Capturing Alpha from Internal Digital Content

Peter Hafez
RavenPack

In this paper, we explore the benefits of using internal digital content, such as emails, attachments and instant messages, to generate differentiated investment insights and trading signals by leveraging our proprietary AI platform and NLP engine. Each organization owns and accumulates massive amounts of digital content, which largely remains under analyzed and untapped. Identifying signals from the noise within the vast volumes of unstructured textual data presents a significant challenge. Intuitively, it is safe to assume that each organization's unique digital footprint contains distinct information that can yield actionable insights.

For instance, it is well-known that institutional investors benefit significantly from corporate access. A substantial amount of analyst insights come from one-on-one meetings with management or private conference calls with sell-side analysts or expert networks. The nature of the discussions during these meetings is likely to be more direct and honest than the scripted rhetoric coming from earnings calls, analyst days and otherwise officially scheduled presentations. Buy-side sector specialists form close relationships with C-level executives and investor relations over the years, and are able to discern nuanced changes in tone or body language that are likely to be shared in internal notes as a part of a "mosaic theory" approach to forming investment decisions. Sell-side analysts, who publicly tend to be more long-biased in their coverage in an effort to preserve investment banking relationships, are frequently more honest about their views on private calls. More often than not, these insights get documented internally in a sea of pre-earnings write-ups and postmortems. The main goal is to capture as much of this information across the entire organization. Additionally, internal research collaborations and content sharing can offer interesting insights into the team's analytical process, adding a human layer to the equation.

The content we analyzed consists of three years of internal data (2016-2019), comprising of hundreds of thousands of emails, attachments and Skype messages in over 1,000 different file formats. While the data is anonymized, we know that the content includes broker research, desk commentary and internal research notes. This content is likely to be highly curated and differentiated by nature, and within the purview of a firm's mandate.

We demonstrate that there is incremental value to be captured within an investment organization's own data as opposed to using public content alone, particularly when considering longer investment horizons. This is a real-world case study conducted on the internal content of a $1 billion European discretionary hedge fund, which focuses on the utilities, infrastructure and commodities sectors in broader Europe.

We collect, store and build a historical archive of all internal content, which we then run through the proprietary NLP engine to capture events and companies tied to those events, structuring and enriching the data in the process. The identified events are categorized under our event taxonomy of over 6,800 events detected by our algorithm, assigning several sentiment scores to each and determining a number of other metrics, such as event novelty and relevance.

We proceed to create a real-world portfolio that captures this incremental information. Exhibit 1 details the incidence of captured internal content that gets passed through our NLP engine for structuring and enrichment.

Methodology

We compute the daily trading signal for each stock by aggregating an Event Sentiment Score (ESS) captured throughout the day, weighted by Event Relevance and Event Similarity. ESS is a granular sentiment score between -1 and 1 that represents the sentiment for a given event identified inside a body of text, and is determined dynamically by our pattern matching algorithm, using scores from similar events categorized by financial experts as having short-term negative or positive impact. Below formula is an example of how granular ESS can be aggregated to a tradeable daily signal by security.

$$signal = \frac{1}{n}\sum_{i=1}^{n} ESS_i \left(\frac{ER_i}{100}\right)^2 \left(\frac{ESD_i}{365}\right)^2 \quad (1)$$

