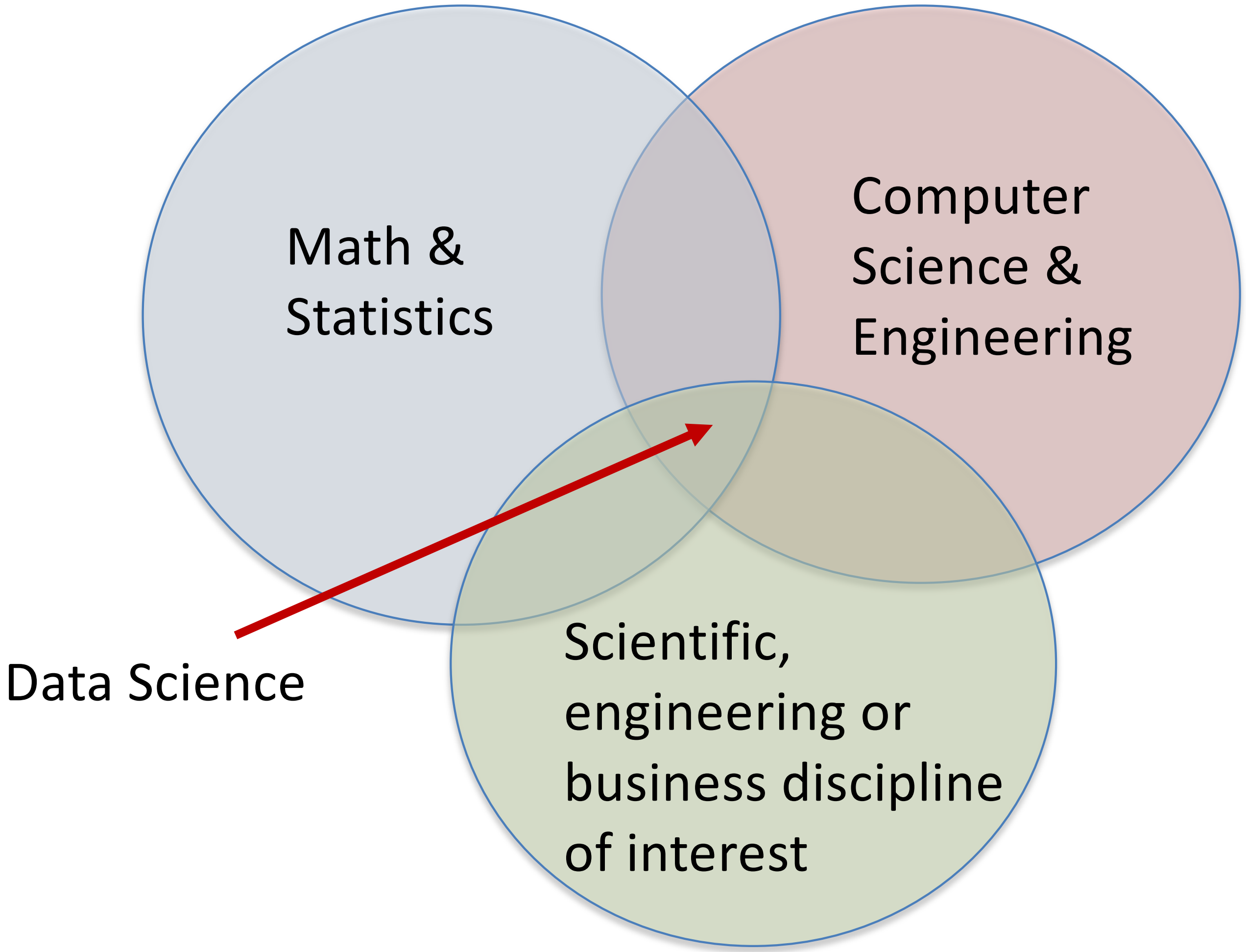
THE UNIVERSITY OF CHICAGO | Center for Translational Data Science

Introduction to the Center for Translational Data Science (CTDS) and Some of the Data Commons It Develops

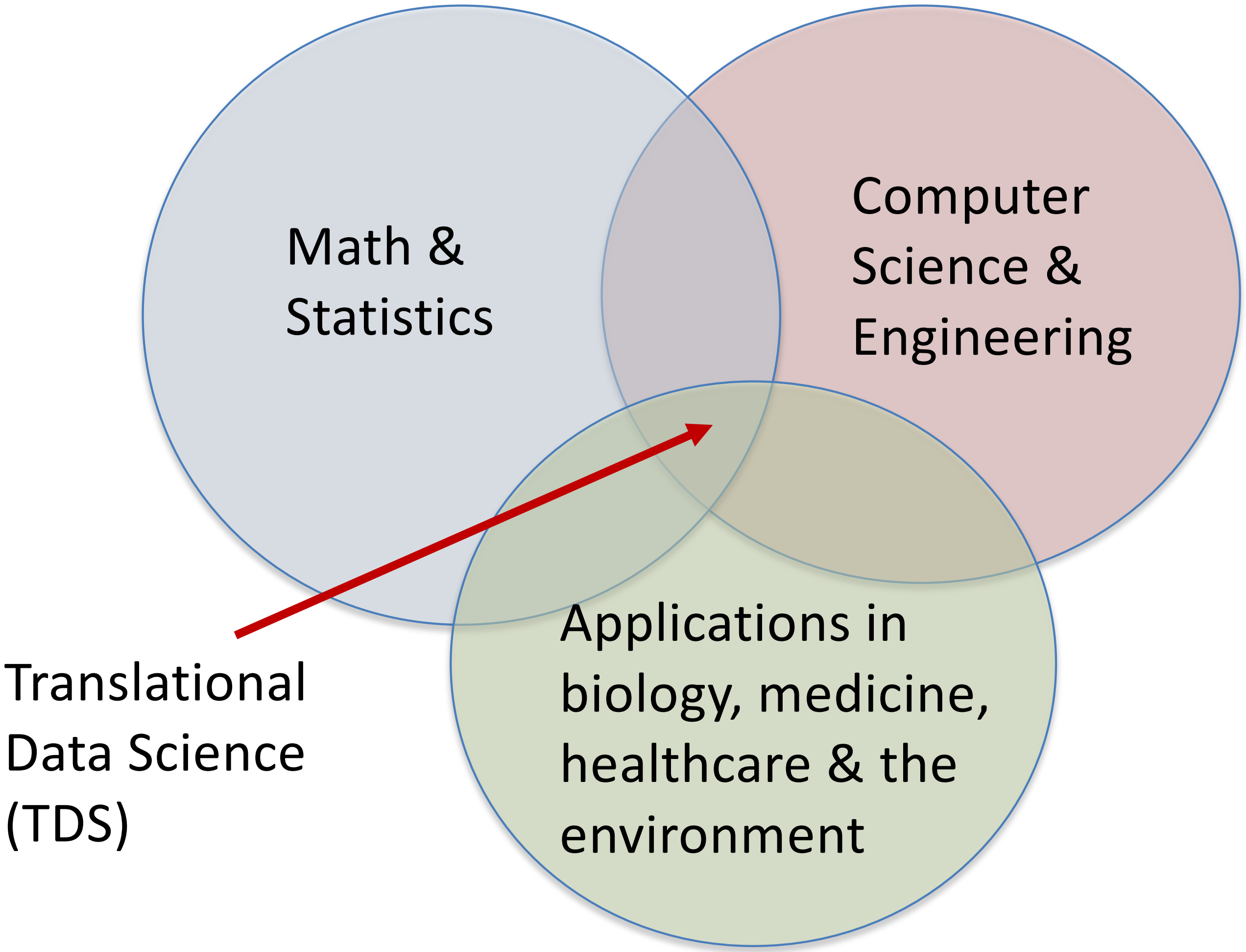
Robert Grossman
Center for Translational Data Science
University of Chicago

January 30, 2019

Part 1: Introduction to the CTDS



- As usual, we view data science as an approach that integrates math/stat, computer science & engineering, and its applications.



