Structuring Time Series Data to Gain Insight into Agent Behaviour

1st Najim Al-baghdadi
Department of Computer Science
Royal Holloway, University of London
Egham, United Kingdom
najim.al-baghdadi.2019@live.rhul.ac.uk

2nd Wojciech Wisniewski
Department of Computer Science
Royal Holloway, University of London
Egham, United Kingdom
wojciech.wisniewski.2019@live.rhul.ac.uk

3rd David Lindsay
AlgoLabs
Bracknell, United Kingdom
david@algolabs.com

4th Siân Lindsay
AlgoLabs
Bracknell, United Kingdom
sian@algolabs.com

5th Yuri Kalnishkan
Department of Computer Science
Royal Holloway, University of London
Egham, United Kingdom
Yuri.Kalnishkan@rhul.ac.uk

6th Chris Watkins
Department of Computer Science
Royal Holloway, University of London
Egham, United Kingdom
C.J.Watkins@rhul.ac.uk

***Abstract*—Here we introduce a data staging algorithm designed to reconstruct multiple time series databases into a partitioned and regularised database. The Data Aggregation Partition Reduction Algorithm, or DAPRA for short, was designed to solve the practical issue of effective and meaningful visualisation of irregularly sampled time series data. This paper firstly discusses the rationale for DAPRA, walking through its design and introduces the theoretical foundation of any DAPRA application. Later we report empirical evidence that demonstrates the practical relevance of DAPRA by its application with large and complex time series datasets from two distinct domains (financial and travel).**

***Index Terms*—Big Data, DAPRA, Time Series Analysis, Data Visualisation, Database Management, ETL, OLAP Cubes, Data Staging, Business Intelligence.**

I. Introduction

In the age of information, big data is the resource that makes the difference between failure and success. Data collection focuses more on the initial task of gathering as much data as possible and ensuring the robustness of data collection and storage systems. Yet the significant task of building usable data structures is by comparison somewhat neglected. The result is that data scientists encounter challenges in producing meaningful visual analyses or implementing machine learning algorithms to predict future data trends. It has become common practice to store data using online analytical processing (OLAP) cubes since these structures allow for fast and effective querying and drill-down analysis. Many commercial products that make use of OLAP cubes are available to purchase for businesses to help them easily evaluate and manage their data. Studies such as [1] have proven the concept to be beneficial for financial analysis. Before data is stored in an OLAP cube, an intermediate process of ETL (Extract, Transform and Load) is carried out. This process (also known as cleansing and staging) involves extracting the raw data from its source to a data staging area (DSA), where the data is manipulated to suit the requirements of the target data [2]. For example, when dealing with data that is not synchronised or when a set of events have closed but many are still in flux, it is useful to first pre-process the data to take account for this. In the context of time series data, we have identified a need for a generally applicable tool capable of structuring irregularly-sampled data gathered from multiple sources into a format that is practically intuitive to understand, and further avoids loss of information because it is designed "from the ground up".

Interval analysis of time series data has been widely explored in papers such as [3] & [4], but we have not been able to find publications showing how to generate the interval data from source data. In this study we introduce and rationalise a novel DSA process that manages the challenges associated with irregularly-sampled time series. We call this the Data Aggregation Partition Reduction Algorithm, or DAPRA for short. We will describe how the DAPRA framework can reformat irregularly sampled data from two distinct domains (financial and travel). We further illustrate how end-users can use DAPRA to gain previously "unseen" insights into agent behaviour to supplement current business intelligence and decision-making capabilities.

II. Data Aggregation Partition Reduction Algorithm (DAPRA)

DAPRA is specific to analysing time series data which is derived from some concept of "agents" interacting with their environment. Fig.1 illustrates this concept across three distinct time series datasets. The agents are (from left to right) foreign exchange traders, licensed taxis or network

This research was sponsored through match-funded PhD agreements between AlgoLabs and RHUL.

IP addresses. These agents interact with their respective environments of an online trading platform, pick-up/drop off geographical locations in New York City (NYC) or a computer network. The main requirement for DAPRA is that each agent's action has a start time at which that action was started (i.e. the start time of a network data transmission) and a finish time at which that action was completed (i.e. the time at which a taxi dropped off a customer and so finished the journey).

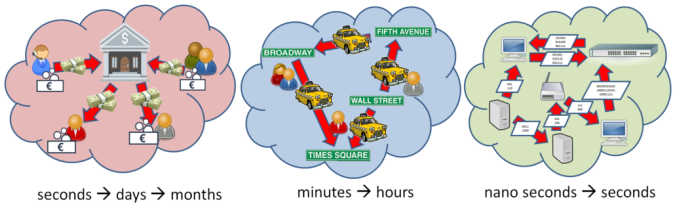

Fig. 1. Agents interacting with their environment

Fig.1 shows that the time spans of such actions by agents vary according to the problem domain, ranging from short lived (i.e. *nanoseconds* for a network data transmission) to very large (e.g. a trader could keep a trade open for *several months* until they speculate on a favourable price movement that will increase the profit on their trade). Time series data can easily contain hundreds if not thousands of agents and all agents are free to carry out as many actions as they are capable of (e.g. an inner city taxi driver makes more short-duration journeys than taxi drivers operating on the city outskirts).

The frequency or regularity of actions by a given agent can vary too, for example a trader can place hundreds of trades in one day but in the week that follows place none at all. There is no obligation on a trader to trade, it is entirely their choice as to when and how often they wish to do so. Similarly if we imagine our agents are licensed taxis they cannot be making journeys all the time, as their drivers will need to rest at predictable periods of the day. Conversely such "dead time" may be unpredictable, for example a taxi may have go into the garage for unexpected repair work. Some actions can also be completed simultaneously, for example a trader can place 10 orders at once. In summary the problem that DAPRA effectively manages is that of irregularity i.e. irregularity between start and finish times of an agent's actions, as well as irregularity in the number and frequency of actions carried out by agents in a time series.

We set out to develop DAPRA to help generate regular time series databases from irregularly-sampled data. Below we explain DAPRA's design and implementation[1] and later show how DAPRA data-restructuring enables the user to easily discern agent behaviours, creating regularised data that is simple to visualise using Business Intelligence tools such as Tableau, [6].