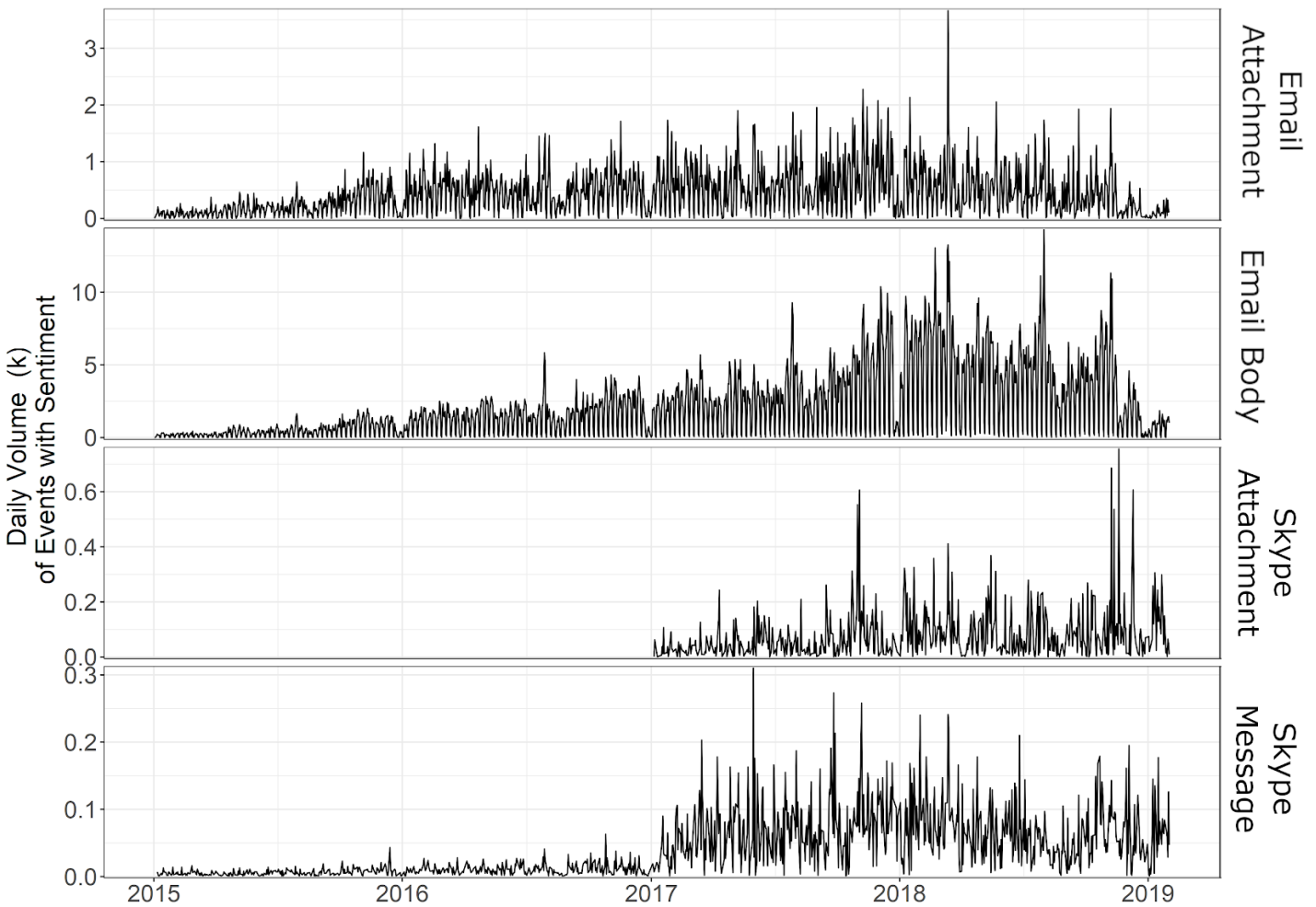

Exhibit 1: Volume of Internal Content Captured Over Time
Source: RavenPack, May 2019

Weighting each ESS score by Event Relevance and Event Similarity Days ensures that relevant and novel events have a higher impact on the daily score, which will be used in a portfolio construction framework. To examine longer investment horizons, we can average the signal over several periods, arriving at a more stable indicator, which results in lower turnover and longer holding periods. Such time-averaging not only lowers portfolio turnover and trading costs, but also allows us to capture some of the post-event drift often missed by fast-moving signals.

While more sophisticated weighting techniques are available, such as using exponentially weighted averages, Kalman filters or utilizing more granular segmented sentiment signals, we illustrate this simple approach to demonstrate the concept in a more straightforward fashion. In this paper, we focus only on internal content. However, a more sensible strategy would combine public news analytics with topical narratives extracted from internal content. We will revisit other methodologies in future research.

