U.S. DEPARTMENT OF ENERGY

Office of ENERGY EFFICIENCY & RENEWABLE ENERGY

# BEDA Accelerator: Washington, D.C.

February 5, 2019

# Agenda

- Introductions
- Overview of UBID generation methodology & data requirements
- Example of UBID Generation: UBID Demonstrator & Drawing Tool
- Analysis of DC UBIDs
- Identify viable datasets for integration of DC UBIDs
- Discussion of Implementation Strategy & IT Requirements

# Problem Statement

The lack of a standardized way to identify buildings makes it difficult to accurately associate data with a specific facility, creating a barrier to effective asset management, research, and analysis.

**Where the current address system breaks down:**

- Different address abbreviation, e.g., st or street; ave or avenue; apt or #;
- Simple misspellings or incorrect addresses
- Large buildings with multiple entrances and possibly multiple addresses

# Problem Statement

The lack of a standardized way to identify buildings makes it difficult to accurately associate data with a specific facility, creating a barrier to effective asset management, research, and analysis.

**How this looks for benchmarking:**


| CBL | vs. | ESPM |
|---|---|---|
| 123 Main St | or | 123 Main Street |
| 456 5th Street | or | 789 Central Ave. |
| Tax ID 5578 | or | Tax Lot 5577 |

# Problem Statement

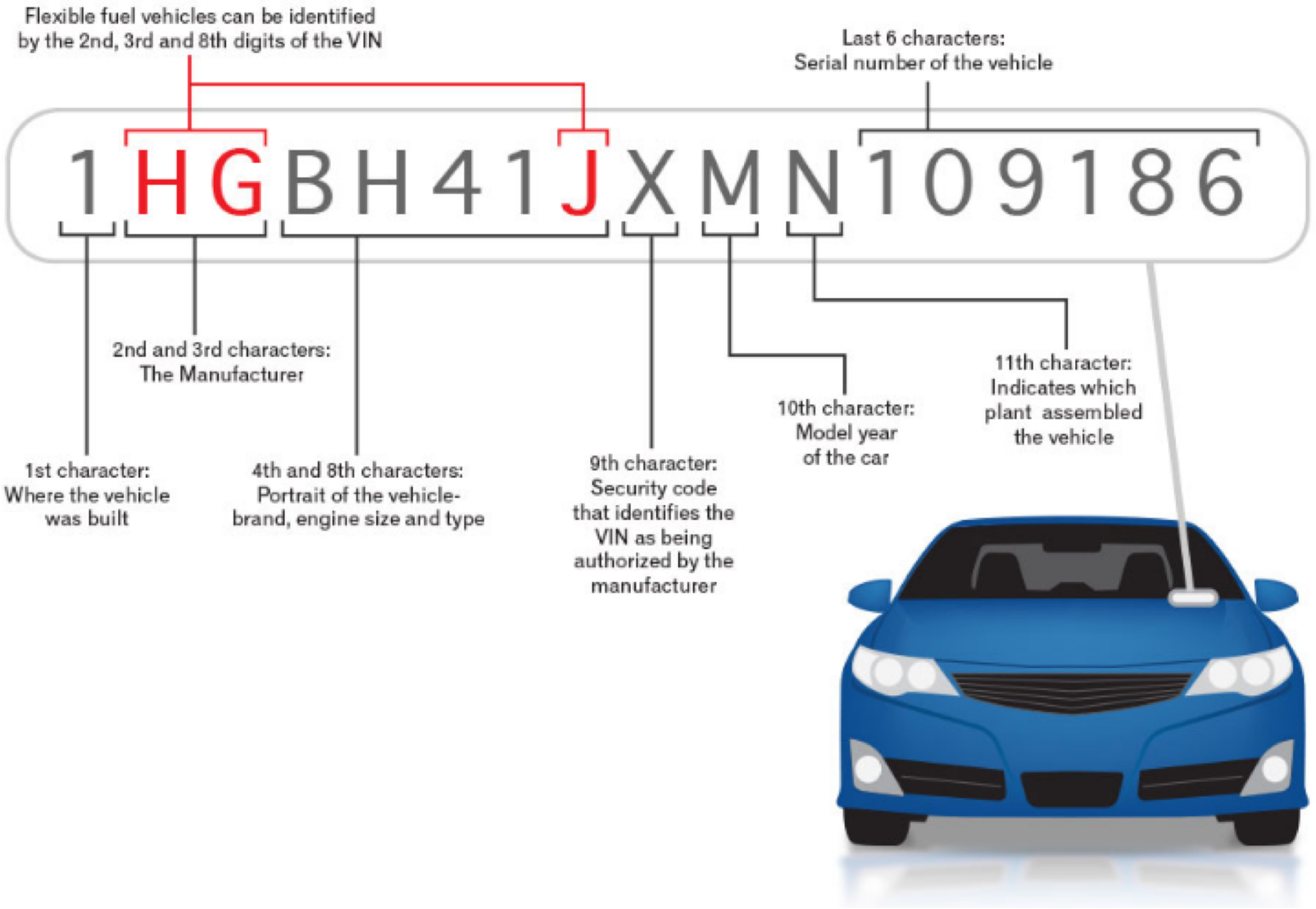
The lack of a standardized way to identify buildings makes it difficult to accurately associate data with a specific facility, creating a barrier to effective asset management, research, and analysis.

