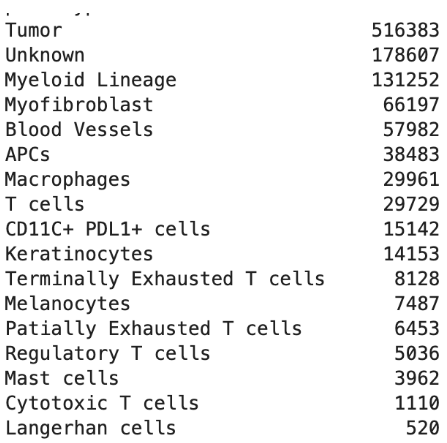
BMI 706 Project: Dataset and Tasks

Nkambule, Lethukuthula
Melnikas, Max
Wan, Guihong

1. **Explore the design space: we want to see what you tried and tell us what you think works and what doesn't.**

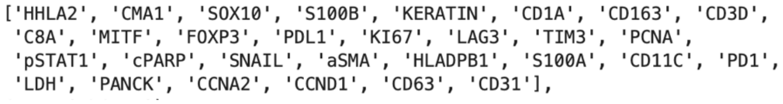
Our dataset consists of 1,110,585 cells in a tissue sample. Each cell has X and Y coordinates of their centroids along with phenotype labels whose distribution is given below. Exploring the spatial location of cells using phenotype as an encoding becomes a natural direction for our project. Additionally, our dataset includes antibody panel data for each cell on 30 key markers like SOX10, Ki67, etc. Understanding the relationship between antibody markers and cell phenotype is a key objective for our audience, making this an important aim of our visualization. Since the number of phenotypes in our dataset is a bit too large for an effective categorical encoding, we will also consider means of clustering cells together based on their antibody marker patterns.

Table 1: Cell Phenotype Counts

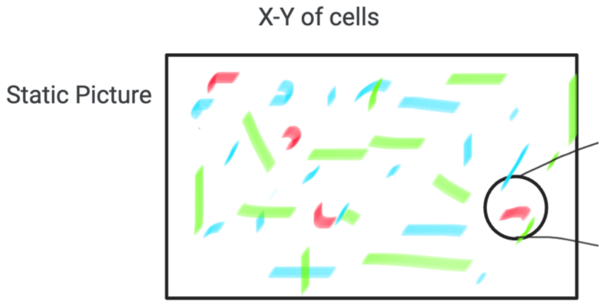


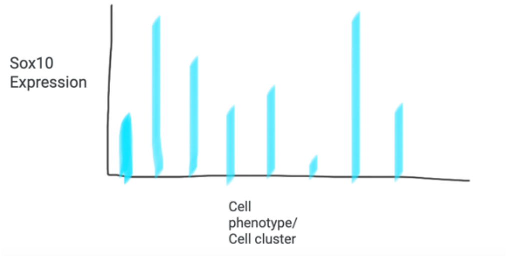
```
Tumor                          516383
Unknown                        178607
Myeloid Lineage                131252
Myofibroblast                   66197
Blood Vessels                   57982
APCs                            38483
Macrophages                     29961
T cells                         29729
CD11C+ PDL1+ cells              15142
Keratinocytes                   14153
Terminally Exhausted T cells     8128
Melanocytes                      7487
Patially Exhausted T cells       6453
Regulatory T cells               5036
Mast cells                       3962
Cytotoxic T cells                1110
Langerhan cells                   520
```

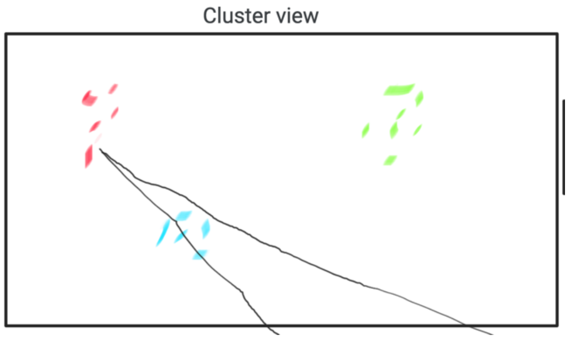
List 1: List of Markers

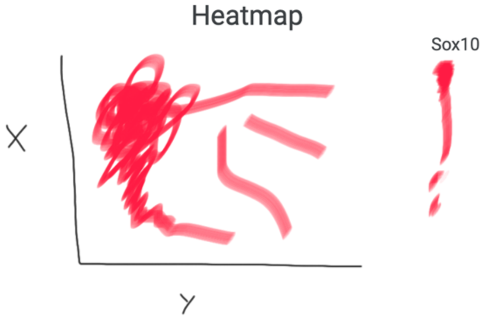


```
['HHLA2', 'CMA1', 'SOX10', 'S100B', 'KERATIN', 'CD1A', 'CD163', 'CD3D',
 'C8A', 'MITF', 'FOXP3', 'PDL1', 'KI67', 'LAG3', 'TIM3', 'PCNA',
 'pSTAT1', 'cPARP', 'SNAIL', 'aSMA', 'HLADPB1', 'S100A', 'CD11C', 'PD1',
 'LDH', 'PANCK', 'CCNA2', 'CCND1', 'CD63', 'CD31'],
```



