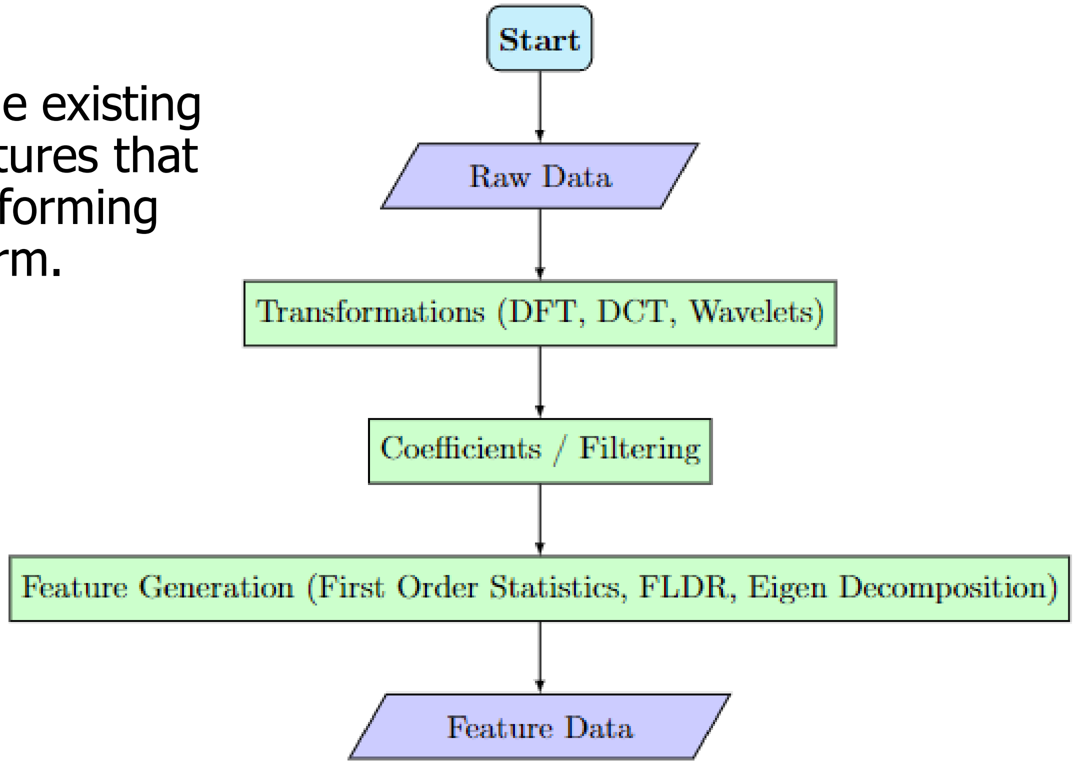
JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Algorithms for Data Science

Feature Engineering, Outliers, and Feature Selection

Feature Engineering

- Creating new features from the existing ones, selecting only those features that are most informative, or transforming features to a more suitable form.



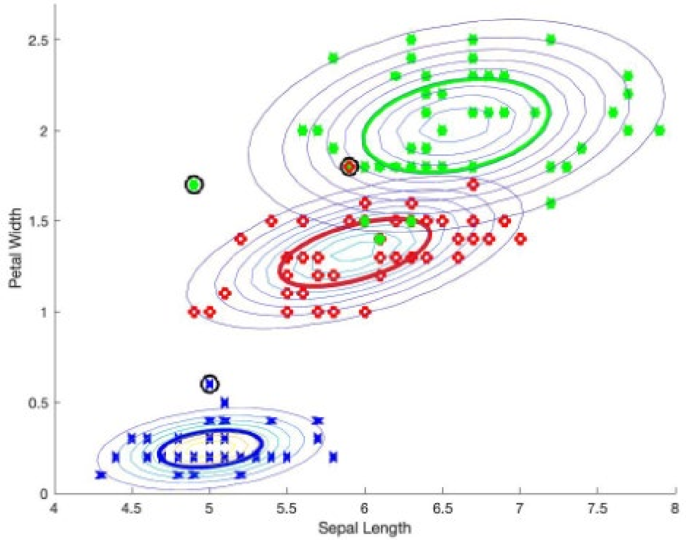
Handling Outliers

- Outliers are data points in a dataset that differ significantly from other observations.
- May have a substantial impact on statistical analyses and machine learning models.
- **Causes of Outliers:**
  - Mistakes in data entry can cause the formation of false outliers.
  - Inaccurate readings can be caused by malfunctioning measurement devices or techniques. Mistakes in measurement can occur.
  - In certain cases, outliers may be indicative of genuine and anticipated variations in the data.

