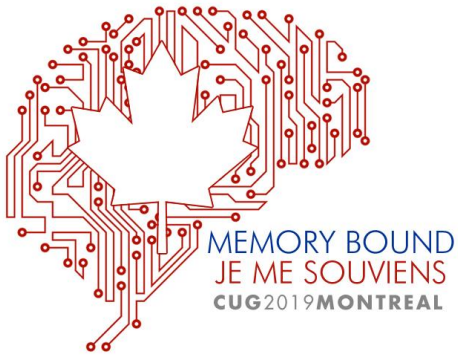
Machine learning on Crays to optimise petrophysical workflows in oil and gas exploration



Nick Brown, Data Architect at EPCC

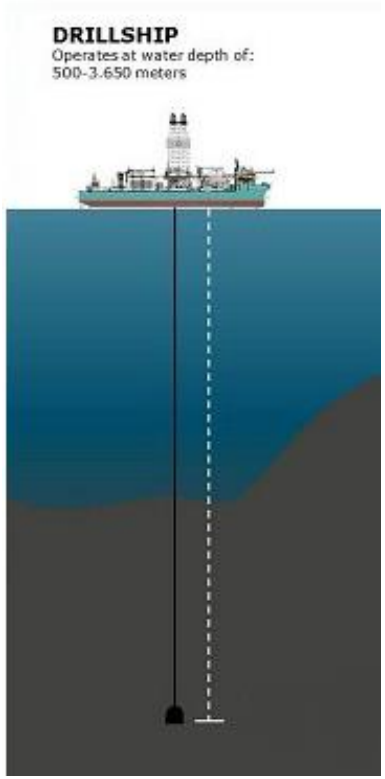
n.brown@epcc.ed.ac.uk

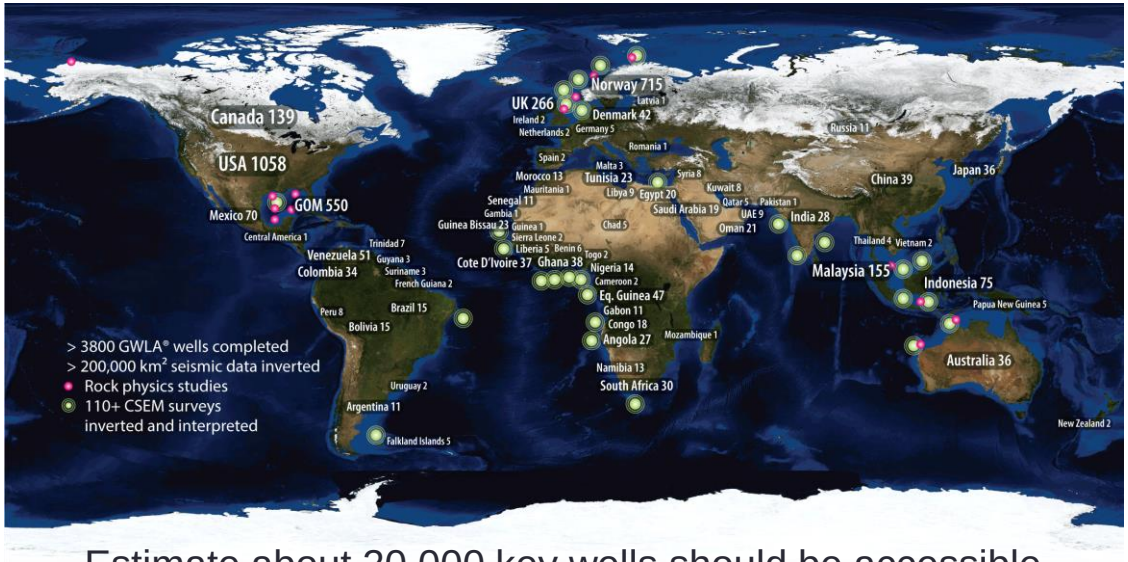
epcc

THE UNIVERSITY OF EDINBURGH

Driven by a collaboration with Rock Solid Images (RSI)

- RSI develop a successful software package known as RockAVO™ which enables oil and gas companies to virtually explore geological regions to inform decision making around exploration and exploitation
  - This is fed with well atlases, which are interpreted from the raw well log data

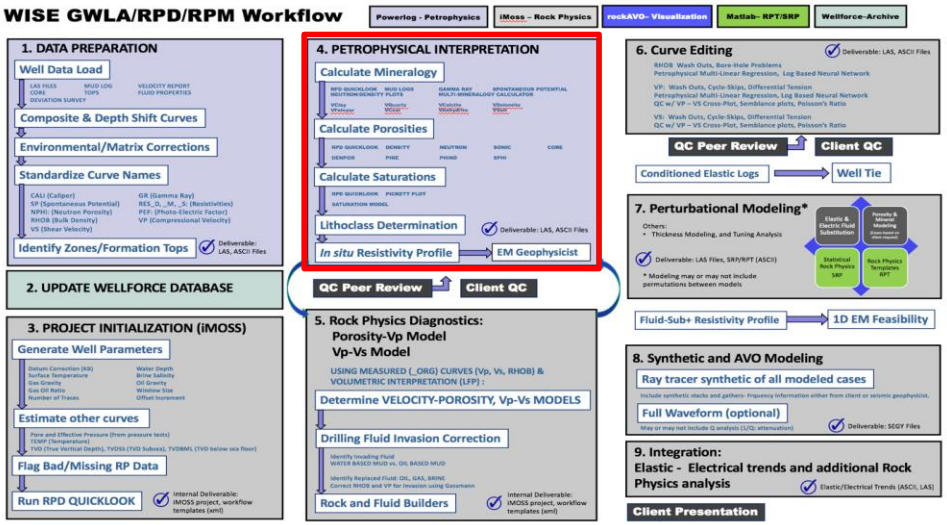




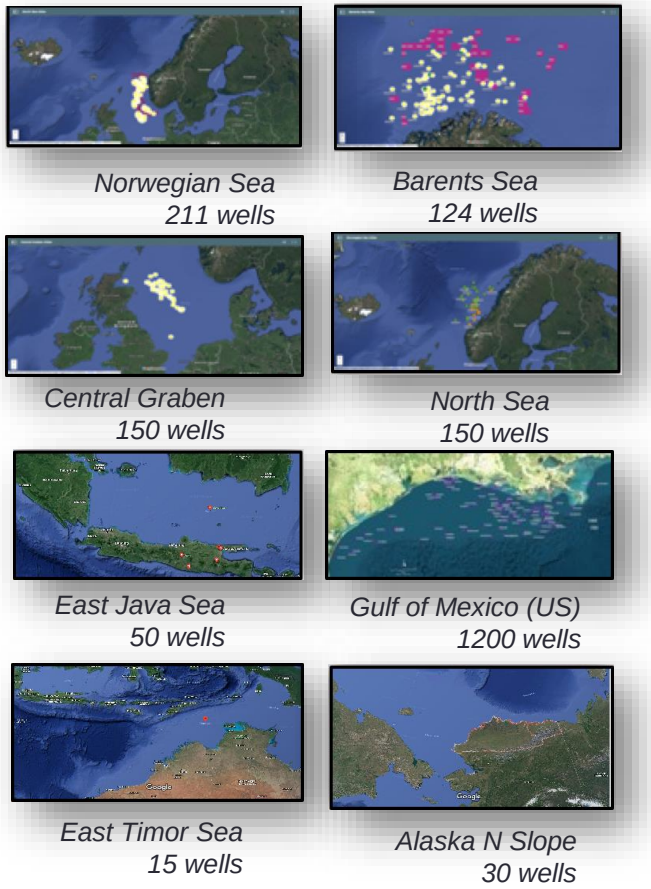
Estimate about 20,000 key wells should be accessible

Where does machine learning come in?

- A time consuming manual interpretation process is needed to convert raw well log data into atlases
  - Over 7 days per well
- This is fundamentally a pattern recognition problem, so can we use machine learning?
  - For instance could we go from over 7 days to 7 minutes per well?



