# Ultra High Frequency Financial Data

Martin V. Sewell
Department of Computer Science
University College London
Gower Street
London
WC1E 6BT
United Kingdom
M.Sewell@cs.ucl.ac.uk

Wei Yan
Department of Computer Science
University College London
Gower Street
London
WC1E 6BT
United Kingdom
W.Yan@cs.ucl.ac.uk

## ABSTRACT

This note is best described as a ‘Research Challenge’, and concerns building an ultra high frequency (UHF) trading system. The emphasis is on addressing the problems posed by UHF data, with a few thoughts on strategy and implementation. The problem may be amenable to evolutionary computation.

## Categories and Subject Descriptors

E.m [**Miscellaneous**]; J.4 [**Social and Behavioral Sciences**]: Economics; I.2.m [**Artificial Intelligence**]: Miscellaneous

## General Terms

Economics

## Keywords

high frequency, data, finance, evolutionary algorithms

## 1. INTRODUCTION

Intuitively, the surest way of making money in the financial markets is to imitate those who make money in the financial markets. Or, rather, replicate the strategy of those who make money due to skill, rather than luck. Distinguishing skill from luck in trading is notoriously difficult, but statistically those who consistently make above-average risk-adjusted returns are more likely to be generating alpha[1]. The set of financial trading companies with the most consistent and profitable track records tend to be ultra high frequency (UHF) statistical arbitrage funds like D. E. Shaw, Renaissance Technologies, Citadel and Man Group’s AHL. For example, Renaissance made approximately a 34 per cent annualized net return since its 1988 inception [12]. So, how do the best players predict the direction of today’s increasingly efficient markets? The answer is, they don’t. Such funds tend to be ‘market neutral’, have huge daily trading volumes, call themselves ‘hedge funds’, and behave more like unofficial market-makers [8].

The key challenge presented by high frequency trading is how one can efficiently exploit a massive data set (e.g. 20 million data points per futures contract per day) to develop profitable trading strategies. Evolutionary algorithms such as GAs and GP are well suited to discovering complex relationships within high dimensional data sets. Firstly, they are often a good choice in real-world problems when there is little domain knowledge; secondly, when used for symbolic regression, information about the target equation is not necessary; thirdly, they support a continuous training mode that can be used when training data is changing dynamically; and finally, the population-based approach is easily amenable to parallel computation and therefore rapid execution. Our task is to analyse the data and develop an automated (unofficial) market-making algorithm.

UHF trading, in common with arbitrage, has become an arms race. The successful trading of UHF data is necessary to consistently generate alpha in financial markets, and the most viable method of achieving this is machine learning. However, the best algorithm in the world is not sufficient. In practice, low latency times are essential if one wishes to utilise a UHF trading system. The edge most companies have is in their infrastructure and location (with a server physically located at the exchange). Don’t try this at home.

## 2. DATA

Today, tick data generated by financial markets quite possibly represents a greater volume than any other source outside high energy physics. Futures markets provide the most data, followed by foreign exchange (FX) markets, followed by stock markets, although most of the literature relates to stock markets. It is likely that the type of modelling required would be similar across all three types of market, and the level of generality in this paper is such that the text applies to all three types of markets, unless otherwise stated. In practice, we are more likely to be concerned with futures markets or FX markets because transaction costs are vanishingly small and leverage is possible. Thanks to Moore’s Law [10], it is now possible to tackle the problem.

In the area of UHF trading we’re interested in tick-by-

[1] *Alpha* is the return over and above that predicted by an equilibrium model like the capital asset pricing model (CAPM) and is a proxy for an (active) investment manager’s skill.

*GECCO’08,* July 12–16, 2008, Atlanta, Georgia, USA.
Copyright 2008 ACM 978-1-60558-131-6/08/07 ...$5.00.

tick data, and domain knowledge comes under the guise of 'market microstructure'. *Market microstructure* is a branch of economics and finance concerned with the details of how exchange occurs in markets, most commonly financial markets. Market microstructure research typically examines the ways in which the working process of a market affects trading costs, prices, volume and trading behaviour. For more information on market microstructure, see [11, 9, 4, 1, 6].

*Good news*:

- Masses of data, e.g. 20 million data points per contract per day. A rich data set with the potential for high statistical significance.
- Costs in futures and FX markets are tiny; FX is the best with $1m of notional costing $3 to trade, whilst futures costs are considerably less than one tick. Costs are dominated by the spread.
- Due to low costs and leverage, it is only necessary to predict with an accuracy slightly better than random.
- The financial rewards are large, with potentially minimal risk.

*Bad news*:

- Masses of data, e.g. 200MB per contract per day. Computationally intensive.
- In common with any financial time series, the data is generated by a nonstationary, stochastic, discontinuous and probably nonlinear dynamic process. Any useful (i.e. profitable) signal is extremely noisy.
- The distributions generated by the stochastic processes can not be described, or even approximated, by any parametric model (the distributions are not even close to Gaussian).
- Because one can trade without putting any significant money up, the concept of 'return' is ill-defined (this is only bad news in the sense that it makes performance measurement and thus comparisons difficult).
- Intra-day volatility is a better forecaster for volatility than ARCH/GARCH, so ARCH/GARCH is useless, whilst the concept of volatility is fairly meaningless anyway. This may not be directly relevant to the prediction of price, but is another example of standard financial assumptions breaking down at high frequency.
- The data is not isochronous (data does not occur at equally spaced time intervals), how can we model this?
- The data is not contemporaneous (different streams of data do not arrive at the same time), so you can not correlate different markets.
- Testing a strategy is difficult, just because a trade took place at a particular price does not mean that we would have got that price.

