Regression:

Predicting House Prices

Emily Fox & Carlos Guestrin

Machine Learning Specialization

University of Washington

Predicting house prices

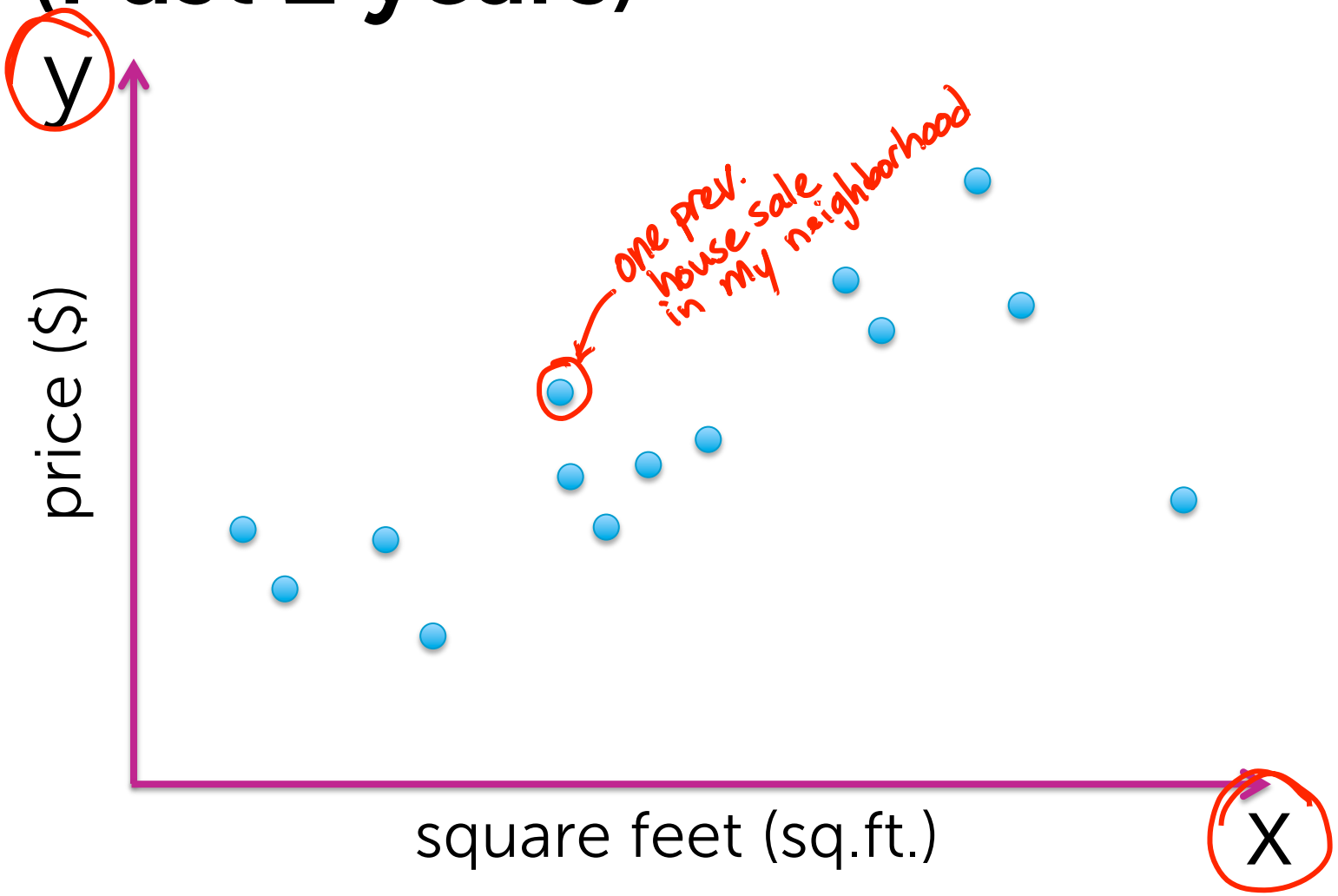
How much is my house worth?

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)

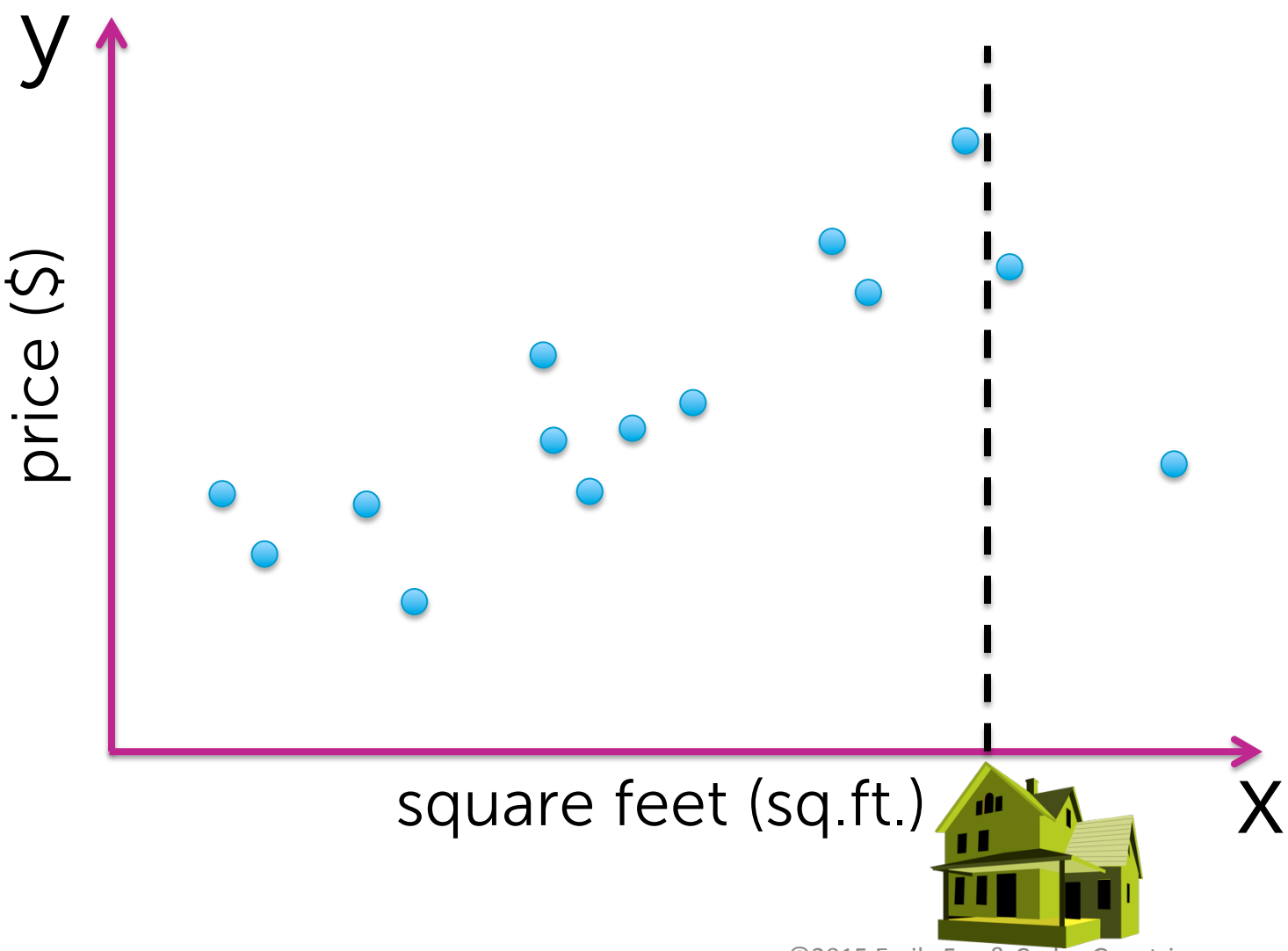


Terminology:

x – feature, covariate, or predictor

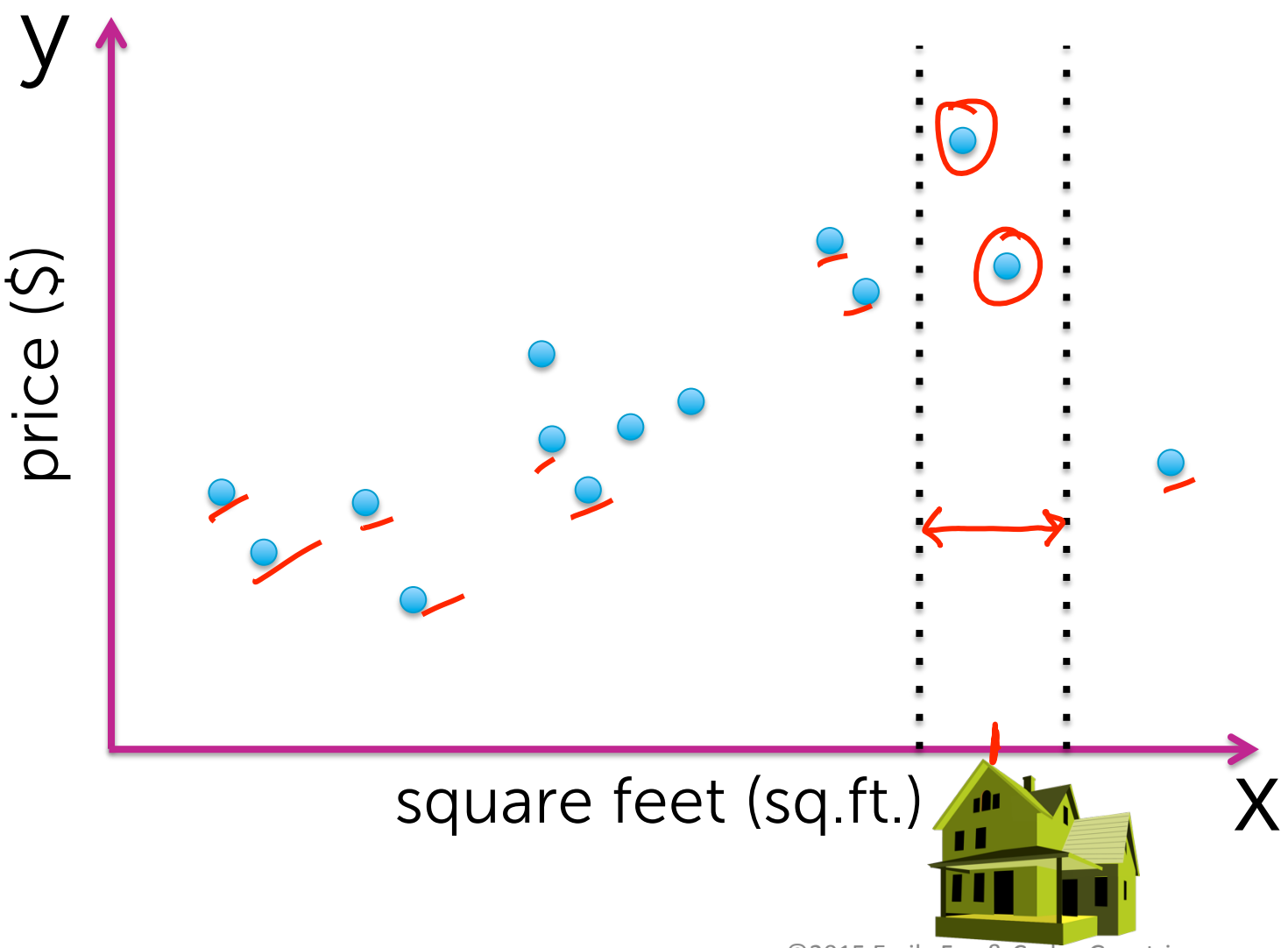
y – observation or response

Predict your house by similar houses



No house sold recently had *exactly* the same sq.ft.

