Chan Zuckerberg Initiative

Meta: a research discovery tool for the biomedical sciences

Ben Nelson
Data Scientist (Product Analytics)
bnelson@chanzuckerberg.com

ATDS Meeting
AAS237

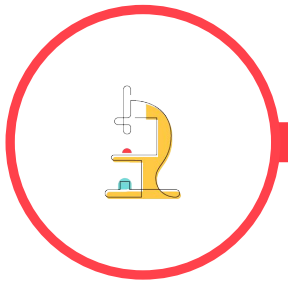
Core Initiatives

"CZI's mission is to build a more inclusive, just, and healthy future for everyone."

Science Initiative Mission

Supporting the science and technology that will make it possible to cure, prevent, or manage all diseases by the end of this century.

Science Goals



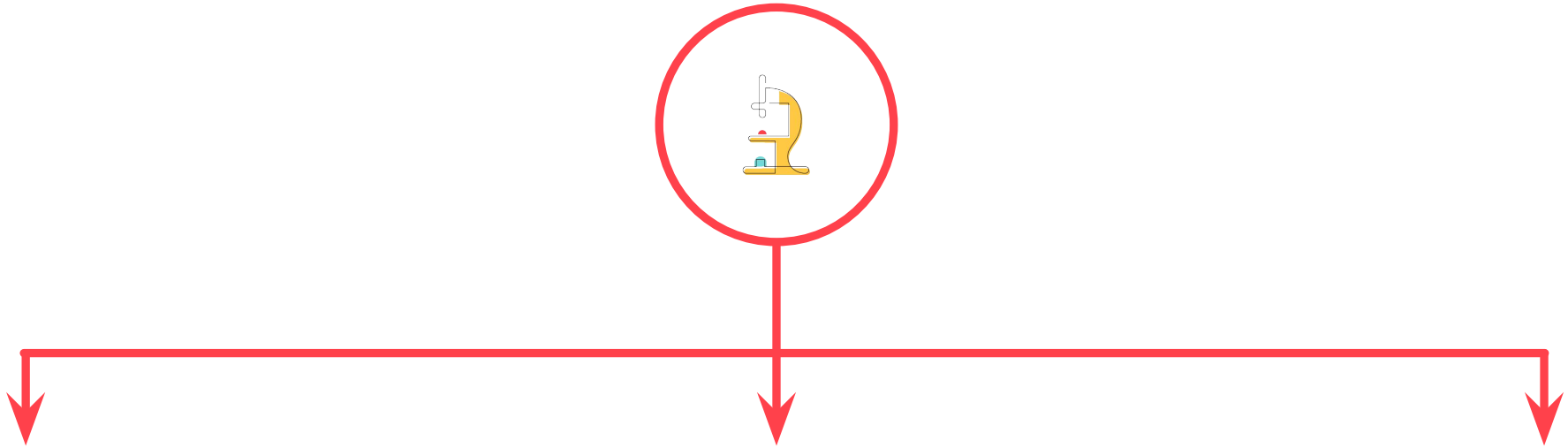
10 YEARS

Accelerate biomedical science with open, collaborative models of research.

80 YEARS

Cure, prevent, or manage all disease by the end of this century.

Science Team



Program

Supports external researchers through grants and collaborations

Technology

Builds software, computational tools, and data platforms for scientists

Policy

Engages patients and the public as partners in the scientific process

Technology: we build tools for science

cellxgene
An open source tool for exploring single-cell transcriptomics datasets.

IDseq
An open source software platform that helps scientists identify pathogens in metagenomic sequencing data.

napari
An interactive, multi-dimensional image viewer for Python.

starfish
An open source Python package that helps biologists create a scalable image processing pipeline for spatial transcriptomics data.

Just to recap

Program

Technology

Policy

...

SciTech Products

...

cellxgene

IDSEQ

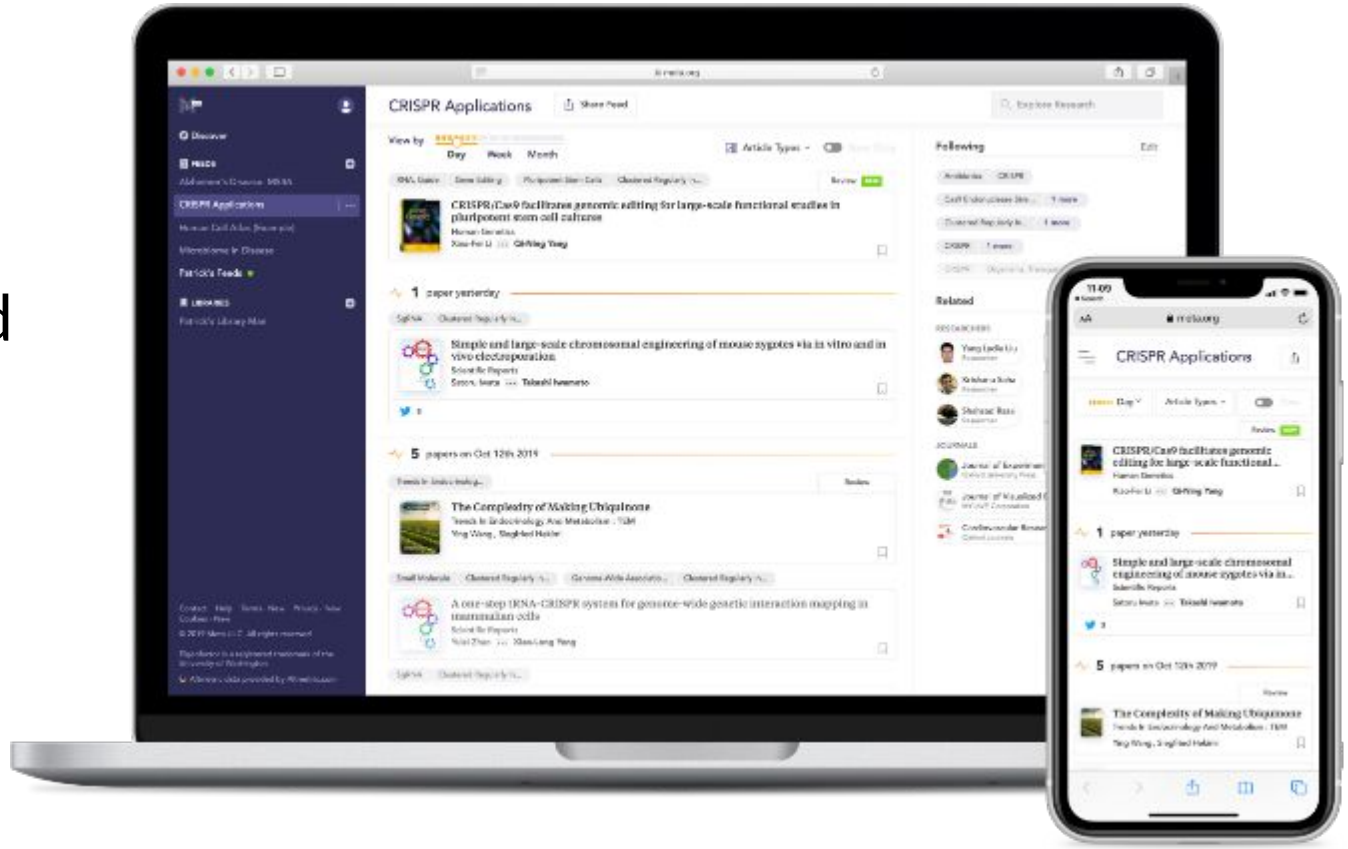
fish

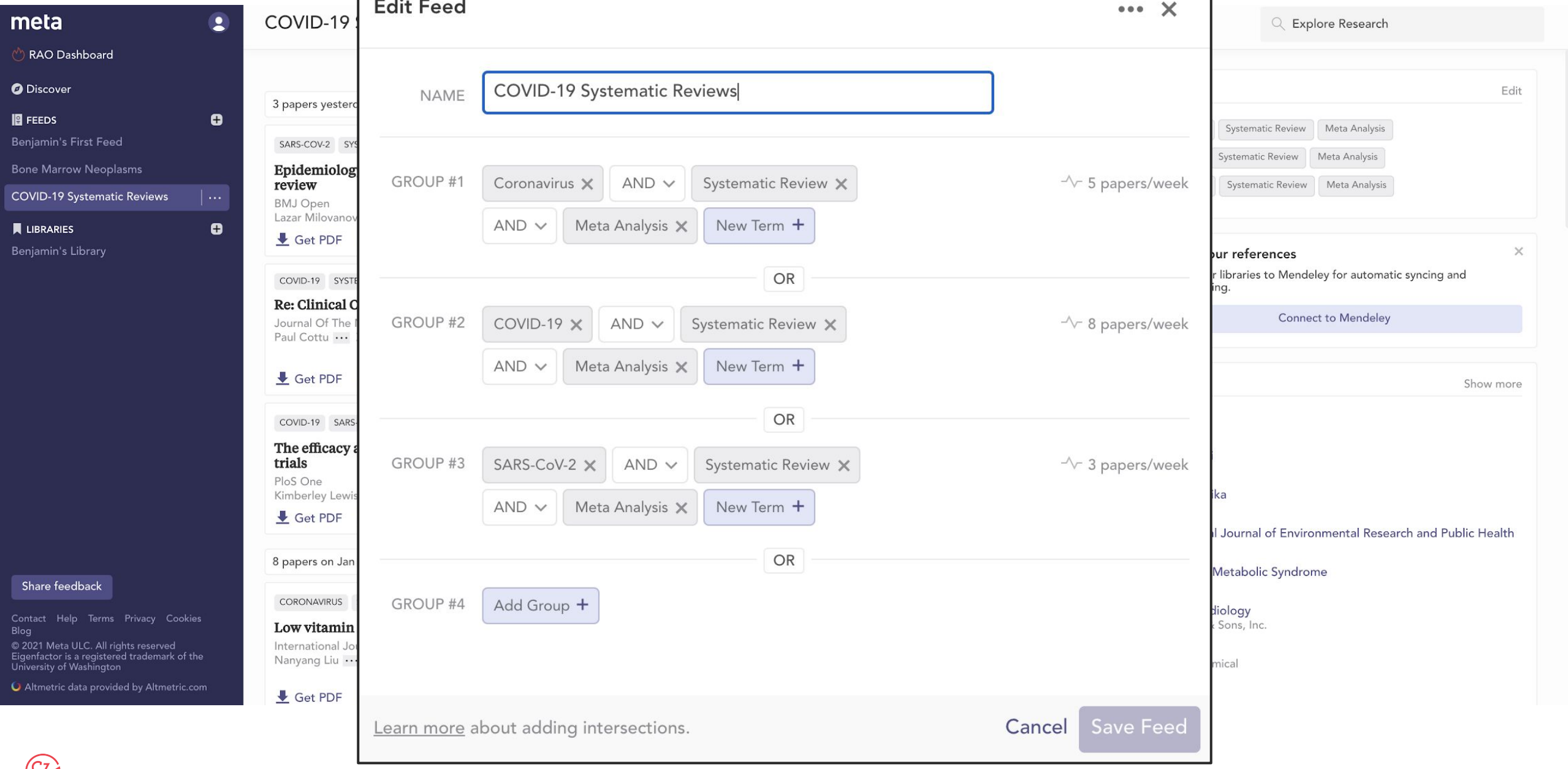
meta

I work here!

