AUGUST 2018

Handwriting Recognition POC

Google Cloud Platform API Assessment

Executive Summary

- A proof of concept is needed to determine the feasibility of **using Google's Cloud API services for text detection and optical character recognition (OCR) applied to common claims forms**
- Google Cloud Platform (GCP) offers a number of pre-trained and available machine learning (ML) models, including the Cloud Vision API, capable of out of the box image labeling, logo detection, OCR, and text detection.
- The proof of concept will aim to **understand the capabilities and limitations of the Cloud Vision API**, and suggest a potential ML pipeline for automation of claim intake.
- The success criteria for the proof of concept will be defined as follows:
  - Determine which available APIs and models within GCP are applicable to automation of claims intake
  - Using determined APIs and GCP tools:
    - Identify various **fields within a given claims form**
    - Identify handwriting and **transcribe into a flat format** (.csv)

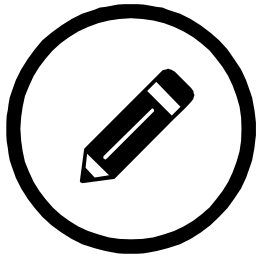
GCP Capabilities Exploration

Plan

Review API Documentation

Identify API pricing, documentation, and samples.

The underlying models from these selections will be used to drive the remainder to the exploration.

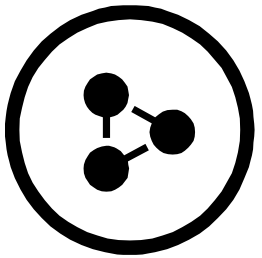


Design

Gather Training Data

Choose several forms of interest and create a training sample to test the output of the selected models.

Having a variety of forms will help understand the limitations of pre-built solutions by determining common areas of weakness.

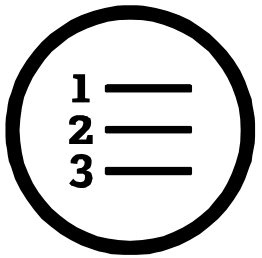


Build

Develop and Test Models

Develop a sample pipeline and test APIs across a variety of criteria including accuracy and compute time.

Outputting the results to a flat format allows for analysis and standardization across forms.



Evaluate

Review Results & Accuracy

Using the criteria outlined in the program objectives, determine whether to continue research into GCP APIs.

Socialize results and suggest areas for future work.

