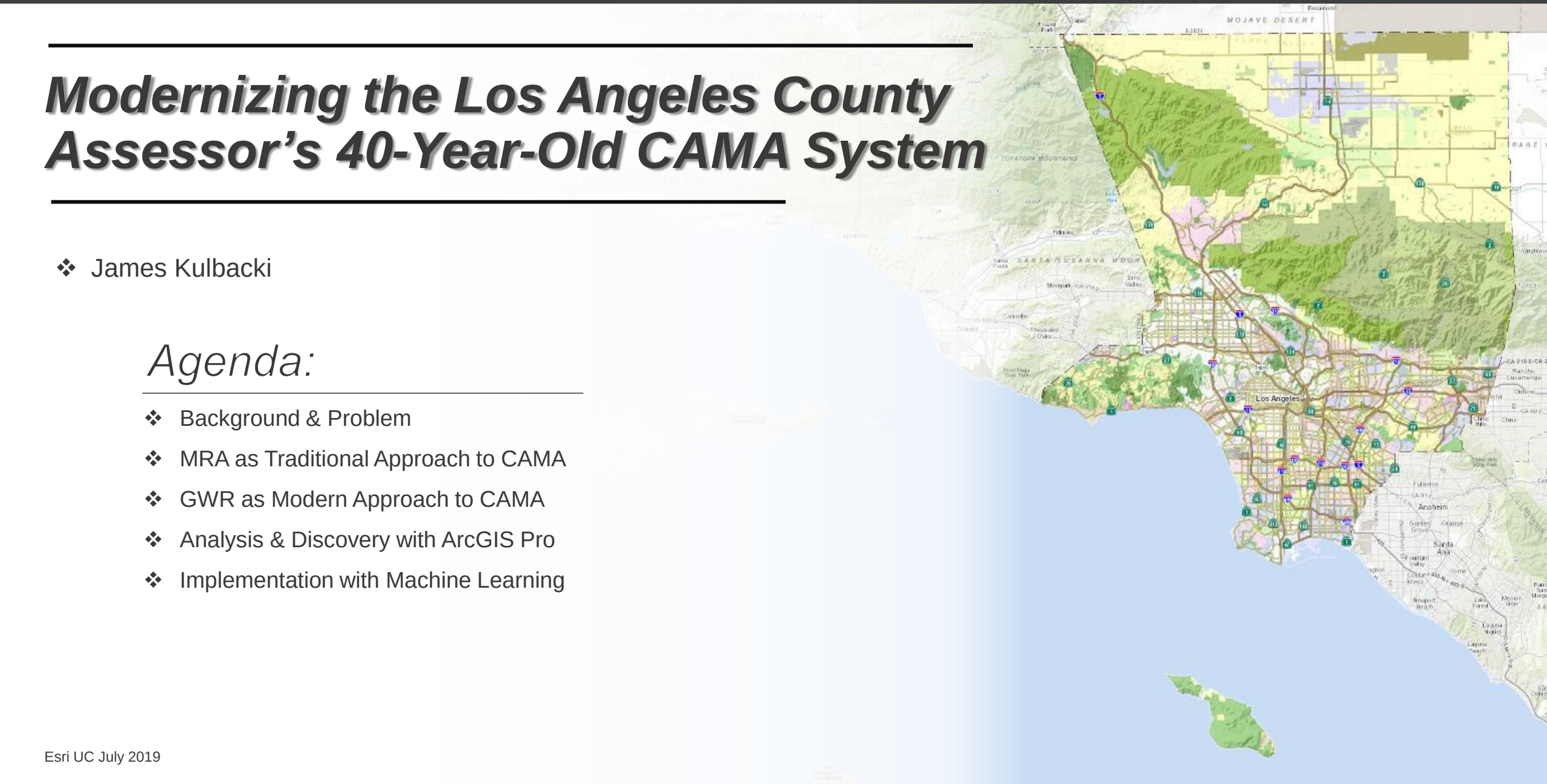
Modernizing the Los Angeles County Assessor's 40-Year-Old CAMA System

- James Kulbacki

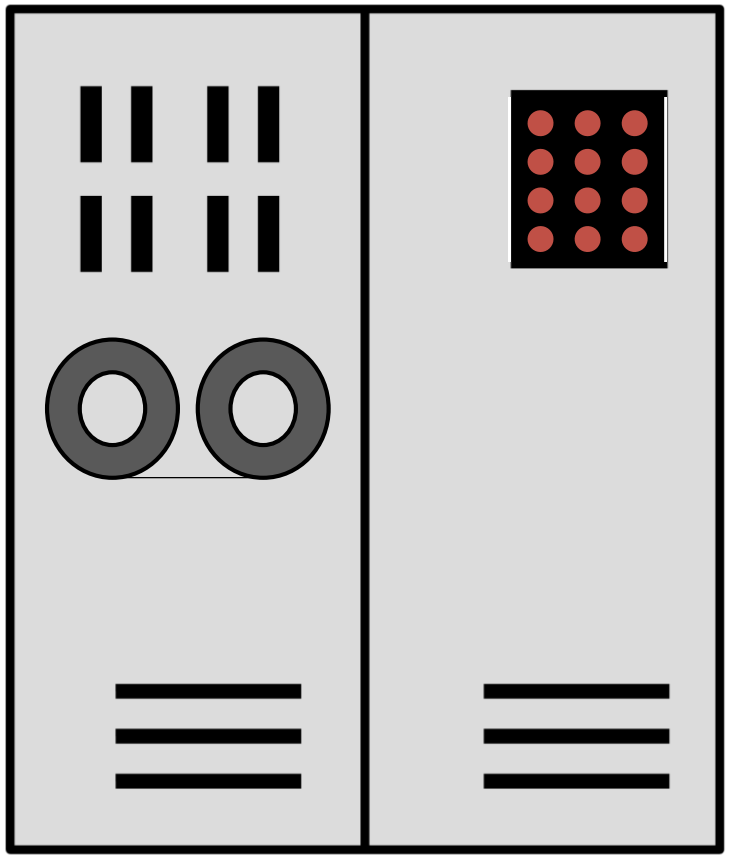
Agenda:

- Background & Problem
- MRA as Traditional Approach to CAMA
- GWR as Modern Approach to CAMA
- Analysis & Discovery with ArcGIS Pro
- Implementation with Machine Learning



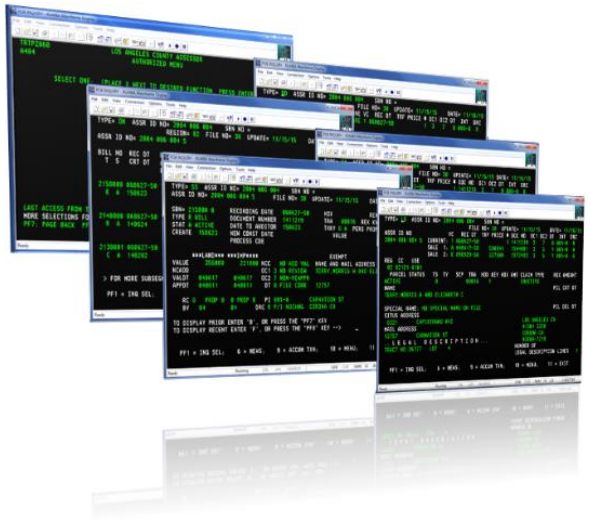
Background & Problem

IBM Mainframe & IMS System



AS400

"Green" Screens



Legacy Systems

- **Technology**: Built in 1980's using 1970's tech.
- **Location**: No spatial intelligence.
- **Methods**: Evolving AVM models unsupported.
- **Data**: Can't adapt to availability of new data.
- **Code**: COBOL, JCL, others…….Fear of change and breaking something.
- **Platform**: Can't scale for economic cycles or alternate uses.

Background & Problem

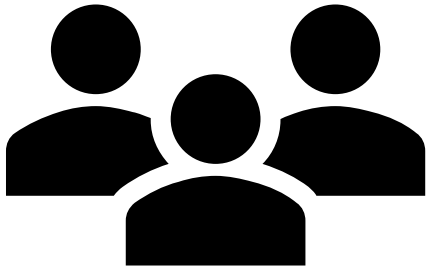
The Mission! Replace the mainframe

Assessor Modernization Project (AMP)

- Multi-Year Project
- $24 Million
- Replace Legacy Systems

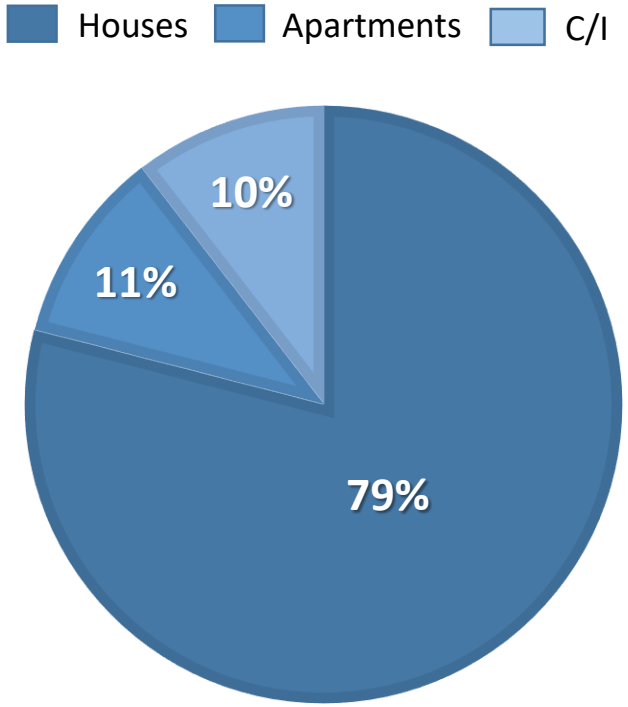
Background & Problem

The County has grown and things have changed over 40 years.

