WHY SMALL DATA IS THE NEW BIG DATA

Marketing material for professional / institutional / accredited investors

Executive Summary

Over the last decade we have seen significant advances in machine learning across a wide range of fields. In many cases, this has come from applying very complex models – often containing tens of thousands of parameters – to extremely large datasets often containing millions of examples. These applications are often described as 'big data' problems.

However, there is a related category of problems where the amount of available data to train a machine learning model is fundamentally limited. In this article, we will refer to these as 'small data problems'. Small data problems are very common in finance and need to be approached in a very specific way since in most cases, techniques designed to solve big data problems simply do not work well when applied to small data sets.

In this paper, we will discuss some examples of small data problems in finance and outline some of the approaches that can be applied to address the challenges posed by small data.

Dr. Chris Longworth
Investment Director,
GAM Systematic

Dr. Silvia Stanescu
Investment Director,
GAM Systematic

Heightened expectations for AI

The last decade has seen a succession of headlines announcing various breakthroughs achieved by machine learning algorithms across a wide range of fields. In 2016, the artificial intelligence (AI) system AlphaGo [1] developed by DeepMind beat Lee Sedol, a 10th Dan Go champion in an exhibition match. There has also been competitive AI performance in less structured games such as DOTA 2 [2]. Most recently, DeepMind announced [3] that it had developed a system for accurately predicting how proteins would fold, effectively solving one of the Holy Grails of AI research. Across fields such as computer vision, translation and speech recognition, AI systems that were once the province of academic research have broken through into the commercial mainstream.

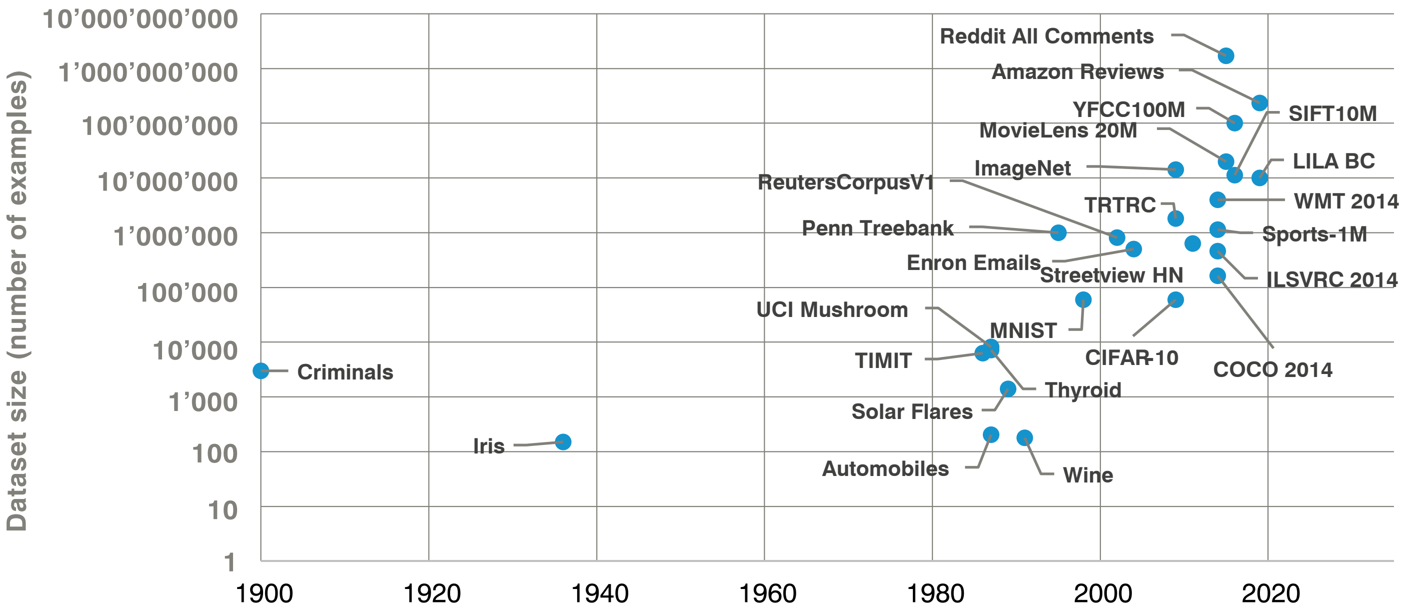
Why are we seeing all these breakthroughs now?

