The Road from Data Commons to Data Ecosystems: Challenges, Opportunities, and Emerging Best Practices

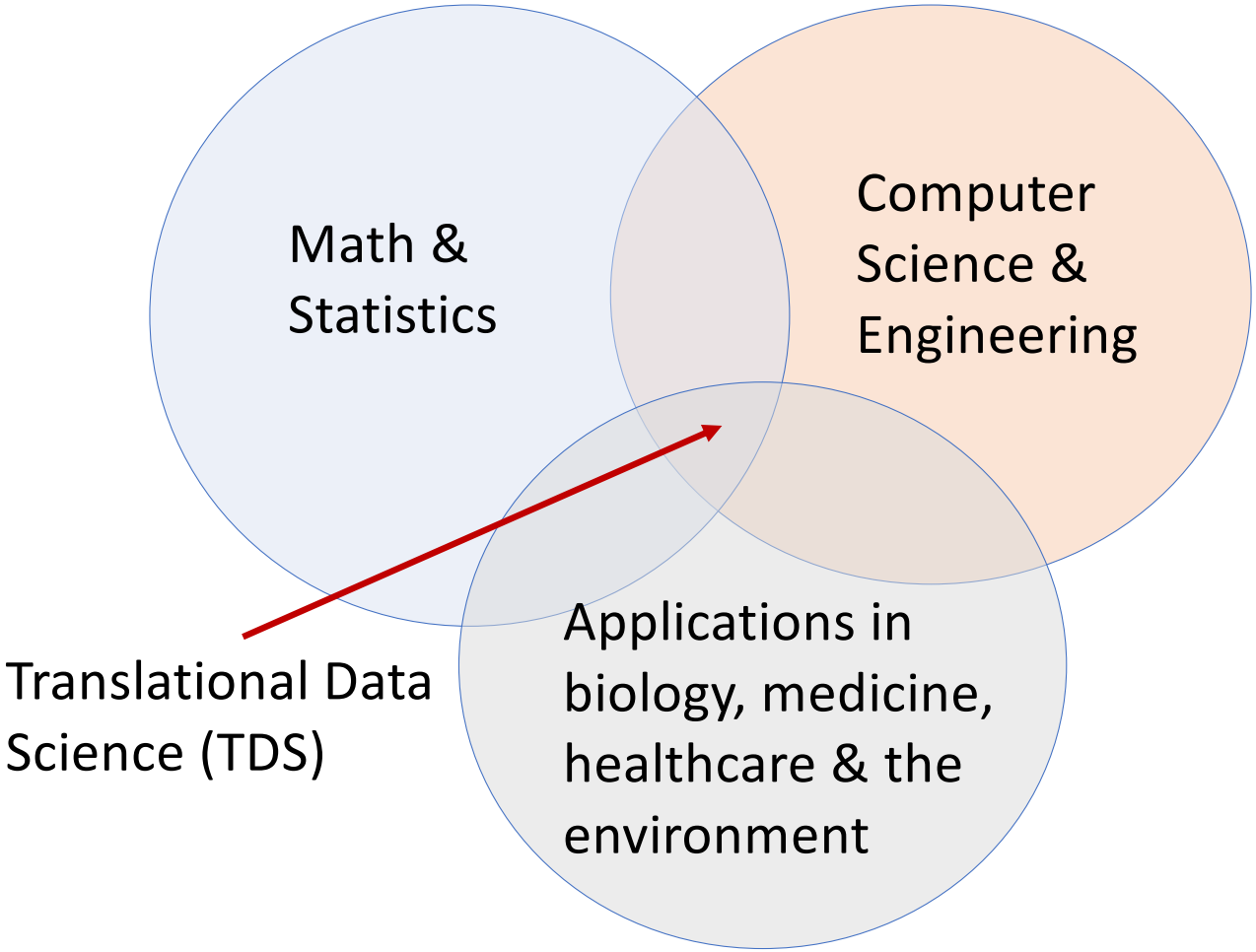
Robert L. Grossman

Center for Translational Data Science

University of Chicago

March 12, 2020

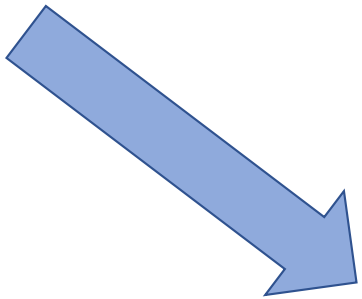
Draft 1.9

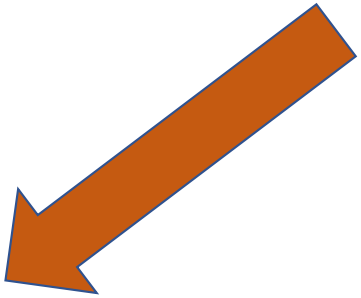


THE UNIVERSITY OF CHICAGO | Center for Translational Data Science

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

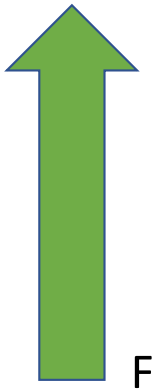
1. What is a Data Commons?





Opportunities:

- Next gen sequencing, mobile devices with sensors, dropping cost of images
- Data as the new oil and AI as the new factory
- Growing importance of open data and reproducible science



Fixed level of funding

Challenges

- IT infrastructure challenges from the size of the data and the complexity of the analysis
- Specialized computing infrastructure required by AI
- Batch effects from different analysis by different groups
- Security & compliance

Opportunities

Challenges

data commons

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.*

Fixed level of funding

*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.

FORCE11
The Future of Research Communications and e-Scholarship
FAIR Data Principles

