

Dimensionality Reduction

Principal Component Analysis (PCA)

t-Stochastic Neighbor Embedding (t-SNE)

Uniform Manifold Approximation and Projection (UMAP)



Quick review

COSMIC DAWN CENTER

DAWN



Wednesday May 3, 2023

Quick review

We've learned several useful methods already.
What sorts of things are we now good at?



Quick review

We've learned several useful methods already.
What sorts of things are we now good at?

Can I always use these? If not, what are the requirements in order to run these methods?



Quick review

We've learned several useful methods already.
What sorts of things are we now good at?

Can I always use these? If not, what are the requirements in order to run these methods?

What should we do if we don't have labels?

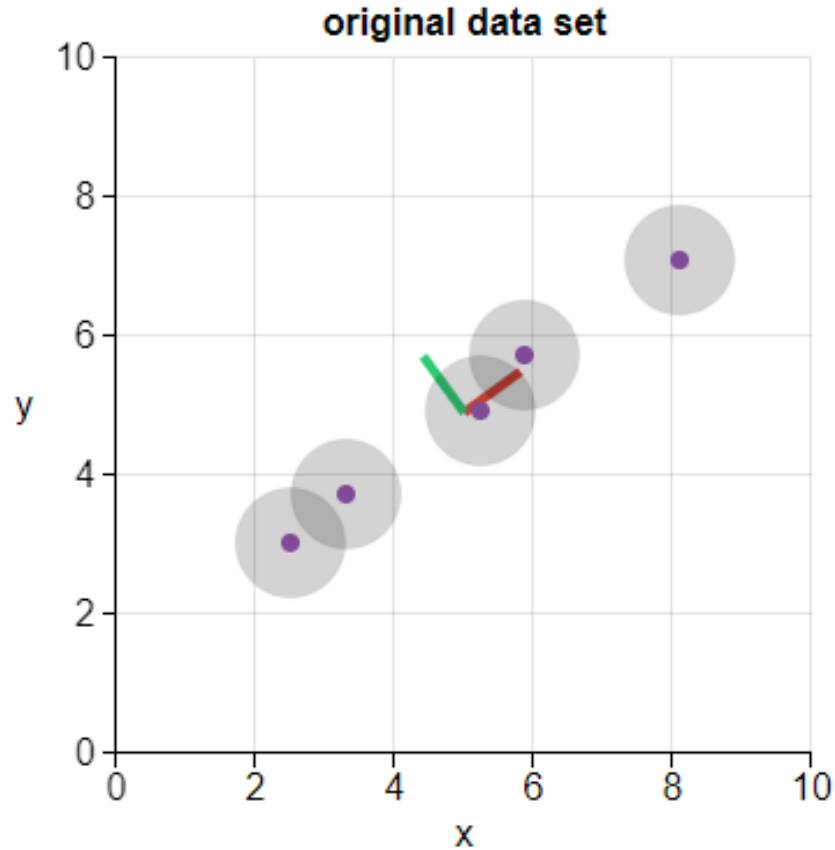


What should we
do if we don't
have labels?



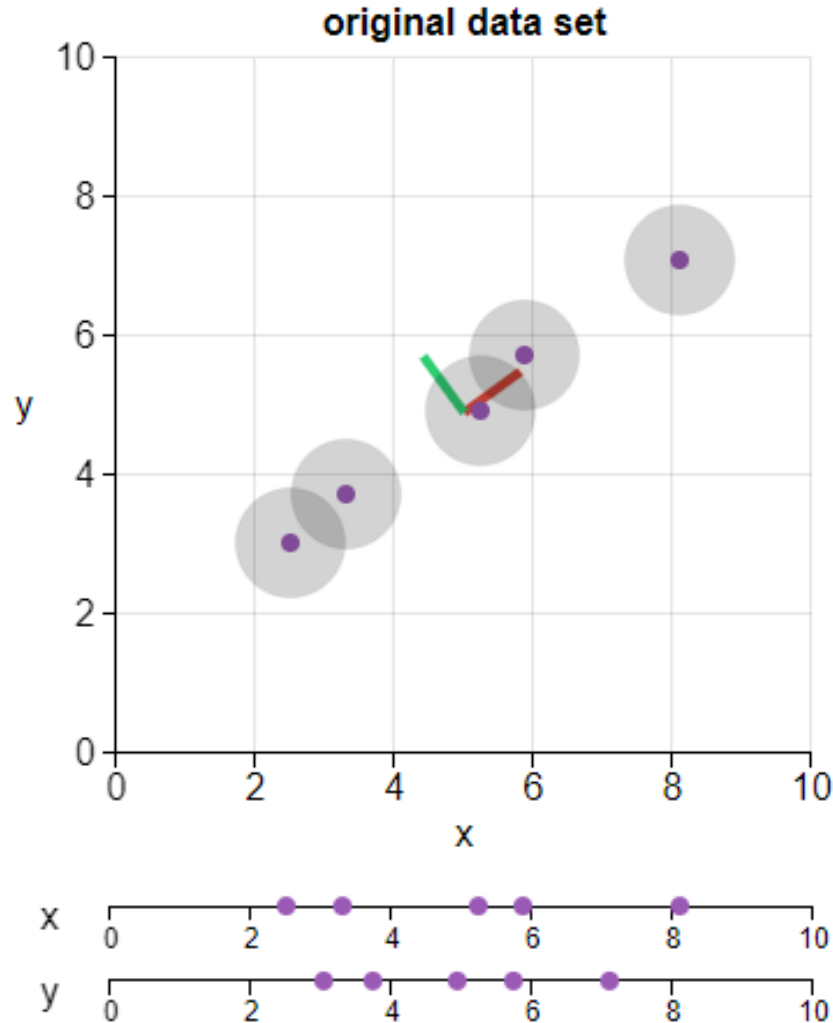
Principal Component Analysis (PCA)

<https://setosa.io/ev/principal-component-analysis/>



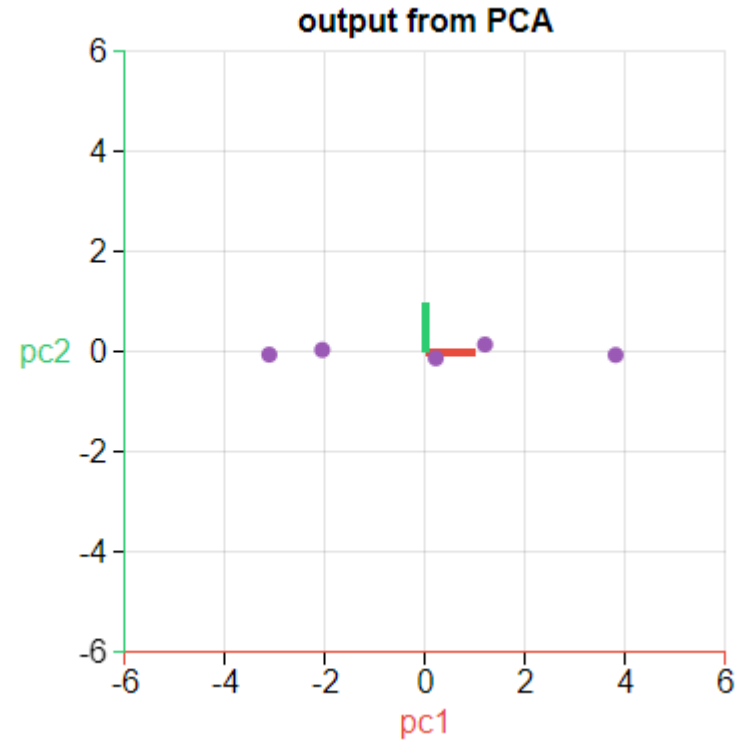
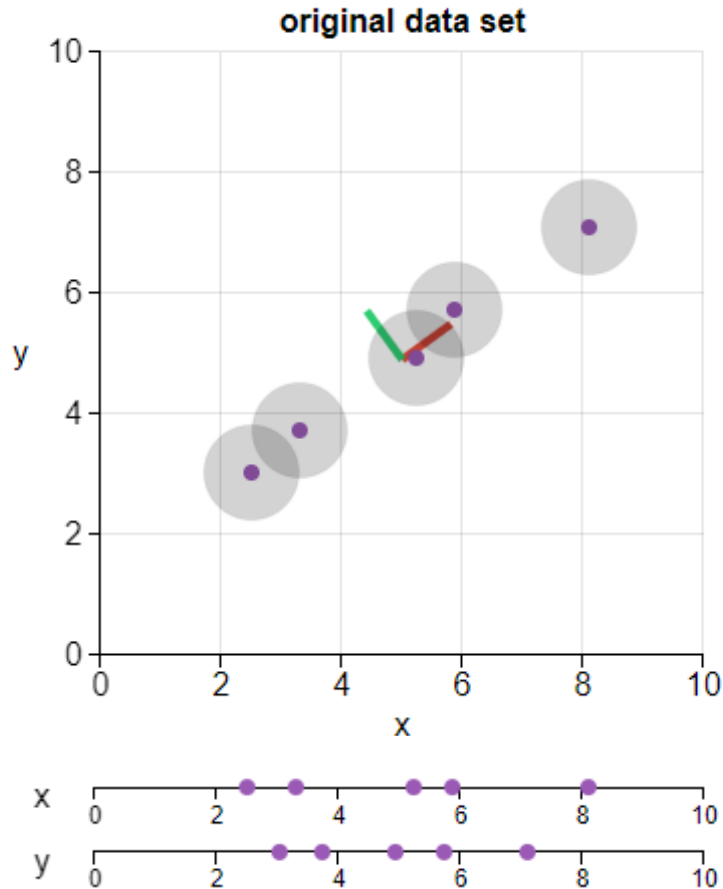
Principal Component Analysis (PCA)

<https://setosa.io/ev/principal-component-analysis/>



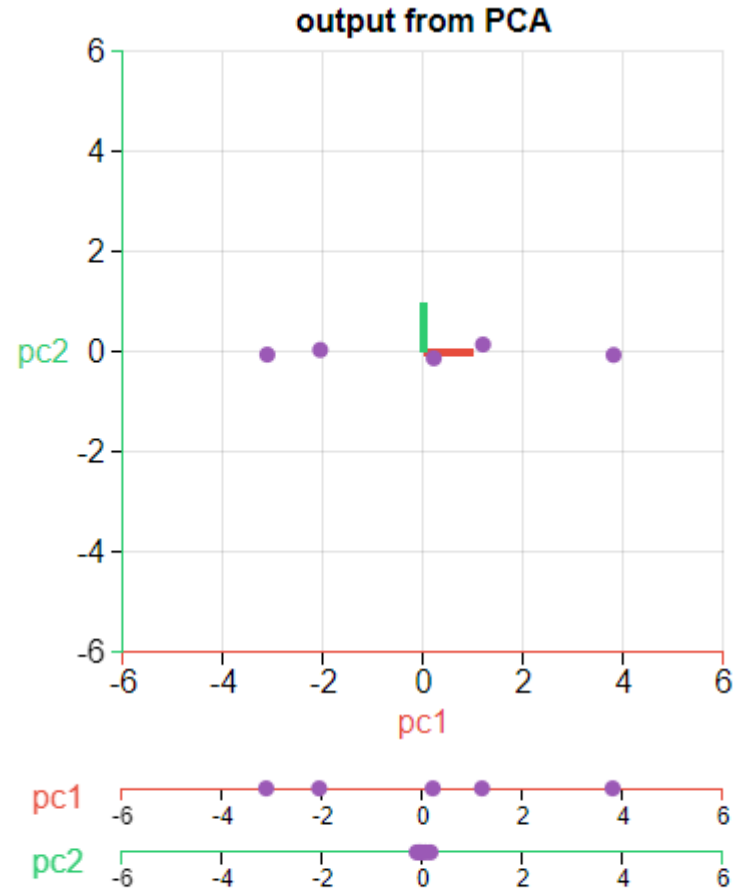
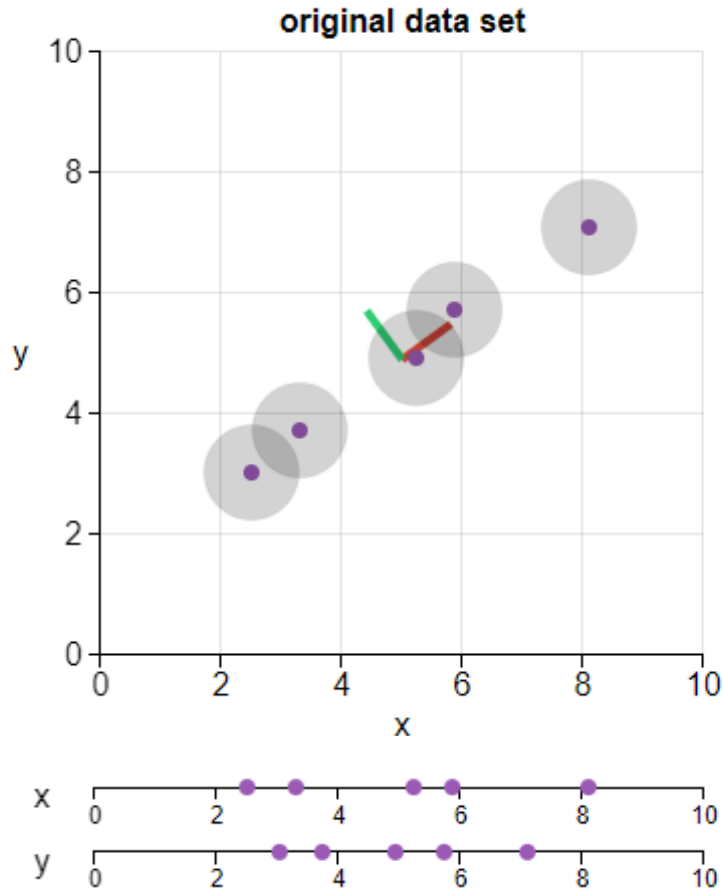
Principal Component Analysis (PCA)

<https://setosa.io/ev/principal-component-analysis/>



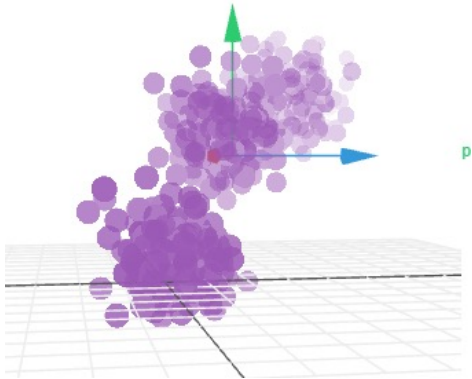
Principal Component Analysis (PCA)

<https://setosa.io/ev/principal-component-analysis/>



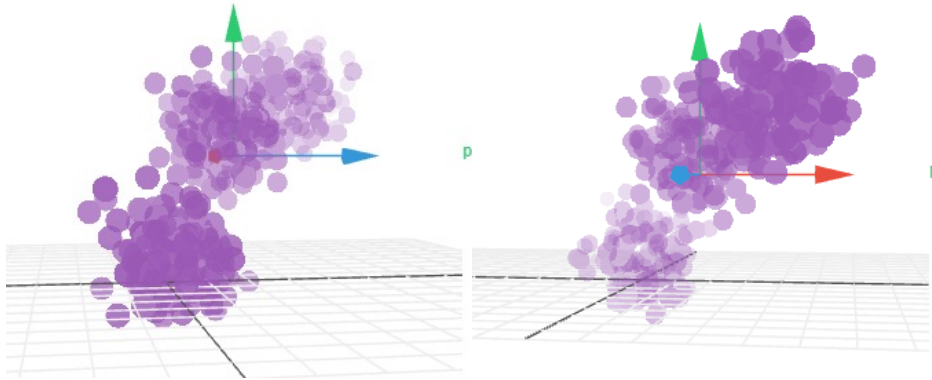
PCA in 3D

<https://setosa.io/ev/principal-component-analysis/>



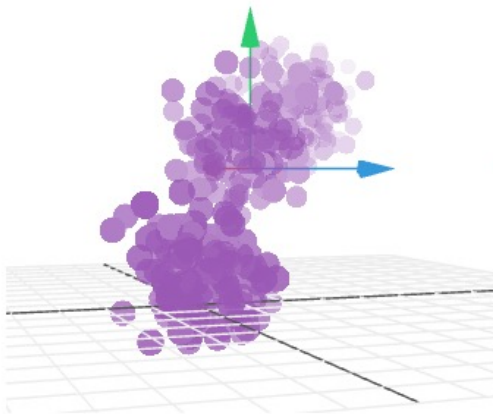
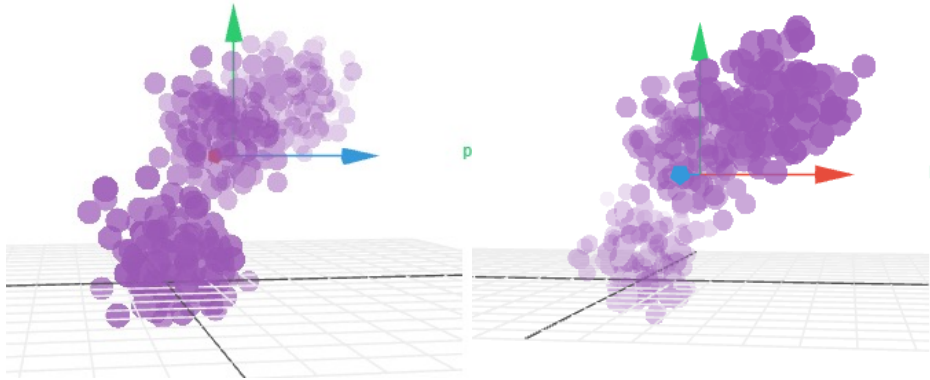
PCA in 3D

