Unique Building Identifiers

Harry Bergmann – Data Tools Fellow, U.S. Department of Energy

Nora Wang, Ph.D. – Pacific Northwest National Laboratory

Alex Vlachokostas, Ph.D. – Pacific Northwest National Laboratory

January 17, 2018

Do not cite without permission

Agenda

- Background

- UBID Methodology

- Analysis of Methodology

- Implementation Mechanism
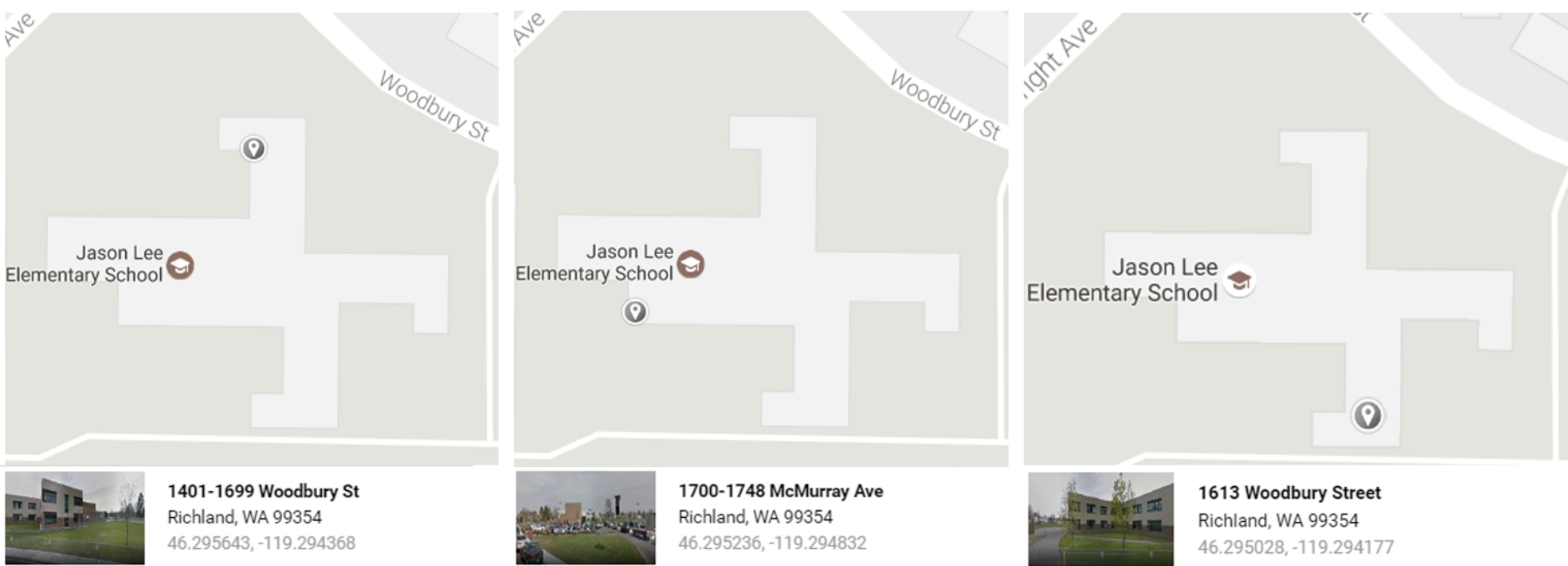
- Q&A

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.

Without standardization, there is no anchor for our building data efforts:

- 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

Example: Geocodes

- A building may have multiple geocodes for the same or multiple addresses, depending on the footprint.
- School Address: 1750 McMurray Ave, Richland, WA 99354

1401-1699 Woodbury St
Richland, WA 99354
46.295643, -119.294368

1700-1748 McMurray Ave
Richland, WA 99354
46.295236, -119.294832

1613 Woodbury Street
Richland, WA 99354
46.295028, -119.294177

Example: Google Place ID

470 L'Enfant Plaza SW
Place ID ChIJD6p3MHa3t4kRYk4VjrIut-E
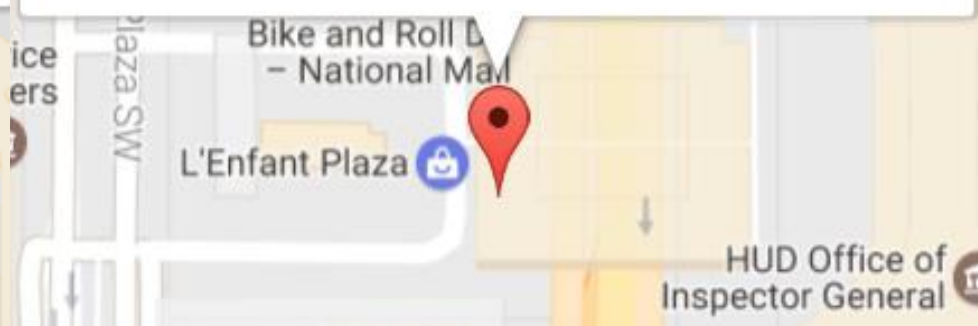470 L'Enfant Plaza SW, Washington, DC 20024, USA
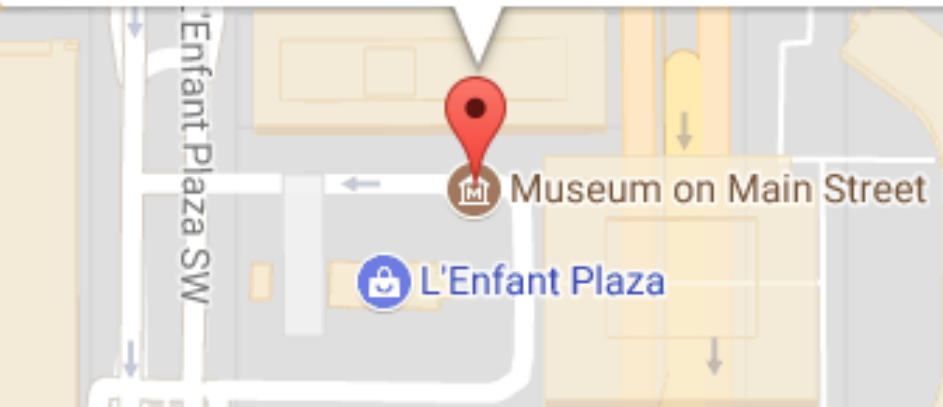
View on Google Maps

490 L'Enfant Plaza SW
Place ID ChIJD9mMzXe3t4kR6-FvrqmdTw0
490 L'Enfant Plaza SW, Washington, DC 20024, USA

View on Google Maps

Example: Pyke and Madan (2013) Space ID

```
var kmlOutput = [
{
```

```
    "SPATIAL"
    "GUID": "1d703a77-5b64-4480-8312-cb49b5693a45",
    "Name": "USGBC Headquarters",
    "Address": "2101 L Street NW",
    "City": "Washington",
    "StateZip": "DC",
    "MinHeightFloor": "5",
    "MaxHeightFloor": "6",
    "TEMPORAL"
    "TimeObservationStartDate": "8/18/2008",
    "TimeObservationEndDate": "4/30/2009",
    "SEMANTIC"
    "PlaceInfo": "U.S. Green Building Council, Delucchi +, American Architectural Foundation",
    "SelectedCenter": "38.904159, -77.04724499999998",
    "Coordinate": [
      {"Latitude": "-77.04725861549377", "Longitude": "38.90443025684092"},
      {"Latitude": "-77.04725325107574", "Longitude": "38.904204830184064"},
      {"Latitude": "-77.04718887805939", "Longitude": "38.904204830184064"},
      {"Latitude": "-77.04718887805939", "Longitude": "38.90406289451437"},
      {"Latitude": "-77.04733908176422", "Longitude": "38.90406289451437"},
      {"Latitude": "-77.04734444618225", "Longitude": "38.90409211658722"},
      {"Latitude": "-77.04755365848541", "Longitude": "38.90409629116806"},
      {"Latitude": "-77.04754292964935", "Longitude": "38.90385416507292"},
      {"Latitude": "-77.04689383506775", "Longitude": "38.90384581588251"},
      {"Latitude": "-77.046673826694489", "Longitude": "38.90382911749872"}
    ]
```

```
}
];
```

Working Group & Major Contributors

Federal Agencies

Department of Energy
General Services Admin.
U.S. EIA
Census
National Institute for
Standards and Technology
Environmental Protection
Agency

Private Sector

Microsoft
Google
CoreLogic

Universities

Northwestern University
UNC Chapel Hill
Massachusetts Institute of
Technology

State & Local Government

City of San Francisco
City of New York
City of Atlanta
City of Washington, D.C.
California Energy
Commission

Non-Profits & NGOs

Institute for Market
Transforation
GRESB
Radiant Earth
New America Foundation
National Association of
Realtors

National Laboratories

PNNL
LBNL
NREL

Working Group Requirements

- **Transparency in methodology**

- **Minimal data requirement**
  - That data which is required should be publicly available

- **Simplicity in execution**
  - The UBID should be human transcribable

Solution

Flexible fuel vehicles can be identified by the 2nd, 3rd and 8th digits of the VIN

Last 6 characters: Serial number of the vehicle

1 H G B H 4 1 J X M N 1 0 9 1 8 6

2nd and 3rd characters: The Manufacturer

1st character: Where the vehicle was built

4th and 8th characters: Portrait of the vehicle-brand, engine size and type

9th character: Security code that identifies the VIN as being authorized by the manufacturer

10th character: Model year of the car

11th character: Indicates which plant assembled the vehicle

But…
- Buildings aren't discrete objects like vehicles.
- So, how do we know we are all talking about the same thing?

What is a building? – Census

Case 1: One address, multiple Structures
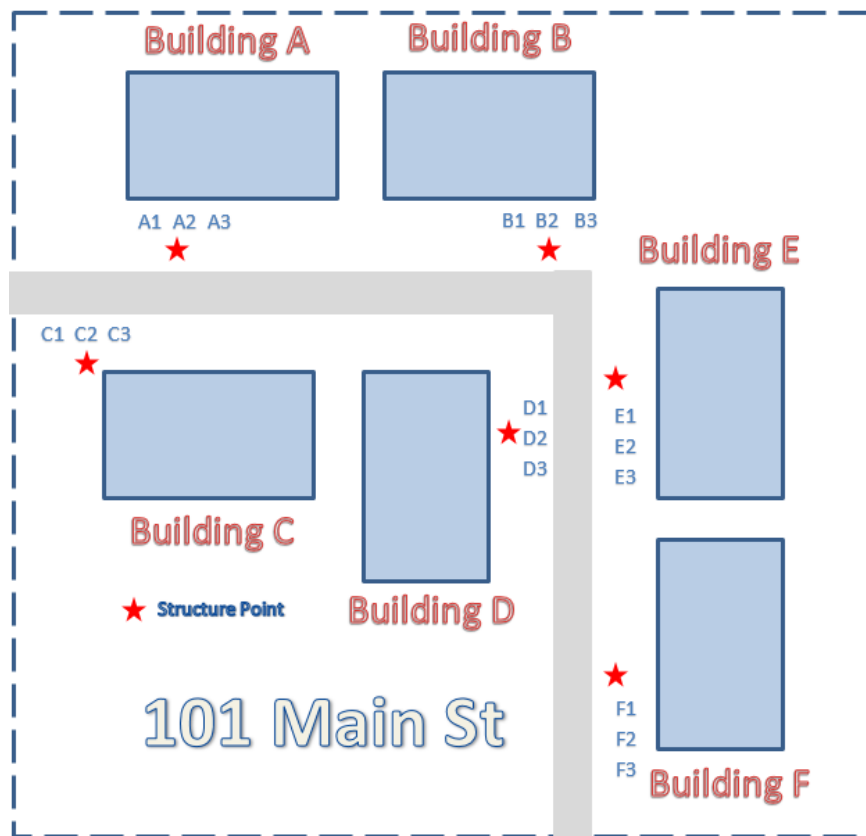
Case 2: One structure, with multiple units.
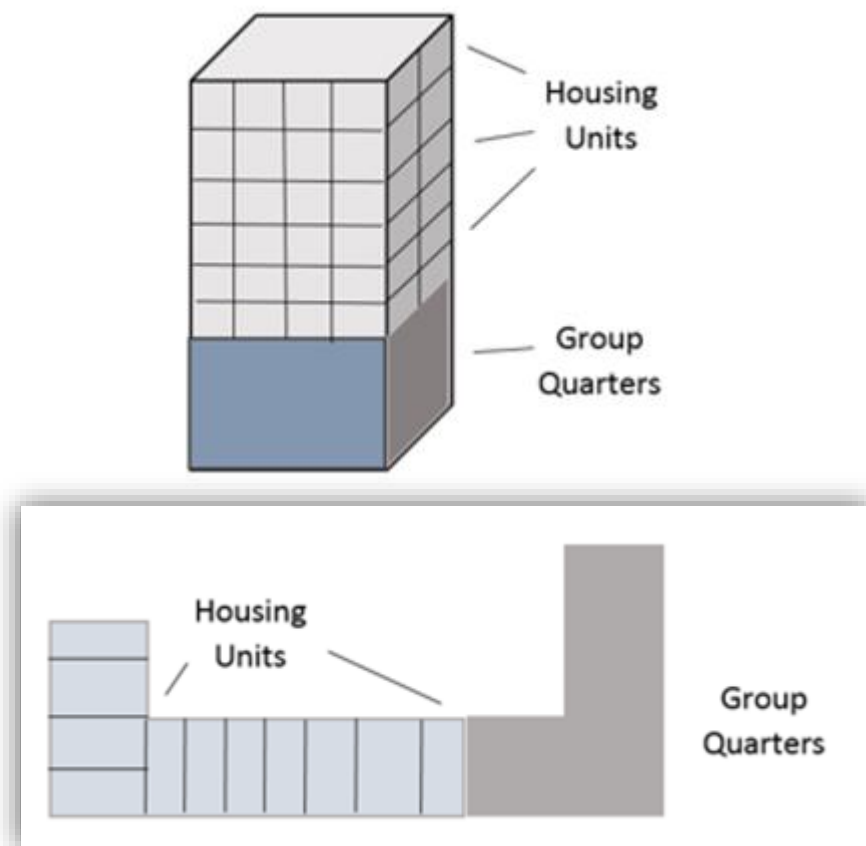
Credit: Census Bureau

What is a building? – U.S. EIA
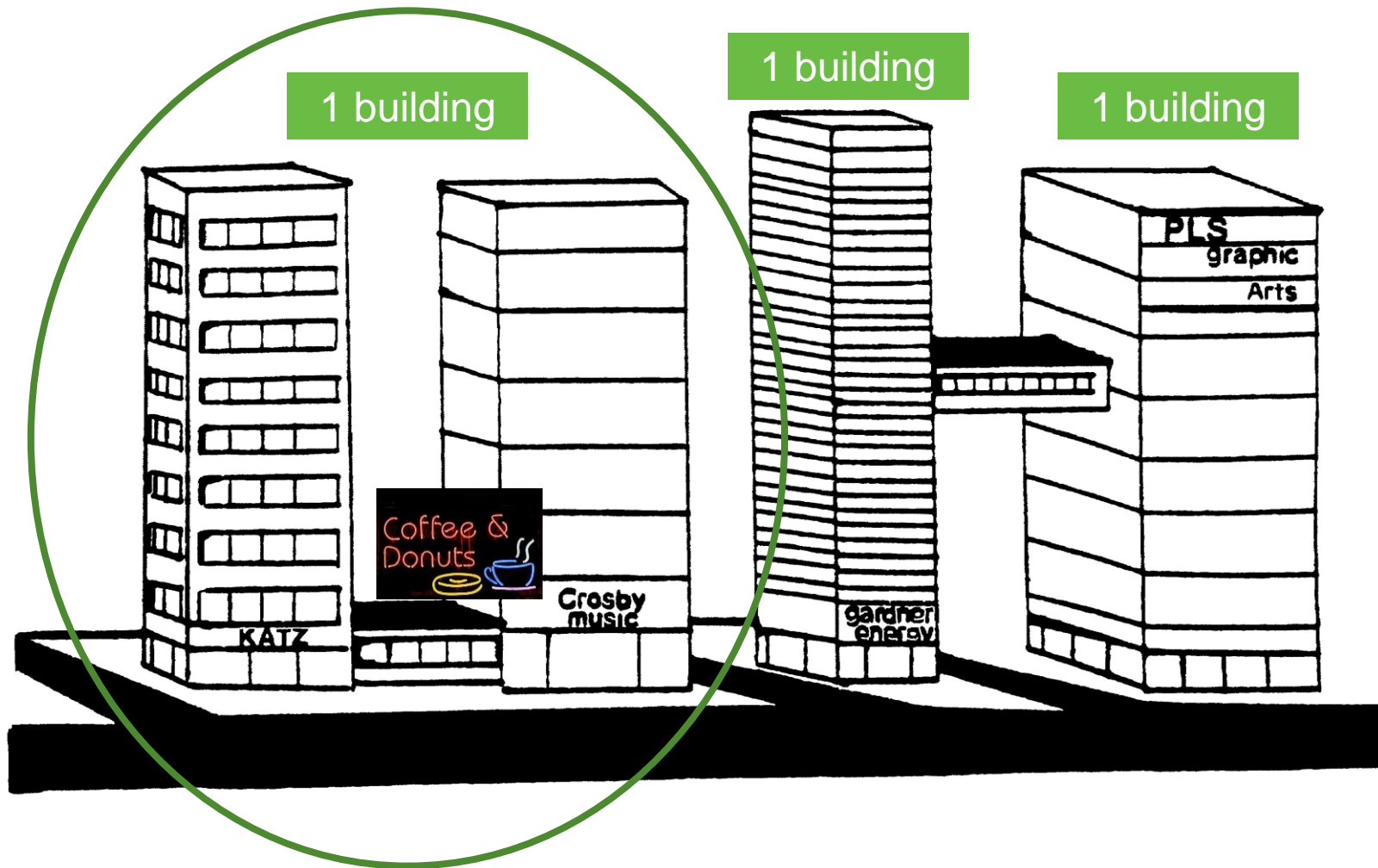
Credit: U.S. EIA

What are we actually identifying?

