NCATS

COLLABORATE. INNOVATE. ACCELERATE.

National COVID Cohort Collaborative (N3C)

Real World Data Roadmap

Background

Governance / Partnership

Data Acquisition/Data Enhancement

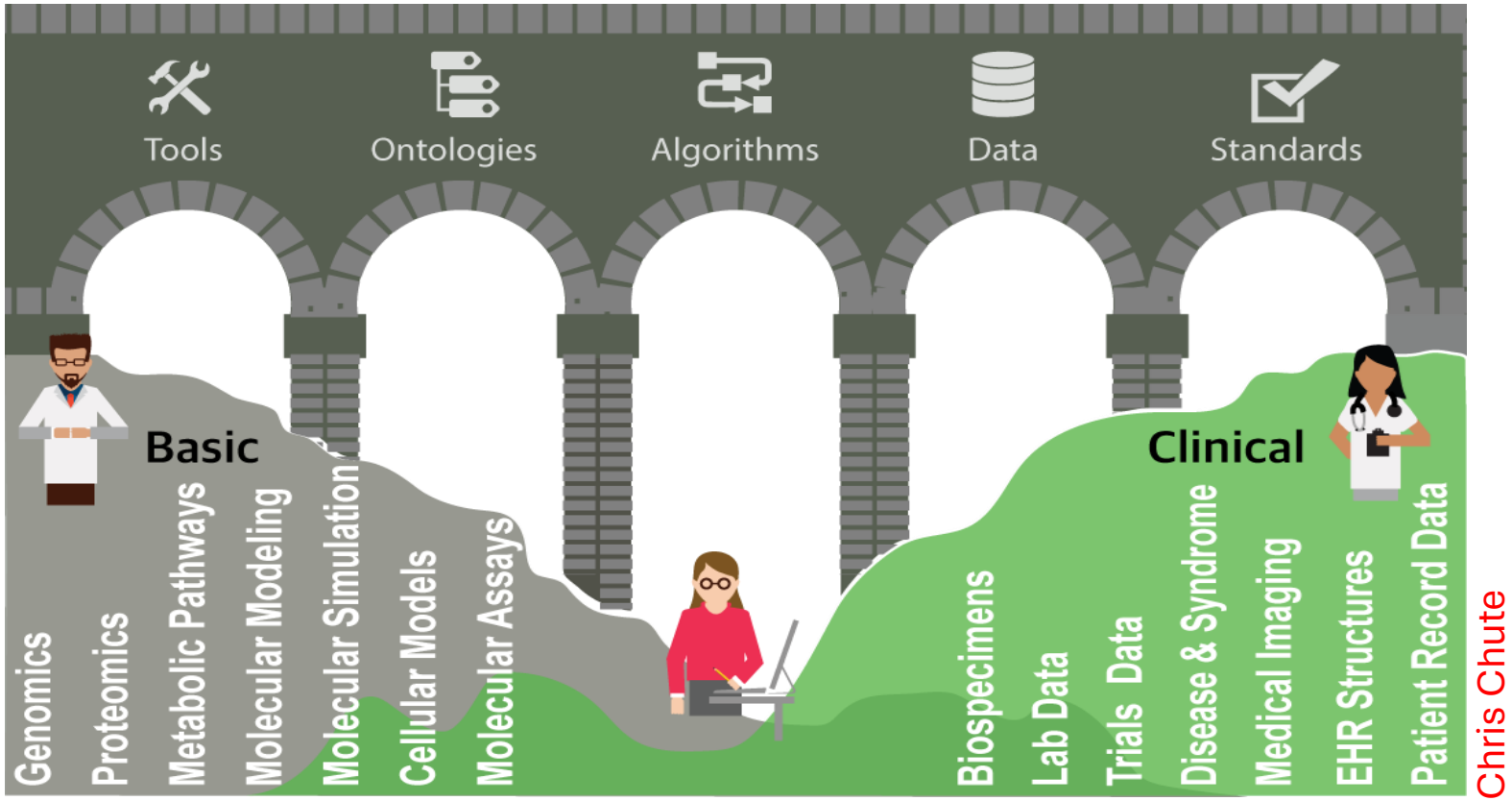
Harmonization, Concept Sets, Quality

Collaborative Analytics

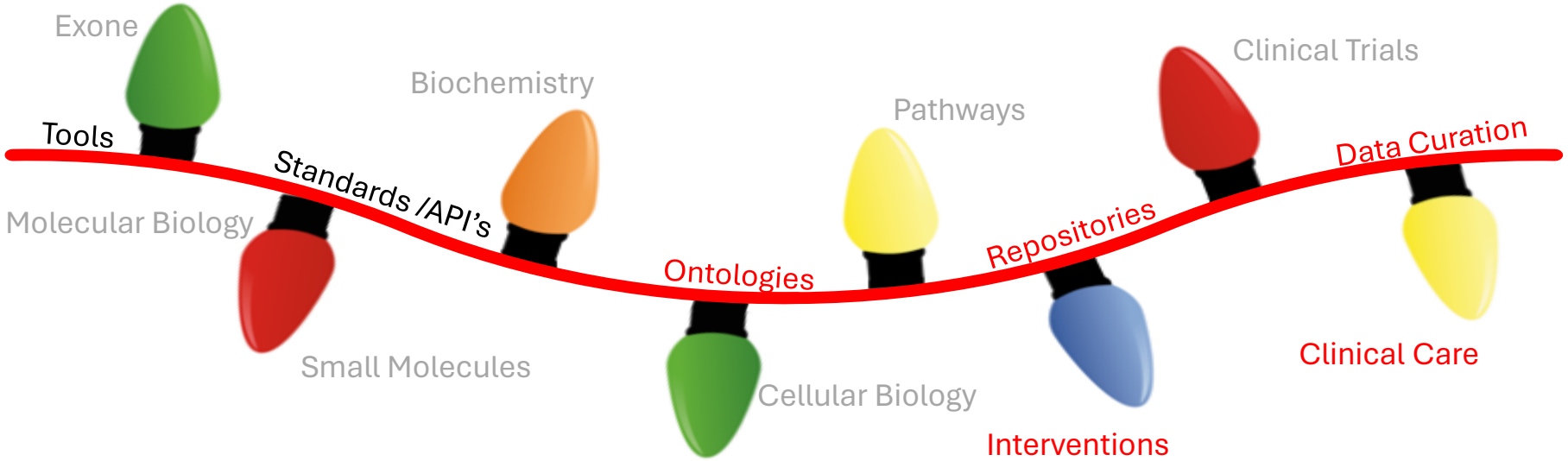
Education/Support

Partnerships

