Interactive Exploration of Large-Scale Datasets with Jupyter-Scatter

Fritz Lekschas

@flekschas
lekschas.de

July 13, 2023

SciPy '23

WORK

Head of Visualization Research at Ozette

EDUCATION

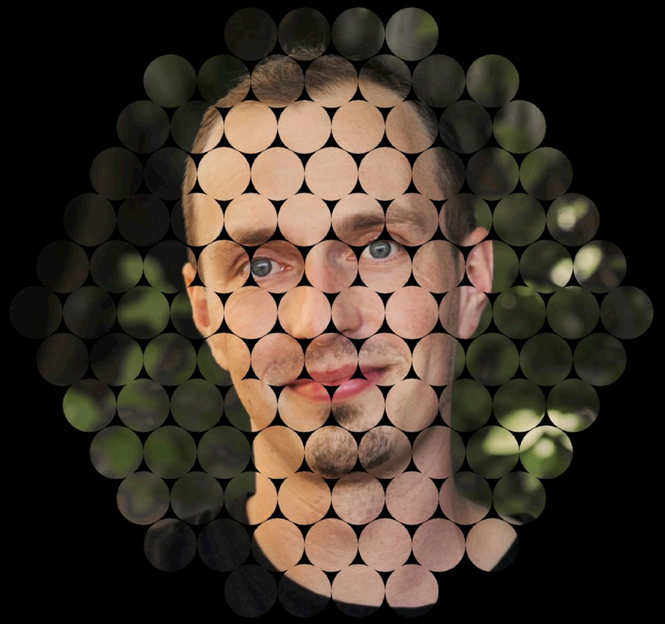
PhD '21 in CS from Harvard University

RESEARCH

Visualization | Human-Centered ML | Design

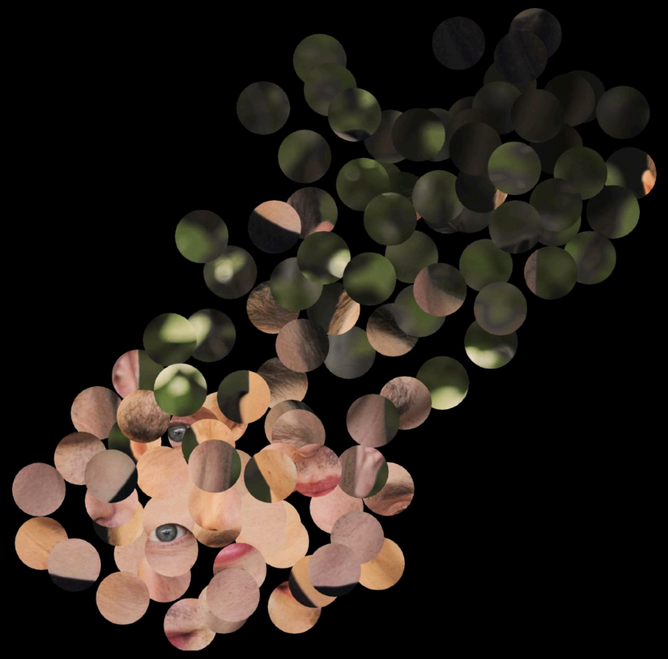
PASSION

Embeddings & Scatter plots!



PASSION

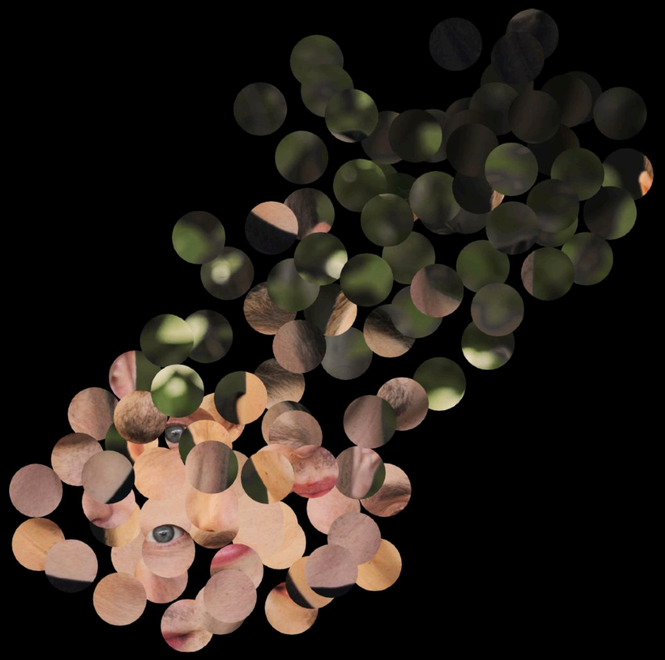
Embeddings & Scatter plots!

