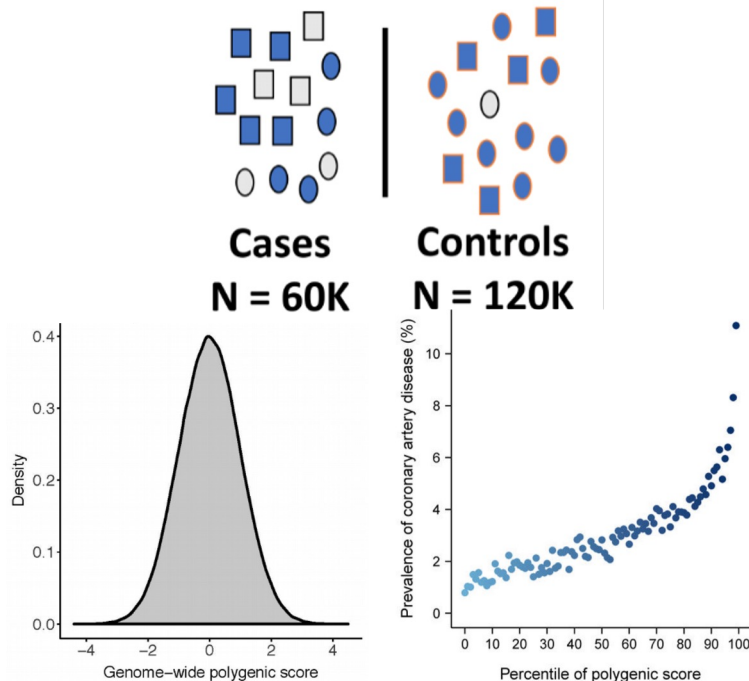


Why care about scientific computing?

Vignette 1: Polygenic Risk Scores & Patient Selection

Khera et al, Nature Genetics, 2018



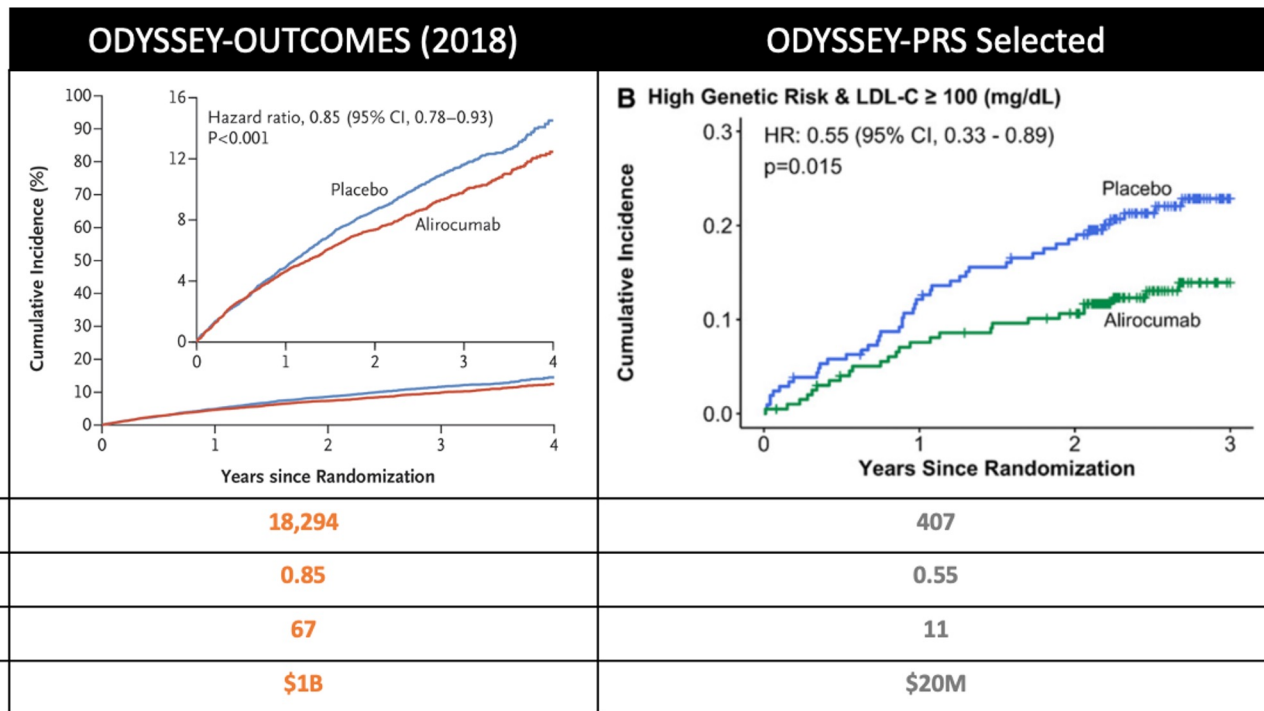
Coronary Artery Disease

	Remainder of population	Top 5% of polygenic score
Family history	35%	44%
Hypertension	28%	32%
Type 2 diabetes	2%	2.7%
Hypercholesterolemia	13%	20%
Current smoking	9.2%	9.5%
Body mass index	27.3	27.7
Systolic blood pressure	140	141

Some traditional risk factors are slightly elevated, but not enough to be useful

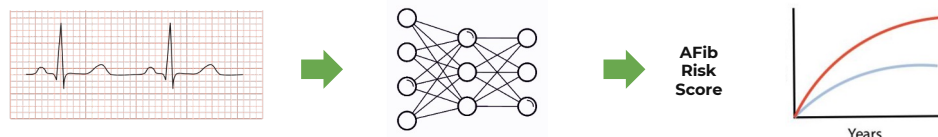
Not a subcluster!

