National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

NIDDK-CR Resources for Research

Data Science and Meet the Expert Webinar Series

REPOSITORY

TECHNOLOGY ADVANCEMENT

BIOSTATISTICS

CLINICAL STUDIES

REGULATORY

Office of Clinical Research Support

NIH National Institute of Diabetes and Digestive and Kidney Diseases

January 21, 2026

NIH National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

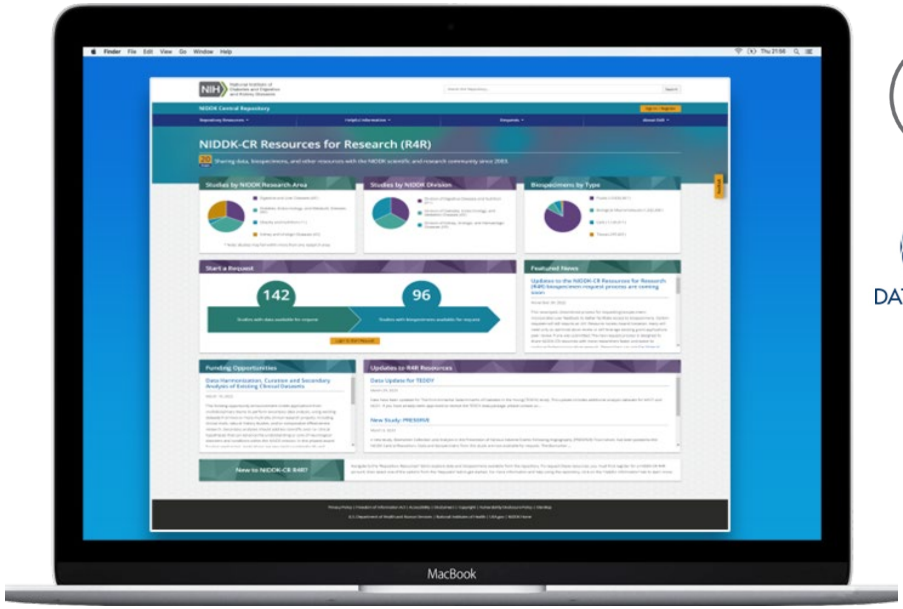
NIDDK Central Repository Overview

Mission

Established in 2003 to **facilitate sharing of data, biospecimens, and other resources** generated from studies supported by NIDDK and within NIDDK's mission by making these **resources available for request to the broader scientific and research community**.

- Supports receipt and distribution of data and biospecimens in a manner that is ethical, equitable, and efficient
- Enables investigators not involved with the original work to test new hypotheses without the need to collect new data or biospecimens
- Promotes FAIR (Findable, Accessible, Interoperable, and Reusable) and TRUST (Transparency, Responsibility, User focus, Sustainability, and Technology) principles

Recorded past tutorials, webinars, and other educational resources can be found on the NIDDK-CR website


Finder File Edit View Go Window Help
NIDDK Central Repository
NIDDK-CR Resources for Research (R4R)
Sharing data, biospecimens, and other resources with the NIDDK scientific and research community since 2003.
Studies by NIDDK Research Area
Studies by NIDDK Division
Biospecimens by Type
Start a Request
142
96
Featured News
Funding Opportunities
Updates to R4R Resources
Data Update for TEDDY
New Study: PRESERVE
New to NIDDK-CR R4R?
MacBook

CORE TRUST SEAL

WORLD DATA SYSTEM

Imaging Data Files 15.8 M	Clinical Datasets >8,600 from 194 clinical studies	Biospecimens >16 M
Registered Users 7,026	Weekly Users >5,000	Public Releases >875

NIH National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Analytics Workbench Functionality

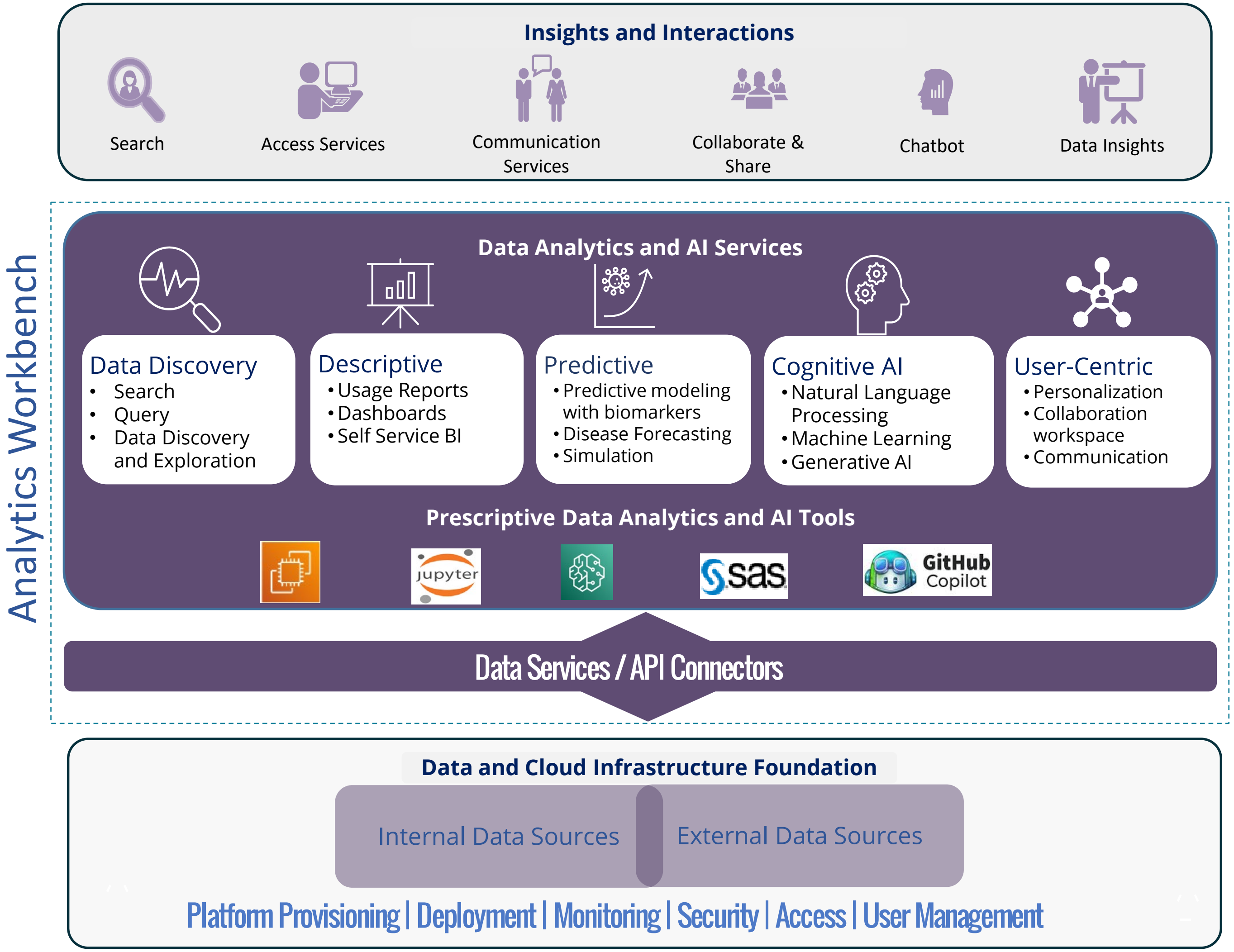
***Streamlining** end-to-end data science lifecycle and discovery of data-driven biomedical insights.*

