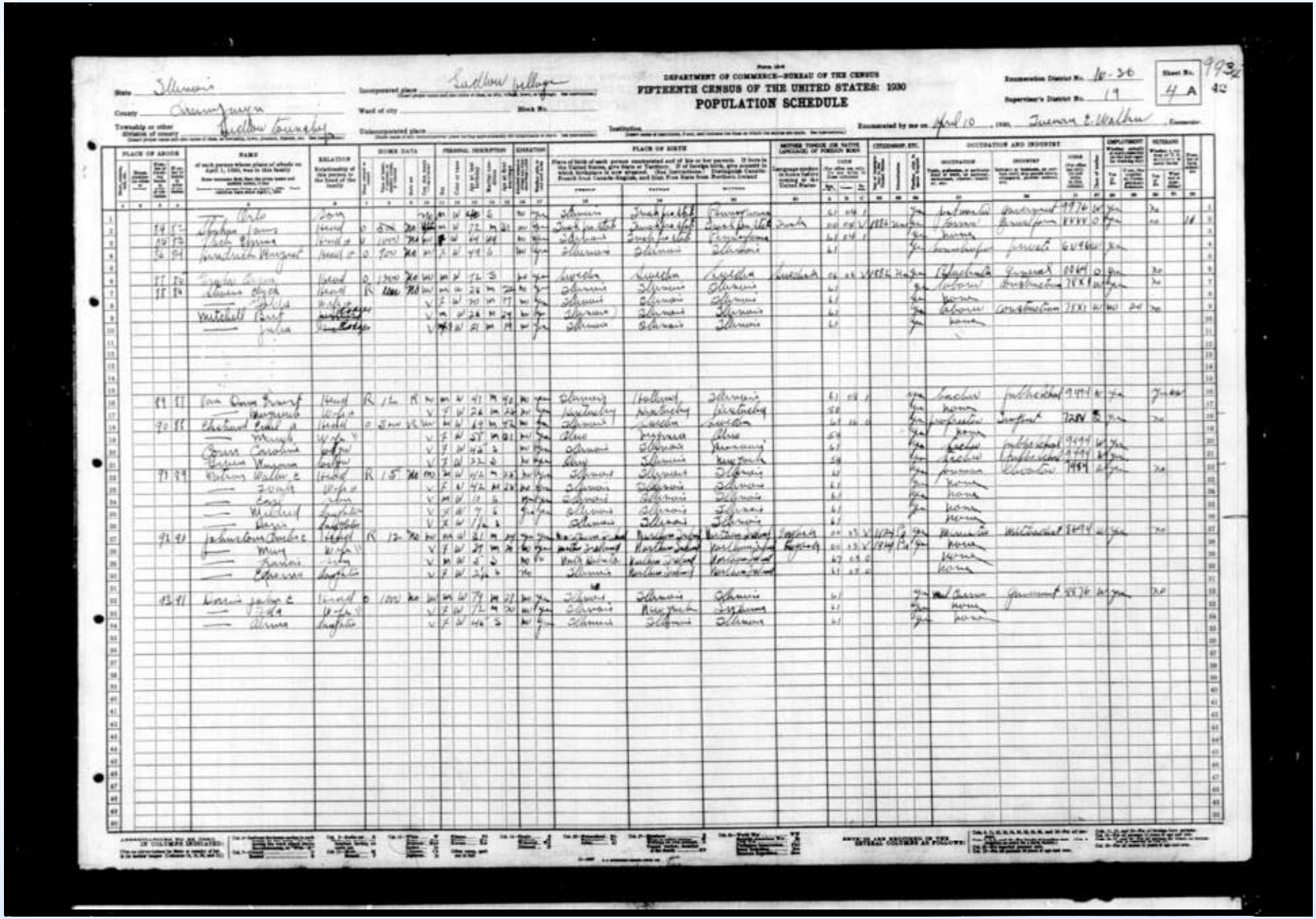
Towards Free and *Searchable* Access of the 1940s Census Data

Kenton McHenry
Luigi Marini
Peter Bajcsy

National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign

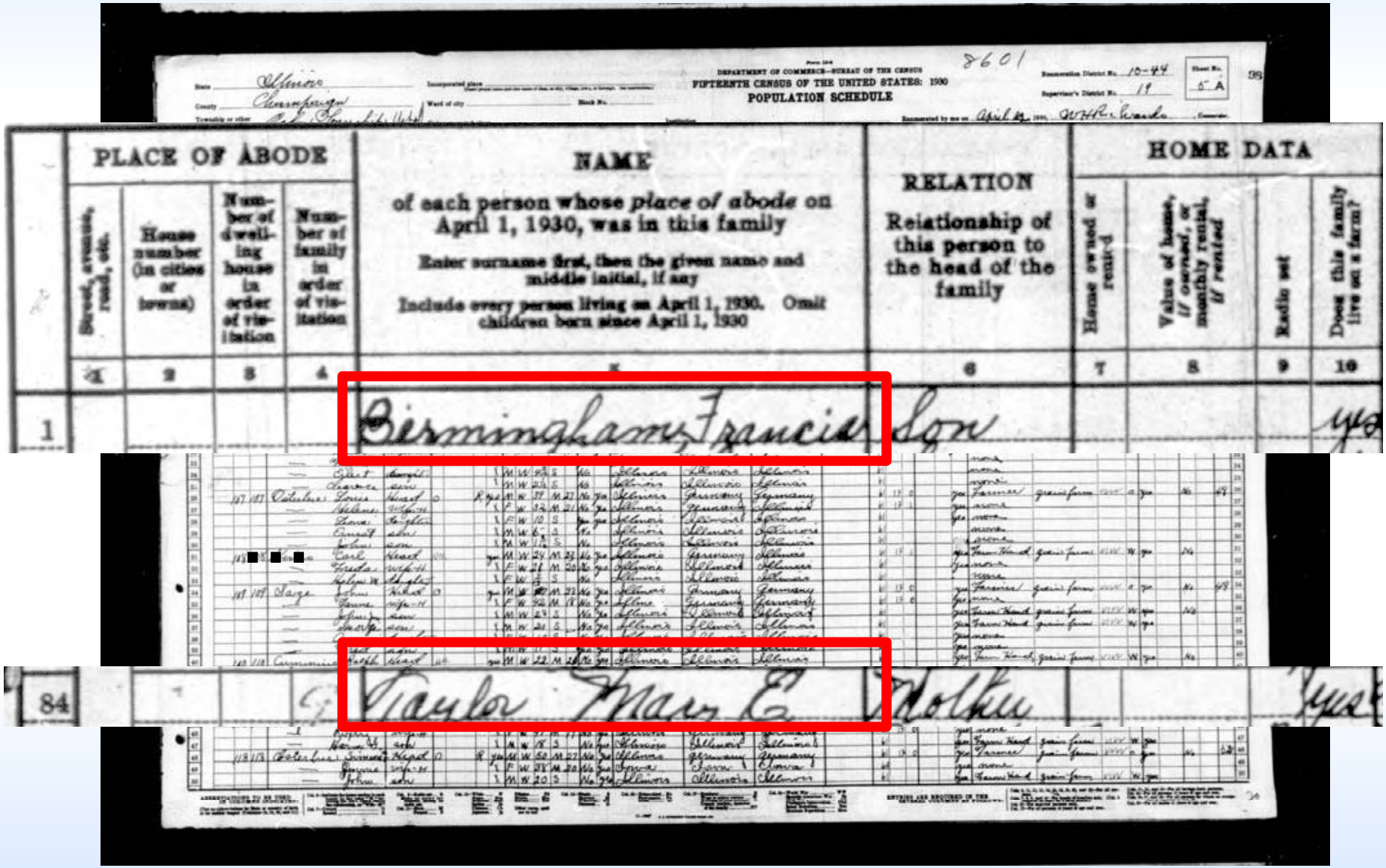
The Problem


DEPARTMENT OF COMMERCE—BUREAU OF THE CENSUS
FIFTEENTH CENSUS OF THE UNITED STATES: 1930
POPULATION SCHEDULE

The Problem

~3.6 million more…

The Data


DEPARTMENT OF COMMERCE—BUREAU OF THE CENSUS
FIFTEENTH CENSUS OF THE UNITED STATES: 1930
POPULATION SCHEDULE
PLACE OF ABODE
NAME
of each person whose place of abode on April 1, 1930, was in this family
Enter surname first, then the given name and middle initial, if any
Include every person living on April 1, 1930. Omit children born since April 1, 1930
RELATION
Relationship of this person to the head of the family
HOME DATA
Street, avenue, road, etc.
House number (in cities or towns)
Number of dwelling house in order of visitation
Number of family in order of visitation
House owned or rented
Value of home, if owned, or monthly rental, if rented
Radio set
Does this family live on a farm?
1
Birmingham, Francis
Son
84
Taylor, Mary E.
Wife

The Data

PERSONAL DESCRIPTION					EDUCATION		PLACE OF BIRTH		
							Place of birth of each person enumerated and of his or her parents. If born in the United States, give State or Territory. If of foreign birth, give country in which birthplace is now situated. (See Instructions.) Distinguish Canada-French from Canada-English, and Irish Free State from Northern Ireland		
Sex	Color or race	Age at last birthday	Marital condition	Age at first marriage	Attended school or college any time since Sept. 1, 1929	Whether able to read and write	PERSON	FATHER	MOTHER
11	12	13	14	15	16	17	18	19	20
M	W	22	S		No	Yes	Illinois	Irish Free State	Irish Free State
...									
F	W	74	Wid		No	Yes	Illinois	Indiana	Irish Free State

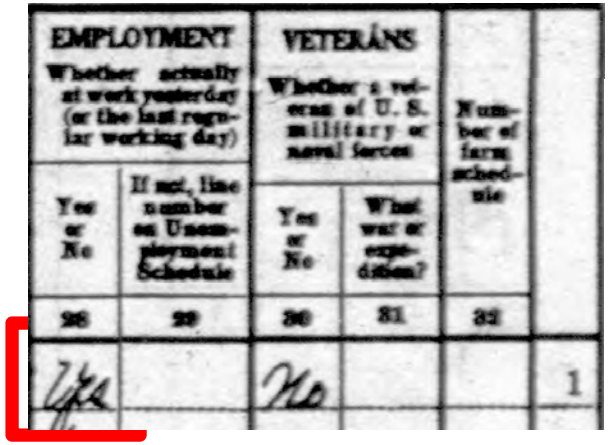
The Data

MOTHER TONGUE (OR NATIVE LANGUAGE) OF FOREIGN BORN				CITIZENSHIP, ETC.			OCCUPATION AND INDUSTRY			
Language spoken in home before coming to the United States	CODE (For office use only. Do not write in these columns)			Year of immigration to the United States	Naturalization	Whether able to speak English	OCCUPATION Trade, profession, or particular kind of work, as *spinner, salesman, riveter, teacher, etc.*	INDUSTRY Industry or business, as *cotton mill, dry-goods store, shipyard, public school, etc.*	CODE (For office use only. Do not write in this column)	Class of worker
	State or N. T.	Country	Religion							
21	A	B	C	22	23	24	25	26	D	27
	41	04	0			Yes	Laborer	Farm	V 0 V V	

...

English	41	04	2			Yes	None			

The Data



EMPLOYMENT		VETERANS			
Whether actually at work yesterday (or the last regular working day)		Whether a veteran of U. S. military or naval forces		Number of farm schedule	
Yes or No	If not, line number on Unemployment Schedule	Yes or No	What war or expedition?		
28	29	30	31	32	
Yes		No			1

...

