Enhancing the Precision of Vast Historical Datasets via Deep Learning and AI: A Case Study on Mitigating Misinterpretation Challenges within the 1950 U.S. Census

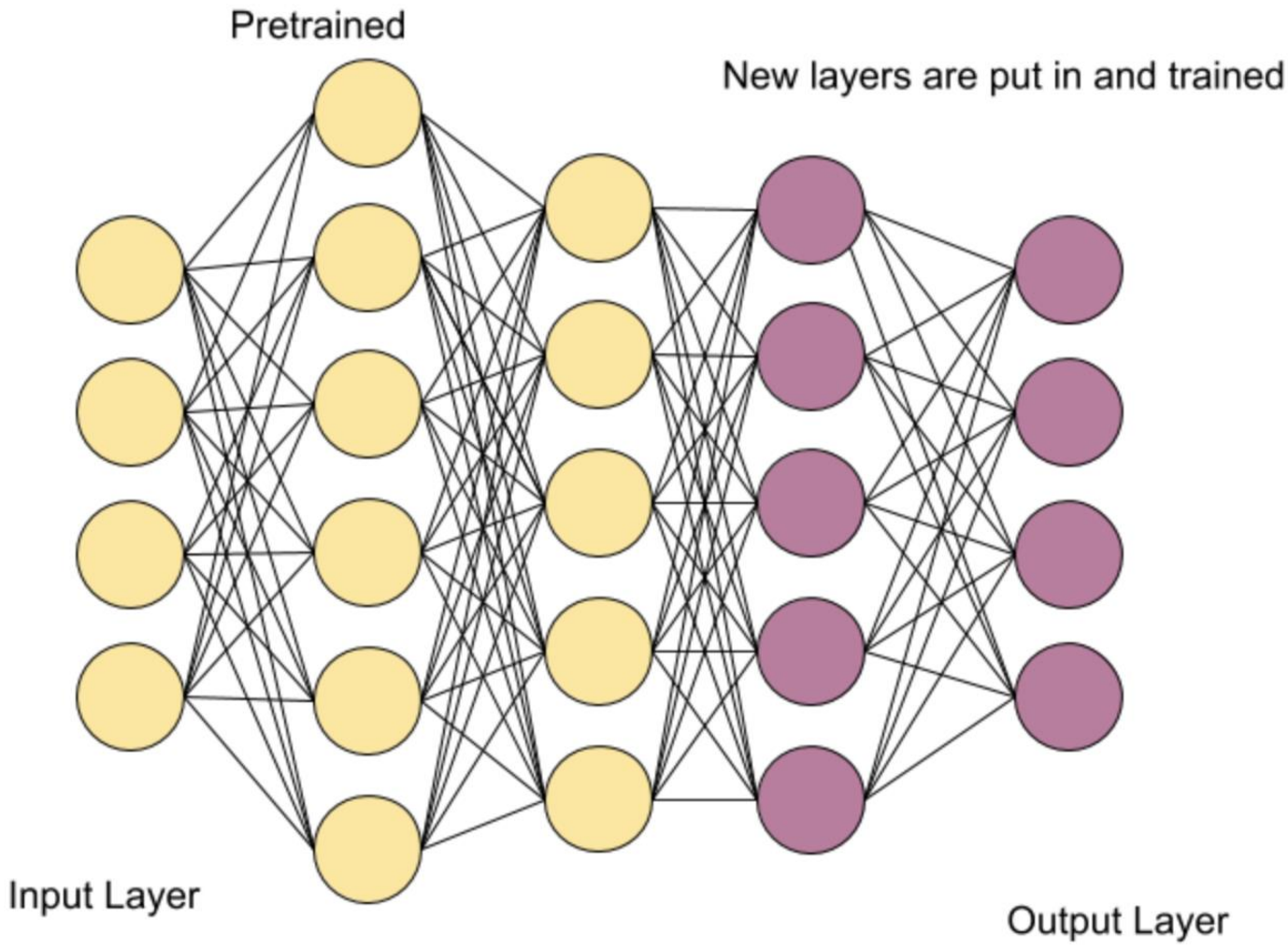
Mengyue Zhao (University of Oxford)

beczhaozmy0@gmail.com

Agenda

- What is deep-learning & transfer learning?
- Deep-Learning Frameworks
  - Machine Learning Iteration Process
  - How robust is this approach?
- Use Cases:
  - Transcribing the Official Register of the United States
  - Reduce 1950 Census Overcounts
- The Importance Page Segmentation
- The role of Generative-AI in Fine-tuning tasks
  - Can it replace traditional ML?

Transfer Learning



The 1950 census data

•an **overcount** of approximately **two million individuals**.

•Source of Error: OCR mistakes

- “No One” -> “Owen” ;
- “Not at Home” -> “Norton”

•“Strike-through” Indicators

- Deep-Learning Solution
  - Page Layout Analysis
  - Image Classification

b. COUNTY SAN MATEO
c. INCORPORATED PLACE OR TOWNSHIP BURLINGAME CITY
d. E. D. NUMBER 41-20

Name / Type / Line Nos. ... to ... inclusive

CONFIDENTIAL This inquiry is authorized by Act of Congress (Title 13, U.S. Code). The information furnished is accorded confidential treatment. The Census report cannot be used for purposes of taxation, investigation, or regulation.

