The Case for Regional Data Commons to Support a Public Health Emergency

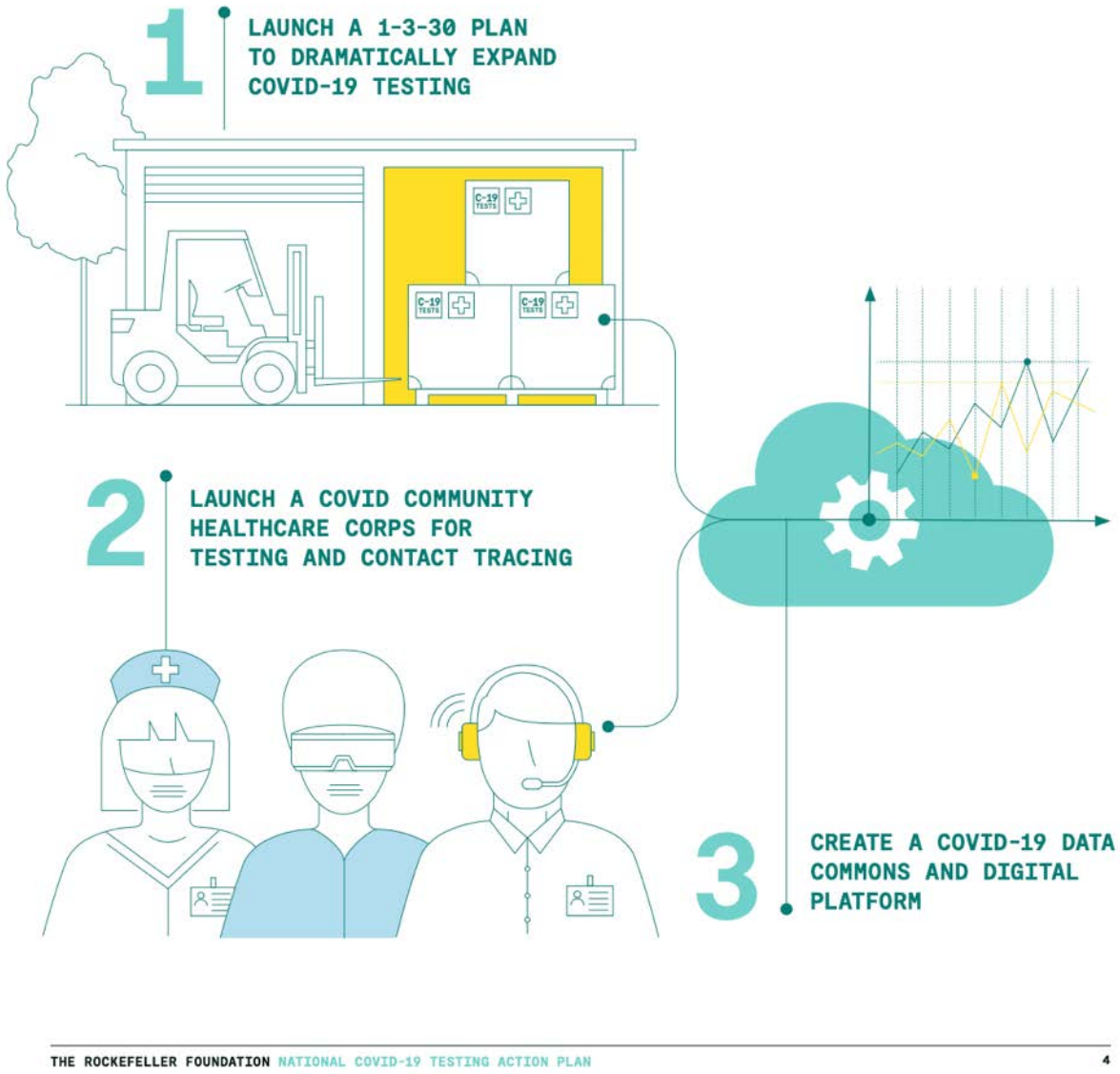
Robert L. Grossman

University of Chicago &

Open Commons Consortium

September 14, 2020

1. Data Access in a Public Health Emergency



Rockefeller Foundation National COVID-19 Testing & Tracing Action Plan

1. Grow testing to 30 million per week
2. Launch a COVID Community Healthcare Corps for testing and contact tracing
3. Create a COVID-19 Data Commons and Digital Platform

The COVID Tracking Project estimates that about 5.3 million tests are performed each week.

Sources: Rockefeller Foundation, Covid-19 National Testing & Tracing Action Plan Action, https://www.rockefellerfoundation.org/national-covid-19-testing-and-tracing-action-plan/, accessed September 1, 2020
The COVID Tracking Project, https://covidtracking.com/

