MSc in Finance & Banking September 25, 2025

Big Data & Machine Learning

FinTech Project

Week 2 - Team C

Sprint Objectives

- Integrate equity & cryptocurrency data into a unified 7-day dataset
- Apply and compare imputation techniques to handle missing values
- Evaluate methods with error metrics and interpolation approaches
- Test advanced imputation (KNN, MICE) and add safeguards against anomalies
- Link data processing choices to investor philosophy (Aggressive Persona)
- Build final dataset with returns and assess portfolio performance

Week 1 - Portfolio construction

Week 2 - Data processing

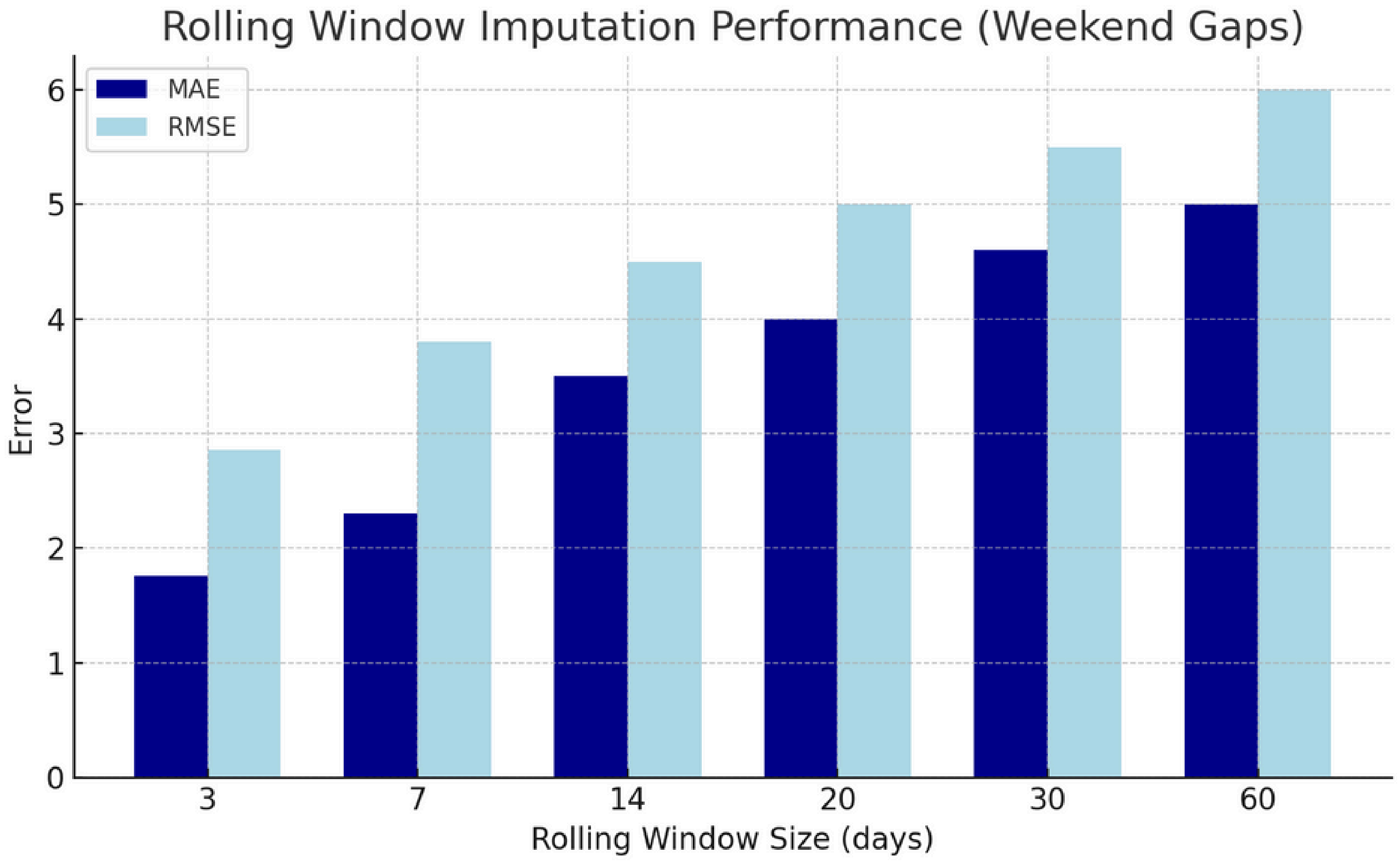
Mean, Median & Rolling Windows

Weekend gaps filled using mean vs median

- **Mean**: preserves volatility → aggresive investors
- **Median**: robust to outliers → risk-averse investors

