Northeastern Systematic Alpha

# Alternative Data Scraping

News Prominence & S&P 500 Hiring Signals


**Filip Ilic · Sean May · Tyme Sutespoon**

Quantitative Development Team | April 2026

# The Problem

## Q: Why build our own alternative-data pipeline?

Alternative data — news sentiment, editorial prominence, hiring patterns — is increasingly central to systematic strategies. However, commercial providers are expensive, delayed, and pre-filtered. The raw signals we need are not available off the shelf.

- **Commercial APIs cost a fortune.** Bloomberg and Refinitiv price out student research teams.
- **BLS hiring data lags 30–60 days.** By the time the release lands, the signal is priced in.
- **No product tracks editorial prominence.** Nobody captures which stories get homepage priority over time.
- **S&P 500 hiring data is fragmented.** 500+ career sites across 15 different ATS providers.
- **Monetary incentive is misaligned.** Vendors sell what packages well — not what moves markets.

# The Pipeline

## Q: What does the system actually do?

**30**

**News sources**

scraped 3× daily for homepage prominence

**470**

**S&P 500 companies**

successfully scraped across 15 ATS providers

**~$10**

**per month infra**

fully serverless, scales to Fortune 1000

---

**1. Collect** Schedulers fire 3× daily — pull RSS, homepages, ATS job boards.

**2. Normalize** Classify finance relevance, extract position, dedupe, timestamp.

**3. Store** Postgres with immutable snapshots — ready for analysis.

# News — Hypotheses

**Q: What testable signals live in the news data?**

## H1 · Front-Page Prominence

Where a story sits on the homepage signals how much attention editors expect it to get.

- Capture homepage rank 3× daily as a time series.
- Rising prominence → editors think it matters more.
- Test: position-delta vs. next-day return and volume.
- Needs 3–6 months of snapshots for significance.

## H2 · Story Convergence

When many outlets tell the same story at once, that agreement is itself a market signal.

- 30 sources across the editorial spectrum.
- Measure how fast a narrative spreads.
- Test: convergence metric vs. sector volatility.
- Needs topic extraction (on roadmap).

# Jobs — Hypotheses

**Q: What testable signals live in the hiring data?**

## H3 · Hiring Velocity

How fast a company is hiring leads earnings by a quarter — visible in job boards before filings.

- 470 S&P 500 companies scraped live.
- 30–60 days ahead of BLS reports.
- Test: sector velocity vs. earnings surprises.
- Needs 6+ months for cross-sectional study.

## H4 · Wage Pressure

Posted salary ranges across the S&P 500 form a real-time wage index by sector and role.

- Pay bands parsed directly from listings.
- Wage acceleration squeezes margins.
- Test: wage delta vs. sector margin change.
- Coverage: ~30–50% of listings include pay.

# Problem & Approach

**Q: Where are the commercial data gaps, and how do we fill them?**

## The Problem

- Commercial APIs are pre-filtered and expensive.
- BLS jobs data is stale by 30–60 days.
- No product tracks editorial homepage prominence.
- S&P 500 hiring data fragmented across 500+ sites.
- Monetary incentive misaligned with research needs.

## Our Approach

- Scrape raw homepages, 30 outlets, 3× daily.
- Track article position as a time series.
- Auto-discover ATS endpoints across 15 providers.
- Fully serverless, ~$10/month.
- Own the data — no vendor gatekeeping.

# Coverage

## Q: How much data do we actually have?

**30**

**news sources**

general + finance

**470**

**S&P 500 live**

of 503 tested

**15**

**ATS providers**

Workday, Greenhouse, +13

**93%**

**success rate**

470 / 503 companies

---

## Notable

- ATS breakdown: Workday 245 · SuccessFactors 55 · iCIMS 31 · OracleCloud 27 · SmartRecruiters 21 · Greenhouse 17 · Phenom 15 · Eightfold 14 · Taleo 9 · Lever 7 · Ashby 6 · Avature 3 · Dayforce 1 · Jobvite 1 · custom 18.
- News: 6,141 articles captured · 6,388 finance snapshots classified · 12 runs across 10 days.
- Top volume: Barron's (900), Guardian (458), AP (430), HuffPost (392).

