

# Cuckoo!

Automated Outlier Analysis for Revit

# Our Team!



**Alex  
Ogata**



**Sarvesh  
Sp**



**Jim  
Peraino**



**Zhenxiang  
Huang**



**Matthew  
Breau**

# Problem definition

Uncaught errors in models cost real \$\$

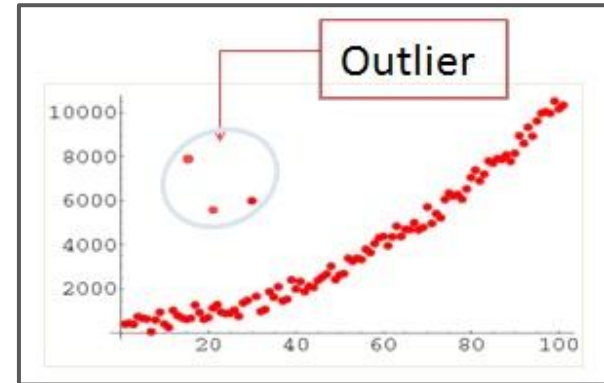
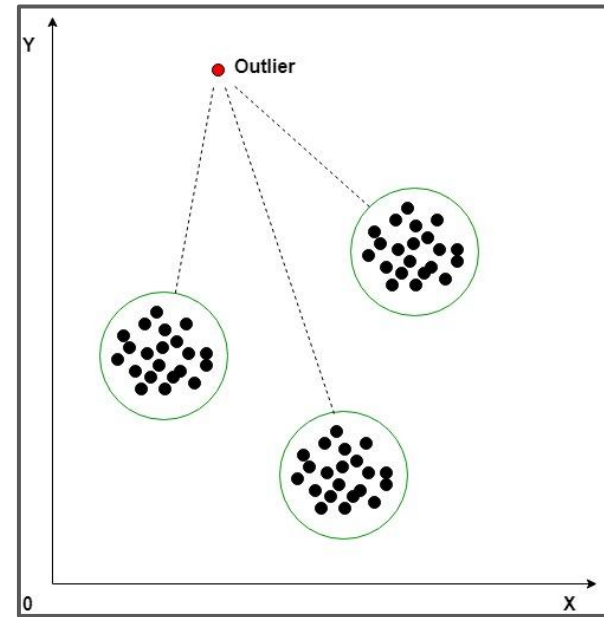
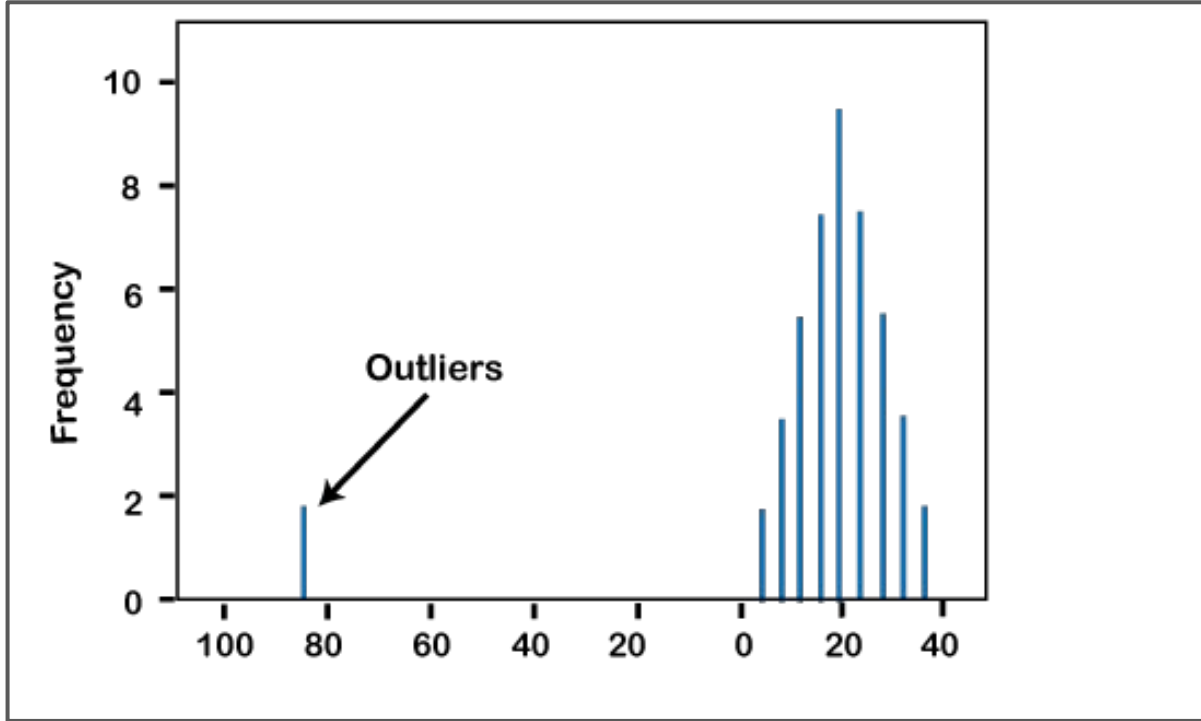
Q:

How do you identify when some feature is likely to be an error?  
(without hard-coding a huge set of rules)

# What is an outlier?

In **statistics**, an **outlier** is a **data point** that differs significantly from other observations.<sup>[1][2]</sup> An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the **data set**.<sup>[3]</sup> An outlier can cause serious problems in statistical analyses.

