TradingTech Insight | Whitepaper

Alternative Data: Application and Best Practices for Investment Management Firms

Commissioned by Dow Jones Newswires

DOW JONES

@TradingTechIns

Search TradingTech Insight

www.tradingtechinsight.com

A-Team Insight

DOW JONES

dowjones.com/newswires

INTRODUCTION

The modern propensity to measure all aspects of human endeavour has created a vast quantity of data as a byproduct of primary commercial activities. In recent years, innovative trading firms involved in quantitative analysis and fund management have tapped into these sources – often known as alternative data – to gain unique insight into underlying industry segments or market behaviours.

The promise of alternative data has expanded significantly as the practice became almost mainstream, particularly as more – and sometimes more exotic – data sets became available. It's now almost commonplace for retail market analysts to monitor satellite imagery of supermarket chains' parking lots for indications of activity levels.

But sourcing, evaluating, integrating and using alternative data is a non-trivial exercise. Many providers of alternative data are experts in their field, but inexpert when it comes to data provision. Alternative data sets may be incomplete or unverifiable; they may be unstructured in format and difficult to integrate; they may include data that isn't permitted for redistribution under new privacy rules. And there may be limited or no archive available for back-testing.

This paper looks at how different alternative data sets may be of use to trading firms. It looks at the issues surrounding effective use of alternative data in gaining competitive insight through indications on market sentiment and predictions of future events. Finally, it discusses best practices and approaches for practitioners attempting to take advantage of this promising data type.

DOW JONES

dowjones.com/newswires

WHAT IS ALTERNATIVE DATA?

Finance has always been a data-driven business. But the data that's historically driven financial markets is data that financial markets and the companies they track themselves produce: prices, volumes, fundamentals, estimates – all are created as a byproduct of trading or by financial institutions participating in those markets.

Recent years, though, have witnessed the emergence of data sets that are not endogenous to financial markets but are instead exogenous. These data sets – commonly referred to as alternative data – don't have their origin in financial markets but are instead the byproduct of the global data explosion of the digital era, generated by the activities of businesses of all types. They are often offered in an unstructured format, making them different from most traditional data sets and challenging to integrate and consume.

The proliferation of these new data sets means that it's now possible to know practically anything anyone would want to know about any aspect of the global economy. Somebody, somewhere is measuring the activity any analyst might want to study. And this has been at the heart of the growing interest in alternative data.

The early adopters were mostly quantitative hedge funds, which are by nature both innovative and curious. These firms had been the first to use microcomputers, the first to use stochastic calculus to price derivatives and the first to adopt high frequency trading techniques. For them, alternative data represented an opportunity to seek out new ways of gaining insight into markets and industries, giving them the kind of competitive advantage their early exploits had given them.

Today, quant funds like WorldQuant – a spinoff of New York-based hedge fund Millennium Management – AQR Capital Management, TwoSigma Investments and Citadel Securities lead the pack as pioneers of how trading firms evaluate and deploy emerging alternative data sets into their trading and investment strategies. Quants are perfectly designed to exploit data. They analyse data sets and systematically look for the signal that comes out of that and they execute a trade against it. This approach maps well to the idea of using alternative data to enhance what they've already been doing.

Alternative data has emerged as a powerful force, not only because the world is awash with new data sets, but also because of the increasing challenge of alpha generation. Knowledge is money in capital markets. But because a lot of traditional data is everywhere, finding a niche to exploit can be difficult. Being first is no longer enough. Low-latency technologies have been democratised, so speed of execution is no longer a guarantee of market success.

Regulation has also played a role in contributing to the interest in alternative data. MiFID II unbundled research services from broker execution, with the result that investment firms now for the first time have to pay for research. This has led to a new scrutiny of the value of broker-supplied research and impetus to look elsewhere.

The above factors have conspired to create opportunity for investment firms to derive value from previously ignored and/or emerging data sets, often termed the 'exhaust' of other industries' activities. These data sets have the potential to offer unique insight into how industry sectors work, indications on market sentiment and predictive data of future events.

