

ACHIEVING NET ZERO: APPLICATION OF MACHINE LEARNING MODELS TO COST-EFFECTIVELY CHARACTERISE THE SUBSURFACE FOR AN EMERGING CCS INDUSTRY.

Research Methods and Professional Development
MSc International Innovation Management

Module code = ACFA7004
Supervisor = Graham Howe

Assignment 1 – Research Proposal
Submission date: 03rd May 2022

John Ford (Student No. = 2115200)
(2115200@student.uwtsd.ac.uk)

Achieving Net Zero: Application of Machine Learning models to cost-effectively characterise the subsurface for an emerging CCS industry.

Contents

1 Introduction	3
1.1 Research project aims.....	3
1.2 Research study rationale	3
1.3 Research study boundaries and key stakeholders.....	3
2 Background and literature review	5
2.1 An overview of Machine Learning	5
2.1.1 Supervised Machine Learning.....	5
2.1.2 Un-supervised Machine Learning	5
2.3 Applications in Earth Science	6
3 Methodology.....	7
3.1 Methodology to be employed in this research.....	7
3.1.1 Training dataset generation: Geological analysis	7
3.1.2 Supervised Machine Learning.....	7
3.1.3 Un-supervised Machine Learning	8
3.2 Ethical issues	8
3.3 Materials, software, costs and timings	8
3.3.1 Materials	8
3.3.2 Software	9
3.3.3 Costs.....	9
3.3.4 Timings	9
4 Results	10
4.1 Qualitative results	10
4.2 Quantitative results	10
5 Evaluation	11
5.1 Potential limitations to this research.....	11
5.1.1 Technical limitations	11
5.1.2 Other constraints	12
5.2 Conclusion.....	13

6 References	14
Appendix 1: Present day emissions, the carbon cycle and effective CO ₂ storage	16

1 Introduction

To reach the UK Government's ambitious target of Net Zero CO₂ emissions by the year 2050 (<https://www.gov.uk/government/publications/net-zero-strategy>), significant progress in CO₂ capture, transport and storage must be made. To this end, British industry plays a critical role, developing the technologies and procedures for effective carbon storage. However, the application of any industrial strategy has to be commercial in nature; i.e. both privatised and public industrial bodies will need to generate a profit whilst implementing schemes to either reduce or capture and store CO₂ emissions.

1.1 Research project aims

This research project aims to solve two problems that persist within the emerging CCS industry:

- Can the application of Machine Learning models provide a more cost-effective benefit to subsurface characterisation?
- Can the application of Machine Learning models provide more robust geological interpretation where direct geological observation is not possible?

To answer these questions, this research project will undertake geological analysis on a subset of core material and then build a series of supervised and un-supervised Machine Learning models to predict Facies Associations outside of cored intervals.

If successful, this research project will reduce geological uncertainty in the subsurface through auditable facies predictions, whilst producing scalable results quickly. Furthermore, it is hoped the results of this research project will provide a robust groundwork for further, large-scale geological modelling which is currently outside the scope of this research proposal, but, will be crucial for advancing development of selected CCS sites.

1.2 Research study rationale

At present, there is no unit price of CO₂ per tonne, meaning a global commodity market (similar to that of Oil & Gas trading) is yet to emerge. The profitability of carbon storage is, therefore, yet to be seen. However, to reach Net Zero CO₂ emission targets by 2050, industrial application of carbon storage must happen now.

This paradox is one of the key reasons why industrial-scale carbon storage is yet to take place in the UK; there is a fundamental need for cost-effective and highly efficient procedures to be created within many parts of British industry that are focussed around Carbon Capture and Storage (CCS) so that projects can move forward.

By implementing multiple Machine Learning models, this research project will create a cost-effective workflow through semi-automation that will help also to reduce uncertainty in CCS storage sites. This is an important step in trying to make carbon storage a more commercially viable prospect and is why this research project is of importance.

1.3 Research study boundaries and key stakeholders

This research project will focus on geological material and wireline log data from the East Irish Sea Basin. It will generate predicted facies association data but will not use this data as input into any larger geo-cellular models. Such work is outside the scope of this project.

The key stakeholders of this project will be geologists and geological companies who are focusing their efforts on screening and developing CCS sites, principally within the East Irish Sea Basin (although the workflow generated in this project can be exported to other geographic regions if required).