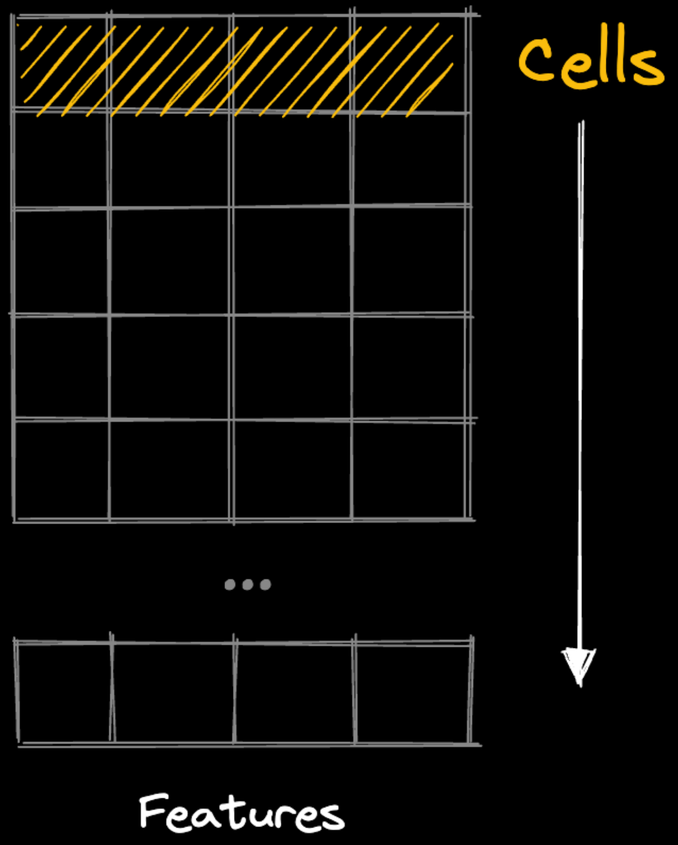


PASSION

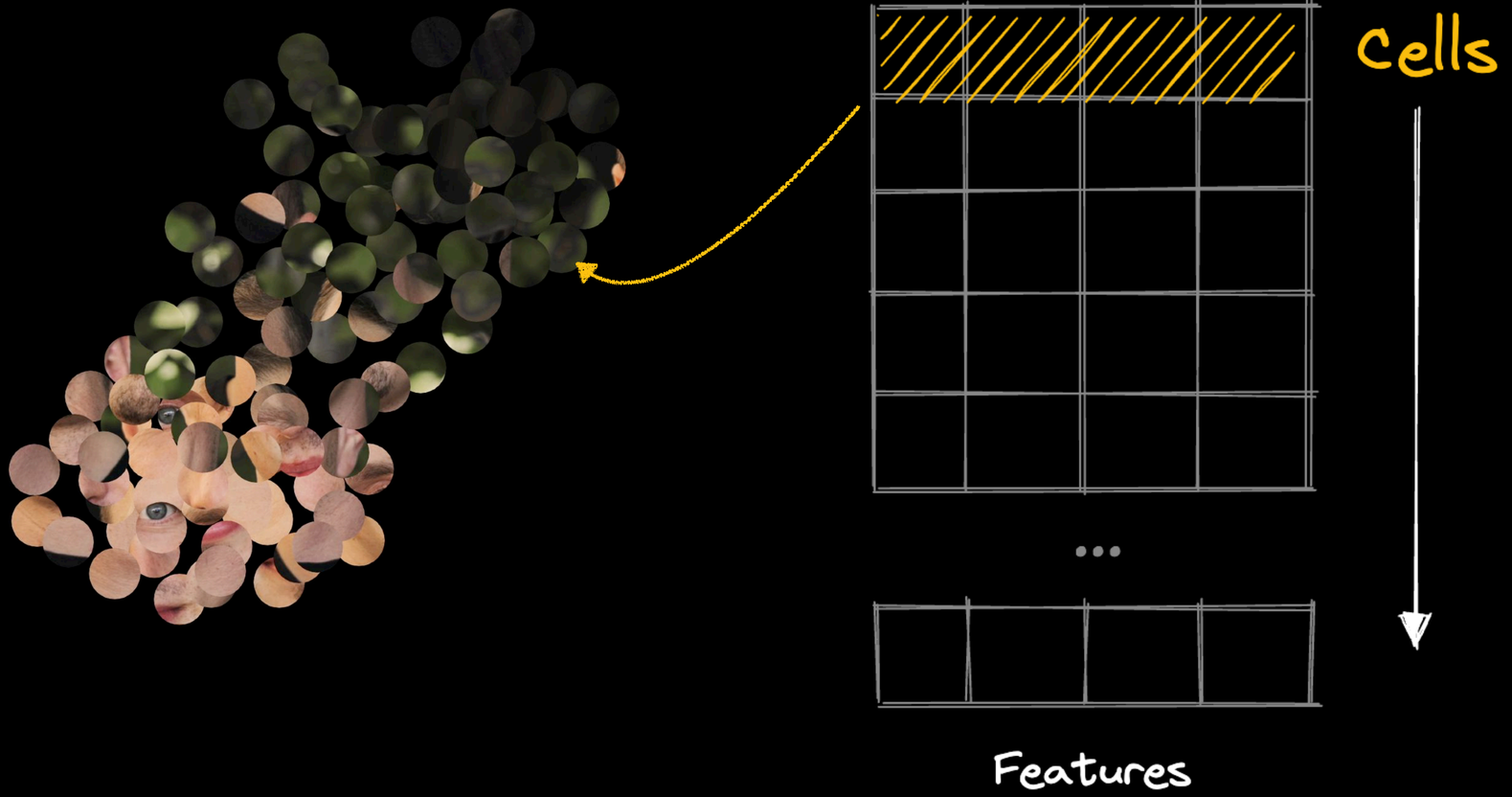
Embeddings & Scatter plots!

