Regression:
Predicting House Prices

Utkarsh Kulshrestha

Predicting house prices

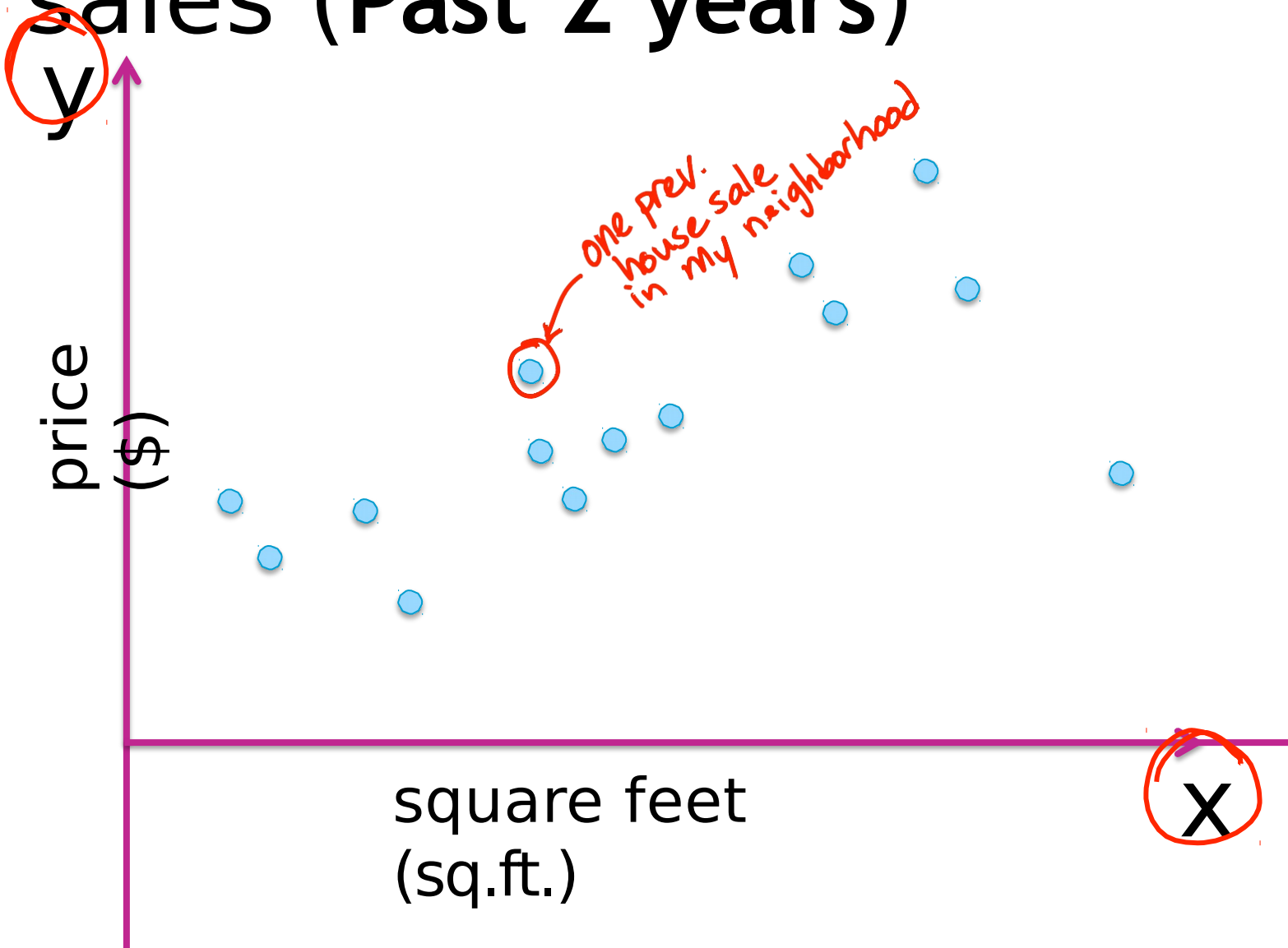
How much is my house worth?

How much is my house worth?

Look at recent sales in my neighborhood

- How much did they sell for?

Plot recent house sales (**Past 2 years**)

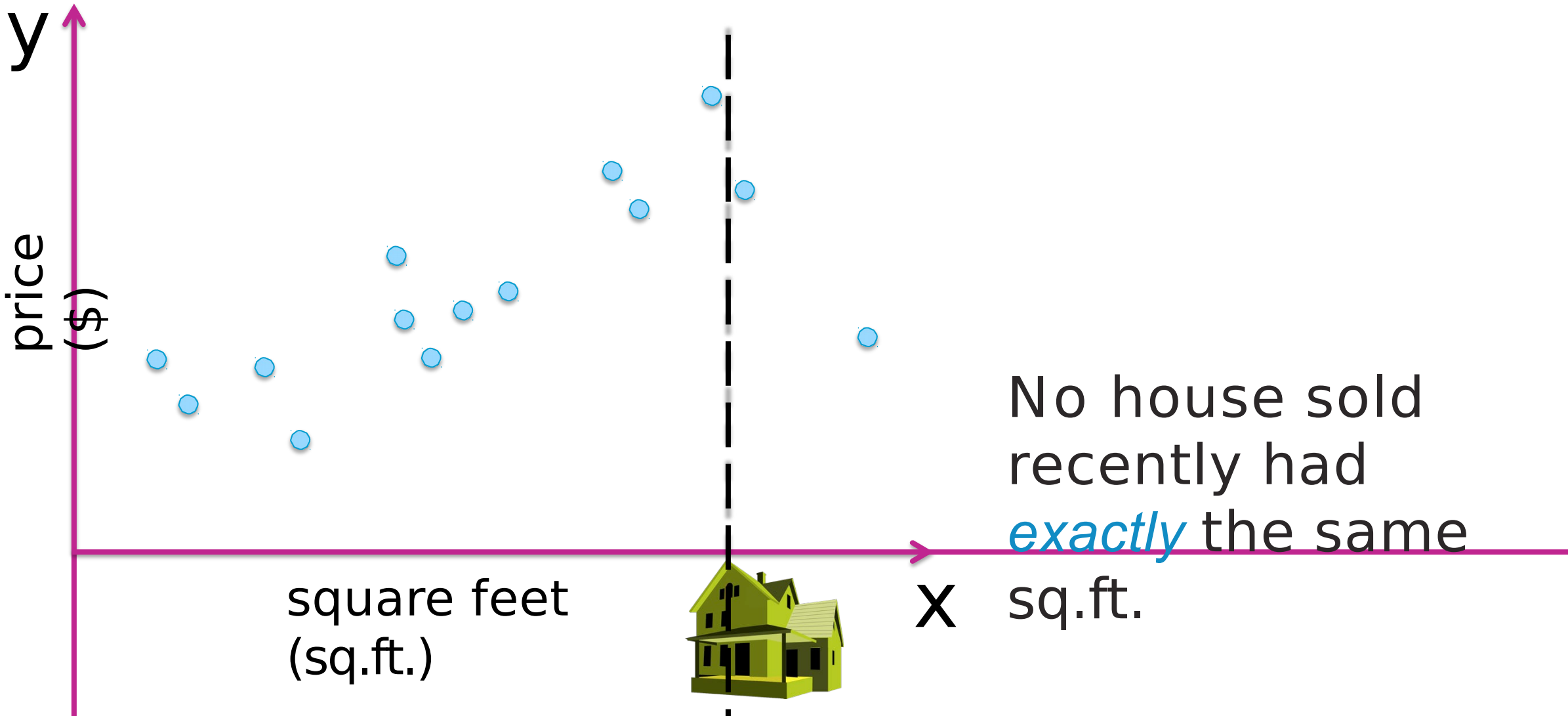


Terminology:

