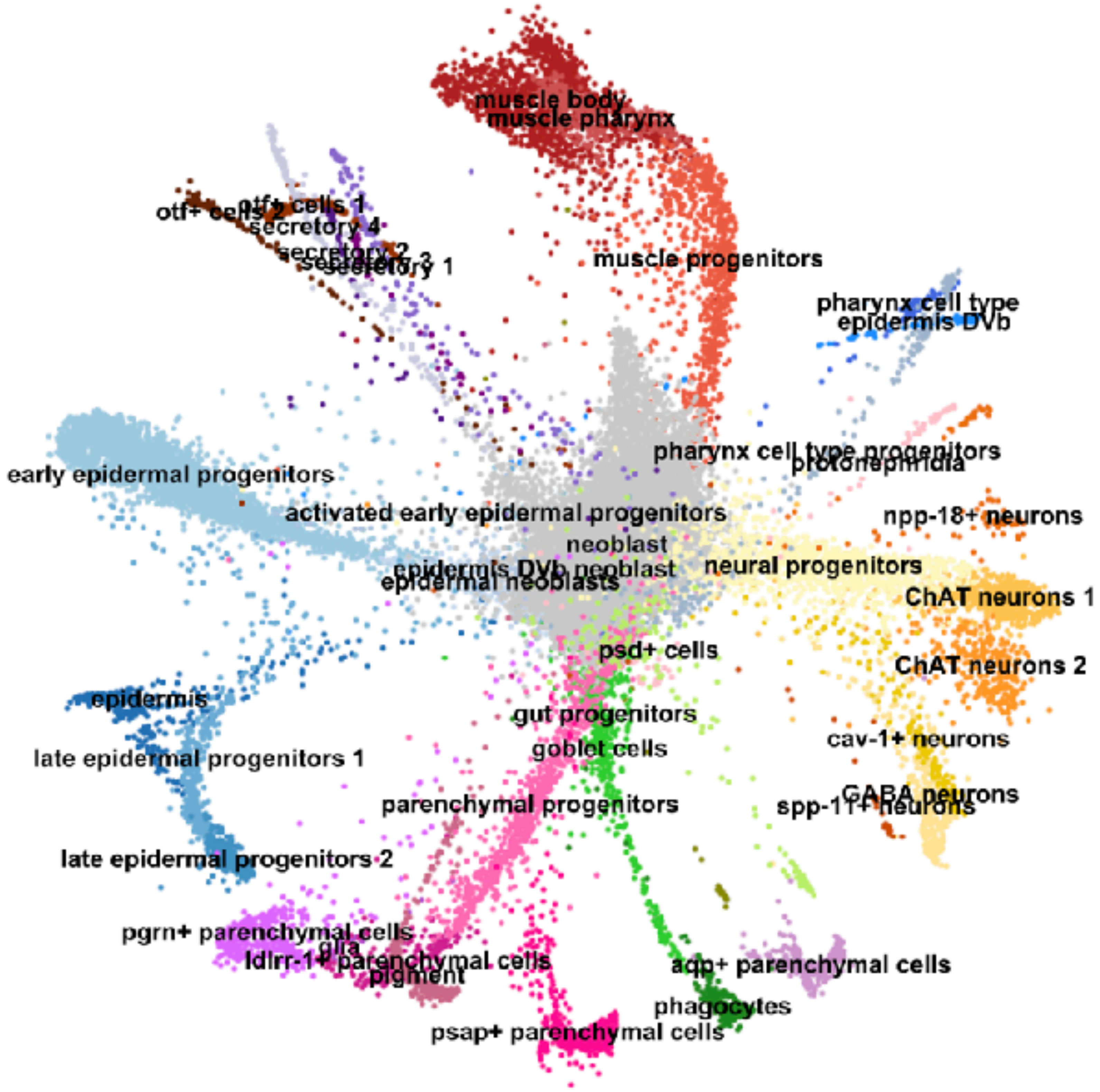
scverse

Scientific Python Sparse Summit

Isaac Virshup – 26th Sept. 2022

What is single cell RNA-seq?