# What is an outlier?



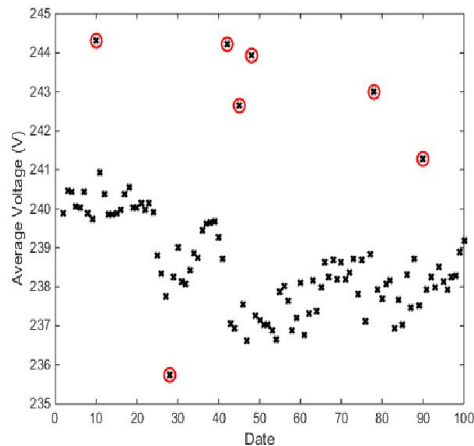


Outliers can be hard to spot

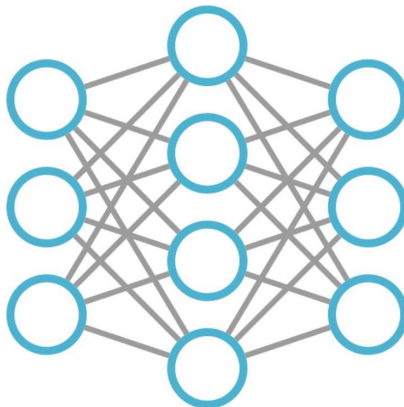




# Outliers are bad for Machine Learning



+

