

# BUILDING AN OCR PIPELINE TO DIGITIZE THE 1970 BLOCK HEADER RECORDS

Decennial Census Digitization and Linkage Project

U.S. Census Bureau

Kelsey Drotning and John Sullivan — Survey Statisticians

coding it forward >



CONNIE HONG

Stanford University  
Math and Computer Science

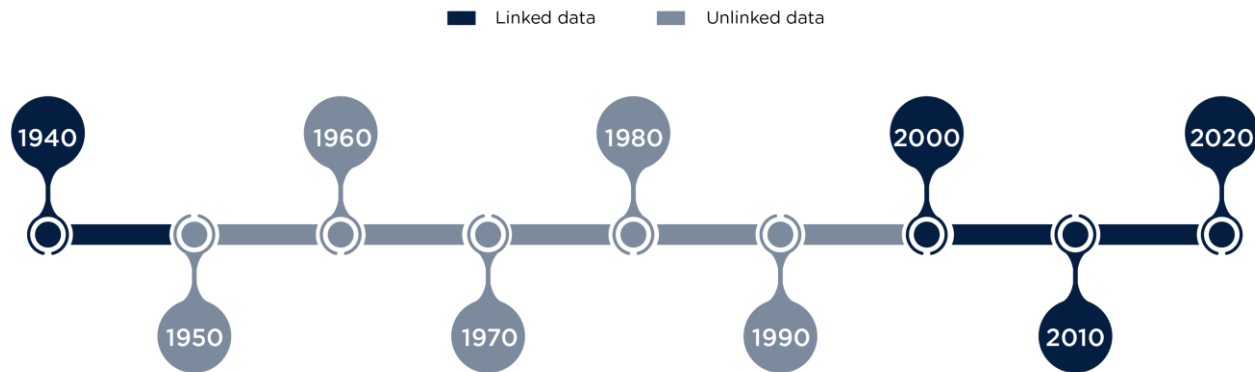
LARA KARACASU

Columbia University  
Computer Science

# Background

**Decennial Census Digitization Linkage Project:** Linking microdata files from the 1960 - 1990 decennial censuses. Will produce large longitudinal dataset to track behaviors across generations in the U.S. population from the 1940s to present-day.

Images for historical census data are limited to scanned TIFs. **Natural language analysis** and **OCR** technologies are needed.



# Goal: Develop OCR pipeline to generate digital data

1970s  
digital BHR  
not  
preserved...

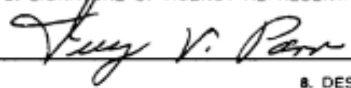


## 6. CERTIFICATE OF AGENCY REPRESENTATIVE

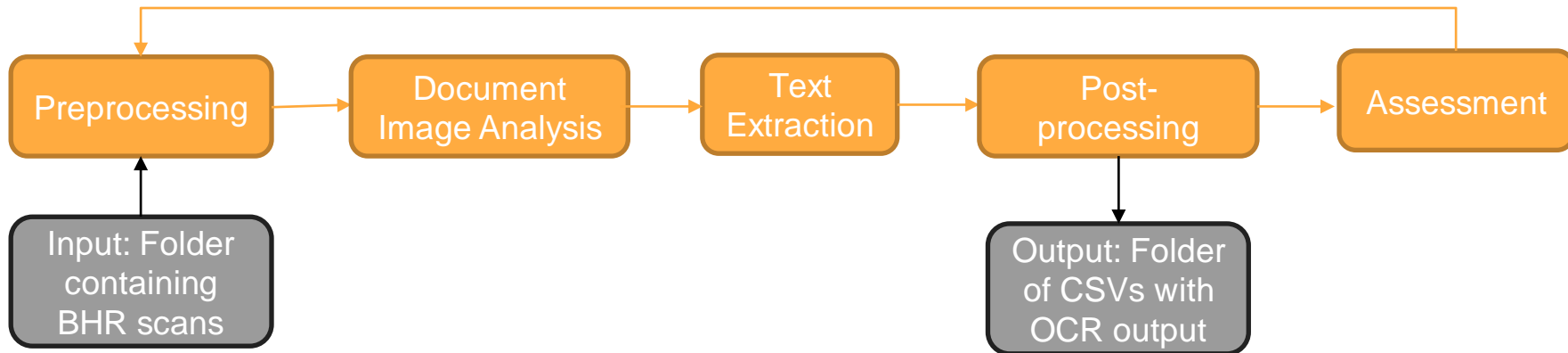
I hereby certify that I am authorized to act for this agency in matters pertaining to the disposal of the agency's records; that the records proposed for disposal in this Request of ~~2~~ 221 page(s) are not now needed for the business of this agency or will not be needed after the retention periods specified.

