NCI Cancer Research Data Commons

Allen Dearry, Ph.D.
Program Director
Center for Biomedical Informatics and Information Technology

NIH NATIONAL CANCER INSTITUTE

Big Data Interagency Working Group
June 28, 2018

Agenda

1. *National Cancer Data Ecosystem*
2. *NCI Cancer Research Data Commons*
3. *Data Linkages*
   - *Cancer Data Aggregator*
   - *Encrypted Unique Patient Identifier*
4. *Collaboration/coordination*
   - *Partnerships*
   - *Office of Data Sharing*

Agenda

1. *National Cancer Data Ecosystem*
2. *NCI Cancer Research Data Commons*
3. *Data Linkages*
   - *Cancer Data Aggregator*
   - *Encrypted Unique Patient Identifier*
4. *Collaboration/coordination*
   - *Partnerships*
   - *Office of Data Sharing*

The Beau Biden Cancer Moonshot^{sm}

Overarching goals – Jan, 2016

- Accelerate progress in cancer, including prevention & screening
    - From cutting edge basic research to wider uptake of standard of care
- Encourage greater cooperation and collaboration
    - Within and between academia, government, and private sector
- Enhance data sharing

Blue Ribbon Panel – October, 2016

- Network for Direct Patient Engagement
- Cancer Immunotherapy Translational Science Network
- Therapeutic Target Identification to Overcome Drug Resistance
- A National Cancer Data Ecosystem for Sharing and Analysis
- Fusion Oncoproteins in Childhood Cancers
- Symptom Management Research
- Prevention and Early Detection – Implementation of Evidence-based Approaches
- Retrospective Analysis of Biospecimens from Patients Treated with Standard of Care
- Generation of 3D Human Tumor Atlas
- Development of New Enabling Cancer Technologies
- Full report: www.cancer.gov/brp

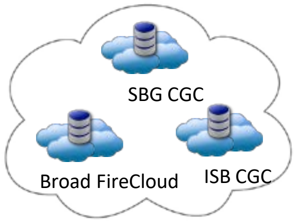
National Cancer Data Ecosystem Recommendations

Overall goal: *"Enable all participants across the cancer research and care continuum to contribute, access, combine and analyze diverse data that will enable new discoveries and lead to lowering the burden of cancer."*

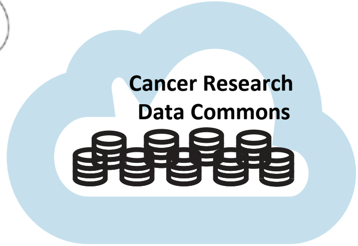
Recommendations

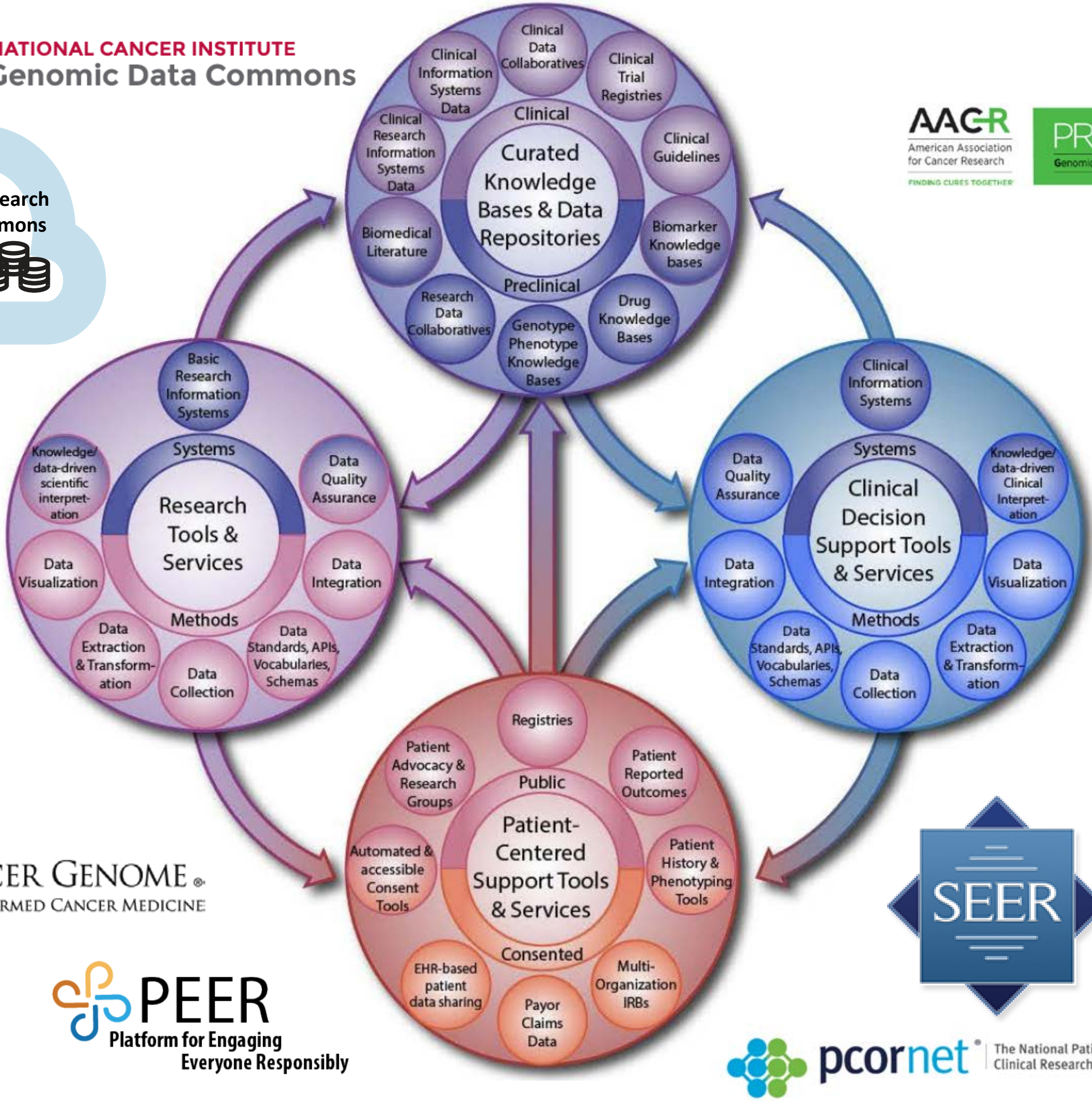
- Build a National Cancer Data Ecosystem
  - Enhanced cloud-computing platforms.
  - Essential underlying data science infrastructure and portals for the Cancer Data Ecosystem.
  - Services that link disparate information, including clinical, image, and molecular data.
  - Develop standards and tools so that data are interoperable.
  - Address sustainability and data governance to ensure long-term health of the Ecosystem.
- The National Cancer Data Ecosystem is broader than NCI
  - An NCI Cancer Research Data Commons is envisioned as part of the National Cancer Data Ecosystem

