



# 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

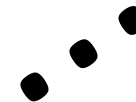
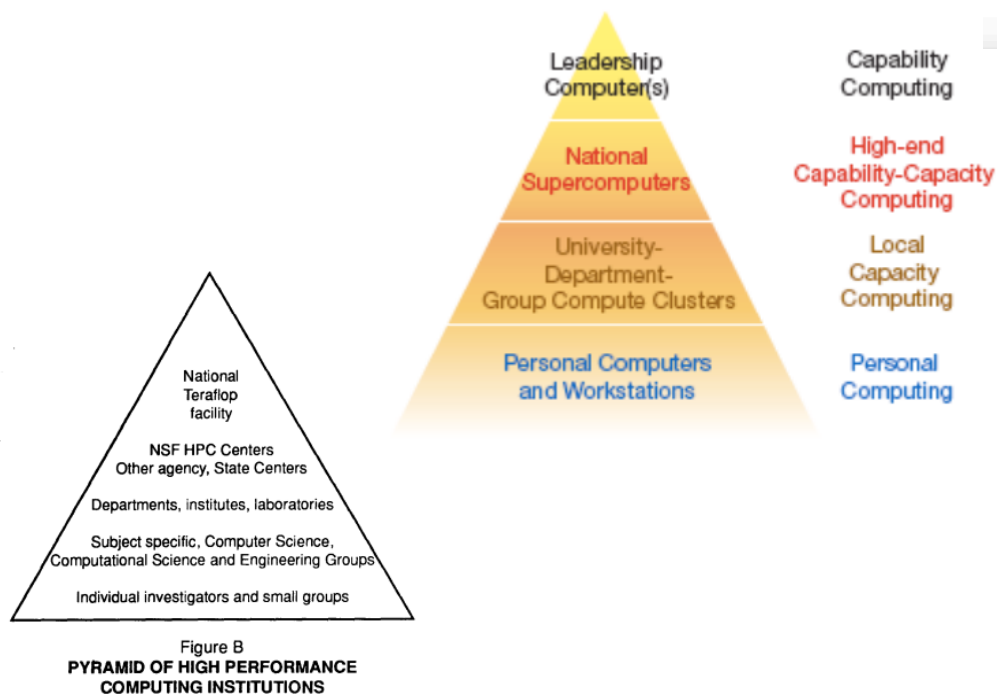
The work discussed here is joint work with the various researchers at the University of Chicago's Center for Data Intensive Science (CDIS) and with the Open Commons Consortium and its Working Groups.

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.

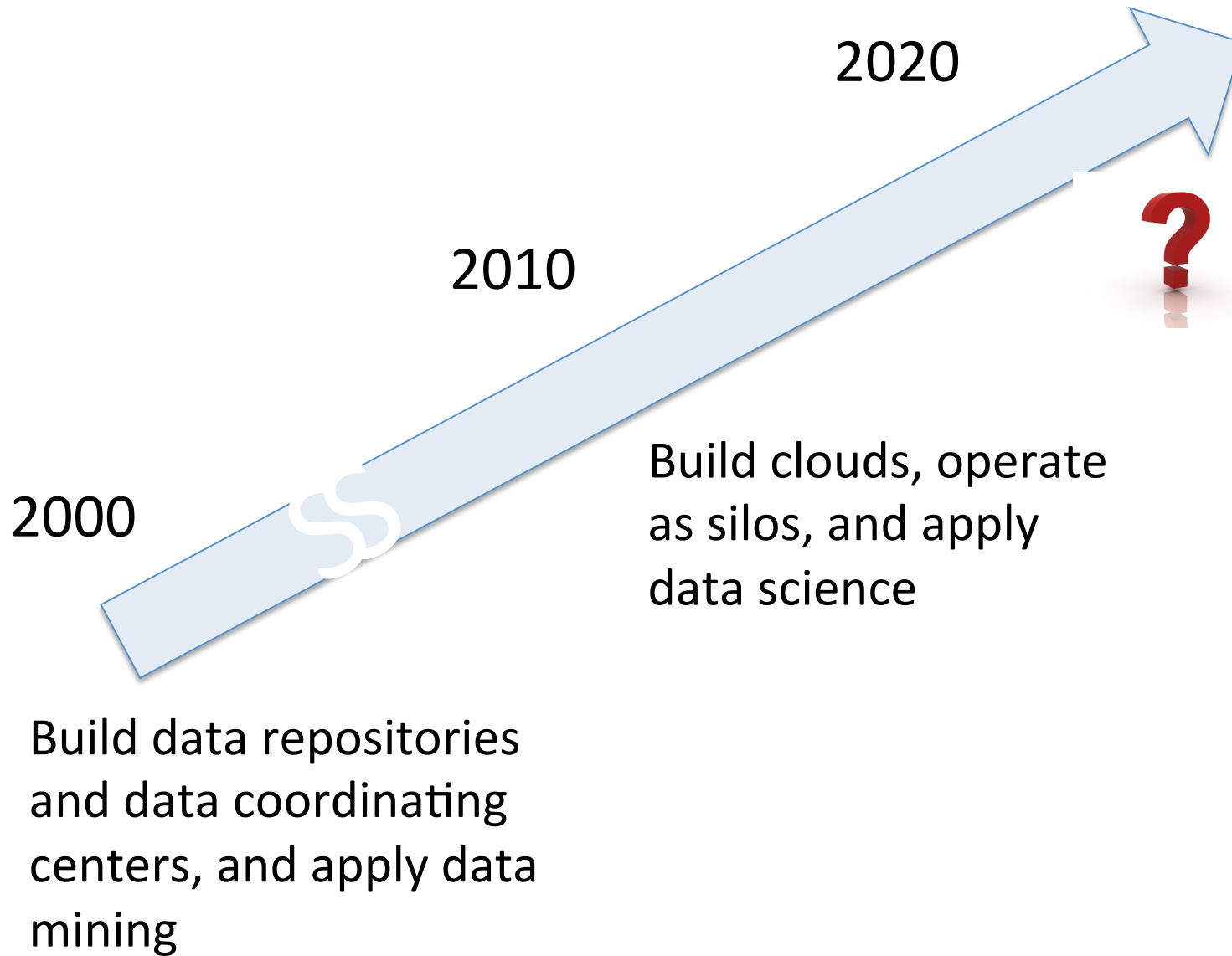
The design and architecture of the Bionimbus Protected Data Cloud is described in: Allison P. Heath, Matthew Greenway, Raymond Powell, Jonathan Spring, Rafael Suarez, David Hanley, Chai Bandlamudi, Megan McNerney, Kevin White and Robert L Grossman, Bionimbus: A Cloud for Managing, Analyzing and Sharing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



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](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.



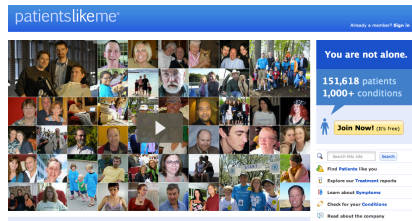


# 1. Data Commons

# 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.

## OCC Project Matsu

An open source project for cloud-based processing of satellite imagery to support the earth sciences.

Project Matsu – NASA  
Collaboration (2009)



Open Science Data Cloud  
(2010)

PDC Console Apply Status

## BIONIMBUS PROTECTED DATA CLOUD

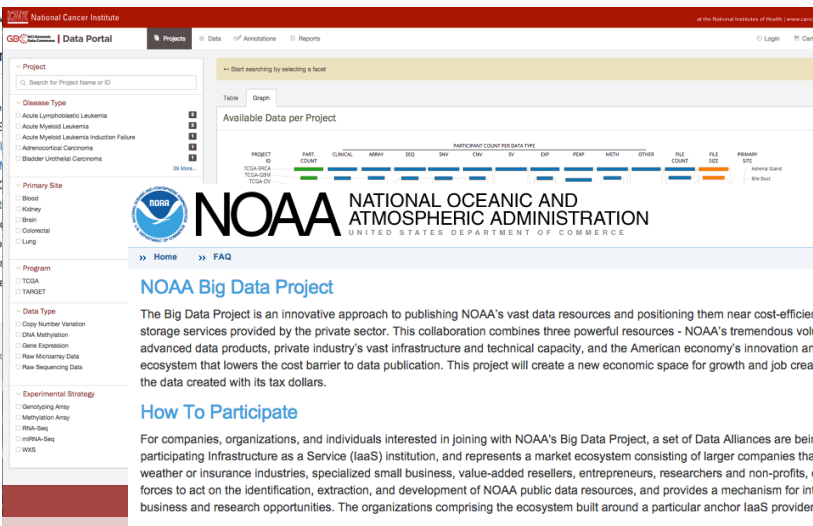
Bionimbus Protected  
Data Cloud\* (2013)

Secure cloud service

What is the Bionimbus

The Bionimbus Protected Data Cloud (OSDC) and the IGS Research Informatics (CRI) Translational Medicine (TM) Chicago Comprehensive Center (CCC) The PDC allows users to securely compute over human genomic datasets from the The Cancer Genome Atlas (TCGA) are available in the

OSDC OPEN SCIENCE DATA CLOUD



NCI Data Commons\*  
(2015)



## NOAA Big Data Project

The Big Data Project is an innovative approach to publishing NOAA's vast data resources and positioning them near cost-efficient high performance computing, analytic, and storage services provided by the private sector. This collaboration combines three powerful resources - NOAA's tremendous volume of high quality environmental data and advanced data products, private industry's vast infrastructure and technical capacity, and the American economy's innovation and energy - to create a sustainable, market-driven ecosystem that lowers the cost barrier to data publication. This project will create a new economic space for growth and job creation while providing the public far greater access to the data created with its tax dollars.

### How To Participate

For companies, organizations, and individuals interested in joining with NOAA's Big Data Project, a set of Data Alliances are being formed. Each Data Alliance is anchored by a participating Infrastructure as a Service (IaaS) institution, and represents a market ecosystem consisting of larger companies that represent various economic sectors, such as the weather or insurance industries, specialized small business, value-added resellers, entrepreneurs, researchers and non-profits, etc. The Data Alliance structure allows market forces to act on the identification, extraction, and development of NOAA public data resources, and provides a mechanism for interested parties to work together to develop new business and research opportunities. The organizations comprising the ecosystem built around a particular anchor IaaS provider are free to participate in multiple Data Alliances.