					84

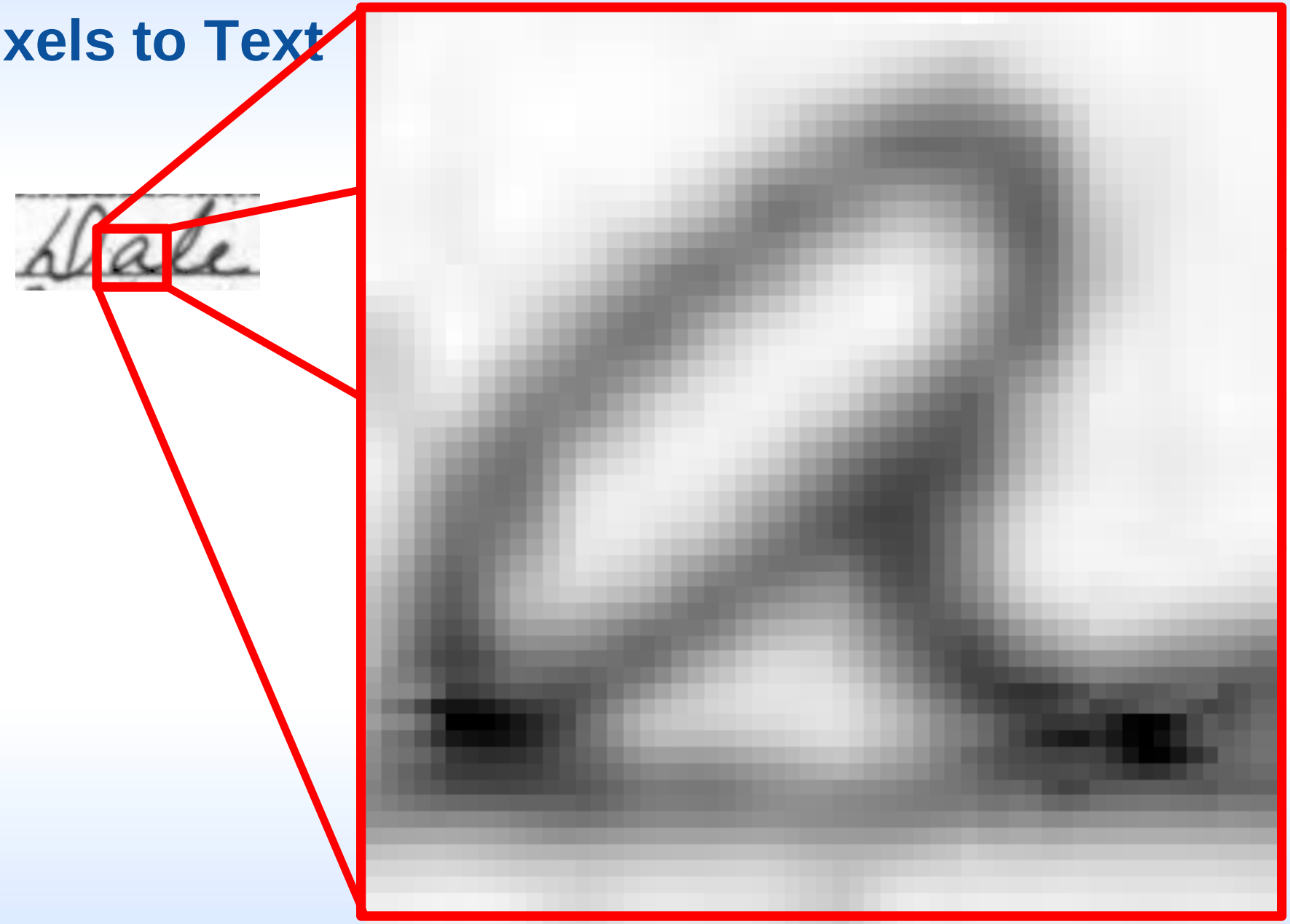
The Information

	PLACE OF ABODE				NAME	RELATION	HOME DATA			
	Street, avenue, road, etc.	House number (in cities or towns)	Number of dwelling house in order of visitation	Number of family in order of visitation	of each person whose *place of abode* on April 1, 1930, was in this family. Enter surname first, then the given name and middle initial, if any. Include every person living on April 1, 1930. Omit children born since April 1, 1930	Relationship of this person to the head of the family	Home owned or rented	Value of home, if owned, or monthly rental, if rented	Radio set	Does this family live on a farm?
	1	2	3	4	5	6	7	8	9	10
1					Birmingham, Francis	Son				Yes

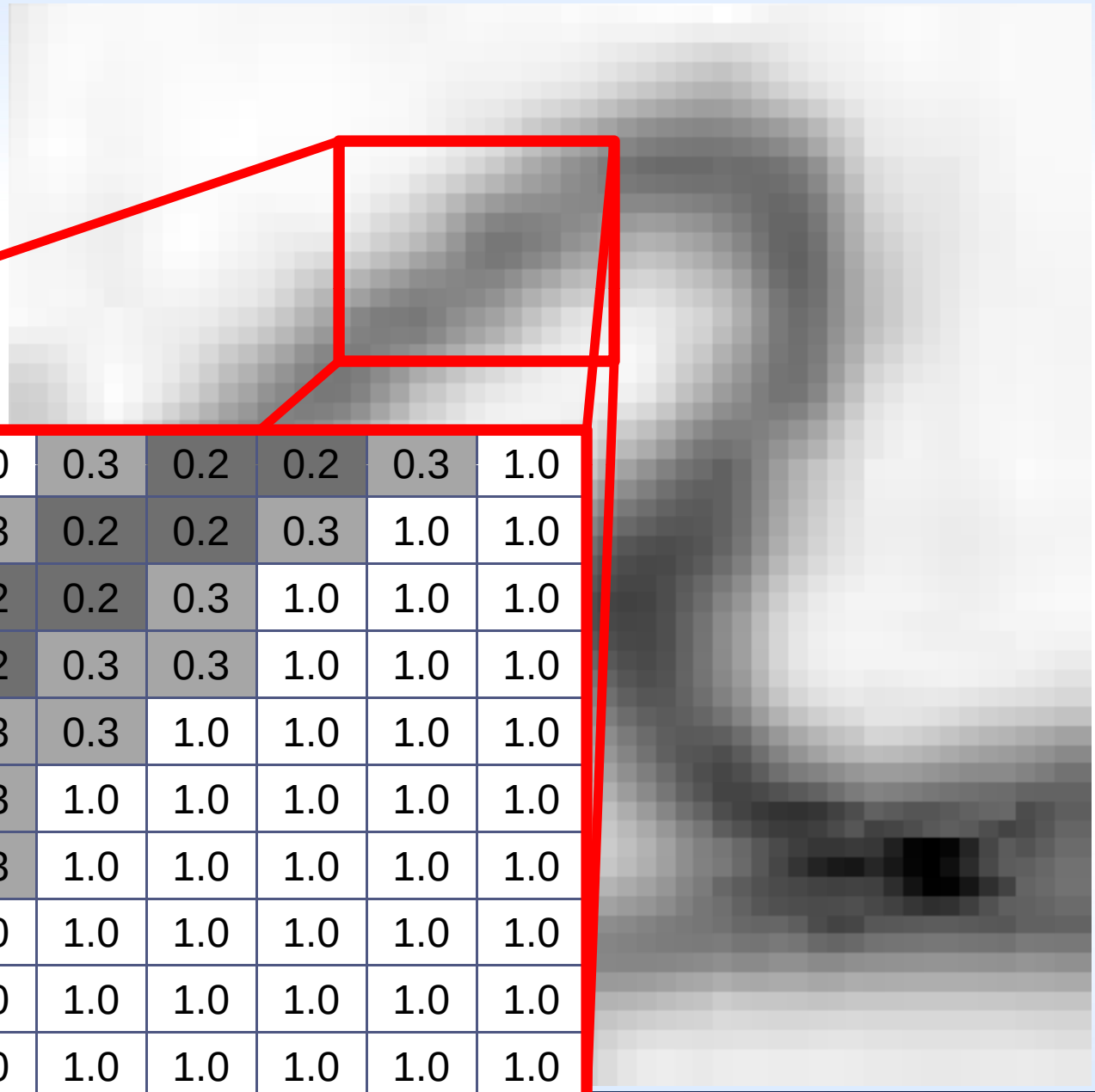
	PERSONAL DESCRIPTION					EDUCATION		PLACE OF BIRTH		
	Sex	Color or race	Age at last birthday	Marital condition	Age at first marriage	Attended school or college any time since Sept. 1, 1929	Whether able to read and write	Place of birth of each person enumerated and of his or her parents. If born in the United States, give State or Territory. If of foreign birth, give country in which birthplace is now situated. (See Instructions.) Distinguish Canada-French from Canada-English, and Irish Free State from Northern Ireland		
								PERSON	FATHER	MOTHER
	11	12	13	14	15	16	17	18	19	20
1	M	W	22	S		No	Yes	Illinois	Irish Free State	Irish Free State

M W 22 S No Yes Illinois Irish Free State Irish Free State

Pixels to Text



Pixels to Text



1.0	1.0	1.0	1.0	1.0	0.3	0.2	0.2	0.3	1.0
1.0	1.0	1.0	1.0	0.3	0.2	0.2	0.3	1.0	1.0
1.0	1.0	1.0	0.3	0.2	0.2	0.3	1.0	1.0	1.0
1.0	1.0	0.3	0.2	0.2	0.3	0.3	1.0	1.0	1.0
1.0	1.0	0.3	0.2	0.3	0.3	1.0	1.0	1.0	1.0
1.0	1.0	0.3	0.2	0.3	1.0	1.0	1.0	1.0	1.0
1.0	0.3	0.2	0.3	0.3	1.0	1.0	1.0	1.0	1.0
1.0	0.3	0.2	0.3	1.0	1.0	1.0	1.0	1.0	1.0
1.0	0.3	0.2	0.3	1.0	1.0	1.0	1.0	1.0	1.0
0.3	0.2	0.2	0.3	1.0	1.0	1.0	1.0	1.0	1.0

NCSA

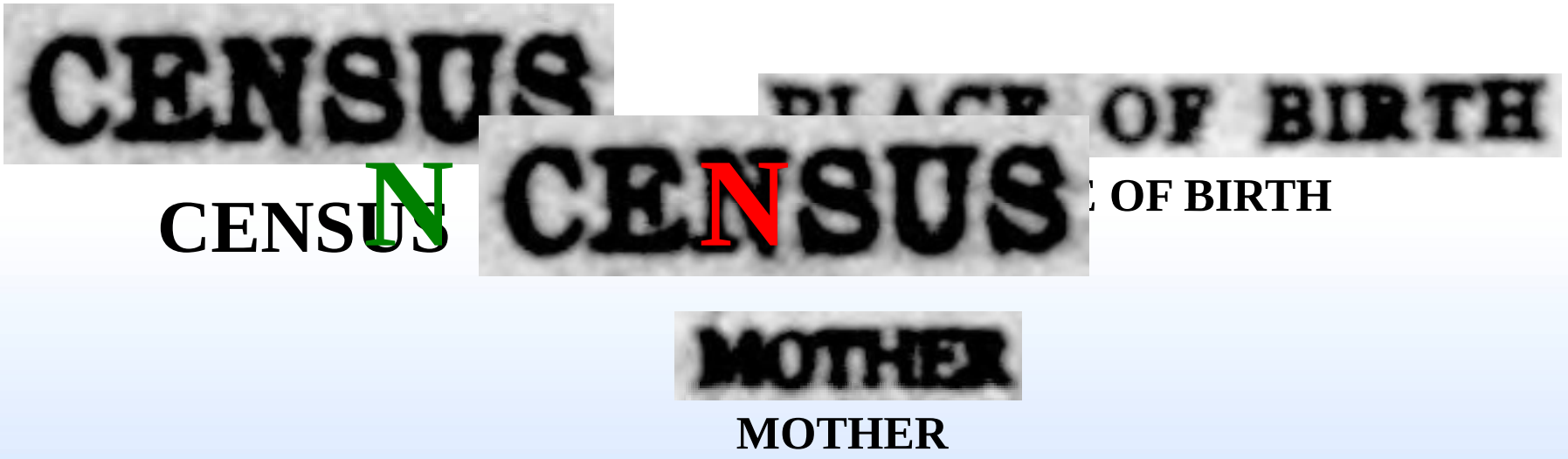
Image sizes

- 9312x6648
  - ~5 MB file
  - 61,906,176 pixels
- 5351x3635
  - ~500 KB file
  - 19,450,885 pixels
- 7984x5608
  - ~1.2 MB file
  - 44,774,272 pixels
- 6288x3704
  - ~600 KB file
  - 23,290,752 pixels

