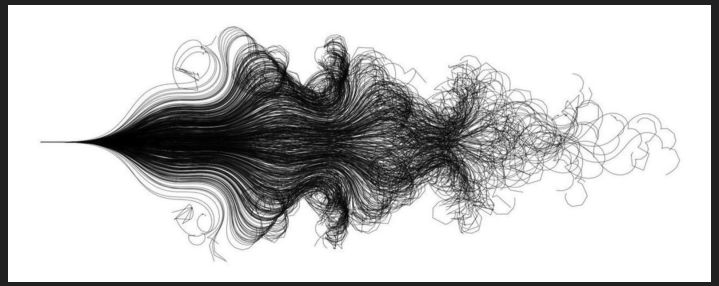


# Visualization of high-dimensional data

May 9th, 2025

IMAT Lecture series

Sandhya Prabhakaran
Anderson lab
Integrated Mathematical Oncology lab
Moffitt Cancer Center

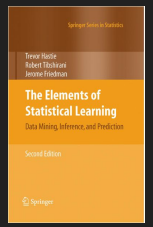
# Objective of this talk

- Different kinds of data
- What is visualization?
- Why is visualization of data important?
- Some examples of high-dimensional data visualization
- Current issues

| | Cell1 | Cell2 | ... | CellN |
|---|---|---|---|---|
| Gene1 | 3 | 2 | . | 13 |
| Gene2 | 2 | 3 | . | 1 |
| Gene3 | 1 | 14 | . | 18 |
| ... | . | . | . | . |
| ... | . | . | . | . |
| ... | . | . | . | . |
| GeneM | 25 | 0 | . | 0 |

# Data

| Feature | High-Dimensional Data | Low-Dimensional Data |
|---|---|---|
| Number of Features | Large, often exceeding the number of samples | Small, fewer features than samples |
| Complexity | High, computationally challenging | Low, easier to analyze |
| Data Sparsity | Data can be sparse in high-dimensional spaces | Data is typically denser |
| Visualization | Difficult to visualize and understand | Easier to visualize and understand |
| Dimensionality Reduction | Often requires dimensionality reduction techniques | Typically not required |




The Elements of Statistical Learning :
https://hastie.su.domains/ElemStatLearn/
(chapter 18, page 649

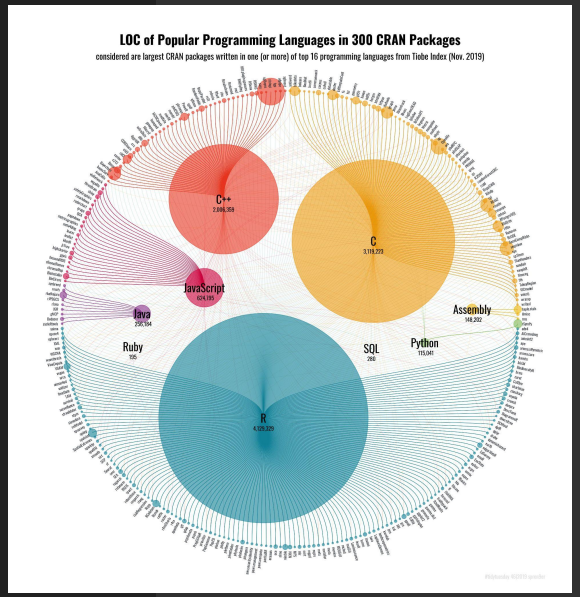
# Visualization

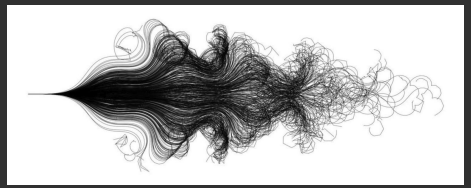
- The act of creating mental images or visual representations to convey information or ideas.

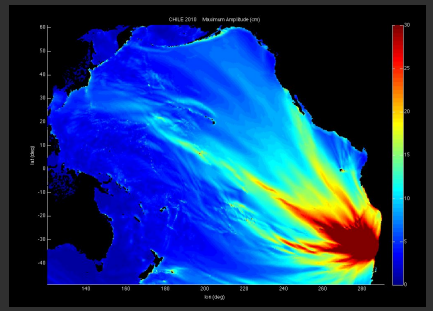
- Mental visualization
- Data visualization
- Creative visualization
- Information visualization

Examples of

# Visualization





https://www.peoplemov.in/

# Why is data visualization important?

*Visualization is an art; to master it one needs patience, focus and dedication*

# Why is data visualization important?

*Visualization is a craft, one has to be sufficiently sane and sober to do this.*


DatasauRus dozen

"When only looking at the statistics, we would have probably wrongly assumed that the datasets were identical. This example highlights why it is important to visualize data and not just rely on summary statistics."

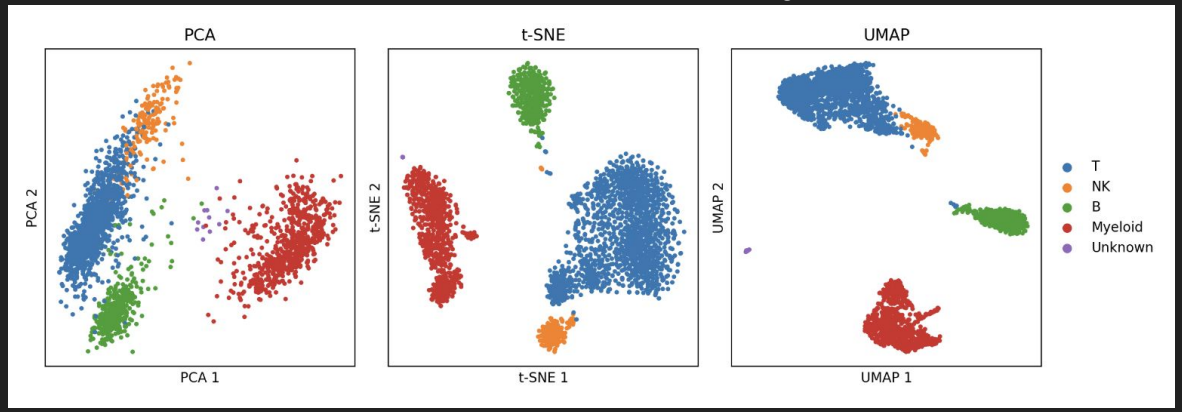
# Why is data visualization important?

- Enhances understanding
- Improves engagement
- Aids in the extraction of insights from data

- It simplifies complex information
- Makes it more accessible
- Facilitates quick comprehension and retention

# Examples of high-dimensional data visualization

# Data visualizations based on dimensionality reduction




```python
# Core
import numpy as np
import pandas as pd
pd.plotting.register_matplotlib_converters()
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
sns.set(style='darkgrid', font_scale=1.4)
import plotly.express as px

# Sklearn
from sklearn.decomposition import PCA
from sklearn.manifold import TSNE
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis as LDA
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler

# UMAP
import umap
import umap.plot
```

**Principal Component Analysis (PCA)**

- First dimensionality reduction technique discovered by Karl Pearson
- ~ **1901**.
- It is **fast**, **easy to implement** and **easy to interpret**.
- PCA works by finding a low dimensional subspace that **maximises the variance** of the data in that subspace and performing a **linear projection**. This basically means the data will be as **spread out** as possible, without changing the relationship between the data points.

**t-distributed Stochastic Neighbor Embedding (t-SNE)**

- By Laurens van der Maaten and Geoffrey Hinton (2008).
- This is a stochastic method
- Pairwise iterative method
  - Datapoints move closer or further away from each other depending on their 'similarity'
- Non-linear mapping
  - Harder to interpret
  - Effective at 'unravelling' highly non-linear data.

**Uniform Manifold Approximation and Projection (UMAP)**

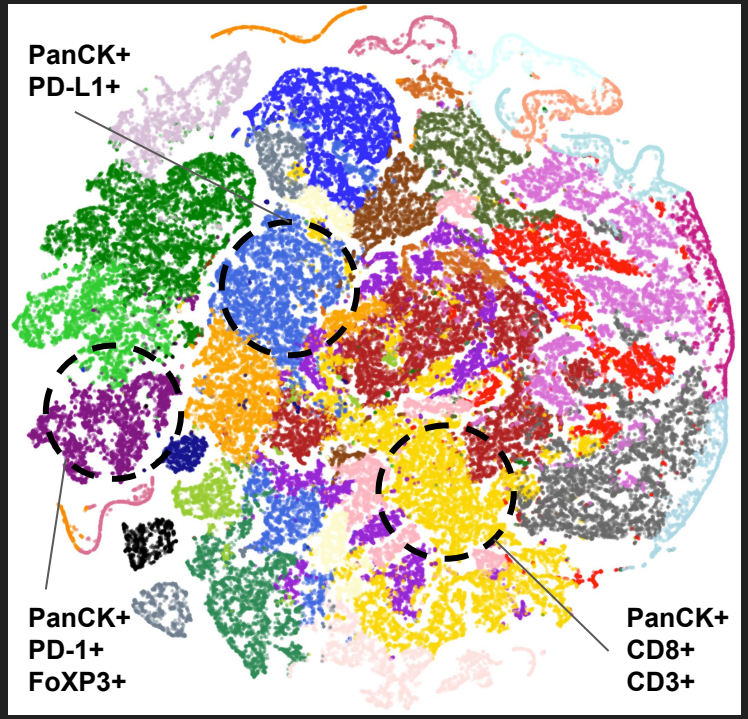
- By Leland McInnes, John Healy and James Melville (2018).
- Non-linear mapping
  - Preserves clusters
  - Is significantly faster.
  - Preserves global structure of the data compared to t-SNE.