<https://setosa.io/ev/principal-component-analysis/>



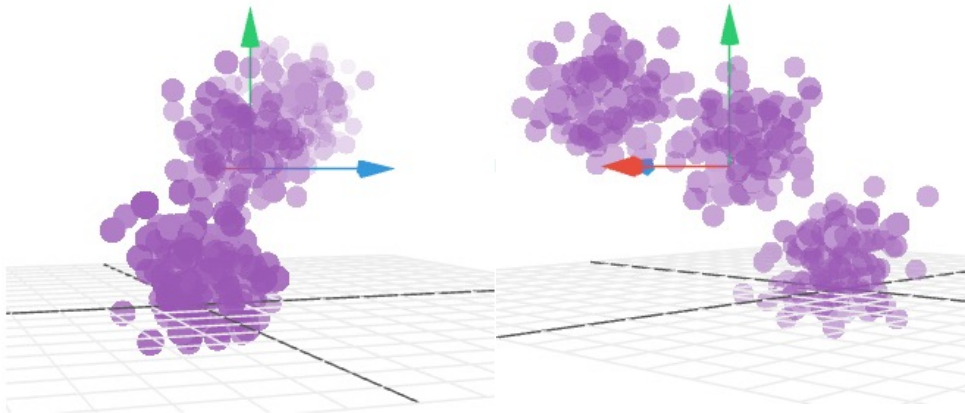
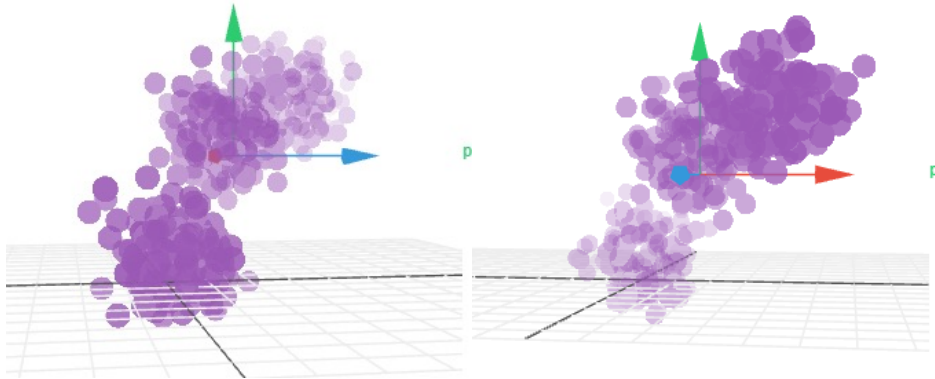
PCA in 3D

<https://setosa.io/ev/principal-component-analysis/>



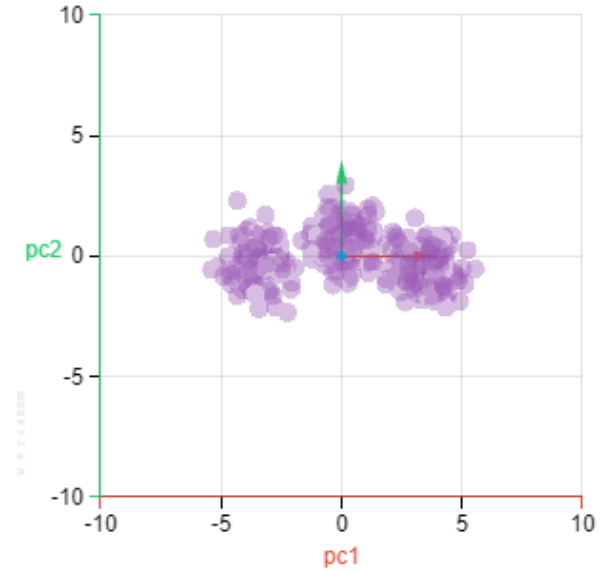
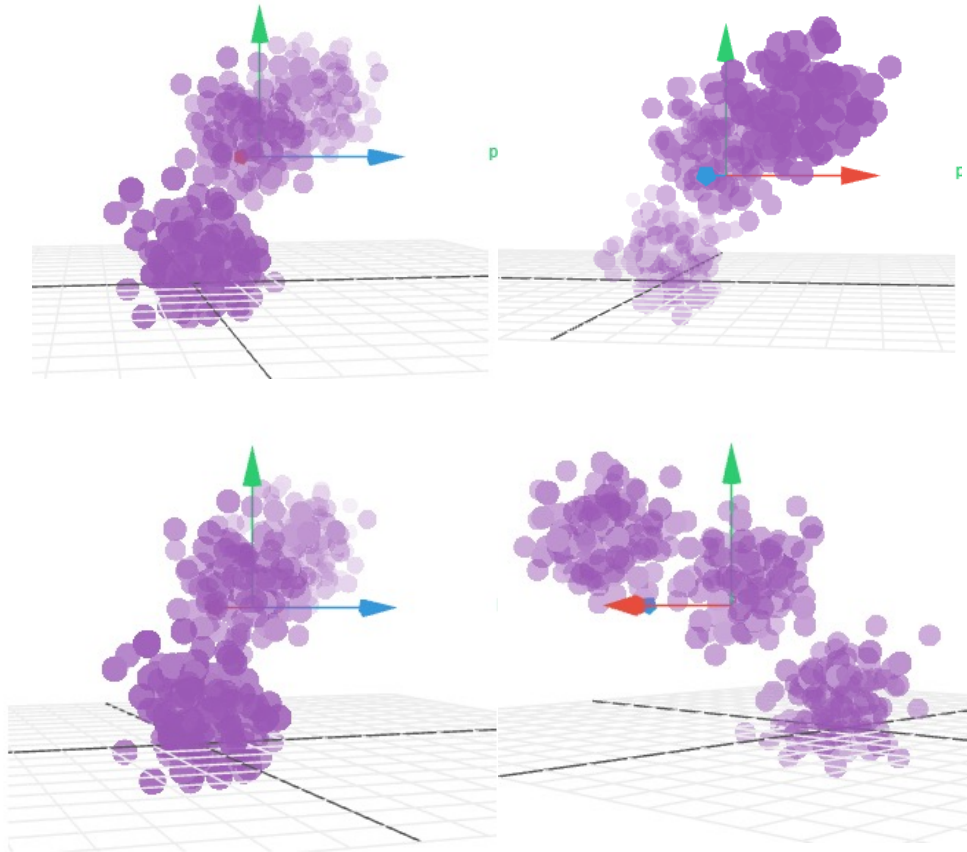
PCA in 3D

<https://setosa.io/ev/principal-component-analysis/>



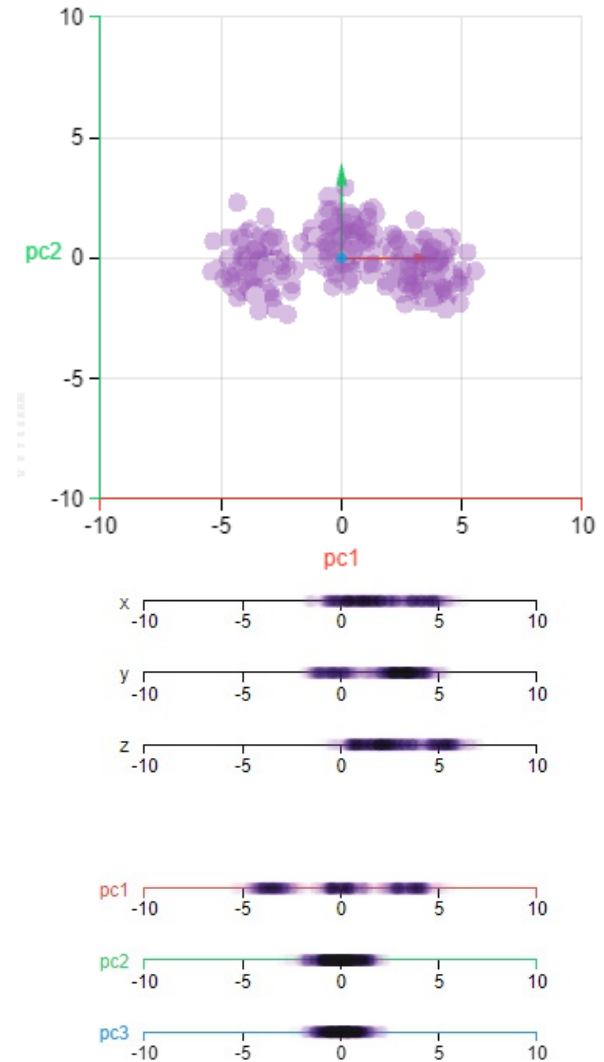
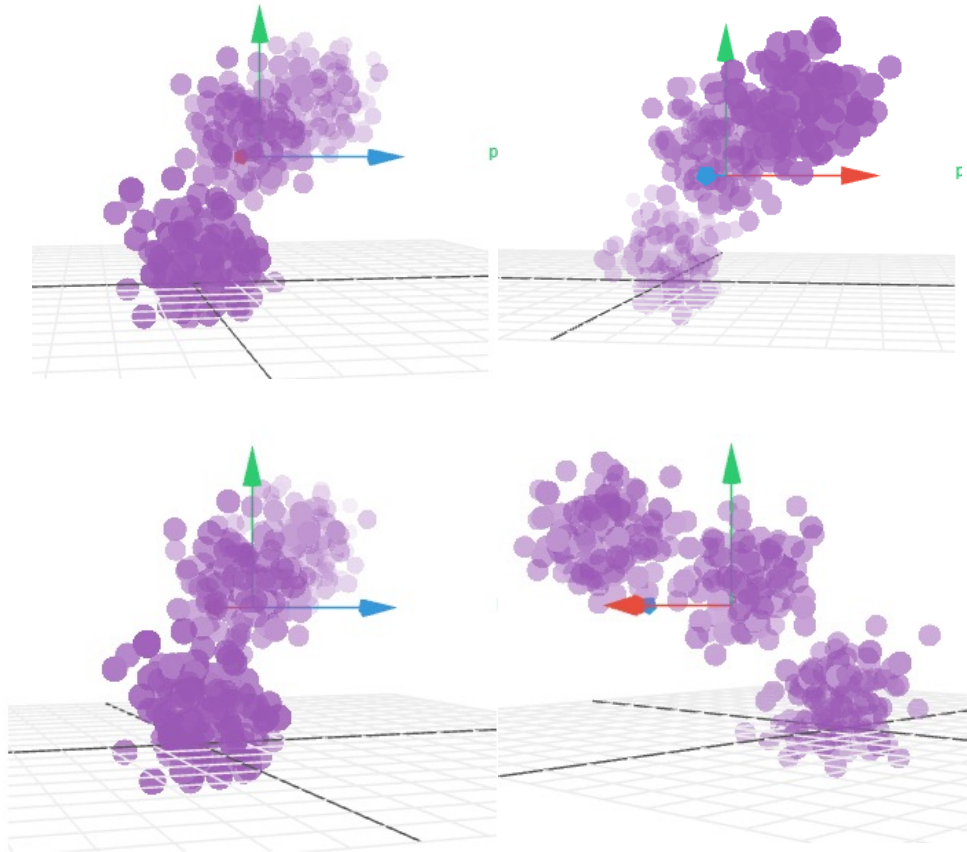
PCA in 3D

<https://setosa.io/ev/principal-component-analysis/>



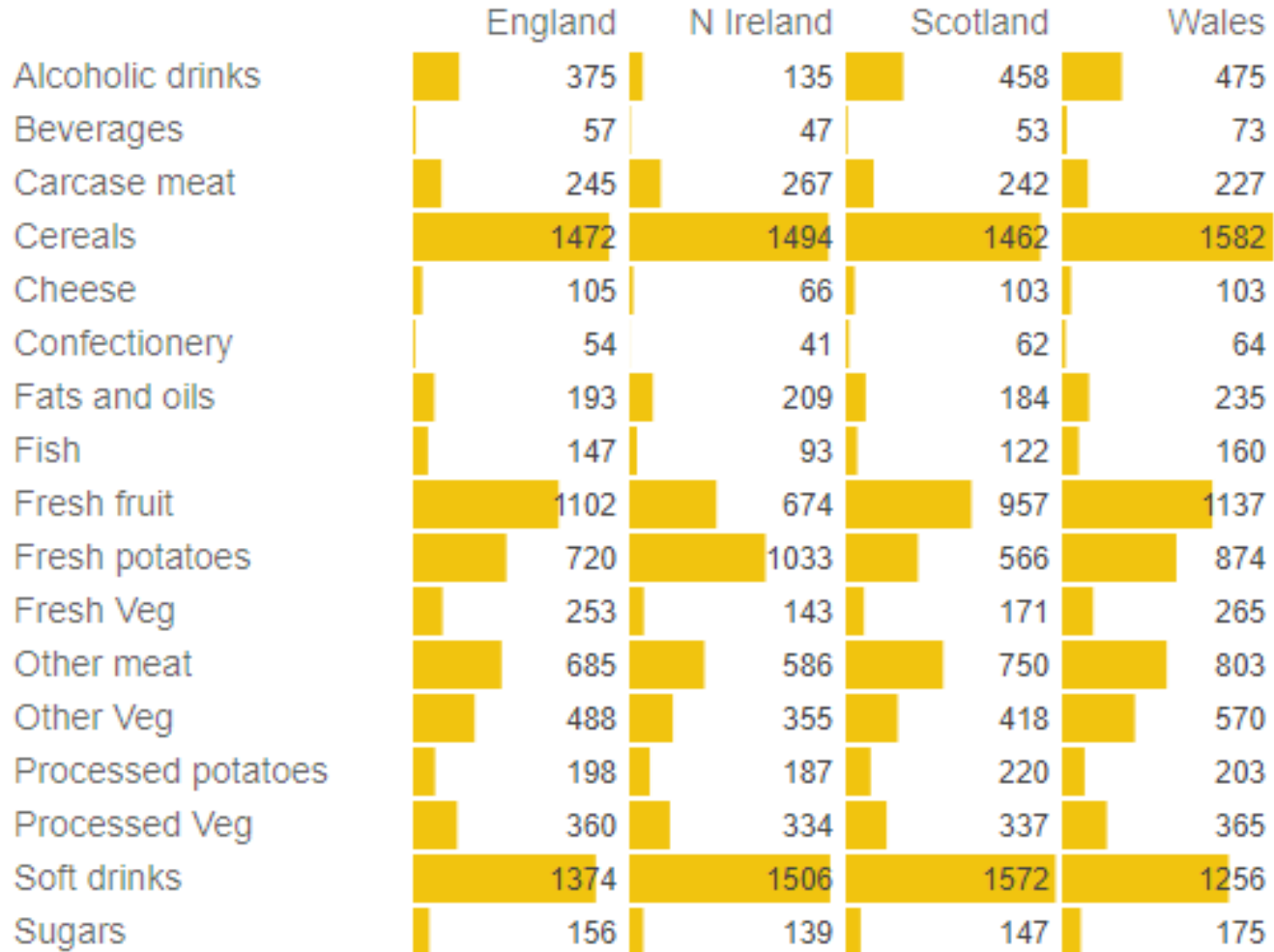
PCA in 3D

<https://setosa.io/ev/principal-component-analysis/>



PCA in 17D

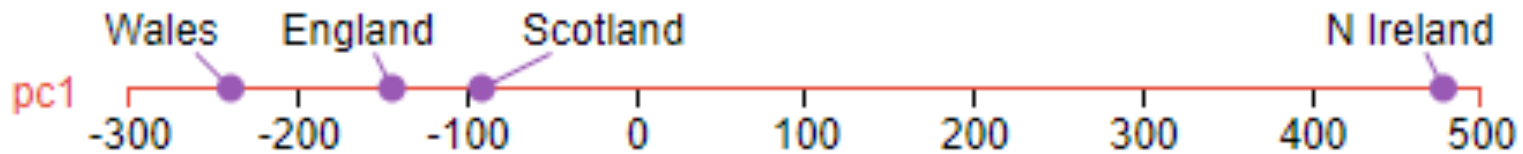
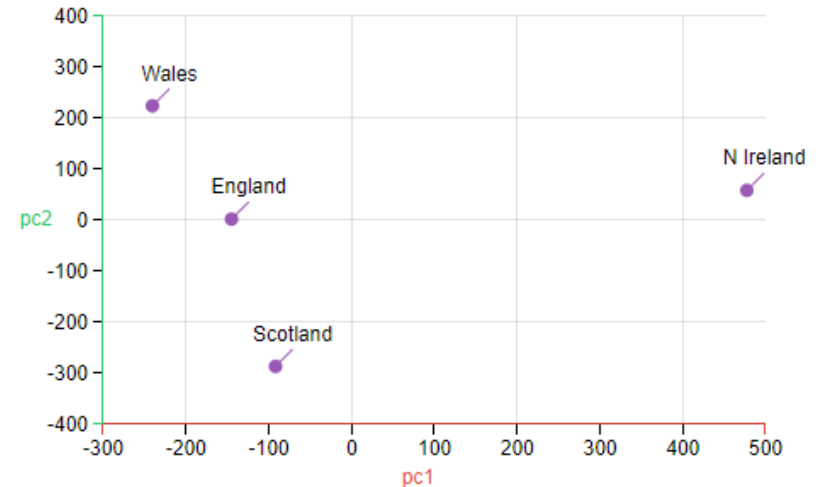
<https://setosa.io/ev/principal-component-analysis/>



PCA in 17D

<https://setosa.io/ev/principal-component-analysis/>

	England	N Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	874
Fresh Veg	253	143	171	265
Other meat	685	586	750	803
Other Veg	488	355	418	570
Processed potatoes	198	187	220	203
Processed Veg	360	334	337	365
Soft drinks	1374	1506	1572	1256
Sugars	156	139	147	175



OK, so how can we find the right basis?