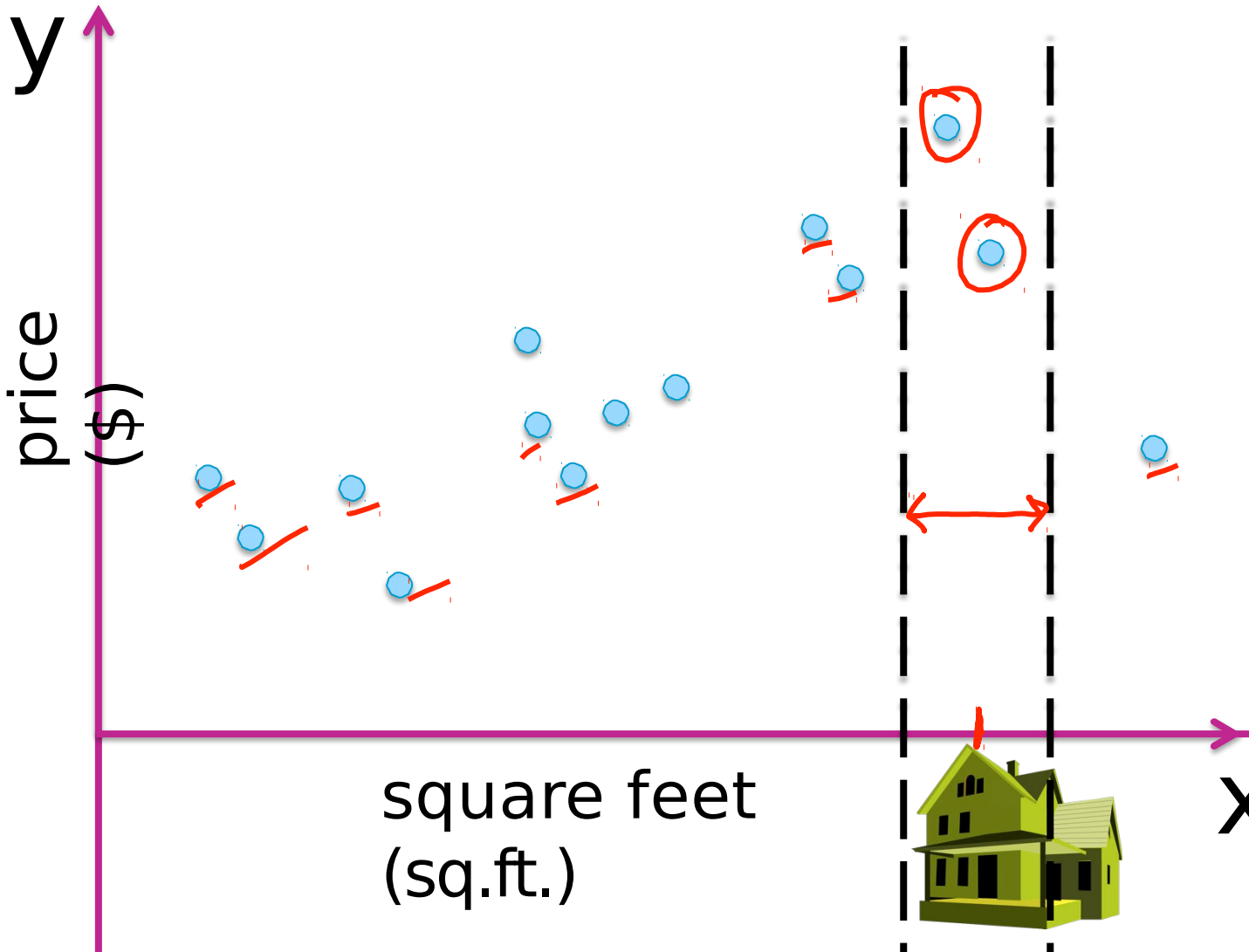
x – feature, covariate, or predictor

y – observation or response

Predict your house by similar houses



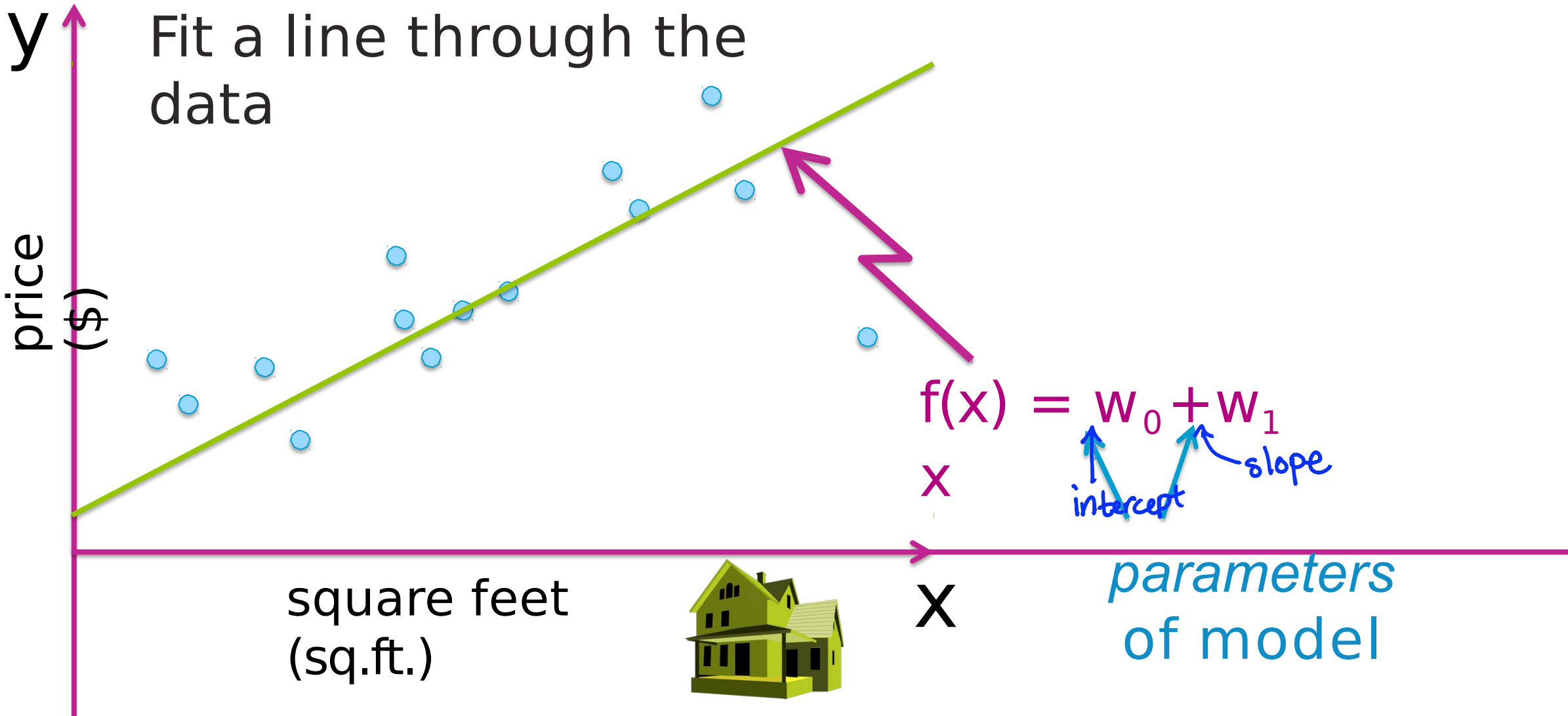
Predict your house by similar houses



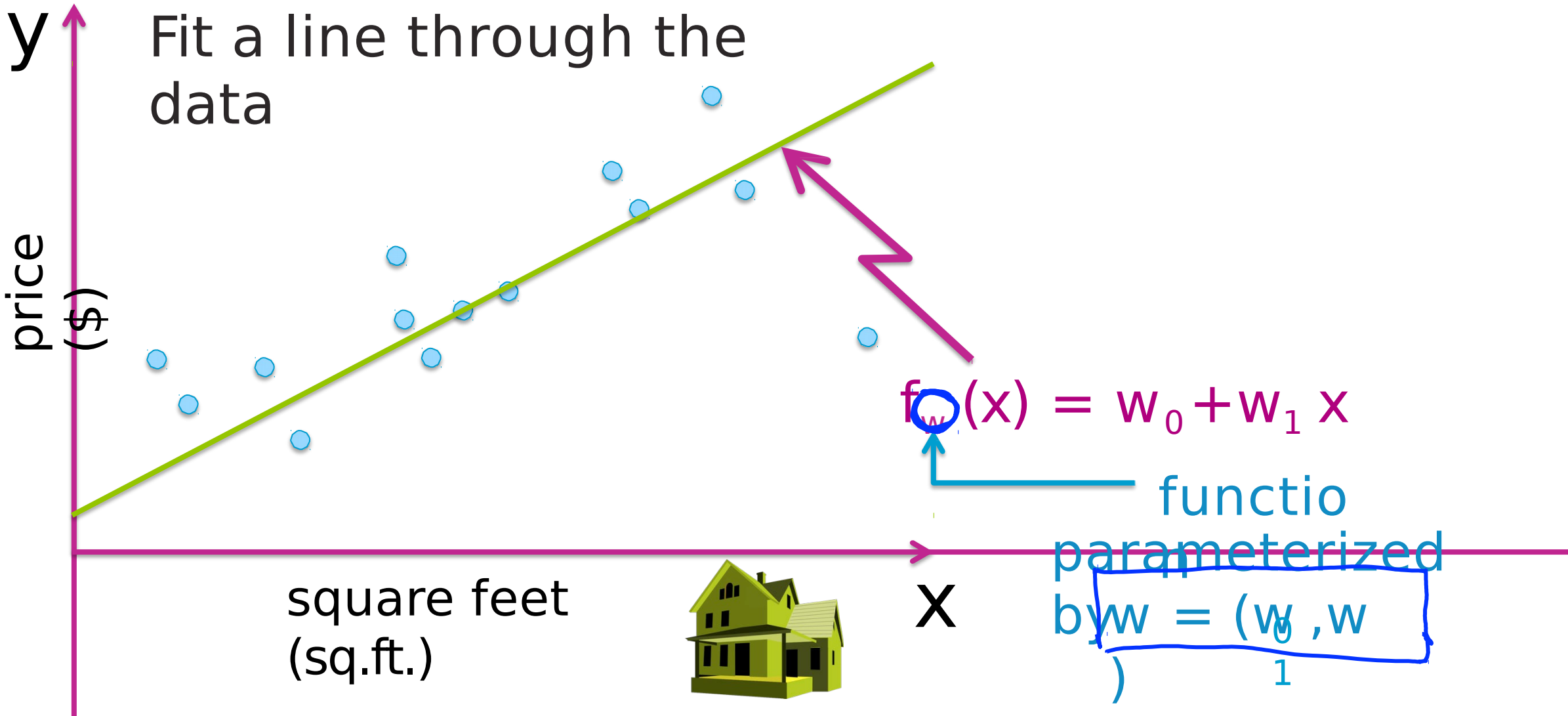
- Look at average price in range
- Still only 2 houses!
- Throwing out

