Challenges of Implementing Alternative Data in Investment Management

Hurdles to Consider When Building your Data Strategy

By: Qaisar Hasan & Stan Altshuller

Maiden Century

Introduction

Hedge fund managers have accepted alternative data as a necessary part of their investment process. One recent survey found that 65% of HFs use alternative data, and the average data spend has increased considerably over the recent years. And yet, many managers still struggle to get the full value out of their alternative data spend.

Why then are managers who are so good at squeezing the highest possible ROI from every investment in their portfolio, so hamstrung to get value from their data? The quick answer is that managing data is not easy. In this article we will unpack this and discuss the top five challenges to extracting value out of data for money managers and how to overcome them.

The Data Discovery process is broken

Managers know exactly where to go to get market data and financial data on their securities. It's firmly ingrained in their process over years of experience in money management. But Alternative Data is relatively new, and it is not obvious where to source data that can give them a definitive edge over the market. Compounding the issue is the fact that different datasets can have very different applicability and relevance to each investment in their portfolio. How does a PM or analyst go about discovering a new data source for their watchlist or even know what type of data can inform them on their most important KPIs?

Currently, to get an idea of their "data coverage" or what percentage of their universe has applicable data, the analyst needs to reach out to multiple data vendors and request their coverage lists. They also need to make their best initial guess as to what type of data will be relevant to their portfolio. Does credit card data pertain to Overstock's (OSTK) total sales or is it better to look at Clickstream data and gauge website traffic? Or maybe the number of app downloads will yield the strongest model for Overstock?
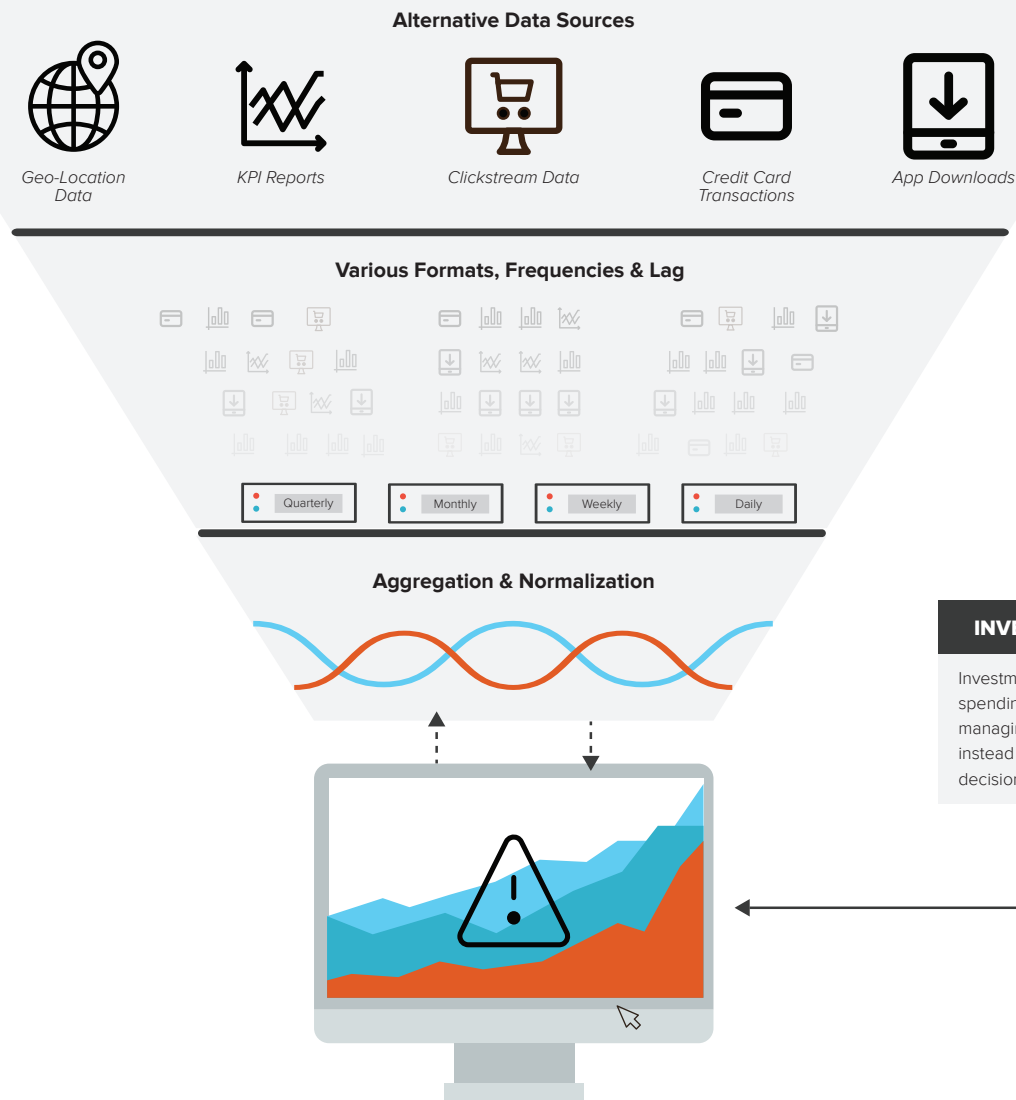
The vendors, in turn, may provide a list of their covered names but they often don't tell you how good the fit of their model is for any individual name or KPI. In the cases where vendors do provide backtests, they are viewed very skeptically by fund managers, since each vendor follows their own methodology and may select the one specific approach or variation out of dozens (sometimes hundreds) tested that happen to yield the best results. The vendors' selection criteria remain a black box, fueling further skepticism from managers.

