Joe Honour B3029153 Newcastle University

# Dissertation Proposal: Develop a Machine Learning Application that is Able to Detect Real-Time Anomalies in User Behaviour.

## Motivation

In the United Kingdom, an estimated 4.2 million surveillance cameras watch us every day (Norris and McCahill, 2006). CCTV is deployed to businesses, homes, shops and on high streets making us one of the most watched nations in the world. CCTV is actively available in 96% of homicide investigations, where it added value to the case 80% of the time (Scotland Yard, 2010). However, this statistic does not hold true for all crimes within the UK, with reports showing that one in every 1000 crimes involves CCTV usage (Hope, 2009). Albeit CCTV may not be required in every crime, it is still fair to say that the large work force allocated to high profile crimes makes it possible to find, collect and search all the archived video footage for evidence. This illuminates the problem at hand; CCTV is called upon when it is often too late to be proactive in stopping a crime, or the physical consequences of a crime have already been felt by the victims, whether this be an individual or business. We have an opportunity to combat this if we can begin to ingest video footage in a smarter capacity. If we add a capability to understand the events occurring in a video; who is present and how they are acting, then we would be able notify a relative party of an anomaly in a much shorter space of time, and even be able to stop certain crimes before they occur.

To achieve this, once presented with a set of data points representing the events that have occurred in a video feed, we need to be able to classify whether what we are shown is anomalous in nature or within the norm. Furthermore, we need to apply this on an individual basis and detect when a single person on video is acting out of the ordinary. If we can achieve this, we are in a capacity to create a system that could alert the appropriate authority of an anomaly in real-time, allowing event investigation and necessary action to be taken. This creates a much more proactive approach to detecting crime along with removing the need for archival of video footage.

Furthermore, a proposed solution could build upon existing CCTV infrastructure, meaning a benefit is felt without a need to purchase or invest in new hardware. This is an improvement over existing smart CCTV systems that often require specialised equipment to operate. (Nest, 2017)

## Approach

My approach will be to develop a system that can detect users in a video stream, extrapolate data points about them, and then apply a variety of machine learning models to classify their behaviour as anomalous or normal. I will be focusing on unsupervised learning methodologies, evaluating which models provide the highest level of accuracy while being as high performant as possible. The system will use a distributed architecture in order to provide scalability and high-performance data throughput, a key success criterion for a real-time system.

The solution will not require storage of any video footage, as the video processing happens at the video source. I am aiming to show that this proposed approach will enable higher performance than existing alternatives while being accurate in its ability to detect anomalies is user behaviour. To develop a suitable and achievable application in the given timeframe I will confine the problem to a single room environment with a set of configured subjects for facial recognition.

## Aim and Objectives

**Aim:** To develop a machine learning application that is able to detect real-time anomalies in user behaviour.

**Objectives:**

1. To research existing video processing techniques and available software in order to obtain as many data points with the highest degree of accuracy from a video stream.
2. To research machine learning techniques for detecting anomalies in time series data that can be applied to data sets produced from objective one.
3. Develop testing scenarios that will allow the evaluation of machine learning models in their ability to detect anomalies in real-time.
4. Develop a solution that provides a Minimum Viable Product of video processing, anomaly classification models, an interactive web interface and an anomaly notification capability. This must be scalable and provide anomaly detection in real-time.
5. Using the test scenarios defined in objective three, evaluate the applications ability to detect anomalies and alert users in real-time.
6. Compare and contrast the performance and storage requirements of the final system against existing CCTV methodologies.

To be considered successful, the proposed solution will need to be able to make accurate anomaly predictions in the shortest space of time feasible. Therefore, evaluation of speed is equally as important as accuracy when considering a machine learning models success within this solution and the success of the solution as a whole.

