Driving Investment Performance with Alternative Data

11th November 2015

Gene Ekster, CFA

Contents

Overview

The rapid uptake of alternative data research by both fundamental and quantitative institutional investors is shifting the investment landscape and creating innovative sources of alpha generation. The field is still in the early phases of development, yet depending on the resources and risk tolerance of a fund, multiple approaches abound to participate in this new paradigm.

Regardless of the path taken, the process to extract benefits from alternative data can be challenging. However, with the right tools and strategy, a fund can mitigate costs while creating an enduring competitive advantage.

This paper will examine the changes alternative data is creating on the buy-side and the strategies hedge funds and mutual funds are using to take advantage of these new research assets. Components of this paper first appeared as articles published in Integrity ResearchWatch.

Alternative Data Research

Many investment professionals are now familiar with the concept of non-traditional research. For the sake of clarity, this paper will establish the term “Alternative Data” to refer to research derived primarily from either raw or aggregated data, in the context of the investment process. Alternative data operations contrast with the established investment research traditions of relying mostly on financial information that’s obtained from sources such as company filings, investor presentations, PR releases, media coverage, management meetings, historic market prices, etc. This information is typically accessed through various financial forums and services such as Bloomberg.

Alternative datasets are often less readily accessible and less structured than existing established commercial sources. Examples of such data include point of sale transactions, web site usage, obscure city hall records and other information which contain ”exhaust data” – datasets that might be overlooked as a byproduct of some operation, yet valuable to investors when aggregated and combined. Other examples include URL clickstream data, social media, satellite imagery, shipping container information, product reviews, price trackers and an ever growing list of novel sources of alpha containing alternative datasets.

In today's alternative data market, a fund can participate in this new field via a combination of:

1. Directly licensing third party datasets and growing an internal R&D practice to mine those datasets.
2. Licensing ready-to-use analysis from intermediary providers.
3. Establishing an internal data gathering capability such as running a web harvesting operation or primary research.

In practice, the mixture of sources is assembled over time and often follows a logical sequence. For instance, licensing ready-made datasets gives funds the lowest cost and most rapidly accessible exposure to alternative data.

If adding an internal R&D team is sensible for the fund, expertise in consuming ready-made alternative data would help guide the new endeavor. Arguably, building a team is the most expensive, but also the most comprehensive method of deriving sustainable value from alternative data.

Building an In-house Capability

An increasing number of funds across the investment spectrum are building internal alternative data groups, sometimes referred to as R&D. A long term vision in a data driven investment process and internal demand for information are some of the factors motivating funds to grow their in-house alternative data competencies. These groups generate value by acquiring unique datasets and analyzing them internally to guide investment decisions.

A full scale R&D group is a considerable capital investment that often requires a significant ramp up time of up to a year to start producing valuable insights. This kind of infrastructure has a different payoff cadence from other research outlays such as expert networks which deliver a faster upfront return. Only select funds are willing or able to invest into a dedicated R&D group, but doing so can provide a substantial lead in the ability to consume alternative data, thus the advantages of scale play a key role.

As with any technology-focused endeavor, the team and the infrastructure can make the difference between an exercise in frustration and a humming alternative data machine. Luckily, a build-out of an R&D group can be a scalable undertaking with a minimum staff of just two or three people. Some funds start there before growing to a 20+ person enterprise.

Team build-outs need upfront capital to create the technology, source data, grow a talent pool and invest in product development. Growth is best assessed by initially emphasizing product output metrics, then using bottom line ROI metrics (P&L impact vs. cost) in the later stages.

Alternative data is poised to revolutionize the research process for both quantitative and fundamental investors. It can be thought of as an arms race or simply a competitive advantage. Either way, once adopted by a portion of the market, the rest of the market has to adopt to remain relevant. Within the scope of an alternative data operation, an R&D team is pivotal in sourcing datasets, developing them, and ultimately contributing to P&L with data driven research.