Portfolio Construction

The investment universe consists of approximately 180 companies belonging to the utilities, infrastructure and commodities sectors in broader Europe, mirroring the fund's asset coverage. In an effort to replicate the discretionary investment approach, we analyze the performance of relatively concentrated long-only and long-short portfolios at a target AUM level of $100 million, while applying the following constraints to ensure a realistic trading framework:

- Exclude illiquid stocks with less than 0.1% of target AUM as a percentage of 21-day trading volume
- Restrict maximum allocations to 10% of target AUM or to 10% of the 21-day average exchange traded volume, whichever is less (client also uses dark pools to tap additional liquidity)
- Assume continuous rebalancing to maintain constant AUM within the strategy

Taking into account the above constraints, we construct long-only and long-short portfolios with allocations proportional to the aggregated daily sentiment signals, focusing on the Top 40 names.

Results

Approximately 80% of identified stock-related events were detected in the firm's internal content and 20% originated from public news and social media. We found strong long-only signals across the universe that extended into holding periods of several weeks. Signals derived from public news show similar performance or outperform those from the fund's internal content over a 1 to 5 day investment horizon. Nevertheless, the positive sentiment signals derived from the hedge fund's internal content provide better value for longer horizons, up to several weeks (for this particular universe). Exhibit 2 shows long vs. short portfolio performance using internal vs. public content for varying sentiment averaging windows in columns.

In order to get a better sense for the types of events that are captured in the dataset and their impact on strategy performance, in Exhibit 3, we plot the volume and annualized returns of the Top 10 event-group based strategies, captured by the RavenPack event detection algorithm. While product-services and acquisitions-mergers generates lots of volume, the analyst-ratings, equity-actions earnings, and assets event groups produce greater returns.

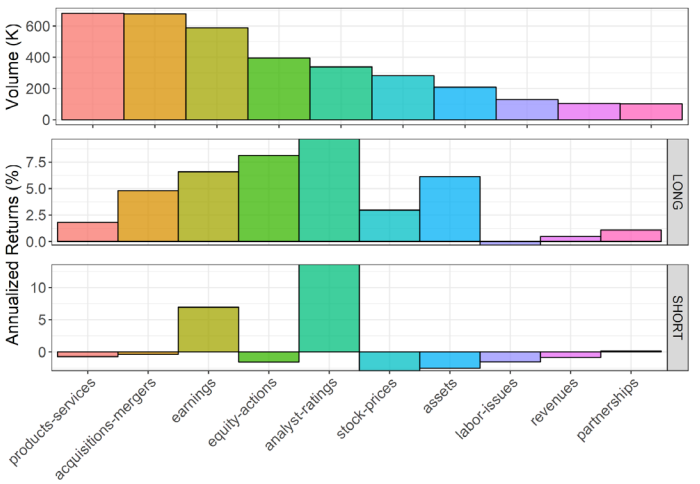

Exhibit 3: Top 10 Event Signals Captured in the Fund's Data
Source: RavenPack, May 2019

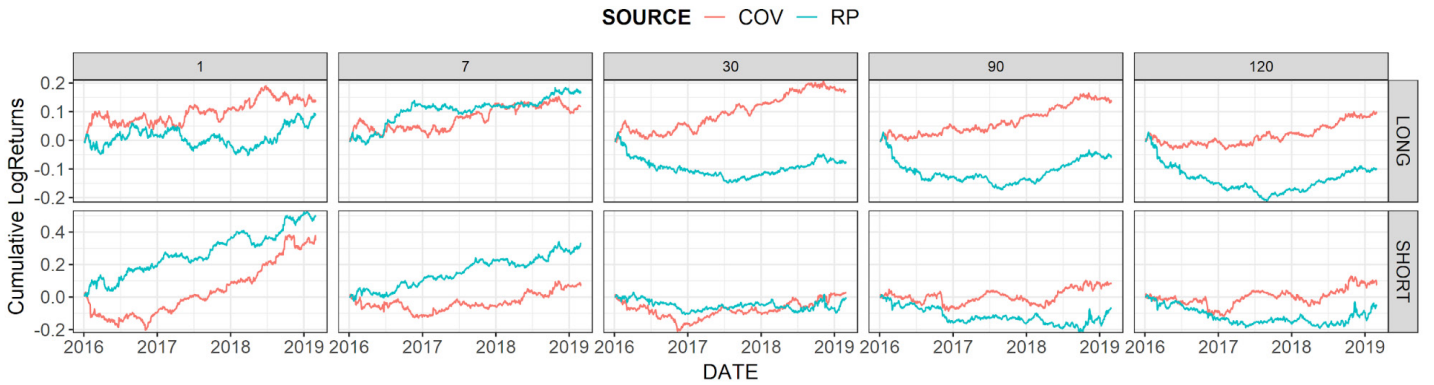

Exhibit 2: Performance Comparison of Internal vs. Public Content
Source: RavenPack, May 2019

