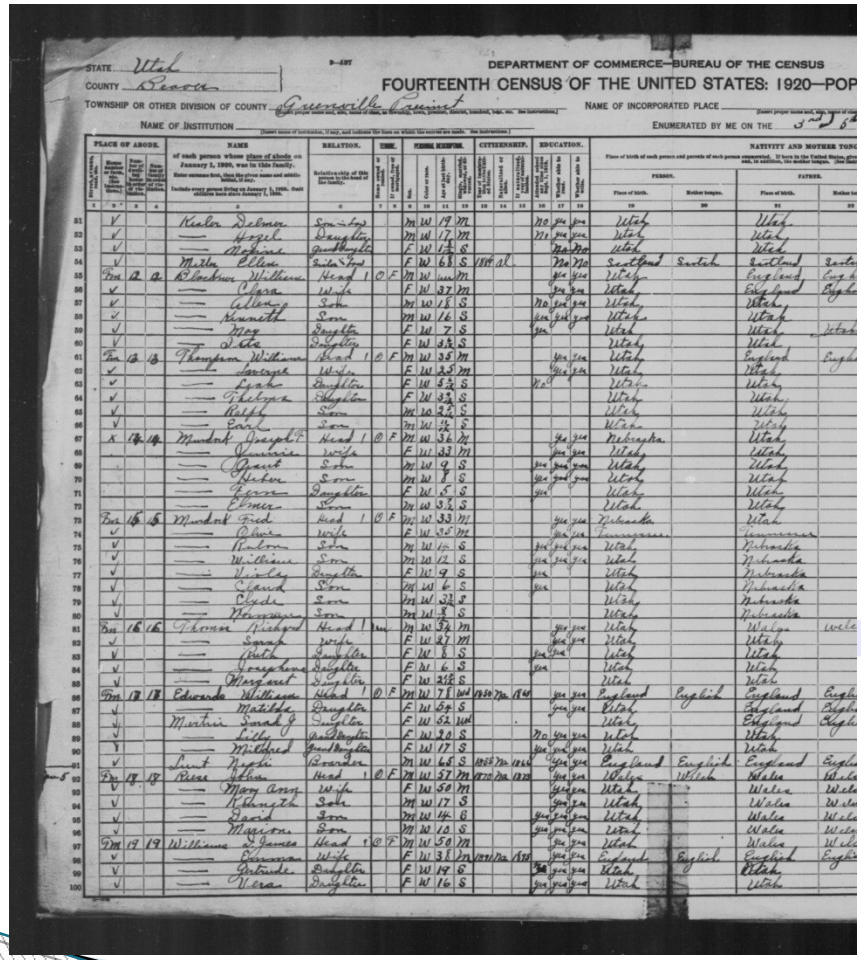
Intelligent Indexing: A Semi-Automated, Trainable System for Field Labeling