Innovation and ease of use

A cloud-based analytics environment where researchers and data scientists can access a suite of integrated analytics tools and cloud computing resources to participate in data challenges and AI innovation.

Expected Benefits of Analytics Workbench:

- Promote Collaboration
- Support AI Innovation
- Minimize Data Movement
- Improve User Experience
- Discover Data Insights
- Advance NIDDK Research Mission



Insights and Interactions

Search | Access Services | Communication Services | Collaborate & Share | Chatbot | Data Insights

Analytics Workbench

Data Analytics and AI Services

Data Discovery
- Search
- Query
- Data Discovery and Exploration

Descriptive
- Usage Reports
- Dashboards
- Self Service BI

Predictive
- Predictive modeling with biomarkers
- Disease Forecasting
- Simulation

Cognitive AI
- Natural Language Processing
- Machine Learning
- Generative AI

User-Centric
- Personalization
- Collaboration workspace
- Communication

Prescriptive Data Analytics and AI Tools

jupyter | sas | GitHub Copilot

Data Services / API Connectors

Data and Cloud Infrastructure Foundation

Internal Data Sources | External Data Sources

Platform Provisioning | Deployment | Monitoring | Security | Access | User Management

NIDDK-CR Data Science Centric Challenge Series

Goals of NIDDK-CR Data-science centric challenge series

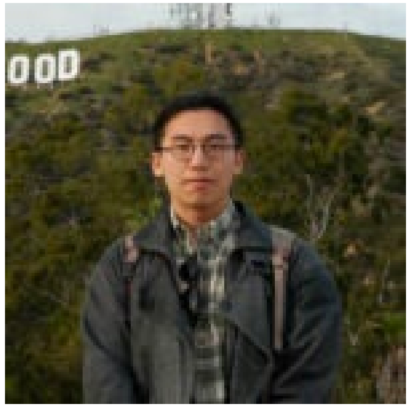
- Develop tools, approaches, models and/or methods to increase data interoperability and usability for artificial intelligence (AI) and machine learning (ML) applications
- Augment and enhance existing data for future secondary research, including data-driven discovery by AI/ML researchers
- Discover innovative approaches to enhance the utility of datasets for AI/ML applications

Visit our website for more information on our data-centric movement and to learn more about our past data-challenges

Meet the Experts

Chen Li is a professor in the Department of Computer Science at UC Irvine. He received his Ph.D. degree in Computer Science from Stanford University, and his M.S. and B.S. in Computer Science from Tsinghua University, China, respectively. He was a recipient of an NSF CAREER award and several test-of-time publication awards, a part-time visiting research scientist at Google, an ACM distinguished member, and an IEEE fellow. He was a co-founder and CTO of a startup to commercialize his research.

Kun Woo Park and Jiadong Bai are PhD students in Computer Science at UC Irvine working on the Texera project.



Advancing Collaborative Data Science with Texera

Prof. Chen Li

with Kun Woo Park and Jiadong Bai

Department of Computer Science, UC Irvine

January 21, 2026

NSF

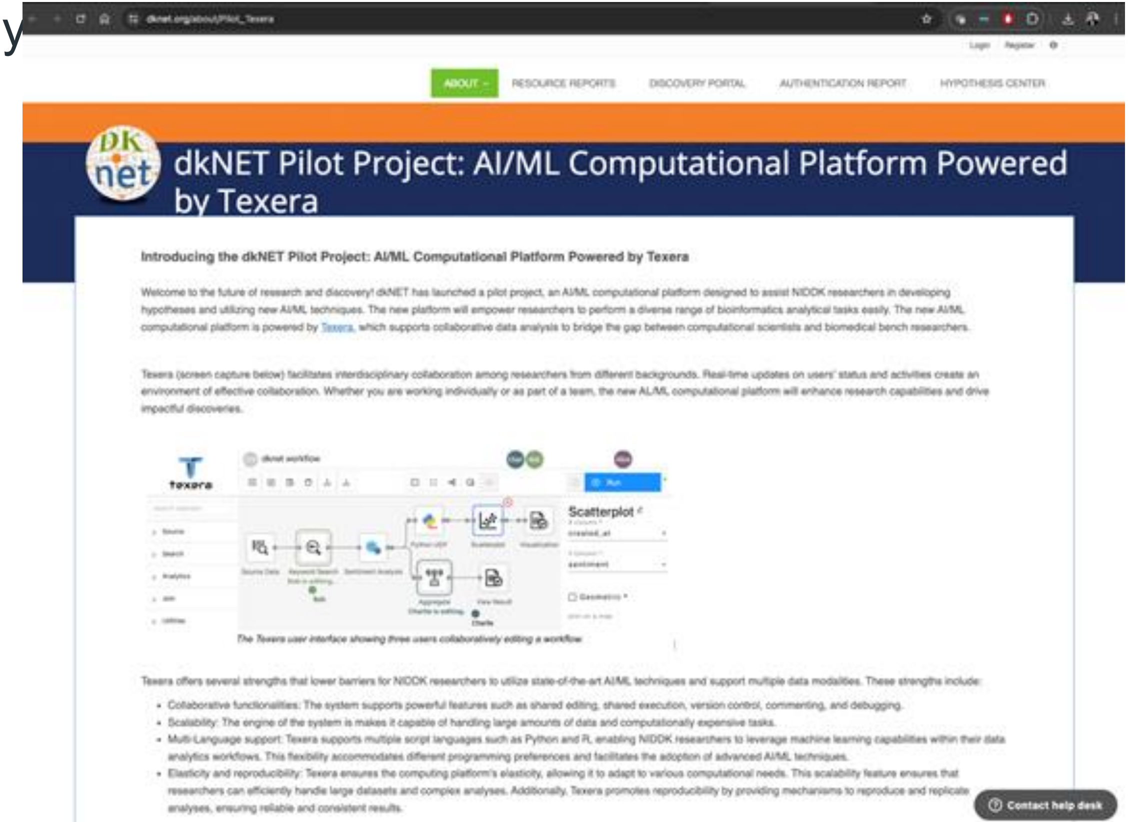
NIH National Institute of Diabetes and Digestive and Kidney Diseases

