# Interpreting Financial Machine Learning Models

## 1. The Need for Interpretability in Finance

Financial institutions rely on machine-learning models for high-stakes decisions (e.g. credit approval, risk scoring). However, *black-box* models (e.g. random forests, neural nets) lack transparency:

- **Regulatory compliance**: Laws (e.g. GDPR) require explanations for automated decisions.
- **Trust & accountability**: Loan officers and customers demand meaningful reasons.
- **Bias detection**: Hidden biases (e.g. demographic) must be discovered and mitigated.

## 2. SHAP and LIME: Local Feature Attribution

Both SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) provide *per-prediction* explanations by attributing the change in model output to each input feature.

- **SHAP** is grounded in cooperative game theory: it computes Shapley values, the *fair* average contribution of each feature across all feature-coalitions.
- **LIME** fits a simple, interpretable surrogate (e.g. linear) model locally around the instance being explained.

## 3. SHAP Algorithm Outline

1. **Train the predictive model** $f : \mathbf{x} \to y$ on data $\{\mathbf{x}_i, y_i\}$.
2. **Choose an explainer:** e.g. for tree-based models, use `shap.TreeExplainer`.
3. **Compute SHAP values:** for each instance $\mathbf{x}$, compute

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[ f_{S \cup \{j\}}(\mathbf{x}) - f_S(\mathbf{x}) \right],$$

   where $F$ is the set of all features and $f_S$ denotes the model with features outside $S$ marginalized out.

4. **Aggregate or display:**
   - *Global summary:* average $|\phi_j|$ across all instances.
   - *Local force plot:* show how each feature pushes the prediction from the base value.

## 4. Integrating SHAP into a Financial Classifier

Below, we demonstrate end-to-end integration: synthetic data, training a Random Forest, computing SHAP values, and visualizing explanations.