# Architecture

## Q: How do the two pipelines fit together?

**NEWS**

**Scheduler**
3× daily

**Scraper**
30 sources

**Classifier**
Finance tag

**Postgres**
Snapshots

**JOBS**

**S&P 500 Fetch**
Weekly

**ATS Finder**
Auto-discover

**15 Scrapers**
Workday, GH…

**Postgres**
Unified jobs

---

**470**

S&P 500 scraped

**30**

news outlets, 3× daily

**~$10/mo**

serverless infra

# Why This Matters for NUSA

Q: What does building our own data pipeline mean for the club?

## Leveling the Playing Field

- Vendors like Revelio Labs charge $5K–$50K/yr for hiring signal data alone
- We build and own our version at infrastructure cost only (~$10/month)

## A Dataset That Compounds

- Each semester adds statistical depth.
- The pipeline supports 15 ATS providers today, with potential for more.
- Future members inherit a live, growing research asset

## Testing Our Hypotheses

- H3 + H4: Hiring velocity and wage pressure visible 30–60 days before earnings filings
- H1 + H2: News prominence and story convergence tracked as quantitative time series

## Real Engineering Experience

- AWS Lambda, PostgreSQL, Docker
- Members graduate with hands-on infrastructure experience that stands out in recruiting

# Appendix

# A1: Case Study — Tesla Q1 Deliveries

## Q: Does this pipeline actually capture real market narratives as they unfold?

One event — TSLA Q1 2026 delivery release — traced across 8 homepage snapshots and 9 articles over 36 hours.

| Time (EST) | Outlet | Headline Fragment | Position | Phase |
|---|---|---|---|---|
| **Apr 2, 08:00** | Reuters | Analysts eye weakest Tesla quarter in 2 years | **#14** | Pre-event |
| **Apr 2, 16:00** | Bloomberg | Tesla delivery estimates walked back again | **#9** | Pre-event |
| **Apr 3, 00:00** | **Tesla IR** | **(Tesla publishes Q1 delivery numbers — 386k vs 408k est.)** | **—** | **Release** |
| **Apr 3, 08:00** | Reuters | Tesla reports weakest quarter, shares slide pre-market | **#3** | Reaction |
| **Apr 3, 08:00** | Bloomberg | Delivery miss reignites demand questions | **#5** | Reaction |
| **Apr 3, 08:00** | WSJ | Tesla Q1 falls short of Wall Street | **#6** | Reaction |
| **Apr 3, 08:00** | FT | Tesla misses deliveries amid EV slowdown | **#8** | Reaction |
| **Apr 3, 16:00** | Reuters | Analysts cut Tesla price targets after miss | **#11** | Reaction |
| **Apr 4, 08:00** | Barron's | Is Tesla's growth story broken? | **#18** | Decay |

**What this shows**

Full narrative arc — anticipation, release, reaction, decay — captured end-to-end in our snapshots.

Peak 9 simultaneous articles across outlets. Position #3 at reaction marks maximum editorial prominence.

Hillert et al. (2014) + Boudoukh et al. (2019) predict momentum amplification at exactly this moment.

# A2: Build vs Buy — Commercial Alternatives

## Q: Why not just pay for Revelio, Lightcast, LinkUp, or Bloomberg?

| Vendor / Source | Cost / yr | Latency | Coverage | Customization | Bias Control |
|---|---|---|---|---|---|
| **Bloomberg Terminal** | $28,000 | Real-time | Global news, all markets | Read-only | Pre-filtered |
| **Refinitiv (LSEG)** | $22,000 | Real-time | Global news, fundamentals | Read-only | Pre-filtered |
| **Revelio Labs** | $50k+ | Daily | Workforce, 1B+ profiles | API access | N/A — hiring only |
| **Lightcast (BG)** | $30k+ | Daily | 2.5B historical job posts | API access | N/A — hiring only |
| **LinkUp** | $20k+ | Daily | S&P 500 jobs index | Index-level only | N/A — hiring only |
| **Indeed Hiring Lab** | Free | Weekly | Aggregate indices only | None | N/A — hiring only |
| **BLS / JOLTS** | Free | 30–60 days | Sector-level | None | N/A |
| **Our Pipeline** | **≈ $150** | **8 hours** | **30 news + 470 S&P 500** | **Full — we own the data** | **Balanced, explicit** |

