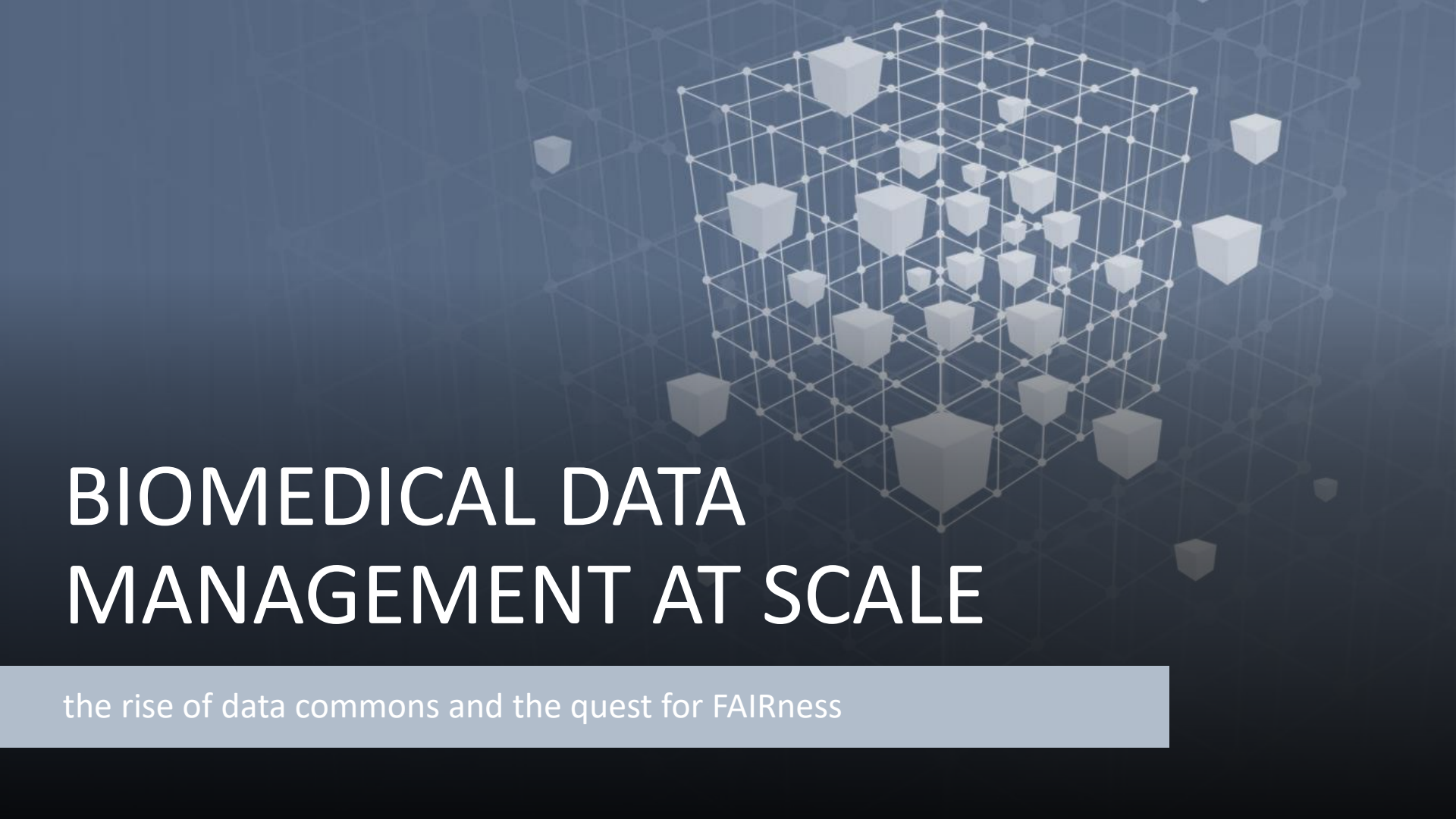


BIOMEDICAL DATA MANAGEMENT AT SCALE

the rise of data commons and the quest for FAIRness

AGENDA

1. Data sharing is difficult
2. Biomedical big data and data commons
3. Metadata and FAIRness
4. The urgency of COVID-19

SHARING BIOMEDICAL DATA IS DIFFICULT

Sharing data is difficult

- Data are rarely collected for reuse
- The effort to make data reusable falls the data producer; the benefit accrues to the data consumer
- Incentives are not aligned to support data reuse