This research proposal will now present a short literatures review, followed by the methodology that will be employed in the project. This is then followed by a brief discussion on the likely results and a final discussion on evaluating the limitations of this project.

2 Background and literature review

2.1 An overview of Machine Learning

The importance and usage of Machine Learning continues to grow and is now applied in most industrial, commercial and scientific/technological sectors across many organisations (Boehmke & Greenwell, 2020). It is commonly used to seek patterns in data that are too subtle for human observation, or to conduct operations that need to 'learn' from large volumes of data (Boehmke & Greenwell, 2020). For effective Machine Learning models to be built, it is also important to know the format of the data that will be 'fed' into the model. Data formats can be categorized into three types; structured data (such as tabular data), un-structured data (such as images, or text strings) or semi-structured data where datasets are connected by nodes (Abukmeil *et al.*, 2021). Structured datasets provide the best inputs, particularly in supervised models.

Machine Learning is classed as a branch of Artificial Intelligence (Tahmasebi *et al.*, 2020) and can be split into two categories:

2.1.1 Supervised Machine Learning

Supervised Machine Learning seeks to build a predictive model (Boehmke & Greenwell, 2020), by generating a mathematical tool that generates an accurate prediction (Kuhn & Johnson, 2013). Within supervised Machine Learning, a predictive model is built using an input dataset which is used to 'train' an algorithm. Commonly it a supervised approach uses numeric data that is 'labelled', and maps this labelled input datapoint to an output datapoint so that a series of input – output pairs are created. Knowing this relationship, it is then possible to predict one when the other is absent. For effective supervised Machine Learning, many, large training datasets need to be generated. Since this labelling is often done manually, it can be quite labour intensive, particularly if the labels become 'outdated' and need to be constantly monitored/updated by humans (Chen & Liu, 2016).

Supervised Machine Learning can be split into two subcategories: Linear regression and Classification, which have been the fundamentals of modern Machine Learning (Watt *et al.*, 2016).

Linear Regression: This type of supervised Machine Learning works best with numeric datasets, where a numeric output can be predicted based on a numeric input (Boehmke & Greenwell, 2020). This type of analysis can be used, for example, to predict house prices based on the square footage of a house.

Classification: This type of supervised Machine Learning works best for non-numeric, categorical problems (Boehmke & Greenwell, 2020) and uses a decision tree to predict an output from a series of inputs. Importantly, the probability of a classification (i.e., no = 0.4, yes = 0.6) is calculated during this modelling. As such, a numeric output (probability) can be predicted, with highest probability becoming the predicted classification (Boehmke & Greenwell, 2020).

2.1.2 Un-supervised Machine Learning

In contrast to supervised Machine Learning, Un-supervised Machine Learning does not need a training dataset. This method learns patterns from input data without any labelling. Fundamentally, un-supervised Machine Learning looks to segregate input data in groups (Boehmke & Greenwell, 2020). It does this using two main methods.

Clustering: This type of un-supervised Machine Learning segregates input data into groups based on similar variables (Boehmke & Greenwell, 2020). K-means clustering and Self Organising Maps are examples of this type of cluster analysis.

Dimensionality reduction: If the dataset is very large, or if there are a large number of features within the input dataset that correlate to one another, it is necessary to reduce the number of variables that are being analysed within the model. This is classed as dimensionality reduction, with Principle Component Analysis (PCA) a common tool for correlation after dimensionality reduction (Abukmeil *et al.*, 2021, Boehmke & Greenwell, 2020).

All in all, the process of applying Machine Learning to any dataset is iterative and, with limited knowledge of the problem, it is difficult to know which Machine Learning process/model to employ (Boehmke & Greenwell, 2020); known as Wolpert's (1996) "no free lunch" theorem. To this end, for this research project, it is vital to know the data structure being fed to the Machine Learning model and what type of results are required to be able to apply the most effective model.

2.3 Applications in Earth Science

Fundamentally, this research project aims to generate Facies Association data from wireline logs, using a small subset of geological descriptions from core material. A Google Scholarly search brings back ~23,500 results for the search term "machine learning facies classification". This indicates that there is a large, collective experience of applying this type of technique to derive Facies Association data from wireline log data, but most of this information is connected with the exploration oil and gas sector.