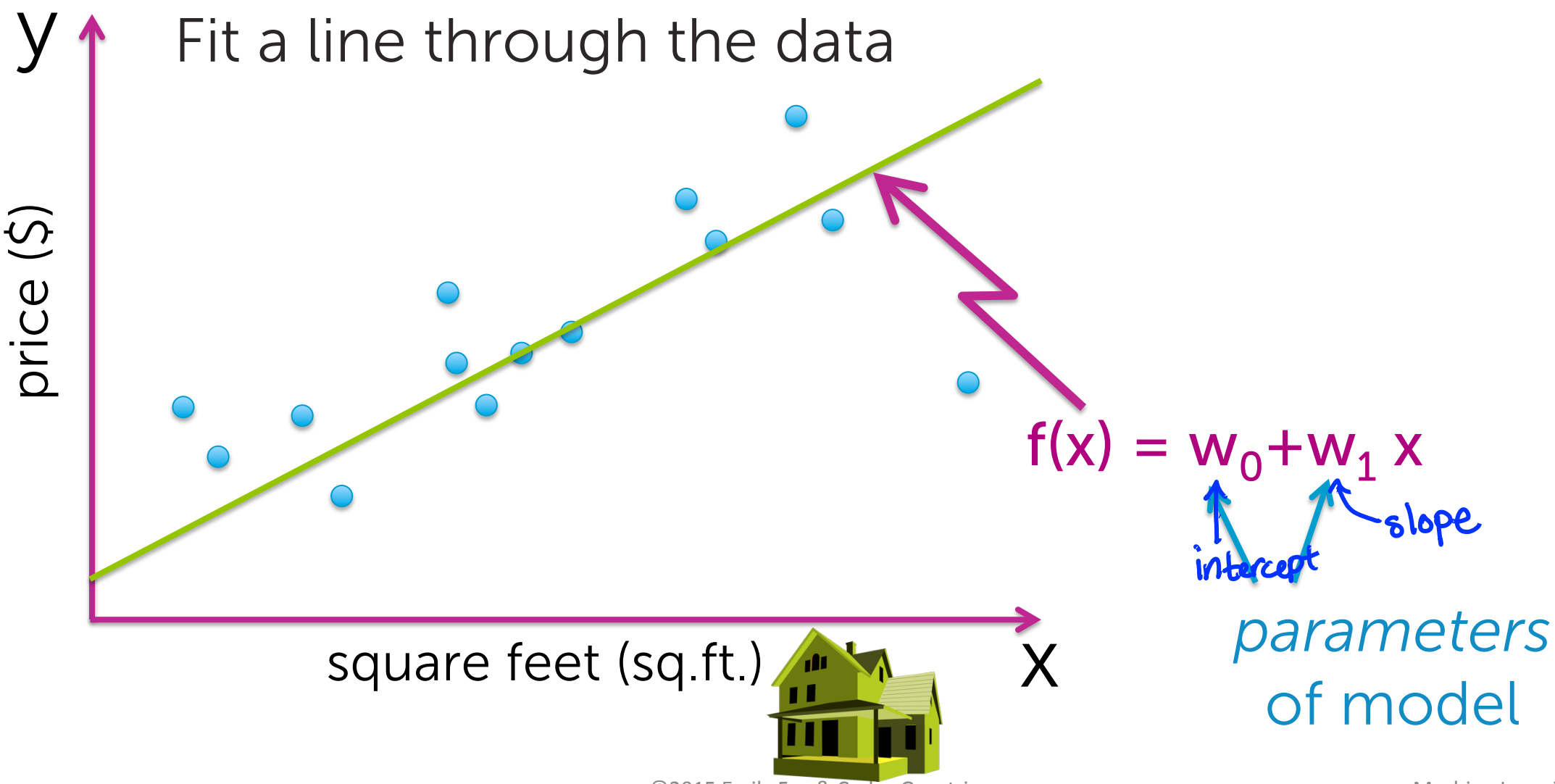
Predict your house by similar houses



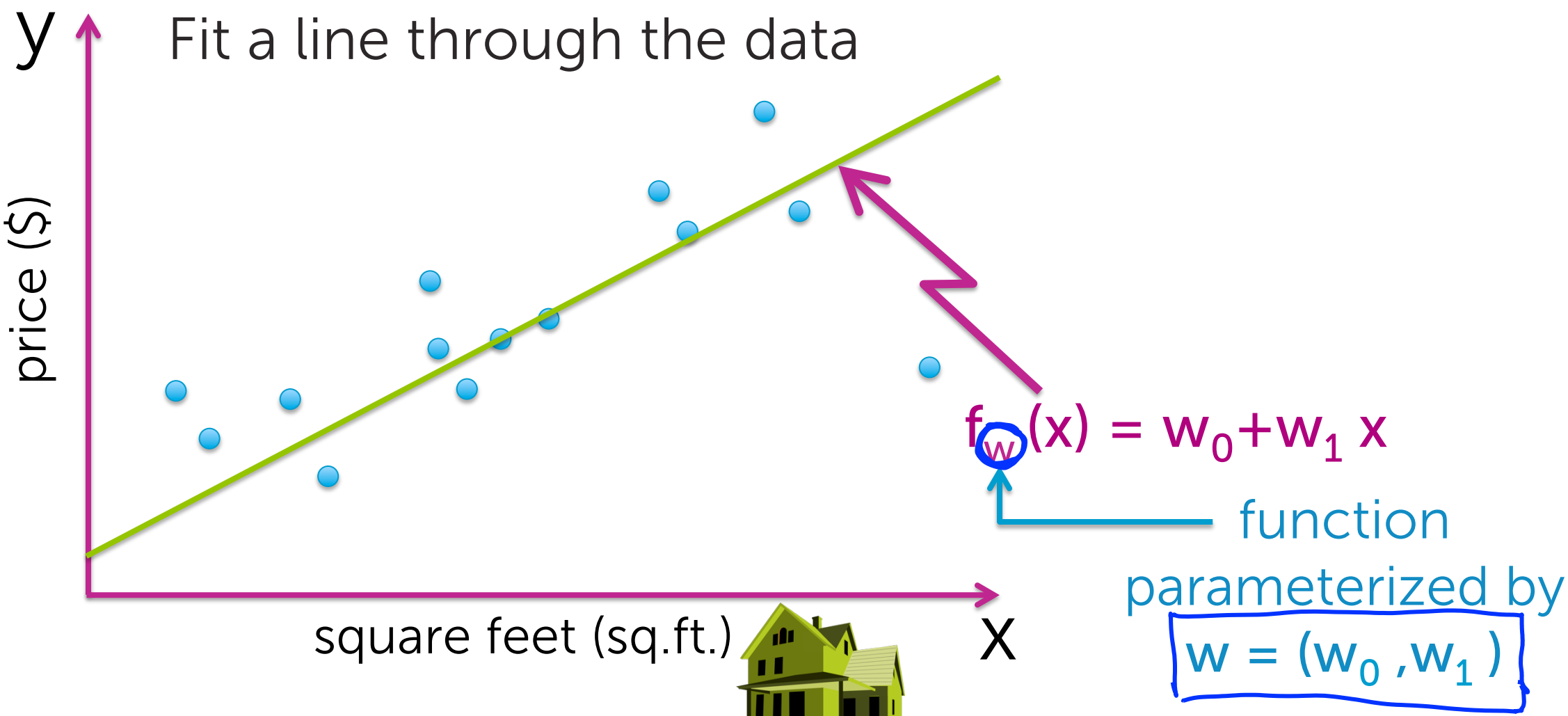
- Look at average price in range
- **Still only 2 houses!**
- Throwing out info from all other sales