Data Driven Research vs. Quantitative Strategies

It would be easy to mistake alternative data driven investing with automated quantitative strategies. In practice however, an experienced specialist is usually needed in order to monetize products from an internal R&D team. That's because the signals generated by alternative data are an order of magnitude more complex, noisy and unstructured than the refined data feeds powering traditional quantitative portfolios.

Despite enormous leaps in big data technology in the recent years, there is currently no reliable way to automatically convert vast pools of heterogeneous and unstructured datasets into automated trading decisions. A number of funds have had success automatically trading social media or news sentiment streams, yet typically these systems are tailored to process only a few limited aspects of the data.

Given the difficulty described above, most alternative data operations amplify the benefit of having fundamentally driven teams, rather than substitute it. At the same time, automating alternative data trading is a small, but growing trend which holds tremendous potential.

The Role of the R&D Team

The overall scope of the R&D team is to generate value by sourcing datasets and then using the information supplied by those data sets to create an internal focused "product." The skill sets located within R&D should be well suited to ensuring that, for any complex dataset, the investment teams receive a product that is consistent and relevant.

R&D outputs are delivered periodically or continuously to internal investment teams and would include various metrics, signals and best ideas. If the fund is fundamental equity, then R&D would provide GAAP and non-GAAP operational metrics such as revenue estimates, unit volumes, prices, inventory, subscriber counts, etc., and any insights that the fundamental portfolio managers can integrate into an investment thesis. If R&D is servicing quantitatively driven investment teams, then R&D's products would include structured data streams tailored to be used as input to automated trading systems.

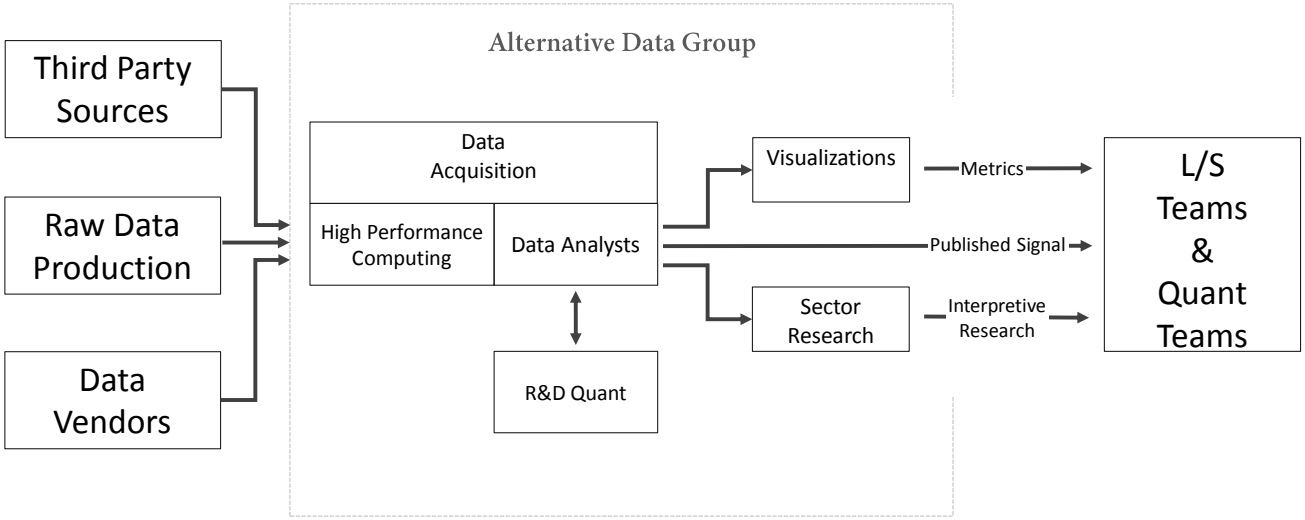
A unique property of alternative data is that the same dataset can generate vastly different insights. It's akin to how the same credit bureau inputs can produce both a FICO score and a Vantage Score which may give very different results. One dataset can even generate multiple dimensions of metrics for different industries. Construction permit data, for example, can

generate both revenue estimates for construction material companies and also inventory estimates for homebuilders.

