EXTERN + OUTAMATION
EXTERNSHIP

ADVANCED AI-POWERED

DOCUMENT INSIGHT AND DATA EXTRACTION PROJECT

SPRINT 2 STATUS UPDATE

by Lashawn Fofung, MBA, PMP, CSM, CSPO

L F

SKILLS

- Python
- AI/ML
- Data Scrapping
- Prompt Engineering
- LLMs
- Text Extraction
- Document Intelligence
- Python Automation
- Optical Character Recognition (OCR)
- Retrieval-Augumented Generation (RAG)

Helping businesses improve, automate, and architect critical tasks (such as document handling, customer service, and data management) by saving thousands of hours and boosting accuracy.

Mortgage Documents

SPRINTS

2025

01 AI Document Intelligence
02 Python & Google Colab
03 Python Data Extraction
04 Optimizing OCR
05 Implementing RAG
06 Optimizing RAG Pipelines
07 Blob Processing & Classification
08 Interactive Chatbot
09 AI-Powered Document Automation Platform

PROJECT TIMELINE

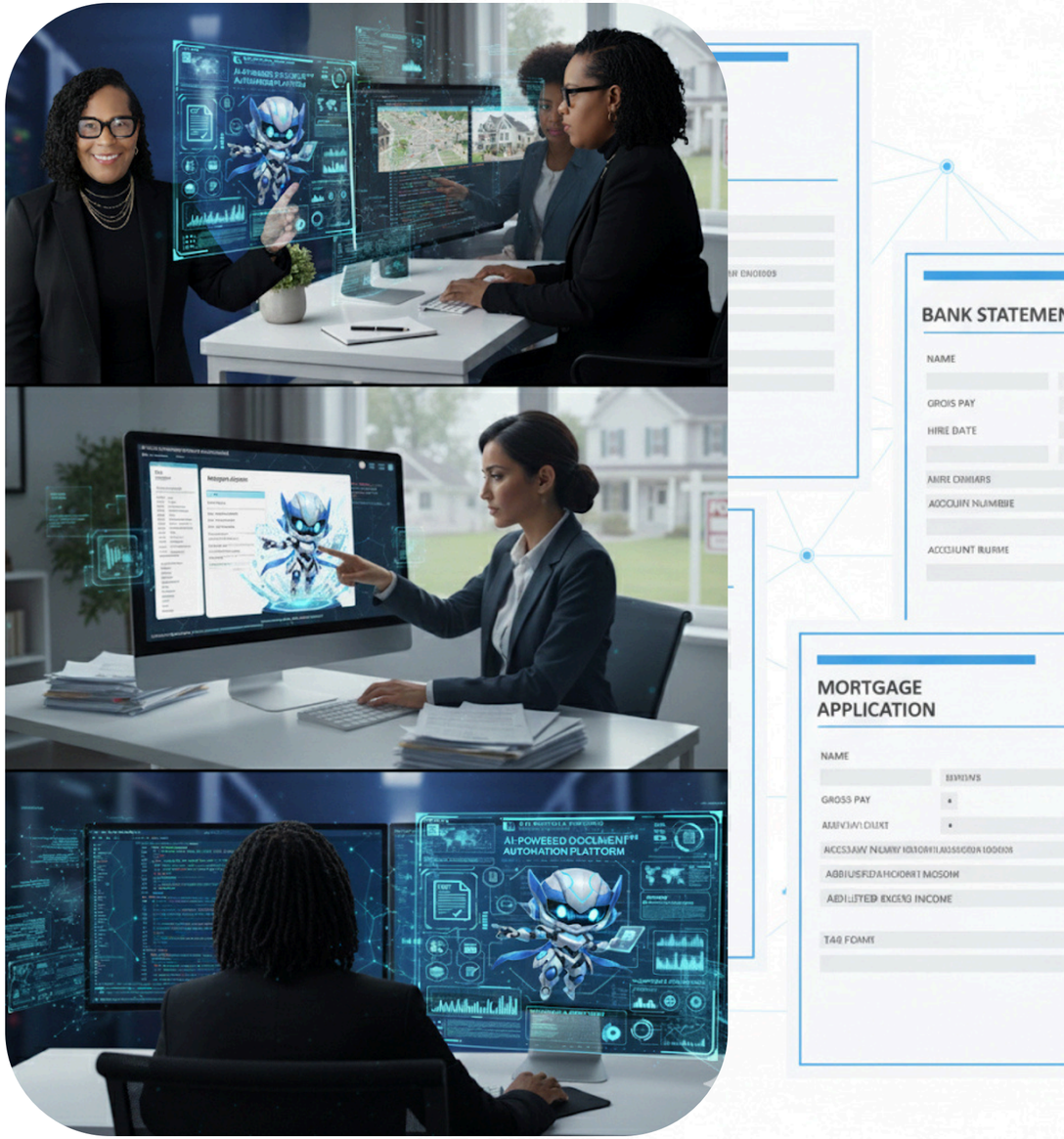
10 WEEKS

Python & Google Colab

Preparing Mortgage Data for AI

Background: 10-Week Project

The Outamation Advanced AI-Powered Document Insights and Data Extraction Externship is a comprehensive 10-week program. Its core goal and purpose are to master the required skills to create an AI-Powered Document Automation Platform. The program focuses on bridging foundational technologies (ML, LLMs, NLP, Computer Vision) with high-value industry applications, such as solving complex, document-heavy challenges in mortgage automation, ultimately transforming slow, manual review processes into automated, efficient workflows.