1. Standardization



OK, so how can we find the right basis?

1. Standardization
2. Compute covariance matrix

$$\begin{bmatrix} \text{Cov}(x, x) & \text{Cov}(x, y) & \text{Cov}(x, z) \\ \text{Cov}(y, x) & \text{Cov}(y, y) & \text{Cov}(y, z) \\ \text{Cov}(z, x) & \text{Cov}(z, y) & \text{Cov}(z, z) \end{bmatrix}$$

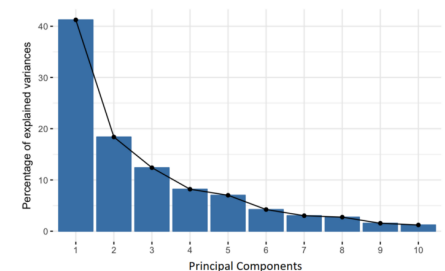


OK, so how can we find the right basis?

1. Standardization
2. Compute covariance matrix

$$\begin{bmatrix} Cov(x, x) & Cov(x, y) & Cov(x, z) \\ Cov(y, x) & Cov(y, y) & Cov(y, z) \\ Cov(z, x) & Cov(z, y) & Cov(z, z) \end{bmatrix}$$

3. Compute eigenvectors and eigenvalues

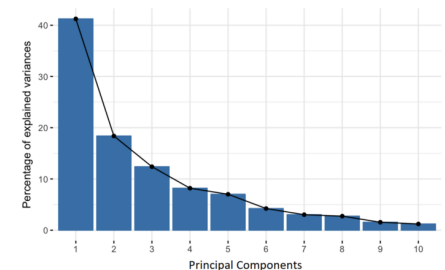


OK, so how can we find the right basis?

1. Standardization
2. Compute covariance matrix

$$\begin{bmatrix} Cov(x, x) & Cov(x, y) & Cov(x, z) \\ Cov(y, x) & Cov(y, y) & Cov(y, z) \\ Cov(z, x) & Cov(z, y) & Cov(z, z) \end{bmatrix}$$

3. Compute eigenvectors and eigenvalues
4. Discard vectors that are not important enough



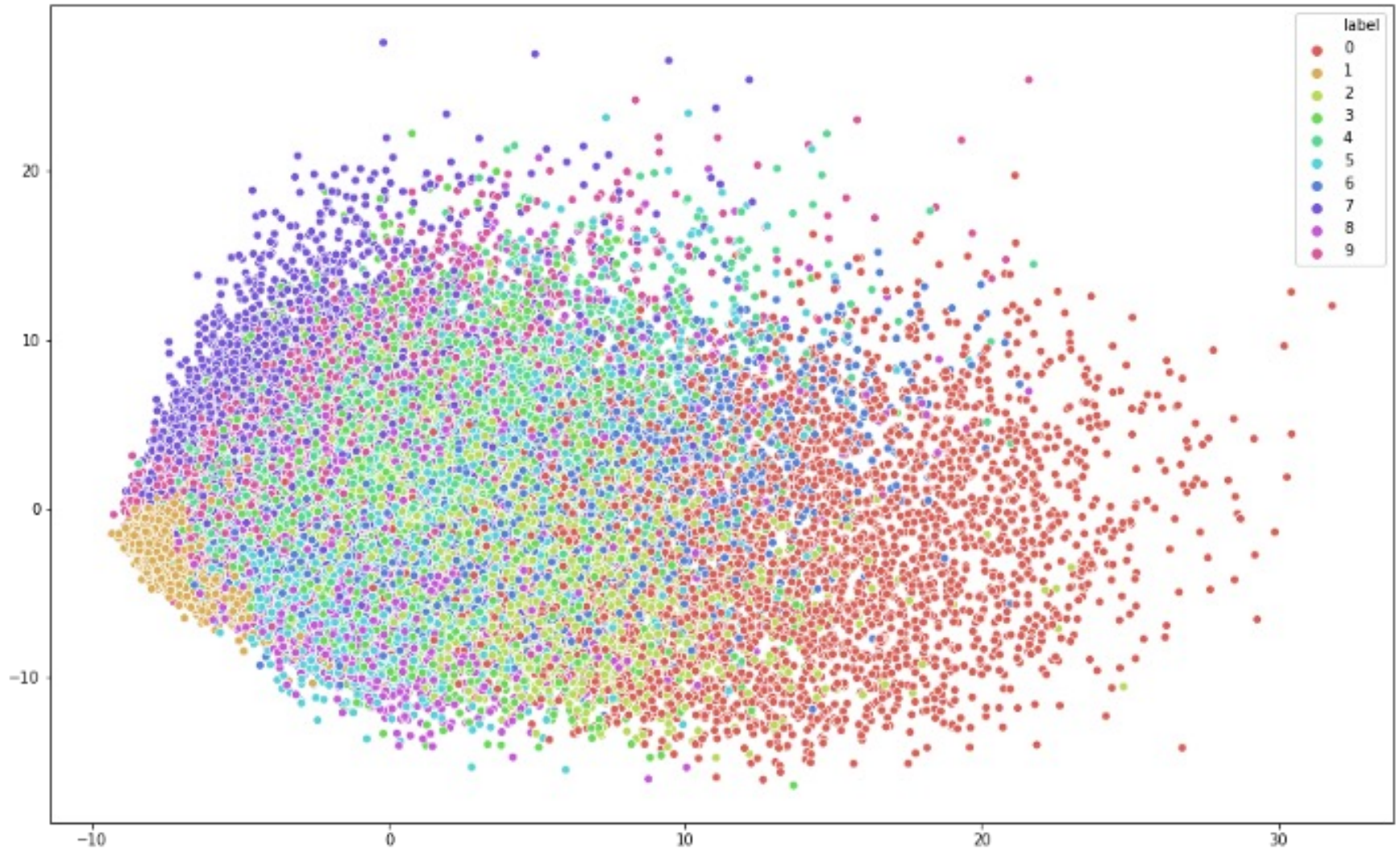
Example: Handwritten Digits

MNIST dataset



Group “similar” things together

Principal Component Analysis



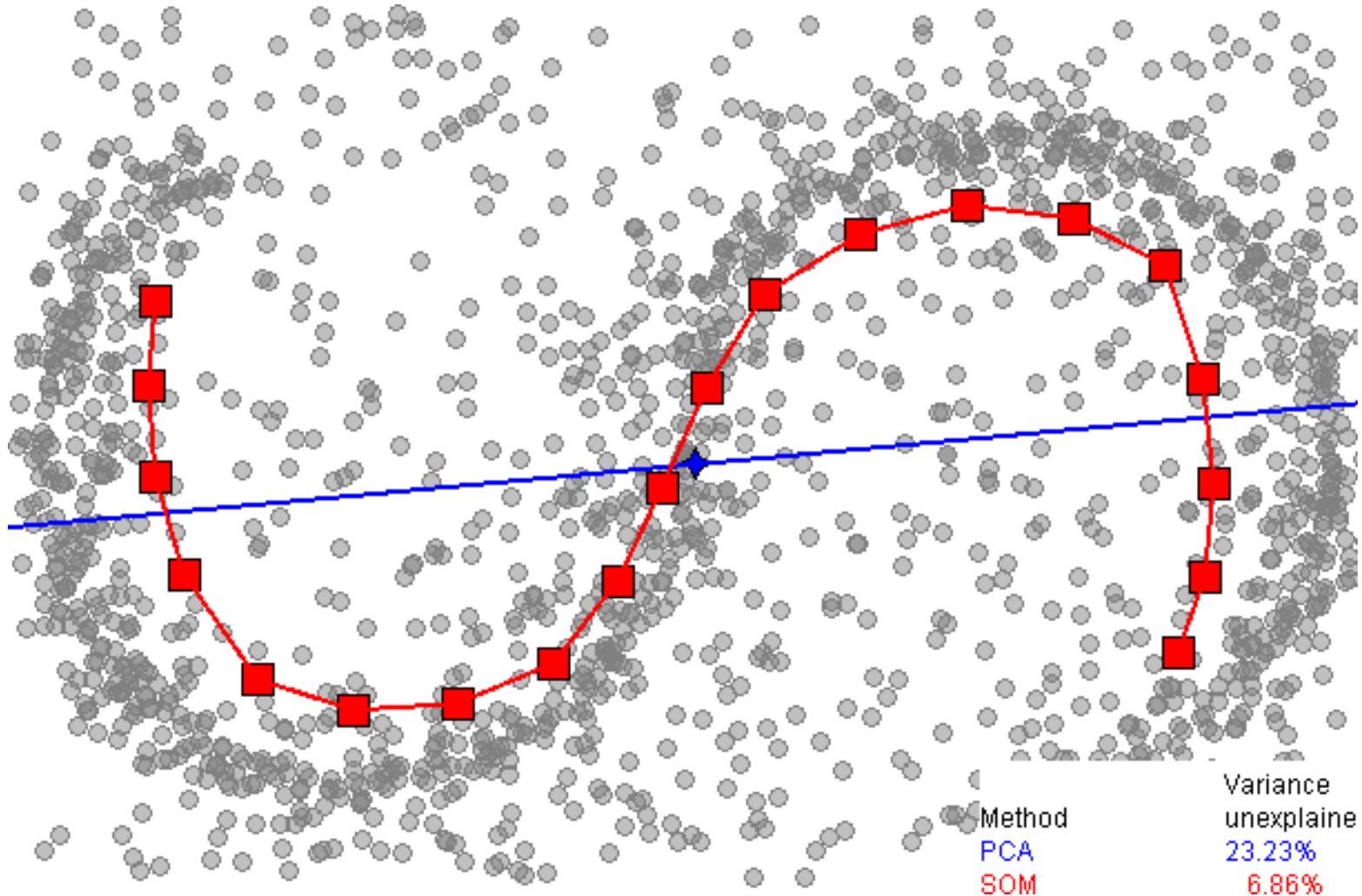
Example: Handwritten Digits

MNIST dataset



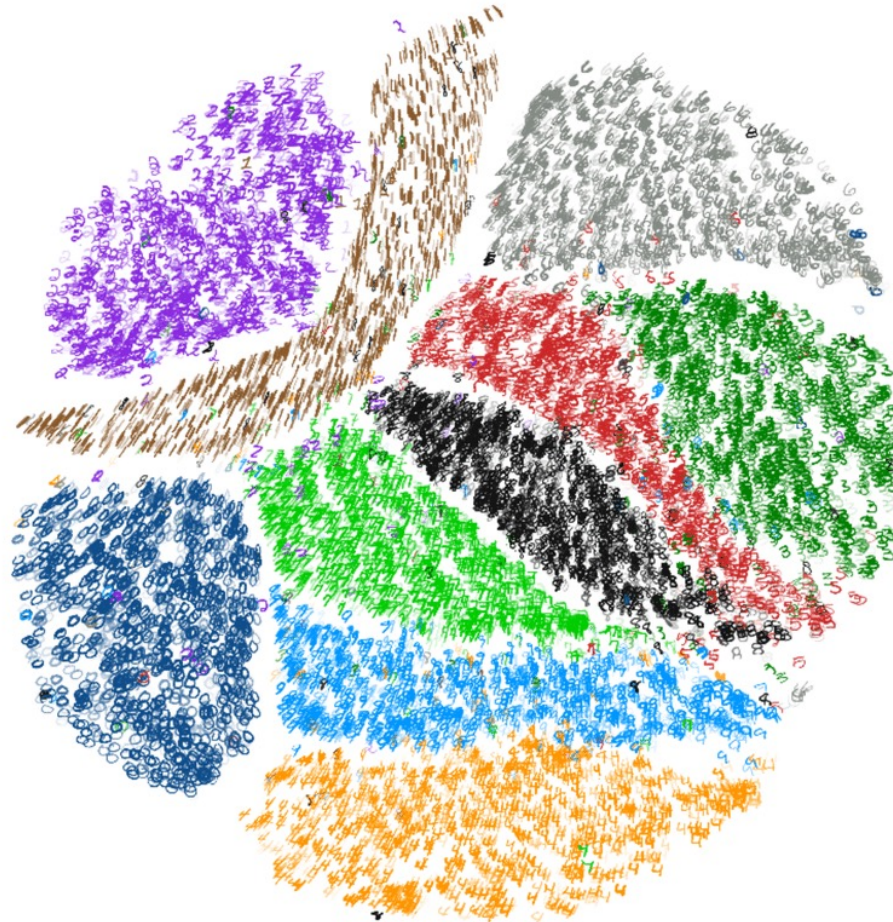
Some things aren't linear!

Wikimedia Commons



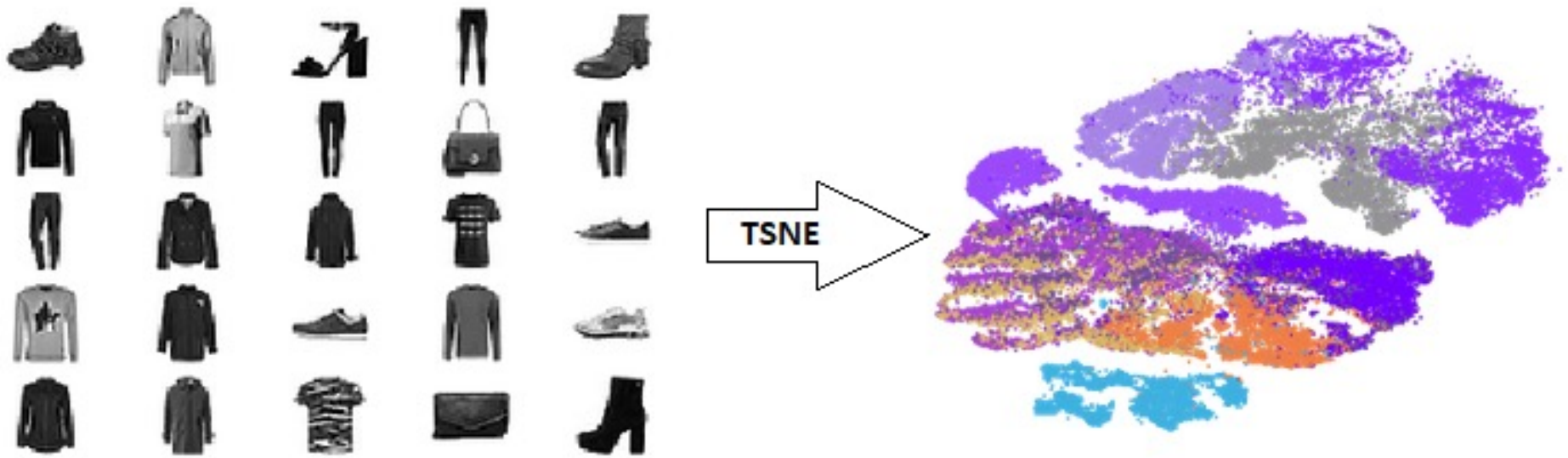
Group “similar” things together

Pezzotti 2019



Group “similar” things together

“Fashion MNIST” datasets, t-SNE



Group “similar” things together

Wang et al. 2020

t-SNE(perplexity=10)



UMAP(n_neighbors=10)



TriMAP(n_inliers=8)



t-SNE(perplexity=20)



UMAP(n_neighbors=20)



TriMAP(n_inliers=10)



PaCMAP



t-SNE(perplexity=40)



UMAP(n_neighbors=40)