## How bad is this problem?

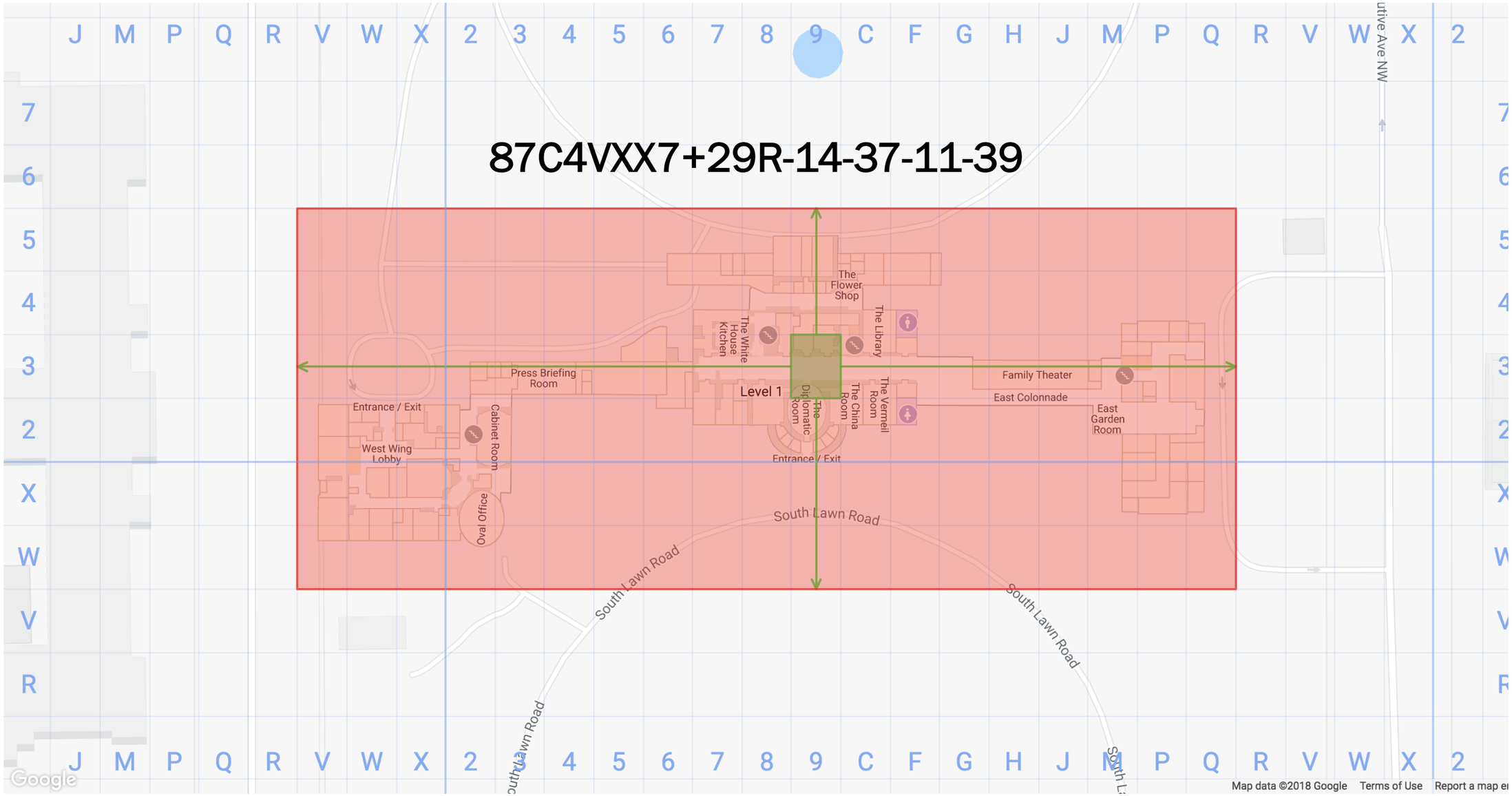
- An analysis of 800k buildings in Houston, TX yielded an 80% match rate based on address from pre-cleansed datasets; an **additional 20-30 person hours** were required to reach a 95% match rate using fuzzy matching algorithms and hand matching.
- Even small towns like Department of Planning in South Burlington, VT estimates **2 hours/month** go into developing data workarounds for bad matches
- According to Ecotope and SF Department of Environment, average match rates are 50-60%. **UBIDs could save days to weeks of manual data matching efforts**.

Acknowledgement: UC Berkeley Student Consulting & Research Group

# Solution: A Natural Key for Buildings



# Solution: A Natural Key for Buildings



# UBID Demonstrator

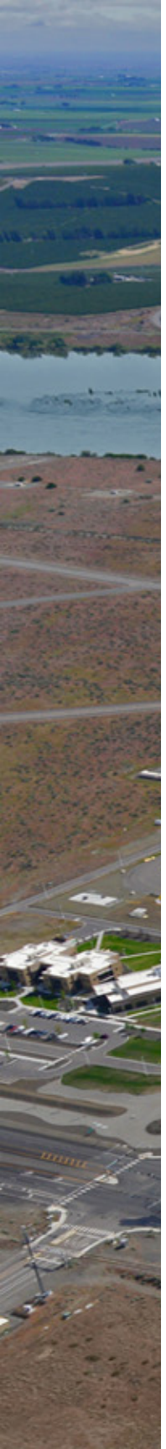
UBID.PNNL.GOV

Pacific Northwest
NATIONAL LABORATORY

# UBID Matching Washington, DC

U.S. DEPARTMENT OF ENERGY BATTELLE

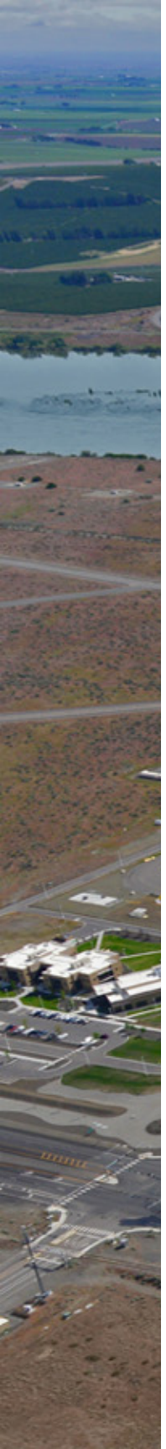
PNNL is operated by Battelle for the U.S. Department of Energy