Solutions

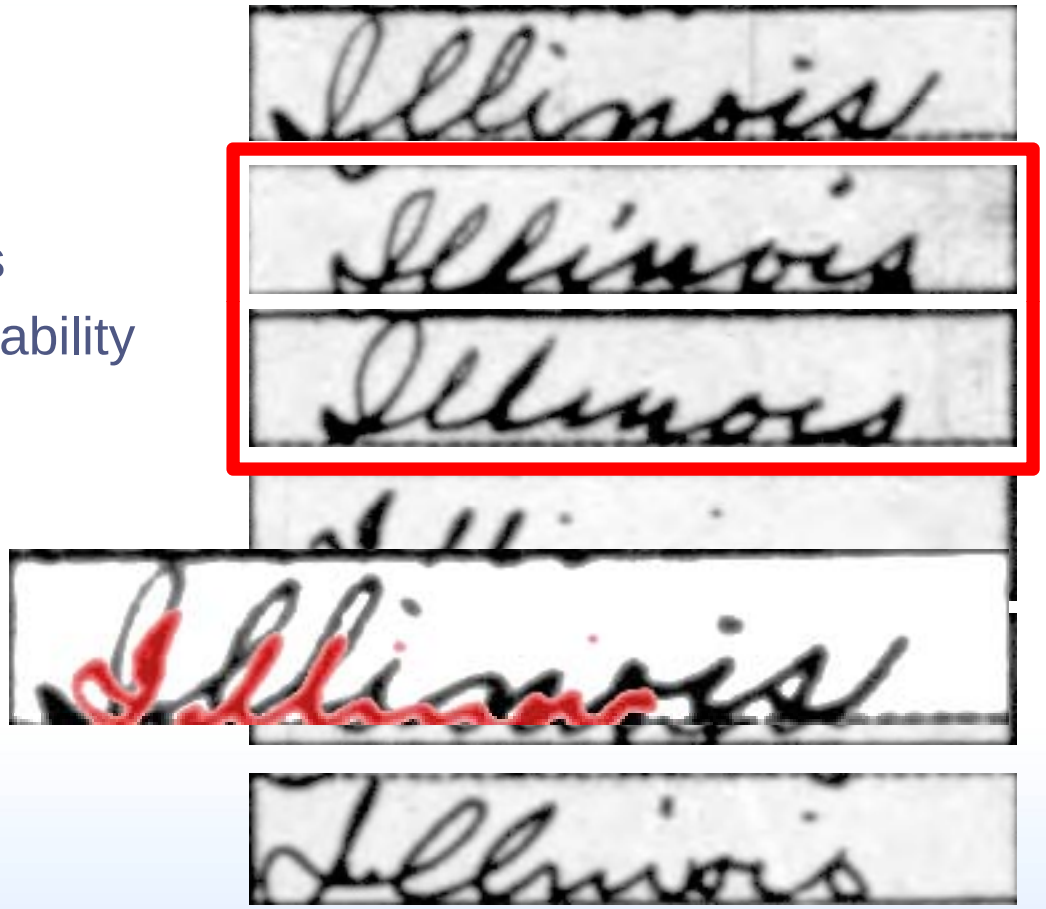
- OCR
  - Type written characters
    - Separation
    - Low variability
    - Pretty good these days
      - 71% - 98% [R. Holley, 2009]

a a a a a a



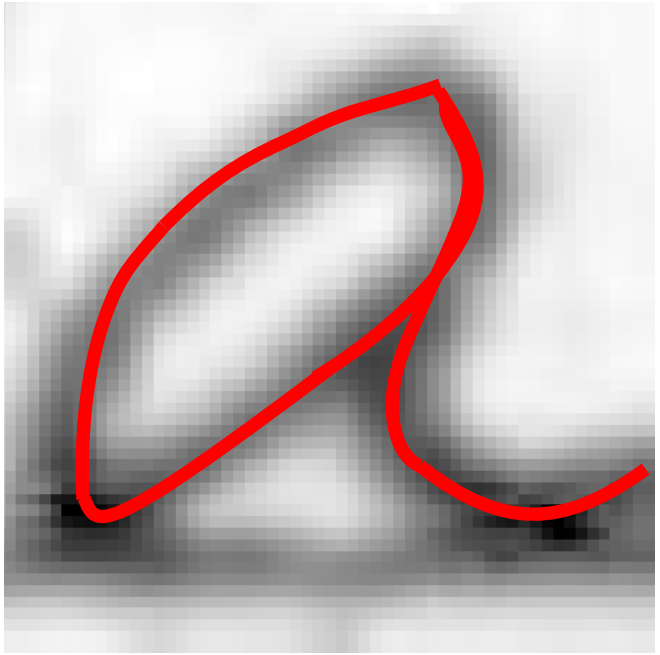
Solutions

- OCR
  - Type written characters
  - Hand written characters
    - Connected characters
    - Large amounts of variability
      - Writer
      - Slant
      - Shapes
      - Sloppiness
      - etc…



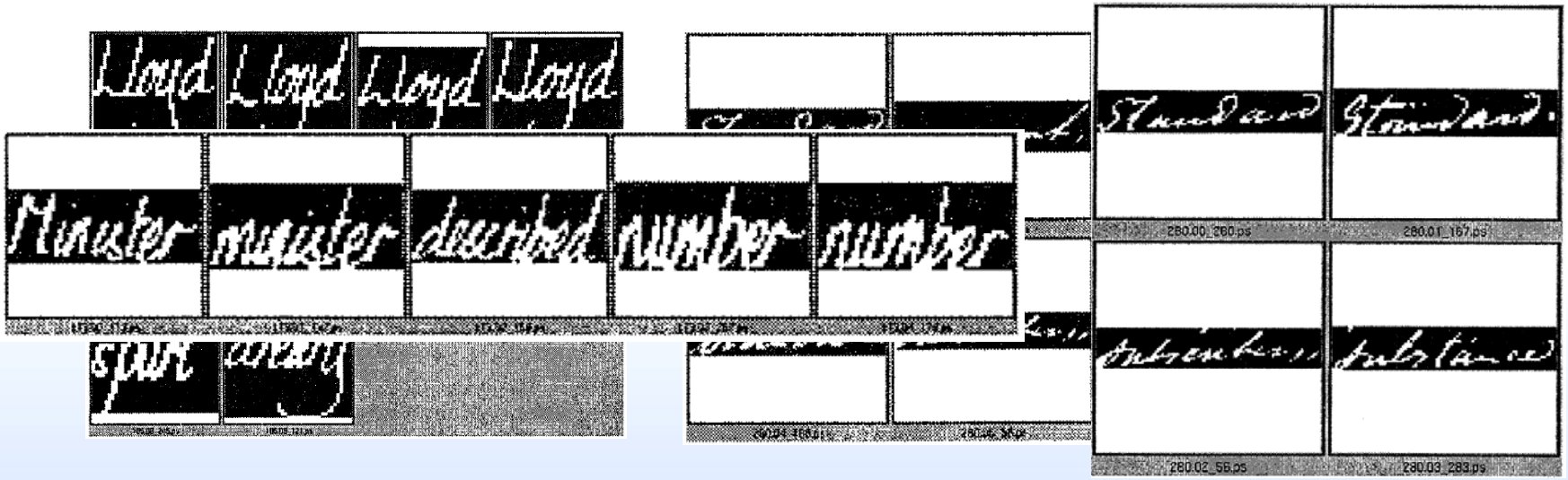
Solutions

- OCR
  - Type written characters
  - Hand written characters
    - Offline methods:
      - 95%, 85%, 75% for 10, 100, 1000 word lexicons [R. Plamondon, 2000]
      - Open research topic
    - Online methods:
      - 80% for 21,000 word lexicon
      - Temporal component
      - Much better accuracy
      - Used in mobile devices



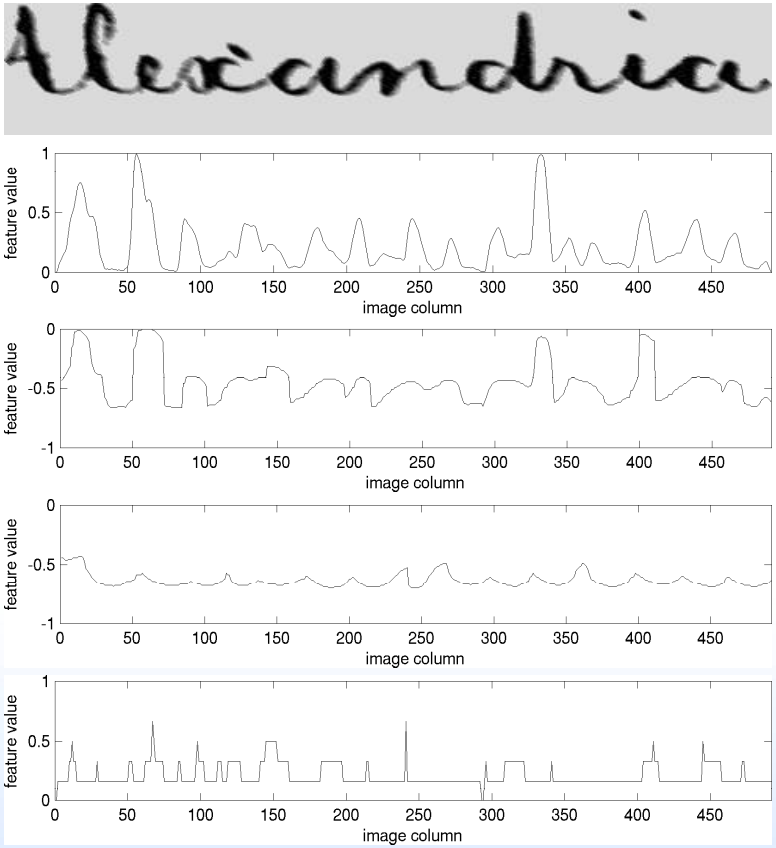
Handwriting Retrieval

- *R. Manmatha, et al.*, “Word Spotting: A New Approach to Indexing Handwriting”, CVPR, 1996
  - Index documents written by $_{\text{same}}$ person
  - Cluster segmented words based on a distance measure
    - Drop top *n* clusters assuming stop words
    - Present the next 2000 clusters to a user for transcription



Handwriting Retrieval

- *T. Rath, et al.,* "Word Image Matching Using Dynamic Time Warping", CVPR, 2003
  - Prune using area, aspect ratio, and decenders
  - Features
    - Projection profile (1)
    - Word profile (2)
    - Ink transitions (1)