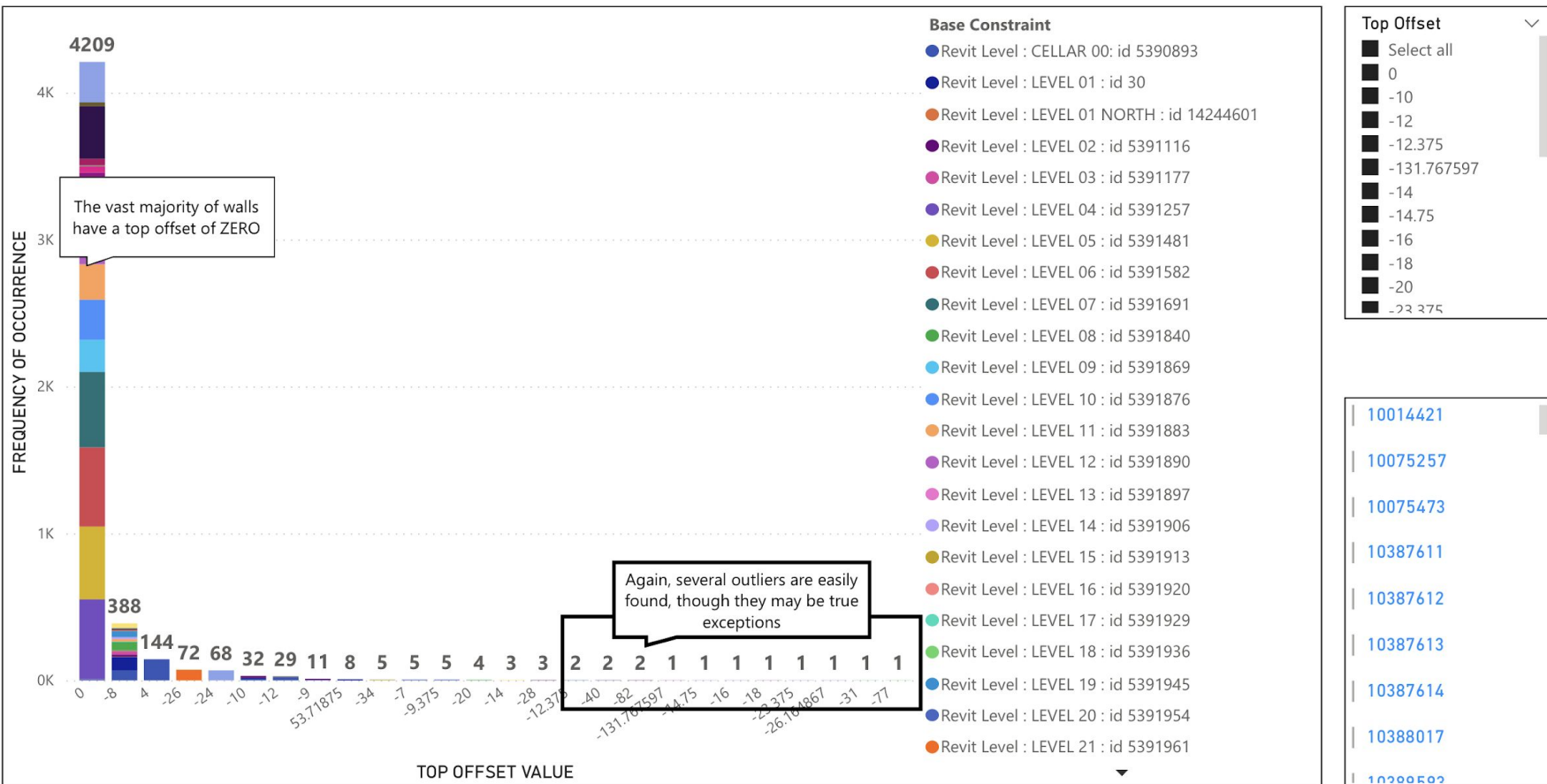


=

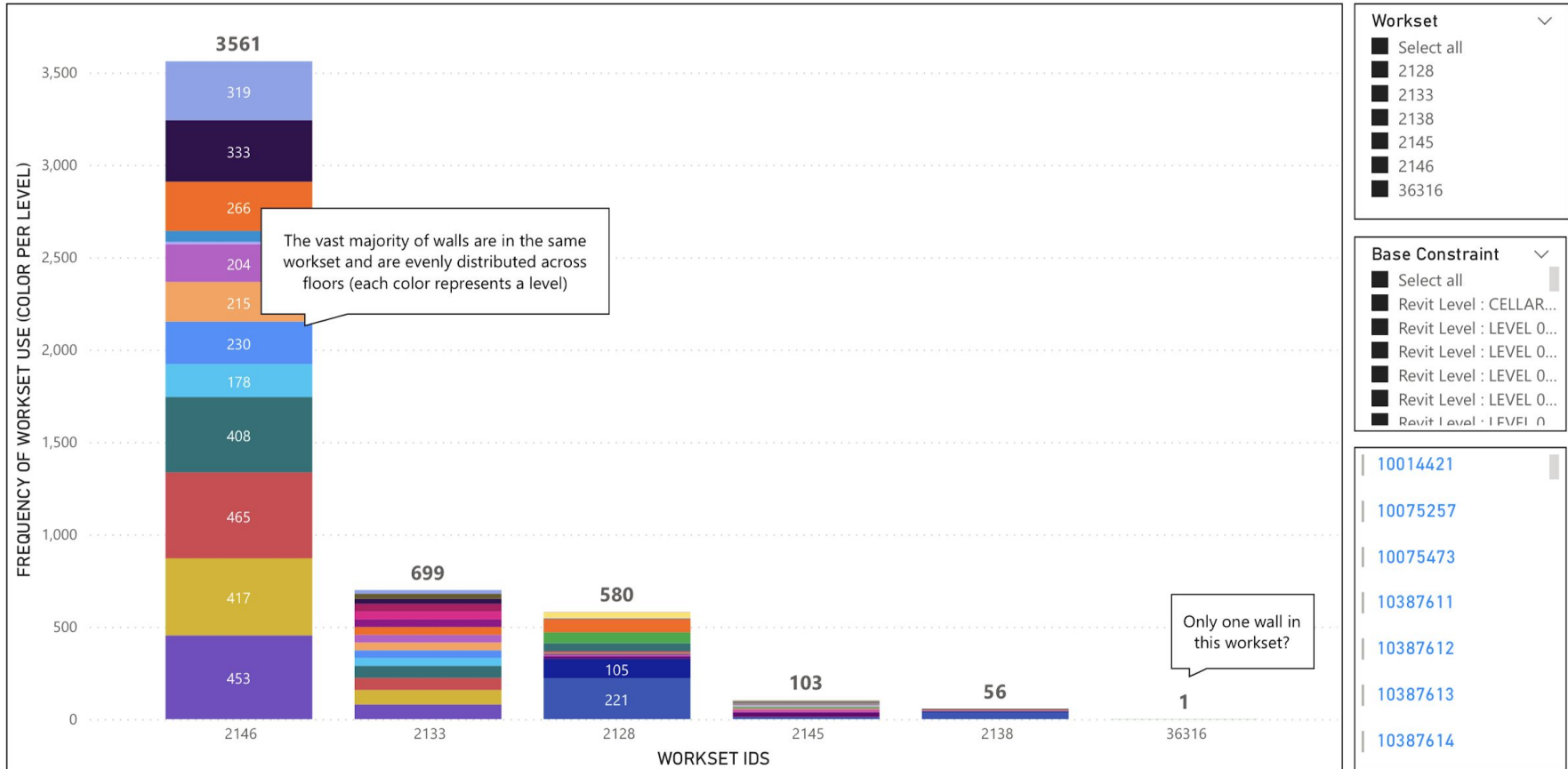




Outliers: TOP OFFSET (each column represents a top offset value, the height of the column shows the qty of walls with that value)