**Takeaway**

Commercial feeds are pre-filtered by vendors — we never see what they cut. Owning the pipeline means owning the editorial decisions — and paying less than a single Bloomberg seat costs for an hour.

# A3: News Source Breakdown

## Q: Which sources are scraped, and how are they chosen to balance editorial bias?

### General News · 15 sources, all-topic homepage tracking

| | |
|---|---|
| **NYTimes** | left / establishment daily |
| **CNN** | lean-left cable network |
| **Washington Post** | left-leaning national |
| **The Guardian (US)** | left international |
| **HuffPost** | left-leaning digital |
| **Slate** | left-leaning magazine |
| **BBC News** | center international |
| **Associated Press** | center wire service |
| **Forbes** | center-right business |
| **Wall Street Journal** | center-right daily |
| **NY Post** | right-leaning tabloid |
| **Fox News** | right-leaning cable |
| **Newsmax** | right-leaning cable |
| **Washington Examiner** | right-leaning daily |
| **Breitbart** | right-leaning digital |

### Finance Sources · bias-balanced, 5 tiers

**LEFT**

ProPublica · The Intercept · Common Dreams

**LEAN LEFT**

Bloomberg · Business Insider · Yahoo Finance

**CENTER**

Reuters Business · MarketWatch · Financial Times

**LEAN RIGHT**

The Economist · Barron's · Investor's Business Daily

**RIGHT**

Fox Business · Daily Caller · Breitbart Economy

Popularity so far: Barron's (900) · Guardian (458) · AP (430) · HuffPost (392) · WSJ · Bloomberg · Reuters. Taxonomy: AllSides media-bias ratings.

# A4: ATS Providers & Company Coverage

## Q: Which S&P 500 companies are scraped, and by which ATS platform?

**503 S&P 500 companies tested · 470 successfully scraped · 15 ATS providers · 92.6% overall success rate**

| PROVIDER | SUCCESS | RATE | EXAMPLE COMPANIES |
|---|---|---|---|
| **Workday** | 245 / 246 | 99.6% | Apple · Microsoft · Google · IBM · Cisco · Adobe · AT&T · 3M · Accenture · Bank of America (+235 more) |
| **SuccessFactors** | 55 / 58 | 94.8% | Altria · Ball · CMS Energy · A.O. Smith · American Water Works (+50 more) |
| **iCIMS** | 31 / 34 | 91.2% | Costco · Dollar General · AMD · EchoStar · GE Healthcare (+26 more) |
| **Oracle Cloud** | 27 / 31 | 87.1% | AutoZone · Carnival · Chubb · Citizens Financial · Cummins (+22 more) |
| **SmartRecruiters** | 21 / 21 | 100% | AbbVie · Arista Networks · Domino's · Expedia (+17 more) |
| **Custom** | 18 / 26 | 69.2% | Amazon · Alphabet · ADP · Crown Castle (+14 more) |
| **Greenhouse** | 17 / 17 | 100% | Airbnb · Block · Coinbase · DoorDash · Databricks (+12 more) |
| **Phenom** | 15 / 17 | 88.2% | Hilton · Lowe's · Newmont · GE HealthCare (+11 more) |
| **Eightfold** | 14 / 16 | 87.5% | American Express · Ford · Kraft Heinz · Lam Research (+10 more) |
| **Taleo** | 9 / 15 | 60% | Legacy Oracle Taleo deployments |
| **Lever** | 7 / 7 | 100% | Berkshire Hathaway · MetLife · Palantir (+4 more) |
| **Ashby** | 6 / 6 | 100% | Union Pacific · Parker Hannifin · Arthur J. Gallagher (+3 more) |
| **Avature** | 3 / 6 | 50% | CBRE · Jack Henry · Lululemon |
| **Jobvite** | 1 / 2 | 50% | Two deployments |
| **Dayforce** | 1 / 1 | 100% | Single deployment |

# A5: Data Quality & Known Limitations

Q: Where does the data fall short, and what do we already know is missing?

