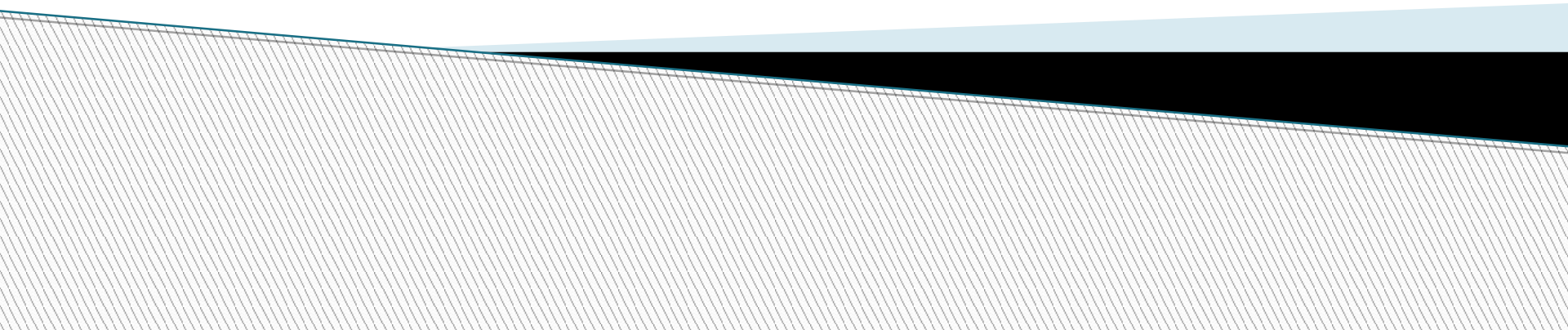


Intelligent Indexing: A Semi-Automated, Trainable System for Field Labeling

Robert Clawson, Bill Barrett
Brigham Young University



STATE Utah COUNTY Garfield TOWNSHIP OR OTHER DIVISION OF COUNTY Garfield NAME OF INSTITUTION Garfield DATE OF RECORD 1920 NAME OF INCORPORATED PLACE Alton ENUMERATED BY ME ON THE 10th DAY OF April 1920

NAME OF INSTITUTION Garfield DATE OF RECORD 1920 NAME OF INCORPORATED PLACE Alton ENUMERATED BY ME ON THE 10th DAY OF April 1920

PAGE OF ABON.	NAME				RELATION.	TIME.	CITYSHIP.										CITYSHIP.	NATIVITY AND BIRTHPLACE.	NATIVITY AND BIRTHPLACE.																																																																																				
	First Name	Middle Name	Last Name	Suffix			1	2	3	4	5	6	7	8	9	10				11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94
1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100			
1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100			
1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100			
1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64</																																							

<TimeToLabel>0</TimeToLabel>

Currently...

FamilySearch Indexing

Close and Return to Start Page Philippines, Metropolitan Manila, Manila—Civil Registration of Birth, 1901-1979 [Part D]/004622543[4] Welcome, Robert Taylor Clawson

100%

CERTIFICATE OF LIVE BIRTH

(FILL OUT COMPLETELY, ACCURATELY, LEGIBLY IN INK OR TYPEWRITER)

Register Number:

Province: _____ (a) Civil Registrar-General No. _____
City or Municipality: MANILA (b) Local Civil Registrar No. 825 (663)