Incidence and other count data

CDC

covid.cdc.gov/covid-data-tracker

The COVID Tracking Project

covidtracking.com

{T} Developers

Covid-19 Data

developer.nytimes.com/covid

Johns Hopkins University & Medicine | Coronavirus Resource Center

coronavirus.jhu.edu/data

Clinical and imaging data

N3C

covid.cd2h.org

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

biodatacatalyst.nhlbi.nih.gov

NIH National Institute of Biomedical Imaging and Bioengineering

Creating Biomedical Technologies to Improve Health

MIDRC

Viral sequence data

Nextstrain

nextstrain.org/sars-cov-2

GISAID

www.gisaid.org

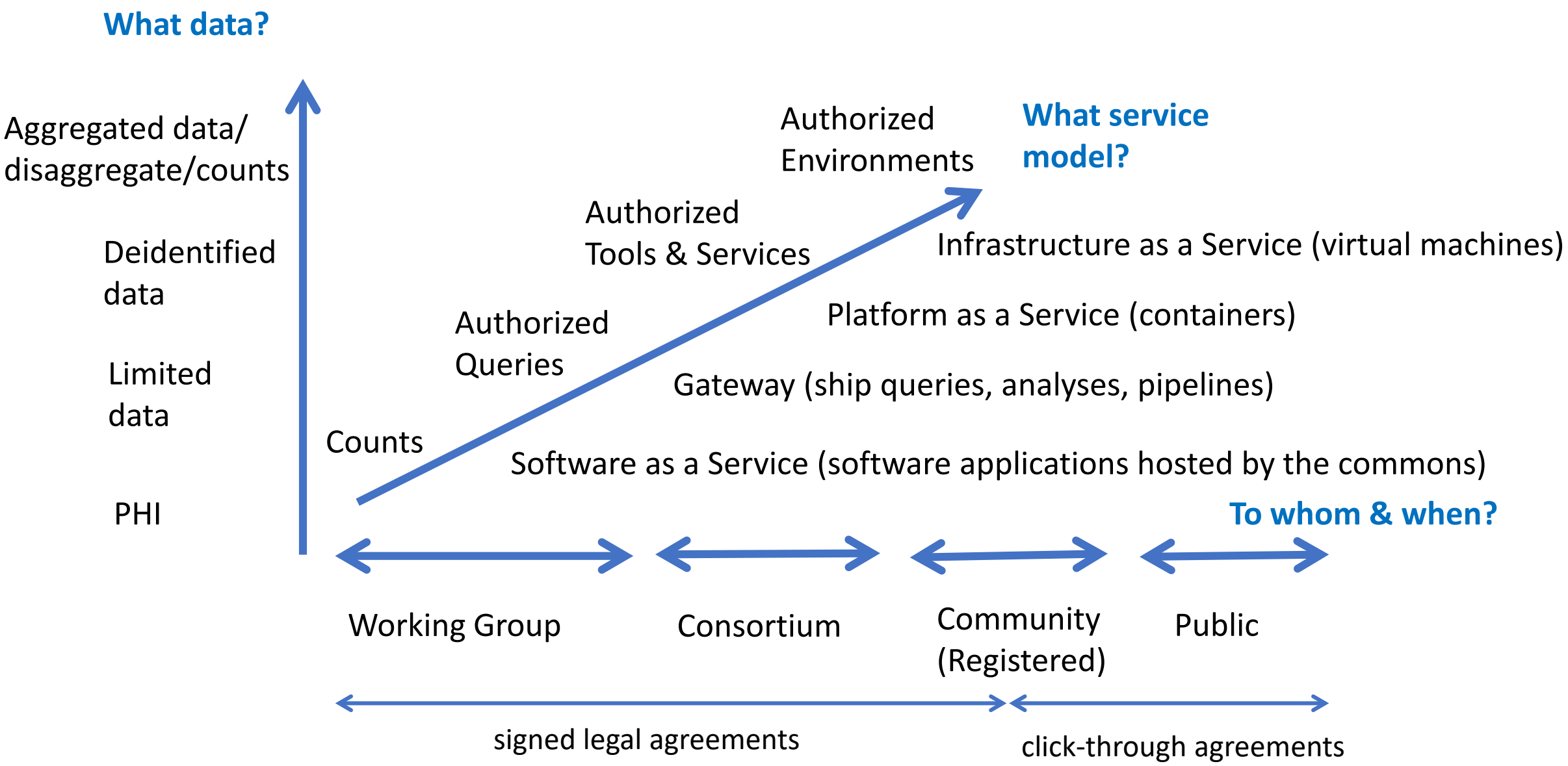
Mobility and other behavioral data

Google COVID-19 Community Mobility Reports

www.google.com/covid19/mobility

COVID-19 Mobility Data Network

www.covid19mobility.org



Source: Robert L. Grossman, Supporting Open Data and Open Science With Data Commons: Some Suggested Guidelines for Funding Organizations, March 23, 2017.

Data Driven Decision Making for a Public Health Emergency

What Data Platform can best support a PHE?

- Testing
- Contact tracing
- Quarantine

public health

- Testing
- Supply chain management
- Surge capacity planning
- Quality improvement

patient care & hospital operations

- Clinical research
- Clinical trials
  - Vaccine
  - drug
- Basic research

medical research

- Incidence levels
- Incidence trends
- Positivity
- Hospital capacity
- Supply chain

data to support governance

- Hot spots
- Hot “vectors”

data to improve community life

2. Data Commons

Data Commons

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

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

NIH NATIONAL CANCER INSTITUTE Genomic Data Commons

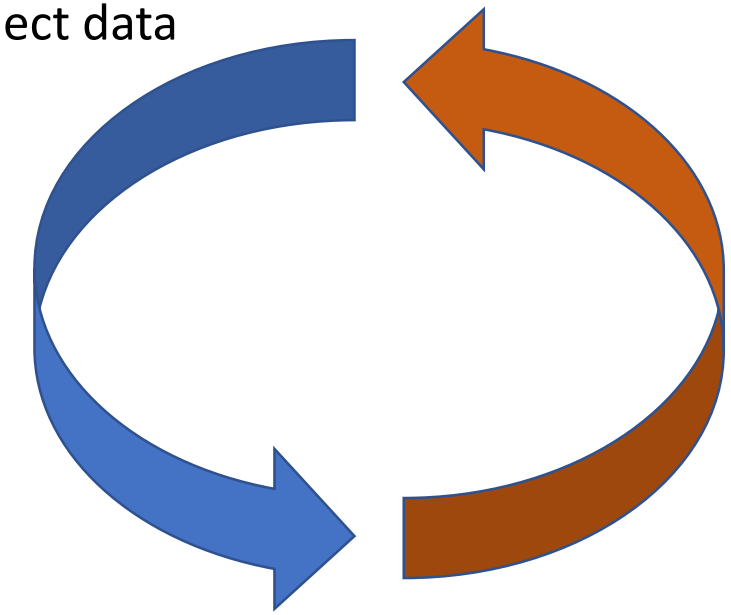
NIH National Heart, Lung, and Blood Institute | BioData CATALYST

NIH NHGRI AnVIL

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

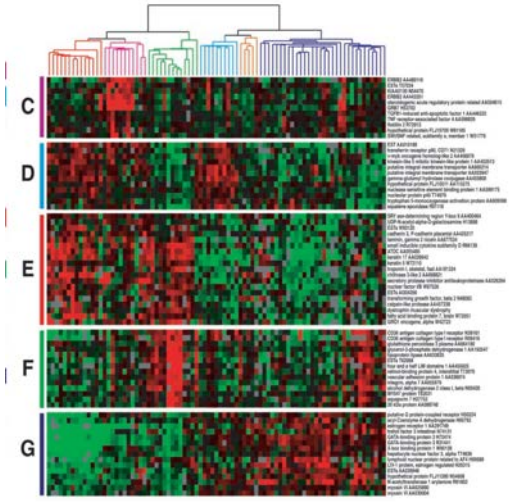


The right of human subjects to **benefit** from research.

