# DI-document analytics

Digital intelligence (**DI**) for automating document control and classification.

INTRODUCTION

Whilst industries strive towards a fully data centric way of working the reality is that documents, as opposed to data is the way most information is still stored and accessed, usually from an Electronic Document Management System (EDMS) with content that has evolved over many years and where it is classified according to meta-data attributes defining, among other things, the document type such as a P&ID.

At DIGATEX we have developed **machine learning** capability to achieve a quantum leap in drawing and document control, not only **saving a significant amount of time and money** but also **vastly improving the accuracy and availability of information**. One important aspect of automating document control is to audit and remediate the document classification and meta-data, which we cover in detail below.
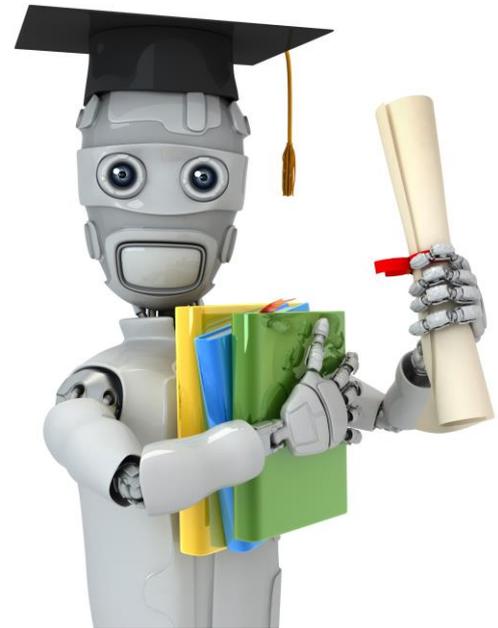
THE CHALLENGE OF DOCUMENT CLASSIFICATION

Many millions of drawings and documents exist and continue to be created for industrial assets, all of which need to be managed. Many of these are paper, dumb scanned documents or PDF's as well as native files. These documents need to be properly digitised, identified, classified, distributed and stored in such a way that they can easily and quickly be retrieved when required. Most of our clients have complained that their staff struggle to find the information they need and from conducting root cause analysis exercises we have identified some common causes:

- Meta-Data attributes do not exist, have not been filled in, or at least not correctly, causing searches to fail.
- Incorrect or NO classification of documents, e.g. a drawing classified as a PFD when it should be a P&ID, so when a user searches for all P&IDs the search engine does not return any of the P&IDs incorrectly classified as PFDs
- Drawings and/or documents filed incorrectly in the wrong folders or wrong projects even.
- Drawings and/or documents stored on local PCs (because users cannot find them in the document management system)
- Paper drawings as well as electronic, multiple versions and copies – document anarchy!

These problems are common to both greenfield and brownfield situations.

THE ANSWER

Machine learning has proven its capability in many fields and document control, which is very much a rules-based discipline, lends itself particularly well to a machine learning approach. At DIGATEX we have developed domain specific rule based mining models, configured to identify key aspects of the target drawings/documents specific to each client and/or asset or business line grouping. We take a small set of correctly classified documents/drawings and teach our system to identify these classes based on the mining models. We then process the documents/drawings in batches and classify them according to these rules, and at the same time we index the tag identifiers providing tag to document relationships, which are a key ingredient for a digital asset.

Each data mining model will predict the classification. The output is then aggregated via a voting process and the classification issued with a level of certainty based on the number of positive identifications by the data mining models.

From experience, we expect to correctly classify greater than 75% of documents after the first teaching rounds. With more teaching, the number classified will increase. The correct classifications and tag to document linkages are then provided back to the client. The remainder, depending on criticality, need some manual intervention.

PROCESS
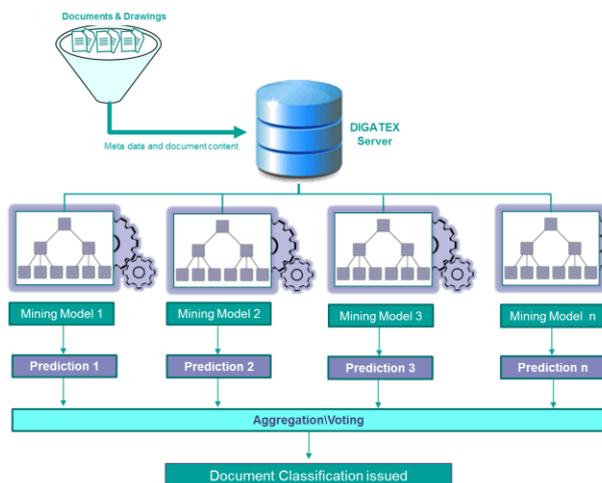
BENEFITS

- Significant reduction in cost of classifying drawings and documents
- Reduce the manual intervention to less than 25%
- Free up precious resources from mundane tasks
- Can be applied to new drawings and documents as well as existing portfolio, hence automating the document registration process and reducing human error

The work is carried out as and end to end process including scanning and OCR as required in our Information Engineering Centre of Excellence and is usually divided into a short assessment and scoping phase (typically around 10-15 days) followed by full-scale execution.

CONTACT US

If you would like to learn more about how we could help your business improve its document control and management activities with our **DI**-document analytics, all our contact information can be found on our website www.digatex.com