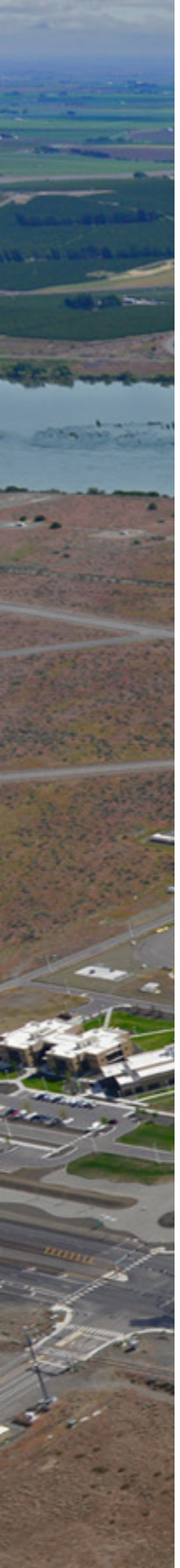


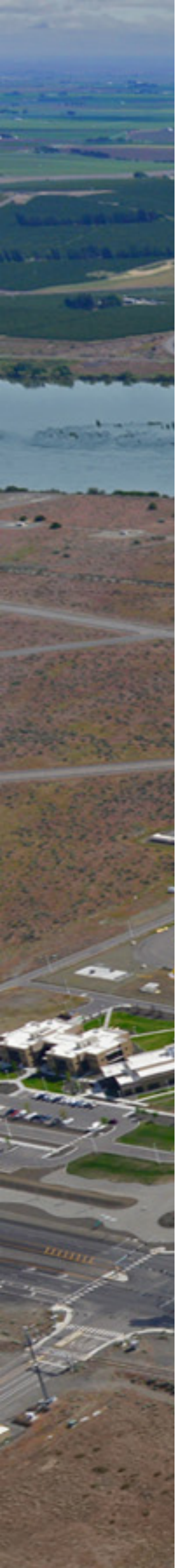
Pacific Northwest
NATIONAL LABORATORY

San Jose Microsoft Cross Reference Analysis

Kevin Keene

U.S. DEPARTMENT OF ENERGY BATTELLE

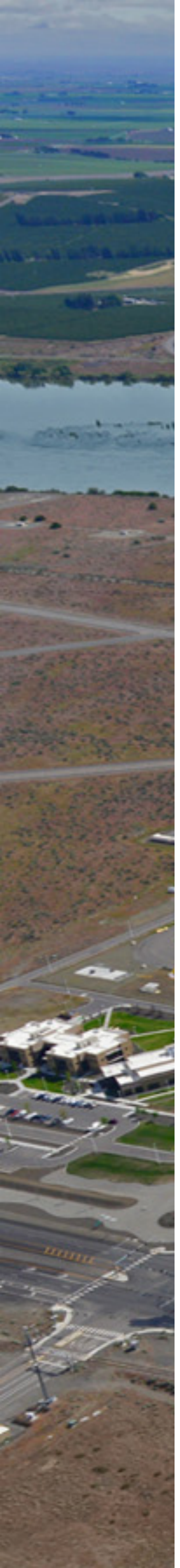
PNNL is operated by Battelle for the U.S. Department of Energy



Pacific Northwest
NATIONAL LABORATORY

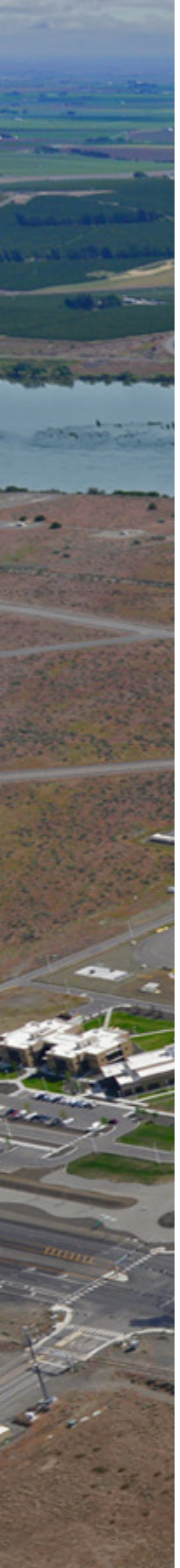
Objectives

1. **Compare Microsoft footprints to SJ footprints in GIS to see how similar the two datasets are**
2. **Investigate UBID one-to-one matching between two building footprint datasets (MS and SJ) and compare to GIS matching**
3. Investigate accuracy of UBID cross referencing for matching address UBID0 to polygon UBIDs
4. Investigate accuracy of UBID cross referencing for buildings to parcels many-to-many matching



Dataset Background

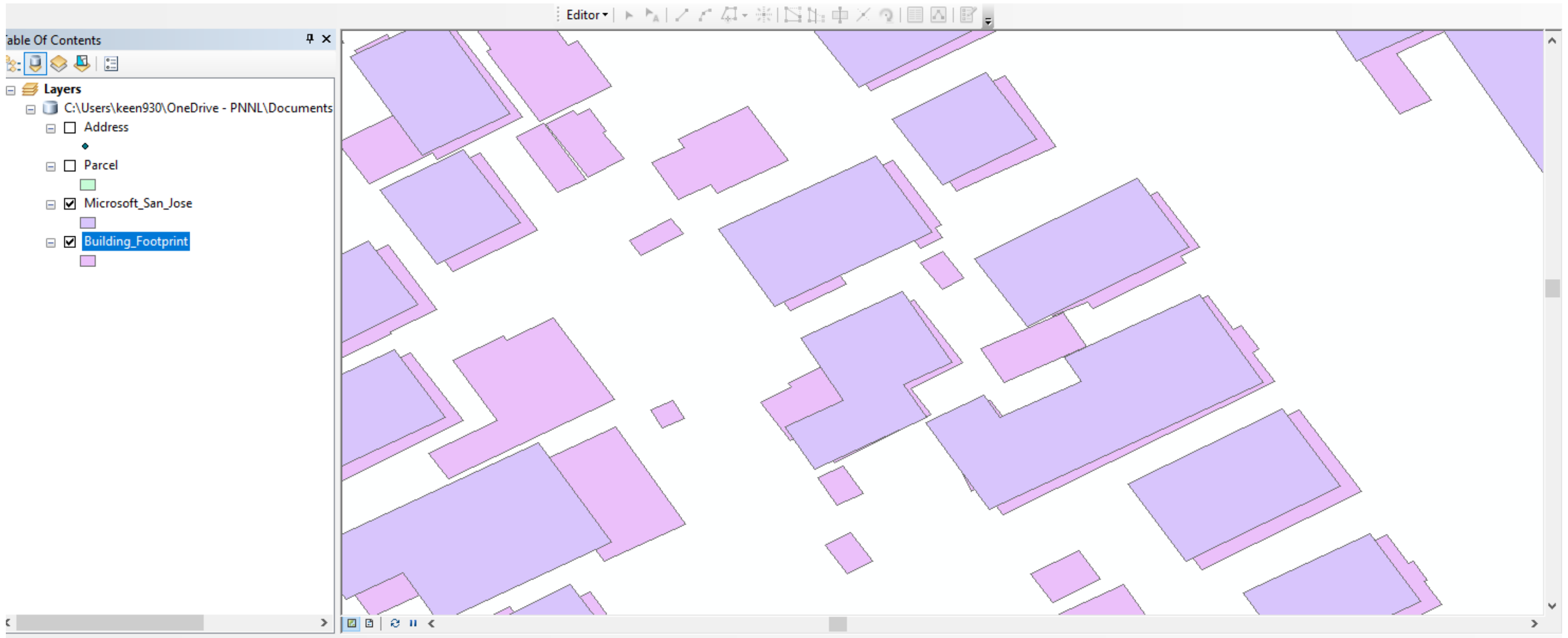
- San Jose Footprints
  - 2006 satellite data
- Microsoft Footprints
  - Nation-wide open source building footprints from satellite data with geometric screening algorithms
  - From 2017
  - https://www.arcgis.com/home/item.html?id=f40326b0dea54330ae39584012807126
  - https://github.com/Microsoft/USBuildingFootprints



Pacific Northwest
NATIONAL LABORATORY

▨ Microsoft
▨ San Jose

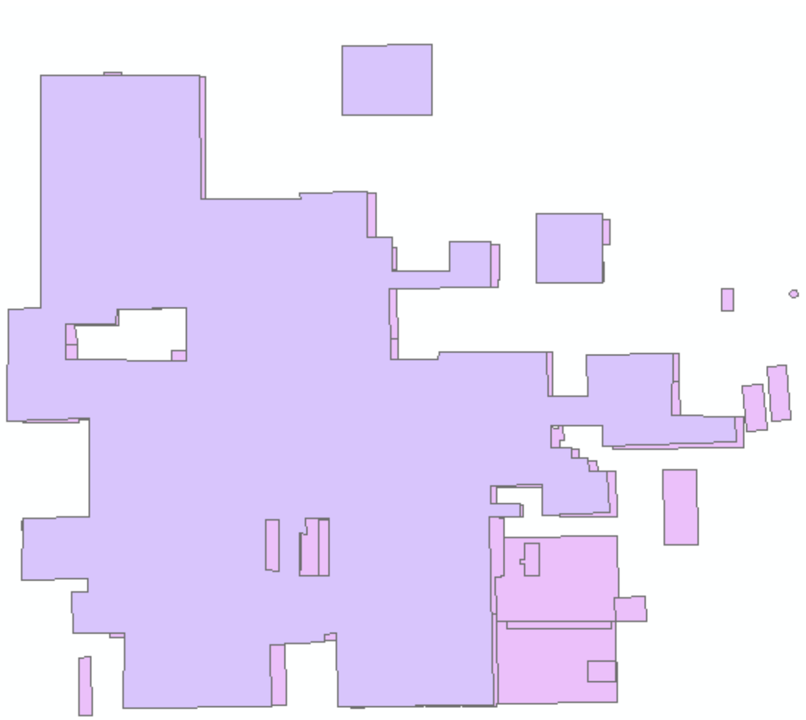
Comparing Datasets: Microsoft vs. San Jose

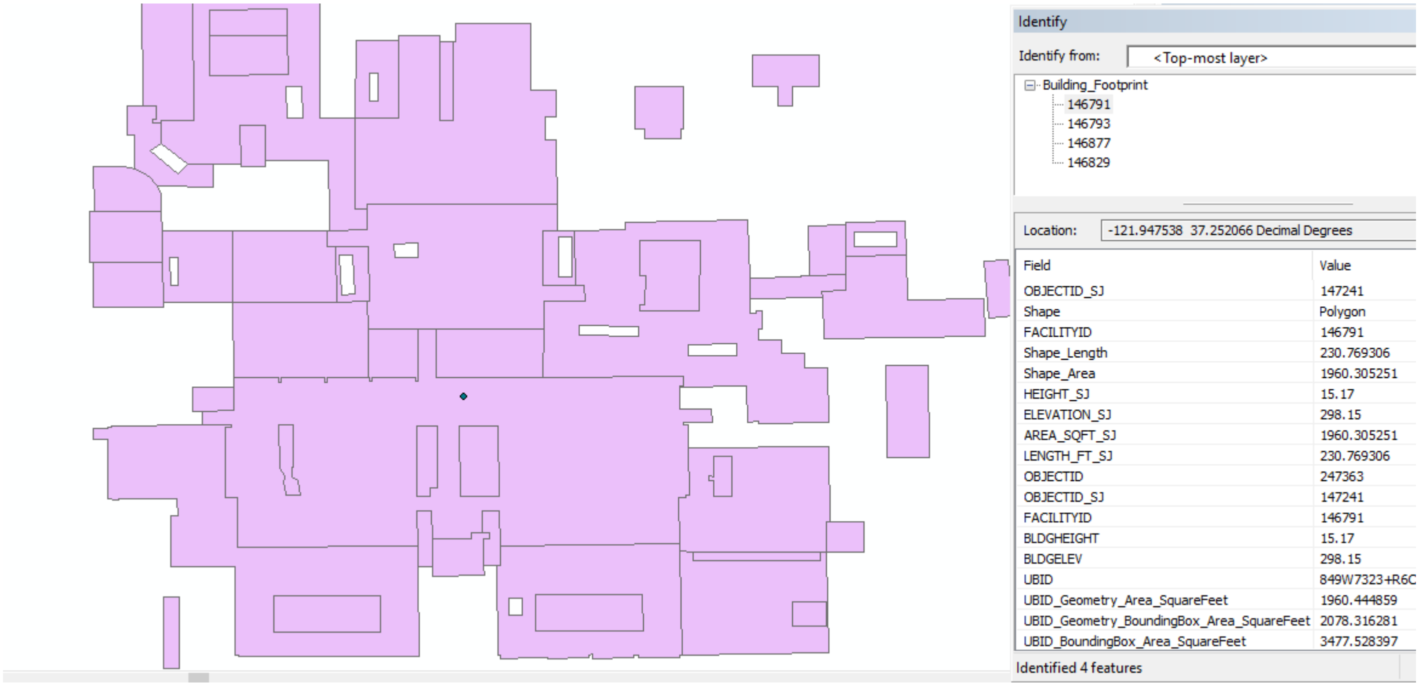


Only includes intersection with IoU > 0.05 (5%) to ignore slight overlaps

	San Jose	Microsoft	Note
A. Buildings with 0 intersections	101,550	5,495	• SJ has many small bldgs MS doesn’t have that appear to be small structures or sheds • Both datasets have some legit buildings the other doesn’t
B. Buildings with one-to-one	208,608	208,608	• One-to-one matches are more likely to be equivalent buildings
C. Buildings with one-to-many or many-to-one	14,059	27,330	• See next slide for examples
Total	**324,217**	**241,433**	