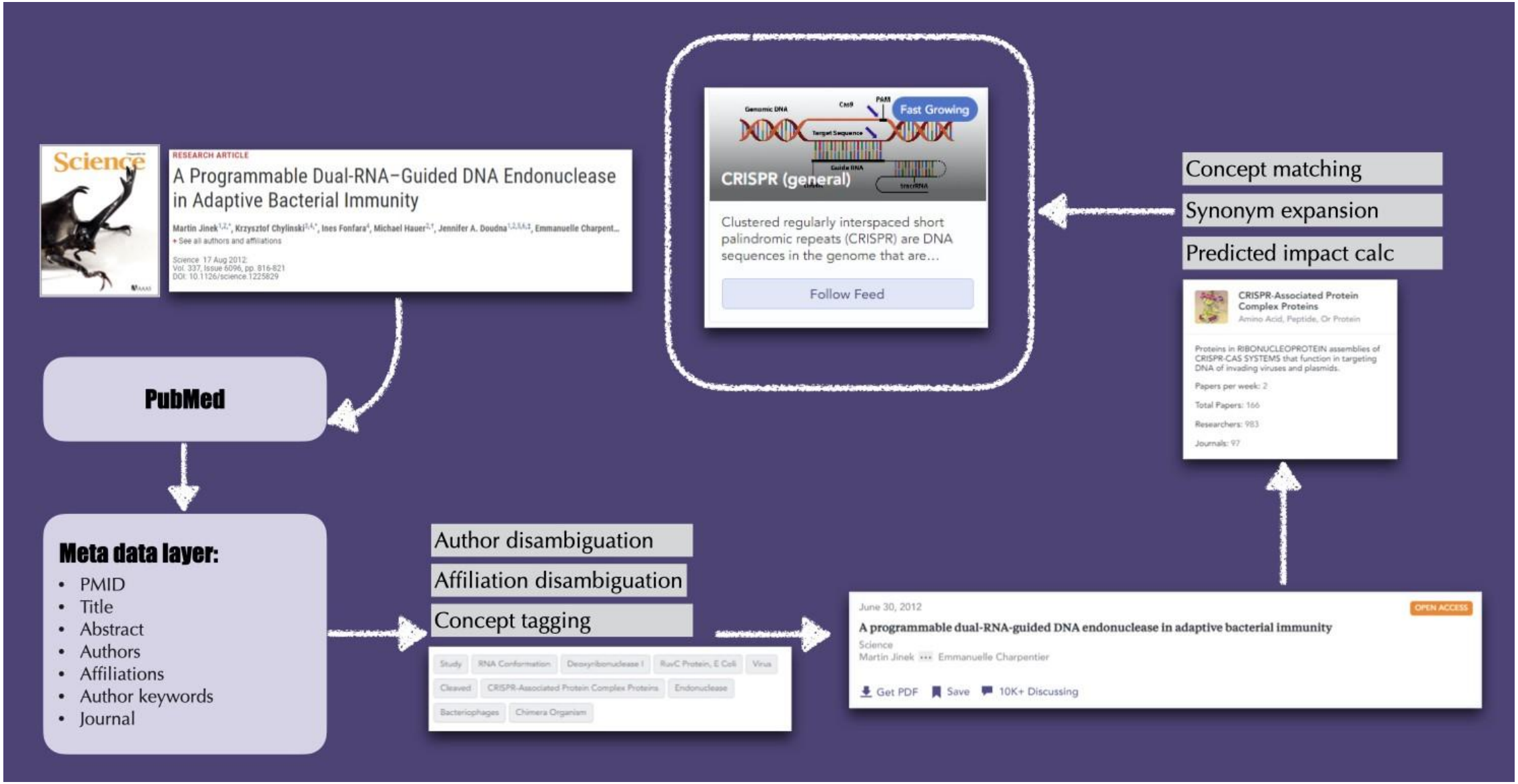
Meta

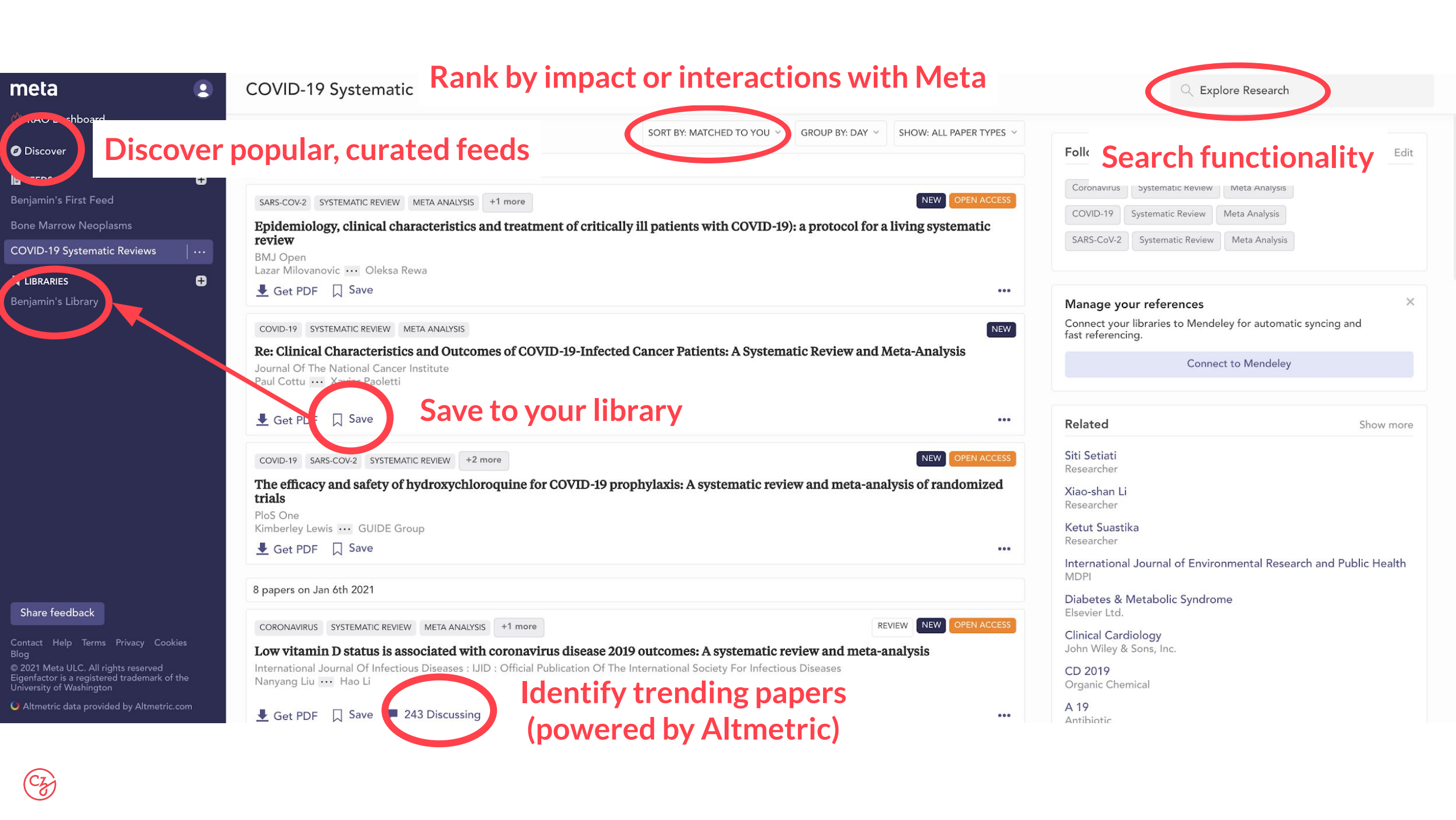
Meta provides a faster way to understand and explore science as it evolves — both at the level of broad research fields and at the level of an individual's specific research interests.




meta
RAO Dashboard
Discover
FEEDS
Benjamin's First Feed
Bone Marrow Neoplasms
COVID-19 Systematic Reviews
LIBRARIES
Benjamin's Library
Share feedback
Contact Help Terms Privacy Cookies
Blog
© 2021 Meta ULC. All rights reserved
Eigenfactor is a registered trademark of the University of Washington
Altmetric data provided by Altmetric.com
COVID-19
Explore Research
3 papers yesterd
SARS-COV-2
Epidemiolog
review
BMJ Open
Lazar Milovanov
Get PDF
COVID-19
Re: Clinical C
Journal Of The
Paul Cottu
Get PDF
COVID-19
The efficacy
trials
PloS One
Kimberley Lewis
Get PDF
8 papers on Jan
CORONAVIRUS
Low vitamin
International Jo
Nanyang Liu
Get PDF
Edit
Systematic Review
Meta Analysis
Systematic Review
Meta Analysis
Systematic Review
Meta Analysis
ur references
libraries to Mendeley for automatic syncing and
ing.
Connect to Mendeley
Show more
ka
l Journal of Environmental Research and Public Health
Metabolic Syndrome
diology
Sons, Inc.
mical
Edit Feed
NAME
COVID-19 Systematic Reviews
GROUP #1
Coronavirus
AND
Systematic Review
5 papers/week
AND
Meta Analysis
New Term
OR
GROUP #2
COVID-19
AND
Systematic Review
8 papers/week
AND
Meta Analysis
New Term
OR
GROUP #3
SARS-CoV-2
AND
Systematic Review
3 papers/week
AND
Meta Analysis
New Term
OR
GROUP #4
Add Group
Learn more about adding intersections.
Cancel
Save Feed