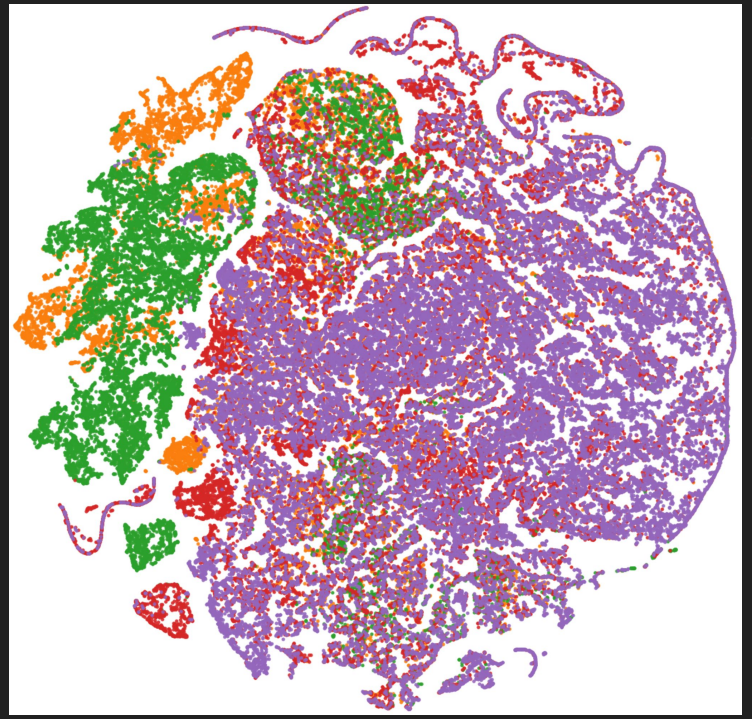
```python
# PCA
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X)
```

```python
# t-SNE
tsne = TSNE(n_components=2)
X_tsne = tsne.fit_transform(X)
```

```python
# UMAP
um = umap.UMAP()
X_fit = um.fit(X)
X_umap = um.transform(X)
```

# Cell segmentation approach shows preserved cellular neighborhoods across PD and SD patients



PD pre (16 FoVs)
PD on (15 FoVs)
SD pre (30 FoVs)
SD on (28 FoVs)

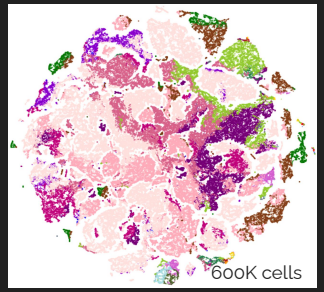
9 patients -
4 PD, 5 SD
121,000
cells

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

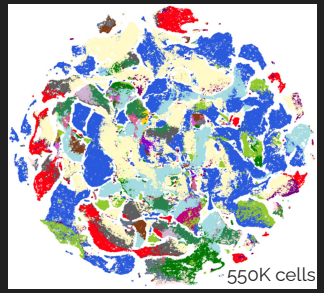
https://www.biorxiv.org/content/10.1101/2022.10.22.513219v2

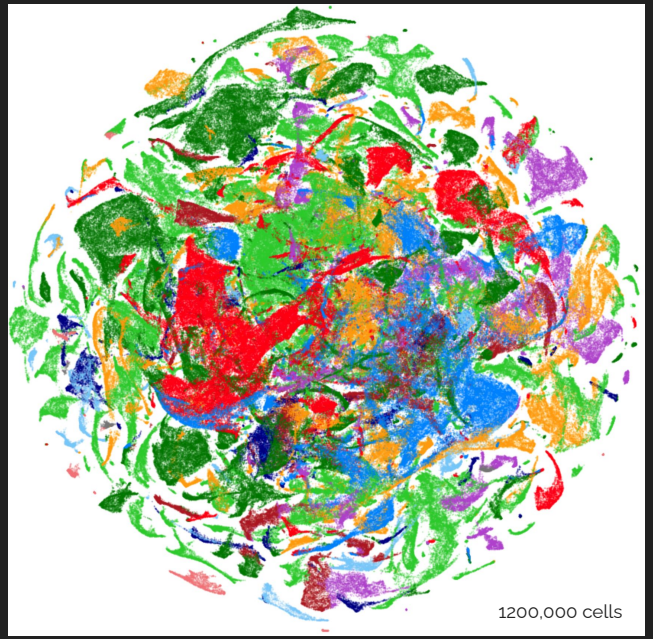
# Example t-SNEs and UMAPs

**Lung2 ArmA Pre-Tx**

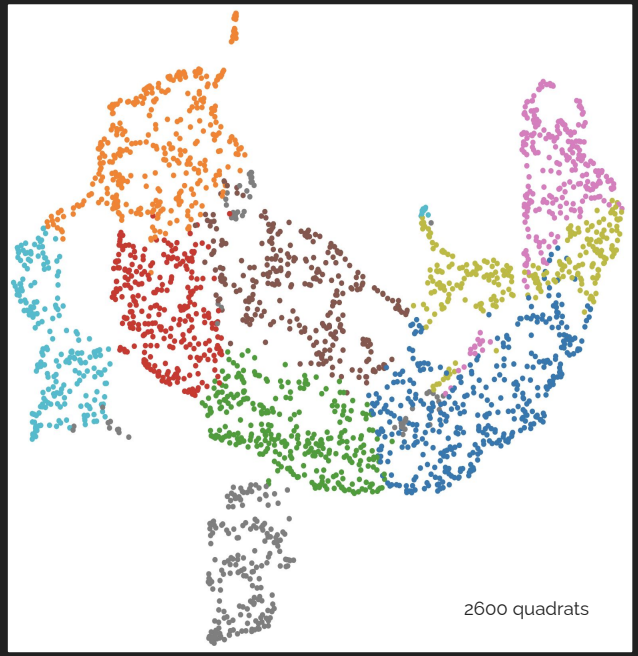


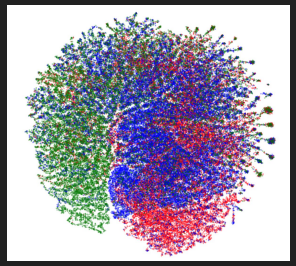
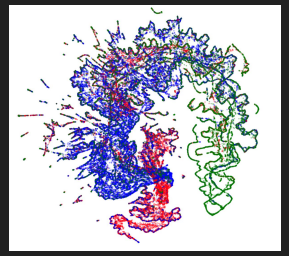
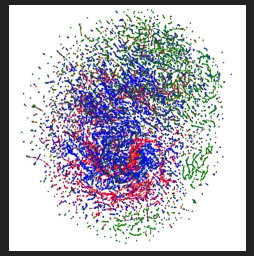
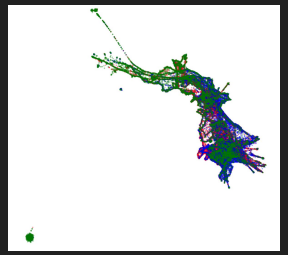
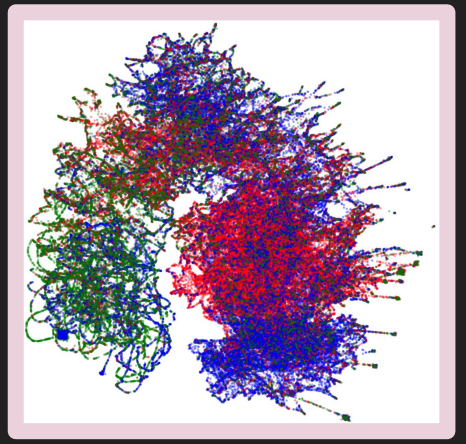
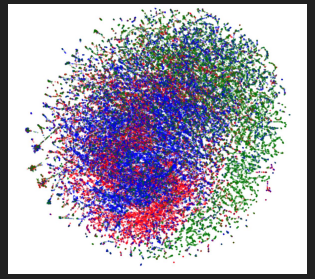
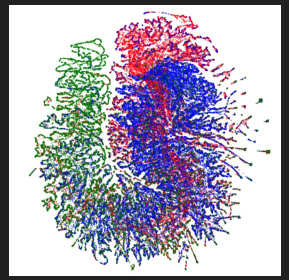
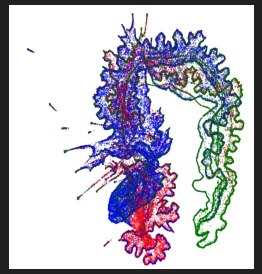
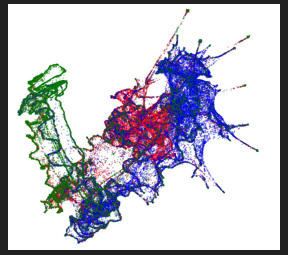
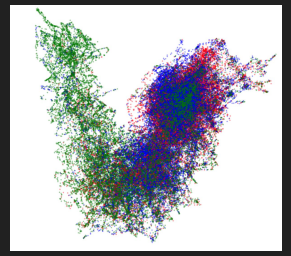
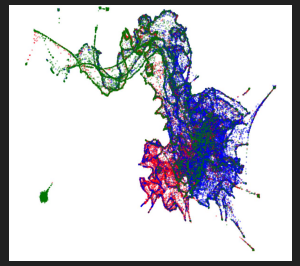
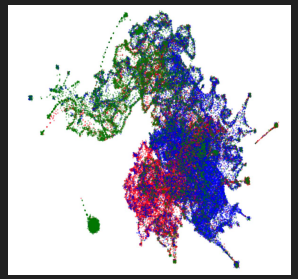
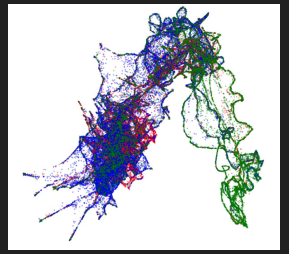
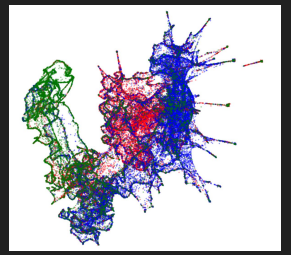