Math & Statistics

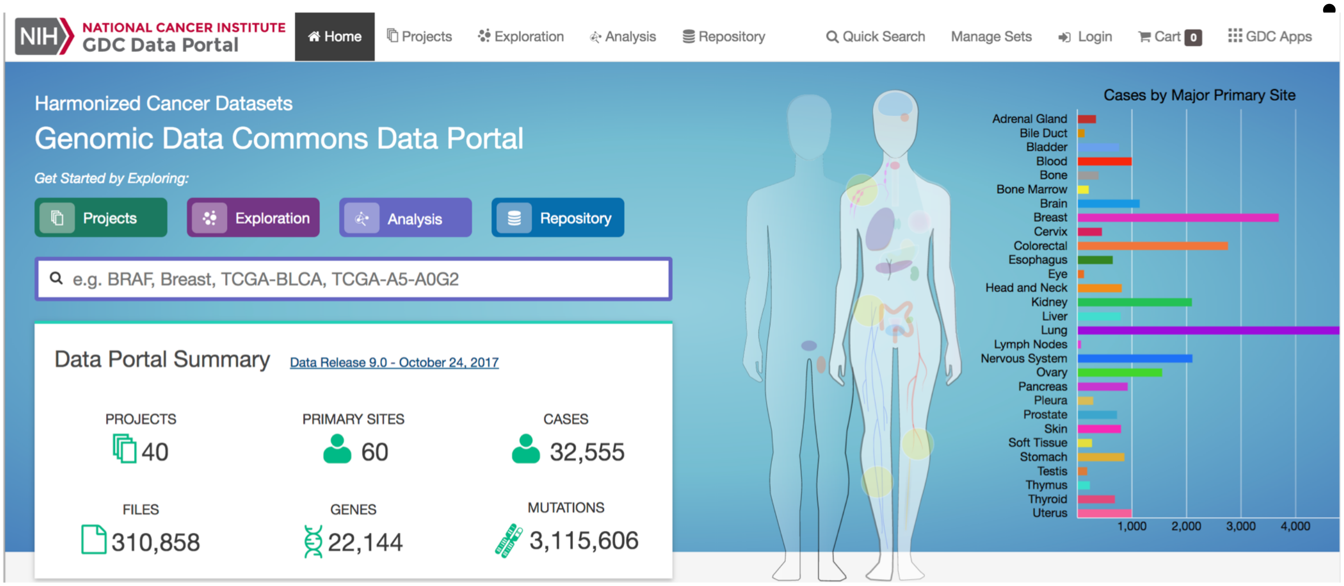
Computer Science & Engineering

Applications in biology, medicine, healthcare & the environment

Translational Data Science (TDS)

- Translation is about the human or societal impact of the data science.
- The challenge is **translating** a discovery in data science to have an impact.
- Translational data science is the discipline that supports this challenge.

NCI Genomic Data Commons*



Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

Data Portal Summary Data Release 9.0 - October 24, 2017

PROJECTS	PRIMARY SITES	CASES
40	60	32,555
FILES	GENES	MUTATIONS
310,858	22,144	3,115,606

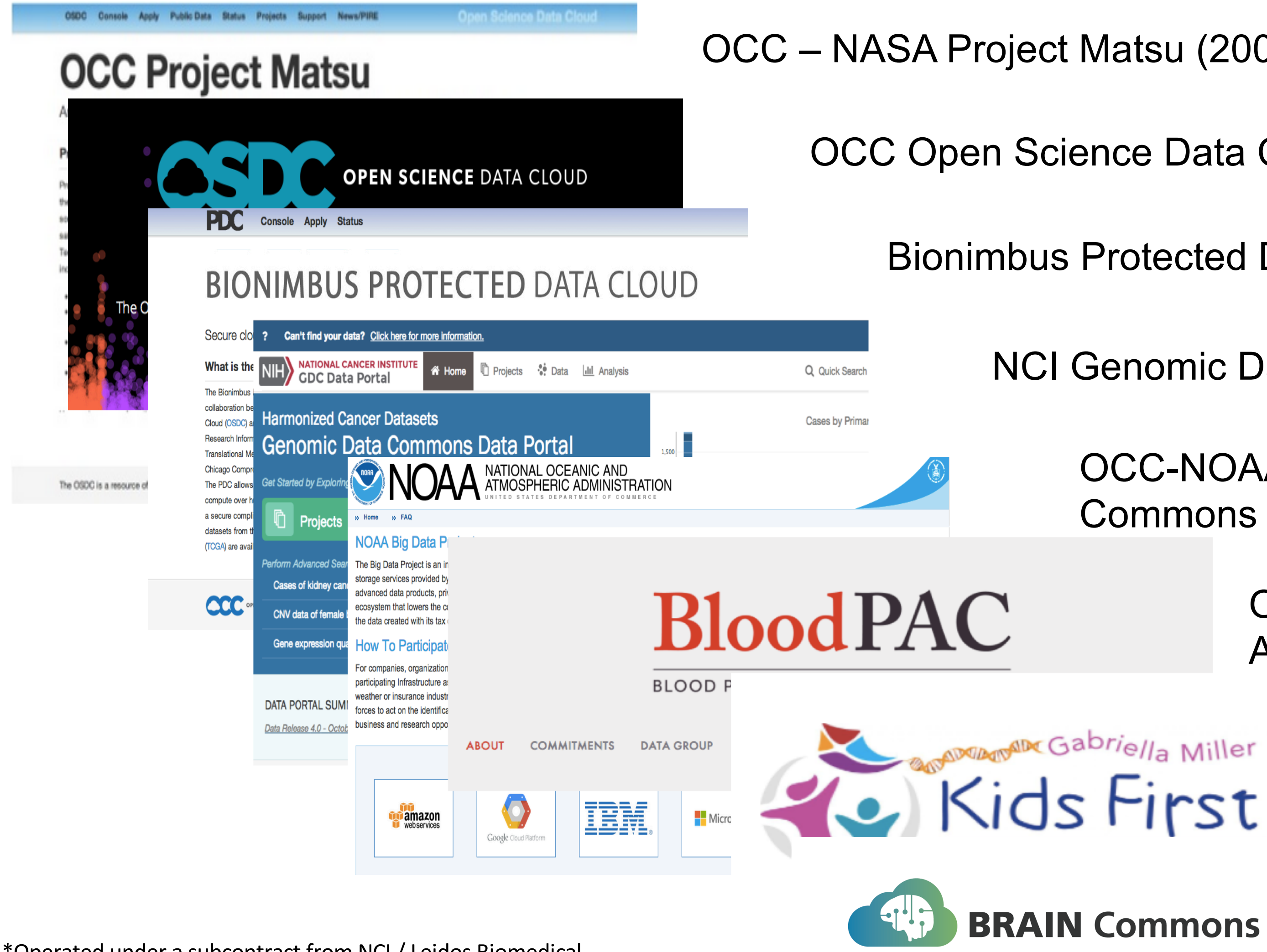
- The GDC makes available over 2.5 PB of data available for access via an API, analysis by cloud resources on public clouds, and downloading.

  In an average month, the GDC is used by over 22,000 users and over 2 PB of data are downloaded.

- The GDC is based upon an open source software stack that can be used to build other data commons.

The GDC consists of a 1) data exploration & visualization portal (DAVE), 2) data submission portal, 3) data analysis and harmonization system system, 4) an API so third party can build applications.

*See: NCI Genomic Data Commons: Grossman, Robert L., et al. "Toward a shared vision for cancer genomic data." New England Journal of Medicine 375.12 (2016): 1109-1112.


OCC Project Matsu
OSDC OPEN SCIENCE DATA CLOUD
PDC Console Apply Status
BIONIMBUS PROTECTED DATA CLOUD
NIH NATIONAL CANCER INSTITUTE GDC Data Portal
Home Projects Data Analysis Quick Search
Harmonized Cancer Datasets
Genomic Data Commons Data Portal
NOAA NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
NOAA Big Data P
How To Participat
BloodPAC
ABOUT COMMITMENTS DATA GROUP
Gabriella Miller Kids First
BRAIN Commons

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)

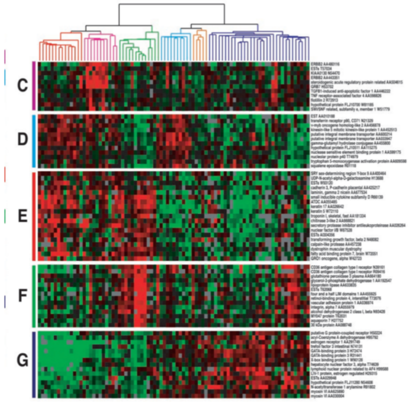
OCC Blood Profiling Atlas in Cancer (2017)

Kids First Data Resource (2017)

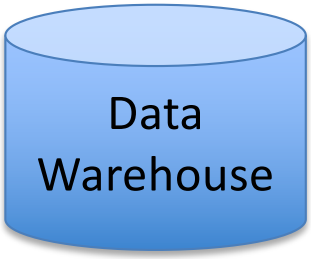
Brain Commons (2017)

Gen3

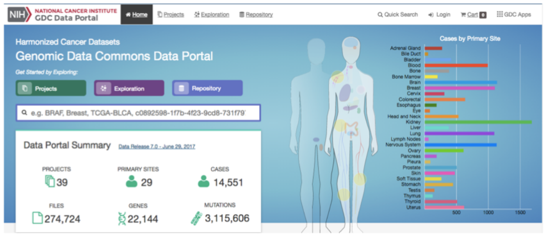
*Operated under a subcontract from NCI / Leidos Biomedical to the University of Chicago with support from the OCC.



