How Data Commons Are Changing the Way We Share Research Data and Make Discoveries: The Open Commons Consortium Perspective

Robert Grossman

University of Chicago, Open Commons Consortium and Open Data Group

NSF Data Science Seminar
July 6, 2016

Thanks

The work discussed here is joint work with the various researchers at the University of Chicago's Center for Data Intensive Science (CDIS) and with the Open Commons Consortium and its Working Groups.

The design and architecture of a data commons is described in: Robert L. Grossman, Allison Heath, Mark Murphy, and Maria Patterson, A Case for Data Commons Towards Data Science as a Service, , IEEE Computing in Science and Engineer, 2016, to appear.

The design and architecture of the Bionimbus Protected Data Cloud is described in: Allison P. Heath, Matthew Greenway, Raymond Powell, Jonathan Spring, Rafael Suarez, David Hanley, Chai Bandlamudi, Megan McNerney, Kevin White and Robert L Grossman, Bionimbus: A Cloud for Managing, Analyzing and Sharing Large Genomics Datasets, Journal of the American Medical Informatics Association, 2014.

The design and architecture of the OCC Matsu Commons is described in: Maria T. Patterson, Nikolas Anderson, Collin Bennett, Jacob Bruggemann, Robert L. Grossman, Matthew Handy, Vuong Ly, Daniel J. Mandl, Shane Pederson, James Pivarski, Ray Powell, Jonathan Spring, Walt Wells, and John Xia, "The Matsu Wheel: A Cloud-based Framework for Efficient Analysis and Reanalysis of Earth Satellite Imagery." In 2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService), pp. 156-165. IEEE, 2016.

A publication describing the design and architecture of the NCI Genomic Data Commons is in preparation.

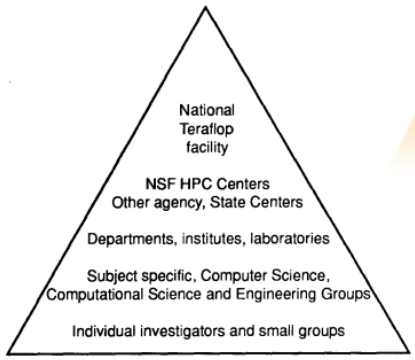
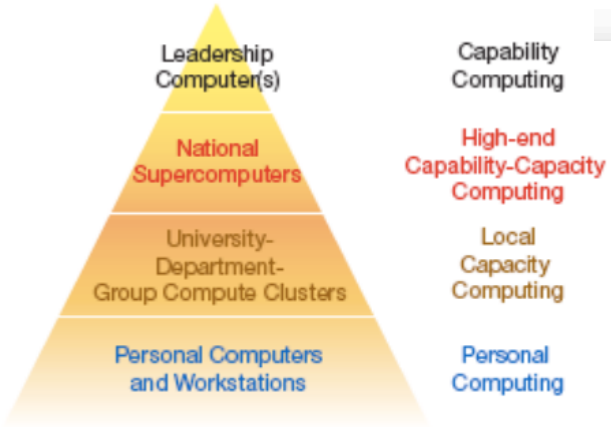
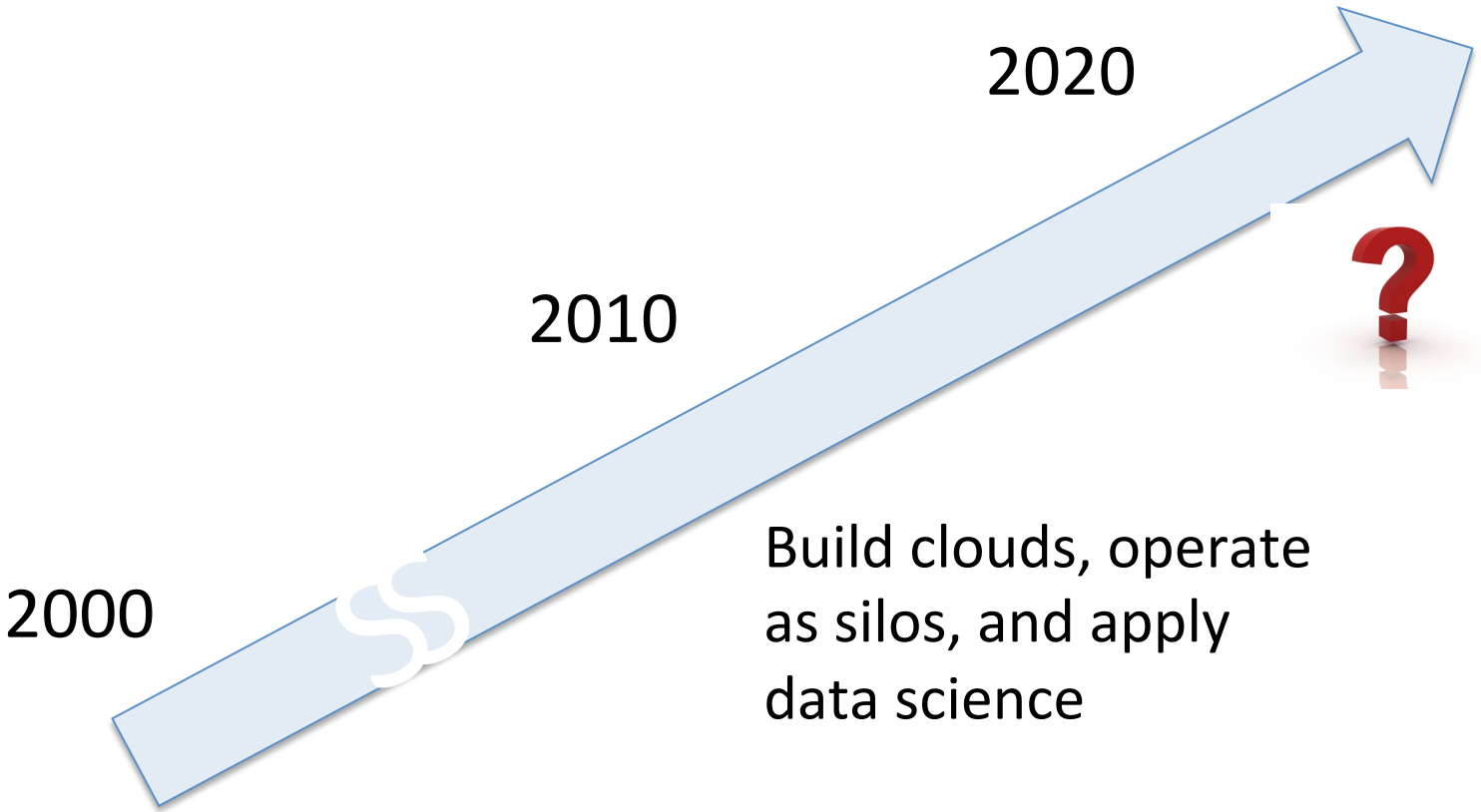


Figure B
PYRAMID OF HIGH PERFORMANCE COMPUTING INSTITUTIONS



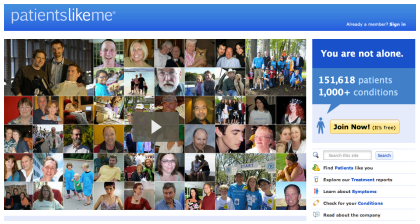
Source of first figure: From Desktop to Teraflop: Exploiting the U.S. Lead in High Performance Computing, NSF Blue Ribbon Panel on High Performance Computing, August 1993. Source of second figure: Mid-Range Computing in Support of Science at Office of Science Laboratories, Report of a Workshop October 21-22, 2008, http://science.energy.gov/~/media/ascr/pdf/program-documents/docs/Midrange_report_final.pdf. Source of third figure: Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science and Engineering in 2017-2020, National Academies Press.


2020
2010
2000
Build clouds, operate as silos, and apply data science
Build data repositories and data coordinating centers, and apply data mining

1. Data Commons

We have a problem …

The commoditization of sensors is creating an explosive growth of data



There is not enough funding for every researcher to house all the data they need

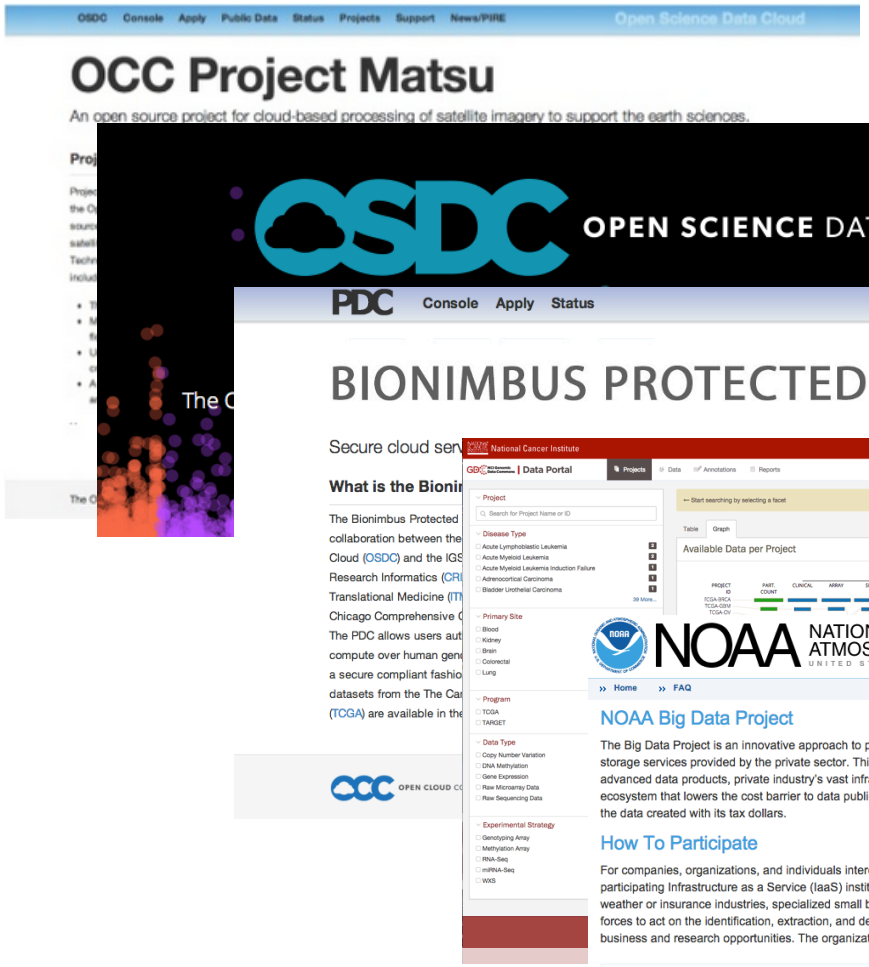
It can take weeks to download large datasets

Analyzing the data is often more expensive than producing it

Data Commons

Data commons co-locate data, storage and computing infrastructure with commonly used services and tools for 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 the University of Chicago Kenwood Data Center.