Existing data base



Norwegian Sea
211 wells

Barents Sea
124 wells

Central Graben
150 wells

North Sea
150 wells

East Java Sea
50 wells

Gulf of Mexico (US)
1200 wells

East Timor Sea
15 wells

Alaska N Slope
30 wells

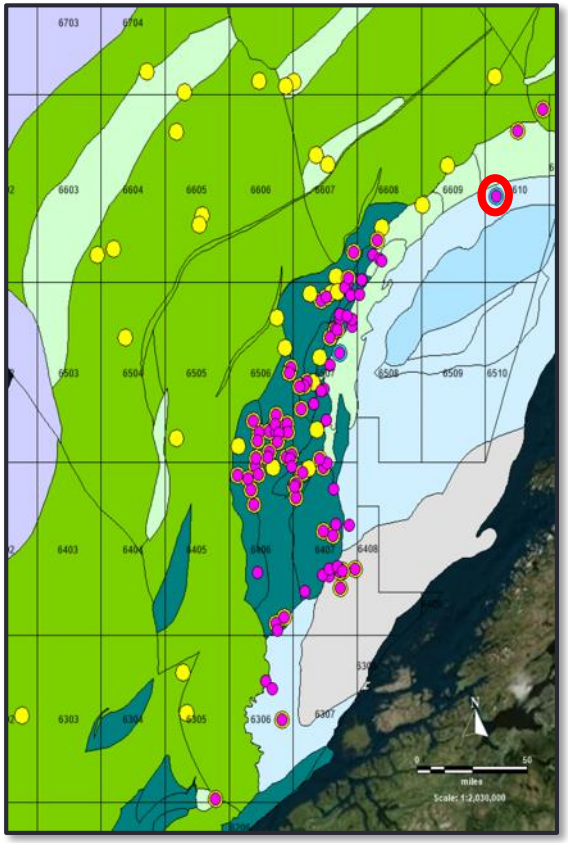
2000 wells in the existing database

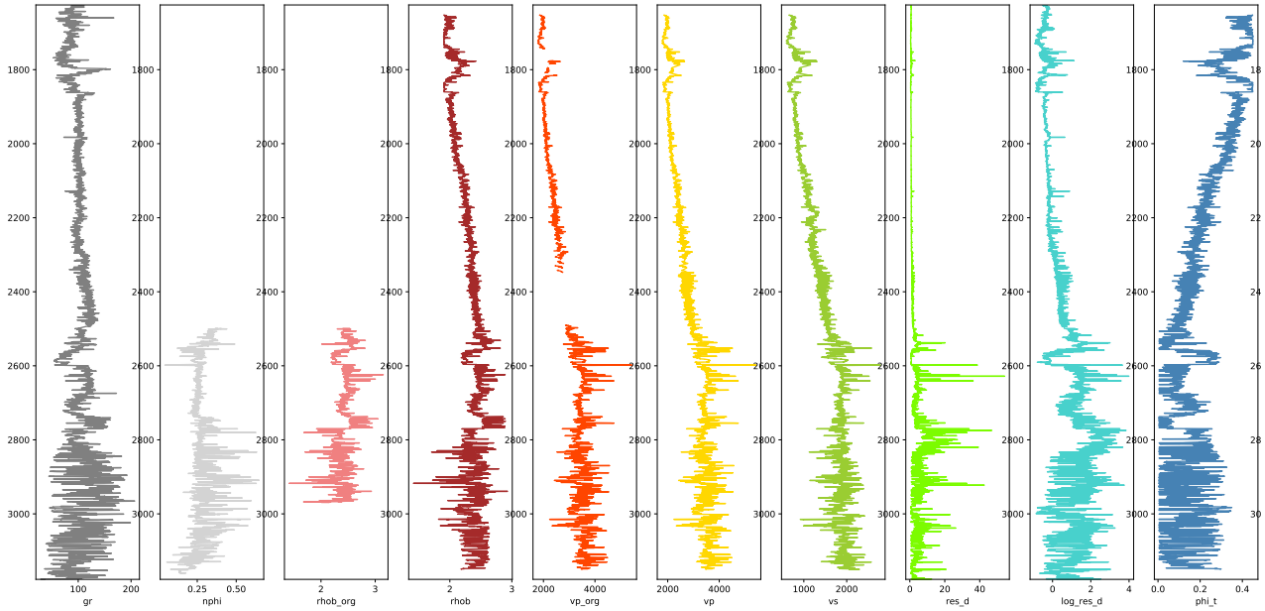
epcc

THE UNIVERSITY OF EDINBURGH

The real world data can be messy!

- The data itself it incomplete and unpredictable
- Our truths are themselves interpreted, and the human doesn't always get it right!

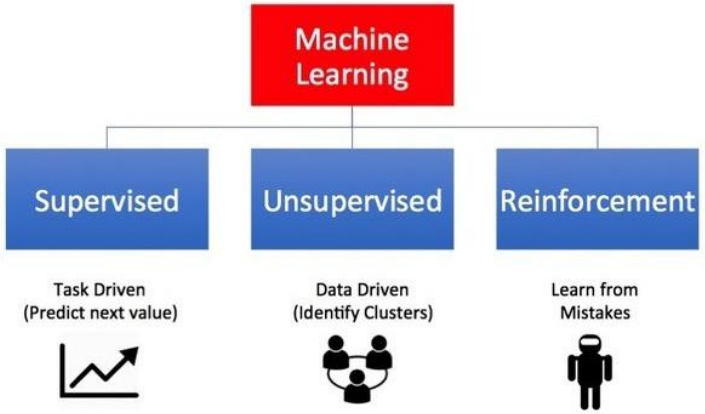




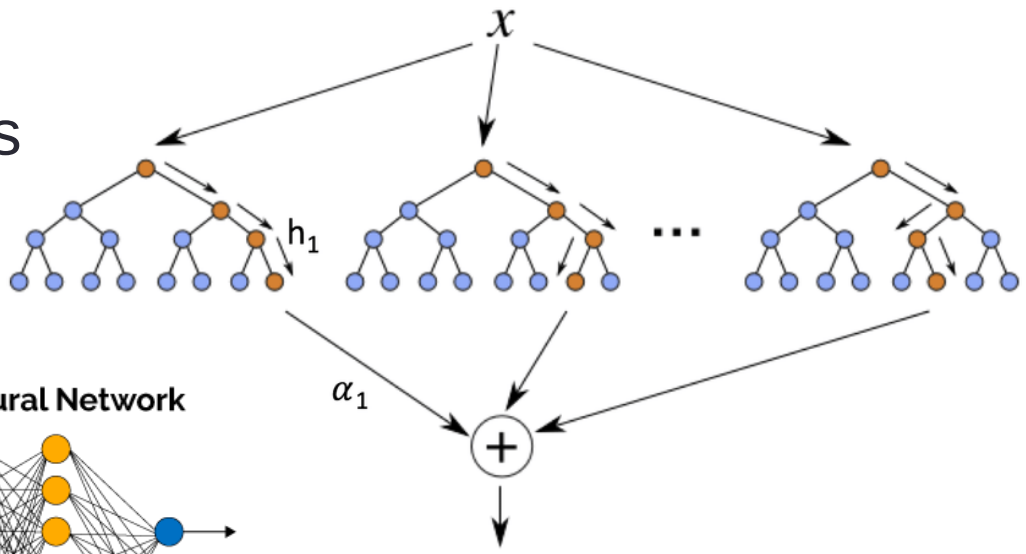
epcc

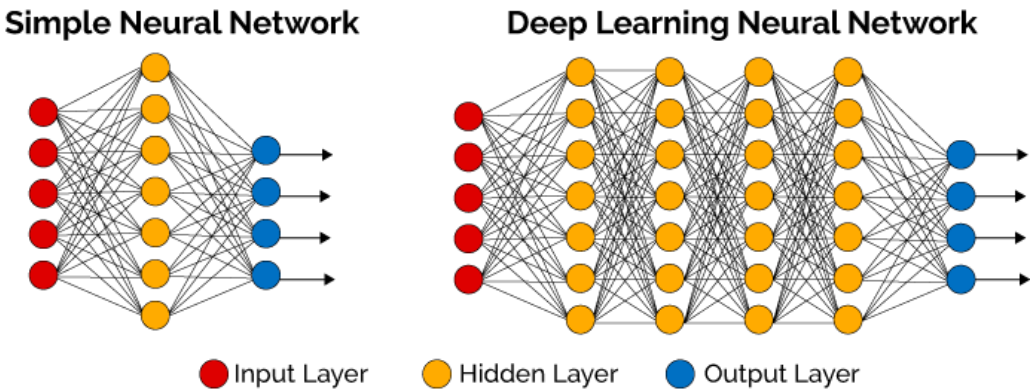
THE UNIVERSITY OF EDINBURGH

Machine learning methods



- Focussed on supervised learning methods here
  - Use 80% of the wells for training and 20% of wells for testing (sight unseen)
- We mostly use boosted trees via the XGBoost library on ARCHER, a Cray XC30





