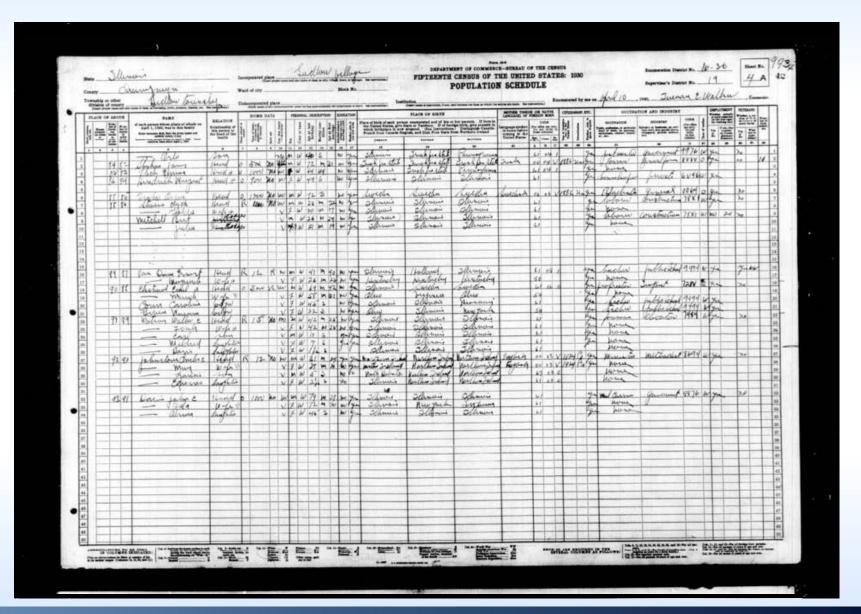


The Problem

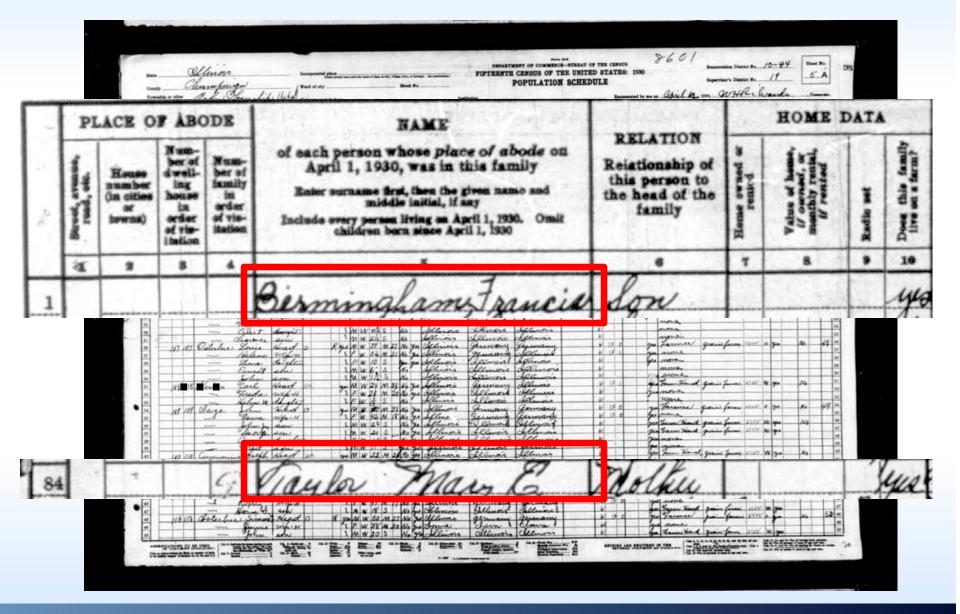




The Problem

~3.6 million more...

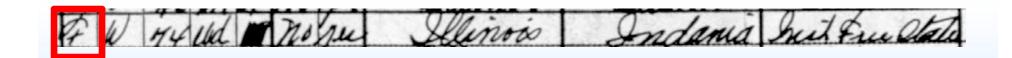






P	ERSON	IL DES	the United States, give State or Territory. If of foreign birth, a which birthplace is now situated. (See Instructions.) Disting												
	r of race	at last	dition	at first	led school or see any time e Sept. 1, 1929	face able to	Place of birth of each person enumerated and of his or her parents. If born is the United States, give State or Territory. If of foreign birth, give country is which birthplace is now situated. (See Instructions.) Distinguish Canada French from Canada-English, and Irish Free State from Northern Ireland								
Sex	Colo	Ag	×	A B	A Selection	71	PERSON	PATRER	MOTHER						
11	12	13	14	15	16	17	18	19	20						
m	H	22	8		76	Yes	Illinois	Irish Fry State	Irish Free State						

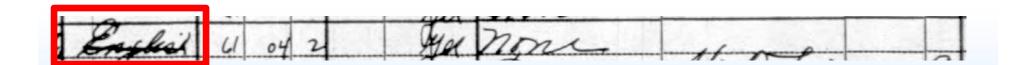
. . .