Apache Texera: Overview

- Supporting data science and AI/ML as workflows
- Cloud services (no installation, software patches)
- Supporting community-based sharing of data and workflows
- Shared editing/execution
- Supporting Python, R, Java as user-defined functions (UDFs)
- Started in 2016
- Open source (being incubated by Apache)
- Parallel engine, scalable

Part of NIDDK dkNET Computational Core

Making Data Science and AI/ML easily available to the NIDDK community



dkNET Pilot Project: AI/ML Computational Platform Powered by Texera

Introducing the dkNET Pilot Project: AI/ML Computational Platform Powered by Texera

Welcome to the future of research and discovery! dkNET has launched a pilot project, an AI/ML computational platform designed to assist NIDDK researchers in developing hypotheses and utilizing new AI/ML techniques. The new platform will empower researchers to perform a diverse range of bioinformatics analytical tasks easily. The new AI/ML computational platform is powered by Texera, which supports collaborative data analysis to bridge the gap between computational scientists and biomedical bench researchers.

Texera (screen capture below) facilitates interdisciplinary collaboration among researchers from different backgrounds. Real-time updates on users' status and activities create an environment of effective collaboration. Whether you are working individually or as part of a team, the new AL/ML computational platform will enhance research capabilities and drive impactful discoveries.

The Texera user interface showing three users collaboratively editing a workflow

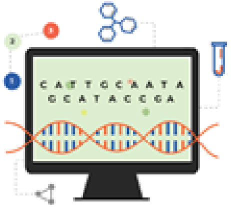
Texera offers several strengths that lower barriers for NIDDK researchers to utilize state-of-the-art AI/ML techniques and support multiple data modalities. These strengths include:

- Collaborative functionalities: The system supports powerful features such as shared editing, shared execution, version control, commenting, and debugging.
- Scalability: The engine of the system is makes it capable of handling large amounts of data and computationally expensive tasks.
- Multi-Language support: Texera supports multiple script languages such as Python and R, enabling NIDDK researchers to leverage machine learning capabilities within their data analytics workflows. This flexibility accommodates different programming preferences and facilitates the adoption of advanced AI/ML techniques.
- Elasticity and reproducibility: Texera ensures the computing platform's elasticity, allowing it to adapt to various computational needs. This scalability feature ensures that researchers can efficiently handle large datasets and complex analyses. Additionally, Texera promotes reproducibility by providing mechanisms to reproduce and replicate analyses, ensuring reliable and consistent results.

Example application: sequence analysis in biology

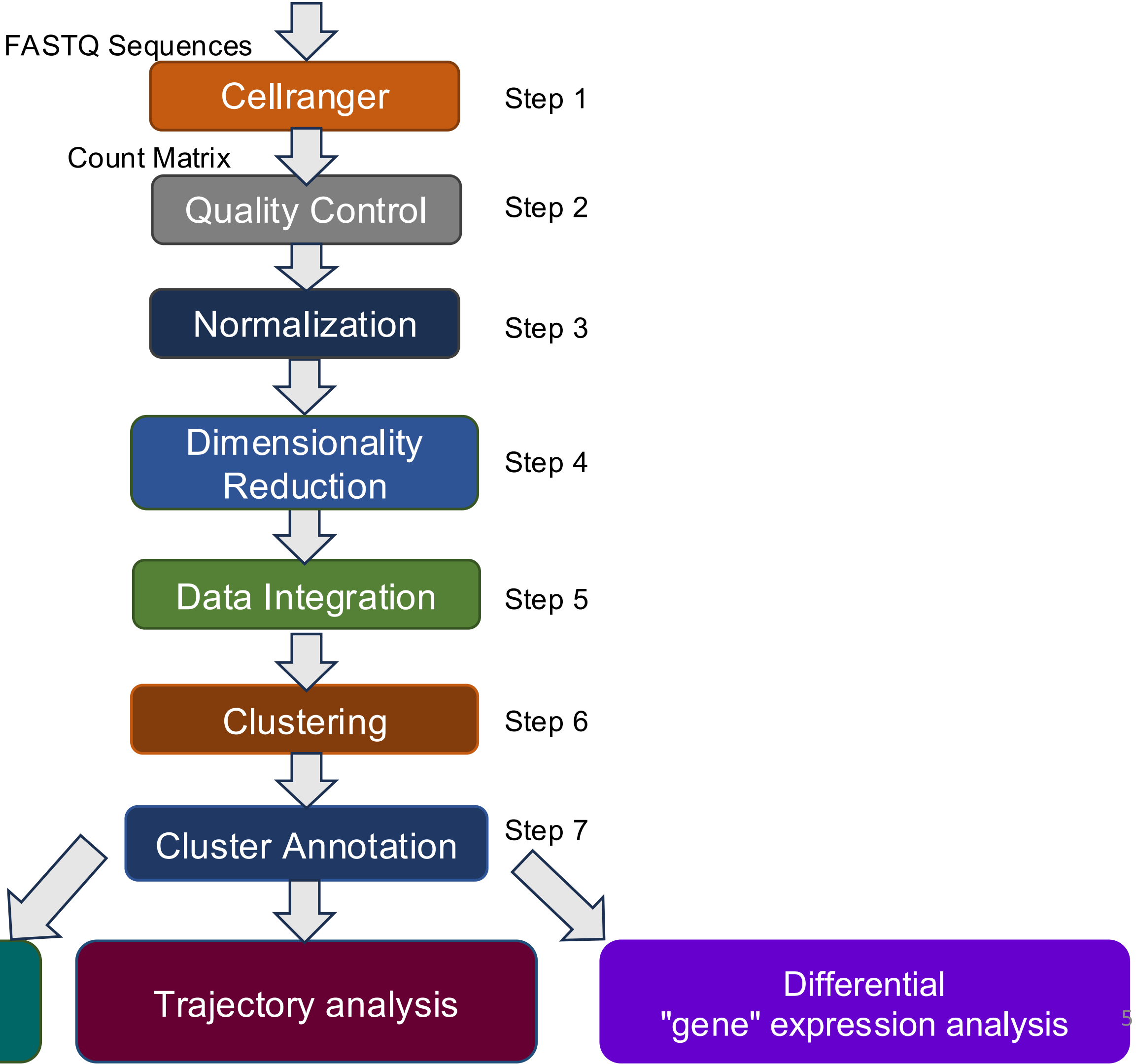
Alice: Biologist (PI)