Firstly it is important to identify and collect the different streams of time series data that DAPRA will be applied to (Fig. 2). Essentially these are the irregular time series that describe agents' actions through time, in addition to some time series of "exogenous" variables that are deemed useful in helping to explain agent behaviour. For example, one of the case studies used in this paper - that of taxi journeys around NYC - comprises irregular time series data pertaining to the pick-up and drop-off time stamps of taxi journeys undertaken by individual taxis. The exogenous data stream in this case describes the weather in NYC at similar points in time to the taxi journey data stream. It is rational to assume that the weather will have some effect on the number and frequency of taxi journeys made. For example heavy rain is likely to generate greater demand for taxis and so an increased number of journeys should take place. In addition to the streams of irregular and exogenous time series, it is possible to derive an extra data stream by recording the outputs of simple calculations performed on the irregular or exogenous time series fields. For example, the time distance between the last drop-off and the next pick-up of a given taxi. Likewise, (as will be discussed later for this paper's second case study on client trades), one can derive extra data by summing the value of individual trades to produce a trader's overall profit and loss (PnL) and position in the market (i.e. long, short or flat). After data collection, Fig.2 illustrates that DAPRA follows a three-step process of:

1) **D**ata **A**ggregation, where data from one or more sources of irregularly sampled time series data are combined into a regular sampled time series. In particular, we are focusing on sequences of variable time length actions and trying to correlate these actions with other useful time series with the aim to explain the agents behaviour.
2) The aggregated stream is **P**artitioned into intuitive epoch types which allow us to re-sample the data into regular time intervals. The size of partitions has an impact on the effectiveness of reduction and we must take care in selecting an appropriate resolution to partition our time series into, this will be discuss further in section IV.
3) Finally this aggregated, partitioned stream of data is **R**educed where the aggregated data is grouped depending on the requirements of the application.

The following sections describe the two irregularly sampled time series datasets that were used as case studies for application of DAPRA. These are also summarised in Table I.

(a) *Case Study 1: Retail Foreign Exchange Broker Client Trades*
Since the proliferation of mobile computing, a plethora

[1]Example DAPRA code can be found at [5]

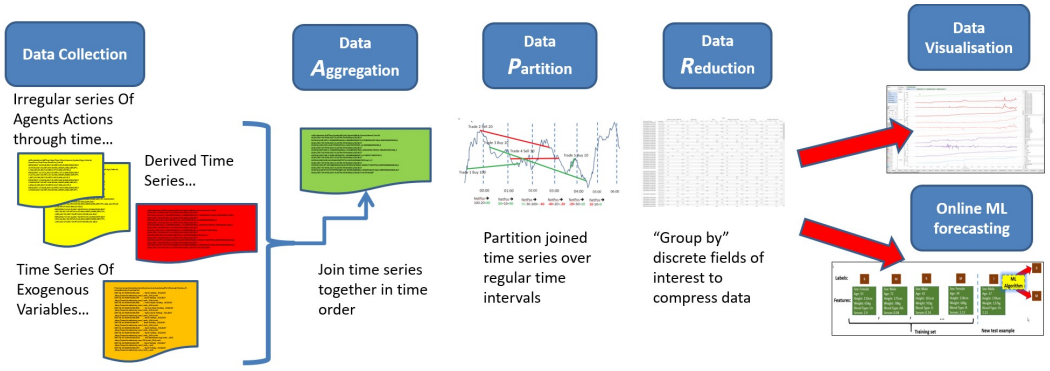

Fig. 2. Diagram of the main processes in the DAPRA framework

of retail Foreign Exchange (FX) brokerages have led the way in enabling retail investors (or clients) from around the world to speculate on the FX market. The FX broker is the 'middleman' – it links to the best liquidity providers (or LP's, such as investment banks), and the LP's stream 'trade-able' currency prices to the broker. The FX broker then passes these prices on to its thousands of clients worldwide all of whom can trade from their mobile phones or personal computers at the click of a button, investing as little as $100 or as much as $1,000,000 on each order. To open an account with the broker, retail clients must deposit initial funds and agree a leverage ratio before they are allowed to trade, typically anywhere between 20-100 to 1. Thus an initial deposit of $1000 will allow the clients to place orders and enter positions up to $100,000. Once an account is live the client is free to speculate trading as many different currency symbols (e.g. GBP, USD or EUR) as they wish provided they remain within the confines of their (leveraged) funds, in addition to any profit and loss (PnL) amounts. A typical retail broker will provide their clients with trading platform software such as MetaTrader 4 (MT4), [7]. Clients use the trading platform to place trades, monitor positions, track both historic and live movements in prices and access the latest world economic news which influence and drive volatility in the markets. In the publicly available dataset [8] we apply the DAPRA framework to analyse the trades of 684 clients during January 2017. The client trades are in MT4 format and relate to 30 of the most liquid FX symbols. Table I outlines the main features of two data streams pertaining to our case study 1: *Client Trades* (the irregularly-sampled time series) and *Prices* (the 'exogenous' dataset in this case).

(b) *Case Study 2: New York City Taxi Rides*
Our second case study describes data pertaining to taxi journeys in NYC. The data was made available by C. Whong, using freedom of information laws to obtain the data from the NYC Taxi and Limousine Commission (TLC), [9].

The TLC provided data in the form of two separate data streams comprising information about individual taxi rides and taxi fare data throughout 2013 (*NYC Taxi Rides* and *NYC Taxi Fares* - see Table I). We used these data streams as the the irregularly-sampled time series data for this case study. For practical reasons we also chose to extract data from January 2013 only. The exogenous dataset for the second case study refers to the weather conditions in NYC over the same time period. This data stream was published by D. Beniaguev on Kaggle [10] and provides an hourly snapshot of the weather conditions featuring the weather classification (such as "Heavy Rain" or "Clear Skies") as well as the temperature and humidity.

III. Data Aggregation

As previously discussed, the aggregation step of DAPRA involves merging the different data streams into one large dataset that effectively helps to explain or describe the actions of agents. Table I shows some key features of each case study's data streams that are aggregated and the derived fields that result. Differences in the number of fields and rows for each case study data stream illustrate clear differences in the irregularity of data sampled over the same time horizons. For example, the regularly sampled Market Prices data has 937,830 rows whereas the irregularly sampled Client Trades data has only 177,132 rows. This difference is what allows us to gain further insight into the trading behaviour of a client. Furthermore, Table I shows how many newly derived fields result from simple computations performed on aggregated data streams. For our first case study we can use aggregated Client Trades and Market Prices data to generate and track derived fields such as PnL and position at varying levels of granularity.

It is important to note the derived fields PnL and position are normalised to a common currency, in this case USD. This allows for meaningful comparison between client trades and analysis of the brokers position. When a client enters a trade their position is evaluated in the base currency of the symbol

they traded. We can therefore use the exchange rate between the base currency and USD (base multiplier) to calculate the position in USD. Likewise, PnL is quoted in the contra currency we use a similar method utilising the exchange rate between the contra currency and USD (contra multiplier) to calculate the PnL in USD.

TABLE I
CASE STUDY DATA TYPES

Case Study	Type of Data	Number of Fields	Number of Rows	Description of Data	Derived Field
Case Study 1: Trade Data	Client Trades	9	177,132	Detailed profile of each client order.	Profit and Loss (PnL)
	Market Prices	5	937,830	1 minute price data sampled for 30 major currency pairs.	Position
Case Study 2: Taxi Data	NYC Taxi Rides	9	14.7 mio	Detail of each taxi ride, with pickup and dropoff location.	Wait Time
	NYC Taxi Fares	4	14.7 mio	Fare prices for corresponding trip data	Location Latitude Location Longitude
	NYC Weather	1	8,760	Weather type sampled each hour.	Speed

IV. DATA PARTITION

The partition step serves to "bucket" the aggregated data into regular time epochs, allowing for effective analysis of each time epoch in detail. For this we must decide the partition size which will vary depending on the data / problem domain. The partition size is ultimately the choice of the end-user, however if it is too small there will not be any meaningful aggregation of the data. Similarly too large a partition size will not enable sufficient breakdown/analysis of the data.

