FEDERAL RESERVE BANK
PHILADELPHIA

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

PhiladelphiaFed.org | @PhiladelphiaFed

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.

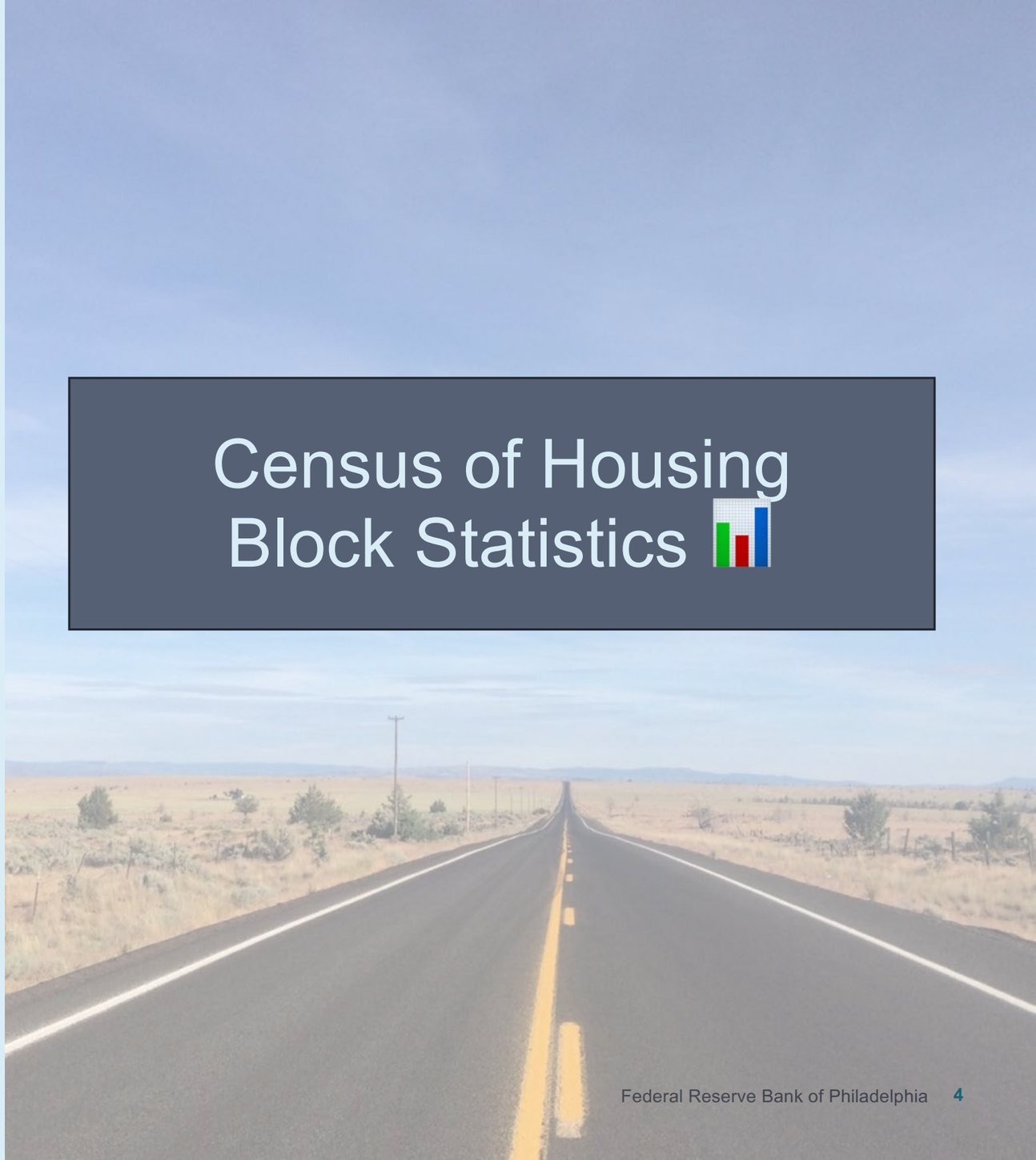
FEDERAL RESERVE BANK
PHILADELPHIA

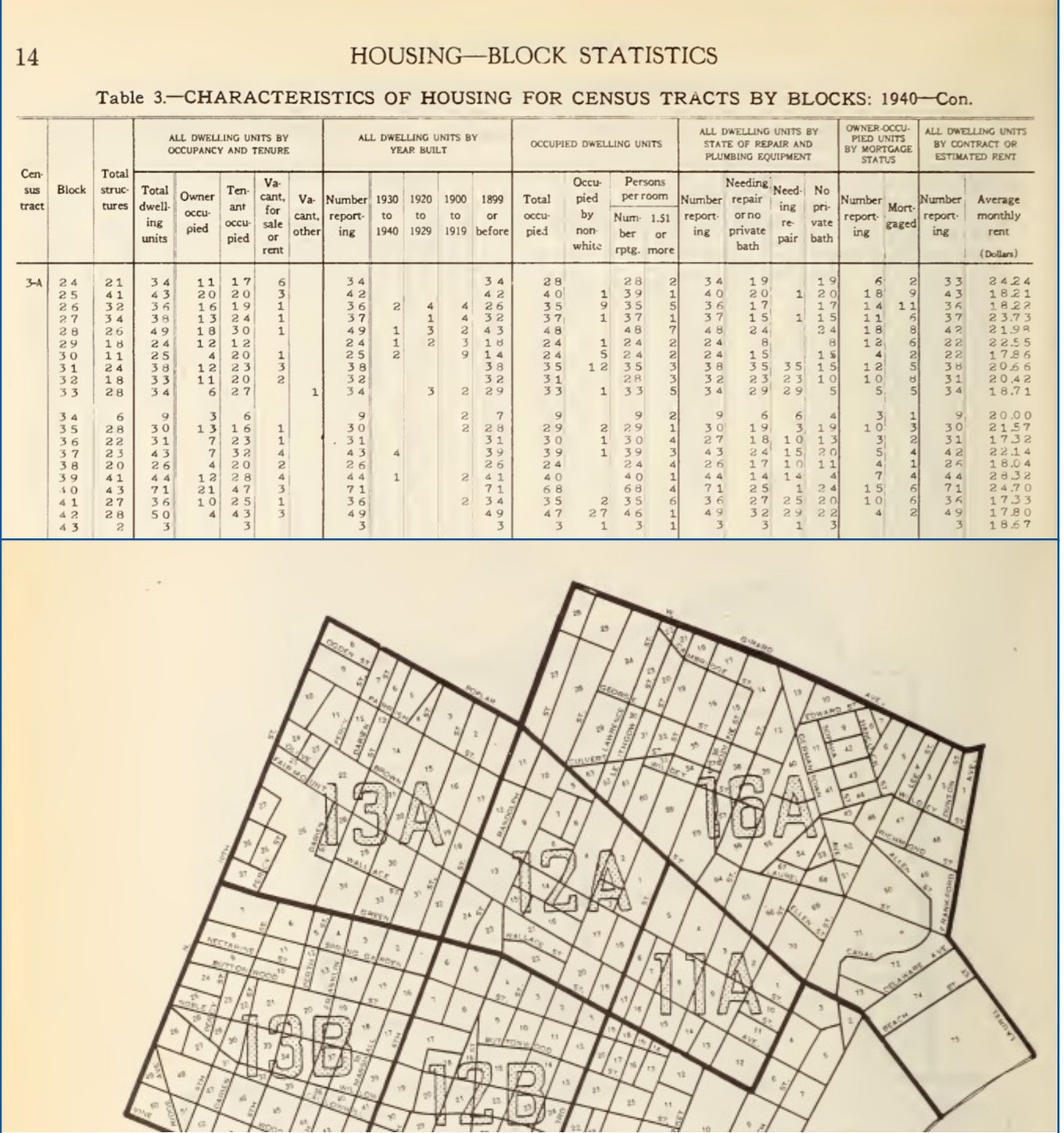
Digitizing Block Statistics

- What
- Why
- Goals
- Tasks and Challenges

Digitizing Block Statistics

- **What**
- Why
- Goals
- Tasks and Challenges





14 HOUSING—BLOCK STATISTICS

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

Census tract	Block	Total structures	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 occupied	Tenant occupied	Vacant, for sale or rent	Vacant, other	Number reporting	1930 to 1940	1920 to 1929	1900 to 1919	1899 or before	Total occupied	Occupied by nonwhite	Persons per room		Number reporting	Needing repair or no private bath	Needing repair	No private bath	Number reporting	Mortgaged	Number reporting	Average monthly rent (Dollars)
															Number rptg.	1.51 or more								
3-A	24	21	34	11	17	6		34				34	28		28	2	34	19		19	6	2	33	24.24
	25	41	43	20	20	3		42				42	40	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	18	8	42	21.98
	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				38	35	12	35	3	38	35	35	15	12	5	38	20.66
	32	18	33	11	20	2		32				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	5	34	18.71
	34	6	9	3	6			9			2	7	9		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	19	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				26	24		24	4	26	17	10	11	4	1	26	18.04
	39	41	44	12	28	4		44	1		2	41	40		40	1	44	14	14	4	7	4	44	28.32
	40	43	71	21	47	3		71				71	68		68	4	71	25	1	24	15	6	71	24.70
	41	27	36	10	25	1		36			2	34	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	18.67

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?

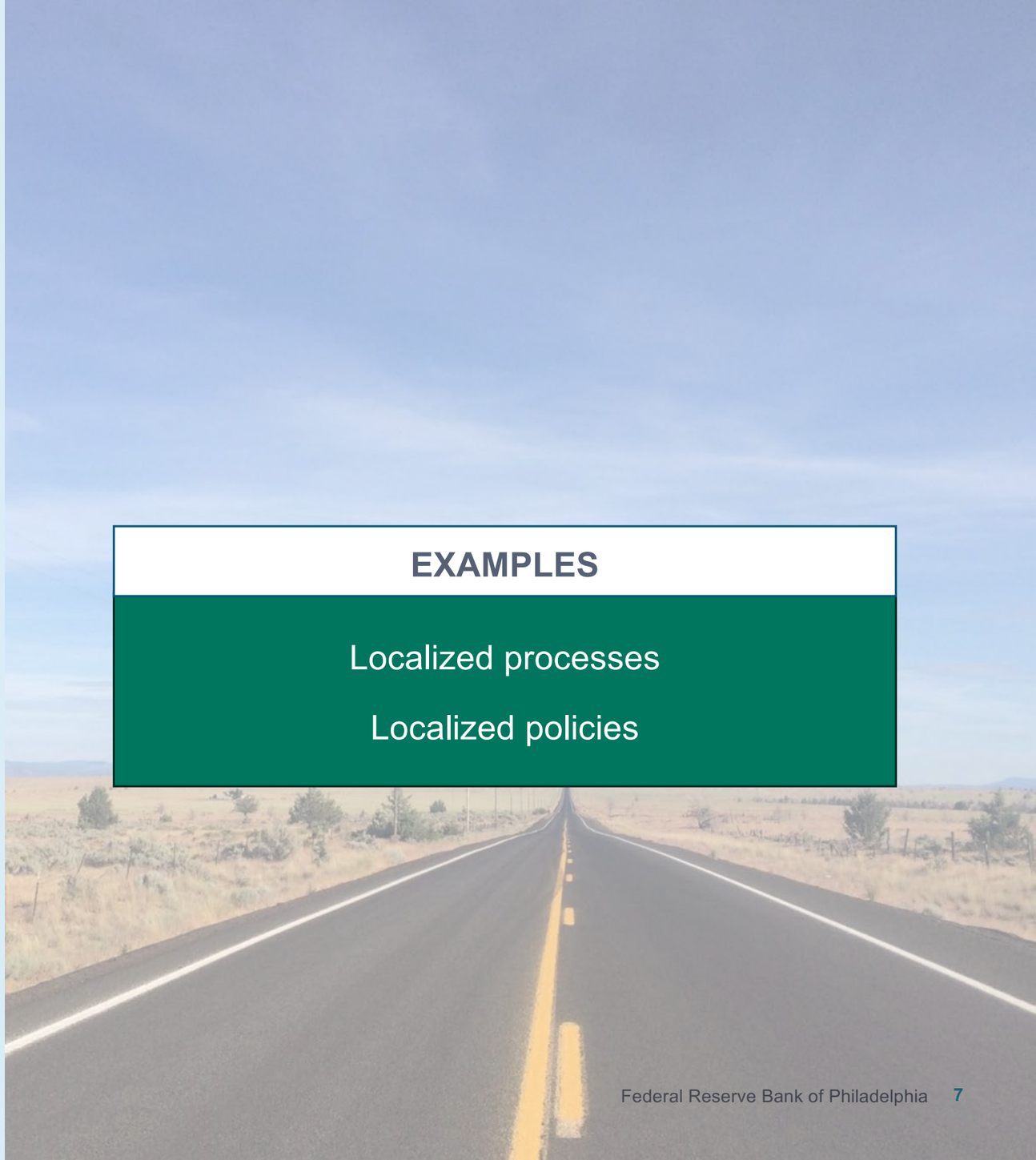
Studies of housing investment and maintenance and long-run urban dynamics.

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

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

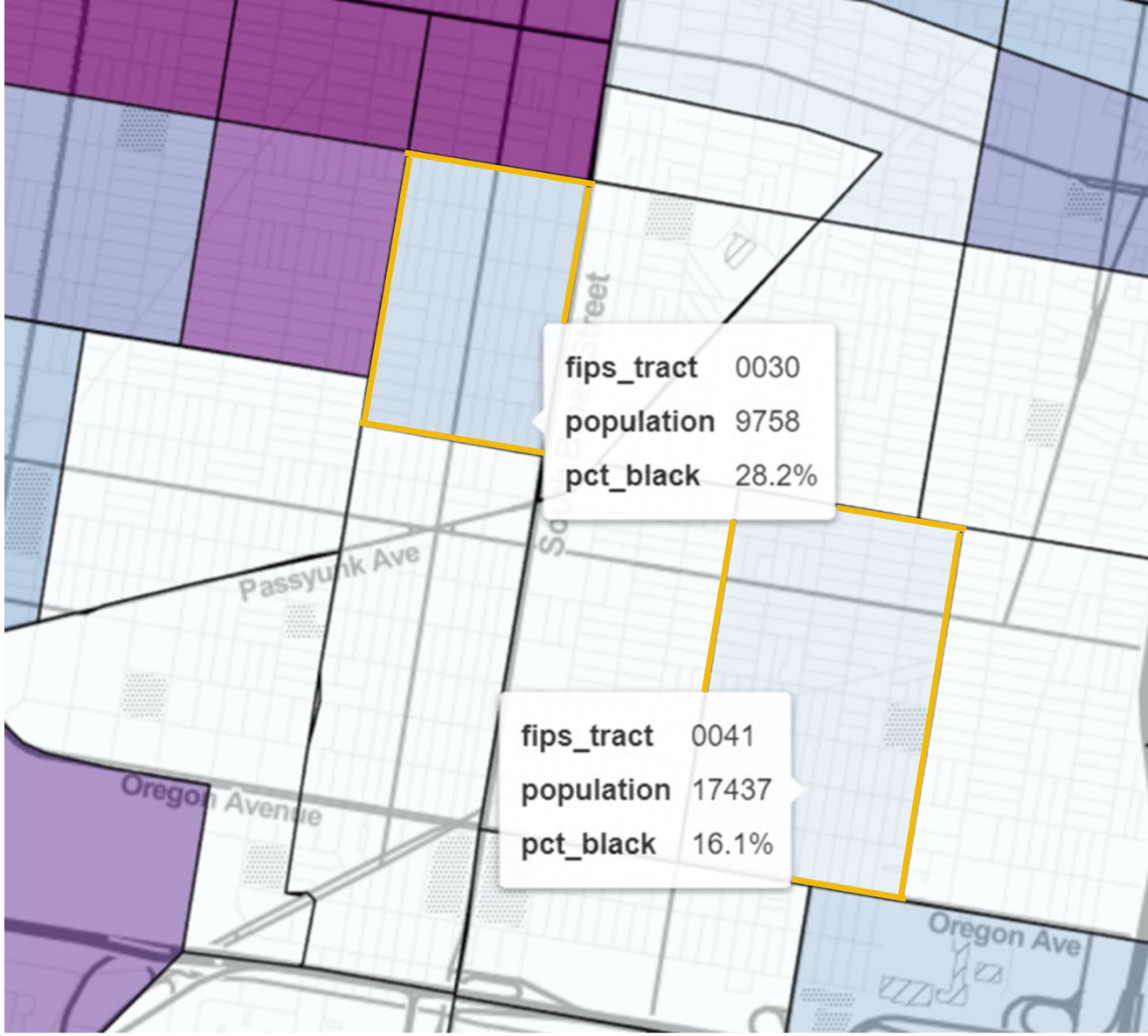
Digitizing Block Statistics

- What
- **Why**
- Goals
- Tasks and Challenges

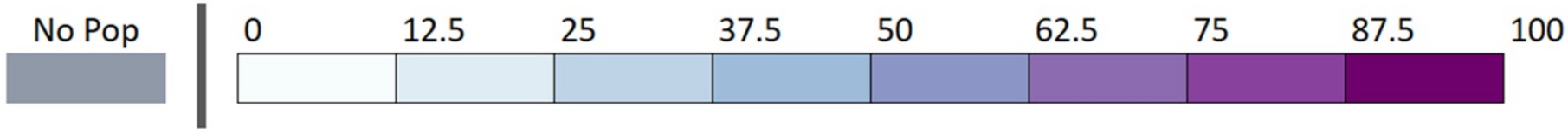


Localized Processes

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

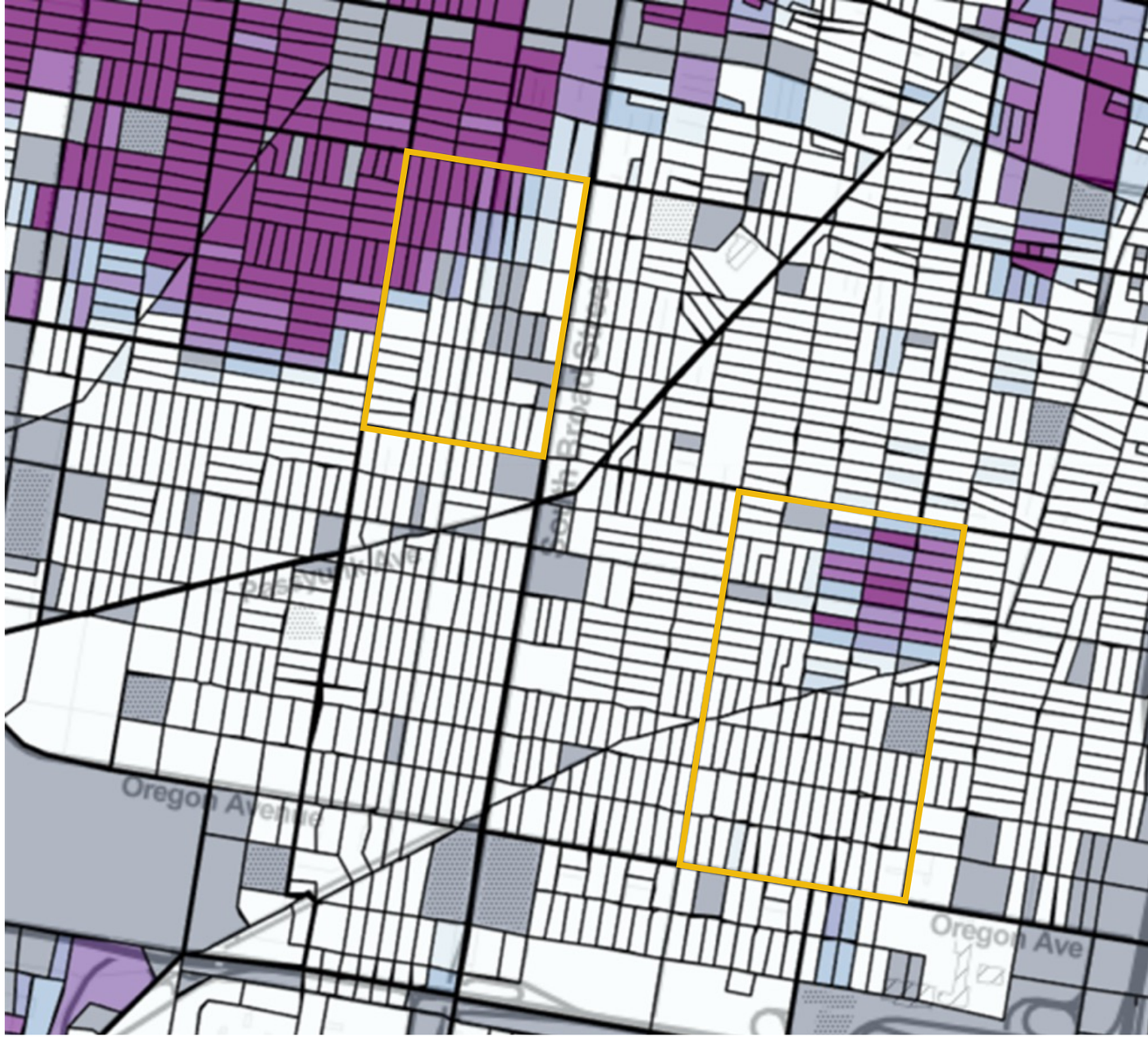


Percent of population Black



Localized Processes

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



Percent of population Black

No Pop | 0 | 12.5 | 25 | 37.5 | 50 | 62.5 | 75 | 87.5 | 100

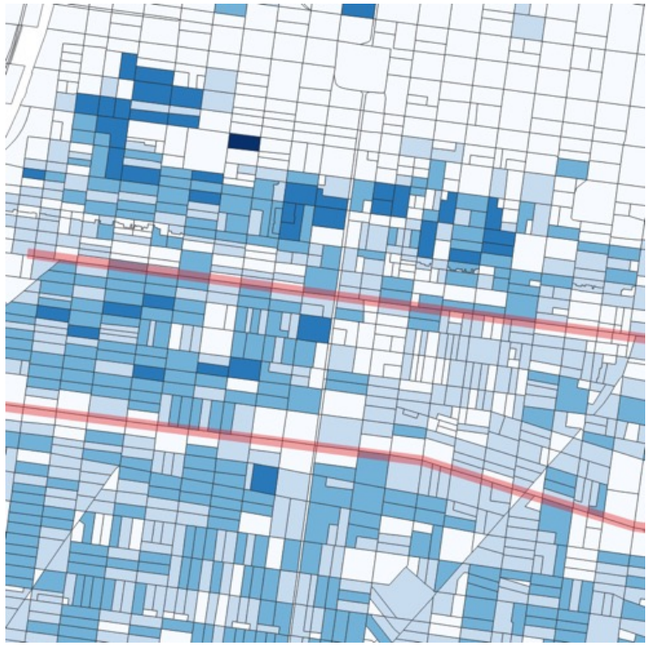
Localized Policies

Runner-up design

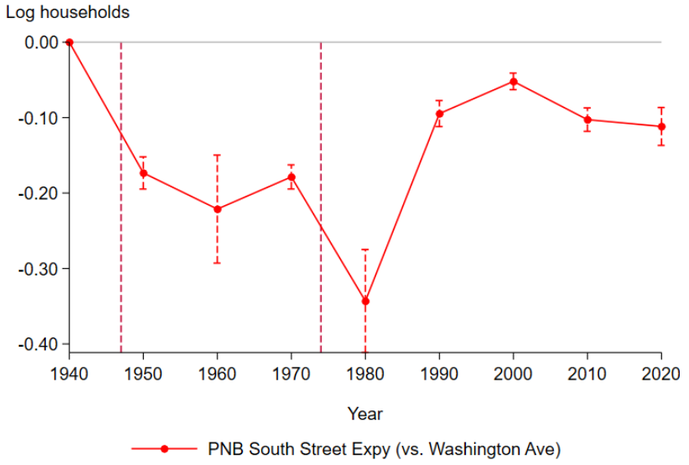
- “Expecting an Expressway” (Brinkman, Lin & Mangum).
- Two proposed routes for the Crosstown Expressway in South Philadelphia.

