Beyond Today's Internet
Experiencing a Smart Future

Prototype SDX Bioinformatics Exchange: Demonstrating an Essential Use-Case for Personalized Medicine

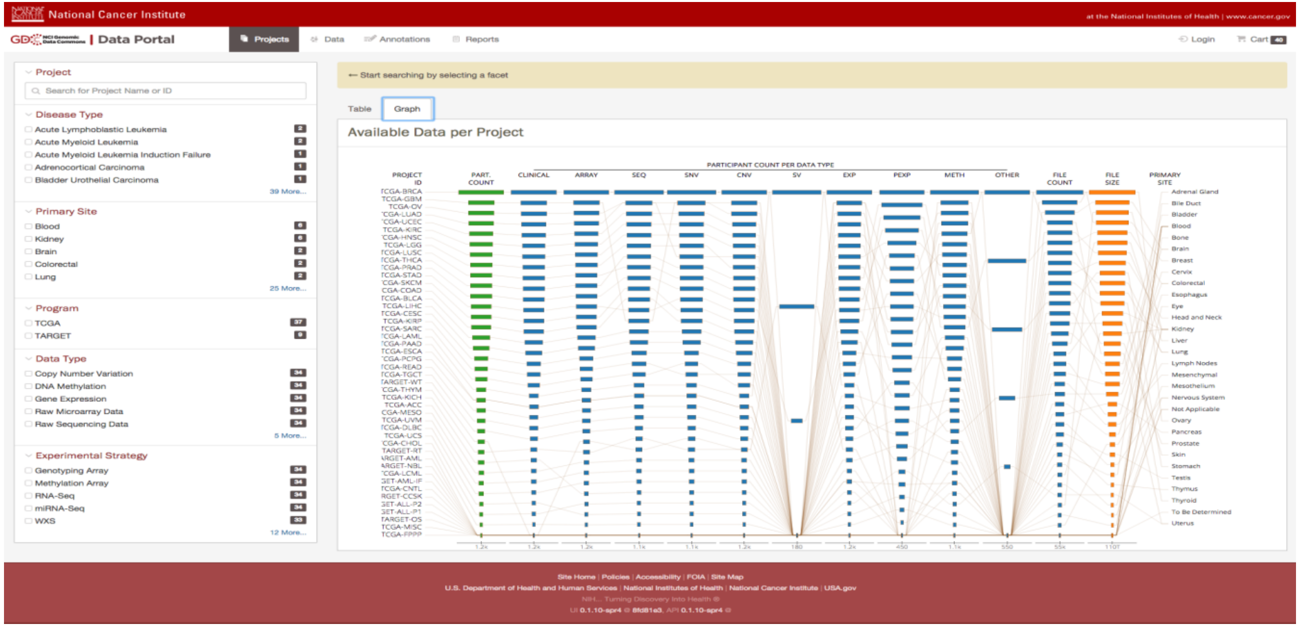
Robert Grossman – University of Chicago
Joe Mambretti – Northwestern University
Piers Nash – University of Chicago
Jim Chen – Northwestern University
Allison Heath – University of Chicago

Precision Medicine

- Precisely match treatments to patients and their specific disease
- Genomic data promises optimal matching.
- 1.7 million cancer cases diagnosed in America each year.
- A single RNA-seq file is 10-20 GB, Whole genome raw data files are > 100 GB.
- Analysis has become the bottleneck and data size is an issue.
    - 2,000,000 genomes ≈ 1 Exabyte (1,000,000,000,000 MB)
    - Cost to sequence 1 genome less than $5,000 and falling fast.
    - Cost to analyze 1 genome is approx. $100,000 and rising.
- A key step towards Algorithm-assisted Personalized medicine is building Data Commons/Cloud analytics and the *Programmable* Networks & Communication Exchanges (SDXs) for high performance, flexible data transport.

Infrastructure for Precision Medicine

NCI Genomic Data Commons



National Cancer Institute
at the National Institutes of Health | www.cancer.gov
GDC NCI Genomic Data Commons | Data Portal
Projects
Data
Annotations
Reports
Login
Cart 40

Project

Search for Project Name or ID

Disease Type

□ Acute Lymphoblastic Leukemia	2
□ Acute Myeloid Leukemia	2
□ Acute Myeloid Leukemia Induction Failure	1
□ Adrenocortical Carcinoma	1
□ Bladder Urothelial Carcinoma	1

39 More...

Primary Site

□ Blood	6
□ Kidney	6
□ Brain	2
□ Colorectal	2
□ Lung	2

25 More...

Program

□ TCGA	37
□ TARGET	9

Data Type

□ Copy Number Variation	34
□ DNA Methylation	34
□ Gene Expression	34
□ Raw Microarray Data	34
□ Raw Sequencing Data	34

5 More...

Experimental Strategy

□ Genotyping Array	34
□ Methylation Array	34
□ RNA-Seq	34
□ miRNA-Seq	34
□ WXS	33

12 More...

← Start searching by selecting a facet

Table | Graph

Available Data per Project

Site Home | Policies | Accessibility | FOIA | Site Map

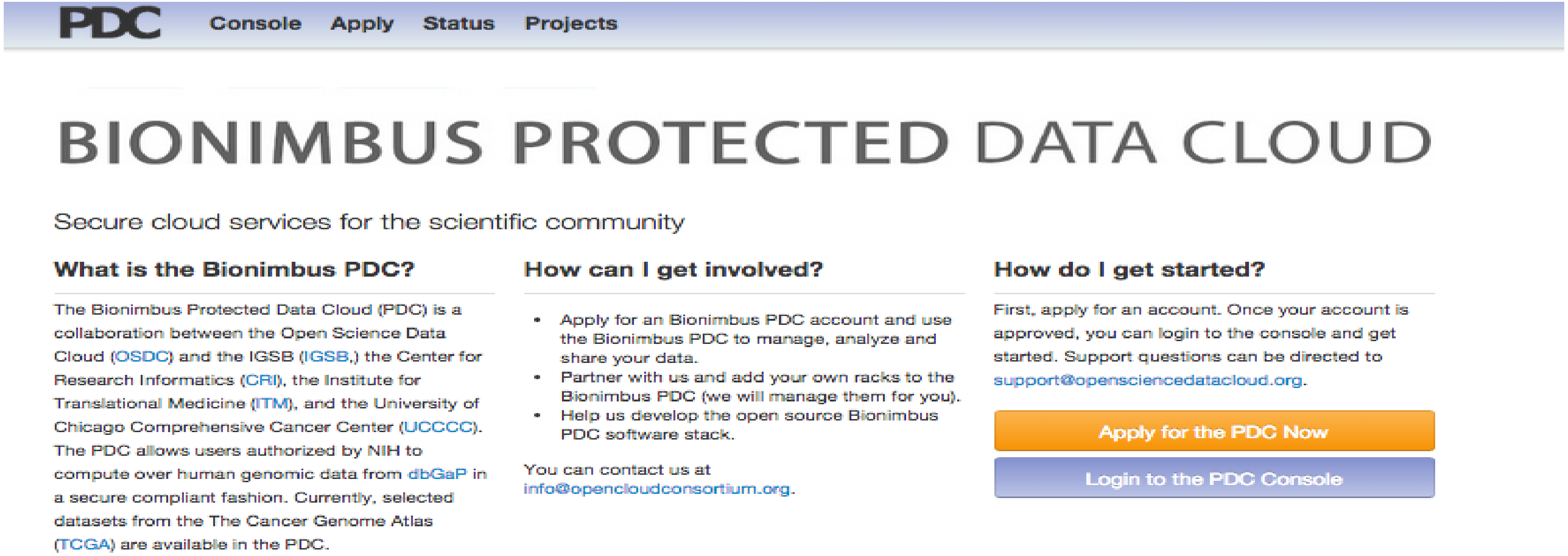
U.S. Department of Health and Human Services | National Institutes of Health | National Cancer Institute | USA.gov