If there are so many ways to define a "building" what does that mean we are trying to identify?

Two buildings can't exist in the same place at the same time, so what if we uniquely identify the space a building occupies rather than the building itself?

We can uniquely identify an area in space using a Grid Reference System, an already established approach in the mapping field.

Open Location Code (OLC)
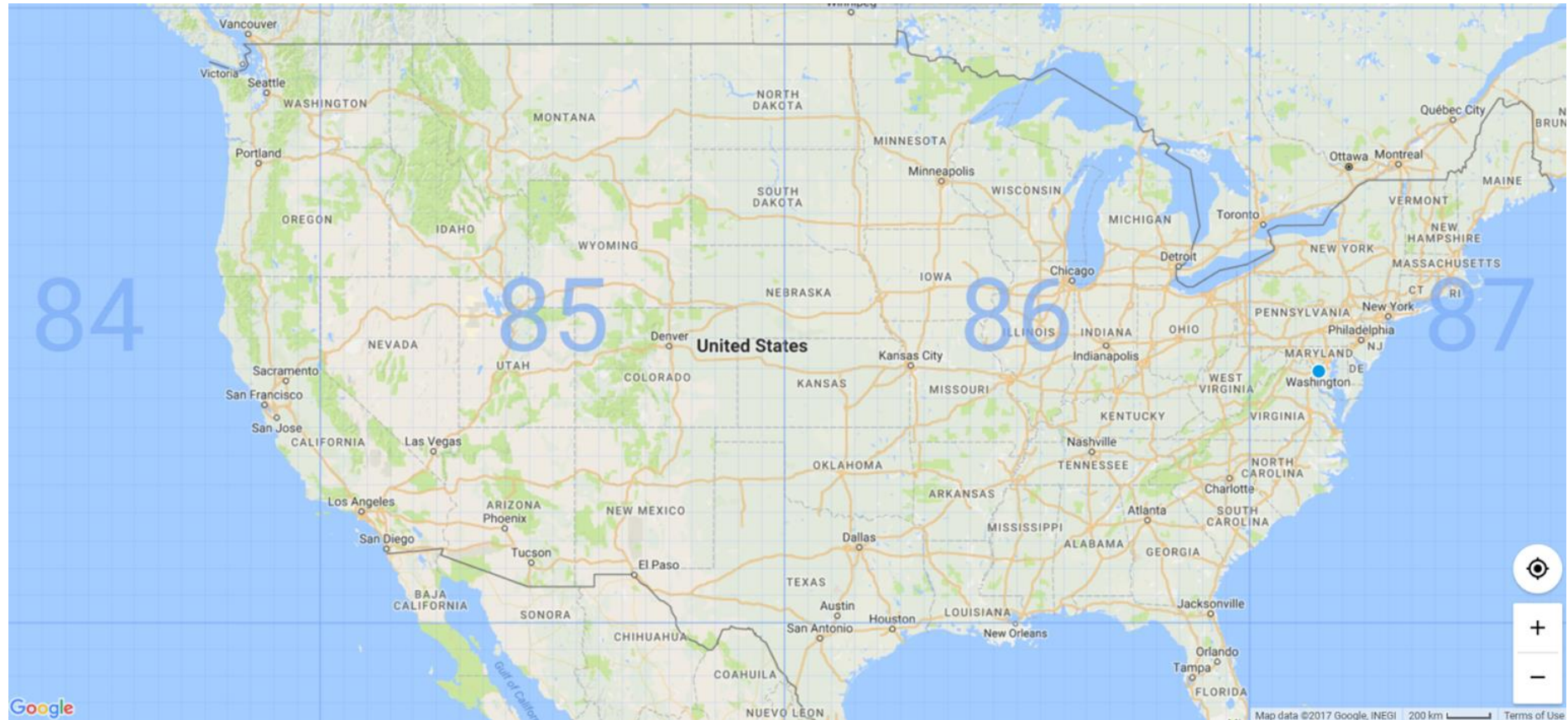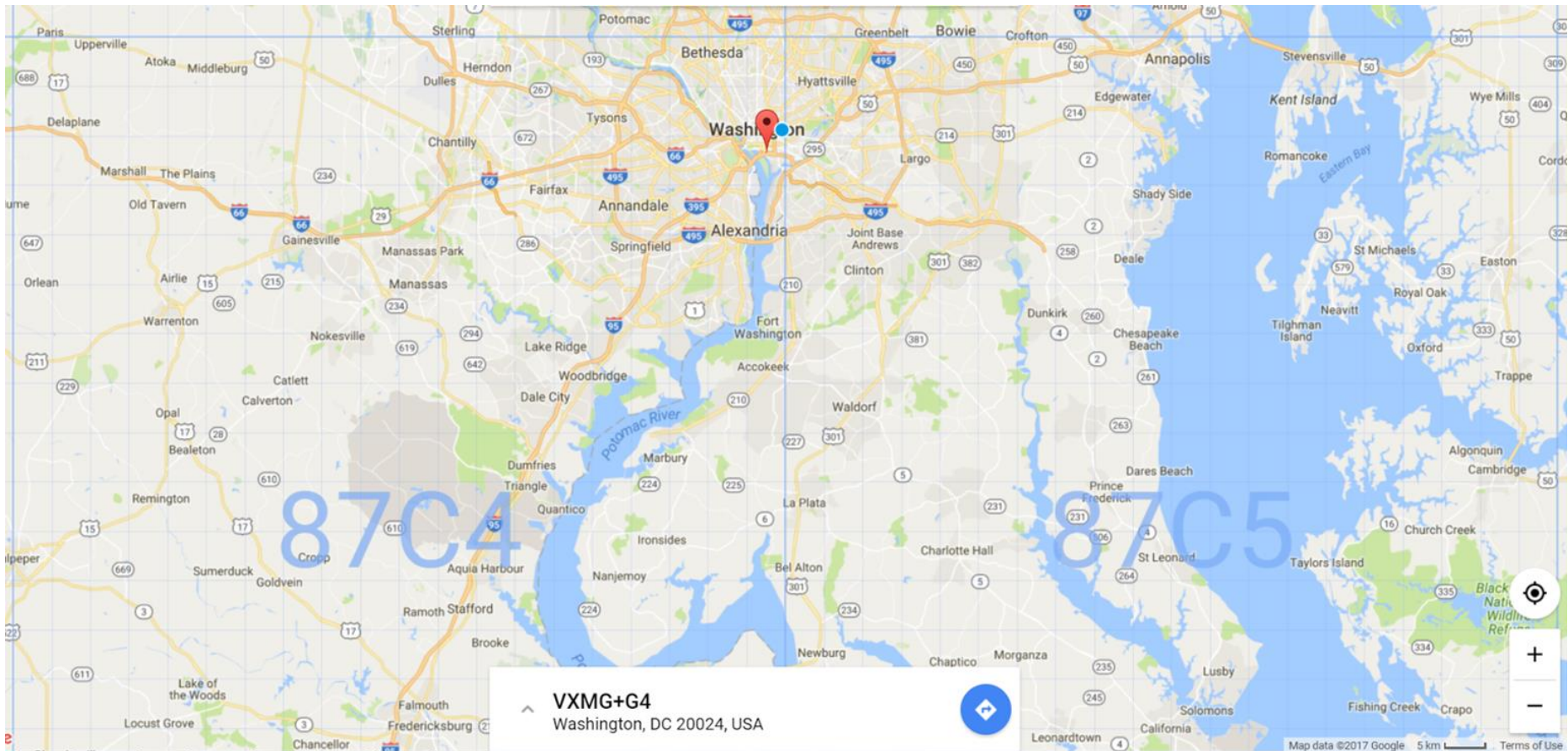
Open Location Code (OLC)

VXMG+G4
Washington, DC 20024, USA

Open Location Code (OLC)
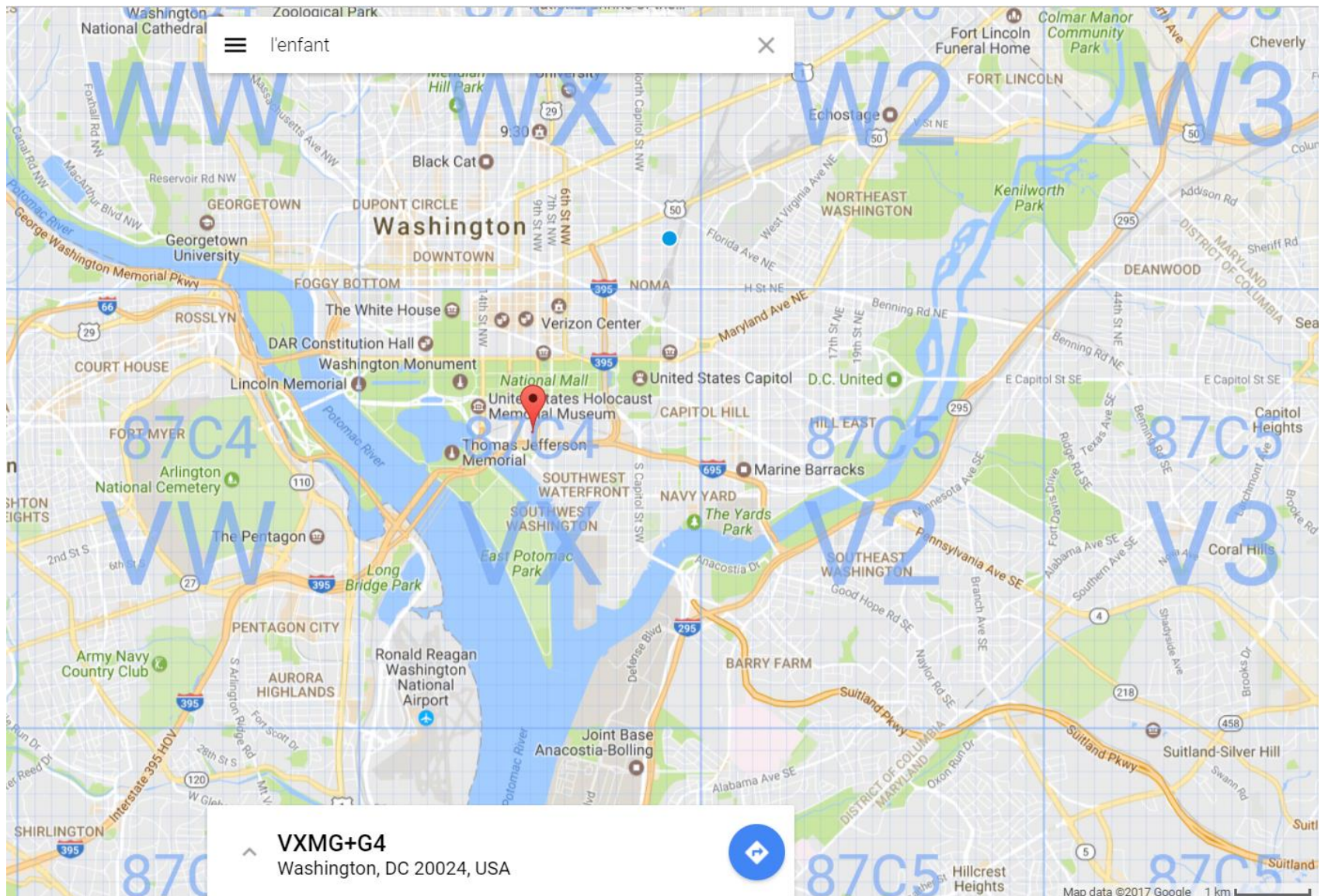
Open Location Code (OLC)
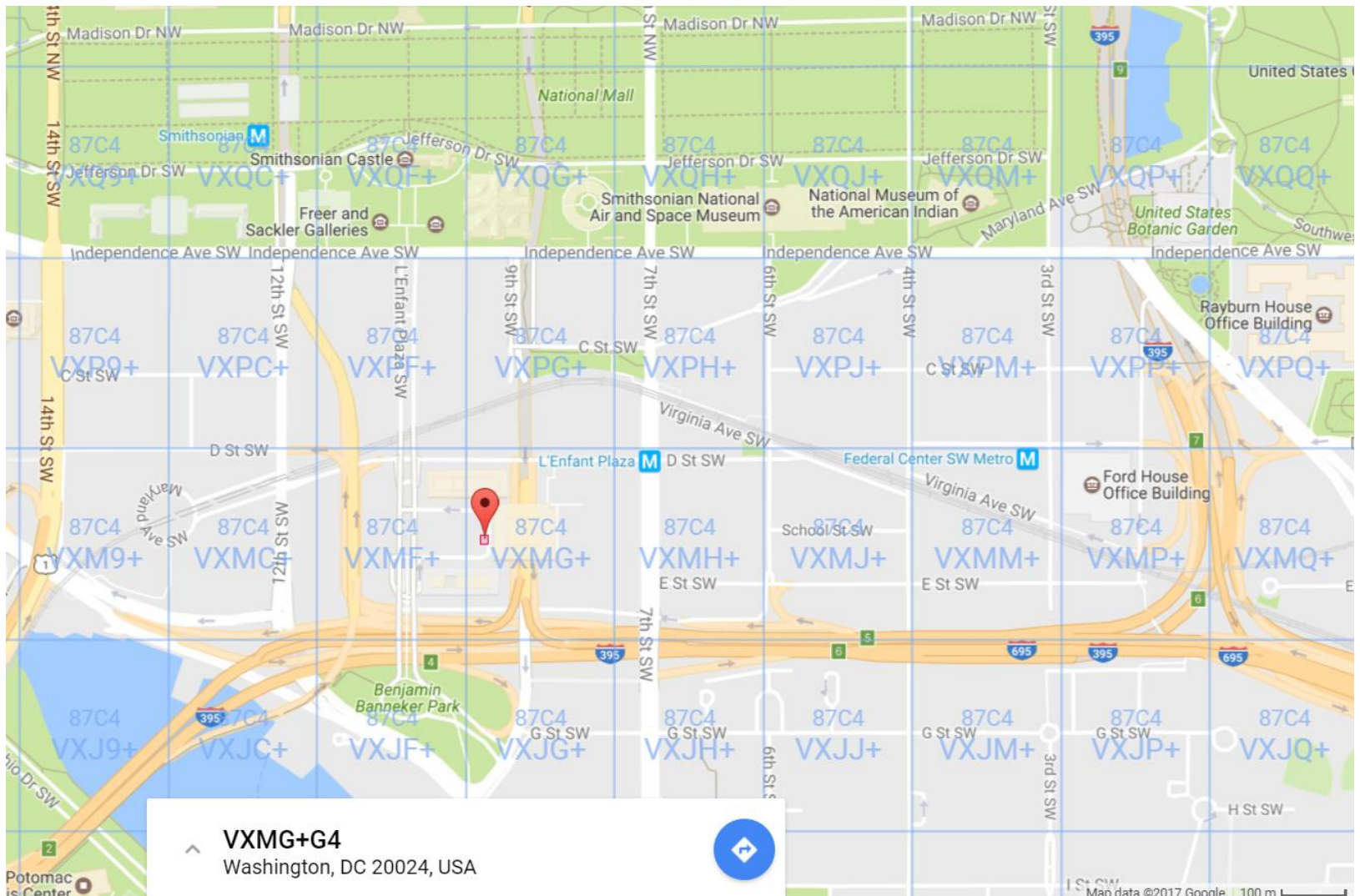
Open Location Code (OLC)
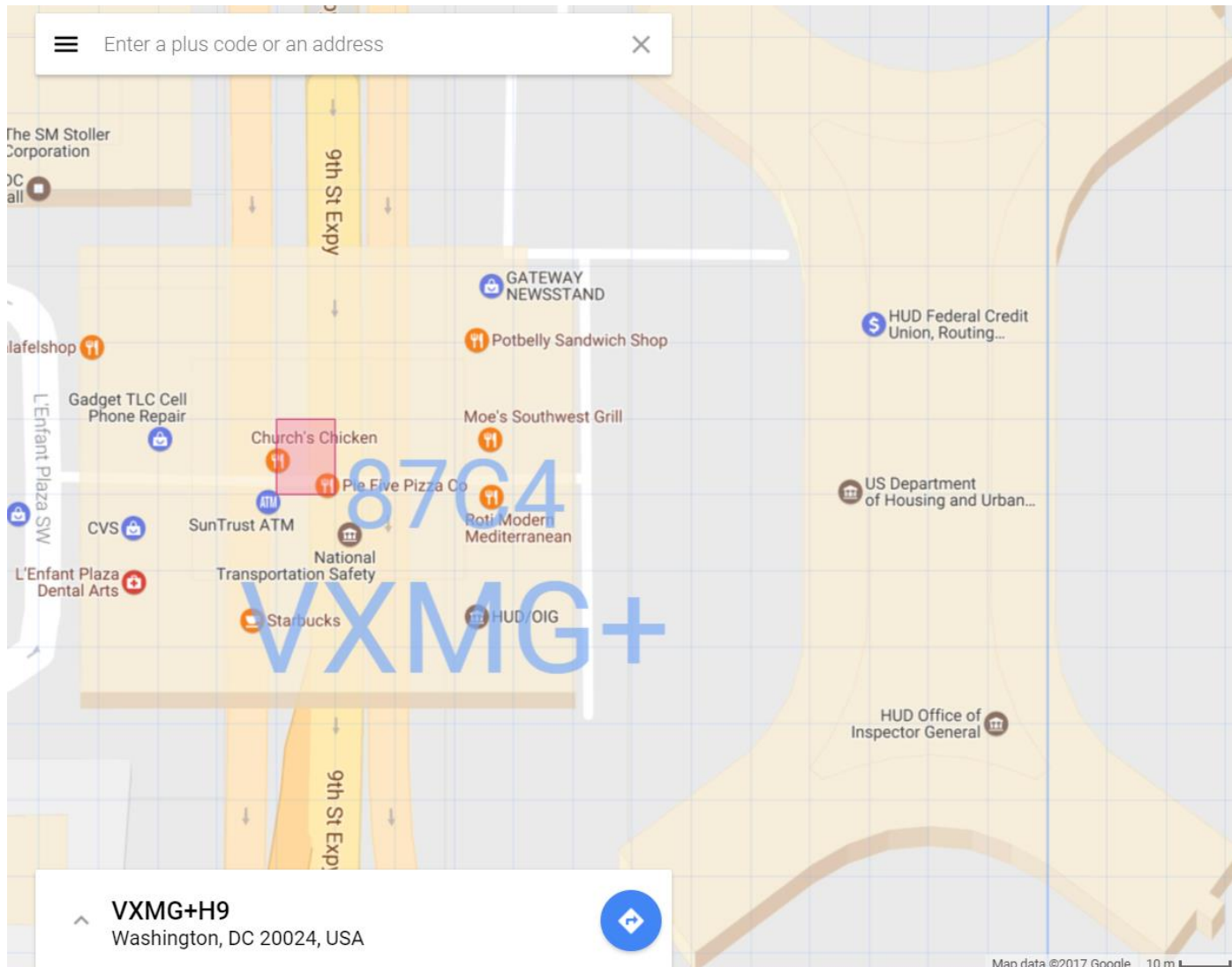
Open Location Code (OLC)
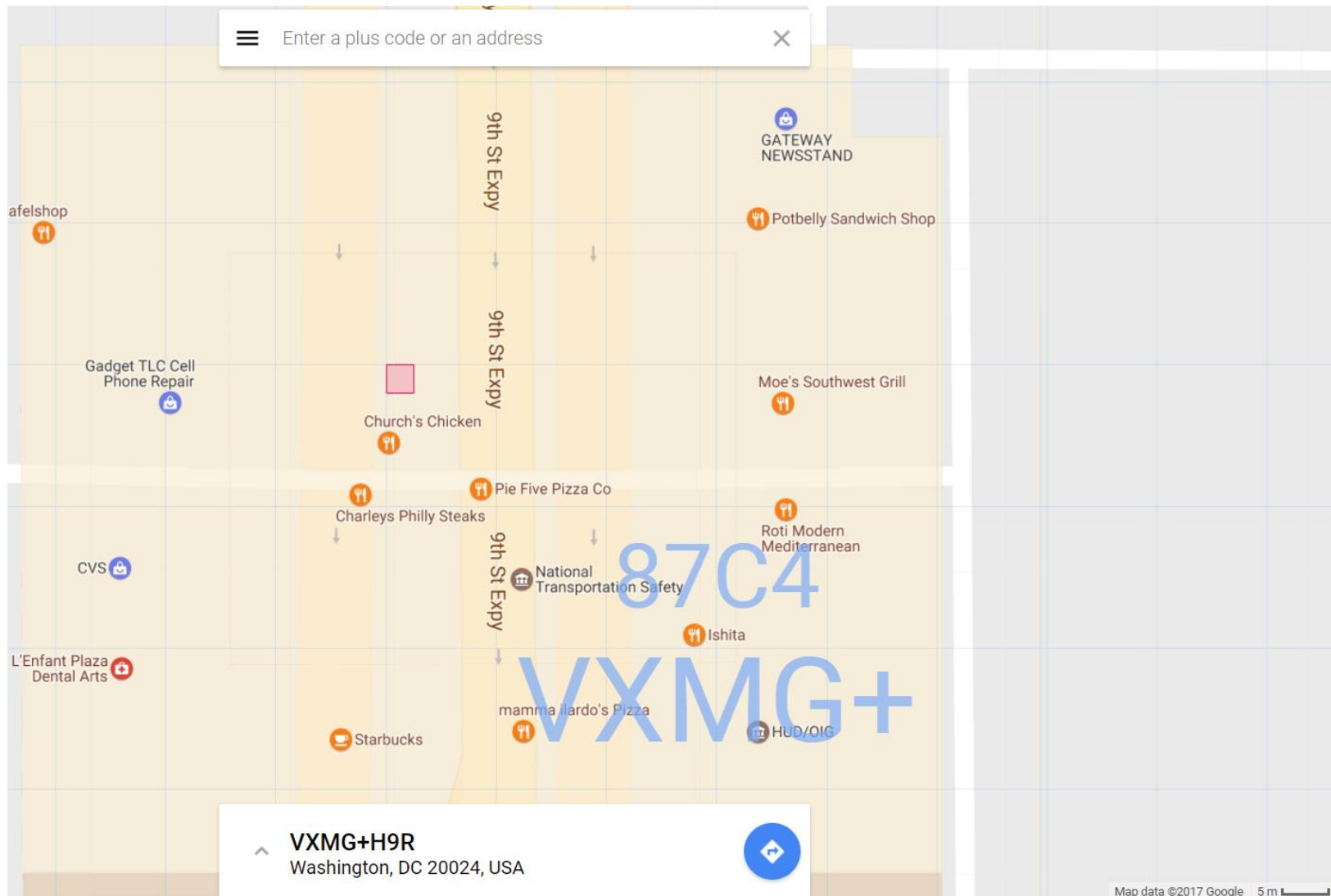
UNIQUE BUILDING IDENTIFICATION
Proof of Concept

