(Unsupervised) # Clustering

A very personal point of view

What do you cluster?

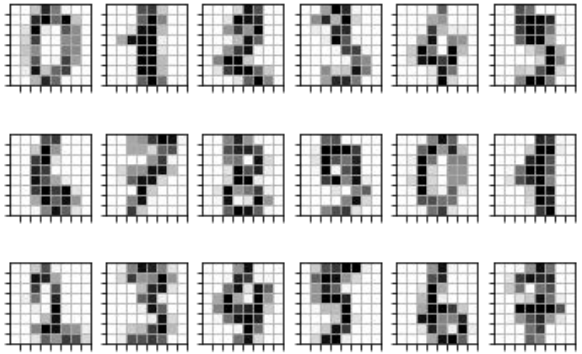
Questions:

- What is your question?
- How is your data?
- How would you measure distances? ...
  - Is it fair?

What do you cluster?

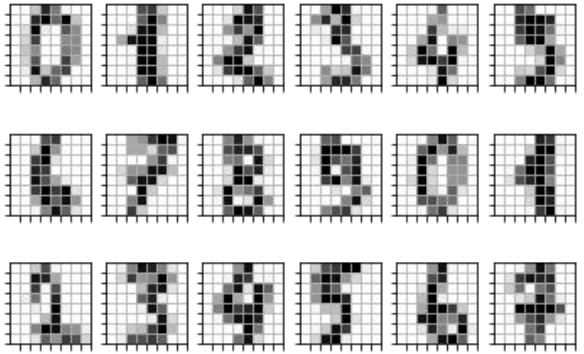
Questions:

- What is your question?
- How is your data?
- How would you measure distances?
  - Is it fair?



...

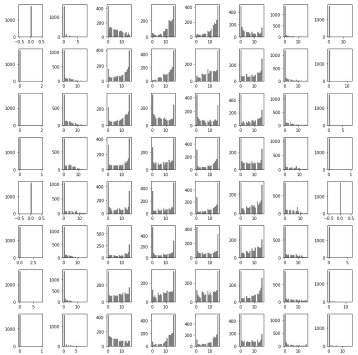
What do you cluster?

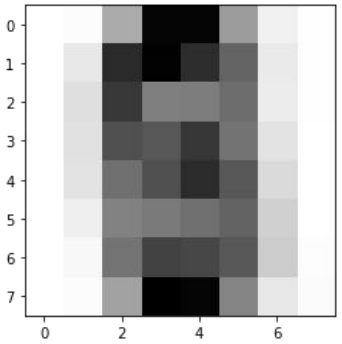


...

Questions:

- What is your question?
- How is your data?
- How would you measure distances?
  - Is it fair?





Solutions:

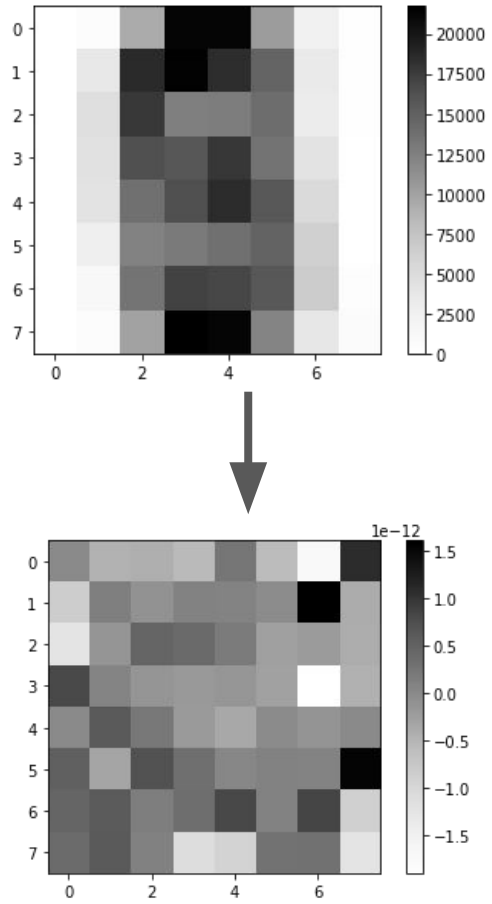
- Think
- Plot
- Think
- Plot

} Patience

Data massaging (rescue your outliers)

