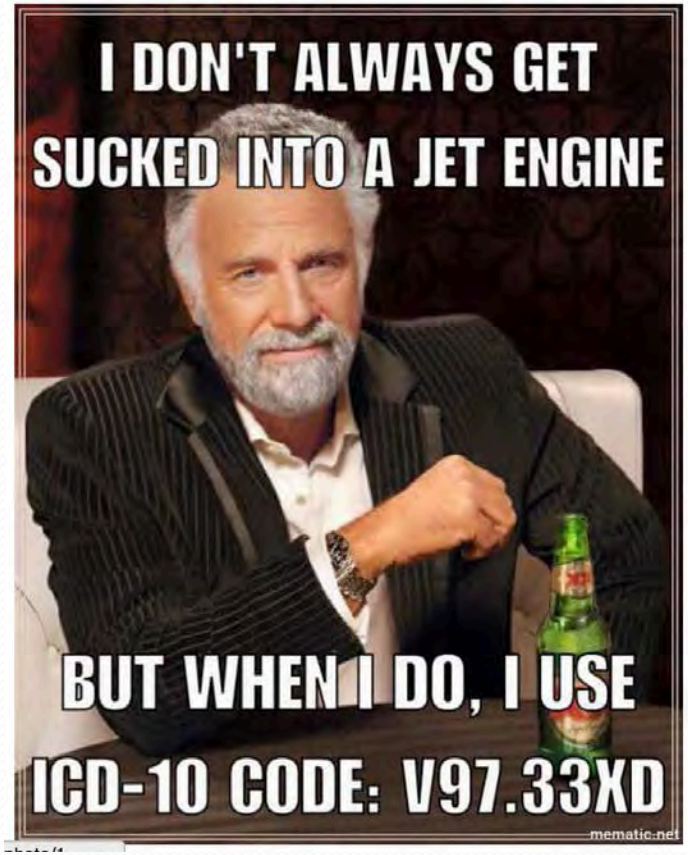
Lead: Dr. Rhonda Schwartz

ESAC
www.esacinc.com

SCC Contacts

- Contact Us: bd2kscc@esacinc.com
- Portal Address: www.bd2kbiostandards.org (coming soon!)
- Our Contacts:
  - Anand Basu (anand.basu@esacinc.com)
  - Laura Sheahan (laura.sheahan@esacinc.com)
  - Rhonda Schwartz (rhonda.schwartz@esacinc.com)
  - Sweta Ladwa (sweta.ladwa@esacinc.com)
  - Karen Ketchum (karen.ketchum@esacinc.com)

Thank You



We'd like to thank the NIH, and the BD2K Program Committee for recognizing the importance of standards – we no longer just want to be the source of endless humor but your partners in shareable, interoperable and repeatable Data Science.

ESAC
www.esacinc.com