



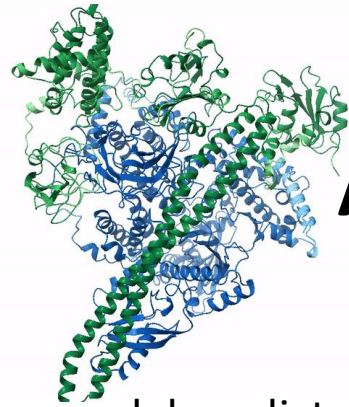
CREATING THRIVING DATA-CENTRIC COMMUNITIES FROM BASIC RESEARCH TO COMMERCIAL APPLICATIONS

Rob Schuler & Carl Kesselman

September 18, 2024

“Impact from eScience” Workshop @ IEEE eScience 2024

Discoveries, broad societal impact, are driven from data...



AlphaFold

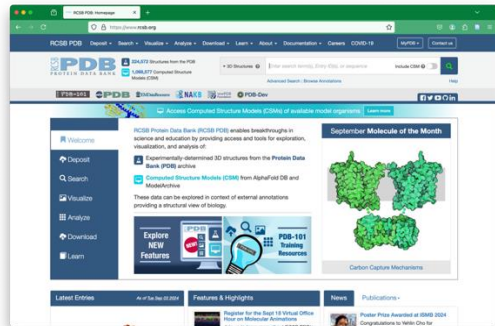
Deep learning model predicts molecular structure and interactions with unprecedented accuracy. Uses in drug design to benefit human health treatments.



ChatGPT

Large language model able to mimic human conversation and adept at a wide range of natural language tasks. Enabling an array of AI assistants.

BUT...!



AlphaFold was trained on 170,000+ protein structure models from open databases, notably the Protein Data Bank (PDB) dating back five decades.

GPT trained on proprietary training data set widely believed to depend critically on Wikipedia content that has been authored and edited for over two decades.



These didn't merely spring forth from an algorithm...

Tremendous high-quality data were used to train these models

Yet there is a lack of usable data “out there”...

“The first thing we’ve learned is the importance of having outstanding data to base your ML on. In our own shop, we’ve been working on a few big projects and we’ve had to spend most of the time just cleaning the data sets before you can even run the algorithm. That’s taken us years just to clean the datasets. I think people underestimate how little clean data there is out there, and how hard it is to clean and link the data.”

Vas Narasimhan, CEO Novartis

How do we get the “Data” in “Data Science”???



AI Work Assistants Need a Lot of Handholding

Getting full value out of AI workplace assistants is turning out to require a heavy lift from enterprises. 'It has been more work than anticipated,' says one CIO.

By Isabelle Bousquette [Follow](#)
June 25, 2024 at 3:33 pm ET

Garbage in, garbage out: mitigating risks and maximizing benefits of AI in research

Brooks Hanson, Shelley Stall, Joel Cutcher-Gershenfeld, Kristi Yuhan (Douglas) Rao & Ge Peng

Artificial-intelligence tools are transforming data-driven science – better ethical standards and more robust

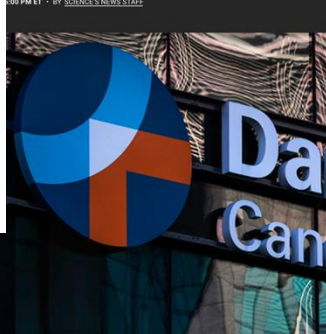
Science is producing so large as to Advances in art are increasingly of all this infor

Errors found in dozens of papers by top scientists at Dana-Farber Cancer Institute

Harvard-affiliated research institute seeks retractions on six papers and corrections for 31 others

June 4, 2024

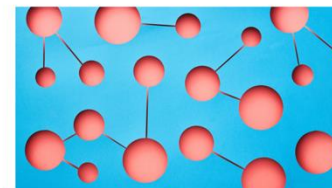
1:00 PM ET · BY SCIENCE'S NEWS STAFF



Why AI Failed to Live Up to Its Potential During the Pandemic

by Bhaskar Chakravorti

March 17, 2022



- Despite the tremendous interest in “data science” many projects fail, often related to data:
 - 80% of time spent on accessing, cleaning, integrating data
 - Scarcity of data sharing
 - 10% reproducibility of data
 - Recent high-profile retractions in COVID-19 research, for example.

It takes a (data) village...

- *Individuals or single labs* cannot unilaterally create the data resources necessary to drive wide societal impact
- *Data-centric communities* needed to produce the scale of data to unlock innovation for broad societal impact
- Yet, *data handling skills are often lacking*, compute and algorithms often chased first, and stewardship later
- Finally, there is an element of serendipity... no one creating PDB or Wikipedia decades ago would have envisioned their usage in deep learning models



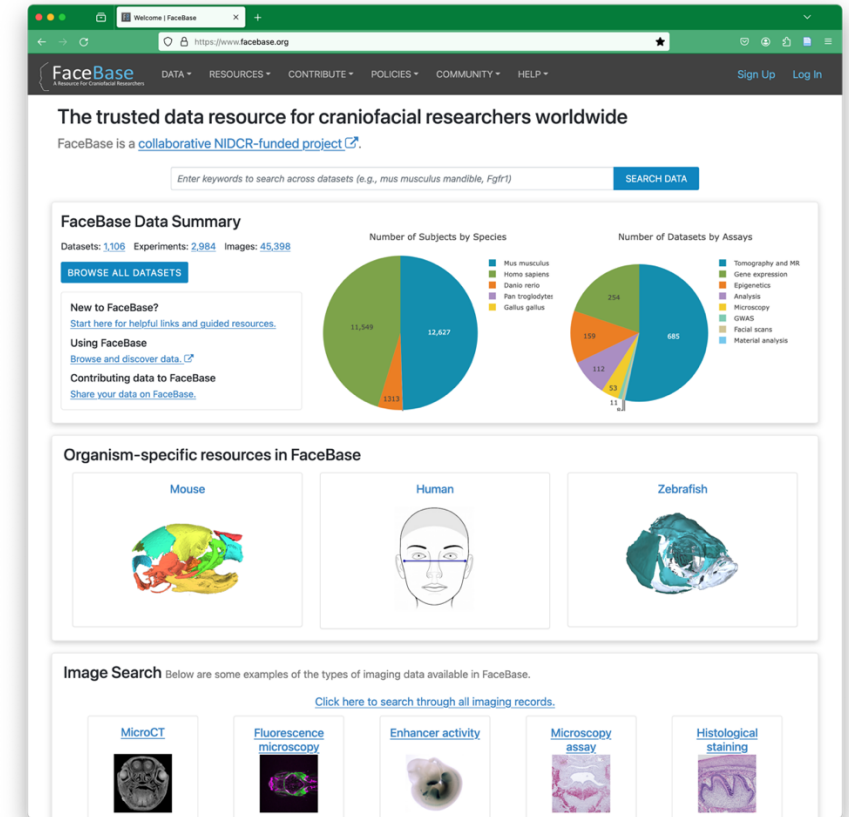
*Opportunities for eScience to play a greater role in data-centric discovery --
socialize and train data scientists on stewardship of data and best practices in usage of data*

An example from Craniofacial Research

To serve as the trusted online data resource for dental, oral, and craniofacial (DOC) researchers worldwide

DOC community is comprised of:

- **Basic researchers**
- **Clinicians**
- **Clinician scientists**
- **Public health researchers**
- **Commercial applications**
- **Trainees**
- **Patient advocates**