Organizing R&D

R&D group activities can be organized into several subdivisions. The sourcing team explores new purchasing opportunities and maintains relationships with vendors. A high performance computing group is needed to set up the databases and toolset environments to ensure smooth operations.

Analysts and data scientists mine the data for insights, remove bias and create estimates. Finally, if the fund utilizes quantitative trading, an internal R&D quant can be employed to create tailored quant strategies directly driven by the alternative data streams.



The investment in an R&D team can provide signals and inputs to both qualitative and quantitative investment processes. R&D teams also can also extract multiple insights from one dataset.

R&D and Buy-Side Investment Teams

Hedge funds' and Mutual Funds' development of in-house alternative data teams is a relatively new phenomenon aimed at helping funds monetize the opportunities presented by alternative data. Those teams, sometimes called "R&D" or simply "Data Groups" generate value by first acquiring unique datasets and then developing them into insights leveraged by internal portfolio management teams.

Much of the R&D process is highly technical, employing terabyte scale databases, machine learning algorithms and data scientists, a world that appears removed from the day-to-day operations of traditional fundamental investors. Yet, for a fund to derive the full benefit of

data, it must integrate the human intelligence of investment managers with the intelligence derived from alternative datasets.

Practically, the integration is realized via a continuous feedback loop between the portfolio management teams and R&D, where the combined experience of data analysts and investment professionals are employed to drive the process. Unlike some sell-side research models, internal R&D teams should not exist as a one-way operation feeding information upstream to investment teams; doing so would defeat the purpose of building in-house capabilities.

Most data research projects are lengthy and resource intensive, therefore, a potential market investment use-case must be taken into account before committing to a dataset. The frequent interaction between the data talent and the investment professionals allows multiple strategies to be explored simultaneously, some failing quickly and others succeeding spectacularly. Ultimately funds gain the most benefit from employing at least one person to bridge the gap between the traditional investment process and the science of data research.

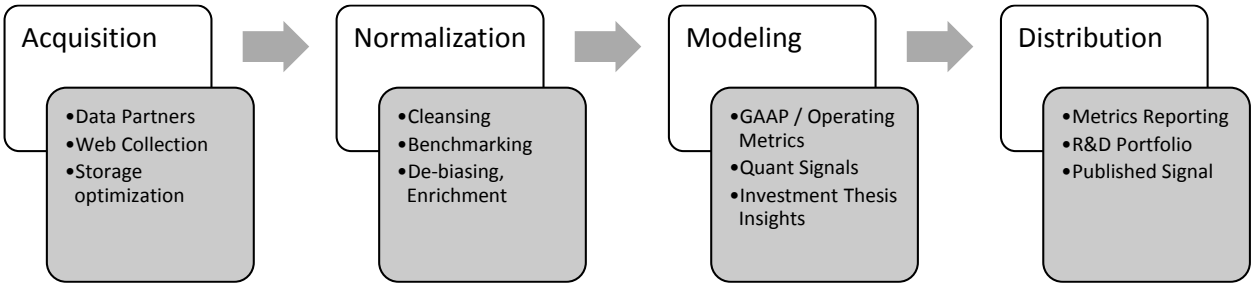
Incubation Programs

An internal alternative data group can supply a non-obvious benefit: A firm-wide uptick in data savviness and an enhanced Portfolio Manager (PM) development program. Alternative data expertise can propagate from the R&D group into the rest of the firm including the investment teams, thereby increasing the fund's overall data aptitude which is quickly becoming a required competency for all investment professionals.

This can be accomplished via an analyst incubation program, where financial analysts from PM teams spend a portion of their time training inside of the R&D group. Once analysts revert back to their original teams, they transfer with them a wealth of technical skill sets that can add significant value to the group, especially if they end up trading with the same dataset on which they received their apprenticeship.

