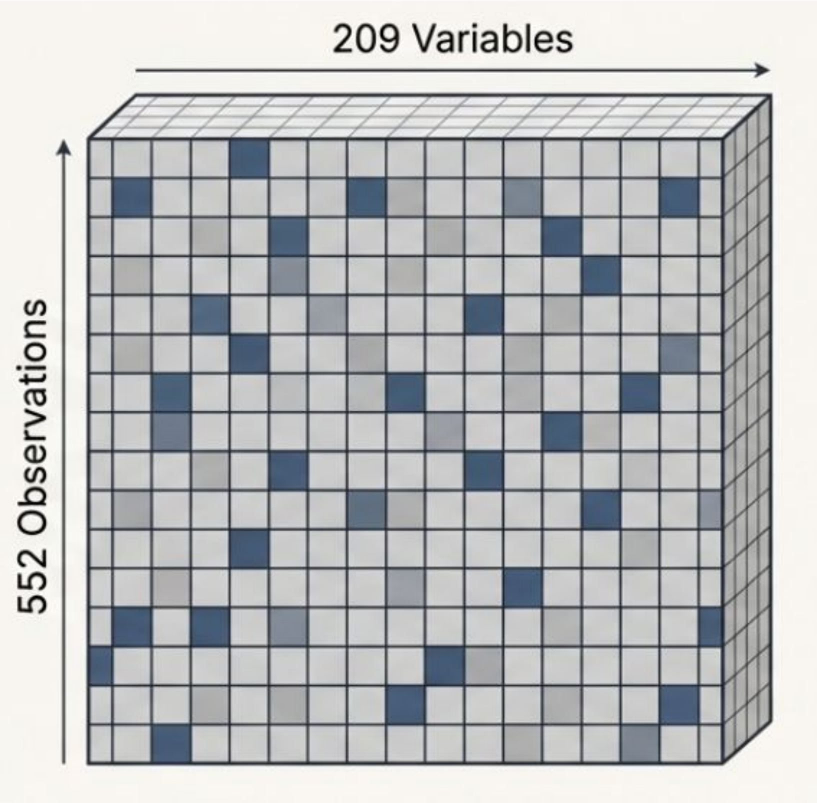
# IENG 551: Quality & Reliability Engineering Project

Submitted by:
Akshay Patel

Instructor: Imtiaz Ahmed, Ph.D.
Assistant Professor, IMSE Department

West Virginia University
December 2025

# The Challenge: Establishing a Stable Process Baseline



**Dataset:** A high-dimensional manufacturing dataset comprising 552 records, each with 209 anonymized variables.

**Problem:** The data is an unknown mixture of 'In-Control' (IC) and 'Out-of-Control' (OOC) observations. The physical meaning of the variables is omitted.

**Objective:** To conduct a robust Phase I analysis. The goal is to isolate the true IC data, estimate its statistical parameters, and establish a validated framework for future (Phase II) process monitoring.

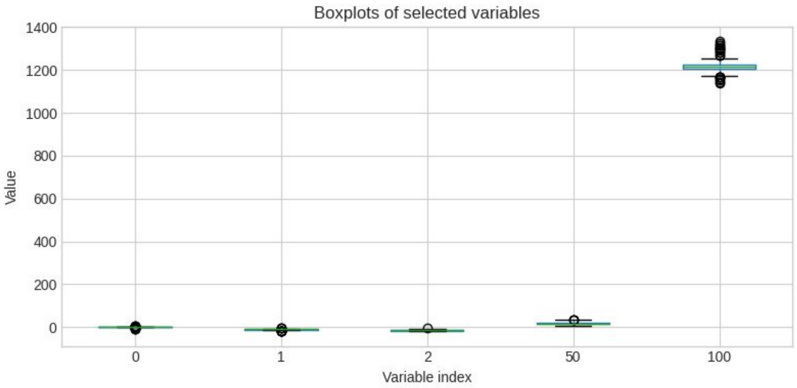
# Initial Investigation: Uncovering Data Characteristics