How Meta populates a feed




meta
COVID-19 Systematic
Rank by impact or interactions with Meta
Explore Research
Discover
Discover popular, curated feeds
SORT BY: MATCHED TO YOU
GROUP BY: DAY
SHOW: ALL PAPER TYPES
Search functionality
Benjamin's First Feed
Bone Marrow Neoplasms
COVID-19 Systematic Reviews
LIBRARIES
Benjamin's Library
SARS-COV-2 SYSTEMATIC REVIEW META ANALYSIS +1 more
NEW OPEN ACCESS
Epidemiology, clinical characteristics and treatment of critically ill patients with COVID-19): a protocol for a living systematic review
BMJ Open
Lazar Milovanovic ··· Oleksa Rewa
Get PDF Save
COVID-19 SYSTEMATIC REVIEW META ANALYSIS
NEW
Re: Clinical Characteristics and Outcomes of COVID-19-Infected Cancer Patients: A Systematic Review and Meta-Analysis
Journal Of The National Cancer Institute
Paul Cottu ···
Get PDF Save
Save to your library
COVID-19 SARS-COV-2 SYSTEMATIC REVIEW +2 more
NEW OPEN ACCESS
The efficacy and safety of hydroxychloroquine for COVID-19 prophylaxis: A systematic review and meta-analysis of randomized trials
PloS One
Kimberley Lewis ··· GUIDE Group
Get PDF Save
8 papers on Jan 6th 2021
CORONAVIRUS SYSTEMATIC REVIEW META ANALYSIS +1 more
REVIEW NEW OPEN ACCESS
Low vitamin D status is associated with coronavirus disease 2019 outcomes: A systematic review and meta-analysis
International Journal Of Infectious Diseases : IJID : Official Publication Of The International Society For Infectious Diseases
Nanyang Liu ··· Hao Li
Get PDF Save 243 Discussing
Identify trending papers (powered by Altmetric)
Share feedback
Contact Help Terms Privacy Cookies Blog
© 2021 Meta ULC. All rights reserved
Eigenfactor is a registered trademark of the University of Washington
Altmetric data provided by Altmetric.com
Edit
Coronavirus Systematic Review Meta Analysis
COVID-19 Systematic Review Meta Analysis
SARS-CoV-2 Systematic Review Meta Analysis
Manage your references
Connect your libraries to Mendeley for automatic syncing and fast referencing.
Connect to Mendeley
Related
Show more
Siti Setiati
Researcher
Xiao-shan Li
Researcher
Ketut Suastika
Researcher
International Journal of Environmental Research and Public Health
MDPI
Diabetes & Metabolic Syndrome
Elsevier Ltd.
Clinical Cardiology
John Wiley & Sons, Inc.
CD 2019
Organic Chemical
A 19
Antibiotic

Product Analytics

Are users getting value out of Meta?
What does that value look like?
How can we measure it?

Recall the 10 year goal: Accelerate biomedical science with open, collaborative models of research.

I think a lot about metrics

Metric design: What makes a good metric?

- It's comparative
  to previous time periods (cohorts), groups of users/objects, performance benchmarks, etc.
- It's understandable
  if you can remember it and discuss it, it's easier to turn a change in data to a change in culture
- It's probably a ratio
  easier to act on, inherently comparative, good for looking at opposing effects
- It changes the way you behave. It's actionable!
  "What will I do differently based on changes in the metric?"

Lean Analytics: How to Focus on What Matters - Ben Yoskovitz - MicroConf 2013
https://www.youtube.com/watch?v=iN5MTyqvjDY

Choosing the right metric

- Qualitative vs Quantitative
  Quantitative data answer "what" or "how much." Qualitative data answer "why."
- Vanity vs Actionable
  Vanity metrics (up and to the right) make you feel good. Actionable metrics change the organization's behavior.
- Exploratory (interesting) vs Reporting (managerial)
  Exploratory metrics are speculative and for finding unknown opportunities. Reporting metrics keep you informed of day-to-day operations.
- Leading vs Lagging
  Predict the future vs learn about the past

Example: measuring engagement with Meta?

How often do we expect users to be active?

daily? weekly? monthly?

average time users spend on Meta this week?

✗ doesn't tell us why they are sticking around

# of paper clicks this week?

✗ more/fewer clicks could be due to more/fewer users that week

# of paper clicks / weekly active user?

✗ not robust to "super users," e.g., a few users clicking on hundreds of papers

% of WAUs clicking on 1+ papers this week?

✓ tells us what fraction of users this week are finding relevant content

But getting accurate event counts is surprisingly difficult

Analytics Hierarchy of Needs

OPTIMIZE AND PREDICT

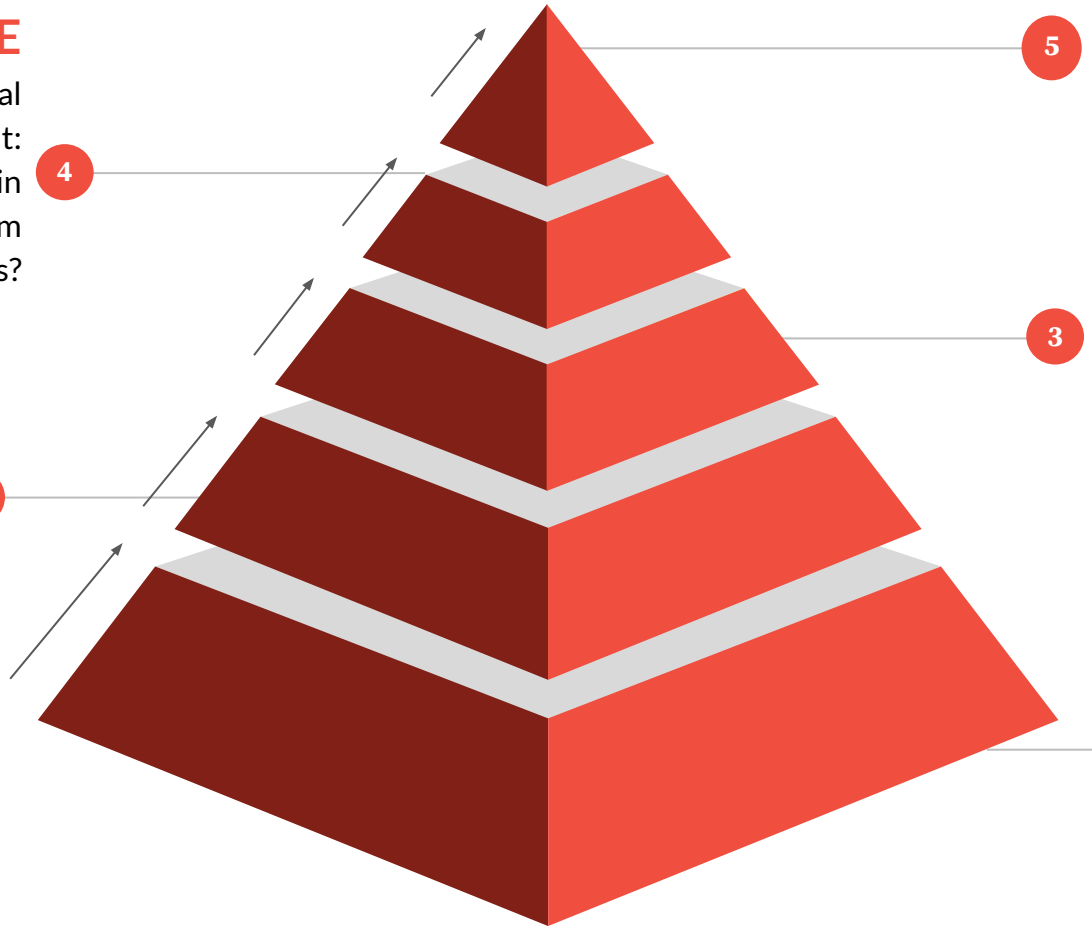
Once you have clean data, tracked across the user funnel, and analyzed rudimentary trend drivers: only then should you look to apply advanced techniques like machine learning. Also consider optimizing the top areas of opportunity in your funnel with A/B testing.

DEFINE AND TRACK

There's an old analytics adage: "what isn't tracked will probably go down." Map out your user journey (acquisition → conversion → usage → retention) and define metrics to track the drivers of your top-line goals (usually revenue and active users)

COLLECT

You can't manufacture data that you wish existed in hindsight! First ensure you have event logging and basic data modeling for key entities. This can be informed by your product intuition and key questions stakeholders have of your data.



ANALYZE

Evaluate your metrics against internal benchmarks, external benchmarks, and your gut: what are the biggest areas of opportunity in your product funnel? What can you learn from cohorting and segmenting your users?

CLEAN

Your engineering team will likely emit data in disparate, narrow tables. ETL your data into wider, standardized tables such that its easily analyzable and queryable. Consider user lifetime summary and user daily summary tables as a place to start.

https://towardsdatascience.com/the-analytics-hierarchy-of-needs-6d57d0e205e2

Summary

CZI Science is tackling a wide variety of problems in the biomedical sciences
    Imaging and Open Science teams might be interest to astronomers
    Current open roles at https://chanzuckerberg.com/careers/

Meta is a researcher discovery tool using ML to recommend papers and preprints.
    My role as a data scientist is in defining and measuring user value from our platform.

Good metric design is an important skill for data scientists (in product analytics or ML eng)
    Start with *Lean Analytics* by Croll & Yoskovitz

Thank you!

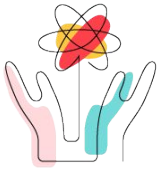
CZI-wide

https://twitter.com/ChanZuckerberg
https://www.facebook.com/chanzuckerberginitiative/
https://www.instagram.com/chanzuckerberginitiative
www.linkedin.com/company/chan-zuckerberg-initiative
https://www.youtube.com/channel/UCZioJ6fb9SuRdLIO7DlE09w
https://medium.com/czi-technology

CZI Science

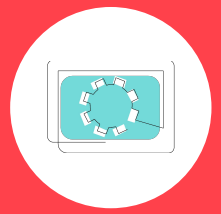
https://twitter.com/cziscience
https://medium.com/@cziscience

Supplemental Slides

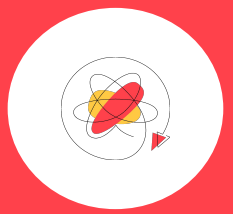


What does success look like?

Productivity
Publications, preprints, software, datasets, protocols, resources



Reach
Deposition in public repositories, requests and re-use citations, clinical applications, commercial development



Collaborative contributions
Leadership, co-authorship, success of students and postdocs, acknowledgments

Choosing Projects

We look for cross-cutting themes across groups of diseases — rather than focusing on single diseases.

Unmet need

Where are the most urgent needs for new advances?

Opportunity

What new advances create opportunities for progress, and can CZI have a systemic effect?

