



Automated Digitization of the Censuses of Housing Block Statistics, 1940-1970

Jeffrey Lin, Dan Moulton, Isaac Rand & Robyn Smith
Federal Reserve Bank of Philadelphia

August 2024



Disclaimer

The views expressed here are those of the authors and do not necessarily represent the views of the Federal Reserve Bank of Philadelphia or the Federal Reserve System.

Digitizing Block Statistics

- What
- Why
- Goals
- Tasks and Challenges



Digitizing Block Statistics

- What
- Why
- Goals
- Tasks and Challenges

Census of Housing
Block Statistics 



HOUSING—BLOCK STATISTICS

Table 3.—CHARACTERISTICS OF HOUSING FOR CENSUS TRACTS BY BLOCKS: 1940—Con.

Cen-sus tract	Block	Total struc-tures	ALL DWELLING UNITS BY OCCUPANCY AND TENURE					ALL DWELLING UNITS BY YEAR BUILT					OCCUPIED DWELLING UNITS			ALL DWELLING UNITS BY STATE OF REPAIR AND PLUMBING EQUIPMENT			OWNER-OCCUPIED UNITS BY MORTGAGE STATUS		ALL DWELLING UNITS BY CONTRACT OR ESTIMATED RENT				
			Total dwelling units	Owner occu-pied	Ten- ant occu-pied	Va-cant, for sale or rent	Va-cant, other	Number report- ing	1930 to 1940	1920 to 1929	1900 to 1919	1899 or before	Total occu-pied	Occu-pied by non- white	Persons per room Num- ber or rptg. more	Needing repai- ring	repa-re pair or no private bath	Need-ing repa-re pair bath	No pri- vate bath	Number report- ing	Mort-gaged	Number report- ing	Average monthly rent (Dollars)		
3-A	24	21	34	11	17	6		34					28	28	2	34	19	1	19	6	2	33	24.24		
	25	41	43	20	20	3		42					42	1	39	1	40	20	1	20	18	9	43	18.21	
	26	32	36	16	19	1		36	2	4	4	26	35	9	35	5	36	17	17	14	11	36	18.22		
	27	34	38	13	24	1		37		1	4	32	37	1	37	1	37	15	1	15	11	6	37	23.73	
	28	26	49	18	30	1		49	1	3	2	43	48	48	7	48	24	24	24	18	8	42	21.99		
	29	18	24	12	12			24	1	2	3	18	24	1	24	2	24	8	8	12	6	22	22.55		
	30	11	25	4	20	1		25	2		9	14	24	5	24	2	24	15	15	4	2	22	17.86		
	31	24	38	12	23	3		38					35	12	35	3	38	35	35	15	12	5	38	20.66	
	32	18	33	11	20	2		32					31	28	3	32	23	23	10	10	8	31	20.42		
	33	28	34	6	27		1	34		3	2	29	33	1	33	5	34	29	29	5	5	34	18.71		
	34	6	9	3	6			9					2	7	9	2	9	6	6	4	3	1	9	20.00	
	35	28	30	13	16	1		30					2	28	29	2	29	1	30	19	3	10	3	30	21.57
	36	22	31	7	23	1		31					31	30	1	30	4	27	18	10	13	3	2	31	17.32
	37	23	43	7	32	4		43	4				39	39	1	39	3	43	24	15	20	5	4	42	22.14
	38	20	26	4	20	2		26					24	24	4	26	17	10	11	4	1	26	18.04		
	39	41	44	12	28	4		44	1				40	40	1	44	14	14	4	7	4	4	44	28.32	
	40	43	71	21	47	3		71					68	68	4	71	25	1	24	15	6	71	24.70		
	41	27	36	10	25	1		36					35	2	35	6	36	27	25	20	10	6	36	17.33	
	42	28	50	4	43	3		49					49	47	27	46	1	49	32	29	22	4	2	49	17.80
	43	2	3	3	3			3					3	1	3	1	3	3	1	3	3	3	3	18.57	



Census of Housing Block Statistics

- Most granular, earliest, extant Census spatial data on housing.
- 1940-1970.
- Tens of thousands+ of scanned pages of tables and maps.

What's in it?

- Tenure, occupancy, structure age and condition, rents and values, race of occupants.
 - All houses, not just occupied ones.
-
- High level of spatial detail: Usually, a city block.
 - Small size (Pop. ~50 vs ~4,000 for ED/Tract).
-
- Coverage of large section of cities over time.
 - 191 cities in 1940 → All 1970 urbanized areas.

What's it good for?

Studies of housing investment and maintenance and long-run urban dynamics.

Studies of policies and processes that occur at extremely localized spatial scales.

Studies of many cities, or a single city's history.

Digitizing Block Statistics

- What
- Why
- Goals
- Tasks and Challenges

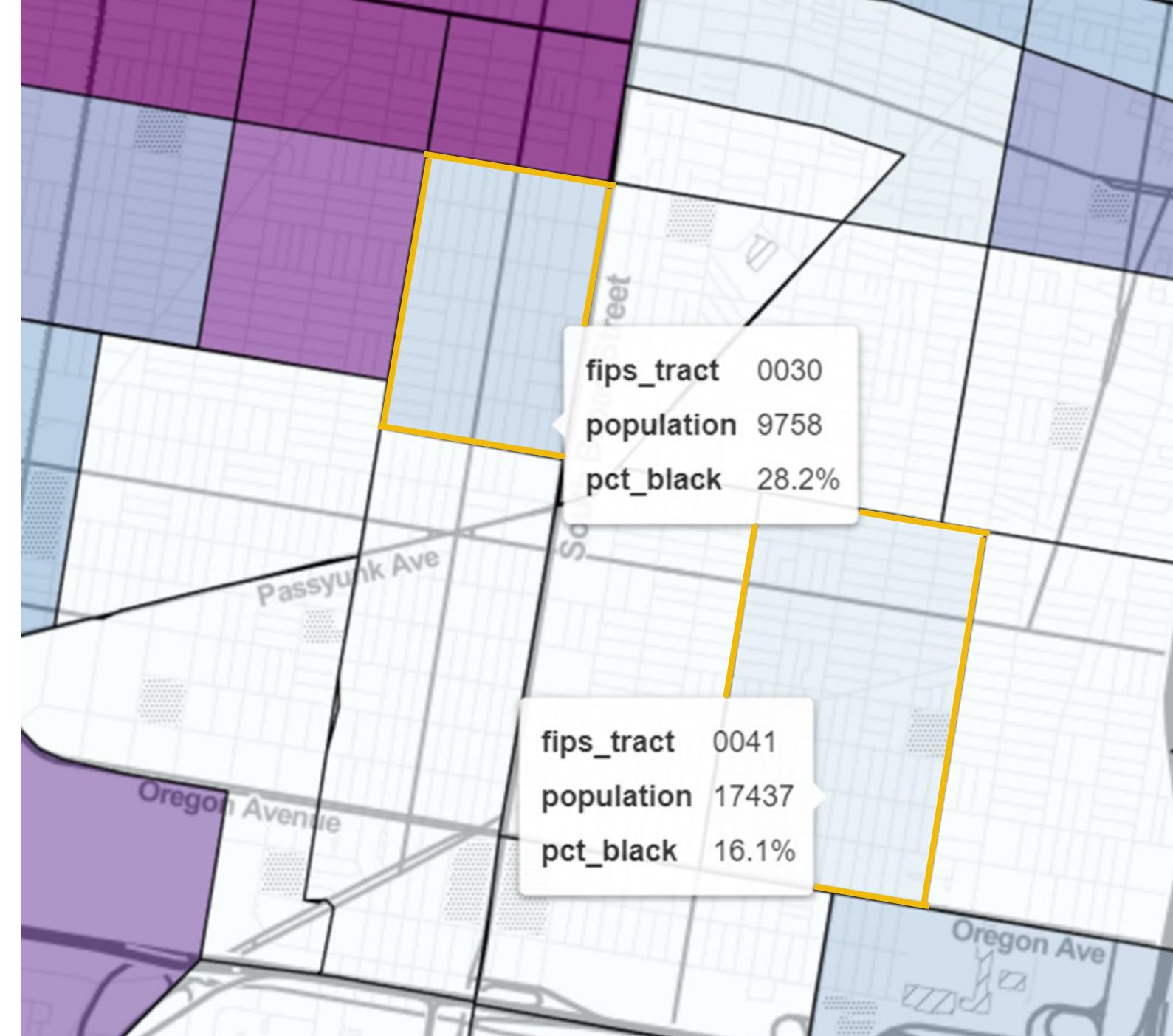
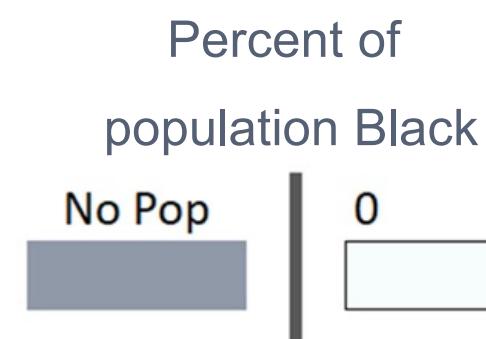
EXAMPLES

Localized processes
Localized policies



Localized Processes

- Residential segregation in South Philadelphia, 1970 at **Census Tract** scale

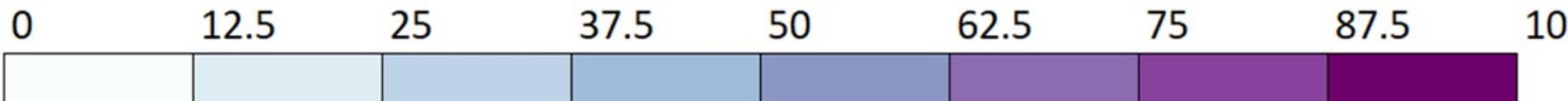


Localized Processes

- Residential segregation in South Philadelphia, 1970 at **Census Block** scale

Percent of
population Black

No Pop

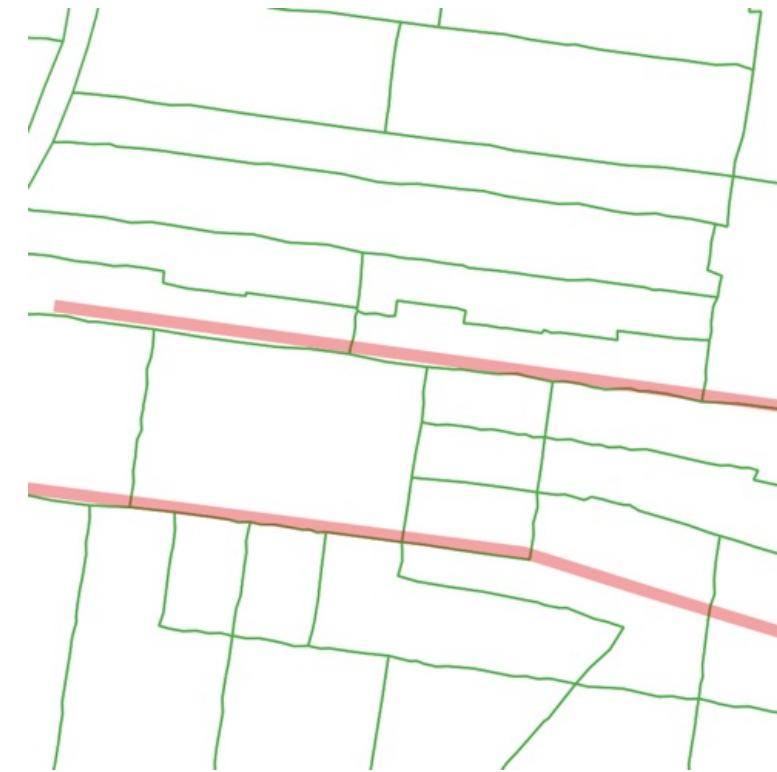


Localized Policies

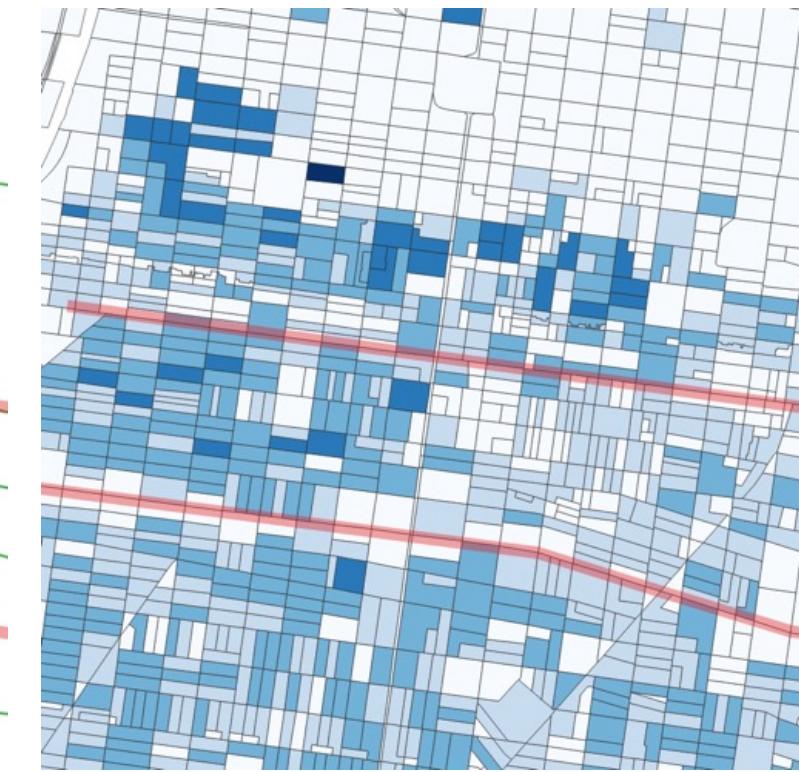
Runner-up design

- “Expecting an Expressway” (Brinkman, Lin & Mangum).
- Two proposed routes for the Crosstown Expressway in South Philadelphia.

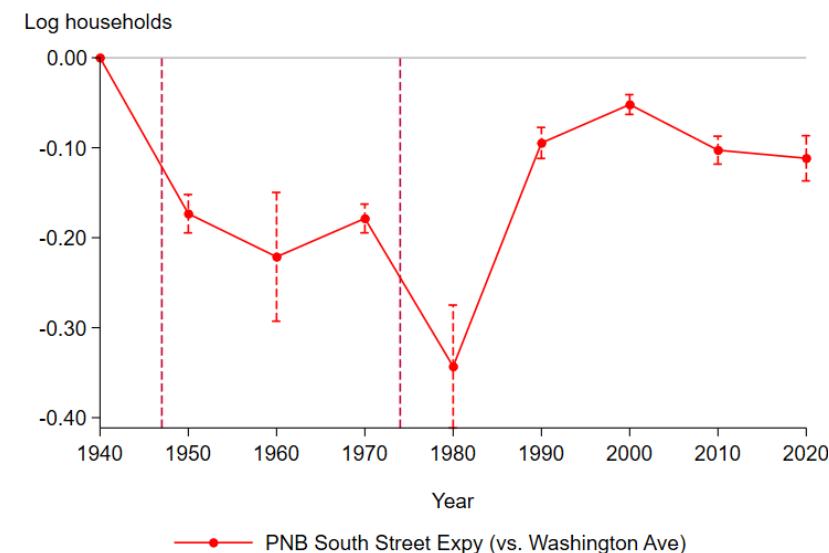
Tracts



Blocks



Difference in differences



Digitizing Block Statistics

- What
- Why
- **Goals**

- Tasks and Challenges



Our Goals



- Block data for **16 cities, 1940-1970.**
- Training and validation data.
- Code and methods.
- **Freely distributed for use and re-use.**

Digitizing Block Statistics

- What
- Why
- Goals
- Tasks and Challenges

Three Tasks	Challenges
Shapes Situations Statistics	Limitations of traditional approaches Our current work



