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 >

United States Census Bureau

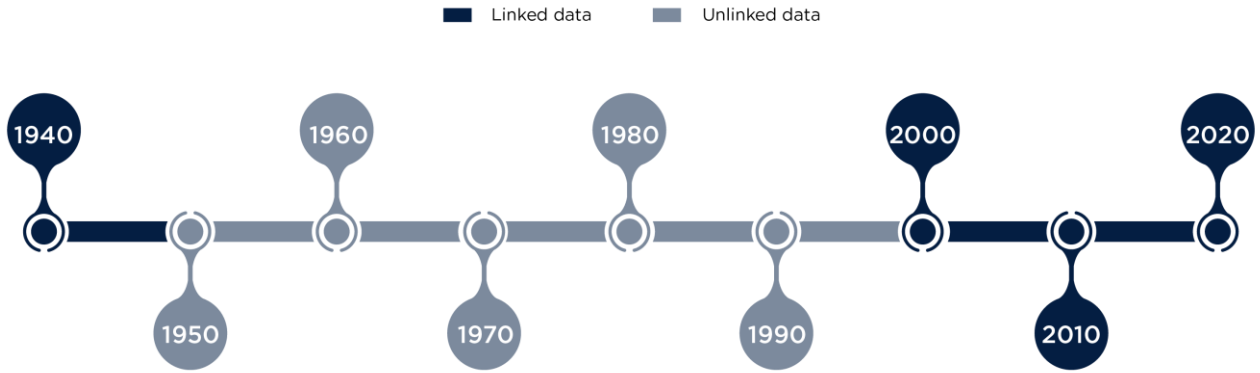
CONNIE HONG
Stanford University
Math and Computer Science

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.



Any opinions and conclusions expressed herein are those of the author(s) and do not reflect the views of the U.S. Census Bureau.

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 page(s) are not now needed for the business of this agency or will not be needed after the retention periods specified.

[X] A Request for immediate disposal.

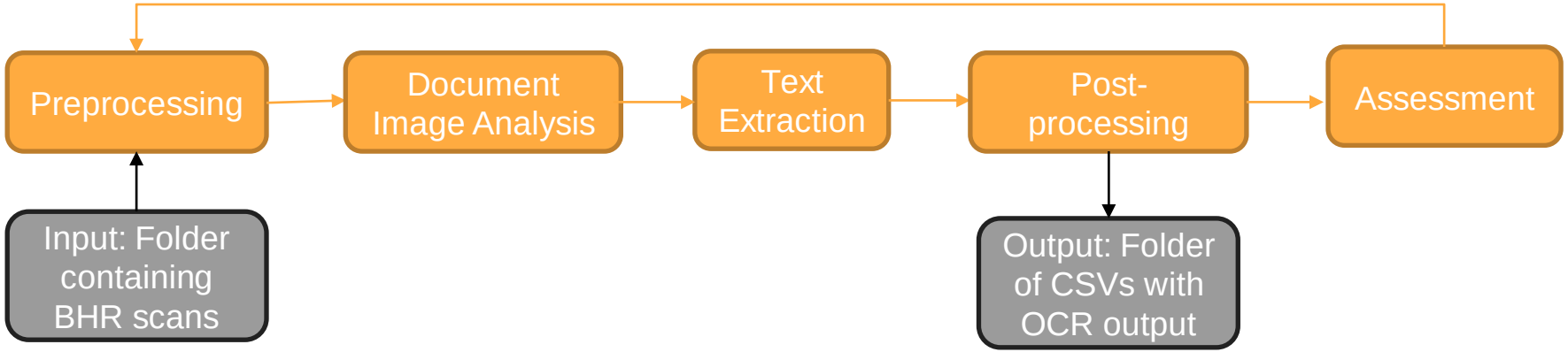
[ ] B Request for disposal after a specified period of time or request for permanent retention.

C. DATE	D. SIGNATURE OF AGENCY REPRESENTATIVE	E. TITLE
7-10-81	Guy V. Pa[illegible]	Departmental Records Officer

7. ITEM NO	8. DESCRIPTION OF ITEM (With Inclusive Dates or Retention Periods)	9. SAMPLE OR JOB NO	10. ACTION TAKEN
	The Geography Division wants to destroy the following microfilm records because they are no longer needed by the Division to carry out current programs.		
1.	1970 ADDRESS CODING GUIDES. 16mm Negative printouts 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		

Any opinions and conclusions expressed herein are those of the author(s) and do not reflect the views of the U.S. Census Bureau.