SJ dataset subdivides some buildings based on height differential





Identify

Identify from: <Top-most layer>

Building_Footprint
- 146791
- 146793
- 146877
- 146829

Location: -121.947538 37.252066 Decimal Degrees

Field	Value
OBJECTID_SJ	147241
Shape	Polygon
FACILITYID	146791
Shape_Length	230.769306
Shape_Area	1960.305251
HEIGHT_SJ	15.17
ELEVATION_SJ	298.15
AREA_SQFT_SJ	1960.305251
LENGTH_FT_SJ	230.769306
OBJECTID	247363
OBJECTID_SJ	147241
FACILITYID	146791
BLDGHEIGHT	15.17
BLDGELEV	298.15
UBID	849W7323+R6C
UBID_Geometry_Area_SquareFeet	1960.444859
UBID_Geometry_BoundingBox_Area_SquareFeet	2078.316281
UBID_BoundingBox_Area_SquareFeet	3477.528397

Identified 4 features

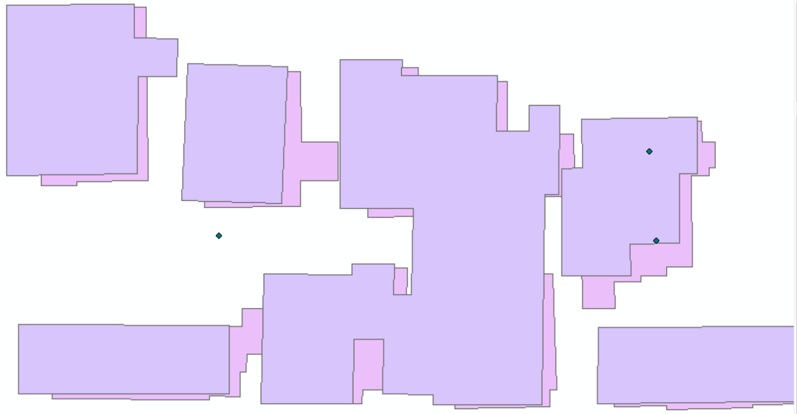
Adjacent polygons follow changes in roof heights, whereas MS has one footprint to represent entire facility

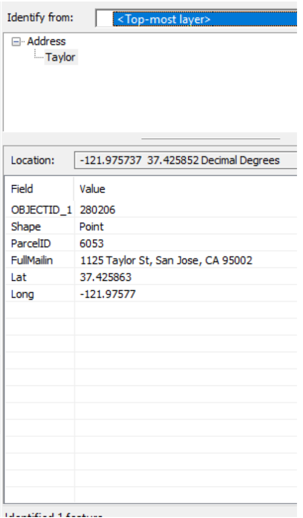
Could dissolve adjacent footprints on same parcel– will cause some problems but may solve more

▨ Microsoft

▨ San Jose

Other incongruencies between SJ and MS





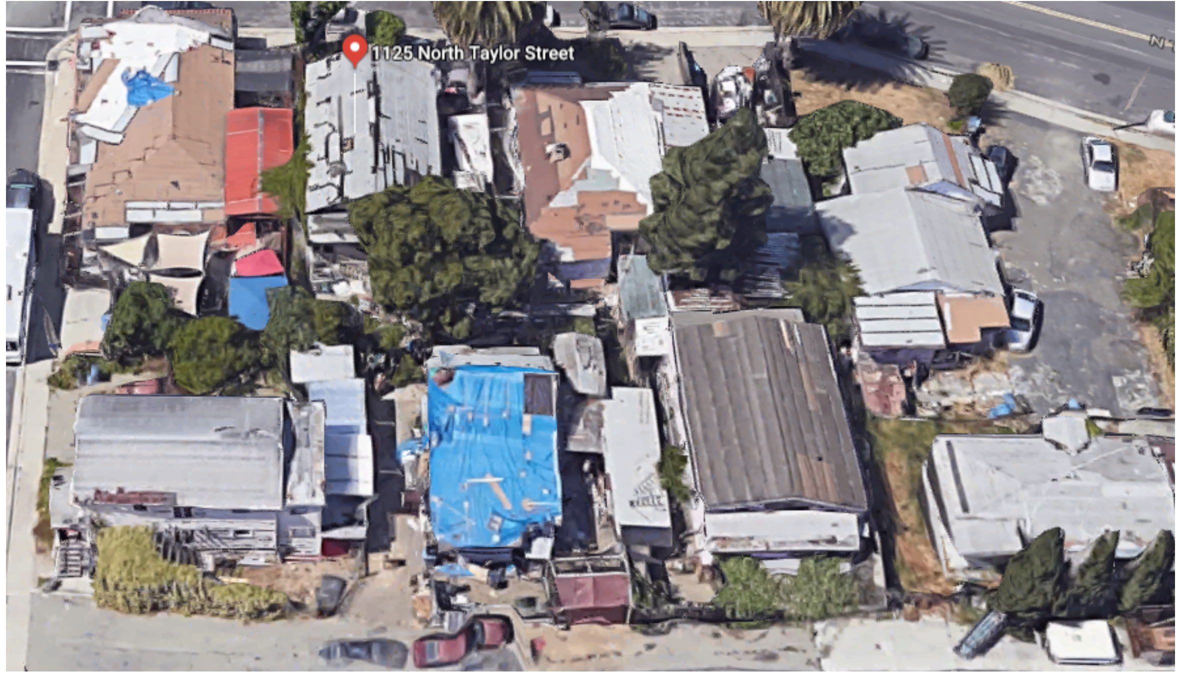
Identify from: <Top-most layer>

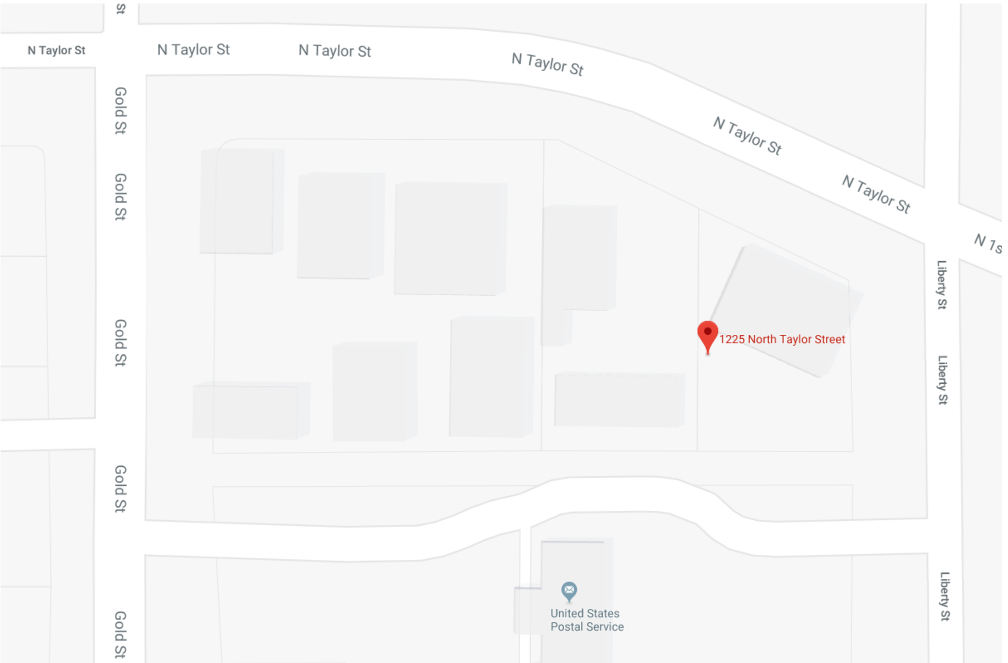
- Address
  - Taylor

Location: -121.975737 37.425852 Decimal Degrees

Field	Value
OBJECTID_1	280206
Shape	Point
ParcelID	6053
FullMailin	1125 Taylor St, San Jose, CA 95002
Lat	37.425863
Long	-121.97577

Identified 1 feature





▨ Microsoft

▨ San Jose

In some cases, the datasets will approximate the building outline and merge nearby buildings, making it difficult to discern if the two datasets are equivalent and if they reflect reality

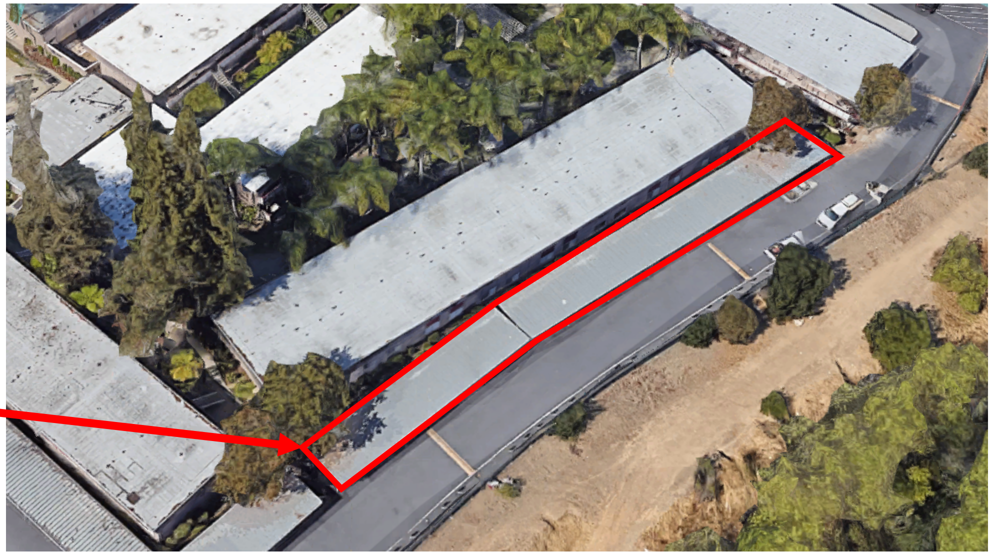
Cleaning and Filtering Datasets

1. Delete buildings with large area increases between the geometry and geometry bounding box

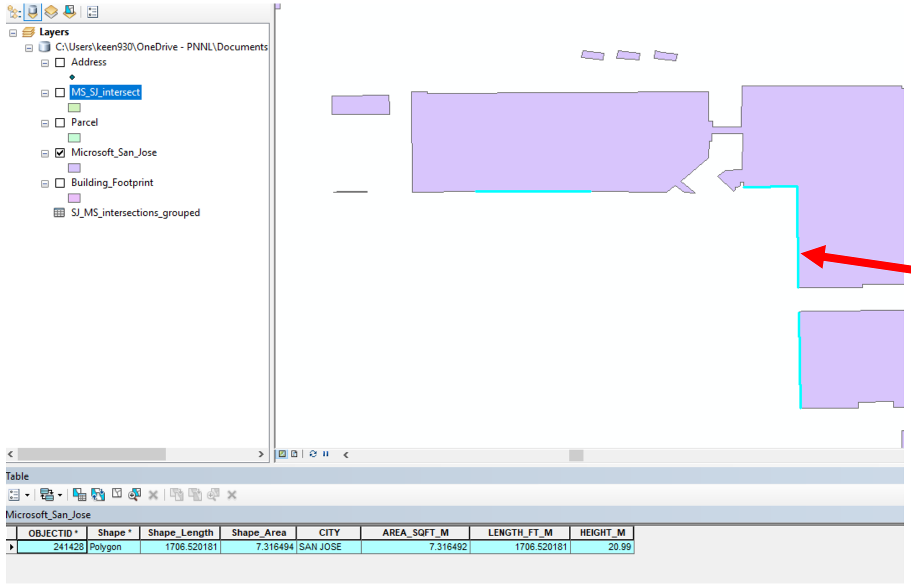
BB Increase	Microsoft	San Jose
> 2500%	45	4
> 2000%	50	8
> 1500%	65	11
> 1000%	**83**	**26**
These buildings are deleted		