Tracts

Blocks



Difference in differences



Digitizing Block Statistics

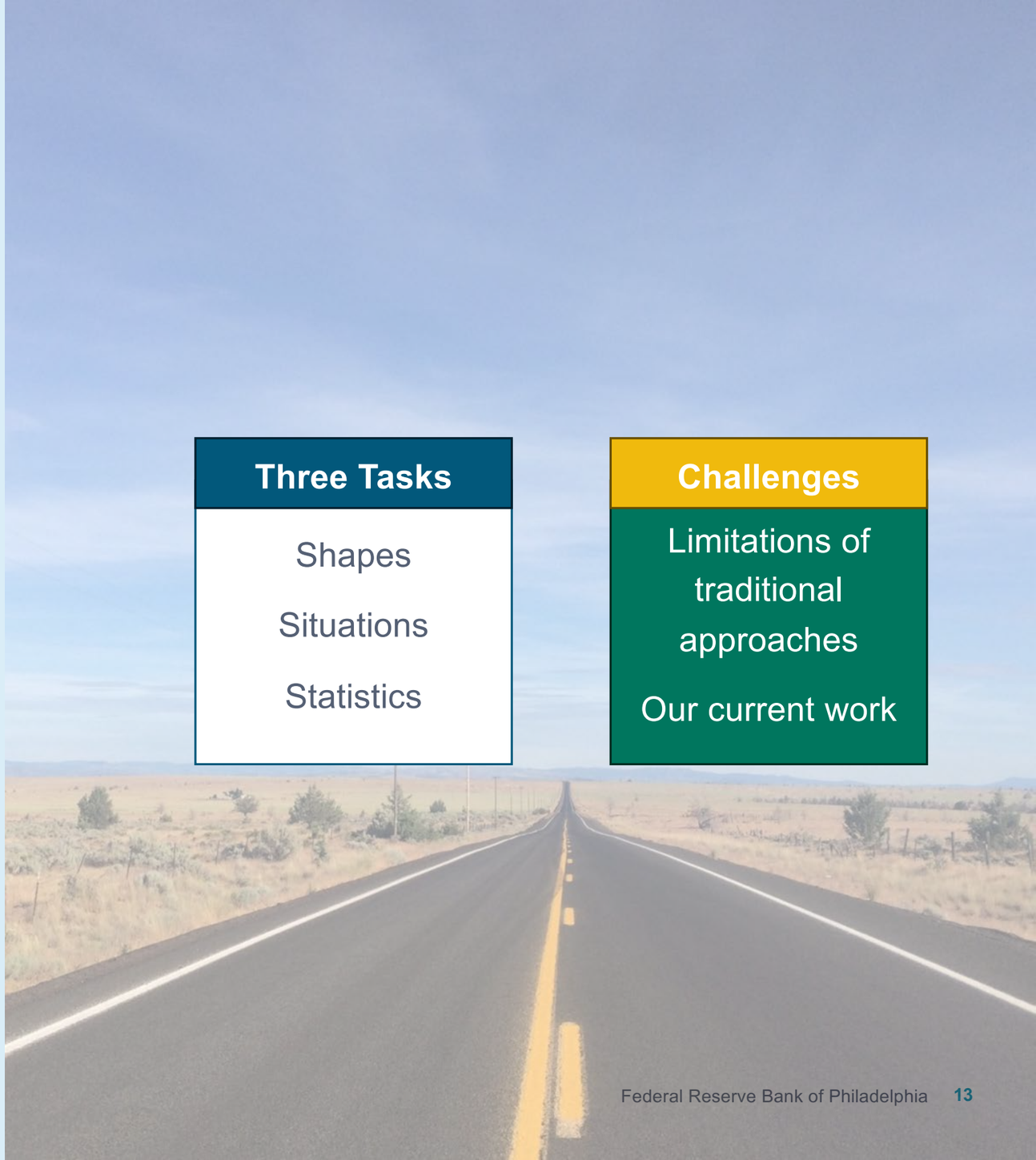
- What
- Why
- **Goals**
- Tasks and Challenges

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

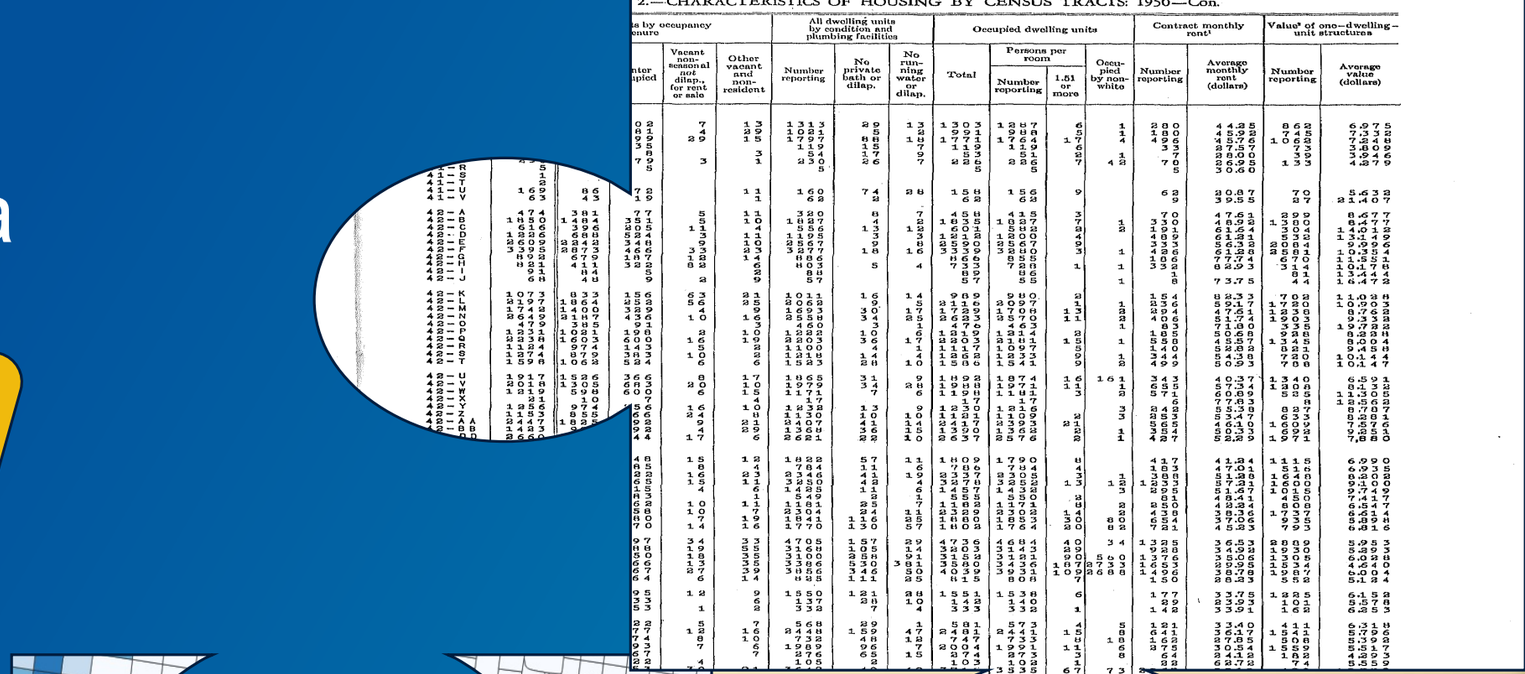
- What
- Why
- Goals
- **Tasks and Challenges**

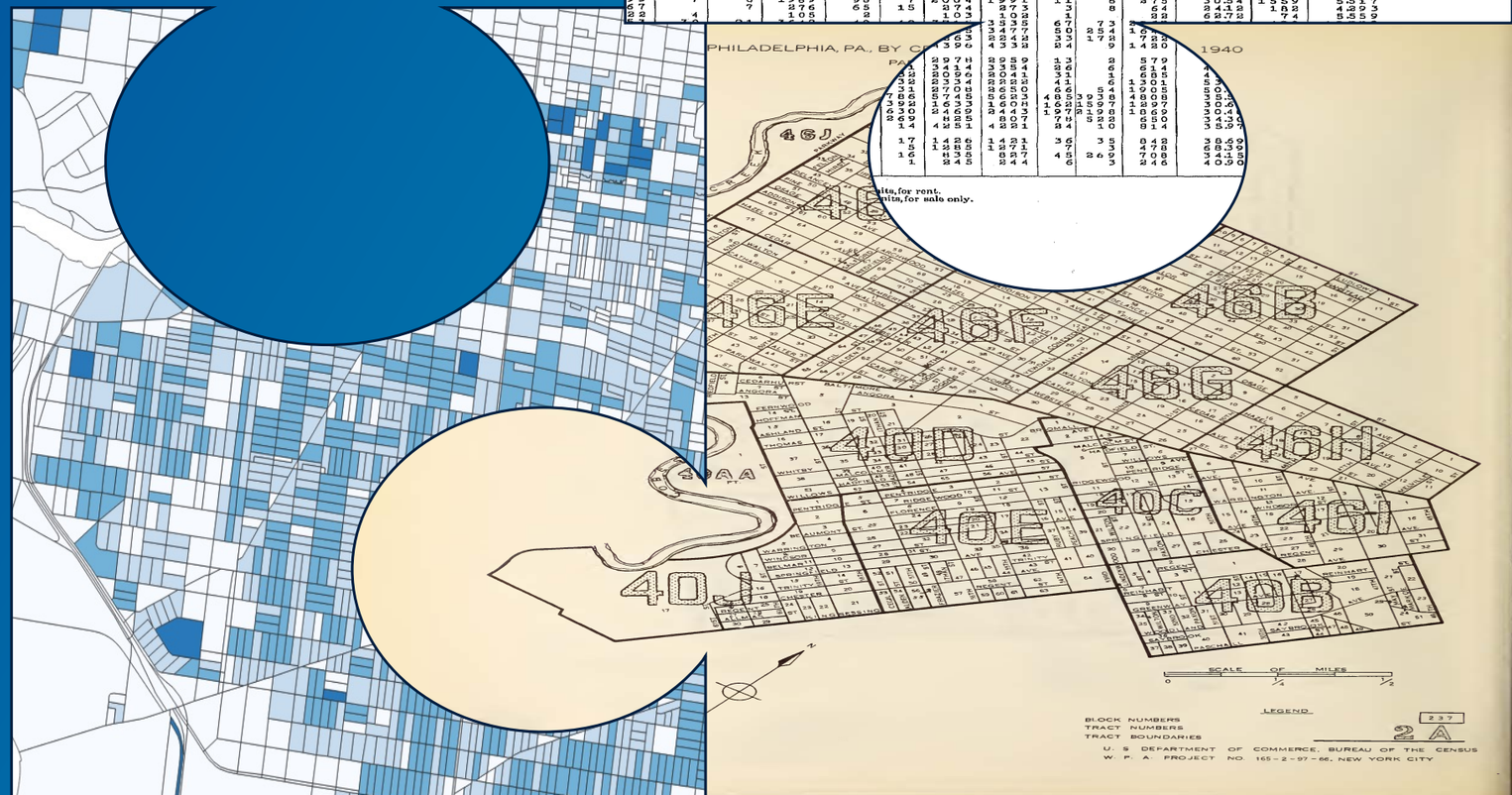


3 Tasks, 3 Pieces of Data

1. Shapes
2. Situations
3. Statistics

Blocks need to know their:





3 Tasks, 3 Pieces of Data

1 **Shape**

2 Situation

3 Statistics

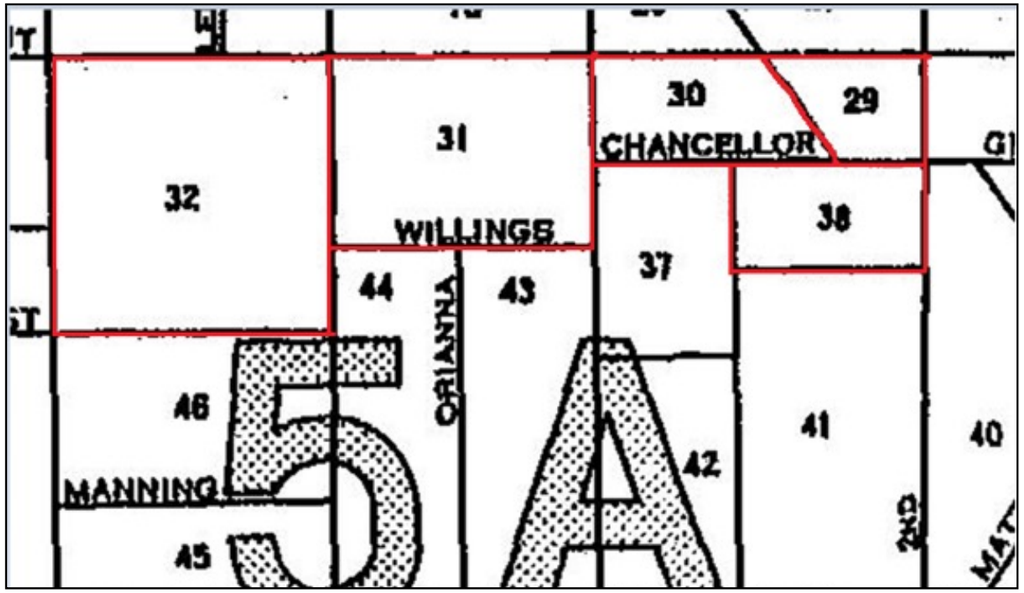
Segmenting Block Shapes from Maps

Our Ideal Process Has Only Three Steps

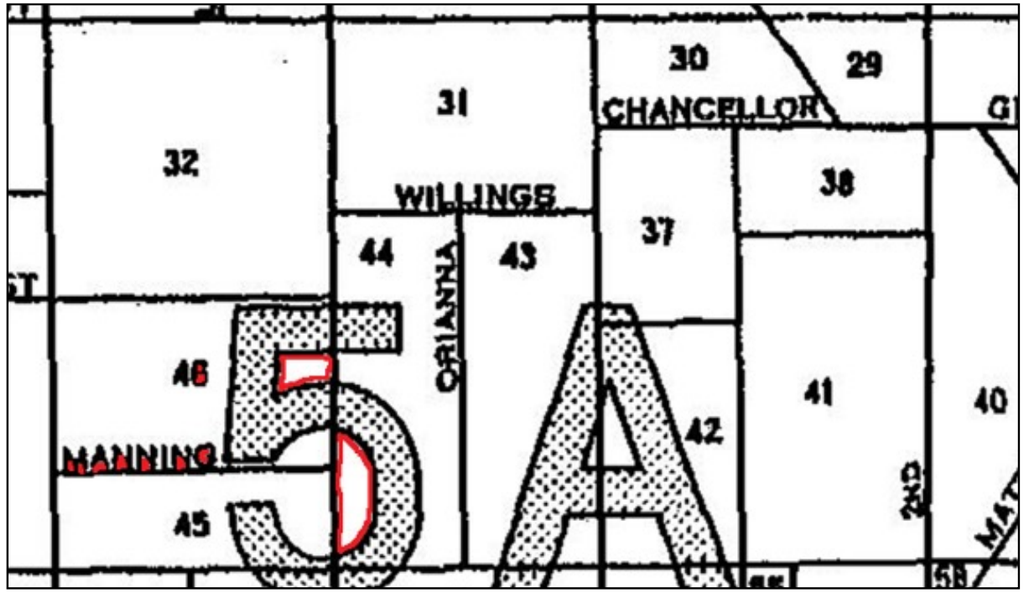
1. Identify closed loops of black ink.
2. Call them all blocks.
3. Declare victory.

Unfortunately, This Process Fails Spectacularly

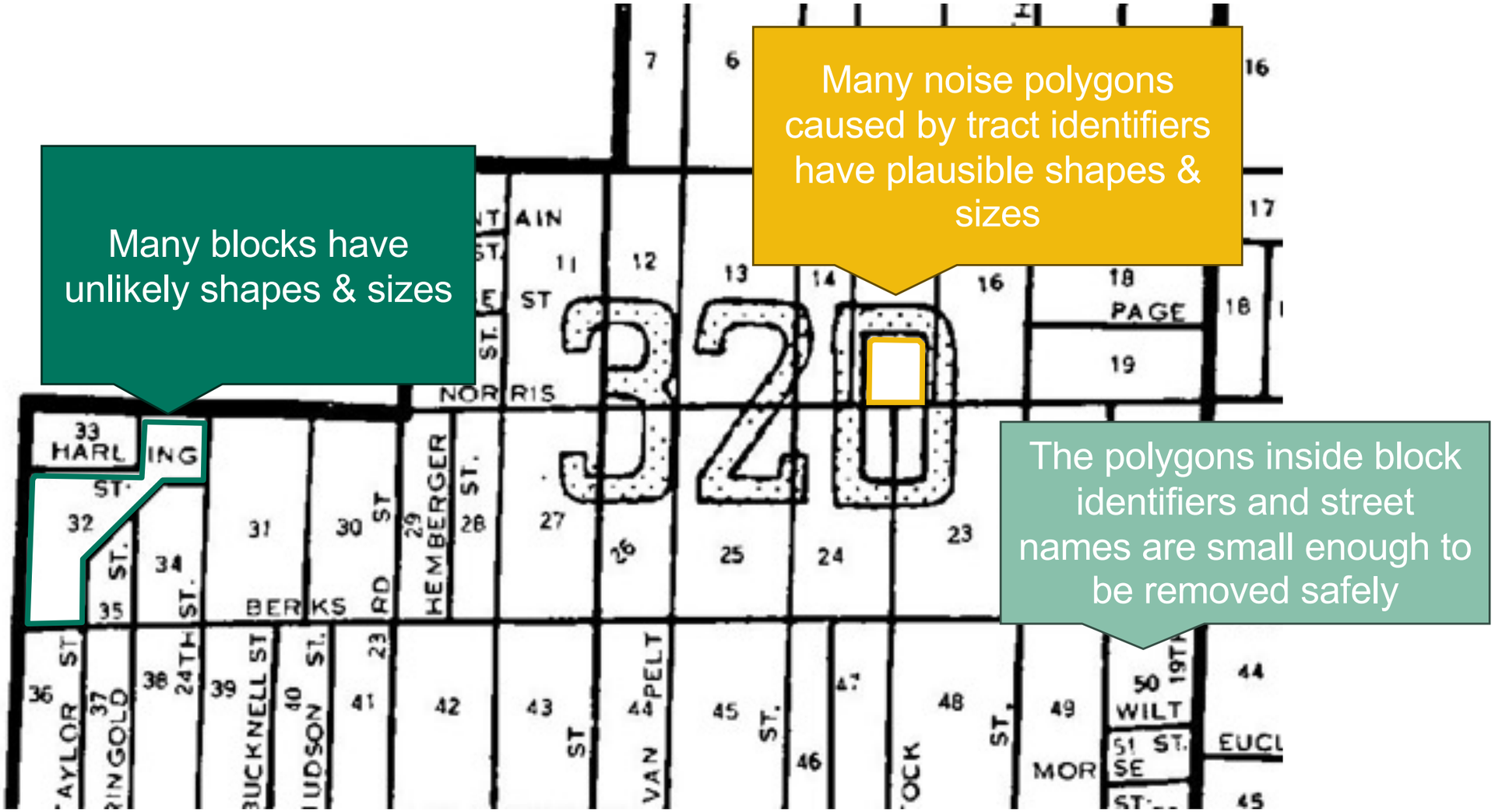
While many closed loops of black ink are blocks….



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?