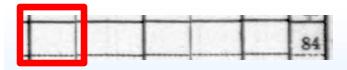
MOTHER TONG LANGUAGE) OF	UE (OR POREK	NATITO	VE RN	CITIZE	NSHIP	, ETC.	OCCUPA	TION AND INDUSTRY	15 44	
Language spoken in home before coming to the	(For or	CODE	only.	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	allastion	be obto	OCCUPATION Trade, profession, or particular kind of work, as spinner, salesman, riveter, teach-	INDUSTRY Industry or business, as cotton mill, dry-poods store, shippard, public school,	(For office use only. De not write	. 1
United States	Sin.	C	No.	F. D	Mate	1	er, etc.	elci	in this column)	8
	A	В	C	22	28	94	25	26	D	27
	61	04	0	in.		Wes	Laborer	Farm	VOVV	AZ

- - -





	Num- ber of farm	RANS of U.S. fary or forces	VETE Whether or no.	OYMENT or actually rk yesterday o last regu- orking day)	Wheth
	nie nie	Visit of the control	Yes No	If mpt, line number on Unem- ployment Schodule	Yes Xe
	32	81	36	20	96
1			No		7/2





The Information

	PI	ACE O	F ABC	DE	NAME		00	HOME	DATA	
2	Street, avenue, read, etc.	House number (in cities or towns)	Number of dwelling house in order of vis-	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 an April 1, 1930. Omit children born since April 1, 1930.	RELATION Relationship of this person to the head of the family	Beme owned or rented	Value of home, U corned, or monthly rental, U rented	Radio set	Does this family live on a farm?
	a	9	8	4	8	6	т	8	9	10
1					Berningham, Francis	low				we

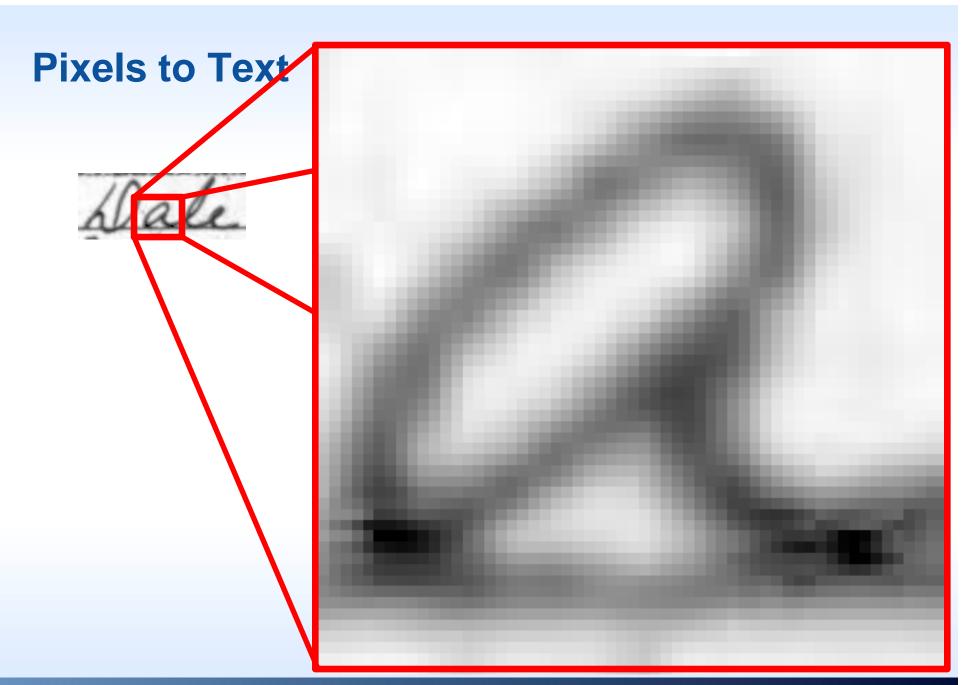
Birmingham, Francis

Son

Yes

P	RSON	L DES	CRIPTI	ON .	EDUC	ATION		PLACE OF BIRTH	
	r of race	at last	dital con-	at Arst	sed school or see any time Sept. 1,1929	ther able to	the United States, give	Son enumerated and of his of State or Territory. If of forew situated. (See Instructioninglish, and Irish Free State for	eign birth, give country in s.) Distinguish Canada-
Sex	Colon	A A A A A A A A A A A A A A A A A A A		11	PERSON	PATHER	MOTHER		
11	19	13	14	15	16	17	18	19	20
m	H	22	8		76	Yes	Illinois	Irish Fry State	Irish Free State
M	W	22	S		No	Yes	Illinois	Irish Free State	Irish Free State







Pixels to Text 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 0.3 0.2 0.2 0.3 1.0 1.0 0.2 1.0 1.0 1.0 0.3 0.2 0.3 1.0 1.0 1.0 0.2 0.3 1.0 1.0 0.3 0.2 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 0.3 0.3 1.0 1.0 1.0 0.2 1.0 1.0 1.0 0.3 0.2 1.0 1.0 1.0 0.3 0.3 1.0 1.0 1.0 0.3 1.0 0.2 0.3 1.0 1.0 1.0 1.0 1.0 1.0 0.3 0.3 1.0 1.0 1.0 0.2 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 NC5A

