The Meaning of Uniform Manifold Approximation and Projection (UMAP)

Nick Lines

JOHNS HOPKINS UNIVERSITY

https://github.com/linesn/the_meaning_of_umap

What is UMAP?

- Non-linear dimensionality reduction
- Invented in 2018 by
  Leland McInnes
  John Healy
  James Melville

  (Tutte Institute) [1]

UNIFORM MANIFOLD
UMAP
APPROXIMATION & PROJECTION

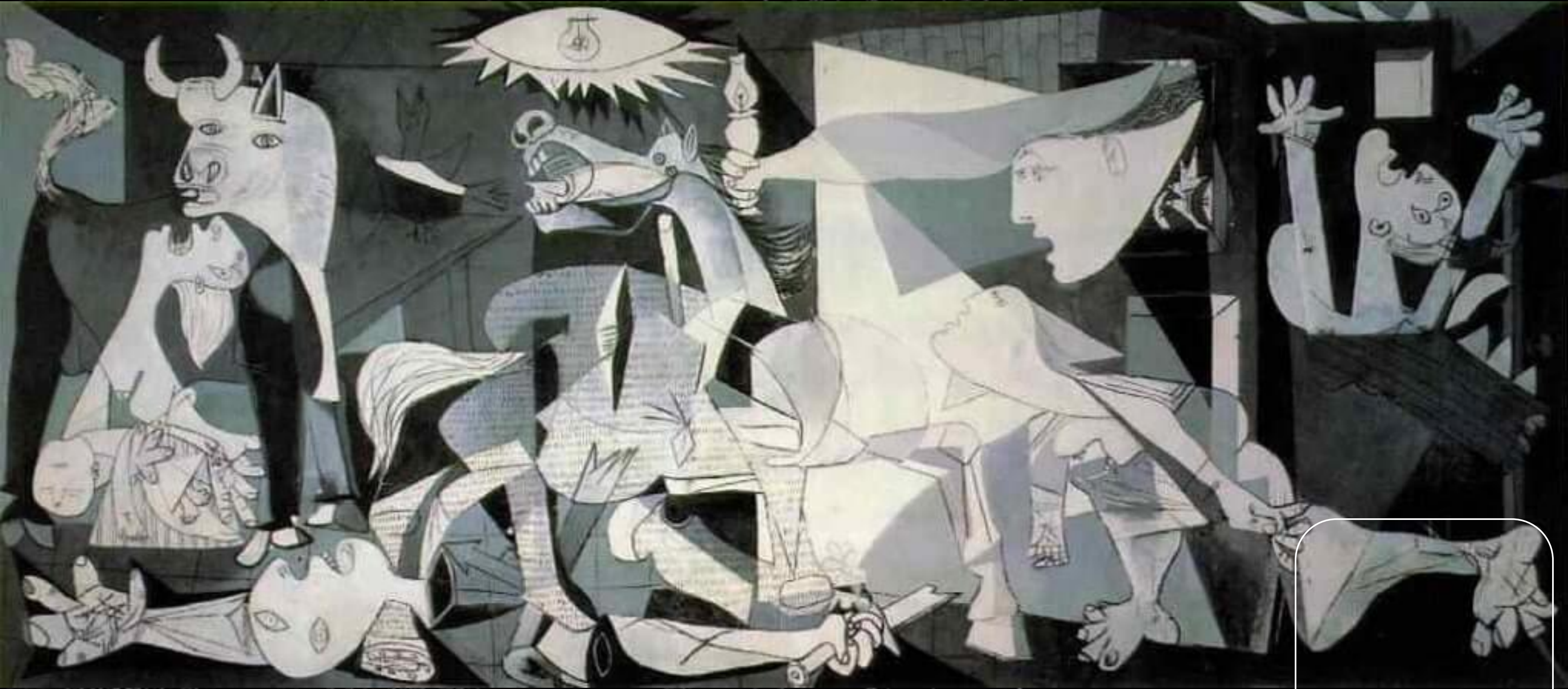
Outline

1. What does UMAP do?
2. How does it do that?
3. Practical considerations
4. Conclusion

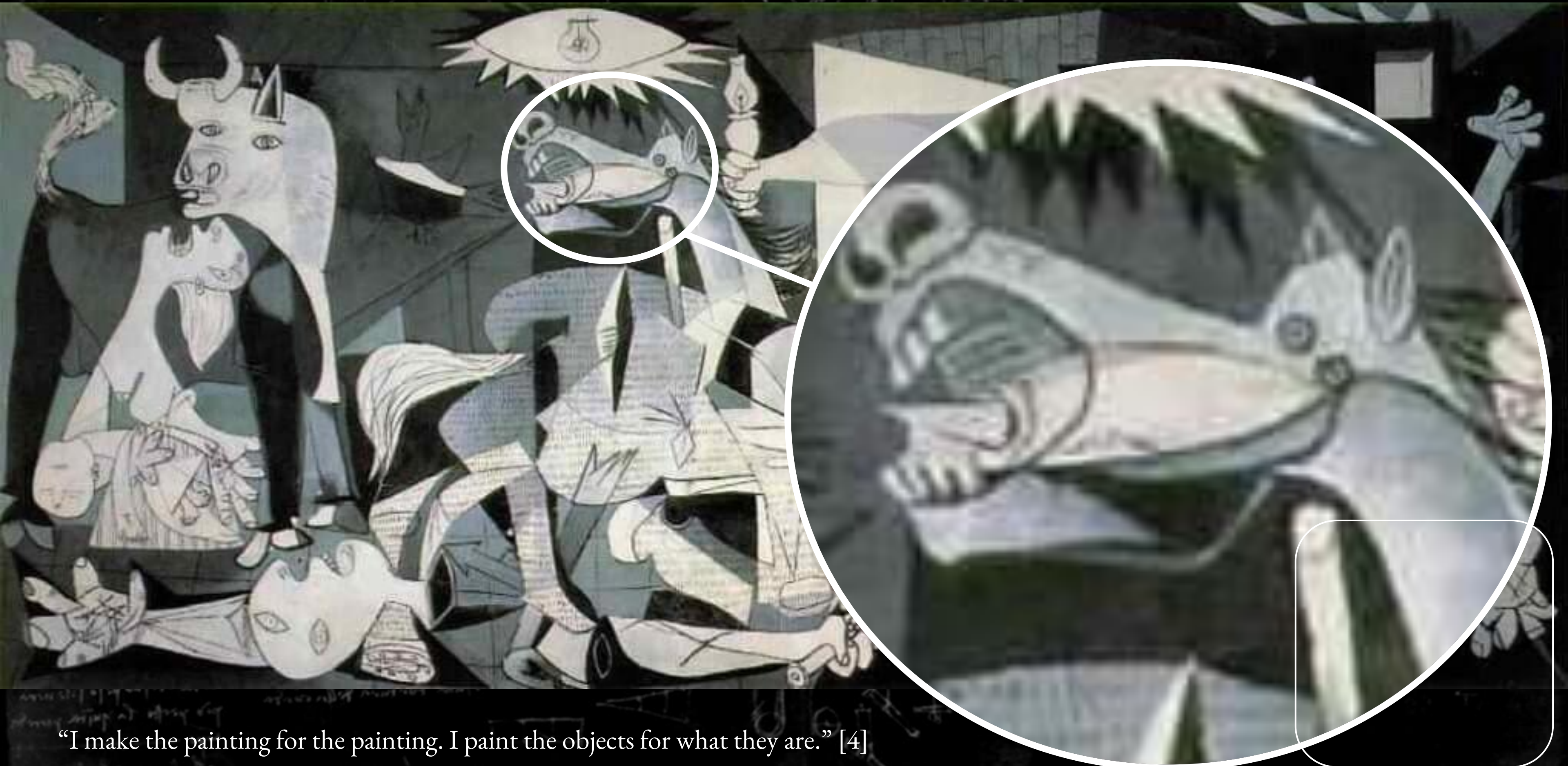
Remember: Nothing I’ll say here is original!