Single-Cell

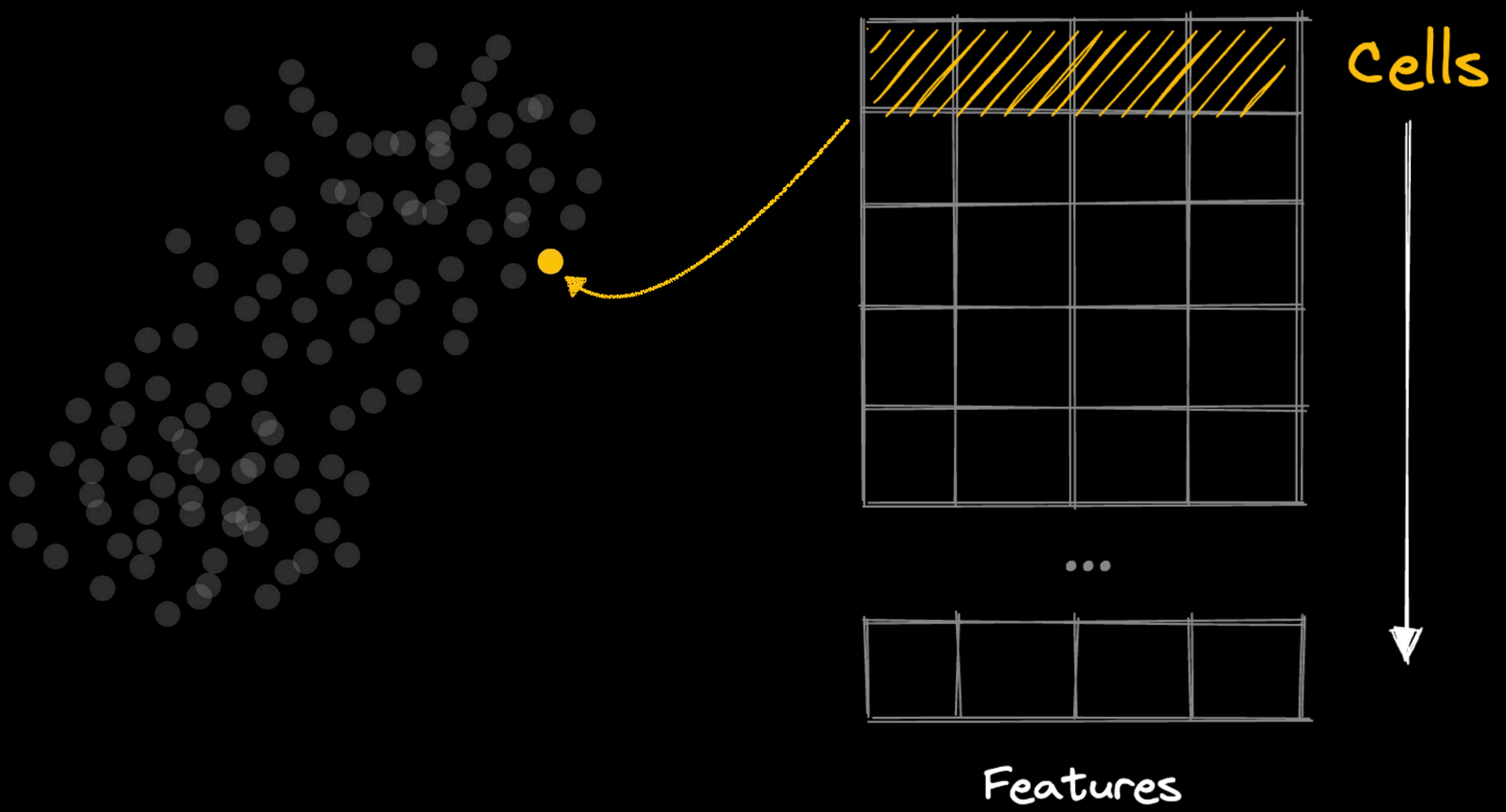


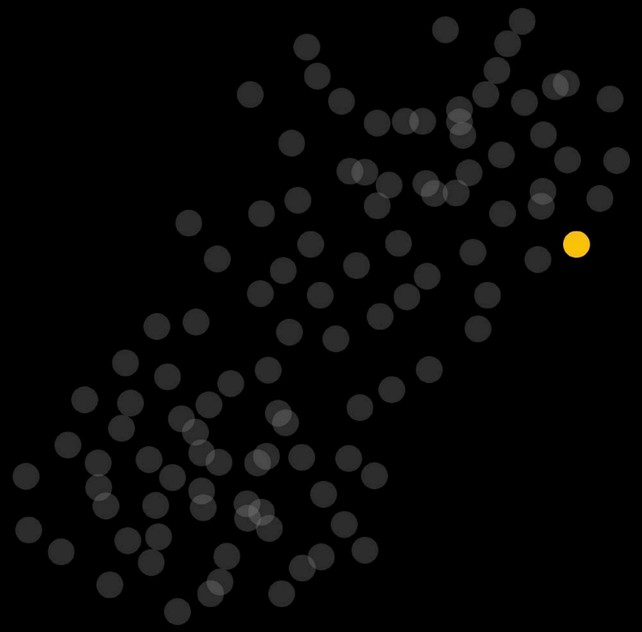


Single-Cell



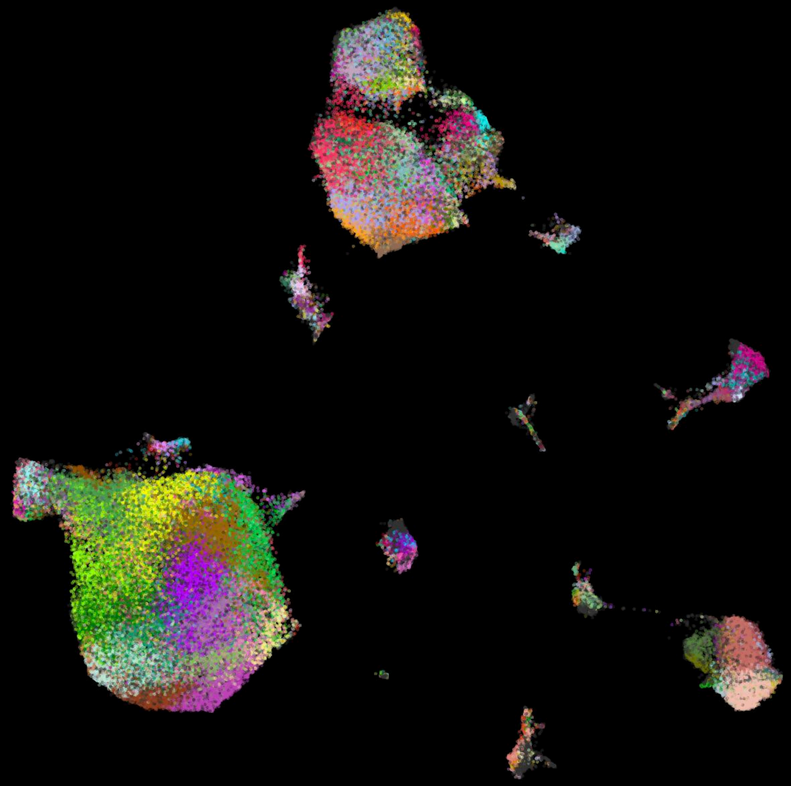