Sally: Bioinformatician



Bob: Bioinformatician

Sequence analysis pipeline



FASTQ Sequences

Cellranger — Step 1

Count Matrix

Quality Control — Step 2

Normalization — Step 3

Dimensionality Reduction — Step 4

Data Integration — Step 5

Clustering — Step 6

Cluster Annotation — Step 7

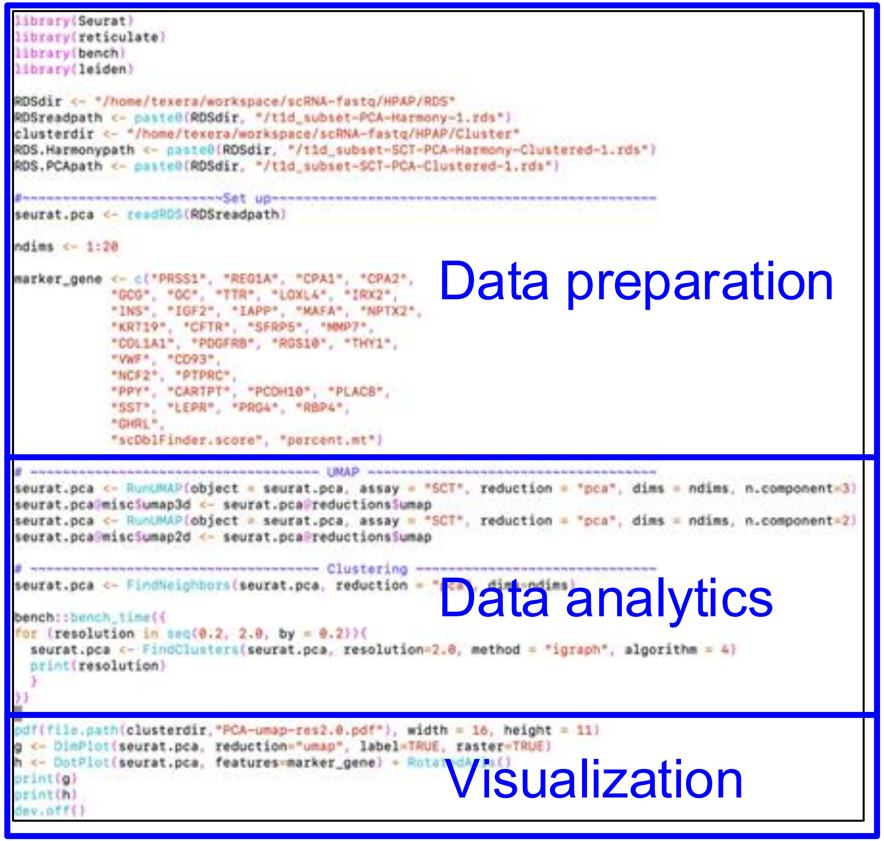
AI/ML analysis

Trajectory analysis

Differential "gene" expression analysis

Coding challenges

- Coding is hard!
- Version control of libraries
- Needs servers
- Slow on large data
- Not every lab can afford a bioinformatician



```
library(Seurat)
library(reticulate)
library(bench)
library(leiden)

RDSdir <- "/home/texera/workspace/scRNA-fastq/HPAP/RDS"
RDSreadpath <- paste0(RDSdir, "/t1d_subset-PCA-Harmony-1.rds")
clusterdir <- "/home/texera/workspace/scRNA-fastq/HPAP/Cluster"
RDS.Harmonypath <- paste0(RDSdir, "/t1d_subset-SCT-PCA-Harmony-Clustered-1.rds")
RDS.PCApath <- paste0(RDSdir, "/t1d_subset-SCT-PCA-Clustered-1.rds")

#-------------------------Set up--------------------------------------------------
seurat.pca <- readRDS(RDSreadpath)

ndims <- 1:20

marker_gene <- c("PRSS1", "REG1A", "CPA1", "CPA2",
             "GCG", "GC", "TTR", "LOXL4", "IRX2",
             "INS", "IGF2", "IAPP", "MAFA", "NPTX2",
             "KRT19", "CFTR", "SFRP5", "MMP7",
             "COL1A1", "PDGFRB", "RGS10", "THY1",
             "VWF", "CD93",
             "NCF2", "PTPRC",
             "PPY", "CARTPT", "PCDH10", "PLAC8",
             "SST", "LEPR", "PRG4", "RBP4",
             "GHRL",
             "scDblFinder.score", "percent.mt")
# ------------------------------------ UMAP ------------------------------------
seurat.pca <- RunUMAP(object = seurat.pca, assay = "SCT", reduction = "pca", dims = ndims, n.component=3)
seurat.pca@misc$umap3d <- seurat.pca@reductions$umap
seurat.pca <- RunUMAP(object = seurat.pca, assay = "SCT", reduction = "pca", dims = ndims, n.component=2)
seurat.pca@misc$umap2d <- seurat.pca@reductions$umap

# ------------------------------------ Clustering ------------------------------
seurat.pca <- FindNeighbors(seurat.pca, reduction = "pca", dims=ndims)

bench::bench_time({
for (resolution in seq(0.2, 2.0, by = 0.2)){
  seurat.pca <- FindClusters(seurat.pca, resolution=2.0, method = "igraph", algorithm = 4)
  print(resolution)
  }
})

pdf(file.path(clusterdir,"PCA-umap-res2.0.pdf"), width = 16, height = 11)
g <- DimPlot(seurat.pca, reduction="umap", label=TRUE, raster=TRUE)
h <- DotPlot(seurat.pca, features=marker_gene) + RotatedAxis()
print(g)
print(h)
dev.off()
```