Enhanced Data Sharing Working Group Recommendation:
The Cancer Data Ecosystem



NIH NATIONAL CANCER INSTITUTE
Genomic Data Commons





Curated Knowledge Bases & Data Repositories

Clinical: Clinical Data Collaboratives; Clinical Trial Registries; Clinical Guidelines; Biomarker Knowledge bases; Clinical Information Systems Data

Preclinical: Drug Knowledge Bases; Genotype Phenotype Knowledge Bases; Research Data Collaboratives; Biomedical Literature; Clinical Research Information Systems Data

Research Tools & Services

Systems: Basic Research Information Systems; Data Quality Assurance; Knowledge/ data-driven scientific interpret-ation

Methods: Data Visualization; Data Integration; Data Extraction & Transform-ation; Data Collection; Data Standards, APIs, Vocabularies, Schemas

Clinical Decision Support Tools & Services

Systems: Clinical Information Systems; Data Quality Assurance; Knowledge/ data-driven Clinical Interpret-ation

Methods: Data Integration; Data Visualization; Data Standards, APIs, Vocabularies, Schemas; Data Collection; Data Extraction & Transform-ation

Patient-Centered Support Tools & Services

Public: Registries; Patient Reported Outcomes; Patient History & Phenotyping Tools; Patient Advocacy & Research Groups

Consented: Automated & accessible Consent Tools; EHR-based patient data sharing; Payor Claims Data; Multi-Organization IRBs

AACR American Association for Cancer Research FINDING CURES TOGETHER

PROJECT GENIE Genomics Evidence Neoplasia Information Exchange

ORIEN ONCOLOGY RESEARCH INFORMATION EXCHANGE NETWORK

NIH NATIONAL CANCER INSTITUTE
Informatics Technology for Cancer Research

monarch INITIATIVE

ASCO CANCER•LINQ Learning Intelligence Network for Quality

MY CANCER GENOME GENETICALLY INFORMED CANCER MEDICINE

SEER

Sage BIONETWORKS

PEER Platform for Engaging Everyone Responsibly

pcornet The National Patient-Centered Clinical Research Network

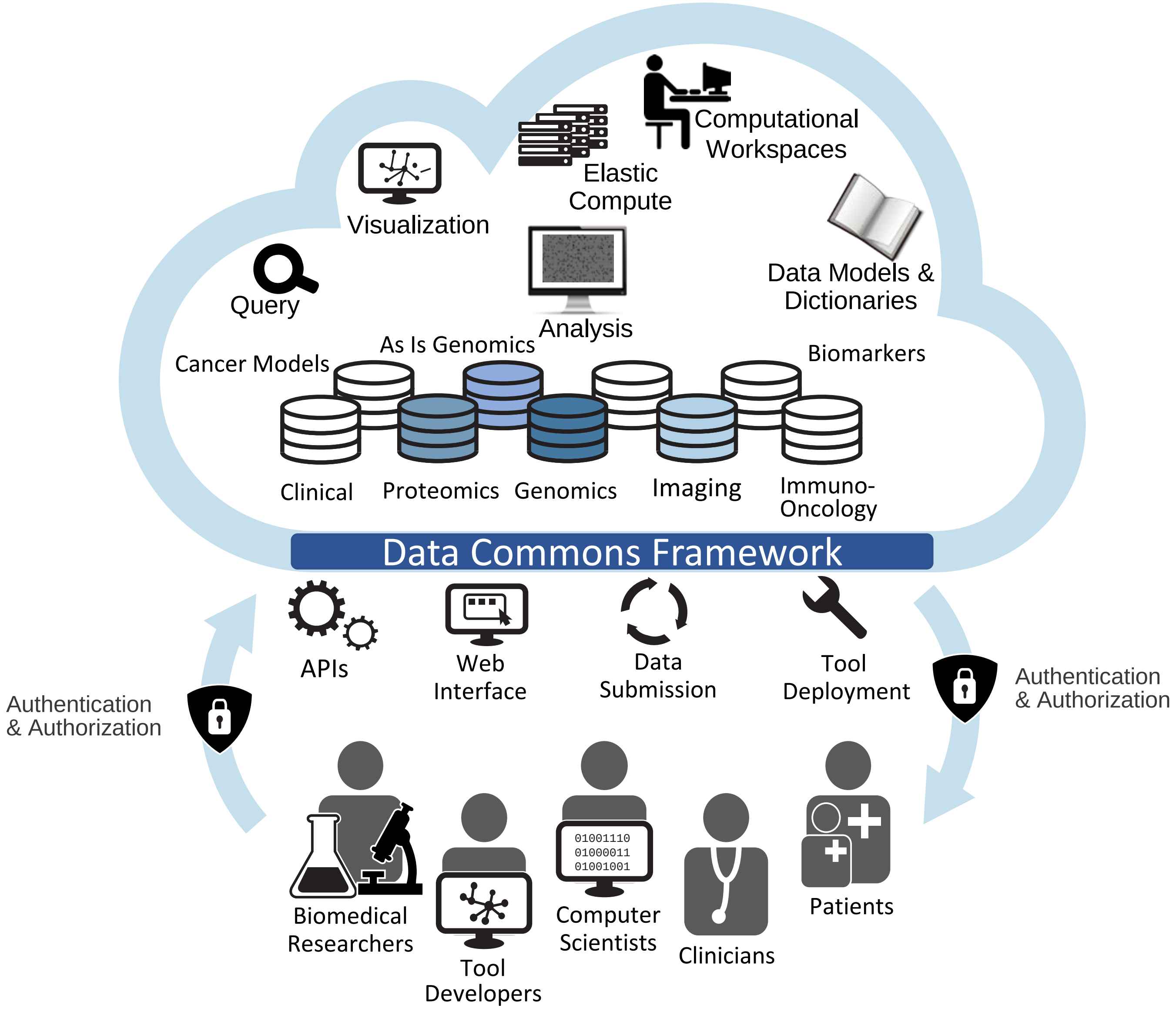
Agenda

1. *National Cancer Data Ecosystem*
2. *NCI Cancer Research Data Commons*
3. *Data Linkages*
   - *Cancer Data Aggregator*
   - *Encrypted Unique Patient Identifier*
4. *Collaboration/coordination*
   - *Partnerships*
   - *Office of Data Sharing*

