# Methods

Past tense

Improvements

* Define paramaters for models, this can be done by printing the models used in the workflow to show the paramaters
* Tom’s input
* Include VM details
* Anything that requires more details tagged with (det)

## General computational details

The workflow constructed was constructed in the jupyter-lab environment using the python programming language(ref). The computationally expensive hyperparameter optimisation and model validations were done using a virtual machine hosted on Azure. The virtual machine employed 8 GB of random access memory (det).

## Data collection method

For the data visualisations and machine learning pipeline we used data from mass-spectrometry analysis of flint samples collected from a number of sites around the UK. (det)

**\*Tom\***

## Sites sampled from

**\*Tom\***

Laser ablation

Laser info details

## Dimensionality reduction techniques

Prior to application of machine learning classifiers to the data we had visualised the structure of the mass-spectrometry data using two dimensionality reduction techniques. These were Principal Component Analysis and t-distributed stochastic neighbour embedding. The structure of the ion abundance data was visualised in two dimensions *via* the dimensionality reduction techniques; Principal Component Analysis and t-distributed stochastic neighbour embedding.

### t-distributed stochastic neighbour embedding

t-SNE was used to alongside PCA to interrogate whether data from different flint sites were distinct from other sites. Visual inspection of the T-SNE projections also allowed for observation of outliers. The t-SNE utilised 750 iterations.

## Machine learning models overview

Supervised Machine learning classifiers were used to map profiles of ion abundances of flint samples to specific named bedrock and superficial sites sourced in the United Kingdom. The models were trained on the optimum features as determined by the feature selection process. The labels into which samples were classified as were either single bedrock/superficial sites or groups of sites.

## Preproccessing and feature selection

### Data preproccessing

**\*Tom\* - what prepoccessing did you do?**

Missing values were imputed with the mean average for that feature. Outliers were defined as values that exceeded 2 times the standard deviation from the mean average for that feature. Outliers were replaced with the mean plus or minus two times the standard deviation. Outliers were also identified by graphical analysis using t-SNE visualisations. These outliers were not used in the models.

### Feature selection

The goal of all model optimisations was to maximise the weighted-F1 score (det). The optimum features were identified by the implementation of recursive feature elimination (RFE) with Random Forest Classifiers. Random Forest Classifiers were chosen because feature importance scores for each feature can be derived from the model. (explanation of feature importance). RFE is an iterative process whereby multiple models are sequentially built. The process builds a model with all features then drops the feature with the lowest ‘feature importance’ score. The feature importance score is a measure of the predictive power for that feature. In random forests the level within the decision trees at which the feature is utilised for splitting the data determines how many observations will utilise that feature. A feature utilised higher up within the decision tress is deemed as more important because more observations are split with reference to that feature (det). Models are iteratively built not including the least important features as calculated from the previous models.

Each model was evaluated by 3-fold stratified cross-validation with weighted-F1 score. The mean averages for each stage of RFE were visualised against all the feature selections and the feature combination with the highest weighted-F1 score was chosen for input into all subsequent models.

## Novelty detection

Due to the fact that we had not sampled all flint sites in the UK and that flint may have been sourced from outside the UK we had always assumed that a proportion of the flint artefacts may have not been sourced from any of the sites that we had collected samples from. It was therefore necessary to identify artefacts that were likely to have been sourced from a flint deposit not sampled in this study to prevent erroneous classification of artefacts. The method indirectly permitted the formation of an ‘other’ class. The Local Outlier Factor model was used for this purpose. (det) Features were standardized so each feature has a mean value of 0 and an unit variance.

Samples that were classified as outliers were not classified by the final classifier but instead classified as ‘other’.

**\*print details of model\***

## Learning Curves

In order to assess the impact of collecting more data for future work training curves were created and assessed. Random forest classifiers were built and evaluated on increasingly larger training datasets, each model was evaluated by stratified x-fold cross-validation with weighted-F1 scores.

## Machine learning models

The performance of 4 different classifiers were assessed. These were Random Forest, Support Vector Machine, Gradient Boosting machine and K-nearest neighbours. (det)

Random Forest outperformed all other models evaluated and so was used as the final model.

## Model evaluation and hyperparameter optimisation

All 4 models were extensively evaluated to compare performances. To assess performance the dataset was split into 80% training and 20% test data randomly 100 times. This process was stratified so that the proportions of classes in the training data was representative of the proportions within the entire dataset. Hyperparamater optimisation was done on the train data fold by 5-fold stratified cross-validation. The original 80% train data was then used as input into a model which was paramaterised with the optimum hyperparamaters. The model was evaluated by weighted-F1 score by comparison of the predictions against the test data labels. The model performances were compared by the visualisation of the 100 weighted-f1 scores *via*  boxplots.

## Data Availability

\*follow guidelines for what to put here