For more information, visit one of the NOAA Big Data Collaborators:



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: OCC 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](http://www.occ-data.org)

# Open Science Data Cloud

Cloud services for the scientific community

## What is the OSDC?

We provide and support cloud computing and storage services for the scientific research community. The OSDC is run by the [Open Cloud Consortium](#), a non-profit organization whose primary goal is to support scientific advances by working with researchers in a variety of disciplines.

## Why use the OSDC?

Our cloud services are based on the principles of openness and interoperability. The OSDC infrastructure is tailored towards the high performance storage and compute resources often required for scientific discovery. We view the OSDC as complementary to commercial cloud services available.

## How do I get started?

First, apply for an account. Once you have an account, you can take a look at the OSDC instructions, login to the console and get started.

[Apply Now](#)

[Login to the OSDC Console](#)

The OSDC is a resource of the [Open Cloud Consortium](#) and made possible by our [sponsors](#).

- 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](http://www.opensciencedatacloud.org)

GORDON AND BETTY  
**MOORE**  
FOUNDATION

# OCC Project Matsu

An open source project for cloud-based processing of satellite imagery to support the earth sciences.

## Project Matsu

Project Matsu is a collaboration between NASA and the Open Cloud Consortium to develop open source technology for cloud-based processing of satellite imagery to support the earth sciences. Technology developed by the collaboration include:

- The Namibia Flood Dashboard.
- MapReduced based analytics for identifying floods and CO<sub>2</sub> concentrations.
- Using elastic infrastructure-as-a-service to create Level 1 images each day.
- A Hadoop-based OGC-compliant tiling service and Web MapService.

## Matsu Resources

- Daily Namibia [Flood Dashboard](#)
- Hadoop-supported Web Map Service Areas of Interest:
  - [Carbon Monoxide cluster centers from volcanic eruption](#)
  - [Irrigation Patterns in the Sahara](#)
  - [Elevation](#)
  - [Water Classifier Namibia](#)
  - [Water Classifier Italian coast](#)
- [Available Matsu images](#)

## Matsu Support

- The code can be found at the [OCC Github site](#).
- Description of Matsu [Tile and WMS Services](#).

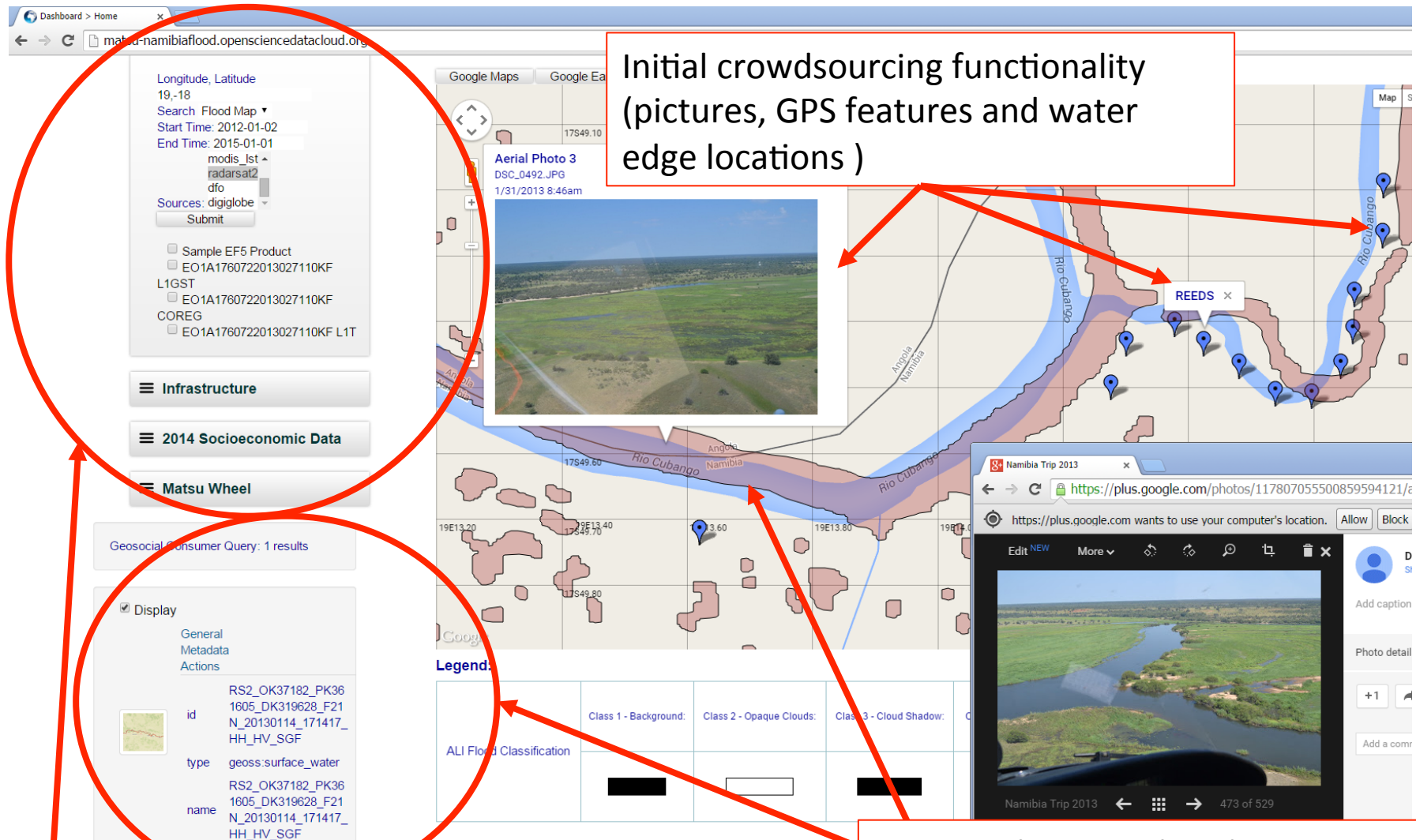
[matsu.opensciencedatacloud.org](https://matsu.opensciencedatacloud.org)

The OSDC is a resource of the [Open Cloud Consortium](#) and made possible by our [sponsors](#).

OCC-NASA Collaboration 2009 - present



# Flood/Drought Dashboard Examples



GeoSocial API Consumer embedded in Dashboard

Source: Flood Dashboard, Dan Mandl, NASA



# The Bionimbus Protected Data Cloud

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](mailto:info@opencloudconsortium.org).

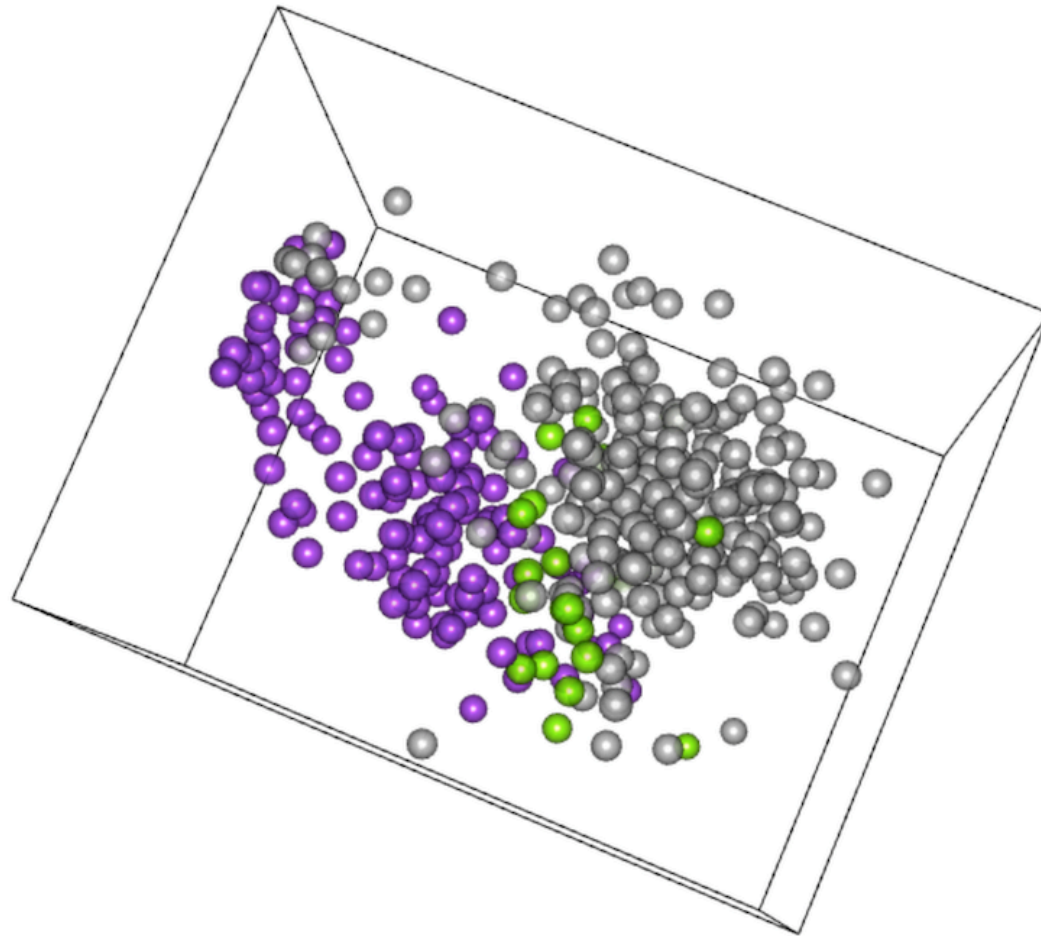
### How do I get started?

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](mailto:support@opencloudconsortium.org).