Handwriting Retrieval

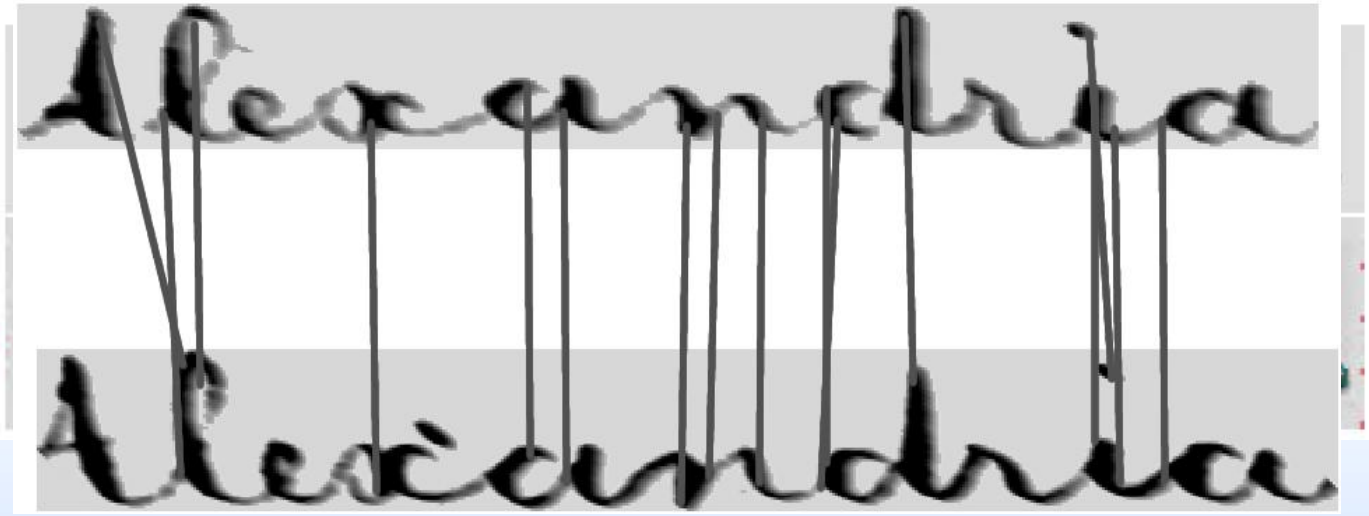
- *T. Rath, et al.,* "Word Image Matching Using Dynamic Time Warping", CVPR, 2003
  - Feature distance: squared Euclidean distance
  - Score obtained from dynamic time warping

XOR	SSD	SLH	SC	EDM	DTW	SC	EDM	DTW
54.14%	52.66%	42.43%	48.67%	72.61%	73.71%	40.58%	67.67%	67.92%

Algor.	XOR	SSD	SLH	SC	EDM	DTW
time [s]	13	72	121	~50	14	~2

Handwriting Retrieval

- *J. Rothfeder, et al.,* "Using Corner Feature Correspondences to Rank Word Images by Similarity", Workshop on Document Image Analysis and Retrieval, 2003
  - Use detected corners as features
  - Small windows around corners used to find matches
    - Constrained to be nearby
  - Score based on sum of distance between matching corners and number of correspondences



Handwriting Retrieval

- *J. Rothfeder, et al.*, "Using Corner Feature Correspondences to Rank Word Images by Similarity", Workshop on Document Image Analysis and Retrieval, 2003
  - Overall fastest but slightly less accurate than DTW

XOR	SSD	SLH	SC	EDM	DTW	CORR	SC	EDM	DTW	CORR
54.14%	52.66%	42.43%	48.67%	72.61%	73.71%	73.95%	40.58%	67.67%	67.92%	69.69%

Algorithm:	XOR	SSD	SLH	SC	EDM	DTW	CORR
Running time [s]:	13	72	121	~50	14	~2	~1

Handwriting Retrieval

- *J. Rodriguez-Serrano, et al.*, “A Similarity Measure Between Vector Sequences with Application to Handwritten Word Image Retrieval”, CVPR, 2009
    - Groups examples into a small sets containing similar features
    - Represents new examples based on these small sets

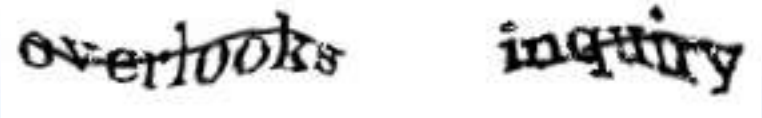
Solutions

- OCR
  - Type written characters
  - Hand written characters
- Human Intelligence
  - Transcription companies
    - High accuracy
    - Costs money
    - Human error
  - Crowd Sourcing
    - Amazon Mechanical Turk
      - Costs money
      - Varying accuracy [J. Downs, 2010]
    - reCAPTCHA

reCAPTCHA

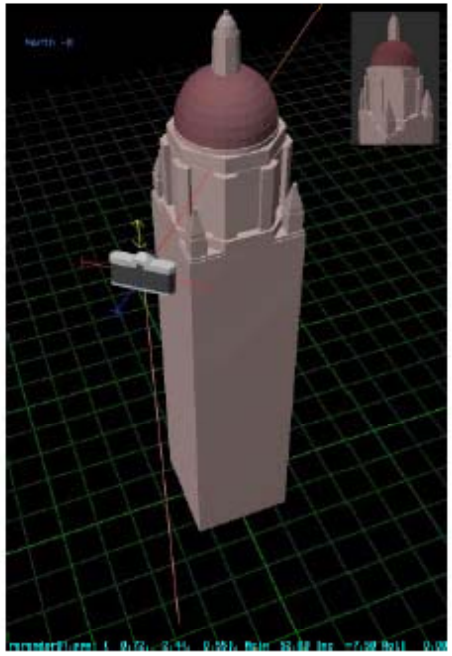
- CAPTCHA [L. von Ahn, 2003]
  - Distinguish computers and humans

- reCAPTCHA [L. von Ahn, 2008]
  - Distinguish computers and humans
  - Transcription



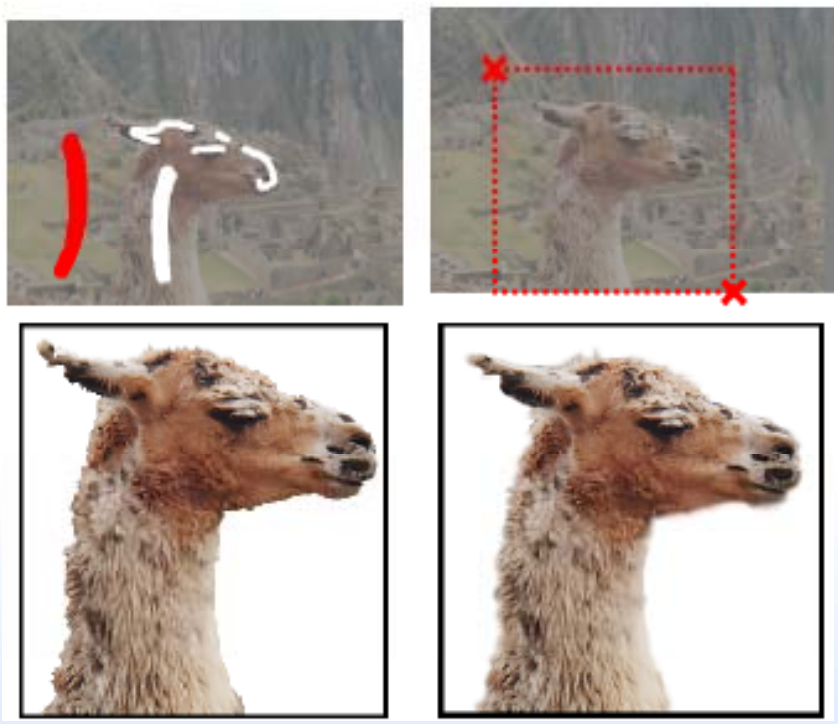
Human Intelligence

- 3D Reconstruction
  - [P. Debevec, 1996], [R. Cipolla, 1999]
  - Mark parrallel lines



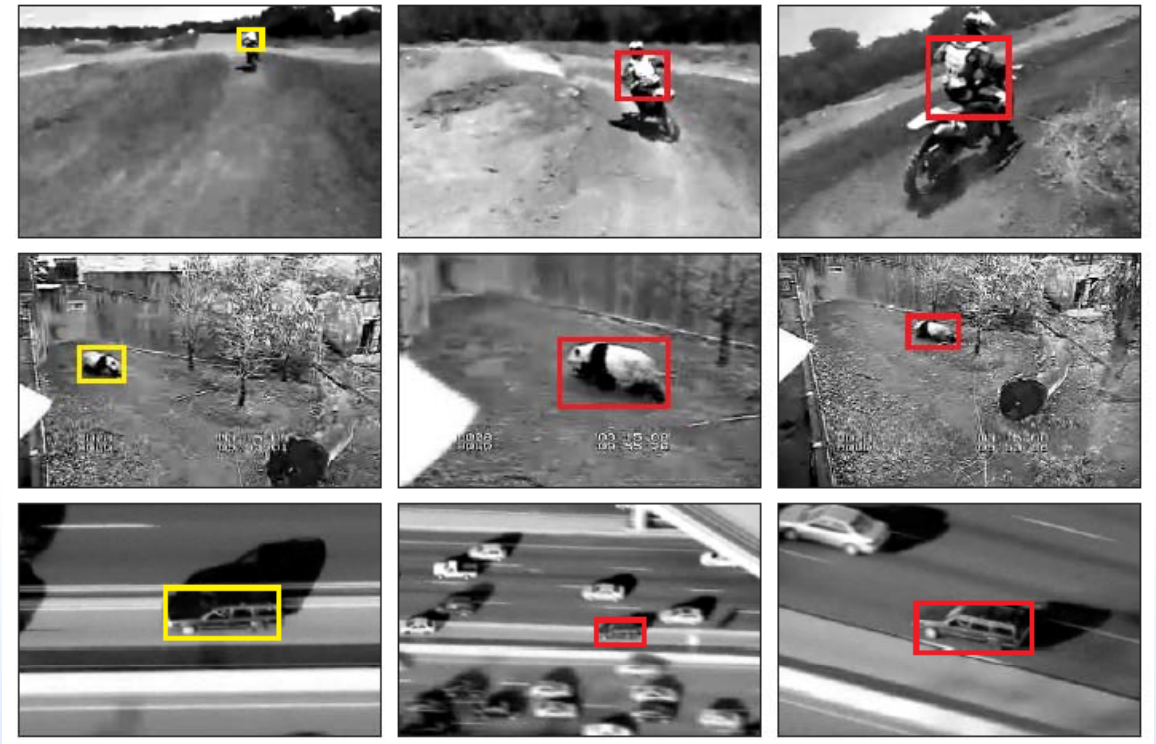
Human Intelligence

- Foreground/Background Segmentation
  - [Y. Boykov, 2001], [C. Rother, 2004], [A. Delong, 2011]
  - Mark examples of foreground and background