3 Tasks, 3 Pieces of Data

1

Shapes

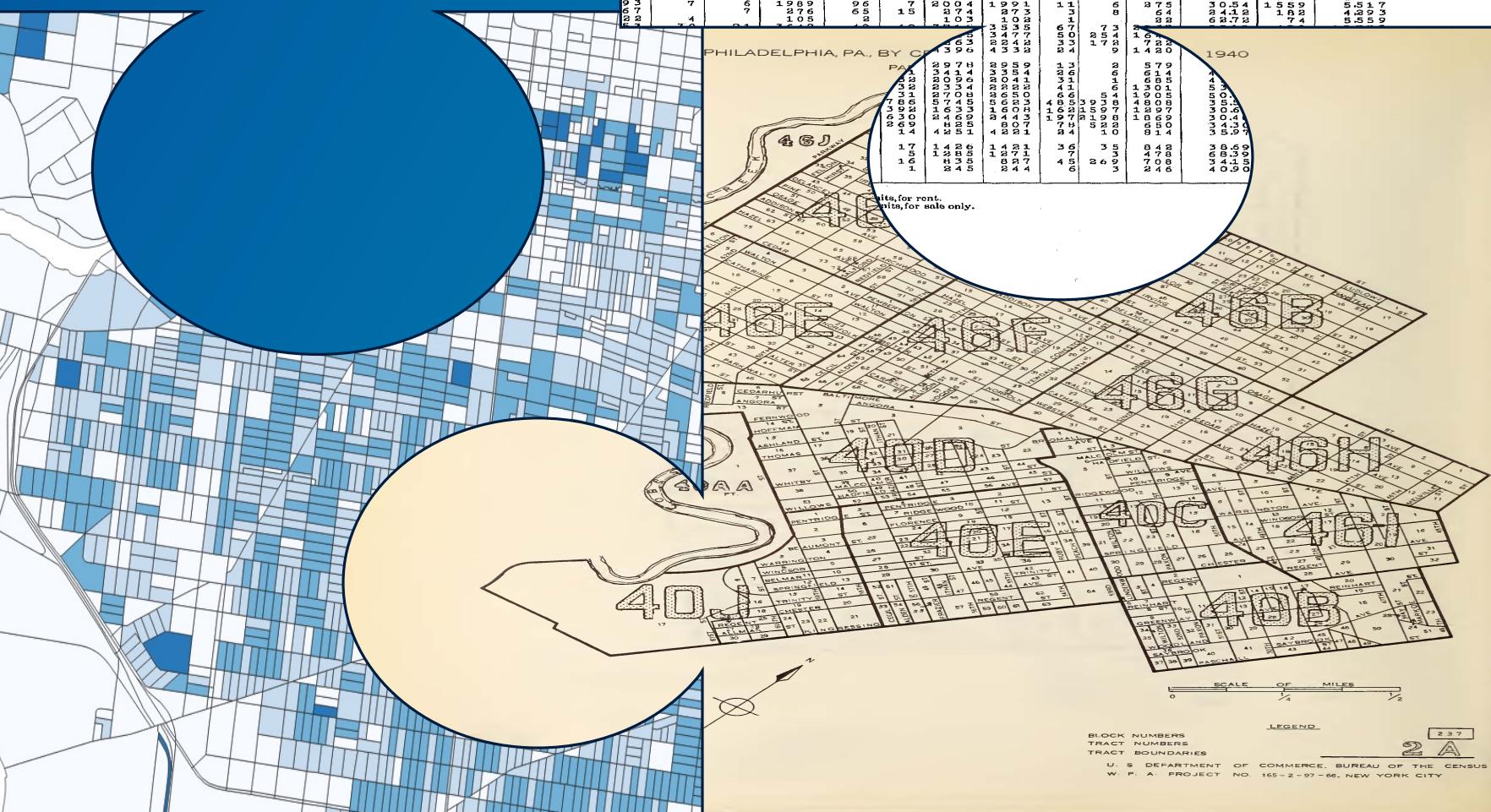
*Blocks need
to know their:*

2

Situations

3

Statistics



3 Tasks, 3 Pieces of Data

1

Shape

Segmenting Block Shapes from Maps

2

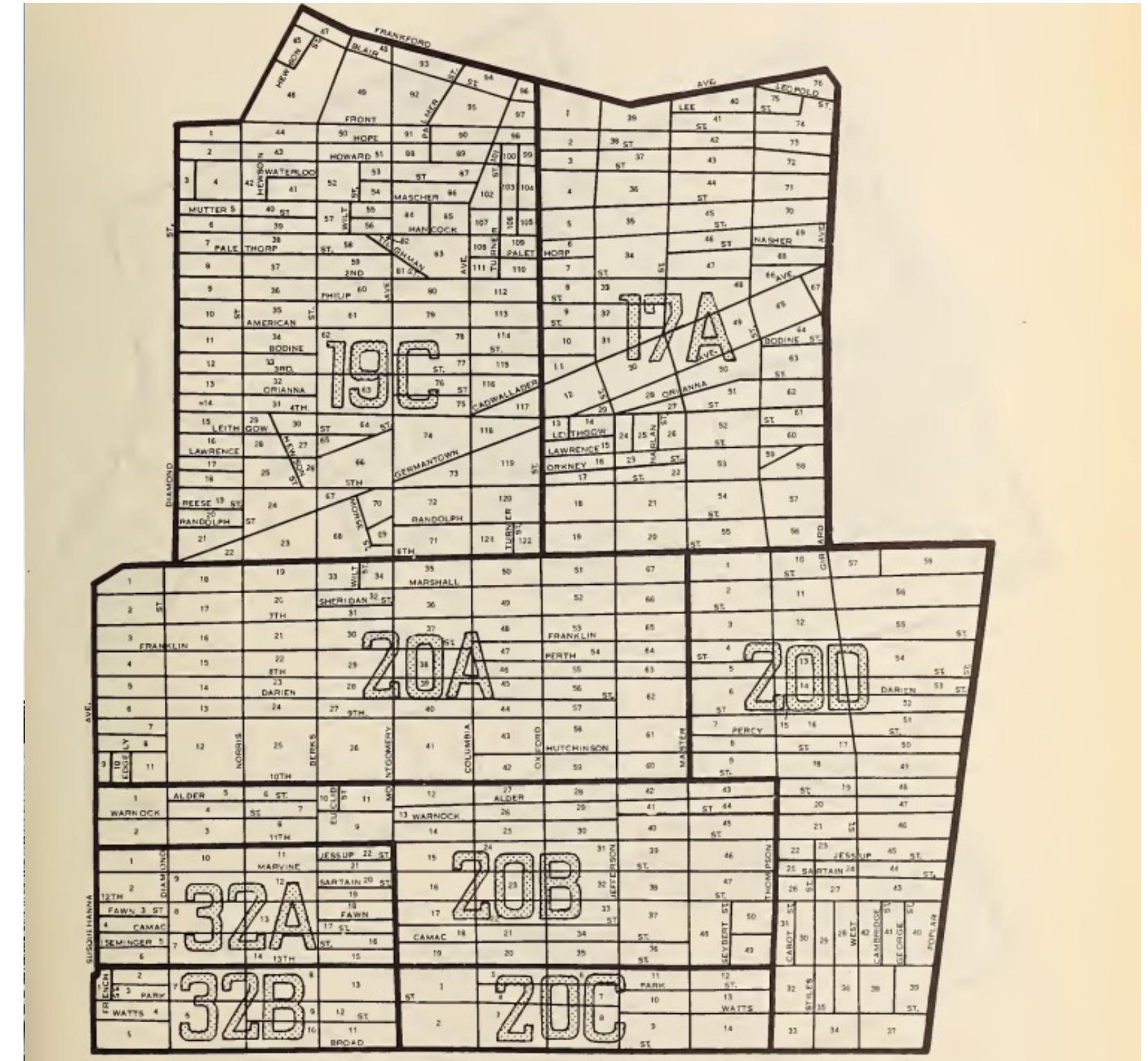
Situation

3

Statistics

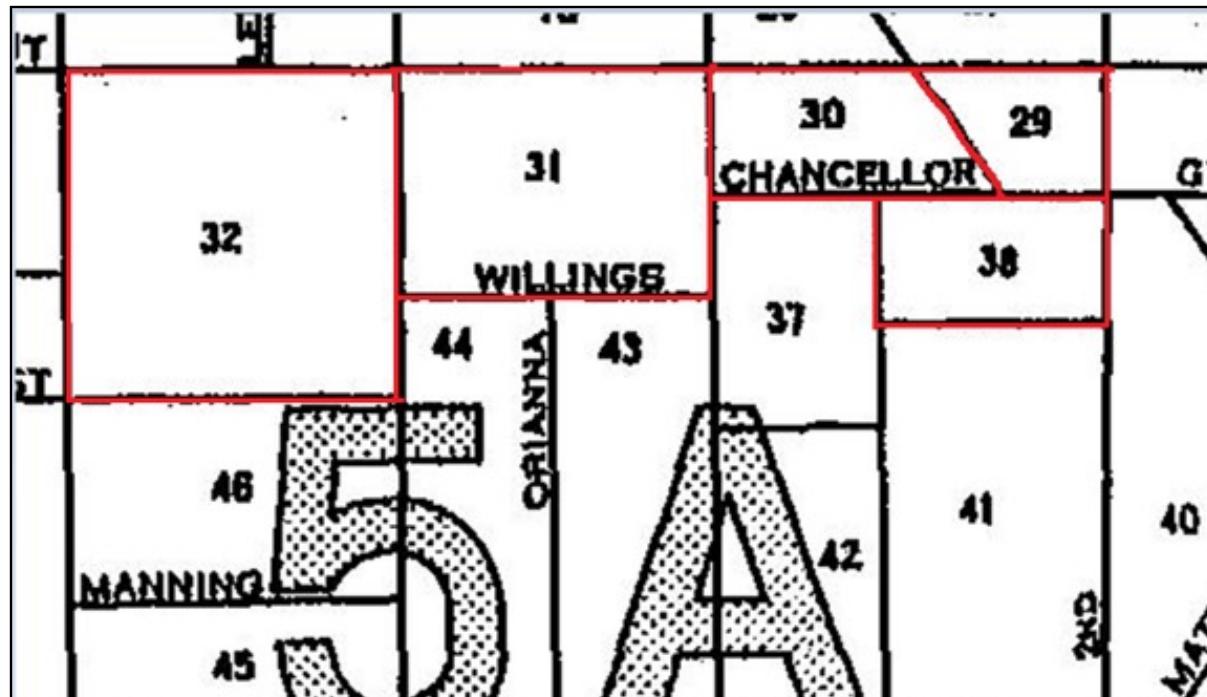
Our Ideal Process Has Only Three Steps

1. Identify closed loops of black ink.
2. Call them all blocks.
3. Declare victory.

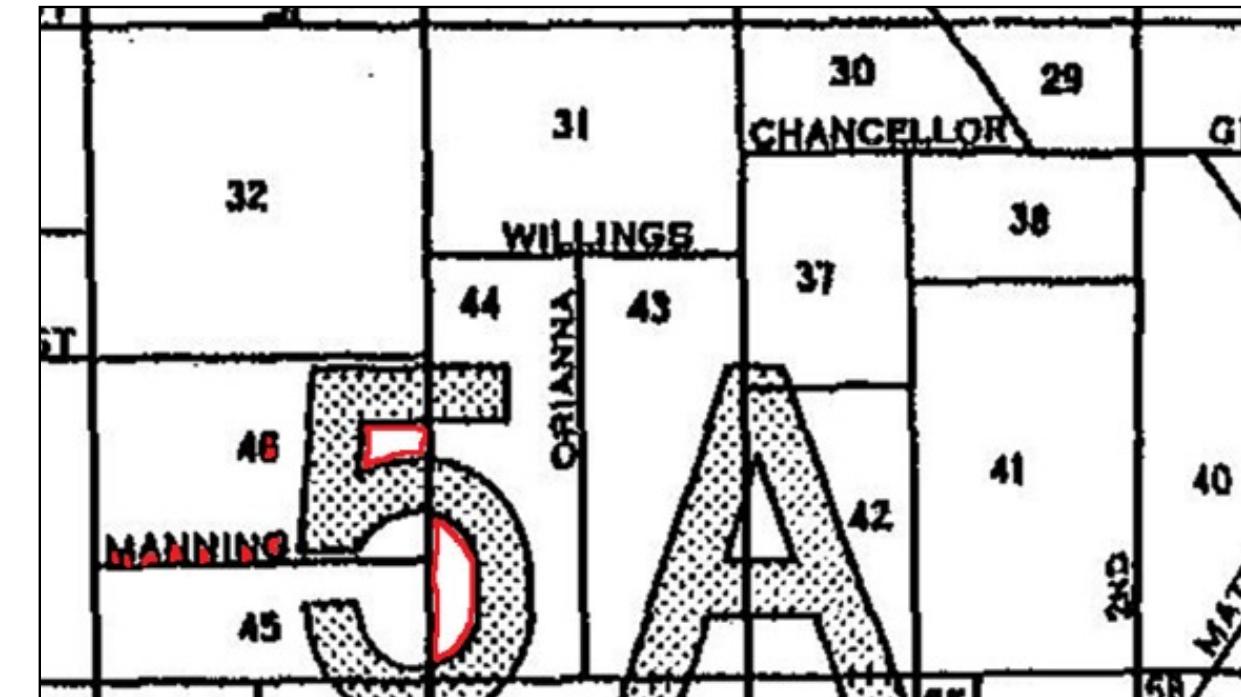


Unfortunately, This Process Fails Spectacularly

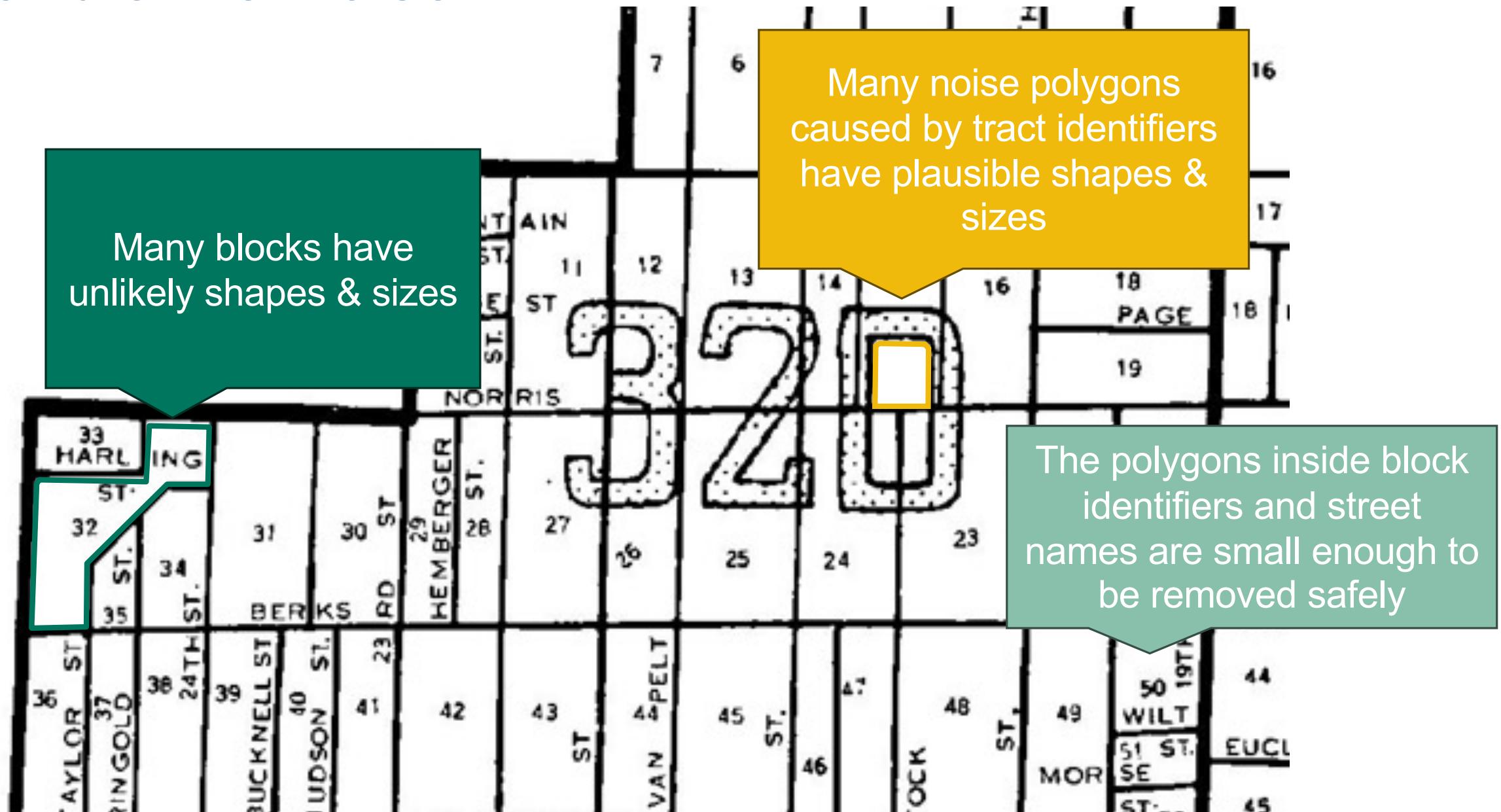
While many closed loops of black ink are blocks....



Many closed loops of black ink are not blocks 😞

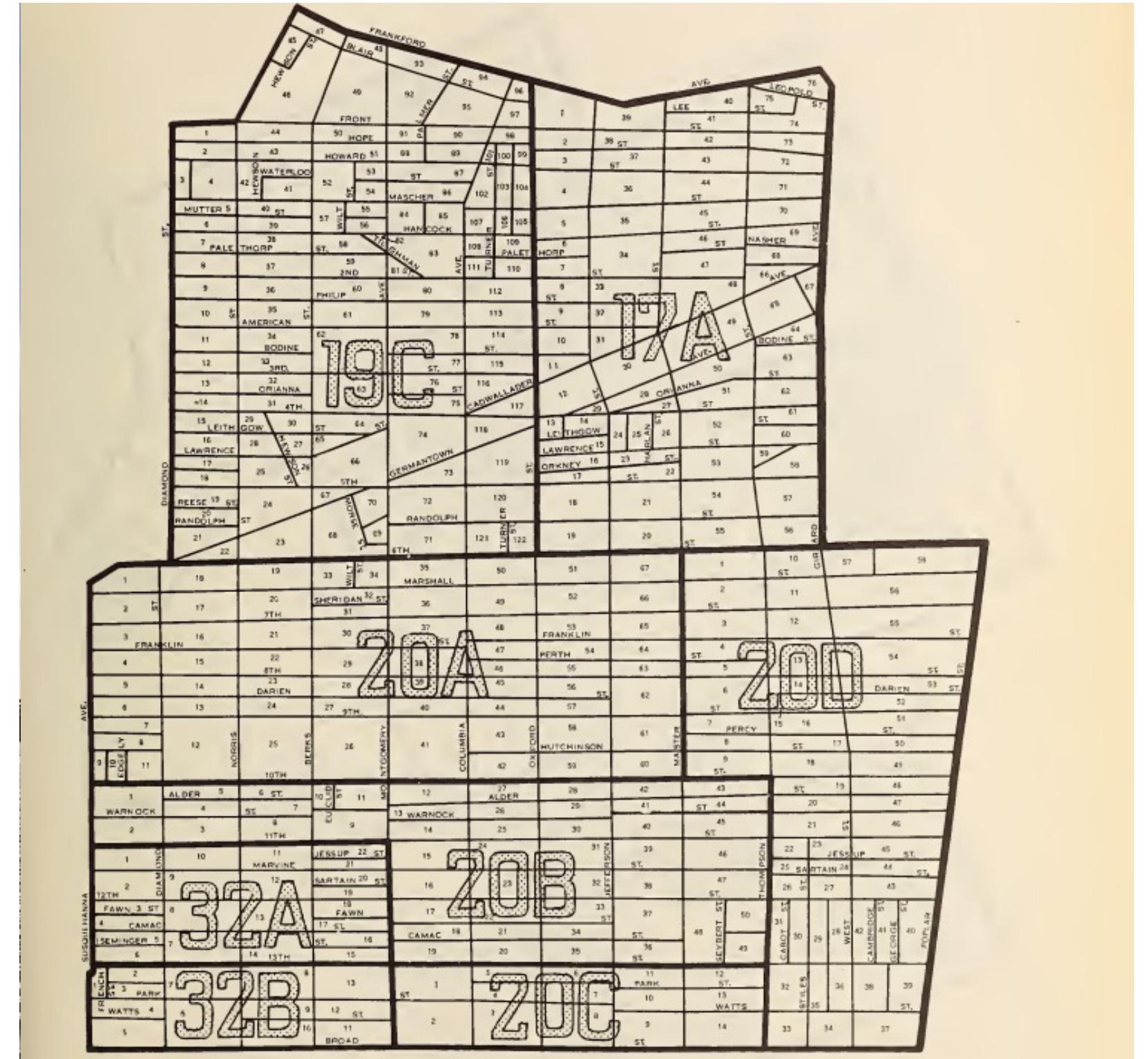


Can We Handle the Noise?



Our Ideal Process Has Only Three Four Steps

1. Remove the tract identifiers from the page
2. Identify remaining closed loops of black ink
3. Call any reasonably large loops blocks
4. Declare victory



How can we remove tract identifiers?

Traditional Method 1: Matching Large Shapes

Issues

- Arbitrary Rotations, Inconsistent Scale, Shape, and Font, Noise/Interference from other features.

