Predicting age of rocks using microfossils: Volve Field case-study

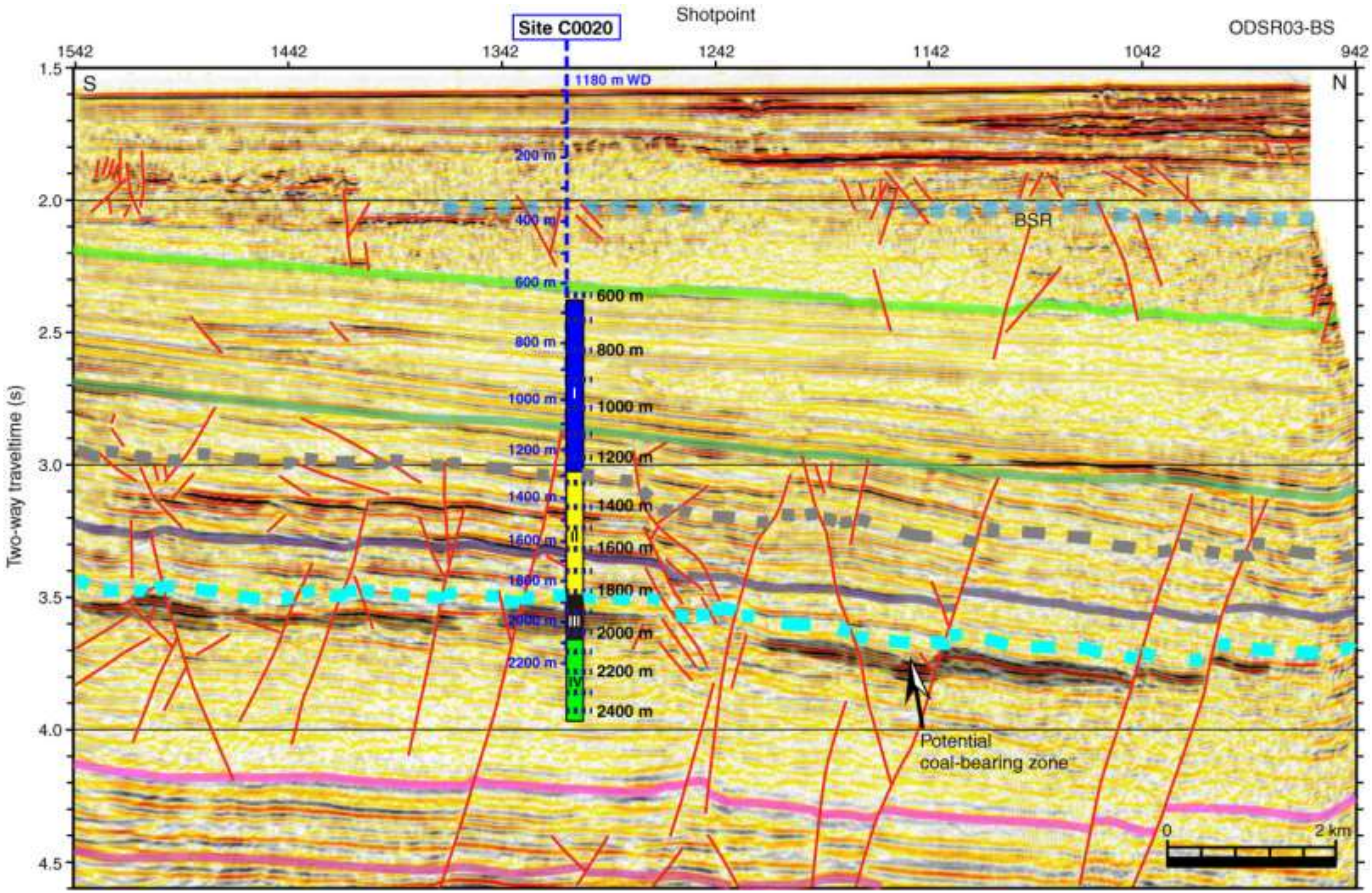
Guy Harrington

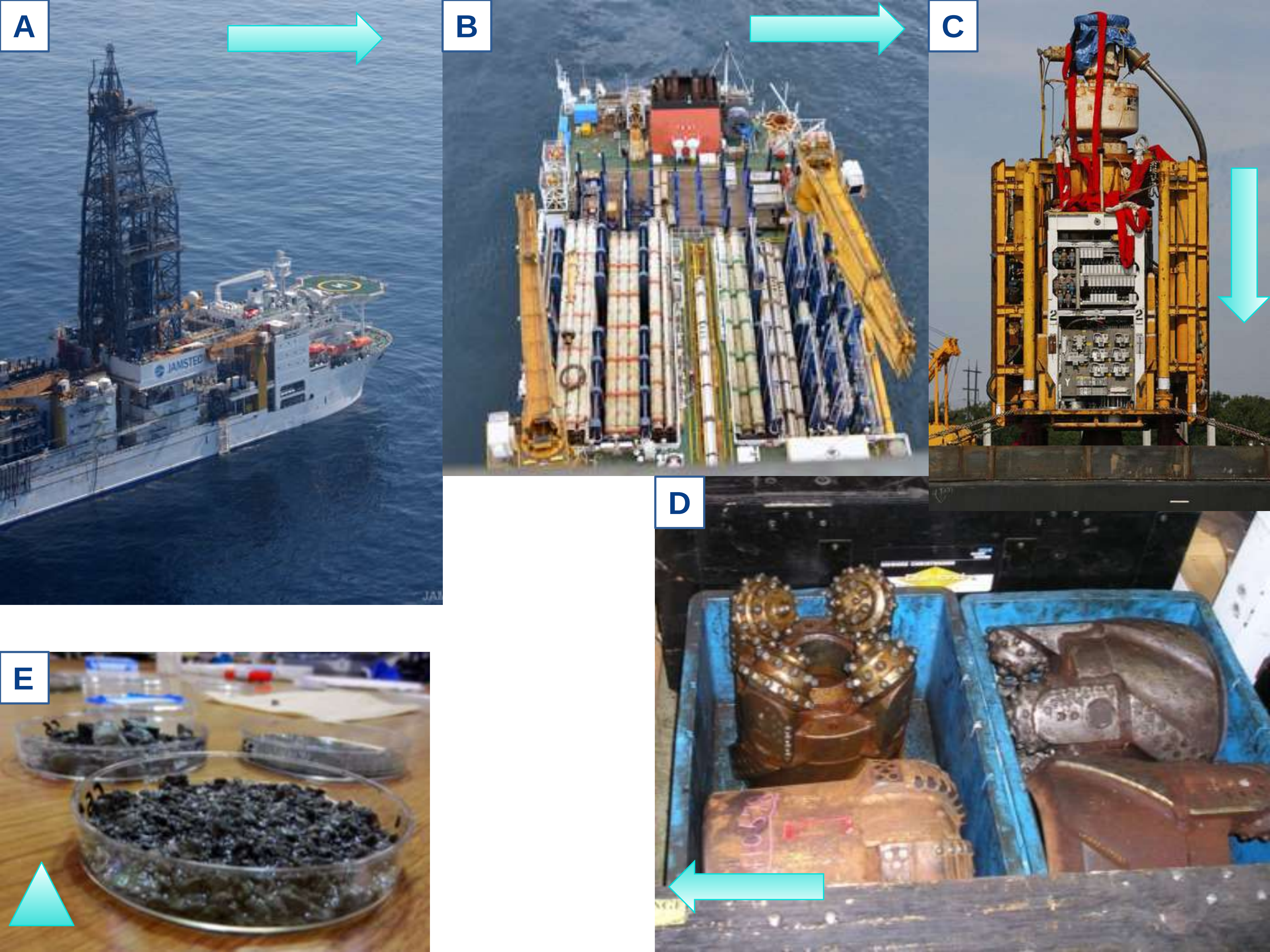
https://github.com/jqbirm75/Unit-7-Capstone

What’s this about?

- **Types of data** collected by oil exploration
  - Biostratigraphic data (= microfossils) and collection
- **Tuning models**
  - What models work with these sparse count data?
  - How best to combat small class sizes and imbalance.
- Ensemble approach – **can we predict age and geological stages using fossil data with the minimum of cleaning and feature engineering?**

Seismic profiles

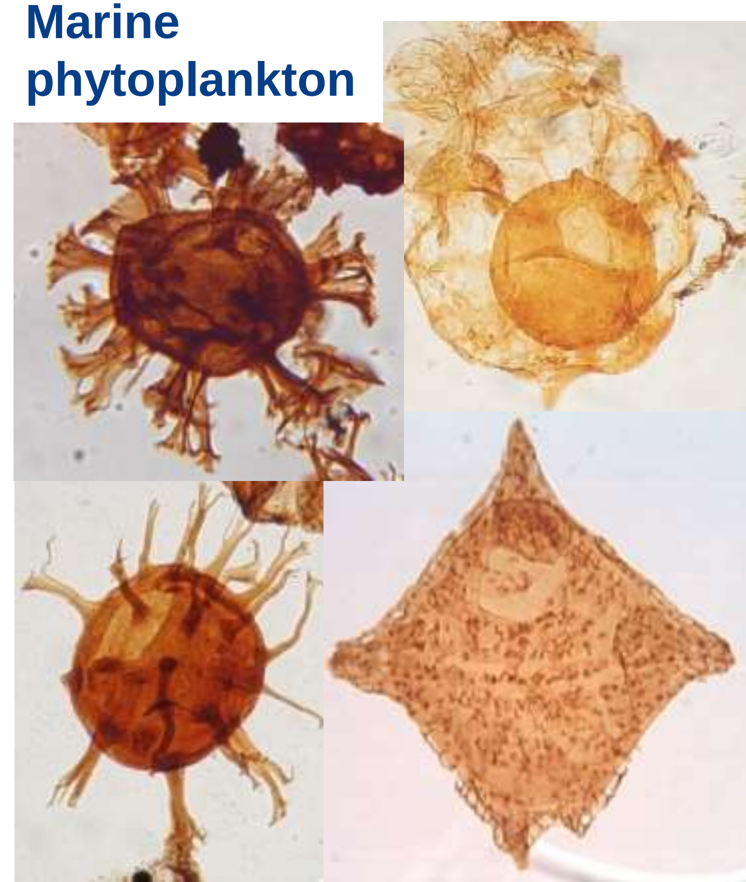



A
B
C
D
E
JAMSTEC

Palynomorphs

pollen & spores

Marine phytoplankton



* Age determination based on fossil components (counts). Analogous to **fingerprints**. Different fossil groups calibrate timescales and zones.