Human Intelligence

- Tracking
  - [Z. Kalal, 2010]
  - Mark object in a frame

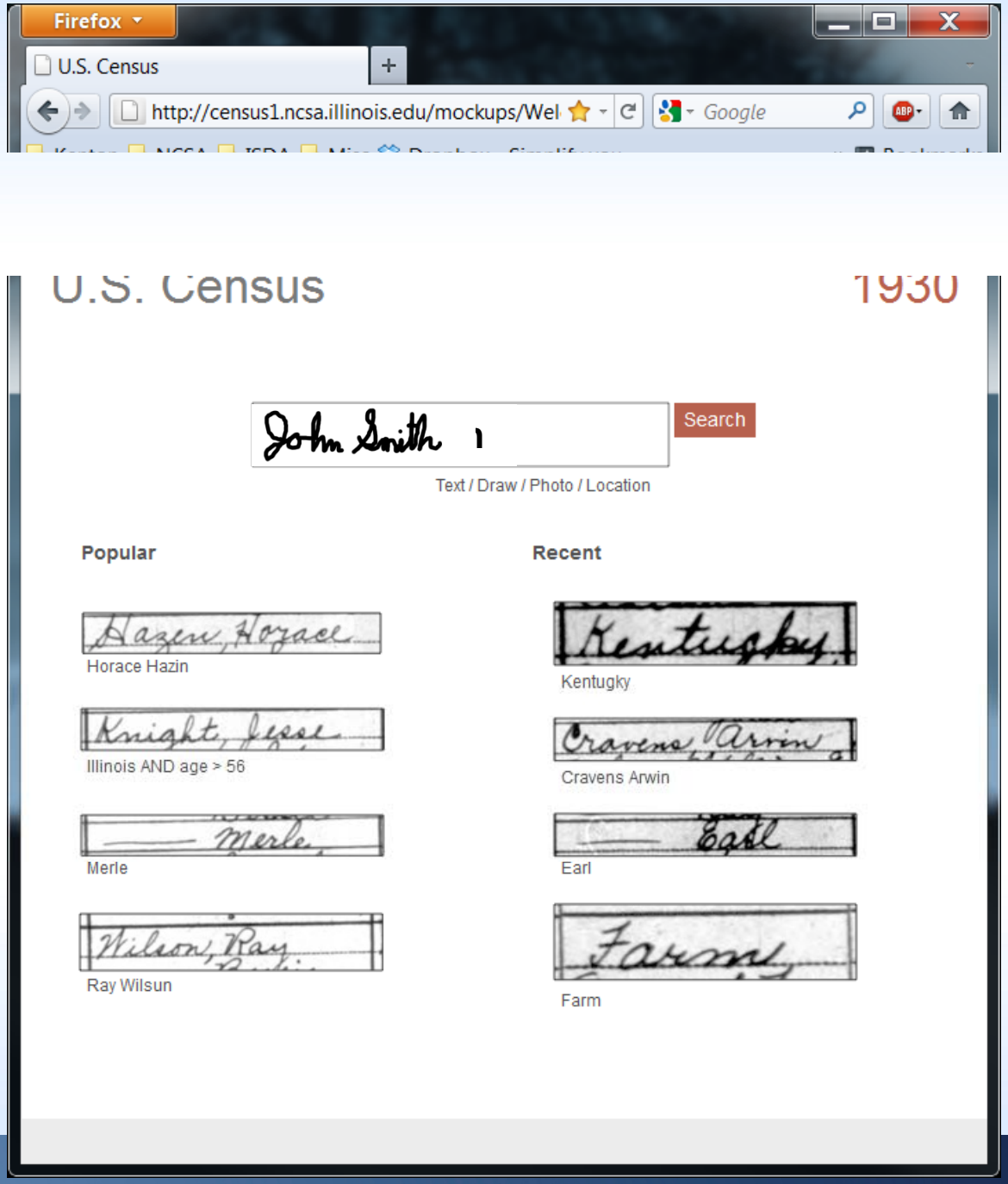


Unorthodox Interfaces [R. Veltkamp, 2000]

- Query image (one or more)
  - From database
  - From URL
  - Sub-images
  - Sub-region (segmented area)
  - Map area
- Clusters and key images
- Refinement
  - Successive filtering and query re-submission
  - Relevance labeling
- Color selection
  - List
  - Percentages
  - Values
  - Dominance hierarchy
- Texture selection
  - List
  - Sub-area
  - Values
- Boundary selection
- Location selection
- Sketching
  - Draw shape
    - Assign weights to convex parts
  - Select internal colors/texture
  - Select region in example image
  - Semantic queries: place icons containing meaning (e.g. faces)
  - Enforce relationships (e.g. next to, inside, outside, size)
- Comparison criteria
  - Feature selection
  - Measure selection (e.g. weighting features)
- Text
  - Keywords
  - Topic (NLP)
  - Boolean operations on feature values
  - SQL queries with keywords and values
  - Voice (voice recognition)

Solutions

- OCR
  - Type written characters
  - Hand written characters
- Human Intelligence
  - Transcription companies
  - Crowd Sourcing
- A Hybrid Approach


Firefox
U.S. Census
http://census1.ncsa.illinois.edu/mockups/Wel
Google
U.S. Census
1930
John Smith
Search
Text / Draw / Photo / Location
Popular
Recent
Horace Hazin
Illinois AND age > 56
Merle
Ray Wilsun
Kentugky
Cravens Arwin
Earl
Farm

Firefox

U.S. Census

http://census1.ncsa.illinois.edu/mockups/Results

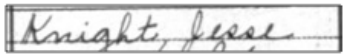
Google

Kenton NCSA ISDA Misc Dropbox - Simplify you... Bookmarks

Welcome Results Form recaptcha

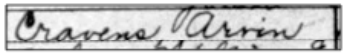
U.S. Census 1930

Showing search results for *Champaign* *(3 out of 5609)*



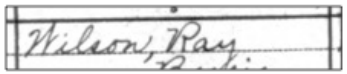
Similar

State: Illinois
Reel: 45
Page: 4
Enumerated District: 10-29
Top 3 Transcriptions:
1. Knight Jessi
2. Kuight, Jesse
3. Kuigtt, Jesse



Similar

State: Michigan
Reel: 11
Page: 2
Enumerated District: 5-13
Top 3 Transcriptions:
1. Cravens Arwin
2. Cmvens Arvin
3. Crvenv Arvini

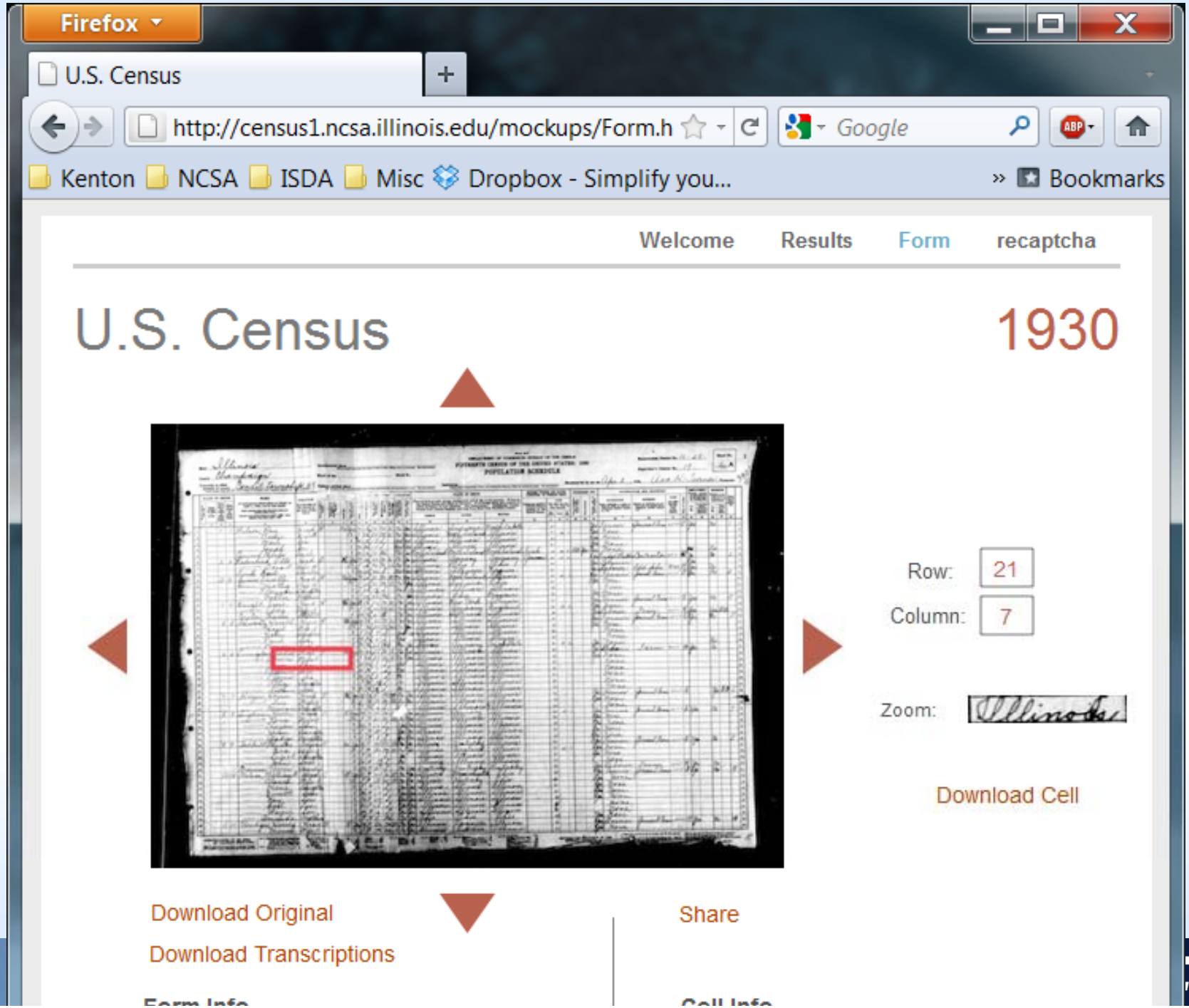


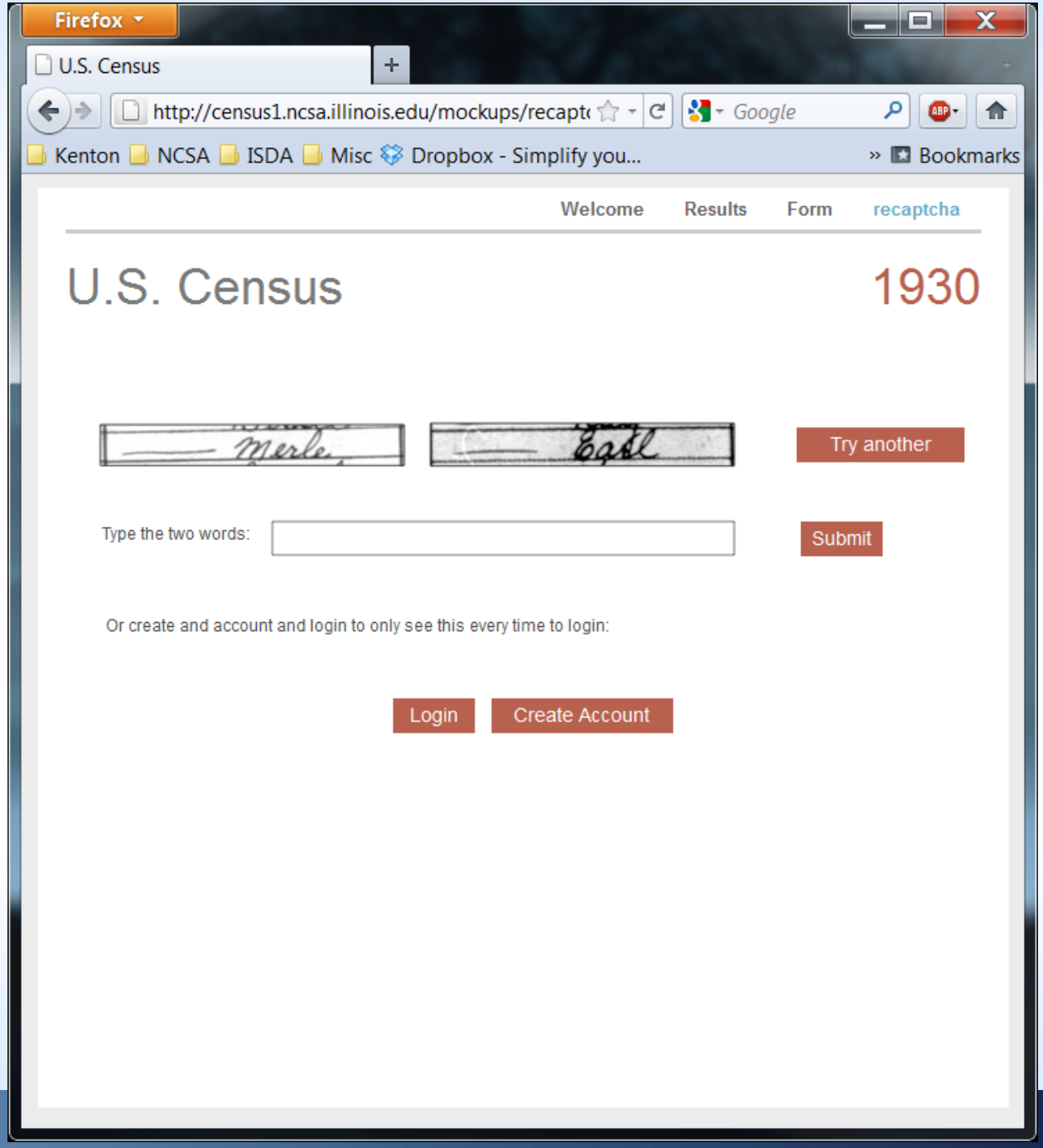
Similar

State: Illinois
Reel: 13
Page: 7
Enumerated District: 11-4
Top 5 Transcriptions:
1. Wilson, Ray
2. Wilsun, Raj
3. Wulsuu, Ray