Nora Wang, Ph.D.
Alex Vlachokostas, Ph.D.
Mark Borkum, Ph.D.

Pacific Northwest National Laboratory

Open Location Code

Plus codes are based on latitude and longitude – the grid that can be used to describe every point on the planet. By using a simpler code system, they end up much shorter and easier to use than traditional global coordinates.

A plus code in its full length is 10 characters long, with a plus sign before the last two. It consists of two parts:

- The first four characters are the area code, describing a region of roughly 100 x 100 kilometers.
- The last six characters are the local code, describing the neighborhood and the building, an area of roughly 14 x 14 meters – about the size of one half of a basketball court.

Each code uses these two parts to locate a larger region and then find the precise location within that region.

If the location is within or near a town, the area code isn't needed. In rural locations, even if the nearest town is up to 25 kilometers away, the area code isn't needed. You can use the local code together with the name of the town.

For those needing more precision, an additional, optional character can be used to improve accuracy to roughly 3 x 3 meters – about the size of a small car.

global code

8GC2CMXR+X6

area code   local code

R	V	W	X
J	M	P	Q
C	F	G	H
6	7	8	9
2	3	4	5

https://plus.code

UBID Methodology Overview

Step 1   Step 2   Step 3   Step 4   Step 5   Step 6
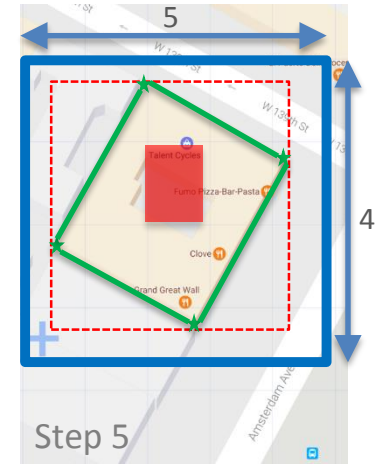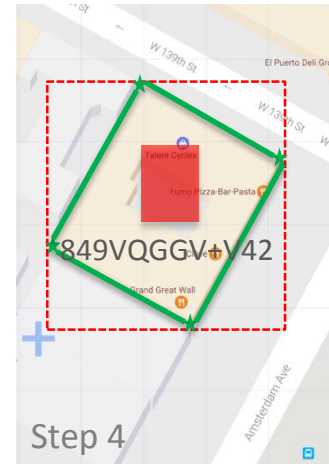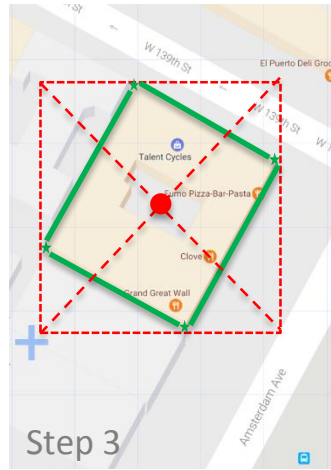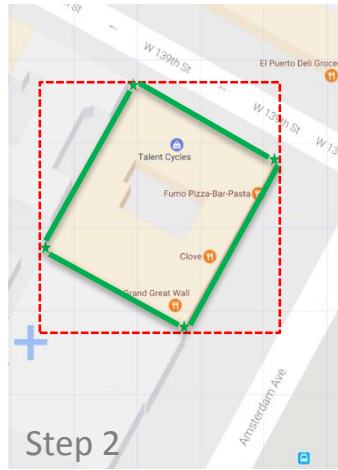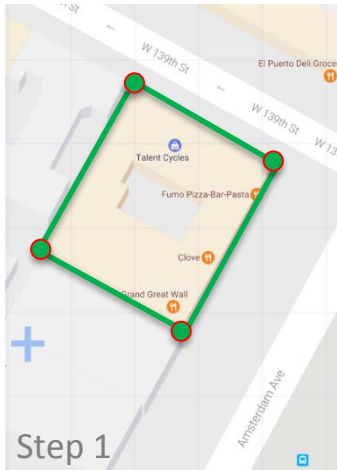
Step 1: Obtain building footprint polygon **(green box)** and calculate the lat/long of footprint corners

Step 2: Create a min bounding box that covers the footprint and along the grid

Step 3: Calculate the lat/long of the centroid the bounding box

Step 4: Convert the lat/long of the centroid to OLC reference cell **(red rectangle)**

Step 5: Calculate the W-E and N-S dimensions of the bounding box on OLC grid **(blue box)**

Step 6: Document the NW, SE OLCs of the bounding box **(blue rectangles)**

Step 9: Generate the UBID using the info from Steps 4, 5, 6.
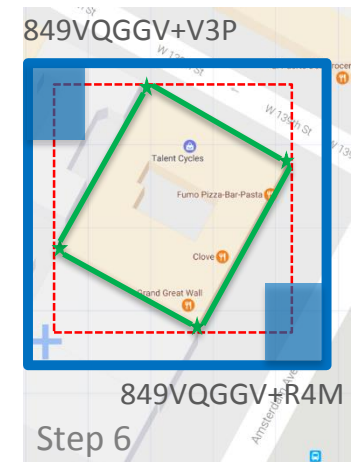
849VQGGV+V3P-849VQGGV+R4M	849VQGGV+V42-005-004
UBID (for Machine Implementation)	UBID Alias (Human Readable)
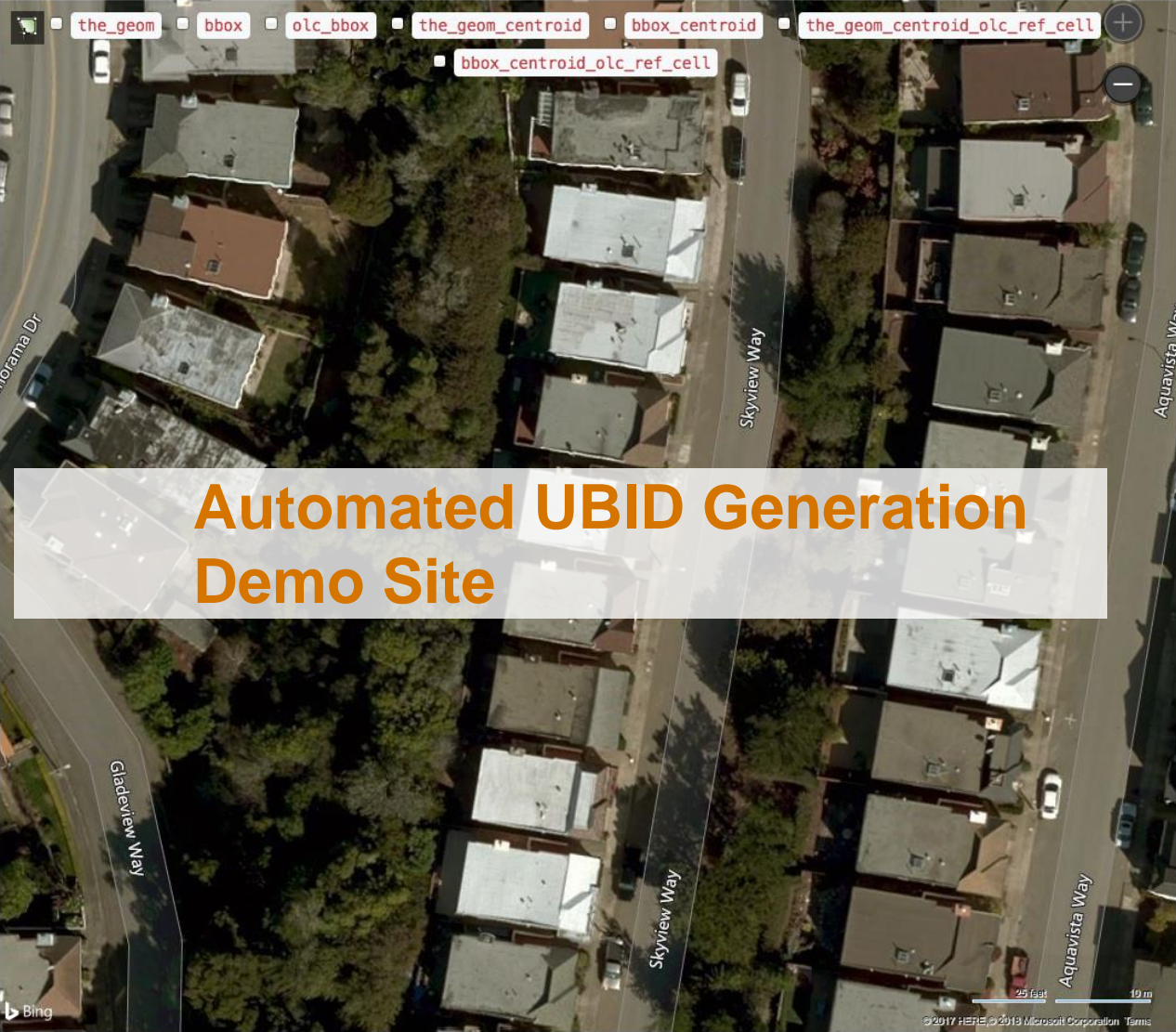
Automated UBID Generation Demo Site

Building footprint polygon in database

Minimum bounding box, centroid, & OLC

Snapped bounding box & UBID

UBID is © 2017 Pacific Northwest National Laboratory.   Back to top

You can also manually trace over map

Generate UBID and add it to the dataset

UBIDs for San Francisco, New York, and Chicago

- PNNL Demo site: https://ubid.emsl.pnl.gov
- Open Data from Cities
  - San Francisco
    - Building footprint polygons are available at: https://data.sfgov.org/Housing-and-Buildings/Building-Footprints/72ai-zege
    - 177,023 records
  - New York
    - Building footprint polygons are available at: https://data.cityofnewyork.us/Housing-Development/Building-Footprints/nqwf-w8eh/data
    - 101,831 records
  - Chicago
    - Building footprint polygons are available at: https://dev.socrata.com/foundry/data.cityofchicago.org/syp8-uezg
    - 820,606 records

