# Future Directions of AnVIL Workshop

29OCT2021

## - Introduction to AnVIL -

Anthony Philippakis and Michael Schatz

# Introductions

**Michael Schatz**
Johns Hopkins University
Computer Science and Biology

**Anthony Philippakis**
Broad Institute
Chief Data Officer & Institute Scientist

# AnVIL Working Groups + Committees

## Technical Working Group

Chairs: Michael Schatz (JHU)
Brian O'Connor (Broad)

## Data Access Working Group

Chairs: Stacey Donnelly (Broad)
Carolyn Hutter (NHGRI)

## Outreach Working Group

Chairs: Jeffrey Leek (JHU)
Frederick Tan (JHU)

## Data Processing Working Group

Chairs: Eric Banks (Broad),
Ira Hall (WashU/Yale)

## Portal Working Group

Chairs: Michael Schatz (JHU)
Benedict Paten (UCSC)

## Phenotype Working Group

Chairs: David Crosslin (eMERGE - UW)
Robert Carroll (VUMC)

## Data Ingestion Committee

Members: Michael Schatz (JHU)
Anthony Philippakis (Broad) et al

## AHA/AnVIL Working Group

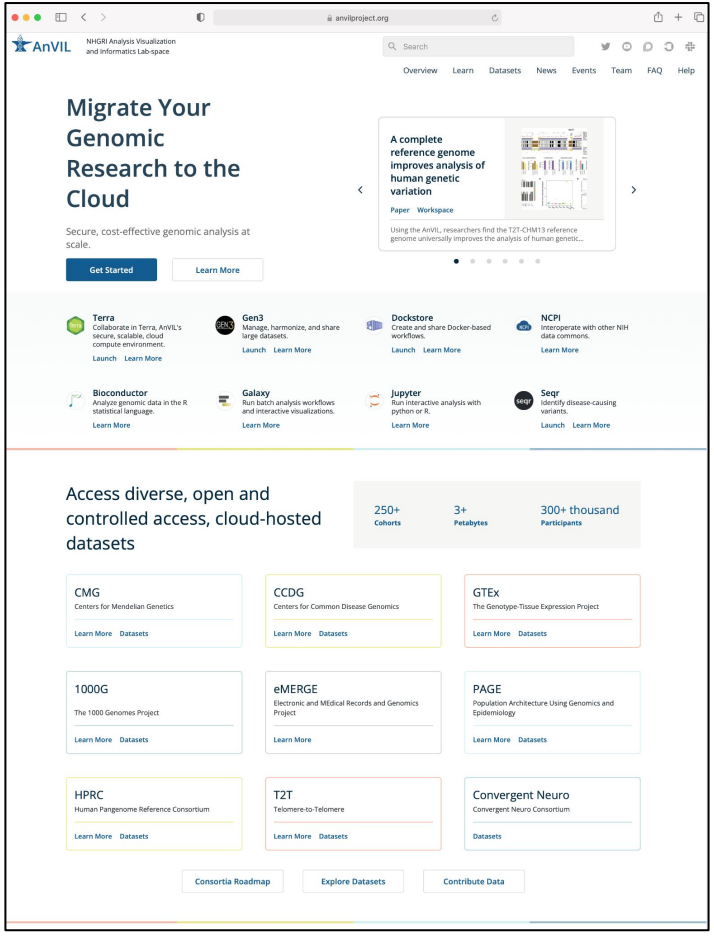
Members: Michael Schatz (JHU)
Anthony Philippakis (Broad) et al

With extensive participation from all sites

# What is the AnVIL?

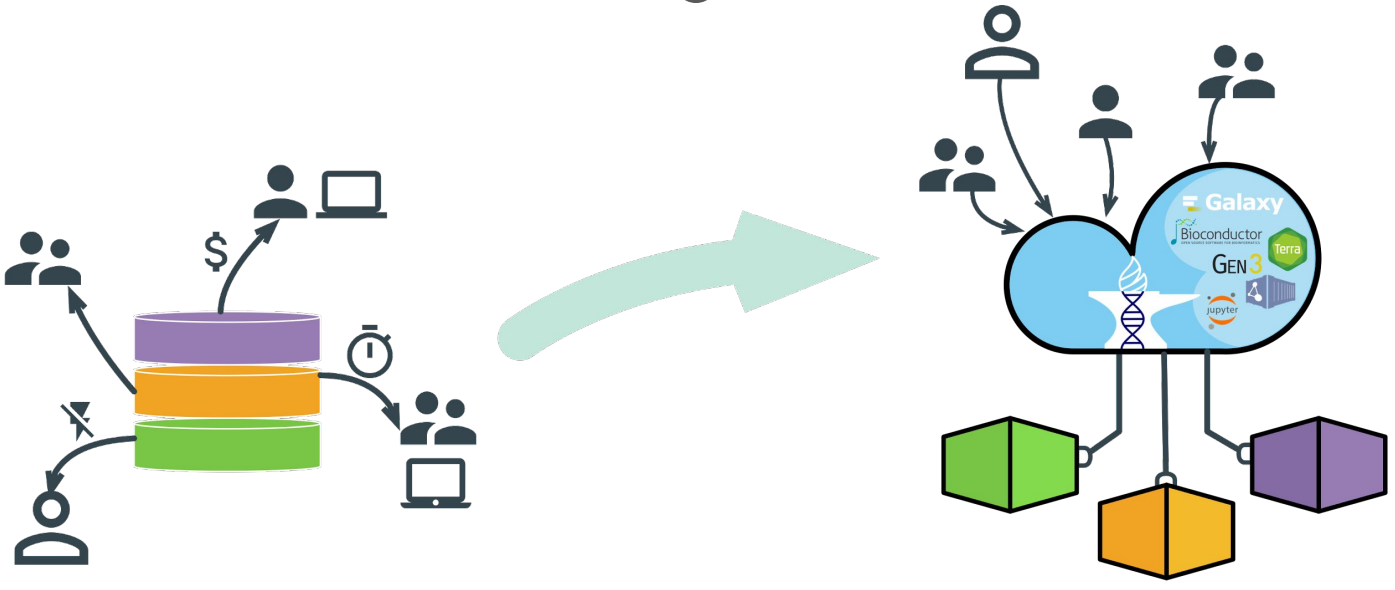
## Scalable and interoperable computing resource for the genomics scientific community

