169

Sustainable data mining: AI/ML-based parameter extraction, data visualization and connectivity to upcycle big-data for basin analysis

T. Looi[1], N.E. Arif[1], N.M. Hernandez[1], F. Baillard[1]

[1] Iraya Energies

Summary

In the upstream oil and gas sector, the processes of data mining, which involves searching, extracting, and validating information that sits within the technical documents, reports, presentations, and studies to understand exploration history and geological parameters are often challenging and requires vast resources to be completed. Yet, many geological information that have already been mined, are stored in spreadsheets or niche databases, that limits their abilities to be recycled for multiple uses within the organization. Basin information such as formation pressure, formation temperature, fracture pressure, drilling rate of penetration, total organic carbon (TOC) and lithologies are some of the typical parameters that are crucial to understand the basin scale geology, reservoir properties and identify opportunities within the area of interest and often needed to perform in depth workflows, such as seismic reservoir characterization, basin modelling and geomechanical studies, which cover multiple cycles of exploration, development and site development of future carbon waste disposals. The research demonstrates the application of AI/ML technologies coupled with interactive data visualization and API connectivity, that can significantly accelerate the extraction of knowledge that are sitting inside the reports and ensure sustainability in data mining activities.

Introduction

In the upstream oil and gas sector, the processes of data mining, which involves searching, extracting, and validating information that sits within the technical documents, reports, presentations, and studies to understand exploration history and geological parameters are often challenging and requires vast resources to be completed. Yet, many geological information that have already been mined, are stored in spreadsheets or niche databases, that limits their abilities to be recycled for multiple uses within the organization. Basin information such as formation pressure, formation temperature, fracture pressure, drilling rate of penetration, total organic carbon (TOC) and lithologies are some of the typical parameters that are crucial to understand the basin scale geology, reservoir properties and identify opportunities within the area of interest and often needed to perform in depth workflows, such as seismic reservoir characterization, basin modelling and geomechanical studies, which cover multiple cycles of exploration, development and site development of future carbon waste disposals.

In this paper, we are going to demonstrate on how we can improve on the traditional data mining workflows and enhance data sustainability by focusing process improvement on three areas: a.) accelerating the speed of geoscience parameter extraction by Artificial Intelligence/ Machine learning (AI/ML) techniques, b.) increase data readability and integrity by effective visualization of both source and output data, and c.) promote data reusability by enabling API access to extracted data.

Methodology

The advancement in ML and AI with the support of advancement in cloud computing has made a significant impact to the traditional data mining workflows. However, AI/ML workflows are typically focused only on the data extraction phase as a one-time effort, whereas an effective data mining strategy should also ensure data sustainability, which means the ability to visualize the data to identify trends and patterns easily, and pass that high-integrity extracted information intact to multiple type of users within the organization.

A sustainable data mining strategy is proposed to accelerate extraction with an automated pipeline using Machine Learning techniques such as Natural Language Processing (NLP) or Deep Convolutional Neural Network (DCNN) ingests all the unstructured data (Hernandez et al., 2019) in steps 1 to 3, followed by human-in-the loop quality control, visualization and data trackability and connectivity in steps 4 and 5. (Figure 1):

1. A vast amount of unstructured data such as final well report, technical reports, working files that varies from .pdf, .docx., .xlsx, .csv .jpg, .png and .tif are used as the main source of information and feed into the production ready ML pipelines for audit, duplicates, and version detections.
2. The unstructured data ingestion starts with the digitalization of data using Optical Character Recognition (OCR). Next, Deep Convolutional Neural Network (DCNN) classifies the extracted images into their respective geological categories such as map, seismic, stratigraphic chart, SEM, thin section, core and well logs. Simultaneously, Natural Language Processing (NLP) pipeline performs automated extraction and tagging of metadata.
3. Further analysis and knowledge extraction are available once the unstructured data is ingested. Data Science analytical tool such as deep search with heat map density allows additional insights where the user can monitor the trend of a parameter regionally. In addition, tables extracted from the reports are post-processed to retain the structure and extract the values.
4. Missing depth interpolation, units' standardization and plotting of values on the gradient isolines are part of the data validation procedures during the Human in the Loop quality control.

