# Task 3: Alternative Data Scenarios

### Questions

1. How would you generate investment ideas from a dataset containing information when executives buy and sell shares in the listed companies they work for?

At its broad concept, the insider trading data indicates executives’ confidence level of the companies they work for. Rationally, they would purchase (sell) shares if they are optimistic (pessimistic) and believe that the asset price would go up (down). Past evidence also shows that insider purchase beat the market in general; meanwhile, insider selling is a less indicative sign as it could be due to several reasons. Thus, insider purchase should be an essential factor in generating investment ideas. When processing data, it is crucial to identify the level of the insider as top executives (e.g., CEO, CFO, etc.) should have a more comprehensive view of the company’s performance as compared with VP’s. Moreover, the historical performance of the specific executives should also be factorised in. The purchase amount and the percentage change in holdings should also be examined because the larger the amount, the more critical the message. When analysing the investment opportunity, these information should be aggregate to firm level to further examine whether there is a crowd purchase/sell over a short period of time.

A practical challenge of insider trading data is on time, as most transactions are reported on the Form-4 filings, and there is often a time lag between the transaction date and the report date. Thus, the investment opportunity might have disappeared as one get hands on the data. While there are databases claiming to provide real-time insider trading data, the credibility of these data providers should be examined before using the figure.

1. Armed with anonymised web traffic data, how could you use the information contained within this dataset to trade in a streaming platform of your choice?

The streaming platform operates based on different kinds of business models (e.g., Ad-based model, subscription model, etc.), which has different revenue-generating methods. Thus, the first step should be to select the appropriate factors to be examined within the web traffic data based on the business model. An important factor is the activity level (new and returning visitors). Activity level is a key factor in revenue generation as it is directly related to more ads played and thus higher revenue for the platform. New visitors might be more important to subscription-based platforms, as they represent potential new subscriptions. Investors looking to invest in a streaming platform should look for a steady, increasing number of visitors, indicating the business is more stable and healthier. It is also important to look at pageviews/unique pageviews to ensure that visitors do not concentrate on specific content. One should also fit the factors into time series models or machine learning models to identify their relationship with the asset price and look for any trading opportunities. Alternatively, trying to predict the price movements after the financial statements are released.

A challenge for web traffic data is how one can factorise in the multi-equipment scenario when analysing the data, i.e., the same user account could use different equipment with different IP addresses and thus, being recognised as different users. This could lead to a bias in evaluating subscribers.

1. Using email-receipt data, how would you decide which company to buy in a fiercely competitive subindustry (food delivery, for example)?

Email receipt data contains information related to order details, which are important factors to examine at a cross-sectional peer level. Using food delivery as an example, email receipt data are expected to contain information such as order amount, order time, delivery time and other merchant identifiers. If one is looking to decide the right company to buy, one way of combining these email receipt data into the analysis is to use a multi-criteria model, along with other financial related factors. The order amount can be viewed as a revenue resource and thus, should be maximised. Order duration can be calculated, and a lower average order duration means more orders capacity within a specific time period and, thus, higher revenue upper bound. Thus, order duration should be minimised within the evaluation model. Along with other cost and revenue related factors, these can be put into a multi-objective optimisation model and evaluate 1) the optimal one or 2) the optimal combination to invest in.

Email-receipt data are often highly personalised and, thus, heavily affected by regulations such as the GDPR. Identifying a proper data provider who can comply with the lates regulations but still deliver as much useful information as possible is a challenging task. Also, as data are mostly on order level or even item level, the total amount is enormous (often in billions), which means processing the data, identifying the valuable information and aggregating data to firm-level would be difficult. The algorithm has to be optimised to reduce the time of processing.

For investment professionals only. Not for use by retail investors.

**JUPITER** 4