Pacific Northwest NATIONAL LABORATORY

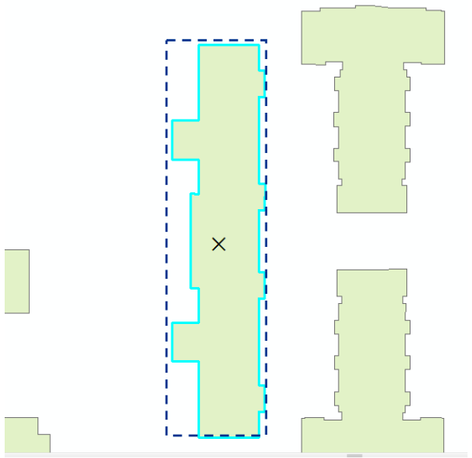
# Datasets

- Open Data Footprints
  - http://opendata.dc.gov/datasets/a657b34942564aa8b06f293cb0934cbd_1
  - 163,467 entries
  - No local ID (“GIS_ID” field empty)
- Energy Benchmarking 2016
  - https://doee.dc.gov/publication/2016-building-benchmarking-dataset
  - 1,846 entries
  - pid, dc_real_pid, and pm_pid are local IDs, pid only with no duplicates and value for every entry
- Other datasets used for analysis:
  - Street Centerlines
    - http://opendata.dc.gov/datasets/street-centerlines
  - Address Points
    - http://opendata.dc.gov/datasets/address-points



Pacific Northwest
NATIONAL LABORATORY

# UBID Matching Process

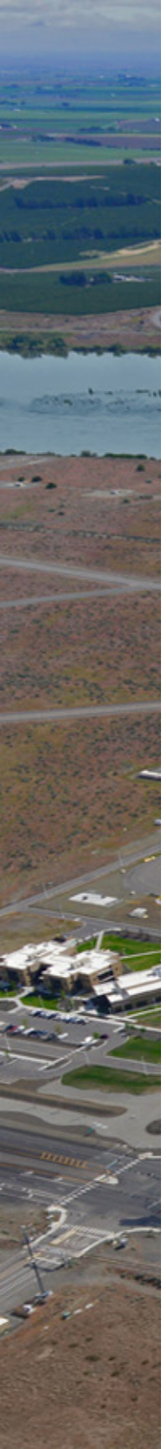


**MATCH**

**Benchmarking Point**
PID: PM05823132
87C4VXJM+452-0-0-0-0

**Footprint (Bounding Box)**
87C4VXJM+456-29-5-29-6

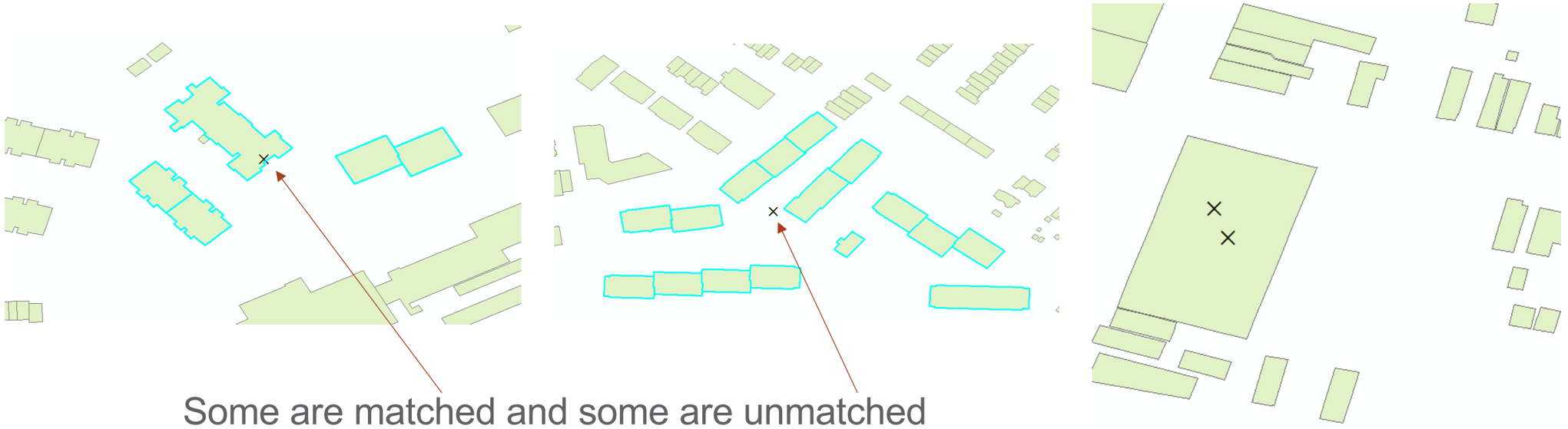
- 0 invalid geometries from footprints and benchmarking
- 1,608 benchmarking points ($UBID_0$) matched to footprint UBIDs
  - 238 $UBID_0$ not matched
  - 191 duplicate $UBID_0$ created



Pacific Northwest
NATIONAL LABORATORY

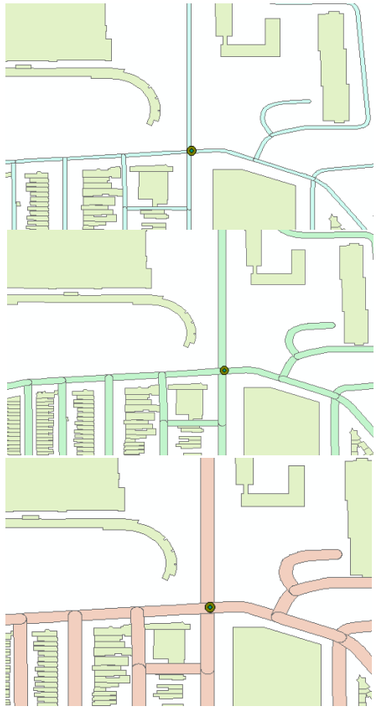
# UBID Matching Main Issue Overview

- Benchmarking points that represent multiple building footprints and multiple benchmarking points that represent the same building footprint