Linear regression

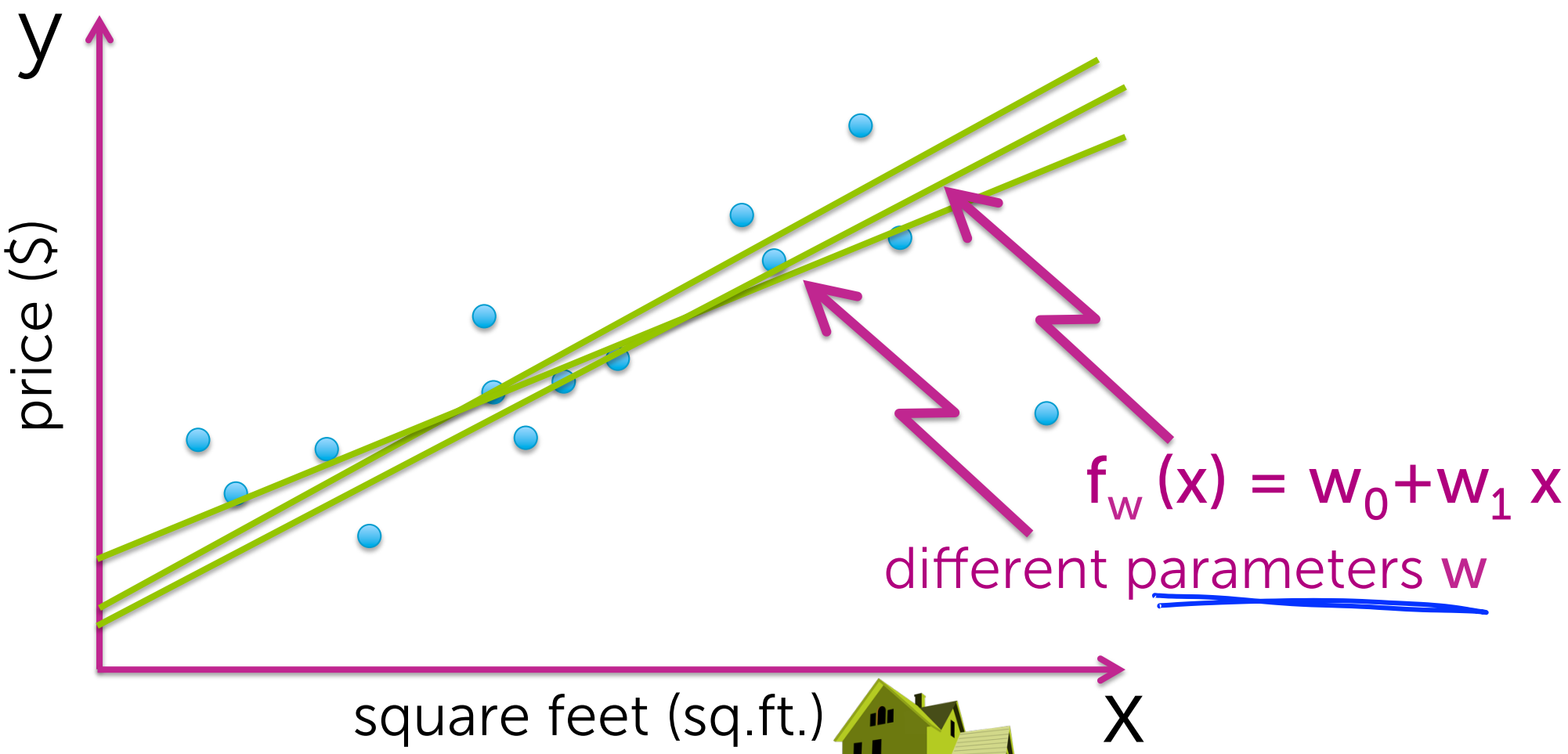
Use a **linear** regression model



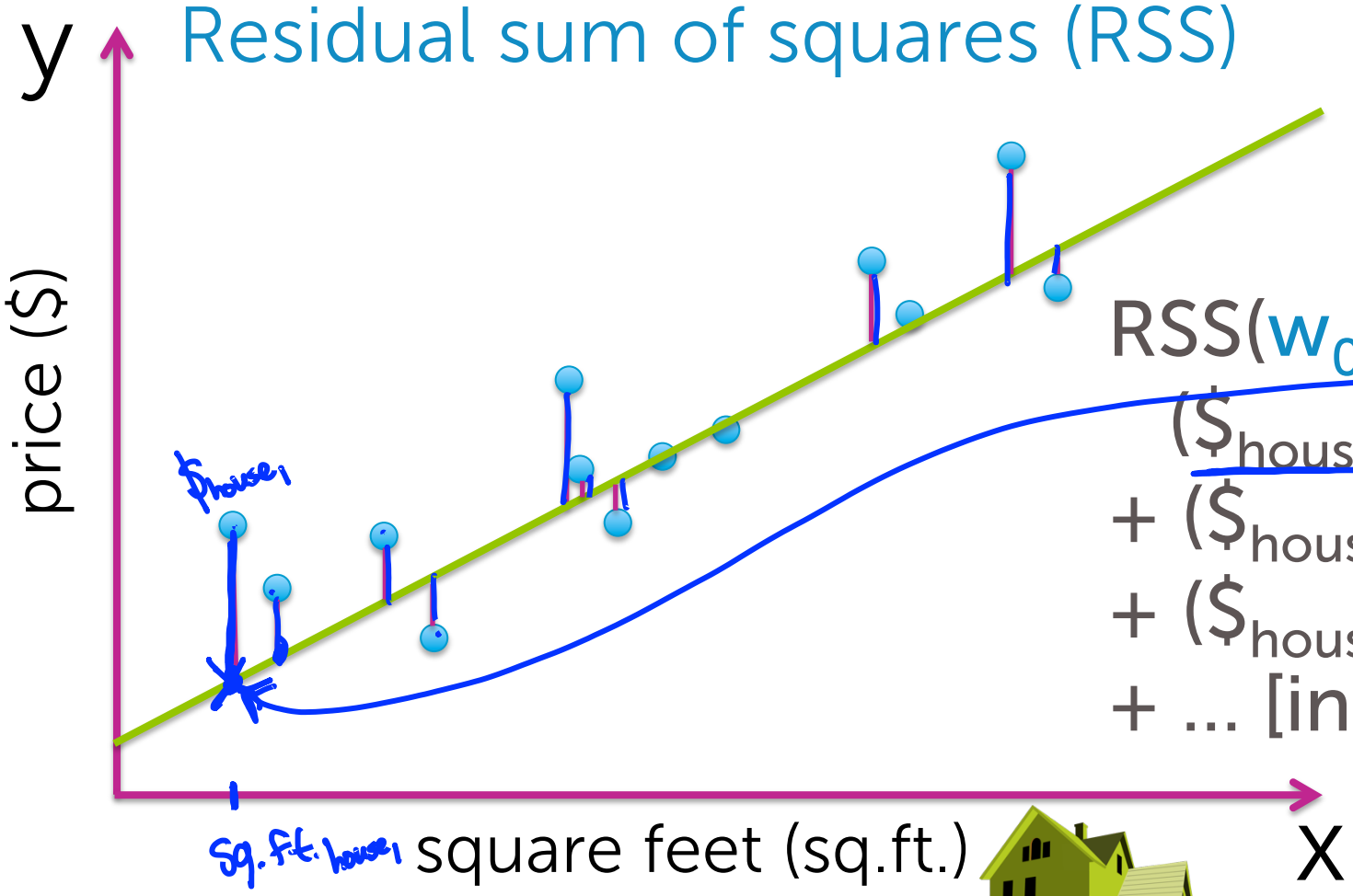
Use a **linear** regression model



Which line?



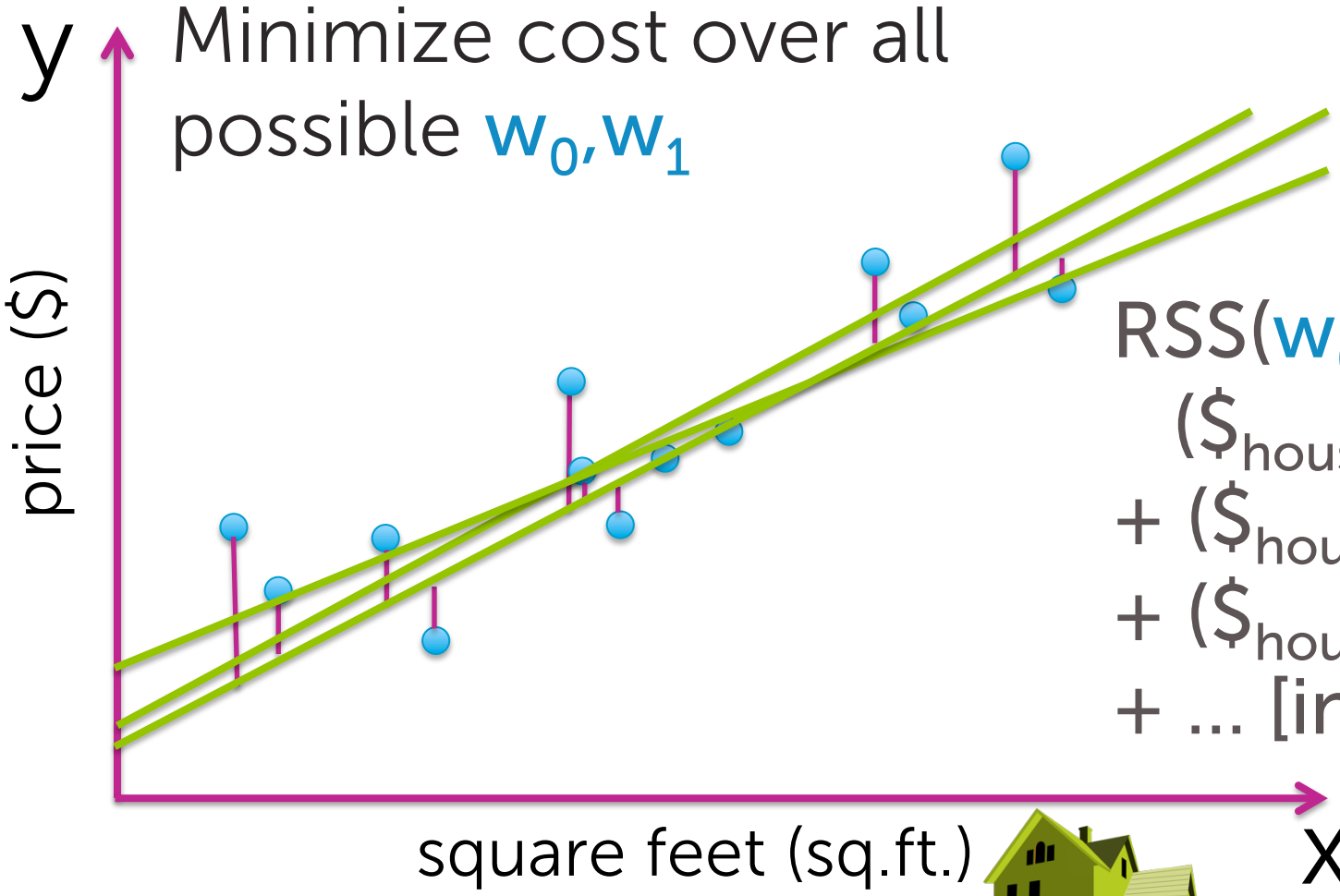
"Cost" of using a given line



Residual sum of squares (RSS)

