

Update from the Associate Director for Data Science

BD2K & ADDS activities

Philip E. Bourne, PhD, FACMI
Associate Director for Data Science

▶ BD2K Multi-Council Working Group ▶ September 1, 2016





Outline

- Scientific Data Council (SDC) Update
- NIH Task Force on Data Sustainability & Maintenance
- BD2K/ADDS highlights since April 2016
- Upcoming events
- Response to prior MCWG requests



Scientific Data Council (SDC) Update

- SDC reorganized to expand NIH Data Sharing and BD2K oversight
- NLM Director, Patti Brennan joins SDC
- RFIs
 - Metrics to Assess Value of Biomedical Digital Repositories (closes 9/30)
 - Data Annotation in Biomedical Core Research Facilities and Related Needs for Community Education and Training (closed 6/30)
 - Forthcoming RFIs
 - Pre-Prints and other Interim Research Products
 - Data Sharing Strategy, including Data and Software Citation



Task Force on Data Sustainability & Maintenance

- To address a pressing problem to which BD2K contributes, but does not solve
- Chaired by the NIH Director
- Agreement to support the Commons and the FAIR principles as an institution
- Addressing:
 - Supplemental data associated with publications
 - Large data sets in which NIH has invested
- Task Force aims are aligned to BD2K
 - e.g., in standards development, cloud procurement, data wrangling, making data and resources FAIR,...

Highlights Since the Last MCWG ...

- The DataMed data indexing tool released – 23 repositories, 650,000 datasets indexed to date – feedback sought
- 10 Commons Conformant providers onboard, including major cloud providers
- Added NIH 64 data repositories to Healthdata.gov (with BMIC)
- Workshop with Common Fund on migrating large datasets to the Commons
- Six finalists of the [Open Science Prize](#) were announced at the 7th Health Datapalooza in Washington DC on 5/9

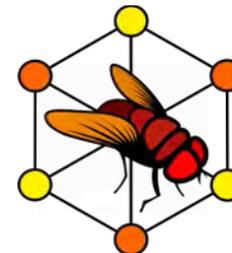


Open AQ



Real-time Pathogen
Surveillance

OpenTrialsFDA



Fruit Fly Brain
Observatory



Open
Neuroimaging
Lab



More Highlights

- **84** trainees (T15/32), **25** courses (R25).
- Software and tools from U54 and U01 Data Science Research Awards
- BD2K Centers of Excellence highlights
 - Development of a prototype of the Mobilize Data Hub
 - Development of analytical and software approaches to integrate LINCS and TCGA expression profiling datasets
 - Cancer Gene Trust
- Special meetings: training, west coast meetings organized by grantees
- Standards Coordinating Center contract not renewed

Collaboration - Count Everything: Integrating Clinical Genomics, and mHealth APIs

- Collaboration across 4 BD2K awards: 3 BD2K Centers and bioCADDIE.