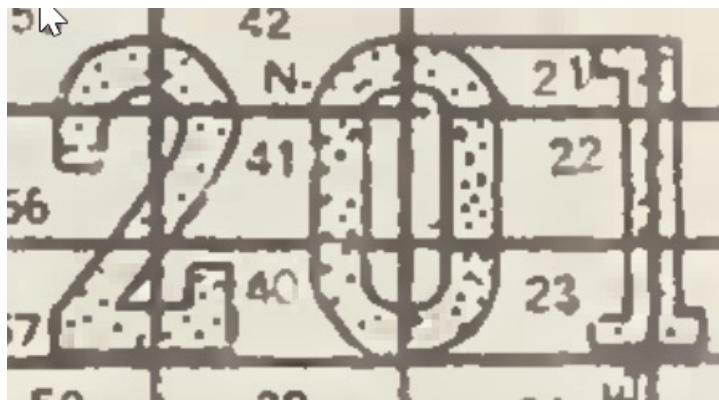
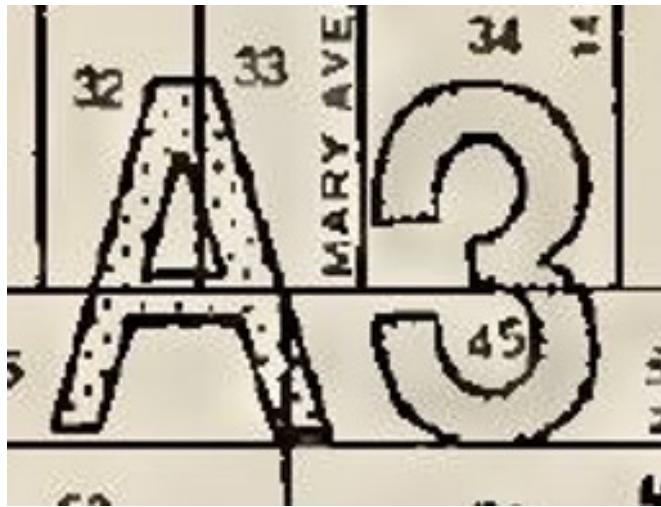
Low-confidence matches

- If we accept low confidence matches of enough templates, everything starts to look like a tract identifier.
- Especially problematic with blocky characters like 1 and E.



Traditional Method 2: Matching Patterns

- Sometimes speckled
- Sometimes no fill
- Block boundaries still a problem
- Sometimes inked



Current Work: CNN

More holistic

- Consider properties of block boundaries as well as properties of tract identifiers.
- Focus on identifying block boundaries, not removing tract identifiers.

More flexible

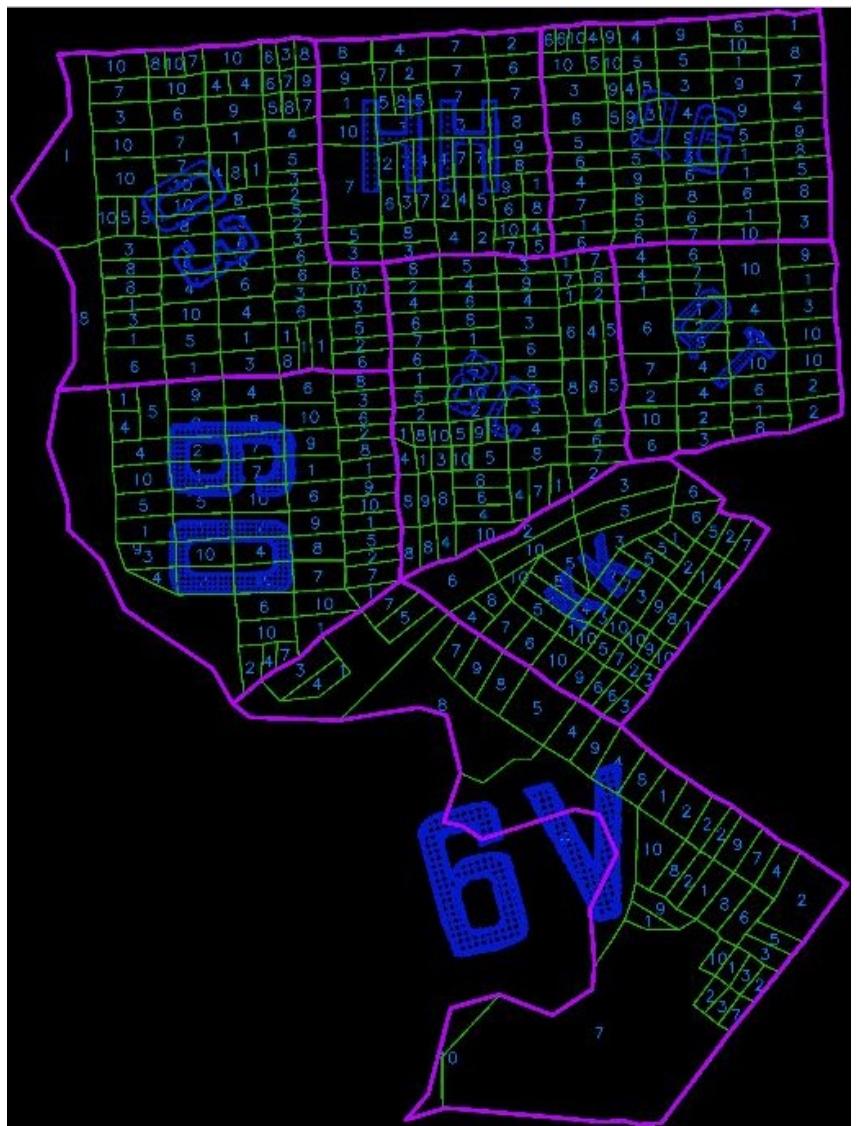
- Can learn more patterns than we can with shape template matching.
- Can address partial shapes.
- Important because of intersections between boundaries and identifiers.



Creating Training Data



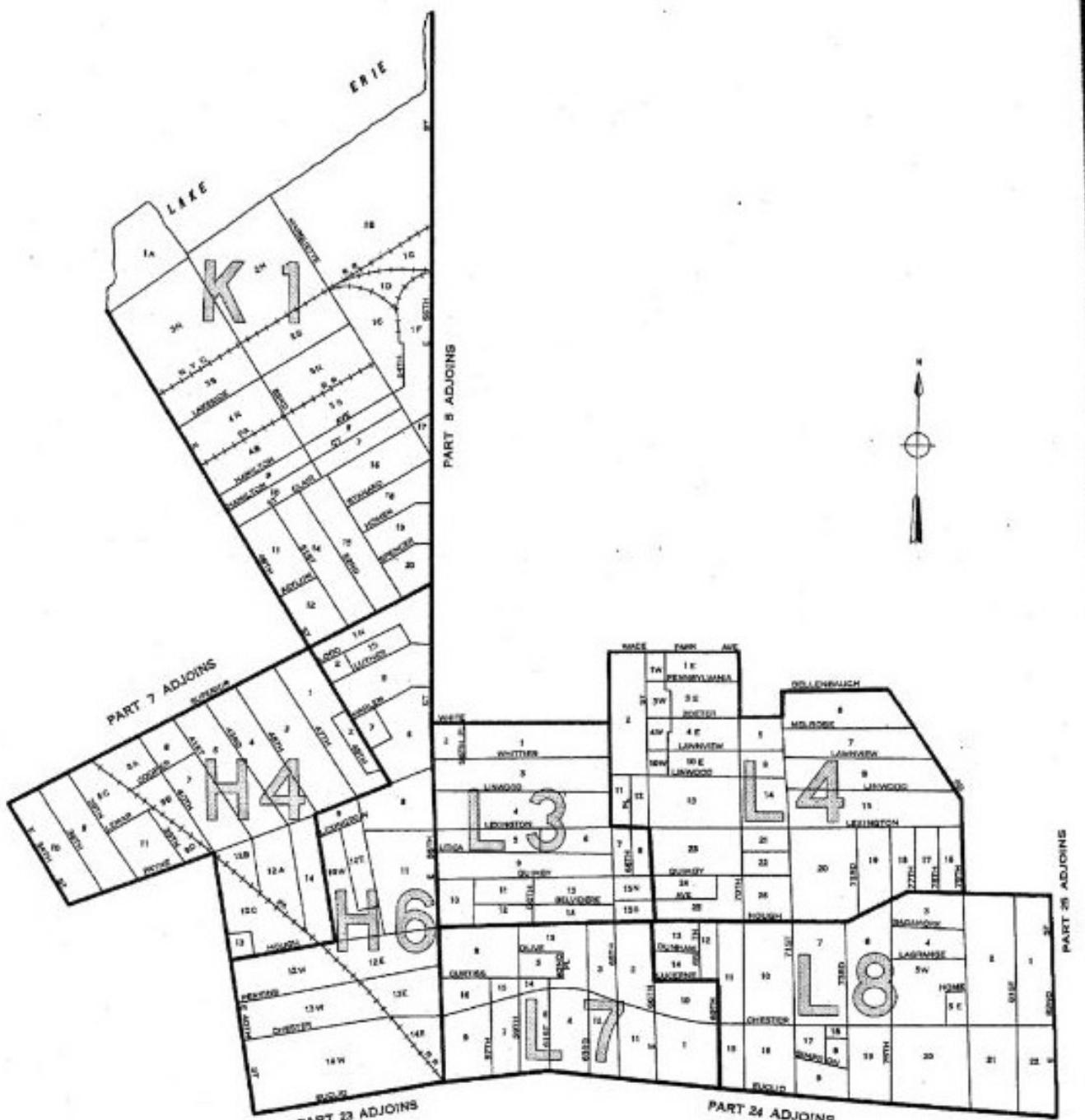
Simulated Map



Training Mask

- Hand annotations are expensive; Simulating maps is cheap.
- Sample 1990 Census block and tract boundaries from NHGIS.
- Sample tract and block identifiers from real 1940 maps.
- Randomly assign speckle density to tract identifiers.

Can the model trained on simulated
maps generalize to real ones?





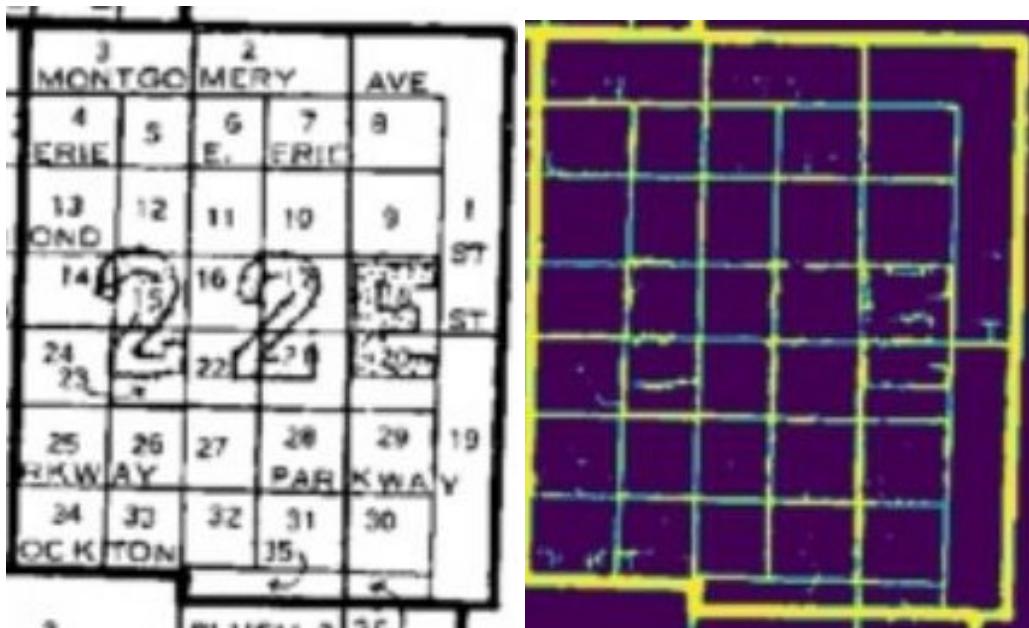
What's Next?

Model and training improvements

- Better simulated maps.
- Augment with hand annotations.

Add more steps

- Inpainting lines erased by CNN.
- Suggestions?



And Now For Something Completely Different

(1970 maps)

Promises

- Tract boundary segmentation is somewhat easier.

Pitfalls

- Which block is this? Block identifiers are inconsistently located, look like street names.
- Too much detail: Block boundaries look like streets.
- "Fishhooks" are important and omnipresent.

Our current approach

- We are relying on hand annotations for training and validating CNN.



3 Tasks, 3 Pieces of Data

1

Shape

2

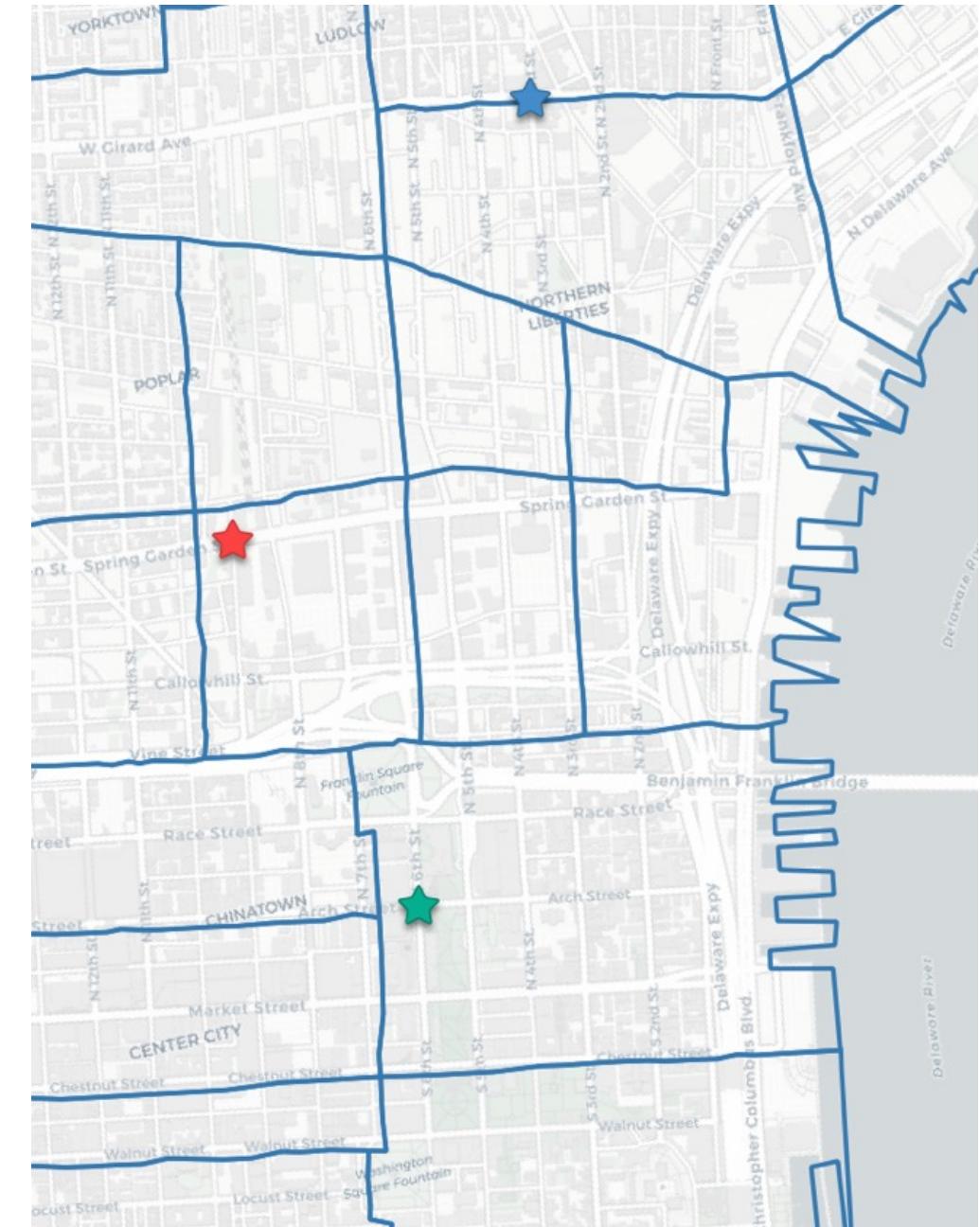
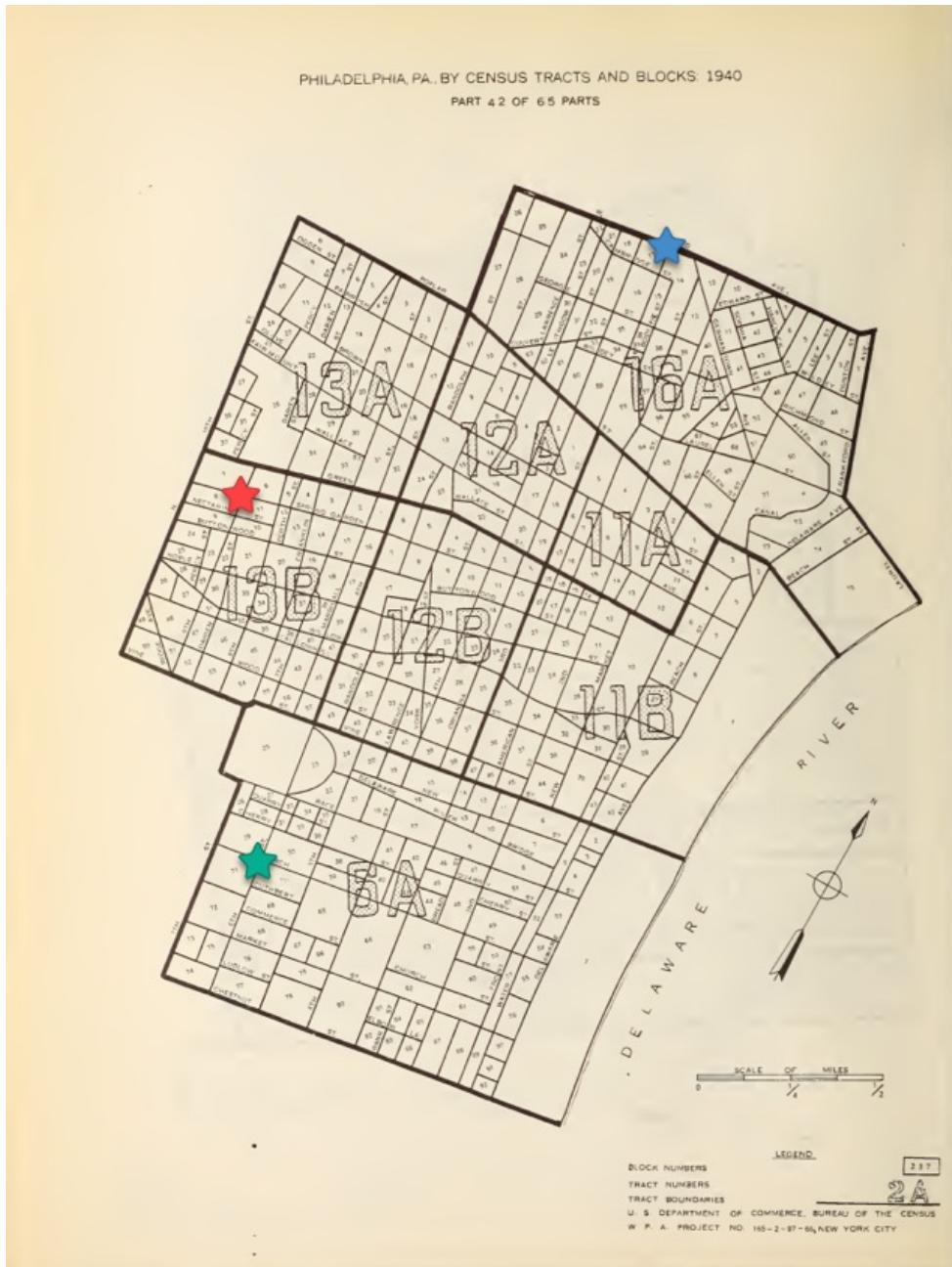
Situation

