

Outline

1

- Introduction
- Exact Query Processing
- Approximate Query Processing
- Selectivity Estimation
- Open Problems

Open Problems / 1

2

- Understanding high-dimensional data
 - ▣ Visualization:
 - PCA, t-SNE, uMA
- Characterizing high-dimensional data
 - ▣ Existing proposals:
 - Intrinsic dimensionality, Relative Contrast, hubness, growth constant
 - Special structures, including sparse ones
- Characterizing the query workload
 - ▣ Tree indexes assume $k=1$ and data-like query workload

Hard to interpret and limited to 2D/3D

Do not correlate with the hardness of the data

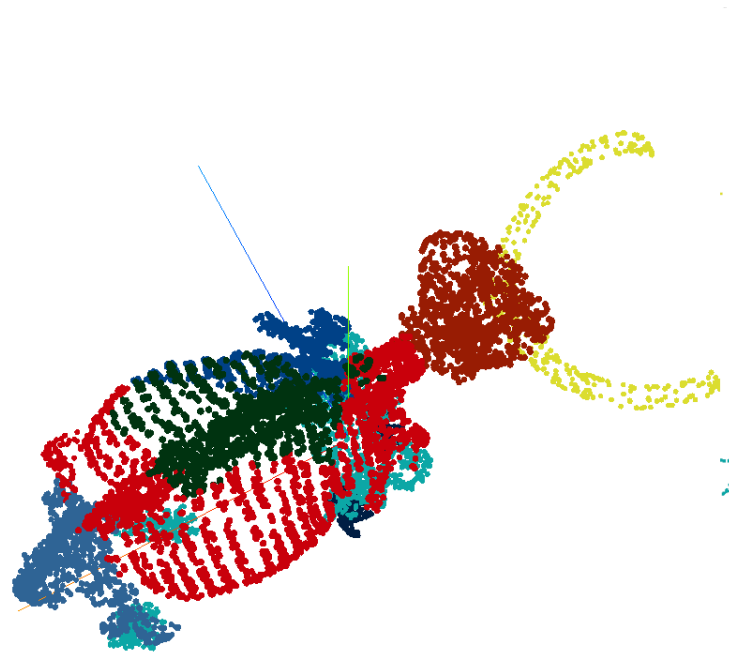
Generative model?

UMAP

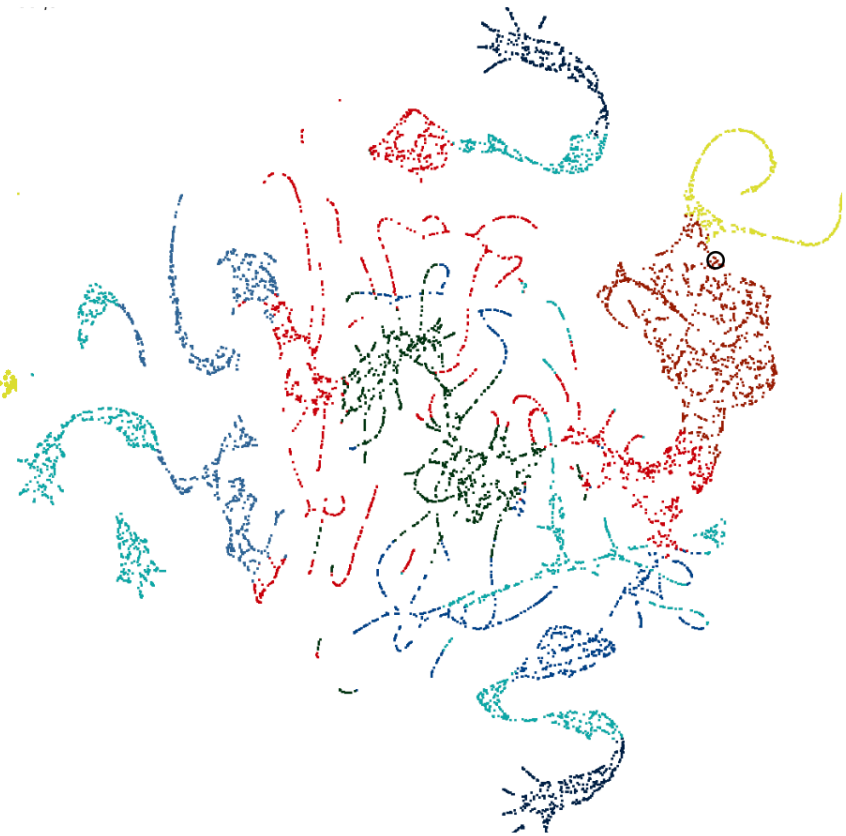
<https://pair-code.github.io/understanding-umap/>

3

Original 3D Data



2D UMAP Projection



3D

n_neighbors: 15

min_dist: 0.1



Open Problems /2

4

- Leveraging Machine/Deep Learning
 - ▣ Huge gap between theory and practice
 - e.g., PCA vs LSH
 - ▣ Many different ideas exist
 - Learning to index
 - Learning to stop
 - Learning to search

Directions:

- New perspectives
- Principled approaches and theories adapted for DS scenarios
- Robustness

Open Problems /3

5

- Handling various hardware and system settings
 - Mixture of
 - CPU/GPU/APU
 - Memory/NVM/SSD/hard disk
 - various distributed computing environments
- Integration with other software stacks
 - With(in) DBMS
 - With(in) big data software stack
 - With(in) machine learning stack
 - With downstream applications

Open Problems /4

6

- Accommodating more down-stream tasks & applications
 - ▣ Non-metric distances
 - ▣ Scores from an evaluation function
- Optimization for similarity queries
 - ▣ Estimating the statistics (cardinality, cost, ...)
 - ▣ Complex join conditions
 - Multiple similarity query predicates
 - Mixed with traditional query predicates
 - ▣ Aggregate queries

Thank You !