Apply for the PDC Now

Login to the PDC Console






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

 **NATIONAL CANCER INSTITUTE**  
GDC Data Portal

[Projects](#) [Data](#) [Annotations](#) [Reports](#)

[Login](#) [Cart](#) [GDC Apps](#)

## Harmonized Cancer Datasets

### Genomic Data Commons Data Portal

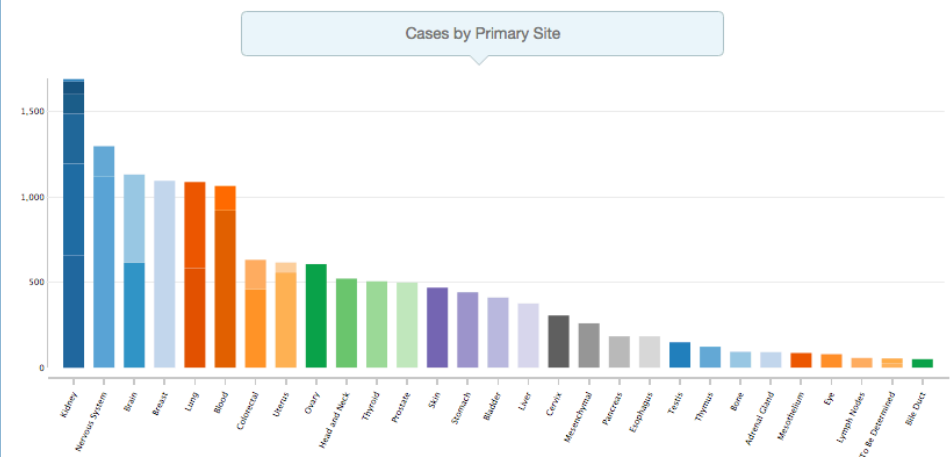
Get Started by Exploring:

[Projects](#) [Data](#)

Perform Advanced Search Queries, such as:

|  |           |             |
|--|-----------|-------------|
| Kidney cancer cases under the age of 20 at diagnosis | 128 Cases | 1,159 Files |
| CNV data of female brain cancer cases                | 459 Cases | 7,809 Files |
| Germline mutation data in TCGA-OV project            | 423 Cases | 1,700 Files |





### Cases by Primary Site



| Primary Site     | Cases |
|------------------|-------|
| Kidney           | 1,159 |
| Nervous System   | 1,159 |
| Brain            | 1,159 |
| Breast           | 1,159 |
| Lung             | 1,159 |
| Blood            | 1,159 |
| Colorectal       | 1,159 |
| Ovary            | 1,159 |
| Head and Neck    | 1,159 |
| Thyroid          | 1,159 |
| Prostate         | 1,159 |
| Skin             | 1,159 |
| Stomach          | 1,159 |
| Bladder          | 1,159 |
| Liver            | 1,159 |
| Cervix           | 1,159 |
| Mesothelioma     | 1,159 |
| Pancreas         | 1,159 |
| Esophagus        | 1,159 |
| Testis           | 1,159 |
| Thyroid          | 1,159 |
| Bone             | 1,159 |
| Adrenal Gland    | 1,159 |
| Melanoma         | 1,159 |
| Eye              | 1,159 |
| Lymph Nodes      | 1,159 |
| To be Determined | 1,159 |
| Bile Duct        | 1,159 |

### Data Portal Summary

Latest Release #0 - May 2, 2016

|  |   |   |  |
|--|---|---|--|
| <b>PROJECTS</b><br> <b>46</b> | <b>PRIMARY SITES</b><br> <b>29</b> | <b>CASES</b><br> <b>14,194</b> | <b>FILES</b><br> <b>594,527</b> |
|--|---|---|--|

### Infrastructure

Data is continuously being processed and harmonized by the GDC.  
System stats:

### Documentation

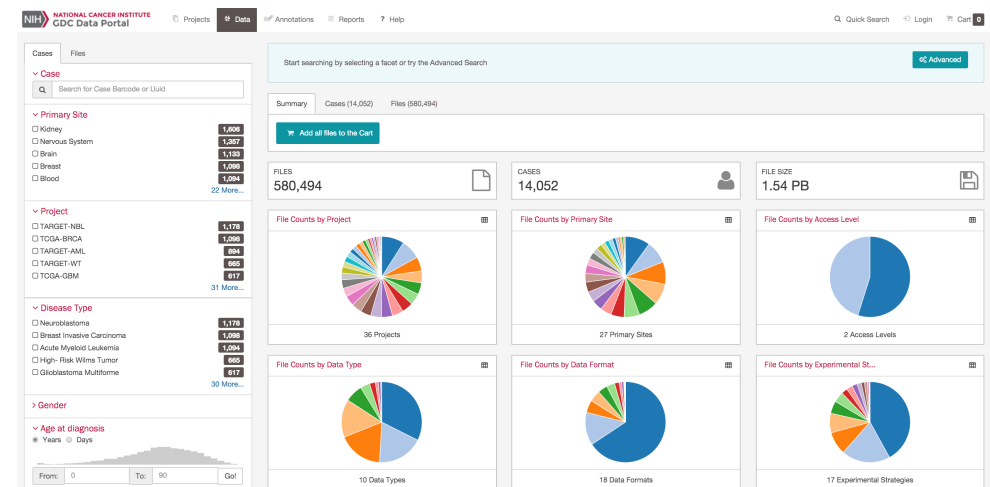
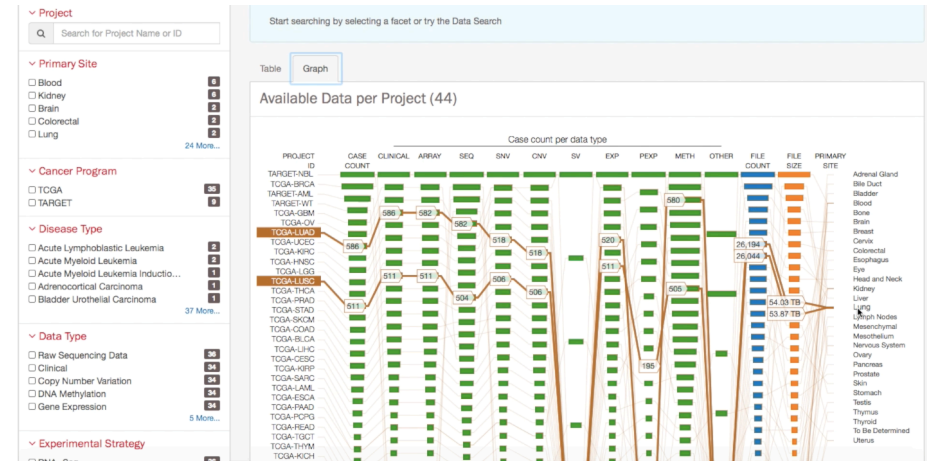
Learn how to use the GDC Data Portal to its full potential with common topics such as:

### GDC Applications

The GDC Data Portal is a robust data-driven platform that allows cancer researchers and bioinformaticians to search and download cancer data for analysis. The GDC applications include:

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

| API URL                     | Endpoint | Optional Entity ID                   | Query parameters |
|-----------------------------|----------|--------------------------------------|------------------|
| https://gdc-api.nci.nih.gov | /files/  | 5003adf1-1cfd-467d-8234-0d396422a4ee | ?fields=state    |



# NOAA

NATIONAL OCEANIC AND  
ATMOSPHERIC ADMINISTRATION  
UNITED STATES DEPARTMENT OF COMMERCE



[» Home](#)   [» FAQ](#)

## NOAA Big Data Project

The Big Data Project is an innovative approach to publishing NOAA's vast data resources and positioning them near cost-efficient high performance computing, analytic, and storage services provided by the private sector. This collaboration combines three powerful resources - NOAA's tremendous volume of high quality environmental data and advanced data products, private industry's vast infrastructure and technical capacity, and the American economy's innovation and energy - to create a sustainable, market-driven ecosystem that lowers the cost barrier to data publication. This project will create a new economic space for growth and job creation while providing the public far greater access to the data created with its tax dollars.

## How To Participate

For companies, organizations, and individuals interested in joining with NOAA's Big Data Project, a set of Data Alliances are being formed. Each Data Alliance is anchored by a participating Infrastructure as a Service (IaaS) institution, and represents a market ecosystem consisting of larger companies that represent various economic sectors, such as the weather or insurance industries, specialized small business, value-added resellers, entrepreneurs, researchers and non-profits, etc. The Data Alliance structure allows market forces to act on the identification, extraction, and development of NOAA public data resources, and provides a mechanism for interested parties to work together to develop new business and research opportunities. The organizations comprising the ecosystem built around a particular anchor IaaS provider are free to participate in multiple Data Alliances.

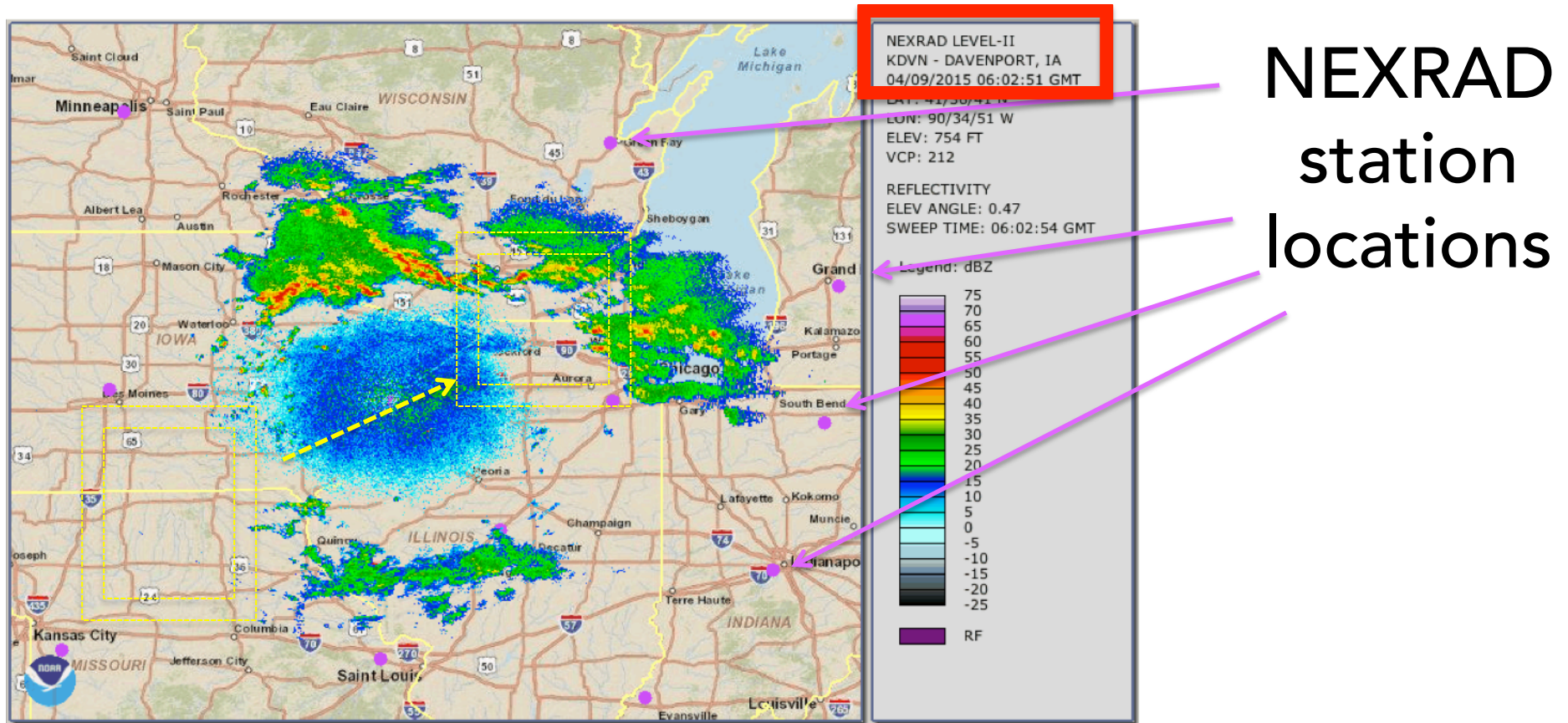
For more information, visit one of the NOAA Big Data Collaborators:



- 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.

