RESEARCH NOTE RN-01 · DATA ENGINEERING & ECONOMETRIC HYGIENE

Point-in-Time Alternative Data Ingestion, Entity Resolution & *Survivorship-Free Panel Construction.*

The easiest way to fool yourself in quantitative finance is to test a strategy on data the market didn't actually have yet, or on a list of companies that only includes the survivors. Here is how we lock every filing to its down-to-the-second SEC `acceptanceDateTime` across 3,607 historical U.S. issuers—including 555 delisted companies.

AUTHOR: **STARLING QUANTITATIVE RESEARCH** PUBLISHED: **NOVEMBER 8, 2026** DESK: **STARLING QUANTITATIVE**

SECTION 01

Three Ways Alternative Data Backtests Accidentally Cheat

Almost every signal looks great in a naive backtest and then falls apart when real money is on the line. When you audit why alternative data strategies fail in production, the culprit is almost never the math—it is data contamination. Three specific leaks account for most of that gap:

1. **The 4:15 PM Filing Trap & Restatement Overwrites:** If you align quarterly numbers by fiscal period-end (fp), you are trading on information 30 to 90 days before it was public. Even using the calendar filingDate introduces a full day of lookahead bias whenever a company files its 10-K, 10-Q, 8-K, or Form 4 after the 16:00 ET closing bell (say, at 16:15 ET), because a daily backtest assumes you bought at the 16:00 ET close—15 minutes before the filing hit the wire. Worse, standard data vendors routinely overwrite original quarterly numbers when a company files a restatement months later.

2. **Mutable Vendor Timestamps:** Many public and commercial data feeds only give you a single last_updated date instead of an immutable event log. If a public registry updates a minor administrative field six months after an event, pulling that record today makes it look as if the original catalyst was missing—or arrived on a completely different date.

3. **LLM Pre-Training Contamination:** If you ask a commercial language model (trained on internet data through 2025) to read a 2023 earnings release and score whether the stock looks bullish, the model already knows how the stock performed in 2023 and 2024. You aren't testing a signal; you are testing the model's memory.

SECTION 02

Down-to-the-Second SEC acceptanceDateTime Alignment & Feed Quarantine

To make sure our system only sees what a trader sitting at a desk could see at that exact second, our ingestion layer enforces a strict point-in-time filtration $\mathcal{F}_t$. Every structured XBRL line item in SEC EDGAR companyfacts (revenue, operating cash flow, working-capital accruals, share counts, and ASC-280 customer footnotes) is joined by accession number (accn) to its EDGAR submissions header to recover the exact down-to-the-second SEC server timestamp τ_{accept} (acceptanceDateTime) and stamped with normalized **One-Document-One-Witness** provenance (source_artifact_id = SEC:<18-digit-accession>) so downstream calibration never double-counts multiple parsers reading the same filing:

$$\mathcal{D}_{i,t} = \left\{ d_{i,k} \;\middle|\; \tau_{\text{accept}}(d_{i,k}) \leq \tau_{\text{cutoff}}(t) \right\}$$

That timestamp routes every disclosure into one of three execution windows: (i) filings accepted before the opening bell ($\tau_{\text{accept}} < 09{:}30{:}00\ \text{ET}$) can trade at the same-day $\text{Open}(T_0)$; (ii) regular-session filings ($09{:}30{:}00\ \text{ET} \leq \tau_{\text{accept}} < 16{:}00{:}00\ \text{ET}$) are benchmarked against live arrival mid-quotes (ArrivalMid) in our intraday daemon and pushed conservatively to $\text{Open}(T+1)$ in daily batch replay; and (iii) after-hours filings ($\tau_{\text{accept}} \geq 16{:}00{:}00\ \text{ET}$) are pushed strictly to the next trading day's $\text{Open}(T+1)$. If a company later files an amendment (10-K/A or 10-Q/A) at $\tau_{\text{amend}} > \tau_{\text{accept}}$, the historical ledger keeps the original unrevised number over $[\tau_{\text{accept}},\ \tau_{\text{amend}})$ and reveals the amended figure only from τ_{amend} onward.