- Cloud-based infrastructure
  - Highly elastic; shared analysis and computing environment
- Data access and security
  - Genomic datasets, phenotypes and metadata
  - Large datasets generated by NHGRI programs, as well as other initiatives / agencies
  - dbGaP Authenticated sharing of primary and derived datasets
- Collaborative computing environment for datasets and analysis workflows
  - Storage, scalable analytics, data visualization
  - Security, training & outreach, with new models of data access
  - ...for both users with limited computational expertise and sophisticated data scientist users




**https://anvilproject.org**

# AnVIL: Inverting the model of genomic data sharing




*Traditional: Bring data to the researcher*

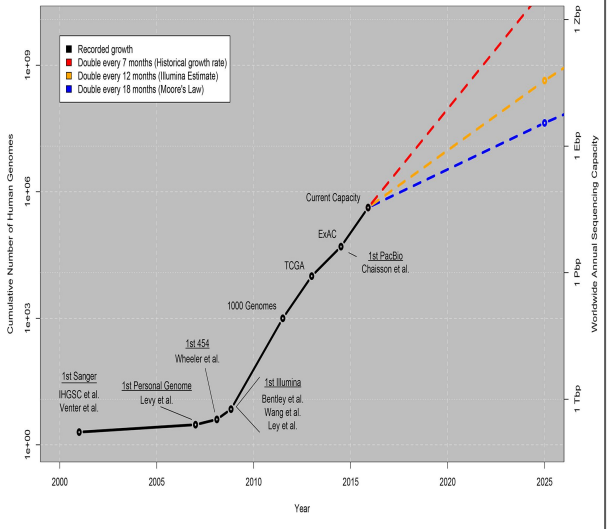
- Copying/moving data is costly
- Harder to enforce security
- Redundant infrastructure
- Siloed compute

*Goal: Bring researcher to the data*

- Reduced redundancy and costs
- Active threat detection and auditing
- Greater accessibility
- Elastic, shared, compute

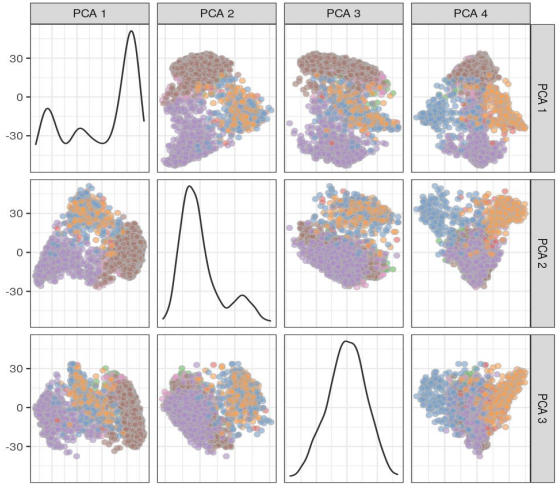
# Why AnVIL?

## *Data*




Scale,
Integration,
Sharing & Reuse

## *Computation*

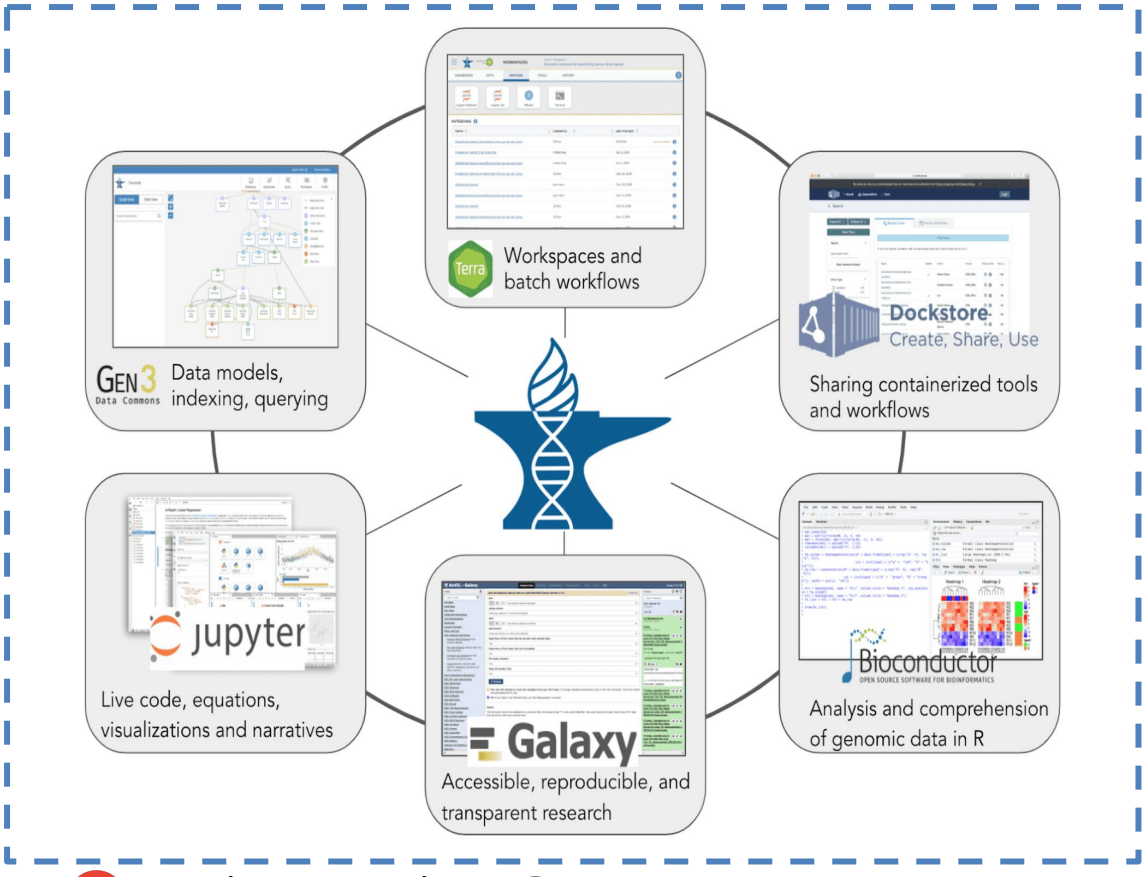



Simplicity,
Reproducibility,
Security

## *Users*

Democratization,
Collaboration,
Discoveries

# Building a Secure Federated Data Ecosystem




FedRAMP

FedRAMP certified
1 ATO

Implemented on Google Cloud Platform

# AnVIL by the numbers

## Data

| | Gen3 | Total |
|---|---|---|
| Consortia | | 9 |
| Cohorts | | 254 |
| Subjects | 22,071 | 291,301 |
| Samples | 69,787 | 314,038 |
| Size | | 3.87 PB |

