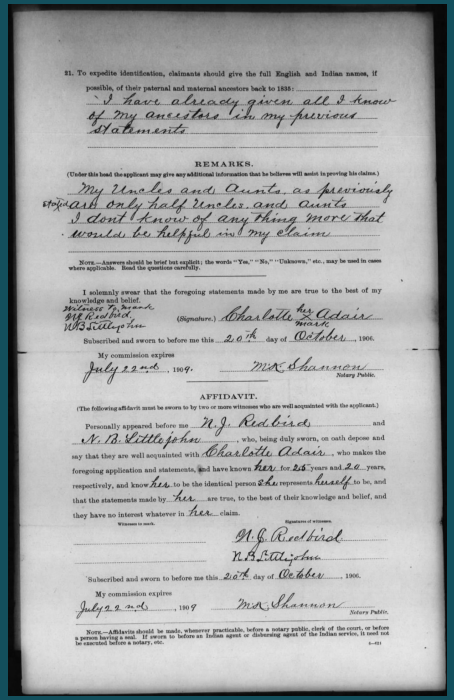
DANIEL NASHED

➔ https://www.fold3.com/publication/73/us-eastern-cherokee-applications-1906-1909

ML PIPELINE FOR DOCUMENT CLUSTERING

An Automated Solution for Signature Detection + Clustering



21. To expedite identification, claimants should give the full English and Indian names, if possible, of their paternal and maternal ancestors back to 1835:

I have already given all I know of my ancestors in my pervious Statements

REMARKS.

(Under this head the applicant may give any additional information that he believes will assist in proving his claims.)

My Uncles and Aunts as previously Stated are only half Uncles and Aunts I dont know of any thing more that would be helpful in my Claim

NOTE.—Answers should be brief but explicit; the words "Yes," "No," "Unknown," etc., may be used in cases where applicable. Read the questions carefully.

I solemnly swear that the foregoing statements made by me are true to the best of my knowledge and belief.

Witness To Mark
N.J. Redbird
N.B. Littlejohn

(Signature.) Charlotte her X mark Adair

Subscribed and sworn to before me this 20th day of October, 1906.

My commission expires July 22nd, 1909. M.K. Shannon Notary Public.

AFFIDAVIT.

(The following affidavit must be sworn to by two or more witnesses who are well acquainted with the applicant.)

Personally appeared before me N.J. Redbird and N.B. Littlejohn, who, being duly sworn, on oath depose and say that they are well acquainted with Charlotte Adair, who makes the foregoing application and statements, and have known her for 25 years and 20 years, respectively, and know her to be the identical person she represents herself to be, and that the statements made by her are true, to the best of their knowledge and belief, and they have no interest whatever in her claim.

Witnesses to mark. Signatures of witnesses.
N.J. Redbird
N.B. Littlejohn

Subscribed and sworn to before me this 20th day of October, 1906.

My commission expires July 22 nds, 1909 M.K. Shannon Notary Public.

NOTE.—Affidavits should be made, wherever practicable, before a notary public, clerk of the court, or before a person having a seal. If sworn to before an Indian agent or disbursing agent of the Indian service, it need not be executed before a notary, etc.

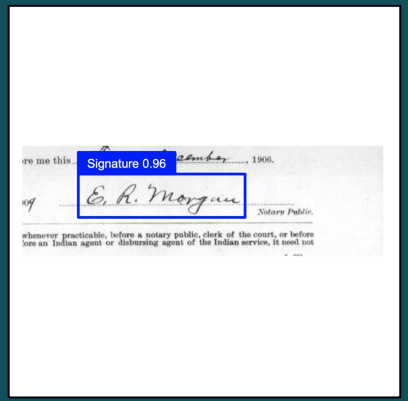


TABLE OF CONTENTS

1 ..... Problem Statement
2 ..... Goals
3 ..... System Overview
4 ..... Demo
5 ..... Future Improvements

