July 2019 Whitepaper

QUOD FINANCIAL

SMART ORDER ROUTING AND AUTOMATED TRADING

A new wave of innovation

Table of Contents

Introduction

Liquidity fragmentation, the phenomenon of multi-listing of an instrument on different venues, creates deeper and broader markets and lowers transaction costs. Despite the benefits, it has become a major challenge to market participants from buy-side to sell- side institutions, as it has led to a more complex trading landscape. Some of the tools used to address this problem have been in place for a long period, while others are new to the market.

Smart Order Routing (SOR), which is undeniably the best solution to tackle liquidity fragmentation, has been around for over a decade. SOR is also increasingly the critical element in the building of any trade automation technology. Current SORs - both the simplistic rule-based and the newer generation using a liquidity seeking algorithmic approach - are offered by bulge-brackets and some vendors with a wide range of capabilities and limitations.

This whitepaper provides an update on our previous Smart Order Routing whitepaper and looks at the innovation in this domain. In addition, this paper is written as Artificial Intelligence (AI), and more precisely Machine Learning (ML), are rapidly reshaping the IT landscape. We will naturally look at the areas of application for ML in SOR and more broadly in automated trading.

Liquidity Fragmentation: An Evolving Challenge

Liquidity is shaped by two major forces; the regulatory framework and technology changes.

Regulation

The two main regulatory frameworks are:

- **US:** RegNMS is a 2005 rule-based equities regulation. It mandates that best price is seeked across all available venues. With nearly 15 years of existence, SEC/FINRA has now embarked on a review process to address its shortcomings, including the high cost of market data, the influence of dark pools on price formation, and the exchange 'favorable bias' towards High Frequency Traders (HFT).

- **Europe:** MiFID II, a 2018 multi-asset regulation, has greatly shaped the latest evolution of the European capital markets. A big push has been to restrict the operations of dark pools, and consequently the emergence of a new class of Systematic Internalizers (SIs), with the aim of increased transparency. The SIs are in fact divided into two distinct categories - the Bulge-bracket banks internal pools, and the principal-based market-makers, such as Jane Street or Virtu Financial. MiFID II comes on the back of 10 years of effort to push OTC trading to clearing houses to address the systemic counterparty risks, as made apparent during the 2008 financial crisis.

These two regional set of regulations are the de-facto global regulatory frameworks, being used often by other countries as the blueprint of their national regulations. For instance, European MiFID II's unbundling of sell-side services (execution, research, …) has been adopted on a global basis by all big buy-sides.

Technology

Fragmentation, by creating competition, enables innovative business and pricing models. It also drove the market for ever faster and ever more performant matching systems which have decreased latency and increased the volume being processed.

Fragmentation cannot happen until there is a mechanism to seek liquidity across venues and liquidity providers. As long as the amount of data and orders were low, traders could manually execute across venues. But with the rapid rise in volume of market data, of orders, and complexity of dealing with a large number of venues on a sub-millisecond time-frame, humans are inoperative. This new environment demands new SOR technology.

Multi-asset fragmentation

Figure 1 shows shows how each region has fared in terms of fragmentation on a per asset class basis.

Liquidity has transformed greatly in the past decade, with the following trends:

- **Liquidity is becoming ever more dynamic:** As competition between liquidity venues increases, price wars are becoming more frequent, and pricing models are being altered to attract more liquidity. For instance, the rebate model for passive orders has frequently been used as an effective promotion tool for new alternative trading systems. Clients are therefore moving their execution on a frequent basis from venue-to-venue, to take advantage of these taker incentive schemes.

- **Liquidity is becoming siloed and specialized:** Retail, ETF, or purely institutional liquidity are being traded increasing in distinct fashion and in dedicated venues. For instance, the biggest part of the retail flow in the US is executed against market makers (and is never sent to a public exchange / venue). The liquidity profile is also very much dependent on the daily timing and trading flow; e.g. as of 2018 NYSE has approx. 13% of the US securities market share, but 80% of NYSE traded volume is either at the beginning of the trading session or at the close. It means that NYSE plays an essential role in the initial price formation, and then in the market close it becomes the centralized market, which is a far more efficient mechanism than seeking liquidity across multiple venues. Interestingly, the closing phase in the equities market has become an essential trading point; this is due to position rebalancing requirement by ETFs and funds during the close; in fact makers know that takers are ready to pay more for liquidity at that moment.