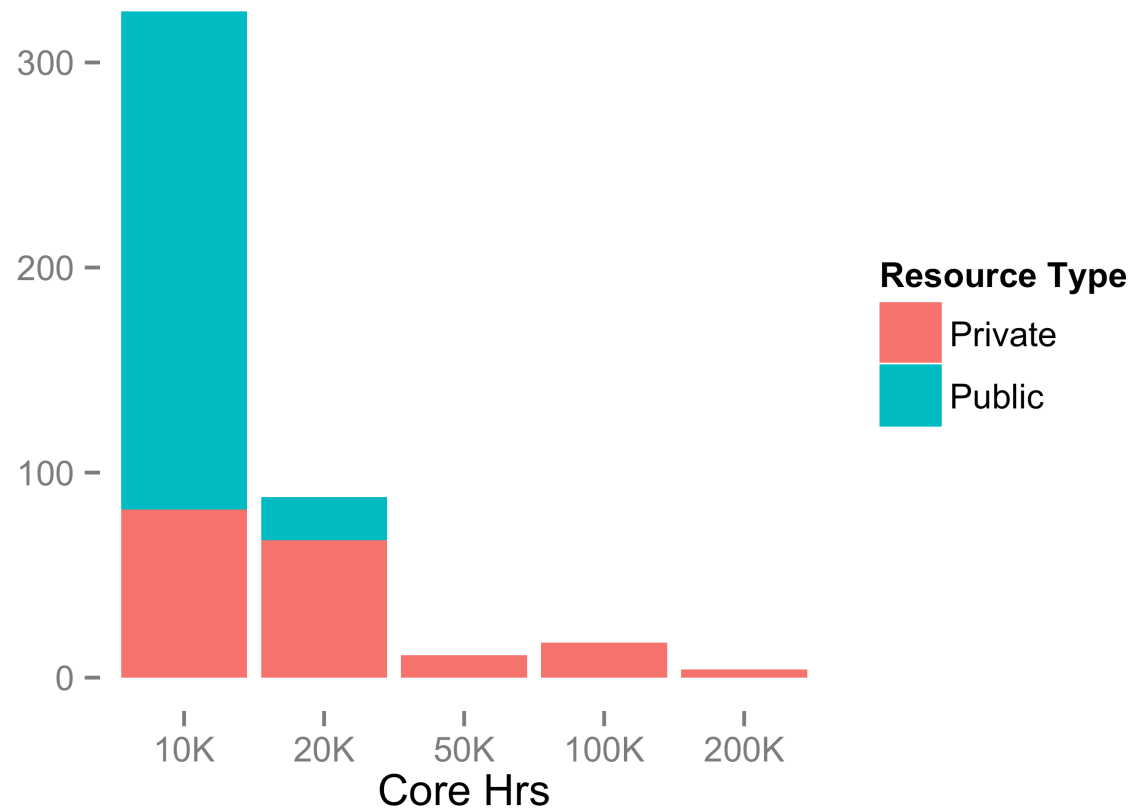


# NOAA NEXRAD level II radar data: Useful for severe weather



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.

Number of times a user exceeded core hrs  
in a given month in 2015 by resource type



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





CWI

Centrum Wiskunde & Informatica

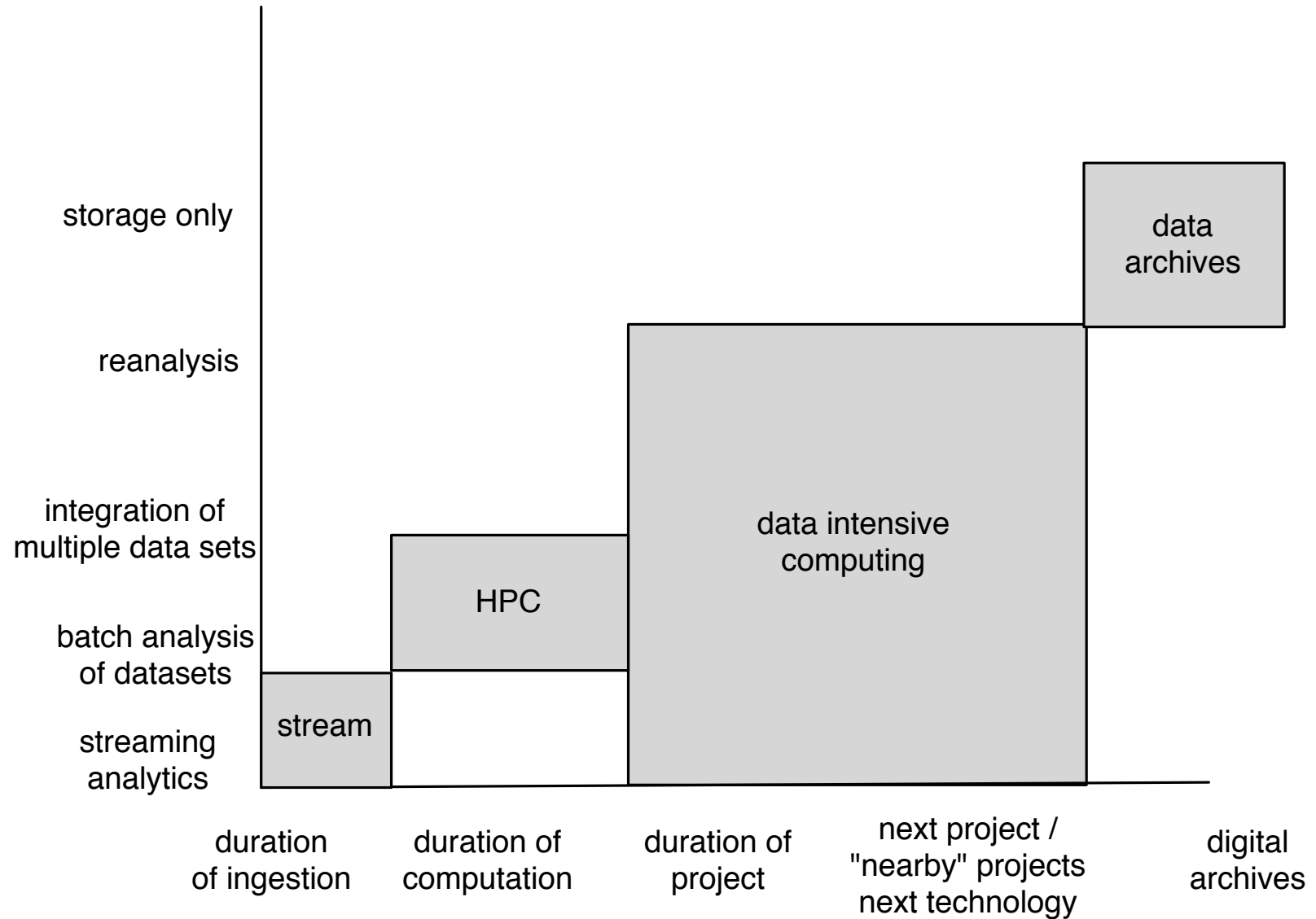
The Open Science Data Cloud PIRE (2010-2015) educated over 70 students with long term placements with our PIRE Partners in Edinburgh, Amsterdam, Tokyo, and Sao Paulo. We also taught over 500 others in one day tutorials.

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.