DOW JONES

dowjones.com/newswires

ALTERNATIVE DATA IN FINANCIAL MARKETS: DATA TYPES AND USE-CASES

The term alternative data largely refers to the information that is generated by the increasingly digital nature of human endeavour. The acceptance of the fact that most human activities are these days measured and monitored has led to an explosion in the amount of quantitative data available to those who require it. And financial markets place great value on new sources of information as a way of gaining an edge over one's competitors.

Informed observers of this new alternative data market estimate the number of databases useful to financial markets participants could be as high as 25,000. But the total population of databases globally is far larger – probably measured in the millions – making the challenge of identifying valuable data sets significant.

Alternative data includes many non-standard data sets, including satellite imagery, social media posts, geolocation tracking, website visit statistics, retail traffic figures, shipping numbers, data from the Internet of Things (IoT) and other databases of industrial/business metrics. A recent email from an alternative data consolidator offering to pay up to $50,000 for useful databases listed a broad array of data types required, among them databases relating to online travel bookings, physical sales of video games and in-app sales on social media sites including Tinder, Uber, Lyft, Instagram, Facebook and SnapChat.

But alternative data can also refer to traditional data, used in a new way. For example, some firms have interpreted cable subscription cancellation rates as an indicator of real estate market performance. Others have tracked car insurance policy take-up rates to provide intra-report predictive analysis of car production, which is often reported by the big auto makers on a monthly basis. And in some respects, news services represent perhaps one of the earliest forms of alternative data, in the sense they are exogenous and often unstructured in format.

Given the broad range of data sets coming onstream, many applications in financial services are being identified, among them alpha identification by quant traders, sentiment analysis, predictive analytics, market surveillance (for identifying financial crime) and research verification.

Importantly, many financial markets practitioners are using alternative data in tandem with traditional data sets. Financial institutions are seeking to 'triangulate' alternative data with traditional data sets in order to augment those data sets or create new and unique derivative insights. The challenge here is linking many sets of potentially unstructured data to the existing core set of databases used in business processes.

An example of this tandem approach is the analysis of transponder data from planes: Plane ownership registration data and corporate linkage data can be combined to understand the travel patterns of corporate jets and executives who may be involved in M&A activity. Similarly, alternative data can be used to augment original research about a particular industry segment and current market data and fundamentals in order to reaffirm a trading signal or trend. Indeed, there is a growing propensity among research shops to support their offerings with alternative data sets to provide color to their narratives.

Firms are mixing freely available data sets that describe industries, products and people with core financial data (prices, interest rates, macroeconomic data, earnings estimates, fundamental data, etc.) and finally, with alternative data sets to derive a unique view of the markets in which they participate.

The alpha-seeking element of alternative data has made most headlines. But in the broader investment management landscape, firms are using alternative data as part of their more run-of-the-mill investment business. One well known hedge fund buys in alternative data as part of perhaps a thousand data points for its research, which informs investment decisions going forward five years or more. It's not always purely about alpha. Many firms are keen not to miss out on an opportunity, in which sense it's a lot like the low-latency space where it's no longer about being first but rather about not being last to hit a fill.

DOW JONES

dowjones.com/newswires

ALTERNATIVE DATA: OBSTACLES TO ADOPTION

The promise of alternative data may be inspiring, but dealing with it is no easy undertaking. There are many obstacles to consuming alternative data as part of a firm's business activities.

Alternative data is unstructured, making it difficult to receive, integrate and consume

By its nature, alternative data comes in many different formats, many of them unstructured. It may take the form of social media posts, video content, images or Excel-based lists, for example. While the bigger financial institutions have invested in so-called Big Data technologies, often based on sophisticated Hadoop and Spark server implementations, second-tier firms that aren't in the quantitative hedge fund segment often don't have the talent or resource to handle this level of pre-filtering/searching for consuming alternative data sets.