* Some groups are cosmopolitan (e.g. plankton), others are highly regional (terrestrial fossils). ∴ Can determine:

1) **Age**
2) **Provenance**

Analogy

Elizabethan (c.1560-1600)




Victorian (c.1840-1900)



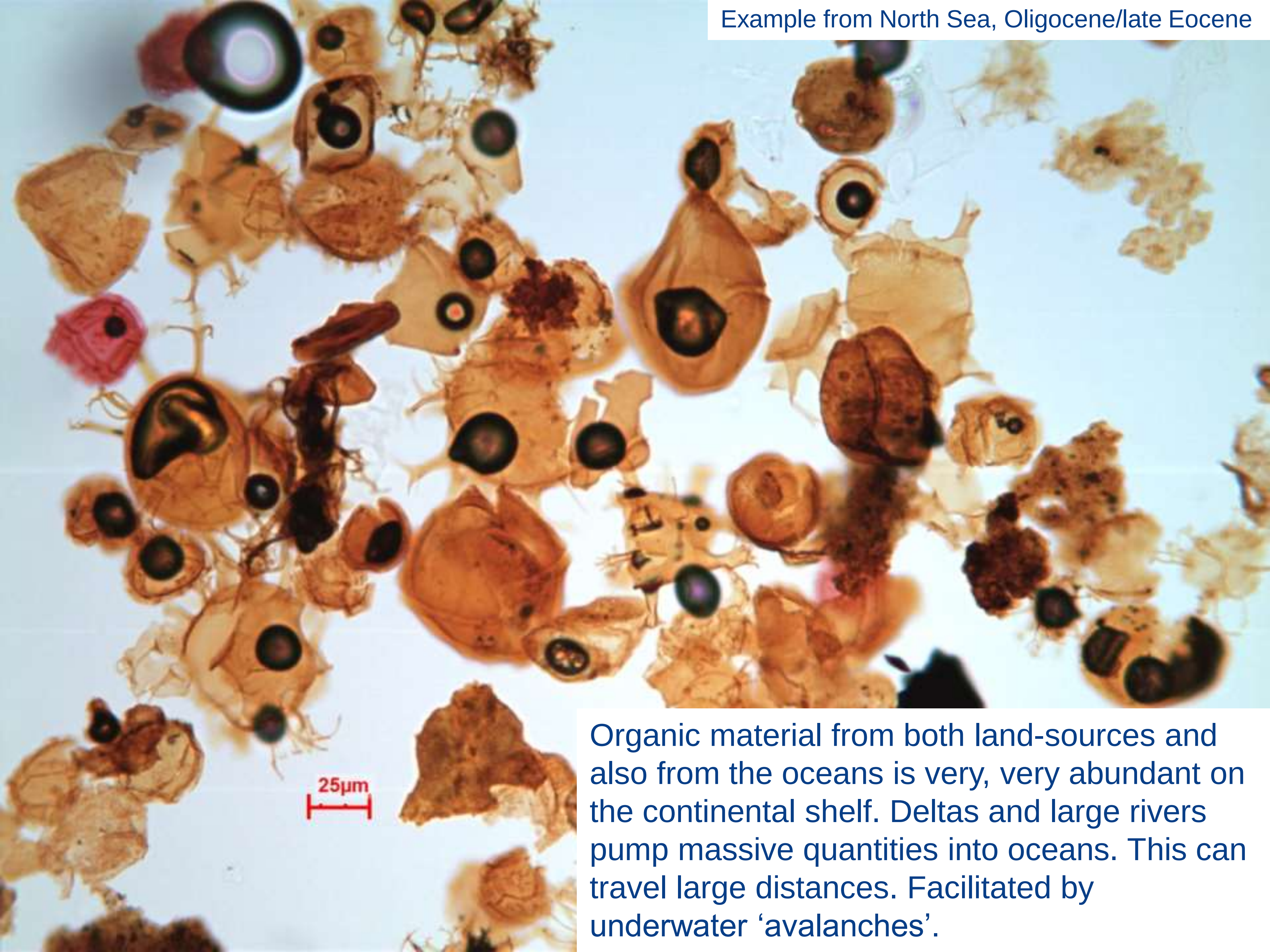

Caving

Contamination – falling rocks

Reworking

Recycling of old rocks and fossils

Example from North Sea, Oligocene/late Eocene



Organic material from both land-sources and also from the oceans is very, very abundant on the continental shelf. Deltas and large rivers pump massive quantities into oceans. This can travel large distances. Facilitated by underwater ‘avalanches’.

What’s the challenge?

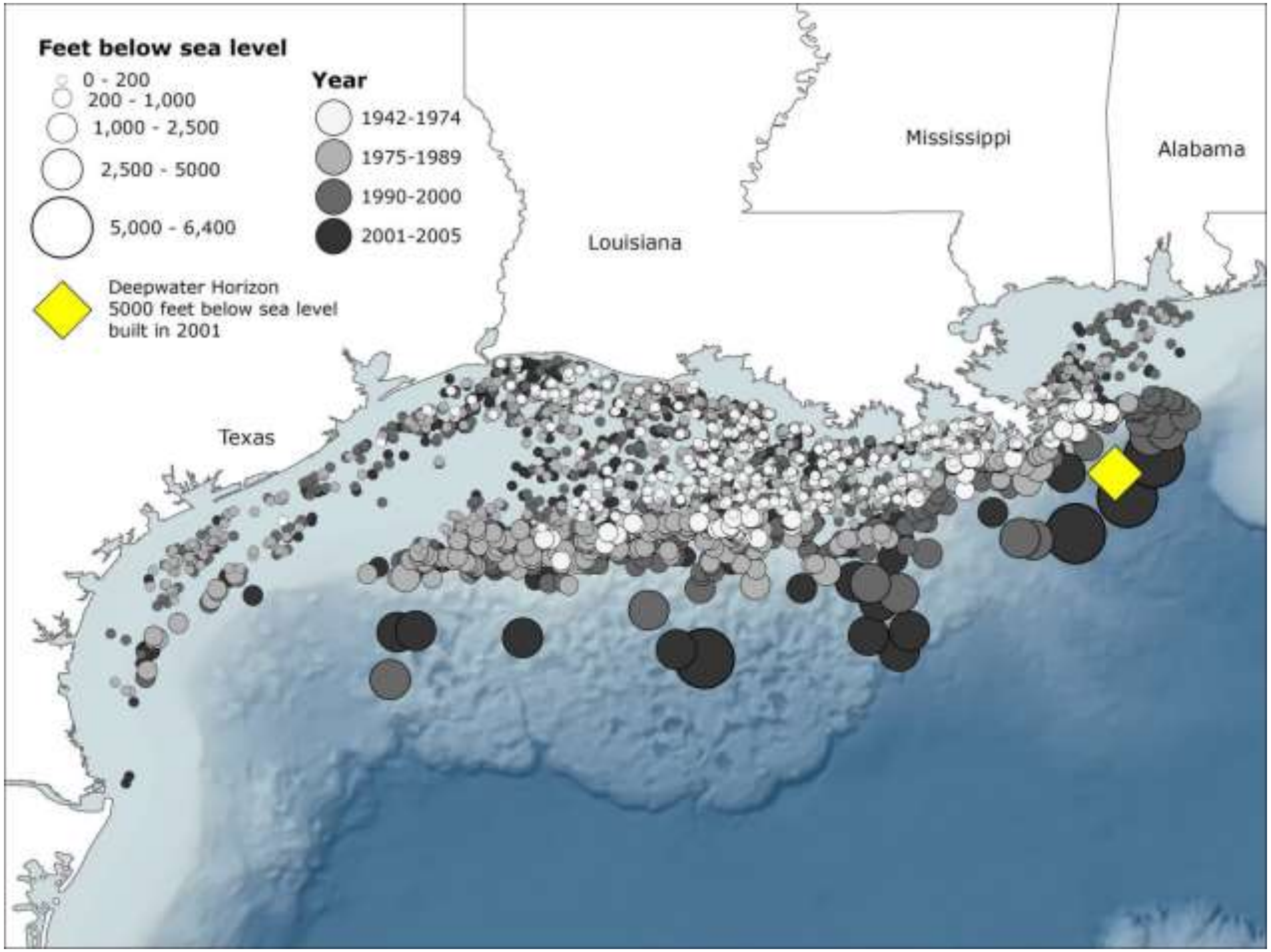
Company viewpoints

- **Company A:** “This kind of stuff is really starting to get lots of attention in industry …With respect to biostrat …, I think that the scope is more limited given the nature of our data.”

- **Company B:** “with regards to biostrat … the results weren't particularly encouraging”

- **Company C:** ”I feel limited success with biostrat as the focus has been on automation of identification and basically there are not enough biostratigraphers around to have significant impact.

Why bother?



The quantity of data!

Data

Volve dataset

Data released for an oil field off Norway by Equinor (formerly Statoil)

Rare opportunity to study industry data from several wells. 13 different wells have fossil data (termed biostratigraphic data).

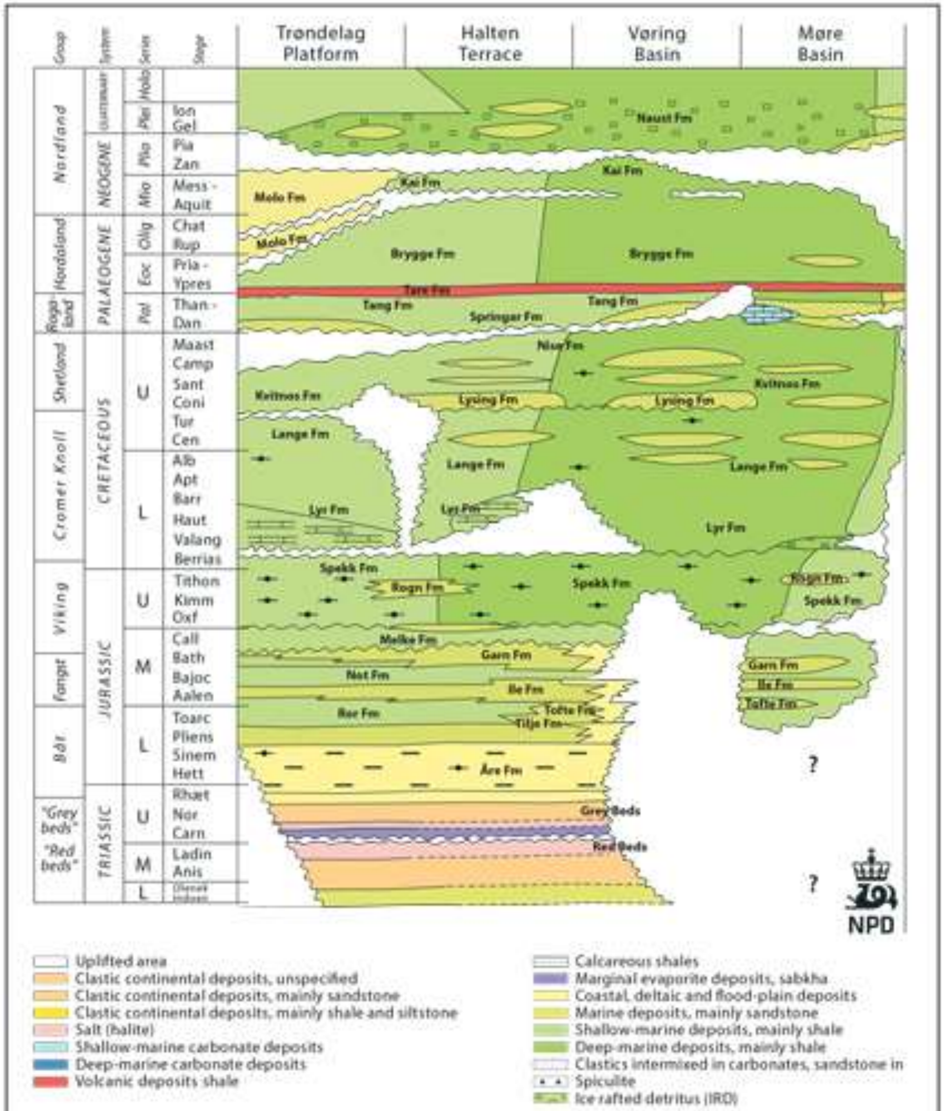
Data in various formats. Biostratigraphic data are in the form of .dex files.