Informatics is the music between the notes



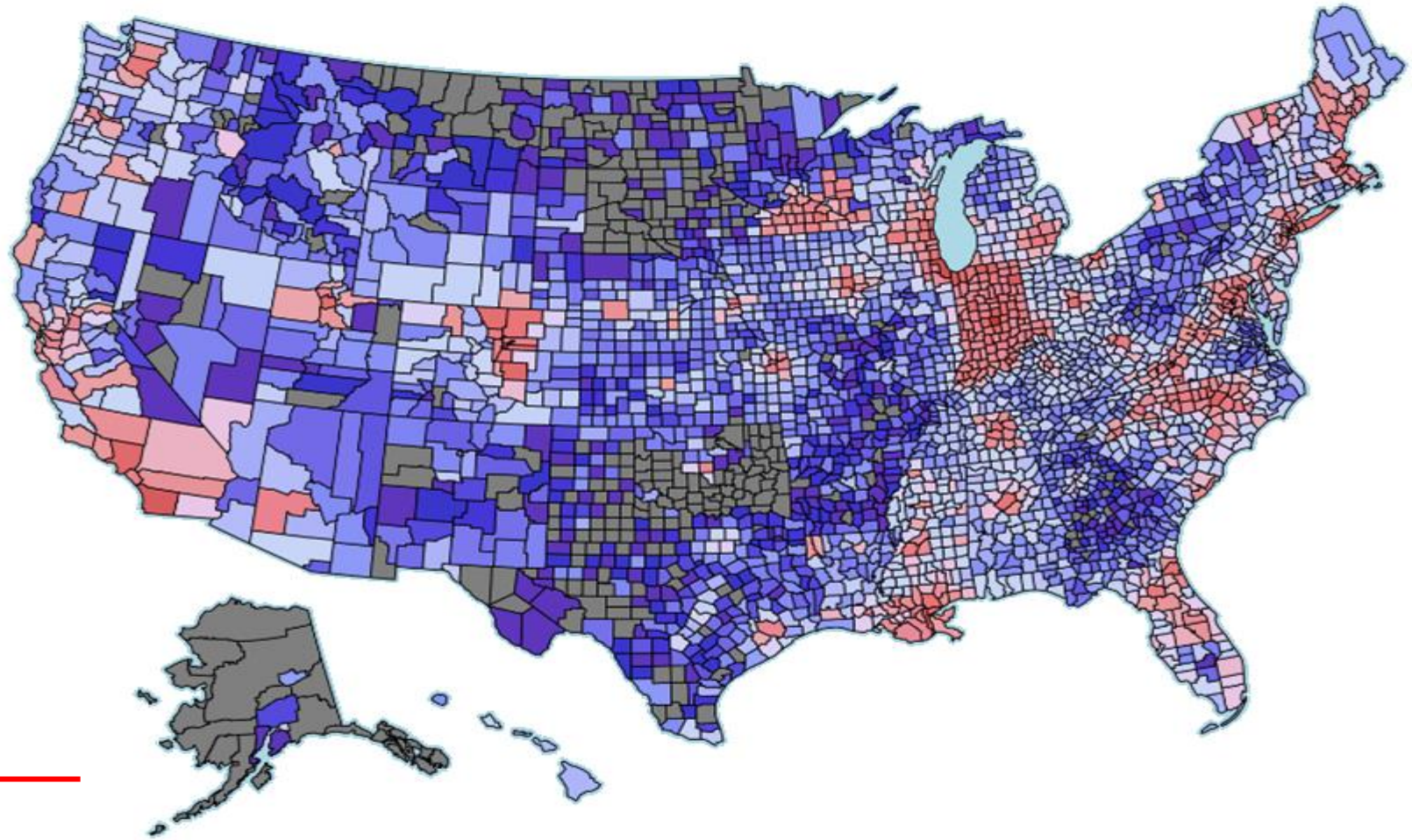
Chris Chute



National COVID Cohort Collaborative (N3C): (01/02/2024)

https://covid.cd2h.org/dashboard/

COVID+ CASES 8.5m	Total Patients 21.5m
Rows of Data 31.0b	Health System Contributors 83
Institutions Using Data 370	Active Investigators >3700
Research Studies 521	Citations & H-Index. 3324/28