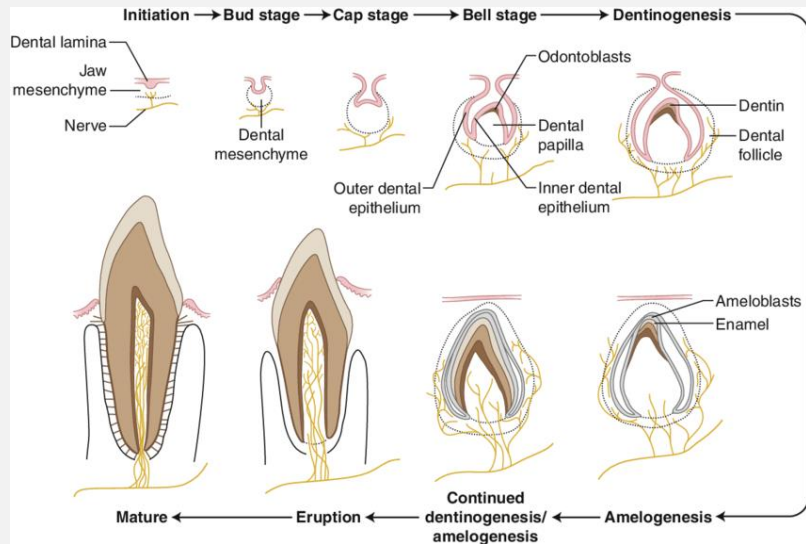


facebase.org

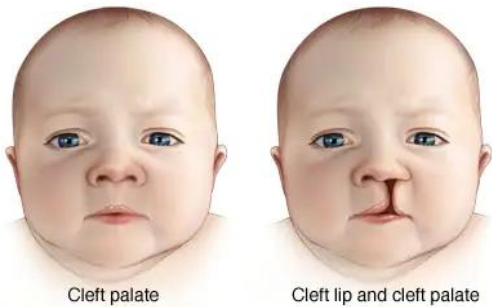
Established in 2009

In service of improving human health...

Tooth development as a model to study the role of nerves in organogenesis



Fried and Gibbs, The Dental Pulp, 2014



Cleft palate Cleft lip and cleft palate

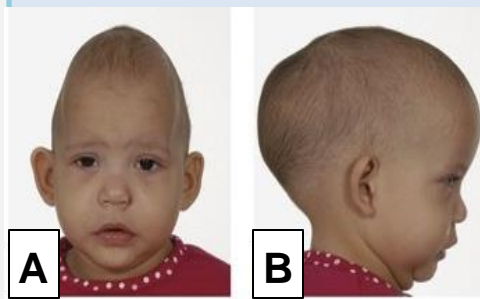


Cleft Lip and Cleft Lip/Palate

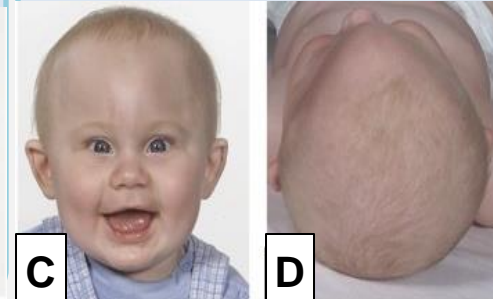
Mayo Clinic, 2024

And many “rare” diseases without known genetic determinants

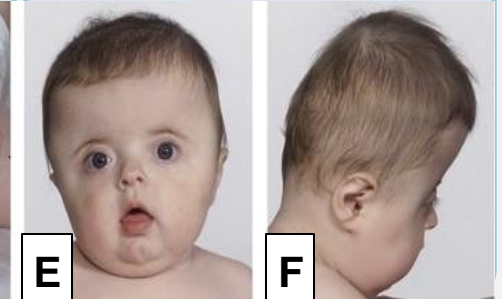
Sagittal synostosis



Metopic synostosis



Bi-coronal synostosis



Craniosynostosis involves premature fusion of the cranial sutures

Johnson and Wilkie, Eur J Hum Genet, 2011; Sidoti et al., Plast. Reconst. Surg., 1996; Boulet et al., Am J Med Genet A, 2008; Twigg and Wilkie, Am J Hum Genet, 2015; Stanton et al., Dis Model Mech, 2022

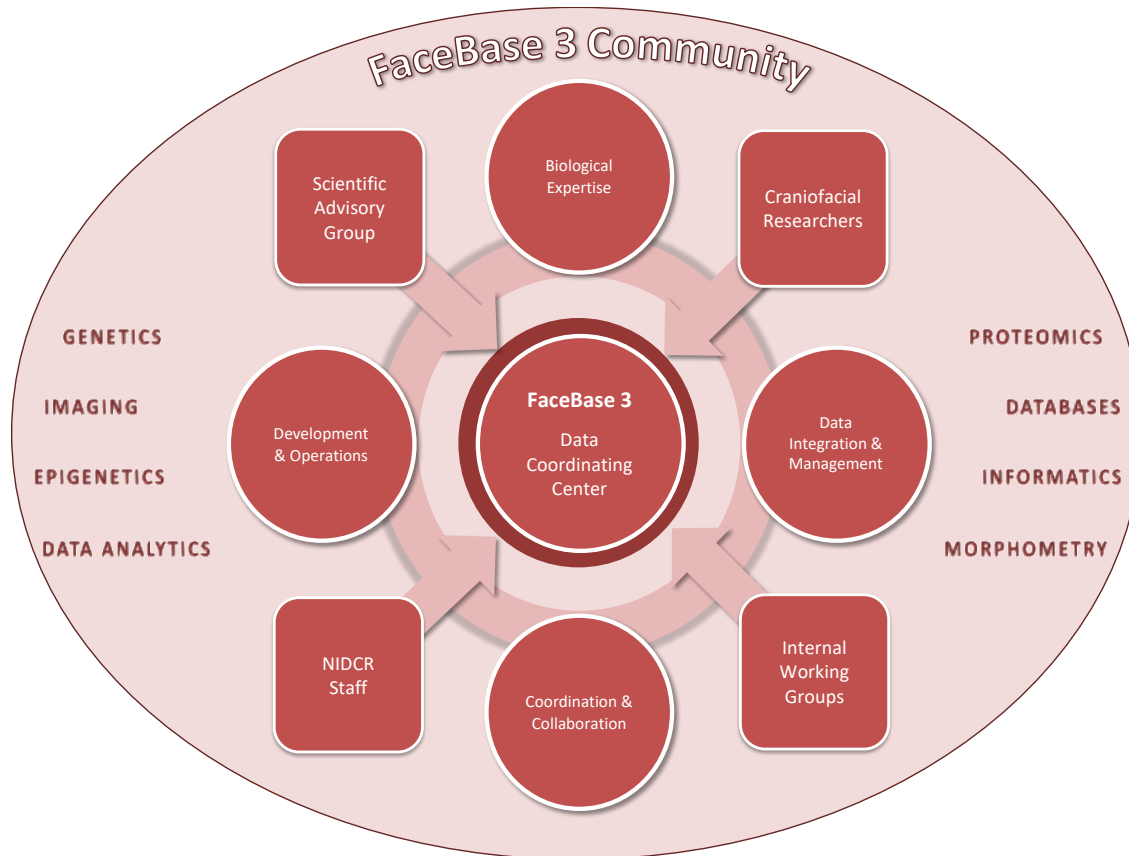
Empowers FAIR Data Usage...