BioData Catalyst
NHLBI


NCI Genomic Data Commons

BaseSpace
illumina®

NHGRI

DNAnexus®

Gabriella Miller Kids First
PEDIATRIC RESEARCH PROGRAM
Data Resource Center

SevenBridges
FireCloud
BROAD INSTITUTE
ISB

Open Science Data Cloud
THE UNIVERSITY OF CHICAGO

Bionimbus Protected Data Cloud
THE UNIVERSITY OF CHICAGO

2010 2014 2016 2018 2020

Projects

Data Clouds
2010 - 2025

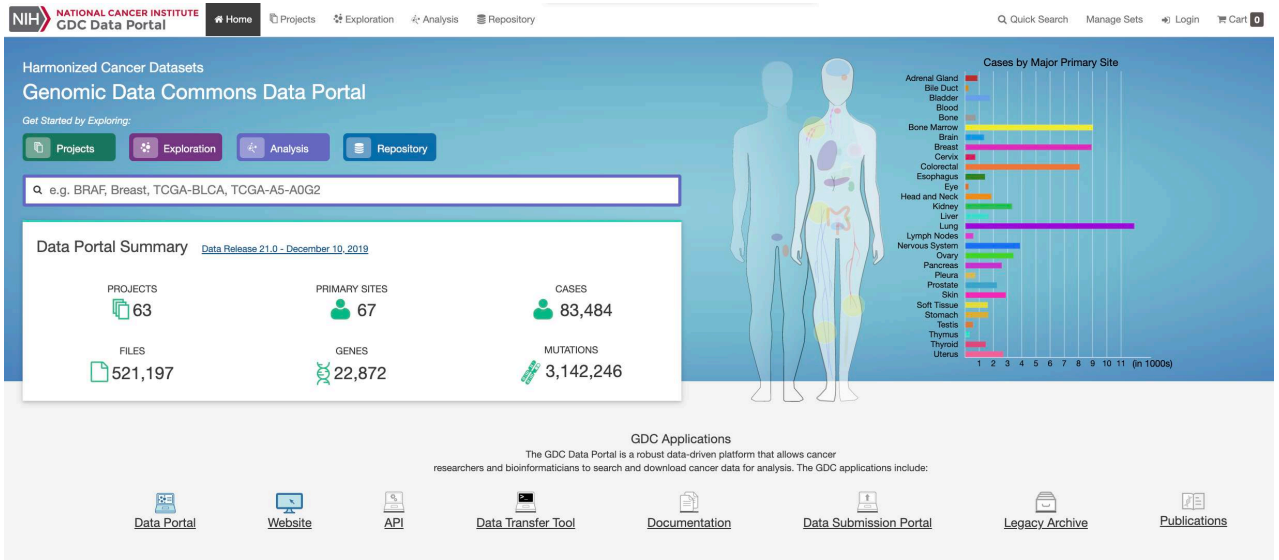
- Persistent Identifiers for **data objects in clouds**
- Researchers can use **cloud computing** to analyze data so it does not have to downloaded
- **Workflow languages** and **container repositories** for large scale computation

Data Cloud Architecture

- Data lake model
- GA4GH advocating for standards
- Standards have emerged for the data objects
- No standards yet for the data object's metadata
- Data is pulled into a computing environment for analysis
- Slow consensus on workflow languages
- No real consensus on work execution orchestration

Source: Robert L. Grossman, Data Lakes, Clouds and Commons: A Review of Platforms for Analyzing and Sharing Genomic Data, Trends in Genetics 35, 2019, pages 223-234. arxiv.org/abs/1809.01699 PMID: 30691868 PMCID: PMC6474403

NCI Genomic Data Commons*



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

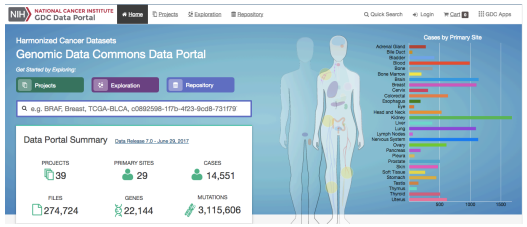
- The GDC makes over 2.5 PB of data available for access via an **API**, analysis by cloud resources on public clouds, and downloading.
- In an average month, the GDC is used by over 20,000 users, over 2 PB of data is accessed, and over 25,000 container based bioinformatics pipelines are run.
- The GDC is based upon an open source software stack that can be used to build other data commons.

*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.

Projects

Data Clouds
2010 - 2025

Communities



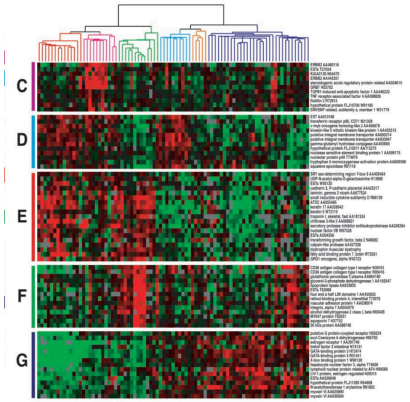
Data Commons
2015 - 2030

- Data objects in clouds
- Data workspaces in clouds
- **Common data models**
- **Harmonized data**
- **Core data services w APIs**
- **Data & Commons Governance**
- **Data sharing**
- **Reproducible research**

Data Commons Architecture

- Data lake model for data objects
- Graph (or other model) for clinical, biospecimen and other structured data
- Container based workflows to uniformly process submitted data (data harmonization)
- Open APIs to support portals, workspaces and third party applications

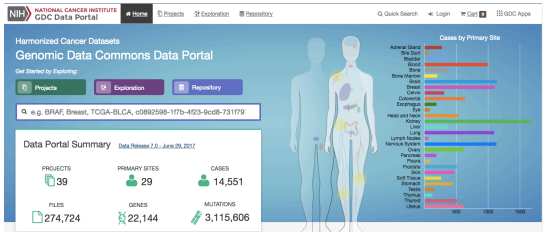
Source: Robert L. Grossman, Data Lakes, Clouds and Commons: A Review of Platforms for Analyzing and Sharing Genomic Data, Trends in Genetics 35, 2019, pages 223-234. arxiv.org/abs/1809.01699 PMID: 30691868 PMCID: PMC6474403



Databases organize data around a **project**.

Data Warehouse

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

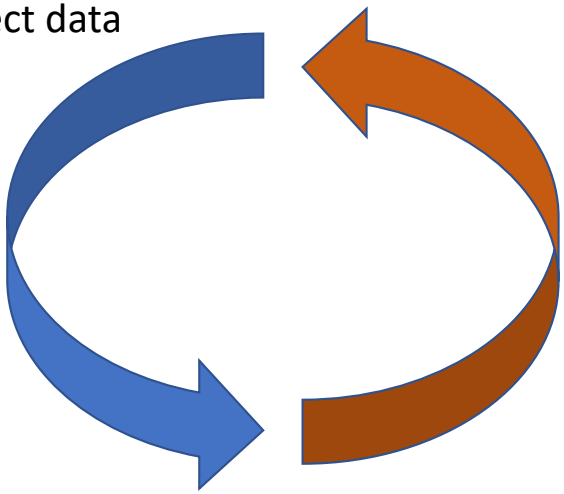


Data commons organize the data for a scientific **discipline** or field

Data commons balance protecting human subject data with open research that benefits patients:

Research ethics committees (RECs) review the ethical acceptability of research involving human participants. Historically, the principal emphases of RECs have been to protect participants from physical harms and to provide assurance as to participants' interests and welfare.*