Linear regression

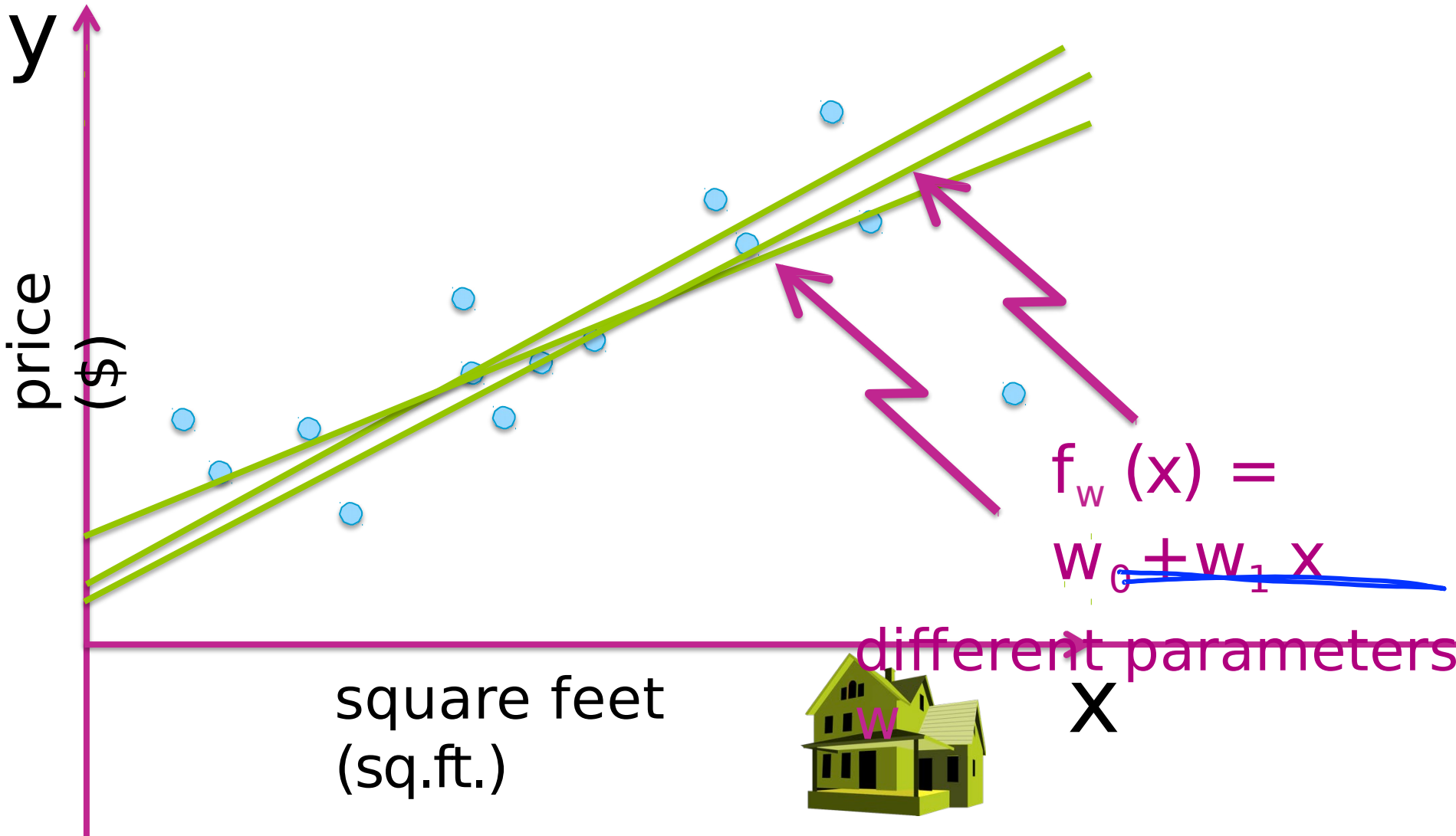
Use a linear regression model



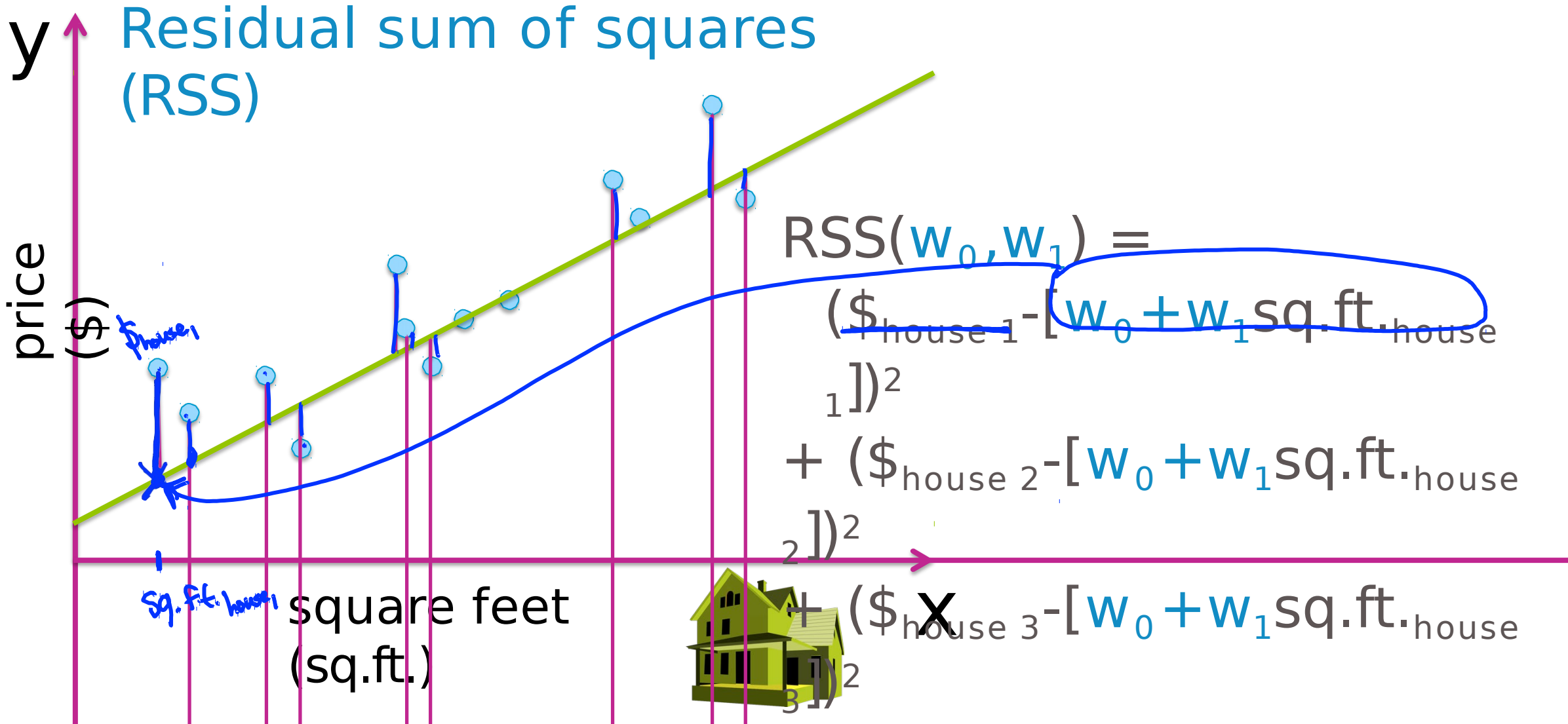
Use a linear regression model



Which line?