NCI Cancer Research Data Commons (CRDC) - Concept

NCI Scope: *“Create a data science infrastructure necessary to connect repositories, analytical tools, and knowledge bases”*

Data commons co-locate data, storage and computing infrastructure with commonly used services, tools & apps for analyzing and sharing data to create an interoperable resource for the research community.*



Data Contributors and Consumers

*Robert L. Grossman, Allison Heath, Mark Murphy, Maria Patterson and Walt Wells, A Case for Data Commons Towards Data Science as a Service, IEEE Computing in Science and Engineer, 2016.   Source of image: The CDIS, GDC, & OCC data commons infrastructure at the University of Chicago Kenwood Data Center.

Goals of the NCI CRDC

- Enable the cancer research community to share diverse data types across programs and institutions.
- Provide easy access to data, regardless of where it is stored.
- Provide mechanisms for innovative tool discovery, access, usage.
- Help NCI Data Coordinating Centers sustain and share their data publicly.
- Develop a set of reusable components - a framework - for the community to use to build interoperable data commons.

CRDC Data Sources / Contributors (Examples)

The Cancer Genome Atlas (TCGA)

Therapeutically Applicable Research to Generate Effective Treatments (TARGET)

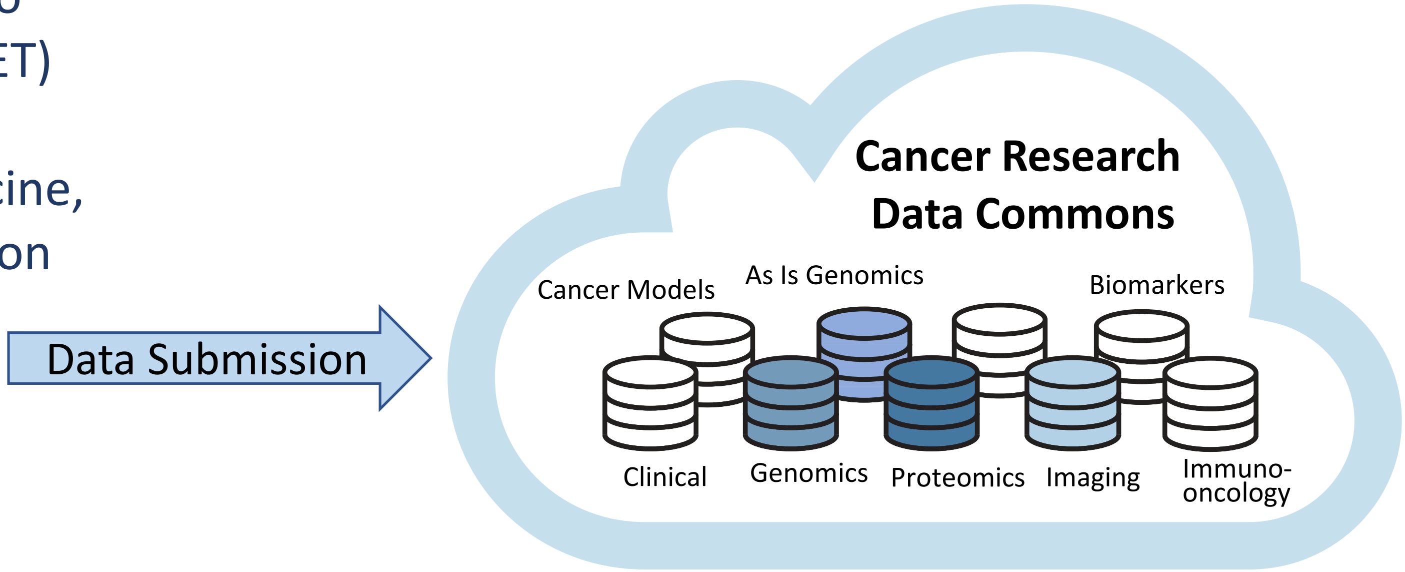
3rd Party Programs: Foundation Medicine, Multiple Myeloma Research Foundation

Clinical Proteomic Tumor Analysis Consortium (CPTAC)

The Cancer Imaging Archive (TCIA)

NCI Individual Labs / Grants / Contracts / Cancer Centers (GENIE)

Collaborative Programs: APOLLO (Applied Proteogenomic OrganizationaL Learning and Outcomes), ICPC (International Cancer Proteogenome Consortium)



Data Commons Framework – What Is It?

Reusable, expandable framework for a Data Commons

Core principles and structures for a Data Commons

Set of modular components that can be leveraged across Data Commons

Modular Components

- Secure user authentication and authorization
- Metadata validation and tools
- Domain-specific, extensible data models and dictionaries
- API and container environment for tools and pipelines
- Access to computational workspaces for storing data, tools, and results

NCI is developing the Framework and will use it to stand up several example Data Commons the community can leverage or use as a model to build their own commons.