Figure 1: Fragmentation in the US & Europe Per Asset Class

Asset class	US	Europe
Equities	**Exchanges:** 13 exchanges operating under the RegNMS (NBBO). Includes Nasdaq, New York Stock Exchange, Cboe Global Markets, IEX, etc	**Primary exchanges (previously national exchanges):** Borsa Italiana, LSE (London), Euronext (AMS, Brussels, Lisbon, Paris, Lisbon), Madrid Stock Exchange, etc
	Dark Pools / Crossing Services: BNY Convergex, BlocSec, Citi (liquifi), CSFB (Cross Finder), GS (SigmaX), Instinet, ITG Posit, Liquidnet, ML (AXP, CrossMix), MS (MS-Pool), Pipeline, Pulse, UBS (PIN).	**MTFs / :** Aquis, Cboe Global Markets, Turquoise.
	Marker makers: Virtu Financial, Jane Street, Citadel, etc	**Dark Pool/SIs:** Bank's SIs, Virtu Financial, Jane Street, Citadel, etc
	Changing landscape: The nature of the venues continues to change as M&A activities bring together multiple capabilities. New entrants such as IEX and Members Exchanges formed by Wall Streets banks.	**Changing landscape:** MiFID II push for decrease of Dark pools and creation of SIs. Europe is at the moment in a low number of MTFs, which results from a period of M&A.
Options	**Highly fragmented**, with competing exchanges - Boston Stock Exchange (BOX), CBOE, PHLX, NASDAQ Options, etc	**No fragmentation.**
Futures	**No fragmentation**, which is due to the barrier to entry erected by the exchanges via vertical clearing thereby arresting the fungibility across different venues, and the exchange IP protection of the trade marked contracts.	**No fragmentation** (same as in the US)
Foreign Exchange and FX products (e.g. Forward, NDFs, Swap, etc)	**Highly fragmented**, the nature of OTC nature of FX makes it fragmented. The technology landscape favors 3 main data centres (Tokyo, Lonodn and New York).	**Highly fragmented** (same as in the US)
	Different methods such as ESP and RFQ add a large layer of complexity.	

- **Liquidity is a zero-sum game and a poor SOR is an invitation to be gamed:** HFTs have perfected the latency game, and will notice if same orders appear on different venues' order books at different times. This case happens when sending multiple simultaneous child orders (for the same parent order) to different venues without considering the microseconds route difference between these venues. Gaming or spoofing techniques utilised by the HFT-like participants will shift the price prior to the liquidity even appearing in the wider market. Therefore, a poor SOR, or SOR which is poorly designed / implemented will therefore end up offering an amazing gaming opportunity to more agile players.

- **Fragmentation into dark pools is decreasing transparency:** As new dark pools flourish, transparency levels are decreasing. The present estimates are that 17% of overall US equities trading occurs in such venues. There is a lively US debate surrounding transparency and price formation on dark pools. We believe that if proper post-trade monitoring is implemented, the advantages of dark pools can outweigh all of its disadvantages. Yet, it is undeniable that the increase of dark pools creates a significant trading challenge: The price discovery model is based on efficient and transparent markets, which is now being pushed into a non-transparent execution venues. As transparency decreases, there is an urgent need to rebuild a real-time picture of liquidity which includes both the transparent and non-transparent liquidity.

A Current Picture of Smart Order Routing

The first effective smart order routing solution was launched in 2001 by Lava. It offered a low cost and efficient way to route orders to US exchanges and newly launched ECNs. The introduction of this solution accelerated the fragmentation away from the primary markets which, until then, had remained in the low single digits percentage. Lava's smart order routing was based on a best price formula, which rapidly led to US fragmentation.