Robert Clawson, Bill Barrett
Brigham Young University

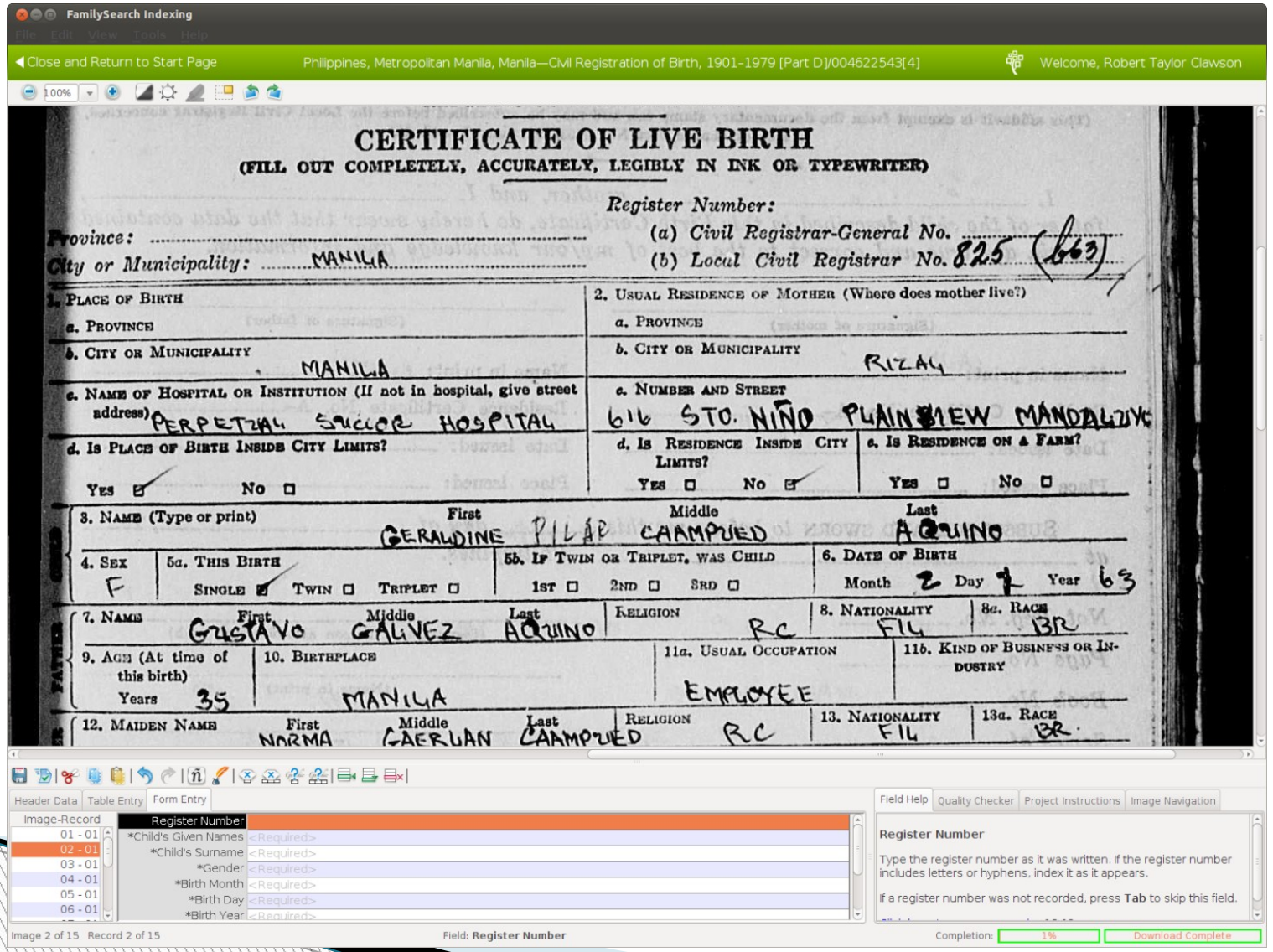
What is indexing?


STATE Utah
COUNTY Beaver
TOWNSHIP OR OTHER DIVISION OF COUNTY Greenville Precinct
NAME OF INCORPORATED PLACE
NAME OF INSTITUTION
ENUMERATED BY ME ON THE 3rd Jan

DEPARTMENT OF COMMERCE—BUREAU OF THE CENSUS
FOURTEENTH CENSUS OF THE UNITED STATES: 1920—POP

```
<Field>
    <BoundingBox>1752,713,249,42</BoundingBox>
    <Category>BIRTH_PLACE</Category>
    <DocumentImageName>004390330_01088</DocumentImageName>
    <Enumerator>Dresden J Cragun</Enumerator>
    <RecordId>004390330_01088_2</RecordId>
    <GroundTruth>Utah</GroundTruth>
    <Label>utah</Label>
    <TimeToLabel>0</TimeToLabel>
</Field>
<Field>
    <BoundingBox>1752,755,249,39</BoundingBox>
    <Category>BIRTH_PLACE</Category>
    <DocumentImageName>004390330_01088</DocumentImageName>
    <Enumerator>Dresden J Cragun</Enumerator>
    <RecordId>004390330_01088_3</RecordId>
    <GroundTruth>Utah</GroundTruth>
    <Label>utah</Label>
    <TimeToLabel>1.008</TimeToLabel>
</Field>
<Field>
    <BoundingBox>1752,794,249,41</BoundingBox>
    <Category>BIRTH_PLACE</Category>
    <DocumentImageName>004390330_01088</DocumentImageName>
    <Enumerator>Dresden J Cragun</Enumerator>
    <RecordId>004390330_01088_4</RecordId>
    <GroundTruth>Utah</GroundTruth>
    <Label>utah</Label>
    <TimeToLabel>0</TimeToLabel>
```