2. Only include buildings with footprint_area * height / 4m > 50,000 (covered buildings)

Geometry Area	Microsoft	San Jose
> 200,000 sqft	559	334
> 150,000 sqft	791	553
> 100,000 sqft	1,247	984
> 50,000 sqft	**3,068**	**1,992**
These buildings are kept		



% Increase >1000% (Parking cover)



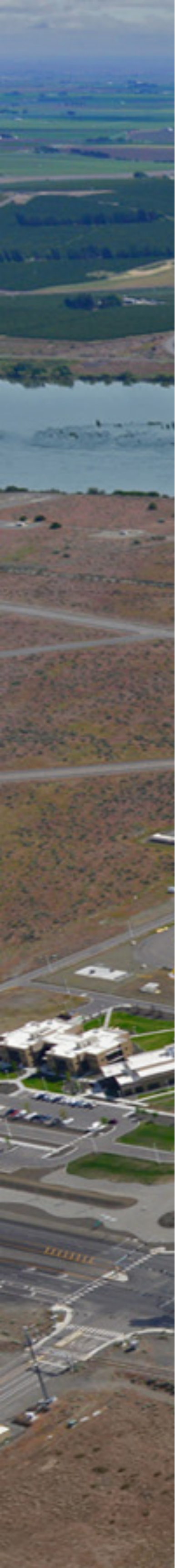
Layers
C:\Users\keen930\OneDrive - PNNL\Documents
□ Address
□ MS_SJ_intersect
□ Parcel
☑ Microsoft_San_Jose
□ Building_Footprint
SJ_MS_intersections_grouped

Table

Microsoft_San_Jose

OBJECTID *	Shape *	Shape_Length	Shape_Area	CITY	AREA_SQFT_M	LENGTH_FT_M	HEIGHT_M
241428	Polygon	1706.520181	7.316494	SAN JOSE	7.316492	1706.520181	20.99

MS has some erroneous shapes



Pacific Northwest
NATIONAL LABORATORY

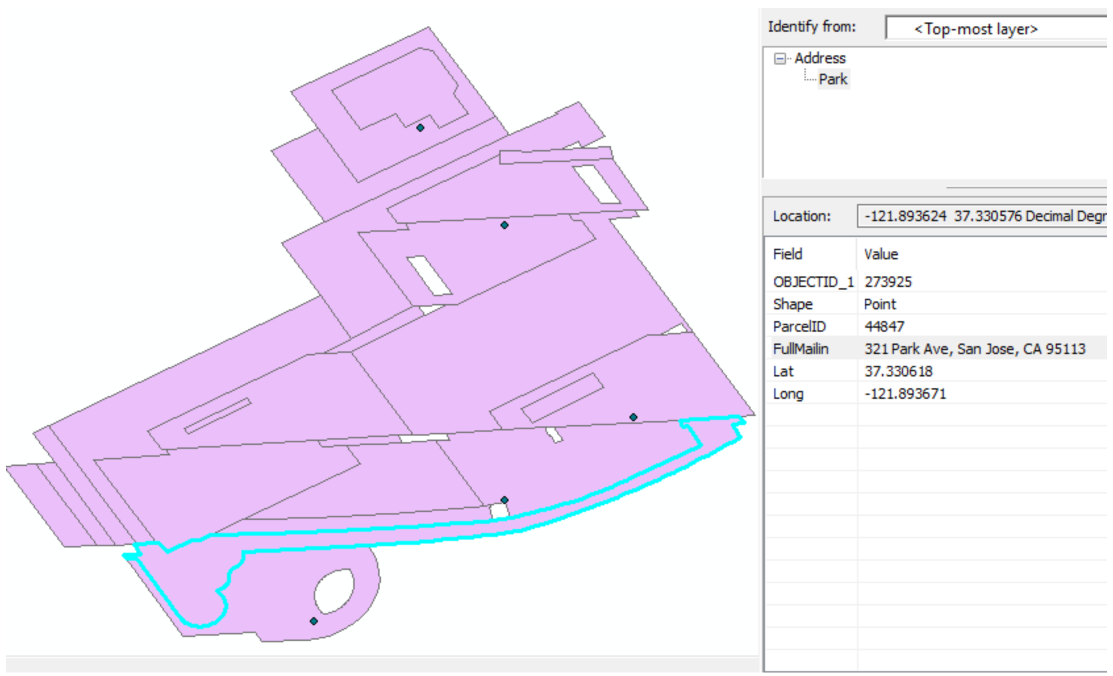
Cleaning and Filtering Datasets

Only includes intersection with IoU > 0.05 (5%)

	San Jose		Microsoft		Note
	Raw	**Clean/Filter***	**Raw**	**Clean/Filter***	
A. Buildings with 0 intersections	101,550	**264**	5,495	**1,362**	• Since SJ footprints are subdivided more, many don’t make 50k sqft cut, leaving many MS footprints without intersections
B. Buildings with one-to-one	208,608	**1,466**	208,608	**1,466**	
C. Buildings with one-to-many or many-to-one	14,059	**262**	27,330	**240**	• Many of these are the buildings that SJ subdivides that are still large enough to be over 50k sqft • Others are overlapping neighboring buildings
Total	324,217	**1,992**	241,433	**3,068**	

*Clean/Filter only includes buildings 50k sqft or more

Difficulties estimating floor area

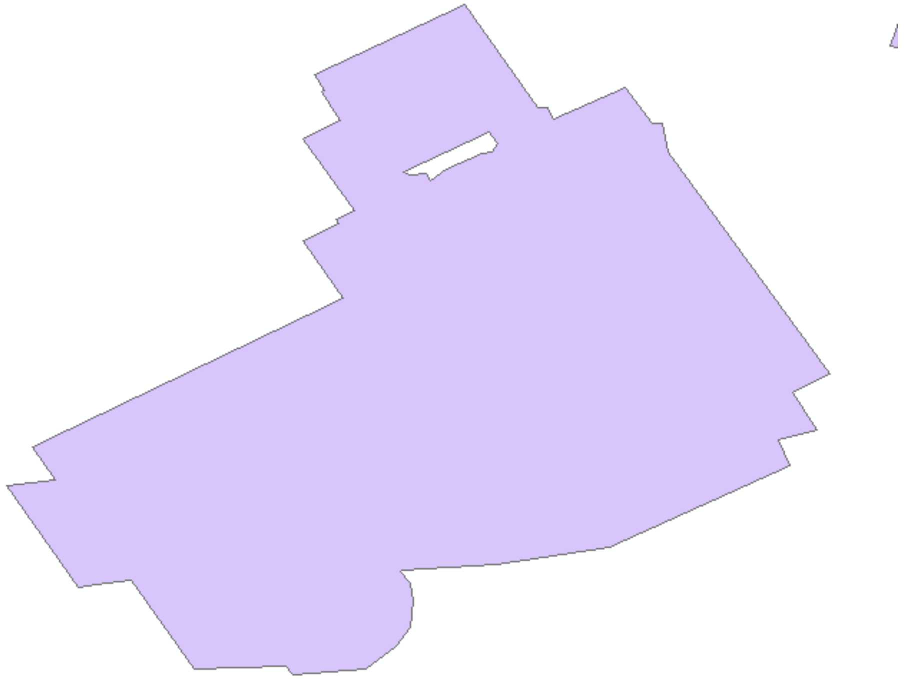


Identify from: <Top-most layer>

- Address
  - Park

Location: -121.893624 37.330576 Decimal Degr

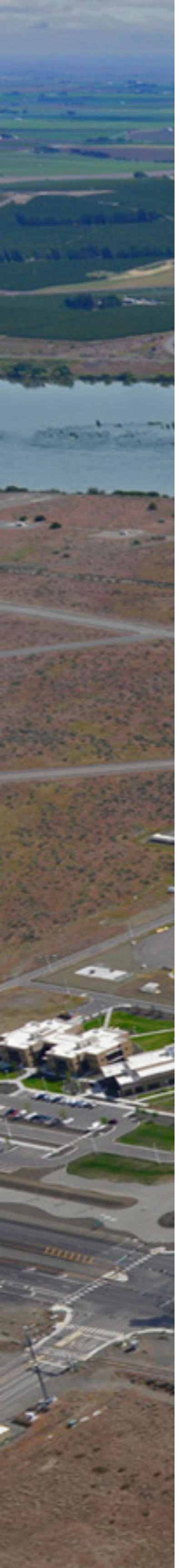
Field	Value
OBJECTID_1	273925
Shape	Point
ParcelID	44847
FullMailin	321 Park Ave, San Jose, CA 95113
Lat	37.330618
Long	-121.893671



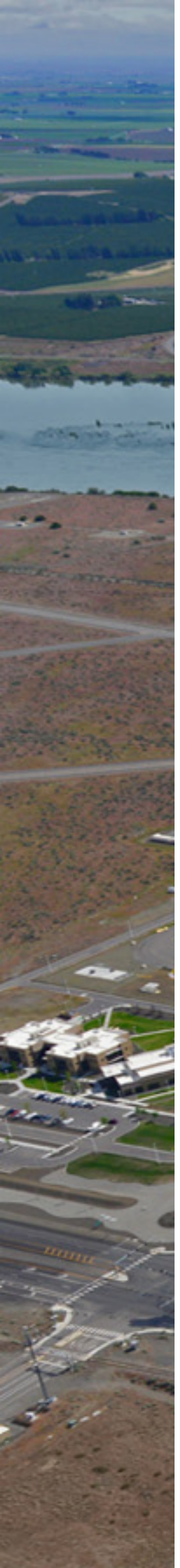
- When one building is subdivided and another isn't, it throws off the floor area approximations
- The two buildings won't match even if some of the subdivided buildings are greater than 50k sqft because the overlaps are all too small

🟪 Microsoft

🟪 San Jose



GIS Matching/Cross Reference



Pacific Northwest
NATIONAL LABORATORY

ArcMap Matching

GOAL:

- Create the “ground truth” (or as best possible) of what buildings are considered “equivalent” by setting an intersection threshold

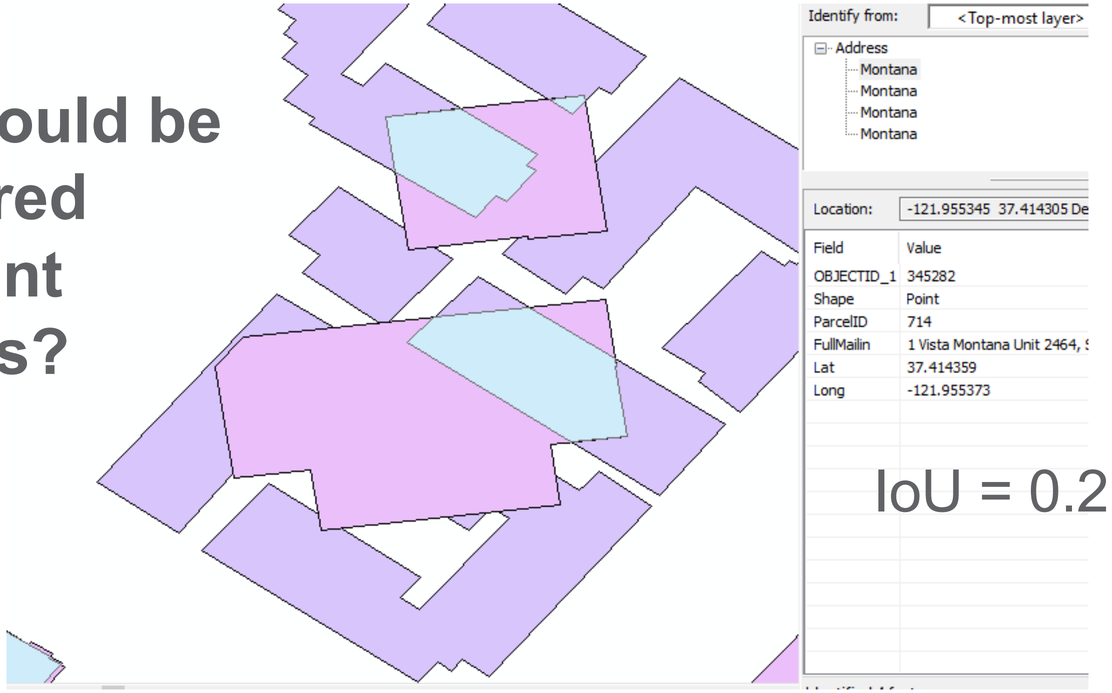
Process

1. Use INTERSECT tool to find all intersections between two polygons
2. Calculate area of intersection polygons
3. Calculate Intersection over Union (IoU)
    - Intersect Area / (Footprint Area 1 + Footprint Area 2 – Intersect Area)
4. Group by matches with same ID and delete multiple matches to keep one-to-one match (with highest intersect)
5. Only keep over certain threshold of IoU (see next slide)

IoU Threshold	Number of Intersects	IoU Threshold	Number of Intersects
0	1,614	**0.5**	1,475
0.1	1,610	**0.6**	1,399
0.2	1,588	**0.7**	1,348
0.3	1,556	**0.8**	1,265
0.4	1,522	**0.9**	898

Investigating IoU GIS Threshold

What should be considered equivalent buildings?



