

Real-world Use Case

House Recommender

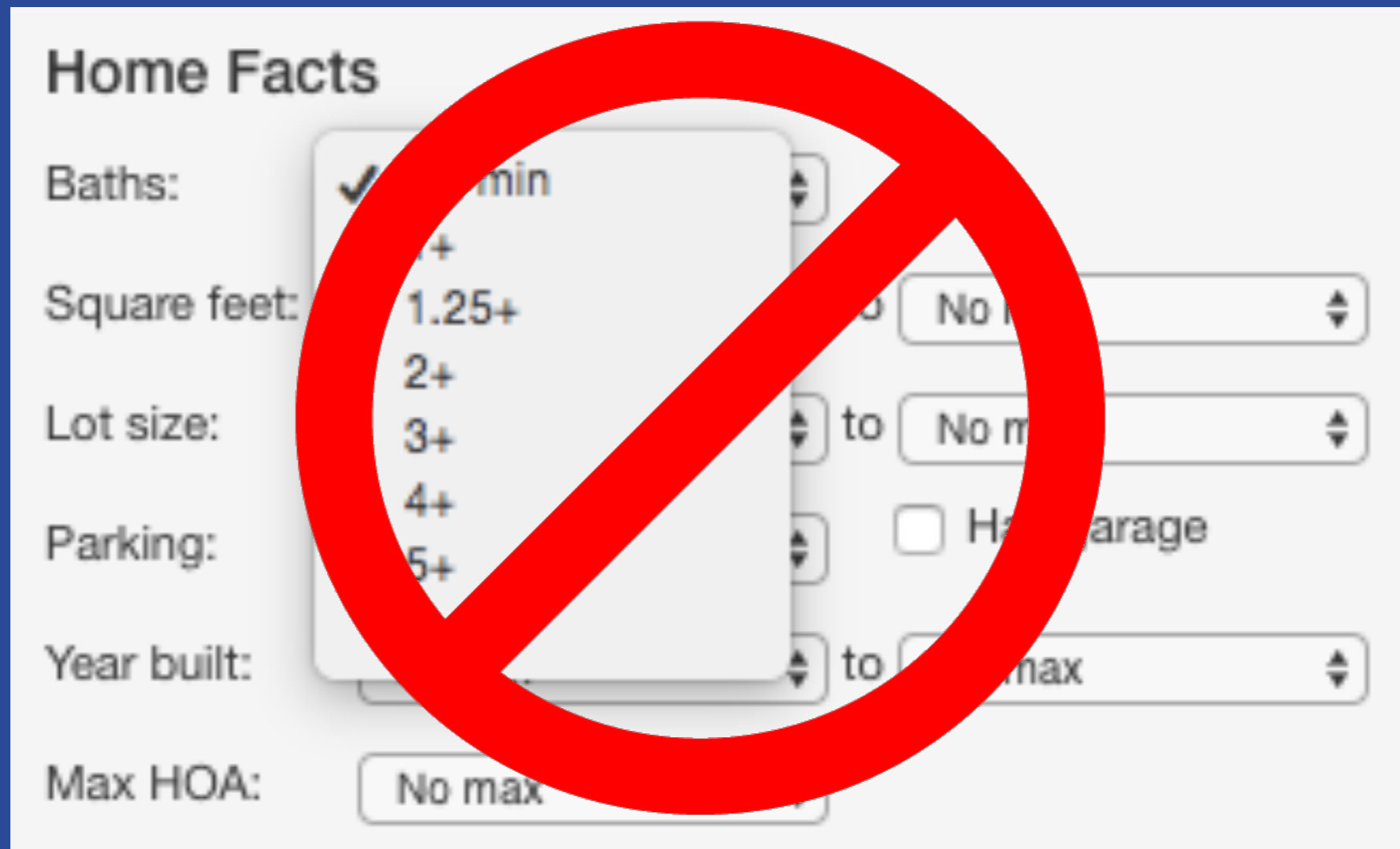
Poul Petersen

CIO, BigML

BigML for Alexa

Let's build a recommender

Typical way to shop for a home...



Recommender Idea

Sample



?



?



?



