[tile]DB DESIGNED FOR DISCOVERY + Chan Zuckerberg Initiative

Training models on atlas-scale single-cell datasets

scverse Conference | September 12, 2024

Maximilian Lombardo
Sr. Product Applications Scientist, CZI

Ryan Williams
SOMA Software Engineer, TileDB

Spencer Seale
Solutions Architect, TileDB

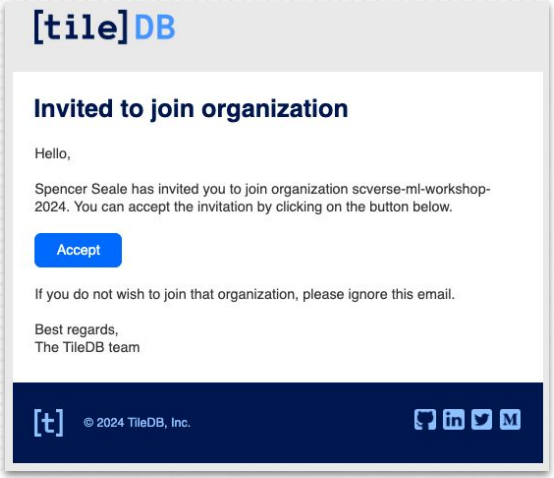
The workshop in a nutshell

WE WILL COVER

- What is CellxGene Census?
- What are TileDB and TileDB-SOMA?
- Understanding the Census and its data.
- Using the Census to:
  - Gain a high-level understanding of its contents.
  - Access and slicing data efficiently.
  - Performing computations efficiently.
  - Export data to single-cell toolkit.
- Utilizing Census PyTorch loaders for scalable modelling.
  - Training a classifier for Cell Type Annotation (immune cells)
- Cell Type Annotation via similarity search (vector search) for immune cells

[t] + CZ
DESIGNED FOR DISCOVERY

This will be a hands-on workshop



[tile]DB

Invited to join organization

Hello,

Spencer Seale has invited you to join organization scverse-ml-workshop-2024. You can accept the invitation by clicking on the button below.

Accept

If you do not wish to join that organization, please ignore this email.

Best regards,
The TileDB team

[t] © 2024 TileDB, Inc.

If you *didn't* receive an invitation to join TileDB Cloud via email, **email spencer.seale@tiledb.com now.**

If you *did* receive an invitation**, accept it, and follow the instructions to create a TileDB Cloud account.**

Alternatively, you can passively along, or view the notebook at TileDB-Inc/scverse-ml-workshop-2024 on GitHub.

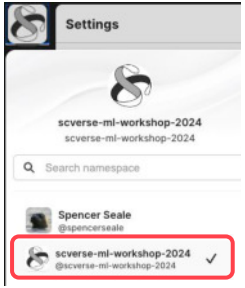
Questions? Message us in Zulip!

[t] + CZ

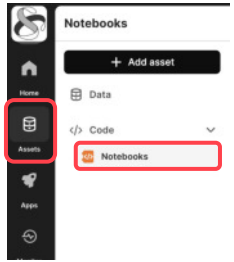
DESIGNED FOR DISCOVERY

1 Login at cloud.tiledb.com

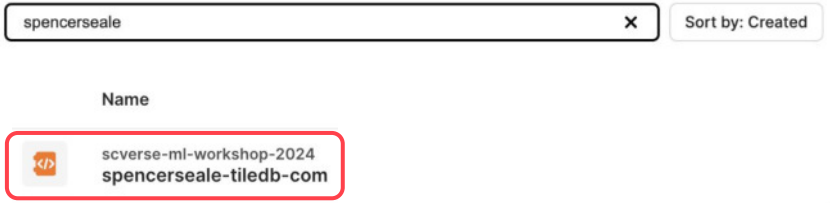
2 Switch to workshop Namespace (top-left)