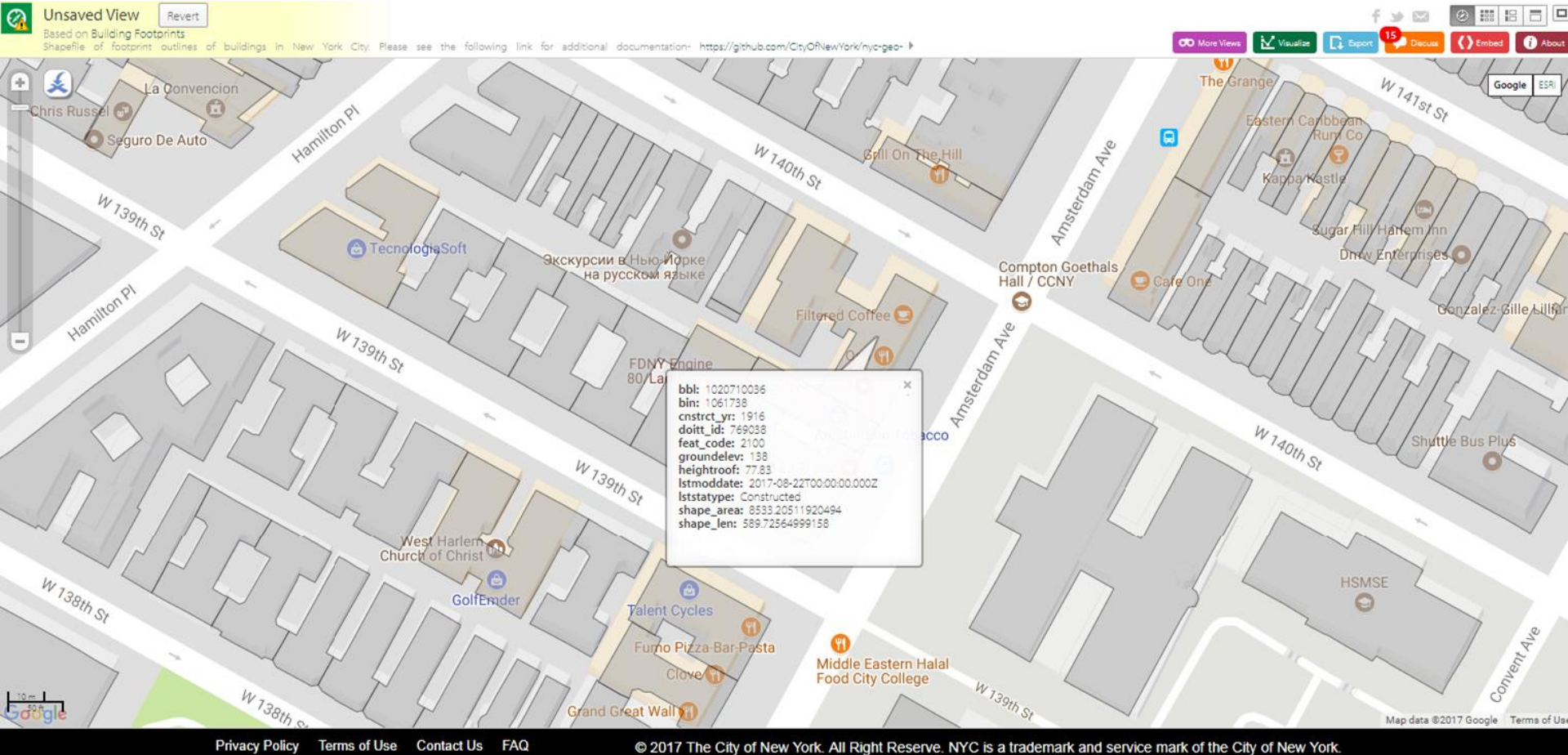# San Francisco Open Dataset
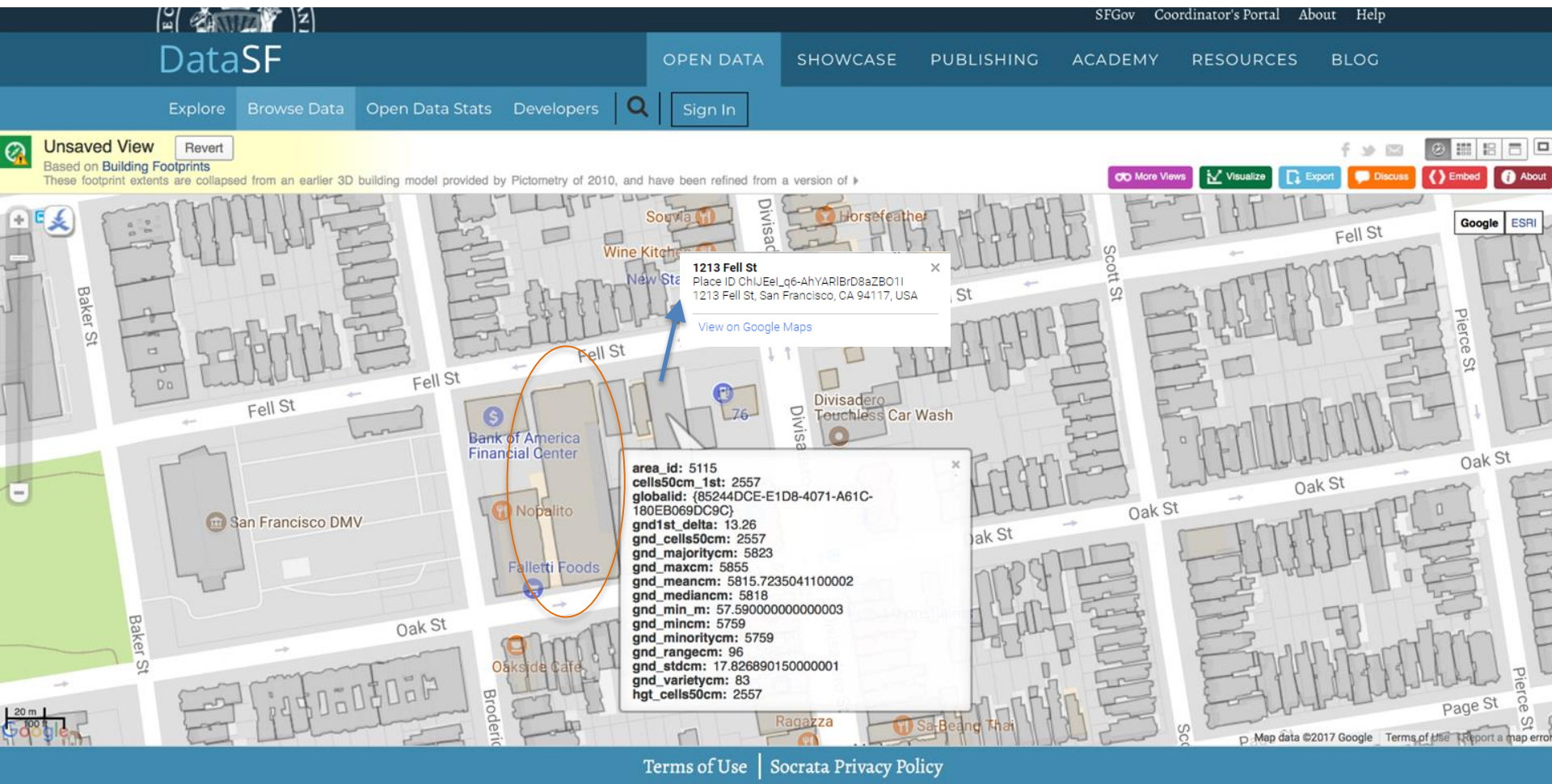
- Each building is identified with
  - **globalid**: 85244DCE-E1D8-4071-A61C-180EB069DC9C (database ID)
  - **MBLR:** SF1215014 (BLOCK= 1215, LOT=014) (Lot ID)
  - **Parcel(ID):** 1215/014
- Footprint polygon does not always fit over the Building footprint on the Google map

NYC Open Dataset

- Each building is identified with
  - **BBL**: 1020710036 (Borough= 1 (Manhattan); Block = 02071; Lot=0036) (Lot ID)
  - **BIN:** 1061738 (Building ID)
  - **doitt_id:** 769038 (Unique, numeric identifier that is assigned by DoITT; Not for all buildings in the dataset)
- Footprint polygon does not always fit over the Building footprint on the Google map

Chicago Open Dataset

- Each building is identified with
  - **Bldg_id**: 368011 (Building ID for internal use only)

UBID quick glance
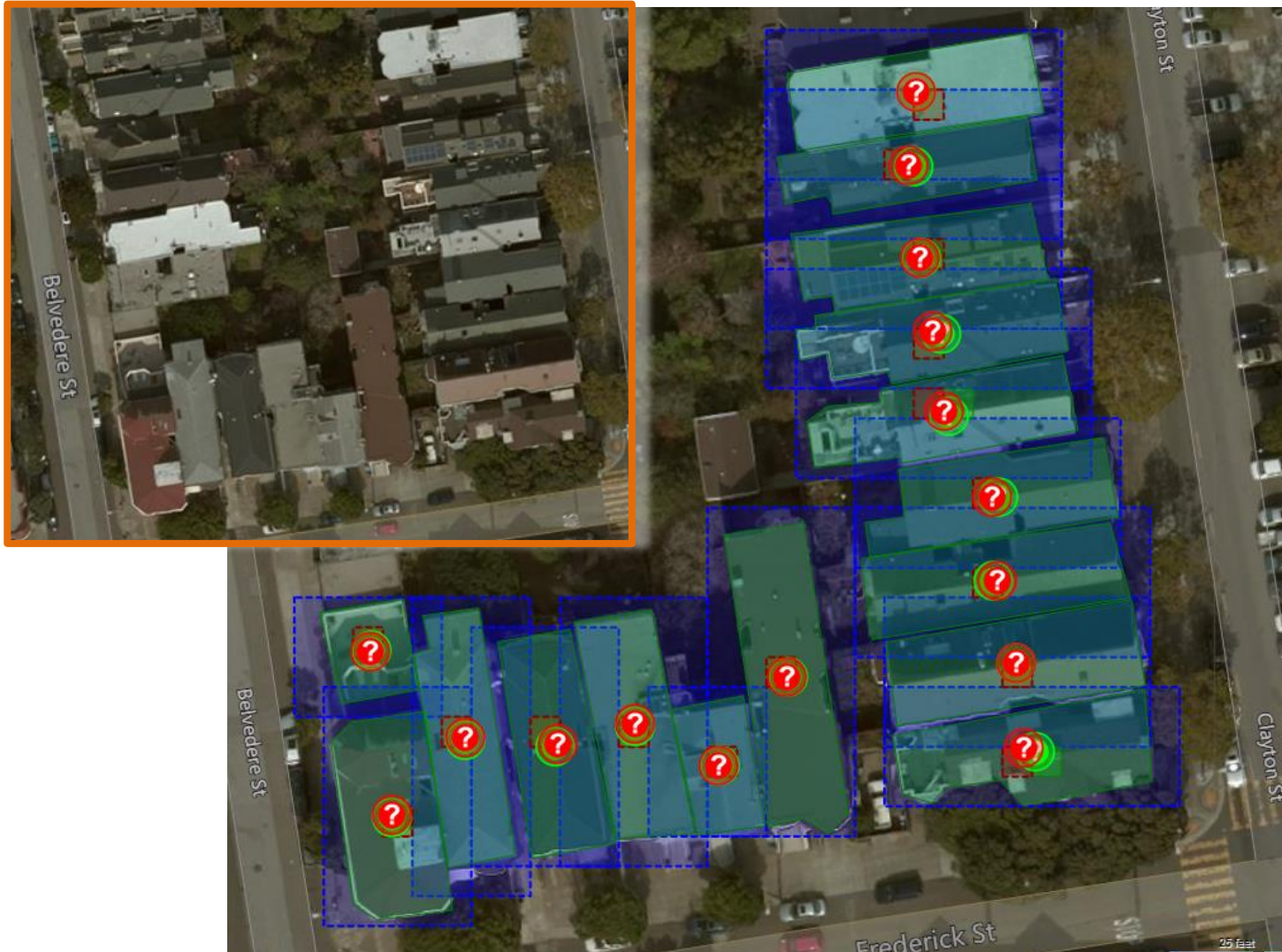
Research Questions

- Are there UBID duplicates in each dataset? Why (bugs, edge cases, bad data)?
  - Yes, but not THAT many. Mostly due to bad data.

- How does UBID reflect edge cases vs bad data? How to quantify the quality of the UBID dataset?
  - It does well, but the "quality" threshold varies by city.
- How do we know if a UBID is "good"?
  - Centroid distance
  - Bounding box scale
  - Building orientation

Duplicates Analysis

	UBID Alias	UBID	Local Building ID	Local Lot ID
	OLC (C) − h − w	OLC (NW) − OLC (SE)		
Chicago	106	22	4	
NYC	0	0	6	6945
San Francisco	72	6	0	44452

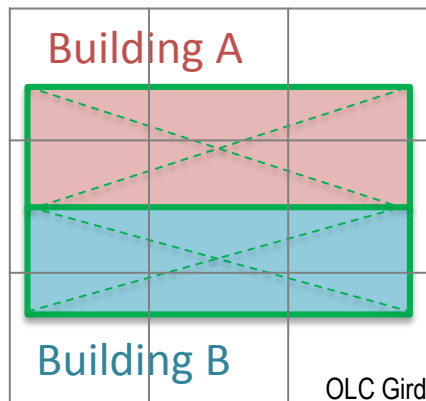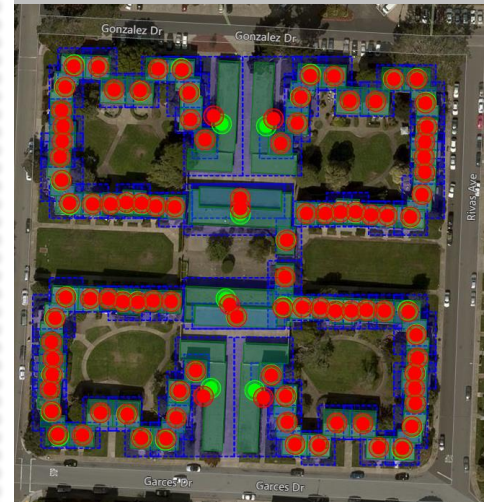**Why does UBID Alias have more duplicates?**
- In our test sets, most are bad data.
- When the buildings are very small (e.g., <30 sq.m./323 sq.ft), it is possible for Building A and Building B to have the same UBID Alias, although their UBIDs are different.

Building A

Building B

OLC Gird

Duplicates caused by bad data

Multiple Buildings on a Lot
E.g. MBLR == "SF7336001"

849VPG6Q+Q3F-2-2

849VPGR2+W3W-2-2

849VPH7Q+886-2-2

849VQH57+WV2-2-4

849VPGVC+2MR-2-2

849VQJ37+WR9-3-2

849VPG4R+W7H-2-2

849VQG7F+8PR-2-2

All shown here are actually bad data.
We can flag the UBID Alias that contains H or W value smaller than 3, and require manual verification of data errors.

Local Building ID duplicates due to data error (example 1):
Same Local Building ID, Different UBID (Chicago)

86HJW947+QR6-
86HJW947+RJM-
86HJW947+PX4

UBID
86HJW947+QR6-86HJW947+RJM-
86HJW947+PX4

UBID (OLC Grid Units)
86HJW947+QR6-14-30

BLDG_ID
882804

86HJX6CR+9R5-
86HJX6CR+CR4-
86HJX6CR+8V7

UBID
86HJX6CR+9R5-86HJX6CR+CR4-
86HJX6CR+8V7

UBID (OLC Grid Units)
86HJX6CR+9R5-10-4
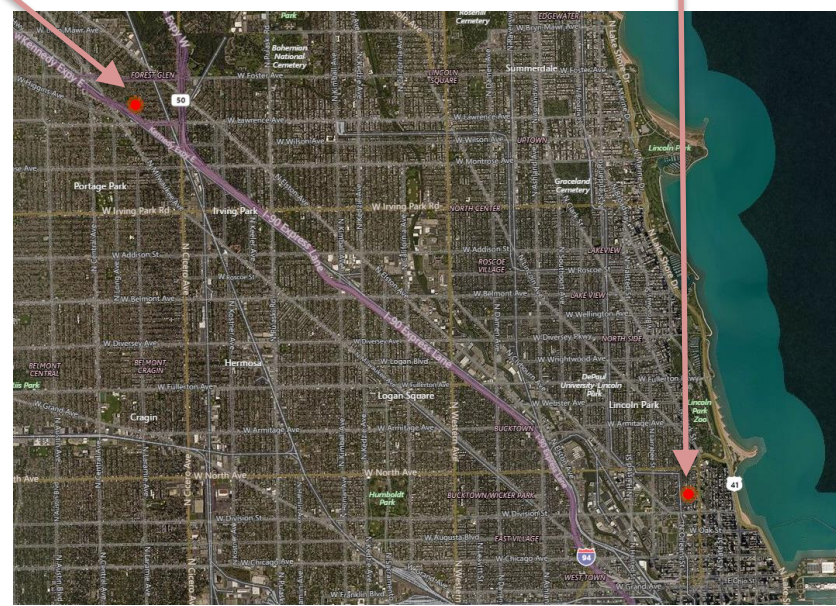
BLDG_ID
882804

Local Building ID duplicates due to data error (example 2):
Same Local Building ID, Different UBID (Chicago)

86HJX4FM+CJ9-
86HJX4FM+GH2-
86HJX4FM+8PG

UBID
86HJX4FM+CJ9-86HJX4FM+GH2-
86HJX4FM+8PG

UBID (OLC Grid Units)
86HJX4FM+CJ9-19-15

BLDG_ID
0

86HJV989+MW6-
86HJV989+MW6-
86HJV989+MW7

UBID
86HJV989+MW6-86HJV989+MW6-
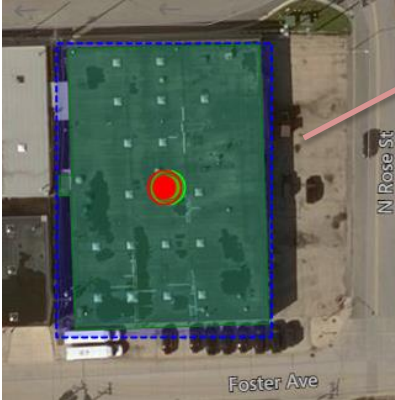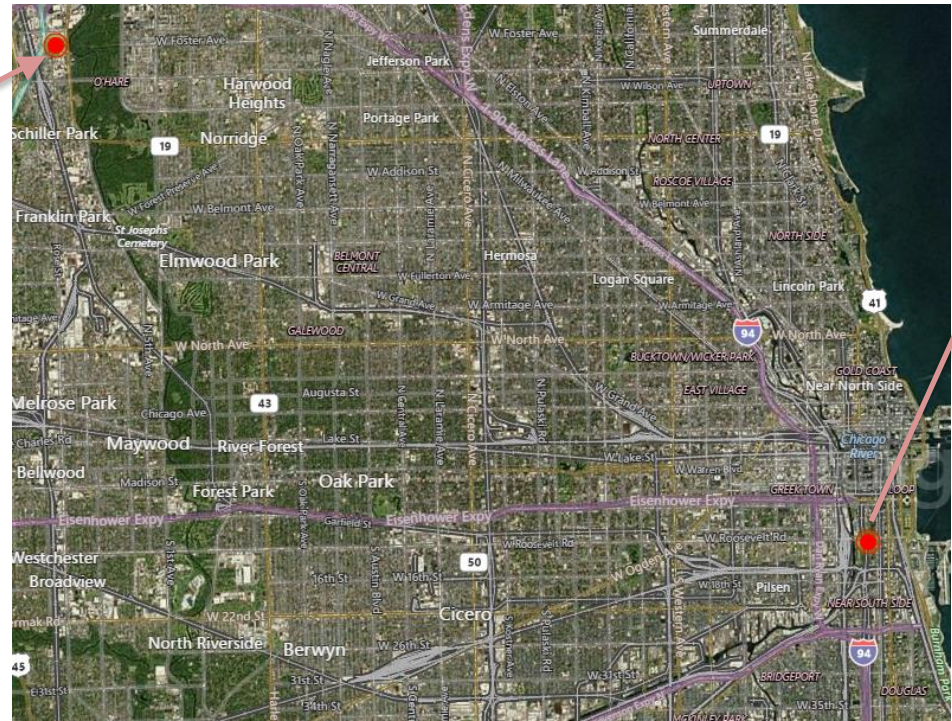86HJV989+MW7