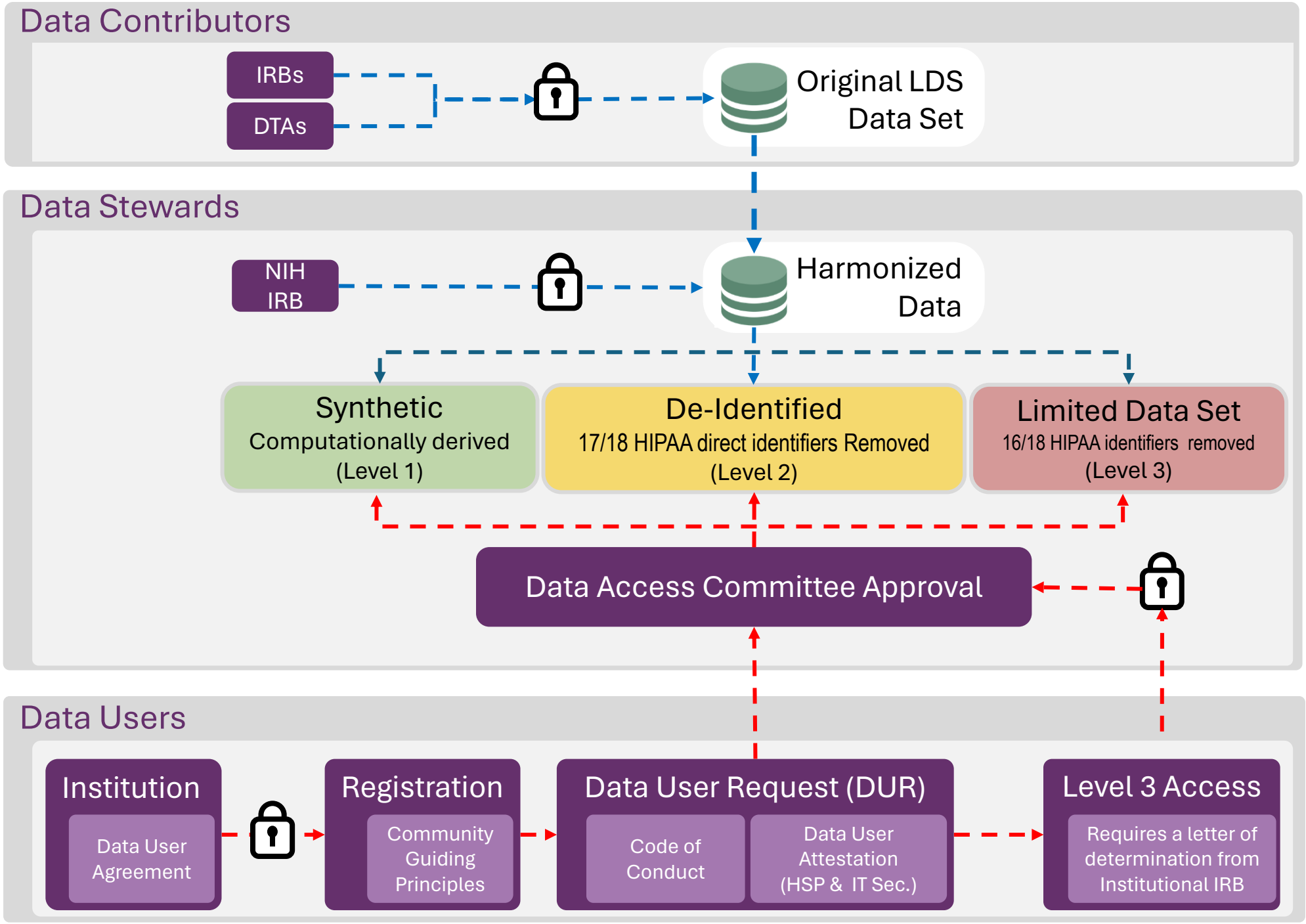


Geographics: 50/50 States >92% of all US Counties in USA
Representative of US population
Source: Community, Academic, FQHCs
Patient Mix: Inpatient ~20%, Outpatient ED ~80%
Longitudinal Data: 1/1/2018 to Present