1. Find: consistent labeling leads to accurate filtering to find data of interest

2. Access: detailed descriptions with well organized details on experiments and biological characteristics

3. Interop and Reuse: many paths for online visualization or download for offline analysis

We “partner” with the community



- **Steering Committee:** frequent (bi-weekly+) meetings with government sponsors, oversight, direction
- **Advisory Committee:** annual+ meetings with diverse group on strategic, scientific direction
- **Users:** events throughout year, UX research, usability input
- **Consortia:** partnering with complementary projects

Partnering keeps us grounded in the problems this scientific community faces and the key innovations in discovery and therapy

Problem solved...?

Bwr?????

0.2982?????

data.csv?????

The screenshot shows a web browser displaying a Figshare dataset page. The main content is a CSV table with 13 rows and 8 columns (A-H). The table data is as follows:

A	B	C	D	E	F	G	H
Bwr	Bwr_visit	Visit	N	sd_InW	sd_InD	slopeD	slopeW
1	1_3	3	36	0.3493	0.3849	0.0469	0.0462
1	1_6	6	48	0.3865	0.2777	0.0663	0.0987
1	1_0	0	69	0.2992	0.3263	0.0633	0.0327
2	2_6	6	8	0.2509	0.2141	0.0436	-0.0083
2	2_3	3	17	0.3164	0.3191	0.017	-0.0243
2	2_0	0	201	0.3117	0.4029	0.0142	0.0101
8	8_3	3	34	0.352	0.3825	0.0302	0.0071
8	8_6	6	69	0.523	0.4146	0.0771	0.1029
8	8_0	0	108	0.4961	0.463	0.0598	0.0866
13	13_6	6	14	0.5272	0.6256	0.0386	0.0259
13	13_3	3	55	0.2982	0.3696	0.0064	0.0274
13	13_0	0	169	0.5603	0.5951	0.0509	0.0351

Below the table, the dataset is identified as 'data.csv - The effect of building ability and object availability on the construction of bowerbirds'. It includes options to Cite, Download (3.43 kB), Share, Embed, and Collect. The dataset was posted on 2023-09-21 by Menno van Berkel. Usage metrics show 1 view, 0 downloads, and 0 citations. Categories include 'Evolutionary biology not elsewhere classified' and 'Animal behaviour'. Keywords include 'Animal architecture', 'Courtship display', 'Extended Phenotype', and 'mate choice behaviour'. The license is CC BY 4.0.

Generalist Repository

?????

KEYWORDS

- Animal architecture
- Courtship display
- Extended Phenotype
- mate choice behaviour

LICENCE

CC BY 4.0

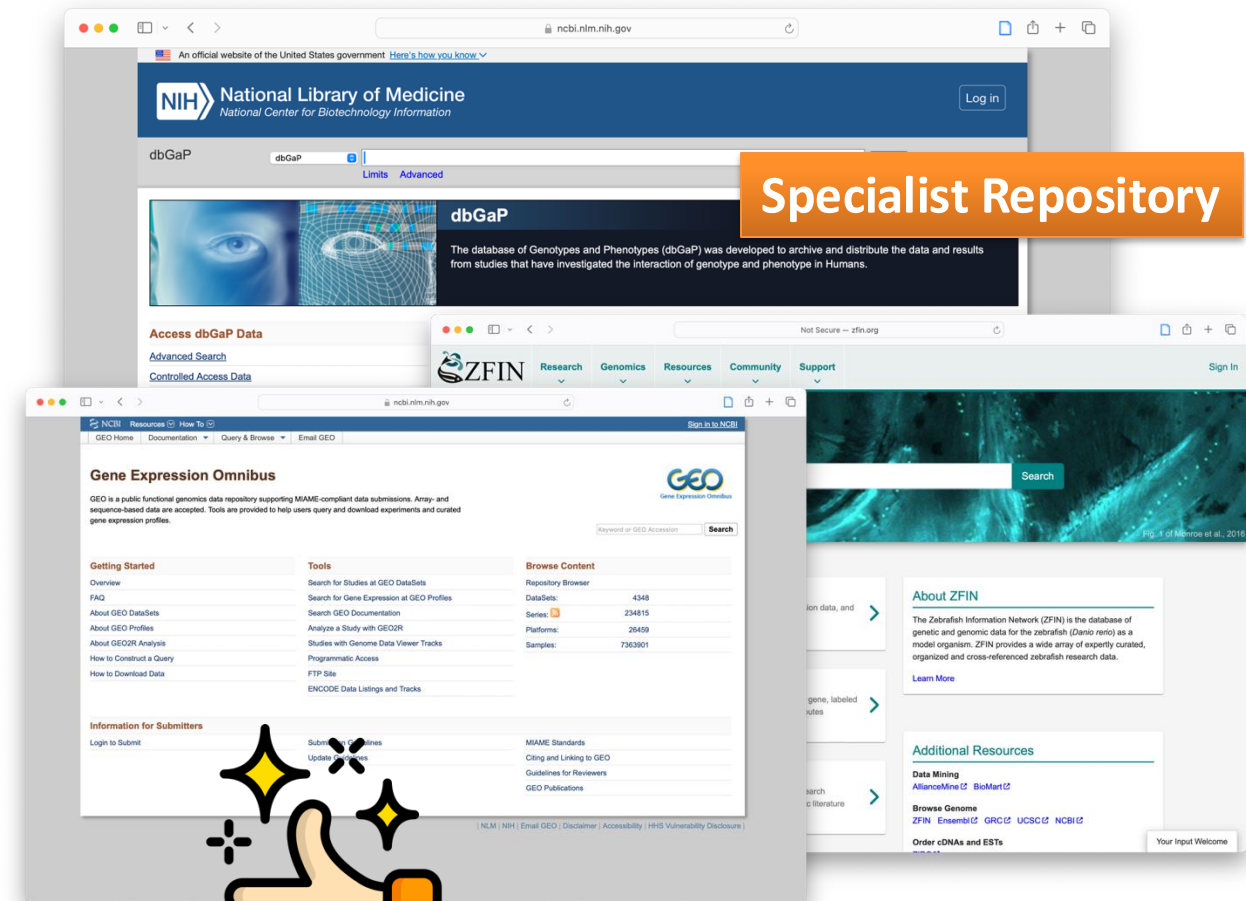
The “Generalist Repository” approach offers little in the way of support for generating “FAIR” data

Let the data curators do it?

Typical approach: “Toss it over the fence” and let curators sort it out

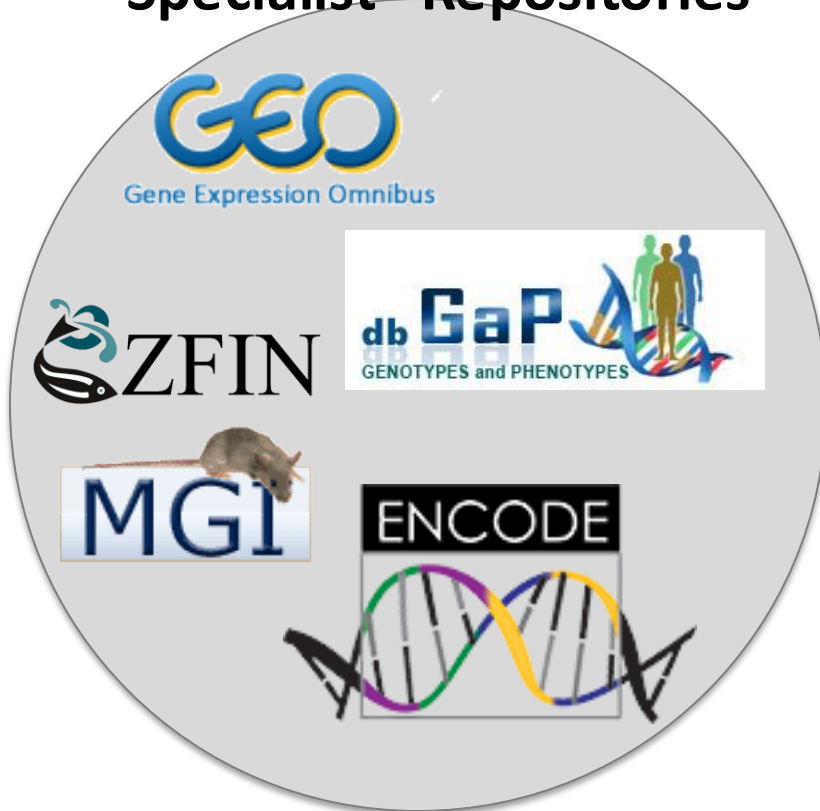
However...