R&D Process Flow

The typical R&D analytical process is highly technical and a full discussion of the details of is beyond the scope of this paper, but a brief highlight follows:



1. **Acquire and Evaluate Datasets**: R&D's sourcing group contacts potential vendors, assesses compliance, acquires data samples, evaluates ROI, and establishes commercial relationships. A slightly modified version of the same process would apply to internally or externally harvested web data.
2. **Normalization**: A broad category which applies to technical aspects of data processing, but not necessarily modeling, including converting unstructured data to structured, cleansing, aggregating. Often R&D must secure datasets just for the purposes of calibration, such as publicly available data from the BLS (for instance the BLS Consumer Expenditures Survey), Census, Federal Reserve and other sources. Removing bias is a key step enabling broader insights to be gleaned from a dataset and involves paneling and creating optimized weights to reduce distance to one or several benchmark datasets. Compliance related scrubbing of datasets is also performed in this step.
3. **Modeling**: Once the dataset sample is normalized and representative of the population being measured, R&D proceeds to devise models to predict both past and future metrics used in the investment process. Typically these are GAAP and non-GAAP operational metrics such as revenues, margins and other KPIs specific to a particular industry, including unit sales, average prices and revenue per user, subscribers, churn etc. Currently, most alternative data driven investment strategies are fundamental not quantitative, thus modeling securities prices directly from the raw data is not yet a common practice.
4. **Publishing and Distribution**: R&D distributes its insights to investing teams via internal products including Excel reports, dashboards, programmatic access to the structured data and qualitative analysis. Well-designed internal distribution systems can help address the thorny issue of attribution of value to the various groups involved in the data driven investment process.

Computing Infrastructure

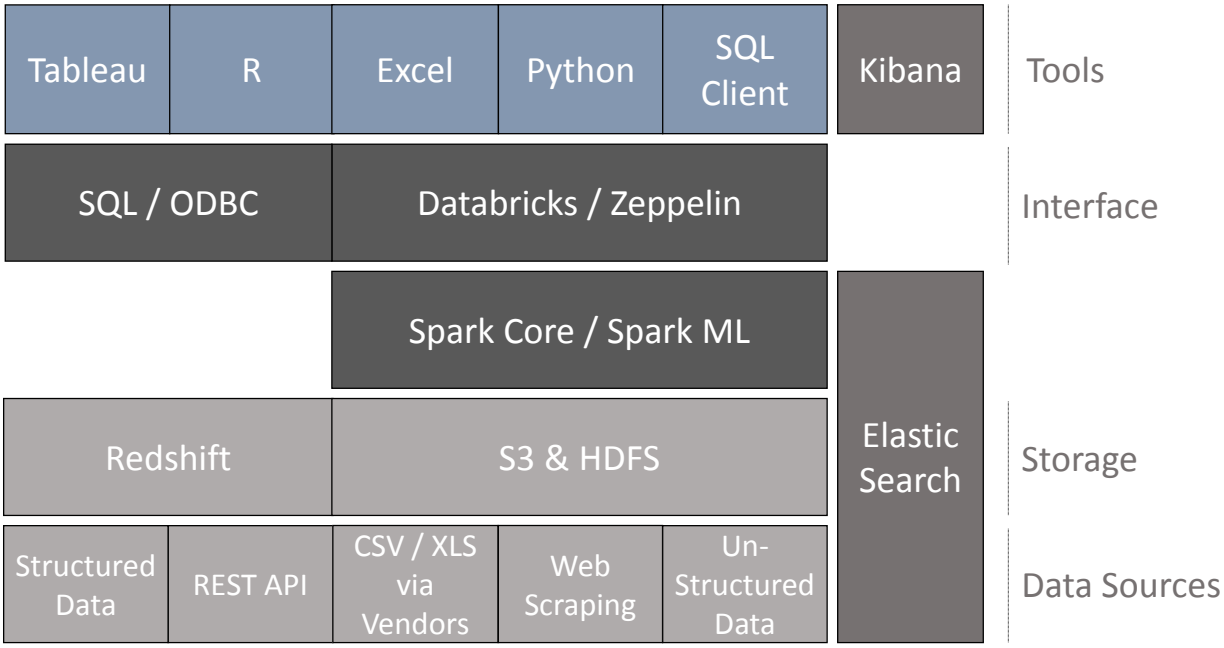
A full discussion of the hardware and software components of well-designed infrastructure is beyond the scope of this paper. The essentials, however, are that a technology stack of an alternative data group needs to fulfill several key requirements.