FORM P1

U. S. DEPARTMENT OF COMMERCE
BUREAU OF THE CENSUS
1950 CENSUS OF POPULATION AND HOUSING

e. ENUMERATOR'S SIGNATURE Marjorie L. Callow
b. CHECKED BY

NUMBER 25

Notes 1/ not known for sure.
1/ see S83-L 3

LINE NUMBER	Name of street, avenue, or road	FOR HEAD OF HOUSEHOLD: House (and apartment) number	Serial number of dwelling unit	Is this house on a farm (or ranch)? (Yes or No)	If No in item 4— Is this house on a place of three or more acres? (Yes or No)	Agriculture Questionnaire Number	FOR ALL PERSONS: NAME — What is the name of the head of this household? What are the names of all other persons who live here? List in this order: The head; His wife; Unmarried sons and daughters (in order of age); Married sons and daughters and their families; Other relatives; Other persons, such as lodgers, roomers, maids or hired hands who live in, and their relatives. (Last name first)	RELATIONSHIP — Enter relationship of person to head of the household, as Head, Wife, Daughter, Grandson, Mother-in-law, Lodger, Lodger's wife, Maid, Hired hand, Patient, etc.	RACE — White (W), Negro (Neg), American Indian (Ind), Japanese (Jap), Chinese (Chi), Filipino (Fil), Other race—spell out	LEAVE BLANK	SEX — Male (M), Female (F)	How old was he on his last birthday? (If under one year of age, enter month of birth as April, May, Dec., etc.)	Is he now married, widowed, divorced, separated, or never married? (Mar., Wd., D., Sep., Nev.)	What State (or foreign country) was he born in? If born outside Continental United States, enter name of Territory, possession, or foreign country. Distinguish Canada-French from Canada-other	If foreign born— Is he naturalized? (Yes, No, or AP for born abroad of American parents)	LEAVE BLANK	FOR PERSONS 14 YEARS OF AGE AND OVER: What was this person doing most of last week—working, keeping house, or something else? (Wk, H, O, or U for unable to work)	If H or O in item 15 — Did this person do any work at all last week, not counting work around the house? (Include work for pay, in own business, profession, on farm, or unpaid family work) (Yes or No)	If No in item 16 — Was this person looking for work? (See Special Cases below) (Yes or No)	If No in item 17 — Even though he didn't work last week, does he have a job or business? (Yes or No)	If Wk in item 15 or Yes in item 16 — How many hours did he work last week? (Include unpaid work on family farm or business) (Number of hours)	What kind of work was he doing? (Occupation)	What kind of business or industry was he working in? (Industry)	Class of worker (For PRIVATE employer (P), For GOVERNMENT (G), In OWN business (O), WITHOUT PAY on family farm or business (NP)) (P, G, O, or NP)	LEAVE BLANK	LINE NUMBER
1							—, LORRAINE K.	daughter	W	F	1	new	Calif.																1
(2)							GEORGE, CLORINDA															48	housework	Private home	P	720826 1		(2)	2
3		1473	290				not at home see sheet																						3
4		1467	291				not at home see sheet																						4
5		1465	292	no	no		DODD, JOHN M.	head	W	M	53	mar	New York	021		WK				40	district manager in charge of sales	soft drink bottling Co	P	290 418 1			5
6							—, FLORENCE A.	wife	W	F	50	mar	New York	021		H	no	no	no									6
(7)							—, PATRICIA	daughter	W	F	11	new	New York	021														(7)	7
8		~~1460~~ 1455	293				not at home see sheet 77 lines		12,13,14,15,16,17,18,19,20,					21															8
9		1458	294				not at home see sheet 77 lines		11 —																				9
10		~~1449~~	295	no	no		~~not at home see sheet~~ BIONDI, DINO H.	head	W	M	40	mar	Pennsylvania	023		OT	no	yes			automotive mechanic	automobile service and sales Co.	P	550 667 1			10
11							—, IRENE P	wife	W	F	36	mar	Calif.			H	no	no	no									11
(12)							—, J. CHRISTIAN	son	W	M	7	new	Calif.															(12)	12
13		1445	296	no	no		EMPEY, WALTER T.	head	W	M	39	mar	Calif.			OT	no	no	yes		PBX installer	Telephone Co	P	540 578 1			13
14							—, ERMA F.	wife	W	F	33	mar	Calif.			H	no	no	no									14
15							—, DARRELL R.	son	W	M	1	new	Calif.																15
16		1441	297	no	no		MARSH, DAVID G.	head	W	M	68	mar	West Virginia	055		OT	no	no	no									16
(17)							—, ANNA C.	wife	W	F	62	mar	Maryland	052		H	no	no	no									(17)	17
18							—, ROBERT D.	son	W	M	34	new	Pennsylvania	023		WK				40	plumbing salesman	wholesale plumbing supplies	P	490 616 1			18
19		1437	298	no	no		—VACANT—																						19
1/ 20		1433	299	no	no		GALITZ, ~~HERBERT~~ EVELYN M.	~~head~~ 2	W	F	41	mar	Illinois	033		H	no	no	no									20
21							—, ~~ETHEL~~ RICHARD L.	~~wife~~ son 3	W	M	13	new	Illinois	033															21
(22)							—, GORDON W.	son	W	M	5	new	Calif.															(22)	22
23							—, KEITH G.	son	W	M	5	new	Calif.																23
24		1429	300	no	no		—VACANT—																						24
25		1425	301	no	no		HOWELL, DELPHUS D.	head								OT	no	no	no									25