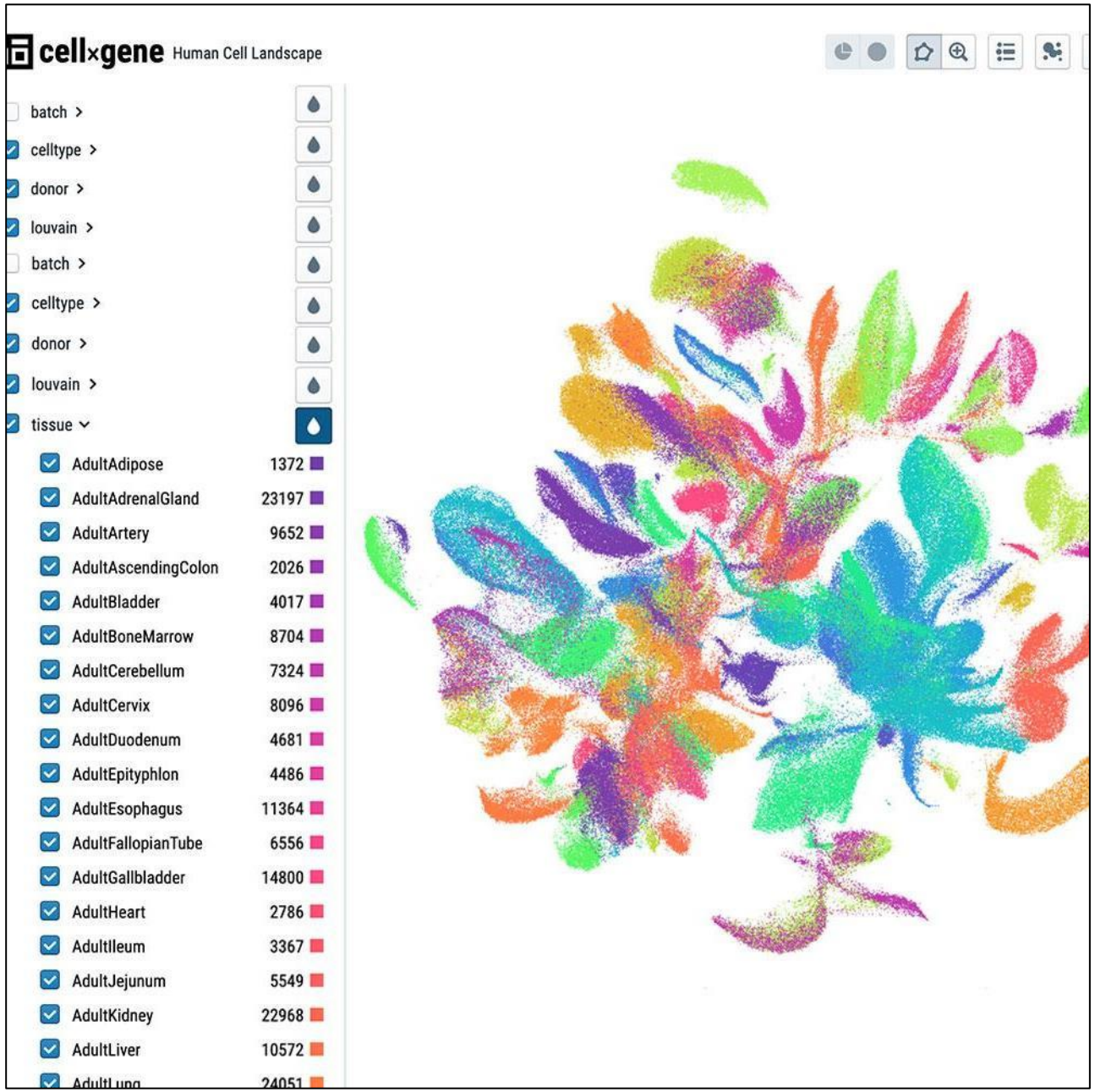
Differentiated Impact

Is CZI uniquely positioned to make an impact?

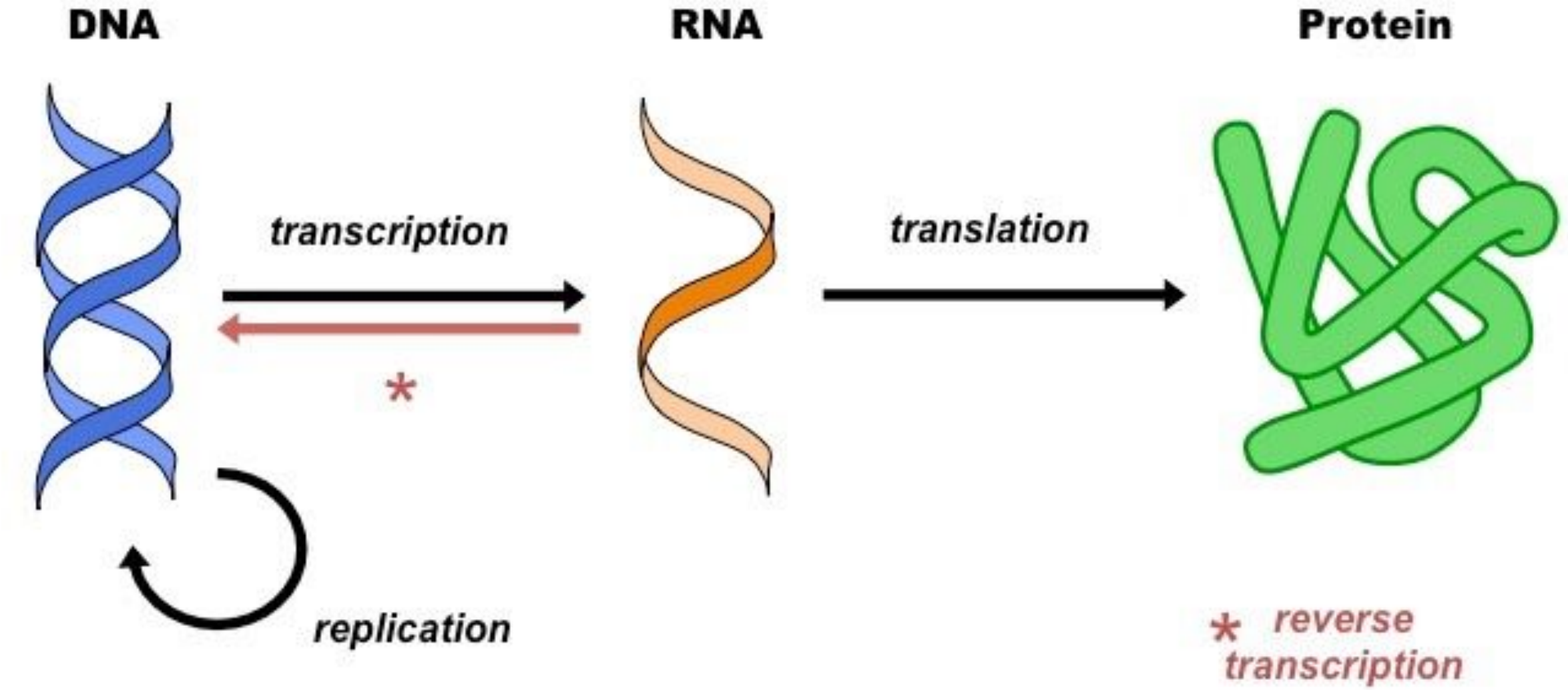
cellxgene

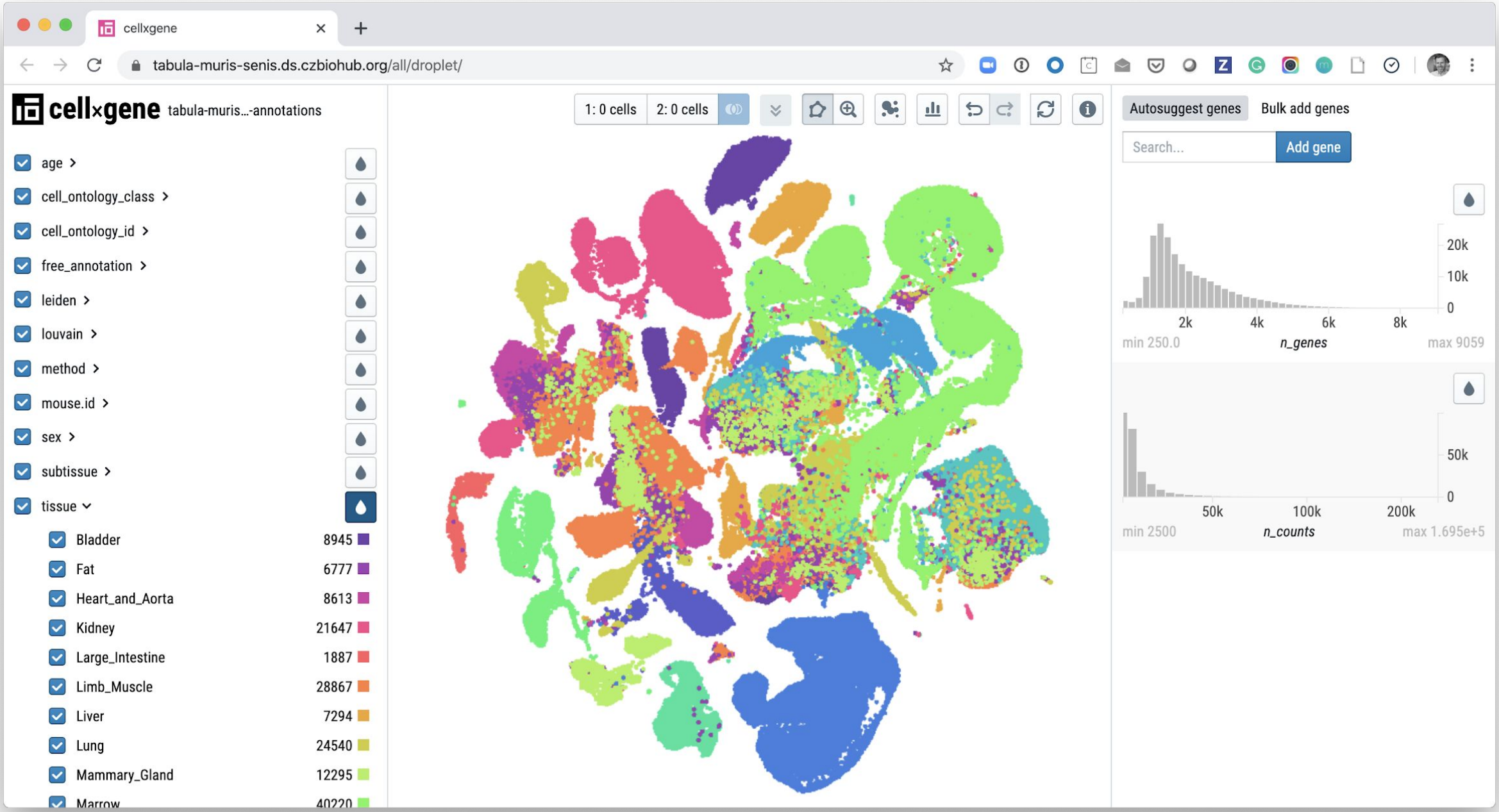
cellxgene is an open source tool for exploring single-cell transcriptomics datasets — including those from the Human Cell Atlas.

cellxgene



What is image-based transcriptomics?