epcc

THE UNIVERSITY OF EDINBURGH

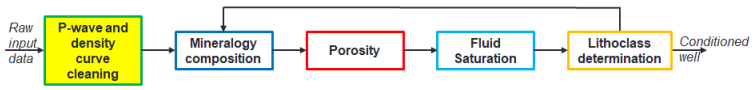
The petrophysical workflow

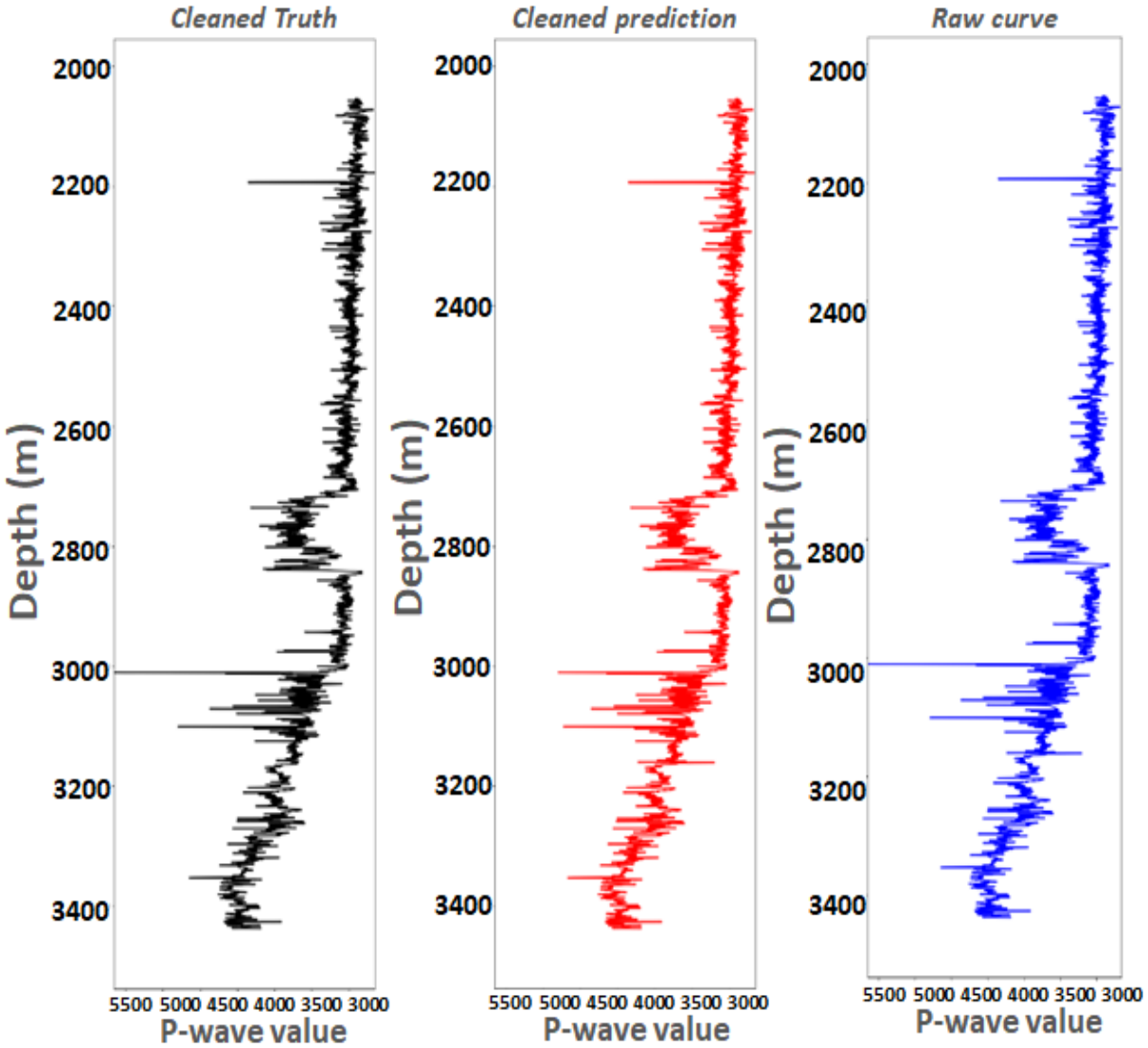
- Starting with raw data a preliminary cleaning phase is performed
- Then the four main phases are performed
  - Other people have looked at ML for of these stages (especially lithoclass determination), but this is the first time that ML has been used for the entire workflow and some of the stages too
- There is an iterative feedback loop when done manually
  - Can this knowledge be captured by a machine learning model?

epcc

THE UNIVERSITY OF EDINBURGH

Data cleaning: p-wave



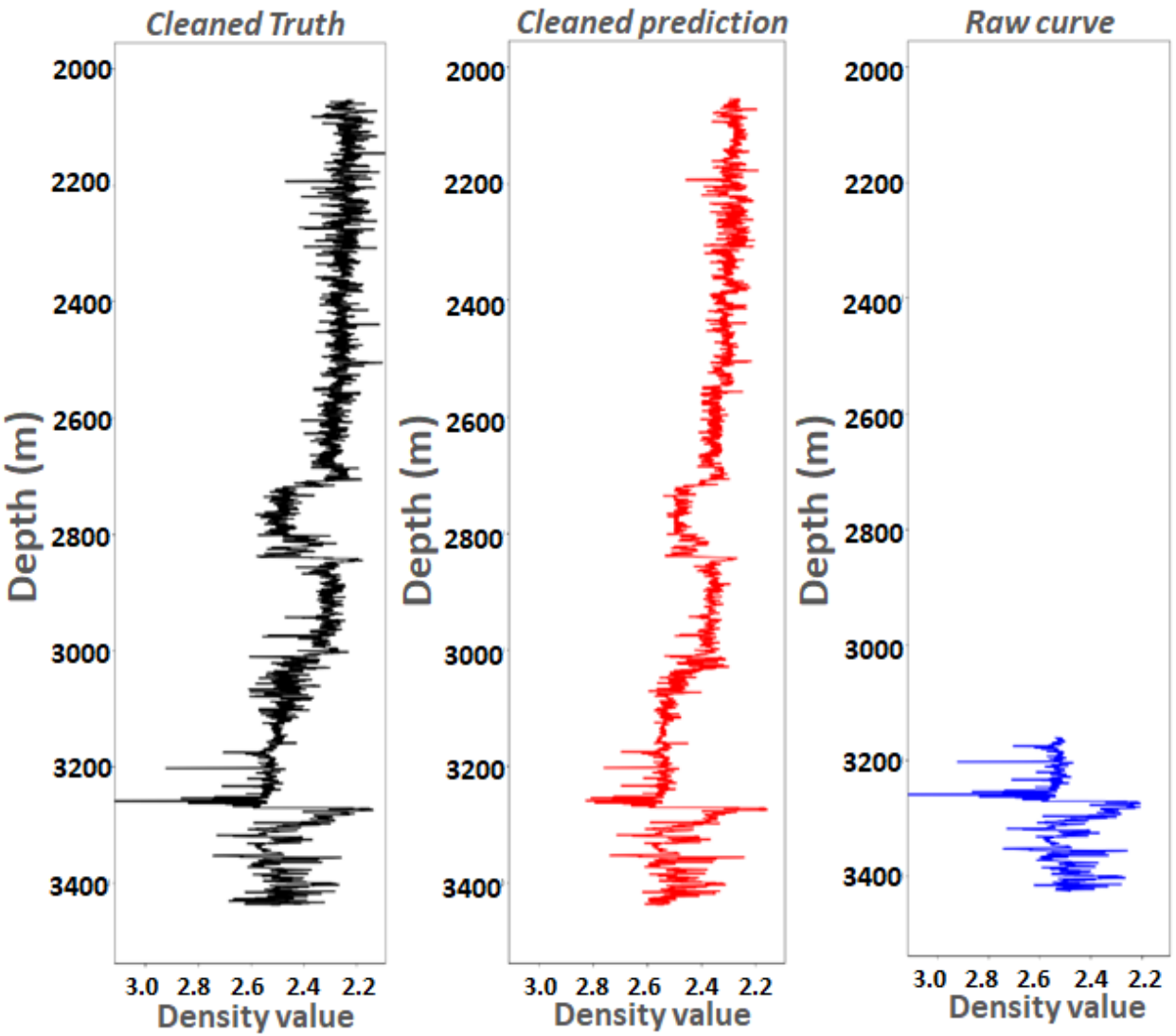


- Some input curves can be cleaned manually, although it can be a time consuming process
  - Here p-wave is being “cleaned”, although there isn’t too much to do (note the reduction in magnitude at 3000m)