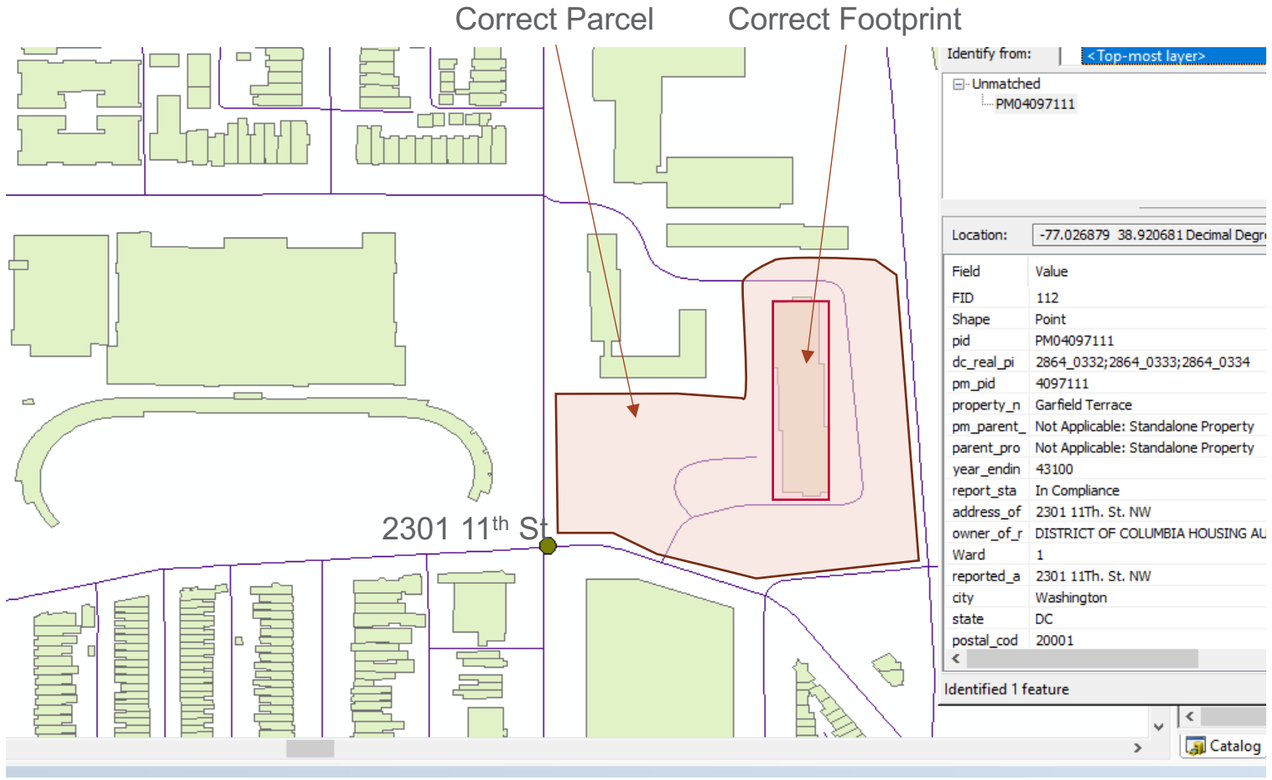
Some are matched and some are unmatched

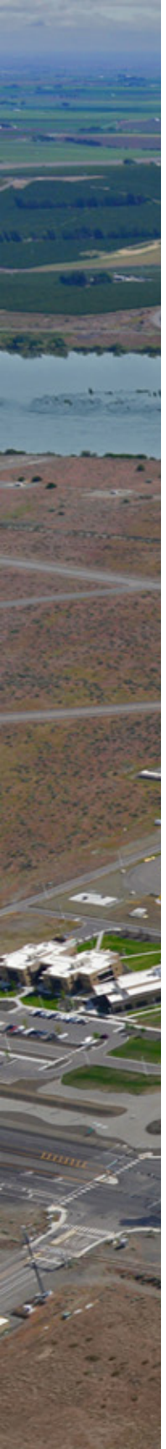
# Unmatched Points: $UBID_0$ in Street



- 5ft buffer: 32 $UBID_0$
- 10 ft buffer: 66 $UBID_0$
- 15ft buffer: 71 $UBID_0$

- Some unmatched points are the “campus” issue on the previous slide
- Others are points that are in the street, and the first step is to differentiate these





Pacific Northwest
NATIONAL LABORATORY

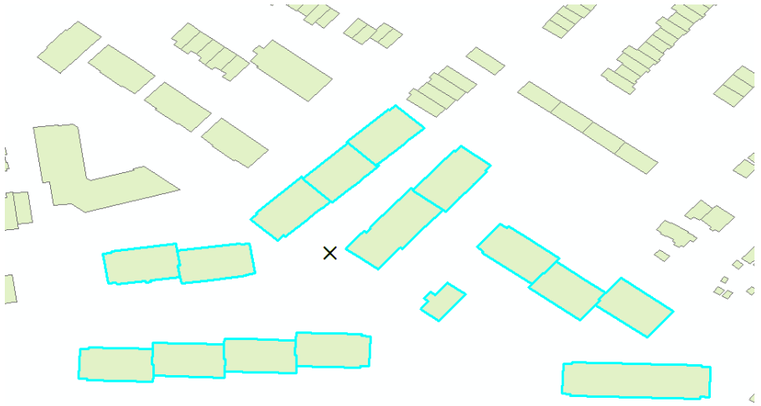
# Solution

- Matching to nearest footprint is a quick solution, but there will be false positives (like example in previous slide)
- Best solution is manual review of the ~70 points
  - a little time consuming, but only needs to be done once
- Other solution is matching addresses
  - Not perfect, usually 60-80% success rate, but 60-80% for 5% of database isn't bad
  - Matching addresses requires some data processing to match the formatting, could be almost as time consuming as manual review
- For future benchmarking, worth making reporter quickly confirm that the geocoded address doesn't lie in street

Pacific Northwest
NATIONAL LABORATORY

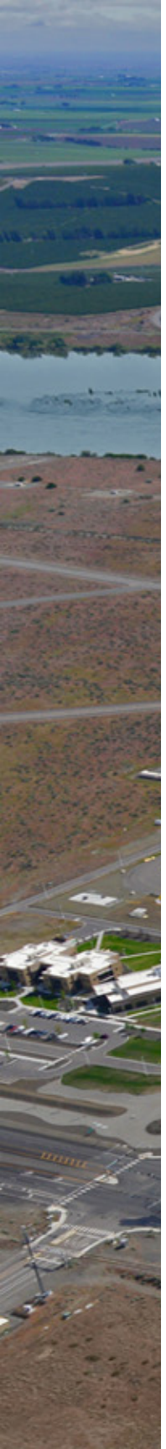
# One (unmatched) $UBID_0$ that represents multiple buildings