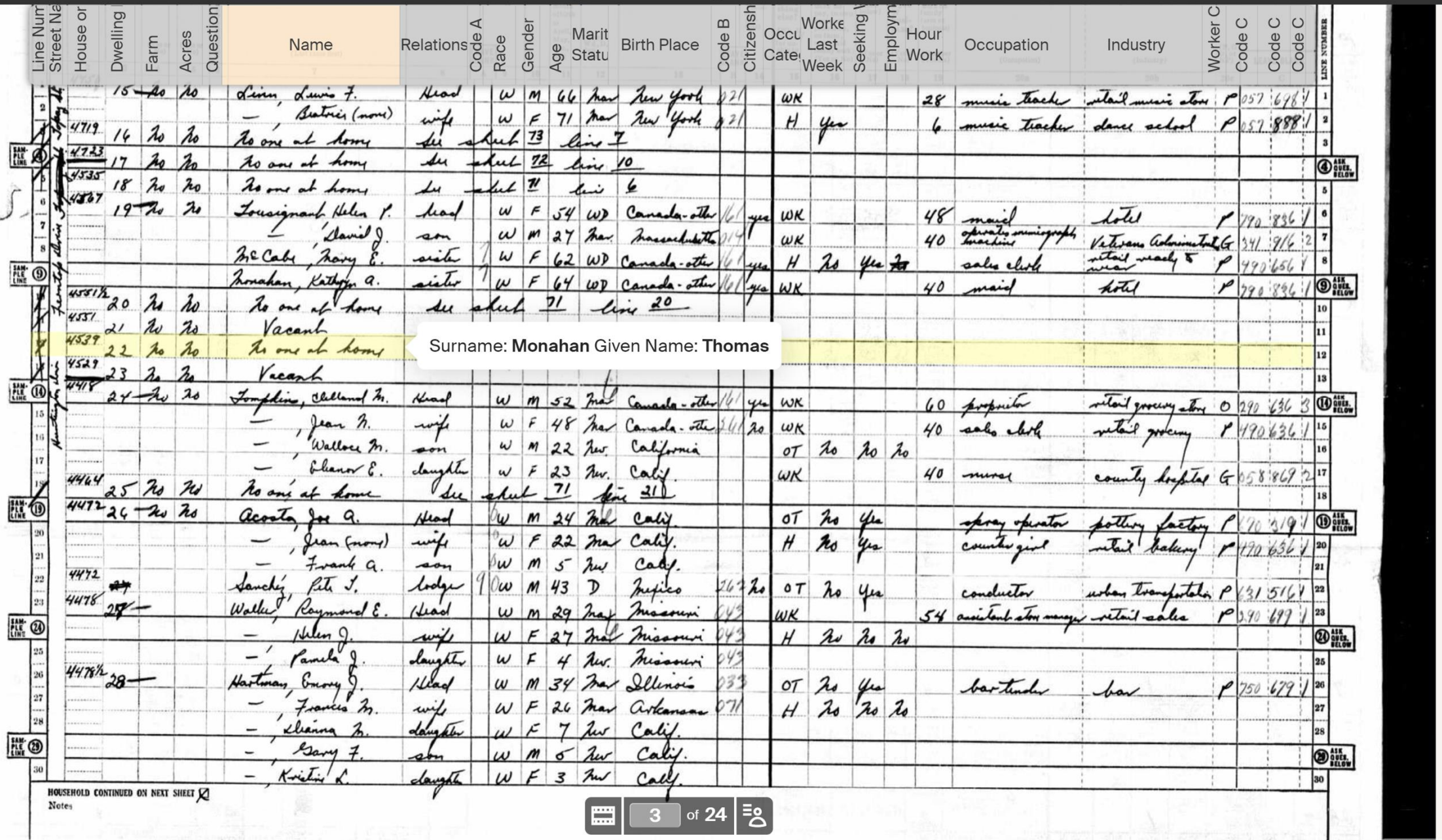
Cabrillo Avenue
3rd 4-17

Surname: **George** Given Name: **Scheel**

41-20

LINE NUMBER	Name of street, avenue, or road	House (and apartment) number	Serial number of dwelling unit	Is this house on a farm (or ranch)? (Yes or No)	If No in item 4 — Is this house on a place of three or more acres? (Yes or No)	Agriculture Questionnaire Number	NAME — What is the name of the head of this household? What are the names of all other persons who live here? List in this order: The head; His wife; Unmarried sons and daughters (in order of age); Married sons and daughters and their families; Other relatives; Other persons, such as lodgers, roomers, maids or hired hands who live in, and their relatives. (Last name first)	RELATIONSHIP — Enter relationship of person to head of the household, as Head, Wife, Daughter, Grandson, Mother-in-law, Lodger, Lodger's wife, Maid, Hired hand, Patient, etc.	LEAVE BLANK	RACE — White (W), Negro (Neg), American Indian (Ind), Japanese (Jap), Chinese (Chi), Filipino (Fil), Other race — spell out	SEX — Male (M), Female (F)	How old was he on his last birthday? (If under one year of age, enter month of birth as April, May, Dec., etc.)	Is he now married, widowed, divorced, separated, or never married? (Mar, Wd, D, Sep, Nev)	What State (or foreign country) was he born in? If born outside Continental United States, enter name of Territory, possession, or foreign country. Distinguish Canada-French from Canada-other	LEAVE BLANK	If foreign born — Is he naturalized? (Yes, No, or AF for born abroad of American parents)	What was this person doing most of last week — working, keeping house, or something else? (Wk, If, Ot, or U for unable to work)	If H or Ot in item 15 — Did this person do any work at all last week, not counting work around the house? (Include work for pay, in own business, profession, on farm, or unpaid family work) (Yes or No)	If No in item 16 — Was this person looking for work? (See Special Cases below) (Yes or No)	If No in item 17 — Even though he didn't work last week, does he have a job or business? (Yes or No)	If Wk in item 15 or Yes in item 16 — How many hours did he work last week? (Include unpaid work on family farm or business) (Number of hours)	FOR PERSONS 14 YEARS OF AGE AND OVER — 1. If employed (Wk in item 15, or Yes in item 16 c... 2. If looking for work (Yes in item 17), describe la... 3. For all other persons, leave blank. What kind of work was he doing? For example: Nails heels on shoes; Chemistry professor; Farmer; Farm helper; Armed forces; Never worked. (Occupation)	20a
1							—, LESLEE ANN	daughter		W	F	mo		Calif.												
2		1407	303				not at home see																			
3		1405	304 ~~no~~	no	no		BEGGS, WILLIAM H.	head		W	M	63	mar	Northern Ireland		13	yes	WK					32 ~~72~~	Carpenter	
4							—, ELIZABETH L.	wife		W	F	65	mar	Northern Ireland		13	yes	H	no	no	no				
5		1800	305 ~~no~~	no	no		CUTSHAW, ROBERT P.	head		W	M	35	mar	North Carolina		056		WK					65	Contractor division sales man	
6							—, LUCY M.	wife		W	F	34	mar	Tennessee		062		H	no	no	no				
7							—, ROBERT ANDREW	son		W	M	4	nev	Tennessee		062		~~H~~							
8		1804	~~306~~ no	no	no		RUSCH, CLARENCE E.	head		W	M	40	mar	Wisconsin		035		WK					60	Proprietor bookkeeping	
9							—, ELINOR M.	wife		W	F	32	mar	Calif.				H	no	no	yes				
10							—, JUDITH A	daughter		W	F	9	nev	Calif.											
11							—, SELMA S.	daughter		W	F	7	nev	Calif.											
12							—, MARTHA E.	daughter		W	F	4	nev	Calif.											
13							—, EMIL	father	6	W	M	64	wd	Wisconsin		035		WK					60	repair man	
14		1810	307	not at home see sheet																						
15		1406	308	not at home see sheet 79										lines 12, 13, 14, 15, 16 / lines 17, 18, 19, 20											

Surname: **Howell** Given Name: **Lome See**



Surname: **Monahan** Given Name: **Thomas**

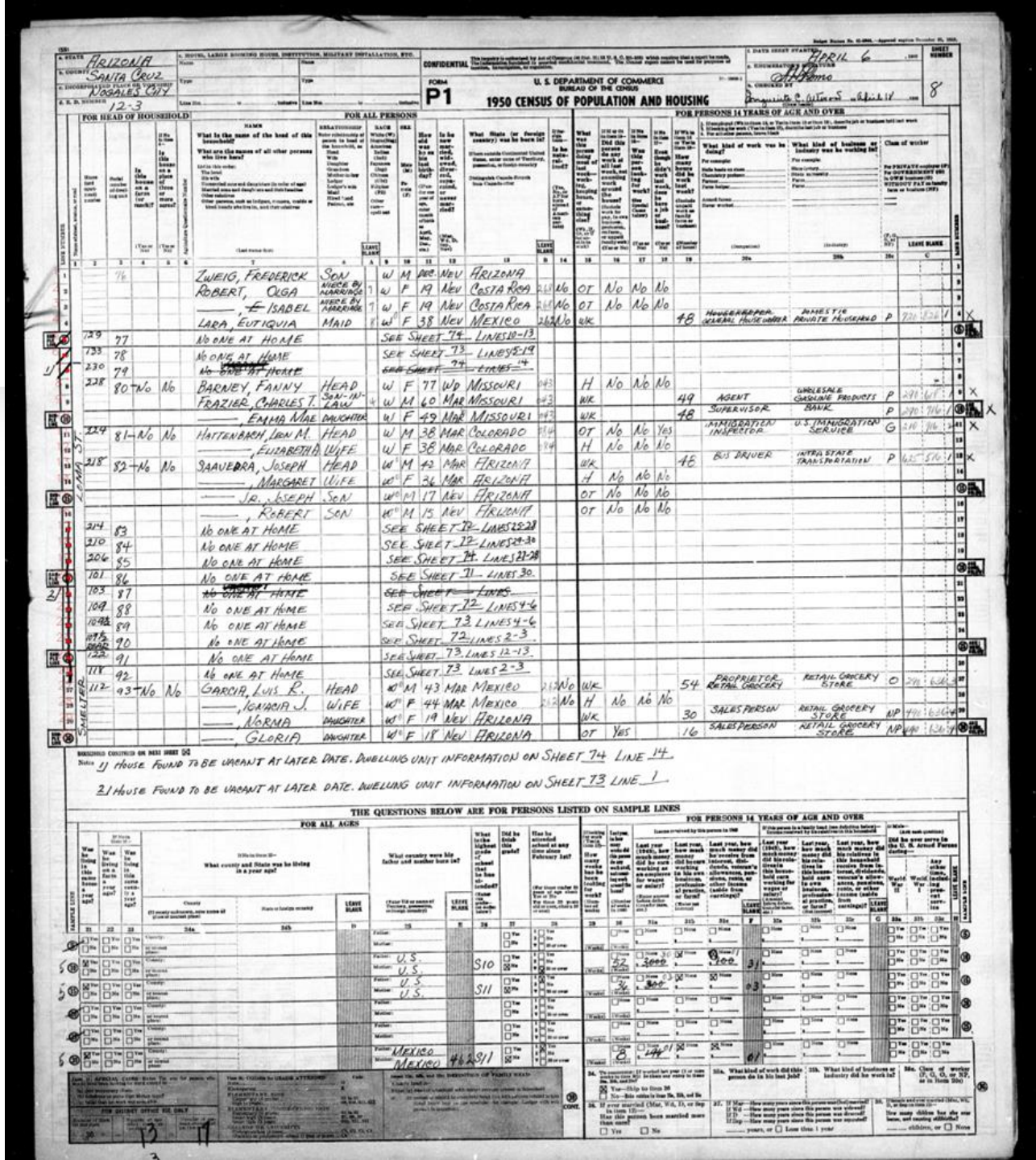
Line Num / Street Na	House or	Dwelling	Farm	Acres	Question	Name	Relations	Code A	Race	Gender	Age	Marit Statu	Birth Place	Code B	Citizensh	Occu Cate	Worke Last Week	Seeking W	Employm	Hour Work	Occupation	Industry	Worker C	Code C	Code C	Code C	LINE NUMBER
1		15	No	No		Linn, Lewis F.	Head		W	M	66	Mar	New York	021		WK				28	music teacher	retail music store	P	057	678	4	1
2						—, Beatrice (mrs)	wife		W	F	71	Mar	New York	021		H	Yes			6	music teacher	dance school	P	057	888	4	2
3	4719	16	No	No		No one at home																					3
4	4723	17	No	No		No one at home																					4
5	4535	18	No	No		No one at home																					5
6	4567	19	No	No		Lousignant, Helen P.	Head		W	F	54	WD	Canada-other	161	yes	WK				48	maid	hotel	P	770	836	1	6
7						—, David J.	son		W	M	27	Mar	Massachusetts	014		WK				40	operates mimeograph machine	Veterans Administration	G	341	916	2	7
8						McCabe, Mary E.	sister		W	F	62	WD	Canada-other	161	yes	H	No	Yes	No		sales clerk	retail ready to wear	P	490	656	1	8
9						Monahan, Kathryn A.	sister		W	F	64	WD	Canada-other	161	yes	WK				40	maid	hotel	P	770	836	1	9
10	4551½	20	No	No		No one at home																					10
11	4551	21	No	No		Vacant																					11
12	4539	22	No	No		No one at home																					12
13	4529	23	No	No		Vacant																					13
14	4418	24	No	No		Tompkins, Clelland M.	Head		W	M	52	Mar	Canada-other	161	yes	WK				60	proprietor	retail grocery store	O	290	636	3	14
15						—, Jean M.	wife		W	F	48	Mar	Canada-other	261	No	WK				40	sales clerk	retail grocery	P	490	636	1	15
16						—, Wallace M.	son		W	M	22	Nev.	California			OT	No	No	No								16
17						—, Eleanor E.	daughter		W	F	23	Nev.	Calif.			WK				40	nurse	county hospital	G	058	869	2	17
18	4464	25	No	No		No one at home																					18
19	4472	26	No	No		Acosta, Joe A.	Head		W	M	24	Mar	Calif.			OT	No	Yes			spray operator	pottery factory	P	470	319	1	19
20						—, Jean (mrs)	wife		W	F	22	Mar	Calif.			H	No	Yes			counter girl	retail bakery	P	490	636	1	20
21						— Frank A.	son		W	M	5	Nev	Calif.														21
22	4472	~~27~~				Sanchez, Pete T.	lodger		W	M	43	D	Mexico	262	No	OT	No	Yes			conductor	urban transportation	P	431	516	4	22
23	4478	~~27~~ 27				Walker, Raymond E.	Head		W	M	29	Mar	Missouri	043		WK				54	assistant store manager	retail sales	P	270	699	1	23
24						— Helen J.	wife		W	F	27	Mar	Missouri	043		H	No	No	No								24
25						—, Pamela J.	daughter		W	F	4	Nev.	Missouri	043													25
26	4478½	28				Hartman, Emory J.	Head		W	M	34	Mar	Illinois	033		OT	No	Yes			bartender	bar	P	750	679	1	26
27						—, Francis M.	wife		W	F	26	Mar	Arkansas	071		H	No	No	No								27
28						—, Diana L.	daughter		W	F	7	Nev	Calif.														28
29						—, Gary F.	son		W	M	5	Nev	Calif.														29
30						—, Kristin L.	daughter		W	F	3	Nev	Calif.														30