1. What does UMAP do?

Guernica, by Picasso

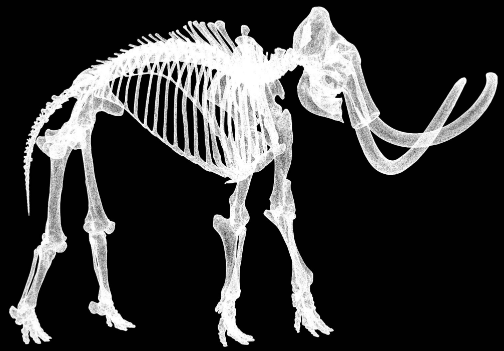


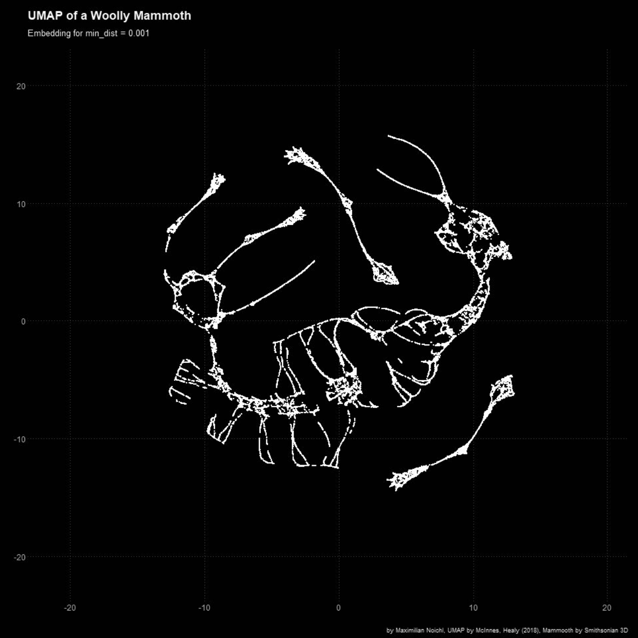
Guernica, by Picasso



“I make the painting for the painting. I paint the objects for what they are.” [4]

Examples of UMAP Dimensionality Reduction applied to Smithsonian 3D models [6]

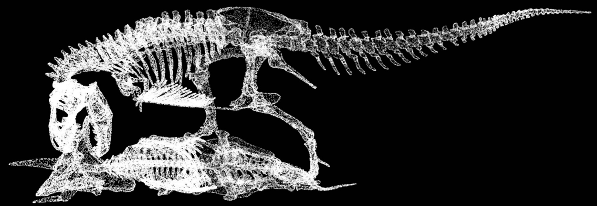


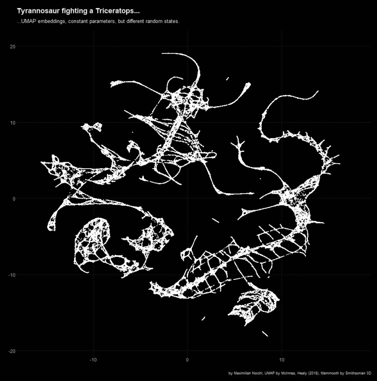


UMAP of a Woolly Mammoth

Embedding for min_dist = 0.001

by Maximilian Noichl, UMAP by McInnes, Healy (2018), Mammooth by Smithsonian 3D



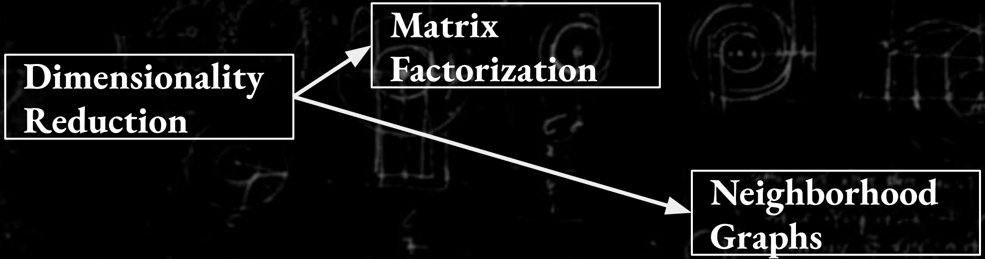


Tyrannosaur fighting a Triceratops...

...UMAP embeddings, constant parameters, but different random states.

by Maximilian Noichl, UMAP by McInnes, Healy (2018), Mammooth by Smithsonian 3D

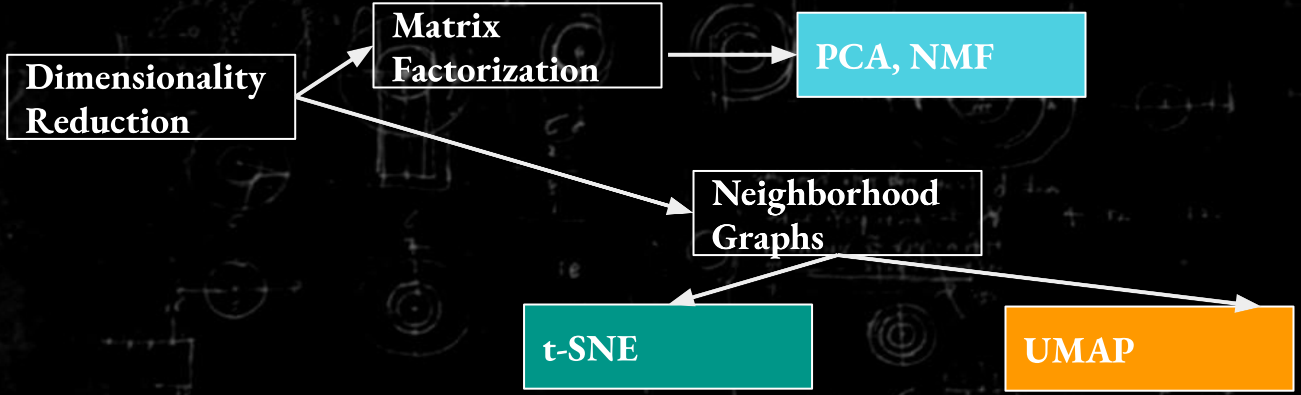
2. How does UMAP do that?