## What Works

- **News deduplication.** Composite key (source, articleid) prevents duplicate ingestion.
- **Jobs deduplication.** UNIQUE(cid, ats_provider, ats_job_id) enforced at DB layer.
- **RSS parsing.** 30 sources with consistent, reliable feed formats.
- **Workday scraping.** 245 / 246 companies — nearly perfect capture rate.
- **Perfect-rate providers.** Greenhouse, SmartRecruiters, Lever, Ashby: 100% success.
- **ATS auto-discovery.** Brave Search + URL classifier finds career pages without manual config.
- **Immutable snapshots.** Every scrape is append-only — back-in-time queries work natively.
- **Schema stability.** Articles + snapshots schema unchanged since Day 1, Alembic-managed on jobs side.

## Known Gaps

- **Salary coverage ~30–50%.** Only a fraction of listings publish pay ranges; sector-biased distribution.
- **Taleo scraping 60%.** Oracle Taleo templates are inconsistent; 6 of 15 deployments fail.
- **Avature / Jobvite 50%.** Small-N providers, two failures each.
- **Custom sites 69%.** 8 of 26 bespoke career portals resist the generic classifier.
- **No ground-truth validation.** Capture completeness is estimated, not verified against an authority.
- **33 companies un-scraped.** 503 tested minus 470 success = 33 failures across ATS providers.
- **No historical backfill.** Data only exists from scraper start date forward — no retrospective.
- **NLP not yet applied.** Finance classifier is keyword/URL based; sentiment + topics still on roadmap.

# A6: Sample Data — What a Row Looks Like

**Q: What does the captured data actually look like once it lands in Postgres?**

### articles (news_homepage.articles)

```
source              = "bloomberg"
articleid           = "fin_20260413_045"
title               = "Oil Surges 3% After Strait of Hormuz Incident"
byline              = "Alex Longley, Bill Lehane"
url                 = "bloomberg.com/news/articles/2026-04-13/..."
published_at        = 2026-04-13 14:55:00+00
summary             = "Brent crude jumped to $92/barrel after..."
tags                = {"energy","oil","geopolitics","commodities"}
```

### finance_snapshots (position over time)

```
id                    = 8234
source                = "bloomberg"
articleid             = "fin_20260413_045"
scraped_at            = 2026-04-13 16:00:00+00
section               = "Markets"
global_position       = 3    ← moved up from 7 in prior snapshot
section_position      = 1
article_updated_at    = 2026-04-13 15:45:00+00
```

### companies (jobs.companies)

```
cid                   = 550e8400-e29b-41d4-a716-446655440000
name                  = "Microsoft Corporation"
ticker                = "MSFT"
in_sp500              = true
ats_provider          = GREENHOUSE
jobsite_url           = "boards.greenhouse.io/microsoft"
ats_searched_at       = 2026-04-13 08:00:00+00
ats_search_attempts   = 5
```

### jobs (jobs.jobs)

```
jid             = 660e8400-e29b-41d4-a716-446655440001
cid             = 550e8400-e29b-41d4-a716-446655440000
title           = "Senior Software Engineer, Azure Infra"
location_raw    = "Redmond, WA"
employment_type = FULL_TIME
pay_salary_low  = 158000      pay_salary_high = 246000
currency        = USD         status          = OPEN
posted_at       = 2026-04-01  first_seen_at   = 2026-04-01
last_seen_at    = 2026-04-13  ats_job_id      = "gh_12345678"
```

Joins: articles ↔ snapshots on (source, articleid) · companies ↔ jobs on cid. Every pipeline write is immutable — back-in-time queries work out of the box.