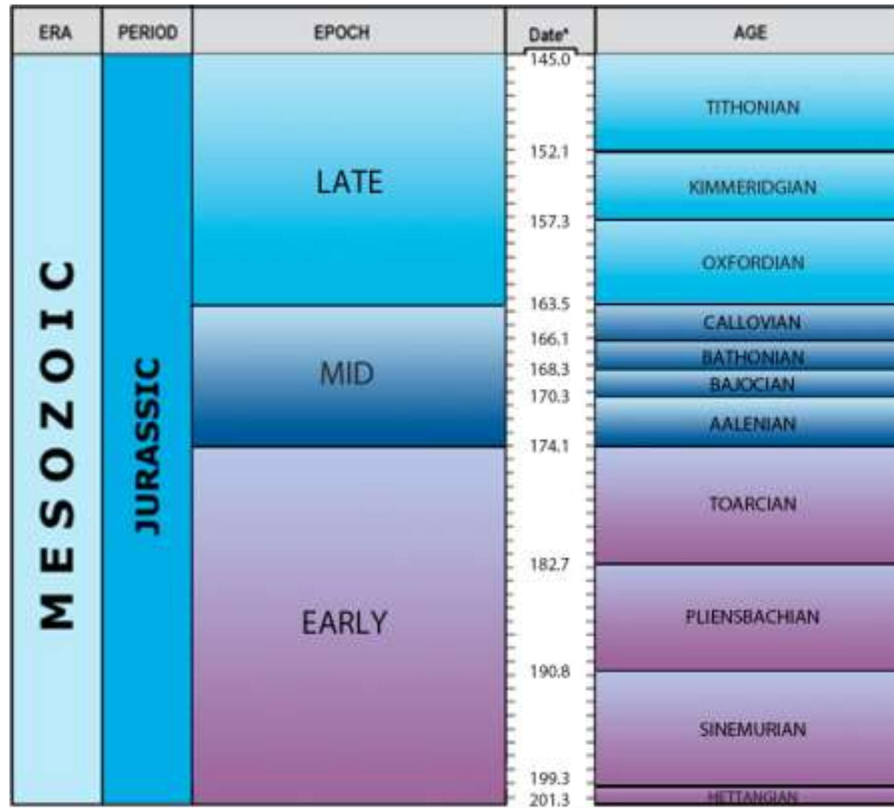
Other data types include seismic profiles and geochemical and physical data.

www.equinor.com

Location: c. 200 kilometres west of Stavanger, Norway
Production start: 12 February 2008
Production end: 17 September 2016

Time period





www.bgs.ac.uk

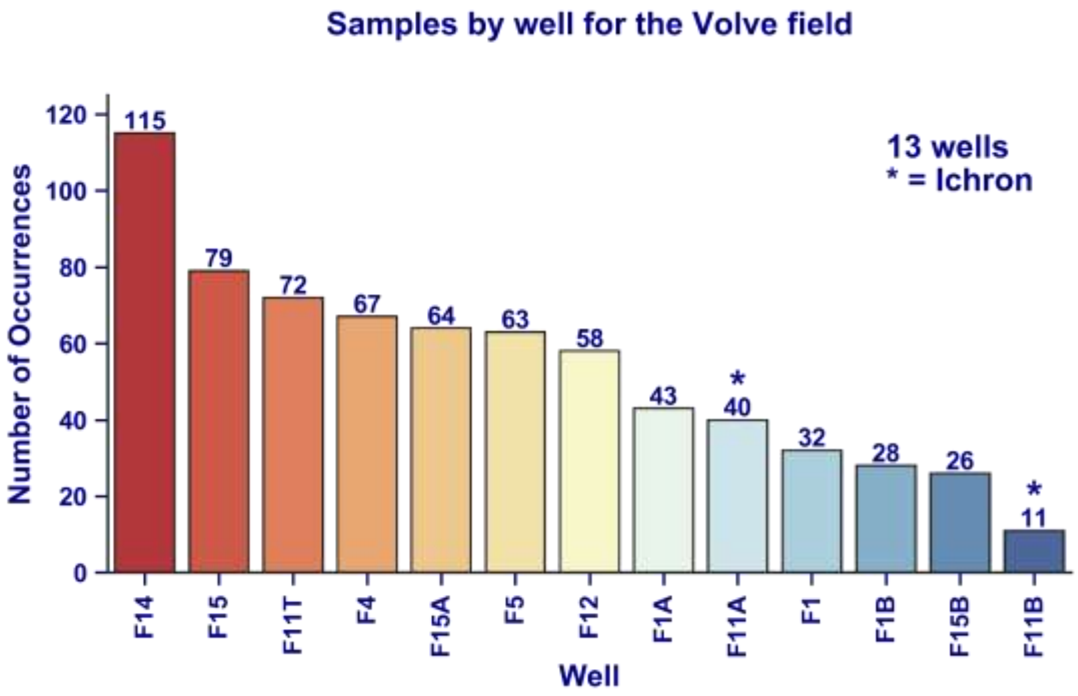
Features

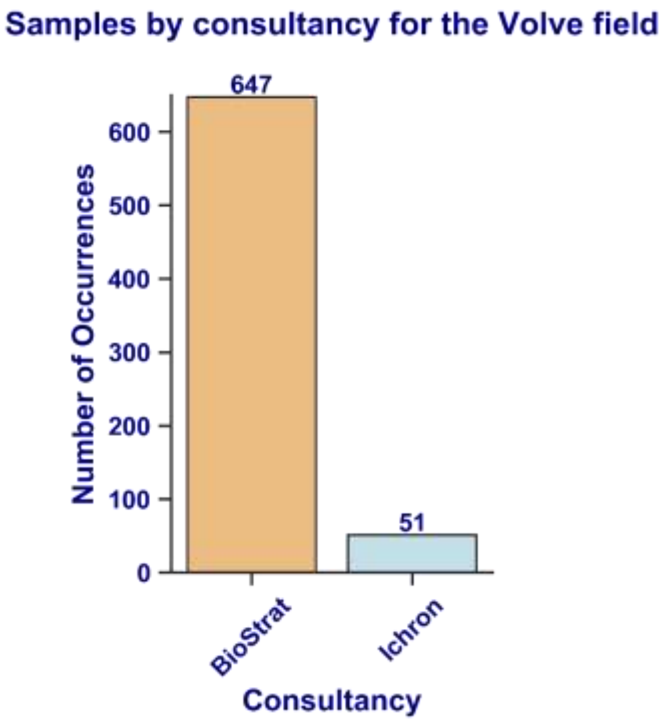
Labels and added features

AGE	Major age division
SITE	Name of well
STAGE	Subdivisions of 'Age'
FORMATION	Depositional group
ZONE	Fossil zone
ANALYST	Who counted the data
AXIS_1	PCoA axis 1 score
AXIS_2	PCoA axis 2 score
AXIS_3	PCoA axis 3 score
N	Count size
S	Raw species richness
R100	Rarefied richness at 100
E1/D	Simpson index

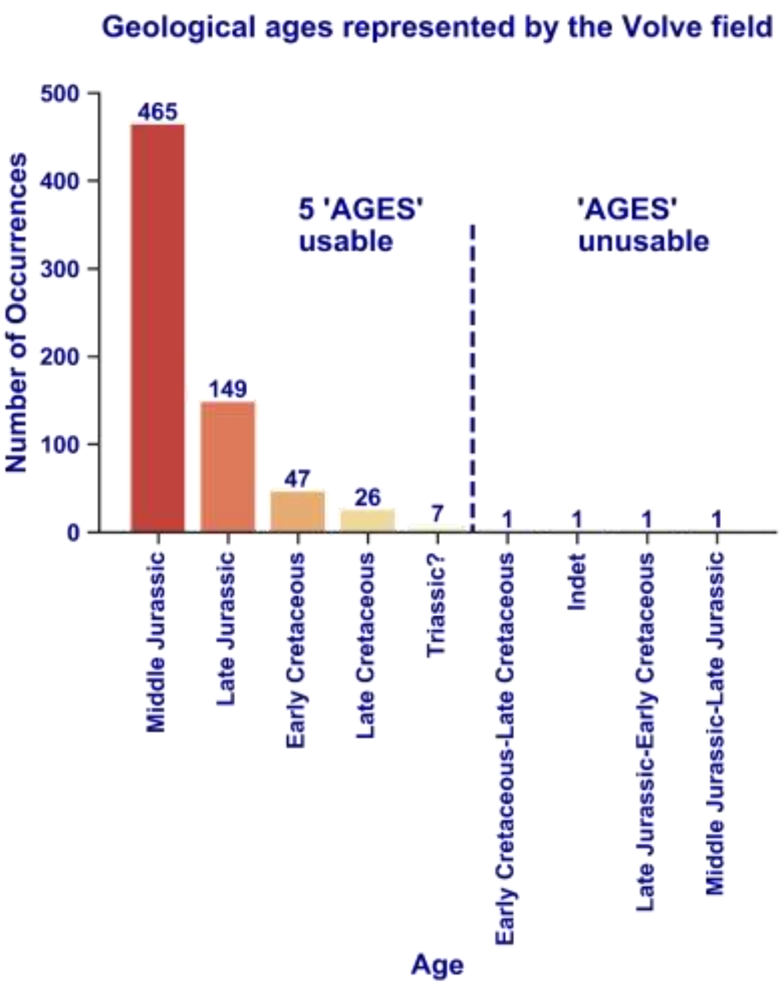
- Counts of major fossil groups
- Whole spreadsheet = 581 taxa and 764 samples.
- Cut to 422 taxa and 694 samples:
  - Cleaned then processed in R using *CullMatrix* package.
  - PCoA run in R using *Vegan* package. Works on a distance metric (Bray-Curtis).
  - Counts scaled using Hellinger transformation.

Data exploration





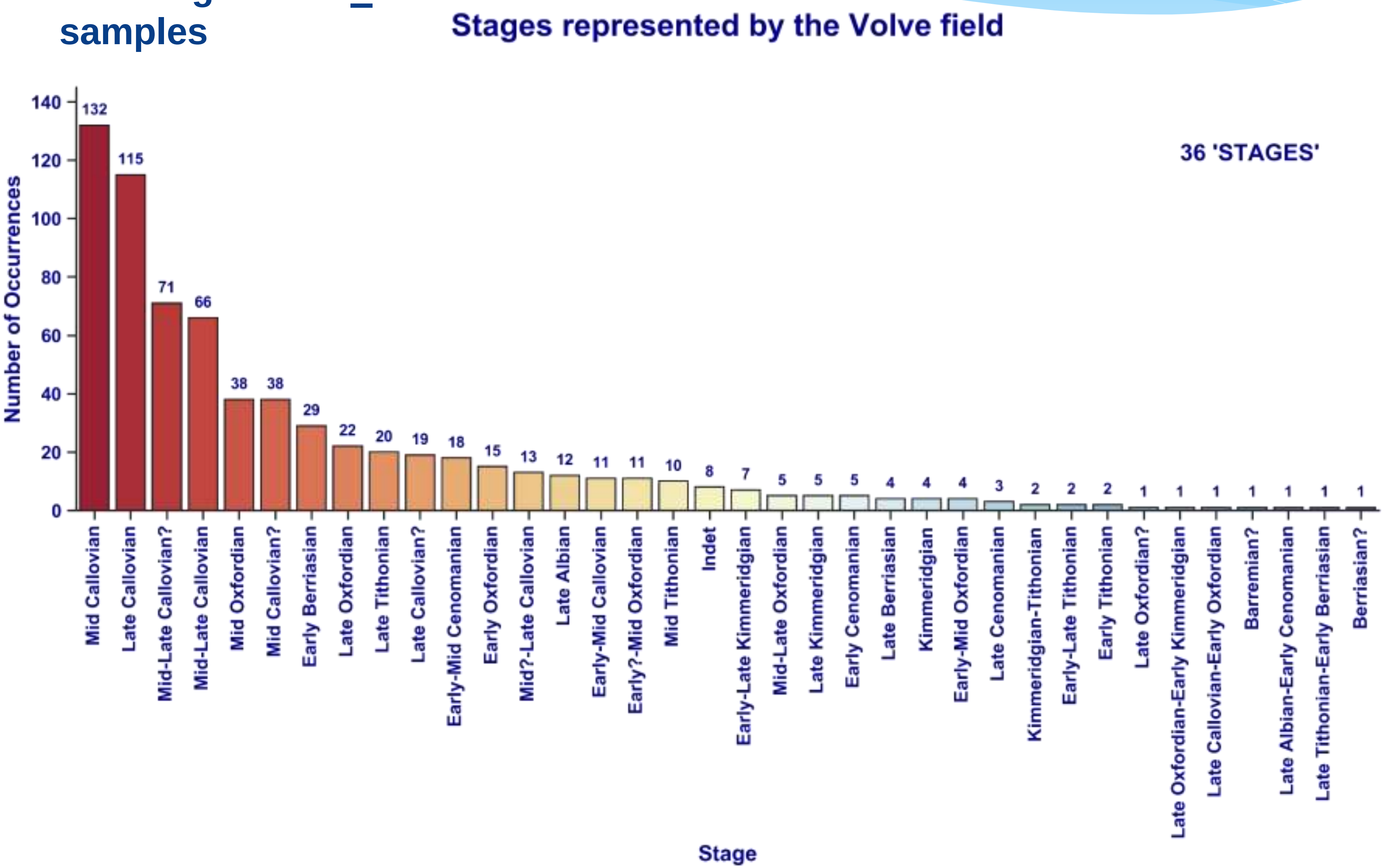
Age information



- Five major time periods have sample representation.
- Should have major assemble differences.
- What shape are they and what models work best?
- How best to compensate for imbalance?