Challenge & Solution: Building The Document Intelligence (AI) Pipeline

- **The Business Challenge:**
  - Automating the highly manual and time-consuming process of mortgage underwriting by handling 200-300 pages of unstructured documents.

- **The Solution:**
  - Designing and implementing a robust, AI-Powered Document Automation Platform to transform raw files into structured, intelligent insights.

- **The 4-Step Document Intelligence Workflow:**
  - Step 1. Document Acquisition & Pre-Processing (The Input)
  - Step 2. Data Extraction & Preparation
  - Step 3. Knowledge Base & Modeling (The AI Engine)
  - Step 4. Output, Review, & Integration (The Results)

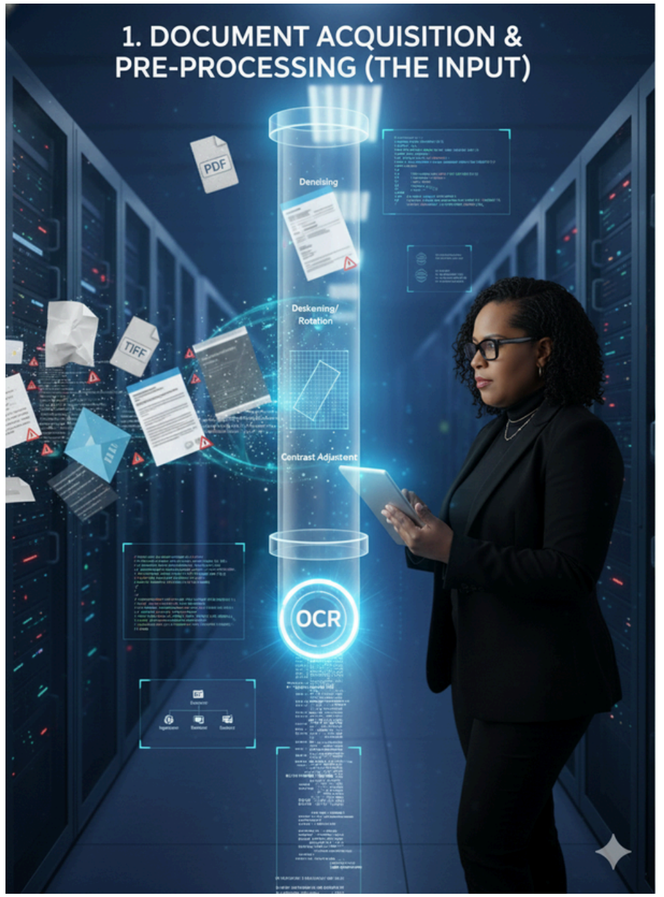
These foundational skills, mastered in Sprint 2, directly enable the 4-Steps Document Intelligence (AI) Pipeline designed to solve our core business challenge.

SPRINT 2 FOCUS

Step 1: The Input

Step 1: Document Acquisiton & Pre-Processing

- Ingest raw, messy files (PDF, TIF, JPG).
- Apply computer vision (Denoising, Deskewing) to enhance document quality.
- Use OCR to extract text and location (coordinates), turning unstructured input into structured data.



Overall Status: Sprint 2 Complete

I've successfully concluded Sprint 2, focused on establishing the foundational Python and Google Colab skills essential for building the AI-Powered Document Automation Platform. The primary goal was to learn the practical coding environment and core data handling techniques necessary to transform complex mortgage data for AI use.