# A7: Infrastructure Cost Breakdown

**Q: What does it actually cost to run this system every month?**

**Target: ~$10–15/month, fully serverless. Current: $0, all services on free tier.**

| SERVICE | DETAIL | COST / MO | NOTES |
|---|---|---|---|
| **AWS Lambda** | 7 functions | $8 – $15 | News: 1536MB × 450s × 3/day. Jobs: 6 funcs, 256–512MB, SQS-driven. |
| **SQS Queues** | 3 queues + 3 DLQs | $0.50 | $0.40 / M requests. Low volume. |
| **EventBridge** | 7 scheduled rules | $0.70 | News 3×/day, jobs dispatch 3×/day, ATS finder weekly. |
| **CloudWatch Logs** | All Lambda output | $1.50 | ~150–300 MB / month at $0.50/GB ingest. |
| **ECR Container** | 2 images, ~250 MB | $0.50 | Shared ARM image across all Lambda functions. |
| **Neon Postgres** | Free tier | $0 | 3 GB storage, 6–12 mo runway. Pro tier $19 when needed. |
| **Brave Search API** | ~2K queries / mo | $0 – $5 | ATS discovery only. Free tier covers current volume. |
| **ESTIMATED TOTAL** | | $11 – $25 / month | Production-scale estimate: $35–50 / mo (Neon Pro + scaled Brave tier). |

# A8: Technology Stack

**Q: What technologies are used and why?**

| Category | Technology | Version | Rationale |
|---|---|---|---|
| Language | Python | 3.11 | Ecosystem breadth, Lambda support, team expertise |
| HTTP | requests | 2.32 | Industry standard, connection pooling, simple API |
| HTML Parsing | BeautifulSoup4 | 4.12–4.14 | Resilient to malformed HTML, flexible selectors |
| XML Parsing | xml.etree.ElementTree | stdlib | RSS/Atom parsing — no external dependency |
| DB (sync) | psycopg2 | 2.9.9 | Mature Postgres driver for news scraper |
| DB (async) | asyncpg | 0.31 | High-perf async Postgres for Lambda concurrency |
| ORM | SQLAlchemy | 2.0.46 | Type-safe models, async sessions, migrations |
| Migrations | Alembic | 1.18 | Schema evolution for jobs pipeline |
| AWS SDK | boto3 | 1.38 | SQS, Secrets Manager integration |
| Search API | Brave Search | v1 | ATS discovery; competitive pricing, no Google rate limits |
| Browser | Selenium + UC | 4.38 | LinkedIn scraper (undetected-chromedriver) |
| IaC | AWS SAM | — | Simplified CloudFormation for serverless |
| Database | Neon Postgres | 16 | Serverless Postgres with connection pooling |

# A9: Finance Classifier Internals

**Q: How does the finance classifier decide what's finance-related?**

## Three-Stage Matching Pipeline

- Stage 1 — Section Match: article.section (lowercased) checked against FINANCE_SECTIONS set
- Stage 2 — Tag Match: each tag checked for FINANCE_SECTIONS substring OR FINANCE_TAG_KEYWORDS match
- Stage 3 — URL Path Match: split URL path by '/', check each segment against FINANCE_SECTIONS
- Performance: O(n) via set membership; optimized from O(n×m) in commit 71aacf8

## Configuration

- FINANCE_SECTIONS: comma-delimited override via Lambda env var
- FINANCE_TAG_KEYWORDS: comma-delimited override via Lambda env var
- Both runtime-configurable without code deploy