Single-Cell





USEFUL FOR

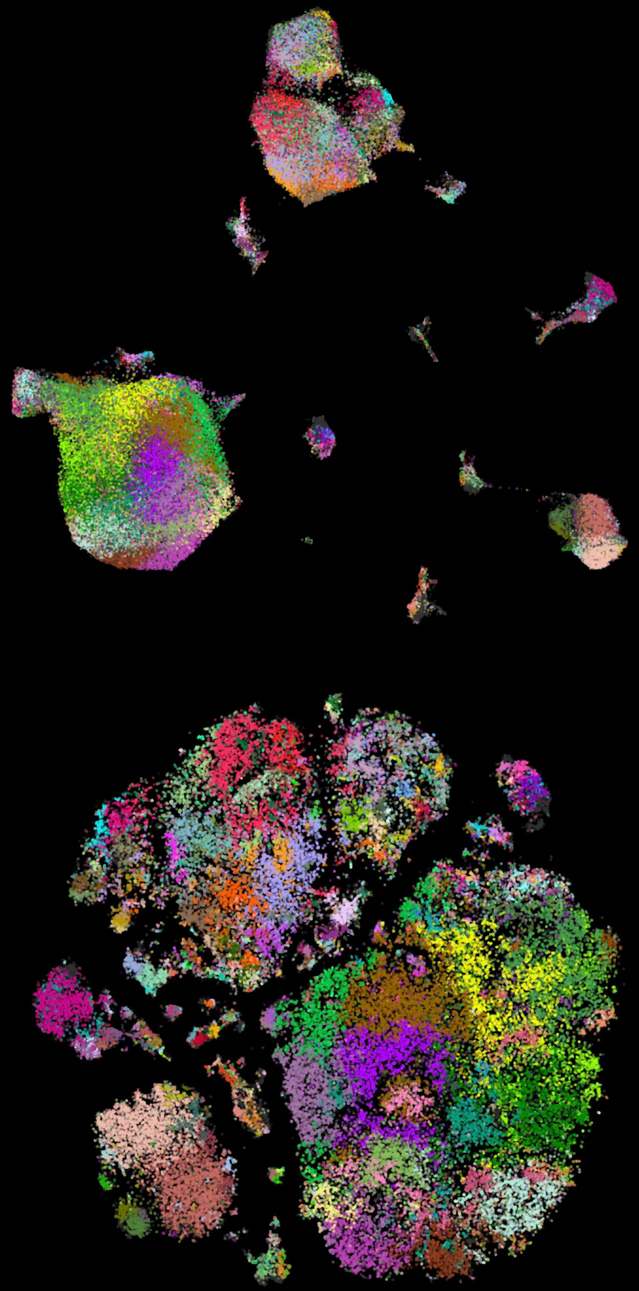
Overview of Data
Explore & Compare Clusters



