PROFIT&LOSS

ISSN: 1467-2650 © 2017 P&L Services Ltd. All Rights Reserved. This article first appeared in: Profit & Loss Magazine, Q4 2017 | issue 172 | vol. 18

Data Drilling

Drilling for the New Oil

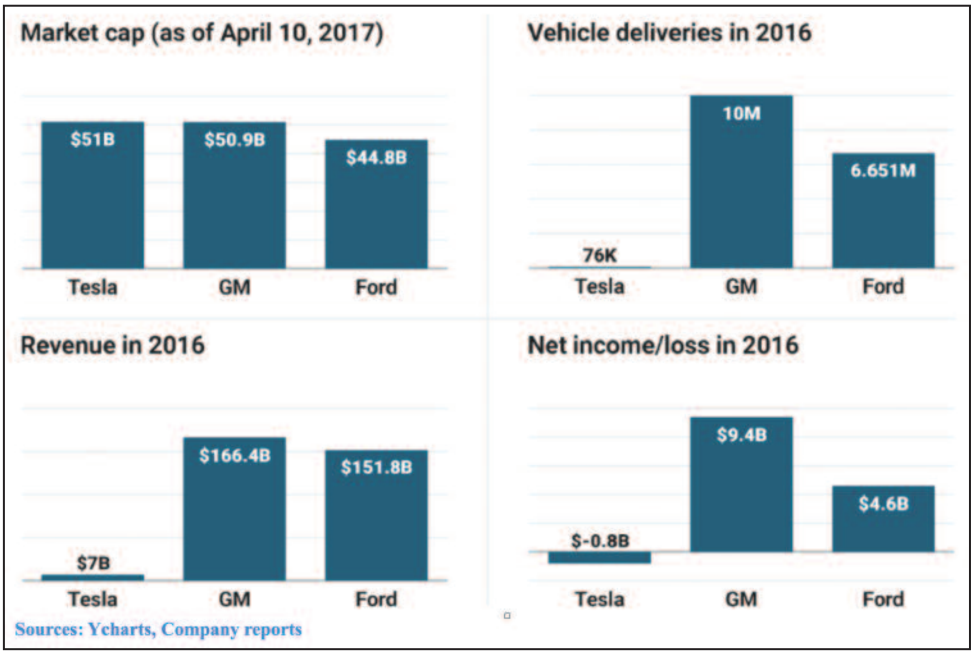
If data is the new oil, then how trading firms "drill" it in order to generate alpha becomes increasingly important. Galen Stops reports.

So often has the phrase been used recently that it's in danger of becoming something of a cliché but, apparently, data is the new oil.

To see evidence of this, look no further than the technology giants that have emerged out of Silicon Valley. Yes, Facebook doesn't charge users money for the social media platform it provides, but is it free? Arguably, users "pay" with the data that they create via their interactions on the platform, which Facebook is then free to use and sell to generate profits.

Uber, with losses totaling $2.8 billion in 2016, is thought to be valued at around $68 billion. Part of the reason for this massive valuation is because it owns a massive amount of data about supply and demand for personal transportation.

Tesla's market capitalisation overtook GE for the first time earlier this year, despite the fact that the vehicle production, revenue and overall net income of the former is but a tiny fraction of the latter's. Why does this disparity exist? Because Tesla's cars collect vast amounts of data that can be used for numerous different purposes, such as improving and optimising the algorithms that underpin the self-driving cars that it is developing.



Sources: Ycharts, Company reports

Not only is data valuable, but there is more of it available than ever before. In a whitepaper published earlier this year, the advisory services firm IDC forecasted that by 2025 the global datasphere will grow to 163 zettabytes (ZB) – the equivalent to a trillion gigabytes, or 10 times the 16.1ZB of data generated in 2016.

Sadly, the increasing recognition of the value of data combined with the expected surge in growth of data in existence does not mean

that everyone is going to get rich. That's because the data only has value once it is analysed and structured in such a way that it can provide insight that in turn can provide value or, in the context of the financial markets, alpha.