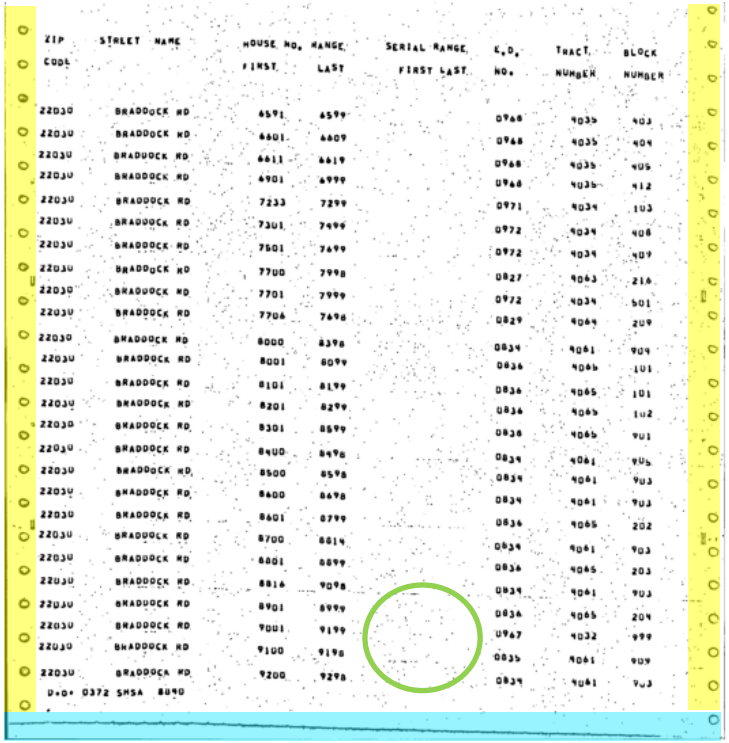
High Level Approach



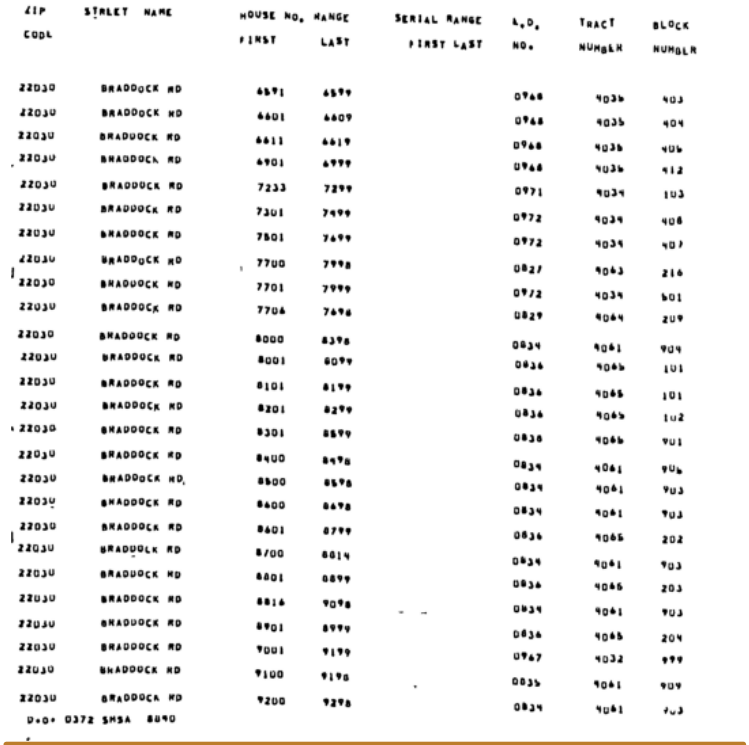
Any opinions and conclusions expressed herein are those of the author(s) and do not reflect the views of the U.S. Census Bureau.