NIH... Turning Discovery Into Health ®

UI 0.1.10-spr4 @ 8fd81e3, API 0.1.10-spr4 @

- Harmonization and storage for the Nations Cancer Genomic Data GDC 1.6PB of cancer genomic data and associated clinical data.
- **Precision Medicine Enabled By Precision Networking**

Bionimbus Protected Data Cloud



BIONIMBUS PROTECTED DATA CLOUD

Secure cloud services for the scientific community

What is the Bionimbus PDC?

The Bionimbus Protected Data Cloud (PDC) is a collaboration between the Open Science Data Cloud (OSDC) and the IGSB (IGSB,) the Center for Research Informatics (CRI), the Institute for Translational Medicine (ITM), and the University of Chicago Comprehensive Cancer Center (UCCCC). The PDC allows users authorized by NIH to compute over human genomic data from dbGaP in a secure compliant fashion. Currently, selected datasets from the The Cancer Genome Atlas (TCGA) are available in the PDC.

How can I get involved?

- Apply for an Bionimbus PDC account and use the Bionimbus PDC to manage, analyze and share your data.
- Partner with us and add your own racks to the Bionimbus PDC (we will manage them for you).
- Help us develop the open source Bionimbus PDC software stack.

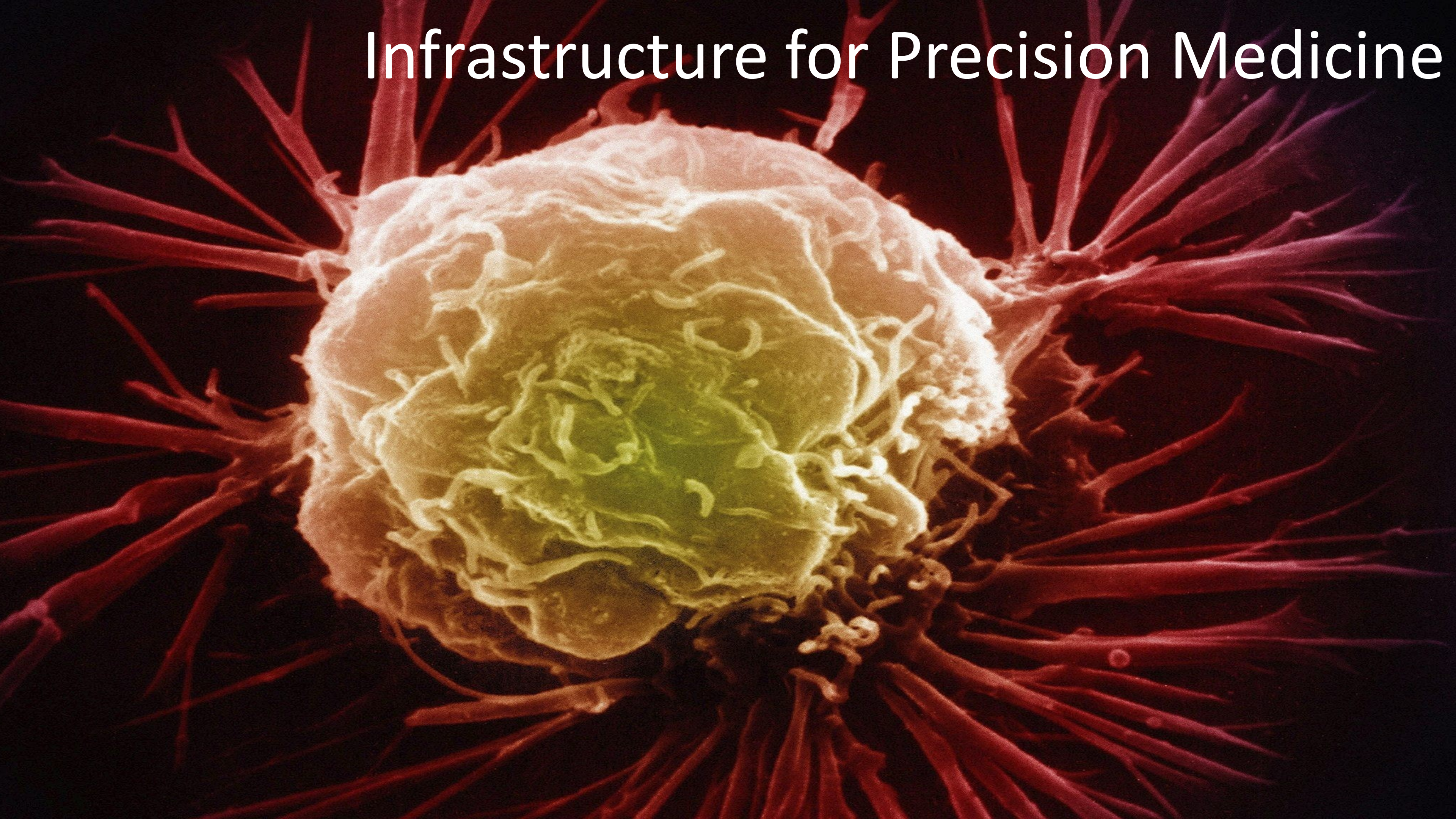
You can contact us at info@opencloudconsortium.org.

How do I get started?