epcc

THE UNIVERSITY OF EDINBURGH

Data cleaning: density

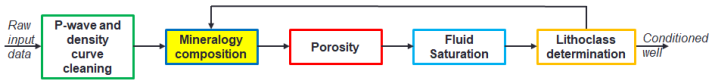


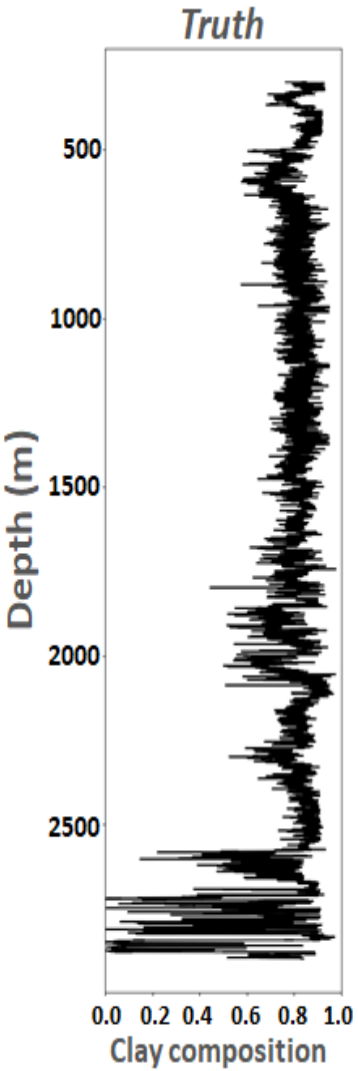
- For this well the density curve is more difficult as there are significant amounts of it that are missing
- Our model is able to *fill in the blanks* and matches reasonably closely to the manual curve which takes many hours to produce.

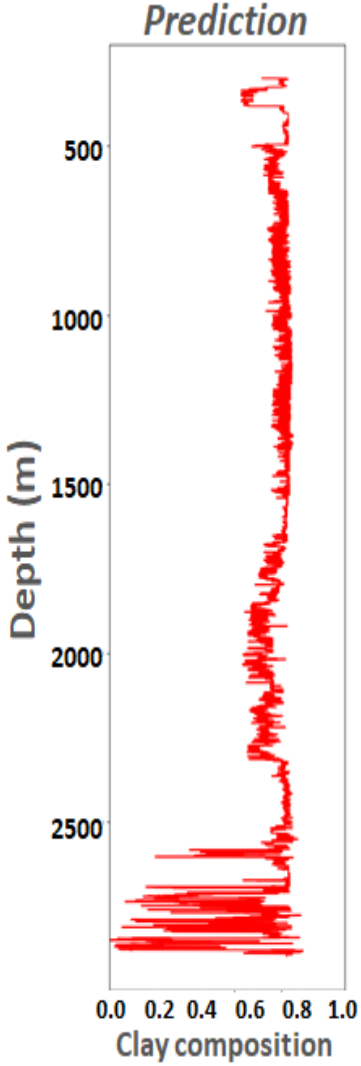
epcc

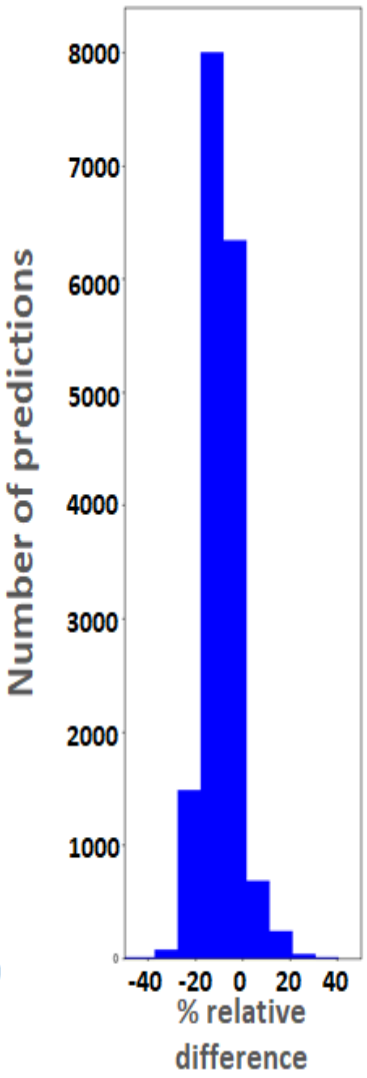
THE UNIVERSITY OF EDINBURGH

Mineralogy composition

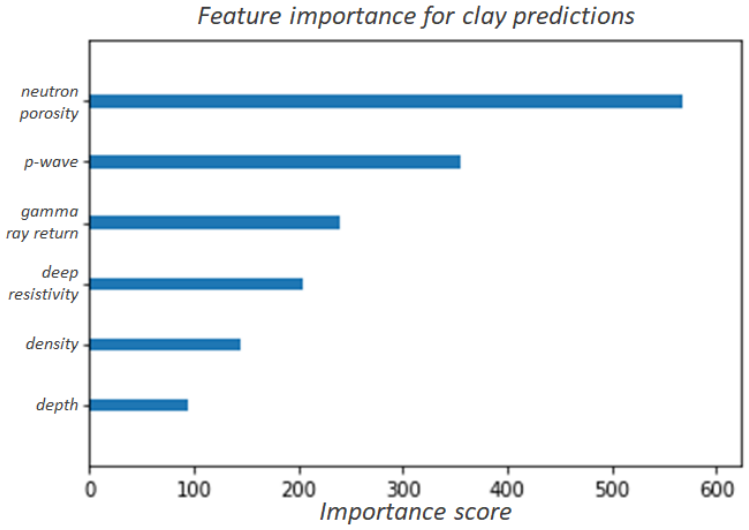








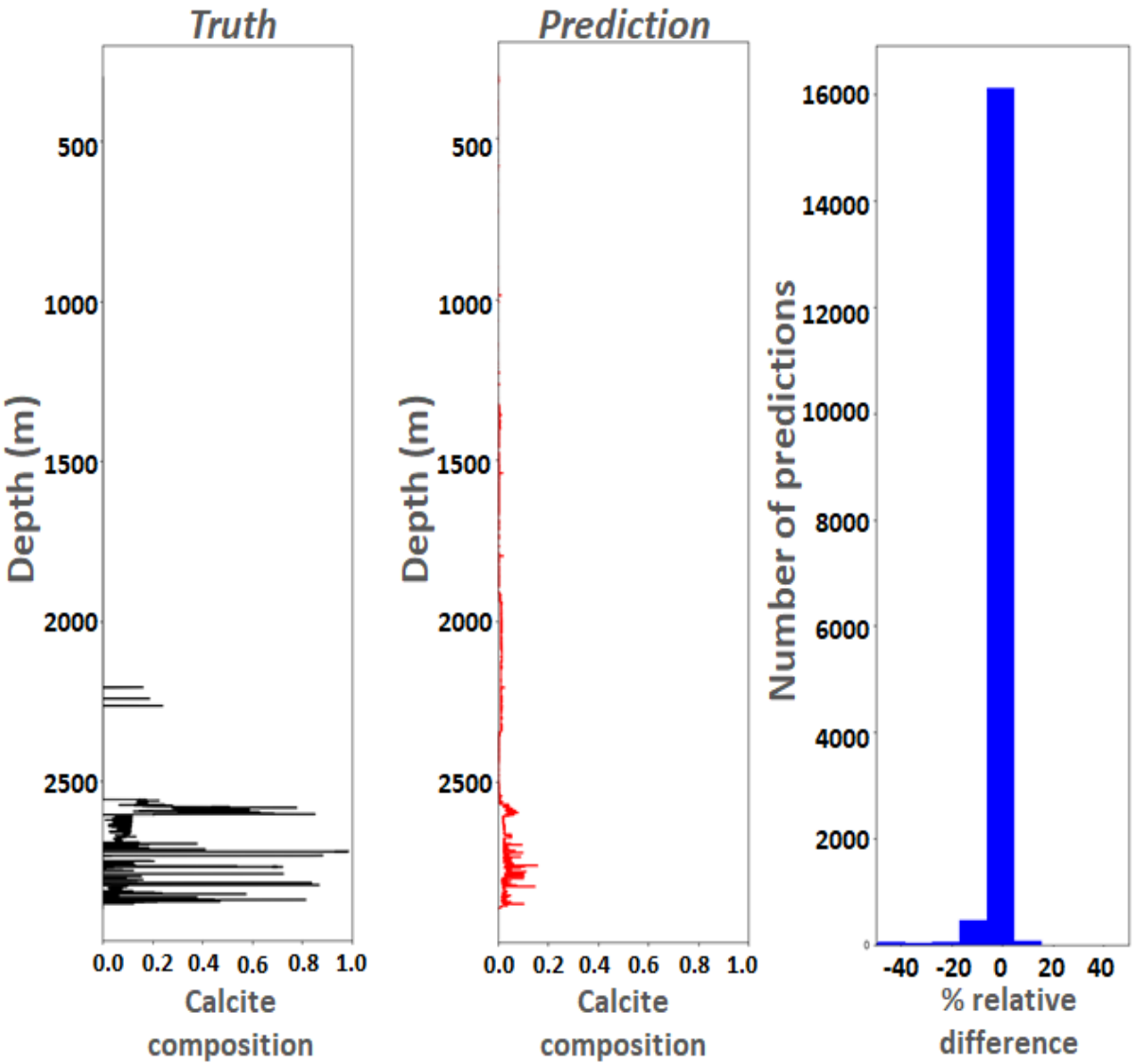
- There are thirteen minerals that we predict
  - With a trained model, inference takes less than a second per mineral



