# CTDS ANNUAL REPORT TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>INTRODUCTION</strong></td>
<td>4</td>
</tr>
<tr>
<td>Message from the Director</td>
<td>4</td>
</tr>
<tr>
<td><strong>OVERVIEW</strong></td>
<td>6</td>
</tr>
<tr>
<td>Center for Translational Data Science At a Glance</td>
<td>6</td>
</tr>
<tr>
<td><strong>RESEARCH</strong></td>
<td>8</td>
</tr>
<tr>
<td>Machine Learning and AI Research</td>
<td>8</td>
</tr>
<tr>
<td>Data Platforms Research</td>
<td>10</td>
</tr>
<tr>
<td>Cancer Research</td>
<td>12</td>
</tr>
<tr>
<td><strong>PROJECT HIGHLIGHTS</strong></td>
<td>14</td>
</tr>
<tr>
<td>BioData Catalyst</td>
<td>14</td>
</tr>
<tr>
<td>Biomedical Research Hub</td>
<td>15</td>
</tr>
<tr>
<td>Data Commons Framework</td>
<td>16</td>
</tr>
<tr>
<td>Geonomic Data Commons</td>
<td>17</td>
</tr>
<tr>
<td>HEAL Data Platform</td>
<td>19</td>
</tr>
<tr>
<td>Medical Imaging and Data Resource Center</td>
<td>20</td>
</tr>
<tr>
<td>Chicagoland COVID-19 Data Commons</td>
<td>21</td>
</tr>
<tr>
<td>Veterans Affairs Data Commons</td>
<td>22</td>
</tr>
<tr>
<td>Veterans Precisions Oncology Data Commons</td>
<td>23</td>
</tr>
<tr>
<td><strong>NEW PROJECTS</strong></td>
<td>24</td>
</tr>
<tr>
<td>CHORDS Resource Catalog</td>
<td>24</td>
</tr>
<tr>
<td>Gen3 Data Hub</td>
<td>25</td>
</tr>
<tr>
<td><strong>GEN3 SOFTWARE PLATFORM</strong></td>
<td>26</td>
</tr>
<tr>
<td>Recent Growth</td>
<td>26</td>
</tr>
<tr>
<td>Improvements to Gen3 in 2023</td>
<td>27</td>
</tr>
<tr>
<td>Minting DOIs</td>
<td>27</td>
</tr>
<tr>
<td>Easier deployment</td>
<td>27</td>
</tr>
<tr>
<td>Running complex analytics</td>
<td>28</td>
</tr>
<tr>
<td>Improved metadata indexing</td>
<td>28</td>
</tr>
<tr>
<td><strong>NEW PUBLICATIONS</strong></td>
<td>29</td>
</tr>
<tr>
<td><strong>NEW KEY PERSONNEL</strong></td>
<td>31</td>
</tr>
<tr>
<td>Phil Schumm</td>
<td>31</td>
</tr>
<tr>
<td><strong>FUNDING SOURCES</strong></td>
<td>32</td>
</tr>
<tr>
<td><strong>OTHER NEWS AND UPDATES</strong></td>
<td>33</td>
</tr>
<tr>
<td>Super Computing Conference</td>
<td>33</td>
</tr>
<tr>
<td>KubeCon and CloudNativeCon</td>
<td>34</td>
</tr>
<tr>
<td>Internship Program</td>
<td>34</td>
</tr>
<tr>
<td>Gen3 Community Forum</td>
<td>35</td>
</tr>
<tr>
<td>Biomedical Research Hub Named as</td>
<td>35</td>
</tr>
<tr>
<td>GA4GH Driver Project</td>
<td>36</td>
</tr>
<tr>
<td><strong>NEW PUBLICATIONS</strong></td>
<td>29</td>
</tr>
<tr>
<td><strong>NEW KEY PERSONNEL</strong></td>
<td>31</td>
</tr>
<tr>
<td><strong>FUNDING SOURCES</strong></td>
<td>32</td>
</tr>
<tr>
<td><strong>OTHER NEWS AND UPDATES</strong></td>
<td>33</td>
</tr>
<tr>
<td><strong>NEW PUBLICATIONS</strong></td>
<td>29</td>
</tr>
<tr>
<td><strong>NEW KEY PERSONNEL</strong></td>
<td>31</td>
</tr>
</tbody>
</table>
Dear Colleagues,

The Center for Translational Data Science (CTDS) had a productive year in 2023 and made important progress on our core mission of harnessing data science to advance the fields of biology, medicine, health care, and the environment. You can see our 2023 research publications on our CTDS web site. In terms of our goal of accelerating research around the world through secure data sharing and cloud-based data exploration and analysis, we added new capabilities to our existing projects, brought new projects online, and added new features to the core underlying Gen3 software platform.