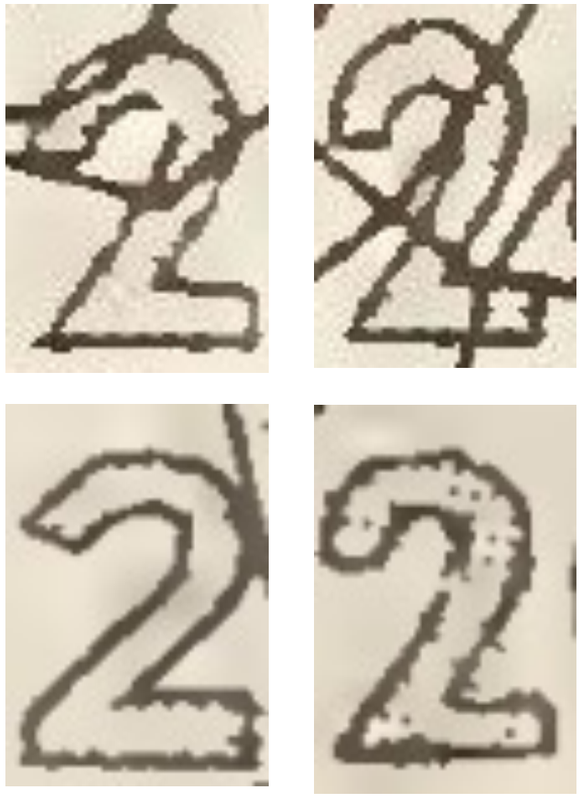
Traditional Method 1: Matching Large Shapes

Issues

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

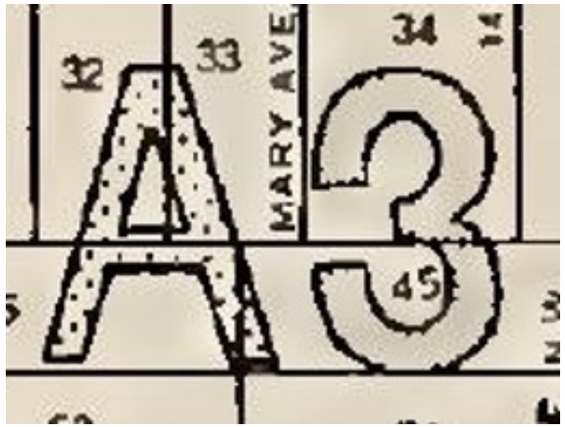
Low-confidence matches

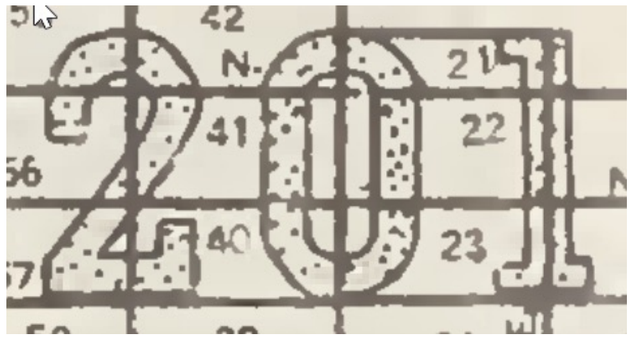
- If we accept low confidence matches of enough templates, everything starts to look like a tract identifier.
- Especially problematic with blocky characters like 1 and E.



Traditional Method 2: Matching Patterns

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





Current Work: CNN

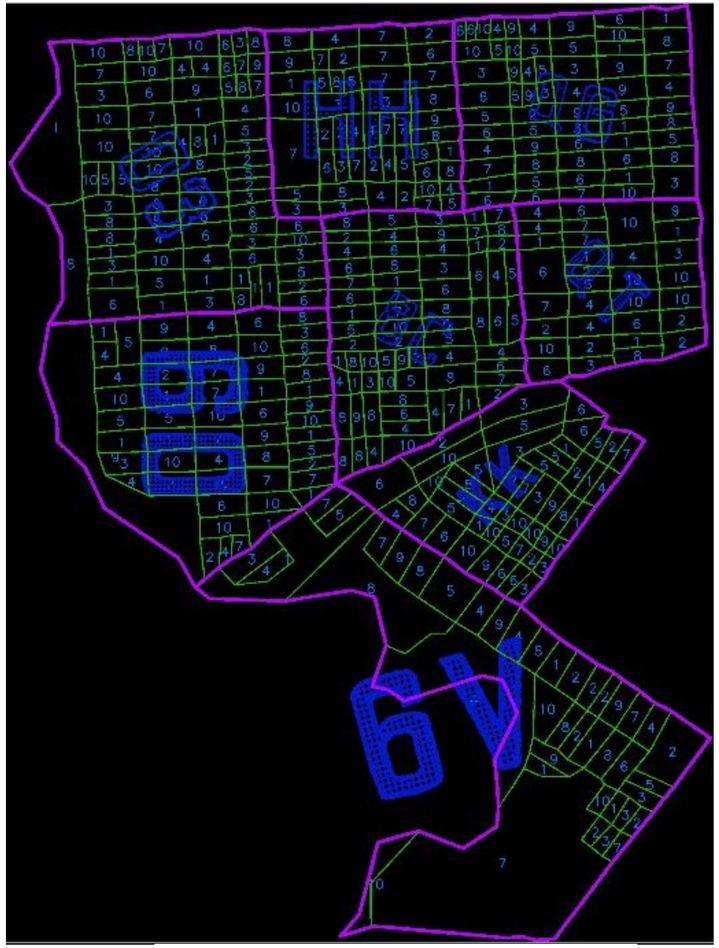
More holistic

- 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.

Simulated Map

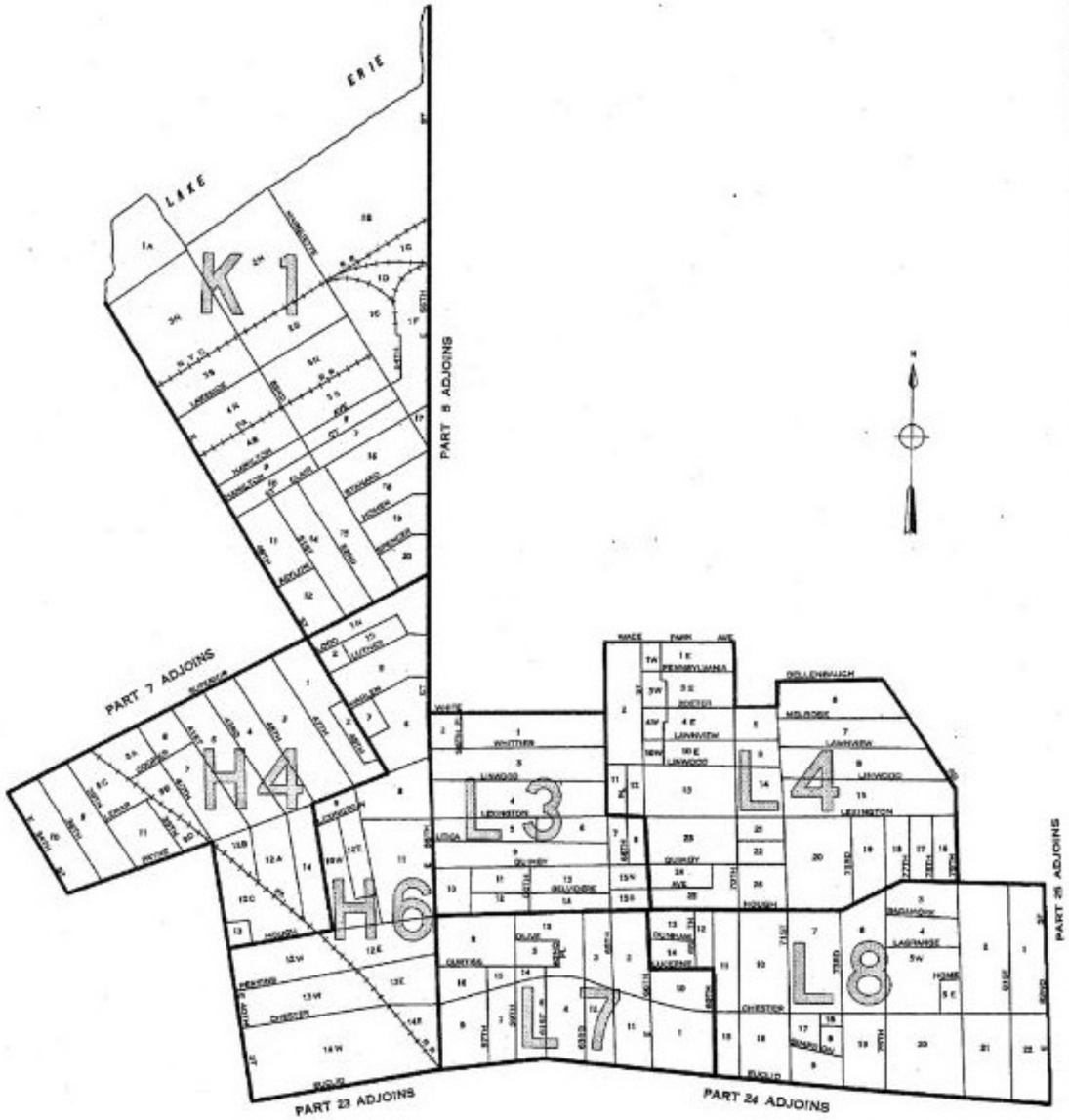


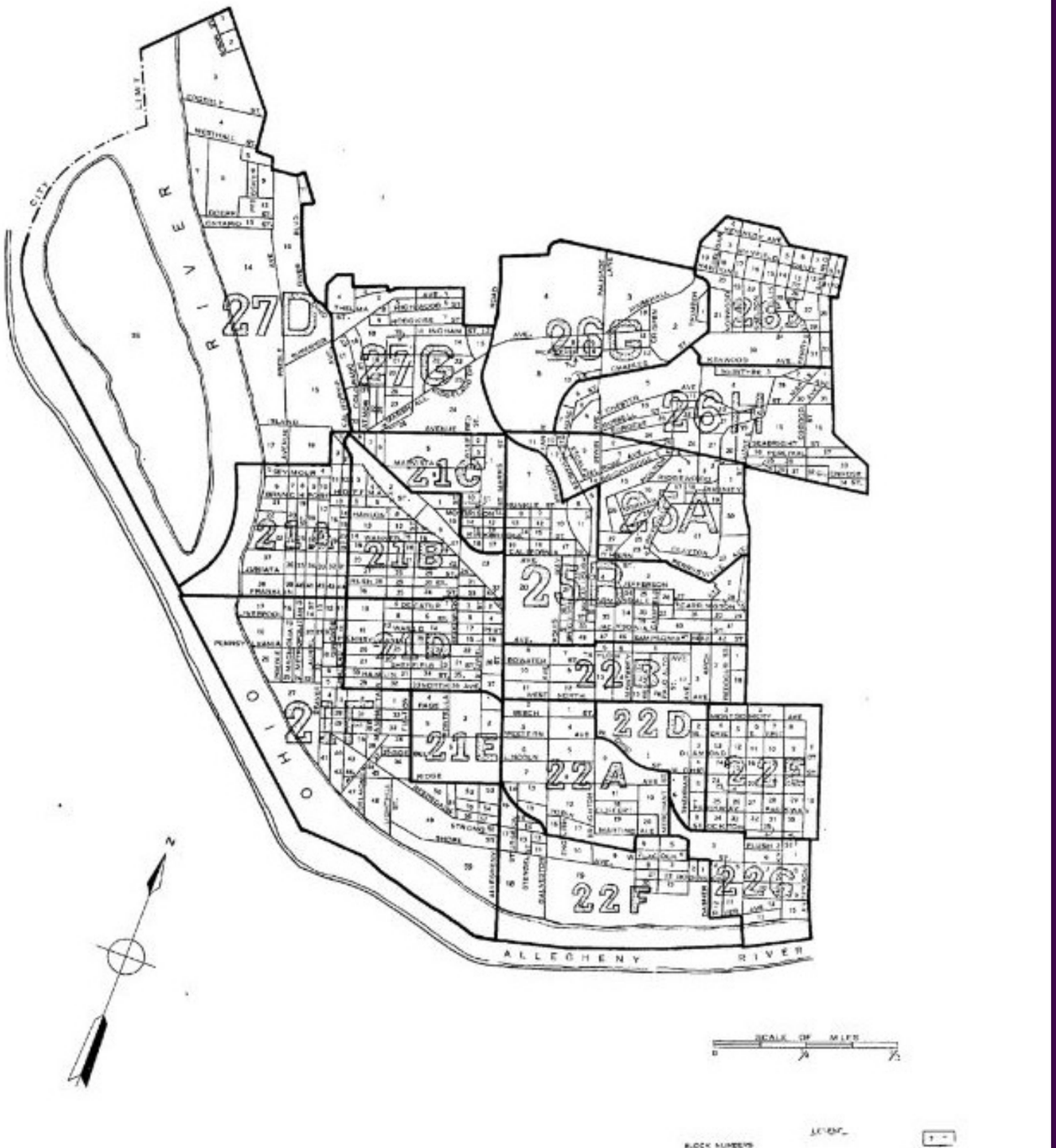
Training Mask

Creating Training Data

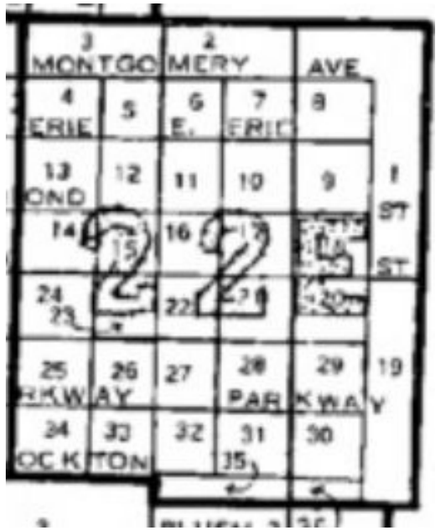
- Hand annotations are expensive; Simulating maps is cheap.
- Sample 1990 Census block and tract boundaries from NHGIS.
- Sample tract and block identifiers from real 1940 maps.
- Randomly assign speckle density to tract identifiers.

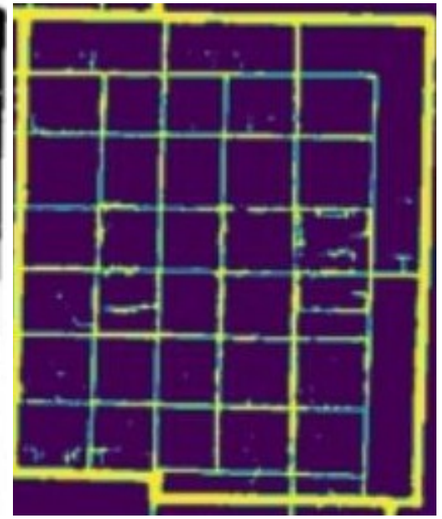
Can the model trained on simulated maps generalize to real ones?


LAKE ERIE
K1
H4
H6
L3
L4
L7
L8
PART 5 ADJOINS
PART 7 ADJOINS
PART 23 ADJOINS
PART 24 ADJOINS
PART 25 ADJOINS
SCALE IN FEET
LEGEND
BLOCK NUMBERS
TRACT NUMBERS
TRACT BOUNDARIES
U.S. DEPARTMENT OF COMMERCE, BUREAU OF THE CENSUS


RIVER
OHIO
ALLEGHENY RIVER
27D
27B
26G
26F
21C
23A
21B
21A
21D
21E
22B
22D
22A
22C
22F
22E
N
SCALE OF MILES

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

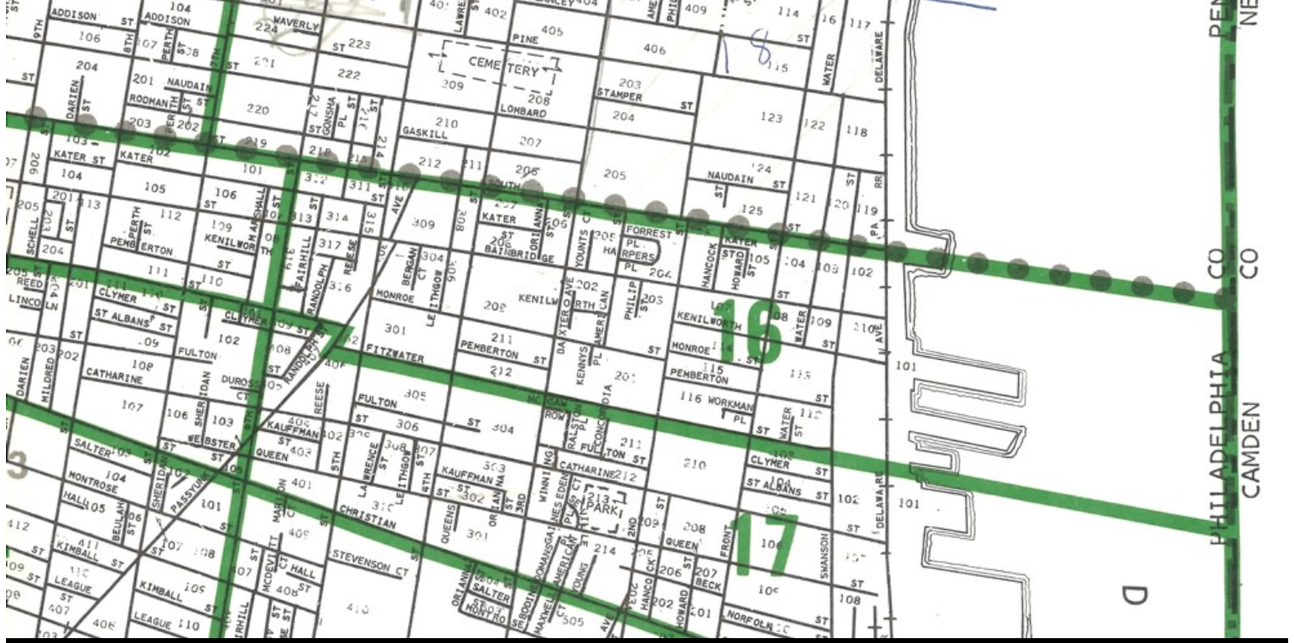
- Tract boundary segmentation is somewhat easier.

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.

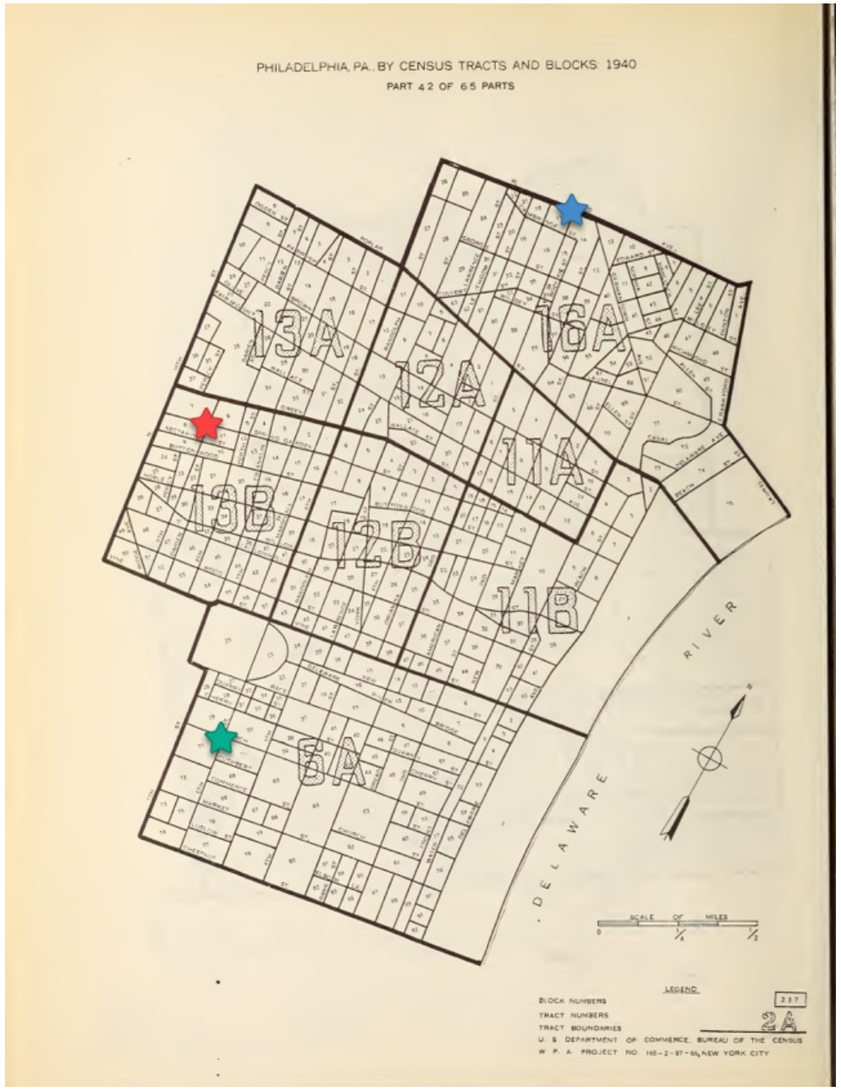


3 Tasks, 3 Pieces of Data

1. Shape
2. **Situation**
3. Statistics

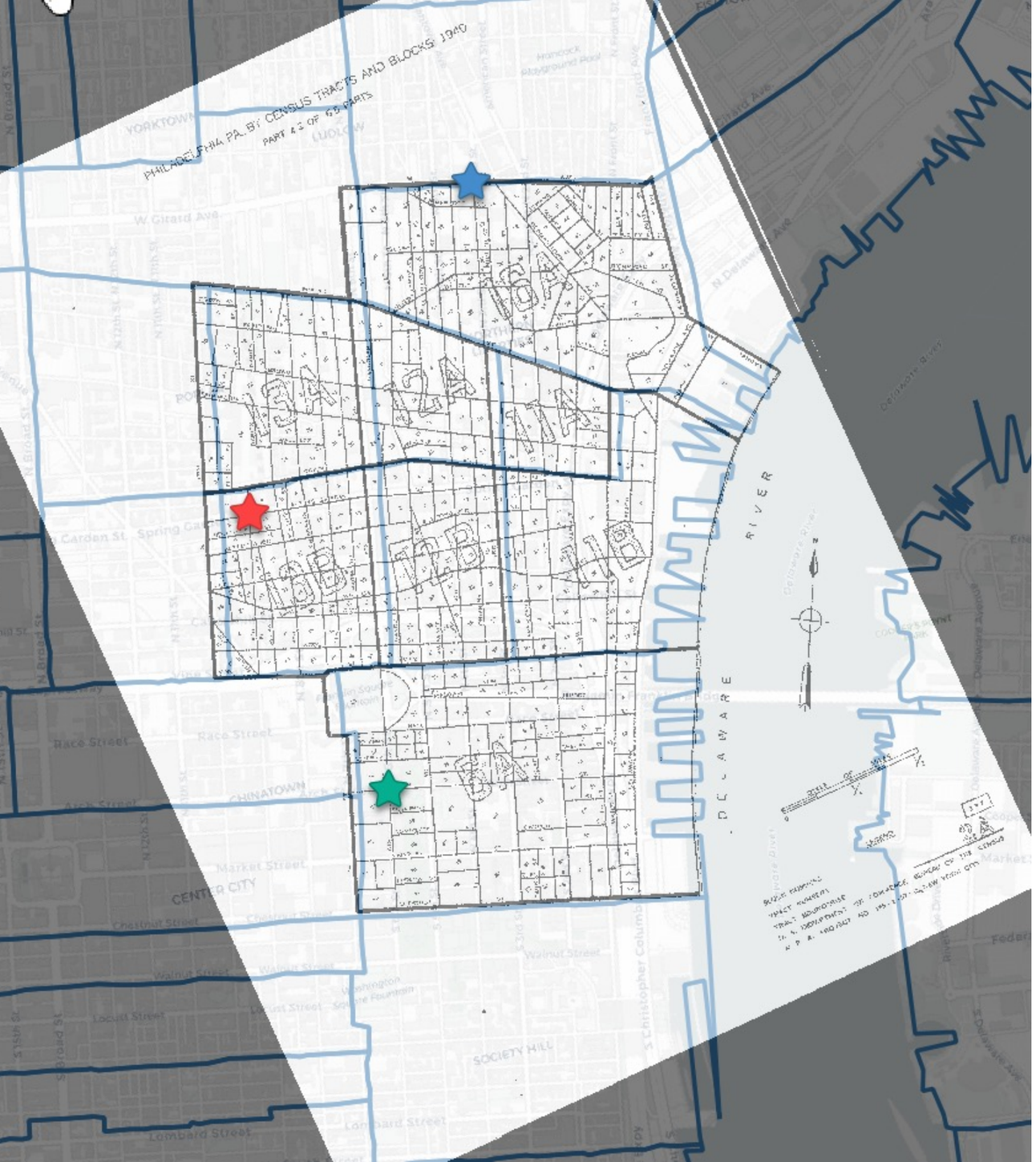
Geo-Referencing Maps

Just Keep Clicking 🐟



+

= . . .