PROBLEM STATEMENT

Need to Scrape + Cluster 50,000 Scanned Documents Based on Signature

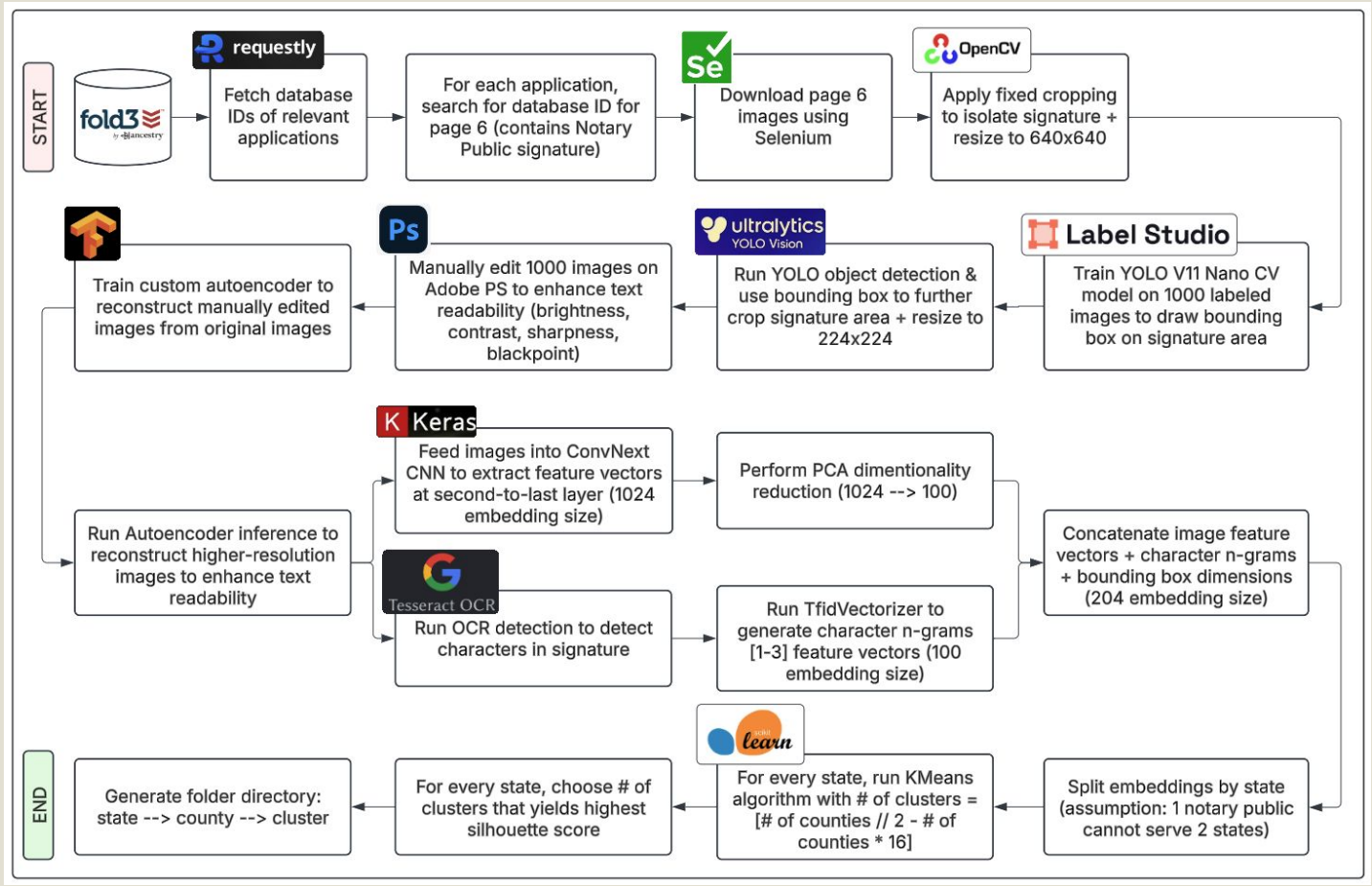
- ➔ Need an automated solution to collect all 50,000 applications from Fold3.com
- ➔ Need an automated solution to cluster documents into folders, where each folder contains same signature by Notary Public

GOALS

- ➔ Fold3.com does not expose API so use **web scraping** methods to navigate through HTML to download required documents
- ➔ Run **object detection** using **computer vision** to isolate and extract signature area from document
- ➔ Transform cropped signature images into **vector embeddings** using pre-trained **CNNs**
- ➔ Run **clustering** on embeddings to group similar signatures together