Currently…


FamilySearch Indexing
File Edit View Tools Help
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:
City or Municipality: MANILA
(a) Civil Registrar-General No.
(b) Local Civil Registrar No. 825 (663)
1. PLACE OF BIRTH
a. PROVINCE
b. CITY OR MUNICIPALITY MANILA
c. NAME OF HOSPITAL OR INSTITUTION (If not in hospital, give street address) PERPETUAL SUCCOR HOSPITAL
d. IS PLACE OF BIRTH INSIDE CITY LIMITS? YES ☑ NO ☐
2. USUAL RESIDENCE OF MOTHER (Where does mother live?)
a. PROVINCE
b. CITY OR MUNICIPALITY RIZAL
c. NUMBER AND STREET 616 STO. NIÑO PLAINVIEW MANDALUYONG
d. IS RESIDENCE INSIDE CITY LIMITS? YES ☐ NO ☑
e. IS RESIDENCE ON A FARM? YES ☐ NO ☐
3. NAME (Type or print) First GERALDINE Middle PILAR CAMPUED Last AQUINO
4. SEX F
5a. THIS BIRTH SINGLE ☑ TWIN ☐ TRIPLET ☐
5b. IF TWIN OR TRIPLET, WAS CHILD 1ST ☐ 2ND ☐ 3RD ☐
6. DATE OF BIRTH Month 2 Day 7 Year 63
7. NAME OF FATHER First GUSTAVO Middle GALVEZ Last AQUINO
RELIGION RC
8. NATIONALITY FIL
8a. RACE BR
9. AGE (At time of this birth) Years 35
10. BIRTHPLACE MANILA
11a. USUAL OCCUPATION EMPLOYEE
11b. KIND OF BUSINESS OR INDUSTRY
12. MAIDEN NAME OF MOTHER First NORMA Middle CAERLAN Last CAMPUED
RELIGION RC
13. NATIONALITY FIL
13a. RACE BR
Header Data
Table Entry
Form Entry
Image-Record
01 - 01
02 - 01
03 - 01
04 - 01
05 - 01
06 - 01
Register Number
*Child's Given Names <Required>
*Child's Surname <Required>
*Gender <Required>
*Birth Month <Required>
*Birth Day <Required>
*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

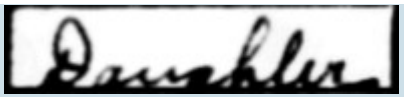
- How can handwriting recognition be used to improve indexing?

Possible Methods

- Machine Learning Approach
- Split training/testset
- 1920 Utah Census
- About50000 fields per category
- Accuracy~80%
  - Errorrate too high
  - Would require lotsof corrections after the fact