First, we have seen a huge increase in the amount of available computing power. The computer in your home is much more powerful at a much cheaper cost than at any time in the past. The rise of the internet has also led to the growth of cloud computing, with easy and cheap access to vast amounts of networked computing power.

A second factor has been the development of new kinds of machine learning techniques. Many of the problems we have discussed seem quite different, yet many of the underlying techniques used to solve them are very similar. A common approach is based around models known as neural networks. When designed with a large number of parameters, neural networks are able to model extremely complex phenomena – a field known as deep learning.

Finally, we have seen the development of much larger sets of data, available for training these models. If you were to go back 50 years, many of the datasets being worked with were extremely small, in part because creating them required significant manual effort to collect the data [4]. But also because processing larger datasets would simply have been beyond the capabilities of computers at the time. More recently, we have seen a steady increase in the size of datasets available, as depicted in Figure 1. This increase has partly been driven by the increased digitisation of our everyday life, with many people leaving a much larger online footprint across photos, videos and other online interactions. While we do not use such personalised data, other advances such as increased access to high-resolution satellite data have also helped to increase the data available to financial researchers. This data can also more easily be transformed into large scale collections for building and training models.

Figure 1: Increase in dataset size over time

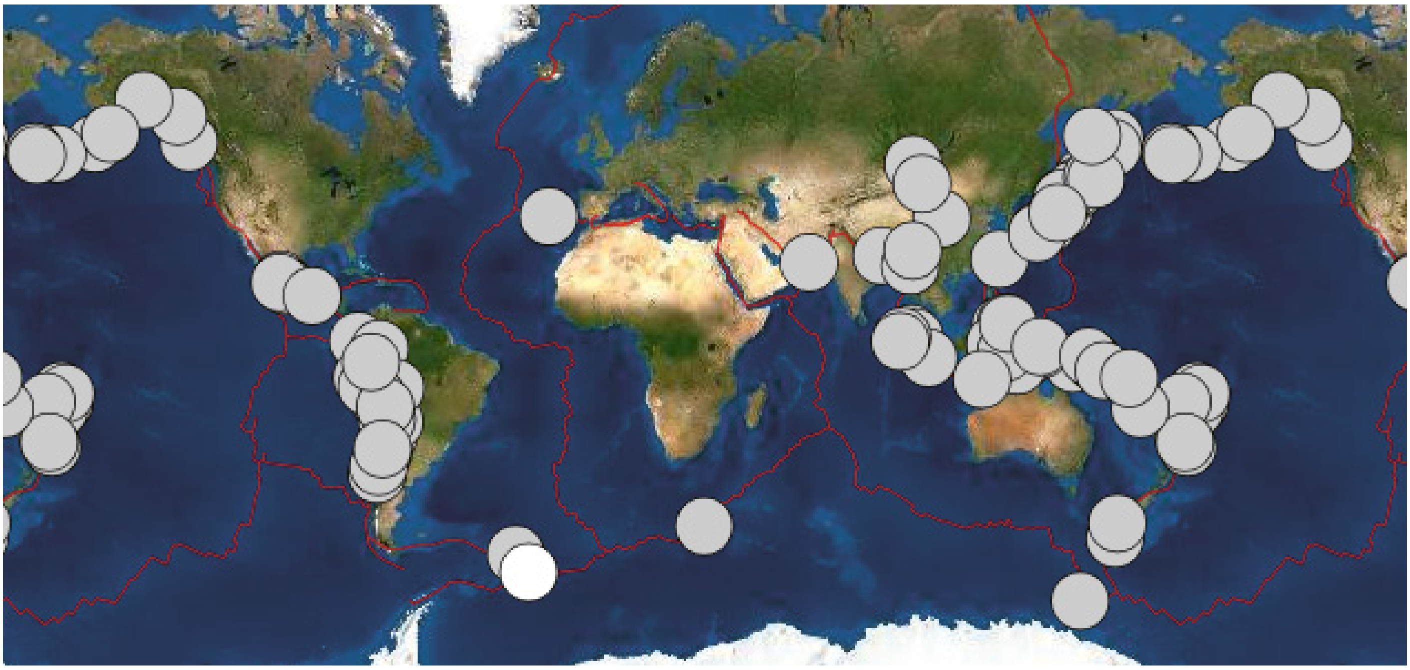


Source: GAM, based on a chart from ‘Deep Learning’ by Ian Goodfellow.

What are small data problems?