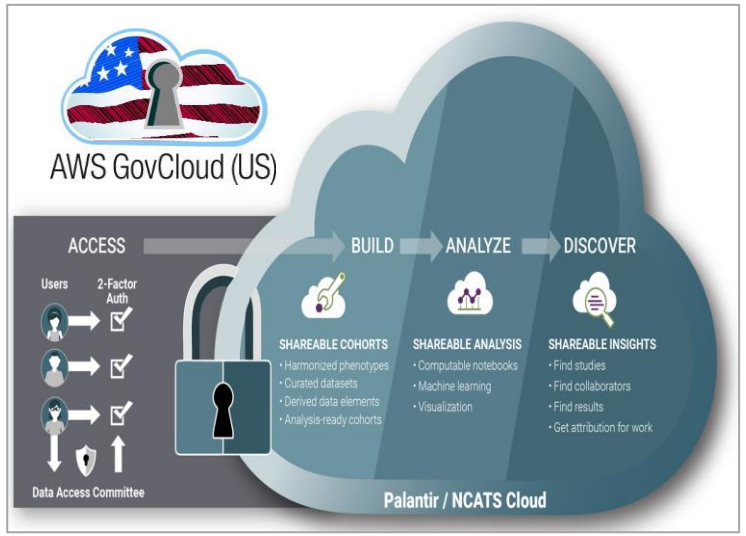
FQHCs/Commu CTR CTS

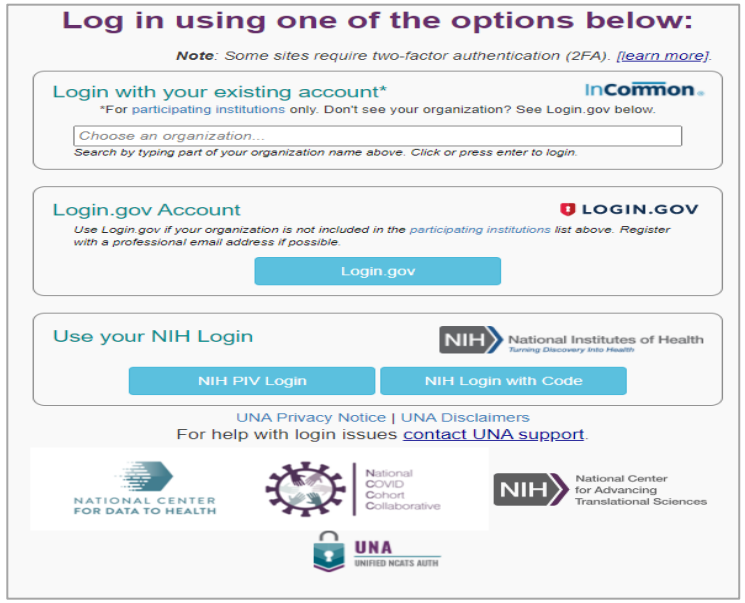
National COVID Cohort Collaborative

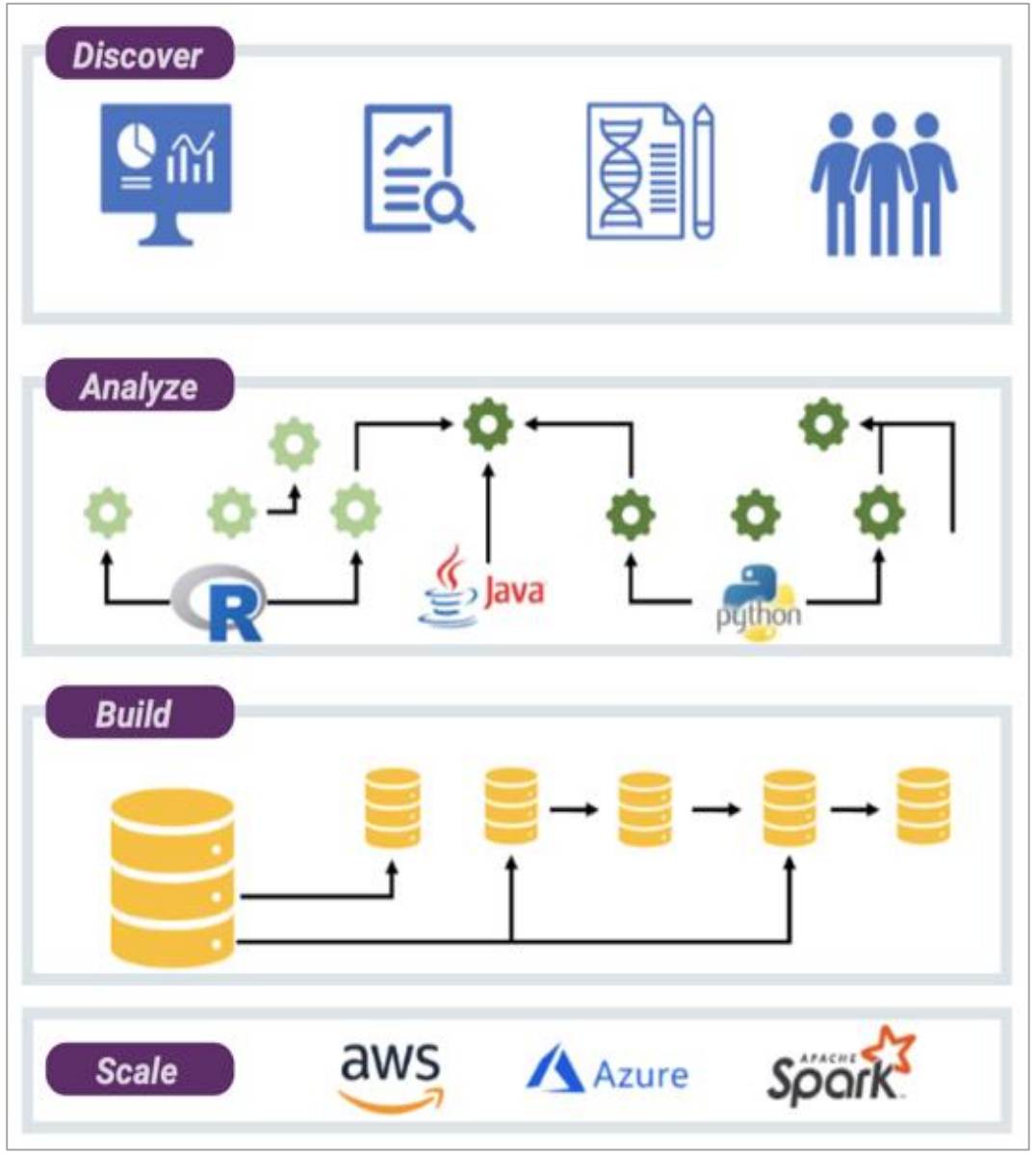
N3C: Governance and Access



Virtual Research Infrastructure







UNA UNIFIED NCATS AUTH

slack

HL7 FHIR

Palantir

zenodo

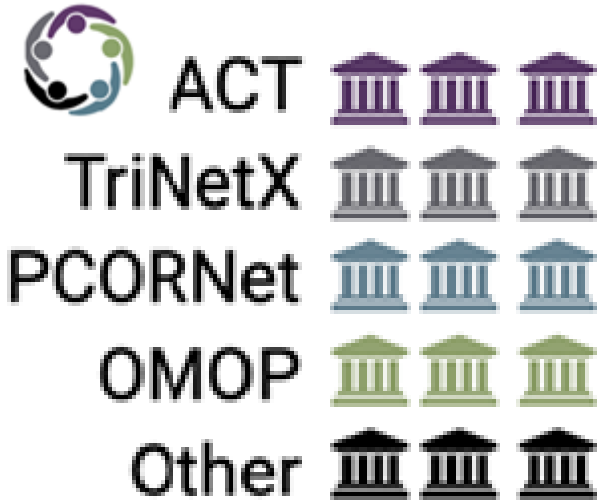
ATLASSIAN Jira Software

Google Workspace

N3C Data Lifecycle

Limited Data Set

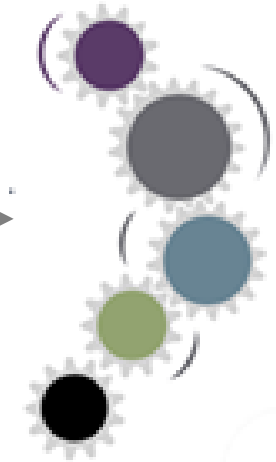
1. Data Partnership & Governance



FHIR

2. Phenotype & Data Acquisition

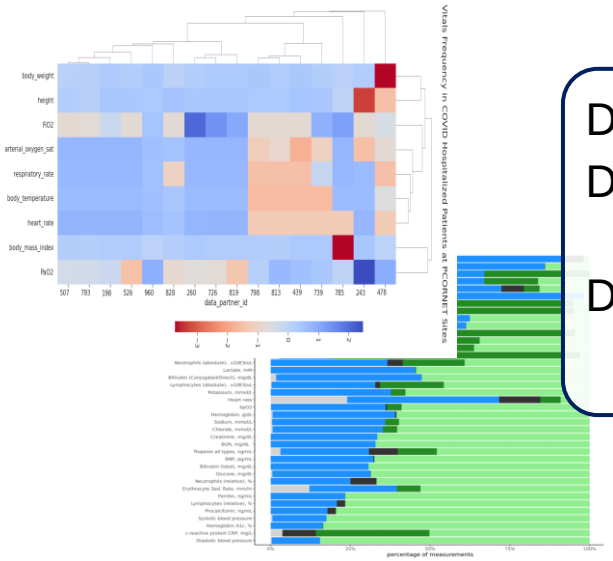
3. Data Ingest & Harmonization



4. Collaborative Analytics & Support

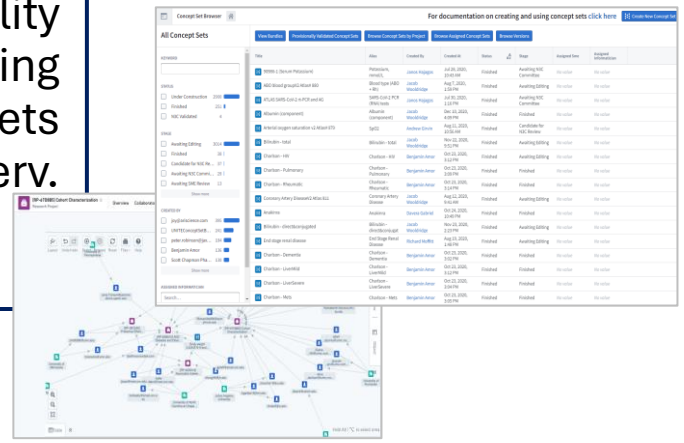
NCATS Cloud