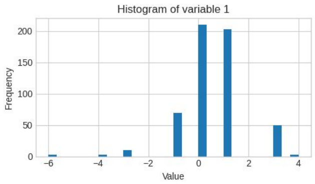
Exploratory Data Analysis (EDA) revealed several key features of the raw data:

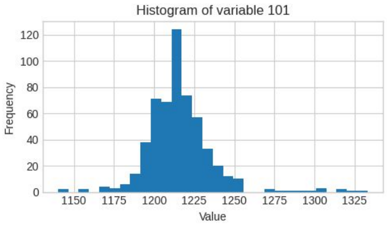
- **Completeness:** No missing values were found across the entire dataset.

- **Scale Disparity:** Variables operate on vastly different scales, as shown in the boxplots (e.g., Var 1 vs. Var 101).

- **Distribution Variety:** Many variables exhibit non-Gaussian distributions, with significant skewness and potential outliers.







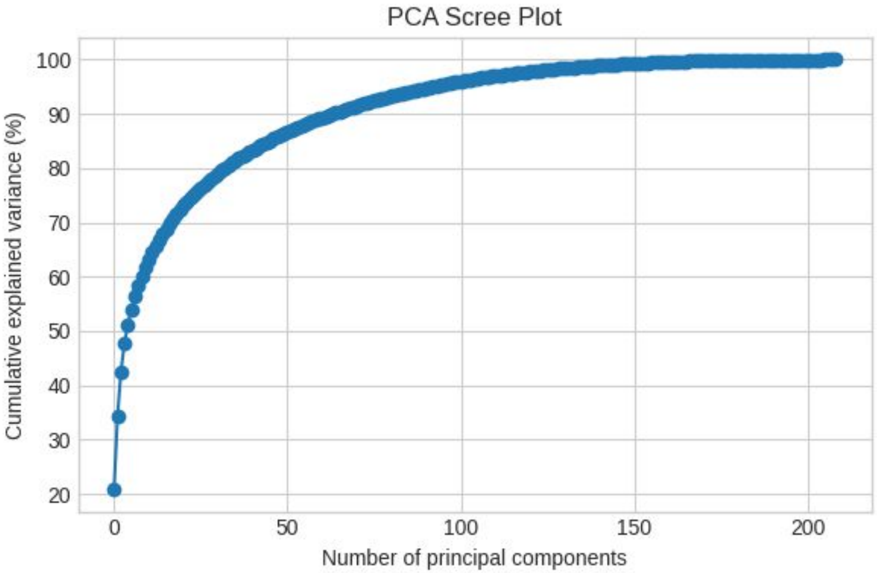
**Justification:** The high dimensionality, significant scaling differences, and non-normality mandate a multi-stage approach beginning with data standardization and dimensionality reduction.

# Taming Complexity: Dimensionality Reduction via PCA

Principal Component Analysis (PCA) was applied to the standardized data to manage the high dimensionality (209 variables).

- PCA transforms the original, correlated variables into a smaller set of uncorrelated Principal Components (PCs), preserving the majority of the process variation.

**Decision:** We retained the first **46 Principal Components**, which collectively capture **85%** of the total process variance. This provides a robust, lower-dimensional space for more effective anomaly detection.