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



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

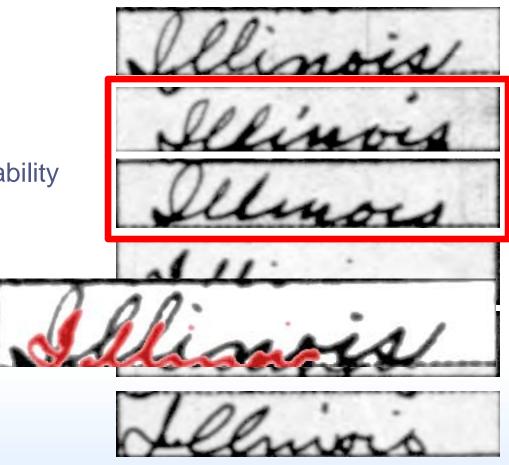
aaaaaa





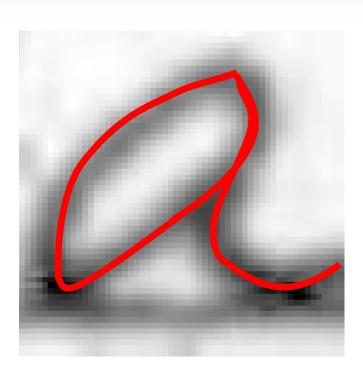


- OCR
 - Type written characters
 - Hand written characters
 - Connected characters
 - Large amounts of variability
 - Writer
 - Slant
 - Shapes
 - Sloppiness
 - etc...



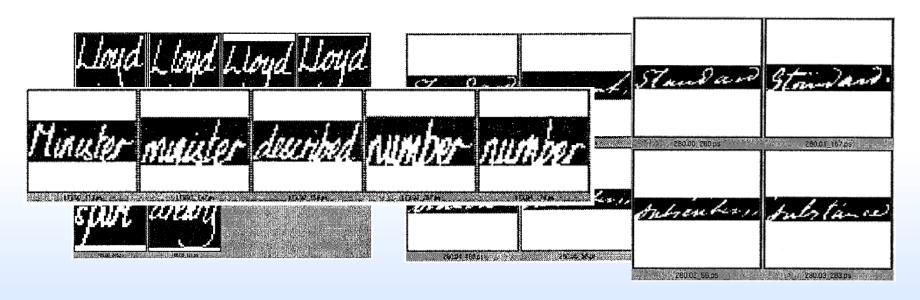


- OCR
 - Type written characters
 - Hand written characters
 - Offline methods:
 - 95%, 85%, 75% for 10, 100, 1000 word lexicons [R. Plamondon, 2000]
 - Open research topic
 - Online methods:
 - 80% for 21,000 word lexicon
 - Temporal component
 - Much better accuracy
 - Used in mobile devices



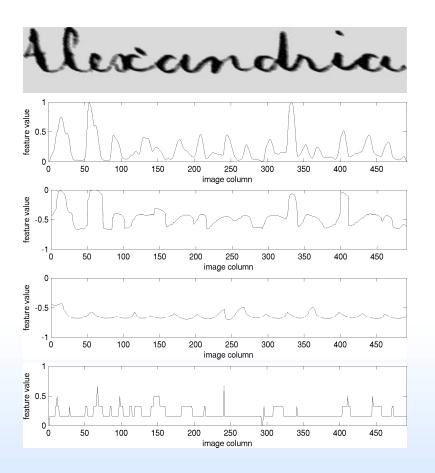


- R. Manmatha, et al., "Word Spotting: A New Approach to Indexing Handwriting", CVPR, 1996
 - Index documents written by person
 - Cluster segmented words based on a distance measure
 - Drop top *n* clusters assuming stop words
 - Present the next 2000 clusters to a user for transcription





- 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)





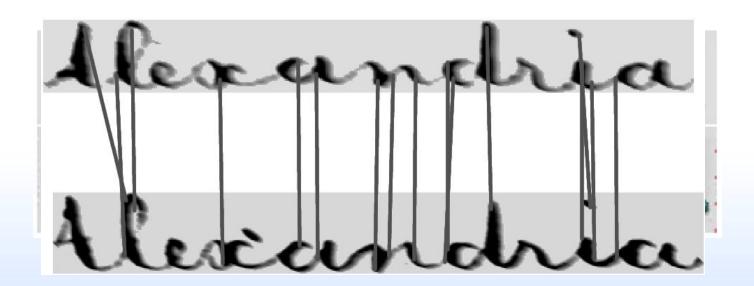
- T. Rath, et al., "Word Image Matching Using Dynamic Time Warping", CVPR, 2003
 - Feature distance: squared Euclidean distance
 - Score obtained from dynamic time warping