Project Matsu – NASA Collaboration (2009)

Open Science Data Cloud (2010)

Bionimbus Protected Data Cloud* (2013)

NCI Data Commons* (2015)

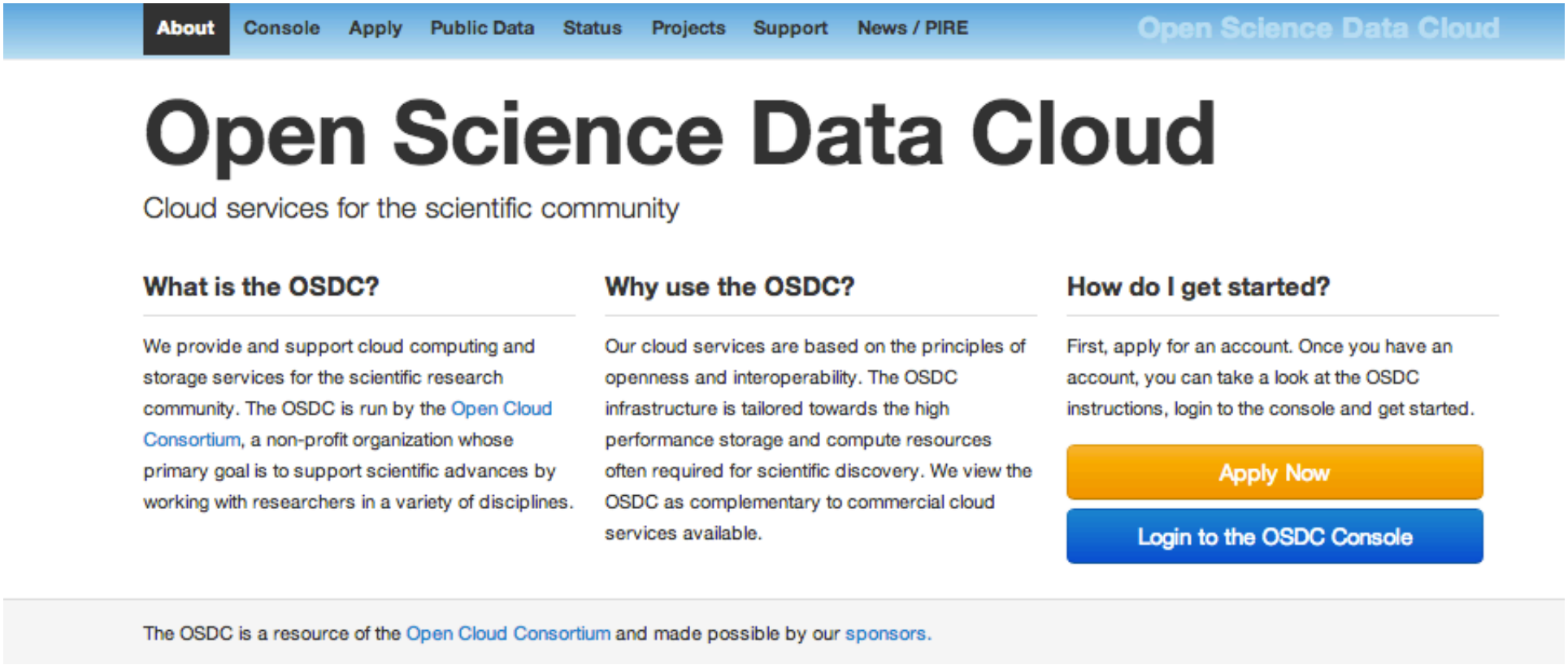
NOAA Data Commons (2015)

- OCC operated testbed from 2008 to 2012 to develop the technology.
- 860+ research projects supported since 2010.
- 210+ currently active researchers from 55 organizations from 14 countries in 215. Over 18 million core hours used by allocation grantees in past year

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

- U.S based 501(c)(3) not-for-profit corporation founded in 2008.
- 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.
- Manages data commons to support medical and health care research: OOC Biomedical Data Commons.
- It is international and includes universities, not-for-profits, companies and government agencies.
- Formerly known as the Open Cloud Consortium (the name was changed in 2015).

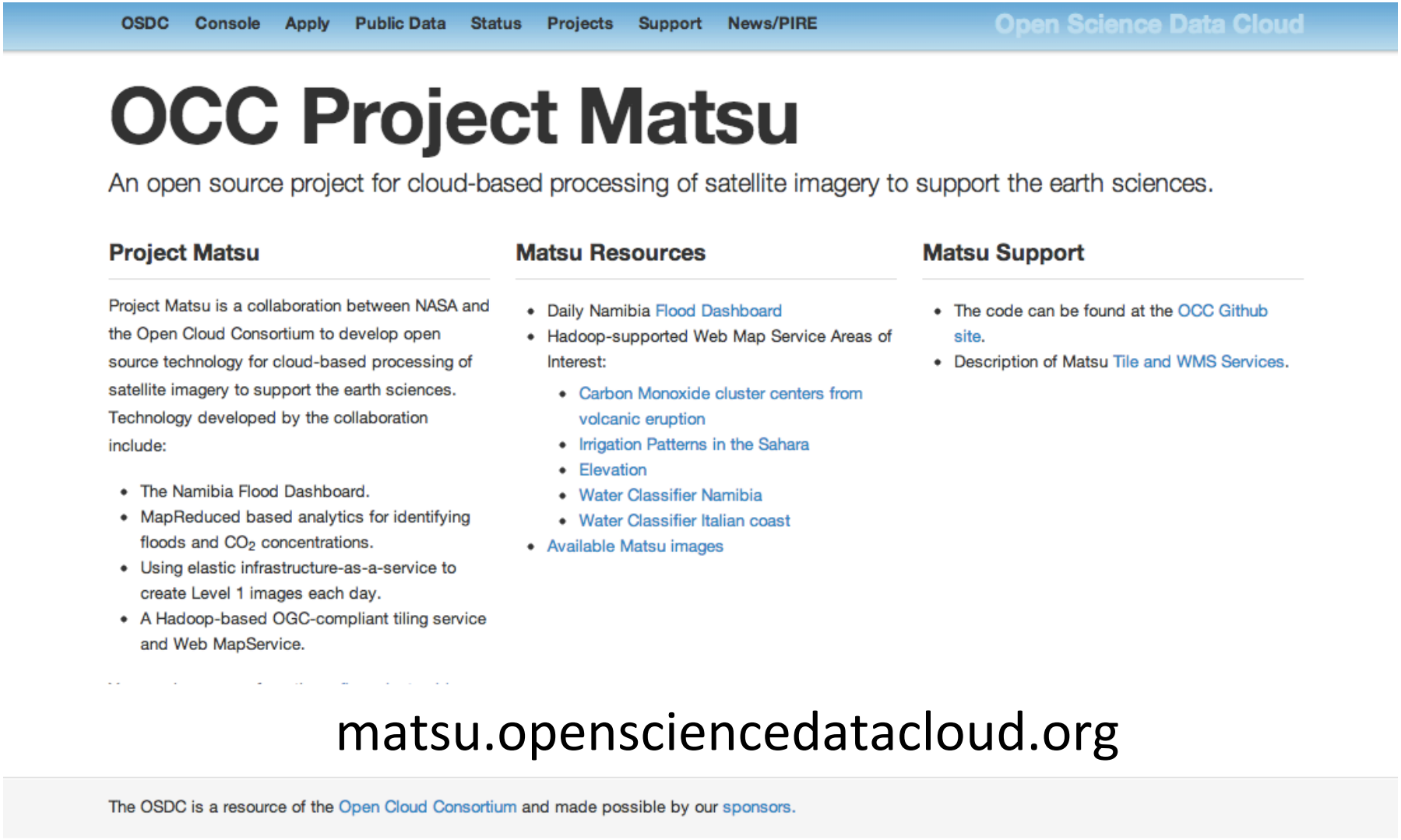
www.occ-data.org



- The Open Science Data Cloud (OSDC) is a production 5PB, 7500 core science cloud with attached data commons.
- 860+ research projects supported since 2010.
- 210+ currently active researchers from 55 organizations spanning 14 countries.
- Most of our users have allocations of 10,000 or more core-hours per month.
- Over 18 million core hours used by researchers in past year.

www.opensciencedatacloud.org

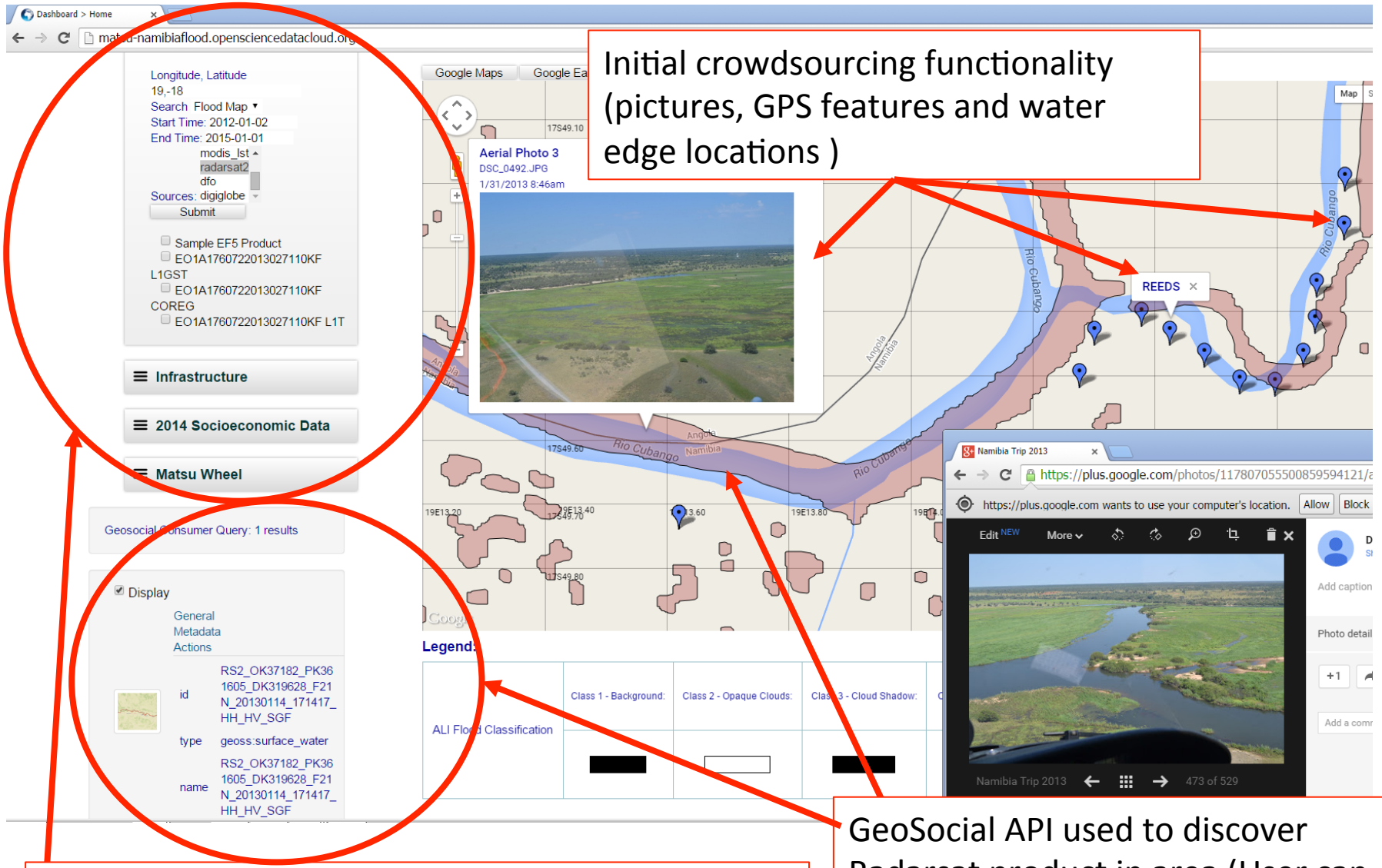
GORDON AND BETTY MOORE FOUNDATION



matsu.opensciencedatacloud.org

OCC-NASA Collaboration 2009 - present

Flood/Drought Dashboard Examples



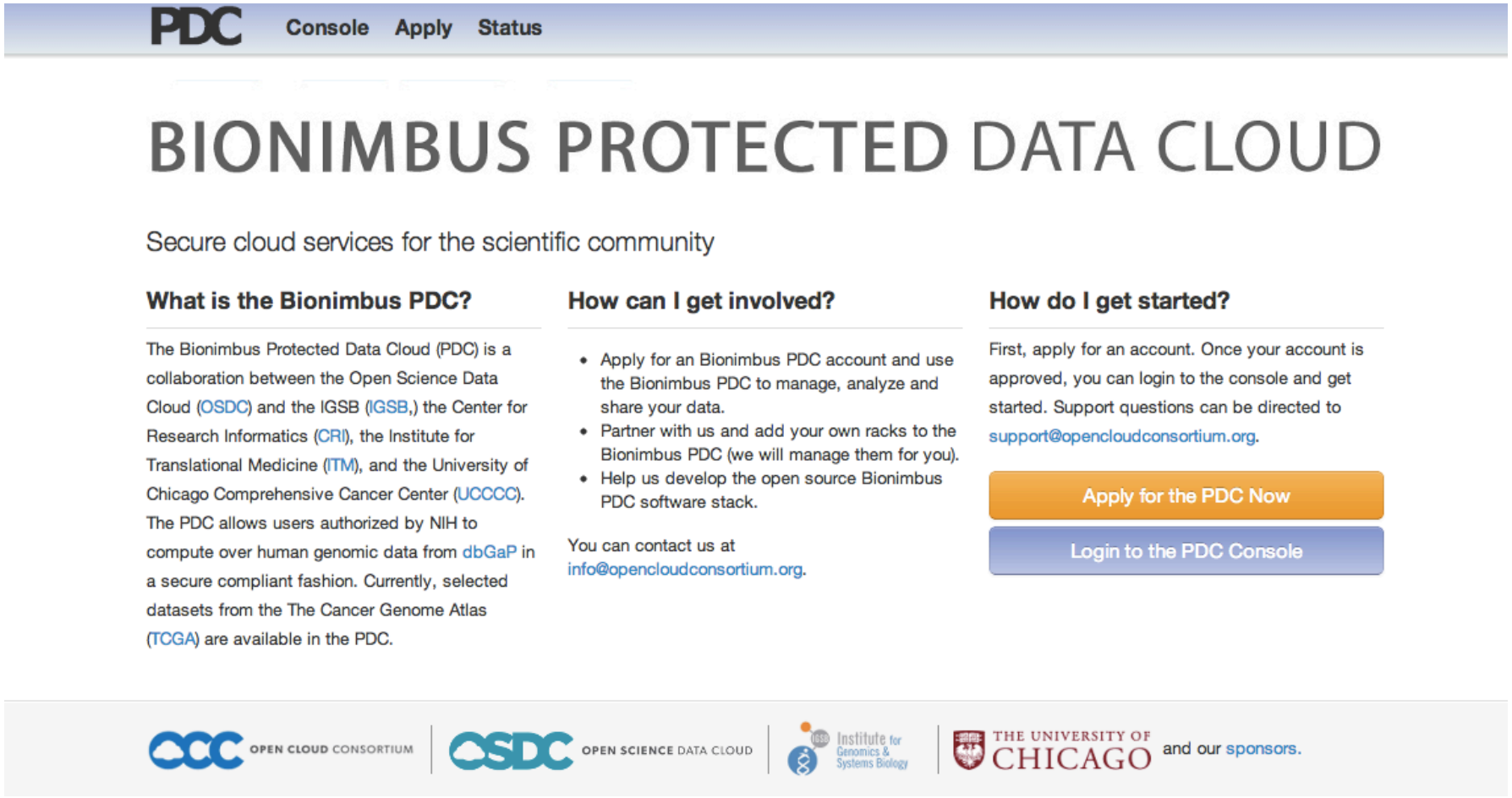
Initial crowdsourcing functionality (pictures, GPS features and water edge locations)