Protect human subject data



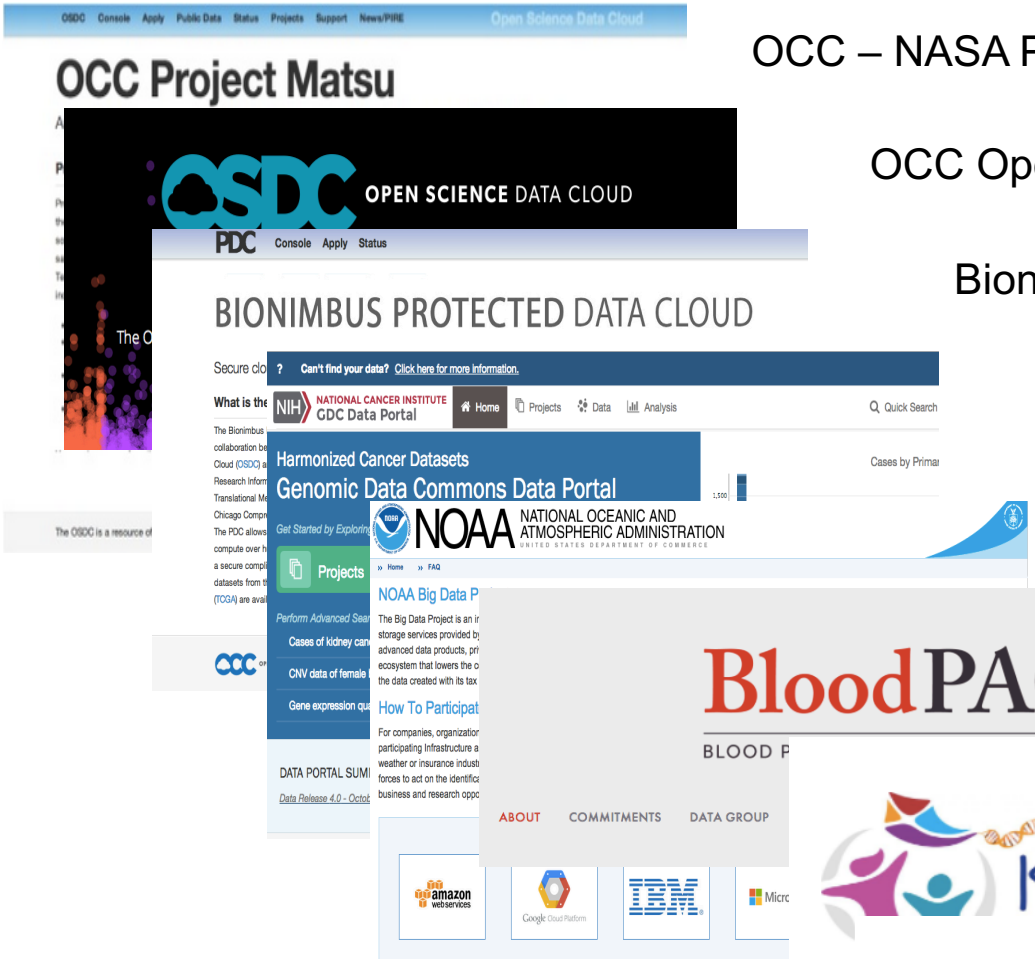
The right of human subjects to **benefit** from research.

[The Framework] is guided by, Article 27 of the 1948 Universal Declaration of Human Rights. Article 27 guarantees the rights of every individual in the world "to share in scientific advancement and its benefits" (including to freely engage in responsible scientific inquiry)…*

Data sharing with **protections** provides the evidence so patients can **benefit** from advances in research.

*GA4GH Framework for Responsible Sharing of Genomic and Health-Related Data, see goo.gl/CTavQR

2. Building a Data Commons with the Open Source Gen3 Data Platform



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)

BloodPAC Data Commons (2017)

Gabriella Miller Kids First

Kids First Data Resource (2017)

BRAIN Commons

Brain Commons (2017)

Gen3

GEN3 Data Commons

About Products Get Started ▼ Resources ▼ Community ▼

Gen3 is how data commons are made.

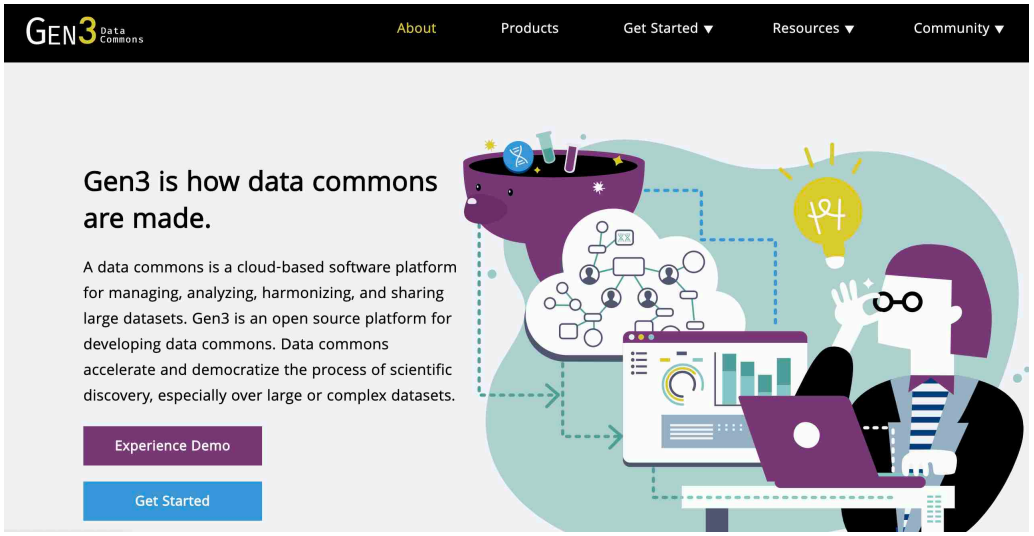
A data commons is a cloud-based software platform for managing, analyzing, harmonizing, and sharing large datasets. Gen3 is an open source platform for developing data commons. Data commons accelerate and democratize the process of scientific discovery, especially over large or complex datasets.

Experience Demo

Get Started

Gen3.org

Five Steps to Build a Gen3 Data Commons



1. Define a data model using Gen3.
2. Use the Gen3 platform to *auto-generate* the data commons and associated API (based upon your data model).
3. Import data into the commons using Gen3 data submission portal or Gen3 data submission API.
4. Use Gen3 data exploration portal to explore your data and create synthetic cohorts.
5. Use existing workspaces, (Jupyter and Rstudio) notebooks and applications to analyze the data or develop your own.