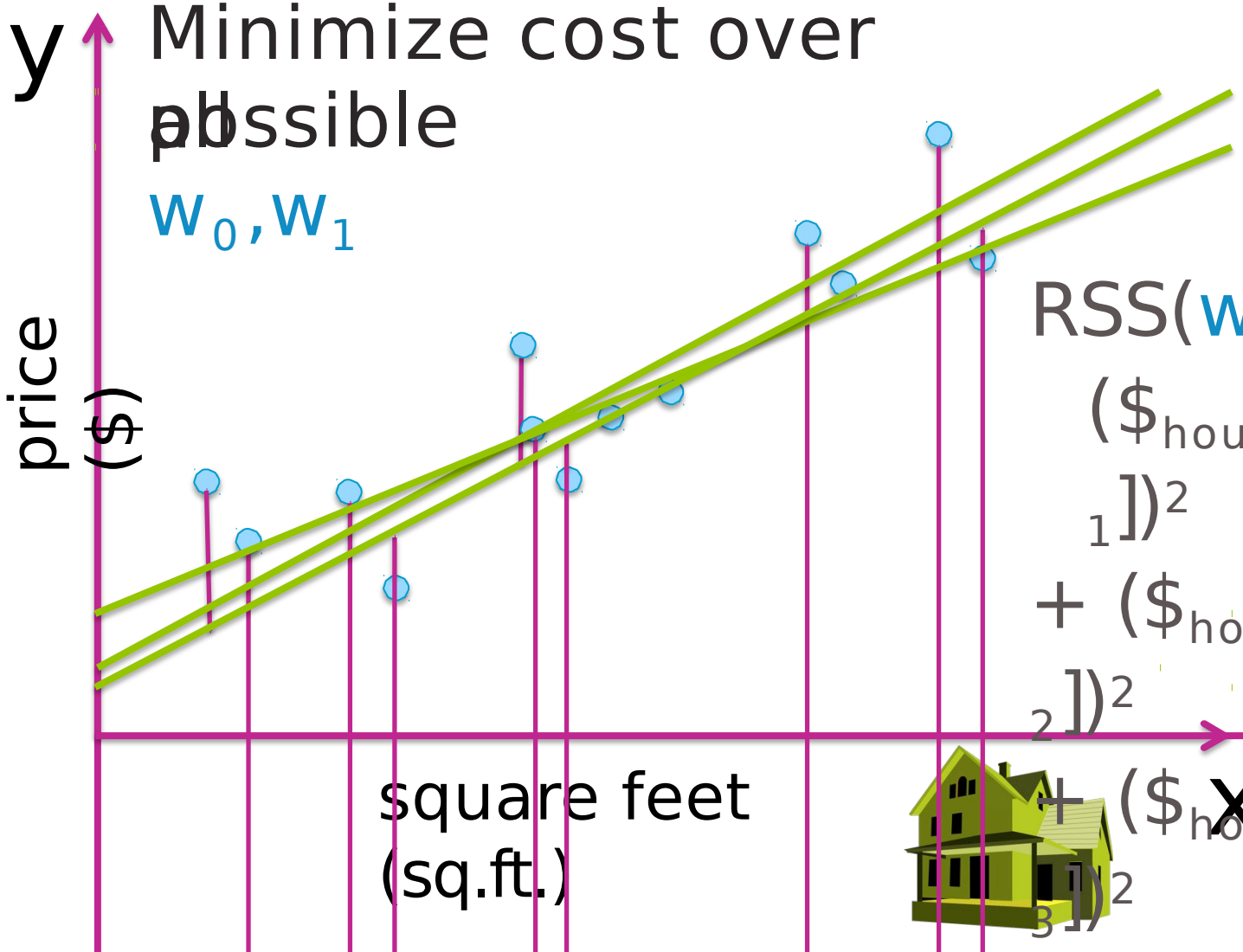


“Cost” of using a given line

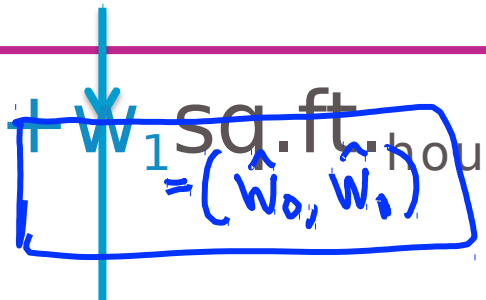


Find “best” line

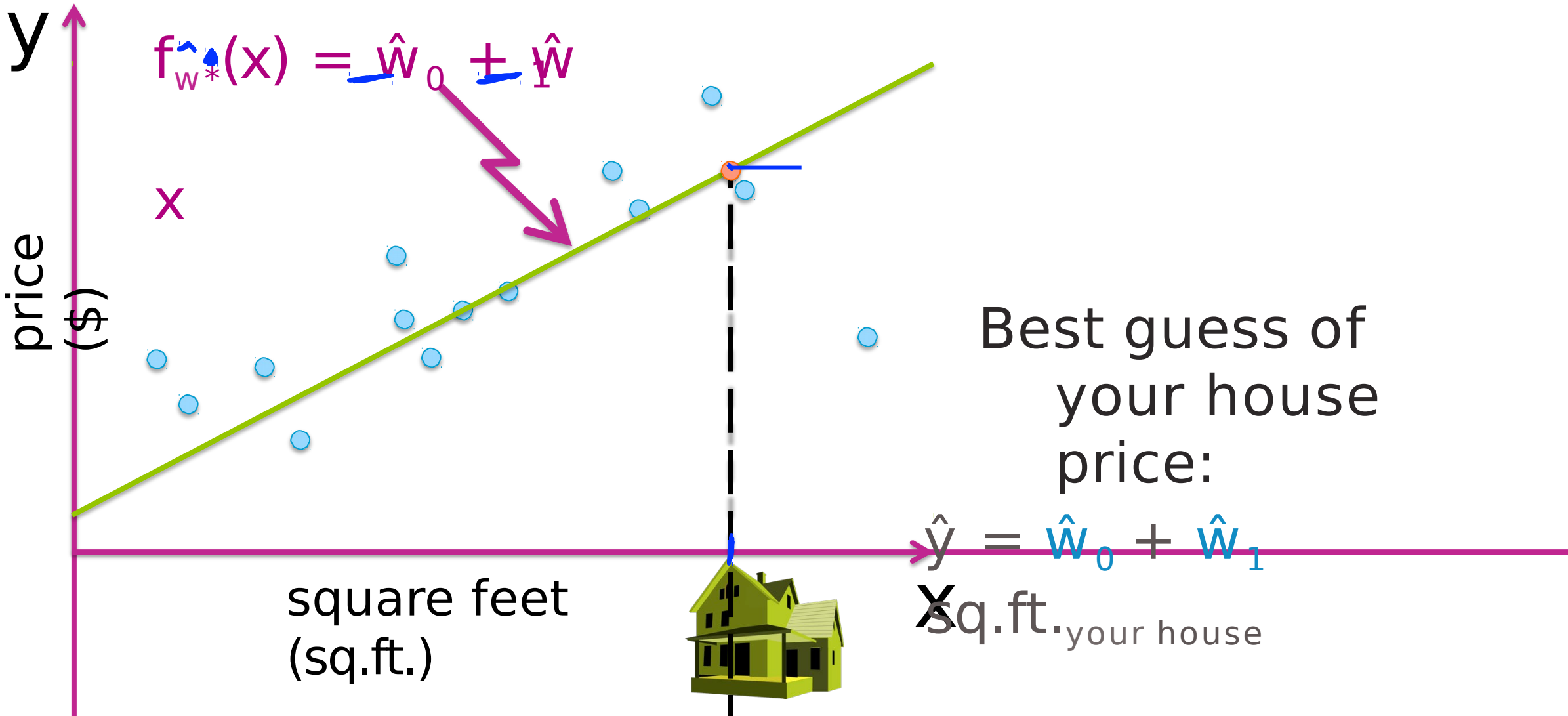
Minimize cost over possible w_0, w_1



$$RSS(w_0,w_1) = (\$_{house\ 1} - [w_0 + w_1 sq.ft._{house\ 1}])^2 + (\$_{house\ 2} - [w_0 + w_1 sq.ft._{house\ 2}])^2 + (\$_{house\ 3} - [w_0 + w_1 sq.ft._{house\ 3}])^2$$

