Advancing the Capabilities of Data Commons

Alastair Thomson BSc(Hons) DipGrad

Director for Data Innovation, ARPA-H

NIH NATIONAL CANCER INSTITUTE

ARPA H

Data Commons

- Demonstrated value to the research community across domains
  - CRDC
  - NHLBI BioData Catalyst™
  - Kids First Data Resource/Cavatica
  - Elixir

But there are challenges

- FAIRness of data has been improved
  - But mostly more Findable and Accessible
  - Interoperability and Reusabality is still a challenge
- We’ve broken down siloes in individual institutions
  - But built still larger siloes
  - Separated by disease focus, funder, etc.

The next phase of evolution: Data Meshes

- Interconnected Data Commons to enable:
  - Building of cross commons cohorts
  - Bridging between domains for analysis
- Common APIs e.g. GA4GH DRS, WES, TES, Beacon
- Standards for automated connectivity: Mesh Cards
- E.g. European Genomic Data Infrastructure

But Data Meshes are not a panacea

- Cross data commons search can be difficult
  - Without a master index, forces separate searches in each commons
- Varying data models and data representations
  - Data requires harmonization, often difficult as experimental conditions and intent are not captured
- Varying polices and processes for data access

Towards a Comprehensive Data Fabric

- A Data Fabric provides the foundations for a highly interoperable data eco-system
  - Common data models e.g. OMOP, FHIR
  - Harmonized data
    - With captured experimental intent
  - Master Data and Metadata Indices

Biomedical Data Fabric (BDF) Toolbox

Vision: Develop a reuseable, easily deployable data fabric to maximize the usability of research data for researchers, patients, and clinicians, while reducing the human effort needed to generalize data fabric capabilities across multiple disease.

Technology focus areas

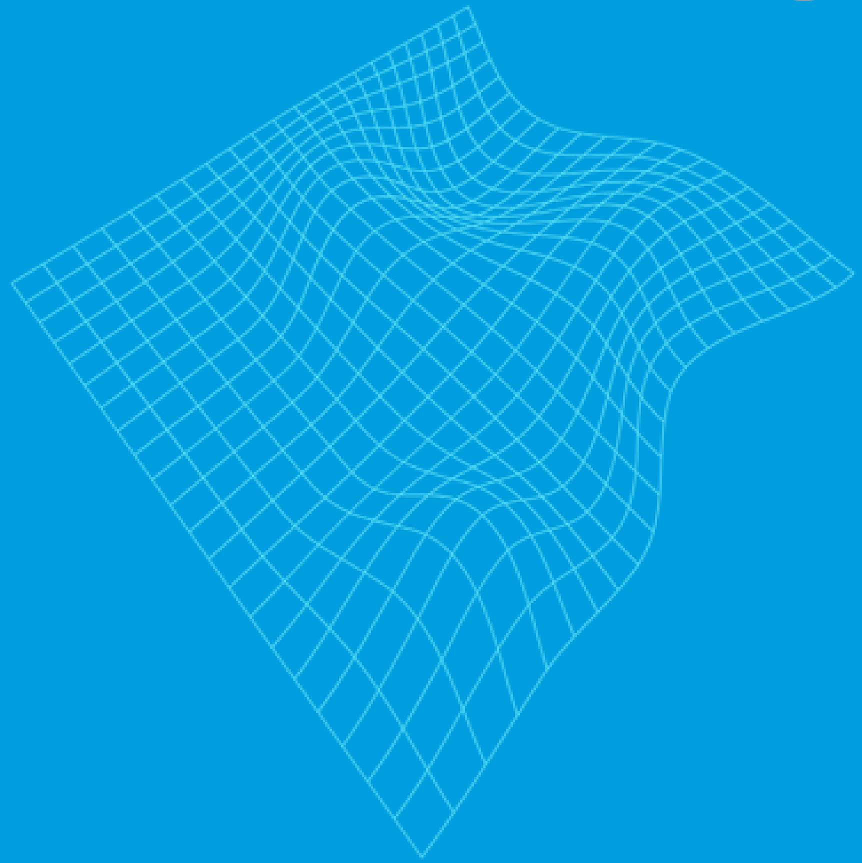
- Automated data collection;
- Machine-assisted data curation;
- Intuitive data exploration;
- User testing;
- Cross-domain generalization of best-in-class capabilities.

Approach

- Develop automated workflows to reduce time and effort to collect data from labs and electronic health records.
- Apply next-generation AI/ML approaches to automate multi-modal data integration and analysis
- Advance capabilities for quality assurance/quality control, de-identification, and equity checks.
- Enhance data discovery and exploration using AI/ML
- Iterative user testing with broad range of stakeholders for continuous feedback and development of tools

Key Dates/Links

Program announcement (September 2023)



What if new data integration tools made it possible to get more value out of the health research data produced by thousands of labs and hospital centers?

ARPA-H

Approved for Public Release: Distribution Unlimited

What if…

- We could use AI to analyze the life story of a person, understand their exposure to pollutants, their risks due to financial pressures, education, housing and food insecurity, their health and healthcare history, and genomic risk factors…
- And then proactively guide them and their physician to the care needed based on their risks e.g. early screening for lung cancer, heart disease etc.

The Big Vision

- A data fabric that connects disparate data resources
  - Clinical and Observational Data across domains and diseases
  - Multi-omics, Imaging
  - Real World Data
    - Electronic Health Records
    - Geolocation data e.g. housing history
    - Social Determinants of Health
    - Pollutant exposure e.g. particulates
    - Toxicology
  - All linked using advanced privacy preserving record linkage
  - Harmonized using AI
  - Indexed into a master data index with descriptive statistics
  - Implemented as federated nodes with local control and sovereignty