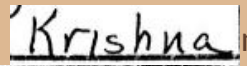
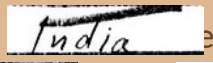
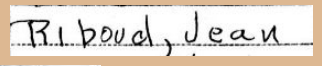
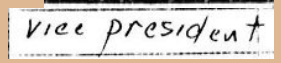
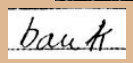
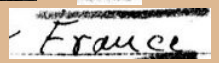
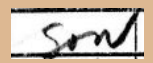
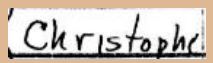
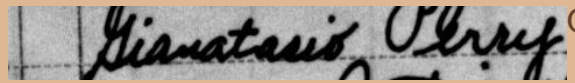
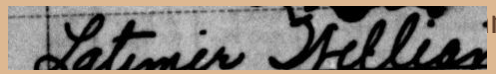
Tracking the Boom: Digitizing the 1950 Census with ML techniques

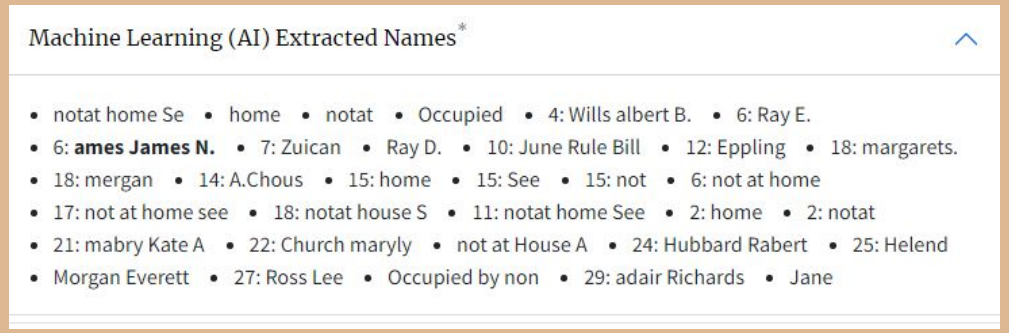
Karthik Nambiar

Why?

-  m  ear  , ma  
     ,  
- Lots of stories like that, which can also provide real micro-data from Census records- the ability to track by job, location, person, etc.
- But not all names are written in plain-case or as easy to
-  Gianatasio
-  m Latimir
-
-  an exercise for the reader

Current gold-standard?

- The 1950 census has used machine-learning techniques to extract names- but no other features- and with limited success



- Earlier censuses either ask you to remember enumeration districts or have been hand-processed by volunteers- and again not all features

Methodology

- Looked at Name/Position/Age/Race/Gender/Job/Job Location
- Tried using a way to clip them down to their tables based on differences in brightness- but it failed on darker image scans:

state	total	failed	percent_failed
Virgin_Islands	50	50	1.0
Panama_Canal_Zone	104	104	1.0
Alaska	33	33	1.0
Guam	50	50	1.0
Puerto_Rico	56	56	1.0
Colorado	32	21	0.66
Wyoming	5	2	0.4
New_Hampshire	61	20	0.33
Georgia	54	9	0.17
New_Jersey	62	10	0.16

Labeling

- Tried to clip based on table cells- but issues with tilt
- Labeled 1407 cells w/ LabelStudio
- Letter Distribution (contained capitals and numbers as well) :

Character	Frequency
a	1028
	950
e	948
-	681
o	634
r	629
n	595
i	518
l	465
t	406

Panta Rhei

- -Heraclitus on cursive
- Flowy nature of cursive can add difficulty in labeling at the character level
- Connectionist Temporal Classification or Hidden Markov Models can mitigate by looking at words at a sequence level

Levenshtein Distance

- Levenstein Distance (LD=1) is number of character edits to transform one word to another
- Best model got an LD of 6.47 on test data!
  - But mean character length is 8.15- so mostly wrong
- Notably more ‘M’s- perhaps to fit with gender category ‘M’ that it assumed was more common than actuality

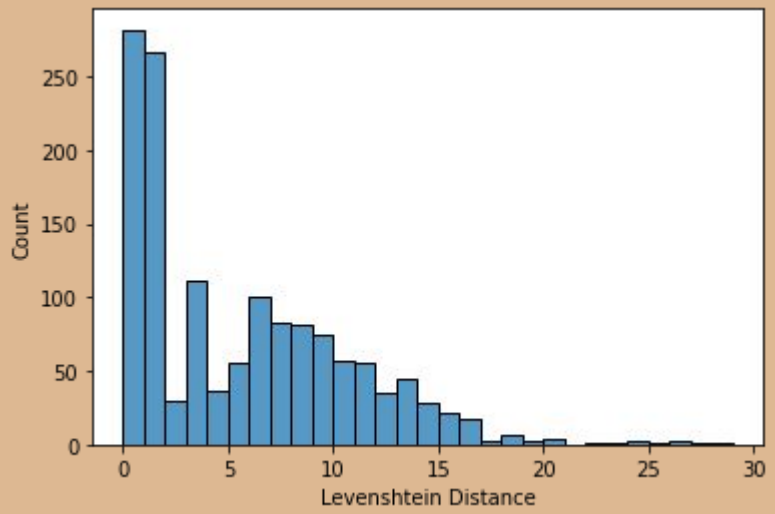
Character	Frequency
a	709
-	671
	567
M	350
o	336
,	323
e	293
n	280
l	189
r	158

At the word level

What the machine guessed

Text	Frequency
---, Ma	90
Alab	69
F	59
M	51
Ne	45
Hea	43
Wif	33
---, M	31
W	29
So	26

L Distance Histogram



True most frequent text cells

Text	Frequency
Alabama	72
M	58
F	58
Neg	52
Head	43
Wife	34
W	28
Son	28
No one at home	21
Daughter	21

Setbacks

- Some sections were written as 2 lines, even at the cell level- makes it difficult to actually use any sequence algorithm
- Inconsistent handwriting suggests it needs larger database
- Even with efforts to include more states, only so many could be processed, and certain states tended to have more tilt and have other problems that made them difficult to process
- CTC Loss and Padding

Next Steps

- Rotating census page based on header/side file
- Drop Job Location, and try to do Age/Race/Gender with standard class-based CV
- Consider using WordBeam for job-matching
- Understand how to use CTC loss, and implement LD calculations in-epoch
- HMMs?

Questions about the model/census?
Suggestions about next steps?
Ask away!