A Georeferenced Map 🎯

We know where this picture goes now!

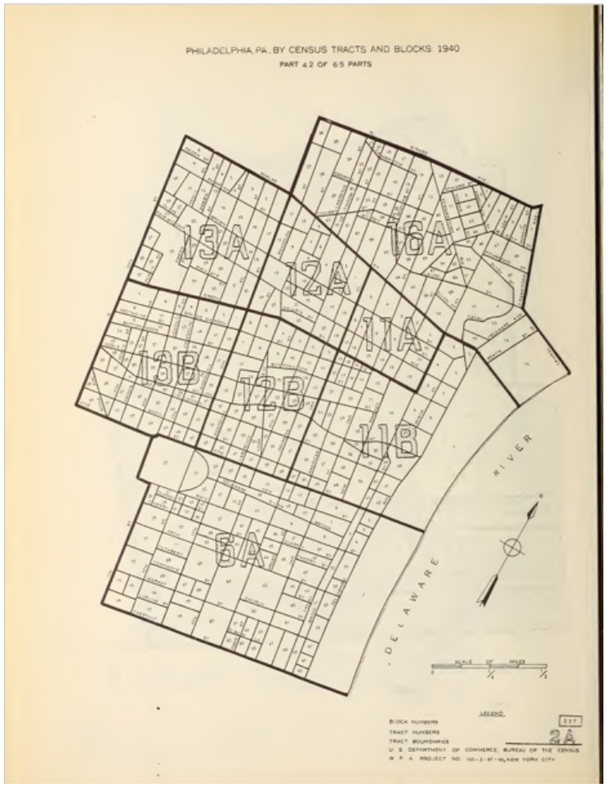
But…

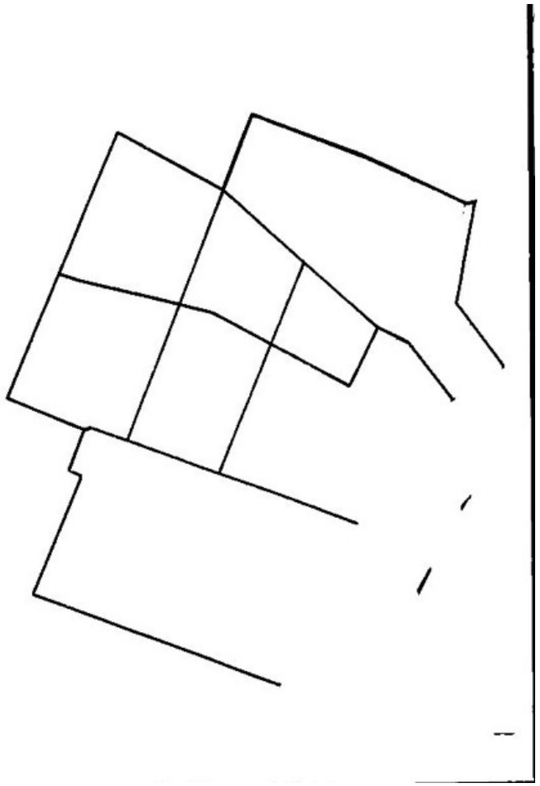
- Slooooooooow.
- Only 3 points doesn’t handle map inaccuracies well.
- Doesn’t match (blue) reference NHGIS shapefile.

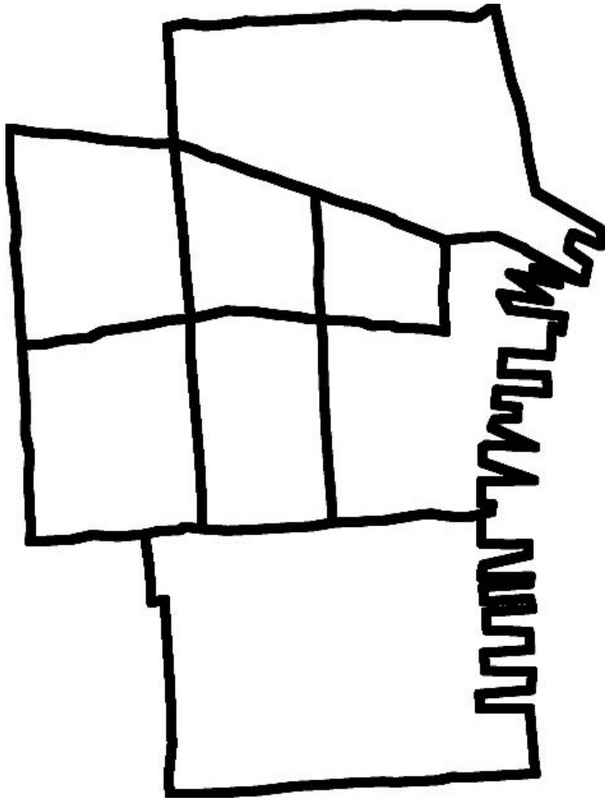
How Can We Get Better?🏋️‍♀️

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

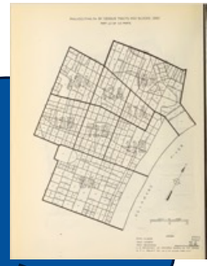
Simplifying the Problem



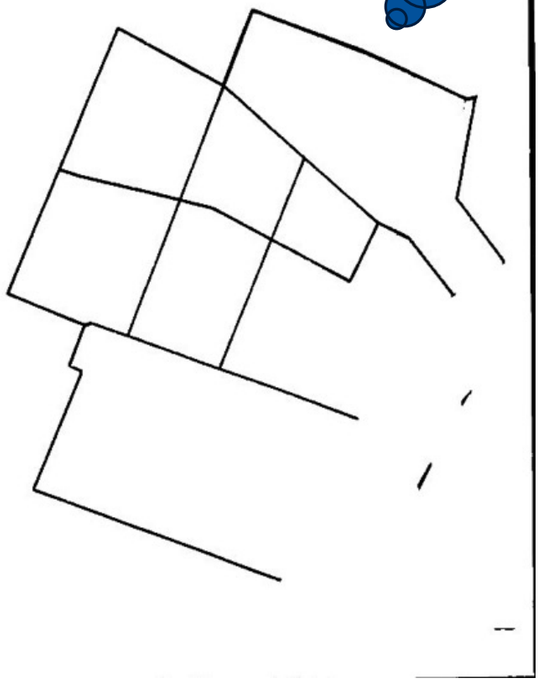




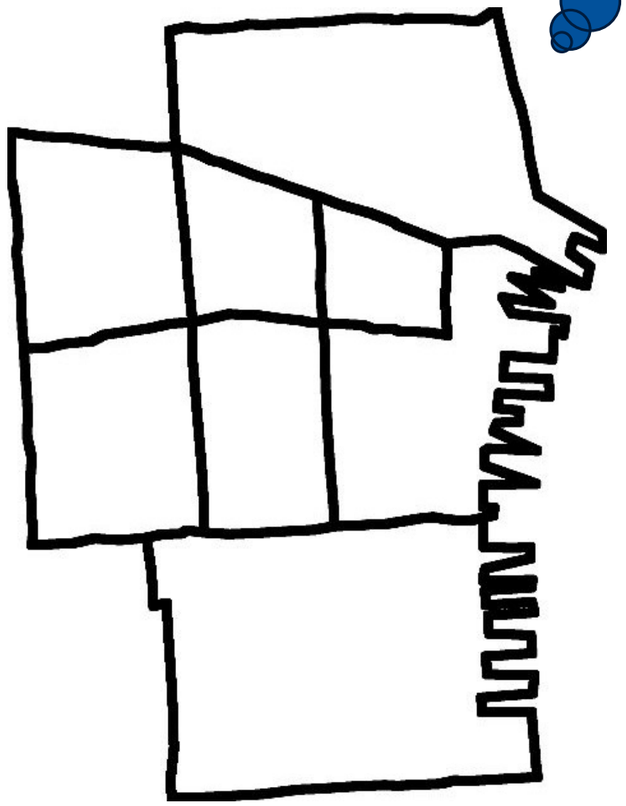
How Hard Could This Be?



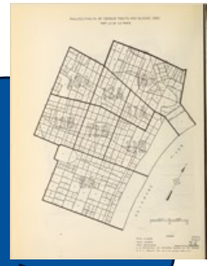
I still remember the block boundaries!



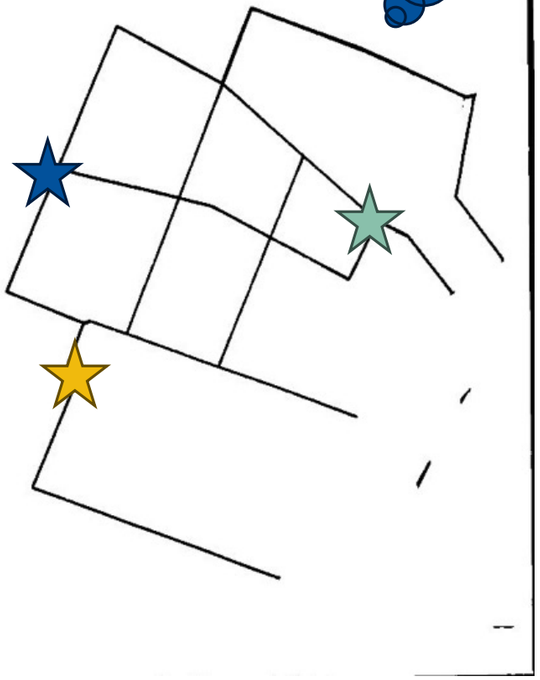
I still remember where I am in the world!



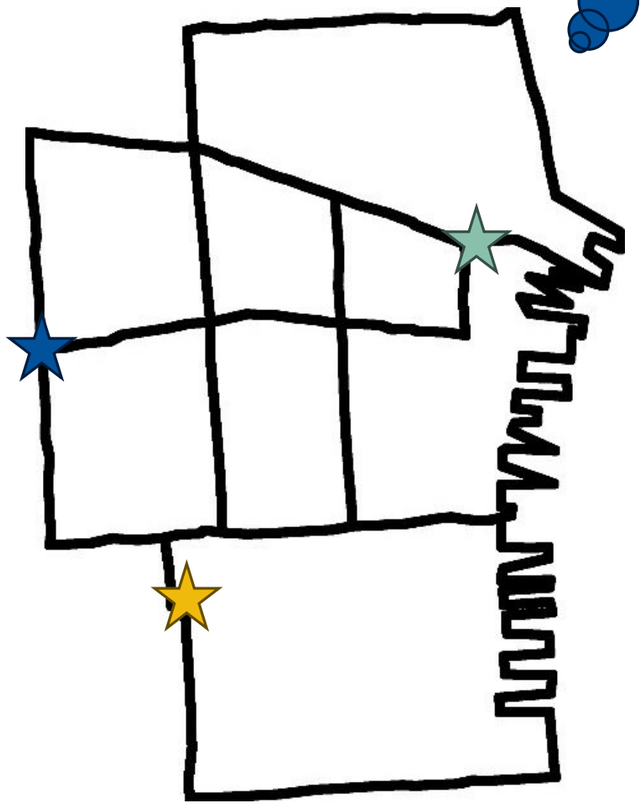
How Hard Could This Be?

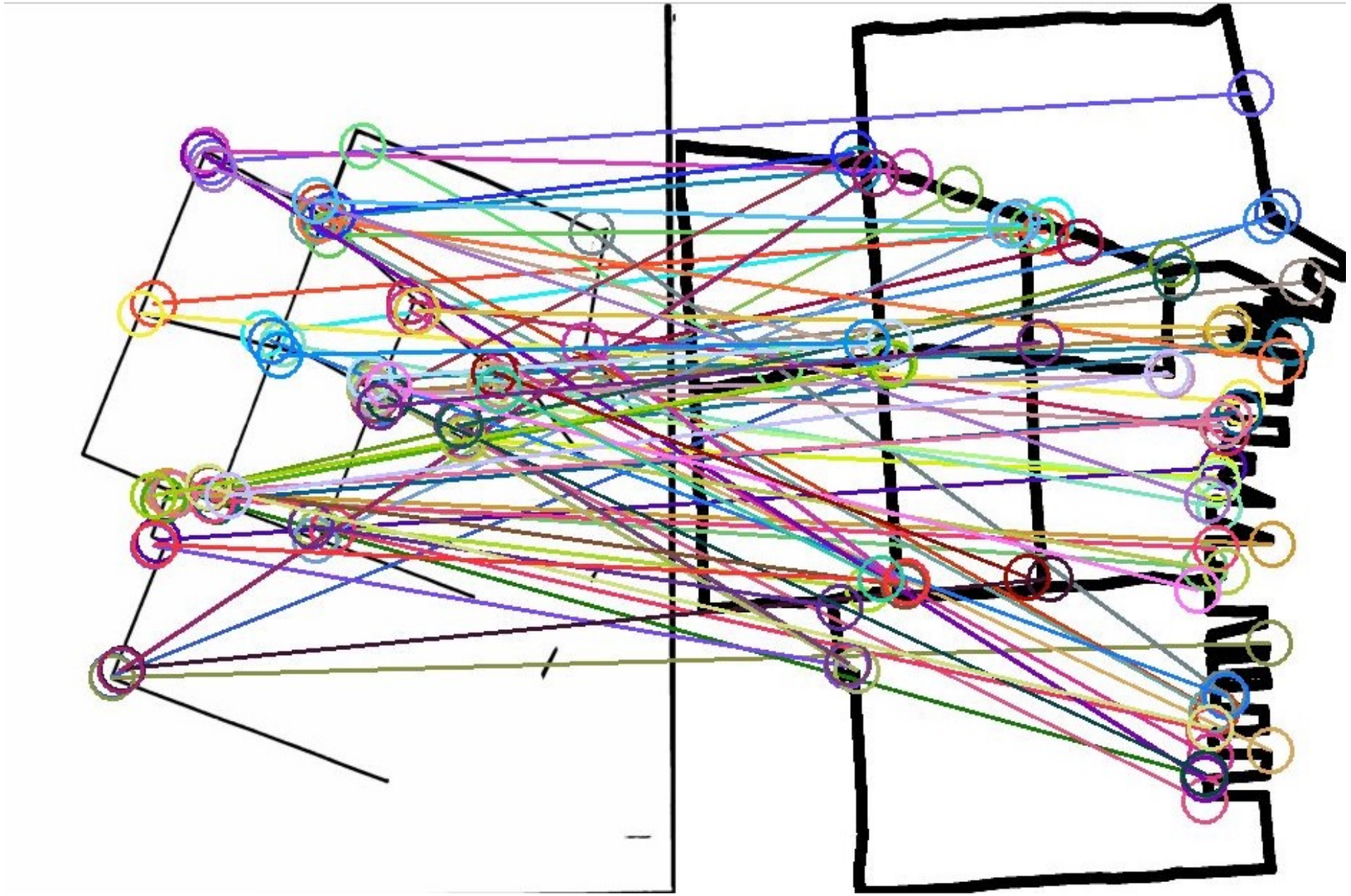


I still remember the block boundaries!



I still remember where I am in the world!





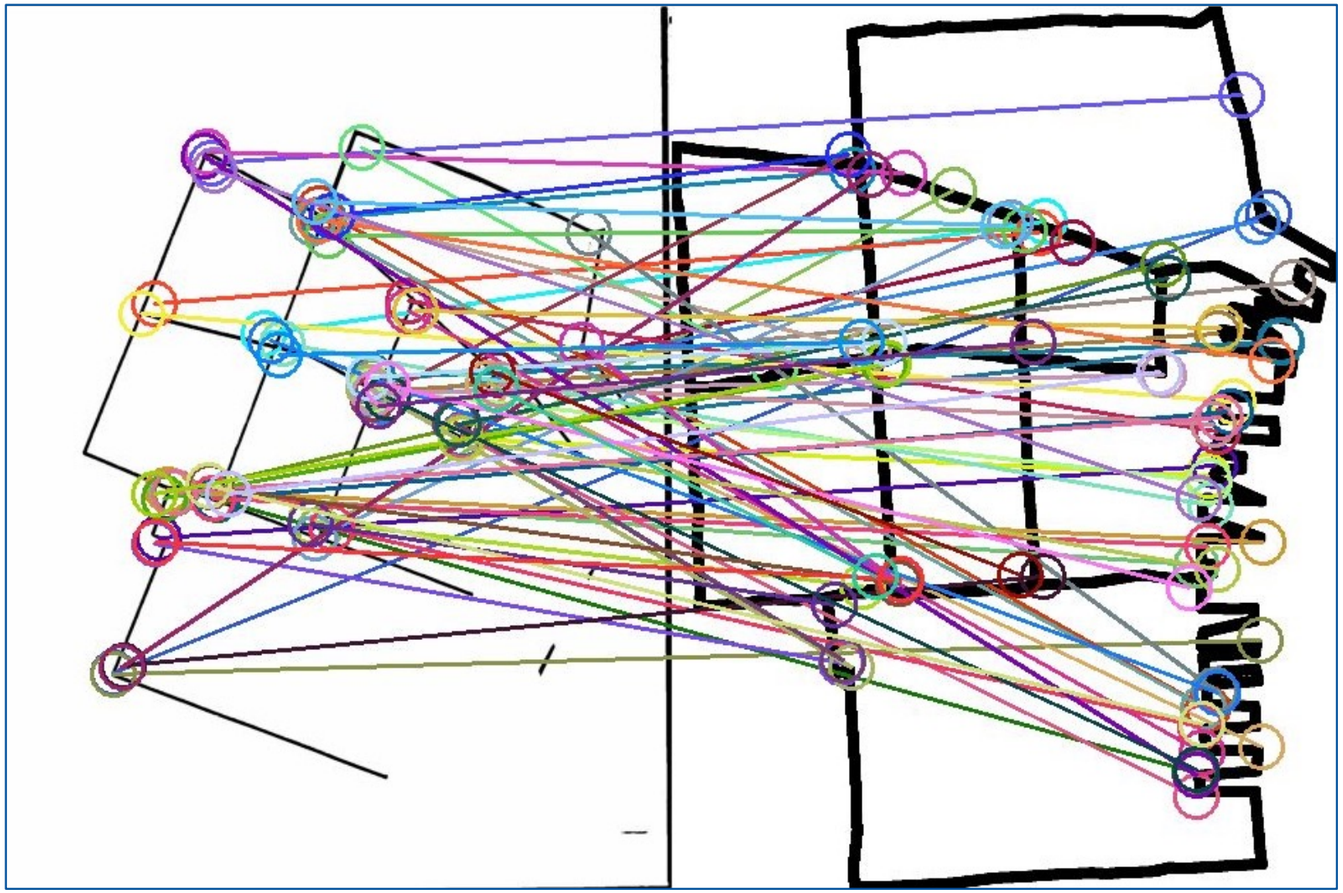
Super. Duper. Hard.

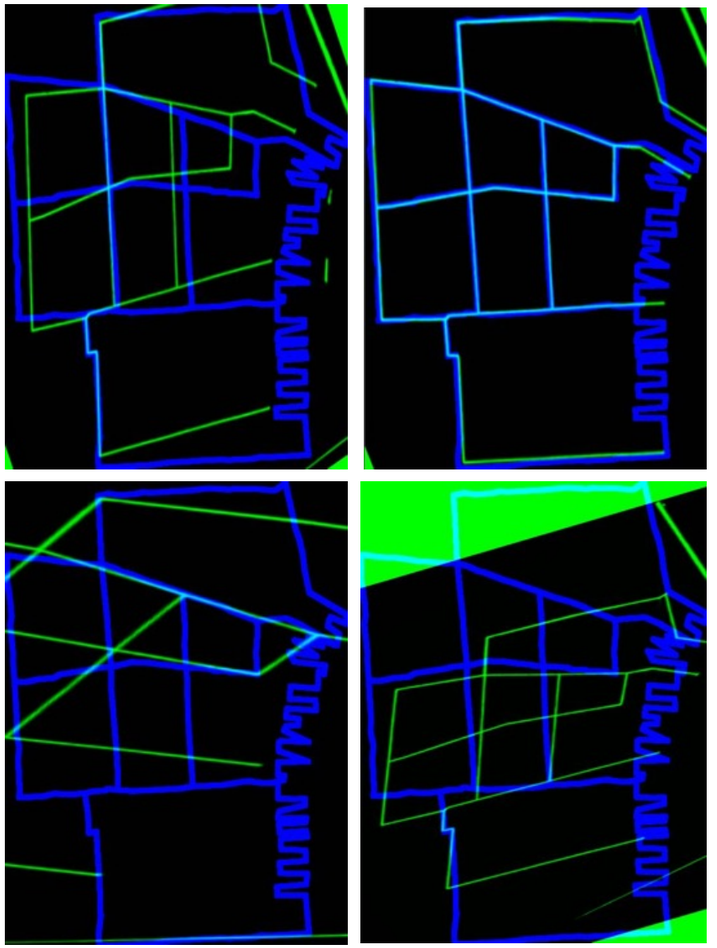
- Too many matches.
- Several per corner!
- We need to narrow these down.