# A10: News Database Schema (DDL)

**Q: What is the exact database schema for the news pipeline?**

```
CREATE TABLE articles (
    source TEXT, articleid TEXT, title TEXT, byline TEXT,
    url TEXT, published_at TIMESTAMPTZ, summary TEXT, tags TEXT[],
    PRIMARY KEY (source, articleid)
);

CREATE TABLE snapshots (
    source TEXT, articleid TEXT, scraped_at TIMESTAMPTZ,
    section TEXT, global_position INT, section_position INT,
    article_updated_at TIMESTAMPTZ,
    UNIQUE (source, scraped_at, articleid)
);
-- Indexes: scraped_at, (scraped_at, global_position),
--          (articleid, scraped_at), published_at

CREATE TABLE finance_snapshots (
    -- Same structure as snapshots
    -- Parallel table for finance-classified articles
);
```

# A11: Jobs Database Schema (ORM)

## Q: What is the jobs database schema and what enums are used?

### SQLAlchemy ORM Models

- Company: UUID pk, name, ticker, in_sp500, ats_provider, company_url, jobsite_url, ats_searched_at, ats_search_attempts, ats_search_last_error
- Job: UUID pk, cid (FK), title, description, location_raw, employment_type, pay_hour_low/high, pay_salary_low/high, currency, status, posted_at, closed_at, ats_provider, ats_job_id, post_url, first_seen_at, last_seen_at
- Unique constraint: (cid, ats_provider, ats_job_id)
- Managed via Alembic (3 migrations: init, add_ticker, remove_lever)

### Enum Definitions (from source code)

- ATSProvider: workday, greenhouse, icims, taleo, successfactors, smartrecruiters, ashby, oraclecloud, phenom, eightfold, lever, avature, jobvite, dayforce, custom
- EmploymentType: full_time, part_time, internship, seasonal
- Currency: usd
- JobStatus: open, closed

# A12: Data Freshness & Latency

Q: How fresh is the data and what is end-to-end latency?

| Data Type | Update Frequency | Typical Latency | Staleness Risk |
|---|---|---|---|
| News articles | 3× daily | ~60s (scrape-to-DB) | 8h max between scrapes |
| News snapshots | 3× daily | ~60s | Position ages with next scrape |
| Finance classification | Per scrape | In-line (0s added) | Follows news latency |
| S&P 500 list | Weekly | ~30s | Low (changes quarterly) |
| ATS URLs | Daily | ~5min per company | URLs rarely change day-to-day |
| ATS health | Daily | ~2s per URL | 24h max stale |
| Job postings | 3× daily | ~30–120s per company | 8h max between scrapes |
| LinkedIn data | Ad-hoc | ~5–10min per search | Manual trigger only |

# A13: Scalability, Growth & Retention

**Q: How does the system scale, what are the bottlenecks, and how is historical data retained?**

## Current Capacity & Growth

- **News snapshots.** ~9,000 rows/day across 30 sources × 3 daily scrapes ≈ 3.3M rows/year.
- **Finance snapshots.** ~500–1,000 rows/day on the finance-filtered parallel table.
- **Articles.** ~500–1,000 newly-deduplicated articles per day.
- **Jobs postings.** 60,367 postings captured across 470 S&P 500 companies and 15 ATS providers.
- **Lambda auto-scale.** Per-invocation scaling, no server management, no provisioning decisions.
- **SQS throughput.** Effectively unlimited for our volume (AWS-managed).
- **Adding Fortune 1000.** Roughly 2× cost, zero architecture changes — the same pipeline runs.

## Bottlenecks & Retention Strategy

- **DB connection pooling.** Neon pooler + SQLAlchemy NullPool prevent exhaustion at Lambda scale.
- **Brave Search rate limit.** 1.5s delay per request, max 500/run — batched into daily ATS-finder job.
- **Lambda timeout.** Single-company per invocation bounds execution to 600s max.
- **No deletion today.** All data retained — this is the research asset, not hot transactional state.
- **SQS DLQ retention.** 14-day default before failed messages are purged.
- **Neon runway.** Free tier (0.5 GB) carries ~6–12 months; Neon Pro ($19/mo, 10 GB) picks up after.
- **Planned.** Monthly partitioning on snapshots; S3 + Athena archive for data older than one year.

