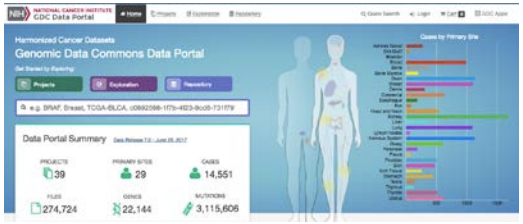
Big Data Sharing Meeting: An Introduction to Today's Meeting

Robert Grossman
Center for Data Intensive Science
University of Chicago
& Open Commons Consortium

HRA Meeting on Big Data Sharing
September 19, 2017

Why Are We Here?

1. Sharing data can advance research discoveries and improve patient outcomes.
2. The foundations in the room today can make decisions that will change the landscape of data sharing in a fundamental one.
3. The technology has been developed and proven within the cancer community (Genomic Data Commons, Cancer Clouds & related projects) and is open source and available to other communities.



Databases

1982 - present

- Data repository
- Researchers download data.

Data Clouds

2010 - 2020

- Supports big data
- Collaborative tools
- Researchers can analyze data (data does **not** have to be downloaded)

Data Commons

2014 - 2024

- Supports big data
- Collaborative tools
- Researchers can analyze data
- Governance
- Common data models
- Harmonized data

Step 3. Fill it with data & develop apps so others can use it.

Start here

Session 3

Step 2. Choose a data commons platform, configure it and operate it.

Sessions 2 & 3

Step 1. Determine an operating, governance & sustainability model.

Step 0. Develop a data commons platform.

This step is **not** necessary for most foundations.

The Three Main Questions

1. **Who shares?**
   - Researchers funded by foundations.

Session 1

2. **How do we share?**
   - Change terms & conditions in grants.
   - Select governance and operating models.
   - Select data commons platform.
   - Fund and build commons (lead, co-lead, or join)
   - Fund bioinformaticians to submit data.

Session 2

Session 3

3. **How do we interoperate commons?**
   - Focus on the large collections / commons of data.
   - Alliances (require commons to interoperate)
   - Fund commons to interoperate.

Session 4

What You Need to Decide

- Will you **build** your own data commons (perhaps with others)?
- Or, do you prefer to **join** an existing data commons?
- Do you prefer to create a commons that is your own foundation's **brand** (Foundation A's Data Commons), to lead or join data commons focused on a disease (Disease A), or a research area (Brain Commons)?

What You **Don't** Need to Do

- You do not have to build a data commons **platform**. You can simply choose one of the six here today.

The Players in the Data Sharing Ecosystem

Researchers

Medical Research Centers

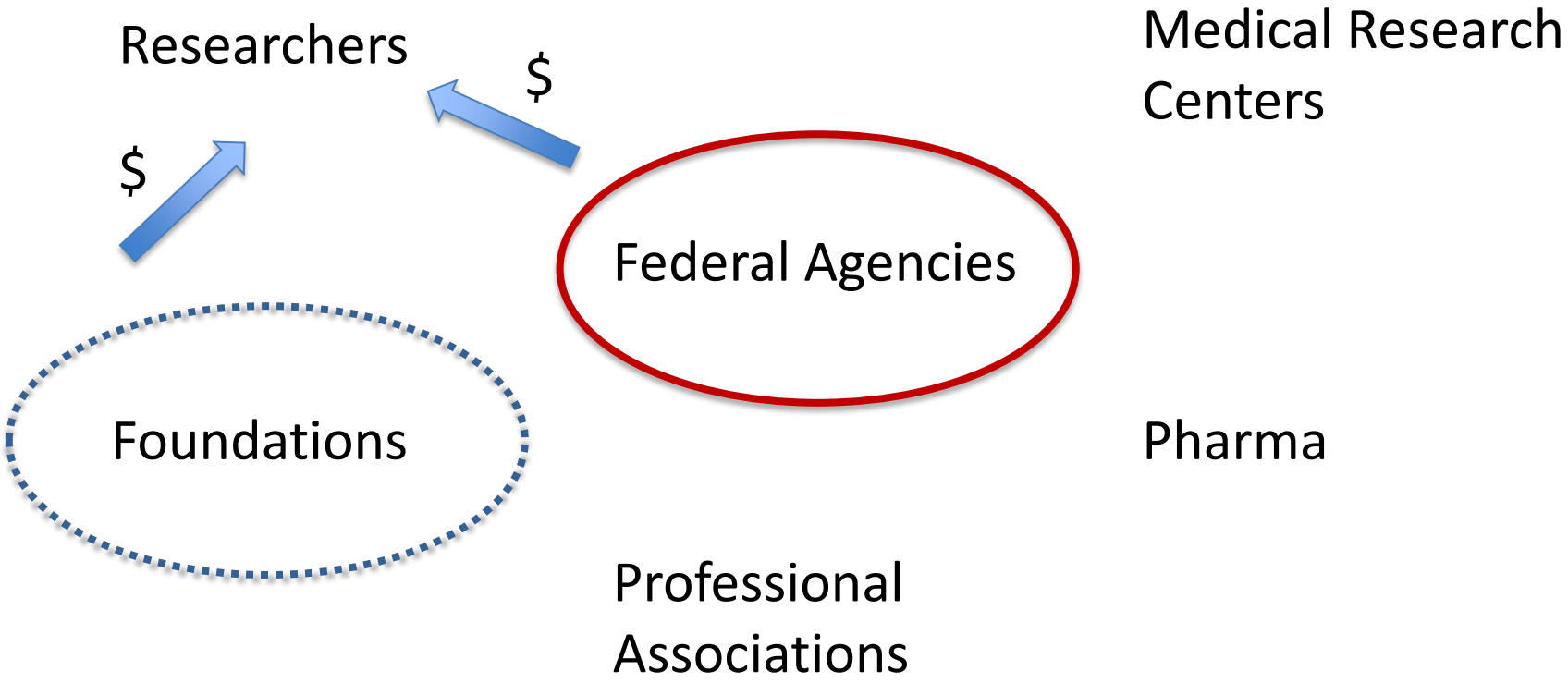
Federal Agencies

Foundations

Pharma

Professional Associations

The Players in the Data Sharing Ecosystem



Today, We Should Not Worry About

- The role of medical research foundations, professional associations, pharma and others in data sharing.
- “Individual, small datasets.”
- The complexity of EMRs, ontologies, etc.
- The failures of the past.
- Whether there will be coffee at the breaks.

Our Focus Today

Share these

10,000's to 100,000's of individual small datasets and databases

100's to 1000's programs/projects/commons with data governance & multiple projects/datasets

Small studies and datasets

Projects / programs with governance & multiple datasets

The Tragedy of the Commons

Individuals when they act independently following their self interests can deplete a deplete a common resource, contrary to a whole group's long-term best interests.

Source: Garrett Hardin, The Tragedy of the Commons, Science, Volume 162, Number 3859, pages 1243-1248, 13 December 1968.

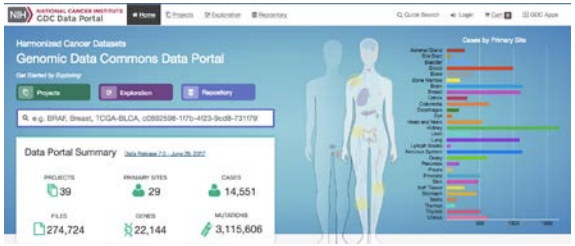
Bermuda Principles
& Genomic Databases
(e.g. GenBank)
1982 - present

PLOS

PMC arXiv.org

bioRxiv

Open Access Principles for Publications
arXiv, PubMed Central
2010 - present



Chicago Principles
Data Commons
2017 -

Lets debate, draft and sign these by Dec 15, 2017

Bermuda Principles

1. Automatic release of sequence assemblies larger than 1 kb (preferably within 24 hours).
2. Immediate publication of finished annotated sequences.
3. Aim to make the entire sequence freely available in the public domain for both research and development in order to maximise benefits to society.

Source: Summary of the Report of the Second International Strategy Meeting on Human Genome Sequencing (Bermuda, 27th February - 2nd March, 1997) as reported by HUGO, http://web.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml

Chicago Principles

1. Require that researchers share the data generated by research that you fund.
2. Foundations should provide the computing infrastructure and bioinformatics resources that is required to support data sharing.
3. The data commons supported by Foundations should themselves share data and interoperate with other data commons.

- U.S based 501(c)(3) not-for-profit corporation founded in 2008.
- Supports data commons to support biological, medical and health care research: BloodPAC Data Commons, Brain Commons and Bionimbus.
- Manages data commons and cloud computing infrastructure to support scientific research: Open Science Data Cloud, Project Matsu (OCC & NASA), and the OCC NOAA Data Commons.
- The OCC is international and includes universities, not-for-profits, companies and government agencies.
- The OCC has templates for building data commons.
- The OCC contributes to the open source software community.

www.occ-data.org

Summary

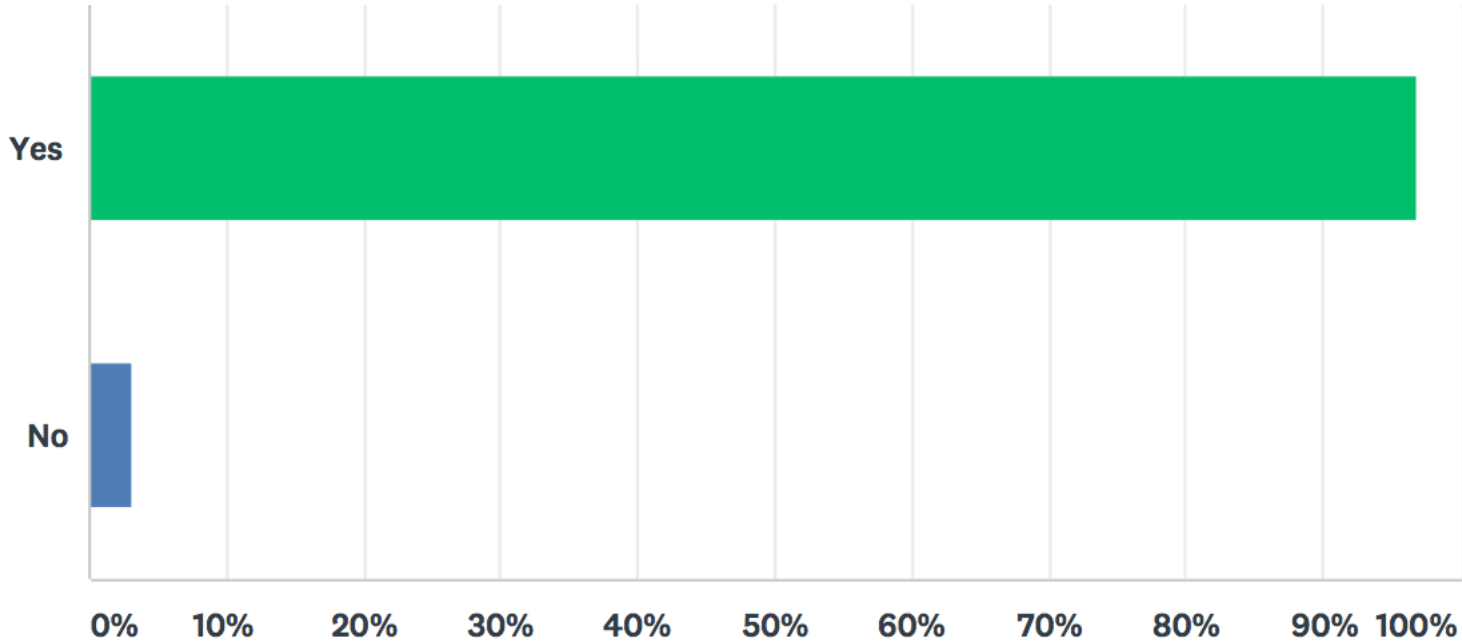
1. Data sharing and open science will accelerate research and improve patient outcomes.
2. Data commons are a proven technology to support data sharing, open data and open science.
3. Foundations have a critical role to play and can disruptively change the open data and open science landscape.
4. There are proven technologies, governance models and operating models for building and operating data commons to support data sharing.

Questions?

rgrossman.com
@bobgrossman

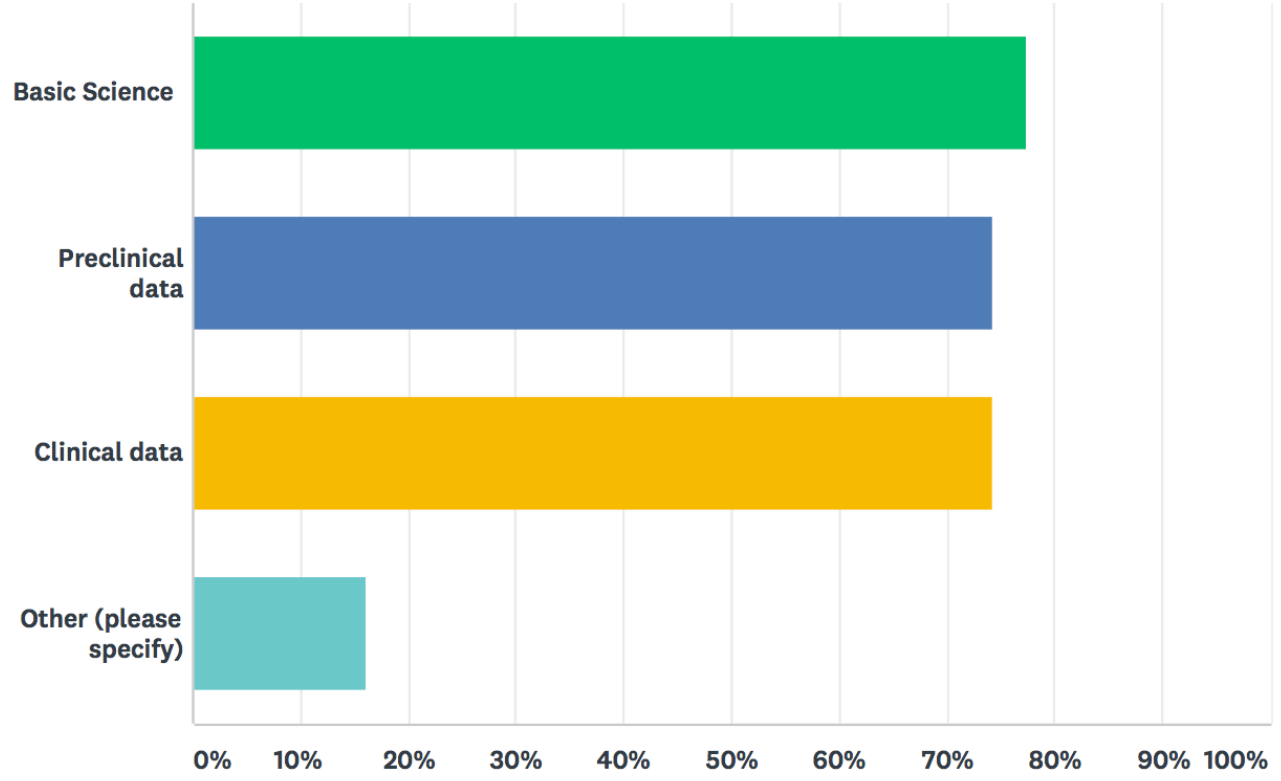
Q1 Is your organization interested in publicly sharing the data generated by your grantees?

Answered: 31 Skipped: 0



Q2 Do you know what kind of data your grantees generate?

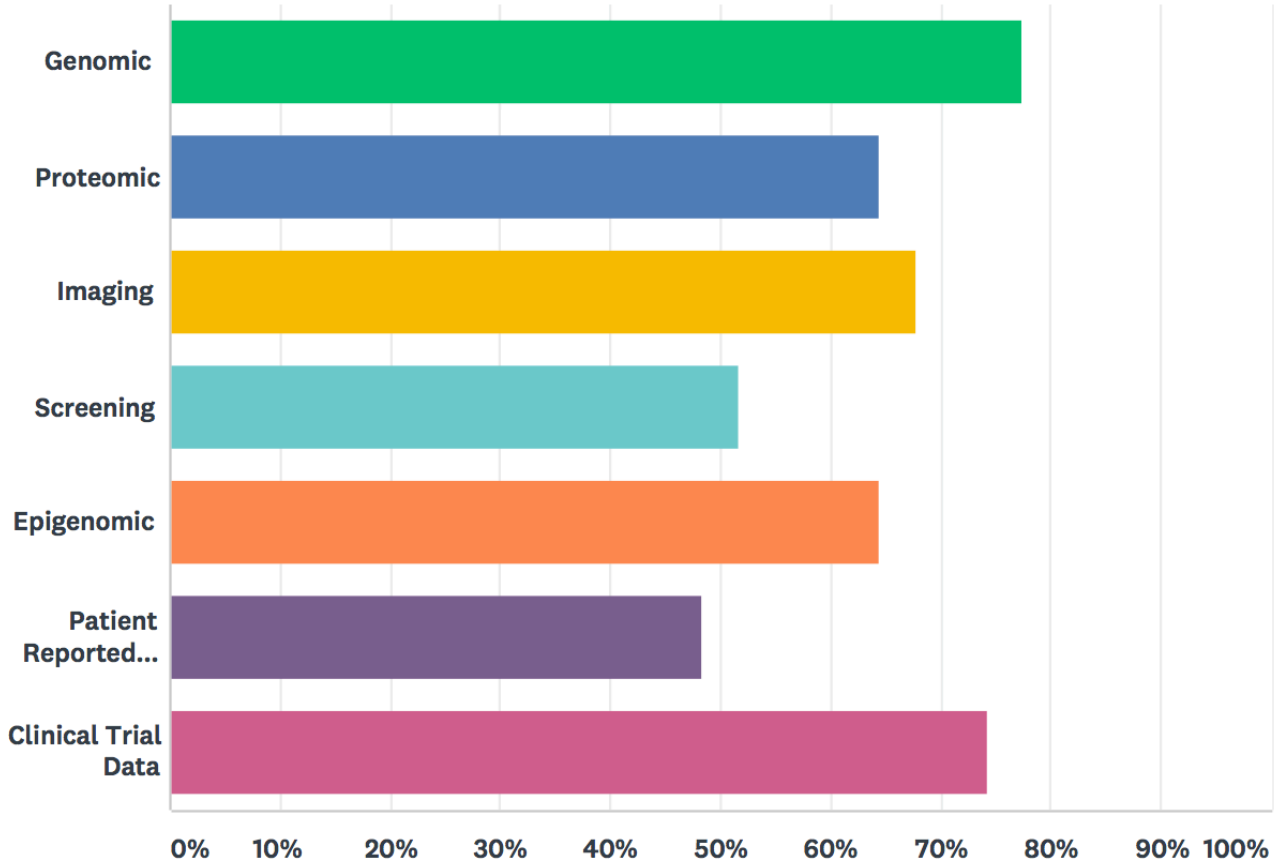
Answered: 31 Skipped: 0



#	OTHER (PLEASE SPECIFY)	DATE
1	All of the above	9/18/2017 2:31 PM
2	CER	9/18/2017 9:25 AM
3	non clinical data	9/15/2017 4:22 PM
4	technical data needed for product development	9/13/2017 6:16 PM
5	Health services	9/13/2017 4:12 AM

Q3 What types of data do your funded research projects produce?

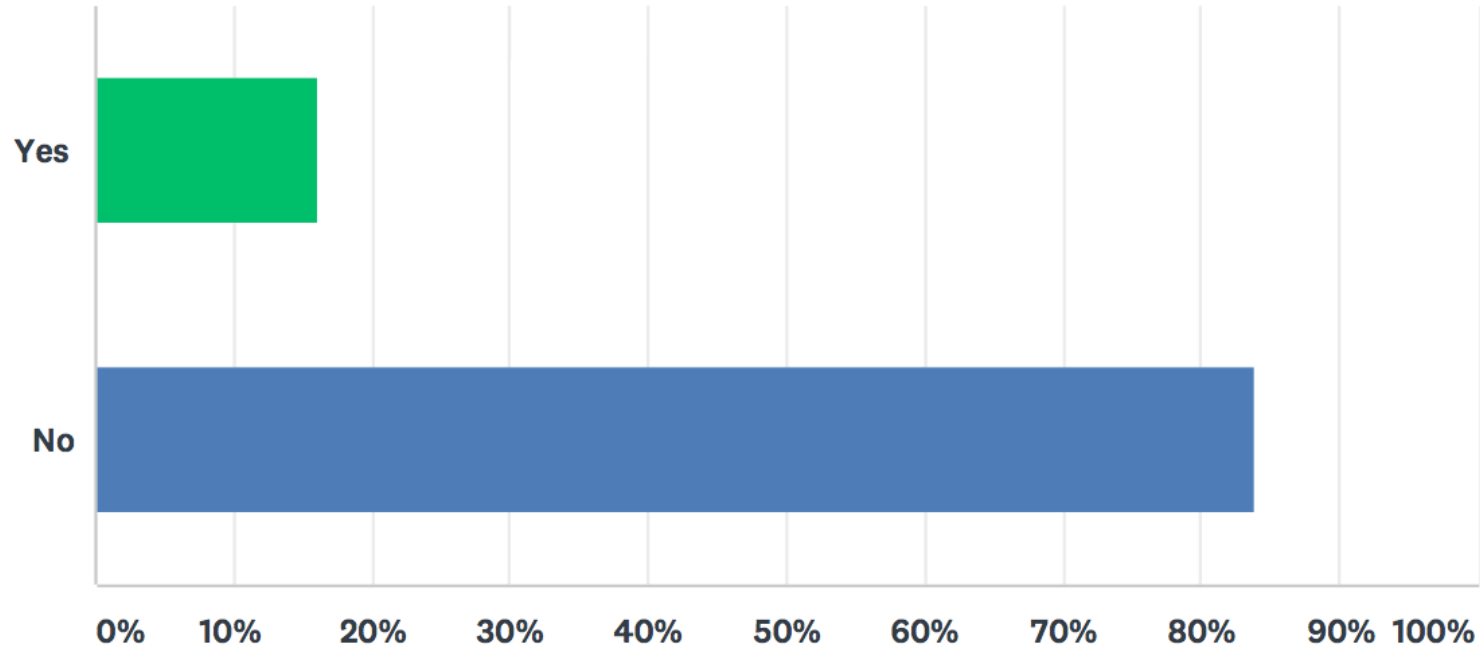
Answered: 31 Skipped: 0



#	OTHER (PLEASE SPECIFY)
1	Electrical recordings
2	preclinical and technical data sets
3	questionnaire data, neurocognitive test data, medical record data, clinical test data (eg overnight polysomnogram)

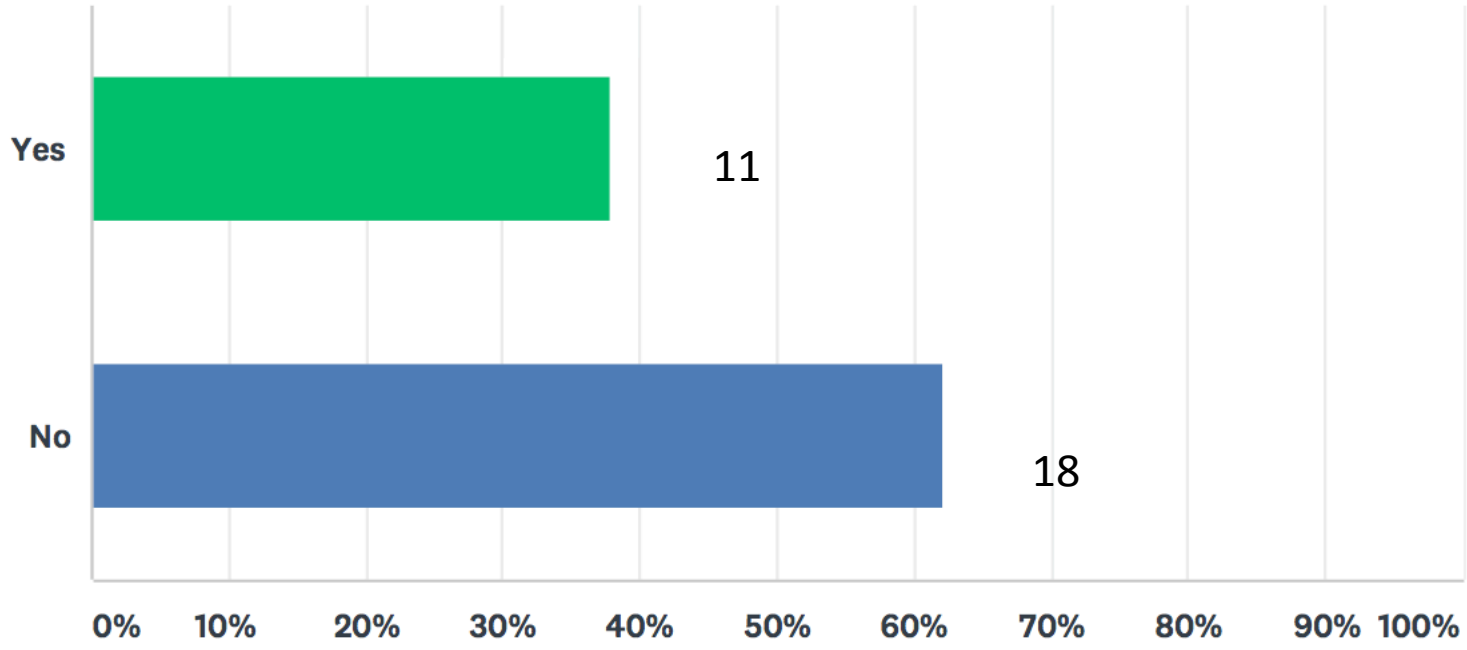
Q4 If your funded researchers do produce genomics, proteomics, imaging, screening, etc. data, do you know how much data they produce?

Answered: 31 Skipped: 0



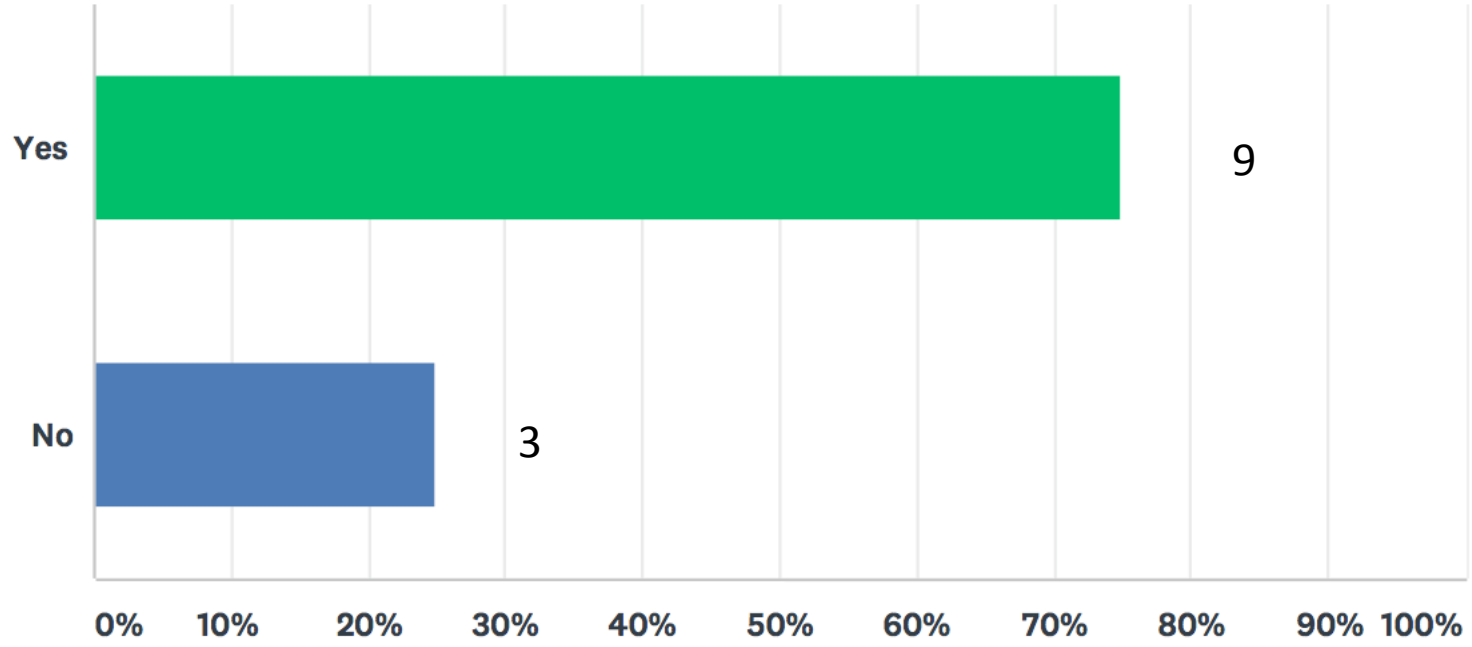
Q5 Do you require that grantees share data generated by research projects that you fund?

Answered: 29 Skipped: 2



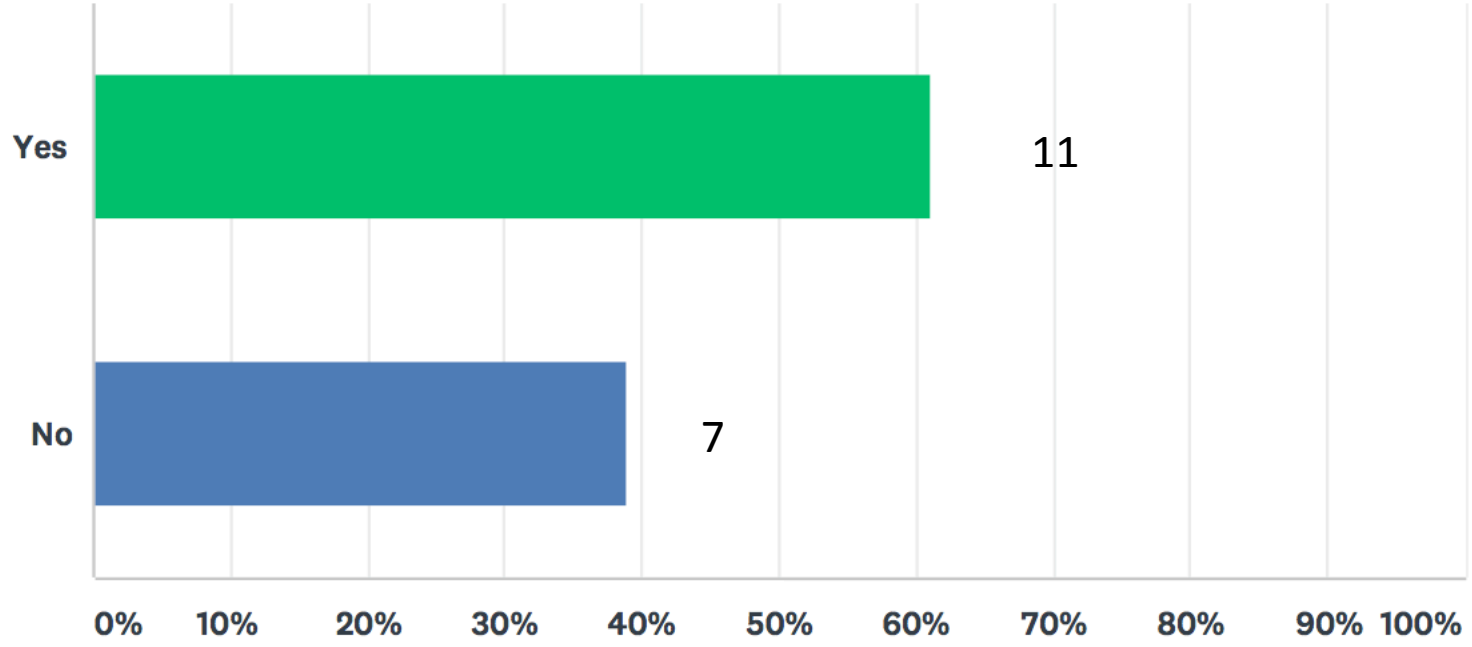
Q6 Do you require data sharing in your grant agreements?

Answered: 12 Skipped: 19



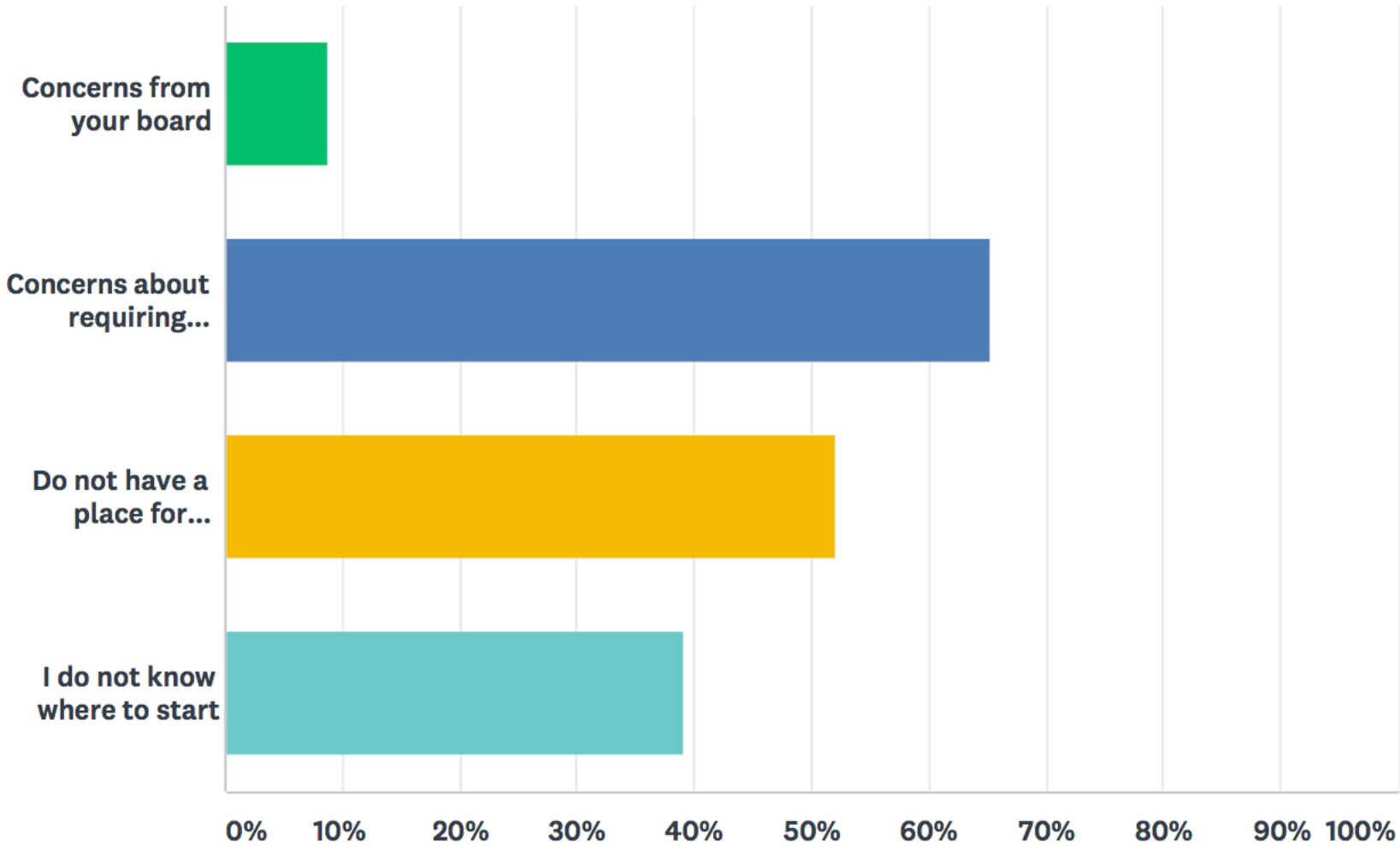
Q7 Do you plan to require grantees to share data in the future?