1. PLACE OF BIRTH		2. USUAL RESIDENCE OF MOTHER (Where does mother live?)	
a. PROVINCE		a. PROVINCE	
b. CITY OR MUNICIPALITY	<u>MANILA</u>	b. CITY OR MUNICIPALITY	<u>RIZAL</u>
c. NAME OF HOSPITAL OR INSTITUTION (If not in hospital, give street address)	<u>PERPETUAL SUCCOR HOSPITAL</u>	c. NUMBER AND STREET	<u>616 STO. NIÑO PLAIN NEW MANDALAY</u>
d. IS PLACE OF BIRTH INSIDE CITY LIMITS?	Yes <input checked="" type="checkbox"/> No <input type="checkbox"/>	d. IS RESIDENCE INSIDE CITY LIMITS?	Yes <input type="checkbox"/> No <input checked="" type="checkbox"/>
3. NAME (Type or print)		e. IS RESIDENCE ON A FARM?	
First <u>GERALDINE</u> Middle <u>PILAR</u> Last <u>CAMPUE</u>		Yes <input type="checkbox"/> No <input type="checkbox"/>	
4. SEX	5a. THIS BIRTH	5b. IF TWIN OR TRIPLET, WAS CHILD	6. DATE OF BIRTH
<u>F</u>	SINGLE <input checked="" type="checkbox"/> TWIN <input type="checkbox"/> TRIPLET <input type="checkbox"/>	1st <input type="checkbox"/> 2nd <input type="checkbox"/> 3rd <input type="checkbox"/>	Month <u>2</u> Day <u>7</u> Year <u>63</u>
7. NAME		RELIGION	8. NATIONALITY
First <u>GUSTAVO</u> Middle <u>GALVEZ</u> Last <u>AQUINO</u>		<u>RC</u>	<u>FIL</u>
9. AGE (At time of this birth)		11a. USUAL OCCUPATION	11b. KIND OF BUSINESS OR INDUSTRY
Years <u>35</u>	10. BIRTHPLACE <u>MANILA</u>	<u>EMPLOYEE</u>	
12. MAIDEN NAME		RELIGION	13. NATIONALITY
First <u>NORMA</u> Middle <u>CAERLAN</u> Last <u>CAMPUE</u>		<u>RC</u>	<u>FIL</u>
		13a. RACE	<u>BR</u>

Header Data Table Entry Form Entry

Image-Record

Register Number
01 - 01
*Child's Given Names <Required>
02 - 01
*Child's Surname <Required>
03 - 01
*Gender <Required>
04 - 01
*Birth Month <Required>
05 - 01
*Birth Day <Required>
06 - 01
*Birth Year <Required>

Field Help Quality Checker Project Instructions Image Navigation

Register Number

Type the register number as it was written. If the register number includes letters or hyphens, index it as it appears.

If a register number was not recorded, press Tab to skip this field.

Image 2 of 15 Record 2 of 15 Field: Register Number Completion: 1% Download Complete

Our Research

- ▶ 5-10 times faster
- ▶ More accurate
- ▶ More enjoyable

Word Morphing

- ▶ <https://www.youtube.com/watch?v=eBQjHgejchA>

Question

- ▶ How can handwriting recognition be used to improve indexing?

Possible Methods

- ▶ Machine Learning Approach
- ▶ Split training/testset
- ▶ 1920 Utah Census
- ▶ About 50000 fields per category
- ▶ Accuracy ~80%
 - Error rate too high
 - Would require lots of corrections after the fact

Possible Methods

- ▶ Pre-clustering
- ▶ Cost Matrix for each category/document

Relationship to head of household				
	0	20.2	15.9	11.4
		0	8.0	9.3
			0	3.2
				0

Possible Methods

Pre-clustering

Green ICR

Enter label:

1920 Utah Census

TABLE 1-20. HOUSEHOLD MEMBERS. (SEE INSTRUCTIONS.)

PLACE OF BIRTH.				NAME		PERSON DESCRIPTION.				CITIZENSHIP.		EDUCATION.		Place of birth of each person and parents of each PERSON.					
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Street, or town, or post office.	House number or farm, etc. (see instructions.)	Number of dwelling house in order of the block.	Number of family in order of the household.	of each person whose January 1, 1920, was	Enter surname first, then the given name, if any.	Sex.	Color or race.	Age at last birth-day.	Single, married, widowed, or divorced.	Year of immigration or last arrival in the United States.	Naturalized or alien.	If naturalized, date of naturalization.	Attended school within year ending Sept. 1, 1918.	Whether able to read.	Whether able to write.	Place of birth.	Mother tongue.		
	Eng 1	1		Morris Mary	Head	1	D F	F	W	69	W				yes	yes	Utah		
				Zealand	Son			M	W	35	S				yes	yes	Utah		
	Eng 2	2		Morris Frank	Head	1	D F	M	W	28	M				yes	yes	Utah		
	V			Margaret	Wife			F	W	30	M				yes	yes	Utah		
	V			Edna	Daughter			F	W	4 1/2	S				yes	yes	Utah		
	V			Birchena	Daughter			F	W	3 1/2	S				yes	yes	Utah		
	Eng 3	3		Sault Harry V.	Head	1	D F	M	W	39	M				yes	yes	Utah		
	V			Lebbie	Wife			F	W	39	M				yes	yes	Utah		
	V			Laura	Daughter			F	W	17	S			yes	yes	yes	Utah		
	V			Velma	Daughter			F	W	15	S			yes	yes	yes	Utah		
	V			Russell	Son			M	W	12	S			yes	yes	yes	Utah		
	V			Edwina	Daughter			F	W	6	S			yes	yes	yes	Utah		
	V			Furnell	Son			M	W	3 1/2	S			yes	yes	yes	Utah		
	Eng 4	4		Morris Joseph H.	Head	1	D F	M	W	46	M				yes	yes	Utah		
	V			Emily	Wife			F	W	40	M				yes	yes	Utah		
	V			Lealand	Son			M	W	17	S			no	yes	yes	Utah		
	V			Violet	Daughter			F	W	15	S			yes	yes	yes	Utah		
	V			Nellie	Daughter			F	W	13	S			yes	yes	yes	Utah		
	V			Ormond	Son			M	W	5	S			no	no	yes	Utah		
	V			Agnes	Daughter			F	W	2 1/2	S			yes	yes	yes	Utah		
	Eng 5	5		Barton Stephen P.	Head	1	D F	M	W	69	M				yes	yes	Utah		
	V			Sarah	Wife			F	W	55	M				yes	yes	Utah		
	V			Sherman	Son			M	W	25	S				yes	yes	Utah		
	V			Amasa	Son			M	W	21	S			no	yes	yes	Utah		