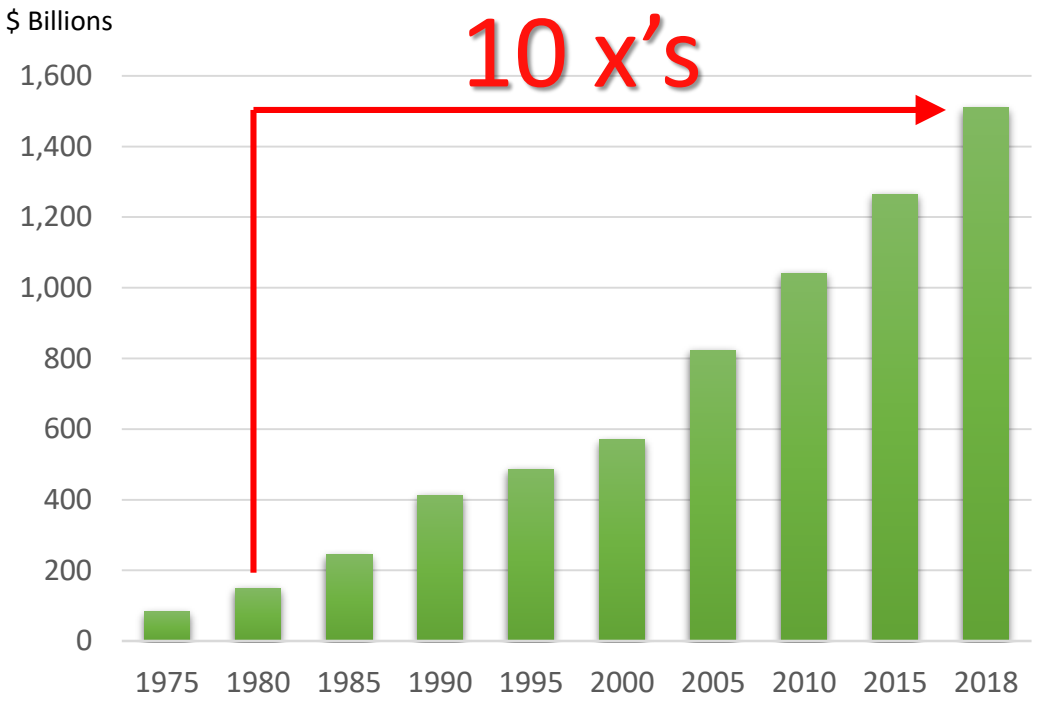


10 Million People

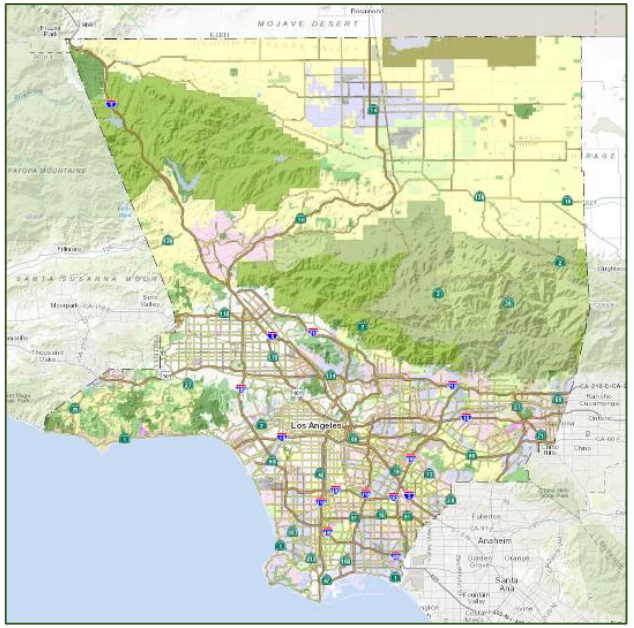
88 Cities



2.4 Million Parcels



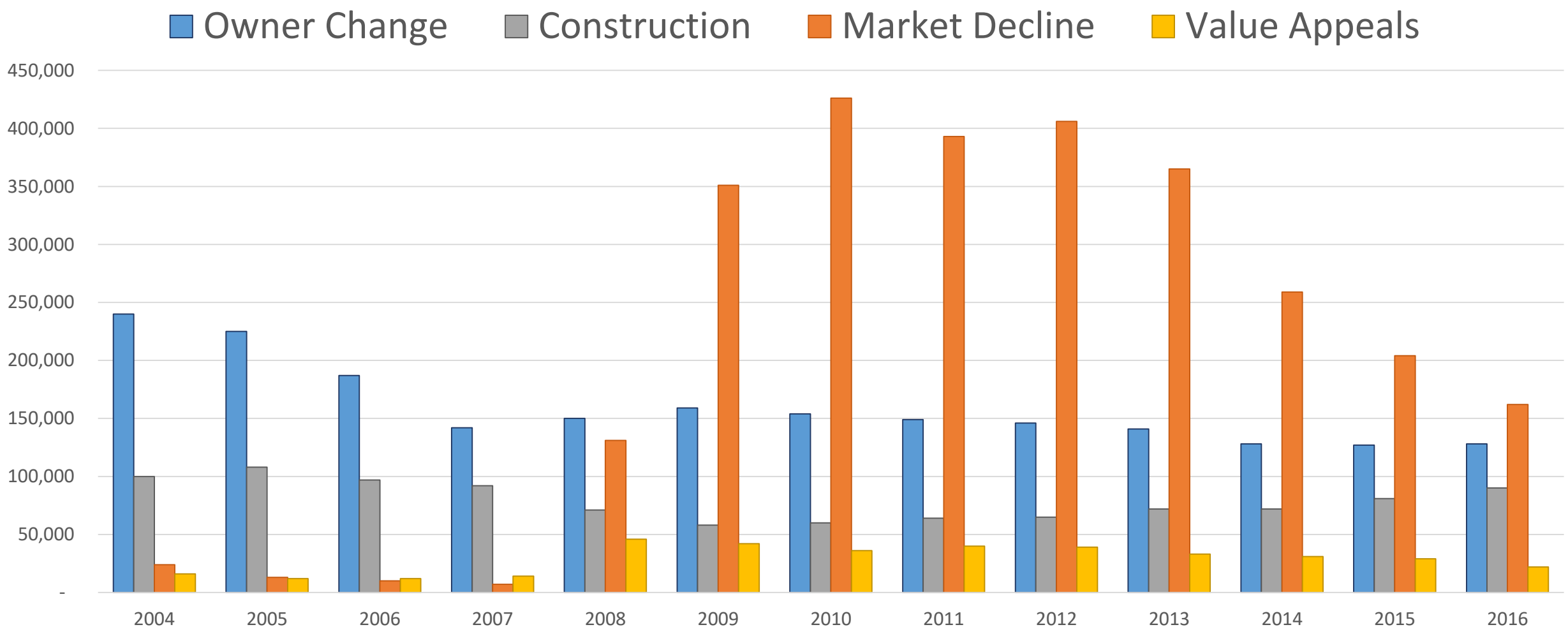
Assessed Value



4,751 Sq. Miles

Background & Problem

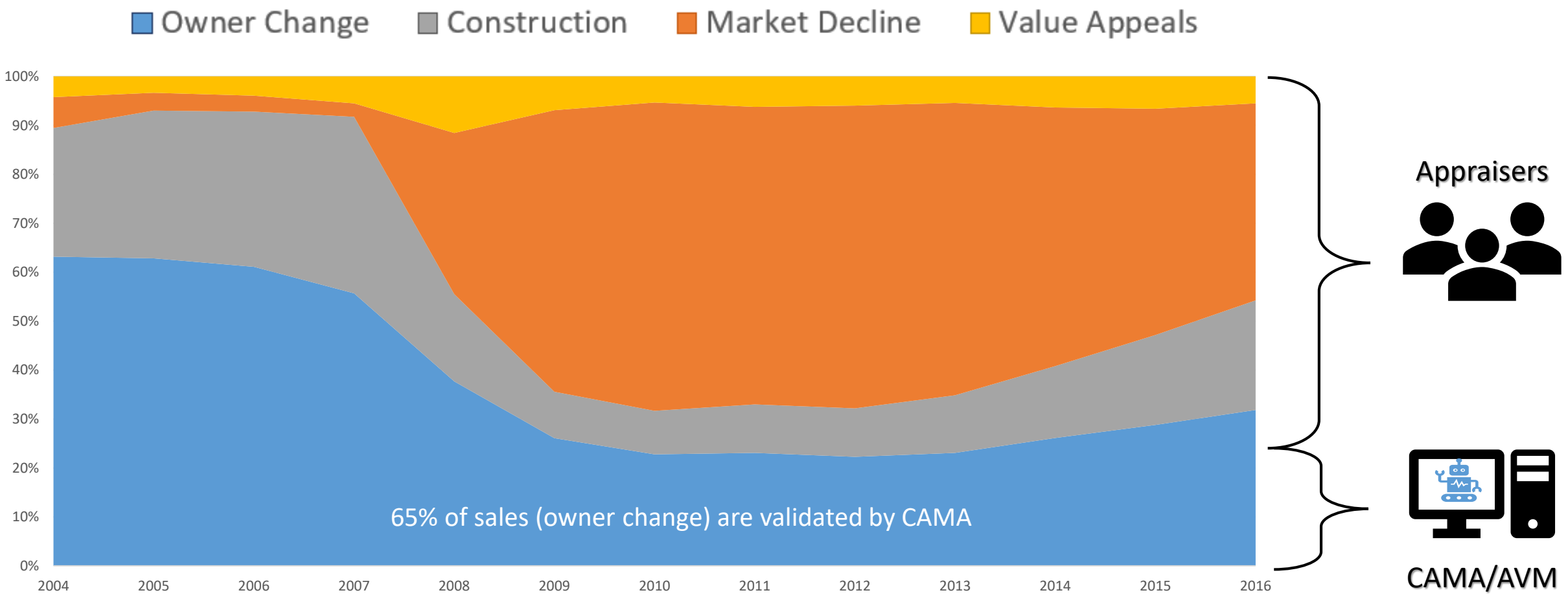
Primary Duty: Determine Assessed Value



Statistics and figures from the 2016, 17, 18 Annual Report

Background & Problem

Annual Appraisal Workload: ≈ 700,000 Appraisals in 2010



Statistics and figures from the 2016, 17, 18 Annual Report

Traditional AVM Approaches

Hedonic Pricing Model

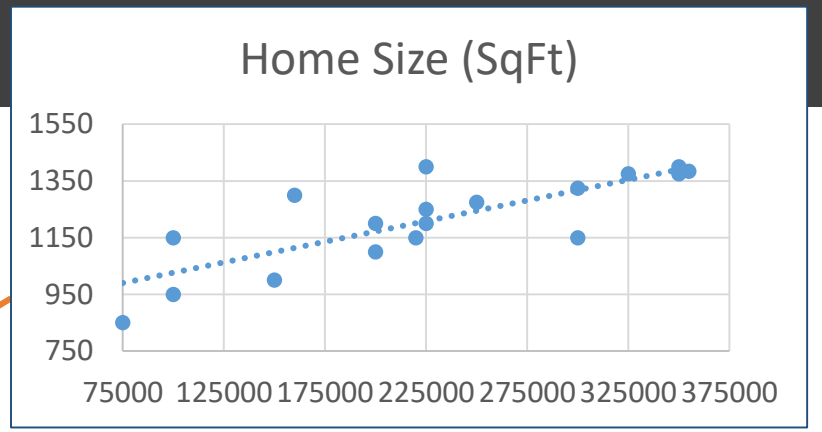
Characteristic	Unit		$ /Unit		Total
Home Size (SqFt)	1,500	x	$75	=	$112,500
# Bedrooms	3	x	$5,000	=	$15,000
# Bathrooms	2	x	$10,000	=	$20,000
Home Age (Years)	50	x	($500)	=	($25,000)
Swimming Pool	Yes	x	$20,000	=	$20,000
Lot Size (SqFt)	7,500	x	$15	=	$112,500
Estimated Price					**$225,000**

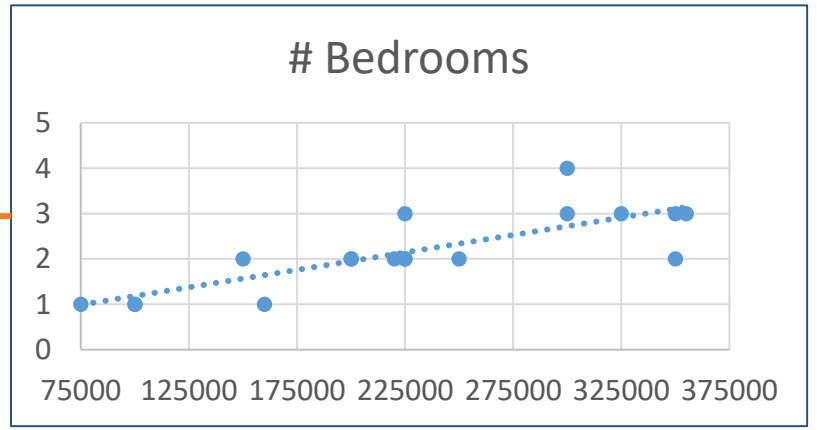
Traditional AVM Approaches

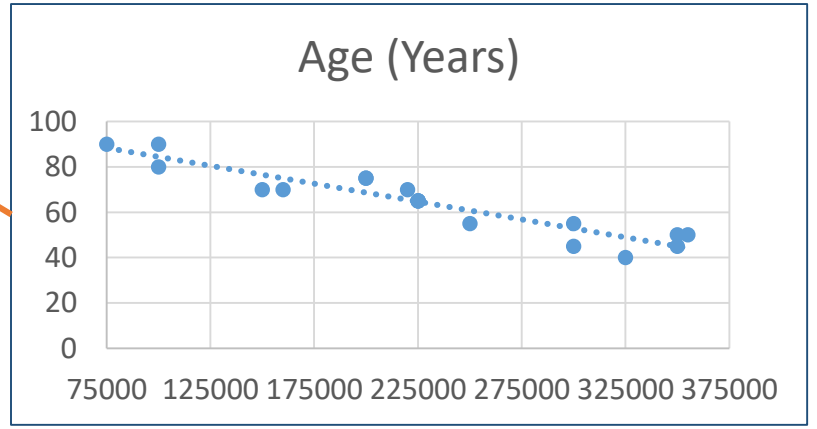
Linear Regression & MRA: Slope of best fitting line

Hedonic Pricing Model

Characteristic	Unit	$ /Unit	Total
Home Size (SqFt.)	1,500	$75	$112,500
# Bedrooms	3	$5,000	$15,000
# Bathrooms	2	$10,000	$20,000
Home Age (Years)	50	($500)	($25,000)
Swimming Pool	Yes	$20,000	$20,000
Lot Size (SqFt)	7,500	$15	$112,500
Price			**$225,000**