Identifying Outliers

- Scatter plots, histograms, and box plots can be used to visually detect outliers
- Z-scores, Tukey's fences, and the **Mahalanobis distance** are all popular statistical techniques used to identify outliers.
- Algorithms such as Isolation Forest and One-Class SVM can be utilized to identify anomalies, particularly in data with a high number of dimensions.

$$D_i = \left((x_i - \mu)\Sigma^{-1}(x_i - \mu)^T \right)^{1/2}$$

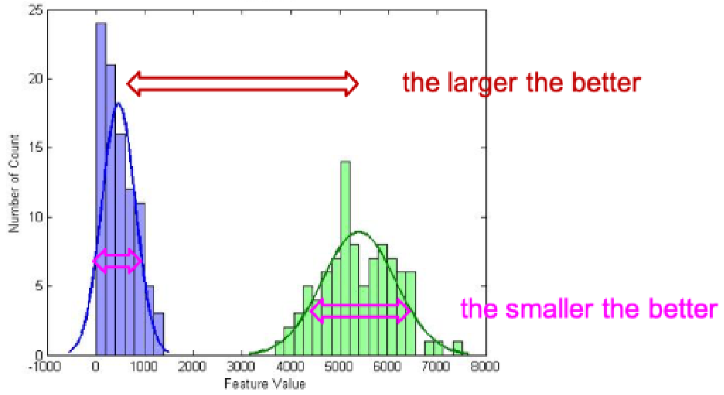


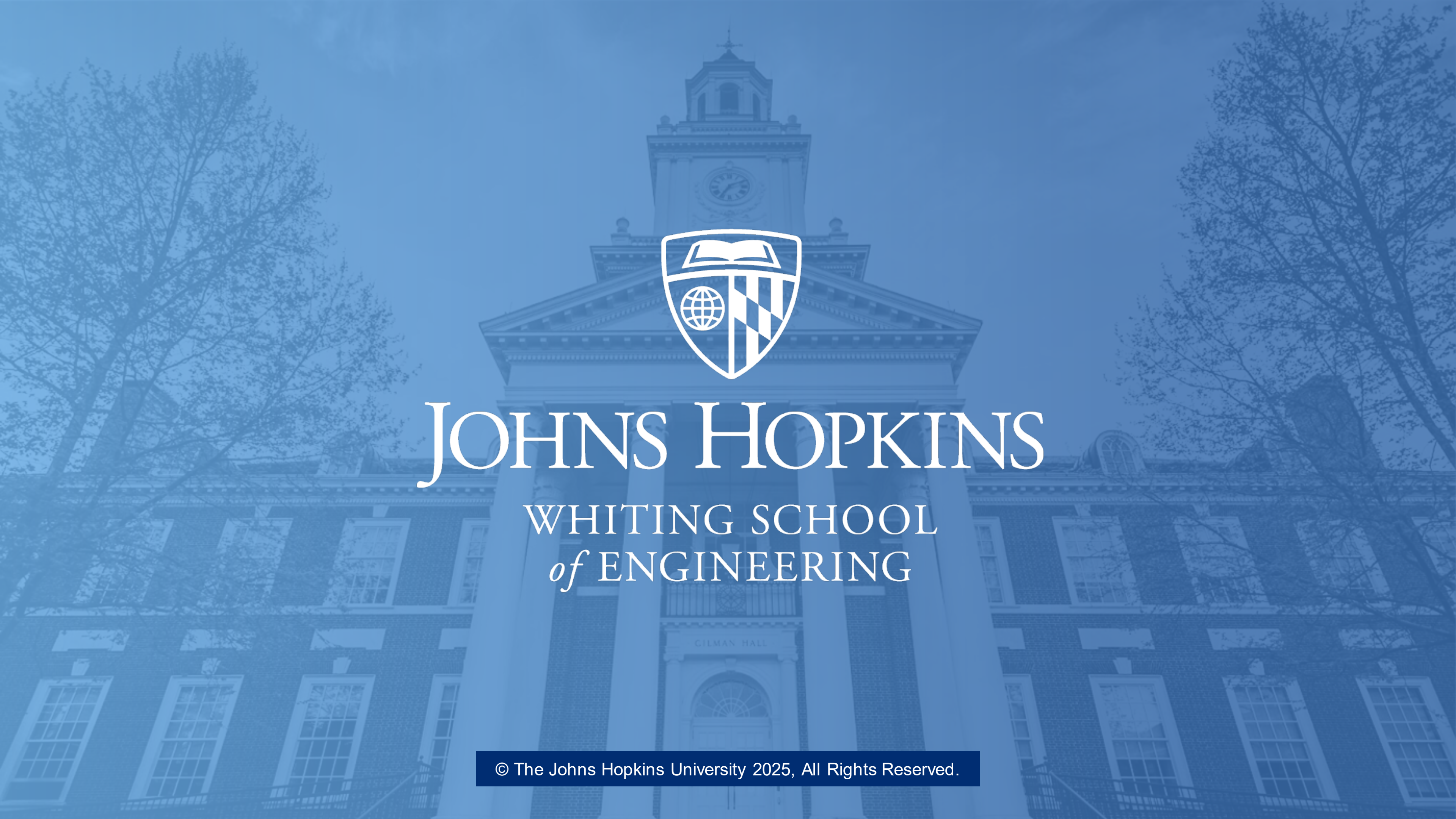
Feature Ranking and Selection

Steps for Feature Ranking and Selection:

1. Features should be ranked in the space that the classifier will be used.
2. Features should be ranked individually with a metric that gives a value on how well the feature classifies the observations.
3. Iterative add features to identify how well the top *n* features perform.
4. Continue until the classification accuracy falls or reaches a steady state.

$$\boldsymbol{FDR} = \frac{\left(\mu_1 - \mu_2\right)^2}{\sigma_1^2 + \sigma_2^2}$$




JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING
GILMAN HALL