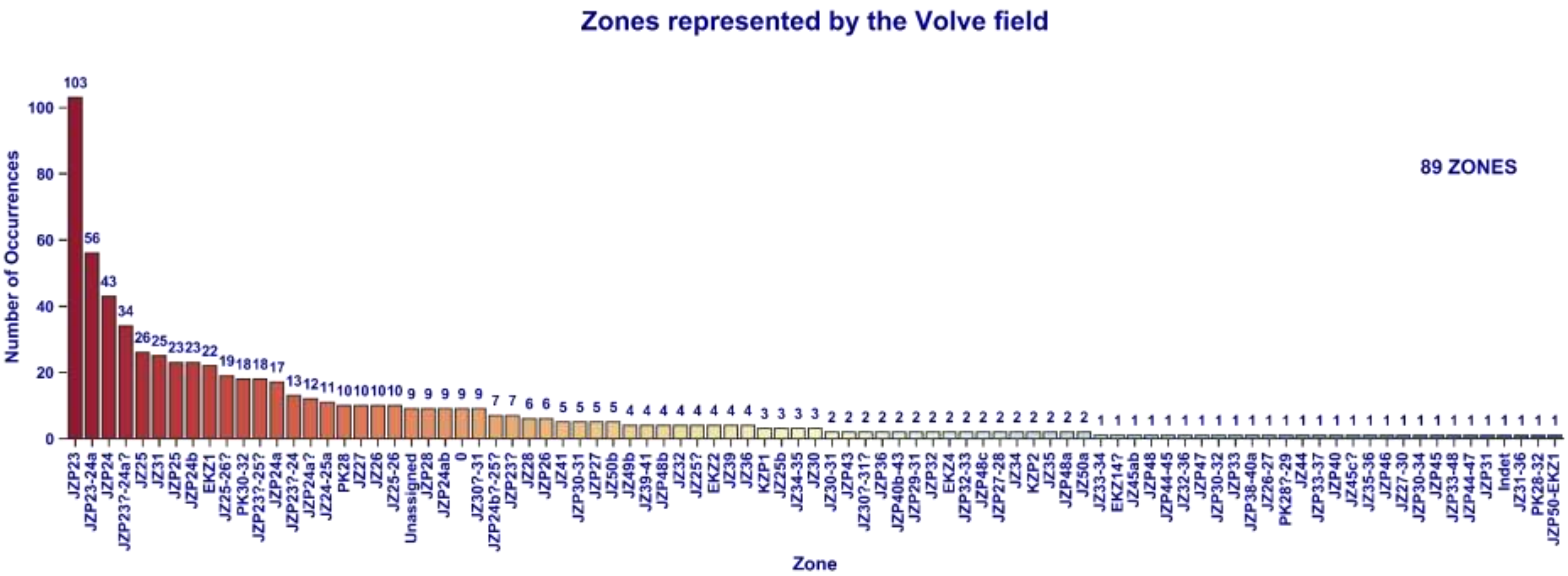
Stage information

Use stages with ≥7 samples

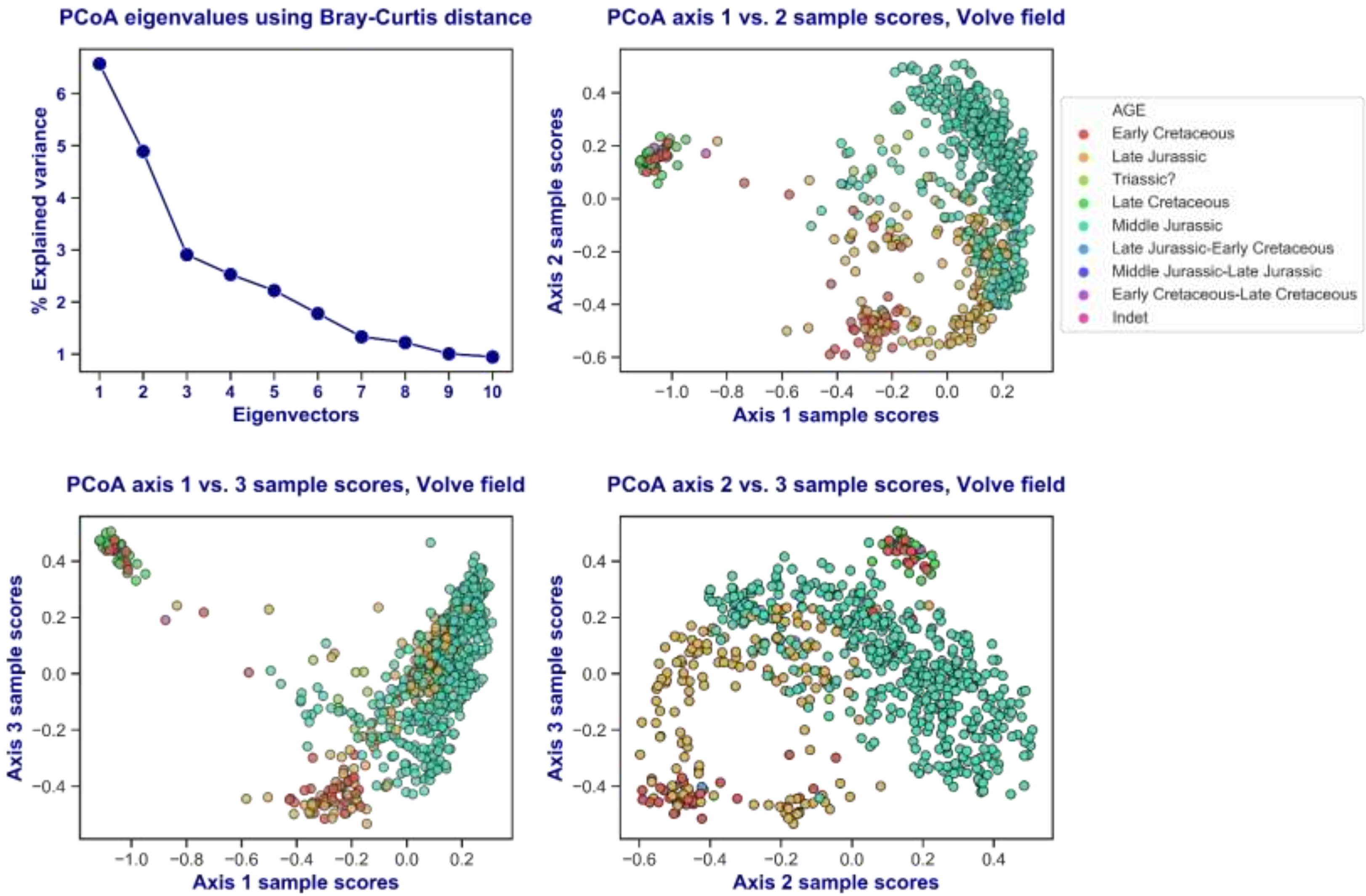


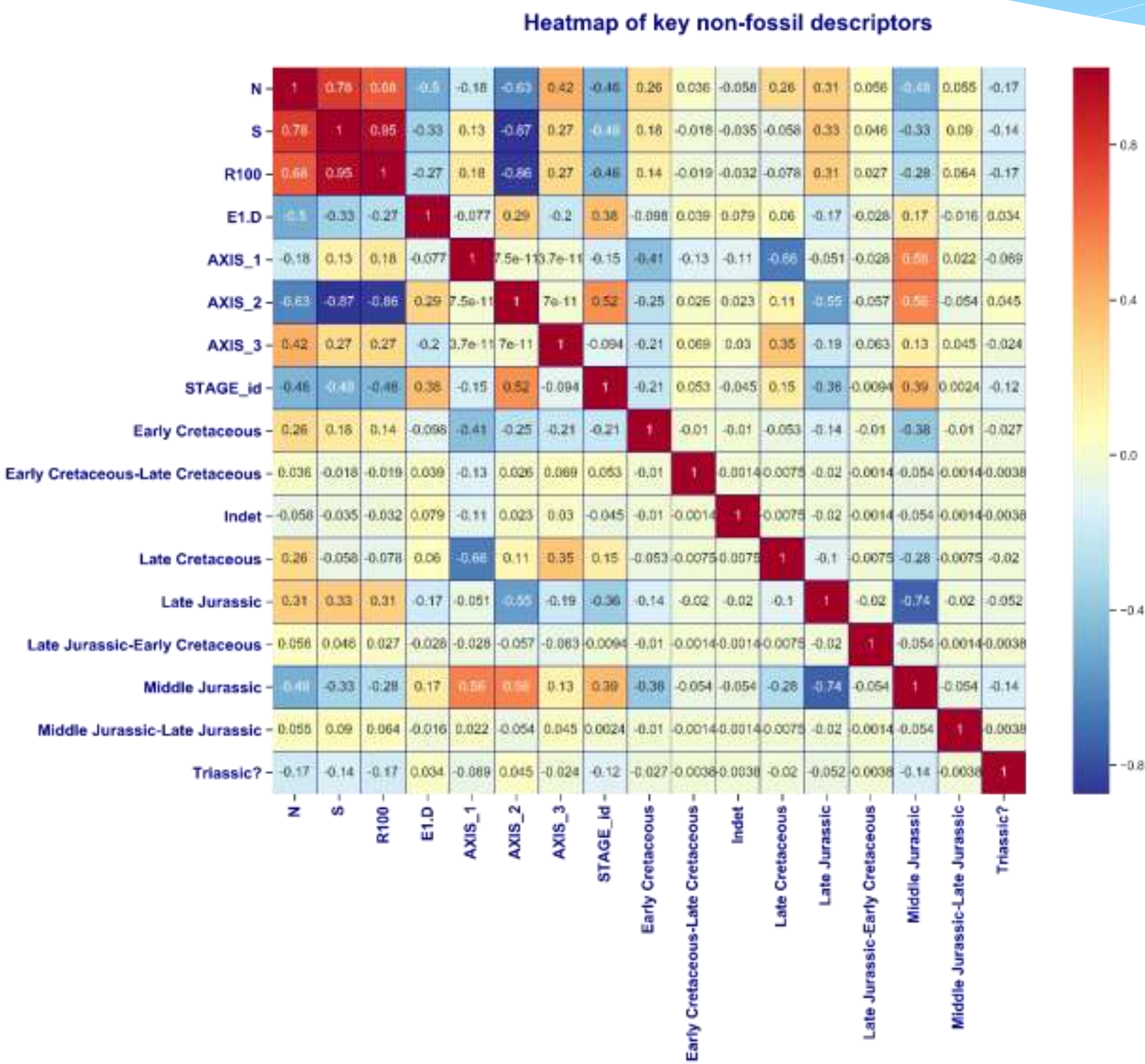
Zone information

Ultimate goal: Find robust zones that could be classified

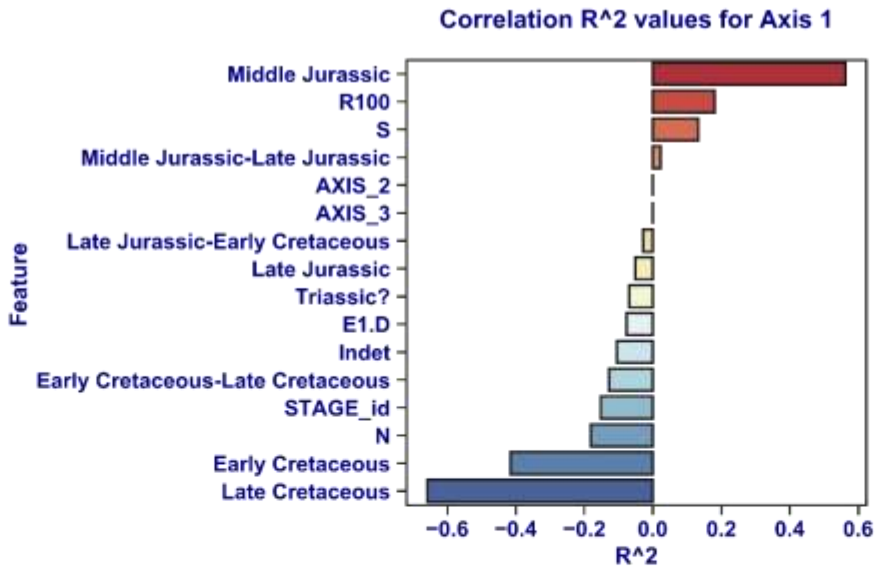


Principal coordinates analysis

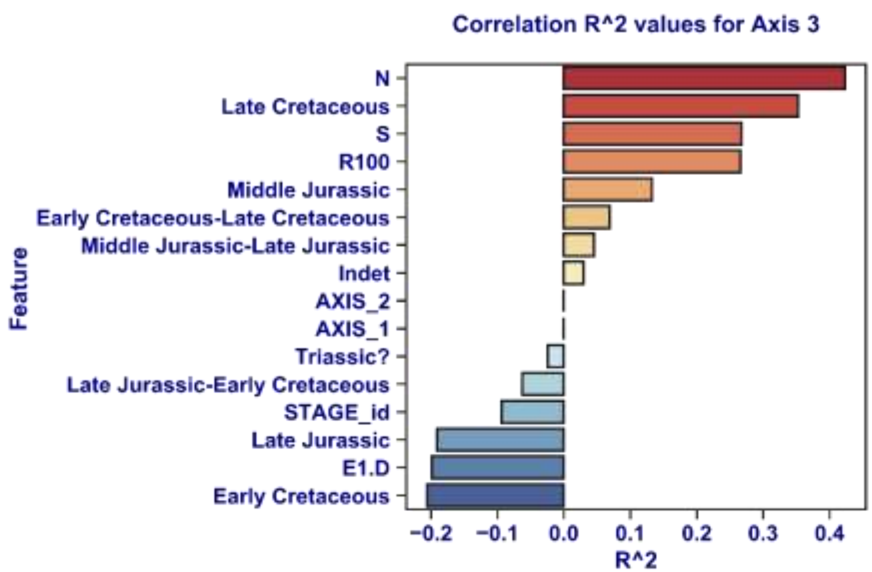