- Data volumes growing rapidly
- New scientific methods and instruments impose changes
- Data (bio)curators stretched thin
- Funding for curation won't keep pace with data growth
- Process is slow, often requires lots of iteration



Addressing the gaps in the data sharing ecosystem...

“Specialist” Repositories



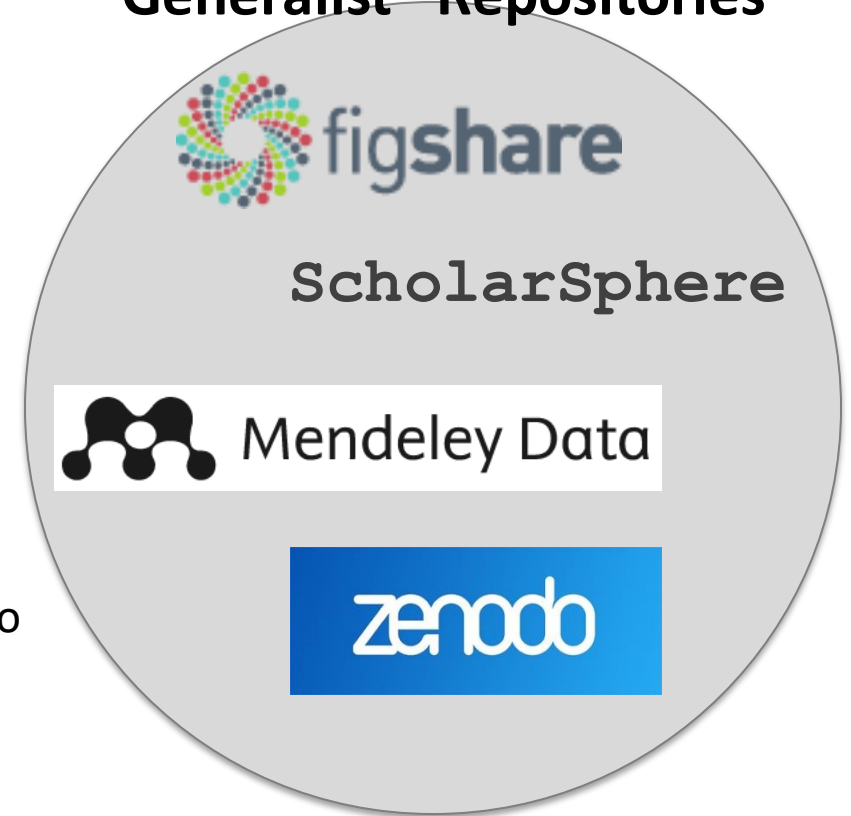
PRO: Highly detailed descriptive information; High quality data
CON: Narrow focus; High cost of biocuration

“Hybrid” Repositories



Addresses Gaps: Flexibility to adapt to new species and assays, with descriptive information for reusability; “Self-serve” data curation to guide scientists to produce quality (meta)data.

“Generalist” Repositories



PRO: All types of data and science; Highly scalable
CON: Minimal structure for data reuse; Quality concerns

Engaging the community

TRAIN

Year-round training events
(bootcamps, office hours, 1-on-1, online)



GUIDE

Rather than DIY or do-it-for-them, we guide
the self-service curation



SOCIALIZE

Participate in key community outreach events
(i.e., conferences, user symposiums)



MOTIVATE

Both “push” and “pull” forms of motivation;
i.e., a combination of incentives and mandates



INCLUDE

User-driven design and user-in-the-loop development
(i.e., usability studies, working groups)



PROVIDE

Domain-agnostic, User-friendly platform;
Tailored to the needs of the DOC research community

Provide: intuitive tools to simplify data sharing

Record Number: 1 | 2

* Dataset: Select a value

Local Identifier: 3XhSox9-E125-S5_pHsp68-lacZ-tdTom_E9.5_I-42 | hSox9_145mb_Enh-pHsp68-lacZ-tdTomato_E9.5_I-42

* Species: Mus musculus

* Specimen: whole organism preparation

Gene: Select a value

Genotype: Select a value

Strain: FVB

Mutation: Select a value

Stage: E9.5

Anatomy: head

Origin: Select a value

Phenotype: Select a value

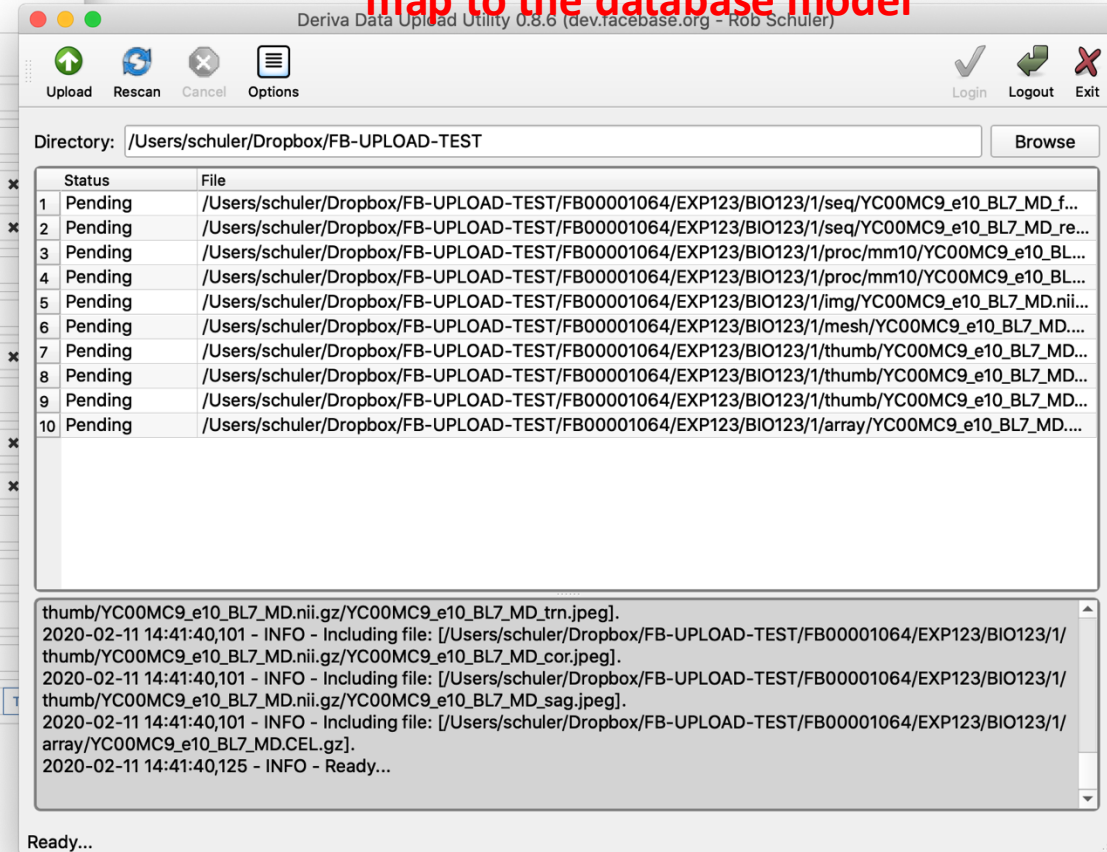
Gender: Select a value

Litter:

Collection Date: YYYY-MM-DD | Today | YYYY-MM-DD

(1) Dynamically-generated forms and views based on database model

(2) Graphical desktop applications that map to the database model



(3) Command-line, Python and Web services interfaces for integrating with analytics and other services (not pictured)

Researchers have submitted their own datasets with 100s to 1000s of files, usually in a few days

Guide: assist but don't do it for them

- Socializing, Training, Documentation, Bootcamps, 1-on-1, Screensharing,... and *Quality Control*
- ...yet effort remains shifted to community

Automated QC Issue Flagging

The screenshot displays the FaceBase project page for 'Project Chai: Integrated research of craniofacial morphogenesis'. The page includes sections for Summary, Project Publication (7), Project Investigator (1), Project Member (2), and Dataset (25+). A table of datasets is shown at the bottom, with columns for Actions, Record ID, Title, Contributors, Release Date, and Quality Control Issues. A red arrow points from the text 'Automated QC Issue Flagging' to the 'Quality Control Issues' column in the dataset table.

Release Date	Quality Control Issues
	Experiment missing at least one replicate
	Experiment missing at least one replicate
	Missing dataset stage tag, Experiment missing at least one replicate
2019-04-15	Replicate(s) missing data files
2019-04-15	Missing dataset anatomy tag, Replicate(s) missing data files
2019-04-15	Replicate(s) missing data files
2019-04-15	Replicate(s) missing data files

Actions	Record ID	Title	Contributors	Release Date	Quality Control Issues
	1-92SR	RNAseq of Ezh2fl/fl and Osr2-Cre;Ezh2fl/fl at PN3 days	Junjun Jing, Pedro Sanchez, Paul Thomas, Yang Chai		Experiment missing at least one replicate
	1-77AR	RNAseq of Nfic-/- and littermate Nfic+/- control at PN4 days	Yang Liu, Pedro Sanchez, Paul Thomas, Yang Chai		Experiment missing at least one replicate
	1-77AB	Histology and schematic overview of mouse molar root development	Jingyuan Li, Carolina Parada, Pedro Sanchez, Paul Thomas,		Missing dataset stage tag, Experiment missing at least one replicate

Reproducible AI/ML (upcoming)

Oft cited risks associated with ML in Science

- Biases inherited from data skew
- Lack of transparency in model training
- Failure to properly validate model performance
- Scarcity of clean data for training models

Garbage in, garbage out: mitigating risks and maximizing benefits of AI in research

Brooks Hanson, Shelley Stall, Joel Cutcher-Gershenfeld, Kristina Vrouwenvelder, Christopher Wirz, Yuhan (Douglas) Rao & Ge Peng

Artificial-intelligence tools are transforming data-driven science – better ethical standards and more robust data curation are needed to fuel the boom and prevent a bust.

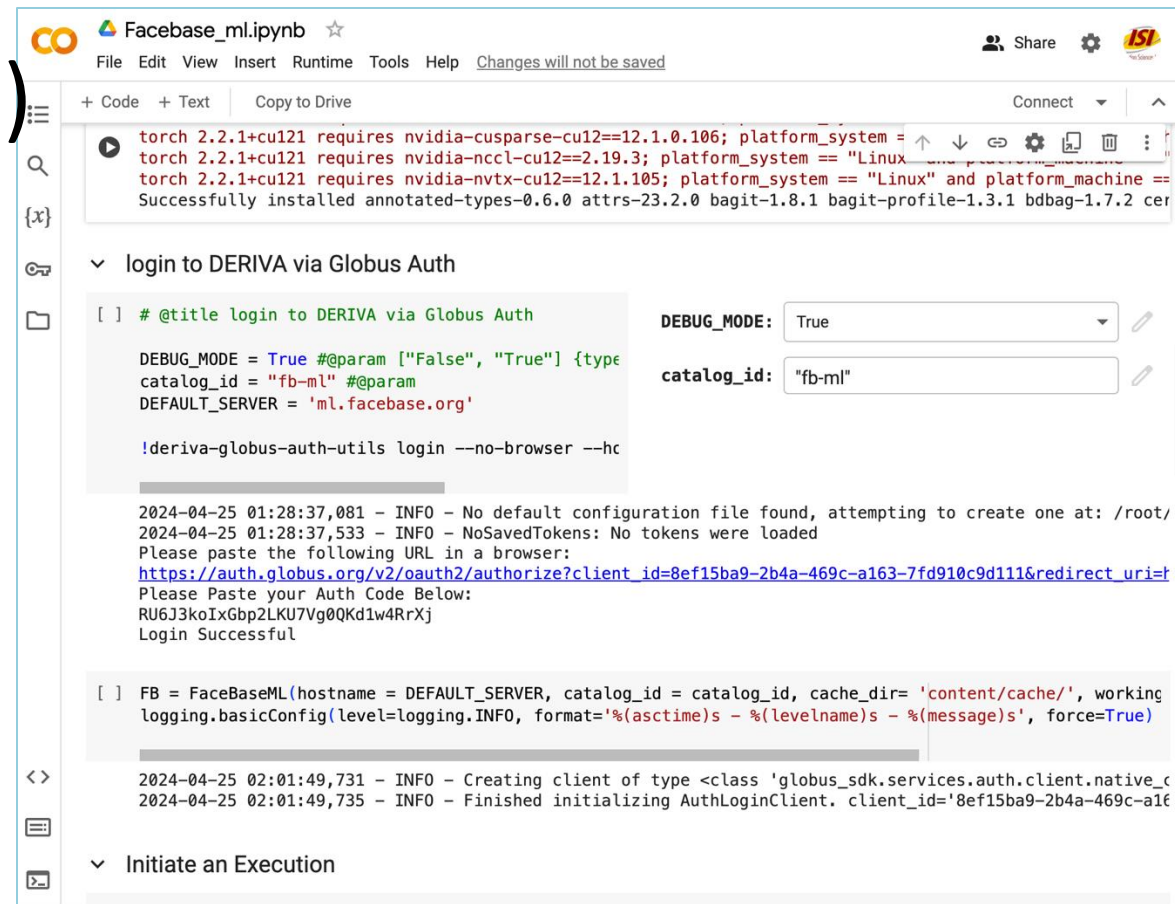
Science is producing data in amounts so large as to be unfathomable. Advances in artificial intelligence (AI) are increasingly needed to make sense of all this information (see ref. 1 and *Nature Rev. Phys.* 4, 353; 2022). For example, through training on copious quantities of data, machine-learning (ML) methods get better at finding patterns without being explicitly programmed to do so.