Identify from: <Top-most layer>

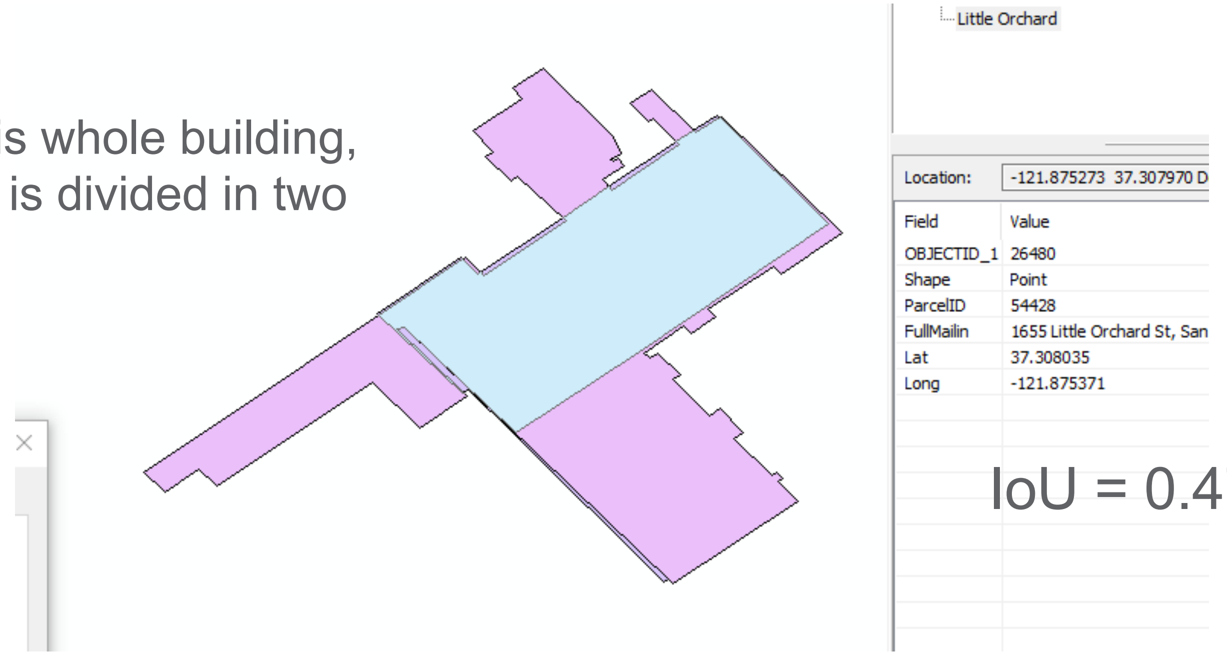
⊟ Address
- Montana
- Montana
- Montana
- Montana

Location: -121.955345 37.414305 De

Field	Value
OBJECTID_1	345282
Shape	Point
ParcelID	714
FullMailin	1 Vista Montana Unit 2464,
Lat	37.414359
Long	-121.955373

IoU = 0.21

SJ is whole building, MS is divided in two

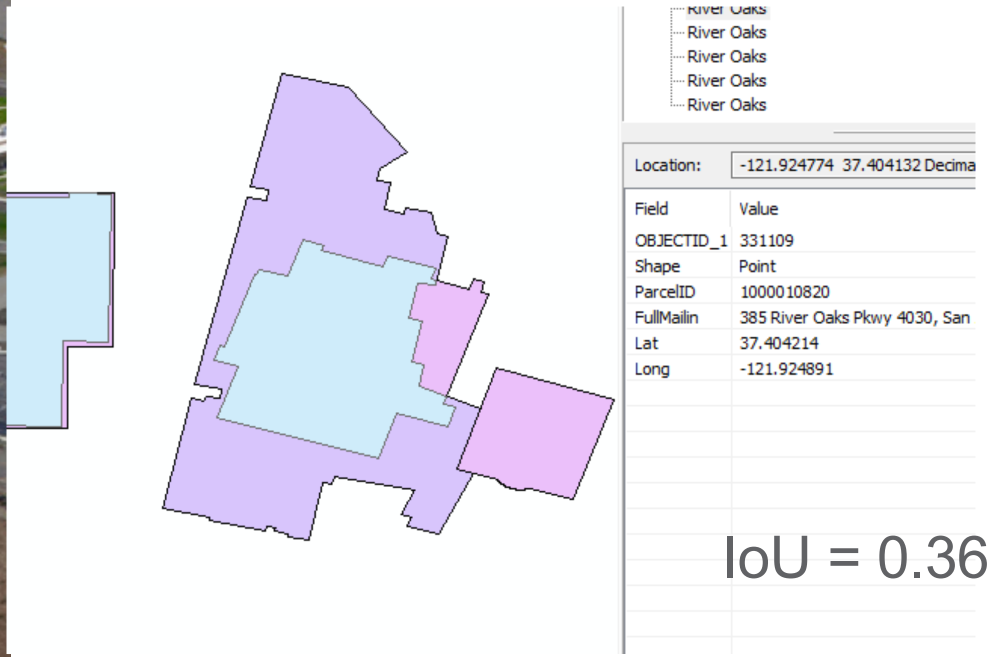


Little Orchard

Location: -121.875273 37.307970 D

Field	Value
OBJECTID_1	26480
Shape	Point
ParcelID	54428
FullMailin	1655 Little Orchard St, San
Lat	37.308035
Long	-121.875371

IoU = 0.47

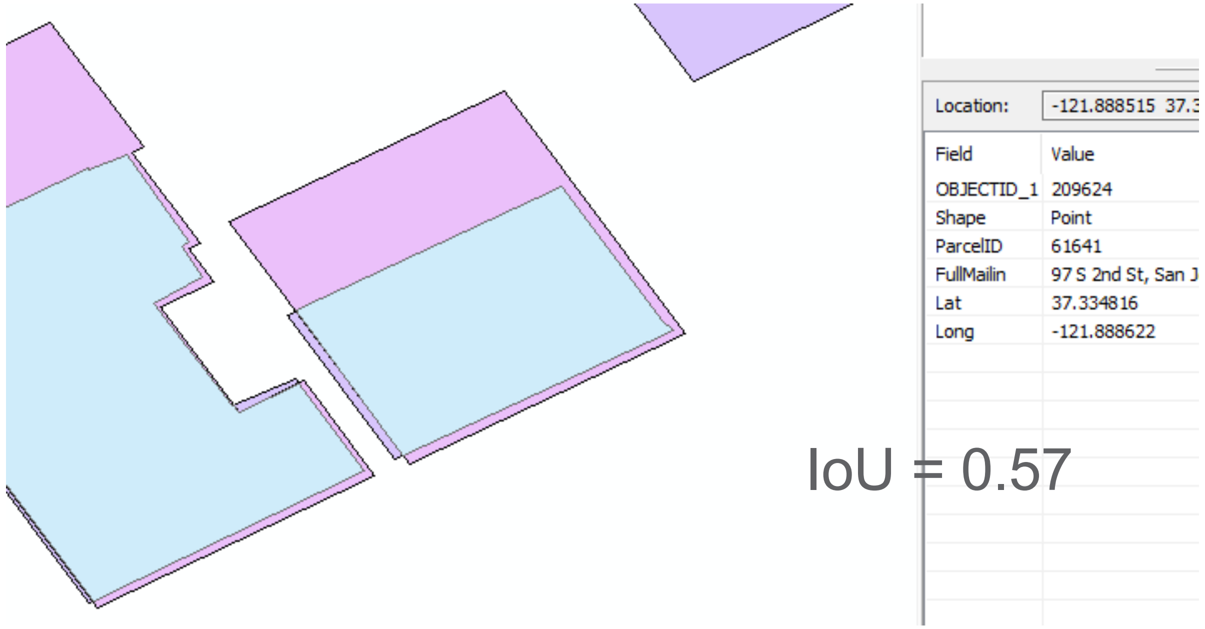


- River Oaks
- River Oaks
- River Oaks
- River Oaks
- River Oaks

Location: -121.924774 37.404132 Decima

Field	Value
OBJECTID_1	331109
Shape	Point
ParcelID	1000010820
FullMailin	385 River Oaks Pkwy 4030, San
Lat	37.404214
Long	-121.924891

IoU = 0.36



Location: -121.888515 37.3

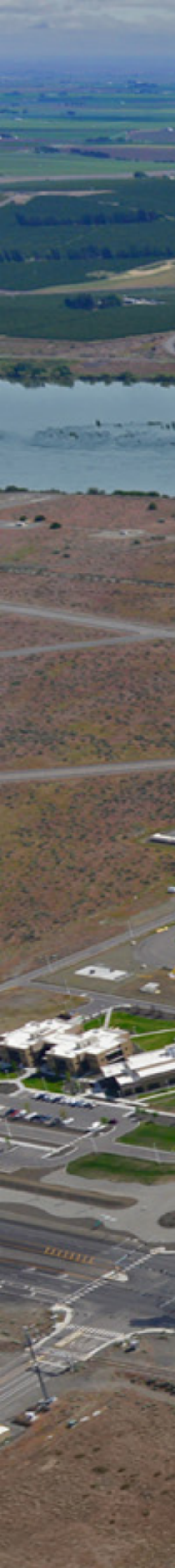
Field	Value
OBJECTID_1	209624
Shape	Point
ParcelID	61641
FullMailin	97 S 2nd St, San J
Lat	37.334816
Long	-121.888622