## Tools & Workflows

| | |
|---|---|
| Dockstore: | WDL: 840 workflows<br>Galaxy: 28 workflows |
| Terra: | 272 public workspaces<br>48 featured workspaces |
| Bioconductor: | 2,041 software packages<br>977 annotation resources<br>406 data collections |
| Galaxy: | 7,829 tools available |

## Users

Visits

| | |
|---|---|
| anvilproject.org | 6731 in Q3 |
| anvil.terra.bio | 440 / month |

Terra usage

| | |
|---|---|
| Users | >15,000 |
| Public Workspaces | 272 |
| Cloned Workspaces | 625 |

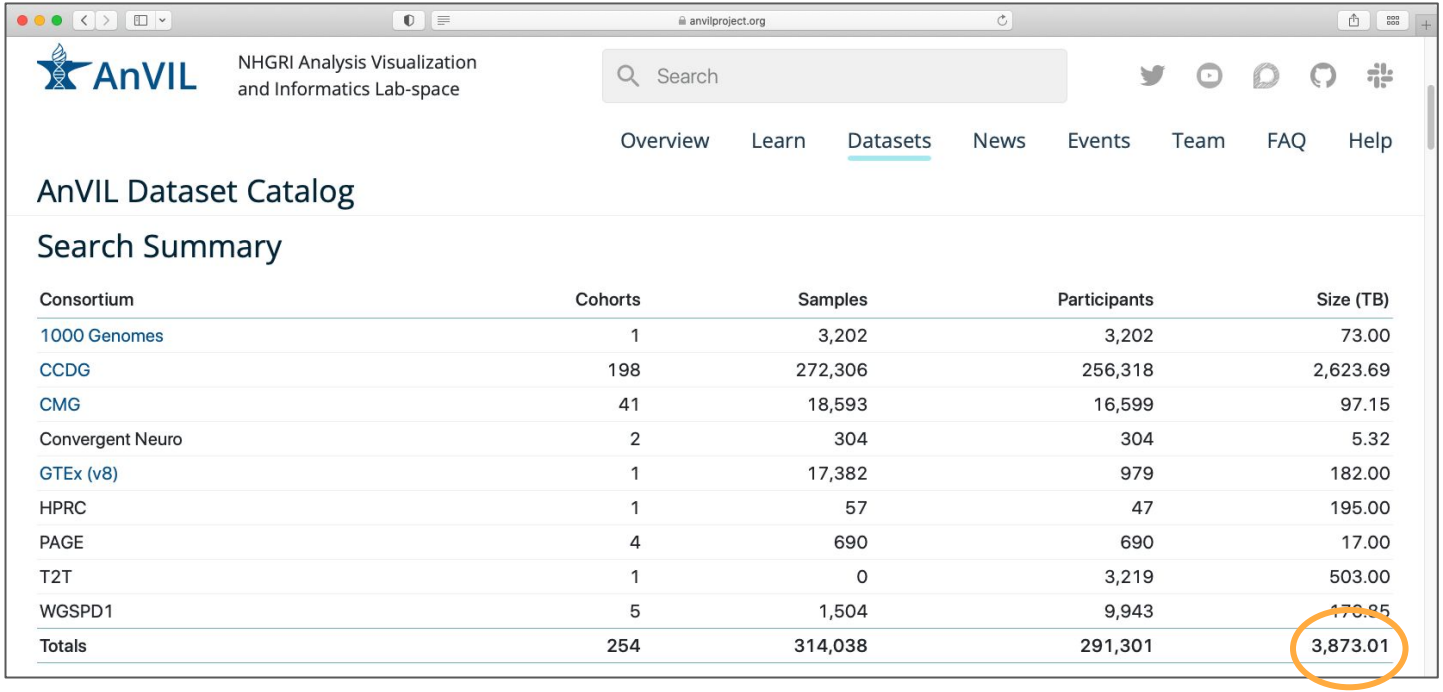
## Terra Launches

| | |
|---|---|
| Workflows | 3082 |
| Jupyter | 1305 |
| RStudio | 1562 |
| Galaxy | 470 |

## Communications

| | |
|---|---|
| Twitter | 648 followers |
| Slack | 234 users |

# AnVIL Dataset Catalog



anvilproject.org

AnVIL — NHGRI Analysis Visualization and Informatics Lab-space

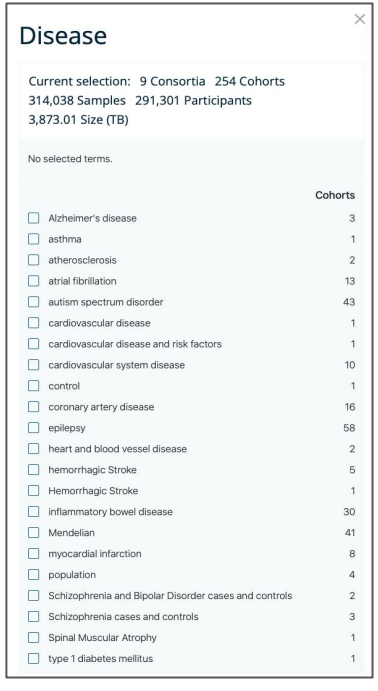
Search

Overview | Learn | Datasets | News | Events | Team | FAQ | Help

## AnVIL Dataset Catalog

### Search Summary

| Consortium | Cohorts | Samples | Participants | Size (TB) |
|---|---|---|---|---|
| 1000 Genomes | 1 | 3,202 | 3,202 | 73.00 |
| CCDG | 198 | 272,306 | 256,318 | 2,623.69 |
| CMG | 41 | 18,593 | 16,599 | 97.15 |
| Convergent Neuro | 2 | 304 | 304 | 5.32 |
| GTEx (v8) | 1 | 17,382 | 979 | 182.00 |
| HPRC | 1 | 57 | 47 | 195.00 |
| PAGE | 4 | 690 | 690 | 17.00 |
| T2T | 1 | 0 | 3,219 | 503.00 |
| WGSPD1 | 5 | 1,504 | 9,943 | 170.85 |
| **Totals** | **254** | **314,038** | **291,301** | **3,873.01** |