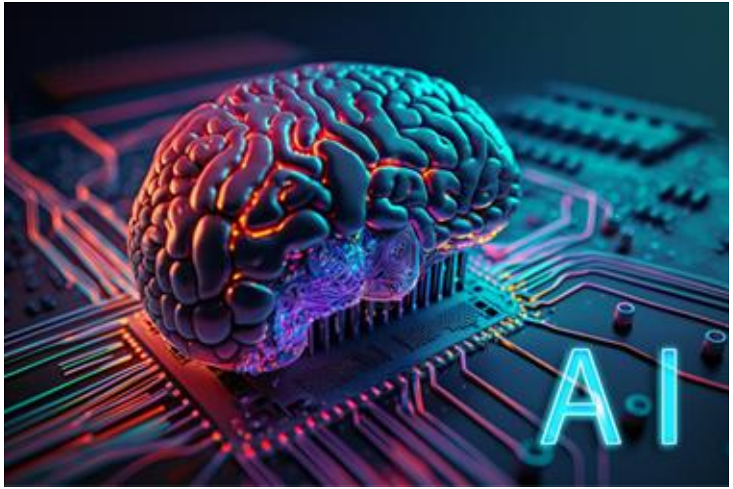
Sally: Bioinformatician

Collaboration challenges

- Collaborators of different backgrounds
  - Biologists
  - Bioinformaticians
  - Computer scientists
- Collaborators from different organizations
  - Same lab: senior students vs new students
  - Other labs

AI/ML opportunities

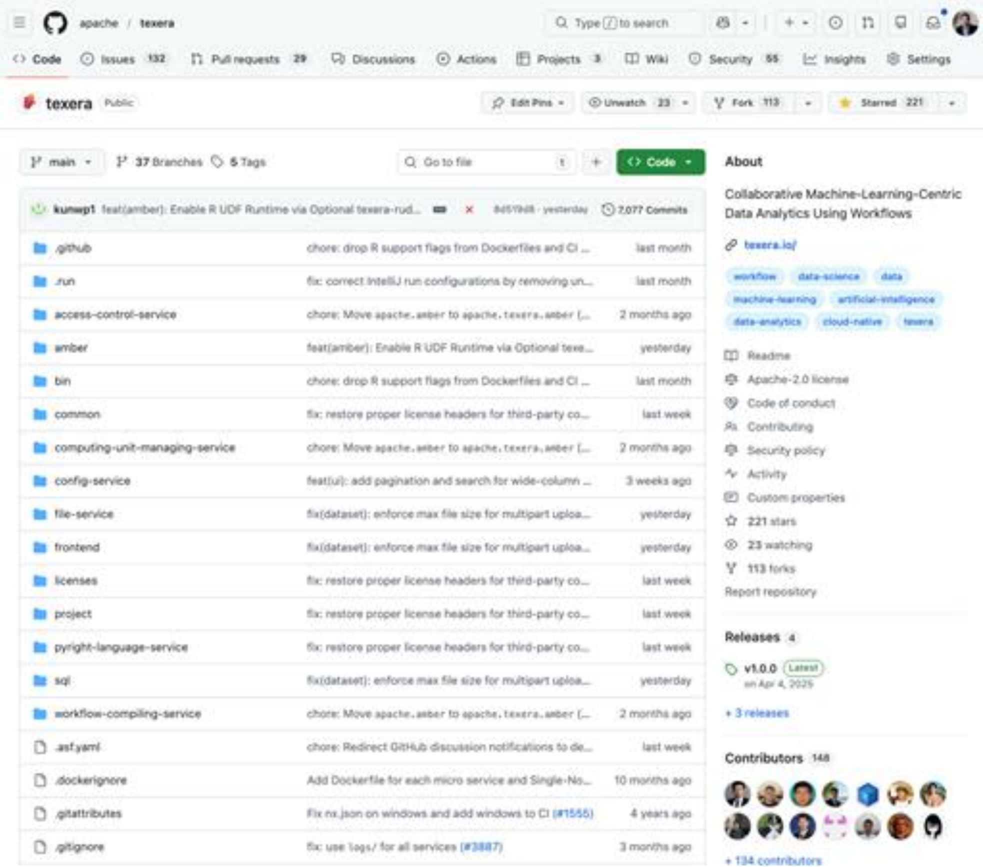
- How to utilize state-of-the-art AI/ML technologies?
- Require advanced coding skills
- Not easily available



Our solution

Collaborative data science and AI/ML using GUI-based workflows

Open source (Apache Incubating)



kunwp1 feat(amber): Enable R UDF Runtime via Optional texera-rud... 8d519d8 · yesterday 2,077 Commits