Exhibit 4 shows the sentiment distribution of the detected events by category. We can see how many groups have a clear positive sentiment imbalance, which is also reflected in the fact that long portfolios benefit of a larger volume of signals. This is one of the reasons why individual groups usually show better performance for the long side. Note how analyst-ratings, which seems to bring good value for both long and short portfolios (as seen in Exhibit 3), is in fact balanced in sentiment direction. Although those group-specific analytics give us an idea of the event composition, it is not immediate to extrapolate any of these results towards longer trading horizons.

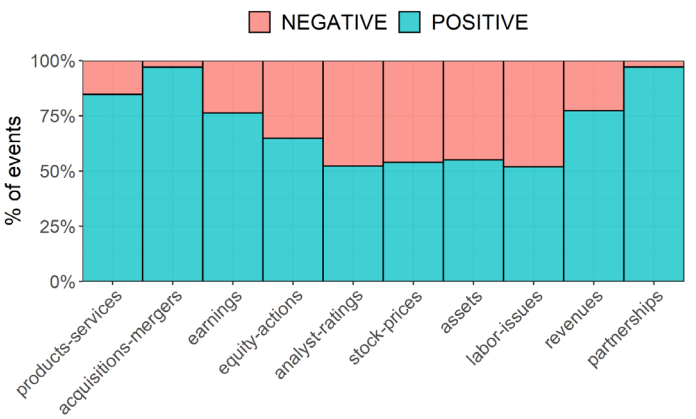

Exhibit 4: Distribution of Top 10 Event Signals Captured in the Fund's Data
Source: RavenPack, May 2019

Strategy Performance

Below, we present the performance metric from 2016-2019 for a long-only $100M portfolio, comprising of 40 stocks, and using a 3-month averaging window for our sentiment indicator. The long-only strategy produces an annualized return of 12.3% (accounting for 8bp of one-way trading costs), with an Information Ratio of 0.8 and average holding period of two to three weeks. The market-neutral strategy passively hedged via the fund's benchmark,[1] produced a 10.6% annualized return with a 1.0 Information Ratio. Exhibit 5 shows cumulative gross P&L of the strategy (before trading costs) versus the benchmark.

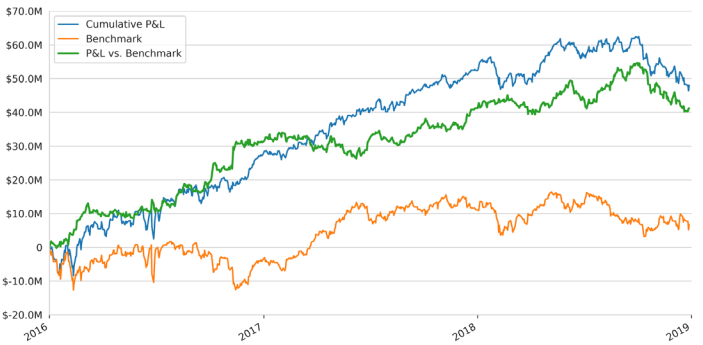

Exhibit 5: Performance of a $100M AUM Long-Only Portfolio Verses Benchmark
Source: RavenPack, May 2019

The specific return of the portfolio, or an equivalent neutral factor-hedged portfolio (assuming a perfect factor hedge), produced an annualized return of 6.4% before trading costs, with an Information Ratio of 1.6. In Exhibit 6, we isolate specific portfolio P&L (alpha) versus traditional market factors, demonstrating persistent alpha over the period.[2]