Challenges with biomedical data

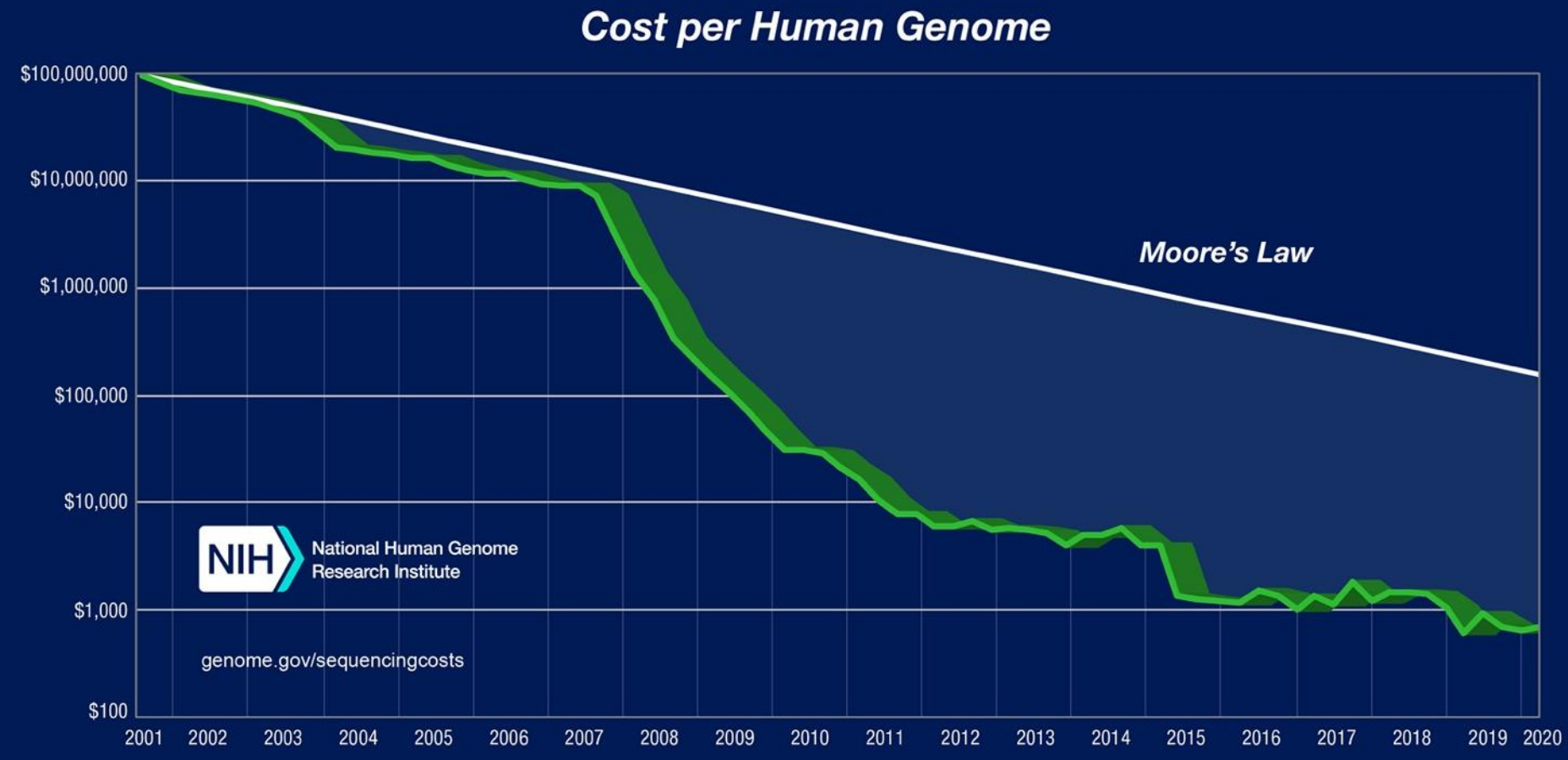
Privacy and regulatory concerns over data from human subjects

Clinical data generally collected for treatment, not reuse

Biologists tend not to converge on standards

Biomedical Big Data

Image source: Piqsels.com. License: CC0 public domain.