Here are a just a few highlights of 2023:

- We made important progress adding data mesh / data fabric services to Gen3. The HEAL data platform is our largest mesh and we will running a workshop in 2024 called Accelerating Research through Data Meshes and Data Fabrics (ARDM-24)
- The Biomedical Research Hub was named a new GA4GH Driver Project.

- We developed a new frontend for the GDC and Gen3, which will be released in 2024.

- We added 1.2M new files / 4.8PB of data to Gen3 commons managed by CTDS

- We added NextFlow, a workflow execution system, to Gen3, which allows for complex and long running analyses with our workspaces.

- Five new third party organizations adopted Gen3 in 2023.

The progress in this report is due to the hard work of dozens of CTDS researchers and staff as well as our many other collaborators and contributors from around the world. It is also due to our sponsors and the enormous opportunity they see in how data commons, and Gen3 in particular, can have a positive impact on research. Thanks also to the readers of this report for your interest in CTDS and our mission of applying data science to hard questions in biomedical research. We look forward to sharing our updates from 2024 with you next year!

[Signature]

Robert Grossman, PhD

Center for Translational Data Science Director
ACTIVITY OF CTDS

The Center for Translational Data Science at the University of Chicago is developing the discipline of data science and its applications to problems in biology, medicine, healthcare, and the environment.

We create and operate large-scale data platforms to support research in topics of societal interest, including cancer, cardiovascular disease, inflammatory bowel disease (IBD), birth defects, veterans’ health, pain management, opioid use disorder, and environmental science. We also develop new machine learning and AI algorithms over the data in our platforms.

Today with our partners, we operate a data ecosystem comprising over 30 data commons that make over 22 PB of data available to the research community from nearly 2 M patients.

We provide access to this data via secure and compliant workspaces, while protecting patient privacy. These are all based on the open-source Gen3 data platform, that includes data commons, framework and mesh services, and workspaces.
Gen3 Data Commons are used by a long list of sponsors in order to enable secure and compliant data sharing and analysis.
Enhancing Instance-Level Image Classification with Set-Level Labels.

Algorithm 1 FACILE algorithm

1: **Input**: loss functions $\ell_{fg}$, $\ell_{cg}$, predictors $E$, $G$, $F$, datasets $D^c_m$ and $D^g_n$
2: obtain feature map $\hat{e} \leftarrow A(E, G, F, D^c_m, D^g_n)$
3: create dataset $D^{fg,\text{aug}}_n = \{(z_i, y_i) : z_i = \hat{e}(x_i), (x_i, y_i) \in D^g_n\}_{i=1}^n$
4: obtain fine-grained label predictor $\hat{f} \circ \hat{e}$, where $\hat{f} \leftarrow A(E, G, F, D^{fg,\text{aug}}_n)$
5: **Return**: $\hat{f} \circ \hat{e}$

Summary: A large amount of labeled data is typically required in traditional machine learning approaches. However, obtaining high quality fine-grained labels for all samples is often challenging or infeasible in real-world scenarios. In this paper, we introduce Fine-grAined representation learning from Coarse-graIned LabEls (FACILE), which allows us to build models to predict image specific labels using only a few examples of each fine-grained label. Our algorithm leverages more widely available coarse-grained image labels to pretrain the model to learn set-level representations, for example "fruit" for images of apples, oranges, and bananas. We apply our new approach to natural image and histopathology image datasets. Notably, our algorithm achieves 13% improvement in classification accuracy compared to the strongest baseline on the histopathology image classification benchmarks.

ATAT: Automated Tissue Alignment and Traversal.