Answered: 18 Skipped: 13



Q8 What are the barriers for your organization to start sharing data?

Answered: 23 Skipped: 8



Q8 What are the barriers for your organization to start sharing data?

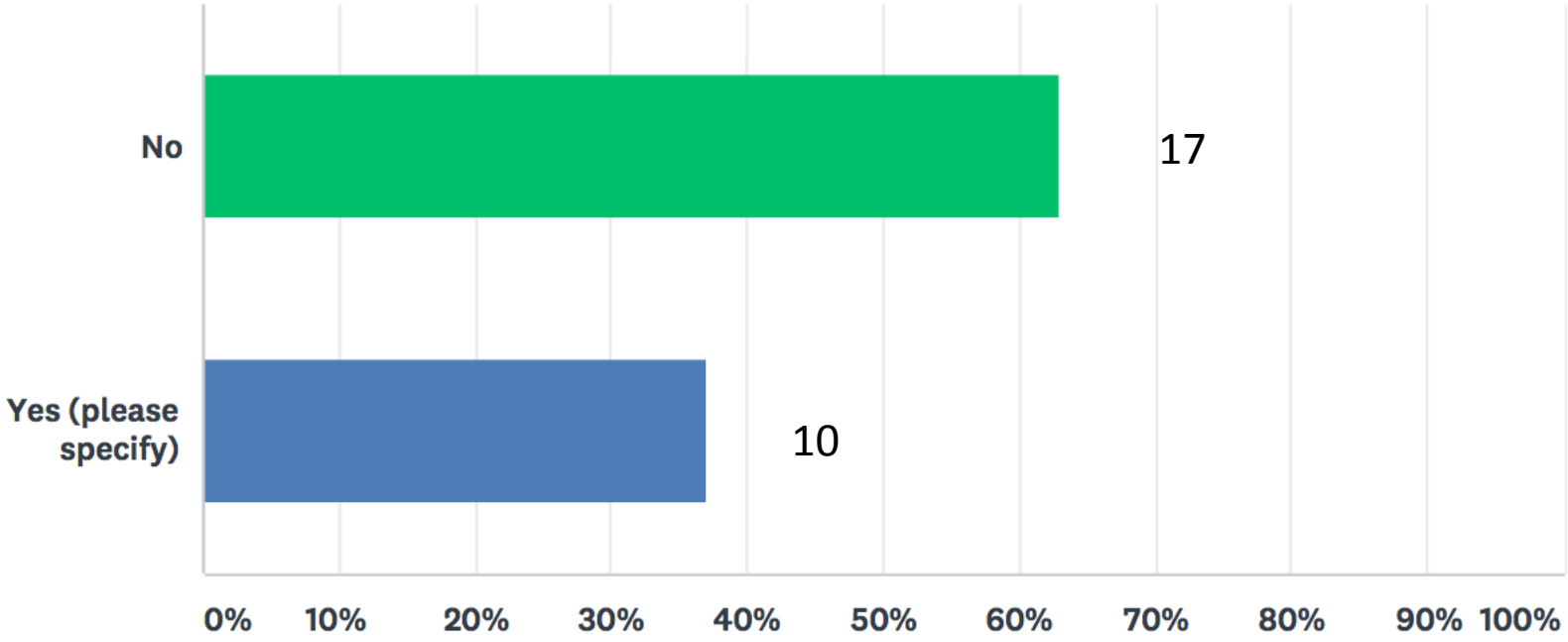
Answered: 23 Skipped: 8

ANSWER CHOICES	RESPONSES	
Concerns from your board	8.70%	2
Concerns about requiring researchers to share data	65.22%	15
Do not have a place for researchers to deposit data	52.17%	12
I do not know where to start	39.13%	9
Total Respondents: 23		

#	OTHER (PLEASE SPECIFY)	DATE
1	Variability in institutional policies, sharing secondary data used in study, defining shared elements	9/18/2017 9:32 AM
2	we are sharing data already	9/18/2017 8:43 AM
3	Privacy rules and other legal/regulatory issues.	9/18/2017 8:39 AM
4	we are relying on publication policies to require data sharing and house it	9/14/2017 6:06 PM
5	specs for how the data should be submitted - format etc	9/13/2017 6:17 PM
6	Researchers have been reticent to share data due to confidentiality	9/13/2017 10:10 AM
7	We don't see barriers -- we require them to work out where data will be stored in advance of signing the grant agreement. Normally in a controlled access repository, preferably government hosted. For smaller studies where the data may not be as valuable we are not as rigorous but still require a sharing plan.	9/13/2017 9:03 AM

Q9 Do you know of a data-sharing platform that your organization would use or suggest to grantees?

Answered: 27 Skipped: 4



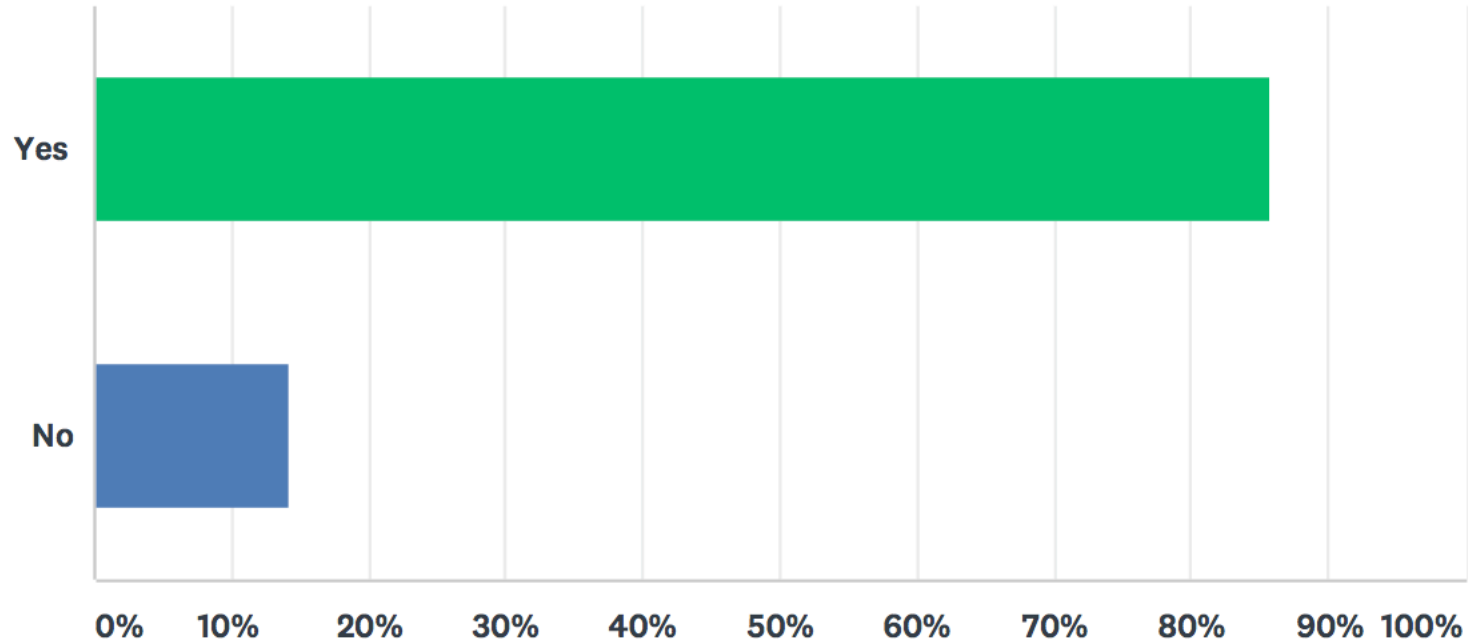
Q9 Do you know of a data-sharing platform that your organization would use or suggest to grantees?

Answered: 27 Skipped: 4

#	YES (PLEASE SPECIFY)	DATE
1	Figshare, AWS, CRCNS.org, others	9/18/2017 2:46 PM
2	Brigham and Young Multi-regional clinical trials center and Inter university consortium for political and social science at UMich	9/18/2017 9:32 AM
3	synapse	9/18/2017 8:43 AM
4	there are many - too confusing	9/18/2017 8:40 AM
5	GAAIN (gaain.org)	9/18/2017 8:39 AM
6	GDC or Figshare	9/15/2017 11:16 AM
7	Center for Open Science	9/13/2017 6:17 PM
8	COS open science framework is an option	9/13/2017 10:10 AM
9	NIMH Data Archive (NDA), dbGaP, NRGR, EBI, etc.	9/13/2017 9:03 AM
10	We know multiple places where they are sharing, need to hone in on what should be required	9/12/2017 6:56 PM

Q10 Would you be willing to partner with another organization in creating a data repository or using the same repository, based on type of data, type of disease area, etc.?

Answered: 28 Skipped: 3



Big Data Sharing Drivers in Oncology and Precision Medicine

Warren A. Kibbe, Ph.D.
Professor, Biostats & Bioinformatics
Chief Data Officer, Duke Cancer Institute
warren.kibbe@duke.edu

@wakibbe

Outline

- Background
- Cancer as a model and driver – focus more on clinical translational research than basic science research
- Buzzword Bingo
- Take homes

Personal & Professional Background

- PhD in Chemistry at Caltech, Postdoc in molecular genetics of RAS
- Cancer research for 20+ years - cancer informatics, data science, healthcare
- Faculty in the Feinberg School of Medicine at Northwestern for 15+ years
- Director NCI CBIIT 2013-2017; Acting NCI Deputy Director 2016-2017
- Lost three grandparents to cancer

Open Data

- Cancer Moonshot Data Sharing Policy
- **Informed Consent** supports sharing
- Balance Risk vs Benefit
- Promotes **Ethical Behavior**
- Speeds **Discovery**
- Enables and Enhances **Collaboration**
- Drives **Innovation**

Cancer is a grand challenge

Courtesy of P. Kuhn (USC)

Requires:

- Deep biological understanding
- Advances in scientific methods
- Advances in instrumentation
- Advances in technology
- Data and computation
- Mathematical models

*Cancer Research and Care generate detailed **data** that is critical to create a learning health system for cancer*

In 2016 there were an estimated

15,500,000

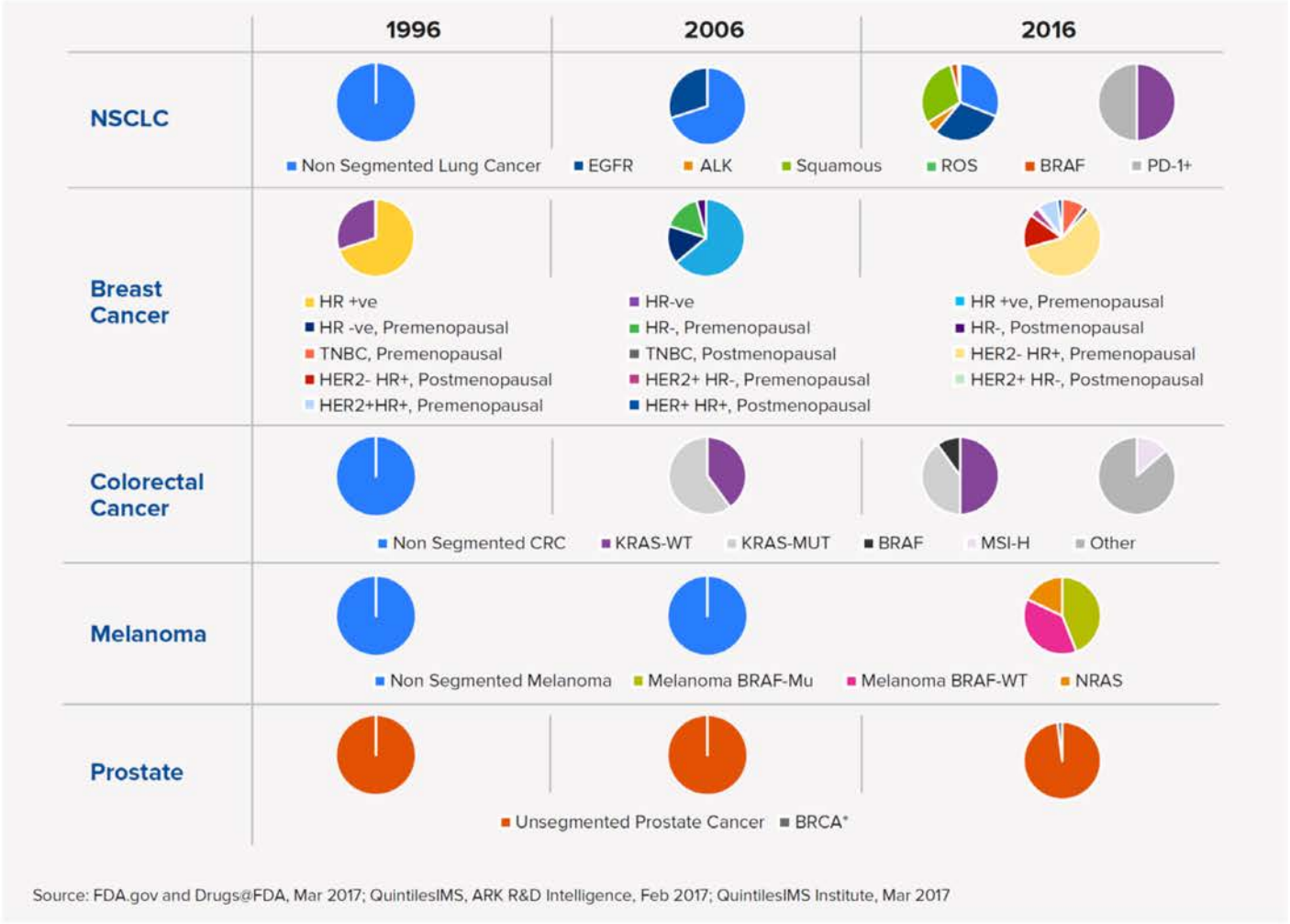
cancer survivors in the U. S.

Understanding Cancer

- **Precision medicine** will lead to **fundamental understanding** of the complex interplay between genetics, epigenetics, nutrition, environment and clinical presentation and **direct effective, evidence-based prevention and treatment**.

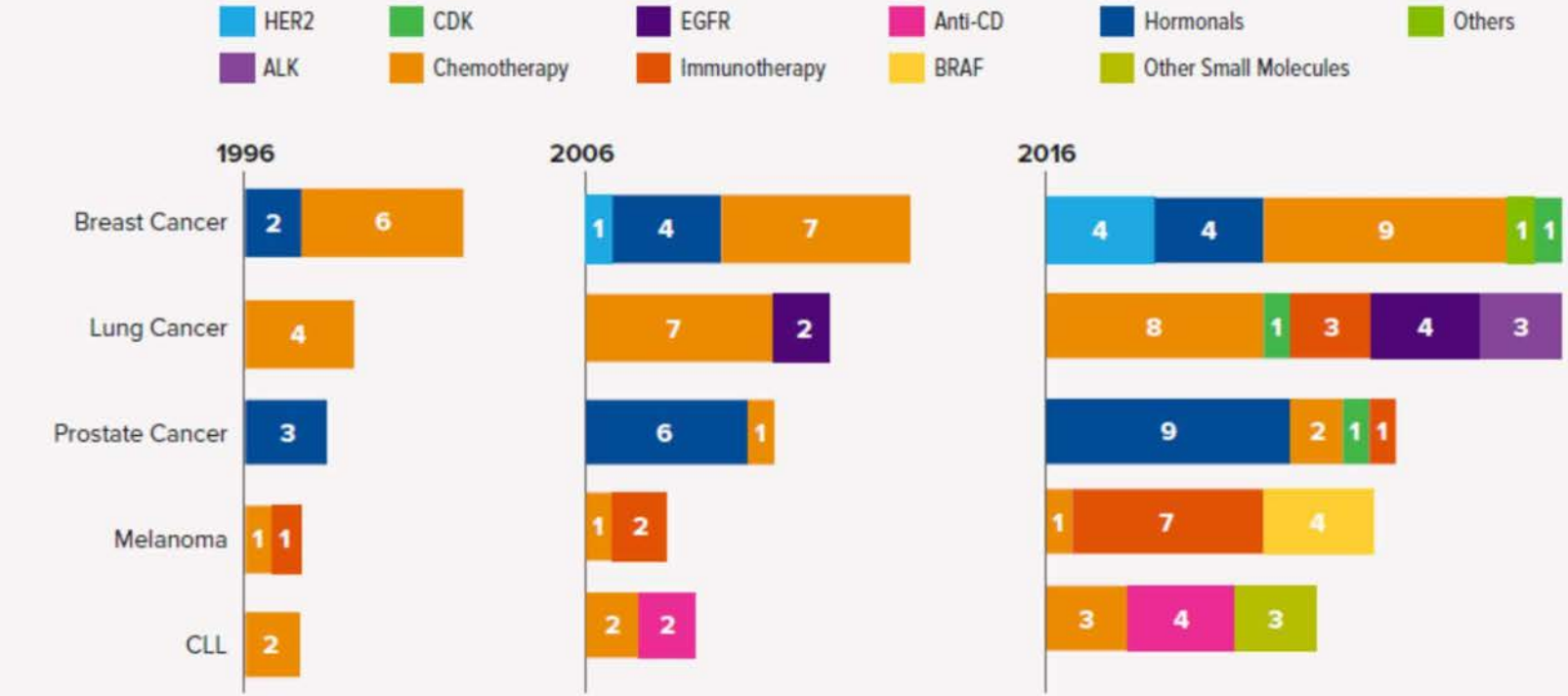
Ramifications across many aspects of health care

Cancer has been progressively redefined over the past 20 years



This redefinition has been driven by improved biological understanding

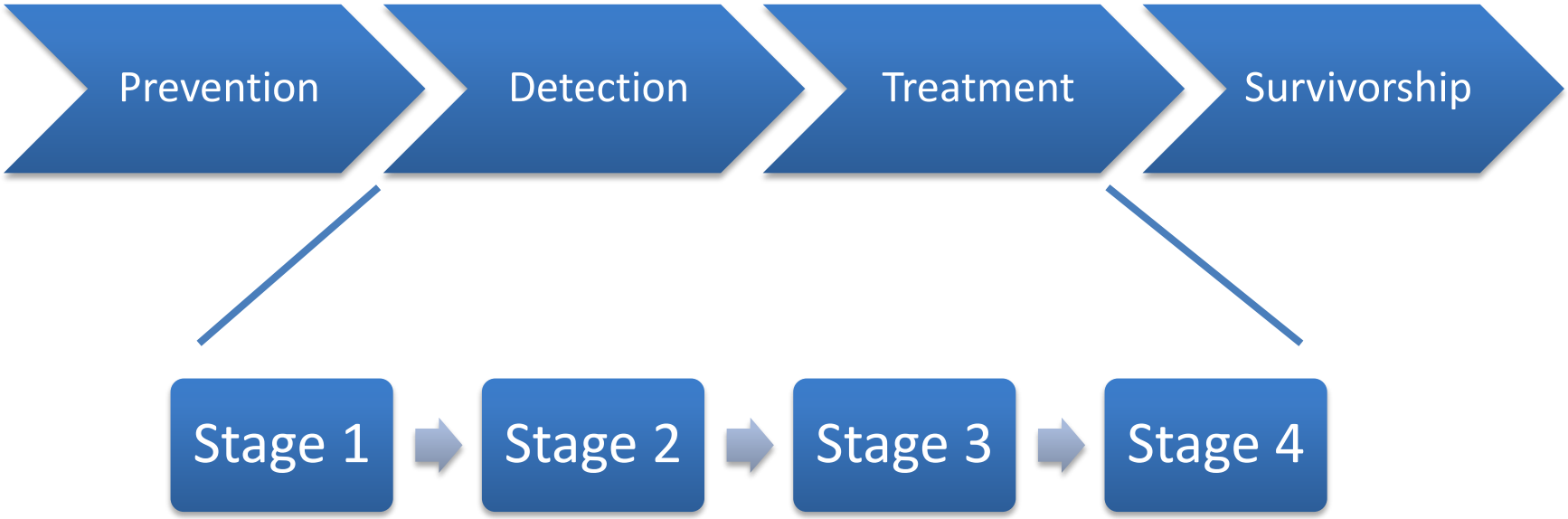
Number of Treatment Options over Time for Selected Tumors (1996–2016)



Source: Drugs@FDA, Feb 2017; QuintilesIMS, ARK R&D Intelligence, Feb 2017; QuintilesIMS Institute, Mar 2017

This change has been driven by improved technology - sequencing, imaging, nanotech, drug developing, computing and the availability of data about patient response to therapy

Open Data drives Innovation



Open Data enables validation
Open Data enables benchmarking

How do we solve problems in Cancer

- Support and incentives for team science, collaboration
- We need FAIR, open data
- Support open source, open science
- Support for rapid innovation

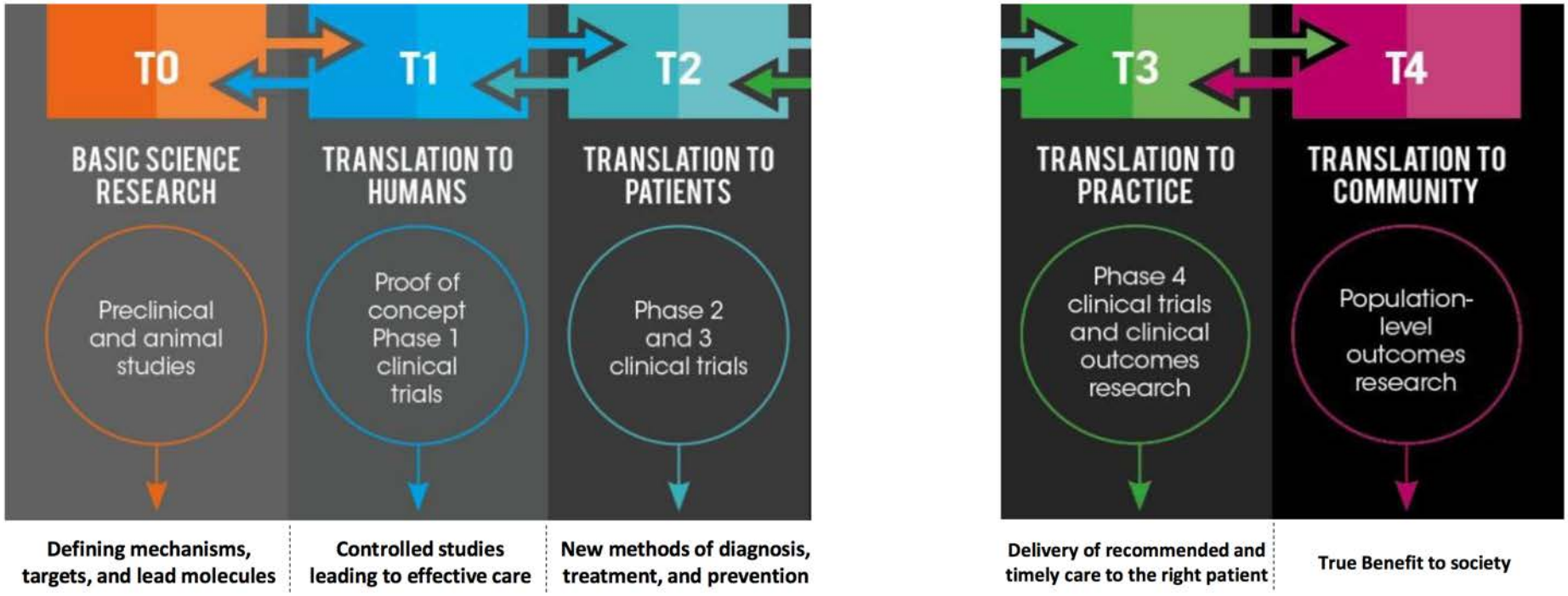
Data Sharing and the FAIR Principles

FAIR –
Making data
Findable,
Accessible,
Attributable,
Interoperable,
Reusable,
and provide Recognition

Force11 white paper
https://www.force11.org/group/fairgroup/fairprinciples

Translational from basic science to human studies

Translational of new interventions into the clinic and health decision making



Pieces of a Learning Healthcare System

Open Data is critical for identifying new opportunities and evaluating the effectiveness of new technologies, procedures, techniques

Vision:

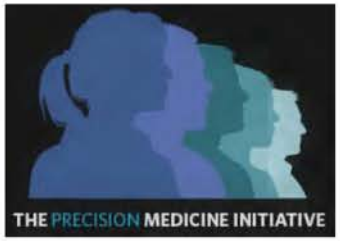
Enable the creation of a *Learning Healthcare System for Cancer,* where as a nation we learn from the *contributed knowledge* and experience of *every cancer patient*. As part of the Cancer Moonshot, we want to *unleash the power of data* to enhance, improve, and inform the journey of every cancer patient from the *point of diagnosis through survivorship*.

Team Science is critical

- Analytics and Visualization
- Bioinformatics
- Biostatists
- Clinical Care
- Clinical Research
- Clinical Trials
- Data Science
- EHRs, Imaging, Lab Systems

Open Data enhances collaboration and team science!

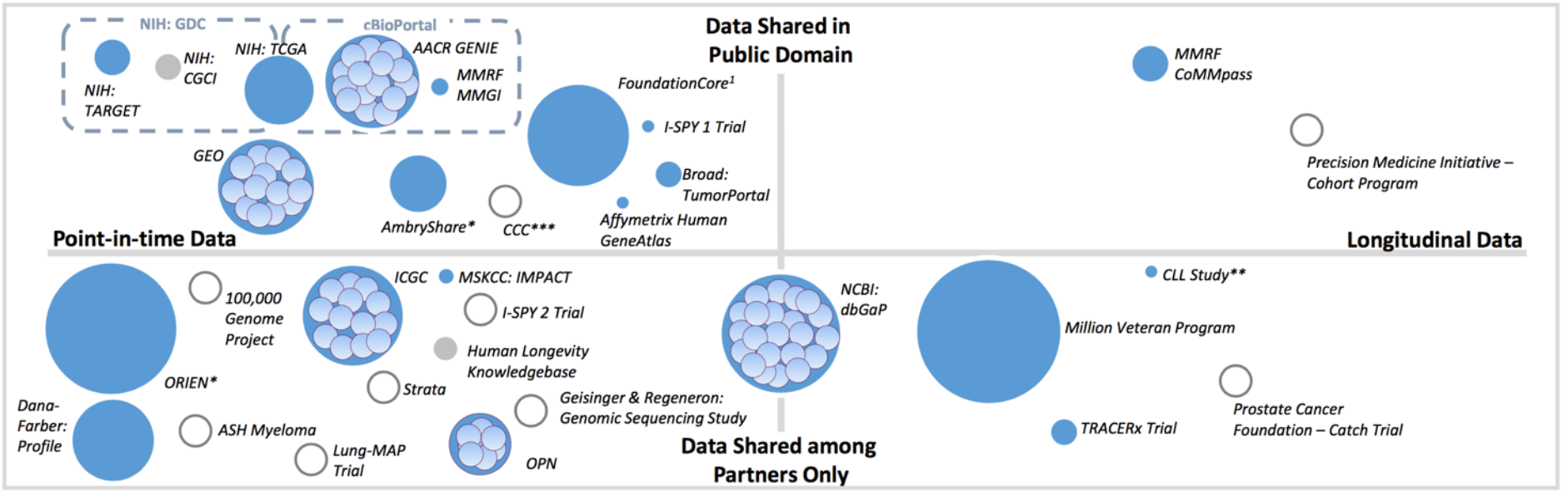
Scale is changing!



1 million healthy genomes

2001 **2010** **2015**

Sharing and complexity



Bubble size = estimated size of available dataset
Note: Representative selection of landscape, not all inclusive

Available datasets, size known
Available datasets, size unknown
Dataset in development
Available datasets that comprise multiple smaller datasets
Portals that are currently active/utilized in the public domain

Opportunity exists to generate publicly available longitudinal data to drive understanding of genetic mutations and find Precision Medicine cures

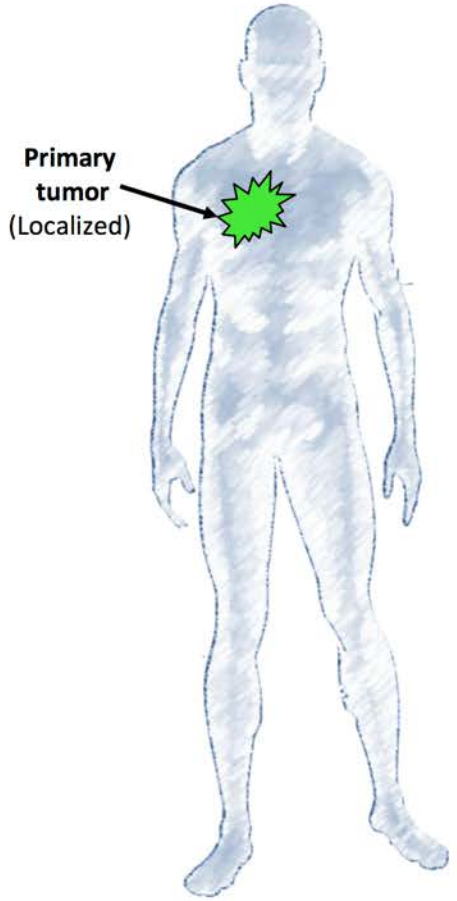
*Datasets have potential to include longitudinal data in the future
**Public/private information not available
***Serves as a portal also, has potential to include longitudinal data in the future

1. FoundationCore's pediatric cancer data has been made public

HARVARD BUSINESS SCHOOL
Kraft Precision Medicine Accelerator

From HBR, June 2016

2006-2015: A Decade of Illuminating the Underlying Causes of **Primary Untreated Tumors**



The Cancer Genome Atlas

CLINICAL PROTEOMIC TUMOR ANALYSIS CONSORTIUM

CANCER IMAGING ARCHIVE

(12,000+ patient tumors and increasing)

Openly shareable, but not always easy to access

Buzzwords

- Big Data – either population based or a large enough sample of people /instruments /activities that it is more than a selected sample
- Open Access – Freely available, may have usage restrictions
- Public Access – need to verify identity of people and sometimes purpose of access before granting access
- Proprietary – requires agreements to access

Buzzwords

- HIPAA – Privacy rule established standards for when and how to share identifiable patient data. https://www.hhs.gov/hipaa/for-professionals/privacy/index.html
- PII – personally identifiable information
- PHI – protected health information, from HIPAA
- Limited data sets – usually controlled access, with much of the granularity of patient data removed
- De-identified data sets – identifiers have been removed and risk of identifying individuals is low. Moving target!

Buzzwords

- EHRs – Electronic Health Records
- EMRs – Electronic Medical Records
- NGS -Next generation sequencing. Can be of many types. Targeted DNA sequencing (panels). Whole Exome (WES), Whole Genome (WGS), RNAseq, the Epigenome (methylSeq)

Buzzwords

- Population-based – data sets and studies that cover *everyone* in a given population – can be geographical, disease-based, or some other characteristic
- Registries – collections of information (patient generated, EHRs, surveys, medicare/medicaid, etc), biospecimens, environmental samples, etc

Machine Learning

- Large data sets, particularly population-based with a well-annotated comparator set, are ideal
- Machine Learning and Deep Learning on image features is feasible, accurate, reproducible and scalable

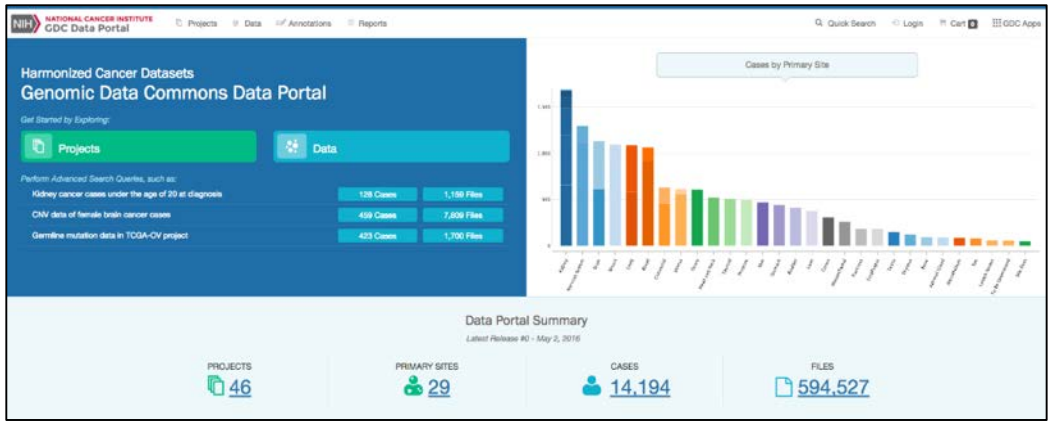
Cancer Genomics

- Several distinct molecular forms of cancer at each organ site
- The genomic abnormalities of each cancer are unique
- The same molecular abnormalities are found in cancers that arise in different organs

Our understanding of biology, cancer, and intervention is changing based data from open resources like TCGA, GENIE, etc!

GDC is an example of a new architecture for storing and sharing cancer data

NCI Genomic Data Commons
launched at ASCO on June 6, 2016



https://gdc.cancer.gov

2.6 PB of legacy data and **1.5** PB of **harmonized** data.

Biology and Medicine are now data intensive enterprises

Scale is rapidly changing

Technology, data, computing and IT are pervasive in the lab, the clinic, in the home, and across the population

Real World Evidence

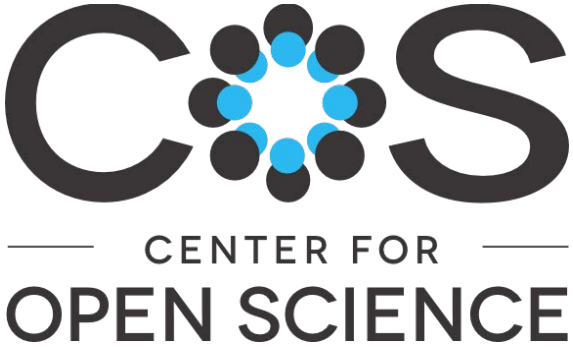
- Needs big data! Big **open data!**
- Needs **population** representation
- Need **epidemiologists** and **statisticians** to understand the potential **biases** in representation
- **EHRs**, **NLP**, **Machine Learning** can power real world evidence learning
- Critical for a **Learning Health System**

Questions?