Disease

Current selection: 9 Consortia 254 Cohorts
314,038 Samples 291,301 Participants
3,873.01 Size (TB)

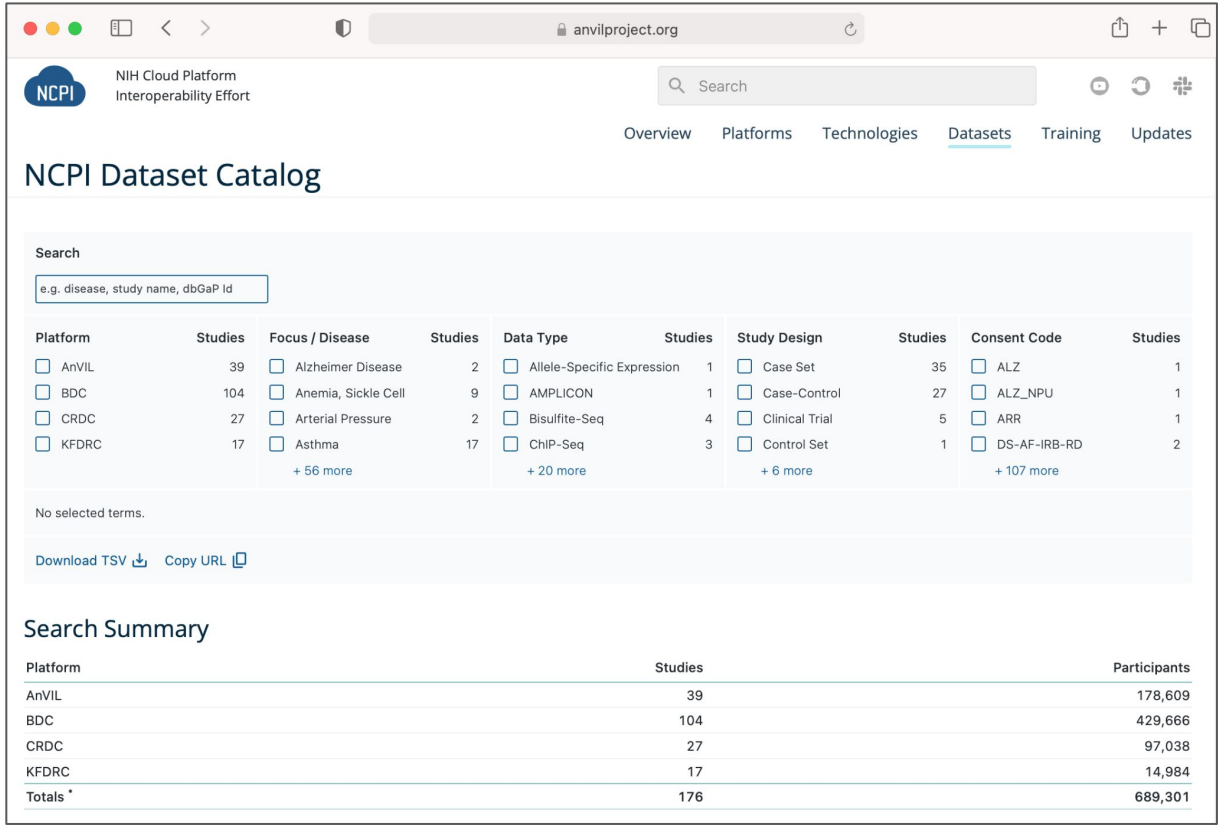
No selected terms.

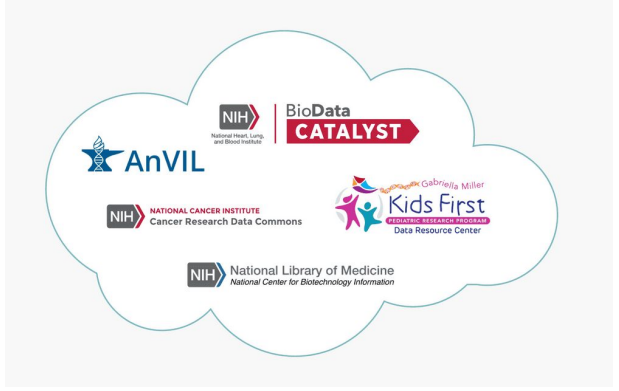
| | Cohorts |
|---|---|
| Alzheimer's disease | 3 |
| asthma | 1 |
| atherosclerosis | 2 |
| atrial fibrillation | 13 |
| autism spectrum disorder | 43 |
| cardiovascular disease | 1 |
| cardiovascular disease and risk factors | 1 |
| cardiovascular system disease | 10 |
| control | 1 |
| coronary artery disease | 16 |
| epilepsy | 58 |
| heart and blood vessel disease | 2 |
| hemorrhagic Stroke | 5 |
| Hemorrhagic Stroke | 1 |
| inflammatory bowel disease | 30 |
| Mendelian | 41 |
| myocardial infarction | 8 |
| population | 4 |
| Schizophrenia and Bipolar Disorder cases and controls | 2 |
| Schizophrenia cases and controls | 3 |
| Spinal Muscular Atrophy | 1 |
| type 1 diabetes mellitus | 1 |

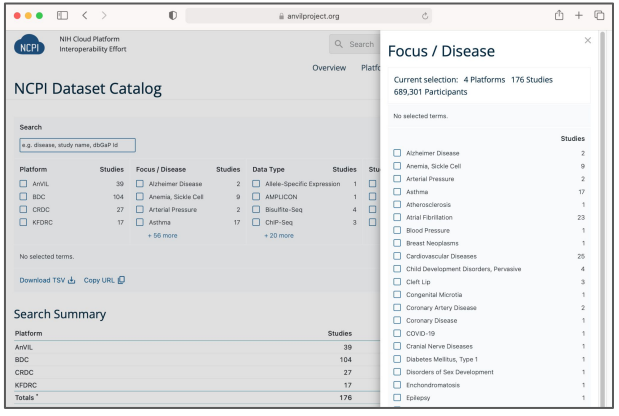
- >3.8Pb of data, >291,000 participants available
- Population-scale cohorts powers disease association studies
- Cross-project synthetic cohorts make existing data more valuable
- Connect multiple datatypes together to make new discoveries

https://anvilproject.org/data

# NIH Cloud Platform Interoperability (NCPI) Dataset Catalog




11Pb / 689k participants and growing!
Cross-platform accessibility through several key technologies (RAS, DRS, FHIR)

# Outreach & User Engagement

Upcoming

03 NOV 2021
Genome Informatics 2021 · POSTER SESSION
**Modeling the computing requirements and costs for genomics analysis in the cloud**
The 2021 Genome Informatics Meeting will cover topics including Microbial and Metagenomics; Sequencing Algorithms, Variant Discovery and Genome Assembly; Evolution, Complex Traits and Phylogenetics; Functional Genomics; Single Cell Genomics; and Epigenetics and Genome Structure.