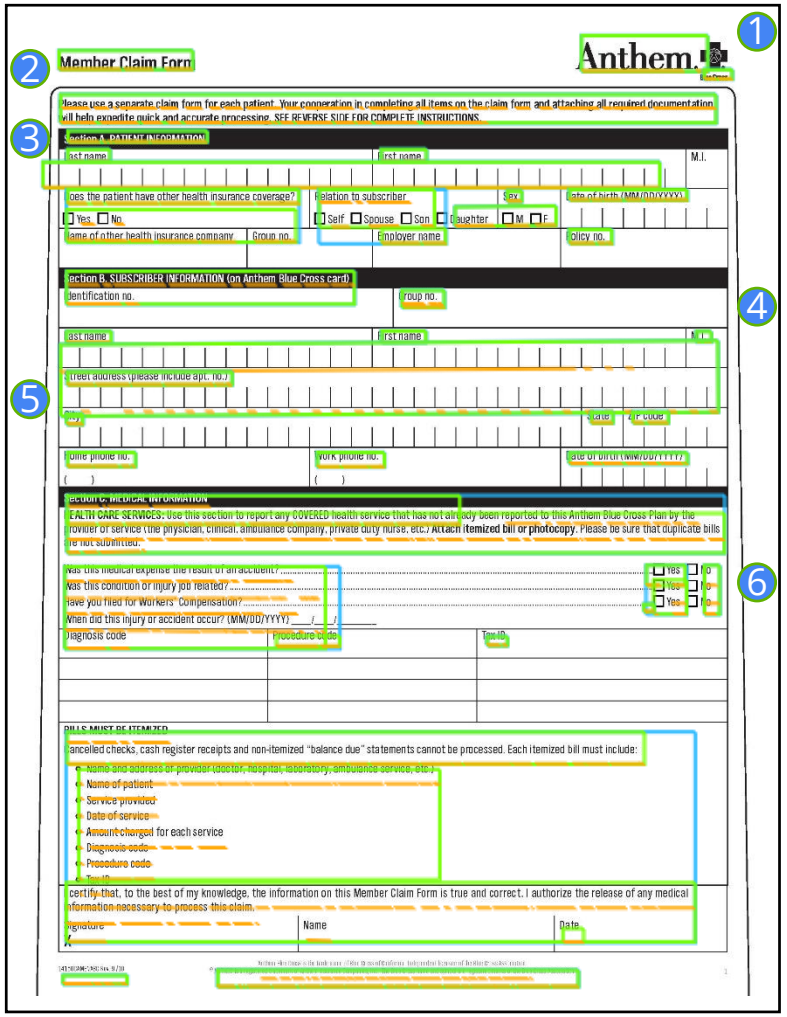
PLAN

Cloud Vision API Pricing

Applicable	Feature Type	Description	Price per 1000 Units	
			Units 1001 - 5,000,000 / month	*Units 5,000,001 - 20,000,000 / month*
✕	LABEL_DETECTION	Add labels based on image content (see Detecting Labels)	$1.50	1.00
✓	TEXT_DETECTION	Perform Optical Character Recognition (OCR) on text within the image (see Detecting Text)	$1.50	$0.60
✓	DOCUMENT_TEXT_DETECTION	Perform OCR on dense text images, such as documents (see Document Text Detection)	$1.50	$0.60
✕	SAFE_SEARCH_DETECTION	Determine image safe search properties on the image (see Detecting Safe Search Properties)	Free with Label Detection, or $1.50	Free with Label Detection, or $0.60
✕	FACE_DETECTION	Detect faces within the image (see Detecting Faces)	$1.50	$0.60
✕	LANDMARK_DETECTION	Detect geographic landmarks within the image (see Detecting Landmarks)	$1.50	$0.60
✓	LOGO_DETECTION	Detect company logos within the image (see Detecting Logos)	$1.50	$0.60
✕	IMAGE_PROPERTIES	Compute a set of properties about the image, such as the image's dominant colors (see Detecting Image Properties)	$1.50	$0.60
✕	WEB_DETECTION	Detect topical entities such as news, events, or celebrities within the image	$3.50	Contact Google for more information
✓	CROP_HINTS	Determine suggested vertices for a crop region on an image (see Detecting Crop Hints)	Free with Image Properties, or $1.50	Free with Image Properties, or $0.60
✕	OBJECT_LOCALIZATION	Detect and extract multiple objects in an image (see Detect Multiple Objects)	$2.25	$1.50

Required Capabilities

A successful extraction system needs to accurately transcribe handwritten content and associate it with the correct field



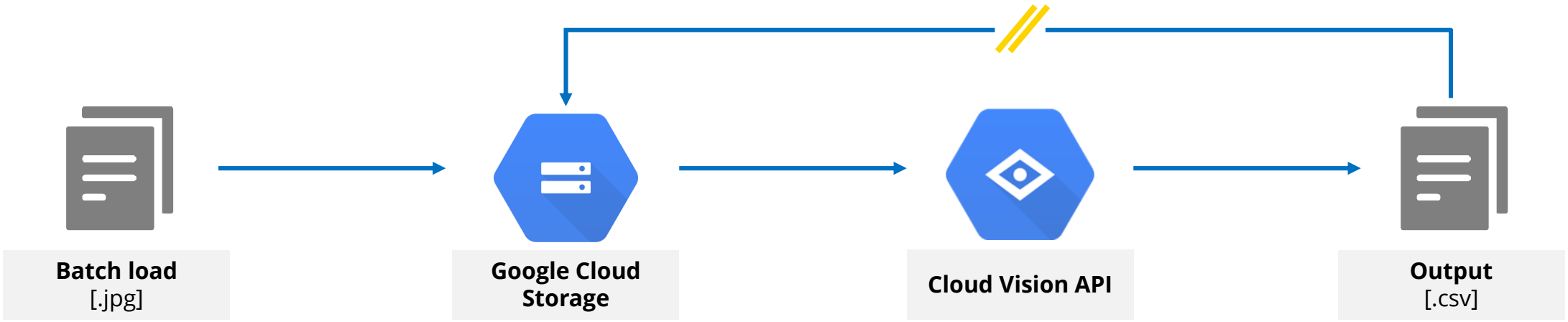
#	Capability	Vision API
1	Correct skewness, remove noise, and crop	✓[1]
2	Detect form template	✓
3	Identify relevant section content	✓
4	Transcribe handwritten text	✓
5	Standardize, validate, and extract content from data fields	X[2]
6	Identify checked boxes and other non-text symbols	✓