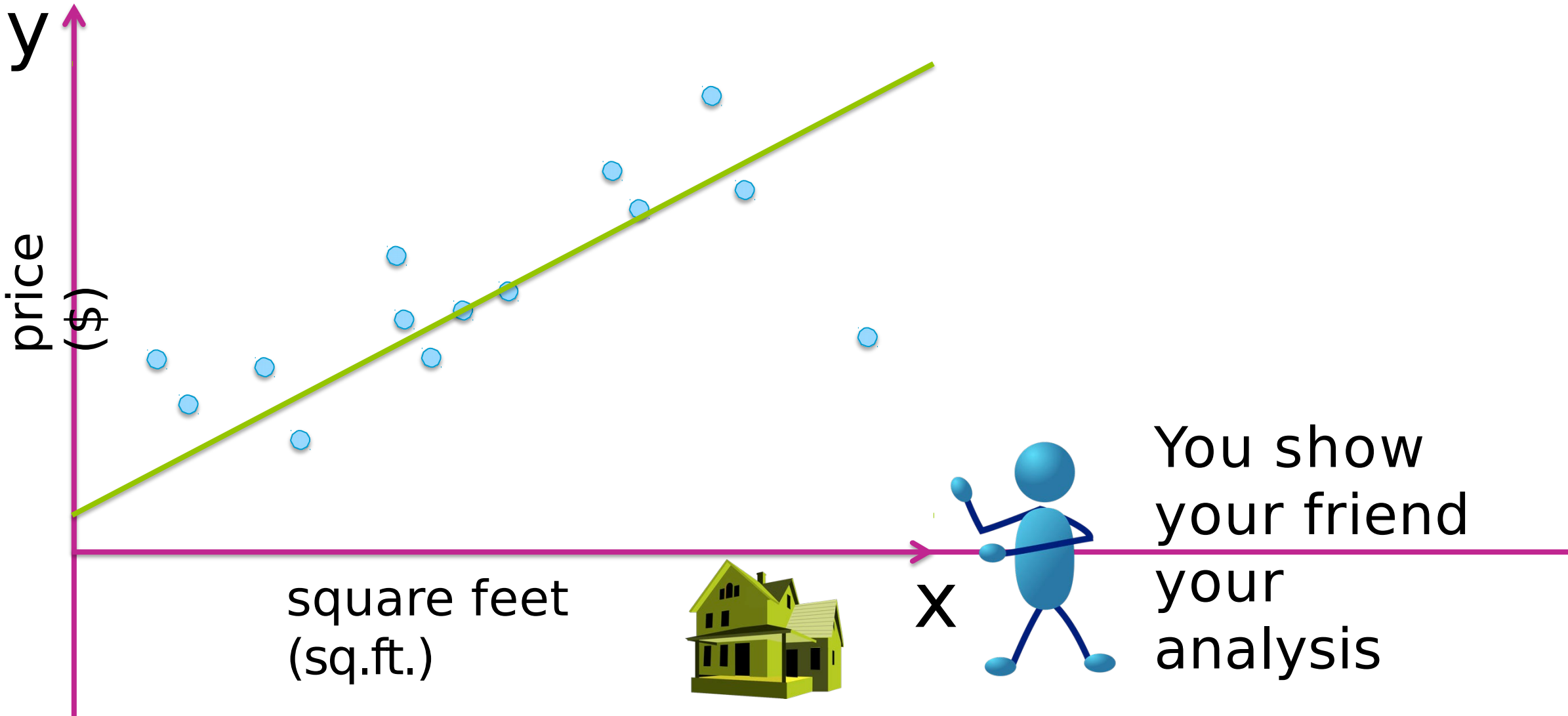


Predicting your house price

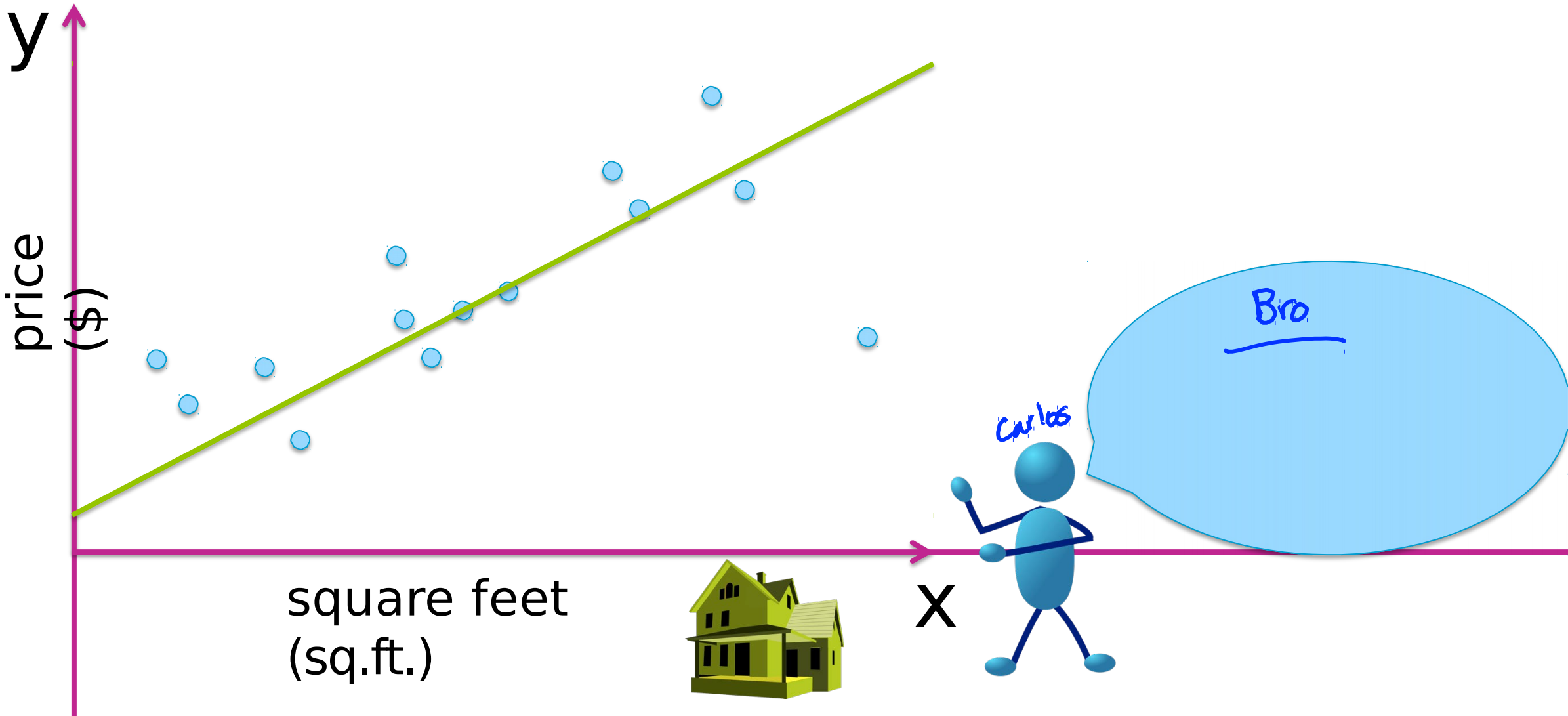


Adding higher order effects

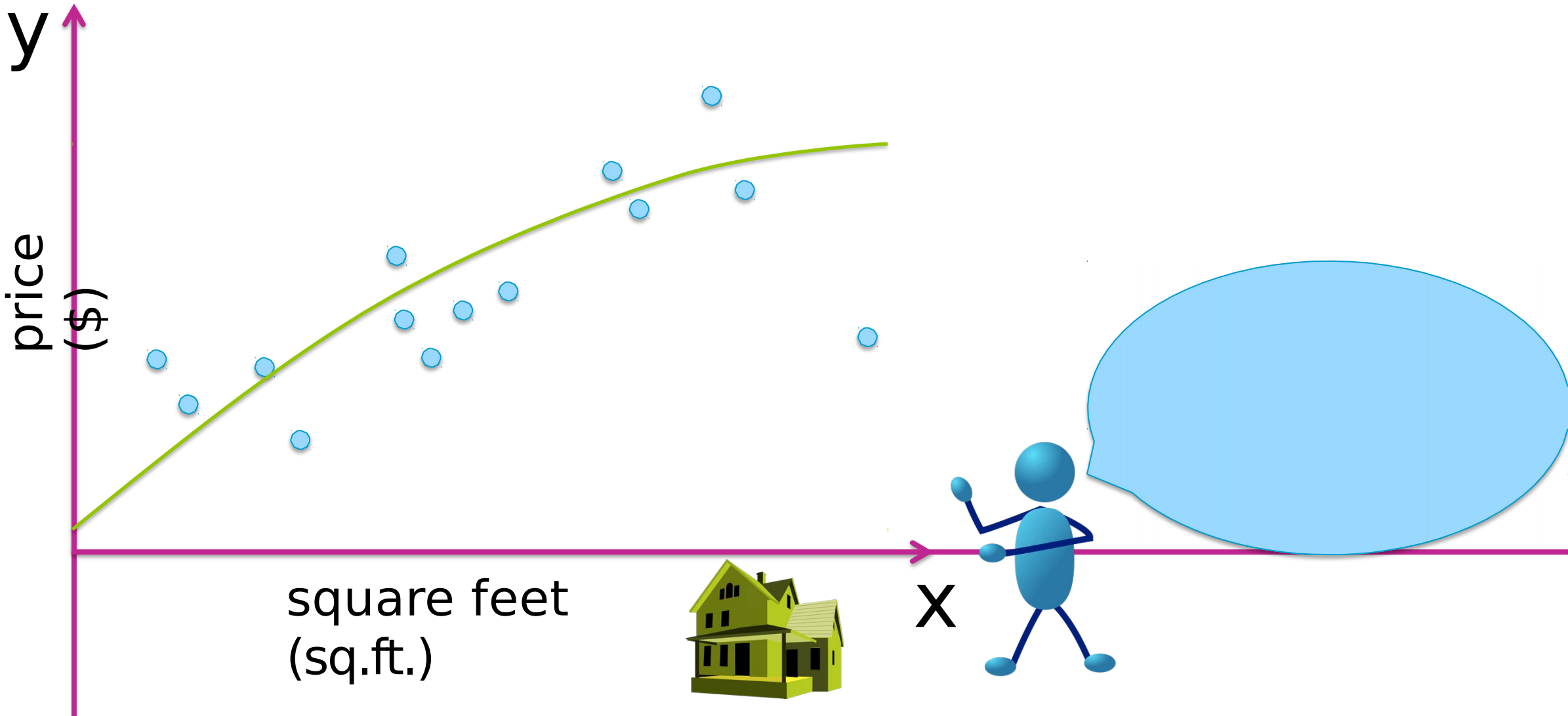
Fit data with a line or … ?



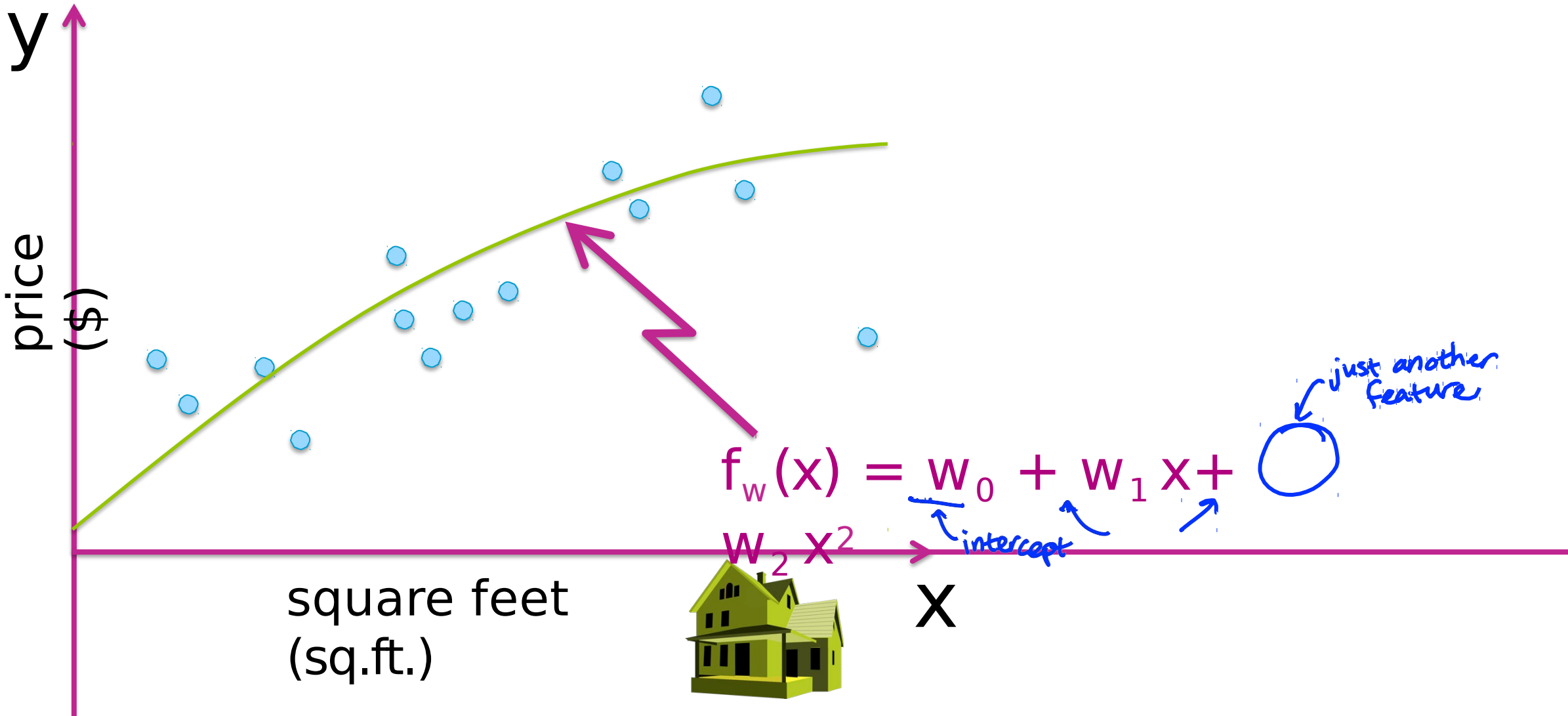
Fit data with a line or … ?



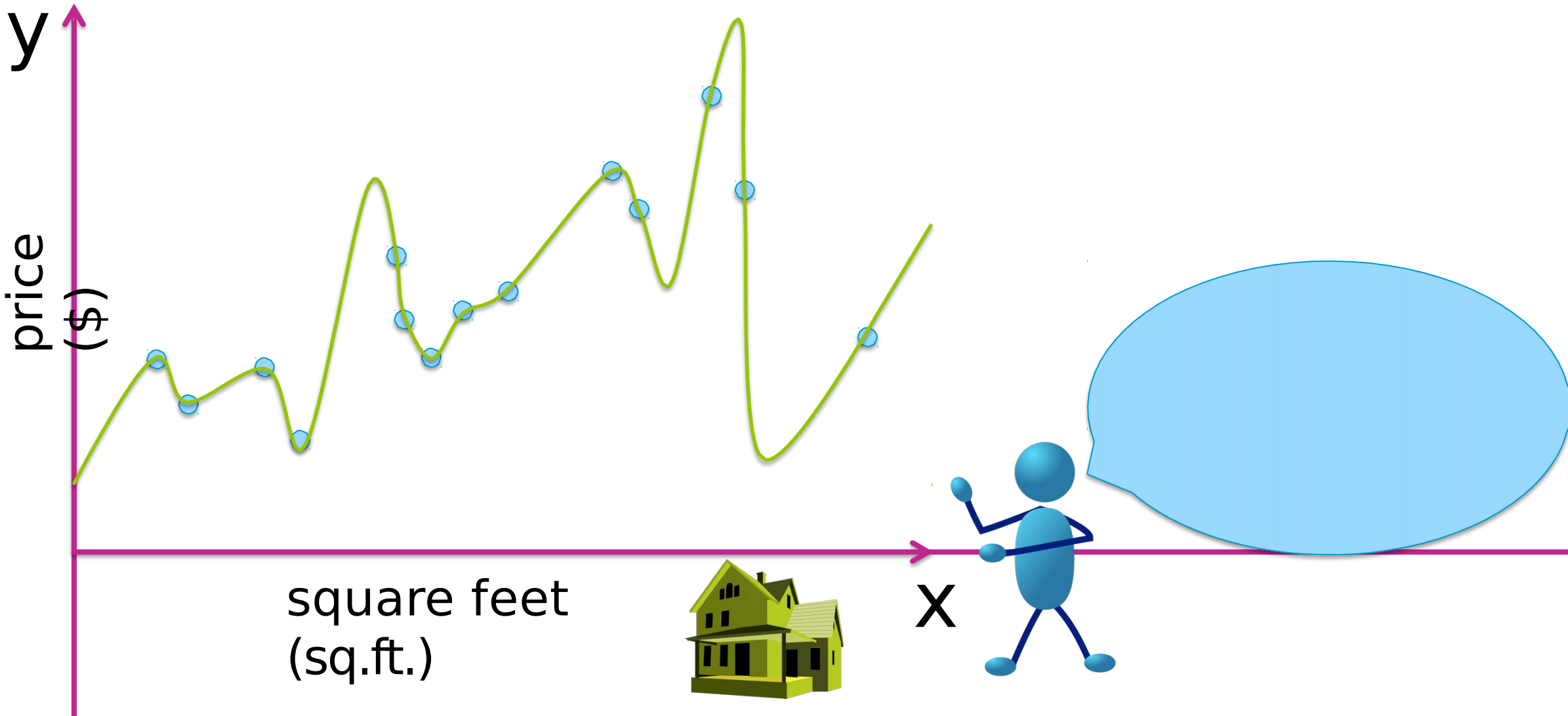
What about a quadratic function?