UBID (OLC Grid Units)
86HJV989+MW6-1-2

BLDG_ID
0

Local Building ID duplicates due to data error (example 3):
Same BIN, Different UBID (NYC)

Local Building ID duplicates due to data error (example 4):
Same BIN, Different UBID (NYC)

87G8P3C6+P49-
87G8P3C6+P4V-
87G8P3C6+M5R

87G8P3C6+P49-87G8P3C6+P4V-
87G8P3C6+M5R

UBID (OLC Grid Units)
87G8P3C6+P49-6-4

BBL
3028340018

BIN
3000000

87G8J2XX+X7F-
87G8M22X+26M-
87G8J2XX+W8C

87G8J2XX+X7F-87G8M22X+26M-
87G8J2XX+W8C

UBID (OLC Grid Units)
87G8J2XX+X7F-12-8
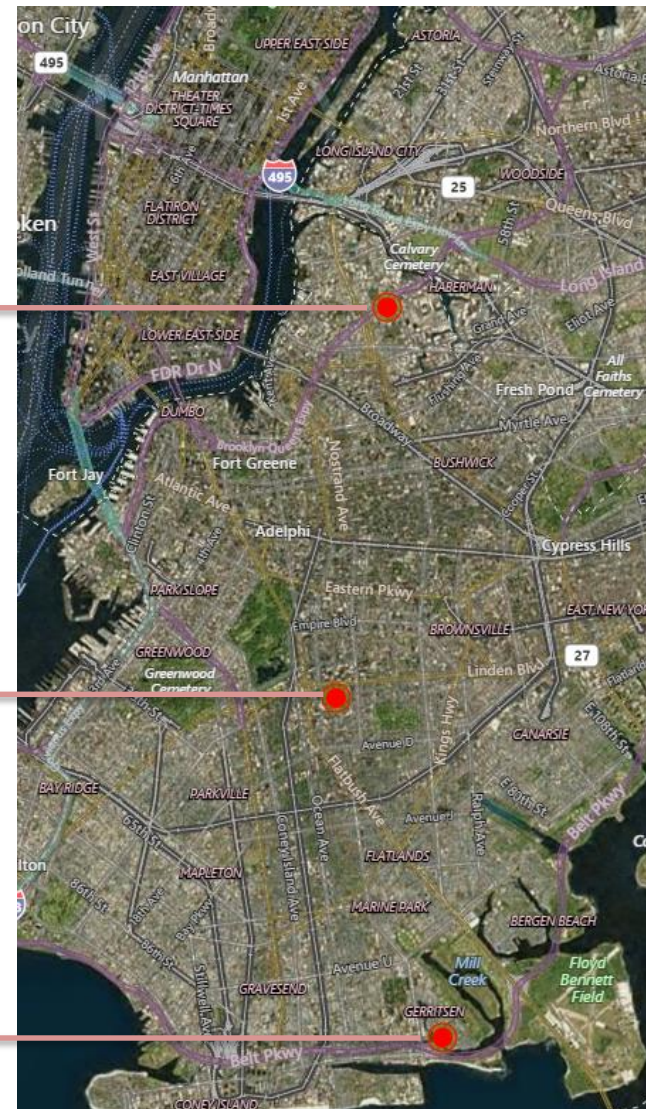
BBL
3051050076

BIN
3000000

87G8H3PF+XHR-
87G8H3QF+2GF-
87G8H3PF+XHG

87G8H3PF+XHR-87G8H3QF+2GF-
87G8H3PF+XHG

UBID (OLC Grid Units)
87G8H3PF+XHR-6-6

BBL
3088661760

BIN
3000000

UBID duplicates due to data error (example 1):
Same UBID, Different Local ID (Chicago)

Different Local ID

UBID duplicates due to data error (example 2):
Same UBID, Different Local ID (Chicago)

Different Local ID

UBID duplicates due to data error (example 3):
Same UBID, Different Local ID (Chicago)

Different Local ID

UBID duplicates due to data error (example 4):
Same UBID, Different Local ID (Chicago)

Different Local ID

UBID duplicates due to data error (example 5):
Same UBID, Different Local ID (San Francisco),
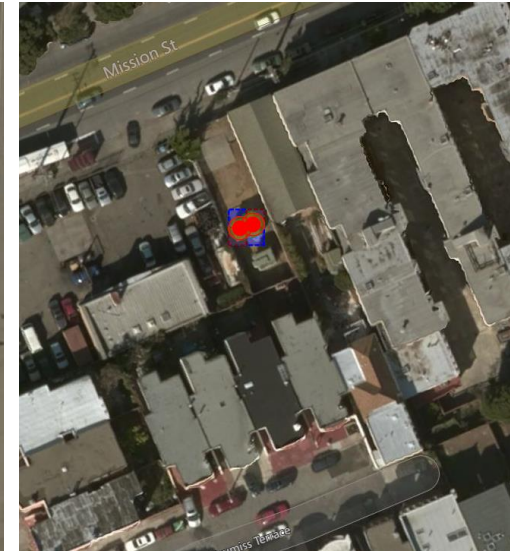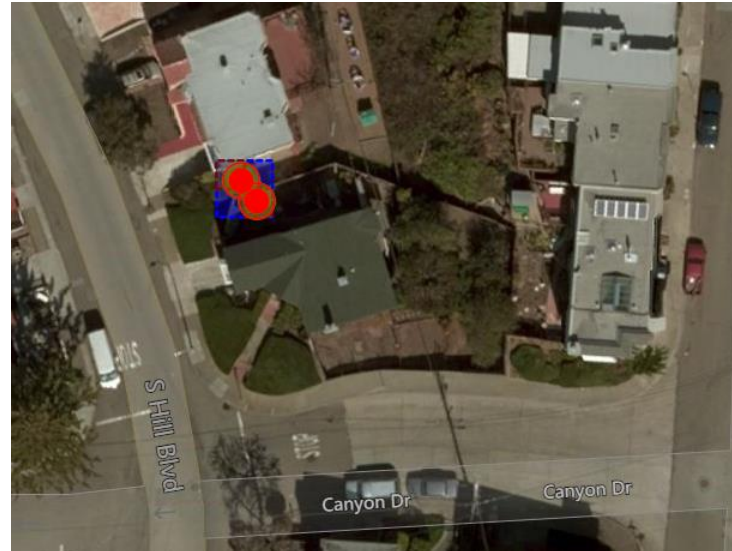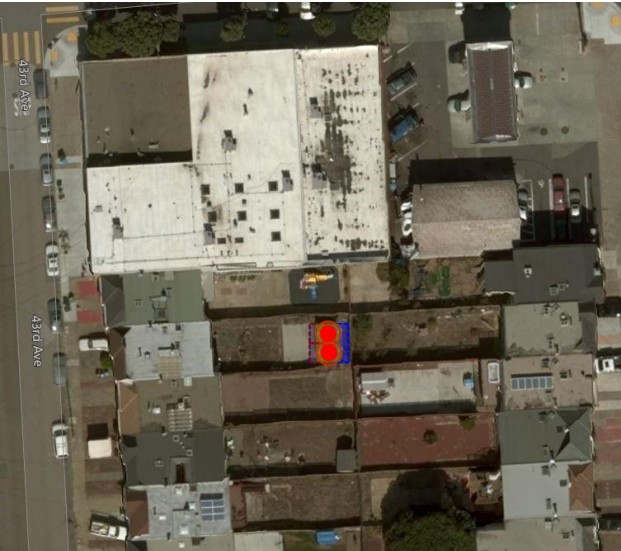
WHERE:
olc_bbox_centroid_UBID == "849VQF4X+G6W-849VQF4X+H63-849VQF4X+G6W"

WHERE:
olc_bbox_centroid_UBID == "849VPH5C+M32-849VPH5C+M32-849VPH5C+J3V"

WHERE:
olc_bbox_centroid_UBID == "849VPG5X+G68-849VPG5X+G67-849VPG5X+G64"

Edge Case Analysis

- Quantify the **distance (absolute and relative) between** the Shape Centroid and the Bounding Box Centroid

- Quantify the **size difference** between the OLC Bounding Box and Building Footprint Shape

- Quantify the **rotation** of the bounding box from the building orientation
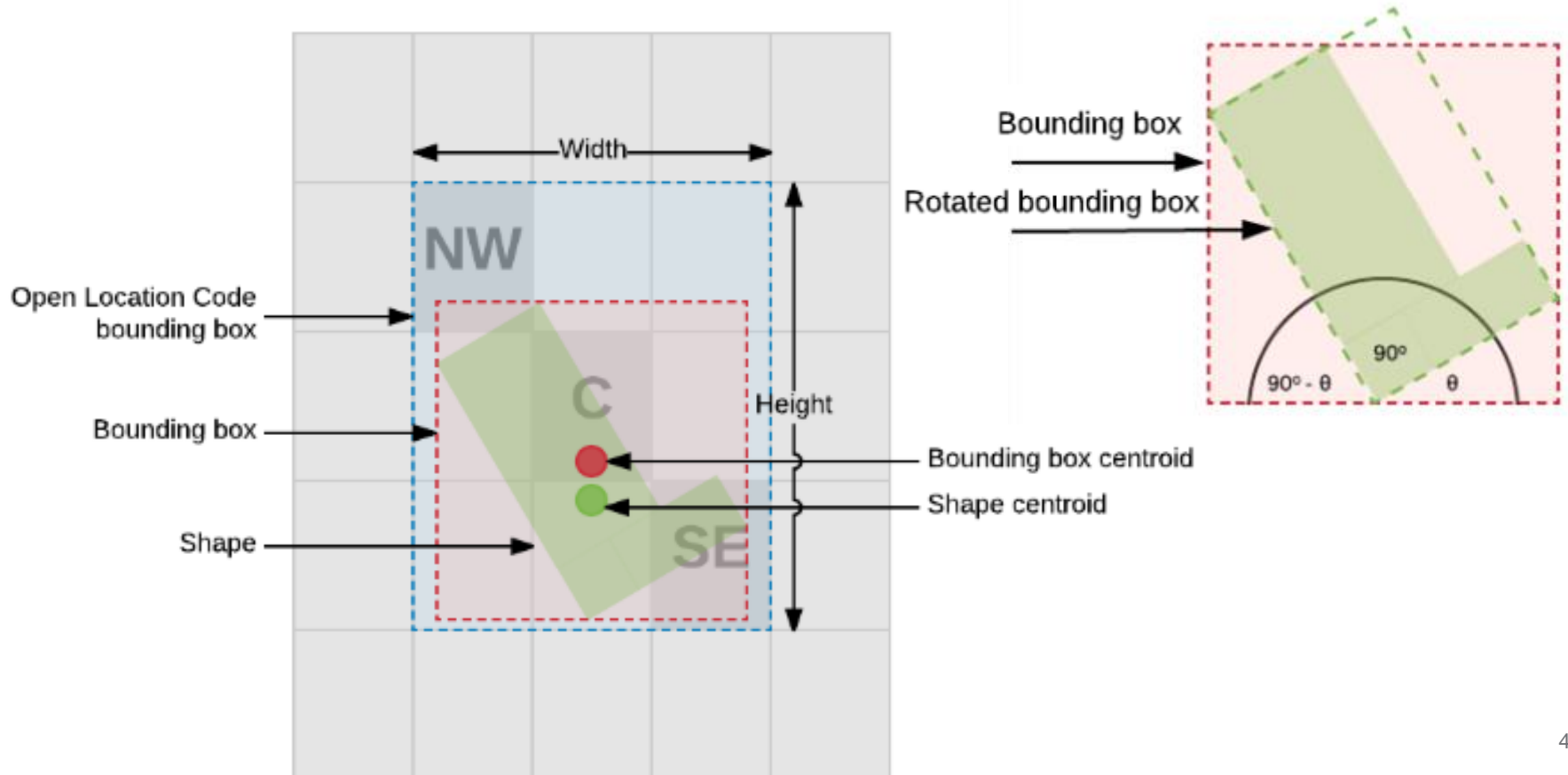
Centroid Offset: Absolute Distance
San Francisco
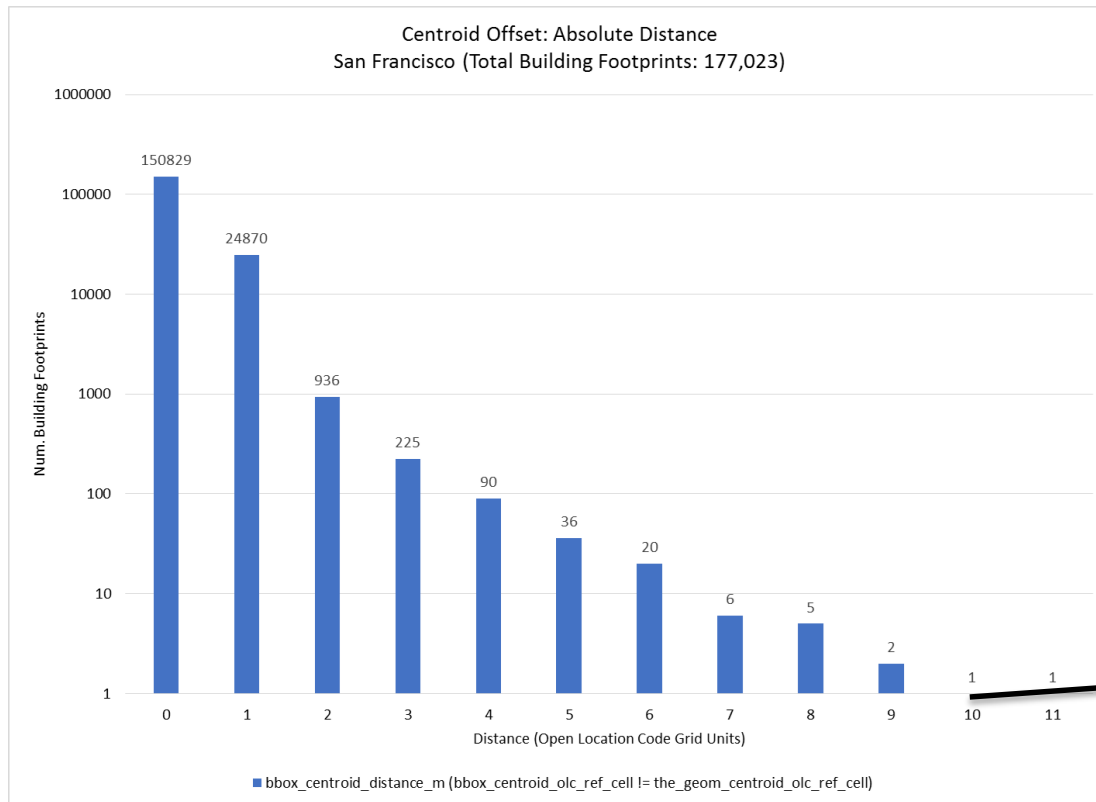
Centroid Offset: Absolute Distance
San Francisco (Total Building Footprints: 177,023)

WHERE:
(bbox_centroid_olc_ref_cell !=
the_geom_centroid_olc_ref_cell) and
(bbox_centroid_distance_m >= 30) and
(bbox_centroid_distance_m < 33)

Source: https://stash.pnnl.gov/users/bork374/repos/ubid/browse/assets/data.sfgov.org/2s2t-jwzp/centroid_comparison.csv
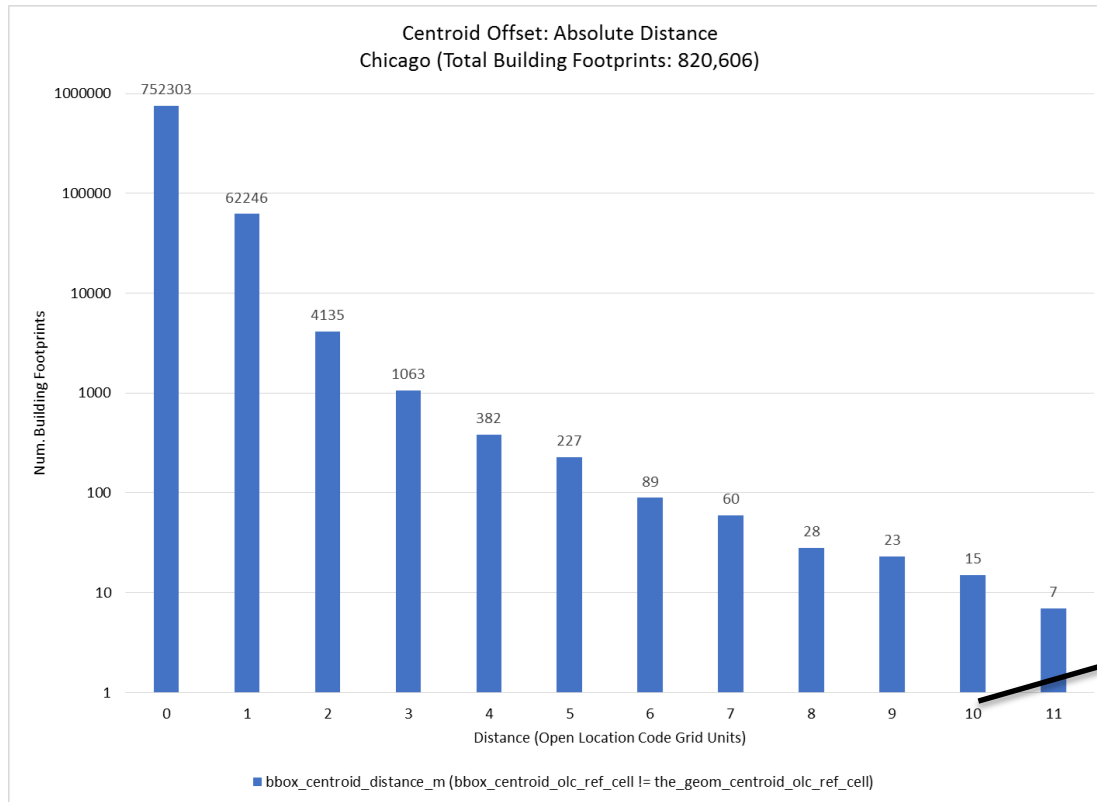
Centroid Offset: Absolute Distance
Chicago