Machine Learning has been applied in the geological sciences at multiple scales, from the micro scale (analysing thin section images; Tahmasebi *et al.*, 2020), to the well log and seismic scales (Alaudah *et al.*, 2019). A combination of supervised and un-supervised Machine Learning models have also been used. For example Hall (2016) uses a classification based approach (using a Support Vector Machine; SVM), to determine the relationship between wireline log data and Facies Association and achieve a good fit between modelled results and actual data. In Hall (2016), all data used within the model was within the public domain and the model was ran using the open-source Python programme language. This is very similar in set-up to this suggested research proposal, and so, to some degree, validates the approach taken here.

3 Methodology

The methodology employed in this study can be broken down into three sequential phases; 1) training dataset generation through geological analysis, 2) supervised Machine Learning, and 3) Un-supervised Machine Learning. Each phase consists of a sequential number of steps, with each step building on the next.

3.1 Methodology to be employed in this research

3.1.1 Training dataset generation: Geological analysis

All successful Machine Learning projects require a robust, non-biased and representative training dataset. For this research project, the training dataset will comprise two datasets:

Dataset 1: geological interpretation of core material: This represents the fundamental building blocks of this study. It is qualitative in nature, based on a varying degree of human interpretation of selected geological material and will be undertaken using the following steps:

- A. First, all core material available within the study will be analysed by applying a non-subjective, fully descriptive scheme. Its purpose is to simplify the complexity of natural world and allow similar intervals of core material to be called the same, making comparison between separate wells more accurate. A descriptive scheme using alphanumeric codes, such as that proposed by Reijers (1993) Will be applied;
- B. Once step A is complete, a more subjective interpretation of the cored material will be made. This will group the results of step A into coarser/less detailed packages that focus on how the geological material formed (e.g. package A = sandstones forming in the centre of a river system, package B = siltstones forming in a floodplain). These packages are typically referred to as 'Facies Associations' and can be prone to operator bias. Suggestions on how to remove this bias are provided in Chapter 5, along with notes on sampling representation.

Step B results are the most useful and will be used as the geological interpretation input within the training dataset.

Dataset 2: wireline log data: These data are quantitative in nature and directly record physical parameters of geological material. For consistency across the study dataset, wireline data will have to be quality controlled, with erroneous values nullified. Calibration and possible normalisation of wireline log responses between wells will also take place, so that any differences observed between wells is geological in nature and not an artefact of initial data acquisition.

3.1.2 Supervised Machine Learning

Once complete, both datasets 1 and 2 will be loaded into an RStudio project and joined together to form the main input/training dataset. The Key point is that each wireline data point will now be labelled with a facies association, and it is this labelling that will 'train' the supervised model.

A classification based supervised model will then be applied to the 80% of the training dataset and this model will predict Facies Association based on wireline log input (Figure 1). The results of the initial model will be iteratively compared to the actual results of the input data, and tweaks made to the input data will be made until there is a good fit between model vs. real outputs. Results will be visualised by a simple correlation matrix and, once a certain confidence value has been reached for each facies association, the model will be propagated across all wireline log data in the study.

3.1.3 Un-supervised Machine Learning

After completion of the supervised Machine Learning model, an additional un-supervised Machine Learning model will be built. A clustering approach, most likely using Self Organising Maps will be applied and results iteratively compared to the input classifications (i.e. Facies Association labels). Results will once again be visualised by a simple correlation matrix, and the model will only be propagated across the full dataset until a certain confidence value has been reached for each Facies Association.

In both the supervised and un-supervised model approaches, the result will be a variably robust, computed Facies Association 'value' that extends outside of any cored interval. This is very important as it allows companies developing CCS sites to gain a greater understanding of the subsurface with potentially limited geological data. This process is also relatively quick (dependant on the geology), resulting in cost savings through reduced personnel time.