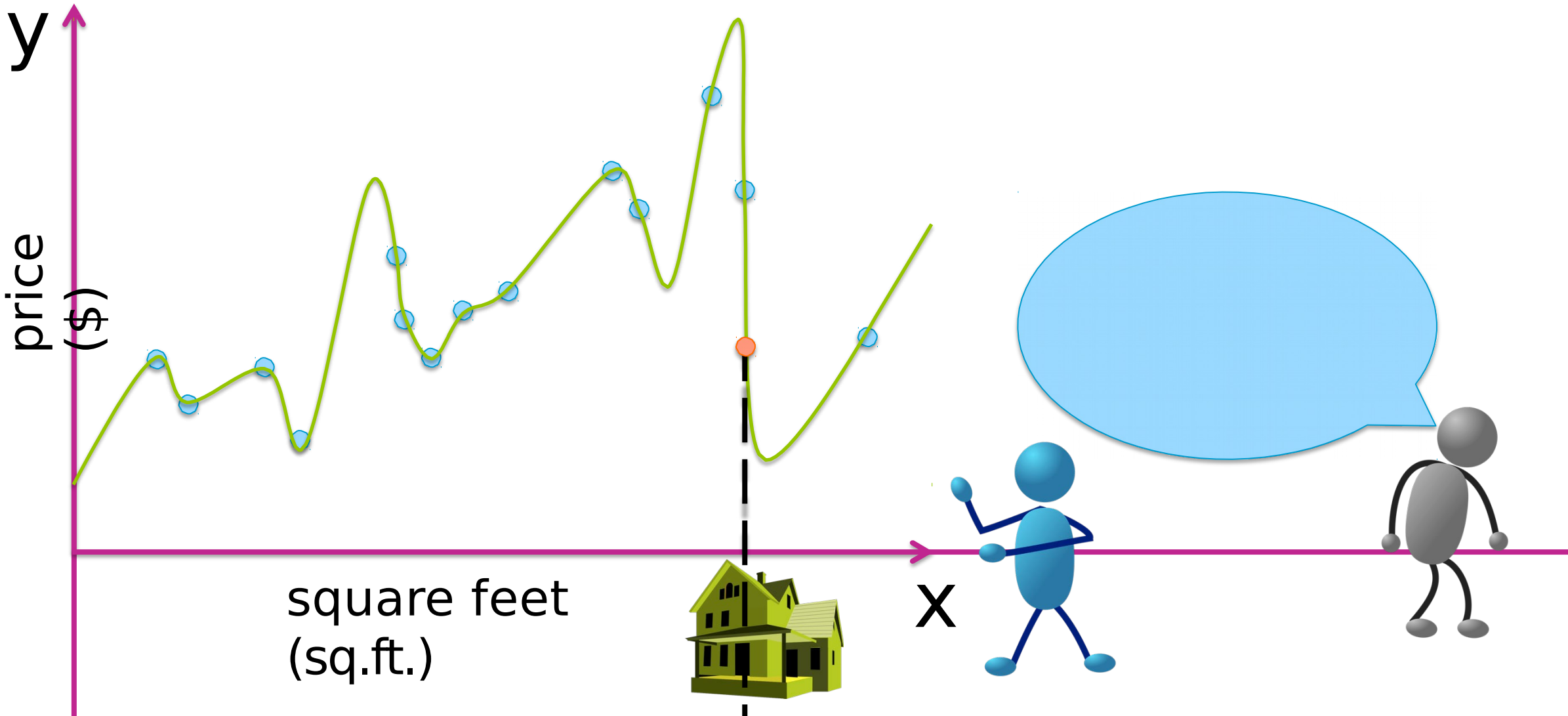
What about a quadratic function?



Even higher order polynomial

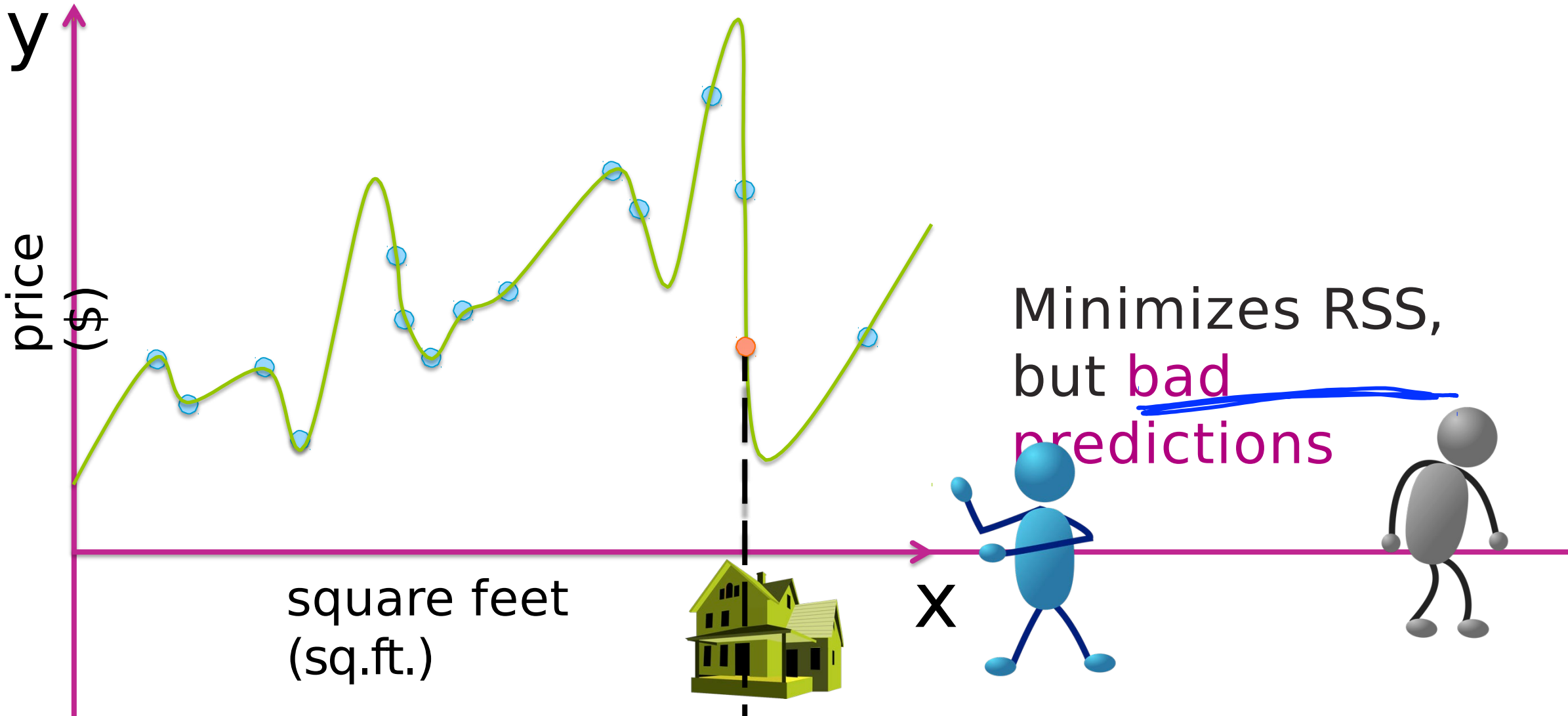


Do you believe this fit?