Quarantining Feeds That Can't Prove Their History: We apply a simple rule to every data feed: if we cannot verify its historical timestamps down to the minute from an immutable primary archive, we do not allow it to promote itself in a backtest. Scouts backed by immutable primary filing timestamps (SEC EDGAR acceptanceDateTime, Form 4 XML ownership disclosures, Schedule 13D XML filings, Form 8-K ASR contracts, 10-K ASC-280 footnotes, FINRA short volume, and exchange OHLCV bars) are eligible for historical walk-forward cross-validation. Any scout whose historical archive relies on mutable vendor watermarks (such as ClinicalTrials.gov lastUpdatePostDateStruct, quarantined at $0 in SHADOW) or retrospective monthly batch files (such as historical FDA 510(k) bulk downloads) is **quarantined from historical backtest promotion**; in forward production, our daemon polls intraday registry deltas and logs exact point-in-time arrival timestamps so those sleeves can only graduate by proving themselves out-of-sample under hard canary caps.

SECTION 03

Point-in-Time Entity Resolution: 10-Digit SEC CIK, Composite OpenFIGI & Ticker Splicing

Ticker symbols are neither unique nor permanent across multi-year equity panels: bankrupt issuers relinquish tickers that are subsequently reassigned to unrelated listings, and corporate spin-offs or

mergers alter symbol mappings mid-sample. Joining alternative disclosures to price series by raw ticker string creates silent cross-entity contamination.

Starling Quantitative resolves every corporate filing, regulatory action, and supply-chain edge at ingestion time to a canonical **10-digit zero-padded SEC Central Index Key (CIK)**, a **composite point-in-time OpenFIGI identifier (BBG...)**, and **point-in-time ticker-regime splicing** (tracking historical symbol-to-CIK validity intervals across corporate actions and delistings for deterministic institutional security-master joins). Cross-entity supply-chain graphs extracted from Form 10-K Item 1 and ASC-280 major-customer footnotes link upstream suppliers i to downstream principals j via verified (CIK_supplier, CIK_principal, tau_accept) edges, ensuring that customer-distress or inventory-glut propagation signals reflect only relationships publicly disclosed prior to evaluation date t.

SECTION 04

Survivorship-Free Historical Universe Construction & Shumway Delisting Haircuts

Evaluating fundamental or event-driven strategies across today's surviving index members systematically truncates left-tail bankruptcy and distressed-acquisition outcomes. To eliminate survivorship bias, Starling's historical validation panel is constructed from the union of historical SEC EDGAR registrant snapshots and broad-market constituent histories from January 2019 through November 2026 (comprising 505 trading days of rolling estimation warmup prior to out-of-sample evaluation):

PANEL DIMENSION	SPECIFICATION	ECONOMETRIC SAFEGUARD
Total Historical Universe	**3,607 Unique CIKs**	Broad U.S. liquid equity universe (Russell 3000 historical union)
Non-Surviving / Delisted Issuers	**555 Delisted CIKs (15.4%)**	Retains bankruptcies, liquidations, mergers & exchange delistings in-sample
Estimation Warmup Buffer	505 Trading Days	252d rolling factor regressions + trailing rank/MAD baselines before Fold 1
Terminal Distress Haircut	Shumway (1997) -30.0%	Applied automatically on unscheduled performance-related delisting events

Consistent with Shumway (1997), whenever a constituent experiences an unscheduled performance-related delisting without an executable terminal closing print, the backtest engine injects an explicit -30.0% delisting return penalty into forward return attribution, preventing distressed value traps from exiting historical simulations at artificially benign last-traded prices. In parallel, an automated OHLCV scrubber cleans 2-day bounce-back bad ticks and unadjusted forward/reverse stock-split discontinuities prior to rolling covariance and Barra 2-Factor (Market + GICS Sector) residual-volatility estimation.

SECTION 05

Deterministic Feature Extraction & Dual Institutional Delivery Schemas

Every stage of Starling Quantitative—across all specialized filing scouts, isolated sub-arms (arm_id), point-in-time primary data harvesters, non-parametric calibration, statutory distress vetoes, and portfolio allocation—is implemented in **100% deterministic quantitative Python** with zero generative LLM calls anywhere in the loop. Given identical point-in-time filing inputs $\mathcal{D}_{i,t}$, the pipeline produces bit-for-bit identical outputs across historical replay and live production.