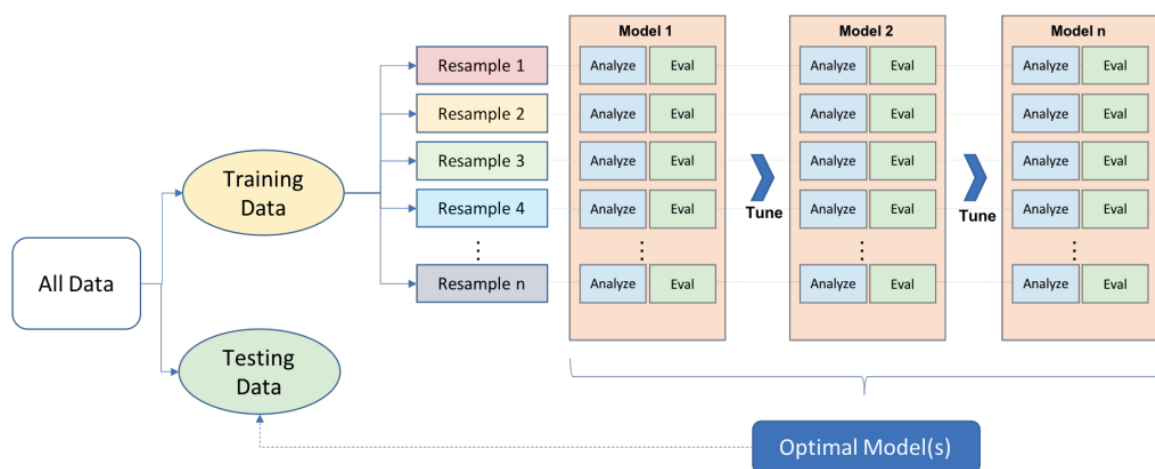


Figure 1: Standard workflow approach for supervised Machine Learning modelling. From Boehmke & Greenwell, 2020

3.2 Ethical issues

There are no direct ethical issues associated with applying Machine Learning in this context, with the exception that increased automation derived from Machine Learning may contribute to human operator redundancies. This is a much larger issue that cannot be addressed in this research project and so will not be discussed any further.

3.3 Materials, software, costs and timings

3.3.1 Materials

The majority of this project will be desk-based research. However, geological description of core material will be required. This will take place at the BGS Keyworth site in Leicester. Wireline log data will also be required. This dataset is in the public domain and can be downloaded from the NDR website (National Data Repository, <https://ndr.nstauthority.co.uk>)

3.3.2 Software

Wireline log data will be loaded into ALT WellCAD software (<https://www.alt.lu/products-wellcad>). Whilst this is not a dedicated petrophysical platform, it will be possible to perform calibration and normalisation of wireline log data between wells using this software.

Both supervised and un-supervised machine learning models will be written using the R programme language and compiled within R Studio. The R programme language is open-source and is ideally suited for statistical analysis (<https://www.r-project.org/about.html>).

Office applications such as MS Excel, MS Word and MS PowerPoint will also be used throughout the project.

3.3.3 Costs

The table below provides estimated costs for this research project:

Task	Unit	Unit rate	Unit No.	Cost
Describe geological material at BGS	day	£200	5	£1,000
Accommodation	day	£150	4	£600
WellCAD licence	Year	£5,000	1	£5,000

Table 1: estimated research project costs

A cost of £6, 600 is estimated for this project. Funding may be acquired through industrial partnership. Some additional costs associated with R programme language training may also be accrued. Estimated cost of this training is around £10 per month (<https://www.datacamp.com/pricing>).

3.3.4 Timings

The Gantt chart below provides and estimated timing for this research project. In total, this research project is estimated to take a minimum of 23 weeks to complete.

	Task	Wk 1	Wk 2	Wk 3	Wk 4	Wk 5	Wk 6	Wk 7	Wk 8	Wk 9	Wk 10	Wk 11	Wk 12	Wk 13	Wk 14	Wk 15	Wk 16	Wk 17	Wk 18	Wk 19	Wk 20	Wk 21	Wk 22	Wk 23
	Source financial support from academic grants or industrial partnership	Unknown, but work sponship is likely																						
Generate training data	Geological description of material at BGS Keyworth site, Leicester																							
	Apply lithofacies scheme to geological description																							
	Apply facies association scheme to geological description																							
	Access and load wireline logs																							
	Calibrate and/or normalise wireline logs																							
	Join datasets together to generate training dataset																							
	R Programme training																							
Supervised model	Run supervised model on training dataset and iterate																							
	Propogate across full wirelinre log dataset																							
Un-supervised model	Run un-supervised model on training dataset and iterate																							
	Propogate across full wirelinre log dataset																							
	Project write-up																							

Table 2: Estimate research project Gantt chart

4 Results

The results collated and generated in this proposal, will be both qualitative and quantitative in nature. Qualitative results will be generated at both an early (input/training-data) and later (Supervised and un-supervised Machine Learning) stage. Qualitative results will be generated only at the later stage of this research, occurring during and after the supervised and unsupervised Machine Learning component.