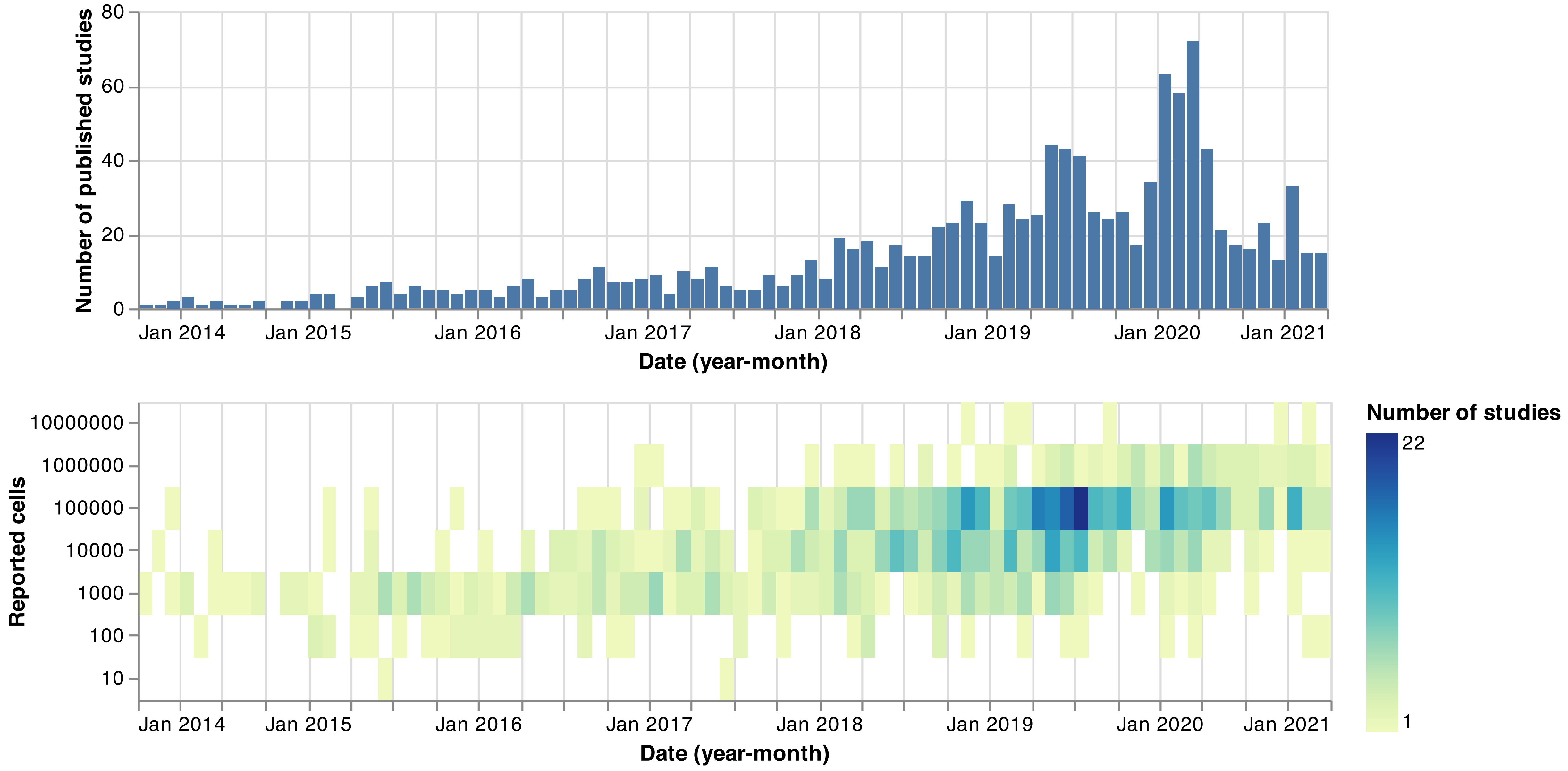
- High throughput assays on single cells
- Thousands to millions of observations (cells) on tens of thousands of variables (genes)
- Identify cell types, states, and dynamics



Wolf 2019, doi: 10.1186/s13059-019-1663-x

single cell data is sparse

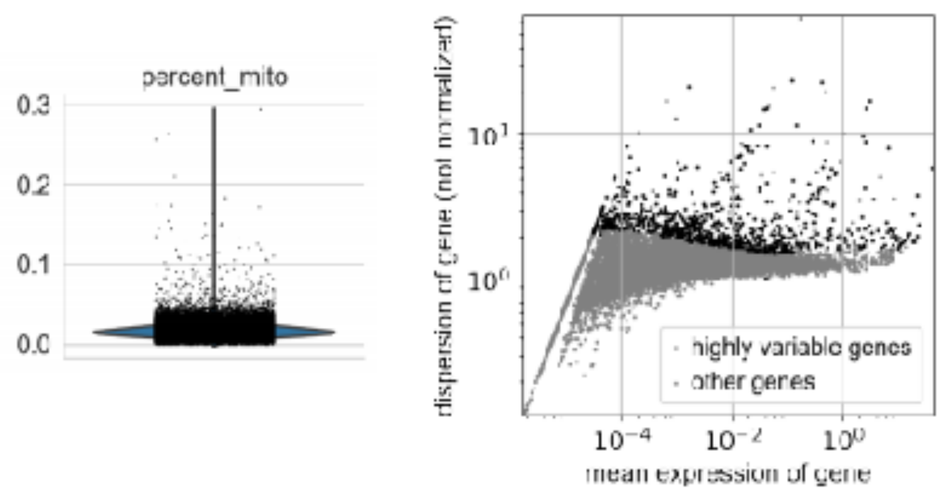
- The data is
  - ~20k possible genes
  - ~3-5k counts per cell
  - Increasing dataset sizes (1m+)
- Benefits of sparse representation
  - Memory usage
  - Compute time