Vignette 1: Polygenic Risk Scores & Patient Selection

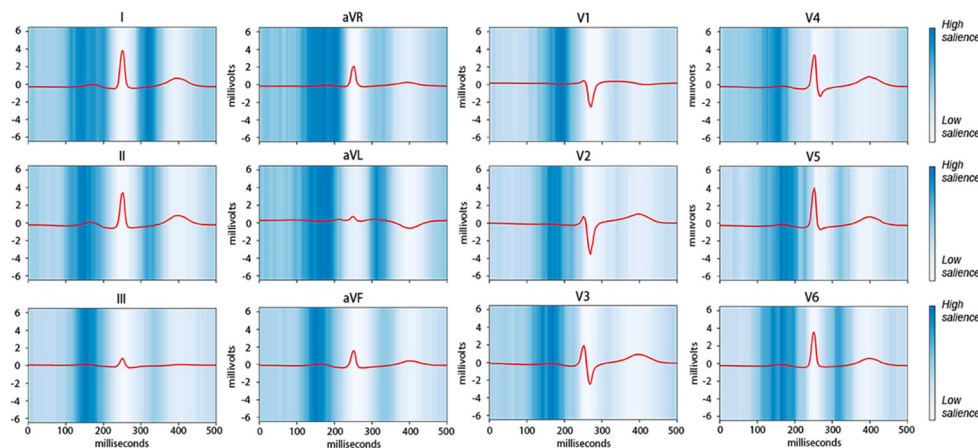


Damask et al, Circulation, 2020

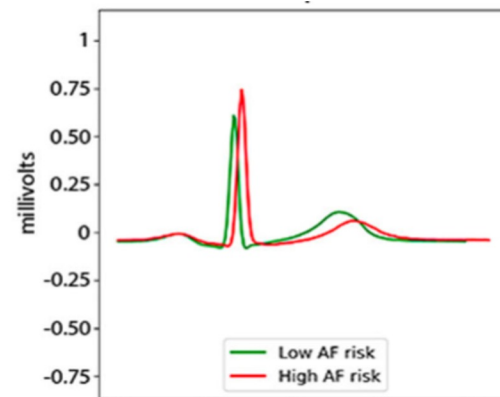
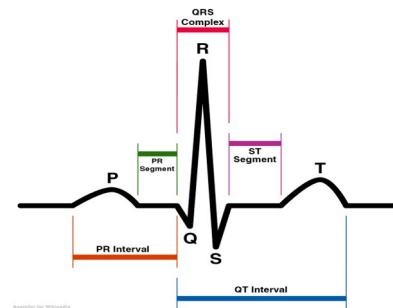
Vignette 2: Risk Prediction of Incident Afib



1-D deep convolutional neural network with **survival curve loss** to predict time to AF.

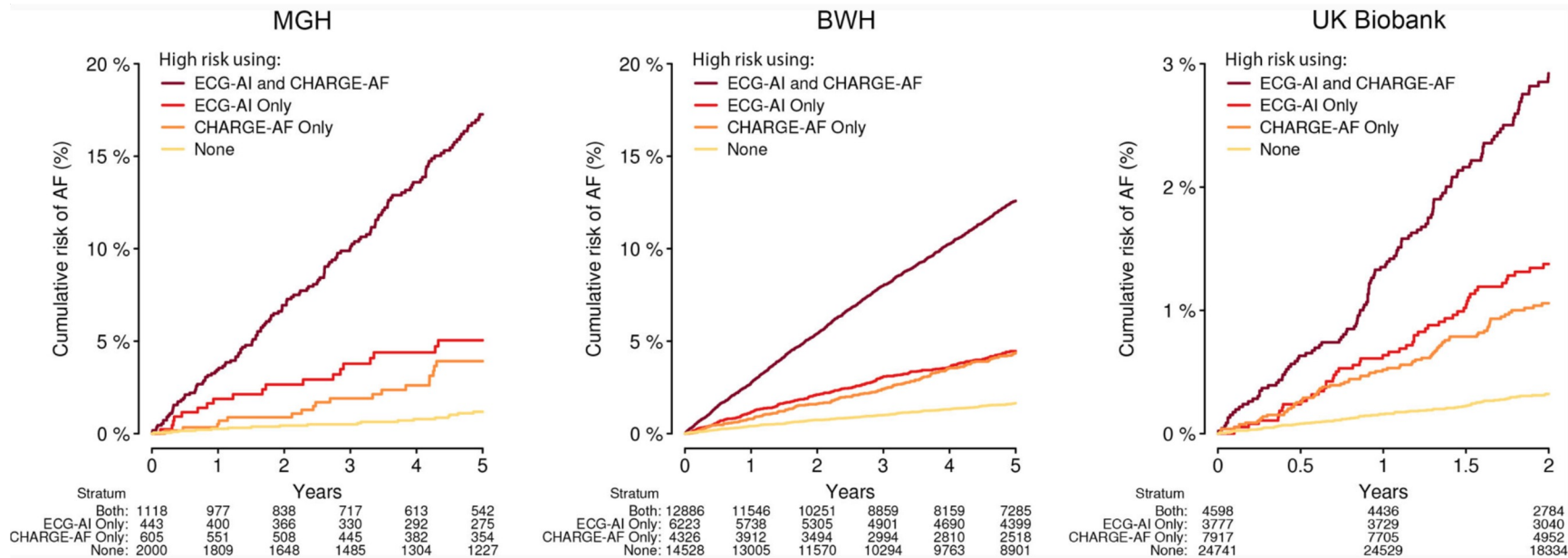


P wave and surrounding regions had the greatest effect on ECG-AI AF risk



Individuals with high estimated AF risk have a longer P wave duration and slightly wider QRS and a flatter ST segment

Vignette 2: Risk Prediction of Incident Afib



Magical and sparkly!

Standalone analytic
software *projects*

Not software *products*
that scale, create
leverage for others, or
accelerate other work



Collective intelligence in culture

~450TB including
all files in
Wikimedia
Commons

~1GB per year of
novel compressed
text

Wikipedia:Size of Wikipedia

🌐 3 languages ▾

[Project page](#) [Talk](#)

[Read](#) [Edit](#) [View history](#) [Tools](#) ▾

From Wikipedia, the free encyclopedia

For the editing guideline on article size, see [Wikipedia:Article size](#).