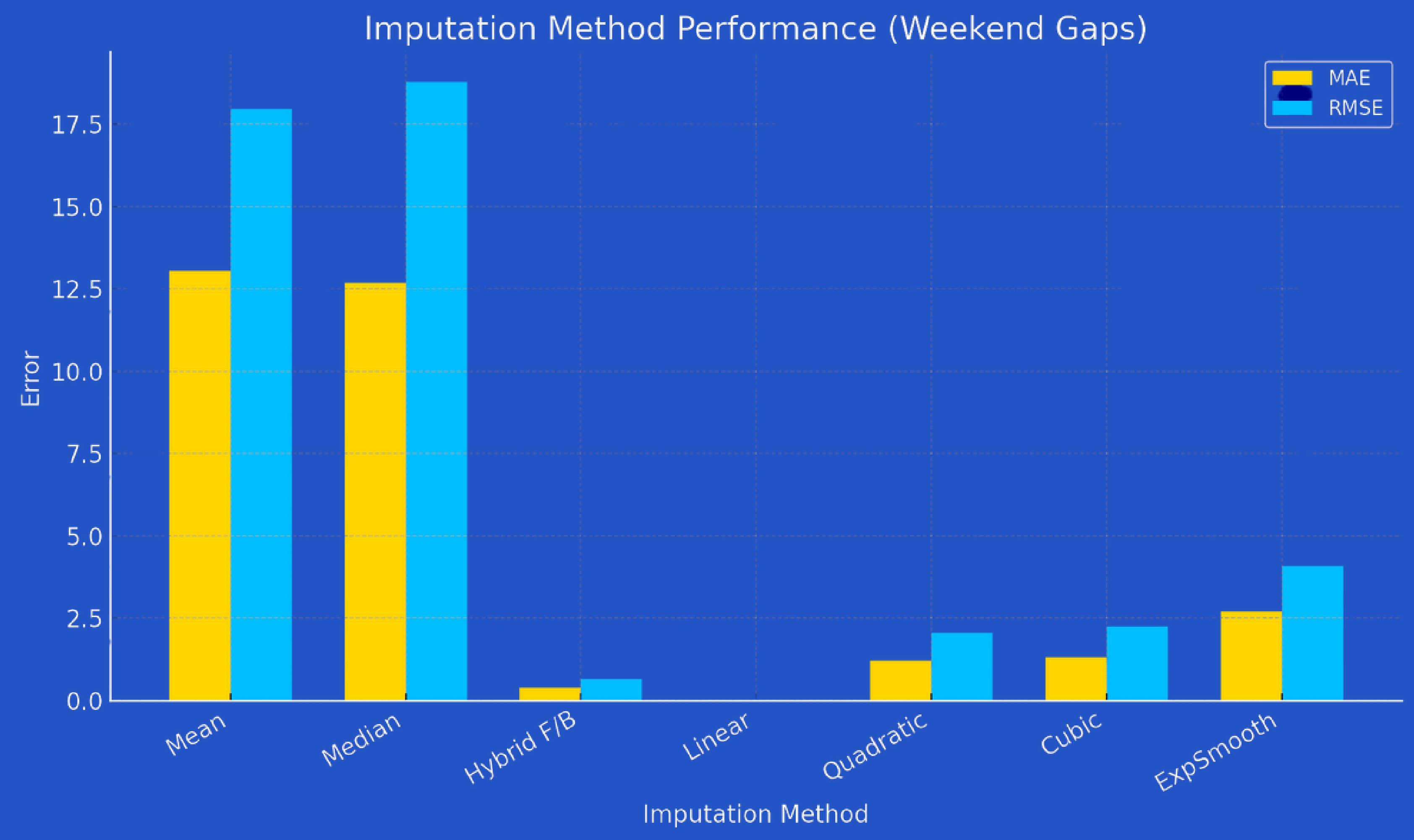
Rolling windows tested (3–60 days) against weekend reference

- **3-day rolling mean** gave best accuracy (MAE ≈ 1.76, RMSE ≈ 2.86)
- **Longer windows** smoothed data but missed short-term signals