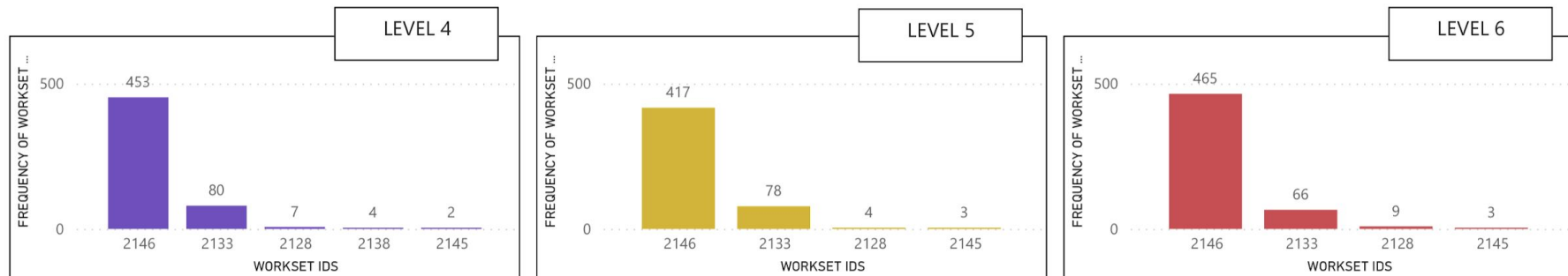


Outliers: WORKSET USAGE (each column represents a workset, height shows number of walls in that workset)



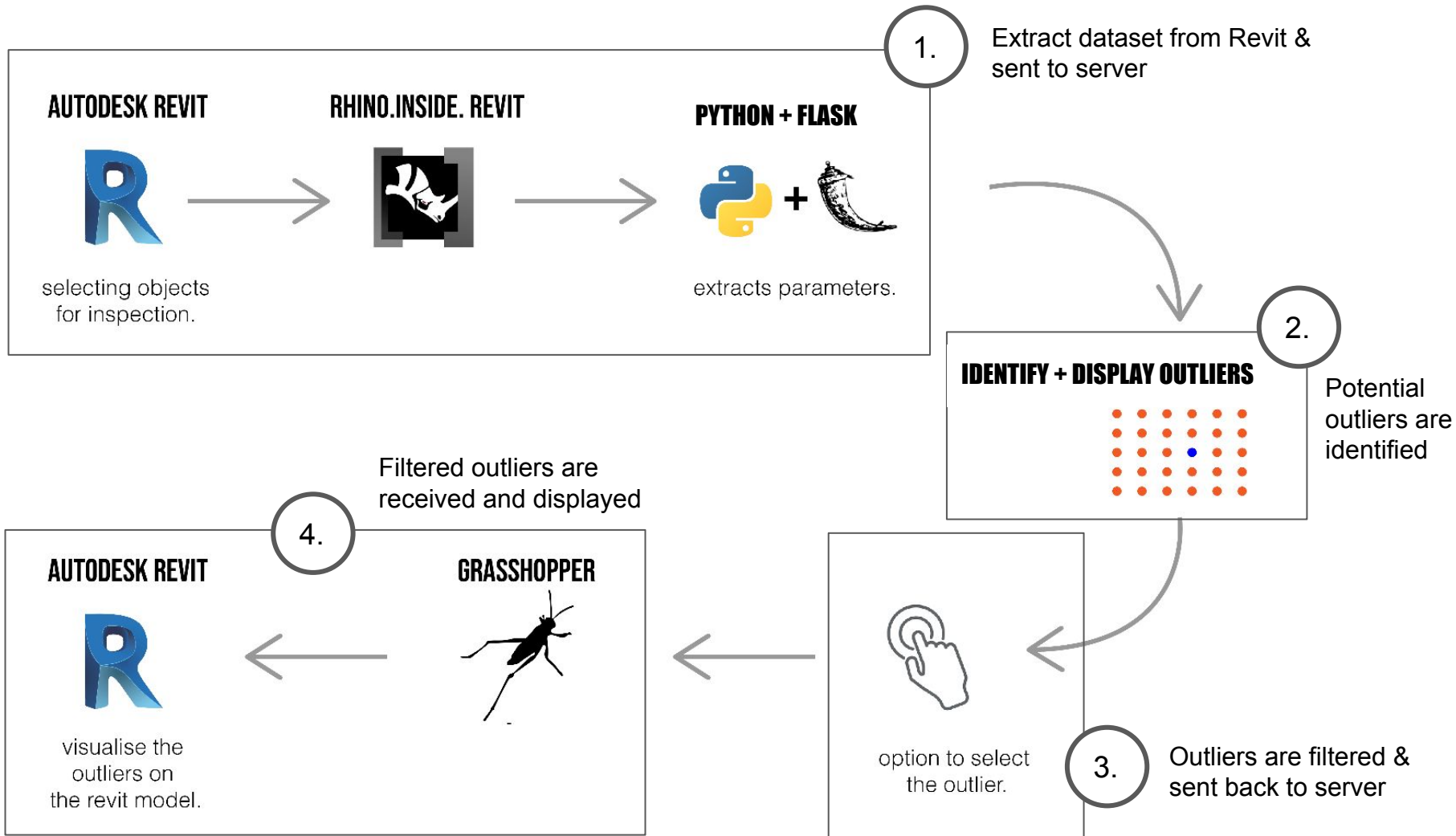
## Outliers: NUMBER OF WORKSETS IN USE BY LEVEL (each column represents a workset)

