2022 EAGE ANNUAL
83RD CONFERENCE & EXHIBITION
MADRID I SPAIN

UNLOCKING VALUE FROM UNSTRUCTURED DOCUMENTS USING MACHINE LEARNING: A GEOCHEMISTRY CASE STUDY, US GULF OF MEXICO

M. Fry[1], S. Badejo[1], J. Richardson[1], K. Petersen[2], H. Justwan[2]

[1] CGG; [2] Hess

Summary

Over two million files, containing geochemical information, have been collected from tens of thousands of wells drilled during decades of exploration in the Gulf of Mexico (GOM) and are available to geoscientists in the public domain. While these files represent a vast knowledgebase covering subsurface geology and petroleum systems, data extraction, systematic compilation and quality control was previously too cumbersome to harness the full power of the data to make basin wide correlations, uncover new trends and ultimately opportunities.

A novel machine learning approach was employed to automate data classification and extraction across three protraction areas for all public domain geochemistry and PVT documents to provide a single consistent database from un-tagged, legacy formats stored in entirely different subfolders. The resulting database provides the ability to rapidly screen and integrate data from multiple disciplines over a large scale, in terms of data volume as well as geospatial coverage. This in turn opens up petroleum systems analysis work to a wider user base by acting as a bridge between disciplines, such as reservoir engineering and geochemistry. Removing disciplines from silos is critical to enhancing collaboration between teams, improving efficiencies around specific workflows such as fluid property prediction and therefore reducing uncertainty.

Unlocking value from unstructured documents using machine learning: A geochemistry case study, US Gulf of Mexico

Introduction

With over **55,000** wells drilled in the US Gulf of Mexico (GOM), extracting, QC'ing and databasing vast amounts of unstructured post-well data in multiple formats is a significant undertaking. To work with such large data volumes in reasonable timeframes, it has been necessary to develop a data processing pipeline leveraging the application of Machine Learning techniques. In this study we will be looking specifically at geochemistry and reservoir engineering data, however, this data pipeline has broader applications across geological and engineering disciplines.

The data for the purpose of this study was sourced from the Bureau of Safety and Environmental Enforcement (BSEE) inventory which contains more than **15,000** geochemistry files, of various formats, across over **800** wells. Approximately **250** of those wells fall within the three protraction areas of interest in this study: Garden Banks, Green Canyon and Mississippi Canyon (Figure 1).

Challenges associated with locating and accessing specific data in such a dataset include:

- Large numbers of both partially and fully duplicated files,
- A limited degree of pre-existing file classification,
- Important contextual information relating to data in one file can be held within another file meaning that any solution needs be holistic in terms of evaluating all files.