Through my initial work on my dissertation, I have been able to produce a first architectural diagram of the proposed solution, based off the papers described in the background section of this document (Figure 1). This should aid in the description of what a distributed based architecture looks like and show initial direction towards the shape of my final solution. You can see we make use of Apache Storms topology to allow distributed processing of incoming data, and then allow a web interface to display anomalies and provide a visual representation of the data.

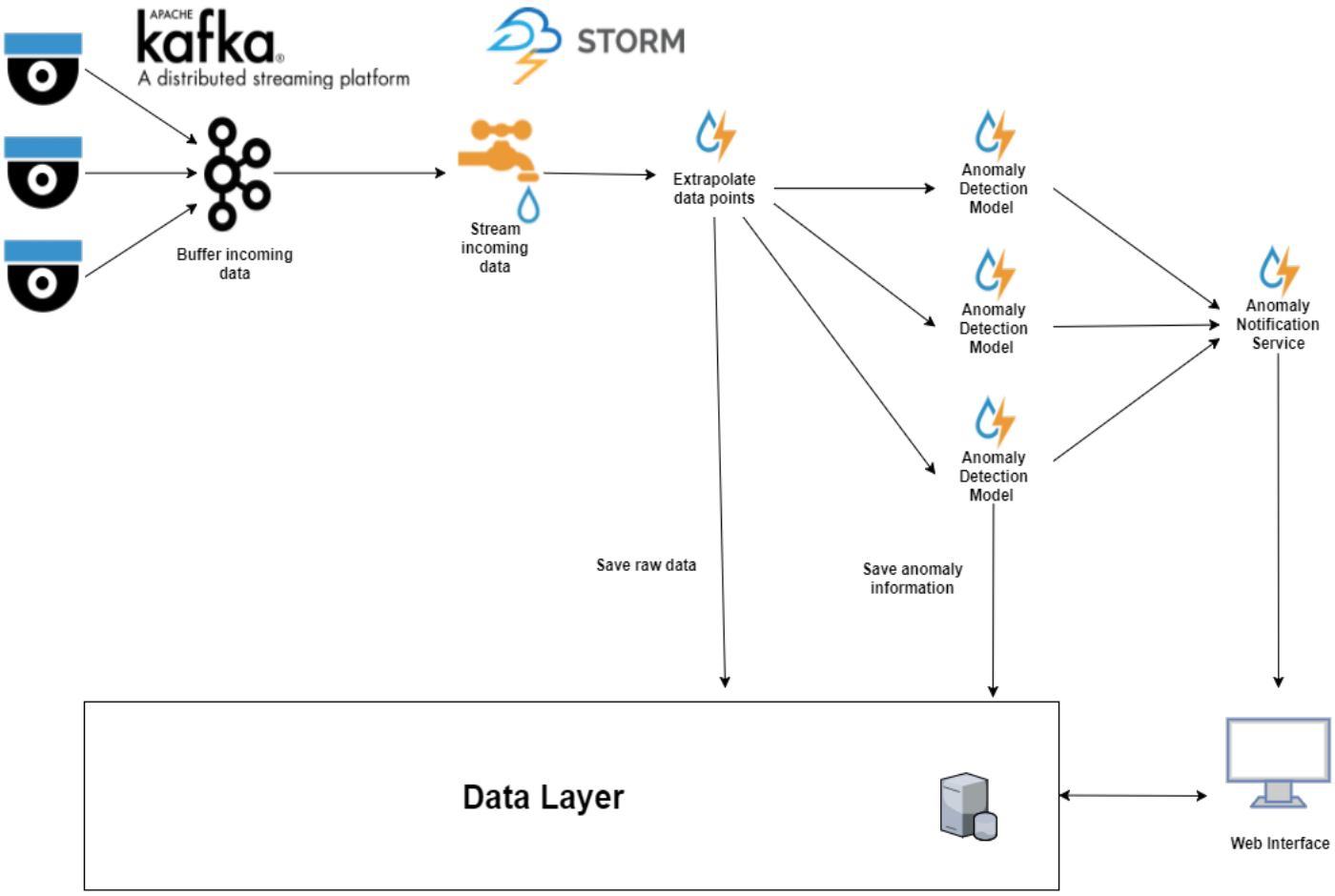


Figure 1: Proposed architecture for real-time analytics and anomaly detection

## Background

**Paper:** A Survey on Behaviour Analysis in Video Surveillance for Homeland Security Applications (Ko, 2008)
**Description:** The paper shows an overall look at the stages involved in creating an intelligent video stream solution. It gives insight into object classification, object tracking, extracting motion information, and behaviour analysis.
**Relevance:** The paper proposes a general architectural design from the initial capture of video, making sense of what is in the feed to making smart decisions based off this information. The paper, with its insights into computer vision, give me a general basis of the requirements found in processing real-time video. I will be using this paper for its capacity in behaviour analysis, giving me a high-level overview of some of the most effective models in behaviour classification.

**Paper:** A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data (Goldstein and Uchida, 2016)
**Description:** The paper aims to achieve a comparative universal evaluation between the most common unsupervised learning models using public domain datasets for comparative evaluations in performance and accuracy. It provides benchmarks in accuracy and performance of these models allowing ranking of models against each other.
**Relevance:** During my work, I will be required to detect anomalies in human behaviour within video feeds. This paper gives a good basis for selecting and exploring known machine learning models within this space, looking at not only accuracy and success of the models but also on their performance. It compares and contrasts unsupervised learning mechanisms on a variety of datasets allowing me a good foothold when choosing what models to adopt for my own

data set. I will be using this paper to influence which models are worth considering, and which may not be appropriate for my classification problem, either due to accuracy issues or performance lags when adopting them.

**Paper:** Unsupervised Learning Techniques for an Intrusion Detection System (Zanero and Savaresi, 2004)
**Description:** The paper proposes a two-tier architecture for detecting anomalies in network traffic using solely unsupervised learning methodologies. It looks at how you can initially perform clustering of data to reduce dimensionality and then enable more sophisticated anomaly detection algorithms to categorise data into anomalous and normal groups.
**Relevance:** The mentioned paper investigates the use of unsupervised machine learning models; K-means algorithm, Principle Direction Partitioning and Self Organising Maps, in order to provide anomaly detection on network traffic. The investigation of these models is extremely useful for my work, when evaluating which methodologies to adopt in detecting anomalous human behaviour. It further proposes an architectural approach that adopts an initial filtering of data using clustering algorithms to then enable more traditional anomaly detection methods. This approach is relevant to my work as I will be looking at how best to gain real-time analysis of video streams, which may require data transformations to make possible, and this approach could be adopted to my work.

**Paper:** Real-Time Network Anomaly Detection System Using Machine Learning (Zhao *et al.*, 2015)
**Description:** The paper proposes a novel framework for real-time network traffic anomaly detection using machine learning algorithms.
**Relevance:** The paper proposes a suitable architecture for my dissertation project, though with a slightly different aim and data source. It will allow me a basis to work from and insight into problems to overcome when attempting to provide real-time processing of data streams. They adopt real-time technologies including Apache Kafka, Apache Storm and Apache Hadoop, which will greatly influence the final design of my dissertation project.

**Paper:** Microservice Architectures for Scalability, Agility and Reliability in E-Commerce (Hasselbring and Steinacker, 2017)
**Description:** The paper presents how microservice architectures facilitate scalability, agility and reliability using an industrial case study.
**Relevance:** My dissertation will need the key features of the microservice-based architecture, mainly focused on scalability and agility. This paper allows a firm understanding of how we can achieve this and the design decision required to create a successful solution. It also provides a real-world success story based on this approach, which allows me confidence in the achievability of my solution.

**Paper:** About Microservices, Containers and their Underestimated Impact on Network Performance (Kratzke, 2017)
**Description:** This paper aims to investigate the cost associated with using a container-based approach within a microservice architecture. It sets out to prove that although microservices provide scalable high-performance systems, there is a cost associated with them.
**Relevance:** My dissertation will require high-performance to be viable; we need to detect anomalies while we still have to act on them. To achieve this, I will be considering a microservice architecture and this paper allows me the appropriate information in the cost I incur in adopting this design. It can also provide reason if performance issues are encountered during the development of the final solution. This is a good counter argument to containerisation, and shows the considerations to account for when trying to achieve horizontal scalability.

## Project Plan

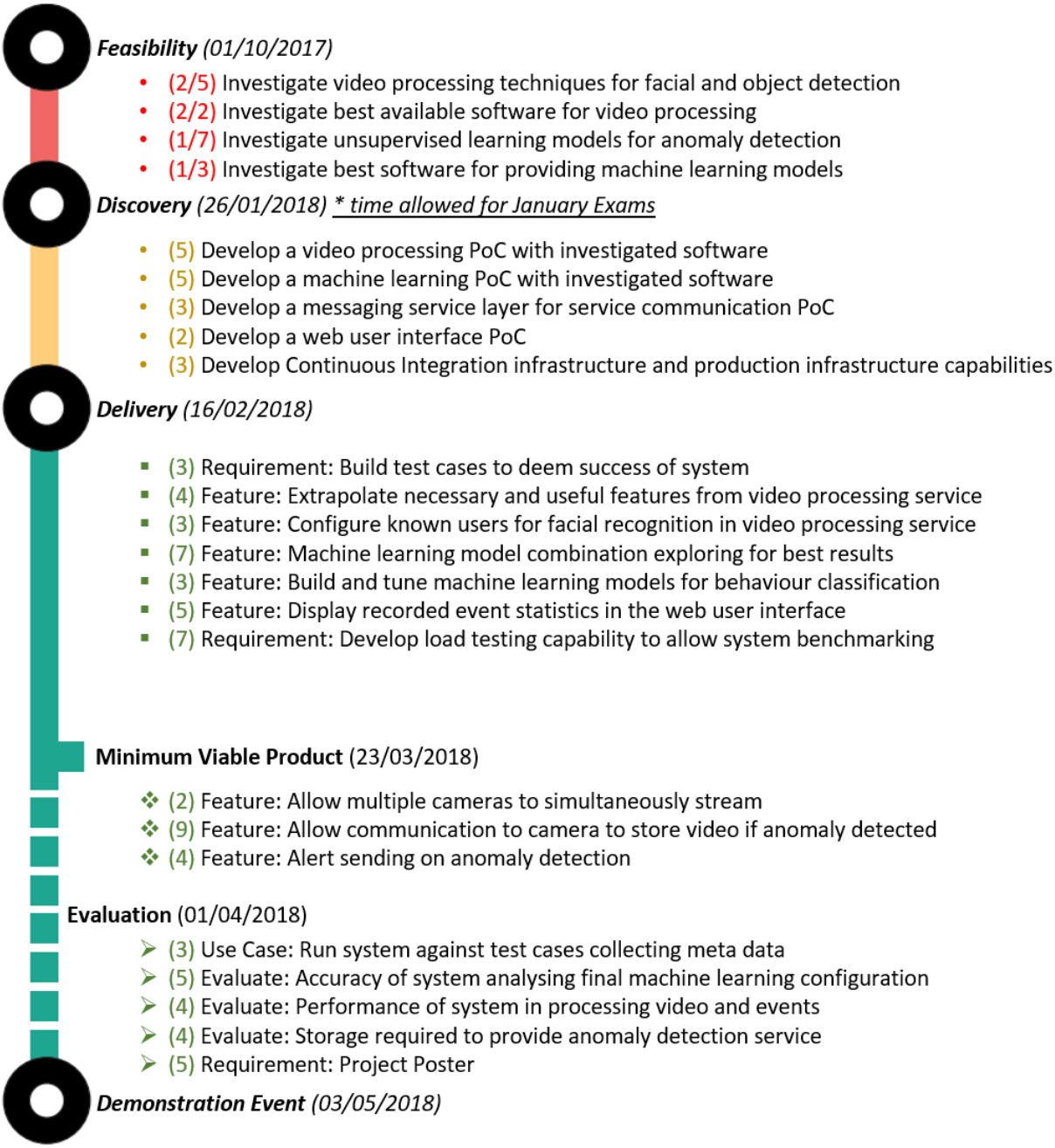


Figure 2: The project plan to completing the dissertation

KEY:

- (NUMBER) = Estimates the amount of work required to complete the task.
- (N/NUMBER) = The amount of work done towards completing the task.
- PoC = Proof of Concept

I am currently working through the feasibility phase of the project lifecycle. You can see from Figure 2, based on the Enterprise Agile model (BJSS, 2016), that I have completed tasks within the feasibility stage of the project and will be moving onto the discovery phase after the January exam period.

I have chosen this delivery methodology as it allows me an achievable and measurable journey to completion, while allowing the flexibility to adapt to change as I move through the phases of delivery. The Discovery phase enables me to prove the technologies are compatible and work for their chosen tasks, and if they do not I am able to pivot quickly to replace them. This development style is known as a fail fast approach.

Once I have a proven technology stack I move on to the core of the implementation, developing features in the priority displayed in the project plan, however this priority can change at any time based on problems I may encounter or future knowledge gain. I believe this gives me the greatest chance of success, not only in completing the dissertation, but also in producing a viable product.

## References

BJSS (2016) *Enterprise Agile*. 4th edn. BJSS.

Goldstein, M. and Uchida, S. (2016) 'A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data', *PLoS ONE*, 11(4). doi: 10.1371/journal.pone.0152173.

Hasselbring, W. and Steinacker, G. (2017) 'Microservice architectures for scalability, agility and reliability in e-commerce', in *Proceedings - 2017 IEEE International Conference on Software Architecture Workshops, ICSAW 2017: Side Track Proceedings*, pp. 243–246. doi: 10.1109/ICSAW.2017.11.

Hope, C. (2009) 'One crime solved for every 1,000 CCTV cameras, senior officer claims', *Telegraph.co.uk*. Available at: http://www.telegraph.co.uk/news/uknews/crime/6081549/One-crime-solved-for-every-1000-CCTV-cameras-senior-officer-claims.html.

Ko, T. (2008) 'A survey on behavior analysis in video surveillance for homeland security applications', *Applied Imagery Pattern Recognition Workshop, 2008. AIPR '08. 37th IEEE*, pp. 1–8. doi: 10.1109/AIPR.2008.4906450.

Kratzke, N. (2017) *About Microservices, Containers and their Underestimated Impact on Network Performance*. Available at: https://www.researchgate.net/publication/273456042_About_Microservices_Containers_and_their_Underestimated_Impact_on_Network_Performance.

Nest (2017) *No Title*. Available at: https://nest.com/uk/cameras/nest-cam-indoor/overview/ (Accessed: 24 November 2017).

Norris, C. and McCahill, M. (2006) 'CCTV: Beyond penal modernism?', *British Journal of Criminology*, 46(1), pp. 97–118. doi: 10.1093/bjc/azi047.

Yard, Scotland. (2010) *CCTV in Homicide Investigations*. Available at: https://goo.gl/oS5Tgn (Accessed: 15 November 2017).

Zanero, S. and Savaresi, S. M. (2004) 'Unsupervised learning techniques for an intrusion detection system', in *Proceedings of the 2004 ACM symposium on Applied computing - SAC '04*, p. 412. doi: 10.1145/967900.967988.

Zhao, S. *et al.* (2015) 'Real-time network anomaly detection system using machine learning', in *2015 11th International Conference on the Design of Reliable Communication Networks (DRCN)*, pp. 267–270. doi: 10.1109/DRCN.2015.7149025.