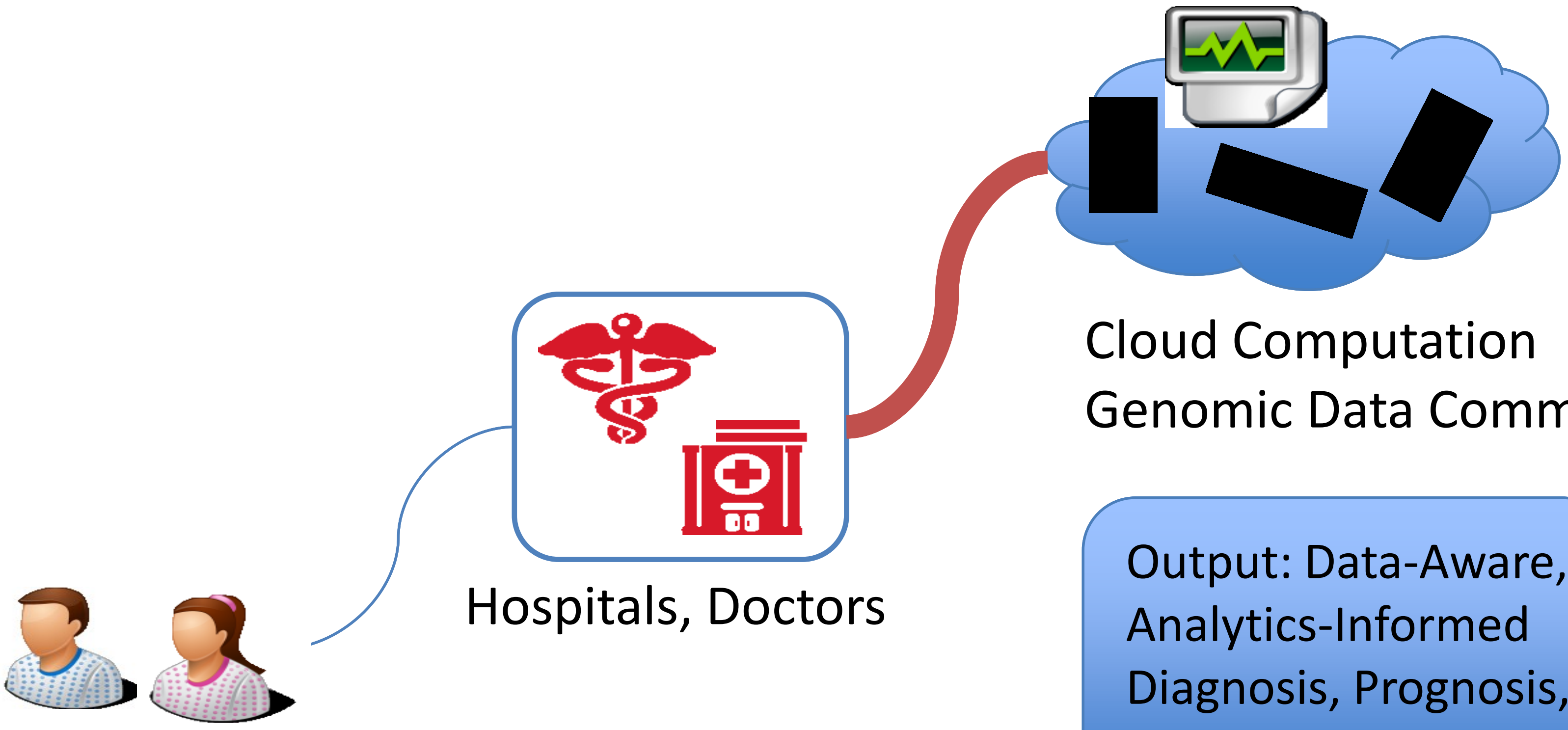
First, apply for an account. Once your account is approved, you can login to the console and get started. Support questions can be directed to support@opensciencedatacloud.org.

- Petabyte-scale, secure compliant biomedical cloud that interoperates with dbGaP controlled access data at NIH.

Infrastructure for Precision Medicine



Future Vision: A Nationwide Virtual Comprehensive Cancer Center



[redacted] [redacted] [redacted]

Cloud Computation
Genomic Data Commons

Hospitals, Doctors

Patients

Output: Data-Aware, Analytics-Informed Diagnosis, Prognosis, Optimal Treatment

Opportunity: Close Integration of Research Workflows and Foundation Networks

- Opportunity: Using GENI To Develop Innovative Techniques for Extremely Close Integration of Research WorkFlows and Dynamic Programmable Network Resources, Enabling Precision Networking
- Network Foundation Architecture: GENI + Innovative Customized Software Defined Networking Exchange (SDX)
- For This Demonstration: Specifically To Meet The Requirements of Bioinformatic Workflows

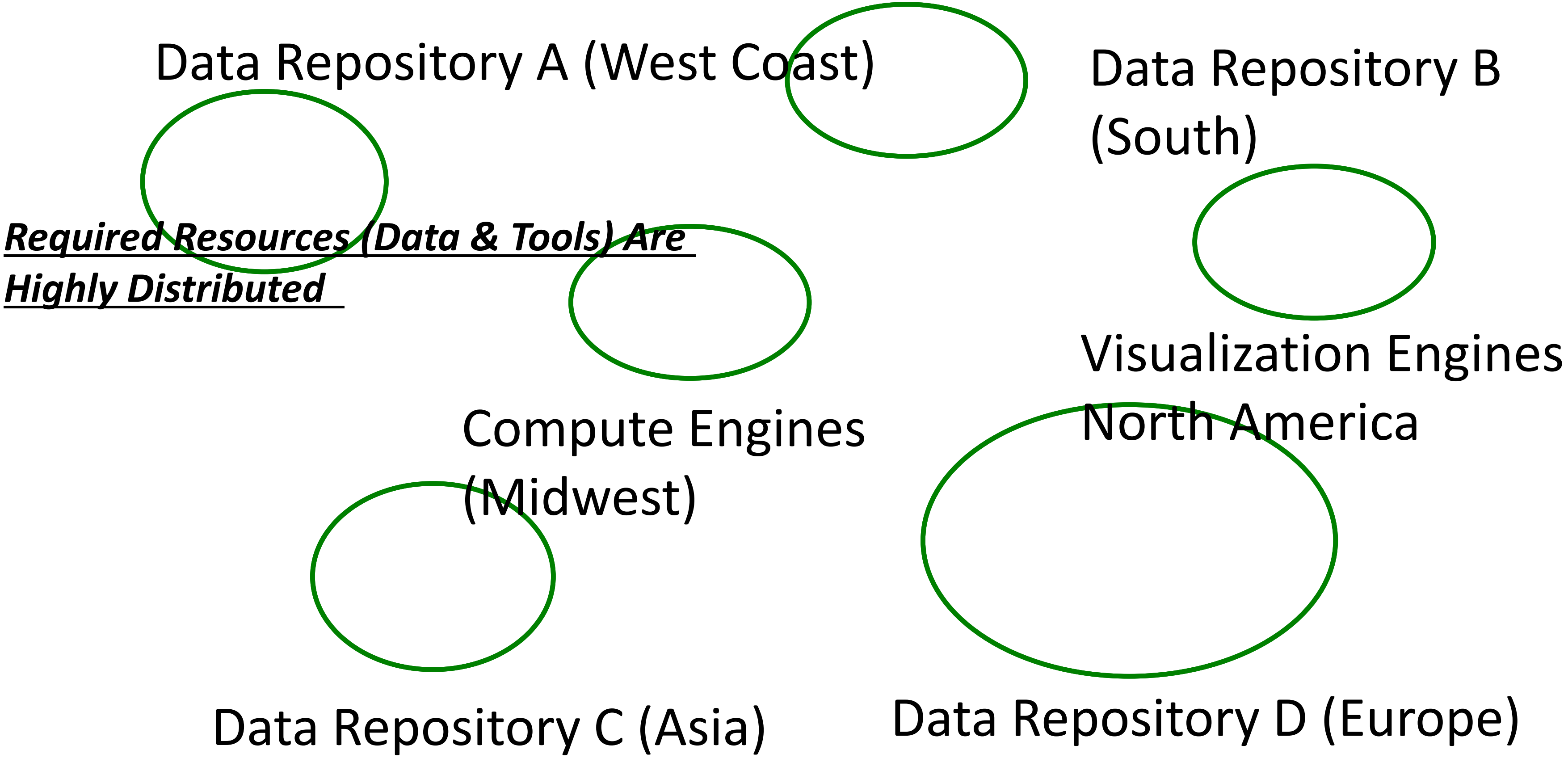
GENI Network Programmability Is Key

GENI Programmability

- GENI Provides A Platform for Building the Required Precision Communication Services, Networks and Exchanges (SDXs)
- GENI OpenFlow Network
  - National Overlay Infrastructure Comprised of Shared VLANs Interconnected With OpenFlow Switches
  - FOAM/FlowVisor Enabling Sliced OF Switches (e.g., via Subnet, VLAN, Tunnel, etc)
- Discoverable, Integratable, Configurable, Programmable, Virtual Devices: Click Routers, OVS Switches, Mobile Devices, Instrumentation, and Other Resources
- Dynamic Edge Process Topology Design and Implementation