[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.

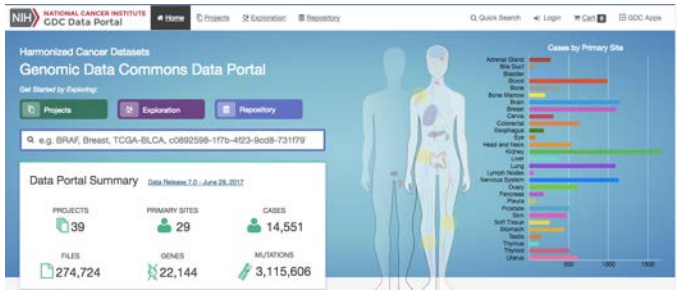
*GA4GH Framework for Responsible Sharing of Genomic and Health-Related Data, see goo.gl/CTavQR



Databases organize data around a **project**.

Data Warehouse

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



Data commons organize data for a **community** and can support scientific, medical and healthcare data for research or support a public health emergency, while providing security and protecting privacy.

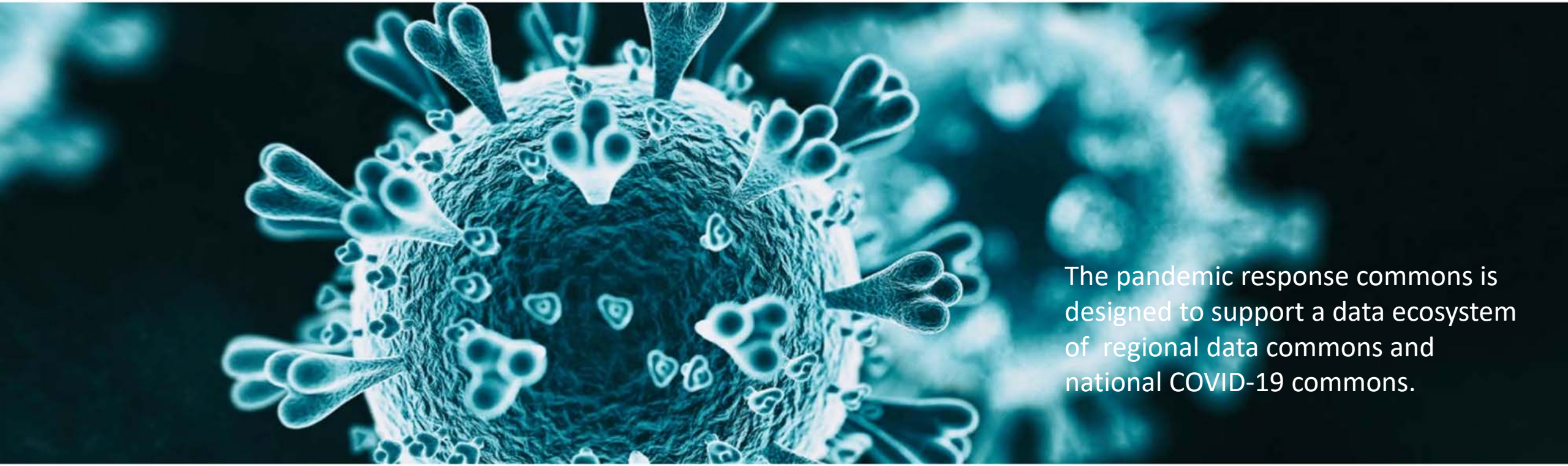
3. Pandemic Response Commons

Pandemic Response Commons

A data commons to improve health outcomes for COVID-19

https://pandemicresponsecommons.org/

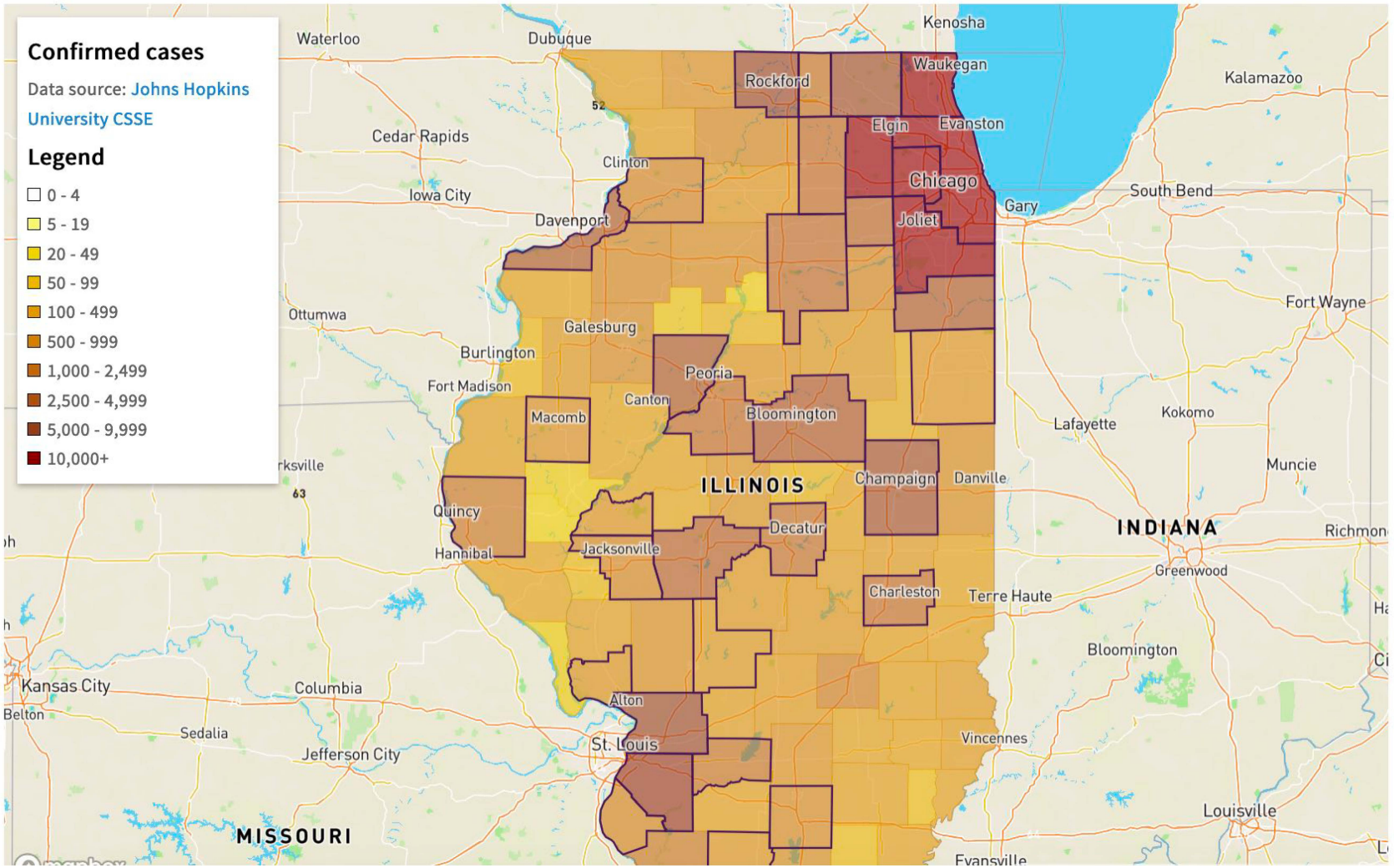
Home About COVID-19 Commons Governance Members and Contributors News Working Groups Contact Us