The data is often incomplete or unverifiable

An incomplete alternative data set can undermine its value. A gap in a time-series database can render it useless for historical back-testing of trading models, for example. Given their unstructured nature, many alternative data sets aren't suitable for driving the kind of intense calculations often needed for quantitative analysis. But certain alternative data sets that are structured – like histories of web-site activity, for example – can be used to design, test and operate unique trading strategies. What's essential, though, is that any such data set has the kind of deep archive available that meets the quants' needs.

Incomplete data can impact a data set's value in other ways. A new data source may be missing a crucial field or data point that would transform it from nice-to-have to must-have. Similarly, it may not be possible, at first sight, to confirm the veracity of a new data set.

The data is of limited use

Many emerging data sets may have a short shelf life, or may only be of use if consumed on an exclusive basis; once others have access to it, its value to a market practitioner may fall off rapidly. Databases may be limited in scope, covering narrow segments of a given marketplace, for example, which may restrict its usefulness.

These concerns may be the biggest hurdle to adopting alternative data. Clearly, the applicability of new data sets is a major factor in firms' investing the resource to evaluate and integrate them, about which more later. Some data is not

immediately tradable, for example, but could be useful as a benchmark for use in investment or trading; although this may not be immediately apparent to the firm. Likewise, these concerns also limit potential revenue for the big data vendors to integrate new data sets, and goes some way in explaining why, to date, they haven't jumped in with both feet.

The right data is hard to find

Before they can start using alternative data, firms first need to find the kinds of data sets they are seeking. Discovery of alternative data is a skillset that's new to most trading organisations. Most firms don't have a resource dedicated to searching out valuable data sets, and so it may fall to a staff member or team who are already deployed (and busy) with other activities.

Privacy concerns

Because much alternative data is the byproduct of a separate commercial activity, it often comes with data that relates to the originator, or customers or partners of the originator. This can pose issues with respect to privacy.

Historically, originators' internal data sets have been subject to only cursory restrictions on usage to protect third parties' data. The EU's General Data Protection Regulation (GDPR), implemented in the spring of 2018, has tightened things up greatly, introducing eye-watering financial penalties for transgressors.

This means that acquirers of alternative data sets that contain client or other third-party details need to ensure any data that could run afoul of GDPR needs to have been anonymised even before it is brought on board. Ensuring this has happened can be challenging when dealing with originating suppliers that are inexperienced in selling data.

The value of the data is not easily verifiable

It's not a trivial task to assess the value of a new data set. It's likely that the data a firm is reviewing hasn't been optimised for external consumption; rather, it is maintained for the originator's own purposes. This means it will require analysis using data science techniques to get a clear view of its value, some of the best practices for which we discuss below.

Inexperience of data originators / providers

Even when an investment firm has identified a desirable data set, there is still the inexperience of the data originators / owners to deal with. Most financial firms are used to dealing with professional data providers; with alternative data they may find themselves negotiating via the bureaucracy of the organisation that owns the data they want. It's often the case that the most valuable data is to be found in the database of a company that isn't primarily in the data business.

This can have major implications in terms of commercials, data quality and delivery issues. Data originators often give the task of monetising data sets to staff without any idea about how to do this, or indeed about the value of the data or how it might be used.

If the firm intends to build a trading strategy around a new data set, then it's imperative that it secures a reliable delivery mechanism to ensure the data arrives at the right place at the right time. This may require the firm to build teams to manage the process or indeed infrastructure to ensure reliable delivery. Finally, there is the question of documentation, to ensure proper use and administration.

General data quality issues

For firms intending to build models and strategies around new data sets, data quality is a key requisite. The challenge of ensuring data quality – cleansing, de-duping and filling gaps – expands significantly as more new and unstructured sources emerge. This may require firms to adopt more robust data validation and cleansing processes in order to make use of the new data sets.

