# Data Analysis Project Report

Team: Ohm's squad
Frizberg Raphael
Rauch Bruno
Biljesko Leo

## 1 Contributions

- Frizberg Raphael:
  - Data preprocessing and Basic Analysis
  - Visualization and Exploratory Analysis
  - Probability Analysis
- Rauch Bruno:
  - Statistical Theory Applications
- Biljesko Leo:
  - Regression Analysis
  - Report Writing and Documentation
- Github Repository: (https://github.com/RF-at-FH-Joanneum/DataAnalysis_Project_EEM24.git)

## 2 Dataset Description

- Dataset name and source: Westermo system test dataset (https://github.com/westermo/test-results-dataset.git)
- Time period and sampling frequency: sampled twice per minute for a month
- Key variables analyzed: server-up, load-15m, memory_used_pct, cpu-user, cpu-system, sys-thermal, sys-interrupt-rate, disk-io-time
- Stats: 86,383 observations, 0 missing values, 18,172 outliers removed.

# 3 Methods and Analysis

## 3.1 Data Preprocessing

- Cleaning procedures involved removing outliers, missing values, duplicate values, invalid values, taking specific range and sorting out by datetime
- Outlier handling: The outliers are removed by IQR method
- Missing value treatment: Removed rows with empty entries (no interpolation to avoid misrepresentation of data characteristics).
- Data transformations: load-15m multiplied by 100, memory_used_pct calculated, timestamp converted to datetime.

## 3.2 Exploratory Data Analysis

- Distribution analysis: .describe() was used to calculate statistics such as: mean, median, standard deviation, min, max, etc.. Values are stored in CSV file
- Time series patterns: Data was grouped by hour to examine trends in variables over time. Mean and standard deviation were computed for metrics and plotted as a line graph. Further variability was added that indicates fluctuations.
- Correlation analysis: Correlation has been plotted using .heatmap() showing the relationship between all metrics.
- Key visualizations: Matplotlib and seaborn have been used to produce plots. Other visualizations have been applied such as: histograms, boxplots, heatmaps, time plots, etc..

## 3.3 Statistical Analysis

- Probability analysis: To examine probability, threshold-based probability estimation is performed. Additionally, crosstable analysis and conditional probability analysis are applied.
- The Law of Large Numbers is demonstrated as the observed probability converges towards the measurements mean (threshold), while the absolute error decreases towards zero as the sample size increases.
- The Central Limit Theorem is demonstrated as the histograms of the sample means become closer to the theoretical normal curves as the sample size increases.

## 3.4 Regression Analysis

- Polynomial regression (deg. 1–6) was used to model non-linear relationships between features and the target variable.
- Models were evaluated using cross-validation (5-fold splits), with $R^2$ and RMSE as performance metrics.
- Training and CV scores were analyzed to assess overfitting or underfitting, identifying degree 5 as a good balance between complexity and performance.
- StandardScaler ensured consistent feature scaling during model fitting.

# 4 Key Findings

## 4.1 Statistical Insights

- Distribution characteristics: cpu-user & -system, as well as sys-interrupt have normal distribution. Load-15m, memory_used_pct, disk-io-time and sys-thermal are multi-modal.
- Significant correlations: cpu-system (-user) & sys-interrupts, cpu-user & memory_used_pct

## 4.2 Pattern Analysis

- Temporal patterns: By displaying the mean and standard deviation, Daily Patterns shows hourly behavior. The processed data seems tighter (lower std), suggesting higher consistency, yet perhaps at the expense of important details like CPU-system behavior and system interrupts.
- Variable relationships: The strongest relationships are between cpu-user and sys-interrupt-rate, as well as cpu-system and sys-interrupt-rate, while moderate relationships are observed between cpu-user and cpu-system, and sys-thermal and sys-interrupt-rate.
- Identified anomalies: system-19 shows a clear problem around 16-17.01. visible in load-15m, memory_used_pct and best visible in disk-io-time (processed data)

## 4.3 Advanced Analysis Results

- For lower polynomials, the gap between training and CV scores implies underfitting. At higher degrees, the gap reduces, but still with slight overfitting. Degree 5 seems like a best solution for a model.
- At degree 1, $R^2$ fails to capture data's curvature, leading to low values. With higher degrees, model starts to capture more data, and $R^2$ increase. At degrees 5 and 6, it stabilizes at around 0.629.
- Residuals are well distributed around zero, resulting in no major errors. In Q-Q plot, residuals align well with red line, with slight tail deviations (could indicated non-normality).

# 5 Summary and Conclusions

- Main insights: Key distribution features were identified by the analysis; load-15m and sys-thermal showed multi-modal behavior, whereas variables like cpu-user and cpu-system showed normal distributions. Significant correlations were found, especially between cpu-user and sys-interrupt-rate, which offered valuable information on the dynamics of system performance. A 5th-degree model, which achieved an $R^2$ of roughly 0.629 and successfully captured the underlying data patterns, was found to be the best compromise between complexity and performance using polynomial regression analysis.
- Limitations: It's possible that potentially useful trends were lost as a result of the choice to exclude missing data rather than interpolate. Furthermore, eliminating outliers might have eliminated important abnormalities that could have provided more in-depth information while also correcting noise. Lastly, during data aggregation, some temporal pattern variability was decreased, which would have masked smaller features that are essential to understanding system actions.