The current equities landscape of smart order routing is quite different between the US, where RegNMS has shaped the way smart order routers function with National Best Bid and Offer (NBBO) routing logic, and Europe, where MiFID I and MiFID II have created more complex SORs.

The majority of current US SOR implementations are based on one round liquidity scanning logic and some simplistic rule-based posting. For instance, many large tier one brokers have comprehensive connectivity to exchanges, ECNs and ATS' (around 80 equities venues to connect to) which is coupled with a SOR. This SOR is prevalently based on a rule-based logic operating on a sequence of actions:

1. **Initial phase:** The SOR takes an initial decision to route orders to different liquidity pools based on a snapshot of the market.

2. **Periodic and discrete decision making:** If the initial routing round leaves unexecuted amounts, new routing decisions are made on a periodic basis (e.g. every n-milliseconds to n-seconds). One of the obvious shortcomings of this approach is that between these two decision periods, execution opportunities will not be considered.

Additionally, the US has a non-negligible number of second and third tier sell-side firms who utilise Exchange and ECN smart order routing mechanisms (also called market route-away). These engines are in fact in place to comply with the Reg NMS top-of-book protection rule (NBBO). In this case, an order is placed on a given market, the market fills the order as best it can or must then routes the remainder of the order (away) to the best price venue. The advantage is that it is a low cost solution to liquidity fragmentation. But this also has a few disadvantages:

1. The market will only route away the top of the market, but otherwise will not seek liquidity on other venues. It is obvious that an exchange will favor primarily its own liquidity to another venue.

2. The NBBO represents only a percentage of the available liquidity, and by focusing too narrowly on this available liquidity, it misses the overall objective of comprehensively seeking all liquidity.

The US picture of smart order routing is best represented by the tier-1 institutions, which effectively play the role of liquidity venues with their algorithmic trading, dark pools and ATS'es. The internalization of orders is the only way to stay competitive, as exchanges fees/costs are too high for a razor thin margin business. Then there are the exchanges and ECNs which, in addition to being liquidity venues, have a routing-away mechanism to comply with Reg NMS. And finally, lower tier sell side institutions which have a mix of tier one smart order routing facilities, exchange / ECN-based routing away and vendor solution.

Europe institutions have a different situation to address. With the more complex execution policy definition in Europe, there has been the necessity to provide more sophisticated technology capable of handling multiple criteria such as price, cost, preference and immediacy. SOR investment, after an initial phase post-MiFID I, has been rather low for mid-tier firms. It was mostly a consequence of the number of European single exchanges and their high cost of non-display market data fees. A remarkable fact is that the non-display charges are identical for all firms regardless of trading/order size, which favors large brokers and banks.

SOR and Algorithmic trading

Sell-side algorithmic trading - for example, execution business used for internal and client purposes, such as an agency broker- has been conceived with the objective of reducing the impact of the order flow. In a simplistic description, algorithmic trading is essentially slicing and scheduling of orders, with rather a limited consideration of real-time liquidity and price formation. In addition, algorithmic trading and smart order routing investments have been made at separate times, and therefore result in two separate infrastructures. The typical architecture is based on a 2-step process: first comes the Slicing and dicing of the orders, and then each child order is sent to a SOR. This is an inefficient, complex, and creates duplicate trading mechanism.

In a revamped architecture, a single execution infrastructure should be in place for algorithmic trading, which has a primary objective of reducing trading impact, and for smart order routing, which has a primary goal of seeking liquidity.

SOR and Internalisation

Sell-side institutions have embarked on pro-actively managing their internal liquidity which can be considered as 3 distinct type of liquidity (1) Incoming client order flow, (2) Internal market-making and (3) Trading activity and proprietary positions. They can then engage in systematic internalisation which is most appropriate for the very few firms that have a sizeable percentage of the liquidity of certain instruments, or in some cases for opportunistic internalisation.