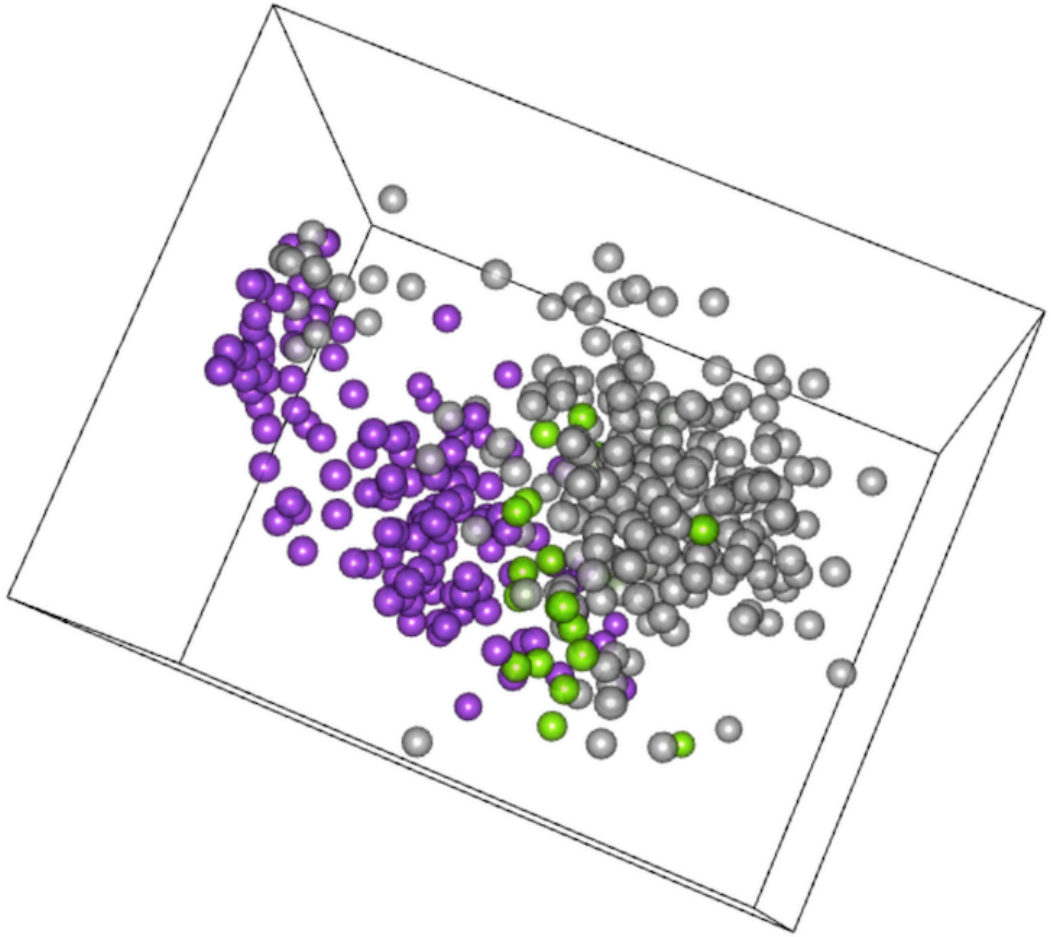
GeoSocial API Consumer embedded in Dashboard

GeoSocial API used to discover Radarsat product in area (User can see registration error)

Source: Flood Dashboard, Dan Mandl, NASA

The Bionimbus Protected Data Cloud

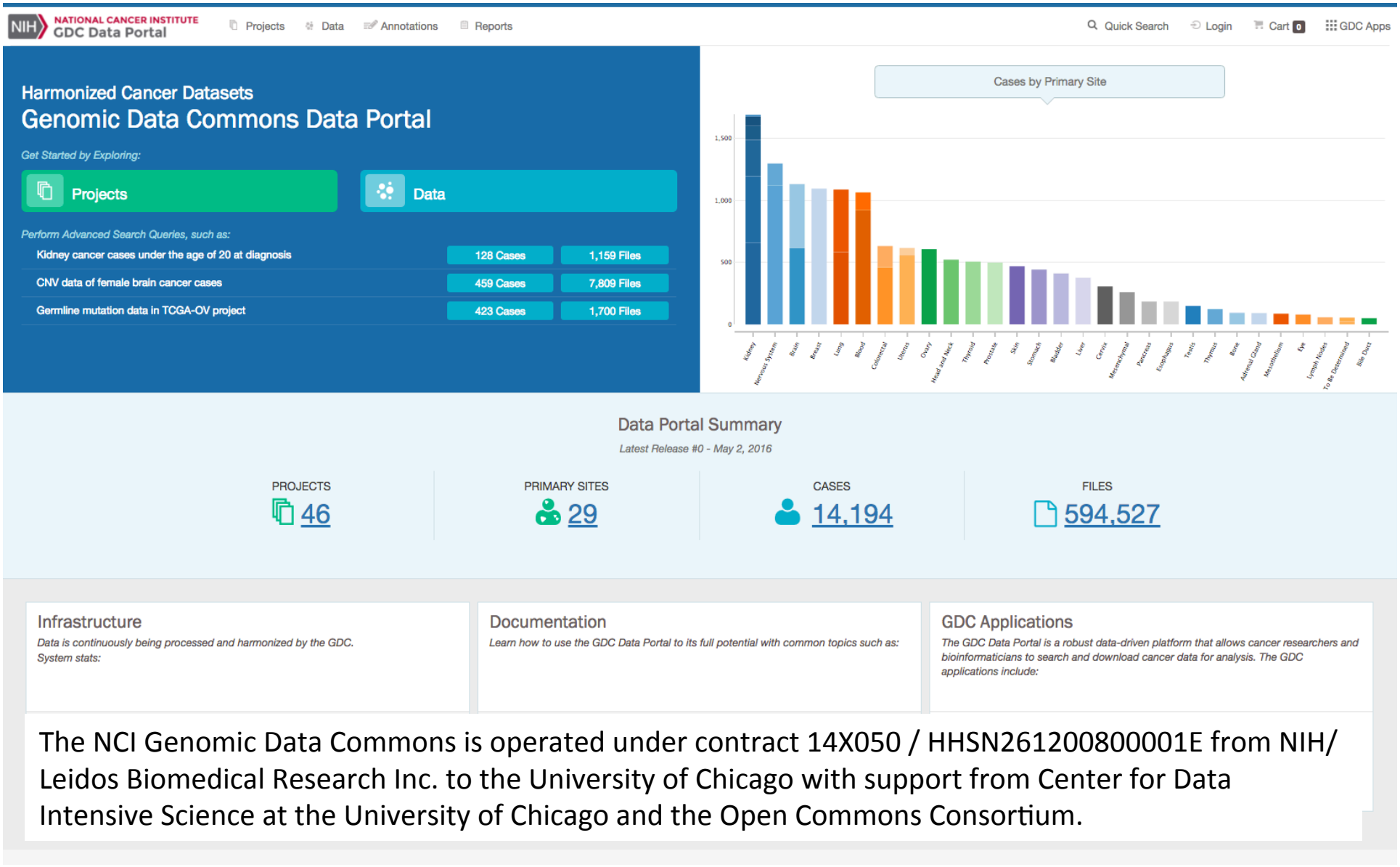




Purple balls are lung adenocarcinoma. Grey are lung squamous cell carcinoma. Green appear to be misdiagnosed.

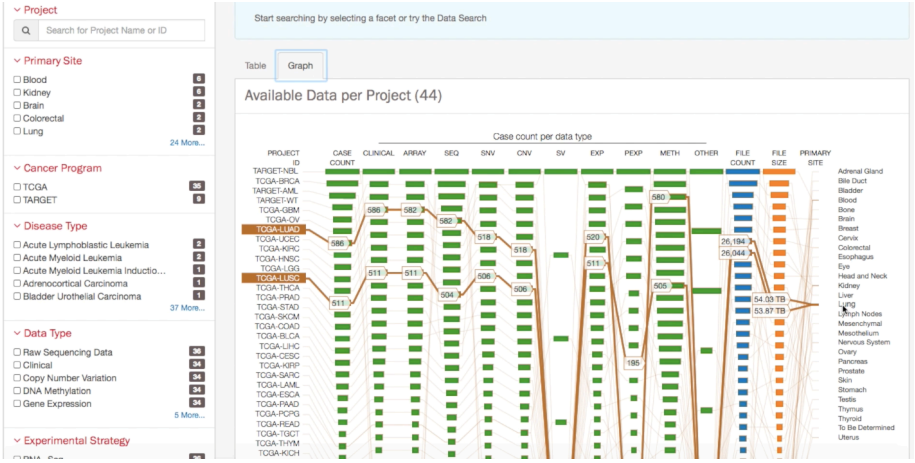
Source: Center for Data Intensive Science, University of Chicago.

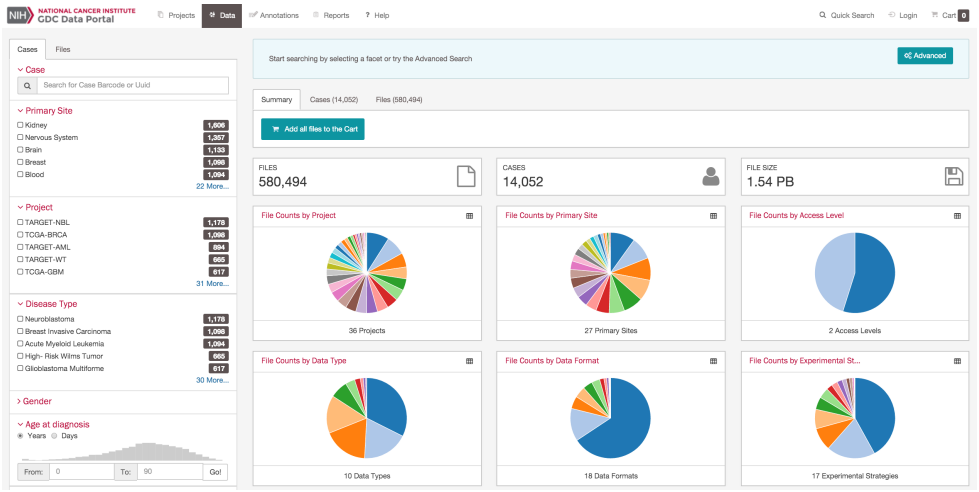
NCI Genomic Data Commons Data Portal



The NCI Genomic Data Commons is operated under contract 14X050 / HHSN261200800001E from NIH/ Leidos Biomedical Research Inc. to the University of Chicago with support from Center for Data Intensive Science at the University of Chicago and the Open Commons Consortium.