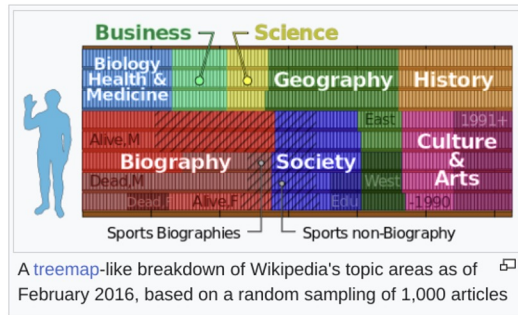
The **size of the English Wikipedia** can be measured in terms of the number of articles, number of words, number of pages, and the size of the database, among other ways. As of 17 May 2024, there are 6,824,497 articles in the [English Wikipedia](#) containing over 4.5 billion words (giving an average of about 668 words per article). The total number of pages is 60,678,100. Articles make up 11.25 percent of all pages on Wikipedia. As of 2 July 2023, the size of the current version of all articles compressed is about 22.14 GB without media.^{[1][2]}

Wikipedia continues to grow, and the number of articles on Wikipedia is increasing by about 14,000 a month (as of January 2024). The number of articles added to Wikipedia every month reached its peak in 2006, at over 50,000 new articles a month, and has been slowly but steadily declining since then. While this might seem to show that Wikipedia's growth is slowing or stopping, it should be noted that the amount of text added to Wikipedia articles every year has been constant since 2006, at roughly 1 gigabyte of

[Shortcuts](#)

[WP:SIZEWP](#)

[WP:WPSIZE](#)



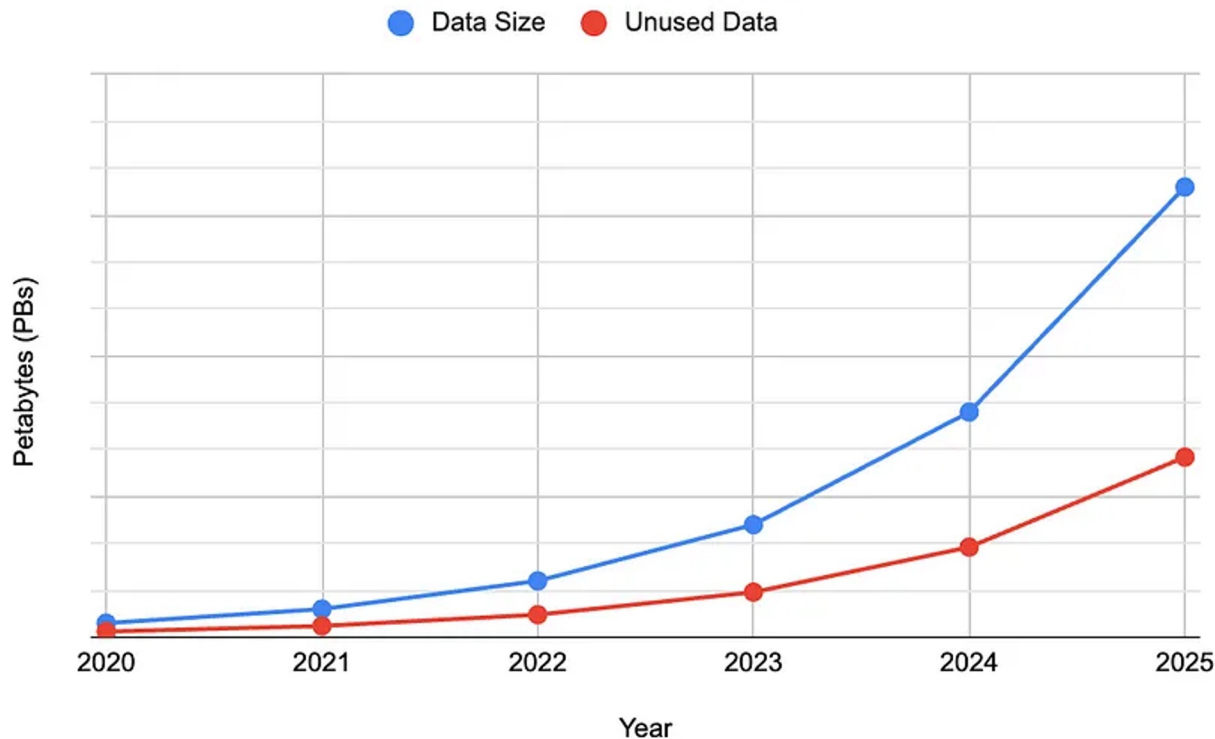
Collective intelligence in culture

NETFLIX

2 petabytes / week

50% increase in storage
cost YoY

40% of data goes unused



<https://netflixtechblog.medium.com/navigating-the-netflix-data-deluge-the-imperative-of-effective-data-management-e39af70f81f7>

We lack the collective intelligence in science that we have in culture

Blockers: data withholding, format drift, biological complexity, sharing knowledge as papers behind paywalls, you name it

Multiple social movements and massive investments made to address these! Some successful!

Figure 1



Parachutes reduce the risk of injury after gravitational challenge, but their effectiveness has not been proved with randomised controlled trials

Credit: HULTON/GETTY

(these are not the blockers)



HIPAA

43. H88: 2B41

The Belmont Report

Ethical Principles
and Guidelines for
the Protection of
Human Subjects
of Research

The National Commission
for the Protection of Human Subjects
of Biomedical and Behavioral
Research

NTSU LIB. DEPOSITORY

My own opinion on collective intelligence blockers:

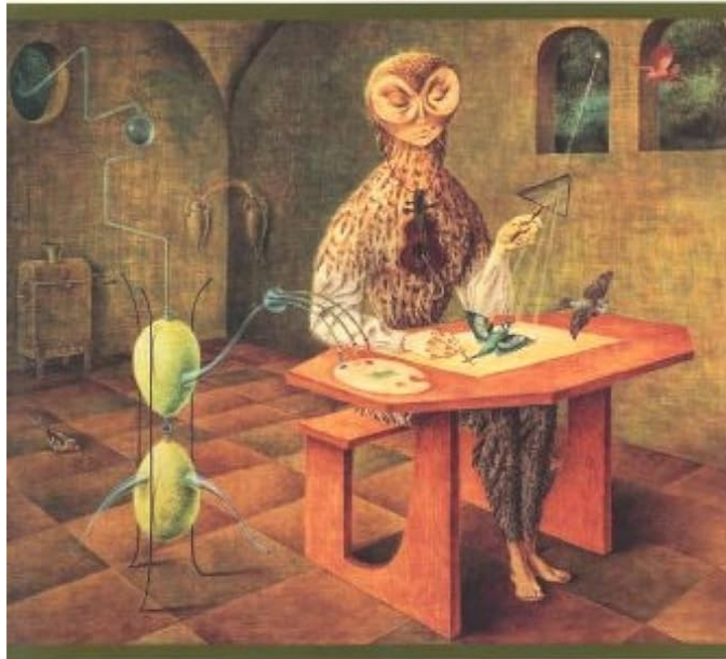
1. It's hard to make (scaled, platform) software that expands the universe of biomedical software beyond the unicorns - because we don't meet scientists where they live / have an expansive idea of "scientist"
2. When we actually do meet scientists where they live, the political economy of biomedical software restricts longitudinal knowledge accumulation in many (most?) large organizations



Copyrighted Material

A SMALL MATTER OF PROGRAMMING

PERSPECTIVES ON END USER COMPUTING



BONNIE A. NARDI

Copyrighted Material



Sound familiar?