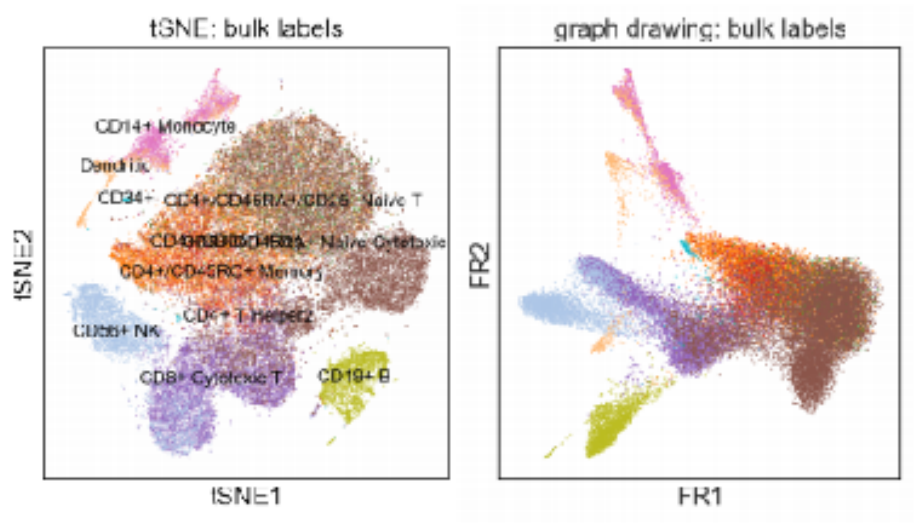
graphs are important in our domain

Normal Workflow

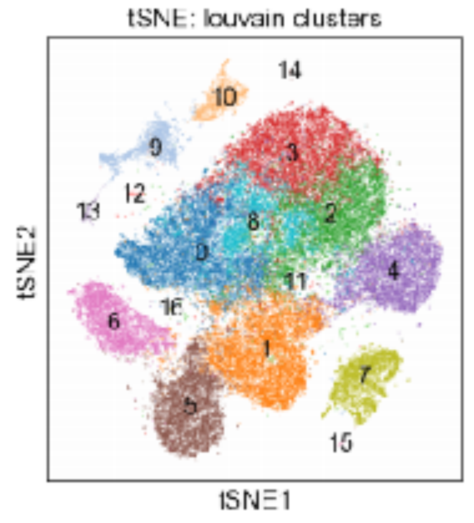
Preprocessing



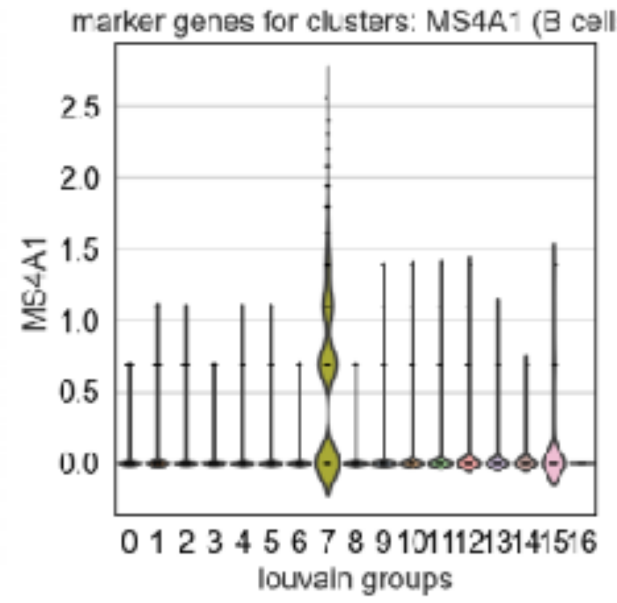
Viz via UMAP



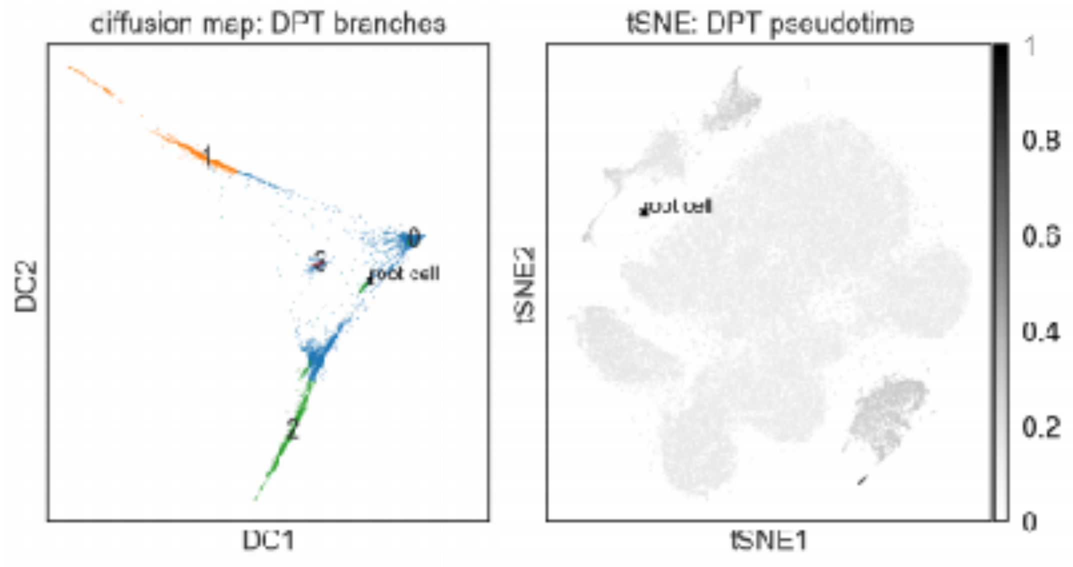