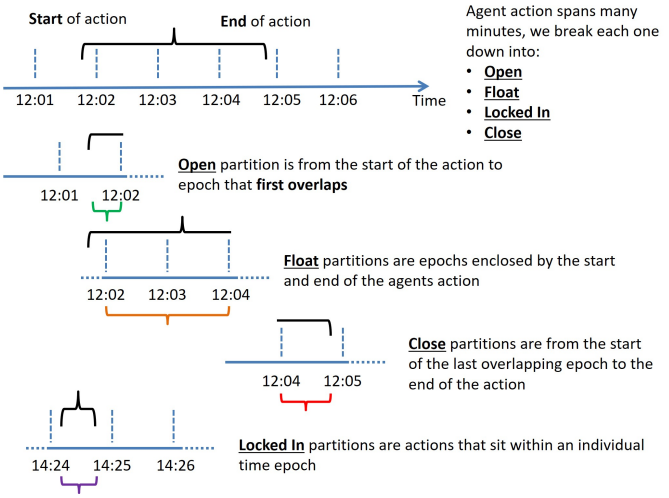

Fig. 3. DAPRA Partition Classification

There are four classifications of epoch partition which we can break each agent's actions into, we refer to these as *type parameters*, τ. Actions of a given agent can span many epochs depending on the resolution chosen and will affect the number of classification types in the resulting partitioned data, as illustrated in Fig.3.

The four type parameters can be defined as follows:

- **Open**: Actions opened in the current epoch.
- **Closed**: Actions closed in the current epoch.
- **Locked**: Actions opened and closed in the current epoch.
- **Float**: Actions that have been opened in a previous epoch and are still open at the end on the current epoch.

Over an interval from t to $t + \Delta$ where Δ is the size of the resolution. Let OT and CT be the open and close times of an agents' action. We assign the type parameter, τ, using a method similar to Allen's interval algebra [11]. We can represent this logic in pseudo code as follows:

```
If OT < t:
    If CT ⩾ t + Δ:
        τ = float
    Else:
        τ = close
Else:
    If CT ⩾ t + Δ:
        τ = open
    Else:
        τ = locked
```

As Table II shows, the choice of epoch resolution intuitively influences how each action breaks down into the different type classifications. For example if the resolution is too small most of the actions will be classified as "floating", and as the resolution increases, an increasing number of actions are classified as "locked in".

TABLE II
NUMBER OF EPOCH TYPES AT VARIOUS RESOLUTIONS

Case Study	Resolution	Open	Closed	Locked	Float
Case Study 1: Trade Data	5 days	42 k	42 k	135 k	29 k
	2.5 days	59 k	59 k	118 k	108 k
	1 day	70 k	70 k	107 k	330 k
	60 min	93 k	93 k	84 k	9.4 mio
	5 min	104 k	104 k	73 k	113 mio
Case Study 2: Taxi Data	1 day	0.1 mio	0.1 mio	14.3 mio	53
	60 min	2.7 mio	2.7 mio	11.7 mio	8 k
	15 min	9.2 mio	9.2 mio	5.2 mio	1.8 mio
	10 min	11.4 mio	11.4 mio	3.0 mio	5.1 mio
	5 min	13.6 mio	13.6 mio	0.8 mio	19.4 mio

We can also see how the choice of resolution depends on the problem domain for example, in the case of trade data a resolution of 2-2.5 days provides an optimal ratio between the type parameters. Whereas for the taxi data an optimal ratio is found at a resolution of around 10 minutes. This intuitively makes sense as taxi rides are shorter on average that trade holding periods. It is also noteworthy that our tolerance for the number of "locked" classifications varies from one problem domain to the next. For example, short trades under 10 minutes are fairly typical, whereas taxi rides under 1 minute are extremely scarce. Therefore, variance in the distribution of the length of agent actions is a key factor when deciding a sensible resolution to partition the data.

V. Data Reduction

Once the staging process is complete, the resulting target data is stored in an OLAP cube. This is depicted in Fig.4 where the Market Prices and Client Trades data are transformed into an OLAP cube with time, symbols and clients as its dimensions.

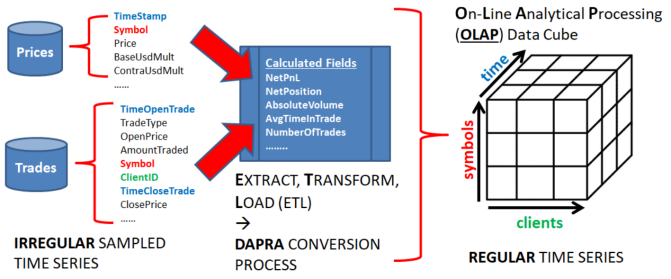

Fig. 4. DAPRA storage in an OLAP cube

The hierarchical dimensions of the OLAP cube allow for straight forward operations to be executed with relative ease, allowing for effective reduction of the data [12]. This step involves focusing on a specific discrete field that we use to regroup the data in a meaningful way. For example, for the taxi case study, we may want to break down the data by location of where pickups are taking place. Despite longitude and latitude being real numbers we can bucket them into rough locations rounding to 2 decimal places. We take the partitioned data and then group by our relevant statistics that we want to compute such as fare amount, distance travelled or weather type. The choice of which discrete field we want to group by and the resolution determines how many rows DAPRA finally compresses down to.

TABLE III
NUMBER OF ROWS IN TARGET DATA USING DIFFERENT GROUPINGS

Case Study 1: Trade Data			
Time Resolution	Group by client + symbol (20,520 Unique)	Group by client (684 Unique)	Group by symbol (30 Unique)
5 days	5,0635	4,264	210
2.5 days	99,661	8,799	411
1 day	217,539	20,267	934
60 min	4,780,691	465,308	21,605
5 min	57,095,054	5,571,898	258,778
Case Study 2: Taxi Data			
Time Resolution	Group by Location to 2 decimal places (3k unique)	Group by taxi license plate (32.5k unique)	Group by weather type (14 unique)
1 day	18,760	714,959	185
60 min	148,861	5,968,452	2,639
15 min	455,175	18,303,614	10,312
10 min	647,205	25,350,520	15,433
5 min	1,210,543	44,095,732	30,793

Table III shows the effectiveness of the data reduction step at various resolutions and with 3 choices of grouping for each case study. For example, grouping by location in the case of taxi journeys enables us analyse locations as zones rather than individual points which offers a more practical overview. In addition (and assuming a 1:1 relationship between taxi driver and taxi licence plate), grouping taxi journeys by license plate allows us to build a profile for each driver offering insights into their behaviour.