.github	chore: drop R support flags from Dockerfiles and CI ...	last month
.run	fix: correct IntelliJ run configurations by removing un...	last month
access-control-service	chore: Move `apache.amber` to `apache.texera.amber` (...	2 months ago
amber	feat(amber): Enable R UDF Runtime via Optional texe...	yesterday
bin	chore: drop R support flags from Dockerfiles and CI ...	last month
common	fix: restore proper license headers for third-party co...	last week
computing-unit-managing-service	chore: Move `apache.amber` to `apache.texera.amber` (...	2 months ago
config-service	feat(ui): add pagination and search for wide-column ...	3 weeks ago
file-service	fix(dataset): enforce max file size for multipart uploa...	yesterday
frontend	fix(dataset): enforce max file size for multipart uploa...	yesterday
licenses	fix: restore proper license headers for third-party co...	last week
project	fix: restore proper license headers for third-party co...	last week
pyright-language-service	fix: restore proper license headers for third-party co...	last week
sql	fix(dataset): enforce max file size for multipart uploa...	yesterday
workflow-compiling-service	chore: Move `apache.amber` to `apache.texera.amber` (...	2 months ago
.asf.yaml	chore: Redirect GitHub discussion notifications to de...	last week
.dockerignore	Add Dockerfile for each micro service and Single-No...	10 months ago
.gitattributes	Fix nx.json on windows and add windows to CI (#1555)	4 years ago
.gitignore	fix: use `logs/` for all services (#3887)	3 months ago

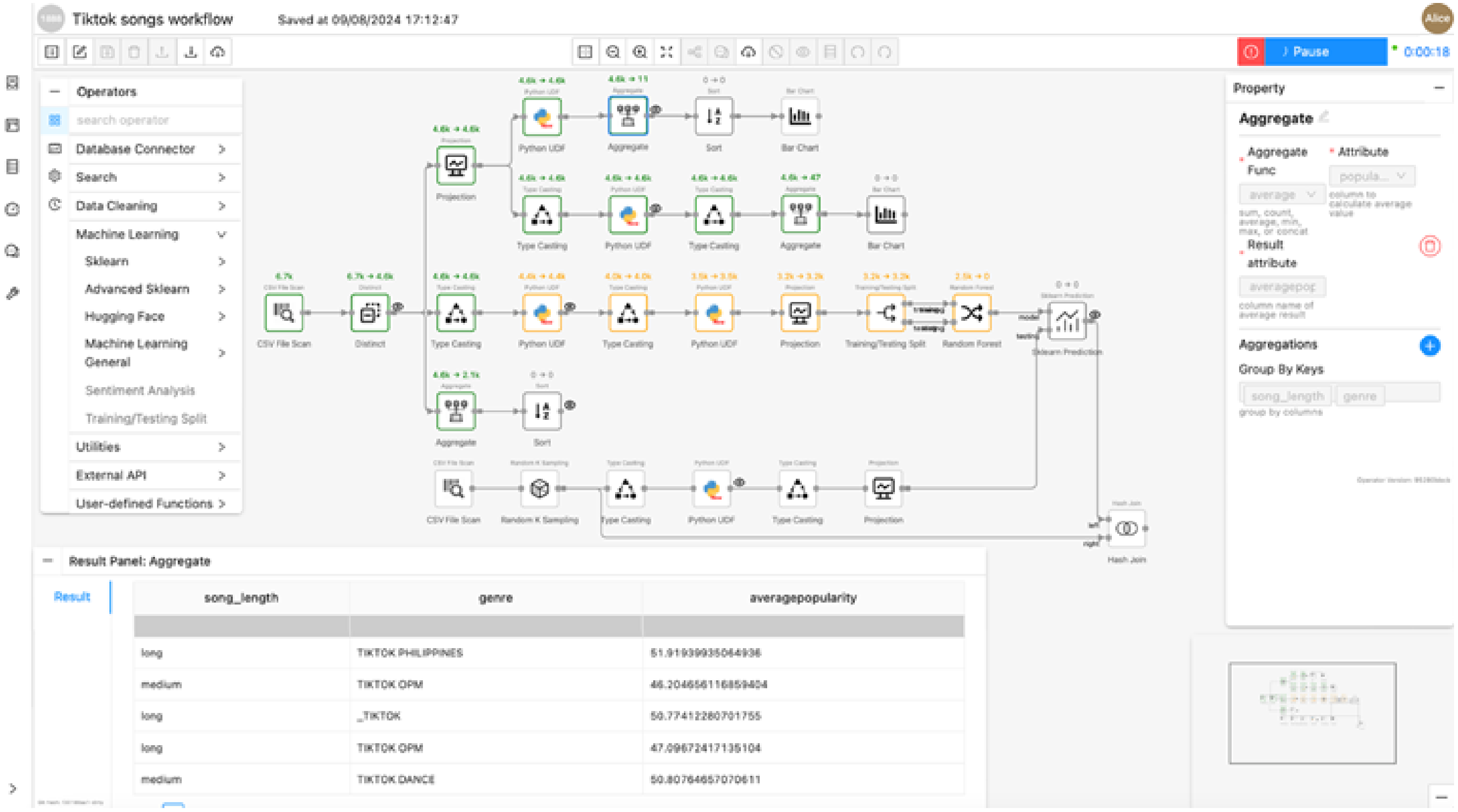
About

Collaborative Machine-Learning-Centric Data Analytics Using Workflows

texera.io/

Chen Li, UCI

Demo!