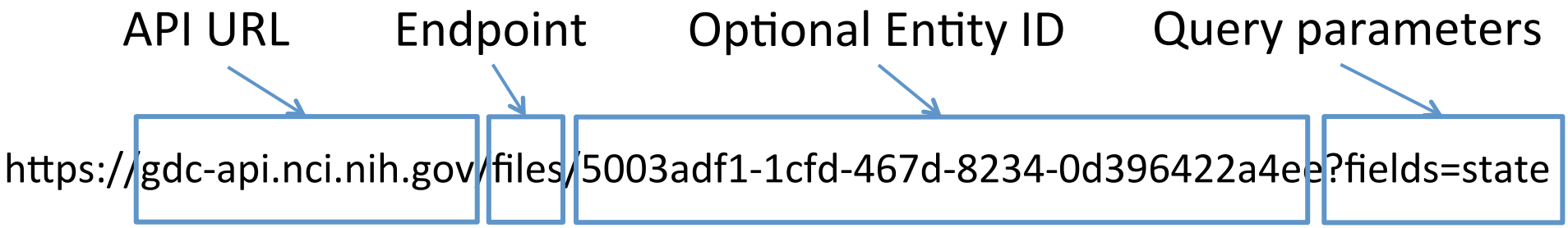
- Designed to store and analyze existing & future NCI genomic and clinical cancer data (5 - 100+ PB)
- Key capability is uniform re-analysis (“data harmonization”)
- Requires interactive system for researchers and eventually patients to upload their data.
- Researchers should be able to interoperate with commons and associated clouds and “pay for compute,” with allocations, credit cards or “chits”.

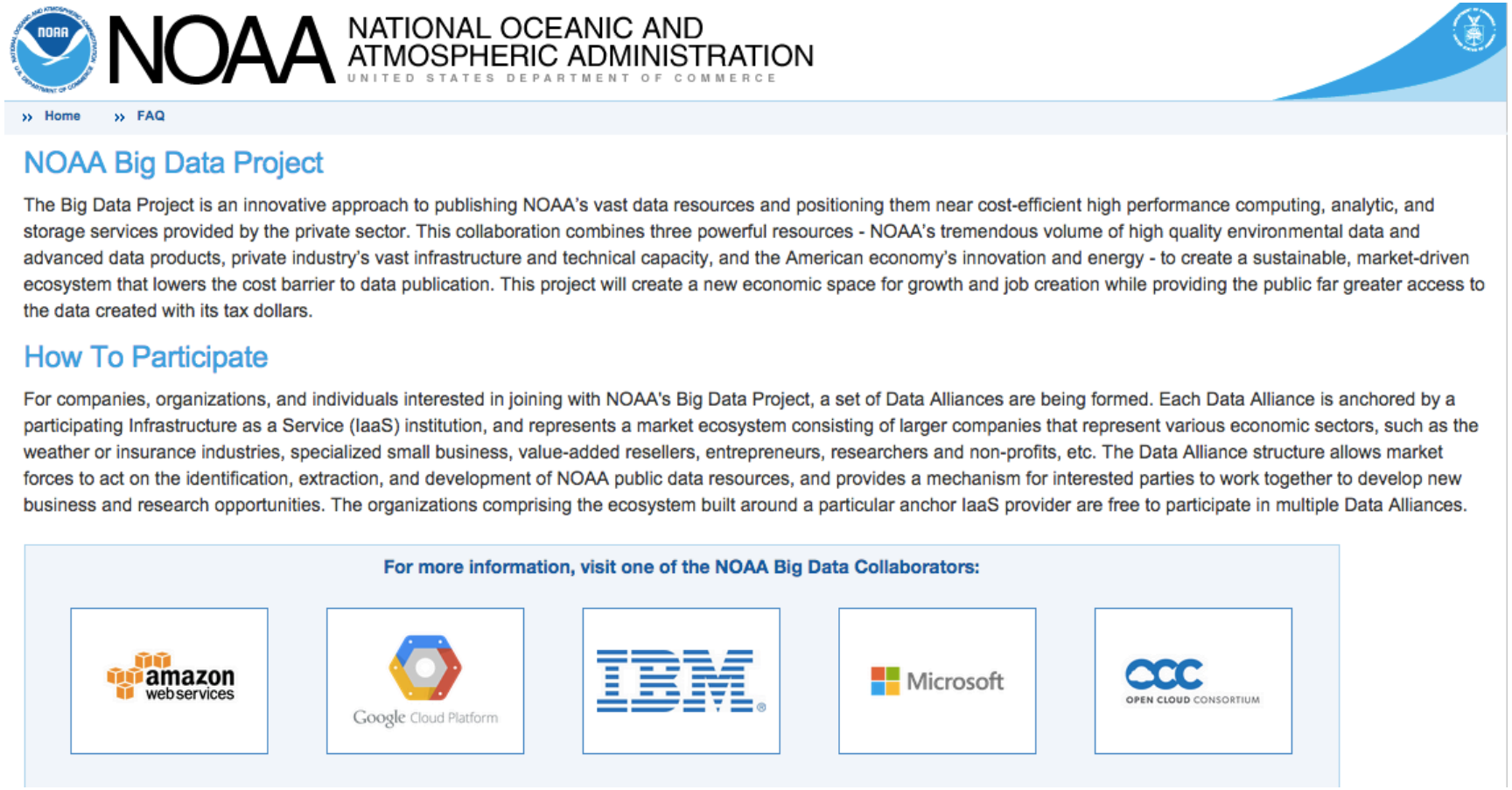




GDC API

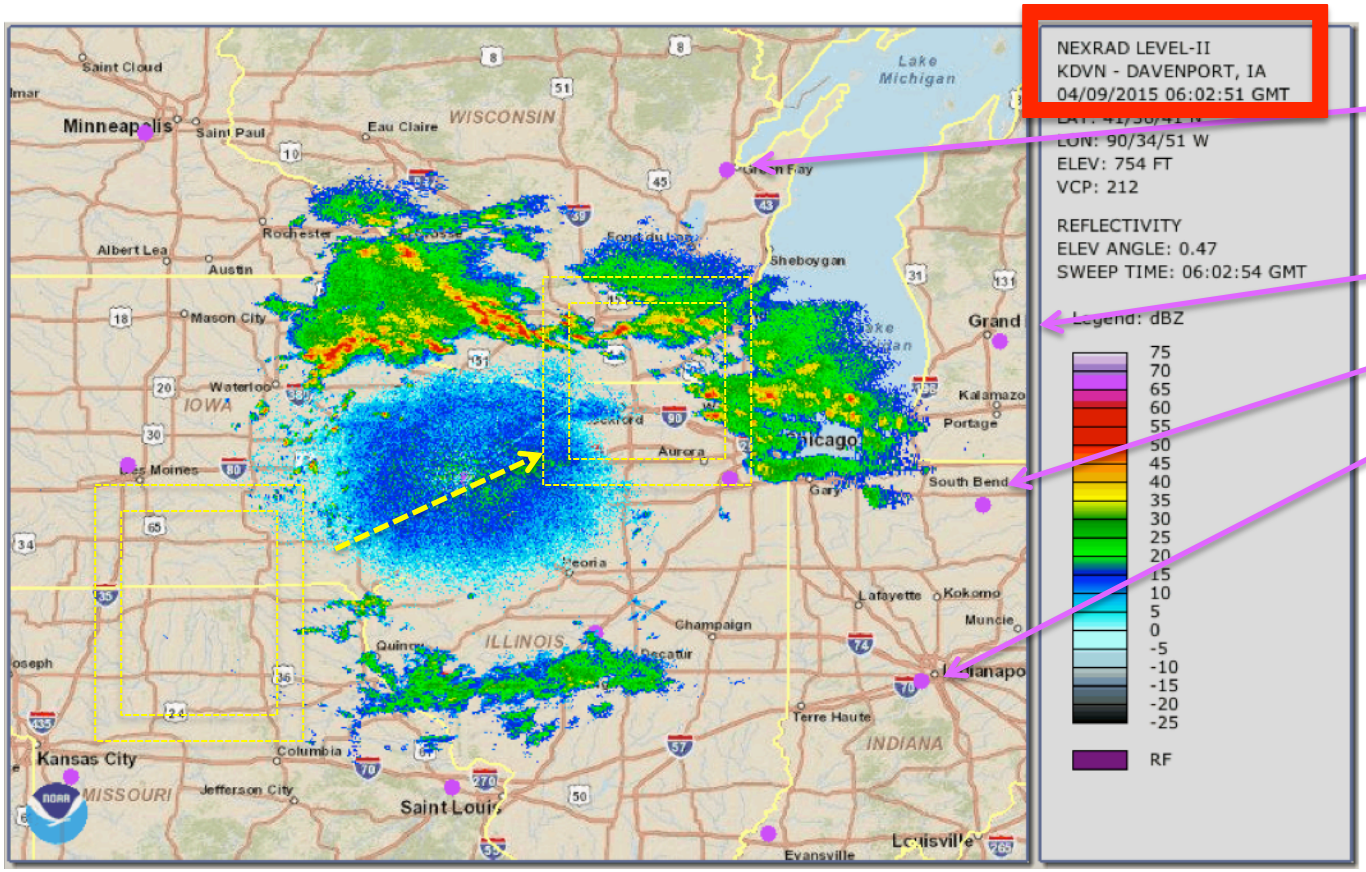
- REST API for programmatically interfacing with the GDC to query and download data
- Drives the current data portal
- Current features include:
- Search an indexed view of the GDC data model for projects, files, cases, and annotations
- Gather details about a project, file, case, or annotation
- Download files





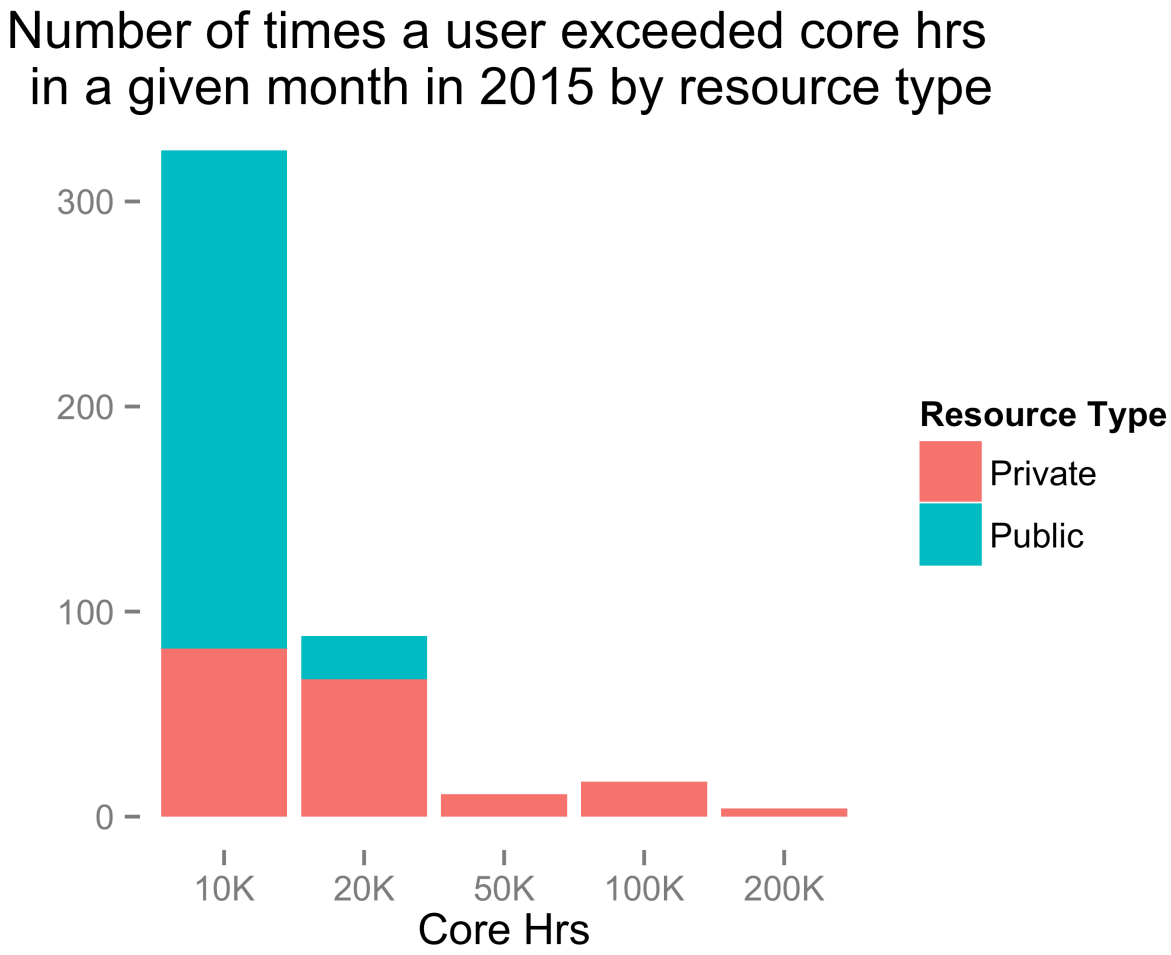
- Public-private data collaborative announced April 21, 2015 by Secretary of Commerce Pritzker.
- AWS, Google, Microsoft, IBM and the Open Commons Consortium will form five alliance partners.