Centroid Offset: Absolute Distance
Chicago (Total Building Footprints: 820,606)

Bar chart values by Distance (Open Location Code Grid Units):
- 0: 752303
- 1: 62246
- 2: 4135
- 3: 1063
- 4: 382
- 5: 227
- 6: 89
- 7: 60
- 8: 28
- 9: 23
- 10: 15
- 11: 7

Y-axis: Num. Building Footprints
X-axis: Distance (Open Location Code Grid Units)

Legend: bbox_centroid_distance_m (bbox_centroid_olc_ref_cell != the_geom_centroid_olc_ref_cell)

WHERE:
(bbox_centroid_olc_ref_cell != the_geom_centroid_olc_ref_cell)
and (bbox_centroid_distance_m >= 30) and
(bbox_centroid_distance_m < 33)

Source: https://stash.pnnl.gov/users/bork374/repos/ubid/browse/assets/data.cityofchicago.org/syp8-uezg/centroid_comparison.csv

Centroid Offset: Absolute Distance
New York City
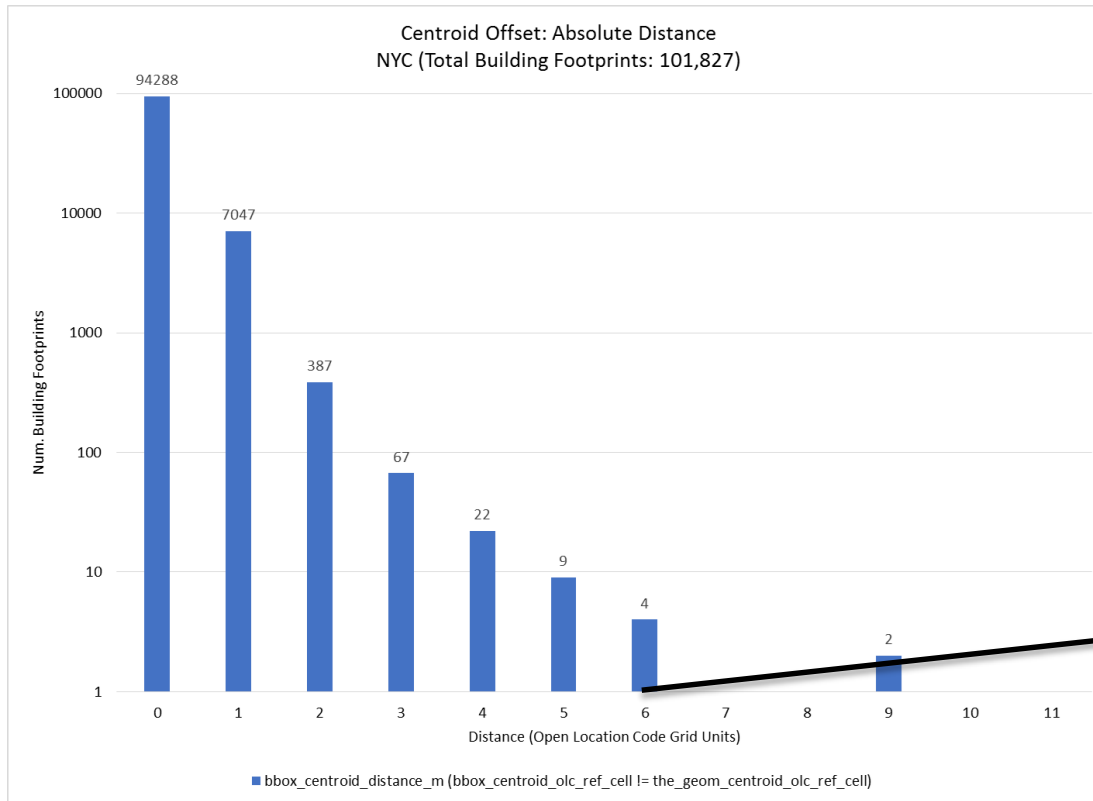
Centroid Offset: Absolute Distance
NYC (Total Building Footprints: 101,827)

WHERE:
(bbox_centroid_olc_ref_cell !=
the_geom_centroid_olc_ref_cell) and
(bbox_centroid_distance_m >= 15) and
(bbox_centroid_distance_m < 18)

Source: https://stash.pnnl.gov/users/bork374/repos/ubid/browse/assets/data.cityofnewyork.us/dszh-ae6r/centroid_comparison.csv

Centroid Offset: Normalized Distance

San Francisco
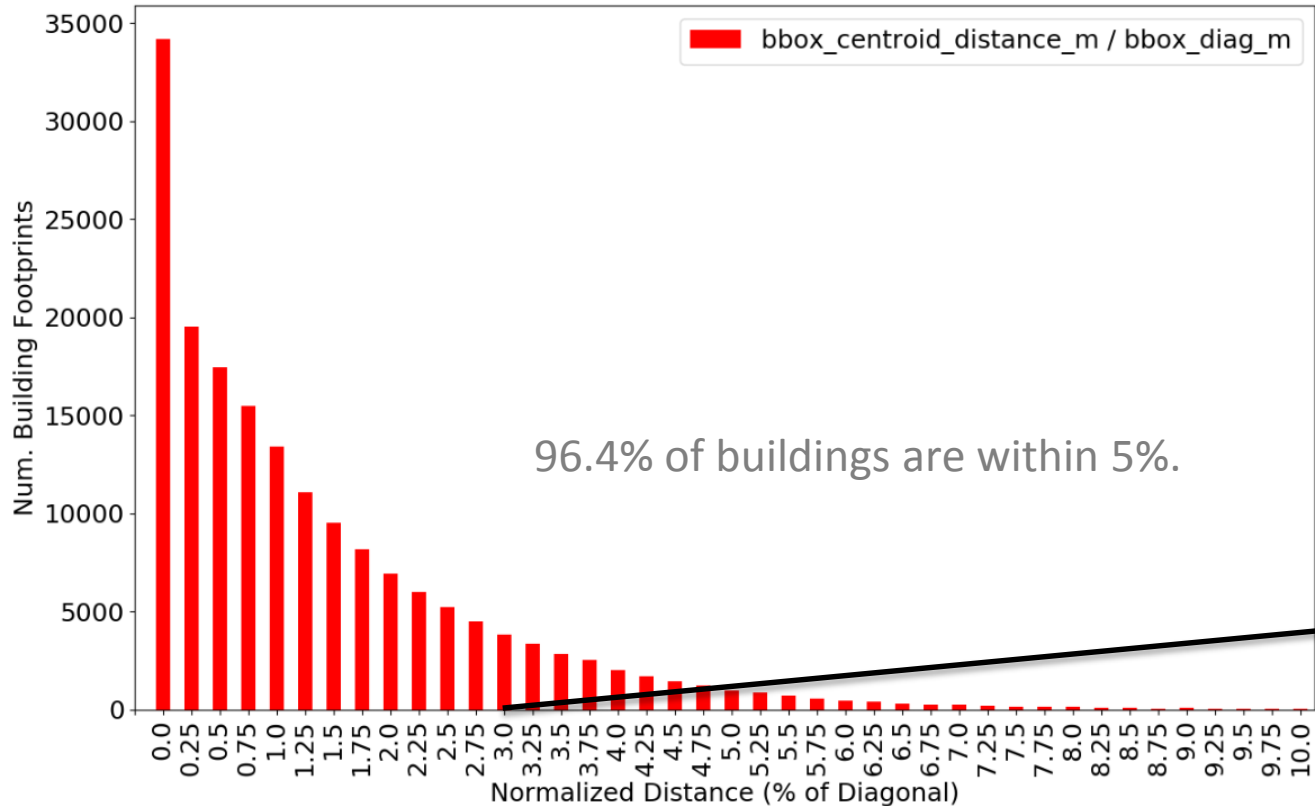
849VPGW2+CR2-
849VPGW2+CQW-
849VPGW2+9RF

96.4% of buildings are within 5%.

SELECT:
bbox_centroid_distance_m /
bbox_diag_m

AS:
x

WHERE:
(x >= 0.03) and (x < 0.0325)

Source: https://stash.pnnl.gov/users/bork374/repos/ubid/browse/assets/data.sfgov.org/2s2t-jwzp/bbox_centroid_distance_m.diag.0.png

Centroid Offset: Normalized Distance

Chicago
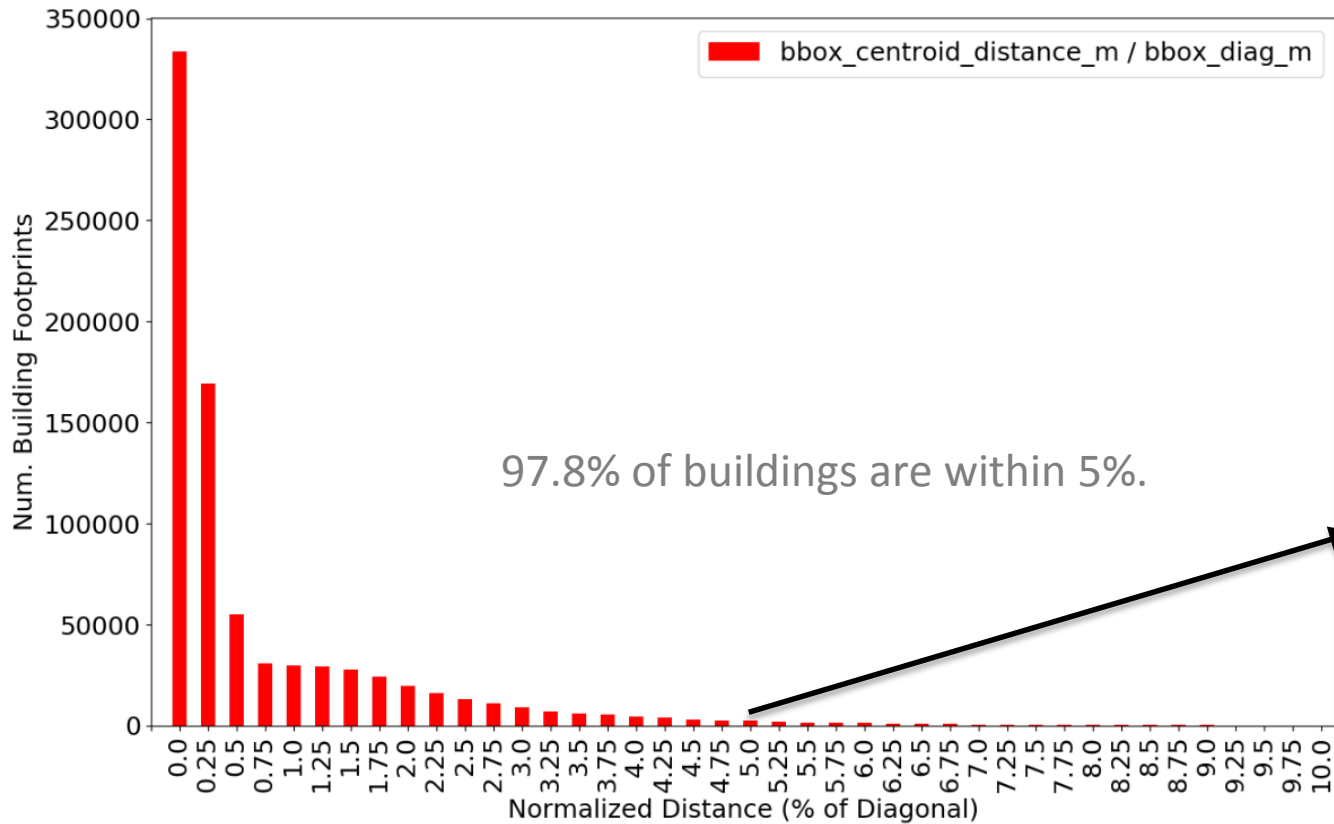
86HJX8CX+7C3-
86HJX8CX+79P-
86HJX8CX+6FC

97.8% of buildings are within 5%.
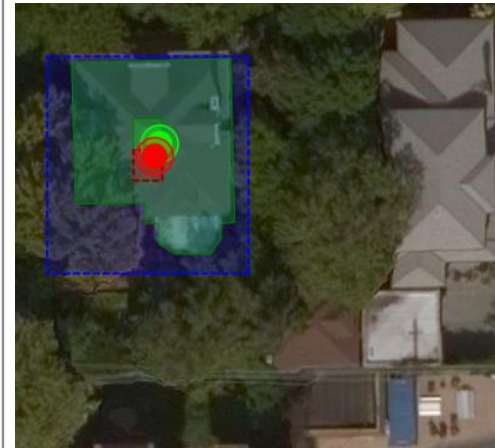
SELECT:
bbox_centroid_distance_m /
bbox_diag_m

AS:
x

WHERE:
(x >= 0.05) and (x < 0.0525)

Source: https://stash.pnnl.gov/users/bork374/repos/ubid/browse/assets/data.cityofchicago.org/syp8-uezg/bbox_centroid_distance_m.diag.0.png

Centroid Offset: Normalized Distance

New York City
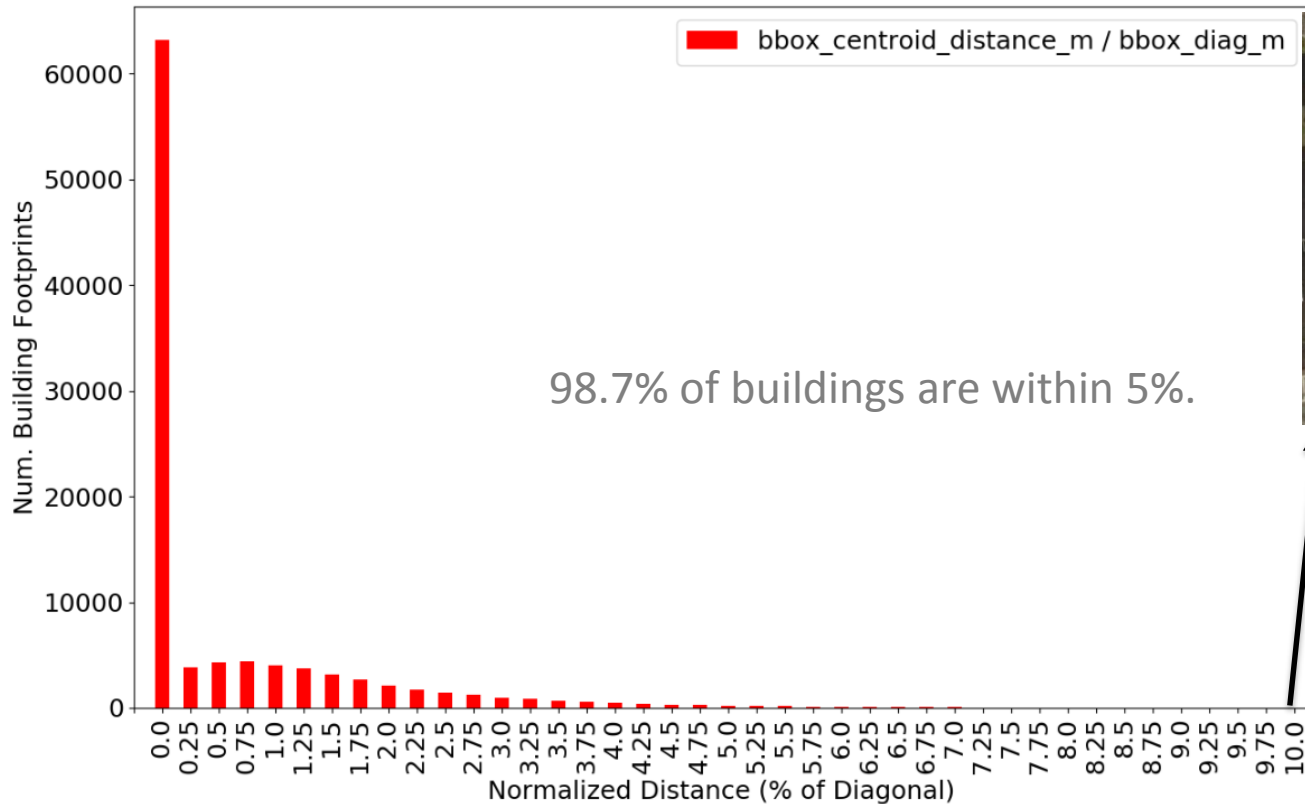
98.7% of buildings are within 5%.