https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

NATIONAL CANCER INSTITUTE
THE CANCER GENOME ATLAS

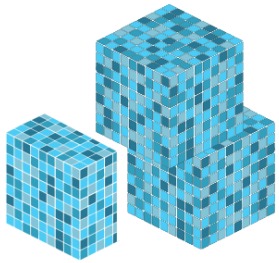
TCGA BY THE NUMBERS

TCGA produced over

2.5

PETABYTES

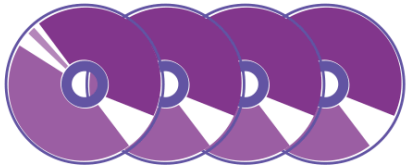
of data



To put this into perspective, **1 petabyte** of data is equal to

212,000

DVDs



TCGA data describes

33

DIFFERENT TUMOR TYPES

...including

10

RARE CANCERS

...based on paired tumor and normal tissue sets collected from

11,000

PATIENTS

...using

7

DIFFERENT DATA TYPES

2006-2015

Sustaining the big-data ecosystem

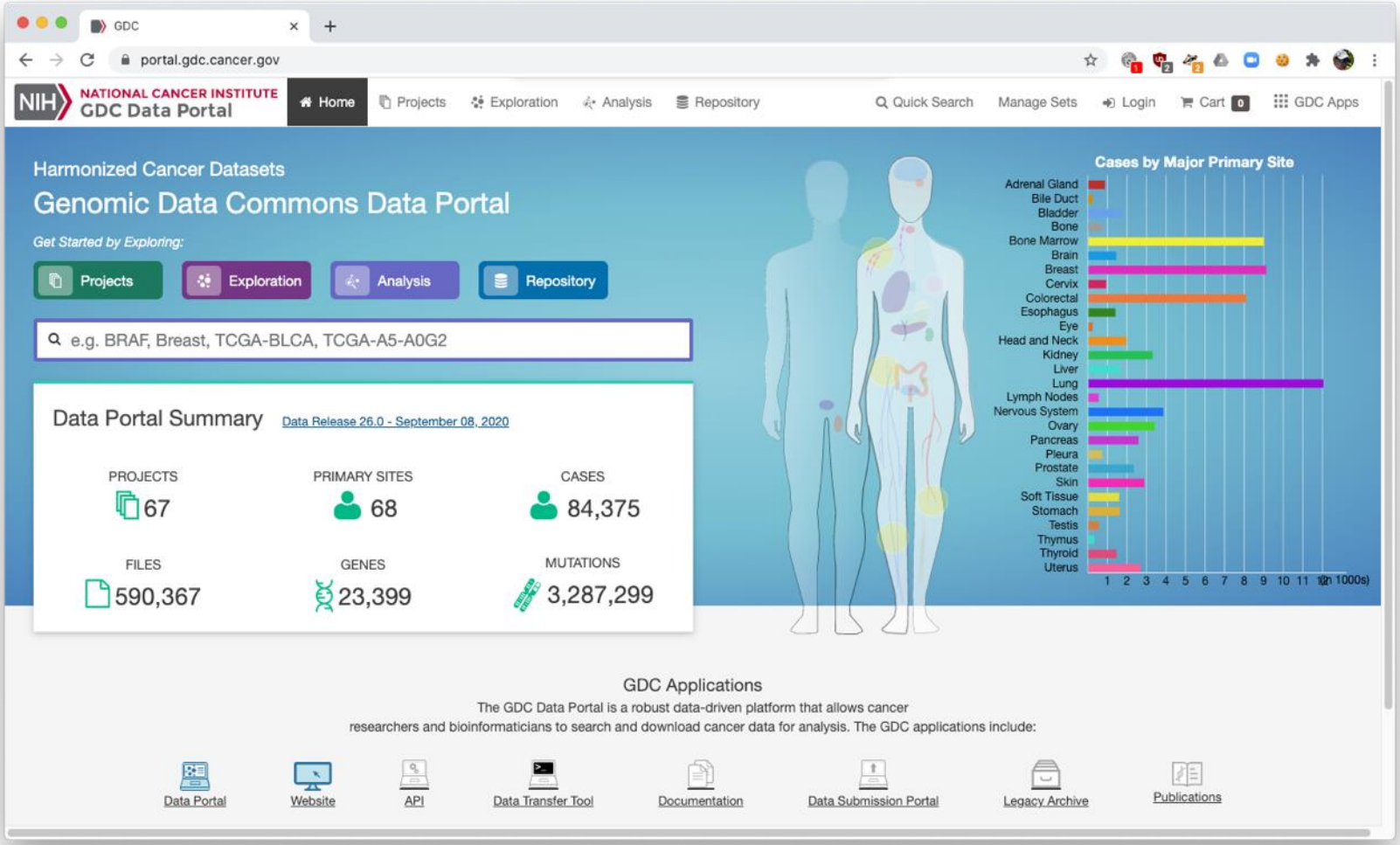
Organizing and accessing biomedical big data will require quite different business models, say **Philip E. Bourne**, **Jon R. Lorsch** *and* **Eric D. Green**.



THE RESEARCH COMMUNITY MUST FIND MORE **EFFICIENT** MODELS FOR **STORING, ORGANIZING AND ACCESSING** BIOMEDICAL DATA.