the "four column type"



the "three column type"







# 1. Extract data from Revit

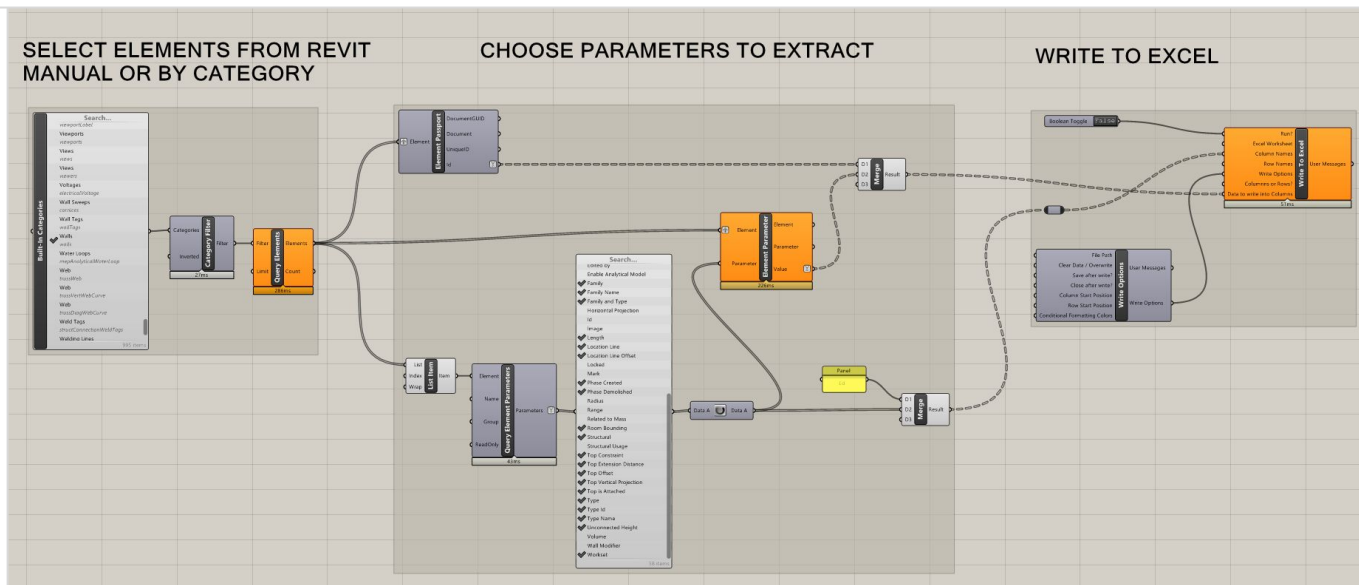
Revit Model



Rhino.Inside  
(Use grasshopper with Revit)



Hops to Flask  
(Send a stream of data to a flask server)



## 2. Identify possible outliers

Flask Server  
(Call server to access JSON)



SK Learn  
(Univariate outliers,  
Mahalanobis, Box Plot,  
Scatter Plot)



Flask Server  
(Send list of ids to server)

```
#Detecting Multivariate Outliers
#each parameter at the input should be numerical
def robust_mahalanobis_method(df):
    #Minimum covariance determinant
    rng = np.random.RandomState(0)
    real_cov = np.cov(df.values.T)
    X = rng.multivariate_normal(mean=np.mean(df, axis=0), cov=real_cov, size=506)
    cov = MinCovDet(random_state=0).fit(X)
    mcd = cov.covariance_ #robust covariance metric
    robust_mean = cov.location_ #robust mean
    inv_covmat = sp.linalg.inv(mcd) #inverse covariance metric

    #Robust M-Distance
    x_minus_mu = df - robust_mean
    left_term = np.dot(x_minus_mu, inv_covmat)
    mahal = np.dot(left_term, x_minus_mu.T)
    md = np.sqrt(mahal.diagonal())

    #Flag as outlier
    outlier = []
    C = np.sqrt(chi2.ppf((1-0.00001), df=df.shape[1])) #degrees of freedom = number of variables
    for index, value in enumerate(md):
        if value > C:
            outlier.append(index)
        else:
            continue
    outlier_id = df.index[outlier]
    return outlier_id
```