About 140 instances

Example Below:

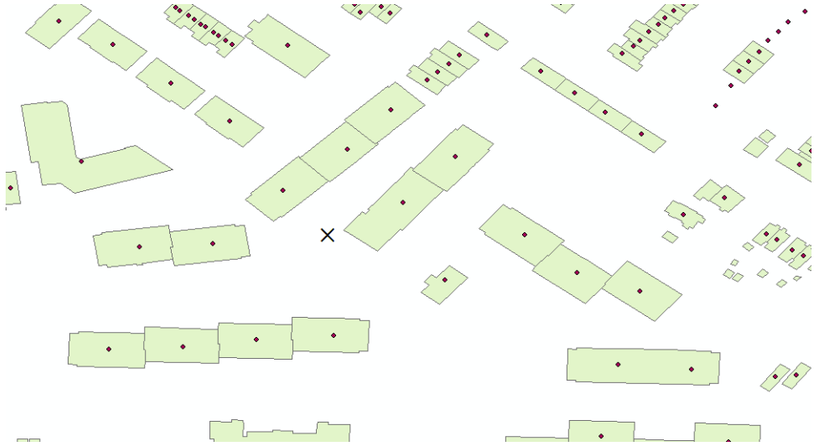
- Estimated area (with area map tool): 77k * 4 floors = ~308k sqft
- Reported area = 280k sqft
- Conclusion: Benchmarking data represents all buildings in this multi-family housing unit but didn't match because fell outside bounding box of all footprints



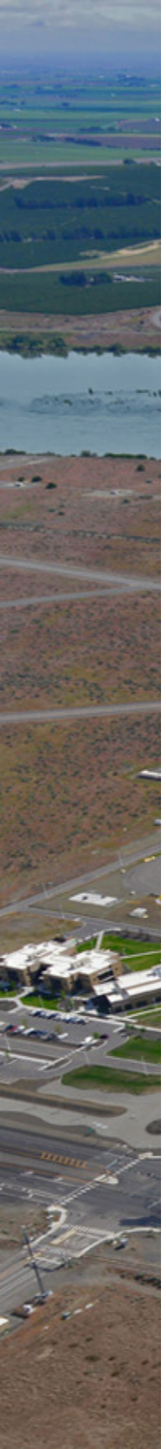
Pacific Northwest
NATIONAL LABORATORY

# Solution

- Merge all footprints that correspond to the $UBID_0$ using the “Square” and “Lot” fields
  - Could use either Parcel Lot or Address Point dataset to facilitate the merge
  - Some data processing labor involved
- Worth doing for $UBID_0$ in street in case they have multiple buildings

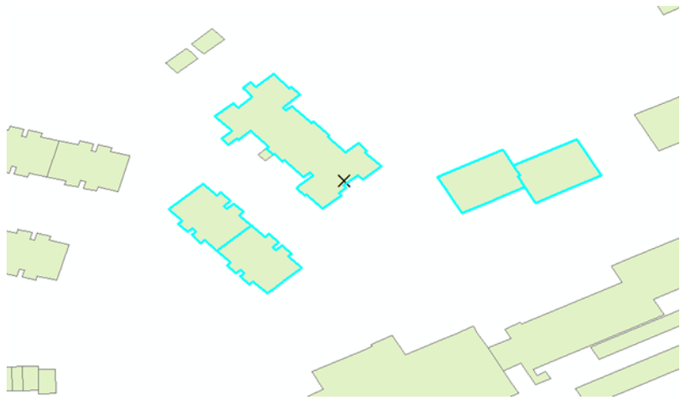


| | |
|---|---|
| FULLADDRES | 2615 BOWEN ROAD SE |
| SQUARE | 5869 |
| SUFFIX | |
| LOT | 0068 |



Pacific Northwest
NATIONAL LABORATORY

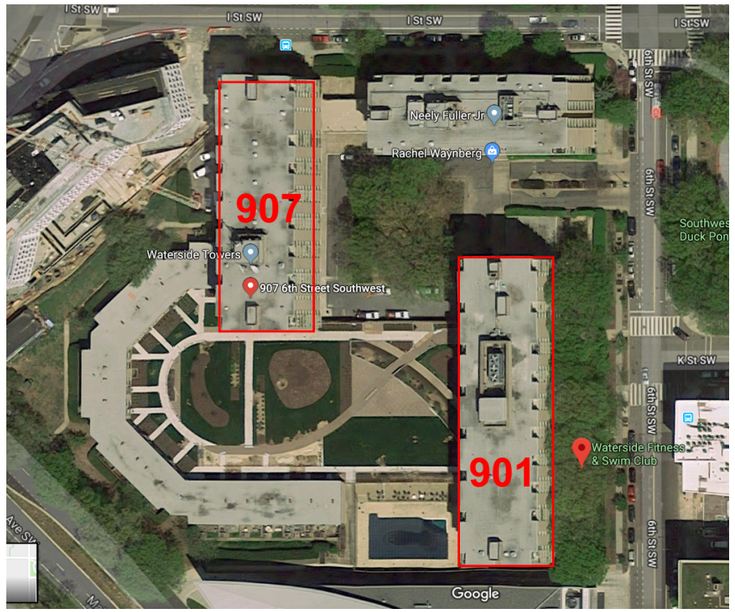
# One (matched) $UBID_0$ for multiple buildings



Hypothetical example

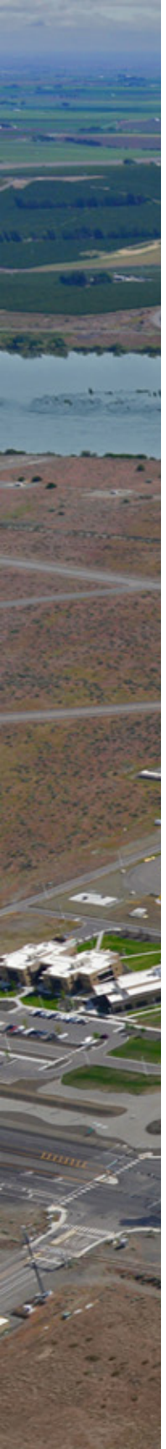
- We assumed there are cases like the unmatched ones, that happened to land within a footprint bounding box – but no way to detect these
- Can be improved in future benchmarking by including critical data to identify these
- Quick estimate (not up to date):
  - 544 $UBID_0$ that are matched to footprints with multiple addresses in the same lot
  - Even if we can flag the $UBID_0$ with multiple addresses in same lot, could be difficult to determine if the $UBID_0$ is only for one in matched to or for all the buildings