HOUSEHOLD CONTINUED ON NEXT SHEET ☒
Notes

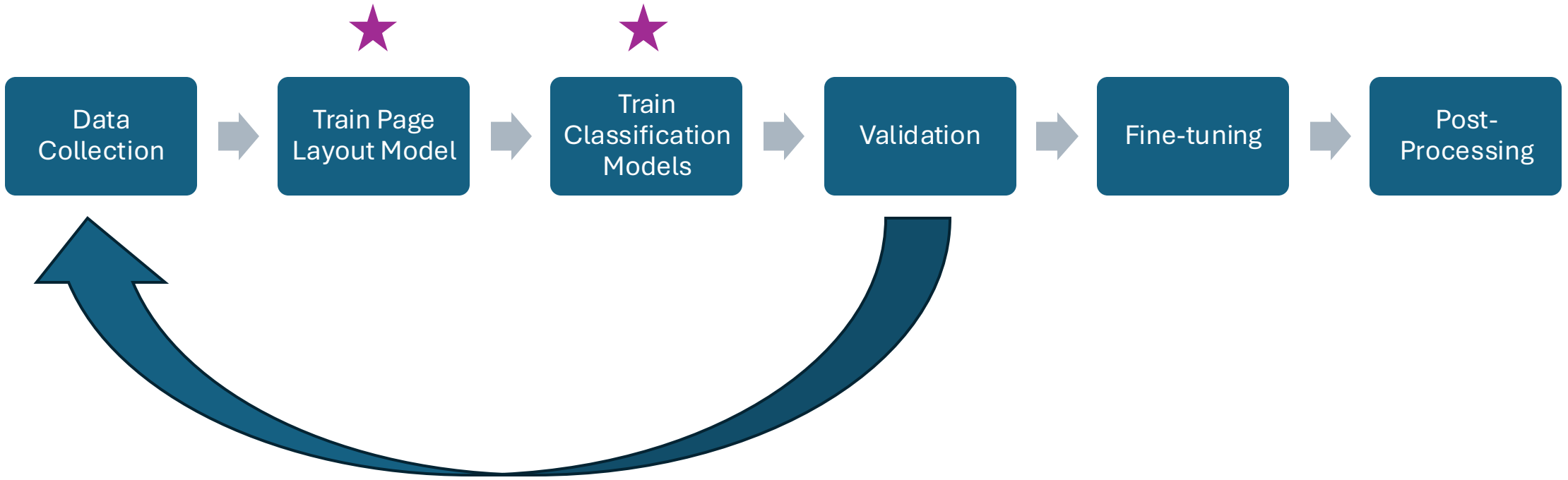
THE QUESTIONS BELOW ARE FOR PERSONS LISTED ON SAMPLE LINES