IoU = 0.57

🟪 Microsoft

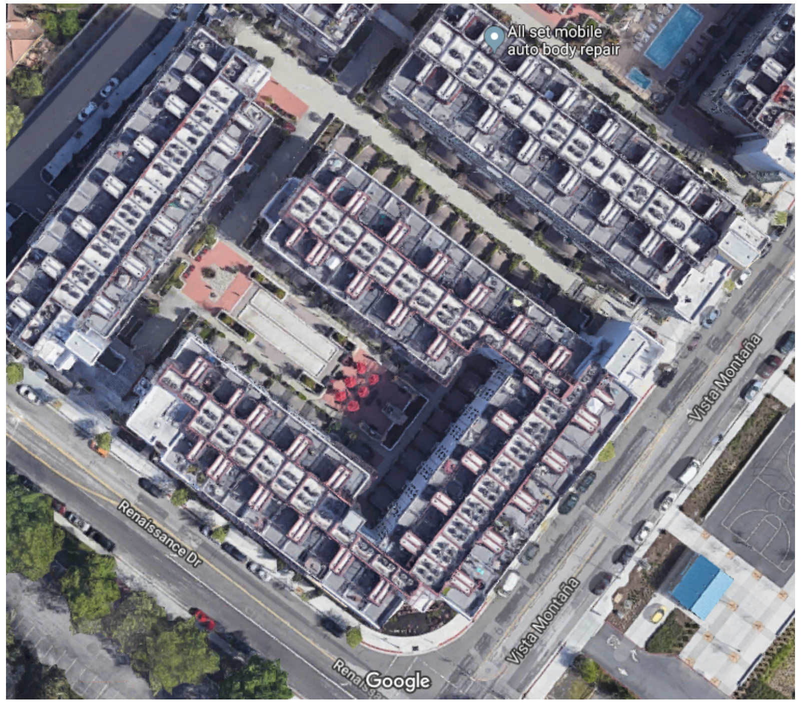
🟪 San Jose

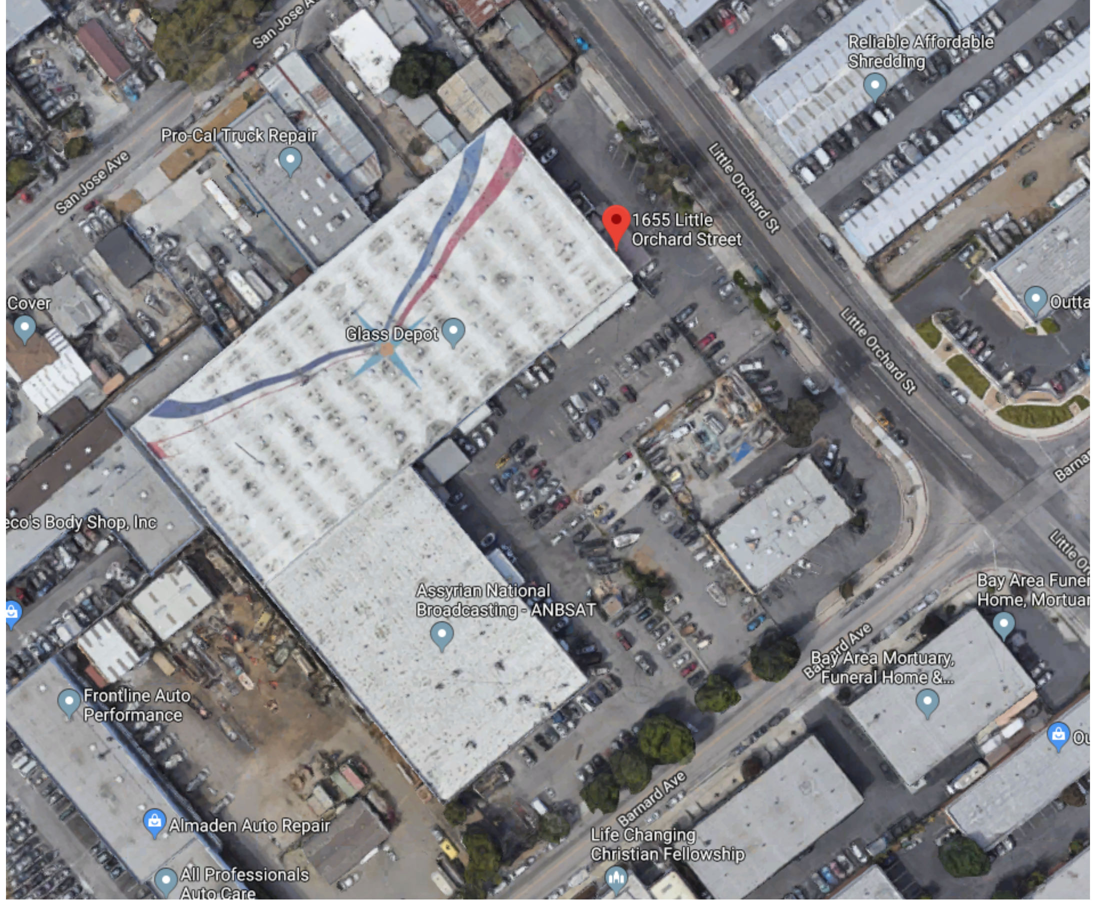
🟦 Intersection

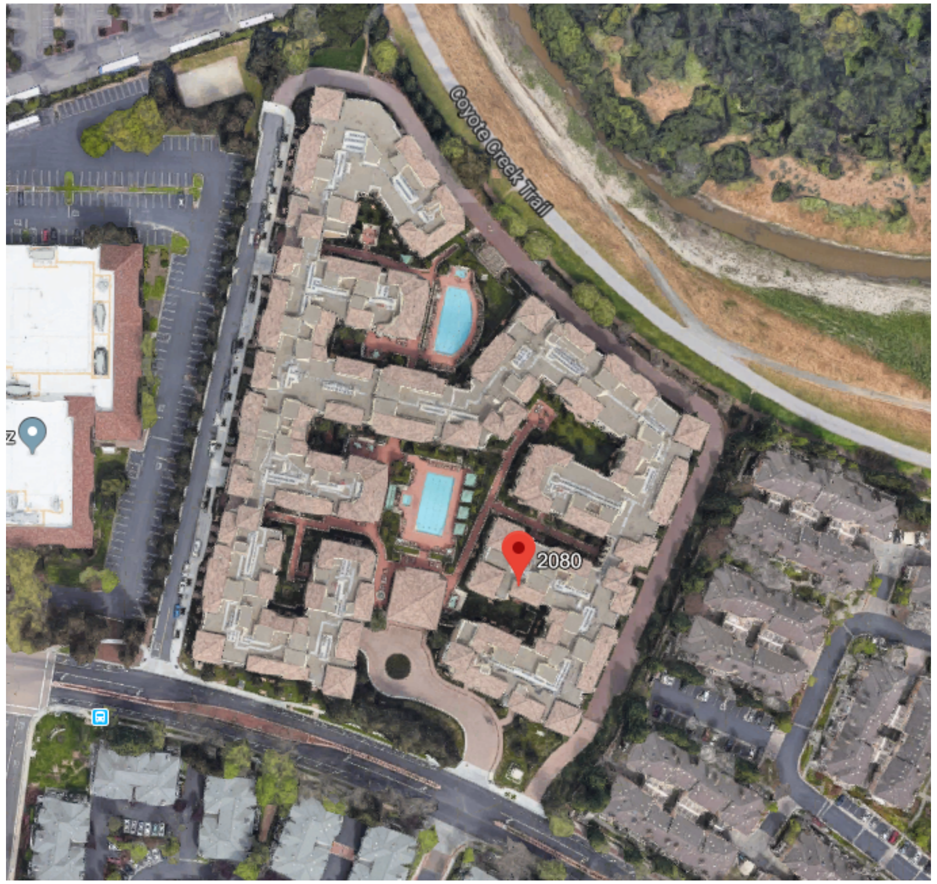


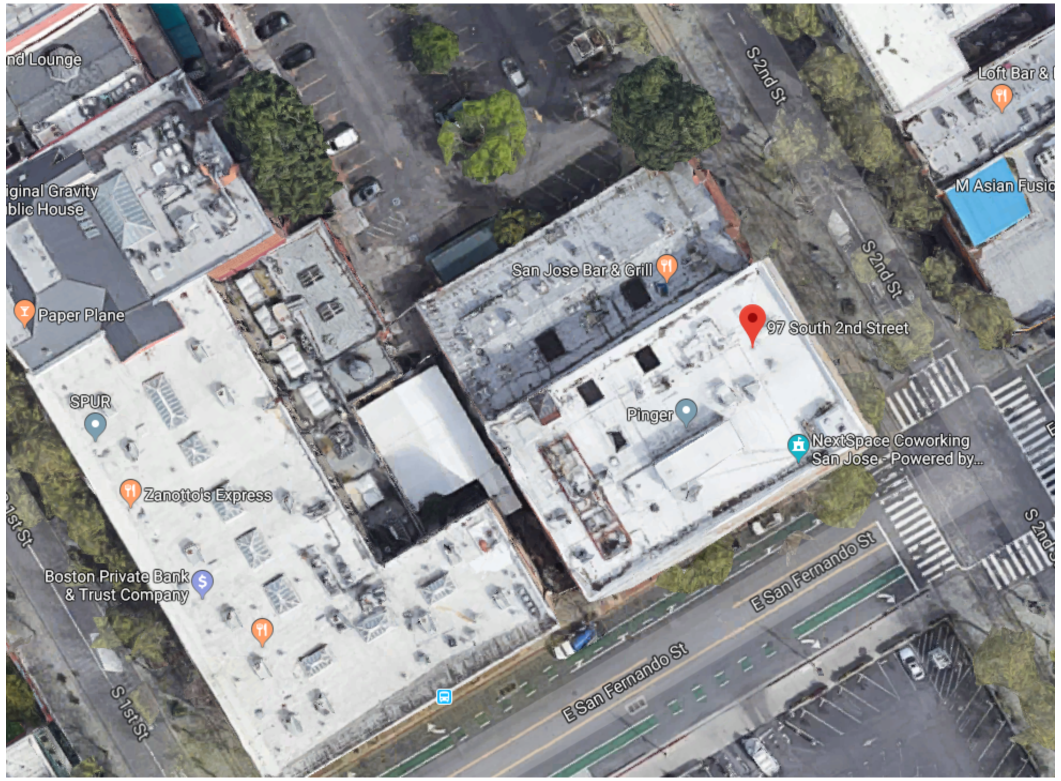
Pacific Northwest
NATIONAL LABORATORY

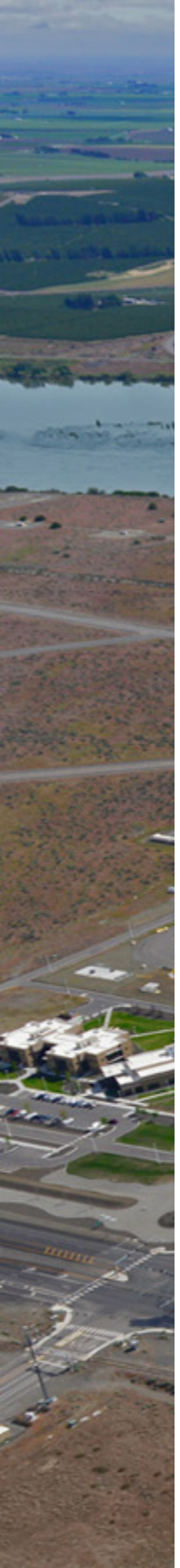
Investigating IoU GIS Threshold




San Jose Ave
Reliable Affordable Shredding
Pro-Cal Truck Repair
San Jose Ave
1655 Little Orchard Street
Little Orchard St
Cover
Glass Depot
Outta
Little Orchard St
eco's Body Shop, Inc
Barnard
Assyrian National Broadcasting - ANBSAT
Bay Area Funeral Home, Mortuar
Bay Area Mortuary, Funeral Home &...
Frontline Auto Performance
Barnard Ave
Almaden Auto Repair
Life Changing Christian Fellowship
All Professionals Auto Care



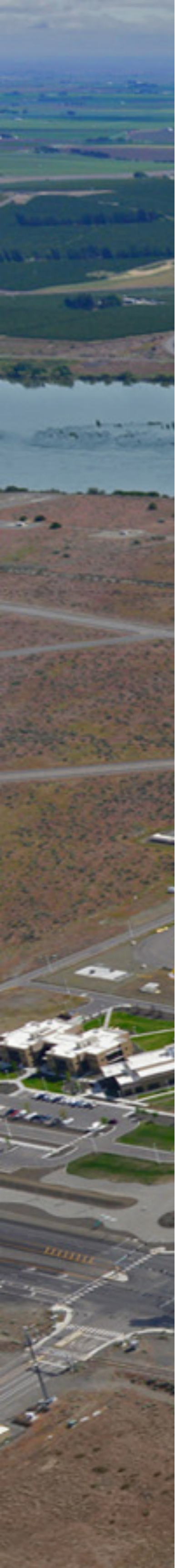




Pacific Northwest
NATIONAL LABORATORY

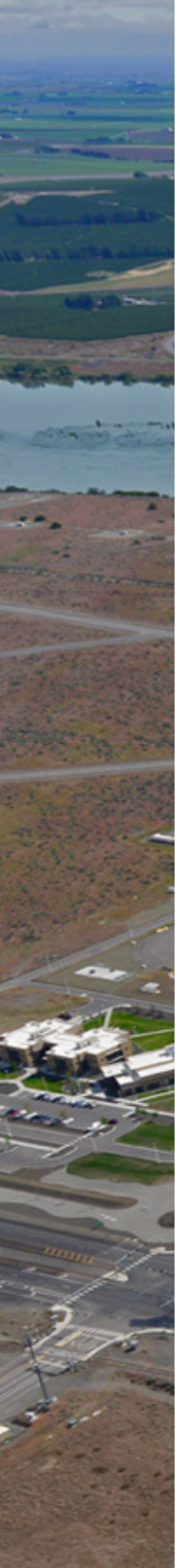
Threshold of 0.5 IoU seems most appropriate

- Number of buildings matched = 1,475
  - Microsoft
    - ✓ Over threshold (>0.5): 1,475
    - ✓ Under threshold (0-0.5) or discarded (multiple) match: 231
    - ✓ No intersection with SJ: 1,362
  - San Jose
    - ✓ Over threshold (>0.5): 1,475
    - ✓ Under threshold (0-0.5) or discarded (multiple) match: 253
    - ✓ No intersection with MS: 264
  - **Conclusion**: 86% of buildings 50,000 sqft or more with intersections were matched (IoU of > 0.5) and 58% of all buildings 50,000 sqft or more were matched
    - ✓ 86%/58% is indicator (sort of) of how similar the 50k sqft + buildings are
    - ✓ Better comparison would be with entire city



Pacific Northwest NATIONAL LABORATORY

UBID (Bounding Box) Matching/Cross Reference

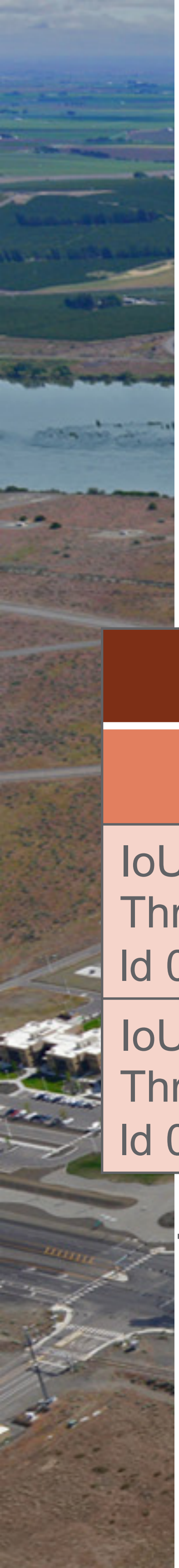


UBID Cross Reference

- This slide (step 1): Filter intersections with **bounding box IoU threshold** to maximize resemblance to GIS matches (previous slide)
- On the next slide (step 2): there could still be many-to-many matches, so **group duplicate matches and only keep best match**