3 Go to "Assets" & "Notebooks" (left)

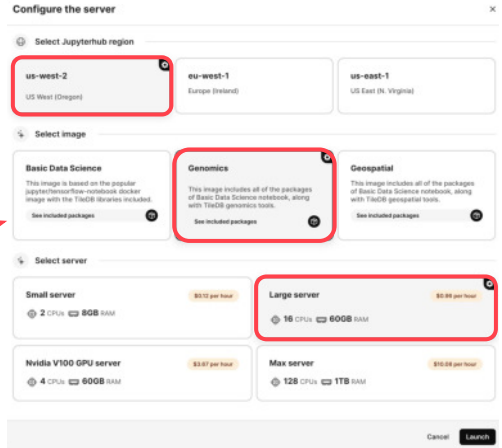


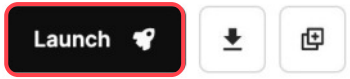
4 Search by your email username and select your notebook (top-right)



5 Launch your notebook with the configurations:

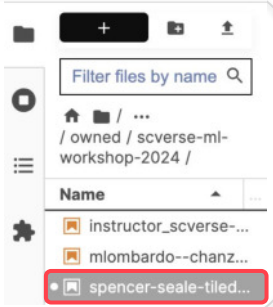
- **us-west-2**
- **Genomics**
- **Large server**





6 Display server

7 Select notebook



[t] + CZ
DESIGNED FOR DISCOVERY

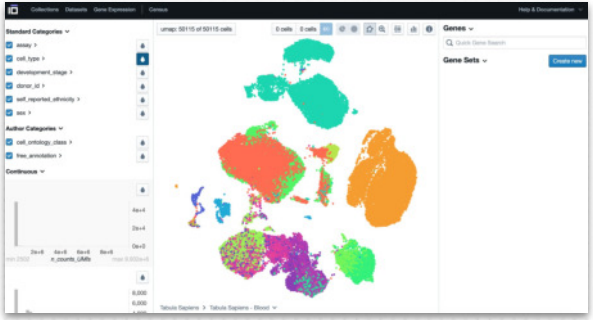
[tile]DB DESIGNED FOR DISCOVERY + Chan Zuckerberg Initiative

CELLxGENE Discover Census

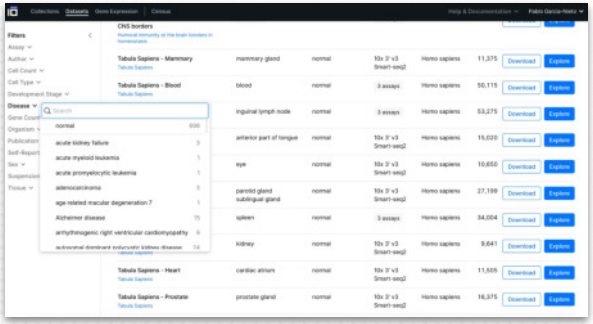
Visual tools democratize access to single-cell data

https://cellxgene.cziscience.com/

Explore individual datasets [tile]DB

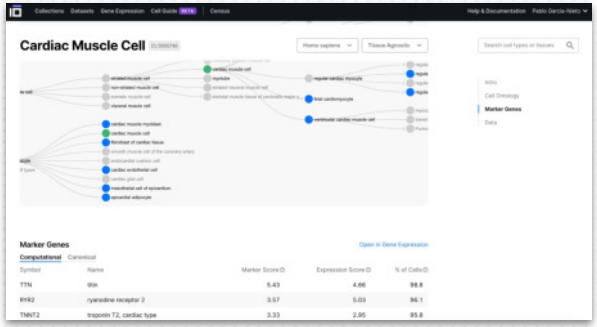


Browse and **download**



Inspect **Gene Expression** per tissue

Learn about cell types with **Cell Guide**



CZ CELL × GENE DISCOVER

https://cellxgene.cziscience.com/

UNIQUE CELLS

90M

Human **55M**

Mouse **30M**

DATASETS

1486

DISEASES

132

CELL TYPES

907

ASSAYS

- 10x RNA-seq
- sci-RNA-seq
- microwell-seq
- Drop-seq
- Seq-Well
- Smart-seq2 and v4
- BD Rhapsody
- GEXSCOPE
- MARS-seq
- Visium 10x