SYSTEM OVERVIEW

SYSTEM DIAGRAM

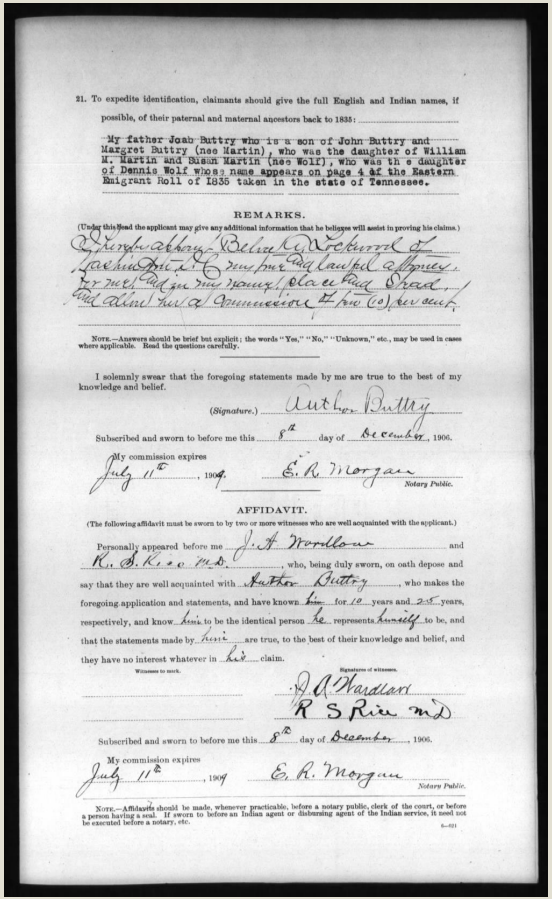


TECH STACK

- Data Collection
  - Requestly to intercept HTTP POST requests and retrieve all database IDs
  - Selenium with ChromeDriver to automate browser navigation for downloading images
- Image Processing + Labeling
  - OpenCV to read images and perform cropping + resizing
  - Adobe Photoshop for manual enhancement of text readability
  - Tesseract OCR to run character recognition on signatures
  - Label Studio for image labeling + annotation
- Computer Vision
  - YOLO V11 Nano (Ultraytics) to run object detection on signatures
  - ConvNext Base (using Keras) to extract deep feature embeddings from images
  - Custom autoencoder (using Tensorflow) to reconstruct high resolution images
- Embeddings Post-processing
  - PCA (using Scikit-learn) for dimensionality reduction
  - TfidVectorizer for generating character n-grams
- Clustering
  - KMeans & DBSCAN for clustering of embeddings

DEMO

DEMO



21. To expedite identification, claimants should give the full English and Indian names, if possible, of their paternal and maternal ancestors back to 1835:

My father Joab Buttry who is a son of John Buttry and Margret Buttry (nee Martin), who was the daughter of William M. Martin and Diana Martin (nee Wolf), who was th e daughter of Dennis Wolf whose name appears on page 4 of the Eastern Emigrant Roll of 1835 taken in the state of Tennessee.

REMARKS.

(Under this head the applicant may give any additional information that he believes will assist in proving his claims.)

Note.—Answers should be brief but explicit; the words "Yes," "No," "Unknown," etc., may be used in cases where applicable. Read the questions carefully.

I solemnly swear that the foregoing statements made by me are true to the best of my knowledge and belief.

(Signature.) Luther Buttry

Subscribed and sworn to before me this 8th day of December, 1906.

My commission expires July 11th, 1909. E. R. Morgan Notary Public.

AFFIDAVIT.

(The following affidavit must be sworn to by two or more witnesses who are well acquainted with the applicant.)

Personally appeared before me J. A. Wardlaw and R. S. Rice M.D., who, being duly sworn, on oath depose and say that they are well acquainted with Luther Buttry, who makes the foregoing application and statements, and have known him for 10 years and 25 years, respectively, and know him to be the identical person he represents himself to be, and that the statements made by him are true, to the best of their knowledge and belief, and they have no interest whatever in his claim.

Witnesses to mark. Signatures of witnesses. J. A. Wardlaw, R. S. Rice M.D.