5. The last step is the discovery stage that allows shareable structured data among the team members for further interpretation and insights. The final shareable structured data allows data trackability within unstructured data, data export functionality, spatial filter to confine the search to the area of interest, depth filter and outliers' identification. The typical outputs of data mining exercise are the standard excel or csv as application agnostic format. In this stage, we also enabled API access to allow for easier connectivity to other geoscience platforms or internally developed digital infrastructure systems.

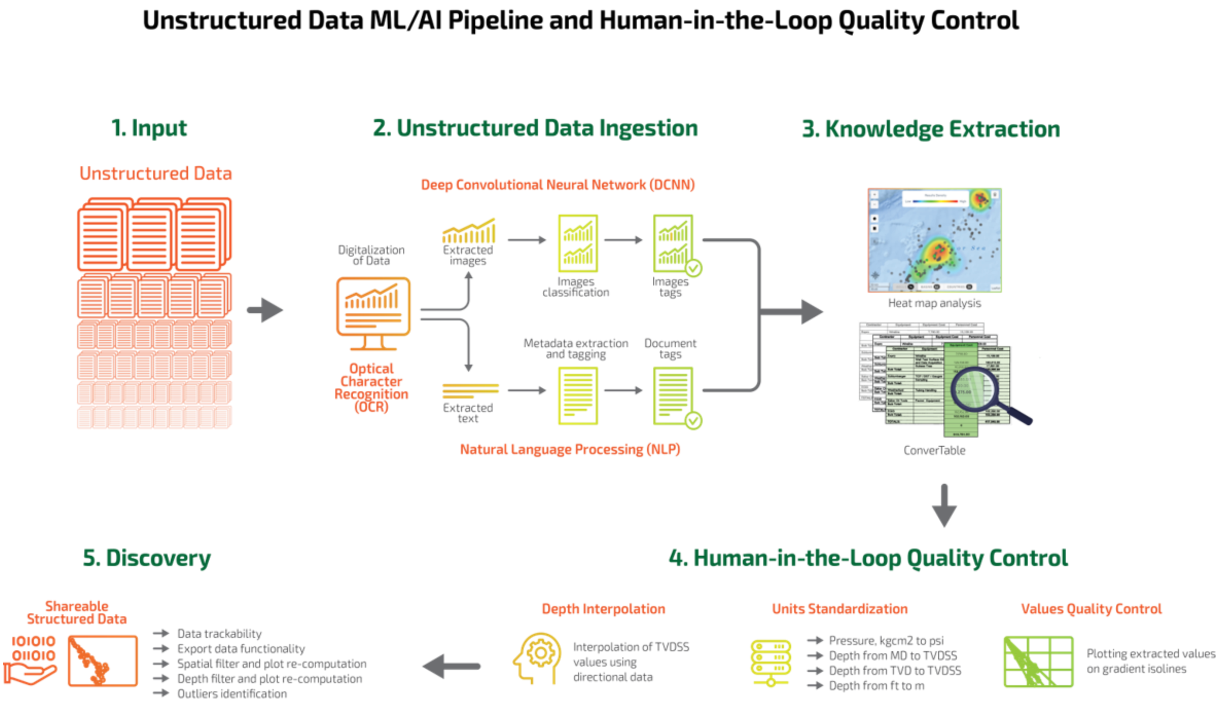
Unstructured Data ML/AI Pipeline and Human-in-the-Loop Quality Control
1. Input
Unstructured Data
2. Unstructured Data Ingestion
Deep Convolutional Neural Network (DCNN)
Digitalization of Data
Extracted images
Images classification
Images tags
Optical Character Recognition (OCR)
Metadata extraction and tagging
Document tags
Extracted text
Natural Language Processing (NLP)
3. Knowledge Extraction
Heat map analysis
ConverTable
5. Discovery
Shareable Structured Data
101010
011010
Data trackability
Export data functionality
Spatial filter and plot re-computation
Depth filter and plot re-computation
Outliers identification
4. Human-in-the-Loop Quality Control
Depth Interpolation
Interpolation of TVDSS values using directional data
Units Standardization
Pressure, kgcm2 to psi
Depth from MD to TVDSS
Depth from TVD to TVDSS
Depth from ft to m
Values Quality Control
Plotting extracted values on gradient isolines

***Figure 1** Full implemented workflow of ingestion and analysis of unstructured data using ML/AI*

We highlight the human-in-the-loop process during the data extraction process in step 4 above, which means there is a component of human interpretation during the geoscience parameter extraction process. Since there could be conflicting understanding on what is considered right or relevant data at the time the extraction was done, it is important to keep the sources of information intact, so these can be reviewed and updated when new additional data is available and the geological understanding of the basin of interest has evolved. The two additional steps introduced in the data mining process increases data integrity.