Different internalisation scenarios are then considered:

- **Crossing against resident orders and dark pools:** Program and block trading represents one of the first cases for internalisation. The program trading resident orders, or large/block client orders are left for a given period of time, which can be from single seconds for pure dark pools, to days for a program. The applied price is taken from the transparent markets, with a vast majority matched at mid-point (mid-spread).

- **Crossing against internal inventory:** This is used when a client order or an opposite side in-house order is executed from an existing proprietary position. Different parameters can be set up in order to decide which positions the firm allows to cross against, at what price and for what quantity. Different decision making criteria based on position keeping/P&L metrics can also be set up. As a general rule, buy-side firms do not want their order flow to be published to the outside trading community or in-house sell-side traders; this results in information leakage that is detrimental to their execution objectives. The leakage affects proprietary trading activity, and also market making, which would look to hedge the position by ‘going’ out to balance the position.

- **Crossing in a dynamic way (Internal pool connected to the external exchange and venues):** Since internal crossing is a difficult endeavor, liquidity management is best achieved when a strategy of combining internal pools and external liquidity. This means that this internal liquidity is treated as yet another venue, with its own ‘favorable’ economic terms. This approach creates an efficient systematic internalisation mechanism, without creating an upstream matching workflow, which would increase latency and performance.

Internalisation and smart order routing are, in effect, complementary. No client wants to be trapped indefinitely within an internal or dark pool, but would like to maximize execution by tapping into the growing liquidity pool. This can be done by creating internalization matching combined with a SOR to simultaneously seek external liquidity. In addition, it has to offer an efficient protection mechanism against information leakage.

SOR and Dark pools and SIs

Dark Pools: The dark pools operate mostly on a mid-point (also called mid-spread) matching logic, without providing any price/transparency. In this case, it is really the role of the SOR to seek liquidity but also to verify that the mid-point price is respected. That requires an efficient way to define a real-time benchmark and monitoring of the spread capture. In addition, as there is no information, the dark pool SOR logic must include a mechanism to decide how much liquidity to route to a given pool, as well as the real-time capacity to re-route more liquidity to the pool demonstrating available liquidity. Finally, one of the issues that has surfaced with certain of these pools is the risk of information leakage. This usually occurs when the pools don’t have enough liquidity and sends orders out of the venue to find more liquidity. This information provides to HTFs signals of the venues’ liquidity, and creates an opportunity to game. So the SOR and more importantly the real-time analysis should include mechanisms to detect information leakage and subsequent decision to disconnect a leaky pool.

SIs Streaming: The increase of Systematic Internalizers (SIs) offering streaming quotes in equities, as it is the norm in the FX Spot market (called ESP) is relatively new. This means that the SOR must be able to function in tandem on a quote and an order basis. This requires a revamp of most equities SORs which have been conceived for order-driven execution mechanisms. For instance, we have used our own FX implementation to adapt to the new equities trading mechanism.

The Case for Adaptive Trading Technologies

The above description of what constitutes the 3rd reincarnation of smart order routing demands a comprehensive revamp of the technology which we grouped under the acronym Adaptive trading technologies. **Adaptive trading technologies** are a set of technologies, encompassing Software, Data management, and Software configuration/management, which at its core provides real-time event processing, natively incorporate Machine Learning methods, and enable change of the behaviour of the overall software (SOR) in near time (or real-time). In addition, it should handle very complex scenarios with multiple algorithms (execution policies) covering the complexity of liquidity venues individual characteristics (e.g. order types) on a multi-asset basis. Realistically, the implementation of these technologies need to leverage the current trading investment in terms of infrastructure, upstream and downstream systems - such as risk management or connectivity - and be as non-disruptive as possible.

In a more granular way, a few core requirements should be added to make a SOR truly adaptive:

- **Ability to dynamically process events:** There is still a lot of SORs that cannot process events in real time. That means that execution is at best poor. The real-time processing should come along the ability to manage different type of execution, i.e. order-driven and quote-driven, as liquidity is now both in exchanges, but also streamed by some market makers.