Summary: Spatial transcriptomics is a powerful new tool that collects high resolution and spatially resolved transcription data within a tissue. It allows for a deeper understanding of cellular interactions than traditional bulk RNA-Seq can allow. However, its utility is currently limited by the requirement for pathologists to annotate and delineate tissues of interest on the source histology slide. In this paper, we introduce ATAT: Automated Tissue Alignment and Traversal, an algorithm which linearizes transcription data into an expression trajectory which spans tissue layers. We utilize self supervised contrastive learning on hematoxylin and eosin (H&E) stained images to teach models the gradual morphological transitions across tissues. With our approach, we automate the integration of H&E images with spatial transcriptomics and simplify the investigation of important biomedical questions, such as characterization of inflammatory conditions across intestinal walls.

Towards self-describing and FAIR bulk formats for biomedical data.

**Summary:** There is currently no widely accepted standard for archiving and storing of bulk structured data. Structured data is typically found in a database and may include data such as Electronic Health Records (EHR), experimental data, or clinical trial records. We introduce a self-describing format based upon Avro for bulk structured biomedical data called the Portable Format for Bioinformatics (PFB) that encapsulates a data model, a data dictionary, the data itself, and pointers to third party controlled vocabularies. Benefits of the PFB format include long term storage of structured data, sharing of structured data, conversion between data standards due to the controlled vocabulary pointers, and versioning.

CNT: Semi-Automatic Translation from CWL to Nextflow for Genomic Workflows.

**Summary**: Analysis of biomedical data (e.g. genomic, proteomic, imaging) can be very complex including the use of many tools and custom data processing tasks. Workflow tools have become very commonplace because they offer consistent processing of batches of data, data provenance, and ease of sharing. However, there are a variety of distinct workflow languages currently available and there is no simple way to convert from one to another. For example, based on our experience, translating a production-grade pipeline, such as the GDC DNA-Seq workflow from CWL to Nextflow, takes approximately 160 person-hours. To address this challenge, we present CNT, a novel, semi-automated translator converting CWL workflows into Nextflow ones. We evaluated CNT with production-grade workflows and found that it can cover up to 81% of the original workflows, substantially reducing development time.


<table>
<thead>
<tr>
<th></th>
<th>DNA-Seq</th>
<th>RNA-Seq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speedup over cwltool [5]</td>
<td>21-39%</td>
<td>33-72%</td>
</tr>
<tr>
<td>CPU Utilization Increase</td>
<td>18-33%</td>
<td>45-85%</td>
</tr>
<tr>
<td>Translation Coverage</td>
<td>73%</td>
<td>81%</td>
</tr>
</tbody>
</table>

**TABLE I. Summary of evaluation results (§1).**
**Summary:** Veterans are at an increased risk for prostate cancer compared with the general population. Using clinical and targeted tumor sequencing data from the National Veterans Affairs health system, we conducted a retrospective cohort study on 45 patients with advanced prostate cancer in the Veterans Precision Oncology Data Commons (VPODC). We characterized the mutational burden in this cohort and conducted unsupervised clustering analysis to stratify patients by molecular alterations. The clustering approach presented in this study can potentially be used to clinically stratify patients based on their distinct mutational profiles and identify actionable somatic mutations for precision oncology.

Summary: Liquid biopsy holds great promise as a non-invasive way of monitoring cancer disease progression or recurrence. The Blood Profiling Atlas in Cancer (BLOODPAC) Consortium is a collaborative effort involving stakeholders from the public, industry, academia, and regulatory agencies focused on developing shared best practices on liquid biopsy. This report describes our recent efforts to develop standards on the use of contrived materials mimicking cell-free circulating tumor DNA, to comparatively evaluate clinical laboratory testing procedures. We present on how different laboratory practices and technologies show broad concordance in measuring variant allele frequencies in a known contrived sample. These results support the practice of using a shared known and well understood DNA standard across technologies as a way to objectively measure the impact of changes in laboratory and analysis protocols.

The NHLBI BioData Catalyst program is a cloud-based ecosystem providing tools, applications, and workflows in secure workspaces. By increasing access to NHLBI datasets and innovative data analysis capabilities, BioData Catalyst accelerates efficient biomedical research that drives discovery and scientific advancement, leading to novel diagnostic tools, therapeutics, and prevention strategies for heart, lung, blood, and sleep disorders.

The BioData Catalyst Gen3 Platform provides data commons services through authentication/authorization, object file indexing, interactive data search and export, and analytical workspaces services. Partner organizations and approved researchers can search and access hosted genomic and phenotypic data, and export selected cohorts to analytical workspaces in a scalable, reproducible, and secure manner.

