# Impact of Dimensionality Reduction with PCA in Cybersecurity

## Data Preprocessing

The dataset used in this analysis is a high-dimensional `.parquet` file containing network traffic features with a target label identifying benign or malicious classes.

### 1. Inspection and Preparation:

The dataset was inspected for structure, missing values, and data types. Non-numeric columns were encoded with `OneHotEncoder`, while numeric columns were scaled to zero mean and unit variance using `StandardScaler`. This standardization is crucial because PCA is sensitive to the scale of features.

### 2. Dimensionality Reduction with PCA:

PCA was initially applied to the preprocessed data to examine the explained variance by each principal component. A cumulative variance plot indicated that [n_components] components were sufficient to retain 90% of the variance, suggesting a significant reduction in dimensions without major information loss.

## Results and Interpretation

### 1. Variance Retention:

By using [n_components] principal components, the PCA-transformed dataset retained approximately 90% of the original variance. This high variance retention indicates that most of the critical information in the original dataset is preserved, allowing us to work with a reduced feature set that maintains data integrity.

### 2. Visualization:

A 2D scatter plot of the first two principal components revealed distinguishable clusters for benign and malicious traffic. This separation suggests that PCA effectively captures the data's essential structure, making it easier to identify patterns associated with malicious activity.

## Applications in Cybersecurity

### 1. Anomaly Detection:

PCA's reduced feature set can help highlight unusual patterns, which may stand out more clearly in a lower-dimensional space. This makes it useful for real-time anomaly detection by enabling faster identification of deviations from normal patterns.

### 2. Malware Classification:

Reduced dimensionality allows machine learning models to focus on key distinguishing features of malware, enhancing classification accuracy and reducing computational costs.

### 3. Intrusion Detection:

The efficiency gained from reduced dimensions can enable quicker analysis in network intrusion detection systems, which often operate under time and resource constraints.

## Limitations and Considerations

While PCA retained a large proportion of variance, some information loss (about 10%) may affect certain detailed analyses, potentially limiting its effectiveness for identifying subtle attack patterns. Additionally, PCA is a linear technique and may not capture non-linear relationships in complex cybersecurity data. In such cases, non-linear methods like Kernel PCA could be more effective.