1 Crop Hints (Beta)
2 Data Validation feature not available in Vision API

BUILD

Proof of Concept Architecture

A basic workflow to test multiple claim types with the ability to scale using products like Google Dataflow and BigQuery



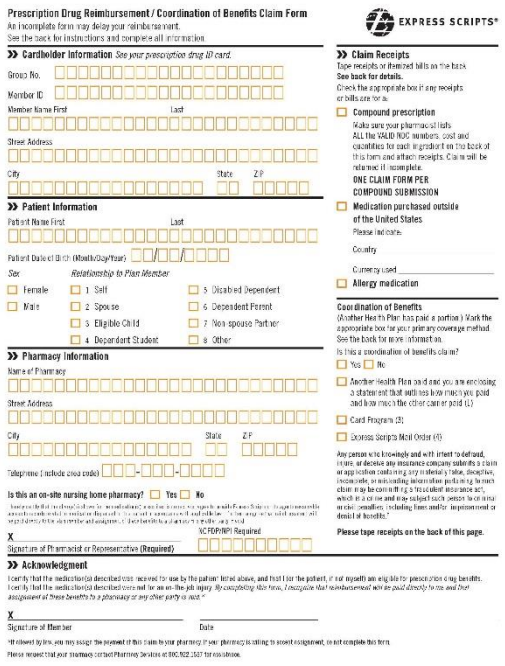
Represents future additions

- Dataflow for batch or streaming data transformations and processing
- BigQuery for inexpensive database storage with interactive analysis tools

Claim Forms

Multiple claim forms were compared to evaluate the performance of the Vision API across different document types

New York Prescription Form


Prescription Drug Reimbursement / Coordination of Benefits Claim Form

EXPRESS SCRIPTS®

» Cardholder Information See your prescription drug ID card.

Group No.

Member ID

Member Name First Last

Street Address

City State ZIP

» Patient Information

Patient Name First Last

Patient Date of Birth (Month/Day/Year)

Sex: Female, Male

Relationship to Plan Member: 1 Self, 2 Spouse, 3 Eligible Child, 4 Dependent Student, 5 Disabled Dependent, 6 Dependent Parent, 7 Non-spouse Partner, 8 Other

» Pharmacy Information

Name of Pharmacy

Street Address

City State ZIP

Telephone (include area code)

Is this an on-site nursing home pharmacy? Yes No

X

Signature of Pharmacist or Representative (Required)

NCPDP/NPI Required

» Acknowledgment

Signature of Member Date

» Claim Receipts

Tape receipts or itemized bills on the back. See back for details.

Check the appropriate box if any receipts or bills are for a:

Compound prescription

Make sure your pharmacist lists ALL the VALID NDC numbers, cost and quantities for each ingredient on the back of this form and attach receipts. Claim will be returned if incomplete.

ONE CLAIM FORM PER COMPOUND SUBMISSION

Medication purchased outside of the United States

Please list date:

Country

Currency used

Allergy medication

Coordination of Benefits

(Another Health Plan has paid a portion.) Mark the appropriate box for your primary coverage method. See the back for more information.

Is this a coordination of benefits claim? Yes No

Another Health Plan paid and you are enclosing a statement that outlines how much you paid and how much the other carrier paid (1)

Card Program (3)

Express Scripts Mail Order (4)

Any person who knowingly and with intent to defraud, insure, or deceive any insurance company submits a claim or application containing any materially false, deceptive, incomplete, or misleading information pertaining to such claim may be committing a fraudulent insurance act, which is a crime and may subject such person to criminal and civil penalties (including fines and/or imprisonment or denial of benefits.)

Please tape receipts on the back of this page.

