Outline

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

Open Problems /1

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

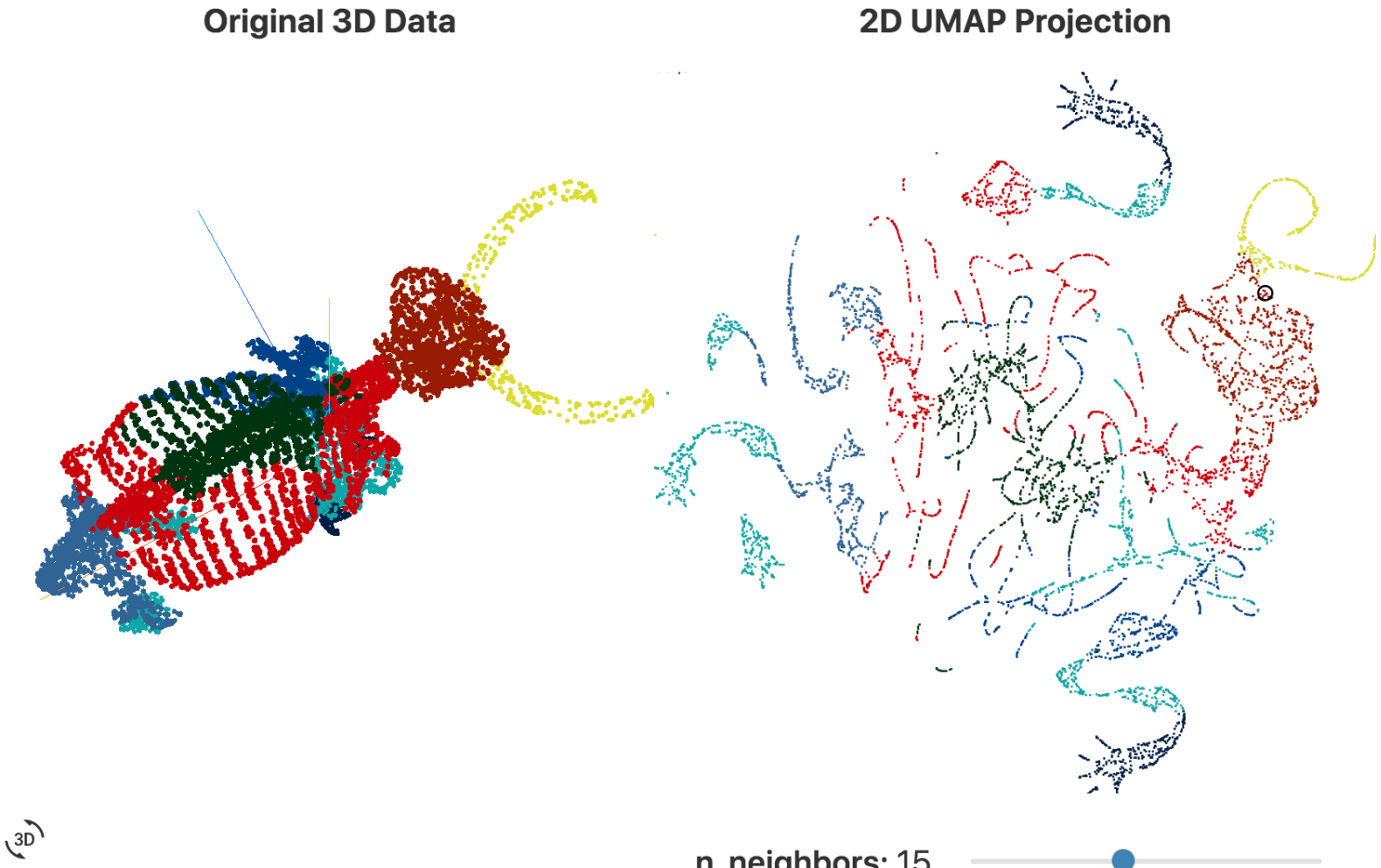
Hard to interpret and limited to 2D/3D

Do not correlate with the hardness of the data

Generative model?

uMAP

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



Open Problems /2

- 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 DB scenarios
- Robustness

Open Problems /3

- 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

- Handle more distance/similarity functions
  - 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

Q&A

Thank You !

VLDB 2020 Tutorial

Similarity Query Processing for High-Dimensional Data

Jianbin Qin, Wei Wang, Chuan Xiao, and Ying Zhang