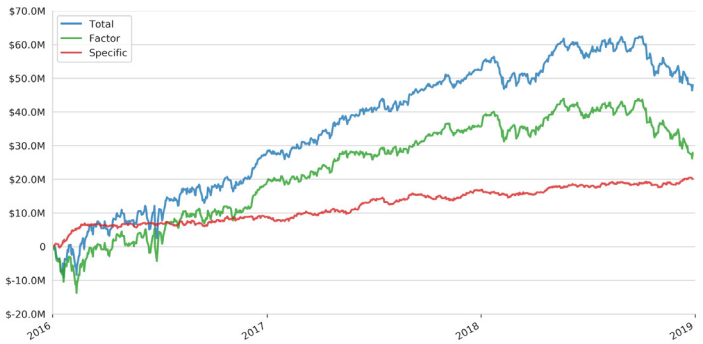

Exhibit 6: Factor Performance Breakdown of a $100M AUM Long-Only Portfolio
Source: RavenPack, May 2019

Portfolio exposures to traditional factors are relatively stable over the period - we can see from Exhibit 7 that the factor component of total portfolio risk ranges between 85-95%, which mainly comprises of broad market beta.

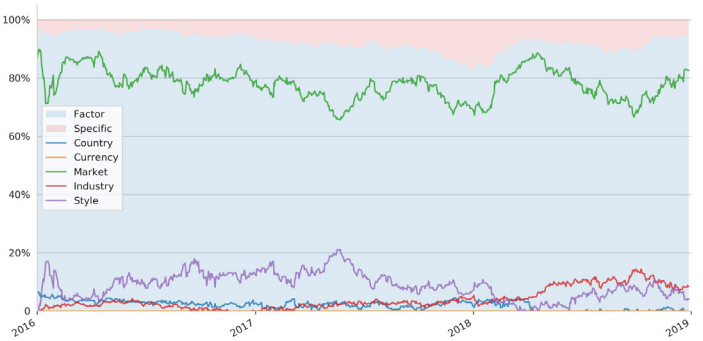

Exhibit 7: Factor Exposure Breakdown of a $100M Long-Only Portfolio
Source: RavenPack, May 2019

To get a better understanding for the high-level composition of the portfolio, we show top position concentrations in Exhibit 8. As can be observed, the portfolios are somewhat concentrated in the extremes with the Top 5 names accounting for about 25% of allocations. However, we still achieve a reasonable degree of diversification including the 40 names into our portfolio, with the Bottom 20 accounting for 30% of allocations.

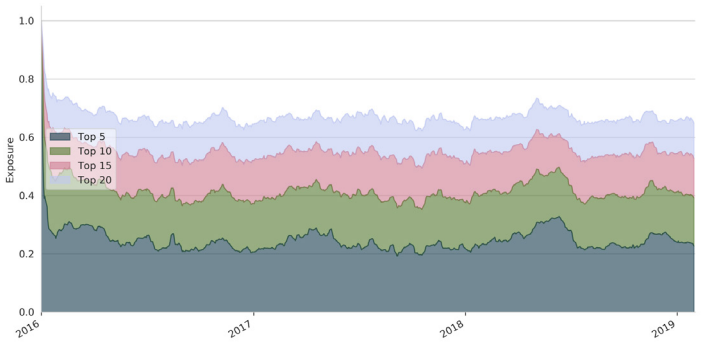

Exhibit 8: Top Position Concentrations Over Time
Source: RavenPack, May 2019

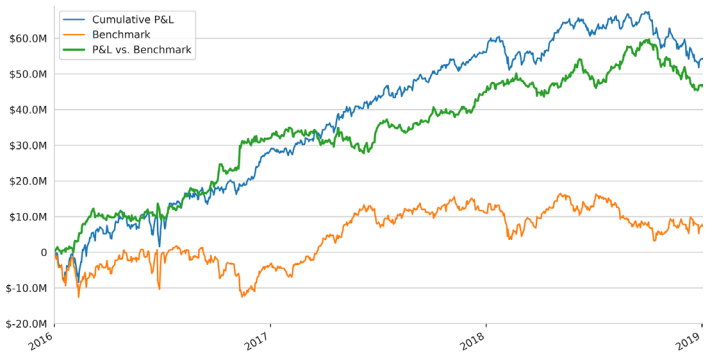

Exhibit 9: Performance of a $100M AUM Long-Only Portfolio with a 3-Day Liquidity Window.
Source: RavenPack, May 2019