Databases organize the data for a project or department.



Data warehouses organize the data for an **organization**



Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

e.g. BRAF, Breast, TCGA-BLCA, c0892598-1f7b-4f23-9cd8-731f79

Data Portal Summary Data Release 7.0 - June 29, 2017

PROJECTS	PRIMARY SITES	CASES
39	29	14,551
FILES	**GENES**	**MUTATIONS**
274,724	22,144	3,115,606

Data commons organize the data for a research **discipline** or field

Project

Databases

1982 - present

- Data repository
- Data catalogs
- Download data

(Virtual) Organization

OSDC OPEN SCIENCE DATA CLOUD

A Petabyte-scale Scientific Community Cloud

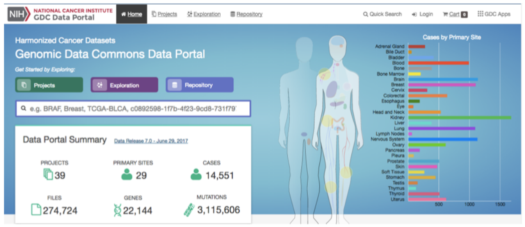
The OSDC enables scientific researchers to easily manage, share, and analyze large datasets.

Data Clouds

2010 - 2020

- Supports large data & data intensive computing with **cloud computing**
- Researchers can analyze data with collaborative tools (**workspaces**) – so data does **not** have to be downloaded)

Discipline



NIH NATIONAL CANCER INSTITUTE GDC Data Portal

Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

e.g. BRAF, Breast, TCGA-BLCA, c0892598-1f7b-4f23-9cd8-731f79

Data Portal Summary Data Release 7.0 - June 29, 2017

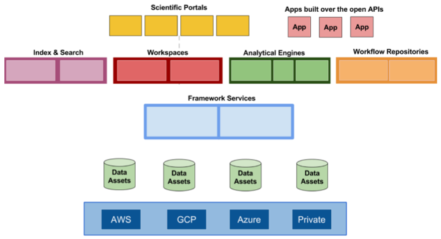
PROJECTS	PRIMARY SITES	CASES
39	29	14,551
FILES	GENES	MUTATIONS
274,724	22,144	3,115,606

Data Commons

2014 - 2024

- Supports large data
- Workspaces
- **Common data models**
- **Core data services**
- **Data & Commons Governance**
- **Harmonized data**
- **Data sharing**
- **Reproducible research**

Multi-Discipline



Data Ecosystems

2018 - 2028

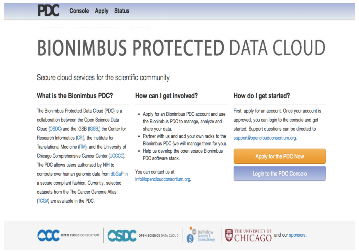
- Interoperates **multiple data commons, databases, knowledge bases**, and other resources
- Supports **ecosystem of commons, portals, notebooks, applications & simulations** across multiple disciplines

Center for Translational Data Science – Selected Firsts

OCC-NASA Project Matsu

First cloud-based processing of satellite images

2009



PDC Console Apply Status

BIONIMBUS PROTECTED DATA CLOUD

Secure cloud services for the scientific community

What is the Bionimbus PDC?

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

How can I get involved?

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

You can contact us at info@opencloudconsortium.org.

How do I get started?

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

Apply for the PDC Now

Login to the PDC Console

OPEN CLOUD CONSORTIUM | OSDC OPEN SCIENCE DATA CLOUD | Institute for Genomics & Systems Biology | THE UNIVERSITY OF CHICAGO and our sponsors.

Bionimbus Protected Data Cloud

First data cloud to earn NIH Trusted Partner status & to operate at FISMA Moderate

2013

NCI Genomic Data Commons

First genomic data commons

2016

BRAIN Commons

2017

NCI DCF – First cancer data ecosystem

2010

OSDC OPEN SCIENCE DATA CLOUD

A Petabyte-scale Scientific Community Cloud

The OSDC enables scientific researchers to easily manage, share, and analyze large datasets.

NEW USERS

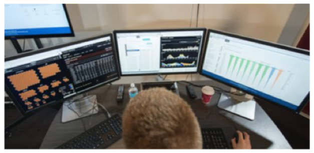
RETURNING USERS

Open Science Data Cloud

First petabyte-scale data cloud (2011)

Open Commons Consortium

First set of open legal and government agreements



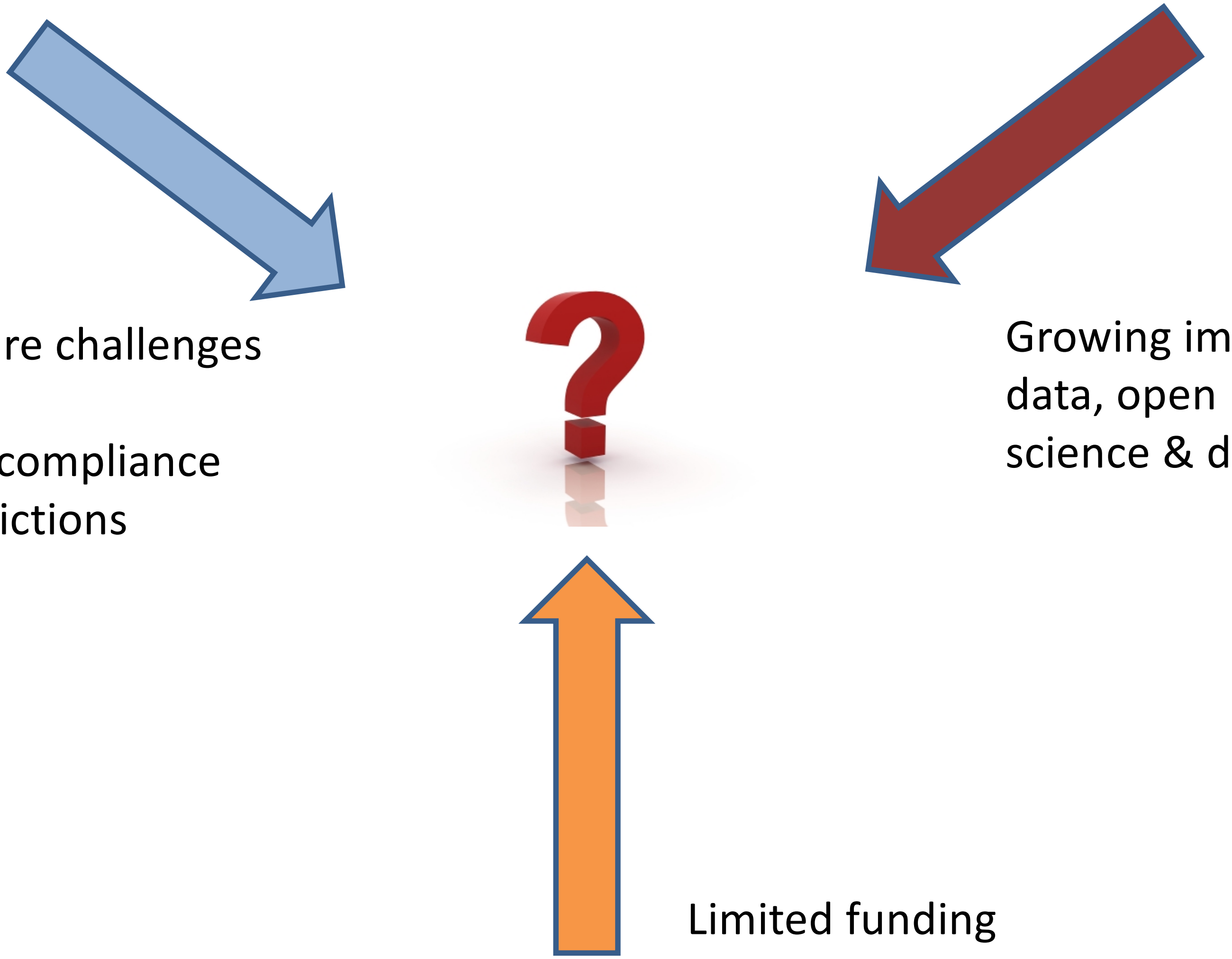
First Commons Services Operations Center (CSOC)