VI. Case Study 1: Foreign Exchange Broker Data

We will now define the fields within each of the datasets for the foreign exchange broker data as described in Table I.

The *Client Trades* data stream has nine fields defined as follows:

- OT: Open time is the time stamp at the open time of the order.
- ID: Is the unique identification number[2] of the client.
- TA: This is the total amount traded in the order, quoted in the base currency.
- SD: This is the side referring to the position the broker takes in the traded, 1 for long and -1 for short.
- SY: Is the symbol traded.
- ON: Is the number assigned to the order.
- OP: Open price is the price of the symbol at the time of the order.
- CT: Close time is the time stamp of the close time of the order.
- CP: Close price is the price of the symbol at the close time of the order.

An example of this from the source data is shown in table IV.

TABLE IV
EXAMPLE OF SOURCE CLIENT TRADE DATA

Open Time	Client	Amount	Side	Symbol	Order Id	Open Price	Close Time	Close Price
03/01/2017 01:01:33	B40	2000	-1	USD/CHF	B40_0	1.0223	03/01/2017 13:04:50	1.0308
03/01/2017 13:04:50	B40	3000	1	USD/CHF	B40_1	1.0308	11/01/2017 00:26:22	1.026
17/01/2017 06:43:26	B40	2000	-1	EUR/CHF	B40_27	1.0729	19/01/2017 00:31:22	1.0723

Next we have the *Market Prices* data stream, containing:

- TS: This is the time stamp at the time of the sample.
- SB: Is the symbol being quoted.
- BM: This is the base multiplier, which refers the the exchange rate from the base currency to USD, used to calculate position in USD.
- CM: Is the contra Multiplier, referring to rate between the contra currency and USD, used the calculate the profit and loss in USD.
- SP: Is the price of the symbol at the time of the observation.

Table V gives an example of the source market data.

[2]This is an encrypted identification number.

TABLE V
EXAMPLE OF SOURCE MARKET PRICE DATA

Datetime	Symbol	Open Usd Mult	Contra Usd Mult	Price
04/01/2017	USD/CHF	1	0.9745	1.0262
06/01/2017	USD/CHF	1	0.9903	1.0098
08/01/2017	USD/CHF	1	0.983	1.0173
10/01/2017	USD/CHF	1	0.9838	1.0164
12/01/2017	USD/CHF	1	0.9894	1.0107
16/01/2017	EUR/CHF	1.0627	0.9902	1.0734
18/01/2017	EUR/CHF	1.0632	0.993	1.0728
20/01/2017	EUR/CHF	1.07	0.9981	1.072

Now we will define the functions used to calculate the derived fields, PnL and position.

$$PnL_t = \begin{cases} (SP_t - OP) \times TA \times SD \times CM & \text{if } \tau = open \\ (SP_{t+1} - SP_t) \times TA \times SD \times CM & \text{if } \tau = float \\ (CP - SP_t) \times TA \times SD \times CM & \text{if } \tau = close \\ (CP - OP) \times TA \times SD \times CM & \text{if } \tau = locked \end{cases}$$

$$POS_t = TA \times SD \times BM$$

It is important to note that the derived fields describe the PnL and position from the perspective of the broker. Table VI shows an example of data populating the derived fields following application of DAPRA to the data in Tables IV and V.

TABLE VI
EXAMPLE OF TARGET DATA AFTER THE APPLICATION OF DAPRA FOR CASE STUDY 1

Order ID	client_ID	symbol	Type	Open	Close	PnL USD	Position USD	Side
B40_0	B40	USD/CHF	LOCKED	02/01/2017	04/01/2017	-16.559	-2000	-1
B40_1	B40	USD/CHF	OPEN	02/01/2017	04/01/2017	-13.532	3000	1
B40_1	B40	USD/CHF	FLOAT	04/01/2017	06/01/2017	-48.678	3000	1
B40_1	B40	USD/CHF	FLOAT	06/01/2017	08/01/2017	23.273	3000	1
B40_1	B40	USD/CHF	FLOAT	08/01/2017	10/01/2017	-9.4088	3000	1
B40_1	B40	USD/CHF	CLOSE	10/01/2017	12/01/2017	35.296	3000	1
B40_27	B40	EUR/CHF	OPEN	16/01/2017	18/01/2017	0.1327	-2139.23001	-1
B40_27	B40	EUR/CHF	CLOSE	18/01/2017	20/01/2017	0.9069	-2131.74999	-1

VII. CASE STUDY 2: NYC TAXI JOURNEY DATA

In this section we will define the fields of the second case study as outlined in Table I .

The *NYC Taxi Ride* data stream has nine fields as described below.

- ID: This is the encrypted license plate for each taxi.
- OT: This is the pick-up time of the trip.
- CT: Is the drop off time of the trip.
- TT: Is the total time of the trip.
- TD: Refers to the total distance of the trip.
- PA: The pick up longitude.
- PL: The pick up latitude.
- DA: The drop off longitude.
- DL: The drop off latitude.

The *NYC Taxi Fares* data stream comprises five fields:

- ID: This is the encrypted license plate for each taxi.

TABLE VII
EXAMPLE OF SOURCE NYC TAXI RIDE DATA

Encrypted license	Pickup datetime	Dropoff datetime	Trip Time in secs	Trip distance	Pickup longitude	Pickup latitude	Dropoff longitude	Dropoff latitude
ED7D81135978A1933D109A6434A7C563	13/01/2013 11:48	13/01/2013 12:10	1320	9.94	-73.98394	40.760399	-73.872818	40.77441
52830E3E5933F38822EF7A73A17C98C2	13/01/2013 12:05	13/01/2013 12:09	240	0.92	-73.983047	40.722694	-73.991402	40.729877
CFF79EB7D13568950B06C4DBC3329A8E	13/01/2013 12:01	13/01/2013 12:16	900	3.7	-73.973038	40.785358	-73.982933	40.745277

- OT: This is the pick-up time of the trip.
- FA: This is the fare amount of each trip.
- TP: Is the tip amount giver to the driver.
- TO: This is the amount the passenger paid as a result of tolls encountered during the journey.

TABLE VIII
EXAMPLE OF SOURCE NYC TAXI FARE DATA

Encrypted license	Pickup datetime	Fare amount	Tip amount	Tolls amount
ED7D81135978A1933D109A6434A7C563	13/01/2013 11:48	30	0	0
52830E3E5933F38822EF7A73A17C98C2	13/01/2013 12:05	5	0	0
CFF79EB7D13568950B06C4DBC3329A8E	13/01/2013 12:01	14	1	0

Next, the exogenous *NYC Weather* data:

- TS: The Time-Stamp of the time the sample was taken.
- WL: This refers the weather conditions at the time of the observation as a Weather Label.

