Matt Turck

The New Gold Rush? Wall Street Wants your Data

A few months ago, Foursquare achieved an impressive feat by predicting, ahead of official company results, that Chipotle's Q1 2016 sales would be down nearly 30%. Because it captures geo-location data from both check-ins and visits through its apps, Foursquare was able to extrapolate foot-traffic stats that turned out to be very accurate predictors of financial performance.

That a social media company could be building a data asset of immense value to Wall Street is part of an accelerating trend known as "alternative data". As just

about everything in our lives is getting sensed and captured by technology, financial services firms have been turning their attention to startups, with the hope of mining their data to extract the type of gold nuggets that will enable them to beat the market.

Could working with Wall Street be a business model for you?

The opportunity is open to a wide range of startups. Many tech companies these days generate an interesting "data exhaust" as a by-product of their core activity. If your company offers a payment solution, you may have interesting data on what people buy. A mobile app may accumulate geo-location data on where people shop or how often they go to the movies. A connected health device may know who gets sick when and where. A commerce company may have data on trends and consumer preferences. A SaaS provider may know what corporations purchase, or how many employees they hire, in which region. And so on and so forth.

At the same time, this is a tricky topic, with a lot of misunderstandings. The hedge fund world is very different from the startup world, and a lot gets lost in translation. Rumors about hedge funds paying "millions" for data sets abound, which has created a distorted perception of the size of the financial opportunity. A fair number of startups I speak with do incorporate idea of selling data to Wall Street into their business plan and VC pitches, but how that would work exactly remains generally very fuzzy.

If you're one of the many startups sitting on a growing data asset and trying to figure out whether you can make money selling it to Wall Street, this post is for you: a deep dive to provide context, clarify concepts and offer some practical tips.

Raw data vs Data products

First, a key principle: the business of selling *raw data* is generally not a great one.

Instead, companies that successfully monetize a data asset tend to offer data-driven *products*. To use an obvious example, Facebook doesn't sell its user data in raw form. Instead, it has built an infinitely more lucrative business around data-driven advertising products that enable brands to target Facebook users, based on the troves of data they provide about themselves.

Even when companies want to license the actual data itself, they tend to do it through data products, rather than in raw form, sometimes with built-in analytical capabilities. For example, Twitter evolved its original firehose business into an full-blown enterprise data platform, GNIP (originally through the acquisition of the startup of the same name), which offers various APIs such as "historical", "real time" and "insights". Mastercard offers data indexes and research products through MasterIntelligence. Foursquare provides its data through a product called Place Insights.

But as a startup entrepreneur, you may have your hands full with your core business, and may not have the luxury to start an ancillary data business. In that case, it may make sense to explore opportunities to monetize your data exhaust by providing it in raw form – something that an increasing number of Wall Street institutions (banks, hedge funds, asset managers) are interested in. In fact, some of the most sophisticated hedge funds will insist on getting raw data over anything else. As hedge funds are very much at the forefront of this trend, the discussion will center primarily on them.

Why do hedge funds care?

First, a little bit of context.

Hedge funds are investment funds that are obsessed about one thing: completely outperforming the broad market, and delivering outsize returns to their investors. They use complex portfolio construction and risk management techniques, and can invest in all sorts of different markets (real estate, stocks, derivatives, currencies, etc,). They're occasionally secretive entities, not very regulated (despite the huge size of the industry – $2.9 trillion), and closed to

the broad public. A good part of the hedge fund mystique comes from the fact that they they have generated extreme wealth not just for their investors, but also for their managers: the top 25 hedge fund managers earned a whopping $13bn in 2015.

But things have been changing in the hedge fund world. The industry was long dominated by "Master of the Universe" personalities, famous for making prescient bets based on contrarian views of the market (think The Big Short). As in many other parts of the economy, however, the computer is gradually taking over and today, Big Data and AI are playing an increasingly crucial role.

This has certainly been true at the major "quant" funds (Renaissance, Winton, D.E. Shaw, AQR, Two Sigma, WorldQuant, etc.) , which have been using mathematical models or algorithms to evaluate investments for a long time, and have been more recently building up significant AI capabilities.

But this is increasingly also true at many of the "fundamental" hedge funds – those funds that traditionally made investment decisions based on analyzing individual stocks or the market as a whole.

This trend has been precipitated by the fact that the hedge fund industry (in general) has been having a tough time lately: low performance led many investors to pull out of a lot of funds.

