Numeraxial

Advanced Data Preprocessing Techniques for Financial & Economic Data

1. Data Cleaning & Alignment

Handling Missing Data

- Forward/Backward fill for time-series gaps.
- Model-based imputation (Kalman filters, EM for state-space models).
- Multiple imputation (Bayesian approaches for macroeconomic series).

Timestamp Alignment

- Align heterogeneous data frequencies (e.g., daily stock prices with monthly macro indicators).
- As-of joins for tick-by-tick vs. daily aggregates.

Corporate Actions Adjustments

- Price adjustment for dividends, stock splits, mergers, ticker changes.
- Rolling adjustment factors for continuity in historical series.

2. Noise Reduction & Signal Extraction

Smoothing Filters

Moving average, Savitzky–Golay filters for trend preservation.

Fourier & Wavelet Transforms

Multi-resolution decomposition for denoising price series.

Empirical Mode Decomposition (EMD)

Decompose non-linear/non-stationary signals into Intrinsic Mode Functions.

Kalman Filtering

Real-time noise reduction and latent state estimation.

3. Stationarity & Transformation

Detrending & Differencing

Log returns (instead of raw prices).
Seasonal decomposition for macroeconomic data.

Normalization & Scaling

Volatility scaling (returns divided by realized volatility).
Z-score scaling for features with different magnitudes (e.g., rates vs. sentiment).
Box-Cox / Yeo-Johnson Transforms
Variance-stabilizing transformations for skewed economic data.

4. Feature Engineering & Enrichment

Market Features

- Volatility clustering (e.g., GARCH residuals as features).
- Higher-order moments (skewness, kurtosis of returns).

- Technical indicators (RSI, MACD, Bollinger Bands).

Macroeconomic Features

- Lagged effects (e.g., interest rates, CPI lags).
- Cyclical indicators (yield curve slope, credit spreads).

Alternative Data Integration

- Social sentiment (NLP preprocessing: embeddings, topic modeling).
- Satellite, ESG, shipping flows, Google Trends (scaled to market calendars).

Market Regimes

- Hidden Markov Models (HMM) or Bayesian Change Point Detection to label bull/bear/stagnant phases.

5. Outlier & Anomaly Handling

Winsorization / Clipping

Reduce impact of extreme tail events.

Robust Scaling

Median & interquartile-based scaling (instead of mean/variance).

Anomaly Detection

Isolation forests, robust PCA, or autoencoders for unusual price/volume/macro shocks.

6. Feature Selection & Dimensionality Reduction

Filter & Wrapper Methods

Mutual information, stability selection for economic variables.

PCA / ICA / RPCA

Remove redundancy from correlated indicators.

Manifold Learning

t-SNE, UMAP for clustering latent regimes.

Sparse Methods

Lasso/ElasticNet to select predictive macro factors.

7. Temporal & Structural Adjustments

Regime-Specific Preprocessing

Normalize/scale separately within regimes (bull vs. bear).

Time-Varying Correlations

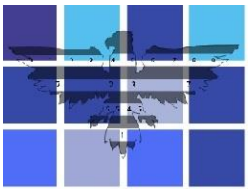
Dynamic Conditional Correlation (DCC-GARCH).

Rolling Window Features

Adaptive statistics (mean/volatility/correlation over rolling periods).

Event-Time Alignment

Align around earnings announcements, FOMC dates, recessions, policy changes.



Numeraxial

8. Cross-Sectional vs. Time-Series Treatment

Cross-Sectional Normalization

Rank or z-score standardization across assets each day.

Panel Data Handling

Fixed effects (sector/country dummies).

Random effects for heterogeneous asset panels.

9. Data Augmentation & Synthetic Data

Bootstrapping & Block Bootstraps

Preserve autocorrelation in returns for resampling.

Generative Models

GANs, VAEs for synthetic return/macro series.

Backtesting Augmentation

Scenario generation (stress events, fat-tail shocks).