Precision Networks for Precision Medicine

Biomedical Data Commons



Data Repository A (West Coast)

Data Repository B (South)

Required Resources (Data & Tools) Are Highly Distributed

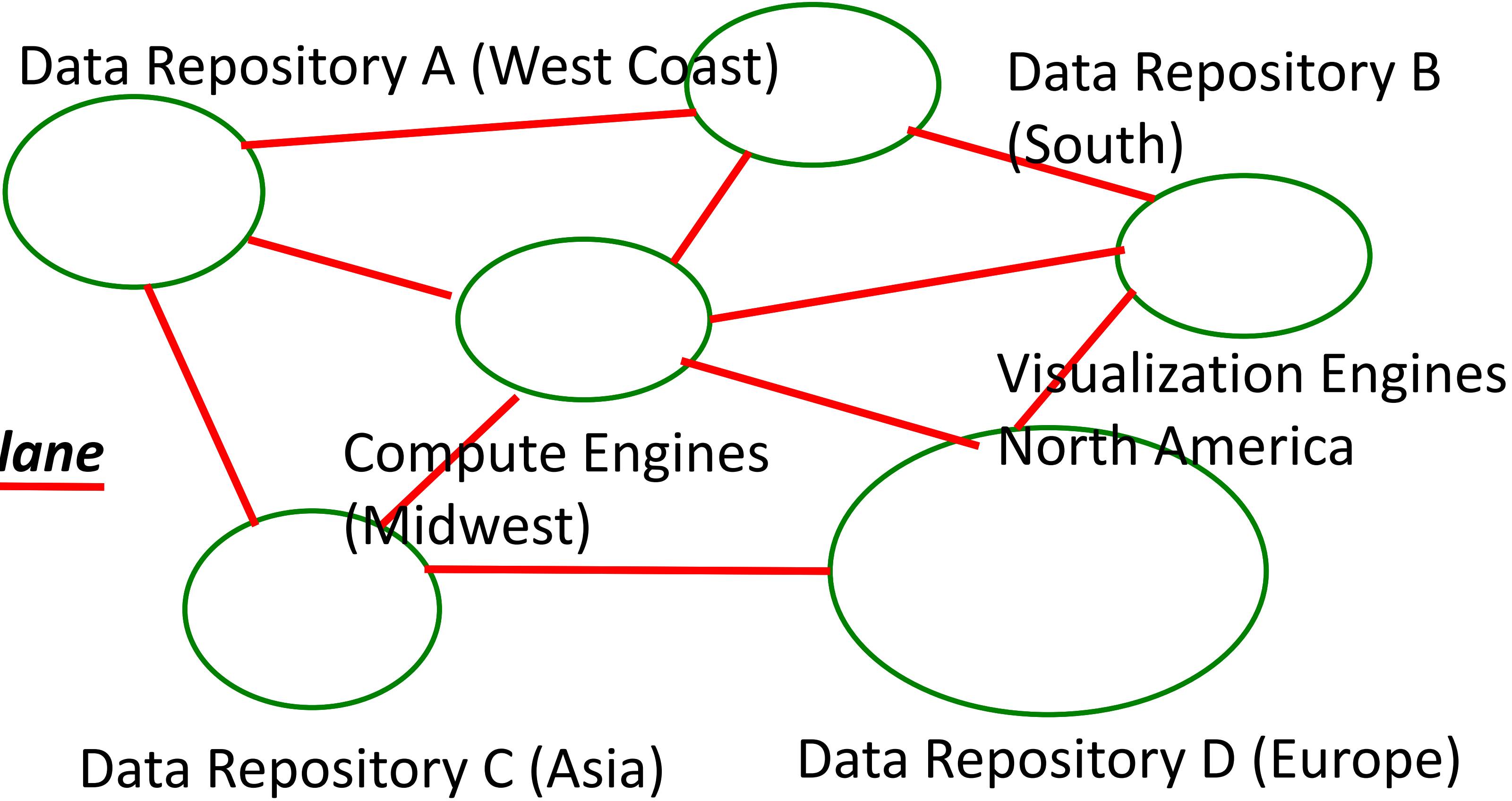
Visualization Engines North America

Compute Engines (Midwest)

Data Repository C (Asia)

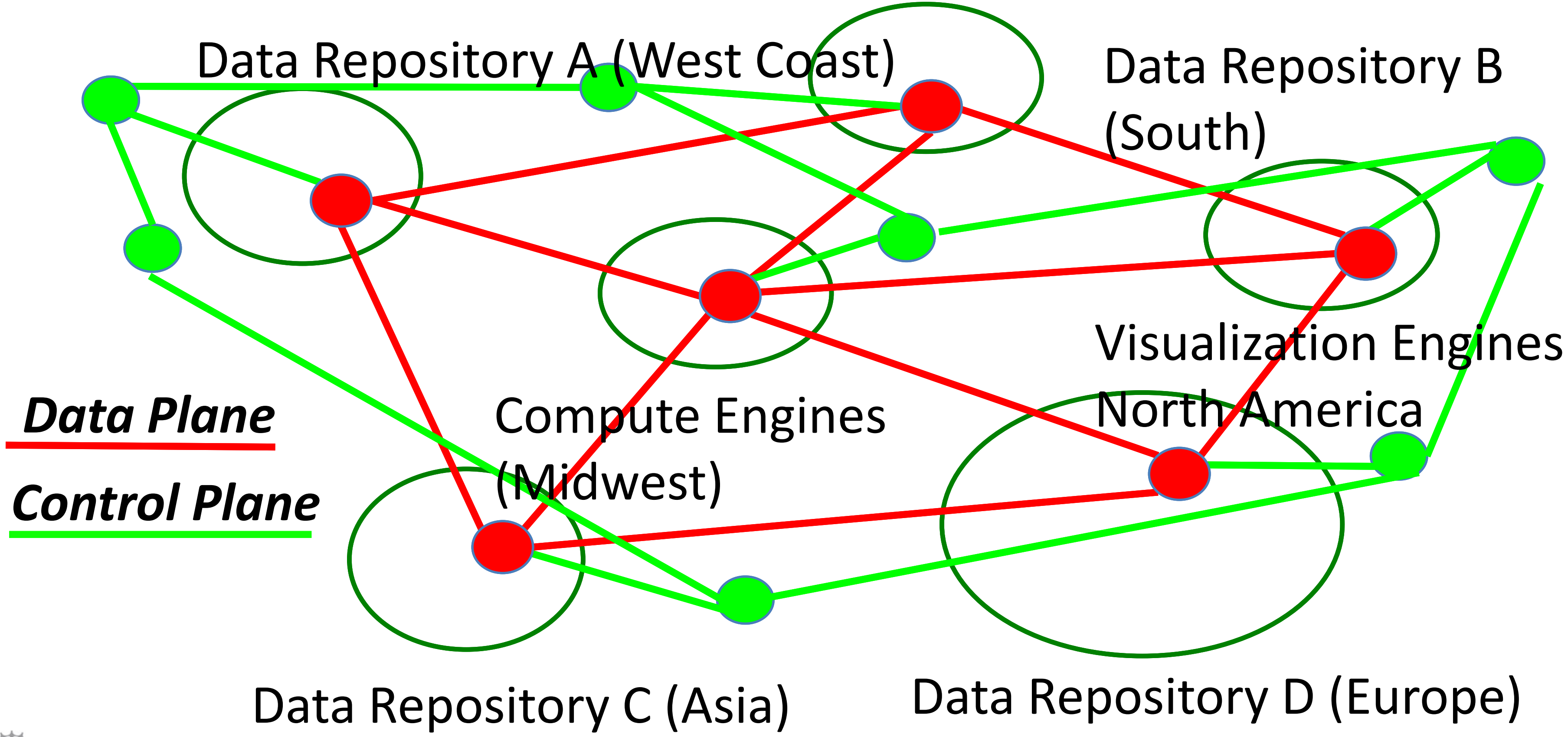
Data Repository D (Europe)