Gen3 Data Commons

BloodPAC — BLOOD PROFILING ATLAS IN CANCER

4,839 Subjects
852 Attributes
28,247 Files
Total Size 9.2 TB

OCC

BRAINCOMMONS

7,175 Subjects
6,919 Attributes
24,728 Files
Total Size 1.1 TB

Private Foundation

Gabriella Miller Kids First — PEDIATRIC RESEARCH PROGRAM — Data Resource Center

9,219 Subjects
622 Attributes
248,103 Files
Total Size 2.89 PB

NIH Common Fund

GenoMEL — the Melanoma Genetics Consortium

1,390 Subjects
387 Attributes
6,555 Files
Total Size 28.74 TB

NCI

BDGC

57,917 Subjects
644 Attributes
11,085 Files
Total Size 928.52 GB

NIDDK

26,636 Subjects
1,596 Attributes
129,379 Files
Total Size 740.75 TB

NHGRI

CANINE Data Commons

1,499 Subjects
1,008 Attributes
3,802 Files
Total Size 1.71 TB

OCC

NIAID DATA HUB

48,268 Subjects
1,693 Attributes
176,593 Files
Total Size 3.4 TB

NIAID

ACCOUNT

1,516 Subjects
936 Attributes
3,554 Files
Total Size 5.02 TB

NIMHD

VPO Veterans Precision Oncology Data Commons

113,154 Subjects
1,606 Attributes
352,458 Files
Total Size 1.68 TB

OCC

NIH NATIONAL CANCER INSTITUTE — Cancer Research Data Commons

83,709 Subjects
622 Attributes
1,986,961 Files
Total Size 2.91 PB

NCI

NIH National Heart, Lung, and Blood Institute | **BioData CATALYST**

240,460 Subjects
928 Attributes
24,215,053 Files
Total Size 1.98 PB

NHLBI

Environmental Data Commons

265 Attributes
18,708,594 Files
Total Size 47.15 TB

OCC

3. From Data Commons to Data Ecosystems

End to End Design Principle

End-To-End Arguments in System Design

J. H. SALTZER, D. P. REED, and D. D. CLARK
Massachusetts Institute of Technology Laboratory for Computer Science

This paper presents a design principle that helps guide placement of functions among the modules of a distributed computer system. The principle, called the end-to-end argument, suggests that functions placed at low levels of a system may be redundant or of little value when compared with the cost of providing them at that low level. Examples discussed in the paper include bit-error recovery, security using encryption, duplicate message suppression, recovery from system crashes, and delivery acknowledgment. Low-level mechanisms to support these functions are justified only as performance enhancements.

CR Categories and Subject Descriptors: C.0 **[General]** Computer System Organization—*system architectures*; C.2.2 **[Computer-Communication Networks]**: Network Protocols—*protocol architecture*; C.2.4 **[Computer-Communication Networks]**: Distributed Systems; D.4.7 **[Operating Systems]**: Organization and Design—*distributed systems*

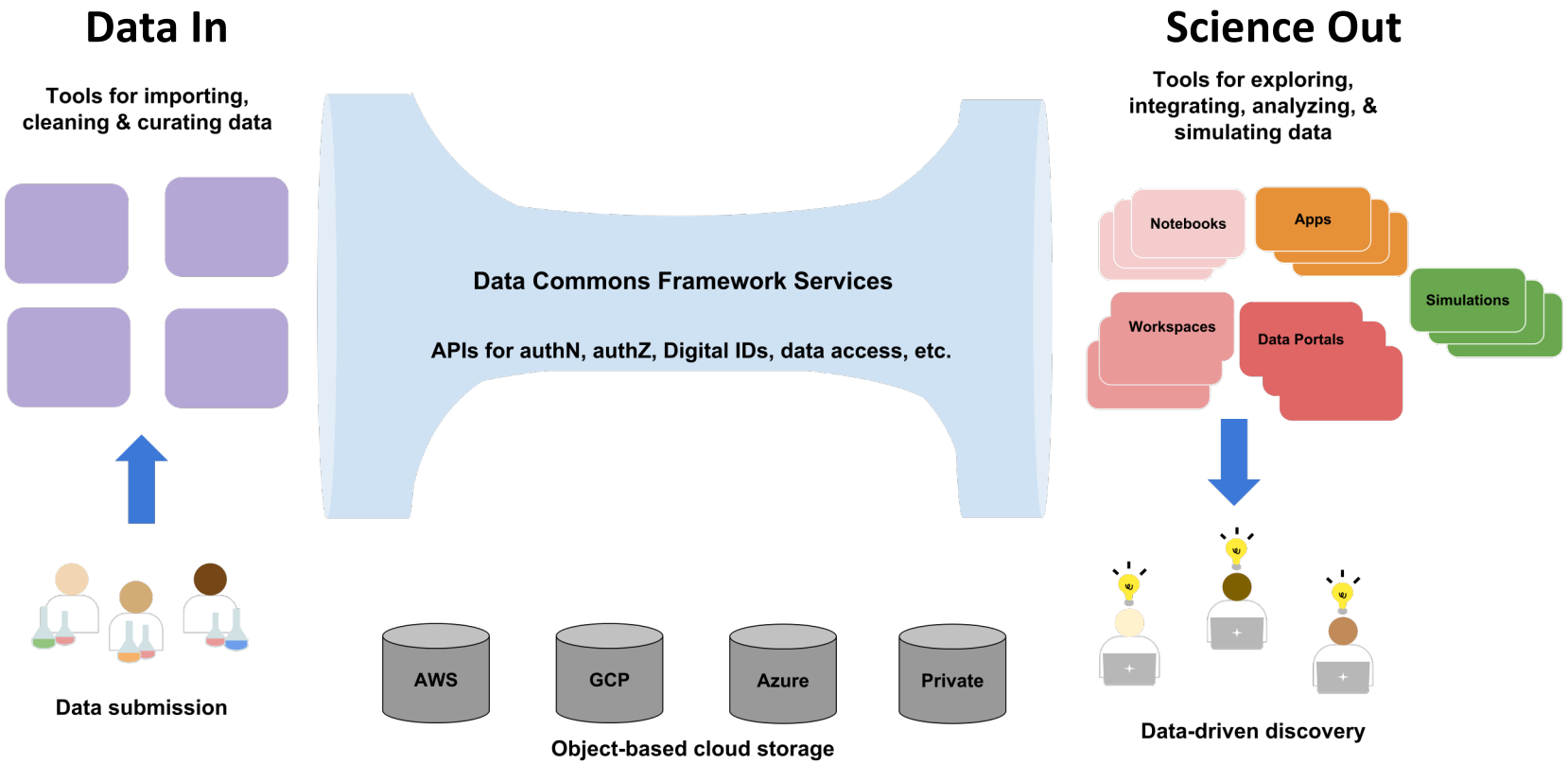
General Terms: Design

Additional Key Words and Phrases: Data communication, protocol design, design principles