19 JAN 2022
ASHG 2021 · INTERACTIVE WORKSHOP
**Structural variant discovery from long-read sequencing data on the cloud with Galaxy in Terra**
In this workshop, we will guide you through an end-to-end SV identification journey using Galaxy, a platform designed to facilitate access to computational methods for researchers without a programming background.

26 JAN 2022
ASHG 2021 · INTERACTIVE WORKSHOP
**Reproducible Analysis of Human Pangenome Data using the AnVIL**
This workshop will explore and demonstrate open access data from the Human Pangenome Research Consortium (HPRC), an NHGRI funded effort to create a more diverse and comprehensive reference human pangenome.

## Announcing the AnVIL Cloud Credits Program (AC2) Awardees

Posted: June 03, 2021

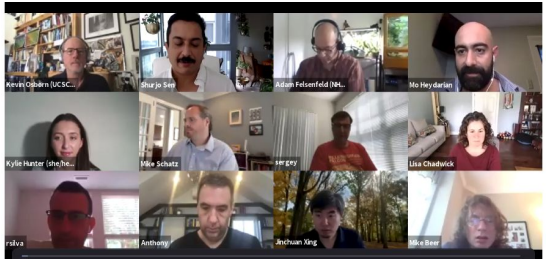
NHGRI's Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL) cloud genomics platform is pleased to announce the awardees of the pilot phase of the AnVIL Cloud Credits (AC2) Program.

### Awardees

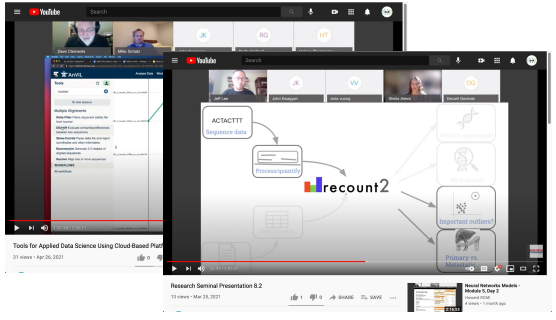
Seventeen proposals were received from 14 different institutions and of these, the AC2 Review Committee (AC2RC) has awarded 6 proposals with cloud credits. Those awardees include:

- Alex Greiner | The University of Iowa | Graduate Student, "Burden analysis of inherited cardiac arrhythmia genes in epilepsy"
- Melissa Suzanne Cline | UC Santa Cruz Genomics Institute | Principal Investigator, "Leveraging AnVIL and Terra for secure collaboration on genetic variant interpretation"
- Andrew Davidson | University of California | Graduate Student, "Comprehensive characterization of transposable element expression across human tissues"
- Anahita Khojandi | University of Tennessee-Knoxville | Associate Professor, "Deep Learning for Accurate Tissue-Specific Prediction of Gene Expression in Large Deeply-Phenotyped Population"
- Anshul Kundaje | Stanford University | Principal Investigator, "Deciphering cis-regulatory syntax of a transcription factor binding atlas with interpretable deep learning models"
- Tychele N. Turner | Washington University in St. Louis | Principal Investigator, "A k-mer based approach to assess copy number in PacBio HiFi data"

GSP/CCDG/CMG MAGIC Jamboree

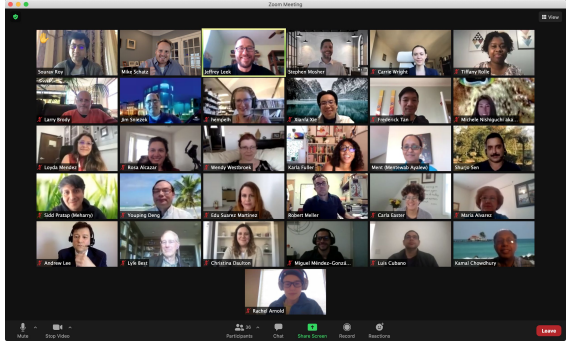



Howard University VADSTI




PRIMED Consortium


Genomic Data Science Community Network



**Portal:** http://anvilproject.org/
**Mailing List:** help@lists.anvilproject.org
**AnVIL Community Slack:** http://bit.ly/anvil-community