One solution for fund managers is to run backtests themselves. But that approach also poses considerable challenges. Even for the largest and most well-resourced investment shops, it can easily take two to three months to run the backtest process rigorously, from start to finish. If you have five data vendors to choose from, that translates to a whole year invested in discovery and evaluation before you can figure out which vendors' data best suits your portfolio. The investment opportunity that triggered the data hunt has probably long withered away by then. Fund managers need answers they can trust, and quickly.

To overcome this challenge some managers choose to partner with a solution provider that has already integrated multiple data sets and gone through the process of data validation, normalization, and back testing to a large number of securities. This way the discovery process becomes more streamlined and efficient - simply go through your watchlist and select the best data source for each security.

Alternative Data Experience

Alternative Data Sources

Geo-Location Data

KPI Reports

Clickstream Data

Credit Card Transactions

App Downloads

Various Formats, Frequencies & Lag

Quarterly   Monthly   Weekly   Daily

Aggregation & Normalization

INVESTMENT TEAM

Investment teams are spending thousands of hours managing a data process instead of making investment decisions.

Disparate data sources compound the problem of aggregation

The average Hedge Fund manager has not one but multiple sources of alternative data. They have good reason to do so. Data shows that by increasing the number of data sources ingested, you increase the accuracy of your models. In addition, we will show later that having different types of data serves another very specific purpose – catching issues before they cause real damage.

Nevertheless, it is a serious challenge to aggregate disparate data sources in one central location, especially if the manger in question is resource constrained.

Imagine an analyst who should be doing due diligence on stocks, spending hours of their day logging into each separate data provider's platform to download spreadsheets, import them into their internal models, update everything, and then run this analysis for multiple stocks.

Another important consideration is that each vendor has different delivery mechanisms, or ways of getting the data to their clients. Consider just some of the delivery mechanisms that managers must work with. Some vendors make the data available through Snowflake, while others work with Amazon's S3 buckets, yet others use their own APIs or excel files that haven't been flattened, and some even send data in PDF

documents. Each one of these delivery mechanisms requires a separate on-boarding and integration process and involves significant time investment and technical skill.

Add to this the fact that different data types, for instance clickstream, app downloads, and e-mail receipts, all have different frequencies, lags, formats, and mapping taxonomies making their aggregation even more of a challenge. Furthermore, as the complexity increases, there is a lot of room for human error to be introduced in a manual data ETL process.

The solution to this problem is automation. Some managers have invested heavily in technology and have built out data pipes (or APIs) to automatically extract and load data directly into their models, validating any anomalies on the fly. While it is a worthwhile endeavor, not every fund has the required technology chops and some prefer to work with a partner to get the same benefit, but faster and cheaper.