Size of the Prize

- overcounts across all Census Collections

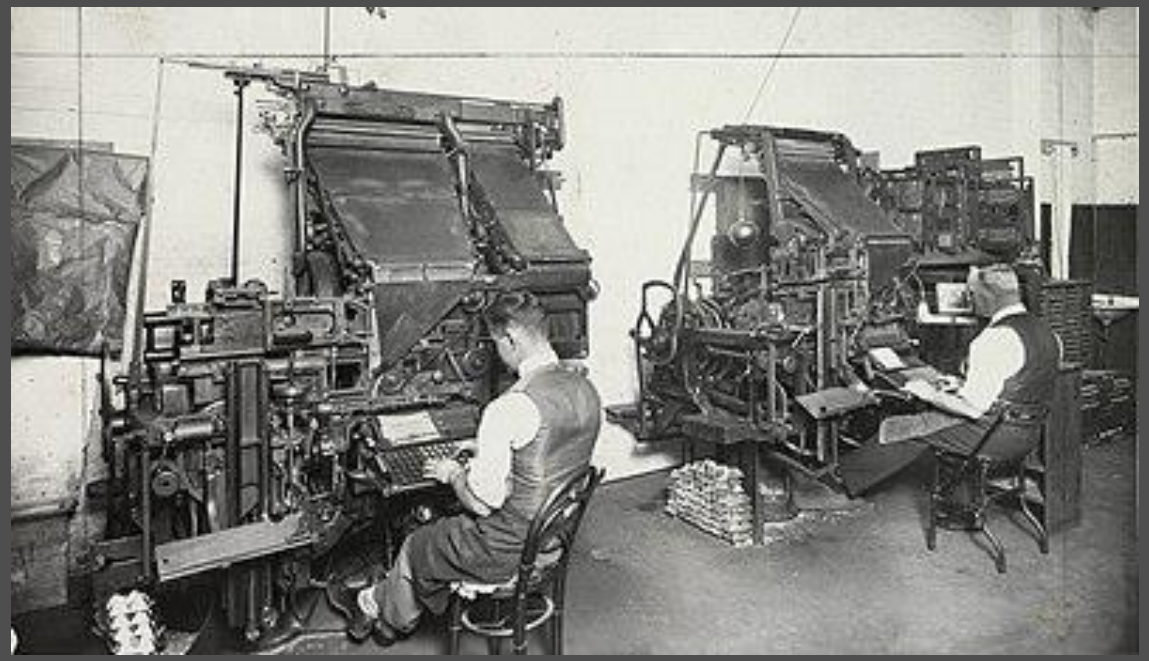


Machine Learning Iteration Process

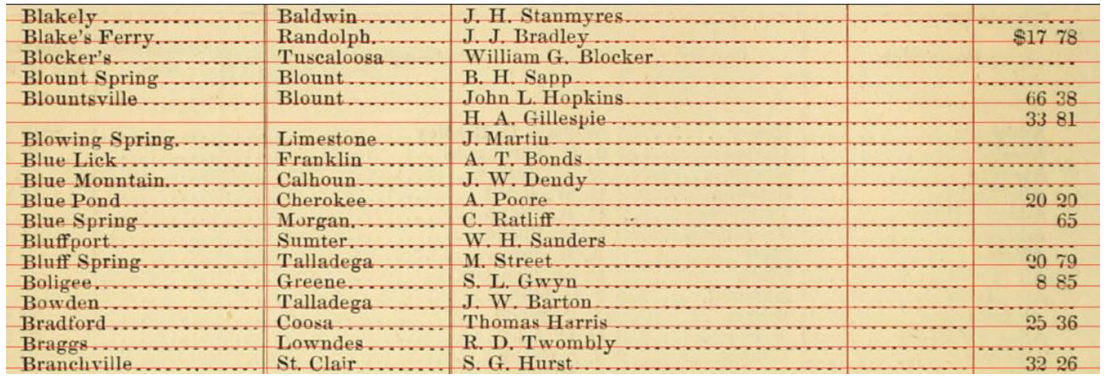


		H. A. Gillespie		33 81
Blowing Spring	Limestone	J. Martin		
Blue Lick	Franklin	A. T. Bonds		
Blue Monntain	Calhoun	J. W. Dendy		
Blue Pond	Cherokee	A. Poore		20 20
Blue Spring	Morgan	C. Ratliff		65
Bluffport	Sumter	W. H. Sanders		
Bluff Spring	Talladega	M. Street		20 79
Boligee	Greene	S. L. Gwyn		8 85
Bowden	Talladega	J. W. Barton		
Bradford	Coosa	Thomas Harris		25 36
Braggs	Lowndes	R. D. Twombly		
Branchville	St. Clair	S. G. Hurst		32 26
Breckinridge	Conecuh	W. W. Johnson		
Brewersville	Sumter	S. J. Arrington		
Brickville	Lawrence	J. N. McCarley		20 86
Bridgeport	Jackson	William C. Glover		19 23
Bridgetown	Shelby	Thomas L. Pledger		14 23
		M. L. Inzer		3 27
Bridgeville	Pickens	M. W. Stephenson		33 06
Broken Arrow	St. Clair	F. B. Walker		
Brooklyn	Conecuh	Y. S. Hirshfelder		26 85
Brooksville	Blount	H. C. Reed		14 64
Broomtown	Cherokee	J. Mills		
Brownsville	Talladega	Daniel Brown		3 38
Bruceville	Pike	Henry Lane		
Brundidge	Pike	G. C. Collier		45 03
Brush Creek	Perry	J. Massengale		13 07
Buchanan	Randolph	A. B. Fincher		
Buckhorn	Pike	J. P. D. Kelley		
Buena Vista	Monroe	M. Patterson		34 27
Buford	Barbour	J. Thomas		
Bulger's Mills	Coosa	J. H. Hubbard		11 12
Bullock	Coffee	A. H. Justice		
Burleson	Franklin	A. A. Hughes		
Burnsville	Dallas	Josiah Dunn		
Burnt Corn	Monroe	J. N. Dennard		
Bushville	Barbour	T. C. Parker		
Butler	Choctaw	J. T. Foster		119 53
		Chas. C. Hill		5 81
Butler Spring	Butler	T. C. Watts		
Buycksville	Coosa	W. H. Spigener		
Buzbeeville	Coffee	Wm. F. Johnson		2 31
Cahaba	Dallas	R. J. Travers		354 32
Cainland	Benton	Richard Jenkins		3 85
Calhoun	Lowndes	A. McCaskill		43 10
Cambridge	Dallas	C. M. Cochran		
Camden	Wilcox	T. C. Brewer		411 07
Campbell's Home	Shelby	H. B. Jennings		
Campbell's Store	Blount	J. H. Campbell		12 15
Camp Hill	Tallapoosa	B. Conire		19 47
Camp Spring	Lawrence	William A. Wilam		
Carlowsville	Dallas	J. D. Alison		
Carrollton	Pickens	William C. Wilson		155 21
Carthage	Tuscaloosa	T. Brown		13 69
Catoma	Montgomery	Benjamin Rodgers		10 63
Cave Spring	Fayette	S. Hamil		
Cedar Bluff	Cherokee	John Lawrence		80 03
Cedar Grove	Jefferson	J. N. Benkett		1 97
		N. W. McDaniel		9 34