USEFUL FOR

Overview of Data
Explore & Compare Clusters

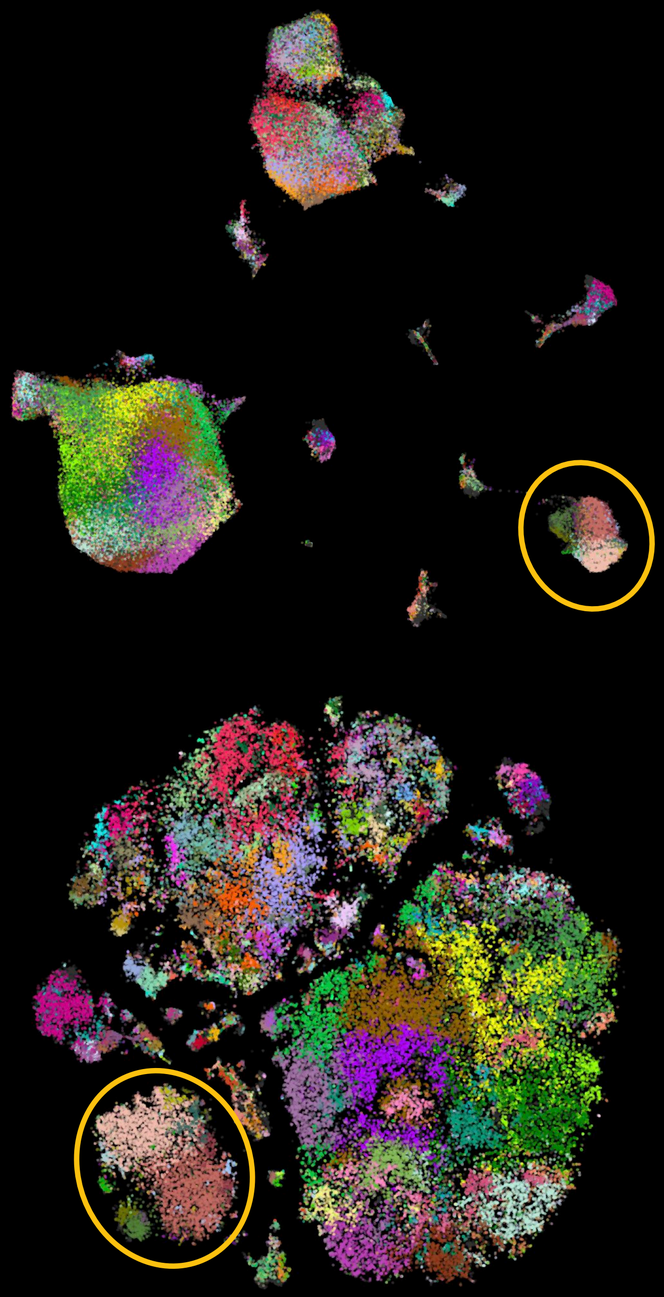
Data from Mair et al., 2022. *Nature*.



