EXPONENTIAL | BY RAVENPACK

FROM DATASET CONSUMER TO DATASET CREATOR

How LLMs and search-native infrastructure are changing alpha construction.

Peter Hafez

Chief Data Scientist,
RavenPack

THE SHIFT

Consuming structure.

→

Creating your own.

It's no longer whether unstructured data produces alpha — it's where in the stack you want to compete.

THE STACK

THREE PLACES TO LIVE

01 · ANALYTICS

Structured

What we built. A decade of analytics.

High-dimensional scores, our taxonomy, point-in-time. Clients filter, aggregate, combine.

02 · ANNOTATIONS

Enriched

Textual content — your infrastructure.

Normalized content, entities and events enriched. Clients own the modeling.

03 · BIGDATA

Search-native

Textual content — our infrastructure.

Hosted retrieval at scale. Custom taxonomies on-the-fly. LLMs creating structure that didn't exist before.

RavenPack's role is shifting with the market — from supplier of structured analytics to infrastructure for dataset creation.

THE WORKFLOW

CUSTOM DATASET CREATION

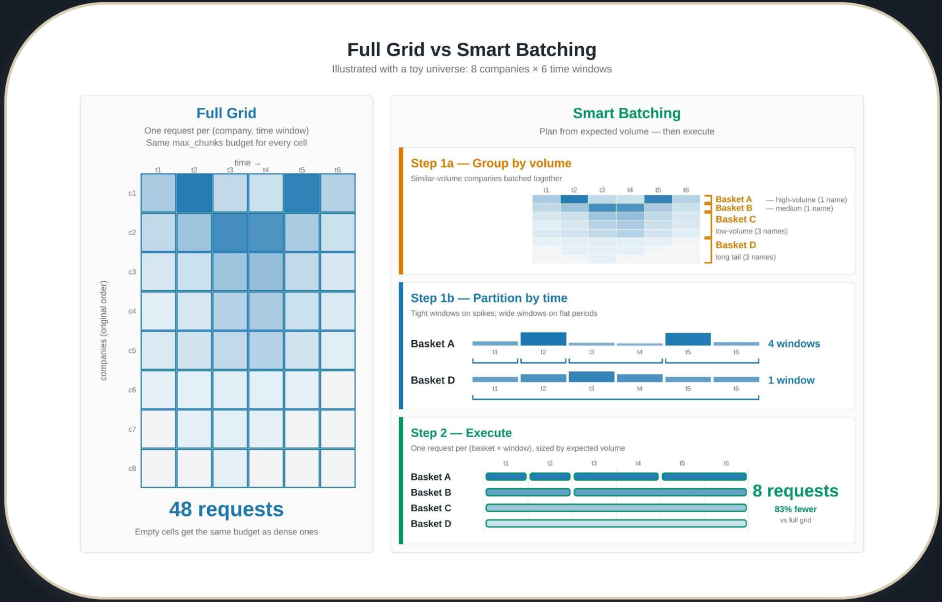
From hypothesis to tradeable signal.

Define your universe	Define your searches	Retrieve at scale	Construct your signal	Run your backtest
Companies, sectors, geographies.	*Themes, events, language patterns.*	*Smart batching + smart sampling.*	*Feature engineering + LLM augmentation/validation.*	*Strategy development + robustness.*

RETRIEVAL WITH SMART BATCHING

RETRIEVAL AT SCALE

Smart batching turns universe-scale screening from hours into minutes, preserving cross-sectional and time-series distributions



01

Faster.

10–50× depending on topic

02

More efficient.

+90% query reduction vs. full grid

03

Smart sampling.

Pick a % of full grid — get the top-relevance chunks.

Under the hood: co-mention endpoint groups companies, volume endpoint splits time. One plan, proportional budget.

FROM THEMES TO EVIDENCE

SIX THEMES, ONE EXAMPLE.

*Each theme retrieves chunks like this one — **Pricing power** shown.*

SHORT
Pricing pressure
Discounting and ASP declines.