NOAA NEXRAD level II radar data: Useful for severe weather



NEXRAD station locations

Seven tornadoes sweep from northern Missouri to north central Illinois April 9, 2015 (~19:00 to 23:45 GMT), including an EF-4 that killed 2, injured 11, and caused $19.25M in property damage.



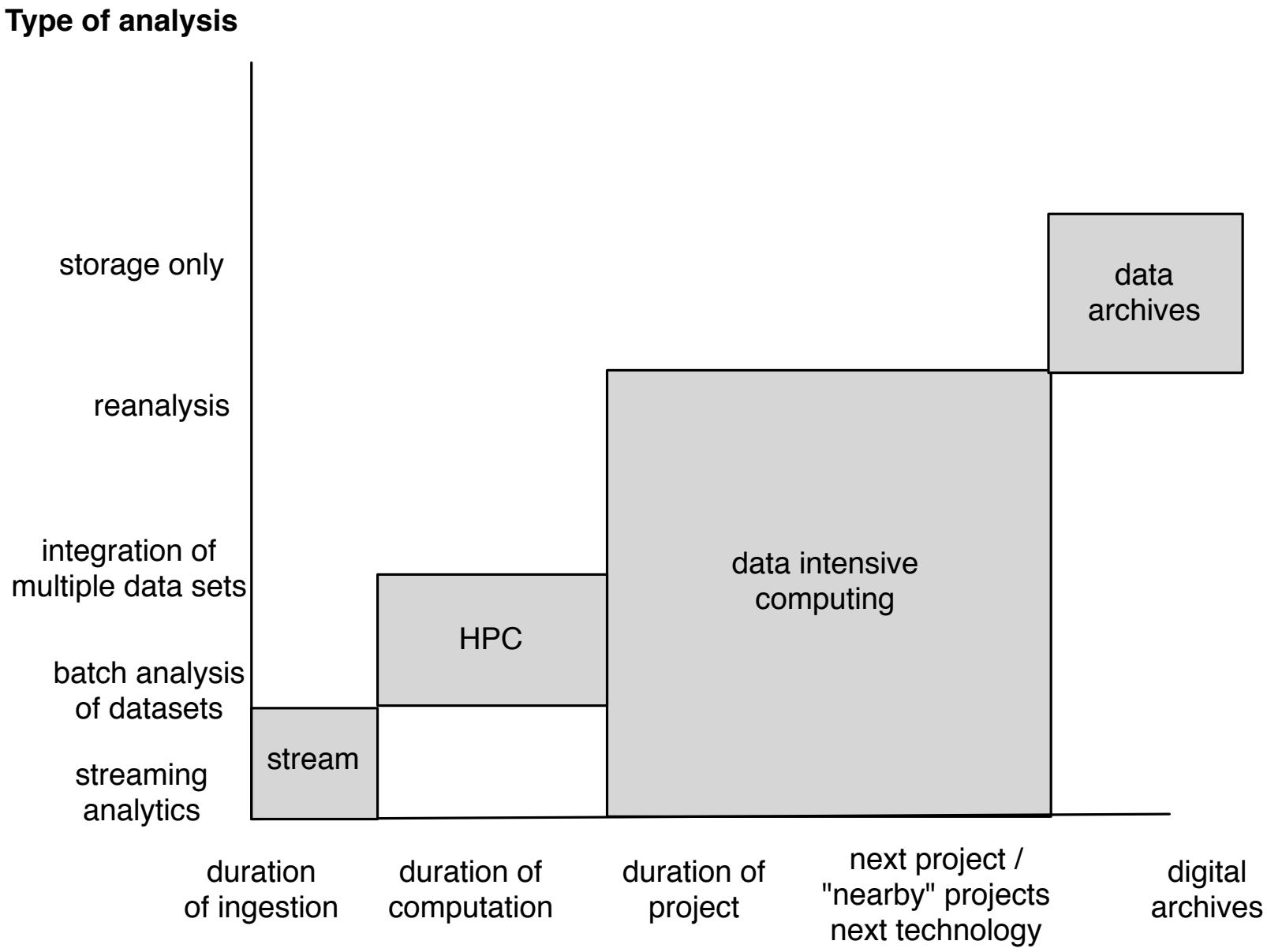
The spectrum of OSDC users – we favor users that require access to large persistent datasets and large amounts of data intensive computing over them.

The Open Science Data Cloud PIRE (2010-2015) educated over 70 students with long term placements with our PIRE Partners in Edinburgh, Amsterdam, Tokyo, and Sao Paulo. We also taught over 500 others in one day tutorials.

Vision: Commons of open data at scale will disruptively accelerate the pace of discovery and improve human welfare.

Strategy:

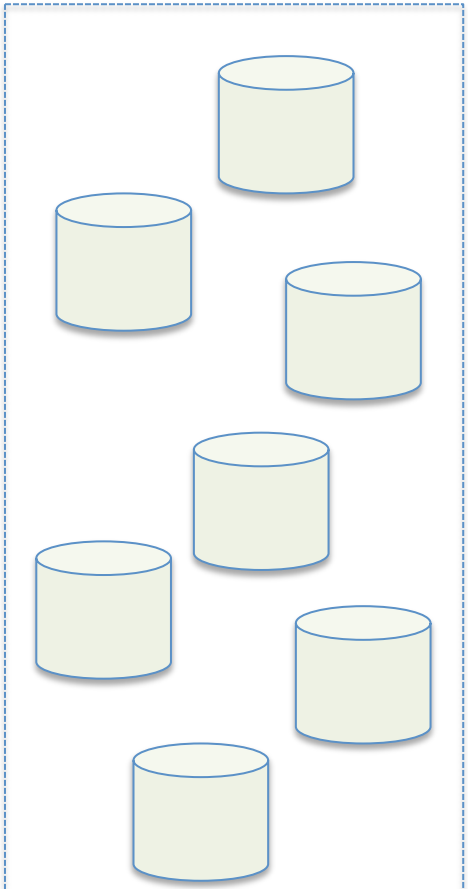
1. Develop an open source platform for open data.
2. Host a critical mass of data (data science discoveries emerge with scale)
3. Give away allocations of resources for discovery (democratize discovery).
4. Collaborate with medium size projects to support their communities.
5. Interoperate with third party applications and resources (pay for compute)
6. Impose a small tax on ecosystem to break even.



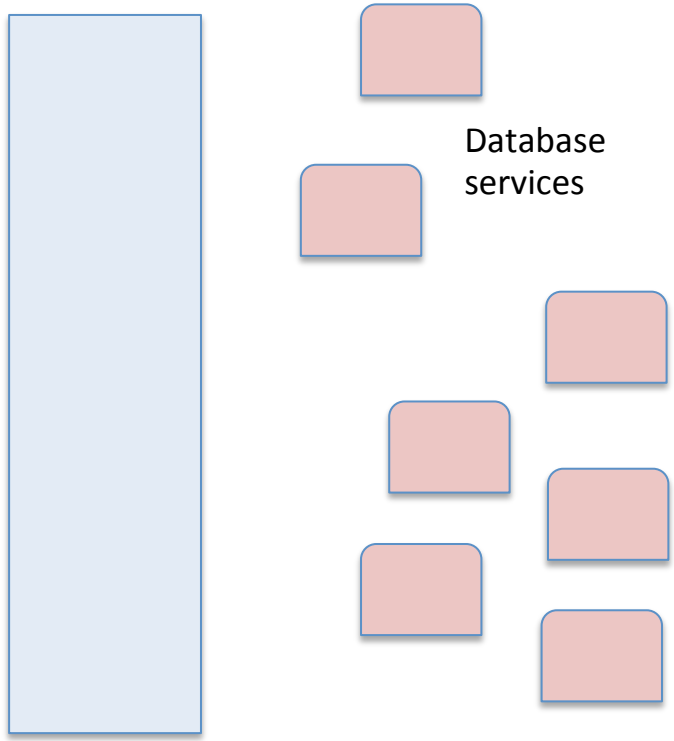
Adapted from: Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science and Engineering in 2017-2020, National Academy Press, DOI: 10.17226/21886, 2016

2. The OCC Technology Stack

The OSDC, Bionimbus and GDC share a common core software stack.



Object-based storage with access control lists



Scalable light weight workflow

Database services

Community data products, including data harmonization

Data access portal

APIs for data access

Data submission portal

APIs for data submission

Co-located "pay for compute"

Long running Middleware Services
(Digital ID, Metadata, Transport)

Devops supporting virtualized environments & containers
(OpenStack, Containers and Scheduling)

The Open Source Technology Stack

- The **open source software stack** for OSDC/ Bionbimbus/GDC consists:
  - Digital ID Service (Signpost)
  - Metadata Service (Sightseer)
  - High performance data transport (Parcel)
  - ETL & lightweight workflow services
  - Data portal and data submission systems
- We expose these services through APIs & interoperate with third party services supporting science as a service.
- The software stack can operate over private or public clouds and uses OpenStack, Ceph, Neo4j, …

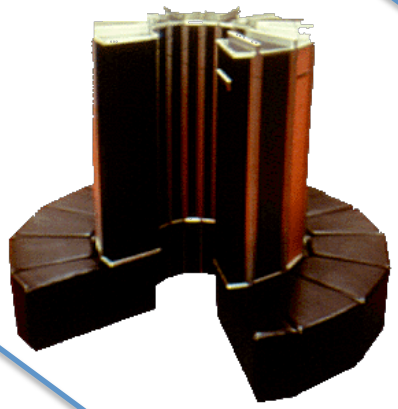
3. What Have We Learned?