SELECT:
bbox_centroid_distance_m /
bbox_diag_m

AS:
x

WHERE:
(x >= 0.1) and (x < 0.125)

Source: https://stash.pnnl.gov/users/bork374/repos/ubid/browse/assets/data.cityofnewyork.us/dszh-ae6r/bbox_centroid_distance_m.diag.0.png

Bounding Box Area Comparison

San Francisco
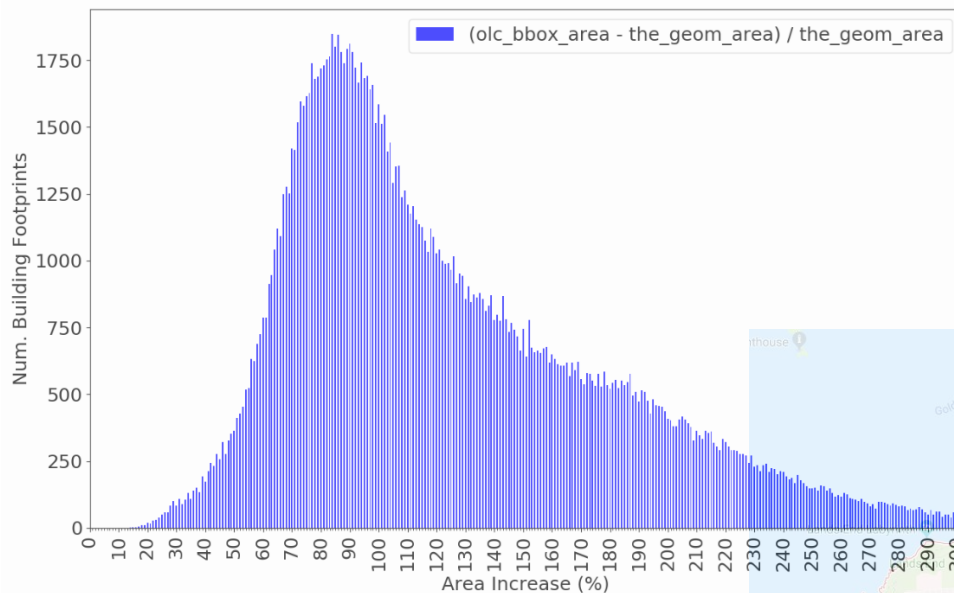
Count: 177,023
Mean: 139%
Std: 122%
25percentile: 84%
Median: 112%
75percentile: 165%
Max: 13,153%

Source: https://stash.pnnl.gov/users/bork374/repos/ubid/browse/assets/data.sfgov.org/2s2t-jwzp/bbox_area_increase.1.png
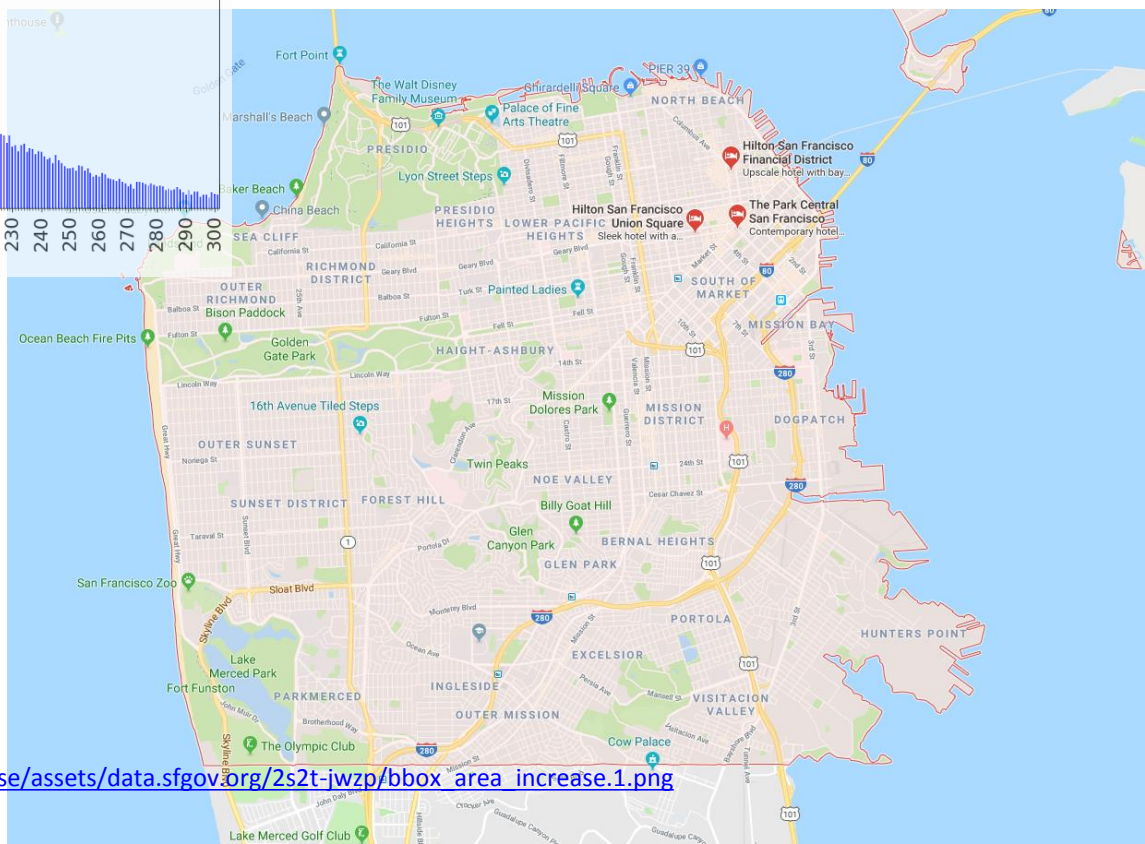
Bounding Box Area Comparison

Chicago

Count: 820,606
Mean: 94%
Std: 61%
25percentile: 61%
Median: 83%
75percentile: 111%
Max: 10,083%
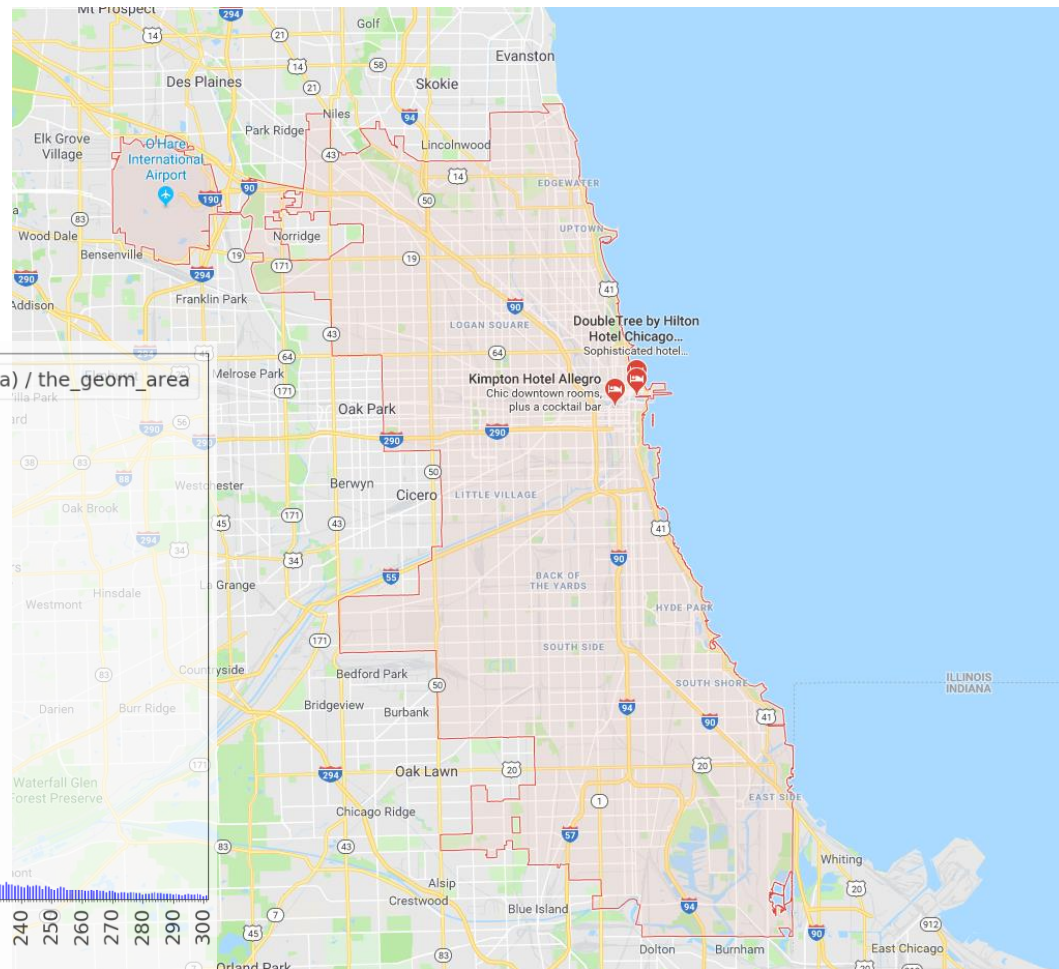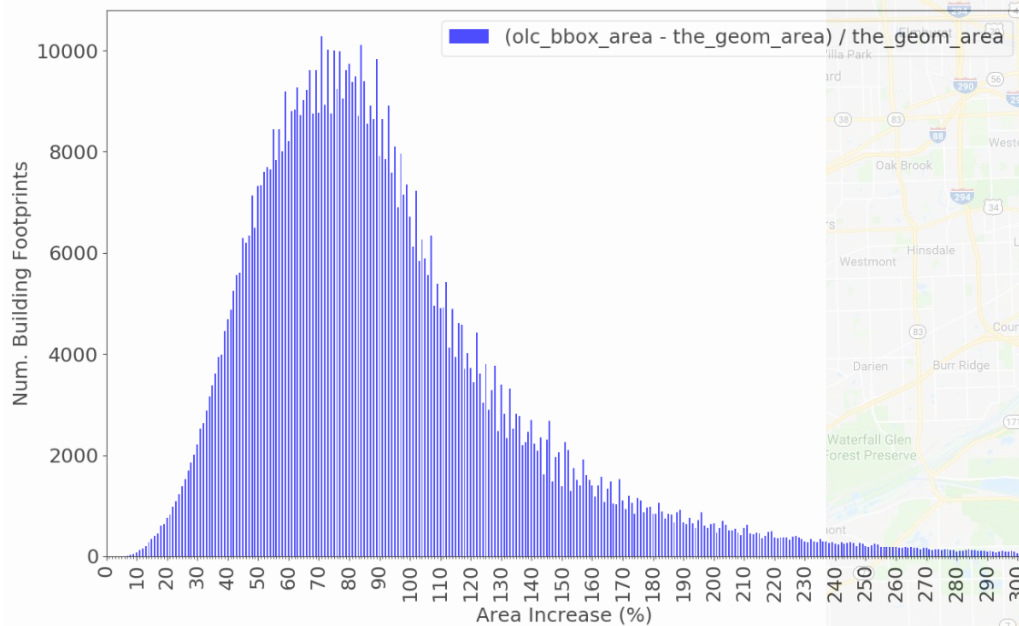
Bounding Box Area Comparison

New York City
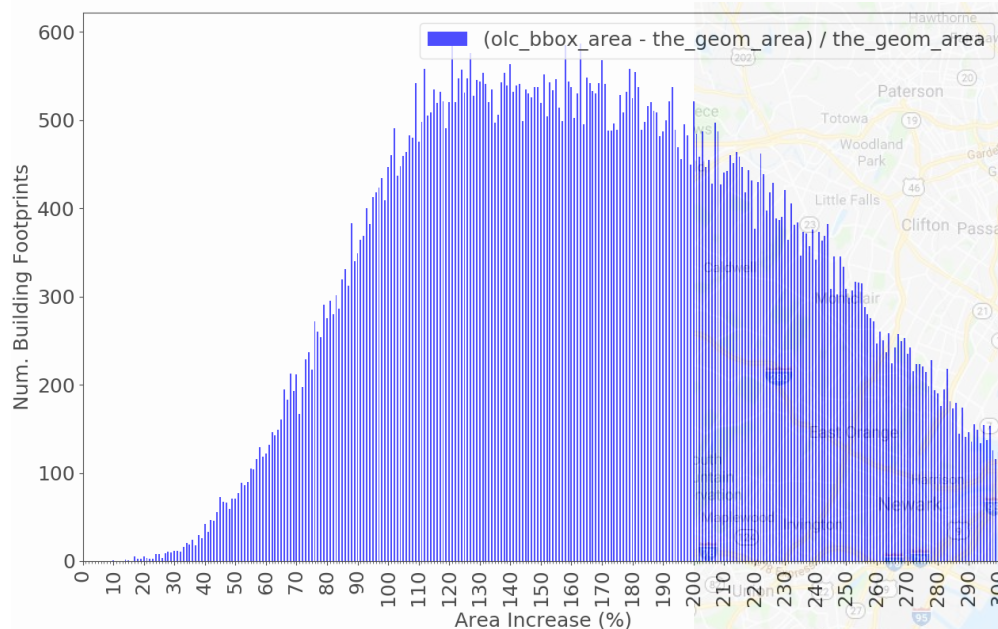
Count 101,831
Mean: 179%
Std: 72%
25percentile: 124%
Median: 172%
75percentile: 225%
Max: 850%

Source: https://stash.pnnl.gov/users/bork374/repos/ubid/browse/assets/data.cityofnewyork.us/dszh-ae6r/bbox_area_increase.1.png
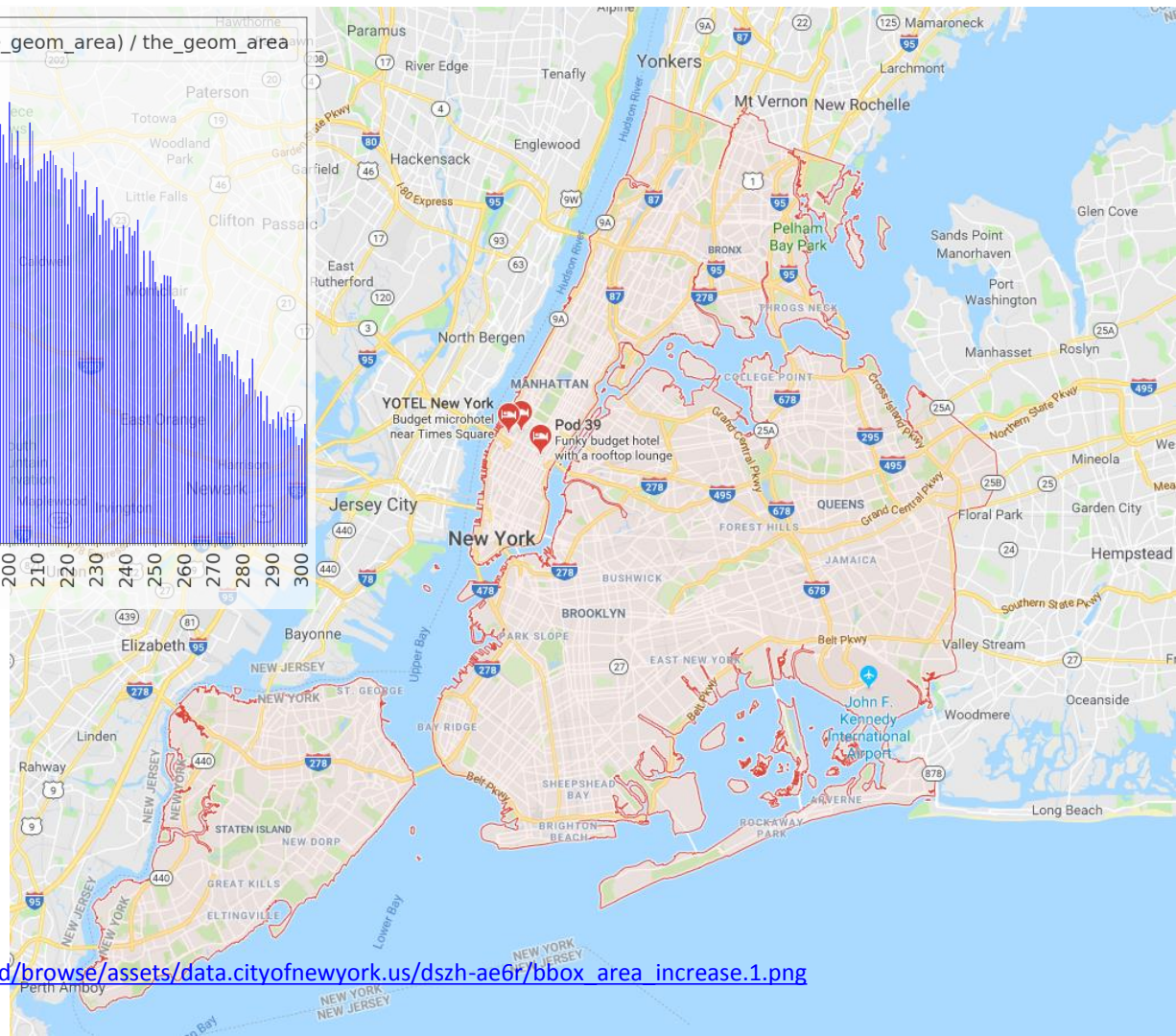
Orientation Analysis

San Francisco

Example of 30 degree rotation

-45

0

Rotation (degrees)