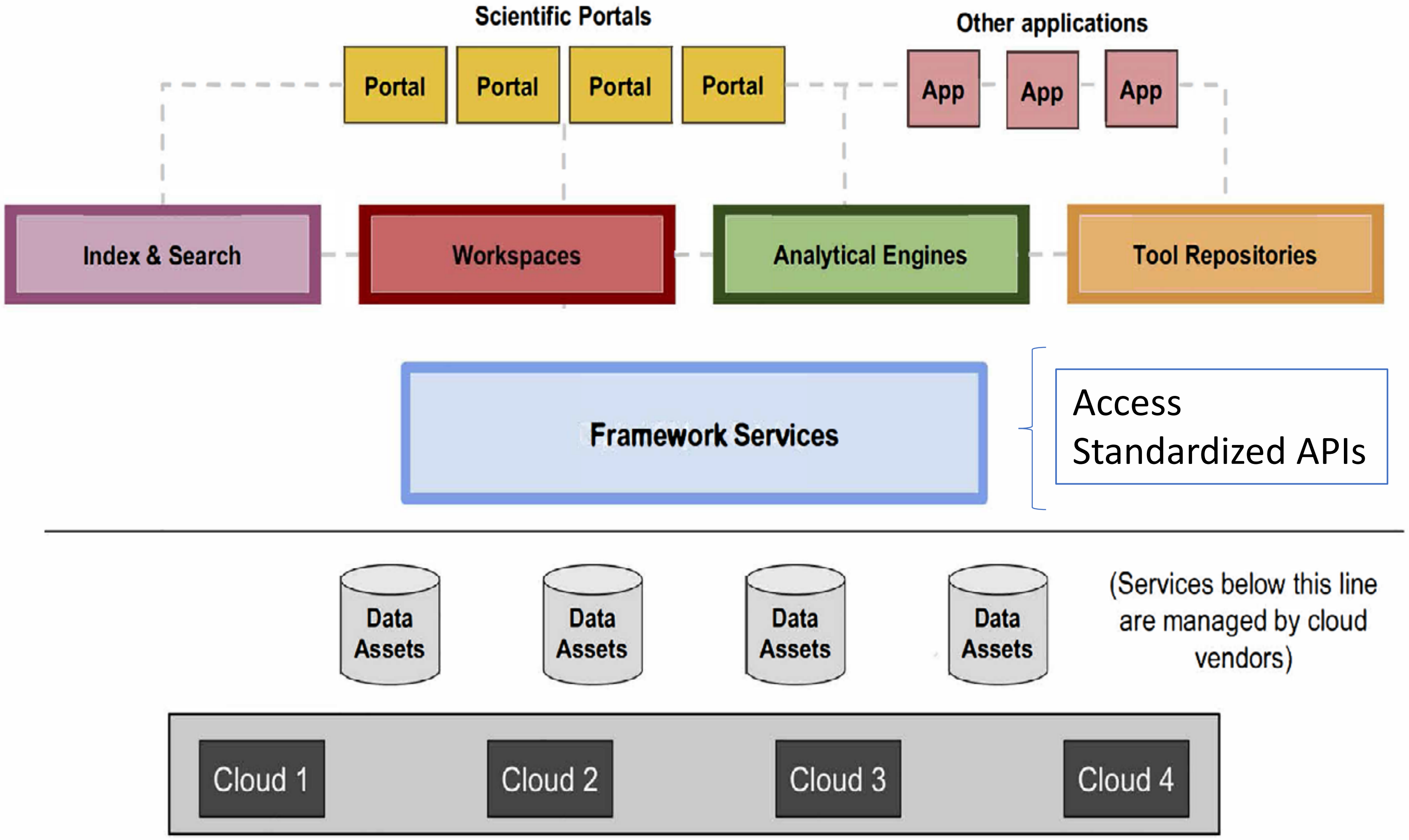
A Commons Alliance and Data Biosphere

NIH NATIONAL CANCER INSTITUTE
Genomic Data Commons

NCI CRDC Framework Services

All of Us RESEARCH PROGRAM

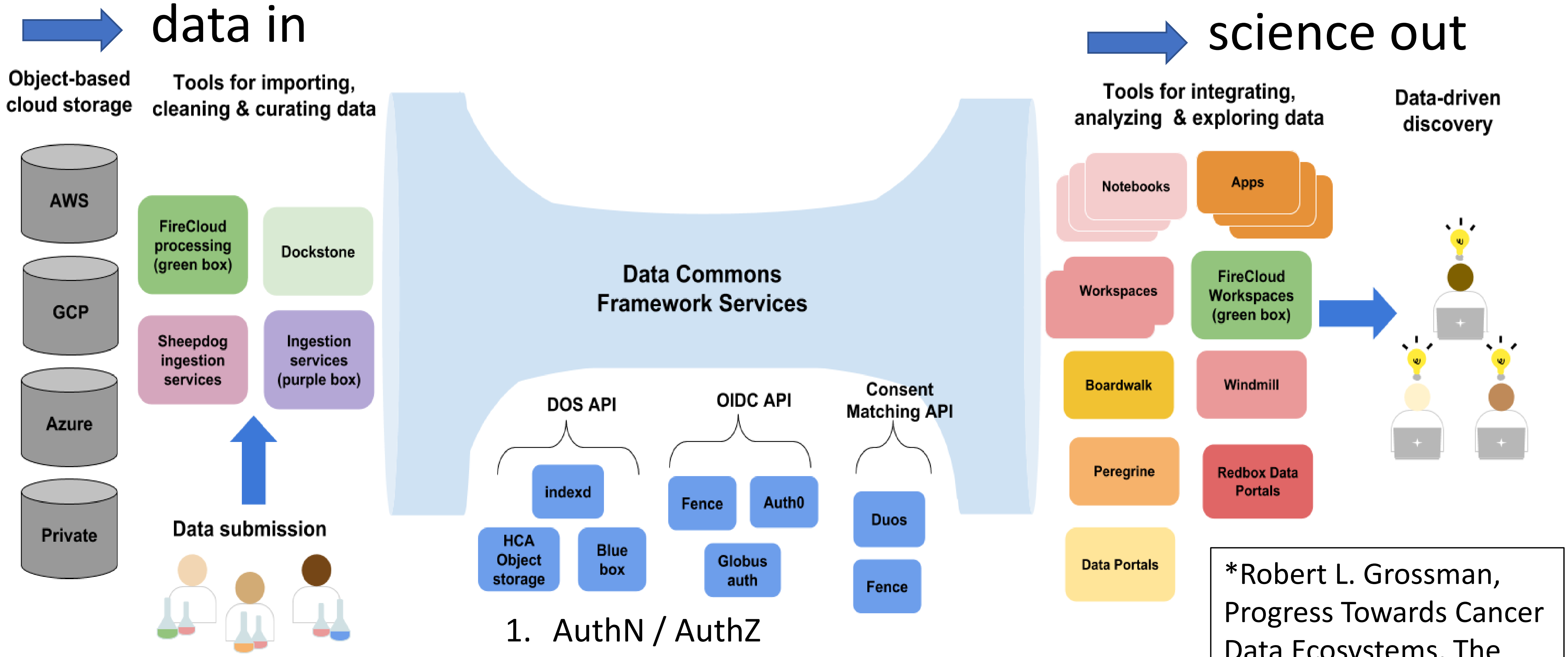
HUMAN CELL ATLAS



1. NCI GDC / NCRDC (UChicago)
2. NIH All of Us (Broad/Verily)
3. CZI HCA Data Platform (UCSC/Broad)

For more information, see: Josh Denny, David Glazer, Robert L. Grossman, Benedict Paten & Anthony Philippakis, A Data Biosphere for Biomedical Research, https://medium.com/@benedictpaten/a-data-biosphere-for-biomedical-research-d212bbfae95d. Also available at: https://goo.gl/9CySeo