Use Case 1:
Transcribing *the Official Register of the United States*

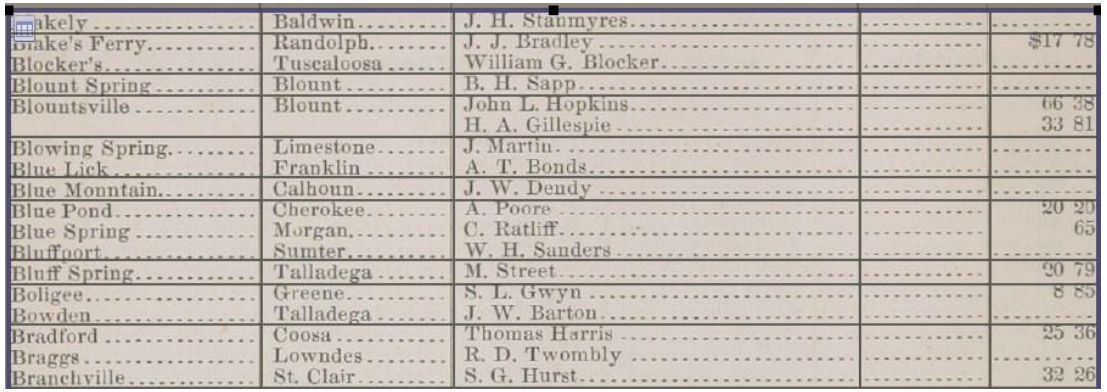


Page Segmentation Results



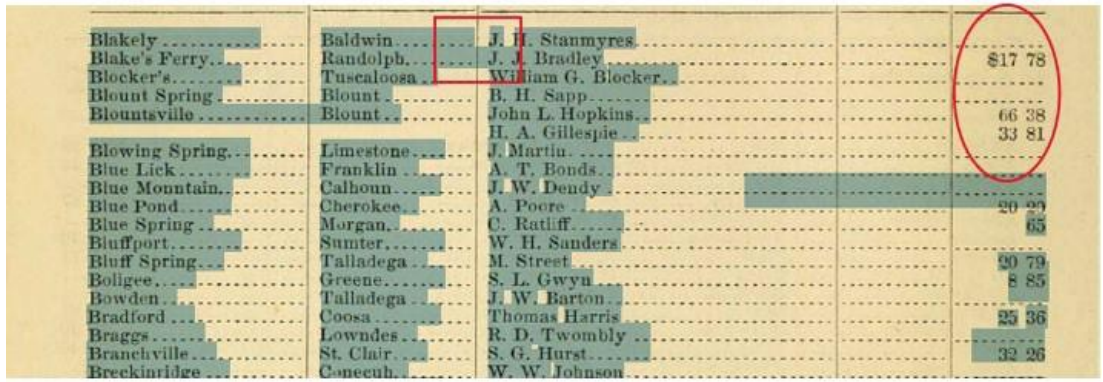
Blakely	Baldwin	J. H. Stanmyres		
Blake's Ferry	Randolph	J. J. Bradley		$17 78
Blocker's	Tuscaloosa	William G. Blocker		
Blount Spring	Blount	B. H. Sapp		
Blountsville	Blount	John L. Hopkins		66 38
		H. A. Gillespie		33 81
Blowing Spring	Limestone	J. Martin		
Blue Lick	Franklin	A. T. Bonds		
Blue Mountain	Calhoun	J. W. Dendy		
Blue Pond	Cherokee	A. Poore		20 20
Blue Spring	Morgan	C. Ratliff		65
Bluffport	Sumter	W. H. Sanders		
Bluff Spring	Talladega	M. Street		20 79
Boligee	Greene	S. L. Gwyn		8 85
Bowden	Talladega	J. W. Barton		
Bradford	Coosa	Thomas Harris		25 36
Braggs	Lowndes	R. D. Twombly		
Branchville	St. Clair	S. G. Hurst		32 26

Page Segmentation Result – My Workflow



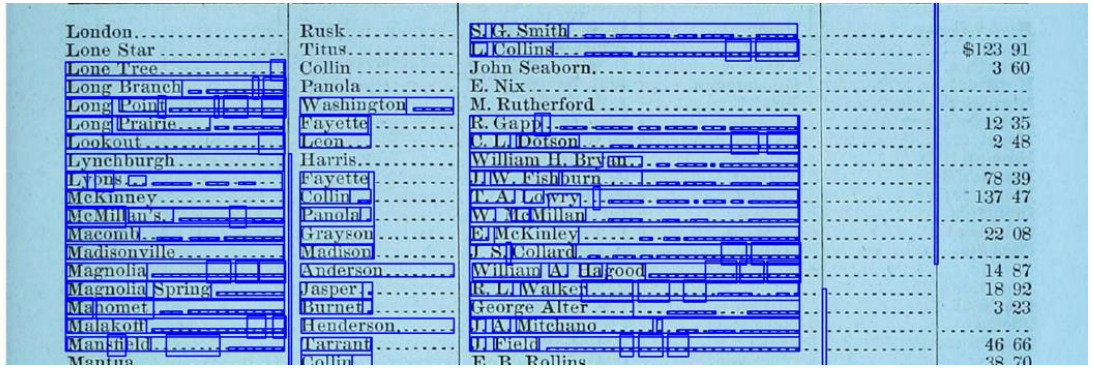
akely	Baldwin	J. H. Stanmyres		
Blake's Ferry	Randolph	J. J. Bradley		$17 78
Blocker's	Tuscaloosa	William G. Blocker		
Blount Spring	Blount	B. H. Sapp		
Blountsville	Blount	John L. Hopkins		66 38
		H. A. Gillespie		33 81
Blowing Spring	Limestone	J. Martin		
Blue Lick	Franklin	A. T. Bonds		
Blue Mountain	Calhoun	J. W. Dendy		
Blue Pond	Cherokee	A. Poore		20 20
Blue Spring	Morgan	C. Ratliff		65
Bluffoort	Sumter	W. H. Sanders		
Bluff Spring	Talladega	M. Street		20 79
Boligee	Greene	S. L. Gwyn		8 85
Bowden	Talladega	J. W. Barton		
Bradford	Coosa	Thomas Harris		25 36
Braggs	Lowndes	R. D. Twombly		
Branchville	St. Clair	S. G. Hurst		32 26

Page Segmentation Result -- Omnipage



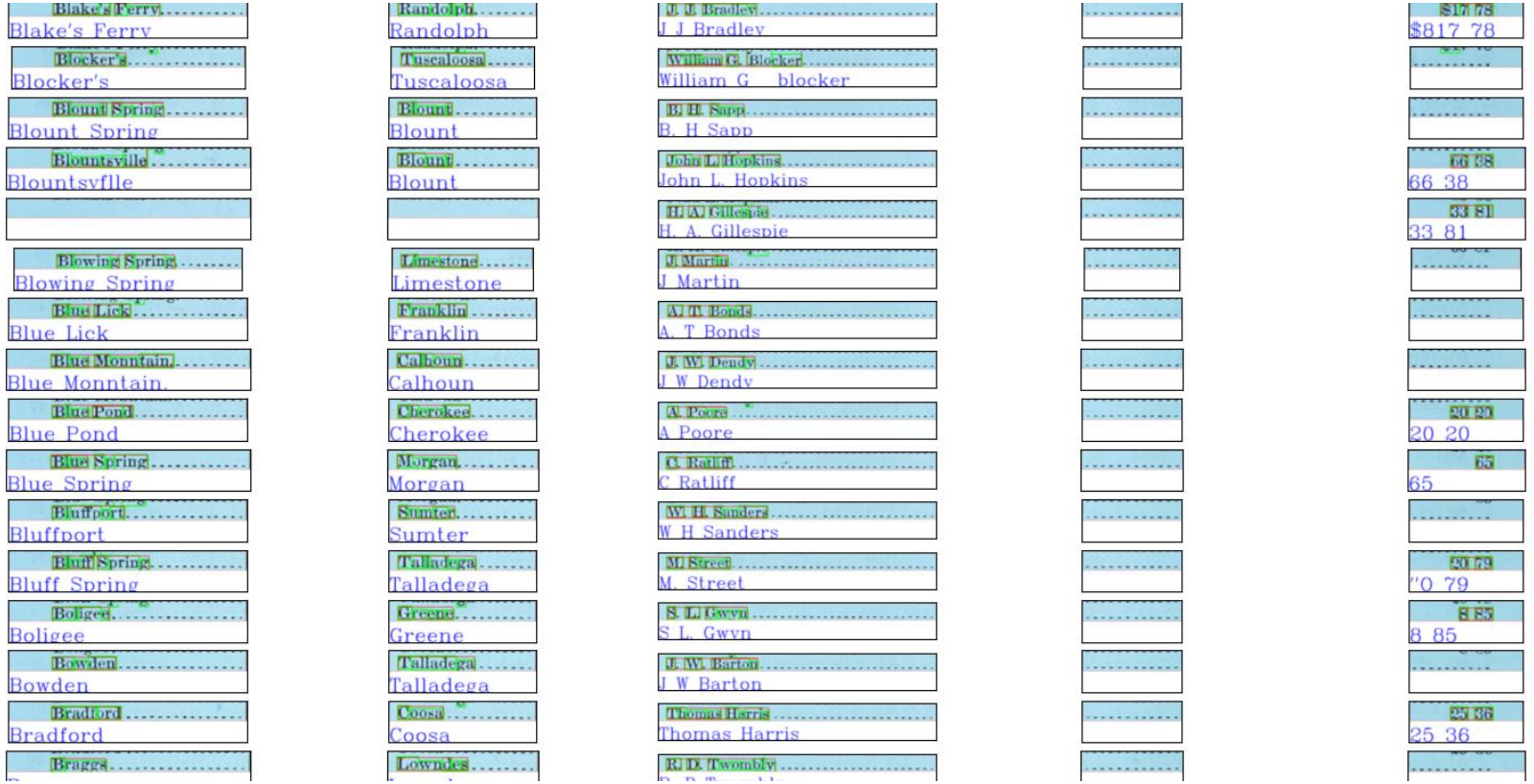
Blakely	Baldwin	J. H. Stanmyres		
Blake's Ferry	Randolph	J. J. Bradley		$17 78
Blocker's	Tuscaloosa	William G. Blocker		
Blount Spring	Blount	B. H. Sapp		
Blountsville	Blount	John L. Hopkins		66 38
		H. A. Gillespie		33 81
Blowing Spring	Limestone	J. Martin		
Blue Lick	Franklin	A. T. Bonds		
Blue Mountain	Calhoun	J. W. Dendy		
Blue Pond	Cherokee	A. Poore		20 20
Blue Spring	Morgan	C. Ratliff		65
Bluffport	Sumter	W. H. Sanders		
Bluff Spring	Talladega	M. Street		20 79
Boligee	Greene	S. L. Gwyn		8 85
Bowden	Talladega	J. W. Barton		
Bradford	Coosa	Thomas Harris		25 36
Braggs	Lowndes	R. D. Twombly		
Branchville	St. Clair	S. G. Hurst		32 26
Breckinridge	Conecuh	W. W. Johnson		

