* When an image is uploaded, we want the system to locate the paper/poster/flyer in the image, detect text present in it, and extract it.
* Once this text has been extracted, we’d like to map attributes to its data, based on the necessity.
  + Receipts/invoices would require us to extract vendor name, total amount, tax applied, and date of the receipt.
  + IDs would require us to extract ID type, name, guardian/parent name, ID number, and other attributes such as date of expiry and address
  + Pictures of books would require us to extract Title, Author and Publisher
  + Nutritional labels would require us to map each nutritional type to its value
* The process to proceed with this would be

| * Image acquisition | * Image acquisition |
| --- | --- |
| * Image Pre-processing | * Image Pre-processing |
| * Segmentation | * Text Detection |
| * Feature Extraction | * Text Extraction (OCR) |
| * Classification | * NLP Information Extraction |
| * Post-processing |  |

[*https://sci-hub.se/10.1109/ICIRCA48905.2020.9183326*](https://sci-hub.se/10.1109/ICIRCA48905.2020.9183326) *,* [*https://stacks.stanford.edu/file/druid:cg133bt2261/Zhang\_Zhang\_Li\_Event\_info\_extraction\_from\_mobile\_camera\_images.pdf*](https://stacks.stanford.edu/file/druid:cg133bt2261/Zhang_Zhang_Li_Event_info_extraction_from_mobile_camera_images.pdf)

* There are a lot of highly optimized tools for text extraction.
  + Tesseract-OCR
* The accuracy of the text extracted is highly dependent on the quality of the image it is

Who’s already doing this?

* Amazon Textract
* Azure Form Recogniser
* Google Cloud Vision API
* Nanonets (limited)

Why do we want to do this then?

* Free/Open Source
* Even though there are existing free&open source application for text detection and extraction, there aren’t any for image preprocessing. And unless the image is of high quality (non digital documents) tesseract-ocr wouldn’t work to its highest efficiency.
* We want to explore image preprocessing methodologies and combine the best techniques to improve the accuracy of text extraction

**PPT**

The problem with historical OCR processes is that their accuracy lies between 70%-80% for a high-quality image. While that may seem high, it can cause significant inaccuracies if used on a large volume of documents or sensitive documents like invoices. Imagine losing 2% of the value on every invoice you generate because your OCR system is not accurate!