» Acknowledgment

I certify that the medication(s) described does received for use by the patient listed above, and that I (or the patient, if not myself) am eligible for prescription drug benefits. I certify that the medication(s) described does not (or are not) for an on-the-job injury. By completing this form, I recognize that reimbursement will be paid directly to me and that assignment of these benefits to a pharmacy or any other party is void.*

X

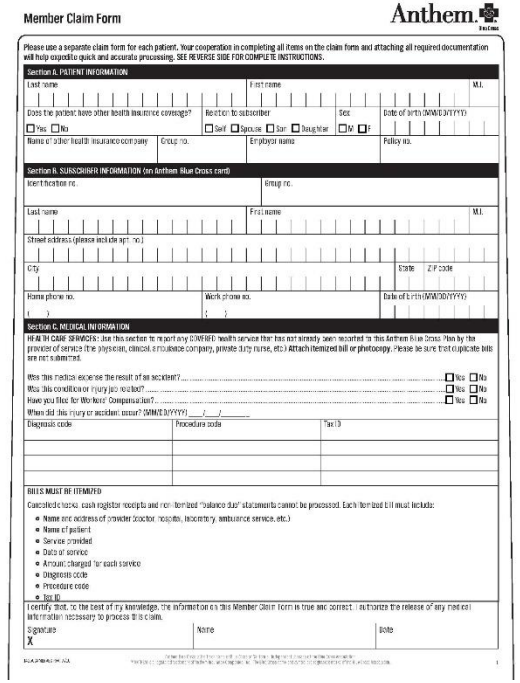
Signature of Member Date

*If allowed by law, you may assign the payment of this claim to your pharmacy if your pharmacy is willing to accept assignment, or not complete this form.

Please request that your pharmacy contact Pharmacy Services at 800.922.1557 for assistance.

- 25 total fields (20 text / 5 symbolic)
- Text and symbolic entry
- **Poses additional difficulty for Vision API due to bounded characters**

California Medical Claim Form


Member Claim Form

Anthem Blue Cross

Please use a separate claim form for each patient. Your cooperation in completing all items on the claim form and attaching all required documentation will help expedite quick and accurate processing. SEE REVERSE SIDE FOR COMPLETE INSTRUCTIONS.

Section A. PATIENT INFORMATION

Last name First name M.I.

Does the patient have other health insurance coverage? Yes No

Relation to subscriber: Self Spouse Son Daughter

Sex: M F

Date of birth (MM/DD/YYYY)

Name of other health insurance company Group no. Employer name Policy no.

Section B. SUBSCRIBER INFORMATION (on Anthem Blue Cross card)

Identification no. Group no.

Last name First name M.I.

Street address (please include apt. no.)

City State ZIP code

Home phone no. Work phone no. Date of birth (MM/DD/YYYY)

Section C. MEDICAL INFORMATION

HEALTH CARE SERVICES: Use this section to report any COVERED health service that has not already been reported to this Anthem Blue Cross Plan by the provider of service (the physician, clinical, ambulance company, private duty nurse, etc.) Attach itemized bill or photocopy. Please be sure that duplicate bills are not submitted.

Was this medical expense the result of an accident? Yes No

Was this condition or injury job related? Yes No

Have you filed for Workers' Compensation? Yes No

When did this injury or accident occur? (MM/DD/YYYY)

Diagnosis code Procedure code Tax ID

BILLS MUST BE ITEMIZED

Canceled checks, cash register receipts and non-itemized "balance due" statements cannot be processed. Each item and bill must include:

- Name and address of provider (doctor, hospital, laboratory, ambulance service, etc.)
- Name of patient
- Service provided
- Date of service
- Amount charged for each service
- Diagnosis code
- Procedure code
- ICD-10