Now, big industry names such as Paul Tudor Jones cut some of their fundamental traders, and are instead switching to a quant strategy, where traders and computer are expected to work in tandem: "No man is better than a machine, and no machine is better than a man with a machine." (Paul Tudor Jones to his investment team, August 2016).

The term du jour for this new approach is "quantimental", a hybrid between quant and fundamental. Convergence has not been always smooth and the jury is out as to its eventual success.

Not everyone in the investment world is a big fan of the trend, but there is enough excitement that hedge funds could now be suing each other over top data science talent: just a few weeks ago, WorldQuant sued Third Point over a 32 year old data scientist, Matthew Ober, who got a 10x salary increase (from $200k to $2M) in the process (the lawsuit was settled, details of the story here).

Things are only accelerating with the emergence of new quantitative hedge funds driven by distributed communities of data scientists and/or quants around the world, such as Quantopian and Numerai.

An arms race has started. As quant talent on Wall Street is a lot more available than it used to be, and models tend to eventually be leaked out as talent switches jobs, data is at the core of this new rush, particularly non-obvious, hard to get, alternative data

What do hedge funds do with the data?

Fundamentally, hedge funds try to leverage alternative data to gain an edge over their competitors and generate "alpha", through accurate predictions. Ultimately, they want knowledge of something that few others know. That way, they can pre-position so that when others do find out, they are holding the exact security that goes up on the news (or executed the opposite trade, in the case of a short).

For anyone outside of the trading world, it's worth emphasizing that on Wall Street, it's not just enough to come up with *strong* predictions. Everyone else on trading floors has their own predictions, using all sorts of methods, so, to make money, you need to come up with *better predictions than anyone else*. The bar is particularly high.

Wall Street has been in the prediction game since its origins, and the idea of obtaining data not available to anyone else is not new. It used to be stock prices and fundamental information. As those became widely available, hedge funds moved on to other forms of data.

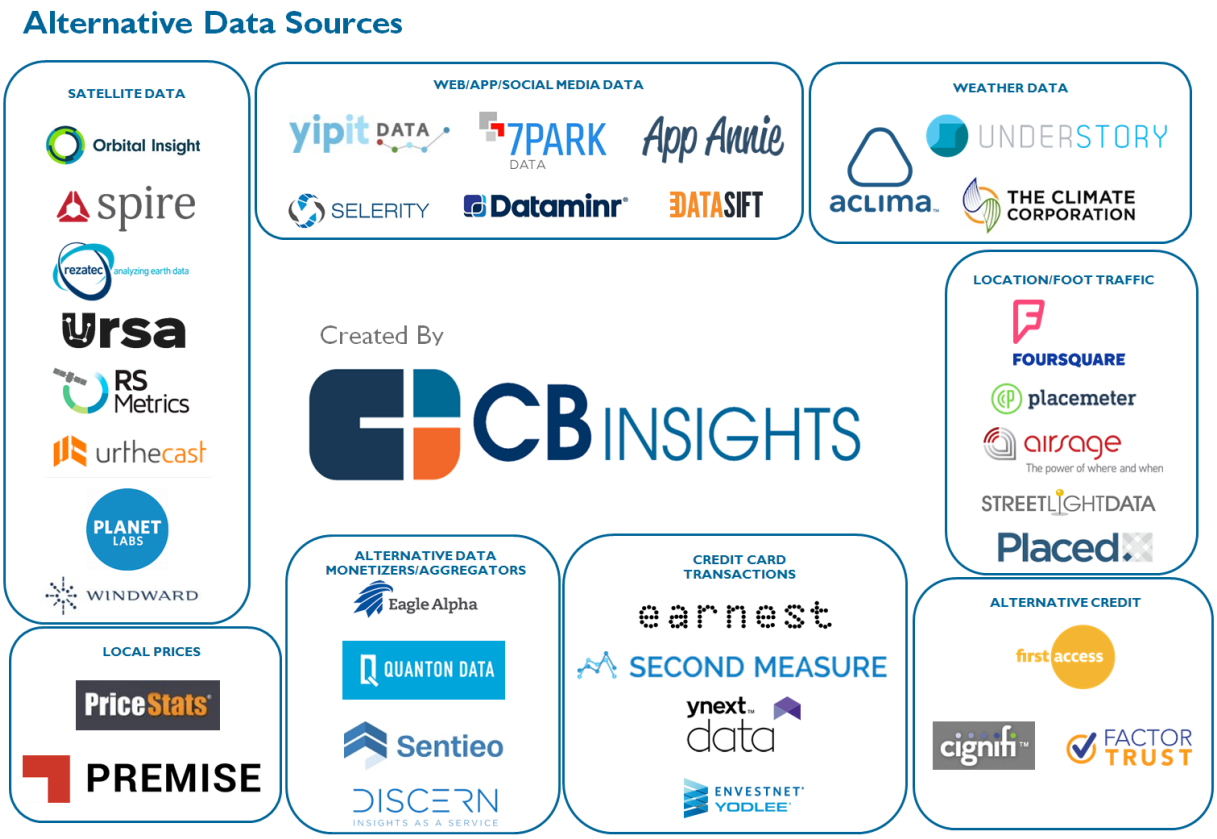
Not that many years ago, some hedge funds would send people to literally stand in front of big-box retail stores and count the number of people coming in and out, and on that basis make predictions about the retail chains themselves and the economy in general.