Percent labeled: Correctly labeled:

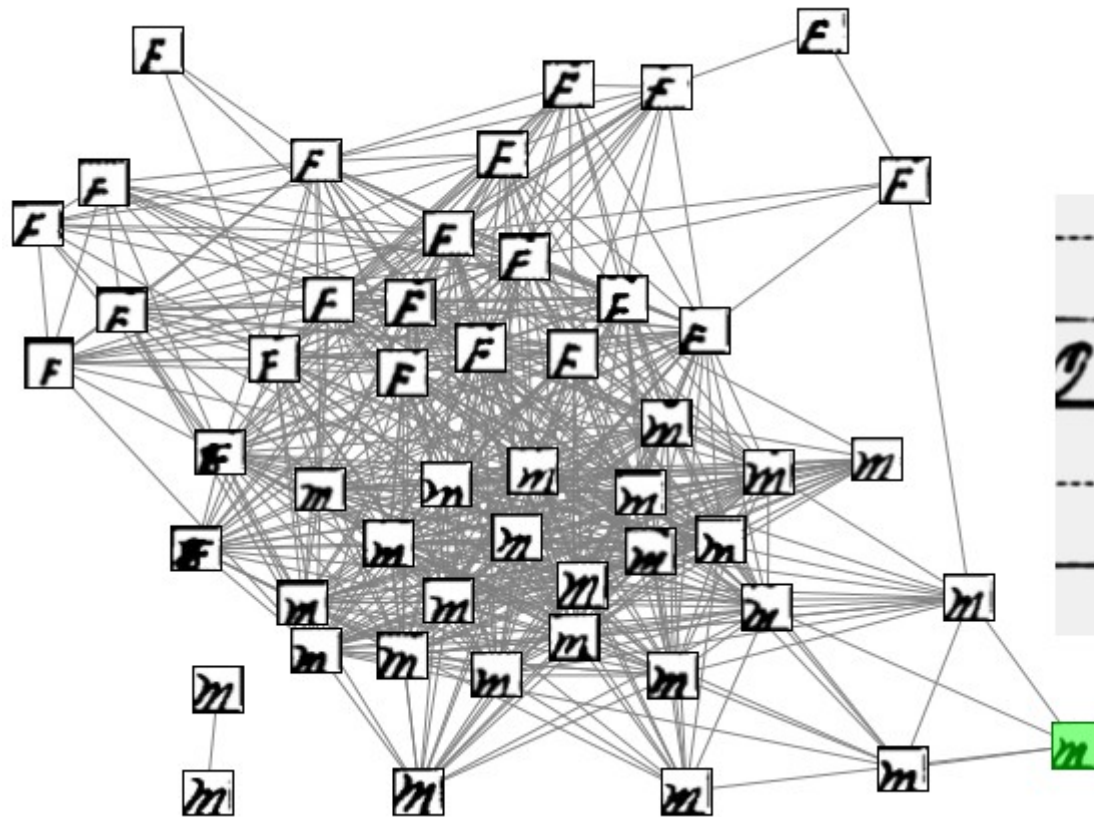
Possible Methods

- ▶ Pre-clustering
- ▶ Problems
 - Still makes frequent errors,
 - Indexer has to scan up and down the page to look for mistakes
 - How to show clusters when there are many different words in the column?