Accomplishments in 2023 included ingesting over 110 TB of data (from TOPMed, COVID-19 C4R and ACTIV, BIOLINCC NSRR Sleep Study, and numerous other smaller project updates). We also brought in additional metadata from each new project from dbGaP and architected many new solutions, including DOIs, to be released in 2024.

The BioData Catalyst provides access to data from programs across the National Heart Lung and Blood Institute (NHLBI). Data can be explored and analyzed within the system.
BIOMEDICAL RESEARCH HUB

The Biomedical Research Hub (BRH) is a cloud-based federated system for managing, analyzing, and sharing patient data for research purposes, while allowing each resource to operate their component based upon their own governance rules. BRH currently interoperates with 11 separate Data Commons.

BRH uses Gen3 mesh framework services including authentication and authorization; services for generating and assessing findable, accessible, interoperable, and reusable (FAIR) data; and services for importing and exporting bulk clinical data. BRH includes analytical workspaces that can access and analyze data from one or more of the data resources in the BRH.

In 2023, BRH was named a GA4GH driver project (see News section) and implemented Direct Pay, which allows for credit card payments to support workspace analytics. The Direct Pay feature is operated in partnership with the Open Commons Consortium (OCC). BRH also supports payment using STRIDES credits from NIH.
The Data Commons Framework plays a central role in connecting the different resources within the Cancer Research Data Commons.

The National Cancer Institute’s Data Commons Framework Services (NCI DCFS) is a set of software services designed to make it easier to develop, operate, and interoperate data commons, data clouds, knowledge bases, and other resources. DCFS enables managing, analyzing, and sharing cancer research data that are part of the Cancer Research Data Commons (CRDC).

The NCI DCFS supports making data Findable, Accessible, Interoperable, and Reusable (FAIR). Data objects are assigned GUIDs and can be stored in one or more private and public clouds and accessed using DCF services. Structured data can be incorporated using data models and enriched with controlled vocabularies and ontologies. Gen3 includes authentication and authorization services so that controlled-access data can be handled securely. Gen3 DCF services also include the ability to define, import, and query against a data model.

Achievements in 2023 included onboarding the Clinical Trials Data Commons (CTDC) and the Immuno-Oncology Data Commons (IODC) to the CRDC, improving data indexing performance, and adding over 44M files and 5.17 PB of data.
The NCI Genomic Data Commons provides harmonized cancer genomic and clinical data to the research community. NCI’s Genomic Data Commons (GDC) is the largest commons managed by CTDS and provides the cancer research community with a harmonized data repository and cancer knowledgebase that enables data sharing across cancer genomic studies in support of precision medicine.

The GDC supports several cancer genome programs including The Cancer Genome Atlas (TCGA), Therapeutically Applicable Research to Generate Effective Treatments (TARGET), Clinical Proteomic Tumor Analysis Consortium (CPTAC), Multiple Myeloma Research Foundation (MMRF), and many others. It includes clinical metadata and a range of raw and derived files from a variety of experimental strategies including whole genome sequencing (WGS), RNA-Seq, whole exome sequencing (WXS), miRNA-Seq, slide images, and many others.

In 2023, the GDC team achieved many important milestones. There were 12 data releases including adding over 6PiB of additional harmonized data. The GDC portal averaged 80,000 unique monthly users who downloaded from 2 to 5PiB in total every month. The Legacy Archive, which stored data from the GDC’s predecessor system, CGHub, was officially retired. And finally, the team made important progress on the new user interface (GDC 2.0), which will be released in early 2024.
The oncomatrix shows allows researchers to visualize the top 50 mutated genes affected by high impact mutations.

The NCI Genomic Data Commons provides harmonized cancer genomic and clinical data to the research community.
The HEAL Data Platform provides a secure environment for discovery and analysis of data generated by NIH HEAL-funded studies. The Platform represents a data ecosystem, or mesh, that aggregates metadata from multiple resources, provides search and query to facilitate data discovery, and interoperates with data repositories to permit secure and easy access to data comprising a diverse array of data types.

The HEAL Data Platform also offers a secure, STRIDES-enabled cloud-computing environment for data analysis, providing researchers with easy access to powerful commercial and open-source software tools, and the ability to perform scalable and collaborative analyses. Computational notebooks and other code and results can be shared with the research community to facilitate reproducibility and to increase visibility of publications with insights that address the opioid crisis.