More

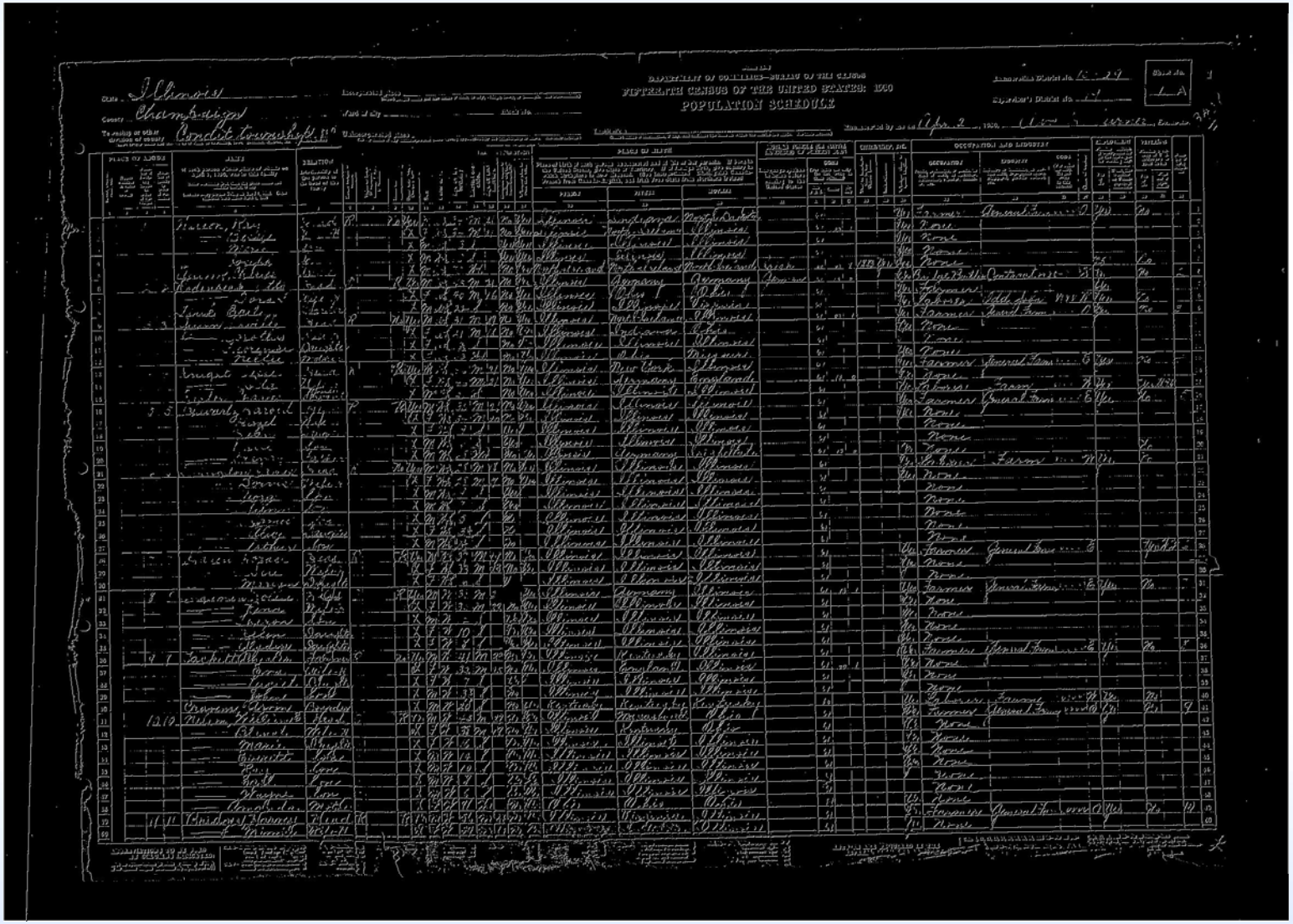
NCSA


Firefox
U.S. Census
http://census1.ncsa.illinois.edu/mockups/Form.h
Google
Kenton
NCSA
ISDA
Misc
Dropbox - Simplify you...
Bookmarks
Welcome
Results
Form
recaptcha
U.S. Census
1930
Row:
21
Column:
7
Zoom:
Download Cell
Download Original
Download Transcriptions
Share

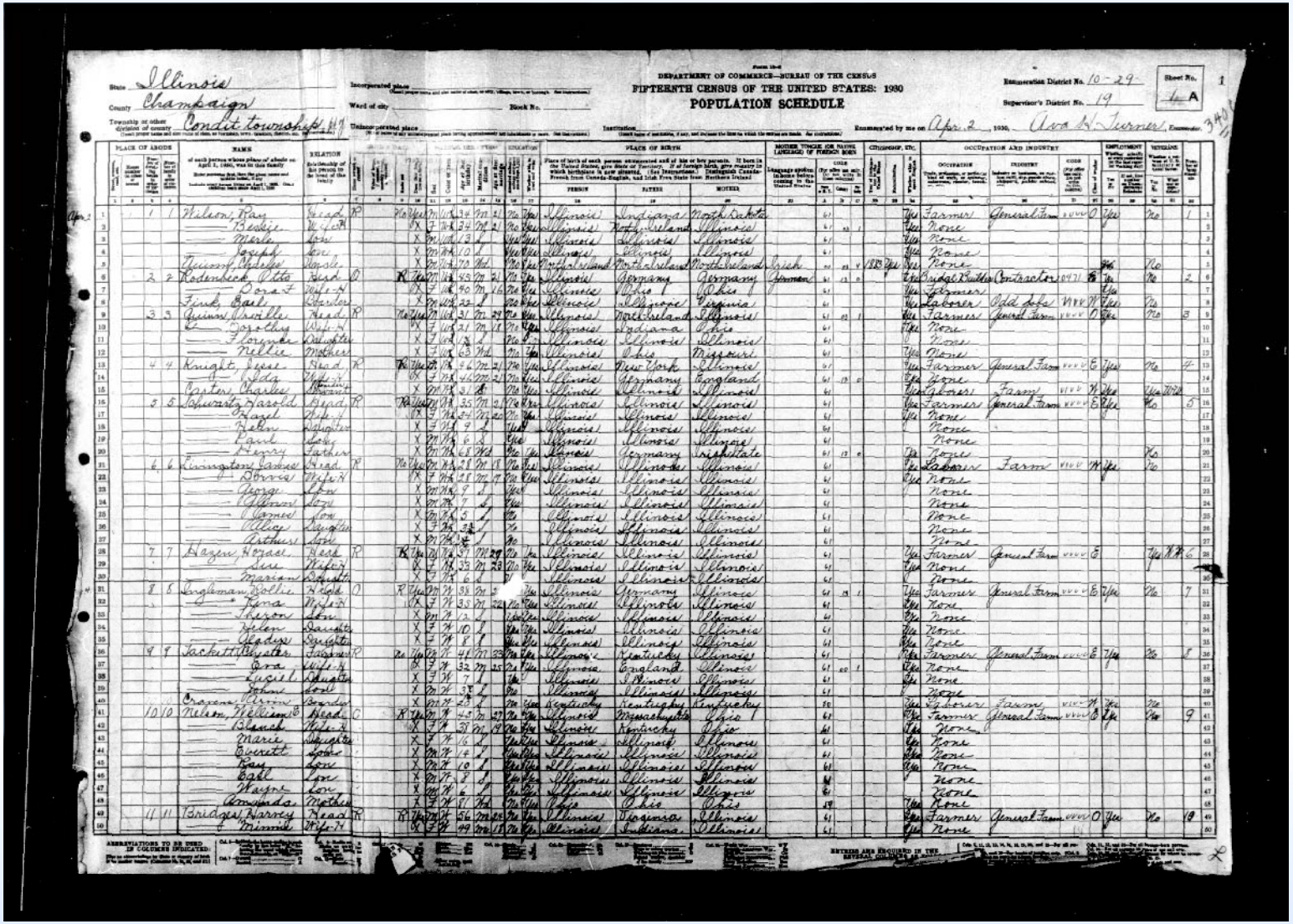

Firefox
U.S. Census
http://census1.ncsa.illinois.edu/mockups/recapt
Google
Kenton
NCSA
ISDA
Misc
Dropbox - Simplify you...
Bookmarks
Welcome
Results
Form
recaptcha
U.S. Census
1930
Merle
Try another
Type the two words:
Submit
Or create and account and login to only see this every time to login:
Login
Create Account