10. Preprocessing Pipelines (Automation)

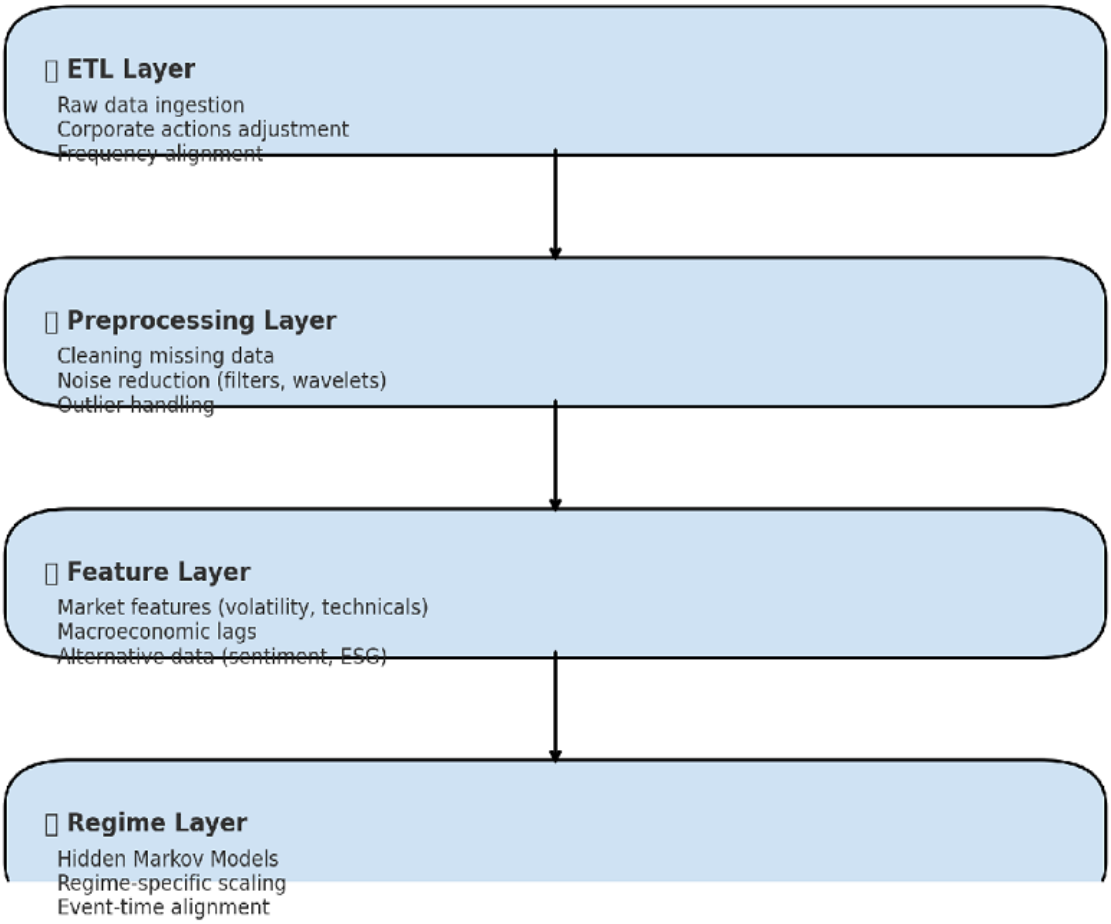
Modular pipelines with:

- **ETL Layer:** Raw data ingestion & corporate actions adjustment.
- **Preprocessing Layer:** Cleaning, scaling, alignment.
- **Feature Layer:** Market/macroeconomic/alternative features.
- **Regime Layer:** State-dependent preprocessing.
- **Model Input Layer:** Final normalized dataset for ML/RL agents.

This framework is the backbone of **quantitative trading, portfolio management, and macroeconomic modeling pipelines**.

Numeraxial

⚙ Advanced Preprocessing Pipeline for Financial & Economic Data



Model Input Layer

Stationarity transformations
Normalization
Final dataset for ML/RL agents