Dimensionality Reduction

Matrix Factorization

Neighborhood Graphs



Dimensionality Reduction

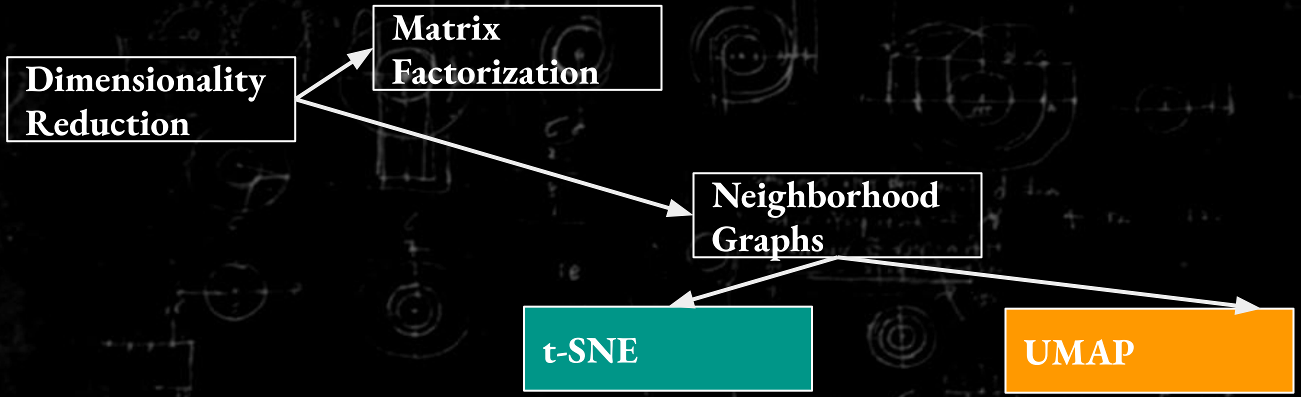
Matrix Factorization

PCA, NMF

Neighborhood Graphs

t-SNE

UMAP



Dimensionality Reduction

Matrix Factorization

Neighborhood Graphs

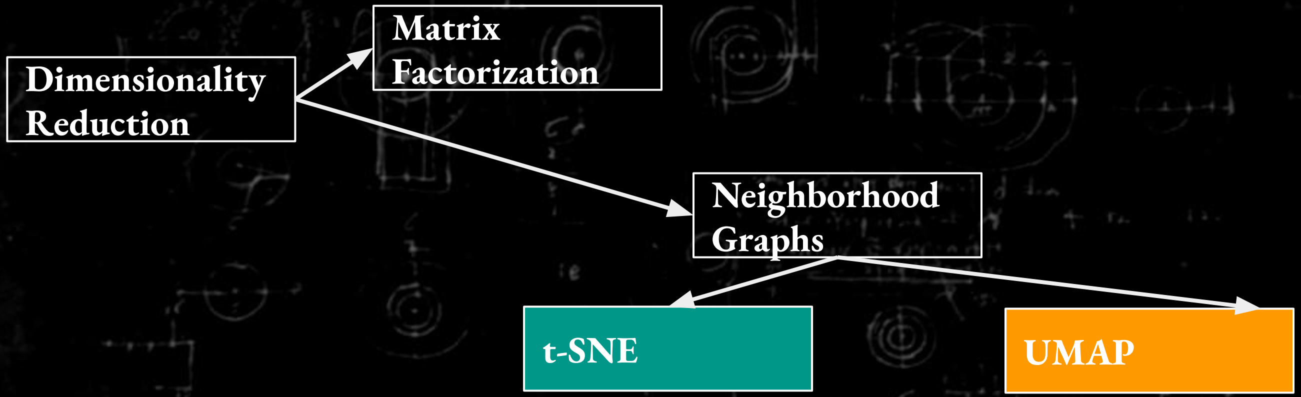
t-SNE

UMAP

Pairwise Distances

Convert to Probability Dists.

Minimize Kullback-Leibler with Gradient Descent



Dimensionality Reduction

Matrix Factorization

Neighborhood Graphs

t-SNE

UMAP

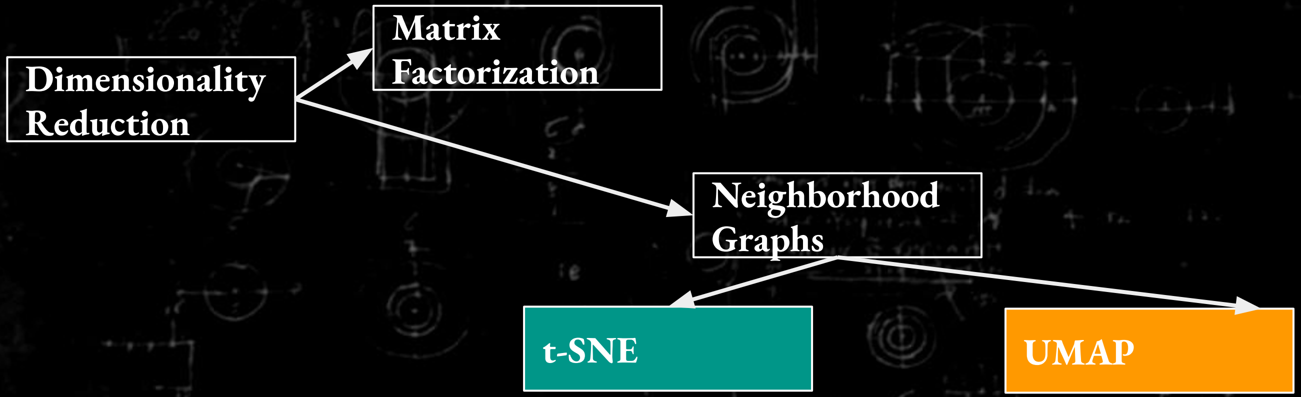
Find the topological structure in high dimension representation

Pairwise Distances

Convert to Probability Dists.

