Semi-supervised manifold learning and other dimension reduction approaches in single-cell data

Benjamin K. Johnson

Triche and Shen laboratories

Center for Epigenetics

Van Andel Research Institute

Grand Rapids, MI

Twitter: @biobenkj
Github: biobenkj.github.io/UMAP

Some additional useful resources and videos describing “why dimension reduction”, “intro to dimension reduction approaches”, and t-SNE and UMAP

Some community commented general resources for dimension reduction and what it’s good for:
https://twitter.com/biobenkj/status/1260652909097697280?s=20

Generally awesome intro to dimension reduction: https://www.youtube.com/watch?v=9iol3Lk6kyU

A PyData talk 2018 from Leland McInnes on UMAP: https://www.youtube.com/watch?v=YPJQydzTLwQ

More details and a great explanation comparing t-SNE and UMAP:
https://jlmelville.github.io/uwot/umap-for-tsne.html

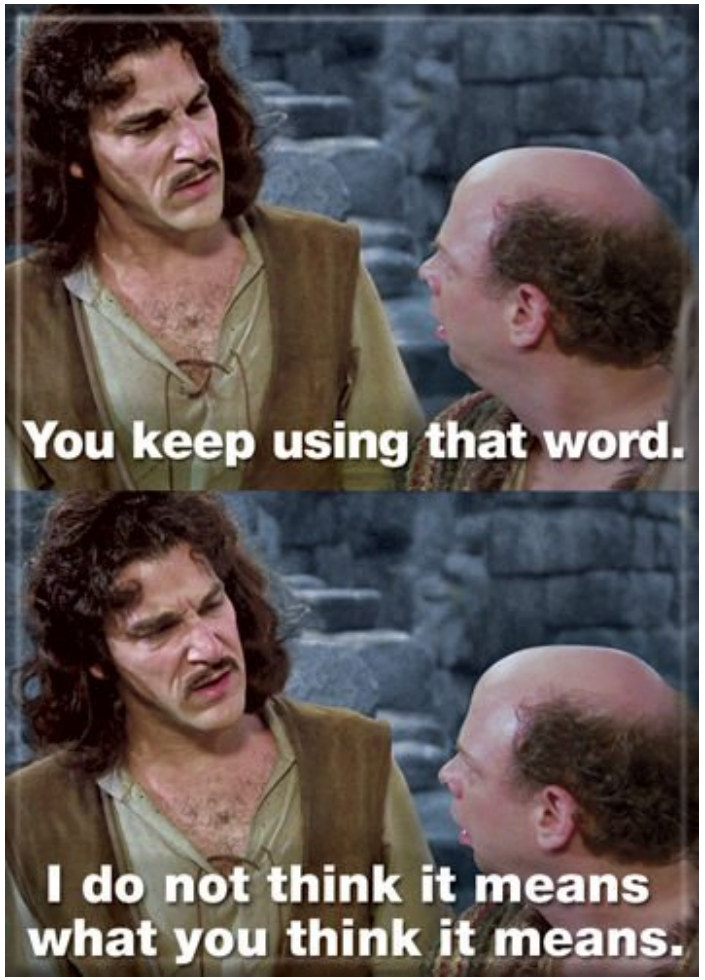
How to use t-SNE effectively (and interactively): https://distill.pub/2016/misread-tsne/

*Most of what’s in these slides was inspired and adapted from the above resources

Overview of semi-supervised Uniform Manifold Approximation and Projection (UMAP)

Why semi-supervised?

- Fully unsupervised: “I know nothing about the system”
- Semi-supervised: “I know some things about the system (e.g. I have a hypothesis)”
- Fully supervised: “I know everything about the system”