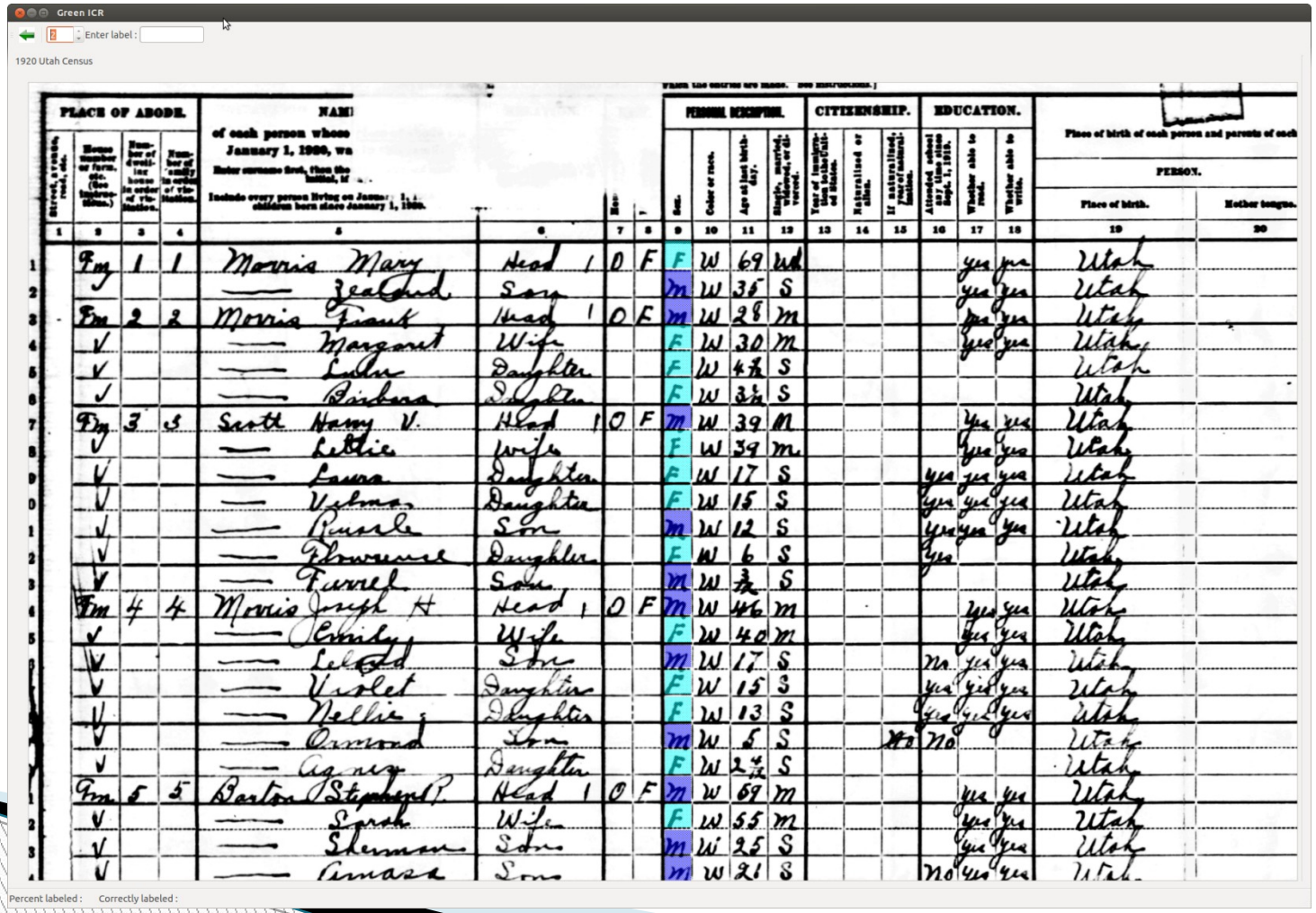
Possible Methods

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

Relationship to headof household				
	0	20.2	15.9	11.4
		0	8.0	9.3
			0	3.2
				0

Possible Methods

- Pre-clustering



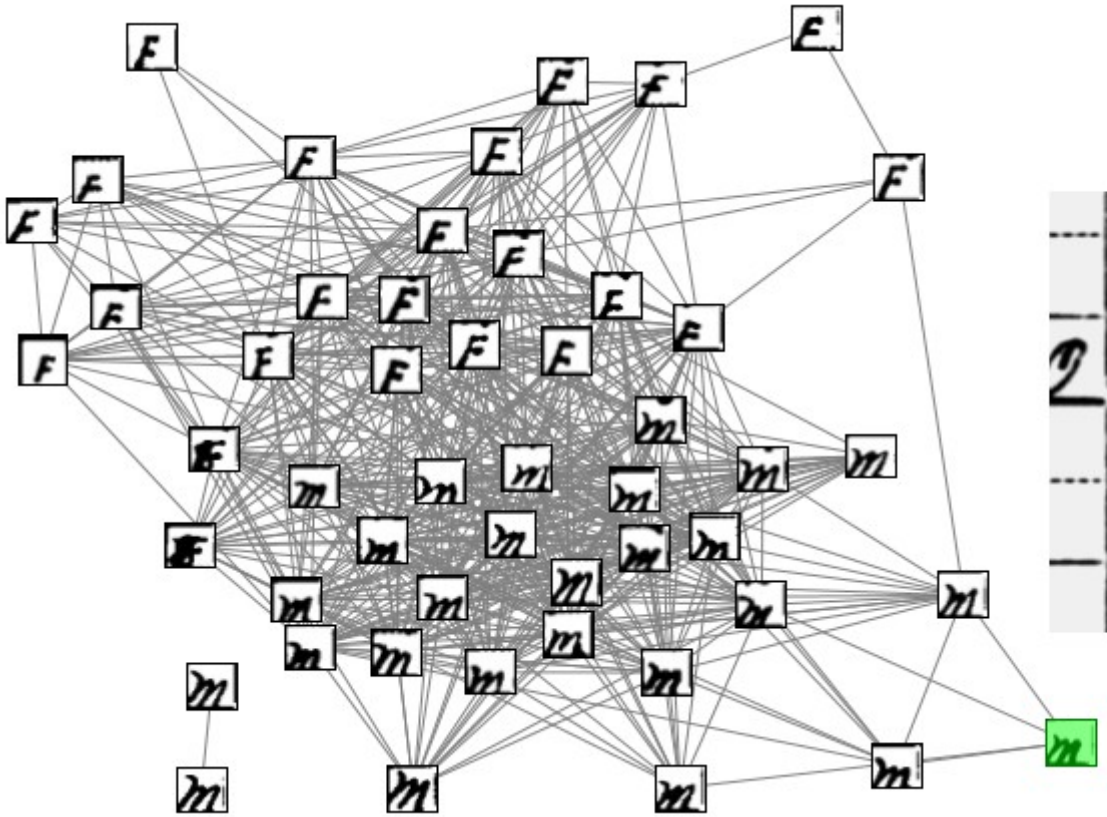
Possible Methods

- 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?