OCC-NOAA

First environmental data commons

2018

2. Data Commons

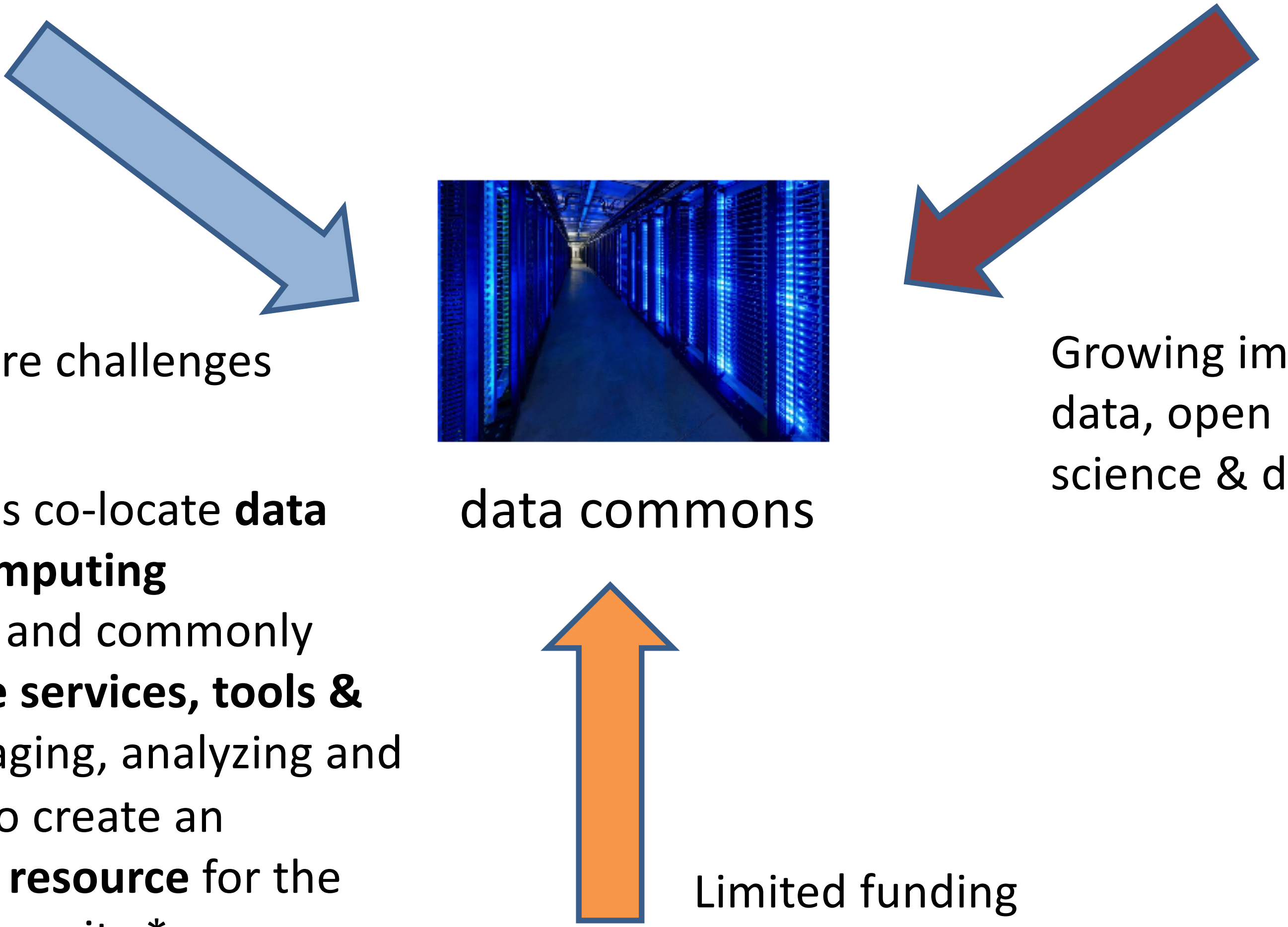


IT infrastructure challenges

- Data size
- Security & compliance
- Policy restrictions

Growing importance of open data, open reproducible science & data ecosystems

Limited funding



IT infrastructure challenges

data commons

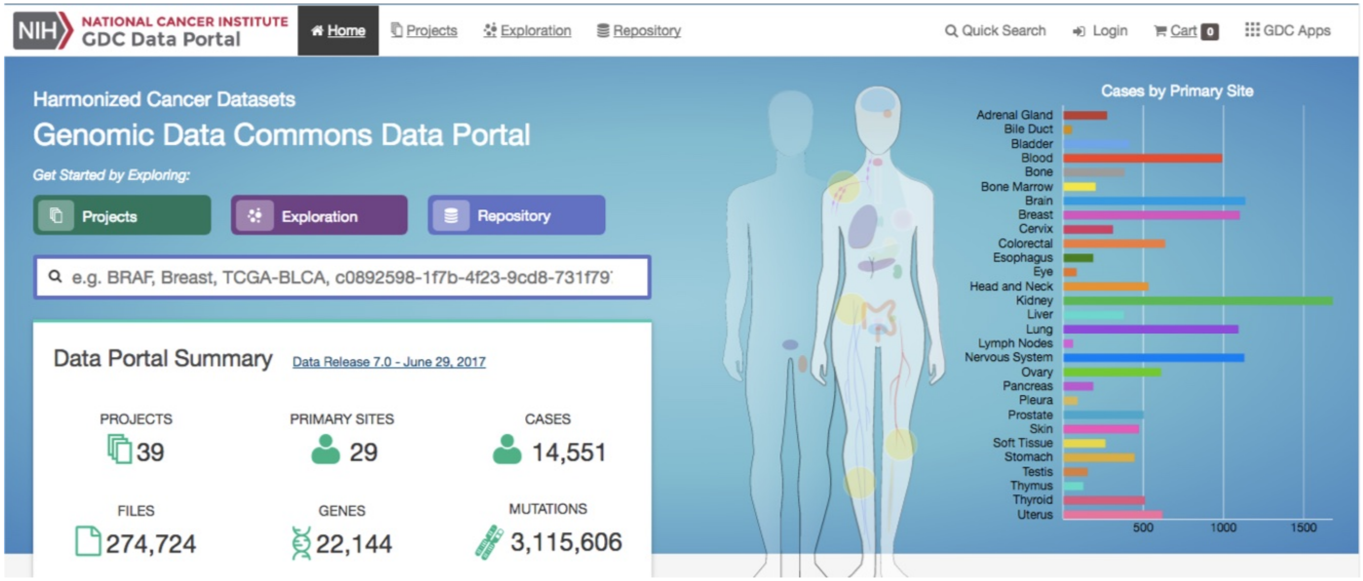
Growing importance of open data, open reproducible science & data ecosystems

Limited funding

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

*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 a University of Chicago data center.

NCI Genomic Data Commons*



Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

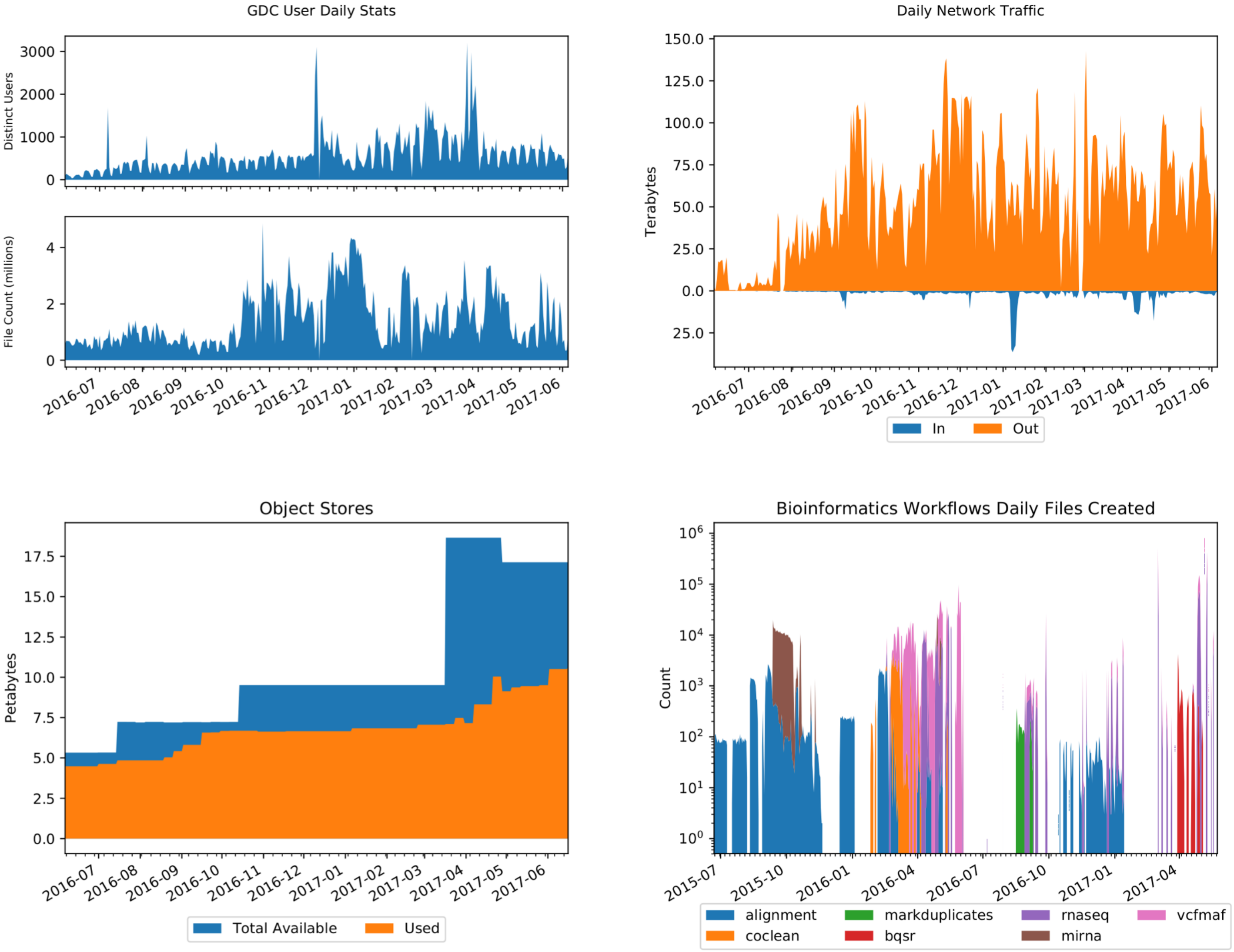
e.g. BRAF, Breast, TCGA-BLCA, c0892598-1f7b-4f23-9cd8-731f79