Milestones in 2023 included interoperating with several new, third-party data repositories (currently including ICPSR/NAHDAP, Vivli, JCOIN Data Commons, and dbGaP); enabling DOI assignment for HEAL study records; several visual and functional enhancements to the Discovery page; refactoring the HEAL Metadata Service (MDS) to facilitate its usability and expansion; and improving support for variable-level metadata (VLMD) creation, submission and management. The HEAL data platform averaged over 500 users per month.
The **Medical Imaging and Data Resource Center (MIDRC)** is a multi-institutional collaborative initiative driven by the medical imaging community yielding an open curated, diverse commons for medical imaging AI research and a sequestered one for translation. MIDRC is funded by the National Institute of Biomedical Imaging and Bioengineering (NIBIB), hosted at the University of Chicago, and resides on the Gen3 platform.

It is co-led by the American College of Radiology® (ACR®), the Radiological Society of North America (RSNA), and the American Association of Physicists in Medicine (AAPM). The aim of MIDRC is to foster machine learning innovation via data sharing through rapid and flexible collection, curation and harmonization, analysis, and dissemination of imaging and associated clinical data by providing researchers with unparalleled resources in the fight against COVID-19. CTDS achievements for MIDRC in 2023 included the release of over 60,000 curated imaging studies from over 45,000 patients; we incorporated sequestration in the pipeline within Gen3, where representative data are partitioned between open and validation data sets; we began utilizing Gen3 DOI capabilities for publication linked data sets; and updated the OHIF DICOM viewer to v3.

For 2023, MIDRC was selected as an early Performer in the ARPA-H Biomedical Data Fabric (BDF) Toolkit with the aim to extend MIDRC to oncology cases. In addition, it was chosen as one of the initial pilots within NAIRR and also, received a DataWorks Prize from the FASEB and NIH Office of Data Science Strategy.
The Chicagoland COVID-19 Commons (CCC) is the first regional data commons launched under the Pandemic Response Commons Consortium. Born out of a consortium of Chicago-area civic and healthcare organizations with support from several technology partners, the CCC represents a persistent data resource for the research community engaging with COVID-19. The founding members of the PRC consist of hospital and nonprofit partners who agree to share and co-locate deidentified patient data and convene to ensure this data can inform predictive models related to diagnostics, treatment and the overall behavior of the virus. In 2023, the Chicagoland COVID-19 Commons built on its previous success by bringing together patient-level de-identified data from Chicago-area medical centers for the purpose of studying long covid.

The Chicagoland Covid-19 Commons, part of the Pandemic Response Commons, brings together public surveillance data related to the COVID-19 pandemic along with de-identified patient data from area medical systems.
The VA Data Commons supports the research and analysis of medical and genomic data from US military Veterans. It aims to accelerate scientific discovery and development of therapies, diagnostic tests, and other technologies for improving the lives of Veterans and beyond. The data commons features GWAS analyses using genetic data from veterans with clinical variables harmonized to the OMOP Common Data Model.

Accomplishments in 2023 included a new GWAS results app that included new features such as an interactive Manhattan plot, pop-up QQ plot, top loci table, and the ability to download results in zip format. Updated data from the Million Veterans Program was loaded. Also in 2023, over 3500 GWAS workflows were analyzed in the platform and additional VA researchers onboarded.
The **VPODC** supports the management and analysis of Veteran oncology data for the research community. This data commons aims to accelerate discovery and development of therapies, diagnostic tests, and other technologies for precision oncology. It supports cross-project analyses by data harmonization through the collaborative development of a data dictionary, providing an API for data queries, and providing an analysis workspace with rich tools and resources. In 2023 the research team published a paper on the stratification of prostate cancer patients based on their tumor mutational profiles.

Hernandez et al. 2023 Molecular Case Studies

The Veterans Precisions Oncology Data Commons distributes clinical and genomic oncology data to researchers with the goal of improving veterans health.
The National Institute of Environmental Health Sciences (NIEHS) has launched the CHORDS Platform to support the Climate and Health Outcomes Research Data Systems (CHORDS) project. The CHORDS Resource Catalog will centralize metadata from a range of federal, state, local, and academic projects and allow users to easily discover and locate resources of interest. This commons represents a unique partnership wherein the commons exists within the CTDS security boundary, but with a majority of operations managed by NIEHS including data and metadata uploads, data dictionary model modifications, portal configurations, and other non-security-related activities.