epcc

THE UNIVERSITY OF EDINBURGH

Mineralogy composition – the limit



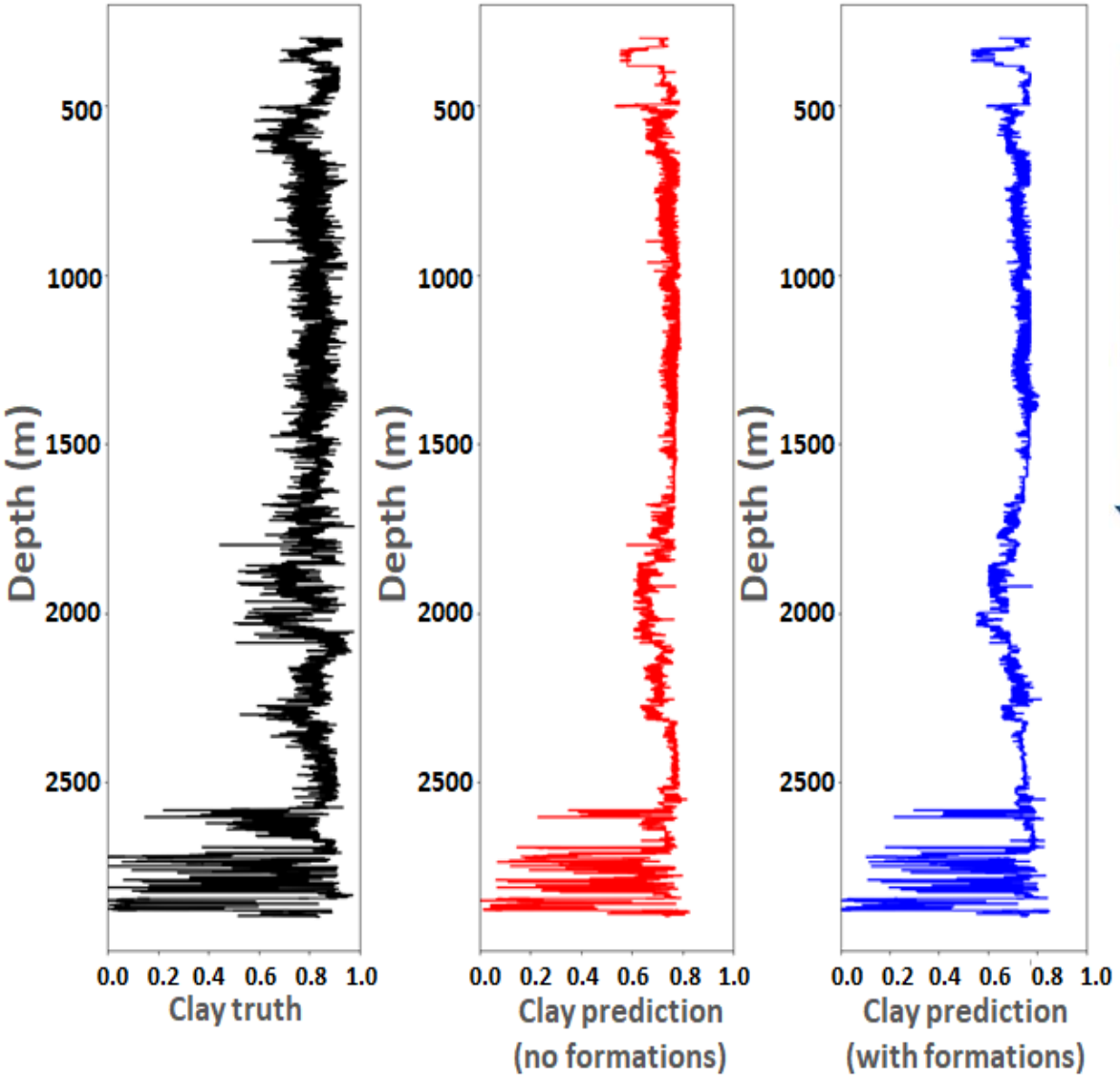
Mineral	RMS error
Clay	0.136427
Quartz	0.145153
Calcite	0.049276
Pyrite	0.004348
Dolomite	0.011489
Coal	0.050087
TOC	0.000394
Anhydrite	0.003198
Volcanic	0.005829
Feldspar	0.023668
Siderite	0.000772
Halite	0.000514

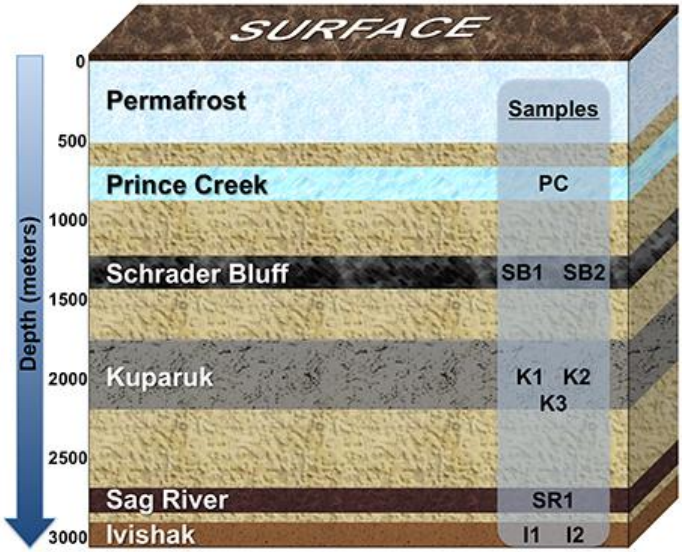
- But geology is inherently biased and as such other minerals don't work as well because the model simply doesn't see them enough
  - This is a fundamental limitation of the approach

epcc

THE UNIVERSITY OF EDINBURGH

Mineralogy: Adding formations



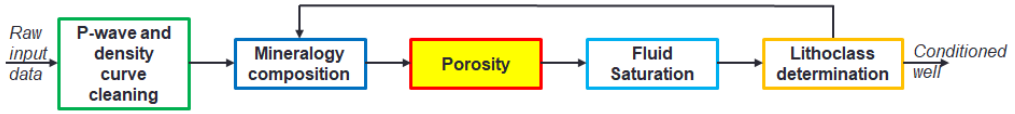


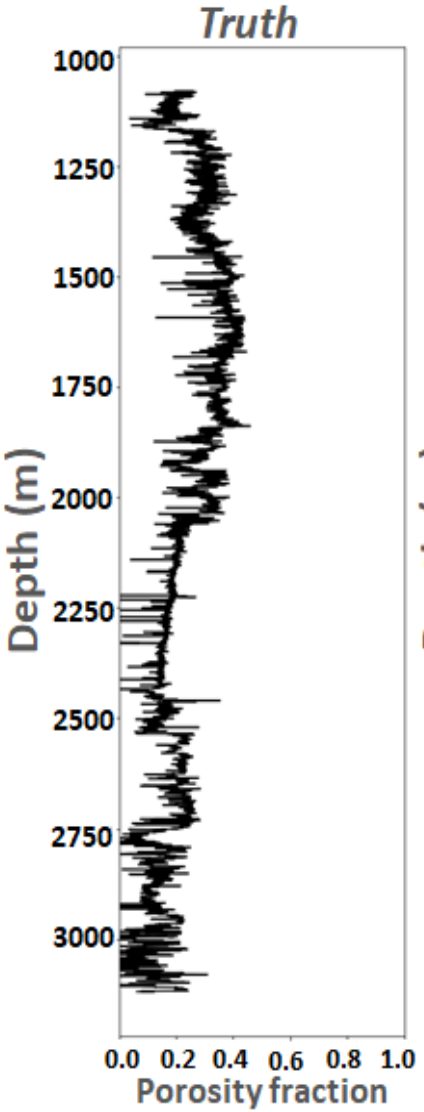
- Added formation information from the NPD website as extra Boolean fields to the data to see if this would help