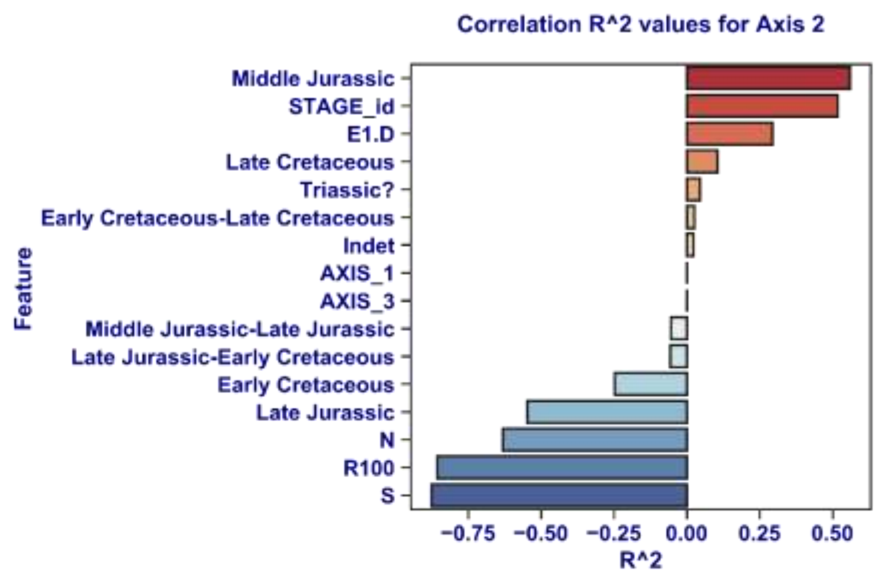


* Richness and count size related
* Axes related to age as well

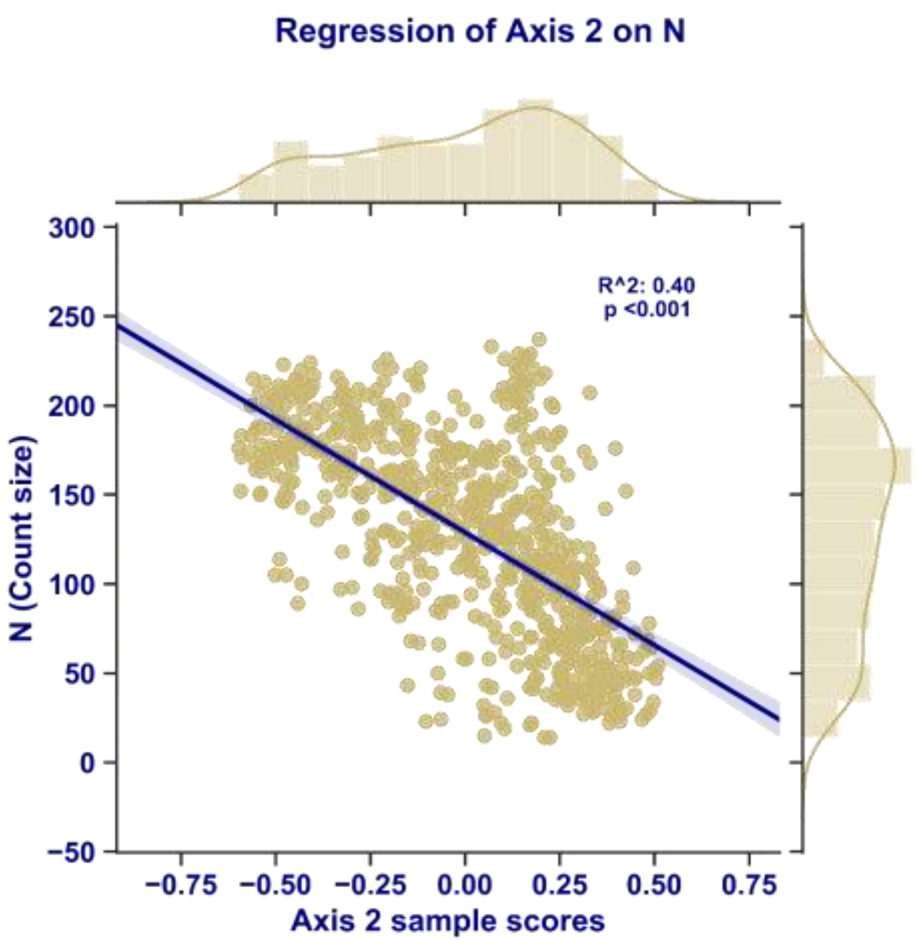


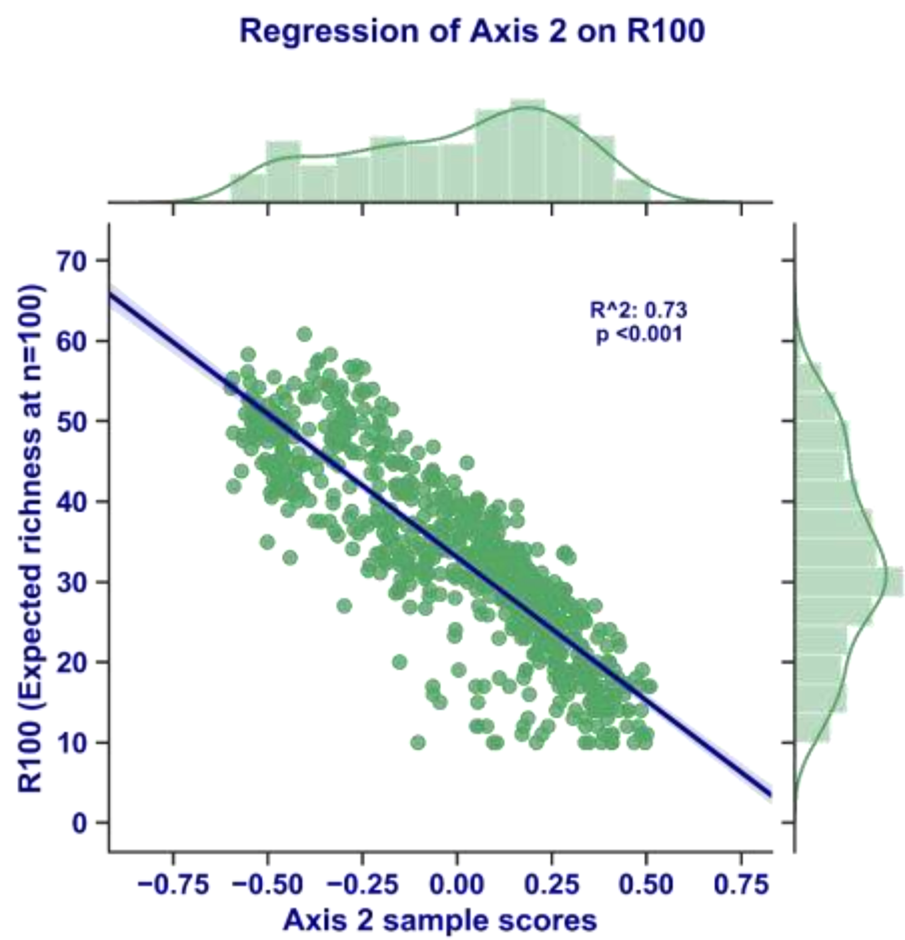
* Axis 1 = Middle Jurassic, and Cretaceous.
* Axis 2 = Middle Jurassic, Stage, richness & late Jurassic.
* Axis 3 = count size (N), richness (S, R100).