TOP TISSUES

Heart 2.2M

Blood 10.2M

Brain 34.2M

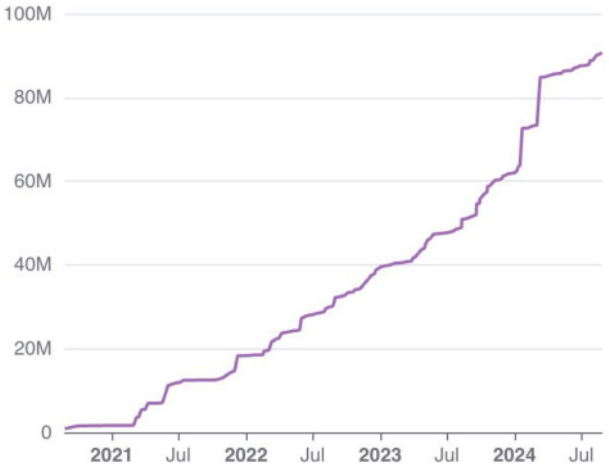
Lung 5.7M

Kidney 4.6M

Intestine 2.0M

Breast 2.6M

CONSTANT RATE OF DATA INGESTION



All data at CZI CELLxGENE Discover is curated for data reuse and integration

https://cellxgene.cziscience.com/

GENE EXPRESSION

Normalized data

Raw counts

FEATURE METADATA

Name Type ID

empowered

SCHEMA

UNIVERSALLY AVAILABLE CELL METADATA

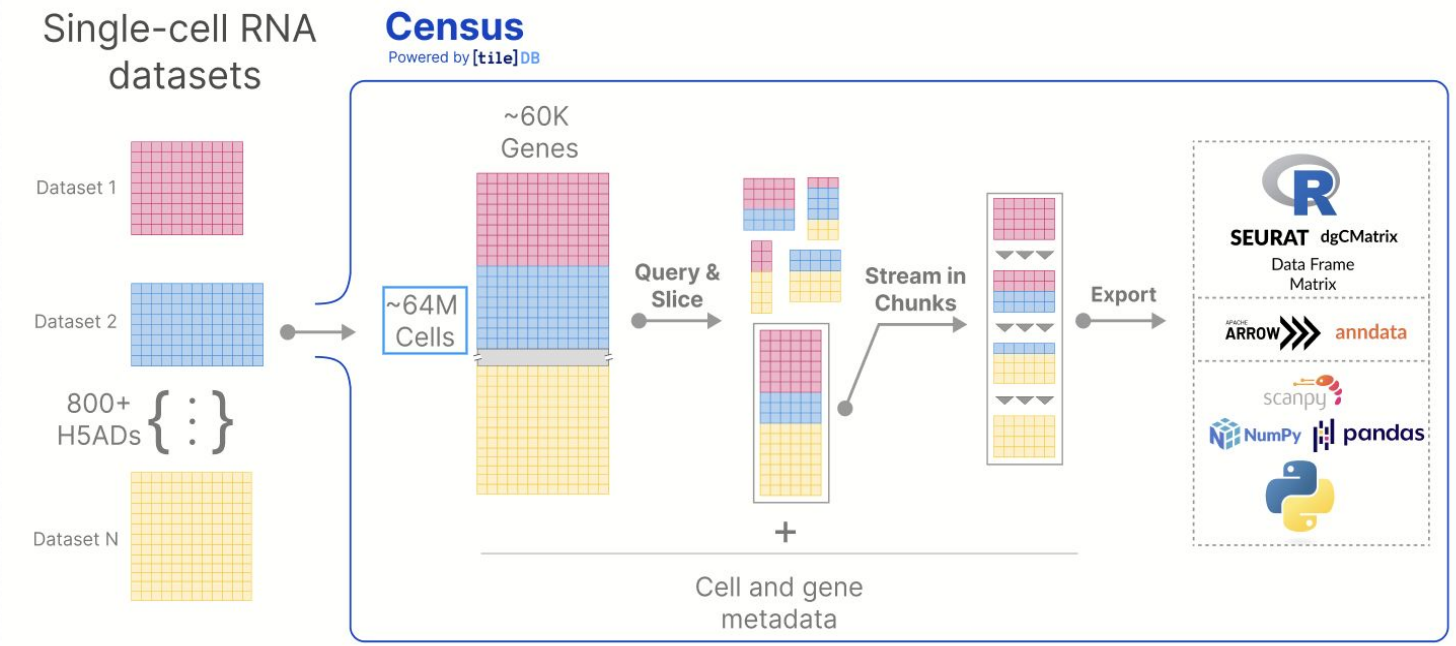
Species Tissue Age Sex Ethnicity

Assay Cell type Suspension Donor Disease

https://github.com/chanzuckerberg/single-cell-curation/blob/main/schema/5.1.0/schema.md

What is Census?

Built from >800 datasets Census is a **data object + API**. It gives efficient access to the largest aggregation of standardized single-cell data ready for analysis and modelling at scale.
https://chanzuckerberg.github.io/cellxgene-census/



Census stack

Census is built atop **TileDB-SOMA**, which is built atop **SOMA** and **TileDB**

https://github.com/chanzuckerberg/cellxgene-census

https://github.com/single-cell-data/TileDB-SOMA

https://github.com/single-cell-data/SOMA

https://github.com/TileDB-Inc/TileDB

For this workshop we will use **TileDB-Cloud** to skip the installation process

https://cloud.tiledb.com/

Coming to Census soon

Newly released **Census Similarity Search**
Now available to use!

- Similarity Search Documentation

[tile]DB DESIGNED FOR DISCOVERY + Chan Zuckerberg Initiative

SOMA (data model) - **TileDB** (database)

What is SOMA?