And when it comes to analysing very large amounts of data very quickly, machines, unsurprisingly, have the edge over humans, hence the current excitement over the practical applications of artificial intelligence (AI) and machine learning in financial markets. To be clear: AI is a broad concept of machines being about to carry out tasks that humans would consider "smart", while machine learning is an application of AI that refers to machines being able to learn from a data set provided to them.

Indeed, even machine learning is actually a fairly broad term – for example, deep learning (short for deep neural network learning) is a type of machine learning that tries to structure the data into something that can be more easily understood by humans. If Janet Yellen, the current chair of the US Federal Reserve, makes a speech, a deep learning network wouldn't try to predict what is going to happen in the markets as a result, but it would be able to summarise that speech in a way that would be understandable to a human.

In the same report, IDC estimates that the amount of the global datasphere subject to data analysis will grow by a factor of 50 to 5.2ZB in 2025, while the amount of analysed data that is "touched" by cognitive, or AI, systems will grow by a factor of 100 to 1.4ZB in 2025.

No Self-Driving Algos Yet

Although AI and machine learning are certainly buzzwords in the financial services industry right now, these tools have actually been around for some time. Charles Ellis, a trader and quantitative strategist at Mediolanum Asset Management, points out that simple linear regression is a form of supervised machine learning and, at a very basic level, this can be done in an Excel spreadsheet.

"On the asset manager side, there is a lot of buzz built up around AI and machine learning, but I think that there's a big misconception about where it is and how it is being used. People look at self-driving cars and a machine beating the best Go player on the planet and think that this is where Wall Street is headed, but that isn't the case. Not many are really doing the equivalent of a self-driving car," says Ellis.

He continues: "That's not to say that at a less sexy, more functional level, machine learning – and especially supervised machine learning – isn't being used. Supervised machine learning is just a mathematical model where you give a computer data and it comes up with, statistically speaking, what it thinks is going to happen in the future. When it comes to alpha generation, this is really just the next step beyond the linear regression models we've been dealing with for the past 20 years or more."

Although AI and machine learning tools have improved in sophistication over time, what is really driving the buzz around them now is that the amount of data that financial services firms are able to access has increased, while the cost of the computing power required to store and analyse the data has reduced significantly. Part of the challenge then becomes finding the right data to use.

Bob Savage, CEO of CCTrack, broadly visualises data types in a grid with four quadrants, divided on one axis by

"There's probably no asset managers out there who are spending less on data than they were five or 10 years ago."

structured versus unstructured data and the other by linear versus non-linear data.

Traditionally, says Savage, the big quant trading firms used structured and linear data – price and time – to generate their alpha. This means that this data has been picked over heavily and that anyone coming to market with a new quant strategy based on price/time data is competing with these firms that have a 35-plus year history in the space.

In the next quadrant is unstructured linear data, such as tick data from FX or equities, volume data across time or data from other markets as it applies to the underlying market being traded.

"The amount of people using this data has really shot up over the past 10 years and it creates a Big Data problem in the sense that it fills servers very quickly, but again, the bigger more established firms have an edge, because they've been collecting this data for longer and have more experience analysing it. However, some funds – such as ourselves – feel like we know how to parse tick data in a better way than a non-FX expert or know how to think about volume in a different way that improves our model," says Savage.

In the next quadrant is structured but non-linear data, which would include things such as meta-data, Google searches and Twitter feeds. This structured data could be looked at across time, but it doesn't have to be in order to provide meaning and value, because in many cases this data helps analyse themes.

In the final quadrant is unstructured linear data, which can take many forms. An example of this type of data could be satellite images that could provide more detailed weather and temperature information, which could have implications for certain commodities markets. This last type of data is more at the bleeding edge when it comes to alpha generation because it is the hardest to analyse effectively and cost efficiently.

"There is some need for this data – obviously when you're trading commodities and you have access to global weather information, it's going to give you a better view on droughts and potential squeezes for particular crops – but the arbitrage that's there is because the markets are inefficient about getting that information, will it sustain the cost of accessing all this data and sifting through it? There is an equation that is a limitation to the cost of parsing through this data, which is the money that you have available versus the money that you might make from doing this," says Savage.