- **Ability to manage all the market specificity and pre-trade risk:** markets provide a wide range of order types, which need to be normalized to a given SOR. In addition, market or regulator rules, including pre-trade risk, such a price control, must be natively supported to avoid a high reject rate.

- **Ability to protect against high market data volume:** the SOR must protect itself against market data surges to remain effective even during high volume market events.

- **Ability to process a lot of concurrent SOR algorithms:** Some passive policies - e.g. Retail type of SOR - will be far from the top of the book passive orders. Mechanism must exist to ensure that the pressure on the SOR is managed.

- **Ability to dynamically change policies:** As explained before, with the fast maturity of ML into the SOR domain, the need to quickly take new parameters/configuration in a given algorithm (or sometimes called policies) is imperative.

Our own experience has demonstrated the following:

- **Speed matters:** There is a trade-off between the complexity of the decision making process (also called an algorithm), and the effectiveness and speed of the algorithm. The right balance is usually empirically found on a case-by-case basis. This requires that the algorithms themselves are benchmarked on a continuous basis. For instance, we consider the Hit-ratio (the percentage of consumed liquidity - most often in an aggressive mode-, expressed as how many Orders were executed compared to the Orders sent, in a given venue) as one of the essential SOR performance metrics.

- **Latency matters, but it is not the only winning formula:** Latency reduction is an eternal goal of all systems and more specifically execution systems. But solely benchmarking a system on latency will lead to false judgements. Latency in trading is always relative (and never absolute); your latency may be good, but is it better than a HFT? In reality, a broker will rarely be able to spend as much on latency as a HFT. In addition, a lot of liquidity today is 'Ghost liquidity' which means that it is a maker liquidity seeking a taker liquidity, and this has low liquidity stickiness. So chasing this type of liquidity is arduous, but also constitutes an opportunity cost of missing a better quality liquidity. Finally, as explained before, micro-second discrepancy of latency in sending an order to multiple venues, will provide information leakage and arbitrage opportunity to other participants. So it is sometimes better to slow down orders so all arrive at the same time to different venues.

- **Testing the 'machine' requires a new approach:** common quality assurance and testing approaches do not apply to these very complex decision making processes. Testing permutations for just three to four liquidity venues and some real-life client trading behavior exceeds hundreds of thousands of cases. The cost and length to arrive to such an extensive level of testing is excessive. It requires a new approach using statistical sampling of test cases (and the art of sampling is difficult) and reflects, as close as possible, the production environment. You then arrive at a stable production algorithm that enables further production testing which shows the effectiveness of overall decisions. In addition, there is a need to build a test harness able to replicate the complexity only available today in production environments.

Machine Learning & Smart Order Routing

As discussed in our two previous whitepapers, Machine Learning (ML) and data-driven software are big leaps in computing. More critically, ML brings Implicit programming to the fore, enabling to code new programs without the necessity to code (or explicitly know) all the behaviour of a machine.

Most advanced SOR's are already making use of the wealth of information in the post-trade, therefore creating an intelligence to the whole decision making process. This is achieved by analysis of available, such as:

- **Probabilities of execution:** This represents the probability of a single instrument on a venue using historical data. This can be refined create the probabilities lifting (also referred to as Aggressive) and resting liquidity (Passive).

- **Calculation of opportunity cost:** This represents the opportunity cost of going to a certain venue over that of another. The opportunity cost would be a function of price, cost (under the transaction cost analysis), immediacy, or slippage

What ML provides in the SOR context is a low cost and efficient mechanism to extract data patterns and implicitly find new trading workflow. It means that trade data derived from the internal EMS, Best Execution reporting and external data providers can be used to find patterns, such as Probability of execution, but also be used to refine some of the behaviour of the underlying SOR algorithm(s). This also creates a demand on the trading and execution platform to be able to provide such granular and specific data in real-time.

One of the consequences of such a shift for the incorporation of ML into the SOR technology is the need to move a SOR technology which is open and flexible to take such dynamic parameters for its decision-making. For instance, if one of the data demonstrates that the toxicity of the venue, the re-routing should be automated and new routes taken directly by the SOR.