From Real-World Data → Research Usable Knowledge



Data Quality
Data
Acquisitions
Data Transform

Reproducibility
Versioning
Code Sets
Vocab Serv.

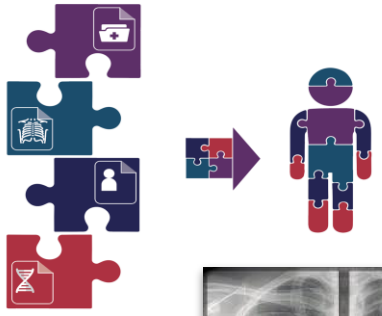


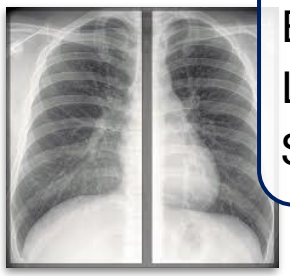
Data Quality Transforms

Validated Knowledge Objects

Data Enhances Missingness

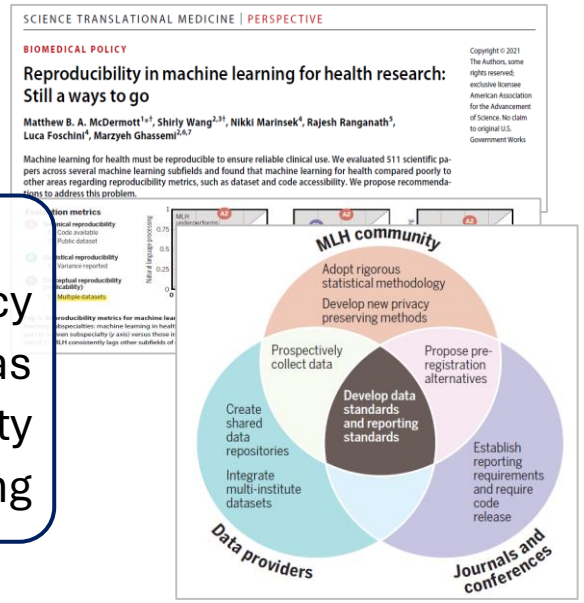
Good Algorithmic Practices





Requirements
External Datasets
Linkages
Site Permission

Transparency
Lack of Bias
Generalizability
Benchmarking



McDermott and colleagues, 2021

Data Enhancement and Patient Deduplication using PPRL

Privacy Preserving Record Linkage, (PPRL) is a de-identified linkage of individuals across different data sources that maintains privacy