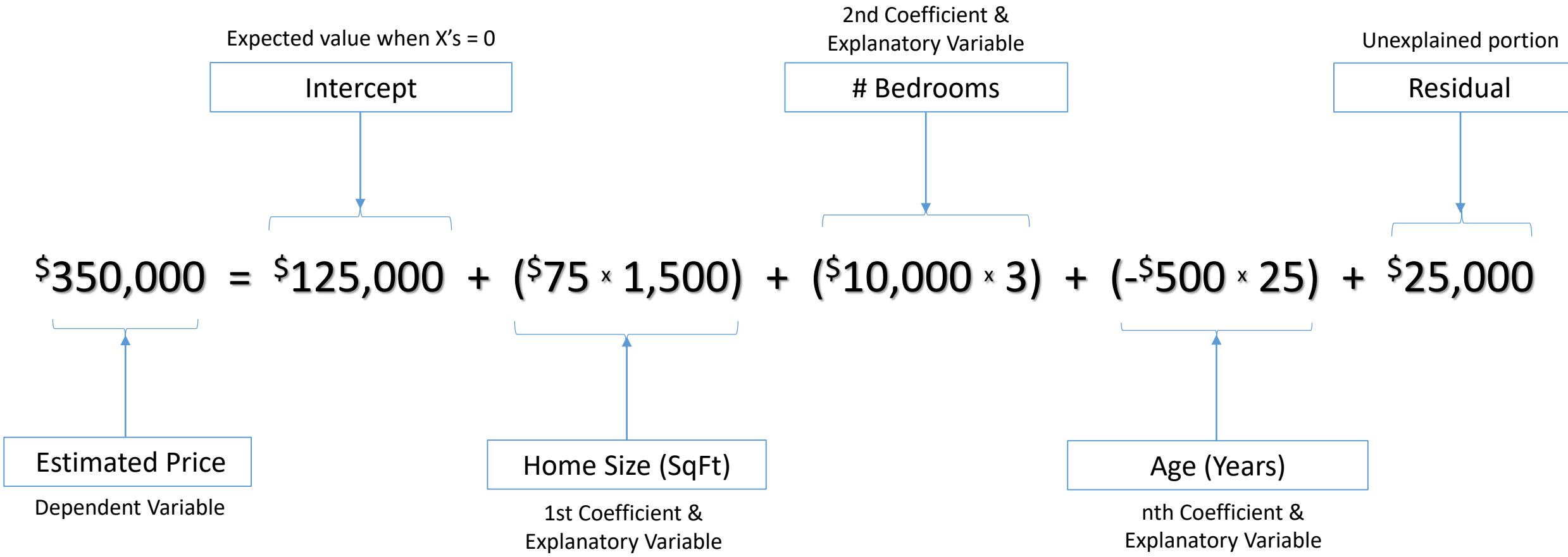


Traditional AVM Approaches

MRA Continued (Example):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots\ldots \beta_n X_n + e$$

Relatively easy to use and understand results.

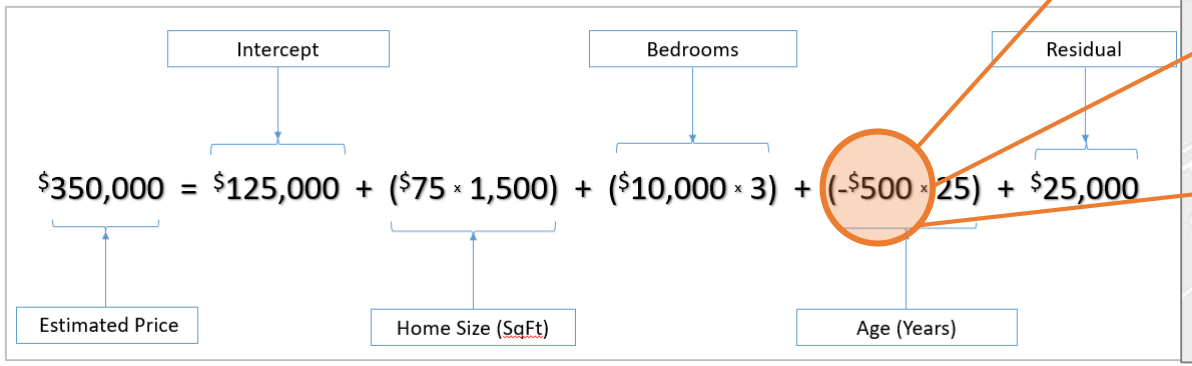


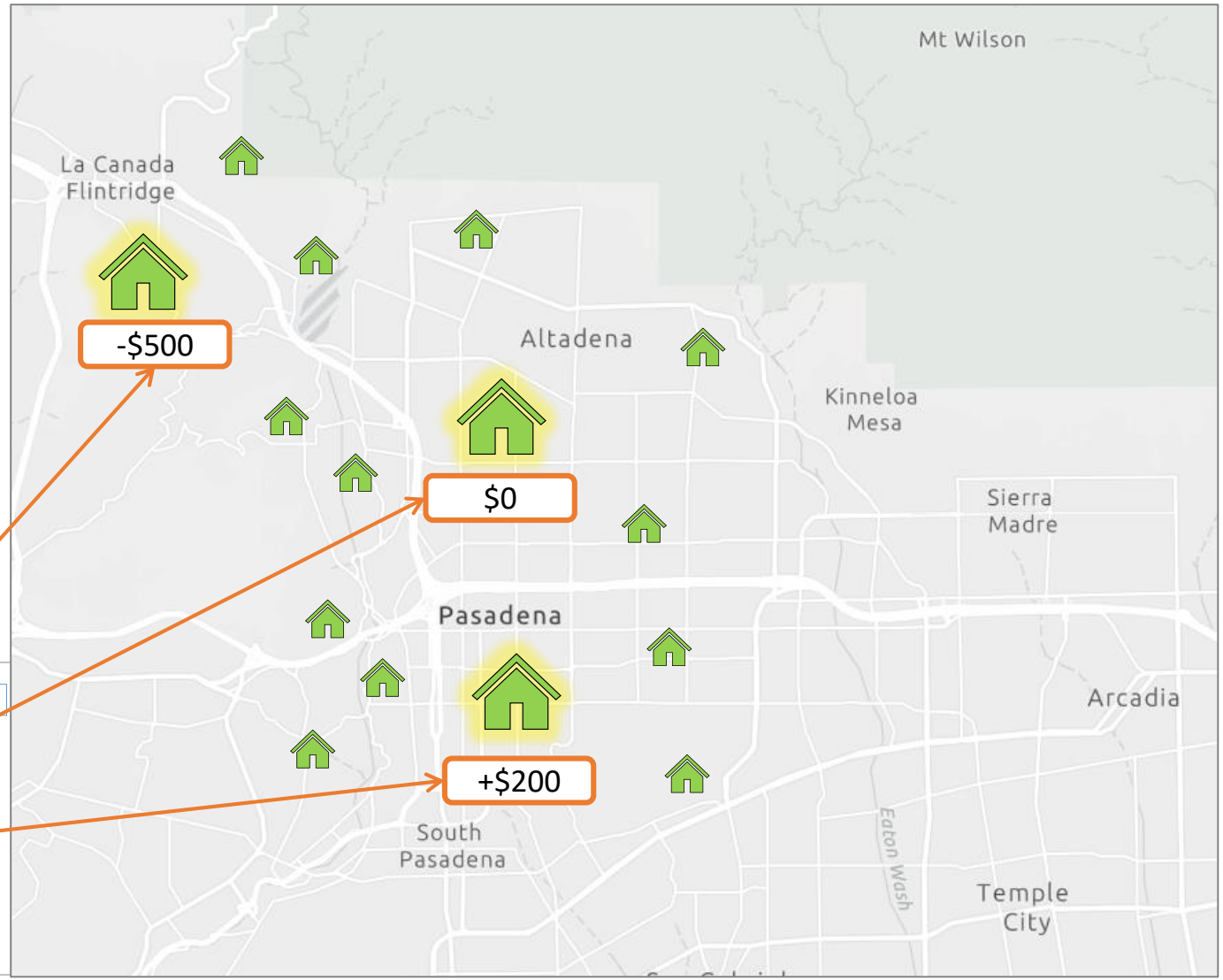
Traditional AVM Approaches

MRA Continued….

Problems!

- Global Model for Local Problem
- Spatial Autocorrelation
- Non-Stationarity



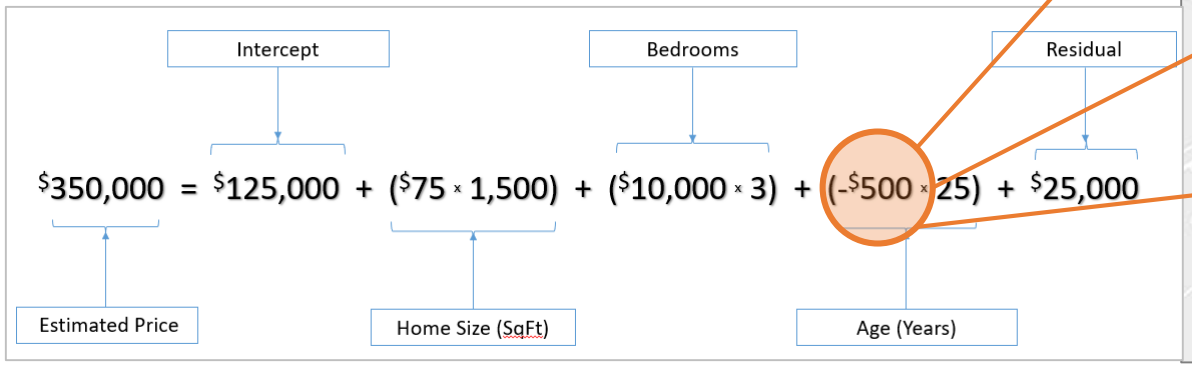


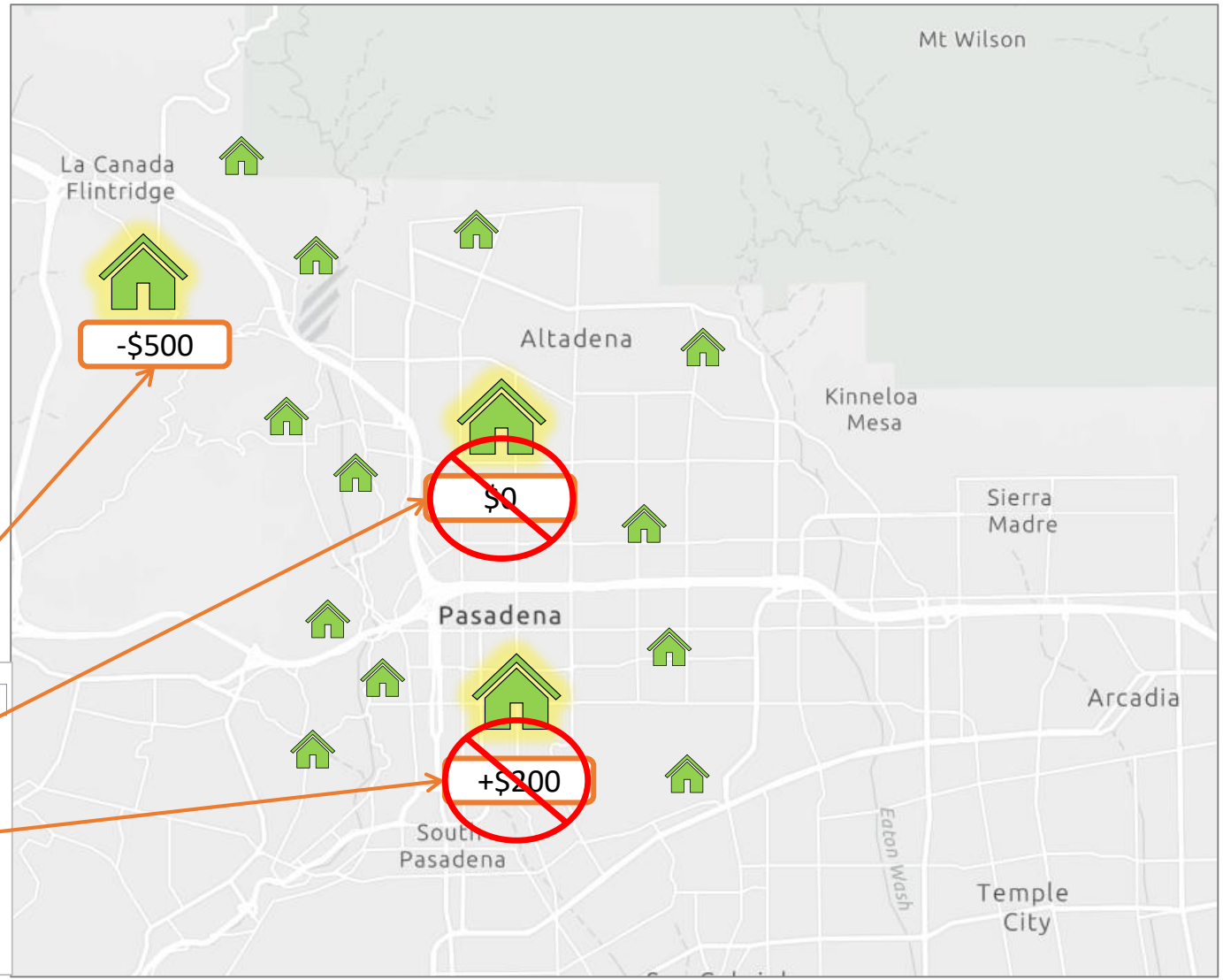
Traditional AVM Approaches

MRA Continued….