Initialize then optimize a lower dimensional representation

Minimize Kullback-Leibler with Gradient Descent



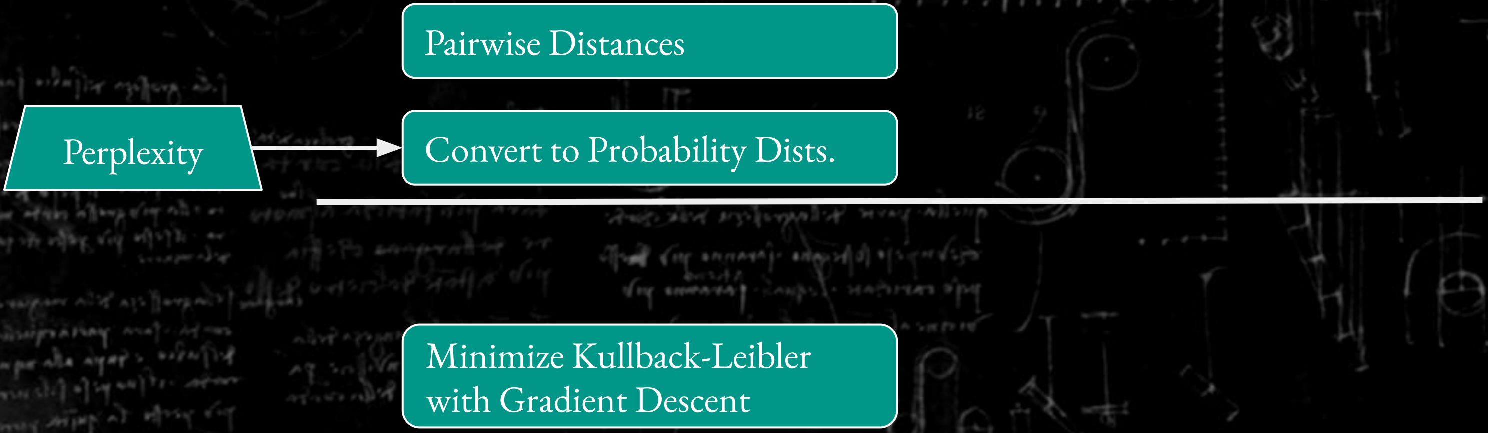
Dimensionality Reduction

Matrix Factorization

Neighborhood Graphs

t-SNE

UMAP

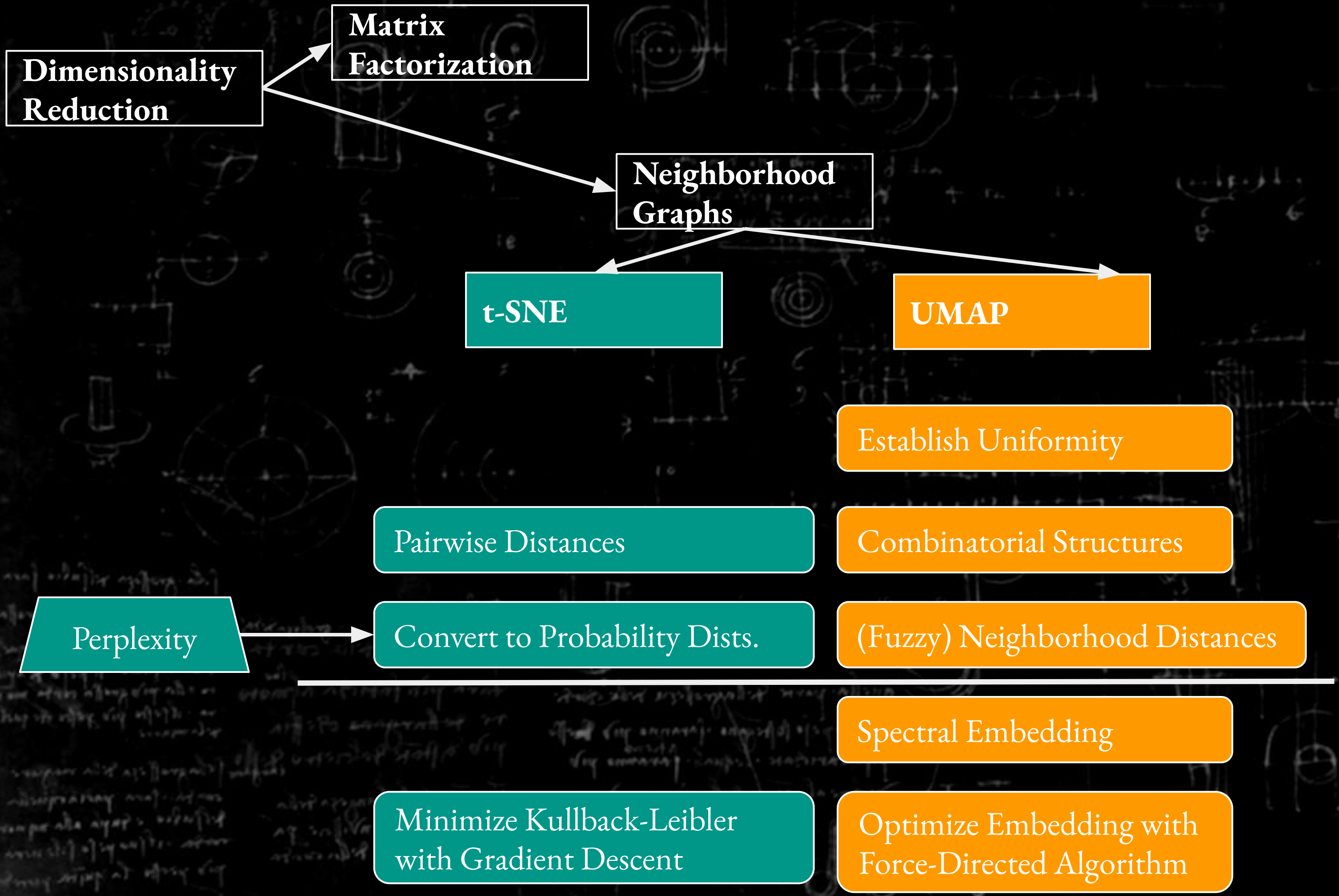


Pairwise Distances

Perplexity

Convert to Probability Dists.

Minimize Kullback-Leibler with Gradient Descent



Dimensionality Reduction

Matrix Factorization

Neighborhood Graphs

t-SNE

UMAP

Establish Uniformity

Pairwise Distances

Combinatorial Structures

Perplexity

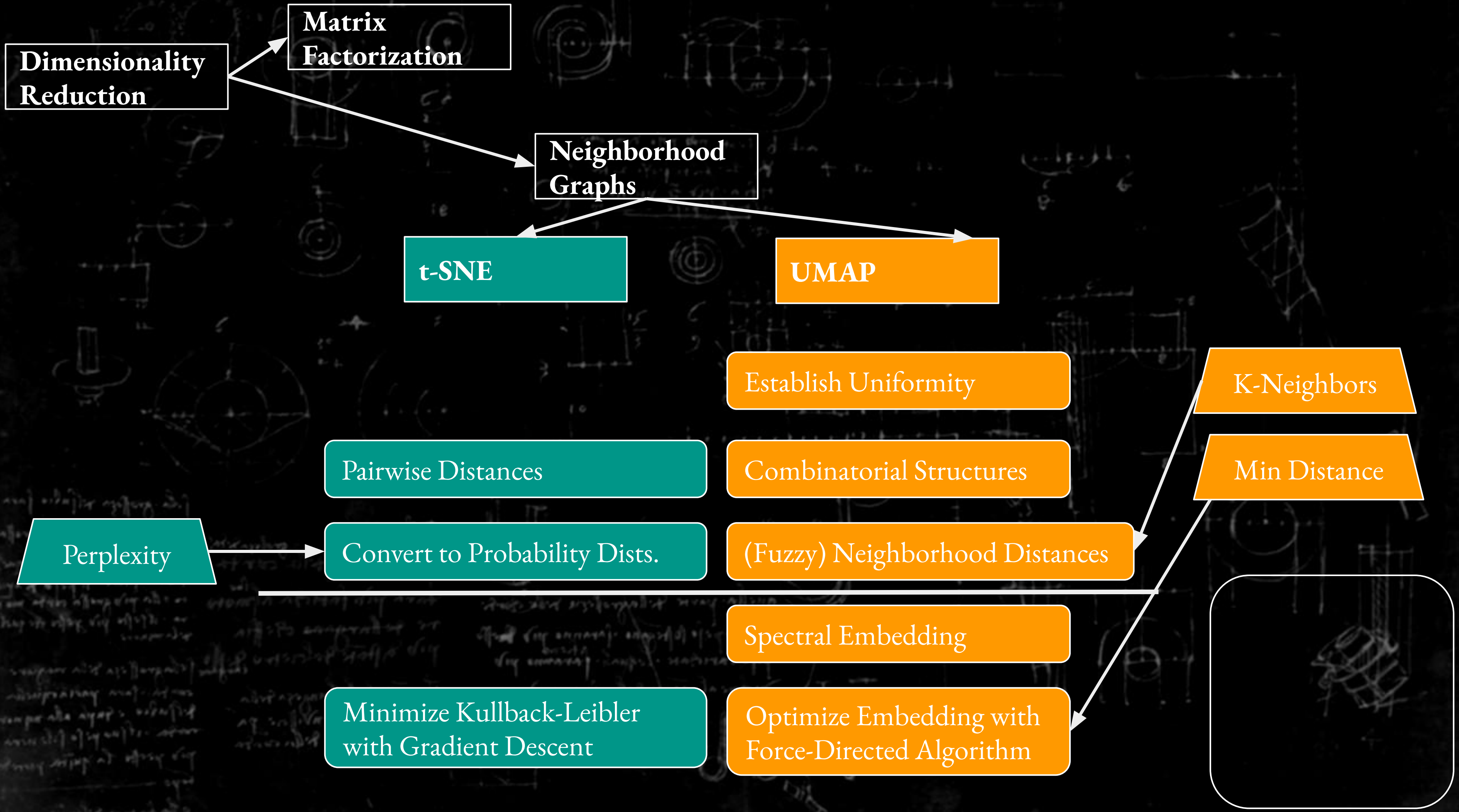
Convert to Probability Dists.