**Lung2 ArmA During-Tx**



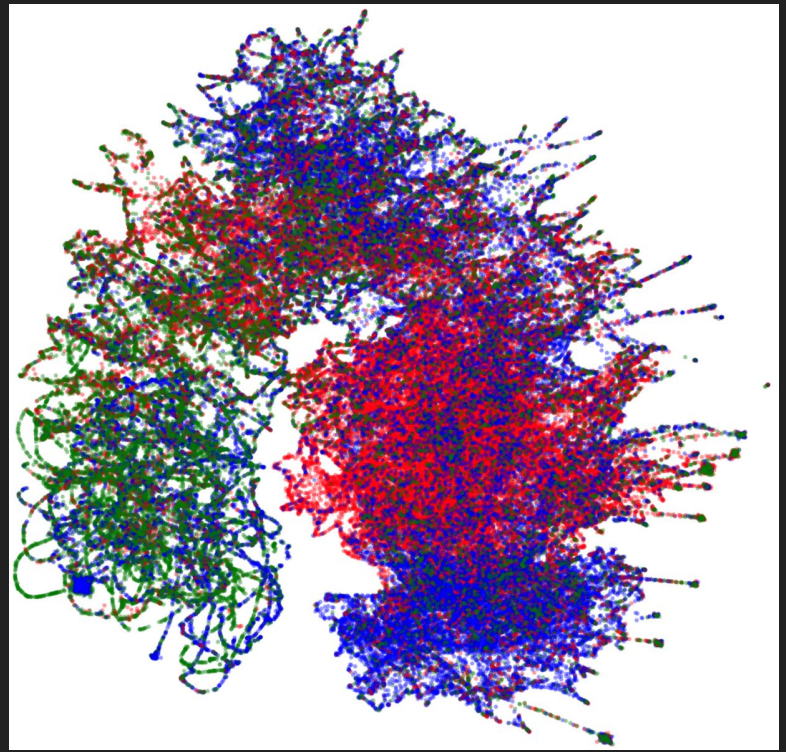

**Lung1 ArmB pre and during-Tx**



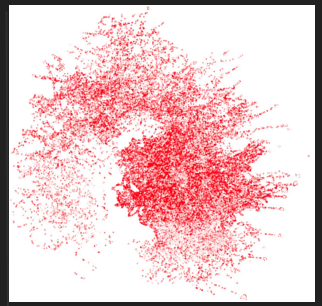
PD
SD
PR

- Lung2 data (3SD/3PD/3PR)
- UMAP renderings for neighbours = 10-100

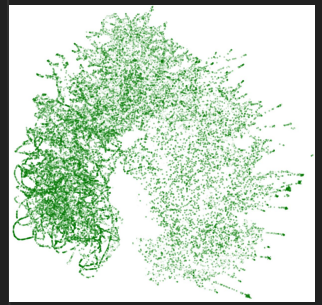


PD
SD
PR

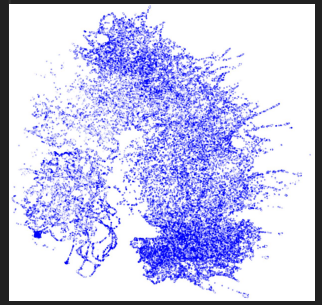
SD



PD



PR

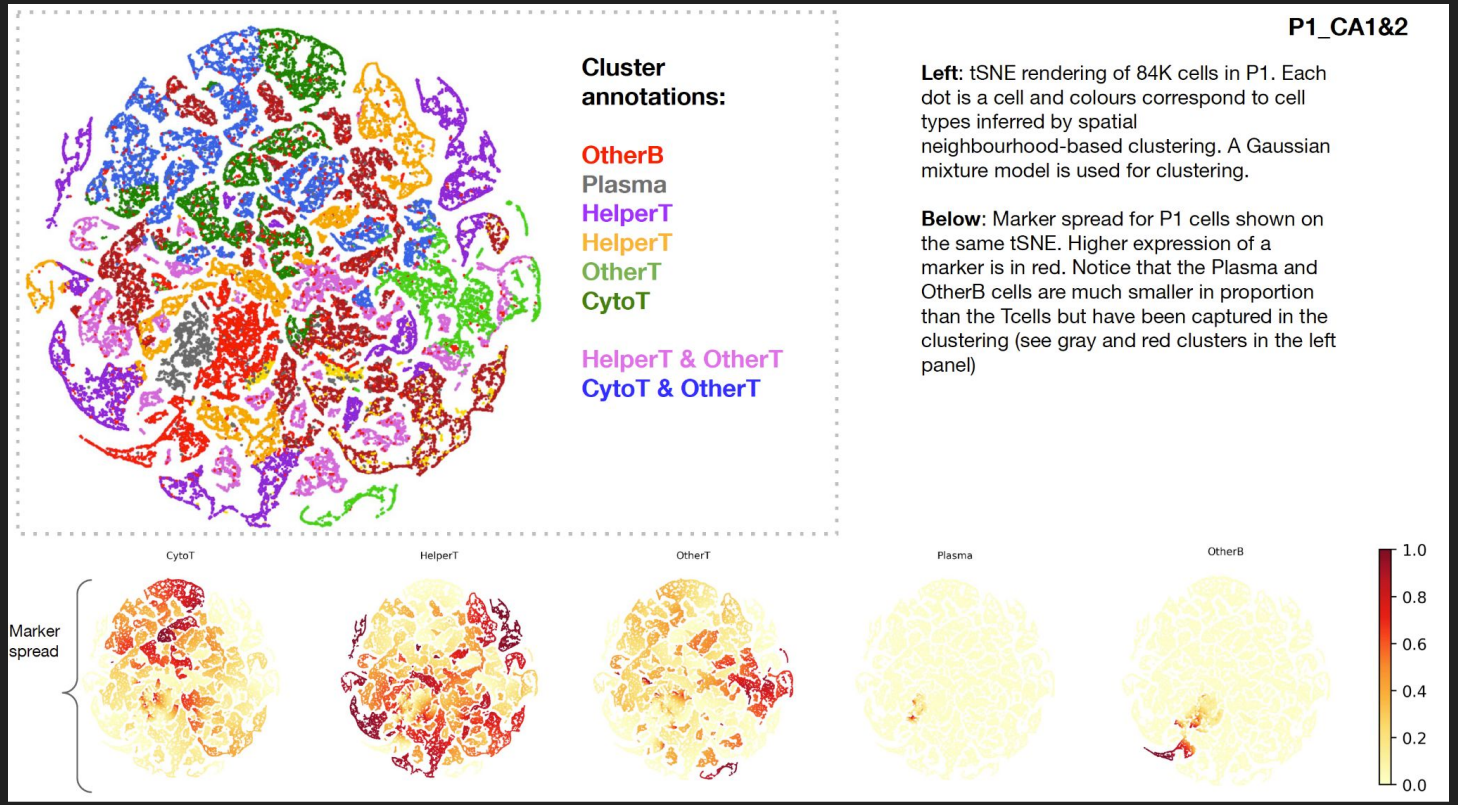


Using a combination of

- Lilliefors test for normality
- Bhattacharyya distance
- One-way ANOVA test

we prove that these three distributions are statistically significant

# Analysis of TMA: Clear cell carcinoma, endometrioid type high grade, endometrioid type low grade and serous




107 patients

84K cells

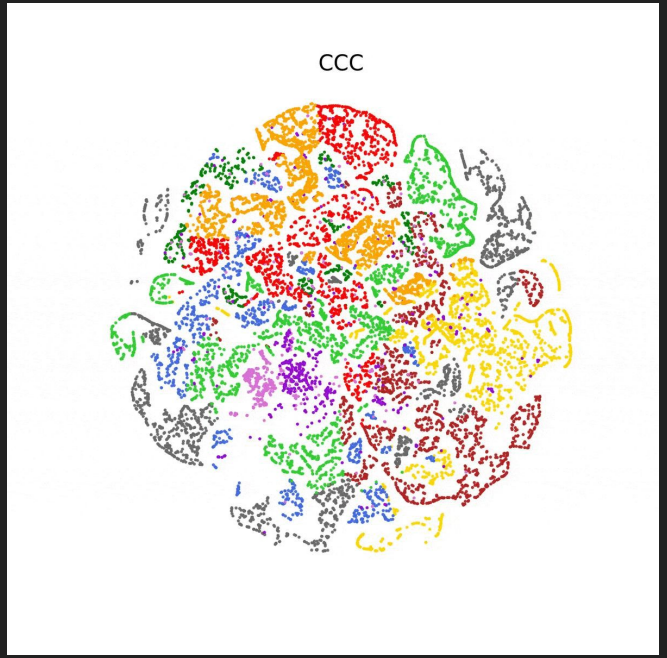
210 TMA cores

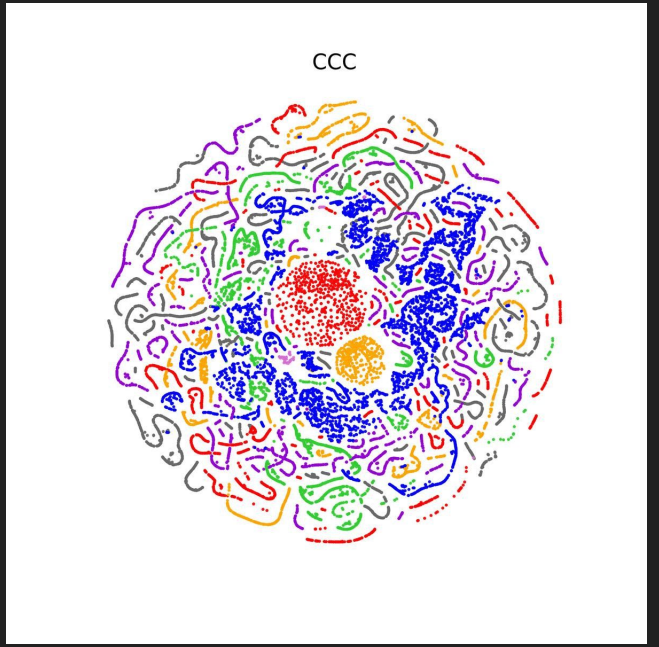
4 ovarian cancer types

We were able to *visually* reinforce that immune protection was associated with polymeric immunoglobulin receptor (pIgR) by tumor cells and its occupancy by IgA and, to a lesser extent, IgG.