Form Segmentation

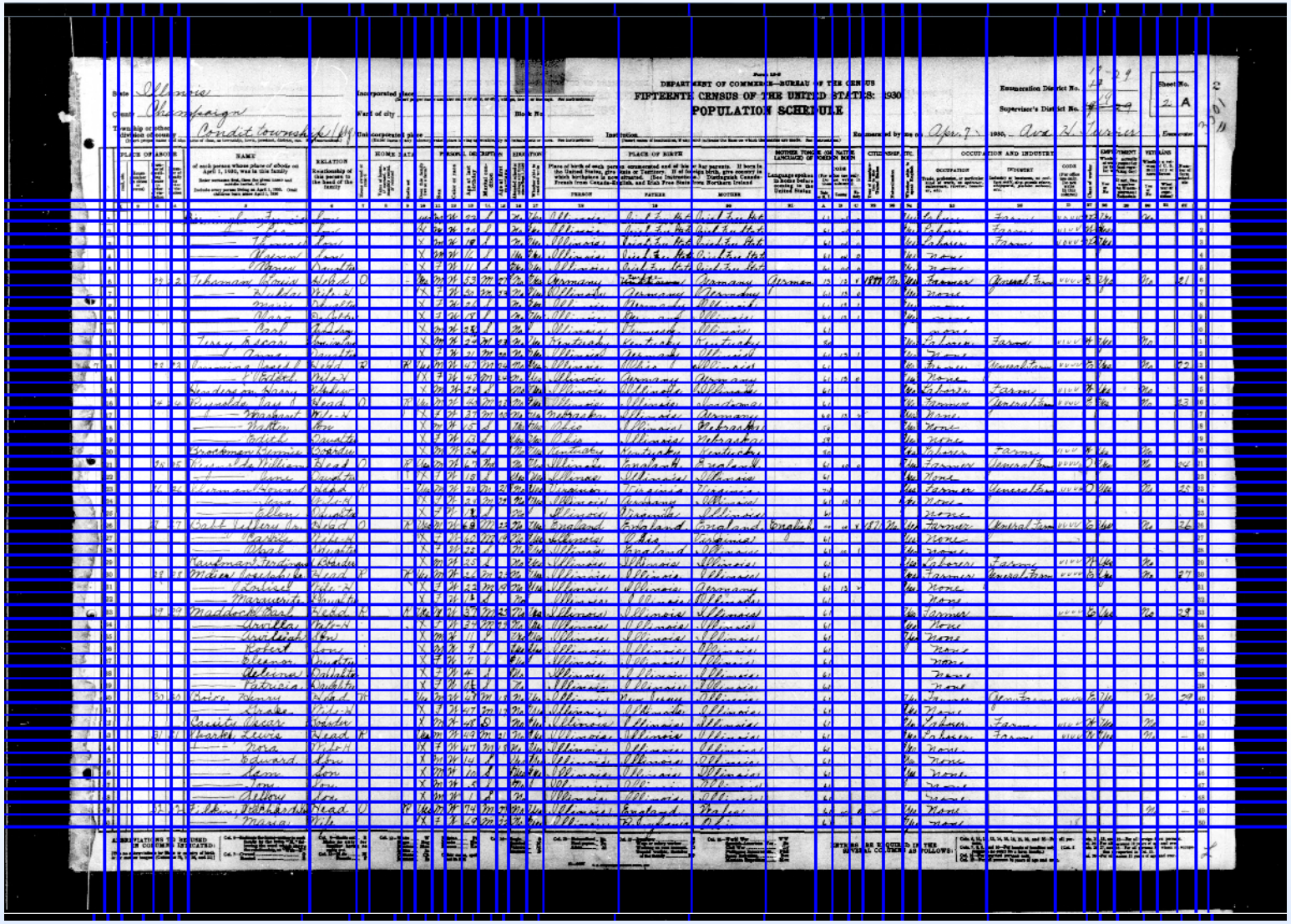
[R. Casey, 1990] [J. Liu, 1995]