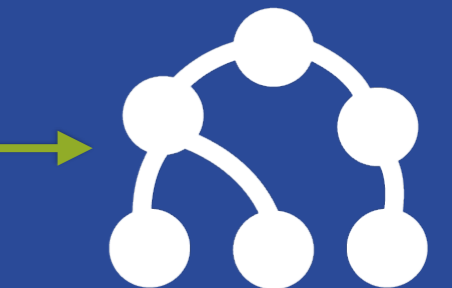
?



All Homes
For sale



Preference
Data



Preference
Model

... then use the Preference Model to filter all the homes on the market

Recommender Problem #1

What if there are really unusual homes in the data?

- A mansion with 20 bathrooms
- A home with no bedrooms
- A lot size that is smaller than the home?



We don't want to show these as suggestions because they are unusual.... How do we detect anomalies?

Anomaly Detection

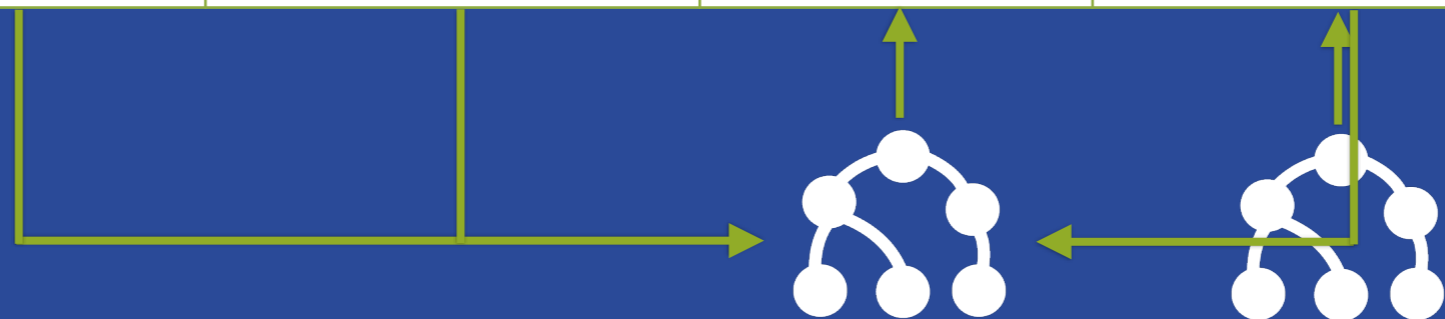


- We want to find and remove unusual houses.
- We create an **Anomaly** Detector and examine the top anomalies.
- We filter out any houses with a “high” anomaly score.

ML to fix missing data...

- Let's use Machine Learning...

SQFT	PRICE	BEDS	BATHS
3,125	\$530,000	5	3
2,100	\$460,000	4	2
1,200	\$250,000	3	1.5
3,950	\$610,000	6	4



What happened to
being *easy*?

BEDS

BATHS

WhizzML Gallery

- We had a **Dataset** with missing values.
- We want to apply an algorithm to fix the missing values with Machine Learning
- Rather than write the algorithm, we can find what we need in the **WhizzML** public gallery.
- Once we clone the **Script** we can use it again and again.
- We can write new ones too!