Despite the increasing availability of large datasets, there still exist problems where the amount of available data is extremely limited. One example of such a small data problem is the study of large earthquakes. High quality historical records of earthquakes start around 1900. However, since then there have been around only 100 earthquakes of magnitude 8.0 or greater worldwide, as shown in Figure 2. Importantly, the issue is not that we did not look hard enough for data. We already have the complete dataset, but it is small.

Figure 2: Locations of the largest earthquakes since 1900



Source GAM, USGS

There are some tell-tale signs that you might be working with small data:

- **Time series:** If you have data that is associated with a particular point in time on a specific date, there a high chance you are dealing with a small data problem. This is especially likely to be the case when dealing with data that is only periodically available which is common for economic data.
- **Rarity:** Does the data represent real world events, and do those events occur rarely in nature? This is the earthquake situation outlined above.
- **Aggregate:** Is your data aggregate data? If your data represents whole countries or already represents a global aggregate, you likely have a small data problem. Unless you are dealing with astronomical data, we normally only have data from one planet to work with.
- **Correlated:** If your data contains a high degree of internal structure or correlation, you are likely to be dealing with fewer independent data samples, particularly if the dataset is noisy.

It turns out that many problems in finance actually satisfy all of these criteria! Finance consists of both big data and small data problems and the challenge is to be able to differentiate one from the other. Some methods that are suitable for big data problems tend to be less suitable for small data problems. Similarly, approaches designed for small data problems often fail to scale up to work on very large data sets. So understanding the distinction between big and small data problems is key to being able to handle both effectively.

Why are small data problems challenging?

To a machine learning researcher, a model is simply a way of describing some state of the world, to which we can then ask questions. Models are usually not designed to reflect absolute truth, but instead attempt to provide a helpful simplification of some real-world behaviour that is still rich enough to encapsulate whatever properties of the data are relevant to the problem at hand. At the same time, models attempt to discard any data that is irrelevant to the problem or is simply noise.

Researchers have access to many different techniques or models to apply in any given situation. But a common property of many models is that they have some dials that can be adjusted to control the 'complexity' of the model. This gives the researcher the choice of building a simple model that can capture the broad strokes of a problem or a complicated model that might represent the world in fine detail but will run the risk of ‘overfitting’ to irrelevant data and failing to generalise well to unseen situations. In general, the risk of overfitting is heightened if the model is overly complex relative to the amount of data available for training.

Why is this relevant? It turns out that many recent AI breakthroughs have come from a particular pattern. Access to large amounts of data for training has enabled the construction of ever more complex models, which have led to significant performance improvements. For small data problems – where only limited amount of data is available for training – there is typically not enough available training data to support highly complex models.

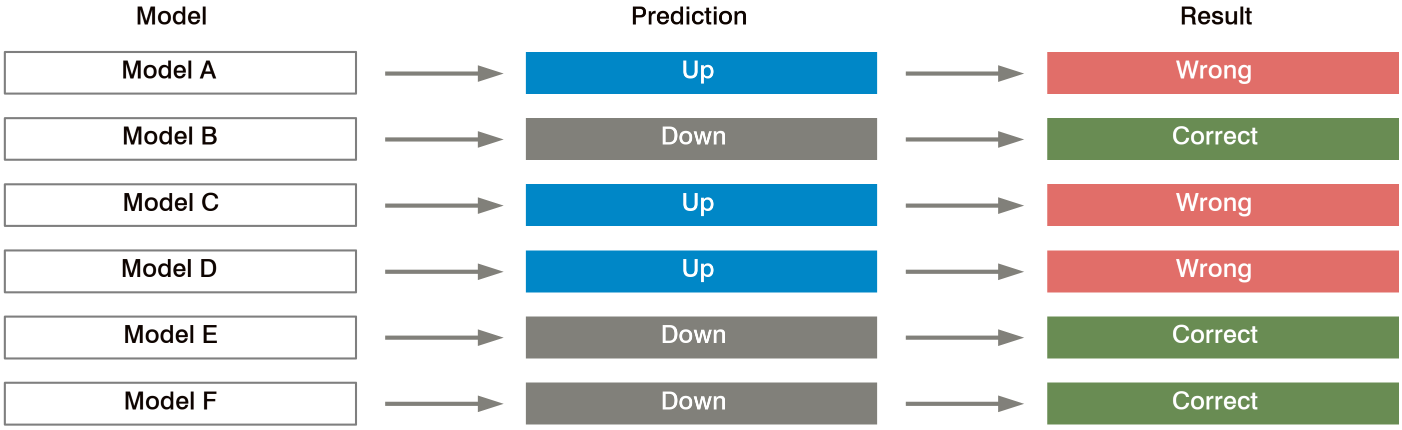
Small data problems in finance

Small data problems are encountered frequently in finance. For example, one of the most common goals in finance is making medium-term price forecasts, eg we could ask, “will the price of coffee go up or down over the next month or year?”