TABLE IX
EXAMPLE OF SOURCE WEATHER DATA

Datetime	Weather description
13/01/2013 10:00	mist
13/01/2013 11:00	overcast clouds
13/01/2013 12:00	mist
13/01/2013 13:00	mist

Following DAPRA we can derive four new fields: (1) wait time, (2) interpolated pick-up / drop off latitude, (3) interpolated pick-up / drop off longitude and (4) average speed for each taxi journey. These can be defined as:

$$WaitTime_t = OT_{t+1} - CT_t$$

$$Speed_t = \frac{TD_t}{CT_t - OT_t}$$

$$Loclat = PL + \mathcal{P} \times (DL - PL)$$

$$Loclong = PA \times \frac{1 - (Loclat - PL)}{DL - PL} + DA \times \frac{Loclat - PL}{DL - PL}$$

where $\mathcal{P}$ = the fraction of the journey complete.

TABLE X
EXAMPLE OF TARGET DATA AFTER APPLICATION OF DAPRA FOR CASE STUDY 2

Encrypted license	Type	Open	Close	Fare	Duration	Distance	Location longitude	Location latitude	Weather	Wait Time	Avg.Speed
ED7D8113 5978A193 3D109A64 34A7C563	OPEN	13/01/2013 11:40	13/01/2013 11:50	2.727273	120	0.903636	-73.9839	40.7604	overcast clouds	01:00:00	27.10909
ED7D8113 5978A193 3D109A64 34A7C563	FLOAT	13/01/2013 11:50	13/01/2013 12:00	13.63636	600	4.518182	-73.9233	40.76804	overcast clouds	01:00:00	27.10909
ED7D8113 5978A193 3D109A64 34A7C563	CLOSE	13/01/2013 12:00	13/01/2013 12:10	13.63636	600	4.518182	-73.8728	40.77441	overcast clouds	01:00:00	27.10909
52830E3E5 933F38822 EF7A73A1 7C98C2	LOCKED	13/01/2013 12:00	13/01/2013 12:10	5	240	0.92	-73.9872	40.72629	mist	02:10:00	13.8
CFF79EB7 D1356895 0B06C4DB C3329A8E	OPEN	13/01/2013 12:00	13/01/2013 12:10	8.4	540	2.22	-73.973	40.78536	mist	00:30:00	14.8
CFF79EB7 D1356895 0B06C4DB C3329A8E	CLOSE	13/01/2013 12:10	13/01/2013 12:20	5.6	360	1.48	-73.9829	40.74528	mist	00:30:00	14.8

VIII. DISCUSSION OF FINDINGS USING DAPRA

Here we present and discuss a fraction of interesting findings following application of DAPRA to the aforementioned case study datasets. We have purposefully focused on findings that have practical and commercial implications for end users in the financial and transport industries and have visually presented these using Tableau, [6].

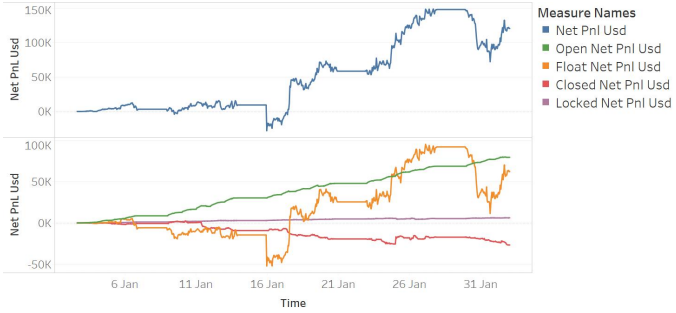

Fig. 5. **TOP:** Broker's Net PnL. **BOTTOM:** Broker's Net PnL grouped by the type parameter

The aim of our first case study was to gain insight into client trading behaviour. Using DAPRA we analysed and compared the trades carried out by 684 clients during January 2017. Each client was allowed to buy or sell any of the 30 available unique currency pairs (e.g. EUR/USD) and they could place trades as many times as they wanted, at anytime of day provided they stayed within the confines of their leveraged funds. During this time the clients' broker accumulates a position in the currency market which is essentially an aggregation of all of its client trades with the LP's. The broker position could be *long* (meaning clients are placing more sell trades), *short* (clients are placing more buy trades) or *flat* (the amount of sell and buy trades by clients are equal). Application of DAPRA enables the broker to visualise how its market position accumulates over time and then answer specific questions about it, as well as predict its position in the future.

Fig.5 shows the broker's Net PnL in USD over the month of January with an interval time of one hour. The application of DAPRA allows the broker to track and analyse their PnL at a high resolution not possible from the source data alone. This strategic information allows the broker to examine the volatility and draw-down of their PnL, which in turn helps to inform hedging strategies. Fig.5 also shows the Net PnL of the broker grouped using the type parameter. This provides an interesting insight into client behaviour: we notice the closed Pnl (in green) is negative, meaning the broker lost on the segment of trades labelled as closed. This pattern is repeated at various interval lengths, suggesting that in general clients tend to wait for positions to rise in value before closing them irrespective of the overall performance of the trade. We also see that the PnL of floating positions far outweighs that of locked positions, suggesting that client trading strategies have a greater potential to be profitable over a longer holding period.

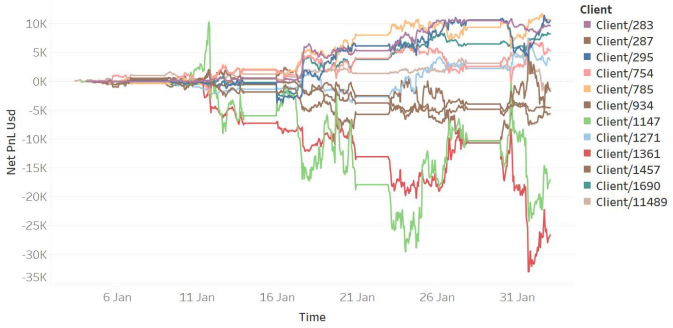

Fig. 6. Broker's Net PnL grouped by client

The storage of the data using an OLAP cube combined with DAPRA staging allows for dynamic insight into individual client trading behaviour and characteristics which were previously unseen. Following DAPRA, Fig.6 shows the net PnL of each client (again from the broker side), for clients with a final PnL of at least 6,000 USD. This allows the broker to identify clients that pose the greatest risk and those from whom they can profit. For example, clients 1147 and 1362 contribute significantly larger losses to the broker than average, thus the broker may decide to use a specific hedging strategy for all orders placed by these clients. This drilling-down of the data also offers the broker a comprehensive view of all trades made by their clients, from their open to their close. By pre-processing the data with DAPRA, we can now examine the variation in floating (or unrealised PnL) across the lifespan of each client trade and see at which point the client decided to close the trade and realise their PnL. This behind-the-scenes intuition into client decision-making is an invaluable resource for the broker that

helps identify and differentiate successful and unsuccessful trading strategies.