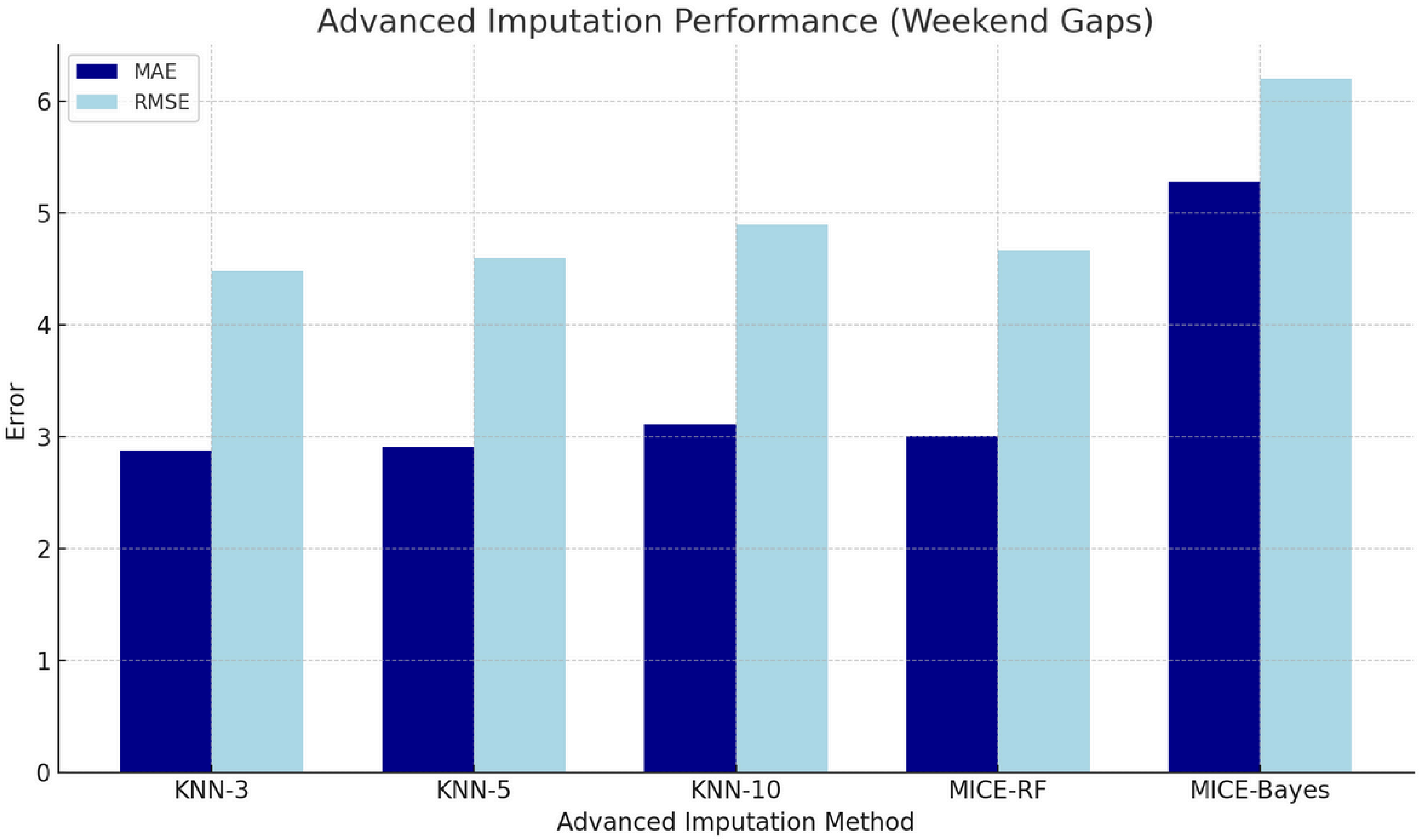
Evaluating Imputation Methods

- Compared accuracy using **MAE, RMSE, MAPE, sMAPE**
- **Mean/median**: weak performance (MAE > 12, RMSE ≈ 18)
- **Hybrid forward/backward fill**: strong baseline (MAE ≈ 0.40, RMSE ≈ 0.66)
- **Linear interpolation**: nearly perfect fit for 2-day weekend gaps (MAE/RMSE ≈ 0)