(Fuzzy) Neighborhood Distances

Spectral Embedding

Minimize Kullback-Leibler with Gradient Descent

Optimize Embedding with Force-Directed Algorithm



Dimensionality Reduction

Matrix Factorization

Neighborhood Graphs

t-SNE

UMAP

Establish Uniformity

K-Neighbors

Pairwise Distances

Combinatorial Structures

Min Distance

Perplexity

Convert to Probability Dists.

(Fuzzy) Neighborhood Distances

Spectral Embedding

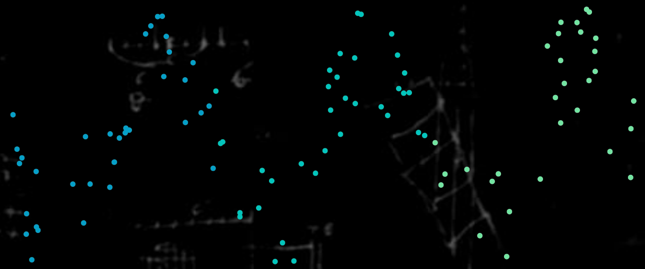
Minimize Kullback-Leibler with Gradient Descent

Optimize Embedding with Force-Directed Algorithm

Justify the assumption “uniformly distributed” data

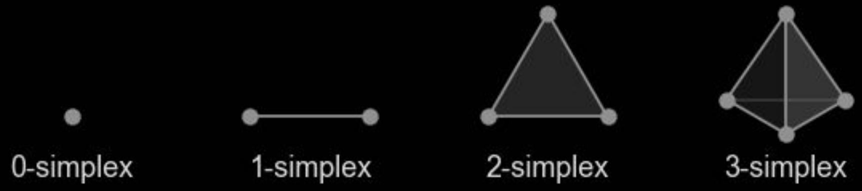
(Builds on David Spivak’s work [5])

$\boldsymbol{x}=x_1, x_2, \dots , x_r$



- To form consistent local neighborhoods that define graph structure, we want uniformly distributed data.
- NOT true by default with finite samples in a high dimension manifold
- So establish a distance function d_i for each x_i so that the unit ball centered at x_i contains the nearest k data points.

Switch to convenient combinatorial structures



[2]

- A *k-simplex* is a structure formed by *k+1* (independent) points in *k* dimensions.
- A *simplicial complex* is a set *K* of simplices such that each face of each simplex in *K* is in *K*, and the (nonempty) intersection of any two simplicies in *K* is a shared face.

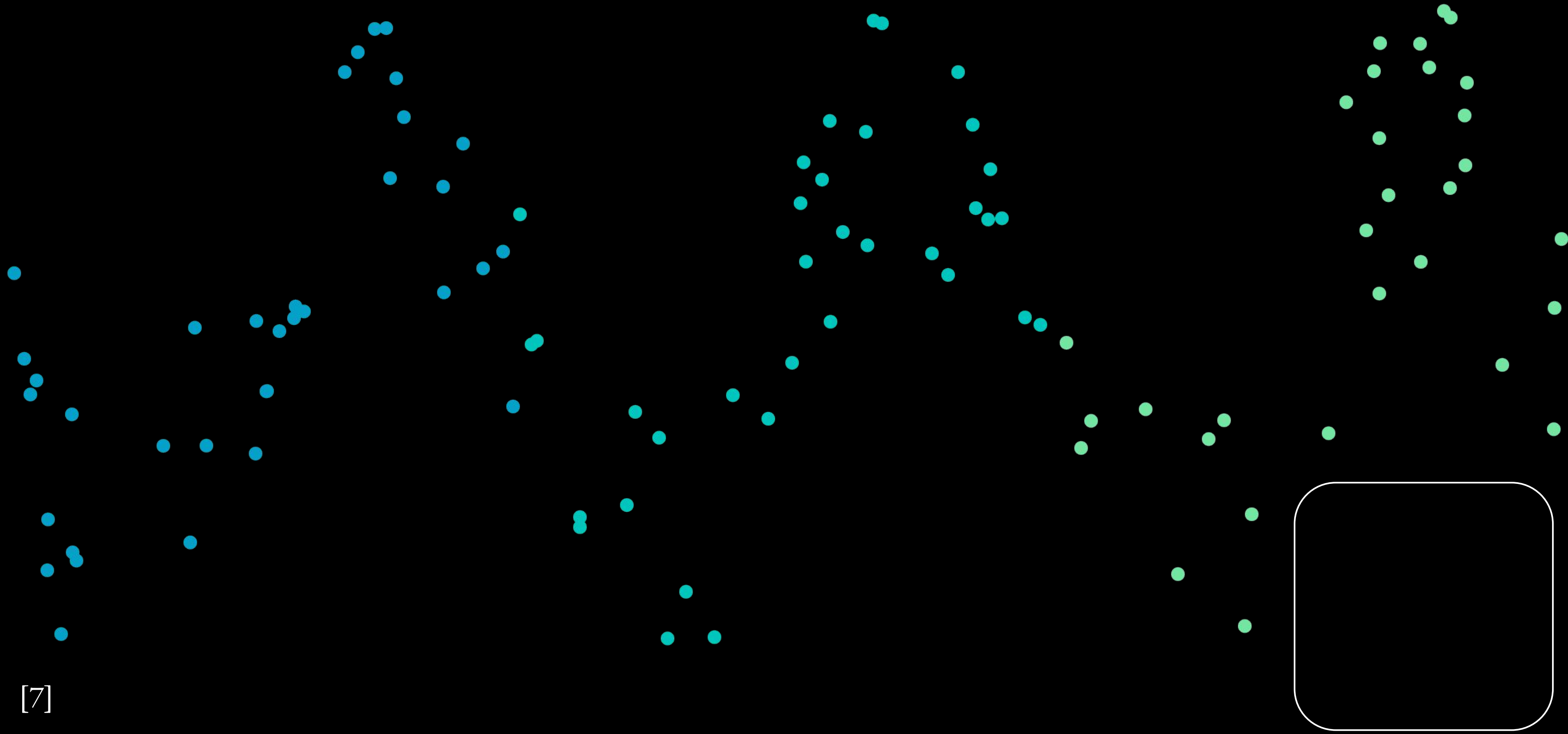
- So we build a simplicial complex called the Čech complex by describing each data point as a 0-simplex. Then we connect each point to other points within the unit ball around it, and build 1-simplices, then 2-simplices, etc. until all points are part of the complex.

- The **Nerve Theorem** guarantees that this preserves the topological structure of the data!