4.1 Qualitative results

Early Stage: A consistent set of input data comprising lithofacies and facies association 'interval' data will be generated. They represent non-subjective, geological description (lithofacies) and potentially more-subjective interpretation of selected vertical sections of rock. These qualitative data will exist as a simple table, comprising four parameters: Well, Top, Base and Code. The code will be alphanumeric in nature, and will reflect either a descriptive or interpretive interval data type.

Later Stage: Both the supervised and un-supervised Machine Learning models generated in this research will generate qualitative data. This will comprise a table output comprising three parameters: Well, Depth and Code. These data will be used for statistical comparison between input and modelled output data, highlighting confidence within the supervised and un-supervised models.

4.2 Quantitative results

Later Stage: As part of the supervised Machine Learning model, probability/confidence scores will be generated. These will be fully quantitative in nature and represented as a correlation matrix comparing input data to supervised Machine Learning modelled data. This matrix and associated scores will be used as a threshold for iteration within the supervised model. A probability/confidence score of $\geq 85\%$ will need to be achieved for each input facies association before propagating the constructed algorithm across the whole dataset, generating a set of automated output data.

5 Evaluation

This research proposal provides a methodology that can help generate a more robust, ground-truthed geological model of the subsurface, using data already available in the public domain in combination with modern advances in Machine Learning capabilities. It focusses on providing a cost-effective approach to characterising key elements related to storing CO₂ at depth by utilising modern-day powerful desktop computation and access to multiple, open-source programme languages. Once sufficiently ‘trained’, both the supervised and un-supervised model can be propagated almost instantaneously over many individual wells. This provides significant cost savings in terms of personnel time, as well as producing results that are ‘auditable’ (i.e., the output results are directly related to the input data and selected algorithm).

However, as in all methodologies, there are potential limitations to this approach as discussed below.

5.1 Potential limitations to this research

It is important to note the potential limitations in this research, which can be broken down into two broad categories:

5.1.1 Technical limitations

There are four main technical limitations to note:

Inherent bias in training data - interpretation: Generating an input dataset that is consistent between wells is vital to the methodology employed in the research proposal. If the initial input dataset is not applied consistently between wells at the start, inherent bias towards one interpretation or exclusion of another interpretation at the input stage will ultimately affect the results of the final Machine Learning models. Suggestions to reduce this inherent bias are listed below:

1. Apply a consistent lithofacies scheme at day 1 of the research project. This lithofacies scheme should be entirely descriptive and able to be applied consistently by different human operators if needed.
2. Ideally, to remove as much interpretation bias as possible, the following should be applied:
 - a. Only one operator should provide the interpreted facies association data. If this is not possible/practical, then an interpretation scheme based exclusively on the non-subjective, descriptive lithofacies should be created and used by all operators involved in the interpretation stage of the study;
 - b. Iterative review and calibration between each operator should be undertaken throughout the interpretation phase. This should reduce operator bias in the input dataset as much as humanly possible (please note: it is very likely that some bias will be present in the final input dataset, even after undertaking these precautions. This is because of the nature of human interpretation and is an unavoidable flaw of using this type of input data within Machine Learning projects).

Inherent bias in input data - calibration: Inconsistent calibration across the input dataset will also lead to artificial variation in some input parameters, increasing uncertainty/reducing the probability scores of the final Machine Learning models by artificially affecting the correlation between input dataset 1

(facies associations) and input dataset 2 (Wireline logs). Suggestions to reduce this inherent bias are listed below:

1. Calibrate wireline log data consistently between individual wells, normalising input values where possible. If applied properly, variation in wireline data between wells should be a product of the natural geological system, not an artefact of data acquisition. Fundamentally this will increase the overall confidence in the supervised and un-supervised model.