Clustering via
Community detection



Feature importance via
differential expression



Trajectory inference via
diffusion and optimal transport

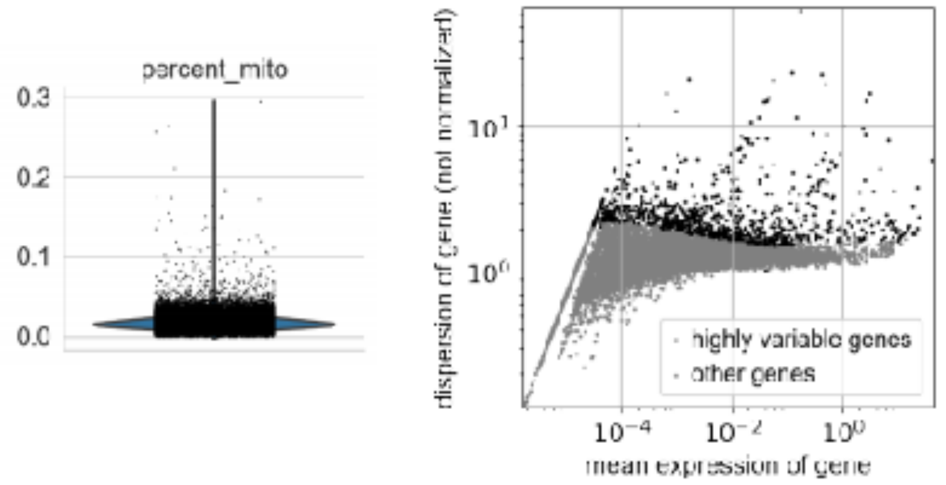


graphs are important in our domain

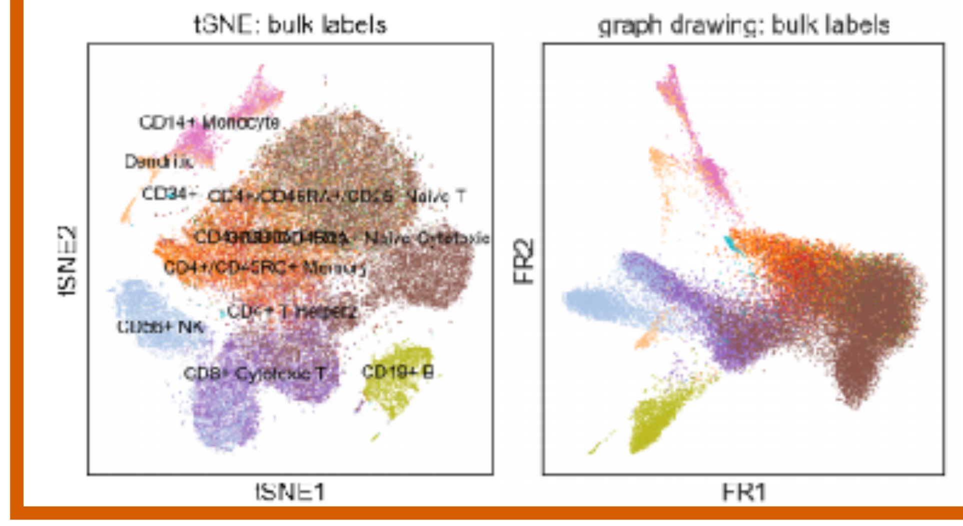