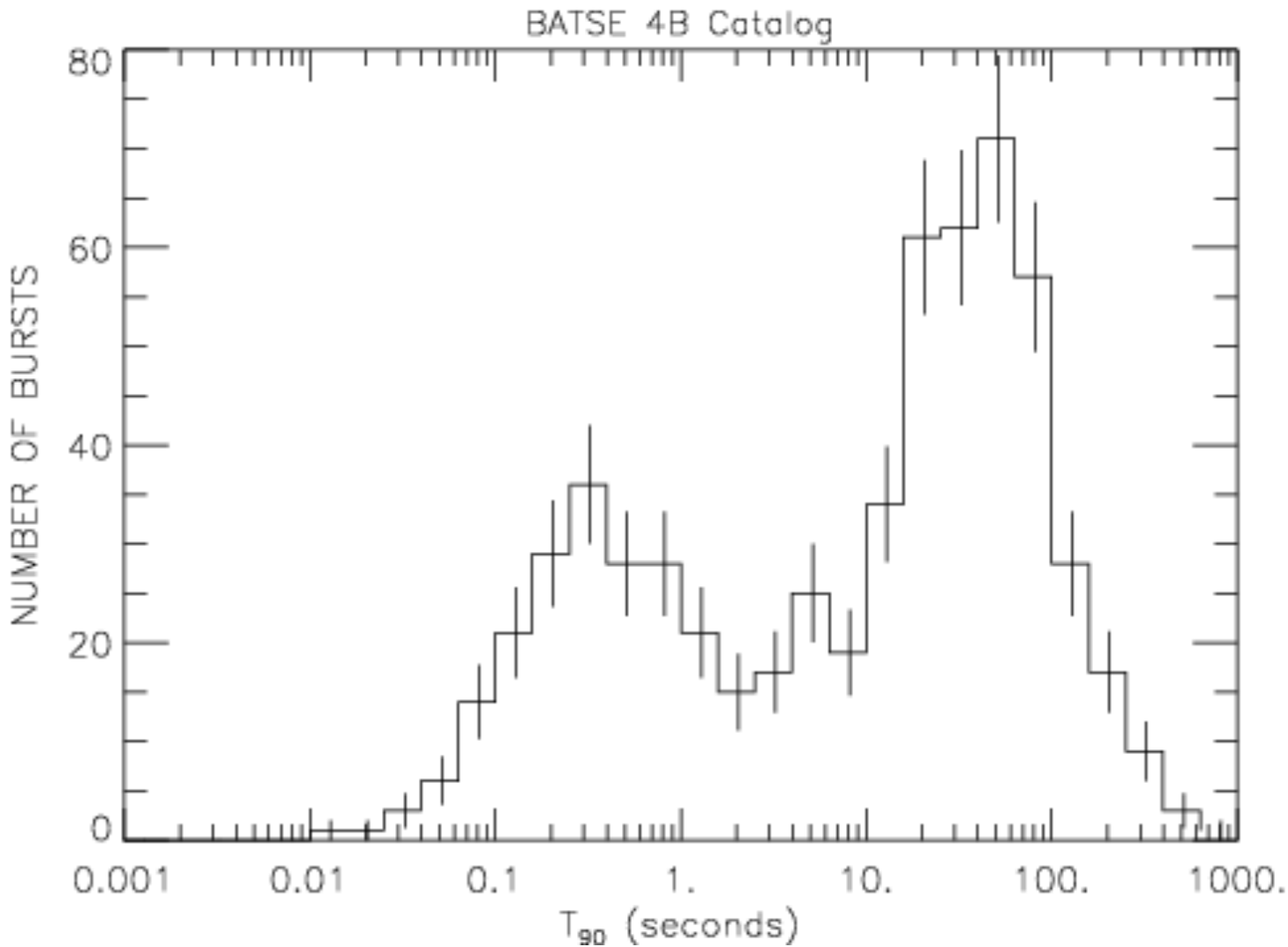


TriMAP(n_inliers=15)