# A14: Monitoring & Alerting

## Q: How is the system monitored in production?

### Current

- CloudWatch Logs: all Lambda stdout/stderr
- Lambda return payload: article/snapshot counts
- SQS metrics: visible messages, DLQ depth
- ATS health check logs: daily URL validation report

### Planned

- Grafana dashboard: real-time status, article counts, error rates
- CloudWatch Alarms: Lambda errors > threshold, DLQ > 0
- SNS notifications: email/Slack for pipeline failures
- Custom metrics: per-source article count, classification hit rate

# A15: Rate Limiting & Throttling

### Q: How does the system manage rate limits and avoid being blocked?

## Rate Limiting Strategy

- News: User-Agent mimics Chrome 123; 10s connect, 30s read timeout
- ATS Finder: 1.5s delay between Brave Search requests, max 500/run
- LinkedIn: SCROLL_PAUSE=5s, DETAIL_PAUSE=2s
- SQS visibility timeout prevents duplicate processing

## Anti-Detection

- RSS/Atom feeds are public APIs — no scraping detection risk
- Workday/Greenhouse: public job board APIs, no auth required
- LinkedIn: undetected-chromedriver patches Chrome fingerprints
- Lambda IP rotation naturally distributes load

# A16: Security Measures

**Q: What security measures are in place?**

## Infrastructure Security

- SQL injection guard: parameterized queries throughout (audit fix in commit 40db84c)
- Credentials via AWS Secrets Manager or Lambda env vars (encrypted at rest)
- Connection leak fix: identified and resolved in news scraper (commit 40db84c)
- SQS encryption at rest (AWS managed); ECR image scanning for CVEs

## Data Security

- No PII collected: job postings contain public data only
- News data is publicly available content — no paywalled content extracted
- .env files gitignored; .env.example provided for onboarding

# A17: S&P 500 Reconstitution Handling

## Q: What happens when companies enter or leave the S&P 500?

### Reconstitution Process

- Wiki Crawler runs weekly (Monday 02:00 UTC) — re-scrapes the full S&P 500 list from Wikipedia
- New additions: inserted with in_sp500=True, queued for ATS discovery automatically
- Removals: in_sp500 set to False — record and historical job data preserved (soft state change)
- Dispatcher only queries WHERE in_sp500=True — removed companies stop being scraped automatically
- Historical data for removed companies remains queryable for backtesting

### Edge Cases

- Reconstitution: quarterly (Mar, Jun, Sep, Dec) + ad-hoc. Weekly crawl catches within 7 days.
- Name changes (mergers, rebrandings): may require manual ATS URL re-discovery
- Ticker changes tracked via ticker column in companies table

# A18: Testing Strategy

**Q: How is the system tested before deployment?**

### Unit & Integration Tests

- test_sources.py: parser tests (NYT, RSS) with checked-in fixtures
- test_finance_classifier.py: classification logic validation
- test_run_integration.py: full pipeline (fetch → classify → store)
- test_ats_finder.py + per-provider tests for all 15 scrapers (workday, greenhouse, icims, ashby, lever, eightfold, phenom, oraclecloud, taleo, smartrecruiters, successfactors, avature, jobvite, dayforce, custom)

### Smoke Test Pipeline

- scripts/smoke_test.py: end-to-end pre-deploy validation
- Spawns ephemeral Docker Postgres + applies schema
- Builds Lambda container (arm64/x86_64), invokes locally
- Validates row counts: articles, snapshots, finance_snapshots

# A19: Roadmap

## Q: What's planned next, and on what timeline?

### NEAR-TERM

Q2 2026

- NLP sentiment analysis (FinBERT) on article bodies
- Named-entity recognition for company and person extraction
- Backfill the remaining 33 failing-ATS companies
- Grafana dashboard for real-time pipeline health
- Historical snapshot backfill where archive.org allows
- Upgrade finance classifier to embedding-based similarity