Normal Workflow

Graph based

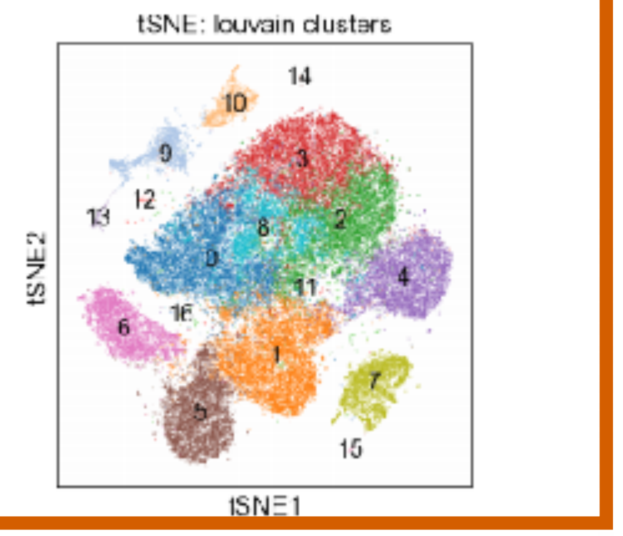
Preprocessing



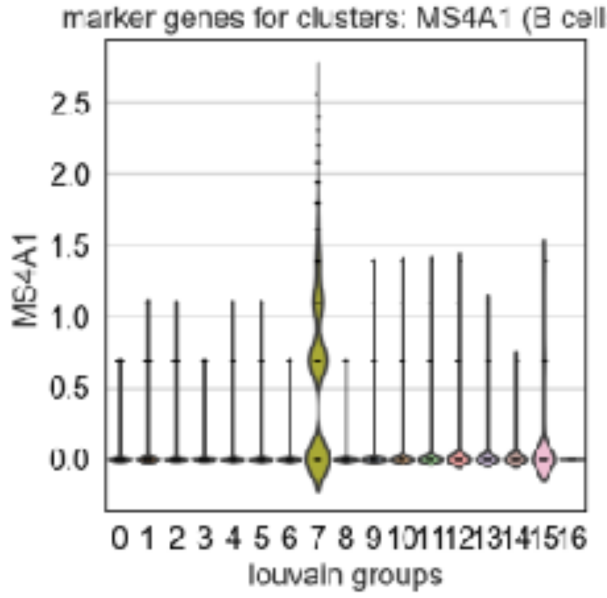
Viz via UMAP



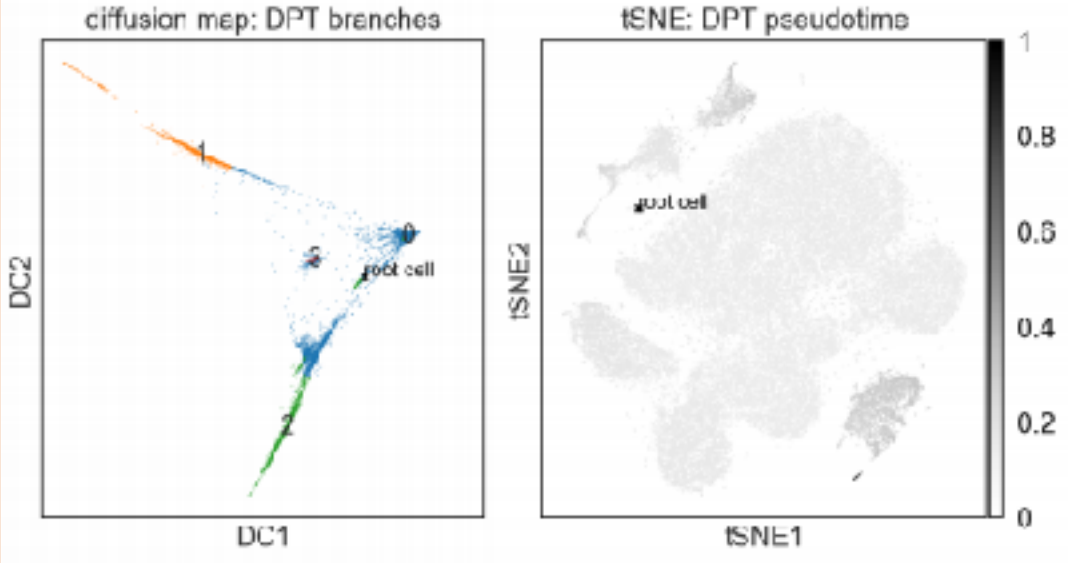
Clustering via Community detection



Feature importance via differential expression