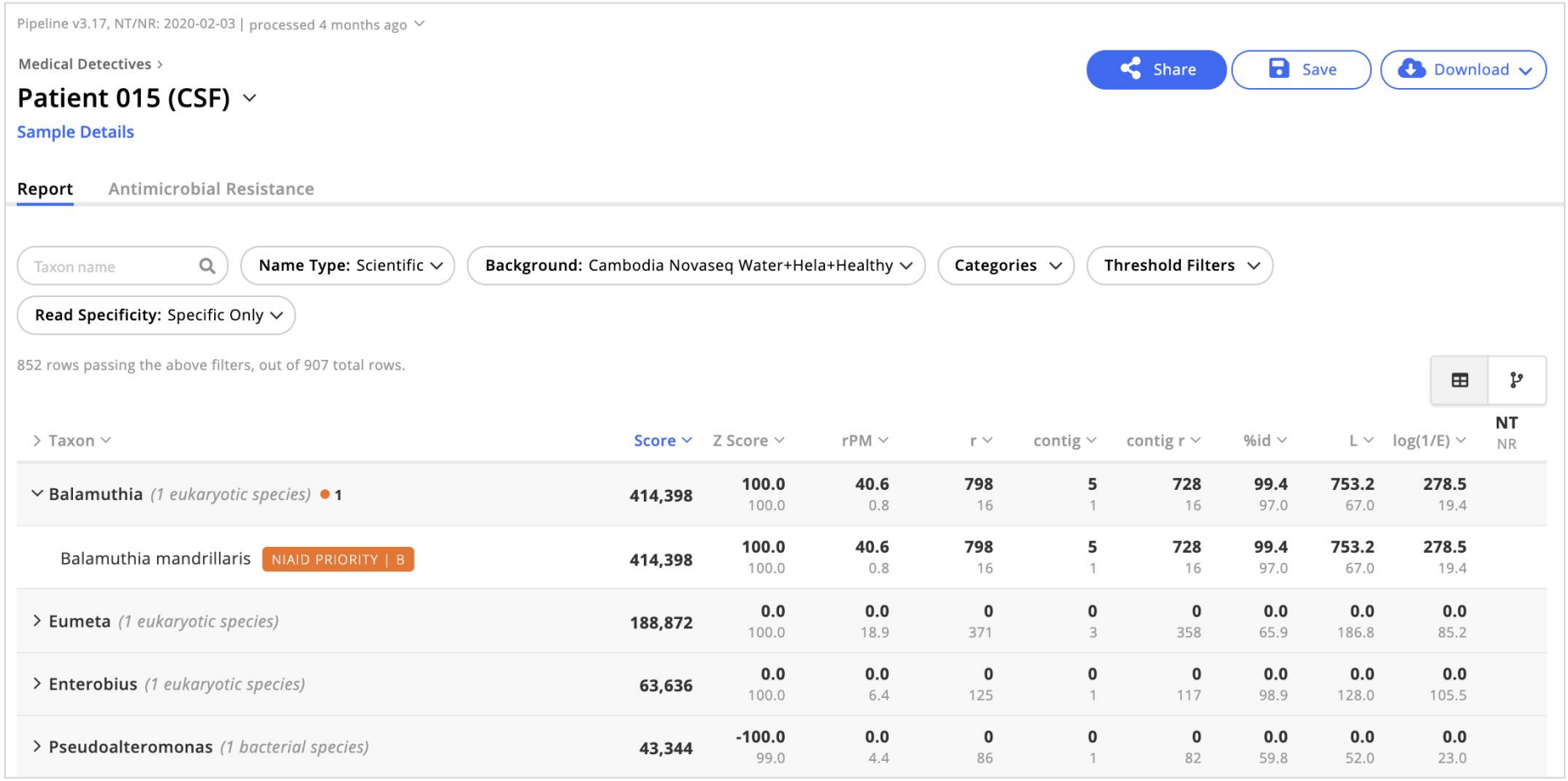

cellxgene
tabula-muris-senis.ds.czbiohub.org/all/droplet/
cellxgene tabula-muris...-annotations
1: 0 cells
2: 0 cells
Autosuggest genes
Bulk add genes
Search...
Add gene
age
cell_ontology_class
cell_ontology_id
free_annotation
leiden
louvain
method
mouse.id
sex
subtissue
tissue
Bladder 8945
Fat 6777
Heart_and_Aorta 8613
Kidney 21647
Large_Intestine 1887
Limb_Muscle 28867
Liver 7294
Lung 24540
Mammary_Gland 12295
Marrow 40220
20k
10k
0
2k
4k
6k
8k
min 250.0
n_genes
max 9059
50k
0
50k
100k
200k
min 2500
n_counts
max 1.695e+5

IDseq

IDseq is an open source software platform that helps scientists worldwide identify pathogens in metagenomic sequencing data.

IDSEQ

IDseq uses metagenomics to enable scientists to rapidly determine what microbes are present in a particular biological sample



Taxon	Score	Z Score	rPM	r	contig	contig r	%id	L	log(1/E)	NT NR
Balamuthia *(1 eukaryotic species)* ● 1	**414,398**	**100.0** 100.0	**40.6** 0.8	**798** 16	**5** 1	**728** 16	**99.4** 97.0	**753.2** 67.0	**278.5** 19.4	
Balamuthia mandrillaris NIAID PRIORITY | B	**414,398**	**100.0** 100.0	**40.6** 0.8	**798** 16	**5** 1	**728** 16	**99.4** 97.0	**753.2** 67.0	**278.5** 19.4	
Eumeta *(1 eukaryotic species)*	**188,872**	**0.0** 100.0	**0.0** 18.9	**0** 371	**0** 3	**0** 358	**0.0** 65.9	**0.0** 186.8	**0.0** 85.2	
Enterobius *(1 eukaryotic species)*	**63,636**	**0.0** 100.0	**0.0** 6.4	**0** 125	**0** 1	**0** 117	**0.0** 98.9	**0.0** 128.0	**0.0** 105.5	
Pseudoalteromonas *(1 bacterial species)*	**43,344**	**-100.0** 99.0	**0.0** 4.4	**0** 86	**0** 1	**0** 82	**0.0** 59.8	**0.0** 52.0	**0.0** 23.0	

Data from: Wilson, *et al.* Ann Neurol. 2015.

Why IDseq?

Standardize workflows and partner with established institutions

Fresh Data Sources

Streamlined and standardized metagenomic sequencing analysis using the most recent versions of publicly available datasets and tools.

Accessible and Free

Available online from anywhere. The computational costs necessary to run samples through the metagenomics pipeline are covered.

Open Source

Facilitates discovery by promoting open science and providing global insights across multiple datasets and projects—accelerating scientific research.

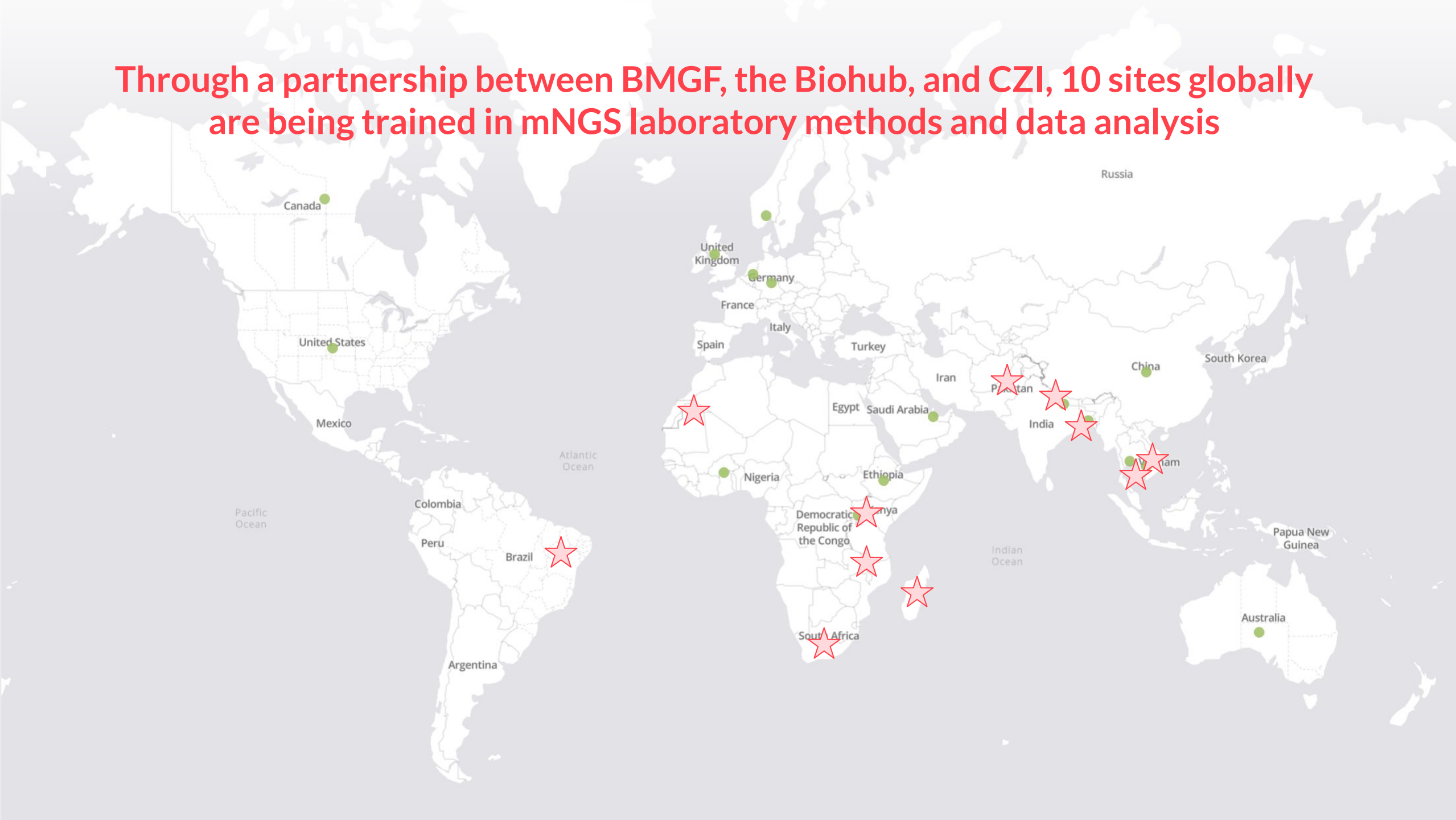
Safe and Secure

Samples are able to remain in country while the data is uploaded to the cloud. IDseq provides a secure environment for data storage that complies with research requirements. Because it's on the cloud all data is backed up and accessible from anywhere.