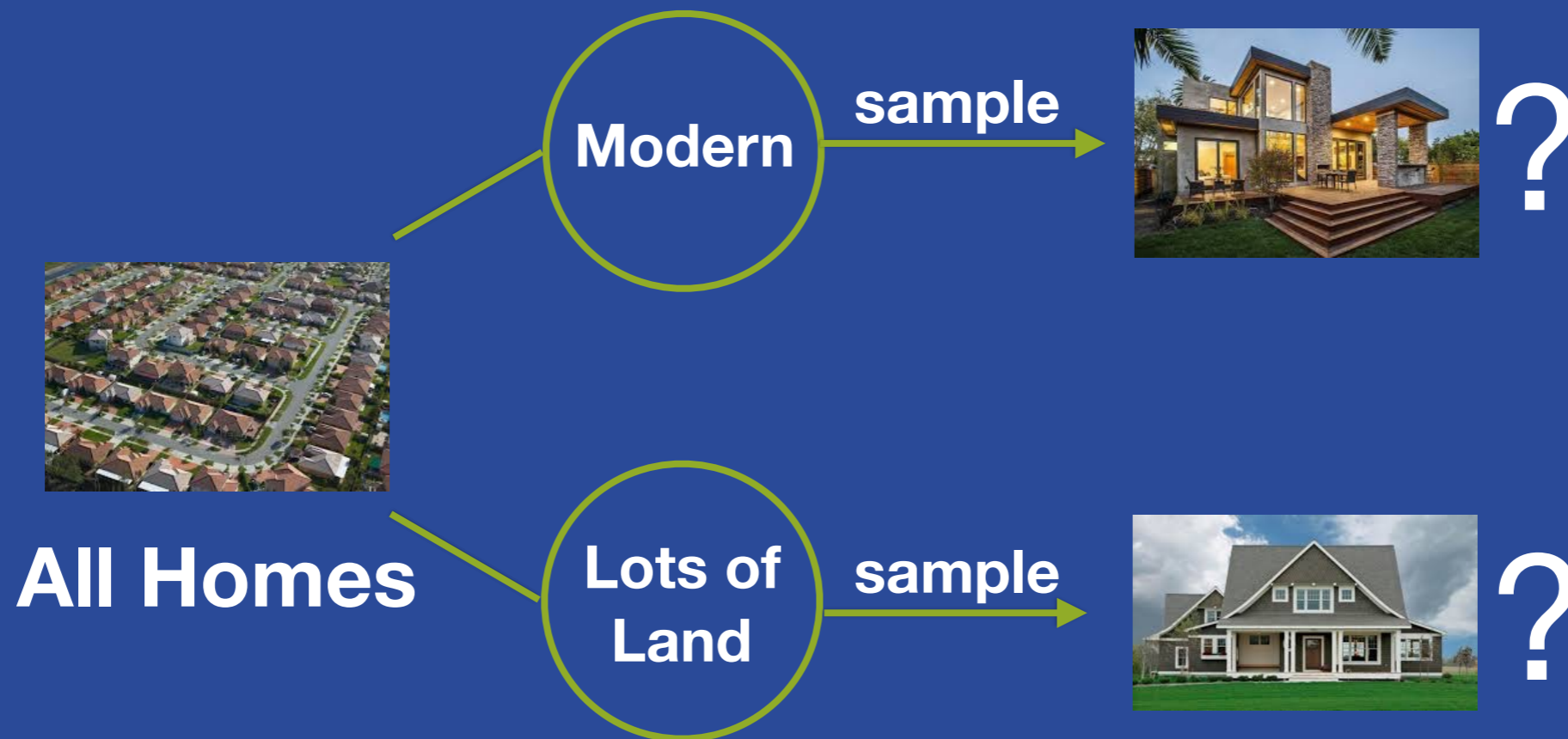
Recommender Problem #2

- How can we avoid showing essentially the same house over and over?



Recommender Problem #2

- How can we avoid showing essentially the same house over and over?



- Great! What if we don't know how to group them? Or how many groups?

- Since we don't know how many groups of homes there should be, we can use G-means **Clustering** to find the optimum number of groups of homes
- Our recommender will use these groups to create a better sampling for user preference
- We can try to understand the home **Clusters** using “model clusters” but the models will be difficult to interpret

Understanding Clusters

SQFT	PRICE	BEDS	BATHS
3,125	\$530,000	5	3
2,100	\$460,000	4	2
1,200	\$250,000	3	1.5
3,950	\$610,000	6	4



What if we could get rules like...

If **SQFT** \geq **3,125** THEN “**Cluster 1**”

- We use a **Batch Centroid** to add the **Cluster** assignment of each home as a feature to the **Dataset**
- We use **Association Discovery** to find “interesting” relationships between the features including the **Cluster** assignment