“Many of these people work on tasks that rapidly vary on a yearly, monthly, or even daily basis. Consequently, their software needs are diverse, complex, and frequently changing. Professional software developers cannot directly meet all of these needs because of their limited domain knowledge and because their development processes are too slow.”

“End-user development (EUD) helps to solve this problem. EUD is "a set of methods, techniques and tools that allow users of software systems, who are acting as non-professional software developers, at some point to create, modify, or extend a software artifact" (Lieberman et al 2006)”

Home

Search

Your Library

Playlists Albums

Liked Songs
Playlist • 116 songs

Psych & Groove
Playlist • Spotify

Neo-Psychedelic Rock
Playlist • Spotify

Discover Weekly
Playlist • Spotify

Graceland
Album • Paul Simon

Unissued Session Copenhagen 1977
Album • Stan Getz

Swamp Gold, Vol. 1

Partition
Bill Callahan



Download the free app



All Music Podcasts Audiobooks

Discover Weekly

Liked Songs

Bossa Nova Classics

Soul Mix

Psych & Groove

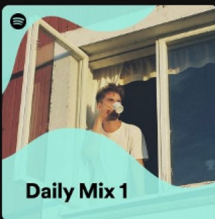
lofi beats

Something More Than Free

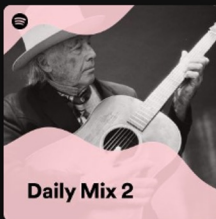
Jon Batiste

Made For John Wilbanks

Show all



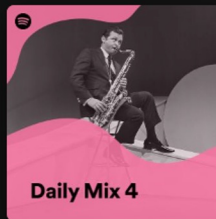
Daily Mix 1



Daily Mix 2



Daily Mix 3



Daily Mix 4



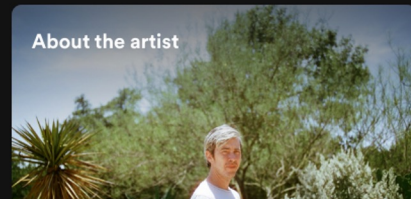
Psych & Groove



Partition

Bill Callahan

About the artist



Hyperbolic Paraboloids

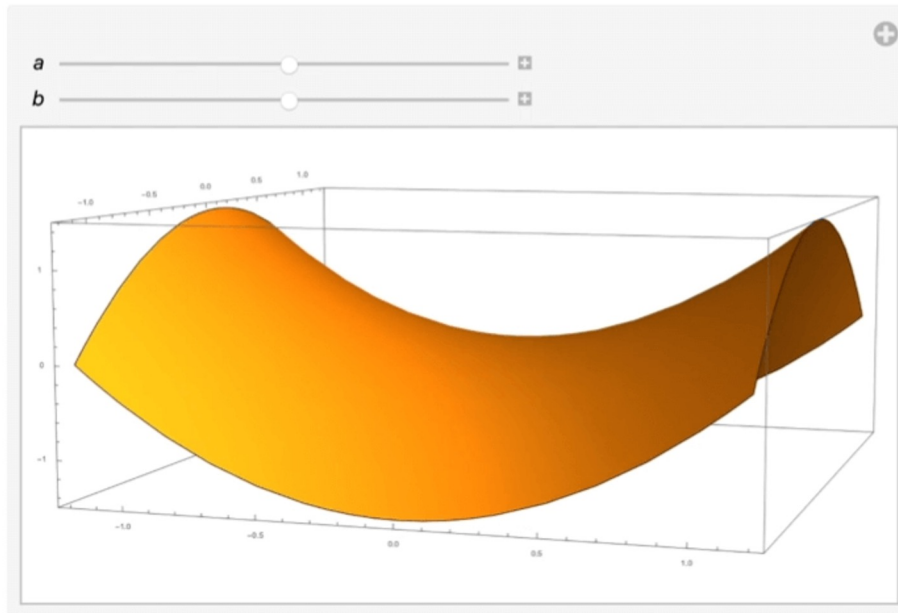
A hyperbolic paraboloid is a conical surface that has the following base form:

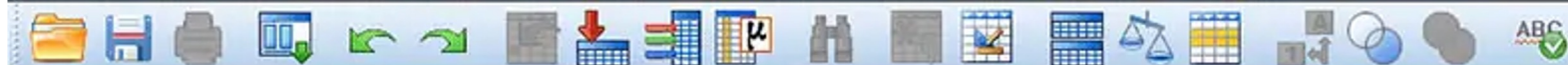
$$z = a x^2 - b y^2$$

Drag the sliders to see how coefficients a and b affect the saddle-like shape:

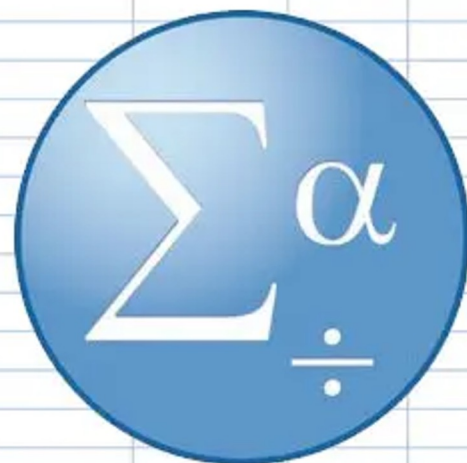
 plot $ax^2 - by^2$ varying a, b 

Out[1]=





	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	InterviewID	Numeric	8	0	Interview ID	None	None	15	Right	Nominal
2	Name	String	8	0	Name	None	None	8	Left	Nominal
3	Gender	Numeric	8	0	Gender	{1, Male}...	None	8	Right	Nominal
4	Age	Numeric	8	2	Age	None	None	8	Right	Scale
5	Rice	Numeric	8	2		None	None	8	Right	Unknown
6										
7										
8										
9										
10										
11										
12										
13										
14										
15										
16										
17										
18										
19										
20										