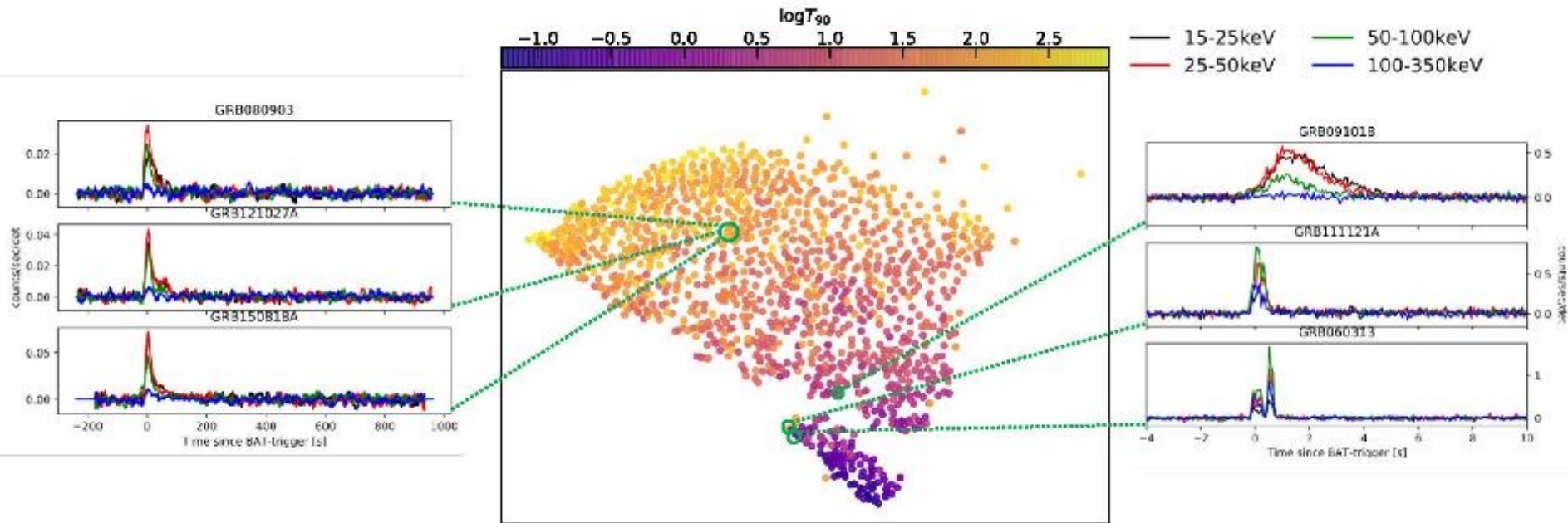
Example: Separate Short and Long GRBs

R. Mallozzi, updated Aug 2018 at <https://gammaray.msfc.nasa.gov/batse/grb/duration/>



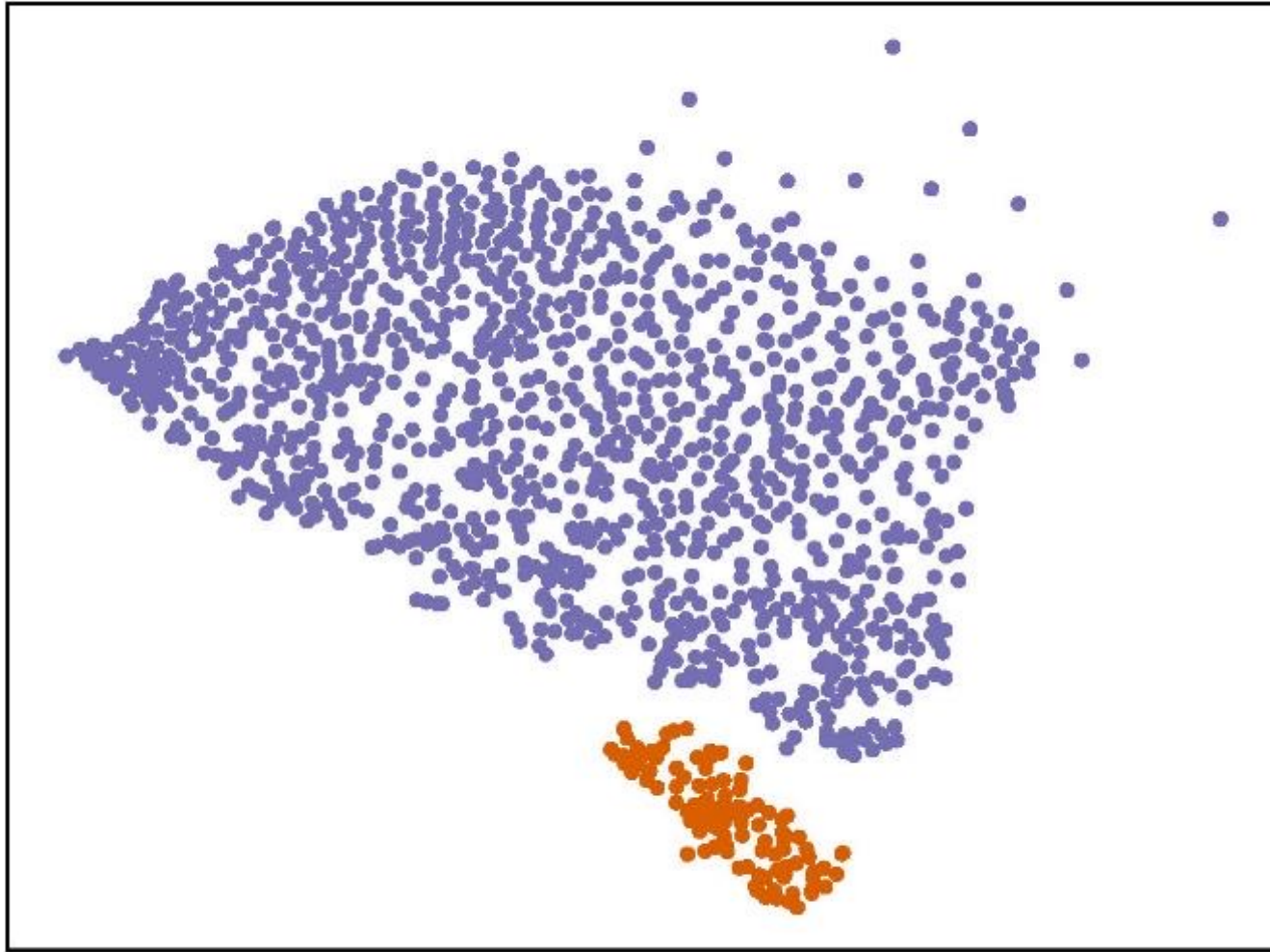
t-SNE map for *Swift* light curves

Kragh Jespersen et al. 2020



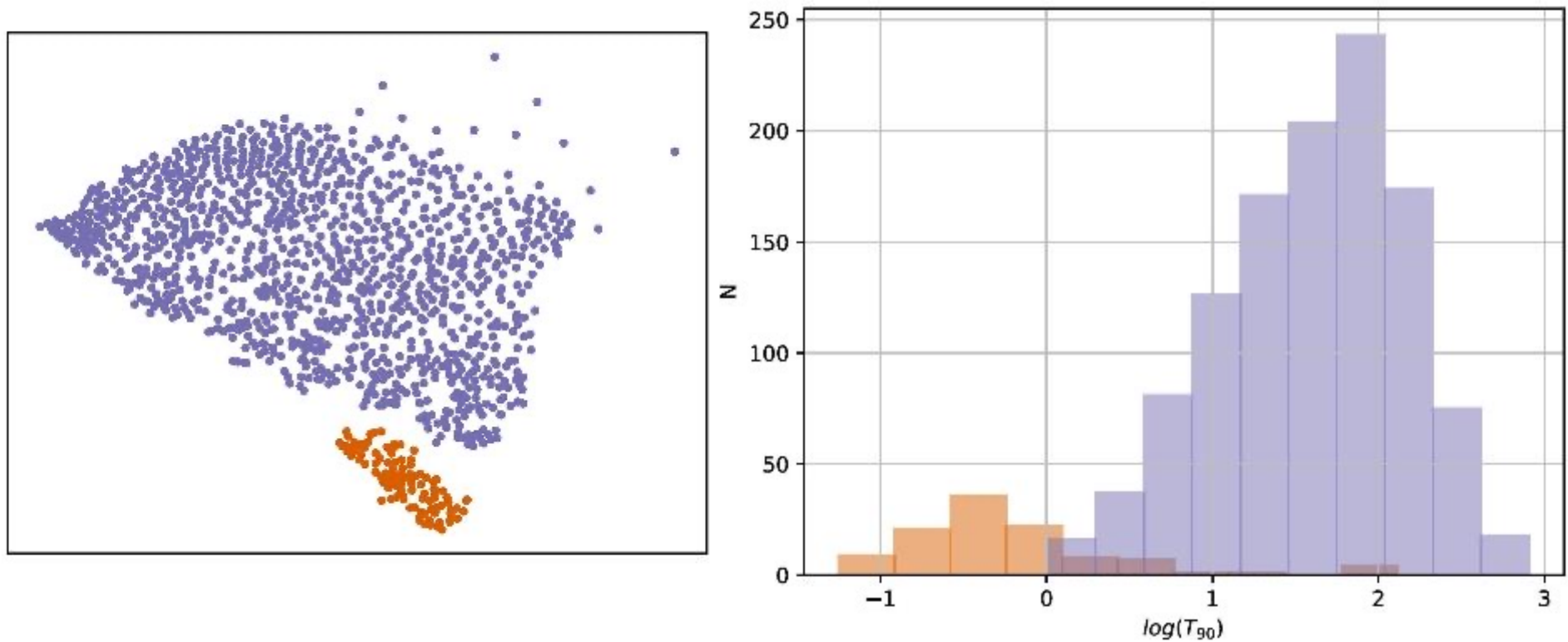
t-SNE map for *Swift* light curves

Kraah Jespersen et al. 2020

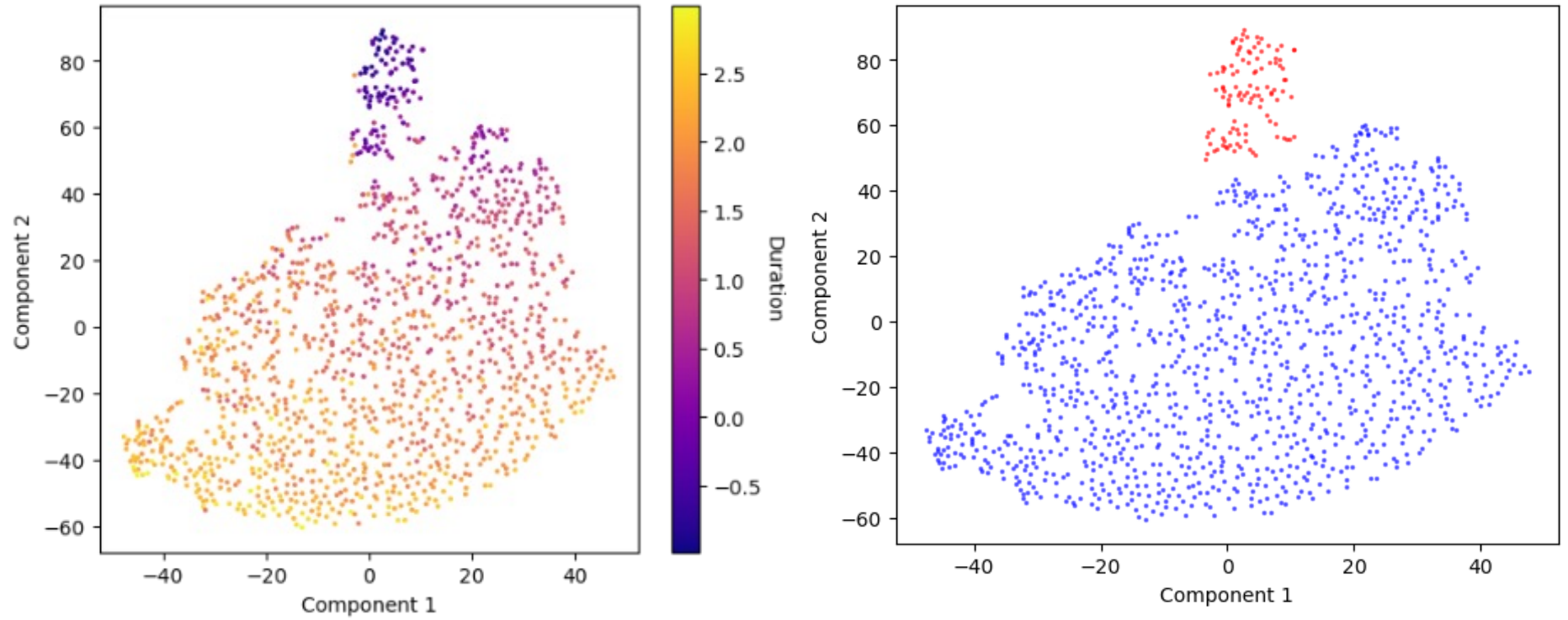


t-SNE map for *Swift* light curves

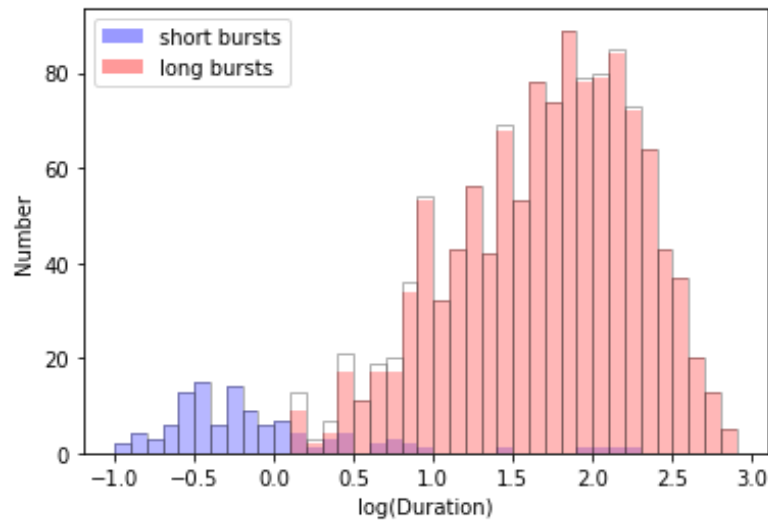
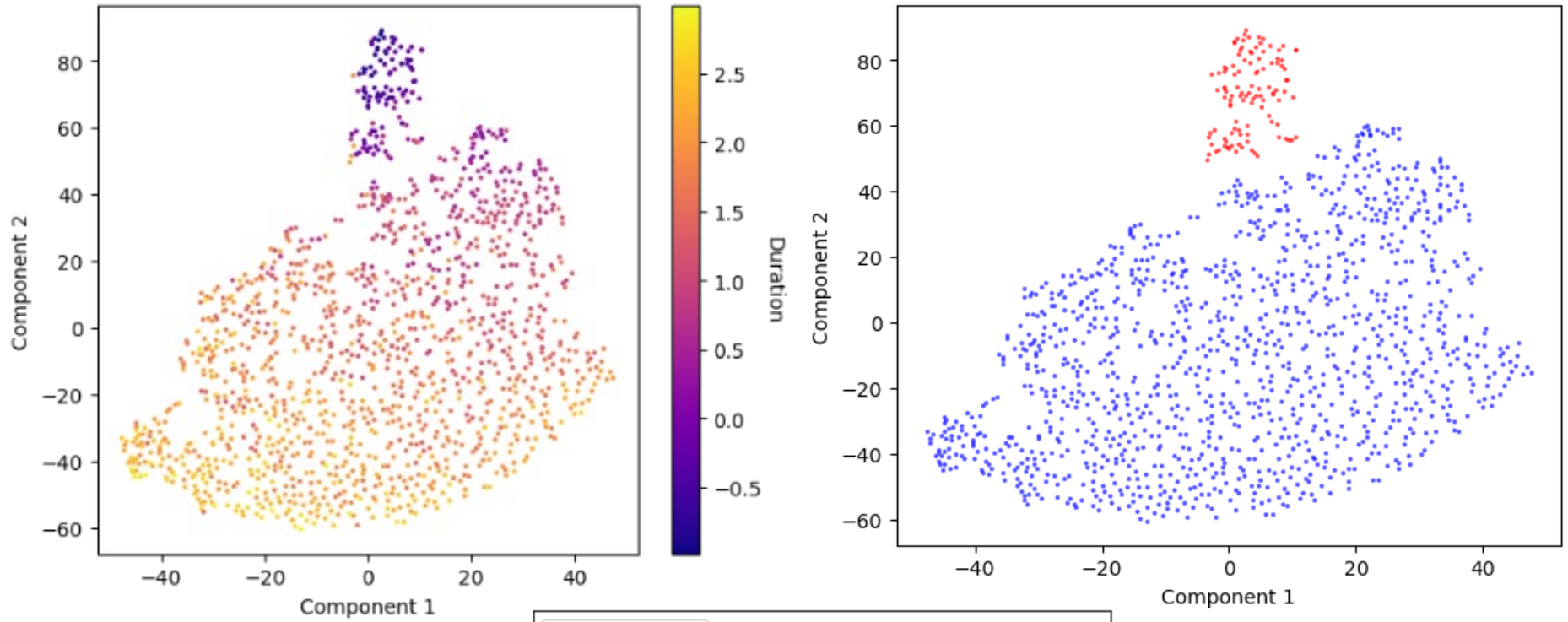
Kragh Jespersen et al. 2020



Embeddings, colored by duration

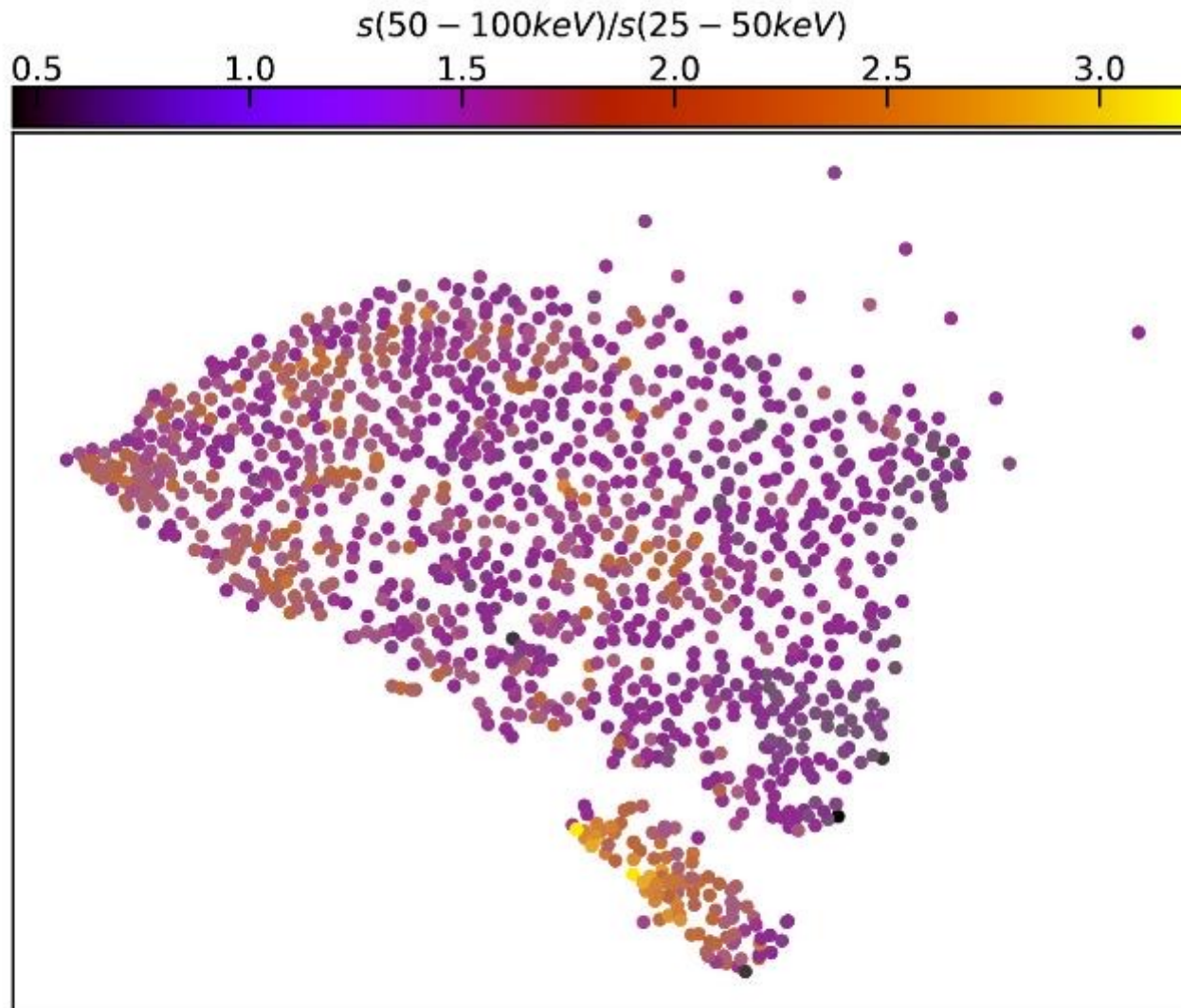


Embeddings, colored by duration



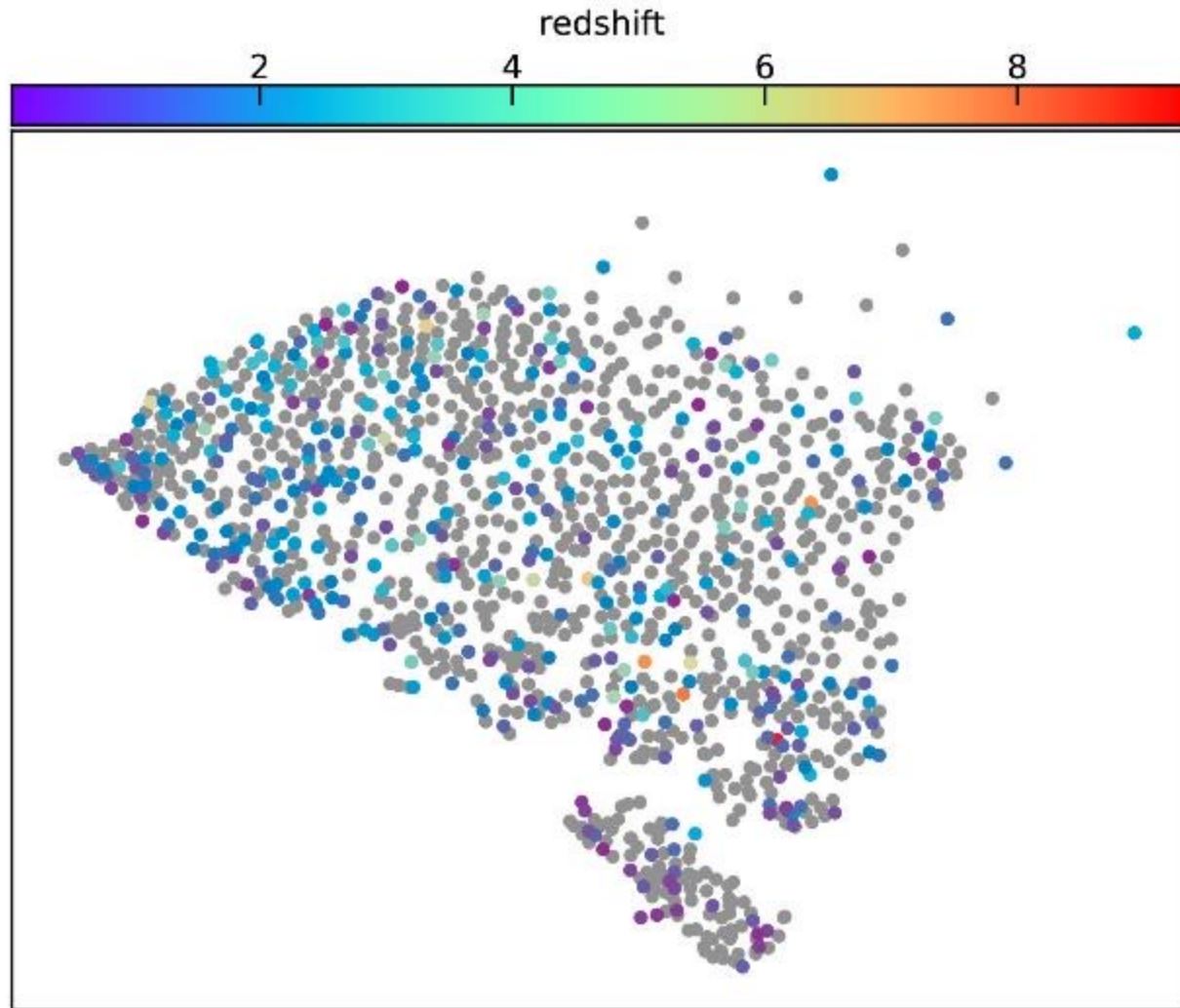
Hardness distribution

Kragh Jespersen et al. 2020



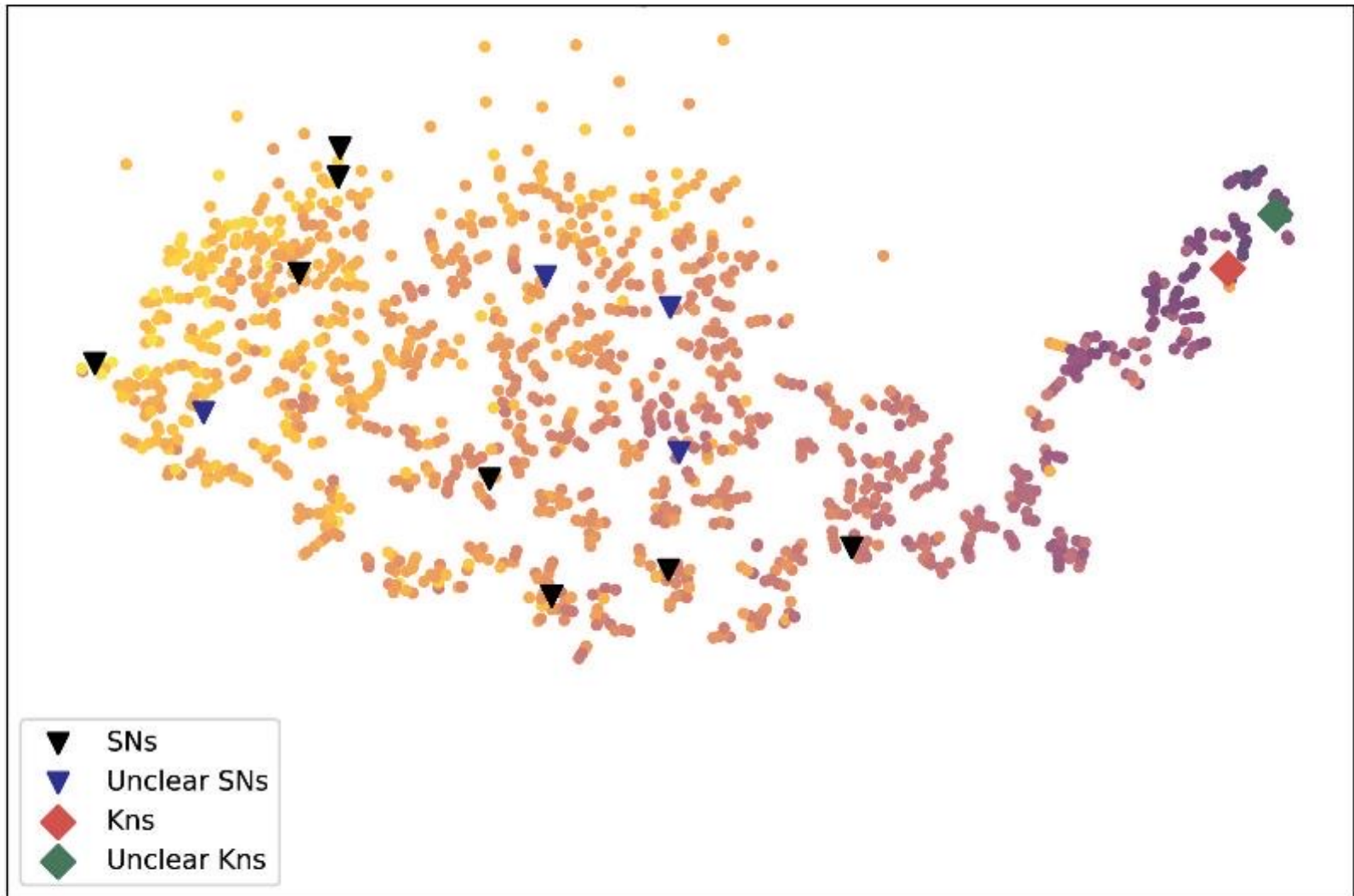
Redshift distribution

Kragh Jespersen et al. 2020



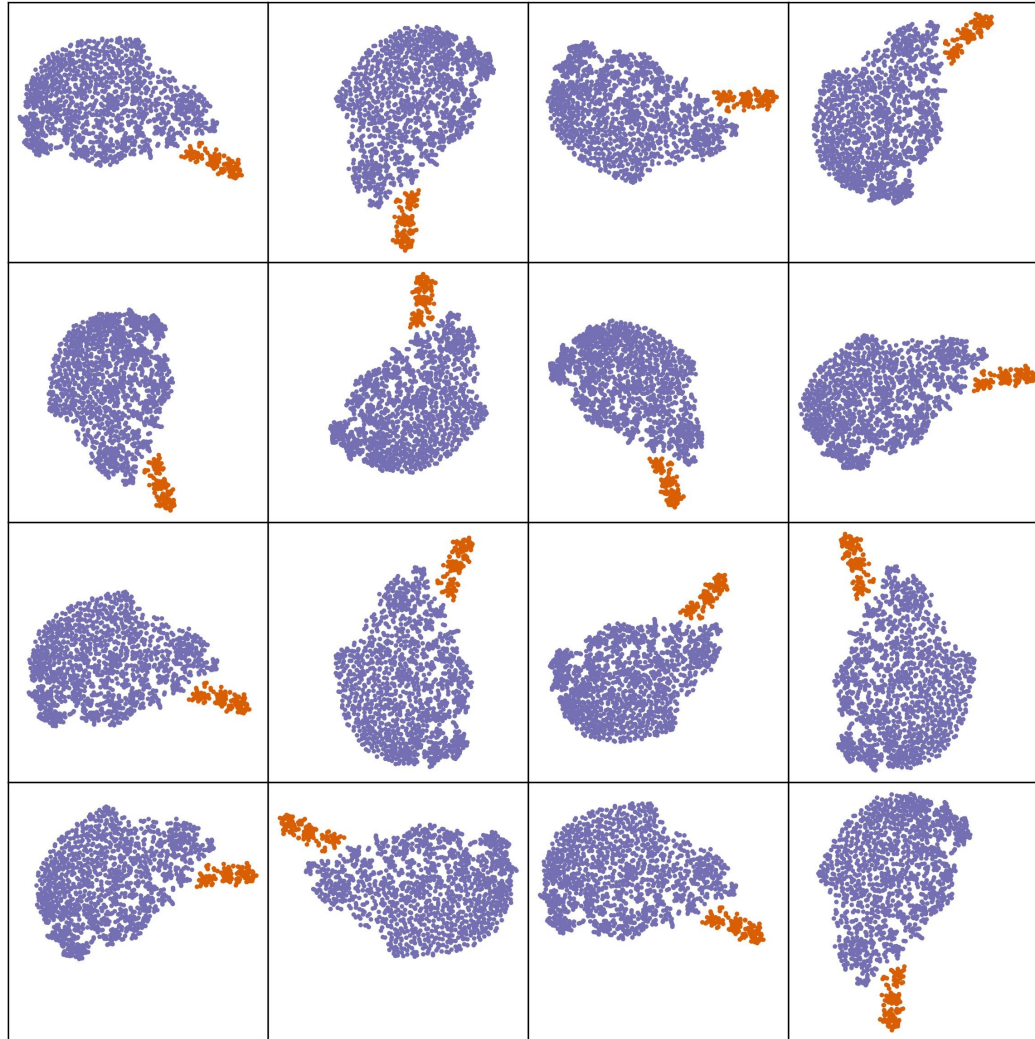
Possible subgroupings?