- Patient Deduplication
- Multiplex Dataset Linkage
- Cohort Discovery

CMS Data

SDOH Data

Viral Variant

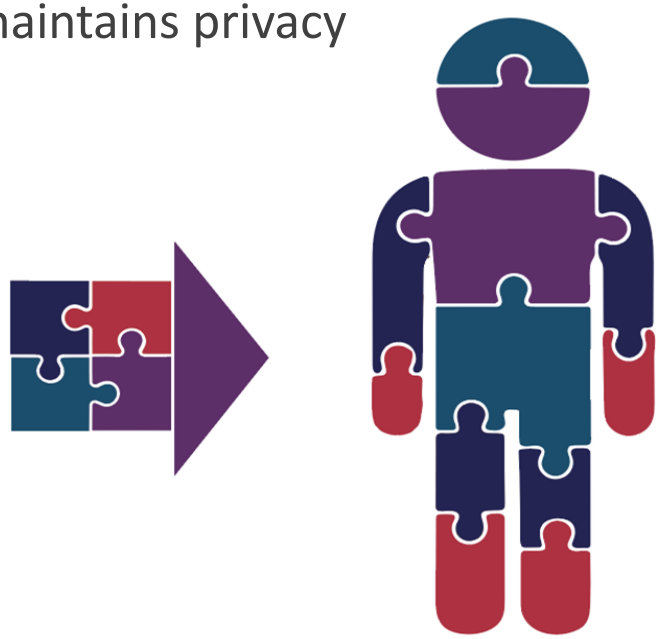
Mortality Data

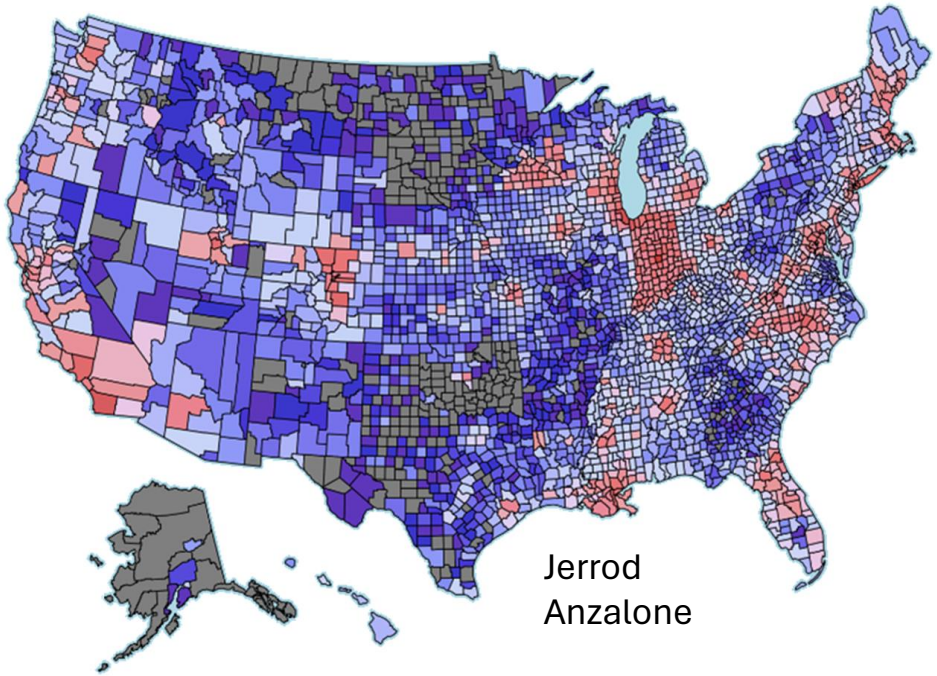
Clinical Data

Vaccine data

**Imaging MIDRC*

SEER, **SRTR*





Jerrod Anzalone

* Pending

Machine Learning and Good Algorithm Practice

A running list of reproducibility failures and overoptimistic claims in applied ML research

The list below consists of papers (especially systematic reviews) that highlight reproducibility failures or pitfalls in applied ML research. We distinguish applied ML research, where the goal is to use ML methods to study some scientific quest research, where the goal is to develop new ML methods, for example the typical NeurIPS paper. We are interested

Field	Paper	Year	Num. papers reviewed	Num. papers w/pitfalls	Pitfalls
Neuroimaging	Whelan et al.	2014	—	4	Incorrect trair
Autism Diagnostics	Bone et al.	2015	2	2	Biased evalua leakage
Bioinformatics	Blagus et al.	2015	—	6	Data leakage
Nutrition research	Ivanescu et al.	2016	—	4	Incorrect trair
Text Mining	Olorisade et al.	2017	30	—	Multiple pitfa
Medicine	Filho et al.	2018	1	1	Data leakage
Software engineering	Tu et al.	2018	58	11	Data leakage
Clinical epidemiology	Christodoulou et al.	2019	71	48	Biased evalua leakage

SCIENCE TRANSLATIONAL MEDICINE | PERSPECTIVE

BIOMEDICAL POLICY

Reproducibility in machine learning for health research: Still a ways to go

Matthew B. A. McDermott[1*†], Shirly Wang[2,3†], Nikki Marinsek[4], Rajesh Ranganath[5], Luca Foschini[4], Marzyeh Ghassemi[2,6,7]

Copyright © 2021 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works

Machine learning for health must be reproducible to ensure reliable clinical use. We evaluated 511 scientific papers across several machine learning subfields and found that machine learning for health compared poorly to other areas regarding reproducibility metrics, such as dataset and code accessibility. We propose recommendations to address this problem.

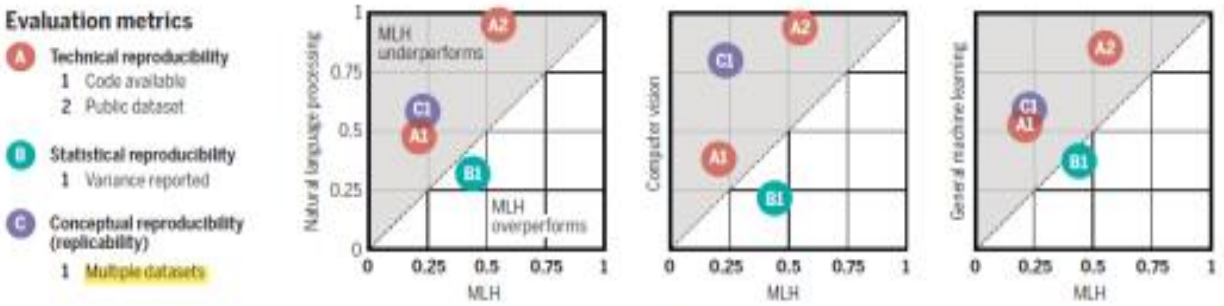
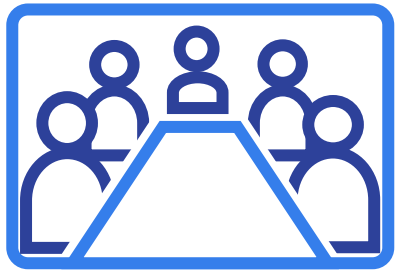