### MID-TERM

Q3–Q4 2026

- Topic modeling (LDA / BERTopic) for narrative clustering
- Cross-source article deduplication (same story, different outlets)
- Expand scope to Fortune 1000 for broader hiring coverage
- Build and backtest factor model: prominence + hiring velocity
- FastAPI layer exposing clean query endpoints for the team
- Internal benchmark: compare our classifier to FinBERT labels

### LONG-TERM

2027+

- Event-driven architecture: news event → signal → execution
- International source expansion (FT, Nikkei, SCMP, Economic Times)
- Social integration (Twitter / X, Reddit, StockTwits)
- ML article-importance scoring (replaces position heuristic)
- Production alpha strategy backtesting on captured data
- Paper / preprint on editorial-prominence signal methodology

# A20: Works Cited

Academic papers and industry sources referenced in this presentation

## News Signals


Tetlock (2007). “Giving Content to Investor Sentiment.” J. Finance 62(3).
Tetlock, Saar-Tsechansky & Macskassy (2008). “More Than Words.” J. Finance 63(3).
Loughran & McDonald (2011). “When Is a Liability Not a Liability?” J. Finance 66(1).
Heston & Sinha (2017). “News vs. Sentiment.” Fin. Analysts J. 73(3).
Garcia (2013). “Sentiment during Recessions.” J. Finance 68(3).
Fang & Peress (2009). “Media Coverage and the Cross-Section.” J. Finance 64(5).
Hillert, Jacobs & Mueller (2014). “Media Makes Momentum.” Rev. Fin. Studies 27(12).
Peress (2014). “The Media and the Diffusion of Information.” J. Finance 69(5).
Boudoukh, Feldman, Kogan & Richardson (2019). “Information, Trading, and Volatility.” Rev. Fin. Studies 32(3).
Calomiris & Mamaysky (2019). “How News and Its Context Drive Risk.” J. Fin. Econ. 133(2).
Gentzkow & Shapiro (2010). “What Drives Media Slant?” Econometrica 78(1).
Araci (2019). “FinBERT.” arXiv:1908.10063.
Lopez-Lira & Tang (2023). “Can ChatGPT Forecast Stock Price Movements?” arXiv:2304.07619.


## Jobs / Hiring Signals


Kothari & O'Doherty (2023). “Job Postings and Aggregate Stock Returns.” J. Fin. Markets 64.
Belo, Lin & Bazdresch (2014). “Labor Hiring, Investment, and Stock Return Predictability.” J. Political Econ. 122(1).
Li, Lourie, Nekrasov & Shevlin (2022). “Employee Turnover and Firm Performance.” Mgmt. Science 68(8).
Green, Huang, Wen & Zhou (2019). “Crowdsourced Employer Reviews and Stock Returns.” J. Fin. Econ. 134(1).
Hazell & Taska (2025). “Downward Rigidity in the Wage for New Hires.” Amer. Econ. Rev. 115(12).
Acemoglu, Autor, Hazell & Restrepo (2022). “AI and Jobs: Evidence from Online Vacancies.” J. Labor Econ. 40(S1).
Babina, Fedyk, He & Hodson (2024). “AI, Firm Growth, and Product Innovation.” J. Fin. Econ.
Hershbein & Kahn (2018). “Do Recessions Accelerate Routine-Biased Tech Change?” Amer. Econ. Rev. 108(7).
Kolanovic & Krishnamachari (2017). “Big Data and AI Strategies.” JP Morgan QDS.


## Industry Data Providers


Revelio Labs — workforce analytics on WRDS platform.
Lightcast (fmr. Burning Glass) — 2.5B+ historical job postings.
LinkUp — S&P 500 Jobs Index (with S&P Dow Jones Indices).
Indeed Hiring Lab — Job Postings Index, Posted Wage Tracker.

# Northeastern Systematic Alpha

## Data Scraping Infrastructure

Questions & Discussion

may.se@northeastern.edu | Quantitative Development Team
ilic.f@northeastern.edu | Quantitative Development Team