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	\sim 50	14	\sim 2



- 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





- 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



- 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





- 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

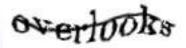


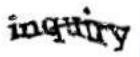
reCAPTCHA

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



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

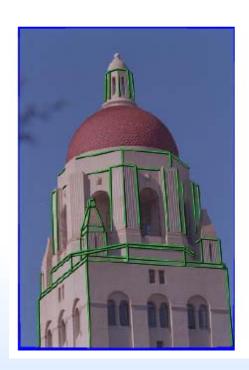


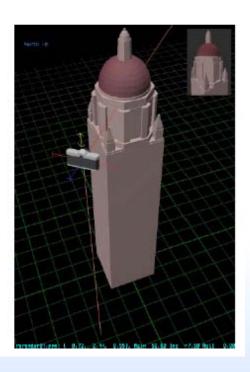




Human Intelligence

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

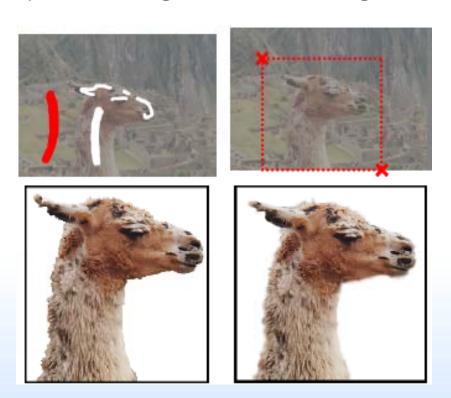






Human Intelligence

- 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]

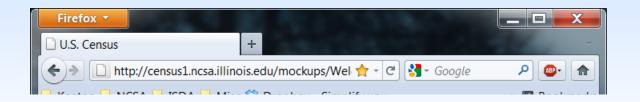
- Query image (one or more)
 - From database
 - From URL
 - Sub-images
 - Sub-region (segmented area)
 - Map area
- Clusters and key images
- Refinement
 - Successive filtering and query re-submission
 - Relevance labeling
- Color selection
 - List
 - Percentages
 - Values
 - Dominance hierarchy
- Texture selection
 - List
 - Sub-area
 - Values
- Boundary selection
- Location selection

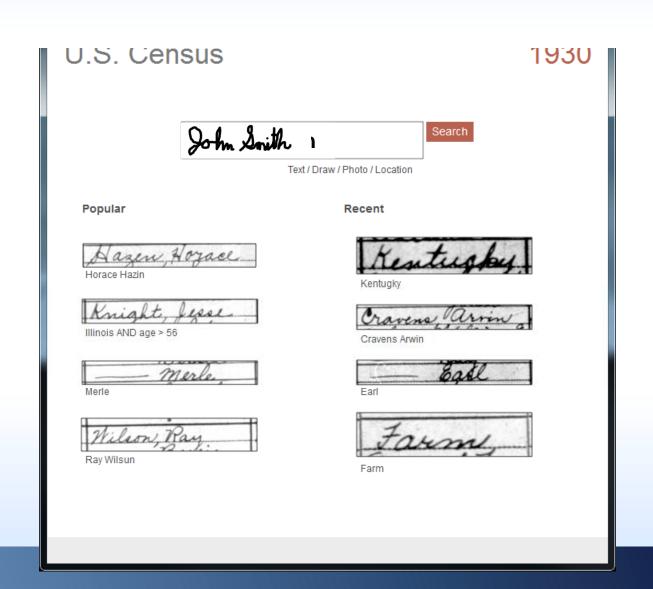
- Sketching
 - Draw shape
 - Assign weights to convex parts
 - Select internal colors/texture
 - Select region in example image
 - Semantic queries: place icons containing meaning (e.g. faces)
 - Enforce relationships (e.g. next to, inside, outside, size)
- Comparison criteria
 - Feature selection
 - Measure selection (e.g. weighting features)
- Text
 - Keywords
 - Topic (NLP)
 - Boolean operations on feature values
 - SQL queries with keywords and values
 - Voice (voice recognition)



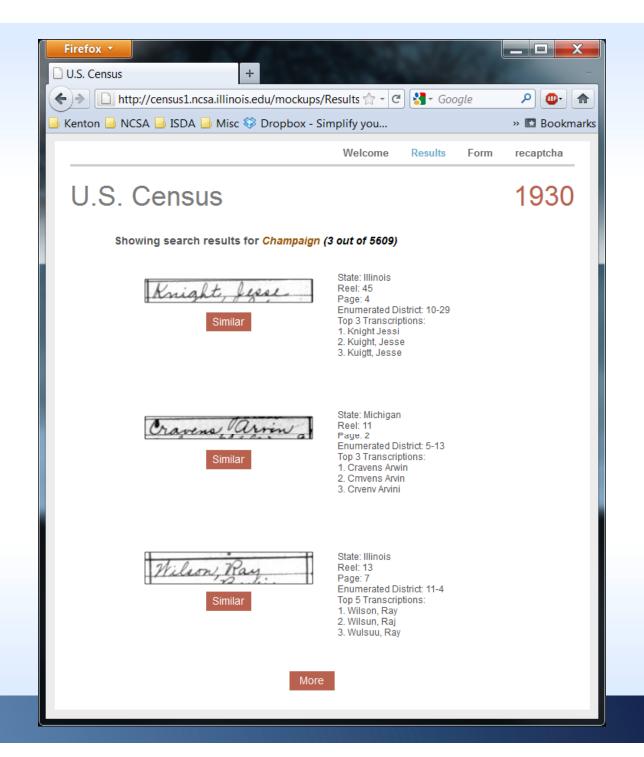
- OCR
 - Type written characters
 - Hand written characters
- Human Intelligence
 - Transcription companies
 - Crowd Sourcing
- A Hybrid Approach