CANCER RESEARCH
https://doi.org/10.1158/0008-5472.CAN-21-2376

# Analysis of TMA: Clear cell carcinoma, endometrioid type high grade, endometrioid type low grade and serous

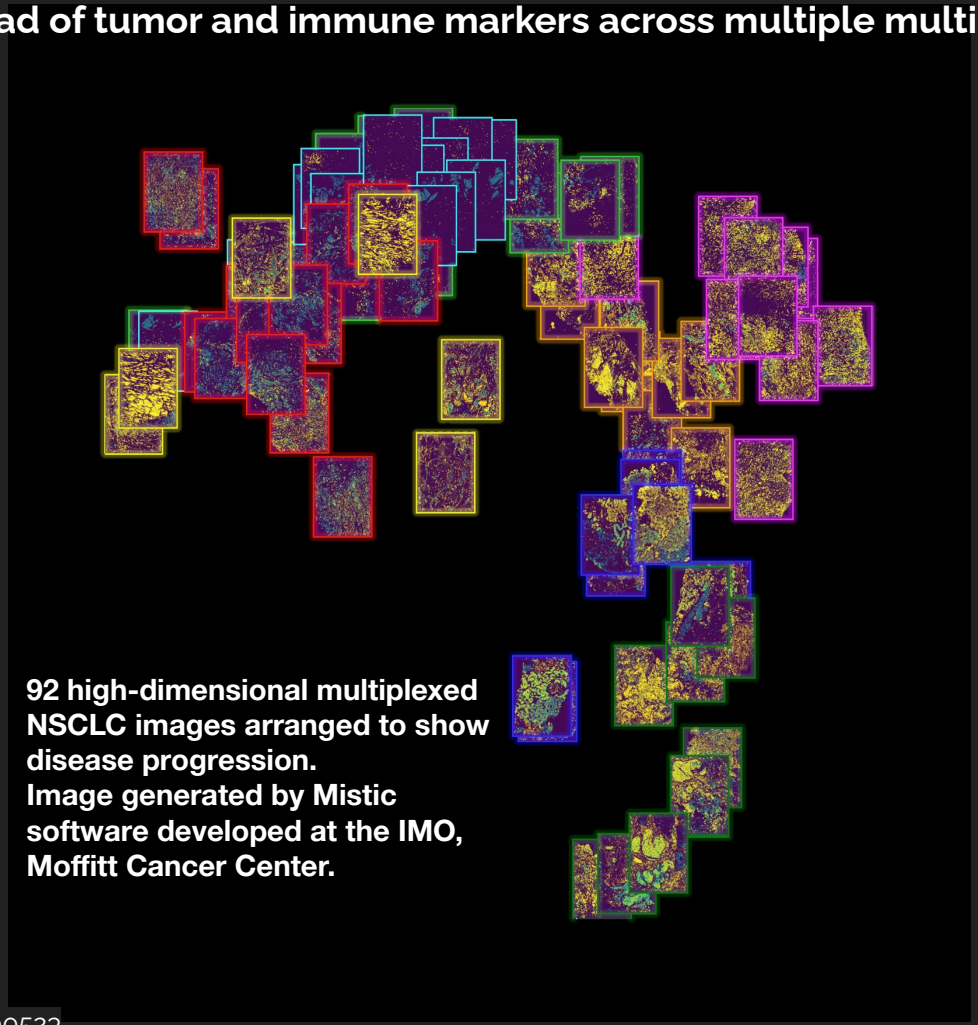



CANCER RESEARCH
https://doi.org/10.1158/0008-5472.CAN-21-2376

# Interactive visualizations

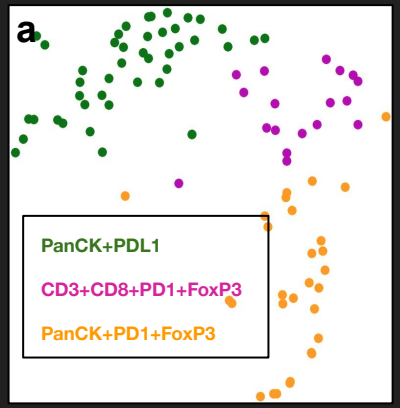
# Mistic: Visualise the spread of tumor and immune markers across multiple multiplexed images

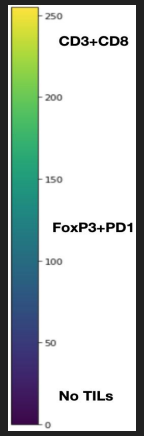


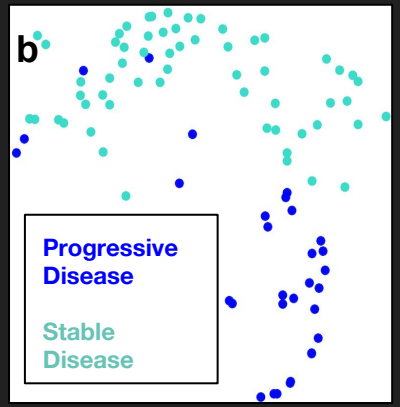
92 high-dimensional multiplexed NSCLC images arranged to show disease progression.
Image generated by Mistic software developed at the IMO, Moffitt Cancer Center.

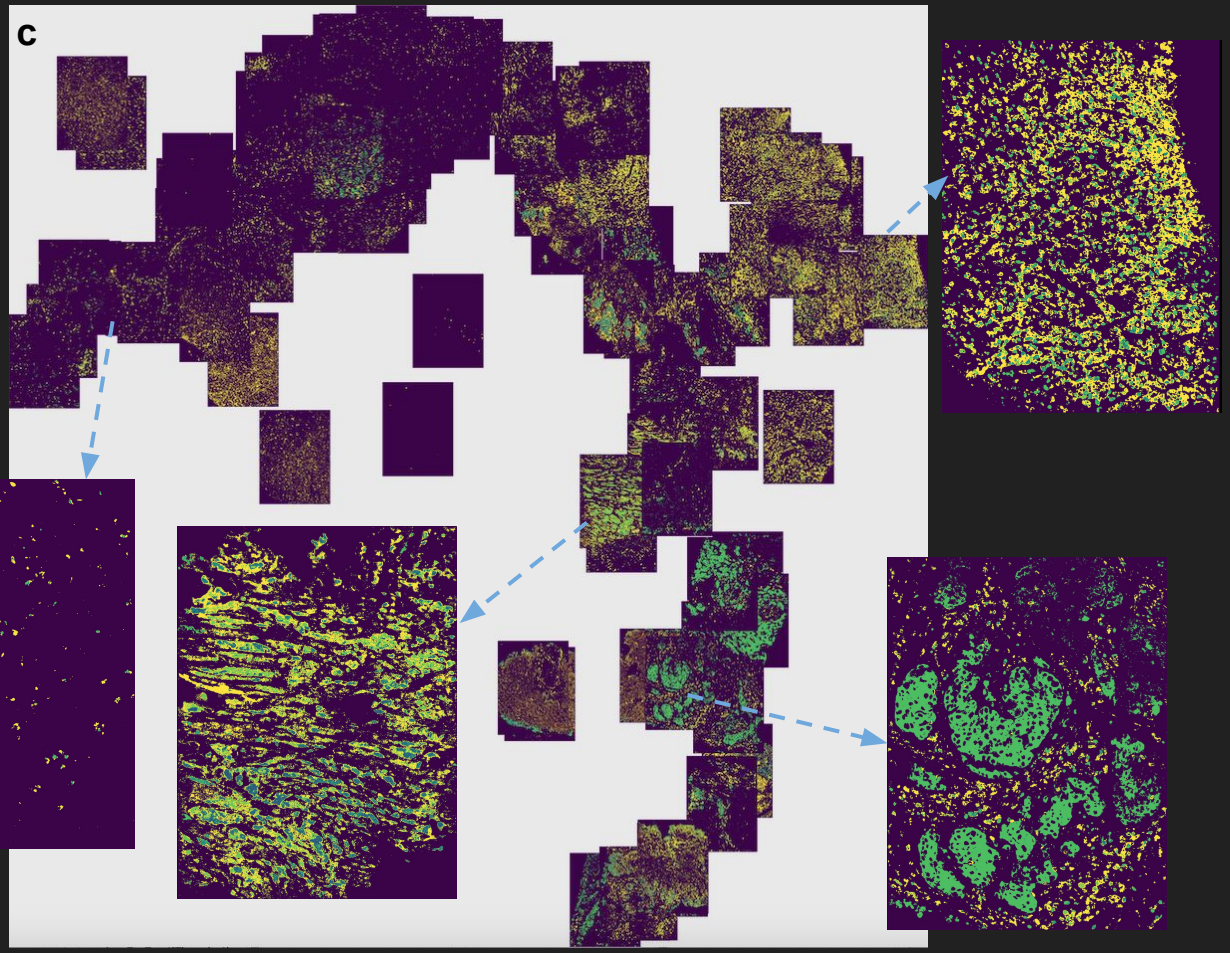
Patterns
https://doi.org/10.1016/j.patter.2022.100523