These kinds of questions are problematic because we do not have much data to work with. If we have access to daily price returns, we would have around 252 data points per year for a particular market. But if we are making a longer term prediction – say a weekly or monthly forecast – we would have far fewer samples.

Note that this problem remains even if we have a lot of data that could potentially be used to predict the market direction. The key restriction here is the number of price movements that we have available to try to predict. The problem arises because having a limited number of examples to predict makes it very difficult to distinguish between good and bad models.

Figure 3: At each timestep, we compare the predictions of six different models. It is difficult to distinguish between models that make the same predictions when we only have a limited amount of time-steps to evaluate over.



Model	Prediction	Result
Model A	Up	Wrong
Model B	Down	Correct
Model C	Up	Wrong
Model D	Up	Wrong
Model E	Down	Correct
Model F	Down	Correct

Source: GAM

Figure 3 illustrates this. It shows six predictive models, each of which is different. Some might be based around simple technical indicators while others could be based around very complex machine learning models. On any given day, each of the models will make a prediction and about half will be correct. The issue is that even if a model got the direction correct, we do not know if it was correct for the right reasons. When models come back with the same answer, we do not have enough information to properly distinguish between them.

In practice, the situation is somewhat more complex, as we will be performing the evaluation over multiple periods with access to more than one return. However, our ability to distinguish between data points is still weak and when the number of available returns is small, we become more limited in the judgements we can make about model performance. This, in turn, means that if we are using these judgements to train our models, it is much harder to have confidence that any model change represents a genuine improvement. This problem is even more acute when trying to build models for newly listed markets that may not come with much historical data, or when working with emerging data sources.

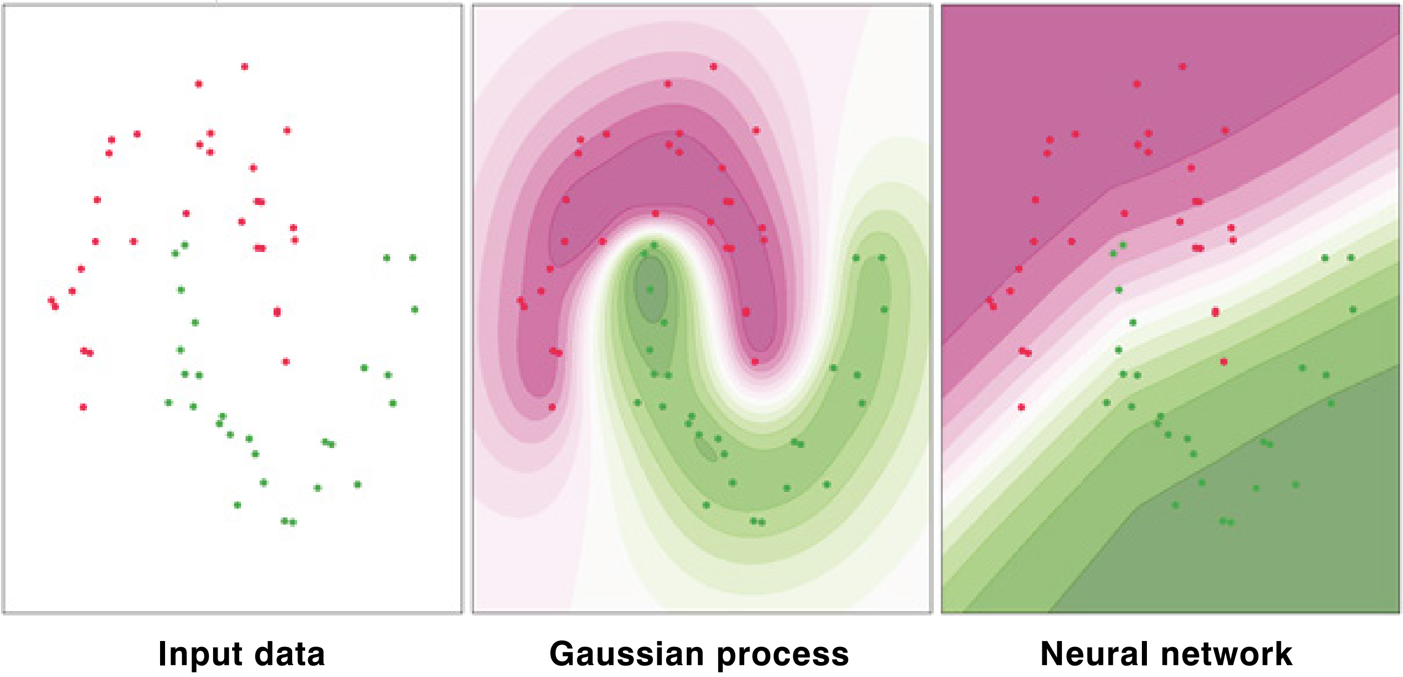
How can we handle small data problems?