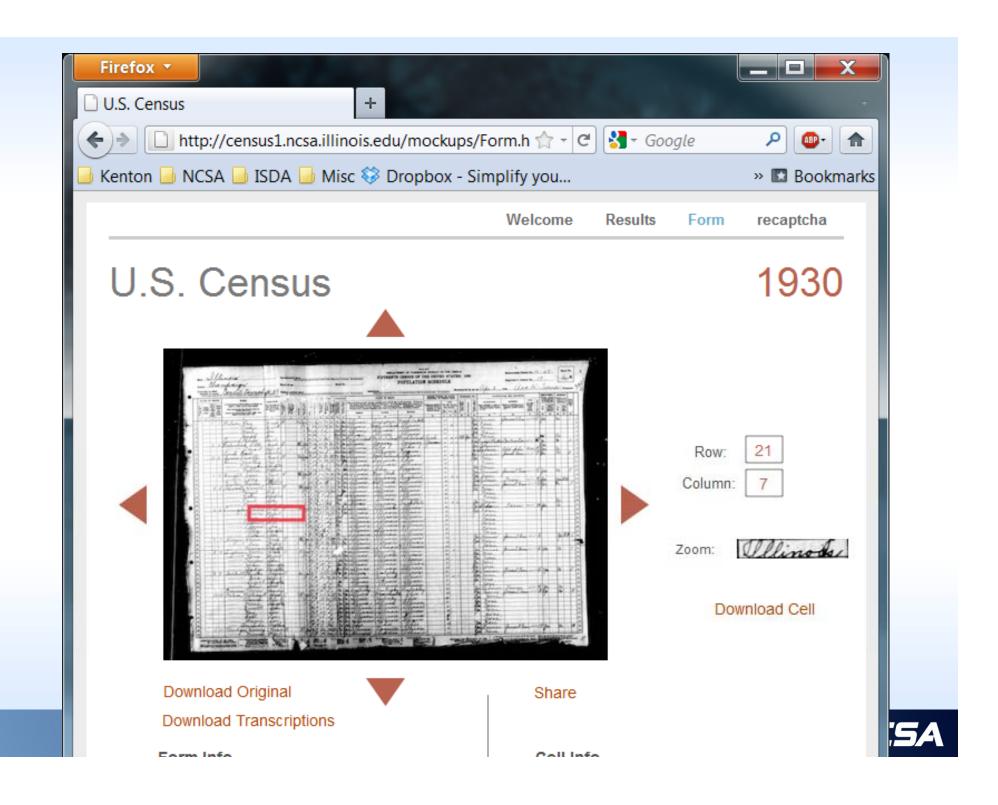


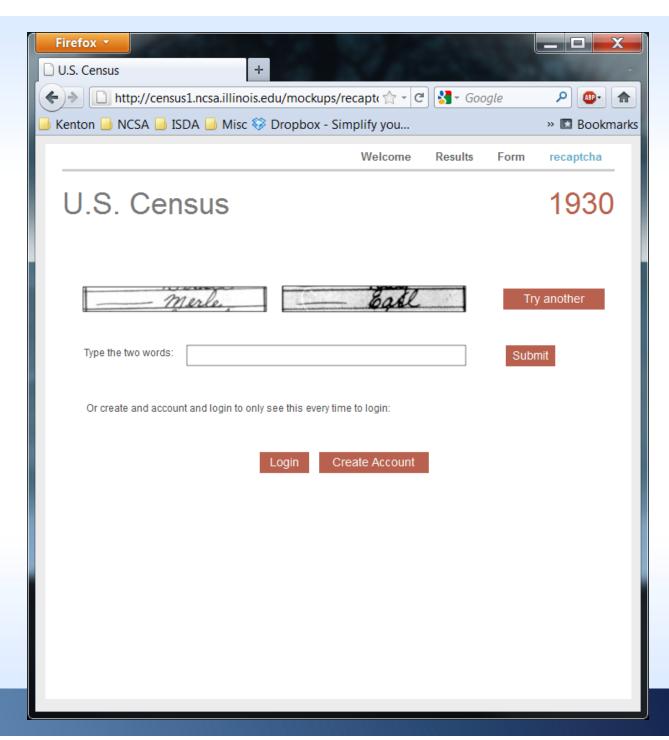










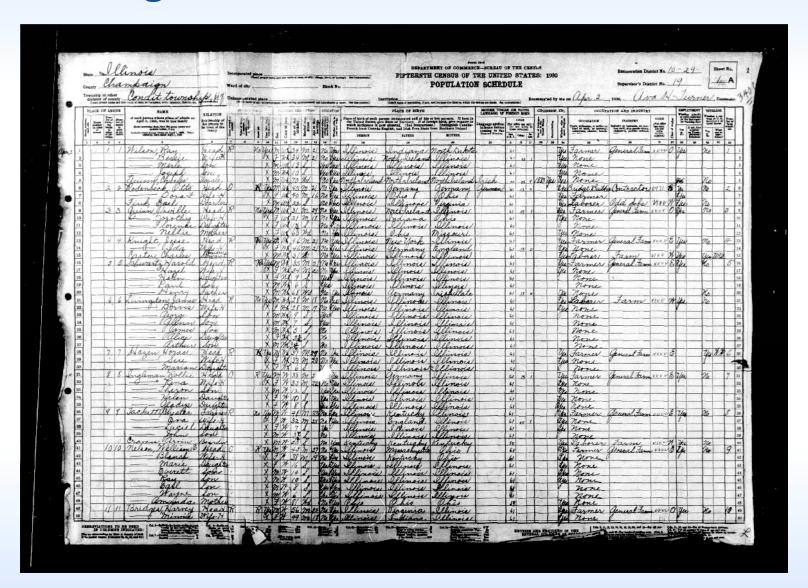




[R. Casey, 1990] [J. Liu, 1995]