GLMSELECT1.sas x PSMATCH8.sas x TTEST05.sas x

CODE LOG OUTPUT DATA

```

99 proc psmatch data=drugs region=cs;
100   class Drug Gender;
101   psmodel Drug(Treated='Drug_X')= Gender Age BMI;
102   match method=optimal(k=1) exact=Gender distance=lps caliper=0.25;
103   output out(obs=match)=Outgs lps=_Lps matchid=_MatchID;
104 run;

```

C:\Gordon\Code_Examples\PSMATCH8.sas UTF-8

RESULTS

Table of Contents

Propensity Score Information											
Observations	Treated (Drug = Drug_X)					Control (Drug = Drug_A)					Treated - Control
	N	Mean	Standard Deviation	Minimum	Maximum	N	Mean	Standard Deviation	Minimum	Maximum	Mean Difference
All	113	0.3108	0.1325	0.0602	0.6411	373	0.2088	0.1320	0.0202	0.6858	0.1020
Region	113	0.3108	0.1325	0.0602	0.6411	361	0.2176	0.1267	0.0510	0.6824	0.0932
Matched	113	0.3108	0.1325	0.0602	0.6411	113	0.3082	0.1310	0.0619	0.6824	0.0025

Matching Information	
Distance Metric	Logit of Propensity Score
Method	Optimal Fixed Ratio Matching
Control:Treated Ratio	1
Caliper (Logit PS)	0.191862
Matched Sets	113
Matched Obs (Treated)	113
Matched Obs (Control)	113



“The earliest spreadsheets were ghastly by today’s user interface design standards (obscure command names, completely text-based, etc) but they were immediate successes with ordinary end users who recognized in them the high-level support for their own problem-solving tasks.”

C11 (L) TOTAL					C1
					25
	A	B	C	D	
1	ITEM	NO.	UNIT	COST	
2	----	----	----	----	
3	MUCK RAKE	43	12.95	556.85	
4	BUZZ CUT	15	6.75	101.25	
5	TOE TONER	250	49.95	12487.50	
6	EYE SNUFF	2	4.95	9.90	
7				----	
8			SUBTOTAL	13155.50	
9			9.75% TAX	1282.66	
10				----	
11			TOTAL	14438.16	
12					
13					
14					
15					
16					
17					
18					
19					
20					

Lotus 1-2-3 wins with
macros and graphics

Success comes from
*giving the user the
ability to do serious
computing*

Not from treating
them like novices who
can't be trusted with
power tools!

A:A1: 'EMP' MENU

	Worksheet	Range	Copy	Move	File	Print	Graph	Data	System	Quit
	Global	Insert	Delete	Column	Erase	Titles	Window	Status	Page	Hide
A	A	B	C	D	E	F	G			
1	EMP	EMP NAME	DEPTNO	JOB	YEARS	SALARY	BONUS			
2	1777	Azibad	4000	Sales	2	40000	10000			
3	81964	Brown	6000	Sales	3	45000	10000			
4	40370	Burns	6000	Mgr	4	75000	25000			
5	50706	Caesar	7000	Mgr	3	65000	25000			
6	49692	Curly	3000	Mgr	5	65000	20000			
7	34791	Dabarrett	7000	Sales	2	45000	10000			
8	84984	Daniels	1000	President	8	150000	100000			
9	59937	Dempsey	3000	Sales	3	40000	10000			
10	51515	Donovan	3000	Sales	2	30000	5000			
11	48338	Fields	4000	Mgr	5	70000	25000			
12	91574	Fiklore	1000	Admin	8	35000	---			
13	64596	Fine	5000	Mgr	3	75000	25000			
14	13729	Green	1000	Mgr	5	90000	25000			
15	55957	Hermann	4000	Sales	4	50000	10000			
16	31619	Hodgedon	5000	Sales	2	40000	10000			
17	1773	Howard	2000	Mgr	3	80000	25000			
18	2165	Hugh	1000	Admin	5	30000	---			
19	23907	Johnson	1000	VP	1	100000	50000			
20	7166	Laflare	2000	Sales	2	35000	5000			