☒ **A Request for immediate disposal.**

☐ **B Request for disposal after a specified period of time or request for permanent retention.**

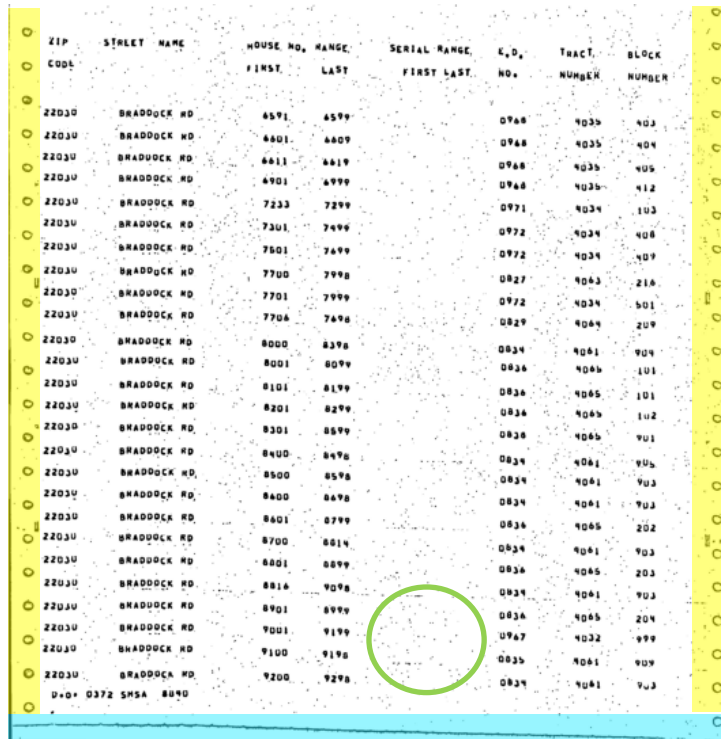
C. DATE	D. SIGNATURE OF AGENCY REPRESENTATIVE	E. TITLE	
7-10-81		Departmental Records Officer	
7. ITEM NO	8. DESCRIPTION OF ITEM (With Inclusive Dates or Retention Periods)	9. SAMPLE OR JOB NO	10. ACTION TAKEN
1.	The Geography Division wants to destroy the following microfilm records because they are no longer needed by the Division to carry out current programs.  1970 ADDRESS CODING GUIDES. 16mm Negative print-outs of block faces (sides of a city block) within SMSA's with each block face listing the street name, the low-high address number range, and geographic codes. This film is probably a COM copy of 1970 ADDRESS CODING GUIDE data also stored in computer tape files. 180 rolls		