# Mistic: Visualise the spread of tumor and immune markers across multiple multiplexed images



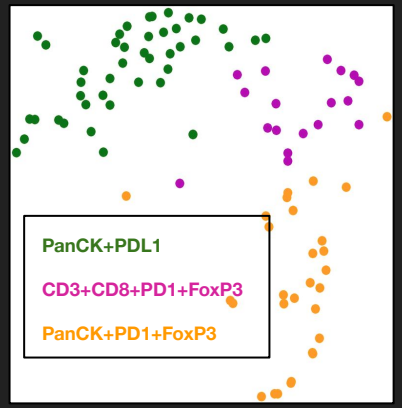



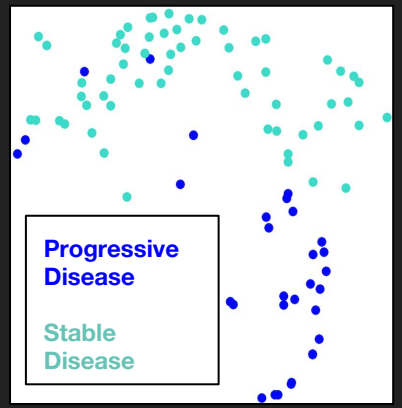



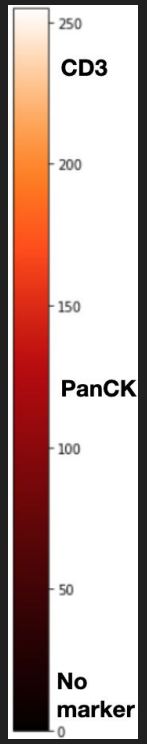



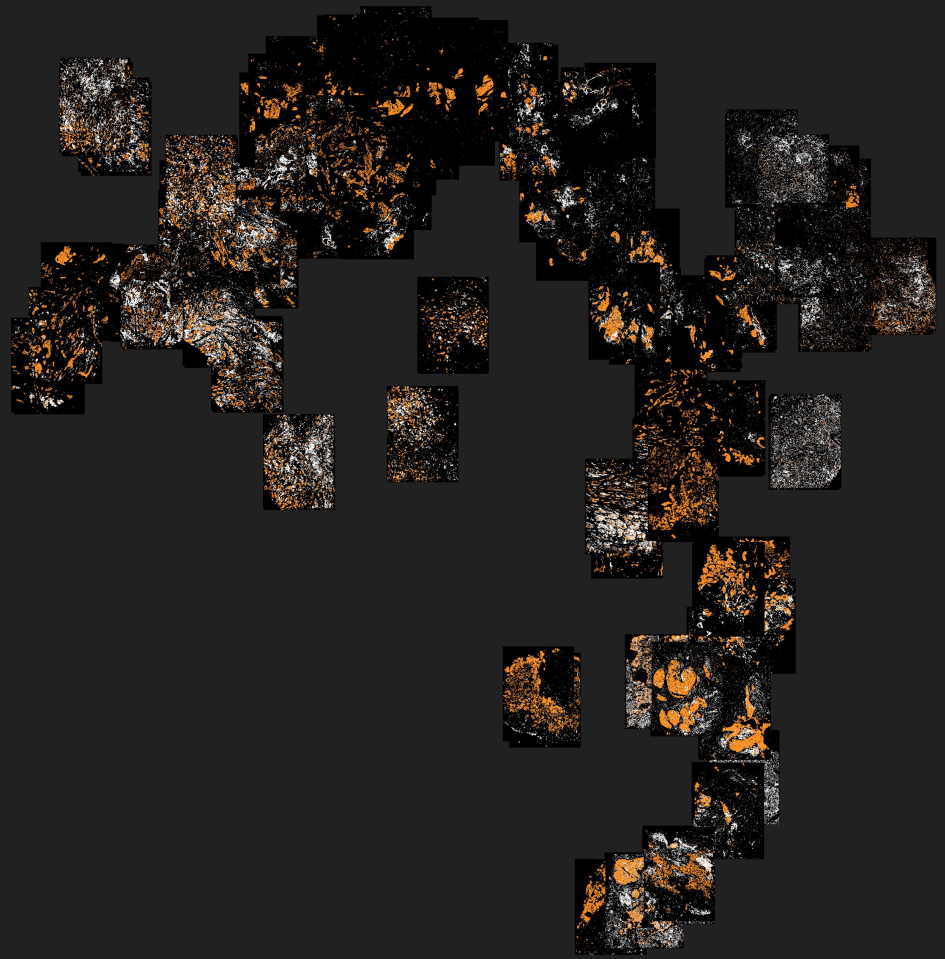

Patterns
https://doi.org/10.1016/j.patter.2022.100523

# Mistic: Visualise the spread of tumor and immune markers across multiple multiplexed images



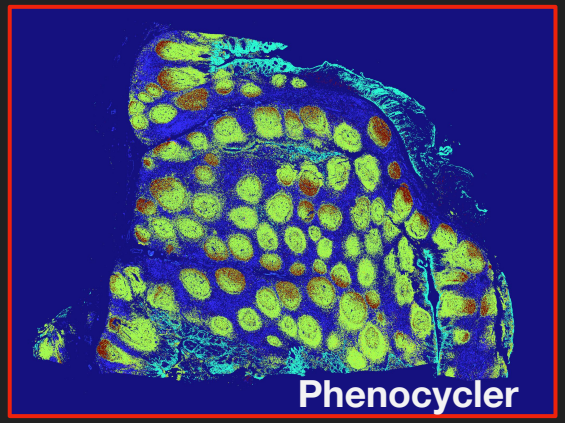



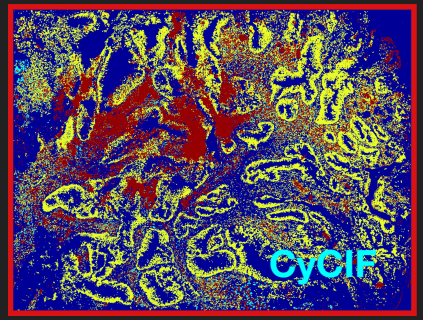
**Mistic can work with images generated from different imaging technologies**