Inherent bias in training data – poor data: Dependant on a number of different parameters, some input data are likely to be poor or, in some circumstances, it may not be possible to record input data over certain intervals. If poor data is included within the training dataset, this may lead to artificial correlations between input datasets, reducing confidence in the Machine Learning model results. Suggestions to reduce this inherent bias are listed below:

1. Poor data should be flagged/tagged, nullified and excluded from the input dataset. Although this may yield ‘gaps’ in the un-supervised modelled output, removing this data will increase overall confidence in the model.

Inherent bias in training data – sample size: The training dataset used in the supervised phase will have a direct impact on the final, supervised Machine Learning model. If a certain interpretation is overrepresented in the input dataset, this may result in over-representation of the same interpretation in the final, supervised model. More worryingly, some interpretations may be under-represented or entirely excluded from the final supervised model. As such, it is key to have an unbiased input data set comprising a representative cross-section across the input dataset. Suggestions to reduce this inherent bias are listed below:

1. Use interpreted facies data that includes a variety of different lithologies (i.e. sandstones, siltstones and mudstones) from as many different wells as possible;
2. Consider omitting datasets that comprise significantly large proportions of sandstones; although these are likely to represent the best sites for CO² storage, if they are overrepresented in the final supervised model, the effective CO² storage capacity of the studied location will be exaggerated.

5.1.2 Other constraints

There are two other potential constraints to be considered:

Data access constraints: Most data required for this research proposal is essentially in the public domain. Access to core data is through the BGS (British Geological Survey, <https://www.bgs.ac.uk/geological-data>) where rock material is stored in the BGS Keyworth site in Leicestershire. If core was cut, a set of this material should be accessible through the BGS. Although the core is available within the public domain, a small admin fee and room hire charge is still required to view the relevant rock material.

Wireline log data is accessible through the UK NDR (National Data Repository, <https://ndr.nstauthority.co.uk>), an online portal hosted by the UK Government that provides access to archived data that has exceeded any initial confidentiality agreement/exclusivity period with the operator that first acquired the data. However, it is not guaranteed that all data that should be

supplied to the NDR has been supplied by independent operators and, furthermore, it is not guaranteed that all data required for this project is available in a ready-to-use digital format. In the eventuality that the required wireline data is not on the NDR portal, one must track down the original operator who acquired the data and request a copy. Given the vintage of this data (from the 1970's – 1990's), tracking down an original copy of the data through the original operator may prove to be very difficult. Furthermore, if the data is available only in PDF format, time (and associated costs) must be made to digitise this data so that it can be used as an input into the supervised and un-supervised model.

Financial constraints: The following financial constraints are likely to relate to this research project:

1. Data access/admin/room hire charges for accessing data through the NDR and BGS. These are not likely to be cost prohibitive;
2. Software licence fees. For loading, manipulation and calibration of wireline log data, access to petrophysical software would be required. An annual licence fee could range from ~£5k for ALT's WellCAD software up to ~£18k - £20k for LR Synergy's IC software;
3. Licensed access to pre-existing, non-exclusive geological data. Over the last 3 decades, multiple consultancies have collected core data that could be easily and quickly incorporated as part of the input dataset required for this research project. By accessing this data, under licence, it will be possible to get access to larger volumes of descriptive geological data quicker and, in most cases, already in a digital format. However, this could be cost prohibitive, with the licence fees to access each well set by each consultancy. Values to access descriptive geological data per well could range from ~£500 - £1.5k, although the actual costs charged by each/any consultancy are unknown.

To overcome these likely financial constraints, an industrial partnership is suggested. This partnership will provide the research project access to relevant software's at either a heavily discounted cost or at no cost at all. Furthermore, with the correct industrial partner, the licence fee to access relevant pre-existing geological descriptive data may be waived, or severely decreased as well. This will benefit the research project both financially, as well as provide quick access to data. Although the 3rd party (i.e., the industrial partner) may attempt to direct the project in a different direction to the original proposal/brief, the potential financial benefits of a partnership outweigh the potential negatives.