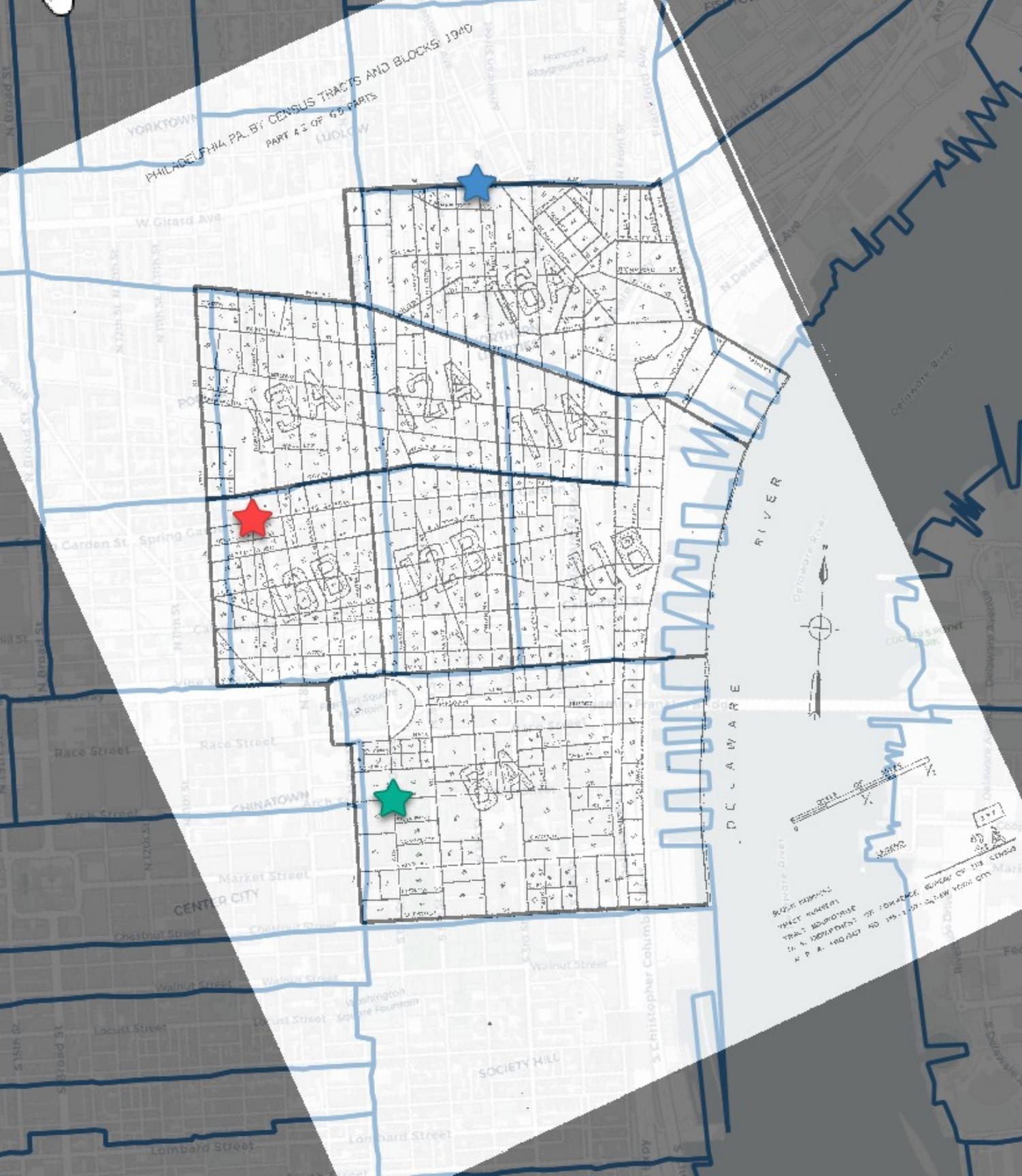
3

Statistics

Geo-Referencing Maps

Just Keep Clicking





A Georeferenced Map

We know where this picture goes now!

But...

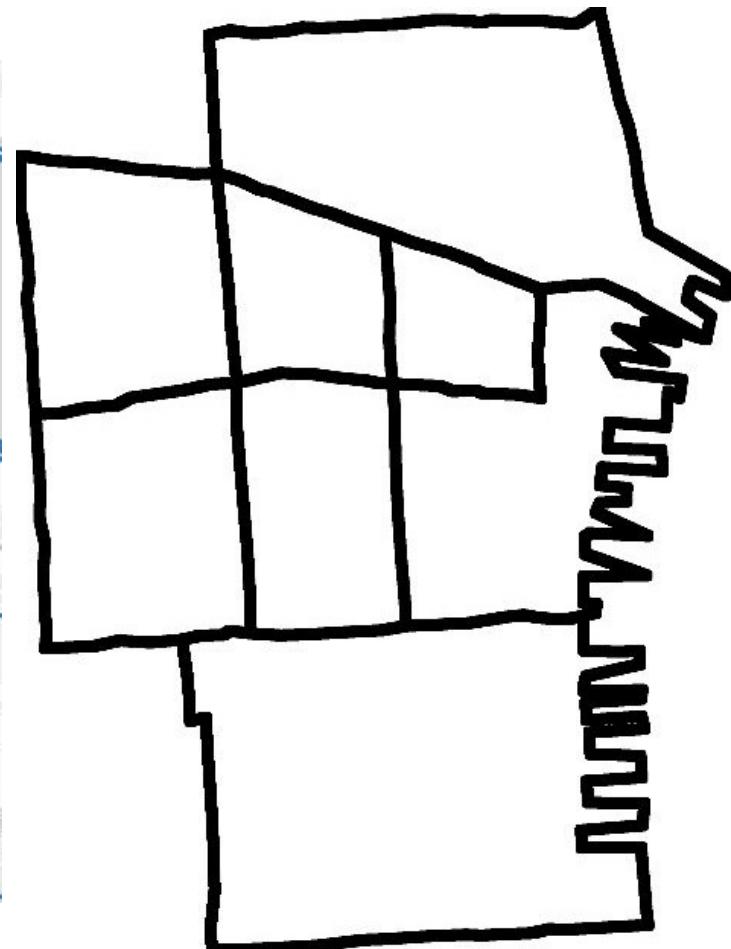
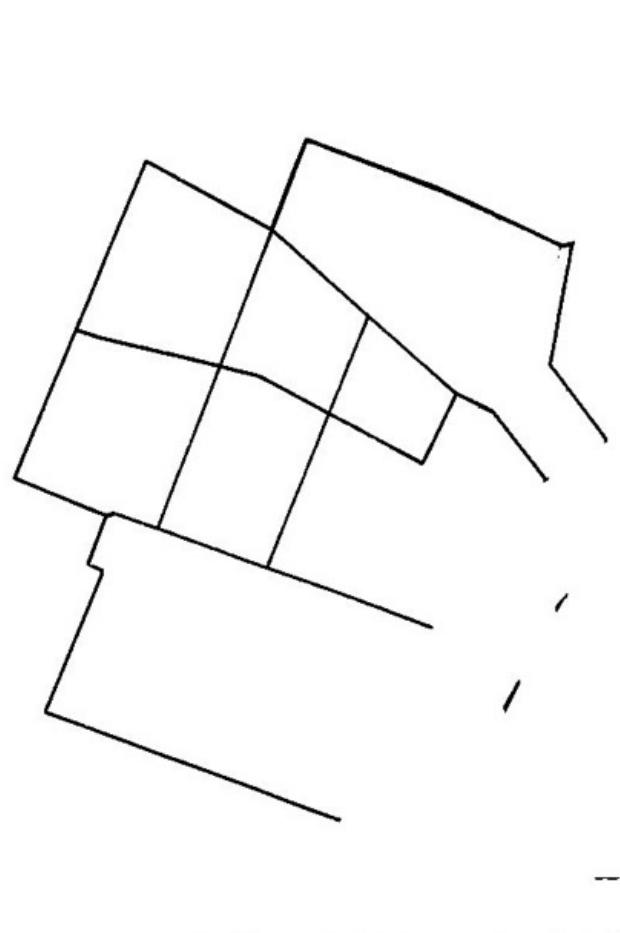
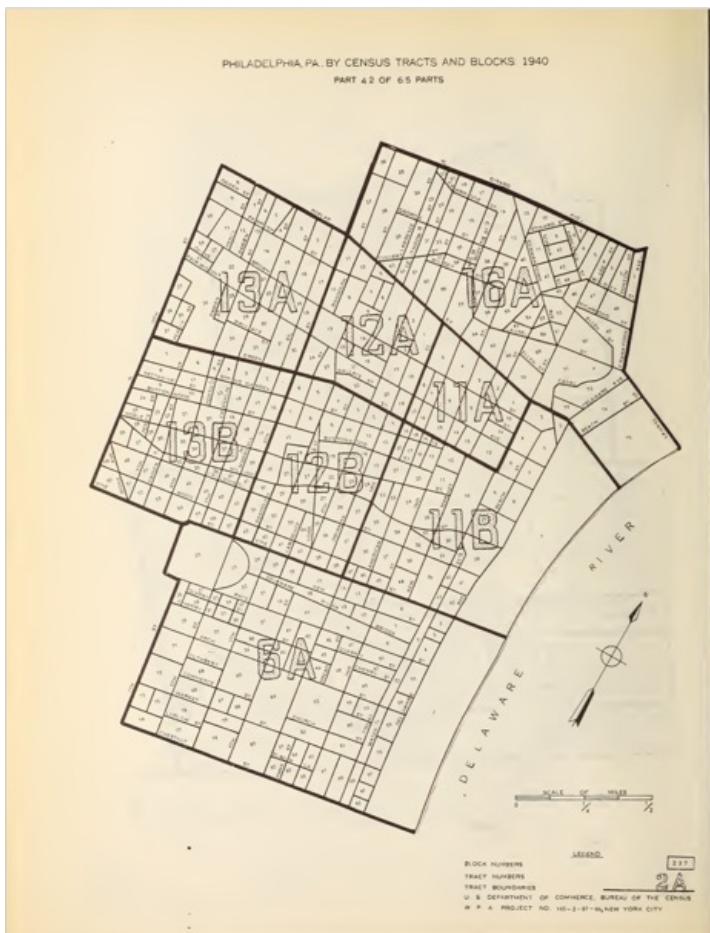
- Sloooooooooow.
- Only 3 points doesn't handle map inaccuracies well.
- Doesn't match (blue) reference NHGIS shapefile.



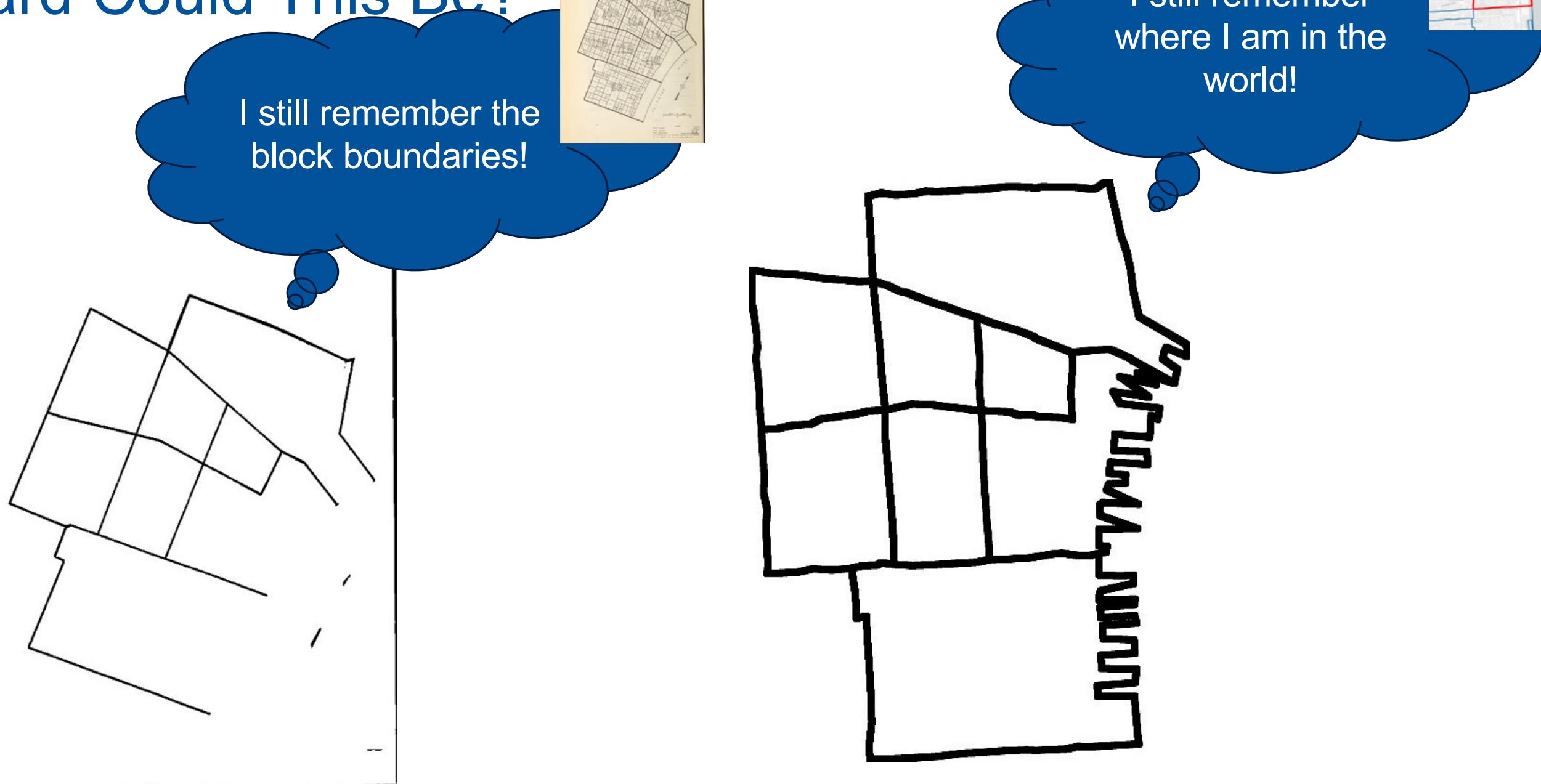
How Can We Get Better? 🏋️

- Faster!
 - We want to process many maps.
- More accurate!
- How much of this can a computer do for us?

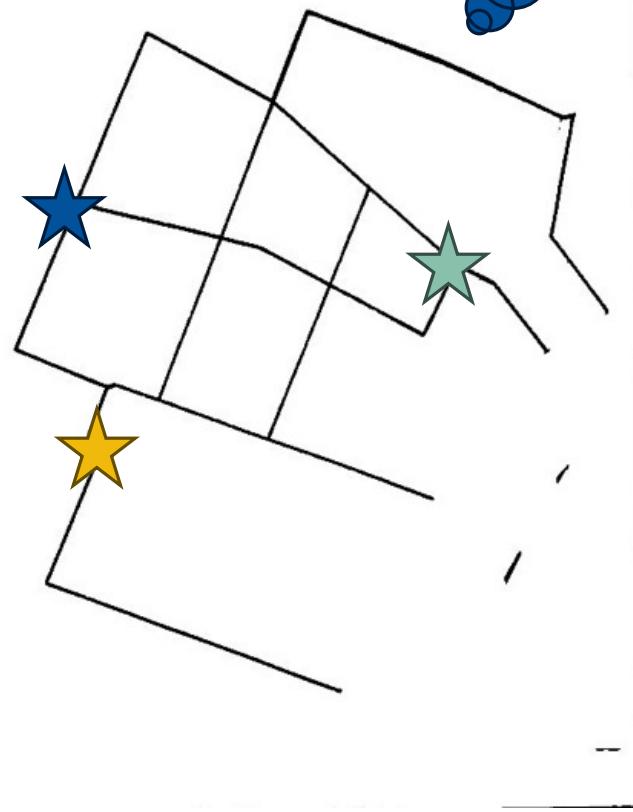
Simplifying the Problem



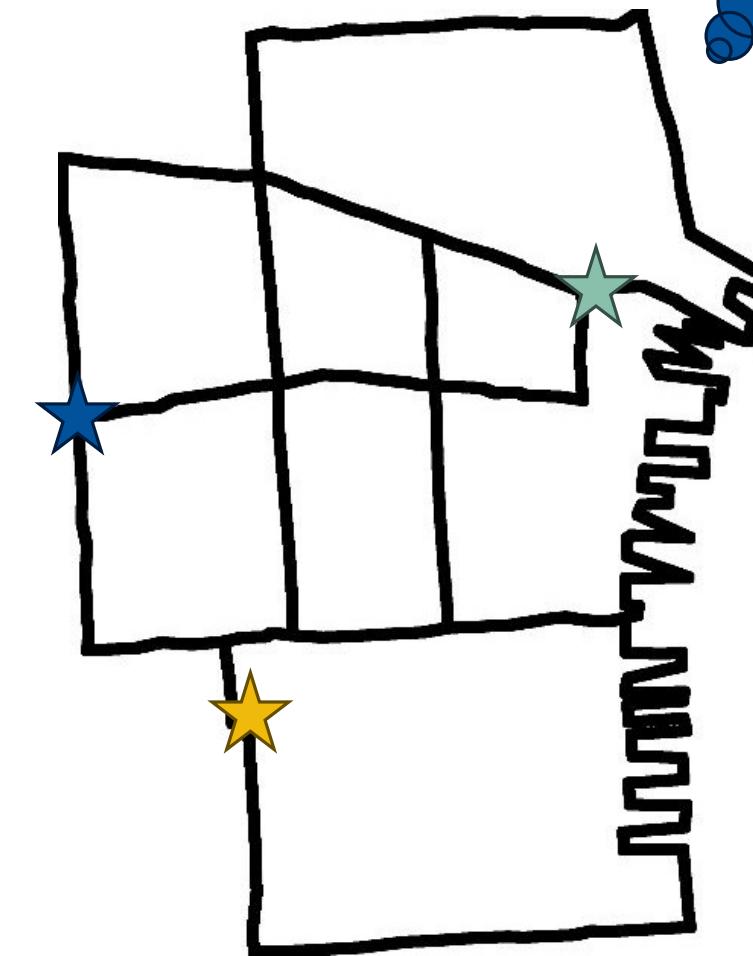
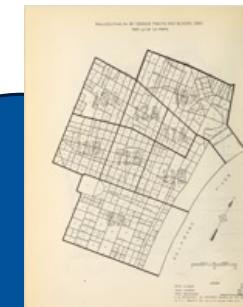
How Hard Could This Be?



How Hard Could This Be?