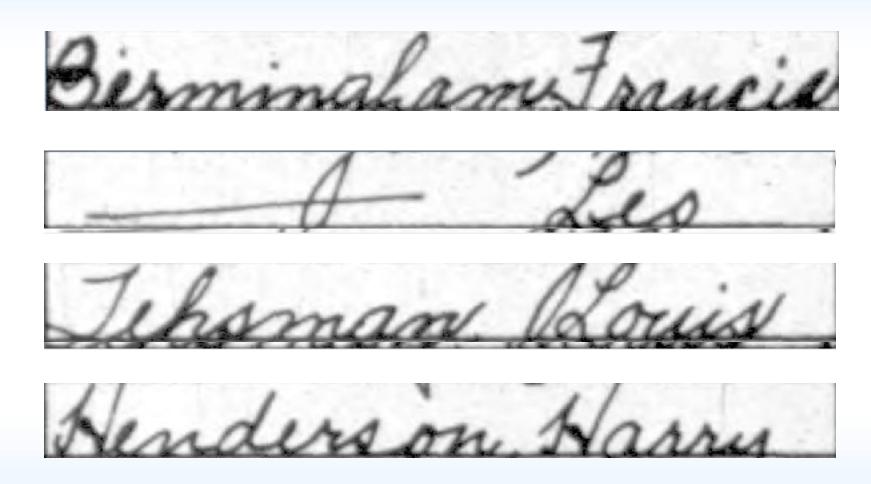






		n ereka kan									8 8 ×	شدات						100 666									
	B etc	Oh.	22	paign		incorporated of Fact of City .	lace ar a par				The same of the sa	i fi	- Nor	DEPART FIFTEENTI	CENSUS OF 1	HE UNIT			1930	trangers	Enumeration Dis	rict No.	19-	29	2	eet No.	2 0
	T of di	m hip or other	Za.	Condit townsh	p/Mg.	Uni corporate (Said lane)	pk ce.	7 83	on L DE	DIFF.	ER	e ere	Ins	PLACE OF RIKTE	eri salasar da karar eta k	MOTHER TON	E (OF NAT		e ed by m	e Ofer 7	1980, Ava	2/.	Zer	in men	VET LIN	8	2,11
1	and, the	- 1	100	of moch persons whose plane of abodic on April 1, 1000, was in this family line remain find. One for given interested models benefit, if my Judicia conceptuate the part 1, 1990. One thinks have been seen April 1, 1990.	RELATION Relationship of this passes to the head of the family	The state of the s		1	Age of last	Martial con- dition	, 1	2-12/mpm	Place of birth of each pur the United States, gire which birthplace is non French from Canada-E Paneon	on commerciated and of his cain or Tentions. If of fa athuraed. (See Instruction plats, and Irlan Free State Fathura	w har parents. If born in sign birth, give consery in a.) Distinguish Canada- less Northern Ireland mornes	Language spokes to home before coming to the United States	Constant of the Constant of th		Parke do v	CONTRACTOR THE PERSON A BRIDE TO THE PERSON AS BRIDE	Discount Control of the Control of t	CODE (For office town state. The new wide in this column)	į	盡			
0.00	, 1		- Uk	·	1		14	in the B	6 00	1	×.	760	all:		Milton At	Service Control	11		- 74	u takur	70000	V 5 V I/2	17. 74	-11	N.	-	1
					low										Quel En State					Laborer	Farmer	VOVV					
				- Wrenn	land		X	28 2	1/2	1	114	de	Minnie	Quel For Het	Oil For State	314037	61 0	0	1/2	11 mares	PH MARLE				-		
		24		Tehrman Pris			- /a	200 2	4 53	m	26.	260 260	Armani	Wish tou State	Armany	anne				La France	Unual to		R 24.		_	2/ 0	
100	7	i E		Nuldas	Willa. 2		- /4	7 2	1 40	20x	12 21	P/2	Williams	annous	Mer Lui	/	60 5		1	11 more	0		1			1	7
		1		Mara	De Alte		X	3 7	W 18	1	10	7/10	.010	De and	000						SEL CONTRACTOR						
	10			- Carl	Ten day	C. Notes	T V	Se 2	6 25	1	21.	7	001:	Henrick.	Mining	BEAUTIC	()			Palmer	4		4 7.60		n e		
1997	12			Tiny Escape	bou into		X	. <i>2m</i> 2	F 24	m	3 W.	26	Minis	ar make	Mentucky	12 32 3	60		2	h more	Jareni!	200	1	100		1 2	
	14	9.2	*2		Weld.	D.	74 V	200 2	4 42	211	4 11/	Ma.	Manage	Whin 1	alloi- aist		61		14	a Trans	Mussal Janin		5 Vac		Ma	72 2	
20%	1.6			Leaders on Horne	Maken		У,	2n 2	4- 29	8	n	The	Minois	allinote	OPP		61		6	to taborer	Farm	110V	1/2 Ch	2.		1	5
				Pregnolds law 1			7P 14	m 2	4 40	200	3 22	26	Ollingia	Minio	Armany	0.0				a none	Jeneralian	8000	E Va		No	23 10	
1	18			- natten	lon	100	X	m 3	15	8	7h	the	Ohio	1 Minaral	Mebranka	121/2016	50	100	24	Mont	811 0				止		9