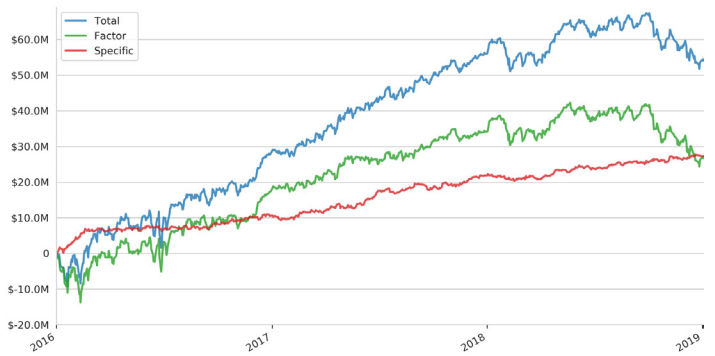

Exhibit 10: Factor Performance Breakdown of a %100M AUM Long-Only Portfolio with a 3-Day Liquidity Window.
Source: RavenPack, May 2019

In an effort to relax some of the liquidity constraints on the portfolio, we decided to examine what happens to performance if we increase the maximum allocation limit to 30% of the 21-day average trading volume from the initial 10%, while still limiting daily trading to 10% of the daily volume. This effectively allows for a 3-day liquidity window. The resulting long-only strategy produced an annualized return of 17.2% (accounting for 8bp one-way trading costs), with an Information Ratio of 1.2 and average holding period of close to three weeks. The market-neutral strategy, passively hedged via the fund's benchmark, produced a 12.9% annualized return with a 1.2 Information Ratio. Exhibit 8 shows cumulative gross P&L of the strategy (before trading costs) versus the benchmark.

The resulting specific return for this strategy was 8.8% before trading costs, with an Information Ratio of 2.2. Exhibit 10 shows the breakdown of factor and specific P&L, demonstrating a more robust alpha over time compared to the original strategy.

Exhibit 11 details the performance metrics for both strategies. Overall, expanding the liquidity window allow us to achieve not only higher returns, but also an improved risk-return trade-off, with the added benefit of reduced turnover.

Conclusion

In this paper, we highlight and uncover the hidden value within the private data assets of a fundamental hedge fund. We demonstrated that there is alpha to be captured in the sea of internal digital content by systematically extracting, structuring and enriching the fund's own content in real time to generate a tradeable investment strategy. The study found strong long-only signals that persist for several weeks, offering fundamental investors a reasonable time frame to act on them. Portfolio risk factor analysis shows stable P&L coming from idiosyncratic price moves, demonstrating persistent alpha generation from a traditional factor model perspective.

The scale of unstructured digital assets across organizations is immense and growing exponentially. Transforming this mountain of data into actionable insights is a real challenge, requiring a combination of reliable technological solutions and sound data science practices. The uniqueness of internal digital footprint within each organization provides an attractive way to harvest differentiated information not available elsewhere and capture incremental alpha that otherwise remains untapped.

Portfolio	Liquidity Window	Information Ratio	Cumulative P&L	Annualized Return (%)	Turnover (%)	Average Holding Period
Long-Only	1-day	0.8	$41.6M	12.3	10.3	~2-3 weeks
	3-day	1.2	$58.3M	17.2	6.8	~3 weeks
Long-Short	1-day	1.0	$34.8M	10.6	10.3	~2-3 weeks
	3-days	1.2	$43.3M	12.9	6.8	~3 weeks
Factor Hedged	1-day	1.6	$20.1M	6.4	-	-
	3-day	2.2	$27.2M	8.8	-	-
Benchmark	-	0.26	$6.8M	3.8	-	-

Exhibit 11: Performance Metric for a $100M Portfolio
Source: RavenPack, May 2019

Disclosure

This White Paper is not intended for trading purposes. The White Paper is not appropriate for the purposes of making a decision to carry out a transaction or trade. Nor does it provide any form of advice (investment, tax, legal) amounting to investment advice, or make any recommendations regarding particular financial instruments, investments or products. RavenPack may discontinue or change the White Paper content at any time, without notice. RavenPack does not guarantee or warrant the accuracy, completeness or timeliness of the White Paper

You may not post any content from this White Paper to forums, websites, newsgroups, mail lists, electronic bulletin boards, or other services, without the prior written consent of RavenPack. To request consent for this and other matters, you may contact RavenPack at research@ravenpack.com.