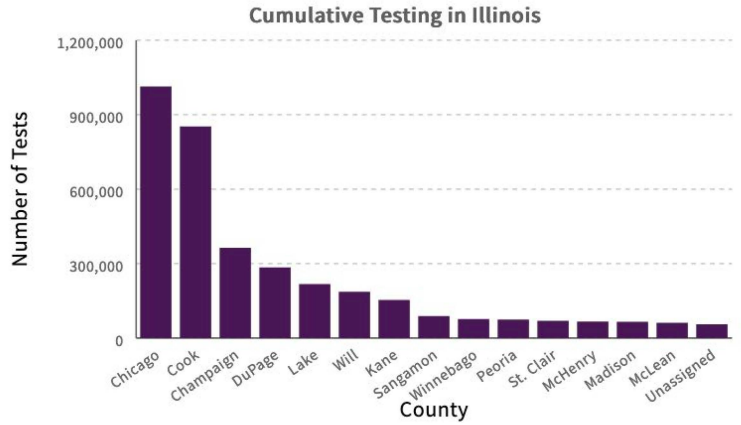


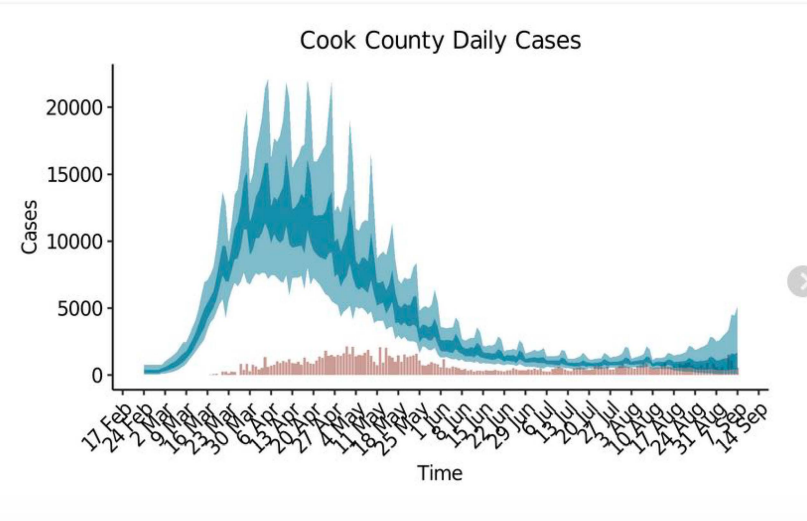
The Pandemic Response Commons represents a collaborative data ecosystem powering research to improve health outcomes, epidemiological models, and back-to-work models. It is developed and operated by a private-public partnership managed by a 501(c)(3) not-for-profit corporation.

Total Confirmed
259,856

Total Deaths
8,211







Pandemic Response Commons Consortium

Chicagoland COVID-19 Consortium Data Partners (regional commons)

RUSH

UIC

SINAI

THE UNIVERSITY OF CHICAGO

SAINT ANTHONY HOSPITAL

NorthShore University HealthSystem

IAFCC
ILLINOIS ASSOCIATION OF FREE & CHARITABLE CLINICS

CommunityHealth

Technology & Consortium Partners

- Open Commons Consortium – 501(c)(3) not-for-profit that operates the commons
- Matter Chicago
- P33 Chicago
- University of Chicago
- BioTeam
- Amazon