				Brockman Bernie	Druglter										Mebraska					a none	+0		8 VA		Or.	15	
Mary St.	21	28	15	Reserved de William	Dead	0	RU	25 2	4 67	The	2	U.	Illinate.	Canalant.	Exalend.	1200	4 .	0 0	12	a Farmer		$y v y y_y$	0 P. de		26	24 1	1
				Herman Howard	Davidta	R	- 77	26 2	16 24	24	21 Fax	1/a	7/10-21-1	700	Minaia	100000000000000000000000000000000000000			- 7	u Hone	General E.	con ich	0 1/4			25 8	
2550	24			. /	Mr. Tack	100	0	72	- 29	20	2 22	Ma	Mineral	armany	Minis	30.00	61 1	5 /	10	now.						. 3	4
	25 36	17	17	Galt Vellery Jr.	Nead	0	72 ZA	7 9 8 20 2	1 / K	200	2 2%	Zle.	Braland	England.	England.	English	60 o	v 187	210 3	none none u Harner u Hone	(Kneral Fara	buvv	6 44		no	26 3	
	27				Mile-H	10	17.	7.2	4-60	m	9 2%	7.64	Allinois!	Phis	Virginia)	0	6/	10		u none						21	7
	28			Taulman Ferdinan	(Boarder		X	20 2	1 25	X	74	260	Allinair	Illimois	Mineral		6.1 .01		- 4	& Laborer		VIVV				2	
	30	29	28	maier Coselahile	Head	P	R 14	m 2	4 26	m	25 0%	Ula	Mining	Officeria.	Plinaria		61	14	- 2	Farmer						27 3	
NA.	31 22			- marauerite	Moualtu	01/10/20	- X	3 3	1 /1	S	12		Ollingia	of Climaia	10 Planakar		61	20		none !		1.8	U	77 -			
	18		19	maddock Barl	Head	P	N 74	0 21 2	1 37	m	5 M	160	Selinous	Illimois	Il lineral	1000	61		2	a Jarmer	- C		E Gu			28 s	
-	15			arvilliah	Son	4	Х	m 2	11	9	24	do	Mineral	Illinois	Allim rist		61	10	- 2	6 none	21 . [2006]	F	1				15
4				- Robert			X	20 2	4 9	1	24	· Elec	Official	Ollinair	Oppinsie			10	1	nane						9	
185	18			deleina	Dallalte		~	B 100 E/	Mr. Acti	- 7	677		110.0	0 00.	11 11 11 11	40.40	61	100		mane	2017 15 4.0					3	18
	19		50	Brice Henry	Ondalita Nond	76	_ 7/4	7 2	1 0±	m	8 20	74.	Minute	Plinain	Allinain		61		7	h. Frances	acousta.	ani in	E 740		26	29 4	
100	12			- Mrake	Milery	a 18.	Ϋ́	37	447	2n	2 2%	Che	Migrain	Min Sunite	Ollinois	0 2 Take	LI		L.	La Taker	0		1	11.5		- 4	0
	3			Casity Oscar	Wead	P	7.6	221 2	48	m .	1 30	de	Volinois	Minair	Minnie	1	U			to Enhance	Farms	110 F	W The		22	- 4	12
1005	4			nora	WiloH	2.6	ΙX	70	447	m	8 3%	U.	Illimois	Officers	Ollinia		41	10	1/2	none.			1			- 4	
	6		H	- bdward	Son		×	221 2	14	1	Die	a Me	000	Ollingia	Dolinain		41		1	a none			Н		+	- 6	15
E	7				Son.	3 00748	X	2n 2	4 .5	7	n,	. /	000	0/10:	700.	1000	41	100		30-00	10 A C				#	- 4	(7
		- 12		- Allow	Nead	0	u u	2 00 0	74	m	20 200	77.	Minnie	England	Malea	A STANFA	40		1 6	10 Mone	BELL THE				90	- 4	
	o i		П	- maria	Wile		_ /x	7 7	1 49	m	12 %	2in	000:	72. Carpenia	01:		41	10 100	9	he now							o
1 1	A 80	VIATIONS IN COLUMN	las Ias	CATED:	-	1 Ca p-1			ě	٩	-		***		Can to Feel Wy		an	MA BE 1	Quin is	THE CALLS	II, N, II, II, II, II, and III-II and IN-Per hands of handles the period for a few marks.) period for the period of the part of	d pe-		Hamilton all manual (8 ye of Toront)	2 2	-	-
	2			K72 (4.5								63				_E				1813	A long in the say of						1