# High Level Approach



# Preprocessing: Clean Images

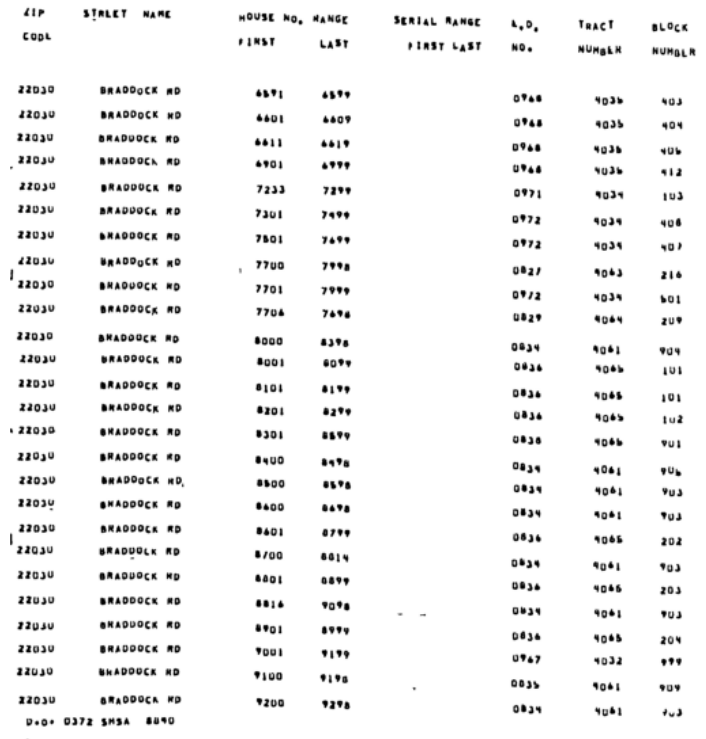
Methods: Binarization → noise reduction → de-skewing → line removal → binder hole removal → increasing DPI



The original image shows a data table with columns: ZIP CODE, STREET NAME, HOUSE NO. RANGE, SERIAL RANGE, L.D. NO., TRACT NUMBER, and BLOCK NUMBER. The table is heavily obscured by noise, including vertical lines and speckles. A green circle highlights a row in the lower half of the table.

ZIP CODE	STREET NAME	HOUSE NO. RANGE	SERIAL RANGE	L.D. NO.	TRACT NUMBER	BLOCK NUMBER
		FIRST LAST	FIRST LAST	NO.	NUMBER	NUMBER
22030	BRADDOCK RD	6591 6599		0768	9036	403
22030	BRADDOCK RD	6601 6609		0768	9035	404
22030	BRADDOCK RD	6611 6619		0768	9035	405
22030	BRADDOCK RD	6901 6999		0768	9035	412
22030	BRADDOCK RD	7233 7299		0971	9034	103
22030	BRADDOCK RD	7301 7499		0972	9034	408
22030	BRADDOCK RD	7501 7699		0972	9034	409
22030	BRADDOCK RD	7700 7998		0827	9043	216
22030	BRADDOCK RD	7701 7999		0972	9034	501
22030	BRADDOCK RD	7706 7698		0829	9044	209
22030	BRADDOCK RD	8000 8398		0834	9041	904
22030	BRADDOCK RD	8001 8099		0836	9045	101
22030	BRADDOCK RD	8101 8199		0836	9045	101
22030	BRADDOCK RD	8201 8299		0836	9045	102
22030	BRADDOCK RD	8301 8599		0836	9045	101
22030	BRADDOCK RD	8400 8498		0834	9041	905
22030	BRADDOCK RD	8500 8598		0834	9041	903
22030	BRADDOCK RD	8600 8698		0834	9041	903
22030	BRADDOCK RD	8601 8799		0836	9045	202
22030	BRADDOCK RD	8700 8814		0834	9041	903
22030	BRADDOCK RD	8801 8899		0836	9045	203
22030	BRADDOCK RD	8816 9098		0834	9041	903
22030	BRADDOCK RD	8901 8999		0836	9045	204
22030	BRADDOCK RD	9001 9199		0967	9032	999
22030	BRADDOCK RD	9100 9198		0835	9041	909
22030	BRADDOCK RD	9200 9298		0834	9041	903

Original Image




The preprocessed image shows the same data table as the original, but with all noise removed. The text is clear and legible, and the table structure is well-defined.