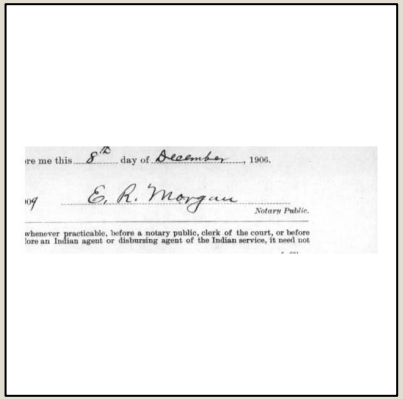
Subscribed and sworn to before me this 8th day of December, 1906.

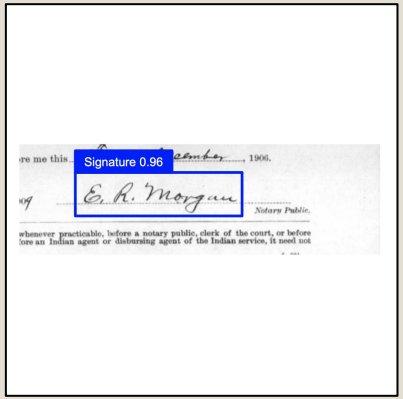
My commission expires July 11th, 1909. E. R. Morgan Notary Public.

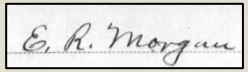
Note.—Affidavits should be made, whenever practicable, before a notary public, clerk of the court, or before a person having a seal. If sworn to before an Indian agent or disbursing agent of the Indian service, it need not be executed before a notary, etc.

6—921

Fixed cropping of right bottom corner	Object detection to draw bounding box	Crop to bounding box size	Resize to 224×224







DEMO

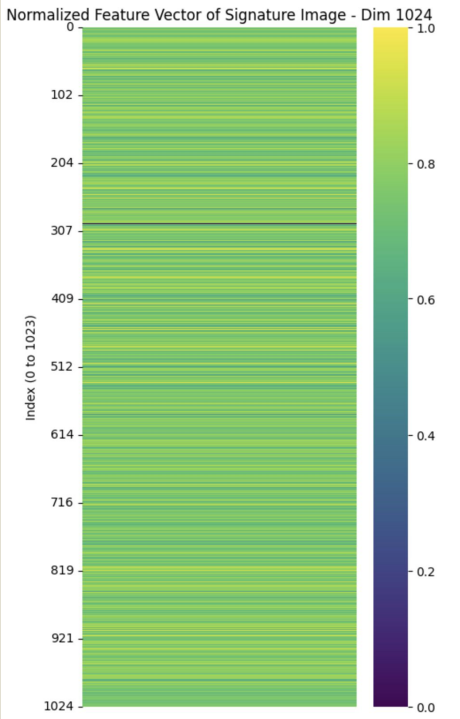
State	(# of images x image height x image width x color channels)
Rhode Island	(2, 224, 224, 3)
Iowa	(36, 224, 224, 3)
Indiana	(378, 224, 224, 3)
Kentucky	(259, 224, 224, 3)
Illinois	(141, 224, 224, 3)
Other	(2354, 224, 224, 3)
Arizona	(17, 224, 224, 3)
New York	(9, 224, 224, 3)
South Carolina	(92, 224, 224, 3)
Georgia	(4393, 224, 224, 3)
Oklahoma	(12024, 224, 224, 3)
Oregon	(20, 224, 224, 3)
Wisconsin	(45, 224, 224, 3)
Colorado	(94, 224, 224, 3)
Tennessee	(3281, 224, 224, 3)
Michigan	(39, 224, 224, 3)
Washington	(63, 224, 224, 3)
District Of Columbia	(6, 224, 224, 3)
North Carolina	(3067, 224, 224, 3)
Florida	(133, 224, 224, 3)
Nevada	(5, 224, 224, 3)
Nebraska	(25, 224, 224, 3)
Texas	(958, 224, 224, 3)
Wyoming	(7, 224, 224, 3)
Missouri	(1308, 224, 224, 3)
Virginia	(722, 224, 224, 3)
West Virginia	(479, 224, 224, 3)
Alabama	(1253, 224, 224, 3)
Louisiana	(47, 224, 224, 3)
California	(177, 224, 224, 3)
New Mexico	(66, 224, 224, 3)
Arkansas	(969, 224, 224, 3)
Pennsylvania	(17, 224, 224, 3)
Maryland	(6, 224, 224, 3)
New Jersey	(7, 224, 224, 3)
Montana	(15, 224, 224, 3)
Minnesota	(7, 224, 224, 3)
Massachusetts	(1, 224, 224, 3)
Alaska Territory	(2, 224, 224, 3)
Ohio	(41, 224, 224, 3)
Idaho	(20, 224, 224, 3)
Kansas	(493, 224, 224, 3)
North Dakota	(3, 224, 224, 3)
Mississippi	(158, 224, 224, 3)
Hawaii Territory	(1, 224, 224, 3)
Utah	(4, 224, 224, 3)