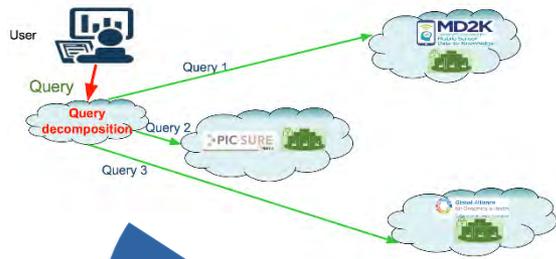


Global Alliance
for Genomics & Health

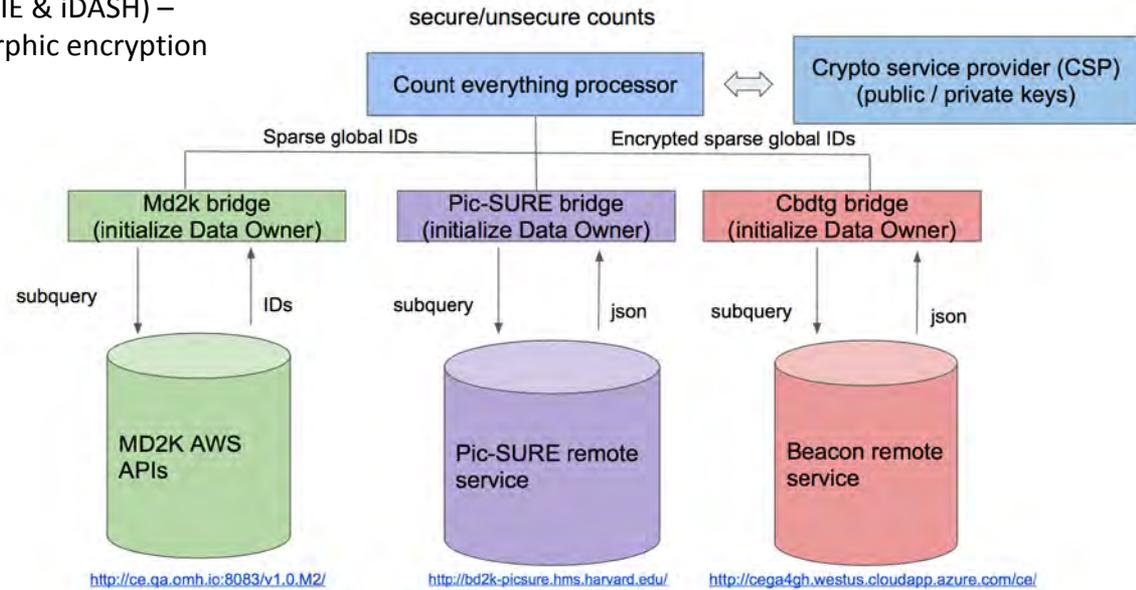


- Developing Interoperable APIs and Federated Security, to find and access data more easily.
 - **Genomic Data** (GA4GH/BDTG) – 1000 Genomes
 - **Clinical Data** (PIC-SURE) – CDC NHANES
 - **mHealth** (MD2K) – Synthetic data, and data from Genomes & Activity
 - **Federated Security and Search** (bioCADDIE & iDASH) – homomorphic encryption

Count Everything: Queries with Homomorphic Encryption Using Interoperable APIs and Federated Security finds and accesses data securely



Federated Security and Search
(bioCADDIE & iDASH) –
homomorphic encryption



**Simple ideas to
Successful Queries**

mHealth (MD2K)
Synthetic data, and data
from Genomes & Activity

**Clinical Data (PIC-
SURE)**
CDC NHANES

**Genomic Data
(GA4GH/BDTG)**
1000 Genomes

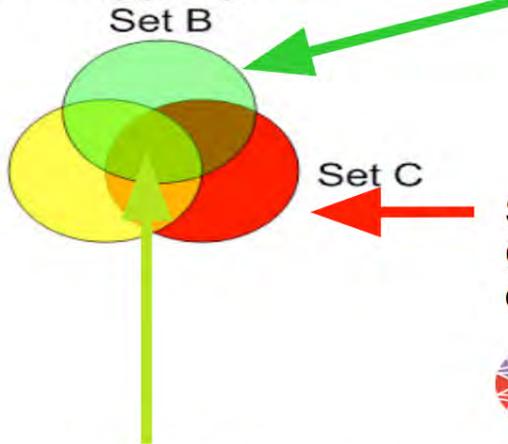
Count Everything: Query Decomposition



Select IDs from MD2K A where $A.\text{bloodGlucose} > 110$ [mg/dL]

Secure aggregation

Set A



Select IDs from PIC-SURE B where $B.\text{smoker} = \text{true}$
Select IDs from GA4GH C where $C.r123140 = T$

Set C

Select count(*) from MD2K A, PIC-SURE B, GA4GH C where $A.\text{bloodGlucose} > 110$ and $B.\text{smoker} = \text{true}$ and $C.r123140 = T$



Global Alliance for Genomics & Health
Collaborate. Innovate. Accelerate.