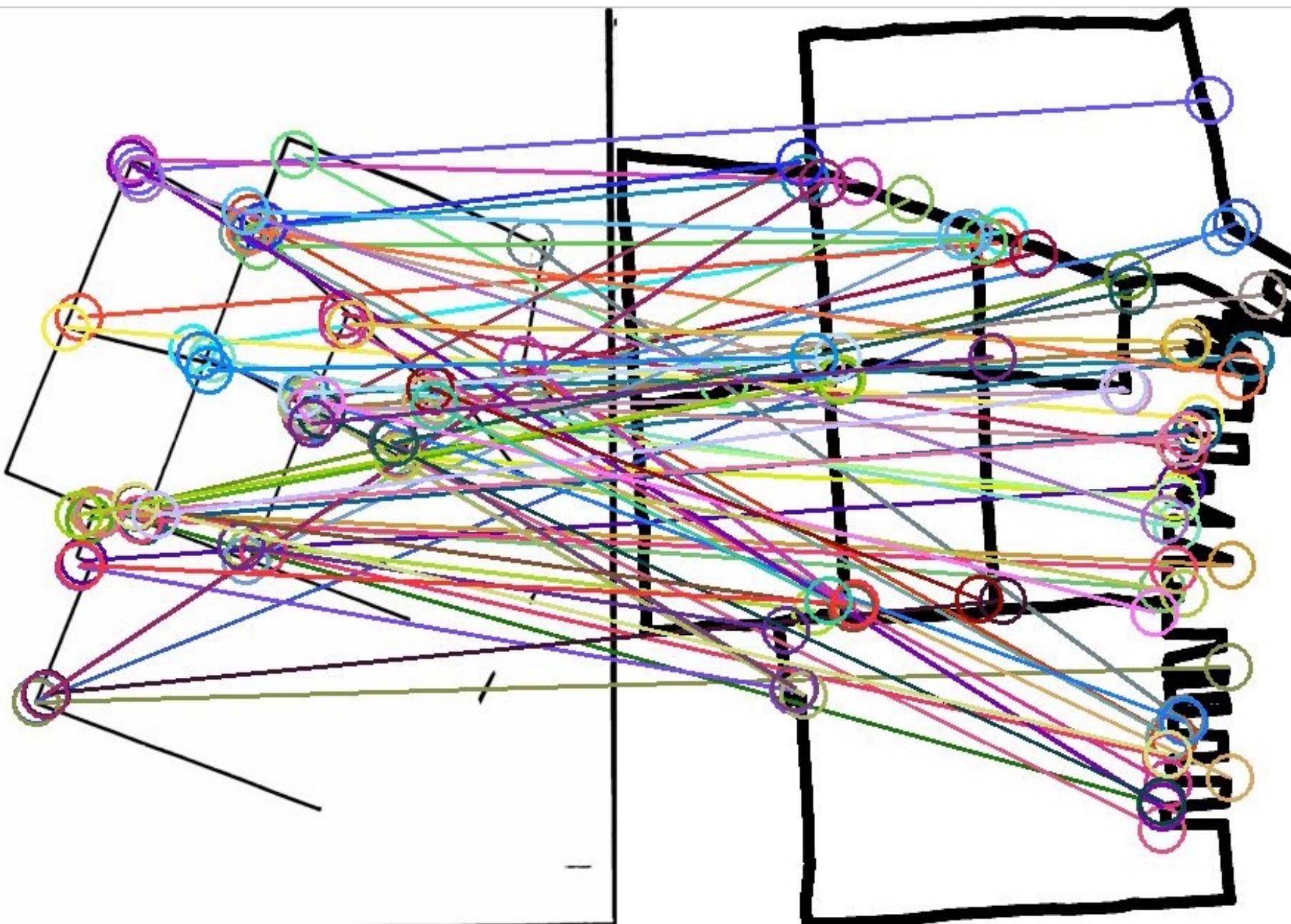


I still remember the
block boundaries!



I still remember
where I am in the
world!





Super. Duper. Hard.

- Too many matches.
- Several per corner!
- We need to narrow these down.

The Basic Steps



1

Make guess

Randomly select internally consistent links.

2

Evaluate
guess

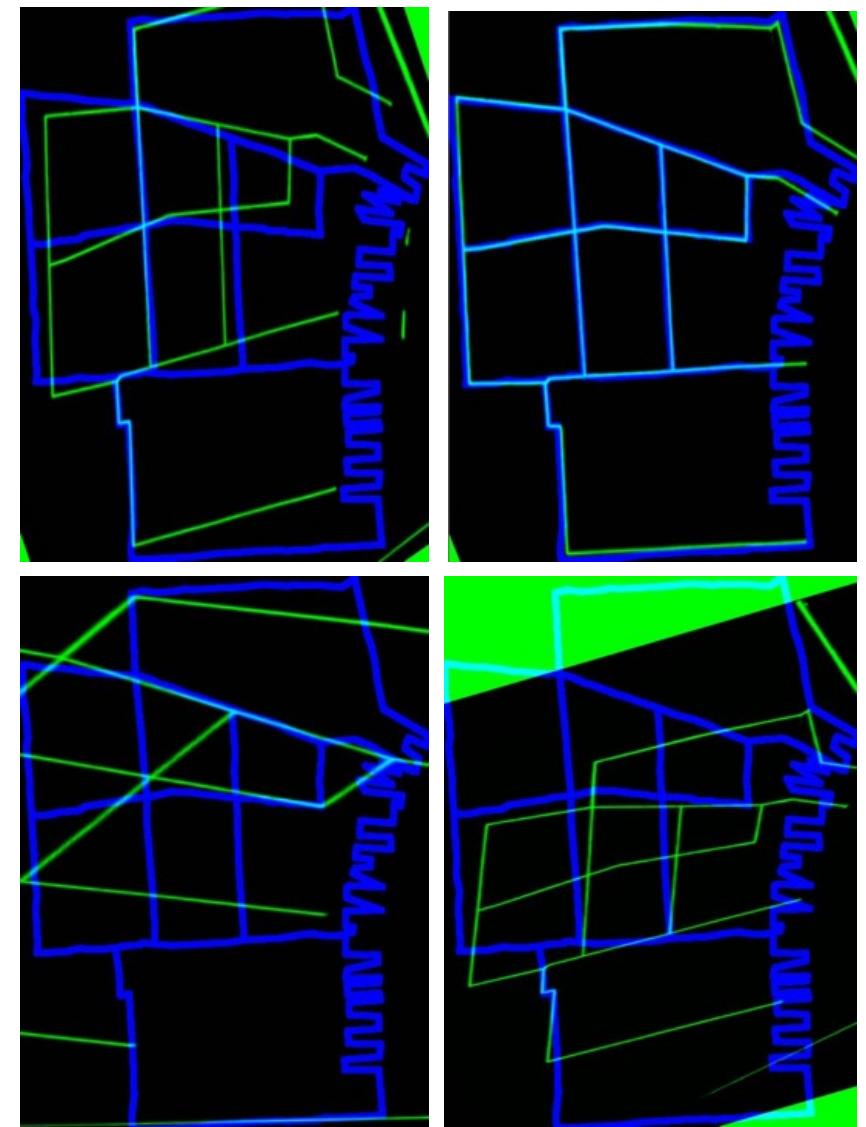
For selected links, how much do maps overlap?

3

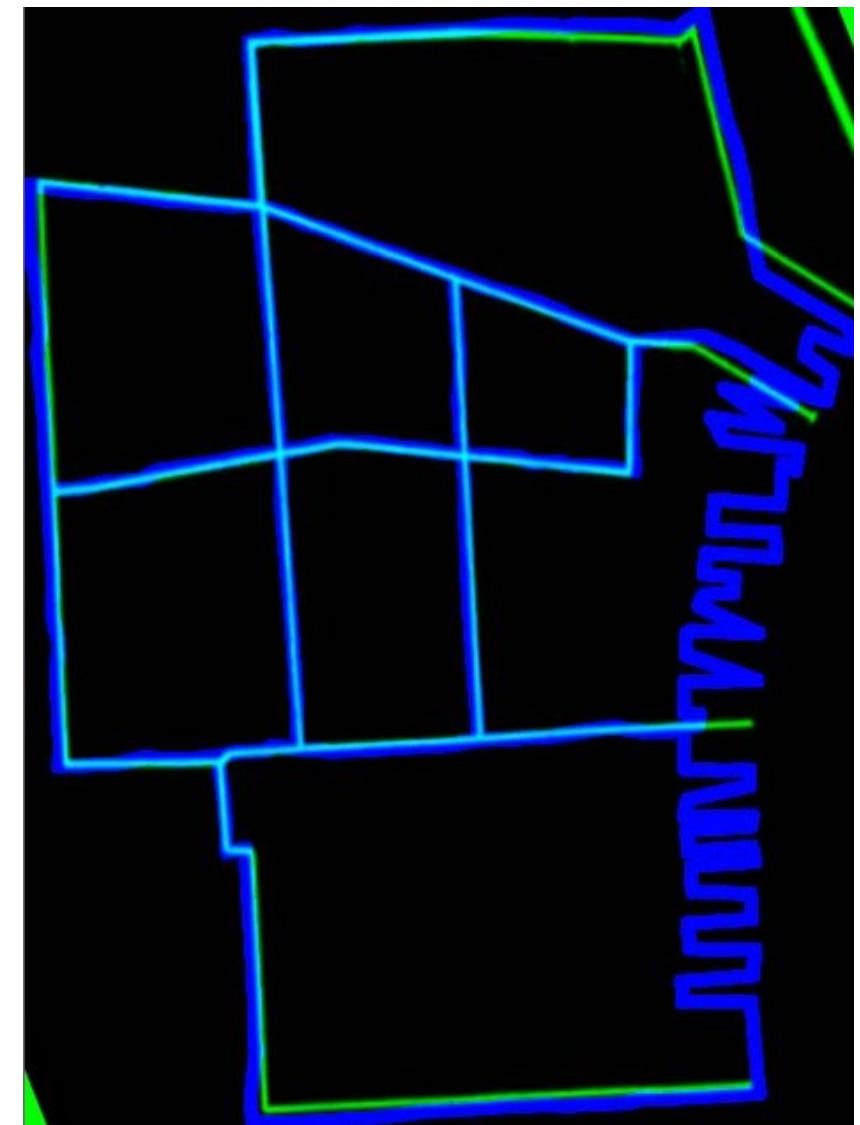
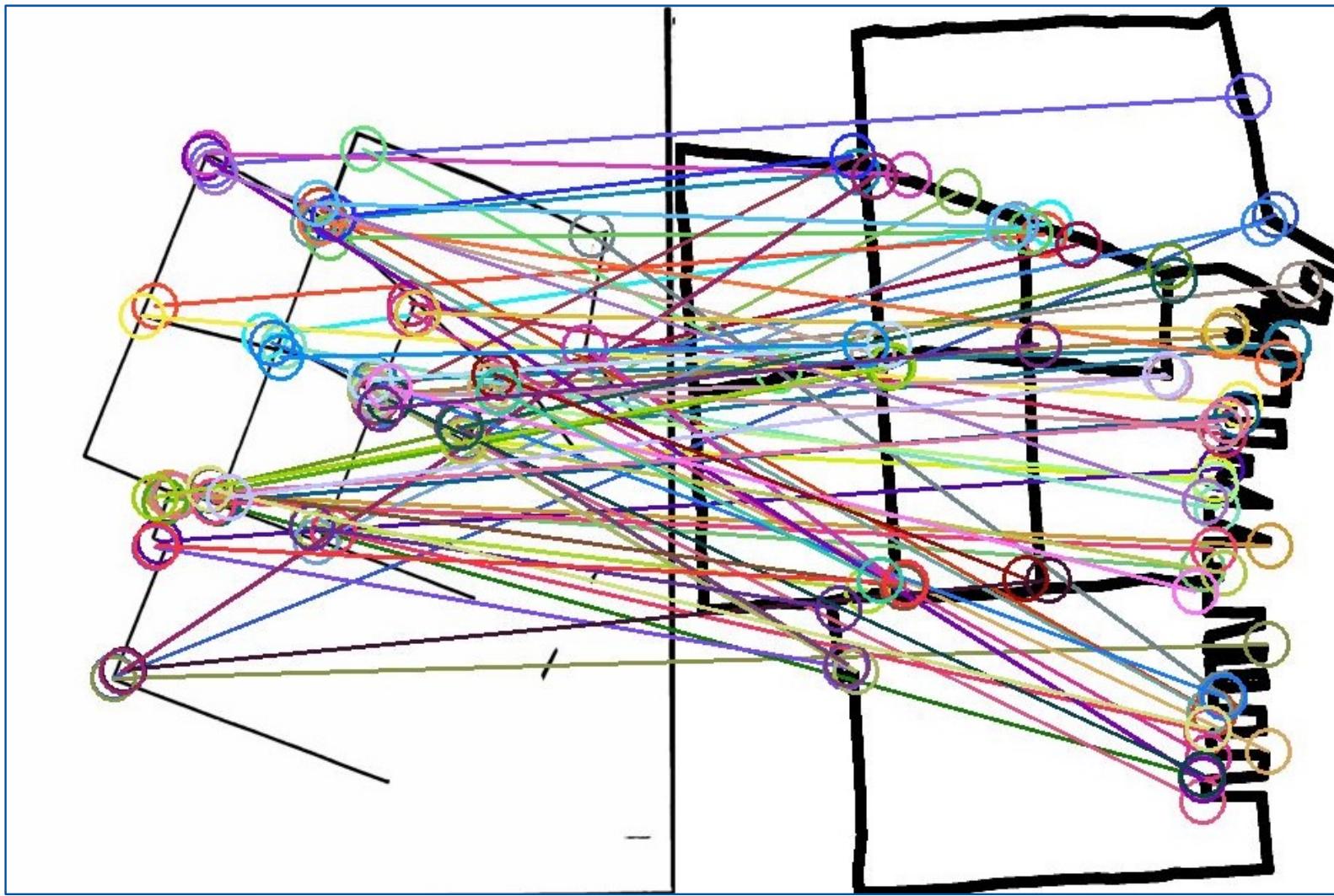
Repeat

Keep best guess.

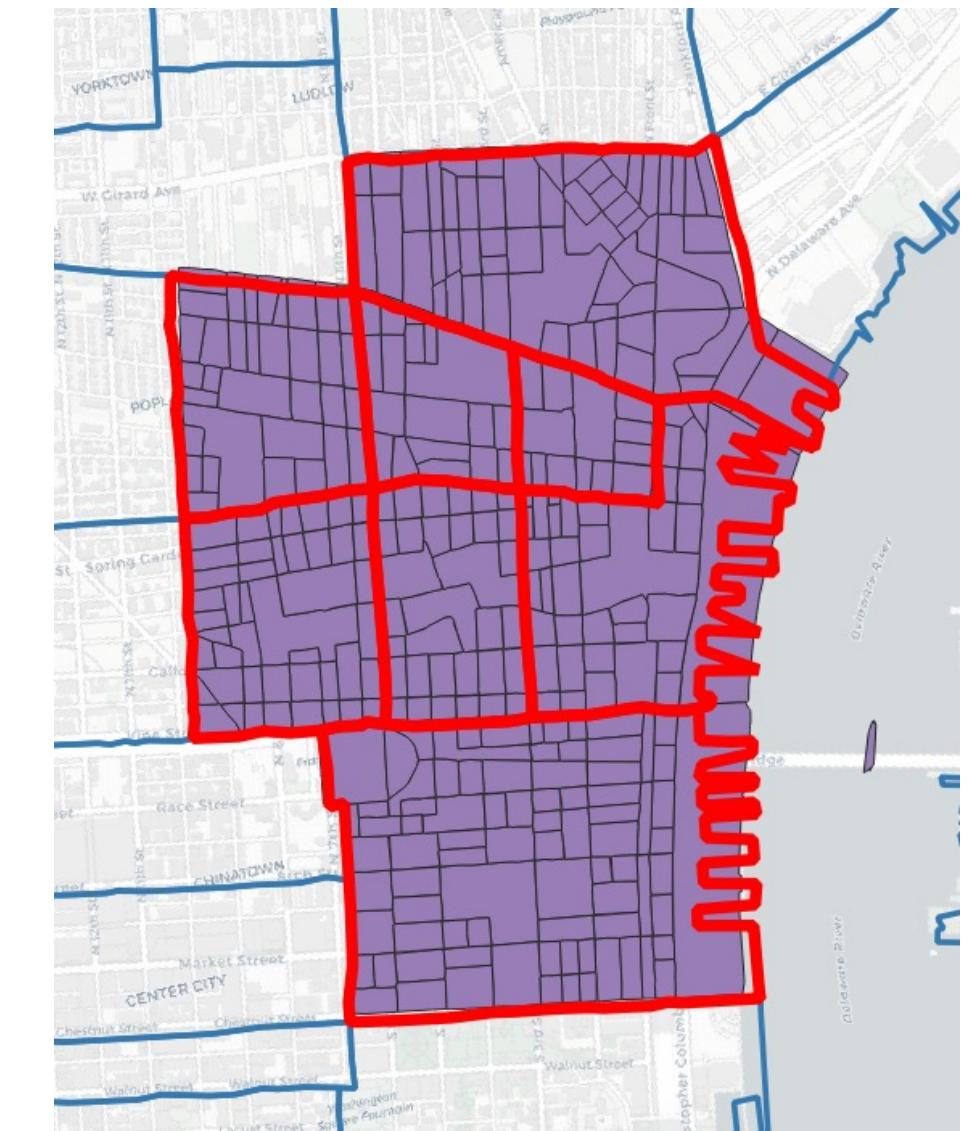
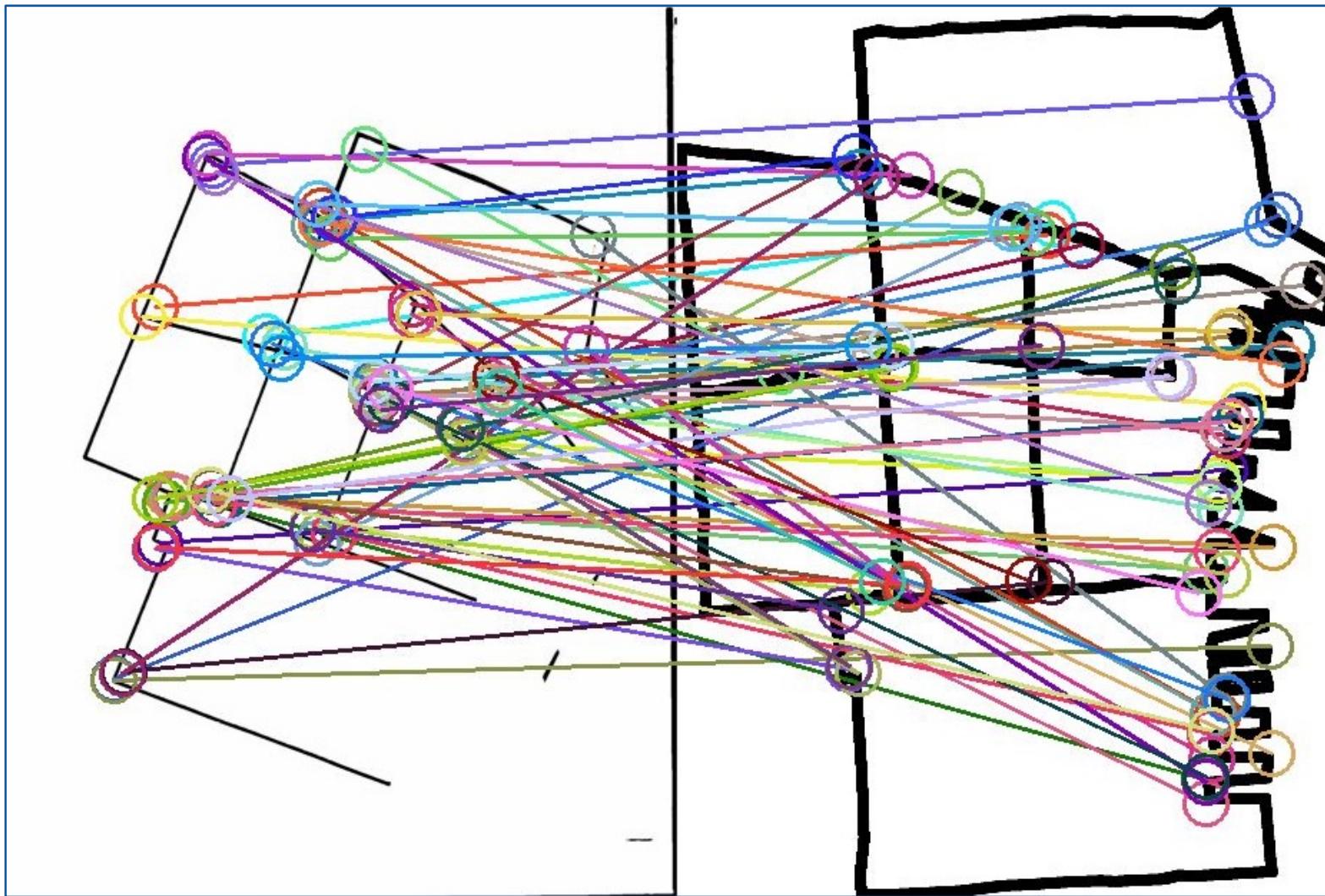
Selecting Links