ZIP CODE	STREET NAME	HOUSE NO. RANGE	SERIAL RANGE	L.D. NO.	TRACT NUMBER	BLOCK NUMBER
		FIRST LAST	FIRST LAST	NO.	NUMBER	NUMBER
22030	BRADDOCK RD	6591 6599		0768	9036	403
22030	BRADDOCK RD	6601 6609		0768	9035	404
22030	BRADDOCK RD	6611 6619		0768	9035	405
22030	BRADDOCK RD	6901 6999		0768	9035	412
22030	BRADDOCK RD	7233 7299		0971	9034	103
22030	BRADDOCK RD	7301 7499		0972	9034	408
22030	BRADDOCK RD	7501 7699		0972	9034	409
22030	BRADDOCK RD	7700 7998		0827	9043	216
22030	BRADDOCK RD	7701 7999		0972	9034	501
22030	BRADDOCK RD	7706 7698		0829	9044	209
22030	BRADDOCK RD	8000 8398		0834	9041	904
22030	BRADDOCK RD	8001 8099		0836	9045	101
22030	BRADDOCK RD	8101 8199		0836	9045	101
22030	BRADDOCK RD	8201 8299		0836	9045	102
22030	BRADDOCK RD	8301 8599		0836	9045	101
22030	BRADDOCK RD	8400 8498		0834	9041	905
22030	BRADDOCK RD	8500 8598		0834	9041	903
22030	BRADDOCK RD	8600 8698		0834	9041	903
22030	BRADDOCK RD	8601 8799		0836	9045	202
22030	BRADDOCK RD	8700 8814		0834	9041	903
22030	BRADDOCK RD	8801 8899		0836	9045	203
22030	BRADDOCK RD	8816 9098		0834	9041	903
22030	BRADDOCK RD	8901 8999		0836	9045	204
22030	BRADDOCK RD	9001 9199		0967	9032	999
22030	BRADDOCK RD	9100 9198		0835	9041	909
22030	BRADDOCK RD	9200 9298		0834	9041	903

Preprocessed Image

# Document Image Analysis: Detect image layout

**Methods:** Annotating 50 scans with Label Studio, splitting up train-test data, training layout detection model.



ZIP CODE	STREET NAME	HOUSE NO. RANGE		SERIAL RANGE		E.O. NO.	TRACT NUMBER	BLOCK NUMBER
		FIRST	LAST	FIRST	LAST			
22041	ARNETT ST	5400	5422		0897	4053	104	
22041	ARNETT ST	5401	5423		0897	4053	104	
22041	ARNETT ST	5424	5498		0898	4053	104	
22041	ARNETT ST	5425	5499		0898	4053	104	
22041	ASHWOOD PLACE	3900	3998		0904	4052	999	
22041	ASHWOOD PLACE	3901	3999		0904	4052	999	
22041	AURA COURT	4100	4198		0902	4051	102	
22041	AURA COURT	4101	4199		0902	4051	102	
22041	BANCROFT TRL	3600	3698		0899	4053	999	
22041	BANCROFT TRL	3601	3699		0899	4053	999	
22041	BAY TREE CT	6500	6598		0908	4054	205	
22041	BAY TREE CT	6501	6599		0908	4054	205	
22041	BAY TREE LANE	6600	6698		0908	4052	205	
22041	BAY TREE LANE	6601	6699		0904	4052	210	
22041	BEACHWAY DR	4100	4126		0904	4051	405	
22041	BEACHWAY DR	4101	4126		0904	4051	403	
22041	BEACHWAY DR	4127	4191		0904	4051	404	
22041	BEACHWAY DR	4128	4198		0902	4051	104	
22041	BEACHWAY DR	4193	4207		0904	4051	405	
22041	BEACHWAY DR	4200	4210		0902	4051	104	
22041	BEACHWAY DR	4209	4213		0904	4051	405	
22041	BEACHWAY DR	4212	4298		0897	4057	24	
22041	BEACHWAY DR	4215	4299		0904	4051	407	
22041	BELLEVIEW DR	4000	4050		0903	4051	207	
22041	BELLEVIEW DR	4001	4049		0903	4051	207	
22041	BELLEVIEW DR	4051	4099		0903	4051	204	