DATA.WK3

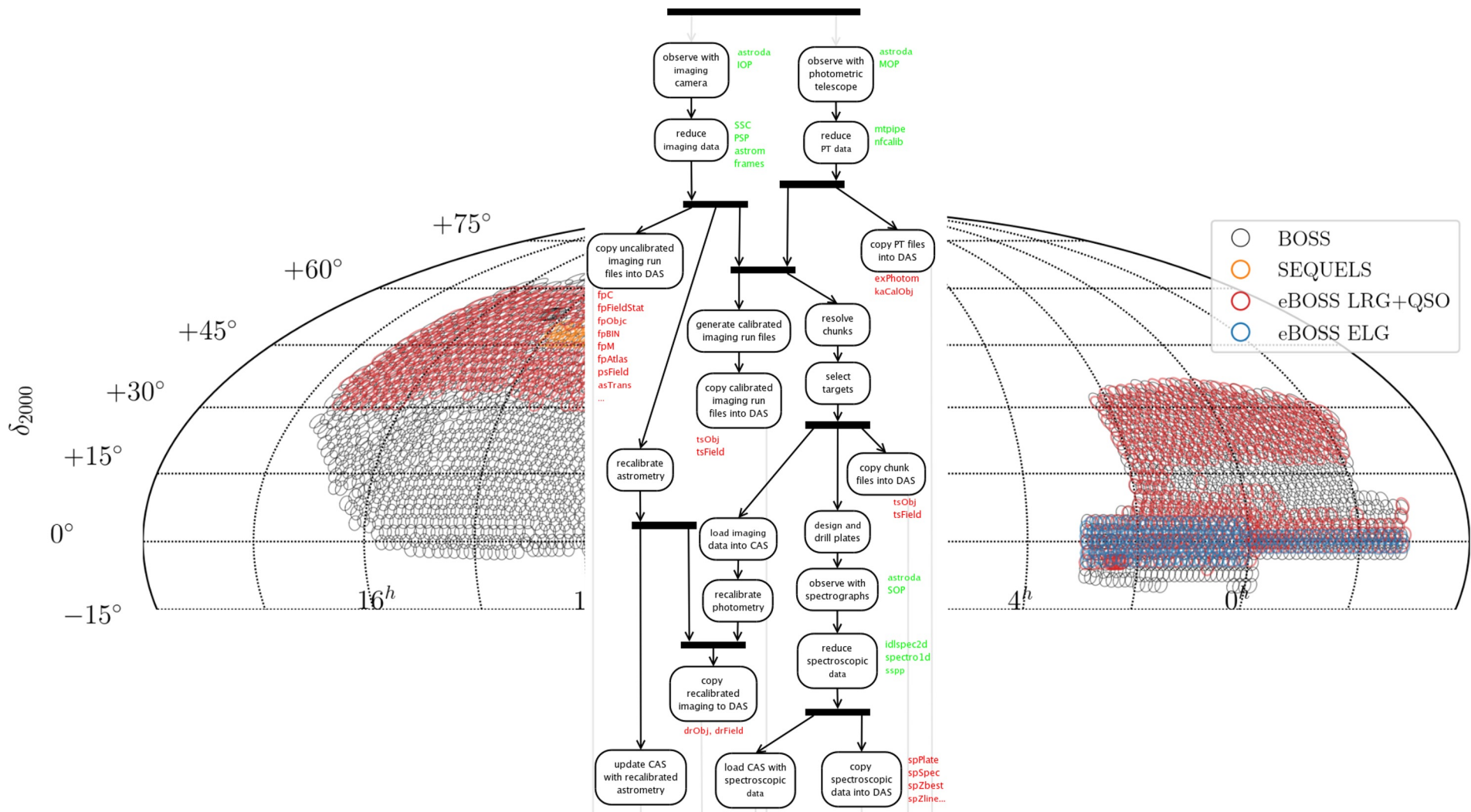


EUD systems also draw on distributed cognition - a built environment that helps users make good decisions as a force of UX and habit



(the opposite of magical
and sparkly)





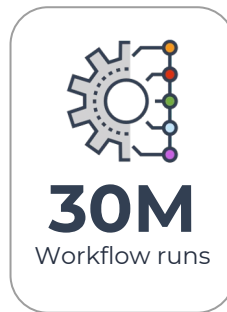
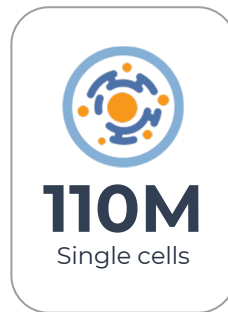
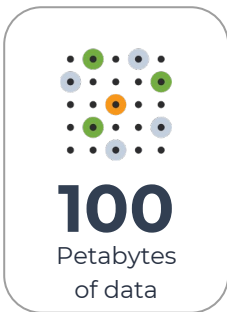
How do you build a
complex and useful data
platform for scientists?

“Five problems at a time,
with technical standards
and steady principles”

It is itself a form of
collective intelligence built
over time



Terra, from the Broad's DSP



Copy-paste-edit as collaboration practice

Terra

WORKSPACES

Workspaces > pathogen-genomic-surveillance/COVID-19_Broad_Viral_NGS > Data

0

DASHBOARD

DATA

ANALYSES

WORKFLOWS

JOB HISTORY

IMPORT DATA

EDIT

OPEN WITH...

EXPORT

SETTINGS

0 rows selected

ADVANCED SEARCH

Search

TABLES

Search all tables

nextstrain_data (9)

sample (2579)

sample_set (1925)

sample_set_set (3)

REFERENCE DATA

No references have been added.
Add reference data

OTHER DATA

sample_id

cleaned_bam

MA_DPH_000...

MA_DPH_00001.IERCC-00012_SSII...

MA_DPH_000...

MA_DPH_00002.IERCC-00031_SSII...

MA_DPH_000...

MA_DPH_00003.IERCC-00051...

MA_DPH_000...

MA_DPH_00004.IERCC-00073_SSII...

MA_DPH_000...

MA_DPH_00005.IERCC-00092_SSII...

MA_DPH_000...

MA_DPH_00006.IERCC-00116_SSII...

MA_DPH_000...

MA_DPH_00007.IERCC-00142_SSII...

MA_DPH_000...

MA_DPH_00008.IERCC-00162_SSII...

MA_DPH_000...

MA_DPH_00009.IERCC-00013_SSII...

1 - 100 of 2579

1 2 3 4 5

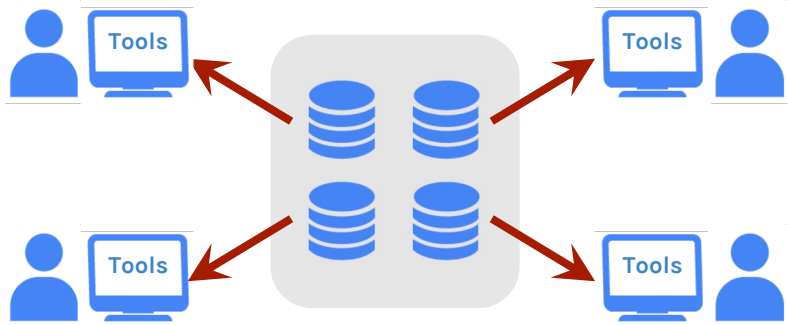
Items per page: 100

Rate: \$0.00 per hour

Software that centers policy

Traditional approach

Bring data to researchers

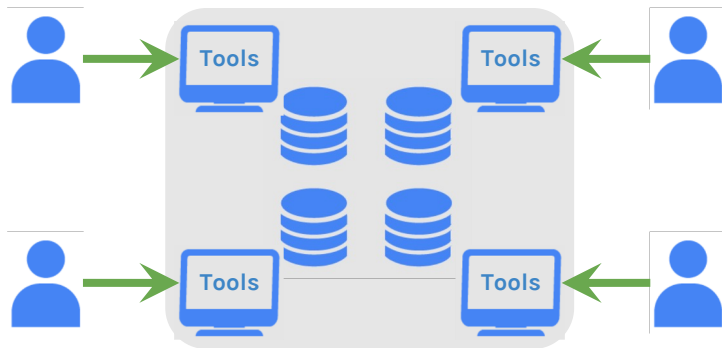


Discourages shared research

- Data sharing = data copying
- Few audit controls
- Huge infrastructure needed
- Siloed compute