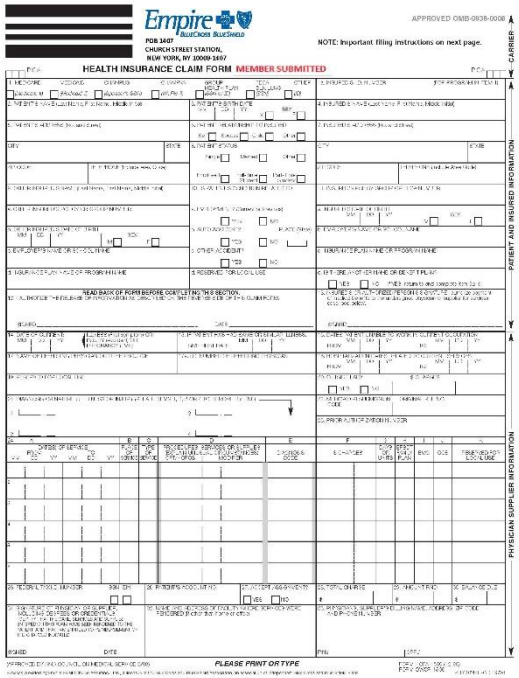
I certify that, to the best of my knowledge, the information on this Member Claim Form is true and correct. I authorize the release of any medical information necessary to process this claim.

Signature Name Date

X

- 29 total fields (23 text / 6 symbolic)
- Bounded text boxes
- Free form text entry
- **Similar character bounding difficulty**

New York Medical Claim Form


Empire BlueCross BlueShield

P.O. BOX 1407
CHURCH STREET STATION,
NEW YORK, NY 10008-1407

APPROVED OMB-0938-0008

NOTE: Important filling instructions on next page.

HEALTH INSURANCE CLAIM FORM MEMBER SUBMITTED

PLEASE PRINT OR TYPE

CARRIER

PATIENT AND INSURED INFORMATION

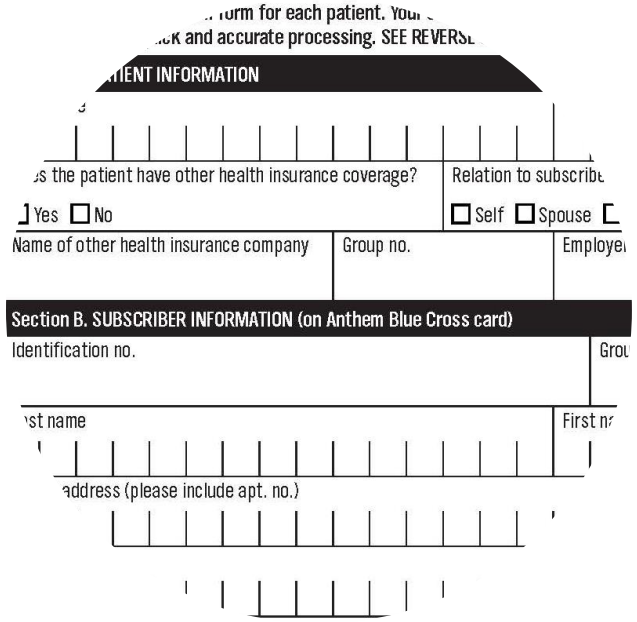
PHYSICIAN SUPPLIER INFORMATION

- 60 total fields (47 text / 13 symbolic)
- Multiple rows / columns
- Numeric and text free form
- **Overlap of handwriting in freeform area**

Entry Types

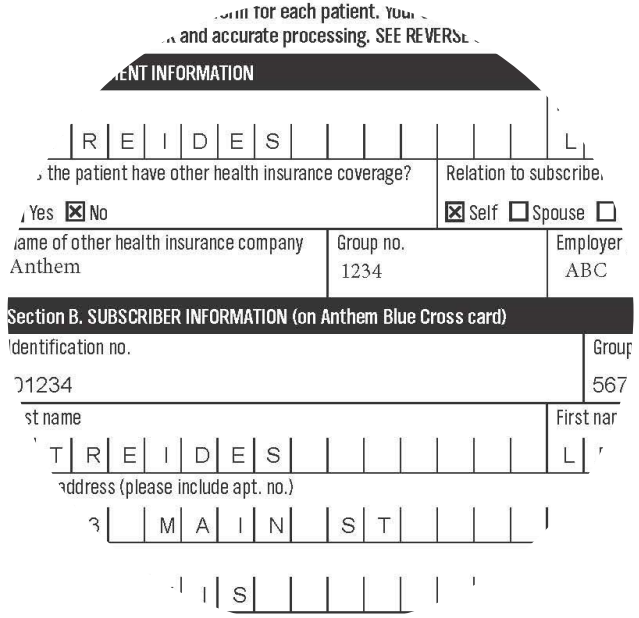
Several entry types were used to understand how model performance differs within identical forms