He adds: "That's why analytics in equities and FX with their tick data, which improves execution first and approximated volume second, is more developed than analytics using this amorphous unstructured data. That doesn't mean that there aren't specific instances where there is money to be made, it's just that parsing this data requires a lot of work and more sophisticated artificial intelligence tools, which is another layer of cost."

A new breed of quants

In terms of how artificial intelligence and machine learning is being applied in new ways in FX, Gaurav Chakravorty, a former Tower Research partner that has now founded his own algorithmic trading firm, qplum, sees a new breed of trading firm emerging. Although these firms use similar quantitative tools to the traditional HFT firms, Chakravorty says that they use machine learning to create higher accuracy trading models.

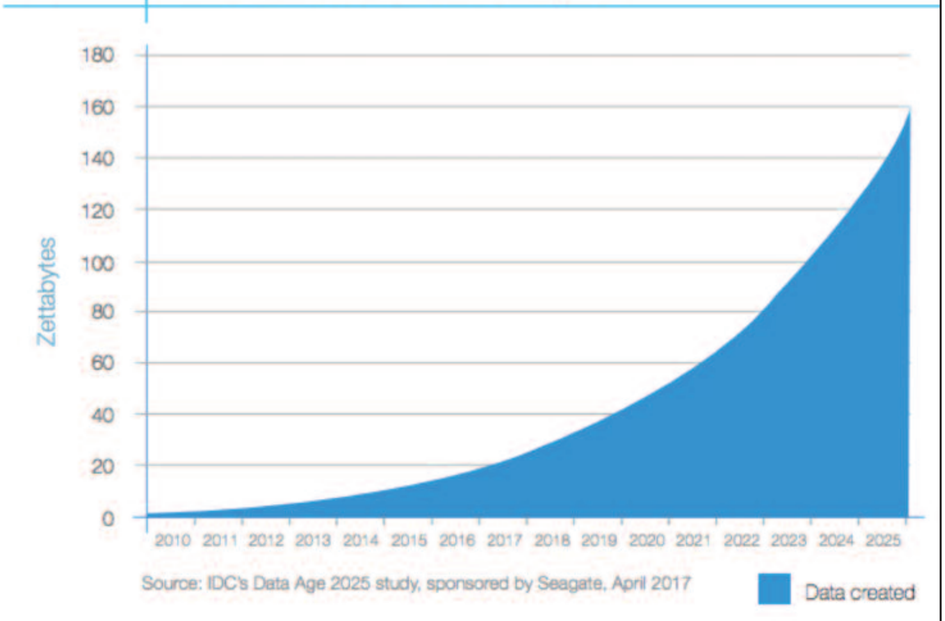
"The traditional HFT firms have very low accuracy models, but they're very fast, that's what works for them. But that's also why they only want to get filled, say, once every 200 times they place an order. That's what happens when you're analysing simple data sets like the bid/ask spread, you're not going to get a very high accuracy model," he explains.

By contrast, Chakravorty says that some firms are using machine learning to build more and more indicators into their models, which makes these models so accurate that every time an order is placed, that trading firm wants to be filled on it. This higher accuracy model requires more capital, because the trading firm is not going in and out of trades all the time, but it yields a higher profit per trade, according to Chakravorty.

"At qplum, we don't look at unstructured data, but instead we focus on data from other asset classes. For example, if payroll data comes in, we look at how that impacts every single currency pair and feature that we're trading. Not only that, but we trade as a basket, meaning that all of our positions are in sync and we try to be hedged multi-asset, whereas a typical HFT firm would trade every currency pair as if it's a stock and so they trade it individually," he adds.

A source at one US-based trading firm agrees that widening the data sources that they use as an input for their algos has enabled them to refine their strategies, even when that data can't generate useful trading signals directly. They found that in some cases, when they put this data into machine learning models in a feedback loop it has

Figure 2. Annual Size of the Global Datasphere



Source: IDC's Data Age 2025 study, sponsored by Seagate, April 2017 made their original primary trading model much better.