U.S. 0372 SMSA 8640

ZIP CODE	STREET NAME	HOUSE NO. RANGE		SERIAL RANGE		E.O. NO.	TRACT NUMBER	BLOCK NUMBER
		FIRST	LAST	FIRST	LAST			
2746	W ALBANY ST	200	210		5735	11200	311	
2746	W ALBANY ST	201	219		5735	11200	316	
2746	ALBERMARLE AVE	2	18		5720	1114	105	
2746	ALBERMARLE AVE	2	18		5720	1114	104	
2746	ALBERMARLE AVE	2	20		5721	1114	301	
2746	ALBERMARLE AVE	3	7		5720	1114	109	
2746	ALBERMARLE AVE	9	27		5720	1114	106	
2746	ALBERMARLE AVE	11	55		5721	1114	210	
2746	ALBERMARLE AVE	15	215		5720	1114	101	
2746	ALBERMARLE AVE	19	71		5720	1114	101	
2746	ALBERMARLE AVE	20	48		5720	1114	106	
2746	ALBERMARLE AVE	26	40		5721	1114	303	
2746	ALBERMARLE AVE	46	46		5721	1114	208	
2746	ALBERMARLE AVE	48	48		5721	1114	302	
2746	ALBERMARLE AVE	49	55		5721	1114	209	
2746	ALBERMARLE AVE	50	70		5720	1114	107	
2746	ALBERMARLE AVE	60	64		5721	1114	302	
2746	ALBERMARLE AVE	66	76		5721	1114	303	
2746	ALBERMARLE AVE	71	79		5721	1114	304	
2746	ALBERMARLE AVE	81	93		5721	1114	304	
2746	ALBERMARLE AVE	161	177		5720	1114	107	
2746	ALBERMARLE AVE	162	214		5720	1114	101	
2746	ALBERMARLE AVE	254	258		5721	1114	208	
2746	ALDERFIELD LANE	1	5		5727	11220	214	
2746	ALDERFIELD LANE	2	22		5727	11220	215	
2746	ALDERFIELD LANE	2	57		5727	11220	215	