Recommender Problem #3



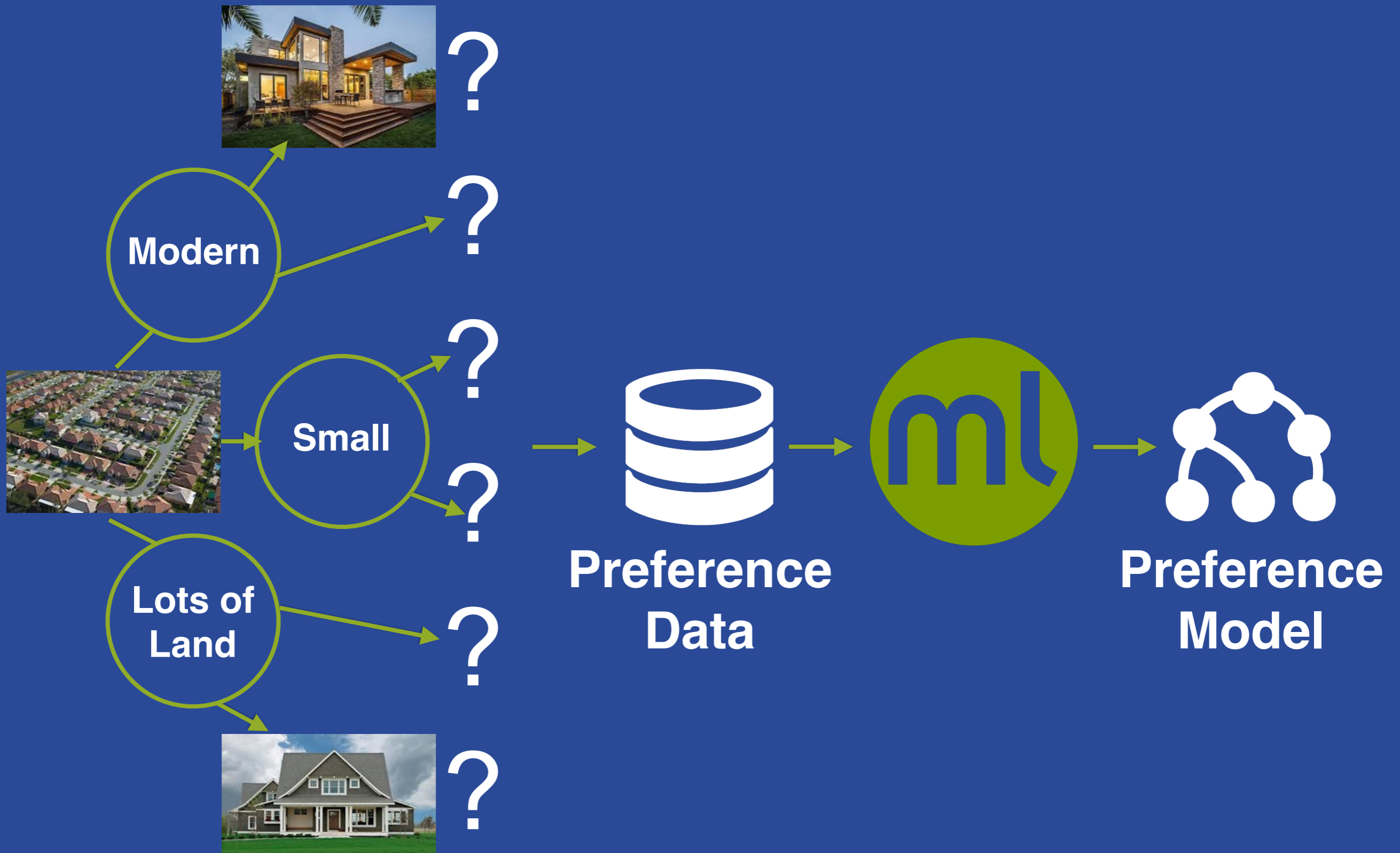
There is much more interesting information than just the number of BEDS, BATHS, etc.

Beautiful, well maintained 3 bedroom, 2.5 bath home on a desirable, quiet cul-de-sac. Bright & open floor plan, A/C, gas fireplace, gas or electric appliance hookups. HUGE landscaped, fenced, west facing backyard with sprinkler systems. Raised garden beds, large patio, water feature. Easy access to Corvallis & Albany. This home is a MUST SEE!

- Unfortunately, these "remarks" are not available in the Redfin download
- Adding them to our dataset requires crawling the website
- Like most ML projects, preparing the data is 80% of the difficulty (fortunately I already did it!)

- We have an extended home dataset with the syndicated remarks text field
- We can use **Topic Modeling** to create a deeper thematic understanding of the remarks
 - Perhaps homes that are "in-town" or "out-of-town"
- We can extend the **Dataset** with fields that represent for each home how related they are to each of these topics
- This will allow our **Clustering** to group homes by a deeper meaning than just **BEDS, BATHS**, etc

Recommender Idea



House Recommender

bigml[®]