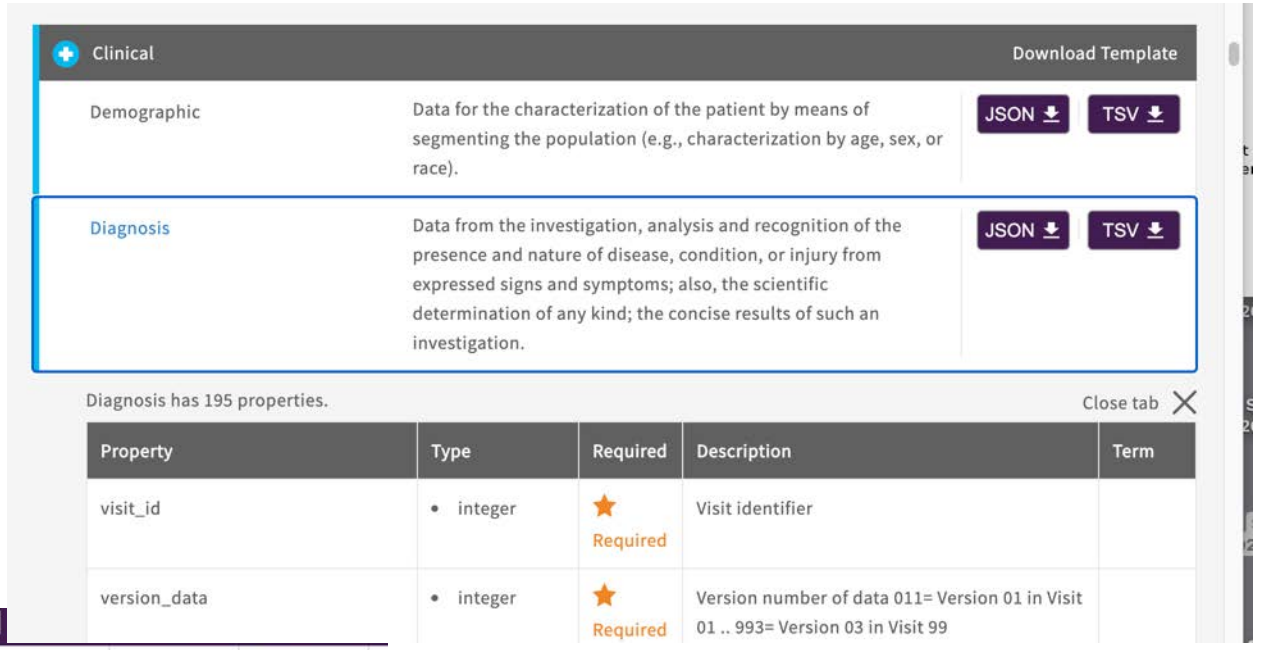
- Our focus is providing the infrastructure, technology and expertise to set up and operate a regional COVID-19 Commons.
- Our first regional commons is in the Chicago region.
- The Consortium is a public-private partnership funded by philanthropy and in-kind donations.

Data Types

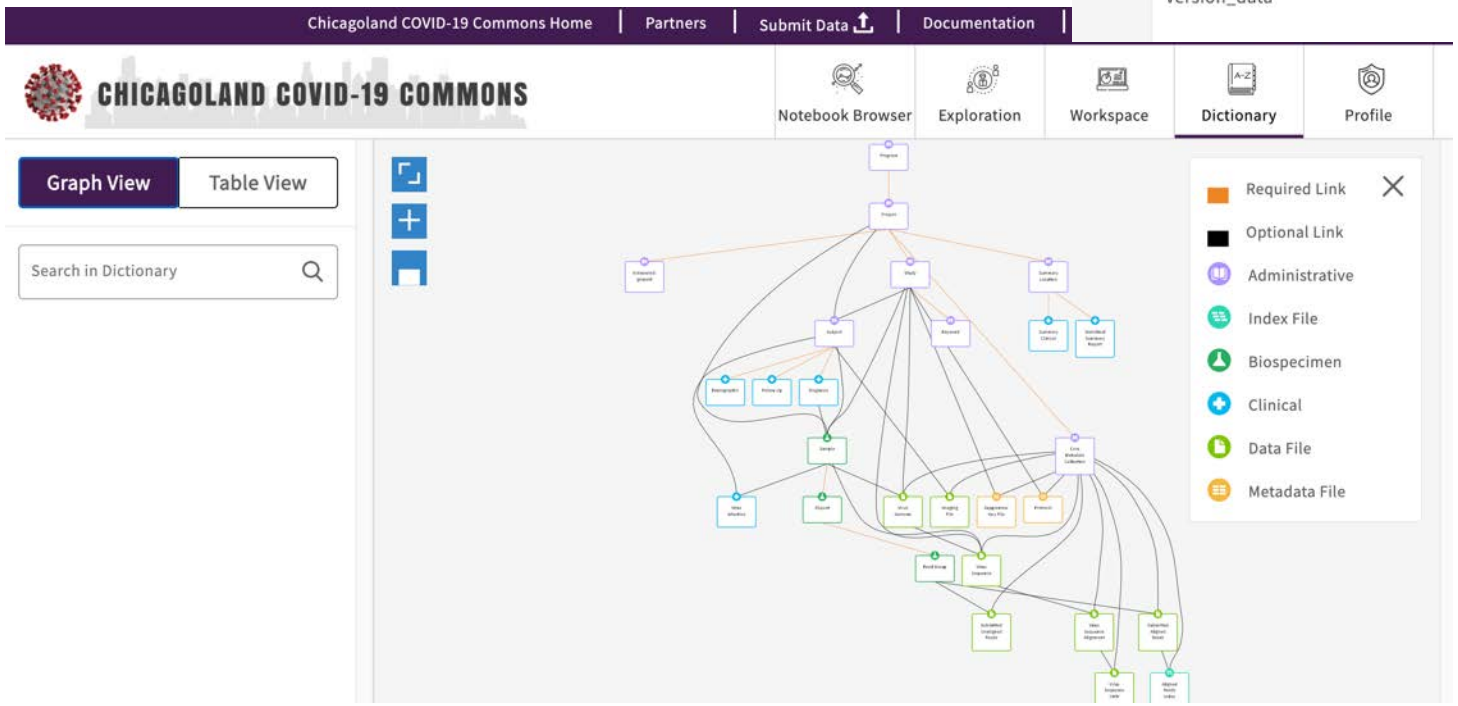
- Clinical data (subject/patient level data)
- Pandemic Response Commons Statistical Summary Reports (SSR)
  - Time series data with accurate dates
  - For epidemiological models and back to work/school models
- Data for researchers
  - De-identified data under safe-harbor (dates are approximate)
  - Clinical/Phenotype data
  - Molecular (viral and host) data
  - Image data
- COVIDStopLight resident-donated data (not associated with a health care provider)
  - Progressive app (works on the web or a smart phone)
  - For identifying hot spots / hot vectors in de-identified / privacy preserving manner
  - How do you feel today?