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

Find "best" line

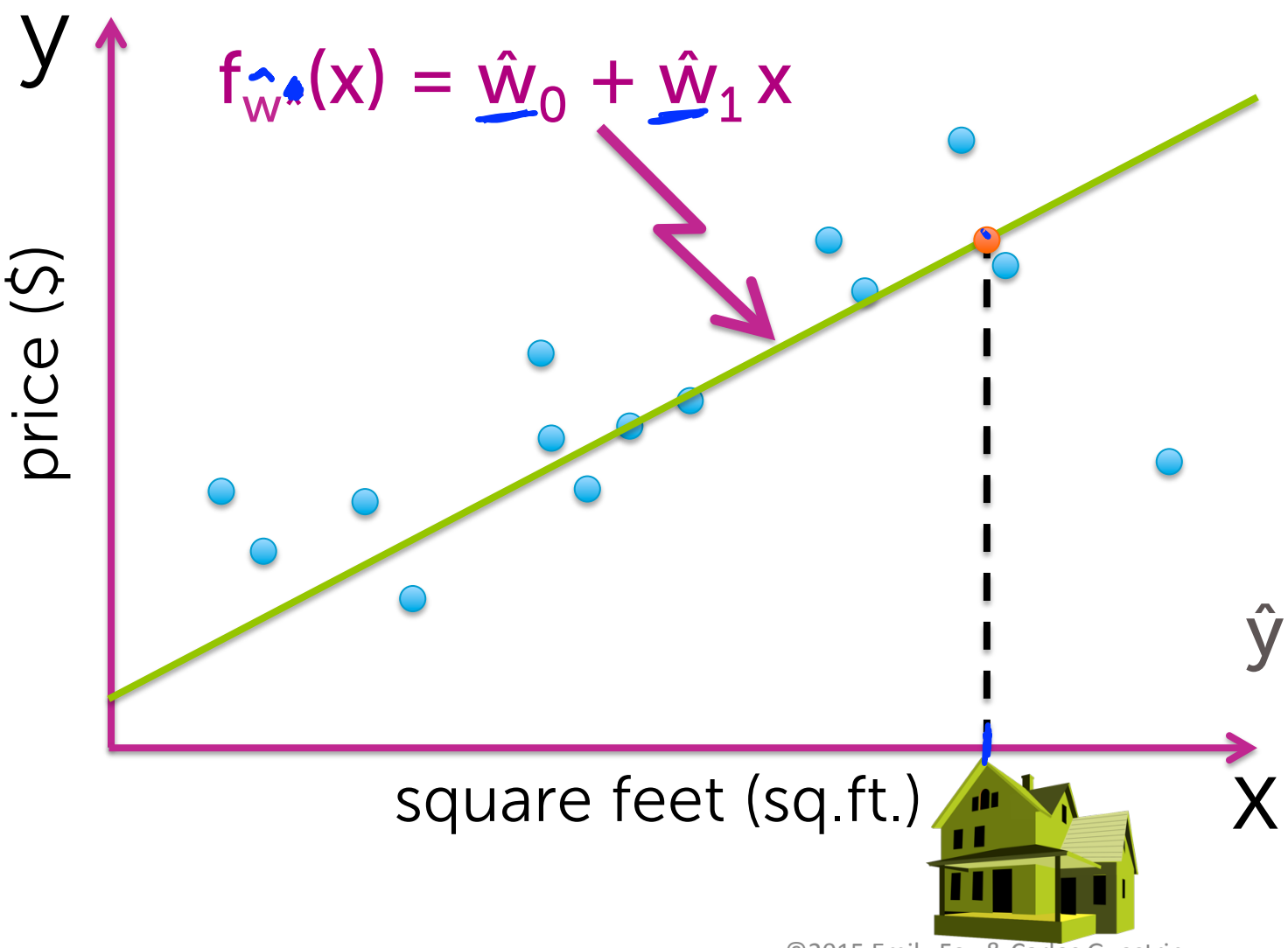


Minimize cost over all possible w_0, w_1

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

$$\hat{W} = (\hat{w}_0, \hat{w}_1)$$

Predicting your house price

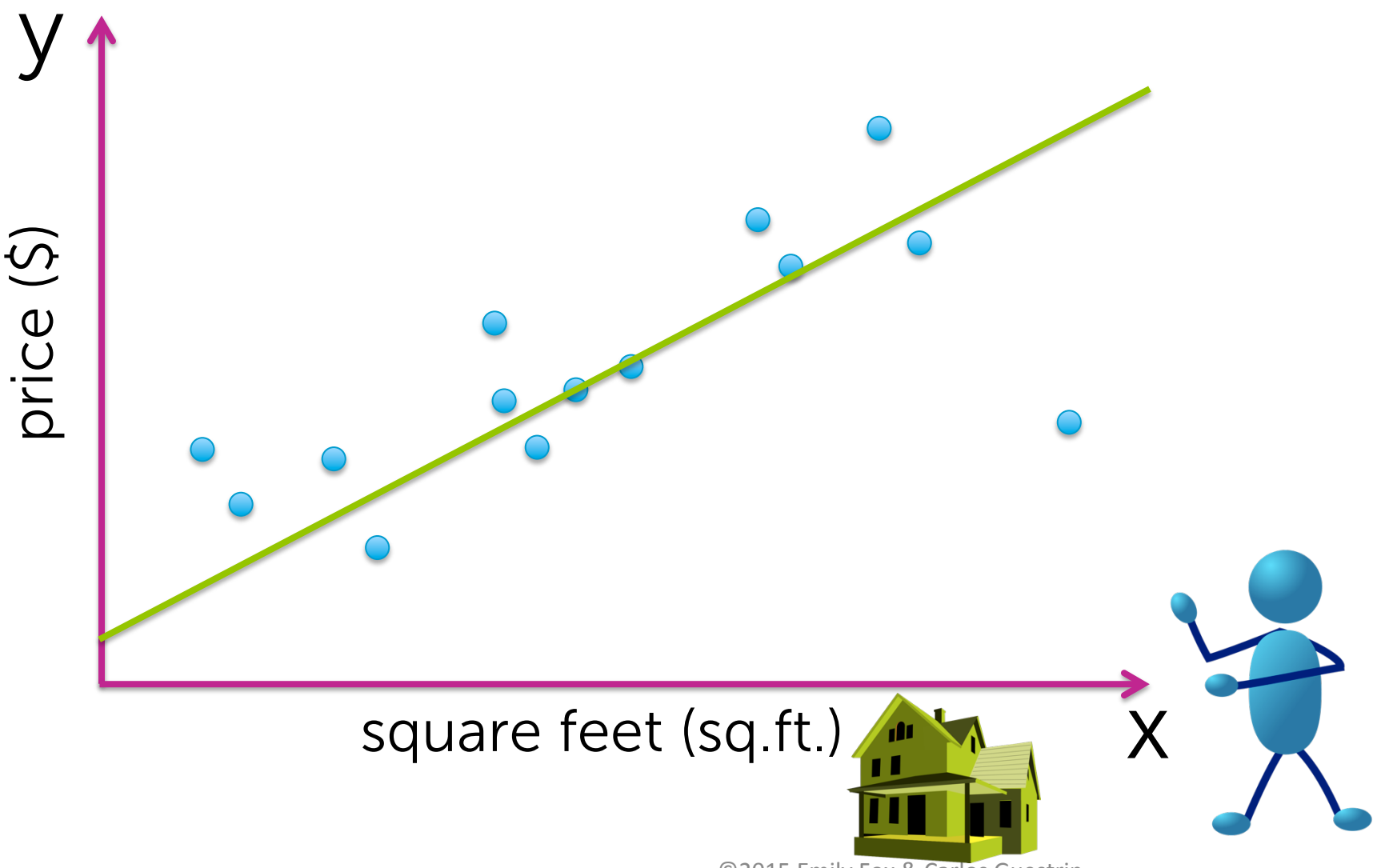


Best guess of your house price:

$\hat{y} = \hat{w}_0 + \hat{w}_1 \text{sq.ft.}_{\text{your house}}$

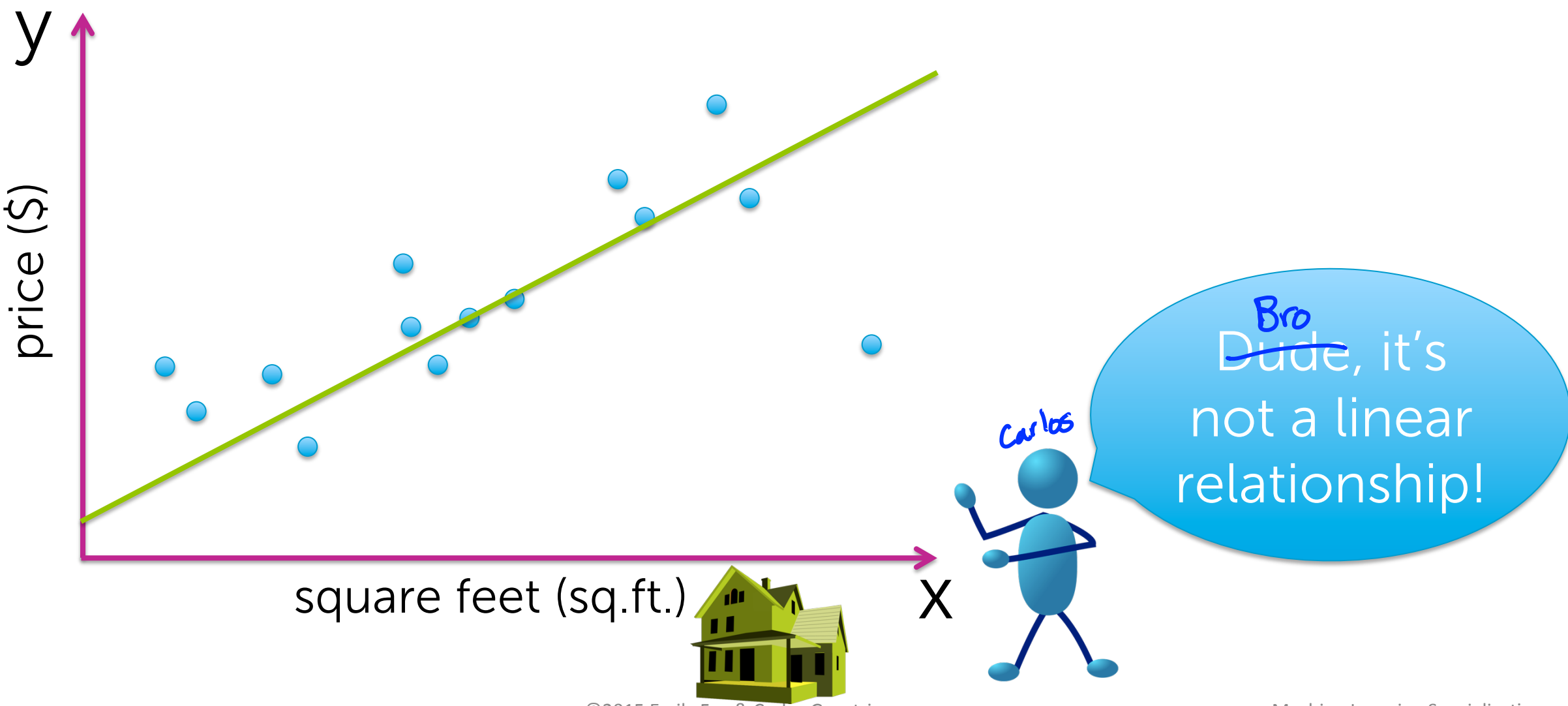
Adding higher order effects

Fit data with a line or … ?

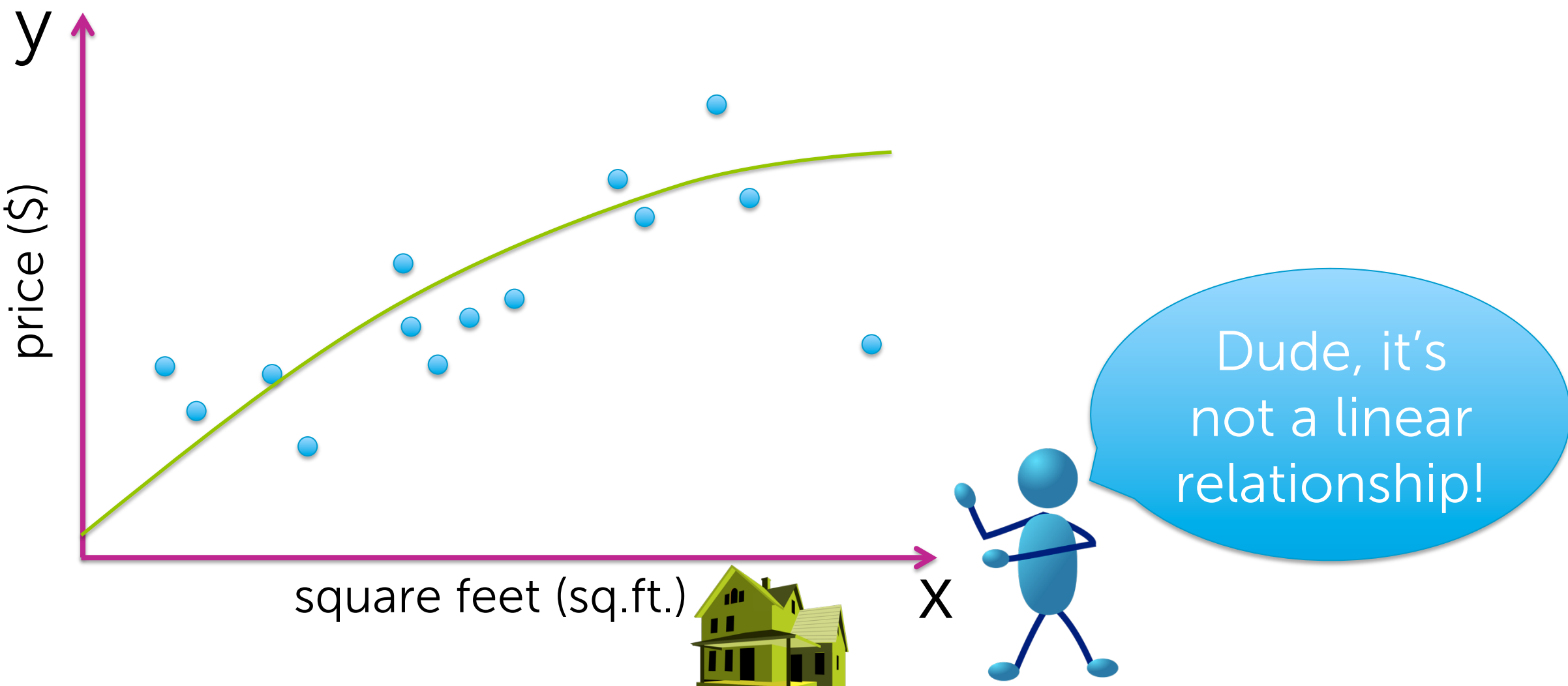


You show your friend your analysis

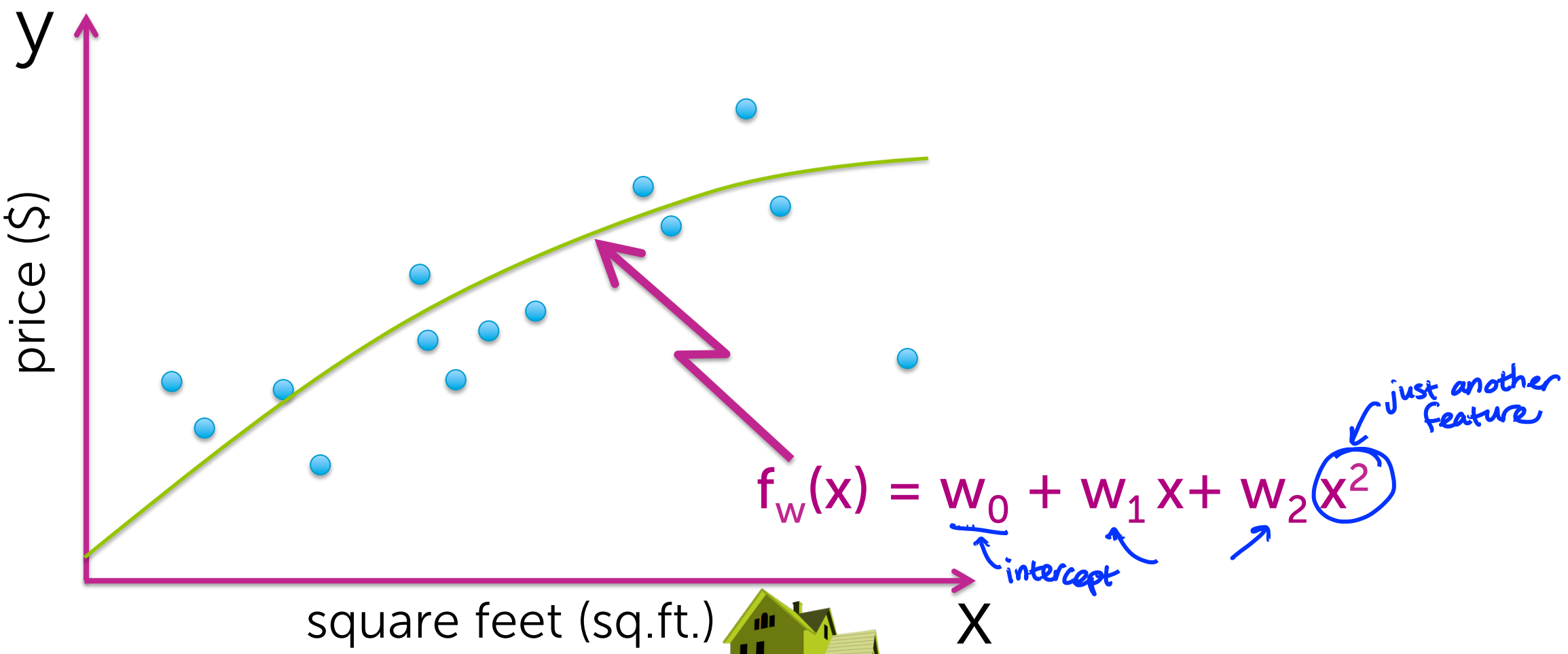
Fit data with a line or … ?