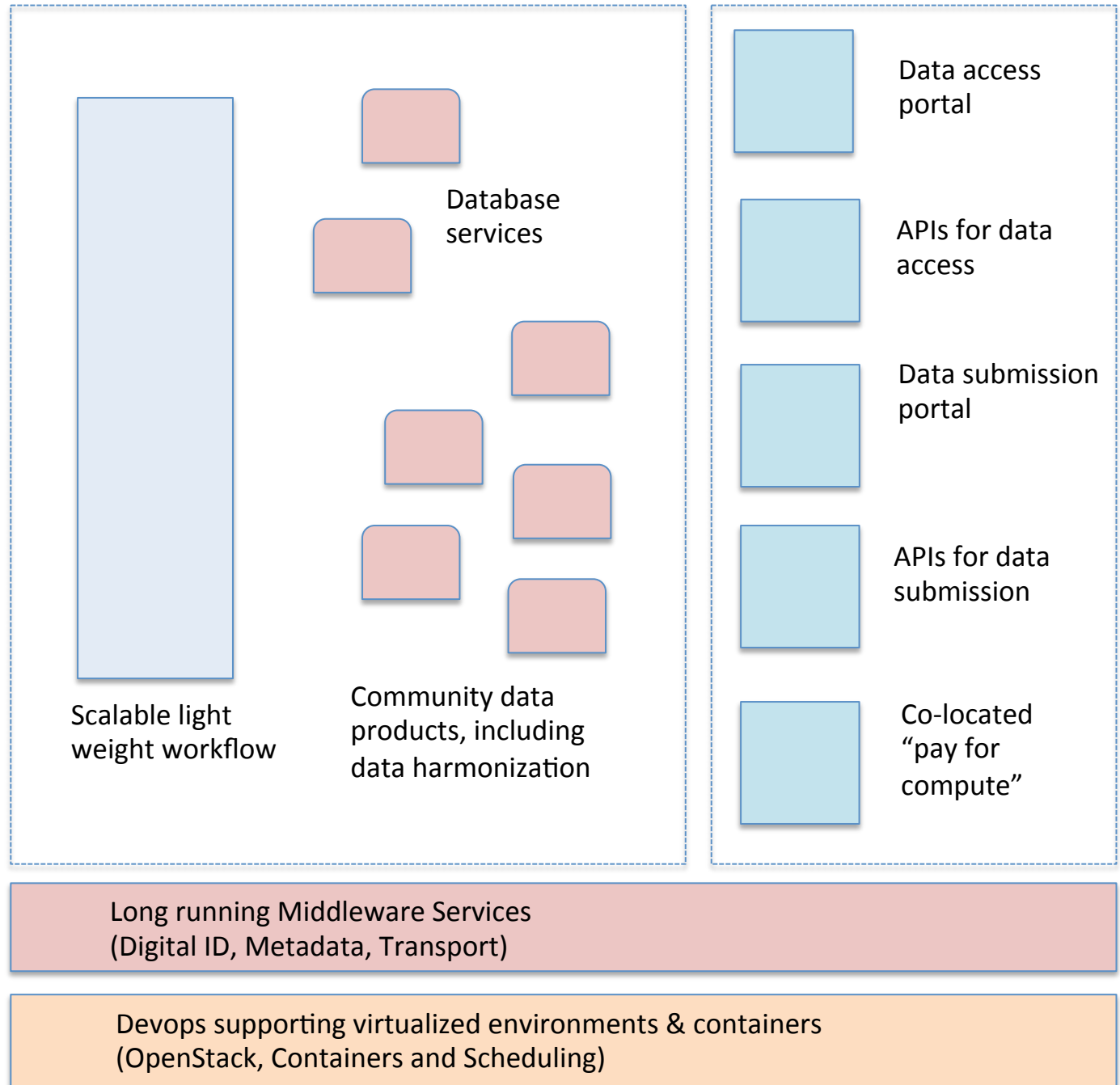
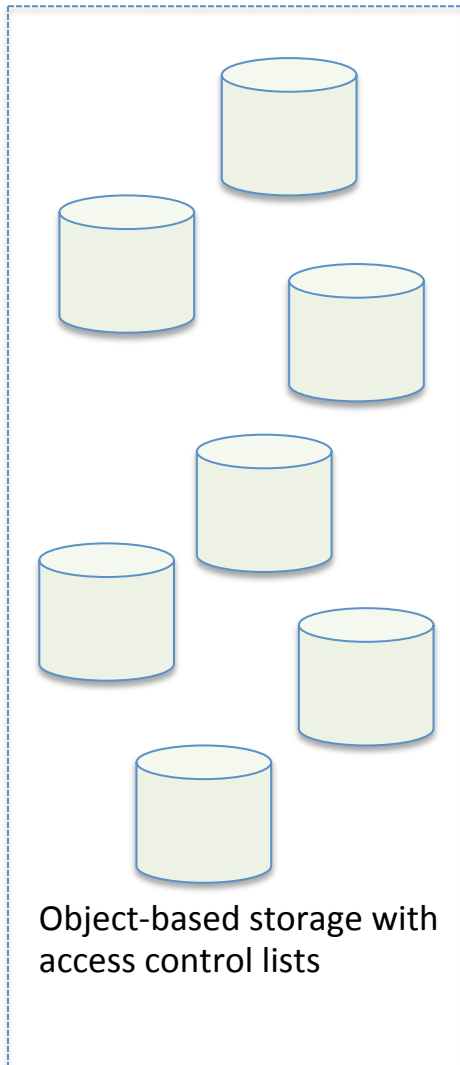
## Type of analysis



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.



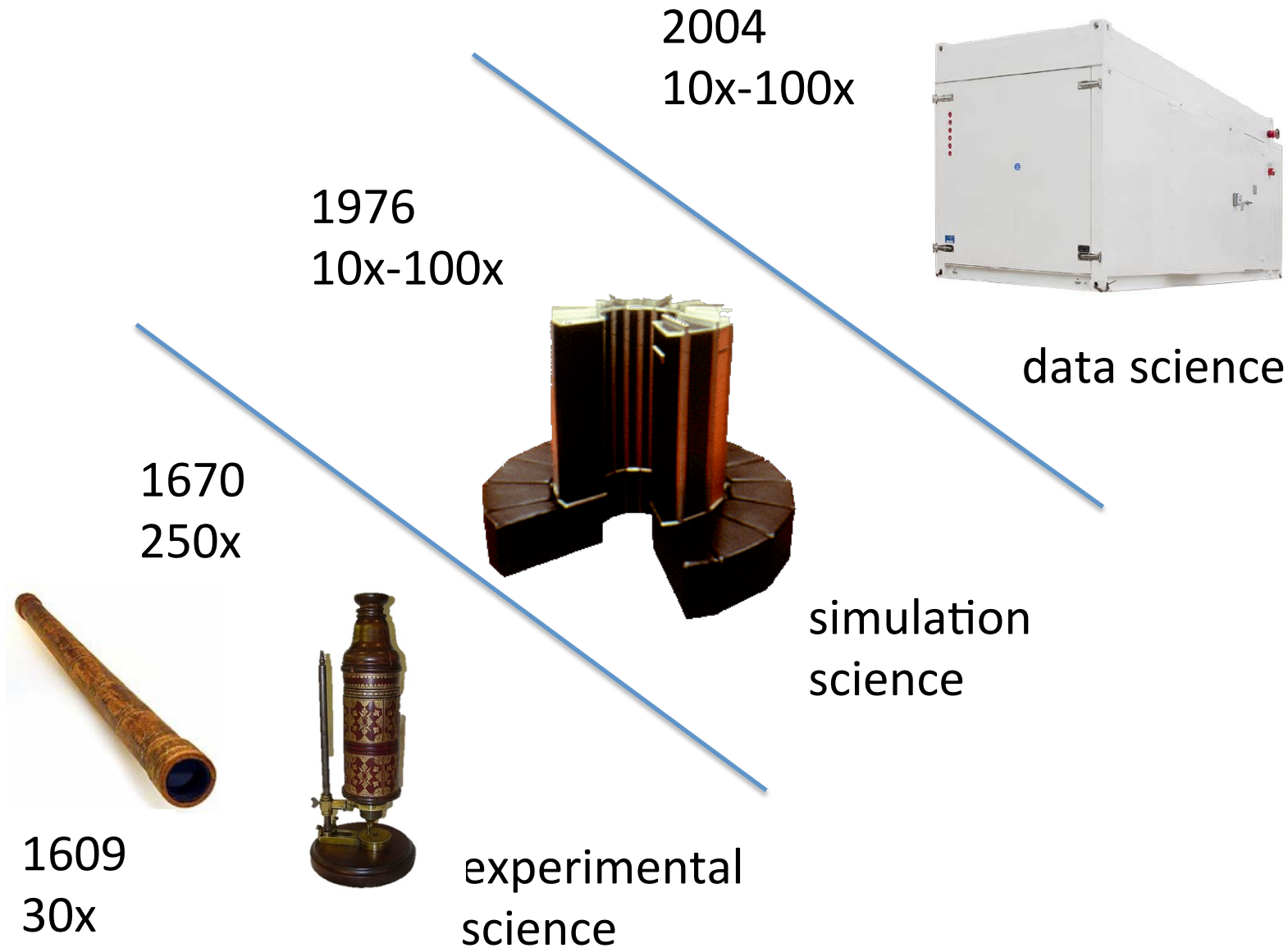
# The Open Source Technology Stack

- 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?



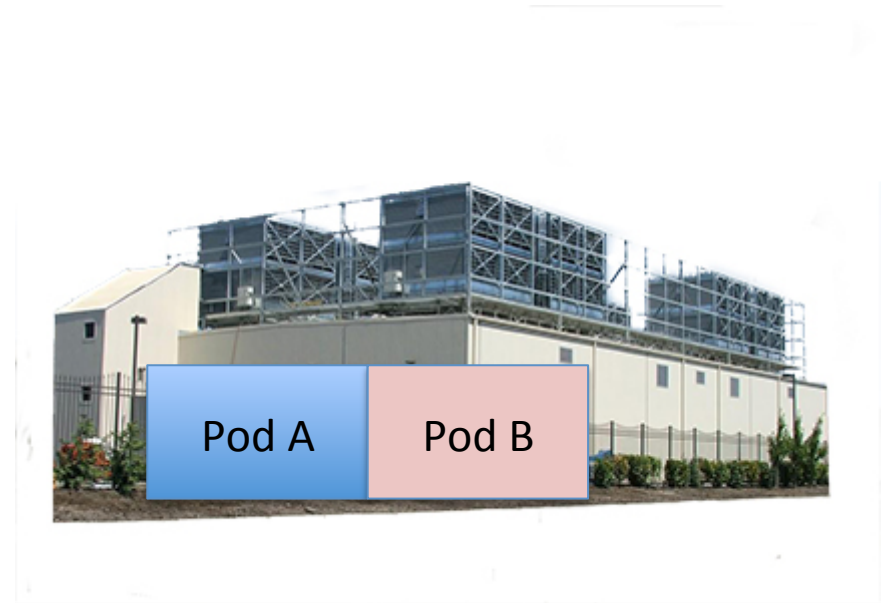
# Lesson 1: Think of a Commons as a New Instrument