Warren Kibbe, Ph.D.

warren.kibbe@duke.edu

@wakibbe



Data Sharing Now and in the Future

Brian Nosek

University of Virginia -- Center for Open Science

http://briannosek.com/ -- http://cos.io/

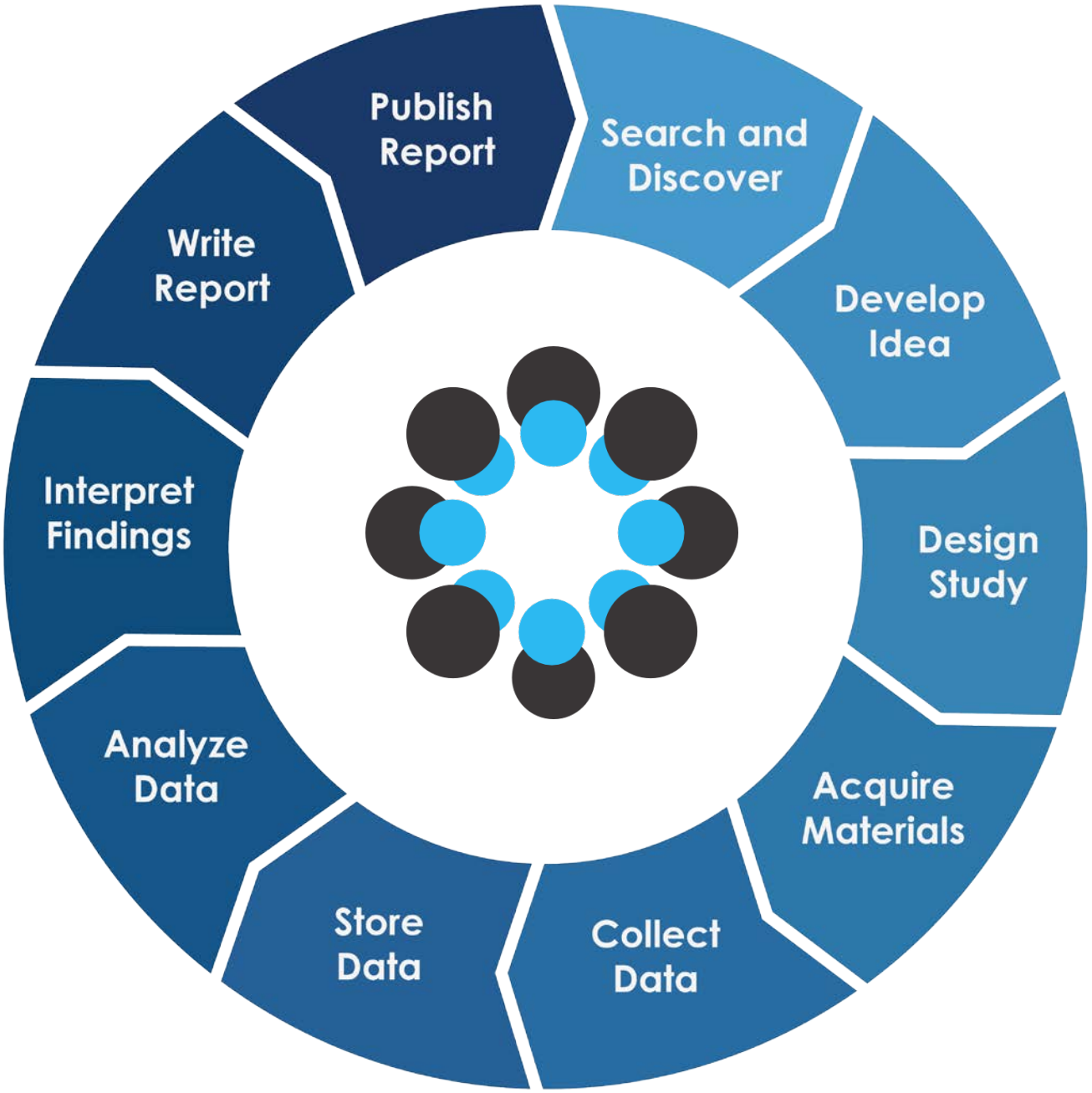
The Kindergartener's Guide to Improving Research

1. Show your work
2. Share

Technology to *enable* sharing

Training to *enact* sharing

Incentives to *embrace* sharing


Search and Discover
Develop Idea
Design Study
Acquire Materials
Collect Data
Store Data
Analyze Data
Interpret Findings
Write Report
Publish Report

Funders Getting Started with Data Sharing

Part 1: Review & update data sharing policy

Part 2: Guidance to grantees for data archiving

Part 3: Training for grantees to do it well

Part 4: Initial steps for monitoring and reporting

Part 1: Review and update data sharing policy

TOP Guidelines http://cos.io/top

1. Data citation
2. Design transparency
3. Research materials transparency
4. Data transparency
5. Analytic methods (code) transparency
6. Preregistration of studies
7. Preregistration of analysis plans
8. Replication

Some TOP Signatory Organizations

- AAAS/Science
- American Academy of Neurology
- American Geophysical Union
- American Heart Association
- American Meterological Society
- American Society for Cell Biology
- Association for Psychological Science
- Association for Research in Personality
- Association of Research Libraries
- Behavioral Science and Policy Association
- BioMed Central
- Cell Press
- Committee on Publication Ethics
- Electrochemical Society
- Elsevier
- Frontiers
- Laura and John Arnold Foundation
- MDPI
- Mind and Life Institute
- Nature-Springer
- PeerJ
- Pensoft Publishers
- Public Library of Science
- The Royal Society
- Society for Personality and Social Psychology
- Society for a Science of Clinical Psychology
- Ubiquity Press
- Wiley

TOP Data Transparency Levels

1	Report states whether data are available, and, if so, where to access them
2	Data must be posted to a trusted repository. Exceptions must be identified at report submission.
3	Data must be posted to a trusted repository, and reported analyses will be reproduced independently prior to publication.

Part 1: Policy To-dos

1. Become TOP signatory

2. Does your data sharing policy meet at least TOP Level 1 on data sharing?

3. Update data sharing policy to align with intentions using TOP framework

- Free consulting: David Mellor, david@cos.io

Part 2: Grantee Guidance

Q: Is there a repository for my kind of data?

A: http://re3data.org/

Q: If not, can a general repository handle my data?

A: OSF, Zenodo, Dataverse, figshare, Dryad

Q: If not, how can I share my data?

A: Direct consult with storage provider (e.g., OSF: Owncloud, Amazon S3, support@osf.io)

Part 2: Guidance To-dos

1. Preferred repositories?

2. Otherwise, give simple sequence:
   re3data.org + generalized repositories

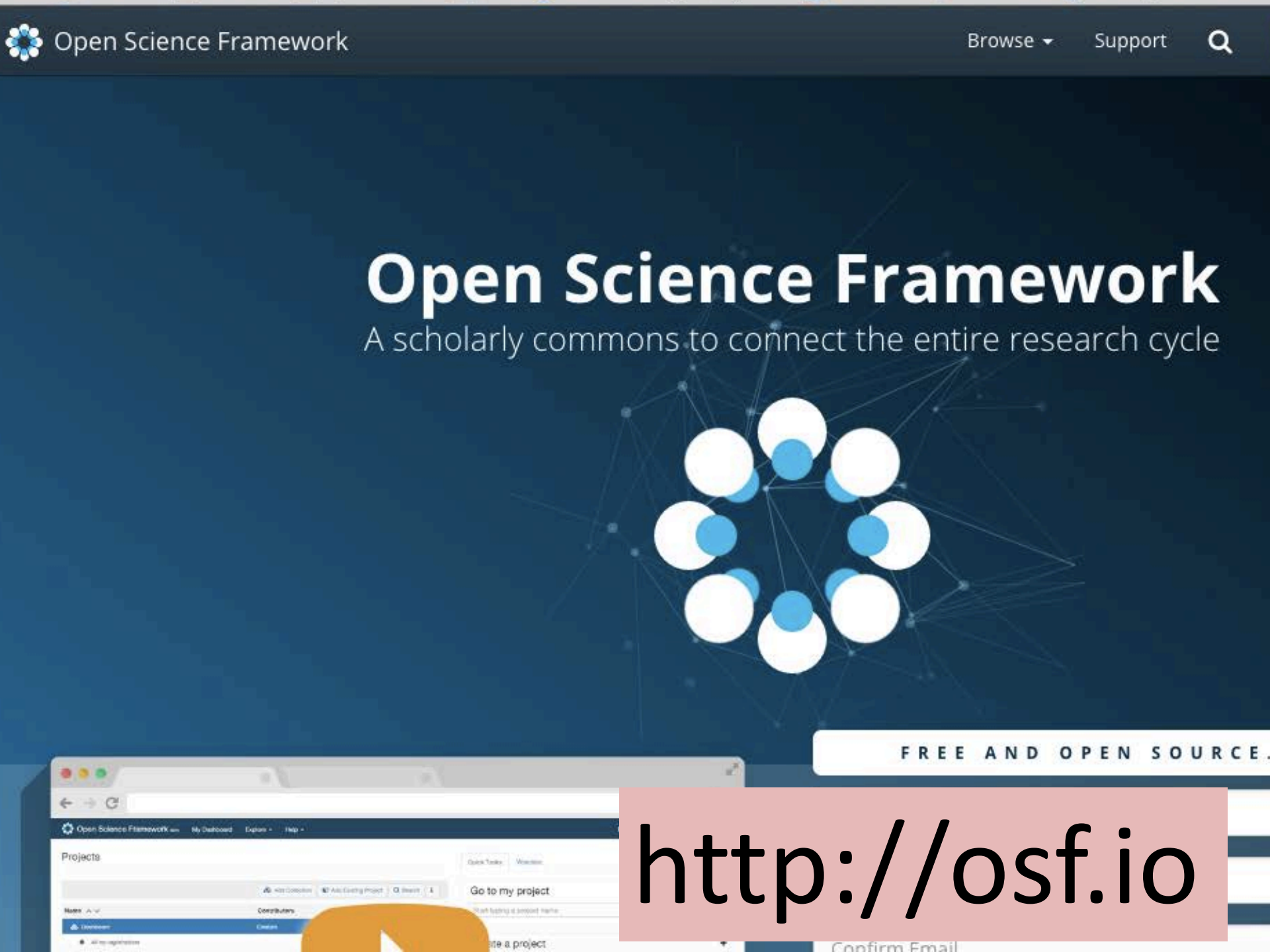
- Free consulting: David Mellor, david@cos.io

Barriers for researchers

I am not organized to share

I don’t have time to share

I am not ready to share


Open Science Framework
Browse
Support
Open Science Framework
A scholarly commons to connect the entire research cycle
FREE AND OPEN SOURCE.
Projects
Go to my project
http://osf.io

Barriers for researchers

I am not organized to share

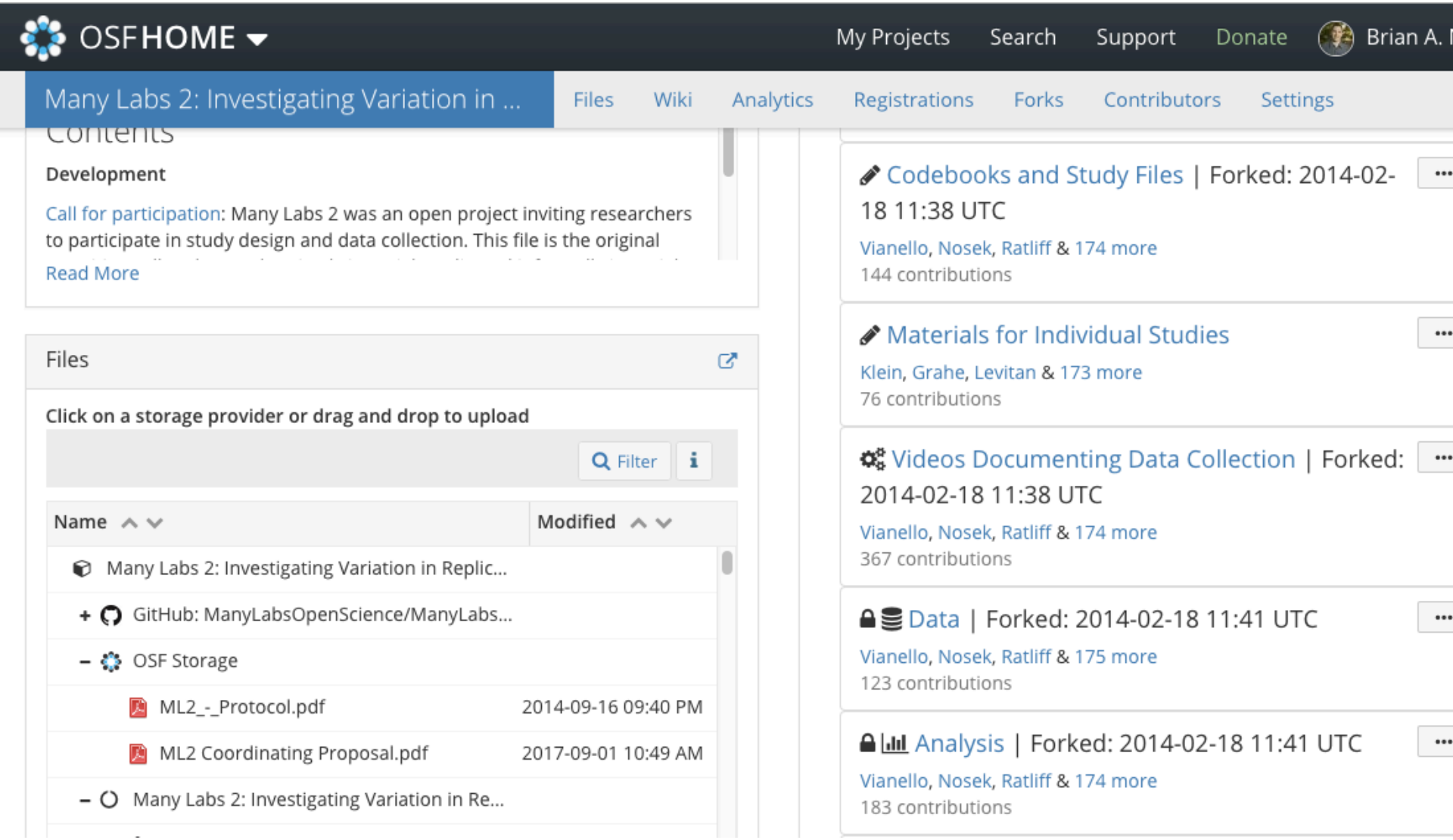
OSF: Project management

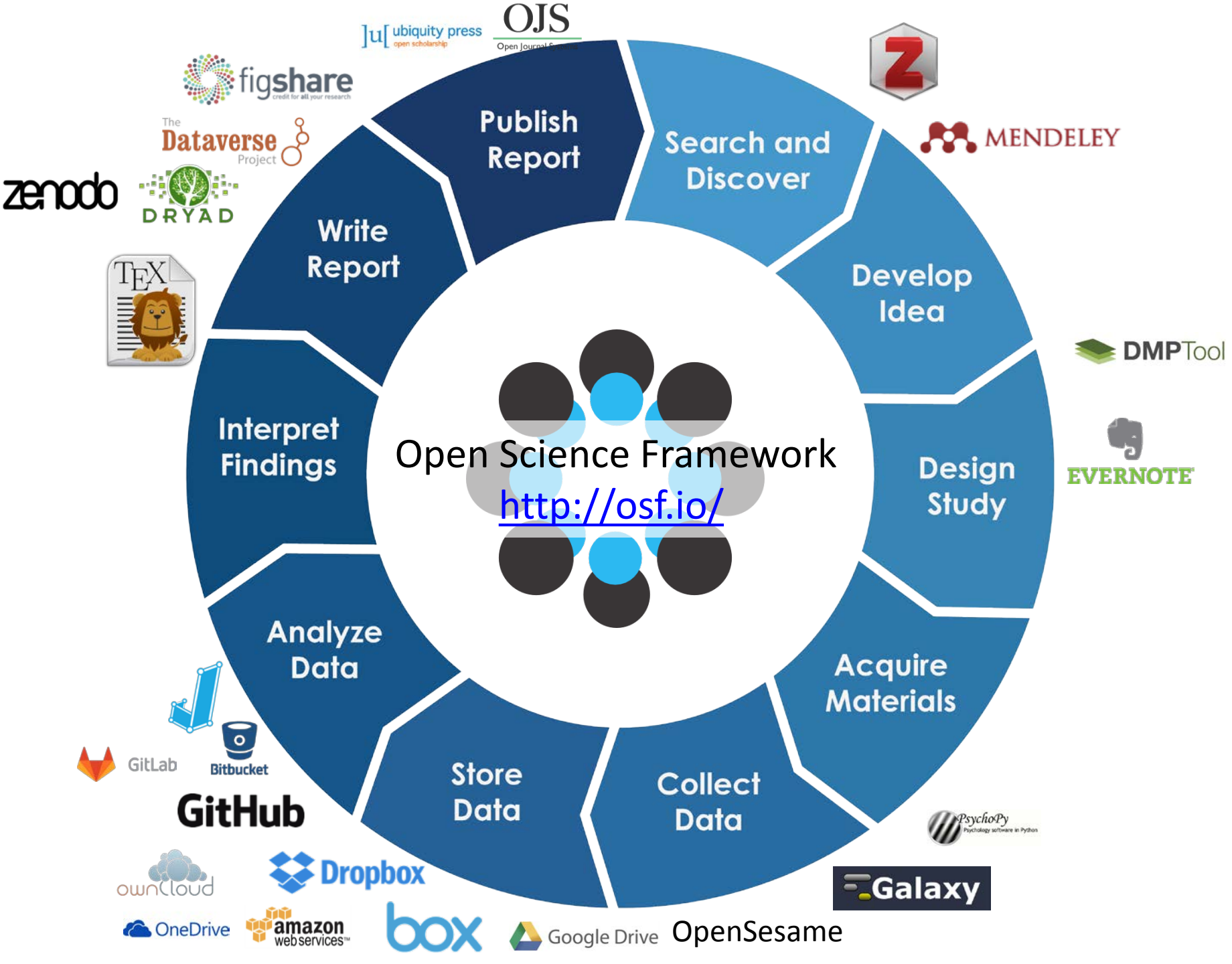
I don't have time to share

OSF: Supports research lifecycle

I am not ready to share

OSF: Integrates private-public workflows


OSFHOME
My Projects
Search
Support
Donate
Brian A.
Many Labs 2: Investigating Variation in ...
Files
Wiki
Analytics
Registrations
Forks
Contributors
Settings
Contents
Development
Call for participation: Many Labs 2 was an open project inviting researchers to participate in study design and data collection. This file is the original
Read More
Files
Click on a storage provider or drag and drop to upload
Filter
Name
Modified
Many Labs 2: Investigating Variation in Replic...
GitHub: ManyLabsOpenScience/ManyLabs...
OSF Storage
ML2_-_Protocol.pdf
2014-09-16 09:40 PM
ML2 Coordinating Proposal.pdf
2017-09-01 10:49 AM
Many Labs 2: Investigating Variation in Re...
Codebooks and Study Files | Forked: 2014-02-18 11:38 UTC
Vianello, Nosek, Ratliff & 174 more
144 contributions
Materials for Individual Studies
Klein, Grahe, Levitan & 173 more
76 contributions
Videos Documenting Data Collection | Forked: 2014-02-18 11:38 UTC
Vianello, Nosek, Ratliff & 174 more
367 contributions
Data | Forked: 2014-02-18 11:41 UTC
Vianello, Nosek, Ratliff & 175 more
123 contributions
Analysis | Forked: 2014-02-18 11:41 UTC
Vianello, Nosek, Ratliff & 174 more
183 contributions


Publish Report
Search and Discover
Develop Idea
Design Study
Acquire Materials
Collect Data
Store Data
Analyze Data
Interpret Findings
Write Report
Open Science Framework
http://osf.io/
ubiquity press
OJS
figshare
The Dataverse Project
zenodo
DRYAD
MENDELEY
DMPTool
EVERNOTE
TEX
GitLab
Bitbucket
GitHub
PsychoPy
Galaxy
ownCloud
Dropbox
OneDrive
amazon web services
box
Google Drive
OpenSesame

Part 3: Grantee Training

- COS
  - In-person training
  - Webinars
  - Web content
  - Individual consultations
- Others
  - Data Carpentry: http://www.datacarpentry.org/

DATA CARPENTRY

BUILDING COMMUNITIES TEACHING UNIVERSAL DATA LITERACY

Part 3: Training to-do

Provide training opportunities!

https://cos.io/our-services/training-services/

Part 4: Monitoring and Reporting

- Is data sharing integrated with reporting requirements?

- OSF Institutions: https://cos.io/our-products/osf-institutions/

- Example: https://osf.io/institutions/ljaf/

Projects listed below are for grants awarded by the Foundation. Please see the LJAF Guidelines for Investments in Research for more information and requirements.

All Projects ›

Filter displayed projects ×

Collections

All Projects

All Registrations

Contributors ‹ 1/27 ›

Stuart Buck

Tim Errington

Nicole Perfito

Elizabeth Iorns

Tags ‹ 1/14 ›

reproducibility

replication

metascience

Reproducibility Project: Cancer Biology

Name	Contributors	Modified
Transparency and Openness Promotion (TOP) Guidelines.	Nosek, Alter + 46	29 minutes ago
TOP 2.0 Promoting Transparency Practices and Diminishing Journal...	Mellor, Nosek + 45	17 hours ago
Maximizing Research Impact	Mellor, Nosek + 84	2 days ago
Preregistration Challenge: Plan, Test, Discover	Mellor, Esposito + 9	3 days ago
Evidence Synthesis/Systematic Reviews of Eyewitness Accuracy	Yaffe, Dodson + 4	4 days ago
Assessing the effectiveness of automatic enrollment at boosting pri...	Cribb, Emmerson	8 days ago
Study 39: Replication of Willingham et al., 2012 (PNAS)	Horrigan, Iorns + 3	11 days ago
Study 21: Replication of Sirota et al., 2011 (Science Translational Me...	Irawati Kandela, Fraser Aird + 4	12 days ago
Study 16: Replication of Ward et al., 2010 (Cancer Cell)	Showalter, Jason Hatakeyama + 8	18 days ago
Statistical Models and Methods for Analyzing Eyewitness Identificat...	Kafadar, Dodson + 4	18 days ago
Injectable Pharmacotherapy for Opioid-Use Disorder (IPOD)	Farabee	21 days ago
A U.S. CARBON TAX AND THE EARNED INCOME TAX CREDIT: An Ana...	Morris, Buck	22 days ago
Measuring the Impact of LTSS Integration on Medicare Utilization	Windh, Buck	a month ago
Reproducibility Project: Cancer Biology	Errington, Tan + 83	a month ago

Make Private | Public | 13

Reproducibility Project: Cancer Biology

Contributors: Tim Errington, Fraser Elisabeth Tan, Joelle Lomax, Nicole Perfito, Elizabeth Iorns, William Gunn, **Brian A. Nosek**, Stuart Buck, Erin Griner, Nimet Maherali, Mathew Veal, Michael McCarthy, Samuel LaBarge, Hyun Yong Jin, Christine Schaner Tooley, Claudia-Gabriela Mitrofan, Tim Smith, Robert L Judson, Matthew Cook, Sarah Statt, Nicole Vasilevsky, Stefano Biressi, Kevin Poindexter, Kartoa Chow, Heidi Hilton, Hildegard Mack, Teresa Krieger, Minyoung Anna Lim, Miguel A. S. Cavadas, Michael V. Gormally,

Affiliated Institutions: Laura and John Arnold Foundation, Center For Open Science

Date created: 2013-10-08 07:31 PM | Last Updated: 2017-08-22 01:08 PM

Create DOI / ARK

Category: Project

Description:

We are conducting a study to investigate the replicability of cancer biology studies. Selected results from a substantial number of high-profile papers in the field of cancer biology published between 2010-2012 are being replicated by the Science Exchange network.

License: Add a license

Wiki

The Reproducibility Project: Cancer Biology is a collaboration between Science Exchange and the Center for Open Science, and is independently replicating a subset of experimental results from a number of high-profile papers in the field of cancer biology published between 2010-2012 using the Science Exchange

Read More

Files

Citation

osf.io/e81xl

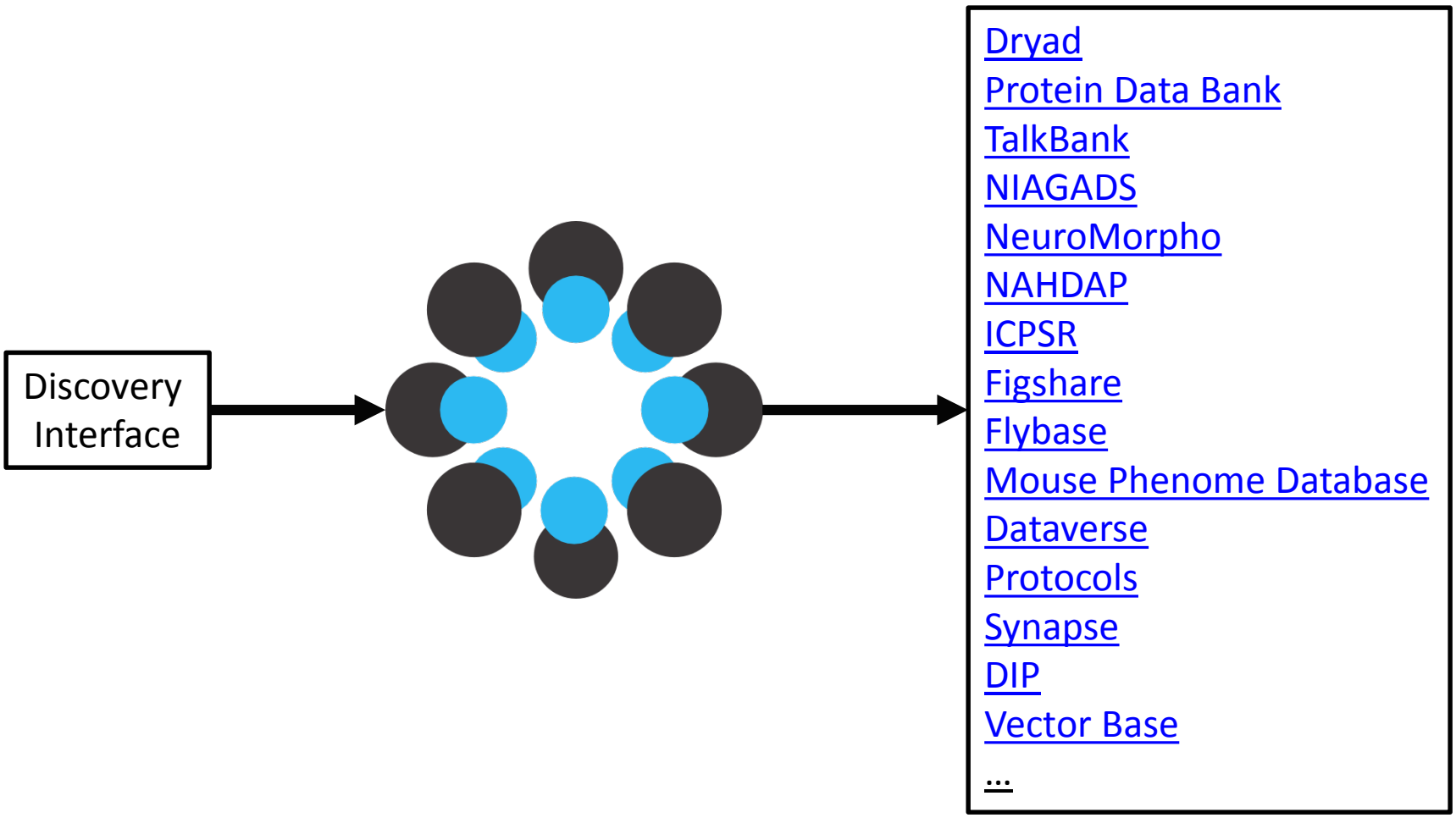
Components

Add Component | Link Projects

- **Replication Studies**
  Errington, Tan, Lomax & 82 more
  222 contributions
- **Identification Analysis of RP:CB**
  Errington, Vasilevsky & Haendel
  45 contributions
- **Data collection and publishing guidelines**
  Tan, Lomax, Errington & 3 more
  78 contributions

Data Sharing in the future

- Part 1: Repository integrations “Data Commons”


Discovery Interface
Dryad
Protein Data Bank
TalkBank
NIAGADS
NeuroMorpho
NAHDAP
ICPSR
Figshare
Flybase
Mouse Phenome Database
Dataverse
Protocols
Synapse
DIP
Vector Base
…

Data Sharing in the future

- Part 1: Repository integrations “Data Commons”
- Part 2: Integrated with grant management workflow
- Goal: Make data sharing natural and easy

COS: http://cos.io/

OSF: http://osf.io/

TOP: http://cos.io/top/

Training: https://cos.io/our-services/training-services/

These slides: https://osf.io/yec47/

HEALTH RESEARCH ALLIANCE – BIG DATA MEETING BRAIN COMMONS

Magali Haas, MD, PhD

CEO & President

September 19, 2017 | 8:30 AM – 3:30 PM CT

Agenda

01.
About CVB
Mission/Programs
Systems Modeling

02.
The Quest
Goals
Needs Assessment
Landscape

03.
Lessons Learned
Scale-ability & Compute
Data Standards
Governance
Cost & Sustainability

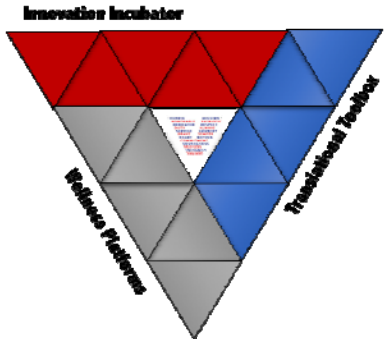
04.
BRAIN Commons
Description
Data Contributors
Data Users

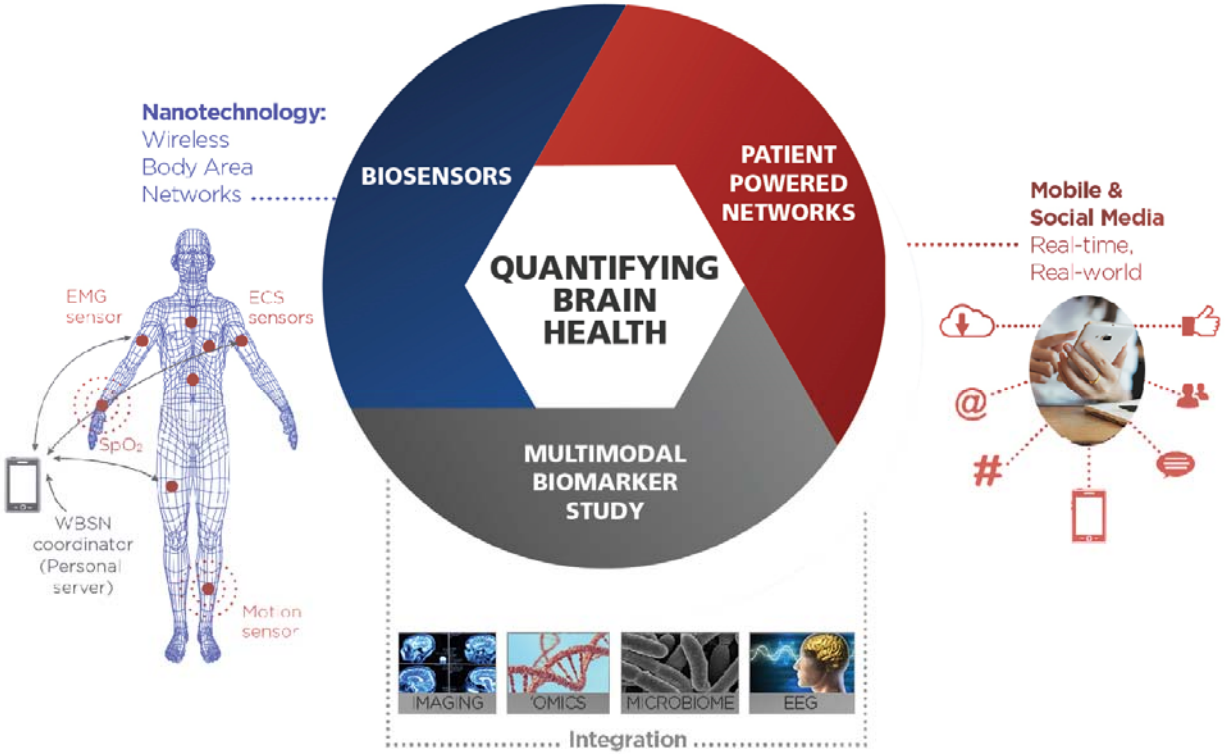
About Cohen Veterans Bioscience

We are a national, nonpartisan 501c3 research organization
dedicated to fast-tracking the development of diagnostic tests and personalized therapeutics
for the millions of veterans and civilians who suffer the devastating effects of trauma-related and other brain disorders.