Fuzzy neighborhood distances

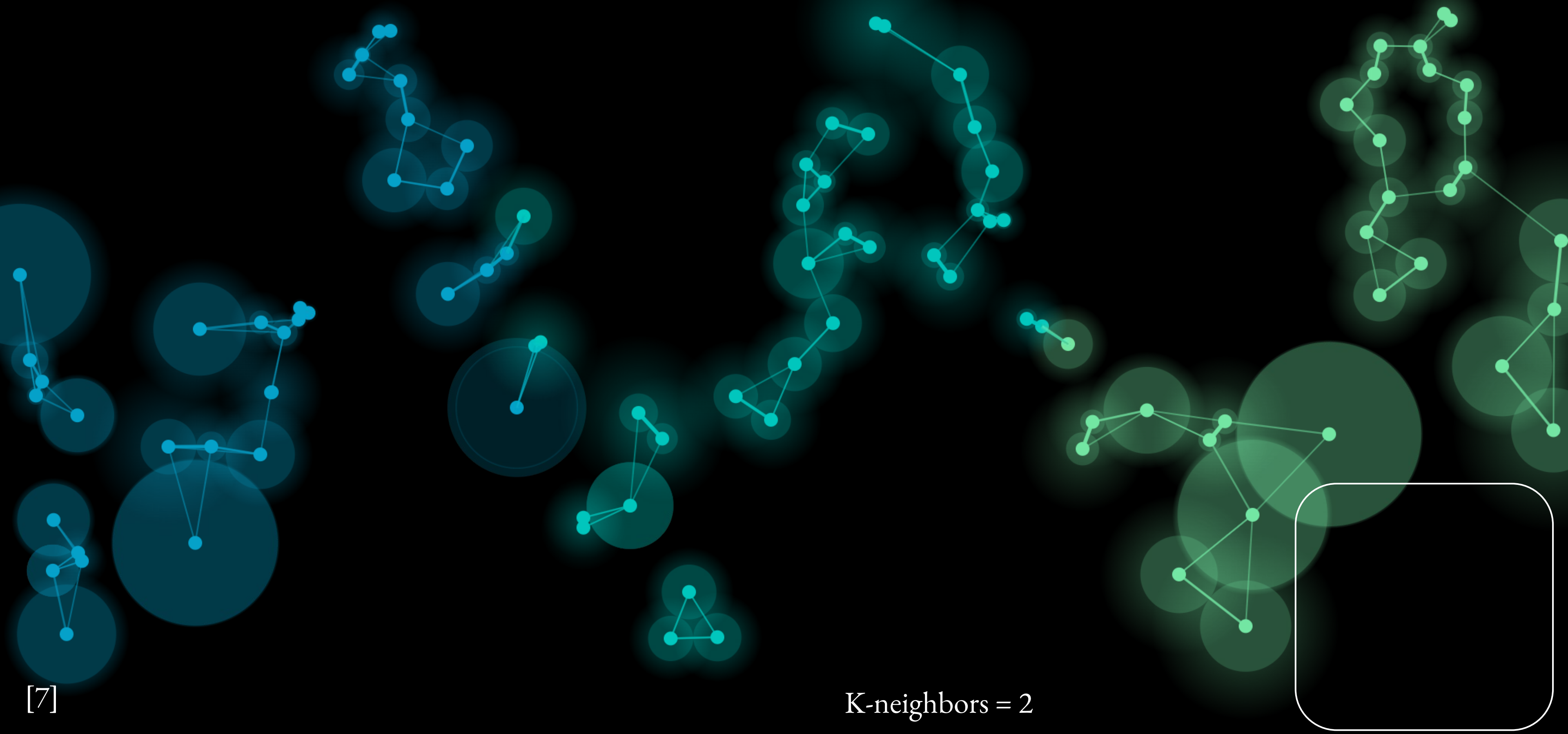
- Probability of an edge x_i to x_j in the sense of d_i is a function of the distance $d_i(x_i, x_j)$. More distant points are less likely to be connected.
- Probability that x_i and x_j share an edge is the probability at least one of (x_i to x_j and x_j to x_i) exists.

Fuzzy neighborhood distances



[7]

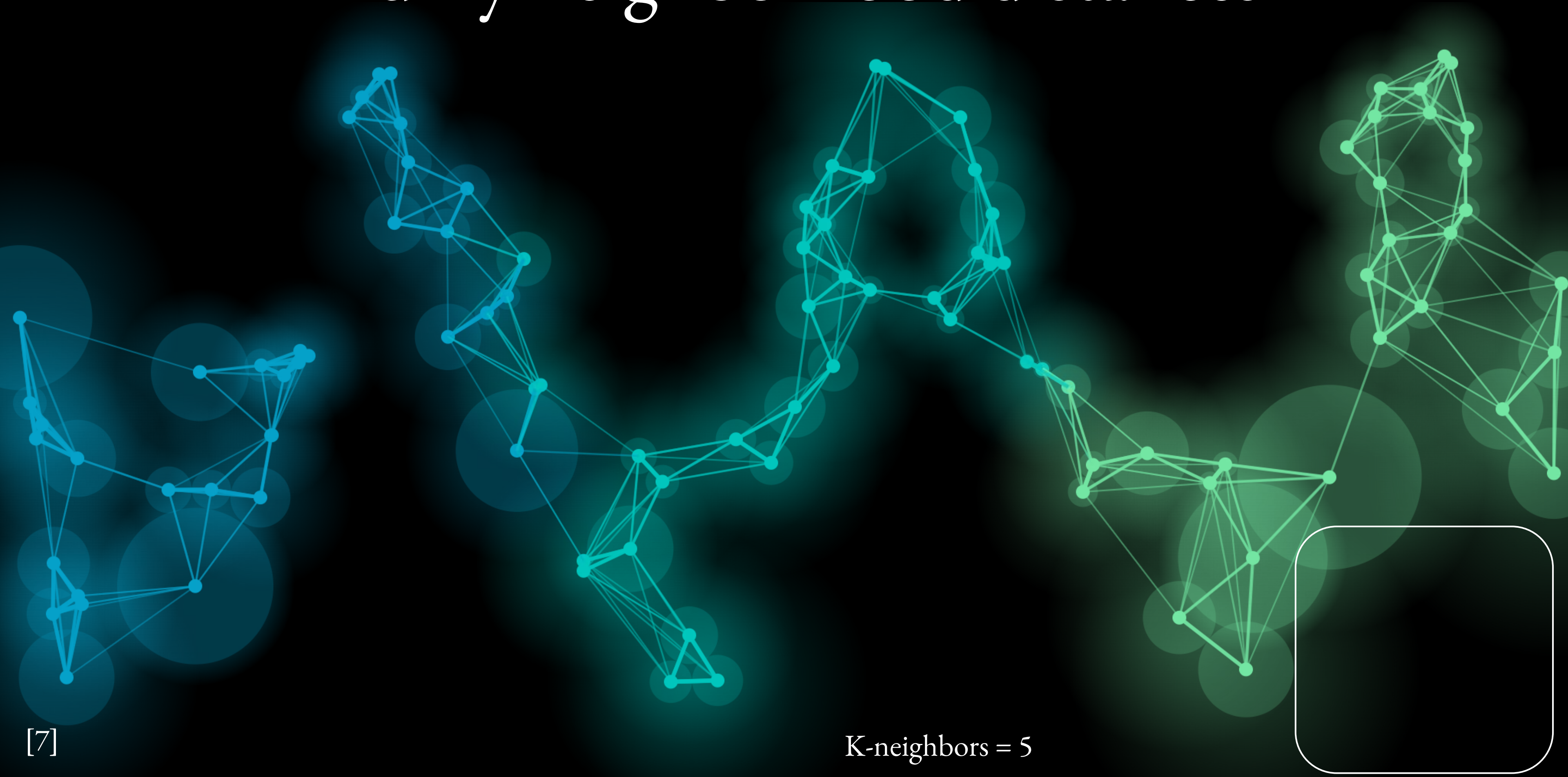
Fuzzy neighborhood distances



[7]