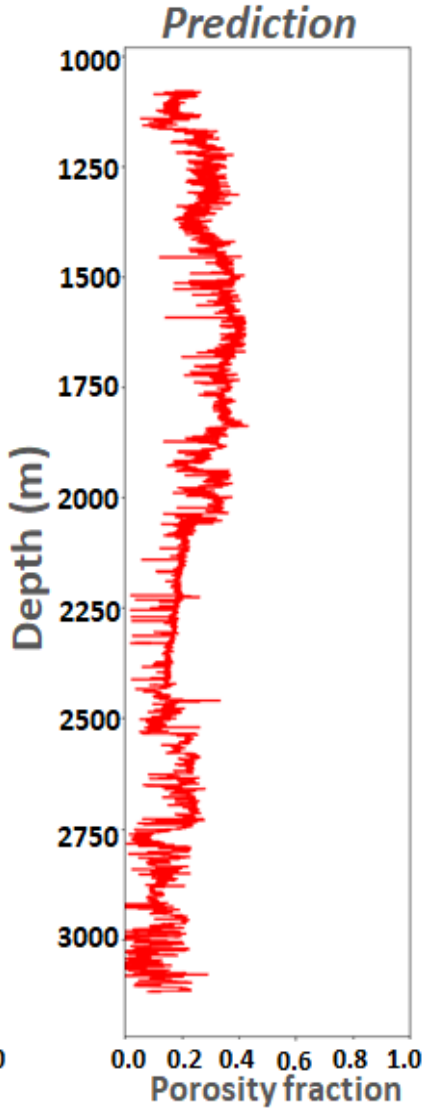
epcc

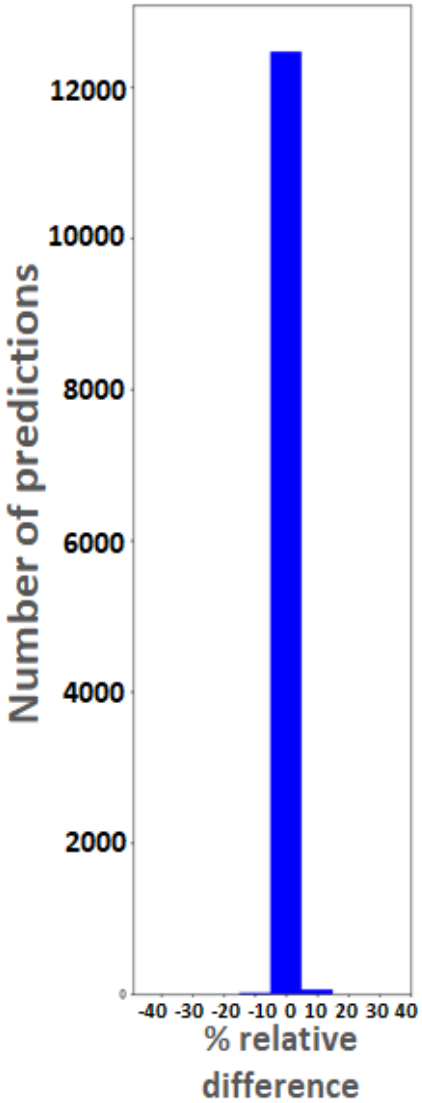
THE UNIVERSITY OF EDINBURGH

Porosity









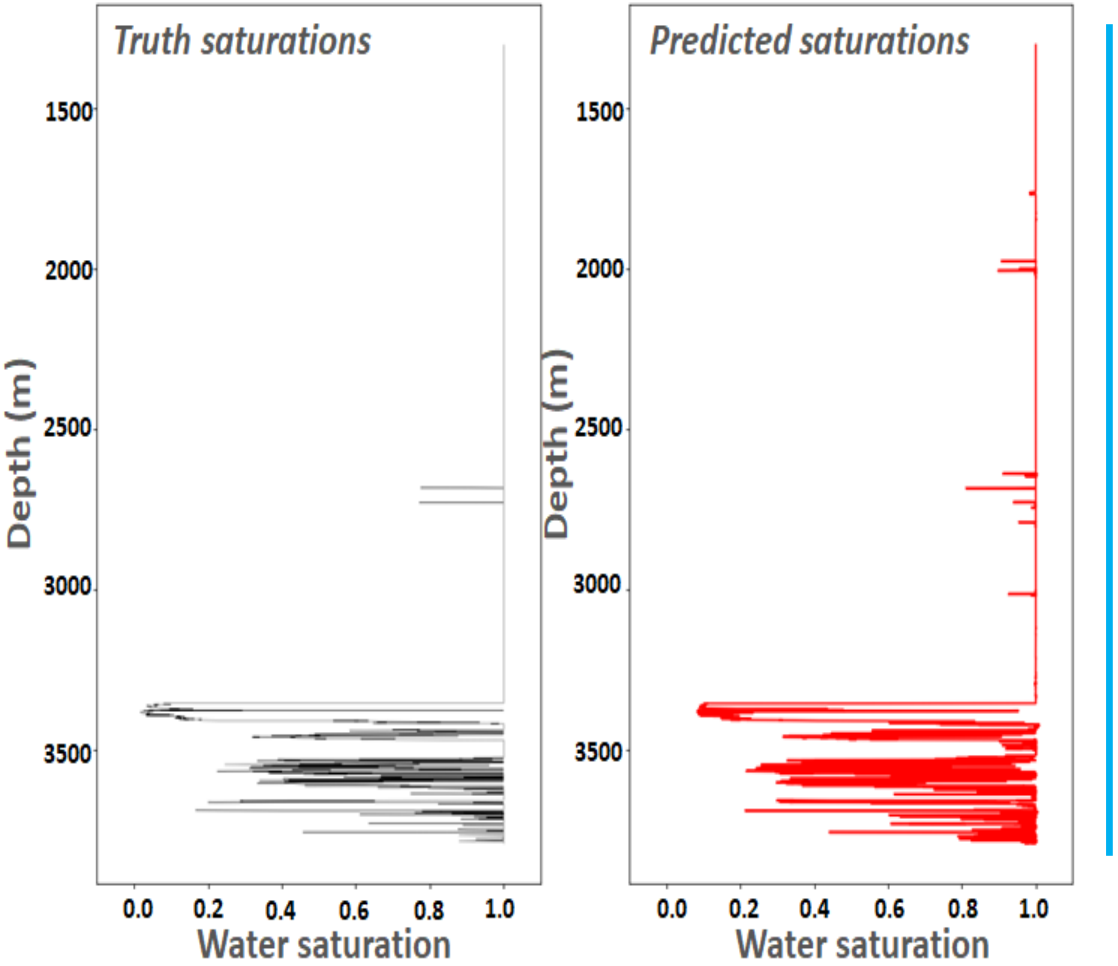
- Just works! The vast majority of our predictions are within 0 to 10% of the truth value and this is highly accurate.
- But it is crucial to use the cleaned p-wave and density curves, but not as important to use the mineralogy and there is only a small impact in prediction accuracy

