Classifying Immigration Documents for Bridge US

By Baolin Liu (MS in Data Science)

BridgeUS

IMMIGRATION SIMPLIFIED

About Bridge US:

- Immigration service provider that streamlines the US visa application process for employers.
- Works with companies like Getaround, Allegiant, and American Family Insurance.
- Over 50k immigration related documents managed.

Motivation

- Typical process includes > 20 files
- Identifying documents takes time
- Files are often misplaced
- No initial feedback

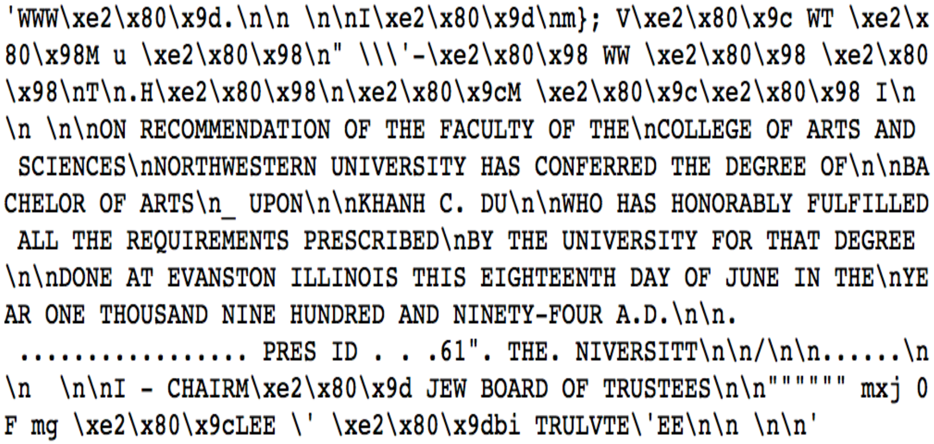
Goals:

- Accurately identify documents
- Guide users to provide what’s missing
- Real-time feedback
- Enable new products based on less intensive document oversight

Data Collection and "Munging"

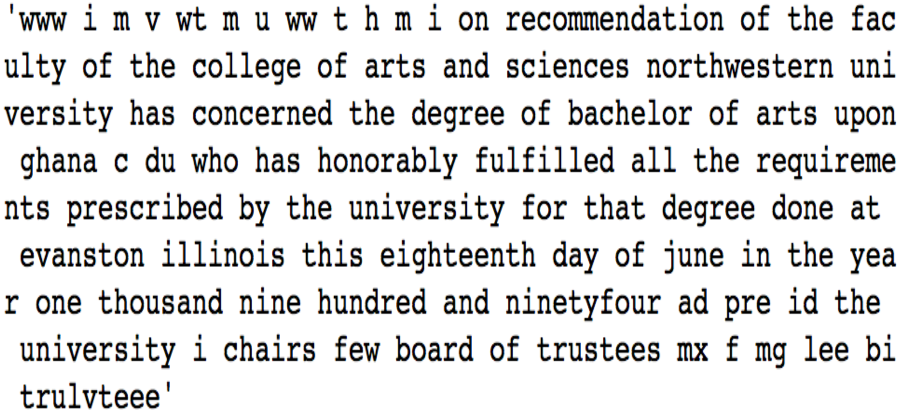
- Text Extraction with Textract
- JPG, PNG,PDF,Doc,Docx, it does it all!
- Regex commands to remove Unicode
- Stemmed the words using a Porter Stemmer
- Ran PySpark to complete the extraction(now Batch Adobe OCR)
- Used a HashVectorizer for my text to matrix transformation(consistent matrix size)
- Saved all my text into a CSV File.

Before:



```
'WWW\xe2\x80\x9d.\n\n \n\nI\xe2\x80\x9d\nm}; V\xe2\x80\x9c WT \xe2\x
80\x98M u \xe2\x80\x98\n" \\\'-\xe2\x80\x98 WW \xe2\x80\x98 \xe2\x80
\x98\nT\n.H\xe2\x80\x98\n\xe2\x80\x9cM \xe2\x80\x9c\xe2\x80\x98 I\n
\n \n\nON RECOMMENDATION OF THE FACULTY OF THE\nCOLLEGE OF ARTS AND
 SCIENCES\nNORTHWESTERN UNIVERSITY HAS CONFERRED THE DEGREE OF\n\nBA
CHELOR OF ARTS\n_ UPON\n\nKHANH C. DU\n\nWHO HAS HONORABLY FULFILLED
 ALL THE REQUIREMENTS PRESCRIBED\nBY THE UNIVERSITY FOR THAT DEGREE
\n\nDONE AT EVANSTON ILLINOIS THIS EIGHTEENTH DAY OF JUNE IN THE\nYE
AR ONE THOUSAND NINE HUNDRED AND NINETY-FOUR A.D.\n\n.
 ................. PRES ID . . .61". THE. NIVERSITT\n\n/\n\n......\n
\n  \n\nI - CHAIRM\xe2\x80\x9d JEW BOARD OF TRUSTEES\n\n"""""" mxj 0
F mg \xe2\x80\x9cLEE \' \xe2\x80\x9dbi TRULVTE\'EE\n\n \n\n'
```