Biomedical Data Commons: Flow Orchestration: Data Plane


Data Repository A (West Coast)
Data Repository B (South)
Visualization Engines North America
Compute Engines (Midwest)
Data Repository C (Asia)
Data Repository D (Europe)

Data Plane

Biomedical Data Commons: Flow Orchestration: Control Plane + Data Plane



Data Repository A (West Coast)

Data Repository B (South)

Visualization Engines North America

Compute Engines (Midwest)

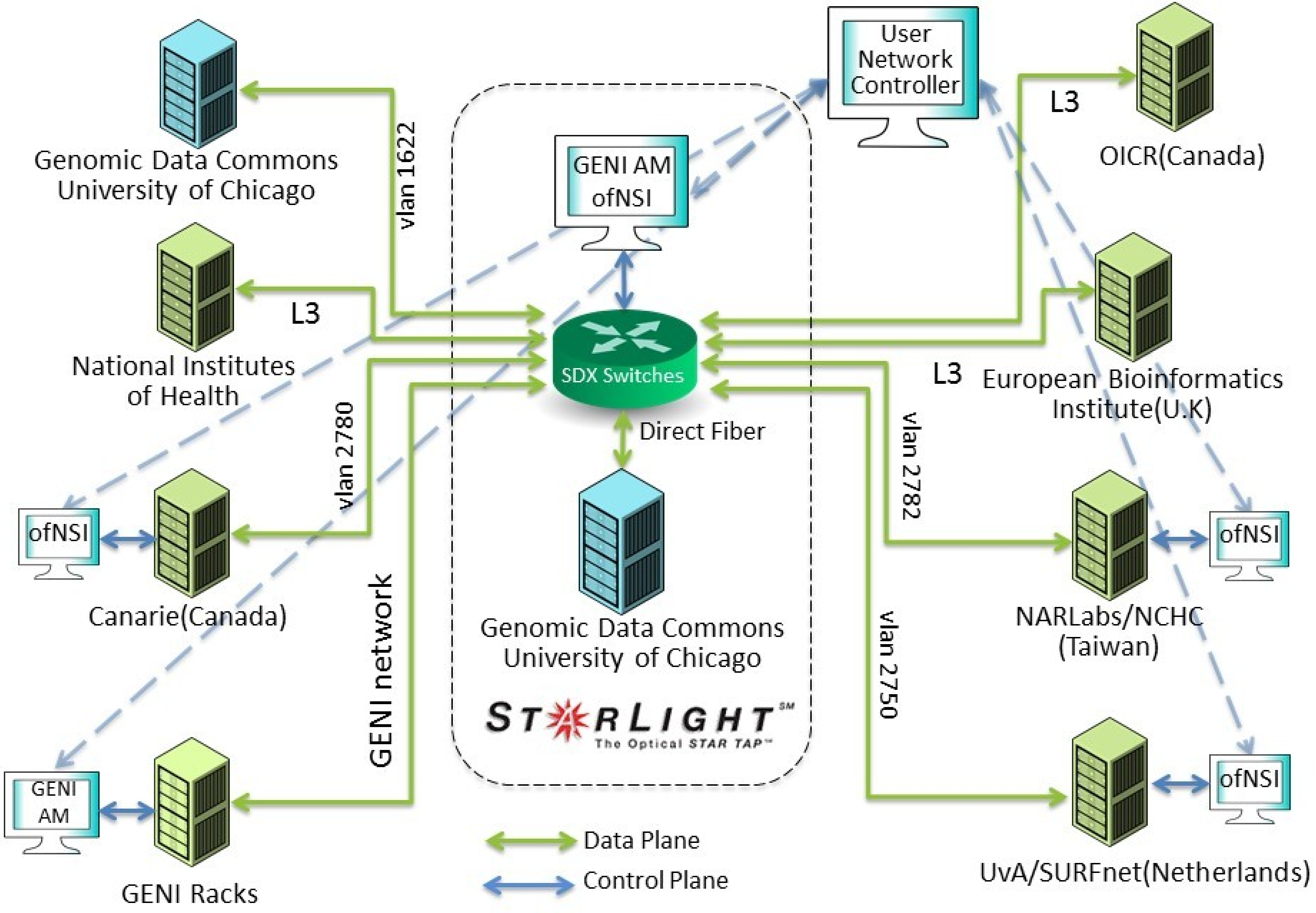
Data Repository C (Asia)

Data Repository D (Europe)