Selecting Links



Selecting Links



3 Tasks, 3 Pieces of Data

1

Shape

2

Situation

3

Statistics

Digitizing Tables

Scale and Scope of Problem

- 1940, 1950, and 1960 Census of Housing, Block Statistics (1970 is digital)
- Sixteen Target Cities
 - New York City
 - Chicago
 - Philadelphia
 - Los Angeles
 - Detroit
 - Baltimore
 - Cleveland
 - St. Louis
 - Washington, DC
 - Boston
 - San Francisco
 - Pittsburgh
 - Houston
 - Cincinnati
 - Columbus, OH
 - Atlanta

Scale and Scope of Problem

- 1940, 1950, and 1960 Census of Housing, Block Statistics
- Sixteen Target Cities
 - New York City
 - Detroit
 - Washington, DC
 - Houston
 - Chicago
 - Baltimore
 - Boston
 - Cincinnati
 - Philadelphia
 - Cleveland
 - San Francisco
 - Columbus, OH
 - Los Angeles
 - St. Louis
 - Pittsburgh
 - Atlanta
- ~2,000 pages of tabular data, ~170,000 blocks, ~2.5 million cells per decade
- Structured, tabular form, with rows and columns properly associated and with accuracy better than 99%

Bottom Line Up Front

- Four stage process
 - Isolate table and each column
 - First pass with Tesseract
 - Algorithm to structure table
 - ML model to correct errors in OCR
- Great results
- Approach only makes sense if dataset is large and consistent

Bottom Line Up Front (1950)

- **Custom Solution**
 - 0.07% Observations with Error
 - 0.03% Character Error Rate

Bottom Line Up Front (1950)

- **Custom Solution**
 - 0.07% Observations with Error
 - 0.03% Character Error Rate
- **Data Entry**
 - 0.12% Observations with Error
 - 0.13% Character Error Rate
- **Tesseract (with assist with table structure)**
 - 12.94% Observations with Error
 - 7.24% Character Error Rate

Why Not Use Out of The Box Solutions?

Why Not Use Out of The Box Solutions?

- Adobe:

Census tract	Block	one-dwelling-structures
		Average value (dollars)
10 - B	11	4,425
	12	
	15	
	16	7,288
	17	8,150
	18	
	19	
	20	
	21	
	22	
	23	
	24	6,366
	25	3,900
	26	
	28	
	29	
	30	
	31	
	32	
	34	

Cen1111 tract	Block	one-dwellInc ture1
		Av. value (dollar)
10-e	11	4,425
11-A	12	7,288
11-e	15	8,150
12-A	16	6,366
	17	3,900
	18	9,500
	19	8,630
	20	6,483
	21	4,777
	22	5,875
	23	4,261
	24	5,345
	25	5,166
	26	4,800
	28	7,800
	29	6,000
	JO 31	4,666
	32	4,125

Why Not Use Out of The Box Solutions?

- Adobe:

Census tract	Block	one-dwelling-structures
		Average value (dollars)
10 - B	11	4,425
	12	
	15	
	16	7,288
	17	8,150
	18	
	19	
	20	
	21	
	22	
	23	
	24	6,366
	25	3,900
	26	
	28	
	29	
	30	
	31	
	32	
	34	

Cen1111 tract	Block	one-dwellInc ture1
		Av. value (dollar)
10-e	11	4,425
11-A	12	7,288
11-e	15	8,150
12-A	16	6,366
	17	3,900
	18	9,500
	19	8,633
	20	6,483
	21	4,777
	22	5,875
	23	4,261
	24	5,345
	25	5,166
	26	4,800
	28	7,800
	29	6,000
	JO 31	4,666
	32	4,125

Why Not Use Out of The Box Solutions?

- Adobe: Bad character recognition, relation of rows lost

Census tract	Block	one-dwelling-structures	Cen1111 tract	Block	one-dwelling-structure1
		Average value (dollars)			Av. value (dollar)
10 - B	11	4,425	10 -	11	4,425
	12		11-A	12	7,288
	15		11-e	13	8,150
	16	7,288	12-A	16	6,366
	17	8,150		17	3,900
	18			18	9,500
	19			19	8,630
	20			20	6,483
	21			21	4,777
	22			22	5,875
	23			23	4,261
	24			24	5,345
	25	6,366		25	5,166
	26	3,900		26	4,800
	28			28	7,800
	29			29	6,000
	30			JO	4,666
	31			31	
	32			32	4,125
	34				

Why Not Use Out of The Box Solutions?

- Textract:

Census tract	Block	All dwelling units by occupancy and tenure					All dwelling units by condition and plumbing facilities			Occupied dwelling units				Contract monthly rent ¹		Value ² of one-dwelling-unit structures	
		Total	Owner occupied	Renter occupied	Vacant non-seasonal not dilap., for rent or sale	Other vacant and non-resident	Number reporting	No private bath or dilap.	No running water or dilap.	Total	Persons per room	Number reporting	1.51 or more	Occupied by non-white	Number reporting	Average monthly rent (dollars)	Number reporting
46-G	15	91	42	46	2	1	75	2	88	86	3	41	4434	23	8630		
	16	80	44	34		2	71	2	78	77	4	34	4320	32	16050		
	17	100	36	57	5	2	94	1	93	93		55	4883	15	10866		
	18	75	43	30		2	66	1	73	73	4	30	5093	31	10064		
	19	79	66	11	2		72	1	77	77	2	7	3828	40	7040		
	20	75	42	19	4	10	56	1	61	56	1	16	4443	32	6921		
	21	66	59	7			62	3	66	66	1	7	4485	48	7479		
	22	42	35	7			42	1	42	42		6	4200	30	8300		
	23	50	46	1		3	42		47	44				26	8192		
	24	64	53	10	1		62	6	63	63		10	5310	23	7434		
	25	57	47	10			54		57	55	1			40	9575		
	26	106	47	58		1	98	2	105	97	2	57	3922	20	9600		
	27	64	39	25			58	3	64	64	1	24	4129	30	6543		
	28	77	32	44		1	76	1	76	76	2	43	4388	20	7570		
	29	79	39	36	3	1	59	2	75	69	3	33	4612	23	7869		
	30	58	42	10	1	5	45		52	47	1	9	4733	26	8323		
	31	49	42	7			49	1	49	49	1	6	4683	41	8402		
	32	62	45	17			60	5	62	61	2	16	4643	39	8128		

Why Not Use Out of The Box Solutions?

- Textract: when it works it works!
- 1.5% error rate, 0.22% ignoring cell alignment errors (stats for this page only)

Census tract	Block	All dwelling units by occupancy and tenure						All dwelling units by condition and plumbing facilities			Occupied dwelling units			Contract monthly rent			Value of one-dwelling unit structures	
		Total	Owner occupied	Renter occupied	Vacant non-seasonal not dilap.	Other vacant and non-resident	Number reporting	No private bath or dilap.	No running water for dilap.	Total	Persons per room	Occupied by non-white	Number reporting	Average monthly rent (dollars)	Number reporting	Average value (dollars)		
46-G	46 - G	15	91	42	46	2	1	75	2	88	8.6	3	41	4,434	23	8,630		
	16	80	44	34	57	5	2	71	1	76	7.7	4	54	4,320	32	16,050		
	17	100	36	57	5	2	94	1	1	93	9.3	1	55	4,883	15	10,866		
	18	75	43	30	2	2	66	1	1	73	7.3	4	30	5,093	31	10,064		
	19	79	66	11	2	2	72	1	1	77	7.7	2	18	3,828	40	7,040		
	20	75	42	19	4	10	56	3	2	61	5.6	1	16	4,443	32	6,921		
	21	66	59	7	2	2	62	3	2	66	6.6	1	7	4,485	48	7,479		
	22	42	35	7	3	3	42	1	1	42	4.2	6	42	4,200	30	8,300		
	23	50	46	10	1	3	42	6	6	47	4.4	10	26	8,192	23	7,434		
	24	64	53	10	1	1	62	6	6	63	6.3	10	53	5,310	23	7,434		
	25	57	47	10	1	1	54	2	2	57	5.5	1	9	4,022	40	9,575		
	26	106	47	58	1	1	98	2	2	105	9.7	2	57	3,922	20	9,600		
	27	64	39	25	1	1	58	3	3	64	6.4	1	24	4,129	30	6,543		
	28	77	32	44	1	1	76	1	1	76	7.6	2	43	4,386	20	7,576		
	29	79	39	36	3	1	59	2	2	75	6.9	3	33	4,612	23	8,864		
	30	58	42	10	1	5	45	1	1	52	4.7	1	9	4,733	26	8,323		
	31	49	42	7	17	1	49	1	1	49	4.9	1	6	4,683	41	8,402		
	32	62	45	17	17	1	60	5	6	62	6.1	2	16	4,643	39	8,128		

Why Not Use Out of The Box Solutions?

- Textract: when it doesn't work...

Census tract	Block	All dwelling units by occupancy and tenure					All dwelling units by condition and plumbing facilities			Occupied dwelling units				Contract monthly rent ¹		Value ² of one-dwelling-unit structures	
		Total	Owner occupied	Renter occupied	Vacant non-seasonal not dilap., for rent or sale	Other vacant and non-resident	Number reporting	No private bath or dilap.	No running water or dilap.	Total	Persons per room		Occupied by non-white	Number reporting	Average monthly rent (dollars)	Number reporting	Average value (dollars)
											Number reporting	1.51 or more					
33-1	2	60	43	17			60		60	60	17	32.41	42	4,821			
	5	35	27	8			34	2	35	34	8	34.62	19	4,410			
	6	77	55	22			76	3	77	75	22	35.45	46	5,260			
	7	37	30	6	1		37	2	36	36	7	31.71	24	5,020			
	8	24	22	2			24		24	24	2	31.00	17	5,500			
	9	26	21	5			26	1	26	26	5	31.00	19	5,894			
	11	10	7	3			10		10	10	3	61.66	7	6,142			
	12	49	7	40	1	1	41	3	47	46	41	25.70	5	3,180			
	14	9	6	3			9		9	8	3	40.00	6	7,483			
	15	49	32	17			49		49	49	17	26.35	32	3,728			
	16	56	34	20	1	1	56		54	54	20	38.30	25	4,980			
	17	18	7	11			18		18	18	11	35.18	4	4,875			
	18	19	11	8			19		19	19	8	30.12	11	6,045			
	19	38	34	4			38		38	37	4	30.50	32	5,812			
	20	30	23	7			30		30	30	4	33.00	22	7,463			
	21	63	40	22			60	1	62	59	20	29.05	35	5,200			
	22	45	17	28			45		45	44	28	29.48	16	7,500			
	23	26	25	1			26		26	26	1	37.50	23	7,026			
	24	26	19	6	1		26		25	25	6	37.50	19	6,868			

Why Not Use Out of The Box Solutions?

- Textract: when it doesn't work... it doesn't work! Small input tweaks do not fix error.

Census tract	Block	All dwelling units by occupancy and tenure				All dwelling units by condition and plumbing facilities			Occupied dwelling units			Contract monthly rent		Value of one-dwelling-unit structures	
		Owner occupied	Renter occupied	Vacant non-seasonal not for rent or sale	Other vacant and non-resident	Number reporting	No private bath or dilap.	No running water or dilap.	Total	Persons per room	Deslindado Occupied by non-white	Number reporting	Average monthly rent (dollars)	Number reporting	Average value (dollars)
33 - 1	25	60	43	17	8	60	00	60	60	1.7	1	17	42	4,821	
	35	35	27	8	22	34	2	35	34	1.8	2	452	19	4,410	
	66	77	55	6	1	76	2	77	75	1.5	1	545	46	5,260	
	77	37	30	2	2	37	3	36	36	1.9	2	211	21	5,020	
	88	24	22	5	1	24	4	24	24	1.9	3	100	19	5,500	
	99	26	21	7	3	26	5	26	26	1.9	4	100	17	5,894	
	111	10	7	4	1	10	6	10	10	1.9	5	100	12	6,142	
	112	49	47	7	1	41	7	47	46	1.9	6	100	10	3,480	
	114	45	44	9	1	9	8	9	8	1.9	7	100	9	2,866	
	115	32	32	49	1	49	9	49	49	1.9	8	100	8	2,866	
	16	56	34	30	1	56	18	54	54	1.9	9	300	25	4,980	
	17	18	7	11	1	18	18	18	11	1.9	10	300	4	4,875	
	18	19	11	8	1	19	19	19	8	1.9	11	300	11	6,045	
	19	38	34	4	1	38	38	37	4	1.9	12	300	32	5,812	
	20	30	23	202	1	30	30	30	30	1.9	13	300	22	7,463	
	21	63	40	32	1	60	62	59	1	1.9	14	300	35	5,200	
	22	45	17	28	1	45	45	44	28	1.9	15	300	16	7,500	
	23	25	14	26	1	26	26	26	1	1.9	16	300	6	6,668	
	24	26	14	25	1	25	25	25	1	1.9	17	300	1	6,668	

Why Not Use Out of The Box Solutions?

- Textract:
 - Sample Size: 169 Pages
 - Catastrophic Failures: 45
 - Moderate Failures: 6
 - **Unacceptable page level error: 30%**
 - Small errors in table layout can be algorithmically corrected, catastrophic failures cannot

Method

- Isolate table
- Isolate columns
- Tesseract columns
- Structure into table
- Match to labeled data
- Train model to correct Tesseract errors
- Visualize and correct issues throughout
- Final check for internal consistency and vs tract

Isolate Table Body, Straighten Image

20

City Block Characteristics

Table 2.—CHARACTERISTICS OF HOUSING UNITS, BY BLOCKS: 1960—Con.

[“Total population” contains no persons in group quarters unless preceded by asterisk: one asterisk (*) denotes less than 10 percent; two asterisks (**), 10 percent or more]

Blocks within census tracts	Total popula- tion	All housing units by condition and plumbing									Occupied housing units							
		Sound			Deteriorating			Dilap- idated	Owner occupied			Renter occupied			Occu- pied by non- white	1.0 or more per- sons per room		
		Total	With all plumb- ing facili- ties	Lack- ing some or all facili- ties	Total	With all plumb- ing facili- ties	Lacking some or all facilities		Total	Average value (dollars)	Aver- age num- ber of rooms	Total	Average con- tract rent (dollars)	Aver- age num- ber of rooms				
21***	30	9	8	8	1	1	1	6	3500	5.3	3	***	***	***	***	***	***	
22***	87	28	22	22	6	5	5	14	6000	7.1	12	43	4.2	3	3	3	3	
23***	124	44	34	30	4	4	5	1	22	6000	7.0	16	28	5.3	1	1	1	
24***	**120	35	26	24	2	9	9	***	20	4500	7.1	12	40	4.4	1	1	1	
25***	247	55	38	38	15	14	1	2	41	5000	7.5	11	47	5.6	2	2	2	
26***	145	40	25	25	10	9	1	5	28	5000	6.9	8	39	5.9	5	5	5	
27***	85	20	12	12	7	7	1	1	16	5000	7.0	4	***	***	5	5	5	
28***	21	7	2	2	5	5	1	1	***	***	4	***	***	***	2	2	2	
29***	14	4	***	***	***	***	***	1	***	***	***	***	***	***	2	2	2	
30***	51	13	12	12	***	***	***	1	9	4000	5.9	2	***	***	***	2	2	
31***	18	4	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	
34***	10	4	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	
35***	9	2	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	
36***	54	18	7	7	8	8	8	3	5	6500	8.4	4	***	***	***	***	***	
37***	60	15	11	11	4	3	1	***	9	5500	7.6	6	37	6.2	3	2	2	
38***	133	52	38	36	2	13	8	5	31	5000	7.4	9	41	5.3	2	3	3	
43***	35	12	5	5	4	2	1	3	2	***	***	5	29	6.0	3	3	3	
44***	6	4	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	
45***	3	1	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	
49***	141	43	17	16	1	26	20	3	15	5500	8.3	21	31	4.9	1	4	4	
50***	84	35	14	13	1	21	4	10	12	3500	5.0	15	22	4.4	4	4	4	
51***	22	9	7	7	2	1	1	***	5	***	7.6	1	***	***	***	***	***	
18-B...*	*8802	2781	2358	2287	71	338	248	57	33	85	1823	5000	6.5	727	40	4.7	2	171
1***	111	53	38	33	5	14	12	2	1	18	5000	6.6	28	25	2.5	2	171	

Isolate Table Body, Straighten Image

Pass through Textract

20

City Block Characteristics

Table 2.—CHARACTERISTICS OF HOUSING UNITS, BY BLOCKS: 1960—Con.

"Total population" contains no persons in group quarters unless preceded by asterisk: one asterisk (*) denotes less than 10 percent; two asterisks (**), 10 percent or more.