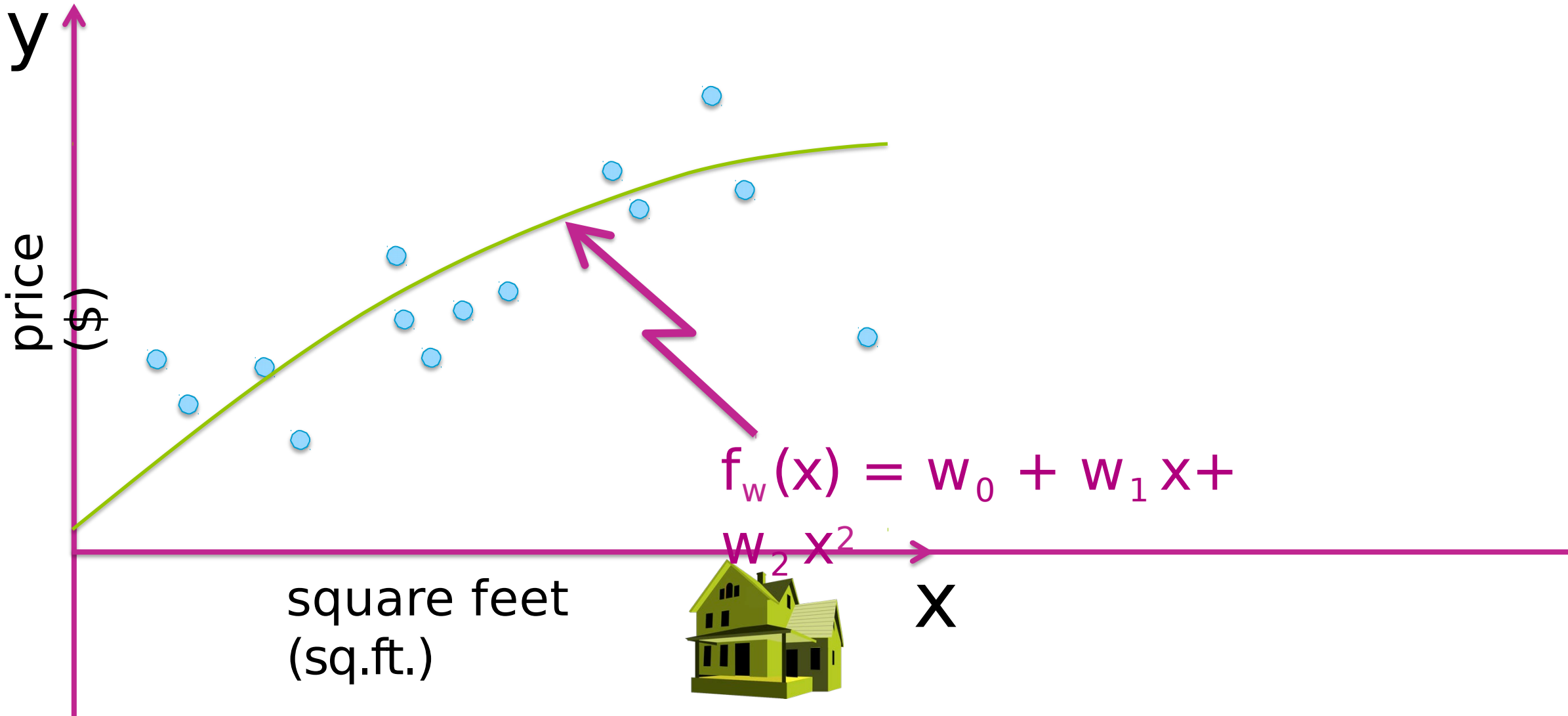


Evaluating overfitting via training/test split

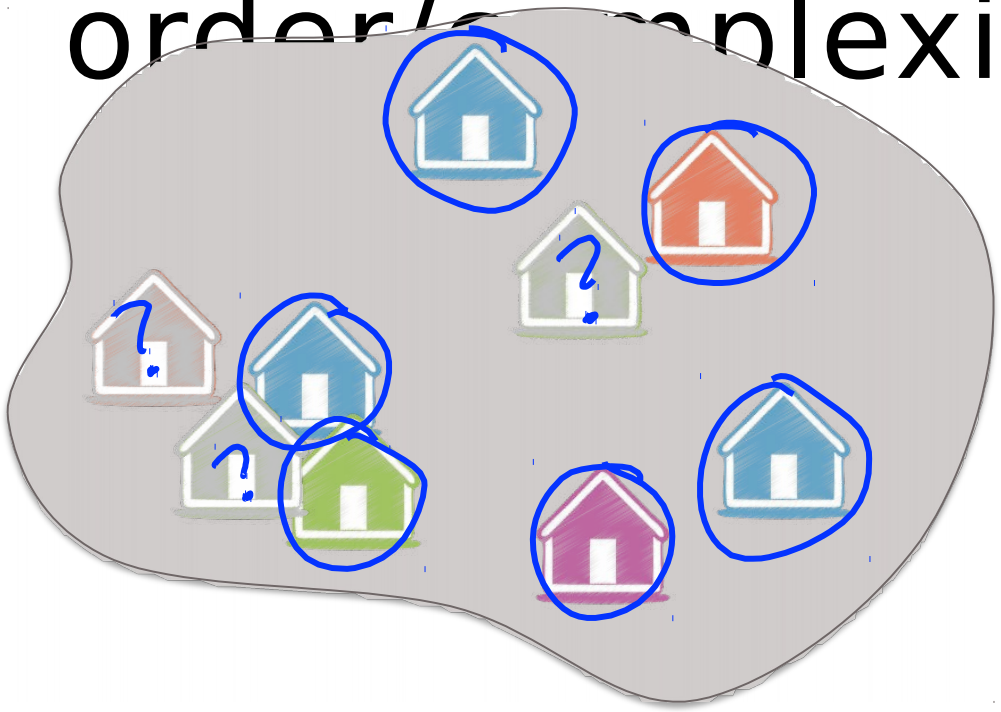
Do you believe this fit?



What about a quadratic function?



How to choose model order/complexity



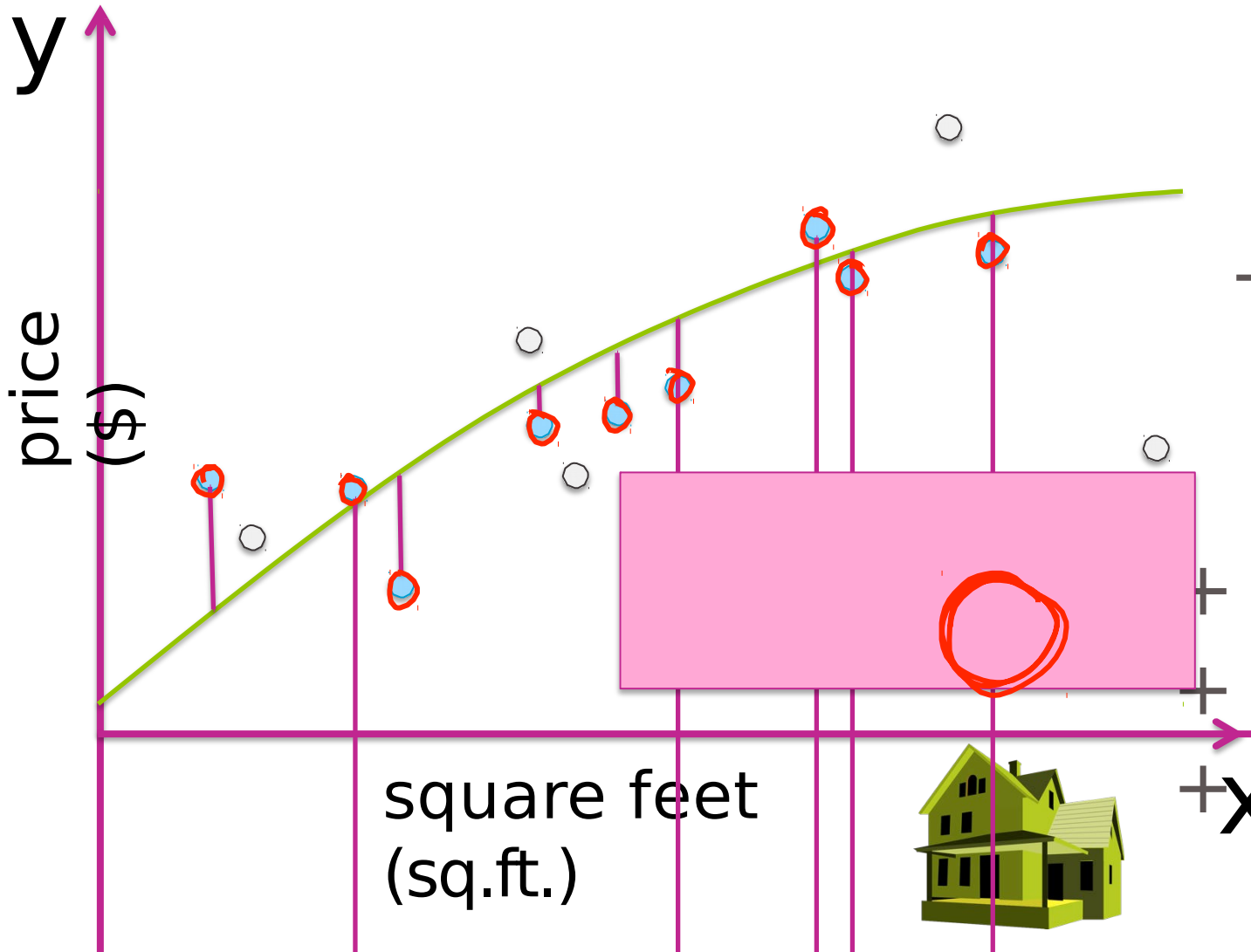
- Want good predictions, but can't observe future
- Simulate predictions

1. Remove some houses
2. Fit model on remaining

Training/test split

Terminology:
- training set
- test set

Training error



Training error (w) =

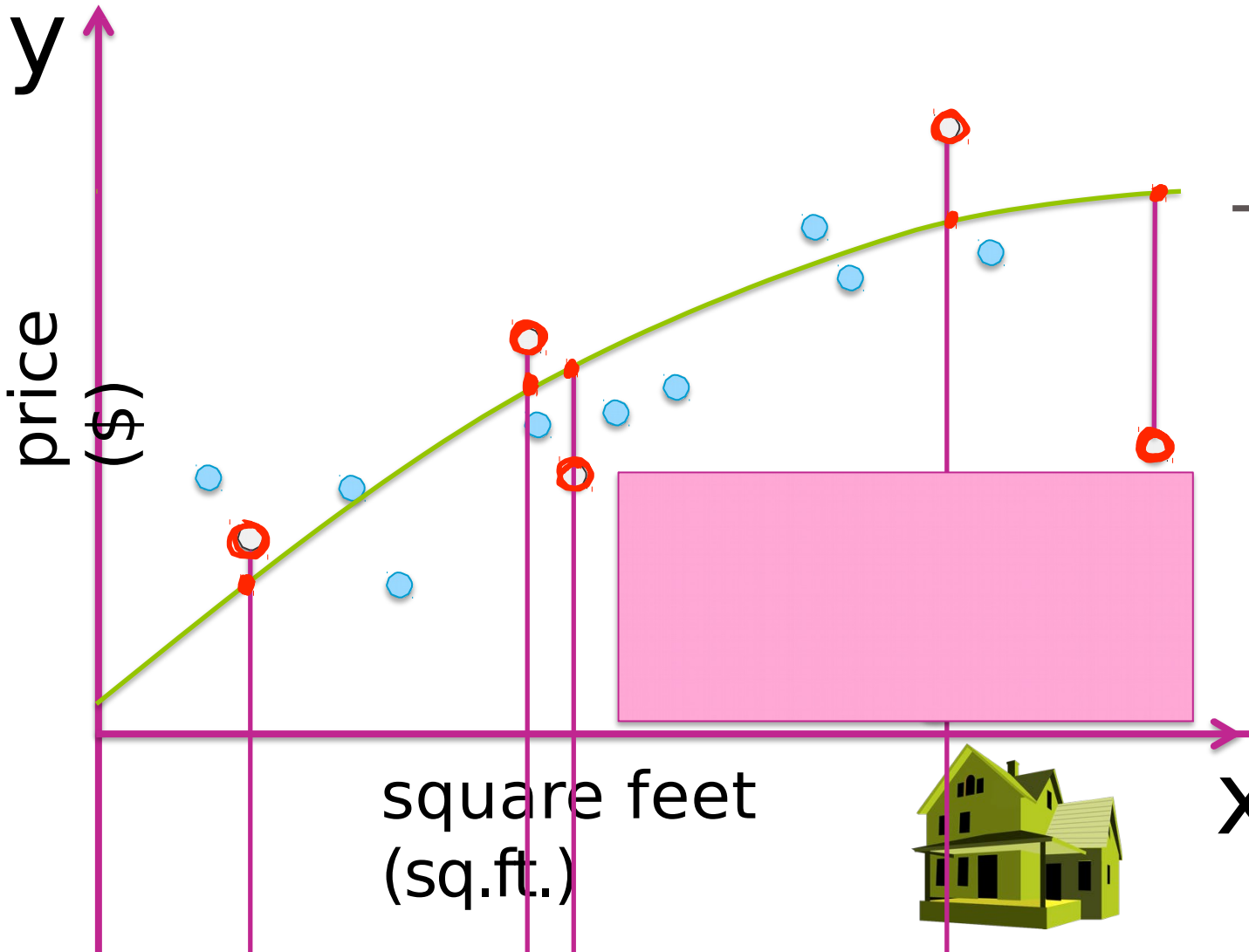
$(\$_{train\ 1} - f_w(sq.ft._{train\ 1}))^2$

$+ (\$_{train\ 2} - f_w(sq.ft._{train\ 2}))^2$

$+ (\$_{train\ 3} - f_w(sq.ft._{train\ 3}))^2$

+ ... [include all training houses]