Tiktok songs workflow

Saved at 09/08/2024 17:12:47

Property

Aggregate

Aggregate Func: average
sum, count, average, min, max, or concat

Attribute: popula...
column to calculate average value

Result attribute: averagepop
column name of average result

Aggregations

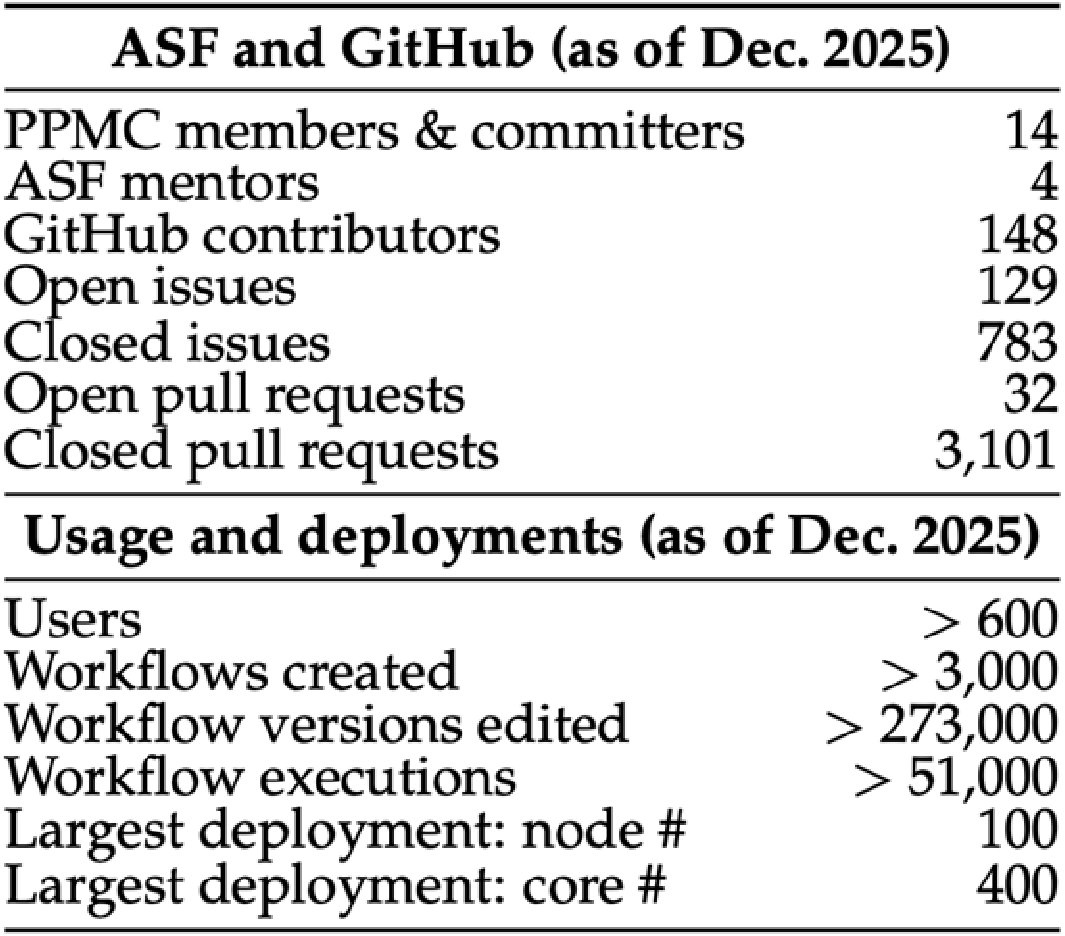
Group By Keys: song_length, genre
group by columns

Result Panel: Aggregate

Result

song_length	genre	averagepopularity
long	TIKTOK PHILIPPINES	51.91939935064936
medium	TIKTOK OPM	46.204656116859404
long	_TIKTOK	50.77412280701755
long	TIKTOK OPM	47.09672417135104
medium	TIKTOK DANCE	50.80764657070611

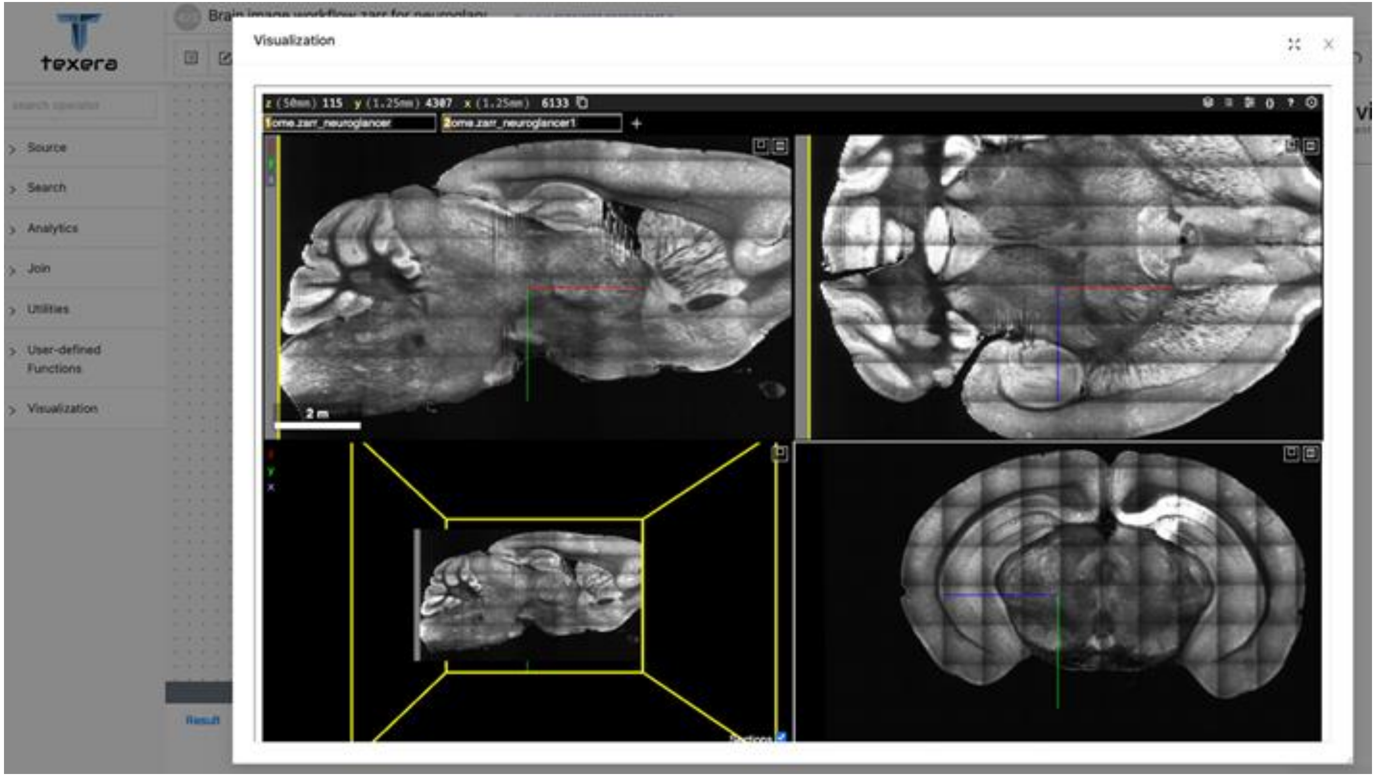
Statistics



ASF and GitHub (as of Dec. 2025)	
PPMC members & committers	14
ASF mentors	4
GitHub contributors	148
Open issues	129
Closed issues	783
Open pull requests	32
Closed pull requests	3,101