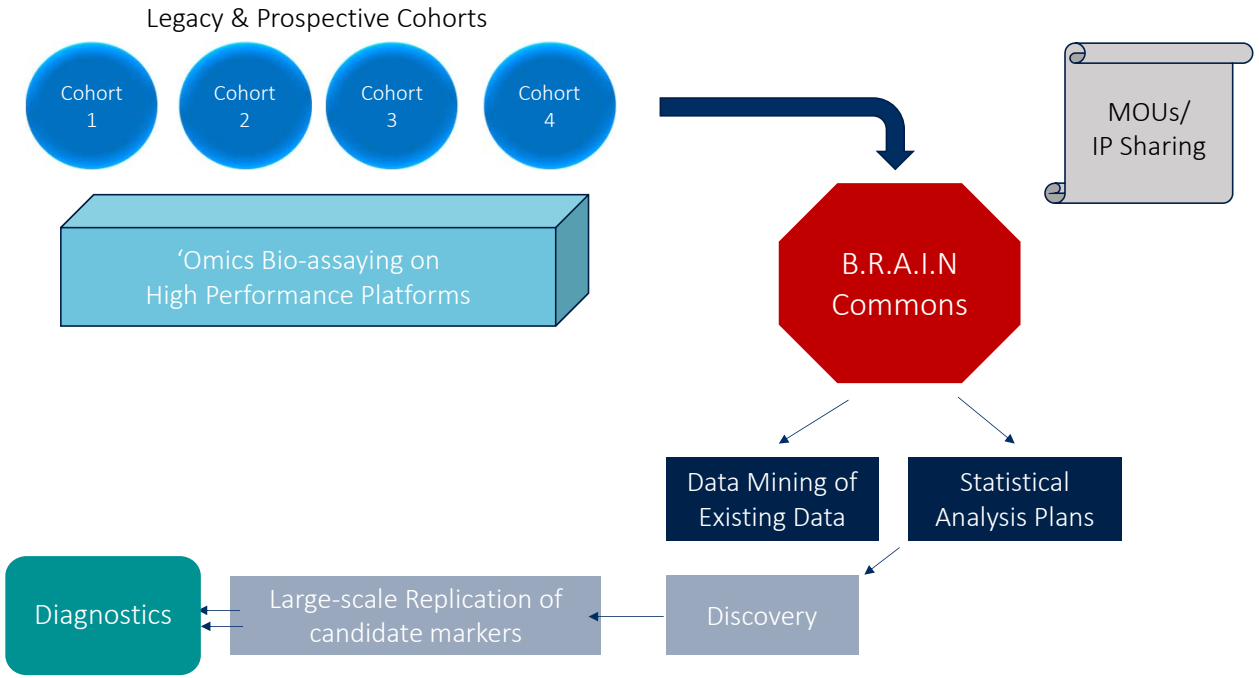
The Ultimate Goal is Prevention & Personalized Medicine for Brain Health

To realize this vision,
we have spearheaded a brain health roadmap
that capitalizes on new technologies and innovative approaches
to foster Wellness & catalyze
Precision Biomarker, Diagnostics and Therapeutics

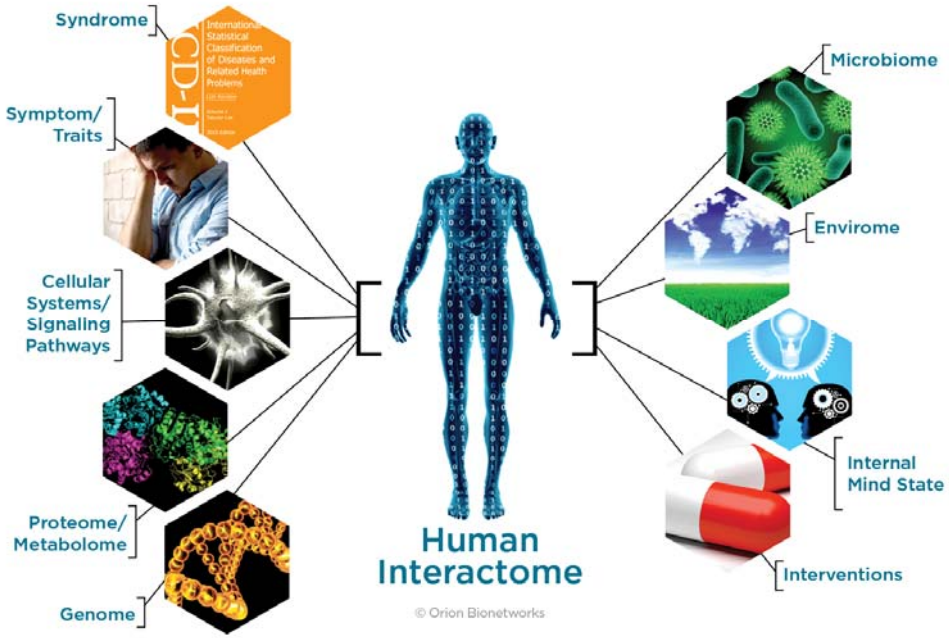



Nanotechnology:
Wireless
Body Area
Networks
EMG sensor
ECS sensors
SpO2
WBSN coordinator (Personal server)
Motion sensor
BIOSENSORS
PATIENT POWERED NETWORKS
QUANTIFYING BRAIN HEALTH
MULTIMODAL BIOMARKER STUDY
Mobile & Social Media
Real-time, Real-world
IMAGING
OMICS
MICROBIOME
EEG
Integration

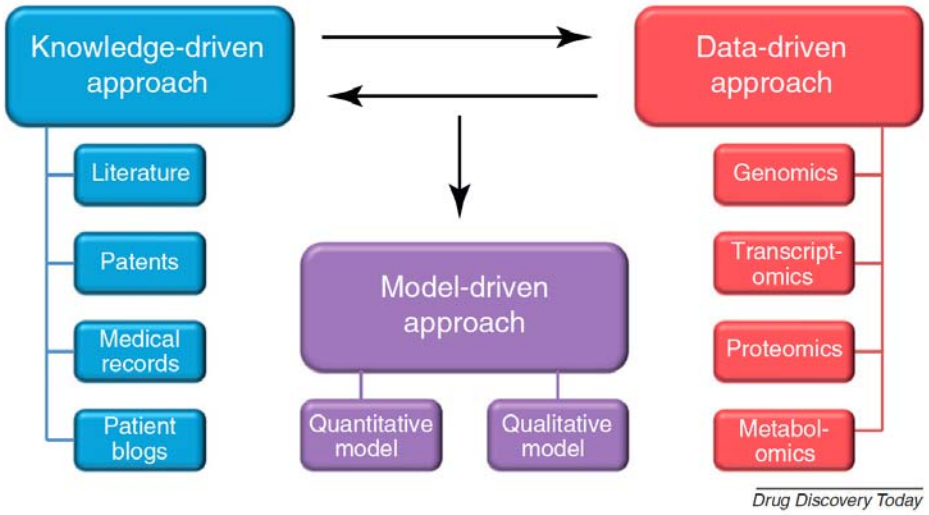
RAPID-Dx Framework



Systems Modeling: The Basis for Understanding the Disease



The Challenge: "Heterogeneity of Data and Knowledge …"



WITH PERMISSION FROM:
Deyati, A., Younesi, E., Hofmann-Apitius, M., & Novac, N. (2013). **Challenges and opportunities for oncology biomarker discovery.** Drug discovery today, 18(13), 614-624.

Agenda

01.

About CVB

Mission/Programs
Systems Modeling

02.

The Quest

Goals
Needs Assessment
Landscape

03.

Lessons Learned

Scale-ability & Compute
Data Standards
Governance
Cost & Sustainability

04.

BRAIN Commons

Description
Data Contributors
Data Users

Critical Resources


SCIENCE AS A SERVICE

A Case for Data Commons:
Toward Data Science as a Service

Robert L. Grossman, Allison Heath, Mark Murphy, and Maria Patterson | University of Chicago
Matt Wells | Center for Computational Science Research

"Tom"

Thomas Oberst
Chief Technology Officer

- 20+ y Computer Industry R&D
- 15+ y Financial Services - Systems Development & Enterprise Architecture
- 7+ years CTO - Bioscience and Healthcare Consulting
- Data-driven discovery: mining emerging technology - matching to mission

CVB's Criteria for Selecting a Data Commons

45 Requirements				
Volume	Variety	Velocity	Veracity	Value
Availability	Durability	Redundancy	Recoverability	Scalability
Computational Capability (3)	Computational Configurability	Extensibility & Adaptability	Interoperability	Searchability & Retrievability
Seamless Integration	Accessibility	Usability	Import-Export	Storing Capability
Affordability	Migratability	Workflow Controllability	Support Capability	Environment Longevity
Retainability	Sustainability	Data Reproducibility	Organizational Survivability	Protectability
Privacy	Cyber Security	Shareability	Transparency	Reproducibility
Compatibility	Global Compliance	Accounting and Auditability	Flexibility & Elasticity	Analytics Visualization
Domain Focus	Open Source	Geographic Diversity		

Comparison of 55+ Available Platforms

Platforms and Software Technologies				
#1. NCI Genomic Data Commons (GDC)	#2. tranSMART Knowledge Management Platform	#3. Informatics for Integrating Biology and the Bedside (i2b2)	#4. Ontario Brain Institute (Brain-CODE)	#5. EU EPILEPSIAE Database
#6. IEEG.org – International Epilepsy Electrophysiology	#7. NSF Cloud Platforms - Computing in the Cloud	#8. NIMH Data Archive - National Institute of Mental Health	#9. MIT "SuperCloud"	#10. HPI Hasso Pattner Institute - Univ of Potsdam
#11. EMC – Pivotal - Large Scale Hadoop Testbed	#12. Perkin Elmer – "Signals"	#13. PMI (Precision Medicine Initiative) New York Genome Center + IBM	#14. The Open Cloud Consortium – Open Science Data Cloud	#15. CG HUB from The Cancer Genome Atlas (TCGA)
#16. Cancer Genome Collaboratory - (Canada)	#17. Blackflynn	#18. "Genome Bridge" – The Broad	#19. IBM Watson Health & IBM Watson Health Cloud	#20. MVP - Million Veterans Program (GenISIS)
#21. Intel PCCSB - Intel Parallel Computing Center Structural Biology	#22. Collaborative Cancer Cloud - Intel	#23. LONI Laboratory of Neuro Imaging - IDA Image and Data Archive (USC)	#24. European Open Science Cloud	#25. ICGC Data Portal

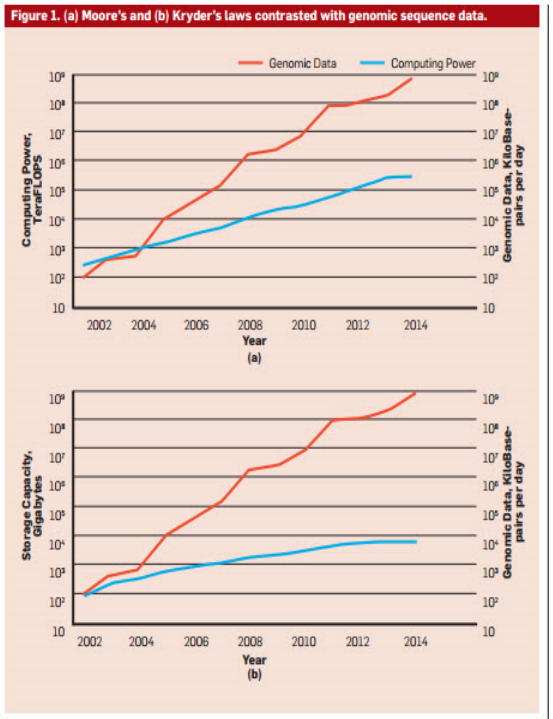
Comparison of 55+ Available Platforms

Platforms and Software Technologies				
#26. LORIS Longitudinal Online Research & Imaging System (Canada)	#27. Frederick National Laboratory FFRDC SysBioCube	#28. DNAnexus Cloud Based Platform	#29. NCBI National Center for Biotechnology Information	#30. GENISIS Project – Cloud Based
#31. cBio Cancer Genomics Portal	#32. Sage Bionetworks - Synapse	#33. Palantir	#34. BioStorage Technologies	#35. BC Platforms – Federated DB
#36. DART –American College of Radiology	#37. Sentinel	#38. XNAT	#39. BioMart	#40. Ensembl Project
#41. REDCap	#42. INCF	#43. NeuroVault	#44. Shanoir Data Management	#45. COINS
#46. NITR	#47. Vivli	#48. Google	#49. Facebook	#50. NIDB Neuro informatics Database
#51. metaCell Analytics	#52. GIFT - Cloud	#53. SciDB	#54. DatStat	#55. HID Human Imaging Database
#56. Dataverse	#57 PopMedNet			

Recommendation: OCC

Brain Commons - University of Chicago

- ✓Petabyte Scale
- ✓Extensibility & Adaptability
- ✓Variety
- ✓Scalability
- ✓Interoperability
- ✓Availability
- ✓Anonymity/Privacy/Security
- ✓Compliance
- ✓Protectability
- ✓Accessibility
- ✓Flexibility & Elasticity
- ✓Retainable and Sustainable
- ✓Pipelines and Workflows
- ✓Domain Focus



Agenda

01.

About CVB

Mission/Programs
Systems Modeling

02.

The Quest

Goals
Needs Assessment
Landscape

03.

Lessons Learned

Scale-ability & Compute
Data Standards
Governance
Cost & Sustainability

04.

BRAIN Commons

Description
Data Contributors
Data Users

Lessons Learned

Data Standards

Data-sharing Incentives

Compute in the Cloud

Sustainable Funding Model

Governance

Compliance/HIPPA

User-Friendly

Agenda

01.

About CVB

Mission/Programs
Systems Modeling

02.

The Quest

Goals
Needs Assessment
Landscape

03.

Lessons Learned

Scale-ability & Compute
Data Standards
Governance
Cost & Sustainability

04.

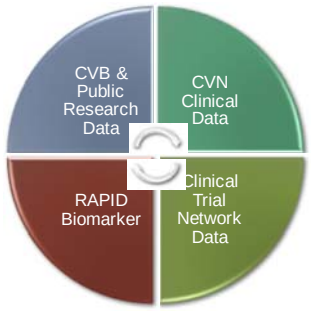
BRAIN Commons

Description
Data Contributors
Data Users

What is the BRAIN Commons

- The BRAIN Commons is a scalable, centralized big data cloud-based platform for computational innovation and data driven discovery
- By partnering with OCC we are leveraging Open Source data models & continued investments in the community platform
- Integrates individual level data across data types (genomics, biomics, imaging, wearables, etc.)
- Able to scale to work with large quantities of data
- Equipped with data-analysis and systems biology tools
- FISMA Moderate Security & HIPAA Compliant

BRAIN Commons



Multiple Sources of Data

Apply CDISC Data Standards

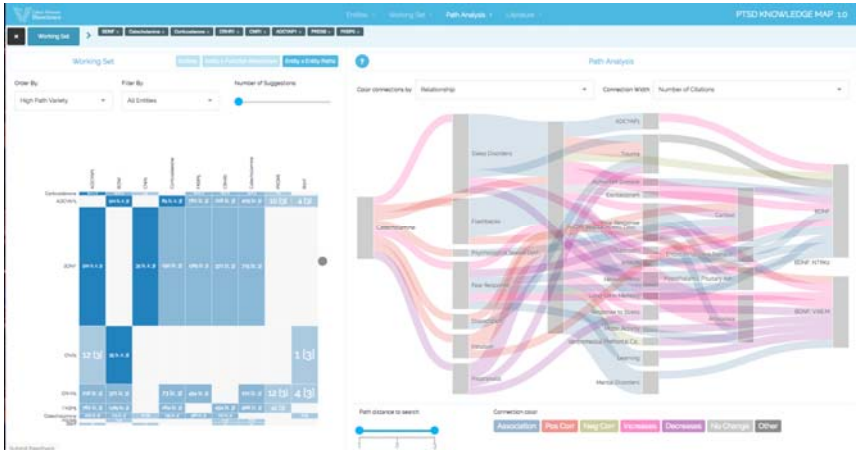
Cloud-Architecture
Big Data Capacity

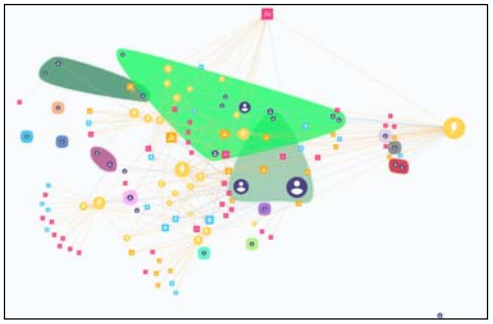
Data Partners Control Access to their data

Visualization & Analytics Built-In

Discovery Accelerated

Cognitive City & KnowledgeMap™





We are creating a collaborative network that is able to suggest tools, datasets, algorithms and even collaborations based on objective measures and usage of the assets that exist in the network. This cognitive city will grow and allow our partners and collaborators to learn from each others.

Why a BRAIN Commons?

The BRAIN Commons will:

1. Enable Data-Driven Accelerated Discovery through a Unified Data Repository.
2. Simplify data sharing hurdles and provide easy-to-use visual, dynamic data tools to spark innovation.
3. Be uniquely positioned to tackle traditional big data challenges such as capture, pedabyte storage, data curation, transfer, search, sharing, data-mining, security and information privacy.
4. Enable the combining, interpreting and analyzing of vast and disparate data types, including imaging, genomic & biomic, wearable and sensor, and clinical data, from different sources with sophisticated visualization and analytics tools.
5. Realize the potential of machine learning and predictive modeling.
6. Safeguard and Protect Data Integrity and Access.
7. Promote Collaboration across the Multidisciplinary Research Community.

A True Commons

We invite other brain-related disease organizations to partner with us!

- ✓ Grow the platform
- ✓ Share 'commons charges' for sustainability
- ✓ Leverage government & CVB investment
- ✓ Build a brain community
- ✓ Integrate knowledge across "diseases"

Cohen Veterans Bioscience

THANK YOU

www.cohenveteransbioscience.org

It's Not Enough To Share: The Funder's Responsibility

Kenna R. Mills Shaw, Ph.D.

Executive Director

MD Anderson Cancer Center

Khalifa Institute for Personalized Cancer Therapy

September 19, 2017

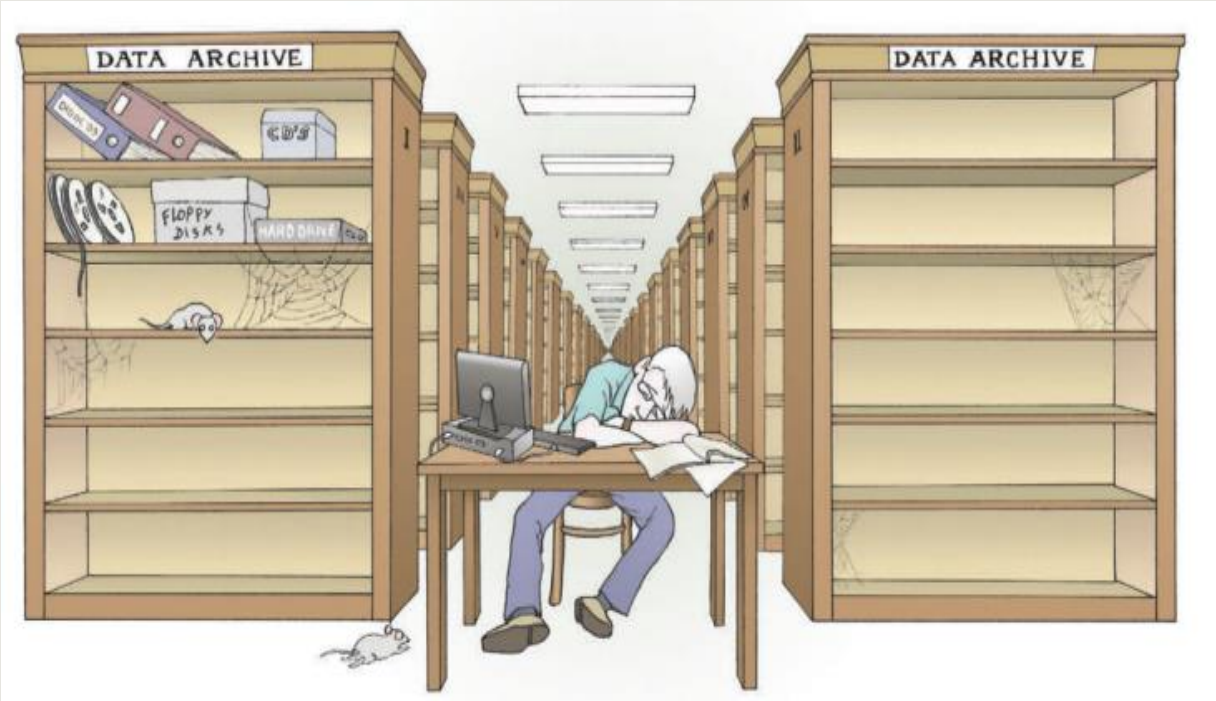
krshaw@mdanderson.org

THE UNIVERSITY OF TEXAS
MD Anderson Cancer Center
Making Cancer History®

Disclosures

- I have no relationships to disclose.
- The thoughts and opinions presented in this talk are my own.

Just Because You Build It, Doesn't Mean They Will Come



Art from http://www.nature.com/news/2009/090909/pdf/461160a.pdf

A warm and fuzzy, "beautiful" ideal that resonates with many scientists… data sharing is often considered, at best an un(der)funded mandate… at worst food for the scientific "parasite."

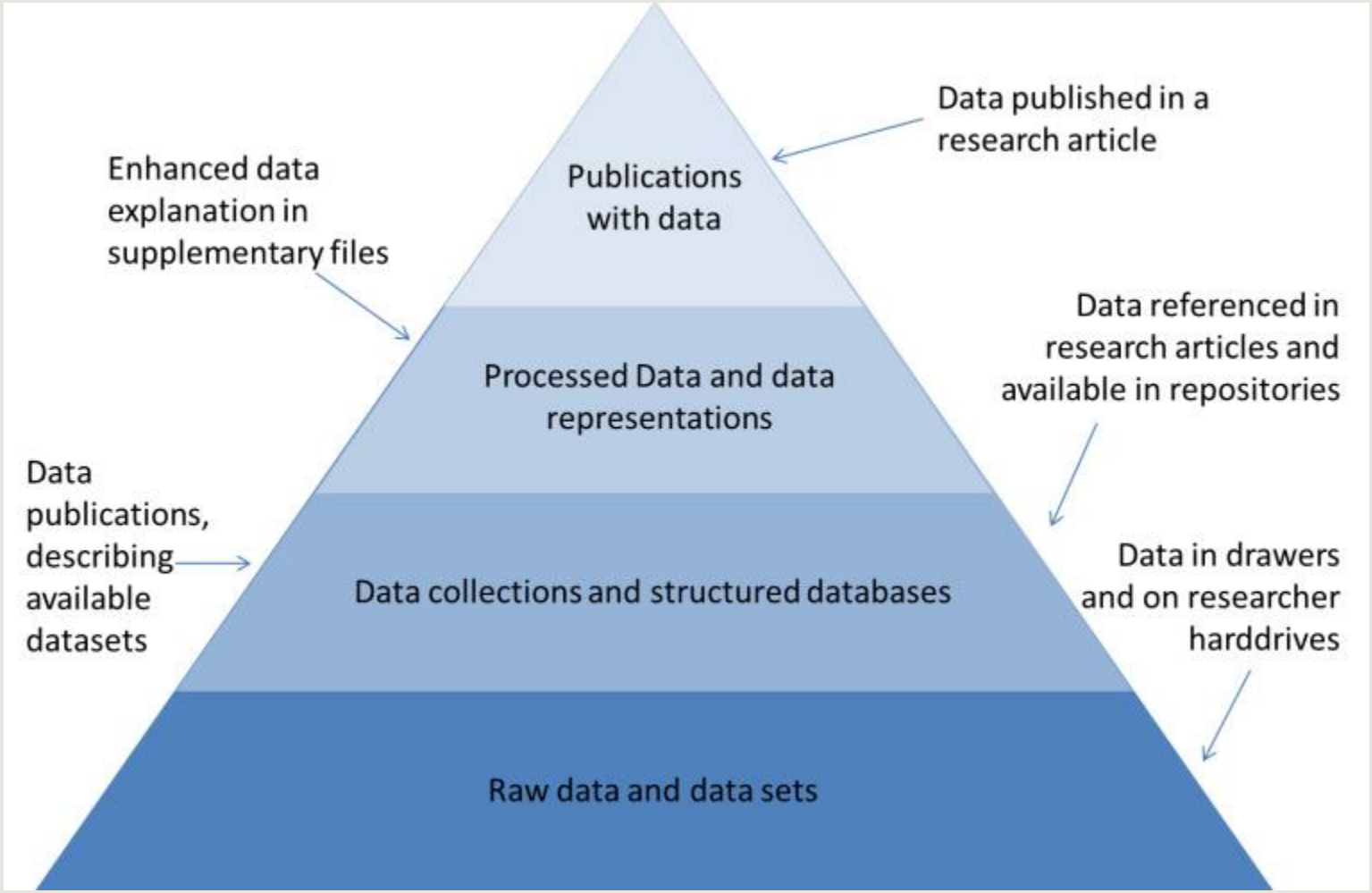
The Research Parasite is Real and Productive

Reanalyzing your data.
Disproving what you posited.
Stealing ideas you haven't yet had.

Scientists Share & Are Incentivized to Share ALL the Time

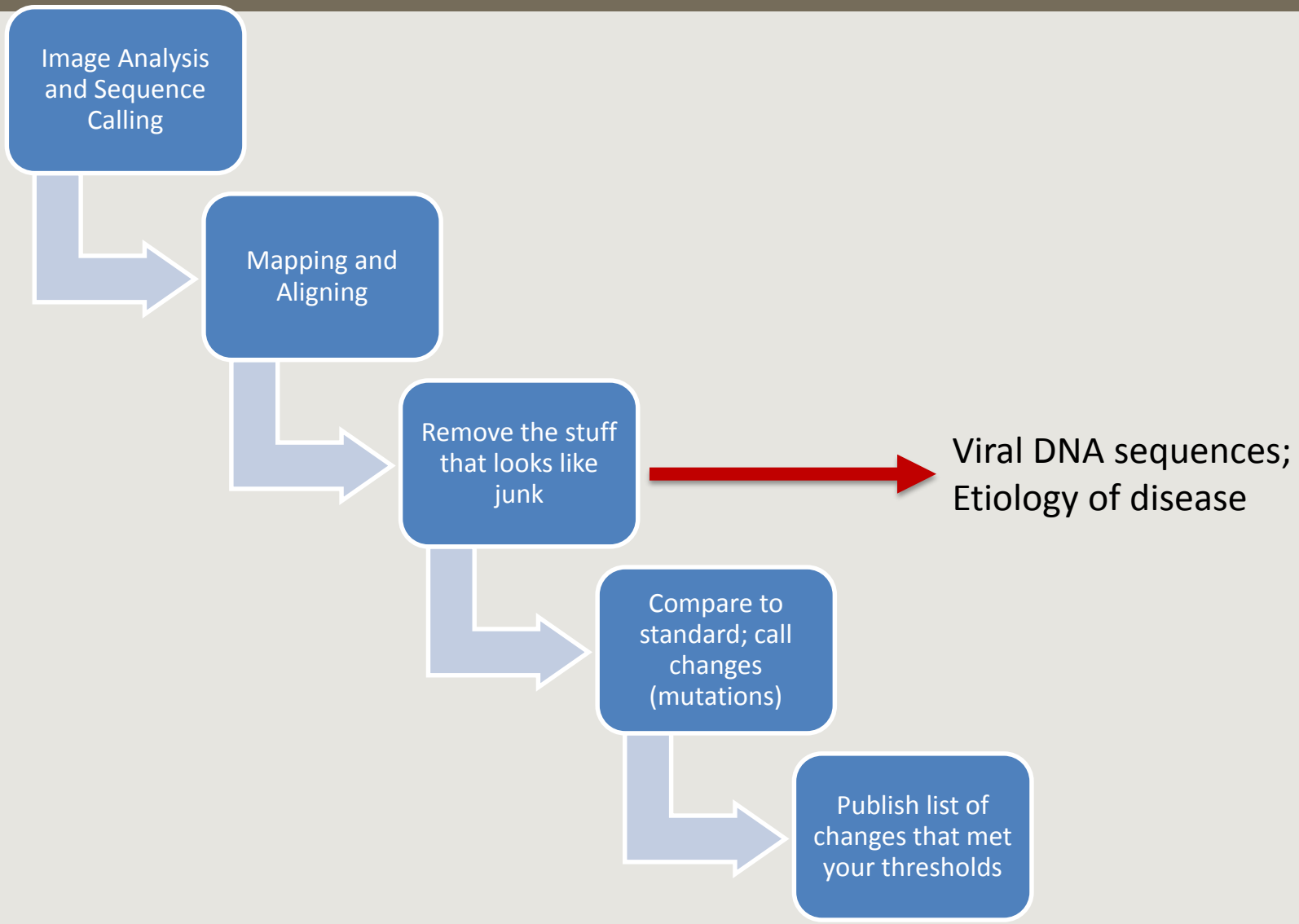
- Processed, analyzed data provided as an aggregate snapshot
- Single method, single point in time

What Opportunities Arise From Sharing More than the Tip of the Iceberg?



https://www.elsevier.com/connect/can-data-be-peer-reviewed

Do We Really NEED Raw Data?

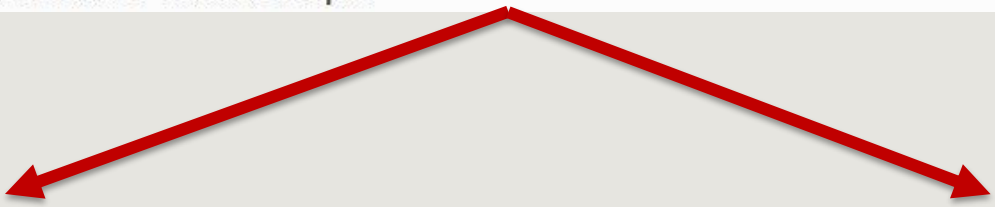


There's Available and There's Available

Availability of data, material and methods

An inherent principle of publication is that others should be able to replicate and build upon the authors' published claims. A condition of publication in a Nature journal is that **authors are required to make materials, data, code, and associated protocols promptly available to readers without undue qualifications**. Any restrictions on the availability of materials or information must be disclosed to the editors at the time of submission. Any restrictions must **also** be disclosed in the submitted manuscript.

(www.nature.com)



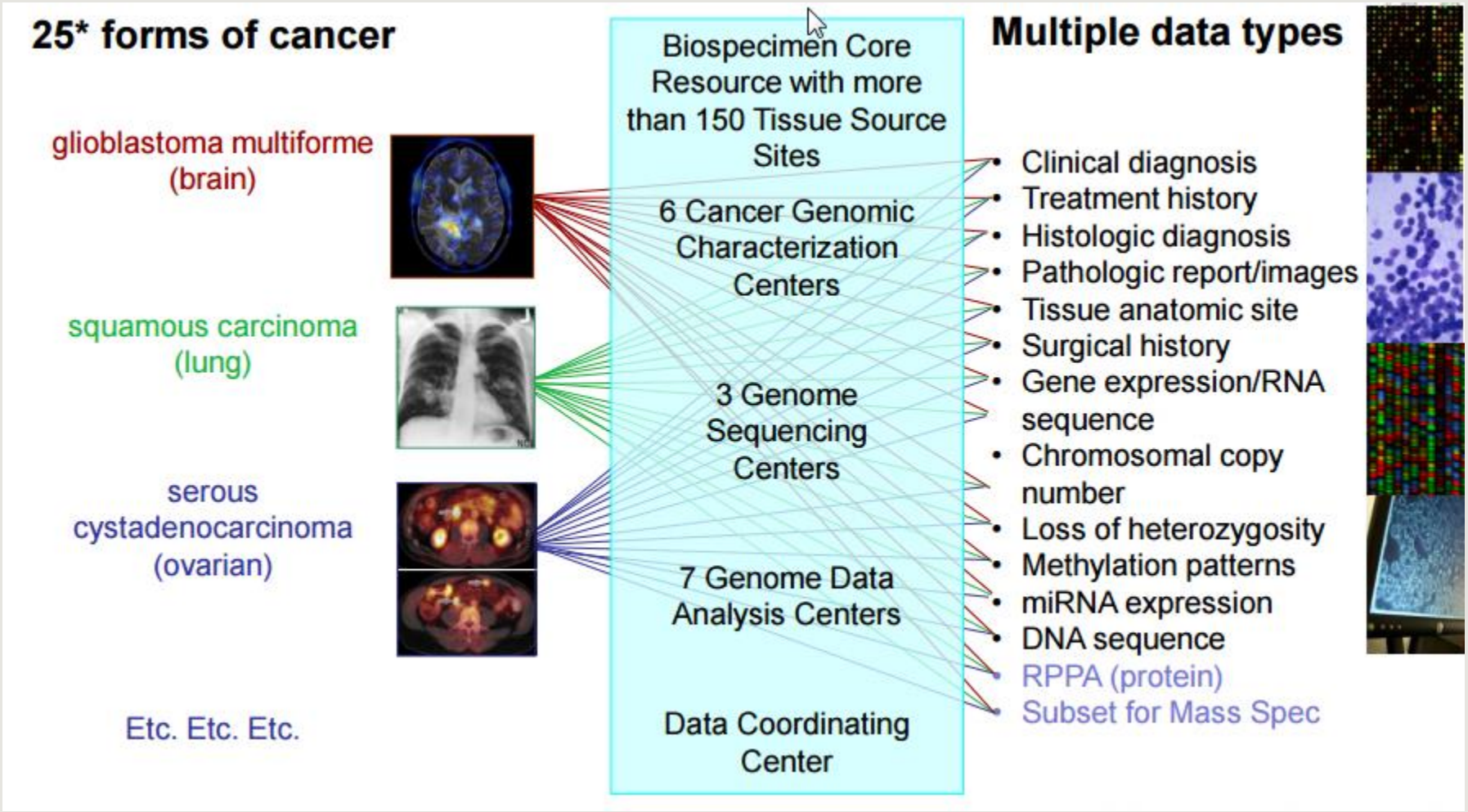
Data is available from authors upon request.

Data <u>not</u> available.

Data is available from accessible environment designed to store and share study data.