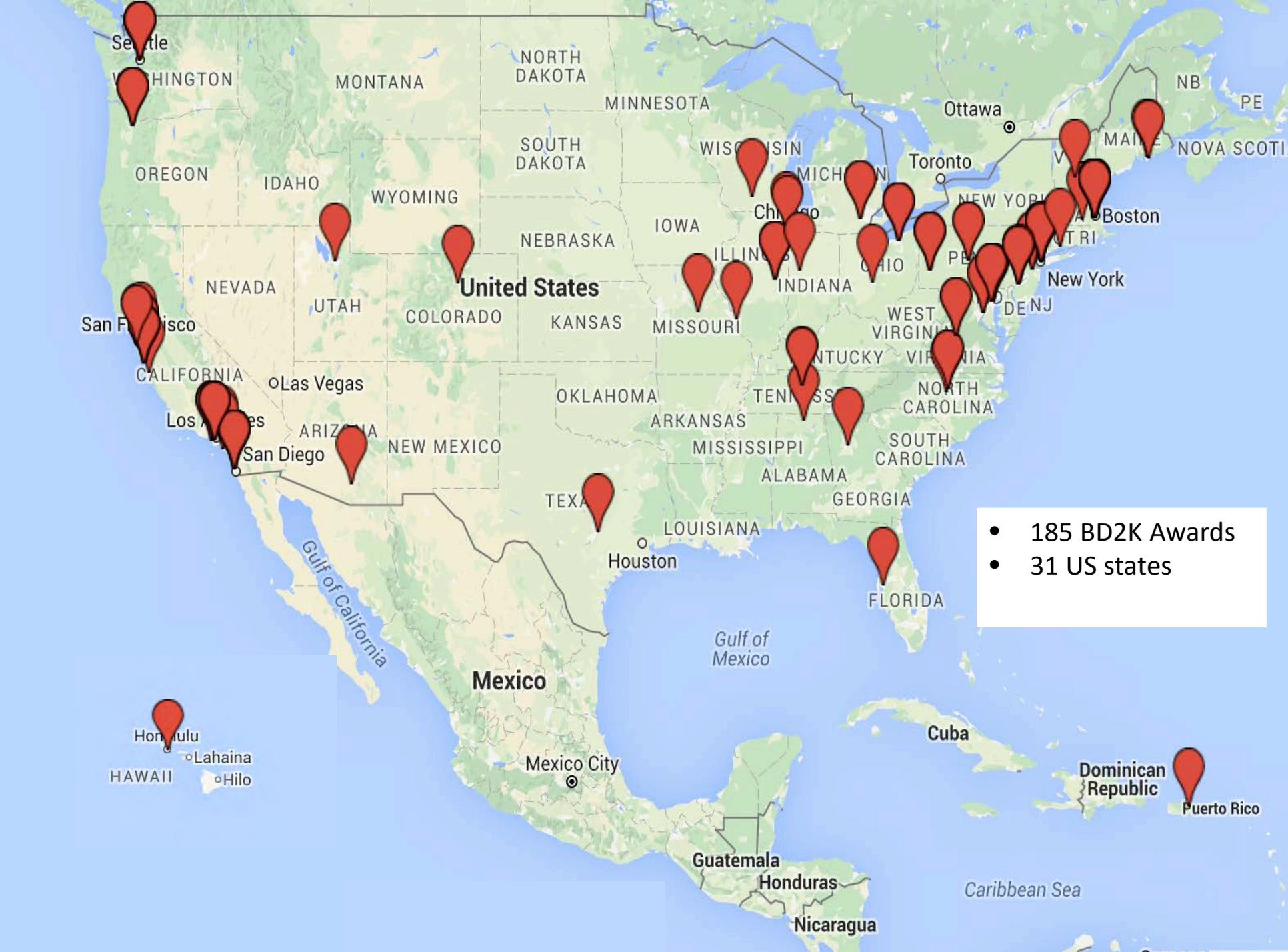




BD2K Funding Update

- **120/185*** awards, **30/31** US states & **3** countries.
- **10/12*** different mechanisms (U54, U01, U41, R01, U24, UH2, T32, T15, K01, R25, R44, P50).
- **14** new FOA concepts
 - 9 published since January 2016
- **27** supported and/or organized workshops and meetings

* May/Now



- 185 BD2K Awards
- 31 US states

Examples of New Data Science Research & Training Opportunities

- Big Data to Knowledge (BD2K) Enhancing the Efficiency and Effectiveness of Digital Curation for Biomedical Big Data (U01) [RFA-LM-17-001](#)
- Big Data to Knowledge (BD2K) Community-Based Data and Metadata Standards Efforts (R24) [RFA-ES-16-010](#)
- NIH Big Data to Knowledge (BD2K) Enhancing Diversity in Biomedical Data Science (R25) [RFA-ES-16-011](#)
- Notice of NIH/BD2K Participation in the Joint NSF/NIH Initiative on Quantitative Approaches to Biomedical Big Data (QuBBD) – Focus on mobile health [NOT-EB-16-008](#)
- Data Science Rotation for Advancing Discovery Trip (RoAD-Trip), through the Training Coordination Center
- Curriculum Development FOA (endorsed by the ACD)

Please promote in your respective communities!