## Interpreting alternative data is nuanced and resource intensive

When you need to know if a stock is expensive or cheap you can go to any data provider like Bloomberg, Refinitiv, or S&P and they will all agree on the company's P/E and EV/EBITDA ratios. Not so with Alternative Data. Different datasets may not tell you a consistent story about a company like traditional data would. Even worse, two models sourced from the same underlying data may disagree. Why?

It really depends on how you look at the data. For instance, when considering spending data, some vendors will combine different panels together and advise customers to analyze their whole portfolio based on that one combination (Usually, their preferred model is based in part on the total panel and part some derivative like a consistent shopper panel). They do not (normally) differentiate models based on the companies you are studying. But for certain companies, the difference can be huge. One model can show significant growth for a company while a different model based on the same data can show deceleration.

The problem is that each company is different, and one data model does not fit all. Expecting one dataset and, furthermore, one variant of that dataset, to be the ultimate answer to all your data related questions is akin to asking your stockbroker for stock tips as well as his opinion on the best fabric softener. The broker is

Data Accuracy VS. Total # of Data Sources
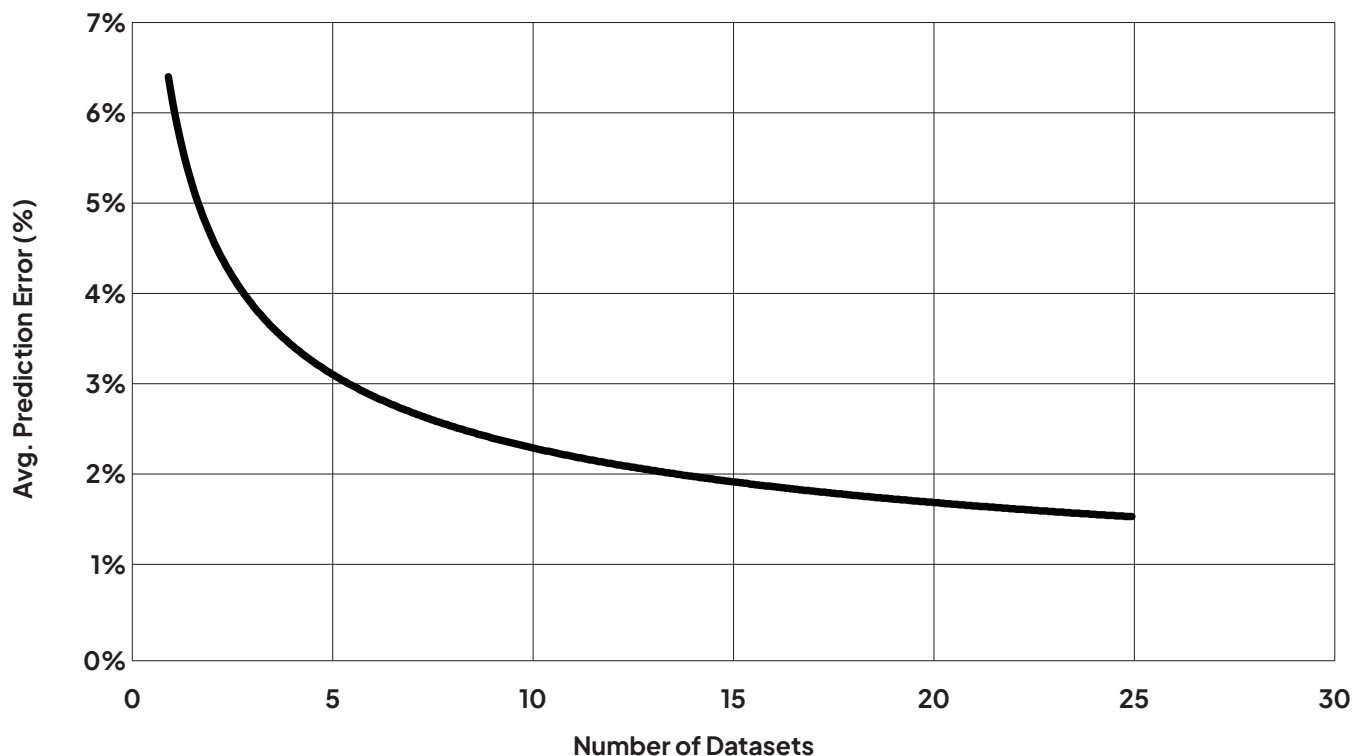
Source: Maiden Century Analysis based on thousands of predictions.

good for some things, but not everything. In the same vein, if you are analyzing a retailer catering to lower income demographics you will require a different cut of the same transaction panel than, say, when analyzing a retailer selling luxury jewelry.