*(**S**tack **O**f **M**atrices, **A**nnotated)*

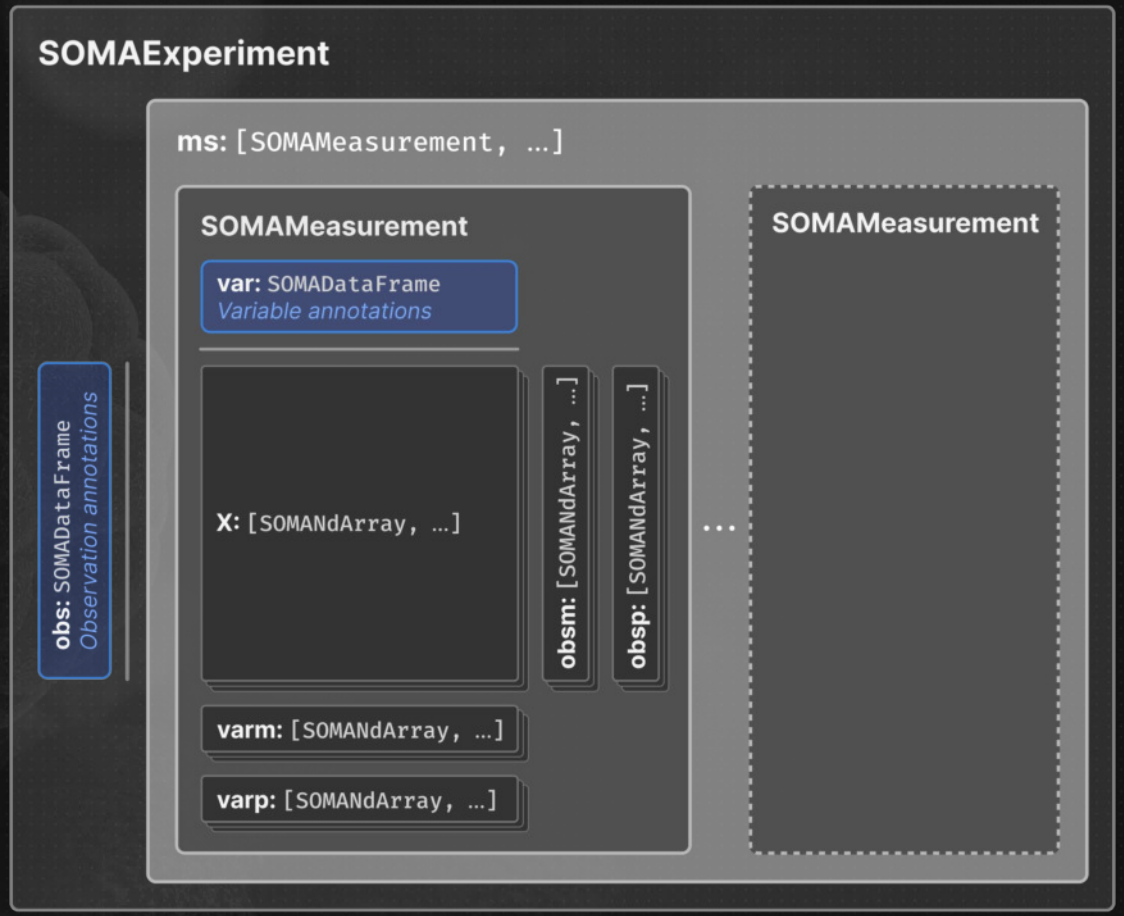
Data Model

A language-agnostic data model for representing collections of annotated matrices and derived results on disk.

API Specification

A language-agnostic API specification for interacting with the data model.

SOMA Data Model



Goals for SOMA

Open source

- specification & reference implementations

Language-independent

- support both R & Python
- option to add more languages in the future

Interoperable with popular single-cell analysis frameworks

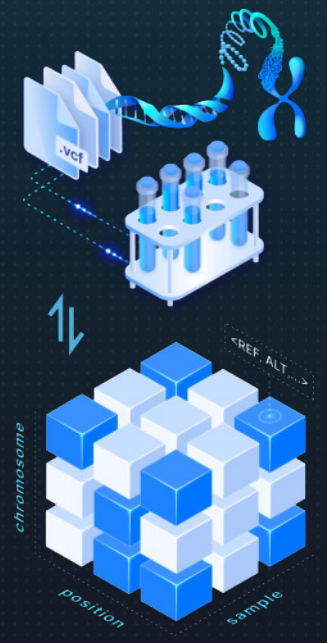
- AnnData/ScanPy (Python)
- Seurat (R)
- Bioconductor (R)

Designed for atlas-scale datasets

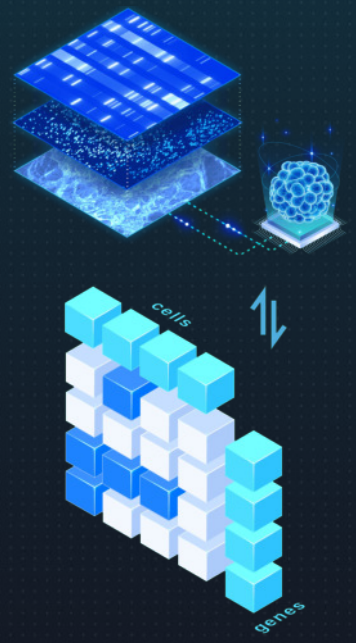
- support for cloud object stores
- out-of-core reads
- efficient querying