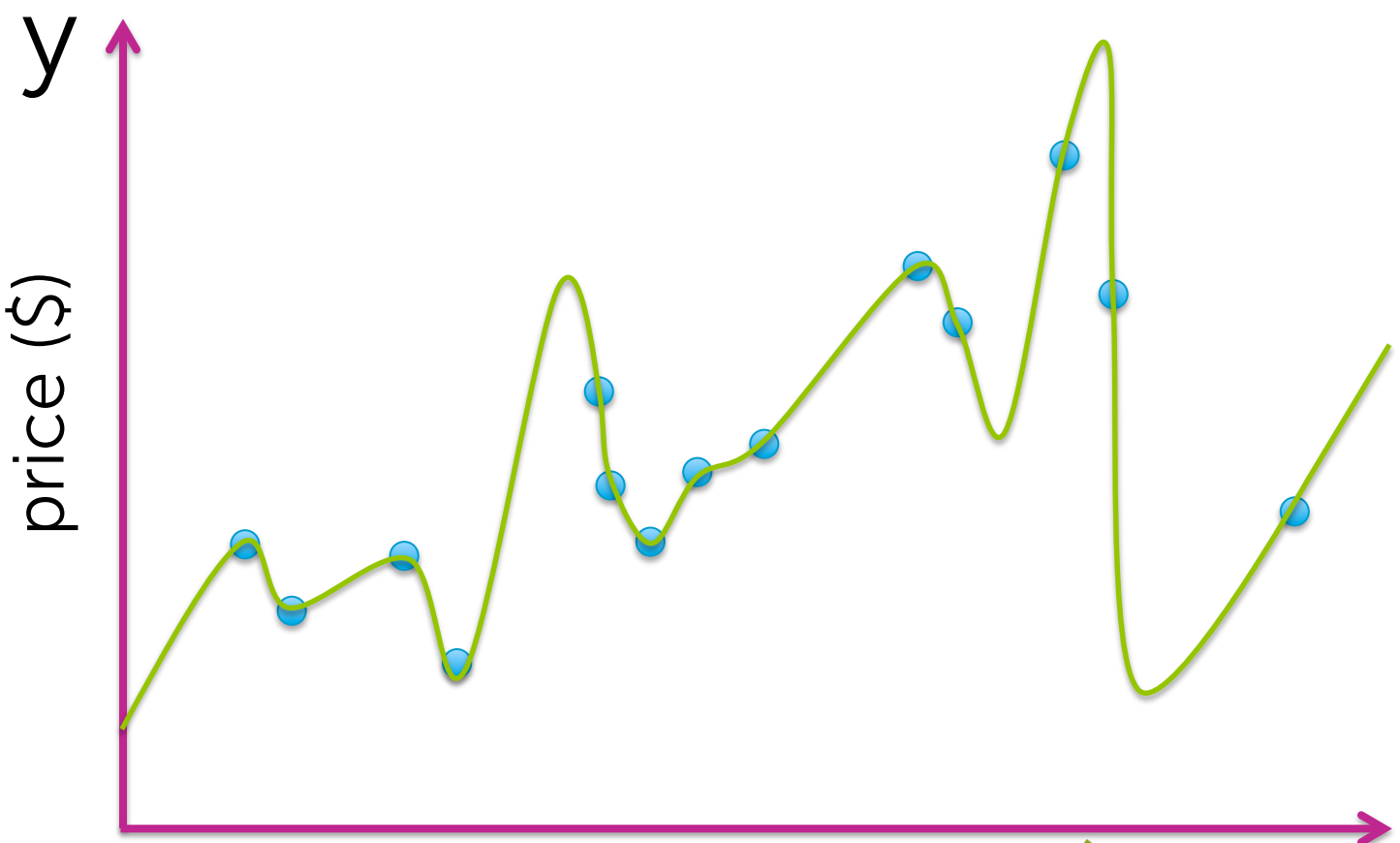
What about a quadratic function?



What about a quadratic function?

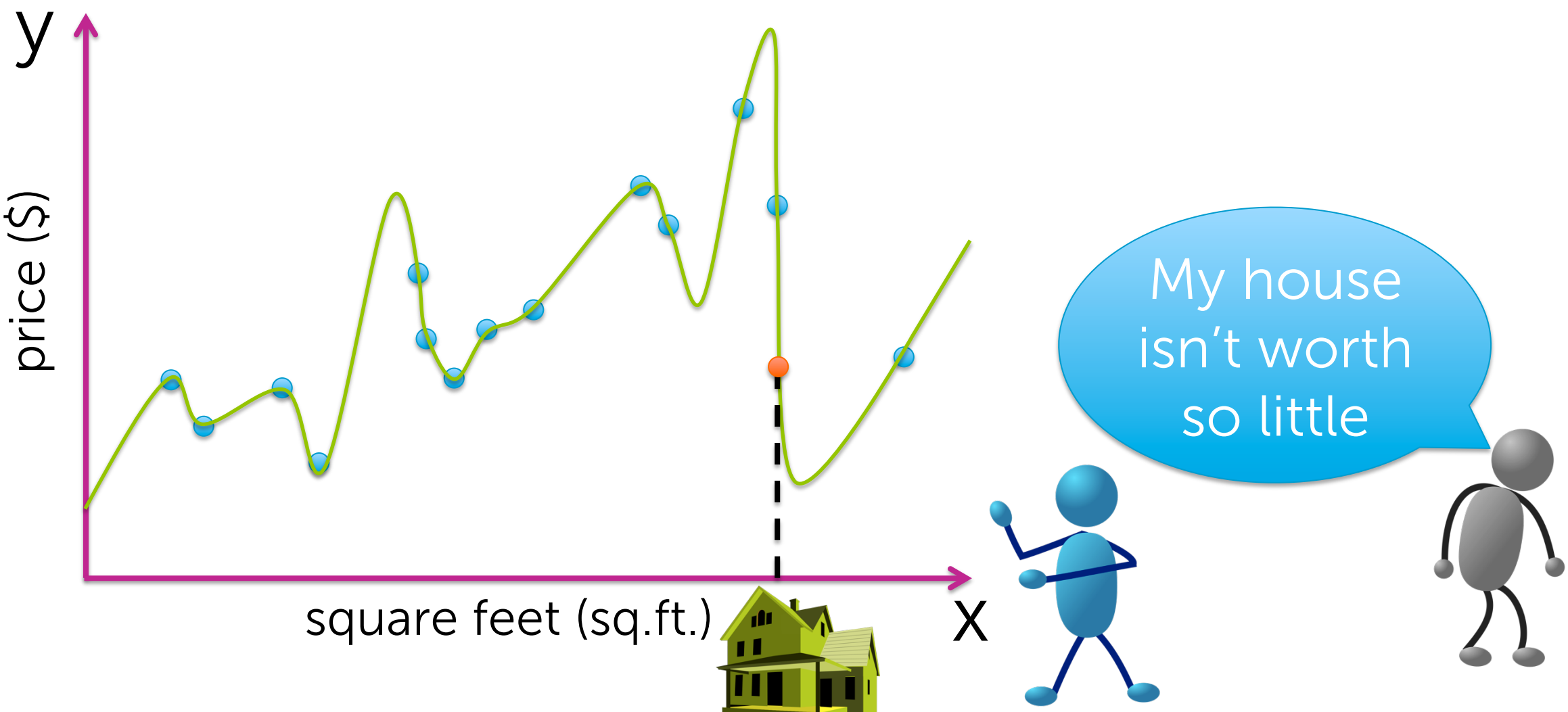


Even higher order polynomial



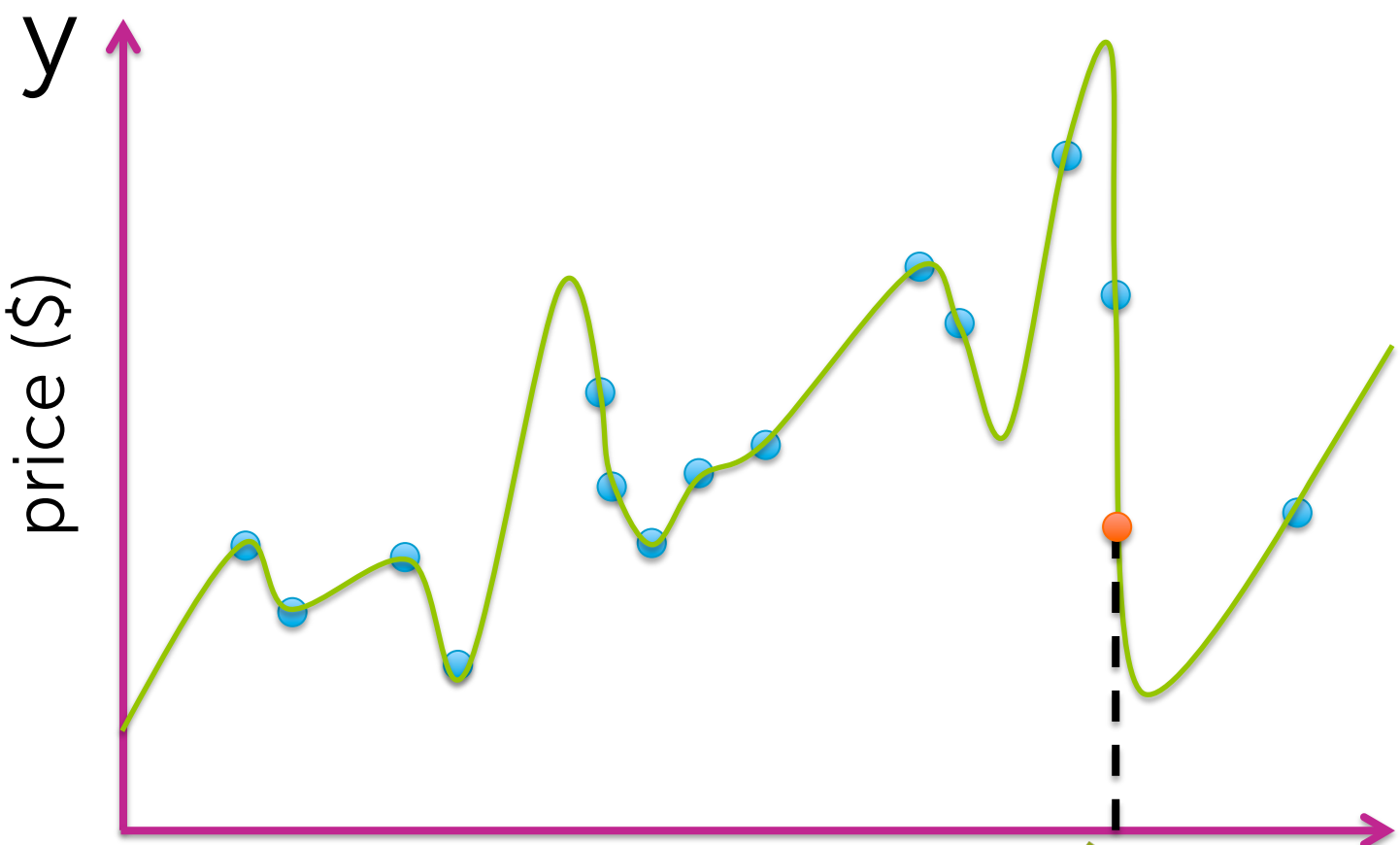
I can minimize your RSS

Do you believe this fit?