Data Plane 🟥

Control Plane 🟩

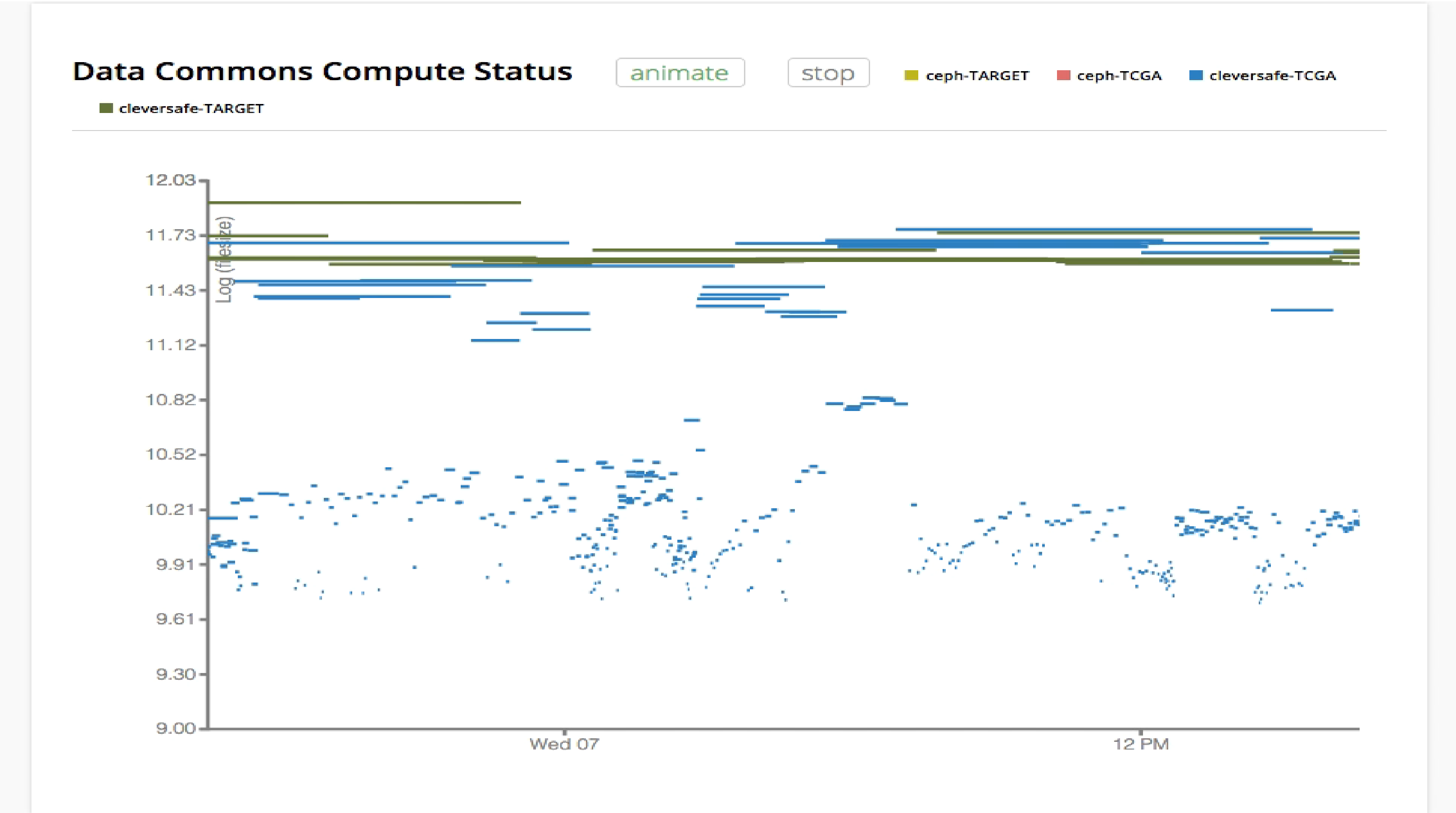
GEC22 Bioinformatics SDXs Demo Network



Today's Demonstration

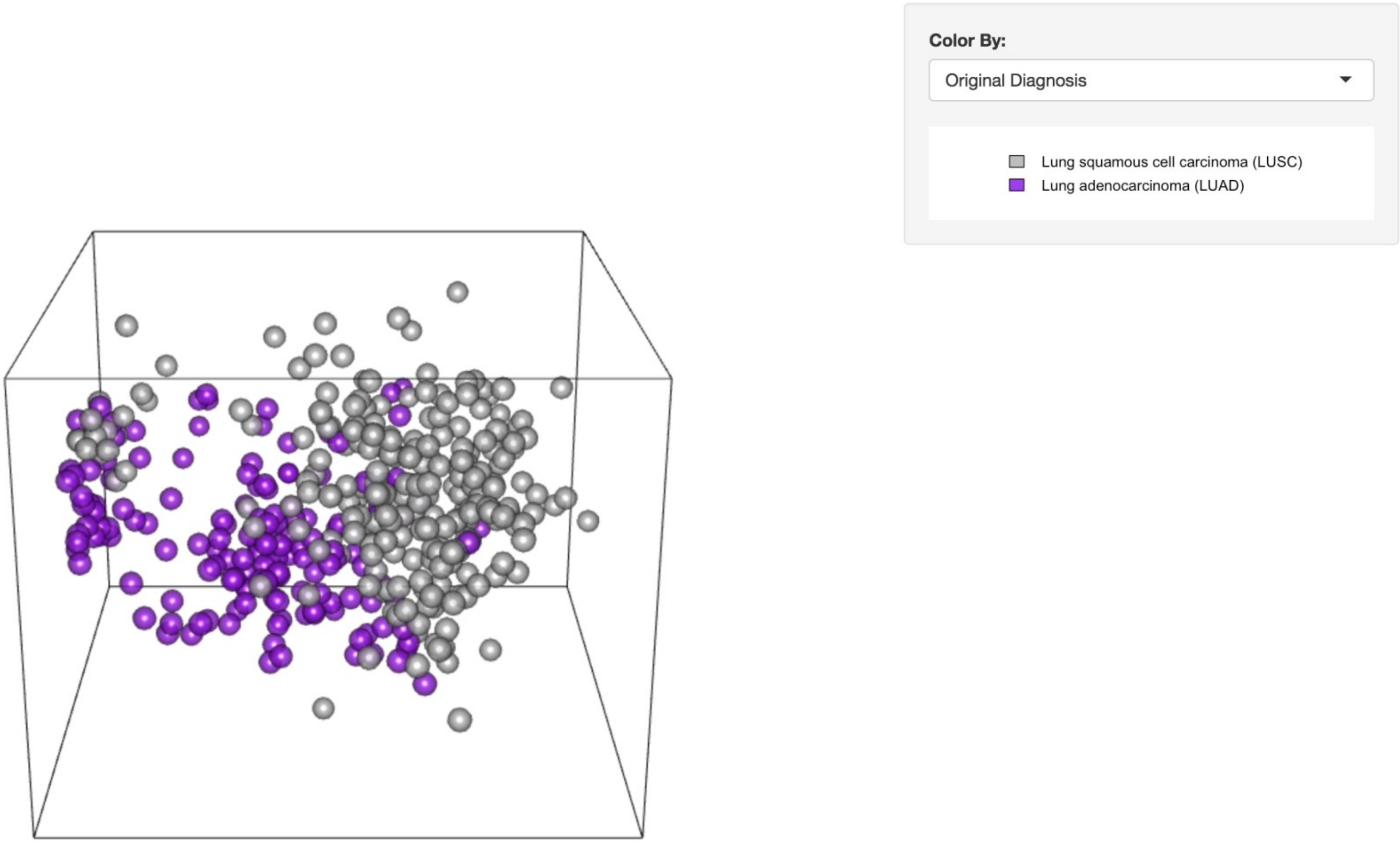
- A) Dynamically Moving Core Data Files Among Multiple Sites Around the World Via StarLight SDX
- B) Moving RNA-seq Data Files From NCI (Bethesda, MD) and EBI (Hinxton, UK) Through SDX Switch/Routers to The University of Chicago.
  - Analysis By Comparison To Known Data Correlated To Drug Response.
  - Determine Possible Actionable Therapeutic Options.
  - Return Viable Treatment Options To the Originating Site.