LONG
Pricing power
Raises prices without losing volume.

SHORT
Demand softening
Weaker orders, destocking.

LONG
New product traction
Ramping volumes from new launches.

SHORT
Guidance credibility
Missed guidance, lowered outlooks.

LONG
Capacity expansion
Growing backlog, new facilities.

FACTSET TRANSCRIPTS · RANK_1 · 2026-05-19 13:10 UTC

Carrier Global Corp. — Wolfe Research Industrials Conference

doc_id 593325A691BD6910019E5036F6164E19

*"… we did some pricing activity in Germany, that's now behind us. So, **we've raised price starting in April. So, a few points of price, few points of surcharges.** So, I think if the volume continues and we're **very disciplined on pricing,** the indicators are positive for Resi."*

RELEVANCE	SENTIMENT	CONTRIBUTION
0.61	+0.25	0.61 × +0.25 = **+0.15**

FROM CHUNK TO POSITION

01 · CHUNKS
News + transcripts
Articles, earnings calls, conference Q&A.

→

02 · DAILY SCORE
Per (entity, theme)
Relevance-weighted sentiment. One number per company per day.

→

03 · POSITION
Rank, neutralize, hold
Top names go long; bottom go short. ~1.6-day holding.

*Carrier's transcript chunk contributes **+0.15** to its pricing-power score for May 19, 2026.*

THE RESULT · 01 OF 02

SIX THEMES, ONE COMBINED BOOK.

Long-short, market-neutral. Russell 1000. 2016 — today.

INFORMATION RATIO

1.01

Combined book.

ANN. RETURN

+6.9% market-neutral

VOLATILITY

6.8% annualized

PORTFOLIO

~166 names 1.2-day holding

Cumulative log-return, daily, equal-weighted across six themes.

*Mean theme IR **0.51**. Combined IR **1.01**. Diversification is doing real work.*

THE RESULT · 02 OF 02

PER-THEME PERFORMANCE.

2016 — today. Russell 1000. Long-only or short-only per theme.

Theme	Side	IR	Ann. ret.	Vol	Avg names	Hold (days)	Gross expo.
Pricing power	LONG	0.40	+3.0%	7.5%	77.1	1.2	1.00
New product traction	LONG	0.72	+6.6%	9.1%	87.7	1.2	1.00
Capacity expansion	LONG	0.69	+5.6%	8.1%	71.3	1.2	0.99
Pricing pressure	SHORT	0.47	+4.2%	9.0%	50.6	1.3	0.99
Demand softening	SHORT	0.38	+3.3%	8.8%	14.4	1.8	0.62
Guidance credibility	SHORT	0.41	+4.2%	10.3%	21.8	1.6	0.78
Combined book	**L/S**	**1.00**	**+6.9%**	**6.9%**	**166.0**	**1.2**	**1.00**

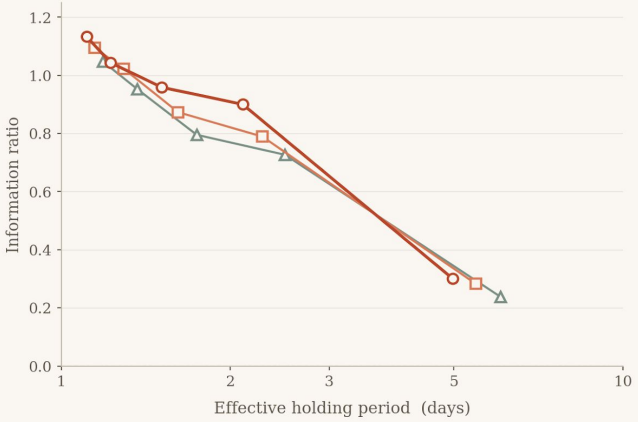
Per-theme returns are modest; the combination is what produces a 1.01 IR.