Blank



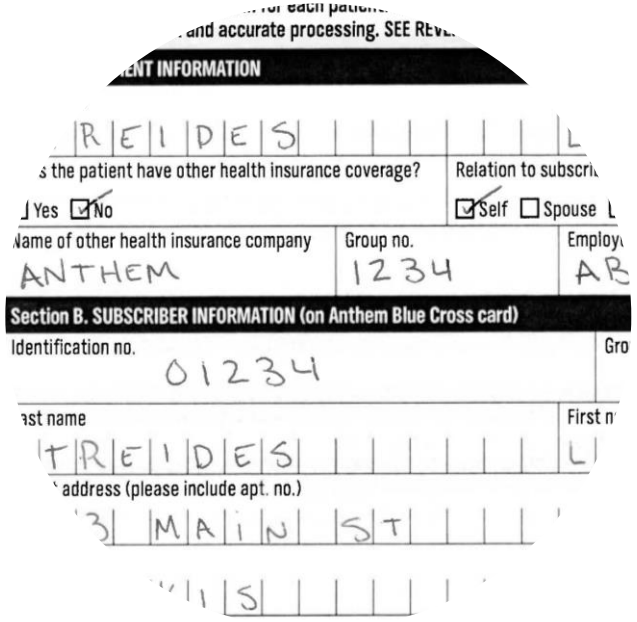
- Identifies bounds for fields
- Baseline for model performance
- Useful for form identification

Typed



- Variable performance across forms
- Affected by distance and spacing between fields