In our field of Earth, space and environmental sciences, technologies ranging from sensors to satellites are providing detailed views of the planet, its life and its history, at all scales. And AI tools are being applied ever more widely

– for weather forecasting² and climate modeling³, for managing energy and water⁴, and for assessing damage during disasters to speed up aid responses and reconstruction efforts.

The rise of AI in the field is clear from tracking abstracts⁵ at the annual conference of the American Geophysical Union (AGU) – which typically gathers some 25,000 Earth and space scientists from more than 100 countries. The number of abstracts that mention AI or ML has increased more than tenfold between 2015 and 2022: from less than 100 to around 1,200 (that is, from 0.4% to more than 6%; see ‘Growing AI use in Earth and space science’⁶).

Yet, despite its power, AI also comes



```
torch 2.2.1+cu121 requires nvidia-cusparse-cu12==12.1.0.106; platform_system =
torch 2.2.1+cu121 requires nvidia-nccl-cu12==2.19.3; platform_system == "Linux
torch 2.2.1+cu121 requires nvidia-nvtx-cu12==12.1.105; platform_system == "Linux" and platform_machine ==
Successfully installed annotated-types-0.6.0 attrs-23.2.0 bagit-1.8.1 bagit-profile-1.3.1 bdbag-1.7.2 cer

login to DERIVA via Globus Auth

[ ] # @title login to DERIVA via Globus Auth
DEBUG_MODE: True
catalog_id: "fb-ml"
DEFAULT_SERVER = 'ml.facebase.org'

!deriva-globus-auth-utils login --no-browser --hc

2024-04-25 01:28:37,081 - INFO - No default configuration file found, attempting to create one at: /root/
2024-04-25 01:28:37,533 - INFO - NoSavedTokens: No tokens were loaded
Please paste the following URL in a browser:
https://auth.globus.org/v2/oauth2/authorize?client_id=8ef15ba9-2b4a-469c-a163-7fd910c9d111&redirect_uri=
Please Paste your Auth Code Below:
RU6J3koIxGbp2LKU7Vg0QKd1w4RrXj
Login Successful

[ ] FB = FaceBaseML(hostname = DEFAULT_SERVER, catalog_id = catalog_id, cache_dir= 'content/cache/', working
logging.basicConfig(level=logging.INFO, format='%(%asctime)s - %%(levelname)s - %%(message)s', force=True)

2024-04-25 02:01:49,731 - INFO - Creating client of type <class 'globus_sdk.services.auth.client.native_c
2024-04-25 02:01:49,735 - INFO - Finished initializing AuthLoginClient. client_id='8ef15ba9-2b4a-469c-a1f

Initiate an Execution
```

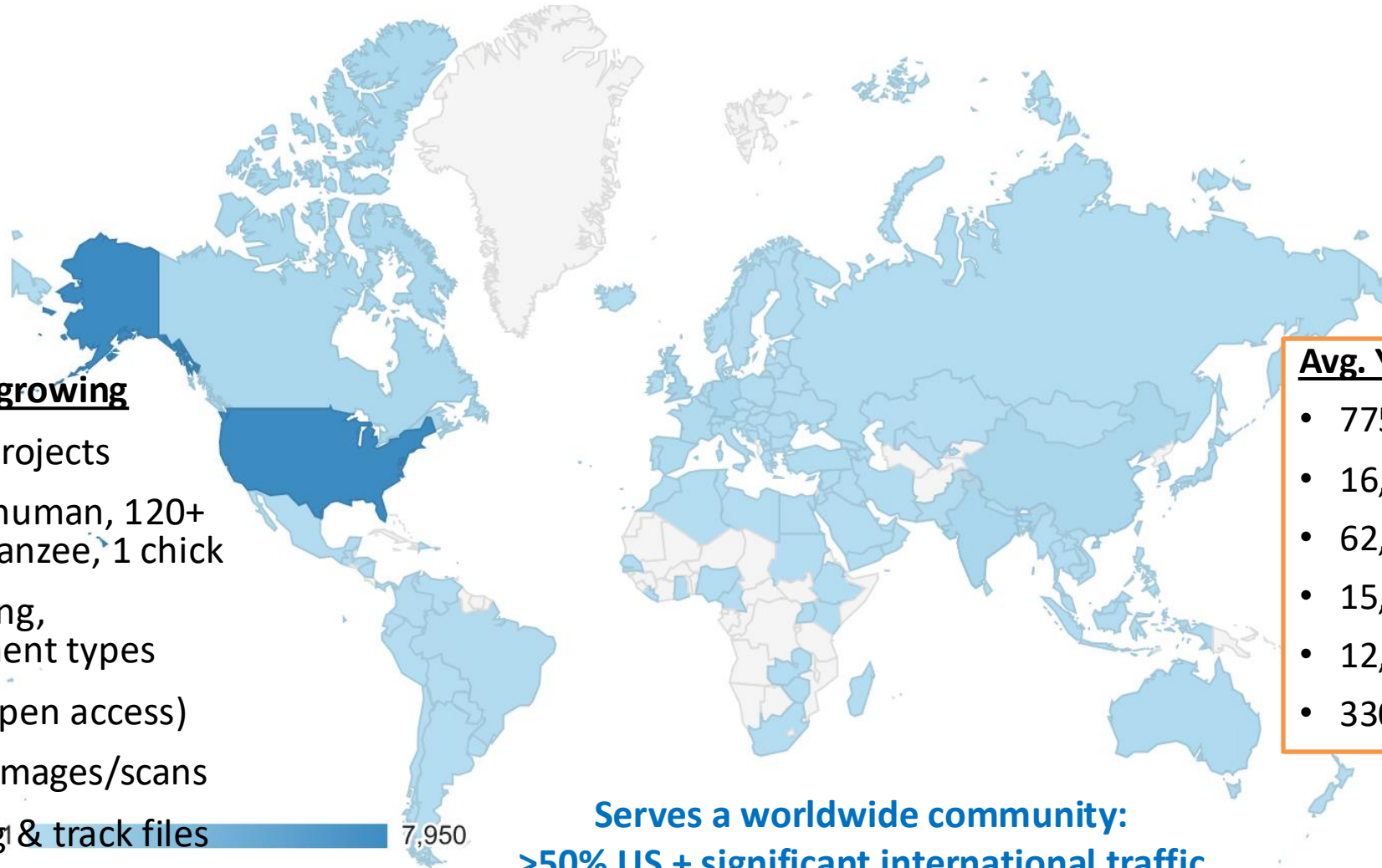
Fig. Example on Google Colab

Reproducible ML pipeline:

- ❖ Integrated with FaceBase
- ❖ Use free and low-cost Google Colab service
- ❖ Use ML-accelerated VM on AWS Cloud
- ❖ Automatically generates train/validate/test splits
- ❖ Catalogs all operations to ensure reproducibility

Li et al., eScience, 2024

Evidence of a thriving community...



