Financial Market Data: From Traditional Structures to AI-Driven Analytics

Exploring Data Types, Bar Structures, and Modern AI Applications

Advanced Financial Machine Learning

September 11, 2025

Understanding Financial Market Data Landscape

Fundamental Data

PBR, PER, ROE, sales metrics, often reinstated or backfilled

Market Data

Price, volume, open interest, spreads, cancellations, aggressor side

Analytics Data

Analyst recommendations, credit ratings, earnings expectations, news sentiment

Alternative Data

Social media, web search, satellite imagery, geolocation, transaction data

AI Enhancement

NLP for text analysis, computer vision for imagery, deep learning for pattern recognition

From Unstructured to ML-Ready: The Data Transformation Challenge

Volume & Variety

Massive amounts of heterogeneous financial data from diverse sources requiring specialized processing

Real-time Processing

Market data arrives at millisecond intervals requiring immediate analysis for timely decisions

Noise & Signal

Distinguishing meaningful patterns from market noise and random fluctuations

AI Solutions

NLP for text analysis, **Computer Vision** for imagery, **Deep Learning** for pattern recognition, **Transformers** for sequential data

Performance Gains

AI-driven preprocessing improves prediction accuracy by 15-30% compared to traditional methods

Evolution of Financial Data Sampling: Beyond Time Bars

Time Bars

Fixed time interval sampling (traditional approach)

Poor statistical properties: auto-correlation, non-normality, GARCH effects

Tick Bars

Record bars when a fixed number of transactions occur

Synchronizes sampling with information arrival

> *"Price changes over a fixed number of transactions may have a Gaussian distribution." — Mandelbrot & Taylor (1967)*

Volume Bars

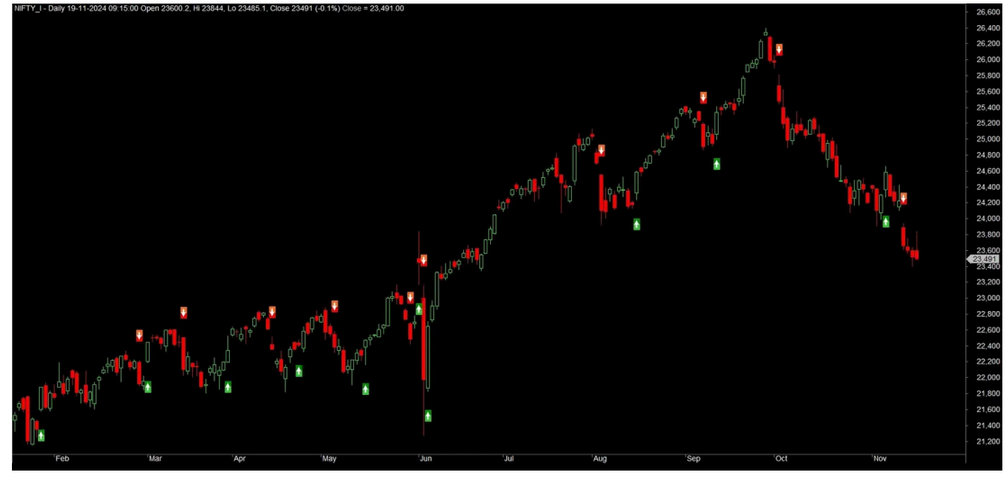
Record observation when pre-determined trading volume occurs

Better statistical properties than tick bars (Clark, 1973)

Dollar Bars

Record observation when pre-determined market value is transacted

Improved normality and reduced autocorrelation



Advanced Bar Construction: Capturing Market Microstructure

⚖ Tick Imbalance Bars (TIB)

Samples when tick imbalance exceeds expectations, capturing order flow dynamics $\theta_T = \Sigma b_i$

Volume Imbalance Bars (VIB)

Measures volume-weighted order flow imbalance, detecting large trader activity $\theta_T = \Sigma b_i v_i$

$ Dollar Imbalance Bars (DIB)