Extract latent feature vectors by passing all images through ConvNext

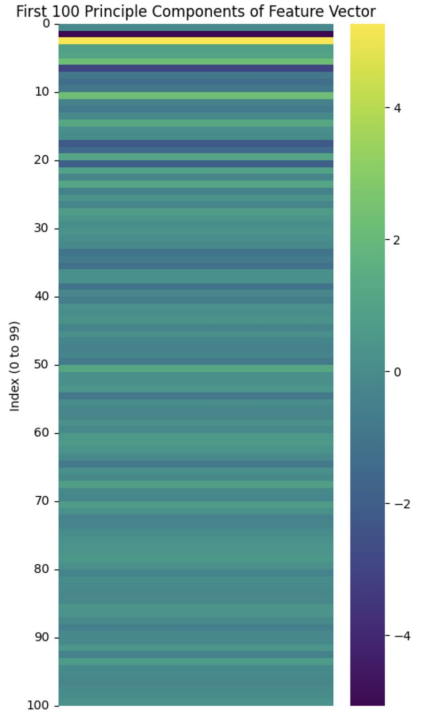
State	(# of images x embedding size)
Rhode Island	(2, 1024)
Iowa	(36, 1024)
Indiana	(378, 1024)
Kentucky	(259, 1024)
Illinois	(141, 1024)
Other	(2354, 1024)
Arizona	(17, 1024)
New York	(9, 1024)
South Carolina	(92, 1024)
Georgia	(4393, 1024)
Oklahoma	(12024, 1024)
Oregon	(20, 1024)
Wisconsin	(45, 1024)
Colorado	(94, 1024)
Tennessee	(3281, 1024)
Michigan	(39, 1024)
Washington	(63, 1024)
District Of Columbia	(6, 1024)
North Carolina	(3067, 1024)
Florida	(133, 1024)
Nevada	(5, 1024)
Nebraska	(25, 1024)
Texas	(958, 1024)
Wyoming	(7, 1024)
Missouri	(1308, 1024)
Virginia	(722, 1024)
West Virginia	(479, 1024)
Alabama	(1253, 1024)
Louisiana	(47, 1024)
California	(177, 1024)
New Mexico	(66, 1024)
Arkansas	(969, 1024)
Pennsylvania	(17, 1024)
Maryland	(6, 1024)
New Jersey	(7, 1024)
Montana	(15, 1024)
Minnesota	(7, 1024)
Massachusetts	(1, 1024)
Alaska Territory	(2, 1024)
Ohio	(41, 1024)
Idaho	(20, 1024)
Kansas	(493, 1024)
North Dakota	(3, 1024)
Mississippi	(158, 1024)
Hawaii Territory	(1, 1024)
Utah	(4, 1024)

DEMO

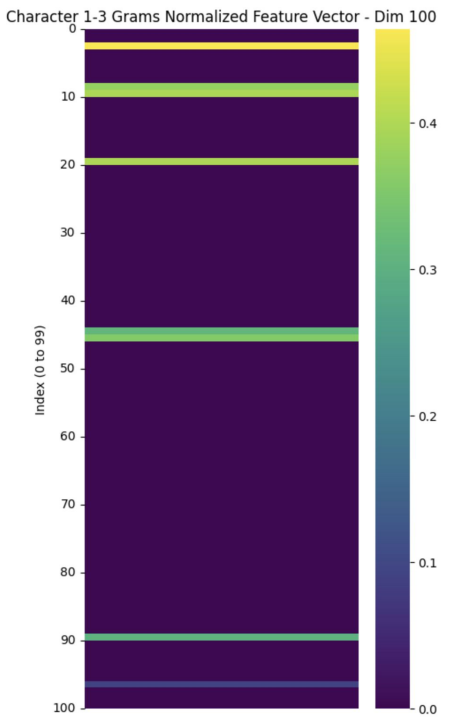
Feature vector of a single signature image