Trajectory inference via diffusion and optimal transport

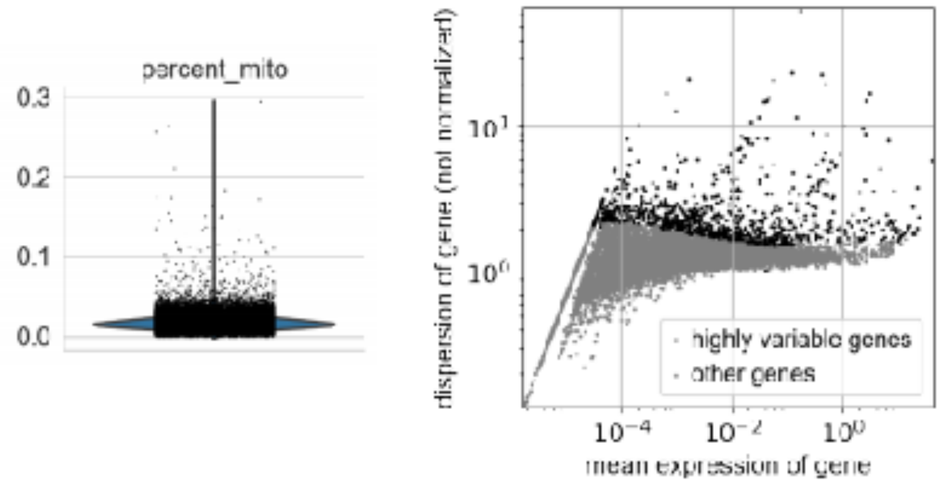


graphs are important in our domain

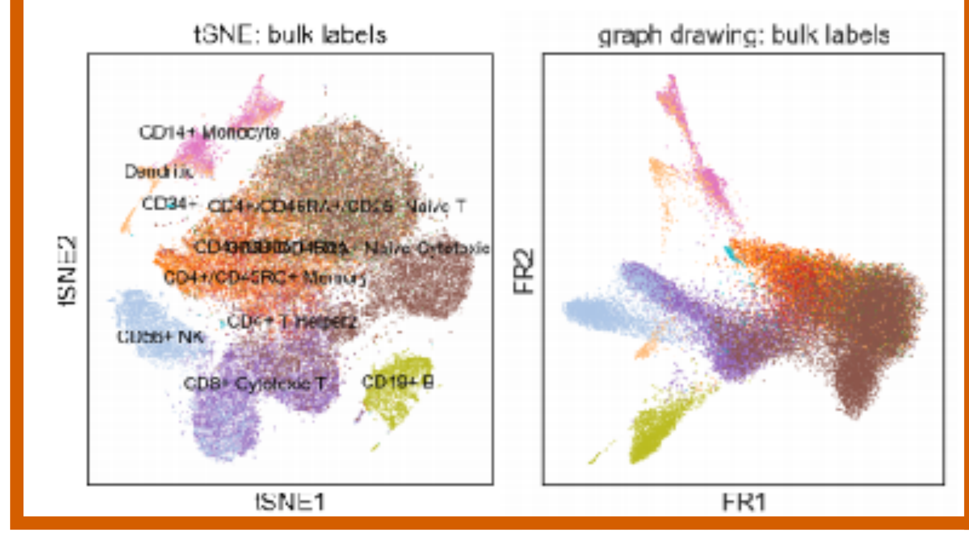
Normal Workflow

Graph based

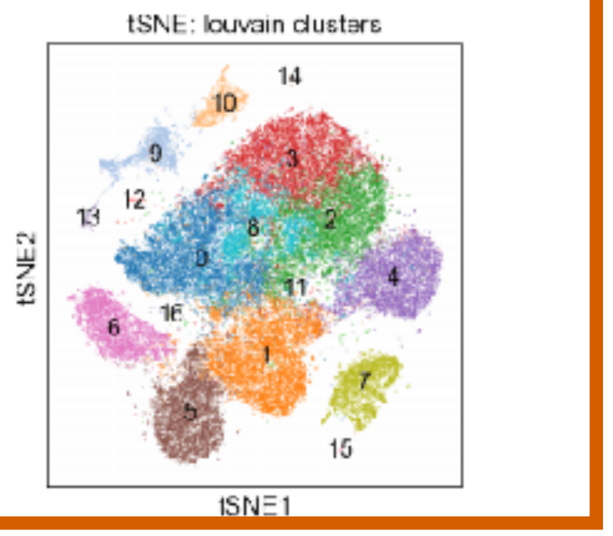