Creating a Collaborative Data Science Community

- Supporting communities e.g. GA4GH, RDA, FORCE11, PLOS.
- Innovation Lab and QuBBD program with NSF.
- International biomedical funder collaborations:
 - Wellcome Trust & HHMI - open science prize.
 - ELIXIR - Alignment of tools and research related to indexing and standards.
 - European Open Science Cloud - Shared vision of implementing the Commons Framework.



Open Science Prize

- 96 submissions received
- Solvers from 45 countries spanning 5 continents
- Over 100 entities represented
- Phase 1 finalists will present their prototypes on December 1 at the Open Data Science Symposium of the AHM
- Overall winner to be announced 2/17





Looking Forward

- Significant BD2K presence:
 - *International Data Week September 11-17, 2016*
- Vibrant exchange of ideas and expression of support for open science:
 - *All Hands Meeting and Open Data Science Symposium, Nov 29-Dec 1, 2016.*

Addressing your Concerns



BD2K increased outreach to PMI and Other Flagship Projects

- ADDS Phil Bourne has engaged:
 - Precision Medicine Initiative (PMI) Director (Eric Dishman)
 - Environmental influences on Child Health Outcomes (ECHO) Director (Matthew Gillman)
 - NLM Director (Patti Brennan)
- Proposing meeting with ADDS/PMI/ECHO/BRAIN staff to foster and strengthen collaboration
- Consulting with the BRAIN initiative on new FOAs with a strong Data Science component



Reference Data Sets

- We have not as yet done anything regarding synthetic datasets
- Working with the Common Fund and considering the following data to be available in the Commons and FAIR:
 - GTEx
 - Epigenomics
 - KOMP
 - LINCS
 - HMP2
 - Metabolomics
- Welcome a discussion of other actions



BD2K Metrics and Evaluation

- May BD2K, MCWG asked BD2K to address metrics of success: How will it be evaluated?
- BD2K/NIH working group developing a plan
 - Allen Dearry
 - Susan Gregurick
 - Mark Guyer
 - Jennie Larkin
 - Elizabeth Kittrie
 - Sonynka Ngosso



BD2K Metrics and Evaluation

Four areas under consideration:

- Assess implementation of the original DIWG recommendations
- Update strategic goals for BD2K (rapidly moving field since 2012)
- Evaluate success of individual BD2K programs/awards
- Evaluate success of the overall BD2K initiative



Assess implementation of the original DIWG recommendations

DIWG Recommendations

1. Promote Data Sharing Through Central and Federated Catalogues
2. Support the Development, Implementation, Evaluation, Maintenance, and Dissemination of Informatics Methods and Applications
3. Build Capacity by Training the Workforce in the Relevant Quantitative Sciences such as Bioinformatics, Biomathematics, Biostatistics, and Clinical Informatics



Assess implementation of the original DIWG recommendations

- Propose mapping of BD2K activities and FOAs to the DIWG recommendations.
 - How was each recommendation addressed?
 - # of FOAs, # of awards, \$
 - Balance of effort across the 3 recommendations.
- Present analysis to MCWG, January 2017.



Updated strategic goals for BD2K

1. Facilitate broad use of biomedical digital assets by making them Findable, Accessible, Interoperable, and Reusable (FAIR).
2. Conduct research and develop the methods, software, and tools needed to analyze biomedical Big Data.
3. Enhance training in the development and use of methods and tools necessary for biomedical Big Data science.
4. Support a data ecosystem that accelerates discovery as part of a digital enterprise.

<https://datascience.nih.gov/bd2k/about>



Evaluate success of individual BD2K programs and the overall BD2K initiative

- Evaluate coherence and success of the overall program by mapping to the updated BD2K Strategic Goals
 - Identify where investments being made
 - Identify possible gaps
- Evaluate Success of each Program against the RFA/Program Goals
 - One set of metrics cannot be used across all awards. (e.g., K01 versus T32 versus UH2 versus U54)



Evaluate success of individual BD2K programs and the overall BD2K initiative

- Evaluate coherence and success of the overall program by mapping to the updated BD2K Strategic Goals
 - January 2017 MCWG, provide initial mapping of Program/RFA goals to BD2K overall strategy
- Evaluate Success of each Program against the RFA/Program Goals
 - January 2017, provide proposed evaluation criteria for success of each program.
 - Actual metrics cannot be measured for some time, as many awards are new.

Data Science at NIH

- ▶ <https://datascience.nih.gov>
- ▶ bd2k@nih.gov
- ▶ @NIH_BD2K
- ▶ #BigData, #NIH_BD2K