K-neighbors = 2

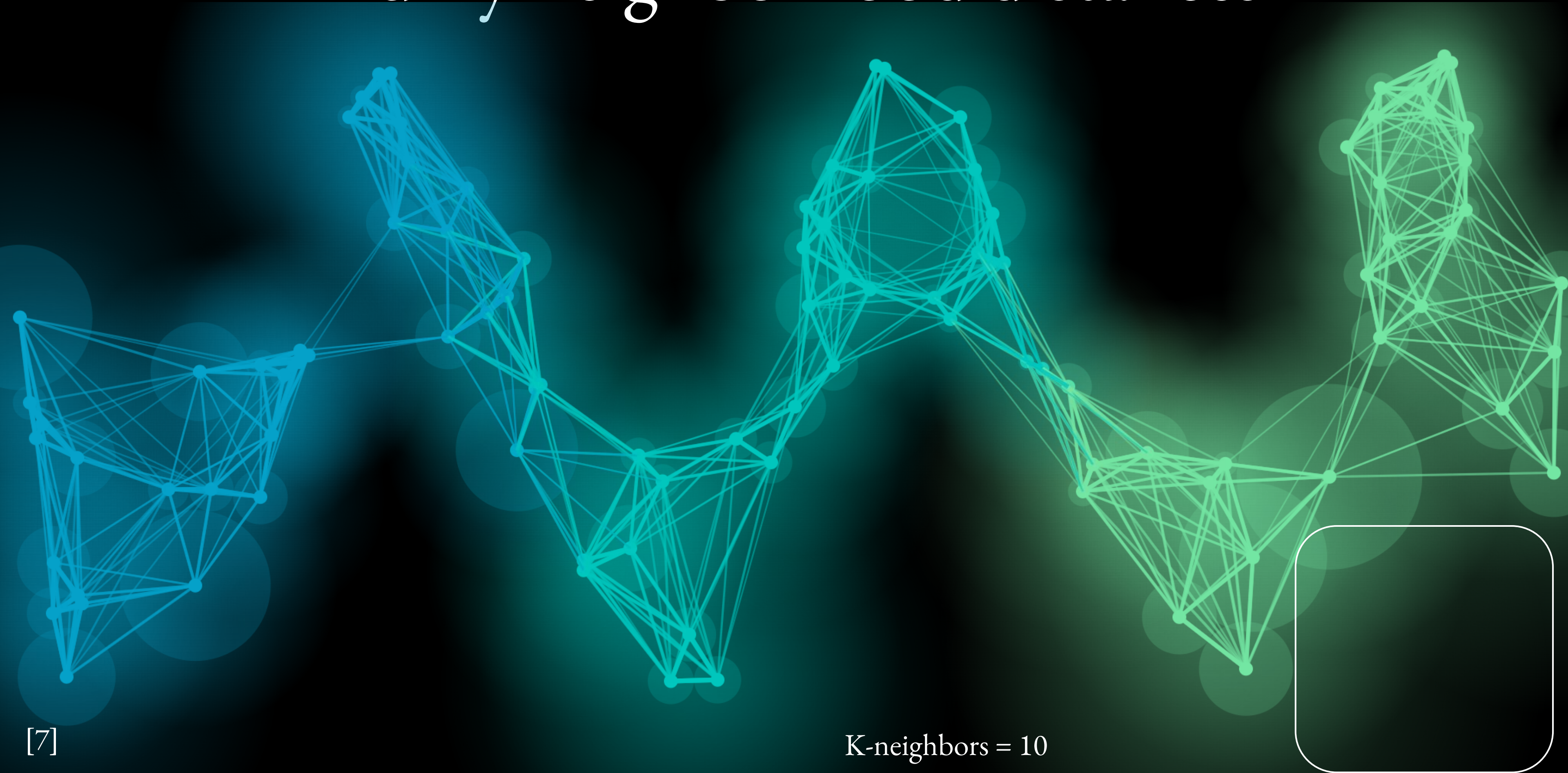
Fuzzy neighborhood distances



[7]

K-neighbors = 5

Fuzzy neighborhood distances



[7]