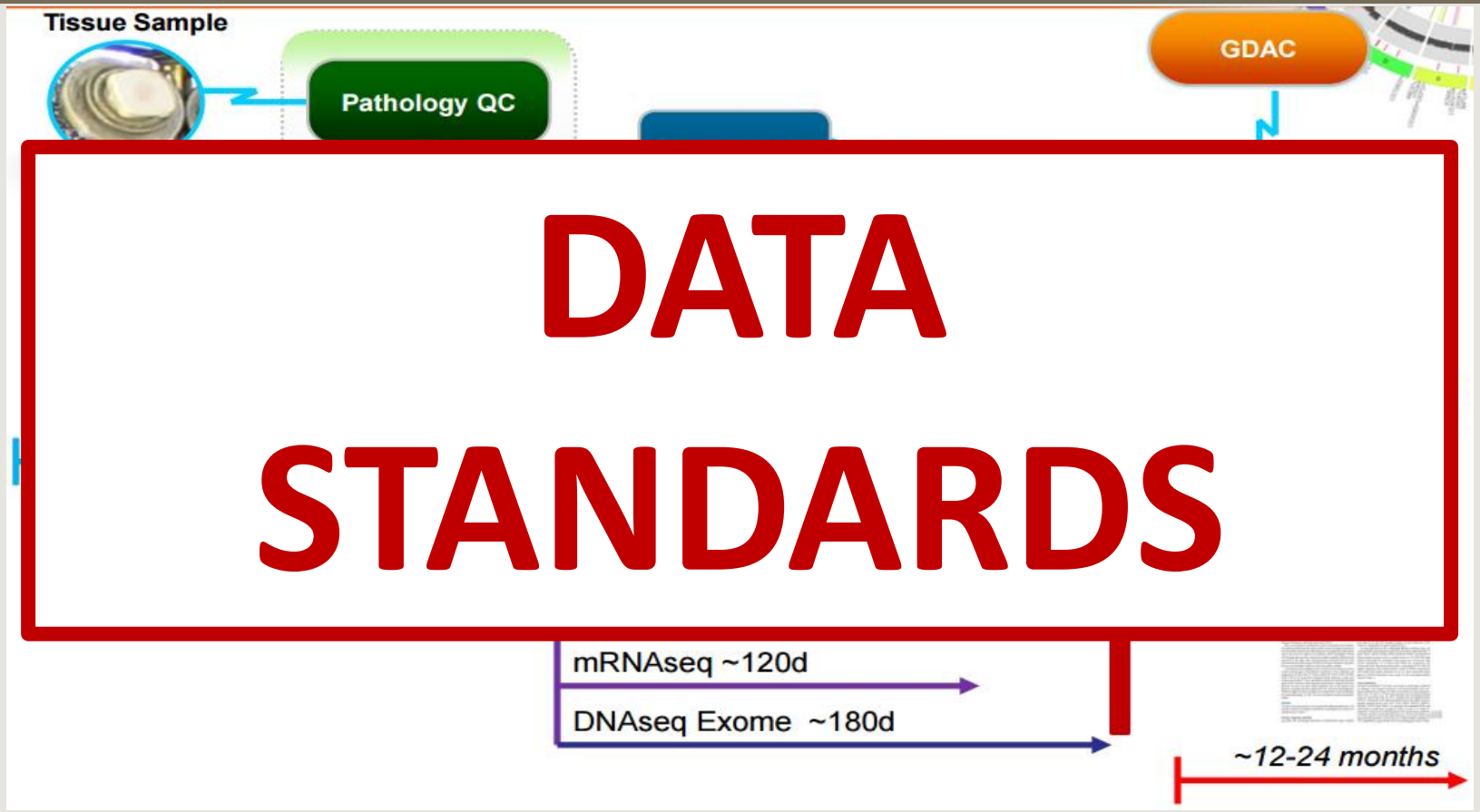
Data available. Hopefully it's usable.

The Cancer Genome Atlas:
Cancer Genomics as Example Domain for Data Sharing



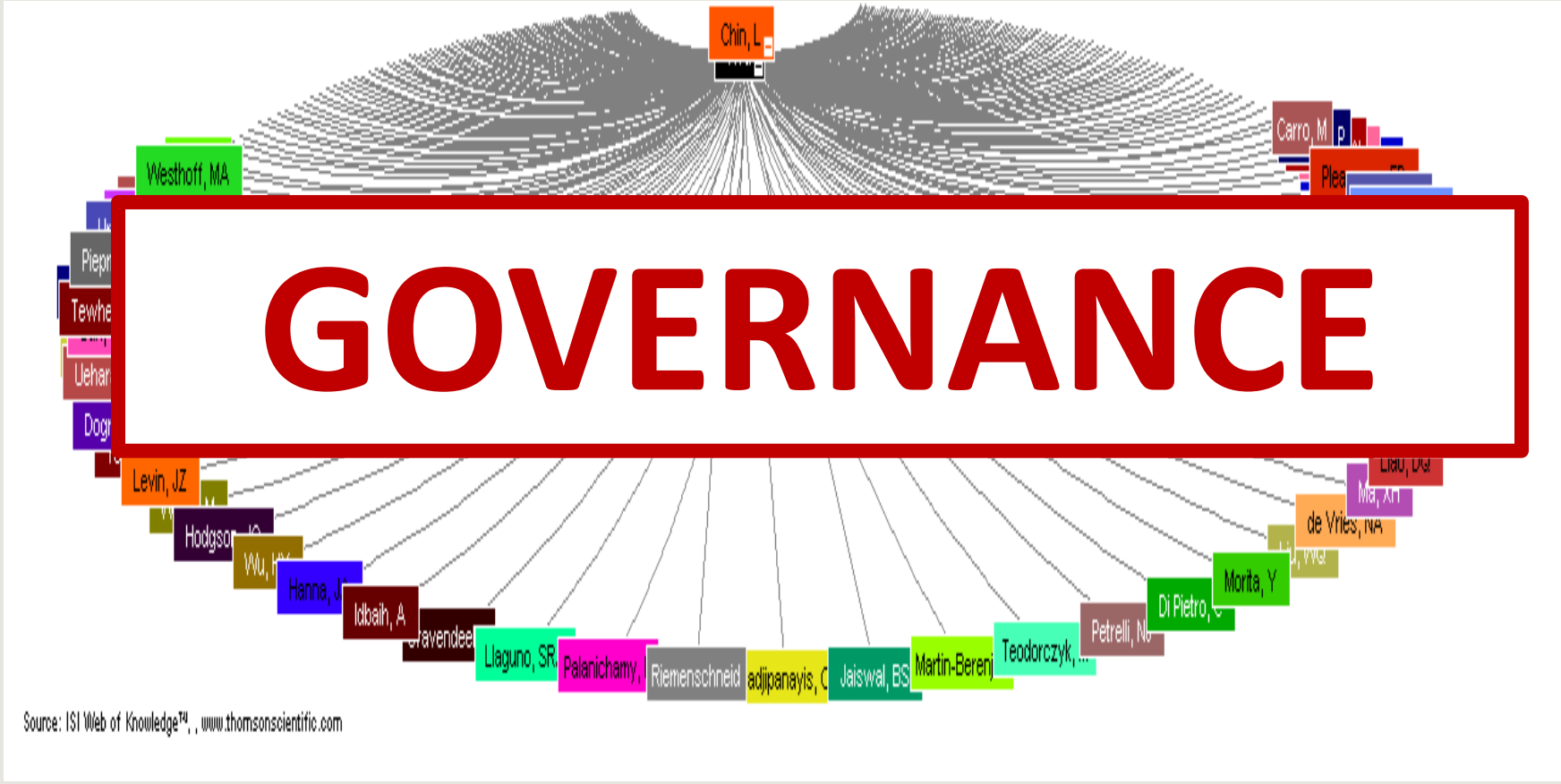
- Goal of the project was to **create a reference data set**

The Data Sharing Life Cycle Adds Time & Cost



- Planning for and actual data sharing requires data standards, robust QC, systematic updates & ability to version (example from The Cancer Genome Atlas)
- Data standards give data a format that is integratable, reusable
- In their data sharing plan do they even mention what standards they will use?

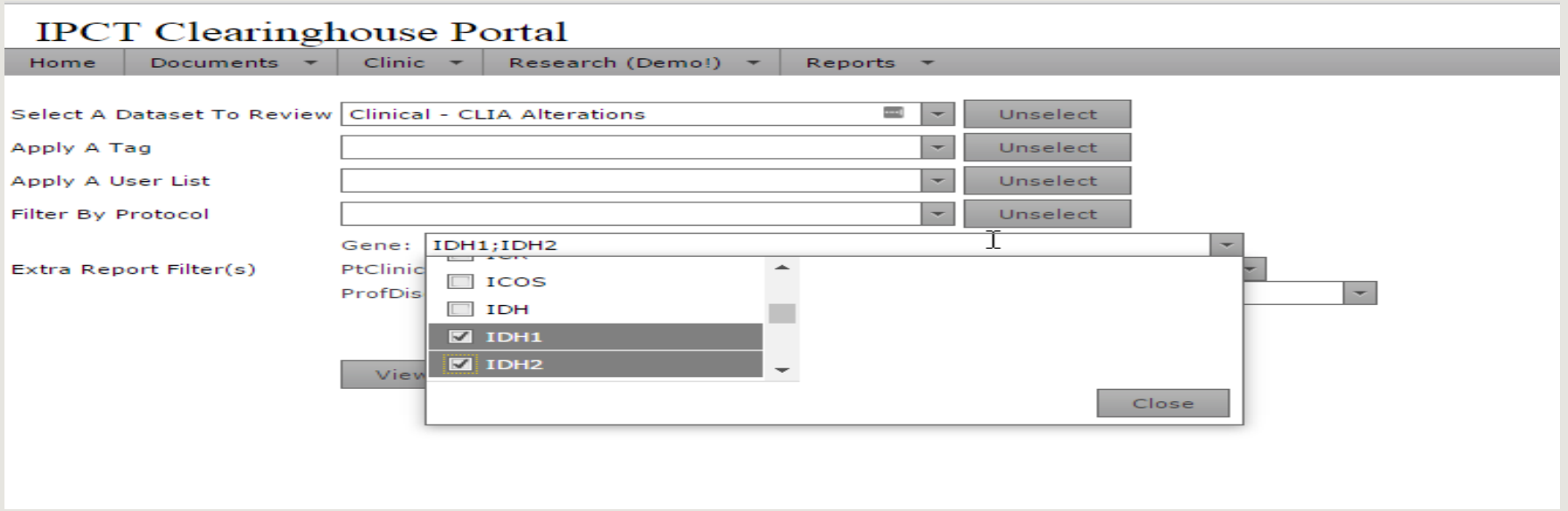
Making an Exhaustible Resource Inexhaustible



- Governance policies make it possible to find and reuse the data
- Unfortunately scientists are often measured by the number of index articles and the impact of the journal, not the number of people who reuse their data

Data Sharing is NOT ‘One-Size Fits All’

- EXISTING database that connects >25,000 patient records with CLIA sequencing data with basic clinical data, research data, trial enrollments, manually aggregated data
- Database supports >300 users across 12 clinics; provides real time clinical mutation frequency data to facilitate feasibility discussions
- Clinicians are often singularly focused, not asking open-ended questions so it’s important to not give them open ended data; Data scientists/bioinformaticists often the opposite



Different Stakeholders Get Different Views of Data

- Data provided in excel/CSV formats for easy sharing with clinicians

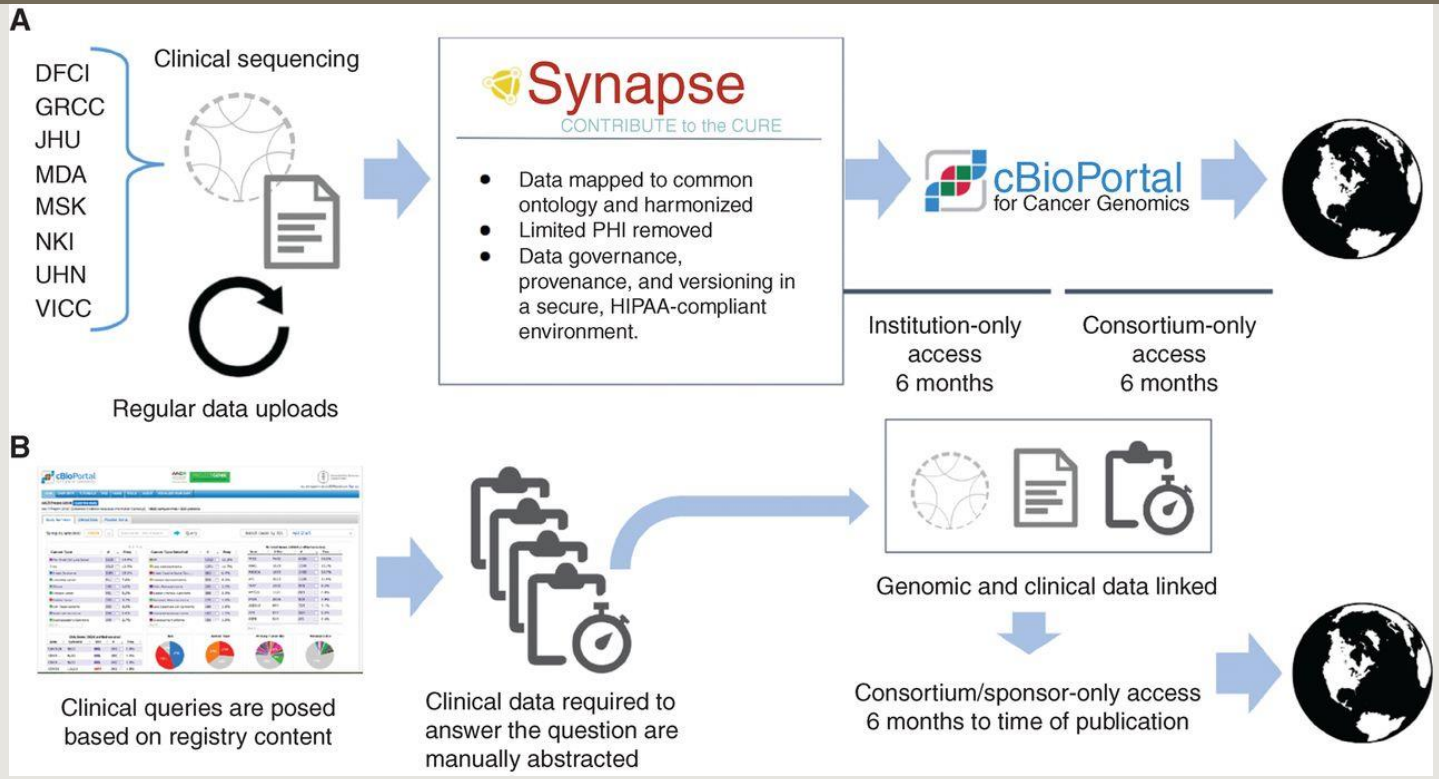
- 735 patients with IDH1 or IDH2 and other alterations provided same day

	A	B	C	D	E	F	G	H
1	IPCT_No	PT_DECEASED	Cancer Type	Gene	Panel	Mutation Type	MU_Date	Alteration
242	22332	N	Glioma	IDH1	CMS50	Somatic	5/12/2015	IDH1_R132H
243	22601	Y	Glioma	IDH1	CMS50	Somatic	5/11/2015	IDH1_R132C
244	19106	Y	Parathyroid	IDH1	CMS50	Somatic	3/31/2014	IDH1_R132H
245	19163	N	Melanoma	IDH1	CMS50	Somatic	4/1/2014	IDH1_R132C
246	22408	Y	Parathyroid	IDH	CMS50	Somatic	3/20/2015	IDH1_R132C
247	21290	N	Parathyroid	IDH1	CMS50	Somatic	8/16/2014	IDH1_R132L
248	21447	Y	Sarcoma	IDH1	CMS50	Somatic	6/13/2014	IDH1_R132C
249	21763	Y	Melanoma	IDH1	CMS50	Somatic	3/16/2015	IDH1_R132C
250	22880	N	Glioma	IDH1	CMS50	Somatic	7/7/2015	IDH1_R132H

- Facilitates real-time ability to determine frequency, ***trial feasibility***
- Identification of patient populations for novel trials
- Pro-active clinical trial alert infrastructure

- Not a good method/mechanism for sharing for to enable *discovery*

Data Sharing of Shared Data



- Clinical (CLIA) genomic data available on >18K patients; minimal clinical data
- 18K records mostly immediately interoperable with other datasets; but not completely
- Clinical data present maps to different lexicon than TCGA, NCI- dictionary, SEER, etc.
- NOT a single FUNDER in my presence*** ever asked that question
- Required overarching data commons to remap data; feasible but likely unfunded

Data Sharing 'Requirements' are Meaningless without Standards, Governance & Monitoring

NIH **(funders)** *could* be a driver for the sharing of clinical trial data by making it a requirement in the grant approval process and funding stipulations, including funding annual increments. Currently, NIH requires grantees to have a plan for data sharing if they request direct costs of $500,000 or more in any budget year, but it does not require data sharing, monitor whether data are shared as planned, ~~or~~ expressly allow a line item for expenses due to data sharing activities (NIH, 2003), or mandate usage of existing community data standards.

https://www.nap.edu/read/18998/chapter/5#60; emphasis, text in red my own

Funder's Responsibility Goes Beyond Writing the Check

How do you measure the success of your grantees?
The value of your investment?

BEFORE you require a data sharing statement in your applications, have you answered the following questions?

- What model of governance of the data is appropriate? Where will they submit? What data model/standards will they use? Does that make sense for your stakeholders, donors?

- Does your grantee/applicant have the appropriate IRB-approved protocol in place to facilitate sharing of data the way YOU envision? If not, are you willing to support reconsenting?

- Do you have anyone at your organization that can validate data exists, the quality is sufficient? Or any way to measure access or use of the data by others or value to other researchers?

Which Kind of Data Do You Expect your Grantees to Share?

https://www.elsevier.com/about/open-science/research-data

Data Sharing Mandates Only As Strong as The Funder's Will to Commit Short & Long Term

- **Know the culture of your community**. Is data sharing already the norm? If not, you might need to reset your expectations re: the level of sharing that is reasonable.

- When you fund a program and "require" data sharing- **ensure the act of data sharing is funded**. Do they need to reformat their data to share it? Do they have the right people on their team that know how to systematically share data (not via FedEx on a thumb drive)

- How many funders **employ** an expert (FTE or contract) to make sure the data you fund are useful? **Need to have someone that can help you know if your goals were met.**

- Do you have multi-year programs? **Make data sharing a requirement for renewal.**

- How committed are you? Do you support the generation of standards for the data in your domain? Do you support the groups that store, forward migrate, share with others? **Develop support for data parasitism, Community standards & Data Commons**

The Cancer Genome Collaboratory

Vincent Ferretti, PhD
Director, Genome Informatics
Ontario Institute for Cancer Research

Chicago
Sept 19th, 2017

International Cancer Genome Consortium

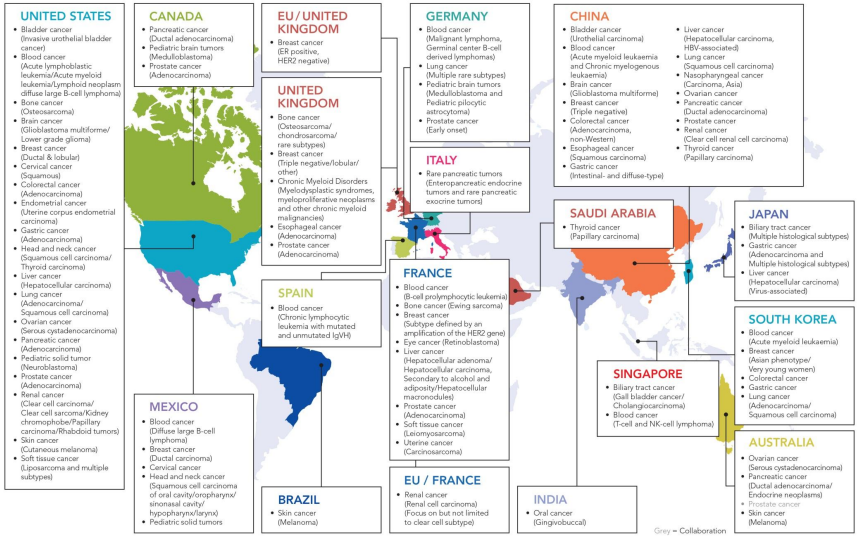
Vol 464|15 April 2010|doi:10.1038/nature08987 nature

Nature (2010) PERSPECTIVES

International network of cancer genome projects

The International Cancer Genome Consortium*

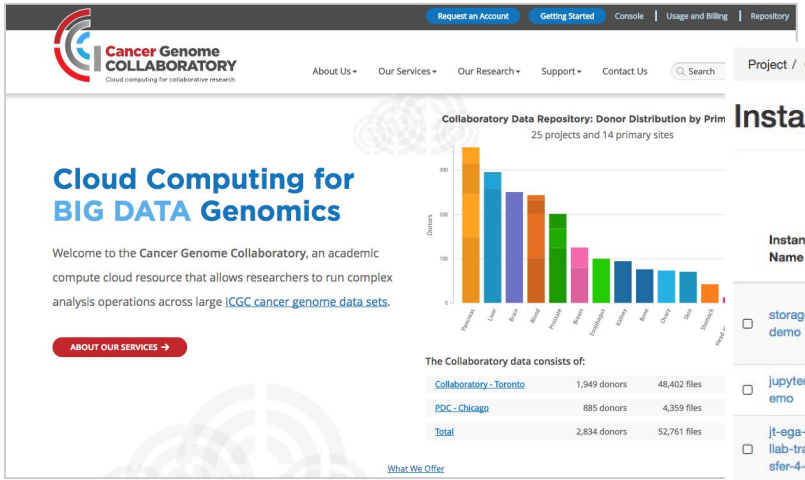
- 107 projects, 17 countries
- Goals: Sequence and analyse **25,000** tumor genomes (with matched normals) across **50** tumor types and **share** data.
- ICGC Data Coordination Center hosted at OICR
  - ICGC Data Common: Data submission, Validation & Annotation, Discovery and Cloud Compute infrastructures
  - Big data, scalable technologies

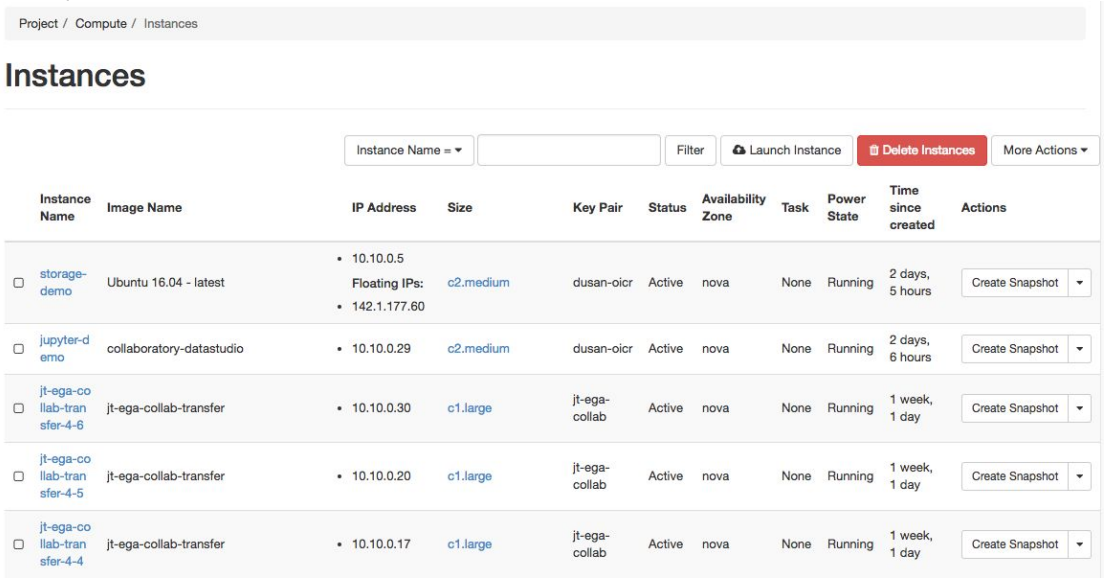


Cancer Genome COLLABORATORY
Cloud computing for collaborative research

A Compute Cloud Resource for ICGC Data

- Accessible to ICGC DACO-approved users
- Self-service infrastructure hosted at Compute Canada, Toronto
  - OpenStack (compute) and Ceph (storage)
  - 2600 CPUs, 7.6 PB raw storage
  - High-performance networking, compute collocated with data
- Data
  - PCAWG harmonized dataset and other ICGC datasets



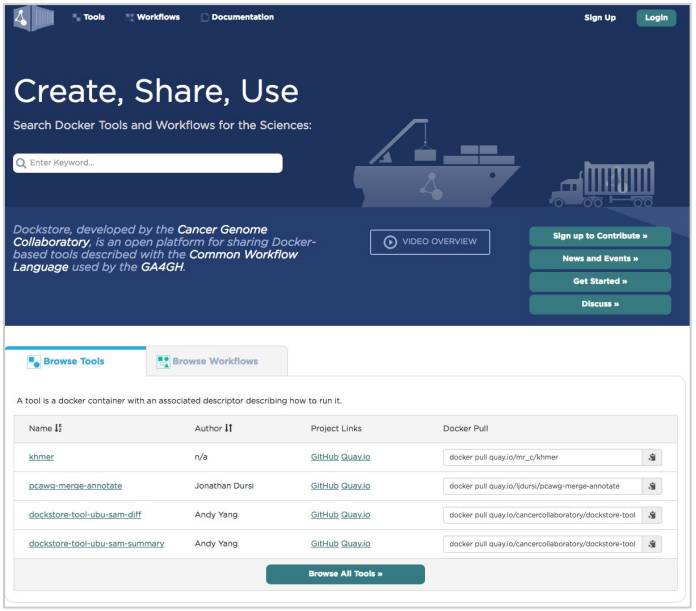


Cancer Genome COLLABORATORY
Cloud computing for collaborative research

Software Infrastructure

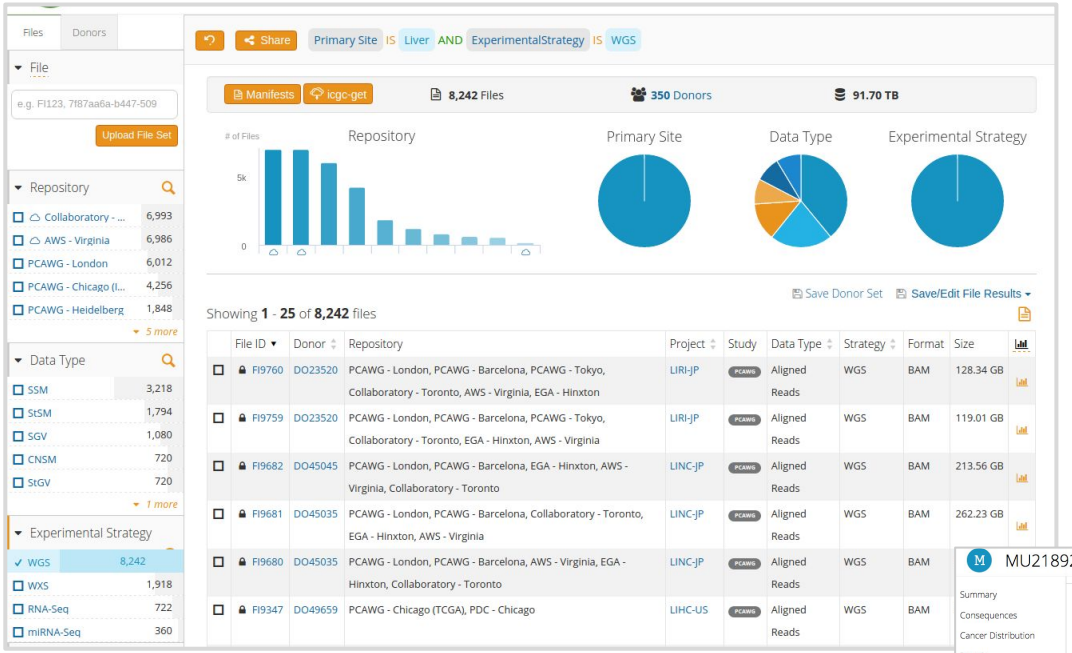
- High Performance Data Management
  - **ICGC-storage**
    - Advanced solution for managing genomic data on cloud platforms
    - Support for Collaboratory (CEPH), Amazon (S3), Microsoft Azure and soon Google Cloud
  - **SONG**
    - Metadata submission and management system

- Workflow Reproducibility
  - **DockStore** (*dockstore.org*)
    A repository of packaged data analytic workflows
  - Standardized and programmatic language for executing them

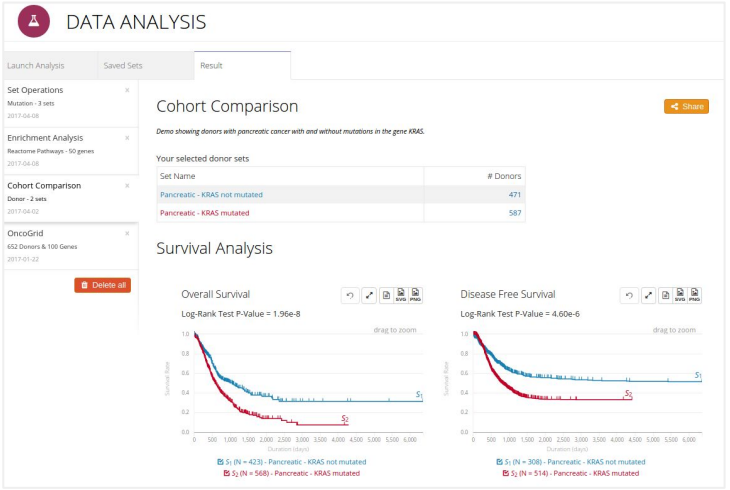


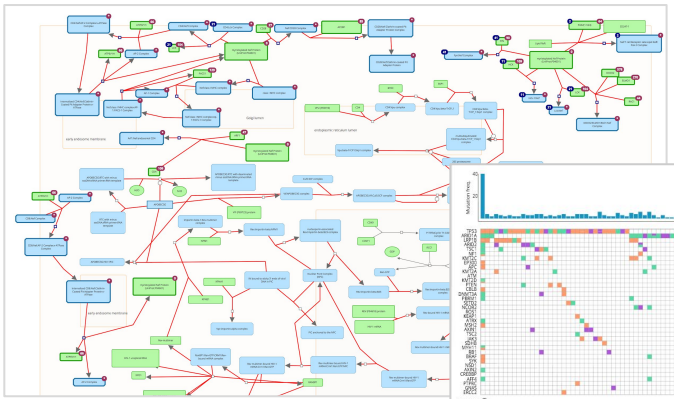
Data Discovery Platform
The ICGC Data Portal

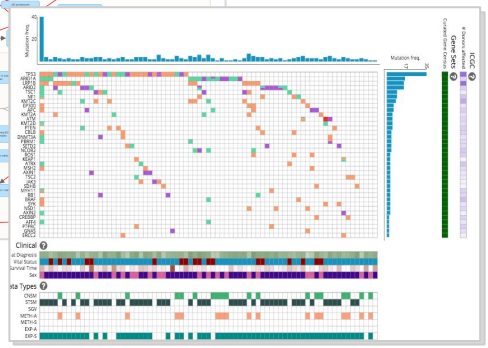
Advanced Search



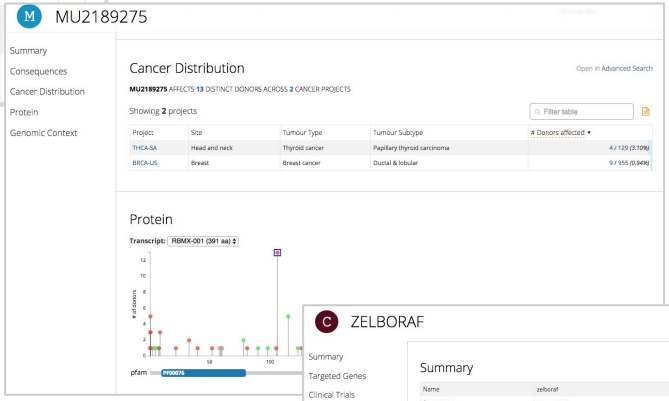
Data Analytics Toolbox

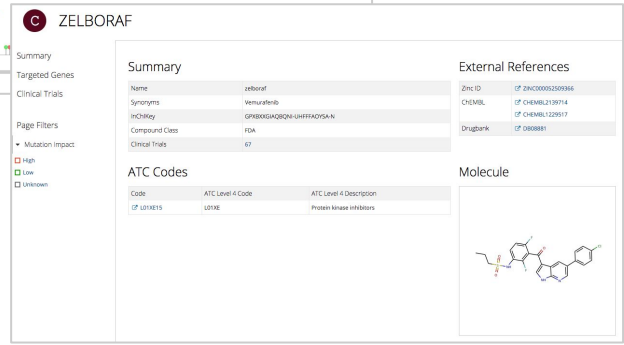






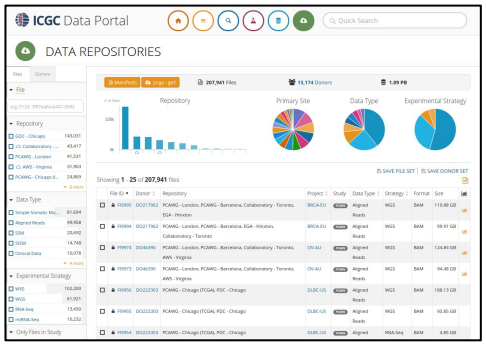
Visualization





Entity Pages

Interoperability



II - Get a `manifest-id`

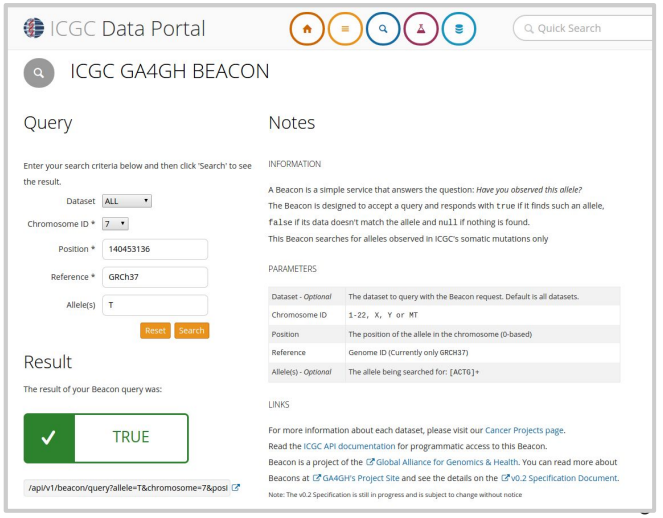
I - Search Data

III - `icgc-get download -m <manifest-id>`

- Universal data transfer software

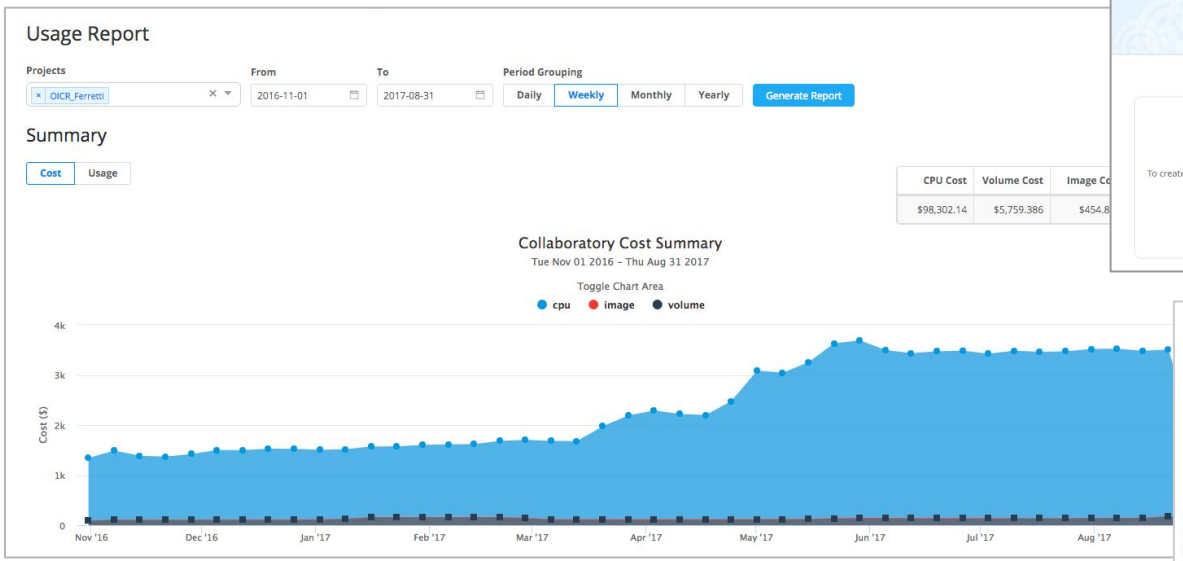
Global Alliance for Genomics & Health

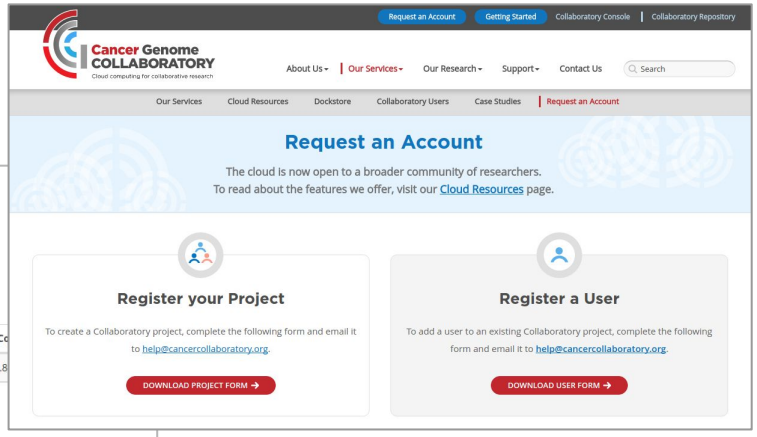
- GA4GH APIs implementation
- GA4GH discovery tool