Alternative data now offers an opportunity to do the same thing at an entirely different scale and level of sophistication.

The trend started a few years ago with social media data. Could one not only access market moving news faster than the regular press, but also gain non-obvious insights by crunching through all tweets relating to a certain topic? Those where the days when some of the larger hedge funds and banks would start licensing the Twitter firehose.

Now hedge funds have broadened their interest to all sorts of other datasets: geo-location, credit card payments, satellite images, IoT sensor data, building permits, health data, etc. Some of this data comes from companies that are just trying to monetize their data exhaust; other data sets are coming from companies whose primary business model is to offer this data (often in the form of data products, as per the above).

A whole cottage industry has now appeared, with some key players nicely highlighted by CB Insights in this landscape:



What hedge funds do with the data depends a little bit on where they fall on the spectrum described in the preceding section.

The more fundamental funds will use the data as an input into human-driven investment decisions. For example, they'll try to predict the sales or churn of a specific company, with the overall gall of of outperforming sell side consensus. Or they'll try to predict macro economic trends, for example through the observation of satellite images. They will also often use models, but what the quant (data scientist) predicts will be generally just one data point that the "PMs" (portfolio managers) will decide to use or ignore in their investment decisions, alongside other inputs (such as what their carefully cultivated professional network thinks).

At the other end of the spectrum, the quantitative funds will take your data set, combine it with other alternative data sets and feed them into very sophisticated models. The growing trend is to completely or partially automate trading strategies on the basis of those models, fed by alternative data.

How interesting is your data?

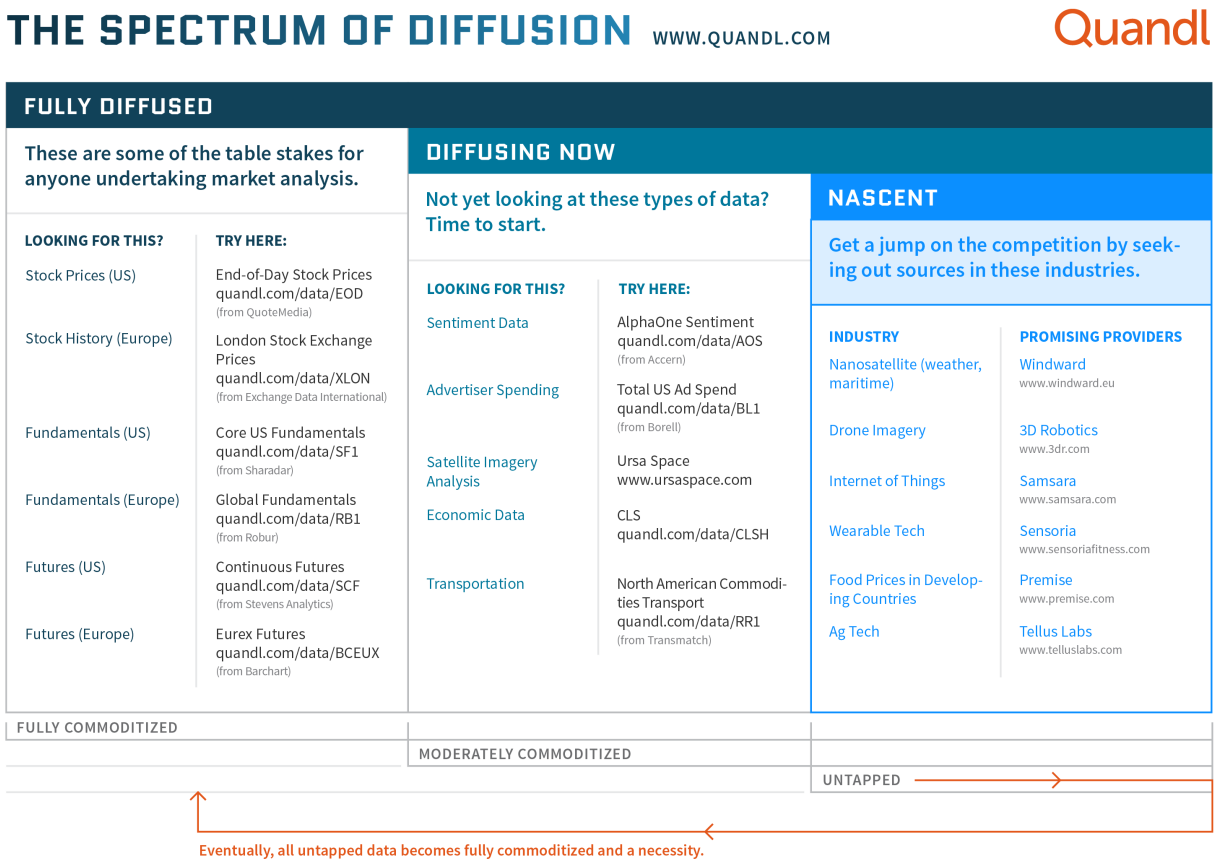
There are a few key characteristics that make your data more or less interesting to hedge funds: level of detail, history, breadth and rarity.