Data Portal Summary Data Release 7.0 - June 29, 2017

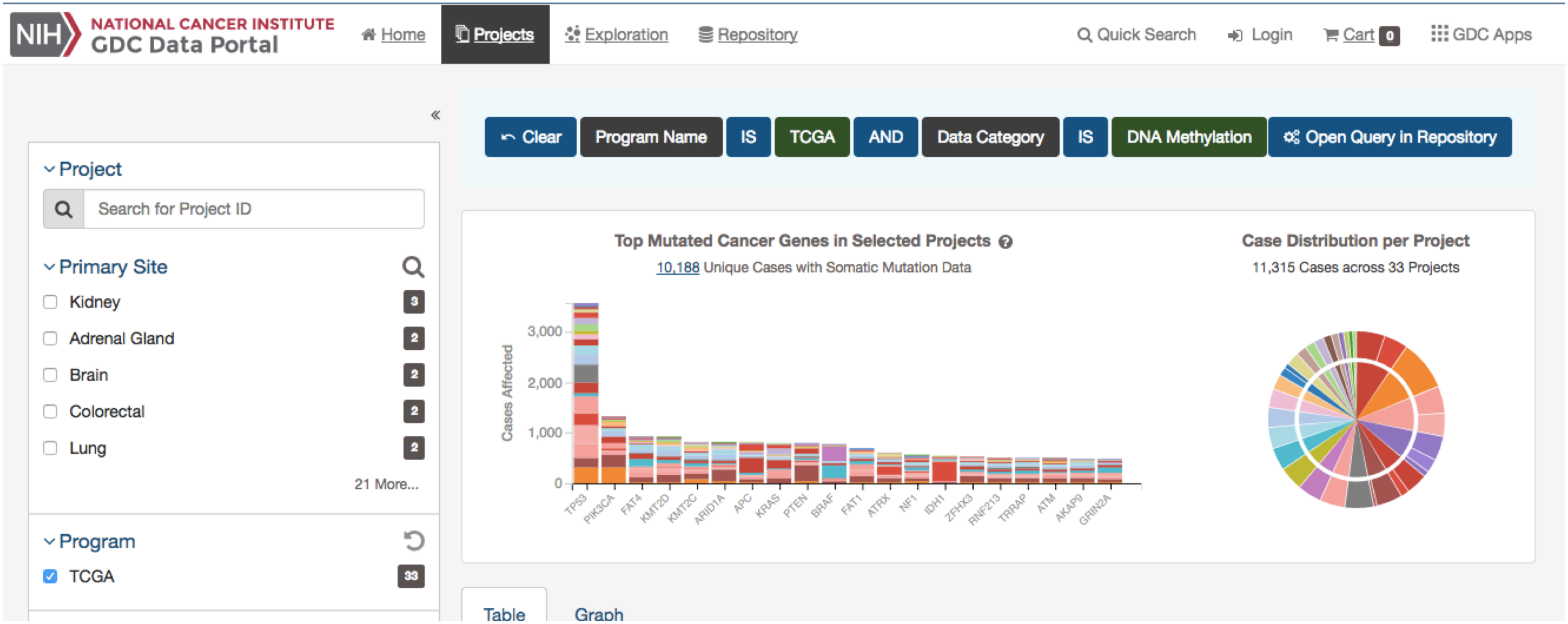
PROJECTS	PRIMARY SITES	CASES
39	29	14,551
FILES	**GENES**	**MUTATIONS**
274,724	22,144	3,115,606

- The GDC was launched in 2016 with over 4 PB of data.
- Used by 1500 - 3000+ users per day and over 36,000 researchers each month.
- Based upon an open source software stack that can be used to build other data commons.

*Source: NCI Genomic Data Commons: Grossman, Robert L., et al. "Toward a shared vision for cancer genomic data." New England Journal of Medicine 375.12 (2016): 1109-1112.


GDC User Daily Stats
Distinct Users
3000
2000
1000
0
File Count (millions)
4
2
0
2016-07
2016-08
2016-09
2016-10
2016-11
2016-12
2017-01
2017-02
2017-03
2017-04
2017-05
2017-06
Daily Network Traffic
Terabytes
150.0
125.0
100.0
75.0
50.0
25.0
0.0
25.0
2016-07
2016-08
2016-09
2016-10
2016-11
2016-12
2017-01
2017-02
2017-03
2017-04
2017-05
2017-06
In
Out
Object Stores
Petabytes
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
2016-07
2016-08
2016-09
2016-10
2016-11
2016-12
2017-01
2017-02
2017-03
2017-04
2017-05
2017-06
Total Available
Used
Bioinformatics Workflows Daily Files Created
Count
10^6
10^5
10^4
10^3
10^2
10^1
10^0
2015-07
2015-10
2016-01
2016-04
2016-07
2016-10
2017-01
2017-04
alignment
coclean
markduplicates
bqsr
rnaseq
mirna
vcfmaf

Apps 1 and 2: Data Portals to Explore and Submit Data



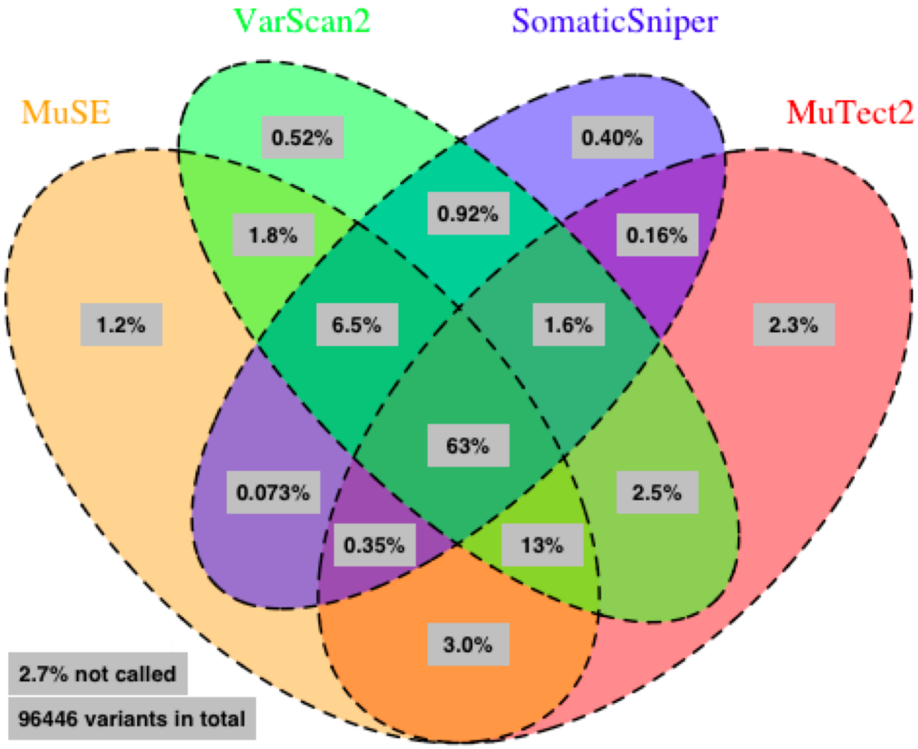
Top Mutated Cancer Genes in Selected Projects

10,188 Unique Cases with Somatic Mutation Data

Case Distribution per Project

11,315 Cases across 33 Projects

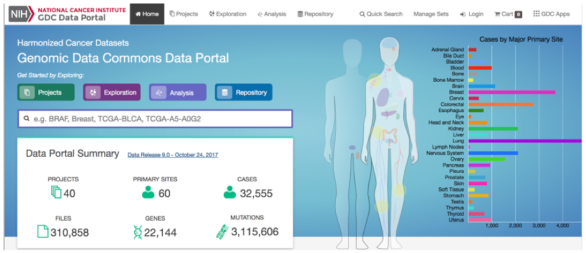
App 3: Analysis & Harmonization of all Submitted Data with a Common Set of Bioinformatics Pipelines



- MuSE
  (MD Anderson)
- VarScan2 (Washington Univ.)
- SomaticSniper
  (Washington Univ.)
- MuTect2
  (Broad Institute)

Source: Zhenyu Zhang, et. al. and the GDC Project Team, Uniform Genomic Data Analysis in the NCI Genomic Data Commons, to appear.