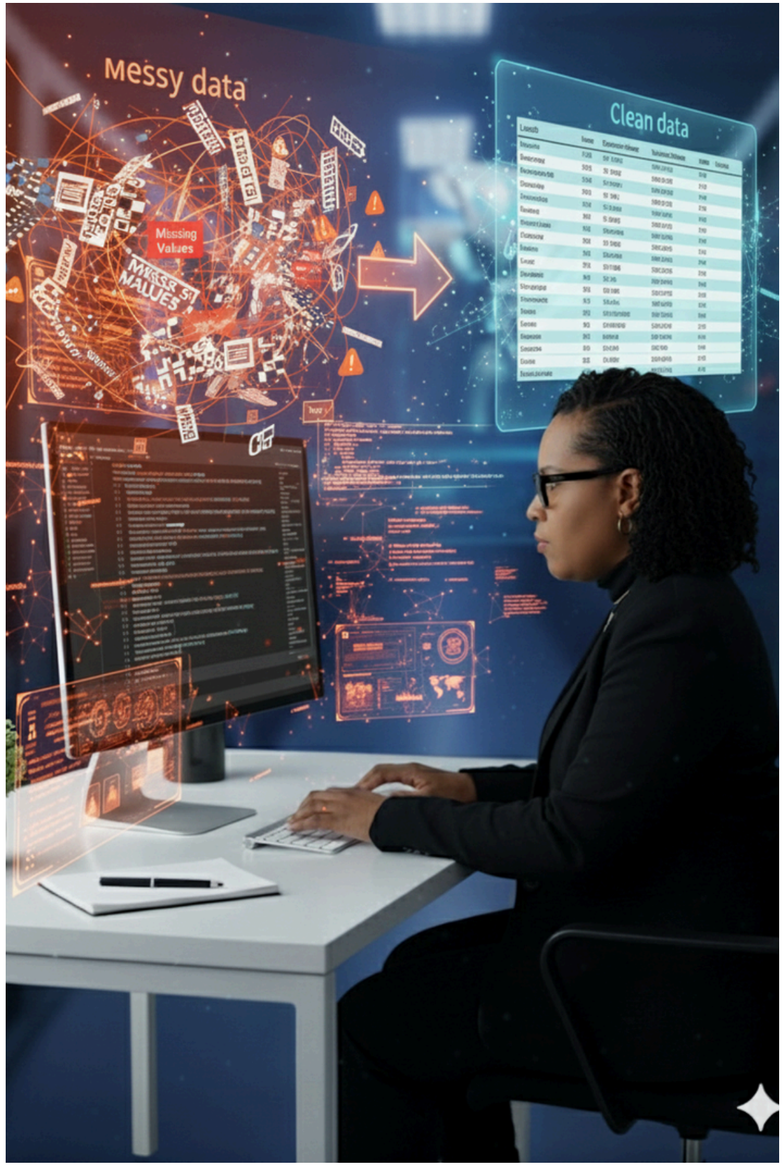
My progress seamlessly moved from learning Python fundamentals to applying these skills across two crucial areas:

- Data Preparation & Standardization (AI Foundation)
- Image Pre-Processing for OCR Accuracy

Key Accomplishments - Data Preparation & Standardization (AI Foundation)

This step was central to preparing inherently messy mortgage data for reliable AI consumption.

- Mastered Data Standardization:
  - Gained proficiency in core data preparation concepts using libraries like Pandas to clean and validate structured data from diverse source materials (CSV, JSON, TXT).
- Complex JSON Parsing:
  - Gained critical experience in handling hierarchical data.
  - This involved writing code to process and flatten complex nested JSON structures—common outputs from commercial OCR services—into simple, single-level tables ready for AI model consumption.
- Enhanced Text Cleaning (NLP Foundation):
  - Learned how to clean raw text extracted from documents.
  - Applied preprocessing techniques (standardizing capitalization, removing noise/special characters, fixing inconsistencies) to legal and specialized mortgage text.
  - This drastically improves the accuracy of downstream Natural Language Processing (NLP) models.



Addressing Data Messiness

The learning centered on transforming raw, unstructured, and inconsistent mortgage data into a reliable format for the AI. The types of data issues addressed included:

- Inconsistent Formatting: Non-standard dates, currencies, and addresses extracted from documents.
- Missing Values: Identifying and handling (imputing or dropping) gaps in data fields before analysis.
- Irregular Text: Correcting typographical errors, line breaks, or non-standard abbreviations resulting from document scanning and OCR extraction.

Impact

- This work is crucial because the integrity of the downstream AI model, which generates accurate insights for mortgage automation, relies entirely on having error-free and well-structured input data.
- The raw input data is now validated and appropriately structured.