1. INTRODUCTION

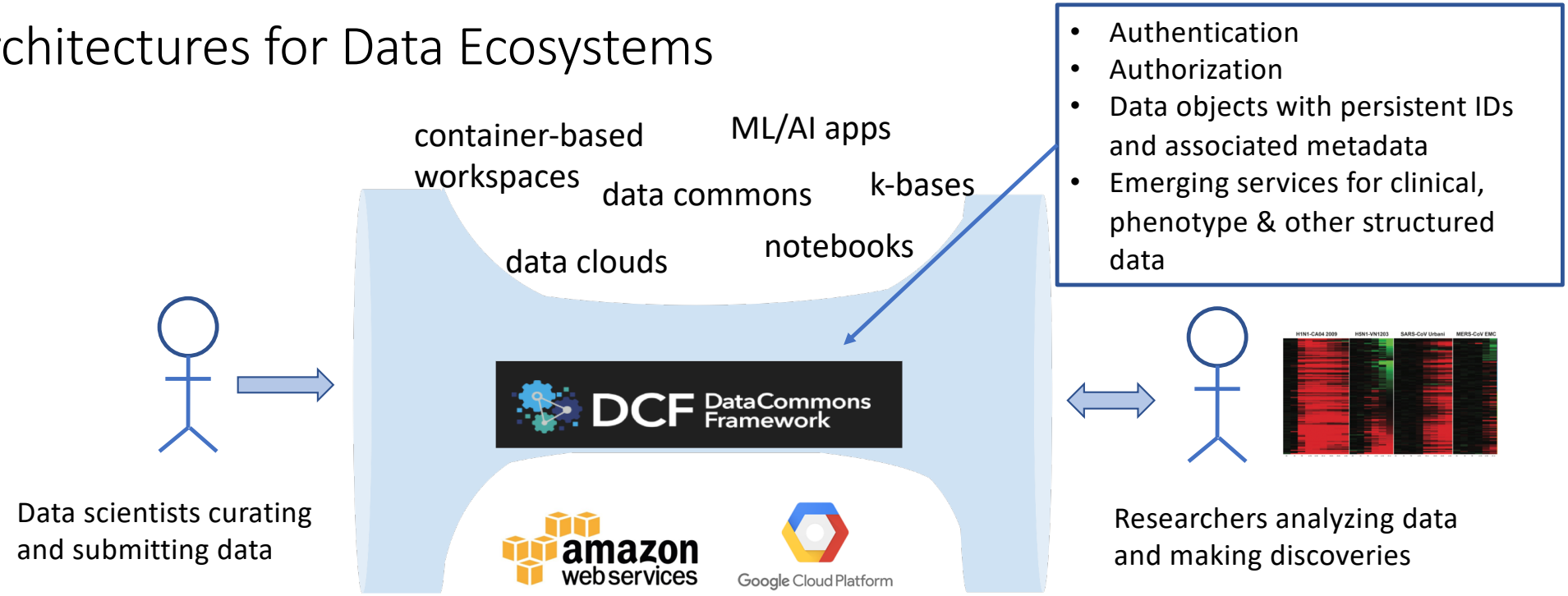
Choosing the proper boundaries between functions is perhaps the primary activity of the computer system designer. Design principles that provide guidance in this choice of function placement are among the most important tools of a system designer. This paper discusses one class of function placement argument that

Source: ACM Transactions on Computer Systems (TOCS), Volume 2 Issue 4, Nov. 1984, Pages 277-288



Source: Robert L. Grossman, Progress Towards Cancer Data Ecosystems, The Cancer Journal: The Journal of Principles & Practice of Oncology, 2018, Volume 24, Number 3, May/June 2018.

Architectures for Data Ecosystems



- A simple data ecosystem can be built when a data commons exposes an API that can support a collection of third party applications that can access data from it.
- More complex data ecosystems arise when multiple data commons and data clouds interoperate and support a collection of third party applications using a common set of core services (called framework services).

Source: Robert L. Grossman, Progress Towards Cancer Data Ecosystems, The Cancer Journal: The Journal of Principles and Practice of Oncology, May/June 2018, Volume 24 Number 3, pages 122-126

Gen3 Data Ecosystem

GEN3 DATA COMMONS

DCF DataCommons Framework

Total Subjects: 616,228
Total Files: 45,895,112
Total File Size: 8.6 PB

BloodPAC
BLOOD PROFILING ATLAS IN CANCER

4,839 Subjects
852 Attributes
28,247 Files
Total Size 9.2 TB

OCC

BRAIN COMMONS

7,175 Subjects
6,919 Attributes
24,728 Files
Total Size 1.1 TB

Private Foundation

Gabriella Miller Kids First
PEDIATRIC RESEARCH PROGRAM
Data Resource Center

9,219 Subjects
622 Attributes
248,103 Files
Total Size 2.89 PB

NIH Common Fund

GenoMEL
the Melanoma Genetics Consortium

1,390 Subjects
387 Attributes
6,555 Files
Total Size 28.74 TB

NCI

BDGC

57,917 Subjects
644 Attributes
11,085 Files
Total Size 928.52 GB

NIDDK

NHGRI

26,636 Subjects
1,596 Attributes
129,379 Files
Total Size 740.75 TB

NHGRI

CANINE Data Commons

1,499 Subjects
1,008 Attributes
3,802 Files
Total Size 1.71 TB

OCC

NIAID DATA HUB

48,268 Subjects
1,693 Attributes
176,593 Files
Total Size 3.4 TB

NIAID

ACCOUNT

1,516 Subjects
936 Attributes
3,554 Files
Total Size 5.02 TB

NIMHD

VPO Veterans Precision Oncology Data Commons

113,154 Subjects
1,606 Attributes
352,458 Files
Total Size 1.68 TB

OCC

NIH NATIONAL CANCER INSTITUTE Cancer Research Data Commons

83,709 Subjects
622 Attributes
1,986,961 Files
Total Size 2.91 PB

NCI

NIH National Heart, Lung, and Blood Institute | BioData CATALYST

240,460 Subjects
928 Attributes
24,215,053 Files
Total Size 1.98 PB

NHLBI

OCC Environmental Data Commons

265 Attributes
18,708,594 Files
Total Size 47.15 TB

OCC

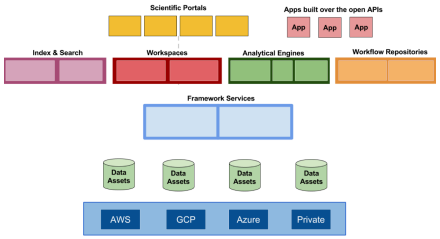
Projects

Data Clouds
2010 - 2025

Communities

Data Commons
2015 - 2030

Multiple Communities



Data Ecosystems
2018 - 2030

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

Data Ecosystem Architecture

- Data lake model for data objects
- Framework services with AuthN/AuthZ, data objects and services for clinical/phenotype data
- Open APIs to support other commons, portals, workspaces and third-party applications
- Container based workflows to uniformly process submitted data (data harmonization)
- Governance model that supports data sharing

