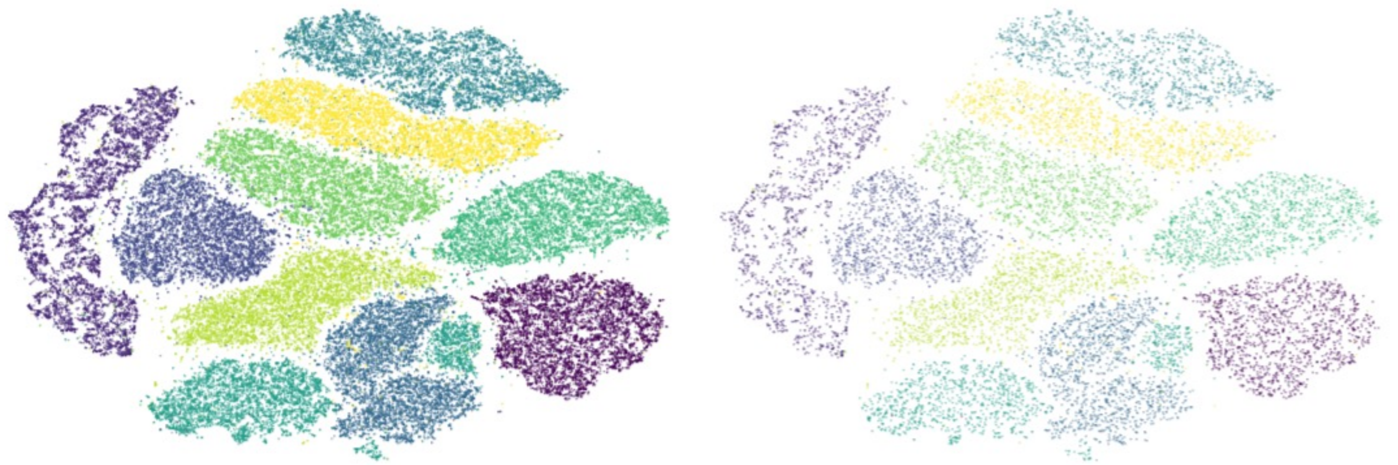
Student Project:

# Outlier Visualization for Multi-dimensional Scatter Plot

University of Zurich UZH



## Context

Outlier detection has been a long-lasting research topic in various fields. However, there has been an imprecise belief that outliers detected in lower dimensional subspace can represent the outliers in the original high dimensional space. In [1], the author showed that it is not reliable to detect outliers for high dimensional data after dimensionality reduction, that is, the corresponding low dimensional subspace cannot precisely capture the outliers in high dimension. This conclusion has a direct influence on many visualization applications, like scatter plot, parallel coordinated plot, and so on.

Scatter plot sampling, on the other hand, is a technique that is used to alleviate the over-plotting issue for high-volume datasets. How to represent the outliers after sampling is a challenging task. Two mostly adopted strategies were applied to visualize the outliers in the sampled scatter plot: first, the outliers are detected before sampling and replotted without being processed (sampled), and second, the outliers are sampled with certain criteria, for example, when sampling the whole dataset, outliers are given a priority to be kept. However, these strategies either cause misleading of the perceived density of points after sampling, or fail to reveal the distribution of all the outliers.

This project focuses on high-dimensional scatter plot sampling and outlier visualization.

## Assignment

This project contains three main tasks: high-dimensional outlier extraction, scatter plot sampling, and outlier visualization. First, you will implement high-dimensional outlier detection algorithm presented in paper [1] using Python. Second, study and implement scatter plot sampling algorithms on the 8 datasets presented in paper [2]. You will need to figure out whether to do sampling before or after dimension reduction. Lastly, visualize the sampled data in 2D with a novel visual design for the outliers.

## Requirements

Preferably you have taken the Data Visualization Concepts or Data Visualization and Analysis courses, with some statistical data analysis background. You should be comfortable with Python programming and JavaScript visualization tools.

## Work Load

- 30% outlier detection
- 40% scatter plot sampling
- 30% outlier visualization

## Student Project Type

This topic can be done as a student project or Master thesis. Detailed requirements are adjusted depending on the project type.

## Supervision

Prof. Dr. Renato Pajarola
Haiyan Yang (assistant)

## Contact

Write an E-Mail to haiyan@ifi.uzh.ch

## References

[1] Wilkinson, Leland. "Visualizing big data outliers through distributed aggregation." *IEEE transactions on visualization and computer graphics* 24.1 (2017): 256-266.

[2] Yuan, Jun, et al. "Evaluation of Sampling Methods for Scatterplots." *IEEE Transactions on Visualization and Computer Graphics* (2020).

VISUALIZATIONAND
MULTIMEDIALAB

Prof. Dr. Renato Pajarola
Visualization and MultiMedia Lab
Department of Informatics
University of Zürich