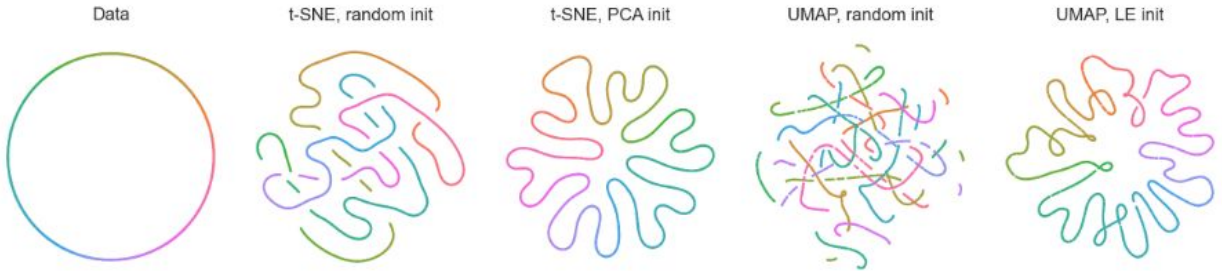
- Data transformation
  - min/max or percentiles 5/95
  - Log
  - Sigmoid (personal favorite)
- Normalization
  - Minus mean over standard-deviation
  - Something else (be creative…)

… try to keep it simple



Dimensionality reduction (and noise removal)