Source: Robert L. Grossman, Data Lakes, Clouds and Commons: A Review of Platforms for Analyzing and Sharing Genomic Data, Trends in Genetics 35, 2019, pages 223-234. arxiv.org/abs/1809.01699 PMID: 30691868 PMCID: PMC6474403

To build a data commons

Gen3.org

Gen3 Data Commons

- Open source
- Define data model
- Import and curate data
- Create and export synthetic cohorts
- Analyze data, share data

Governance & best practices for building data commons & ecosystems

OPEN COMMONS CONSORTIUM

OCC-data.org

Open Commons Consortium

- Not-for-profit
- Data commons governance
- Data ecosystems governance
- Security & compliance services
- Legal templates
- Outsource operating data commons & ecosystems

To build a data ecosystem

DCF DataCommons Framework

DCF.Gen3.org

Gen3 Data Commons Framework Services (DCFS)

- AuthN/AuthZ
- Digital ID and metadata services for data objects
- Emerging services for clinical, phenotype & other structured data

4. Framework Services for (Bulk) Clinical and Phenotype Data

Related Work

- FHIR 

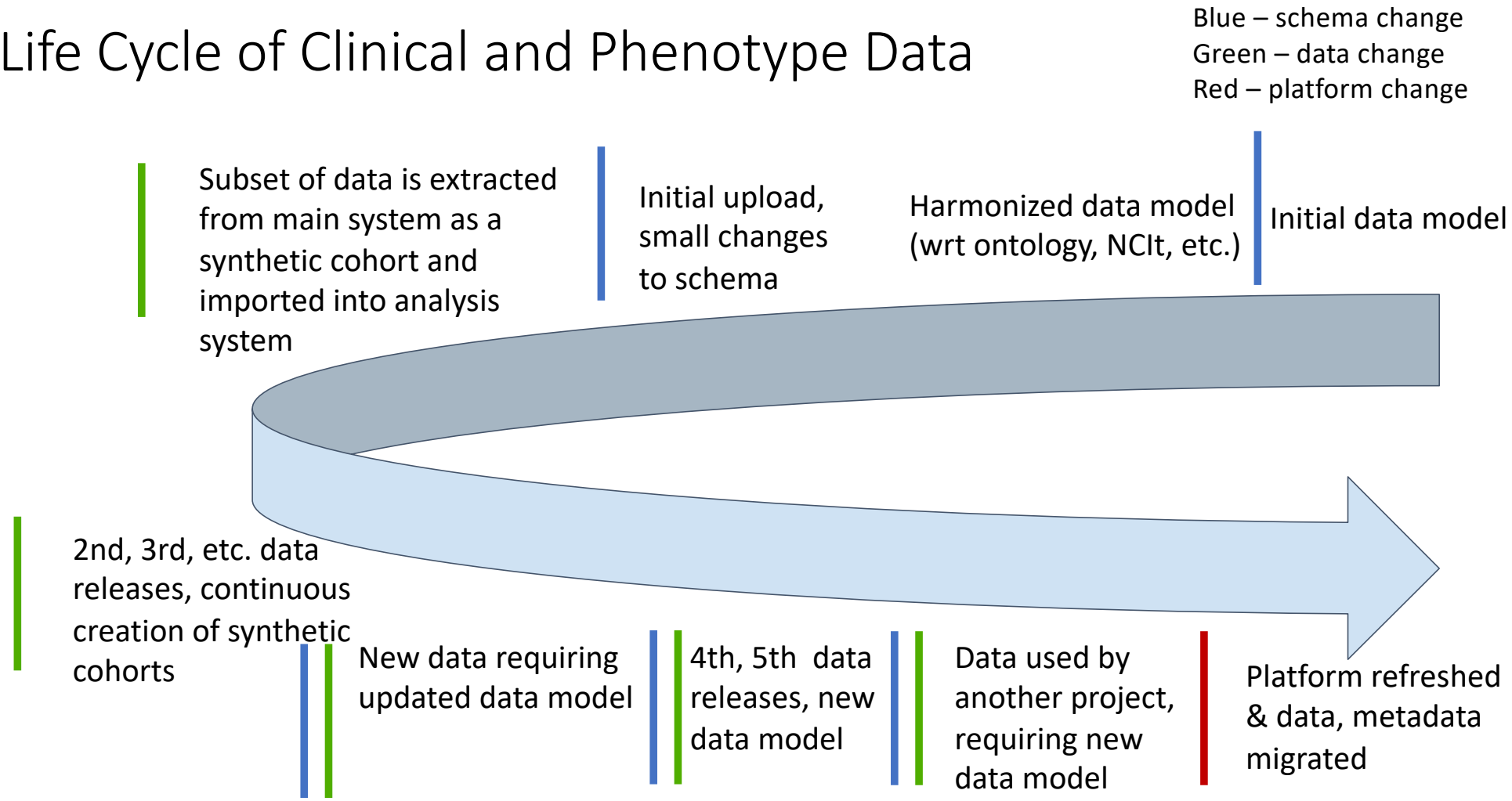
- OMOP Common Data Model 

- bioCADDIE 

- PhenoPackets 

Life Cycle of Clinical and Phenotype Data

Blue – schema change
Green – data change
Red – platform change



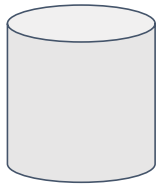
Linking Structured Clinical Data with Genomic Data

Object data - CRAM/BAM genomic data files, DICOM image files, anything stored as data objects in clouds

Clinical data / graph data / core data / **structured data** - data that are harmonized to a data model and searchable using a data model and related APIs. Gen3 uses a graph data model as the logical model.