Data Model

- 1,100 data attributes
- Clinical and phenotype data
- Molecular data (host and viral)
- Resident donated status data
- Biospecimen data

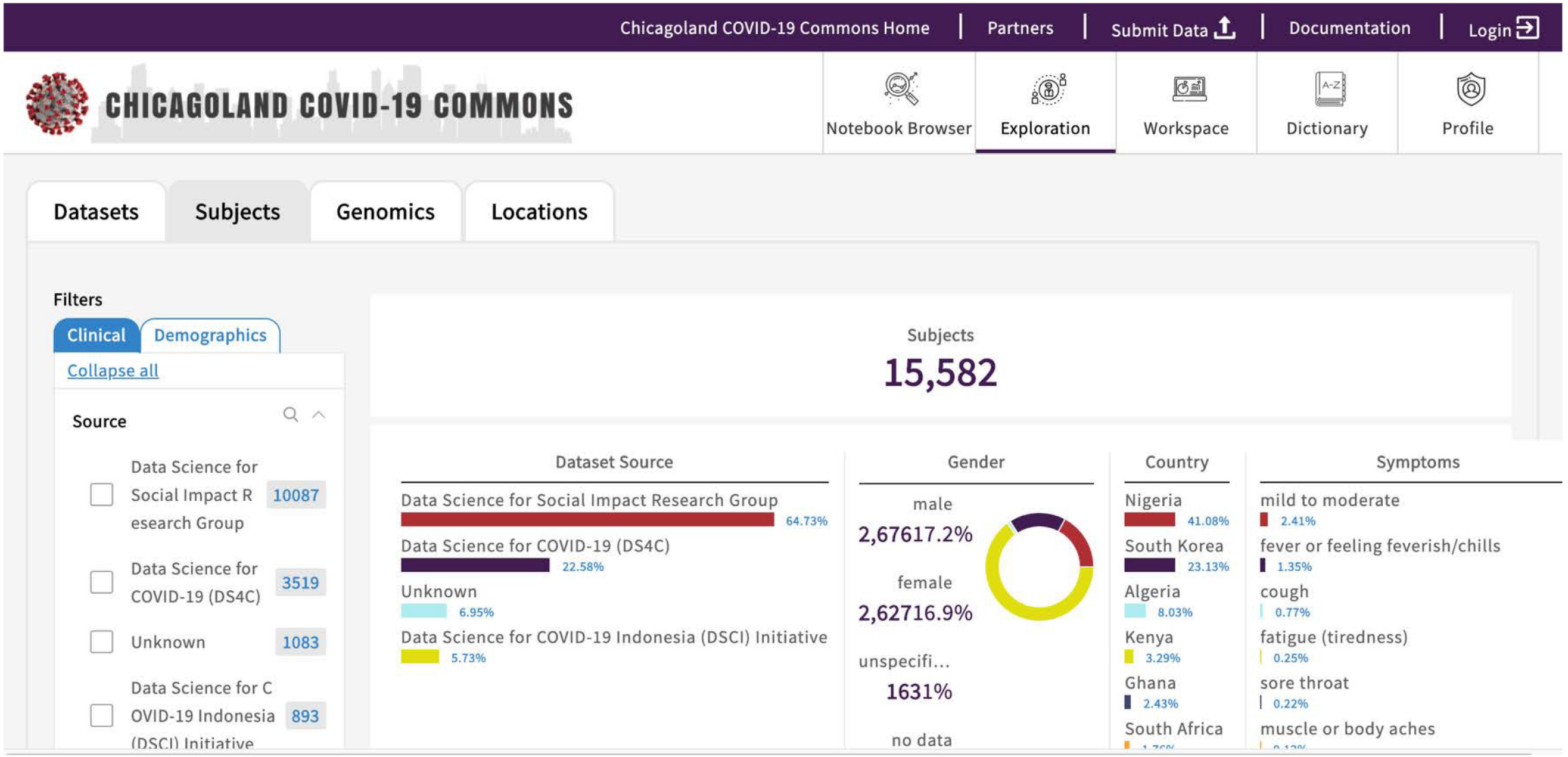


Property	Type	Required	Description	Term
visit_id	• integer	★ Required	Visit identifier	
version_data	• integer	★ Required	Version number of data 011= Version 01 in Visit 01 .. 993= Version 03 in Visit 99	



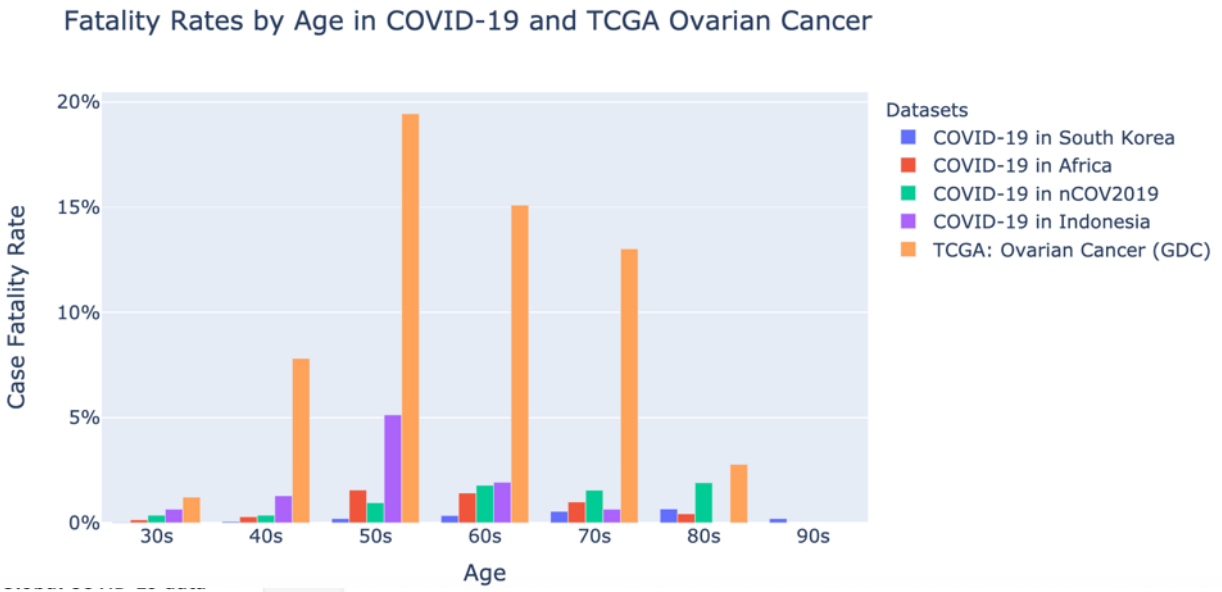
cld_yn	• Yes • No • Unknown	No	Chronic lung disease (asthma/emphysema/COPD)
collected_from	• array	No	Collected from (check all that apply): [Patient interview] [Medical record review]
comorbidity	• boolean	No	The simultaneous presence of two chronic diseases or conditions in a patient
comorbidity_anemia	• boolean	No	Presence of anemia as comorbidity
comorbidity_diabetes	• boolean	No	Presence of diabetes as comorbidity
comorbidity_hepatic_disease	• boolean	No	Presence of hepatic disease as comorbidity
comorbidity_hiv	• boolean	No	Presence of HIV as comorbidity
comorbidity_renal_disease	• boolean	No	Presence of renal disease as comorbidity