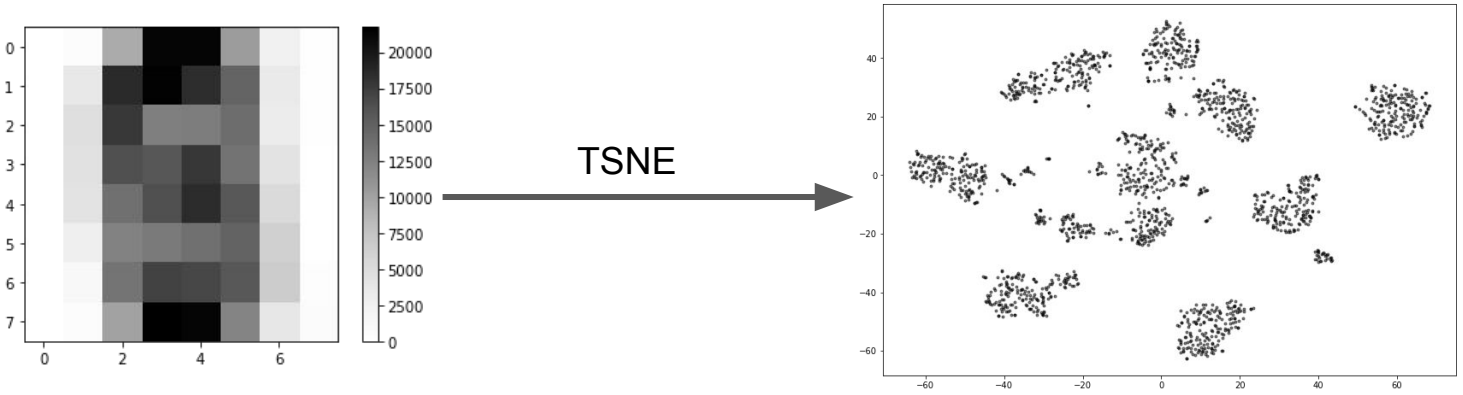
- PCA
- MDS
- t-SNE, **UMAP**

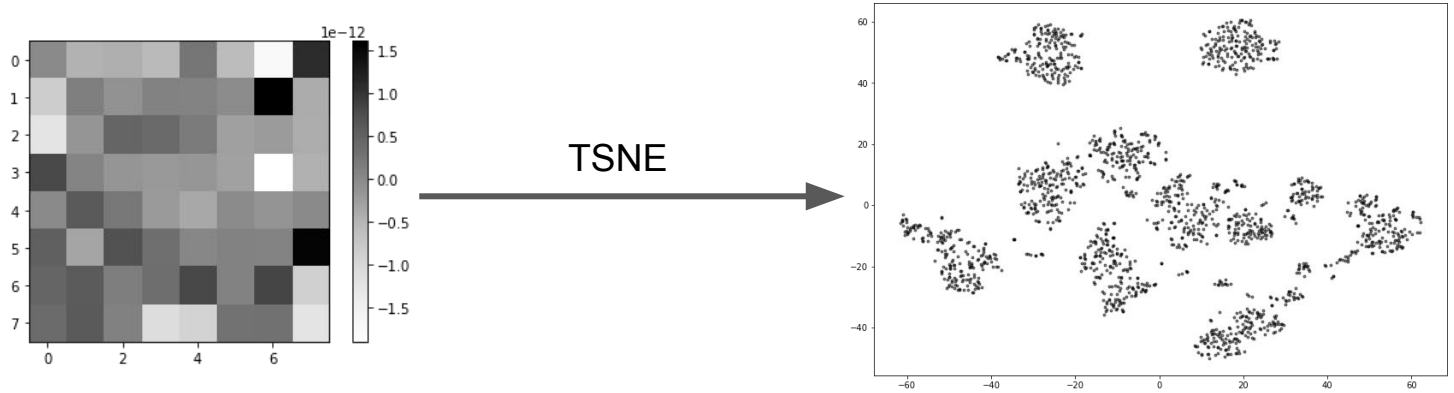


https://github.com/dkobak/tsne-umap-init

- Something else depending on the question…
  - SOM (https://en.wikipedia.org/wiki/Self-organizing_map)
  - sub-sampling

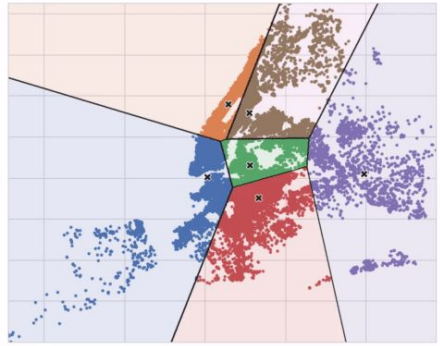
Dimensionality reduction (also a good way to explore)

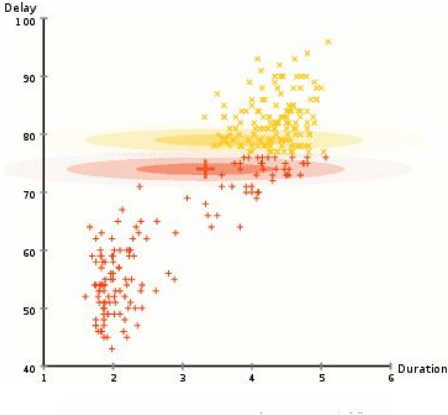




Clustering

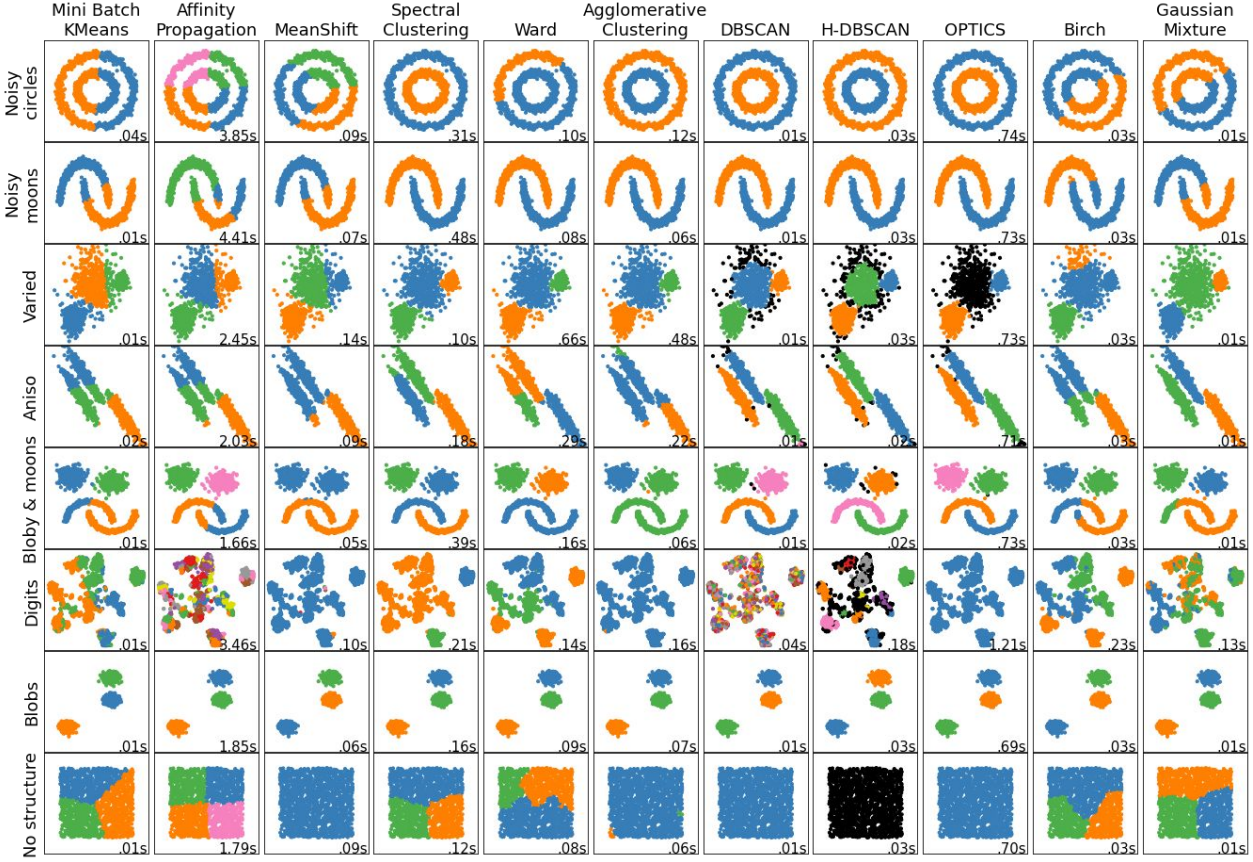
- Pick a type of clustering
  - Hierarchical (single, complete, **Ward**...)
    - Agglomerative
    - **Divisive**
  - Centroid/distribution/density based (K-means, GMM, **DBSCAN**...)
  - Something else
    - MCL (https://micans.org/mcl/)
    - **HDBSCAN** (https://hdbscan.readthedocs.io)