UBID IoU Threshold	Total matches found	Same matches as GIS	Matches missing from GIS	Extra matches not in GIS	Success Rate
0.0	3,600	1,475	0	2,125	58.1%
0.1	2,093	1,475	0	618	82.7%
0.2	1,760	1,475	0	285	91.2%
0.3	1,636	1,474	1	162	94.8%
0.4	**1,555**	**1,470**	**5**	**85**	**97.0%**
0.5	**1,484**	**1,449**	**26**	**35**	**97.9%**
0.6	1,438	1,421	54	17	97.6%
0.7	1,402	1,390	85	12	96.6%
0.8	1,359	1,354	121	5	95.6%
0.9	1,199	1,198	277	1	89.6%

- Want to reduce “Missing” (matches in GIS but not UBID) primarily because this number can’t be reduced in step 2
- “Extra” (matches in UBID but not GIS) could be one building gets matched to two buildings, and in the next step the extra match is removed and the correct match is kept
    - Reducing the extra matches has some importance, because they could be false matches that have no correct match



Pacific Northwest
NATIONAL LABORATORY

- 3 metrics used to find best match (if multiple) – area intersect percentage, distance between centroids, and IoU
- Point isn't to show UBID is better than GIS or vice versa, but to show if UBID is able to produce similar results – there is no "correct" results so we can't know which is better

UBID Cross Reference with Grouping

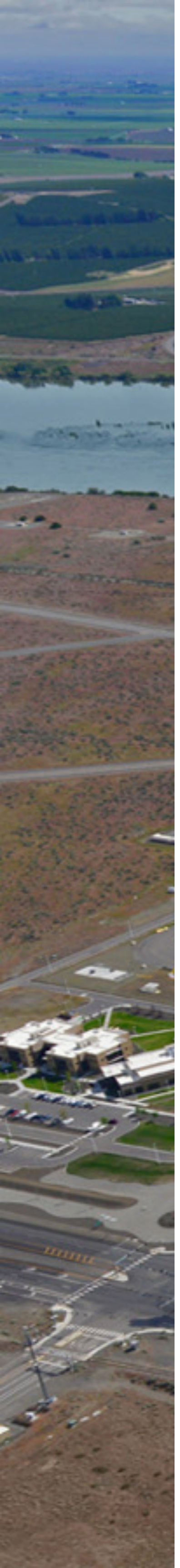
	No Grouping				
	Total	Same	Extra	Missing	Success
IoU Threshold 0.4	1,555	1,470	85	5	97.0%
IoU Threshold 0.5	1,484	1,449	35	26	97.9%

→ Success rate does not increase much with grouping or vary much between different grouping metrics – if using bigger or less similar datasets then the differences would be pronounced

If number of extra matches doesn't decrease significantly with grouping, meaning there is a lot of false matching

Missing matches can't decrease with grouping – if increases it means increase then wrong match was selected

	Area Intersect Percent				
	Total	Same	Extra	Missing	Success
IoU Threshold 0.4	1,526	1,467	8	59	97.8%
IoU Threshold 0.5	1,477	1,447	28	30	98.0%
	Centroid Distance				
	Total	Same	Extra	Missing	Success
IoU Threshold 0.4	1,526	1,468	7	58	97.8%
IoU Threshold 0.5	1,477	1,448	27	29	98.1%
	Intersect over Union (IoU)				
	Total	Same	Extra	Missing	Success
IoU Threshold 0.4	1,526	1,468	7	58	97.8%
IoU Threshold 0.5	1,477	1,447	28	30	98.0%



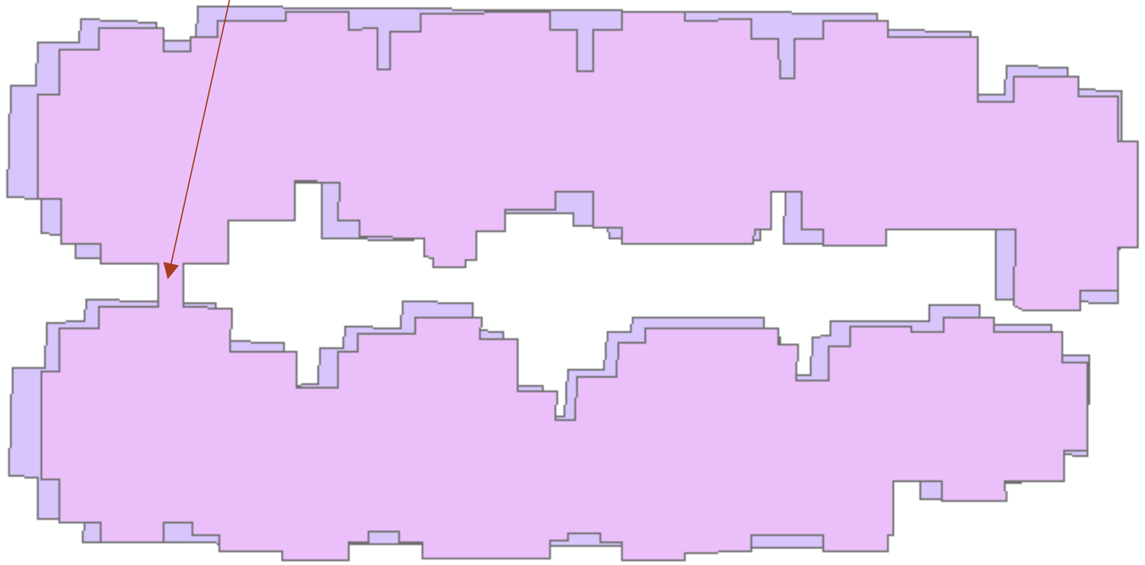
🟪 Microsoft

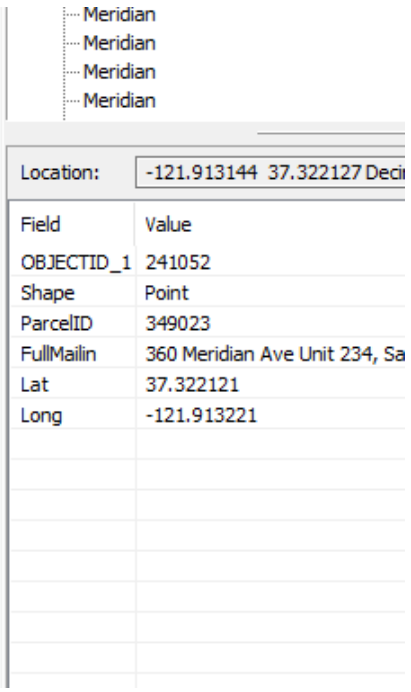
🟪 San Jose

Example: Extra Match (not in GIS Cross Reference)

Likely on borderline of GIS and UBID thresholds – satisfies one but not the other

Two MS footprints, and one connected SJ footprint

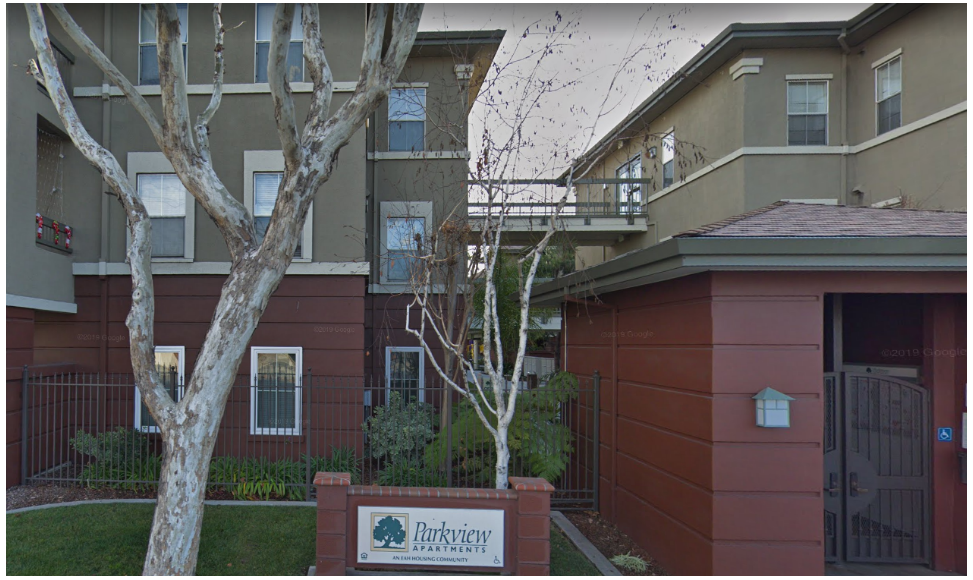


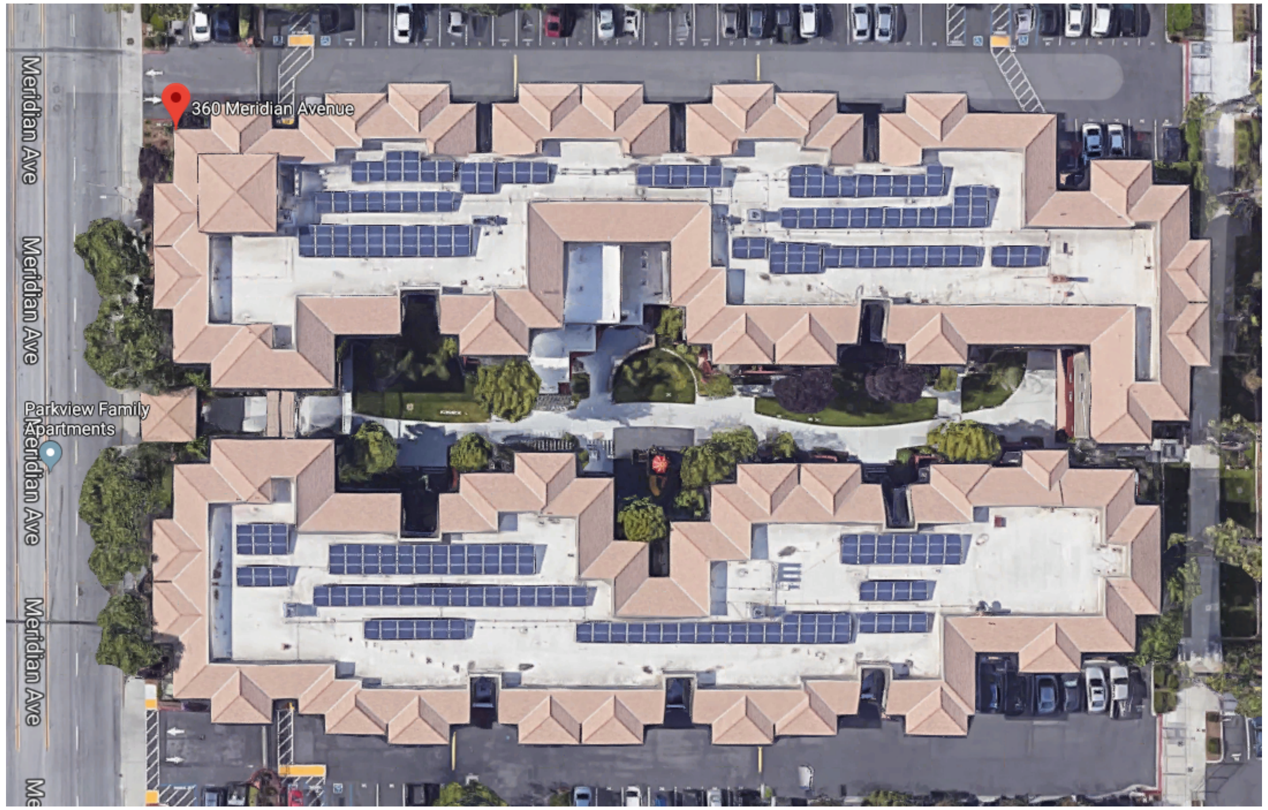


Meridian
Meridian
Meridian
Meridian

Location: -121.913144 37.322127 Deci

Field	Value
OBJECTID_1	241052
Shape	Point
ParcelID	349023
FullMailin	360 Meridian Ave Unit 234, Sa
Lat	37.322121
Long	-121.913221