Page Segmentation Result – Adobe Acrobat



London	Rusk	S. G. Smith		
Lone Star	Titus	L. J. Collins		$123 91
Lone Tree	Collin	John Seaborn		3 60
Long Branch	Panola	E. Nix		
Long Point	Washington	M. Rutherford		
Long Prairie	Fayette	R. Gapp		12 35
Lookout	Leon	C. L. Dotson		2 48
Lynchburgh	Harris	William H. Bryan		
Lyons	Fayette	J. W. Fishburn		78 39
McKinney	Collin	T. A. Lowry		137 47
McMillan's	Panola	W. McMillan		
Macomb	Grayson	E. McKinley		22 08
Madisonville	Madison	J. S. Collard		
Magnolia	Anderson	William A. Hagood		14 87
Magnolia Spring	Jasper	R. L. Walker		18 92
Mahomet	Burnet	George Alter		3 23
Malakoff	Henderson	J. A. Mitchano		
Mansfield	Tarrant	J. Field		46 66
Mantua	Collin	E. B. Rollins		38 70

Page Segmentation Result – Tesseract



Character Error Rate 1.54%

Table 1. OCR Output – Customized Tool

1	Post Office	County	Postmaster	Compensation
2	Blake's Ferry	Randolph	J J Bradley	$17 78
3	Blocker's	Tuscaloosa	William G Blocker	
4	Blount Spring	Blount	B. H Sapp	
5	Blountsville	Blount	John L. Hopkins	66 38
6			H. A Gillespie	33 81
7	Blowing Spring	Limestone	J Martin	
8	Blue Lick	Franklin	A T Bonds	
9	Blue Mountain.	Calhoun	J W Dendy	
10	Blue Pond	Cherokee	A Poore	20 20
11	Blue Spring	Morgan.	C Ratliff	65
12	Bluffport	Sumter	W H Sanders	
13	Bluff Spring	Talladega	M. Street	79
14	Boligee	Greene	8 L. Gwyn	8 85
15	Bowden	Talladega	J W Barton	
16	Bradford	Coosa	Thomas Harris	25 36
17	Bragas	Lowndes	R. D Twombly	

Table 2. OCR output – OmniPage

1	Post Office	County	Postmaster	Compensation
2	Blake, Ferry	Randolph	J. J. Bradley	$17 78
3	Blocker,	Tuscaloosa	William G. Blocker	
4	Blount Spring	Blount	B. H. Sapp	
5	Blountsville	Blount	John L. Hopkins	66 38
6		H. A. Gillespie	33 81	
7	Blowing Spring	Limestone	J. Martin.	
8	Blue Lick	Franklin	A. T. Bonds	
9	Blue Mountain	Calhoun	J. W. Deady	
10	Blue Pond	Cherokee	A. Poore	20 20
11	Blue Spring	Morgan	C. Ratliff .	65
12	Bluffpor t	Sumter	W. H. Sanders	
13	Bluff Spring	Talladega	M. Street	20 79
14	Boligee	Greene	S. L. Gwyn	8 85
15	Bowden	Talladega	J. W. Barton	
16	Bradford	Coosa	Thomas Harris	25 36
17	Braggs	Lowndes	R. D. Twombly	32 26

OCR Output ~ Customized Tool v.s. OmniPage

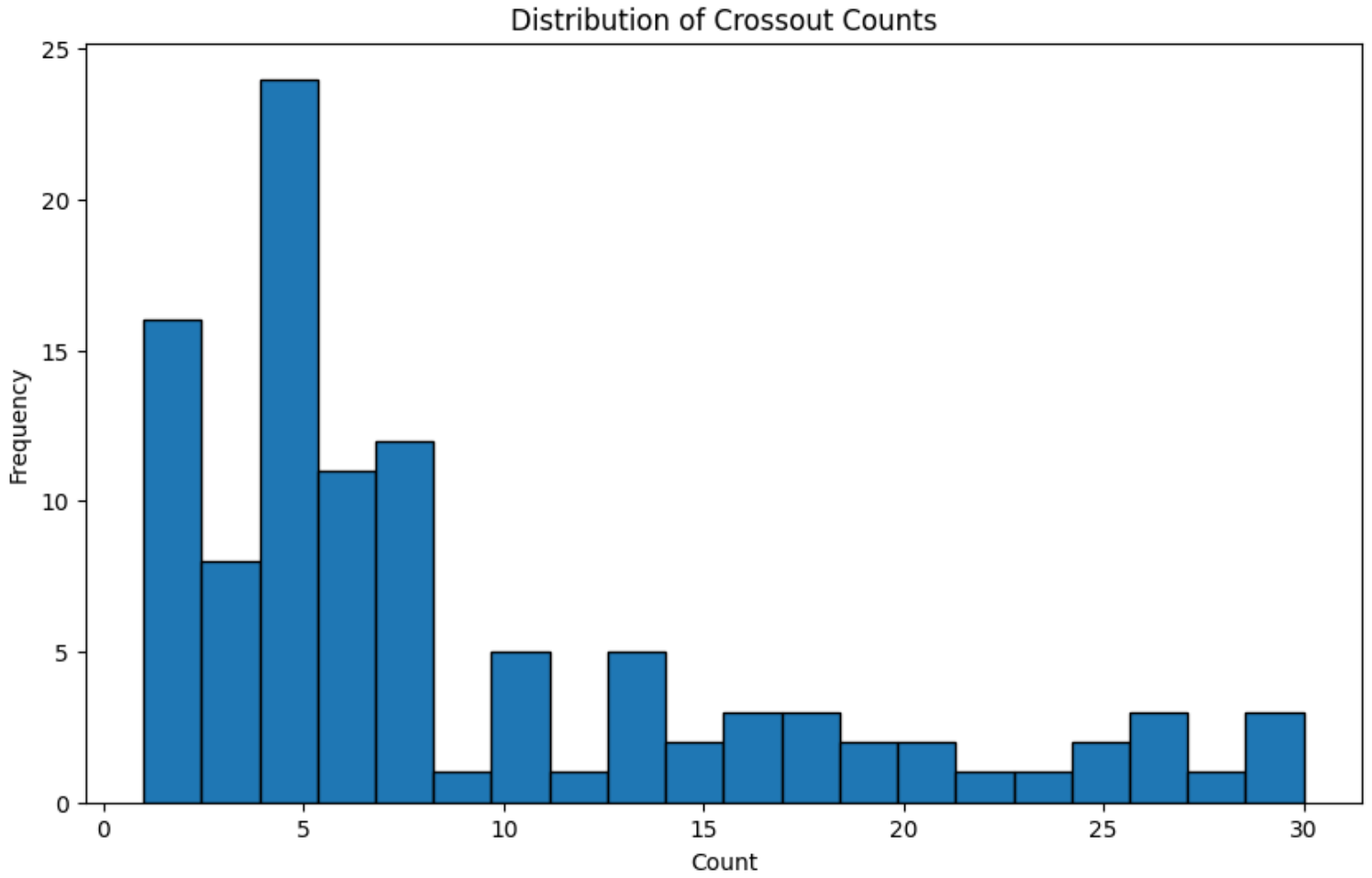
- Page segmentation error by OmniPage
  - Column Splitting Errors
  - Column Insertion Errors
  - Column Shuffling Errors
  - Mixed Errors
- Hard-coded post-processing rules will not work.
  - E.g. “Combine two columns if one is empty.”