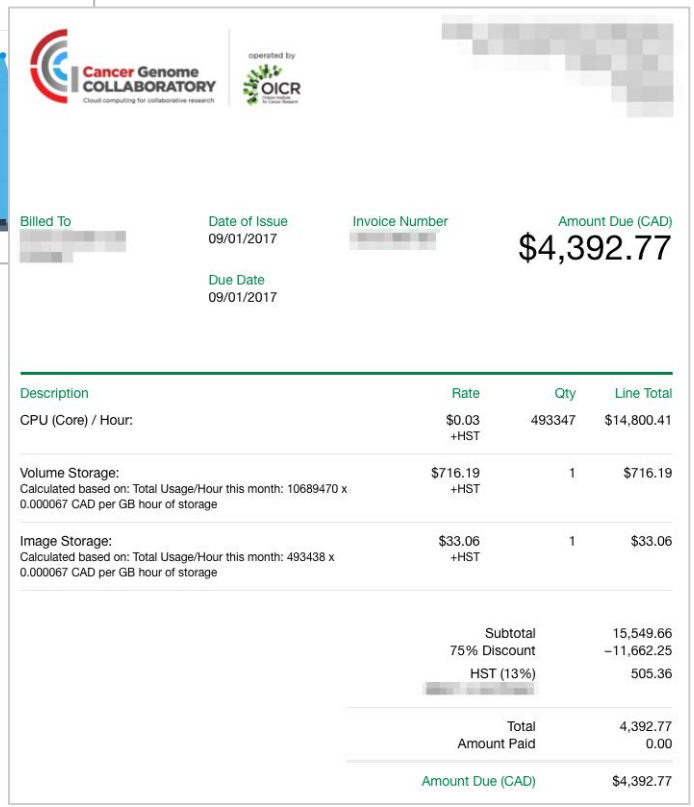


Cancer Genome COLLABORATORY
Cloud computing for collaborative research

Sustainability





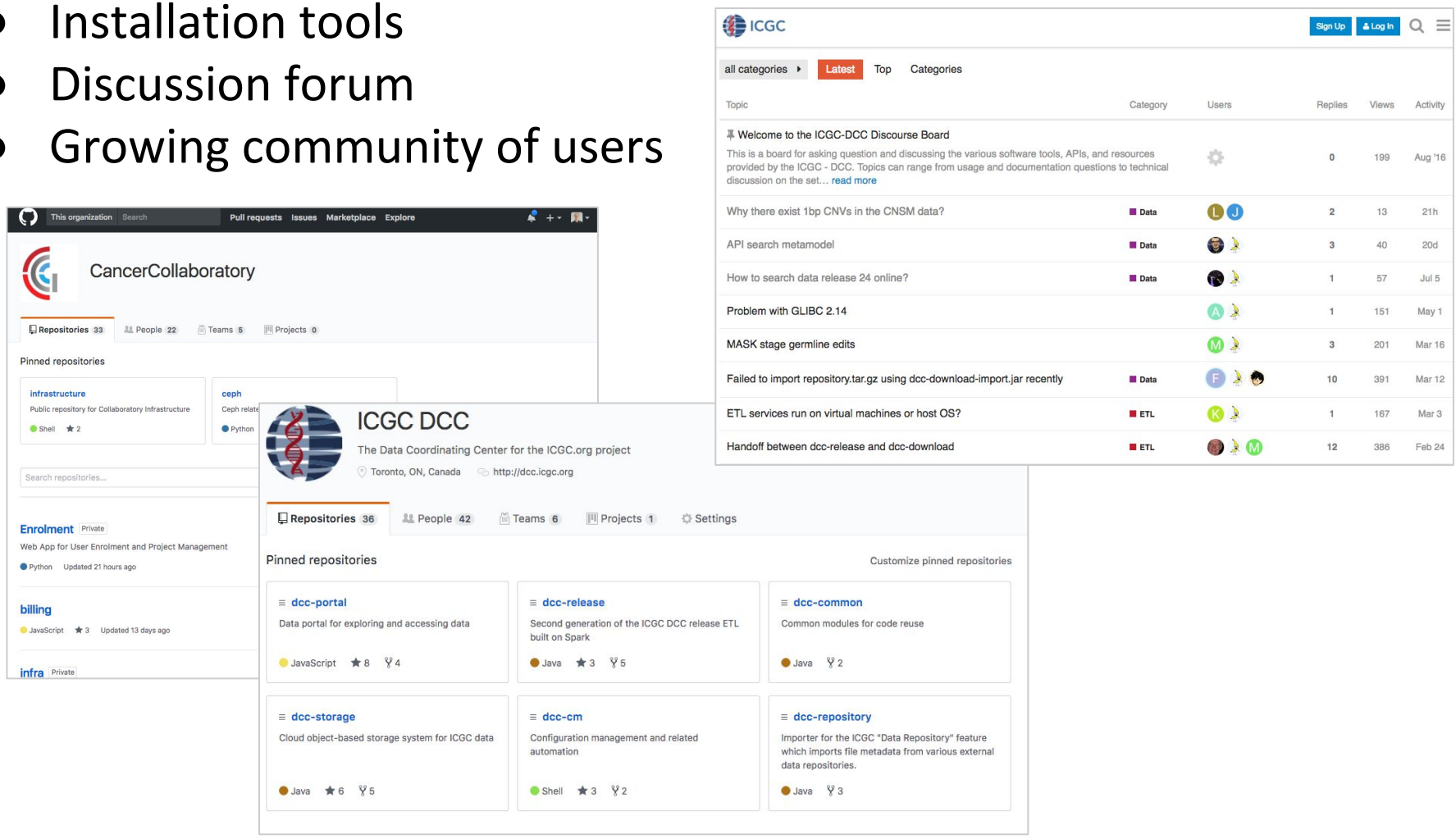


- Real-time usage monitoring (cpu, disk space, image)
- Monthly invoices
- Income reinvested in the infrastructure
- Software contribution to the OpenStack community

Cancer Genome COLLABORATORY
Cloud computing for collaborative research

Open Source Software

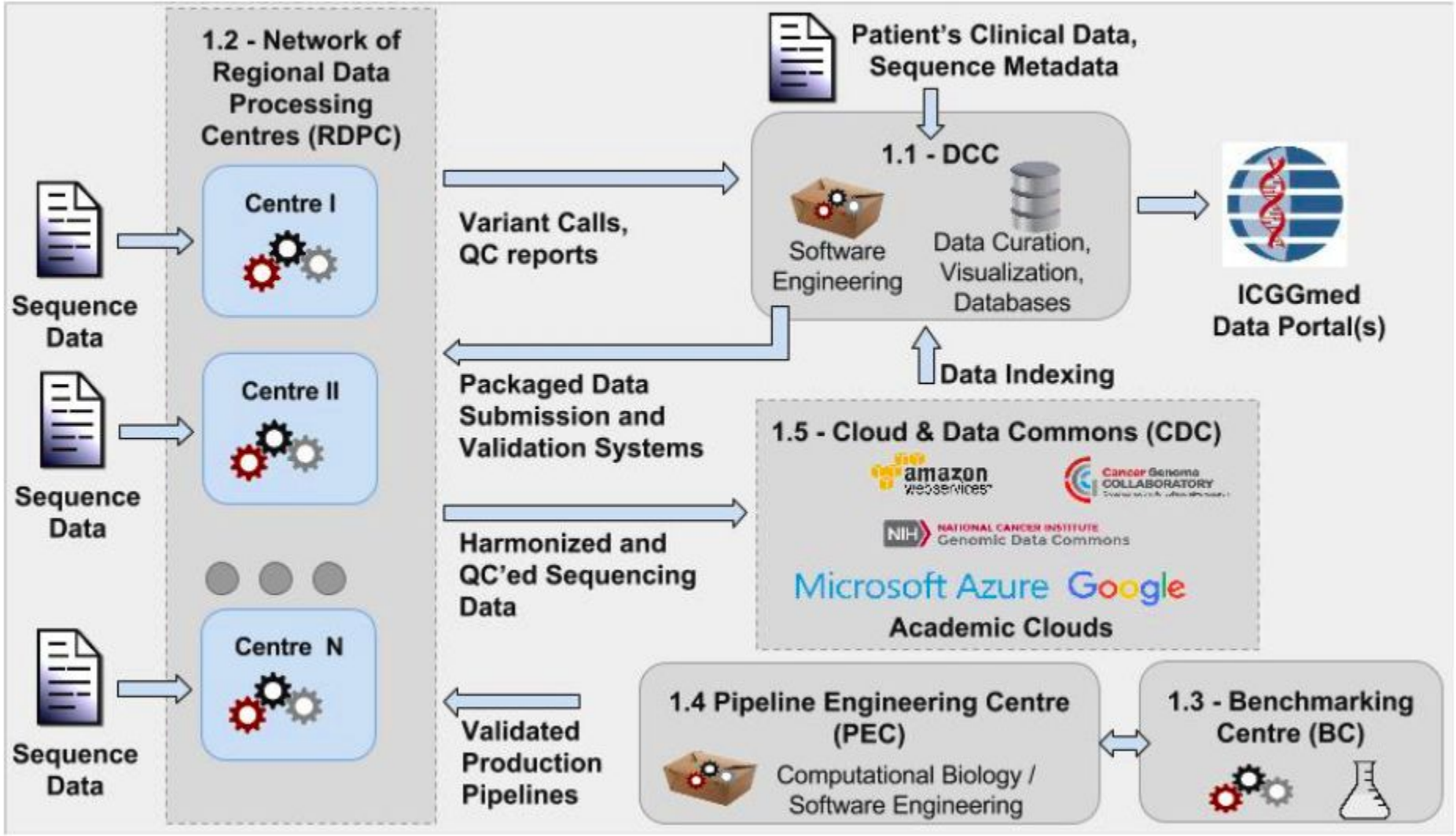
- Freely available on GitHub
- Installation tools
- Discussion forum
- Growing community of users



International Cancer Genome Consortium for Medicine

Next Steps

200,000 patients from clinical trials



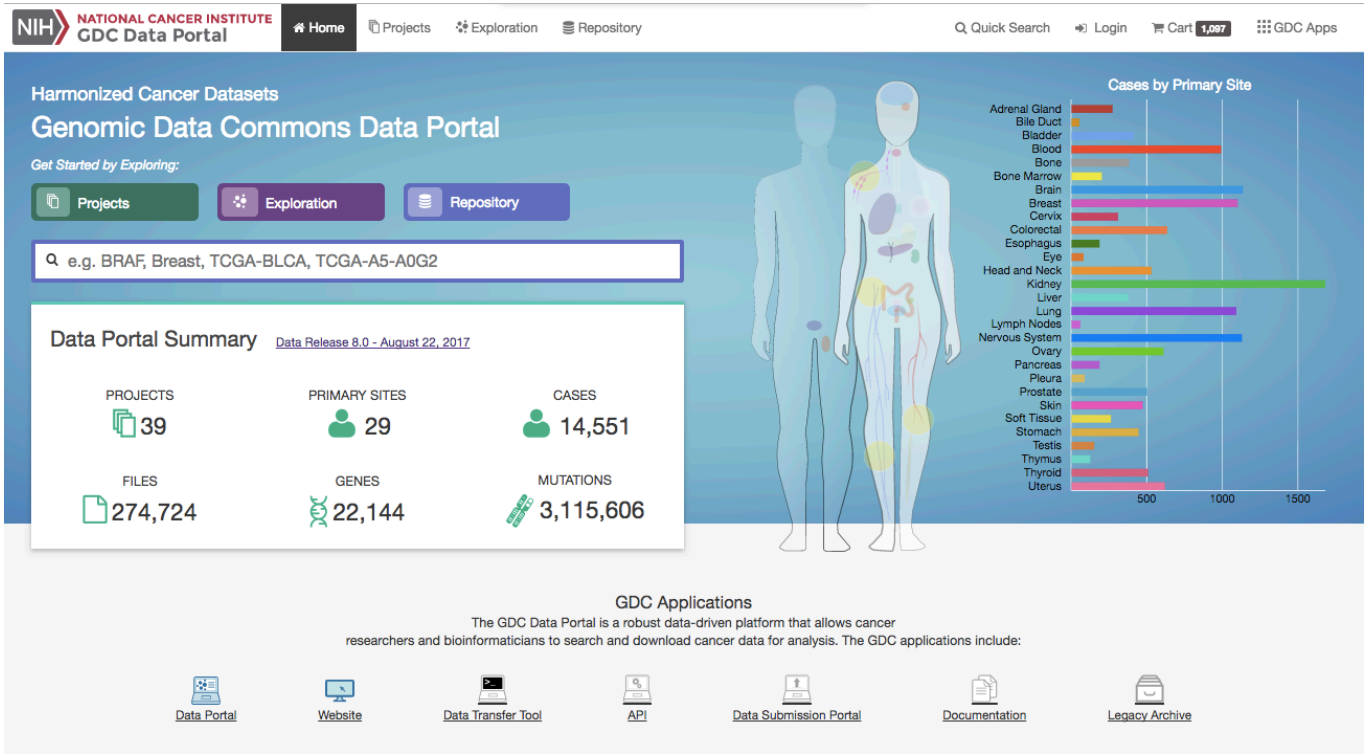
Cancer Genome COLLABORATORY
Cloud computing for collaborative research

Funders

The NCI Genomic Data Commons and Beyond

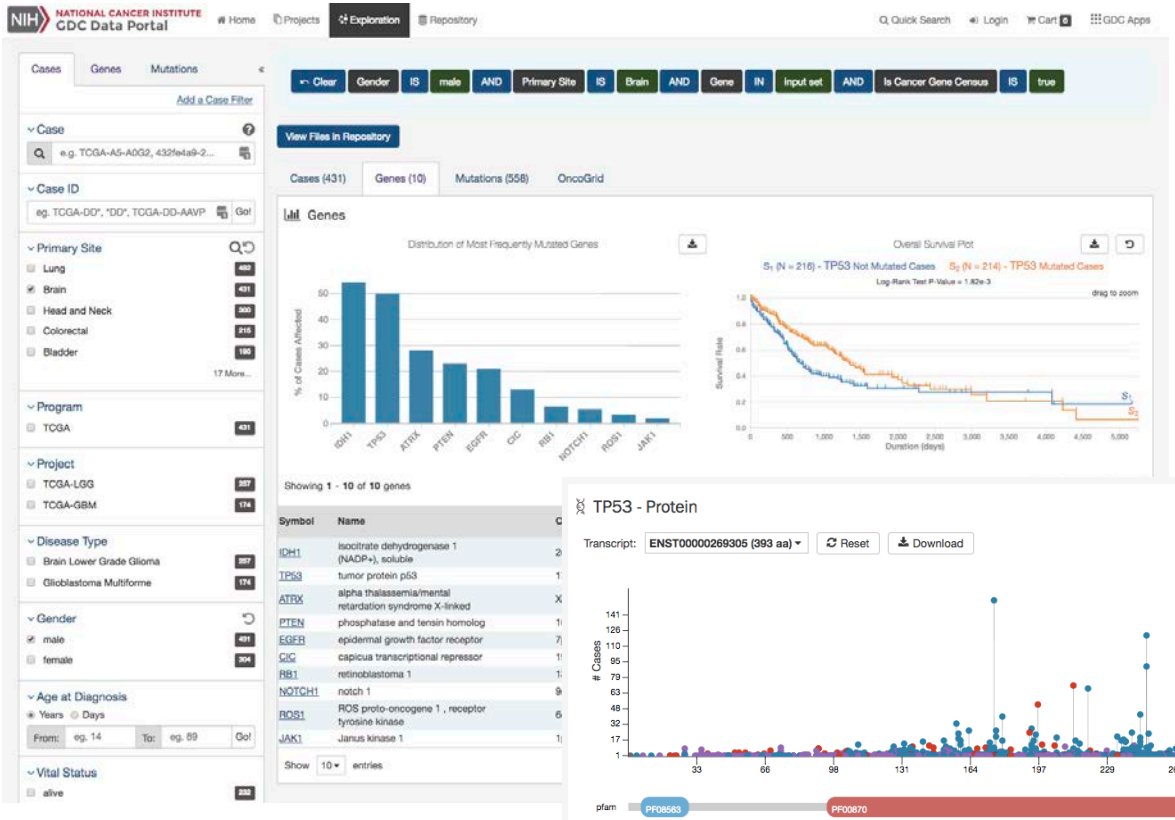
Michael Fitzsimons & Robert Grossman
Center for Data Intensive Science
University of Chicago
September 19, 2017

NCI Genomic Data Commons



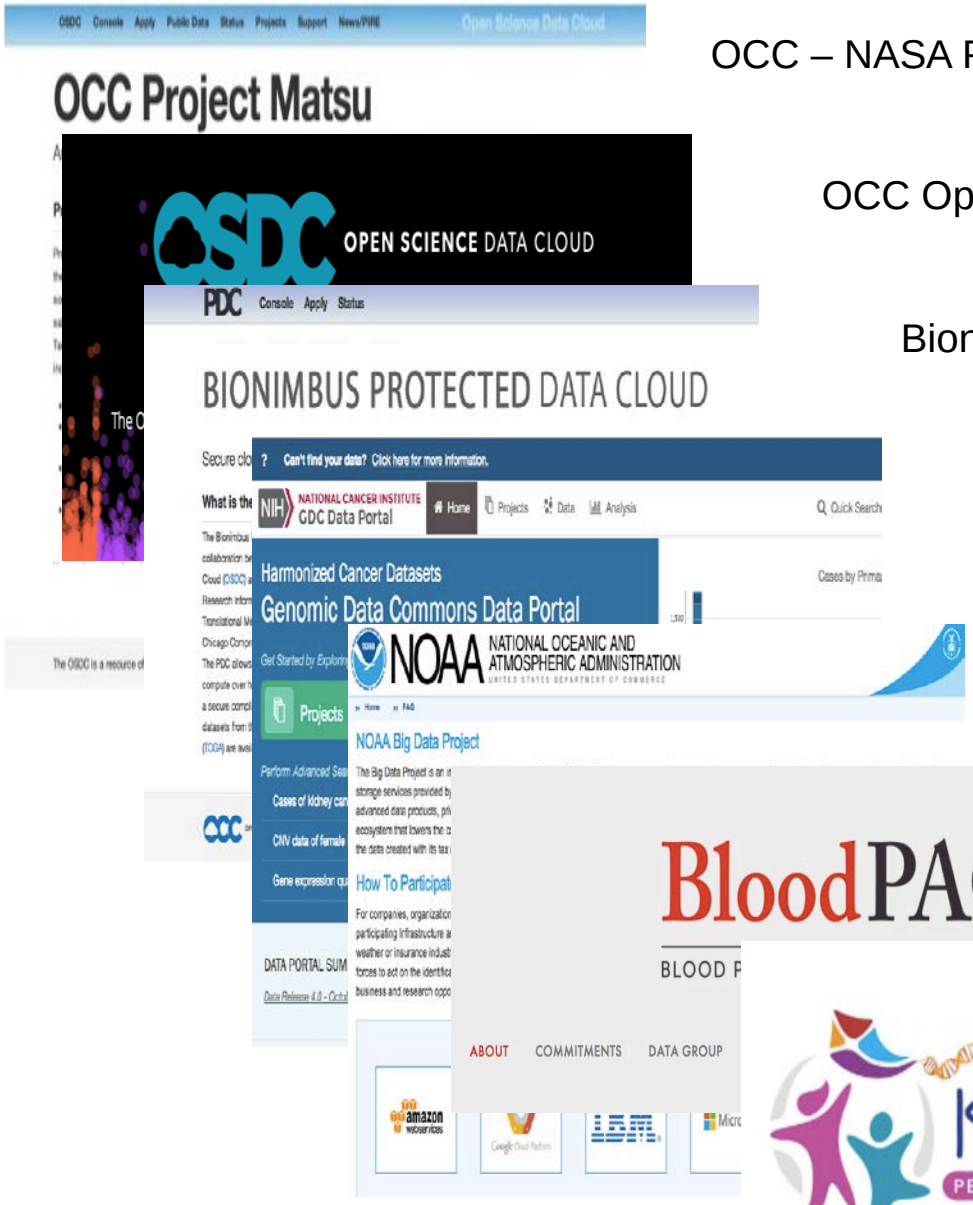
- Launched in 2016 with over 4 PB of data (equivalent of 1.5 billion eBooks).
- Joint project with Ontario Institute of Cancer Research
- Used by 1000-2000+ users per day.
- Based upon an open source software stack that can be used to build other data commons.

NCI Genomic Data Commons Updates



New Visualizations and analyses added in June 2017

18000+ patient data from Foundation Medicine Coming Soon!



OCC – NASA Project Matsu (2009)

OCC Open Science Data Cloud (2010)

Gen1

Bionimbus Protected Data Cloud* (2013)

NCI Genomic Data Commons* (2016)

Gen2

OCC-NOAA Environmental Data Commons (2016)

OCC Blood Profiling Atlas in Cancer (2017)

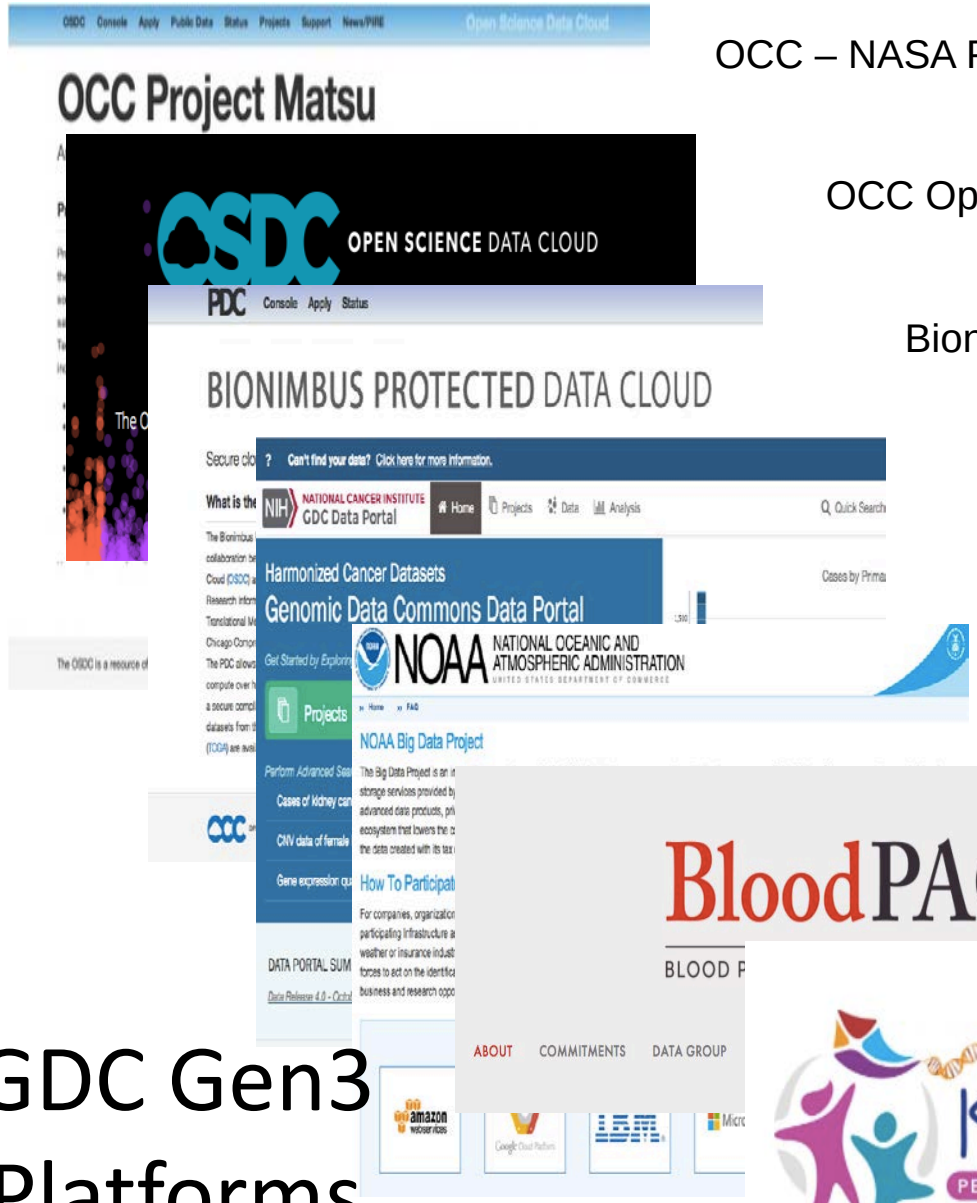
Kids First Data Resource (2017)

Brain Commons (2017)

Gen3

*Operated under a subcontract from NCI / Leidos Biomedical to the University of Chicago with support from the OCC.

GDC Gen3 Platforms



OCC – NASA Project Matsu (2009)

OCC Open Science Data Cloud (2010)

Gen1

Bionimbus Protected Data Cloud* (2013)

NCI Genomic Data Commons* (2016)

Gen2

OCC-NOAA Environmental Data Commons (2016)

OCC Blood Profiling Atlas in Cancer (2017)

Kids First Data Resource (2017)

Brain Commons (2017)

Gen3

*Operated under a subcontract from NCI / Leidos Biomedical to the University of Chicago with support from the OCC.

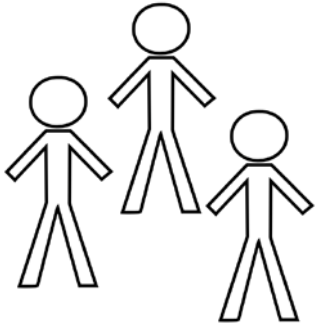
Summary

- Proven architecture – Technology has benefited from 10 years of experience building and operating data commons
- Scalable and flexible – Can be used to build large scale data commons (e.g. Genomic Data Commons and Brain Commons) as well as more targeted data commons (e.g. BloodPAC, Kids First Data Resource, Bionimbus Protected Data Cloud).
- Open and Modular – Foundations using GDC Gen3 technology can can select different features and functionality.
- Be Part of a Ecosystem of Commons – i) Gen3 platforms can peer with Gen3. Ii) We are committed to interoperating with the Broad's All of Us Data Platform and the CZI HCA Data Platform.

non-profit research
speeding medical **insight**
using **open** science

Principles

People

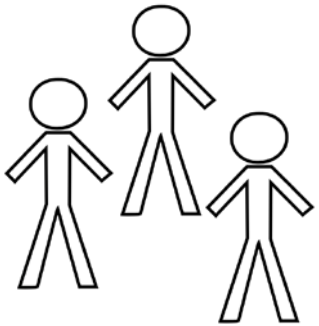


Platform

Principles

- Data sharing
- Reproducibility
- Consent

People



- Disease and biomedical science
- Data science
- Governance
- Open science advocacy

Platform

Synapse

Bridge

Cloud-based **data** and
knowledge sharing system for
collaboration

Portal Overview

This portal provides access to genomic and physical cell line characterization studies funded by the National Cancer Institute.

The Physical Sciences - Oncology Center (PS-OC) Bioresource Core Facility (PBCF) was established by the NCI to create a panel of model cell lines available to PS-OC investigators. These PBCF consists of 39 cell lines from a variety of tissue types and includes standard operating procedures for the growth and handling of each line. The cell lines are intended to be well-characterized biological reagents for the PS-OC Network on which they can develop novel 1) biological models of the physical processes associated with the pathogenesis of neoplastic diseases and 2) analytical systems for characterizing the associated cellular phenomena.

Data Summary

Study	Description	Num. cell lines	Num. experimental conditions	Available files
Morphology	This study uses images of cells collected using brightfield microscopy at 10x magnification, with ImageJ software used to trace the outline of single cells as well as to report area, circularity and aspect ratio.	30	7	tif, jpg, Excel, txt
Motility	This study uses images of cells collected using brightfield microscopy at 10x magnification, with ImageJ software used to track motility.	30	7	tif, Excel, txt
Atomic Force Microscopy	This study uses atomic force microscopy (AFM) to measure the the deflection of a cantilever upon contact with the cells.	30	7	Excel, txt
Traction force and	In this study, live cells were plated on fluorescent beads and fluorescently labeled with CellTracker Green CMFDA (Invitrogen) and DRAQ5 (Cell Signaling Technology) to label cytoplasm and cell	29	1	tif, jpg,

PSON Cell Line Characterization

Search | Justin Guinney (Justin.Guinney) | Help

Synapse ID: syn7248578 **Storage Location:** Synapse Storage

Share | Annotations | Tools

Wiki | Files | Tables | Discussion

Name	Modified On	Size	MD5	ID
Atomic Force Microscopy	09/15/2016 5:05:46 PM			syn7248585
AFM.2.zip	03/08/2017 4:44:47 PM	45.873 MB	56faaeel	syn7696862
PC-3 private folder	03/17/2017 4:28:21 PM			syn8477208
README.txt	11/21/2016 2:46:15 AM	7.263 KB	babe4ba	syn7720660
analysis	11/20/2016 11:55:07 PM			syn7695116
ascii	11/20/2016 11:55:02 PM			syn7695108
summary	11/20/2016 11:55:13 PM			syn7695124
Exome	09/15/2016 5:05:16 PM			syn7248584
Morphology	09/15/2016 5:06:27 PM			syn7248591
Motility	09/15/2016 5:06:17 PM			syn7248586
Proteomics	04/25/2017 8:44:12 AM			syn9697791
Traction Force and Volume	11/21/2016 9:35:55 AM			syn7248592
linking files	02/14/2017 4:19:10 PM			syn8259616
mRNA	09/15/2016 5:05:11 PM			syn7248583
miRNA	09/15/2016 5:05:05 PM			syn7248581

Data Available By Cell Line ☆

Synapse ID: syn10496425 **Conditions for use:** None report issue

Show advanced search

cellLine

- 22Rv1 (1)
- A375 (1)
- Caco-2 (1)
- Caov-3 (1)
- Capan-1 (1)

Show all 39

Clear all

catalogNumber	cellLine	miRNA	mRNA	Exome	Atomic_Force_Microscopy	Motility	Morphology	Traction_Force_and_Volume	Proteomics
NCI-PBCF-HTB14	U-87	true	true	true	true	true	true	true	true
NCI-PBCF-CRL1690	T98G	true	true	true	true	true	true	true	true
NCI-PBCF-CRL4010	hTERT-HME1	true	true	true	true	true	true	true	false
NCI-PBCF-HTB22	MCF-7	true	true	true	true	true	true	true	false
NCI-PBCF-1001	MCF7-B7-TS	true	true	true	false	false	false	false	false
NCI-PBCF-1000	MCF10A-JSB	true	true	true	true	true	true	true	false
NCI-PBCF-HTB133	T-47D	true	true	true	true	true	true	true	true
NCI-PBCF-CRL1500	ZR-75-1	true	true	true	false	false	false	false	false
NCI-PBCF-HTB26	MDA-MB-231	true	true	true	true	true	true	true	true
NCI-PBCF-HTB123	DU4475	true	true	true	false	false	false	false	false

Portal Overview

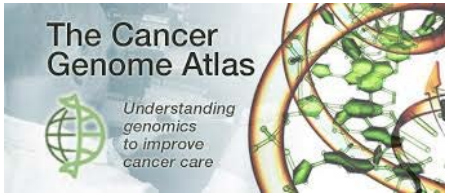
This portal provides access to genomic and physical cell line characterization studies funded by the National Cancer Institute.

The Physical Sciences - Oncology Center (PS-OC) Bioresource Core Facility (PBCF) was established by the NCI to create a panel of model cell lines available to PS-OC investigators. These PBCF consists of 39 cell lines from a variety of tissue types and includes standard operating procedures for the growth and handling of each line. The cell lines are intended to be well-characterized biological reagents for the PS-OC Network on which they can develop novel 1) biological models of the physical processes associated with the pathogenesis of neoplastic diseases and 2) analytical systems for characterizing the associated cellular phenomena.