Data objects stored with GUIDS in one or more clouds

Data objects and clinical data linked in data model



Clinical data and other structured data stored in a database

Requirement	Approach	Gen3 Services
1. Make the data FAIR	Data objects are assigned GUID & metadata and placed in multiple clouds	IndexD, Fence, Metadata services
2. Express the pipelines in a workflow language and making them FAIR	We support Common Workflow Language (CWL)	We support Dockstore, CWL & cwltool, use object services to manage CWL files
3. Encapsulate the code and tools	We encapsulate code in virtual machines & containers	We use Kubernetes, Docker, Dockstore and WES
4. Link data and code	Use notebooks	We support Jupyter notebooks and RStudio
5. Make clinical data portable	???	???

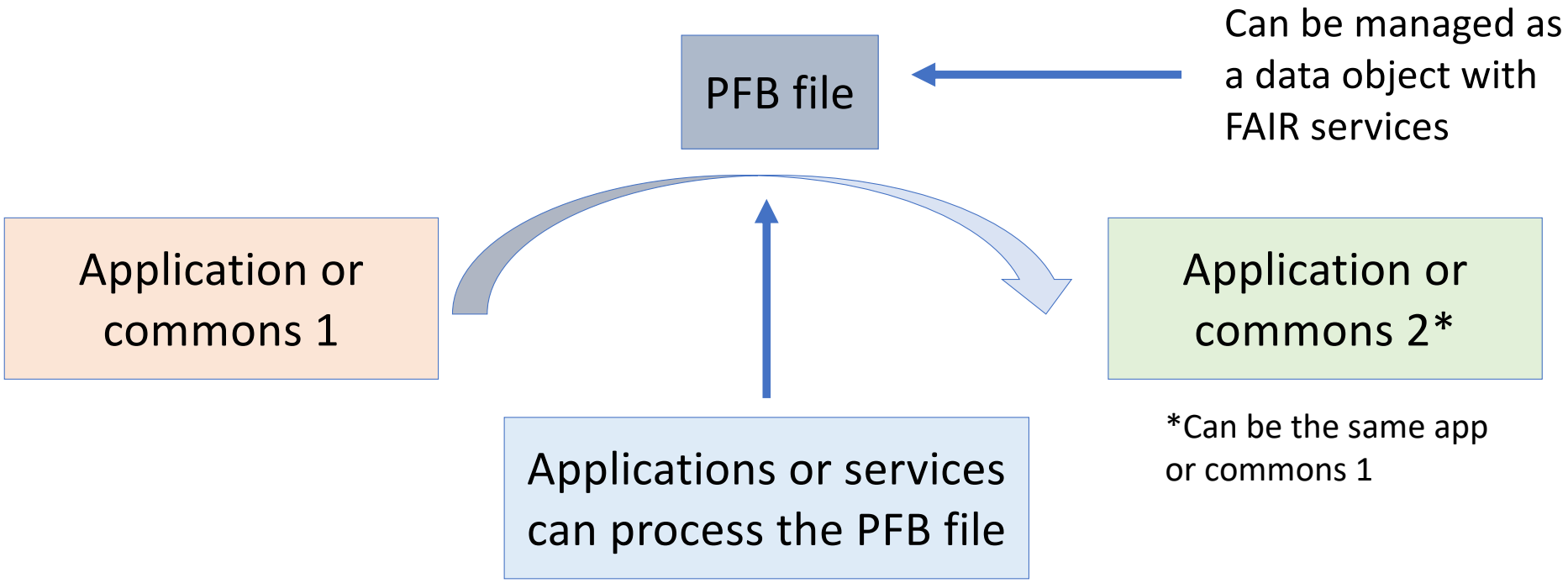
What is Portable Format for Biomedical (PFB) Data?

- PFB is an Avro-based serialization format with a specific schema to import, export and evolve biomedical data.
- PFB specifies metadata and data in one file. Metadata includes a data dictionary, a data model, and pointers to third-party ontology references and controlled vocabularies
- PFB is:
  - Portable: supporting import & export.
  - Extensible: data model changes, versioning, back- and forward compatibility;
  - Efficient: Avro binary format.

Why Avro?

	Avro	**Protobuf**
Self-describing	✓	✗
Schema evolution	✓	✓
Dynamic schema	✓	Partially, needs recompilation
No need to compile	✓	✗
Hadoop support	✓, built-in	✓, third-party libraries
JSON schema	✓	✗, special IDL for schema

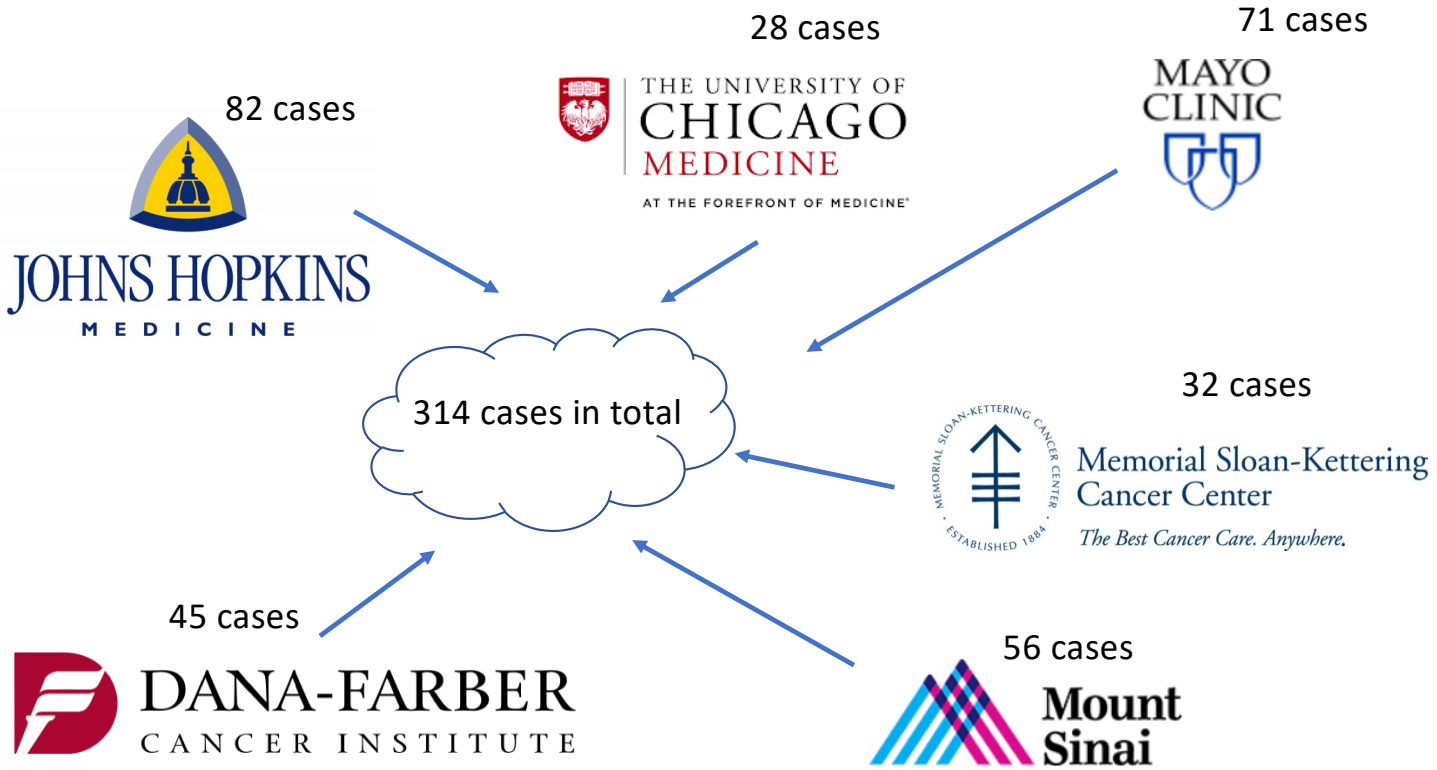
Portable Format for Biomedical Data (PFB)