Kragh Jespersen et al. 2020



Structure is durable, not location!

Steinhardt, Mann, Rusakov, and Kragh Jespersen 2023



Objects can “jump”

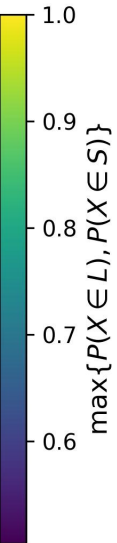
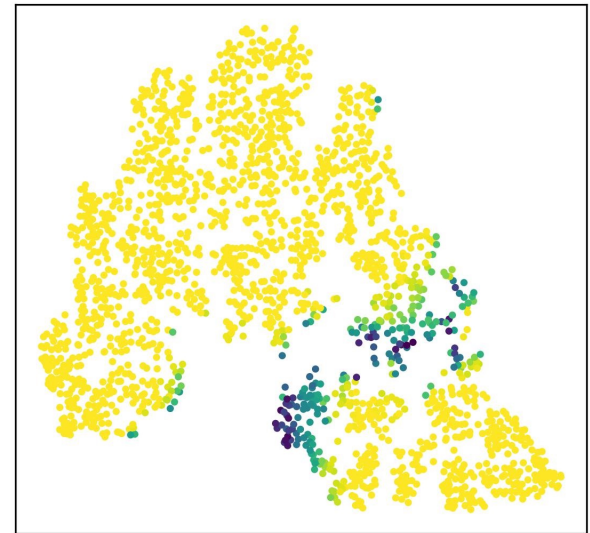
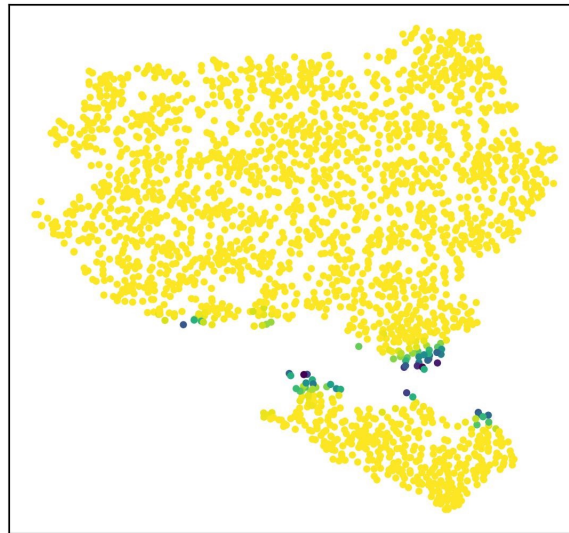
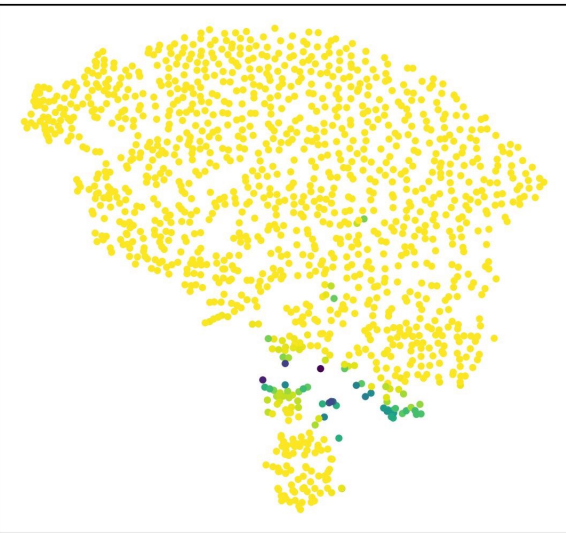
Steinhardt, Mann, Rusakov, and Kragh Jespersen 2023

Swift

Stability

Fermi

BATSE



Example: Photometry



Example: Photometry



V

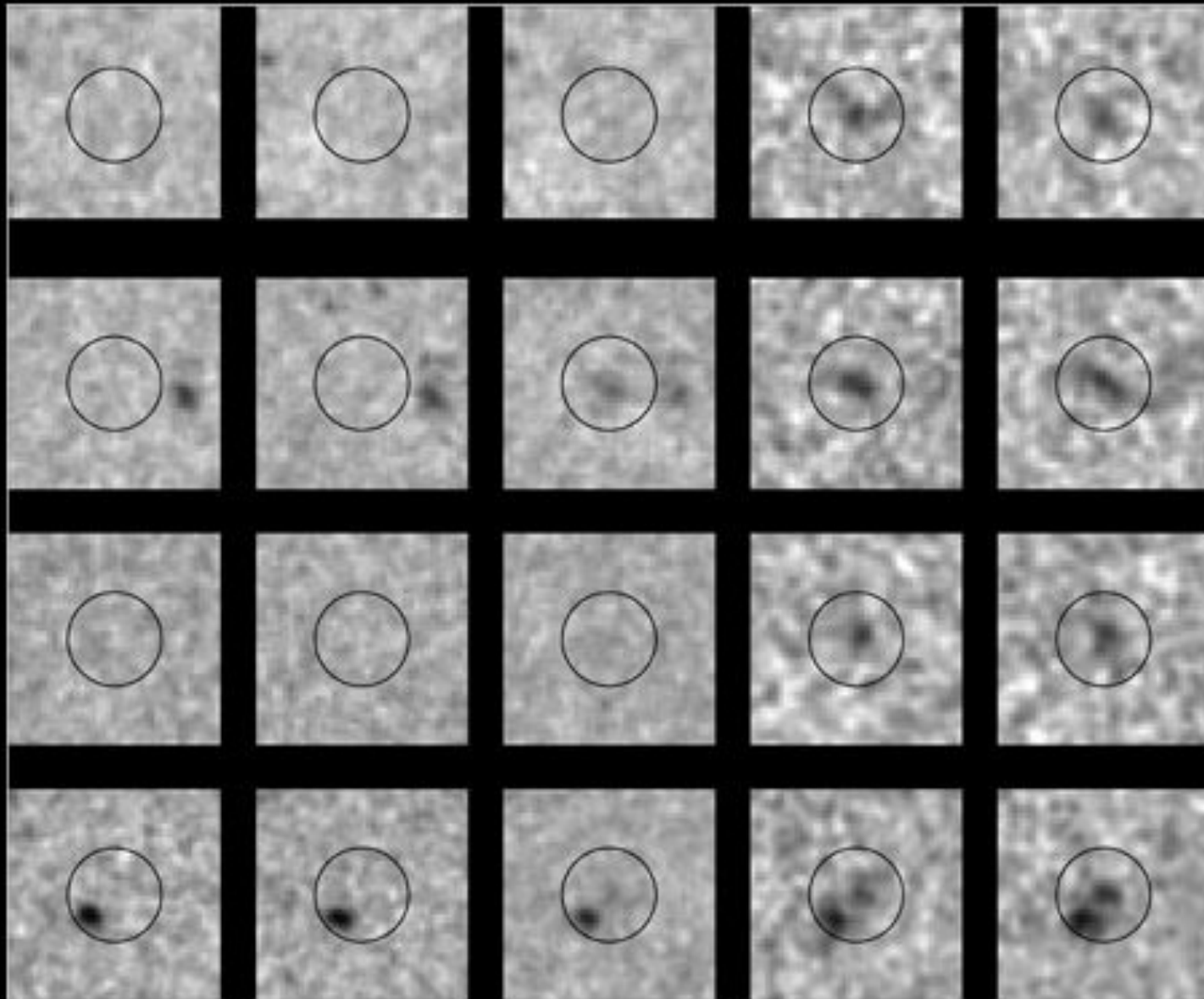
i

Z

J

H

Bouwens et al. 2006



$z \sim 7.4$

$z \sim 6.8$

$z \sim 6.8$

$z \sim 6.8$

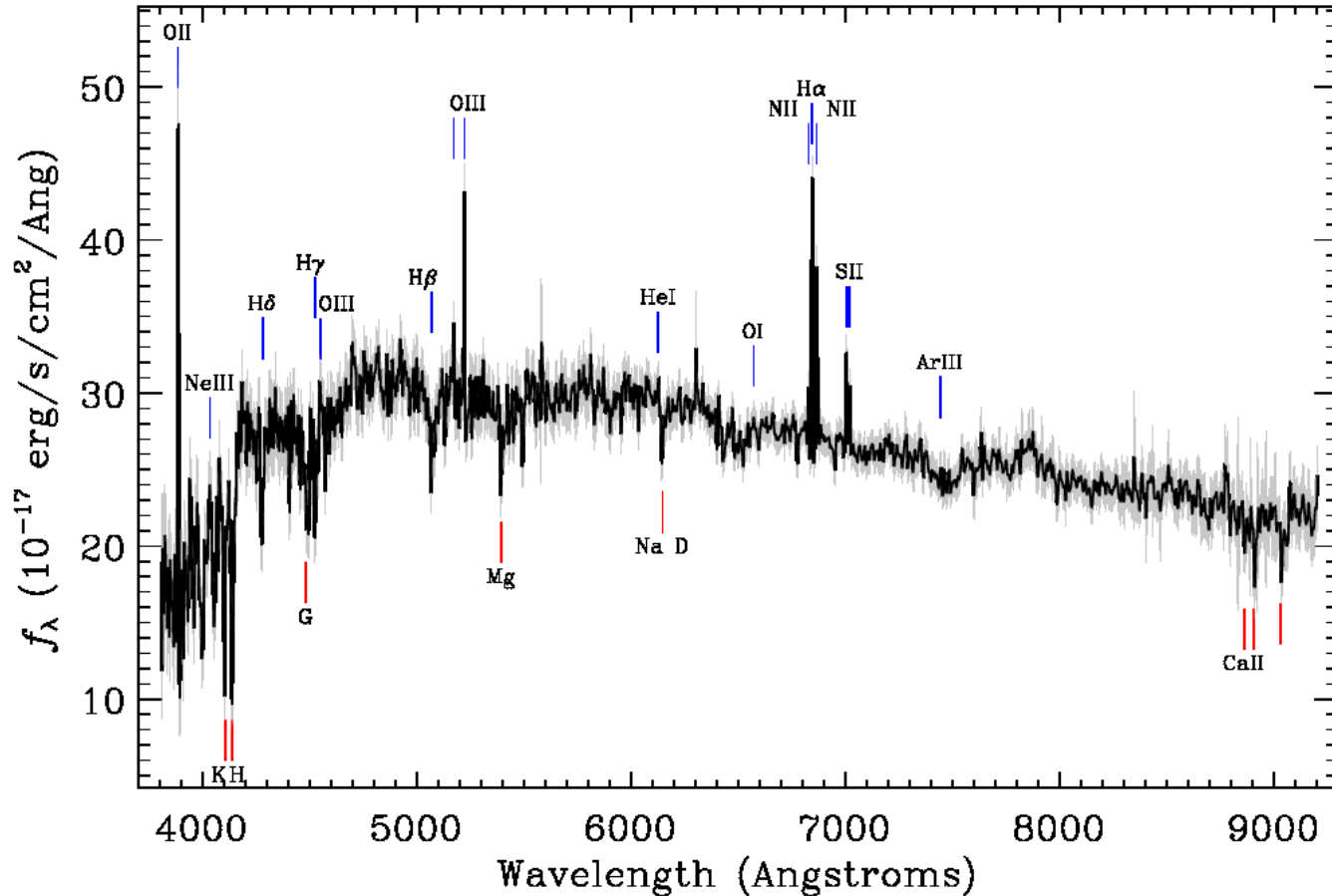
no detection

detection

Example: Galaxy spectra

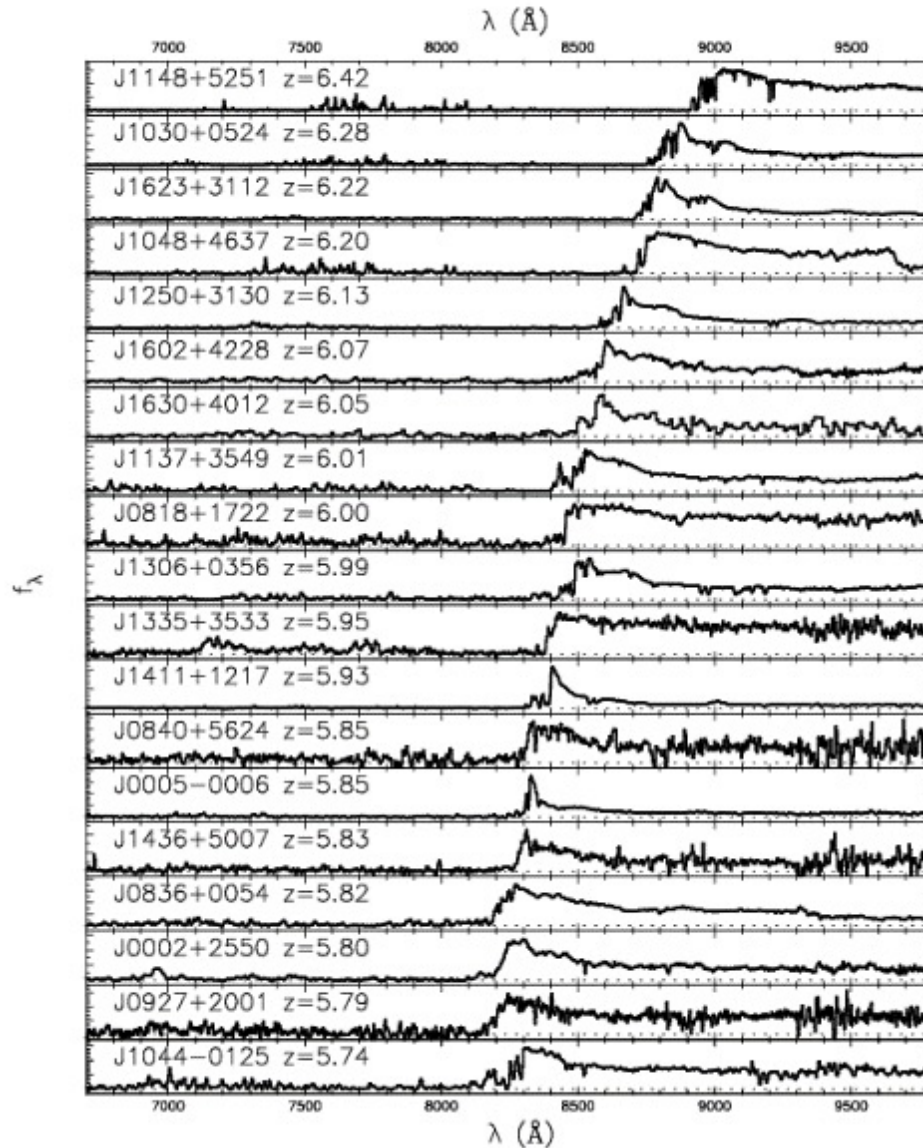
SDSS/Galaxy Zoo

Survey: *sdss* Program: *legacy* Target: *GALAXY ROSAT_D ROSAT_E*
RA=25.85806, Dec=-1.22998, Plate=401, Fiber=125, MJD=51788
 $z=0.04263 \pm 0.00002$ Class=GALAXY AGN
No warnings.