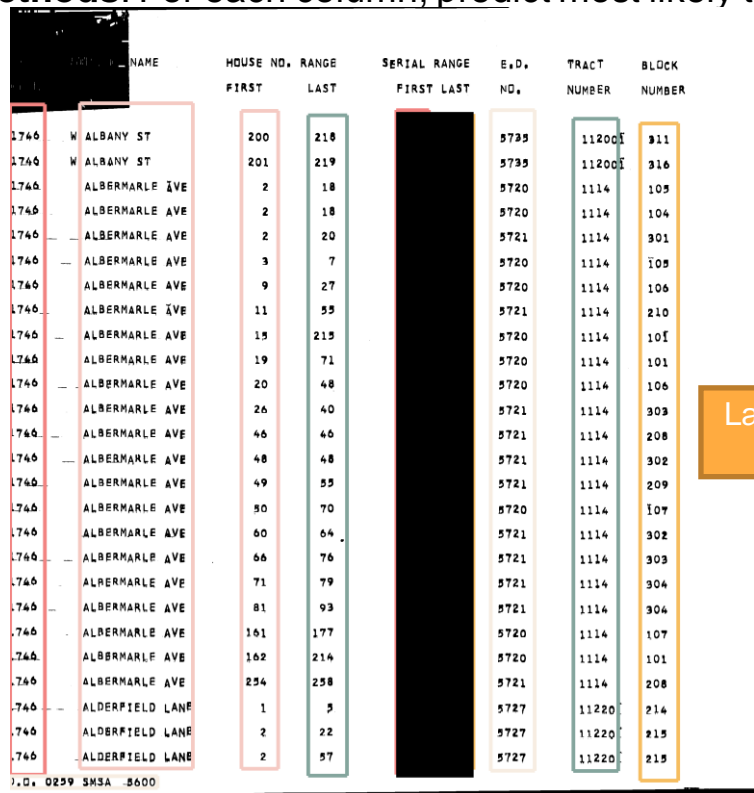
U.S. 0259 SMSA 8600

Label Studio Annotations

Detectron2 Model Detection

# Text Extraction: Tesseract OCR Engine

**Methods:** For each column, predict most likely text and transform raw output into readable data frame.



	NAME	HOUSE NO. RANGE		SERIAL RANGE	E.D. NO.	TRACT NUMBER	BLOCK NUMBER
		FIRST	LAST	FIRST LAST			
1746	W ALBANY ST	200	218		5735	11200	311
1746	W ALBANY ST	201	219		5735	11200	316
1746	ALBERMARLE AVE	2	18		5720	1114	105
1746	ALBERMARLE AVE	2	18		5720	1114	104
1746	ALBERMARLE AVE	2	20		5721	1114	301
1746	ALBERMARLE AVE	3	7		5720	1114	109
1746	ALBERMARLE AVE	9	27		5720	1114	106
1746	ALBERMARLE AVE	11	55		5721	1114	210
1746	ALBERMARLE AVE	15	215		5720	1114	101
1746	ALBERMARLE AVE	19	71		5720	1114	101
1746	ALBERMARLE AVE	20	48		5720	1114	106
1746	ALBERMARLE AVE	26	40		5721	1114	303
1746	ALBERMARLE AVE	46	46		5721	1114	208
1746	ALBERMARLE AVE	48	48		5721	1114	302
1746	ALBERMARLE AVE	49	55		5721	1114	209
1746	ALBERMARLE AVE	50	70		5720	1114	107
1746	ALBERMARLE AVE	60	64		5721	1114	302
1746	ALBERMARLE AVE	66	76		5721	1114	303
1746	ALBERMARLE AVE	71	79		5721	1114	304
1746	ALBERMARLE AVE	81	93		5721	1114	304
1746	ALBERMARLE AVE	161	177		5720	1114	107
1746	ALBERMARLE AVE	162	214		5720	1114	101
1746	ALBERMARLE AVE	254	258		5721	1114	208
1746	ALDERFIELD LANE	1	5		5727	11220	214
1746	ALDERFIELD LANE	2	22		5727	11220	215
1746	ALDERFIELD LANE	2	57		5727	11220	215

Layout Parser  
Interface

	zip	street_name	hn_first	hn_last	serial_first	serial_last	ed	tract	block
0	1746	ILE	200	218			5735	11200	311
1	1744	ALBANY ST	201	219			5735	112001	316
2	1746	ALBANY ST	11				53720	1114	109
3	1746	ALBERMARLE AVE	13	NaN			5720	1114	104
4	1746	ALBERMARLE AVE	19	NaN			57231	1114	301
5	1746	ALBERMARLE AVE	20	NaN			5720	1114	108
6	1746	ALBERMARLE AVE	26	NaN			5720	1114	106
7	1746	ALBERMARLE AVE	46	NaN			5721	1114	210
8	1746	ALBERMARLE AVE	48	NaN			5720	1114	10
9	1746	ALBERMARLE AVE	49	NaN			5720	1114	101
10	1746	ALBERMARLE AVE	50	NaN			53720	1114	106
11	1746	ALBERMARLE AVE	60	NaN			5721	1114	303
12	1766	ALBERMARLE AVE	66	NaN			5721	1114	208
13	1746	ALBERMARLE AVE	71	NaN			5721	1114	302
14	1766	ALBERMARLE AVE	81	NaN			5721	1114	209
15	1744	ALBERMARLE AVE	161	NaN			8720	1114	7
16	1746	ALBERMARLE AVE	162	NaN			5721	1114	302
17	17646	ALBERMARLE AVE	34	NaN			5721	1114	303
18	1746	ALBERMARLE AVE		NaN			3721	1114	304
19	1746	ALBERMARLE AVE	NaN	NaN			8721	1114	107
20	746	ALBERMARLE AVE	NaN	NaN			5720	1114	101
21	1746	ALBERMARLE AVE	NaN	NaN			3720	1114	208
22	1746	ALBERMARLE AVE	NaN	NaN			5721	1114	214
23	1746	ALBERMARLE AVE	NaN	NaN			5727	112201	2153
24	1746	ALDERFIELD LANE	NaN	NaN			5727	112207	218
25	7406	ALDERFIELD LANE	NaN	NaN			5727	112201	

Detectron2 Model Detection

OCR Output

# Postprocessing: Correct OCR Errors

- **Methods:** Map true information [City, State, SMSA] to output columns [Zipcode, Street Name, etc.]
- Based on **Levenshtein distance & neighbor similarity**.
- Techniques differ by column.





# Future Endeavors

## Conclusion:

- Built a bash script that executes Python files for each step.

## Next steps: Improving accuracy. How?

- Training Tesseract OCR engine on custom 1970s-BHR data.
- Improving layout detection accuracy by expanding training-testing dataset.



# THANK YOU!

Any questions?