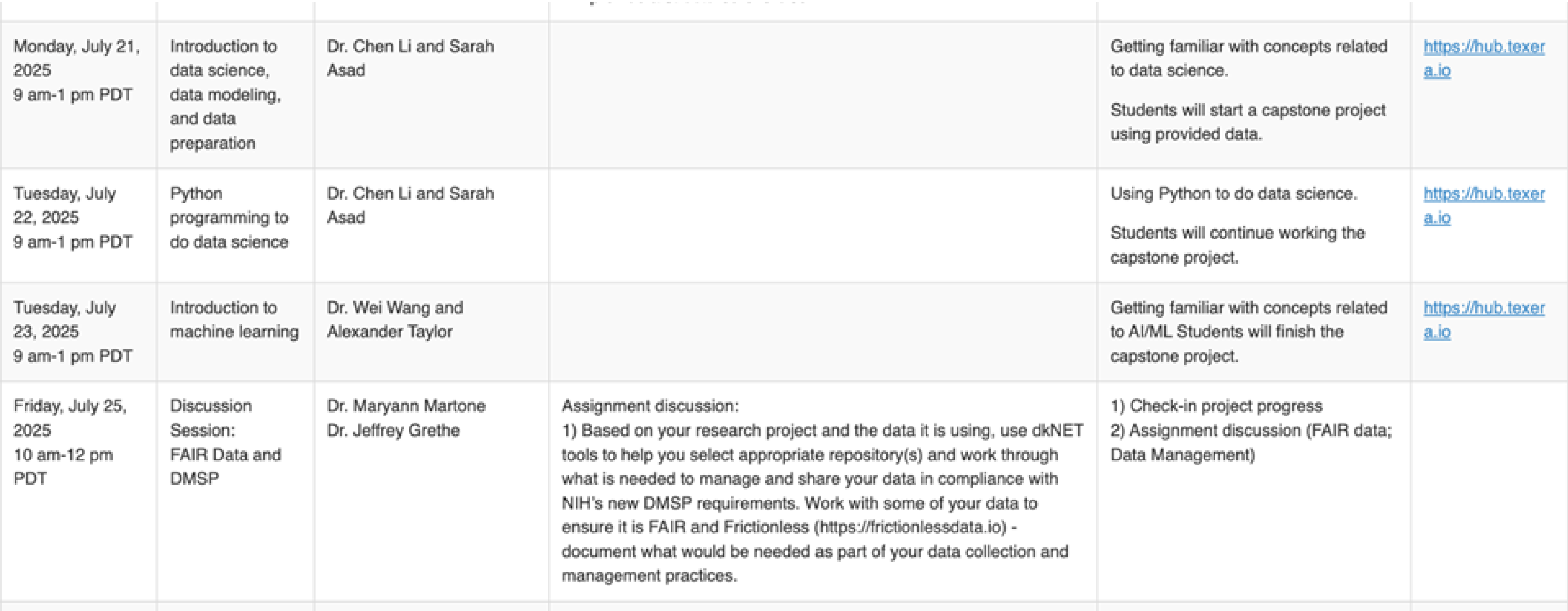
Usage and deployments (as of Dec. 2025)	
Users	> 600
Workflows created	> 3,000
Workflow versions edited	> 273,000
Workflow executions	> 51,000
Largest deployment: node #	100
Largest deployment: core #	400

Example: analyzing brain images, 256GB



Teaching non-STEM students AI/ML using Texera

2025 dkNET Summer Bootcamp



Monday, July 21, 2025 9 am-1 pm PDT	Introduction to data science, data modeling, and data preparation	Dr. Chen Li and Sarah Asad		Getting familiar with concepts related to data science. Students will start a capstone project using provided data.	https://hub.texera.io
Tuesday, July 22, 2025 9 am-1 pm PDT	Python programming to do data science	Dr. Chen Li and Sarah Asad		Using Python to do data science. Students will continue working the capstone project.	https://hub.texera.io
Tuesday, July 23, 2025 9 am-1 pm PDT	Introduction to machine learning	Dr. Wei Wang and Alexander Taylor		Getting familiar with concepts related to AI/ML Students will finish the capstone project.	https://hub.texera.io
Friday, July 25, 2025 10 am-12 pm PDT	Discussion Session: FAIR Data and DMSP	Dr. Maryann Martone Dr. Jeffrey Grethe	Assignment discussion: 1) Based on your research project and the data it is using, use dkNET tools to help you select appropriate repository(s) and work through what is needed to manage and share your data in compliance with NIH's new DMSP requirements. Work with some of your data to ensure it is FAIR and Frictionless (https://frictionlessdata.io) - document what would be needed as part of your data collection and management practices.	1) Check-in project progress 2) Assignment discussion (FAIR data; Data Management)	

https://dknet.org/about/summer_course_2025

Ongoing efforts

- Support management of ML models
- Incorporate more AI techniques to the platform
- Make analysis pipelines to the community
- Improve security and privacy
- High performance and scalability
- Elastic computing using cloud resources
- …

Summary: Apache Texera

- Cloud-computing platform
- GUI-based workflows (no coding needed)
- Collaboration and sharing of data/analyses
- Parallel computing: for big data
- Supporting multiple languages: Python, R, Java, …
- Supporting AI/ML (training, inference, …)

Advancing Collaborative Data Science with Texera

Prof. Chen Li

Department of Computer Science

UC Irvine

NSF

NIH National Institute of Diabetes and Digestive and Kidney Diseases