Test error



Test error ($\hat{w}$) =

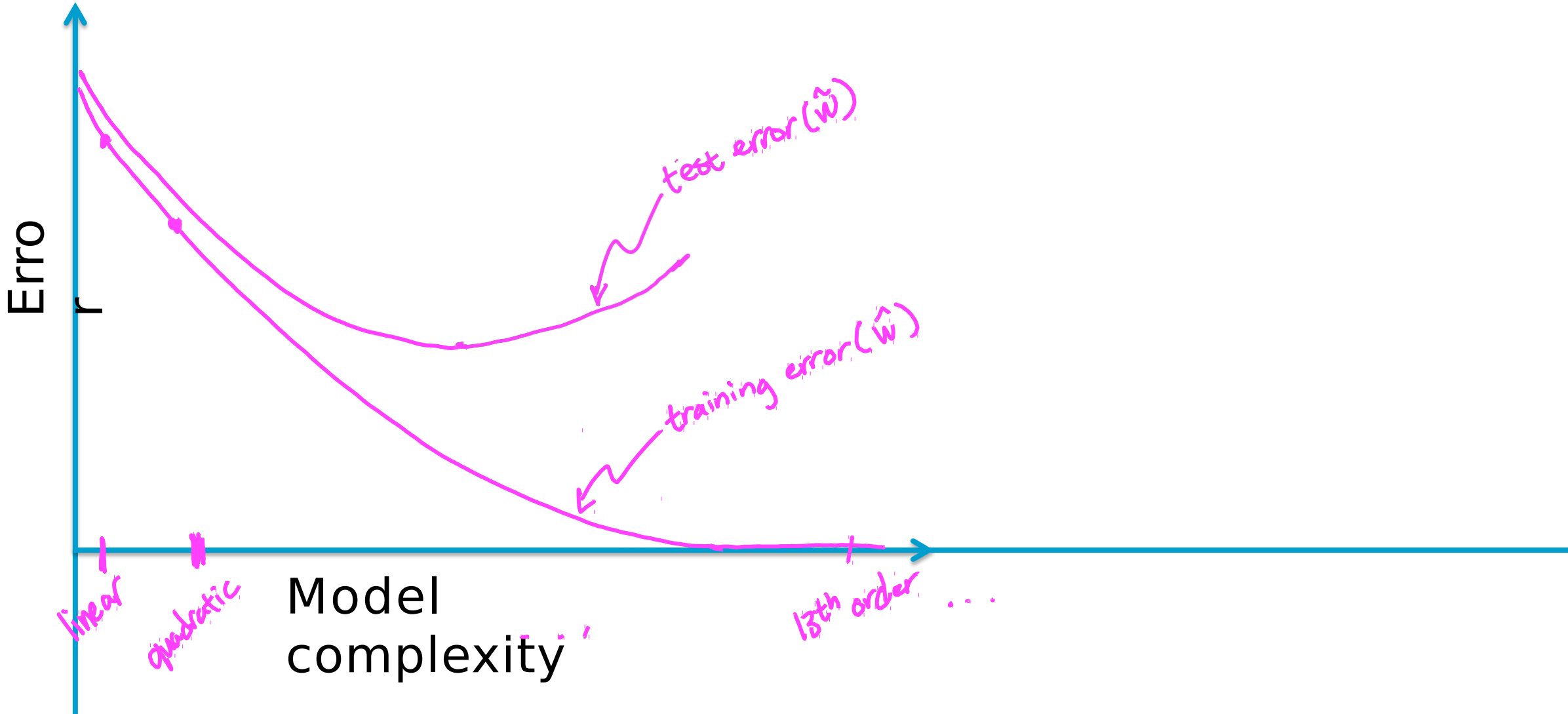
$(\$_{test\ 1} - f_{\hat{w}}(sq.ft._{test\ 1}))^2$

$+ (\$_{test\ 2} - f_{\hat{w}}(sq.ft._{test\ 2}))^2$

$+ (\$_{test\ 3} - f_{\hat{w}}(sq.ft._{test\ 3}))^2$

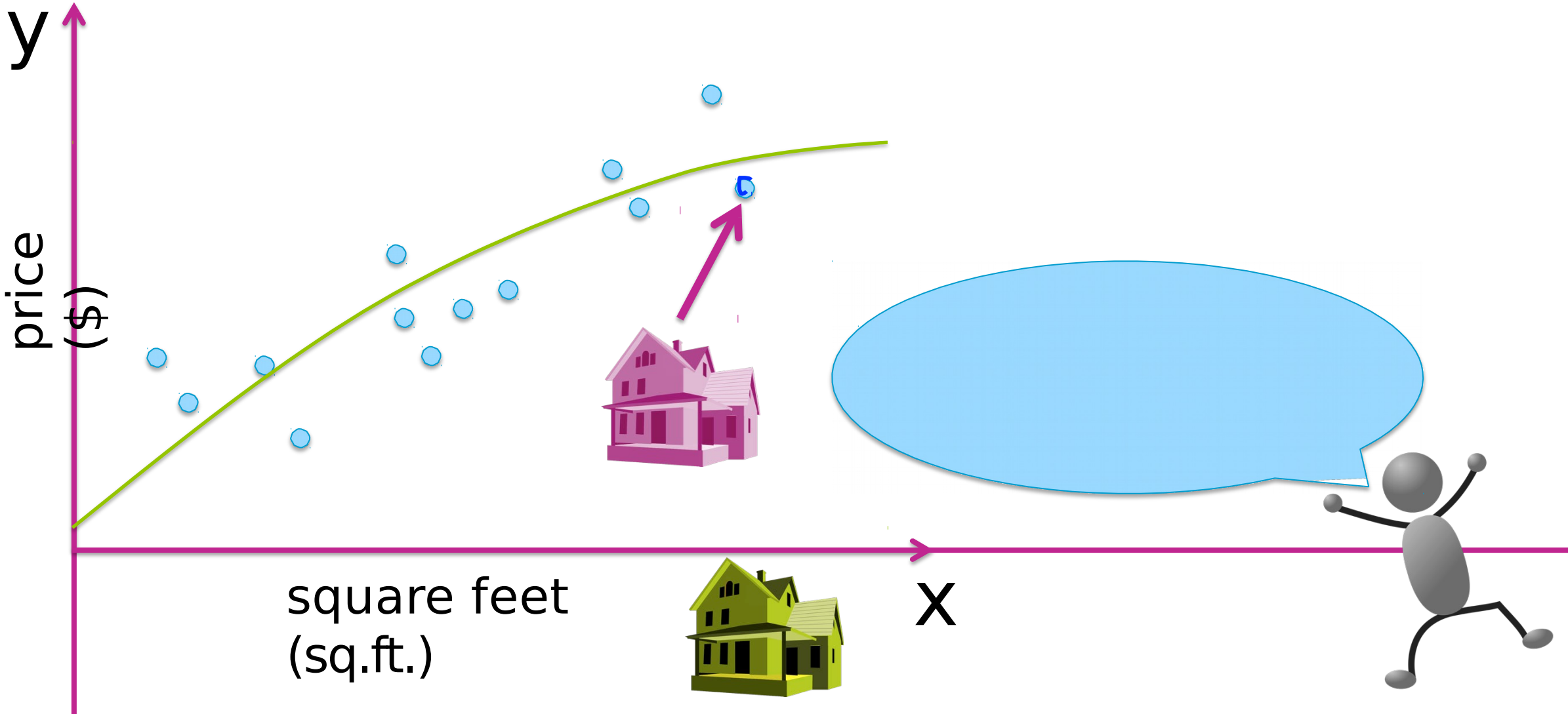
+ ... [include all test houses]

Training/Test Curves

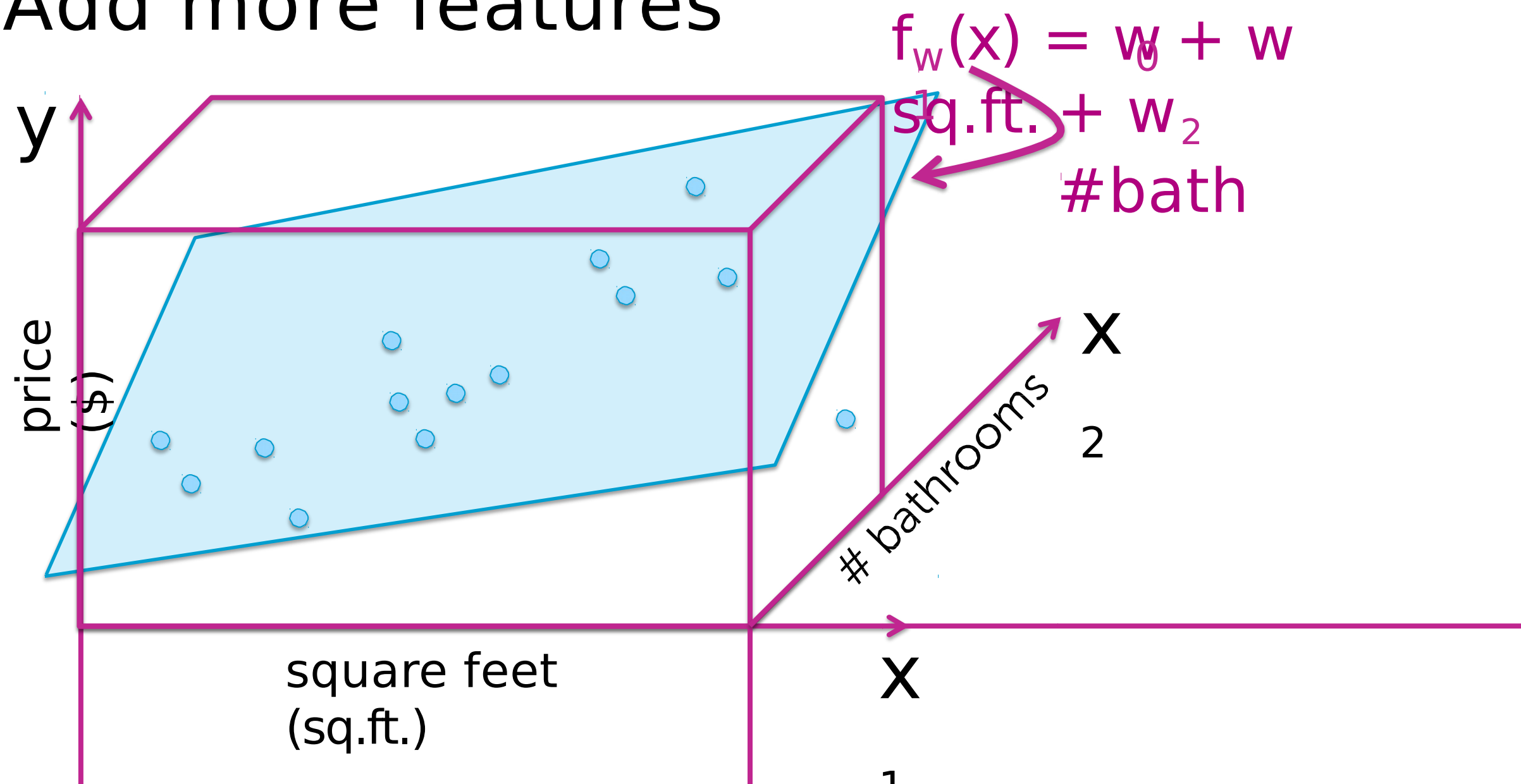


Adding other features

Predictions just based on house size



Add more features



How many features to use?

- Possible choices:
  - Square feet
  - # bathrooms
  - # bedrooms
  - Lot size
  - Year built
  - …
- See Regression Course!