# DATA ANALYSIS AS A WEB SERVICE: A CASE STUDY USING IOT SENSOR DATA

*Alireza Ahrabian*[1], *Sefki Kolozali*[1], *Shirin Enshaeifar*[1], *Clive Cheong-Took*[2] *and Payam Barnaghi*[1]

[1]University of Surrey, Electrical and Electronic Engineering, Institute For Communication Systems
[2]University of Surrey, Department of Computer Science
Email: {a.ahrabian, s.kolozali, s.enshaeifar, c.cheongtook, p.barnaghi}@surrey.ac.uk

## ABSTRACT

The advent of Internet of Things, has resulted in the development of infrastructure for capturing and storing data from domains ranging from smart devices (e.g. smartphones) to smart cities. This data is often available publicly and has enabled a wider range of data consumers to utilise such data sets for applications ranging from scientific experimentation to enhancing commercial activity for businesses. Accordingly this has resulted in the need for the development data analysis tools that are both simple to use and provide the most effective tools for a given data set. To this end, we introduce data analysis tools as web service, that enables the data consumer to make a simple HTTP request for processing data over the internet. By providing such tools as a web service, we demonstrate the potential of such a system to aid both the advanced and novice data consumer. Furthermore, this work provides an use case example of the proposed tool on publicly available data extracted from the smart city CityPulse IoT project.

***Index Terms*****—** Data Analysis web service, Internet of Things, CityPulse Data set, Knowledge Acquisition Toolkit.

## 1. INTRODUCTION

The rapid proliferation of Internet of Things infrastructure to capture and store large sets of data, has resulted in the availability of access to information for large number of data consumers. This has resulted in the need for both data analysis and visualisation methods. Where such data consumers range from quantitative researchers familiar with advanced data analysis tools to consumers with limited experience. To this end, we seek to introduce the concept of data analysis as a web service, that benefits both the advanced and beginner data consumer.

Internet of things has revolutionised the acquisition of data from real-world processes, thus enabling a set of new technologies referred to as smart objects to arise [1]. Such objects can be used to enhance consumers/users experience when interacting with a service. In particular, a specific example of large scale IoT systems with access to large quantities of potentially useful data for data consumers are smart cities [2]. Where the concept of smart cities have arisen from the need to exploit digital technology in order to address urban challenges such as traffic congestion and environmental pollution. While the collection of data by various city departments and services exists, these are often proprietary and the infrastructure for accessing data from these different data sources is often unavailable; owing to the lack of interoperability, between the relevant technologies (that is the infrastructure and data base storage of data). To this end, smart city projects such as iCity [3] and SmartSantander [4] have been proposed where multiple sensors are placed within the boundary of the respective cities that enable the collection and storage of large sets of data. However, such smart city projects do not offer data analytics (that is both data analysis and visualisation) or decision making systems for information extraction such as in CityPulse [5] [6]; where such a project has arisen in order to provide a unified data collection and decision making platform. By providing the relevant infrastructure for the open access of data, such projects seek to address the aforementioned urban challenges.

Internet of Things has enabled a wider range of users access to large sets of data. As a result, it is imperative to provide data analysis tools. In particular both private companies (such as IBM Watson analytics) and academic institutions [7] [8] are now providing both analysis and visualisation tools that can be installed on the users machines. Where such methods enable the user with limited experience to maximise the value of information extraction from their data sets. In particular the following works have focused on developing purpose built software for data analysis: the work in [8] developed a Java based machine learning/data mining user interface for processing data sets. While the knowledge Acquisition toolkit (KAT) [7] is a information extraction and data analysis platform (where both tools require the software to be downloaded on the data consumers machine). However, the download and install concept for data analysis tools has a fundamental drawback; that is such tools would require regular software updates in order to incorporate the latest tools and techniques for data analysis. To this end, this work seeks to introduce a web based system that enables data analysis (where this includes pre-processing methods such as filtering and machine learning algorithms) so as to benefit both the novice and advanced data consumer. We have developed a

978-1-5090-4117-6/17/$31.00 ©2017 IEEE

| Sensor 1 | Time-Stamp 1 | Sensor 2 | Time-Stamp 2 | ... |
|---|---|---|---|---|
| 1.43 | 2014-02-13T11:30:00 | 0.02 | 2014-02-13T11:40:00 | |
| 1.21 | 2014-02-13T11:35:00 | -0.4 | 2014-02-13T11:50:00 | |

(a)

```
{
 "Method": ["method 1","method 2","method 3"],
 "Parameters":["parameter 1","parameter 2","parameter 3"],
 "DataPointer":["http://link to CSV data"]
}
```

(b)

**Fig. 1**: (a) CSV format for both the sensor observations and time stamps. (b) The HTTP request format, for the relevant data analysis tools.

Python based web service that enables the user to select from a list of documented techniques and methods so as to carry out data analysis over the internet. Where we illustrate the advantages of the proposed web based data analysis (KAT web service) tool on data pertaining to traffic data obtained from the CityPulse data repository.

The organisation of this paper is as follows: Section 2 presents an overview of the current data analysis platforms. Section 3 describes the proposed data analysis web service, while Section 4 provides a brief description of the CityPulse data set and Section 5 offers an example of the data analysis on the CityPulse data.

## 2. EXISTING DATA ANALYSIS TOOLS

Open access IoT data providers such as CityPulse are enabling the consumer to both collect and analyse a deluge of data points pertaining to a large number of phenomena. As a result, there is an increasing trend in providing data analysis tools specialising in machine learning to data consumers for more rapid processing of data sets. Examples of existing data analysis tools include the following: the initial implementation of the Knowledge Acquisition Toolkit (KAT) developed as a Python application [7] (which the user has to download), where the tool provides a set of methods for annotating data sets after initial pre-processing and data abstraction. In particular the tool provides an algorithm workflow that is implemented sequentially: that is a pre-processing method, along with dimensionality reduction, (example: principal component analysis), data abstraction (examples include: clustering) and finally the labelling/annotation of the resulting abstracted data is carried out (an example is assigning for temperature data the labels, hot and cold. More details on labelling/annotation can be found in [7]). The work in [8] created a data mining software implemented in Java, referred to as Weka 3.0. Where this tool is downloaded on the clients machine, where the functionalities include, pre-processing and a range of machine learning algorithms (that is, regression, clustering and supervised learning to name but a few). A drawback of these tools is the requirement for the data consumer to download such tools, which limits the class of users due to the requirements of installing additional software over existing technologies (such as MATLAB and R). The CityPulse project seeks to partially address this drawback, by not only providing IoT data, but it also provides a set application specific data analysis tools as a set of online tools for data consumers. In particular, the tools focus on event detection on data streams along with complex event processing of such events [9].

| | |
|---|---|
| Pre-processing | 1) Outlier Removal (Winsorization)<br>2) Digital Filtering (Low,Band,High-pass) |
| Supervised Learning | 1) Linear Regression<br>2) k-Nearest Neighbour |
| Unsupervised Learning | 1) Principal Component Analysis<br>2) k-means Clustering |
| Other | 1) Correlation Coefficient<br>2) Spectral Estimation |

**Fig. 2**: Table showing the data analysis category along with the corresponding examples of methods included in the KAT web service.

## 3. DATA ANALYSIS WEB SERVICE

Existing data analysis tools offer such systems as software that requires the user/data consumer to download the product. However, such systems may require the user to learn a domain specific language (such as Java and Python) for the use of such systems, while also requiring software updates so as to access the latest data analysis techniques. As a result, it is imperative to provide open access data analysis tools for consumers as a web service, that provide the following benefits: namely for novice/beginner data consumer, the tools that would enable them to analyse and obtain useful information. While for the more advanced/experienced user providing the most effective tools for a given data set (we generally refer to a data set as information being obtained from IoT sensors). For example, such a tool would provide relevant documentation for the beginner data consumer, with examples of data processing work flows; while for the more experienced user the most advanced tools developed in academic institutions can be evaluated by a wide range of users and the most useful data analysis tool for a given data set will be identified.

To this end, this work presents a web service (known as the KAT web service, as the proposed tool is an extension and modification of the original KAT tool) for data consumers that provides access to range of tools and techniques for processing time series data. The proposed KAT web service is a Python implemented system, where a data consumer would send an HTTP request of the relevant algorithms and corresponding parameters they seek to have executed on a data set. In particular the data analysis tools are grouped into the following categories, namely: 1) pre-processing methods, 2) supervised learning, 3) unsupervised learning algorithms and finally 4) other techniques and methods (examples of techniques belonging to this class are spectral and dependence estimation methods) [10] [11] [12]. Fig. 2 presents examples of the relevant algorithms for each group, where users will find relevant information for each method via online documentation. Furthermore, the proposed data analysis tool as a web service would enable more rapid updates of the list of algorithms (as algorithms proposed by new research groups will also be included). We now provide an overview of the requirements for the data consumer to utilise the proposed KAT web service[1] :

1. Create a comma separated value (CSV) file with the following format. Where the sensor columns alternate between data and time-stamps (that is the following column header: sensor 1 data, time-stamp sensor 1, sensor 2 data, time-stamp sensor 2 data, etc), an example of which is shown in Fig. 1 (a). Furthermore, the CSV file needs to be available as web link, such that the proposed KAT web service can discover the data.

2. Send an HTTP request to the KAT web service, where an example of the exact format for making the request is shown in Fig. 1 (b). Where the list of methods and corresponding parameters are included (online documentation would help the user determine the number of parameters required for each method). Furthermore, an HTTP link to the data source also needs to be provided.

It should be noted that, the CSV format was selected so as to readily identify columns pertaining to the time-stamp of the corresponding output observations, such that re-sampling of the observations to common sampling frequency is achieved (we align the sampling frequency to the sensor with the lowest sampling frequency). Furthermore, the proposed KAT web service requires the data consumer to select the correct sequence of methods. For example, one cannot execute the Fourier transform on a dataset followed by the execution of a pre-processing algorithm (an example being bandpass filtering). To this end, we provide an error message when such incorrect sequences of data analysis methods are selected and accordingly refer the user to the documentation. Finally, the output of the processed time series data is returned to the user as a CSV file.

## 4. CITYPULSE DATASET

We first provide an overview of the the CityPulse dataset, where the data sources are categorised into eight different domains: vehicle traffic, pollution, weather, cultural events, social events, library event, parking and meeting room data. Among these data sets, road traffic and parking data sets are real-time data streams that are available to the public for collection from the City of Aarhus, Denmark. The City of Aarhus is amongst the smart cities that provide an open data platform called Open Data Aarhus (ODAA)[2], which contains city related information generated by various sensors deployed within the city. Furthermore, for each domain a set of historical data is also available for public use at the following link[3].

```
{
 "Methods": ["TrafficAnalyser"],
 "Parameters":["3 ,P,0 ,1 ,2"],
 "DataPointer":["http ://link to CSV data"]
}
```

**Fig. 3**: An example of the input to the KAT web service using the TrafficAnalyser function. There are 5 input parameters. The first parameter corresponds to $l$, the second to the distribution (in this case ”P” corresponds to Poisson) and the last three parameters correspond to the labels $a, b, c$.

## 5. CASE STUDY: CITYPULSE DATA SET

We now present use case examples for the processing of data obtained from the CityPulse data set, using the proposed KAT web service. The data used for this simulation pertained to traffic data collected from link[4] with the following file name: trafficData158324. The traffic data CSV file contains the following measurements: 1) average vehicle speed, and 2) the vehicle count between two points on a street.

In this particular use case example, we provide a data analysis tool (referred to as: TrafficAnalyser) that obtains the relevant statistics such that labelling/annotation of both the average vehicle speed and vehicle count data can be assigned, thereby providing human interpretable information from traffic sensor data [13]. That is, we seek to identify a confidence interval for the average of the measurements $x_d(t)$, where $t$ corresponds to the time in hours and minutes over one day, and $d$ corresponds to the set of dates. Accordingly, we define the confidence interval around $x_d(t)$ as follows

$$\mu(t) - l\sigma(t) \leq x_d(t) \leq \mu(t) + l\sigma(t)$$

where $\mu(t)$ is the sample mean

$$\mu(t) = \frac{1}{D} \sum_d x_d(t)$$

[1] The link to the documentation is provided in the following: http://iot.ee.surrey.ac.uk:8080/datasets/traffic/traffic_feb_june/index.html

[2] http://www.odaa.dk

[3] http://iot.ee.surrey.ac.uk:8080/datasets.html

[4] http://iot.ee.surrey.ac.uk:8080/datasets/traffic/traffic_feb_june/index.html

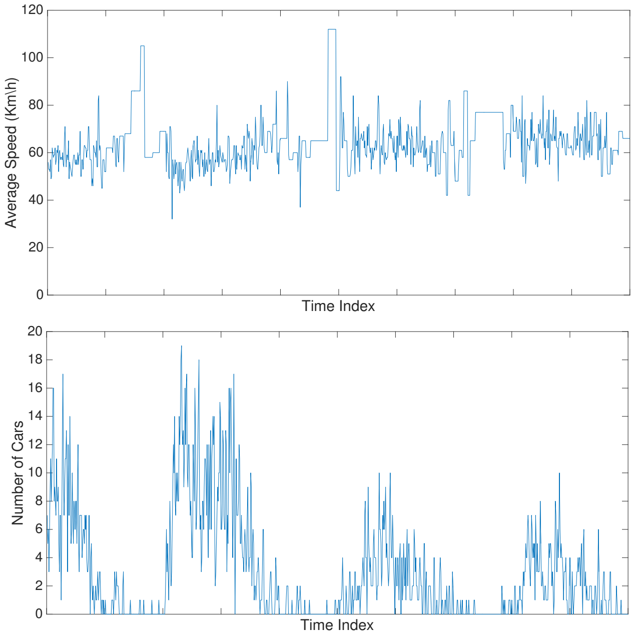

**Fig. 4**: (Upper panel) The average vehicle speed and (lower panel) the vehicle count traffic data.

and $\sigma(t)$ is the sample standard deviation, such that depending on the underlying statistical distribution assumed for the data (that is for vehicle count we assume a Poisson distributed data set, while for the average vehicle speed we assume a Normal distribution),

$$\sigma^2(t) = \begin{cases} \frac{1}{D}\sum_d x_d(t) & \text{if } x_d(t) \text{ is Poisson Dist.} \\ \frac{1}{D}\sum_d (x_d(t) - \mu(t))^2 & \text{if } x_d(t) \text{ is Normal Dist.} \end{cases}$$

where $D$ corresponds to the number of dates and the user defined parameter $l$ corresponds to the significance level of the Normal test. Furthermore, the user then defines three input labels $\{a, b, c\}$, such that annotation, $x_d^{sem}(t)$ , of the data set $x_d(t)$ can then be carried out, that is

$$x_d^{sem}(t) = \begin{cases} a & \text{if } x_d(t) - \mu(t) \leq -l\sigma(t) \\ b & \text{if } |x_d(t) - \mu(t)| < l\sigma(t) \\ c & \text{if } x_d(t) - \mu(t) \geq l\sigma(t) \end{cases}$$

As an example for the vehicle count data, the following labels can be selected: $\{a = \text{“}BelowAverage\text{”}, b = \text{“}Normal\text{”}, c = \text{“}AboveAverage\text{”}\}$, thereby providing an human interpretable meaning to the numerical vehicle count data. Finally, examples of the input parameters and methods selected for the traffic count data set is shown in Fig. 3.

Finally, Fig. 4 illustrates the numerical representation for both the average vehicle speed (upper panel) and vehicle count (lower panel). While Fig. 5 presents the corresponding output of the TrafficAnalyser method stated above, where it can be observed that such a representation can be interpreted more effectively, than the original data set.

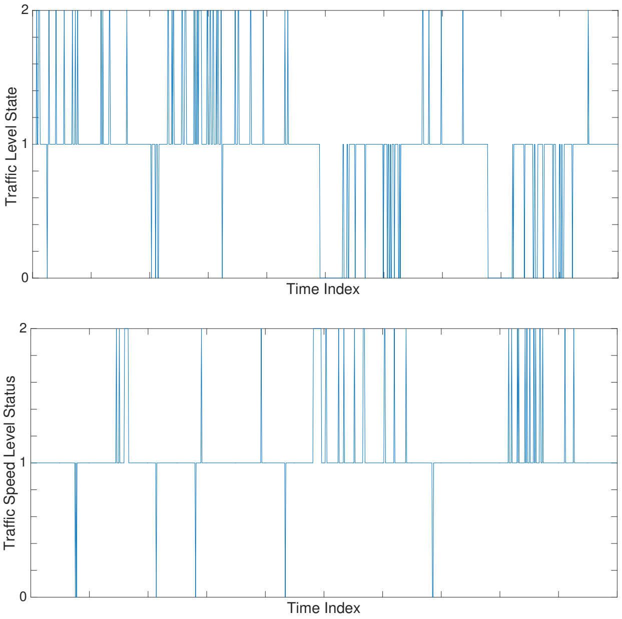

**Fig. 5**: (Upper panel) Figure showing traffic level status and (lower panel) shows the traffic average speed level status, where for both figures the following labels were used $\{0 = \text{“}BelowAverage\text{”}, 1 = \text{“}Normal\text{”}, 2 = \text{“}AboveAverage\text{”}\}$.

## 6. CONCLUSION

In this work we have proposed a data analysis web service tool referred to as the KAT web service that enables the data consumer to have access to a range of algorithms for the analysis of sensor data. Furthermore, the proposed tool implemented in Python requires the data consumer to access the service via simple HTTP requests. Thus providing data analysis over the internet, a capability that would enable the proliferation of data analysis algorithms for a wide range of users, thus benefiting both the advanced and novice data consumer. Thereby enabling more efficient identification of effective algorithms for the processing of a given data set. Future work will focus on increasing the set of data analysis techniques made available, along with improving and refining the documentation.

## 7. ACKNOWLEDGEMENTS

The research leading to this work has received funding from the European Commissions in the H2020 for FIESTA- IoT project under grant agreement no. CNECT-ICT-643943.

## 8. REFERENCES

[1] S. De, T. Elsaleh, P. Barnaghi, and S. Meissner, "An internet of things platform for real-world and digital objects," *Scalable Computing: Practice and Experience*, vol. 13, no. 1, pp. 45–48, 2012.

[2] B. Firner, R. S. Moore, R. Howard, R. P. Martin, and Y. Zhang, "Poster: Smart buildings, sensor networks, and the internet of things," *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*, pp. 337–338, 2011.

[3] iCity Consortium. (2014, June). [Online]. Available: http://www.icityproject.eu/

[4] B. Cheng, S. Longo, F. Cirillo, M. Bauer, and E. Kovacs, "Building a big data platform for smart cities: Experience and lessons from santander," *2015 IEEE International Congress on Big Data*, pp. 592–599, 2015.

[5] D. Puiu, P. Barnaghi, R. Tönjes, D. Kümper, M. I. Ali, A. Mileo, J. X. Parreira, M. Fischer, S. Kolozali, N. Farajidavar, F. Gao, T. Iggena, T. L. Pham, C. S. Nechifor, D. Puschmann, and J. Fernandes, "Citypulse: Large scale data analytics framework for smart cities," *IEEE Access*, vol. 4, pp. 1086–1108, 2016.

[6] S. Kolozali, D. Puschmann, M. Bermudez-Edo, and P. Barnaghi, "On the effect of adaptive and non-adaptive analysis of time-series sensory data," *IEEE Internet of Things Journal*, no. 99, pp. 1–1, 2016.

[7] F. Ganz, D. Puschmann, P. Barnaghi, and F. Carrez, "A practical evaluation of information processing and abstraction techniques for the internet of things," *IEEE Internet of Things Journal*, vol. 2, no. 4, pp. 340–354, 2015.

[8] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.

[9] F. Gao, E. Curry, and S. Bhiri, "Complex event service provision and composition based on event pattern matchmaking," *Proceedings of the 8th ACM International Conference on Distributed Event-Based Systems*, vol. 11, pp. 71–82, 2014.

[10] S. Garcia, J. Luengo, and F. Herrera, *Data preprocessing in data mining*. Verlag, NY: Springer, 2014.

[11] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Verlag, NY: Springer, 2015.

[12] P. Stoica and R. L. Moses, *Spectral analysis of signals*. Upper Saddle River, NJ: Prentice Hall, 2005.

[13] P. Anantharam, P. Barnaghi, K. Thirunarayan, and A. Sheth, "Extracting city traffic events from social streams," *ACM Transactions on Intelligent Systems and Technology*, vol. 6, no. 4, pp. 1086–1108, 2015.