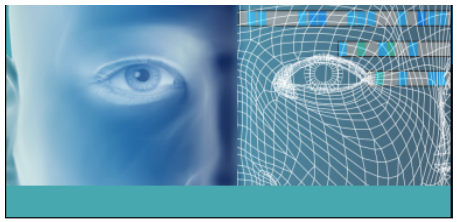


- PFB is an application independent and system independent serialization format for importing and exporting: 1) schema and other metadata, 2) pointers to third party ontologies and authorities, and 3) data.
- PFB services can export to JSON

5. Principles to Support Open Science with Data Commons and Data Ecosystems

Without Data Sharing, We Don't Have the Scientific Evidence We Need for a Learning HealthCare System





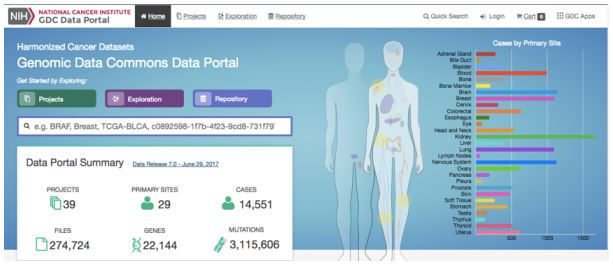
Bermuda Principles
& Genomic Databases
(e.g. GenBank)
1982 - present

PLOS

PMC arXiv.org

bioRxiv

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



What are the principles that might govern data commons and data ecosystems to support open science?

Bermuda Principles

1. Automatic release of genomic 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 maximize 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

Proposed Open Commons Principles* (2017)

1. Agencies and foundations that fund biomedical research should require that researchers share the data generated.

2. Agencies and foundations that fund biomedical research should provide the computing infrastructure ("commons") and bioinformatics resources that are required to support data sharing.

3. The data commons developed by agencies and foundations should themselves share data and interoperate with other data commons to create a data ecosystem.

*Source: Robert L. Grossman, Supporting Open Data and Open Science With Data Commons: Some Suggested Guidelines for Funding Organizations, March 23, 2017, https://www.healthra.org/download-resource/?resource-url=/wp-content/uploads/2017/08/Data-Commons-Guidelines_Grossman_8_2017.pdf

Benefits of Data Commons and Data Sharing

1. Move the research **field forward faster**.
2. Support **repeatable, reproducible and open** research.
3. We have the statistical power to study **weaker effects.**
4. Researchers can work with **large datasets at much lower cost** and make discoveries of phenomena that are not evident at smaller scale.
5. Data commons can **interoperate** with each other to create a data ecosystem so that over time data sharing can benefit from a “network effect”

Questions?

rgrossman.com
@bobgrossman

We are hiring and also looking for volunteers that want to impact biology, medicine and healthcare using data science and cloud computing. Please contact us at the CTDS or the OCC.

For More Information

[1] Robert L. Grossman, Data Lakes, Clouds and Commons: A Review of Platforms for Analyzing and Sharing Genomic Data, Trends in Genetics 35, 2019, pages 223-234.

[2] Robert L. Grossman, Some Proposed Principles for Interoperating Data Commons, Medium, October 1, 2019.

[3] Robert L. Grossman, Progress Towards Cancer Data Ecosystems, The Cancer Journal: The Journal of Principles and Practice of Oncology, May/June 2018, Volume 24 Number 3, pages 122-126 doi: 10.1097/PPO.0000000000000318. PMID: 29794537

- 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
- **Bionimbus**. 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. DOI: 10.1136/amiajnl-2013-002155. This article describes Bionimbus, which was a Gen1 system.
- **GDC**. 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. https://www.nejm.org/doi/full/10.1056/NEJMp1607591. The GDC was developed using Bionimbus Gen2.
- **BloodPAC**. 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 Gen3.
- **GDC API**. 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. PMC: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5683428/ and pdf
- **Medical Science DMZ**. Sean Peisert, Eli Dart, William Barnett, Edward Balas, James Cuff, Robert L Grossman, Ari Berman, Anurag Shankar, Brian Tierney; The medical science DMZ: a network design pattern for data-intensive medical science, Journal of the American Medical Informatics Association, ocx104, published 6 October 2017, https://doi.org/10.1093/jamia/ocx104 pdf
- **Data Ecosystem.** Robert L. Grossman, Progress Towards Cancer Data Ecosystems, The Cancer Journal: The Journal of Principles and Practice of Oncology, May/June 2018, Volume 24 Number 3, pages 122-126 doi: 10.1097/PPO.0000000000000318. PMID: 29794537 pdf
- **Review of Clouds and Commons.** Robert L. Grossman, Data Lakes, Clouds and Commons: A Review of Platforms for Analyzing and Sharing Genomic Data, arXiv:1809.01699v1 [q-bio.GN]

Contact Information

Robert L. Grossman
rgrossman.com

@BobGrossman
robert.grossman@uchicago.edu

THE UNIVERSITY OF CHICAGO | Center for Translational Data Science

ctds.uchicago.edu

OPEN COMMONS CONSORTIUM

occ-data.org