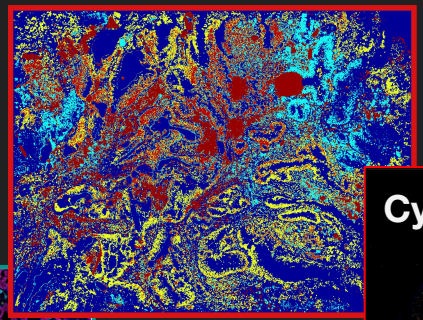

Endometrial cancer

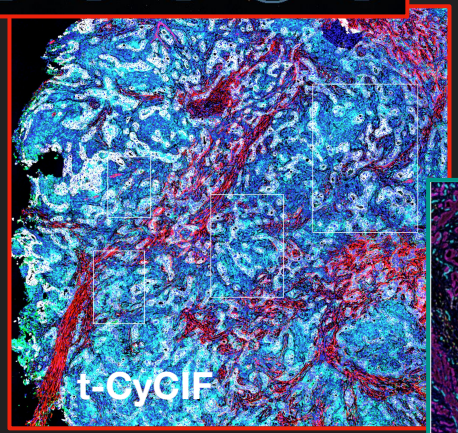
Human Tonsil




Human Colorectal carcinoma

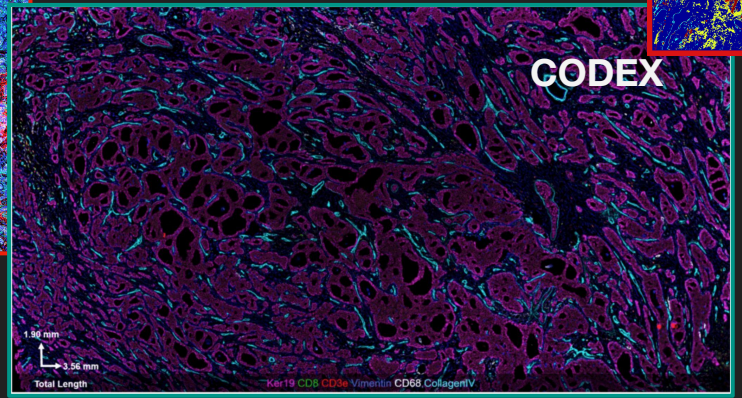


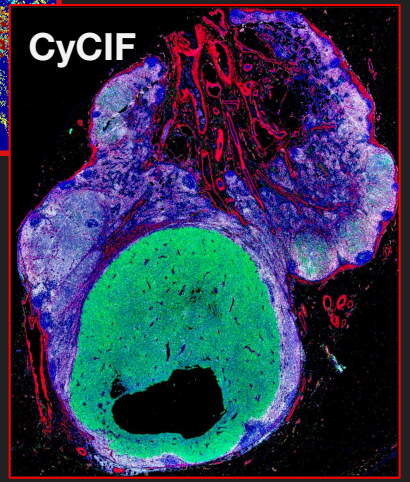



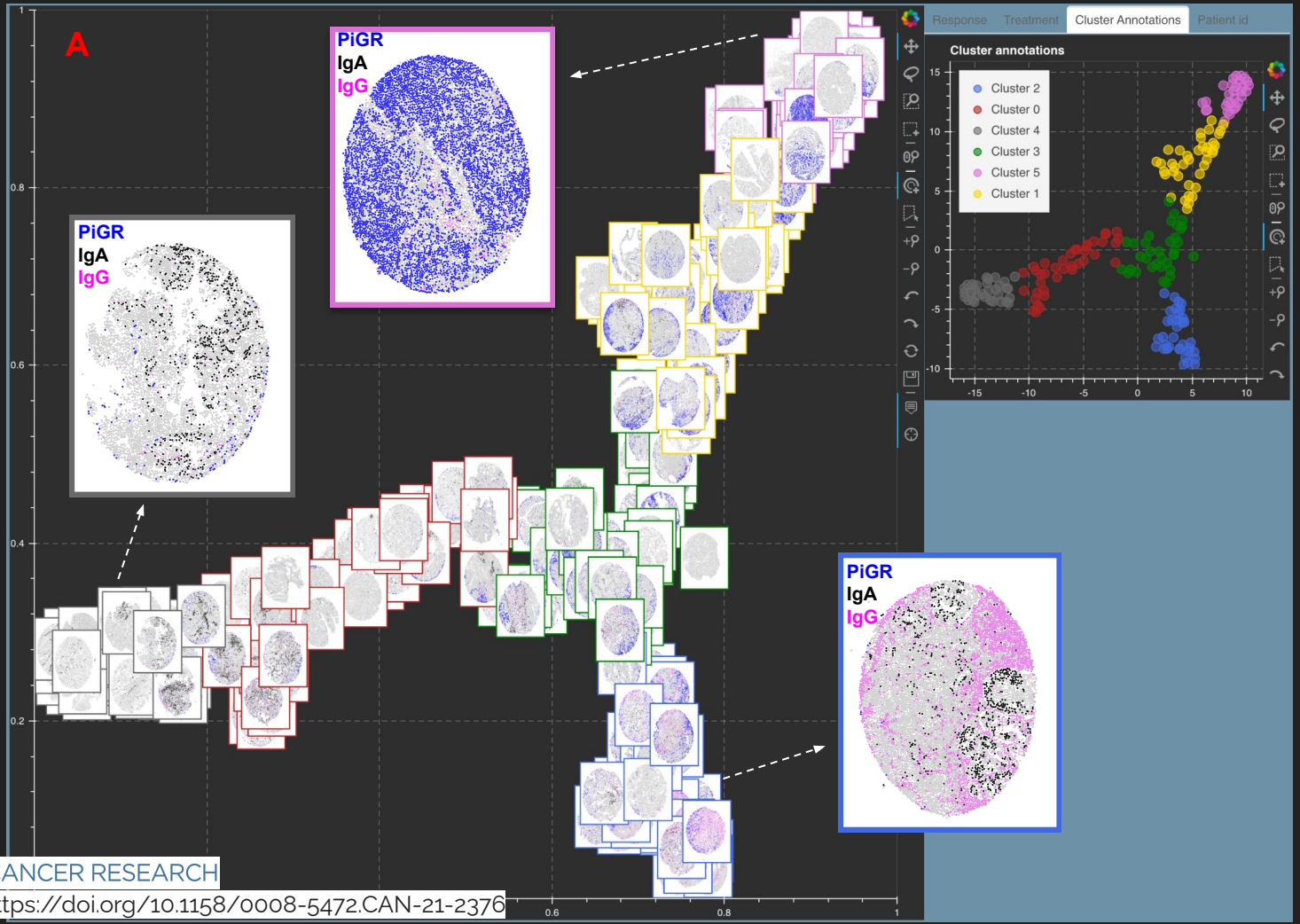



Lung adenocarcinoma metastasis to lymph

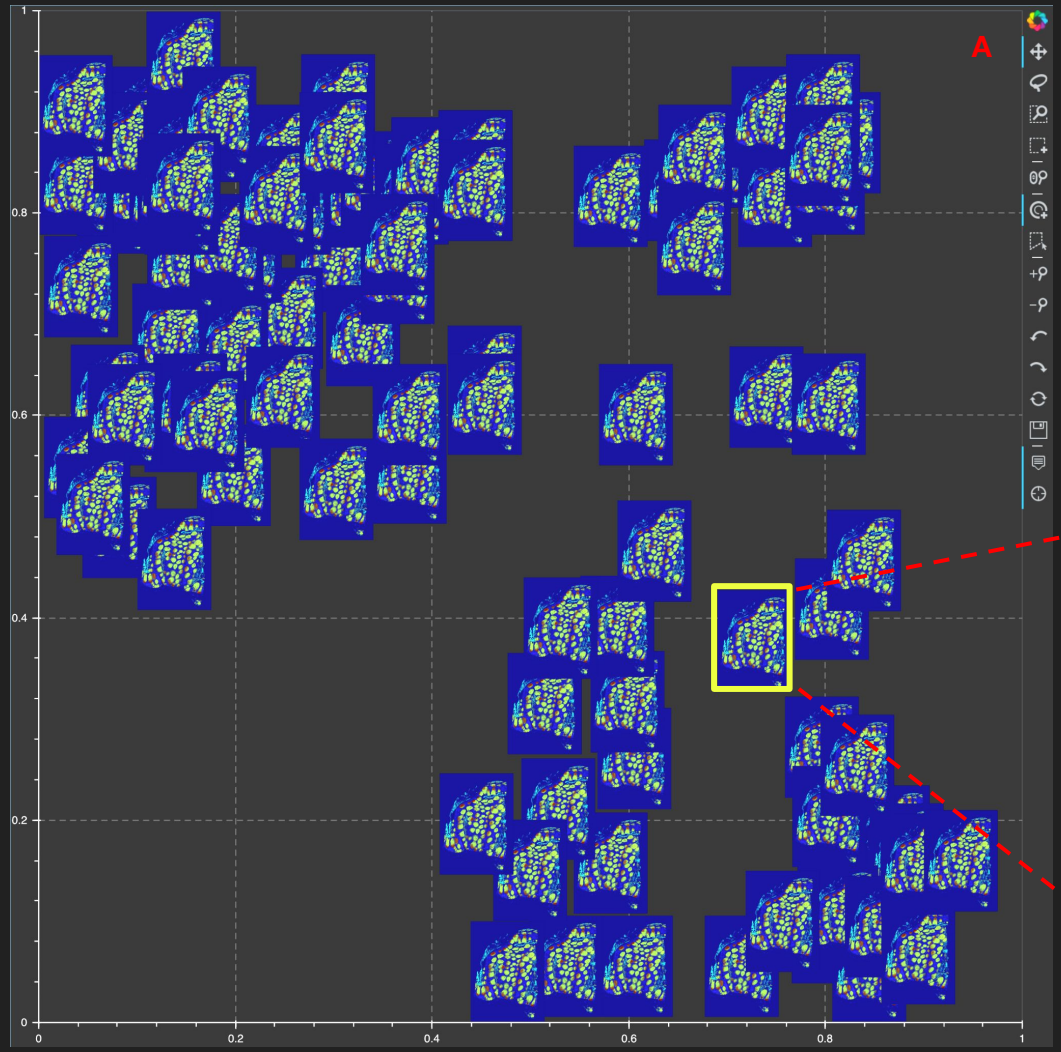
Human Breast adenocarcinoma



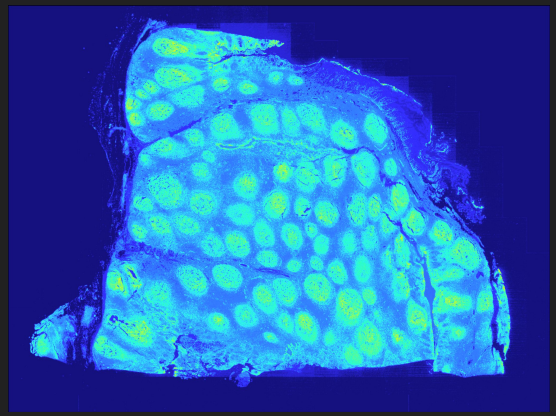





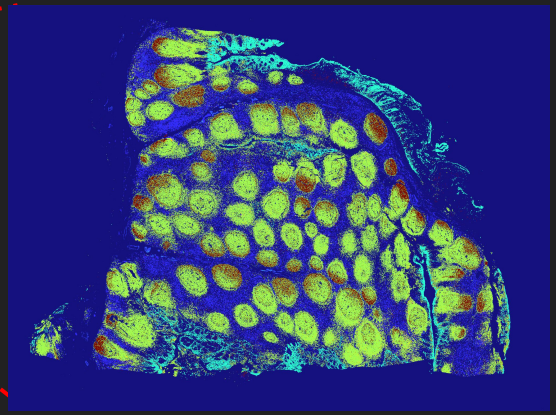

CANCER RESEARCH
https://doi.org/10.1158/0008-5472.CAN-21-2376




C

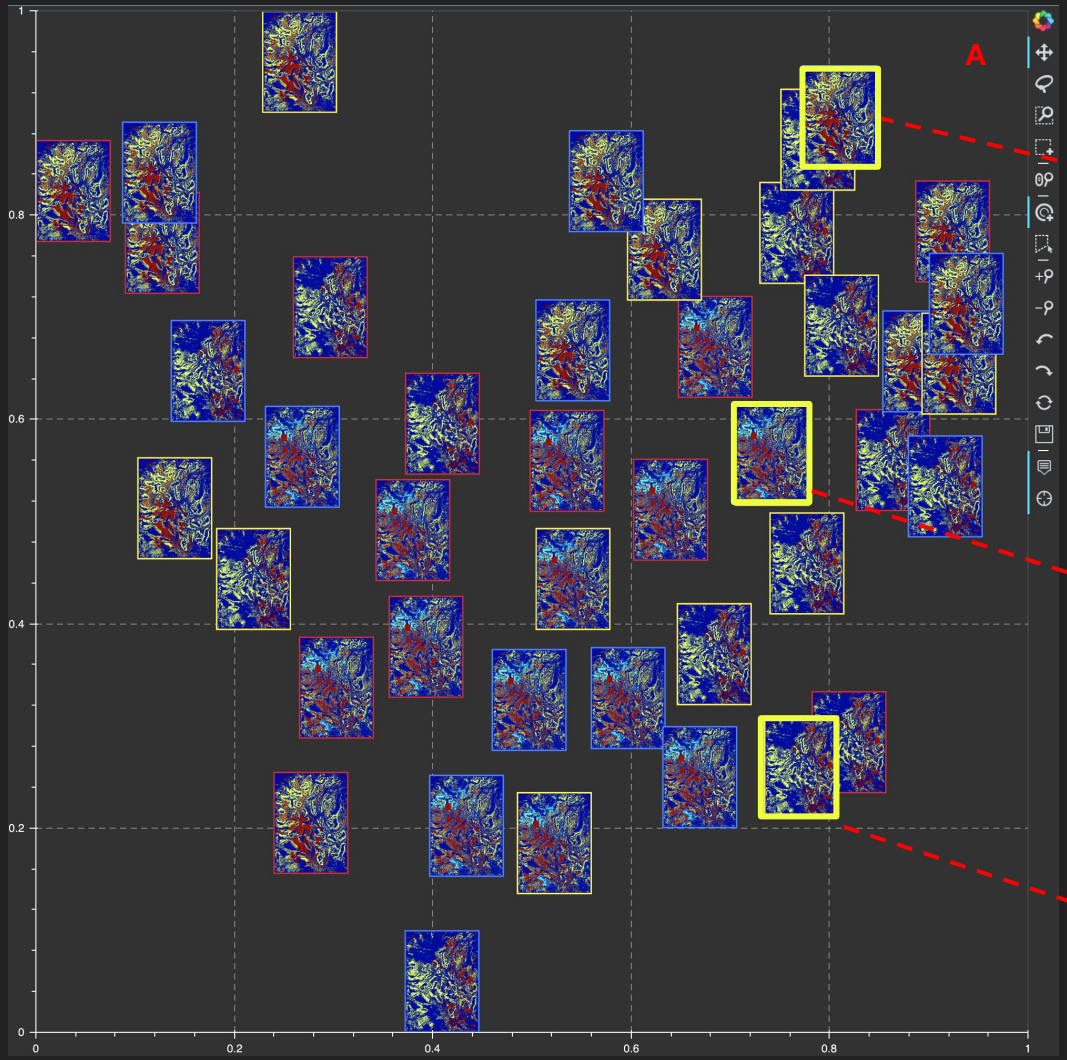


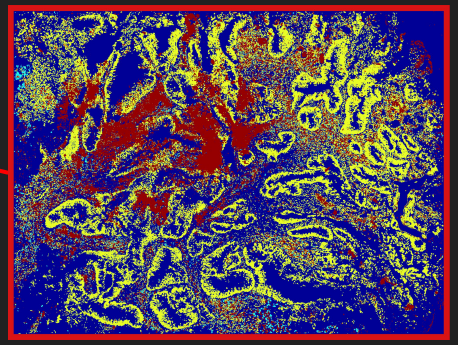
B

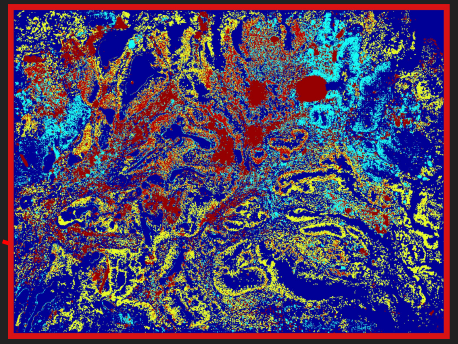


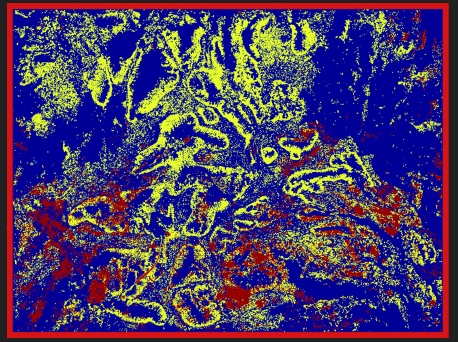
Human FFPE Tonsil Phenocycler

6 markers: PanCK, CD31, SMA, Ki67, CD8, CD20



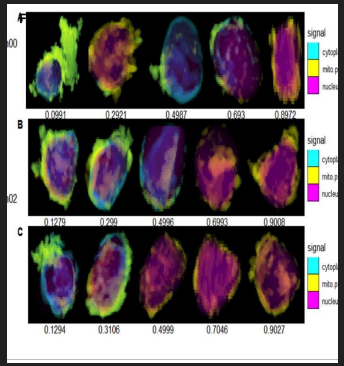



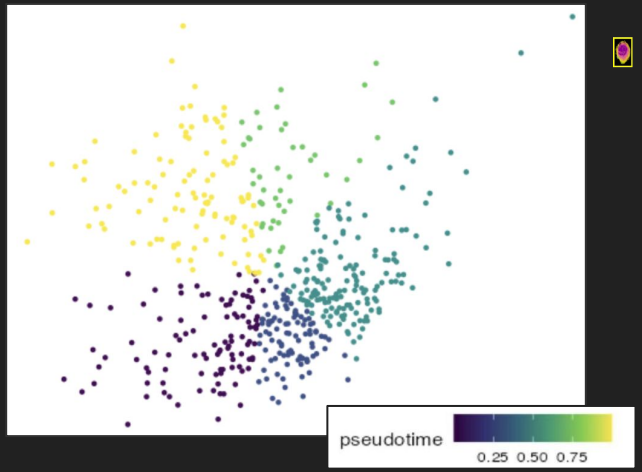



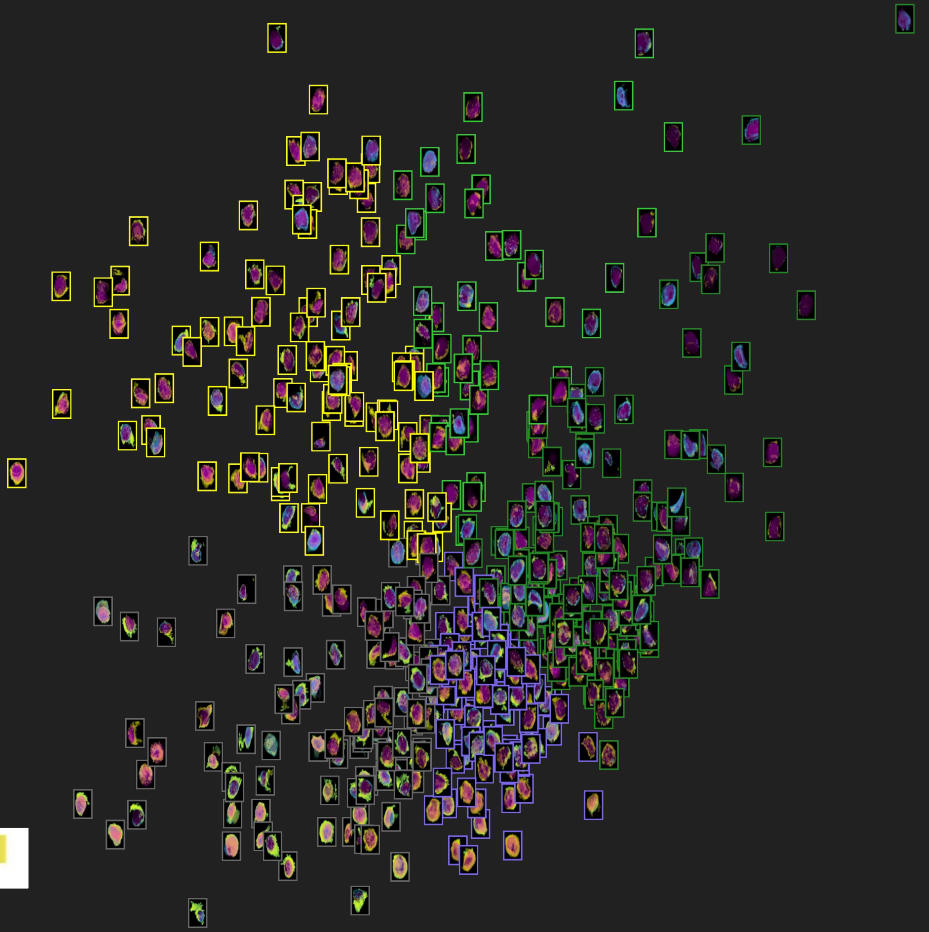

Three Human colorectal carcinoma CyCIF (OME-TIFF) images

7 markers: DAPI, CD3, CD4, CD8, CD20, CD68, FoxP3

# Integrating 693 sequenced cells and imaging derived cell cycle pseudo-times for 449 imaged cells



bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

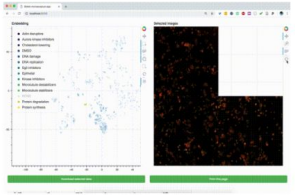
https://www.biorxiv.org/content/10.1101/2024.09.04.610488v1


## User Showcase

### Dask

Dask is a tool for scaling out PyData projects like NumPy, Pandas, Scikit-Learn, and RAPIDS. It is supported by Nvidia, Quansight, and Anaconda.

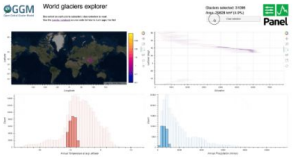
The Dask Dashboard is a diagnostic tool that helps you monitor and debug live cluster performance.



### Microscopium

Microscopium is a project maintained by researchers at Monash University.

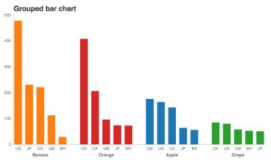
It allows researchers to discover new gene or drug functions by exploring large image datasets with Bokeh's interactive tools.




### Panel

Panel is a tool for polished data presentation that utilizes the Bokeh server. It is created and supported by Anaconda.

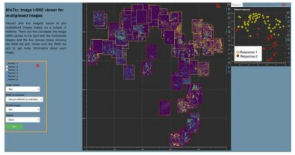