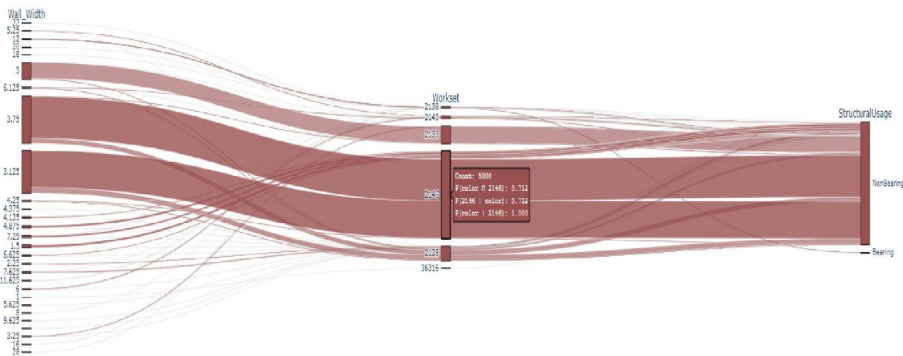
```
#Outlier Functions...
#Detecting Univariate Outliers
def univariate_outlier(df, variable, q1_percent, q3_percent):
    q1 = df[variable].quantile(q1_percent)
    q3 = df[variable].quantile(q3_percent)
    iqr = q3-q1
    inner_fence = 1.5*iqr
    outer_fence = 3*iqr

    #inner fence lower and upper end
    inner_fence_le = q1-inner_fence
    inner_fence_ue = q3+inner_fence

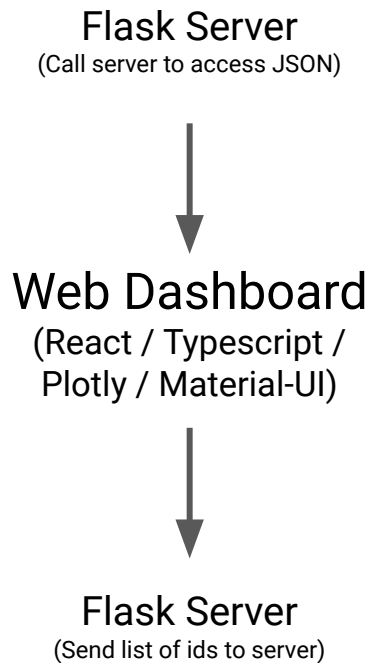
    #outer fence lower and upper end
    outer_fence_le = q1-outer_fence
    outer_fence_ue = q3+outer_fence

    outliers_prob = []