The Basic Steps 🏇🕺

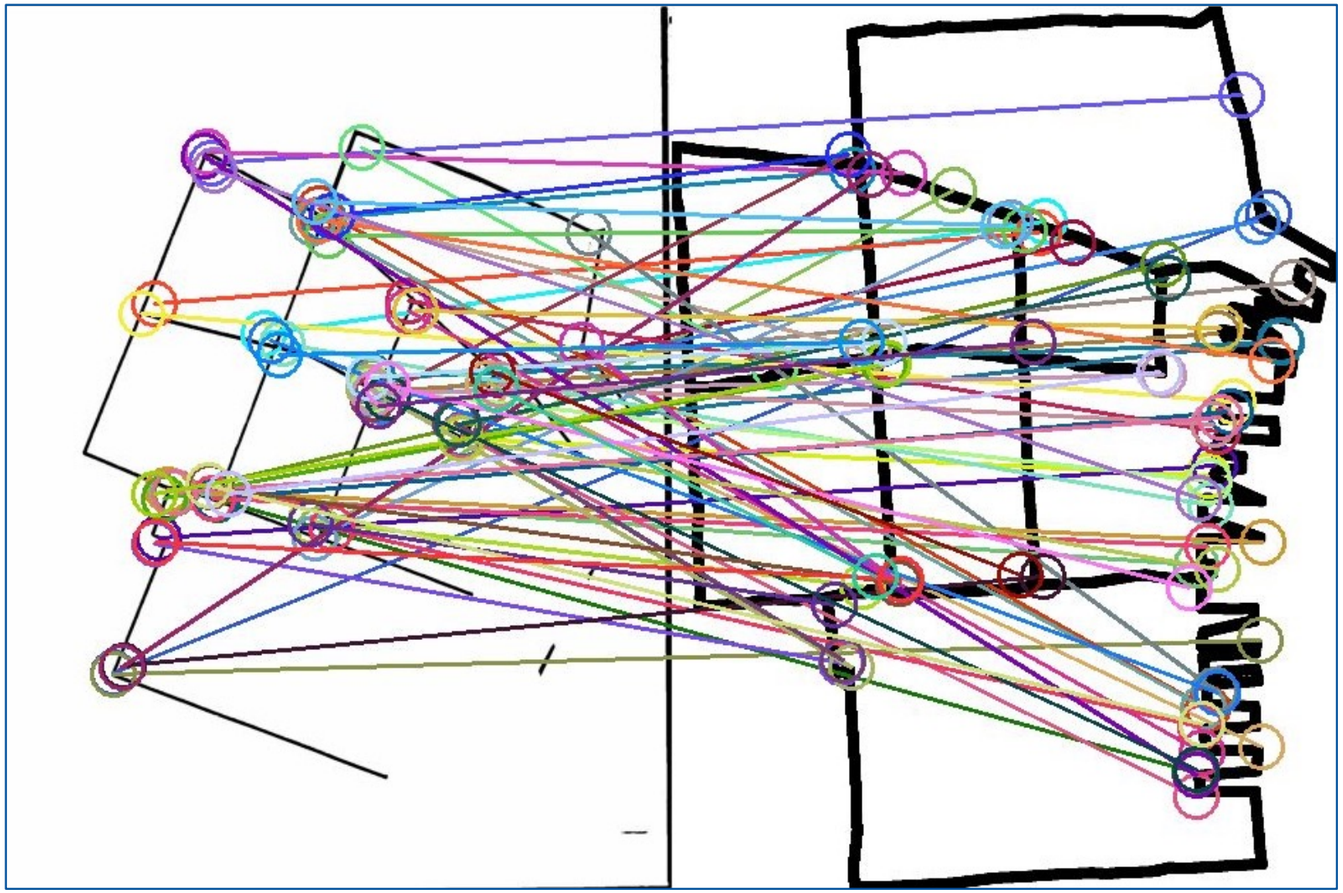
1	Make guess	Randomly select internally consistent links.
2	Evaluate guess	For selected links, how much do maps overlap?
3	Repeat	Keep best guess.

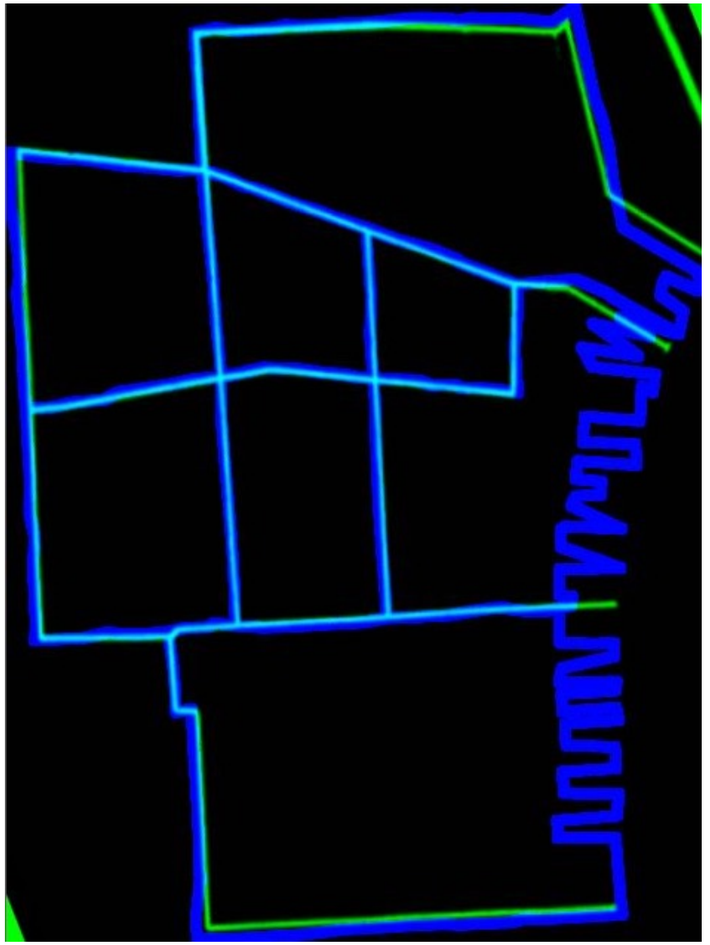
Selecting Links



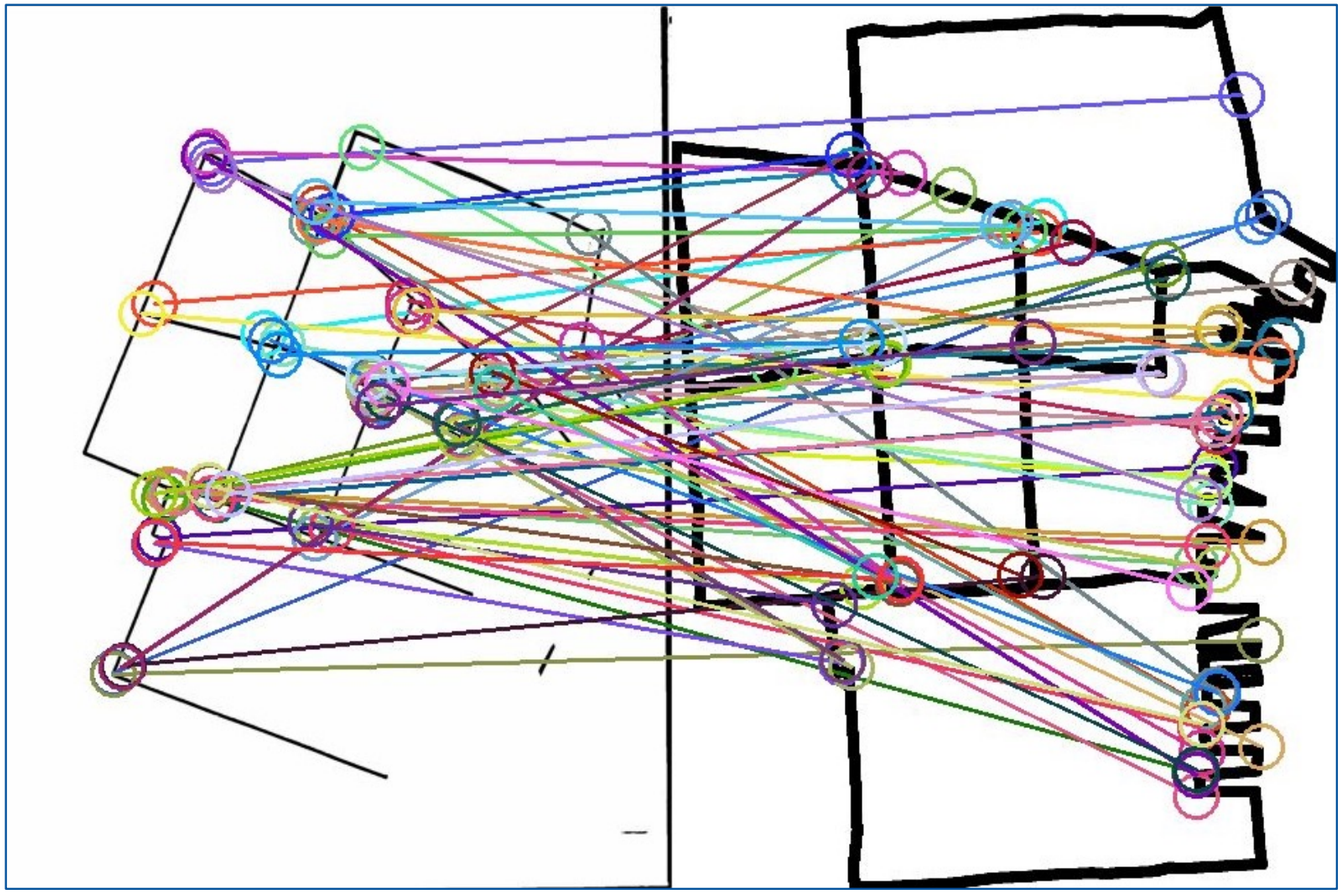


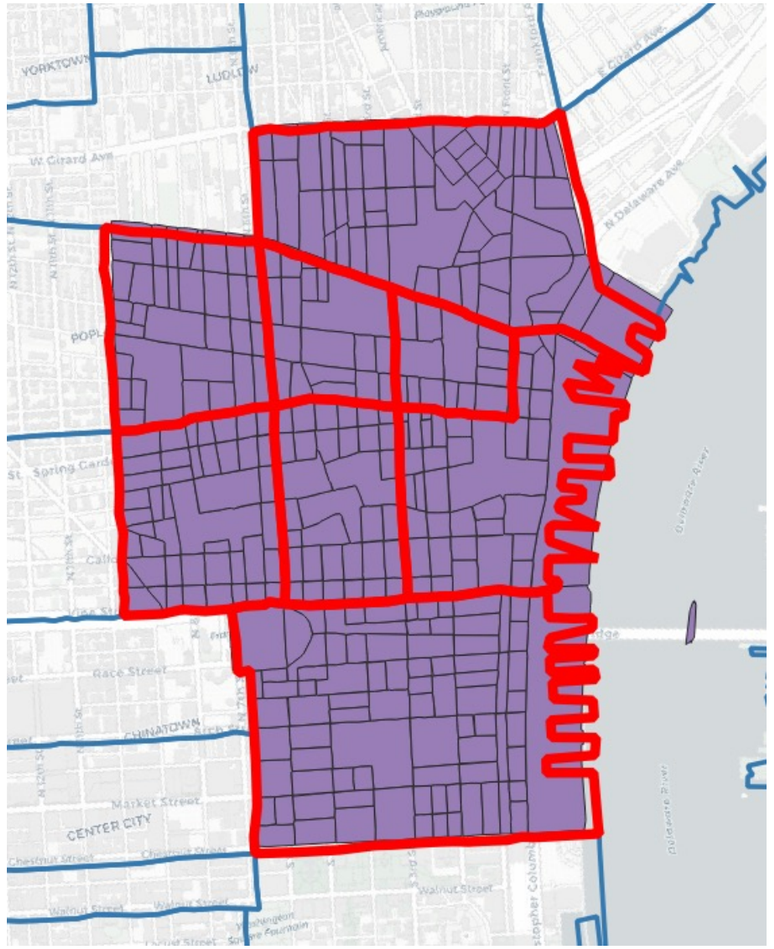
Selecting Links 





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
  - Chicago
  - Philadelphia
  - Los Angeles
  - Detroit
  - Baltimore
  - Cleveland
  - St. Louis
  - Washington, DC
  - Boston
  - San Francisco
  - Pittsburgh
  - Houston
  - Cincinnati
  - Columbus, OH
  - 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	ne-dwelling-tructures Average value (dollars)
10-8	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 (dollan)
10-e	11	4.425
11-A	12	7.288
11-e	15	8.15 0
12-A	16	6,366
	17	3,900
	18	9.SO 0
	19	8,6 3 J
	20	6.483
	21	◄.JS 7
	22	5.87 5
	23	◄.261
	24	5.34 5
	25	5,166
	26	4.80 0
	28	7.800
	29	6.0 0 0
	JO 31	◄.66 6
	32	4,125

Why Not Use Out of The Box Solutions?

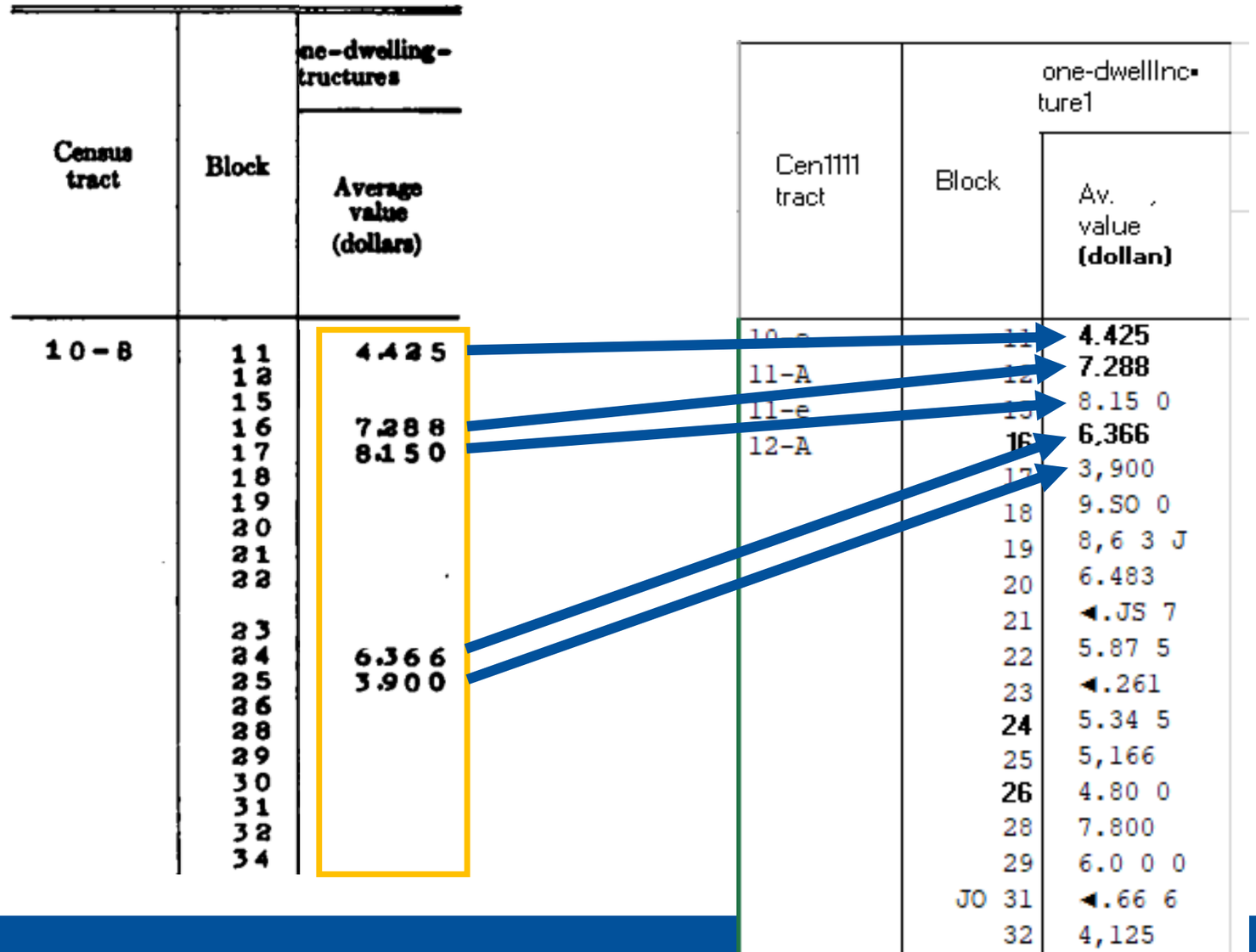
- Adobe:

Census tract	Block	ne-dwelling-tructures Average value (dollars)
10-8	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 (dollan)
10-e	11	4.425
11-A	12	7.288
11-e	15	8.15 0
12-A	16	6,366
	17	3,900
	18	9.SO 0
	19	8,6 3 J
	20	6.483
	21	◄.JS 7
	22	5.87 5
	23	◄.261
	24	5.34 5
	25	5,166
	26	4.80 0
	28	7.800
	29	6.0 0 0
	JO 31	◄.66 6
	32	4,125

Why Not Use Out of The Box Solutions?

- Adobe: Bad character recognition, relation of rows lost



Census tract	Block	ne-dwelling-tructures Average value (dollars)
10-8	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 (dollan)
	11	4.425
11-A	12	7.288
11-e	15	8.15 0
12-A	16	6,366
	17	3,900
	18	9.SO 0
	19	8,6 3 J
	20	6.483
	21	◄.JS 7
	22	5.87 5
	23	◄.261
	24	5.34 5
	25	5,166
	26	4.80 0
	28	7.800
	29	6.0 0 0
JO	31	◄.66 6
	32	4,125

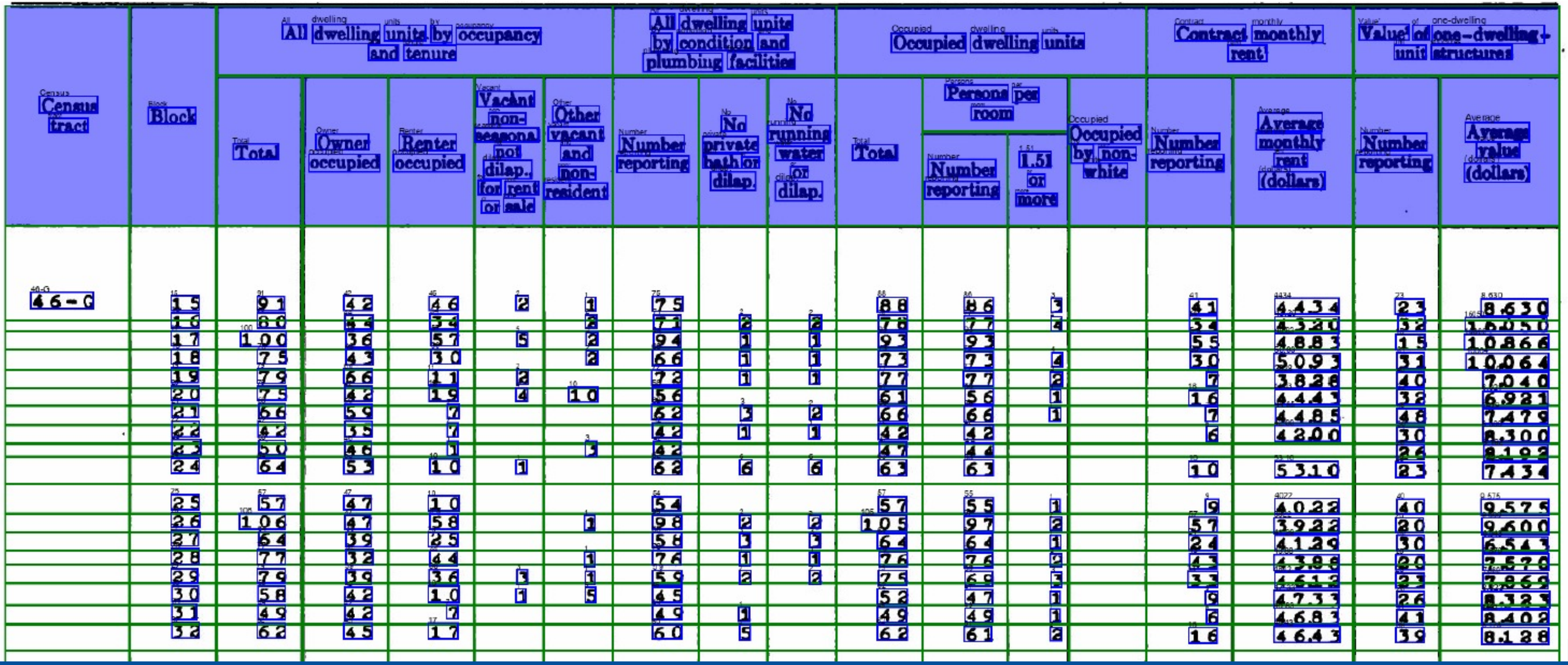
Why Not Use Out of The Box Solutions?