BEST PRACTICES IN ALTERNATIVE DATA USE

It's clear that alternative data – across its various forms and formats – holds great promise for investment managers looking for new opportunities, whether in immediate trading opportunities or more contemplative approaches to long-term portfolio strategy. It's also clear that by its very nature, alternative data poses challenges for those seeking to make use of it, for all of the reasons outlined above.

Investment management firms seeking to exploit the opportunities alternative data may present, however, can take proactive steps to ensure they get return on their investment in time and resources as they approach new data sets.

Assessing the value of alternative data sets

The first step is to establish the requirement for a non-traditional data set to fulfil a business need, and the type of data that would meet that need. But procuring the data itself is often not straightforward, as it is unlikely to be available via the investment firm's main data suppliers; the investment manager must go and find the data, a function most firms don't have on tap.

Complicating the picture: undiscovered data is what's frequently needed. The originator or source may not be aware that their data is in demand. Even if they do, it's likely that the data set in question is created as a byproduct of the principal business activities of the source, which means it's highly improbable there is a data sales function to deal with.

In these circumstances, it's also improbable that the originator has 'productised' the data set to make it consumable by others. For the investment firm this means that in order to determine whether it can be put to the use intended, they will need to employ data science processes to ascertain the quality, completeness and timeliness of the data to assess its value.

Market success rates indicate that it may require analysis of 50 or so data sources in order to arrive at one or two usable data sets. Says one practitioner: "Everyone is looking for that unicorn, but it's hard to assess the quality. It's costly to do, and you could easily draw a blank."

To cut through the noise, this practitioner and others recommend a series of filters to whittle initial lists of prospects down to a more manageable set of potential sources. It's practicable to identify five or 10 prospect sources, and apply the data to models to test their usefulness. These testing models should analyse how the data will be used and whether the actual data is fit for this purpose. They should also take into account whether the originating sources are capable of or willing to perform the necessary ongoing data quality functions and delivery responsibilities.

Even where the value of the data set is compelling, dealing with a difficult or inept supplier may eradicate that value.

As the procurement team gets closer to selecting the data set in question, more resources should be devoted to quality checks. This may include analysis of underlying code where appropriate, back-testing against models and applications that will use the data set, and talking to internal users about the potential value. While there are tools now available to help in this process, it remains a time-consuming task.

Approaches to integration

Alternative data comes in many different formats. Drawing these data sets into a form that can be applied to analytical trading and investment models can be challenging. Alternative data is often unstructured, and there is no general purpose messaging middleware or other platform solution for this kind of data, as there is for real-time data or historical/reference data.

As a result, early adopters have often built their own technology platforms in order to consume alternative data, creating proprietary solutions at significant expense. Some have turned to generic technology providers like Palantir and others, while others still have tapped into innovative point solutions like knowledge graph integrators and visualisation tools.

These early adopters have implemented Big Data-type technologies like Hadoop and Spark servers to meet the challenge of integrating disparately formatted data types. The less structured the format, the greater the challenge of successful integration, especially where back-testing of models is involved. Multimedia data like video and voice are increasingly searchable and manipulable but it remains difficult if not impossible to manage them as time-series data that can support defensible analysis for trading and investment purposes.

But not all alternative data is unstructured by nature. Firms routinely collect histories of telephone call records, web-site requests and forms, and other types of structured but non-traditional data. These types of data may represent the 'low-hanging fruit' of alternative data that is often overlooked because it's not as glamorous or unusual as satellite images or video audience records.

Indeed, it's often these structured but perhaps less eye-catching data sets that can be most easily incorporated into trading models. As mentioned above, a deep archive of consistent, validated and anonymised time-series data, sourced from a mundane activity like tracking residential house insurance policies or website behaviours, can be more useful to quants than the more arcane data types often cited by promoters of alternative data.

These types of data sets have the potential to move rapidly to mainstream acceptance as indicators of future behaviour, ultimately moving out of the shadow

of alternative data designation. In one well known example, ADP payroll data – a data set created as a byproduct of a separate commercial activity – emerged as a prime indicator of net new jobs in the US economy that analysts ignore at their peril. Similarly, news produced to fill newspapers is widely accepted as a form of market information – particularly in its real-time newswire format – rather than as an alternative data set.