Panel makes it simple to create custom interactive web apps and dashboards by connecting user-defined widgets to plots, images, tables, or text.




### Chartify

Chartify is an opinionated high-level charting API built on top of Bokeh, created by Spotify.

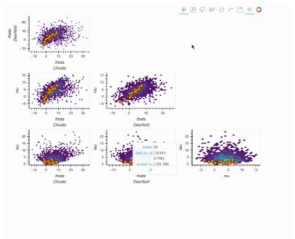
With smart default styles, consistent tidy data format, and a simple API, it's easy for you to concentrate on your work.



### Mistic

Mistic is a software package written in Python and uses the visualization library Bokeh.

Mistic can be used to simultaneously view multiple multiplexed 2D images using pre-defined coordinates (e.g. t-SNE or UMAP), randomly generated coordinates, or as vertical grids to provide an overall visual preview of the entire multiplexed image dataset.



### ArviZ

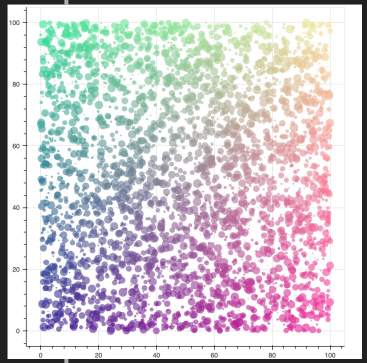
ArviZ is a community-led package for exploratory analysis of Bayesian models in Python.

It Includes functions for posterior analysis, data storage, sample diagnostics, model checking, and comparison. The goal is to provide backend-agnostic tools for diagnostics and visualizations of Bayesian inference.

Patterns
https://doi.org/10.1016/j.patter.2022.100523


```
import numpy as np

from bokeh.plotting import figure, show

N = 4000
x = np.random.random(size=N) * 100
y = np.random.random(size=N) * 100
radii = np.random.random(size=N) * 1.5
colors = np.array([(r, g, 150) for r, g in zip(50+2*x, 30+2*y)], dtype=np.uint8)

TOOLS="hover,crosshair,pan,wheel_zoom,zoom_in,zoom_out,box_zoom,undo,redo,reset,tap,save,box_select,poly_select,lasso_select,help"

p = figure(tools=TOOLS)

p.circle(x, y, radius=radii,
        fill_color=colors, fill_alpha=0.6,
        line_color=None)

show(p)
```

**<<Run Python notebook>>**

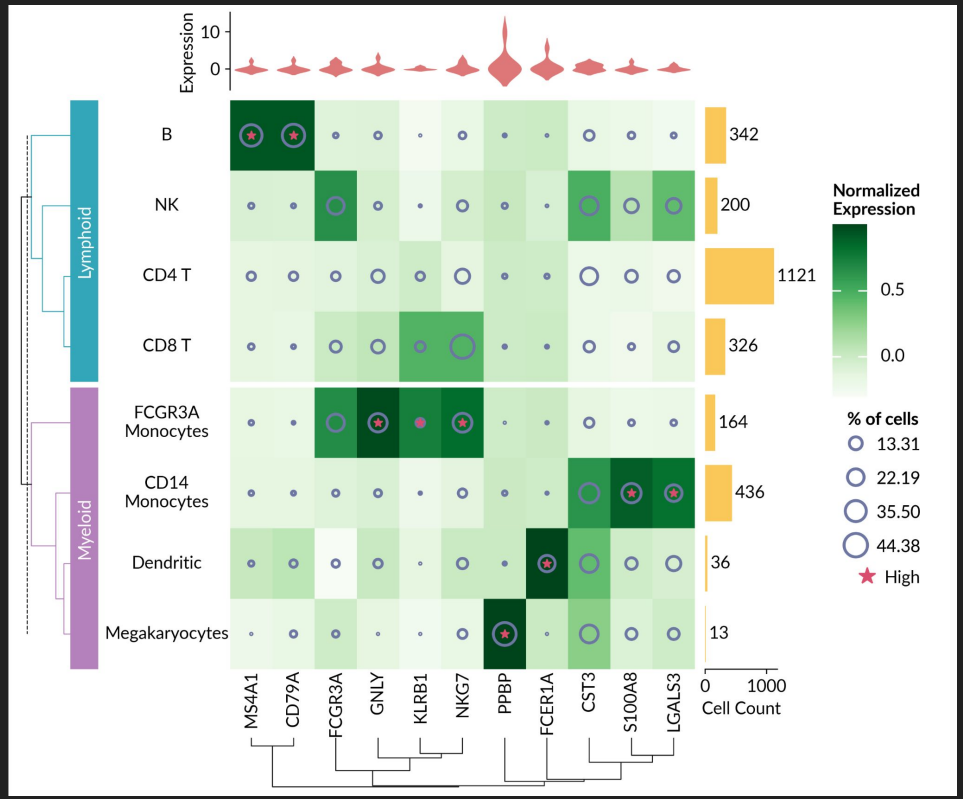


# Composable visualizations

# Composable visualizations: a visual with multiple plots

When we do visualization, we often need to combine multiple plots to show different aspects of the data.

For example, we may need to create a **heatmap** to show the expression of genes in different cells, and then create a **bar chart** to show the expression of genes in different cell types.