TileDB is a multi-modal database that morphs for any application

Population Genomics



Single-cell Biology

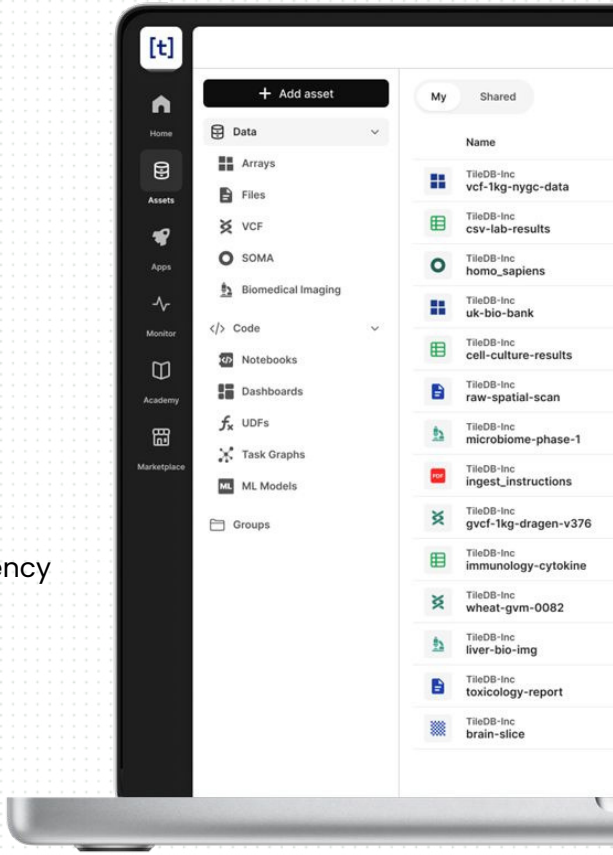


Biomedical Imaging

Why TileDB?

KEY TECHNICAL FEATURES

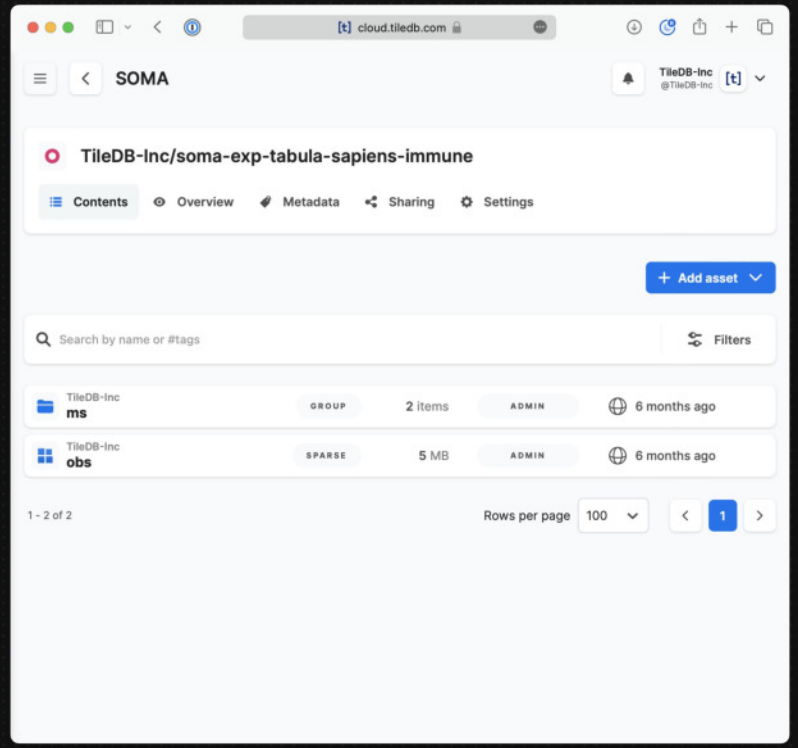
- ✔ **Universal Format** Sparse and dense ND arrays, key-value data, data frames
- ✔ **Flexible Indexing** Supports ints, floats, dates, and strings for versatile queries
- ✔ **Scalable Design** Uses tiling for selective memory loading during queries
- ✔ **Columnar** Enables efficient compression and selective attribute queries
- ✔ **Cloud-Native** Seamlessly works with local and cloud-based storage
- ✔ **High Performance** Fully parallelized I/O operations, multi-reader/writer concurrency
- ✔ **Cross-Platform** APIs for Python, R, Java, and many other languages