Problems!

- Global Model for Local Problem
- Spatial Autocorrelation
- Non-Stationarity
- One (1) Equation for All

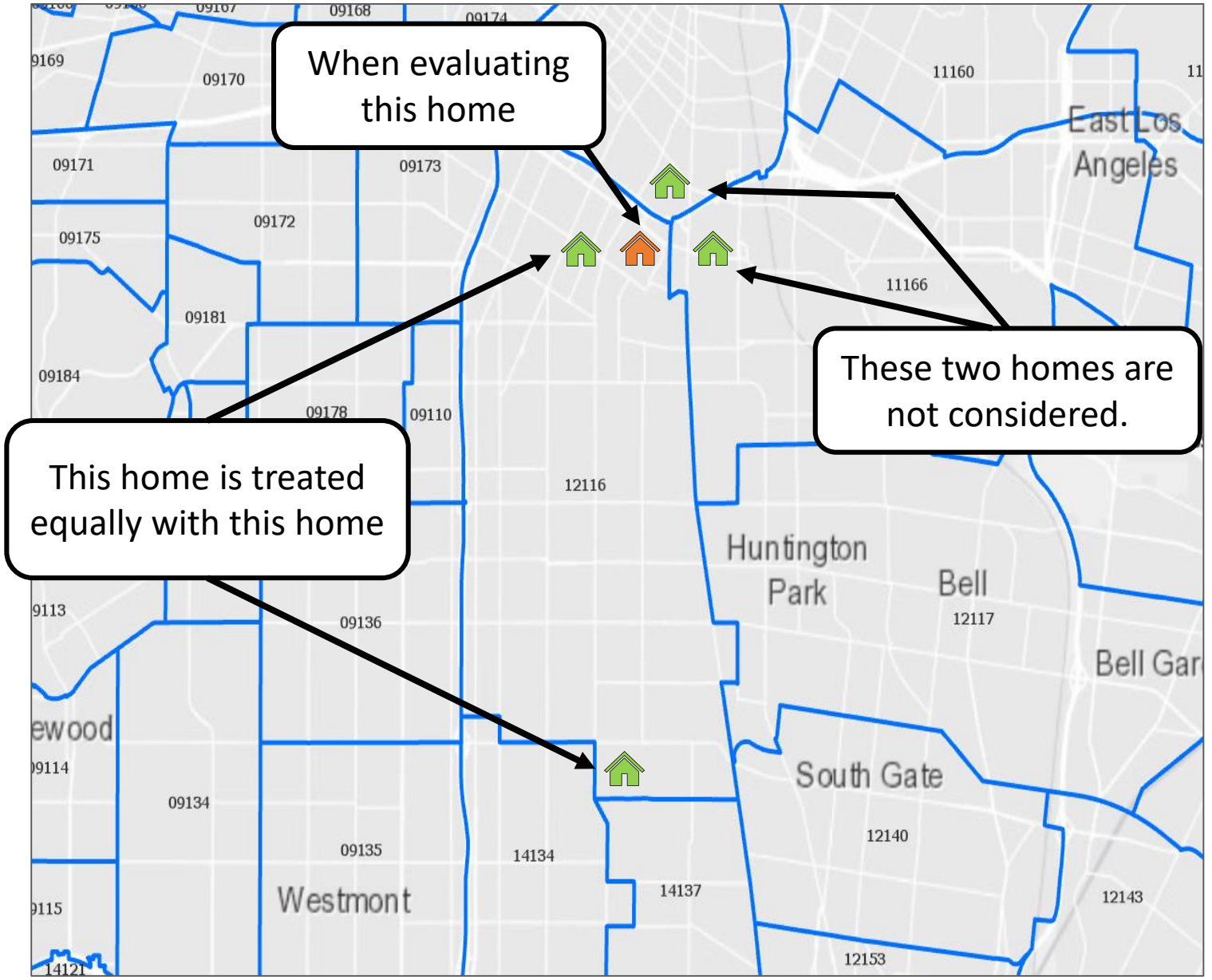




“Clusters”

a solution to non-stationarity & spatial autocorrelation?

- Subjectively Defined
- Potential Edge Effects
- Issues with Sample Size
- Location & Distance All Equal



Geographically Weighted Regression (GWR)

Solution?

- Local Model for Local Problem
- Individual Equations for All
- Reduce Spatial Autocorrelation
- Not Perfect….

$350,000 = $125,000 + ($75 × 1,500) + ($10,000 × 3) + (-$500 × 25) + $25,000

$254,000 = $130,000 + ($50 × 1,500) + ($8,000 × 3) + ($0 × 25) + $30,000

$297,000 = $135,000 + ($65 × 1,500) + ($9,000 × 3) + (+$200 × 25) + $35,000

Geographically Weighted Regression (GWR)

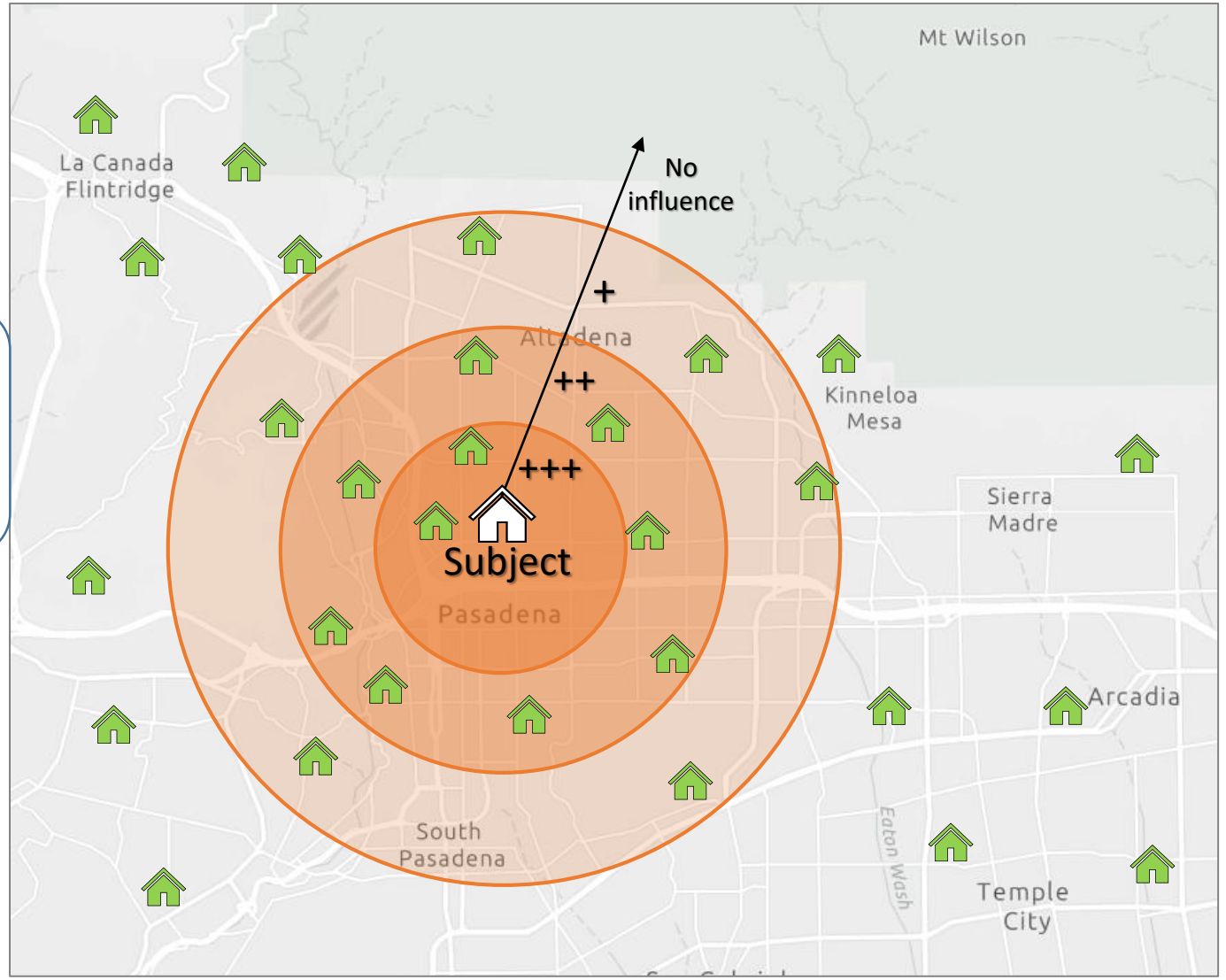
and....

- Near homes given more weight

Everything is related to everything else, but near things are more related than distant things (Tobler, 1970)

Image courtesy of URISA.org https://www.urisa.org/awards/waldo-tobler/. Used here for educational purposes.

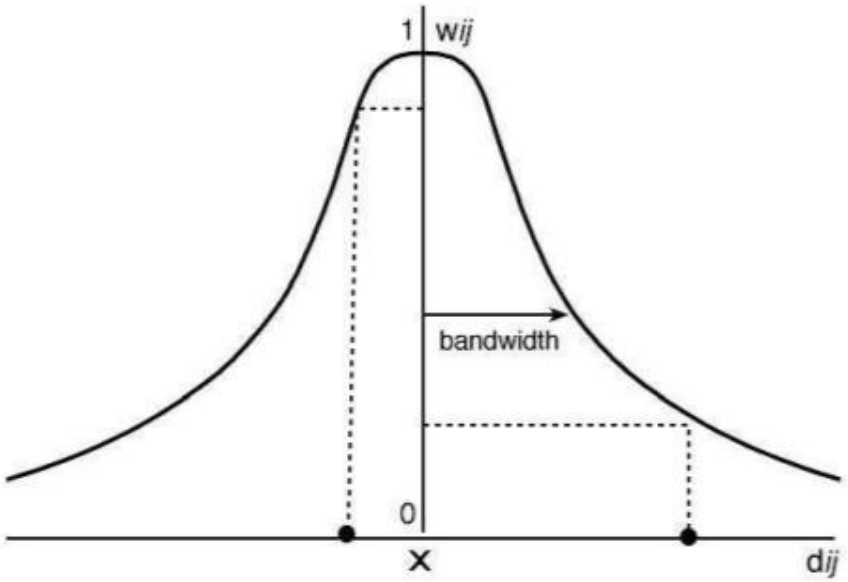
Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46, 234–240.