Breakthrough

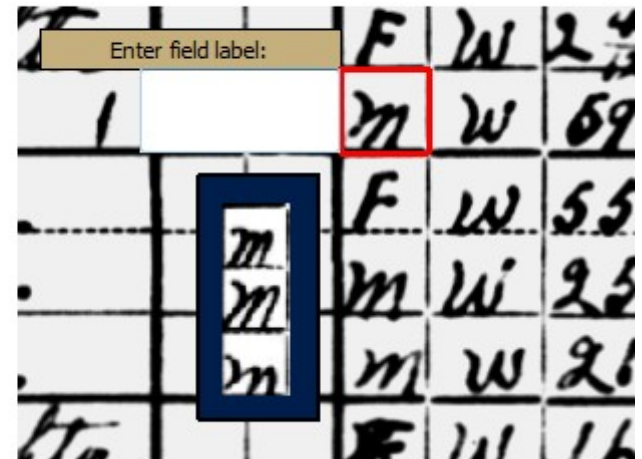
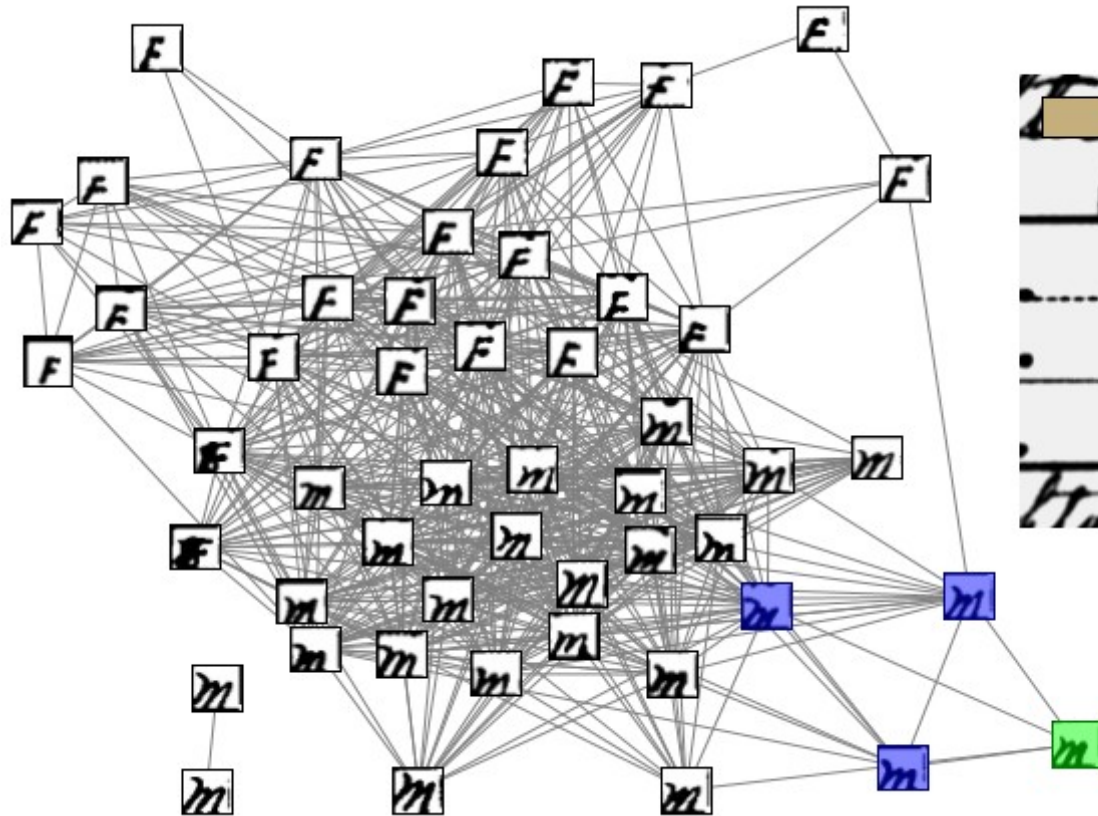
- ▶ Interactive training: learn as you go, correct as you go
- ▶ Indexer drives
- ▶ Training set can start empty, quick ramp up
- ▶ Still use cost matrix
 - Switch from per document to per enumerator
- ▶ Introduce threshold
 - All fields that match under a threshold are labeled
 - Learn the threshold
- ▶ Demo