USEFUL FOR

Overview of Data
Explore & Compare Clusters

Data from Mair et al., 2022. *Nature*.



USEFUL FOR

Overview of Data
Explore & Compare Clusters

Data from Mair et al., 2022. *Nature*.

Jupyter Scatter

A widget for interactive exploration of large-scale scatter plots.

pip install jupyter-scatter

github.com/flekschas/jupyter-scatter

GOALS

1. Scale to millions of points
2. Support interactive pan+zoom and selections
3. Offer perceptually-effective defaults
4. Allow linking multiple scatter plots
5. Expose via an easy-to-use API

GOALS

1. Scale to millions of points
2. Support interactive pan+zoom and selections
3. Offer perceptually-effective defaults
4. Allow linking multiple scatter plots
5. Expose via an easy-to-use API

Biology

- Single-Cell
- Genomics

Natural Language Processing

- Word Embeddings
- Document Embeddings

Computer Vision

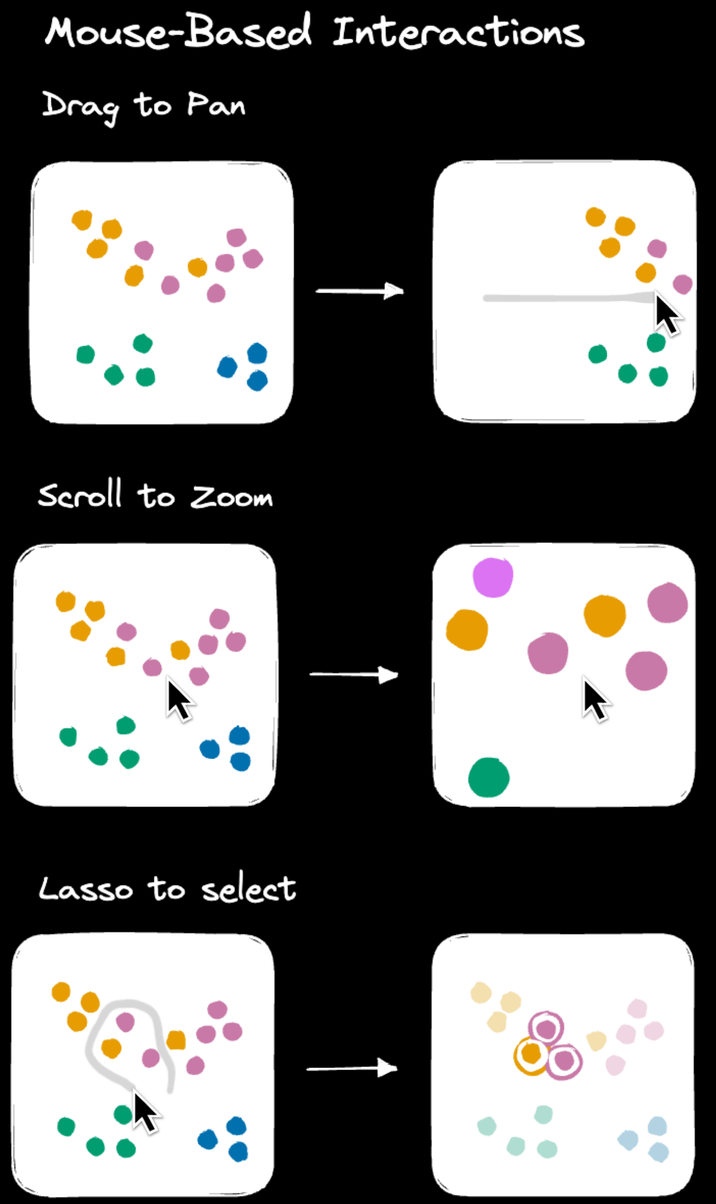
- Image Embeddings
- Generative AI

Geospatial Data

...