- Read the documentation

Clustering (nobody is perfect)



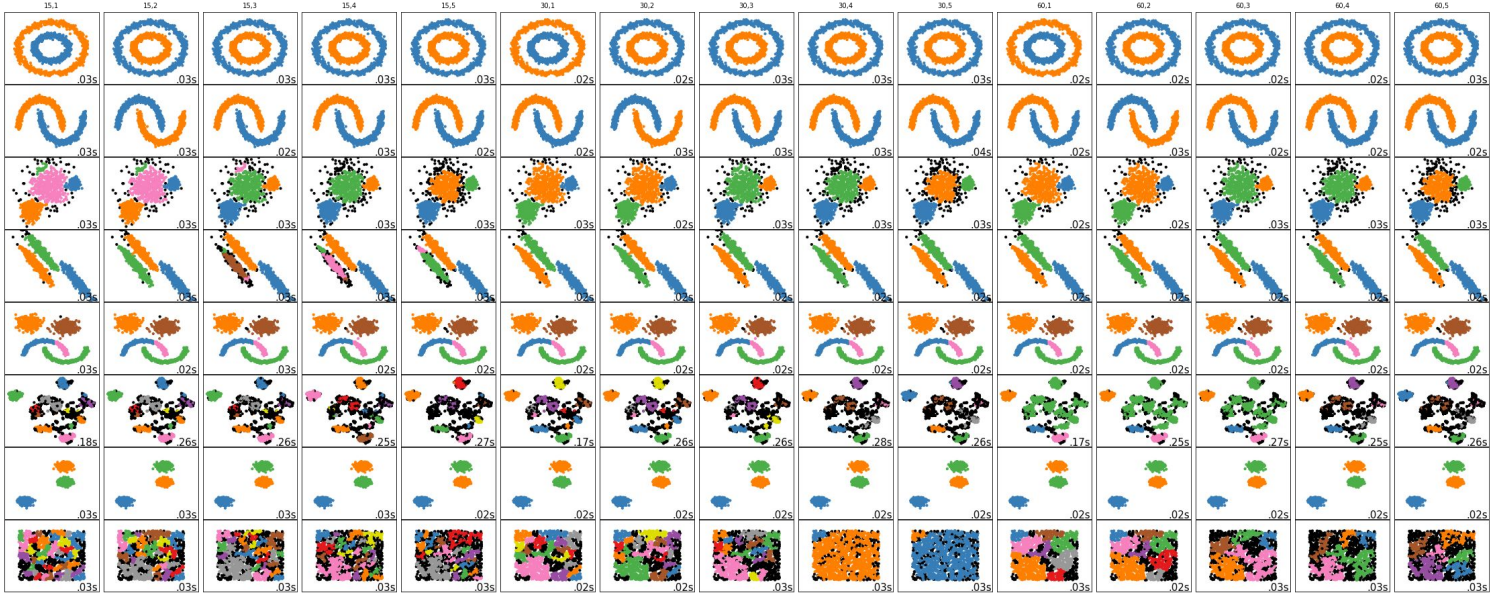
Blind evaluation

- Overall clustering quality
  - Silhouette
  - **Calinski-Harabasz**
  - **Davies-Bouldin**