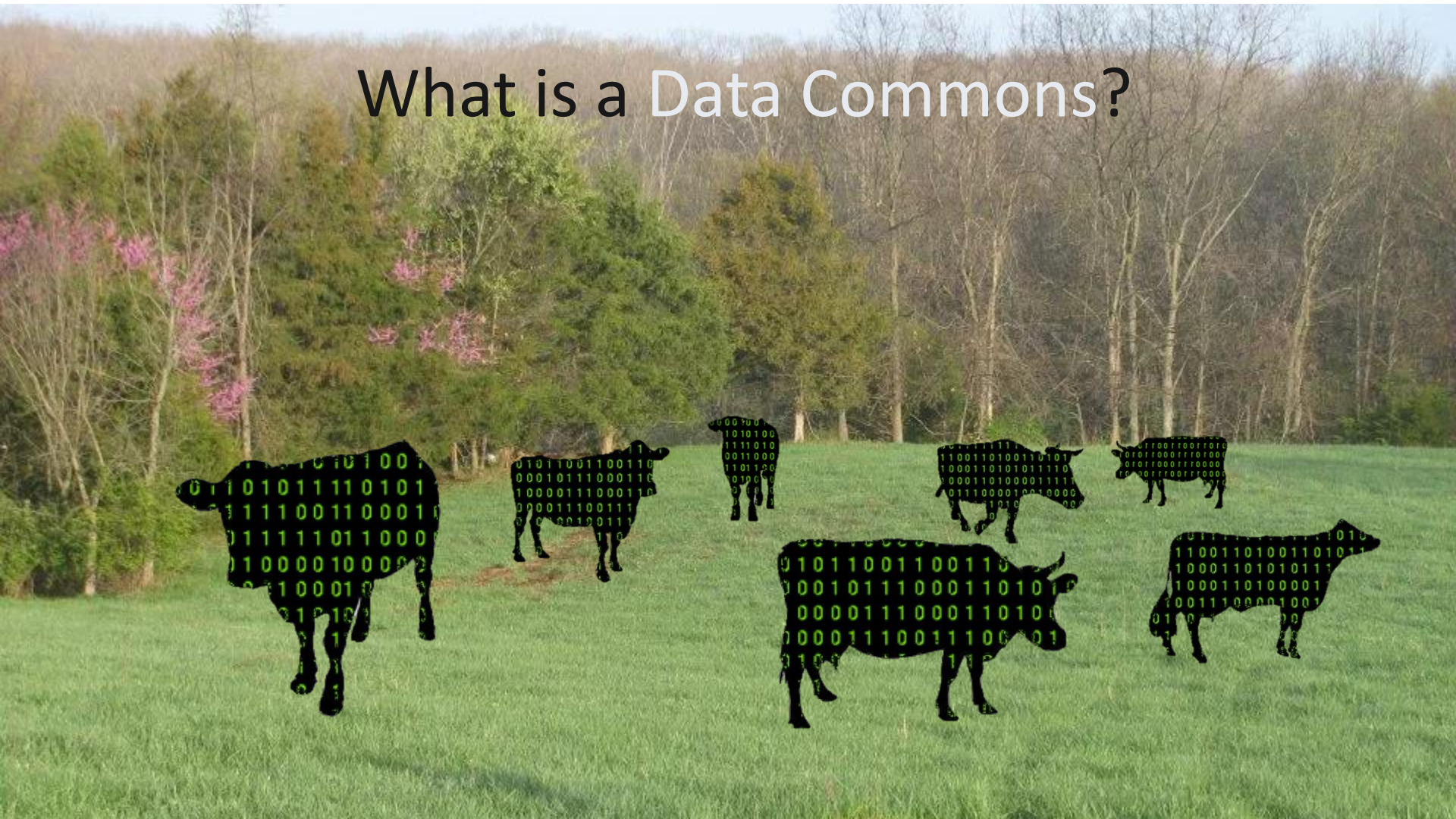
2015

Bourne, P., Lorsch, J. & Green, E. Perspective: Sustaining the big-data ecosystem. *Nature* **527,** S16–S17 (2015). https://doi.org/10.1038/527S16a


GDC
portal.gdc.cancer.gov
NIH NATIONAL CANCER INSTITUTE GDC Data Portal
Home
Projects
Exploration
Analysis
Repository
Quick Search
Manage Sets
Login
Cart 0
GDC Apps
Harmonized Cancer Datasets
Genomic Data Commons Data Portal
Get Started by Exploring:
Projects
Exploration
Analysis
Repository
e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-A0G2
Data Portal Summary
Data Release 26.0 - September 08, 2020
PROJECTS 67
PRIMARY SITES 68
CASES 84,375
FILES 590,367
GENES 23,399
MUTATIONS 3,287,299
Cases by Major Primary Site
Adrenal Gland
Bile Duct
Bladder
Bone
Bone Marrow
Brain
Breast
Cervix
Colorectal
Esophagus
Eye
Head and Neck
Kidney
Liver
Lung
Lymph Nodes
Nervous System
Ovary
Pancreas
Pleura
Prostate
Skin
Soft Tissue
Stomach
Testis
Thymus
Thyroid
Uterus
1 2 3 4 5 6 7 8 9 10 11 12 (in 1000s)
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:
Data Portal
Website
API
Data Transfer Tool
Documentation
Data Submission Portal
Legacy Archive
Publications

2016

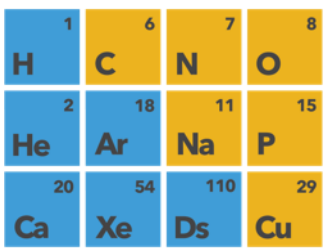
What is a Data Commons?



A Functional Definition

"Data commons are cloud-based software platforms that co-locate: 1) data, 2) computing infrastructure, and 3) commonly used software applications, tools and services to create a resource for managing, analyzing, integrating and sharing data with a community."

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. PMID: 30691868 PMCID: PMC6474403



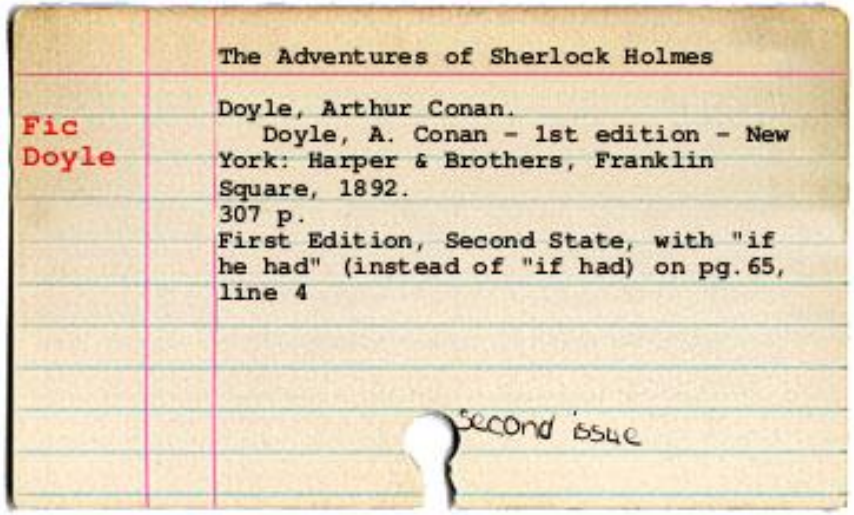
DCPPC

DATA COMMONS PILOT PHASE CONSORTIUM

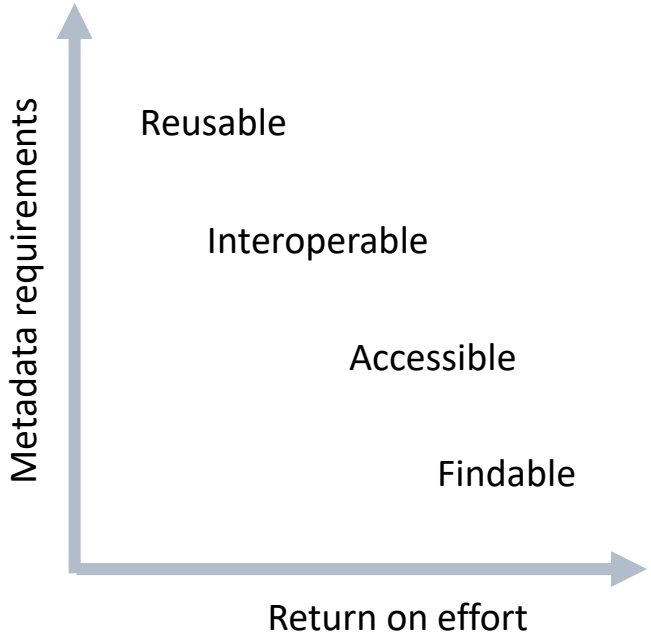
”Our guiding principles for data access, use, and reuse will adhere to the Findable, Accessible, Interoperable, and Reusable (FAIR) guidelines. While the FAIR guidelines are well-accepted, to bring them to practice will require defining and adopting community-based metrics and rubrics so these can be applied to data, and other types of digital objects, hosted within or available through the Data Commons. At the same time, once FAIR metrics and rubrics are defined, these will be used to measure the level of “FAIRness” of repositories, datasets, and other digital objects. Such evaluations will inform and engage both Data Commons users and digital objects producers.”