Geographically Weighted Regression (GWR)

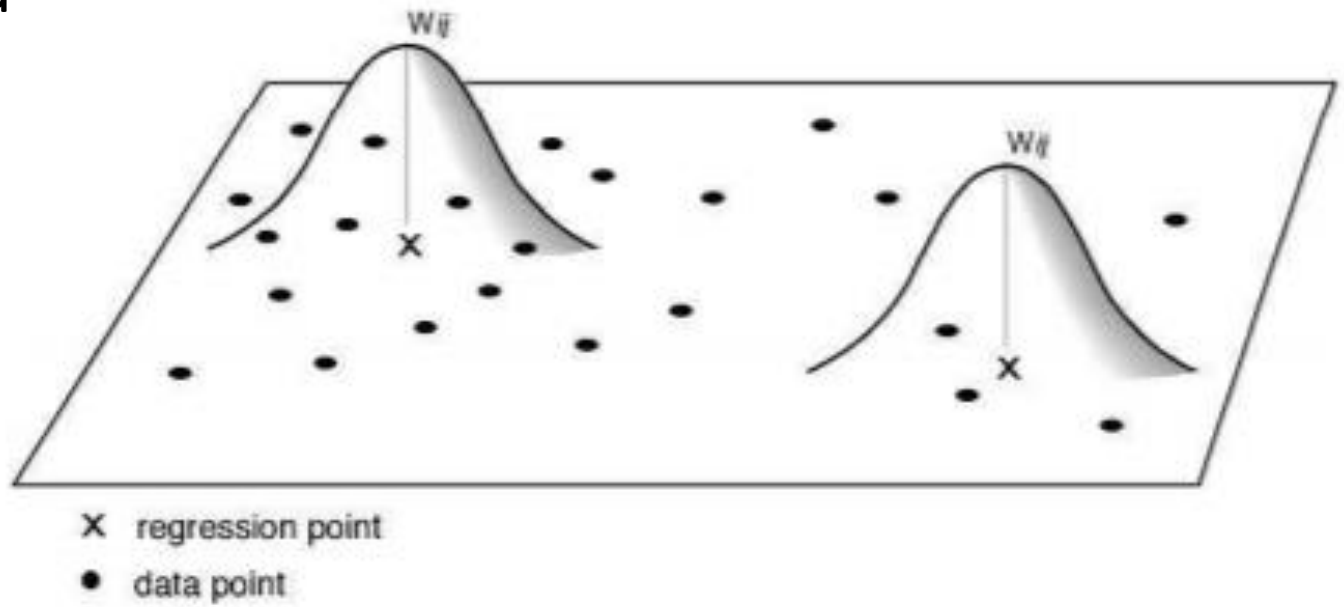
….more accurately

Weighting, or decay, is determined by the “kernel” and “bandwidth”



x regression point wij is the weight of data point *j* at regression point *i*

• data point dij is the distance between regression point *i* and data point *j*



Source: Fotheringham, A. Stewart, Chris Brunsdon, and Martin Charlton. 2002. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Chichester, United Kingdom: John Wiley & Sons*

Goals & Objectives

1. Proof of concept for stated goal (replace the mainframe).
2. Demonstrate replacement AVM process (GWR).
3. Demonstrate county-wide scalability.
4. Provide analysts a place to implement perfection.
5. ~~A perfect solution for estimating home values.~~

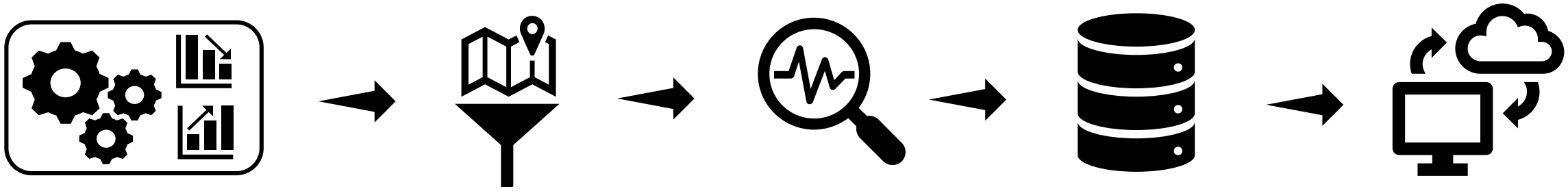
Guiding Principles

- Remember the Mission!
- ALL models are wrong…..but some may be useful.
- PERFECT is the enemy of GOOD.
- Make it run first, make it good later, make it perfect someday.

Development Tasks

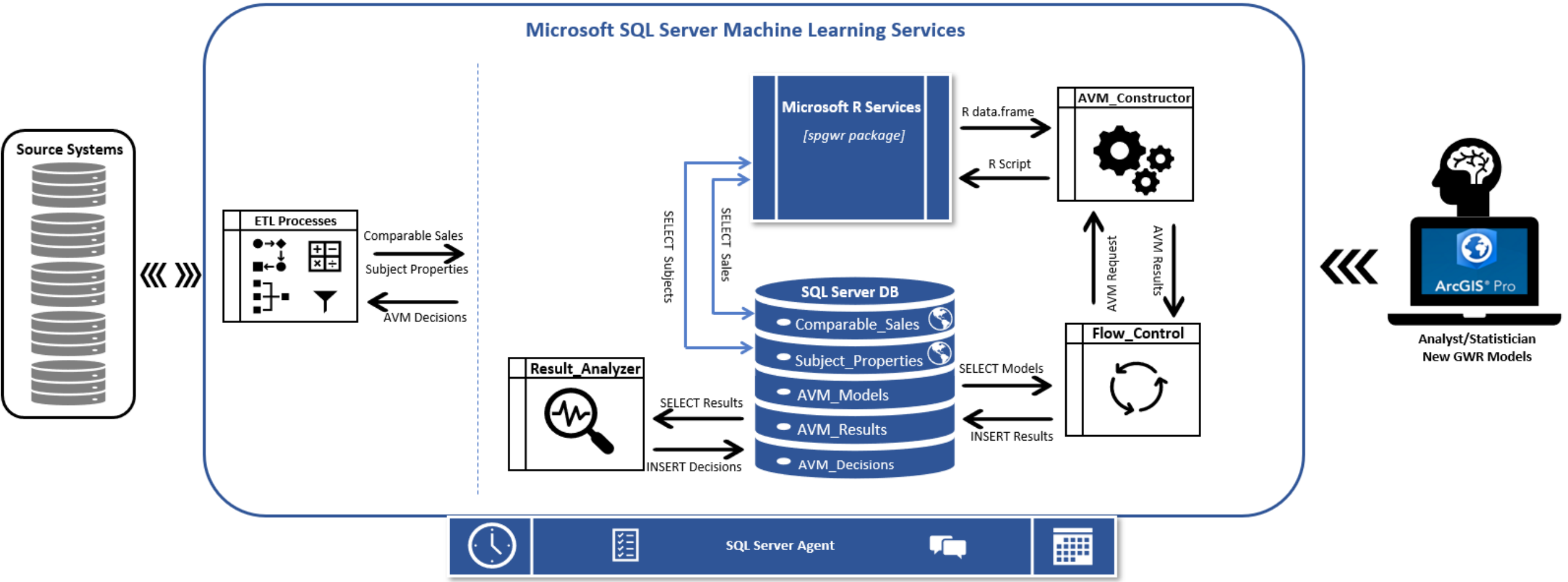
Create a prototype CAMA process.

1. **Identify GWR Models & Create AVM Process (Attributes)**
2. **Control Sample Population (by Boundaries)**
3. **Identify Best Valuations (Results)**
4. **Store & Process Data (Workflow)**
5. **Consider Software & Scalability (Cloud & ML)**



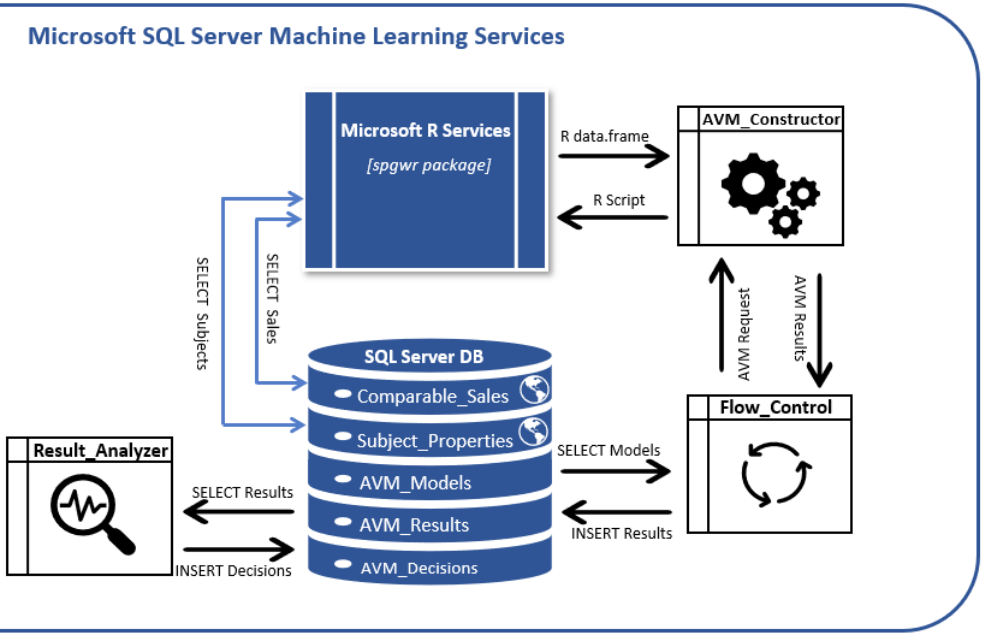
CAMA Prototype

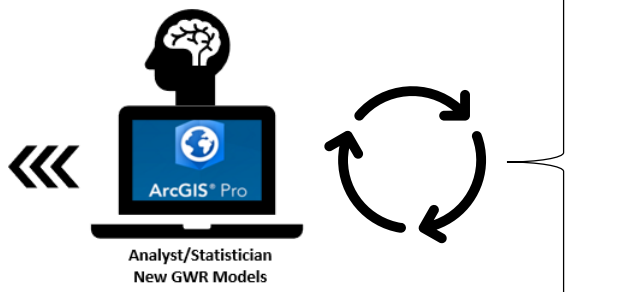
CAMA System



Creating a CAMA Prototype

1. **Identify GWR Models & Create AVM Process:** *Using multiple attribute models*.
2. Control Sample Population (by Boundaries)
3. Identify Best Valuations (Results)
4. Store & Process Data (Workflow)
5. Consider Software & Scalability (Cloud & ML)





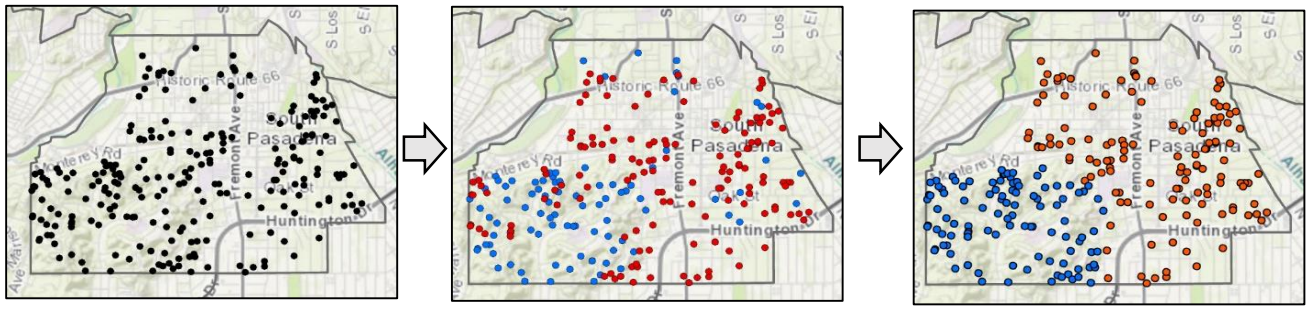
Determine GWR Models

- Identify sample neighborhoods
- Explore housing attributes
- Test attribute models
- Diagnose issues
- Transform variables
- Identify/create missing attributes

Identify GWR Models

ArcGIS Pro Analysis Tools:

- Exploratory Regression
- Clustering & Hot Spot
- OLS (MRA)
- Moran's I
- GWR



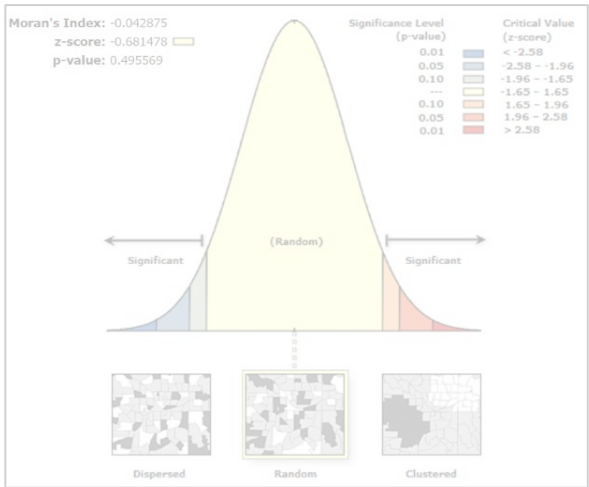
Summary of OLS Results - Model Variables