# AnVIL Manuscript

Cold Spring Harbor Laboratory

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

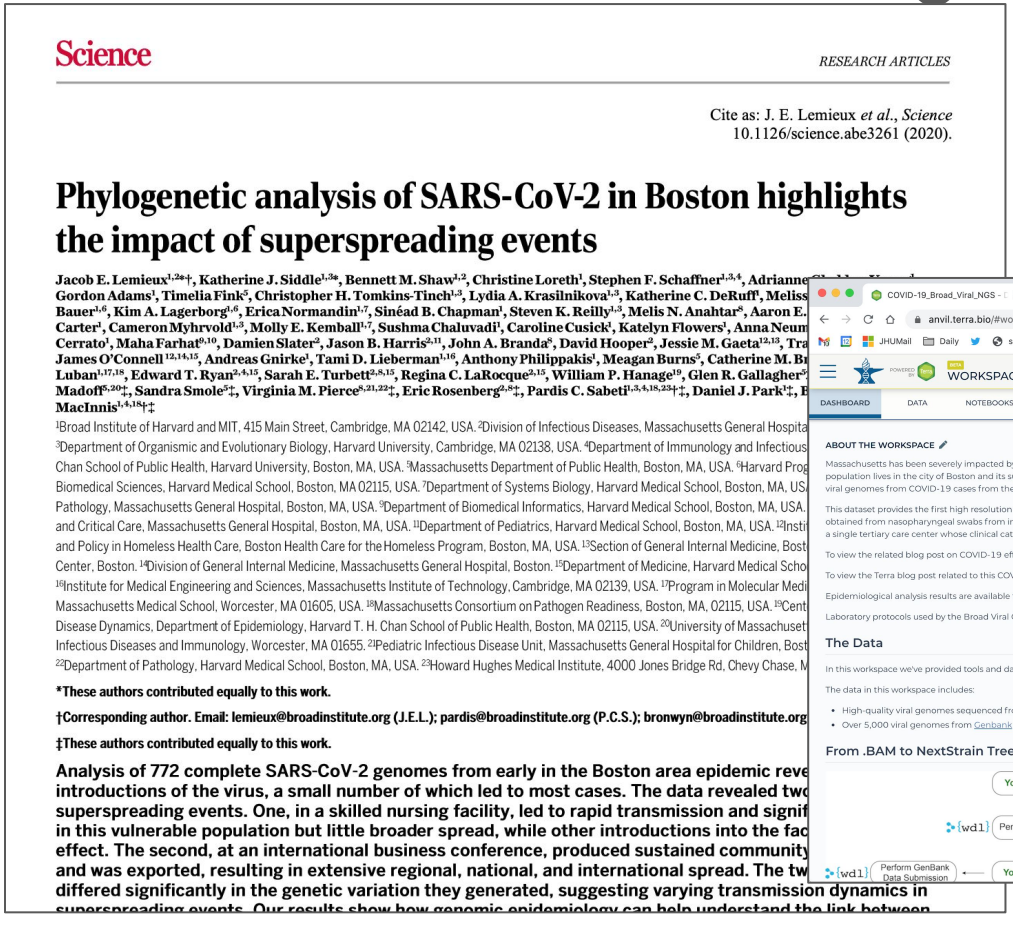
New Results

Follow this preprint

## Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL)

Michael C. Schatz, Anthony A. Philippakis, Enis Afgan, Eric Banks, Vincent J. Carey, Robert J. Carroll, Alessandro Culotti, Kyle Ellrott, Jeremy Goecks, Robert L. Grossman, Ira M. Hall, Kasper D. Hansen, Jonathan Lawson, Jeffrey T. Leek, Anne O'Donnell Luria, Stephen Mosher, Martin Morgan, Anton Nekrutenko, Brian D. O'Connor, Kevin Osborn, Benedict Paten, Candace Patterson, Frederick J. Tan, Casey Overby Taylor, Jennifer Vessio, Levi Waldron, Ting Wang, Kristin Wuichet, AnVIL Team

# AnVIL-powered COVID19 Analysis



Science

RESEARCH ARTICLES

Cite as: J. E. Lemieux *et al.*, *Science*
10.1126/science.abe3261 (2020).

## Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events


Jacob E. Lemieux[1,2*†], Katherine J. Siddle[1,3*], Bennett M. Shaw[1,2], Christine Loreth[1], Stephen F. Schaffner[1,3,4], Adrianne
Gordon Adams[1], Timelia Fink[5], Christopher H. Tomkins-Tinch[1,3], Lydia A. Krasilnikova[1,3], Katherine C. DeRuff[3], Meliss
Bauer[1,6], Kim A. Lagerborg[1,6], Erica Normandin[1,7], Sinéad B. Chapman[1], Steven K. Reilly[1,3], Melis N. Anahtar[8], Aaron E.,
Carter[1], Cameron Myhrvold[1,3], Molly E. Kemball[1,7], Sushma Chaluvadi[1], Caroline Cusick[1], Katelyn Flowers[1], Anna Neum
Cerrato[1], Maha Farhat[9,10], Damien Slater[2], Jason B. Harris[2,11], John A. Branda[8], David Hooper[2], Jessie M. Gaeta[12,13], Tra
James O'Connell[12,14,15], Andreas Gnirke[1], Tami D. Lieberman[1,16], Anthony Philippakis[1], Meagan Burns[5], Catherine M. Br
Luban[1,17,18], Edward T. Ryan[2,4,15], Sarah E. Turbett[2,8,15], Regina C. LaRocque[2,15], William P. Hanage[19], Glen R. Gallagher[5]
Madoff[5,20‡], Sandra Smole[5‡], Virginia M. Pierce[8,21,22‡], Eric Rosenberg[2,8*‡], Pardis C. Sabeti[1,3,4,18,23*‡], Daniel J. Park[1*‡], B
MacInnis[1,4,18*‡]

[1]Broad Institute of Harvard and MIT, 415 Main Street, Cambridge, MA 02142, USA. [2]Division of Infectious Diseases, Massachusetts General Hospita
[3]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA. [4]Department of Immunology and Infectious
Chan School of Public Health, Harvard University, Boston, MA, USA. [5]Massachusetts Department of Public Health, Boston, MA, USA. [6]Harvard Prog
Biomedical Sciences, Harvard Medical School, Boston, MA 02115, USA. [7]Department of Systems Biology, Harvard Medical School, Boston, MA, US
Pathology, Massachusetts General Hospital, Boston, MA, USA. [9]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA.
and Critical Care, Massachusetts General Hospital, Boston, MA, USA. [11]Department of Pediatrics, Harvard Medical School, Boston, MA, USA. [12]Insti
and Policy in Homeless Health Care, Boston Health Care for the Homeless Program, Boston, MA, USA. [13]Section of General Internal Medicine, Bost
Center, Boston. [14]Division of General Internal Medicine, Massachusetts General Hospital, Boston. [15]Department of Medicine, Harvard Medical Scho
[16]Institute for Medical Engineering and Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [17]Program in Molecular Medi
Massachusetts Medical School, Worcester, MA 01605, USA. [18]Massachusetts Consortium on Pathogen Readiness, Boston, MA, 02115, USA. [19]Cent
Disease Dynamics, Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA 02115, USA. [20]University of Massachuset
Infectious Diseases and Immunology, Worcester, MA 01655. [21]Pediatric Infectious Disease Unit, Massachusetts General Hospital for Children, Bost
[22]Department of Pathology, Harvard Medical School, Boston, MA, USA. [23]Howard Hughes Medical Institute, 4000 Jones Bridge Rd, Chevy Chase, M