Blocks within census tracts	Total popu- lation	All housing units by condition and plumbing												Occupied housing units											
		Total		Sound				Deteriorating				Dilap- idated	Owner occupied			Renter occupied				Total	Average age num- ber of rooms	Aver- age con- tract rent (dollars)	Aver- age num- ber of rooms	Occu- pied by non- white per- sons per room	
		Total	With all plumb- ing facili- ties	ack- ing some or all facili- ties	With some or all facili- ties	Total	With all plumb- ing facili- ties	Lacking some or all facilities	With flush toilet	No flush toilet	Total		Average value (dollars)	Aver- age num- ber of rooms											
21***	30	9	8	8	22	22	...	15	11	5	5	...	6	3500	5.3	3	43	4.2	3	1.0	
22***	87	28	22	22	22	22	...	14	14	14	14	...	14	6000	7.1	12	45	4.2	3	1.0	
23***	124	44	34	34	30	30	2	14	14	14	14	...	22	6000	7.0	16	40	4.4	12	1.0	
24***	**120	35	26	24	24	24	2	15	14	14	14	...	20	4500	7.1	12	47	4.7	12	1.0	
25***	247	55	38	38	38	38	...	14	14	14	14	...	41	5000	7.6	11	47	4.7	11	1.0	
26***	145	40	25	25	25	25	...	15	14	14	14	...	28	5000	6.9	8	39	5.6	8	1.0	
27***	85	20	12	12	12	12	...	12	11	11	11	...	16	5000	7.0	4	33	5.9	5	1.0	
28***	21	7	2	2	2	2	...	5	5	5	5	...	11	4	33	5.9	2	1.0	
29***	14	5	3	3	3	3	...	12	12	12	12	...	15	4000	5.9	2	33	5.9	2	1.0	
30***	6	3	3	3	3	3	...	12	12	12	12	...	15	4000	5.9	2	33	5.9	2	1.0	
31***	16	4	3	3	3	3	...	12	12	12	12	...	15	1.0	
34***	10	4	2	2	2	2	...	12	12	12	12	...	15	1.0	
35***	9	2	2	2	2	2	...	12	12	12	12	...	15	1.0	
36***	54	18	7	7	7	7	...	12	12	12	12	...	15	6500	8.4	4	37	6.2	2	1.0	
37***	60	15	11	11	11	11	...	12	12	12	12	...	15	5500	7.6	6	41	5.3	2	1.0	
38***	132	52	38	38	36	36	2	11	11	11	11	...	15	5000	7.4	9	41	5.3	2	1.0	
43***	35	12	5	5	5	5	...	12	12	12	12	...	15	5000	7.4	5	29	6.0	3	1.0	
44***	6	4	1	1	1	1	...	12	12	12	12	...	15	5	31	4.9	4	1.0	
45***	3	1	1	1	1	1	...	12	12	12	12	...	15	5	31	4.9	4	1.0	
49***	141	45	17	16	16	16	1	26	20	10	7	...	15	5500	8.3	21	31	4.9	4	1.0	
50***	84	35	14	13	11	11	2	21	14	10	7	...	15	3500	5.0	15	22	4.4	4	1.0	
51***	22	8	7	7	7	7	...	12	12	12	12	...	15	3500	5.0	1	22	4.4	4	1.0	
18-B...*	18802	2781	2358	2287	71	538	538	1	338	248	57	33	85	1823	5000	6.5	727	40	4.7	2	1.0
	1...	111	53	38	38	6	14	12	2	57	11	11	11	18	5000	6.6	32	40	4.7	2	1.0

Isolate Table Body, Straighten Image

Find where, vertically, we go from mostly alpha to mostly numeric

20

City Block Characteristics

Table 2.—CHARACTERISTICS OF HOUSING UNITS, BY BLOCKS: 1960—Con.

"Total population" contains no persons in group quarters unless preceded by asterisk: one asterisk (*) denotes less than 10 percent; two asterisks (**), 10 percent or more.

Mostly Text

Blocks within census tracts	Total popu- lation	All housing units by condition and plumbing												Occupied housing units							
		Total		Sound				Deteriorating				Dilap- idated	Owner occupied			Renter occupied			OCCU- PATION more per room		
		Total	Total	With all plumb- ing facili- ties	ack- ing some or all facili- ties	With some or all facili- ties	With all plumb- ing facili- ties	Lacking some or all facilities	With flush toilet	No flush toilet	Total	Average value (dollars)	Aver- age num- ber of rooms	Total	Average con- tract rent (dollars)	Aver- age num- ber of rooms	Occupied by non- white per- sons per room				
21***	30	9	8	22	8	...	15	11	5	5	6	3500	5.3	3	45	4.2	1.5	1.5			
22***	87	28	22	22	22	...	16	14	12	12	14	6000	7.1	12	45	4.2	1.5	1.5			
23***	124	44	34	34	30	2	14	14	12	12	22	6000	7.0	12	45	4.2	1.5	1.5			
24***	**120	35	26	24	24	2	15	14	12	12	20	4500	7.1	12	40	5.3	1.1	1.2			
25***	247	55	38	38	38	...	15	14	12	12	41	5000	7.6	12	47	4.4	1.1	1.2			
26***	145	40	25	25	25	...	16	14	12	12	28	5000	6.9	8	39	5.6	1.1	1.5			
27***	85	20	12	12	12	...	17	16	14	14	16	5000	7.0	4	44	5.9	1.1	1.5			
28***	21	7	2	2	2	...	15	14	12	12	11	4000	5.9	2	44	5.9	1.1	1.2			
29***	14	5	3	3	3	...	12	12	12	12	12	4000	5.9	2	44	5.9	1.1	1.2			
30***	6	3	3	3	3	...	12	12	12	12	12	4000	5.9	2	44	5.9	1.1	1.2			
31***	16	4	3	3	3	...	12	12	12	12	12	4000	5.9	2	44	5.9	1.1	1.2			
34***	10	4	2	2	2	...	12	12	12	12	12	4000	5.9	2	44	5.9	1.1	1.2			
35***	9	2	2	2	2	...	12	12	12	12	12	4000	5.9	2	44	5.9	1.1	1.2			
36***	54	18	7	7	7	...	12	12	12	12	12	6500	8.4	4	44	6.2	1.1	1.2			
37***	60	15	11	11	11	...	12	12	12	12	12	5500	7.6	6	41	5.3	1.1	1.2			
38***	132	52	38	36	36	2	12	12	12	12	51	5000	7.4	9	41	5.3	1.1	1.2			
43***	35	2	5	5	5	...	12	12	12	12	12	5000	7.4	5	29	6.0	1.1	1.2			
44***	6	1	1	1	1	...	12	12	12	12	12	5000	7.4	5	31	4.9	1.1	1.2			
45***	3	1	1	1	1	...	12	12	12	12	12	5500	8.3	2	31	4.9	1.1	1.2			
49***	141	45	17	16	16	1	26	20	10	7	15	5500	8.3	21	31	4.9	1.1	1.2			
50***	84	35	14	13	11	11	21	14	10	7	12	3500	5.0	15	22	4.4	1.1	1.2			
51***	22	8	7	7	7	...	12	12	12	12	12	3500	5.0	11	22	4.4	1.1	1.2			
18-B-***	**8802	2781	2358	2287	71	53	111	53	57	33	1823	5000	6.5	727	40	4.7	2	171			

Mostly Numbers

Isolate Table Body, Straighten Image

Table is isolated

20

City Block Characteristics

Table 2.—CHARACTERISTICS OF HOUSING UNITS, BY BLOCKS: 1960—Con.

"Total population" contains no persons in group quarters unless preceded by asterisk: one asterisk (*) denotes less than 10 percent; two asterisks (**), 10 percent or more.

Blocks within census tracts	Total popu- lation	All housing units by condition and plumbing												Occupied housing units									
		Sound		Deteriorating				Dilap- idated	Owner occupied			Renter occupied			Occupied by more per- sons per room								
		Total	Total	With all plumb- ing facili- ties	ack- ing some or all facili- ties	Total	With all plumb- ing facili- ties	Lacking some or all facilities	With flush toilet	No flush toilet	Total	Average value (dollars)	Aver- age num- ber of rooms	Total	Average con- tract rent (dollars)	Aver- age num- ber of rooms	Total	Occupied by non- white	Occupied by white				
21***	50	9	8	22	8	1	1		6	3500	5.3	5	4	3500	5.3	5	4	4.2	4.2	1	1		
22***	87	28	32	22	22	1	1		14	6000	7.0	12	45	6000	7.0	12	45	4.2	4.2	1	1		
23***	124	44	32	30	24	2	1		22	6000	7.0	16	40	4500	7.1	16	40	4.4	4.4	1	1		
24***	**120	35	26	24	24	2	1		20	5000	7.1	12	47	5000	7.1	12	47	5.3	5.3	1	1		
25***	247	55	38	38	38	1	1		41	5000	7.4	11	39	5000	7.4	11	39	5.6	5.6	1	1		
26***	145	40	25	25	25	1	1		28	5000	6.9	8	39	5000	6.9	8	39	5.9	5.9	1	1		
27***	85	20	12	12	12	1	1		16	5000	7.0	4	33	5000	7.0	4	33	5.9	5.9	1	1		
28***	21	7	2	2	2	1	1		11	4000	5.9	2	32	4000	5.9	2	32	5.9	5.9	1	1		
29***	14	4	1	1	1	1	1		1	3500	5.3	1	31	3500	5.3	1	31	5.3	5.3	1	1		
30***	5	3	1	1	1	1	1		1	3500	5.3	1	31	3500	5.3	1	31	5.3	5.3	1	1		
31***	18	4	1	1	1	1	1		1	3500	5.3	1	31	3500	5.3	1	31	5.3	5.3	1	1		
34***	10	4	1	1	1	1	1		1	3500	5.3	1	31	3500	5.3	1	31	5.3	5.3	1	1		
35***	8	2	1	1	1	1	1		1	3500	5.3	1	31	3500	5.3	1	31	5.3	5.3	1	1		
36***	54	16	7	7	7	1	1		5	6500	8.4	4	37	5500	7.6	6	37	6.2	6.2	1	1		
37***	60	15	7	7	7	1	1		9	5500	7.6	9	41	5000	7.4	9	41	5.3	5.3	1	1		
38***	132	52	38	36	36	2	1		31	5000	7.4	9	37	5000	7.4	9	37	6.0	6.0	1	1		
43***	35	12	5	5	5	1	1		2	3500	5.3	1	31	3500	5.3	1	31	5.3	5.3	1	1		
44***	6	4	1	1	1	1	1		1	3500	5.3	1	31	3500	5.3	1	31	5.3	5.3	1	1		
45***	3	1	1	1	1	1	1		1	3500	5.3	1	31	3500	5.3	1	31	5.3	5.3	1	1		
49***	141	43	17	16	16	1	1		15	3500	5.3	1	31	3500	5.3	1	31	5.3	5.3	1	1		
50***	84	35	14	13	13	1	1		12	3500	5.0	15	22	4,4	4.4	15	22	4.4	4.4	1	1		
51***	22	9	7	7	7	1	1		5	3500	5.0	11	11	3500	5.0	11	11	3500	5.0	11	11		
18-B***	28802	2781	2358	2284	71	536	248	57	33	1823	5000	6.5	727	40	4.7	2	171	5.0	5.0	1	1		
1***	111	53	38	33	5	14	12	2	11	5000	6.6	29	40	4.7	2	171	5.0	5.0	1	1			

Isolate Table Body, Straighten Image

Find the rotated bounding box that contains all the body bounding boxes

20

City Block Characteristics

Table 2.—CHARACTERISTICS OF HOUSING UNITS, BY BLOCKS: 1960—Con.

"Total population" contains no persons in group quarters unless preceded by asterisk: one asterisk (*) denotes less than 10 percent; two asterisks (**), 10 percent or more.

Blocks within census tracts	Total popu- lation	All housing units by condition and plumbing										Occupied housing units									
		Sound		Deteriorating						Owner occupied				Renter occupied							
		Total	Total	With all plumb- ing facili- ties	Lack- ing some or all facili- ties	Total	With all plumb- ing facili- ties	Lacking some or all facilities	Dilap- idated	Total	Average value (dollars)	Aver- age num- ber of rooms	Total	Average con- tract rent (dollars)	Aver- age num- ber of rooms	Occu- pied by non- white per- sons per room					
21***	50	9	8	8	1	1	1	1	1	6	3500	5.3	3	11	43	4.2	11	1	1		
22***	87	28	22	22	1	1	1	1	1	14	6000	7.1	2	40	4.4	12	1	1			
23***	124	44	31	30	4	1	1	1	1	22	6000	7.0	1	47	5.6	12	1	1			
24***	**120	95	26	29	2	1	1	1	1	20	4500	7.1	1	39	5.9	11	1	1			
25***	247	55	38	38	1	1	1	1	1	41	5000	7.5	1	47	4.4	11	1	1			
26***	145	40	25	25	1	1	1	1	1	28	5000	6.9	1	39	5.6	11	1	1			
27***	85	20	12	12	1	1	1	1	1	16	5000	7.0	1	44	5.9	11	1	1			
28***	21	7	2	2	1	1	1	1	1	11	4000	5.9	1	33	5.2	11	1	1			
29***	14	5	3	3	1	1	1	1	1	12	4000	5.9	2	33	5.2	11	1	1			
30***	5	3	1	1	1	1	1	1	1	1	4000	5.9	1	33	5.2	11	1	1			
31***	18	4	1	1	1	1	1	1	1	1	4000	5.9	1	33	5.2	11	1	1			
34***	10	4	2	2	1	1	1	1	1	1	4000	5.9	1	33	5.2	11	1	1			
35***	9	2	1	1	1	1	1	1	1	1	4000	5.9	1	33	5.2	11	1	1			
36***	54	16	7	7	1	1	1	1	1	5	6500	8.4	4	37	6.2	11	2	1			
37***	60	15	5	5	1	1	1	1	1	5	5500	7.6	6	41	5.3	11	2	1			
38***	132	52	2	2	1	1	1	1	1	5	5000	7.4	9	41	6.0	11	3	1			
43***	35	2	1	1	1	1	1	1	1	2	4500	6.9	5	29	5.3	11	2	1			
44***	6	1	1	1	1	1	1	1	1	1	4500	6.9	5	31	4.9	11	4	1			
45***	3	1	1	1	1	1	1	1	1	1	4500	6.9	1	33	5.2	11	1	1			
49***	141	43	1	1	1	1	1	1	1	15	5500	8.3	21	31	4.9	11	4	1			
50***	84	35	14	13	1	1	1	1	1	12	3500	5.0	15	22	4.4	11	4	1			
51***	22	9	7	7	1	1	1	1	1	5	3500	5.6	11	33	5.2	11	4	1			
18-B***	**8602	2781	2358	2287	71	338	248	57	35	1823	5000	6.5	727	40	4.7	2	171	1	1		
1***	111	53	38	33	5	14	12	2	11	18	5000	6.6	28	41	4.7	2	171	1	1		

Isolate Table Body, Straighten Image

Rotate image around center of table – image is straightened

20

City Block Characteristics

Table 2.—CHARACTERISTICS OF HOUSING UNITS, BY BLOCKS: 1960—Con.

"Total population" contains no persons in group quarters unless preceded by asterisk: one asterisk (*) denotes less than 10 percent; two asterisks (**), 10 percent or more.

Blocks within census tracts	Total popu- lation	All housing units by condition and plumbing												Occupied housing units							
		Total		Sound				Deteriorating				Dilap- idated	Owner occupied			Renter occupied			1.0 or more per room		
		Total	Total	With all plumb- ing facili- ties	ack- ing some or all facili- ties	Total	With all plumb- ing facili- ties	Lacking some or all facilities	With flush toilet	No flush toilet	Total		Average value (dollars)	Aver- age num- ber of rooms	Total	Average con- tract rent (dollars)	Aver- age num- ber of rooms				
		Total	Total	Total	Total	Total	Total	Total	Total	Total	Total		Total	Total	Total	Total	Total	Total			
21***	30	9	8	8	8	...	8	8	8	8	8	...	6	3500	5.3	5	4	4	...		
22***	87	28	22	22	22	...	22	22	22	22	22	...	14	6000	7.1	12	11	11	...		
23***	124	44	34	34	30	4	4	4	4	4	4	...	22	6000	7.0	12	11	11	...		
24***	**120	55	26	24	24	2	2	2	2	2	2	...	20	4500	7.1	12	11	11	...		
25***	247	55	38	38	38	...	38	38	38	38	38	...	41	5000	7.5	11	10	10	...		
26***	145	40	25	25	25	...	25	25	25	25	25	...	28	5000	6.9	8	7	7	...		
27***	85	20	12	12	12	...	12	12	12	12	12	...	16	5000	7.0	4	4	4	...		
28***	21	7	2	2	2	...	2	2	2	2	2	...	11	4	4	4	...		
29***	14	5	3	3	3	...	3	3	3	3	3	...	5	4000	5.9	2	2	2	...		
30***	51	12	12	12	12	...	12	12	12	12	12	...	5	4000	5.9	2	2	2	...		
31***	16	4	3	3	3	...	3	3	3	3	3		
34***	10	4	2	2	2	...	2	2	2	2	2		
35***	9	2	1	1	1	...	1	1	1	1	1	...	5	4	4	4	...		
36***	54	18	7	7	7	...	7	7	7	7	7	...	5	6500	8.4	4	4	4	...		
37***	60	15	11	11	11	...	11	11	11	11	11	...	5	5500	7.6	6	6	6	...		
38***	132	52	38	36	36	2	2	2	2	2	2	...	51	5000	7.4	9	9	9	...		
43***	35	2	1	1	1	...	1	1	1	1	1	...	2	5	5	5	...		
44***	6	4	1	1	1	...	1	1	1	1	1	29	29	29	...		
45***	3	1	1	1	1	...	1	1	1	1	1		
49***	141	45	17	16	16	1	1	26	20	3	10	7	...	15	5500	8.9	21	21	21	...	
50***	84	35	14	13	13	1	1	21	4	10	7	7	...	12	3500	5.0	15	14	14	...	
51***	22	8	7	7	7	...	2	1	1	1	1	...	5	3500	5.6	1	22	4.4	...		
18-B*	**802	2781	2358	2287	71	538	38	37	6	14	12	11	85	1823	5000	6.5	727	40	4.7	171	