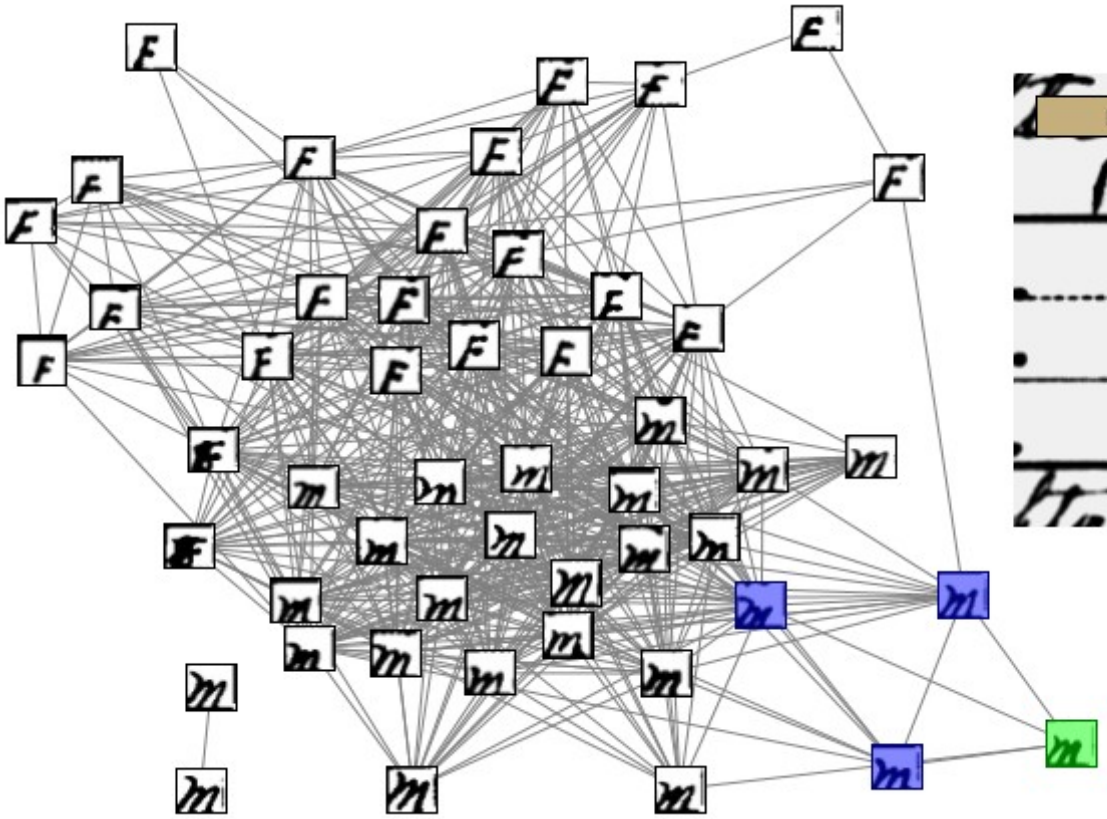
Breakthrough

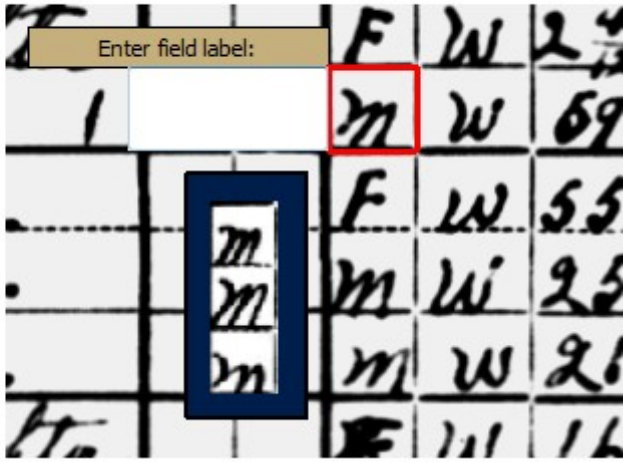
- Interactive training: learn as you go, correct as yougo
- Indexer drives
- Training set can startempty, quick ramp up
- Still use cost matrix
  - Switch from per document to per enumerator
- Introduce threshold
  - Allfields that match under a threshold are labeled
  - Learn thethreshold
- Demo