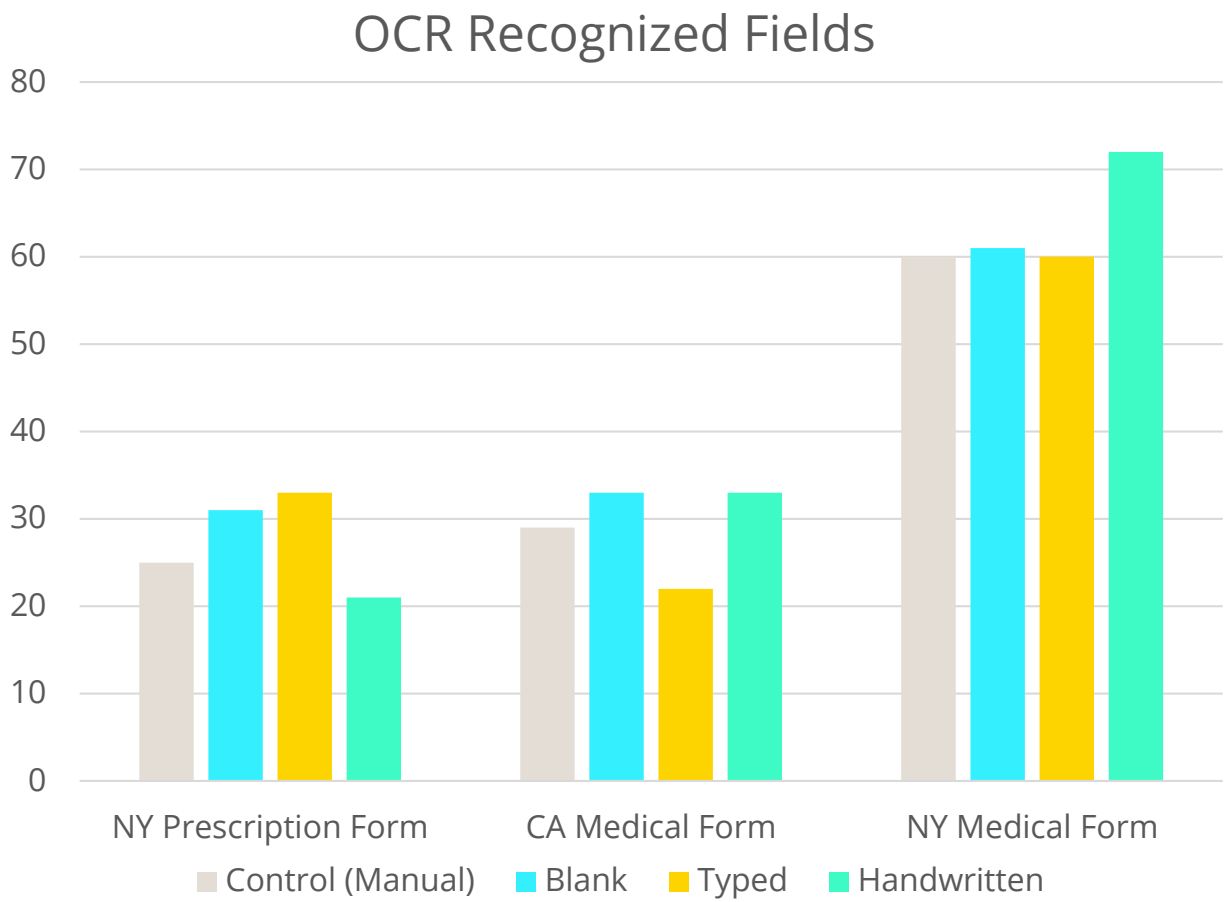
Handwritten



- Highly variable performance within forms
- Affected by a variety of factors (i.e. pen color, size, etc)

OCR: Segmentation

Testing the API's auto segmentation feature



API Segmentation is Highly Variable

Even in forms where the amount of fields is relatively similar, the content included in each segment is inconsistent and/or overlapping

Not All Information is Captured

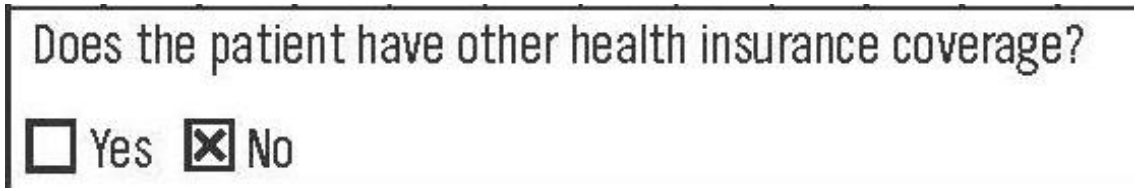
When automatic segmentation is performed by the API, some information is omitted or truncated due to being outside of the bounds

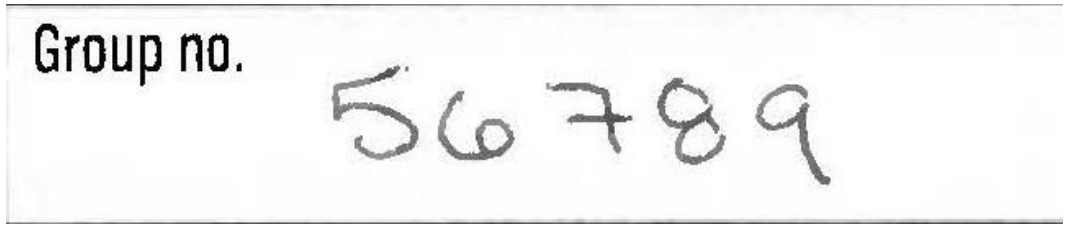
Preprocessing suggested to define field boundaries for each form type

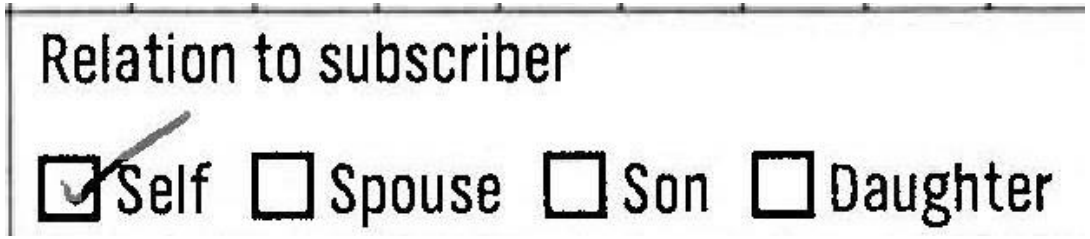
OCR: Accuracy

Isolated fields were compared to determine accuracy of handwritten vs. typed text