Results

In this study, unstructured data from over 500 oil and gas wells are processed on a regional scale this includes a total of 300,000 pages and 140,000 images. Six geological parameters, formation pressure, formation temperature, fracture pressure, drilling rate of penetration, total organic carbon (TOC) and lithologies were efficiently extracted, aggregated, validated, and finally visualized on scatter plots and pie charts. The output from the case study shows notable knowledge analysis (Figure 2) that provides insights on the regional consistency and information distribution of the area of interest and can be used as input into petroleum system modelling, reservoir characterization, idle wells review, geomechanical studies or potential carbon storage studies.

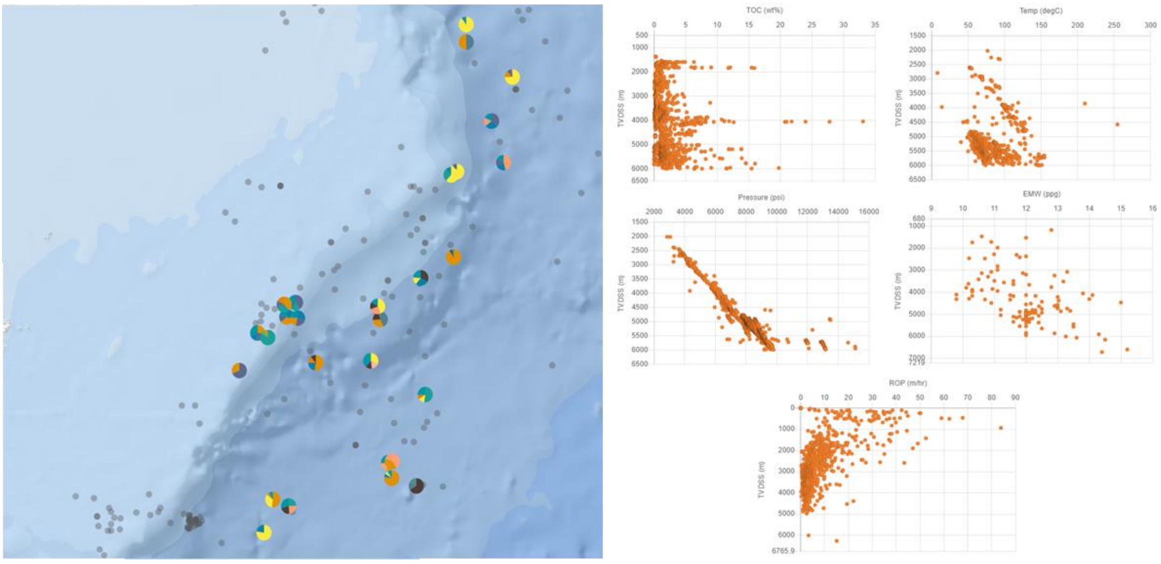

Figure 2 *Lithological pie chart distribution (left) and scatter plots of total organic carbon (TOC), rate of penetration (ROP), formation temperature, fracture pressure and formation pressure (right)*

Conclusions

The research demonstrates the application of AI/ML technologies coupled with interactive data visualization and API connectivity, that can significantly accelerate the extraction of knowledge that are sitting inside the reports and ensure sustainability in data mining activities. In the case study, we have presented a workflow on how six (6) geological parameters from over 500 wells were successfully extracted from unstructured data using ML and AI pipelines that can be used by multi-faceted subsurface teams for more in-depth analysis within the area, resulting to the upcycling of geological data across the full life cycle of a basin.

References

Hernandez, N. M., Lucañas, P. J., Mamador, C., & Panganiban, L. [2019]. Automated Information Retrieval from Unstructured Documents Utilizing a Sequence of Smart Machine Learning Methods within a Hybrid Cloud Container. *EAGE Workshop on Big Data and Machine Learning for E&P Efficiency 25-27 February.*

Mamador, C., Aranda, J. O., Arif, N. E., Hernandez, N. M., & Baillard, F. [2020]. A Geological Regional Case Study for Pressure, Temperature, and Salinity for the GoM using Machine Learning Technology on Unstructured Data. *AAPG Digital Subsurface for Asia Pacific Conference.* Kuala Lumpur, Malaysia.