University of Bristol MSc in Data Science; DSMP (Data Science Mini Project; EMATM0050), January 2024.

Problem B: Leading-edge data analytics for Level-2 financial-market data
Problem owner: Dr Ash Booth, Global Lead in Applied AI/ML, JP Morgan, London.
The global financial markets are obvious sources of "big data". If we look at the market for only one tradeable asset, such as shares in Amazon.com, there are so many people buying and selling the asset that the share price can potentially move up or down (although typically each move is only by a small amount) several times per second for all the time that the market is open, and hence in one trading-day there could be 20,000 or more time-points for movements in the price of an asset. This would be quite a lot to process if the data of interest at each time-point was only a single value, only the share-price in dollars and cents, but very often we are interested in much more data than just the share-price for an asset. Traders in financial markets commonly work with data that summarises all bids (orders to buy) and asks (orders to sell) currently resting at the exchange: any trader looking to buy can post a "bid limit order" at the exchange, saying what price they are prepared to pay per share, and how many shares they wish to buy; similarly any seller can post an "ask limit order" showing how many shares they want to sell, and the per-share price they are seeking. Different buyers will have different price and quantity needs, as will different sellers, and so at any one time the stock-exchange summarises all of the currently-received orders by publishing its "Limit Order Book" (LOB), sometimes also called the ladder, which shows the total quantity of units of the asset available to buy or to sell at each price which has been quoted. The LOB at any one time will typically involve tens of different (price, quantity) pairs – and the LOB may change several times before any transaction takes place that results in a change in the share-price, so there might plausibly be 100,000 data-points in a one-day time-series for the LOB for a highly-traded asset such as Amazon stock, and each of those data-points would be a snapshot of the LOB as it is updated, so each of the 100,000 data-points will itself be a structure involving perhaps 50 numeric values or more, so in approximate figures we can plausibly expect data-files of 5million values from any one such stock, in any one day. Industry practitioners refer to this whole-LOB data as "Level2 data". There are good reasons to believe that executing appropriately advanced data-analytics on Level2 time-series data could identify opportunities for usefully predicting near-term movements in price, and hence for profitable automated trading from those signals.

The problem, put simply, is for you to implement and evaluate data-analytics techniques that could be useful in identifying trading signals ("buy" or "sell") in Level2 data. You will be issued with Level2 data-sets for this project, although the identity of the asset will have been deleted. Some data will be made available as soon as the project commences, for you to start work on, and then additional data may be released at later stages in the project: that data may not be for the same asset, or the same market-period, as the initial data-set, and so it is likely not to be statistically identical to the initial data-set, so you should plan accordingly.

My team has an ongoing research interest in exploring how well various reinforcement learning approaches perform at finding good trading strategies when working with Level2 time-series data. For example, we have an interest in the A3C approach, although we recognise that there is probably not enough time in your mini-project to research, design,

implement, and evaluate a full A3C system. Nevertheless reinforcement learning is a long-established field with a very large academic literature, and there may be simpler methods, or freely-available source code-libraries, that you can use to make good progress in the time available. You might want to start by
implementing an elementary time-series analysis approach such as ARIMA1, which is relatively simple and very well known, and which could serve as a useful baseline for comparing against, but our interests lie beyond such a commonly-used approach; and so should yours.

Remember that we do not just want to see a system that does time-series predictions, we want to see what profit your system might generate from actually trading on the basis of its signals: you'll need to reserve some of the Level2 data as a test-set, and to write (or find) a simple trading simulator so we can see how well an automated trading system would do when using the signals that your analysis identifies.

One final thing: the raw data sets that you will be supplied with may need some initial wrangling (cleaning, extraction, processing etc) before you can use them, and you will probably find that some initial exploratory visualization and data-mining is useful too.

Good luck!

Example Readings/Bibliography

There is a vast literature out there which is potentially relevant to this project. The readings listed here are likely to be helpful, but there are many others that could be just as useful.

Abergel, F., Anane, M., Chakrabort, A., Jedidi, A., and Toke, I. (2016) *Limit Order Books*, Cambridge University Press.

Bouchaud, J., Bonart, J., Donier, J., and Gould, M. (2018), *Trades, Quotes, and Prices: Financial Markets Under the Microscope*. Cambridge University Press.

Gould, M., Porter., M. Williams, S., McDonald, M., Fenn, D., and Howison, S. (2013) Limit Order Books. *Quantitative Finance*, 13(11):1709-1742. Available from: https://people.maths.ox.ac.uk/porterm/papers/gould-qf-final.pdf

Data

The trading data you will receive will be for a single tradeable asset over a period of time. For each time-period there are two data types you need: one is the exchange's "tape", a record of the time, price, and quantity of each transaction for the asset; the other is the exchange's LOB record, showing time-stamped records of the state of the exchange's Limit Order Book for the asset.

Data Link: Problem B data