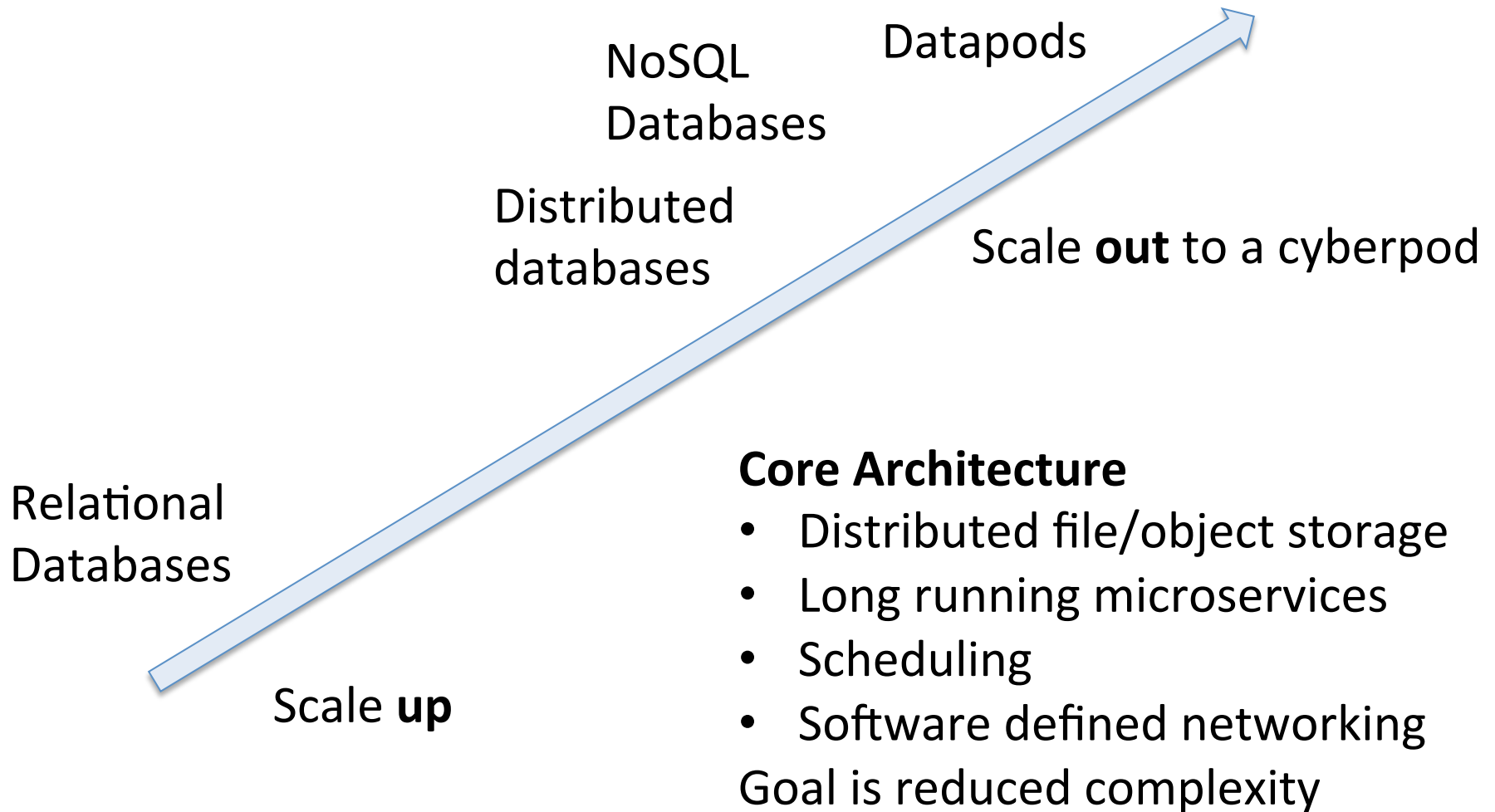




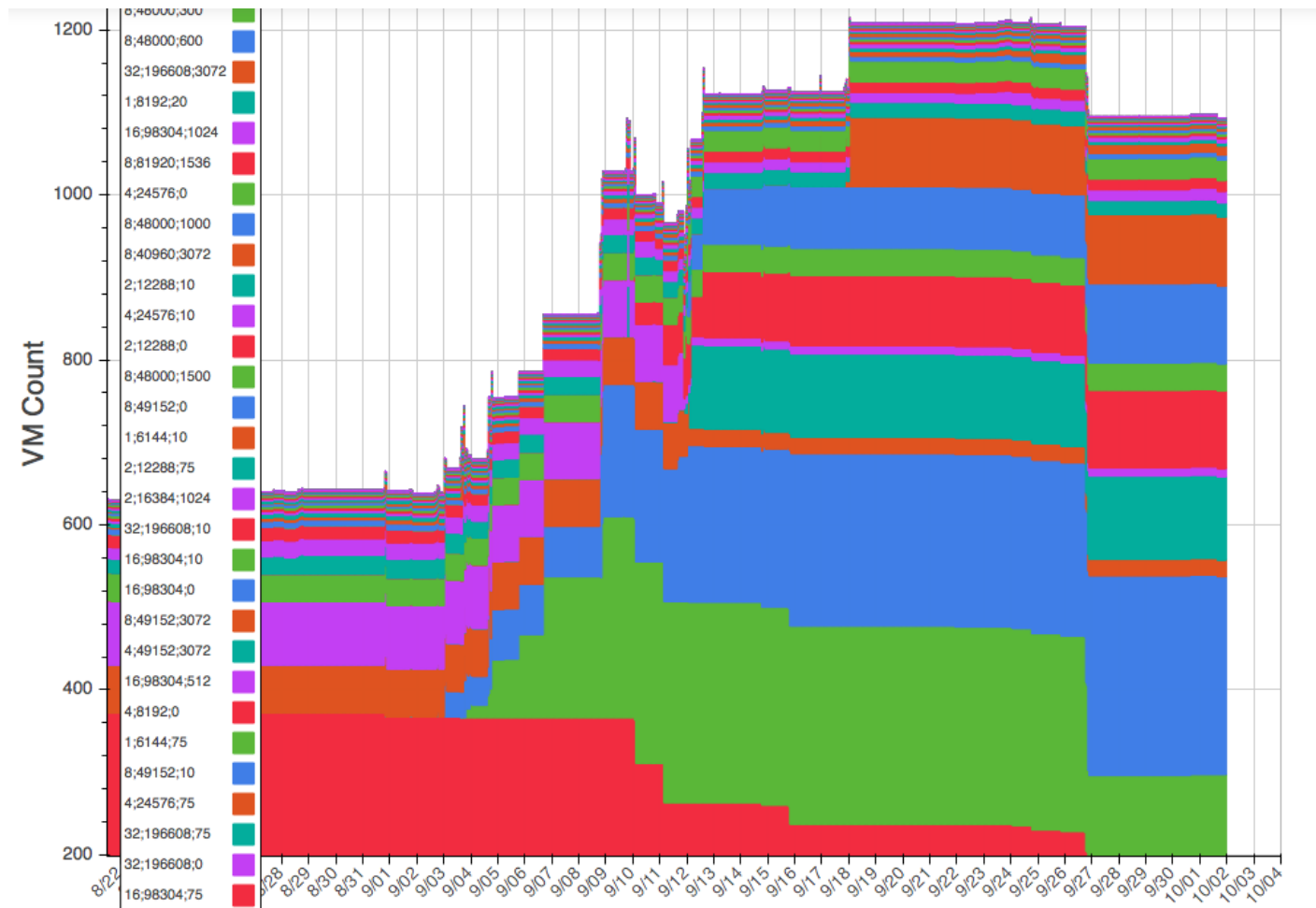
# What Scale?

- 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.