- 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		Occupied by non-white	Number reporting	Average monthly rent (dollars)	Number reporting	Average value (dollars)
											Number reporting	1.51 or more					
46-G	15	91	42	46	2	1	75			88	86	3		41	443	23	8630
	16	80	44	34		2	71	2	2	78	77	4		34	432	32	16050
	17	100	36	57	5	2	94	1	1	93	93			55	488	15	10866
	18	75	43	30		2	66	1	1	73	73	4		30	509	31	10064
	19	79	66	11	2		72	1	1	77	77	2		7	382	40	7040
	20	75	42	19	4	10	56			61	56	1		16	443	32	6921
	21	66	59	7			62	3	2	66	66	1		7	448	48	7479
	22	42	35	7			42	1	1	42	42			6	420	30	8300
	23	50	46	1		3	42			47	44					26	8192
	24	64	53	10	1		62	6	6	63	63			10	531	23	7434
	25	57	47	10			54			57	55	1		9	402	40	9575
	26	106	47	58		1	98	2	2	105	97	2		57	392	20	9600
	27	64	39	25			58	3	3	64	64	1		24	412	30	6543
	28	77	32	44		1	76	1	1	76	76	2		43	438	20	7570
	29	79	39	36	3	1	59	2	2	75	69	3		33	461	23	7869
	30	58	42	10	1	5	45			52	47	1		9	473	26	8323
	31	49	42	7			49	1		49	49	1		6	468	41	8402
	32	62	45	17			60	5		62	61	2		16	464	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)



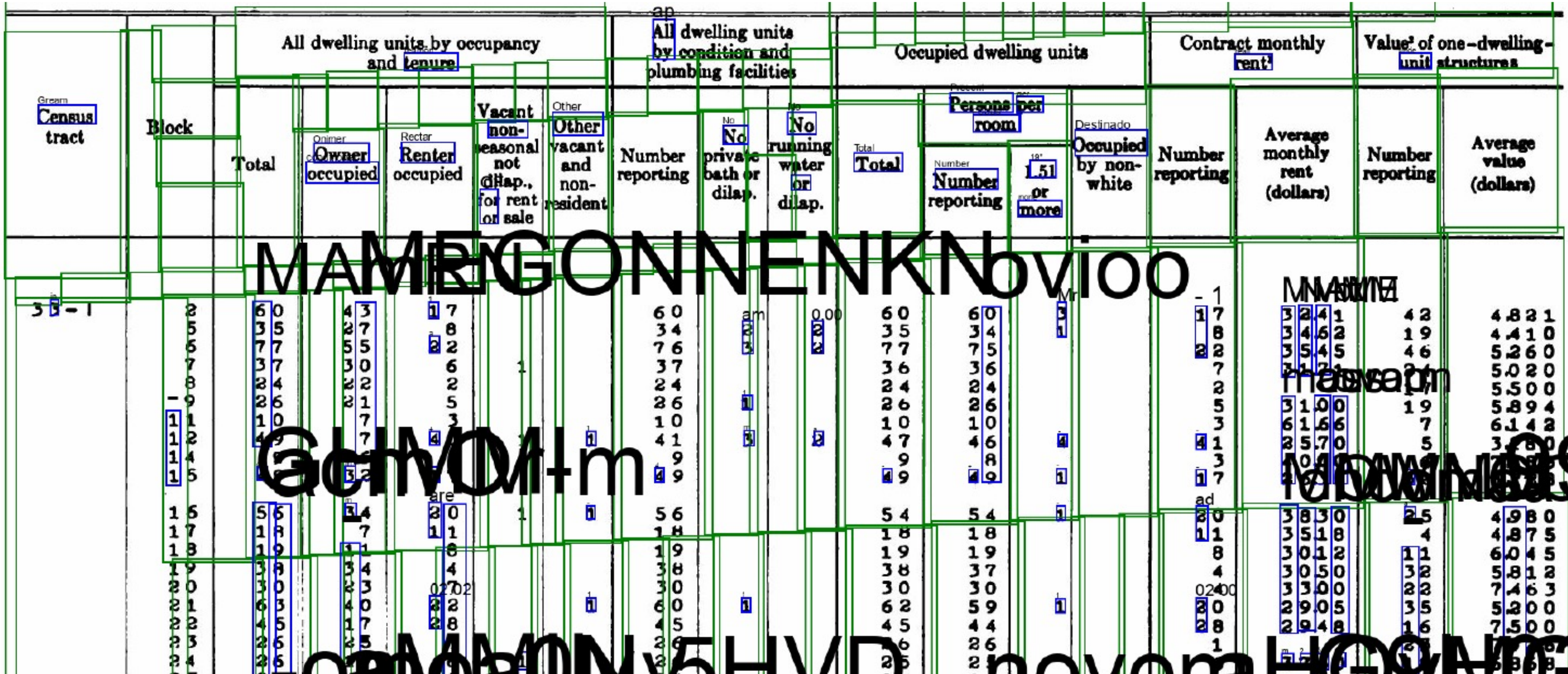
Why Not Use Out of The Box Solutions?

- 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	3		17	32.41	42	4821
	5	35	27	8			34	2	2	35	34	1		8	34.62	19	4410
	6	77	55	22			76	3	2	77	75			22	35.45	46	5260
	7	37	30	6	1		37			36	36			7	31.71	24	5020
	8	24	22	2			24			24	24			2		17	5500
	9	26	21	5			26	1		26	26			5	31.00	19	5894
	11	10	7	3			10			10	10			3	61.66	7	6142
	12	49	7	40	1	1	41	3	2	47	46	4		41	35.70	5	3180
	14	9	6	3			9			9	8			3	40.00	6	7483
	15	49	32	17			49			49	49	1		17	26.35	32	3728
	16	56	34	20	1	1	56			54	54	1		20	38.30	25	4980
	17	18	7	11			18			18	18			11	35.18	4	4875
	18	19	11	8			19			19	19			8	30.12	11	6045
	19	38	34	4			38			38	37			4	30.50	32	5812
	20	30	23	7			30			30	30			4	33.00	22	7463
	21	63	40	22		1	60	1		62	59	1		20	29.05	35	5200
	22	45	17	28			45			45	44			28	29.48	16	7500
	23	26	25	1			26			26	26			1		23	7026
	24	26	19	6	1		26			25	25			6	37.50	19	6868

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.



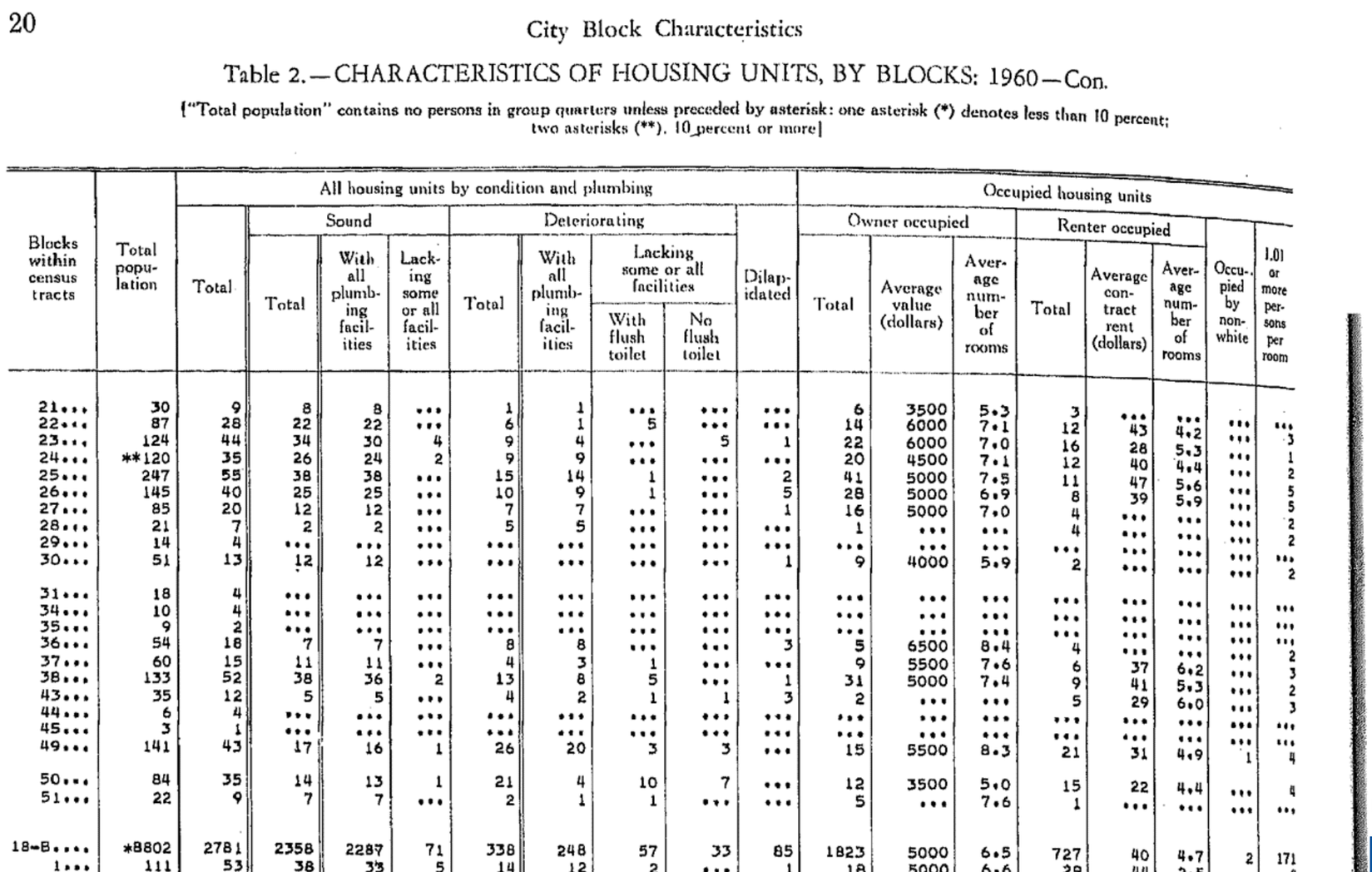
Why Not Use Out of The Box Solutions?

- 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

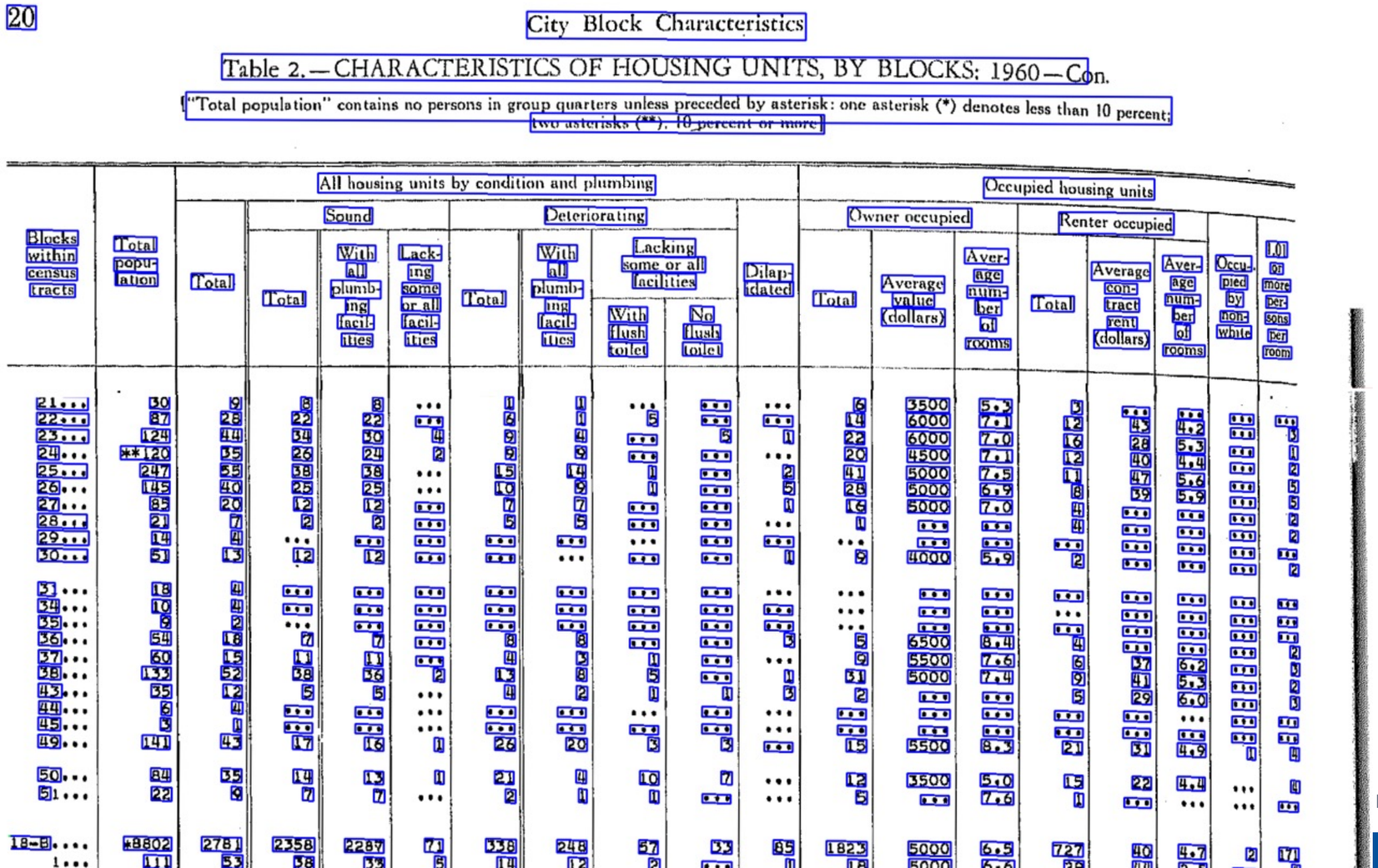
Table 2.—CHARACTERISTICS OF HOUSING UNITS, BY BLOCKS: 1960—Con.

["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 population	All housing units by condition and plumbing									Occupied housing units							
		Total	Sound			Deteriorating				Dilapidated	Owner occupied			Renter occupied			Occupied by nonwhite	1.01 or more persons per room
			Total	With all plumbing facilities	Lacking some or all facilities	Total	With all plumbing facilities	Lacking some or all facilities			Total	Average value (dollars)	Average number of rooms	Total	Average contract rent (dollars)	Average number of rooms		
								With flush toilet	No flush toilet									
21	30	9	8	8	...	1	1	...	...	...	6	3500	5.3	3	...	...	...	...
22	87	28	22	22	...	6	1	5	...	...	14	6000	7.1	12	43	4.2	...	...
23	124	44	34	30	4	9	4	...	5	1	22	6000	7.0	16	28	5.3	...	3
24	**120	35	26	24	2	9	9	...	...	...	20	4500	7.1	12	40	4.4	...	1
25	247	55	38	38	...	15	14	1	...	2	41	5000	7.5	11	47	5.6	...	2
26	145	40	25	25	...	10	9	1	...	5	28	5000	6.9	8	39	5.9	...	5
27	85	20	12	12	...	7	7	...	...	1	16	5000	7.0	4	...	...	...	5
28	21	7	2	2	...	5	5	...	...	...	1	...	...	4	...	...	...	2
29	14	4	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	2
30	51	13	12	12	...	...	...	...	...	1	9	4000	5.9	2	...	...	...	2
31	18	4	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
34	10	4	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
35	9	2	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
36	54	18	7	7	...	8	8	...	...	3	5	6500	8.4	4	...	...	...	...
37	60	15	11	11	...	4	3	1	...	...	9	5500	7.6	6	37	6.2	...	2
38	133	52	38	36	2	13	8	5	...	1	31	5000	7.4	9	41	5.3	...	3
43	35	12	5	5	...	4	2	1	1	3	2	...	...	5	29	6.0	...	2
44	6	4	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	3
45	3	1	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
49	141	43	17	16	1	26	20	3	3	...	15	5500	8.3	21	31	4.9	1	4
50	84	35	14	13	1	21	4	10	7	...	12	3500	5.0	15	22	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	...	...	...	...

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]

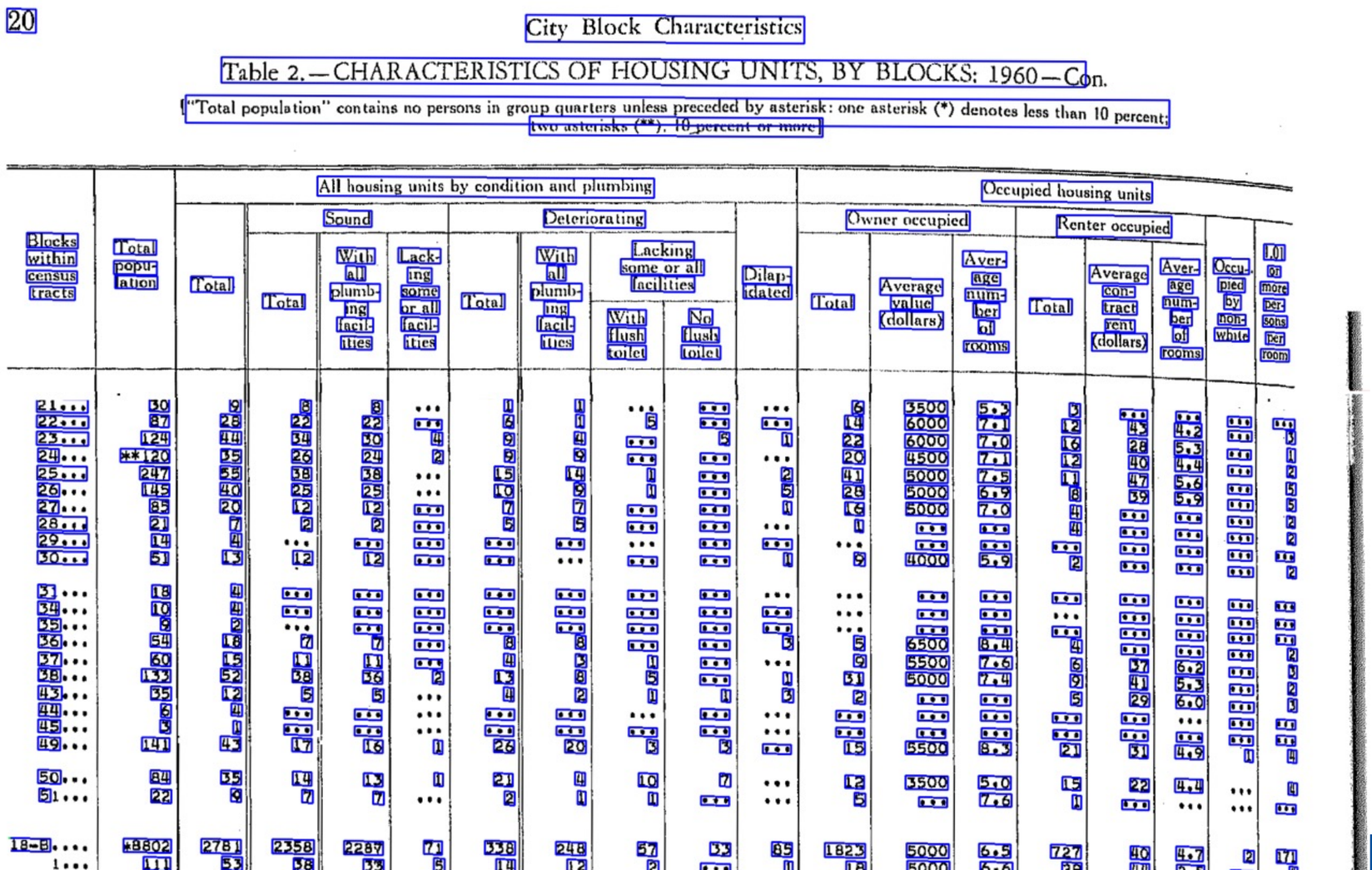
Blocks within census tracts	Total population	All housing units by condition and plumbing								Occupied housing units							
		Total	Sound			Deteriorating				Owner occupied			Renter occupied				
			Total	With all plumbing facilities	Lacking some or all facilities	Total	With all plumbing facilities	Lacking some or all facilities	Dilapidated	Total	Average value (dollars)	Average number of rooms	Total	Average contract rent (dollars)	Average number of rooms	Occupied by nonwhite	1.01 or more persons per room
								With flush toilet	No flush toilet								
21	30	9	8	8	...	1	1	...	...	6	3500	5.3	3	...	...	...	...
22	87	28	22	22	...	6	1	5	...	14	6000	7.1	12	45	4.2	...	...
23	124	44	34	30	4	9	4	...	5	22	6000	7.0	16	28	5.3	...	...
24	**120	35	26	24	2	9	9	...	...	20	4500	7.1	12	40	4.4	...	...
25	247	55	38	38	...	15	14	1	2	41	5000	7.5	11	47	5.6	...	...
26	145	40	25	25	...	10	9	1	5	28	5000	6.9	8	39	5.9	...	...
27	85	20	12	12	...	7	7	...	1	16	5000	7.0	4	...	...	...	...
28	21	7	2	2	...	5	5	...	...	1	...	...	4	...	...	...	...
29	14	4	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
30	51	13	12	12	...	...	...	...	1	9	4000	5.9	2	...	...	...	...
31	18	4	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
34	10	4	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
35	6	2	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
36	54	18	7	7	...	8	8	...	3	5	6500	8.4	4	...	...	...	...
37	60	15	11	11	...	4	3	1	...	9	5500	7.6	6	37	6.2	...	...
38	133	52	38	36	2	13	8	5	1	51	5000	7.4	9	41	5.3	...	...
43	35	12	5	5	...	4	2	1	3	2	...	...	5	29	6.0	...	...
44	9	4	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
45	9	1	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
49	141	43	17	16	1	26	20	3	3	15	5500	8.3	21	31	4.9	11	4
50	84	35	14	13	1	21	4	10	7	12	3500	5.0	15	22	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	5000	6.6	38	...	...	...	...	...

Isolate Table Body, Straighten Image

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

Mostly Text

Mostly Numbers



20

City Block Characteristics

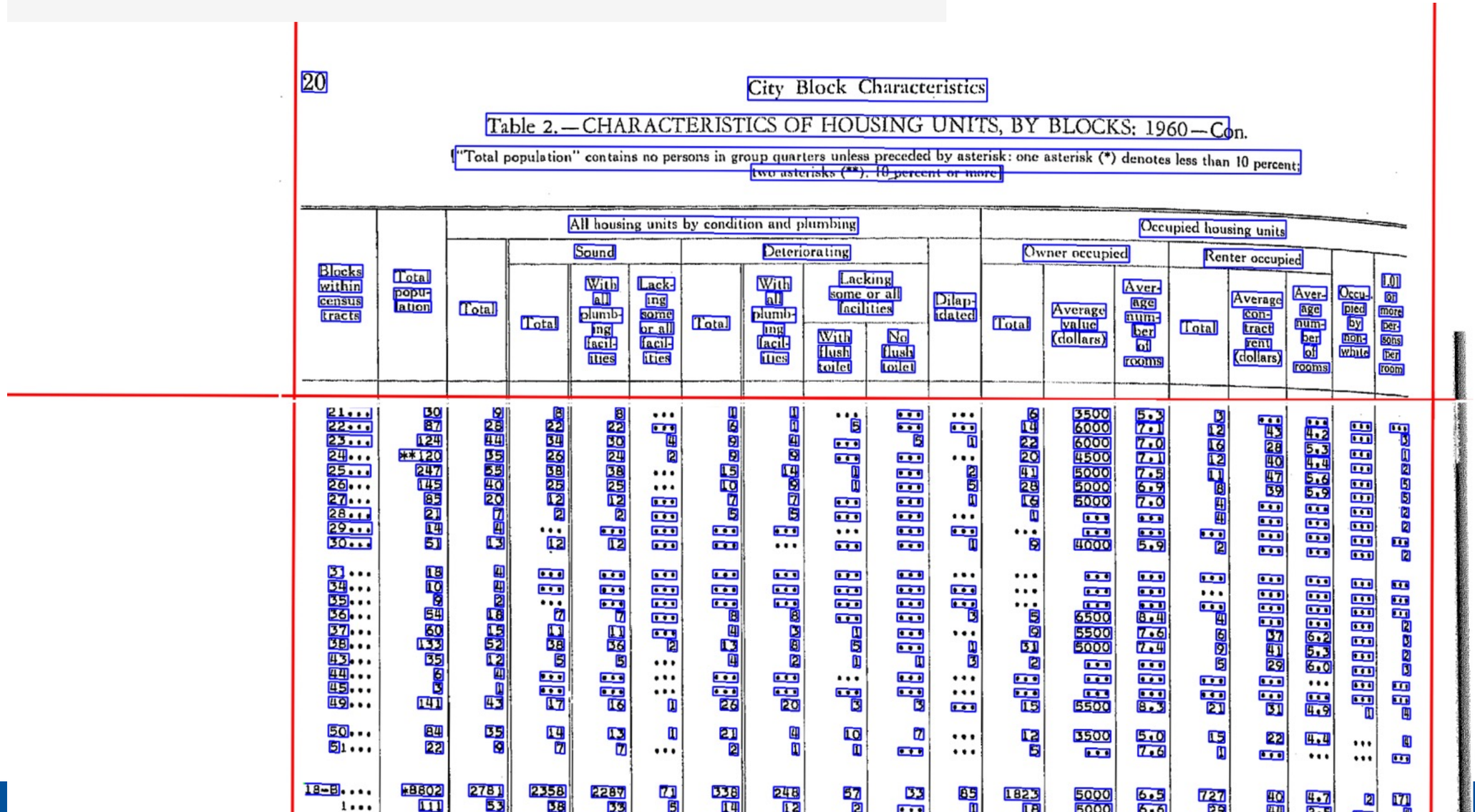
Table 2.—CHARACTERISTICS OF HOUSING UNITS, BY BLOCKS: 1960—Con.