Preprocessing: Clean Images

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



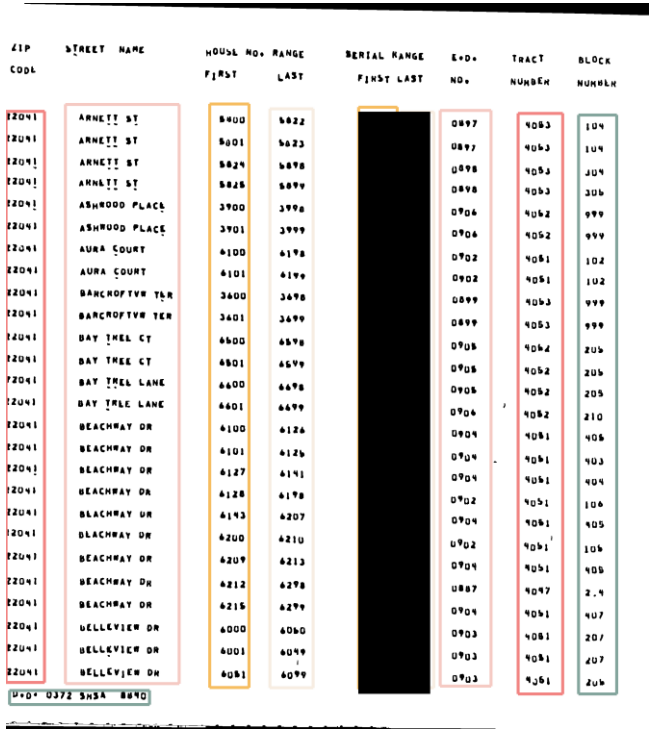
Original Image



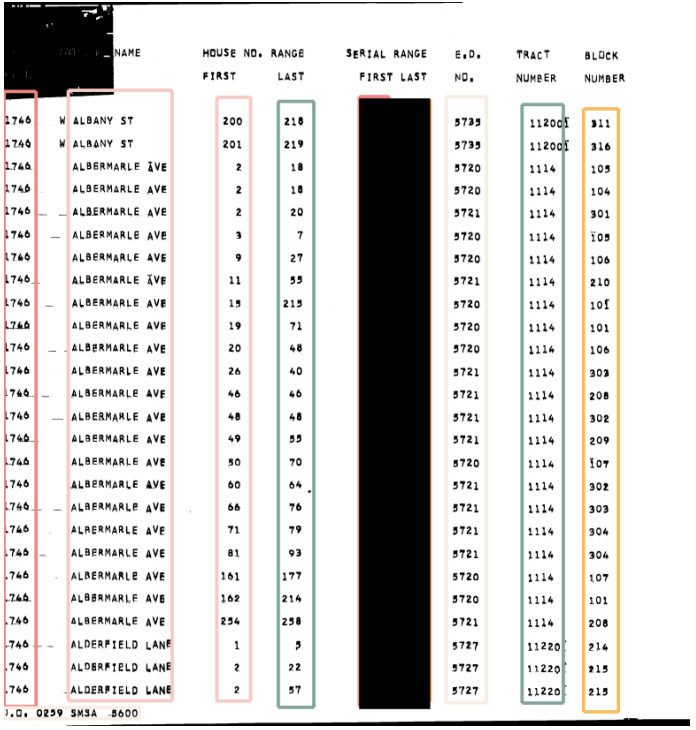
Preprocessed Image

Document Image Analysis: Detect image layout

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



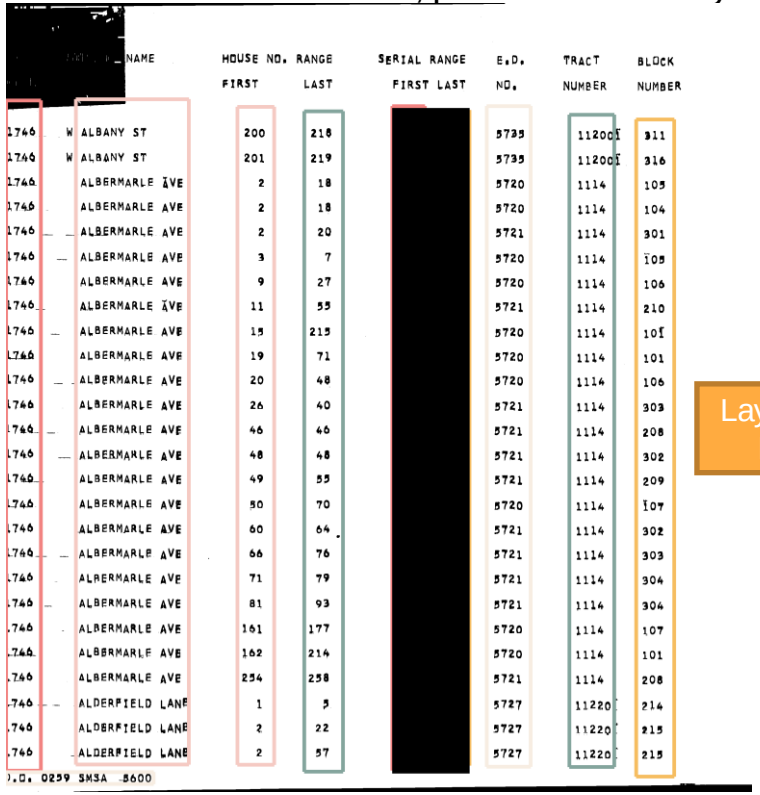
Label Studio Annotations



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 FIRST	LAST	SERIAL RANGE FIRST LAST	E.D. NO.	TRACT NUMBER	BLOCK NUMBER
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	105
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 ZVE	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	ALRERMARLE 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.

Example filename:
ALEXANDRA_VA_DO_0372_VOL_1_OF_2_SMSA_8840_A01.tif

→

smsaToStreet = {…, 8840 : {'ABC STREET' 'DEF LANE'}, …}

→

'ABD STREET' → 'ABC STREET'

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?

coding it forward >

United States Census Bureau

CONNIE HONG
Stanford University
Math and Computer Science

LARA KARACASU
Columbia University
Computer Science