1100+ Datasets and growing

- 60~ contributing projects
- 890+ mouse, 80+ human, 120+ zebrafish, 2 chimpanzee, 1 chick
- Imaging, sequencing, and other experiment types
- 45,000+ images (open access)
- 22,000+ (human) images/scans
- 7,400+ sequencing & track files

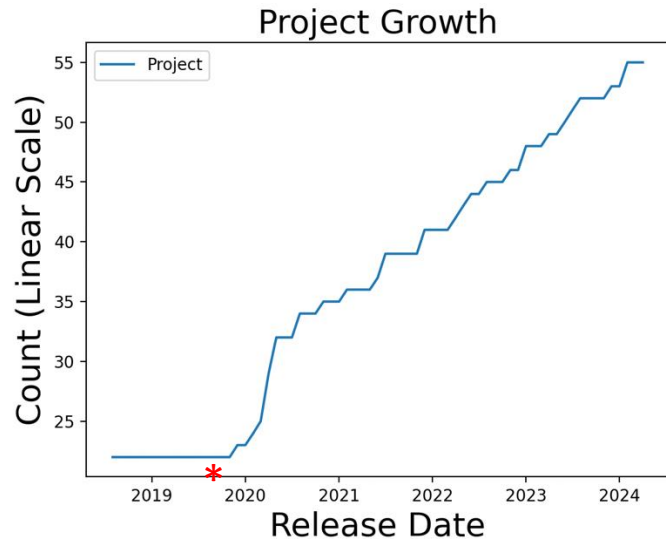
Avg. Yearly Usage

- 775 registered users
- 16,000+ visitors
- 62,000+ page views
- 15,000+ downloads
- 12,000+ image views
- 330,000+ track views

**Serves a worldwide community:
>50% US + significant international traffic**

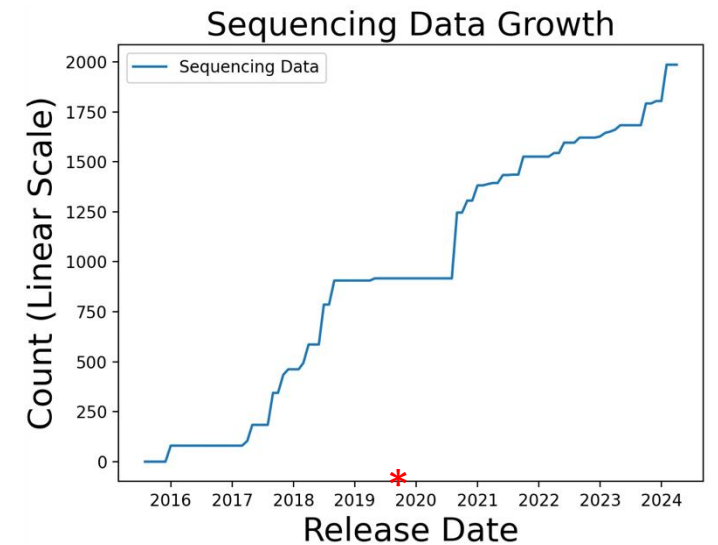
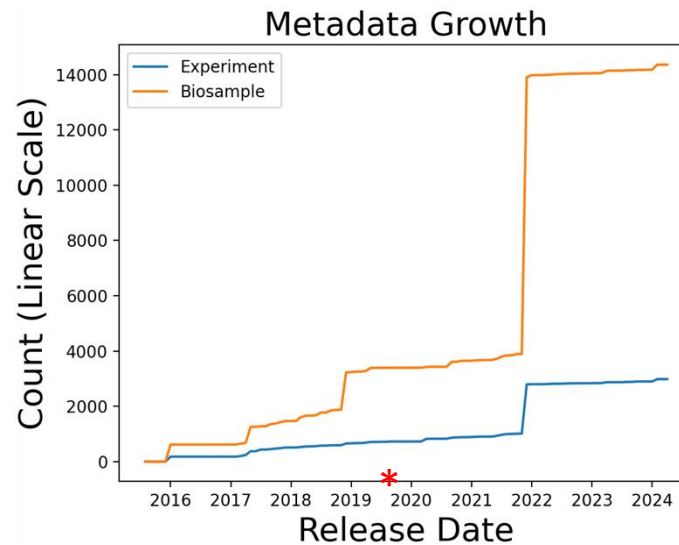
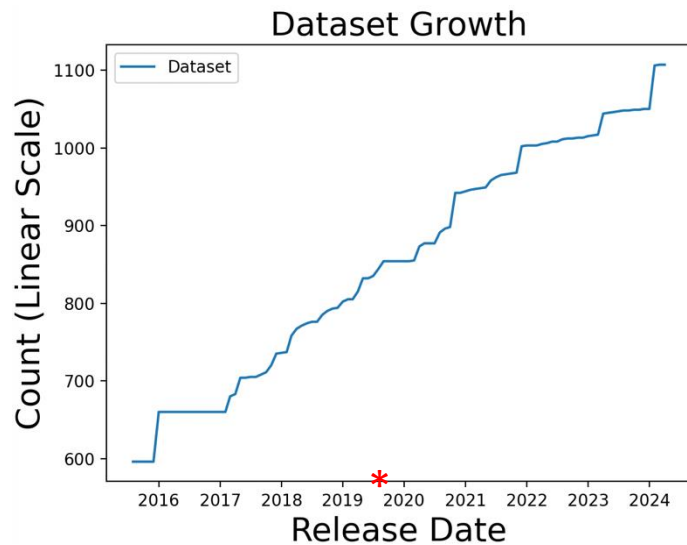
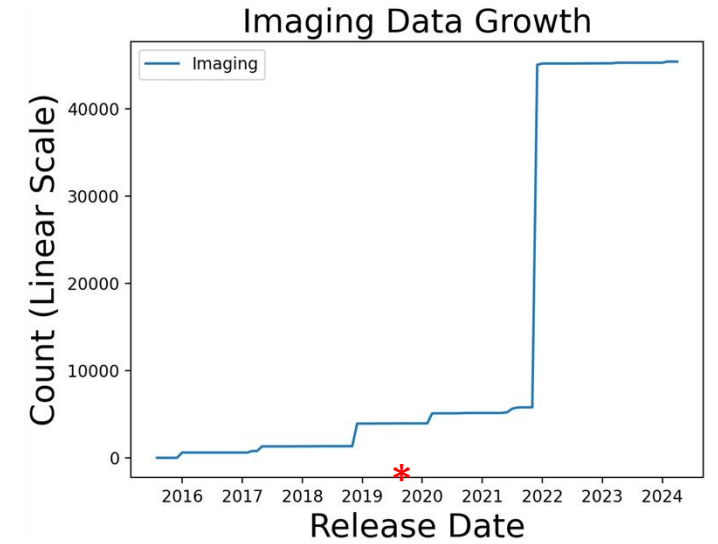
Total number of publications: 230+

Recent 5-Years of Strong Data Sharing



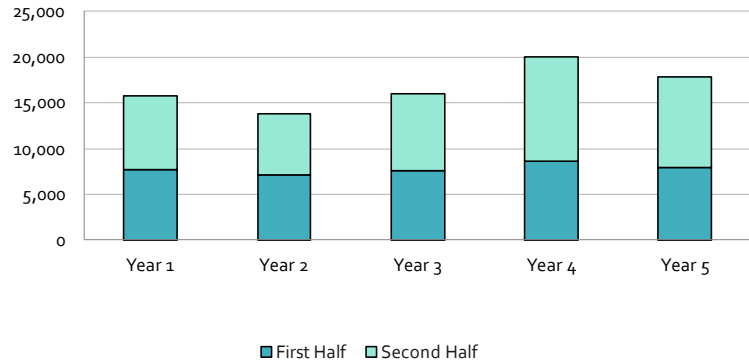
Continued strong contributions from the community even after the transition from hub-spoke to open community model