After:

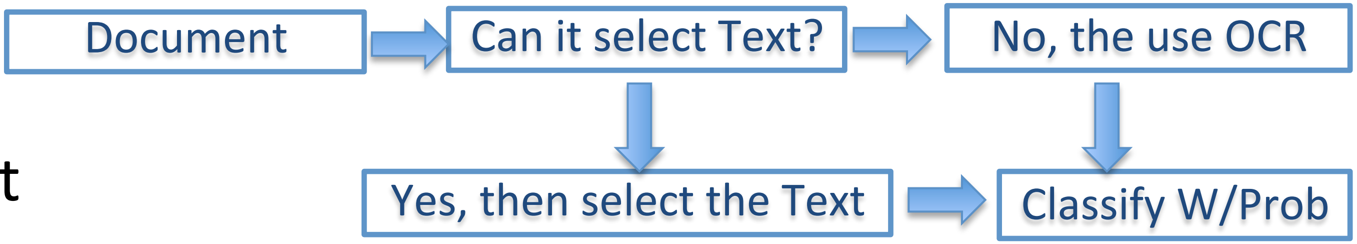


```
'www i m v wt m u ww t h m i on recommendation of the fac
ulty of the college of arts and sciences northwestern uni
versity has concerned the degree of bachelor of arts upon
 ghana c du who has honorably fulfilled all the requireme
nts prescribed by the university for that degree done at
 evanston illinois this eighteenth day of june in the yea
r one thousand nine hundred and ninetyfour ad pre id the
 university i chairs few board of trustees mx f mg lee bi
 trulvteee'
```

DataFrame

- Appended everything together in one big column
- Properly labeled everything
- Multi-Class Classification (One vs. Rest) solution was used

Classification Algorithm


Document

Can it select Text?

No, the use OCR

Yes, then select the Text

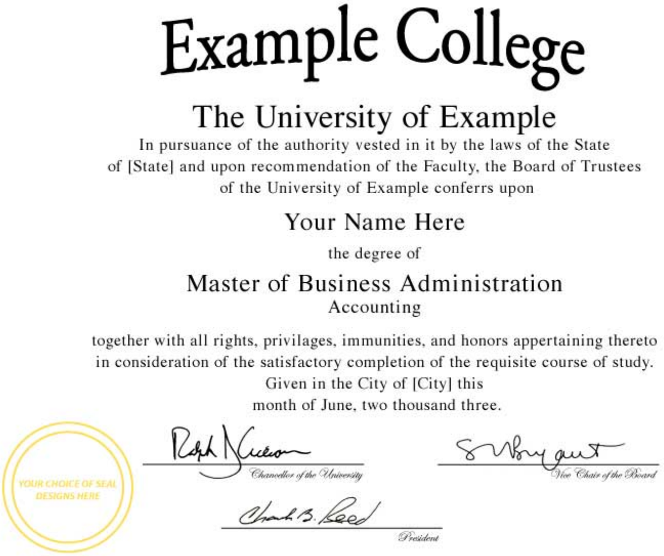
Classify W/Prob

- 2 pathways:

-selectable text

-image text

- Assigned probability scores ranked highest to lowest


Example College

The University of Example

In pursuance of the authority vested in it by the laws of the State of [State] and upon recommendation of the Faculty, the Board of Trustees of the University of Example confers upon

Your Name Here

the degree of

Master of Business Administration
Accounting

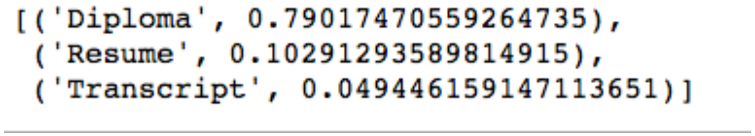
together with all rights, privilages, immunities, and honors appertaining thereto in consideration of the satisfactory completion of the requisite course of study.
Given in the City of [City] this
month of June, two thousand three.

YOUR CHOICE OF SEAL DESIGNS HERE

Chancellor of the University

Vice Chair of the Board

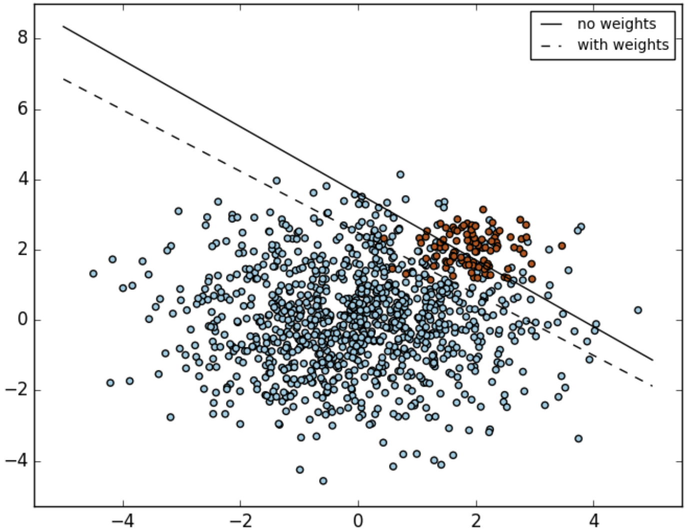
President


```
[('Diploma', 0.79017470559264735),
 ('Resume', 0.10291293589814915),
 ('Transcript', 0.049446159147113651)]
```