Current IDseq Theories of Impact

80-Year Plan	**Cure, manage, or prevent all disease** Currently: Infectious disease = 20% of human deaths worldwide	
20-Year Plan	Eliminate unknowns about the identity, origin, and spread of infection	
3-Year Plan	**Theory 1: Build global capacity for pathogen detection** Provide (exceptional) informatics, training, and compute by enabling scientists, who otherwise wouldn't have the capacity, to do metagenomics to track and eliminate the threat of infectious disease	**Theory 2: Build a compounding data asset for epidemiology** Assemble a global, public data repository of pathogenic agents that facilitate scientific collaboration, real-time surveillance, and rapid response

Through a partnership between BMGF, the Biohub, and CZI, 10 sites globally are being trained in mNGS laboratory methods and data analysis



Examples of Impact to Date

2019 Dengue outbreak in Bangladesh

Senjuti sequenced the first complete genome and added it to public databases

Partnership with Bill & Melinda Gates Foundation

Deploying IDseq to 10 sites: Cambodia, South Africa, Madagascar, Nepal, Malawi, Kenya (finished training); Brazil, Pakistan, The Gambia, Vietnam (to be trained)

First COVID-19 case in Cambodia

Characterized by local Gates-supported team using IDseq in February 2020

Shared on public.idseq.net and in WIRED

First COVID-19 sequenced in Bangladesh

COVID-19 in San Francisco Bay Area

Biohub is using IDseq to characterize viral mutations and co-infections

Much usage and interest from institutions around the world

Links Related to IDseq

Spotlights the work of collaborators including those in Bangladesh studying etiology of pediatric meningitis: https://www.discoveridseq.com/

Public site spotlights work done by GCE Cambodia at the start of the COVID-19 outbreak: https://public.idseq.net/

http://idseq.net/

Biorxiv Preprint outlines the technical details of the IDseq Pipeline: https://www.biorxiv.org/content/10.1101/2020.04.07.030551v3

Help Center contains documentation: http://help.idseq.net/

Github Repos are Open Source: https://github.com/chanzuckerberg/idseq-dag, https://github.com/chanzuckerberg/idseq-web

Patient 015

74 year old woman living in San Francisco.

Initially presented with altered mental state and **fever.** Diagnosed with a UTI and given azithromycin on discharge. Presented 3 days later with rapid **vision loss** in the left eye, no external wound. Asked to return to SF General 2 days later for follow up. At SF General, arrived **comatose.** MRI revealed **destruction throughout all territories of the brain.** Treated empirically with antibiotics, anti-parasitics, anti-fungals, with no improvement. Brain biopsy revealed necrotizing vasculitis. **All sent out diagnostics, cultures, and microscopic examinations of tissue were negative for pathogens.**

Patient 015 (CSF)

Sample Details

Report | Antimicrobial Resistance

Taxon name | Name Type: Scientific | Background: NID Human CSF v3 | Categories | Th

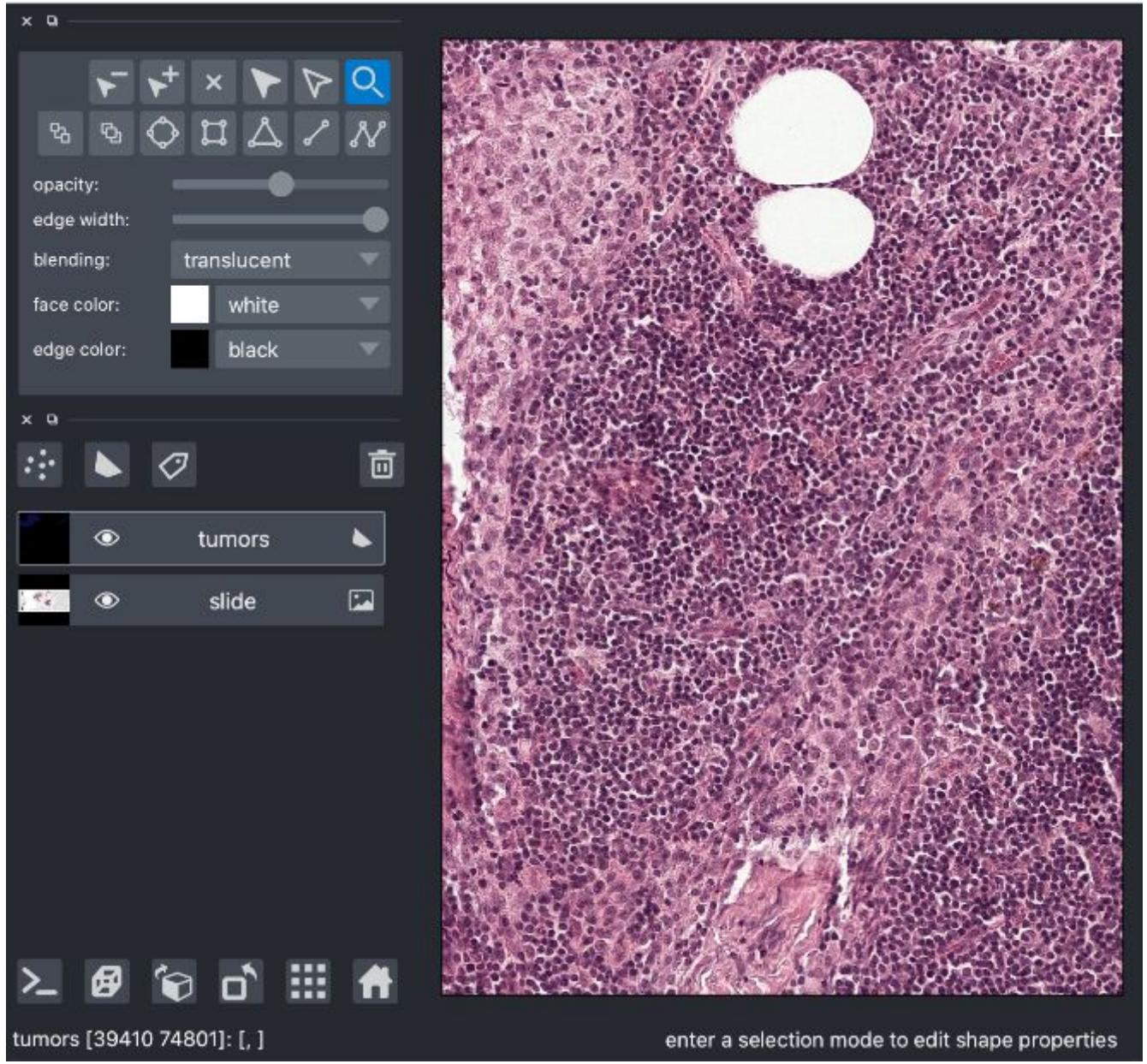
1141 rows passing the above filters, out of 1141 total rows.

Taxon	Score	Z	rPM	r
Balamuthia *(1 eukaryotic species)* 1	406,144	99.0 99.0	40.6 0.8	798 16
Balamuthia mandrillaris NIAID PRIORITY | B	406,144	99.0 99.0	40.6 0.8	798 16