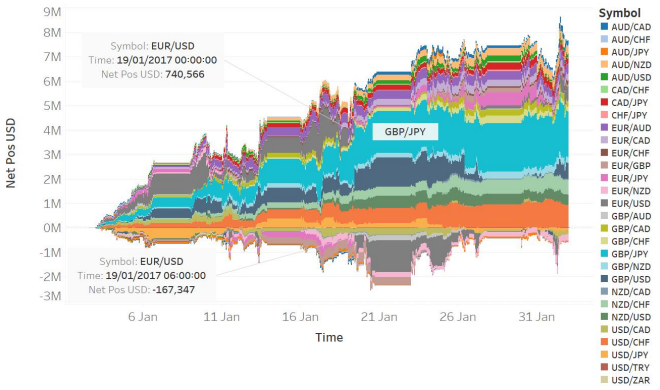

Fig. 7. Broker's Net position grouped by symbol

Equally important is the broker's position - as previously discussed DAPRA can be used to calculate the position of all symbols in one common currency (USD). In Fig.7 we observe the broker's net position over the sample period, as grouped by symbol. A broker would undoubtedly find this information invaluable to ensure that it limits its exposure to unexpected changes in a given currency market. Such information becomes even more meaningful when it is continually re-evaluated according to the current exchange rate and at an interval length specified by the broker. This is all made possible with DAPRA. For example, net position grouped by GBP/JPY (in blue in Fig.7) shows a steady accumulation over time which would likely compel the broker to flatten its position (to reduce its market exposure) by hedging against it. In another example, net position grouped by EUR/USD (in grey) shows a dramatic switch from long to short during the first 6 hours of 19th Jan 2017. This demonstrates that during this time, the clients' combined view of the market moved to predict a rise in the value of EUR/USD. This information would be of great interest to the broker and when combined with individual client PnL would offer considerable insight into client views of the market and what may drive them to open and close their positions.

Fig. 8 shows application of DAPRA to data of the NYC taxi journey case study where 10 minute intervals have been used to partition the data, further grouped by the type parameter. Grouping of the partitioned data reveals some interesting aspects regarding taxi journeys which can be visualised by plotting the interpolated location over a map of NYC. For instance, we can see what appears as a small cluster in the bottom right corner of the open (green) and closed (blue) maps. This cluster maps to JFK airport and as the parameters of this cluster are present in open and closed types we would identify these locations to be airport pick-up / drop off points (note this finding supports those of a 2018 study in [13]). This is to be expected however the versatility of DAPRA can be seen when we analyse the floating data groupings. This reveals a collection of location points from the island of Manhattan to JFK airport. This data not only provides information on the taxis travelling to and from the airport, but also classifies a section of land between Manhattan and JFK airport where a driver would have little chance of securing a pickup. Finally, it is apparent that the intensity of locked-in trips is highest in Manhattan. Intuitively this shows that taxi rides under 10 minutes are more common in built up areas than the surrounding suburbs.

Fig. 9 shows the results of grouping 10-minute partitioned taxi journey data with weather classifications. The groupings are as follows: average waiting around time, average tip amount, average tolls amount and average speed. Here we gain an intuitive insight into how taxi drivers and passengers react to changing weather conditions. Perhaps the most striking result is the correlation between the weather label and the average wait time from the last drop-off to the next pick-up. This grouping clearly shows taxi drivers experience far longer wait times between pickup's during dry weather ("clear sky") than wet weather ("wet"). This is probably because passengers are less inclined to require a taxi when the weather is good, yet demand for taxis increases when passengers want to avoid getting wet and so waiting time between jobs is clearly reduced. An equally intuitive reason for shorter waiting times is that drivers reduce their speed in bad weather, this is reflected in the DAPRA data showing a significantly lower speed of 27kph in wet weather.

Continuing with Fig.9, the DAPRA framework was used to analyse data about the tolls paid by passengers with weather condition data. The findings revealed a less intuitive story, suggesting that in foggy and misty conditions, taxi drivers are more likely to choose routes with tolls in place. This is accompanied well by the data on passenger tips which reveal an interesting finding. When the weather is poor (wet or foggy/misty) drivers can expect to receive higher tips, despite passengers being charged extra for tolls in foggy weather. However, in good weather when tips are generally lower, the choice of a driver to impose a toll charge on a passenger has a detrimental effect on their likelihood of receiving a tip. This is demonstrated by the correlations between average toll amount / average tip amounts with the broken clouds and clear skies weather groupings.

Finally, we used DAPRA to examine patterns of taxi journeys and their features (e.g. duration, fare paid) of individual taxi drivers across January 2013. These findings are illustrated in Fig.10. Here we grouped the data by taxi ID and used a bubble diagram to clearly visualise an obvious working pattern throughout time for each driver. The size of each bubble represents the distance of each trip and the vertical axis shows the computed fare of the whole trip. This concise representation of each

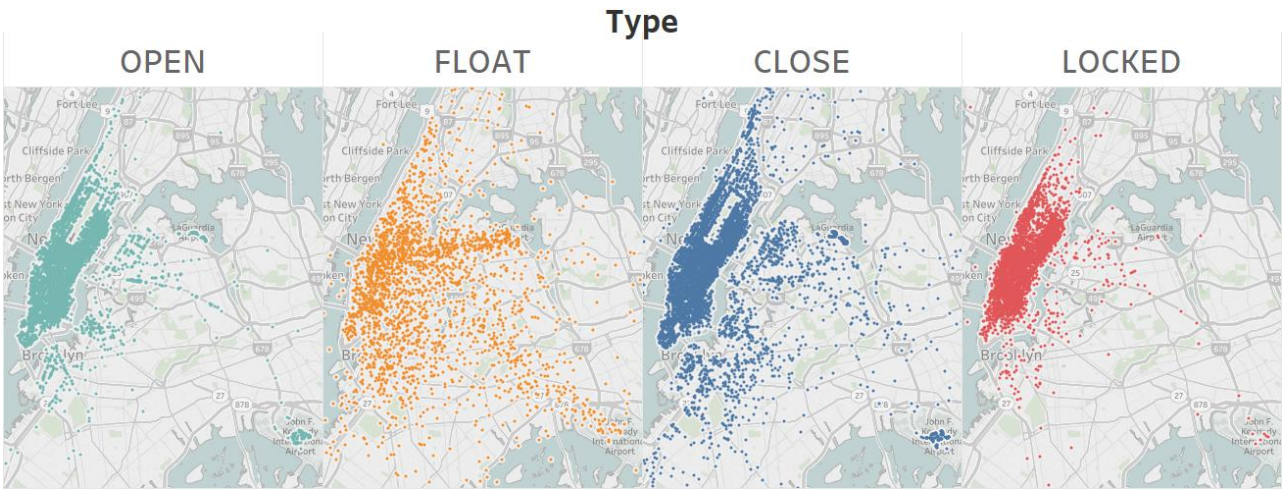

Fig. 8. Visualisation of taxi journeys in NYC during January 2013 for each of the four type parameters

driver's week provides valuable insights, not only can we see when a driver was active but we can also quickly assess the success of each day. For example, we can see driver "0A0B39F7A97A6CFAA62F11C4BDA6BBF8" consistently works a six day week whereas, driver "0A1A0478120C8A7B0C035A06321D3B91" rarely works more that two days a week. We can also see which days are good days for drivers. For example Monday 21^{st} of January was a slow day for driver "0A0B39F7A97A6CFAA62F11C4BDA6BBF8", differing from their usual busy days (described by dense concentration of circles). On the other hand, driver "0A2A8EB88CDB6F1287C1DC86309DC047" had a fare of over $200 on Friday 11^{th} of January, which from looking at the size of the circle is clearly a result of an unusually long trip distance.