45

Source:
https://stash.pnnl.gov/users/bork374/repos/ubid/browse/assets/data.sfgov.org/2s2t-jwzp/rotation_degrees.0.png
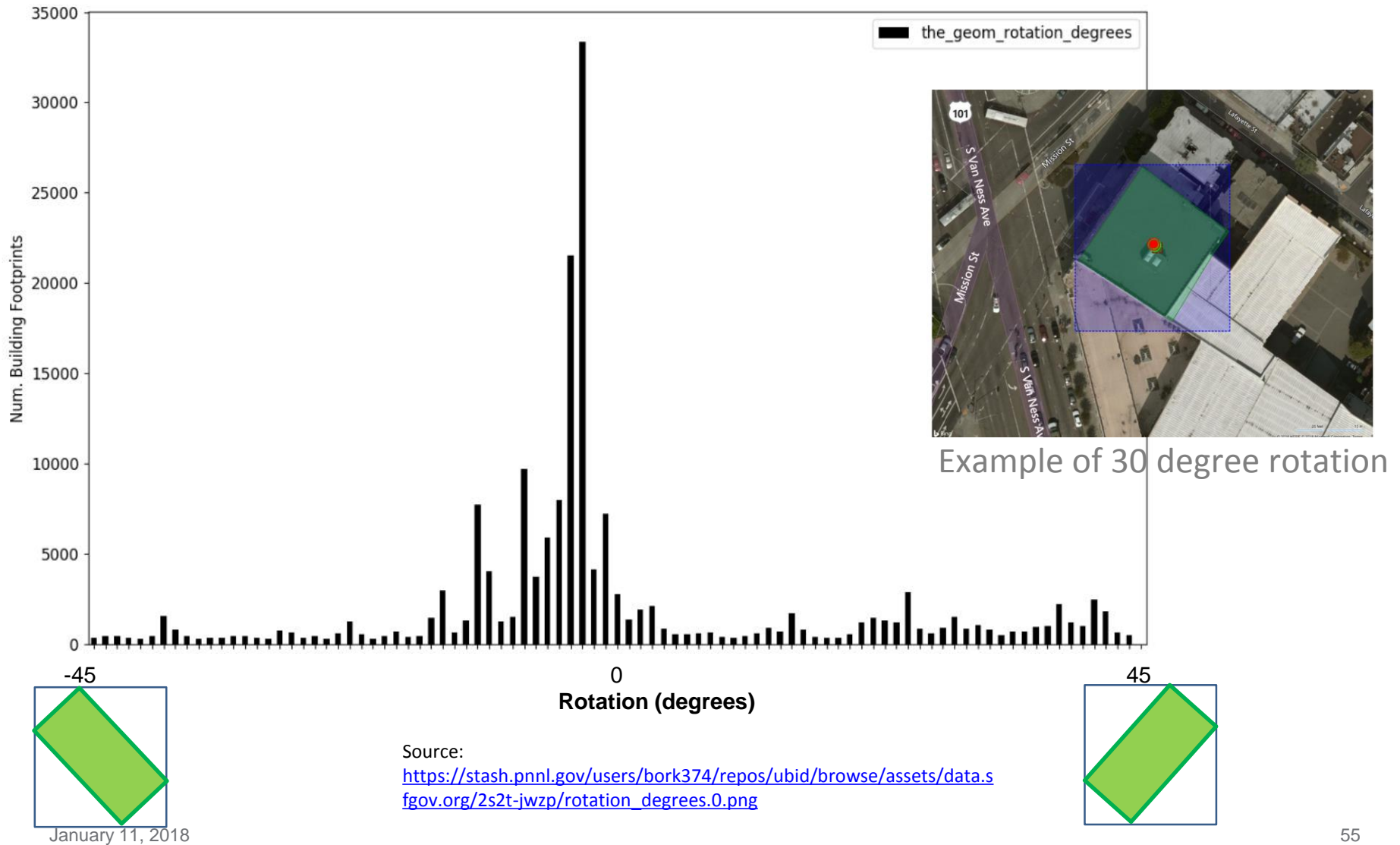
Orientation Analysis

Chicago
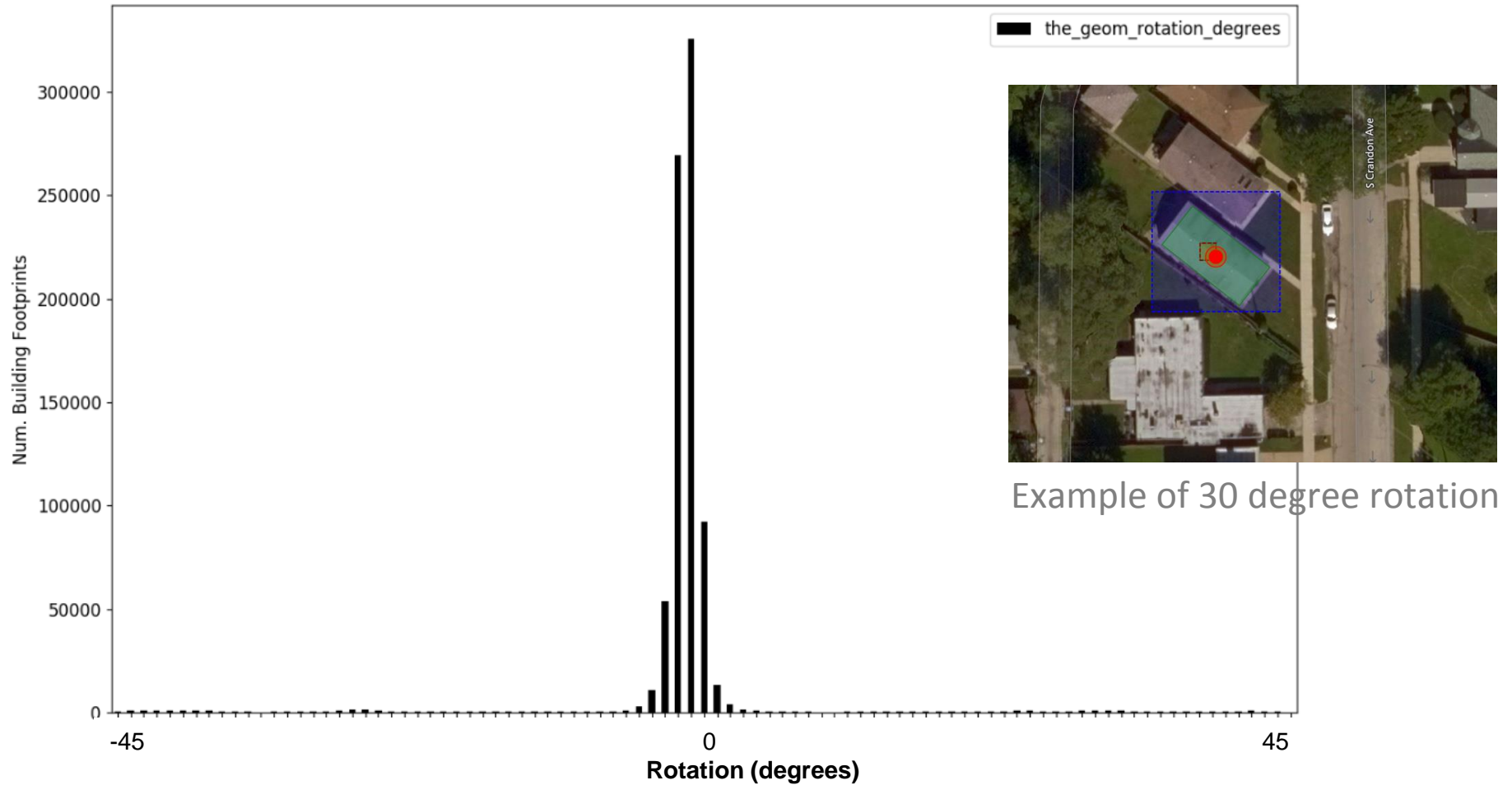
Example of 30 degree rotation

Source: https://stash.pnnl.gov/users/bork374/repos/ubid/browse/assets/data.cityofchicago.org/syp8-uezg/rotation_degrees.0.png

Orientation Analysis

New York City
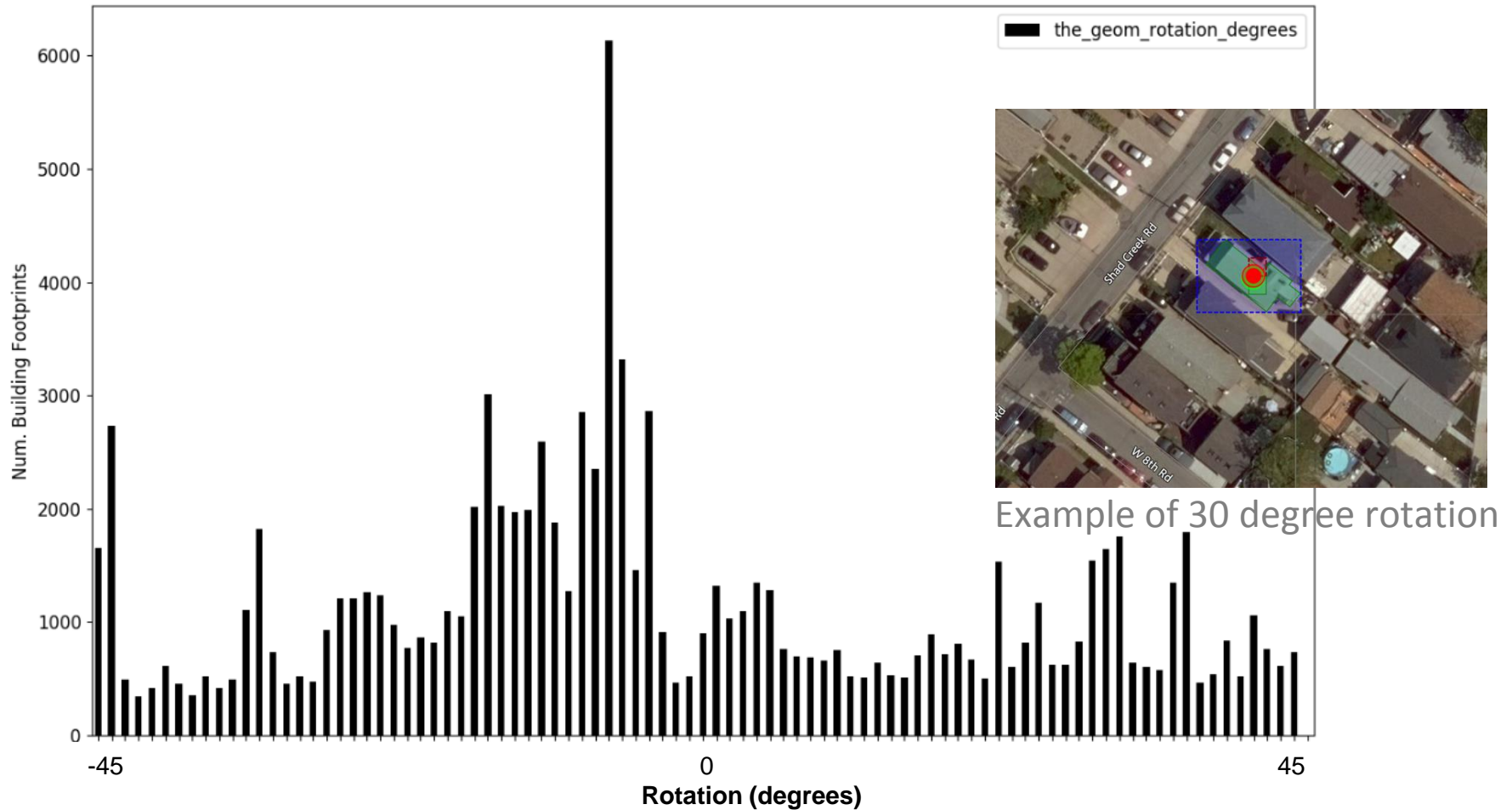
Example of 30 degree rotation

Source: https://stash.pnnl.gov/users/bork374/repos/ubid/browse/assets/data.cityofnewyork.us/dszh-ae6r/rotation_degrees.0.png

Example thresholds to remove duplicates and detect bad data

Data error 1: small objects are mistaken as buildings.
Threshold: Minimum h and w of the bounding box > 3

Data error 3: streets are mistaken as buildings.

Threshold:
Max area increase: 500%
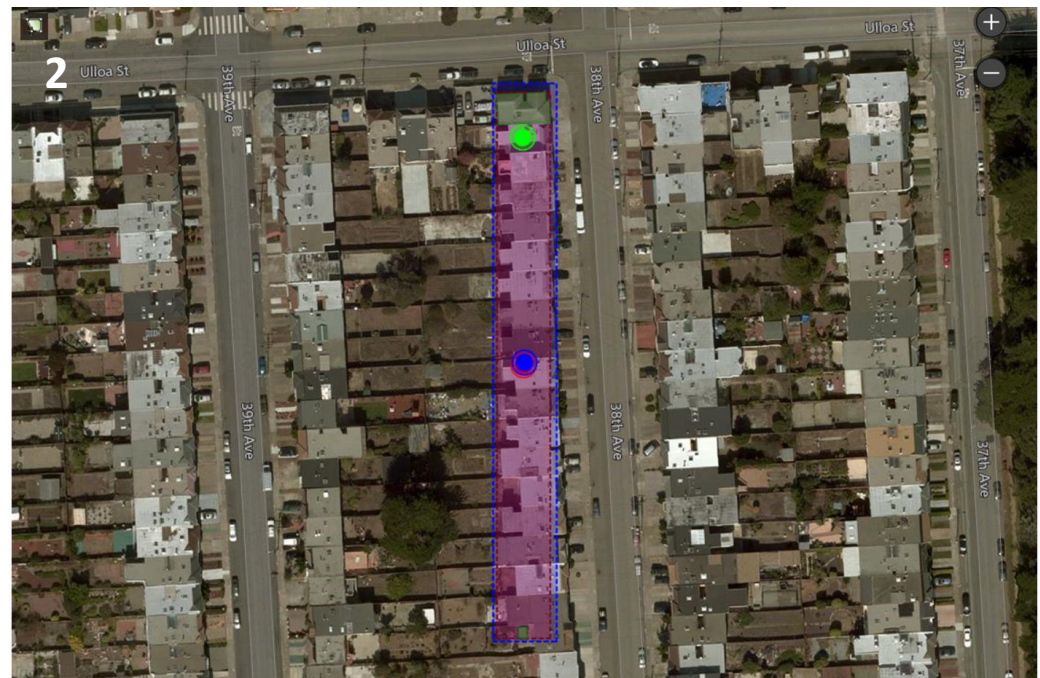
Data error 2: two buildings far away are mistaken as one building.
Threshold: Max centroid offset < 50%

Rules to match UBID (Future Work)

Develop a validation tool to check the likelihood of multiple UBIDs refer to the same building. This is an important aspect of the block chain application.

- **Same:** what does it mean to be the 'same' building?
    - Query: Here is my building, show me all of the UBID's that are the 'same' based on some error metric(s).
    - Can a (large) building own 9 or more UBIDs (3x3 OLC grid)?
- **Within:** Is my UBID within another UBID, if UBID is used to identify land, a building complex, etc.,
    - Query: Here is a UBID for a lot, complex or borough, what building (UBIDs) does it contain?
- **Contained by:** inverse of 'within'
    - Query: Here is the UBID for my building. What lot, complex, or borough UBIDs contain my building?

UBID Implementation

Leveraging Blockchain to solve for persistence

UBID Blockchain Principles:

Implementation Goals

- Persistence
  - Transactions cannot be erased or altered
- Transparency
  - All transactions & associated data are globally visible
- Security & Privacy
  - Non-critical data will not be exposed
- Availability
  - Must reduce cost to users for building identification in data sources
- Connectivity
  - Relevant mappings & keys will be stored on the chain

Public Transactions

- UBID Creation
  - Construction of a new building
  - Initial association of existing building with UBID
- UBID Retirement
  - Demolition of building previously assigned a UBID
- UBID Reassociation
  - Building alteration requires an adjustment to existing UBID
  - New construction on site of retired UBID requires new assignment separate from past data
- Other Possible Transactions
  - Inspections, certifications, contracts, payments
  - Devices, infrastructure changes

Blockchain + API Service

User Ecosystem & Public Accessibility

Blockchain Node Hosts

General Public

Public Agencies

NGOs

Contractors & Service Companies

API Services & Sidechains

Web-based GIS

Data Providers & Aggregators

MLS

Technology Companies

Blockchain Application & API

UBID Blockchain

Benefits of UBID Blockchain

- Documents the creation of UBIDs and enables other transactions to key on them
  - Builds in conflict resolution mechanism
- Creates a unique blockchain "address" for each building
  - Unique private/public key and certificate
- Opens numerous possibilities for other custom services
  - Mapping & GIS
  - Land management & city planning
  - Building certification & maintenance programs

What's next for UBIDs?

- How effectively can UBID match buildings from different datasets (vs. addresses, etc.)?

- How do we ensure persistence of the UBID?

- How much ambiguity should UBID tolerate?

Data Matching Pilots

Premise: UBIDs are better than addresses for matching records across datasets.

Pilots:

- City of San Francisco & BRICR
- City of Chicago
- SEED Platform Partners
- GSA GRESB Portfolio Analysis with USGBC

Full Implementation Pilots

Premise: UBIDs, stored in the blockchain, can serve as a record of note that persists to support data-driven systems and programs.

We are actively seeking pilot partners!

- Data matching

- UBID retirement & reassignment opportunities

- Stand up a blockchain node

Thank You

To learn more about UBIDs visit us at:
https://buildingid.pnnl.gov

Or contact us directly:
Harry.Bergmann@ee.doe.gov
Nora.Wang@PNNL.gov