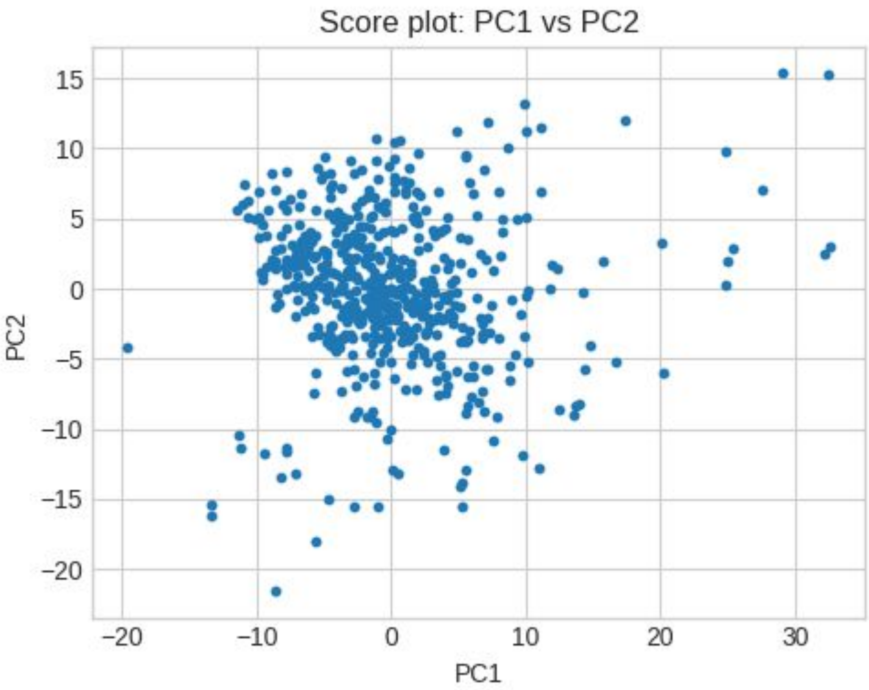
# Anomaly Detection, Part 1: Identifying Outliers with Elliptic Envelope

The `EllipticEnvelope` algorithm, a robust method for outlier detection, was applied to the 46-dimensional PC data.

This method assumes the core data is Gaussian and fits an ellipse to it, classifying observations falling outside this shape as anomalies.

Based on initial EDA, an expected `contamination` level of 10% was used as a key parameter.

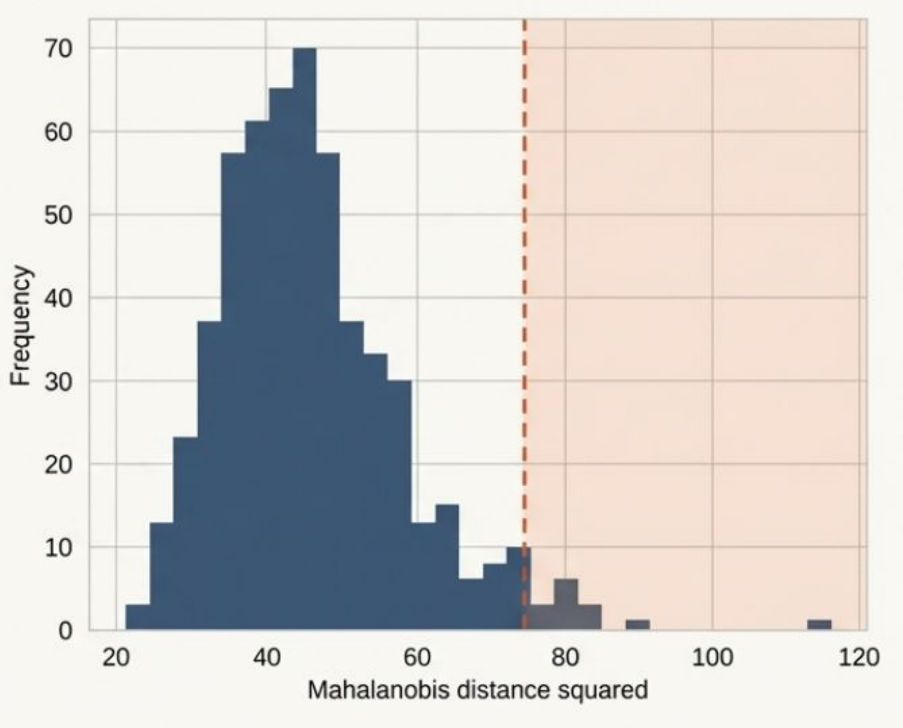
**Result:** This method identified **56** potential Out-of-Control (OOC) observations.