["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 population	All housing units by condition and plumbing								Occupied housing units							
		Total	Sound			Deteriorating				Dilapidated	Owner occupied			Renter occupied				
			Total	With all plumbing facilities	Lacking some or all facilities	Total	With all plumbing facilities	Lacking some or all facilities			Total	Average value (dollars)	Average number of rooms	Total	Average contract rent (dollars)	Average number of rooms	Occupied by nonwhite	1.01 or more persons per room
								With flush toilet	No flush toilet									
21	30	9	8	8	...	1	1	...	...	...	6	3500	5.3	3	...	4.3	...	...
22	87	28	22	22	...	6	1	5	...	...	14	6000	7.1	12	45	4.2	...	...
23	124	44	34	30	4	9	4	...	5	1	22	6000	7.0	16	28	5.3	...	...
24	**120	35	26	24	2	9	4	...	5	...	20	4500	7.1	12	40	4.4	...	...
25	247	55	38	38	...	15	14	1	...	2	41	5000	7.5	11	47	5.6	...	...
26	145	40	25	25	...	10	9	1	...	5	28	5000	6.9	8	39	5.9	...	...
27	85	20	12	12	...	7	7	...	...	1	16	5000	7.0	4	...	...	...	...
28	21	7	2	2	...	5	5	...	...	...	1	...	...	4	...	...	...	...
29	14	4	...	...	...	...	...	...	...	...	...	...	...	4	...	...	...	...
30	51	13	12	12	...	...	...	...	...	1	9	4000	5.9	2	...	...	...	...
31	18	4	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
34	10	4	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
35	6	2	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
36	54	18	7	7	...	8	8	...	...	3	5	6500	8.4	4	...	...	...	...
37	60	15	11	11	...	4	3	1	...	...	9	5500	7.6	6	37	6.2	...	...
38	133	52	38	36	2	13	8	5	...	1	51	5000	7.4	9	41	5.3	...	...
43	35	12	5	5	...	4	2	1	...	3	2	...	...	5	29	6.0	...	...
44	9	4	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
45	9	1	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
49	141	43	17	16	1	26	20	5	5	...	15	5500	8.3	21	31	4.9	11	4
50	84	35	14	13	1	21	4	10	7	...	12	3500	5.0	15	22	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	38	...	...	...	...

Isolate Table Body, Straighten Image

Table is isolated



20

City Block Characteristics

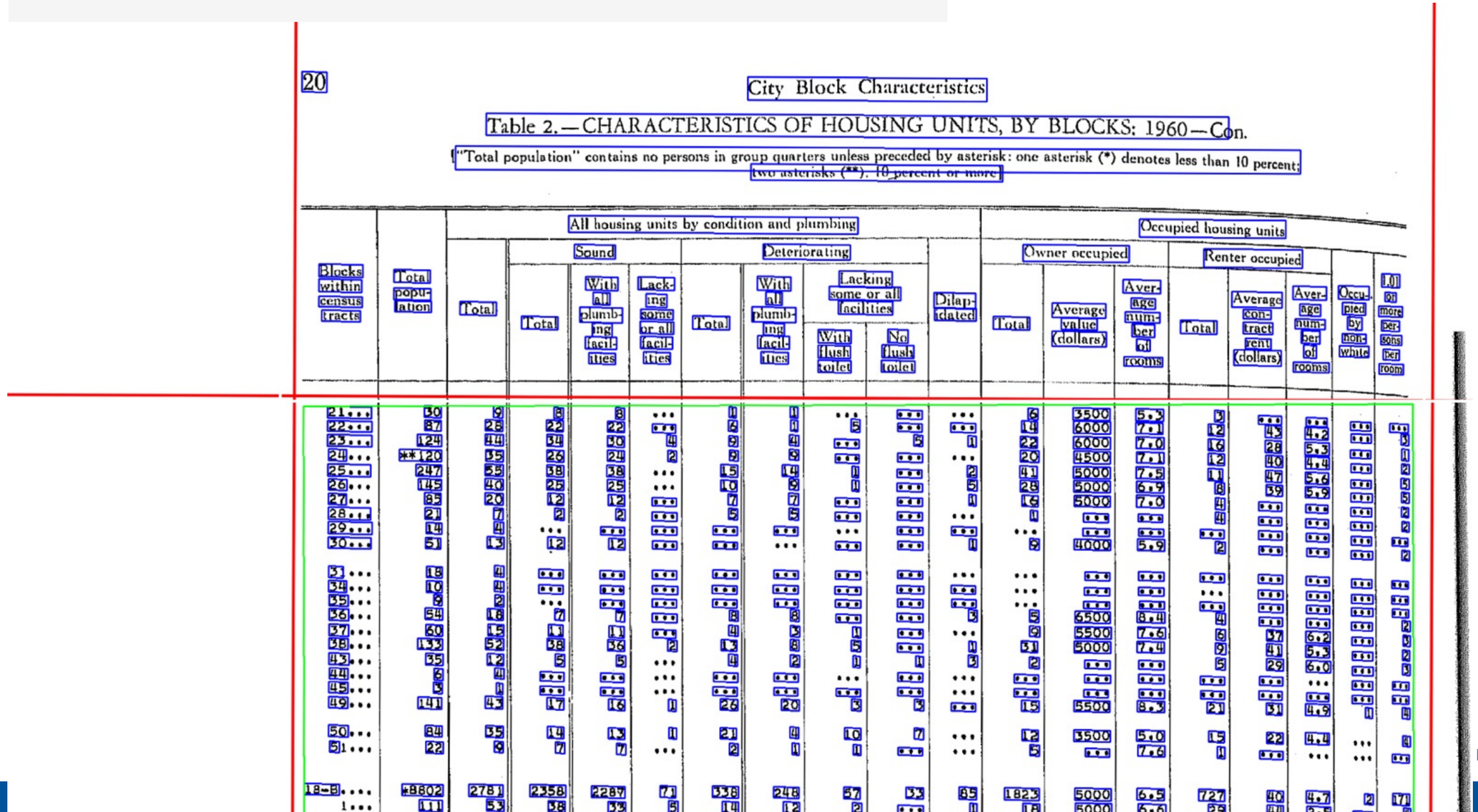
Table 2.—CHARACTERISTICS OF HOUSING UNITS, BY BLOCKS: 1960—Con.

"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 population	All housing units by condition and plumbing									Occupied housing units							
		Total	Sound			Deteriorating				Dilapidated	Owner occupied			Renter occupied				
			Total	With all plumbing facilities	Lacking some or all facilities	Total	With all plumbing facilities	Lacking some or all facilities			Total	Average value (dollars)	Average number of rooms	Total	Average contract rent (dollars)	Average number of rooms	Occupied by nonwhite	1.01 or more persons per room
								With flush toilet	No flush toilet									
21	30	9	8	8	...	1	1	...	...	...	6	3500	5.3	3	...	...	...	...
22	87	28	22	22	...	6	1	5	...	...	14	6000	7.1	12	45	4.2	...	...
23	124	44	34	30	4	9	4	...	5	1	22	6000	7.0	10	28	5.3	...	...
24	**120	35	26	24	2	9	9	...	...	...	20	4500	7.1	12	40	4.4	...	...
25	247	55	38	38	...	15	14	1	...	2	41	5000	7.5	11	47	5.6	...	...
26	145	40	25	25	...	10	9	1	...	5	28	5000	6.9	8	39	5.9	...	...
27	85	20	12	12	...	7	7	...	...	1	16	5000	7.0	4	...	...	...	...
28	21	7	2	2	...	5	5	...	...	...	1	...	...	4	...	...	...	...
29	14	4	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
30	51	13	12	12	...	...	...	...	...	1	9	4000	5.9	2	...	...	...	...
31	18	4	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
34	10	4	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
35	9	2	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
36	54	18	7	7	...	8	8	...	...	3	5	6500	8.4	4	...	...	...	...
37	60	15	11	11	...	4	3	1	...	...	9	5500	7.6	6	37	6.2	...	...
38	133	52	38	36	2	13	8	5	...	1	51	5000	7.4	9	41	5.3	...	...
43	35	12	5	5	...	4	2	1	1	3	2	...	...	5	29	6.0	...	...
44	6	4	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
45	6	1	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
49	141	43	17	16	1	26	20	5	3	...	15	5500	8.3	21	31	4.9	1	...
50	84	35	14	13	1	21	4	10	7	...	12	3500	5.0	15	22	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	17
1	111	53	38	33	5	14	12	2	...	1	18	5000	6.6	28	41	4.8	...	1

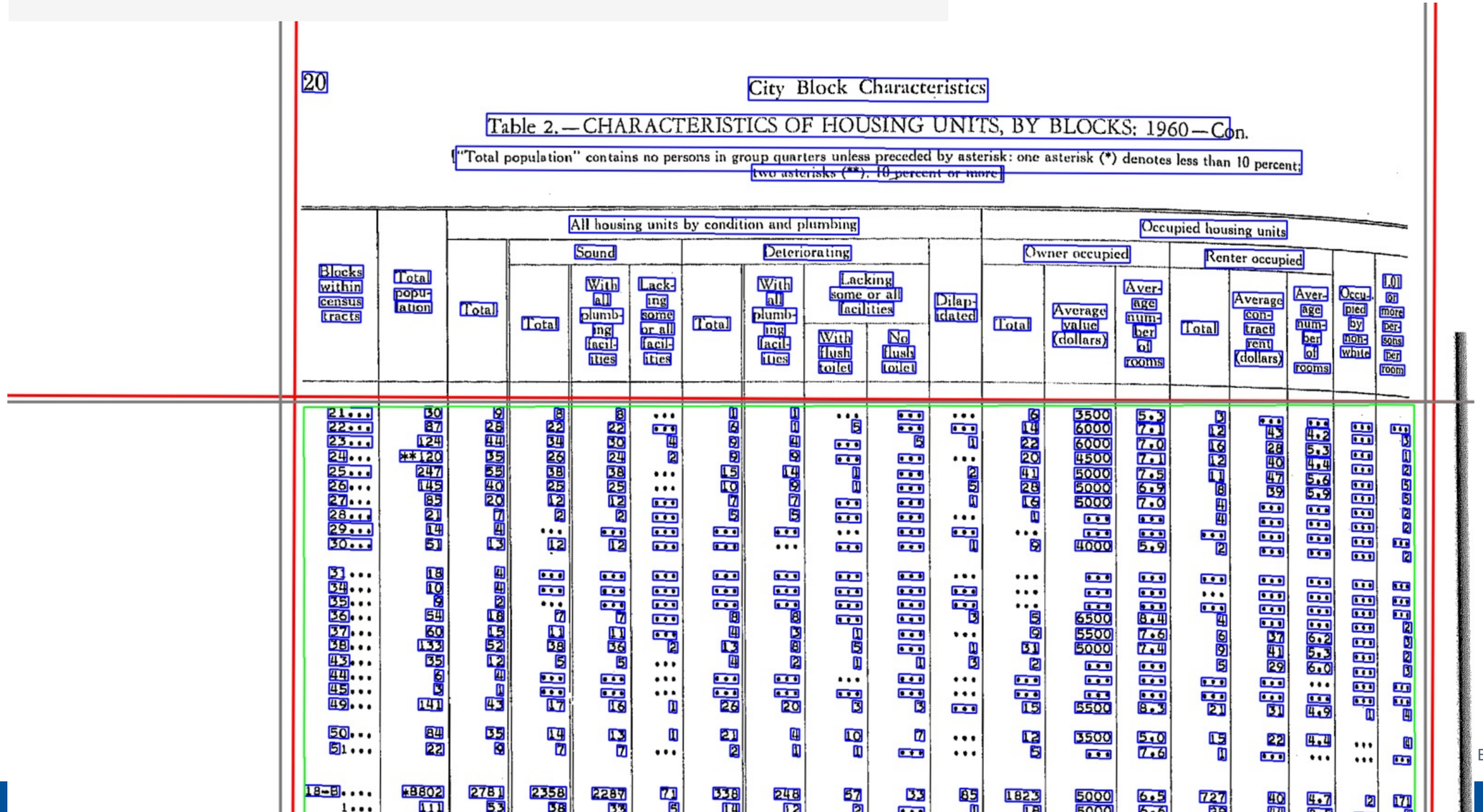
Isolate Table Body, Straighten Image

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



Isolate Table Body, Straighten Image

Rotate image around center of table – image is straightened



20

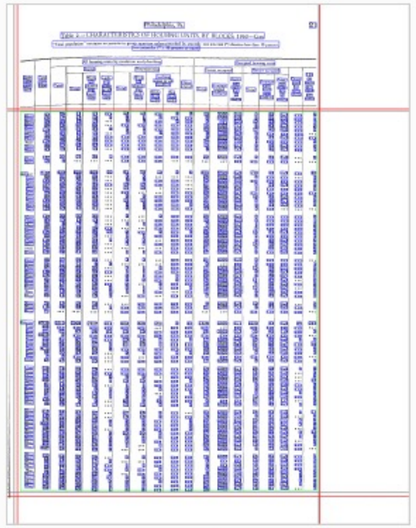
City Block Characteristics

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

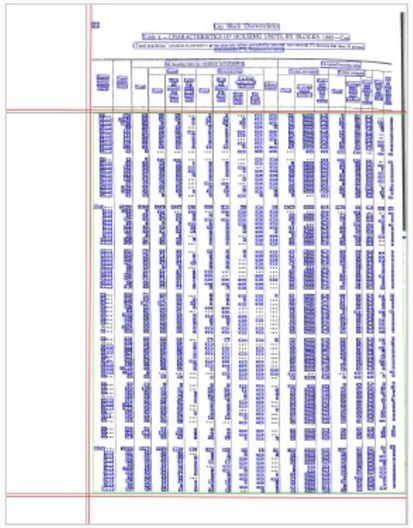
"Total population" contains no persons in group quarters unless preceded by asterisk: one asterisk (*) denotes less than 10 percent; two asterisks (**), 10 percent or more.

Blocks within census tracts	Total population	All housing units by condition and plumbing									Occupied housing units							
		Total	Sound			Deteriorating				Dilapidated	Owner occupied			Renter occupied				
			Total	With all plumbing facilities	Lacking some or all facilities	Total	With all plumbing facilities	Lacking some or all facilities			Total	Average value (dollars)	Average number of rooms	Total	Average contract rent (dollars)	Average number of rooms	Occupied by nonwhite	1.01 or more persons per room
								With flush toilet	No flush toilet									
21	30	9	8	8	...	1	1	...	...	...	6	3500	5.3	3	...	...	...	...
22	87	28	22	22	...	6	1	5	...	...	14	6000	7.1	12	43	4.2	...	...
23	124	44	34	30	4	9	4	...	5	1	22	6000	7.0	16	28	5.3	...	3
24	**120	35	26	24	2	9	9	...	...	...	20	4500	7.1	12	40	4.4	...	1
25	247	55	38	38	...	15	14	1	...	2	41	5000	7.5	11	47	5.6	...	2
26	145	40	25	25	...	10	9	1	...	5	28	5000	6.9	8	39	5.9	...	5
27	85	20	12	12	...	7	7	...	...	1	16	5000	7.0	4	...	...	...	5
28	21	7	2	2	...	5	5	...	...	...	1	...	...	4	...	...	...	2
29	14	4	...	...	...	...	...	...	...	...	...	...	...	4	...	...	...	2
30	51	13	12	12	...	...	...	...	...	1	9	4000	5.9	2	...	...	...	2
31	18	4	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
34	10	4	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
35	6	2	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
36	54	18	7	7	...	8	8	...	...	3	5	6500	8.4	4	...	...	...	2
37	60	15	11	11	...	4	3	1	...	...	9	5500	7.6	6	37	6.2	...	3
38	133	52	38	36	2	13	8	5	...	1	31	5000	7.4	9	41	5.3	...	2
43	35	12	5	5	...	4	2	1	1	3	2	...	...	5	29	6.0	...	3
44	9	4	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
45	5	1	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
49	141	43	17	16	1	26	20	5	3	...	15	5500	8.9	21	31	4.9	1	4
50	84	35	14	13	1	21	4	10	7	...	12	3500	5.0	15	22	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	20	38	5.2	...	2