*Six dimensional data*:

- trade (price)
- bid
- offer/ask
- traded volume
- bid size
- ask size

The above is an over-simplification, and only includes the top of the 'order book'. An *order book* is a compiled list of orders (prices at which traders are willing to buy or sell) received. There is structure and some information contained in the order book. There is evidence that the order book beyond the first step provides 30% of the information [2].

Data would be prohibitively expensive, so collaboration with a bank or hedge fund who have the ability to capture and store order book data is essential.

## 3. ALGORITHM

A *market-maker* is an intermediary who creates a market for a financial obligation. In a given market, he must quote two prices: the lower is the *bid* (the price at which he is willing to buy) and the higher is the *offer* (or *ask*) (the price at which he is willing to sell). The difference between an offer price and the bid price is known as the *spread*. A market-maker receives the full order flow, so is in a unique position to profit from the stream of data received.

The task is essentially to design an automated market-making algorithm; it would need to accommodate the following three objectives: (1) attract order flow, (2) control inventories and (3) avoid losses to informed traders ('adverse selection'). At least in stock markets, limit orders are disproportionately more likely to come from informed traders [5, 7]. *Bluffers* are profit-motivated traders who try to fool other traders into trading unwisely; to avoid losing to bluffers, market-makers must adjust their prices so that buy and sell orders have equal (but opposite) market impact per quantity traded. If most traders use market orders, spreads will be narrow; if most traders use limit orders, spreads will be wide. A market-maker may discover the *equilibrium spread* by adjusting his spread so that limit orders and market orders are equally likely. Spreads increase with (1) the degree of information asymmetry among traders, (2) volatility, and (3) utilitarian trading interest (a *utilitarian trader* trades because they expect to obtain some benefit from trading besides profits).

Unlike an official market-making program which is obligated to offer a bid and an offer whilst the market is open, our algorithm has the option of being more dynamic and reactive by taking prices which are offered and/or only trading when it is optimal to do so.

Broadly speaking, there are two types of market analysis. *Fundamental analysis* is a method of forecasting markets through the analysis of relevant news; whilst *technical analysis* is a method of forecasting markets through the analysis of data generated from the activity of trading itself. The shorter the trade length, the more significant technical analysis becomes, so for UHF data, only technical analysis shall be considered.

In the London Stock Exchange, when a market order removes all the volume at the best price, it creates a change in the best price equal to the size of the gap, so large price fluctuations occur when there are gaps in the occupied price levels in the limit order book [3]. Similarly, with US stock markets, low density of limit orders in the order book, i.e. a small liquidity, is a necessary prerequisite for the occurrence of extreme price fluctuations [13]. At least in a stock market, one could aim to exploit gaps in the order book. Futures and FX markets may be too liquid to have 'gaps' in the order book, but the density of orders may contain useful information.

Evolutionary algorithms involve a population of candidate solutions and a fitness function. In this case, the candidate solutions would, in general, be transformations of the six inputs listed above (and optionally, information from deeper in the order book) used to construct bids and offers (limit orders or market orders). The fitness function would be some proxy for profit (the calculation of which is non-trivial).

## 4. CONCLUSIONS

The Research Challenge described in this paper sought to address the problems presented by UHF data. Modelling data of this nature is extremely difficult, but the rewards are high. A good algorithm is necessary, but not sufficient.

## 5. REFERENCES

[1] B. Biais, L. Glosten, and C. Spatt. Market microstructure: A survey of microfoundations, empirical results, and policy implications. *Journal of Financial Markets*, 8(2):217–264, May 2005.

[2] C. Cao, O. Hansch, and X. Wang. The informational content of an open limit order book. 31st EFA Annual Meeting - Maastricht, 18-21 August 2004 and Sixty Fifth Annual Meeting of the American Finance Association, Philadelphia, PA January 7-9, 2005, Mar. 2004.

[3] J. D. Farmer, L. Gillemot, F. Lillo, S. Mike, and A. Sen. What really causes large price changes? *Quantitative Finance*, 4(4):383–397, Aug. 2004.

[4] L. Harris. *Trading and Exchanges: Market Microstructure for Practitioners*. Financial Management Association Survey and Synthesis Series. Oxford University Press, New York, Sept. 2002.

[5] L. Harris and J. Hasbrouck. Market vs. limit orders: The SuperDOT evidence on order submission strategy. *The Journal of Financial and Quantitative Analysis*, 31(2):213–231, June 1996.

[6] J. Hasbrouck. *Empirical Market Microstructure: The Institutions, Economics, and Econometrics of Securities Trading*. Oxford University Press, New York, Dec. 2006.

[7] R. Kaniel and H. Liu. So what orders do informed traders use? *The Journal of Business*, 79(4):1867–1913, July 2006.

[8] H. M. Kat. Of market makers and hedge funds. Cass Business School, City University, London, Feb. 2007.

[9] R. K. Lyons. *The Microstructure Approach to Exchange Rates*. The MIT Press, Cambridge, MA, Dec. 2001.

[10] G. E. Moore. Cramming more components onto integrated circuits. *Electronics*, 38(8):114–117, Apr. 1965.

[11] M. O'Hara. *Market Microstructure Theory*. Blackwell Publishing, Malden, MA, Feb. 1995.

[12] S. Taub. Really big bucks. *Alpha*, May 2006.

[13] P. Weber and B. Rosenow. Large stock price changes: Volume or liquidity? *Quantitative Finance*, 6(1):7–14, Feb. 2006.