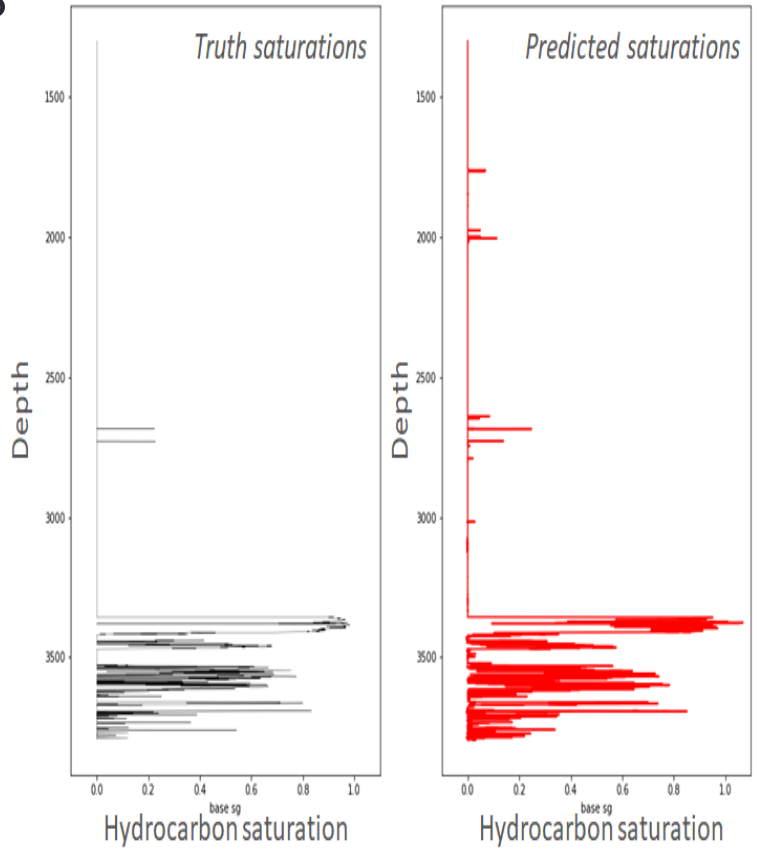
epcc

THE UNIVERSITY OF EDINBURGH

Fluid saturations

- The petrophysicists suspect a reservoir contains fluid
  - Is it hydrocarbons (oil and gas) or water?



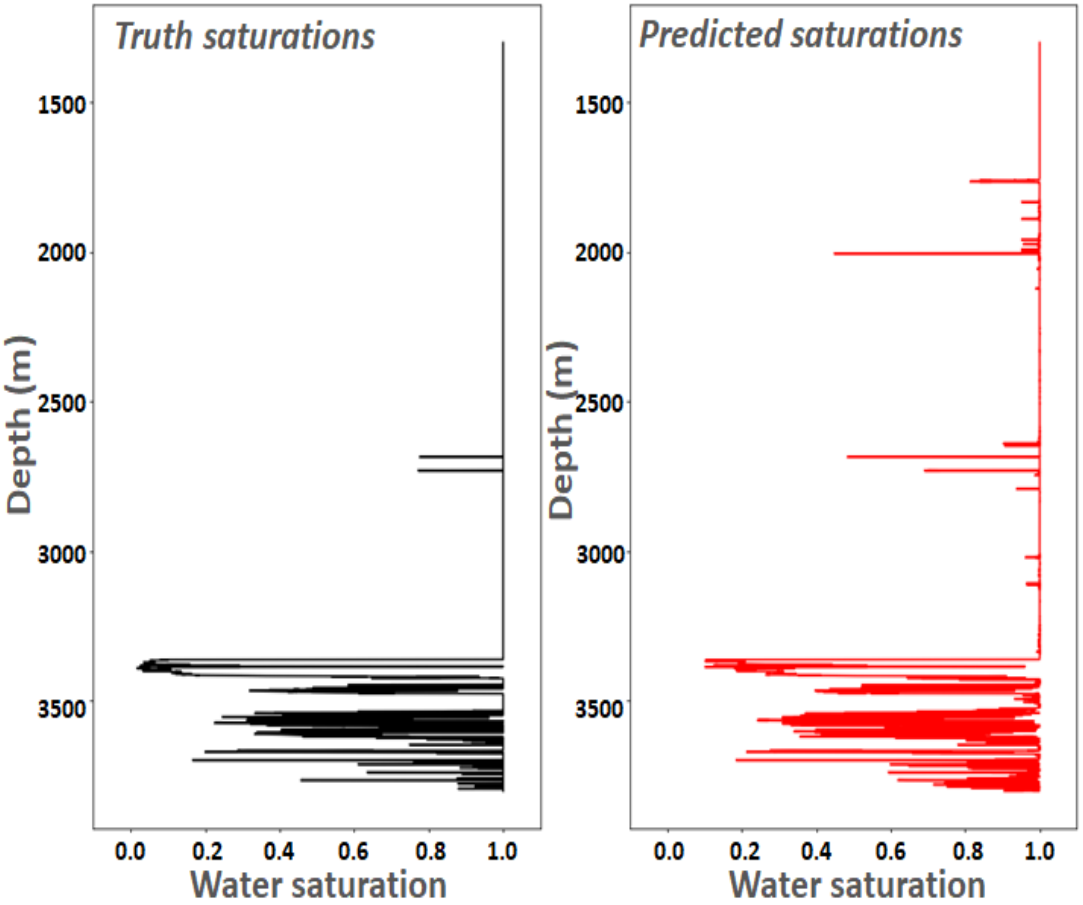


epcc

THE UNIVERSITY OF EDINBURGH

Fluid saturations

- It is either water or hydrocarbons. Water is far more common, so the model is more familiar with this quantity. Hence predict water and invert it to get hydrocarbons



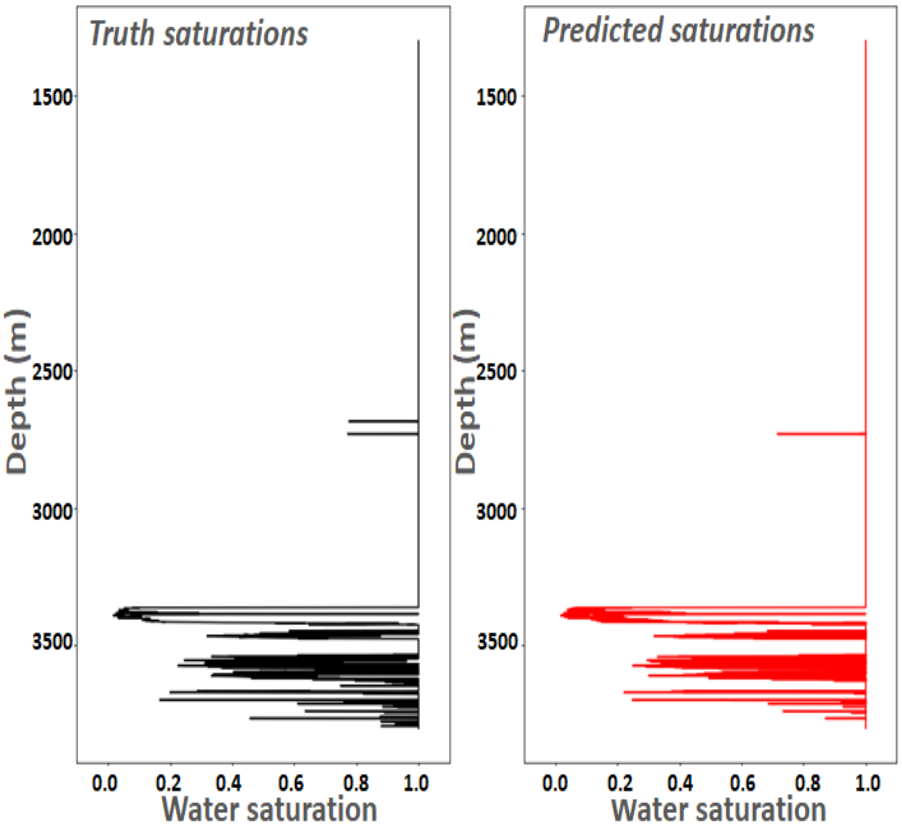
- The regression model wiggles around 100%, so we use a boosted trees classifier and then only regression on points not entirely water
  - But this tends to under predict water, as the classifier thinks it sees something and then the regressor predicts some value here