Brain-eating amoeba

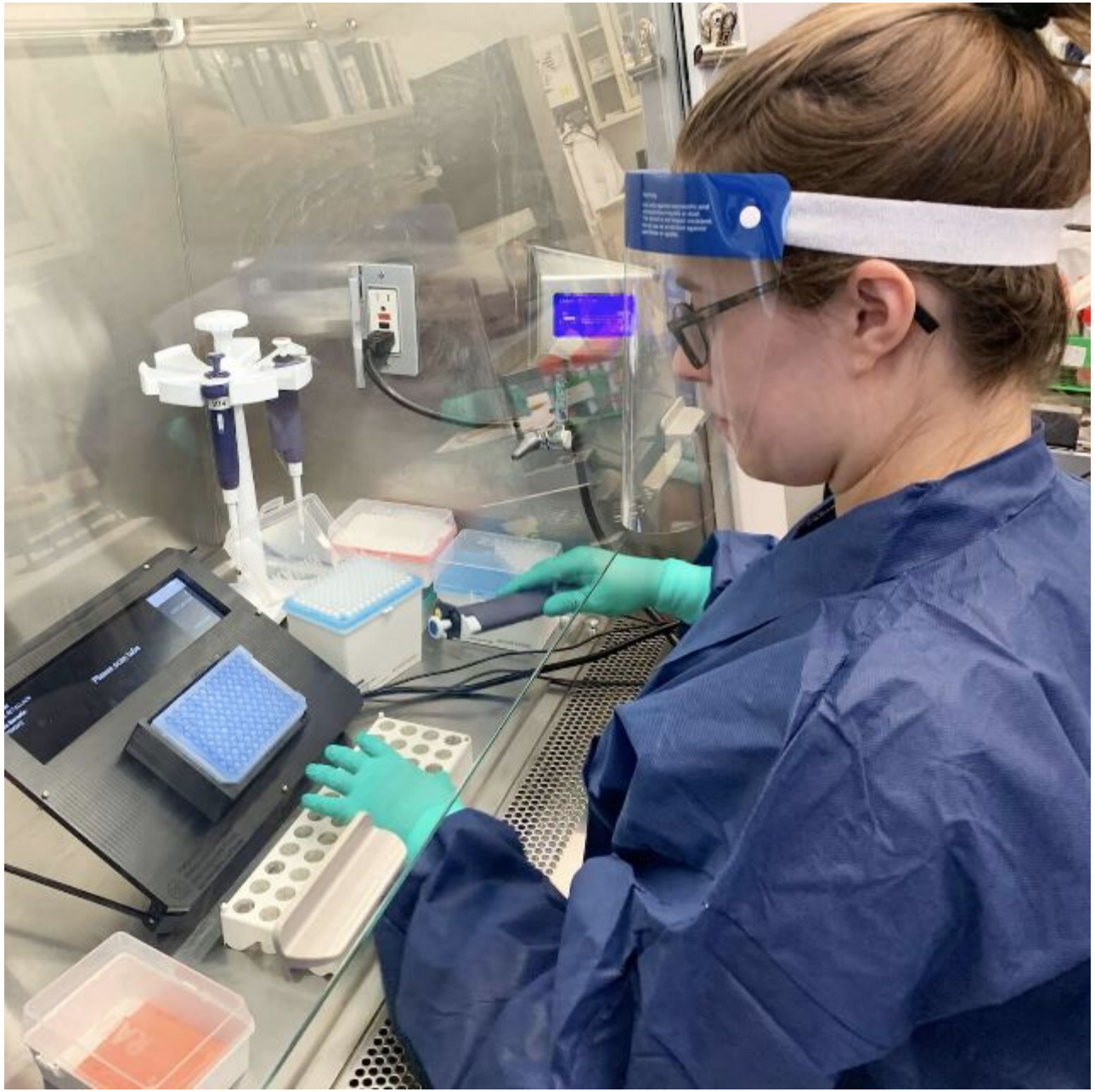
napari

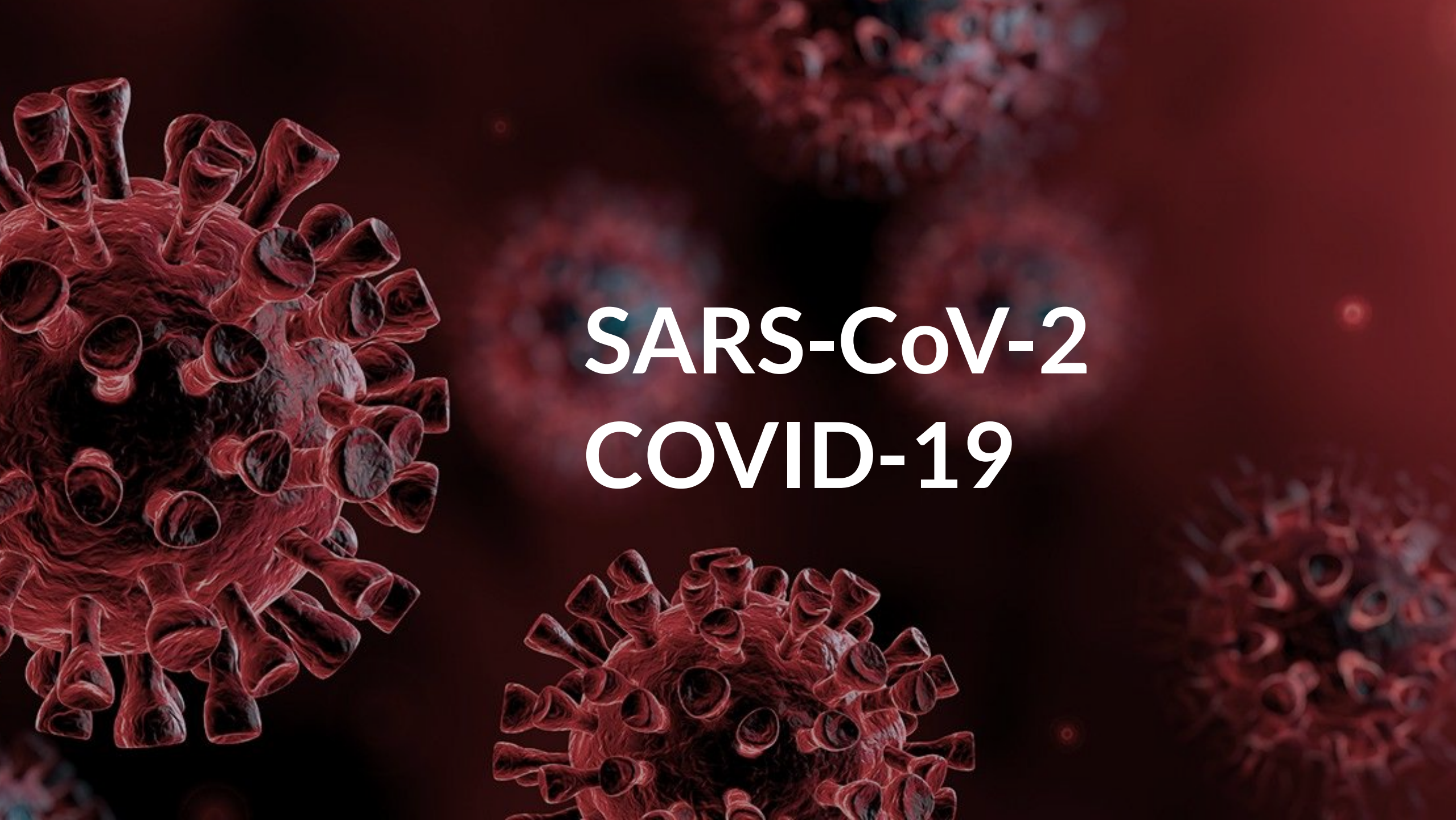
napari is an interactive, multi-dimensional image viewer for Python. It's designed for browsing, annotating, and analyzing large, multi-dimensional images.



COVID-19 Response

We are leveraging open science, technology, and collaboration to accelerate our shared understanding of COVID-19 by increasing access to testing, genomic sequencing, research, community support, and more.




SARS-CoV-2
COVID-19

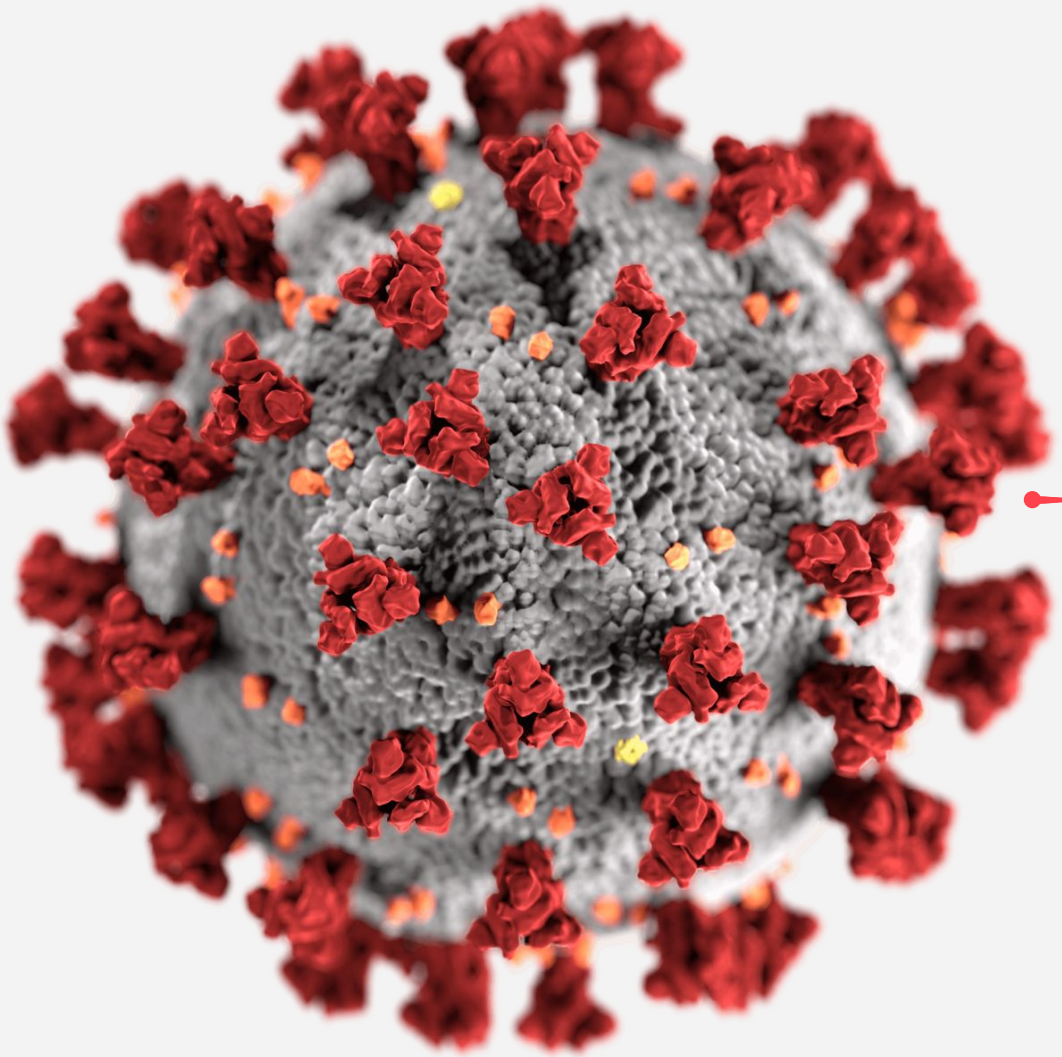
Science COVID-19 Response Work

Guiding Principles

Discrete projects with near-term impact (6 to 12 months), spend funds by July

Leverage our strengths and existing programs where possible (e.g. Single-Cell)

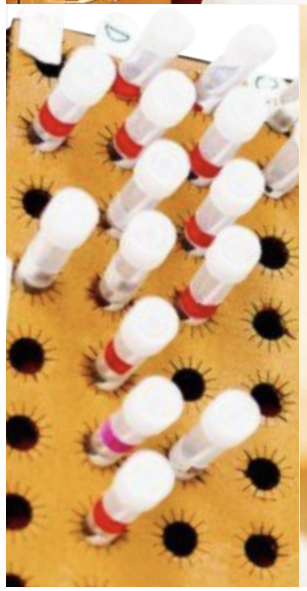
After the immediate crisis is over, then focus on longer-term strategy