http://cheaper-than-tuition.com/shop/fake-diploma/

Modeling: SVM

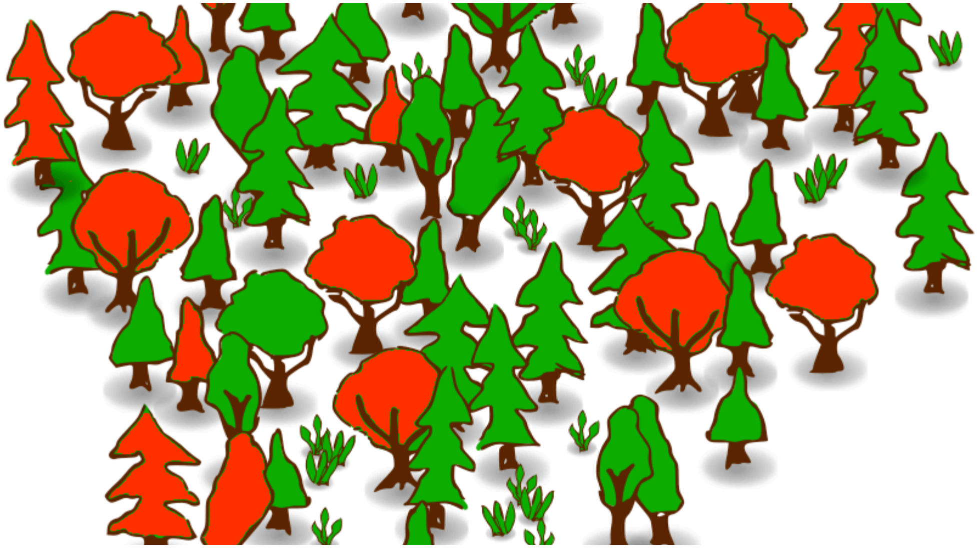
- Multi-class classification model using Support Vector Classifier(SVM) with a Linear Kernel
- Handled unbalanced classes, Resumes
- Dealt with categories with low samples.



http://scikit-learn.org/stable/auto_examples/svm/plot_separating_hyperplane_unbalanced.html