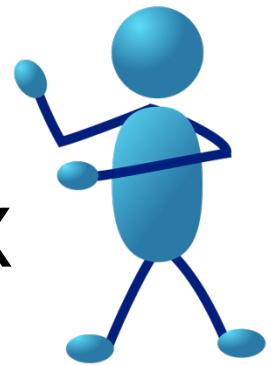


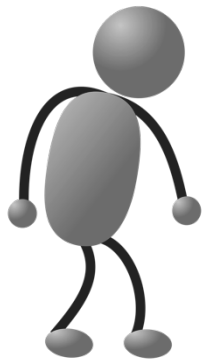
Evaluating overfitting via training/test split

Do you believe this fit?

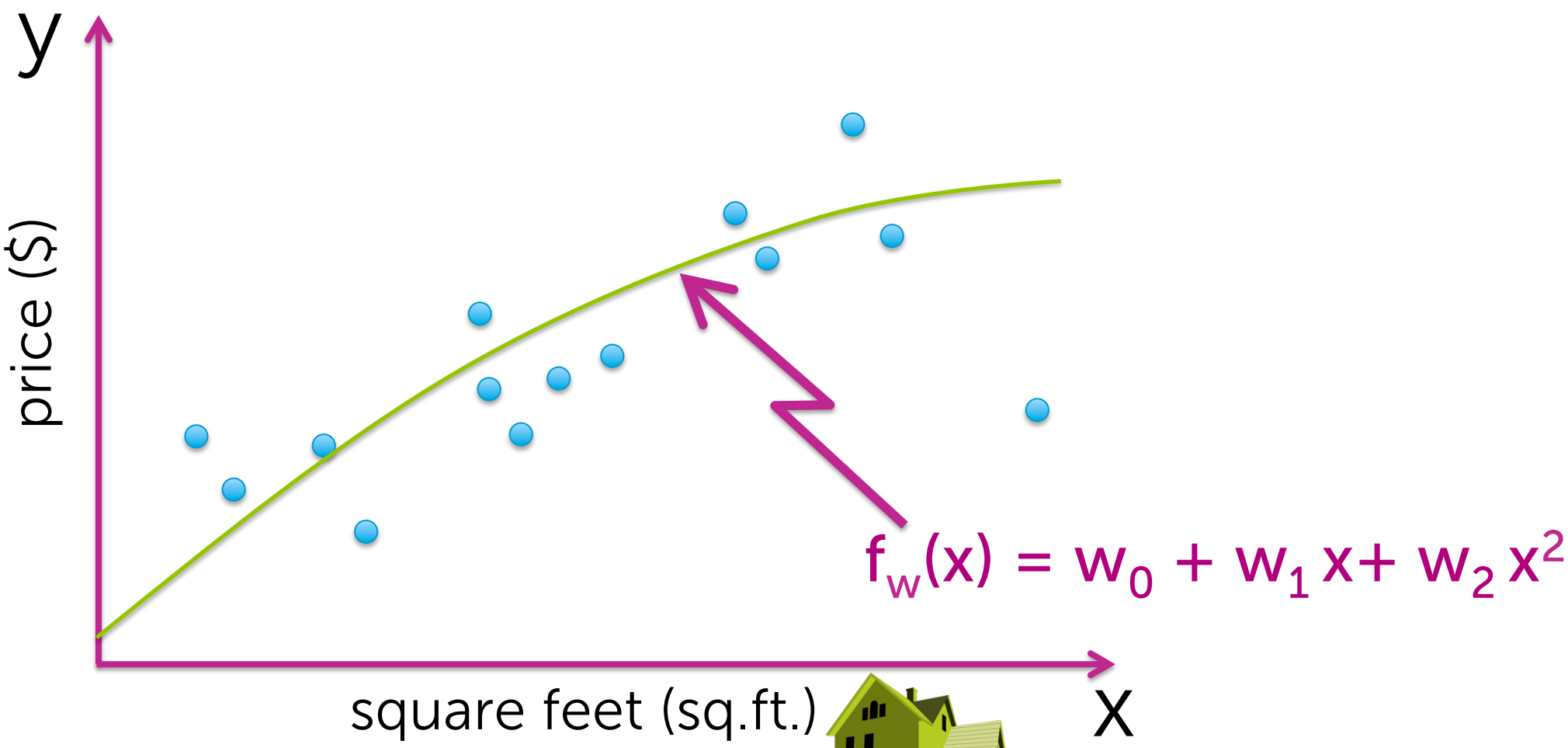


Minimizes RSS,
but bad predictions

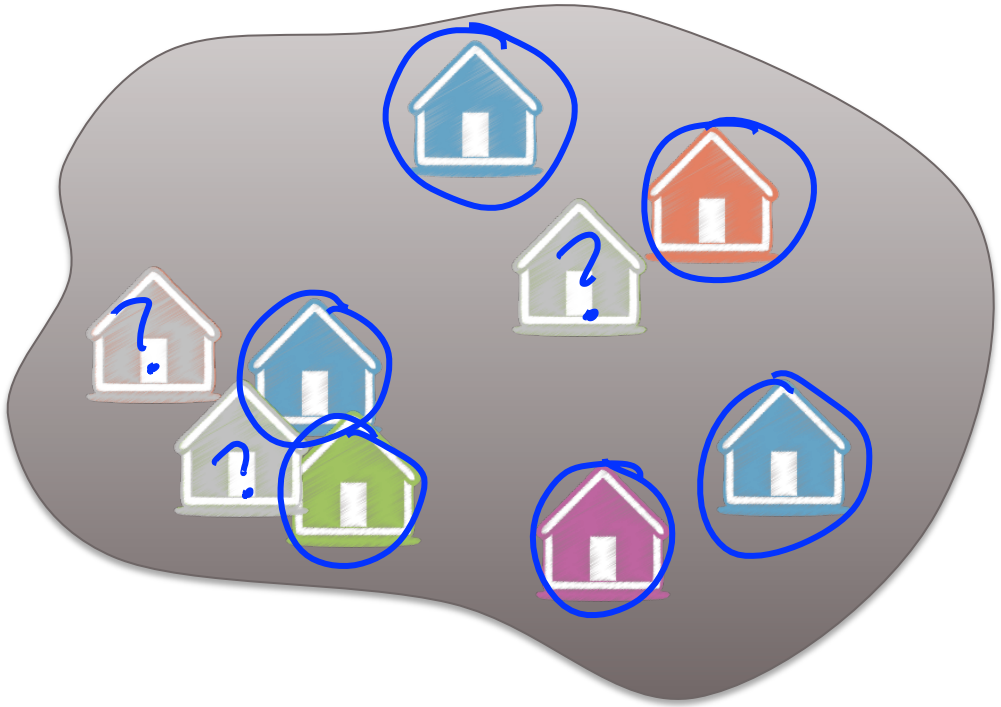




What about a quadratic function?



How to choose model order/complexity



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

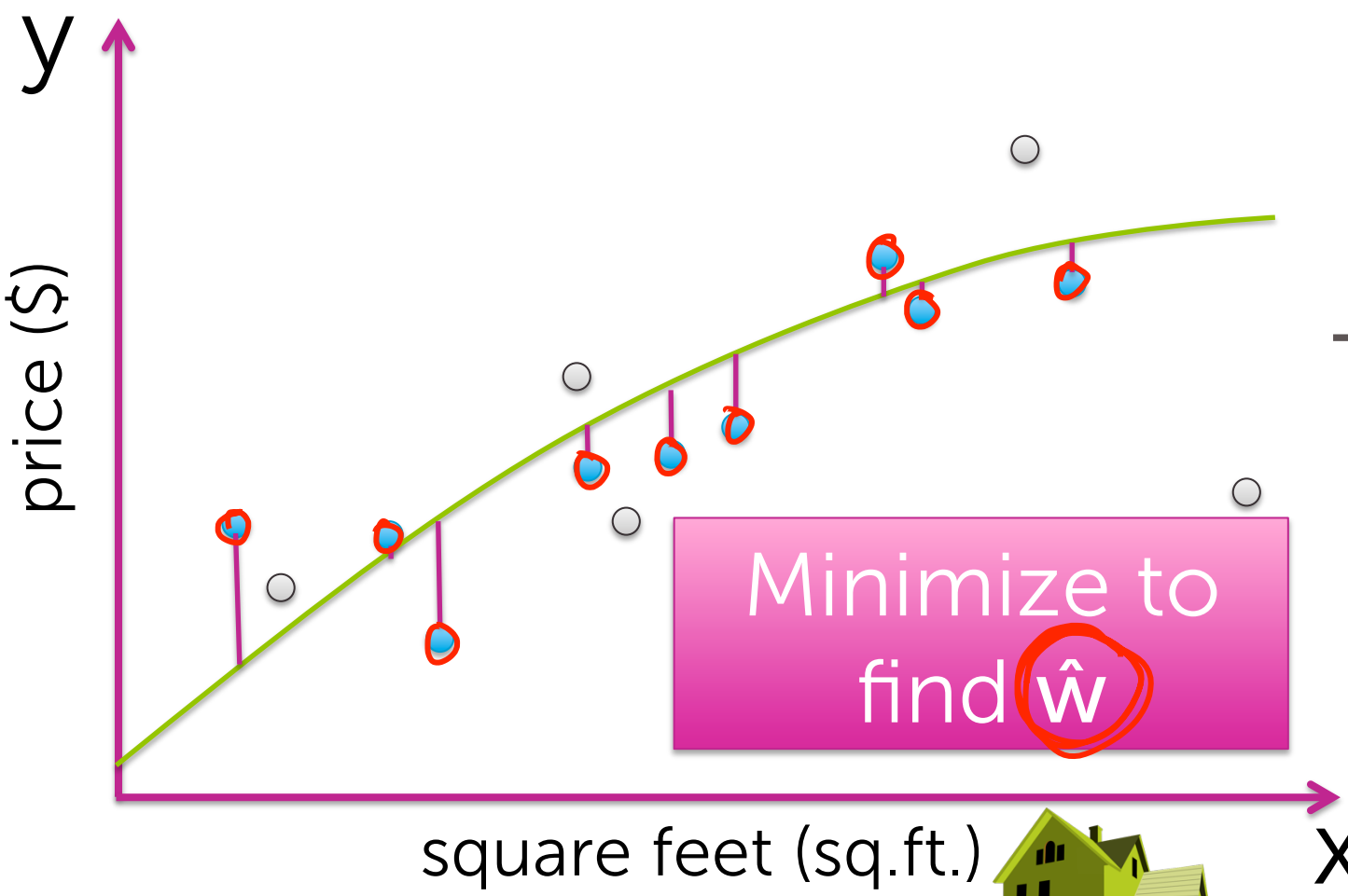
1. Remove some houses
2. Fit model on remaining
3. Predict heldout houses