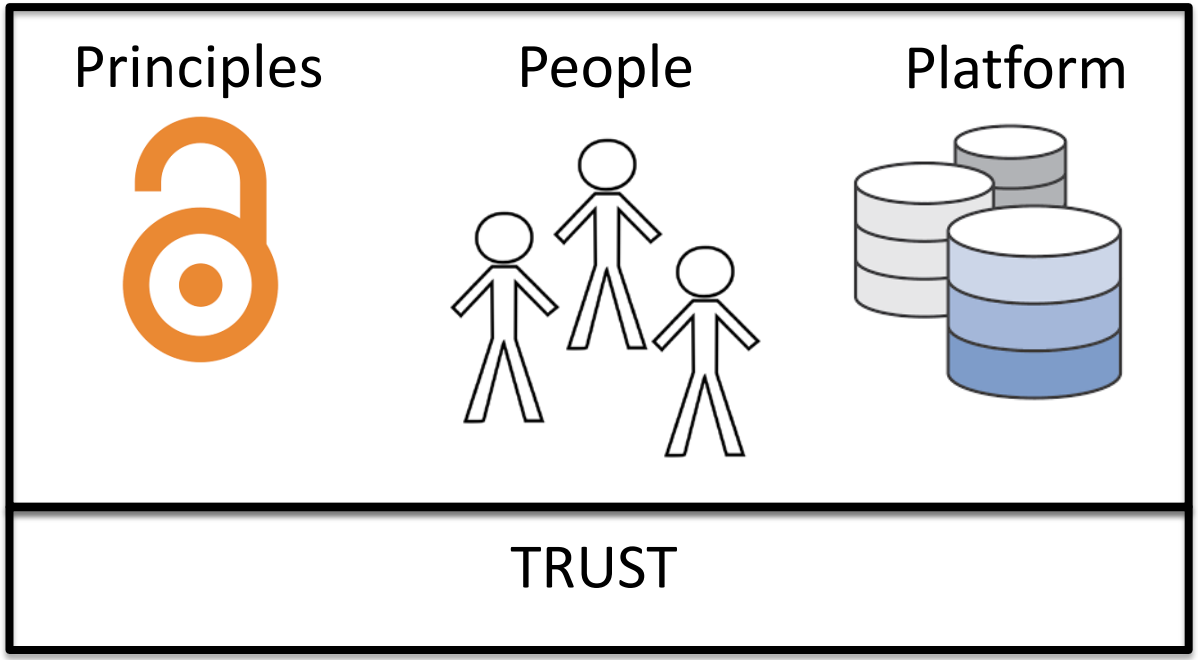
Data Summary

Study	Description	Num. cell lines	Num. experimental conditions	Available files
Morphology	This study uses images of cells collected using brightfield microscopy at 10x magnification, with ImageJ software used to trace the outline of single cells as well as to report area, circularity and aspect ratio.	30	7	tif, jpg, Excel, txt
Motility	This study uses images of cells collected using brightfield microscopy at 10x magnification, with ImageJ software used to track motility.	30	7	tif, Excel, txt
Atomic Force Microscopy	This study uses atomic force microscopy (AFM) to measure the the deflection of a cantilever upon contact with the cells.	30	7	Excel, txt
Traction force and	In this study, live cells were plated on fluorescent beads and fluorescently labeled with CellTracker Green CMFDA (Invitrogen) and DRAQ5 (Cell Signaling Technology) to label cytoplasm and cell	29	1	tif, jpg,

Communities we support…



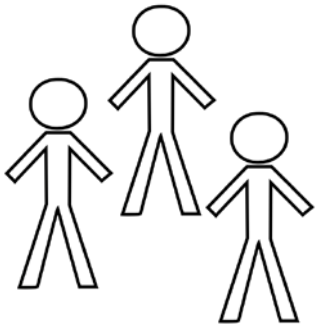
… and many others


Principles
People
Platform
TRUST

INCENTIVES

Principles

People



Platform

TRUST

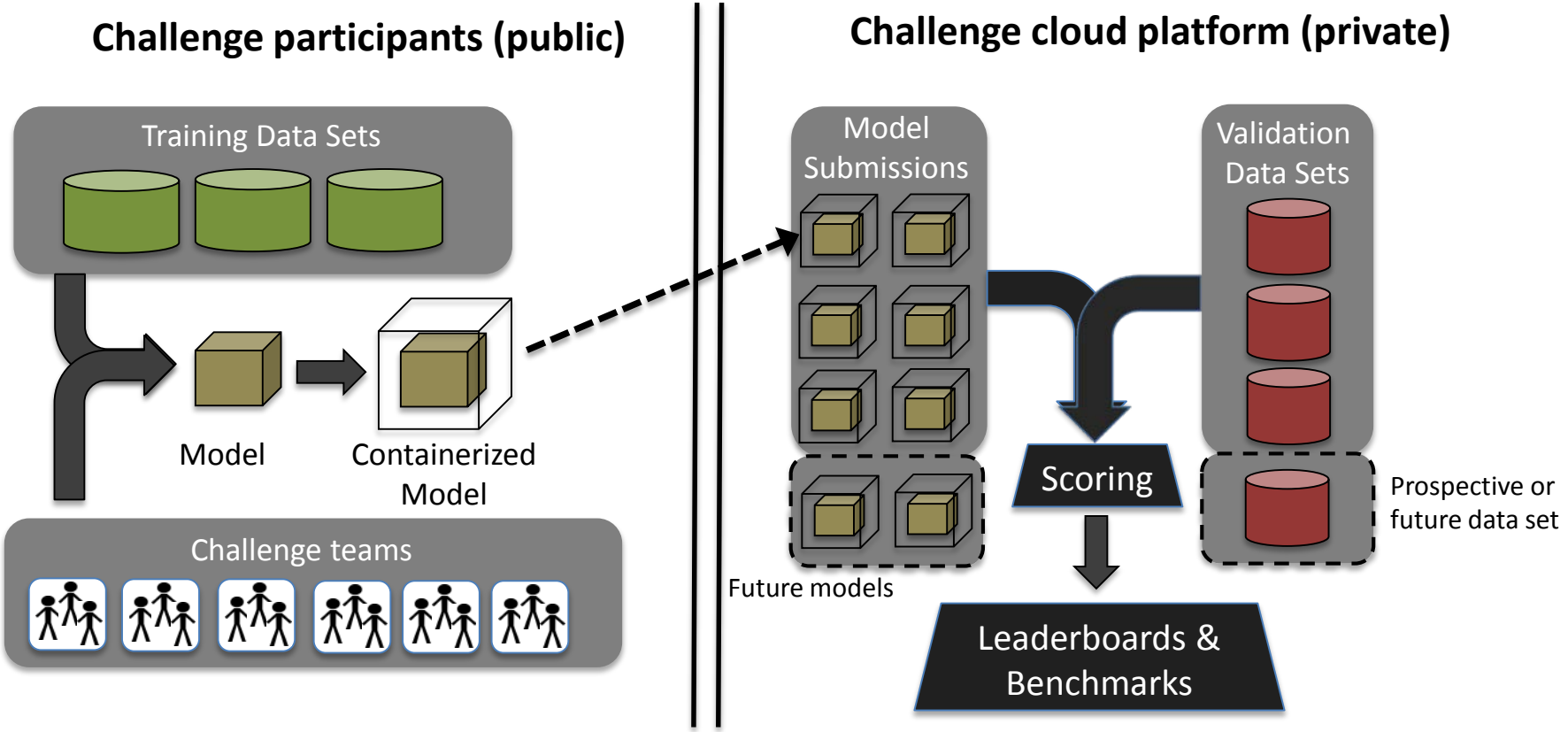
INCENTIVES

1. $$$
2. **Citizen science**
3. **Crowd-sourced competitions**

Challenge platform

DREAM CHALLENGES

CAFA3


Challenge participants (public)
Challenge cloud platform (private)
Training Data Sets
Model
Containerized Model
Challenge teams
Model Submissions
Validation Data Sets
Scoring
Prospective or future data set
Future models
Leaderboards & Benchmarks

Improve accuracy of digital mammogram screening

1 in 10 women are falsely diagnosed with breast cancer

640k

10k

600k

500k

Over 1.7 million digital mammograms

~~big~~ huge data

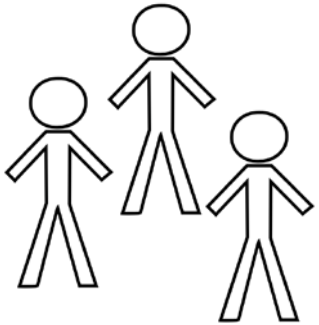
key stats

1k participants
10k model submissions
874k CPU-hours
1,000 TB data usage

Sage BIONETWORKS

Principles

People



Platform

Seven Bridges

The Cancer Genomics Cloud

An NIH Commons Conformant Cloud Service Provider

CANCER GENOMICS CLOUD
SEVEN BRIDGES

The Seven Bridges Cancer Genomics Cloud

- A stable, secure, and highly customizable cloud storage and computing platform
- A user-friendly portal for collaborative analysis of petabytes of public data alongside private data
- An optimized venue for reproducible data analysis using validated tools and pipelines

Easy data management

Scalable computation

Secure collaboration

Optimized bioinformatics algorithms

Flexible & fully reproducible methods

Extensible & developer-friendly platform

Access Petabytes of Public Data

TCGA

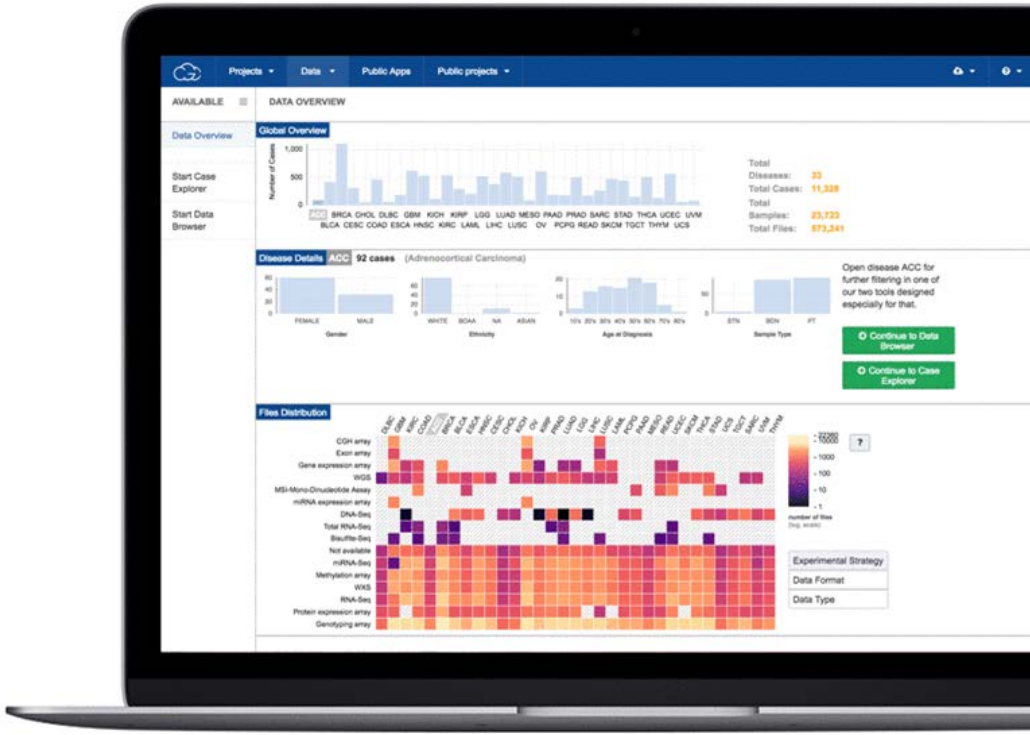
TARGET

CLINICAL PROTEOMIC TUMOR ANALYSIS CONSORTIUM

CCLE Cancer Cell Line Encyclopedia

SIMONS FOUNDATION

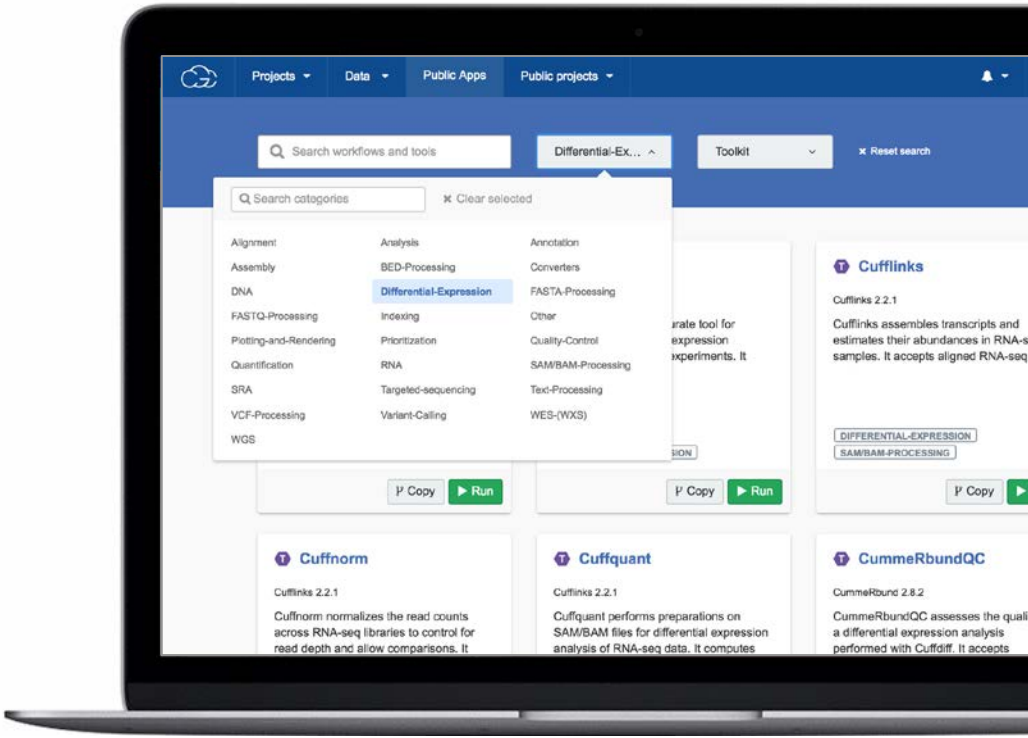
THE CANCER IMAGING ARCHIVE



Access Best-Practices Bioinformatics Workflows

- Select from >250 tools & workflows that are:
  - fully parameterized & customizable
  - optimized for speed & cost on the cloud
  - accessible via the GUI and API

- Align private and public data analysis results using common workflows.



Researchers Have Used the Cancer Genomics Cloud To...

- Detect aberrant splice junctions and splicing profiles across cancer types
- Identify neoantigens arising from novel gene fusion events
- Profile miRNA expression across cancer types
- Conduct HLA typing to identify neoantigens
- Compare viral infection patterns across cancer types
- Detect novel gene fusions from RNA-Seq data with a near-zero false positive rate
- Identify cis-regulatory region variants across cancer types
- ...and much more

NATIONAL CANCER INSTITUTE

This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN261201400008C.

cancergenomicscloud.org
cgc@sbgenomics.com

NATIONAL CANCER INSTITUTE

This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN261201400008C.



Big Data Sharing Meeting

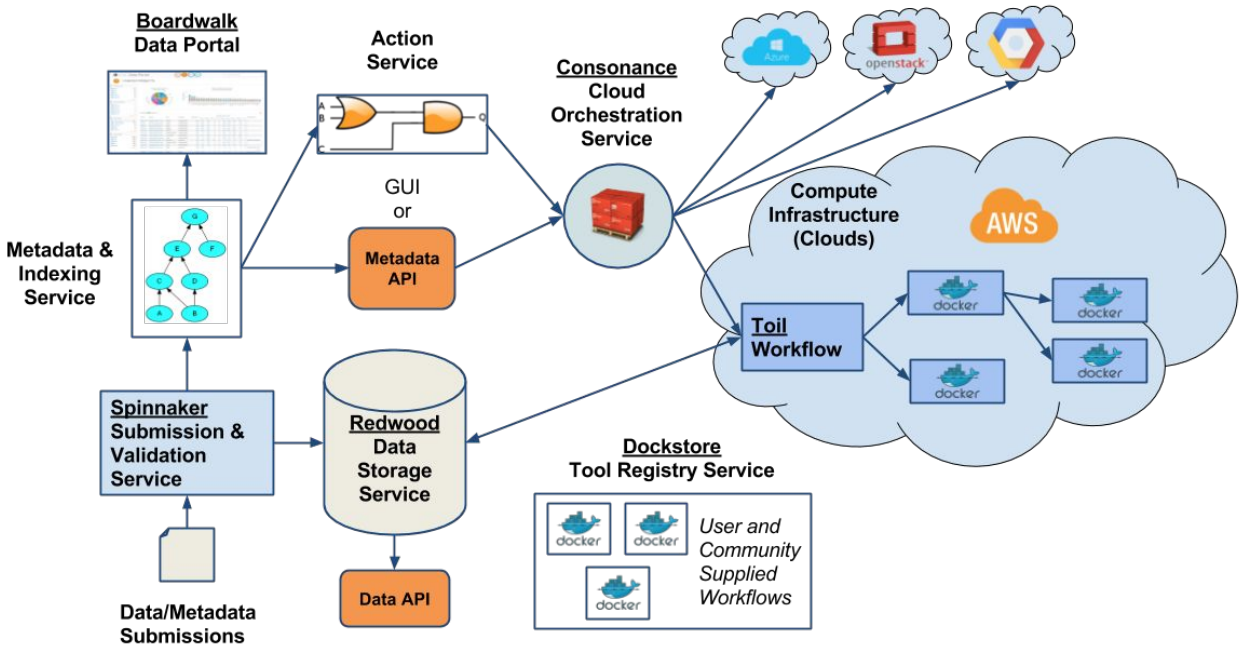
Benedict Paten
Director - Computational Genomics Lab
UCSC Genomics Institute
September 19th, 2017

UC SANTA CRUZ

Computational Genomics Platform (CGP)

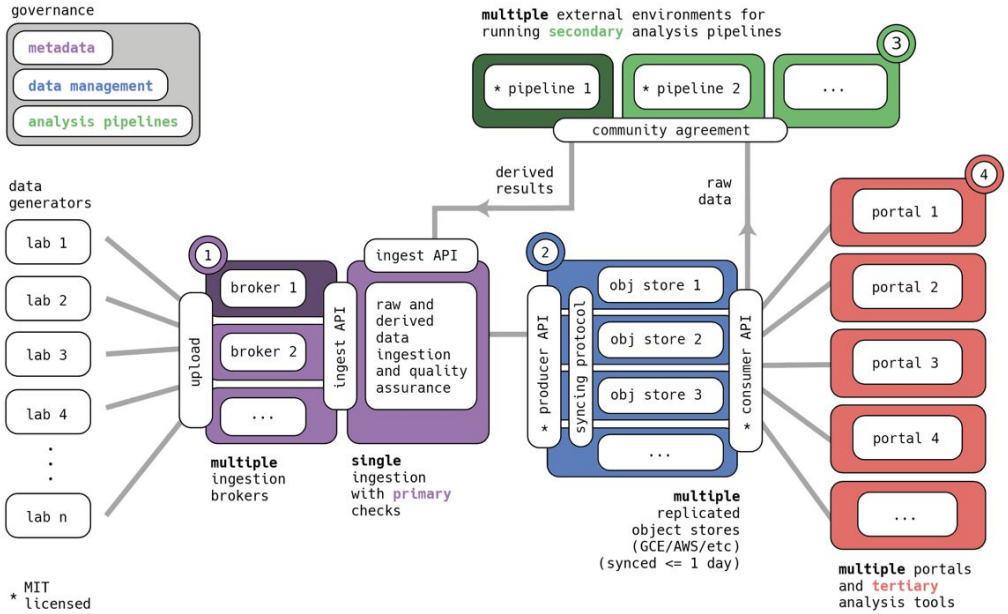
A Framework for Cloud Data Commons

- Platform used by:
  - St Baldrick's Treehouse project
  - SU2C West Coast Dream Team



Human Cell Atlas

To create comprehensive reference maps of all human cells—the fundamental units of life—as a basis for both understanding human health and diagnosing, monitoring, and treating disease.

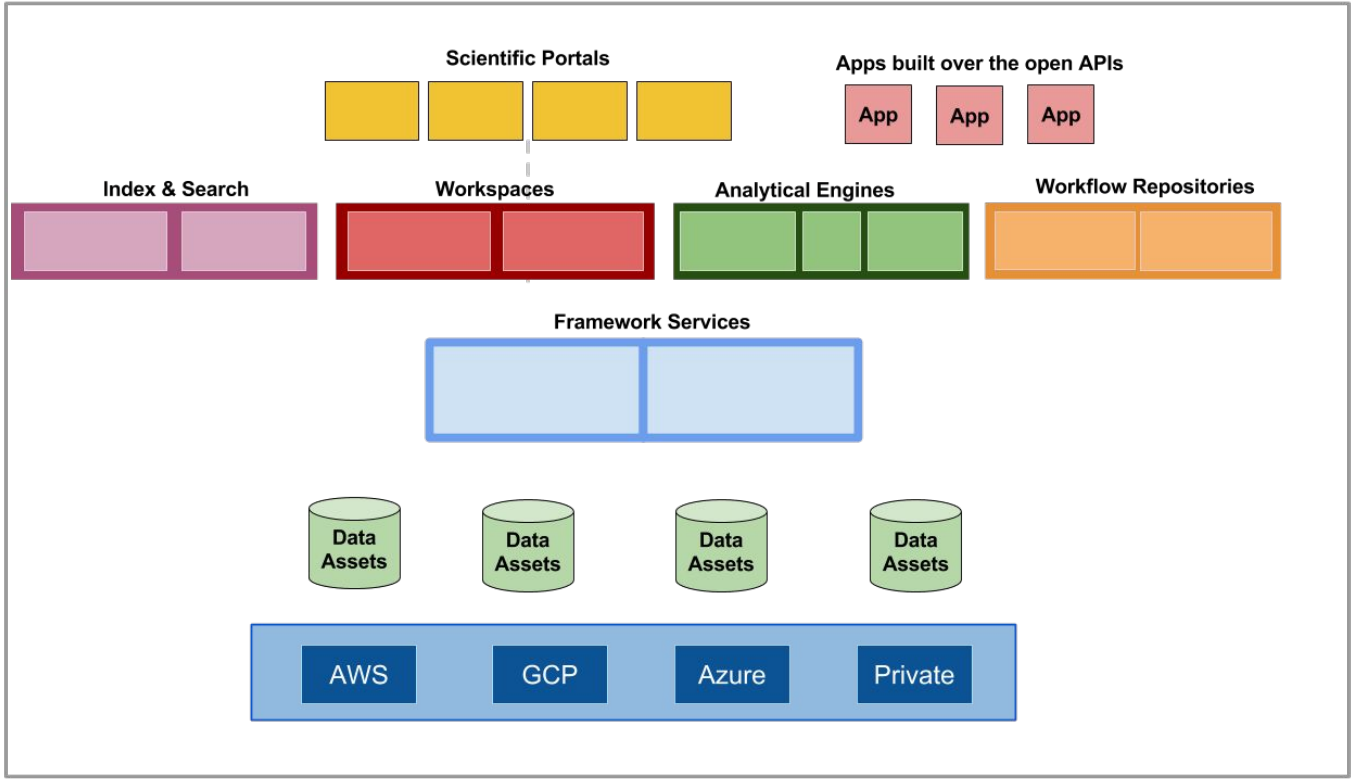


Infrastructure** as a possible bridge project between the **NIH Cloud Commons** efforts and **HCA**, made interoperable through **GA4GH API standards

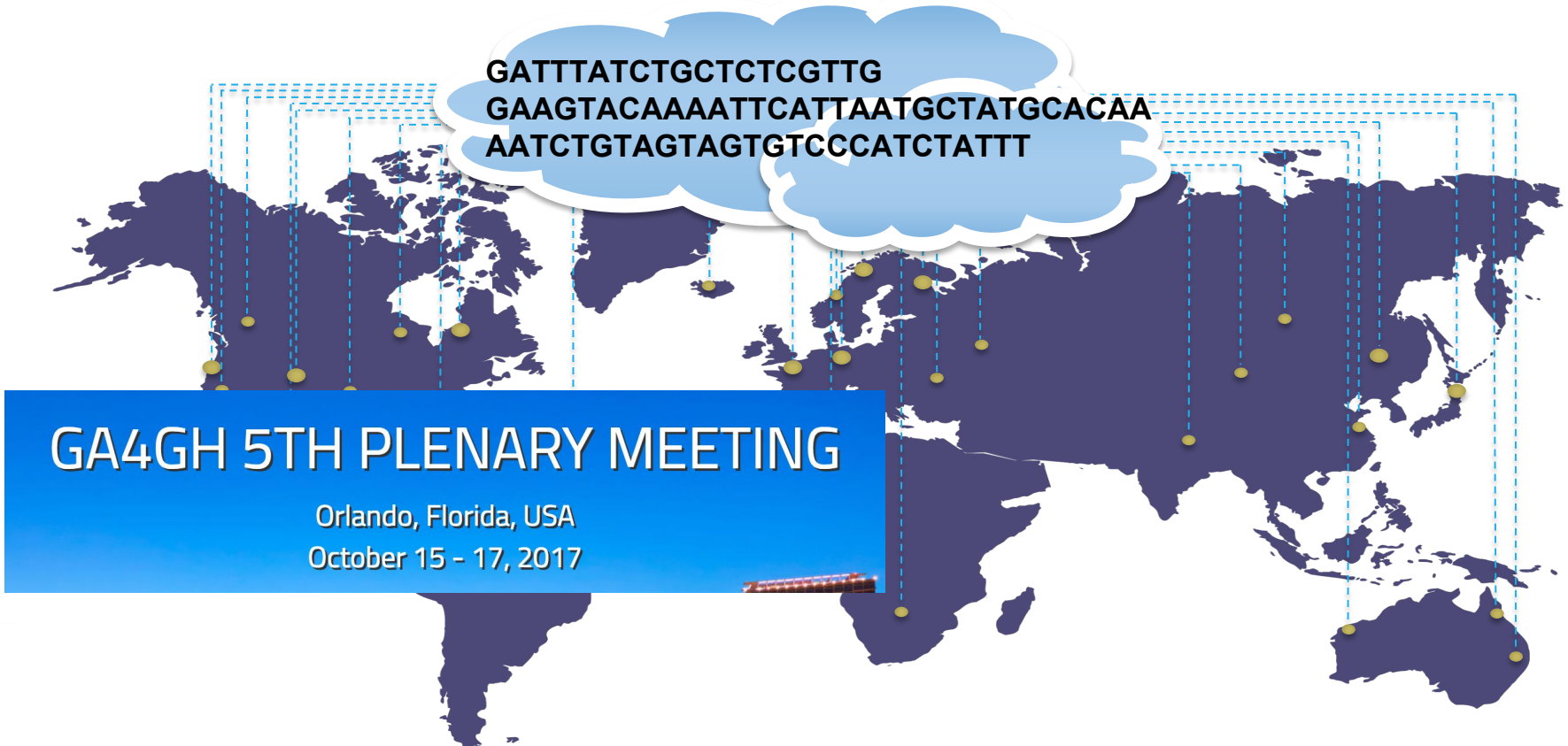
prepared jointly by **HCA Broad UCSC EBI/Sanger CZI**

Alliances Committed to Interoperating Large-Scale Commons

- NCI GDC / Cloud Resources
- NIH All of Us
- CZI HCA Data Platform



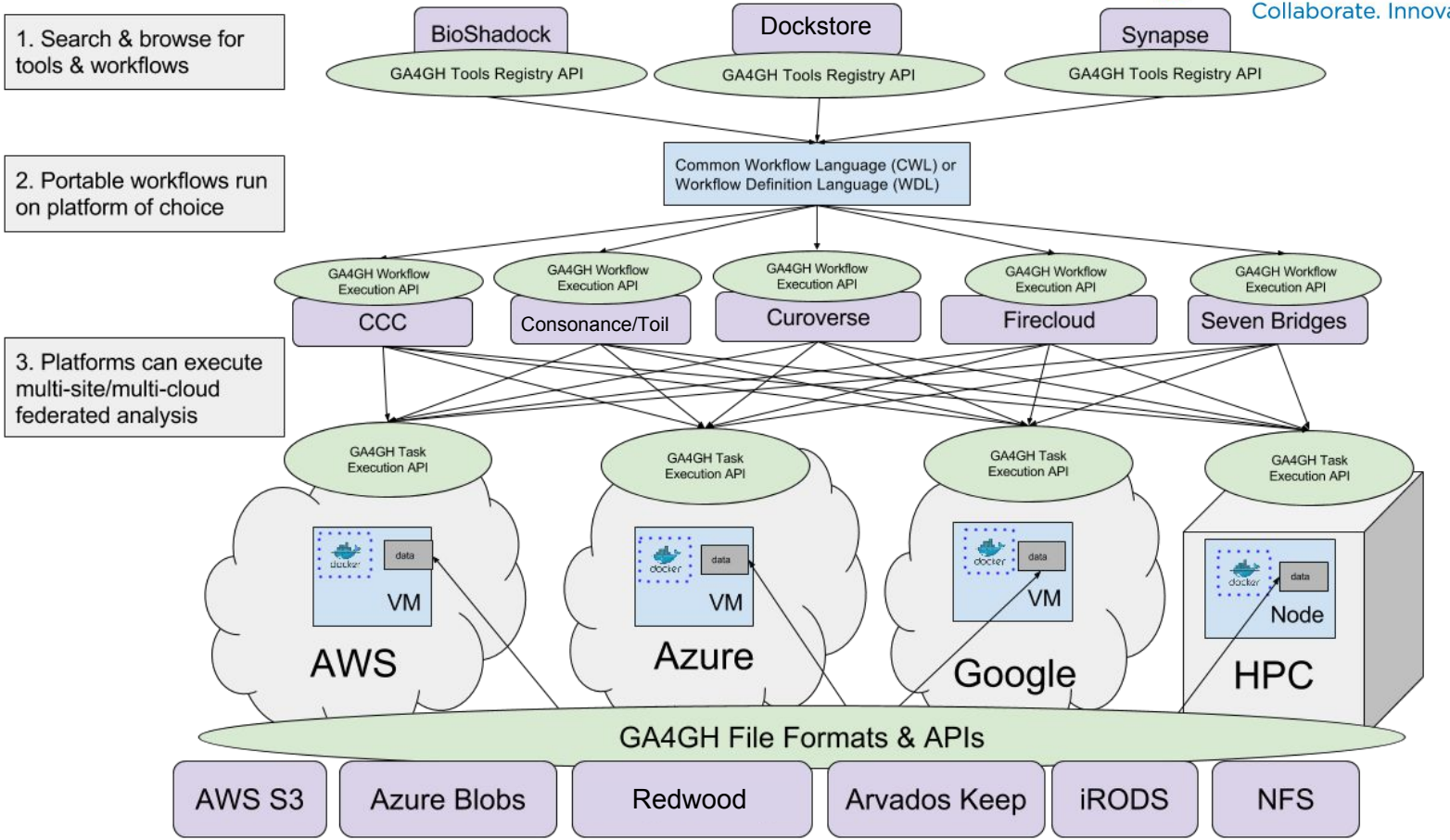
Data Exchange Standards - GA4GH APIs



GA4GH Ecosystem for Cloud Commons

Global Alliance for Genomics & Health

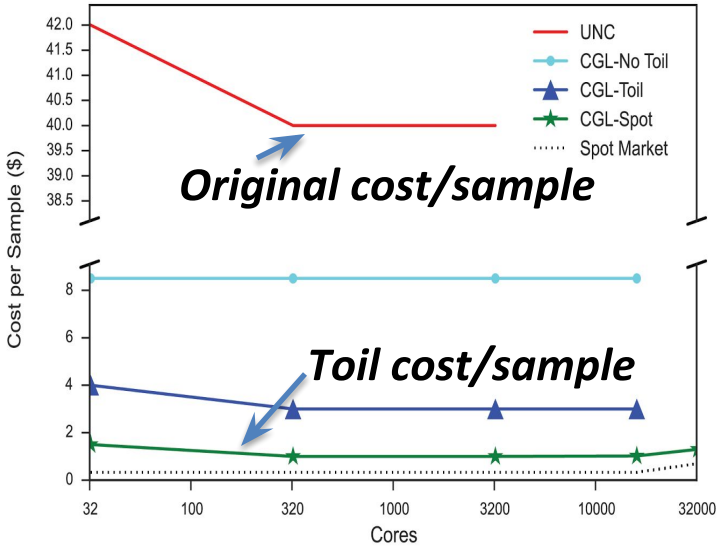
Collaborate. Innovate. Accelerate.