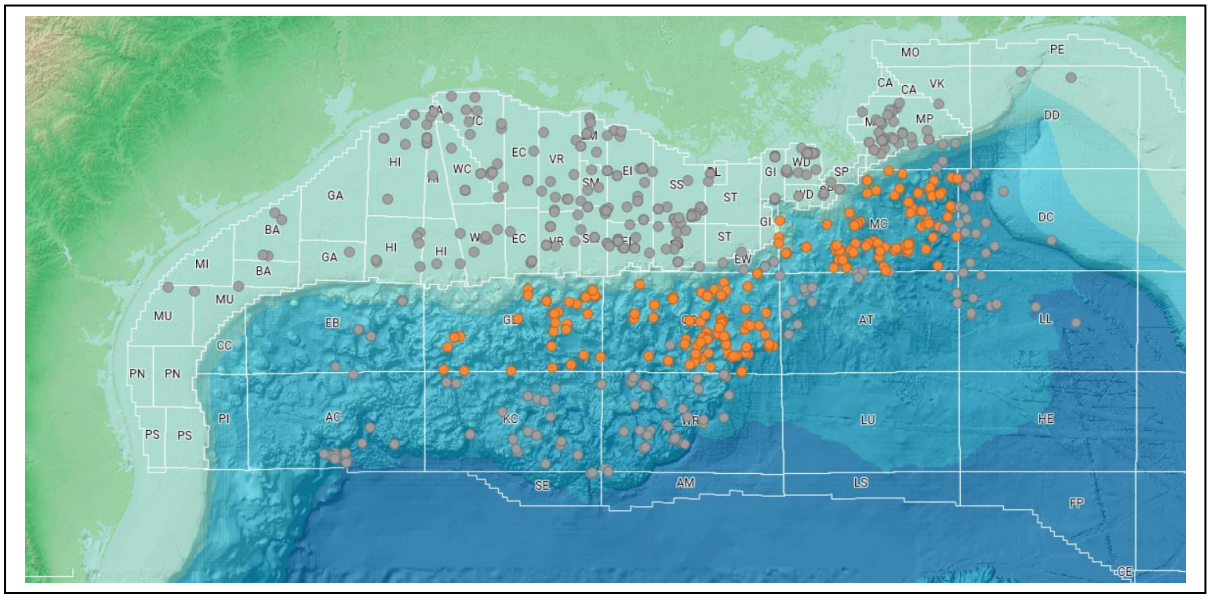

Figure 1 *In excess of 800 wells in the US GOM containing geochemistry data, with wells within the three protraction areas of Garden Banks, Green Canyon and Mississippi Canyon coloured orange.*

Method

An overview of the data classification and extraction pipeline employed in this study is shown in Figure 2. The process begins with classification, utilizing document segmentation, topic modelling and image classification to build a highly granular picture of the data types and their distribution within each document. Machine learning approaches are seeded by an extensive taxonomy of subsurface terms containing hierarchical relationships of data types as well as parent/child and master/alias relationships to capture variabilities in terminology across vintages, regions and/or companies. The results are stored in a knowledge graph representing the connections of data within and between documents.

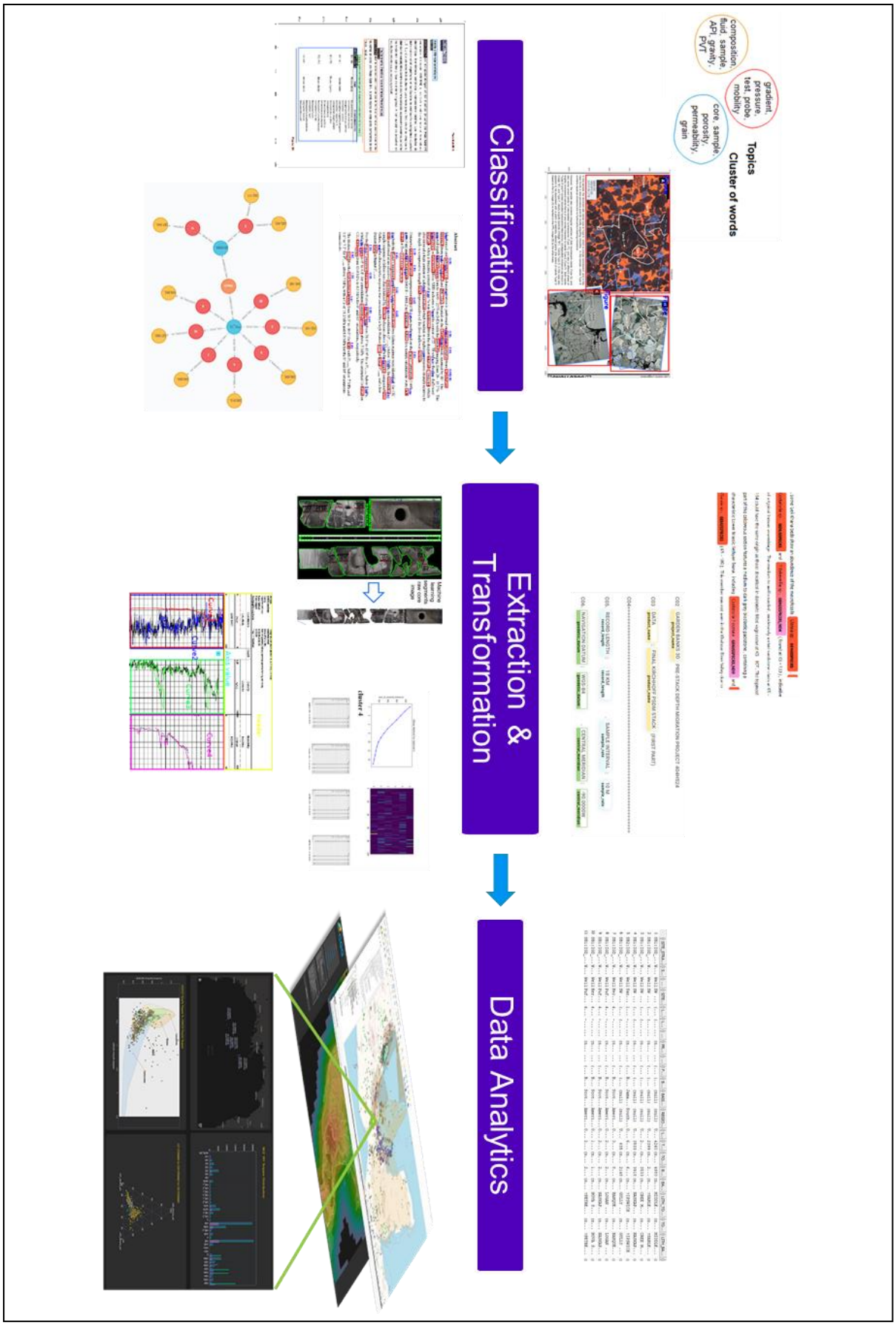

Figure 2 *Summary of the data pipeline, taking documents from a very granular classification through to extraction and transformation and data visualisation and analytics.*