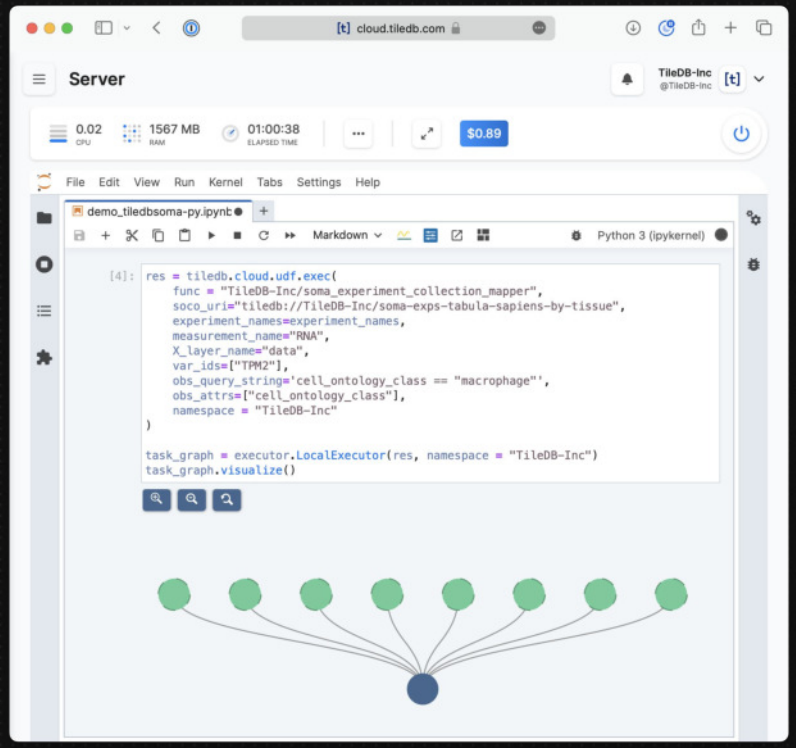


SUMMARY

TileDB offers **unparalleled flexibility** and **scalability**, enabling researchers to **handle diverse and growing datasets** with ease.

Scaling SOMA with TileDB Cloud





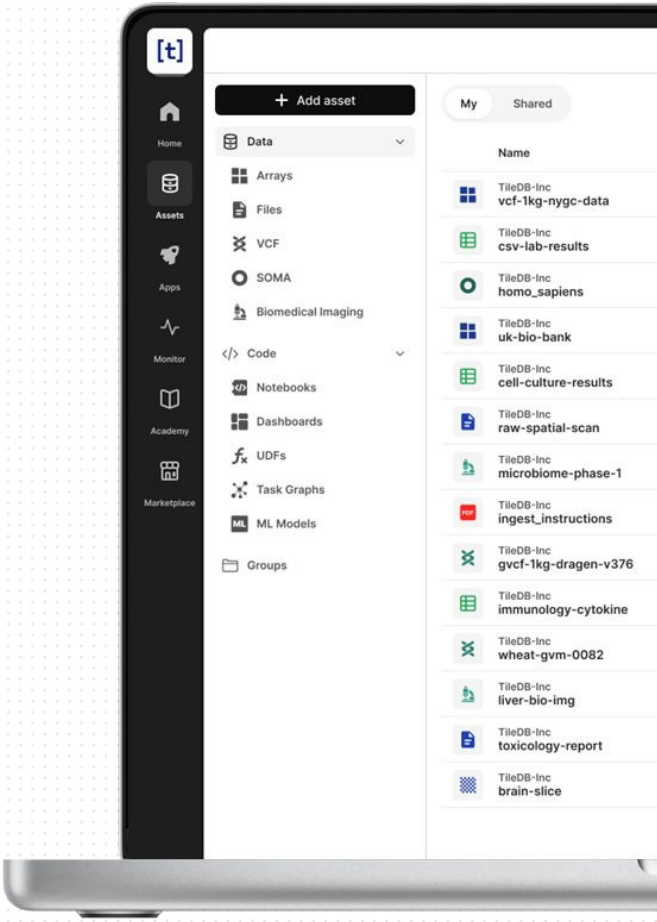
Why TileDB Cloud?