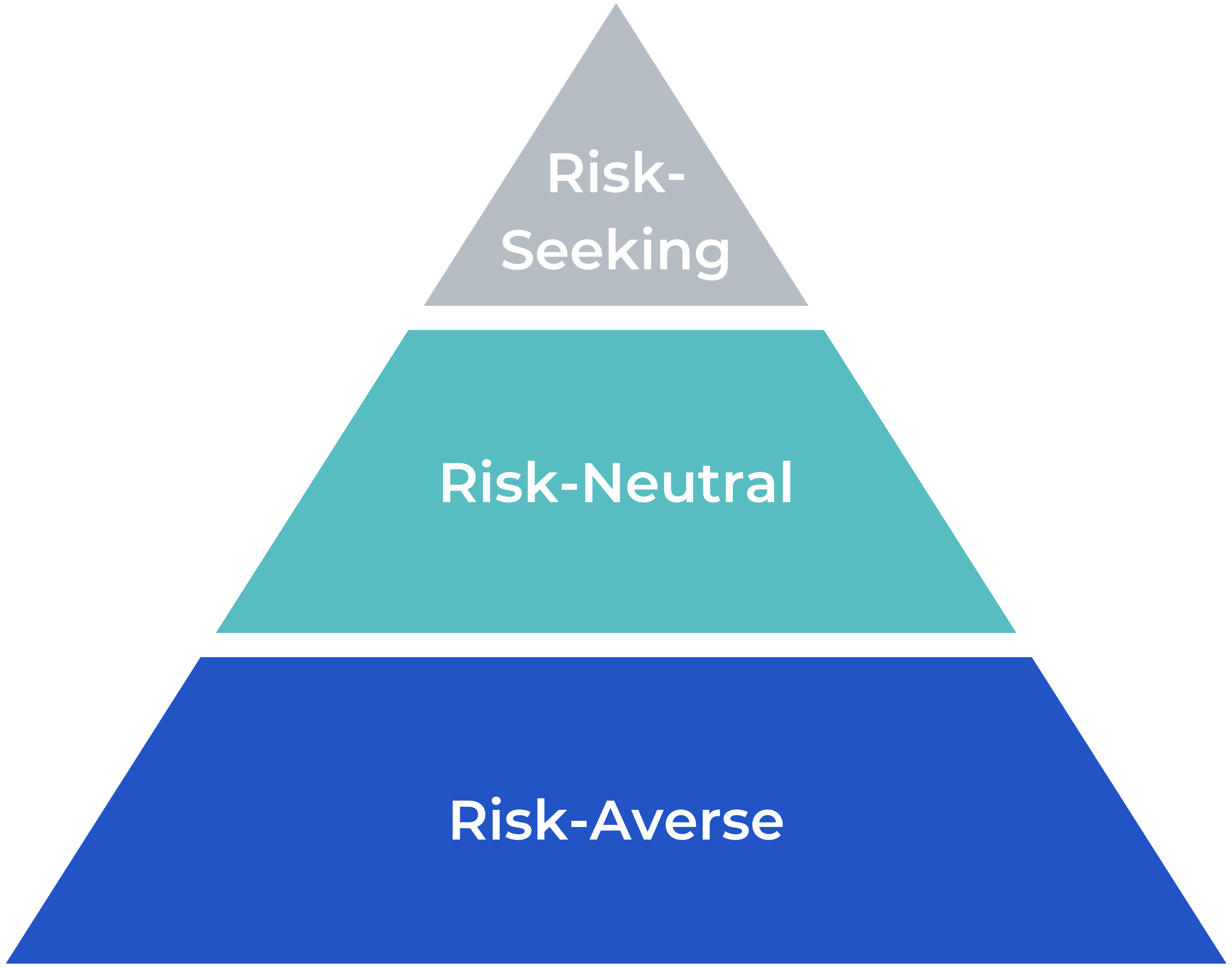
Advanced Techniques & Safeguards

- **Guardrails:** denominator floor + sMAPE → no infinities/NaNs
- **Correlations** recomputed valid **within [-1, 1]**; clipping safeguard added
- Advanced methods:
  - **KNN (k=3)**: strong results (MAE ≈ 2.88, RMSE ≈ 4.48)
  - MICE (**RandomForest**): effective but heavier
  - **Bayesian Ridge**: underperformed on sharp weekend moves



Utility Functions & Imputation Choices

- **Risk-Seeking**
  - Value upside potential, embrace volatility and noise
  - Imputation style: **extrapolation across correlated assets** (e.g. ETH from BTC/SOL), tolerate bias for higher gains
- **Risk-Neutral**
  - Linear view of outcomes, focus on expected return
  - Imputation style: **mean or regression-based fills** (balanced, average-oriented)
- **Risk-Averse**
  - Prefer capital preservation, dislike volatility
  - Imputation style: **median fills** (conservative, robust to outliers)



Final Dataset & Returns

Chosen method: Linear Interpolation (validated via Monte Carlo)

- Robust against repeated resampling, lowest errors across trials
- Best fit for 2-day weekend gaps (Fri → Mon)

Simple & log returns calculated → log more stable

Positive Sharpe ratio despite high variance

High volatility from crypto/tech balanced by defense & infrastructure

Thank You!

Hang tight 9 sprints left