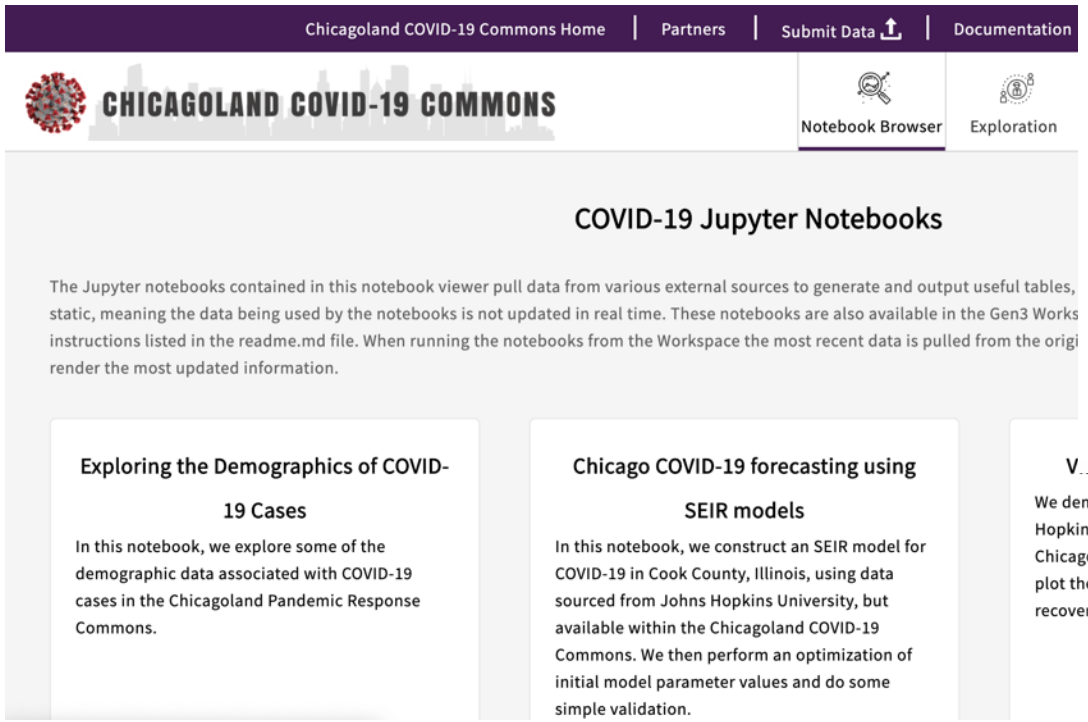
Gen3 Data Explorer



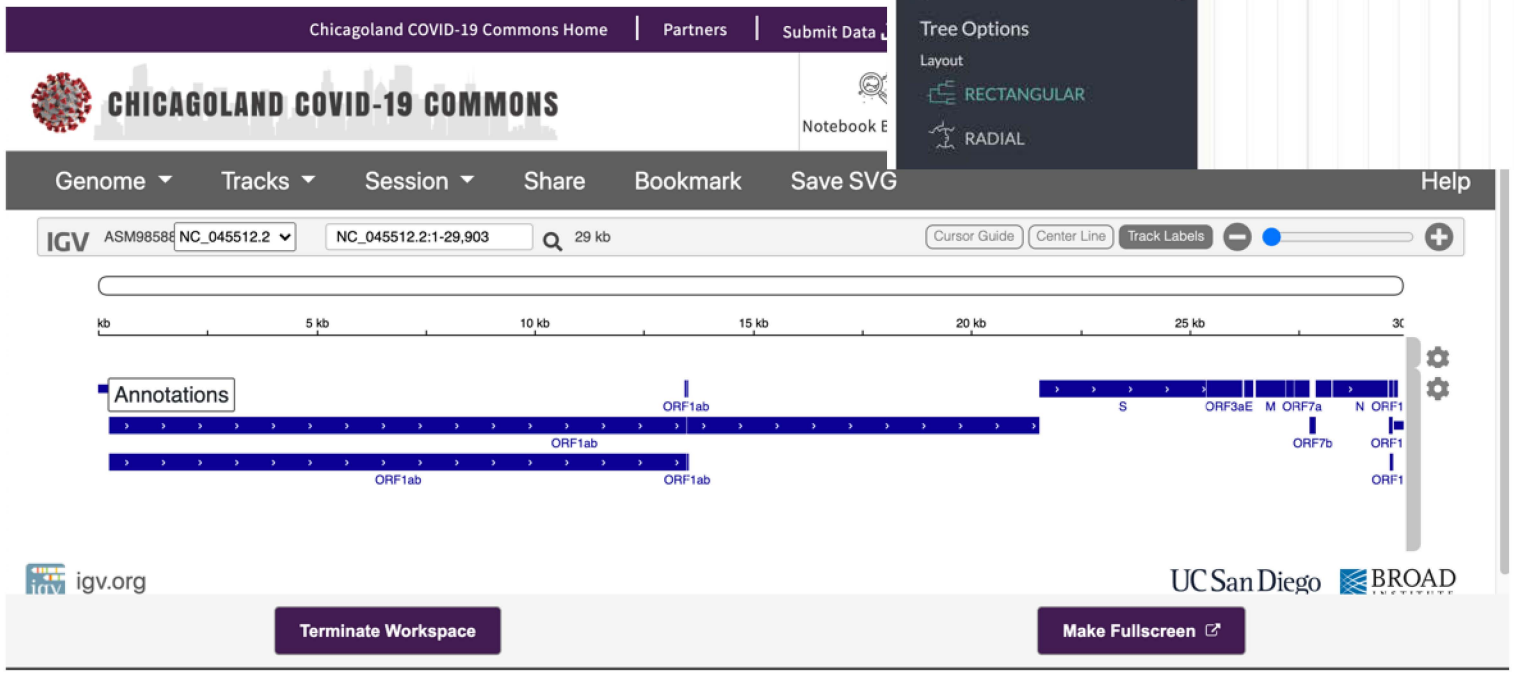
Gen3 Workspaces – Jupyter Notebooks R-Studio