The level of detail and specificity of your data matters a lot. For example, user-level (anonymized) credit card statements with detailed purchases are many times more interesting than high indices and totals, particularly if you get them frequently.

Another key criteria is history: how far back in your time does your data set go? This is often an issue for startups, which by definition haven't been around for very long. In an ideal world, hedge funds would want to see 5 to 10 years of history. Having said that, depending on the specific use case, some may be able to work with one or two years of history, especially if your data is more scarce and interesting. Startups would be well advised to store and retain all their data from the very beginning (which is viable considering the near-zero cost of storage).

Breadth of coverage also matters – not just geographical coverage (making sure you data set covers is representative, say, of the US population), but also coverage in terms of how many stocks your data could possibly cover. While some fundamental analysts will only care about data for the few stocks they cover, quants may want data that can be relevant to hundreds or thousands of stock.

Finally, considering hedge funds are trying to gain insights that their competitors won't, the more unique and original your data set is, the better. An interesting consequence is that the value of your data set is likely to *decay* over time. While they may obtain their data exhaust from a completely different use case, other companies will eventually be able to provide a data set that is comparable enough to yours, and over time most data sources will get commoditized. This phenomenon is well captured in the following chart from Quandl:



In general, very few data sets are going to be the be-all and end-all for any given investment decision, however rare and comprehensive they are. In most cases, hedge funds will want to combine a number of different data sets. To understand QSR sales for example, prudent data scientists will want to combine foot-traffic trends (as provided by a Foursquare or our portfolio company Sense360) with data on credit card transactions to understand if customers also had more meals delivered (colder weather, availability of services like Postmates, etc.).

Are you actually allowed to sell the data?

At a high level, three key concepts (the details are beyond the scope of this post):

Data should be strictly anonymized. You cannot, or shouldn't, sell "Personally identifiable information" (PII), meaning any data that could potentially identify a specific individual. The good news is that hedge funds are not

advertisers and don't care about particular individuals, so there is no economic pressure to provide PII. This seems obvious enough, except that hedge funds report that data sellers routinely fail to hide personal information, resulting in necessary effort and time to clean up (see this Financial times piece on the topic).

You cannot sell data you don't own. Hedge funds care immensely about the legality of the data. How you obtain data from your users is regulated by your Terms of Service (TOS), and they need to enable you to sell it. You should get your TOS to cover this from the beginning, otherwise it's a headache to have to figure out which data was obtained under which TOS (and take out the data you're not allowed to sell). There are some gray areas around the definition of what "consent" and "opting in" mean.