2018

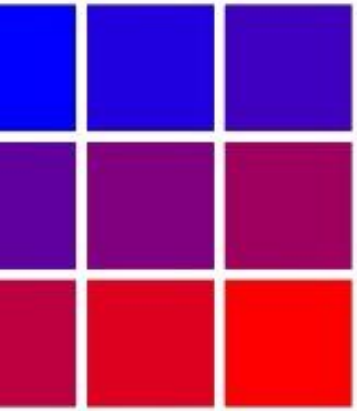
#metadata and FAIRness



Metadata & FAIRness

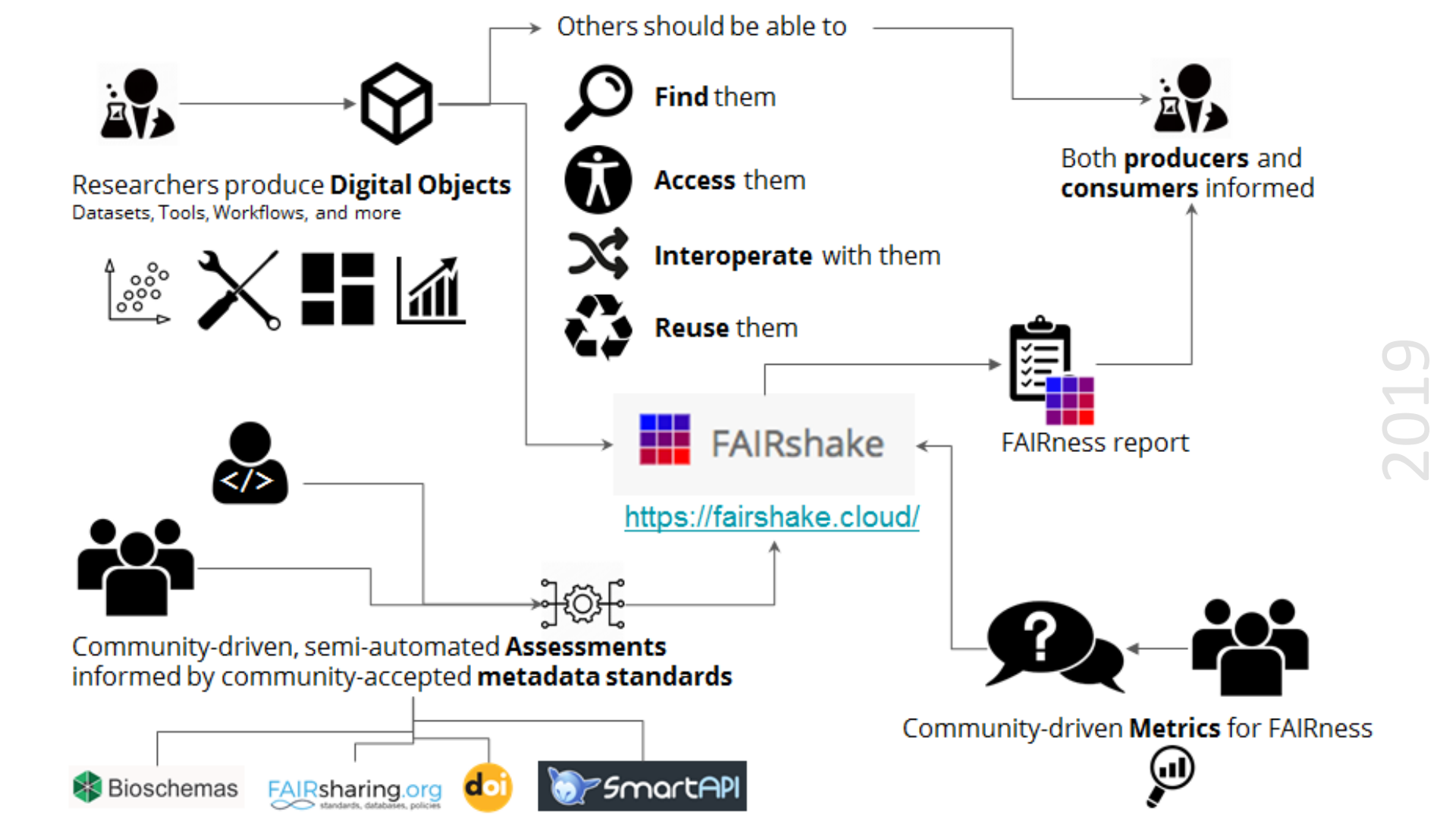


Quantifying FAIRness



FAIRshake

2019


Others should be able to
Find them
Access them
Interoperate with them
Reuse them
Researchers produce Digital Objects
Datasets, Tools, Workflows, and more
Both producers and consumers informed
FAIRness report
FAIRshake
https://fairshake.cloud/
Community-driven, semi-automated Assessments informed by community-accepted metadata standards
Community-driven Metrics for FAIRness
Bioschemas
FAIRsharing.org
standards, databases, policies
doi
SmartAPI
2019