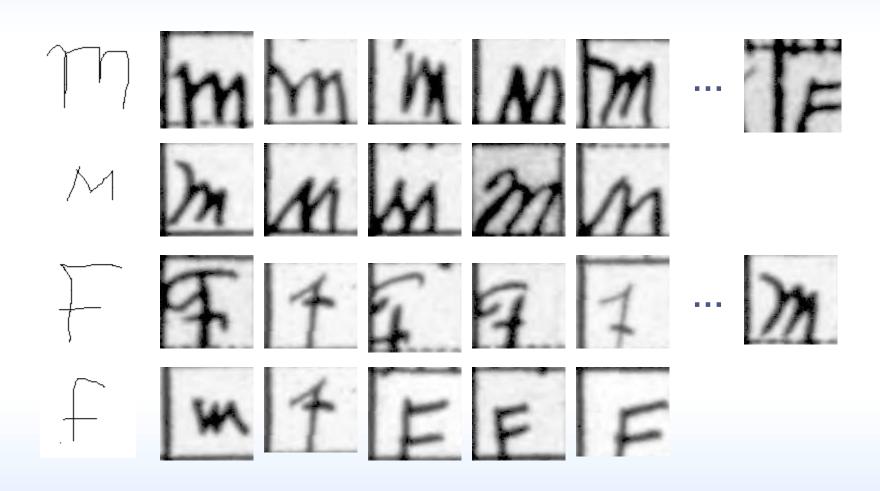
Sondra Jerry Jerry Jen JEN Tessie 1 e 556



Jerry Jerry Jessi Jen Jen Ror Tex Tex Sandra Sandra

Jessie Tessie Rory Pory. Keith Keith







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

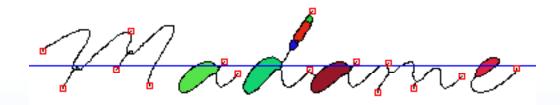




Image Pyramid

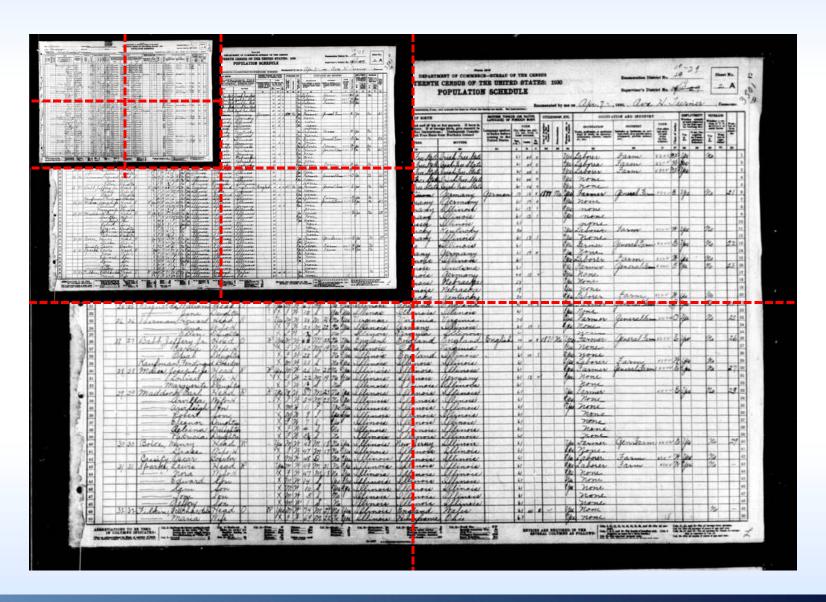
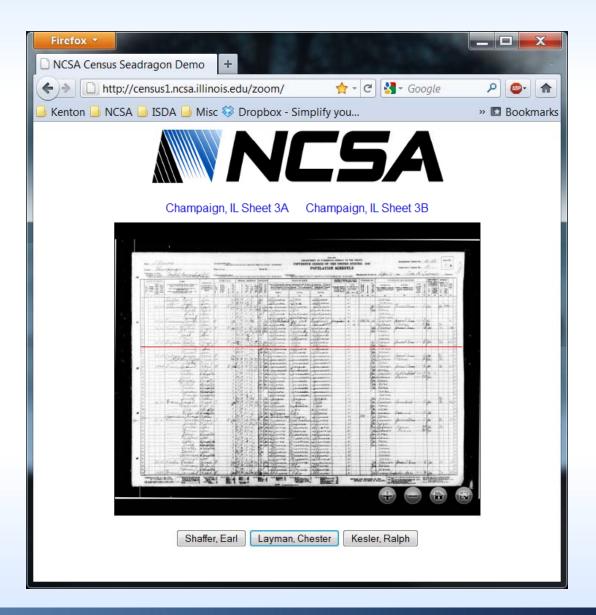




Image Pyramid





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

~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

