Caterpillar Inc. is the world’s leading manufacturer of construction and mining equipment, off-highway diesel and natural gas engines, industrial gas turbines, and diesel-electric locomotives. For nearly 100 years, we've been helping customers build a better, more sustainable world and are committed and contributing to a reduced-carbon future. Our innovative products and services, backed by our global dealer network, provide exceptional value that helps customers succeed.

**INTRODUCTION**

• **About Caterpillar & CAT Digital**
  - Caterpillar Inc.
  - Cat Digital is a digital and technology business unit within Caterpillar Inc.

  - o We’re focused on using data, technology, advanced analytics and AI capabilities to help our customers build a better world.
  - o Cat Digital embodies the bits-and-bytes coming from pieces of machinery, customers and dealers, and is charged with ensuring data flows smoothly and safely through our platform and to applications.

• **Motivation & Problem**
  - Currently, Caterpillar is using a clustering model based primarily on industry- and company-based rules to categorize invoices and parts. This strategy is effective for bootstrapping and is easy to understand and explain to their business partners. However, these rules are not necessarily backed by any scientific processes and trends/rules can change often. Hence, Caterpillar needs a more robust solution with proper validation to classify and categorize invoices from parts sales to customers.

• **Objective**
  - The goal of this project is to explore a novel document clustering algorithm to be used with invoices. The new algorithm should leverage a similarity metric built upon the numerical and categorical features associated with the line items listed on the invoice documents.

**RESEARCH METHODOLOGY**

• **Research Design Ideas**
  - Since we want to better cluster the invoices, we are trying to determine distance metrics for this type of data specifically. ⇒ The similarity metric should work with the "invoice" data type, namely a mix of numerical, categorical, and textual data. ⇒ "Similarity metric" feature that given two invoices it can determine how similar or dissimilar they are.

• **Key Approaches**
  - o Similarity metrics
  - o TF-IDF clustering
  - o Natural Language Processing

**FUTURE GOALS**

• **Include more advanced clustering ideas such as Gower Distance, and k-medoids, etc.**
• **Further, to improve upon the algorithms/model with a computationally efficient metric in a way that millions of invoices can be processed by computers compared to traditional document clustering algorithms.**

**KEY FINDINGS & SOLUTIONS**

• **Data Description & Data Attributes**
  - Column descriptions.
  - Sales_method (Business or internal purchase)
  - Sales_model (Sales model of the machine(s) being repaired).
  - Part_Pricing_Code (Category of parts used to determine its price)
  - Part_name (Fine component where this part was installed)
  - Part_component_L2 (Coarse component where this part is installed)
  - Part_component_L3 (Coarse category of parts)
  - Sales_model (Intermediate category of parts)
  - Source_of_supply (Standardized generic description of the part)
  - Part_Description (How the parts were sold: OTC= over the counter, WO= work order with the repair performed)
  - Dealership (Primary key to identify the dealership where the part invoice was issued)
  - Sales_method (Sales and revenue)
  - Invoice_ID (Primary key to identify an invoice)
  - Bill_to (Customer/individual)
  - Ship_to (Customer/individual)
  - Machine_Component_L1 (Fine category of parts)
  - Machine_Component_L2 (Intermediate category of parts)
  - Machine_Component_L3 (Coarse category of parts)
  - Commercial_Group_Type (Coarse category of parts)
  - Part_Pricing_Code (Category of parts used to determine its price)
  - Sales_method (Sales model of the machine(s) being repaired).

• **Data Cleaning Principles**
  - o Reduce the noise from negative price and quantity
  - o Clarify electric machine component numbers
  - o Improve null values
  - o Delete non-Caterpillar parts in the dataset
  - o Create dummy variables

• **Data Construction**
  - o Group by invoice_ids
  - o Group by major_class
  - o Carry out stratified sampling

• **Data Format**
  - o Create Sales_method as dummy variables

• **Models**
  - o DBSCAN model (Figure 2)
  - o K-Means with 3 clusters (Figure 3)
  - o Hierarchical clustering (Figure 4)

• **Improvement of clustering algorithm**
  - o Cosine similarity is more useful than Euclidean distance.
  - o Principal Component Analysis did not seem to improve the models much in reducing the number of dimensions.
  - o To evaluate, we have been looking at the number of data points in each cluster to see how evenly they are spread and also looking at inter-cluster distance.

• **Conclusions**
  - o Optimal number of clusters to be 62
  - o Chosen based off an elbow curve

• **Solutions**
  - o No significant correlations between attributes
  - o Most numerical attributes are hashed
  - o Others are categorical
  - o Sales_method & sales_method are two specific attributes of relevance for data mining and business purposes since they provide insights into how big/small the repair was and the method in which it was accomplished