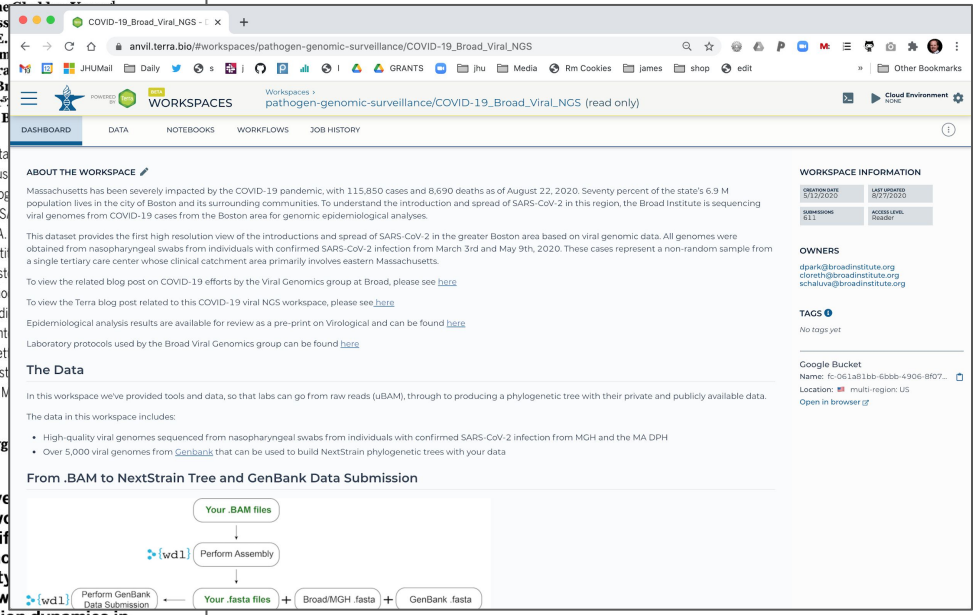
***These authors contributed equally to this work.**

**†Corresponding author. Email: lemieux@broadinstitute.org (J.E.L.); pardis@broadinstitute.org (P.C.S.); bronwyn@broadinstitute.org**

**‡These authors contributed equally to this work.**



**Analysis of 772 complete SARS-CoV-2 genomes from early in the Boston area epidemic reve
introductions of the virus, a small number of which led to most cases. The data revealed tw
superspreading events. One, in a skilled nursing facility, led to rapid transmission and signif
in this vulnerable population but little broader spread, while other introductions into the fac
effect. The second, at an international business conference, produced sustained community
and was exported, resulting in extensive regional, national, and international spread. The tw
differed significantly in the genetic variation they generated, suggesting varying transmission dynamics in
superspreading events. Our results show how genomic epidemiology can help understand the link between**

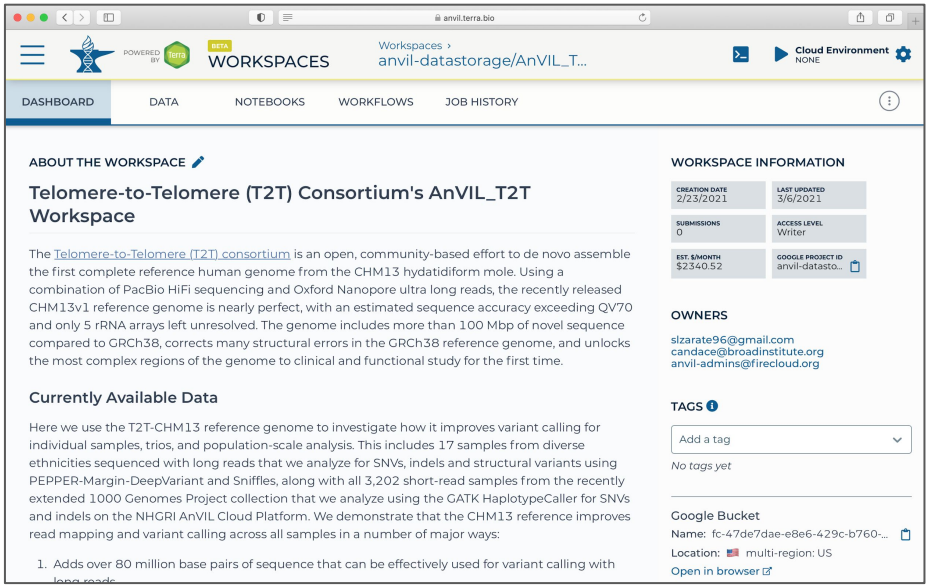



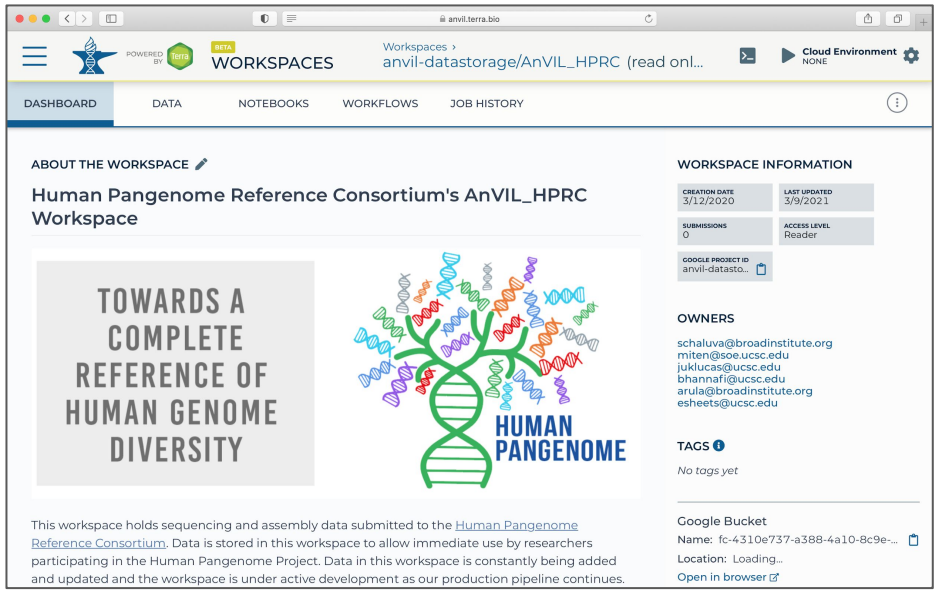

https://anvil.terra.bio/#workspaces/pathogen-genomic-surveillance/COVID-19_Broad_Viral_NGS