Lesson 1: Think of a Commons as a New Instrument

2004
10x-100x

data science

1976
10x-100x



simulation science

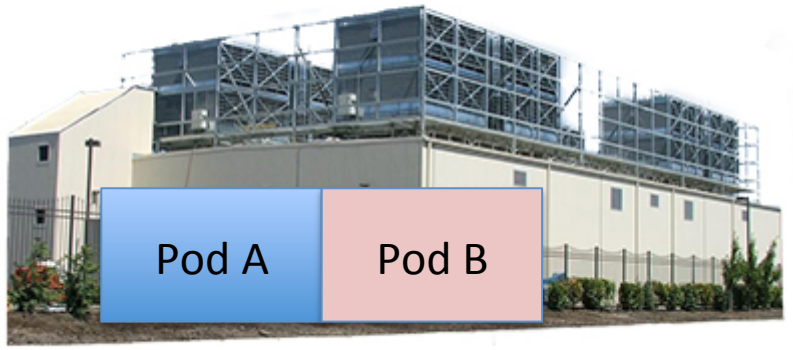
1670
250x

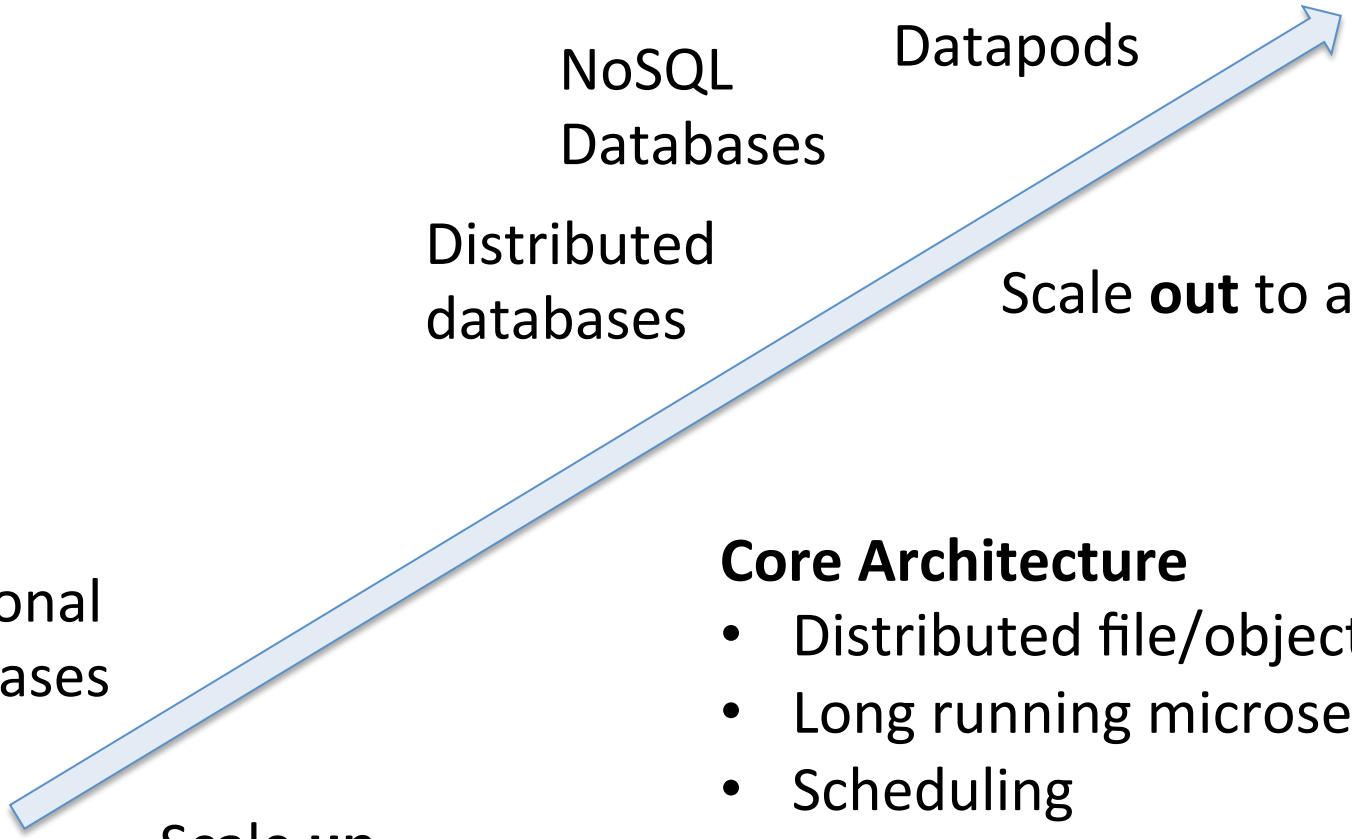
1609
30x

experimental science

What Scale?

- New data centers are sometimes divided into “pods,” which can be built out as needed.
- A reasonable scale for what is needed for a commons is one of these pods (“cyberpod”). It’s an example of mid-scale cyber infrastructure.
- Let’s use the term “datapod” for the analytic infrastructure that scales to a cyberpod.
- Think of as the scale out of a database.



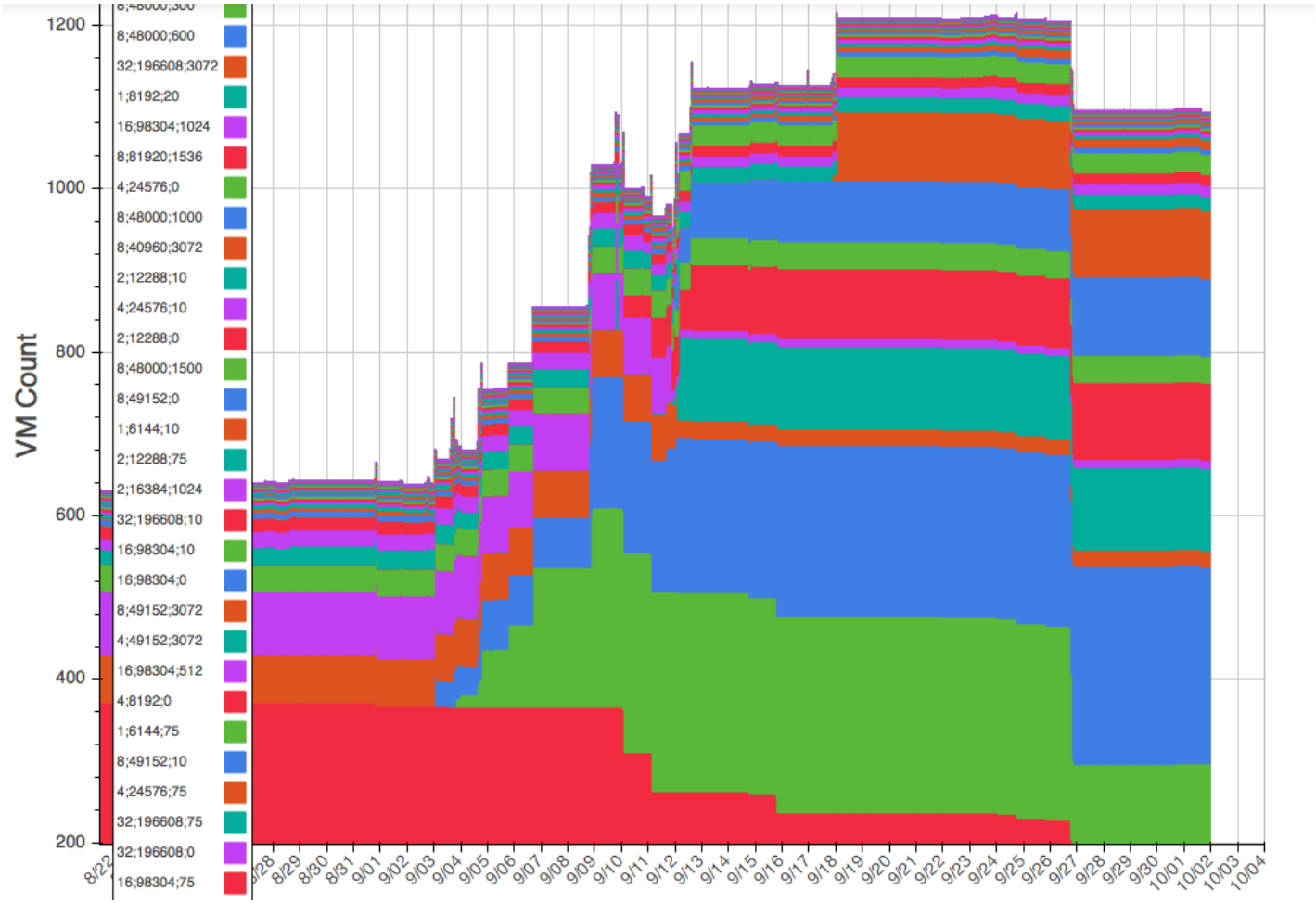


Core Architecture

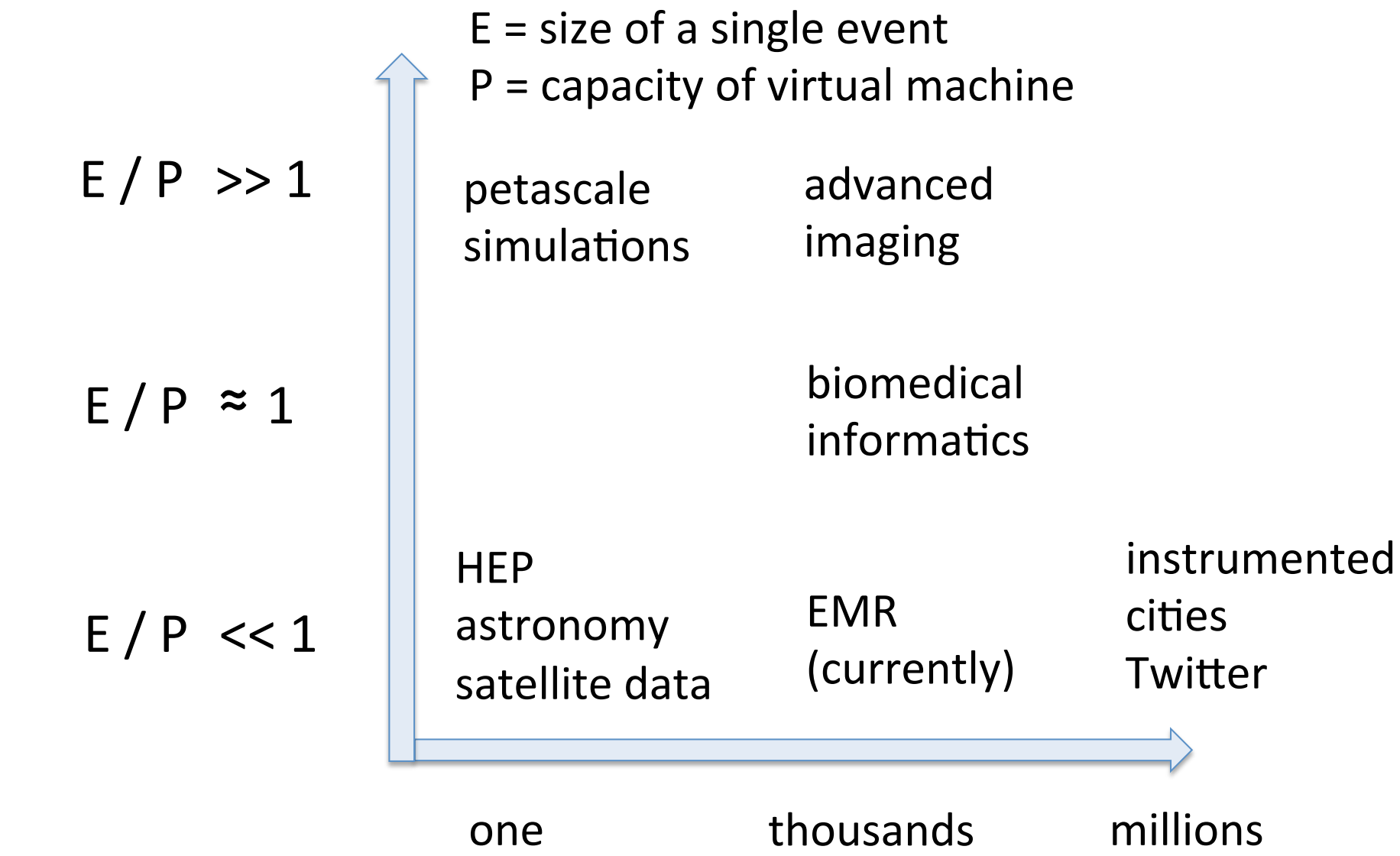
- Distributed file/object storage
- Long running microservices
- Scheduling
- Software defined networking

Goal is reduced complexity

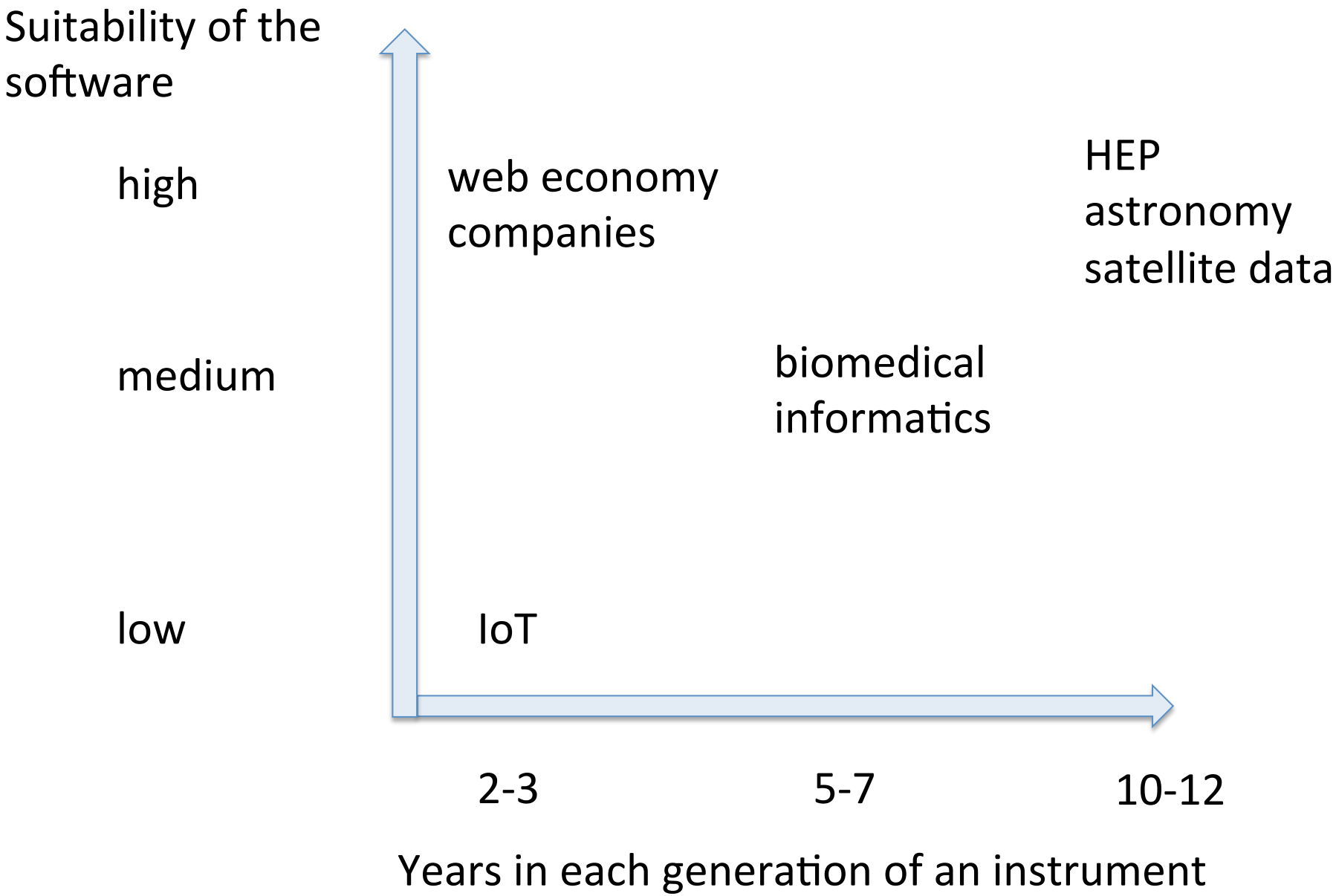
Lesson 2: We Need DevOps for Analytics That Supports Re-analysis, Scans & Mashups