SENSITIVITY

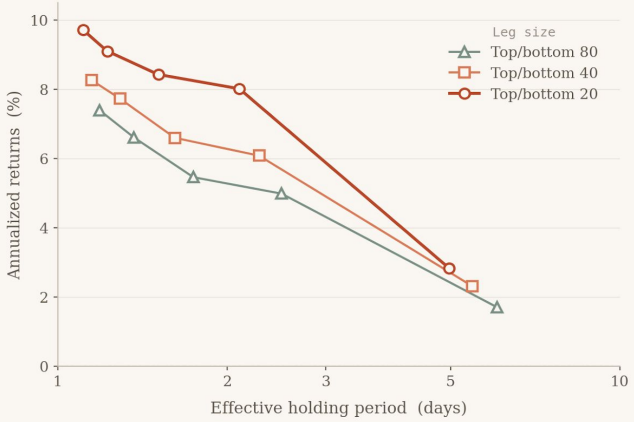
SMALLER, FASTER, STRONGER.

Same combined book. Signal half-life (exponential decay) and portfolio breadth move performance the same way.

INFORMATION RATIO



ANNUALIZED RETURN



WHAT IT TELLS US

CONCENTRATED

Smaller leg, more alpha.

IR **1.13** / return **+9.7%** at leg 20, vs **1.05** / **+7.4%** at leg 80.

TIMELY

Faster signal, sharper edge.

Shorter half-life produces stronger signals; smoothing reduces the edge.

ROBUST

Monotonic behavior.

Performance falls smoothly with breadth and decay.

METHOD

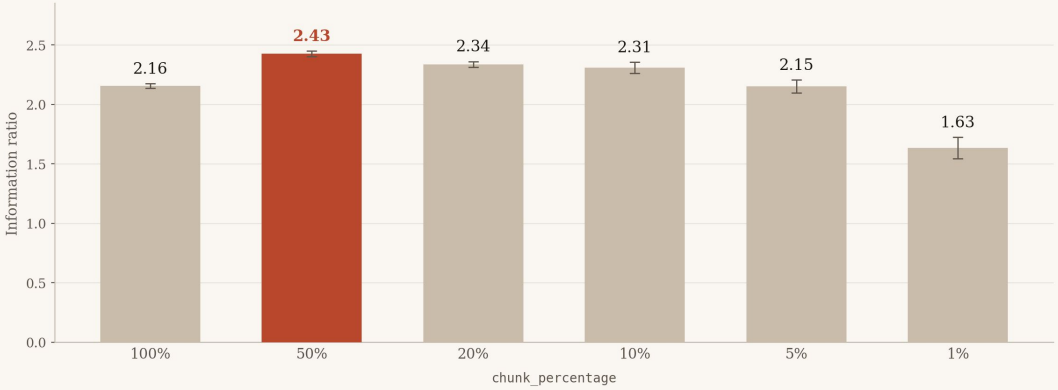
Exponential decay of the daily signal with calendar-day half-life ∈ {**0.25, 0.5, 1, 2, 7**}. Top/bottom selection at leg ∈ {**20, 40, 80**}. Russell 1000, 2016…2026. Effective holding period = 1 / turnover.

*Concentrated, timely, **robust** — sensitivity is monotonic in both breadth and decay.*

ROBUSTNESS

SAMPLING IS ROBUST AND THERE'S A SWEET SPOT.

Pricing power, US Top 1000, last 12 months. Five samples per chunk_percentage.
***98.5%** mean doc overlap across samples.*



INTERPRETATION

100%

Noisy ceiling.

Highest retrieval volume; IR mean diluted by marginal chunks.

50% – 10%

Sweet spot.

Best tradeoff between coverage and signal quality.