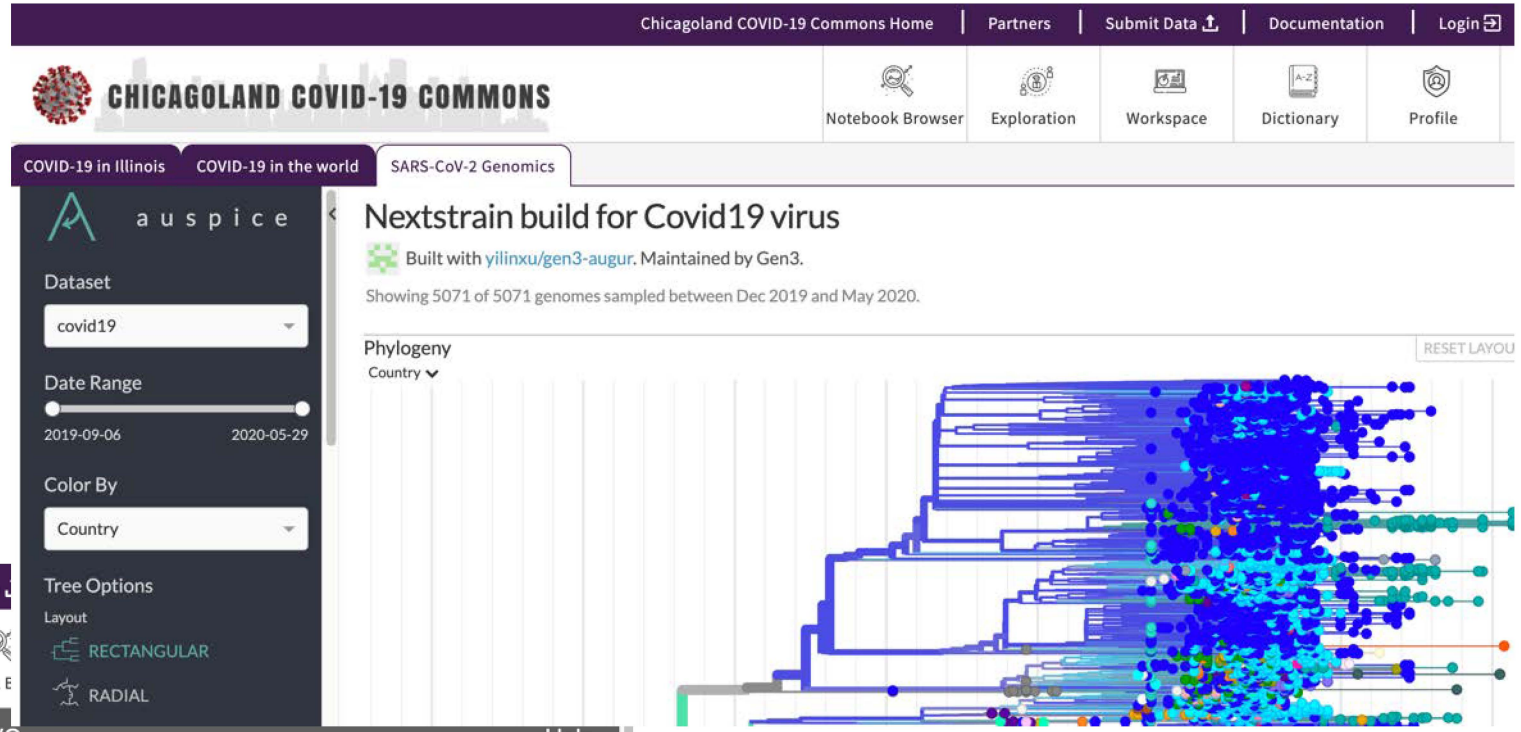
```
        go.Bar(name="TCGA: Ovarian Cancer (GDC)", x=age_decade, y=fatality_all['TCGA: Ovarian Cancer']),
    ]
)

fig.update_layout(
    yaxis=dict(tickformat=".0%"),
    barmode="group",
    title="Fatality Rates by Age in COVID-19 and TCGA Ovarian Cancer",
    xaxis_title="Age",
    yaxis_title="Case Fatality Rate",
    legend_title="Datasets",
    font=dict(size=15),
)
fig.show("notebook")
```





Viral Sequence Data





COVID StopLight - Data Donated by Area Residents

COVID stoplight

CovidStopLight.org

- Progressive app so it can be run from a browser or downloaded to your smartphone
- Second app being beta tested to identify hot spots in buildings when workers or residents opt in

4. Recommendations

Recommendations

1. Set up national and regional commons to support public health emergencies.
2. Establish trust relationship between two commons that have similar privacy and security policies and share data through APIs. Support federated queries in other cases. In all cases, balance privacy and security with the need to improve health outcomes and improve community life.
3. Set up a persistent data commons infrastructure that is ready for future responses to epidemics and pandemics. Have data sharing agreements in place and data infrastructure ready to go.
4. Agencies and foundations that fund biomedical research should require that researchers share data and provide the computing infrastructure (“commons”) and bioinformatics resources that are required to support secure, compliant data sharing that respects privacy.

- **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]
- **Data Ecosystem of multiples commons.** 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
- **Interoperating data dommons.** Robert L. Grossman, Some Proposed Principles for Interoperating Data Commons, Medium, October 1, 2019.
- 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
- **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.
- **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
- **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 is a Gen3 data commons.
- **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 first generation data commons.

ctds.uchicago.edu

occ-data.org