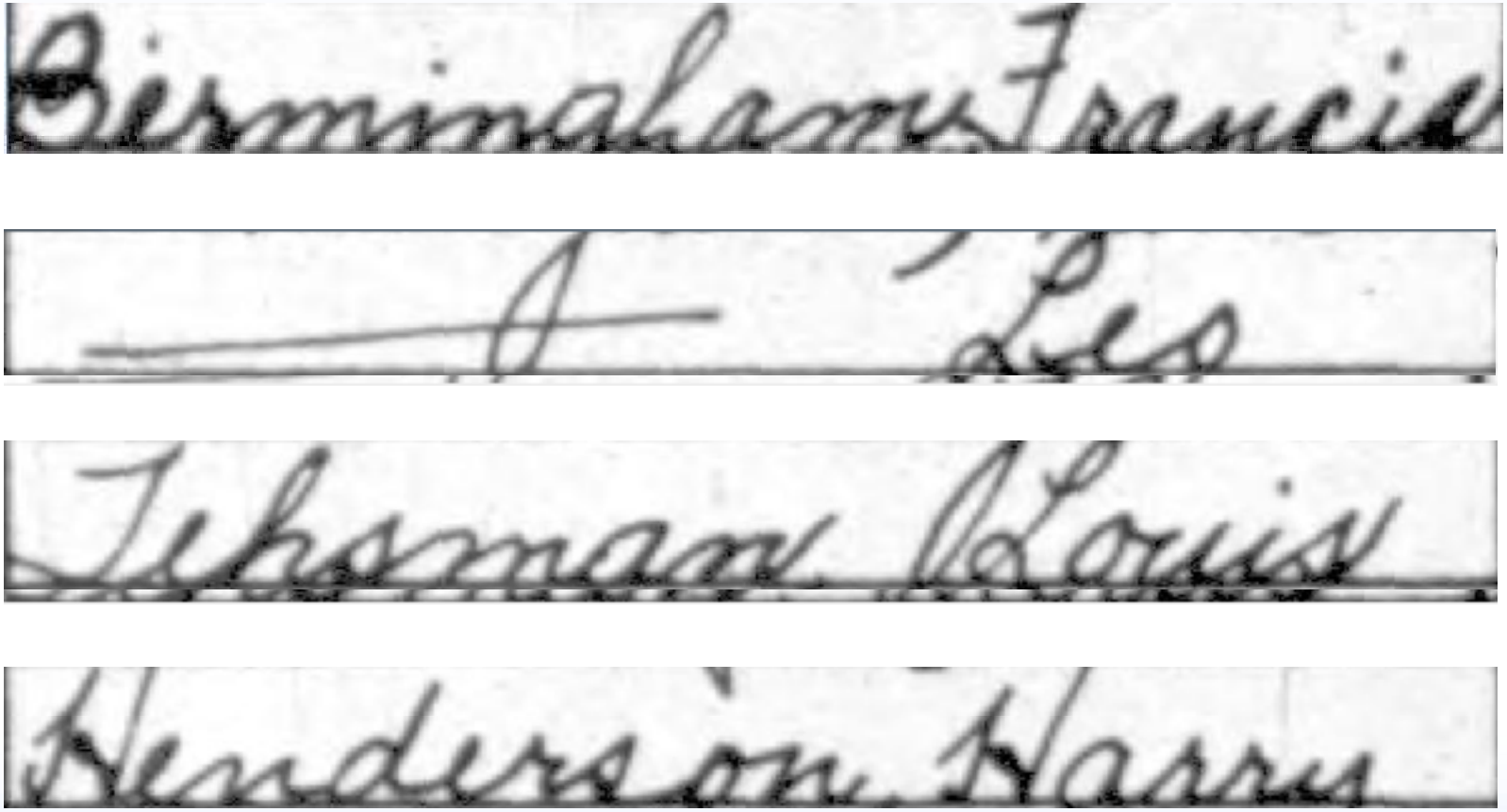
Form Segmentation



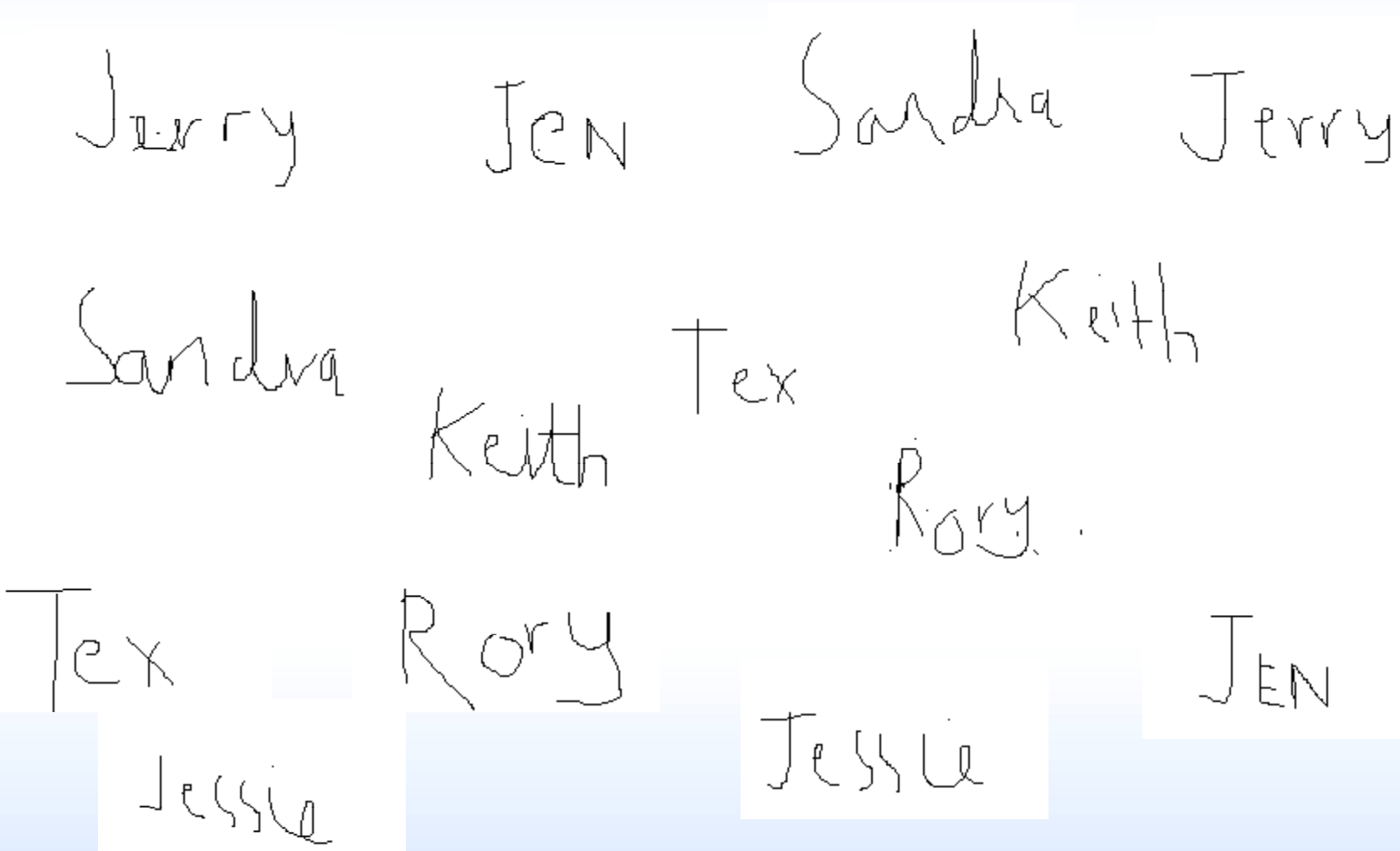
Form Segmentation



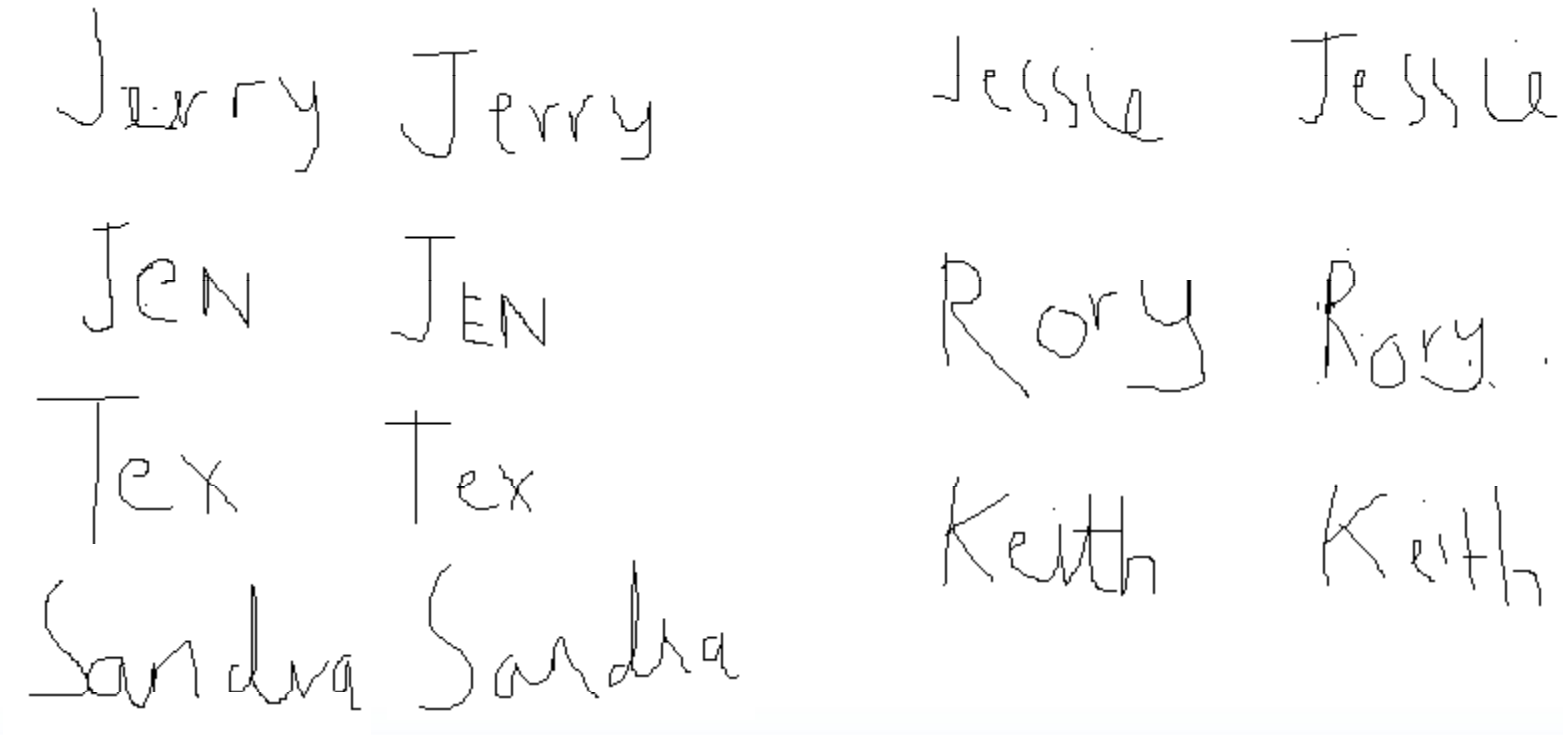
Form Segmentation



Handwriting Retrieval



Handwriting Retrieval



Handwriting Retrieval

Indexing

- DTW
  - Flexible with regards to writing variations
  - Non-linear (i.e. features not independent)
    - Costly retrievals
- Linear features
  - Less accurate
  - Can build a tree of indices for efficient retrieval

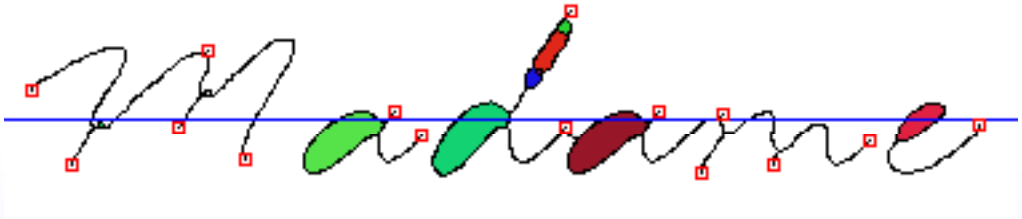


Image Pyramid

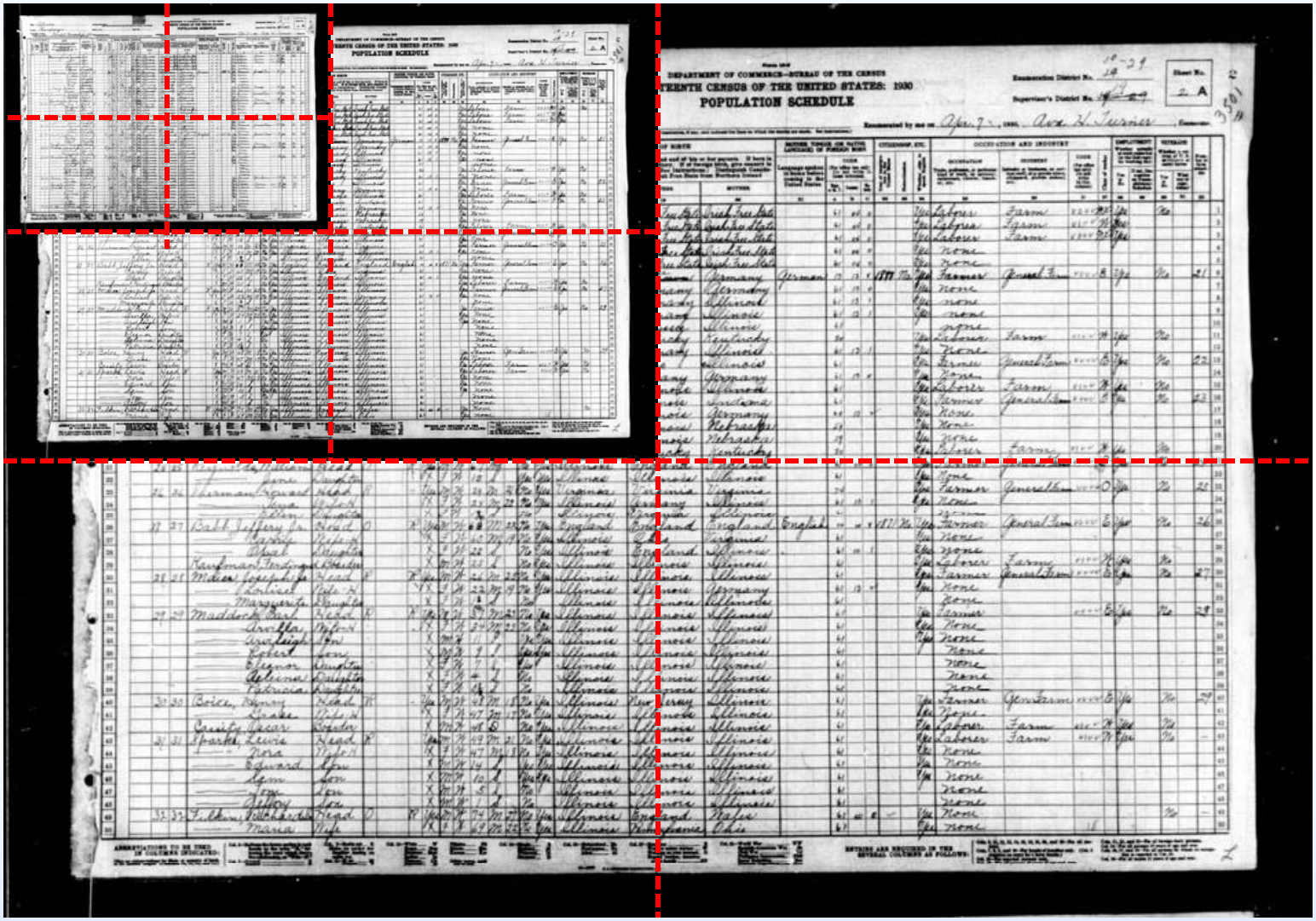
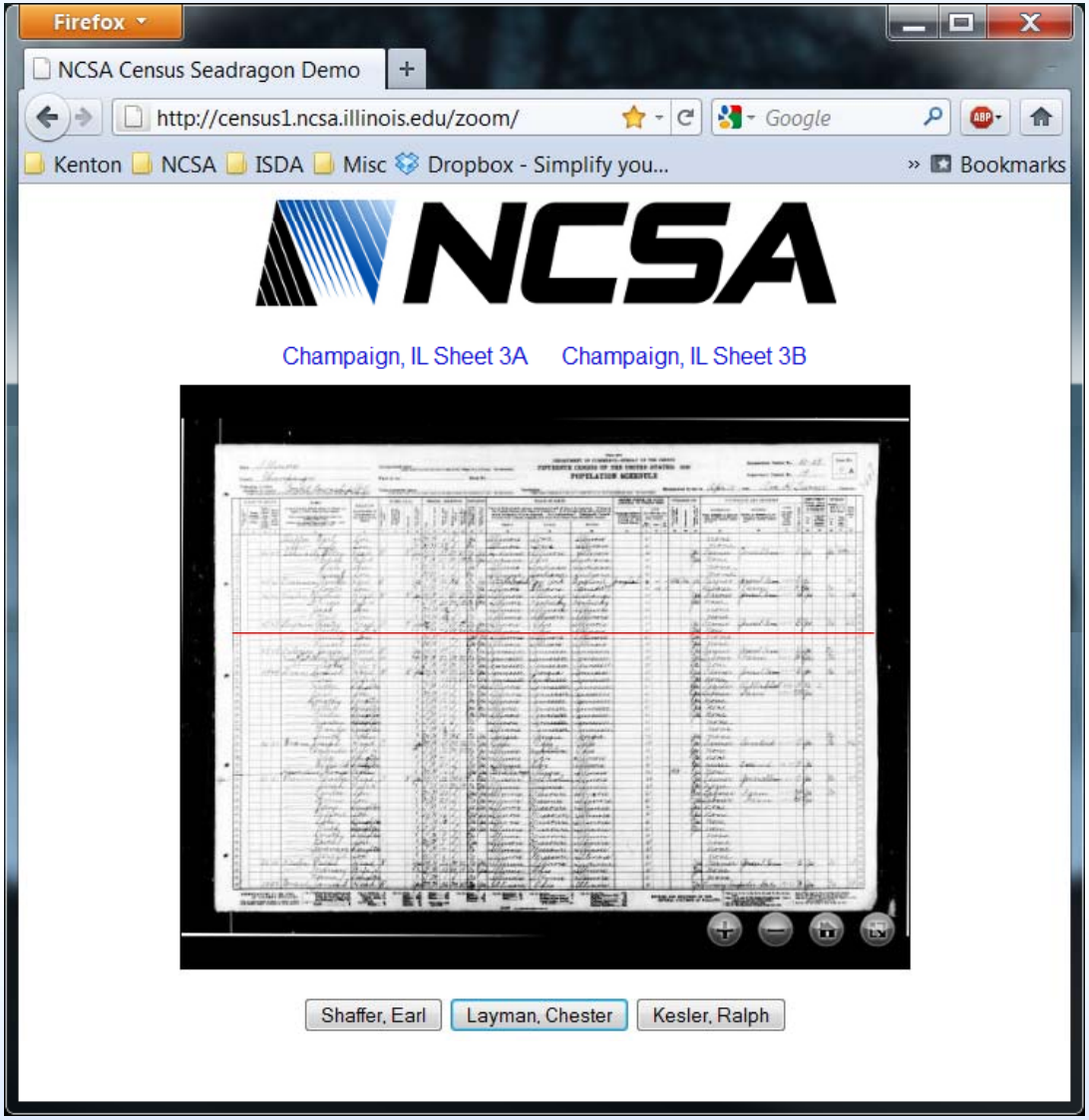


Image Pyramid



Data

- Images
  - **7.0 TB**
- Cell boundaries
  - $(52 + 39)_{x\ 2\ b}$ ytes = 4,056 bytes
  - **14.7 GB** for all images
- Index
  - 50 x 38 cells (1900)
  - 8 x 4 bytes per feature vector
  - **220.3 GB**
- Annotations
  - 10 characters per annotation
  - Top 5 annotations
  - **344.2 GB**
- Image pyramids
  - 1.3 times original size
  - **2.1 TB**

Total of 9.7 TB
or
~40% more data

Acknowledgements

- *This research was supported by a National Archive and Records Administration (**NARA**).*

- *The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the National Science Foundation, the National Archive and Records Administration, or the U.S. government.*

Free and Searchable Access to the 1940 Census Data

Image, Spatial, and Data Analysis Group

http://isda.ncsa.illinois.edu

Kenton McHenry
Rob Kooper
Michal Ondrejcek
Luigi Marini
Peter Bajsy