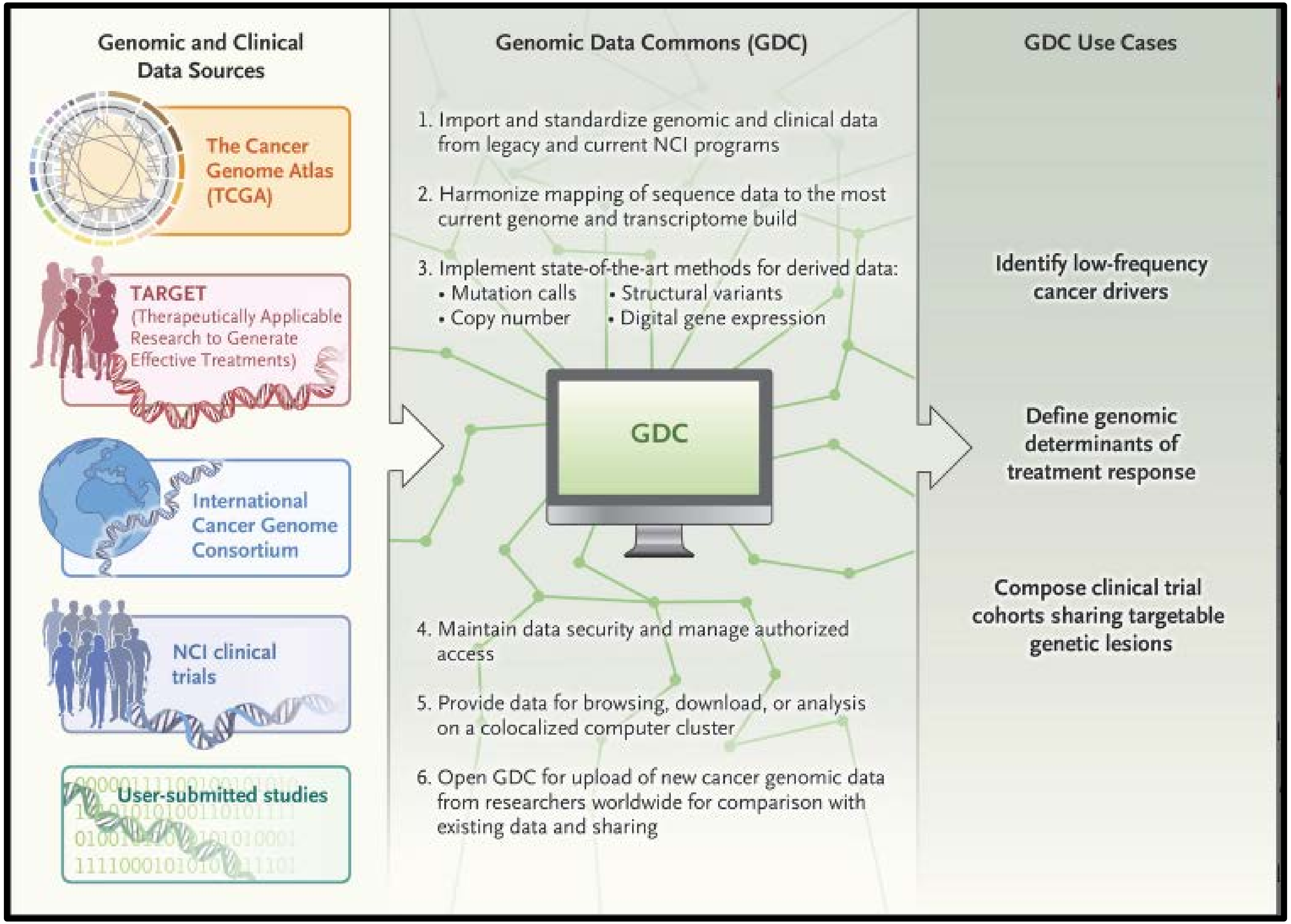
Narrow Middle Architecture (End-to-End Design)


data in
science out
Object-based cloud storage
AWS
GCP
Azure
Private
Tools for importing, cleaning & curating data
FireCloud processing (green box)
Dockstone
Sheepdog ingestion services
Ingestion services (purple box)
Data submission
Data Commons Framework Services
DOS API
indexd
HCA Object storage
Blue box
OIDC API
Fence
Auth0
Globus auth
Consent Matching API
Duos
Fence
Tools for integrating, analyzing & exploring data
Notebooks
Apps
Workspaces
FireCloud Workspaces (green box)
Boardwalk
Windmill
Peregrine
Redbox Data Portals
Data Portals
Data-driven discovery

1. AuthN / AuthZ
2. Metadata validation
3. Extensible data model
4. APIs for containers, workflows & tools
5. Workspaces

*Robert L. Grossman, Progress Towards Cancer Data Ecosystems, The Cancer Journal: The Journal of Principles & Practice of Oncology, 2018, to appear.