    outliers_poss = []
    for i in range(len(df[variable])):
        x = df[variable].values[i]
        index = df[variable].take([i]).index[0]
        if x < outer_fence_le or x > outer_fence_ue:
            outliers_prob.append(index)

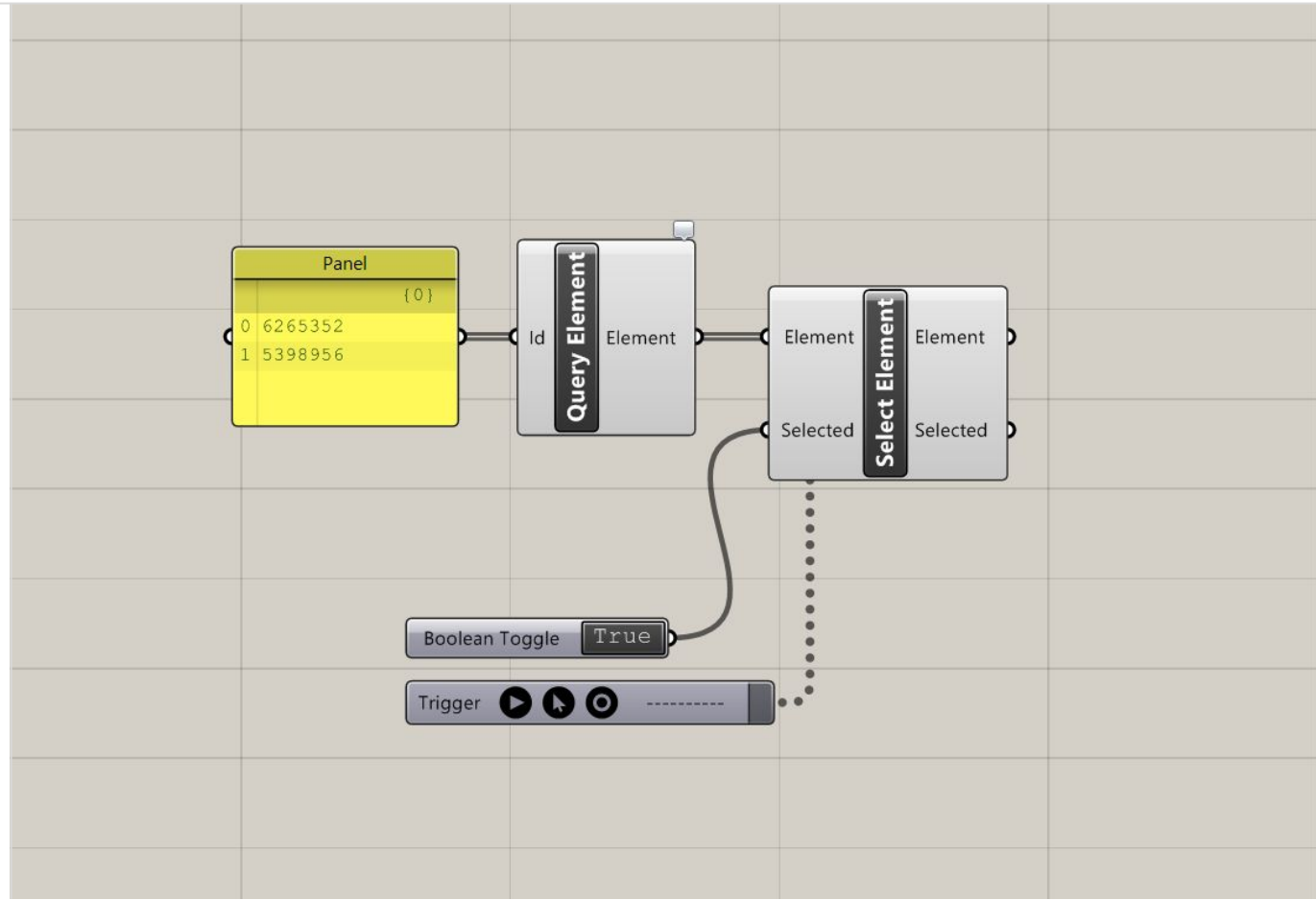
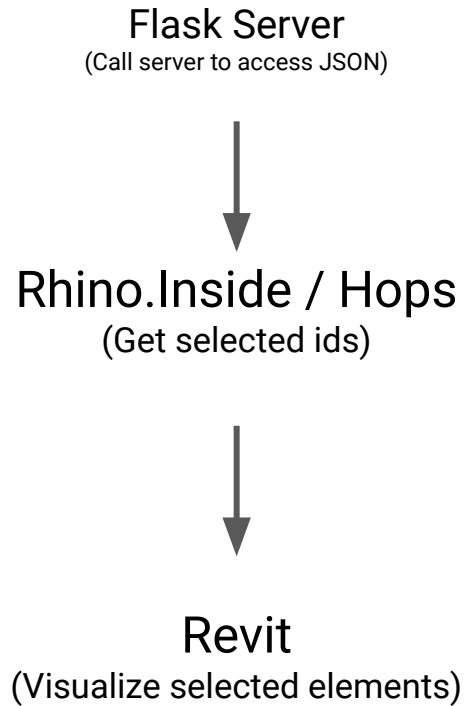
    for i in range(len(df[variable])):
        x = df[variable].values[i]
        index = df[variable].take([i]).index[0]
        if x <= inner_fence_le or x >= inner_fence_ue:
            outliers_poss.append(index)
    return outliers_prob, outliers_poss
```