"So we're looking at a lot of data to see if we can use it to make other models smarter rather than just seeing if it works by itself. As a case in point, we haven't found a way to trade social media sentiment data by itself, but we've found it helpful in other ways," the source says.

Internal Data Sources

While new ways of analysing external data sources via AI and machine learning techniques to generate alpha are likely to continue grabbing headlines, the structure of the FX market means that external trade data sources are limited in value.

Charles Ellis

Stuart Farr

Firstly, only a fraction of the roughly $5.1 trillion notional of FX executed in the market each day is broadly visible to market participants. Secondly, even if all of this trade data was available, it wouldn't necessarily have the same value that it has in a market like equities, because FX is primarily traded on an OTC bilateral basis, meaning that the prices that were traded on won't necessarily be available to any firm in the market.

The answer to this problem, according to Stuart Farr, president of Deltix, is to look internally for the data that trading firms need to build their strategies around.

"Firms should be recording all the data that they get from liquidity providers, from banks, from ECNs, anyone that they're dealing with, and even those that they're just interested in dealing with. As part of your commercial discussions, you should be getting streaming data from them as an evaluation option. Then you're working with data sets that are specific to you and therefore can conduct analysis on actionable data," he says.

Farr likens firms that aren't recording all this FX data to the gas that sometimes flares off at oil refineries, arguing that these firms are "wasting" market data which, as has already been established, is a valuable commodity.

Recording internal data sounds easy enough on paper, so why aren't many more firms doing this? Farr explains that in the past it would have required a sizable financial expenditure on hardware and software to be able to record this data, which is why some firms haven't traditionally done so. However, he points out that although advances in technology that can enable firms to do this are now within everyone's reach, he suggests that perhaps there isn't an awareness of this fact.

Of course, the other problem is for new firms that don't have internal data, but Farr argues that this is a short-term problem.

"If you record your data from day one, then yes, you do have an issue at the start, but this only exists for a few months and then diminishes over time until it becomes non-existent," he says.

When it comes to new ways of using data to generate alpha, Farr says that many of the conversations he's having with clients have moved away from using alternative data sets and towards using the data to improve existing strategies and execution.

"I think there wasn't as much potential in the alternative data sets as everyone hoped for. We all hoped for something exciting and consistent, and although using alternative data to inform strategies is not dead by any means, I don't think the potential has been realised. So, the conversation has moved on from trying to find an all-singing, all-dancing new source of data from which to generate alpha to: how do I preserve the alpha that I'm getting from more traditional strategies?

"Often that is about getting more out of the execution, getting more bang for your buck – if you have a certain amount of alpha, how do I preserve that by losing as little as possible in my execution? That's a very valid question and one that is difficult to answer, partly because the answer changes as the market changes," says Farr.

The Cost of Doing Business

But although the cost of accessing and analysing data has come down dramatically over the past 10 years, this doesn't necessarily mean that firms will be spending less money on data and data analytics.

In a recent survey conducted by *Profit & Loss*, the overwhelming majority of firms – 67% – said that they expect to spend more money on data over the next three years. Compare that to just 5% who said they expect to spend less, 12% who said they expect their spending to remain the same over this period, and 16% who said they were unsure.

"I think there's probably no asset managers out there who are spending less on data than they were five or 10 years ago," says Ellis, adding that Mediolanum's data costs have trebled in the past couple of years.

It's worth noting that the costs associated with this are not all geared towards AI and machine learning tools. The data feeds cost money, handling the sheer volume of data being produced and consumed costs money, so as the value of data comes into sharper focus, more firms are looking to leverage the value of the data they have to charge a fee and, ultimately, it still costs money to find and hire talented individuals who understand both trading, programming and AI.

It seems certain that, in the quest for alpha, data is going to be in increasingly high demand, but also supply. It then seems that the ones that will be most successful are likely to be the ones that become most proficient at, firstly, finding the relevant data within the raw mass of data being produced, and secondly, feeding this relevant data into AI and machine learning tools that will in turn provide them with information that they can use to trade. And, incidentally, it seems that this will probably require a sizable budget for data.

"There wasn't as much potential in the alternative data sets as everyone hoped for...although using alternative data to inform strategies is not dead by any means"