- Ingest and store vastly heterogeneous sources of data. When enumerating the different formats that a system might have to consume, it's useful to include edge cases as well as common scenarios.
    - Structured relational data – CSV, direct ODBC/JDBC
    - JSON files – custom formats and nested
    - Streaming event data – i.e. Twitter, Clickstreams
    - HTML file dumps from a web harvesting operation
    - Sell-side research notes – PDF
    - Excel – Various vendor specific layouts
- Perform ad-hoc analysis – Simply put, data analysts must be able to run queries against the datasets in "human time", results being returned in a few minutes at most.
- Security and access control – Comply with internal control guidelines and adhere to compliance standards. Capable of PII (Personally Identifiable Information) scrubbing processes.
- Disseminate the results of the analysis to internal users, typically via Excel and BI tools akin to Tableau including access, when possible, to the source or WIP data, to enable an ecosystem of downstream innovation.

Technology Stack

There is no such thing as a one size-fits-all technology stack. Mostly for reasons of talent acquisition and retention, it's healthy to have a strong bias for technologies that enjoy mass adoption (and a moderate bias for my Alma mater's Berkeley Data Analytics Stack). The below recommendations can be used to get up and running on a budget that gravitates towards caution.

Alternative Data R&D – Technology Stack



- Infrastructure
  - Cloud – Amazon's array of cloud AWS infrastructure services scales up for all but the largest of alternative data operations. Juggling between EC2, S3, Redshift and EMR isn't without challenges, but typically the on-demand capability wins over in-house alternatives. Strong competitors include Google and Microsoft.
  - Local – Workgroup servers are quick to set up and can accommodate ever larger datasets. Sub $10k workgroup stations with 384 GB of RAM are commonplace and offer significant cost savings over a comparable distributed cloud capability.
    - For datasets between 0.5 and 4.0 TB, local storage can offer incredible performance and flexibility, especially if it's served up by a FusionIO or similar PCIe based SSD technology. Expensive per unit of storage, this local setup is especially relevant for R&D operations in need of extremely rapid product development.

        - SSD Raid is an unproven technology option which combines the speeds of individual SAS/SATA SSDs with significant cost savings via a hardware RAID card. The main bottlenecks in this setup are the limited IOPS that all but the most cutting edge controller cards can manage and more importantly the non-cached random read latency, arguably the most important benchmark, cannot compete with the PCIe SSD option. However, SSD RAIDS offer quite a big bang for the buck.
- Data storage engine – Where and how your data waits to be structured, analyzed and disseminated. Options include:
    - AWS: Redshift, S3
    - Local: Postgres, SQL server
    - HDFS (local or cloud): Unstructured or ready for Spark in Parquet files.
- Compute framework
    - Spark: An alternative to batch processing oriented MapReduce, Spark has massive community support, full documentation and easy to use libraries in Python, R, Java and native Scala. Built-in support for machine learning means that data analysts can run complex algorithms directly on the data without having to sample.
    - Elasticsearch: More of a search engine than a compute framework, Elasticsearch excels at extracting information from documents containing dated events and can help an investment team find quick answers in a hodgepodge of datasets, spreadsheets and research documents.

- Machine learning and query languages
    - SQL – SQL based systems still account for the majority of the heavy lifting in data operations. The common denominator for all data analysts, this is the bread and butter of talking to structured data.
    - Python – Quickly becoming the ubiquitous language for all but the highest performance-hungry ML tasks.
    - R – A slightly more scientific language compared to the more general Python; R is often the fastest language for prototyping data products; it has more data science community support than Python (for now), meaning that complex algorithms can simply be downloaded (from CRAN) rather than developed from scratch.