Genomic Data Commons Data Transfer



Data Commons Compute Status

Gene Expression Clustering of Lung Cancers



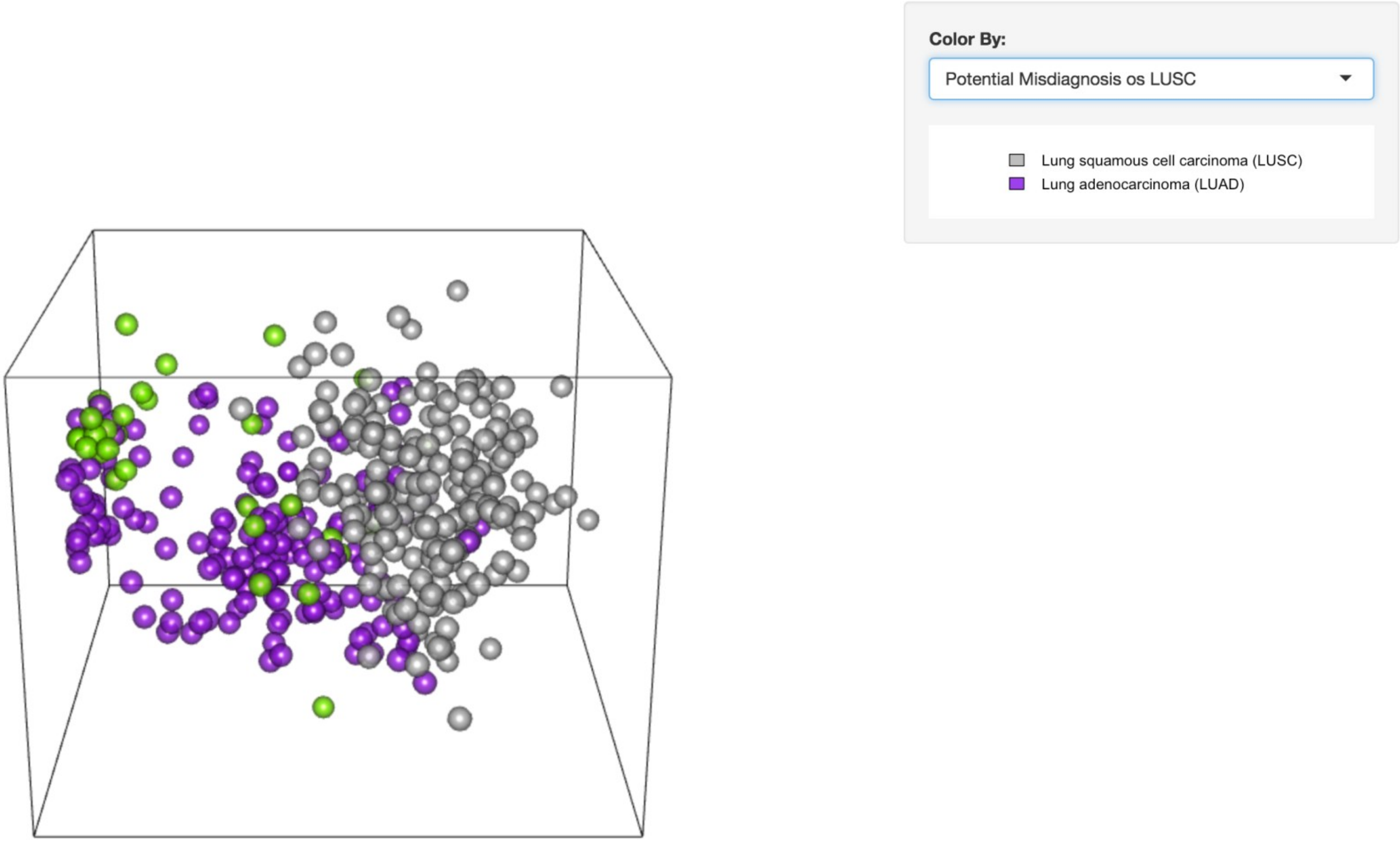
Color By:

Original Diagnosis ▾

⬜ Lung squamous cell carcinoma (LUSC)

🟪 Lung adenocarcinoma (LUAD)

Gene Expression Clustering of Lung Cancers



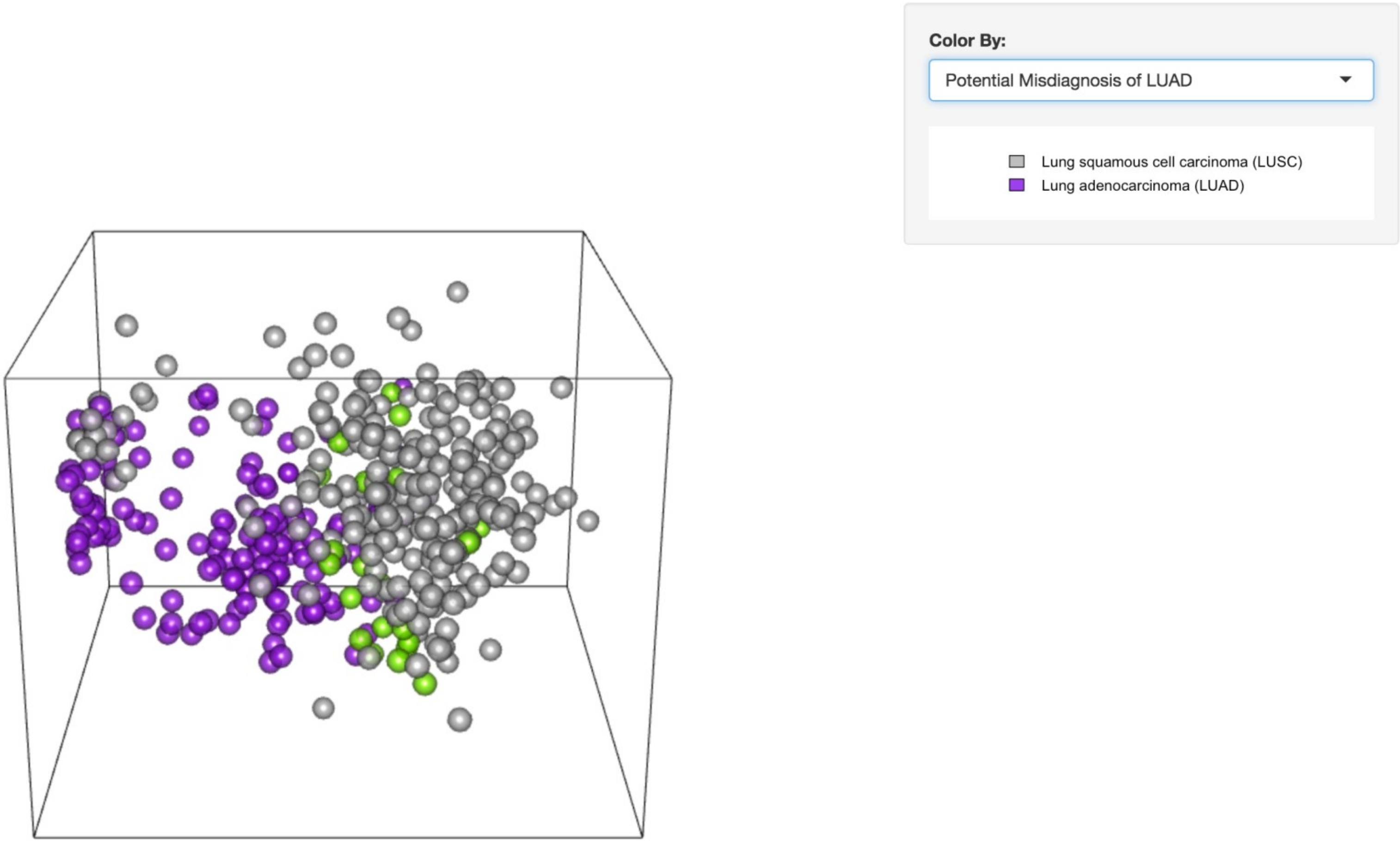
Color By:

