Problem
● Multidimensional datasets are hard to understand and interpret.
● Visualization helps, but we are limited by 3-dimensional geometry.

Solution
● Apply a dimensionality reduction algorithm that decreases the number of dimensions from P to 3.
● Treat values of new dimensions as x, y, and z coordinates and plot all data points in 3D.

Principal Component Analysis
● Represents original data using new linearly uncorrelated dimensions (principal components).
● Principal components are sorted by variance, so the first few components carry most information.
● Requires computing covariance matrix and its eigenvalue decomposition.
● Computationally expensive, runs in $O(n^3)$.

Implementation
Just drag-and-drop your dataset into a browser window to have it visualized!

Back-end is a Python server that analyzes a dataset, performs dimensionality reduction, and returns a JSON file with x, y, and z coordinates of each data point.

Front-end uses WebGL (three.js) to create an interactive visualization of all data points in space, allowing users to zoom and rotate to explore clustered areas.

Conclusion
● Visualization is a powerful technique to understand data.
● Dimensionality reduction can reveal hidden patterns and relations among data points in high-dimensional space.
● Different algorithms should be applied to different types of datasets, but there is no simple rule to determine which algorithm would work best.

Acknowledgements
I would like to thank Professor Peter Yoon for guidance and encouragement. I would also like to thank my seminar classmates for their feedback and suggestions.

References