Source: Center for Data Intensive Science, University of Chicago.

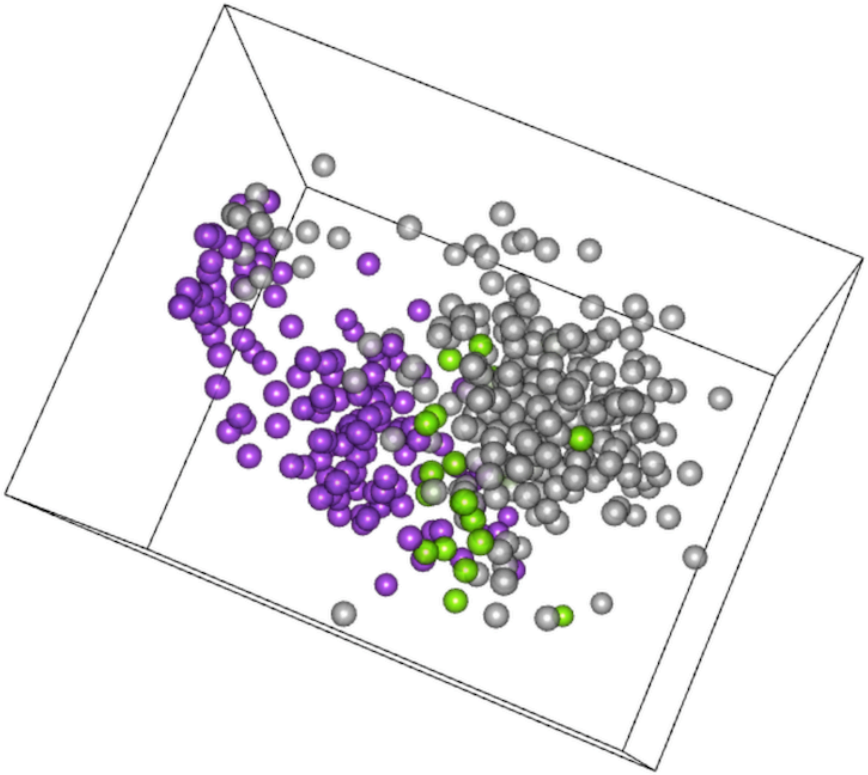


Source: Center for Data Intensive Science, University of Chicago.



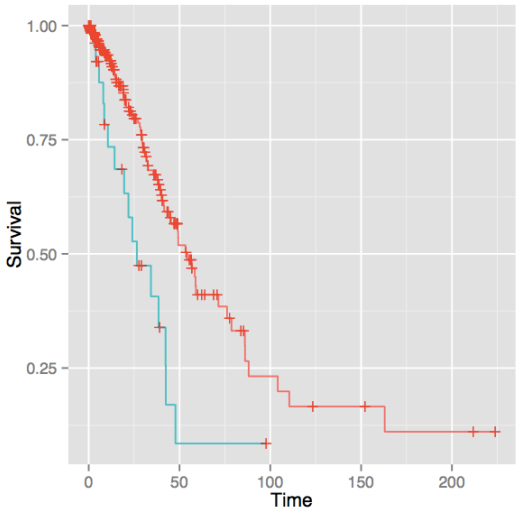
Source: Center for Data Intensive Science, University of Chicago.

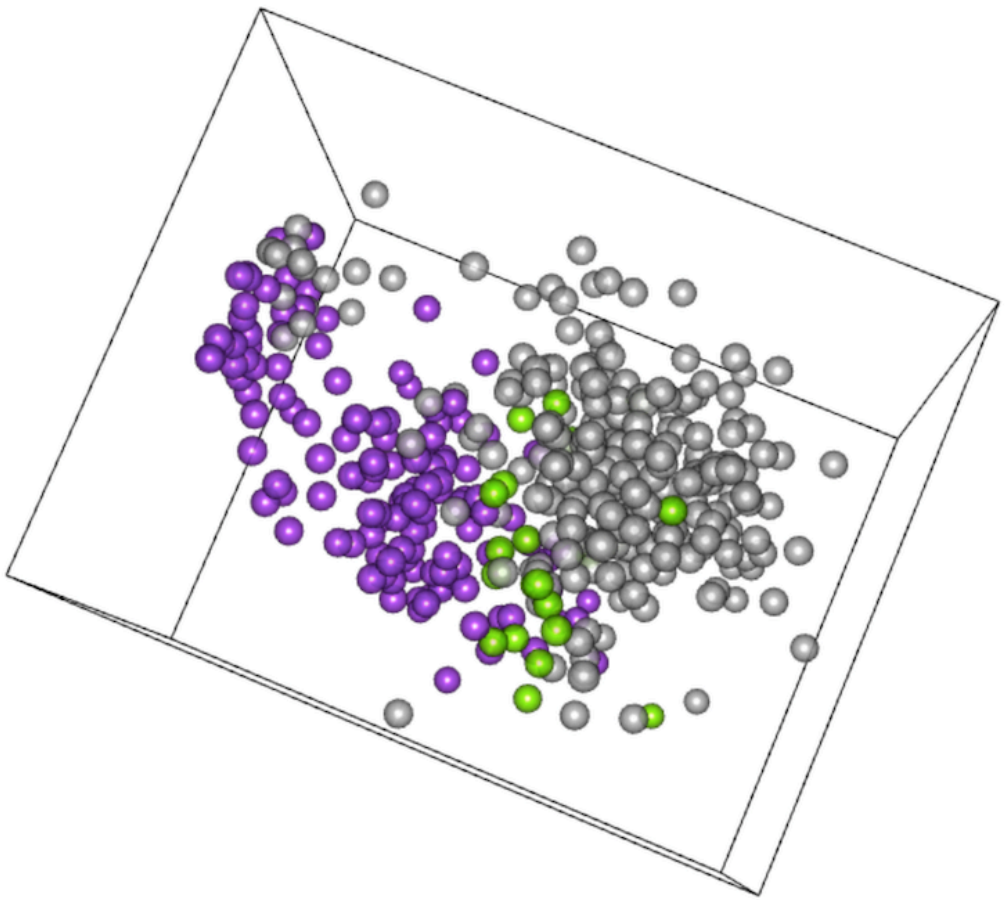
Lesson 3: Discoveries Will Best be Enabled by Higher Order and Cross Common Services



Purple balls are lung adenocarcinoma. Grey are lung squamous cell carcinoma. Green are misdiagnosed.

Source: Center for Data Intensive Science, University of Chicago.

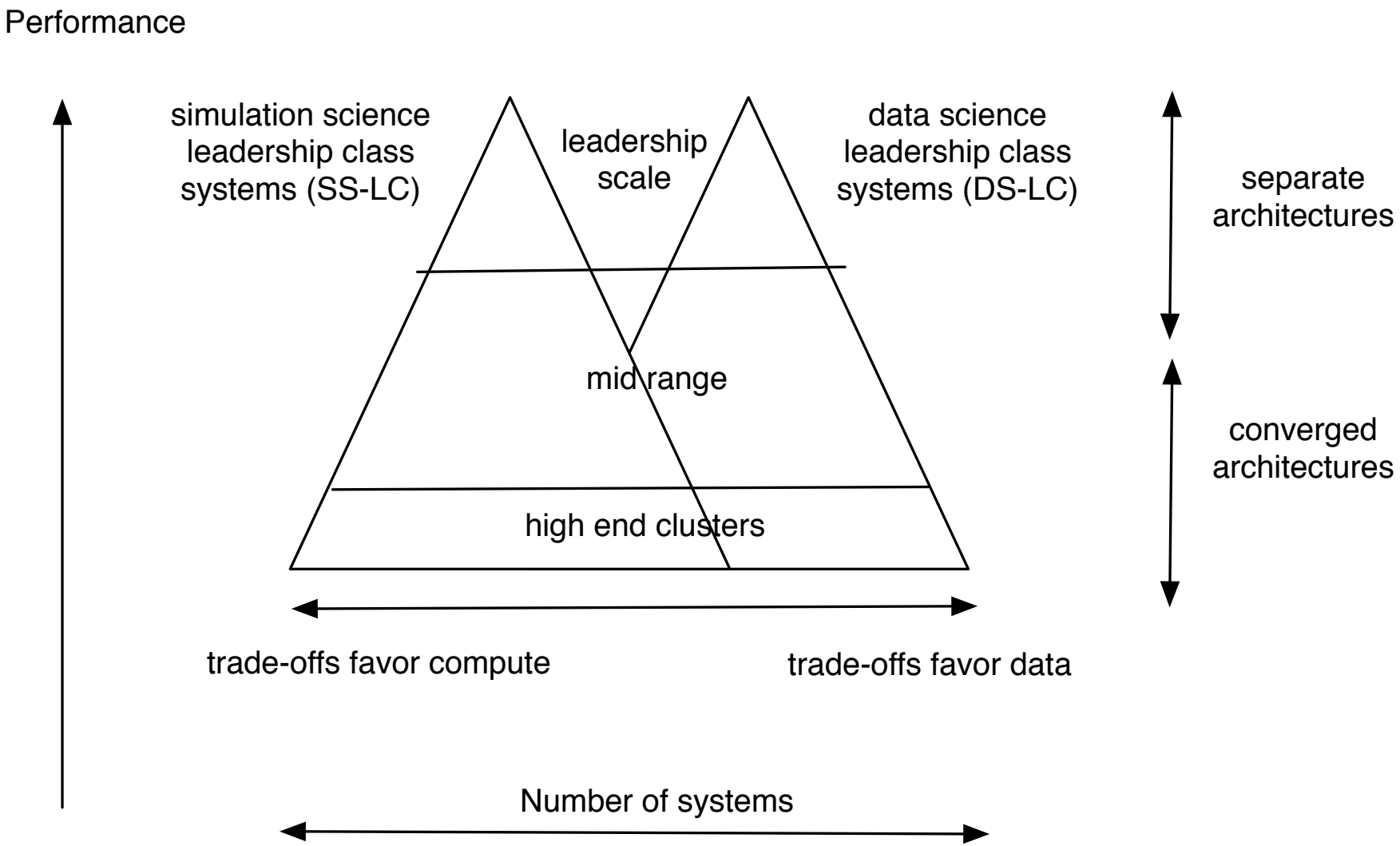




Purple balls are lung adenocarcinoma. Grey are lung squamous cell carcinoma. Green are misdiagnosed.

Source: Center for Data Intensive Science, University of Chicago.

Lesson 4: The Data Science Community Must Learn to to Design, Develop and Operate Their Own Data Commons.



4. “Chicago Model” for a Data Commons

This section is adapted from Robert L. Grossman, Allison Heath, Mark Murphy, Maria Patterson, A Case for Data Commons Towards Data Science as a Service, IEEE Computing in Science and Engineer, 2016.