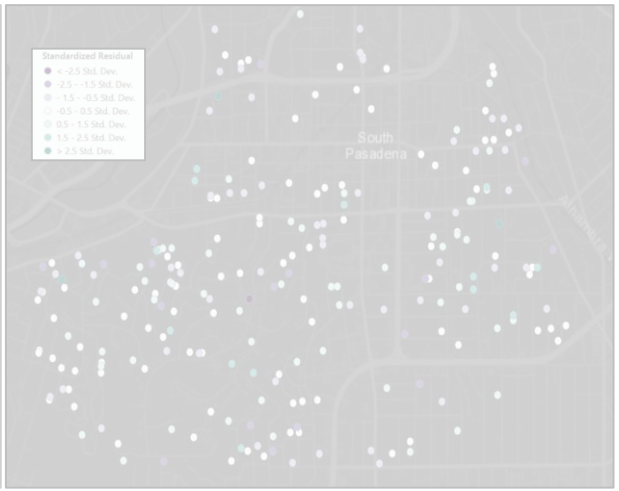
Variable	Coefficient [a]	StdError	t-Statistic	Probability [b]	Robust_SE	Robust_t	Robust_Pr [b]	VIF [c]
Intercept	433660.70803	23939.199111	18.115088	0.000000*	23296.898504	18.614525	0.000000*	--------
REVERSETIMEO	-4103.044006	541.842492	-7.572392	0.000000*	542.981524	-7.556508	0.000000*	1.009788
BEDPLUSBATHC	15804.615233	6779.090481	2.331377	0.020849*	7114.500332	2.221465	0.027577*	3.618845
MAINSTRUCTUR	125.348087	15.500976	8.086464	0.000000*	15.897259	7.884887	0.000000*	3.640644
RCNOTHERTREN	0.662750	0.198789	3.333934	0.001053*	0.186988	3.544353	0.000514*	1.020220

OLS Diagnostics

Input Features:	CerritosNorth	Dependent Variable:	SUGGESTEDPRICE
Number of Observations:	182	Akaike's Information Criterion (AICc) [d]:	4459.413599
Multiple R-Squared [d]:	0.712009	Adjusted R-Squared [d]:	0.705500
Joint F-Statistic [e]:	109.400420	Prob(>F), (4,177) degrees of freedom:	0.000000*
Joint Wald Statistic [e]:	479.504600	Prob(>chi-squared), (4) degrees of freedom:	0.000000*
Koenker (BP) Statistic [f]:	6.459220	Prob(>chi-squared), (4) degrees of freedom:	0.167378
Jarque-Bera Statistic [g]:	5.228495	Prob(>chi-squared), (2) degrees of freedom:	0.073223

```
********************************************************************************
Choose 4 of 22 Summary
                                  Highest Adjusted R-Squared Results
AdjR2    AICc   JB K(BP)  VIF   SA   Model
 0.79 3755.78 0.00  0.00 3.94 0.00  -REVERSETIMEOFSALE***  +LOTSIZEGIS**  +MAINSTRUCTURESQFT***  +RCNTOTAL***
 0.78 3758.94 0.00  0.00 3.93 0.00  -REVERSETIMEOFSALE***  +LOTSIZEUSEABLE**  +MAINSTRUCTURESQFT***  +RCNTOTAL***
 0.78 3761.06 0.01  0.00 4.20 0.00  -REVERSETIMEOFSALE***  +LOTSIZEGIS***  +MAINSTRUCTURESQFT***  +RCNMAINSTRUCTURE***
                                     Passing Models
AdjR2    AICc        JB       K(BP)    VIF      SA         Model
0.687582 3808.576135 0.110114 0.004730 4.639329 0.110412  +LOTSIZEUSEABLE***  +QUALITYCLASSNUMBER***  +EFFECTIVEAGE***
********************************************************************************
```

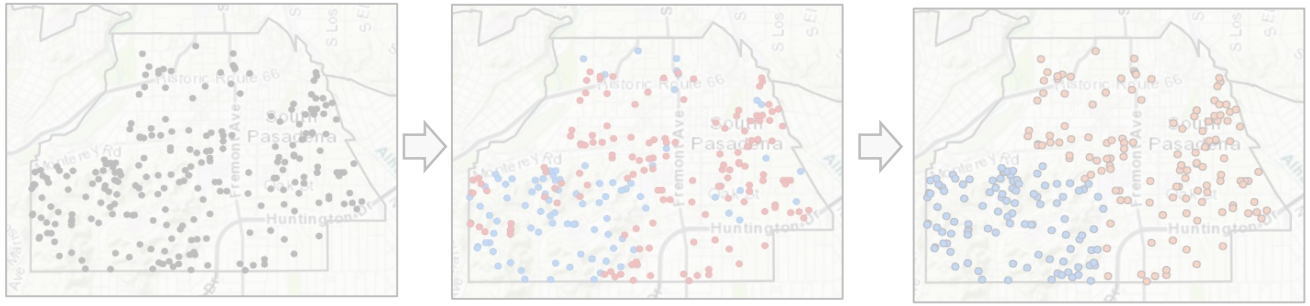




Identify GWR Models

ArcGIS Pro Analysis Tools:

- Exploratory Regression
- Clustering & Hot Spot
- OLS (MRA)
- Moran's I
- GWR



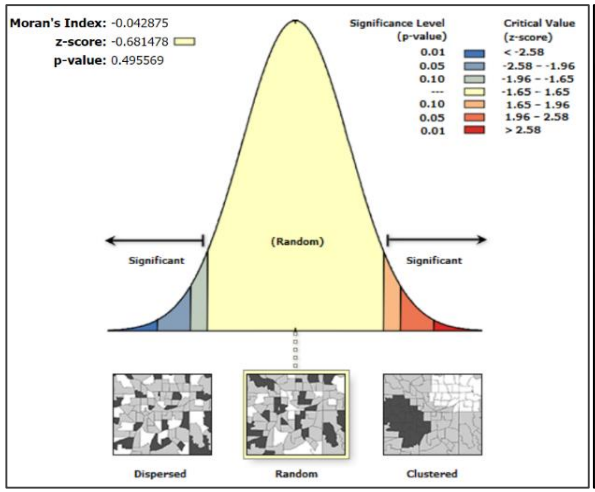
Summary of OLS Results - Model Variables

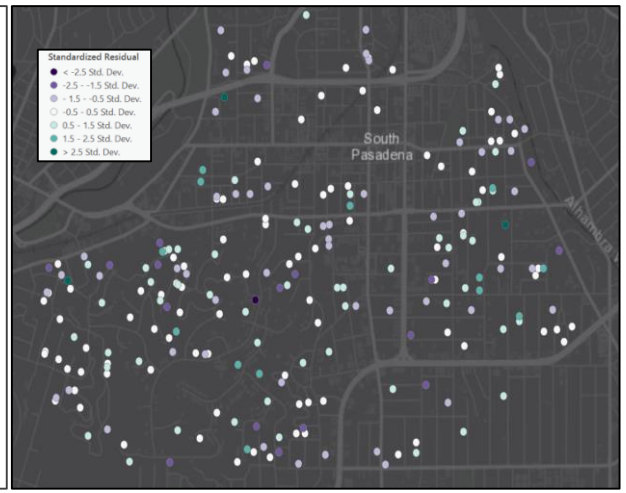
Variable	Coefficient [a]	StdError	t-Statistic	Probability [b]	Robust_SE	Robust_t	Robust_Pr [b]	VIF [c]
Intercept	433660.70803	23939.199111	18.115088	0.000000*	23296.898504	18.614525	0.000000*	--------
REVERSETIMEO	-4103.044006	541.842492	-7.572392	0.000000*	542.981524	-7.556508	0.000000*	1.009788
BEDPLUSBATHC	15804.615233	6779.090481	2.331377	0.020849*	7114.500332	2.221465	0.027577*	3.618845
MAINSTRUCTUR	125.348087	15.500976	8.086464	0.000000*	15.897259	7.884887	0.000000*	3.640644
RCNOTHERTREN	0.662750	0.198789	3.333934	0.001053*	0.186988	3.544353	0.000514*	1.020220

OLS Diagnostics

Input Features:	CerritosNorth	Dependent Variable:	SUGGESTEDPRICE
Number of Observations:	182	Akaike's Information Criterion (AICc) [d]:	4459.413599
Multiple R-Squared [d]:	0.712009	Adjusted R-Squared [d]:	0.705500
Joint F-Statistic [e]:	109.400420	Prob(>F), (4,177) degrees of freedom:	0.000000*
Joint Wald Statistic [e]:	479.504600	Prob(>chi-squared), (4) degrees of freedom:	0.000000*
Koenker (BP) Statistic [f]:	6.459220	Prob(>chi-squared), (4) degrees of freedom:	0.167378
Jarque-Bera Statistic [g]:	5.228495	Prob(>chi-squared), (2) degrees of freedom:	0.073223

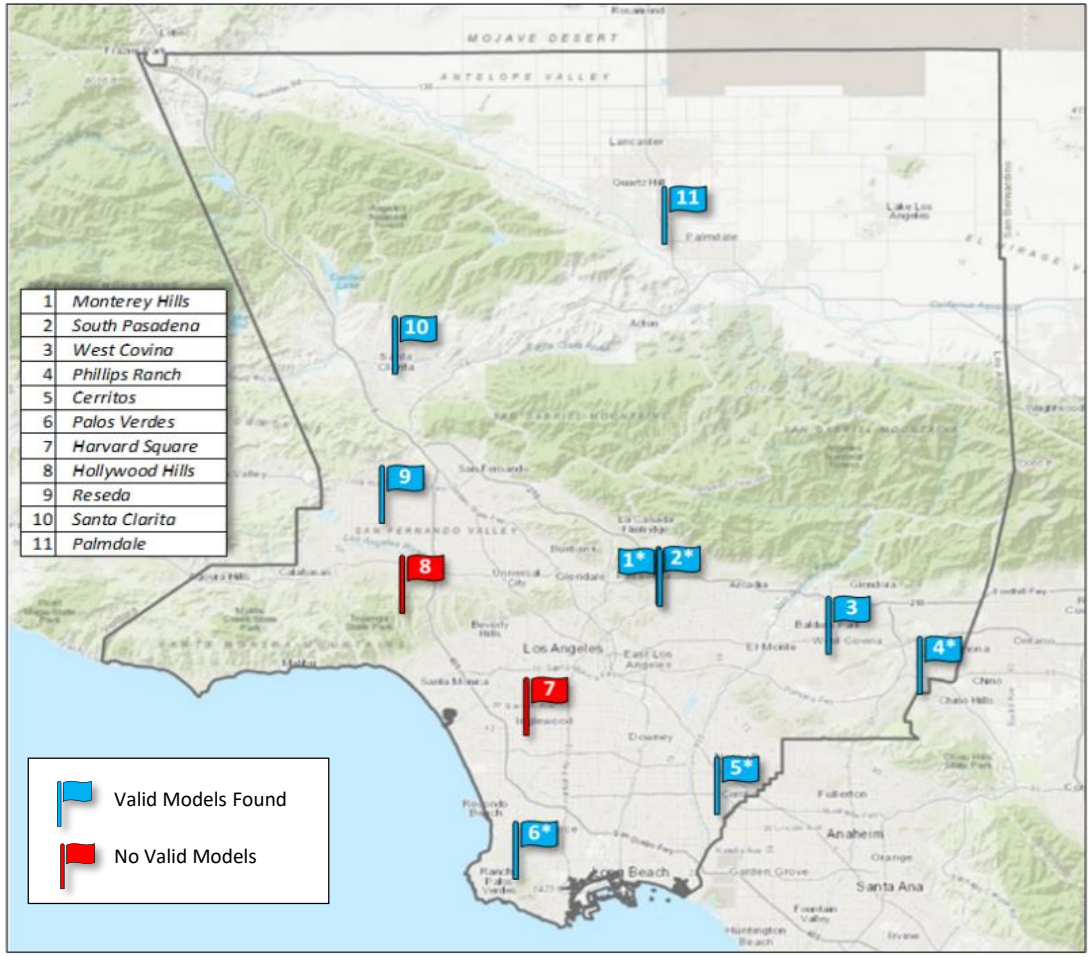
```
********************************************************************************
Choose 4 of 22 Summary
                                Highest Adjusted R-Squared Results
AdjR2    AICc   JB K(BP)  VIF   SA   Model
 0.79 3755.78 0.00  0.00 3.94 0.00  -REVERSETIMEOFSALE***  +LOTSIZEGIS**  +MAINSTRUCTURESQFT***  +RCNTOTAL***
 0.78 3758.94 0.00  0.00 3.93 0.00  -REVERSETIMEOFSALE***  +LOTSIZEUSEABLE**  +MAINSTRUCTURESQFT***  +RCNTOTAL***
 0.78 3761.06 0.01  0.00 4.20 0.00  -REVERSETIMEOFSALE***  +LOTSIZEGIS***  +MAINSTRUCTURESQFT***  +RCNMAINSTRUCTURE***
                                      Passing Models
AdjR2    AICc        JB       K(BP)    VIF      SA         Model
0.687582 3808.576135 0.110114 0.004730 4.639329 0.110412  +LOTSIZEUSEABLE***  +QUALITYCLASSNUMBER***  +EFFECTIVEAGE***
********************************************************************************
```





Identify GWR Models

Sample Neighborhoods



Property Characteristics

Sales from 2016 & 2017 ≈ ***100,000*** **sales**.