*FB3: >2019 (Aug)
FB2: 2014-19
FB1: < 2014

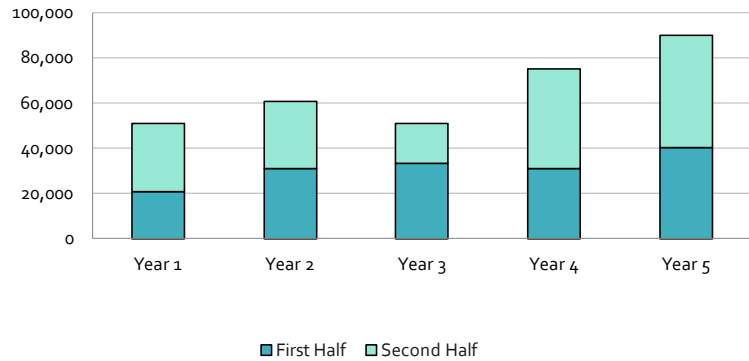


Recent 5-Years of Strong Data Reuse

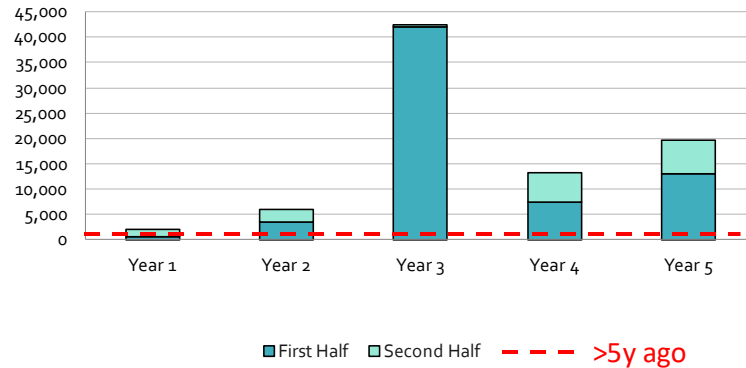
Unique Visitors to the Site



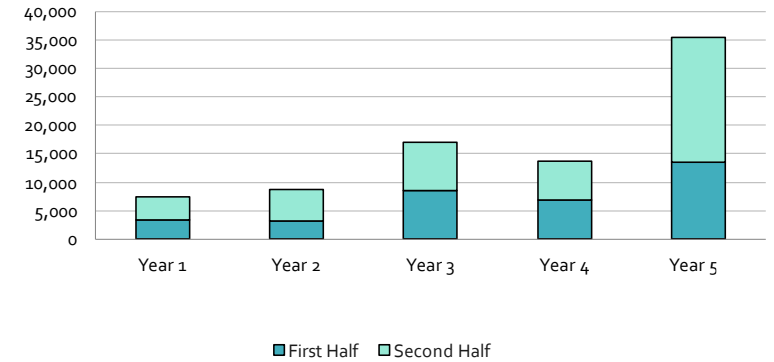
Unique Pageviews



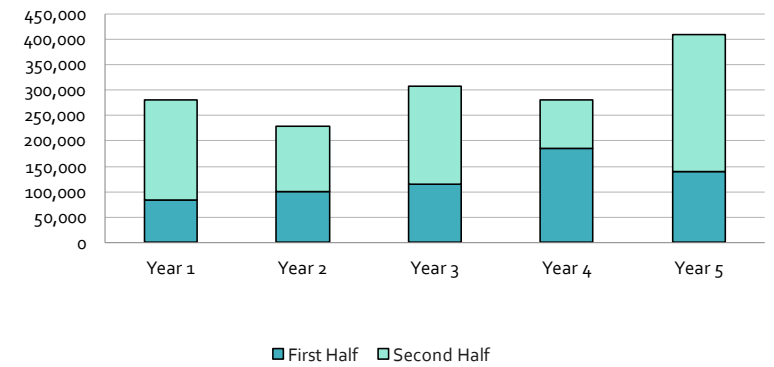
Data Downloads



Imaging Data Thumbnail Views



Track Reads (Partial or whole)



Overall positive trendlines across online data reuse and downloads over the past 5 years (>5y ago, 1K/y downloads)

Beyond basic research... clinical and commercial use cases

- ✓ Eyewear – industrial grade eyewear/helmets, require precise fit
- ✓ Masks – during COVID pandemic, fitted masks for children
- ✓ Patient counseling – pre-/post-operative severity score (CranioRate)
- ✓ Cosmetic surgery – inform the planning of surgical procedures
- ✓ Rare disease – requires ‘creative’ search for information
- ✓ ML models – predictions from models trained on GWAS data
- ✓ Public health – underrepresented populations, health disparities
 - References for clinicians – e.g., atlases for specific syndromes
 - Clinical genetics – craniofacial relevance of genes & variants
 - Facial scans apps – automated syndrome/phenotype identification

Conclusions

- Data availability is the major gating factor for modern scientific discovery and societal impact from data science
- It will take data-centric communities (data villages) to create the scale of data, often without clear prediction of how it will get used some day
- Partner with the community – rather than DIY or do-it-for-them
- Engage actively through domain science venues, “guiding”, socializing, training in better data stewardship and usage
- We have shown in FaceBase how we have partnered with and engaged our community, demonstrating a thriving community over 5, 10, + years

Acknowledgments

FaceBase Team

Alejandro Bugacov

Yang Chai (co-PI)

Jifan Feng

Joe Hacia

Thach-Vu Ho

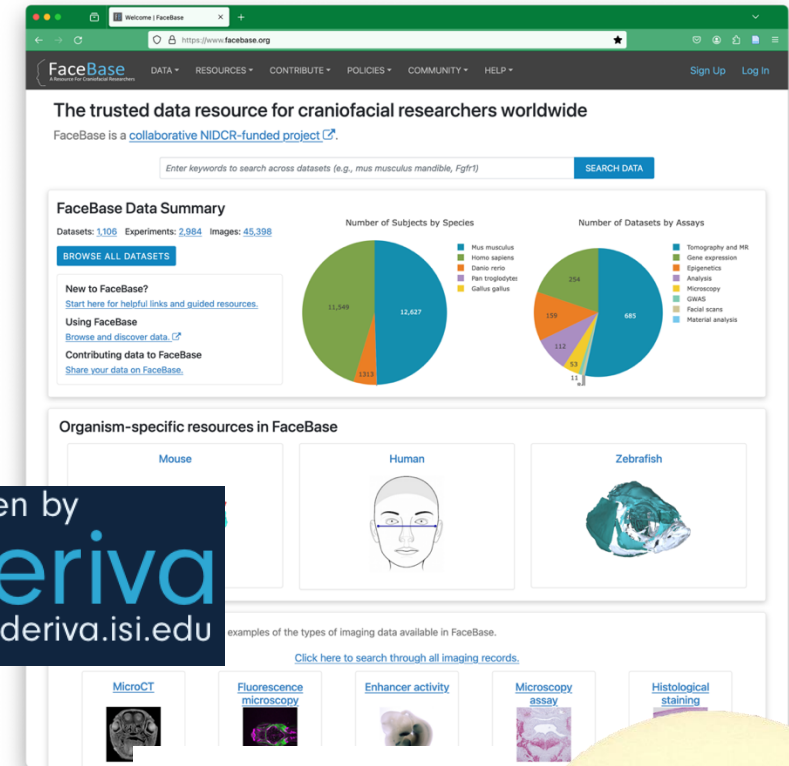
Carl Kesselman (co-PI)

VyVy Nguyen

Laura Pearlman

Rob Schuler

Cris Williams



Sponsors: NIH / NIDCR / ODSS (U01DE028729)

USC
Viterbi
*Information
Sciences Institute*

**Center for
Craniofacial
Molecular
Biology**