### 3. Visualize, Explore, and Select Outliers



## 4. View outliers in Revit





# 4. View outliers in Revit

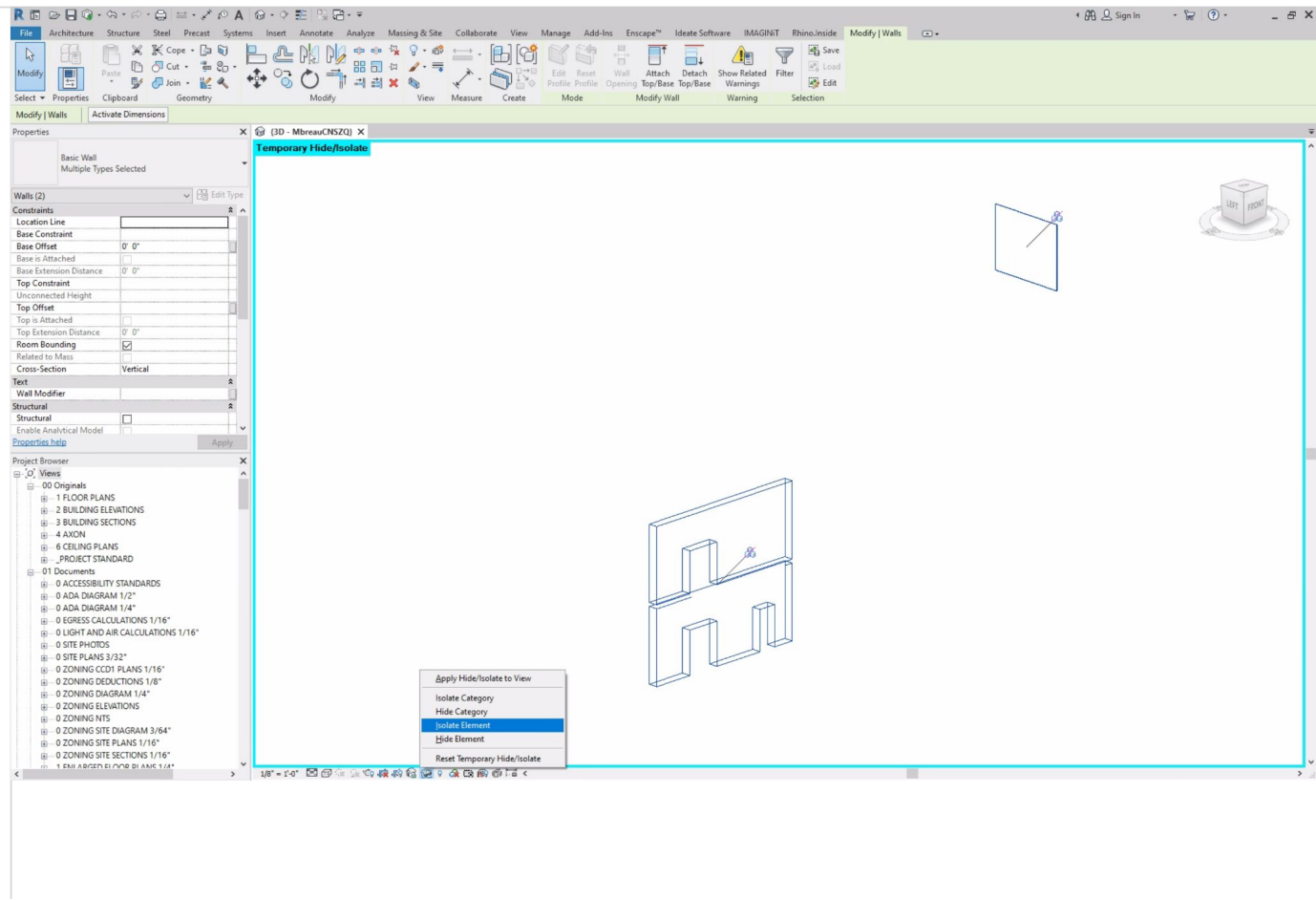
Flask Server  
(Call server to access JSON)



Rhino.Inside / Hops  
(Get selected ids)



Revit  
(Visualize selected elements)



# Future Development

- **Use Hops instead of Excel for data export**

Automate export/ analysis procedure

- **Automatic highlighting in Revit**

Hops should return IDs into Grasshopper so the user does not need to copy & paste

- **“Remember” User’s Input**

Once the user has confirmed that a particular parameter on a particular element its not an outlier, that element/ parameter would not be flagged as an outlier again

- **Recipes**

Learn from experience which parameters are likely to be worthwhile to cross-reference, and prioritize these for new projects.

- **Automatic recognition of meta-patterns**

Implement multi-variable regressions



# Thanks!

<https://github.com/jperaino/cuckoo>

# PRESENTATION PITCH CONCEPT OUTLINE (for discussion)

Clearly Identify:

Key Idea & value proposition

Real problem. Own the “non-flashy”

NOTHING existed before saturday

Errors & Omissions are a real problem. (RFIs, litigation)

Specific uses: data consistency in naming, wrong family use, i.

Incorrect size

We understand not all “outliers” are mistakes. Filtering is key (enter ML)

Key tech innovations used (NEED TO HIT THIS... what are they????)

Identify open source libraries used

Any ML usage? (PYOD)

Key future potential

Model training from multiple models

Key Idea & value proposition  
Key future potential

MB

1. MAIN VISION  
MAIN INTRO  
Introduce team,  
main idea, vision,  
base concepts, etc

2. VALUE POTENTIAL  
Base proof of concept:  
pbi diagrams

AO

3. HOW IT IS DONE  
Data flow diagram  
-this is key. Show whole  
picture, understanding

SSP

Key Idea & value proposition  
Key tech innovations used  
Key future potential

6. FINAL SELL PITCH  
Close  
RECAP, VISION,  
FUTURE POTENTIAL

MB

Key Idea & value proposition  
Key tech innovations used  
Key future potential

5. TECHNOLOGICAL  
INNOVATION Our Hack...  
processes & results

ZH

Key tech innovations used

4. TECHNOLOGICAL  
INNOVATION  
Our Hack...  
processes & results

JP