Always Be Checking

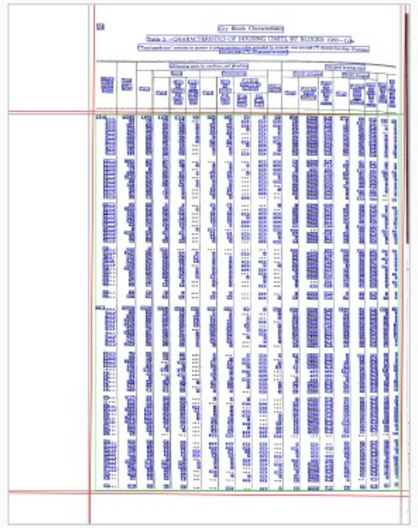


21

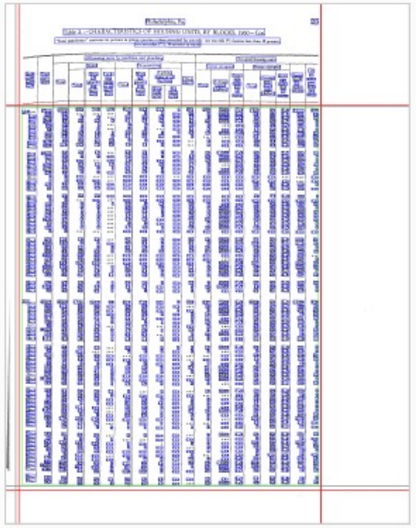


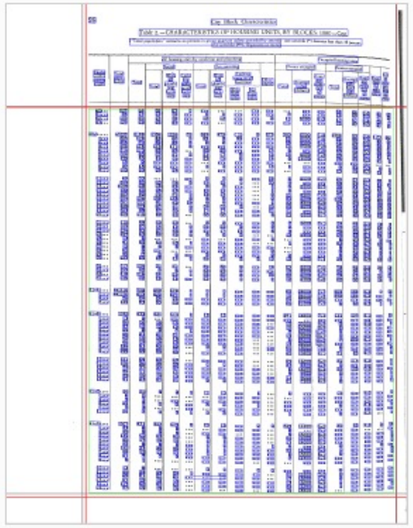
22

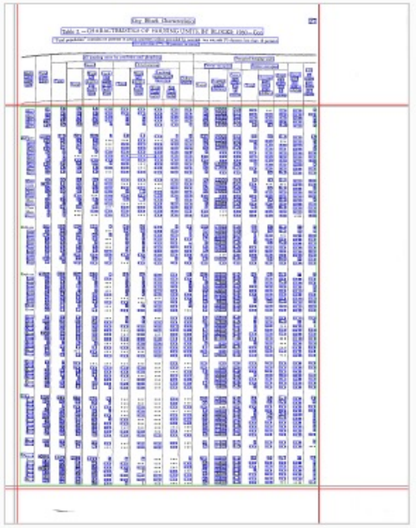
23

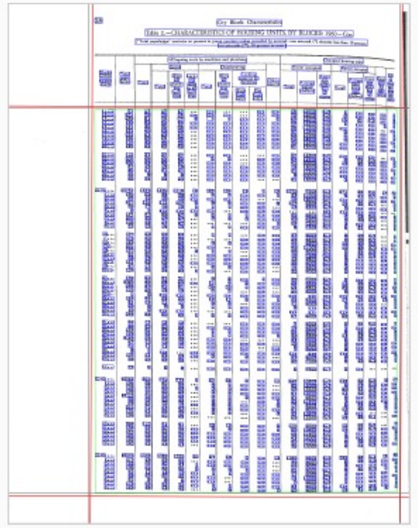


24

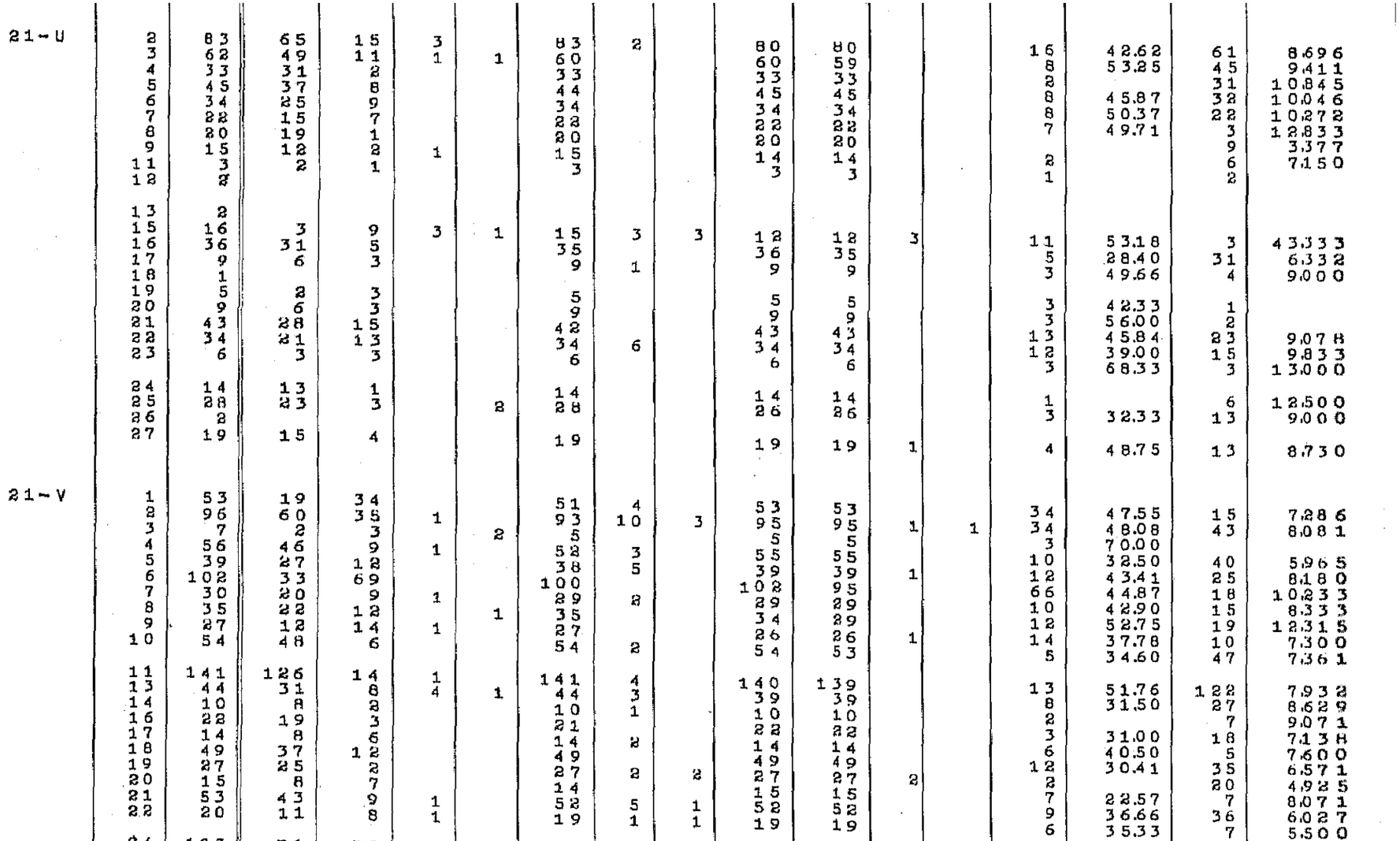






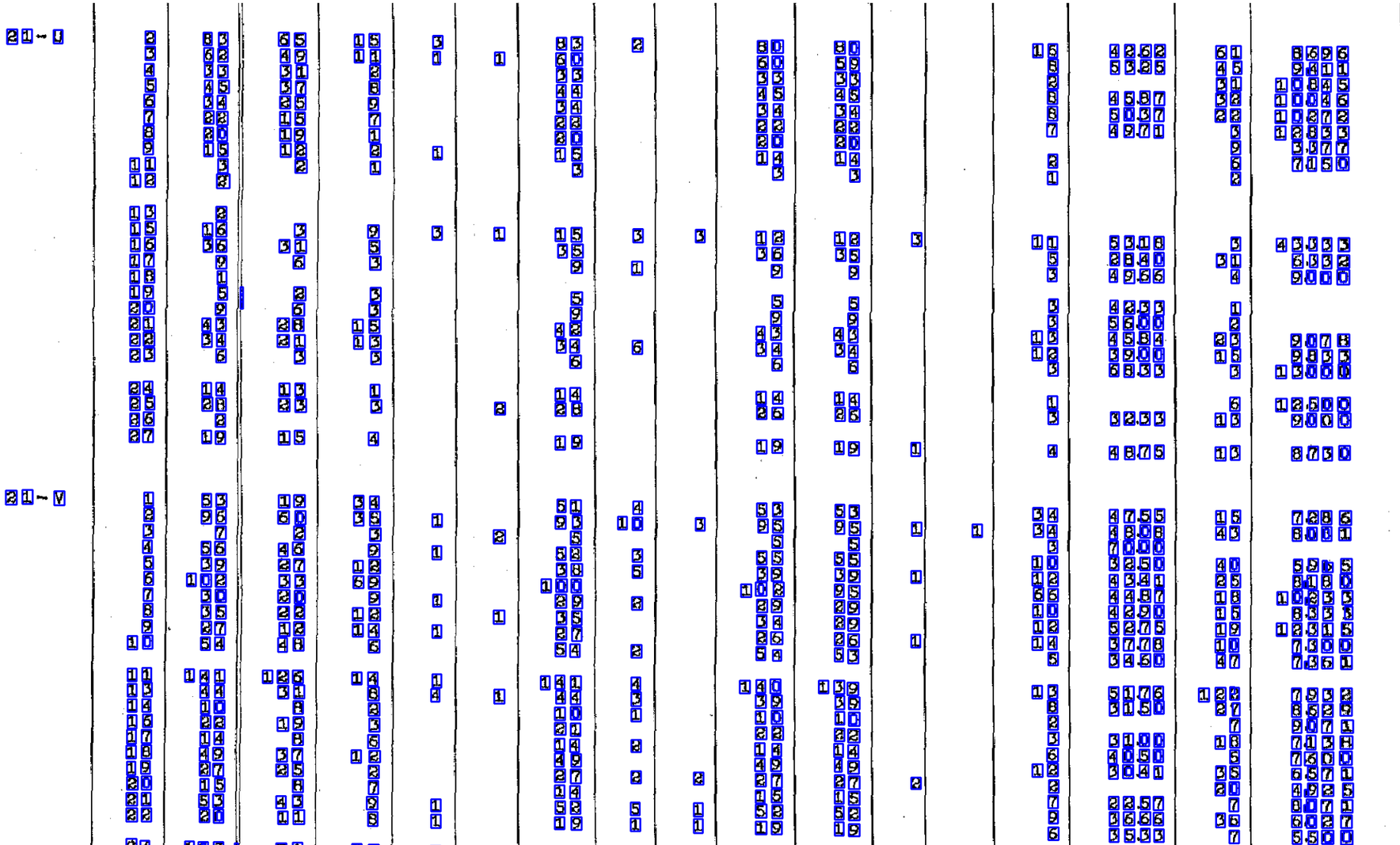


Isolate Columns



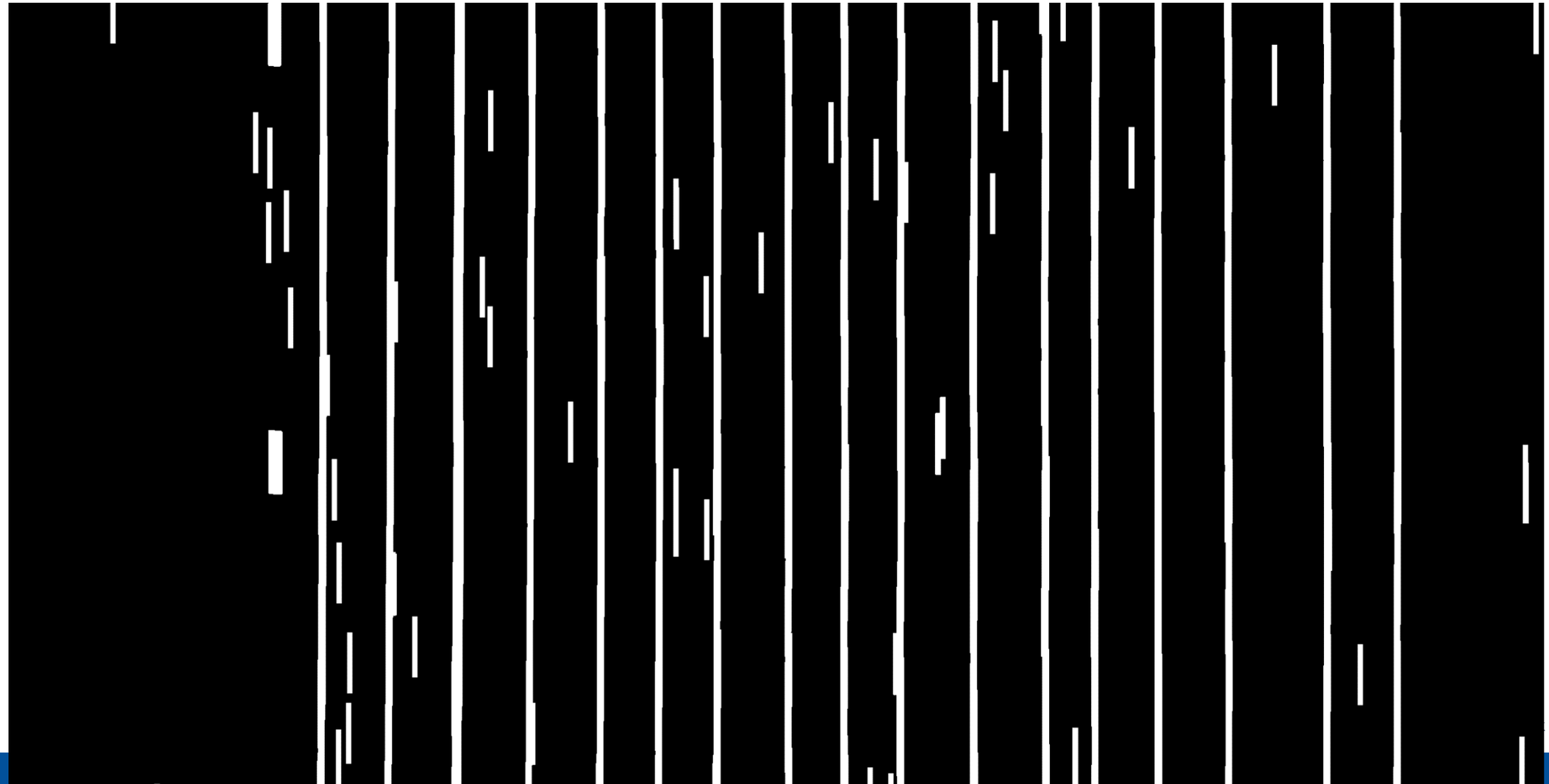
Isolate Columns

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



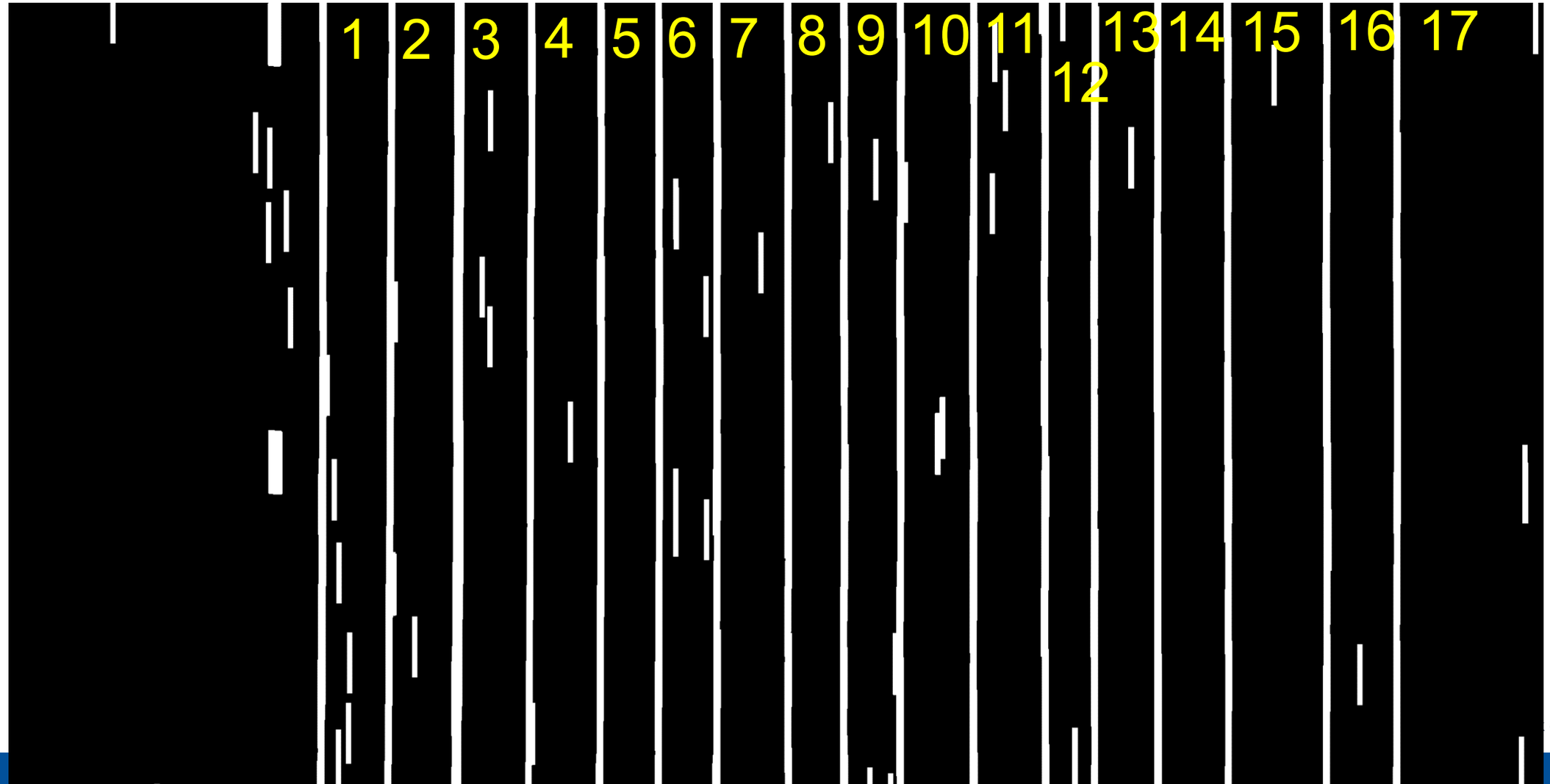
Isolate Columns

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



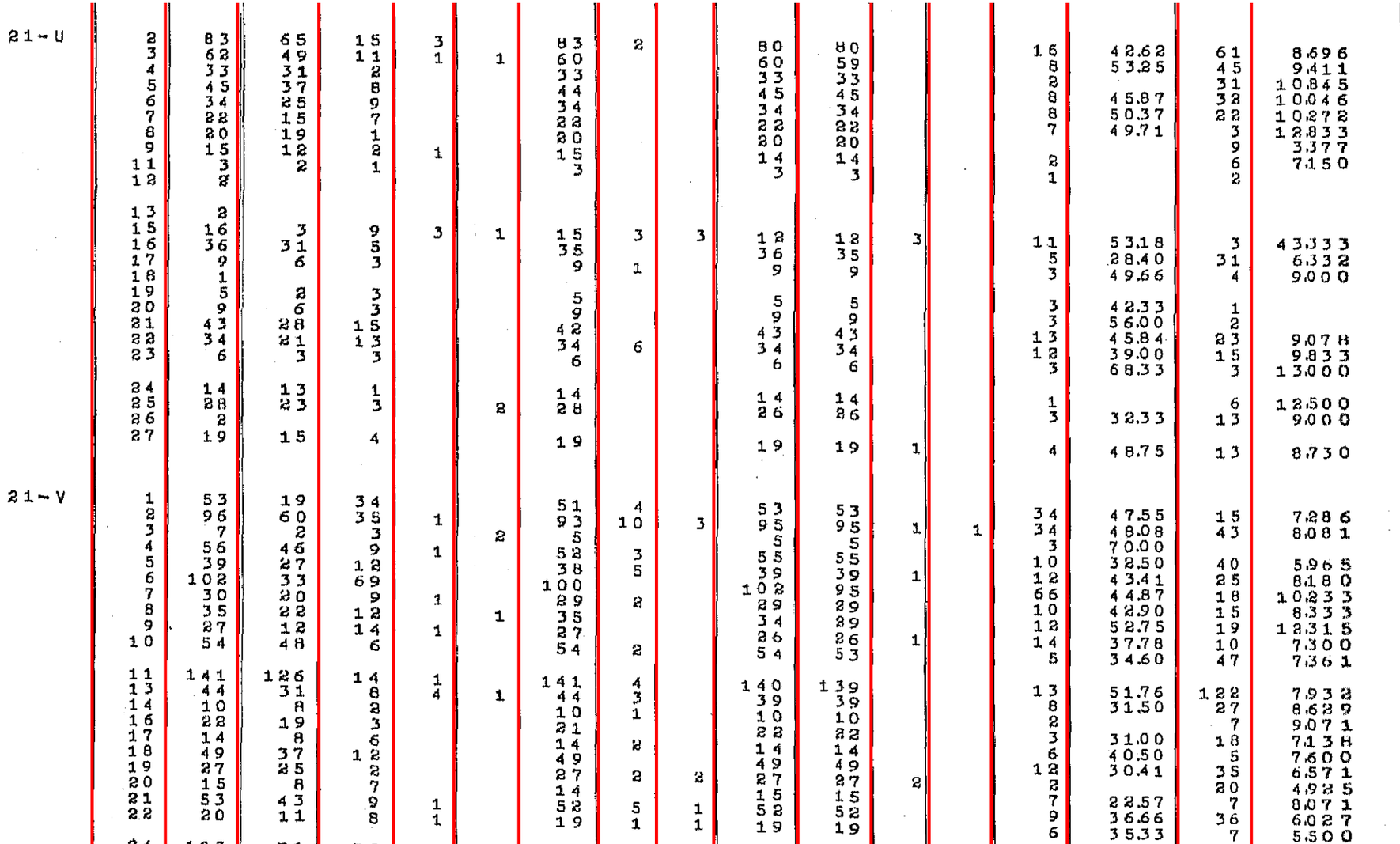
Isolate Columns

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



Isolate Columns

Columns are isolated



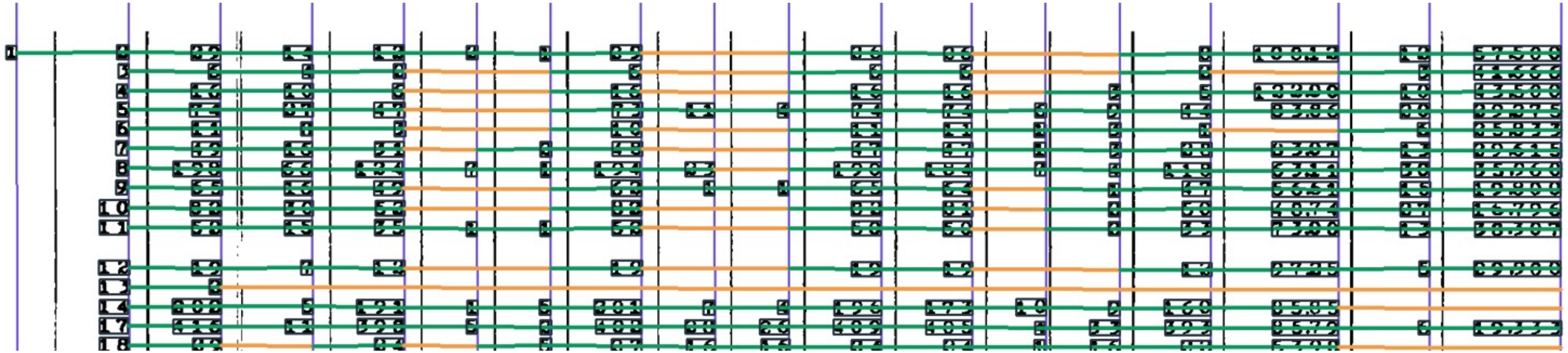
Tesseract Each Row in Each Column

- 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

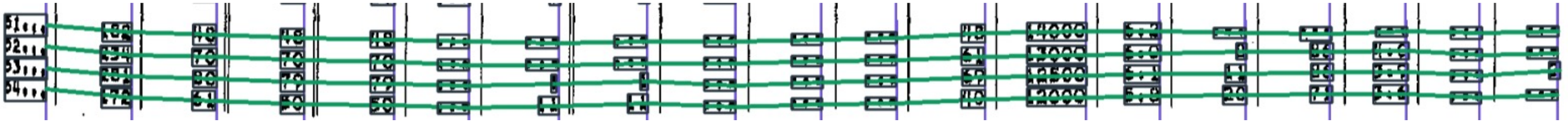
Use Table's Internal Structure to Build Rows and Columns

- 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

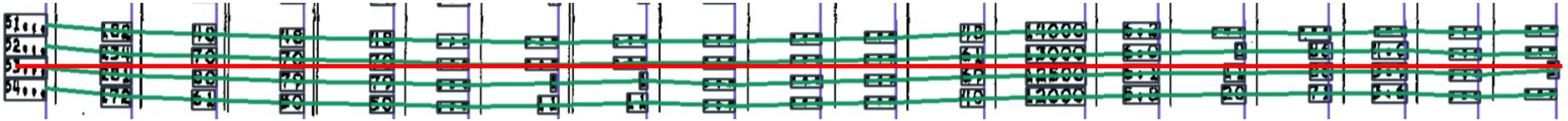
Use Table’s Internal Structure to Build Rows and Columns



Use Table's Internal Structure to Build Rows and Columns



Use Table's Internal Structure to Build Rows and Columns

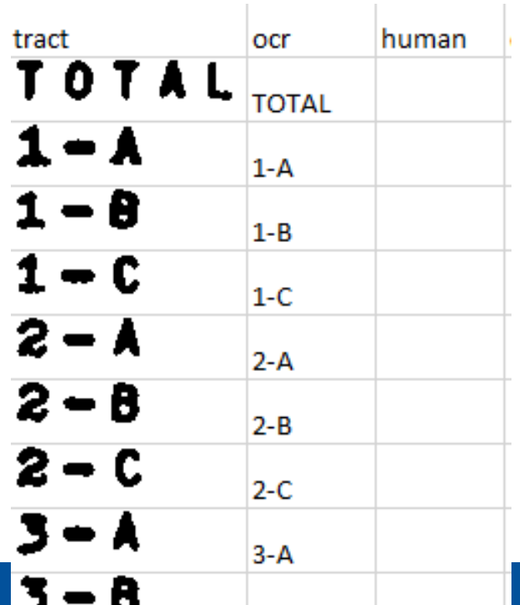


Train and Apply Custom Model

- 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!

FEDERAL RESERVE BANK
PHILADELPHIA