Because institutional systematic desks differ in whether they consume raw alternative feature vectors inside their own portfolio optimizer or finished risk-gated trade signals, Starling exposes two distinct machine-readable schemas:

1. **Feed Tier A — Pre-Quorum Point-in-Time Feature Stream (ScoutPitch + Institutional Export Envelope):** Emits un-netted, scout-level records containing top-level ScoutPitch fields (pitch_id, scout_name, target_ticker, instrument_to_buy, action, holding_period_days, stop_loss_pct, take_profit_pct, and metric_evidence) wrapped with standardized institutional Parquet/JSONL export envelope columns (cik, openfigi, arm_id, source_artifact_id, action_polarity ["LONG" vs. "SHORT_OR_LONG_VETO"], acceptance_datetime_et, acceptance_datetime_utc, executable_bar_ts, and raw_payload_sha256) alongside standardized inner metric_evidence keys (pillar_id, feature_domain, primary_metric_name, primary_metric_value, norm_mode, polarity, as_of_ts, panel_n_common, half_life_days, z_event $z_{k,i,t}$, composite_z, purged_cv_tier, ic_post, sigma_resid_1od_bps, alpha_bps, post_accept_resid_move_bps, alpha_rem_bps, rt_cost_bps, and net_alpha_rank_bps). Crucially, Tier A preserves the exact target_ticker on both positive and negative (bearish) corporate catalysts—even when our internal book routes bearish forensic or product-recall disclosures strictly into $o long-veto shields (with zero single-stock inverse-ETF routing)—allowing institutional long/short equity desks to harvest single-name short alpha directly using prime-broker borrow.

2. **Feed Tier B — Post-Quorum Calibrated Consensus Stream (MasterDecision):** Emits the post-calibration, post-veto signal vector after One-Document-One-Witness deduplication, 4-pillar concave pooling behind the One-Way Canary Glass Wall, statutory/credit distress kill-switches, 6-factor Barra residualization, and volatility-parity sizing (detailed in Research Note RN-02 and Research Note RN-03).

SECTION 06

Selected References & Technical FAQ

• Bernard, V. L., & Thomas, J. K. (1989). Post-earnings-announcement drift: Delayed price response or risk premium? *Journal of Accounting Research*, 27, 1–36.

• Cohen, L., & Frazzini, A. (2008). Economic links and predictable returns. *Journal of Finance*, 63(4), 1977–2011.

• Shumway, T. (1997). The delisting bias in CRSP data. *Journal of Finance*, 52(1), 327–340.

• Sloan, R. G. (1996). Do stock prices fully reflect information in accruals and cash flows about future earnings? *The Accounting Review*, 71(3), 289–315.

Why is indexing SEC filings by period-of-report or filingDate insufficient for quantitative backtesting?

Indexing financial statements by fiscal period-end (fp) introduces 30 to 90 days of lookahead bias, while calendar filingDate ignores whether a disclosure arrived before the 16:00 ET market close or after hours. Starling joins SEC XBRL companyfacts items directly to EDGAR submissions metadata by accession number to enforce down-to-the-second acceptanceDateTime gating, shifting post-16:00 ET filings to $T + 1$.

How does Starling handle alternative feeds whose vendor APIs overwrite historical timestamps?

Feeds with mutable vendor watermarks or delayed batch aggregates are quarantined from historical walk-forward cross-validation unless anchored to immutable primary filing timestamps, and are evaluated exclusively in forward out-of-sample shadow ledgers.

What is the difference between Feed Tier A and Feed Tier B in Starling's institutional schema?

Feed Tier A exports pre-quorum ScoutPitch records wrapped in an Institutional Export Envelope (cik, openfigi, arm_id, source_artifact_id, action_polarity, acceptance_datetime_et, acceptance_datetime_utc, executable_bar_ts, raw_payload_sha256) with typed metric_evidence dictionaries preserving the true point-in-time target_ticker (including bearish single-name short signals prior to any internal long-veto routing, with zero single-stock inverse-ETF routing) for institutional long/short factor ingestion. Feed Tier B exports post-quorum MasterDecision records after One-Document-One-Witness deduplication, 4-pillar pooling behind the One-Way Canary Glass Wall, statutory distress veto checks, and portfolio covariance gating.

NEXT RESEARCH NOTE · RN-02

© 2026 STARLING QUANTITATIVE · A BIWADA COMPANY

REGULATORY DISCLOSURE: Starling Quantitative is a proprietary quantitative research and internal capital allocation desk of Biwada LLC. Nothing in this monograph constitutes investment advice or an offer to solicit outside retail capital or manage third-party funds.