# Anomaly Detection, Part 2: Quantifying Deviation with Mahalanobis Distance

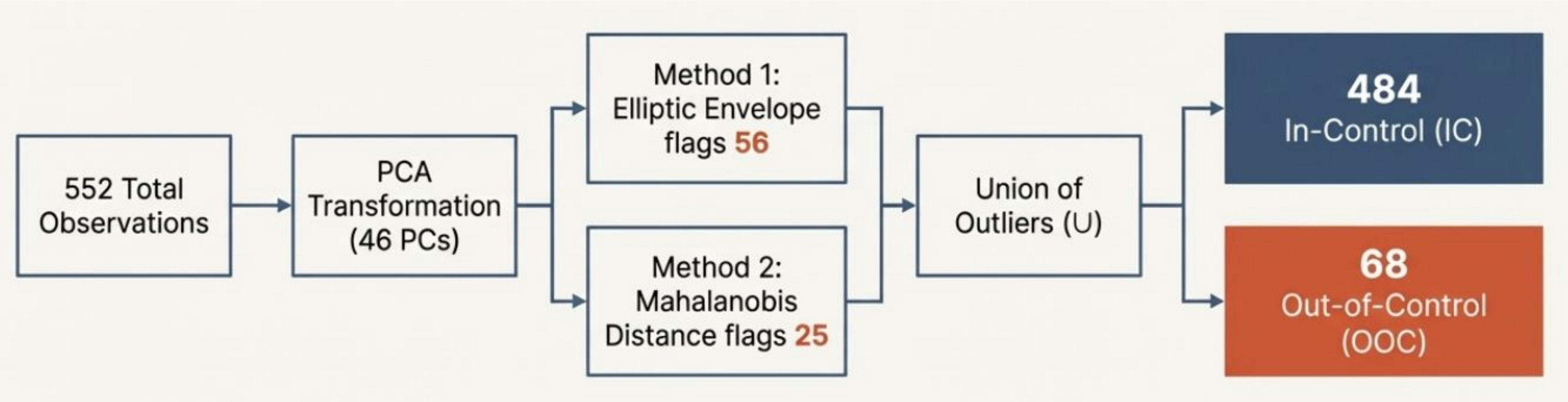
Mahalanobis Distance ($MD^2$) was calculated for each observation in the 46-dimensional PC space. $MD^2$ measures the distance from the multivariate mean, accounting for correlations.

A statistical control limit was set using a Chi-squared ($\chi^2$) distribution with 46 degrees of freedom ($k = 46$ PCs) and a significance level of $\alpha = 0.01$.



**Result:** The calculated $\chi^2$ threshold of **71.2** flagged **25** observations as statistically significant outliers.

# Synthesizing the Evidence: Defining the Final ‘In-Control’ Set


552 Total Observations
PCA Transformation (46 PCs)
Method 1: Elliptic Envelope flags 56
Method 2: Mahalanobis Distance flags 25
Union of Outliers (U)
484
In-Control (IC)
68
Out-of-Control (OOC)

To establish the most conservative and reliable baseline, an observation was classified as Out-of-Control (OOC) if it was flagged by **either** the Elliptic Envelope **or** the Mahalanobis Distance method.

This ‘union’ approach ensures that any point considered anomalous by either robust technique is removed from the initial in-control set.