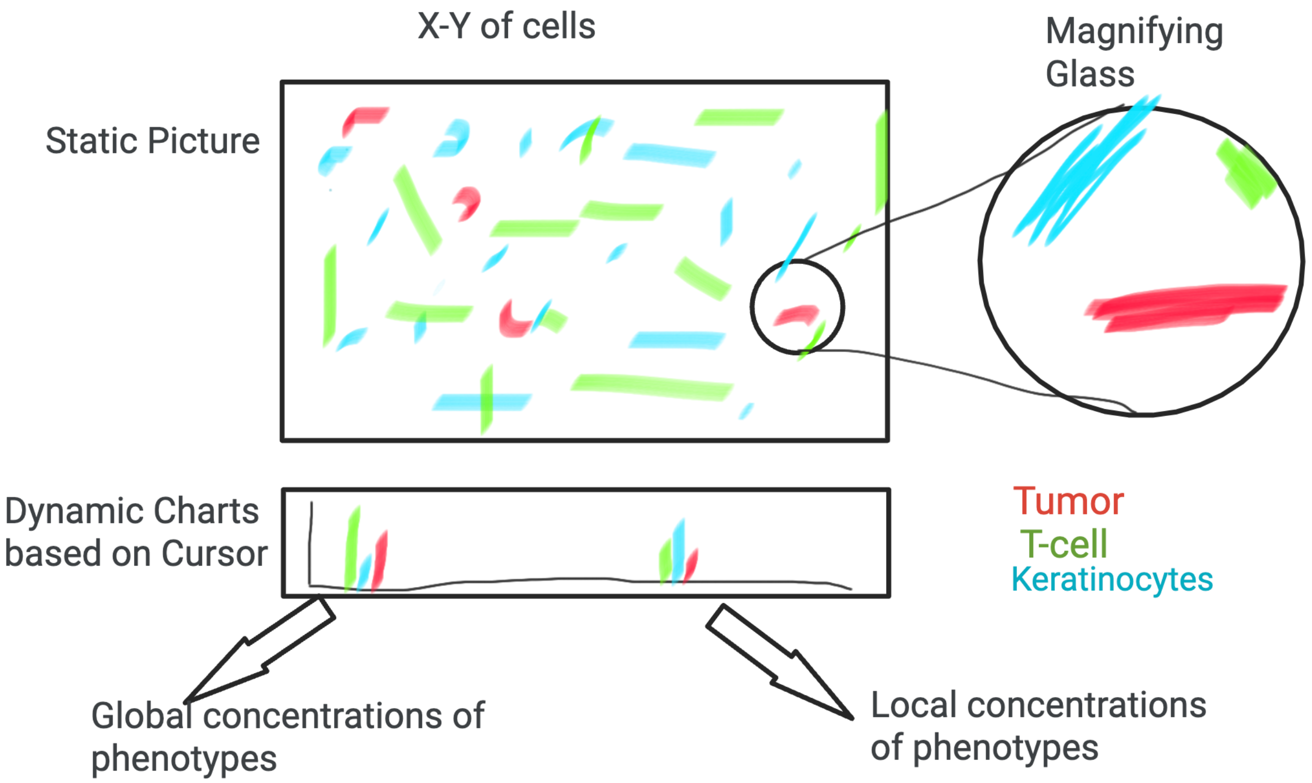


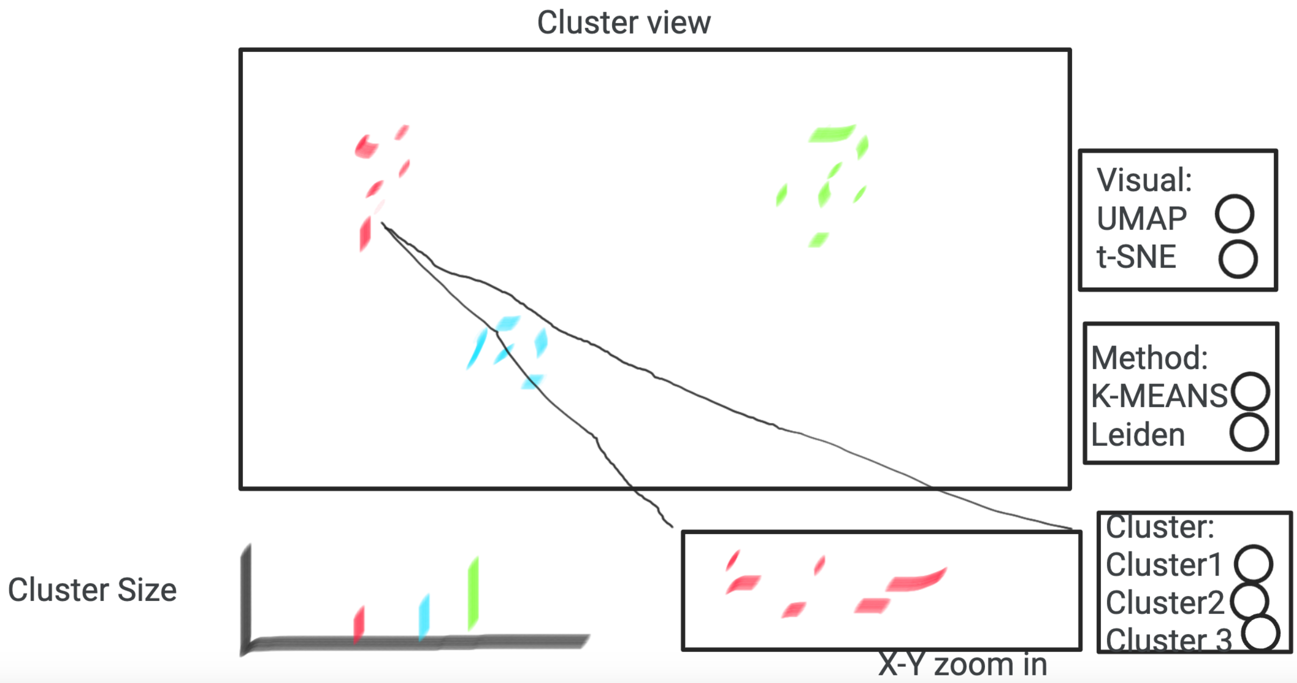




Cluster Size

2. Converge on a solution and annotate your sketches to show how the proposed visualizations, user interfaces, and other features support the tasks that you identified.


X-Y of cells
Magnifying Glass
Static Picture
Dynamic Charts based on Cursor
Tumor
T-cell
Keratinocytes
Global concentrations of phenotypes
Local concentrations of phenotypes



3. Describe potential visualization challenges

A challenge of this project is the complexity of the data. The original image is enormous in size (>100gb), which creates a barrier to using the original image within our visualization. As a workaround, we will likely have to recreate the spatial view of the data using the X, Y coordinates of the cells. Furthermore, the large number of individual cells (>1,000,000) will likely pose a challenge for effectively communicating information visually. We will consider sampling the cells or constricting the dataset to a certain X-Y grid.

One hope for our project is to highlight local patterns of cell phenotypes. Our vision involves the user hovering their cursor over spatial neighborhoods that they want to explore. Since this feature is dynamic and uses the mouse location as a parameter, producing on-demand statistics of the local neighborhood may be computationally challenging, if not unfeasible.

4. Write two or three short paragraphs describing how you are planning to implement your application and how different components of your visualization will be interacting with each other. Note that your design may go beyond what you will actually be able to implement in your Streamlit app.

We hope to develop two main views of the data, each with its own supplementary figures. First will be a X,Y coordinate map that shows the spatial relationships between individual cells. Discrete colors can be used to encode one more categorical dimension into the visual that can be used for cell phenotypes or cell clusters. The user will be able to interact with the map by havering their mouse over the visual. By changing the location of mouse, we hope to implement a dynamic bar chart that shows how the current local region compares to the global dataset in terms of phenotype/cluster concentrations. We envision two ways of defining local regions: uniform and gaussian kernels. An additional opportunity for interaction may be for the user to manipulate the parameters of these kernels via scroll bars.

The second main view of the data will be a spatial relationship of cells' gene expression patterns in a lower dimensional space, like UMAP or t-SNE. Furthermore, clustering techniques can be applied to segment cells. Radio button selections will be available for the user to toggle between dimensional reduction methods (UMAP, t-SNE) and clustering methods (K-means, Leiden). A supplementary bar chart can inform about relative sizes of the clusters. There will also be a radio button toggle that the user can use to highlight X-Y coordinates of cells in a cluster or interest.