5.2 Conclusion

To conclude; the effective storage of CO₂ at depth will require detailed characterisation of the subsurface, particularly of rocks that are likely to trap and store CO₂ in a supercritical state. Furthermore, the trapping of CO₂ at depth will be on an industrial scale and so will need to be profitable at an industrial scale as well. This research proposal provides a method to generate robust geological data, through supervised and un-supervised Machine Learning that can act as direct inputs into larger geo-cellular storage models of the subsurface. Whilst there are limitations to the methodology suggested in this research proposal, namely potential inherent bias within the training dataset, it does provide a cost-effective way to generate auditable geological data at scale, with limited human operator requirements. Fundamentally, this key for industrial adoption, effective CO₂ storage and a better world.

6 References

Abukmeil, M., Ferrari, S., Genovese, A., Piuri, V., and Scotti, Fabio. 2021. A survey of Unsupervised Generative Models for Exploratory Data Analysis and Representation Learning. ACM Comput. Surv. Vol 1.

Alaudah, Y., Michalowicz, P., Alfarraj, M., and AlRegib, G., 2019. A Machine Learning Benchmark for Facies Classification. <https://arxiv.org/pdf/1901.07659.pdf> [accessed 03/05/2022]

ALT WellCAD Software information:

<https://www.alt.lu/products-wellcad> [accessed 02/05/2022]

Atmospheric CO₂ emissions:

<https://ourworldindata.org> [accessed 02/05/2022]

Average temperature variation:

<http://berkeleyearth.org/global-temperature-report-for-2020> [accessed 02/05/2022]

Boehmke. B., and Greenwell, B., 2020. Hands-On Machine Learning with R. in: Chambers, J.M., Hothorn, T., Lang, T.D., and Wickham, Hadely., (Eds.) The R Series. Chapman & Hall/CRC, Taylor & Francis Group

British Geological Survey:

<https://www.bgs.ac.uk/geological-data> [accessed 02/05/2022]

Chen. Z., and Liu. B., 2016. Lifelong Machine Learning. In: Synthesis Lectures on Artificial Intelligence and Machine Learning. Eds. Brachman. R. J., Cohen. W. W., and Stone. P. Morgan and Claypool Publishers

Ciais, P., Sabine, C., Bala, G., Bopp, L., Brovkin, V., Canadell, J., Chhabra, A., DeFries, R., Galloway, J., Heimann, M., Jones, C., Le Quéré, C., Myneni, R.B., Piao, S., Thornton, P., 2013. Carbon and Other Biogeochemical Cycles. Chapter 6, in: Stocker, T.F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S.K., Boschung, J., Nauels, A., Xia, Y., Bex, V., Midgley, P.M. (Eds.), Climate Change 2013: The Physical Science Basis. Contribution of Working Group to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, pp. 465-570.

CO₂ Direct Air Capture:

<https://www.thinkgeoenergy.com/worlds-largest-direct-air-capture-and-co2-storage-plant-on-in-iceland> [accessed 02/05/2022]

Datacamp pricing:

<https://www.datacamp.com/pricing> [accessed 03/05/2022]

Doughty, C., Freifeld, B.M., and Trautz, R.C., 2008, Site characterization for CO₂ geologic storage and vice versa: the Frio brine pilot, Texas, USA as a case study: Environmental Geology, v. 54, p. 1635-1656.

Espinoza, D.N., Kim, S.H., Santamarina, J.C., 2011. CO₂ Geological Storage - Geotechnical Implications. Ksce Journal of Civil Engineering 15, 707-719.

GCCSI, 2020. Global status of CCS 2020. Global CCS Institute.

Hall, B., 2016. Facies classification using machine learning. Leading Edge. <https://library.seg.org/doi/pdfplus/10.1190/tle35100906.1> [accessed 02/05/2022]

Hepburn, C., Adlen, E., Beddington, J., Carter, E.A., Fuss, S., Mac Dowell, N., Minx, J.C., Smith, P., Williams, C.K., 2019. The technological and economic prospects for CO₂ utilization and removal. Nature 575, 87-97.

Kuhn, M., and Johnson, K., 2013. Applied Predictive Modelling. Vol 26, Springer.

National Data Repository:
<https://ndr.nstauthority.co.uk> [accessed 02/05/2022]

R programme language: open-source access
<https://www.r-project.org/about.html> [accessed on 02/05/2022]

Ringrose, P., 2020. How to store CO₂ underground: Insights from early-mover CCS projects. Springer, Switzerland.