# T2T & HPRC on AnVIL

TELOMERE-TO-TELOMERE CONSORTIUM

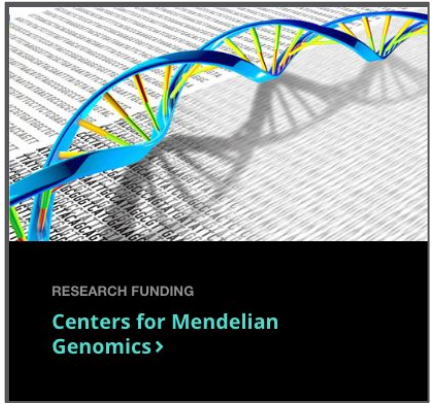
HUMAN PANGENOME



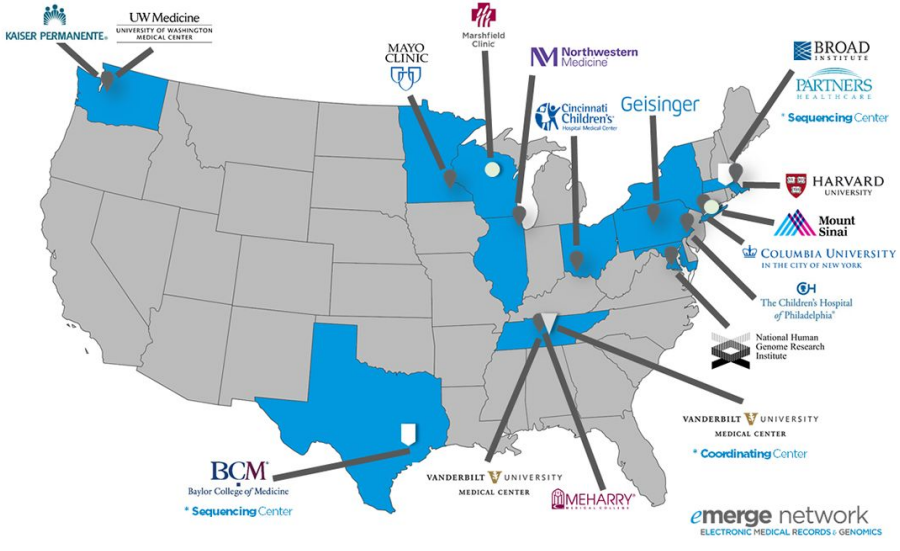



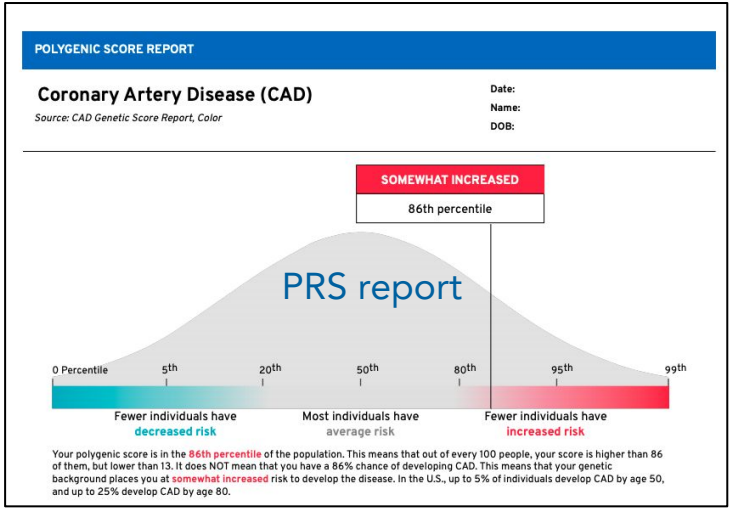

https://anvil.terra.bio/#workspaces/anvil-datastorage/AnVIL_T2T
https://anvil.terra.bio/#workspaces/anvil-datastorage/AnVIL_HPRC

# Clinical Engagements



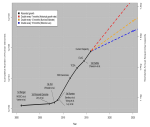


# Our Vision

AnVIL

## *Data*

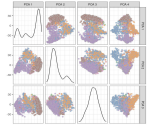


Data as a product to lower barriers to discovery

Support many (all!) NHGRI consortia

Pioneer new models of data use oversight and data governance

## *Computation*



Increased number of tools & analysis types

Interactive, batch, and visual analytics

Expand capabilities for predictive biology & medicine

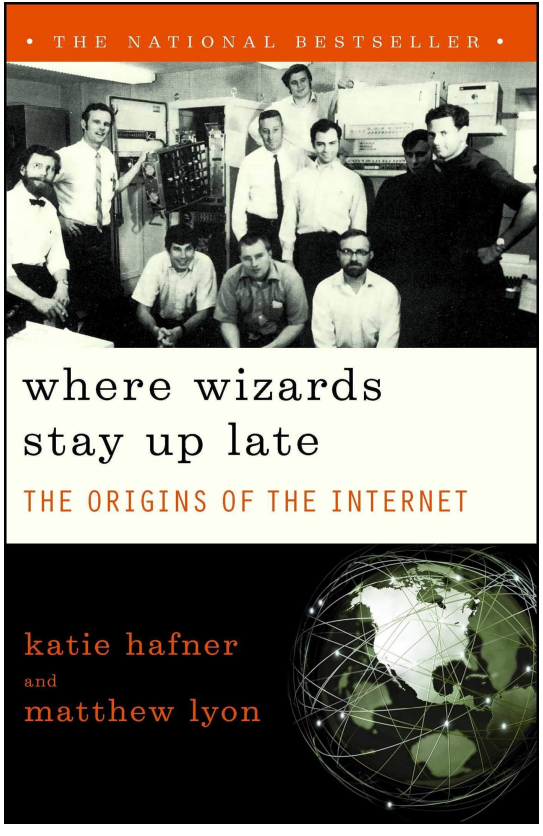
## *Users*

Increase number of users & consortiums

Empower users to do more on their own

Serve as the platform for cutting edge biomedical research

# Closing thought on interoperability




*We should not underestimate the importance of this effort...*

- If we are successful, we will catalyze the creation of an open and federated data ecosystem.
  - Others have done it before (SWIFT, the internet, the web).

- If we fail, we will degenerate into a collection of monolithic data silos
  - Others have done this before too (medical records in US hospitals)...

# Today's deeper dives

- Data submission and consortia engagement
- Analysis tools

Concurrent Breakouts -- Session 1

- Infrastructure
- Outreach and training

Concurrent Breakouts -- Session 2