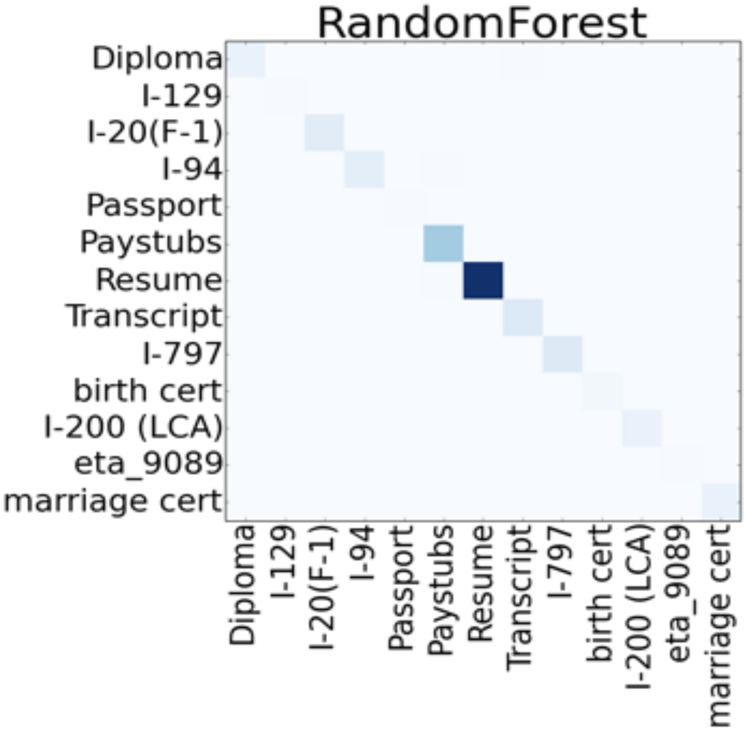
Modeling: Random Forest

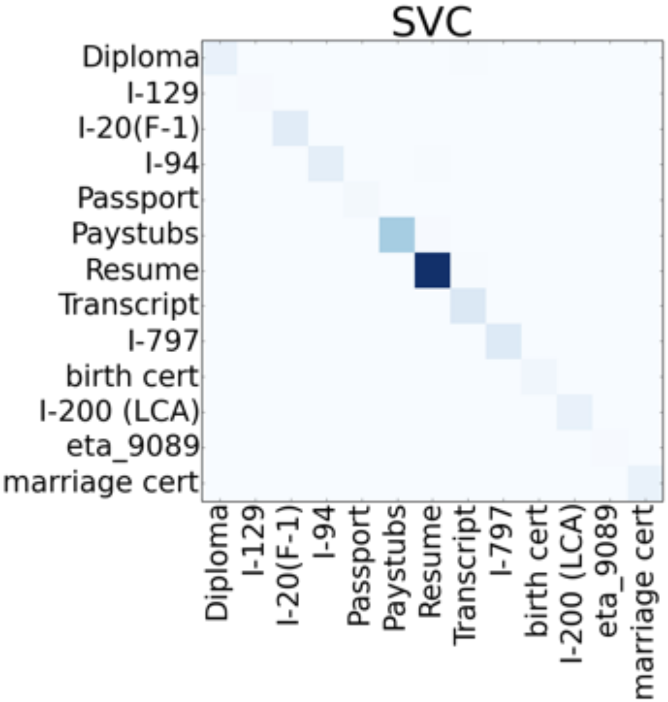
- Repeated sampling, or “bootstrapping”
- Had the “Wisdom of the Crowd”
- Gini/Entropy split created decisions well