Key Accomplishments - Image Pre-Processing For OCR Accuracy

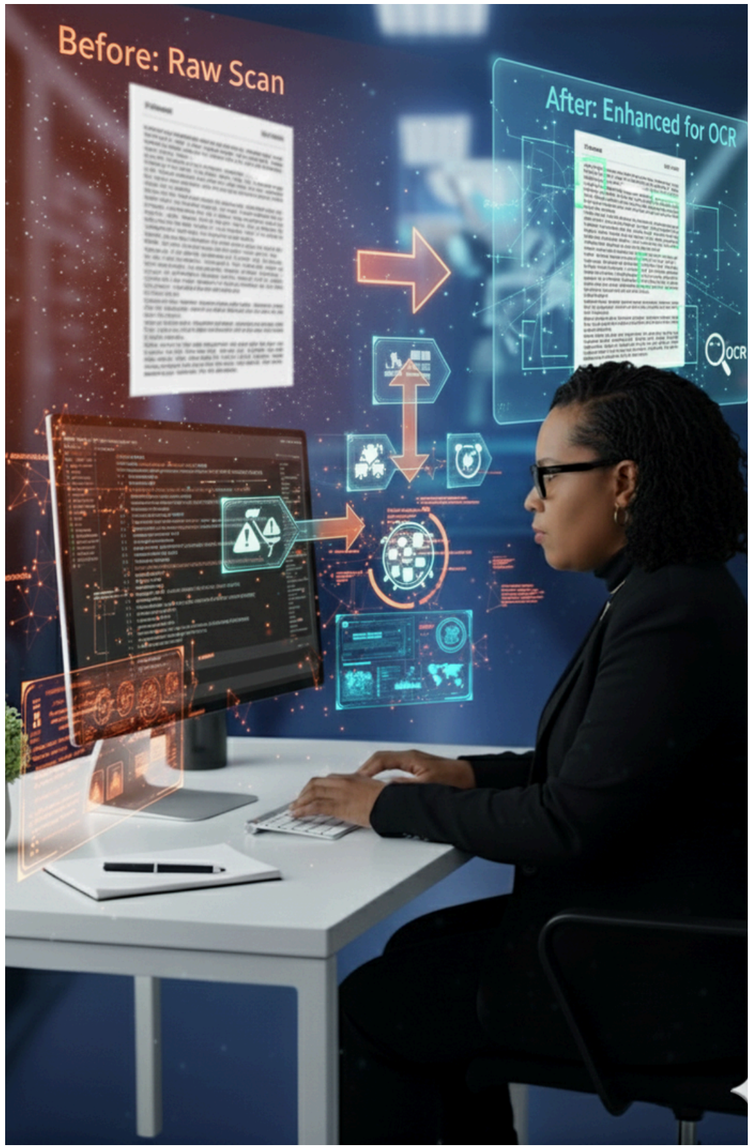
This section details the work focused on ensuring high-quality input for the automation pipeline by handling scanned and image-based components of mortgage files.

- Core Skill:
  - Gained proficiency in image processing techniques using libraries like OpenCV and PIL.
- Mortgage Application:
  - Applied various techniques to prepare complex inputs—such as poor-quality scans, or blurred/dark JPG, PDF, or TIFF document pages—for optimal extraction.