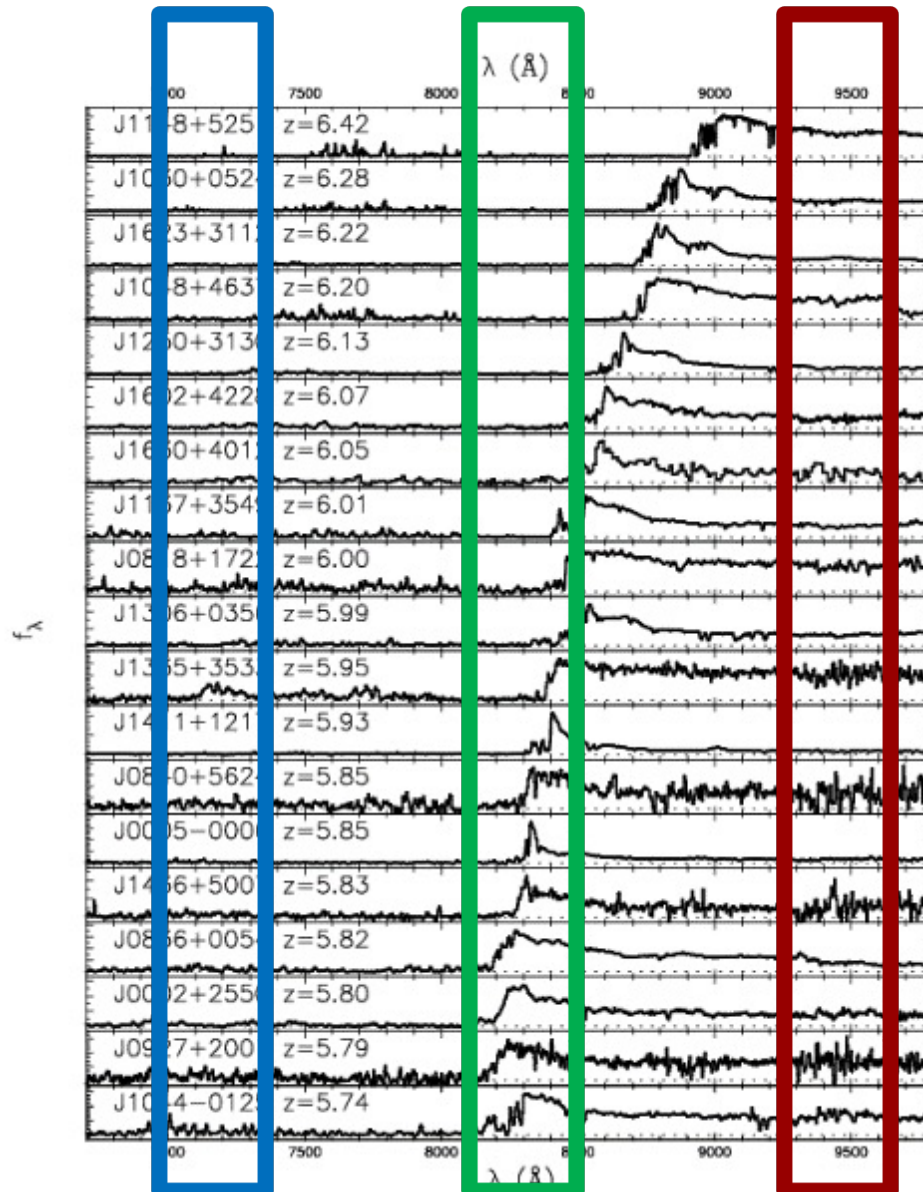
Spectra have similar features!

Zaroubi et al. 2013

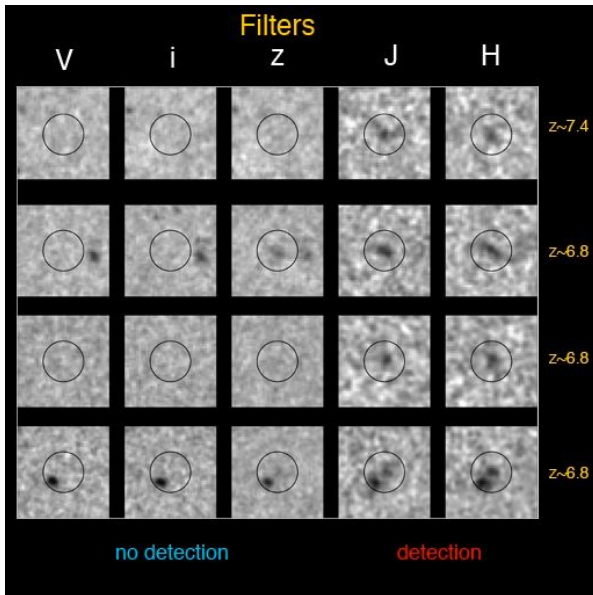


Maybe limited information is enough

Zaroubi et al. 2013



Our goal:



"V band"
"i band"
"z band"
"J band"
"H band"

etc.



(Stellar) Mass
Star formation rate
Star formation history
Distance/Redshift
Age

etc.



Two Fundamental Assumptions for Photometry

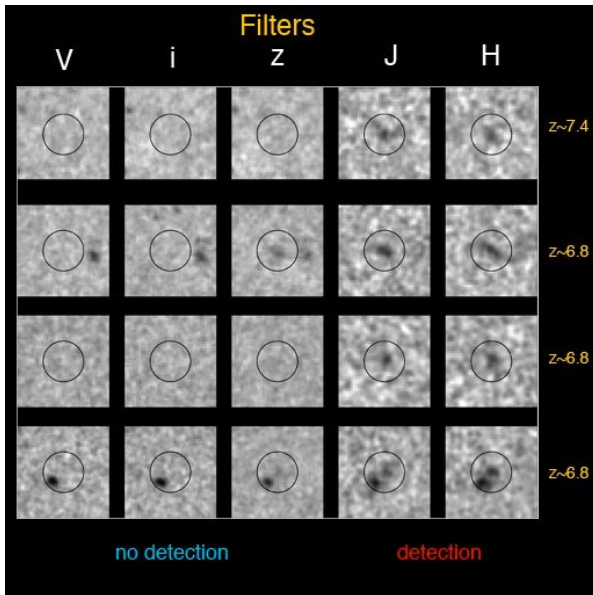


Two Fundamental Assumptions for Photometry

1. If an object is sufficiently well-measured, there is a surjective (one-to-one or many-to-one, but not one-to-many) mapping from photometric fluxes to astrophysical properties



Our goal:



"V band"
"i band"
"z band"
"J band"
"H band"

etc.



(Stellar) Mass
Star formation rate
Star formation history
Distance/Redshift
Age

etc.

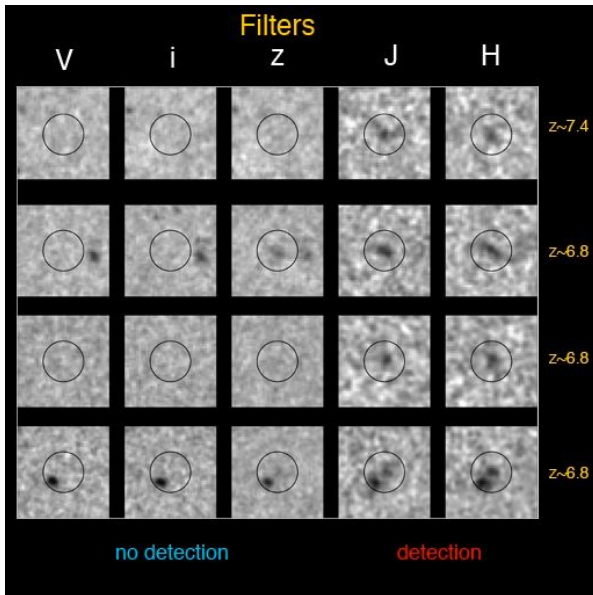


Two Fundamental Assumptions for Photometry

1. If an object is sufficiently well-measured, there is a surjective (one-to-one or many-to-one, but not one-to-many) mapping from photometric fluxes to astrophysical properties.
2. Objects with sufficiently similar photometry should be mapped to similar astrophysical properties.



Our goal:



"V band"
"i band"
"z band"
"J band"
"H band"

etc.



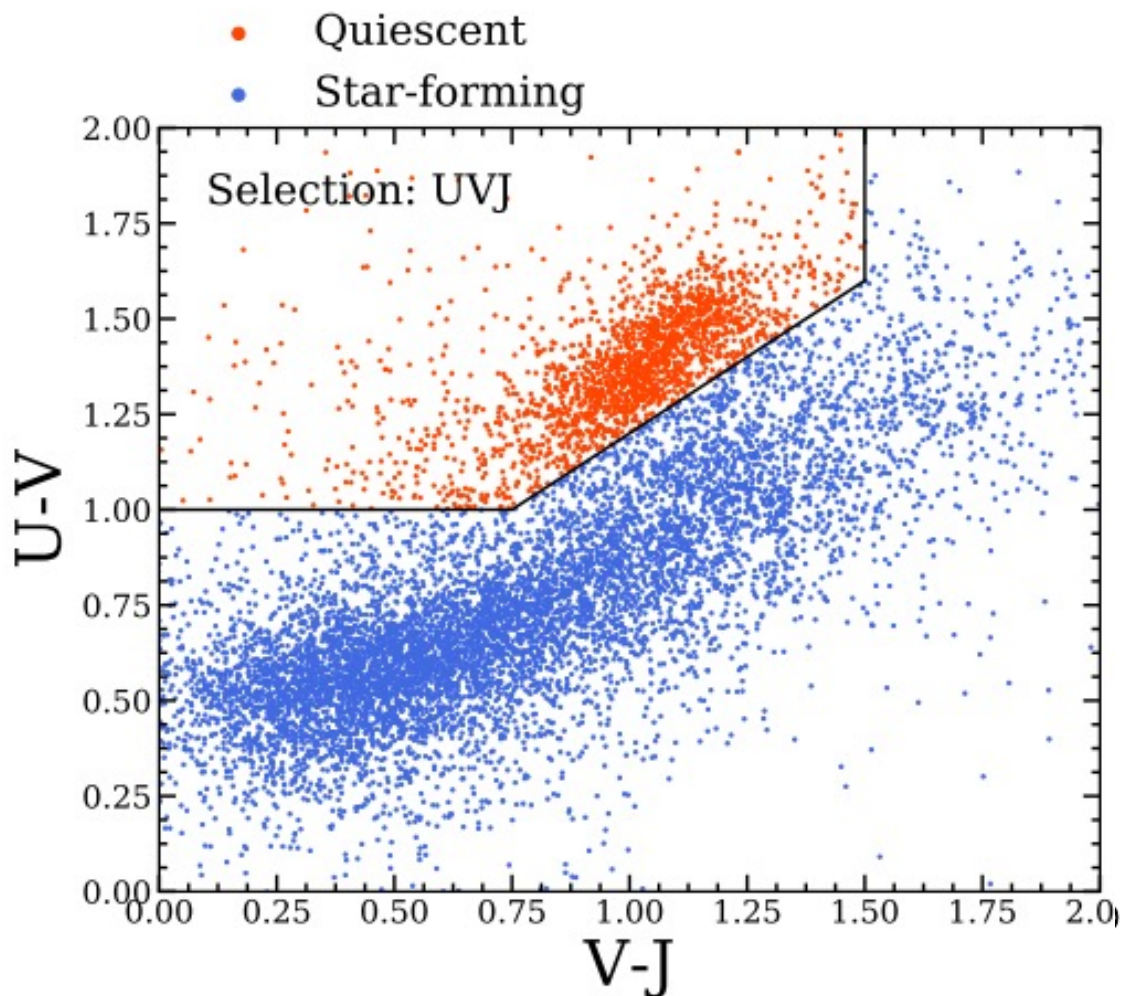
(Stellar) Mass
Star formation rate
Star formation history
Distance/Redshift
Age

etc.



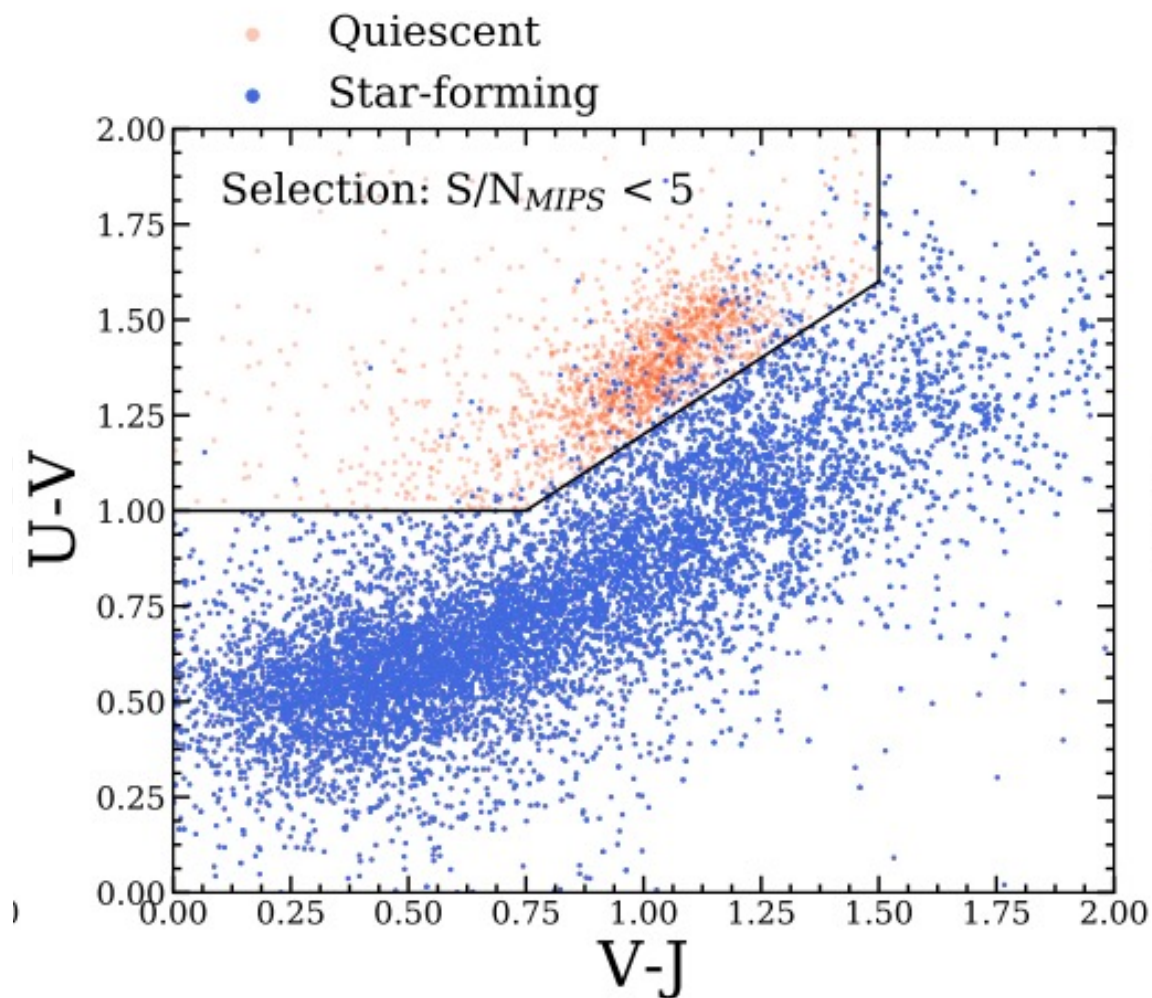
Approach 1: Color space map using two colors (three bands)

COSMOS2015 catalog, objects at $z \approx 1$



Approach 1: Color space map using two colors (three bands)

COSMOS2015 catalog, objects at $z \approx 1$



Approach 2: Color space map using all bands



~~Two~~ Three Fundamental Assumptions for Photometry

1. If an object is sufficiently well-measured, there is a surjective (one-to-one or many-to-one, but not one-to-many) mapping from photometric fluxes to astrophysical properties.
2. Objects with sufficiently similar photometry should be mapped to similar astrophysical properties.
3. We can map objects from the full, n-dimensional space with all bands to a smaller one with many neighbors, and the other two assumptions will continue to hold.