http://blog.yhat.com/posts/random-forests-in-python.html

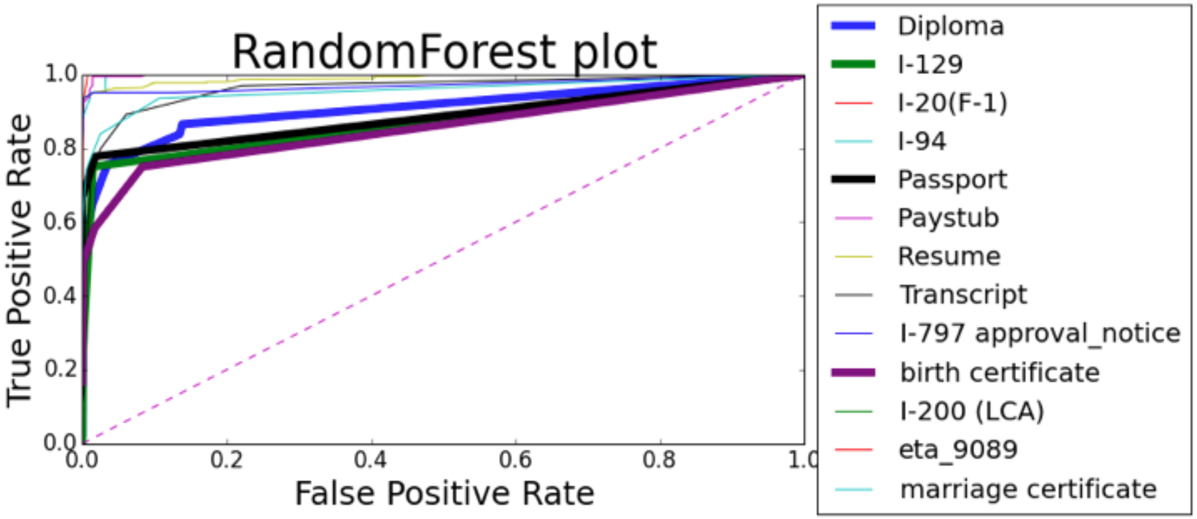
Confusion Matrix

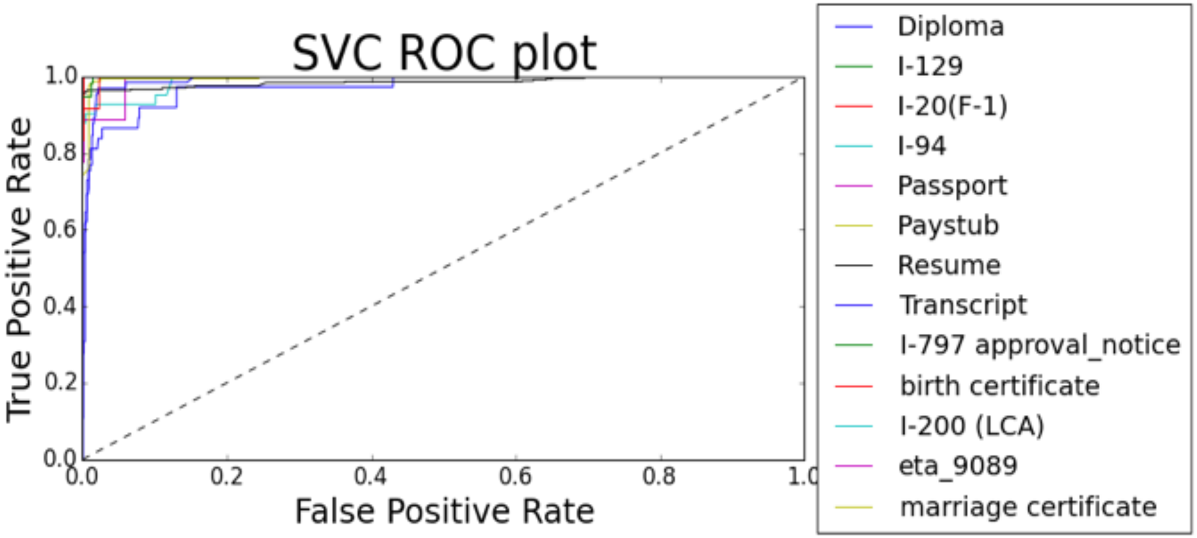




For SVC, there were 48 documents that were misclassified as opposed to 60 documents for Random Forest.

ROC Plots


RandomForest plot
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Diploma
I-129
I-20(F-1)
I-94
Passport
Paystub
Resume
Transcript
I-797 approval_notice
birth certificate
I-200 (LCA)
eta_9089
marriage certificate


SVC ROC plot
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Diploma
I-129
I-20(F-1)
I-94
Passport
Paystub
Resume
Transcript
I-797 approval_notice
birth certificate
I-200 (LCA)
eta_9089
marriage certificate

Cons for both models:

- Gave high confidence scores for documents that did not belong to any class, no good way for probabilities
- Although Random Forest did better at probabilities using ensemble method, it did not perform well for low samples
- Ultimately chose SVC because it handled the low samples well.

Moving code to production:

- Document splits and classifies only the first page
- Classify the document as an Image
- Model Persistence, pickling
- Py files to launch from command line
- AWS

Production Testing

Data:

- 1664 PDF files from 87 case files

Initial Findings:

- need more classes: job offer, wage surveys, tax returns, and more

Did Well:

- government forms, pay stubs and resumes

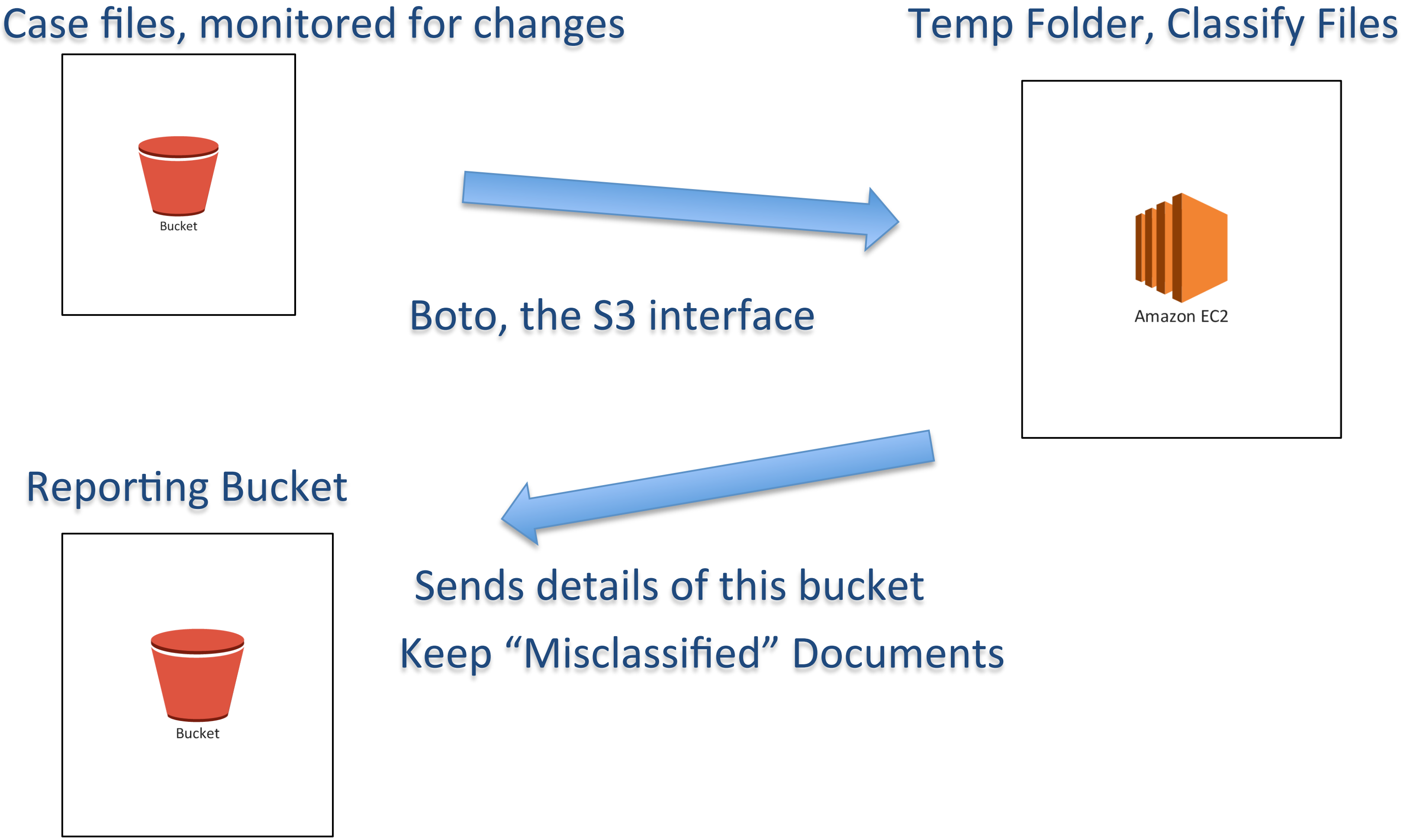
Did Moderate:

- Transcripts, Diplomas, Birth Certificates, Marriage Certificates

Did Poor:

- Passports

Data Engineering Architecture (Proof of Concept)



Case files, monitored for changes

Temp Folder, Classify Files

Boto, the S3 interface

Reporting Bucket

Sends details of this bucket

Keep “Misclassified” Documents

Final Categories

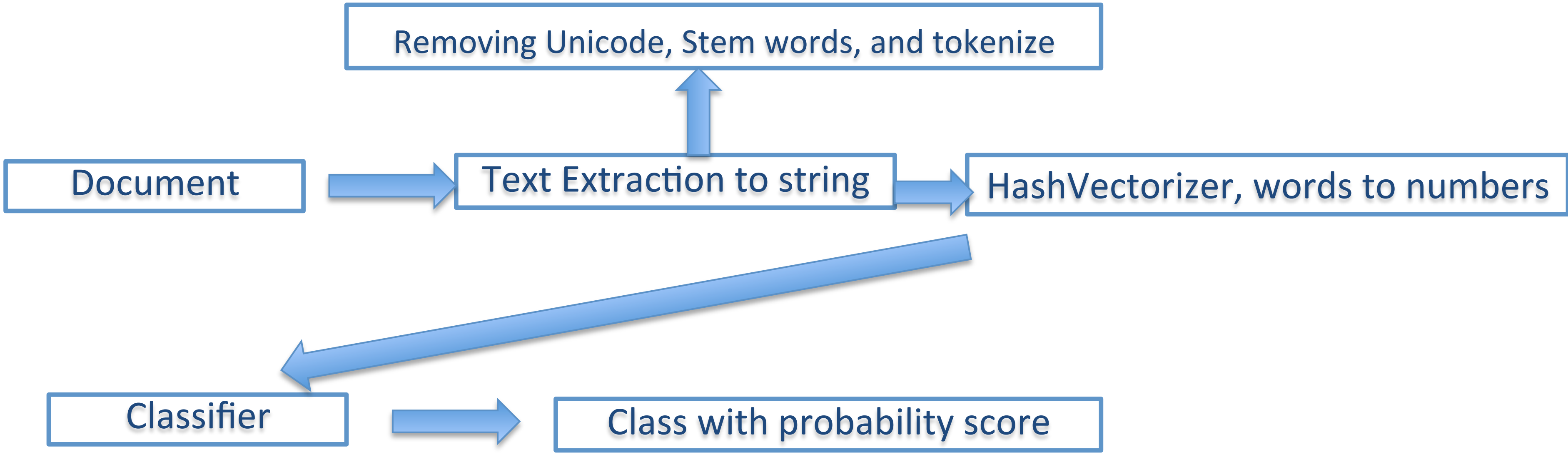
Text Classifier:

- Academic Equivalency
- Attorney Client Agreement
- I -200(LCA)
- ETA 9089
- ETA 9141
- I 20
- I 94
- I 140
- I 129
- Itinerary
- Job Offer
- Lease Agreement
- MSA
- Paystubs
- POA
- Resume
- SOW
- Tax return
- W2
- Wage Survey
- Termination Letter

Image Classifier:

- Academic Equivalency
- Birth certificate
- Diploma
- Employment Authorization Card
- ETA 9141
- ETA 9089
- ETA 140
- I 20
- I 129
- I 797 Approval
- I 797 Receipt
- I 94
- LCA
- Marriage Certificate
- Org Chart
- Passport
- Paystub
- Permanent Card
- Transcript
- W2
- Wage Survey