System 4: An API to Support User Defined Applications and Notebooks to Create a Data Ecosyst



Harmonized Cancer Datasets

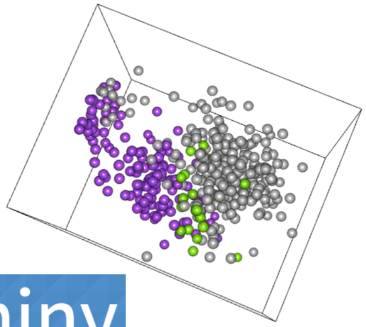
Genomic Data Commons Data Portal

Data Portal Summary — Data Release 9.0 - October 24, 2017

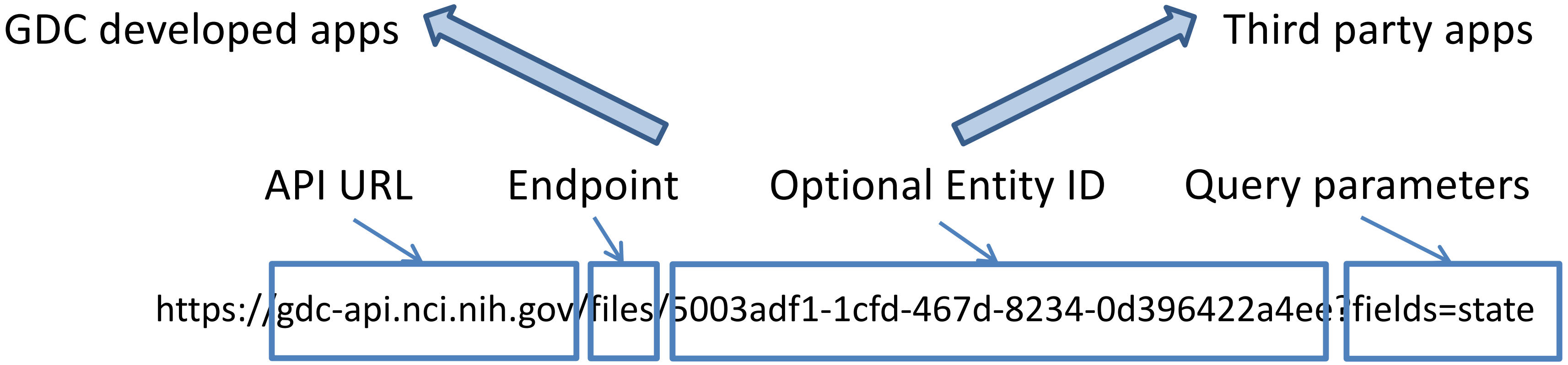
PROJECTS	PRIMARY SITES	CASES
40	60	32,555
FILES	GENES	MUTATIONS
310,858	22,144	3,115,606

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

jupyter



Shiny
by RStudio



GDC developed apps

Third party apps

API URL

Endpoint

Optional Entity ID

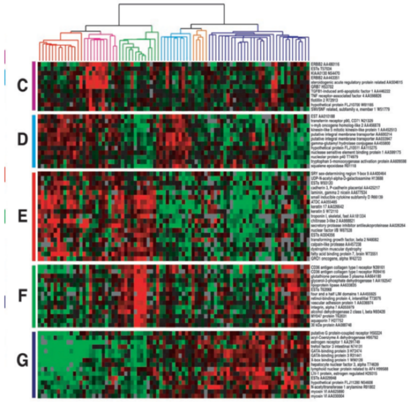
Query parameters

https://gdc-api.nci.nih.gov/files/5003adf1-1cfd-467d-8234-0d396422a4ee?fields=state

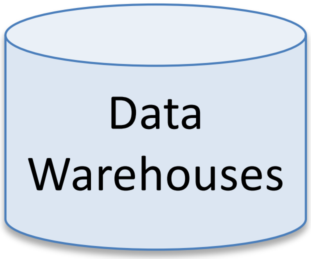
For more about the API, see: Shane Wilson, Michael Fitzsimons, Martin Ferguson, Allison Heath, Mark Jensen, Josh Miller, Mark W. Murphy, James Porter, Himanso Sahni, Louis Staudt, Yajing Tang, Zhining Wang, Christine Yu, Junjun Zhang, Vincent Ferretti and Robert L. Grossman, Developing Cancer Informatics Applications and Tools Using the NCI Genomic Data Commons API, Cancer Research, volume 77, number 21, 2017, pages e15-e18.

Commons

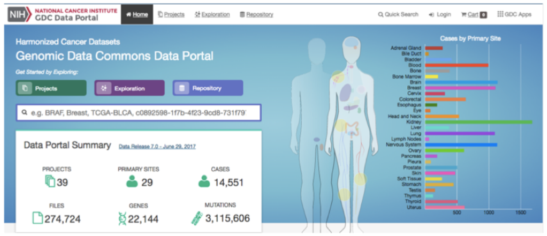
Commons are **resources** that are held in common (and not owned privately) that a group or **community** manage for individual and collective benefit.



Databases organize the data around a project or department.



Data warehouses organize the data for an **organization** (and are enabled by enterprise computing)



Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

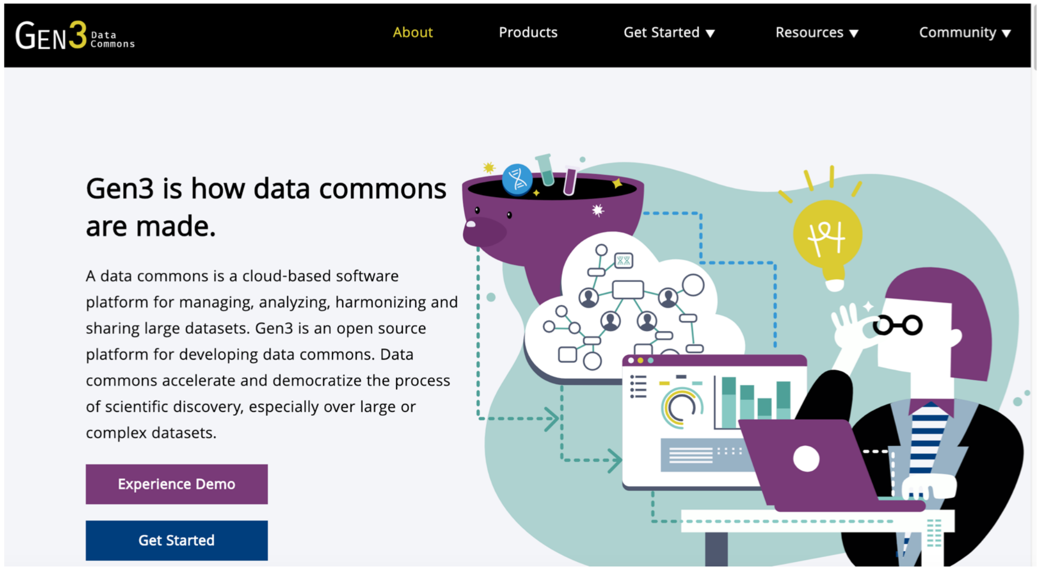
Data Portal Summary Data Release 7.0 - June 29, 2017

PROJECTS	PRIMARY SITES	CASES
39	29	14,551
FILES	GENES	MUTATIONS
274,724	22,144	3,115,606

Data Commons organize the data for a scientific discipline, **community**, or field and support an information platform.

3. Gen3 Data Commons

A Gen3 Data Commons Platform in Six Steps



Gen3 is how data commons are made.

