NIH Big Data to Knowledge
BD2K

Belinda Seto, Ph.D.
Deputy Director
National Eye Institute
Data Trends

- Large datasets
  - Human connectome, 2TB in 1st data release (500 releases so far)
  - 46 million EHRs with Aetna
- Diverse data types: genomics, imaging, electronic health records, mobile health
- Data integration
- Analysis: need for computation resources
- Collaboration: data sharing
  - *NIH Plan for Increasing Access to Scientific Publications and Digital Scientific Data from NIH Funded Scientific Research*
Human Connectome Project

Human Connectome Background

- The Phase 1 of the Human Connectome Project (HCP) was completed in October 2015.
- Final data release (including GWAS, imaging, and behavioral data) for 1200 subjects will occur no later than June 2016.
- The research participants were 22-35 years old.
- Over 100 papers have appeared from the n=500 data release.
Opportunities and Challenges in Integration and Exchange of EHR data

- Large clinical databases can be used to assess risk factors for disease in lieu of costly population-based studies
- EHR-based phenotype: retrospective case studies or predictive
- Multiple EHR systems used in hospitals and medical practices
- Interoperability: syntax; data format
  - Visual acuity 20/40 in a single field vs 2 fields 20 and 40
  - Semantic interoperability with unstructured data: 20/40 with correction vs without correction
  - Sharing clinical data is challenging
Integration of Brain Images and Genomic Data to Predict Disease Risks
Brain measures versus epidemiological studies to find genetic variants that directly affect the brain
difficult
May require 10,000-30,000 people
e.g., the Psychiatric Genetics Consortium studies
easier?
Gene variants (SNP’s) may affect brain measures directly, many brain measures relate to disease status.
Finding Genetic Variants Influencing Brain Structure

Phenotype

Genotype

Intracranial Volume

C/C  A/C  A/A

Association
One SNP
“Candidate gene” approach

Screening
500,000 SNPs – 2,000,000 SNPs

NIH-funded database of genotypes and phenotypes enabling searches to find where in the genome a variant is associated with a trait.
Alzheimer’s risk gene carriers (**CLU-C**) have lower fiber integrity even when young (N=398), 50 years before disease typically hits. Voxels where **CLU** allele C (at rs11136000) is associated with lower FA after adjusting for age, sex, and kinship in 398 young adults (68 T/T; 220 C/T; 110 C/C). FDR critical p = 0.023. Left hem. on Right. 

[Image: Brain MRI scan showing regions with lower FA associated with **CLU-C** allele C.]
BD2K: An ecosystem to extract knowledge from large, structured or unstructured datasets
BD2K Funded Programs

- NIH commitment for 7 years
- 12 centers of excellence: research programs to develop new approaches, methods and software tools to address specific biomedical questions
- Data discovery index coordination consortium: to readily find, cite and access existing data and software
- 15 targeted software development awards: tools to address data compression, Provenance, visualization, and wrangling
- 21 training awards
Centers of Excellence

➢ $31.4M in FY 2016
➢ Investigators initiated programs
➢ Addressing range of data science challenges:
  ▪ Large volume of unstructured data from mobile devices
  ▪ Data from electronic health records
  ▪ Integrating genomic data with phenotypes, including images
➢ Share data and tools
mHealth Big Data

- 2 Centers of Excellence funded on mHealth big data
- High volume, velocity, variety, variability, versatility
- Data analytics to address medical research questions
  - Reducing relapse of smoking cessation
  - Reducing readmission among congestive heart failure patients
**BD2K Training Goals**

<table>
<thead>
<tr>
<th>Goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Establish biomedical data science as a career path</td>
</tr>
<tr>
<td>Foster collaborations between biomedical scientists and data scientists</td>
</tr>
<tr>
<td>Develop and improve data science skills in the biomedical workforce</td>
</tr>
<tr>
<td>Enhance diversity in the biomedical data science workforce</td>
</tr>
<tr>
<td>Ensure training opportunities and resources are more readily discovered and accessed</td>
</tr>
</tbody>
</table>
Programs Address Varying Audiences

- Short Courses (R25) - 11
- Open Educational Resources (R25) - 8
- Diversity (R25) - 4
- All Biomedical Scientists
- Biomedical Data Scientists
- Training Programs (T32/T15) – 6+10
- Career Development Awards (K01) - 21

Audiences:
- Undergrad
- Graduate
- Postdoc
- Junior Faculty
- Senior Faculty
Current Funding Opportunity Announcements

- Early stage development of technology in biomedical computing, informatics, and big data science (R01)
  - PA-14-154, 5/08/2017
  - PA-14-155
  - PA-14-156
  - PA-14-157

- Revisions to add biomedical big data training to active institutional training grants (T32)
  - RFA-HG-14-005, 7/29/2016
NIH Data Commons Concept

- Treats products of research – data and associated meta-data, methods, papers etc. as digital objects (machine-readable)
- These digital objects exist in shared virtual space
- Collections of public and commercial resources (including cloud resources) for storing data and computing tools
  - Uniquely identified, sharable, and re-usable
FAIR Principles on Big Data

- **Findable**
  Uniquely and persistently identifiable
- **Accessible**
  Retrievable by machine or human
- **Interoperable**
  Open, well-defined vocabulary
- **Reusable**
  Machine processable with data citation

source: go.nature.com/axkjiv
Implementing the Common

- Digital object (data and meta-data) should comply with FAIR principle
- Interoperability
  - Data format should use shared vocabularies /ontologies
  - Enable connectivity and shareability between digital objects
  - Data is machine-actionable
- Computing platform
  - Cloud or HPC (High Performance Computing)
  - Hardware that supports access, utilization, sharing and storage of digital objects
The Commons: Pilots Underway

- BD2K Centers (e.g., GA4GH Beacons)
- Model Organism Database Interoperability
- HMP Data and tools available in the cloud
- NCI Cloud Pilots & Genomic Data Commons
- Commons business model – credits
NCI Genomic Data Commons (May 2016 Launch)

- A centralized and standardized cancer knowledge base with genomic and clinical data
- More than a data repository
- Also provides analysis tools and opportunities to harmonize data from multiple datasets
- Data from 4 NCI programs and other cancer genomic projects
  - The cancer genome atlas: adult genome data, will be completed in 2016, 2.5 petabytes data
  - Therapeutically Applicable Research to Generate Effective Therapies (TARGET), childhood cancer data
  - Cancer Genome Characterization Initiative
  - The Cancer Cell Line Encyclopedia
Cancer Genomics Cloud Pilots

- A single platform with co-located data, computational tools and API to enable access to data repositories
- 3 contracts: Broad Institute; Institute for systems biology; Seven Bridges Genomic
- Commercial cloud providers
- Available to researchers in early 2016
- Will enable researchers to search, analyze and visualize own data in combination with NCI data
Current Topics under Discussion

- **Data sharing**
  - Protect privacy, proprietary interests, preserve balance between the benefits of access/preservation and costs
  - Encourage all NIH-funded researchers submit data management and sharing plans with application for peer review
  - Encourage use of established repositories and community-based standards

- **Data resource sustainability**

- **Data citation:** citing data in the same way as we do for publications

- **Common data elements**
Data Resource Sustainability

- Largest 50 NIH funded data resources, annual budget estimate, $110M
- Need efficient models for storing, organizing and accessing data (located in a variety of repositories, different modes of access)
- Data usage, supply/demand
- Possible business models?
- Computing credits: grantees receive credits for computational time instead of funds, spend credits with any Commons-compliant resource providers (including cloud services); pay only for services used
Stay tuned to: https://datascience.nih.gov