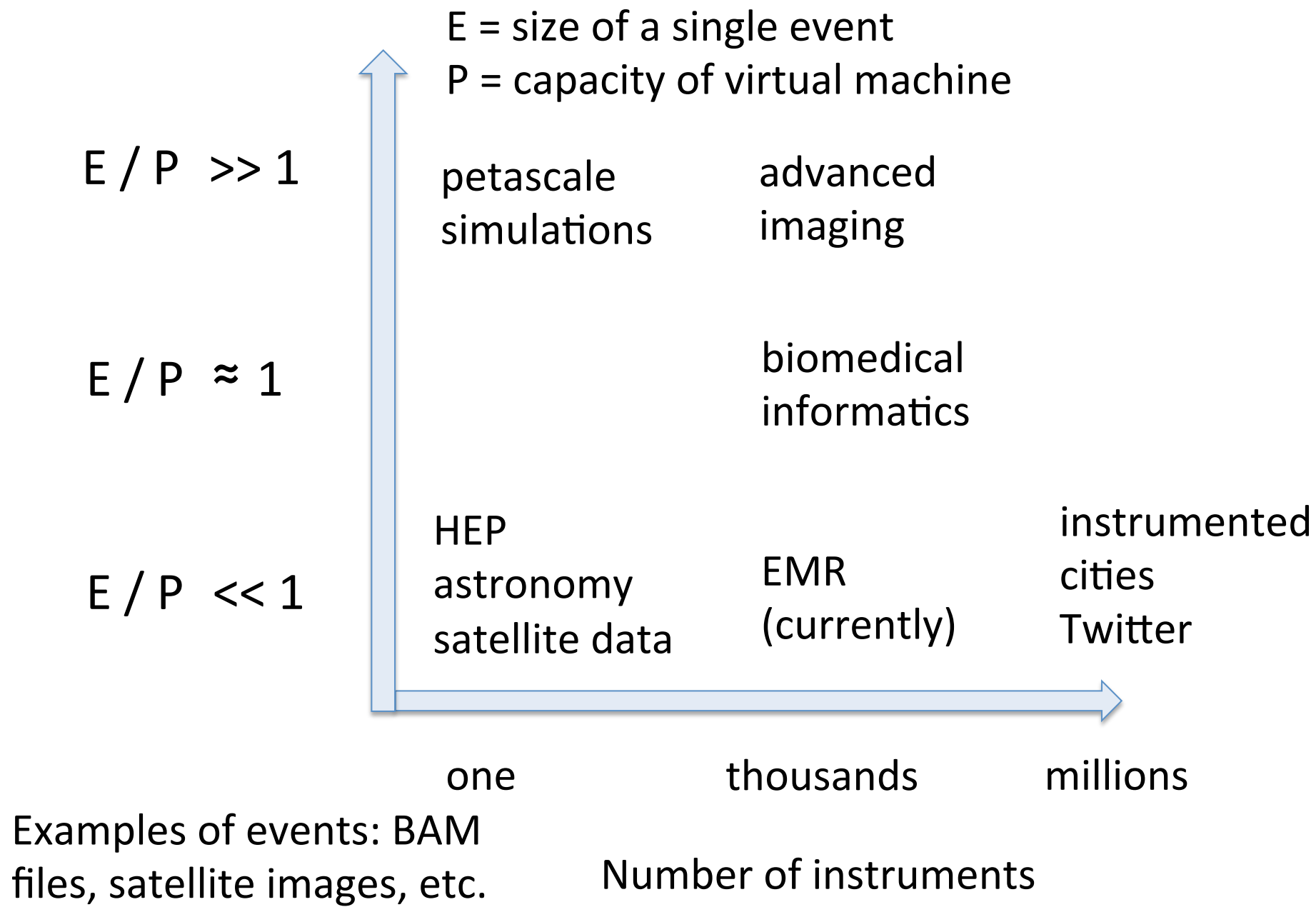




# Lesson 2: We Need DevOps for Analytics That Supports Re-analysis, Scans & Mashups



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



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

Suitability of the  
software

high

medium

low

web economy  
companies

biomedical  
informatics

HEP  
astronomy  
satellite data

IoT

2-3

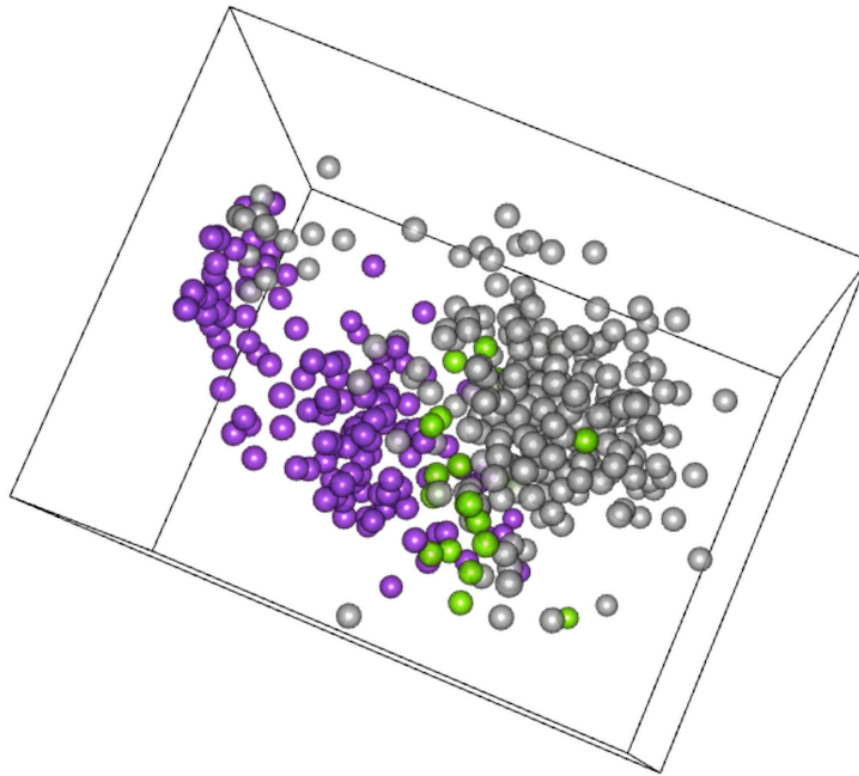
5-7

10-12

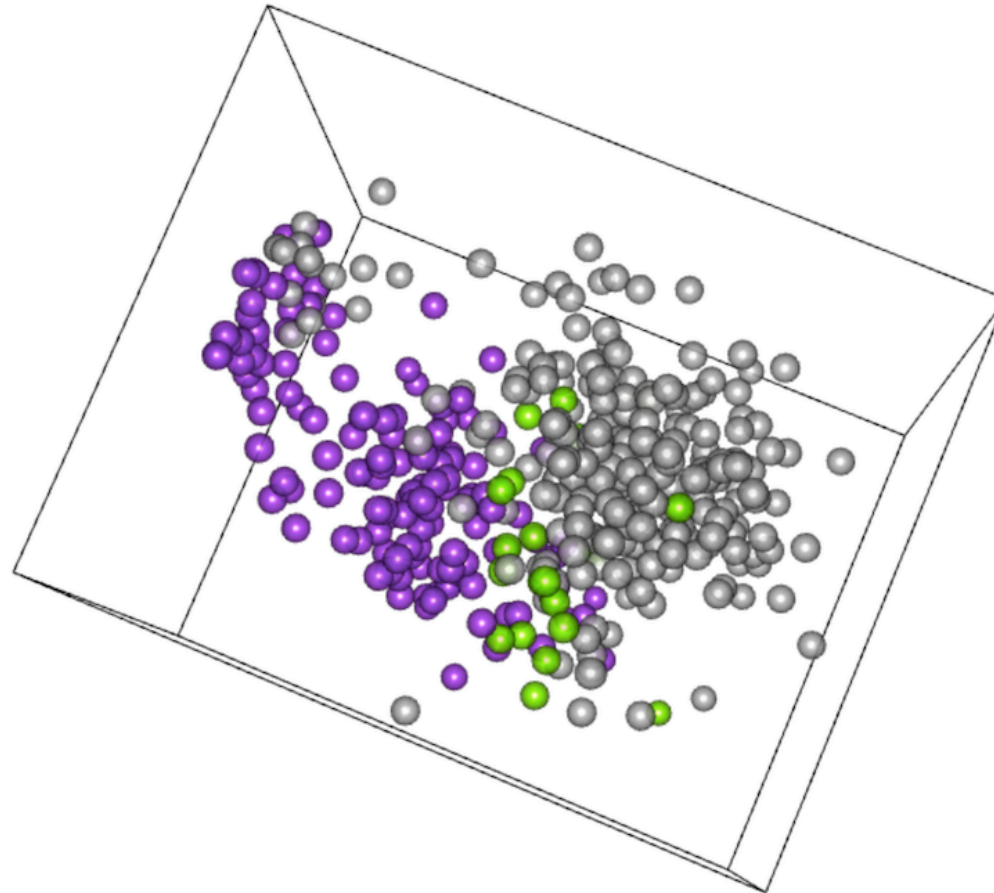
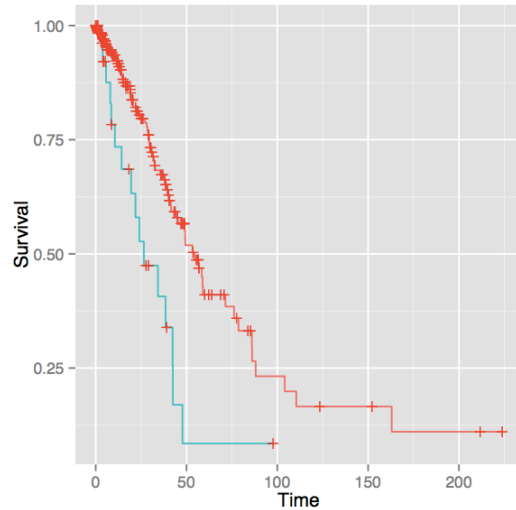
Years in each generation of an instrument

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.

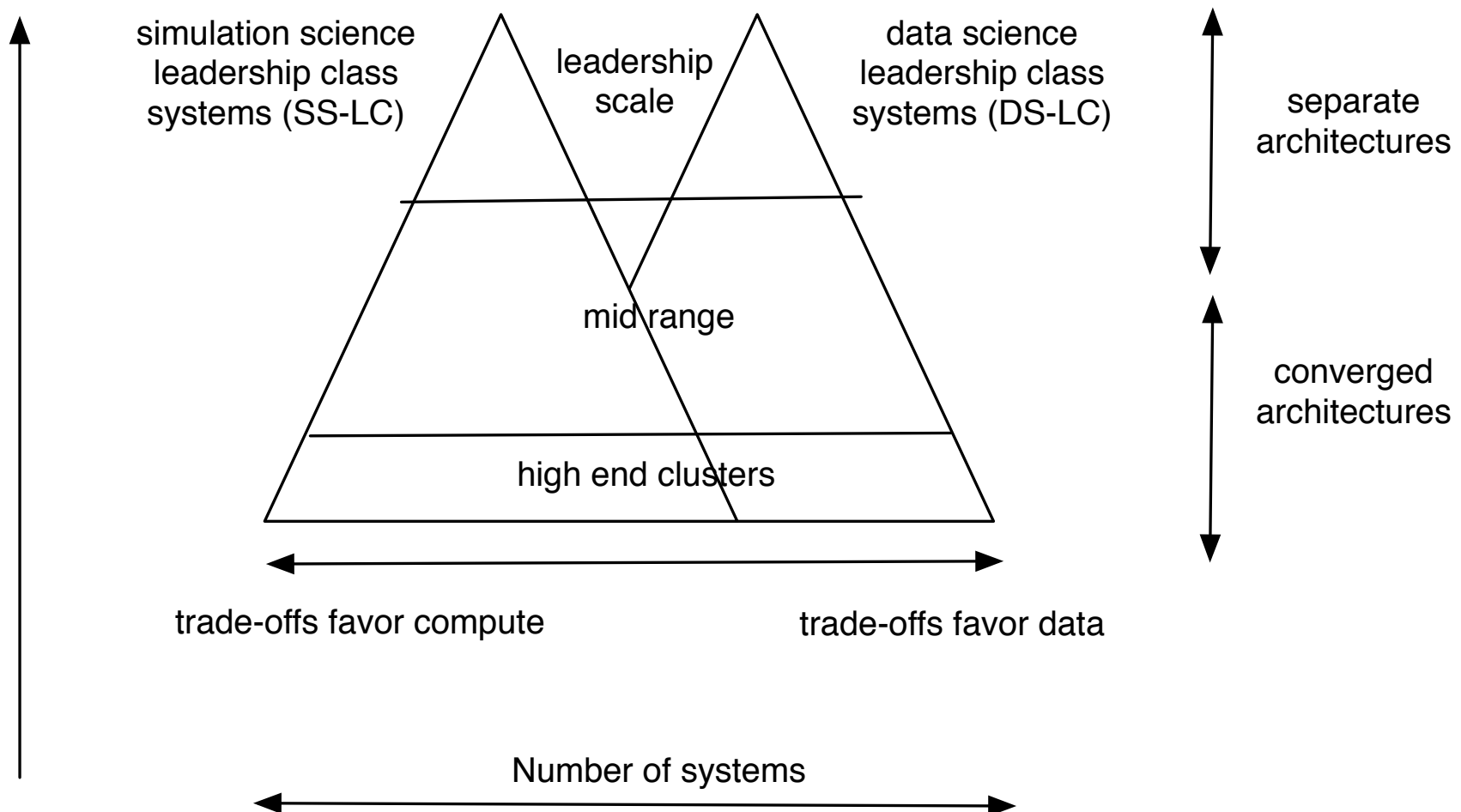


Purple balls are lung adenocarcinoma. Grey are lung squamous cell carcinoma. Green are misdiagnosed.