Preprocessing



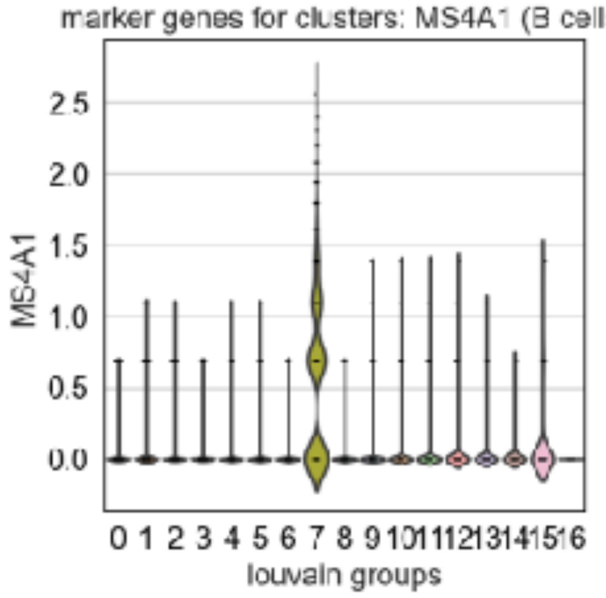
Viz via UMAP



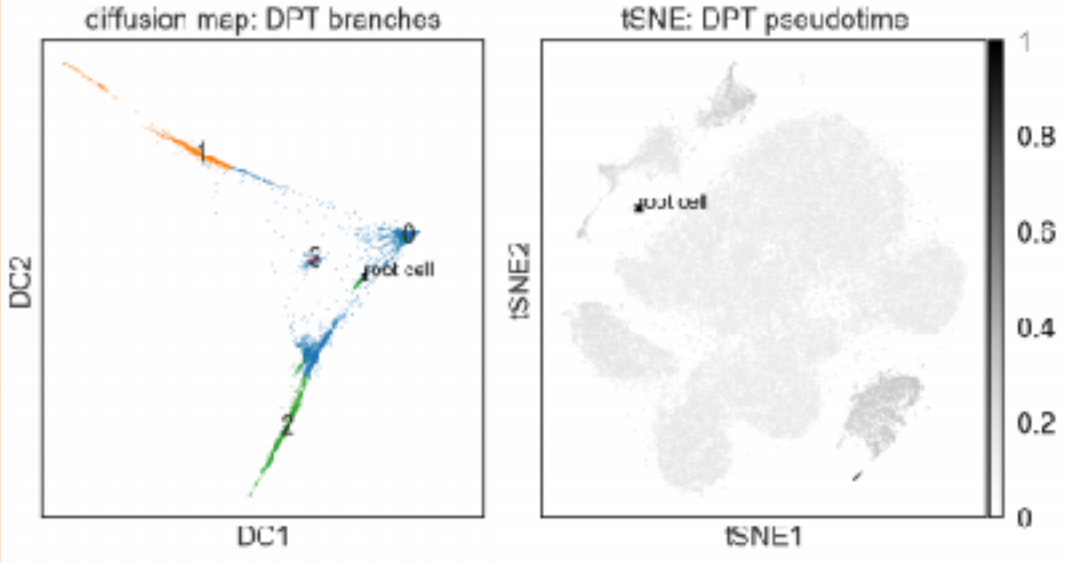
Clustering via Community detection



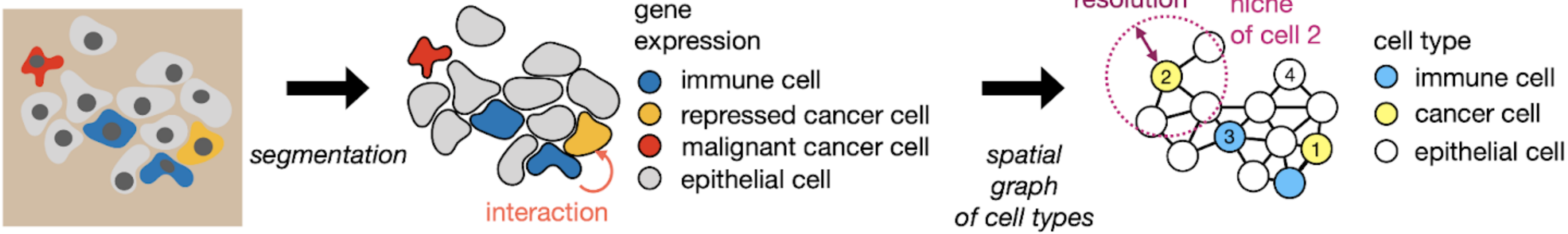
Feature importance via differential expression



