# Results

## Data sourcing

\*Tom’s section\*

* How many and where flint samples are from?
* Mass spectrometry details, laser ablation etc. (not too detailed, bulk goes in methods)
* Ladr processing (not too detailed, bulk goes in methods)

## Visualisation of structure within data

* Why visualise structure of data
* What data was used, bedrock, superficial, both, artefacts etc?

It is difficult to get an understanding of the structure of multivariate data. Dimensionality reduction techniques allow the visualisation of underlying structure of multivariate data. If the data is colour-coded then the distinctness between different sites or flint sites can be visually inspected. The dimensionality reduction techniques Principal Component Analysis and t-SNE were utilised. Four datasets were visualised in this way, bedrock and superficial together, bedrock alone, superficial alone and all samples plus the artefacts.

Show plots, discuss how superficial data not well separated so guided formation of classes eg some hybrids.

### Principal Component Analysis

* Introduction to what PCA is used for
* Do groups resolve well in both techniques? Which groups resolve best/worst? Dies this link back to domain knowledge?

### t-distributed stochastic neighbour embedding

* Introduction to what t-SNE is used for
* Do groups resolve well in both techniques? Which groups resolve best/worst? Dies this link back to domain knowledge?

## Feature selection

* Why do feature selection and what is feature selection?
* Explain method used
* Which features used and why based on graph

The abundances of 53 different elements were measured by mass-spectrometry for each sample and flint artefact. Each of these element abundances could be used as a feature to be inputted into the final model. However, more features doesn’tnecessarily equate to a more predictive model. Some features may just add noise to the model and so could potentially decrease the model’s accuracy. In order to input the best set of features into the model a process known as feature selection was used. There are many ways to evaluate the predictive power of individual and groups of features. In this study recursive feature elimination (RFE) was used. The process of RFE iteratively builds models one at a time, for each model the feature with the lowest score is not included in any future model builds. Each model can be evaluated and then the set of features that gave rise to the best evaluated model can be identified. In this case the models used were random forest classifiers and the metrics for evaluating the feature to be dropped was feature importance. RFE demonstrated that the best features were (det)

\*\*\*Tom\*\*\*

Perhaps you can give a bit of context as to why it makes sense these elements were most predictive, perhaps in the context of flint geology?

## Model comparisons

## Novelty detection

* Outline problem aka only sampled from number of sites within UK, what if artefacts came from sites not sampled?
* With any of these models all artefacts would be classified into the sites sampled from so need method of identifying which artefacts cannot be classified into these sites.
* Explain what local outlier factor model is
* How many artefacts could be classified and what proportion is this of total artefacts
* The non-classifiable artefacts were classified as ‘other’

It was very likely that a proportion of the artefacts to be analysed were not originally sourced from sites that had been sampled in this study. For example, some artefacts may have been sourced from outside the UK or from bedrock sites within the UK not sampled by us.

\*\*\*Tom\*\*\*

Could you expand on this and talk about some popular sites for flint that we had not sampled.?

All 4 models evaluated in this study would classify all artefacts into classes associated with sites that had been sampled for this study. A proportion of tehse classifications would therefore be incorrect. There are different metrics that can be used to quantify model classification confidence such as (det) list examples eg the probabilities outputted by random forest model. However, it would be difficult to set a threshold for which artefacts can be said to confidently be sourced from one of the sites sampled.

The local outlier factor model is a model that can classify observations as outliers given examples of inliers. This model was fitted to the training data aka the flint samples then used to predict the artefact samples as inliers or outliers. Outliers were classified into the site ‘other’ inliers were carried forward to be classified by the final model.

## Final model evaluation

* Which model used and why
* How was it evaluated
* Overall F1-score
* Class-specific F1 scores
* Feature importances, link in with domain knowledge?

Four different machine learning models were assessed for their suitability in classifying artefacts to their source. The models, Random Forest, Gradient Boosting Machine, K-nearest Neighbours and Support Vector Machine were each evaluated. The performance of each model was tested by 100-fold cross-validation. The dataset was split into 80% train data and 20% test data. Hyperparamaters were optimised using the 80% train data, the final model was then evaluated by comparing the predictions against the ground truth in the test data. The boxplots show (det).

The Random Forest Model outperformed all the other models tested. The median weighted-F1 score was higher than all other models.

Explanation of class-specific F1-scores(det)

\*\*\*\*Tom\*\*\*

Feature importances, link in with domain knowledge?

## Classification of artefact sourcing

* Final model built and artefacts classified
* \*Tom how deal with probabilities at end\*

The Random Forest was taken forward as the model of choice for classifying the artefacts. A more detailed evaluation of the Random Forest’s performance was completed. 10-fold stratified cross validation was done. The stratified nature of the train test split ensured the train folds contained representative proportions of the different classes.

\*\*\*Talk about results\*\*\*

## Implications for archeology

\*Tom’s section\*

* Drop the mic.
* Compare to methods within PhD, how improved?