Managers should scrutinize each name and KPI separately and take the time to build the best model for each company. They will also find that models perform best if they set the weights of each of the inputs to fit the specific company. While you certainly need to be mindful of false precision and overfitting, if you simply rely on one model, even the one suggested by your vendor, you will see disappointing results in certain names, and you may blame the data while it's actually the fault of the model. At a minimum, a robust model will allow multiple inputs and a custom weighting for those inputs based on the security being modeled.

Unfortunately, very often managers do blame the data when in reality they simply lacked the tools to help their data tell a better story, which brings us to our next challenge.

Bad process or poor technology can erode confidence in good data

A bad process can erode trust even in the highest quality dataset.

Several years ago, a large number of fund managers following a certain retailer, opened short positions in the stock at around the same time, a few weeks before the earnings announcement. Their data showed a precipitous decline in credit card spending for the retailer and they wagered that the company will massively miss earnings. When the retailer smashed earnings expectations, a miniature short squeeze ensued and many managers lost money, disappointed in the data that inspired them to make the trade.

It turns out a small change in the way the retailer was classified on credit card transaction records caused the tagging process to miss a huge portion of transactions. Managers falsely interpreted this as an abrupt change in customer behavior when it was really a tagging issue. Tagging deals with reading in a text or numerical descriptor and assigning it to a corporate entity or brand. Mapping refers to taking that tagged merchant,

mapping it to a company and adjusting for corporate actions. While the mapping logic worked fine, the tagging process let through a critical error. This mistake not only happened in managers' proprietary models but also afflicted several data vendors.

After realizing a loss, some managers chose never to trust alternative data again, and others invested in improving their tagging process. But some were able to avoid the trade all together.

The ones who avoided the trade realized something interesting. In this case, the tagging mistake could have been detected. How? By looking at other types of data.

The credit card data was tanking (due to the tagging issue) but foot traffic, Google search, and web traffic for the retailer's data didn't budge. Even though those datasets were historically "noisier", they were able to raise a red flag alerting the data savvy managers to an issue with the credit card data.

This is just one example of countless stories where a better process could have saved (or made) fortunes.

Of course, no data is perfect, and each data set has its biases and limitations, but the vast majority of critical mistakes happen due to a poor process or inadequate technology, specifically around tagging data to securities.

We can't underscore enough how important the tagging and mapping processes are. It is the underpinning of any successful Alternative Data operation. Investing in the right technologies and processes and coming back to it often, will help you make data-driven trade decisions with confidence.

Maiden Century IDEA Platform

Click on the image to schedule a free consultation of your alternative data needs.

Insight remains hidden and untapped in most cases

Even if managers move past the technology, resources, and operations hurdles, surfacing actionable insights remains a challenge for most. Simple concepts such as lining up all your data with your companies' financial reporting calendars and choosing the right KPI to model are just some of the reasons why signal is continuously lost in the noise. For most managers, it is just too time intensive to sift through all the available data on all their names.

Frustratingly, very often the signal is actually there in the data, but mangers lack the right mechanisms like alerts, screens, and monitoring tools to surface signal at just the right time.

In truth, a robust analytics engine has its rightful place sitting on top of a solid alternative data operation. Reporting, alerts, and charting tools are all necessary for managers who are too busy to comb through the data constantly looking for needles in the haystack. Insights should be surfaced and easily consumed by the analyst.

If building intuitive front ends is not a core competency for a manager, they may choose to partner with a technology company that can help them achieve that in a more cost and time-effective way.

Conclusion

How do you tackle these challenges in your alternative data operation? Contact us for a free evaluation of your data ensemble and we will provide a score for each of the above categories based on our work with hundreds of fund managers and data vendors.