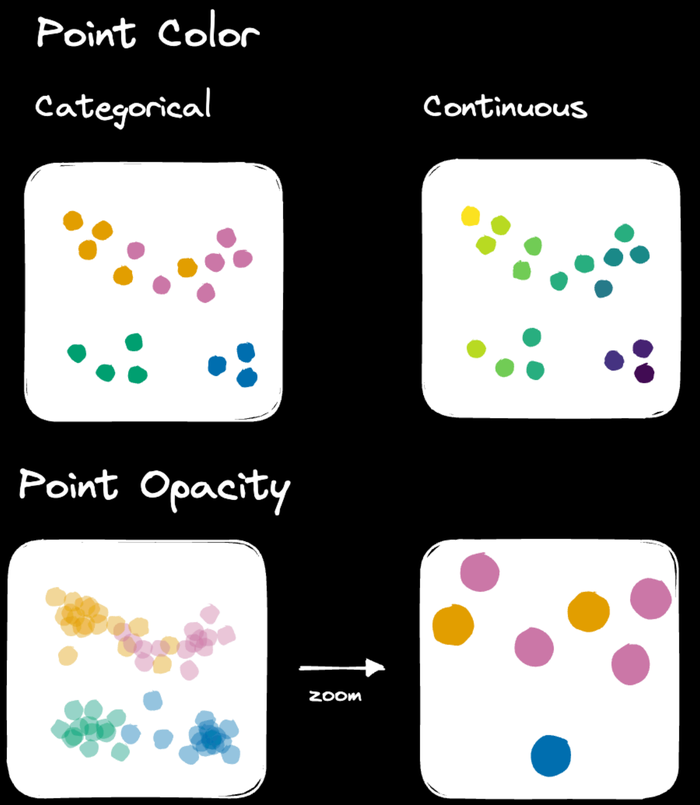
GOALS

1. Scale to millions of points
2. **Support interactive pan+zoom and selections**
3. Offer perceptually-effective defaults
4. Allow linking multiple scatter plots
5. Expose via an easy-to-use API



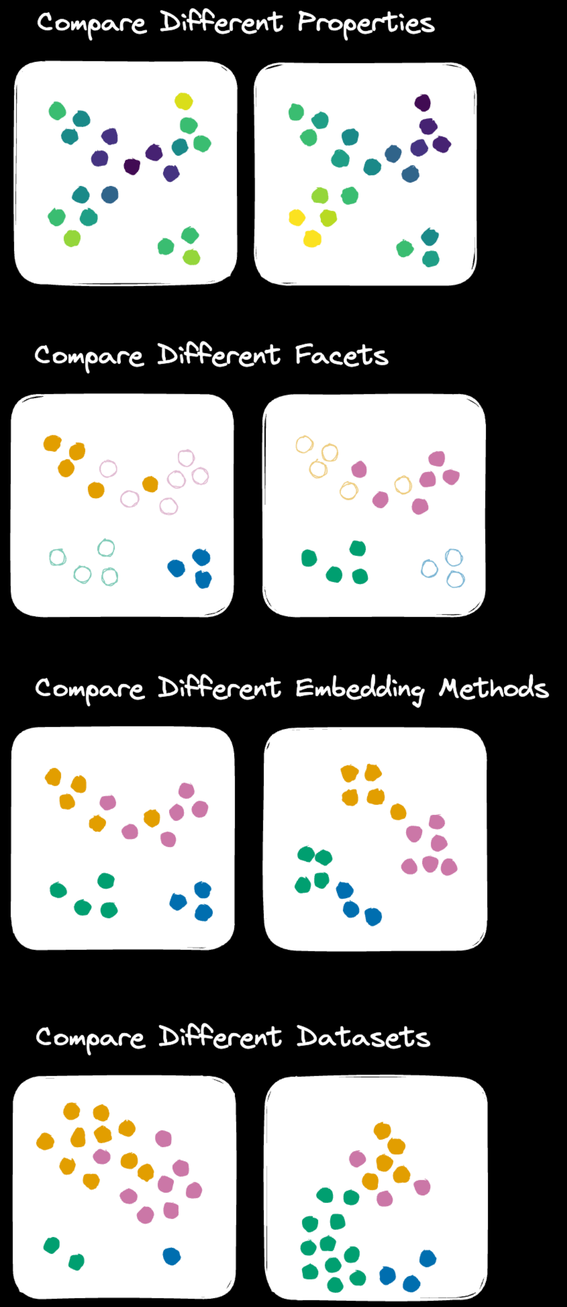
GOALS

1. Scale to millions of points
2. Support interactive pan+zoom and selections
3. **Offer perceptually-effective defaults**
4. Allow linking multiple scatter plots
5. Expose via an easy-to-use API