Text Classification Process



Removing Unicode, Stem words, and tokenize

Document

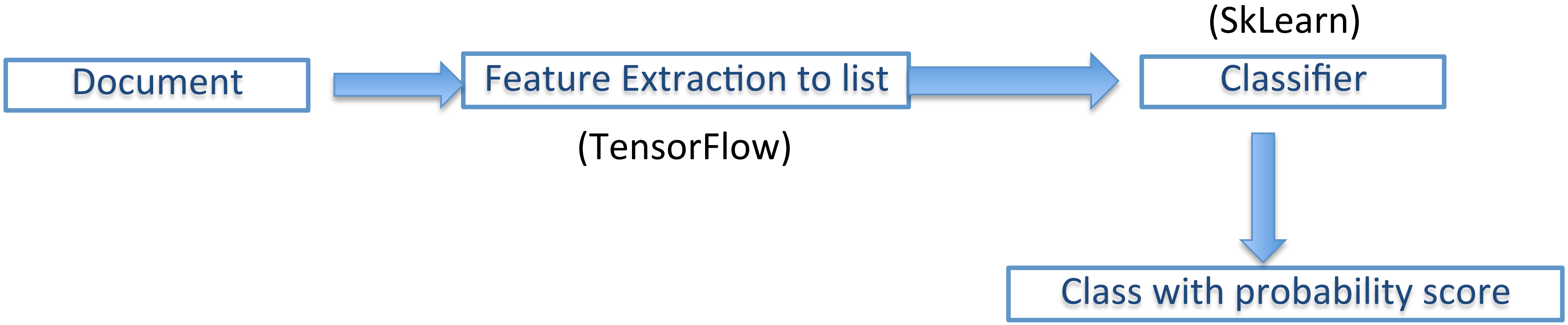
Text Extraction to string

HashVectorizer, words to numbers

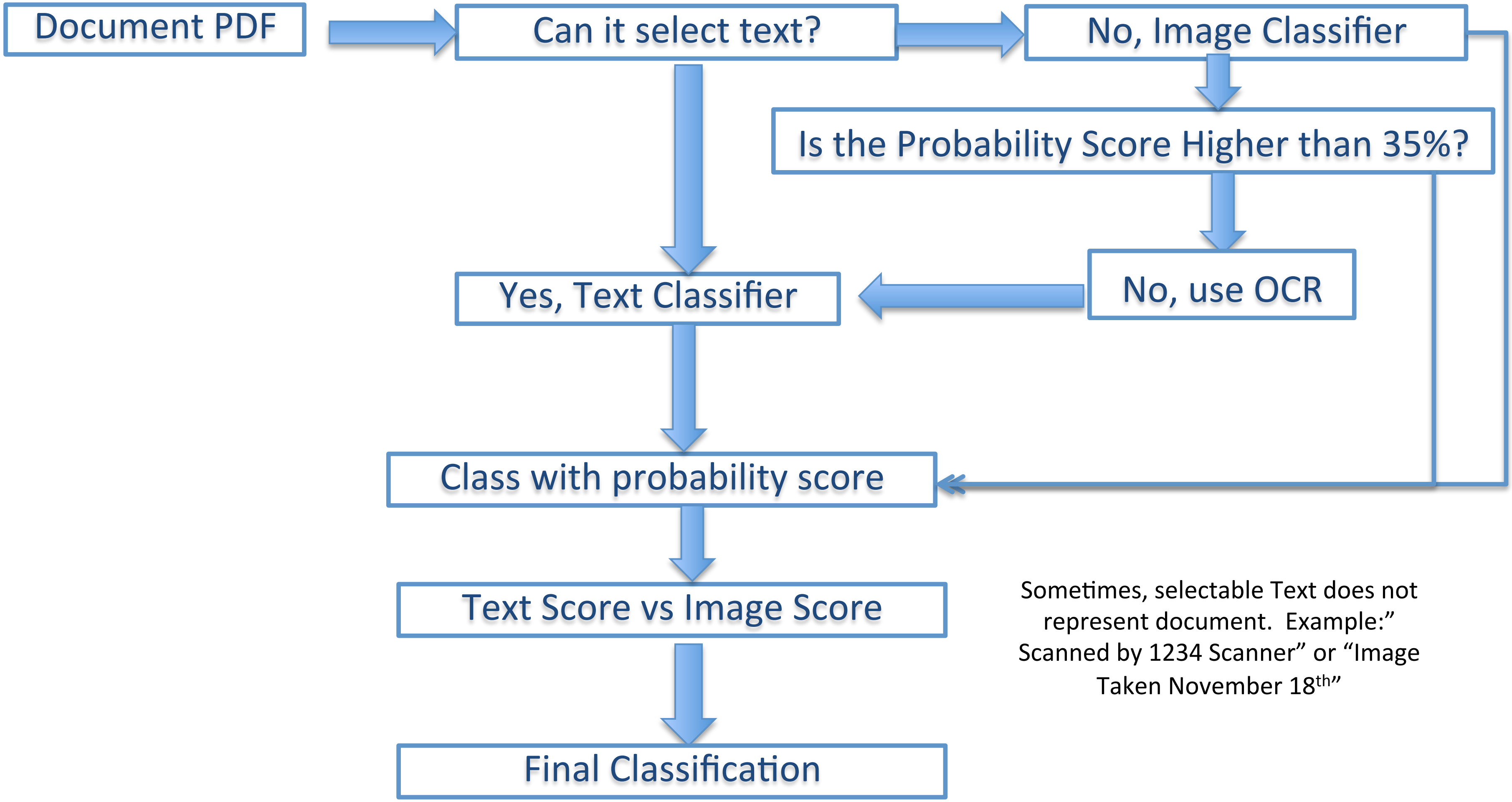
Classifier

Class with probability score

Image Classification Process

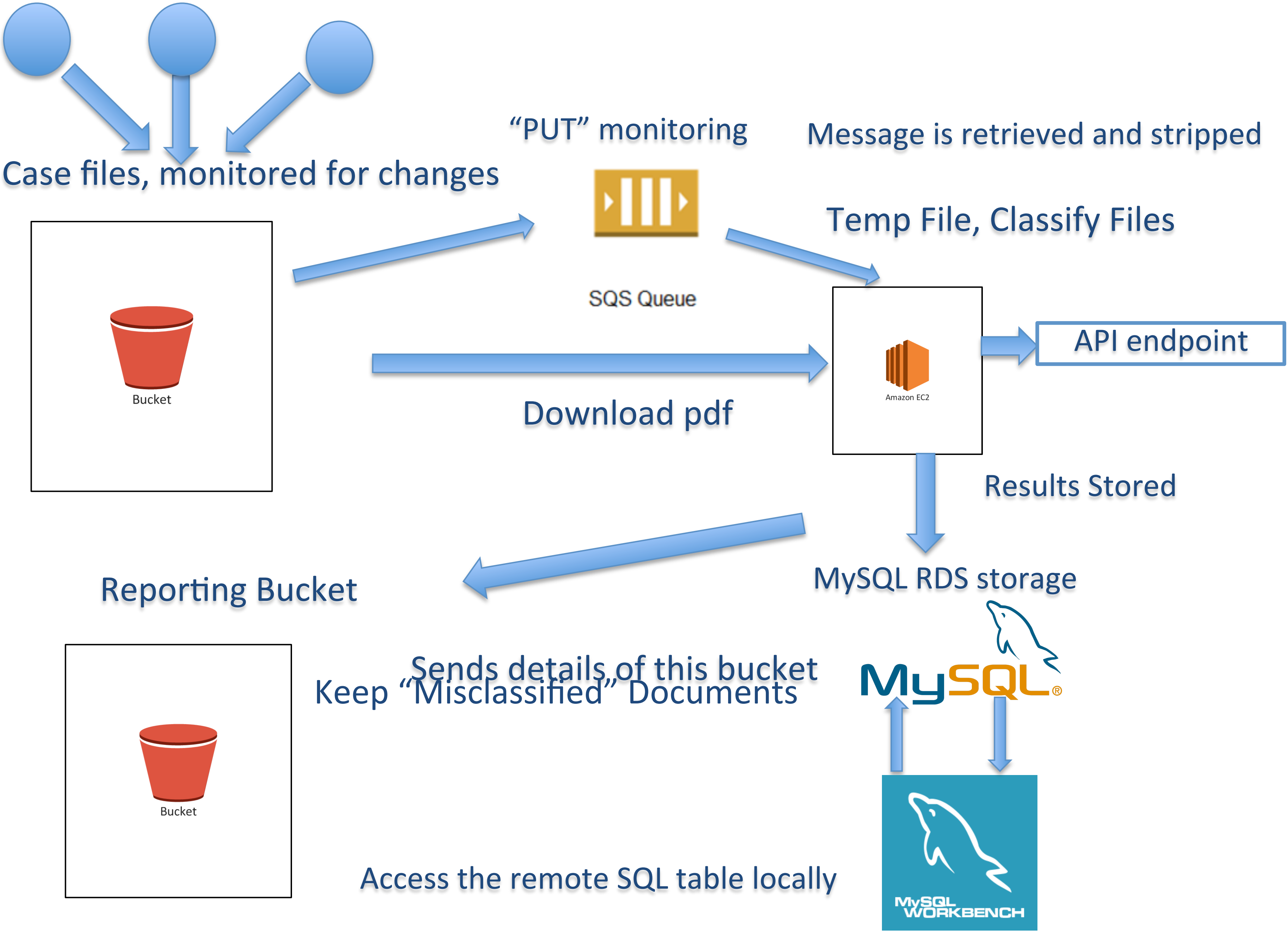


Classification Workflow



Sometimes, selectable Text does not represent document. Example:" Scanned by 1234 Scanner" or "Image Taken November 18th"

DE Architecture



Files are dropped

Case files, monitored for changes

“PUT” monitoring

Message is retrieved and stripped

Temp File, Classify Files

Download pdf

API endpoint

Results Stored

MySQL RDS storage

Reporting Bucket

Sends details of this bucket

Keep “Misclassified” Documents

Access the remote SQL table locally

The End ☺

Contacts:

- GalvanizeU Master's in Data Science partnership:
- Bonnie Xie(Student Director) bonny.xie@galvanize.com
- Bridge US: Forrest Blount (CTO): forrest@bridge.us

  https://www.bridge.us/
- Baolin Liu-Galvanize Master's in Data Science: baolin.liu@gmail.com

(Looking for Opportunities!)