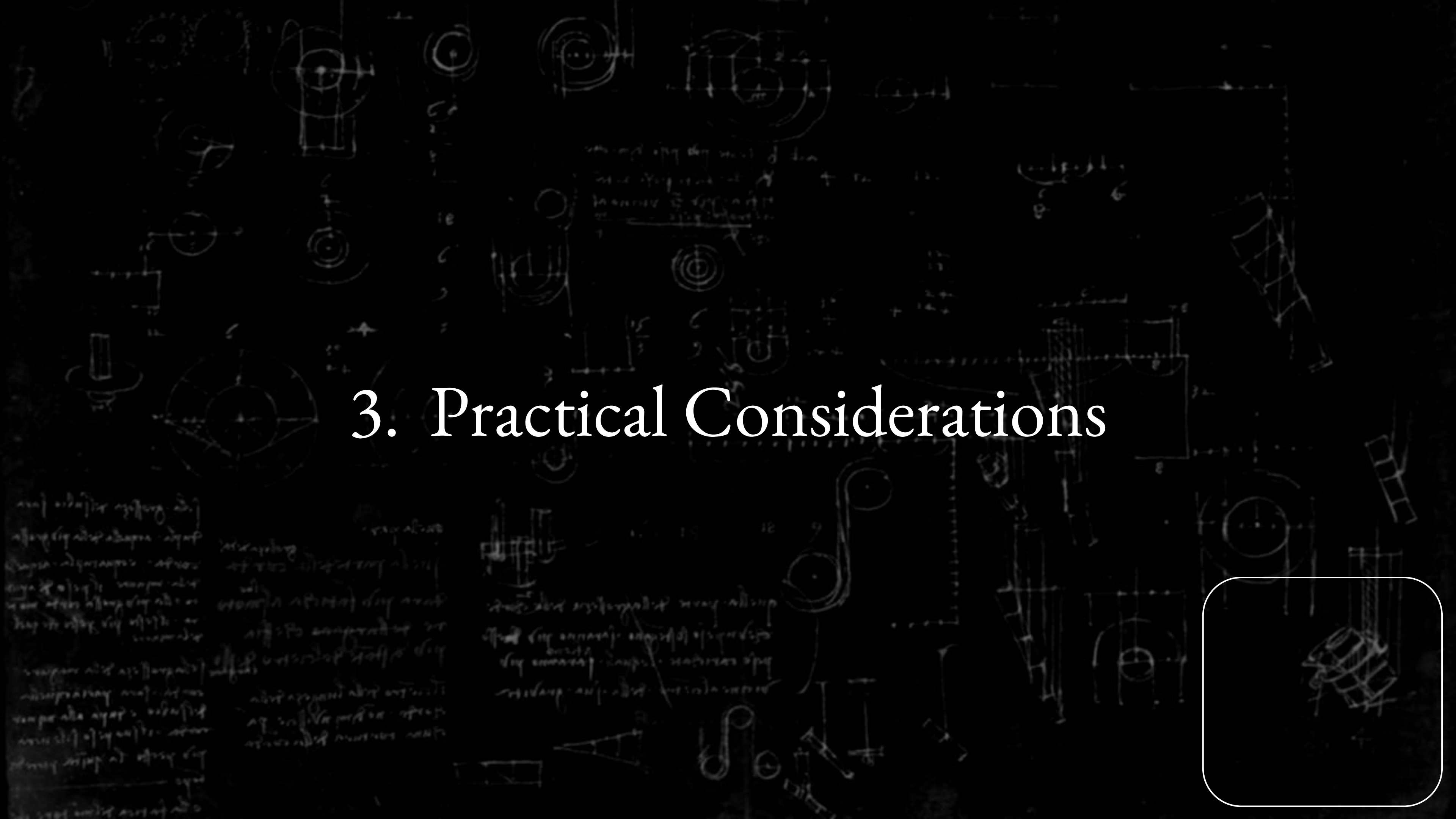
K-neighbors = 10

Initialize and optimize low-dimensional representation

Suppose the weight (probability of an edge $e \in E$ *is* $w_h(e)$ in the original (high dimensional) embedding and $w_l(e)$ in the new (low dimensional) embedding.

We want to minimize the cross entropy,

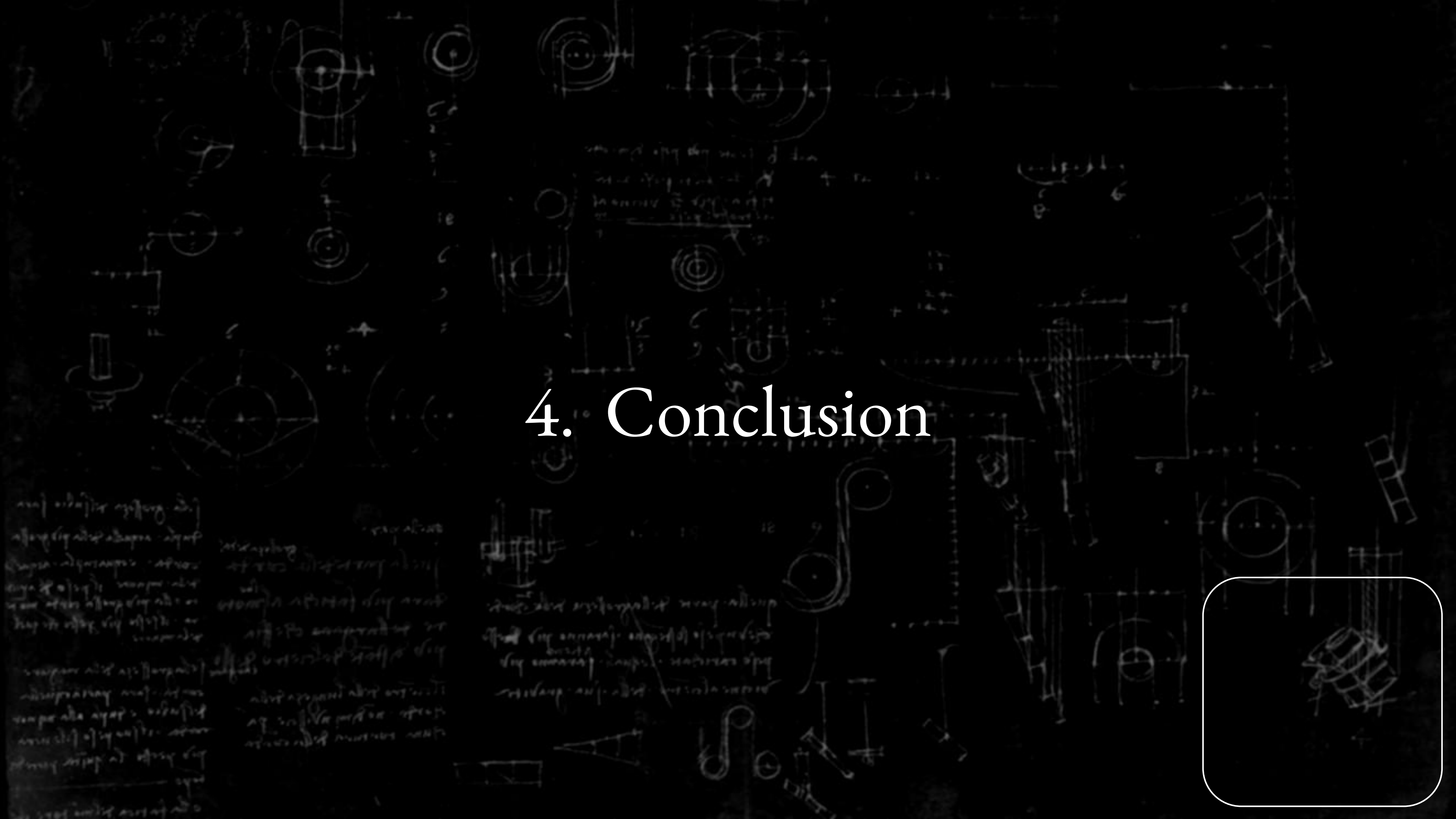
$$\underbrace{\sum_{e \in E} w_h(e) \log\left(\frac{w_h(e)}{w_l(e)}\right)}_{\text{Attractive Force}} + \underbrace{(1 - w_h(e)) \log\left(\frac{1 - w_h(e)}{1 - w_l(e)}\right)}_{\text{Repulsive Force}}$$



3. Practical Considerations

Practical Considerations

- Parameter choice can significantly impact runtime.
- Interpretation is tricky: orientation doesn’t matter, sizes and distances of clusters don’t mean much, and you may see patterns where none exist.
- The burden is on the user to pick parameters that fit your needs.
- It is tempting to use UMAP in places you should not.



4. Conclusion

To sum up:

- UMAP rocks.
- It provides (somewhat hefty) mathematical justification to get theoretically guaranteed results.
- It’s a plug-in-place replacement for t-SNE and uniformly superior for visualizing high-dimensional data

Thanks for listening!

REFERENCES

[1] L. McInnes, J. Healy, and J. Melville, *Umap: Uniform manifold approximation and projection for dimension reduction,* arXiv preprint arXiv:1802.03426, 2018.

[2] L. McInnes and J. Healy, *How umap works*, How UMAP Works - umap 0.5 documentation. [Online]. Available: https://umap-learn.readthedocs.io/en/latest/how_umap_works.html

[3] D. Angelov, *Top2vec: Distributed representations of topics*, 2020.

[4] Wikipedia contributors, *Guernica (picasso) Wikipedia, the free encyclopedia*, 2021, [Online; accessed 17-November-2021]. [Online]. Available: https://en.wikipedia.org/w/index. php?title=Guernica_(Picasso)&oldid=1054068902

[5] D. I. Spivak, *Metric realization of fuzzy simplicial sets*, Self published notes, 2012.

[6] M. Noichl, *Examples for umap reduction using 3d models of prehistoric animals*, GitHub repository, 2019. [Online]. Available: https://github.com/MNoichl/ UMAP-examples-mammoth\-

[7] A. Coenen and A. Pearce, *Understanding umap* [Online]. Available: https://pair-code.github.io/understanding-umap/

QUESTIONS

Please reach out to me at nicholasalines@gmail.com or drop by the github project!