CHORDS PLATFORM

The National Institute of Environmental Health Sciences (NIEHS) has launched the CHORDS Platform to support the Climate and Health Outcomes Research Data Systems (CHORDS) project. The CHORDS Resource Catalog will centralize metadata from a range of federal, state, local, and academic projects and allow users to easily discover and locate resources of interest. This commons represents a unique partnership wherein the commons exists within the CTDS security boundary, but with a majority of operations managed by NIEHS including data and metadata uploads, data dictionary model modifications, portal configurations, and other non-security-related activities.

The CHORDS Resource Catalog will create a catalog of climate change and health resources from a range of federal, state, local, and academic projects, allowing users to easily discover and locate resources of interest.
The Gen3 Data Hub contains both open-access projects as well as archived data from sunsetted data commons and repositories. The Open Access Data Commons (OADC) has been relaunched as the Gen3 Data Hub (G3DH). In the past, the OADC housed only open-access data. The G3DH will now also serve as an archive for any Gen3 data commons that have been sunsetted due to the end of a project. Such data will be available in either a zip or PFB (Portable Format for Bioinformatics) format and may be open or controlled access depending on the type of data. This will serve as an important asset to the research community as research projects inevitably come to an end and an economical and scalable method is needed to make sure the data can remain available into the future.
Gen3 is an open-source and free to use data platform for building data commons and data meshes. It allows users to store and retrieve rich metadata and files, provide authentication and authorization for managing access, and supports both API and graphical user interface-based exploration. CTDS is the maintainer for Gen3 and uses this product as the basis for most of the projects it develops or manages.

RECENT GROWTH

Gen3 continues to support data sharing and analysis with new projects being added in 2023. An important recent change is the number of commons that are managed by external organizations has started to grow substantially. This is due to the versatility and power of Gen3, but also due to improvements in the ease with which Gen3 can be deployed and maintained. In upcoming years we look forward to an exponential growth in the use of Gen3 by external organizations and will continue to simplify Gen3 deployment and maintenance.
IMPROVEMENTS TO GEN3 IN 2023

MINTING DOIs

Digital object identifiers (DOIs) promote research reproducibility and data FAIR-ness by providing a persistent identifier in a format standardized by the International Organization for Standardization (ISO). DOIs are in wide use mainly to identify academic, professional, and government information, such as journal articles, research reports, data sets, and official publications. In Gen3, you can now mint a datacite DOI for a file or group of files by using our python SDK tool. The DOI can then be displayed on a Discovery page within a particular data commons.

EASIER GEN3 DEPLOYMENT

Creating your own Gen3 commons has gotten much easier as we have now adopted Helm Charts for installing and managing Kubernetes applications in the cloud. Helm is a tool that streamlines installing and managing Kubernetes, which itself is a system for automating deployment, scaling, and management of containerized applications. The use of Helm has greatly simplified standing up, configuring, and maintaining your own Gen3 Data Commons, which allows Gen3 to be used by smaller organizations with fewer technological resources.
RUNNING COMPLEX ANALYTICS IN GEN3

In 2023 we made important improvements to our workspaces to allow users to run complex and long-running analysis workflows, which can be commonplace for genomics and imaging datasets. We have worked to create a general purpose workflow execution system that can run containers on the cloud for various applications in a secure and isolated manner. In the first of several related releases, Gen3 operators can now run NextFlow workflows and containers on behalf of users. In early 2024, we will release the ability for researchers to run their own workflows.

IMPROVED METADATA INDEXING

Gen3 creates an elasticsearch index from data stored in its graph data model, which allows for quick retrieval of specific metadata on the explorer page as well as through the API. This year we upgraded our elasticsearch version with the expectation of future additional upgrades in 2024.


NEW KEY PERSONNEL

PHIL SCHUMM

L. Philip Schumm, M.A., is the Director of Biostatistics and Statistical Computing in the Center for Translational Data Science. Prior to this he was the Director of the Research Computing Group and Assistant Director of the Biostatistics Laboratory in the Department of Public Health Sciences, where he had been since 1996.