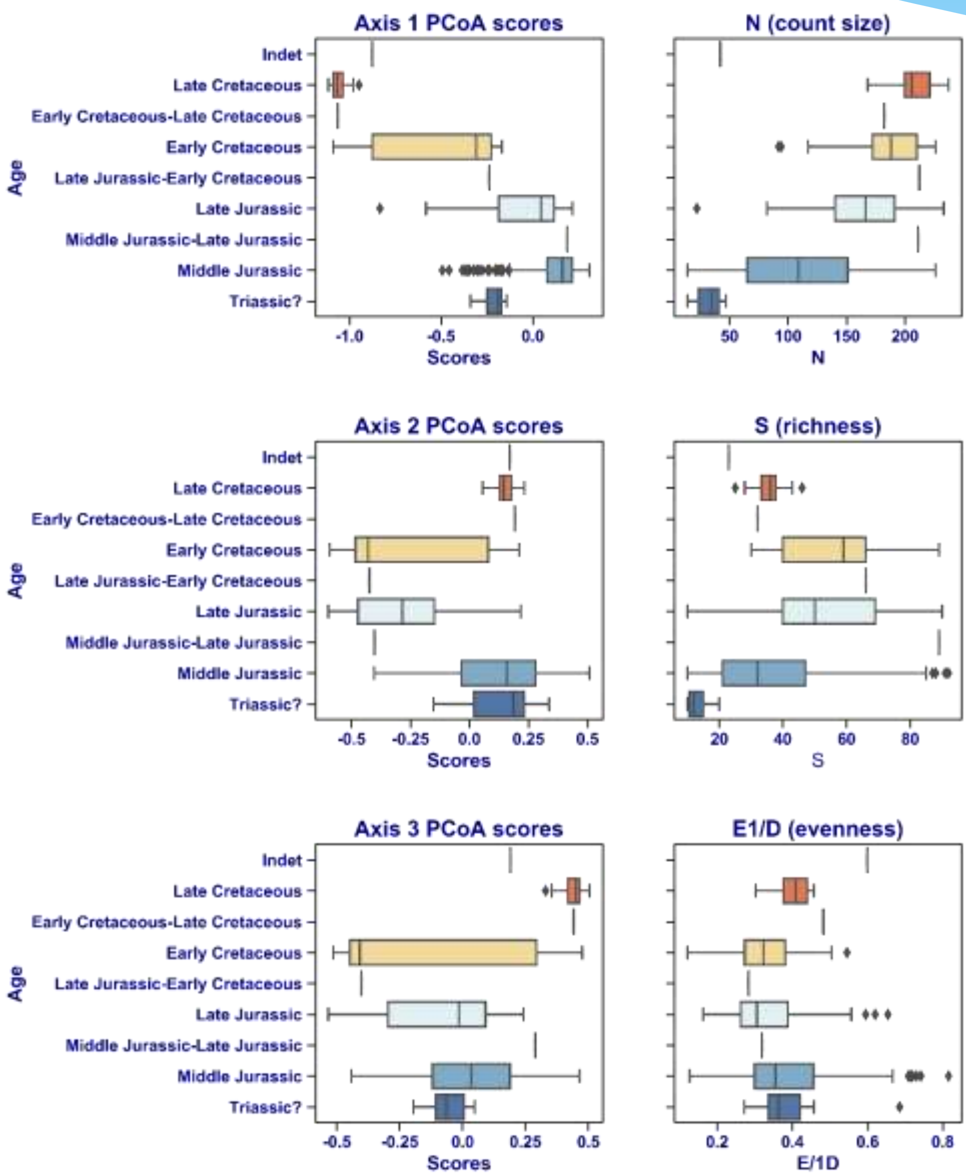




Age, count size and richness are important



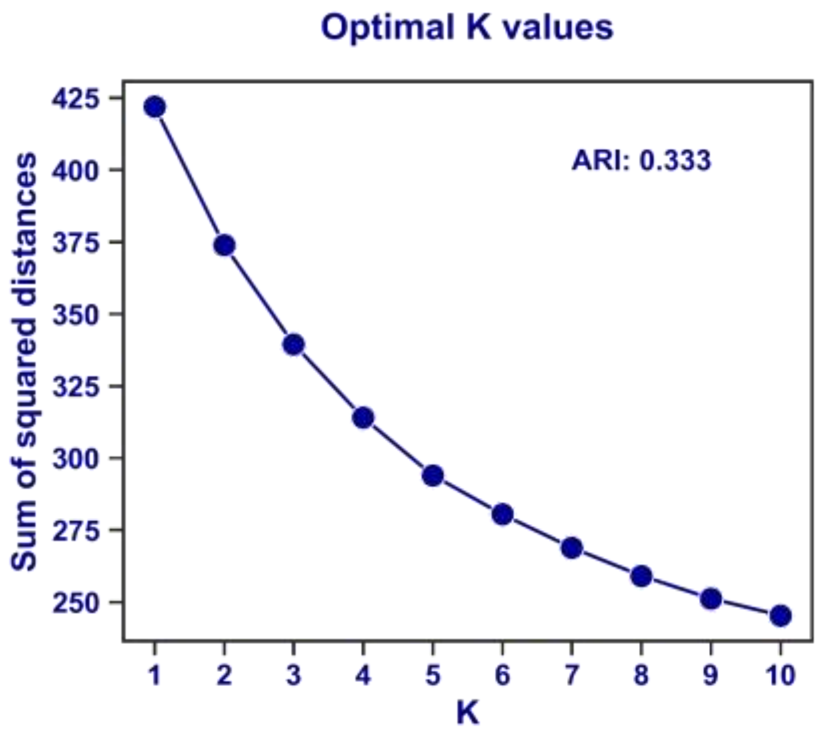


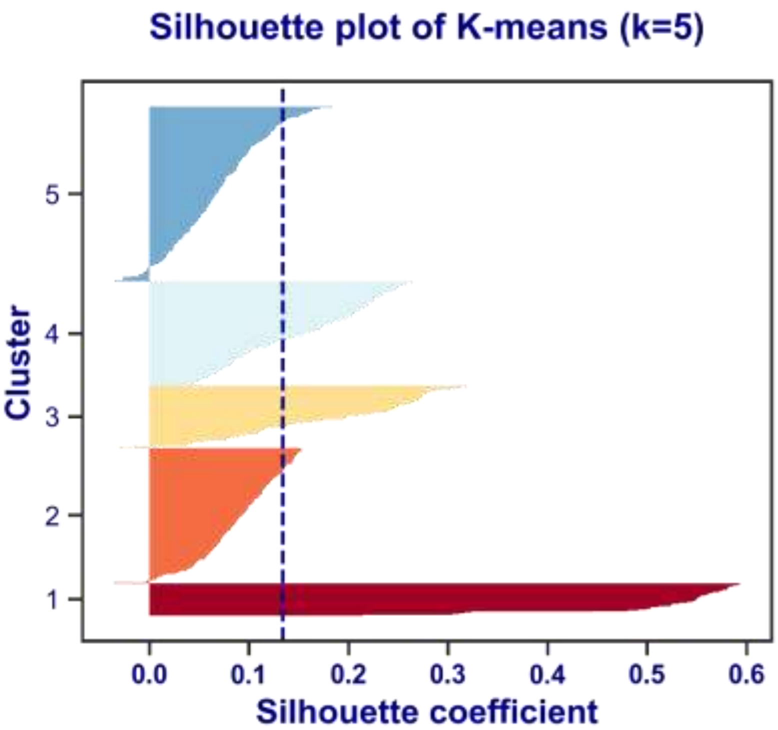


- Axis 1 = Significant changes over time.
- Axis 2 = Same stages are significantly different.
- N (count size) varies over time.
- S changes over time (related to N)
- E1/D shows no notable changes

K-means failure

- Unsurprising – low PCoA axes and complex interactions between age and count size imply low density clusters of linear/complex shape.





Supervised modelling

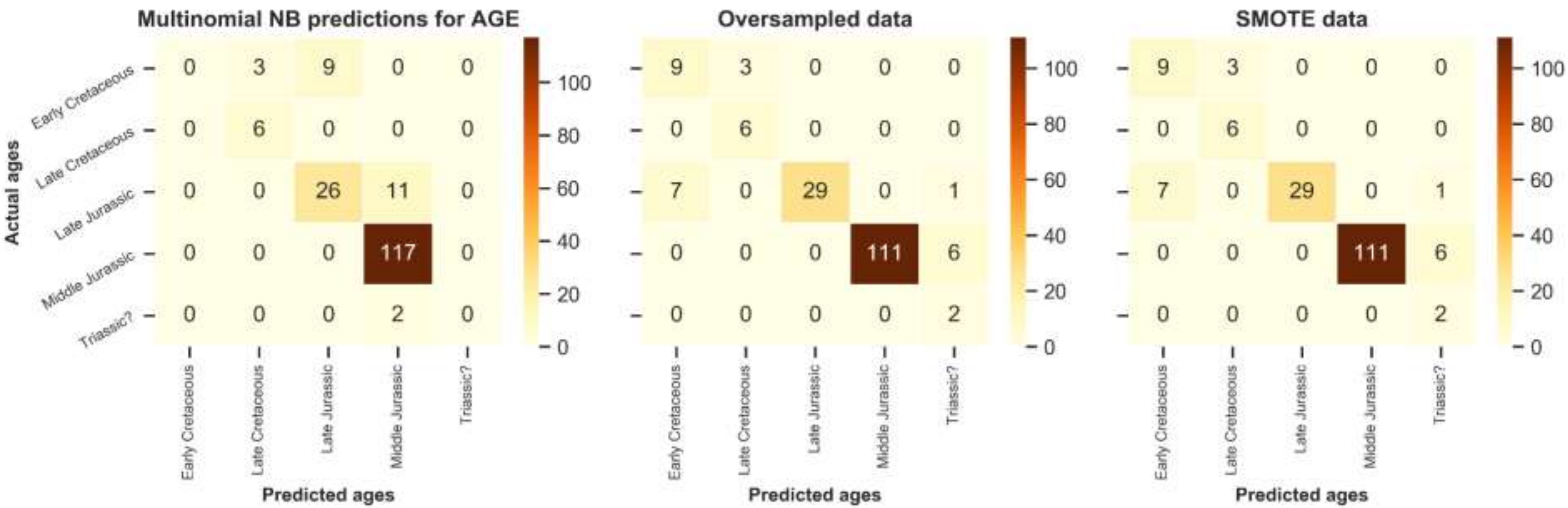
- Unclear what methods will work best – there is no 'best-practice' with these data types.
  - Linear trends – but not cyclic trends

- Start simple and tune more complex models or different types
  - Start with biggest classes – AGE (e.g. Middle Jurassic etc.)