5% – 1%

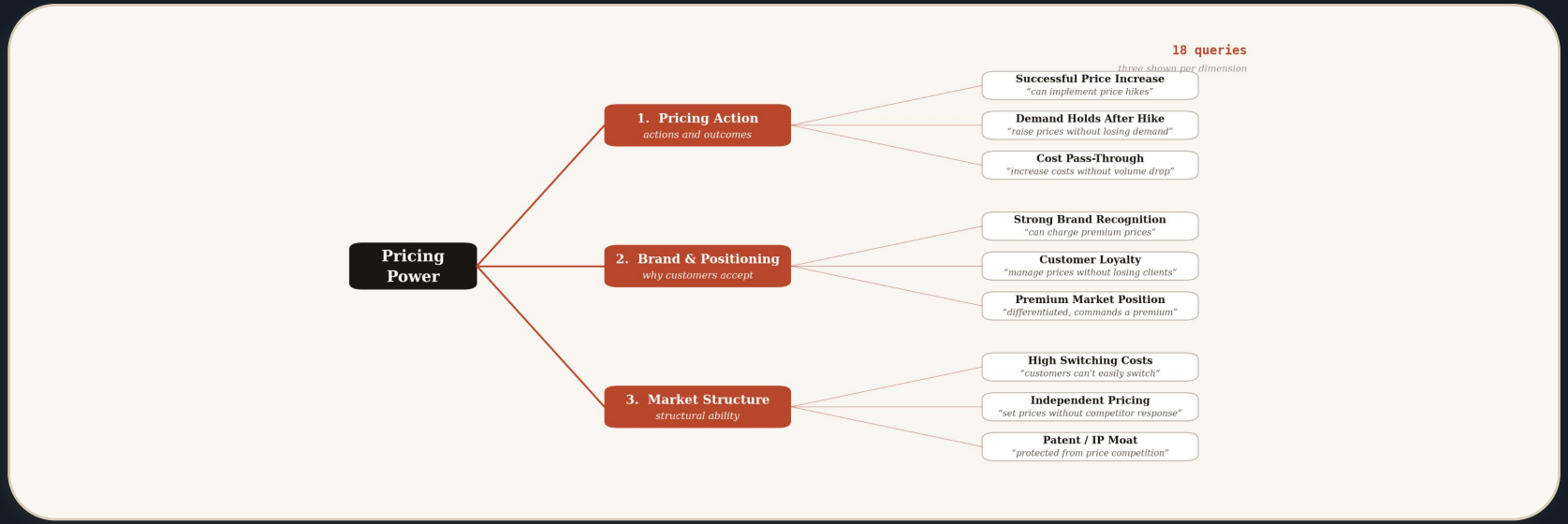
Thin retrieval.

Signal degrades as chunk volume thins; useful lower bound.

PRICING POWER · 01 OF 03

RECALL BOOSTING.

One theme is many things in text. A wider semantic net covers more of how the idea is actually written.



WHY MANY ANGLES

01 · ONE THEME

Many surface forms

Cost pass-through, brand premium, switching costs.

→

02 · MANY ANGLES

Distinct embeddings

Each phrasing maps to a different point in semantic space.

→

03 · HIGHER RECALL

More relevant chunks

Parallel queries pull a broader chunk universe.

PRICING POWER · 02 OF 03

EACH CHUNK GETS A VERDICT.

*An LLM reads each retrieved chunk and returns a **classification**, a **magnitude**, and a **motivation** per company.*

WINNER

MEDIUM

BENZINGA · 2018-10-24

Acuity Brands Reports Fiscal 2019 Q1

"…we implemented two price increases to recover higher costs for components and other input items due to inflation and government tariffs…"

REASONING

Multiple price increases to offset cost and tariff pressure — can raise prices, not just absorb.

LOSER

HIGH

FINANCIAL TIMES · 2023-08-09

Lyft: Fare cuts bring little cheer

"…Lyft has cut its fares and reduced surge pricing. The strategy resulted in a 5% drop in revenue per active user to $47.51…"

REASONING

Cutting fares and reducing surge pricing; RPU falling — pressured on monetization.