Trajectory inference via diffusion and optimal transport



Spatial Graphs



gene expression

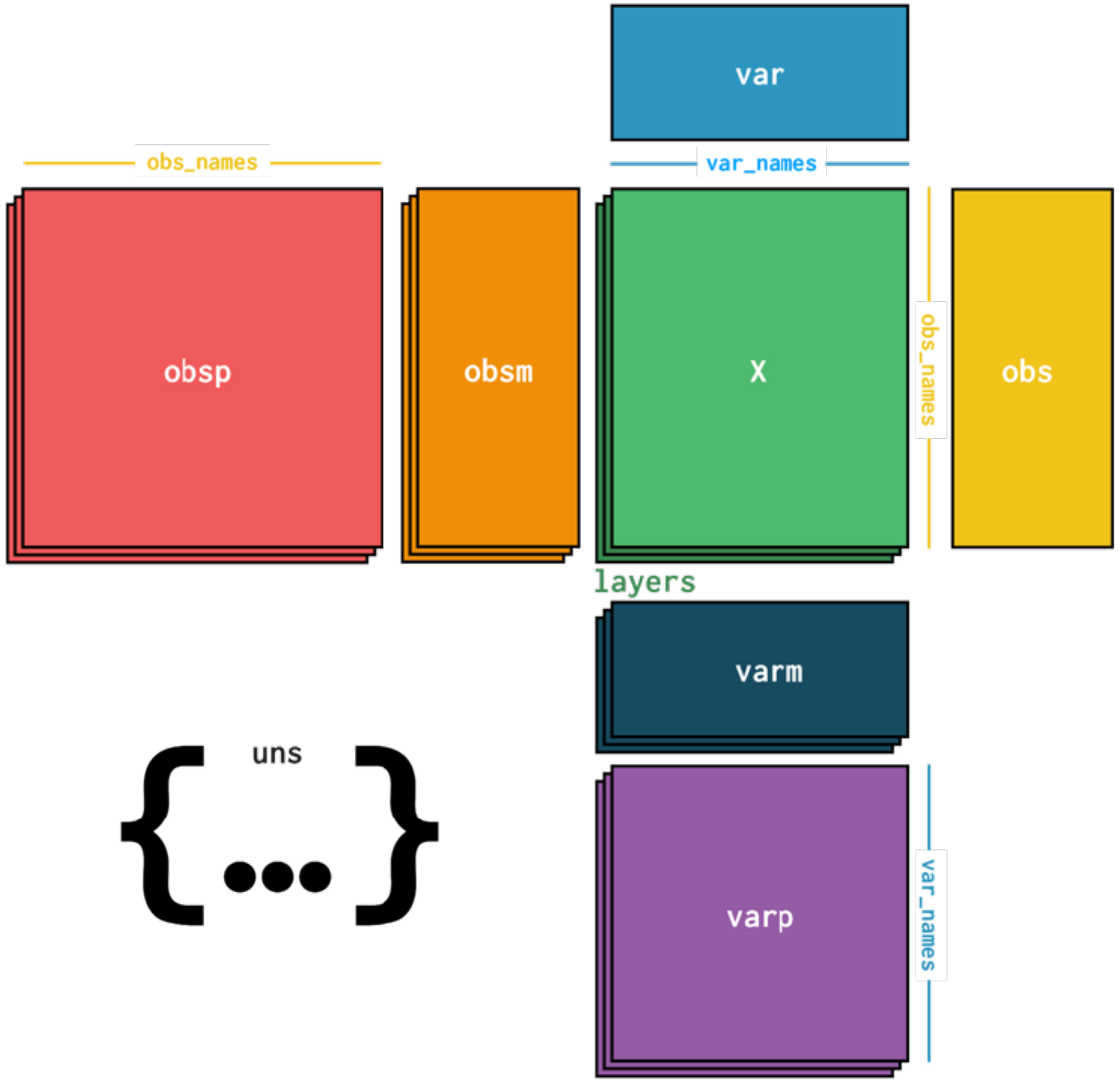
- immune cell
- repressed cancer cell
- malignant cancer cell
- epithelial cell

cell type

- immune cell
- cancer cell
- epithelial cell

How we handle sparse data

- AnnData container
  - Special casing sparse access
  - On disk representation (zarr, hdf5)
    - with out of core access (largely used for viz)
- Compute
  - PCA on sparse data
  - Fast sparse statistics (mean, var)
  - Fast Graph statistics (Morans I, etc)



What do we want

- Array API friendly sparse arrays
  - Eventual compatibility with xarray
  - Dask support (?)
- Consistent support throughout ecosystem (e.g. sklearn, pytorch-geometric)
- Out of core sparse arrays