## Introduction

The objective of this task was to predict **Crowd Energy (0–100)** for future concert venues using historical tour data provided by Electric Omen. The dataset was intentionally noisy and inconsistent, requiring careful data cleaning, exploratory analysis, feature engineering, and model validation. Additionally, the Lead Singer's notes were treated as hypotheses rather than ground truth and were tested against real data.

---

## Data Cleaning & Preprocessing

The raw dataset contained several data quality issues:

- **Data leakage risks:**
  Columns such as Crowd_Size and Merch_Sales_Post_Show were collected after the show and were removed to prevent leakage. The identifier column Gig_ID was also dropped.
- **Date & time inconsistencies:**
  Show_DateTime was converted into meaningful features (Show_Hour, Show_Month, Is_Weekend). Missing values were imputed using median or mode-based strategies.
- **Ticket price normalization:**
  Ticket prices were recorded in mixed currencies (£, €, $) and included malformed values such as "Free" and VIP prices. All prices were converted to USD using provided exchange rates, free tickets were mapped to $0, and invalid values were handled safely.
- **Sensor & numeric errors:**
  Zero or negative sensor readings were treated as missing and imputed. Outliers were capped using the IQR method.
- **Target validation:**
  Since Crowd Energy is defined on a 0–100 scale, rows containing negative or extremely large values (e.g., 999) were removed to avoid corrupt labels.

---

## Exploratory Data Analysis (EDA)

EDA revealed strong **venue-specific behaviour**:

- Crowd energy distributions varied significantly across venues, confirming that each venue has unique dynamics.
- Increasing volume level showed **diminishing returns**, especially in constrained venues, partially supporting the singer's noise-limit theory.

- Ticket pricing influenced crowd energy differently across venues. Premium venues such as **V_Gamma (The Snob Pit)** showed higher energy at higher price ranges, while other venues showed weaker or noisier relationships.

These findings confirmed that venue-specific interactions are important for modelling .

---

**Feature Engineering**

Key engineered features included:

- Temporal features: show hour, month, and weekend indicator.
- Normalized ticket prices in USD.
- Safe handling of categorical variables (Venue_ID, Weather, Moon_Phase, Band_Outfit) with unseen-category support.
- Interaction-ready structure to capture venue-dependent effects.

---

**Modelling & Hyperparameter Tuning**

A **Ridge Regression** model was used as a baseline due to its simplicity and interpretability.

- **Baseline RMSE: ≈ 18.13**

A **Random Forest Regressor** was then trained using **GridSearchCV with 5-fold cross-validation**. Hyperparameters such as number of trees, depth, and minimum samples were tuned.

- **Tuned Random Forest RMSE: ≈ 16.80**

The tuned model outperformed the baseline, demonstrating the benefit of non-linear modelling for this problem.

---

**Conclusion**

This project demonstrated the importance of robust data cleaning, leakage prevention, and careful validation. Venue-specific dynamics play a major role in determining crowd energy, and ensemble models with hyperparameter tuning generalize better than linear baselines. The final tuned Random Forest model was selected for prediction due to its superior performance.