He is the Co-PI for the HEAL Data Platform, a Gen3 data mesh providing access to data from studies funded by NIH’s Helping to End Addiction Long-term (HEAL) Initiative. He also co-leads the Data and Analytics Support Core (DASC) within NIDA’s Justice Community Opioid Innovation Network’s (JCOIN) Methodology and Advanced Analytics Resource Center (MAARC), for which he has built a Gen3 data commons. He has built data commons and analytic platforms for several other NIH-funded research consortia and groups, including a platform at the University of Chicago for secure management and analysis of Medicare and for the NIA-funded National Social Life, Health and Aging Project (NSHAP). His current methodological work focuses on the measurement and modeling of social networks, cognitive and sensory function, and physical activity, all of which are critical to understanding differences in health trajectories at older ages.
National Institutes of Health
  ARPA-H
  Common Fund
  National Cancer Institute
  National Human Genome Research Institute
  National Heart, Lung, and Blood Institute
  National Institute of Allergy and Infectious Diseases
  National Institute of Biomedical Imaging and Bioengineering
  National Institute of Diabetes and Digestive and Kidney Diseases
  National Institute of Drug Abuse
  Office of the Director
  STRIDES
Veterans Administration
OTHER NEWS AND UPDATES

Super Computing Conference

SC provides the leading technical program for professionals and students in the HPC community. As usual, several members of the CTDS team attended and we shared a booth with the Open Commons Consortium. At SC23 we networked and promoted the use of Gen3 by new groups. CTDS Data Scientist, Anirudh Subramanyam, presented the poster: **GDC-GPT (v0.2): A large language model for querying the Genomic Data Commons** at the Computational Approaches for Cancer Workshop.
KubeCon and CloudNativeCon

The platform engineering team attended the joint KubeCon/CloudNativeCon North America meeting in Chicago this year. It is the Cloud Native Computing Foundation’s flagship conference and gathers adopters and technologists from leading open source and cloud native communities. Lead Platform Engineer, Jawad Qureshi, gave a talk with our partners at Krumware on “Gen3: Advancing Biomedical Research with an Open Source Cloud-Native Platform.”

Internship program

CTDS has an annual internship program for undergraduate students. This is a way for us to expose students to the innovative technologies we were working with and are developing in order to improve biomedical data sharing. This past summer we had seven students working on projects ranging from data security to AI to analytics workflows.
Gen3 Community Forum

In 2022, we launched the Gen3 Community Forum. In 2023, we have built on this success and have had six more community meetings over the year, which covered topics ranging from Helm charts, security practices, Gen3 data models, and the soon-to-be-released Gen3 Frontend Framework. In 2023 we had an average of 54 attendees at each meeting representing 26 organizations from around

Biomedical Research Hub Named as GA4GH Driver Project

The Global Alliance for Genomics and Health (GA4GH) is working to create frameworks and standards to enable the responsible, voluntary, and secure sharing of genomic and health-related data. CTDS is an active member of multiple workstreams and a member of the steering committee. This year, the Biomedical Research Hub (created by CTDS using Gen3 Mesh/Fabric services) was named as a new driver project to help set standards for data sharing by the international community.
CTDS is one of the world leaders developing systems and platforms for biomedical data science.

In particular, our open-source Gen3 data platform has been used to build over 30 data commons around the world, including those that CTDS operates, that we operate with our partners, and that third parties operate.

One of our goals in 2024 is to make additional progress on simplifying deployment and maintenance of Gen3. We made major progress on this goal in 2023 with the roll out of helm charts, but in 2024 we will move this service to production for CTDS thereby ensuring that all Gen3 services can be easily and seamlessly deployed in this manner.

In 2024, another goal is to develop tools and techniques to build large language models and other deep learning models over the public data in our systems and to use these tools and techniques to make research discoveries.

A third major goal in 2024 is to develop and release a new frontend for the GDC and Gen3 so that third party data analysis tools can be easily integrated into these data platforms. The GDC frontend is due to be released in early 2024 with the Gen3 frontend coming in phases over the year.

If you are interested in working with us on these or any of other projects, please don’t hesitate to reach out!
CTDS is composed of faculty, scientists, engineers, project managers, and administrators who are all helping to solve the difficult challenges of working with and learning from large biomedical data sets.

Visit our website: ctds.uchicago.edu
Explore Gen3 Data Platform: gen3.org
Follow us on Twitter and LinkedIn:
  twitter.com/UChicagoCTDS
  linkedin.com/company/center-for-translational-data-science/
Explore Gen3 source code on GitHub:
  https://github.com/orgs/uc-cdis/

Questions about getting involved or supporting CTDS: ctds@uchicago.edu