Fig. 1. Reproducibility metrics for machine learning applications. Shown are reproducibility metrics (A, B, and C) for evaluating scientific papers from four machine learning subspecialties: machine learning in health (MLH), natural language processing, computer vision, and general machine learning. Presented is the fraction of papers in a given subspecialty (y axis) versus those in MLH (x axis) that release their code (A1), release their data (A2), report their variance (B1), and leverage multiple datasets (C1). MLH consistently lags other subfields of machine learning on all measures of reproducibility apart from inclusion of proper statistical variance.

Educational Resource for Data Science (AIM-AHEAD, NCATS, NIGMS)

Lectures

Office Hours



Domain Teams

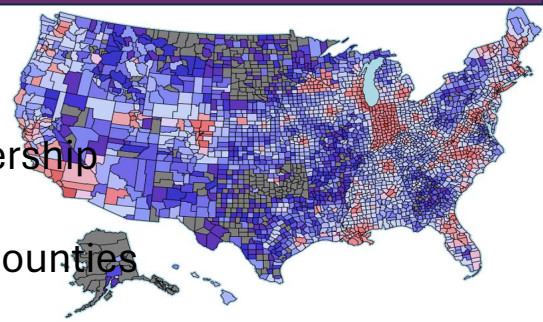
Assignments

Interactive

763+	27	39	9,600	Single
Training Resources Available	Training Modules	Training Webinars	Training video YouTube views	EHR RWD analytics best practices Guidebook

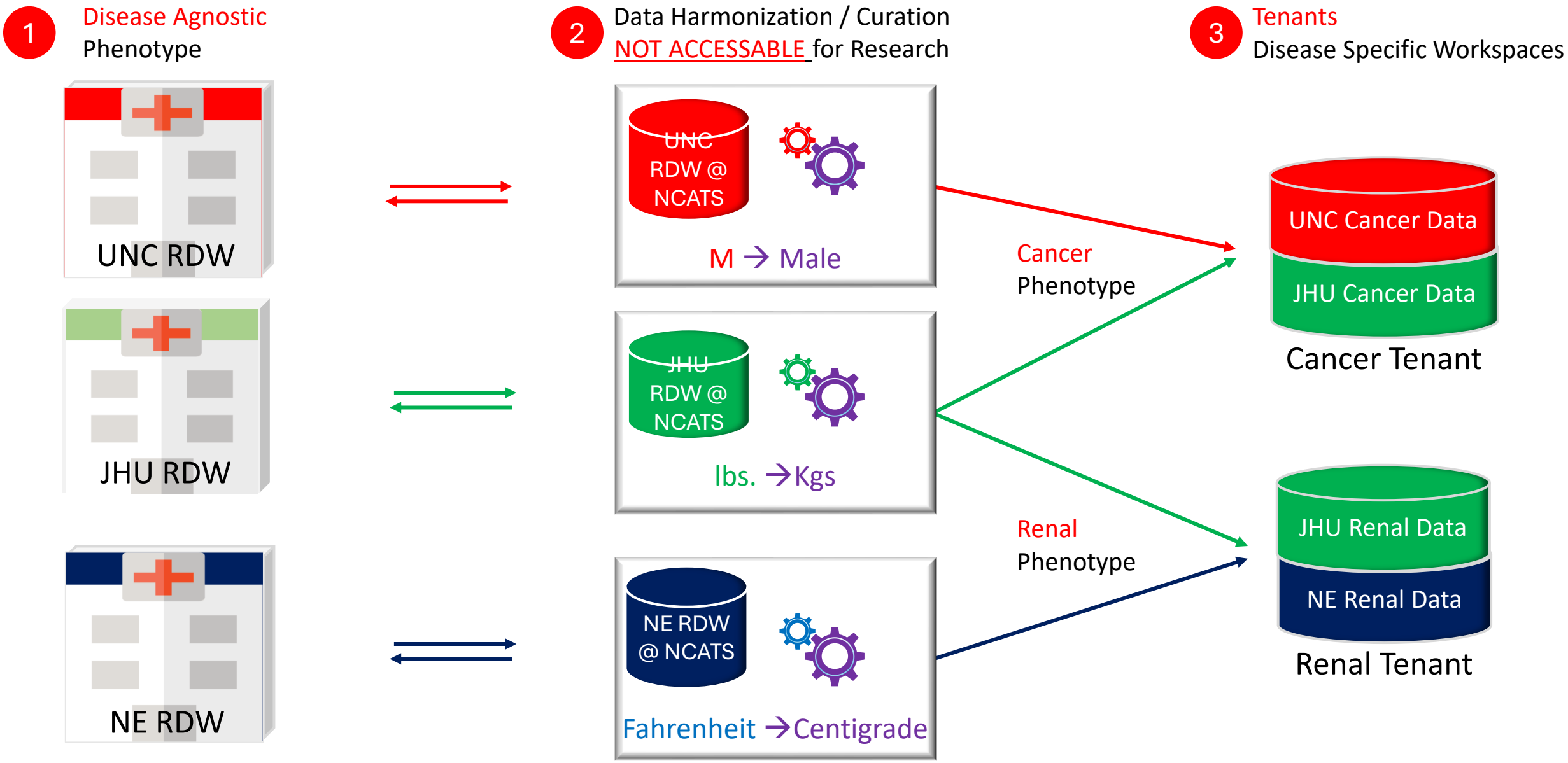
Available ML and RWD Training Resources

N3C Accomplishments & Partnerships (01/02/2024)