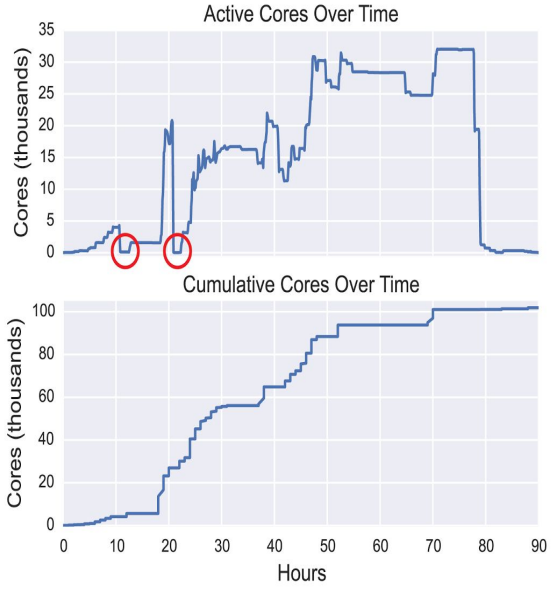


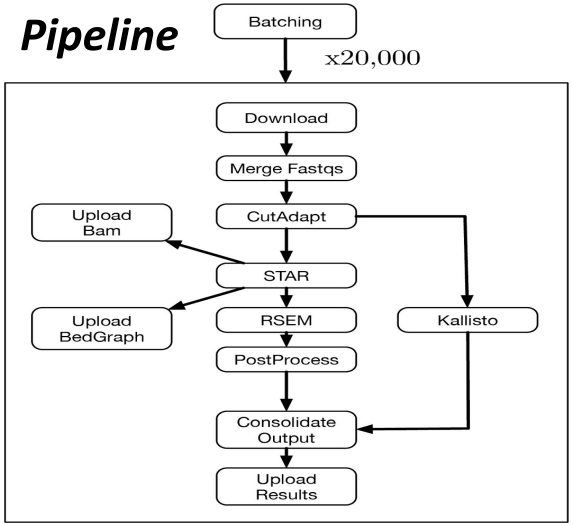
Large-Scale, Distributed Analysis Efforts - Toil Compute

- *20,000 RNA Seq samples – all TCGA, gTEX, PNOC, TARGET and I-SPY2.*
- *Computed in < 4 days, 30,000+ Cores, 1/40th the cost of previous pipeline*
- *Hundreds of registered users*

TOIL
Massively Parallel Workflows
(and fire-breathing dragon slugs)





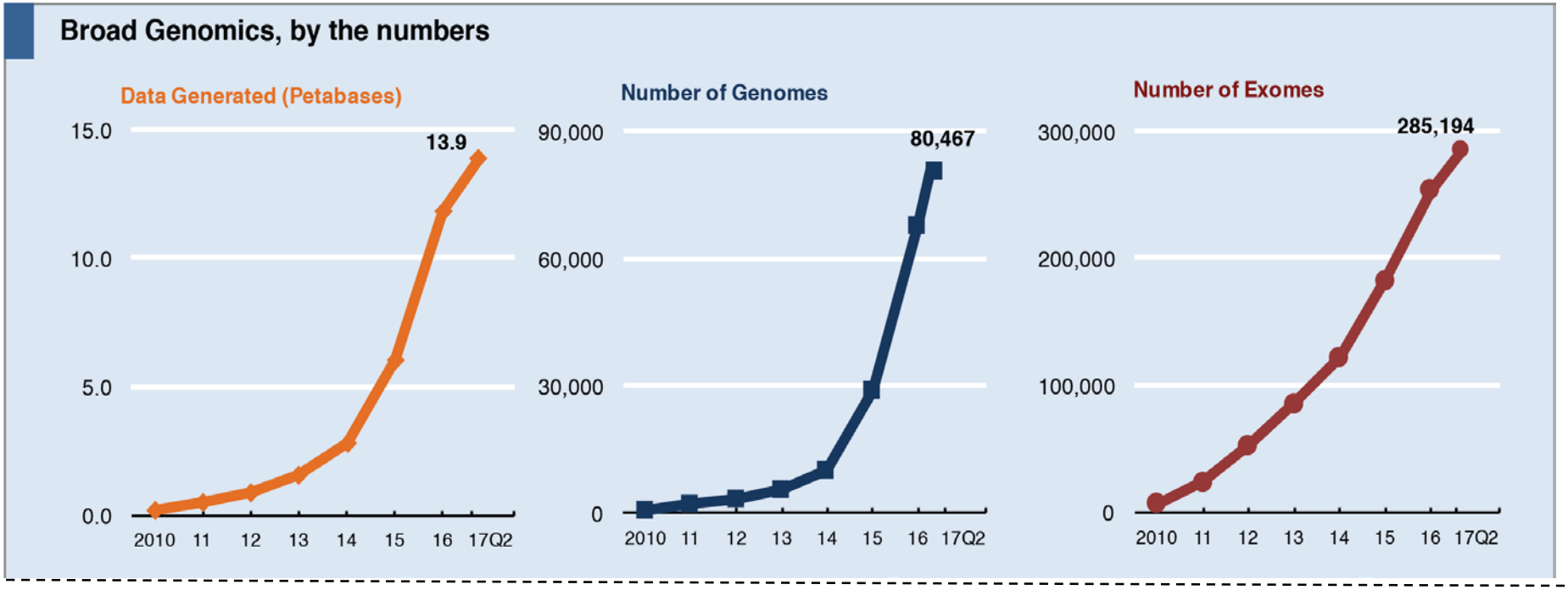


See Vivian et al. 2016, Rapid and efficient analysis of 20,000 RNA-seq samples with Toil, Nature Biotech in press

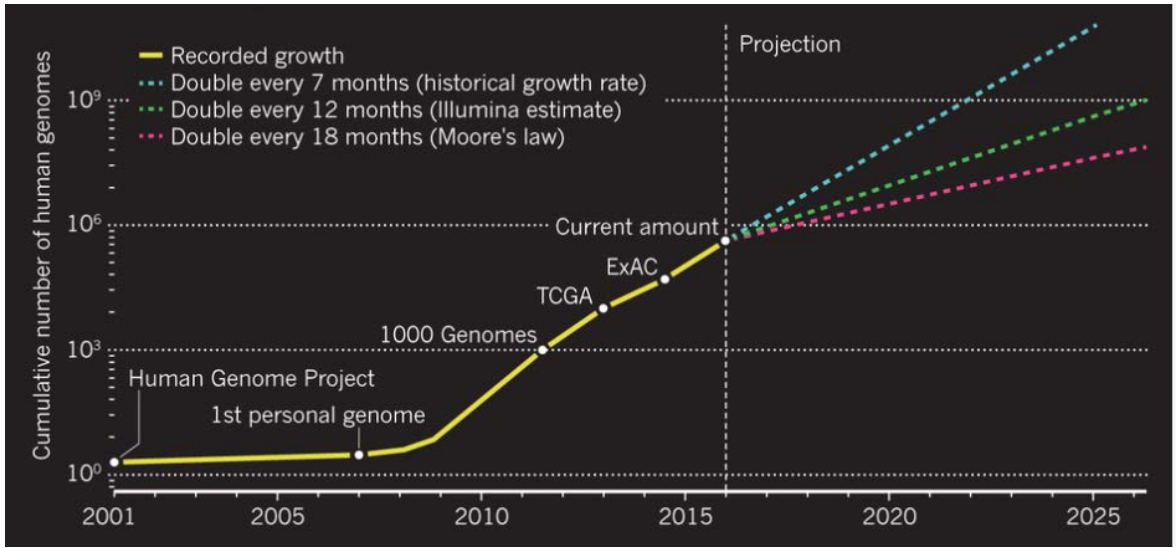
Data Sharing at the Broad Institute

Anthony Philippakis, MD PhD
- Cardiologist, Harvard Medical School
- Chief Data Officer, Broad Institute

The Challenge of Scalability

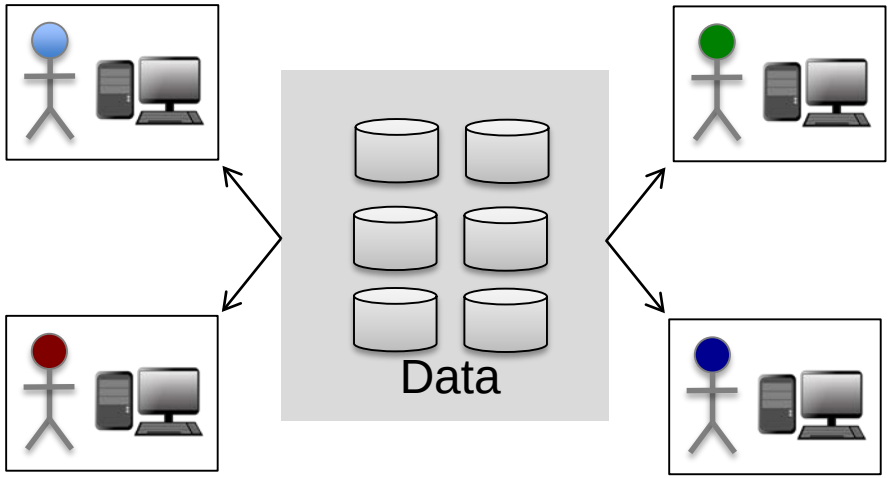


Globally, genomic data doubles every 8 months

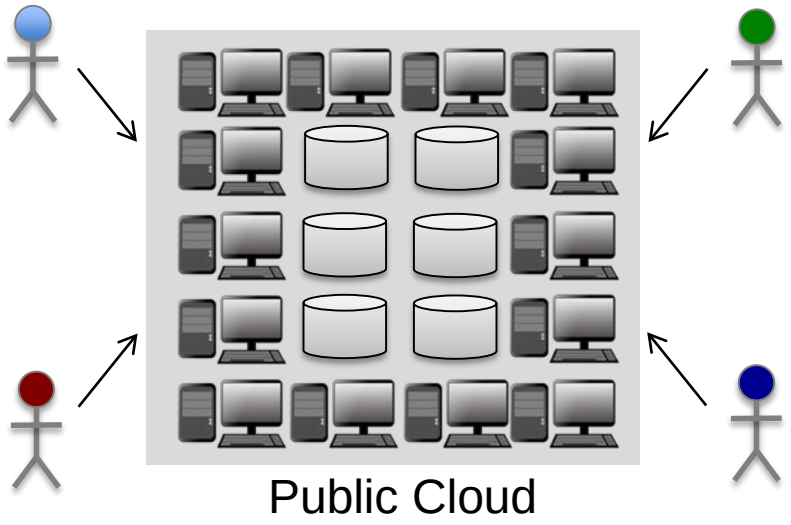


Inverting the Model of Genomic Data Sharing

Traditional Approach: **Bring data to researchers**



Opportunity: **Bring researchers to the data**



Problems

- ***Data sharing = data copying***
- ***Security (data handoffs)***
- ***Huge infrastructure needed***
- ***Siloed compute***

Advantages

- ***Cost***
- ***Threat detection and auditing***
- ***Increased Accessibility***
- ***Shared compute***

NCI Cloud Pilots

FireCloud

An open-source platform for securely managing, sharing and analyzing ***data and tools***.

Initially funded by NCI to host the entire TCGA dataset (2.3PB).

Currently in operational use (ATO granted May 2016), and is the environment where Broad will store and use all of its sequencing data

All of Us: Background

The Precision Medicine Initiative Cohort Program – Building a Research Foundation for 21st Century Medicine

Precision Medicine Initiative (PMI) Working Group Report to the Advisory Committee to the Director, NIH

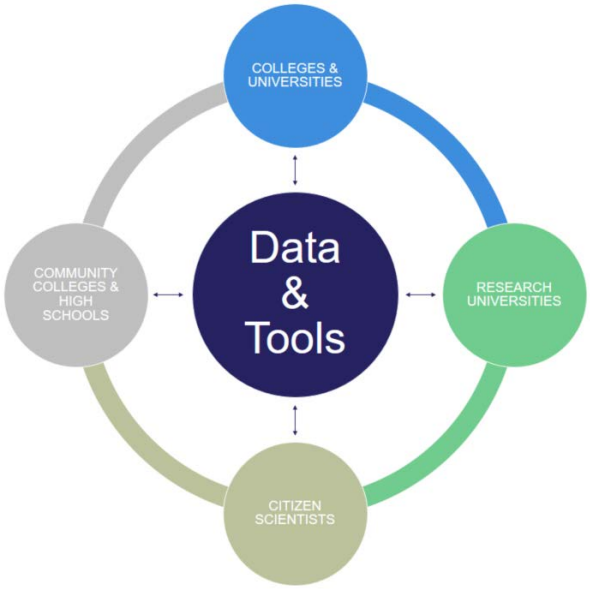
September 17, 2015

- 1 million or more participants
- Longitudinal, re-contactable
- EHR data, biospecimens, baseline exams
- Focus on engagement
- Two methods of enrollment
  - Direct volunteers
  - Healthcare provider organizations

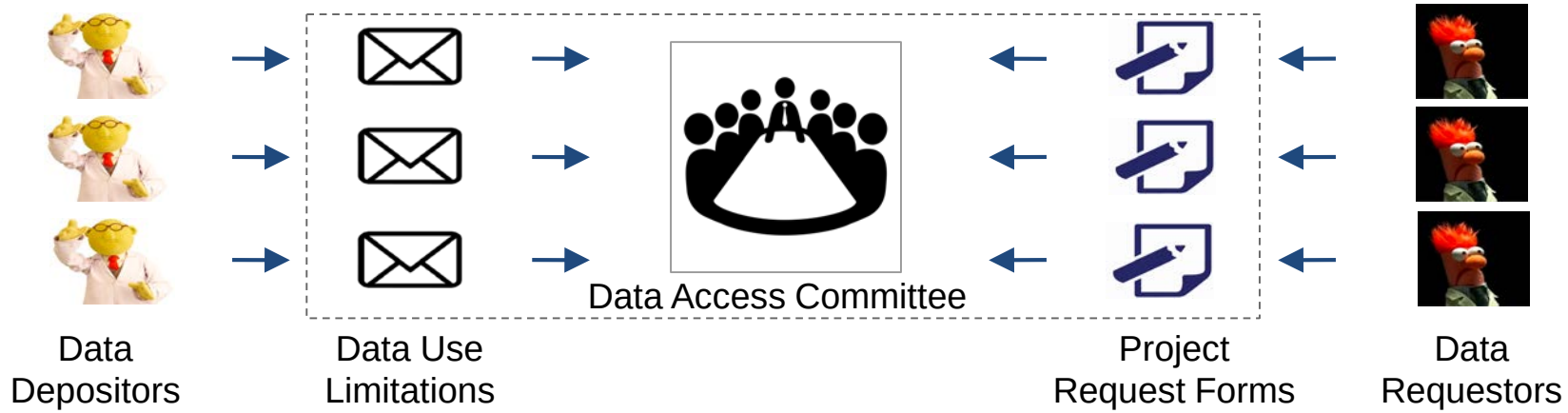
All of Us: Mandate for Innovation

Innovating on Data & Data Access

- All data collected is returned to research participants (including genome sequences)
- Data will be rapidly shared with researchers (All of Us sites do not have privileged access)
- Privacy and security will adhere to the highest standards
- Will create a data platform to expand access and promote utilization.



Our Current Protocol for Data Access



No!

This data is available for cancer research in a non-profit setting.

I am studying Breast cancer at a company.

Our Current Protocol for Data Access

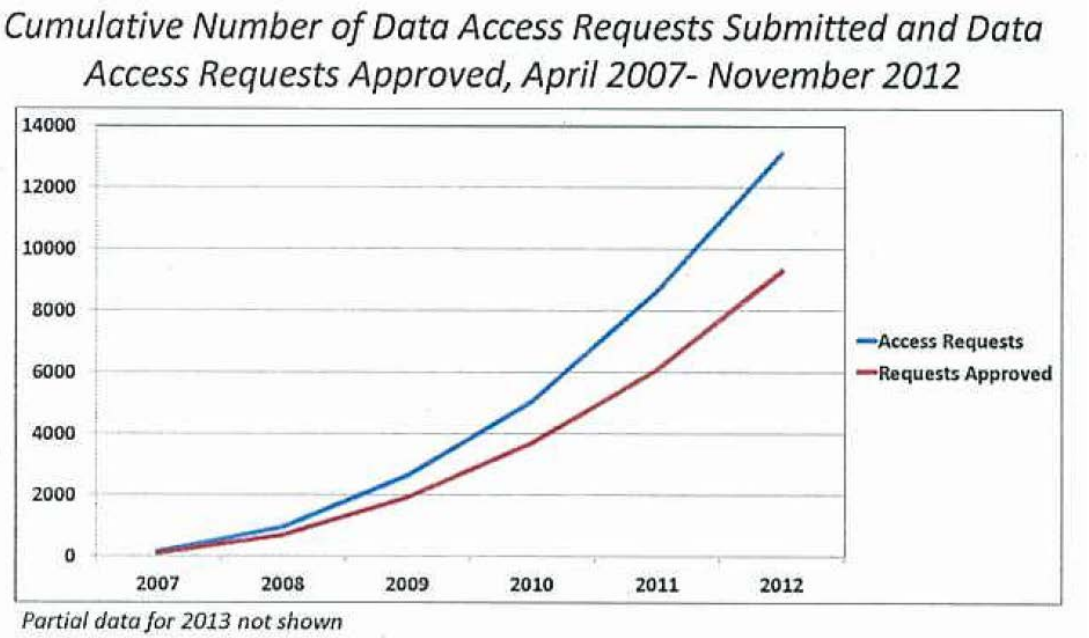
Yes!

This data is available for cancer research in a non-profit setting.

I am studying Breast cancer at a non-profit.

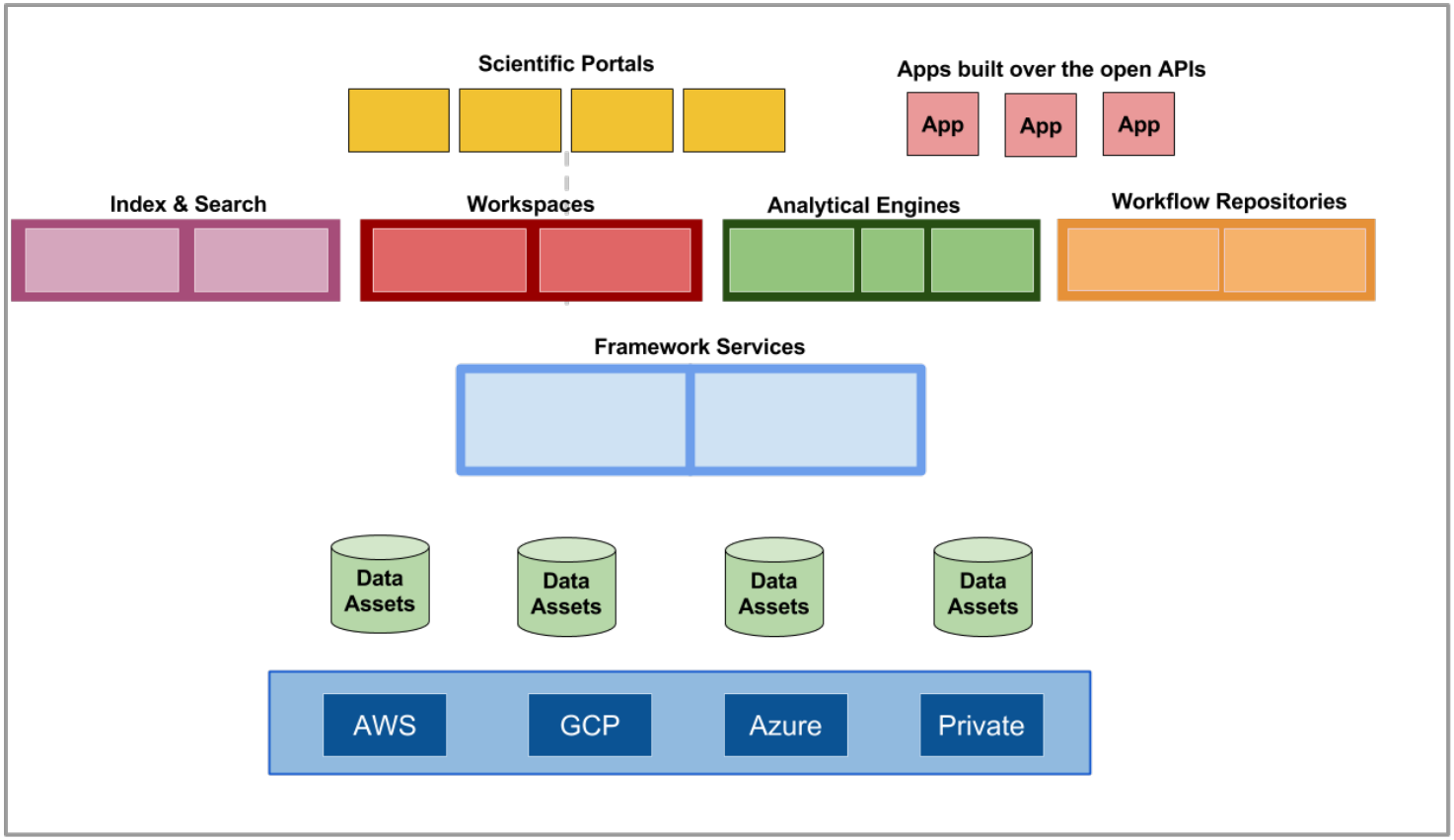
Our Current Protocol for Data Access

Scales Poorly!!
$O(N^2)$



dbGaP at PRIMr 2013

From Data Commons to an Ecosystem



- NCI GDC / Cloud Resources (UChicago / Broad)
- NIH All of Us (Vanderbilt / Broad / Verily)
- CZI HCA Data Platform (UCSC / Broad)

Establishing Data Commons

Robert Grossman
Center for Data Intensive Science
University of Chicago
& Open Commons Consortium

HRA Meeting on Big Data Sharing
September 19, 2017

- Governance
- Data standards
- Funding data commons & assoc. bioinformaticians
- Adding new data types & analysis pipelines
- Enforcement of data sharing

data sharing ≠ data copying

Our focus today

10,000's to 100,000's of individual small datasets and databases

10's to 100's of commons with governance, standards & multiple projects/datasets

Data repositories for small studies and datasets

cos.io, re3data.org
in Session 1

Data commons

Six data common platforms
in Session 3

Key Issues

- Data sharing is not as simple as data copying
- Governance
- Data standards
- It is important to fund both the commons and the bioinformaticians who power it
- Data commons don’t care about the disease, as long as the support the required data types (genomic, clinical, imaging, wearable).
- Each disease doesn’t need to build their own commons (this is called multi-tenancy.

You don’t have to spend ten millions of dollars building a data commons from scratch, because that hard work has already been done.

Talk to several of the data commons service providers from Session 3. Some compete and some are complementary.

You do have to set up and operate a governance structure, establish data standards, fund the bioinformaticians to clean, format & submit the data and do hard work to enforce open data.

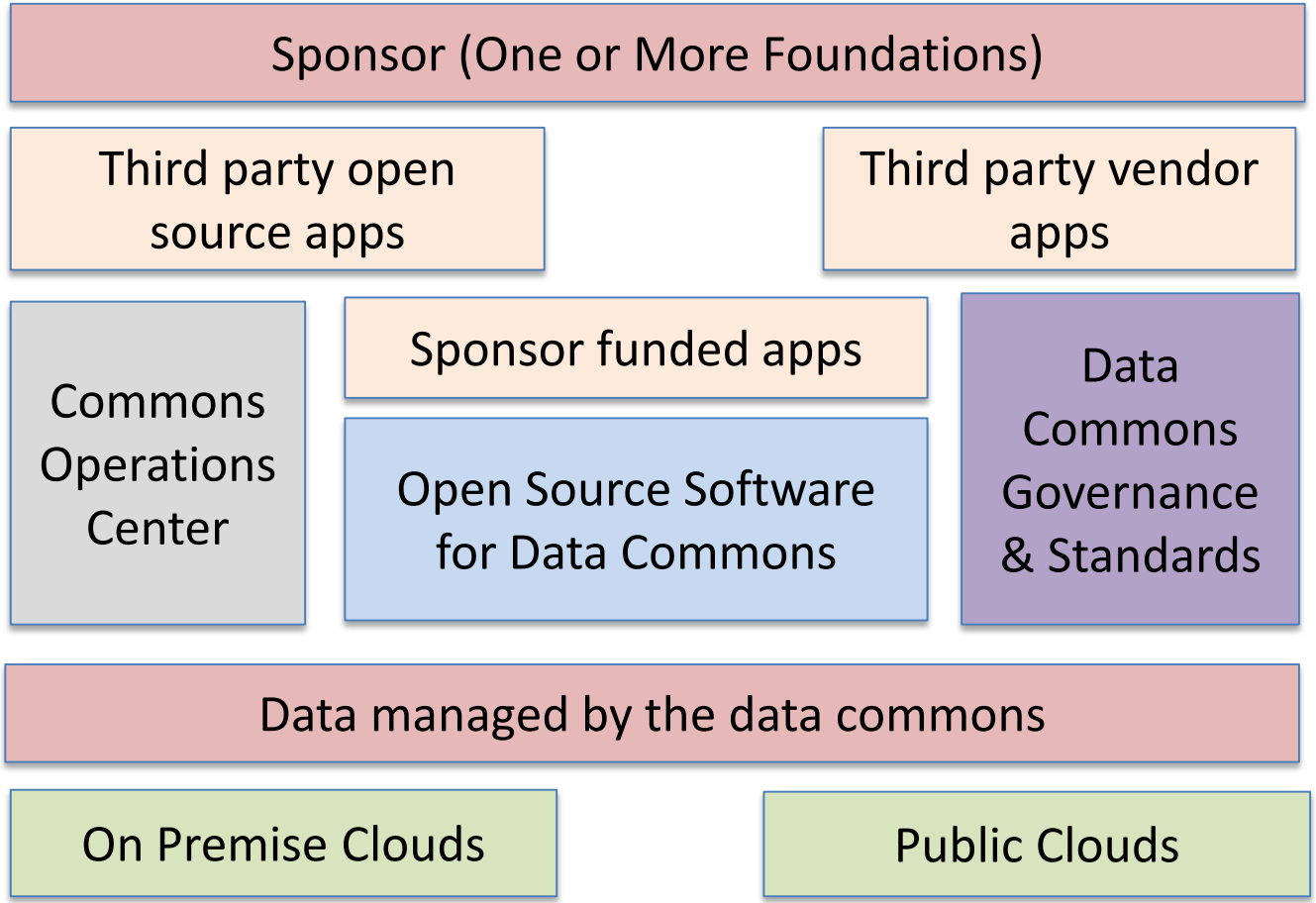
How Do We Organize?

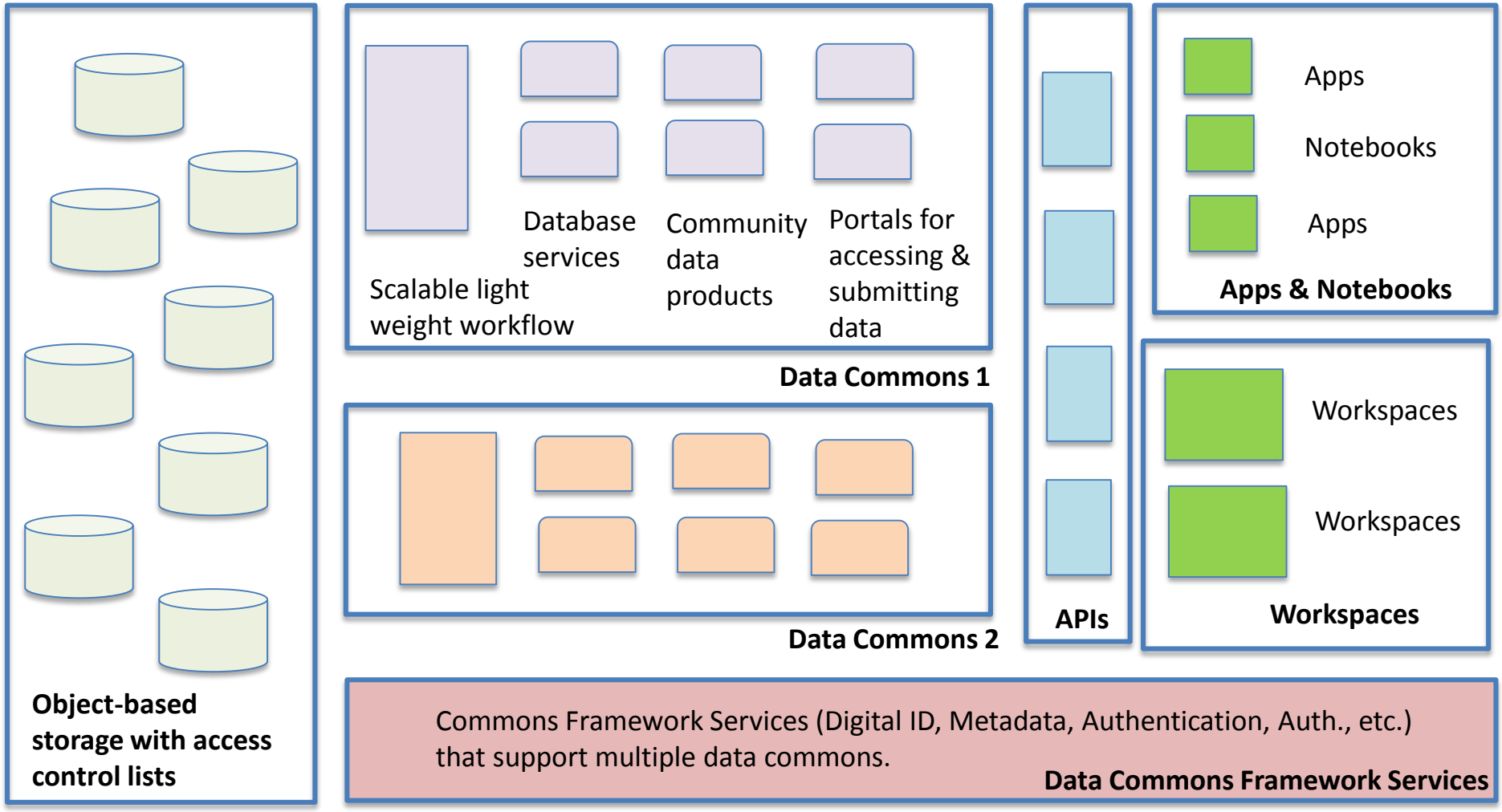
- By foundation?
- By disease?
- By data type (genomic, proteomic, imaging, etc.)
- By broad area (brain, cancer, etc.)
- Some other way
- We will come back to this in the action items.

Sharing Data with Data Commons – the Main Steps

1. **Require data sharing.** Put data sharing requirements into your grant agreements. We can work out some common language.
2. **Build a commons.** Lead, co-lead or join a data commons, fund it, and develop an operating plan, governance structure, and a sustainability plan.
3. **Populate the commons.** Provide resources to your researchers to get the data into data commons.
4. **Interoperate with other commons.** Fund your commons developers and operators to interoperate with other commons that can accelerate research discoveries.
5. **Support commons use.** Support applications that ask for support to build apps over commons.

The Components of a Data Commons




Object-based storage with access control lists
Scalable light weight workflow
Database services
Community data products
Portals for accessing & submitting data
Data Commons 1
Data Commons 2
APIs
Apps
Notebooks
Apps
Apps & Notebooks
Workspaces
Workspaces
Workspaces
Commons Framework Services (Digital ID, Metadata, Authentication, Auth., etc.) that support multiple data commons.
Data Commons Framework Services

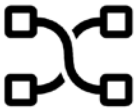
Building the BloodPAC Data Commons (Examplar of Principles 1 & 2)

amazon web services

THE UNIVERSITY OF CHICAGO

OCC

BloodPAC
BLOOD PROFILING ATLAS IN CANCER

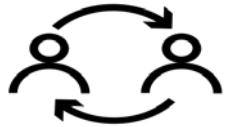


Adapt the data model to your project

Set up & configure the data commons (CSOC)

Put in place the OCC data governance model

Research groups submit data



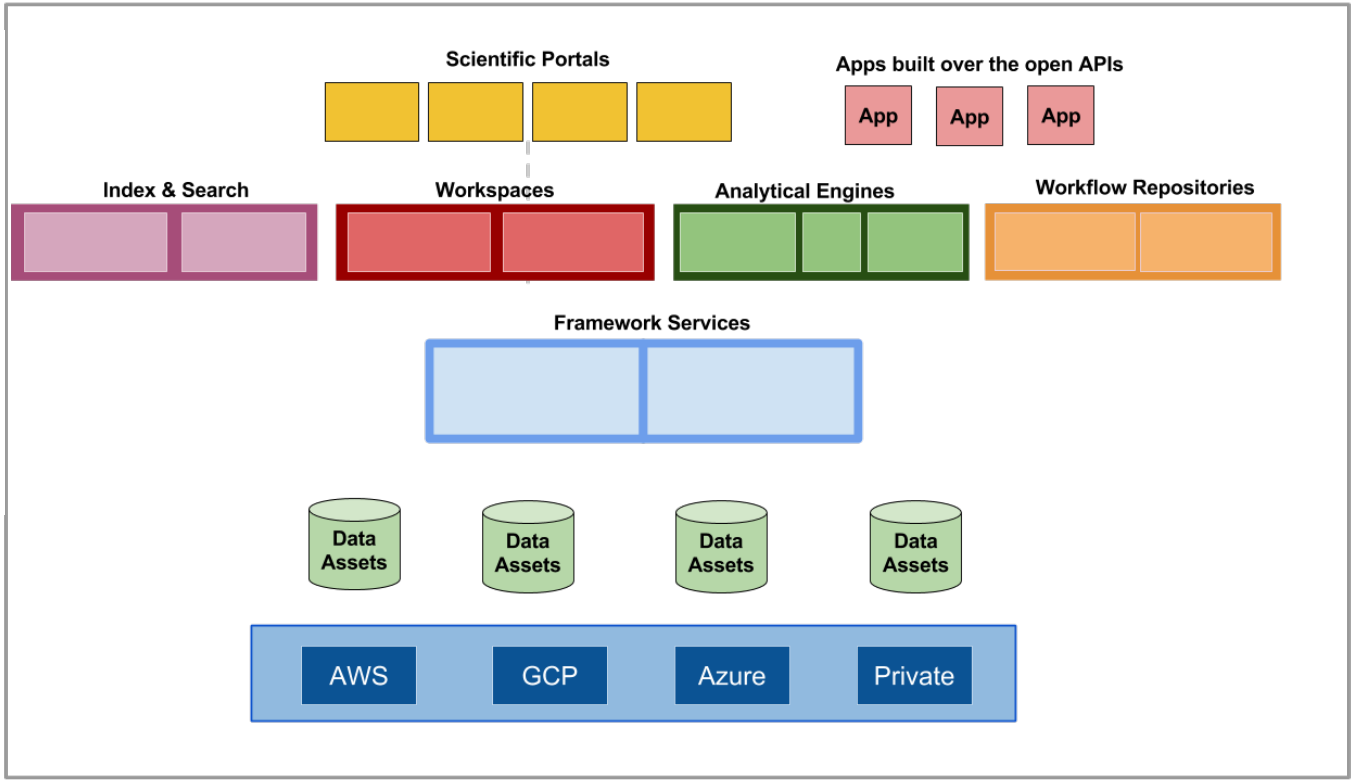
Clean and process the data following the standards

New research discoveries

Researchers use the commons for data analysis

Alliances Committed to Interoperating Large-Scale Commons (Examplar of Principle 3)

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



Databases

- Data repository
- Researchers download data.

Data Clouds

- Supports big data
- Collaborative tools
- Researchers can analyze data (data does **not** have to be downloaded)

Data Commons

- Supports big data
- Collaborative tools
- Researchers can analyze data
- Common data models
- Harmonized data
- Ecosystem of apps

Data Ecosystems
2018 - 2028

Data Commons
2014 - 2024

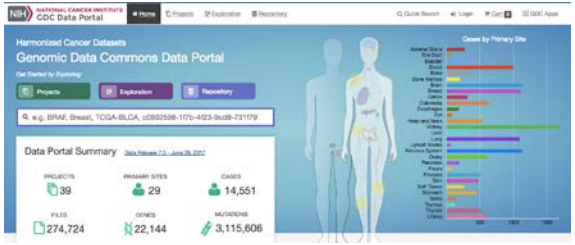
Data Clouds
2010 - 2020

Databases
1982 - present

Bermuda Principles
& Genomic Databases
(e.g. GenBank)
1982 - present

PMC arXiv.org bioRxiv

Open Access Principles for Publications
arXiv, PubMed Central
2010 - present



Chicago Principles
Data Commons
2017 -

Lets debate, draft and sign these by Dec, 2017.

Chicago Principles

1. Require that researchers share the data generated by research that you fund.
2. Foundations should provide the computing infrastructure and bioinformatics resources that is required to support data sharing.
3. The data commons supported by Foundations should themselves share data and interoperate with other data commons.

Four Follow Up Actions

1. **Chicago Principles** for Sharing Research Data funded
2. **Workshops**
   - Workshop 1. How to build a data commons – Administrative, Governance, Sustainability Issues.
   - Workshop 2. How to build a data commons – Technical issues and options.
3. **Partnerships**. We will work to foster partnerships to build and operate data commons.
4. **Ombudsman**. We will establish a POC for data commons who can make introductions, link you with other foundations to create a critical mass of data & keep you doing from doing stupid things..

Questions?