#	Field Name		
1	AIN	13	Quality Class Number
2	Sale Price	14	Central Air
3	Reverse Time of Sale	15	Pool
4	Lot Size	16	Has View
5	Lot Size Useable	17	Has Nuisance
6	Lot Size Excess	18	RCN Main Structure
7	Main Structure SqFt	19	RCN Other Structures
8	Bedroom Count	20	Total RCN
9	Bathroom Count	21	Total RCNLD
10	Room Count	22	LocationID
11	Age	23	X/Y Point
12	Effective Age		

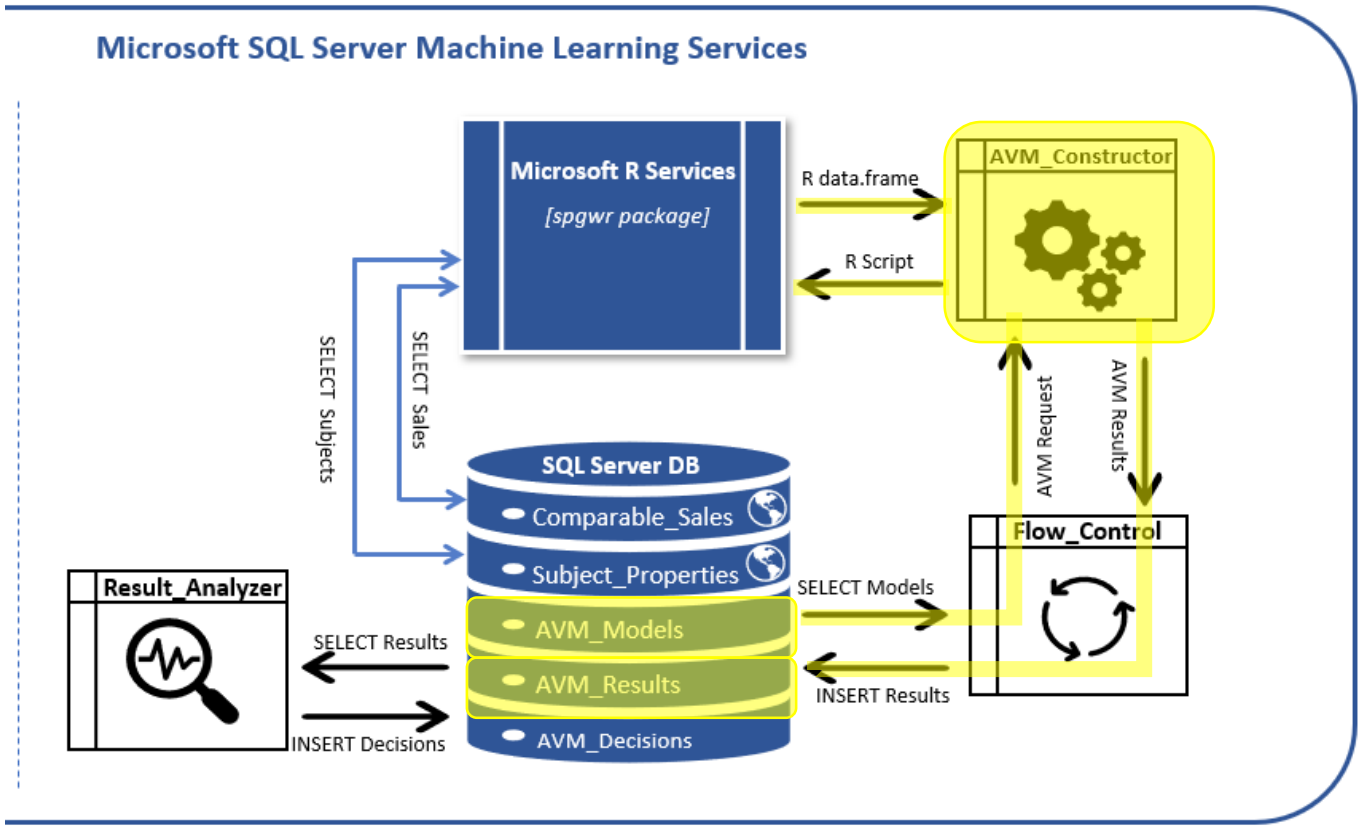
Identify GWR Models

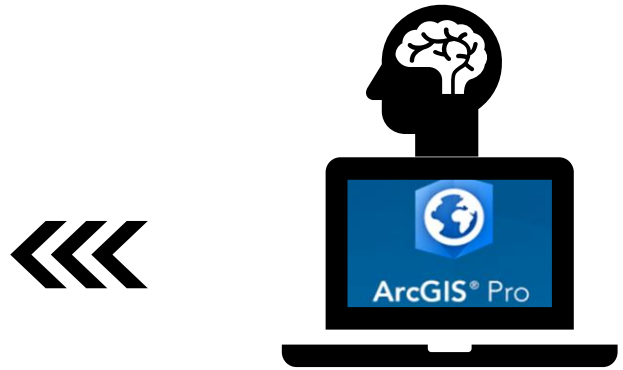
Five GWR Models Selected

Model #	**1**	**2**	**3**	**4**	**5**
Area	Phillips Ranch	Monterey Hills	South Pasadena	Cerritos	Palos Verdes
V1	*ReverseTimeOfSale*	*ReverseTimeOfSale*	*ReverseTimeOfSale*	*ReverseTimeOfSale*	*ReverseTimeOfSale*
V2	*LotSizeUseable*	*LotSizeGIS*	*LotSizeUseable*	*MainStructureSqFT*	*LotSizeGIS*
V3	*Age*	*LotSizeExcess*	*EffectiveAge*	*RCNOtherTrended*	*Age*
V4	*MainStructureSqFt*	*Age*	*MainStructureSqFt*	*BedPlusBathCount*	*RCNLDMainStructure*
V5	*HasPool*	*MainStructureSqFt*	*RCNTotal*		*RCNLDOtherStructure*
V6		*RCNTotal*			
# Samples	*169*	*110*	*135*	*182*	*194*
OLS AIC/R2	*4169 / .749*	*3039 / .661*	*3760 / .782*	*4459 / .705*	*5225 / .717*
GWR AIC/R2	*4170 / .748*	*3036 / .673*	*3736 / .834*	*4443 / .741*	*5210 / .748*

Creating a CAMA Prototype

1. **Identify GWR Models & Create AVM Process:** *Using multiple attribute models.*
2. Control Sample Population (by Boundaries)
3. Identify Best Valuations (Results)
4. Store & Process Data (Workflow)
5. Consider Software & Scalability (Cloud & ML)





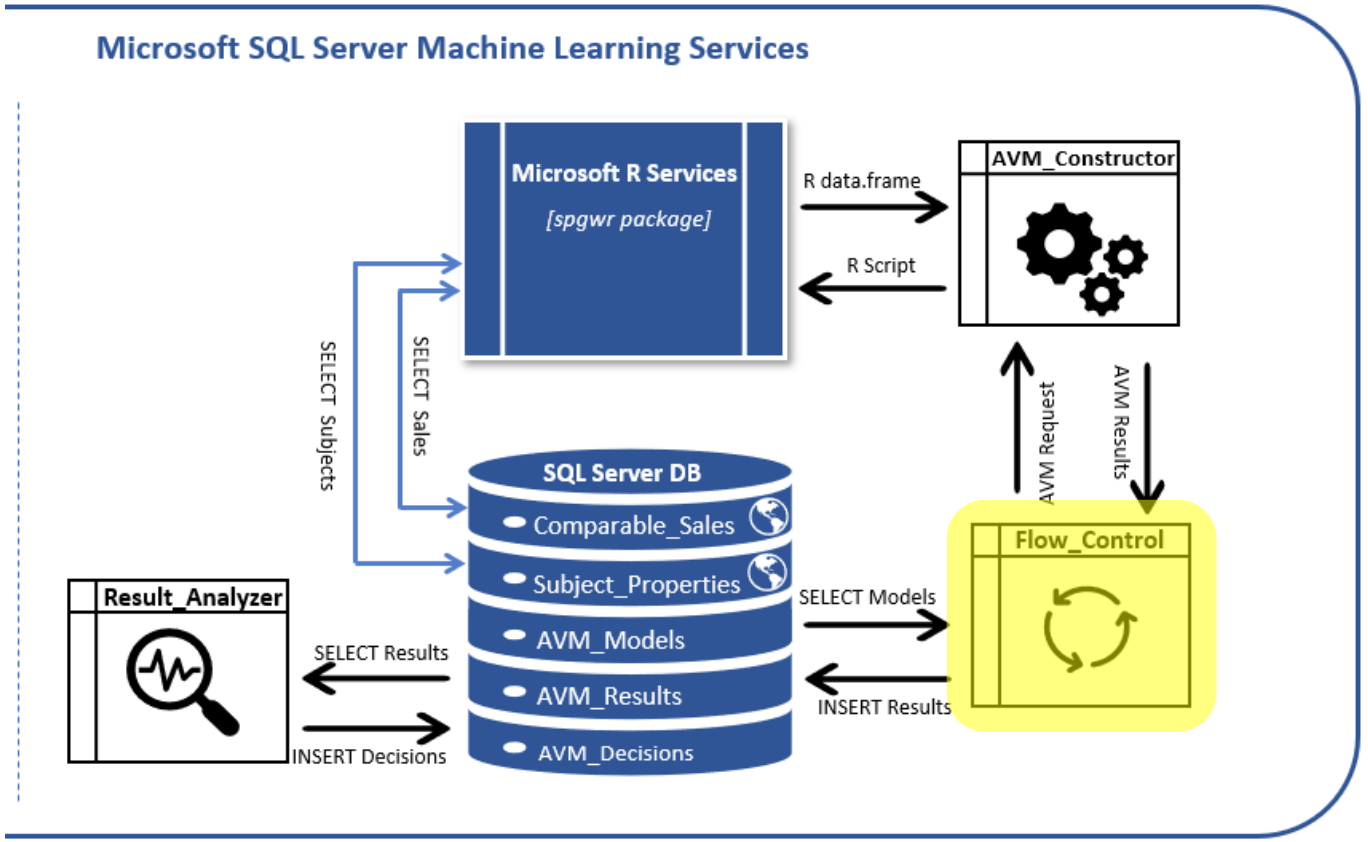
Model #	1	2	3	4	5
Area	Phillips Ranch	Monterey Hills	South Pasadena	Cerritos	Palos Verdes
V1	*ReverseTimeOfSale*	*ReverseTimeOfSale*	*ReverseTimeOfSale*	*ReverseTimeOfSale*	*ReverseTimeOfSale*
V2	*LotSizeUseable*	*LotSizeGIS*	*LotSizeUseable*	*MainStructureSqFT*	*LotSizeGIS*
V3	*Age*	*LotSizeExcess*	*EffectiveAge*	*RCNOtherTrended*	*Age*
V4	*MainStructureSqFt*	*Age*	*MainStructureSqFt*	*BedPlusBathCount*	*RCNLDMainStructure*
V5	*HasPool*	*MainStructureSqFt*	*RCNTotal*		*RCNLDOtherStructure*
V6		*RCNTotal*			
# Samples	*169*	*110*	*135*	*182*	*194*
OLS AIC/R2	*4169 / .749*	*3039 / .661*	*3760 / .782*	*4459 / .705*	*5225 / .717*
GWR AIC/R2	*4170 / .748*	*3036 /.673*	*3736 / .834*	*4443 / .741*	*5210 / .748*

Creating a CAMA Prototype

1. Identify GWR Models & Create AVM Process: Using multiple attribute models.
2. **Control Sample Population (by Boundaries):** *Filter & Loop*
3. Identify Best Valuations (Results)
4. Store & Process Data (Workflow)
5. Consider Software & Scalability