# Multiple $UBID_0$ for multiple buildings




- Example:
  - 901, 907, and 907 6th St SW
- Benchmark XY all on 907 address
- Area
  - 901 area: 20,450*9 = ~184k
  - 907 area: 17,400*9 = ~157k
  - 3 reported areas (381k, 100k, and 53k) and a tax record of 1M sqft
- The two are very similar architecturally and to the other buildings on the plot
- Conclusion: The taxable area (1Msqft) represents all buildings on the property and the three benchmarking are some combination of sub spaces

| | A | I | L | O | P | Q | S | T | U | V | W | X | Y | Z | AA | AB | AC | AD | AE | AL | AP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | pid | address_of_record | reported_address | posta | year | primary_p | tax_rec | reporte | energy | site_eu | weathe | source | weathe | total_g | total_g | water_ | water_ | electric | natural | UBID | |
| 7 | PM03531644 | 0901 - 0947 6TH ST SW | 901 6th Street SW | 20024 | 1971 | Multifamily H | 1037766 | 381600 | 17 | 90.6 | 97.4 | 164.4 | 169.2 | 2567.3 | 6.7 | 22938.8 | | 3703830 | 219421 | 87C4VXHH+FG9-0-0-0-0 | |
| 8 | PM04007394 | 0901 - 0947 6TH ST SW | 907 6th Street SW | 20024 | 1965 | Multifamily H | 1037766 | 100326 | 18 | 83.2 | 89.3 | 148 | 151.2 | 611.5 | 6.1 | 4820.2 | | 853432.8 | 54313.18 | 87C4VXHH+FG9-0-0-0-0 | |
| 9 | PM04007391 | 0901 - 0947 6TH ST SW | 907 6th Street SW | 20024 | 1971 | Multifamily H | 1037766 | 53000 | 50 | 75.8 | 83.6 | 120 | 128.2 | 272.6 | 5.1 | 3127.6 | | 300367.4 | 29922 | 87C4VXHH+FG9-0-0-0-0 | |

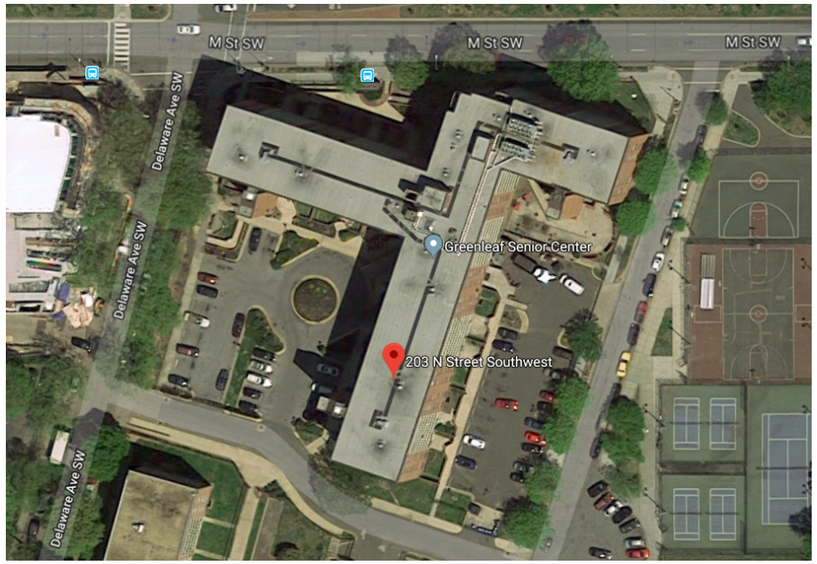


Pacific Northwest
NATIONAL LABORATORY

# Solution

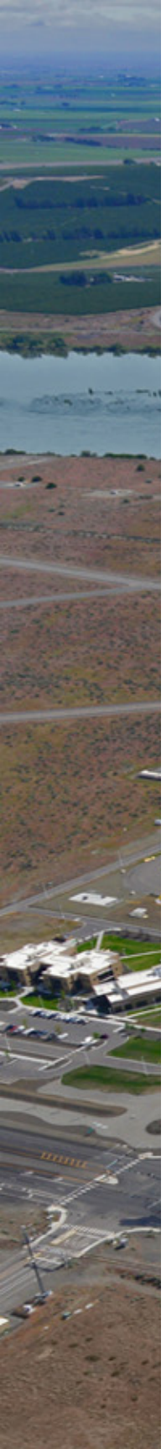
- Detection: Duplicate $UBID_0$ that also have multiple buildings on parcel
  - Some labor involved in this detection process
- Impossible to know, even with manual inspection, what benchmarking entries represent which spaces
- Question for DC: What would be the appropriate solution for this example?
  - Idea for future: mark as not compliant because impossible to know which spaces are being benchmarked
  - Idea 1: merge footprints and create one UBID for the parcel
    - match all benchmarking entries
    - Aggregate benchmark data and match only one entry

# Multiple $UBID_0$ for single building




- Example: 203 N St SW
- Calculate Area in Google: 21,700*8 = ~173k sqft
- Reported Area:
  - 115,323
  - 23,876
  - 35,992
  - Total: 175k
- Conclusion: multiple spaces benchmarked separately

| | A | I | L | O | P | Q | S | T | U | V | W | X | Y | Z | AA | AB | AC | AD | AE | AL | AP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | pid | address_of_record | reported_address | posta | year_ | primary_p | tax_rec | reporte | energy | site_eu | weathe | source_ | weathe | total_g | total_g | water_ | water_ | electric | natural | UBID | |
| 2 | PM04178733 | 203 N St. SW | 203 N St. SW | 20024 | 1959 | Multifamily Housing | | 115323 | 1 | 148.3 | 158.8 | 259.3 | 267.6 | 1239 | 10.7 | 23241.2 | | 1674599 | 113929 | 87C4VXGP+7JX-0-0-0-0 | |
| 3 | PM04178731 | 203 N St. SW | 203 N St. SW | 20024 | 1960 | Multifamily Housing | | 23876 | 46 | 120.4 | 133.6 | 154.2 | 167.6 | 171.1 | 7.2 | 4729.2 | | 93000 | 25585.01 | 87C4VXGP+7JX-0-0-0-0 | |
| 4 | PM04178732 | 203 N St. SW | 203 N St. SW | 20024 | 1965 | Multifamily Housing | | 35992 | | 10.4 | 11.3 | 17.6 | 18.3 | 26.5 | 0.7 | 669.5 | | 33485 | 2601 | 87C4VXGP+7JX-0-0-0-0 | |