As the market matures, more help is emerging, often from specialist firms that are focusing on helping data originators shape their data offerings in order to make them more easily consumable by investment firms. Consultancies and data platforms like NeuData, Eagle Alpha, Yipit and Quandl (which was recently acquired by Nasdaq) help investment firms find and integrate useful data sets and aid data originators in making their offerings palatable to end-consumers. These kinds of integrators often use entity resolution systems to help map alternative data sets with more traditional ones.

Moreover, established data providers are jumping on the alternative data bandwagon by offering pre-integration with their own traditional services or in some cases supporting the use of cross-data identifiers to connect and integrate different, often-unstructured data sets.

Elsewhere, established data providers like Dun & Bradstreet and Dow Jones are packaging their non-traditional data sets into consumable data products that can be more easily integrated into investment firms' internal systems. Non-data companies that find themselves overseeing the sale of newly valuable data sets – firms like office supplies company Pitney Bowes, whose sales information can act as a leading indicator of business activity – are taking a more proactive and data-centric approach to packaging their information for sale.

Finally, the unbundling of research from executions has led to the productization of independent research services, with aggregators and redistributors like Visible Alpha, Nucleus 195 and others offering integration of unstructured research reports with more traditional fundamental and market information.

For investment managers, the Holy Grail is to be able to receive non-traditional data services in a standardised format via an industry-accepted data platform or API, making it available through the same mechanisms used to consume market prices, earnings estimates and ratings information. To achieve this, data teams at financial institutions are making use of analytics languages like R and Python, as well as tools like Tableau, Excel and Mathworks' MatLab to create a professional consumption experience and simplify the life of the analyst or other user that has to deal with the data. The goal is data that looks and feels like data consumed via traditional means, and that is documented and supported by the supplier.

Ensuring data quality

Assuming the firm has been able to find the data, validate its utility to the business, and integrate it with internal systems to make it consumable, there remains the issue of ongoing data quality.

There are no short cuts here. As with any data, quality is fundamental. Assuring the data set in question is of sufficient quality in terms of accuracy, completeness and timeliness to be fit for purpose requires a form of due diligence. It's important to ensure that the owner of the data is capable of the continued harvesting of that data from the originating process, and that it has invested sufficient resource on the physical collection and delivery of that data. It's also imperative to check the data hygiene of the originator, by checking data validation processes, reviewing time-stamp mechanisms, and eradicating selection bias where applicable.

Dealing with inexperienced data sources

Setting aside the issues directly affecting data quality, embracing alternative data necessarily involves dealing with inexperienced data providers. This means investment firms need to ensure that reliable pipelines are in place to deliver the data, particularly if there are plans to build a trading strategy based on its availability. In short, the data needs to be there when it's needed. Similarly, the data needs to be documented for both use and administration. There needs to be some form of customer support to step in when things go wrong.

Given the inexperience of many alternative suppliers, investment firms may need to build their own teams and infrastructure to deal with the gaps in experience or resource among its data providers. Or else turn to one of the middle men emerging who can help ensure reliability and availability of service.

As one practitioner puts it: "Data sources are inexperienced – often people are given the task of monetising data sets without having a clue about how to do this, or indeed about the value of the data or how it might be used. The middle men help a lot. There are handbooks and kits available that help on the mechanics of commercialising databases. So the point is that people are becoming rapidly educated in this, and help is available."

Sell-side and other third-party integrators

Helping to deal with inexperienced data suppliers is becoming something of a career path for traditional market data professionals looking for a new challenge. For these executives, there may be a market opportunity in helping firms to package up their data and administer it. But the caveat is that you need to kiss a lot of frogs before you find a prince, in terms of a valuable data set. It's the proverbial needle in a haystack.