Boundary Choices

1. School Districts
2. Assessor Clusters
3. Communities
4. No Boundaries
5. Others…….??



Creating a CAMA Prototype

1. Identify GWR Models & Create AVM Process: Using multiple attribute models.
2. Control Sample Population (by Boundaries): Filter & Loop
3. **Identify Best Valuations (Results):** *For each home in the population.*
4. Store & Process Data (Workflow)
5. Consider Software & Scalability

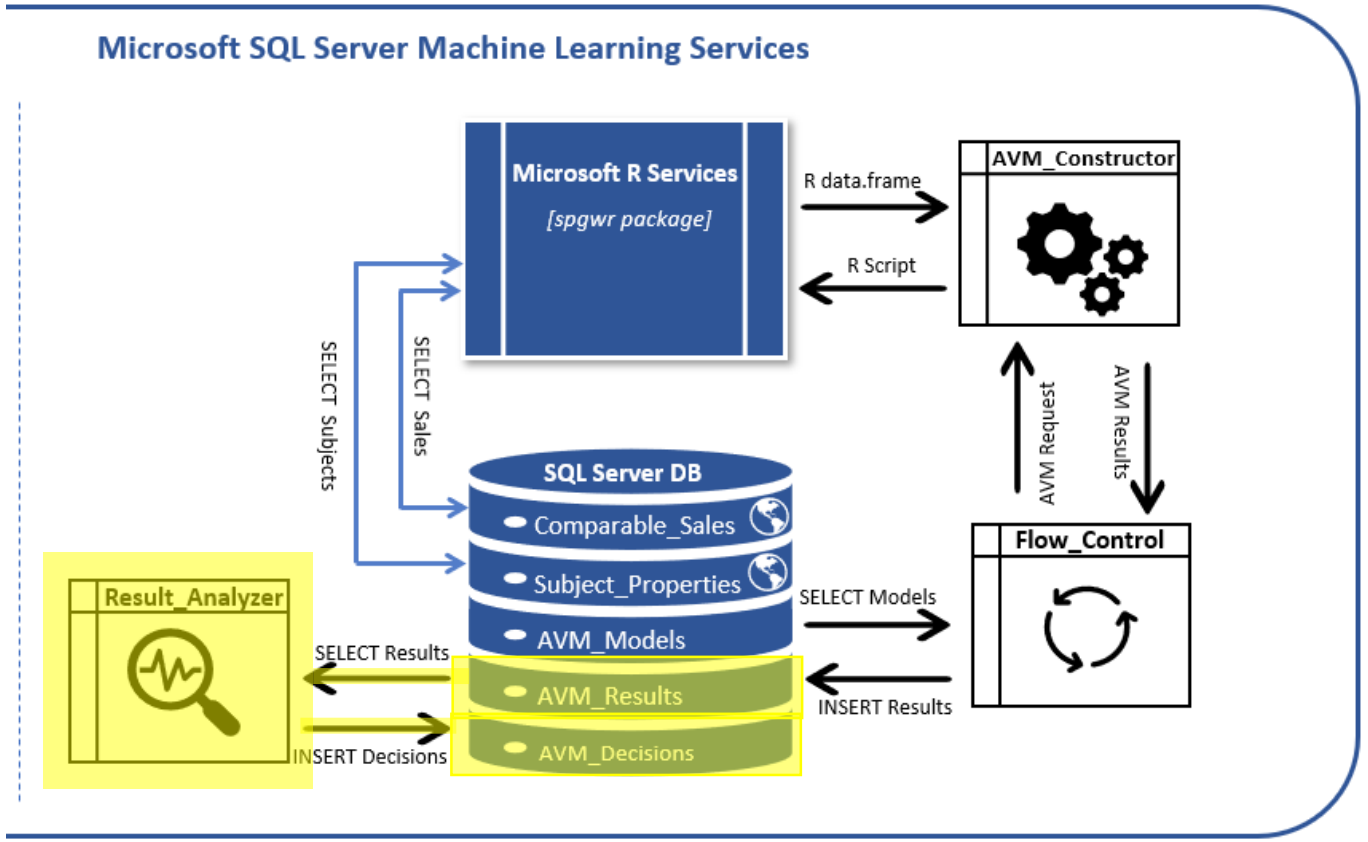
Select Best of 5 Valuations *(Best?)*

First Cut:

1. R2 > =.70 AND Variance <= .05

Second Cut **(in order):**

1. Lowest AIC Score (+-3)
2. Highest R2
3. Lowest Variance

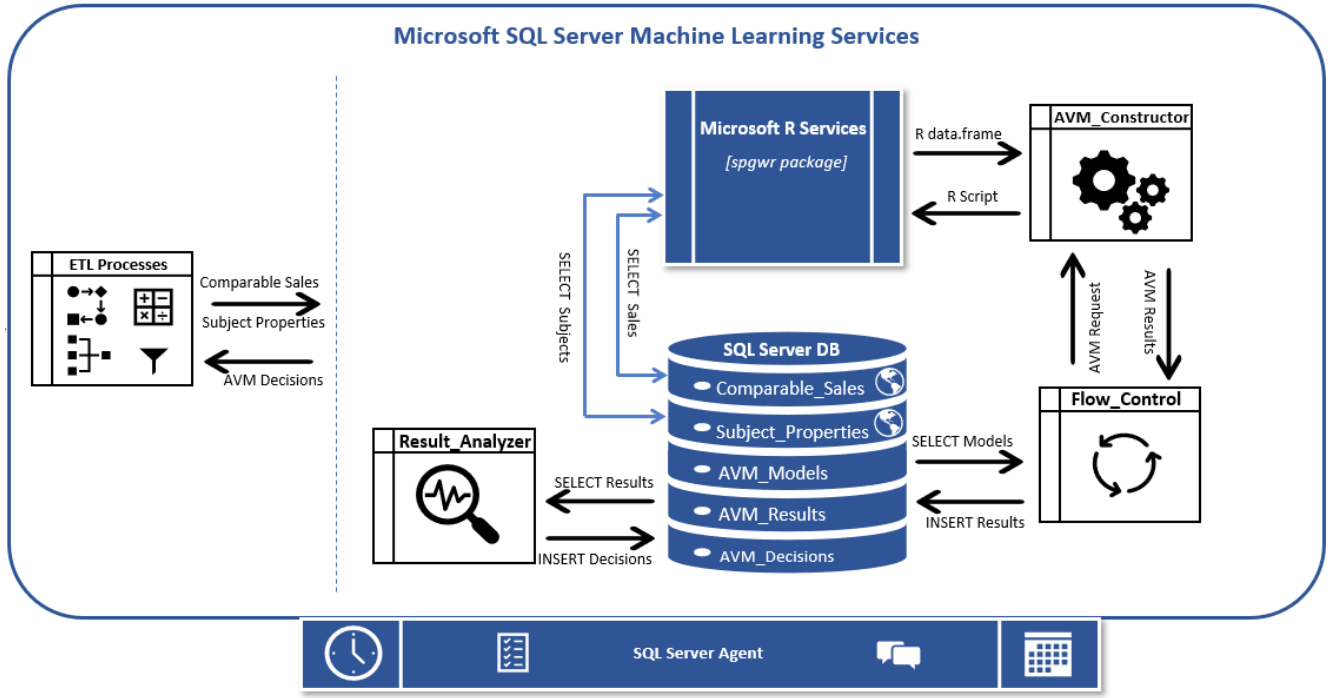


Creating a CAMA Prototype

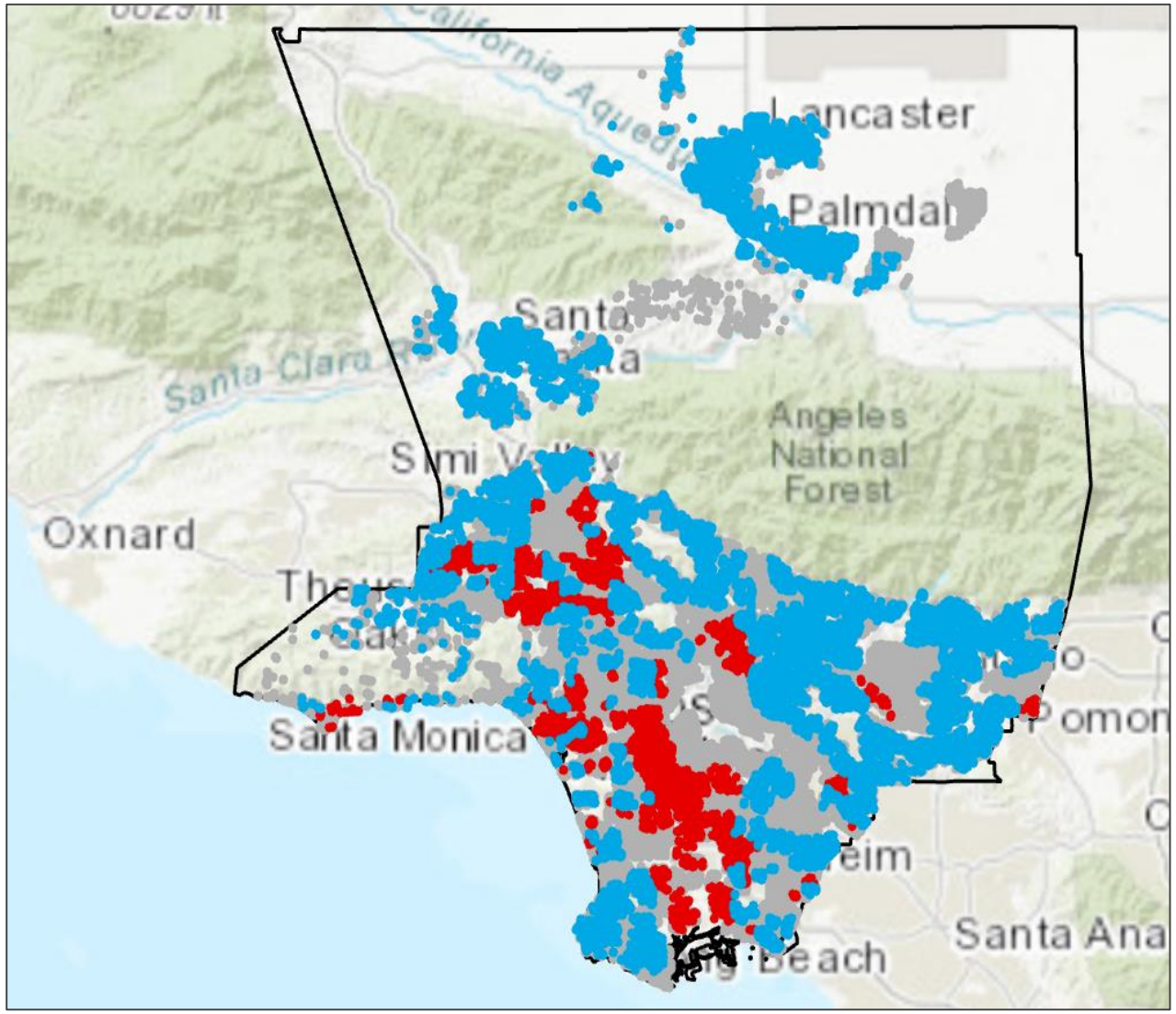
1. Identify GWR Models & Create AVM Process: Using multiple attribute models.
2. Control Sample Population (by Boundaries): Filter & Loop
3. Identify Best Valuations (Results): For each home in the population.
4. **Store & Process Data** (*Workflow*)
5. **Consider Software & Scalability** (*Cloud & ML*)

MS Machine Learning Services

1. Storage: SQL Server database
2. Statistical Process: R using spgwr package
3. Other Processes: T-SQL & Stored Procedures
4. Scheduling & Automation: SQL Server Agent
5. Scalability: Repeated in Azure Cloud



Sample Results & Work Needed



Good Value Estimate
R2 >= .70 AND Variance <= .05

Poor Value Estimate
R2 <= .50 AND Variance >= .15

Neutral Value Estimate
Everything else

Work Needed

- ❑ Add Test Scenarios & Clean-Up Workflow
- ❑ Compare Results to Existing Process
- ❑ Analyze/Improve Boundaries
- ❑ Data Quality, Cleansing, & Collection Standards
- ❑ Comparable Sales Validation Process
- ❑ Guestimate -> Prediction -> IAAO standards