Cloud-centric approach

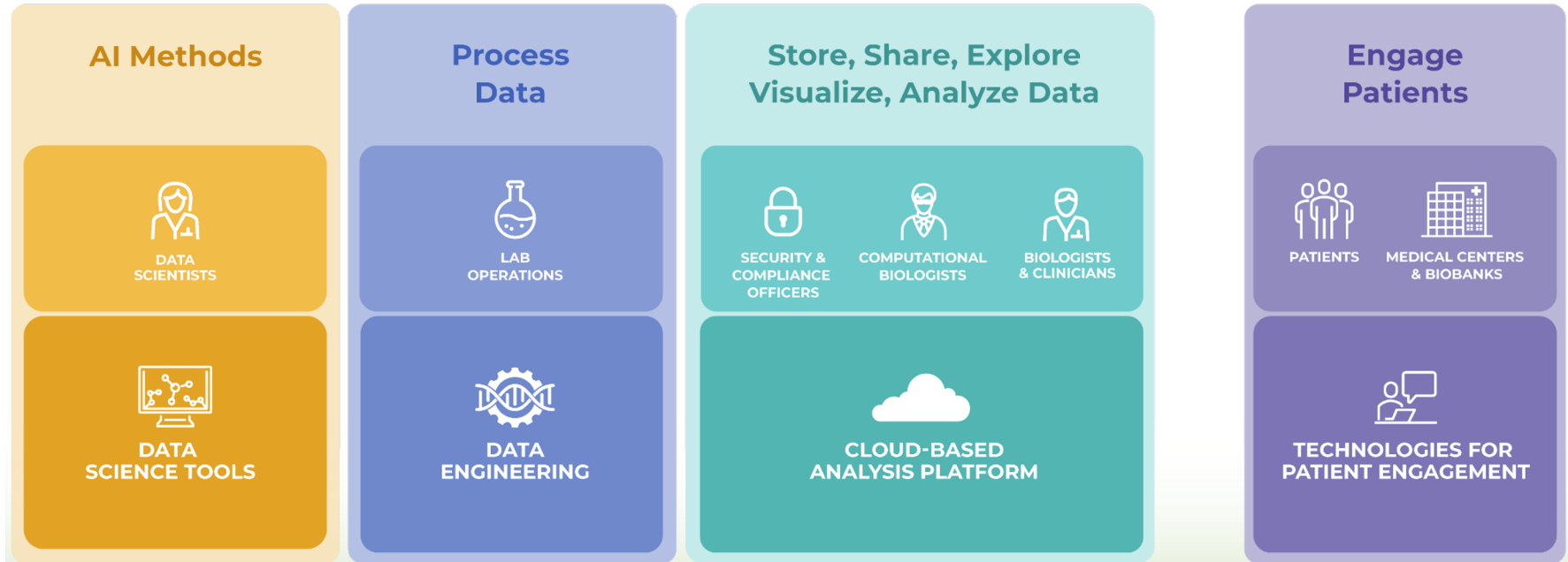
Bring researchers to data



Facilitates collaboration

- Cost
- Threat Detection and auditing
- Increased accessibility
- Shared & elastic compute

8 years of platform development for end users



An organization of ~230 software engineers and ML experts, building a software platform that spans the lifecycle of biomedical data.

It is remarkably hard to build and maintain this product. Why?

We fund biology on predetermined multi year time cycles, and software on the same circadians

Which leads to some truly difficult software development models





Stakeholder-led software:

“Here is a laundry list of things that some folks wanted, please add them”

“Why can’t you put a green button here?”

“Project’s over, time to transfer to sustainability!”



Stakeholder-led software:

“Here is a laundry list of things that some folks wanted, please add them”

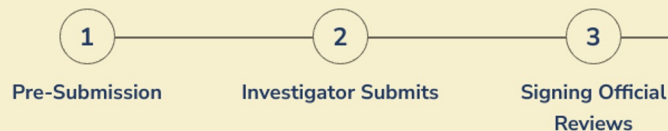
“Why can’t you put a green button here?”

“Project’s over, time to transfer to sustainability!”

“Can you put some AI in it?”



Data data everywhere, let's all write 15 data access requests



Most Common Reasons for Rejection of Data Access Requests

- The Institutional Review Board (IRB) approval letter does not satisfy requirements
 - Some datasets require local IRB approval for use, as noted on the dbGaP website
- The Research Use Statement in the request is not consistent with the dataset's terms of use
 - To understand data use limitations on datasets, please review the Research Use Statement for the dataset
- **Personnel errors** (for example: requestor is not the PI, collaborators are not listed, or the requestor is not the PI, collaborator, or the signing official or the IT director for the project)
 - For all non-intramural investigators, dbGaP receives information on the requestor's affiliation and role from the Institutional Review Board (IRB) approval letter
 - In addition, the signing official's and IT director's emails must be from the institution's email domain
 - For more information on who may request data from dbGaP, check the dbGaP website. For a list of dbGaP-approved signing officials and IT directors, see [How Can I Find the SO and IT Director for a Dataset?](#)

Was this page helpful?

Yes


No



A black and white cartoon illustration. On the right, a police officer in a uniform and cap stands with his hands on his hips, looking towards the left. A speech bubble from him says, "THIS IS WHERE YOU LOST YOUR WALLET?". On the left, a man in a suit is kneeling on the ground, looking up at the officer. A speech bubble from him says, "NO, I LOST IT IN THE PARK. BUT THIS IS WHERE THE LIGHT IS." The man is positioned next to a vertical pole and a horizontal light fixture. The background is dark and textured, suggesting a night scene or a dimly lit area. The drawing style is simple and expressive, with bold lines and a limited color palette (black, white, and grey).