Training Set

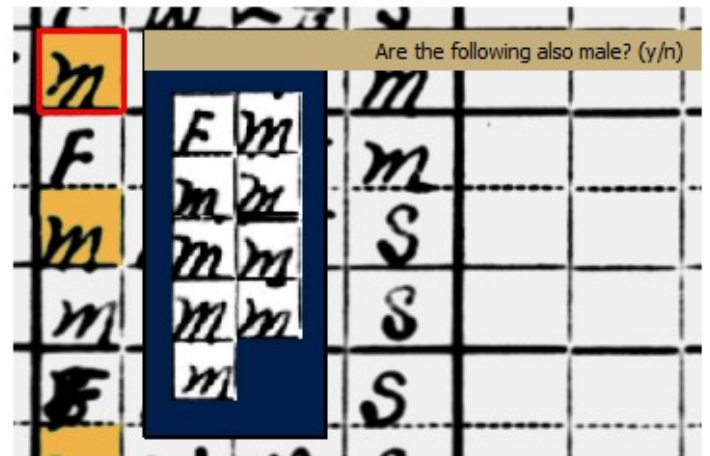
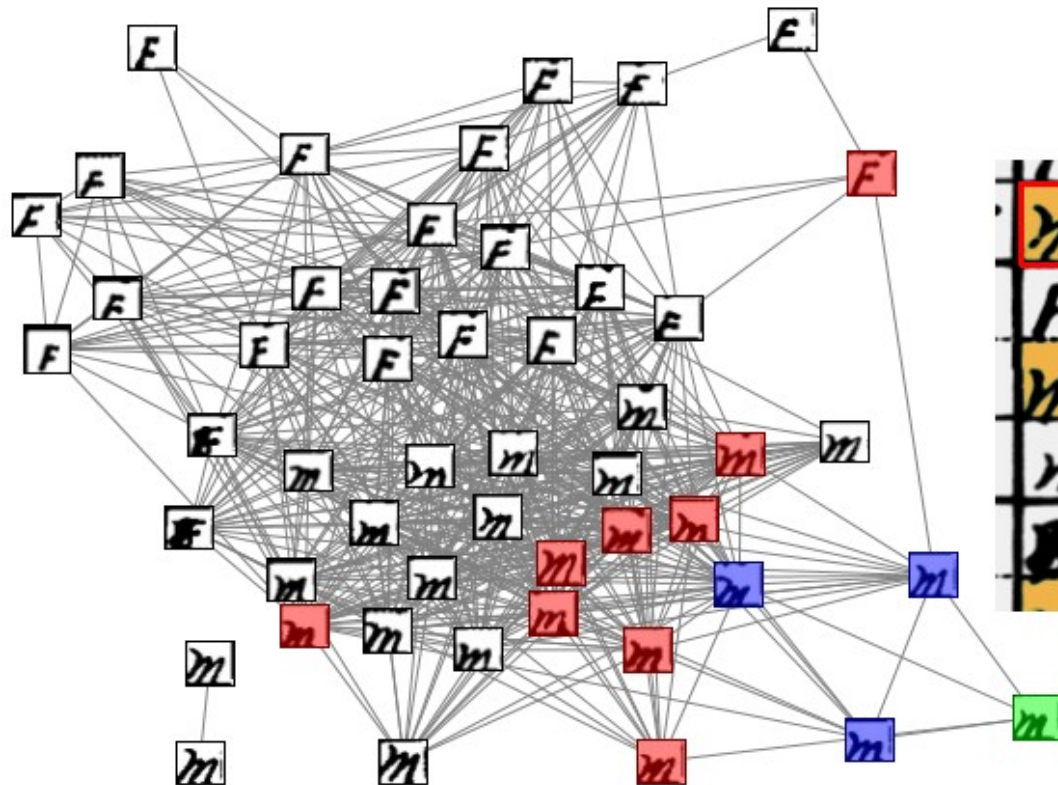


	m	w	2	2
	F	w	2.4	S
9	F	m	69	m
	F	w	55	m
	m	w	25	S
	m	w	26	S

Training Set



Training Set



You can help

- ▶ We need volunteers to test Intelligent Indexing
- ▶ To volunteer: email me (Robert Clawson) at:

intelligentindexing@gmail.com