Cell Systems
Volume 9, Issue 5, 27 November 2019, Pages 417-421

Commentary

FAIRshake: Toolkit to Evaluate the FAIRness of Research Digital Resources

Daniel J.B. Clarke [1], Lily Wang [1], Alex Jones [2], Megan L. Wojciechowicz [1], Denis Torre [1], Kathleen M. Jagodnik [1], Sherry L. Jenkins [1], Peter McQuilton [3], Zachary Flamholz [1], Moshe C. Silverstein [1], Brian M. Schilder [1], Kimberly Robasky [4], Claris Castillo [4], Ray Idaszak [4], Stanley C. Ahalt [4], Jason Williams [5], Stephan Schurer [6], Daniel J. Cooper [6] ... Avi Ma'ayan [1]

Show more

https://doi.org/10.1016/j.cels.2019.09.011

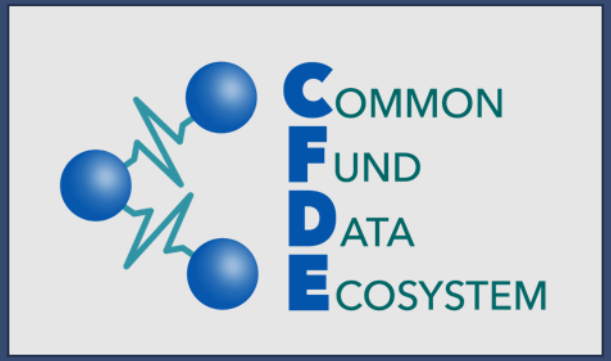
Get rights and content

https://www.alliancegenome.org/

Score: 78%
Information is provided describing how to cite the dataset.

View Associations | View Assessments

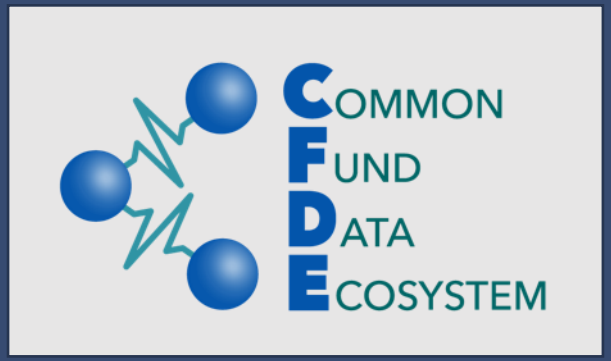
FAIRshake



The goal of Common Fund Data Ecosystem CFDE is to federate data from a number of Common Fund Data Coordinating Centers (DCCs) to improve access to data derived from hundreds of studies and samples collected from thousands of human subjects. This project involves a diversity of datatypes has been generated at the genomic, expression, proteomic, metagenomic, and imaging levels, and the DCCs support a tremendous range of scientific discovery efforts.

2020

https://cfde-published-documentation.readthedocs-hosted.com/en/latest/



Crosscut Metadata Model "C2M2"

C2M2 Level 0 defines a **minimal valid C2M2 instance.** Data submissions at this level of metadata richness will be the easiest to produce, and will support the simplest available functionality implemented by downstream applications.

C2M2 Level 1 models **basic experimental resources and associations between them**. This level of metadata richness is more difficult to produce than Level 0's flat inventory of digital file assets. As a result, Level 1 metadata offers users more powerful downstream tools than are available for Level 0.