Building on Foundation of the NCI Genomic Data Commons



Genomic and Clinical Data Sources

The Cancer Genome Atlas (TCGA)

TARGET (Therapeutically Applicable Research to Generate Effective Treatments)

International Cancer Genome Consortium

NCI clinical trials

User-submitted studies

Genomic Data Commons (GDC)

1. Import and standardize genomic and clinical data from legacy and current NCI programs
2. Harmonize mapping of sequence data to the most current genome and transcriptome build
3. Implement state-of-the-art methods for derived data:
   - Mutation calls
   - Structural variants
   - Copy number
   - Digital gene expression

GDC

4. Maintain data security and manage authorized access
5. Provide data for browsing, download, or analysis on a colocalized computer cluster
6. Open GDC for upload of new cancer genomic data from researchers worldwide for comparison with existing data and sharing

GDC Use Cases

Identify low-frequency cancer drivers

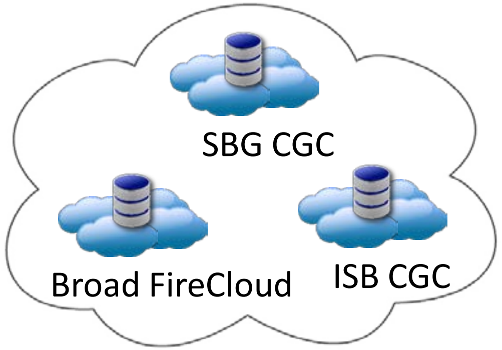
Define genomic determinants of treatment response

Compose clinical trial cohorts sharing targetable genetic lesions

https://gdc.cancer.gov/

From Grossman et al.
N Engl J Med 2016; 375:1109-1112September 22, 2016DOI: 10.1056/NEJMp1607591

NCI Cloud Resources



Cloud Resources provide:

- Access to large genomic data sets without need to download
- Ability for researchers to bring their own tools and pipelines to the data
- Ability for researchers to bring their own data and analyze in combination with existing genomic data
- Workspaces, for researchers to save and share their data and results of analyses

Data

- Access and analyze 11,000 TCGA samples without having to download data
- Upload your own data for analysis

Compute

- Perform large scale analysis using the elastic compute power of commercial cloud platforms

Security

- dbGaP-authorized users can access controlled TCGA data
- Systems meet strict Federal security guidelines

Democratize access to NCI-generated genomic and related data, and to create a cost-effective way to provide scalable computational capacity to the cancer research community.