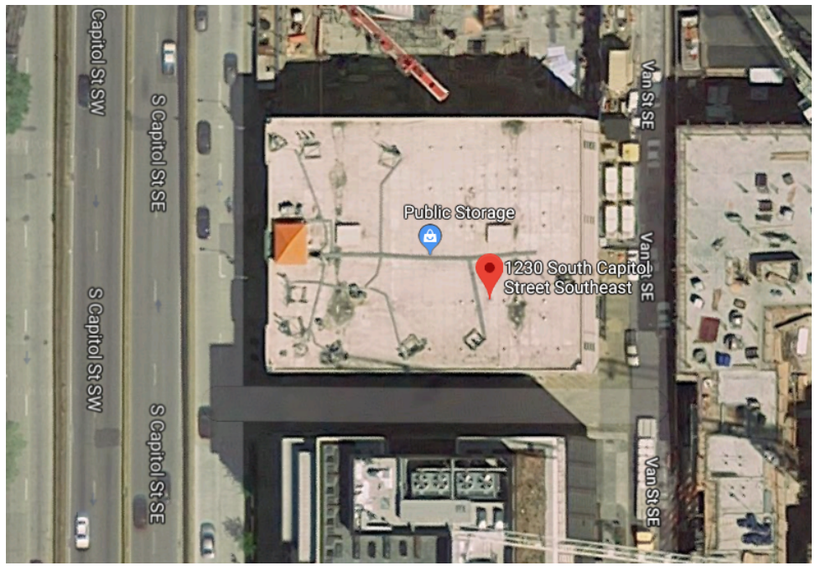


Pacific Northwest
NATIONAL LABORATORY

# Solution

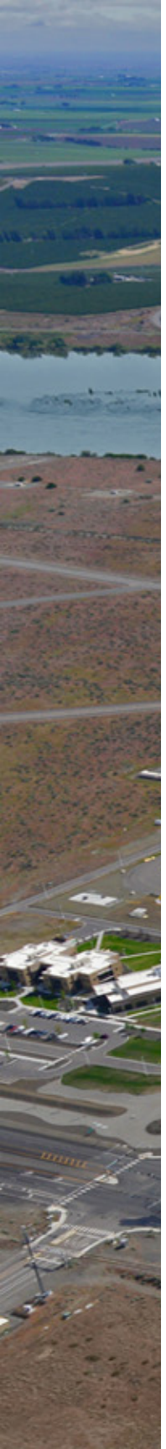
- Similar to previous case – impossible to know which spaces in the building are being benchmarked
- Question for DC: What would be the appropriate solution for this example?
  - Idea 1: no action (i.e. match all benchmarking $UBID_0$ to one footprint UBID)
  - Idea 2: Aggregate data and match only one entry

# False Matching: Incorrect Location




- Example:
  - 1230 S Capitol SE
  - 1263-1265A A 1st St SE
- $UBID_0$ location and use type match first address, second address is a few blocks away
- Conclusion: Incorrect coordinates entered for second address

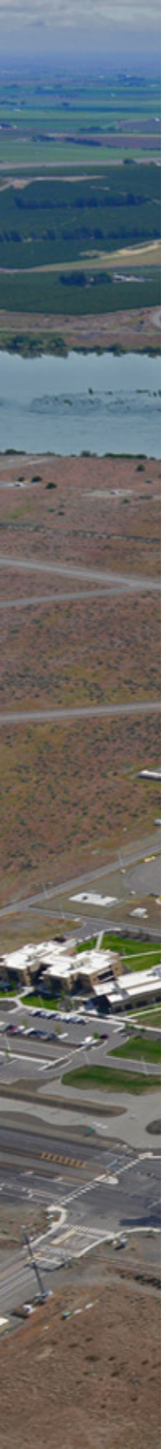
| | A | I | L | O | P | Q | S | T | U | V | W | X | Y | Z | AA | AB | AC | AD | AE | AL | AP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | pid | address_of_record | reported_address | posta | year | primary_p | tax_rec | reporte | energy | site_eu | weathe | source | weathe | total_g | total_g | water_ | water_ | electric | natural | UBID | |
| 5 | PM05932679 | 1263-1265 1ST ST SE | 1263-1265Â Â 1ST ST SE | 20003 | 2015 | Hotel | 118944 | 118944 | 85 | 54.2 | 54.2 | 131.9 | 131.9 | 589.2 | 5 | 3065.4 | | 1249441 | 21889.93 | 87C4VXGR+6HF-0-0-0-0 | |
| 6 | PM03518921 | 1230 SOUTH CAPITOL ST | 1230 South Capitol Stree | 20003 | 1991 | Non-Refriger | 108000 | 89999 | 16 | 33.6 | 37.2 | 67 | 71.1 | 239.6 | 2.7 | 42 | | 400980 | 16530 | 87C4VXGR+6HF-0-0-0-0 | |



