# Discovery and Classification of Images and Video content

**Mentor Name** : Arvind Ashtekar ( arvind.ashtekar@in.ibm.com )

**Project Members :**
Arya S. Deshmukh (arya.deshmukh@cumminscollege.in)
Anusha Karwa (anusha.karwa@cumminscollege.in)
Shreya Mote (shreya.mote@cumminscollege.in)
Asavari Thorat (asavari.thorat@cumminscollege.in)

**Objective:** A program that can discover images and video files in a given data storage device/filesystem and help classify the discovered content in various classification categories- PII, Sensitive, Confidential etc. Possibly use ML to improve the accuracy of classification.

## What can we do, if from scratch

OCR (Optical Character Recognition): Use an OCR tool (like Tesseract, AWS Textract, or Google Vision) to extract any textual information from images.

## Paid Platforms available

### 1. Amazon Web Services (AWS):

- **Amazon Macie**:
  - A machine learning-powered security service that discovers, classifies, and protects sensitive data, such as PII, stored in AWS S3.
  - Automatically detects PII in data (names, emails, credit card numbers, etc.).
- **Amazon Textract**:
  - Extracts text and structured data from scanned documents (images or PDFs), allowing you to process and detect PII.
- **AWS Comprehend**:
  - A natural language processing (NLP) service that uses machine learning to find insights in text, including detecting PII entities in the text.
- **AWS Lambda**:
  - You can use this serverless computing service to run custom logic for further processing and classifying PII data after it's extracted.

## 2. Google Cloud Platform (GCP):

- **Cloud Data Loss Prevention (DLP)**:
    - Scans, classifies, and redacts sensitive data (PII) across a variety of data sources (e.g., GCS, Cloud SQL, BigQuery).
    - Detects sensitive information such as names, credit card numbers, or national IDs.
- **Vision AI**:
    - Detects objects and text within images, and you can combine it with Google DLP to detect sensitive information.
- **Cloud AutoML Vision**:
    - Build custom image classification models (you can use it to classify images that might contain PII data).
- **Document AI**:
    - Extracts structured information from images or PDFs, making it easier to identify PII.

## 3. Microsoft Azure:

- **Azure Cognitive Services - Text Analytics**:
    - Extracts key phrases, entities (including PII), and more from text within documents or images.
- **Azure Form Recognizer**:
    - Automates the extraction of text, key-value pairs, and tables from documents, and can be used to identify sensitive information.
- **Azure Data Lake Analytics with Data Classification**:
    - Detects and classifies sensitive data in large-scale datasets.
- **Azure Computer Vision**:
    - Detects text and objects within images, and can be combined with other services to classify PII data.

## 4. IBM Cloud:

- **IBM Watson Natural Language Understanding**:
    - Can be used to extract and classify PII from text data.
- **IBM Watson Visual Recognition**:
    - Detects objects and text within images, and can be customized to detect specific categories, including sensitive data like PII.

## Platforms available for free

### 1. Tesseract OCR (Text Extraction from Images):

- **What it is**: Tesseract is an open-source OCR engine maintained by Google. It extracts text from images or PDFs.
- **Usage**: You can use Tesseract to extract text from documents or images locally, and then analyze that text for PII.
- **License**: Apache License 2.0 (Free and open-source).
- **Language Support**: Multiple languages, including custom training for specific cases.
- **Link**: Tesseract GitHub

### 2. Presidio (Microsoft Open-Source Tool for PII Detection):

- **What it is**: Presidio is an open-source tool developed by Microsoft for detecting and anonymizing PII in text and structured data. While it's from Microsoft, it's fully free and can be run locally.
- **Usage**: It detects PII like names, emails, phone numbers, and more within text data.
- **License**: MIT License (Free and open-source).
- **Integrations**: You can combine it with tools like Tesseract to process images with text and run PII detection.
- **Link**: Presidio GitHub

### 3. DLPy (Data Loss Prevention with Python):

- **What it is**: DLPy is an open-source library in Python that allows you to create custom PII detection models. You can train models to detect sensitive information from text and images.
- **Usage**: It provides out-of-the-box deep learning tools for detecting text in images and customizing your detection of PII.
- **License**: Open-source (MIT License).
- **Link**: DLPy GitHub

### 4. OpenCV (Image Processing and Classification):

- **What it is**: OpenCV is an open-source computer vision library that can be used for image classification, text detection, and object recognition.
- **Usage**: With OpenCV, you can perform image processing tasks, such as detecting documents, faces, or specific objects within images. Combine it with an OCR tool for PII detection.
- **License**: BSD License (Free and open-source).
- **Link**: OpenCV GitHub

## 5. Fawkes (Privacy Protection for Faces in Images):

- **What it is**: Fawkes is an open-source tool designed to protect the privacy of individuals by obfuscating facial recognition models, making it harder for AI systems to identify people in photos.
- **Usage**: You can use Fawkes to anonymize faces in images, a useful tool for GDPR compliance when handling image data containing PII.
- **License**: MIT License (Free and open-source).
- **Link**: Fawkes GitHub

## 6. spaCy (Entity Recognition and PII Detection):

- **What it is**: spaCy is an open-source natural language processing (NLP) library in Python. It includes powerful tools for detecting entities in text, including PII like names, addresses, and phone numbers.
- **Usage**: spaCy can be combined with OCR tools to extract text from images and detect PII.
- **License**: MIT License (Free and open-source).
- **Link**: spaCy GitHub

## 7. Label Studio (Data Annotation for PII Detection):

- **What it is**: Label Studio is an open-source data labeling tool that supports annotating images, text, and more. You can use it to label data for training PII detection models locally.
- **Usage**: Train machine learning models to detect sensitive data (PII) in images.
- **License**: Apache License 2.0 (Free and open-source).
- **Link**: Label Studio GitHub

## 8. Piiano Vault (Self-Hosted PII Management):

- **What it is**: Piiano Vault is a self-hosted solution that helps manage sensitive data and PII securely. It provides a fully local approach to storing and processing PII, helping with GDPR compliance.
- **Usage**: It's mainly used for structured PII data storage, but could be integrated with OCR and image tools for a complete pipeline.
- **License**: Free to use with enterprise features available.
- **Link**: Piiano Vault

### 9. Octopii :

- **What it is**:Octopii is a Personally Identifiable Information (PII) scanner that uses Optical Character Recognition (OCR), regular expression lists and Natural Language Processing (NLP) to search public-facing locations for Government ID, addresses, emails etc in images, PDFs and documents.
- **License**: MIT, opensource
- **Link**: https://www.researchgate.net/publication/381021821_Detection_and_Classification_of_Personally_Identifiable_Information_in_Images_Using_Artificial_Intelligence

### Combining Tools for a Workflow:

1. **Text Extraction from Images**: Use Tesseract OCR or OpenCV for extracting text from images.
2. **PII Detection in Text**: Run the extracted text through Presidio or spaCy to detect PII entities.
3. **Custom Image Classification**: Use OpenCV for image analysis or train a model using DLPy or Label Studio to identify images containing sensitive content.
4. **Face Redaction/Anonymization**: Use Fawkes to blur or obfuscate faces in images for privacy.

## Open Source Availability

1. https://github.com/microsoft/presidio: presidio
2. https://github.com/google/magritte: google magritte
3. https://github.com/redhuntlabs/Octopii?tab=readme-ov-file: octopii
4. https://redhuntlabs.com/blog/octopii-an-opensource-pii-scanner-for-images/ : octopii demo