Training/test split

Terminology: – training set

– test set

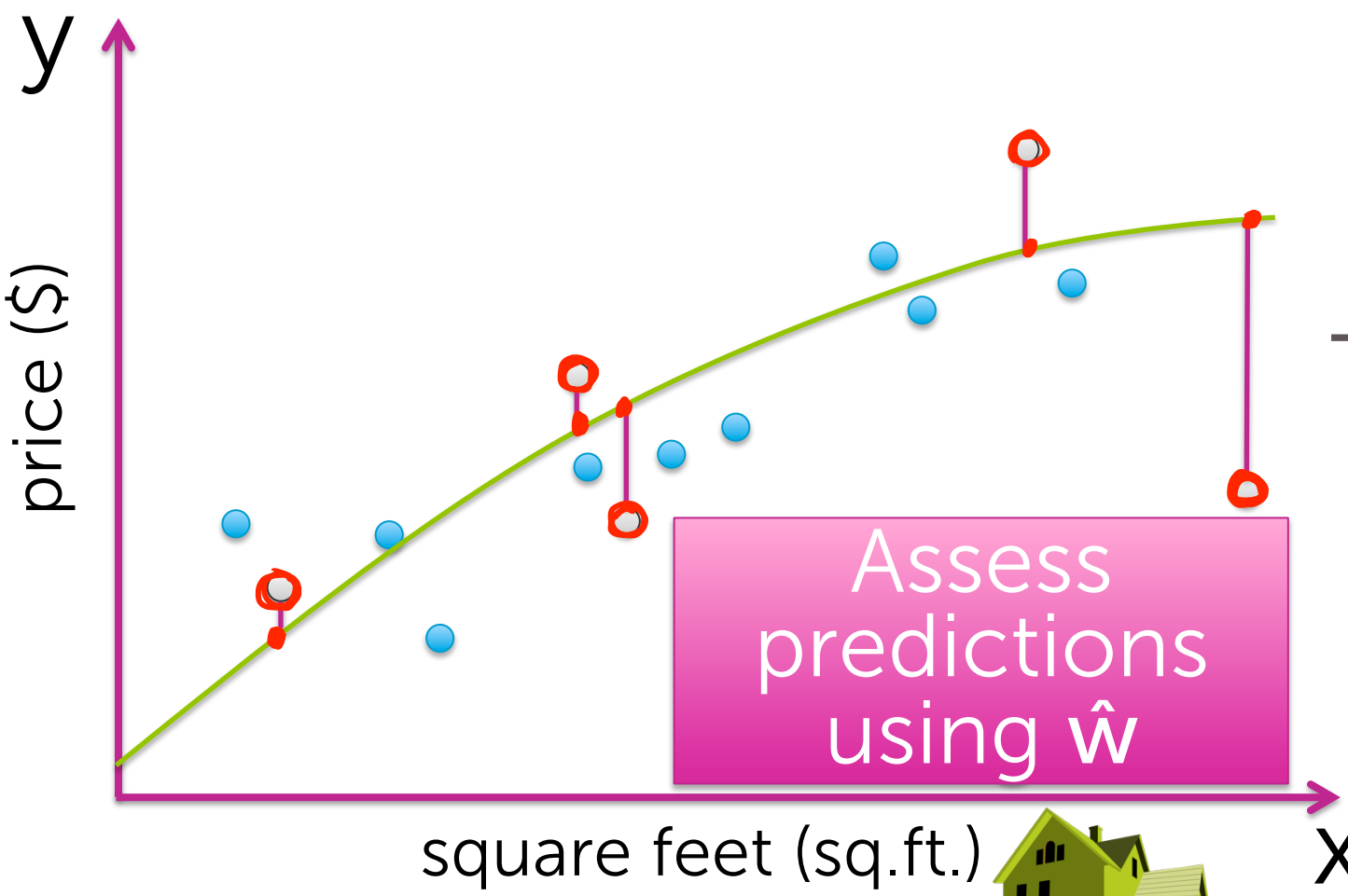
Training error



Training error (w) =

$$(\$_{\text{train 1}} - f_w(\text{sq.ft.}_{\text{train 1}}))^2$$
$$+ (\$_{\text{train 2}} - f_w(\text{sq.ft.}_{\text{train 2}}))^2$$
$$+ (\$_{\text{train 3}} - f_w(\text{sq.ft.}_{\text{train 3}}))^2$$
$$+ \ldots \text{ [include all training houses]}$$

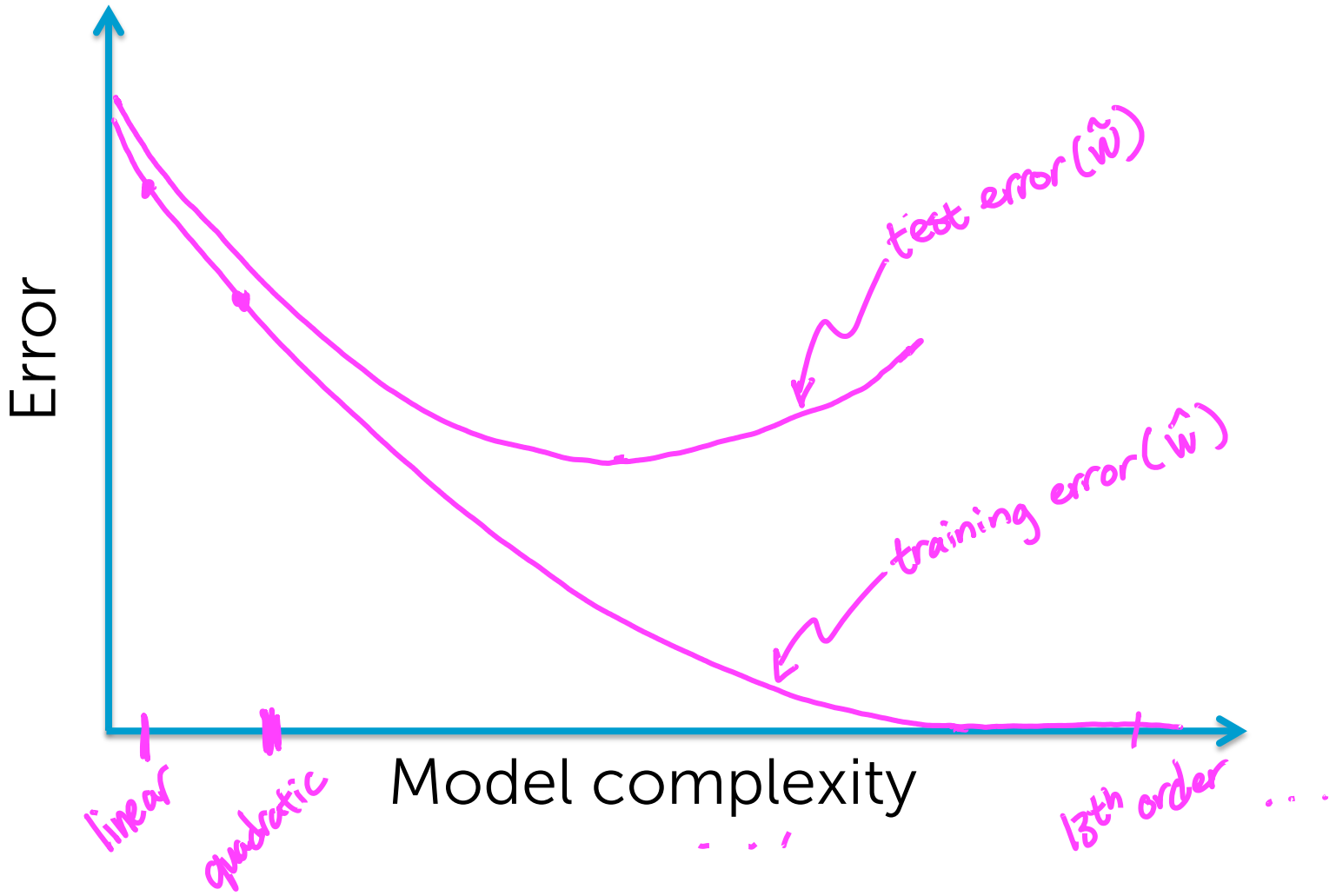
Test error



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

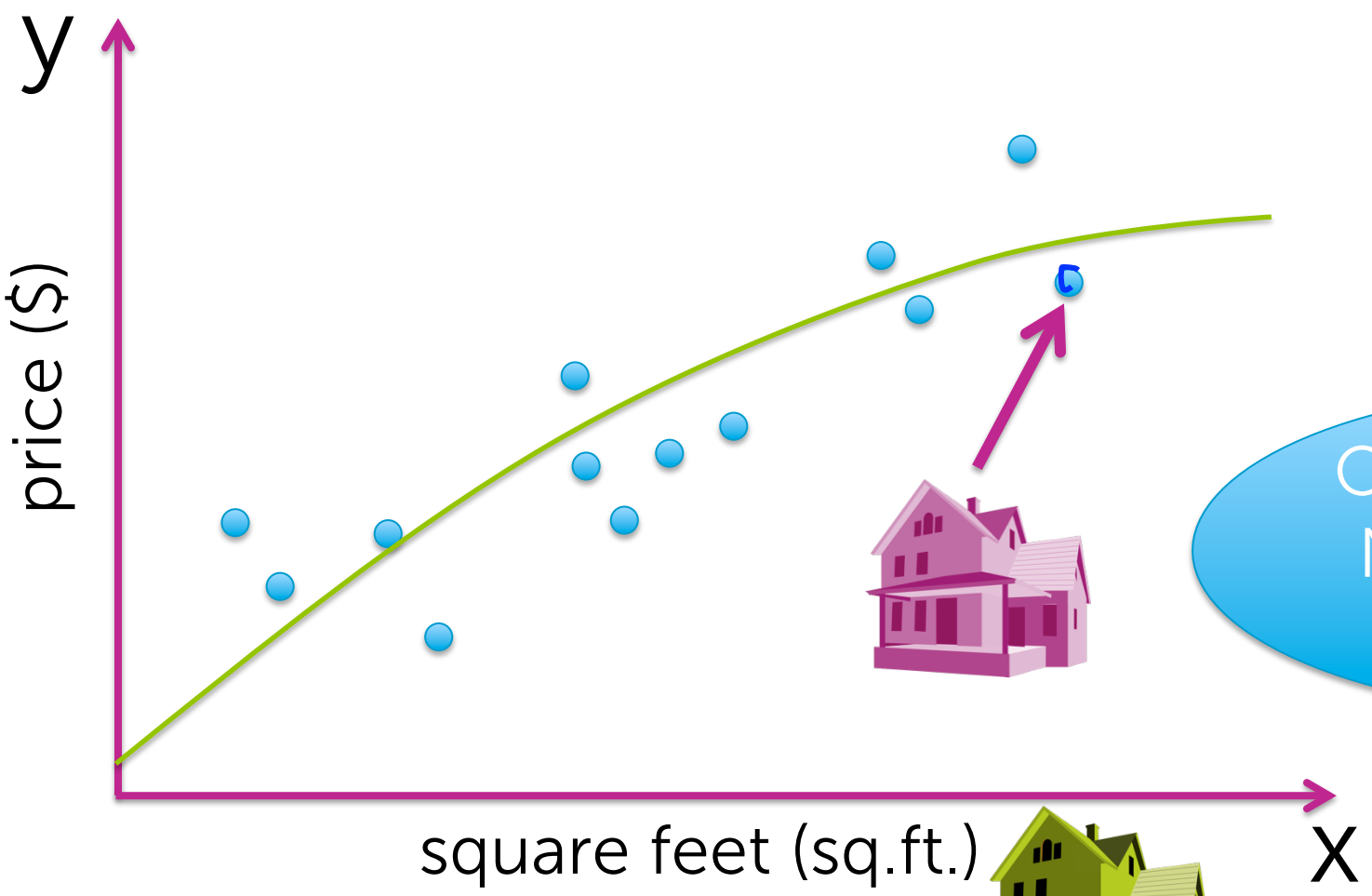
$(\$_{test\ 1} - f_{\hat{w}}(sq.ft._{test\ 1}))^2$
$+ (\$_{test\ 2} - f_{\hat{w}}(sq.ft._{test\ 2}))^2$
$+ (\$_{test\ 3} - f_{\hat{w}}(sq.ft._{test\ 3}))^2$
+ ... [include all test houses]