Pacific Northwest
NATIONAL LABORATORY

# Solution

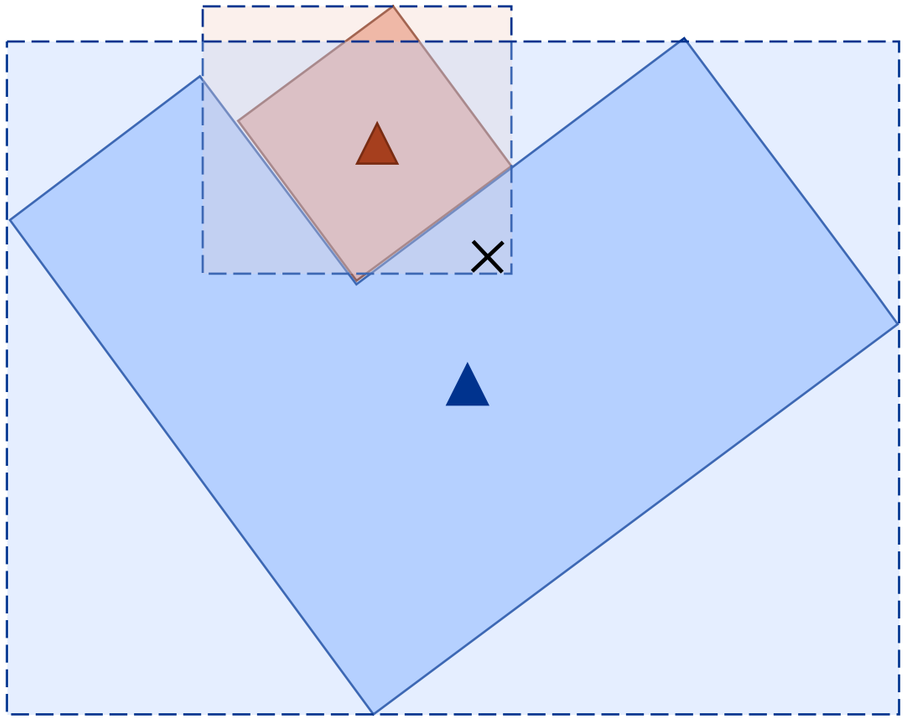
- Garbage in – garbage out?
- Potentially flag (with address?) for revision
  - Just looking through this will not be clean because addresses can vary slightly in zipcode, address number, street format
- For future: when benchmarking ask reporters to confirm geolocation on map (5 seconds)



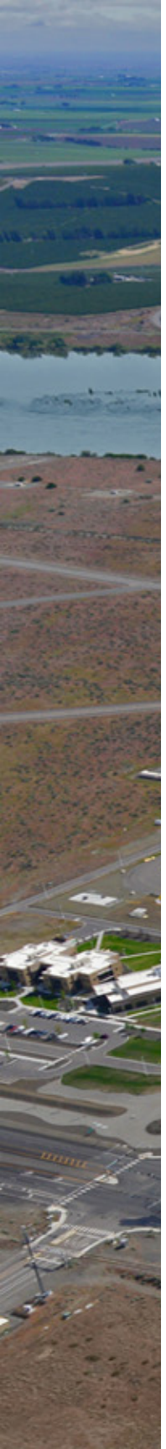
Pacific Northwest
NATIONAL LABORATORY

# False Matching: Incorrectly Matched

- If benchmarking geolocation isn't well aligned with center of the bounding box of the footprint, there is a chance it could be falsely matched to neighboring footprints



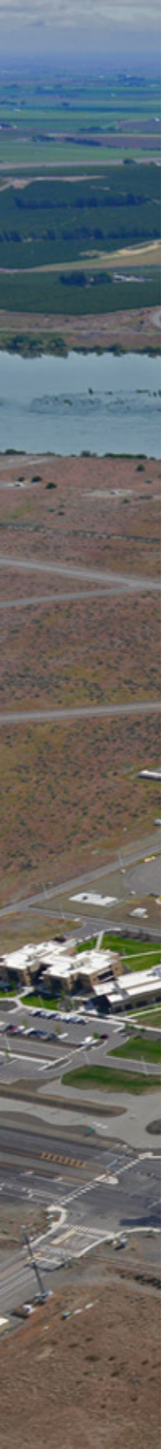
Benchmarking Location
Footprint
Footprint Bounding Box
Footprint Bounding Box Centroid



Pacific Northwest
NATIONAL LABORATORY

# Solution

- We can't definitively find or fix every false positive, but it's possible to look at a subset of data to extrapolate our success rate
- For every benchmarking $UBID_0$ that intersects with multiple bounding boxes, compare the closest and 2$^{nd}$ closest centroids. If the distances are close (say within ~20%) we can flag these for manual review
- Another possibility: Compare distance between $UBID_0$ and matched centroid to the area of the bounding box or the percent area increase between the footprint and the bounding box
- Another possibility: Look at edge cases with large percent area increase between footprint and bounding box



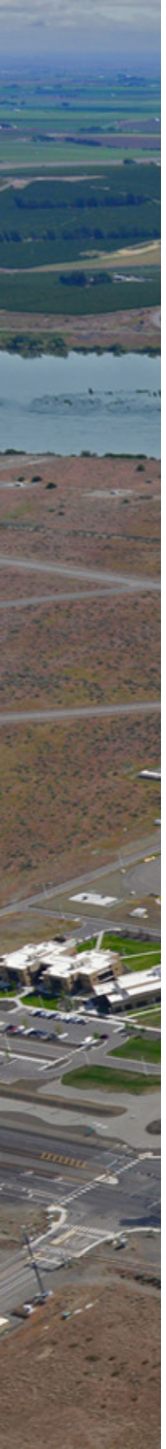
Pacific Northwest
NATIONAL LABORATORY

# Other Possible Cases

- Other cases that may be worth investigating, but would require more time to detect these
  1. Multiple $UBID_0$ with different location but on same building
  2. Multiple $UBID_0$ with different exact location on same property with multiple buildings
  3. One $UBID_0$ represents subsection of building

Pacific Northwest
NATIONAL LABORATORY

# Thank you



# Engagement & Implementation

- **What do we need?**
  - Technical Leads – who are the folks programming and supporting your database infrastructure?
  - Two+ databases – where do you want to see UBIDs incorporated and matched to each other?
- **Process:**
  - Mark will Skype/WebEx/etc. in with your technical team to understand your database architecture
  - Using the tooling developed at PNNL, UBIDs can be added into your existing systems. In the process, Mark can develop a replicable process for use by additional stakeholders in your organization.

# Next Steps

- Timeline
- Points of Contact
- Relevant Datasets for UBID Integration
- Desired Outcomes & Metrics for Success