For financial applications, we typically handle small data problems using a combination of approaches. The first task is to choose an appropriate technique to use. While there are many different types of model available in the toolkit of a machine learning researcher, some are more appropriate for small data problems than others. It is also important to match the model complexity closely to the amount of training data available. Often in finance, simpler models can be more robust and more effective than complex models, particularly when faced with new, unseen situations.

i. Model selection

While there is a degree of skill in choosing appropriate techniques, often some of the best approaches are those specifically designed for sparse, low-data situations. Probabilistic techniques, especially those based on Bayesian techniques can be some of the most effective approaches to use as they are able to explicitly model the uncertainty in the data.

Figure 4: Comparison of modelling techniques to fit two classes of data. The probabilistic Gaussian process (centre) as able to fit the data more effectively than the non-probabilistic neural network (right).



Source: GAM

An example of this is shown in Figure 4. We have a set of data drawn from two distinct classes, red and green. The two classes form a half-moon shape that is difficult to separate using simple linear algorithms. Figure 4 shows how two types of model attempt to distinguish between the classes. The approach in the centre image is probabilistic, based on a technique called Gaussian processes. On the right, we have a non-probabilistic approach based on a neural network. Although both approaches have access to the same data, the probabilistic approach in the centre is generally providing a better representation of the underlying data. Note that if we were to make more data available, we would expect to see the neural network approach improve, and eventually surpass the probabilistic approach. But in our limited, small data world, the probabilistic approach gives us a more robust fit.

ii. Data selection

If you have any additional data that is potentially relevant, it is normally a good idea to try to incorporate it into your model. In finance, this is most likely to involve taking data from other, similar markets. For example, to build a model for the S&P 500 you could also incorporate data from similar equity index futures. Incorporating data from a much broader universe of markets could help. However, this data may be of limited use if the market is very different to the S&P 500. In this case, a common paradigm is to take data from multiple diverse markets in order to build a single robust model. This can then be adapted to match a particular market using a smaller amount of market-specific information.

The macro investment space is often thought of as being highly liquid, with many futures markets having trading volumes orders of magnitude higher than the most liquid single stocks. However, the universe of available macro markets is also narrower, and liquidity tails off quickly. This means that there are many markets that have some trading volume but would not be sufficiently liquid to be traded in size as part of a systematic macro portfolio. Yet, these markets can still provide useful additional data for building robust models for more liquid markets.

A common trick when working with image data is to supplement the dataset by including additional versions of each image that have been transformed in some fashion, for example by mirroring or rotating the images. Similar transformations can be used to enhance a dataset of time-series data. A simple example is to add noise to each time-series. This can often improve models by making them more robust to unseen data. It is also possible to generate completely synthetic time-series data to use in model training. While this is often helpful for testing model robustness, it does have limitations. For example, it is very difficult to capture true out-of-sample black swan events when using synthetic data.