Training/Test Curves



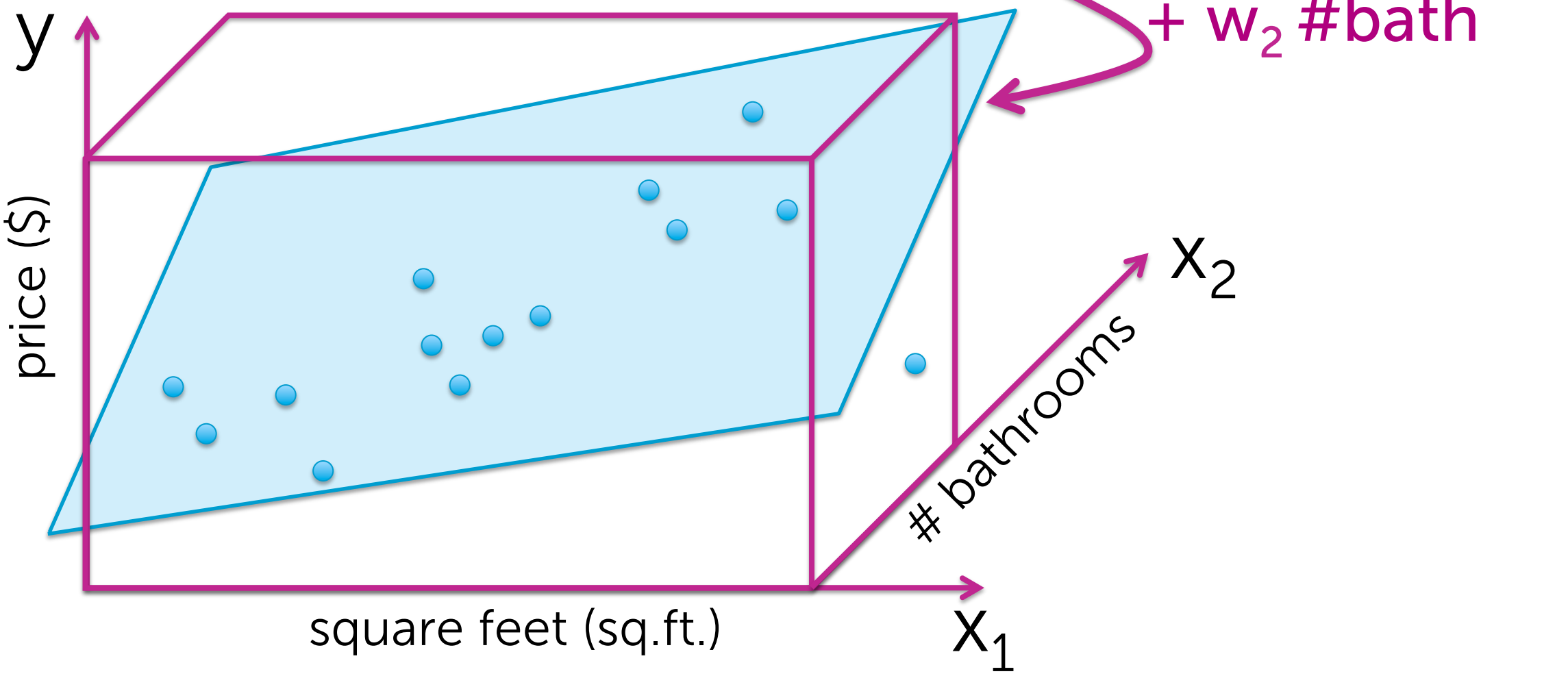
Adding other features

Predictions just based on house size



Add more features

$f_w(x) = w_0 + w_1$ sq.ft. $+ w_2$ #bath



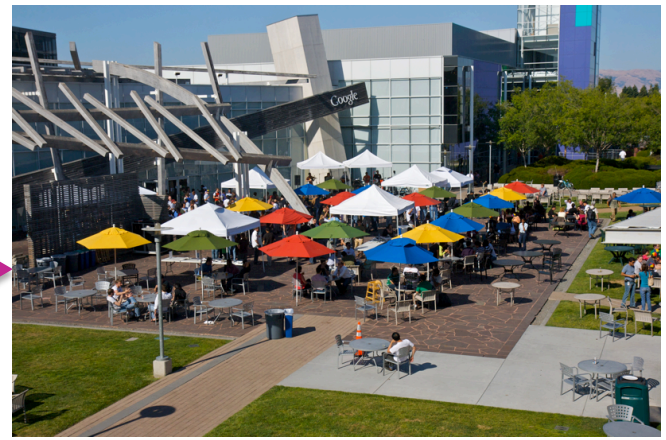
How many features to use?

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

Other regression examples

Salary after ML specialization

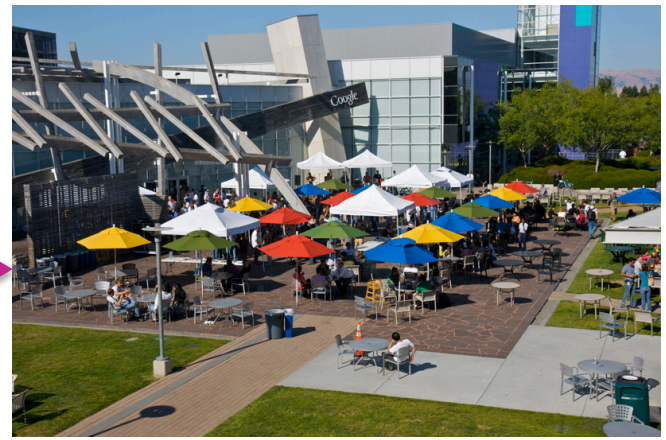
hard work



- How much will your salary be? (**y** = $$)
- Depends on **x** = performance in courses, quality of capstone project, # of forum responses, …

Salary after ML specialization

hard work

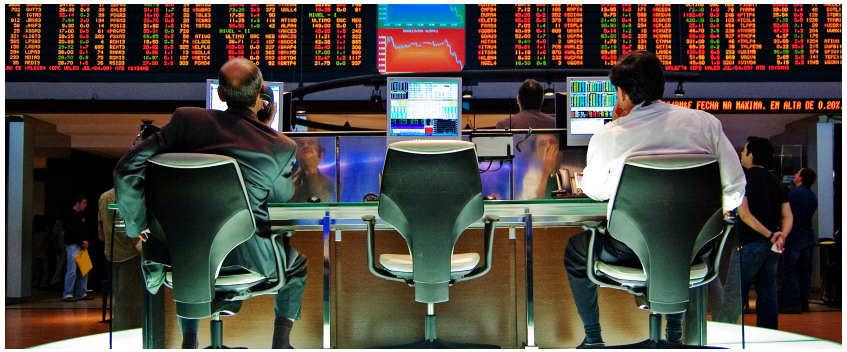


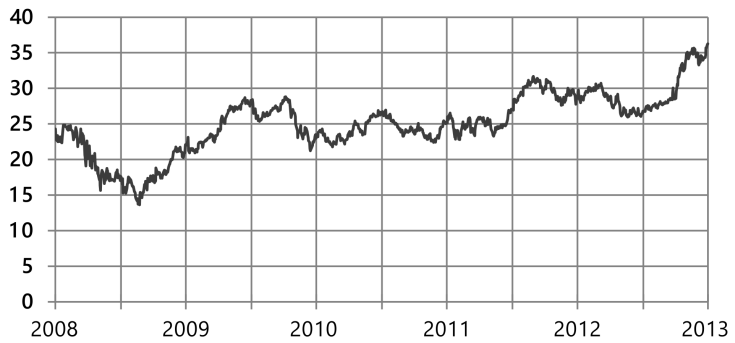
$$\hat{y} = \hat{w}_0 + \hat{w}_1 \text{ performance} + \hat{w}_2 \text{ capstone} + \hat{w}_3 \text{ forum}$$

informed by other students who completed specialization

Stock prediction

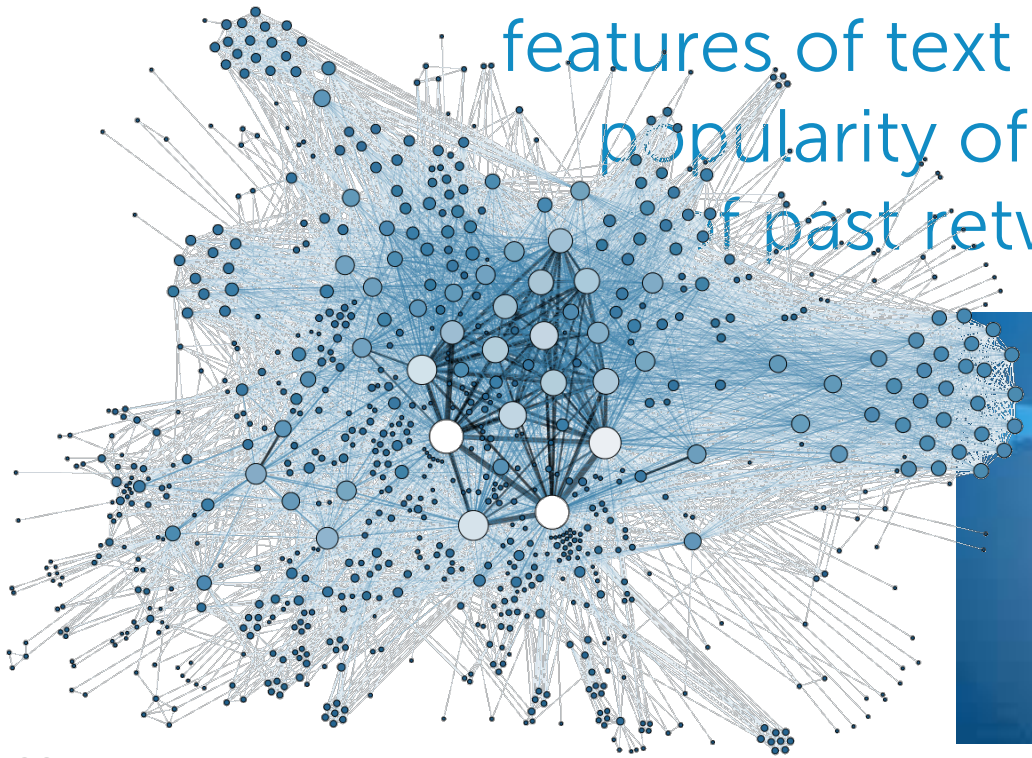
- Predict the price of a stock
- Depends on
  - Recent history of stock price
  - News events
  - Related commodities





Tweet popularity

- How many people will retweet your tweet?
- Depends on # followers,
  # of followers of followers,
  features of text tweeted,
  popularity of hashtag,
  # of past retweets,…



Smart houses

- Smart houses have many distributed sensors
- What's the temperature at your desk? (no sensor)
  - Learn spatial function to predict temp
- Also depends on
  - Thermostat setting
  - Blinds open/closed or window tint
  - Vents
  - Temperature outside
  - Time of day

Summary for regression

What you can do now...

- Describe the input (features) and output (real-valued predictions) of a regression model
- Calculate a goodness-of-fit metric (e.g., RSS)
- Estimate model parameters by minimizing RSS (algorithms to come...)
- Exploit the estimated model to form predictions
- Perform a training/test split of the data
- Analyze performance of various regression models in terms of test error
- Use test error to avoid overfitting when selecting amongst candidate models
- Describe a regression model using multiple features
- Describe other applications where regression is useful