#NCICloud

NCI Cancer Research Data Commons

NIH NATIONAL CANCER INSTITUTE Genomic Data Commons

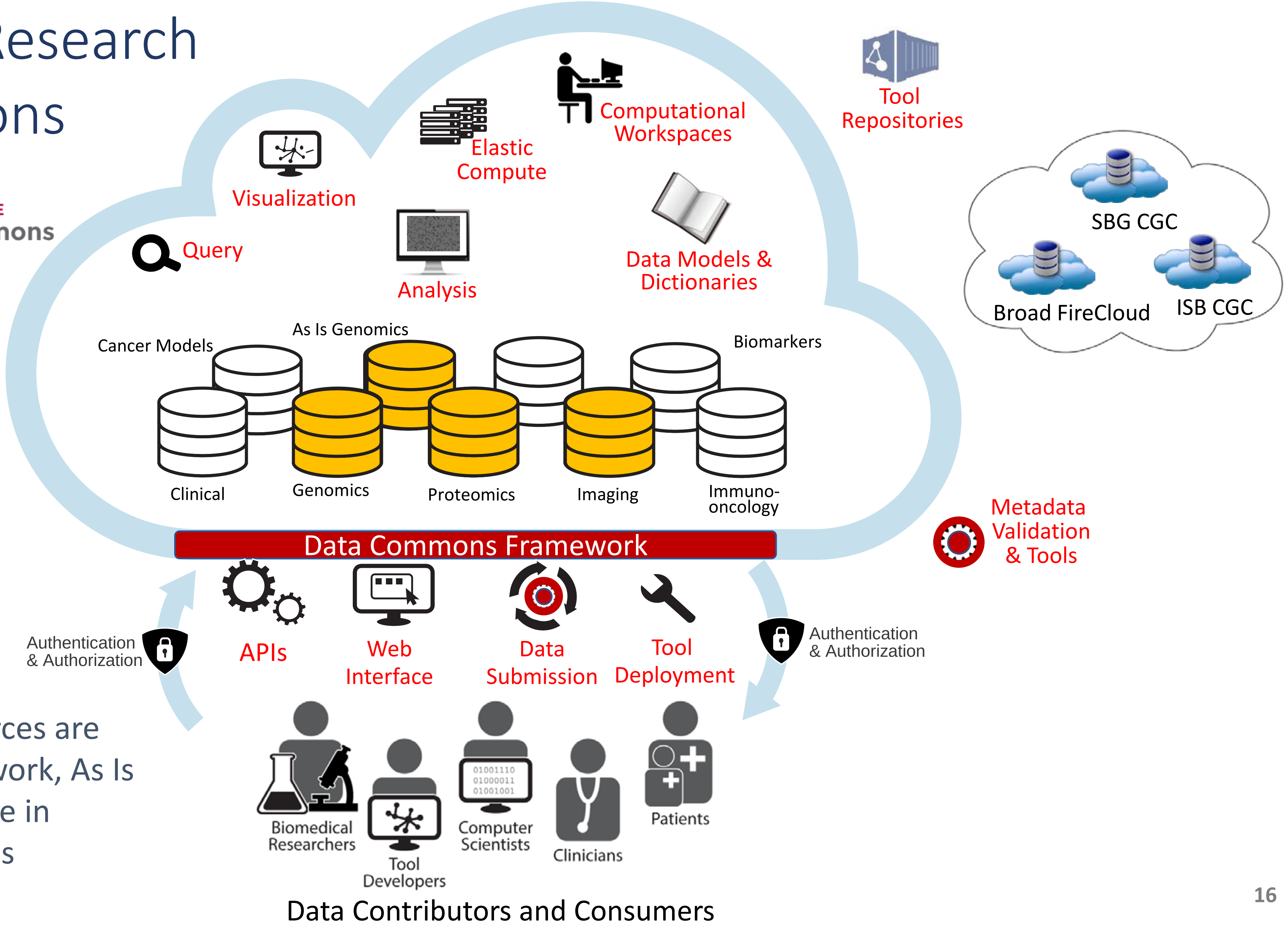
dbGaP GENOTYPES and PHENOTYPES

CPTAC

*Clinical Proteomic Tumor Analysis Consortium**

TCIA

*The Cancer Imaging Archive**



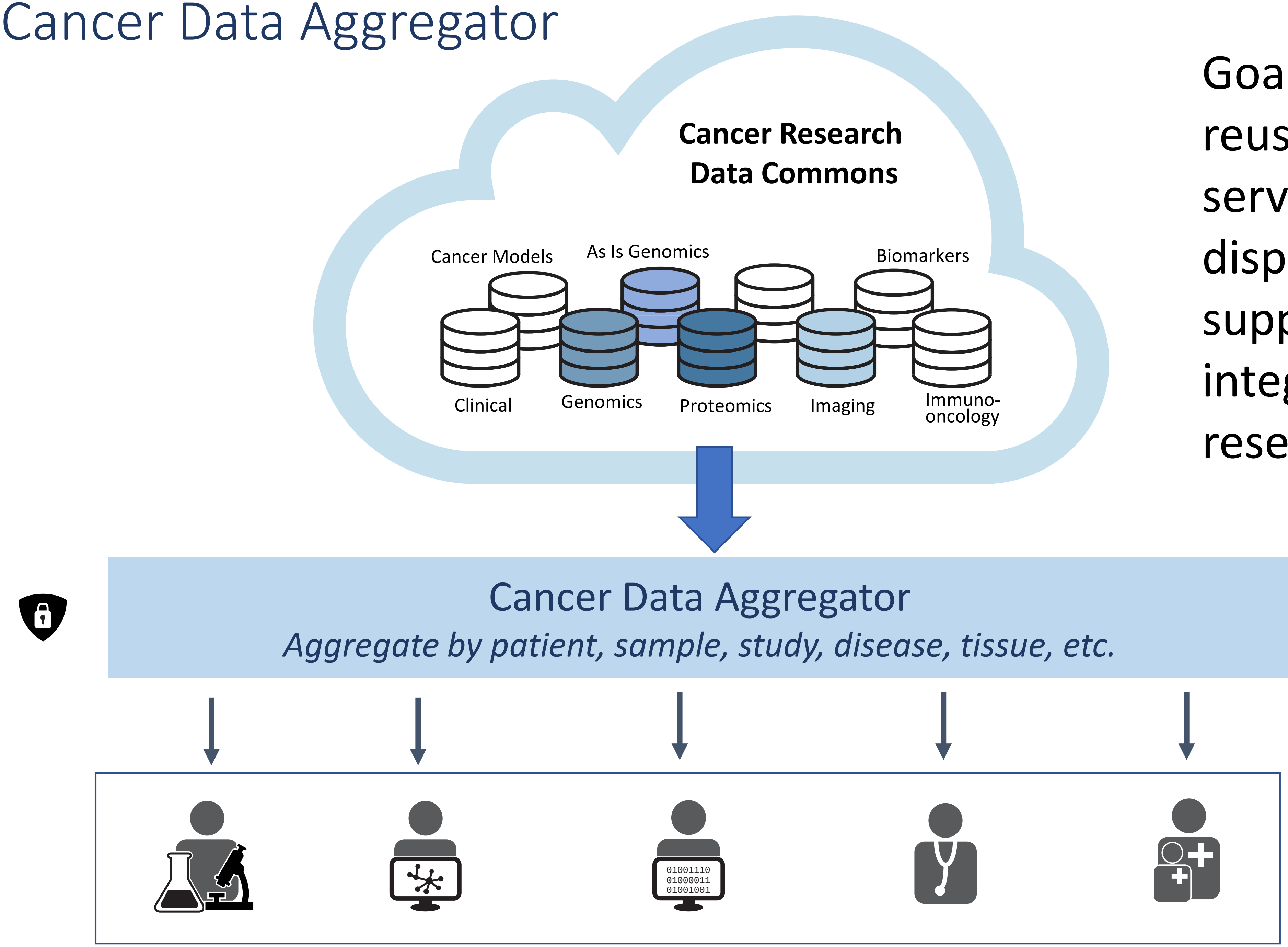
GDC and Cloud Resources are available now; Framework, As Is Genomics, PDC, IDC are in development; all else is notional.

Agenda

1. *National Cancer Data Ecosystem*
2. *NCI Cancer Research Data Commons*
3. *Data Linkages*
   - *Cancer Data Aggregator*
   - *Encrypted Unique Patient Identifier*
4. *Collaboration/coordination*
   - *Partnerships*
   - *Office of Data Sharing*

Cancer Data Aggregator

Goal: Provide a reusable informatics service to connect disparate data in support of integrative cancer research



Cancer Research Data Commons

Cancer Models

As Is Genomics

Biomarkers

Clinical

Genomics

Proteomics

Imaging

Immuno-oncology

Cancer Data Aggregator

Aggregate by patient, sample, study, disease, tissue, etc.