2020

https://cfde-published-documentation.readthedocs-hosted.com/en/latest/

The Urgency of COVID-19

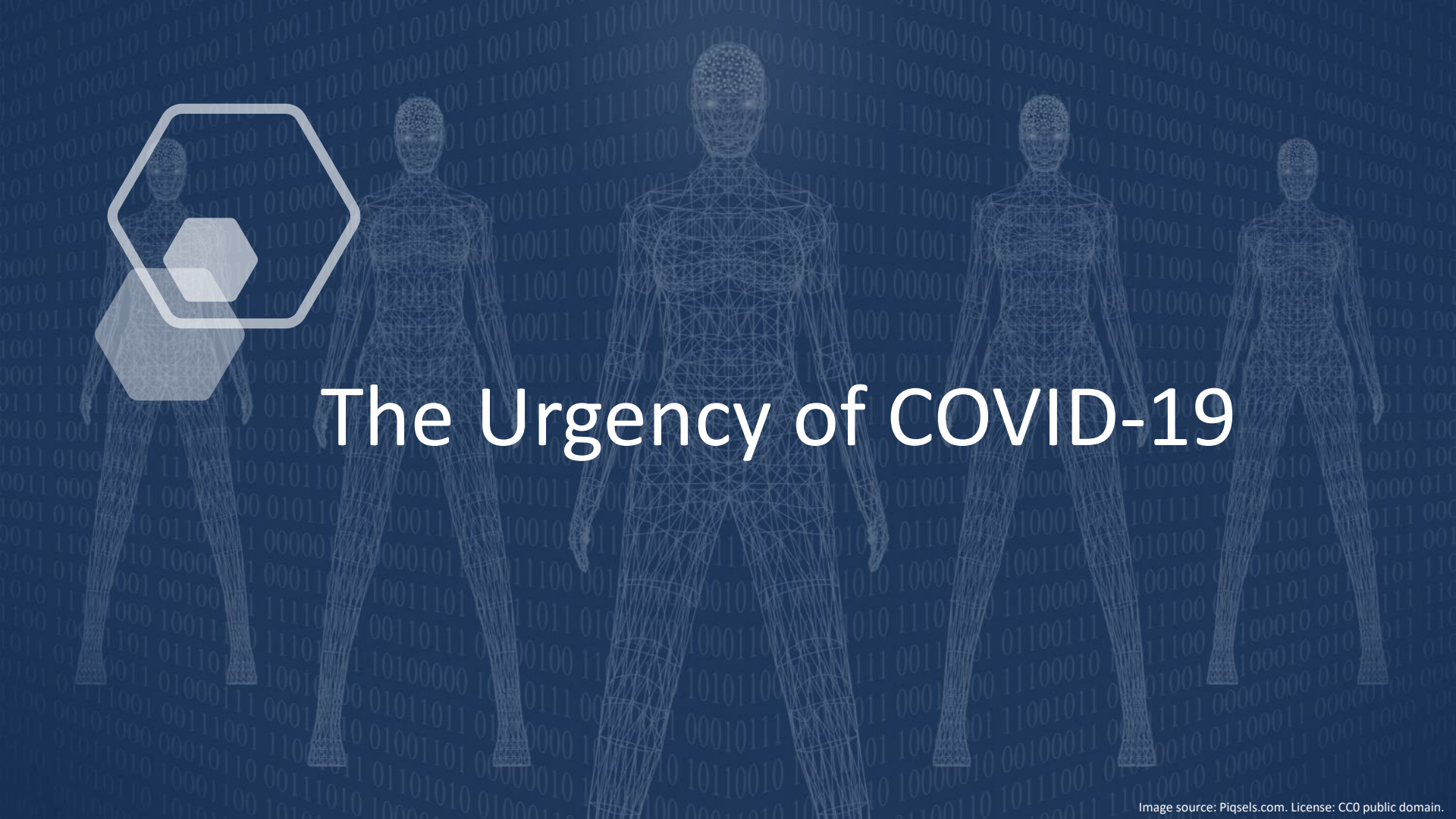
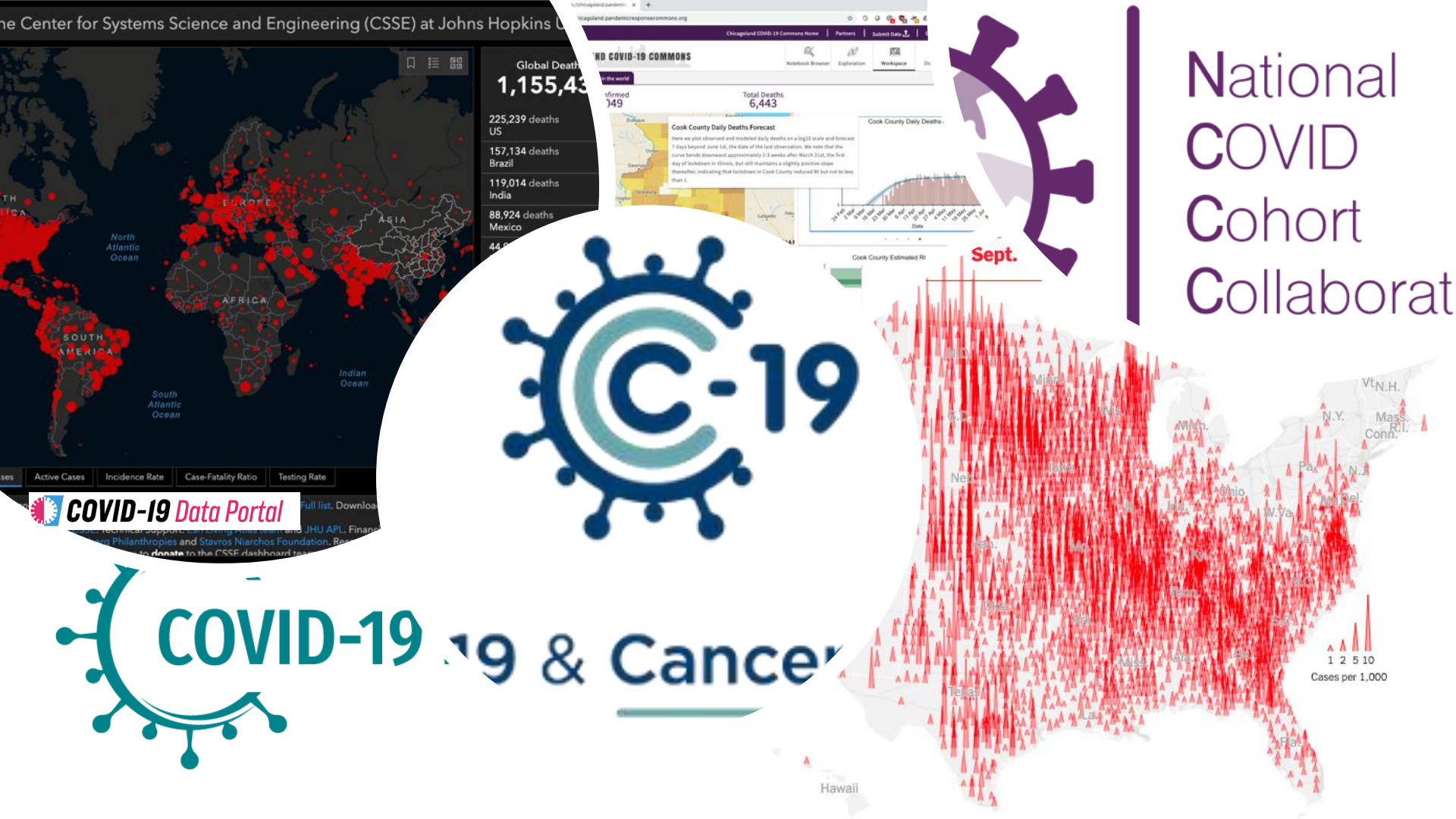
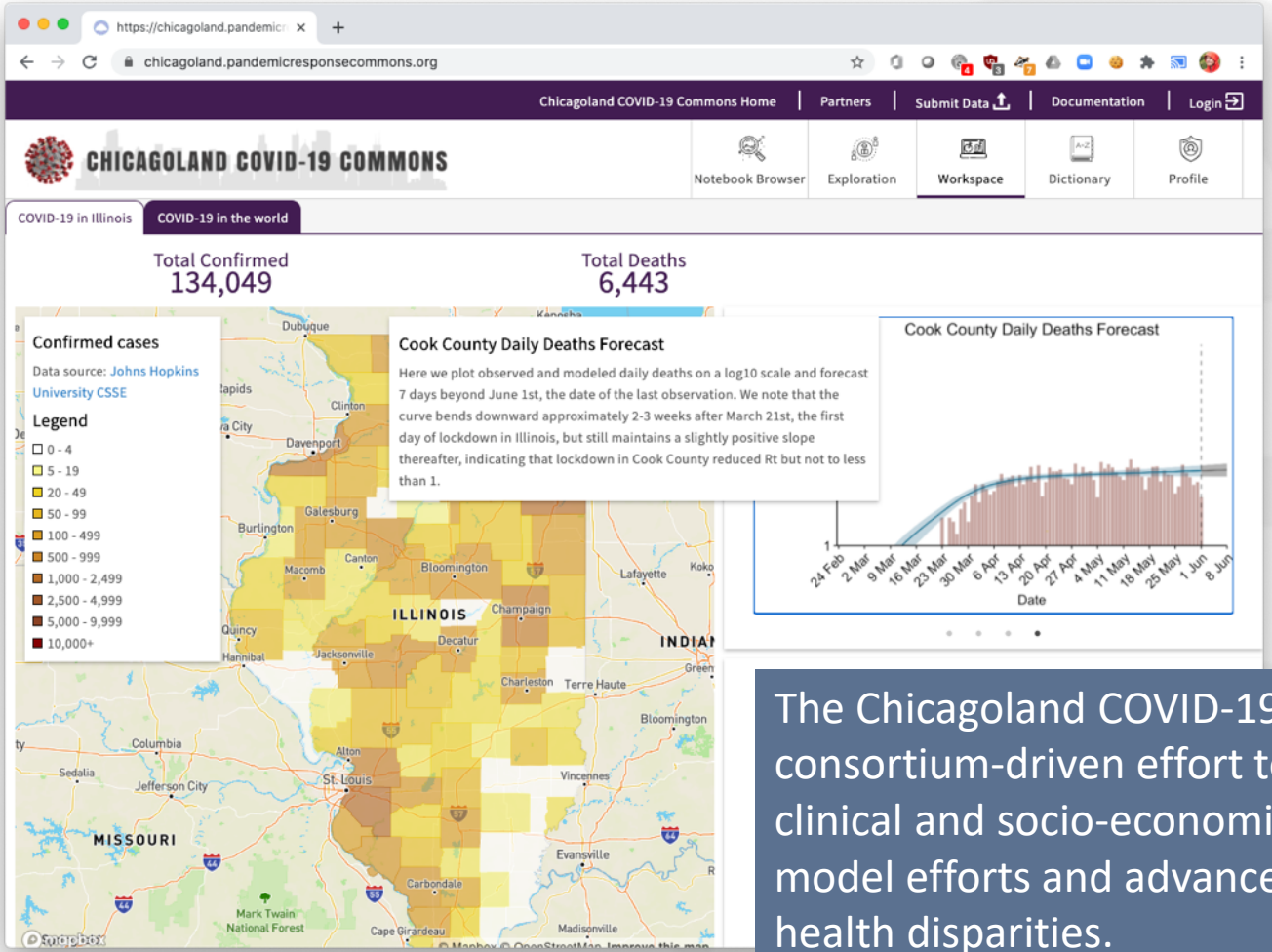