A data commons is a cloud-based software platform for managing, analyzing, harmonizing and sharing large datasets. Gen3 is an open source platform for developing data commons. Data commons accelerate and democratize the process of scientific discovery, especially over large or complex datasets.

Gen3.org
(an open source platform for developing & operating data commons)

1. Define a data model.
2. Use the Gen3 software to auto-generate the data commons and associated API.
3. Import data into the commons using Gen3 import application.
4. Use Gen3 to explore your data and create synthetic cohorts.
5. Use cloud based platforms, tools and workspaces to analyze the synthetic cohosts.
6. Develop your own container-based workflows, applications and Jupyter Notebooks.

(Selected) GEN3 Data Commons

BRAIN Commons

Total Files: 16,733
Total Size: 270.14 GB

GenoMEL
the Melanoma Genetics Consortium

Total Files: 4,008
Total Size: 20.77 TB

BloodPAC
BLOOD PROFILING ATLAS IN CANCER

Total Files: 1,855
Total Size: 295.73 GB

ACCOUNT

Total Files: 1,952
Total Size: 3.77 TB

NHLBI
DATA STAGE

Total Files: 71,368
Total Size: 344.03 TB

Gabriella Miller
Kids First
PEDIATRIC RESEARCH PROGRAM
Data Resource Center

Total Files: 134,531
Total Size: 2.33 PB

NIAID DATA HUB

Total Files: 156,368
Total Size: 304.59 GB

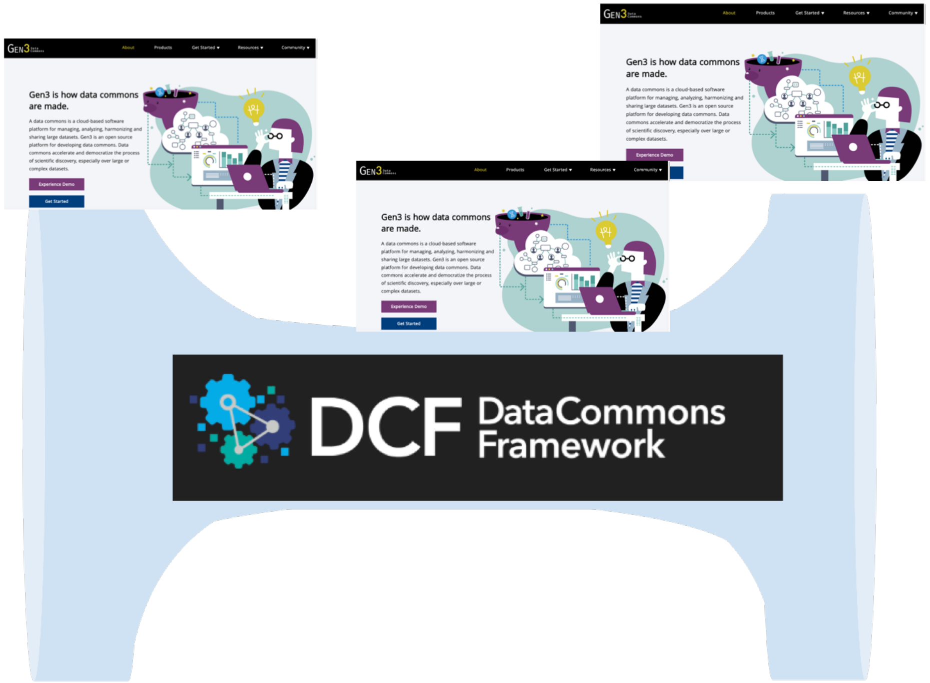
NIH NATIONAL CANCER INSTITUTE
Cancer Research Data Commons

Total Files: 1,688,568
Total Size: 2.2 PB

Environmental
Data Commons

Total Files: 12,317,334
Total Size: 28.93 TB

From Data Commons to Data Ecosystems of Interoperating Data Commons



Gen3 is how data commons are made.

A data commons is a cloud-based software platform for managing, analyzing, harmonizing and sharing large datasets. Gen3 is an open source platform for developing data commons. Data commons accelerate and democratize the process of scientific discovery, especially over large or complex datasets.

1. Build data commons over hosted Data Commons Framework Services
2. Interoperate your data commons with other DCFS compliant data commons.

4. Suggested Guidelines for Foundations for Data Sharing

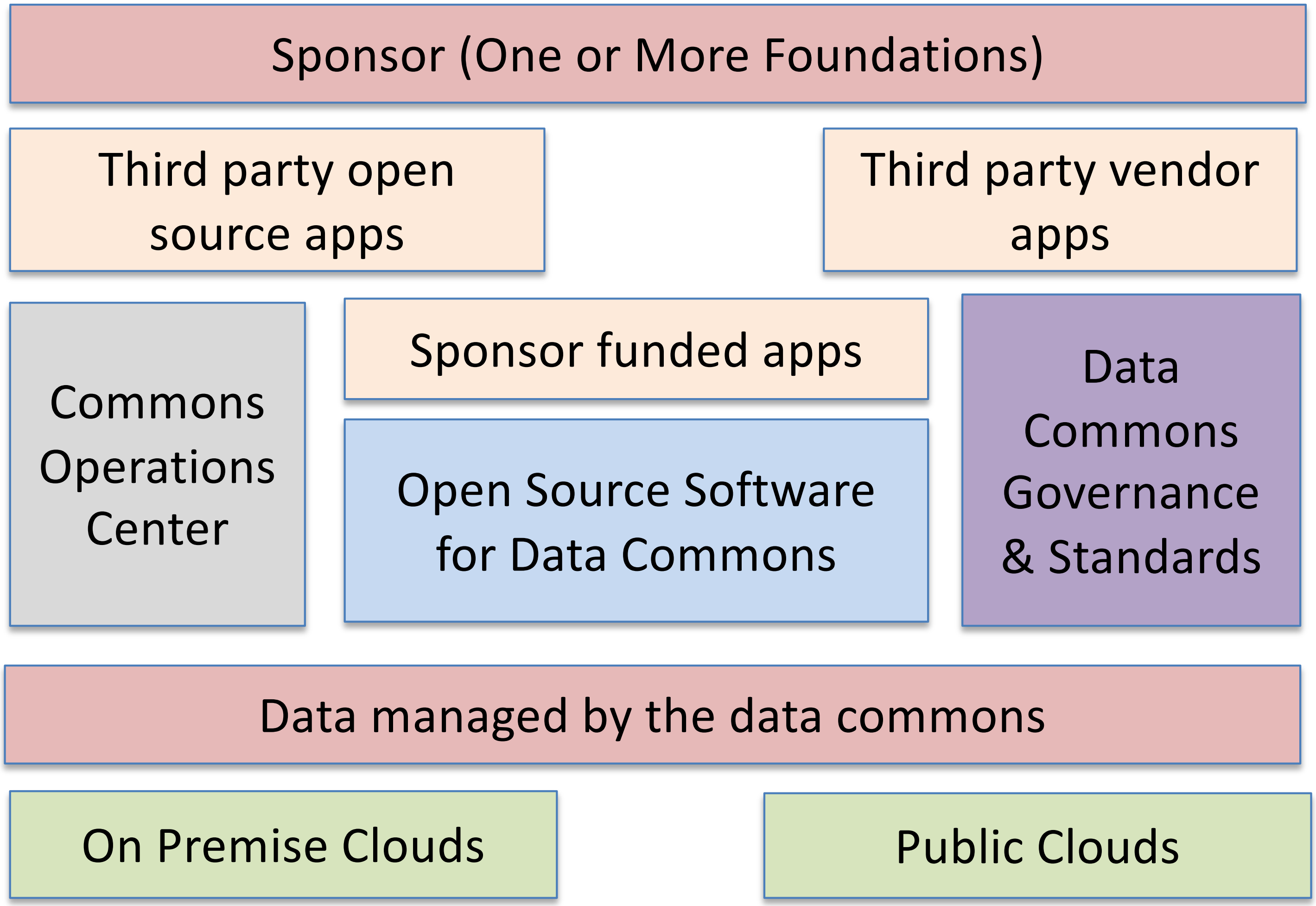
Key Issues

- Data sharing is not as simple as copying data or data & metadata from a repository
- Data Governance and Commons Governance is required
- Data standards
- It is important to fund both the commons and the data scientists that curate the data and build applications for the end users
- Data commons don’t care about the specific data, as long as the support the required data types and applications.
- Each research group doesn’t need to build their own commons (this is called multi-tenancy).