IRRELEVANT

THE FLY · 2018-10-24

NY regulators probe CVS drug prices

"…regulators are examining CVS's fluctuation in pricing of drugs as they review CVS Health's $69B proposed merger with Aetna…"

REASONING

Aetna mentioned only in merger context, no specific pricing-power signal.

SIGNAL CONSTRUCTION

$$contribution = sign(class) \times magnitude$$

WINNER = +1 · LOSER = −1 · LOW = 1 MID = 2 HIGH = 3

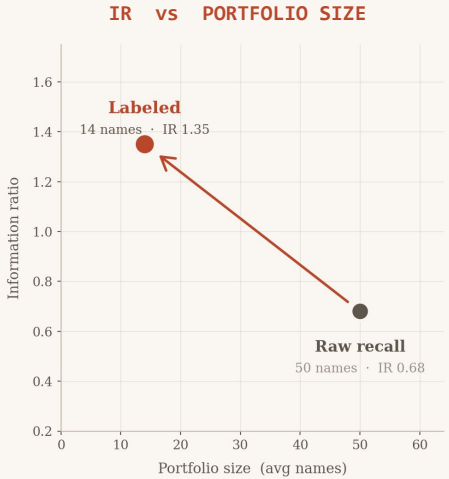
*Winner and loser chunks feed the daily score, weighted by magnitude. **Irrelevant chunks are dropped**.*

PRICING POWER · 03 OF 03

FROM BROAD CROSS-SECTION TO TARGETED SIGNAL.

*Same retrieval pool. Same rules. Labeling moves the book to fewer names with stronger evidence — **IR doubles**.*

Cumulative log-return, daily.



THE TRADE

Raw recall is a quant signal — broad cross-section, diluted edge per name. Labeling concentrates the book on the **names with verified evidence**, trading breadth for conviction.

High recall finds the chunks. ***Labeling picks the conviction names.***

EXPONENTIAL | BY RAVENPACK

EXTERNAL VALIDATION · WORLDQUANT BRAIN

DATASET CREATION AT SCALE.

A real-world experiment in alpha generation.

PARTICIPANTS

5,632

across 17 countries

DATAFIELDS BUILT

46,423

custom, not vendor-fed

ALPHAS TAGGED

5,210

from those datafields

OUT-OF-SAMPLE SHARPE RATIO DISTRIBUTION

SHARPE RATIO > 0

68.4%

produced positive alpha

SHARPE RATIO > 1.0

29.2%

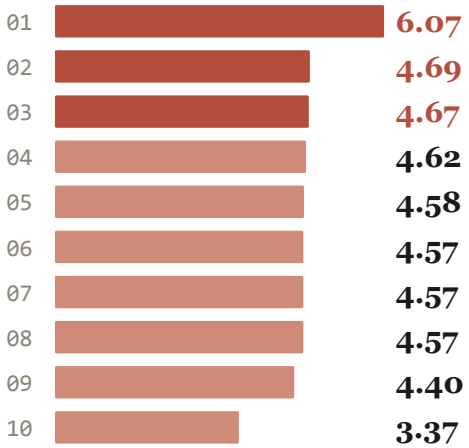
institutional-grade

SHARPE RATIO > 1.5

15.9%

exceptional

TOP 10 ALPHAS BY OOS SHARPE RATIO



SYNTHESIS

The market is moving toward raw text.

Toward LLMs creating structure on the fly.

Toward custom datasets that *didn't exist before.*

RavenPack's role is shifting with it. We built the enriched & structured layers. With Bigdata.com, we're building the search-native layer.

The work that used to live with vendors is moving closer to the use-case — where it belongs.

SYNTHESIS

The market is moving toward raw text.

Toward LLMs creating structure on the fly.

Toward custom datasets that *didn't exist before.*

RavenPack's role is shifting with it. We built the enriched & structured layers. With Bigdata.com, we're building the search-native layer.

The work that used to live with vendors is moving closer to the use-case — where it belongs.

Stop consuming someone else's world.

Start *creating* your own.

Peter Hafez · Chief Data Scientist · RavenPack