IX. Conclusion

Here we presented DAPRA - Data Aggregation Partition Reduction Algorithm - to manage, examine and uncover insights into the time-sequenced actions of agents in an irregular time series. DAPRA originated from our need to find a framework capable of regularising data into a format more amenable to visual analysis and machine learning techniques. We used two case studies to demonstrate the practical application of DAPRA.

Our first case study applied DAPRA in an attempt to better understand the factors influencing client trading behaviours and the downstream effects of these on the broker's PnL and position. We demonstrated how using DAPRA enabled us to progress from the painstaking task of wading through millions of irregularly-sampled client trades, to the generation of a detailed and meaningful data stream that when visualised provided commercially-valuable information regarding client / broker positions and PnL over time. DAPRA requires some

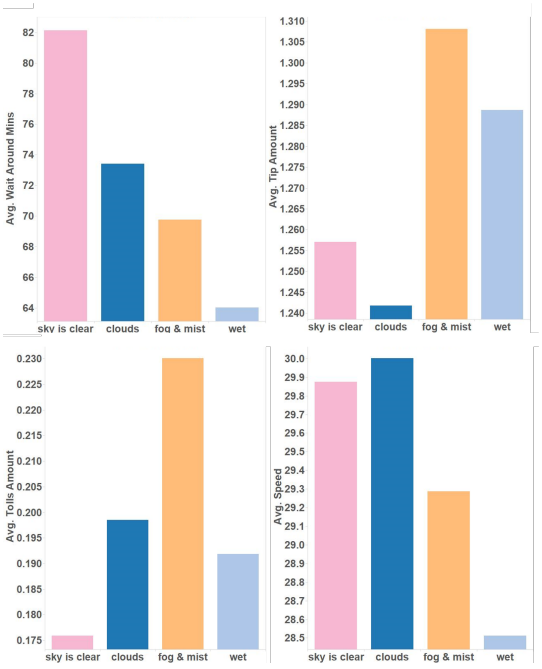

Fig. 9. Grouping taxi journey data by weather

exogenous data stream and in this study we referred to a simple one that described currency prices over the same time period as the client trades. However it is important to note that many more exogenous data streams could have been aggregated with. For example, one major driver of intra-day pricing volatility in the FX market is the economic release calendar [15]. This calendar dataset is known ahead of time and is essentially a schedule outlining when the worlds major economies will publish their latest macro-economic data. Sometimes estimated figures for economic announcements are provided in the calendar too. For example on the first Friday of every month most traders will await the release of so called "Non-Farm Payrolls" (average earnings) which act as a major

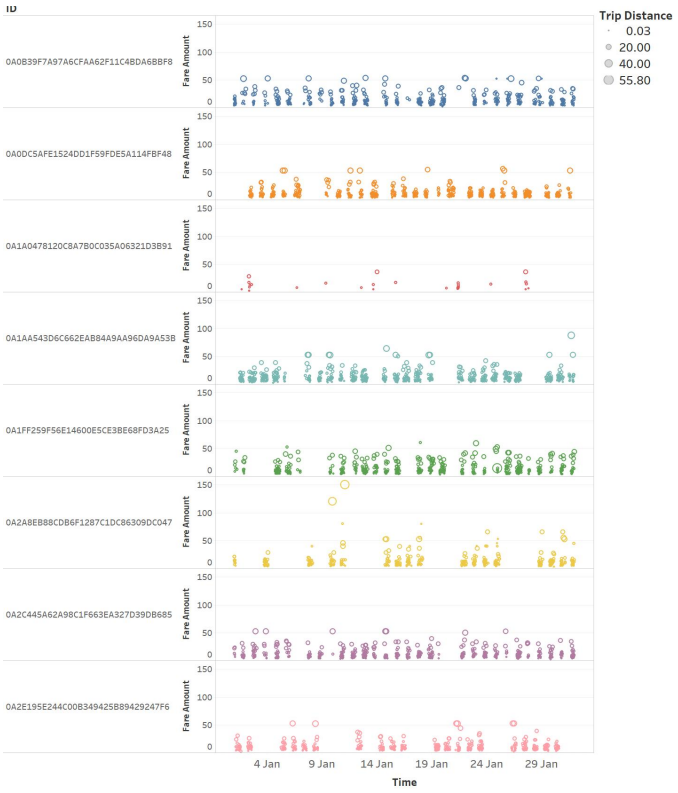

Fig. 10. Visualising patterns of taxi journey information for individual taxi drivers

indicator of the strength of the US economy. The authors have investigated in detail the use of economic release data to provide deeper understanding of the underlying market conditions through time, adding features such as counting how many high importance USD and EUR economic releases happened in the previous and next hour. Given such data we could identify clients that trade in and around specific economic releases and separate these from clients that are trading based on longer term price movements. This would provide even further insights into client trading behaviour.

In our second case study we looked at NYC taxi journeys during January 2013 and attempted to correlate this activity with a weather dataset. Intuitively, a taxi company will want to know the best approaches to allocate its fleet of taxis to avoid too much downtime, meet demand and thus maximise profitability. Even on an an individual level, a taxi driver may want to leverage such information for example to know where best to drive to given specific weather conditions in order to maximise their chances of getting follow on fares in the next hour. It should be noted that we did make a crude assumption to the taxi journey dataset so that it would fit within the DAPRA framework. This assumption was to interpolate the location through time linearly between pickup and drop-off. One improvement to this would be to use a more realistic path through the city using actual road map data using algorithms such as Googles Poly Lines service [16], or better still in the future have full telemetary samples of each taxi ride throughout the journey such as the data used in [17]. It is even conceivable that with smartphone app-based taxi hailing / booking services such as Uber [18] we also could collect and use encoded customer passenger data and link this with taxi ID and journey data. This would open up another rich set of agents that we could cross-correlate behaviours with, perhaps even recording which journeys the taxi drivers had to pickup and which at any given time they chose to carry out.

One area for future work relates to the choice of optimal time resolution during the partition step of DAPRA. This choice has to consider: (1) the nature of the raw data (low/high frequency, origin and size) and (2) the type of analysis we intend to carry out, be it forecasting, visualization or for guiding some actions. In each of our case studies we decided on arbitrary equal width time intervals, indeed other studies such as [19] and [20] look at methods of using variable width intervals that increase sampling during more active regions in the time series. In other work such as [21], [22], [23], the bin intervals are chosen using an iterative process that results in an interval with the lowest forecast error. Future implementations of DAPRA could look to use such techniques of variable width time intervals to see if better performance can be achieved.

In future work it is our aim to take DAPRA-structured datasets and use them for online machine learning tasks. The beauty of the DAPRA approach is that it lends itself perfectly to training classical online machine learning techniques [24]. Classification or regression labels can be easily constructed using features of future DAPRA data instances and detailed features can be derived from the Open/Float/Locked In/Closed formulation of the statistics. For example, in our recent work we have used DAPRA in portfolio optimisation models. This is achieved by grouping orders by client and treating each client as an “expert” that makes a prediction of the future state of the market. Furthermore we are currently researching the use of online machine learning techniques such as the aggregating algorithm [14] - early results show promising potential to outperform the market.

By referring to time series data from two distinct and complex domains we have tried to present DAPRA as a universal framework for all sorts of time series data streams. Time horizons can be of any size provided agent actions with defined start- and stop-times can be identified and agents are free to perform their actions as and when they choose to. Furthermore, agents do not need to relate to human-based behaviours such as trading or taxi driving. Indeed in other work we have analysed network traffic data, where the agents are IP addresses of different machines on a computer network and the duration of sending data is measured in nanoseconds. In this era of Big Data, there is certainly no shortage of data streams that DAPRA can be applied to both now and

in the future. As our world becomes increasingly connected with smart devices in homes and cities, the "sea" of tracked data measurements that results remains ever-expanding. It is an exciting time to find out how, using frameworks such as DAPRA, we will use insights into agent behaviour to our own benefit.

ACKNOWLEDGMENTS

The authors acknowledge the support of Algorithmic Laboratories Ltd (AlgoLabs) and their parent group Equiti Global Markets in establishing and developing this research. Special thanks go to Xudong Li, Tzyy Tong and Samuel Manoharan for setting up the servers necessary to run our experiments.

REFERENCES

[1] Lauría, E.J., Greco, C.A, "OLAP for Financial Analysis and Planning - A Proof of Concept", *ICSOFT*, 2010.

[2] Rasmussen, N., Goldy, P. S. and Solli, P. O, "Financial Business Intelligence : Trends, Technology, Software Selection, and Implementation", *New York : NY, John Wiley and Sons*, 2002.

[3] J. Arroyo and C. Mate, "Introducing Interval Time Series: Accuracy Measures", *Proceedings in Computational Statistics*, pp. 1139-1146, 2006.

[4] P. Rodrigues and N. Salish, "Modeling and Forecasting Interval Time Series with Threshold Models: An Application to SP500 Index Returns", *Banco de Portugal Economics and Research Department Working Papers*, vol. 28, Oct. 2011.

[5] W. Wisniewski, *DAPRA*, GitHub repository, 2019. [Online]. Available: https://github.com/Wisniewskiw/DAPRA [Accessed on 19 November 2019.]

[6] *Tableau*, Seattle WA, Tableau Software. 2019. [Online]. Available: https://www.tableau.com

[7] *MetaTrader 4 Trading Platform*, Metaquotes Software Corp. 2019. [Online]. Available: http://www.metaquotes.net/en/metatrader4

[8] D. Lindsay, *FXClientTrades*, London, UK, Kaggle, 2019. [Online]. Available: https://www.kaggle.com/ davidlindsay1979/ toptradingclient-data/ kernels [Accessed on 28 August 2019.]

[9] C. Whong, "FOILing NYCs Taxi Trip Data", *Chris Whong*. 2014. [Online]. Available: https://chriswhong.com/open-data/ [Accessed: 1 August 2019].

[10] D. Beniaguev, *Historical Hourly Weather Data 2012-2017*, Kaggle, Dec. 2017. [Online]. Available: https://www.kaggle.com/selfishgene/historical-hourly-weather-data/version/2 [Accessed: 1 August 2019].

[11] Allen, J. F, "Maintaining Knowledge About Temporal Intervals", *Communications of the ACM*, 26 (11), 832–843,1983.

[12] Gómez, Leticia I. and Gómez, Silvia A. and Vaisman, Alejandro A, "A Generic Data Model and Query Language for Spatiotemporal OLAP Cube Analysis", *Proceedings of the 15th International Conference on Extending Database Technology*, ACM, 300–311, 2012.

[13] T. Schneider. "Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance", *Todd W. Schneider*. 2018. [Online]. Available: https://toddwschneider.com/posts/analyzing-1-1-billion-nyc-taxi-and-uber-trips-with-a-vengeance/ [Accessed 28 July 2019].

[14] Y. Kalnishkan. and M. Vyugin. M, "The weak aggregating algorithm and weak mixability", *Journal of Computer and System Sciences*, vol. 74, no. 8, pp.1228-1244, 2018.

[15] Forex Factory, *Forex Calendar*, Fair Economy Inc, 2019. [Online]. Available: https://www.forexfactory.com/calendar.php [Accessed: 1 August 2019].

[16] Google Developers, "Geometry Library", 2019. [Online]. Available: https://developers.google.com/maps/documentation/java script/geometry [Accessed 15 August 2019].

[17] Leonardi. L, Orlando. S, Raffaet. A, et al, "A general framework for trajectory data warehousing and visual OLAP". *Geoinformatica*, 18(2) 273312, 2014, [Online] Available: https://doi.org/10.1007/s10707-013-0181-3

[18] *Uber*, Uber Technologies Inc. 2019. [Online]. Available: https://www.uber.com

[19] G.P. Nason, B. Powell, D. Elliott and P.A. Smith, "Should we Sample a Time Series More Frequently? : Decision Support via Multirate Spectrum Estimation", *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 180, no.2, pp.353407, 2017.

[20] G. Xuedong and C. Hailan, "The Method of Time Granularity Determination on Time SeriesBased on Structural Similarity Measure Algorithm", *Proceedings of the International Symposium on the Analytic Hierarchy Process: the 14th ISAHP Conference*, London, 2016. [Online] Available: https://doi.org/10.13033/isahp.y2016.117

[21] R. Al-Hmouz, W. Pedrycz, A. Balamash and A. Morfeq, "Granular representation schemes of time series: A study in an optimal allocation of information granularity", *IEEE Symposium on Foundations of Computational Intelligence (FOCI)*, Singapore, 2013, pp. 44-51.

[22] R. J. Hyndmana, N .Kourentzesb, F. Petropoulosc, G. Athanasopoulosa, "Forecasting with Temporal Hierarchies", *European Journal of Operational Research*, vol. 262, pp. 60-74, 2017.

[23] X. Wu, B. Shi, Y. Dong, C. Huang, L. Faust, and N.V Chawla. "REST-Ful: Resolution-Aware Forecasting of Behavioral Time Series Data", *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* New York, pp. 1073-1082. 2018.

[24] L. Bottou, "On-line Learning and Stochastic Approximations", in *On-Line Learning in Neural Networks*, D. Saad, Ed. Cambridge: Cambridge University Press, pp. 942, 1999.