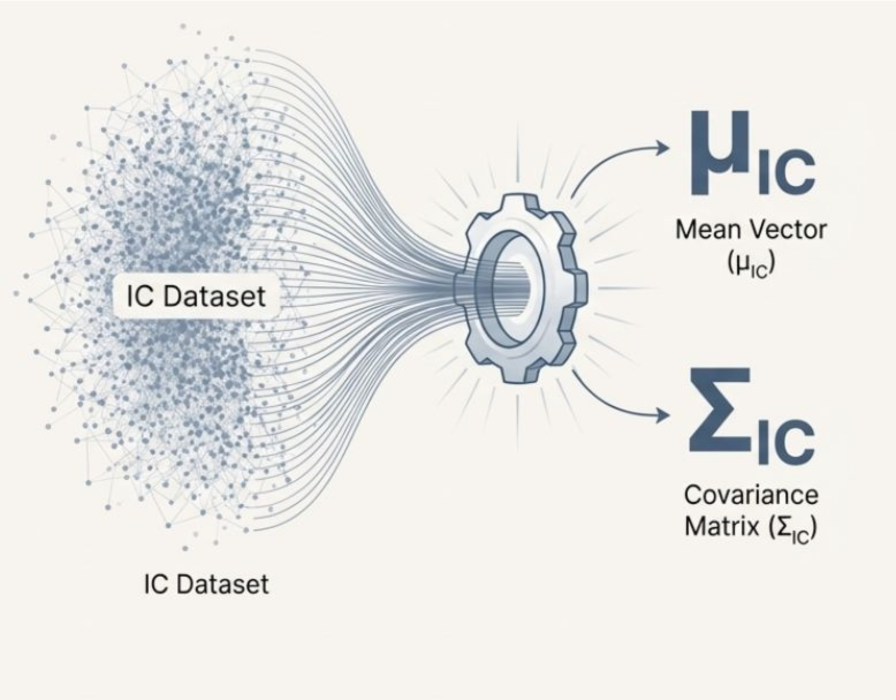
**Final Classification:** This combined approach yielded **68 unique OOC points**, leaving a core dataset of **484 In-Control (IC)** observations.

# Building the Foundation: Characterizing the Stable Process

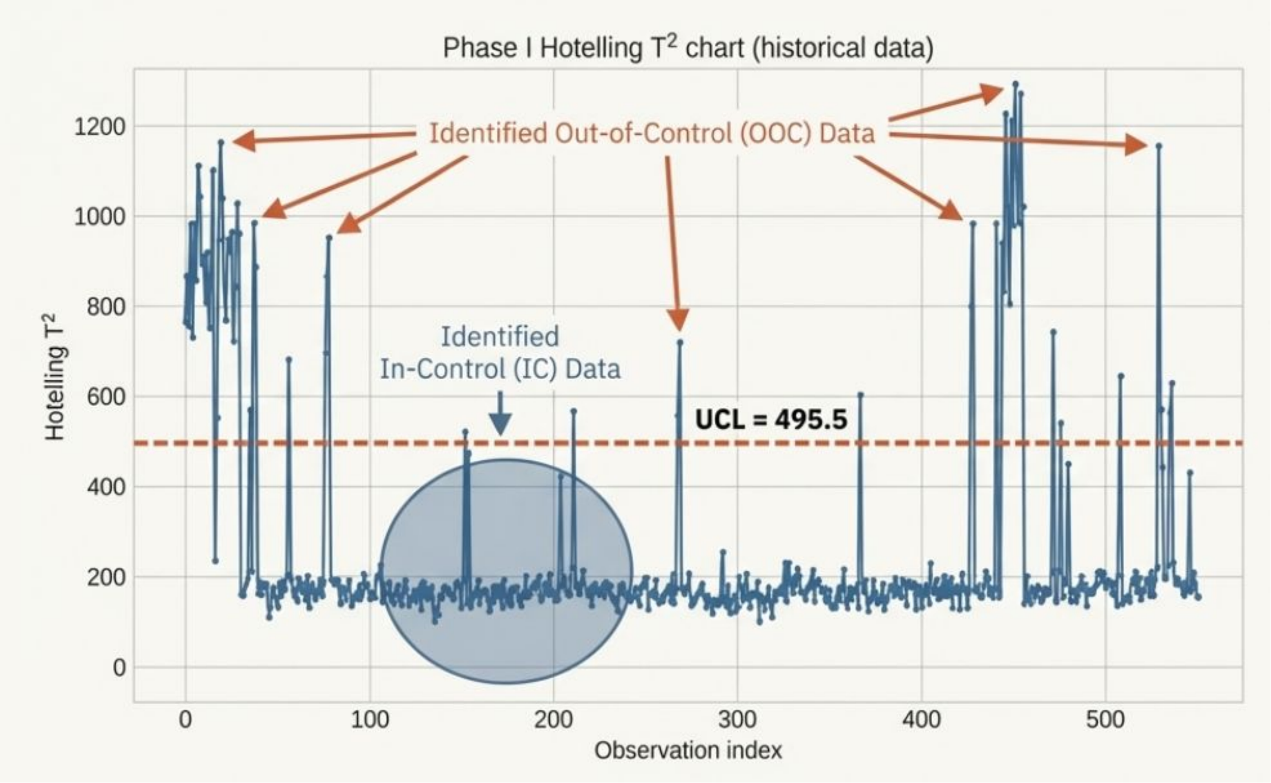
The 484 identified IC observations are assumed to represent the stable, common-cause variation of the manufacturing process. This cleaned dataset was used to calculate the essential parameters required for a multivariate control chart:

- **In-Control Mean Vector ($\mu_{ic}$):** A 209×1 vector defining the process center.

- **In-Control Covariance Matrix ($\Sigma_{ic}$):** A 209×209 matrix describing the process variability and the relationships between variables.



# The Verdict: Validating the Baseline with a $T^2$ Control Chart



Using the estimated $\mu_{ic}$ and $\Sigma_{ic}$, a Hotelling's $T^2$ value was calculated for all 552 original observations.

The Upper Control Limit (UCL) was established at **495.5** based on the F-distribution, using the number of IC samples ($m = 484$), variables ($p = 209$), and a significance level of $\alpha = 0.01$.

**Validation:** The chart provides clear visual confirmation of our analysis. The OOC points identified by our procedure correspond directly to the high $T^2$ values that breach the statistical control limit.

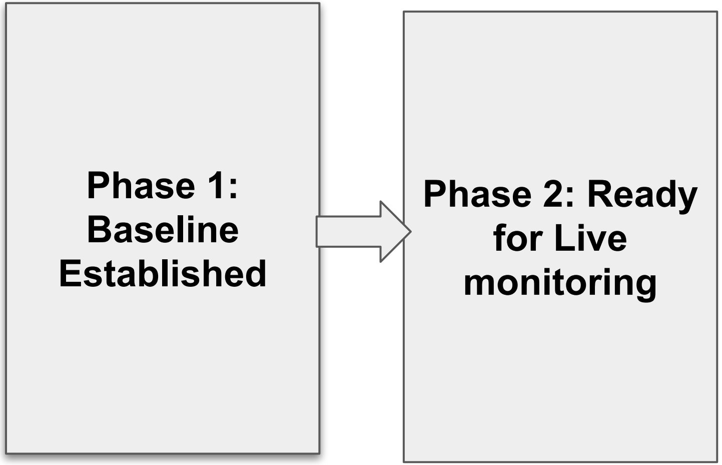
# Conclusion & The Path to Phase II Monitoring

**Project Success:**
A rigorous, multi-stage procedure successfully isolated a stable in-control baseline from a high-dimensional, unlabeled dataset. This validated approach is essential for effective statistical process control.

**Key Deliverables:**

1. A robust estimate of the in-control process mean vector ($\mu_{ic}$) and covariance matrix ($\Sigma_{ic}$).

2. A statistically-derived Hotelling's $T^2$ Upper Control Limit (UCL) for future anomaly detection.


Phase 1: Baseline Established
Phase 2: Ready for Live monitoring

**Path Forward:**The established parameters ($\mu_{ic}$, $\Sigma_{ic}$) and the UCL (**495.5**) form a complete and validated framework. This system is now ready for implementation in Phase II for the real-time monitoring of new manufacturing data.