- Compensate for class imbalance (see GitHub appendix)
    - Random oversampling
    - SMOTE (synthetic Minority Over-sampling Technique)

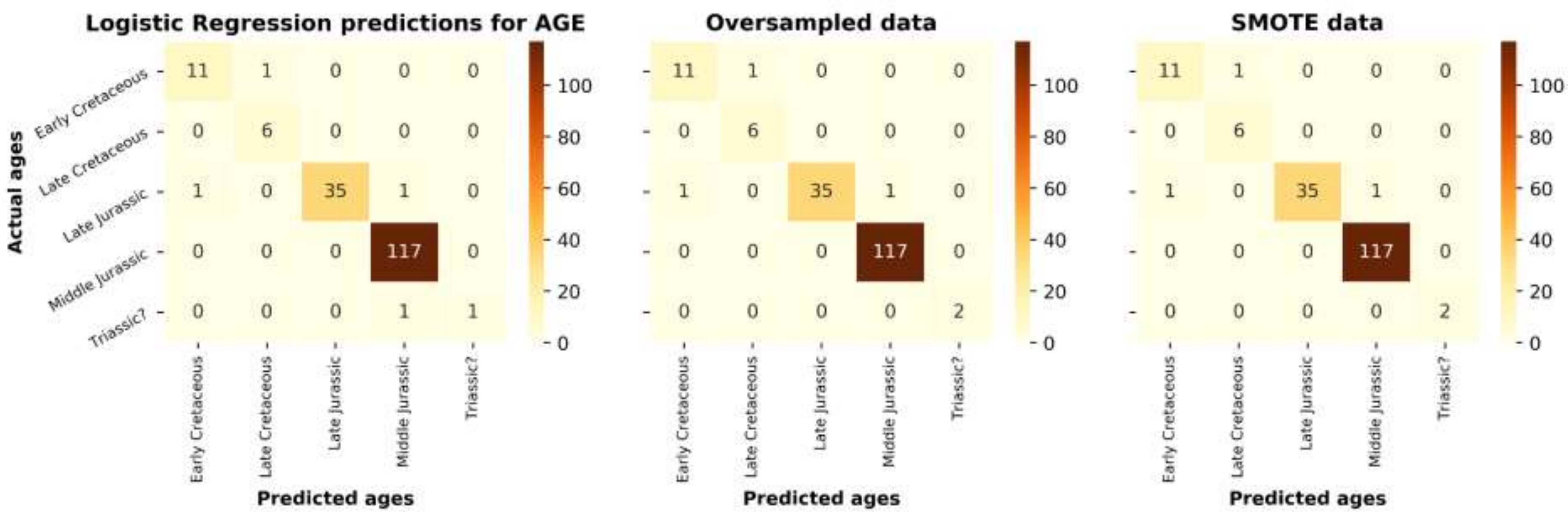
Naive Bayes

- Simplest model – oversampling improves accuracy
  - F1 score = 0.86
  - Oversampled F1 = 0.92



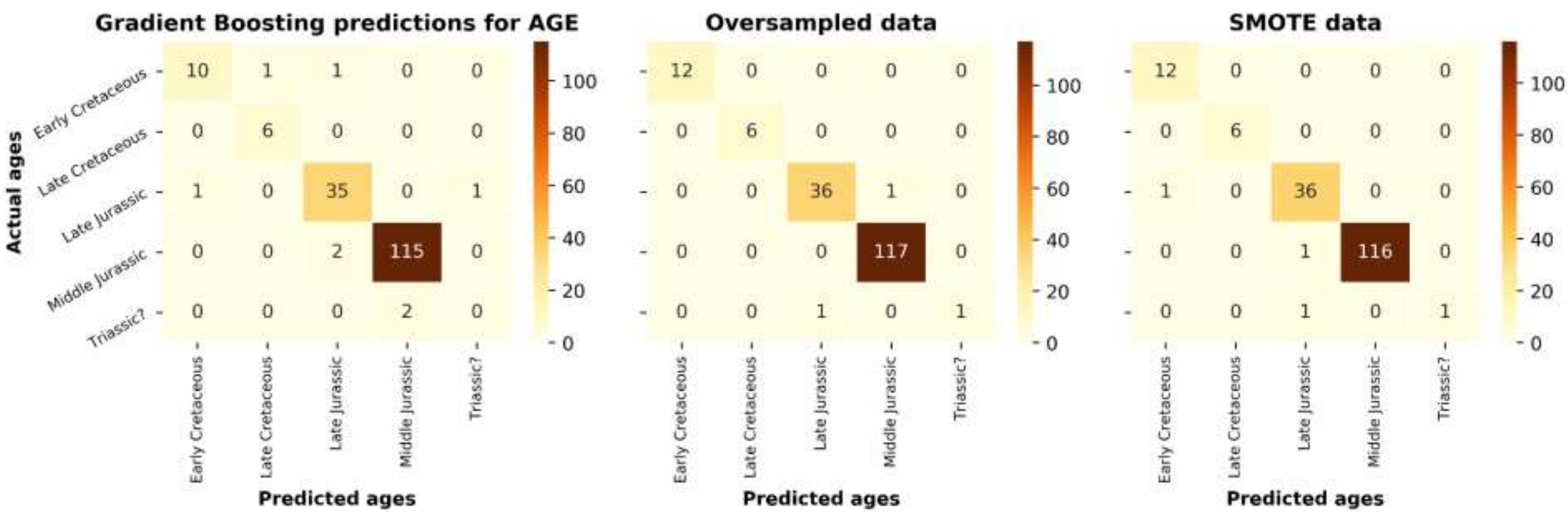
Logistic regression

- Oversampling doesn’t lift performance after tuning
  - F1 score = 0.98
  - Oversampled F1 score = 0.98



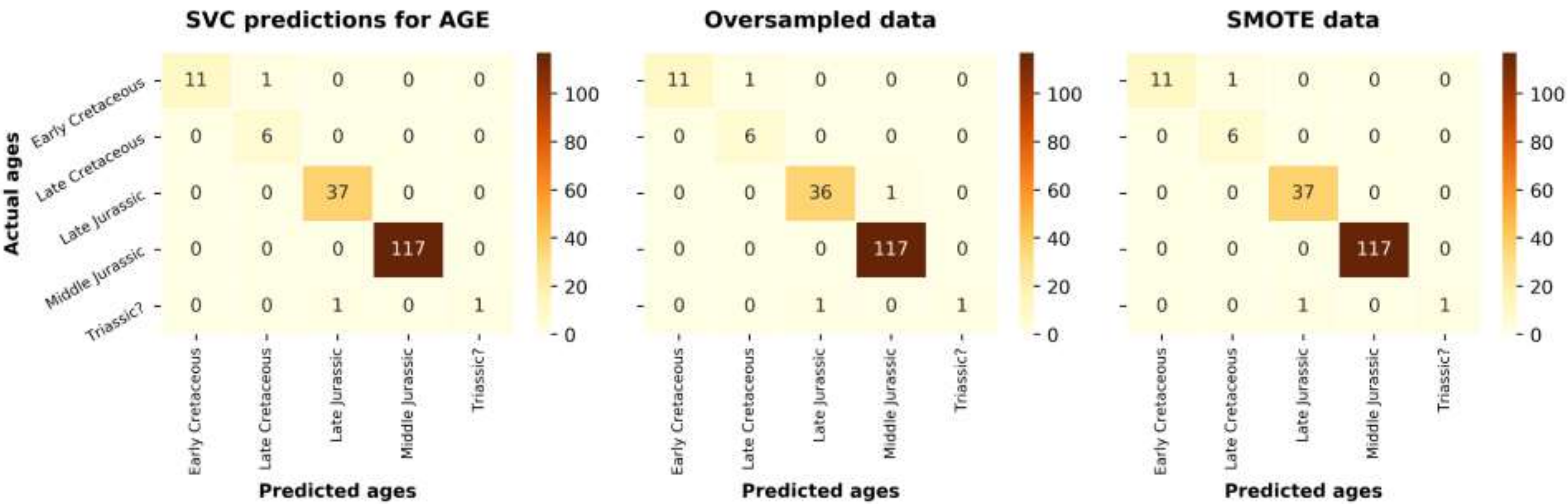
Gradient boosting classifier

- Slow to train
- Oversampling does lift performance after tuning
  - F1 score = 0.95
  - Oversampled F1 score = 0.99



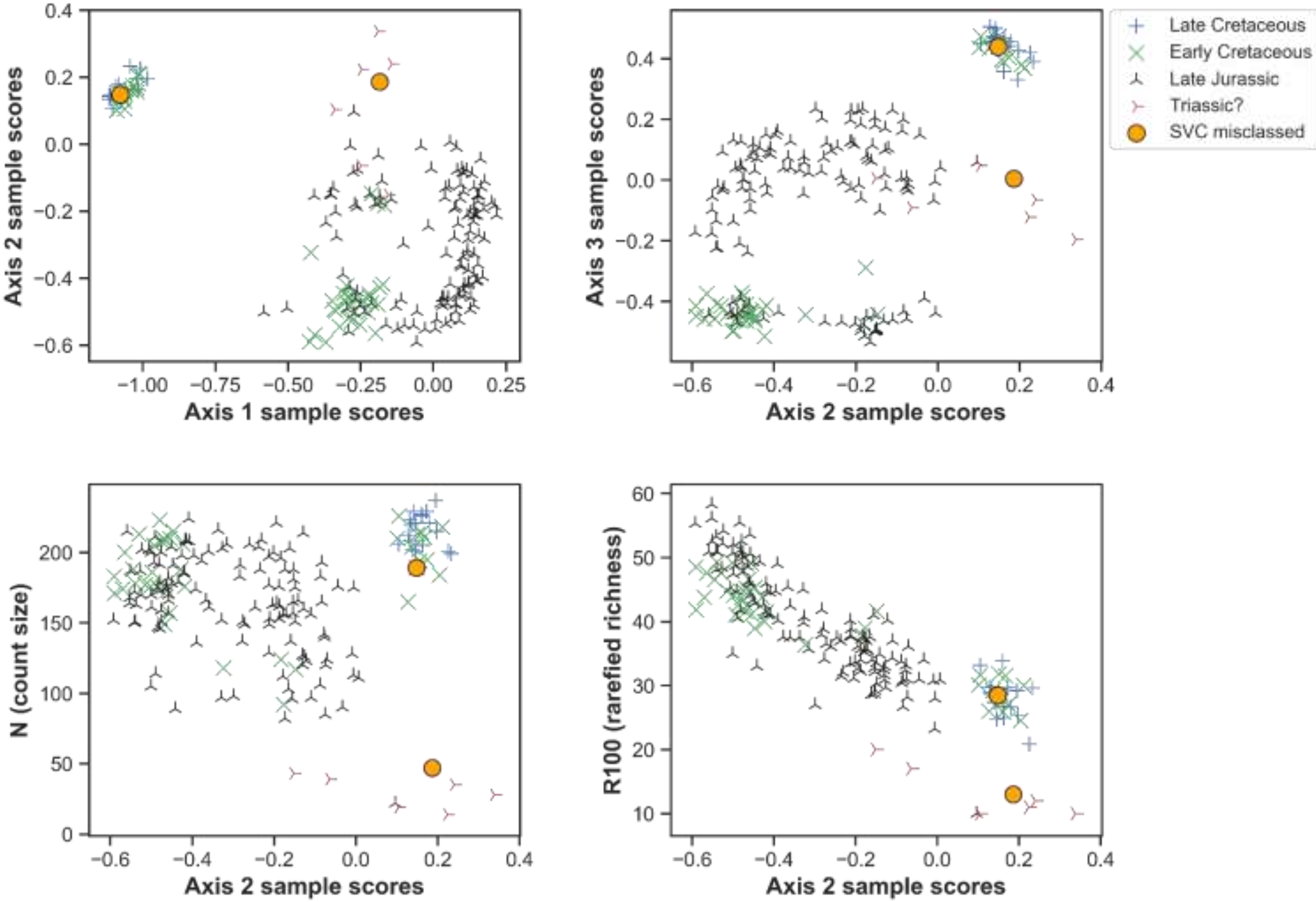
Support Vector Machine

- rbf kernel (untreated data)
- poly kernel for oversampled and smote
- Oversampling doesn't lift performance after tuning
  - F1 score = 0.99
  - Oversampled F1 score = 0.98



	Triassic?	Early Cretaceous
Early_K_prob_SVC	0.0271014	0.56869
Late_K_prob_SVC	0.0227135	0.392575
Late_J_prob_SVC	0.398913	0.0197647
Middle_J_prob_SVC	0.0625966	0.0099806
Triassic?_prob_SVC	0.488676	0.00898937