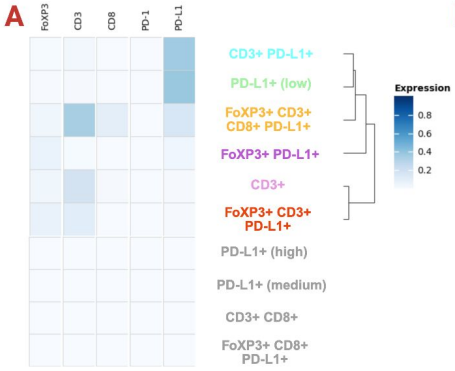



**Marsilea**, Python package to create composable visualization incrementally.

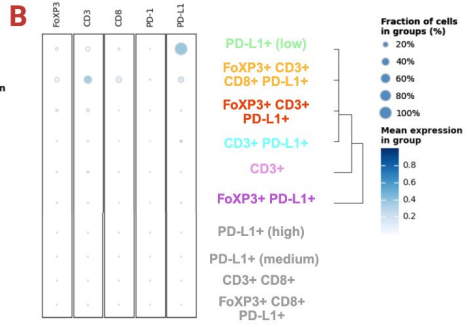
Example:
https://github.com/Marsilea-viz/marsilea?tab=readme-ov-file

## Lung2 ArmA During-Tx

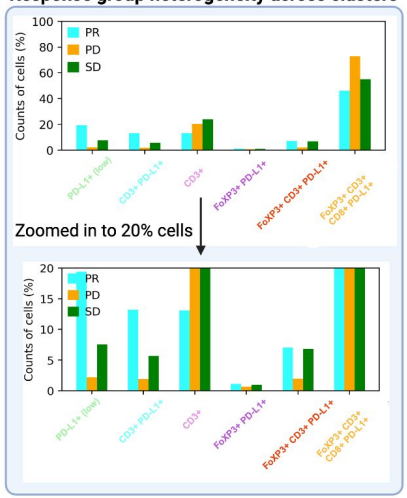
### Heatmap

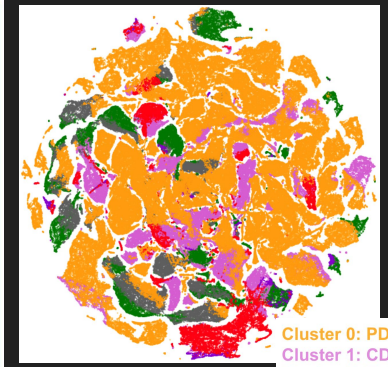



### Dotplot




### Response group heterogeneity across clusters

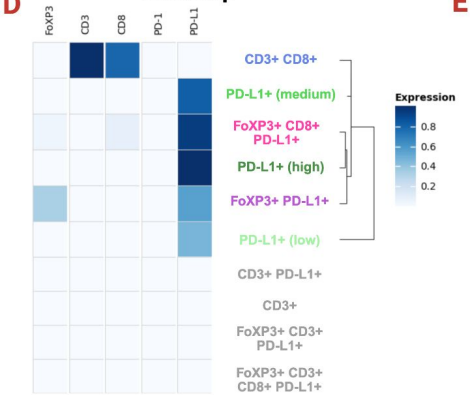


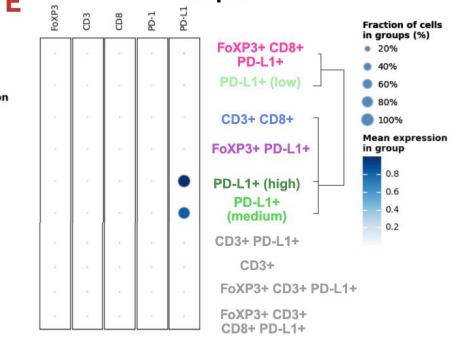
Cluster 0: PD-L1+
Cluster 1: CD3+ PD-L1+
Cluster 2: FoxP3+ CD3+
Cluster 3: FoxP3+ PD-L1+
Cluster 4: FoxP3+ CD3+ PD-L1
Cluster 5: FoxP3+ CD3+ CD8+ PD-L1

## Lung2 ArmB During-Tx

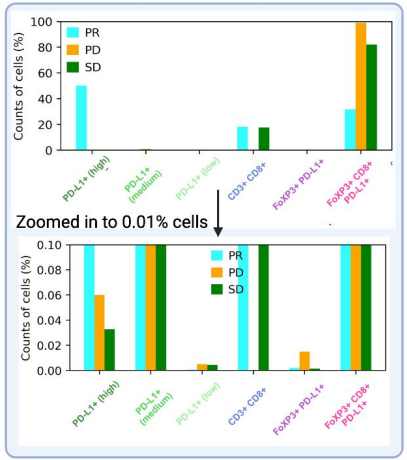
### Heatmap

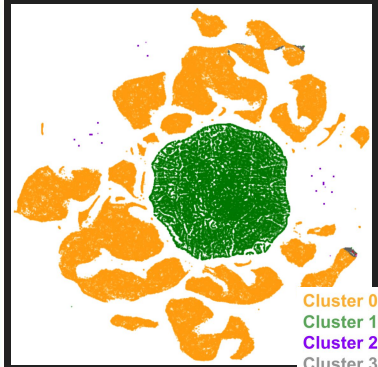



### Dotplot




### Response group heterogeneity across clusters



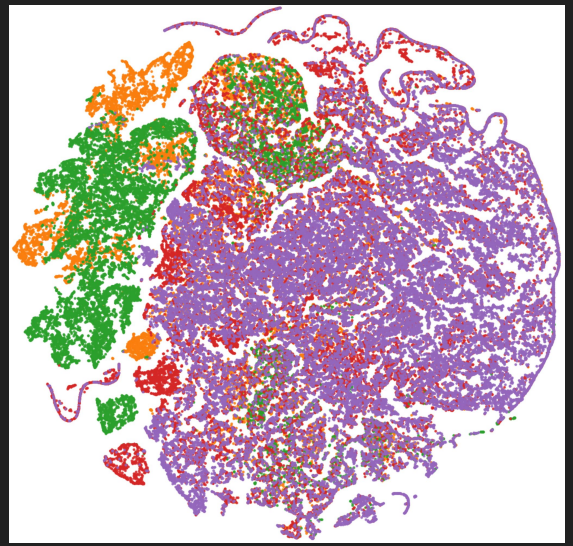
Cluster 0: PD-L1+ high
Cluster 1: PD-L1+ medium
Cluster 2: PD-L1+ low
Cluster 3: CD3+ CD8+
Cluster 4: FoxP3+ PD-L1+
Cluster 5: FoxP3+ CD8+ PD-L1+

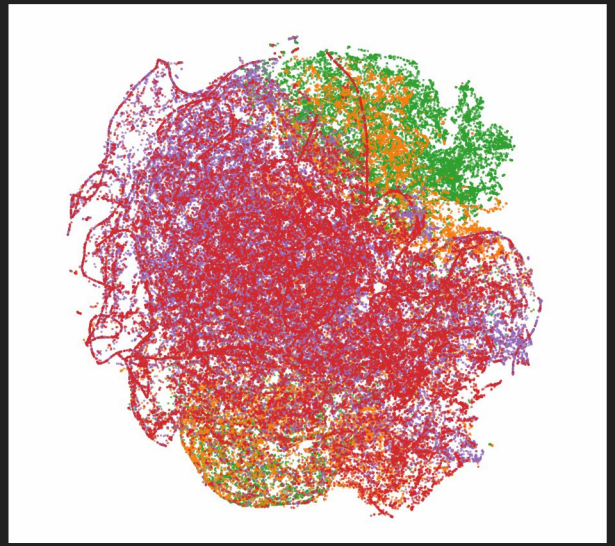

**Observable:**

https://observablehq.com/@vizgen/mouse-brain-transcripts-of-interest-view

# 3D visualizations

# Cell segmentation approach shows preserved cellular neighborhoods across PD and SD patients





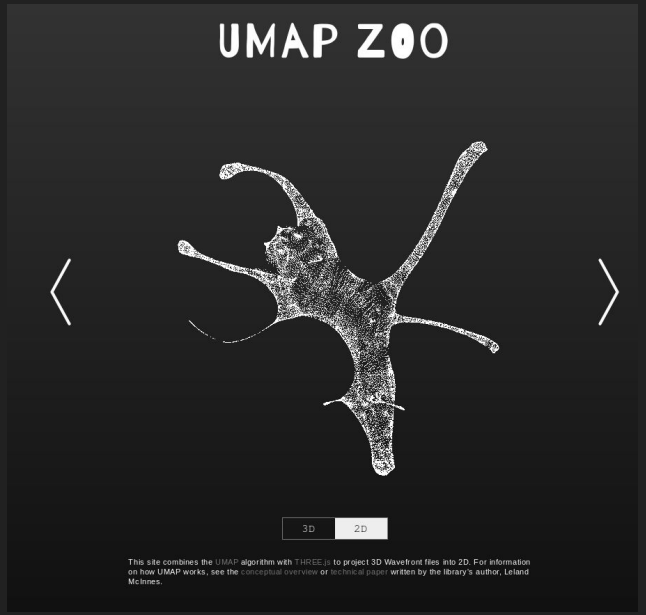
PD pre (16 FoVs)
PD on (15 FoVs)
SD pre (30 FoVs)
SD on (28 FoVs)

9 patients -
4 PD, 5 SD
121,000 cells

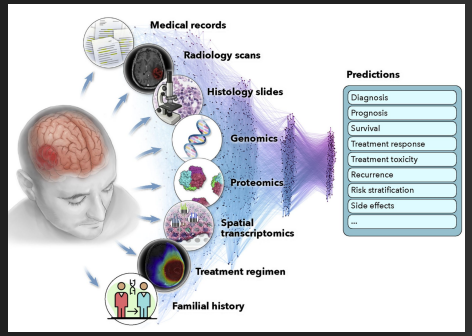
bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

https://www.biorxiv.org/content/10.1101/2022.10.22.513219v2

- UMAP Zoo
- Tensorflow Embedding Projector
- PixPlot
- UMAP Explorer
- Audio Explorer
- Orion Search
- Exploring Fashion MNIST
- ESM Metagenomic Atlas




https://umap-learn.readthedocs.io/en/latest/interactive_viz.html




# Current bottlenecks in visualization

- Handling large and complex datasets
  - Multimodal datasets
  - Various file formats
- Ensuring data quality and integrity
- Selecting the right visualization tools
  - Art of tailoring the tools to data and questions
- Effectively conveying the intended message.

- Dealing with real-time data and addressing privacy concerns pose significant challenges
- Integrate analysis-agnostic visualizations into computational applications

- Human factors:
  - Provide both software libraries and end-user facing visualizations
  - Share browser-based interactive visualizations
  - Ignorance, laziness

The importance of data visualization

Enables descriptive learning

PCA, t-SNE, UMAP

Many software packages that aid in generating these visualizations

- **Not all visualization methods are created equal.**
  - A bad vis can lead to erroneous interpretations
  - A good vis can lead to insights otherwise missed
- **Not all visualization methods have been invented yet.**
  - The space of possible vis methods is astronomical
- **Visualization choice should match the purpose**
  - Vis can be used to consume, produce, or enjoy

Visualization → Interactive → Composable → 3D → Future

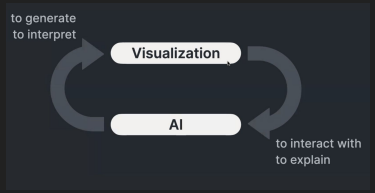
Detailed handshakes with the data

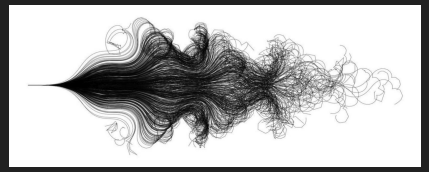
Mistic

Observable

3D versions of 2d reduced data

Good check





Thank you IMO for your attention!