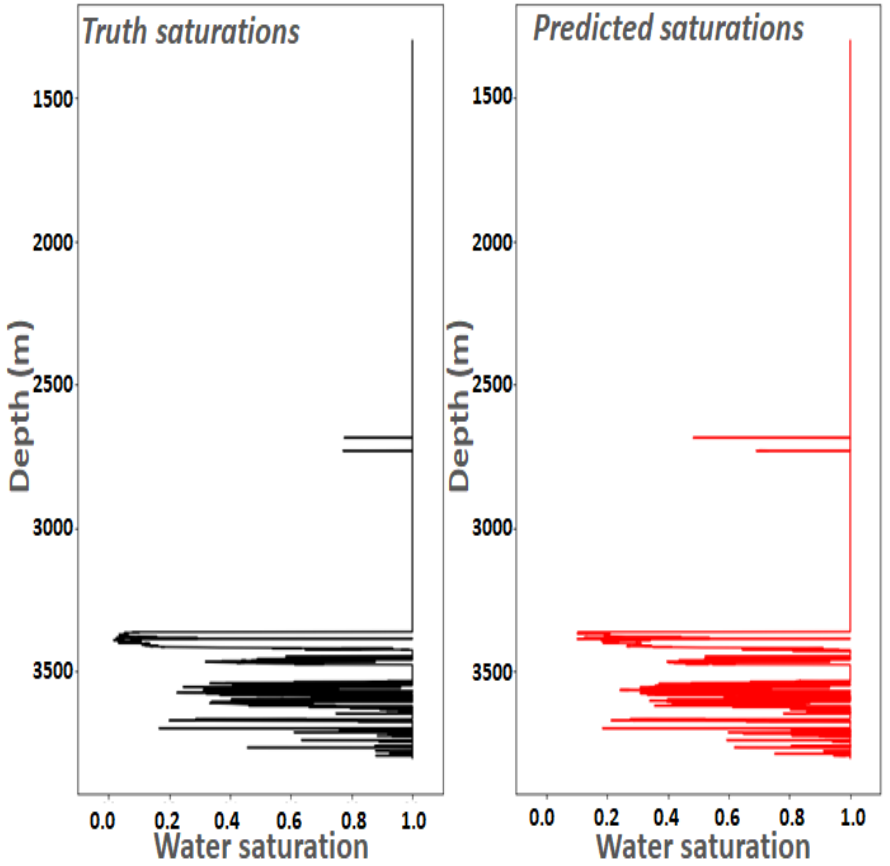
epcc

THE UNIVERSITY OF EDINBURGH

Fluid saturations

- Instead, a deep neural network, but this is the opposite; good at classification but not so much regression



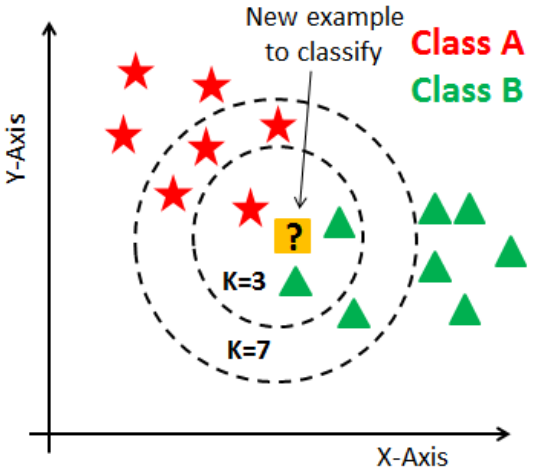


- In the end we found the best approach was to mix the models
  - Use a DNN for the classification and then boosted trees for regression

epcc

THE UNIVERSITY OF EDINBURGH

Lithoclass determination



- The lithology, or facies, is the geological rock type
  - This is fairly simple to classify and many people have had success here
  - But how simple can we get?
- We started with data without any lithology information (which is common), applied some very general rules to label this and trained our models on this data using K nearest neighbours.
  - The question was how much accuracy this would give us
  - It is not perfect but is reasonable!

Facies	HC Sand	Shale	Shaly Sand	Wet Sand
HC Sand	673	8	0	13
Shale	2	27833	2259	653
Shaly Sand	4	1735	2431	1228
Wet Sand	18	306	2313	9369

epcc

THE UNIVERSITY OF EDINBURGH

Taking advantage of the XC30

- Our boosted trees models are highly sensitive to seven hyper-parameters
  - These are interlinked and changing one will impact the optimal value of others

Name	Description
colsample_bytree	Sub sample ratio of columns when constructing each tree
eta	Step size shrinkage, to prevent over-fitting
gamma	Minimum loss reduction required to further partition a node
max_depth	Maximum tree depth, the deeper the tree the more complex the model and likely to overfit
min_child_weight	Minimum sum of instance weight needed in a child
num_rounds	Number of boosting rounds to perform
subsample	Sub sample ratio of the training instances, useful to prevent over-fitting

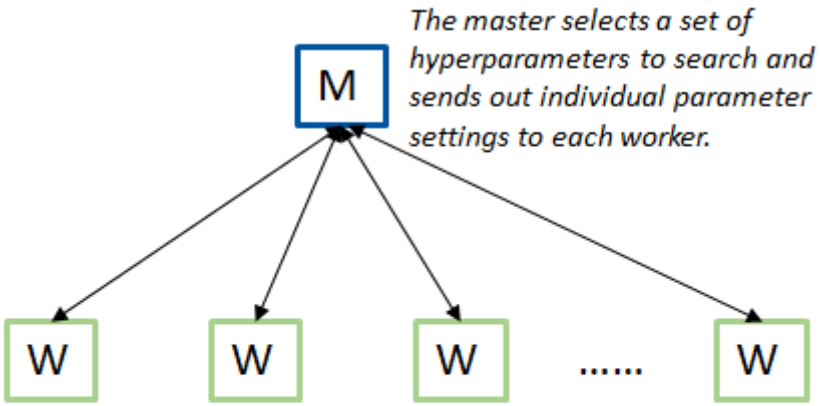
- Used Hyper-opt Python library to search through the parameter space
  - Each run generates an error metric, which it then uses to intelligently chose the next hyper parameters and runs again
    - Via a tree-structured Parzen estimator
  - Typically it takes between 120 to 160 runs to find *good* hyperparameters, with each individual run taking around 15 minutes.
  - Hyper-opt is serial, so it takes a long time to train the models
    - Once they are trained, runtime for inference is less than a second

epcc | THE UNIVERSITY OF EDINBURGH

Taking advantage of the XC30

- Model training time was a major issue
  - Especially for experimenting with the data, as if using an untuned model we don't know if it's the data or the lack of tuning

- We parallelised Hyperopt using MPI4Py and typically run over twenty nodes of ARCHER



Workers execute a single boosted trees model per NUMA region, using the hyper-parameters supplied by the master. Once the model training has completed they send back the corresponding absolute error of the model and wait for the hyper-parameter settings for the next model training iteration.

- Very simply parallelisation via the master/worker pattern and MPI4Py meant it this was less than an hours work
  - But model training time went from between 20 and 40 hours, to between a half and one hour.

epcc

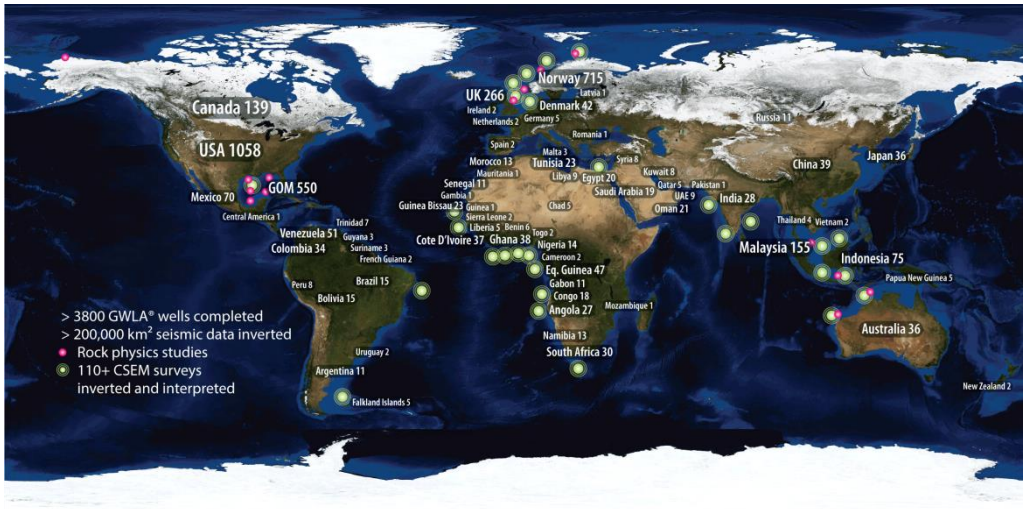
THE UNIVERSITY OF EDINBURGH

Conclusions & further work

- The good news for the petrophysicists is that they won't be replaced any time soon!

Method	Clay prediction RMS error
Base	0.1364
MLP (Sklearn)	0.1772
DNN (Pytorch)	0.0651
Boosted trees	0.1003
Boosted trees (missing data for training)	0.0838

  - This is a complex and involved process which requires their knowledge and expertise, but we believe ML can assist here
  - There is significant low hanging fruit when it comes to machine learning and data science frameworks that we are the community can help with



This works as a general tool

- A first pass to refine the data fairly significantly before interpretation
  - Optimise the use of the human
- Quick pass for a client to check whether further analysis is worth it

epcc | THE UNIVERSITY OF EDINBURGH