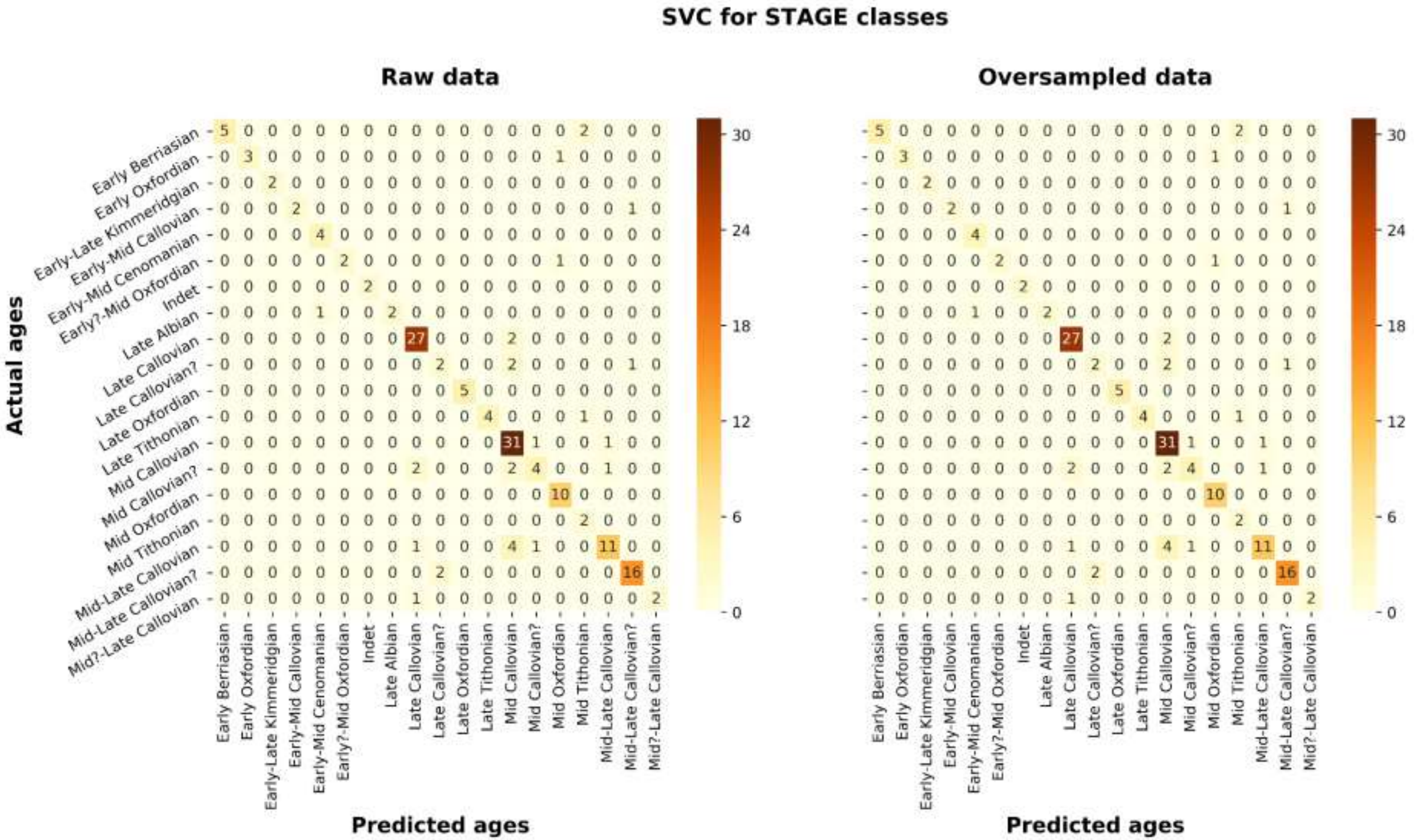
SVC misclassed samples



Modelling stage bins

- More time bins (36) but only those with 7 or more samples (n=19) were included (5 samples in Y_train, 2 in y_test).
- Oversampling employed to mitigate imbalance.
- Tune models
  - Best overall is SVC (not oversampled) F1 score = 0.83.
  - Logistic regression F1 score = 0.81
- More variability to models – some notably much weaker.
- Use ensemble approach (VotingClassifier)

SVC – can performance be improved?

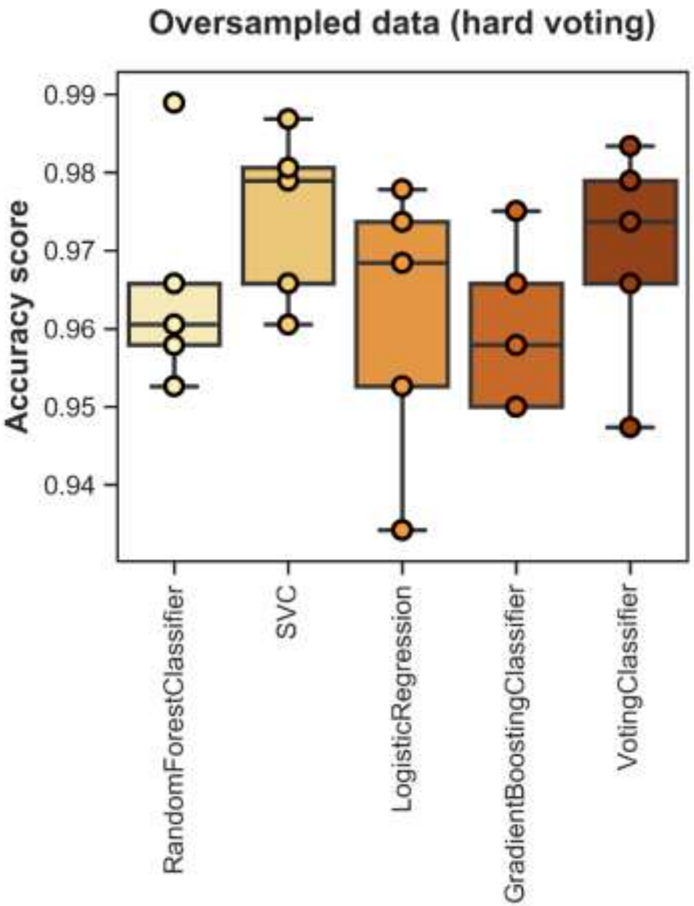


Ensemble models

* Data not suitable for KNN, Naïve Bayes is weak, and Gradient Boosting not particularly effective in this case – classes too small and data too sparse.

* Use different tree, kernel and linear models after appropriate tuning with GridSearch on oversampled data.
  * Random Forest (bootstrapping approach)
  * Support Vector Classifier (“poly” kernel)
  * Logistic regression (linear L2 regularization)
  * Gradient Boosting Classifier (boosting, not boostrap)

Accuracy shows classification potential of majority classes



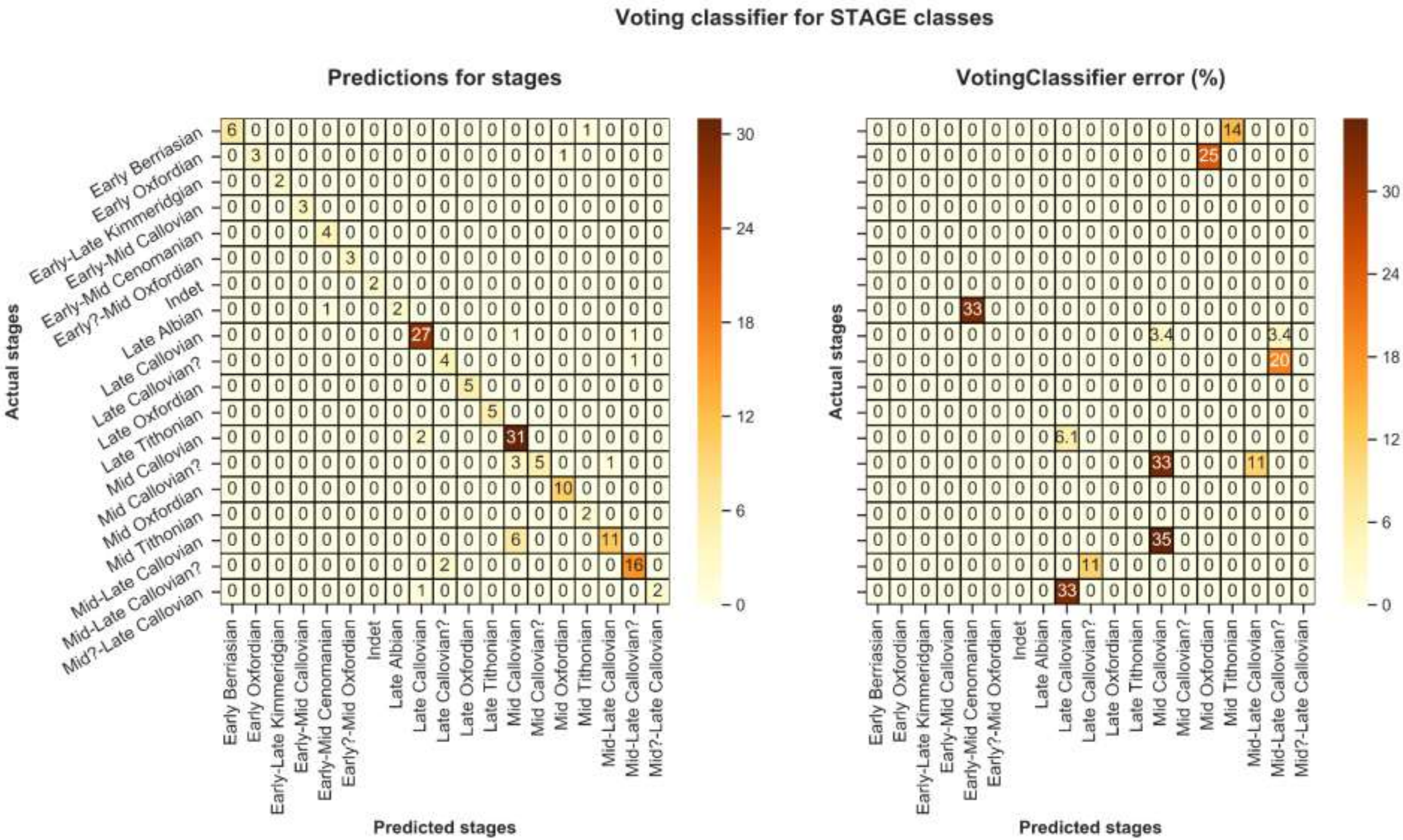

```
Mean of model CVs:

model_name
SVC                           0.974543
VotingClassifier              0.969834
RandomForestClassifier        0.965152
LogisticRegression            0.961357
GradientBoostingClassifier    0.959751
Name: accuracy, dtype: float64

STD of model CVs:

model_name
LogisticRegression            0.017935
VotingClassifier              0.014163
RandomForestClassifier        0.014108
SVC                           0.010961
GradientBoostingClassifier    0.010779
Name: accuracy, dtype: float64
```

* Precision = 0.89
* Recall = 0.87
* F1 score = 0.87 ∴ lift of 4% from SVC.



Concluding remarks

* **Oversampling** can help classification of minority classes significantly.
* Possible to classify classes with just 7 samples
* Performance of models **dependent on size of classes, contrasts between them, and the shape of the data** (i.e. dimensions).
* **Ensemble** of weak- and strong-learners improves model performance by 4%
* **Errors likely due to original analyst error in labelling, and caving not indicated on reports.**