Use Case 2: Reduce Census Overcounts

Experiment Design

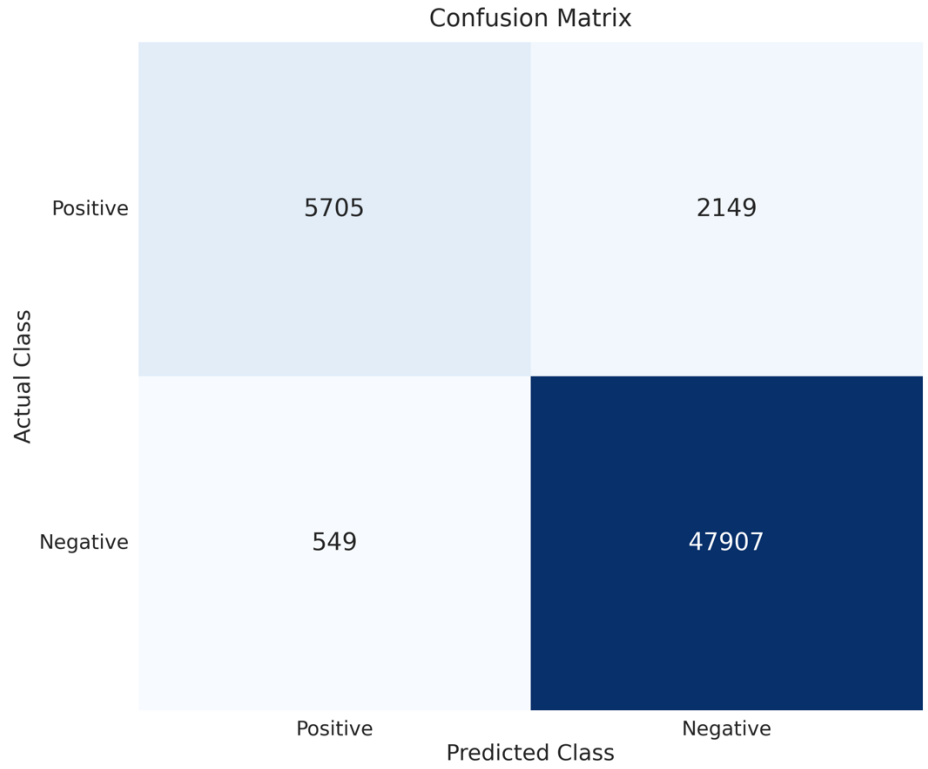
- 1877 (56310 rows) Census Record pages selected from all states with bias for pages with more strikethroughs.
- Fallback to negatives if an image failed at any stage of the pipeline.
- Calculate validation metrics against manually collected strikethrough information.

Distribution of Cross-out Counts in the Validation Dataset



Experiment Results

- Precision: 72.64%
- Recall: 91.22%
- **Accuracy: 95.21%**



- **Precision** measures the accuracy of positive predictions:

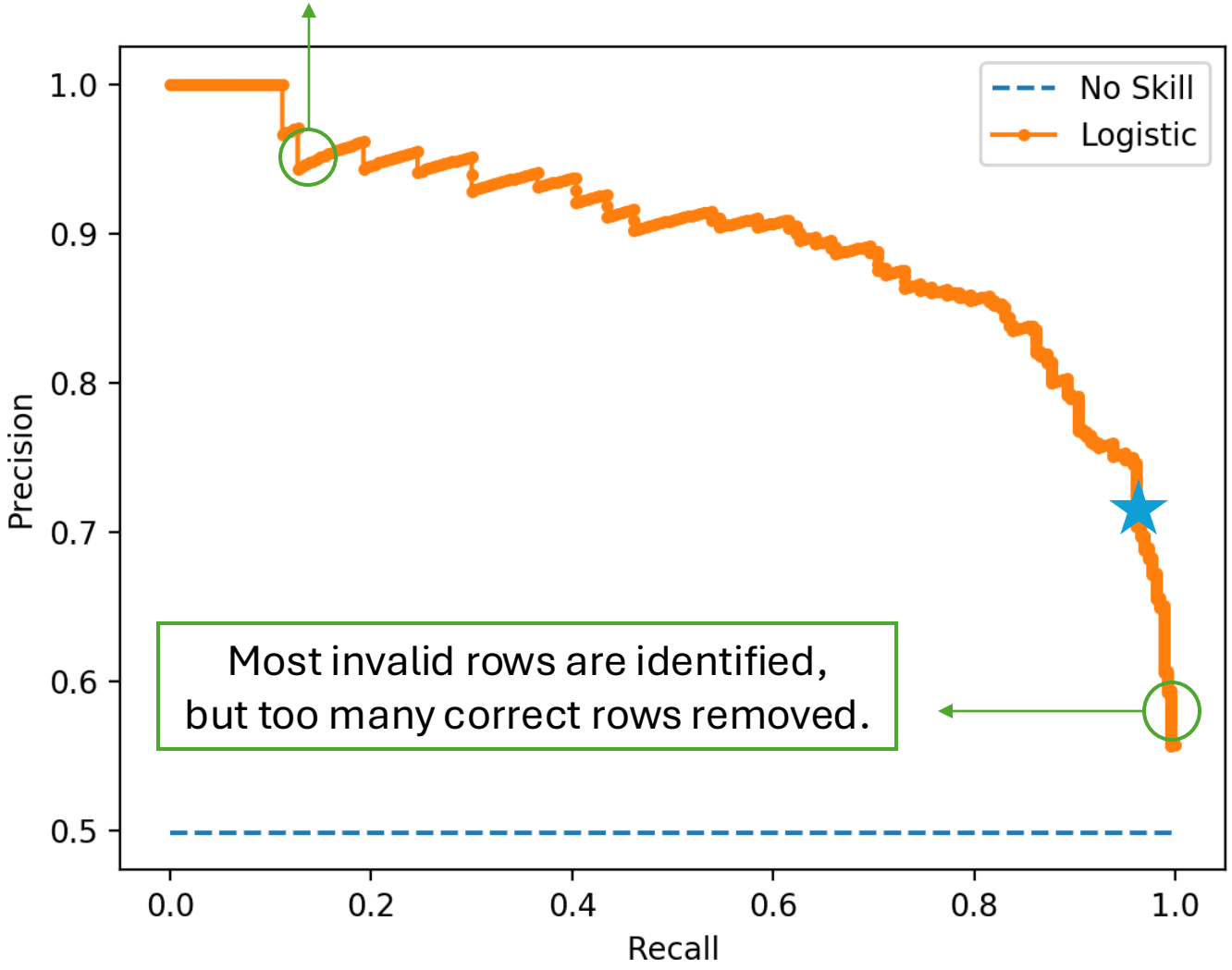
$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

- **Recall** (or Sensitivity) measures the coverage of actual positives:

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

Precision & Recall Curve

Not enough Rows Identified,
But most rows identified are correct.



Most invalid rows are identified,
but too many correct rows removed.

Thank you!

Email: beczhaozmy0@gmail.com

LinkedIn: https://www.linkedin.com/in/mengyue-rebecca-z-a15bb8111/