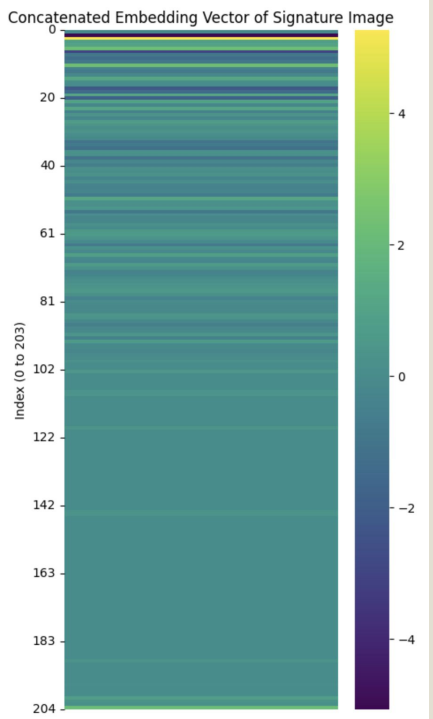
Apply PCA to reduce dim from 1024 to 100

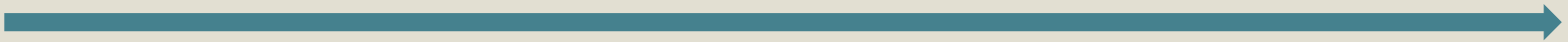


Apply OCR & generate 1-3 character grams



Concatenate 100 PCs + 1-3 n-grams + BB dimensions





DEMO

Clustering results: weighted average silhouette score of **0.680**

Score range: [-1, +1]

< 0.25: weak clustering
> 0.5: fair clustering
> 0.7: strong clustering
= 1: perfect clustering

State	Number of Applications	Number of Counties	Optimal Num of Clusters	Avg # of Notaries/County	Silhouette Score
Oklahoma	12024	96	1816	18.9	0.677
Georgia	4393	101	937	9.3	0.740
Tennessee	3281	92	798	8.7	0.625
North Carolina	3067	78	726	9.3	0.689
Missouri	1308	77	460	6.0	0.635
Alabama	1253	55	429	7.8	0.660
Arkansas	969	59	63	1.1	0.702
Texas	958	145	316	2.2	0.685
Virginia	722	33	22	0.7	0.636
Kansas	493	47	207	4.4	0.748
West Virginia	479	26	20	0.8	0.622
Indiana	378	32	37	1.2	0.748
Kentucky	259	49	140	2.9	0.712
California	177	31	68	2.2	0.695
Mississippi	158	28	55	2.0	0.634
Illinois	141	30	41	1.4	0.623
Florida	133	13	22	1.7	0.654
Colorado	94	24	48	2.0	0.777
South Carolina	92	14	8	0.6	0.660
New Mexico	66	16	25	1.6	0.710
Washington	63	17	30	1.8	0.639
Louisiana	47	15	22	1.5	0.670
Wisconsin	45	14	23	1.6	0.742
Ohio	41	17	22	1.3	0.630
Michigan	39	11	20	1.8	0.681
Iowa	36	13	17	1.3	0.633
Nebraska	25	11	12	1.1	0.751
Oregon	20	9	10	1.1	0.726
Idaho	20	8	10	1.3	0.765
Arizona	17	6	8	1.3	0.670
Pennsylvania	17	8	5	0.6	0.683
Montana	15	8	7	0.9	0.671

FUTURE IMPROVEMENTS

FUTURE IMPROVEMENT

- ➔ **Hybrid OCR Models**: ensemble model combining Tesseract OCR with AWS Textract & Keras-OCR with majority voting to generate more robust OCR
- ➔ **Improved Clustering**: experiment with HDBSCAN to handle varying cluster density distributions
- ➔ **Contrastive Learning**: perform human labeling of few clusters, then perform self-supervised contrastive learning by fine-tuning ConvNext on positive pairs (images inside same human cluster) and negative pairs (images from different human clusters) to make feature space more discriminative for signatures
- ➔ **Active Learning**: perform clustering with basic model then send low-confidence images (near cluster boundaries) to human oracle for manual clustering, then use new annotations to refine clustering model