GOALS

1. Scale to millions of points
2. Support interactive pan+zoom and selections
3. Offer perceptually-effective defaults
4. **Allow linking multiple scatter plots**
5. Expose via an easy-to-use API



GOALS

1. Scale to millions of points
2. Support interactive pan+zoom and selections
3. Offer perceptually-effective defaults
4. Allow linking of multiple scatter plots
5. **Expose via an easy-to-use API**

Integrate with Pandas DataFrame

```
Scatter(
    data=df,
    x="column_a"
    y="column_b"
)
```

Readable API

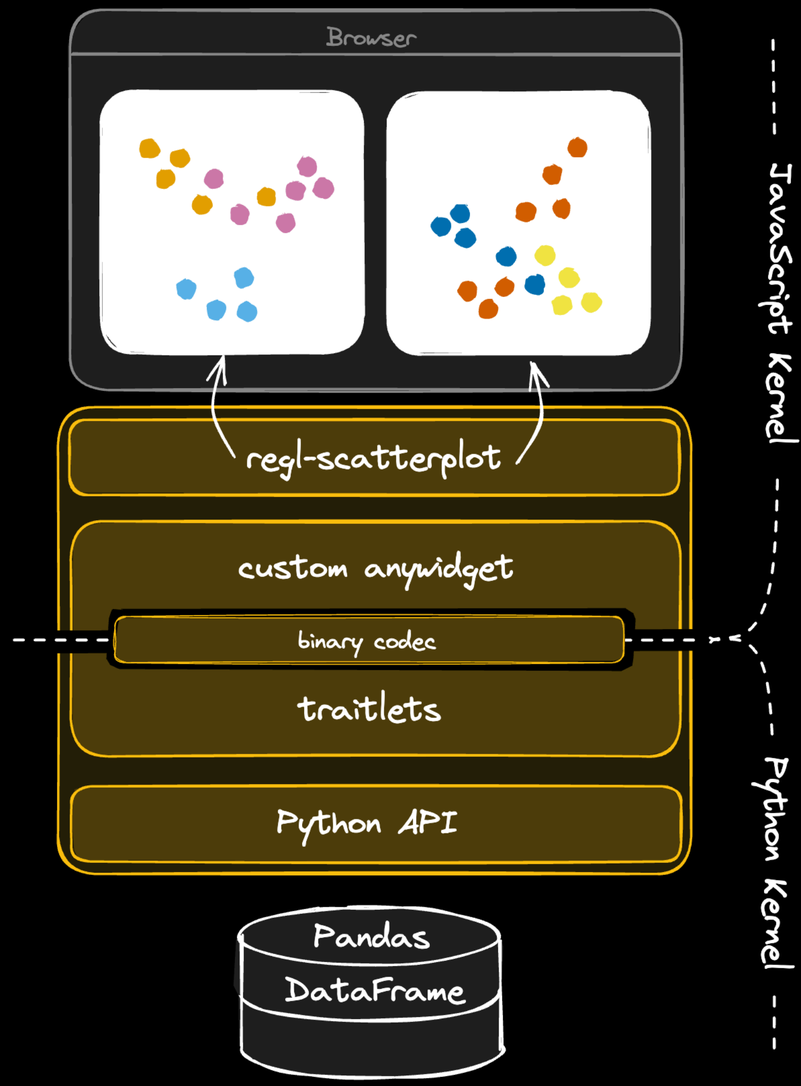
```
scatter.color(by="column_c")
scatter.opacity(0.5)
scatter.size(12)
scatter.height(320)
scatter.selection([1, 2, 3])
```

Live Demos!

https://github.com/flekschas/jupyter-scatter-tutorial

ARCHITECTURE

1. WebGL Rendering via `regl-scatterplot`[1] for fast plotting
2. Python API layer for integrating with Pandas and configuring `regl-scatterplot`[1]
3. Ipywidgets for communication with Jupyter via `anywidget`[2]



1) https://github.com/flekschas/regl-scatterplot/
2) https://github.com/manzt/anywidget/

🚨 MASSIVE SHOUT OUTS!

Trevor Manz for the codec design, anywidget integration, & tutorial setup

Nezar Abdennur for feedback on the API design

Ricky Reusser for his inspirational work on selecting the right point opacity

Rye Terrell for his beautiful multi-instance WebGL rendering approach

Thanks!

- pip install jupyter-scatter
- github.com/flekschas/jupyter-scatter
- github.com/flekschas/jupyter-scatter-tutorial

@flekschas
lekschas.de