THE WHITE PAPER IS PROVIDED "AS IS", WITHOUT ANY WARRANTIES. RAVENPACK AND ITS AFFILIATES, AGENTS AND LICENSORS CANNOT AND DO NOT WARRANT THE ACCURACY, COMPLETENESS, CURRENTNESS, TIMELINESS, NONINFRINGEMENT, TITLE, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE OF THE WHITE PAPER, AND RAVENPACK HEREBY DISCLAIMS ANY SUCH EXPRESS OR IMPLIED WARRANTIES. NEITHER RAVENPACK NOR ANY OF ITS AFFILIATES, AGENTS OR LICENSORS SHALL BE LIABLE TO YOU OR ANYONE ELSE FOR ANY LOSS OR INJURY, OTHER THAN DEATH OR PERSONAL INJURY RESULTING DIRECTLY FROM USE OF THE WHITE PAPER, CAUSED IN WHOLE OR PART BY ITS NEGLIGENCE OR CONTINGENCIES BEYOND ITS CONTROL IN PROCURING, COMPILING, INTERPRETING, REPORTING OR DELIVERING THE WHITE PAPER. IN NO EVENT WILL RAVENPACK, ITS AFFILIATES, AGENTS OR LICENSORS BE LIABLE TO YOU OR ANYONE ELSE FOR ANY DECISION MADE OR ACTION TAKEN BY YOU IN RELIANCE ON SUCH WHITE PAPER. RAVENPACK AND ITS AFFILIATES, AGENTS AND LICENSORS SHALL NOT BE LIABLE TO YOU OR ANYONE ELSE FOR ANY DAMAGES (INCLUDING, WITHOUT LIMITATION, CONSEQUENTIAL, SPECIAL, INCIDENTAL, INDIRECT, OR SIMILAR DAMAGES), OTHER THAN DIRECT DAMAGES, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. IN NO EVENT SHALL THE LIABILITY OF RAVENPACK, ITS AFFILIATES, AGENTS AND LICENSORS ARISING OUT OF ANY CLAIM RELATED TO THIS AGREEMENT EXCEED THE AGGREGATE AMOUNT PAID BY YOU FOR THE WHITE PAPER. JURISDICTIONS DO NOT ALLOW THE EXCLUSION OR LIMITATION OF LIABILITY FOR DAMAGES OR THE EXCLUSION OF CERTAIN TYPES OF WARRANTIES, PARTS OR ALL OF THE ABOVE LIMITATION MAY NOT APPLY TO YOU.

These Terms of Use, your rights and obligations, and all actions contemplated by these Terms of Use will be governed by the laws of New York, NY, USA and You and RavenPack agree to submit to the exclusive jurisdiction of the New York Courts. If any provision in these Terms of Use is invalid or unenforceable under applicable law, the remaining provisions will continue in full force and effect, and the invalid or unenforceable provision will be deemed superseded by a valid, enforceable provision that most closely matches the intent of the original provision.

Endnotes

1. The fund uses a blend of 70% MSCI European Utilities Index and 30% MSCI European Transport and Infrastructure Index as a benchmark.
2. Using the Axioma European Medium-Term Factor Model.
3. Based on a constant $100M AUM and using trading costs of 16bp (two-way)
4. Zero trading costs and assuming a "perfect" factor hedge.

Author Bio

Peter Hafez
RavenPack

"Peter is the head of data science at RavenPack. Since joining RavenPack in 2008, he's been a pioneer in the field of applied news analytics bringing alternative data insights to the world's top banks and hedge funds. Peter has more than 15 years of experience in quantitative finance with companies such as Standard & Poor's, Credit Suisse First Boston, and Saxo Bank.

He holds a Master's degree in Quantitative Finance from Sir John Cass Business School along with an undergraduate degree in Economics from Copenhagen University. Peter is a recognized speaker at quant finance conferences on alternative data and AI, and has given lectures at some of the world's top academic institutions including London Business School, Courant Institute of Mathematics at NYU, Columbia Business School, and Imperial College London."