To give another example, it is already possible to automate decision taking, such as:

- A process-driven discovery of liquidity across the dark and grey pools. This enrichment may allow the use of near time post trade execution data to direct orders to a pool that has demonstrated residual liquidity. Examples of data which can be collected include resting time on a pool which often varies for each venue.

- A similar process can also include transparent markets in a hybrid mode where the smart order router mixes execution across dark pools, transparent markets while retaining the remainder inside the router itself. This ensures that all available liquidity is secured with minimum leakage.

As pre- and post-trade data are analysed in near-/real-time, decisions about which pools to direct orders to are based on micro-data, which will provide a true intelligence to the SOR.

On a grander scheme, the SOR is part of an EMS. And here, we foresee a more radical evolution, which we call the New Trading Architecture. This New Architecture will then be constructed around:

- A set of price and liquidity **Predictive Agents** (which today would be the rudimentary pre-trade TCA) which is designed to predict the outcome of the execution. The prediction deals simultaneously with the price optimisation for timing / impact, and liquidity optimisation for the selection of venues and levels of liquidity.

- A set of **Execution Agents** which are the execution methods used, including the algorithmic trading / SOR capabilities,

- A calculation and data presentation set of **Data Agents**, which work in a feedback loop. They aim to quickly update the predictive models, as new data points become available. To be clear, as a "dirty" but fast and frequently updated prediction, is far better than a superior but slow model.

The whole process is tied together by Actors, which can / will be human users, or other machines in conjunction with human actors, taking **Meta-decisions** on the path of the execution.