    - Development environments
        - The notebook concept that IPython helped bring to the main-stream is quickly gaining popularity for good reason, it allows for web-based interactive data analytics, thus enabling rapid prototyping and deployment to production. Jupyter (the evolution of IPython's notebook), Apache Zeppelin, Databricks and to a lesser extent Rstudio, are all interactive development shells capable of doing ad-hoc data analysis and visualization. Some, like Databricks and Zeppelin, also manage parallel computing infrastructure and make it easier for teams to develop together.
- Output – Ultimately insights from R&D must be presented to the investment teams via a human readable interface, be it revenue forecasts or investment theme metrics. Excel, Tableau and Kibana (paired with Elasticsearch) are just some of the tools used to communicate finalized results to the portfolio teams.

The implementation details of a technology infrastructure is one of the most important decisions when building an in-house alternative data capability. Key stakeholders in the buildout must include those who are familiar which the entirety of R&D supply chain and investment process.

Research & Analytics Providers

Even if a fund does not have the capital, skillset or simply the strategic need to build an internal group, it can enjoy partial benefits of alternative data without the sizable investment. Funds can turn to intermediaries who are in the business of acquiring third party datasets, analyzing them and selling custom or syndicated research to the buy-side clients. Research offerings from UBS' Evidence Lab and Eagle Alpha are good examples of alternative data intermediaries driven by analysis of many alternative data sources including web harvesting blogs, review forums and social media among many other sources. In addition, Eagle Alpha provides access to the underlying data and enables buy-side firms to do proprietary analysis using their online platform. Bloomberg's Polarlake acquisition is an example of ever larger data focused players joining in the rapidly evolving field of intermediary alternative data providers. Many aspects of the R&D process flow detailed above apply to research sourced from external providers as well as to the raw data.

Alternative Data Research Compliance

Alternative data research compliance is an increasingly important topic and subject of intense discussion in part due to its regulatory ambiguity. Risks are sometimes broken down into regulatory compliance risk and headline risk, but with little guidance to go on, research practitioners are often left to speculate on the proper course of action.

One compliance concern relates to access investors might have to personally identifiable information (PII) in datasets of consumer behavior, however, unlike direct marketing or

advertising, the investment research industry has no clear incentive to possess or analyze individuals' information. Investing decisions are based solely on aggregated data where the behavior of any one individual is fully anonymized. Investors are concerned with the trends of the many, not the trends of the one or the few.

Moreover, compliance teams on both the vendor and client sides tend to be highly proactive in ensuring that data sources are compliant with access controls and PII scrubbing processes in place. Therefore, data is typically aggregated and scrubbed in the early parts of the data supply chain, far upstream from investment professionals.

Information security has long been a top priority in the investment industry and many funds mandate their internal teams as well as their data suppliers to adhere to strict risk control standards. Lack of incentive, active PII scrubbing across the supply chain, stringent security standards and ubiquitously proactive compliance teams all mean that identifiable consumer information is reassuringly not a cause for significant alarm in the alternative data ecosystem.

Insider Trading

All investment research including alternative data research is governed by a set of principles addressing insider trading issues. Distinct compliance risks are inherent and are tackled in each of the various research channels. For instance, expert networks deserve an extra degree of diligence around the sensitive information discussed during the phone interview sessions. [Integrity Research's compliance best practices for expert networks can be viewed here.]

One comforting aspect of alternative datasets is that they are sourced from parties that do not have a fiduciary relationship with publicly traded firms and are broadly available (yet are difficult to aggregate). Some examples are web harvested information or municipalities' public property records.

Another facet is that consumer behavior datasets are typically sourced from specific spending channels, have relatively small sample sizes and significant biases. Therefore an analyst has to commit hard work, using their education, judgment and expertise to create a meaningful investment thesis. These are insights derived, not obtained, which is a key difference in applying the insider trading legal lens.

Best Practices

While compliance and internal regulation are widely practiced in the alternative data research field, there exists a need for an industry-wide best practices standard. Such a standard should address PII obfuscation and access scheme requirements among other issues. It would help round out the uncertain interpretations of existing regulatory guidance.