Field Name	Data Type	Actual	Handwritten	Typed
Last Name	Text	ATREIDES	ATRIEL PIESE	ALTRE DIES |
Group Number	Numeric	56789	56789	56789
Coverage*	Symbolic (Yes / No)	No	N/A	N/A
Subscriber Relation*	Symbolic (Multiple)	Self	N/A	N/A



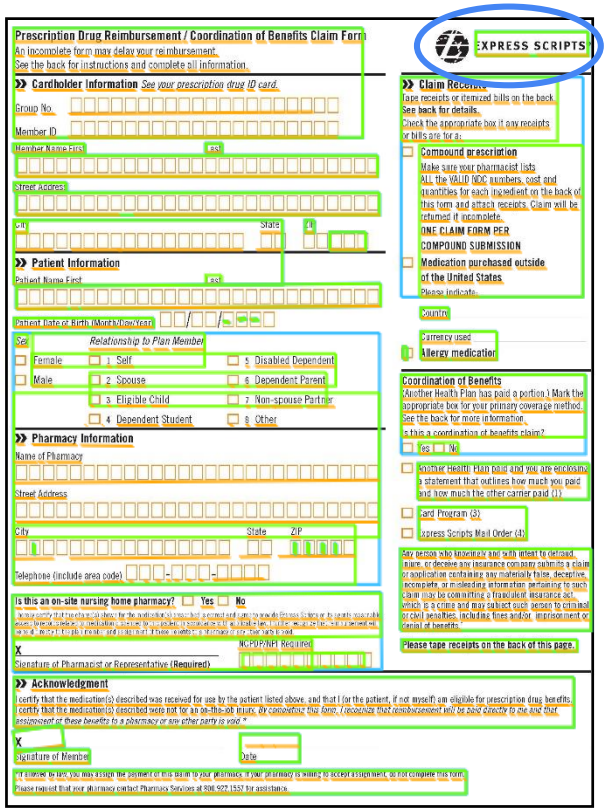




***Symbols to be handled by AutoML preprocessing**

Suggestions for Future Work

Addition of pre-processing and form validation would dramatically improve ability to discern individual fields

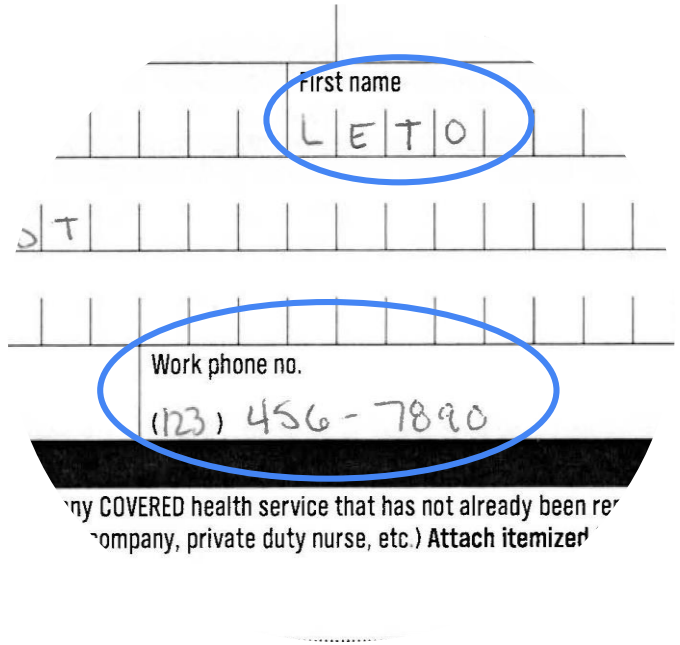


Segmentation Based on Form Type

- Removes variance within forms caused by font placement, size, field overlap
- Requires additional API calls
  - API Call #1: Form recognition
  - API Call #2: OCR

Data Validation

- Preprocessing includes data validation rules to guide field specific values and recognize checkboxes
- Suggested validation rules:
  - Data value (alphanumeric, symbol)
  - Data format (dd/mm/yy, phone #)



Thank you.

Philip Mohun

Deloitte Consulting LLP

GCP Professional Data Engineer *(Candidate)*

Contact: phmohun@deloitte.com