Image source: Piqsels.com. License: CC0 public domain.


ne Center for Systems Science and Engineering (CSSE) at Johns Hopkins U
Global Death
1,155,43
225,239 deaths
US
157,134 deaths
Brazil
119,014 deaths
India
88,924 deaths
Mexico
EUROPE
ASIA
AFRICA
North Atlantic Ocean
SOUTH AMERICA
Indian Ocean
South Atlantic Ocean
Active Cases
Incidence Rate
Case-Fatality Ratio
Testing Rate
COVID-19 Data Portal
Philanthropies and Stavros Niarchos Foundation.
donate to the CSSE dashboard
ND COVID-19 COMMONS
Notebook Browser
Exploration
Workspace
Total Deaths
6,443
Cook County Daily Deaths Forecast
Cook County Daily Deaths
Cook County Estimated R
National
COVID
Cohort
Collaborat
C-19
COVID-19
19 & Cancer
Sept.
Vt.
N.H.
N.Y.
Mass.
R.I.
Conn.
Ohio
Texas
Hawaii
1 2 5 10
Cases per 1,000



The Chicagoland COVID-19 Commons is a consortium-driven effort to collect regional clinical and socio-economic data to drive local model efforts and advance the understanding of health disparities.

Final thoughts

- Data sharing (or enabling data access) is key to advancing biomedical research
- Data sharing is difficult; it requires
  - Active data management
  - Effective data governance
  - Adoption of standards, structured vocabulary
- Most importantly, it requires alignment of incentives

Thank you

https://chicagoland.pandemicresponsecommons.org/

Matthew Trunnell
matthew@occ-data.org
@MatthewTrunnell