Multi-modal data aggregation

Data Exploration

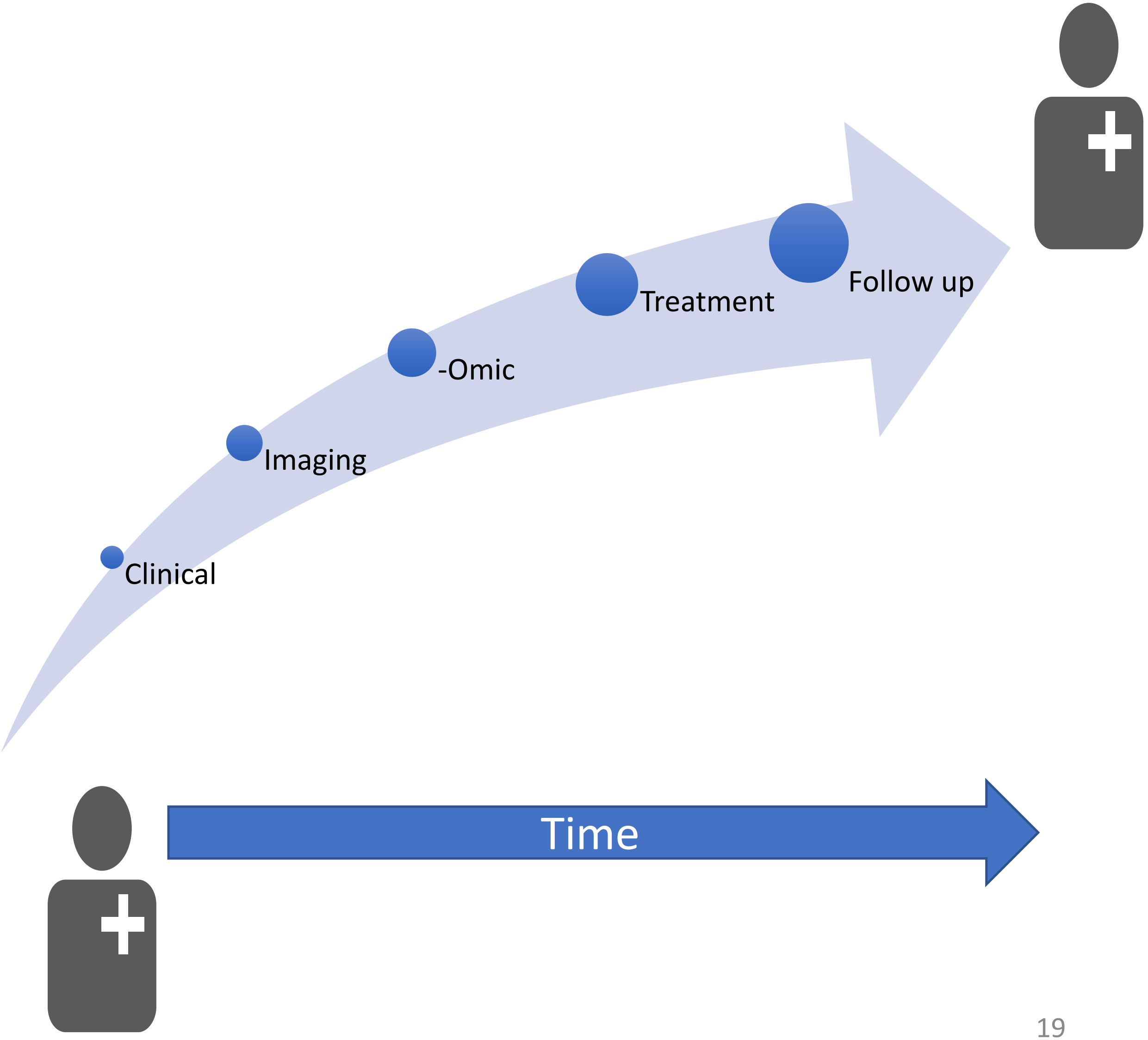
Elastic Compute

Query

Analyze

Development of an Encrypted Unique Patient Identifier

- Pressing need to connect patient-level data across multiple data sources, data types and research studies—over time.
- Challenges include:
  - Protecting patient confidentiality
  - Consistency of identifying data (personally identifiable information, PII) available across diverse sources
  - Accuracy of linkage with varying PII
  - Scalability
- Encrypted hashed token
  - Allows linkage of diverse data.
  - Permits data sharing across multiple sources without release of PII.



Agenda

1. *National Cancer Data Ecosystem*
2. *NCI Cancer Research Data Commons*
3. *Data Linkages*
   - *Cancer Data Aggregator*
   - *Encrypted Unique Patient Identifier*
4. *Collaboration/coordination*
   - *Partnerships*
   - *Office of Data Sharing*

Creating Partnerships

- Administrative supplements for Cancer Centers in GENIE and GA4GH coordination.
- Coordination with and support of Moonshot Programs
  - Assistance for U24 programs, e.g., Human Tumor Atlas & Immuno-oncology Data Coordinating Centers
- Work across related initiatives/programs
  - NCI, other NIH Institutes, NIH Data Commons Pilot Phase Consortium, All of Us, Chan Zuckerberg Initiative, GA4GH
- Establishing NCI Office of Data Sharing as a resource to NCI staff and extramural investigators.
- Workshops and RFIs to gather community input, feedback, and participation
  - Semantics infrastructure workshop
  - Imaging RFI: https://grants.nih.gov/grants/guide/notice-files/NOT-CA-18-060.html
- Establish CRDC governance process, including Scientific and Technical Advisory Board and Steering Committee.

Office of Data Sharing Activities

- Coordinates the interpretation and implementation of data sharing policies across NCI
- Provides workflow management and coordination of NCI data/metadata submissions and access processes relative to NIH databases, including dbGaP
- Advocates for the proper balance of open access, open source, broad data sharing policies
- Outreach and education on NCI data sharing policies and processes; central clearing house for knowledge management
- Develops and monitors metrics relevant for understanding influence, uptake, and compliance regarding NCI data/metadata usage
- Coordinates with and provides leadership as appropriate to other key organizations within NIH and the research community

Office of Data Sharing Tentative Schedule

Jan – May 2018

Begin process of standing up the NCI Office of Data Sharing within CBIIT (approve organizational structure and begin to hire key personnel)

June – Sept 2018

Work with NCI Divisions/Offices/Centers to develop a plan for centralizing current NCI Data Sharing activities

Oct – Dec 2018

Implement centralization plan for NCI Data Sharing activities moving to the ODS

January 2019 –

Work within NCI and the NIH, as well as the external research community, to best implement NIH GDS and other Data Sharing policies as relevant to cancer data and metadata

Cancer Research Data Commons Project Teams

CRDC Framework Principal Investigators

- Robert Grossman - University of Chicago
- Anthony Philippakis - Broad Institute
- Ilya Shmulevich – Institute for Systems Biology
- Brandi Davis-Dusenbery - Seven Bridges

CBIIT Data Commons Team

- Tanja Davidsen
- Ian Fore
- Izumi Hinkson
- Betsy Hsu
- Steve Jett
- Tony Kerlavage
- Juli Klemm
- David Patton

Surveillance Research Program

- Lynne Penberthy
- Paul Fearn

Center for Cancer Genomics

- Lou Staudt
- JC Zenklusen
- Daniela Gerhard
- Zhining Wang
- Liming Yang
- Martin Ferguson

Center for Strategic Scientific Initiatives

- Chris Kinsinger
- Henry Rodriguez

Cancer Imaging Program

- Paula Jacobs
- John Freymann
- Justin Kirby

Leidos Biomedical Data Commons Team

- John Otridge
- Sima Pandya
- Todd Pihl

DEPARTMENT OF HEALTH & HUMAN SERVICES • USA

NIH NATIONAL CANCER INSTITUTE

allen.dearry@nih.gov

www.cancer.gov   www.cancer.gov/espanol

"Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Networking and Information Technology Research and Development Program."

The Networking and Information Technology Research and Development (NITRD) Program

Mailing Address: NCO/NITRD, 2415 Eisenhower Avenue, Alexandria, VA 22314

Physical Address: 490 L'Enfant Plaza SW, Suite 8001, Washington, DC 20024, USA Tel: 202-459-9674, Fax: 202-459-9673, Email: nco@nitrd.gov, Website: https://www.nitrd.gov