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

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

Performance





## 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.

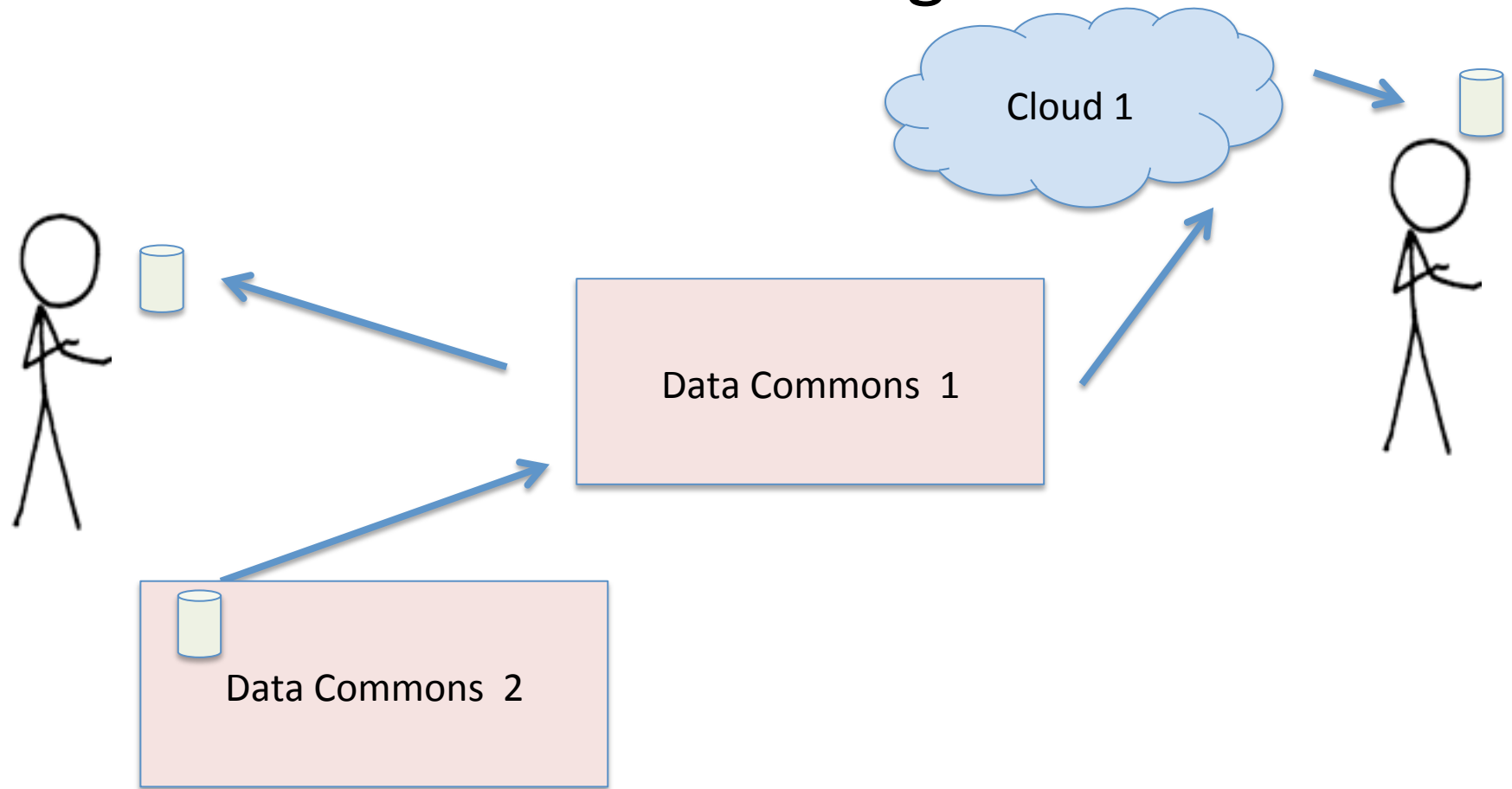
## Requirement 2: Scale to Cyber Pods and Datapods

- 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

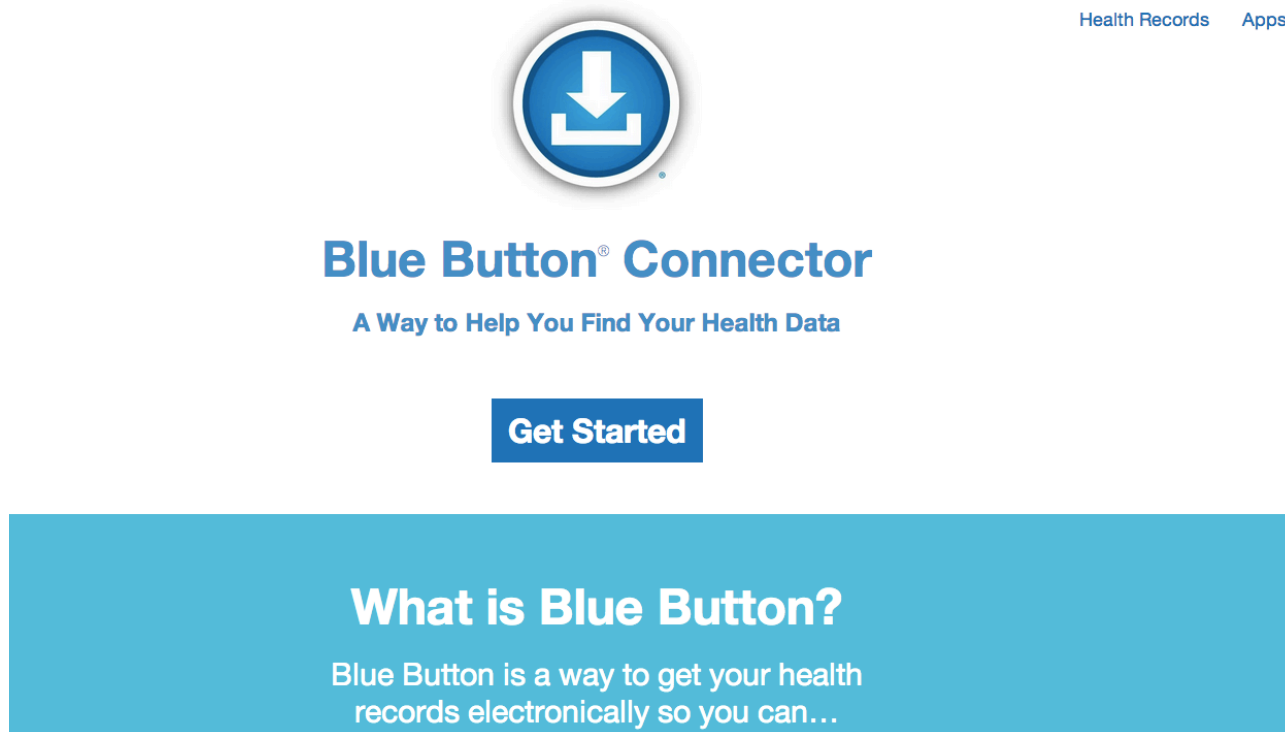
- How can a critical mass of data commons support data peering so that a research at one of the commons can transparently access data managed by one of the other commons
  - We need to access data independent of where it is stored
  - “Tier 1 data commons” need to pass community managed data at no cost
  - 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.

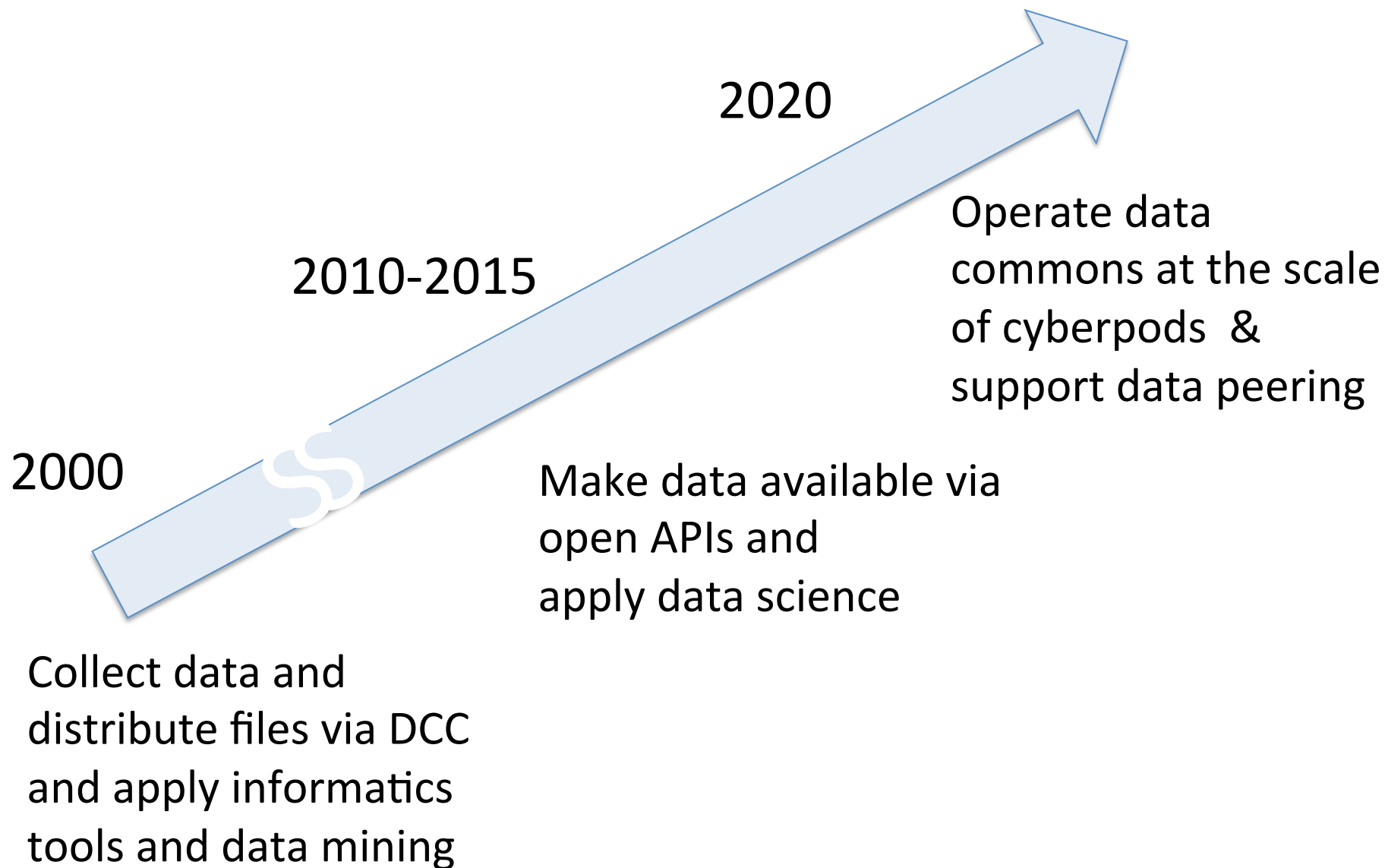
# Requirement 6: Application Program Interfaces

- 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



# Questions?



rgrossman.com  
@bobgrossman

# For More Information

Robert L. Grossman

[rgrossman.com](http://rgrossman.com)

@BobGrossman

[robert.grossman@uchicago.edu](mailto:robert.grossman@uchicago.edu)



Division of the Biological Sciences  
Center for Data Intensive Science

[cdis.uchicago.edu](http://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](mailto: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.