THIS IS WHERE YOU
LOST YOUR WALLET?

NO, I LOST IT IN THE PARK.
BUT THIS IS WHERE THE LIGHT IS.



This is the data
you need?

No, this is the data I can
access and compute on

(this is all going to get much worse because of scale pressure)



Can I download GTEx V8 protected data from SRA?

Protected data for GTEx V8 and future releases are only available in [AnVIL](#). Due to the large size of these data, we recommend against downloading the data, as that will incur significant egress charges. Instead, we recommend that you consider performing your computations in the AnVIL/Terra environment. That will incur compute charges, but those might be significantly less than the egress charges (depending upon your computation).

Access to protected data is described [here](#)

Expression Program (GTEx)

For the Public

Health Relevance

Science Highlights

For Researchers

Funding Opportunities

Funded Research

The Common Fund's **Genotype-Tissue Expression (GTEx) Program** established a data resource and tissue bank to study the relationship between genetic variants (inherited changes in DNA sequence) and gene expression (how genes are turned on and off) in multiple human tissues and across individuals. GTEx also increased our understanding of how gene expression varies between male and female.

The GTEx program has transitioned from Common Fund support. Common Fund programs are strategic investments that achieve a set of high-impact goals within a 5-10 year timeframe. At the conclusion of each program, deliverables will transition to other sources of support or use within the scientific community.

The GTEx program supported by the Common Fund from 2010 to 2019. Currently, GTEx data are widely used as a reference dataset to design new methods and tools, such as a statistical method called [PrediXcan](#). This novel method is used to predict the expression of a gene using DNA sequence data. PrediXcan also predicts visible traits of diseases. GTEx researchers used this method to identify specific genes associated with five diseases.

fold.

Last month, NHGRI issued version 8 of GTEx, the first "free" release on the Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL) platform. NHGRI established AnVIL in late 2018 to create a cloud-based environment for working with the GTEx dataset, which includes genotype data from 838 donors plus 17,382 RNA sequences across 54 tissue sites and two cell lines.

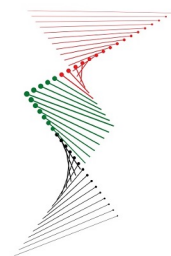
The National Institutes of Health Common Fund established GTEx in 2010 as a 10-year multi-institutional research effort to present a comprehensive atlas of genetic regulatory variation across cell types and tissues and an analysis of how these changes in regulation can contribute to risk for disease and the development of traits. The [research concluded in September](#), but the dataset endures to assist outside scientists.

Michael Schatz, program director for the AnVIL platform, said that GTEx is the "most highly requested" dataset throughout the entire National Institutes of Health.

"We started to take it for granted that you can just go to the web at any time and download it, but in reality, there's a lot of infrastructure costs," said Schatz, an associate professor of computer science and biology at Johns Hopkins University. The full GTEx dataset contains about 40,000 individual files and requires about 150 terabytes of storage.

"If you want the whole collection, it's going to take realistically several days to download," Schatz said. "Researchers constantly streaming those data would consume all of [the National Center for Biotechnology Information's] bandwidth. There's just a lot of overhead with that."

States are figuring this out!



EMIRATI
GENOME

FUTURE OF OUR GENERATIONS



Indiana
Biobank

Indiana Clinical and Translational
Sciences Institute



estonian genome center
university of tartu



Welcome to Application Submission System & Interface for Submission Tracking (ASSIST)
Online Help

Application Structures in ASSIST

Latest Updates

Using ASSIST

Using ASSIST to Apply for an LRP Award

Prepare an Application

Forms Data Entry

Application Submission Status Workflow

Generate a Preview of the Application

Non-Research Amendments (for Non-Research Agencies)

Non-Competing Continuation Overview

Validate the Application

Prepare an OTA Application

Initiating an OTA Application

OTA Summary Screen

OTA Form Screen

OTA Application Submission Status

Application Errors and Warnings Results

Verify Senior Key Personnel

Application Submission

Application Submission System & Interface for Submission

Tracking (ASSIST): Revised May 13, 2024



For additional assistance, please contact the [eRA Service Desk](#).

[PDF version](#)

About Other Transactional Authority (OTA) Awards

What is an Other Transactions Authority?

An Other Transactions Authority (OTA) allows for Federal Government agencies to enter into Other Transactions (OTs).

What is an Other Transaction?

An Other Transaction (OT) is a unique type of legal instrument other than a contract, grant, or cooperative agreement. Generally, this awarding instrument is not subject to the FAR, nor grant regulations unless otherwise noted for certain provisions in the terms and conditions of award. It is, however, subject to the OT authority that governs the initiative as well as applicable legislative mandates.

Why are Other Transactions used instead of traditional funding mechanisms?

Reasons to use an OT may include a combination of the following, among others:

An Other Transactions Authority (statute's citation);

Need for flexibility to negotiate terms and conditions appropriate for the specific program requiring fluid implementation;

Nontraditional review and award management practices are needed because the science is expected to be highly evolving, with requirements for additional aims or expertise added to, or removed from, the project throughout the term of execution;

Collaborative involvement by the NIH in the technical direction and oversight of the research, which can be akin to partnering (e.g., participation in progress reviews and decisions on future efforts or direction; government may be a voting or nonvoting member of the consortium);

Negotiate intellectual property rights; and/or

Communities are figuring this out!



Native **BioData**
consortium



The opportunity / the risk

Growing conviction by researchers of the value of genotyped patients for research and clinical development



A growing number of biobanks around the world collecting exponential genetic data on patients



The opportunity / the risk

Growing conviction by researchers of the value of genotyped patients for research and clinical development



A growing number of biobanks around the world collecting exponential genetic data on patients



The opportunity exists, now, to connect the explosion in data to the collective intelligence of biology writ large - and to make it fair and equitable

. But we won't meet that opportunity without a systemic and cultural shift in how we fund, build, staff, and sustain biological software platforms.

“Most importantly, the advanced expertise gained by individual users becomes more than just the knowledge of a single individual; it is a shared community resource”

(please go read Bonnie Nardi!)



Thank you!

jwilbank@broadinstitute.org