Seven Requirements for the "Chicago" Data Commons Model*

1. Use **P**ermanent Digital IDs and metadata.
2. Support data **P**eering with other commons and clouds.
3. Support Application **P**rogram Interfaces for third party apps.
4. Make **P**ortable data easy.
5. Interoperate with **P**ublic and on premise clouds
6. Support cyber**P**od-scale storage and analytic services.
7. Support **P**ay for compute, with free access to research data.

*Robert L. Grossman, Allison Heath, Mark Murphy, Maria Patterson, A Case for Data Commons Towards Data Science as a Service, IEEE Computing in Science and Engineer, 2016.

Requirement 1: Permanent Secure Objects

- How do I assign Digital IDs and key metadata to data objects and collections of data objects to support distributed computation of large datasets by communities of researchers?
- The test: One commons serving the research community can transfer 1 PB of data files to another commons and no researcher needs to change their code.

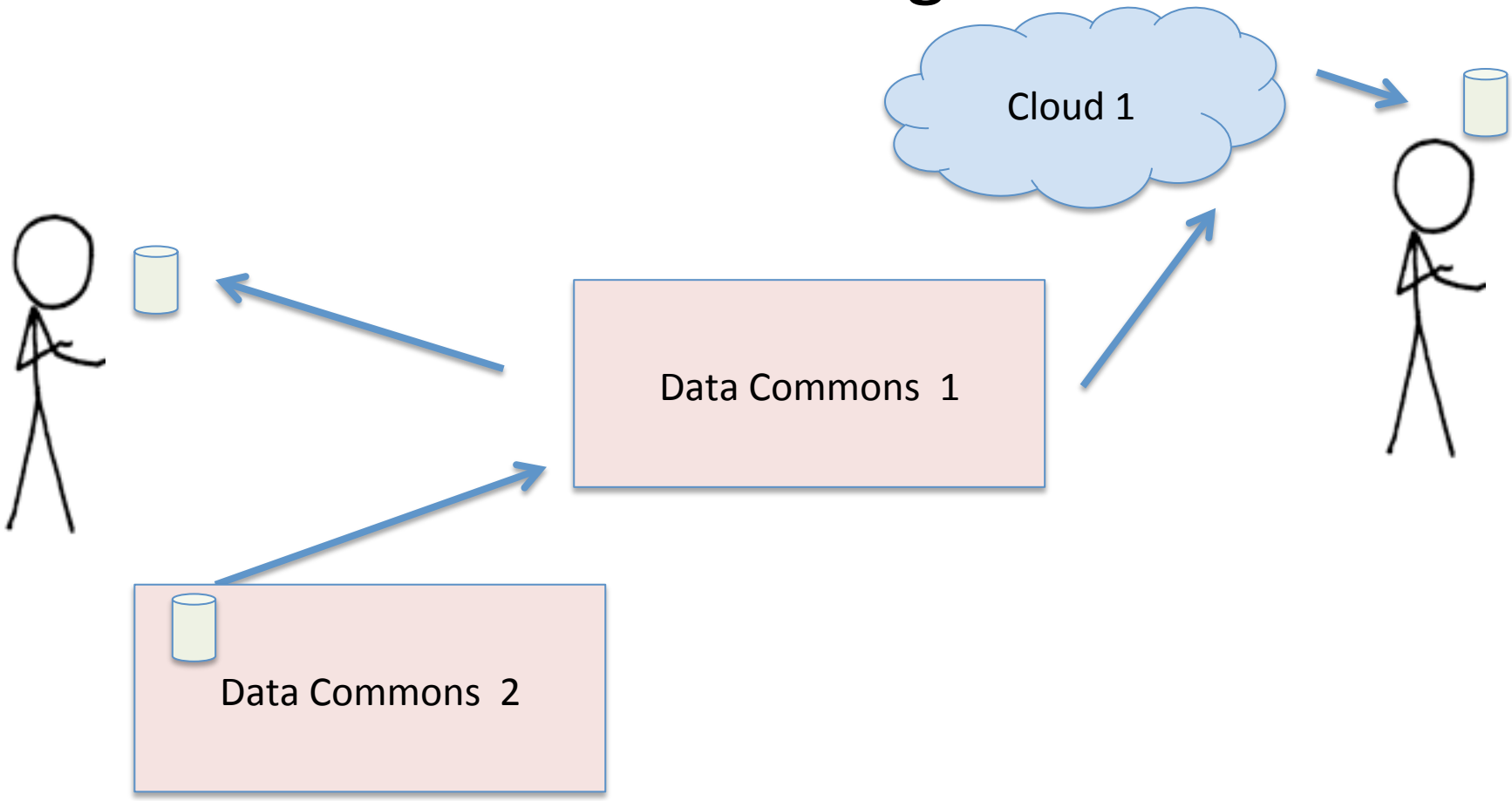
Requirement 2: Scale to Cyber Pods and Datapods

- Recall the term cyberpod refers to a computing infrastructure at the scale of a "pod" or area in a data center, say 20+ racks.
- How can I add a rack of computing/storage/networking equipment to a cyber pod (that has a manifest) so that
  - After attaching to power
  - After attaching to network
  - No other manual configuration is required
  - The data services can make use of the additional infrastructure
  - The compute services can make use of the additional infrastructure
- In other words, we need an open source software stack that scales to cyberpods and data analysis that scales to datapods.

Requirement 3: Data Peering with Other Commons

- How can a critical mass of data commons support data peering so that a research at one of the commons can transparently access data managed by one of the other commons
  - We need to access data independent of where it is stored
  - "Tier 1 data commons" need to pass community managed data at no cost
  - We need to be able to transport large data efficiently "end to end" between commons

Data Peering



1. To transfer research data between them at no cost.
2. To peer with at least two other Tier 1 Research Data Commons at 10 Gbps or higher.
3. To support Digital IDs (of a form to be determined by mutual agreement) so that a researcher using infrastructure associated with one Tier 1 Research Data Commons can access data transparently from any of the Tier 1 Research Data Commons that hold the desired data.

Requirement 4: Data Portability

- We need a simple “button” to move our data between two commons that peer.

Requirement 5: Interoperate with Public Clouds

- With high utilization of compute and sufficient scale, it is less expensive to use dedicated hardware in a well run data data center (private cloud) than to use a public cloud.
- On the other hand, large scale public clouds provide the ability to "burst" to a very large scale capacity with excellent security, whenever is needed for as long as needed.
- A hybrid model that integrates a private cloud with a public cloud is a good foundation for a large scale commons.

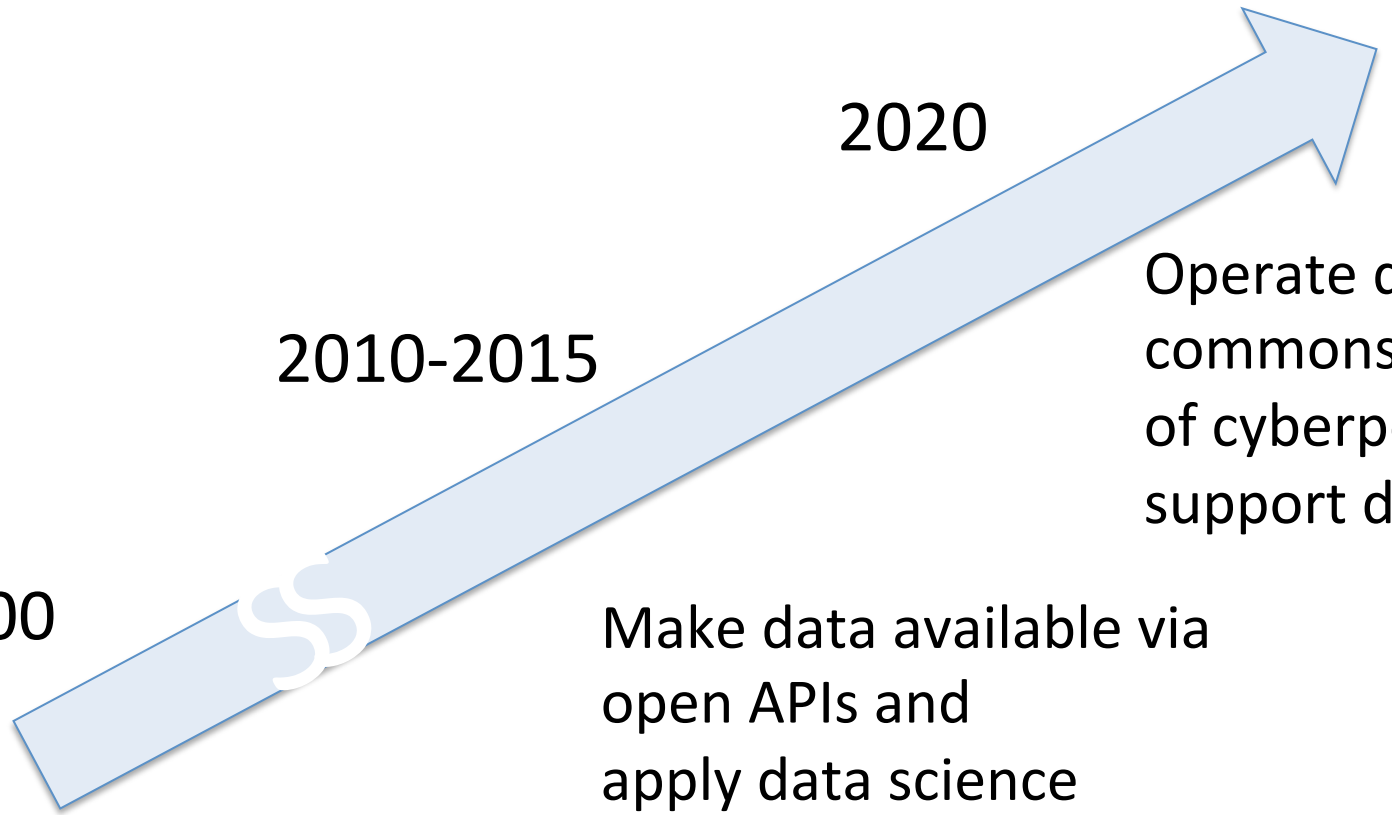
Requirement 6: Application Program Interfaces

- For data access and data queries.
- For high performance data transport.
- For access to the computing infrastructure.
- etc.

Requirement 7: Support Pay for Compute

- Commons should support a “free peering for research data, pay for compute,” model.
- Pay can be via allocations, grants, credit cards or “chits.”
- Without, some allocation mechanism there is a quick transition to the tragedy of the commons.

5. Summary


2020
2010-2015
2000
Operate data commons at the scale of cyberpods & support data peering
Make data available via open APIs and apply data science
Collect data and distribute files via DCC and apply informatics tools and data mining

Questions?

rgrossman.com
@bobgrossman

For More Information

Robert L. Grossman

rgrossman.com

@BobGrossman

robert.grossman@uchicago.edu

Division of the Biological Sciences
Center for Data Intensive Science

cdis.uchicago.edu

Funding

Funding for the projects described in this talk have been provided by the following:

- The Genomics Data Commons is supported in part by NIH/Leidos Biomedical Research, Inc. through contract 14X050 / HHSN261200800001E to the University of Chicago
- The Bionimbus Protected Data Cloud is supported in by part by NIH/NCI through NIH/Leidos Contract 13XS021 / HHSN261200800001E to the University of Chicago
- Major funding and support for the Open Science Data Cloud (OSDC) is provided by the Gordon and Betty Moore Foundation (2010-2015)
- The OSDC is supported by a 5-year (2010-2016) PIRE award (OISE – 1129076) to train scientists to use the OSDC and to further develop the underlying technology.
- Technology for high performance data transport used by the clouds and commons discussed were supported in part by NSF Award 1127316.
- This work was supported in part by BIGDATA: Small: DCM: Open Flow Enabled Hadoop over Local and Wide Area Clusters, NSF Award 1251201.
- Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, NIH or other funders of this research.

The OSDC, Project Matsu, OCC NOAA Data Commons, and Biomedical Data Commons are managed by the Open Commons Consortium, a 501(c)(3) not-for-profit corporation. If you are interested in providing funding or donating equipment or services, please contact us at info@occ-data.org.

Abstract

How Data Commons Are Changing the Way We Share Research Data and Make Discoveries

Large scale data commons are beginning to change the way that we share research data and make scientific discoveries. In this talk, we survey some of the large scale scientific clouds and data commons, highlight some case studies, and discuss some of the emerging opportunities and challenges.