Tracks dollar value imbalance, ideal for high-value securities

$$T^* = \text{argmin}\ E[|\theta_T| \geq E(T)|2v^+ - E(v)]$$

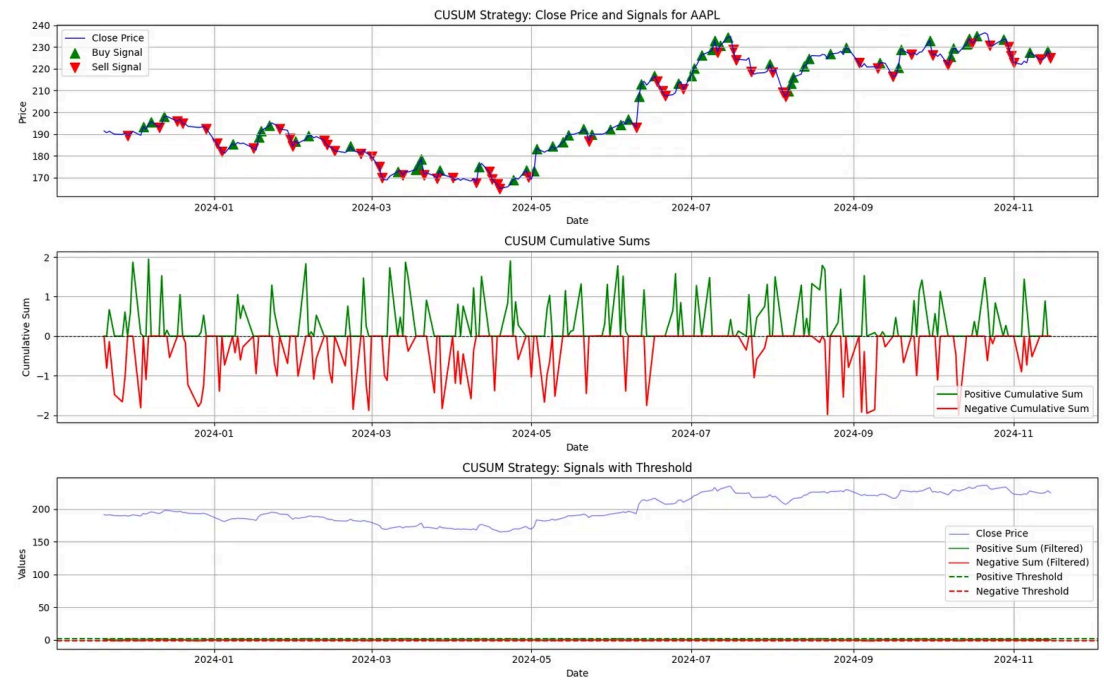
Run Bars

Detects sequences of buys/sells, revealing informed trading patterns and market sweeps

$$\theta_T = \max\{\Sigma b_t \mid b_t=1, \Sigma b_t \mid b_t=-1\}$$

AI Enhancement

Machine learning optimizes thresholds dynamically based on market conditions



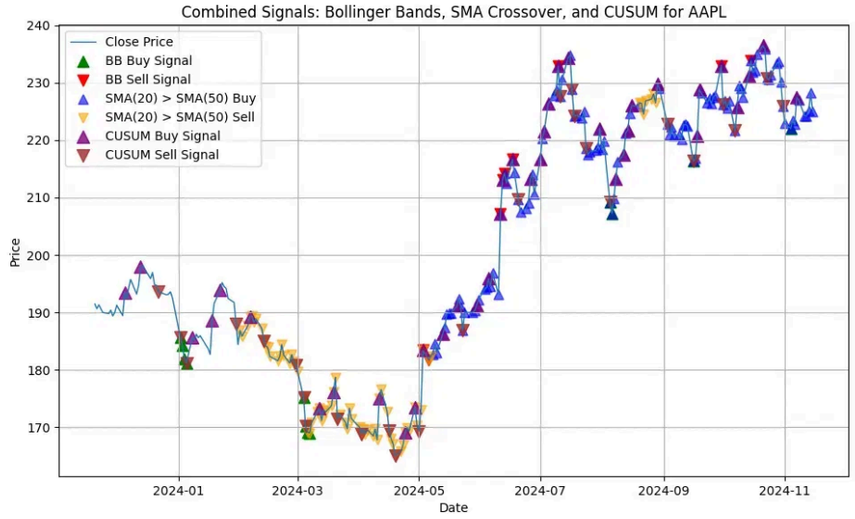
CUSUM Filters: Statistical Approach to Market Event Detection

CUSUM Methodology

Cumulative Sum filter detects shifts in mean value away from target: $S_t = \max\{0, S_{t-1} + y_t - E_{t-1}[y_t]\}$ Signals event when $S_t \geq$ threshold h

Advantages Over Bollinger Bands

- More robust to noise and false signals
- Better detection of persistent shifts
- Adaptive to market volatility



Event-Based Sampling

Sampling triggered by meaningful market events rather than arbitrary time intervals, capturing:

- Structural breaks
- Regime changes
- Microstructural phenomena

AI Enhancement

Deep learning for adaptive threshold setting and **reinforcement learning** for parameter optimization

Alternative Data Revolution: AI at the Forefront

Satellite Imagery

AI-powered analysis of oil storage levels, agricultural yields, retail parking lot traffic, and construction activity

Social Media & News

NLP for sentiment analysis, topic modeling, and event detection from Twitter, Reddit, and financial news

Geolocation Data

Foot traffic analysis, supply chain monitoring, and consumer behavior patterns from mobile devices

Transaction Data

Credit card spending patterns, invoice processing, and B2B transaction analysis for economic indicators

Performance Impact

Hedge funds using alternative data report **20-35% improvement** in alpha generation compared to traditional data sources

Neural Networks in High-Frequency Finance

LSTM Networks

Capture long-term dependencies in financial time series, ideal for sequential market data with memory effects

Convolutional Networks

Extract local patterns and features from market data, effective for detecting chart patterns and price formations

Transformer Models

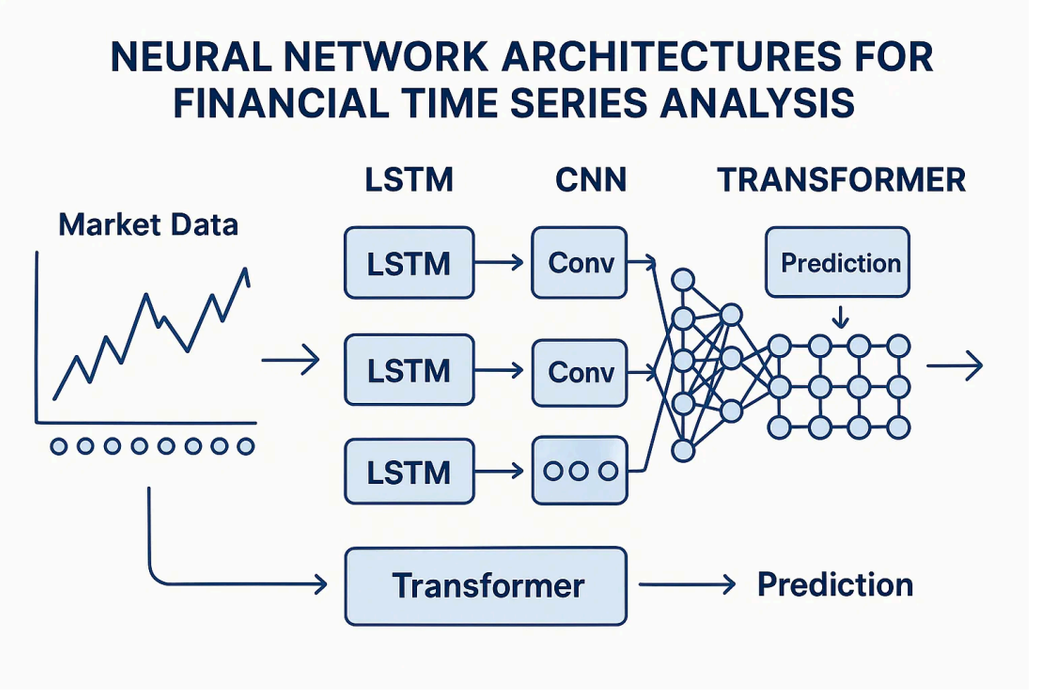
Self-attention mechanisms capture relationships between distant time points, outperforming traditional models for long-range dependencies

Hybrid Architectures

Combining CNN + LSTM + Transformer models to capture multi-scale patterns in market microstructure

Applications

Order flow prediction, **market making**, **price forecasting**, **volatility estimation**, and **anomaly detection**



The Future of Financial Data Analytics

Foundation Models

Large language models fine-tuned for financial domain, enabling advanced document analysis, market commentary, and multi-modal data integration

Quantum Computing

Exponential speedup for portfolio optimization, risk modeling, and Monte Carlo simulations, with early applications in derivatives pricing

Federated Learning

Privacy-preserving collaborative model training across financial institutions without sharing sensitive data, addressing regulatory constraints

Edge Computing

Ultra-low latency processing for high-frequency trading, bringing AI inference closer to data sources for millisecond-level decisions

Research Frontiers

Generative AI for synthetic data, **graph neural networks** for market relationships, **explainable AI** for regulatory compliance