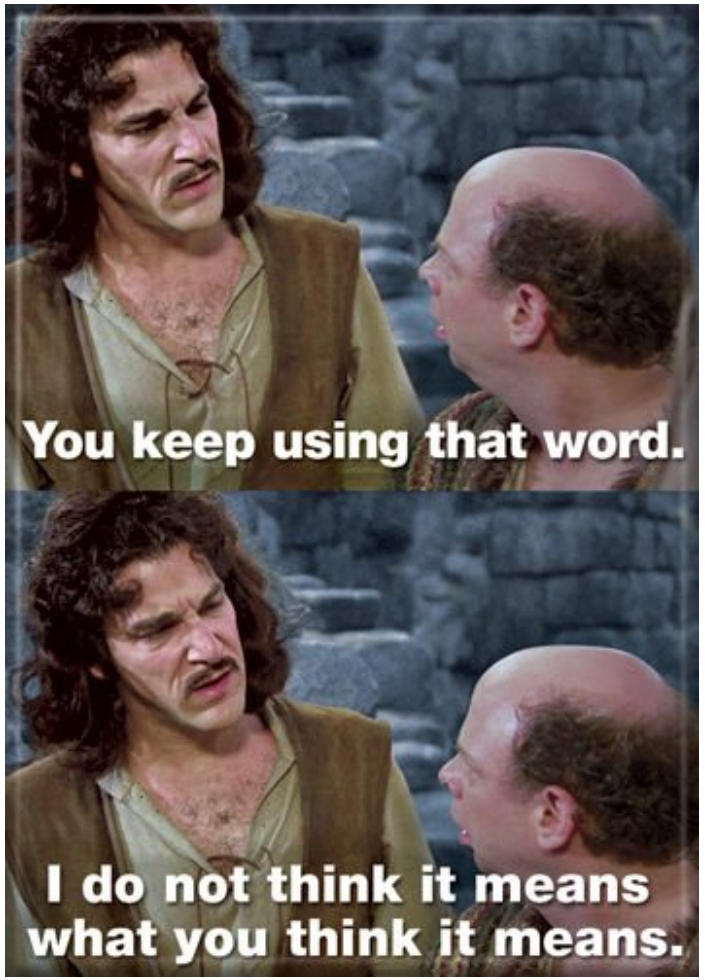
https://images.app.goo.gl/CPGzF733m2VwPCKz9

Overview of semi-supervised Uniform Manifold Approximation and Projection (UMAP)

Why semi-supervised?

- Fully unsupervised: "I know nothing about the system"
- **Semi-supervised: "I know some things about the system (e.g. I have a hypothesis)"**
- Fully supervised: "I know everything about the system"

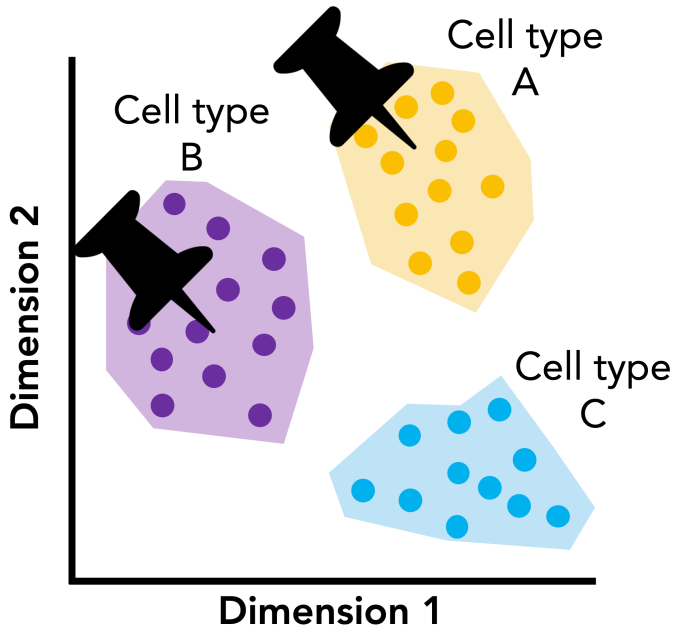
We usually have some knowledge of the system that we want to incorporate (e.g. cell types, treatments, controls, etc.) which is an inherently semi-supervised approach.



https://images.app.goo.gl/CPGzF733m2VwPCKz9

Overview of semi-supervised Uniform Manifold Approximation and Projection (UMAP)

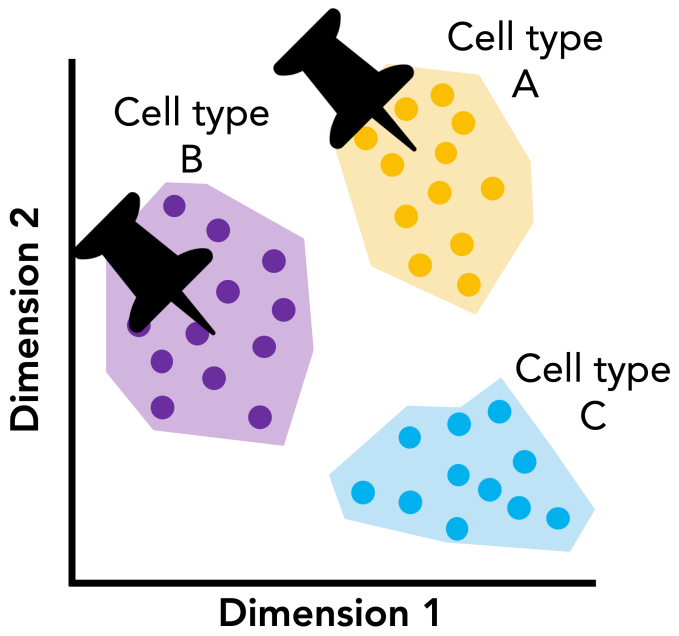
How does it work?