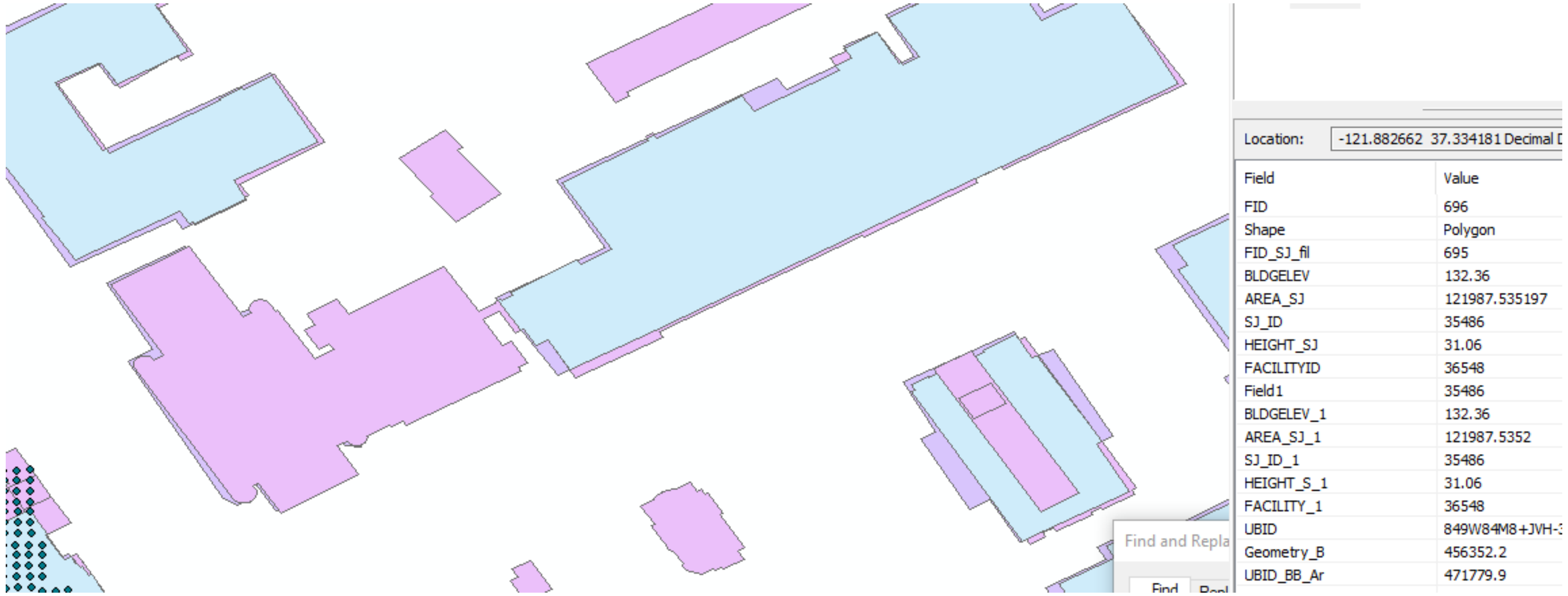




Pacific Northwest
NATIONAL LABORATORY

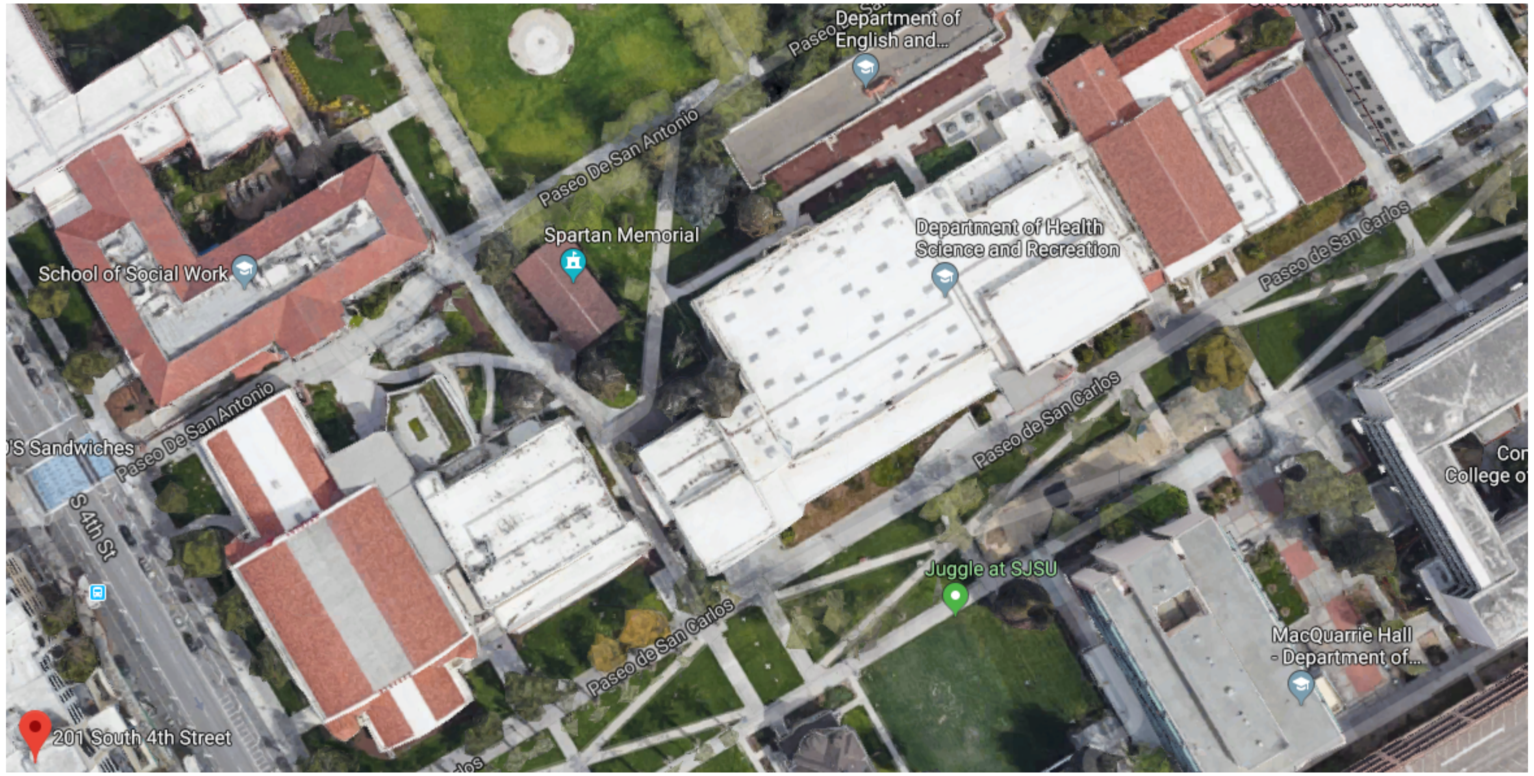
🟪	Microsoft
🟪	San Jose
🟦	Intersection

Example: Missing Match (not in UBID Cross Reference)

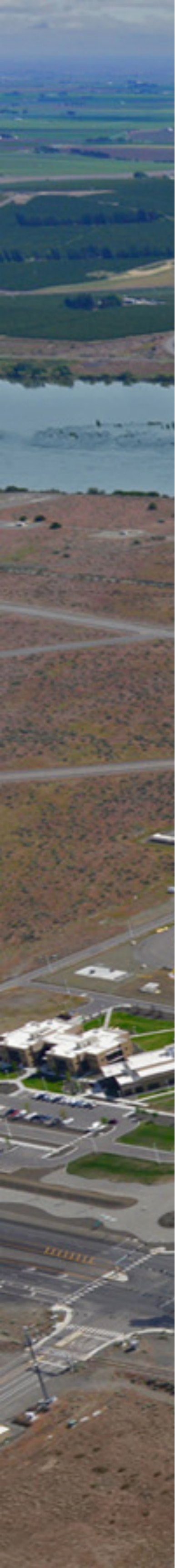


Location: -121.882662 37.334181 Decimal D

Field	Value
FID	696
Shape	Polygon
FID_SJ_fil	695
BLDGELEV	132.36
AREA_SJ	121987.535197
SJ_ID	35486
HEIGHT_SJ	31.06
FACILITYID	36548
Field1	35486
BLDGELEV_1	132.36
AREA_SJ_1	121987.5352
SJ_ID_1	35486
HEIGHT_S_1	31.06
FACILITY_1	36548
UBID	849W84M8+JVH-3
Geometry_B	456352.2
UBID_BB_Ar	471779.9



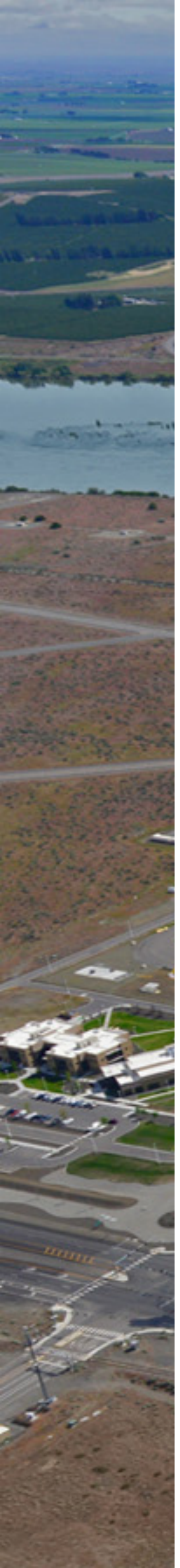
Another example of two MS footprints and one SJ footprint – this one is more spread out and not oriented N-S so UBID doesn't match, but GIS does



Pacific Northwest
NATIONAL LABORATORY

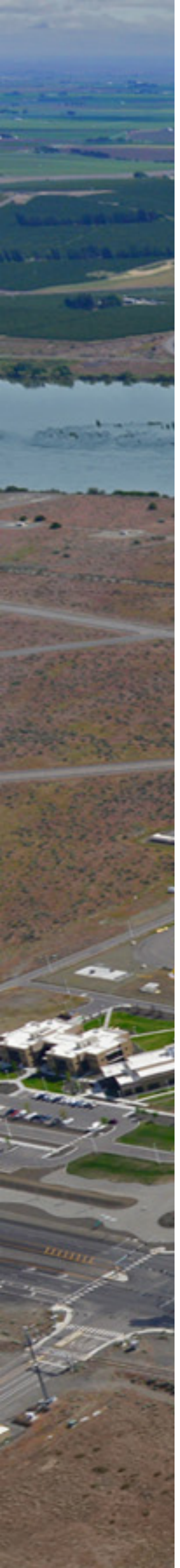
Conclusions

- UBID cross reference can achieve 98.1% correspondence to GIS cross reference for subset of SJ buildings (50k sqft+)
  - If UBID cross reference can achieve similar results to GIS (which is the current best practice for spatial matching [w/o machine learning]), then UBID is a feasible mechanism for establishing equivalency between similar datasets
  - UBID has advantages like transcribability, natural key, universal coding/decoding, etc.
  - The incorrect matches tended to be close to the threshold so the incongruency is more due to the ambiguity of what is considered a match, not due to the methodology for finding matches
- Different grouping metrics (area of intersection, IoU, and centroid distance didn’t greatly alter matching success
  - Look into combinations of these, other heuristics, or machine learning algorithms to find matches in more unique situations
  - Could be more pronounced for larger or messier datasets



Objectives

1. Compare Microsoft footprints to SJ footprints in GIS to see how similar the two datasets are
2. Investigate UBID one-to-one matching between two building footprint datasets (MS and SJ) and compare to GIS matching
3. **Investigate accuracy of UBID cross referencing for matching address UBID0 to polygon UBIDs**
4. **Investigate accuracy of UBID cross referencing for buildings to parcels many-to-many matching**



Match Address to Parcel

INTERSECTIONS	Address	Parcel
No intersections	96	32,633
One-to-one	198,708	198,708
One parcel-to-many address)	177,058	16,259
Total	**375,862**	**247,600**

GIS:

1. SPATIAL JOIN addresses to parcels
2. Address and Parcel both have ParcelID field – compare accuracy:

	Correct	Incorrect	No Match	Success Rate
One-to-one (select random address if multiple)	214,872	95	32,633 parcels/ 160,895 addresses	**99.96%**
One-to-many (validate all address independently if multiple)	375,333	433	32,633 parcels/96 addresses	**99.88%**

UBID:

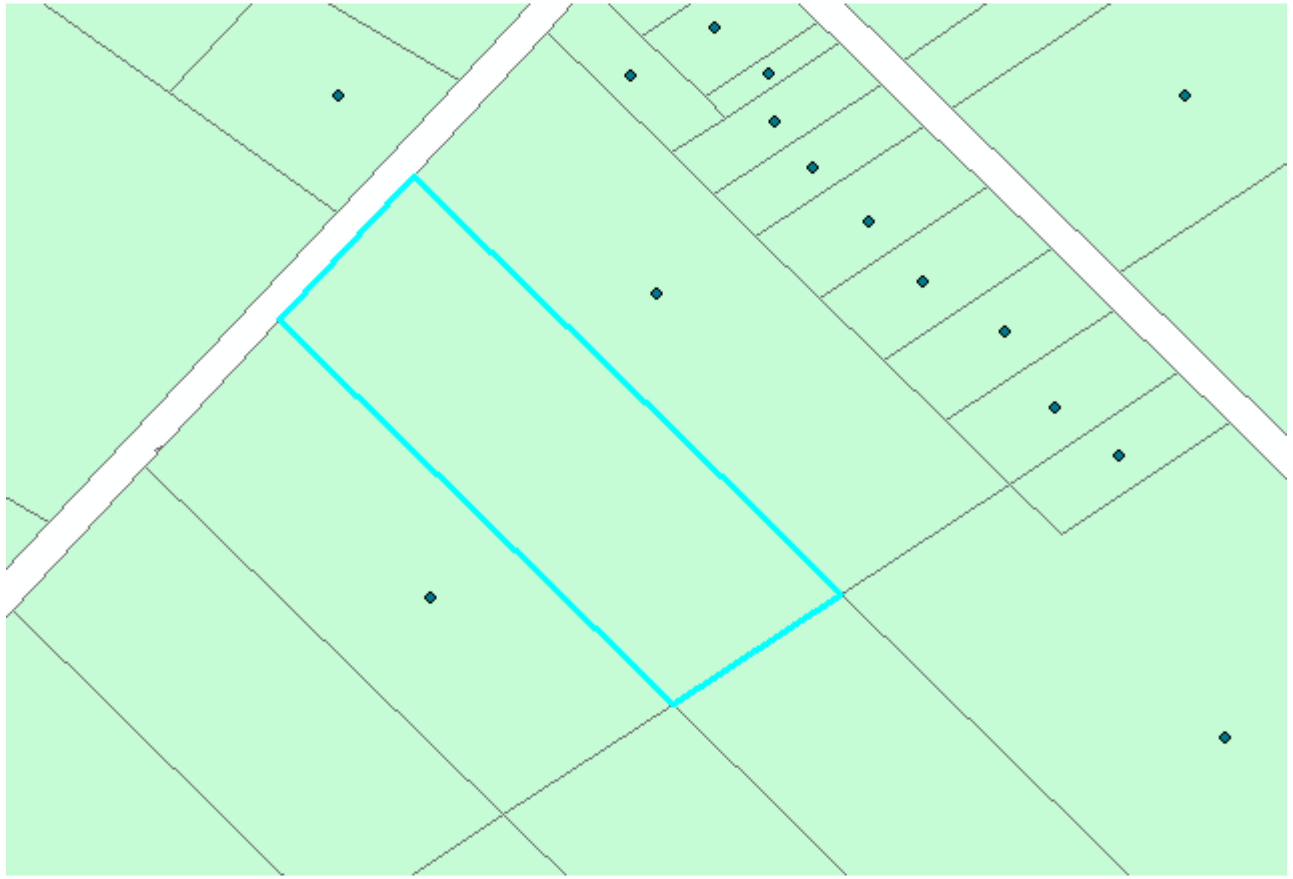
1. Cross reference with IoU > 0, group by centroid radius

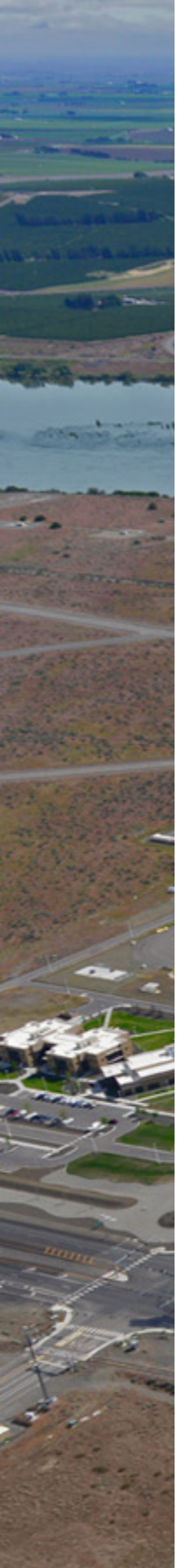
	Correct	Incorrect	No Match	Success Rate
Best centroid radius	199,423	10,309	37,868 parcels	95%

- Cannot be significantly improved with centroid radius max

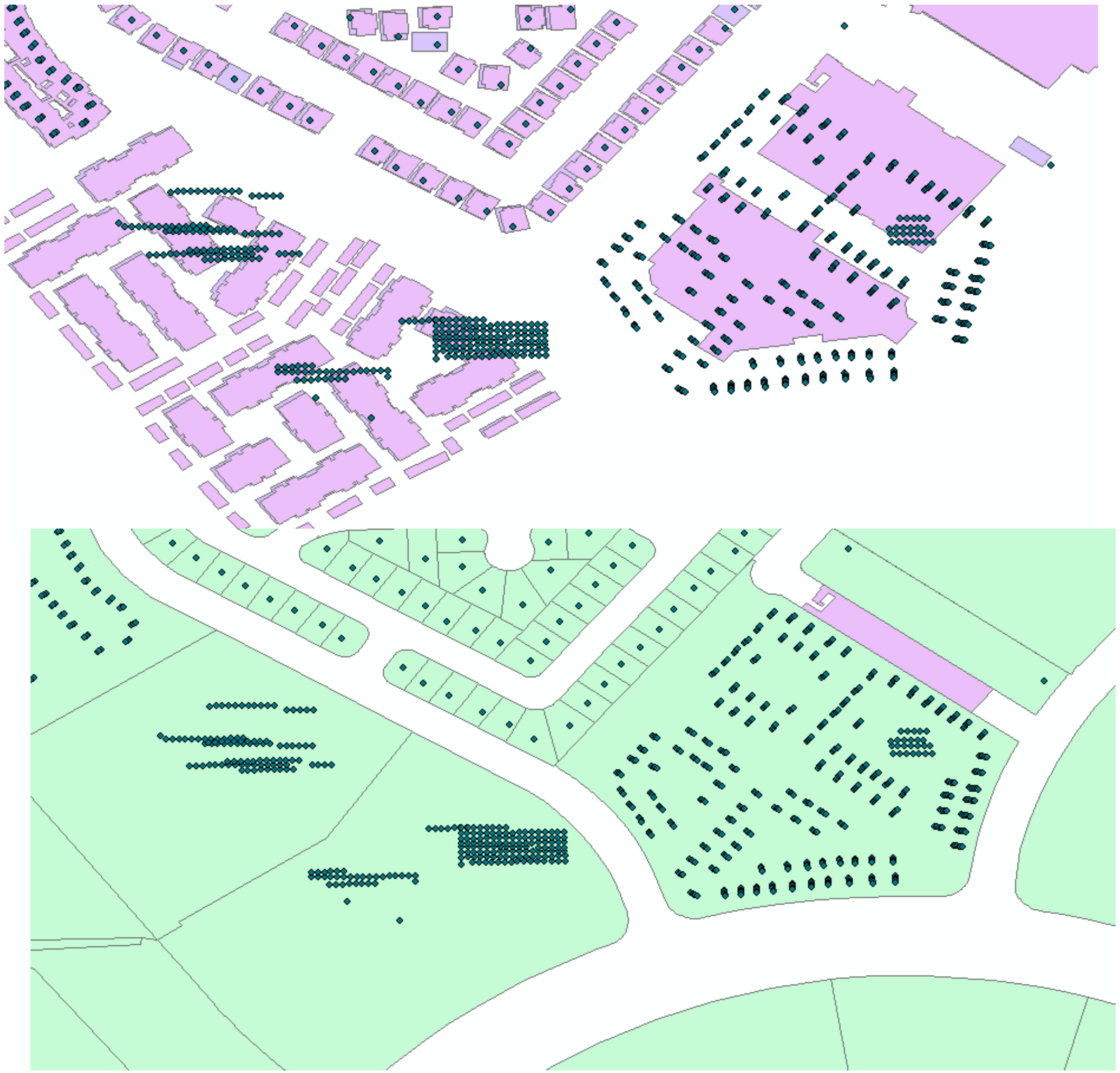
Improvements

- 95% is good but ways to improve?
- Example Incorrect: Highlighted Parcel (8565412) was matched to Address to right (8565414)
    - Why????

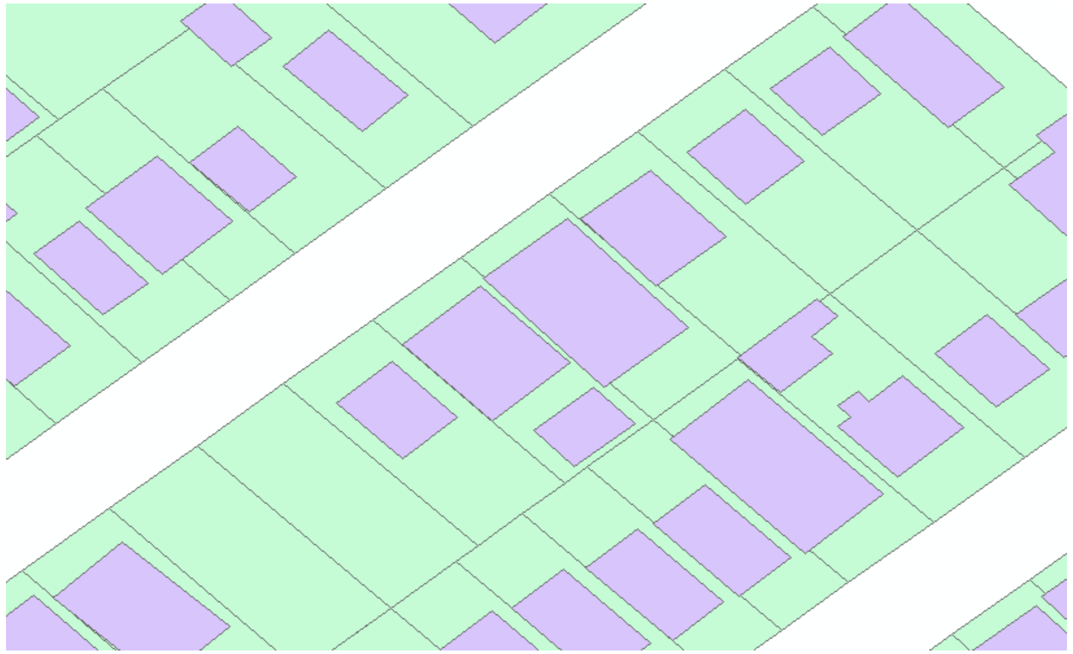




- Address matching works well with parcels for SJ but not always well aligned for buildings



Match Footprint (MS) to Parcel



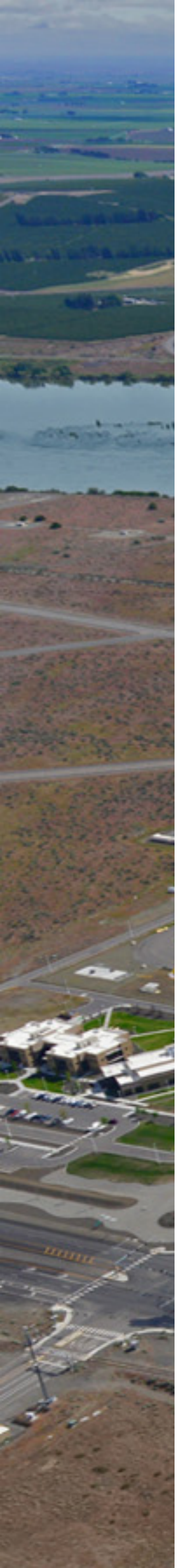
GIS matching to create ground truth of parcel-building relationships

1. INTERSECT MS buildings to SJ parcels
2. Calculate percent overlap
    1. Intersect area / building footprint area
3. Filter out overlaps less than 10% (slight overlap is likely mistake

UBID cross reference at difference IoU thresholds

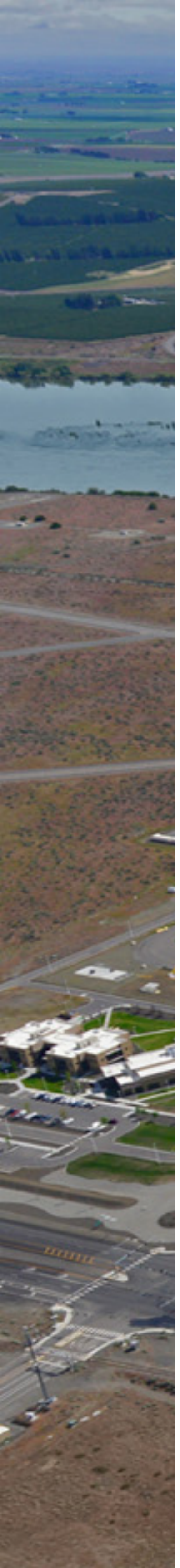
- No grouping because many-to-many
- Should consider other matching criteria…

	Correct Matches	Incorrect Matches	Success Rate	Notes
IoU > 0	195,175	802,207	20%	• Includes all intersections



Next Steps:

- Match Jacob’s addresses to Parcels and Buildings
  - Make UBID0’s from the geolocations
  - Parcels will be easy, buildings more problematic
  - Or just take subset from previous analysis



Pacific Northwest
NATIONAL LABORATORY

Thank you