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



Sponsor (One or More Foundations)

Third party open source apps

Third party vendor apps

Commons Operations Center

Sponsor funded apps

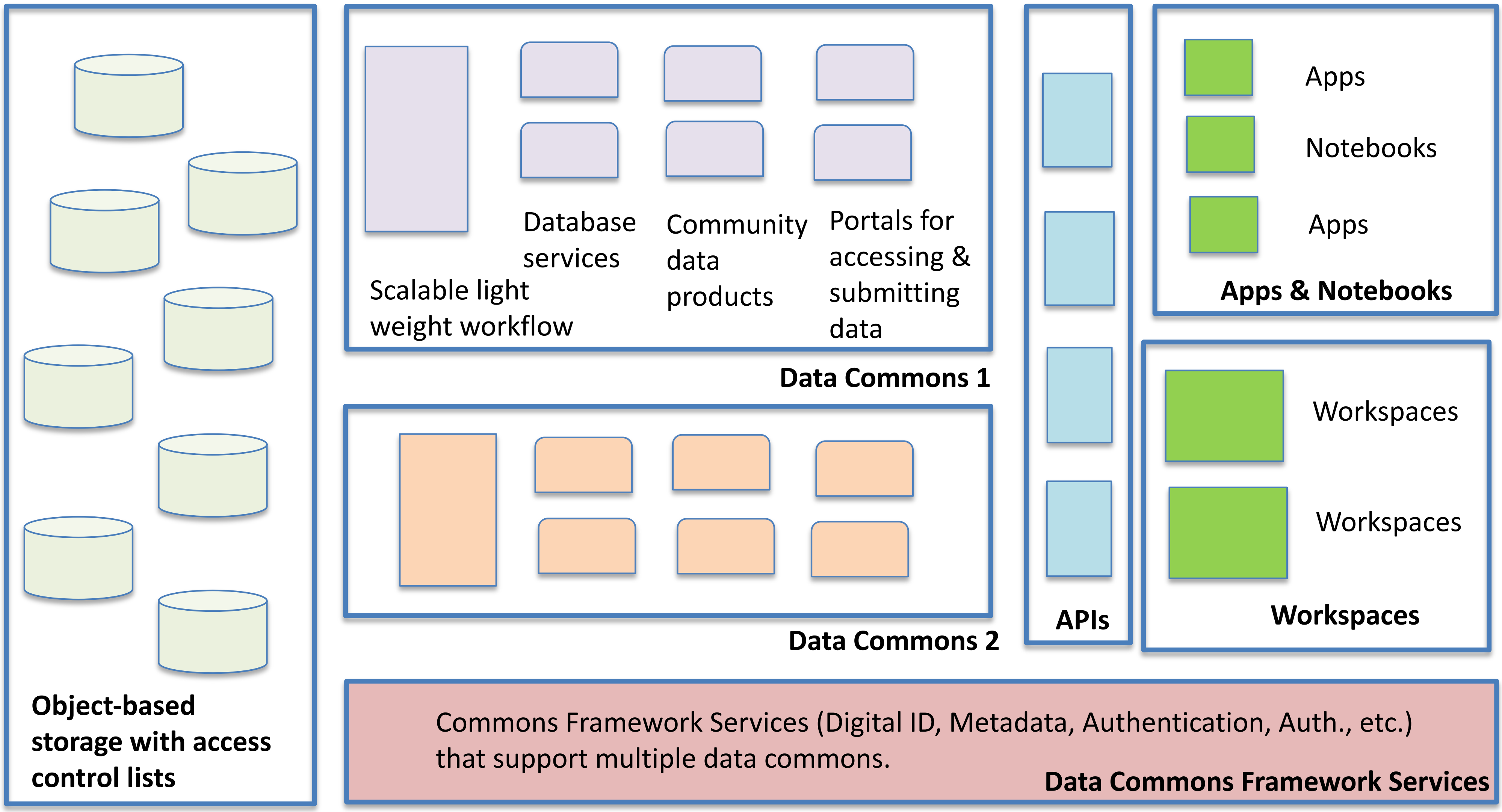
Open Source Software for Data Commons

Data Commons Governance & Standards

Data managed by the data commons

On Premise Clouds

Public Clouds



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

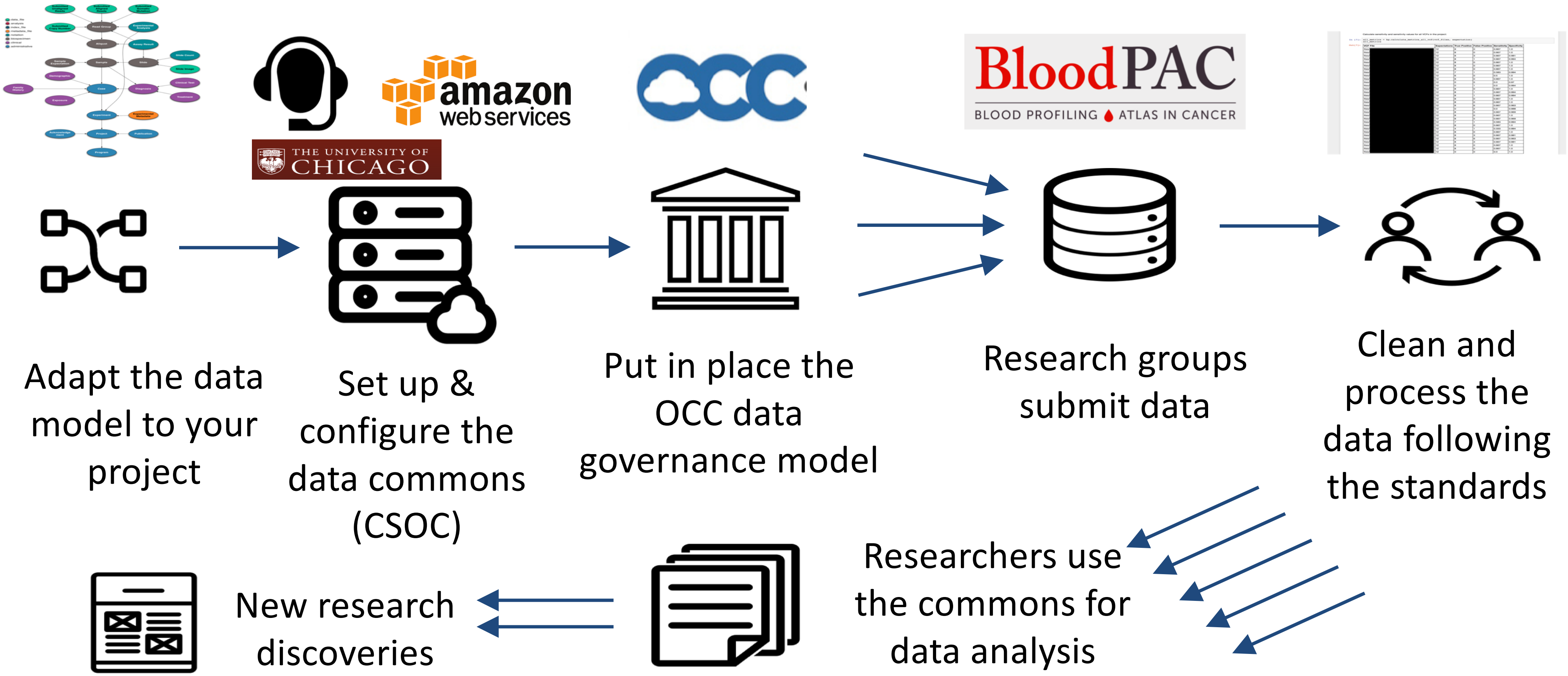
Workspaces

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)



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

Researchers use the commons for data analysis

New research discoveries

Three Principles For Foundations Funding Research

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.

Questions?

rgrossman.com
@bobgrossman

References

For more information:

- To learn more about data commons: Robert L. Grossman, et. al. A Case for Data Commons: Toward Data Science as a Service, Computing in Science & Engineering 18.5 (2016): 10-20. Also https://arxiv.org/abs/1604.02608
- To large more about large scale, secure compliant cloud based computing environments for biomedical data, see: Heath, Allison P., et al. "Bionimbus: a cloud for managing, analyzing and sharing large genomics datasets." Journal of the American Medical Informatics Association 21.6 (2014): 969-975. This article describes Bionimbus Gen1.
- To learn more about the NCI Genomic Data Commons: Grossman, Robert L., et al. "Toward a shared vision for cancer genomic data." New England Journal of Medicine 375.12 (2016): 1109-1112. The GDC was developed using Bionimbus Gen2.
- To learn more about BloodPAC, Grossman, R. L., et al. "Collaborating to compete: Blood Profiling Atlas in Cancer (BloodPAC) Consortium." Clinical Pharmacology & Therapeutics (2017). BloodPAC was developed using the GDC Community Edition (CE) aka Bionimbus Gen3

Contact Information

Robert L. Grossman
rgrossman.com

@BobGrossman
robert.grossman@uchicago.edu

THE UNIVERSITY OF CHICAGO | Center for Translational Data Science

ctds.uchicago.edu