Commercial product for pharma, biotechs, and research institutions

- ✔ stores all types of multi-omics data as arrays and groups
- ✔ multiple ingestors
- ✔ provides a holistic catalog
- ✔ decentralizes data ownership
- ✔ centralizes governance and sharing of notebooks and dataset
- ✔ provides a common and scalable compute infrastructure and APIs
- ✔ allows definition and documentation of “data products” with the concept of shareable virtual “groups”

BENEFIT

Significantly reduced data engineering and infrastructure hassles



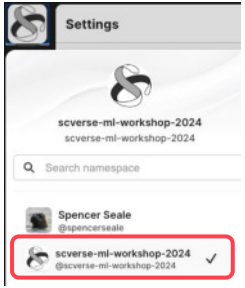
Let’s dive in!

Training and inference

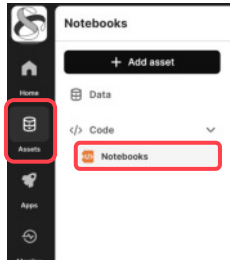
[tile]DB DESIGNED FOR DISCOVERY + Chan Zuckerberg Initiative

1 Login at cloud.tiledb.com

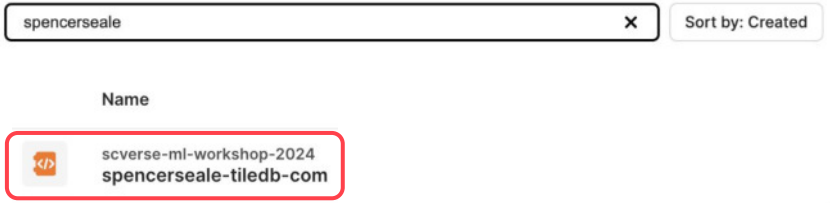
2 Switch to workshop Namespace (top-left)



3 Go to "Assets" & "Notebooks" (left)

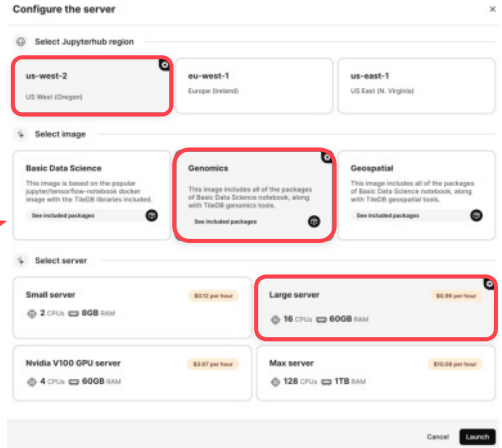


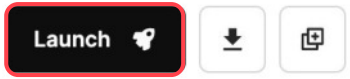
4 Search by your email username and select your notebook (top-right)



5 Launch your notebook with the configurations:

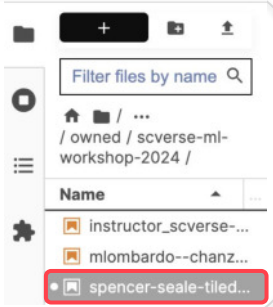
- **us-west-2**
- **Genomics**
- **Large server**





6 Display server

7 Select notebook



[t] + CZ
DESIGNED FOR DISCOVERY

[tile]DB DESIGNED FOR DISCOVERY + Chan Zuckerberg Initiative

Thank you!

TileDB

@TileDB

company/tiledb-inc

Ryan Williams, TileDB

ryan.williams@tiledb.com

@ryan-williams

CZI-wide

@ChanZuckerberg

company/chan-zuckerberg-initiative

Maximilian Lombardo, CZI

mlombardo@chanzuckerberg.com

@MaximilianLombardo

Similarity search UX survey: