Research Methods and Professional Issues
Research Project Proposal

SCOUTING POTENTIAL HIGH PREFORMANCE ATHLETES

I. Introduction

In sports scouting plays an important role in identifying talent to build a successful team. Scouting can be viewed in terms of the short-run and the long-run. Scouting talent with specific attributes required to strength the current team is seen as short-term scouting and while long-run is seen as scouting youthful prospects at an earlier stage in their career in comparison to the short-term. The aim of this task is to investigate short-term scouting for current athletes who are on track to become elite athletes based on their performance.

The application can be adopted to the different stages of the scouting process for team sports as described by most authors [1]:

- Talent detection Stage – Detecting talents that are estimated to possess adequate predisposition.
- Talent Identification Stage – The process of identification of potential elite athletes.
- Talent development Stage – Development of their potential in line with their characteristics.
- Confirmation Stage – Monitoring, verification, and confirmation of the development speed of the athlete.
- Selection Stage – Identification of athletes who will be able to execute set tasks in a certain sports context.

Identifying potential elite athletes is potentially one of the most important topics in sports. Historically in some sports such as football, the tools used for analysis were at the level of "GCSE maths as opposed to PhD Maths" as suggested by James smith of Everton FC [5]. In recent years this has completely changed, the emergence of big data and the success of data scientist at recruiting unrecognised talent has led to a revolution within sports to adopt more advance analytical techniques.

Not lot of information is available about how current sports teams forecast athlete development, and how they categories these growths. Looking at academic reading not much can be found on techniques applied in this context. Most forecasting papers concentrate on predicting the outcomes of games, tournaments or injuries using random forest [6]. Some papers have adopted Kmeans clustering on identifying top current elite athletes [4], but this doesn't consider future potential. The content of this research proposal will look to address this future potential and time lag, how different athletes develop at different time periods in their career.

For this proposal, two aspects will be addressed to help aid in the scouting process. Firstly, we will be forecasting athlete's development using a SARIMAX model which supports a seasonal component of a univariate time series data and then cluster the results using

Dynamic time warping. Following this the research question is: Can we build a meaningful forecasting model for athlete development based on historical performance? And can we successfully build a clustering model to cluster their developments into different performing groups?

The desire initially was to run this project on football player athlete data, but due to the lack of available data, I will switch to using basketball player athlete data instead. The intention is the applied method can be extracted and used for any sports and any dataset. The first aim is to understand the dataset, its structure and content. Due to the scope of the question addressed, not every athlete will be required for the analysis, only those with a few years' experience as we look to forecast their future. The second step will be to understand the metrics involved, as we will be required to fashion a metric to help us generate a general metric to evaluate the overall ability. The next step will be to understand how to prepare the data before implementing a time series model and how to improve on it. Once the forecasting model is successful, we can then implement the clustering model and then profile the clusters.

This proposed attempt can be valuable if successful as a method of identifying potential talent in any sports such as football, American football, basketball, and rugby. It can also be valuable in other fields where any metric is recorded over time and a classification may be required after such as stocks, student exam results and others. It can also be the base of further advance development and potentially an integration of the two techniques, where one model can forecast and classify given a set of data.

II. Critical Content

The first aim to address application of the forecasting model, even though there are plenty of academic paper regarding forecasting and the application of SARIMAX, none are applied in the sports field of athlete development. So, we will look at how it has been applied outside of sports and whether it is the correct model to use. One of the many applications of SARIMAX has been in forecasting daily retail sales [7]. The foundation of forecasting is laid out initially "The forecastability of demand highly depends on the volatility of demand", so if demand is flat and steady it can easily be forecasted but if demand is irregular and arbitrary then it cannot be expected to obtain an accurate forecast. The paper also discusses improving forecasting models not only by improving accuracy but by increasing efficiency, as there is a trade off in modelling between accuracy and complexity. According to Herbig [8] when selecting a forecasting model, one should focus on certain aspects, here are some of those:

- Time to forecast
- Technical resources
- Variability and consistency of data
- Accuracy
- Form

Some of the advantages of implementing time series forecasting according to some [9],[10] and [11] are as follows:

- Easy Implementation
- Better Interpretation
- Reasonable accuracy compared to other approaches

The next aim is to understand the different types of ARIMA models. To establish if our selection is indeed the best model to use for our forecasting model. A stochastic time series model which is an autoregressive integrated moving average model is known as an ARIMA model [12], this type of model is usually successful if the time series is linear and follows a normal distribution. As we can see from figure.1 an ARIMA is applied to a seasonal time series but doesn't fair well as it can only predict the trend of the future forecast. The ARIMA model can be adapted for a seasonal time series and is known as SARIMA, but if the original ARIMA model includes other timeseries inputs it can be also be adapted and is known as an ARIMAX. This is where our model comes in SARIMAX, which is a combination of the former two adaptations, a seasonal autoregressive integrated moving average with an exogenous variable. The combination of both the seasonality and the exogenous variable help the model to overcome the disadvantages of the original ARIMA model in retail and allows the model to incorporate external factors that have an influence on forecast [9]. Figure.2 shows how a SARIMAX model can predict forecast by capturing the seasonality trend and the overall downward trend.

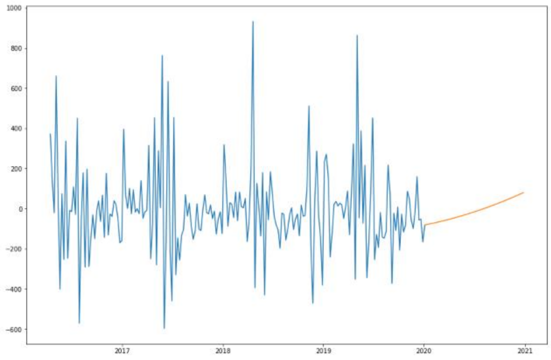


Figure.1

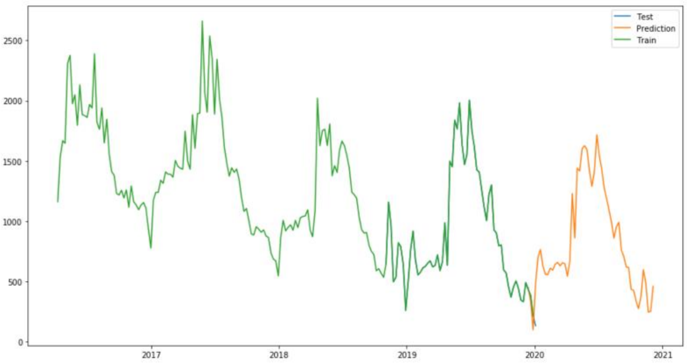
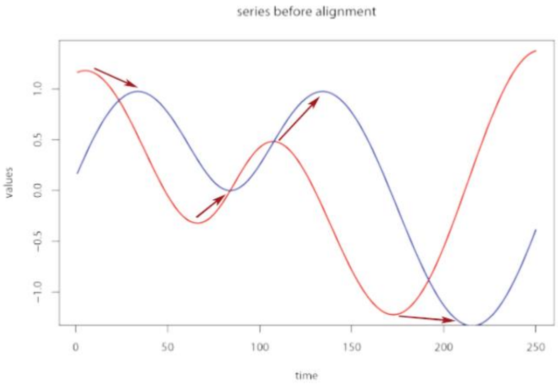


Figure.2

The application of SARIMAX models have been seen in a wide spread of industries forecasting completely different variables from traffic counts [13], weather [14], electricity peaks [15], animal migration [16] and coin circulation [17].

The second aim is to understand the clustering algorithm that will group the athletes forecasted performances. We will be looking at DTW, which is known as Dynamic time warping. This algorithm is used for detecting the similarity between two sequences which may vary in time and space. This concept was introduced in the 60s and was later explored in the 70s, it is now applied to many fields such as speech recognition, computer vision and of course time series clustering.



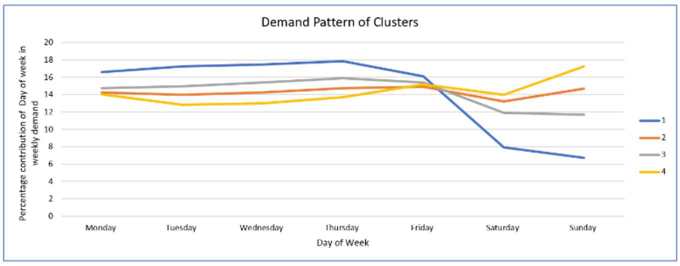


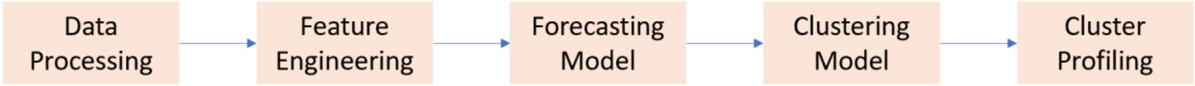
Figure.3 [18]

Figure.4

As a time-similarity measure, DTW has earned a reputation of being extremely efficient by using elastic transformation to minimise the effects of distortion and shifting in time, this allows it to detect different phases with similar shapes like in figure.3. The one requirement for this algorithm to work is that the data sequences should be at equidistant points, if not can be solved be re-sampling. In the core of the DTW algorithm is a cost function which is distance based, where if "d" is great, the less similar the sequences. Optimal alignment is achieved by minimising the cost function through arranging the sequences [20]. In figure.4, we can see the outcome of the DTW clustering algorithm, it shows the average pattern detected for each cluster. In this example, the algorithm was able to detect a weekly pattern for most retail stores except each cluster was trading at a different volume, but where the DTW was successful was clustering the four different divergences of the sequences during the weekend. The next step is then to discover and profile what is in each cluster, this can be achieved using simple statistics and data manipulation.

III. Approaches: Methods & Tools for Design, Analysis & Evaluation

For this section we discuss the methods used to address our project proposal. We visually represent this using a project pipeline shown in figure.5.


Data Processing

Feature Engineering

Forecasting Model

Clustering Model

Cluster Profiling

Figure.5

For the data processing stage, we will look to source the identified dataset from Kaggle an online data science and machine learning community website. The dataset in question is named as "NBA Players stats since 1950" which is a collection of comprehensive datasets about individual player statistics from the inception of the NBA league in 1950 until 2018. The next steps will be to explore and clean the dataset, once this is achieved, we can prepare the dataset for analysis by manipulating it. Players who have retired or near the end of their career will not be needed in the dataset as our aim is to predict future performance. We will look to keep those players who have over half of their career to play, this will be measured against the average playing career in the league. For this stage we will look at using python as the coding language due to familiarity and to avoid any hardware limitation we will be using google colab on google drive as the notebook. We will be using packages such as pandas and numpy for importing, cleaning and manipulation. Packages such a seaborn and matplotlib for visuals.

The next stage is feature engineering. The first step is to analyse the different possibilities of the time frame we will use to predict on, for example by season, month or per game. Once this is finalised, we can then focus on creating a metric to forecast on. In basketball the important fundamental stats are points, assists, blocks, total rebounds, and turnovers. We will then use these against minutes played or games played to create a stat per time-period. The idea for a metric is to help reduce the data and complexity of the model. We will continue to use python in google colab here.

The next stage is the forecasting model. We will adopt the SARIMAX model to forecast the future development but before hand we will need to decomposition the data. This stage is the process of deconstructing the time series into three components trend, seasonality, and noise. The next step is to check if our data is stationary, which means that overtime the mean, covariance or standard deviation should be constant, or auto-covariance should not depend on time. The two components that cause a time series not to be stationary are trend and seasonality. When a model is stationary its behaviour over a time interval may repeat again, which helps with forecast accuracy. A simple KPSS test will determine if the time series is stationary or not. We will continue to use python and google colab for this section and we will use statsmodel package for the decomposition, checking the stationarity and the SARIMAX model.

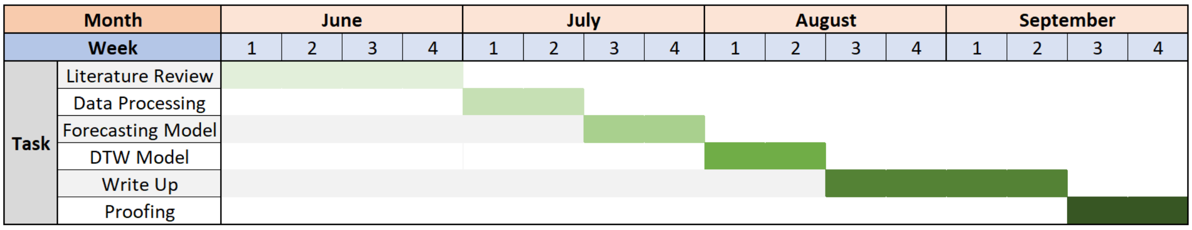
For the Clustering model we will be using a package called dtwclust in R language using R Studios which is perfect for our task. Its described as “Time series clustering along with optimized techniques related to the Dynamic Time Warping distance and its corresponding lower bounds.” [19]. Some transformation or pivoting of the data may be required due to the model output structure.

The last step is cluster profiling, this is one of the fundamental steps in in general clustering. The process is about discovering the attributes in each cluster using basic statistics and data manipulation, such as count, maximum and minimum value and mean. This process will give a general theme of the content of each cluster, which will enrich our final insight. For this section I will revert to python and Google Colab.

The last part of this section is Ethics and privacy. We will be using a public dataset from a public community website “Kaggle” which is renown within data science community. The dataset was originally scraped from www.sports-reference.com. In terms of privacy, no private or sensitive data is present within the dataset and nothing which isn’t already in the public domain. All data is old historic data which means it cannot affect current affairs or harm anyone is any shape or form.

IV. Work Plan

Due to the magnitude of the research project in relation to other projects its important to develop a work plan to ensure quality of each task by spending adequate time. To help implement this a Gantt chart below (figure.6) has been produced to help manage the timeline for the project. The project is due to start in June 2021 and has an end deadline of 1st of October 2021. I have divided the tasks required into six steps, giving each step more than enough adequate time taking into consideration any risk that may arise and cause delays. More information about some of the tasks below.



Month		June				July				August				September			
Week		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Task	Literature Review	🟩	🟩	🟩	🟩												
	Data Processing					🟩	🟩										
	Forecasting Model							🟩	🟩								
	DTW Model									🟩	🟩						
	Write Up											🟩	🟩	🟩	🟩		
	Proofing															🟩	🟩

Figure.6

Literature Review

- Update, amend and confirm the scope of the project if required.
- In-depth reading on the algorithms, understanding the mathematics behind them and finalise choices.
- Ensure the dataset meets the requirements of the project aim, if not find an alternative or better dataset.

Result – Potential update to the project proposal

Data Processing

- Source the data.
- Explore and clean the data.
- Prepare data for analysis. This including singling out current athletes who have time to develop.

Result – Final data frame and its visual representation.

Forecasting Model

- Feature engineer a new metric to be the basis of the forecasting model.
- Apply the forecasting model to the dataset.
- Improve the forecasting model.

Result – Final model forecasting athletes' future development.

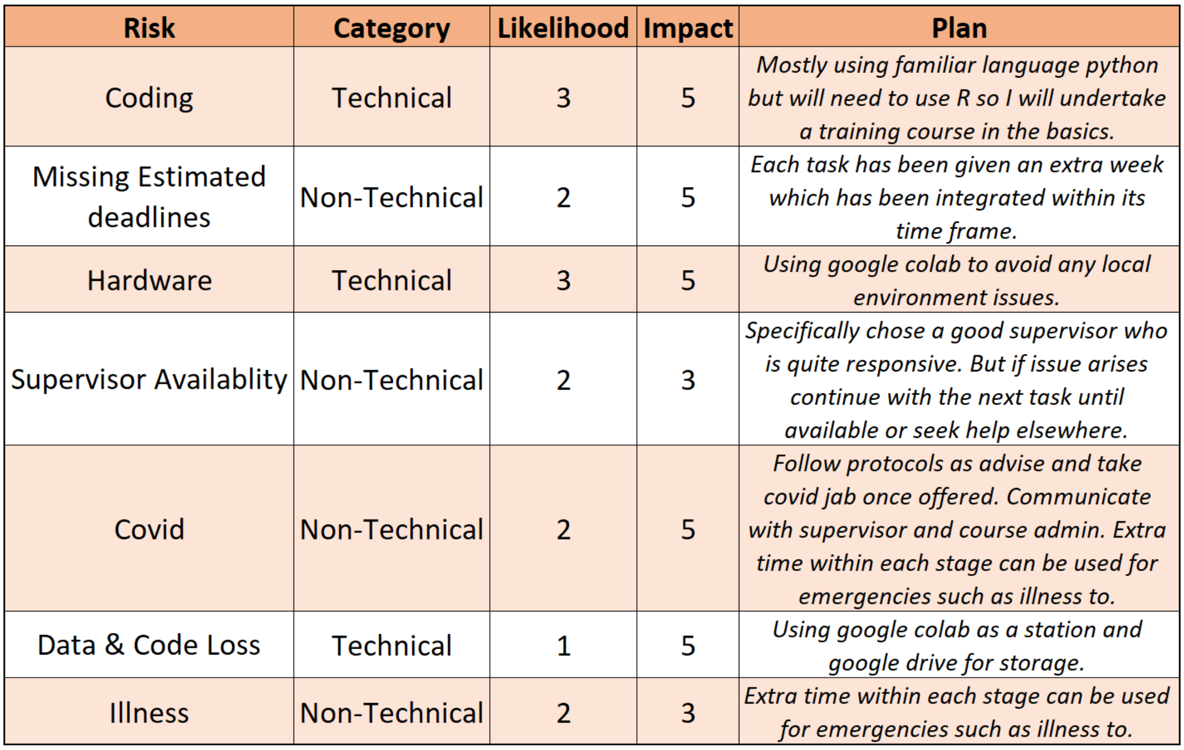
DTW Model

- Applying the DTW clustering model to the dataset.
- Improve the clustering model.
- Profile the resulting clusters.

Result – Final model clustering athletes forecasted development into profiled groups.

V. Risks

For this section we register each potential risk. For each risk we set the category whether it's a technical issue or non-technical, likelihood of this issue occurring with a rating between 1 to 5, where the latter meaning 100% Likelihood. We will also look at the impact of each risk with a given rating between 1 to 5, with the latter meaning damaging impact. Finally, we look at what plan we have in place to avoid these issues.



Risk	Category	Likelihood	Impact	Plan
Coding	Technical	3	5	*Mostly using familiar language python but will need to use R so I will undertake a training course in the basics.*
Missing Estimated deadlines	Non-Technical	2	5	*Each task has been given an extra week which has been integrated within its time frame.*
Hardware	Technical	3	5	*Using google colab to avoid any local environment issues.*
Supervisor Availablity	Non-Technical	2	3	*Specifically chose a good supervisor who is quite responsive. But if issue arises continue with the next task until available or seek help elsewhere.*
Covid	Non-Technical	2	5	*Follow protocols as advise and take covid jab once offered. Communicate with supervisor and course admin. Extra time within each stage can be used for emergencies such as illness to.*
Data & Code Loss	Technical	1	5	*Using google colab as a station and google drive for storage.*
Illness	Non-Technical	2	3	*Extra time within each stage can be used for emergencies such as illness to.*

VI. References

[1] Vaeyens, R., Lenoir, M. and Williams, A.M., 2008. *Talent Identification and Development Programmes in Sport*, sports med 38.

[2] Lazarević, S., Lukić, J. and Mirkovic, V., 2020. *Role of Football Scouts in Player Transformation Process: From Talented to Elite Athlete*.

[3] Szczepański, Ł., 2015. *Assessing the skill of football players using statistical methods*.

[4] Gaurav, V. and Chakraborty, G., 2019. *Scouting in Soccer with Applied Machine Learning*.

[5] Kuper, S. (2013). *Everton: how the blues made good*. http://www.ft.com/cms/s/2/2fa7ef1e-b2c0-11e2-8540-00144feabdc0.

[6] Luan, Z., 2021. *Big Data Prediction of Sports Injury Based on Random Forest Algorithm and Computer Simulation, Microprocessors and Microsystems*.

[7] Arunraj, N., Ahrens, D. and Fernandes, M., 2016. *Application of SARIMAX Model to Forecast Daily Sales in Food Retail Industry*. International Journal of Operations Research and Information Systems.

[8] Herbig, P. a., Milewicz, J. and Golden, J. E., 1993. *The Do's and Don'ts of Sales Forecasting*. Industrial Marketing Management.

[9] Lee, M. and Hamzah, N., 2010. *Calendar Variation Model Based on ARIMAX for Forecasting Sales Data with Ramadhan Effect*. Proceedings of the Regional.

[10] Liu, L.-M., Bhattacharyya, S., Sclove, S. L., Chen, R., and Lattyak, W. J., 2001. *Data Mining on Time Series: An Illustration Using Fast-Food Restaurant Franchise Data*. Computational Statistics & Data Analysis.

[11] Shukla, M. and Jharkharia, S., 201. *Applicability of ARIMA Models in Wholesale Vegetable Market: An Investigation*. International Journal of Information Systems and Supply Chain Management.

[12] Nontapa, C., Kesamoon, C., Kaewhawong, N. and Intrapaiboon, P., 2020. *A New Time Series Forecasting Using Decomposition Method with SARIMAX Model*. In: Yang, H., Pasupa, K., Leung, A.CS., Kwok, J.T., Chan, J.H. and King, I. Neural Information Processing. ICONIP 2020. Communications in Computer and Information Science, vol 1333. Springer, Cham.

[13] Cools, M., Moons, E. and Wets, G., 2009. *Investigating Variability in Daily Traffic Counts Using ARIMAX and SARIMA (X) Models: Assessing Impact of Holidays on Two Divergent Site Locations*, (X).

[14] Agnew, M., and Thornes, J., 1995. *The Weather Sensitivity of the UK Food Retail and Distribution Industry*. Meteorological Applications.

[15] Chikobvu, D., and Sigauke, C., 2012. *Regression-SARIMA Modelling of Daily Peak Electricity Demand in South Africa*. Journal of Energy in South Africa.

[16] Trancart, T., Acou, A., De Oliveira, E., and Feunteun, E., 2013. *Forecasting Animal Migration Using SARIMAX: An Efficient Means of Reducing Silver Eel Mortality Caused by Turbines*. Endangered Species Research.

[17] Monthly Report, 2013. *Current and projected development of coin circulation in Germany*. Deutsche Bundesbank. https://www.bundesbank.de/en/publications/reports/monthly-reports/monthly-report-january-2013-670526

[18] Senin, P., 2009. *Dynamic Time Warping Algorithm Review*.

[19] Sarda-Espinosa. A., 2019. *Package ‘dtwclust’*. https://cran.r-project.org/web/packages/dtwclust/dtwclust.pdf

Research Ethics Review Form: BSc, MSc and MA Projects
Computer Science Research Ethics Committee (CSREC)
http://www.city.ac.uk/department-computer-science/research-ethics

Undergraduate and postgraduate students undertaking their final project in the Department of Computer Science are required to consider the ethics of their project work and to ensure that it complies with research ethics guidelines. In some cases, a project will need approval from an ethics committee before it can proceed. Usually, but not always, this will be because the student is involving other people (“participants”) in the project.

In order to ensure that appropriate consideration is given to ethical issues, all students must complete this form and attach it to their project proposal document. There are two parts:

PART A: Ethics Checklist. All students must complete this part.
The checklist identifies whether the project requires ethical approval and, if so, where to apply for approval.

PART B: Ethics Proportionate Review Form*.* Students who have answered “no” to all questions in A1, A2 and A3 and “yes” to question 4 in A4 in the ethics checklist must complete this part. The project supervisor has delegated authority to provide approval in such cases that are considered to involve MINIMAL risk. The approval may be ***provisional*** *– identifying the planned research as* likely to involve MINIMAL RISK. In such cases you must additionally seek ***full approval*** from the supervisor as the project progresses and details are established. ***Full approval*** must be acquired in writing, before beginning the planned research.

A.1 If you answer YES to any of the questions in this block, you must apply to an appropriate external ethics committee for approval and log this approval as an External Application through Research Ethics Online - https://ethics.city.ac.uk/		*Delete as appropriate*
1.1	Does your research require approval from the National Research Ethics Service (NRES)? *e.g. because you are recruiting current NHS patients or staff?* *If you are unsure try - https://www.hra.nhs.uk/approvals-amendments/what-approvals-do-i-need/*	**NO**
1.2	Will you recruit participants who fall under the auspices of the Mental Capacity Act? *Such research needs to be approved by an external ethics committee such as NRES or the Social Care Research Ethics Committee - http://www.scie.org.uk/research/ethics-committee/*	**NO**
1.3	Will you recruit any participants who are currently under the auspices of the Criminal Justice System, for example, but not limited to, people on remand, prisoners and those on probation? *Such research needs to be authorised by the ethics approval system of the National Offender Management Service.*	**NO**
A.2 If you answer YES to any of the questions in this block, then unless you are applying to an external ethics committee, you must apply for approval from the Senate Research Ethics Committee (SREC) through Research Ethics Online -		*Delete as appropriate*

https://ethics.city.ac.uk/		
2.1	Does your research involve participants who are unable to give informed consent? *For example, but not limited to, people who may have a degree of learning disability or mental health problem, that means they are unable to make an informed decision on their own behalf.*	**NO**
2.2	Is there a risk that your research might lead to disclosures from participants concerning their involvement in illegal activities?	**NO**
2.3	Is there a risk that obscene and or illegal material may need to be accessed for your research study (including online content and other material)?	**NO**
2.4	Does your project involve participants disclosing information about special category or sensitive subjects? *For example, but not limited to: racial or ethnic origin; political opinions; religious beliefs; trade union membership; physical or mental health; sexual life; criminal offences and proceedings*	**NO**
2.5	Does your research involve you travelling to another country outside of the UK, where the Foreign & Commonwealth Office has issued a travel warning that affects the area in which you will study? *Please check the latest guidance from the FCO - http://www.fco.gov.uk/en/*	**NO**
2.6	Does your research involve invasive or intrusive procedures? *These may include, but are not limited to, electrical stimulation, heat, cold or bruising.*	**NO**
2.7	Does your research involve animals?	**NO**
2.8	Does your research involve the administration of drugs, placebos or other substances to study participants?	**NO**
A.3 If you answer YES to any of the questions in this block, then unless you are applying to an external ethics committee or the SREC, you must apply for approval from the Computer Science Research Ethics Committee (CSREC) through Research Ethics Online - https://ethics.city.ac.uk/ **Depending on the level of risk associated with your application, it may be referred to the Senate Research Ethics Committee.**		*Delete as appropriate*
3.1	Does your research involve participants who are under the age of 18?	**NO**
3.2	Does your research involve adults who are vulnerable because of their social, psychological or medical circumstances (vulnerable adults)? *This includes adults with cognitive and / or learning disabilities, adults with physical disabilities and older people.*	**NO**
3.3	Are participants recruited because they are staff or students of City, University of London? *For example, students studying on a particular course or module.*	**NO**

	If yes, then approval is also required from the Head of Department or Programme Director.	
3.4	Does your research involve intentional deception of participants?	**NO**
3.5	Does your research involve participants taking part without their informed consent?	**NO**
3.5	Is the risk posed to participants greater than that in normal working life?	**NO**
3.7	Is the risk posed to you, the researcher(s), greater than that in normal working life?	**NO**
A.4 If you answer YES to the following question and your answers to all other questions in sections A1, A2 and A3 are NO, then your project is deemed to be of MINIMAL RISK. **If this is the case, then you can apply for approval through your supervisor under PROPORTIONATE REVIEW. You do so by completing PART B of this form.** **If you have answered NO to all questions on this form, then your project does not require ethical approval. You should submit and retain this form as evidence of this.**		*Delete as appropriate*
4	Does your project involve human participants or their identifiable personal data? *For example, as interviewees, respondents to a survey or participants in testing.*	**NO**