Always Be Checking

21

2

23

24

2

2

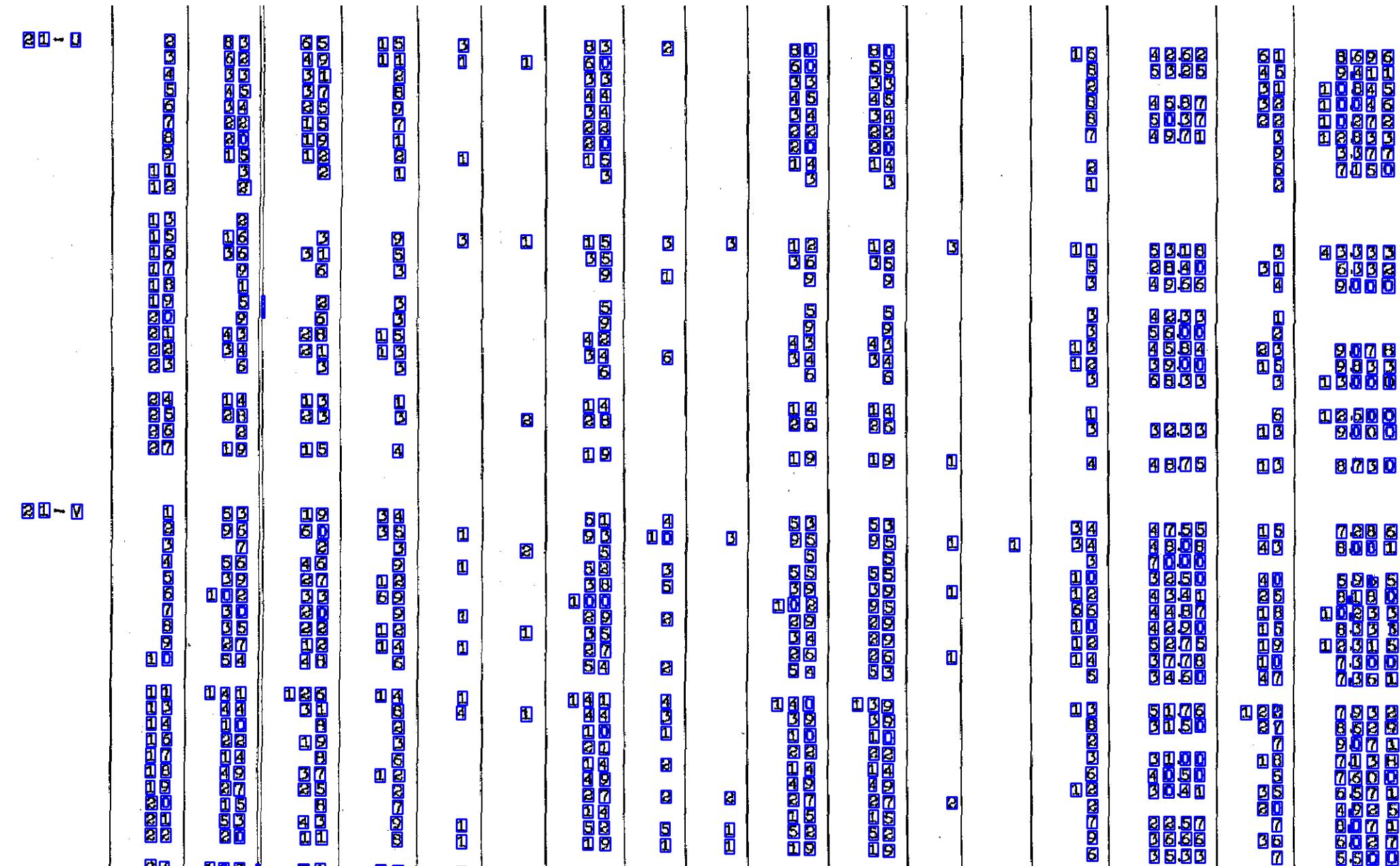
八

phia

Isolate Columns

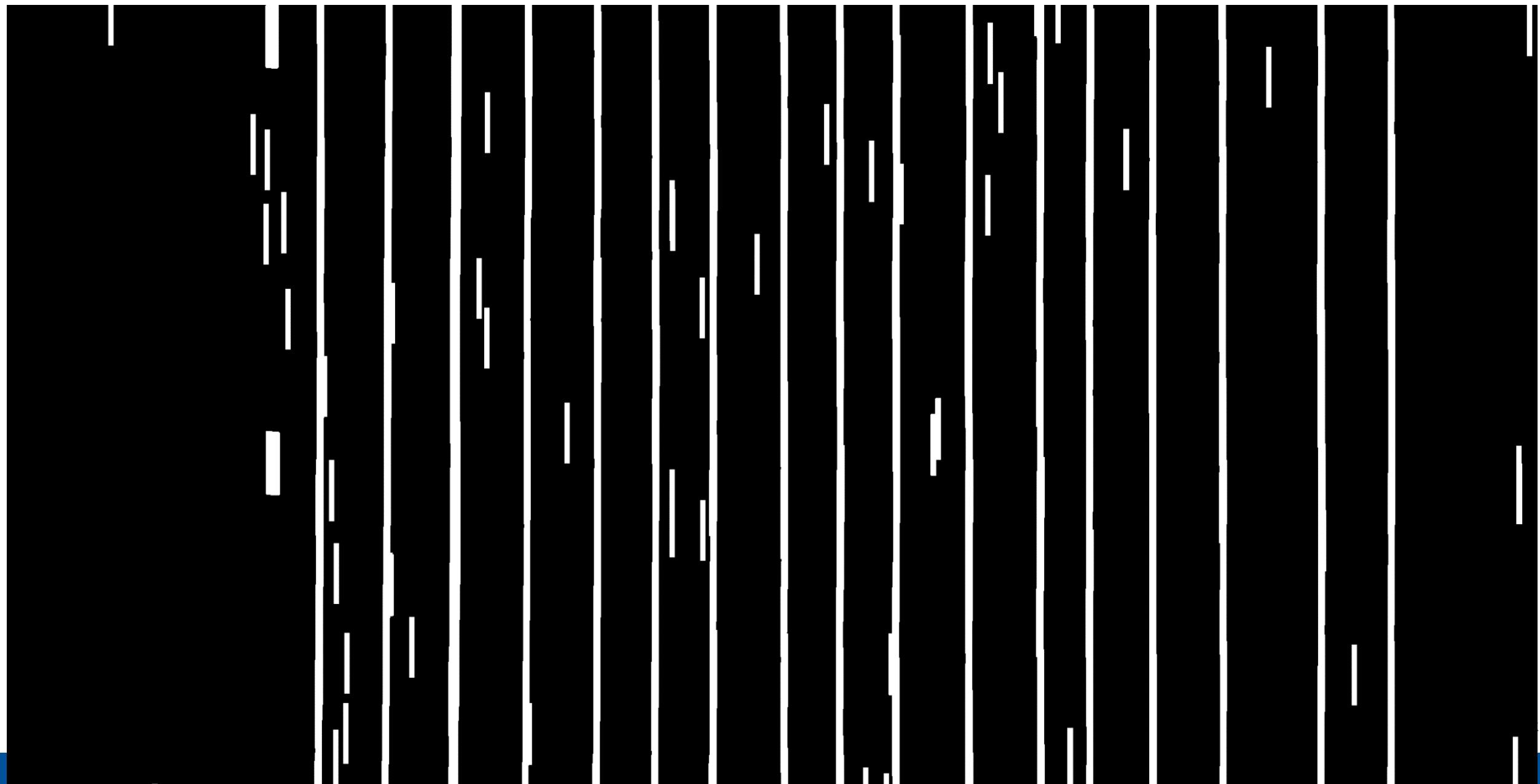
Isolate Columns

Find everything that *could* be a character. Be aggressive, recall is important



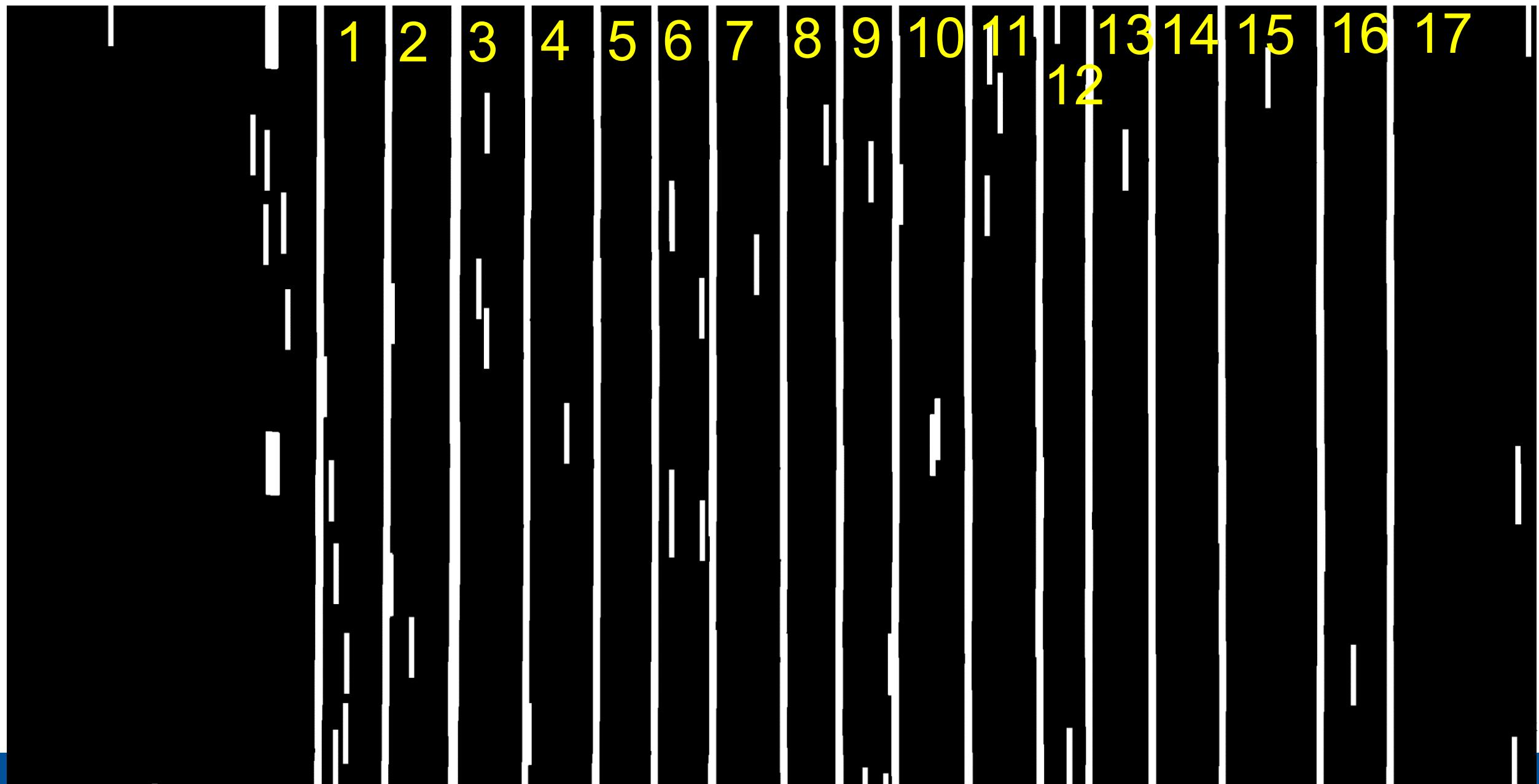
Isolate Columns

Isolate and smear (slightly horizontally, aggressively vertically) what is left



Isolate Columns

Find $(N - 1)$ longest lines that are nearly vertical, $N = \#$ of columns



Isolate Columns

Columns are isolated

Tesseract Each Row in Each Column

- Tesseract *highly* sensitive to input parameters, but flexible and governable
- Use restricted character set and character level confidence
- Collect character level text, bounding boxes and confidence

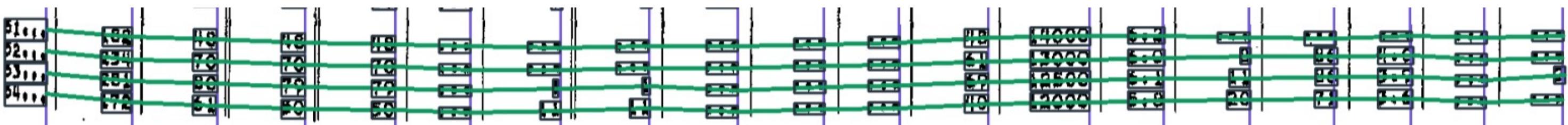
Use Table's Internal Structure to Build Rows and Columns

- Start with block column (always populated)
- Look right to find the two-way unique nearest neighbor for each row, requiring the angle to the nearest neighbor be similar for all rows
- Create a synthetic cell for cells that do not have a nearest neighbor conforming to angle and distances of other cells in column
- Repeat moving out left and right to cover all columns
- Create PDF of all pages to scan for errors

Use Table's Internal Structure to Build Rows and Columns

1	10012	57500
10	3300	51500
11	3.86	52075
12	6.281	52115
13	6.51	52260
14	6.73	52299
15	6.76	52326
16	6.97	52357
17	7.00	52387
18	7.25	52420
19	7.35	52453
20	7.42	52479
21	7.52	52506
22	7.55	52533
23	7.58	52559
24	7.61	52575
25	7.64	52591
26	7.67	52607
27	7.70	52623
28	7.73	52639
29	7.76	52655
30	7.79	52671
31	7.82	52687
32	7.85	52703
33	7.88	52719
34	7.91	52735
35	7.94	52751
36	7.97	52767
37	8.00	52783
38	8.03	52799
39	8.06	52815
40	8.09	52831
41	8.12	52847
42	8.15	52863
43	8.18	52879
44	8.21	52895
45	8.24	52911
46	8.27	52927
47	8.30	52943
48	8.33	52959
49	8.36	52975
50	8.39	52991
51	8.42	53007
52	8.45	53023
53	8.48	53039
54	8.51	53055
55	8.54	53071
56	8.57	53087
57	8.60	53103
58	8.63	53119
59	8.66	53135
60	8.69	53151
61	8.72	53167
62	8.75	53183
63	8.78	53199
64	8.81	53215
65	8.84	53231
66	8.87	53247
67	8.90	53263
68	8.93	53279
69	8.96	53295
70	9.00	53311
71	9.03	53327
72	9.06	53343
73	9.09	53359
74	9.12	53375
75	9.15	53391
76	9.18	53407
77	9.21	53423
78	9.24	53439
79	9.27	53455
80	9.30	53471
81	9.33	53487
82	9.36	53503
83	9.39	53519
84	9.42	53535
85	9.45	53551
86	9.48	53567
87	9.51	53583
88	9.54	53599
89	9.57	53615
90	9.60	53631
91	9.63	53647
92	9.66	53663
93	9.69	53679
94	9.72	53695
95	9.75	53711
96	9.78	53727
97	9.81	53743
98	9.84	53759
99	9.87	53775
100	9.90	53791
101	9.93	53807
102	9.96	53823
103	9.99	53839
104	10.02	53855
105	10.05	53871
106	10.08	53887
107	10.11	53903
108	10.14	53919
109	10.17	53935
110	10.20	53951
111	10.23	53967
112	10.26	53983
113	10.29	53999
114	10.32	54015
115	10.35	54031
116	10.38	54047
117	10.41	54063
118	10.44	54079
119	10.47	54095
120	10.50	54111
121	10.53	54127
122	10.56	54143
123	10.59	54159
124	10.62	54175
125	10.65	54191
126	10.68	54207
127	10.71	54223
128	10.74	54239
129	10.77	54255
130	10.80	54271
131	10.83	54287
132	10.86	54303
133	10.89	54319
134	10.92	54335
135	10.95	54351
136	10.98	54367
137	11.01	54383
138	11.04	54399
139	11.07	54415
140	11.10	54431
141	11.13	54447
142	11.16	54463
143	11.19	54479
144	11.22	54495
145	11.25	54511
146	11.28	54527
147	11.31	54543
148	11.34	54559
149	11.37	54575
150	11.40	54591
151	11.43	54607
152	11.46	54623
153	11.49	54639
154	11.52	54655
155	11.55	54671
156	11.58	54687
157	11.61	54703
158	11.64	54719
159	11.67	54735
160	11.70	54751
161	11.73	54767
162	11.76	54783
163	11.79	54799
164	11.82	54815
165	11.85	54831
166	11.88	54847
167	11.91	54863
168	11.94	54879
169	11.97	54895
170	12.00	54911
171	12.03	54927
172	12.06	54943
173	12.09	54959
174	12.12	54975
175	12.15	54991
176	12.18	55007
177	12.21	55023
178	12.24	55039
179	12.27	55055
180	12.30	55071
181	12.33	55087
182	12.36	55103
183	12.39	55119
184	12.42	55135
185	12.45	55151
186	12.48	55167
187	12.51	55183
188	12.54	55199
189	12.57	55215
190	12.60	55231
191	12.63	55247
192	12.66	55263
193	12.69	55279
194	12.72	55295
195	12.75	55311
196	12.78	55327
197	12.81	55343
198	12.84	55359
199	12.87	55375
200	12.90	55391
201	12.93	55407
202	12.96	55423
203	12.99	55439
204	13.02	55455
205	13.05	55471
206	13.08	55487
207	13.11	55503
208	13.14	55519
209	13.17	55535
210	13.20	55551
211	13.23	55567
212	13.26	55583
213	13.29	55599
214	13.32	55615
215	13.35	55631
216	13.38	55647
217	13.41	55663
218	13.44	55679
219	13.47	55695
220	13.50	55711
221	13.53	55727
222	13.56	55743
223	13.59	55759
224	13.62	55775
225	13.65	55791
226	13.68	55807
227	13.71	55823
228	13.74	55839
229	13.77	55855
230	13.80	55871
231	13.83	55887
232	13.86	55903
233	13.89	55919
234	13.92	55935
235	13.95	55951
236	13.98	55967
237	14.01	55983
238	14.04	55999
239	14.07	56015
240	14.10	56031
241	14.13	56047
242	14.16	56063
243	14.19	56079
244	14.22	56095
245	14.25	56111
246	14.28	56127
247	14.31	56143
248	14.34	56159
249	14.37	56175
250	14.40	56191
251	14.43	56207
252	14.46	56223
253	14.49	56239
254	14.52	56255
255	14.55	56271
256	14.58	56287
257	14.61	56303
258	14.64	56319
259	14.67	56335
260	14.70	56351
261	14.73	56367
262	14.76	56383
263	14.79	56399
264	14.82	56415
265	14.85	56431
266	14.88	56447
267	14.91	56463
268	14.94	56479
269	14.97	56495
270	15.00	56511
271	15.03	56527
272	15.06	56543
273	15.09	56559
274	15.12	56575
275	15.15	56591
276	15.18	56607
277	15.21	56623
278	15.24	56639
279	15.27	56655
280	15.30	56671
281	15.33	56687
282	15.36	56703
283	15.39	56719
284	15.42	5

Use Table's Internal Structure to Build Rows and Columns



Use Table's Internal Structure to Build Rows and Columns

Train and Apply Custom Model

- Match cells to training data – Washington DC, Mapping Segregation
- Train random forest model at the character level
 - Pixel value by position in bounding box
 - Tesseract predicted text
 - Tesseract confidence
- Grid search with cross validation to tune hyperparameters
- Apply model (out of sample) to remaining cities

Identify Internal Inconsistencies, Compare to Tract Totals

- Internal consistency, e.g. Owner Occupied + Renter Occupied = Occupied
- Check for outliers at column level
- Compare stats to tract totals, accounting for suppression
- Make corrections easy with Excel tool

tract	ocr	human
TOTAL	TOTAL	
1 - A	1-A	
1 - B	1-B	
1 - C	1-C	
2 - A	2-A	
2 - B	2-B	
2 - C	2-C	
3 - A	3-A	
3 - B		

Caveats

- Approach requires some customization per dataset
- Manual steps remain (and probably always will)
 - Identifying unusable scans
 - Identification of page ranges in source documents (missing pages)
 - Always be checking
 - Tract transcription is still manual

Current State

- Scaling work to all 16 cities for 1950
- Refining issues with 1960 model
- Starting 1940 work
- Textract for assist with tract identifiers?
- Claude or other LLM based service for first cut?

Summary

Summary

- We are working on **digitizing** the historical Censuses of Housing **Block Statistics**, 1940 to 1970.
- Our goal: Develop & release data for 16 cities, training & validation data, and methods & code.
- The three major tasks are digitizing block **shapes**, the block **situations**, and the block **statistics**.
- This is a work in progress; Questions and comments welcome!

Thanks!