Finally, you should be aware of the all-important concept of Material Non-Public Information (MNPI), which is an insider trading concept – basically, any non-public information regarding a specific company that would be included in your dataset that could provide the hedge fund an advantage when buying or selling the company's stock. Particularly if your data set includes some third party data that you combine with yours, expect compliance departments of hedge funds to do some serious digging.

How much money can you make?

Now, the big question: how significant a business can this be for you?

First of all, unless it's already commoditized, you probably don't want to sell your data to the Bloombergs of the world. They'll pay small amounts of money (low tens of thousands per year), and the data will immediately become available to everyone on Wall Street, since everyone uses the Bloomberg terminal. There are nuances to this (a separate Bloomberg team used to resell data to hedge funds on a more discrete basis, but it may not still exist)

As to how much money you can make selling data directly to hedge funds, this is where things get tricky. It's a very opaque industry, so it's generally hard to

know. Hedge funds will not share with you what exactly they're going to do with your data, so it's hard to ascribe a value to it. It's also hard to build a repeatable use case that you can then go sell to the next hedge fund.

You do hear the occasional story where a hedge fund paid a couple of millions a year to obtain a specific data set, and occasionally more. But there's a reasonable chance that this type of price came with some kind of exclusive. Also, those contracts also probably have a limited shelf life, as the value of a data set decays over time, as per the above.

For the most part, the reality is that most hedge funds are cost-sensitive, and will be negotiating fees down quite aggressively. From what I hear, for the most part average annual fees will range from somewhere in the high tens of thousands of dollars to somewhere in the low hundreds of thousands of dollars, with perhaps $100,000 per hedge fund being a fair median number.

To get a sense of overall market size, there are probably 10,000-15,000 hedge funds. As mentioned, banks and asset managers can also be customers for your data asset.

How do you get started?

The hedge fund world can be tricky to navigate. There are significant variations among funds – in terms of strategies, as described above, but also in terms of overall sophistication and readiness to leverage alternative data. Point72 has a whole team, led by Matthew Granade, focusing on Big Data and AI. Two Sigma has hundreds of PhDs with machine learning backgrounds. Many other firms are on the other end of the spectrum.

Also there is a wide cultural gap between the tech world and the hedge fund world. For starters, the vast majority of hedge funds, and the financial services industry in general, is going to be in New York, and in places like Stamford or Old Greenwich, CT – far from the Silicon Valley. The financial services world has its own strong identity, handles massive amounts of money, and will not

necessarily be in awe of your cool startup. When I was at Bloomberg, I used to cringe when startup founders would show up in hoodies, essentially destroying their credibility before the meeting even started.

For all those reasons, it probably makes sense, at least initially, to work with an intermediary of some sort.

There are new firms, such as Matei Zatreanu's System2, that advise hedge funds on incorporating alternative data. And they often help startups create data products that hedge funds will find valuable.

Companies like GuidePoint have been accelerating their efforts, in particular through the acquisition of Quanton Data (see here).

A few startups have smartly positioning themselves at the intersection of this growing trend. EagleAlpha, for example, is a name that comes back frequently in conversations.

Last but certainly not least, Quandl is a fast-growing marketplace for both regular data sets and alternative data, which takes a lot of the headache involved on both sides around access and legality. They're increasingly recognized as a key thought leader in the space (and they also kindly invited me to speak at their Alternative Data Conference tonight!).

Conclusion

The interest in alternative data is surging, and it's a good time for startups to explore whether they can leverage this trend.

Perhaps counter-intuitively considering how much money hedge funds handle, selling your raw data exhaust to them will probably be "just" an ancillary revenue line, at best.

However, If done right, it can be a first step towards building a proper data business (based on data products, rather than raw data). Some hedge funds will

help you understand how to package and organize your data so it can be of interest to a much broader cross-section of industries, so you can explore selling the data not just to Wall Street but also, for example, to retail or pharma, or whichever vertical your data set offers particular relevance.

Many thanks to Matei Zatreanu (founder of Augvest and System2, former Head of Data Science at King Street) and Tammer Kammel (CEO of Quandl) for reviewing a draft of this blog post and providing helpful feedback.

POPULAR POSTS

Growing Pains: The 2018 Internet of Things Landscape
The New Gold Rush? Wall Street Wants your Data
The Power of Data Network Effects
AI & Blockchain: An Introduction
Building an AI Startup: Realities & Tactics

CATEGORIES

Select Category

SEARCH THE ARCHIVES

Search …

Matt Turck / Proudly powered by WordPress