Figure 2: Logical View of the New Architecture

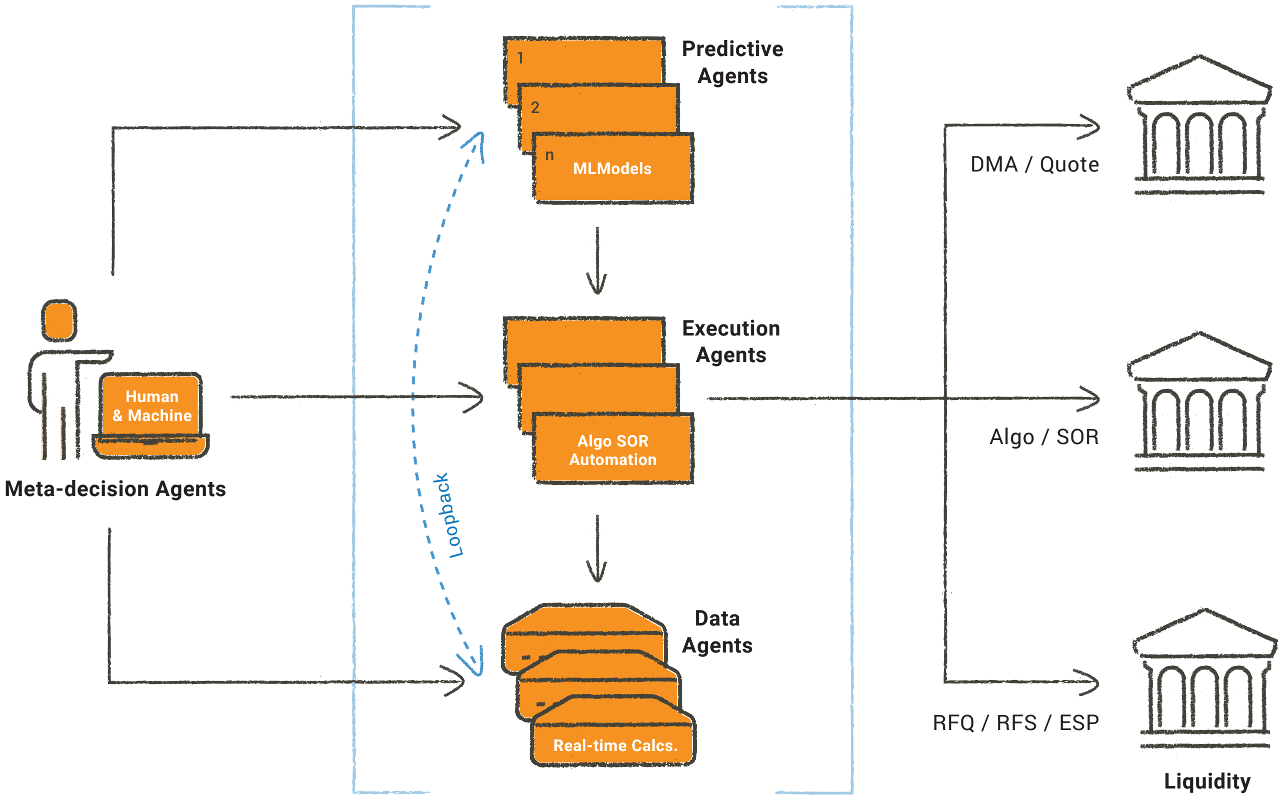


Figure 2 shows a proposed New Architecture is possible because Machine Learning allows the migration away from the current expensive demand of explicitly building and updating all the predictive models. Instead, the new ML techniques can implicitly update the models that are first trained with big data sets. Then the trained models are tightly integrated into the execution software to provide new predictions but also be re-trained with each execution.

Conclusion

Mainstream sentiment, driven by the post-financial meltdown backlash (and popular books such as *Flash Boys*), is that trading and capital markets are beyond control and need to revert to a simpler model. This is a simplistic narrative. In reality, while the level of financial innovation increased, technology investment didn't always keep pace. It is unrealistic to manage a far more complex landscape, such as the current level of liquidity fragmentation, with the same old technology.

A SOR must be based on available Adaptive technologies which provides the following out-of-box capabilities:

1. To algorithmically process under high market data volume all events to take real-time decisions. This also means the capacity to manage different type of order flows (retail, institutional DMA, Algorithmic, etc) and a set complex decisions (lit, lit+dark, lit+internalization, lit+dark+Streaming, etc for a variety of order types) for a high number of concurrent orders.

2. To enable near time or real time changes to the underlying processes (usually via configuration) without code change and / or complex change processes.

3. To provide granular traceability on how the execution decisions were made. It usually comes within the Best execution reporting, but is now an essential part of the real-time execution performance analytics. It should also include real-time monitoring of the 'machine state' and ability to manage all exceptions for such a highly automated trading.

4. And with the advent of Machine learning and data-driven computing, to provide the ability to incorporate ML-type decisions or predictions (e.g. a ML peg) directly within the SOR.