In the meantime, compliance professionals and decision makers can benefit from proactively creating internal guidelines for data operations. Publications like the NIST 800-122, which provide guidelines for protecting PII, are useful when developing internal best practices.

However, due to the relative lack of case precedent, a great degree of diligence needs to be exercised to assess the risks.

An example of a best practice is for a company to set up separate control environments in its data flow architecture. As the raw data enters the organization, it can be initially staged in a restricted access zone. This is where the PII scrubbing and other privacy related data processes are performed. Once cleansed, data can move into the general analyst access environment where human and machine intelligence can manipulate the compliant data.

If a company is directly or indirectly engaged in web harvesting, then best practices include appraising the website's terms and conditions, paying special attention to clickwrap agreements, having a written policy on handling incoming complaints and limiting the outgoing HTTP traffic. Researching the following cases is the first step to understanding the legal gray area around web harvesting:

- Cvent, Inc. v. Eventbrite: The Computer Fraud and Abuse act (CFAA) was used to argue that Eventbrite's harvesting of Cvent venue data was prohibited by the TOU should be considered unauthorized access thus in violation of the CFAA. However the courts dismissed the case on the grounds that Cvent's website did not require a login or a browser wrap, therefore the TOU contract is not enforceable.
- Craigslist v. 3Taps : Craigslist sent 3Taps a cease and desist after learning that 3Taps was harvesting Craigslist's listings. 3Taps continued scraping and was successfully sued and stopped by Craigslist. The courts found that while Craigslist has no browser wrap agreement, it has the right to revoke access to its services via the cease and desist; 3Taps should have complied.
- Other notable cases: CollegeSource v. AcademyOne; Pacific Stock, Inc. v. MacArthur & Company, Inc; Fidlar v. LPS; Ticketmaster v. Tickets.com; White Buffalo Ventures LLC v. University of Texas at Austin; Facebook Inc. v. Power Ventures Inc and many more cases shed light on the courts' current thinking regarding harvesting web information.

Unfortunately there are no well-known cases which directly address the use of web harvesting in an investment context. Courts' opinions need to be individually interpreted by compliance professionals.

Critically, best practices ought to diligently measure the individual risks of a data operation, with each risk evaluated separately and explicitly. Due to its fragmented mosaic of widely available sources and arms-length distance from any given traded securities, alternative data research can exhibit lower levels of compliance risk than some other forms of primary research.

Nevertheless, being up to date with the latest laws and having a rigorous compliance strategy are key success factors for any firm in the alternative data supply chain. Compiling those strategies into a best practices document is a smart move for individual organizations and creating industry-wide guidelines is a sensible step for the alternative data research field as a whole.

Conclusion

For a fund wishing to participate in the growing alternative data world, building an in-house R&D team is a competitive advantage. Even then, it is important to have a strategy which addresses data sourcing, internal team collaboration and compliance. In the near future, as more funds consume unique data sources, competing against similarly endowed rivals will be the norm and funds without an alternative data plan may find themselves at a disadvantage.

These alternative data sources are also being incorporated into the fundamental buy-side decision making process. While not without its challenges, this move into alternative data is an unstoppable transformation with benefits that will ripple across the investment spectrum.

About the Author

Gene Ekster, CFA has been involved with the alternative data industry on the buy-side as head of R&D at Point72 Asset Management (formerly SAC Capital) and in independent research as a Director of Data Products at 1010Data and Senior Analyst at Majestic Research / ITG. Gene serves as an advisor at Eagle Alpha. He holds a degree in Artificial Intelligence and Cognitive Science from UC Berkeley and an MBA from Cornell University.

Currently, Gene works with asset management firms and data providers in a consulting capacity to help integrate alternative data into the investment process. He also helps organizations evaluate data assets' application to the investment industry and facilitate potential introductions to funds interested in alternative data research. He can be reached via LinkedIn. Additional information about alternative data can be seen in his presentation on the topic.