- Diagnostics
- Treatments
- Patients
- Research
- Data and Tools

Chan Zuckerberg Biohub

- CLIAHub
- California COVID Tracker
- California Pandemic Consortium

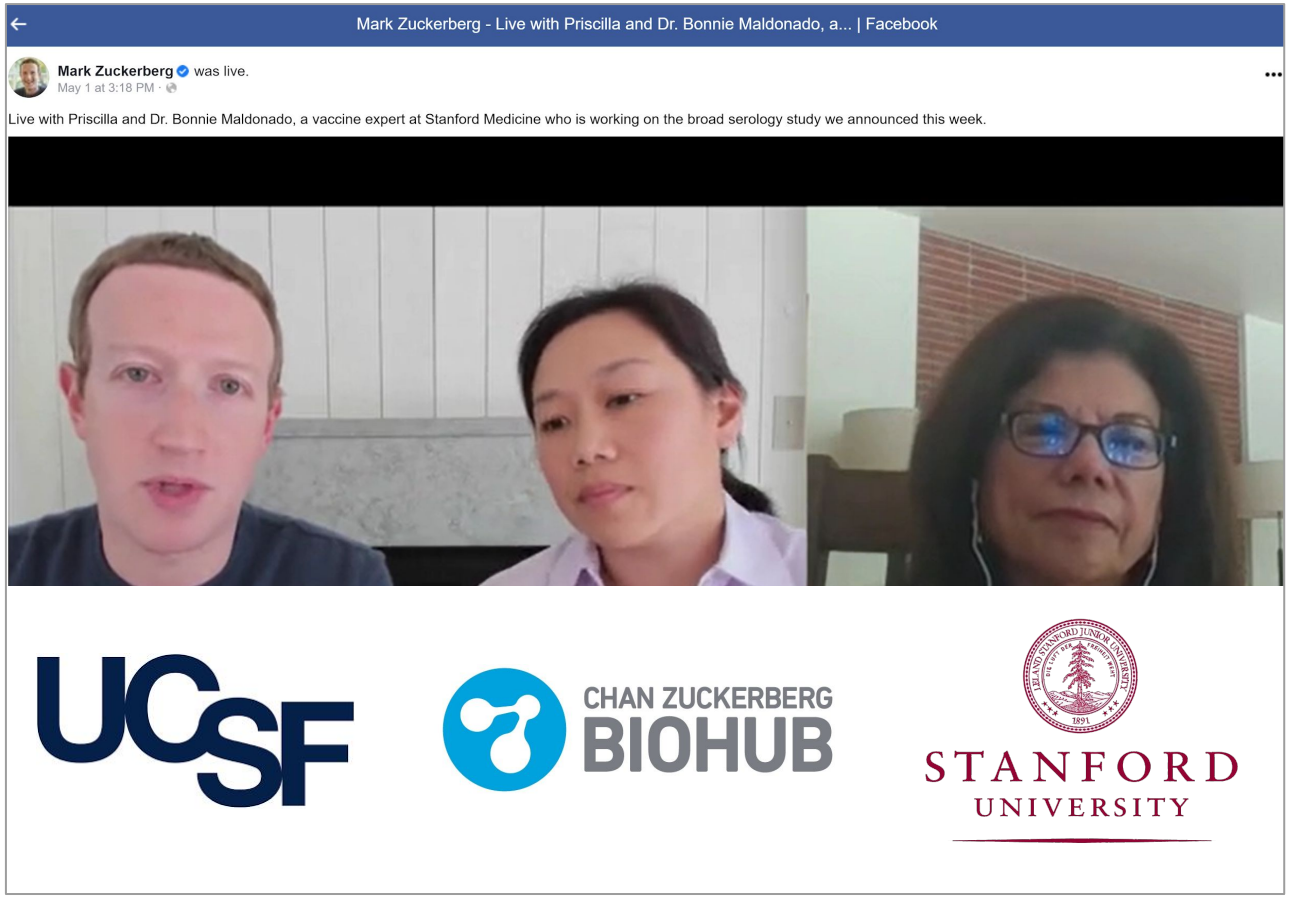


California Pandemic Consortium

- UCSF, Stanford, CZ Biohub
- PCR and antibody testing
- Two-studies

Community subjects
What is the prevalence in representative population in the Bay Area?

Health care workers
What is the rate at which healthcare workers acquire COVID-19 with or without symptoms?



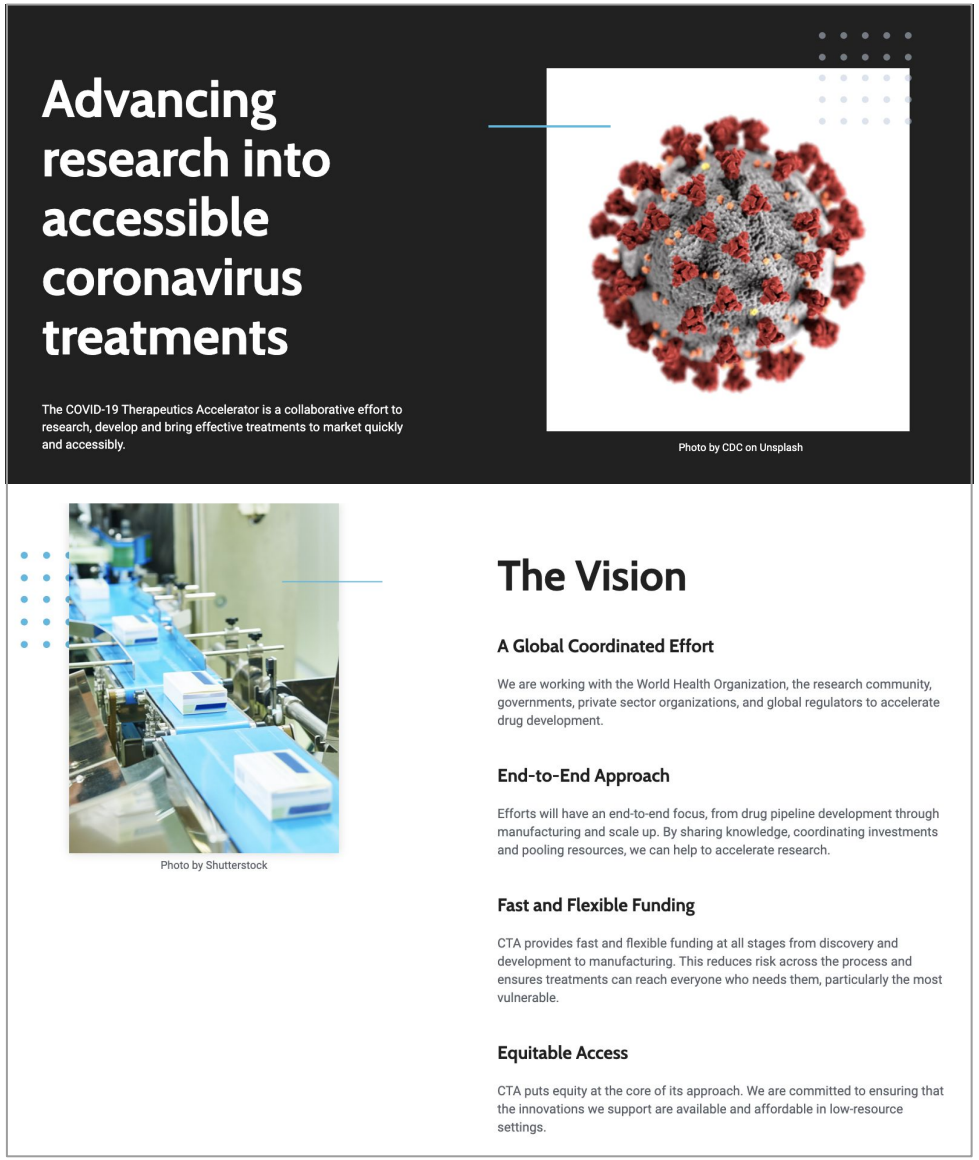
COVID-19 Therapeutics Accelerator

Do any existing drugs work for COVID-19?

Can we test drugs that may be effective?

Can we develop new drugs?

Values: Open data and open access and access in low-income and vulnerable populations and countries



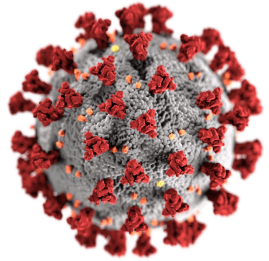
Open Science: COVID-19 Open Research Dataset

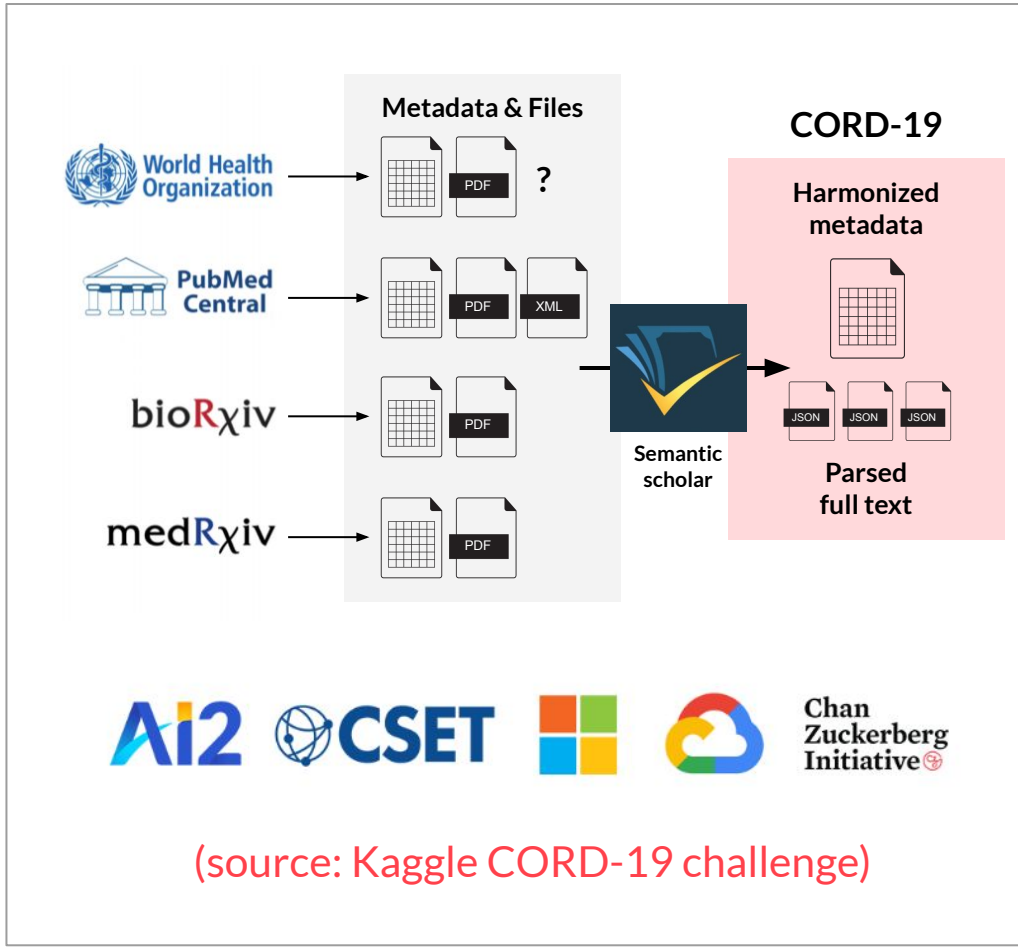
CORD-19

A openly-licensed, machine-readable corpus of full-text research papers and preprints on COVID-19, SARS-CoV-2 and the related coronavirus family

Released in collaboration with AllenAI, Georgetown CSET, Google, and Microsoft in response to White House OSTP call-targets AI/ML community

- 2 million views
- 90,000 downloads
- 1,500+ notebooks contributed
- Daily data releases





(source: Kaggle CORD-19 challenge)

Setting a **Target**

What's a good value and what's a bad value?

Depends on what stage in product lifecycle you're in

Having a target can help with **decision making**

For example:

*User value measured by **user retention compared to competitor**.*

A. *If user retention < competitor → Focus on increasing value to user.*
B. *If user retention > competitor → Focus on increasing user base.*