Computer Vision Techniques for Enhancement

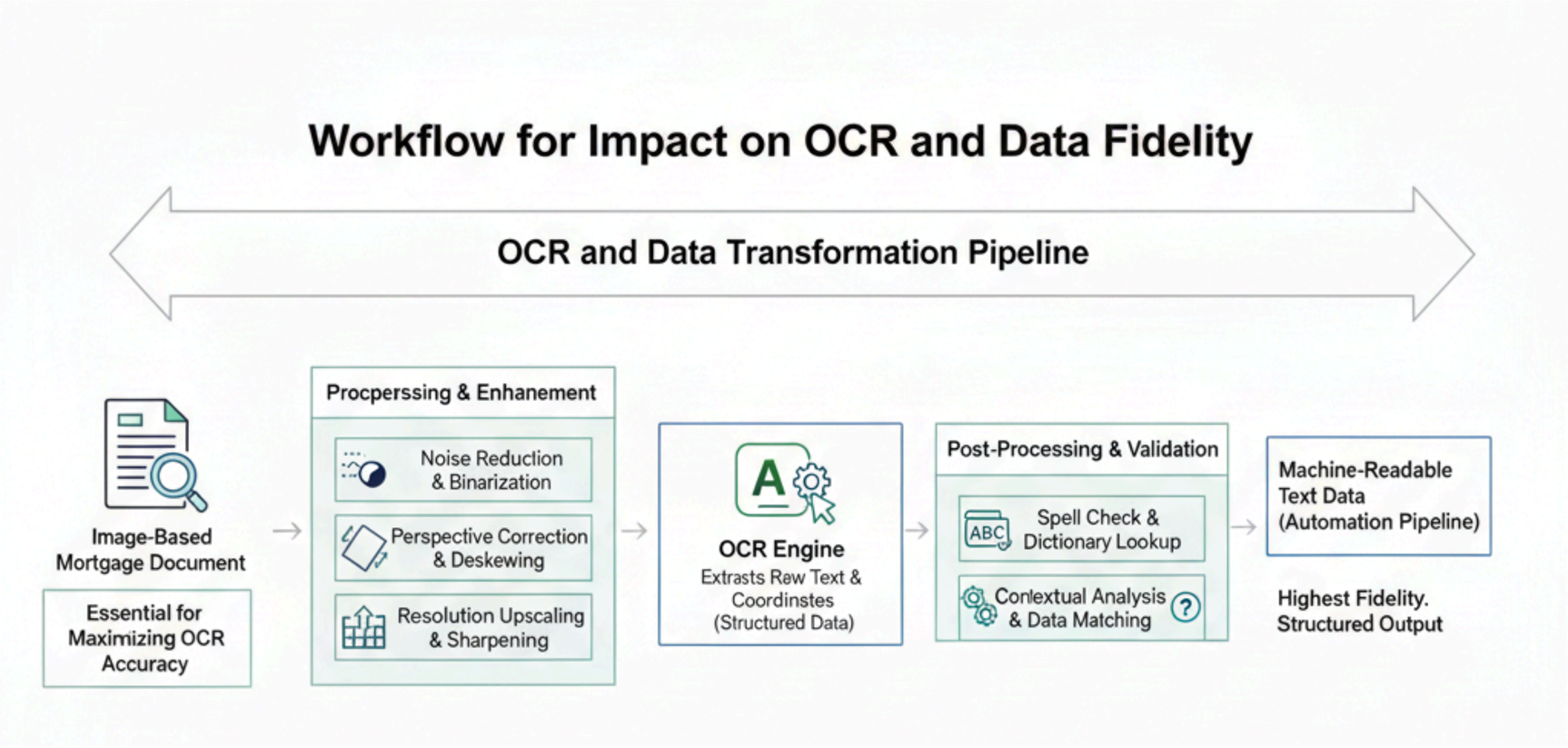
To enhance the quality of poor-quality scans or images for optimal text extraction, the following computer vision techniques were used or studied:

- Denoising: Removing digital "noise" or speckles.
- Deskewing/Rotation: Correcting document alignment.
- Contrast Adjustment/Thresholding: Making the text stand out for better extraction.



Impact on OCR and Data

- These skills are essential for maximizing Optical Character Recognition (OCR) accuracy.
- A clean image is run through an OCR engine to extract all raw text and, critically, the location (coordinates) of that text, which transforms the unstructured image into structured text data.
- This ensures the highest fidelity when turning an image-based mortgage document into machine-readable text for the automation pipeline.



Key Metrics & Progress

This sprint resulted in measurable progress in process maturity, establishing the data foundation, and technical skills necessary for the AI-Powered Document Automation Platform.

- **Process & Data Foundation**
  - Workflow Completion:
    - Python & CoLab Basics (Tooling)
    - Data Preparation (Cleaning & Structuring Data)
  - Data Readiness:
    - Raw, inconsistent mortgage data is now validated, cleaned, and structured.
  - Status:
    - Confirmed Ready to transition to the next phase: Python Data Extraction & AI Model Building.

Key Metrics & Progress

- **Technical Proficiency & Skills**
  - Python Foundation:
    - Mastered Python fundamentals (data structures, functions) and established the Google Colab development environment.
  - Data Cleaning:
    - Achieved proficiency in Pandas for loading, inspecting, and standardizing structured mortgage data.
  - Complex Parsing:
    - Gained the skill to process and flatten complex nested JSON output from OCR APIs.
  - Image Preparation:
    - Acquired knowledge of OpenCV/PIL principles for image enhancement (denoising, contrast) to maximize OCR accuracy.
  - Impact:
    - These skills position the project to help implement solutions that can shift document processing time from weeks to minutes.

THANK YOU

https://github.com/LashawnFofung/AI-Powered-Document-Automation-Platform