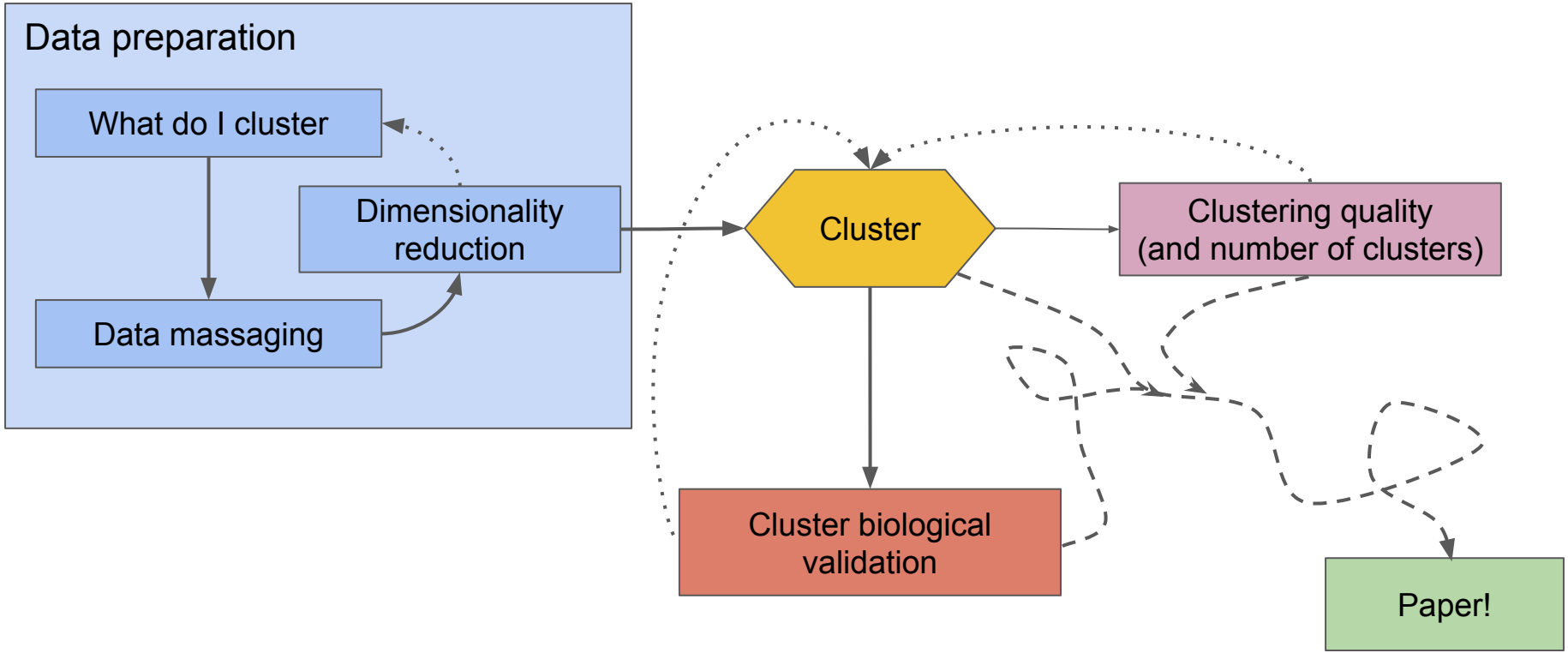
- Specific clustering quality
  - Robustness to bootstrapping (https://github.com/shimo-lab/pvclust)
  - Normal distribution of data at each cluster split (https://github.com/pkimes/sigclust2)

Try to supervise it a bit...

- Enrichment in specific biological function
- **Define a quality metric** (e.g. “how homogeneously active are genes in each cluster”)
- Test many combinations of parameters using this metric



Repeat...


Data preparation
What do I cluster
Dimensionality reduction
Data massaging
Cluster
Clustering quality (and number of clusters)
Cluster biological validation
Paper!

Links

- Scikit-learn summary: https://scikit-learn.org/stable/modules/clustering.html
- t-SNE vs UMAP: https://towardsdatascience.com/tsne-vs-umap-global-structure-4d8045acba17
- General course: https://liulab-dfci.github.io/bioinfo-combio/