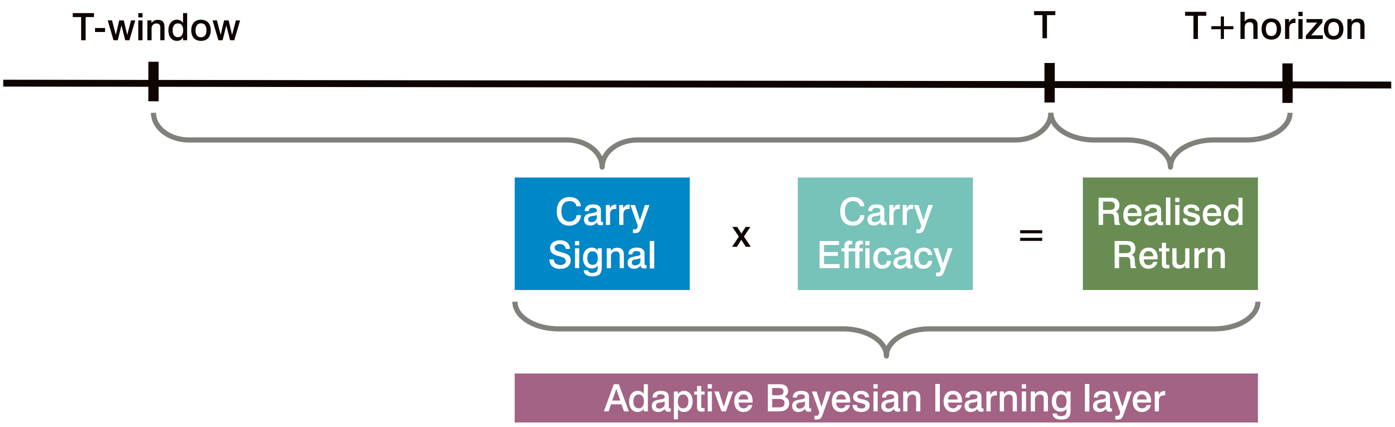
Example: Building a robust model for carry

We can demonstrate some of these ideas using a real-world example. A classic problem that we face in financial investing is to build a model that can pick up the carry in a given market. Here the carry is simply the yield we expect to realise from holding a long position in the market. In the case of foreign exchange, the carry is a function of the interest rate differential between the two currencies. We could expect to earn this difference by going long a high-yielding market and funding this via a short in a lower yielding currency.

We can also extend this idea to other asset classes. For example, in equities, the yield is a function of the dividends we would expect to receive and for bond futures, the carry is a function of how the coupon payments impact the shape of the futures curve. Similar definitions are also available for commodity futures.

It should be noted that there is a risk associated with trading carry. It is possible that the price of the asset will change while we are holding it and we will consequently lose money on our trade. What we would like to do is to build a model that can predict what the actual return is likely to be from our carry investment. This turns out to be an extremely hard problem! Not only are we interested in making forecasts over a medium-term horizon, in addition to that we are looking at indicators – such as dividend yields – that only change slowly over time. This is an extreme example of a small data problem.

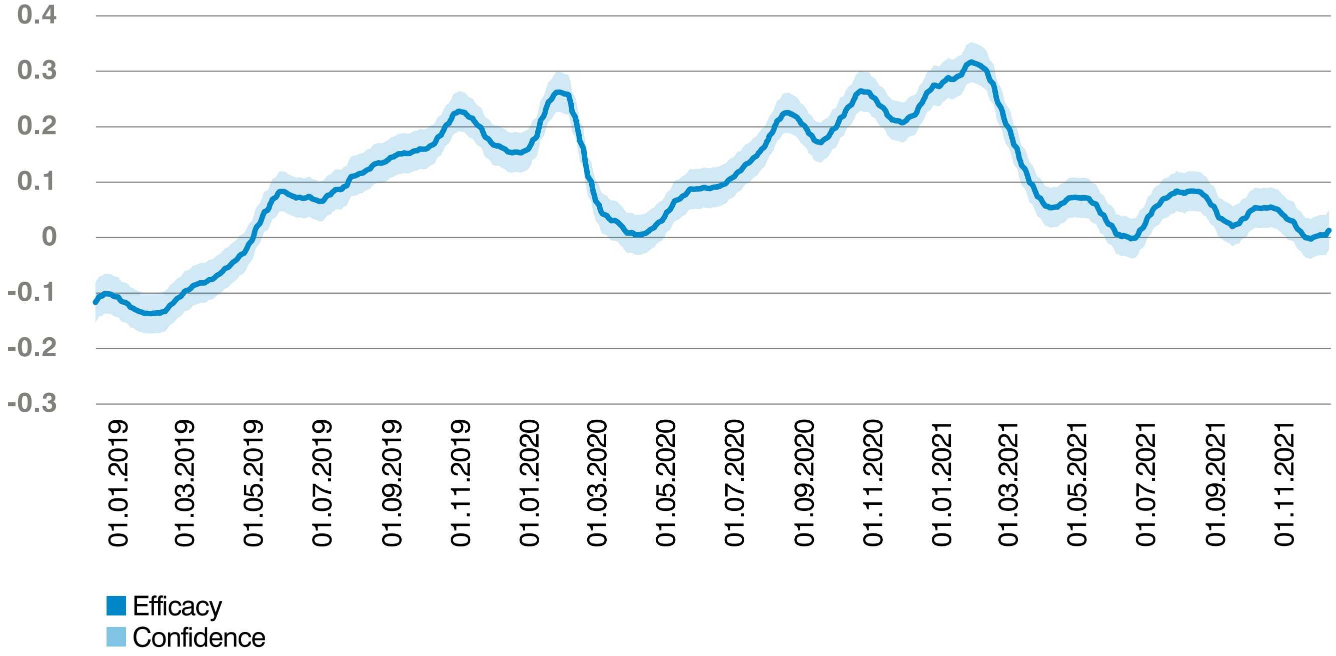
Figure 5: Diagram of a model for learning the predictive power of a carry indicator