Training Set

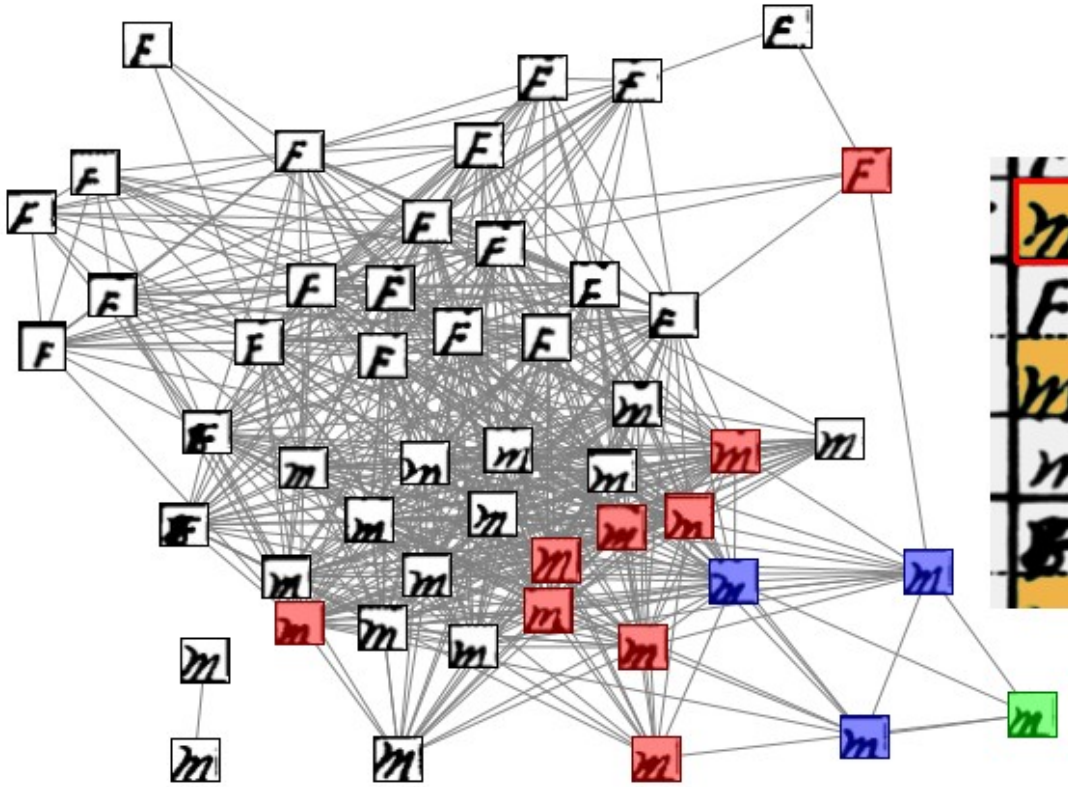


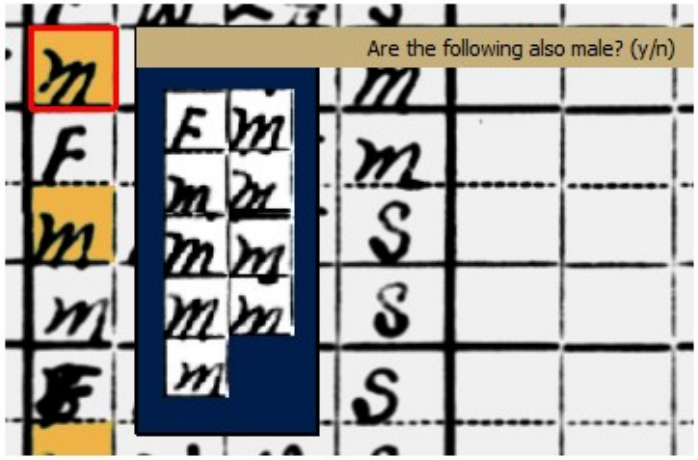
Training Set





Training Set





You can help

- We need volunteers to test Intelligent Indexing

- To volunteer: email me (Robert Clawson) at:

intelligentindexing@gmail.com