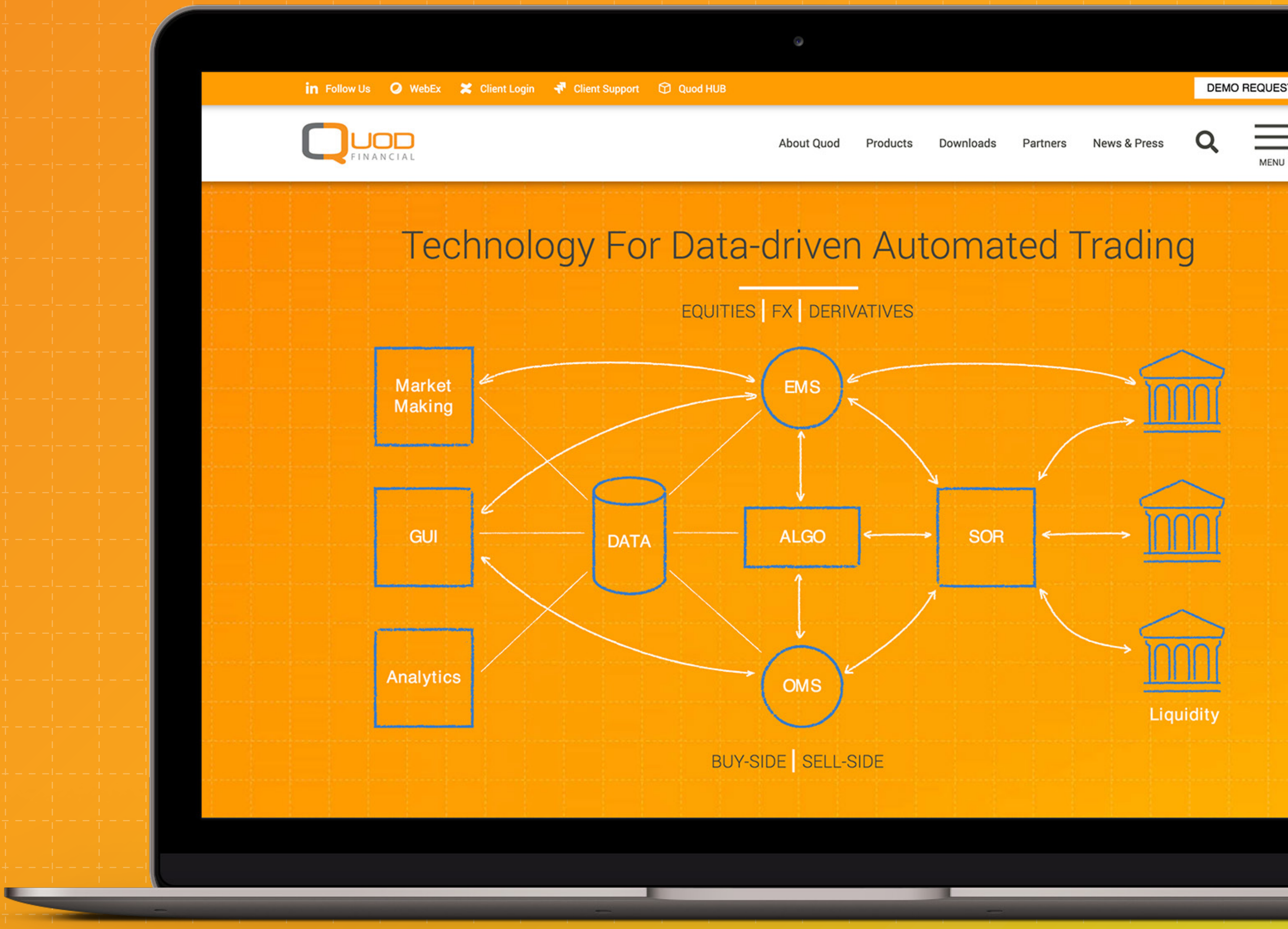
We are at the beginning of a new, and quite certainly fast-paced, Data-driven (Machine Learning) computing revolution. It will impact all aspects, including technology, but also the underlying processes and human assets. Low or no investment at this innovation phase will quite certainly mean a death sentence for a business which is now a fully technology-driven sector.

> "I do not fear computers. I fear the lack of them."
> — Isaac Asimov

Quod Financial is the fastest growing Multi-Asset Trading Platform in Europe. Having consistently prioritised pure performance, flexibility, impartiality and customisation for 14+ years.

Quod has become an inevitable go-to technology when it comes to sophisticated Smart Order Routing and Algo capabilities, helping automate trading and optimise execution.

LONDON | PARIS | NEW YORK | HONG KONG | DUBAI
+44 20 7997 7020 | +33 974 59 4445 | +1 92 92 92 8090 | +852 300 83 775 | +971 8000 320 113
info@quodfinancial.com | www.quodfinancial.com



Follow Us
WebEx
Client Login
Client Support
Quod HUB
DEMO REQUES
About Quod
Products
Downloads
Partners
News & Press
MENU
Technology For Data-driven Automated Trading
EQUITIES | FX | DERIVATIVES
Market Making
GUI
Analytics
DATA
EMS
ALGO
OMS
SOR
Liquidity
BUY-SIDE | SELL-SIDE