We "pin" or "label" cells together that we know *a priori* should stay together (e.g. control treated cells). We leave our experimental cells "unlabeled" to see where they end up amongst the data in the embedding

Overview of semi-supervised Uniform Manifold Approximation and Projection (UMAP)

How does it work?



We "pin" or "label" cells together that we know *a priori* should stay together (e.g. control treated cells). We leave our experimental cells "unlabeled" to see where they end up amongst the data in the embedding

Semi-supervised* UMAP

High dimensional data

Partial data labels (e.g. cell type)

Construct fuzzy simplicial sets

$$\bigcup_{i=1}^{n} FinSing((X, d_i))$$

Intersect the fuzzy simplicial sets

Reset local connectivity of data and reconstruct fuzzy simplicial set

Initialize coordinates (e.g. spectral embedding)

UMAP

$$C_{UMAP} = \sum_{ij}\left[v_{ij}log\left(\frac{v_{ij}}{w_{ij}}\right) + (1-v_{ij})log\left(\frac{1-v_{ij}}{1-w_{ij}}\right)\right]$$

Get clumps right

Get gaps right

Semi-supervised

Embed

Dimension Reduction

Neat, but what about t-distributed Stochastic Neighbor Embedding (t-SNE)??

What do we know about UMAP's cost function?

$$C_{UMAP} = \sum_{ij}\left[v_{ij}log\left(\frac{v_{ij}}{w_{ij}}\right) + (1-v_{ij})log\left(\frac{1-v_{ij}}{1-w_{ij}}\right)\right]$$

Get clumps right

Get gaps right

What does t-SNE's cost function look like?

$$C_{tSNE} = \sum_{i}\sum_{j} p_{ij}\, log\frac{p_{ij}}{w_{ij}}$$

Attractive force

Repulsive force

These two terms look awfully similar (like KL divergence)
and both UMAP and t-SNE can be thought of as force directed (push and pull points)
graph-based approaches

Perhaps one way to loosely think about them is that t-SNE is kind of like a special case of UMAP,
but differ in the way that they handle the global repulsive force
depending on the data points ("get the gaps right") versus the uniform repulsion that t-SNE has.

So which one should I use?? In many cases, UMAP is ideal

UMAP

Can embed **new** data into the existing model (or your data into publicly available data)

Computationally more tractable as data points increase (see Human cell atlas project)

Less sensitive to initialization and produces more stable solutions

t-SNE

Cannot readily embed new data into the existing model

Doesn't quite scale as well (inherently $O(N^2)$ with pairwise distance calcs)

More sensitive to hyperparameter tuning (e.g. perplexity)

So which one should I use?? In many cases, UMAP is ideal

UMAP

Can embed **new** data

Human cell atlas project)

Less sensitive to initialization and produces more stable solutions

t-SNE

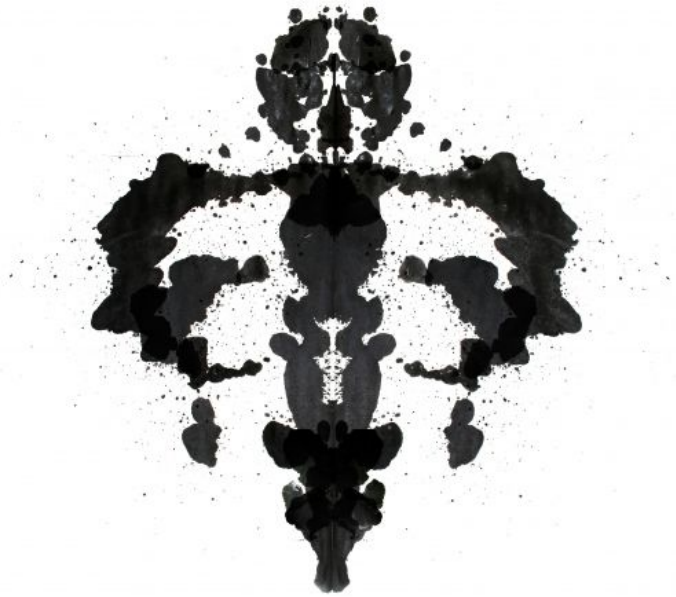
Cannot readily embed

pairwise distance calcs)

More sensitive to hyperparameter tuning (e.g. perplexity)

There are extensions of t-SNE that help remediate many of these "cons" and both can be tuned to produce similar results

Word of caution though, t-SNE and UMAP being non-linear dimension reduction techniques may give the data an appearance of structure that may not be there



Our eyes are *very* good at finding patterns, whether or not they mean anything in the context of the data

https://images.app.goo.gl/FkHbTH2sKU4WB8WF8