Source: GAM

Figure 5 shows an outline of a model designed to address this task. It consists of an Adaptive Bayesian Learning layer, which sits of top of the raw carry indicator and gives a prediction of the efficacy of the signal at each point in time. This gives an indication as to whether we are likely to actually receive the market-implied yield, as well as giving some measure of confidence of its predictions. We can also make use of some of the data augmentation approaches discussed earlier and train this model using data from multiple different assets. This helps to build a more robust model, despite the limited data available.

Figure 6: Example of the output of the adaptive learning layer. A prediction of the efficacy (solid blue line) and confidence (blue band) of a carry signal for the S&P 500 future.



Source: GAM

Figure 6 shows the output of such a model for the S&P 500. The blue line represents a prediction of how the carry indicator relates to the expected returns. It can also adapt as the strength of this information changes over time. In equities most of the time the relationship between the carry and our expected returns is positive – indicated by the fact that the blue line is generally positive.

However, this is not always the case. A good example of this occurs in March 2020 with the onset of the Covid-19 pandemic. We can see that the model quickly realises that this is not a good time to be long equities. The predicted strength of the relationship between our carry indicator and the expected returns falls to zero, which would take us out of the market. In addition to the relationship itself, we also get a measure of the model's confidence in its ability to model this relationship, represented by the uncertainty bounds in Figure 6. These are useful for interpreting the strength of the fit that we learn, as well as providing additional information that can be propagated into other parts of our trading systems.

Conclusion

In this paper, we have discussed a category of problems where the amount of data available for building a model is inherently limited – we call these 'small data problems'. This contrasts with 'big data problems' where a lot of data is available for building very large, complex models. We deal with both big and small data problems in finance, but small data presents a particularly interesting challenge, as it is one of the fundamental reasons why finance is such a challenging application for machine learning.

The key to managing these problems effectively in systematic investing is to be able to clearly differentiate between big data problems and small data problems and to use the right tools for each. This is a distinction which is often overlooked. Many approaches which have been successful when applied to big data problems do not work very well when applied to small data problems. But, by focusing on approaches specifically designed to deal with the challenges of small data, we can enhance our portfolios, well positioning them to understand and navigate uncertainty in the times ahead.

References

[1] Silver, D., Huang, A., Maddison, C. et al. Mastering the game of Go with deep neural networks and tree search. Nature 529, 484–489 (2016).

[2] Berner, C et al. Dota 2 with Large Scale Deep Reinforcement Learning. OpenAI.

[3] Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021)

[4] Anderson, E. 1935. "The Irises of the Gaspe Peninsula." Bulletin of the American Iris Society 59: 2–5]

For more information, please visit GAM.com

Important legal information:

The information in this document is given for information purposes only and does not qualify as investment advice. Opinions and assessments contained in this document may change and reflect the point of view of GAM in the current economic environment. No liability shall be accepted for the accuracy and completeness of the information. Past performance is not a reliable indicator of future results or current or future trends. The mentioned financial instruments are provided for illustrative purposes only and shall not be considered as a direct offering, investment recommendation or investment advice. The securities listed were selected from the universe of securities covered by the portfolio managers to assist the reader in better understanding the themes presented and are not necessarily held by any portfolio or represent any recommendations by the portfolio managers. There is no guarantee that forecasts will be realised.

2465 21/11/2022

G|A|M Investments