Reijers, T.J.A, Pestman, P.J., Mijnlief, H.F., and Kouwe W.F.P., 1993. Lithofacies and their interpretation: a guide to standardised description of sedimentary deposits. Meded. Rijks Geol. Dienst 49.

Tahmasebi, P., Kamrava, S., Bai, T., and Sahimi, M., 2020. Machine learning in geo- and environmental sciences: From small to large scale. Advances in Water Resources. Vol. 142.

UK Government 'Track 1' CCS sites:
<https://www.gov.uk/government/publications/cluster-sequencing-for-carbon-capture-usage-and-storage-ccus-deployment-phase-1-expressions-of-interest/october-2021-update-track-1-clusters-confirmed> [accessed 02/05/2022]

Watt, J., Borhani, R., and Katsaggelos, A.K., 2016. Machine Learning Refined: Foundations, Algorithms and Applications. Cambridge University Press

Wolpert, D.H., 1996. The lack of a priori distinction between learning algorithms. Neural Computation. Vol. 8

Appendix 1: Present day emissions, the carbon cycle and effective CO₂ storage

Annual global temperatures are rising quicker than at any other time in human history, with 2020 recording an average warming of 1.27°C above the expected average (<http://berkeleyearth.org/global-temperature-report-for-2020>). This temperature rise corresponds with a significant increase in atmospheric CO₂ concentration, with modern concentrations exceeding 400ppm (~45% increase over the last ~200 years; <https://ourworldindata.org>). This increase in atmospheric CO₂ coincides with adoption and mass use of internal combustion engines, industrialisation and burning of fossil fuels to generate power.

Today, we emit nearly 40 Gigatonnes of CO₂ annually. To avoid catastrophic climate change, we need to immediately reduce emissions by 50% or more (Ringrose, 2020). Whilst a natural carbon cycle exists, whereby atmospheric CO₂ is captured through biological processes (both on land and offshore) and chemical weathering, increased emissions from human activity result in a net atmospheric increase (Ciais *et al.*, 2013; Espinoza *et al.*, 2011). To avert any additional damage to our global climate, industrial nations need to act fast, reducing emissions and implementing legislation to force large-scale industrial action.

Carbon Capture and Sequestration (CCS) is an industrial process for capturing and storing carbon, commonly at depth in the subsurface. Although there are many sites globally that might act as subsurface carbon stores (GCCSI, 2020), at present there is no large-scale industrial storage of CO₂ in the UK. However, as of 2021, two geographic regions have been 'high-graded' by the UK government for CO₂ storage. These are called 'Track 1 CCS clusters' and are focussed on Ellesmere Port and the East Irish Sea Basin, as well as offshore Hull and Newcastle in the Central North Sea (<https://www.gov.uk/government/publications/cluster-sequencing-for-carbon-capture-usage-and-storage-ccus-deployment-phase-1-expressions-of-interest/october-2021-update-track-1-clusters-confirmed>).

The industrial application of CCS can be broken into two broad elements:

Capture: Capturing CO₂ either at the source of emission or from direct air capture (e.g. the Orca plant on Iceland, built by Climeworks and Carbfix; <https://www.thinkgeoenergy.com/worlds-largest-direct-air-capture-and-co2-storage-plant-on-in-iceland> [accessed 02/05/2022]). CO₂ capture is primarily an engineering problem, using a mixture of mechanical-engineering, chemical-engineering and, more recently bio-engineering.

Storage: Trapping and Storing CO₂. This can be done through many different ways, from changing forestry and land management systems, to advanced geological weathering, to geological storage of CO₂ in the subsurface at depth (Hepburn *et al.*, 2019).

In this research proposal the focus is on geological storage at depth. Subsurface storage is most effective below 1km underground (Doughty *et al.*, 2008) in sandstones that are porous (i.e., sandstones that contain voids between grains). In this storage mechanism, CO₂ is commonly injected in a liquid form where it will convert into a super-critical state at depth owing to increased temperatures and pressures (Ringrose, 2020). This is very important; in a super-critical state it is much easier for CO₂ to move into the pores/voids within a given sandstone compared to CO₂ as a gas and as such be trapped more quickly. As CO₂ will be trapped in porous sandstones at depth, it is, therefore, very important to

understand the distribution of sandstones and the system that controlled their deposition and distribution.