Given the lack of resource devoted to external data procurement and associated quality checks within the investment management community, some sell-side firms and other integrators (including those discussed above) are stepping in to help connect alternative data buyers and sellers. For these intermediaries, helping originators package their data into service offerings is a revenue source, or else a new way for sell-side firms to cement the relationship with buy-side clients. This is why the digital airwaves are full of messages from consultants exhorting potential data originators to get in touch in order to monetise their data sets.

The aggregators, sell-side data teams and other alternative data providers have a kind of 'frenemy' relationship with their investment management clients. Many of those customers are seeking to go direct to the data sources in order to gain an information edge over their competitors. While third-party integrators of all stripes (including traditional market data vendors) can help them identify, take delivery and consume new and valuable data sets, the value of that data may be diluted if it's available to all. Those seeking to seal distribution deals with originators are often keen to lock in exclusive arrangements, whether it's for consumption or for broad distribution to the marketplace.

Reliable data is hard work

In summary, some investment firms are equipped to source and integrate these data sets themselves, while others want to buy product solutions. Whichever route they decide to take, the reality is that firms are better equipped than they were just a couple of years ago, as new GUI tools, reporting systems and assessment groups emerge in the marketplace.

Getting data in shape for consumption takes hard work. Data sets may need an additional source to make it useful, other data may need to be anonymised before it can be consumed. Some data is unverifiable.

But data suppliers and consumers and the middlemen that serve them are going to great lengths to ensure the new alternative data sets are reliable and accurate. There is an issue of trust and suppliers increasingly understand the need to stand behind the data they provide.

DOW JONES

dowjones.com/newswires

ABOUT DOW JONES NEWSWIRES

Dow Jones Newswires has provided unrivaled, real-time news for decades, providing financial professionals and their clients actionable information, rich data and expert insights for making market decisions, informing investment strategy and building customer relationships.

A pioneer in news as an alternative data source, today Dow Jones fuels quantitative strategies with ultra-fast, machine-readable news feeds, corporate and economic data, and a deep archives for testing. Drawing on trusted, renowned sources—Newswires, The Wall Street Journal, Barron's and MarketWatch—Dow Jones offers market-moving news, global multi-asset coverage, exclusive analysis and a richly tagged archive in the formats and at the high speed needed to develop profitable trading models, optimize portfolio allocations and manage risk. Dow Jones' data-as-a-service platform, DNA, is designed to support advanced analytics initiatives.

Learn more at **dowjones.com/newswires**.

ABOUT A-TEAM GROUP

A-Team Group helps financial technology vendors and consultants – large and small – to grow their businesses with content marketing. We leverage our deep industry knowledge, ability to generate high quality media across digital, print and live platforms, and our industry-leading database of contacts to deliver results for our clients. For more information visit **www.a-teamgroup.com**

A-Team Group's content platform is A-Team Insight, encompassing our RegTech Insight, Data Management Insight and TradingTech Insight channels.

A-Team Insight

A-Team Insight is your single destination for in-depth knowledge and resources across all aspects of regulation, enterprise data management and trading technology in financial markets. It brings together our expertise across our well-established brands, it includes:

RegTech Insight

RegTech Insight focuses on how data, technology and processes at financial institutions are impacted by regulations. **www.regtechinsight.com**

DataManagement Insight

Data Management Insight delivers insight into how financial institutions are working to best manage data quality across the enterprise. **www.datamanagementinsight.com**

TradingTech Insight

TradingTech Insight keeps you up to speed with the dynamic world of front office trading technology and market data. **www.tradingtechinsight.com**

You can tailor your experience by filtering our content based on the topics you are specifically interested in, across our range of blogs with expert opinions from our editors, in-depth white papers, supplements and handbooks, and interactive webinars, and you can join us in person at our range of A-Team Summits and briefings. Visit **www.a-teaminsight.com**

Become an A-Team Insight member – it's free!
Visit: **www.a-teaminsight.com/membership**.