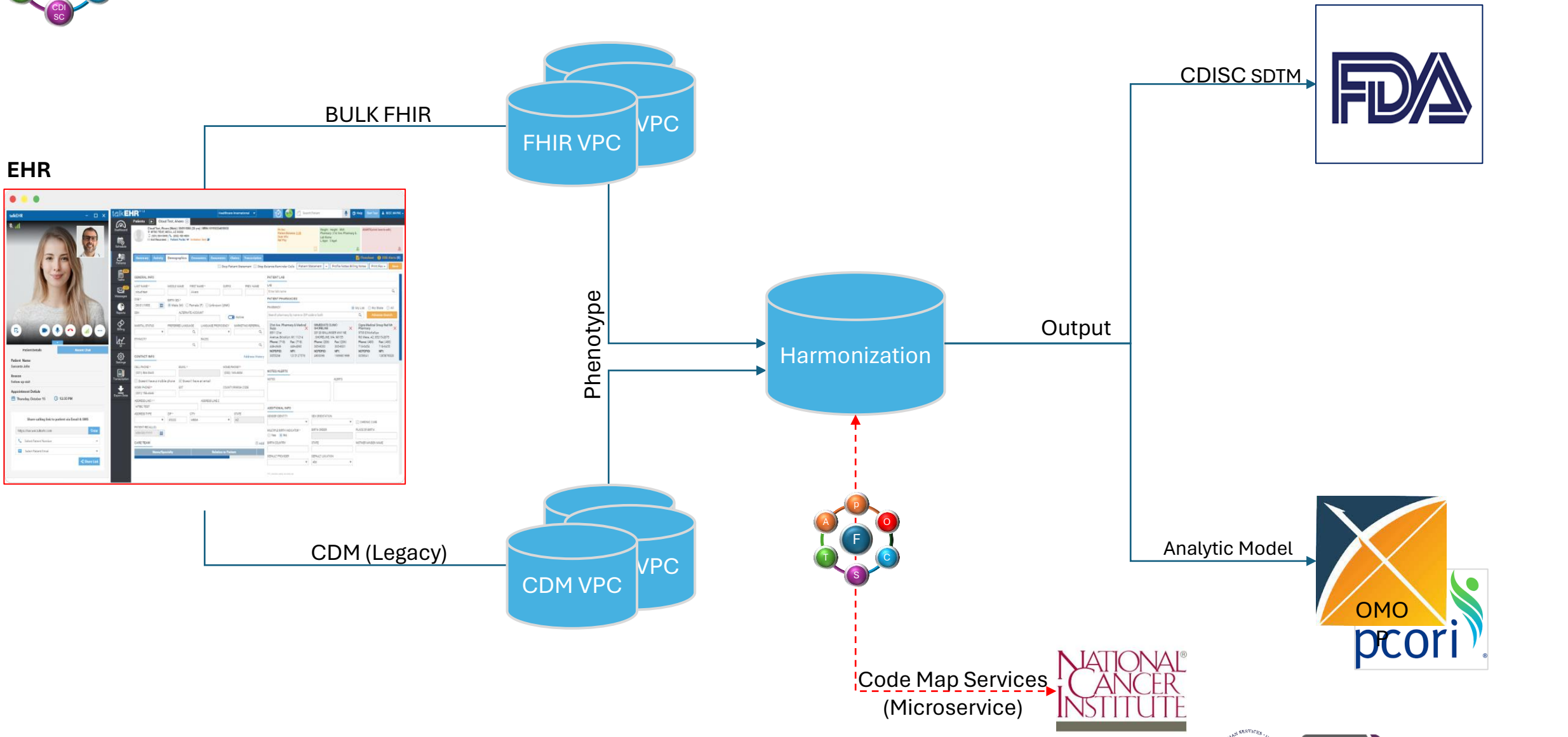


- **Team Science:** > 3600 Users, >500 studies, N3C Leadership is predominantly Women and Minority Leadership
- **Citations:** 3321 Citations, H-index 28, 126 Press Articles, 79 Publications
- **Largess:** Largest COVID repository in the USA >21M patients, 31B rows of data, 83 health systems, 92% Counties
- **Data Quality/Harmonization:** Score Card, Data Quality Checks, CTSA/CTR network harmonization
- **Inclusive Networks:** Only Network that includes: PCORNET, OMOP, ACT, TriNetX.
- **Education/support:** 763 training resources, personal help, office hours, best practice, tickets, website, newsletter, video, office hours, Domain Team, Forum
- **Partners:** ONC, FDA, NCI, ASPE, ASPR, AHRQ, NIBIB, All of Us, NHLBI, NIDDK, NIGMS, ARPA-H, ODSS, NAIRR, AIM-Ahead
- **Recognition:** Biden administration guidance, Dataworks! Grand prize, NIH director's blog, NPR, HHS Distinguished Award, The Journal of Rural Health 2023 “article of the year”
- **SDoH:** AI/AN, 60+ public data sets, CMS Medicare and Medicaid data, Collect 6 Gravity Domains
- **Funding:** 72 awards, $> 109,000,000 from CLC, NCATS, NHGRI, NIA, NIAID, NICHD, NIDA, NIDCD, NIDDK, NIGMS, NIMH, NIMHD, NLM
- **Other** Synthetic Data Validation, Machine Learning Validation (GAP), NAIRR

Tenants: Efficient, Scalable, Support Team Science, Data Interoperability, and Institutional Control



Code Map Services



National Center for Advancing Translational Sciences

Thank you!