~~Two~~ Three Fundamental Assumptions for Photometry

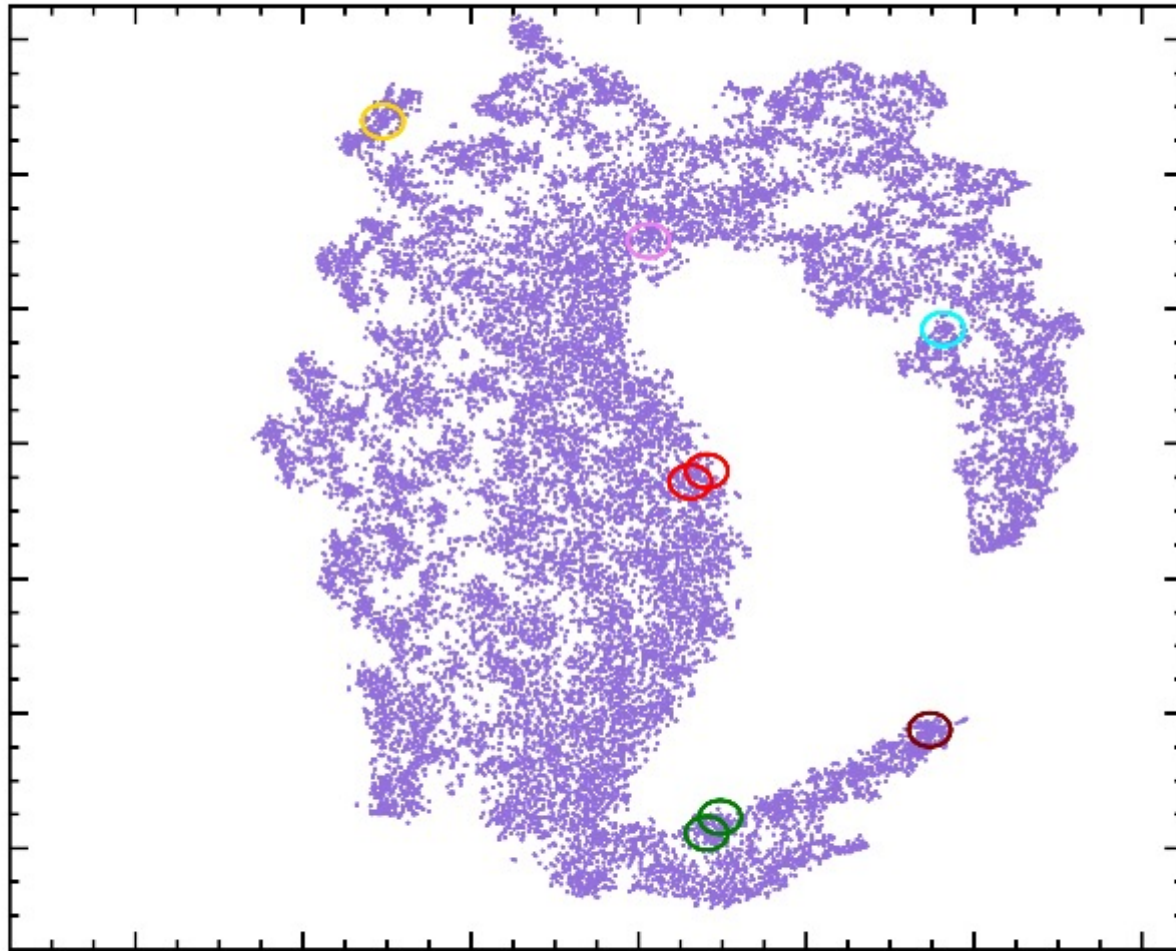
1. If an object is sufficiently well-measured, there is a surjective (one-to-one or many-to-one, but not one-to-many) mapping from photometric fluxes to astrophysical properties.
2. Objects with sufficiently similar photometry should be mapped to similar astrophysical properties.
3. We can map objects from the full, n-dimensional space with all bands to a smaller one with many neighbors, and the other two assumptions will continue to hold.

Can we somehow decide what information is “important” even without labels?



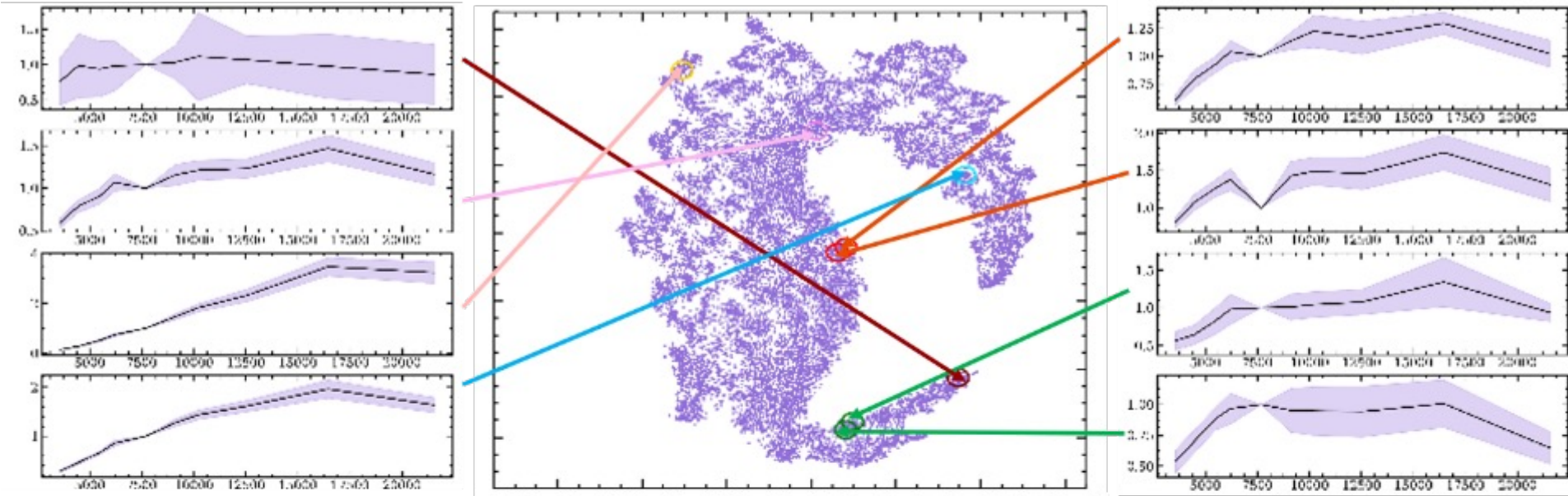
Approach 2: First, make a t-SNE map reducing to two dimensions

COSMOS2015 catalog, objects at $z \approx 1$



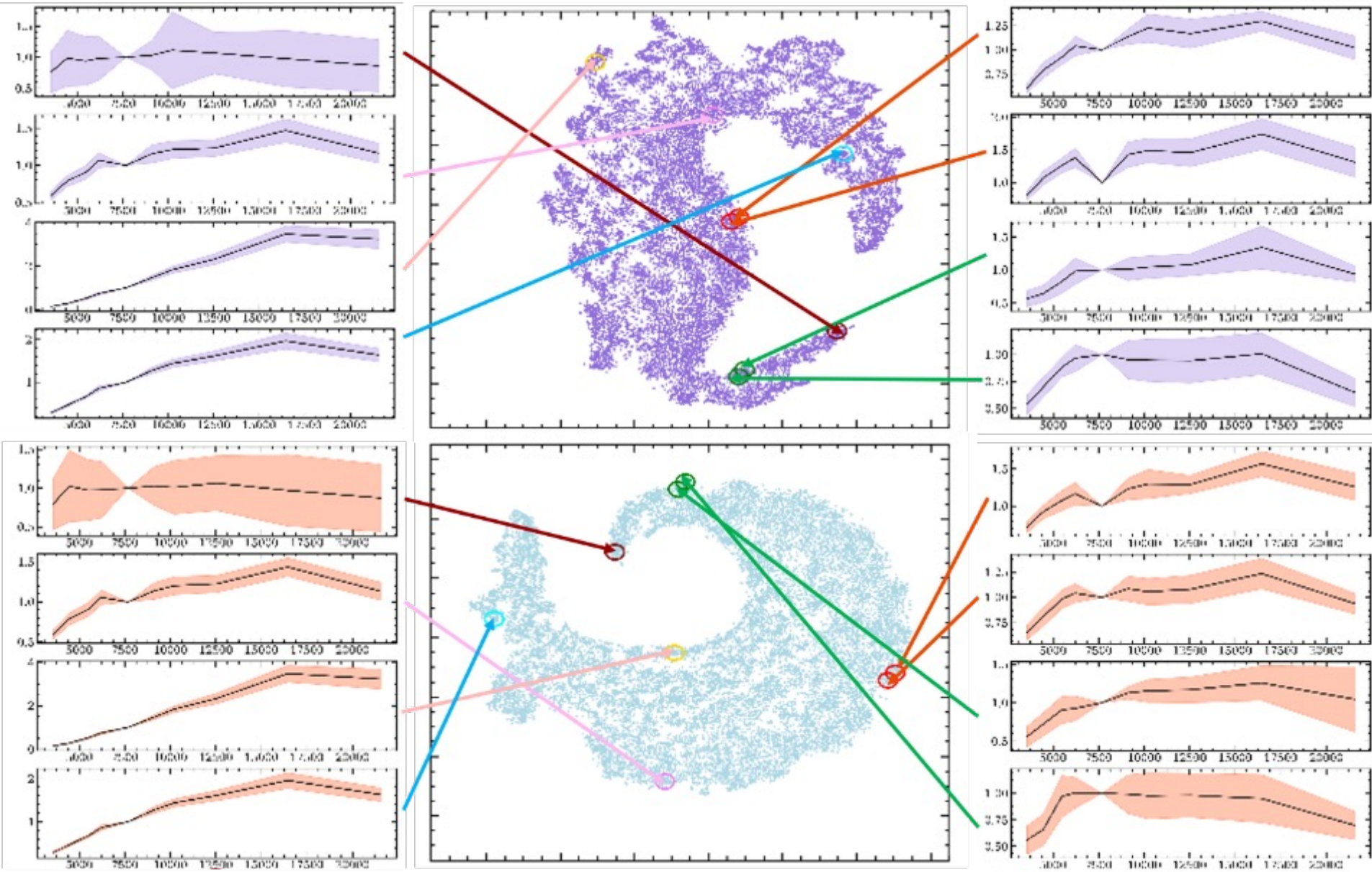
Approach 2: Similar galaxies are nearby

COSMOS2015 catalog, objects at $z \approx 1$



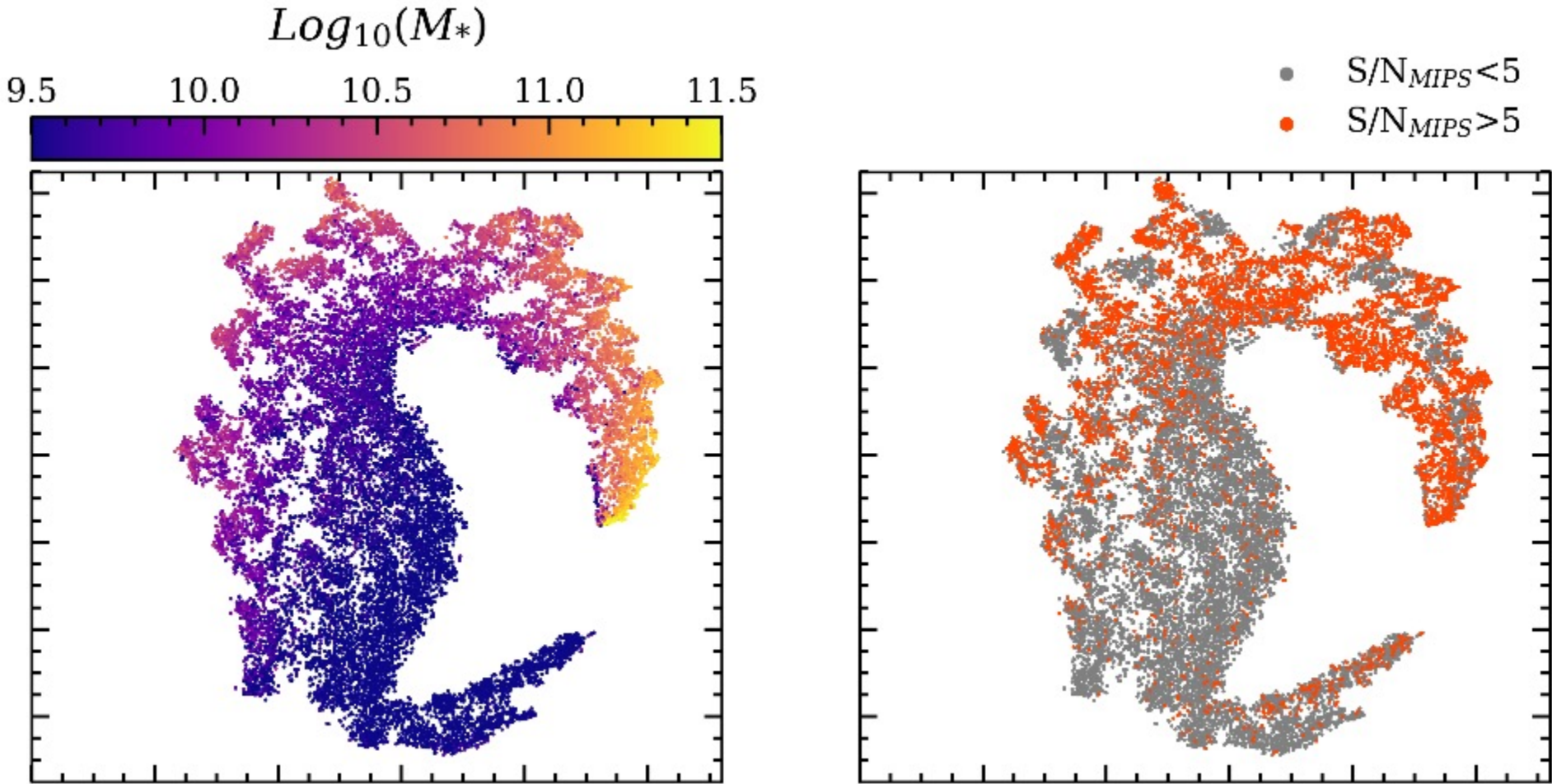
Approach 2: Similar galaxies are nearby

WARNING: positions are neither fixed nor meaningful. Topology is meaningful.



Approach 2: Arranging by photometry also calculates other useful things!

COSMOS2015 catalog, objects at $z \approx 1$



projector.tensorflow.org

Embedding Projector



DATA

Points: 10000 | Dimension: 200 | Selected 101 points

5 tensors found
Word2Vec 10K

Label by word
Color by No color map

Edit by word
Tag selection as

Load Publish Download Label

Spheroize data

Checkpoint: Demo datasets

Metadata: oss_data/word2vec_10000_200d_labels.tsv

UMAP T-SNE PCA CUSTOM

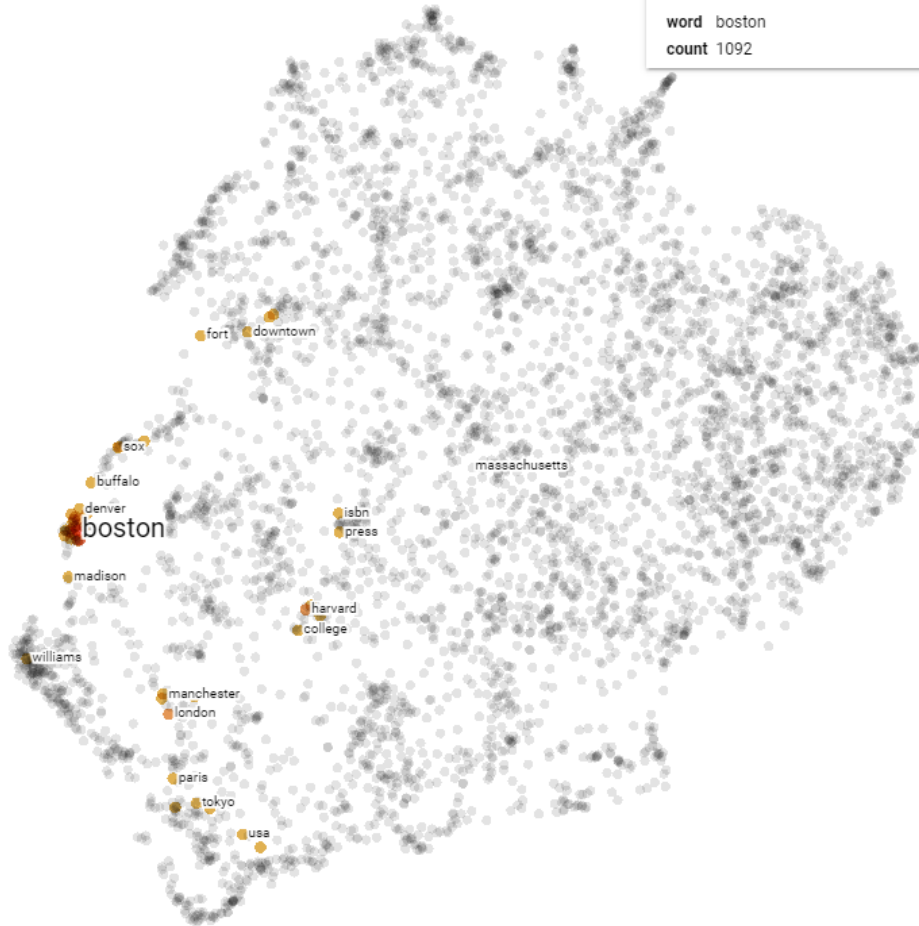
Dimension 2D 3D

Neighbors 15

Run

For faster results, the data will be sampled down to 5,000 points.

[Learn more about UMAP](#)



boston

word	boston
count	1092

Show All Data Isolate 101 points Clear selection

Search by word

neighbors 100

distance COSINE EUCLIDEAN

Nearest points in the original space:

chicago	0.406
massachusetts	0.406
philadelphia	0.413
atlanta	0.497
harvard	0.502
london	0.508
illinois	0.512
baltimore	0.514
maryland	0.546
york	0.551
cincinnati	0.554
toronto	0.556
seattle	0.569
brooklyn	0.578
miami	0.581
cambridge	0.583
pennsylvania	0.584
pittsburgh	0.587
detroit	0.593
california	0.595
sox	0.607
kansas	0.608

BOOKMARKS (0)