Potential Misdiagnosis os LUSC ▾

▨ Lung squamous cell carcinoma (LUSC)

🟪 Lung adenocarcinoma (LUAD)

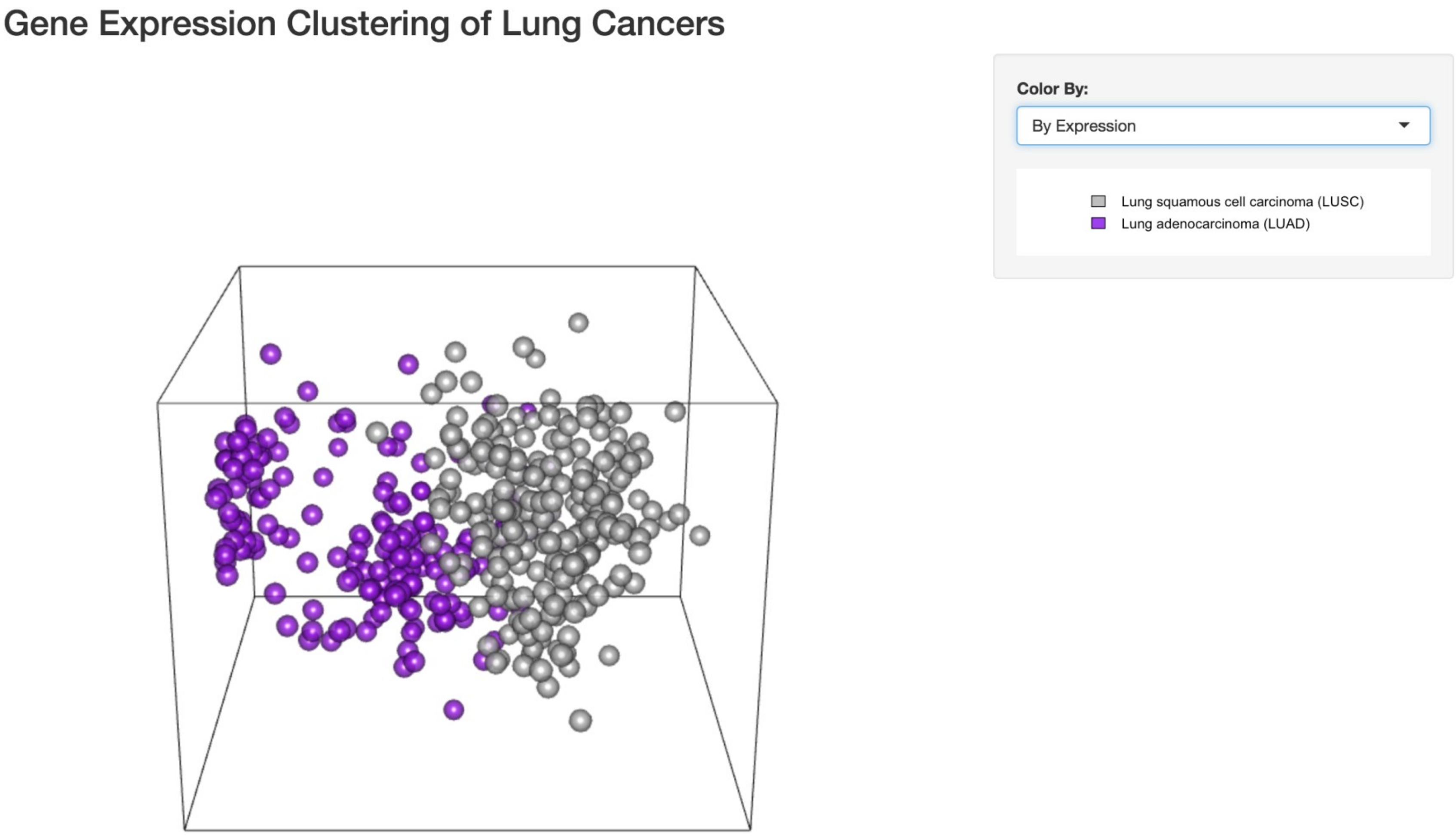
Gene Expression Clustering of Lung Cancers



Color By:

Potential Misdiagnosis of LUAD ▾

- ▨ Lung squamous cell carcinoma (LUSC)
- ■ Lung adenocarcinoma (LUAD)



Gene Expression Clustering of Lung Cancers

Color By: By Expression ▾

- ▨ Lung squamous cell carcinoma (LUSC)
- 🟪 Lung adenocarcinoma (LUAD)

Results

- Precision medicine requires data commons that scale to hundreds of petabytes scale, with programmable networks and data peering to support data sharing.
- Speed discovery and support analytics-driven healthcare to recommend treatment.
- Large Scale Data Analysis and Dynamic Pipelines For Workflows Are Essential For Determining Optimal Results.

Summary and Future

- What you saw: An innovative approach to advanced knowledge discovery and medical treatment***: Precision medicine being supported by precision networking***
- Why GENI/US Ignite is important: Precision mapping of communication services to BI workflow requirements across the world using advanced analytics, the Genomics Data Commons & a programmable dynamic SDX
- What happens looking forward, for the application and its integration with GENI:
  - A) Further development/refinement of basic capabilities
  - B) Transition to ***actual production services***
  - **C) The Genomics Data Commons and Bionimbus Protected Data Cloud is Being Developed As a Key Production Knowledge Discovery/Transformational Medical Treatment Facility**

Using GENI To Invent the Future…

Thank You!

GDC NCI Genomic Data Commons

NIH National Institutes of Health
Turning Discovery Into Health

The University of Chicago

Northwestern University

EBI

iCAIR
International Center for Advanced Internet Research

StarLight℠
The Optical STAR TAP™

OICR
Ontario Institute for Cancer Research

canarie

NARLabs
National Applied Research Laboratories
National Center for High-performance Computing

Universiteit van Amsterdam

SURF NET

MREN
Metropolitan Research & Education Network

usignite geni