In addition to the classification results, the graph also highlights the relationships between files which contain duplicated data. Exact duplicates are detected using a file hash approach to ensure that only unique files are carried forwards for processing. Partial duplication can be identified using a vector similarity approach based on file content, though this provides a more difficult challenge as it is not easy to automatically define which version of a file is more complete from a geoscience perspective. Partial duplicates are flagged with a similarity score and the final decision on which data to progress is passed to a geoscientist to review. This detailed classification approach improves the initial low detail file classification, converting the inventory to a tool that is useful for end users by improving the efficiency and means of exploring the data landscape in detail.

The whole data processing pipeline is guided by Subject Matter Experts (SMEs) to ensure that all data and contextual information is extracted and standardized, such that all intersections are reliable, and any analytics have the necessary metadata to be meaningful. This data pipeline facilitates processing of a large number of diverse files, facilitating construction of a robust multi-disciplinary database, enabling detailed analytics and reliable interpretations that would not otherwise be possible.

Application

Application of the data pipeline allowed extraction and databasing of **43,000** rows of geochemical data and **13,000** rows of reservoir engineering data from the **250** wells in the selected 3 protraction areas. Data was standardized to a taxonomy with multiple intersections possible across the dataset due to various available meta-data. The processing of the data took only a series of months to complete. The resulting dataset was then available for appropriate data analytics including geochemical inversion.

Geochemical inversion is the practice of using hydrocarbon chemistry to correlate recovered samples back to their likely source, which is particularly useful where only few well penetrations of the primary source interval exist, such as within the US GOM. Employing well established geochemical methods such as geochemical inversion on a consistent dataset, in conjunction with data visualization tools, makes it possible to rapidly screen fluids data dynamically across the whole basin. This enables the determination, for instance, of representative biomarker fingerprints for likely depositional environment and therefore organofacies (Figure 3). Possible applications of this include:

1) Creating source facies or oil family maps in real time,
2) Evaluating overarching trends and processes by tapping into a basin-wide dataset,
3) Determining relationships between source and maturity, alteration parameters and bulk properties. Thereby creating fluid analogue tools for phases of exploration (pre-drill) to high-grade opportunities and provide input for pre-development engineering studies.

Further steps can be taken by integrating bulk parameters such as GOR, API, Viscosity and Sulphur content to support these biomarker profiles through application of, for instance, clustering techniques to group samples with similar biomarker fingerprints (Justwan et al., 2019). The empirical relationship between these biomarker fingerprints and bulk parameters may provide a powerful fluid analogue tool to predict fluid properties from minimal geochemical information.

Figure 3 *Data analytics on the saturate GC-MS dataset in the US GOM, including relative abundance diagrams, tricylic cross plots and Sterane ternary diagrams.*

Conclusions

Using the latest data science techniques, it was possible to create a structured **250** well database that comprised in excess of **43,000** rows of geochemical and **13,000** rows of reservoir engineering data. The creation of which, at such data volumes, represented a step increase in efficiency given the monthlong time frames involved. A reduced amount of geoscience resource and manual techniques were required in data identification and extraction, thus freeing up subject matter expertise to focus on data quality and interpretation. The resulting database enabled geological investigation, such as oil to source correlation and fluid property prediction, to be addressed robustly. With such a large resulting dataset it is therefore important to continue to utilize the latest technology in conducting further geological interpretation. In this case, the high data row counts are such that individual well interrogation is insufficient in identifying trends in reasonable timeframes.

Data-driven analytical interpretation requires a clean and standardized dataset to be conducted efficiently and benefits significantly from larger datasets. This enables rapid interrogation and information sharing within project teams, enhancing collaboration, and improving the understanding of the petroleum system.

References

Justwan, H.K., Guthrie, J.M., Petersen, K., and Schmidt, D.P. [2019] The Application of Tricyclic and Tetracyclic Biomarker profiles to Infer the Organofacies of Source Rocks and Oils Using a Global Geochemistry Database. *Search and Discovery*, Article #42460.