INTRODUCTION TO DATA SCIENCE

This lecture is
based on course by E. Fox and C. Guestrin, Univ of Washington

19/10/2022

WFAiS UJ, Informatyka Stosowana
I stopień studiów

Regression for predictions

- **Primer**
- **Advanced**
  - **Linear regression**
  - **Multiple regression**
  - **Accesing performance**
  - **Ridge regression**
  - **Feature selection and lasso regression**
  - **Nearest neighbor and kernel regression**

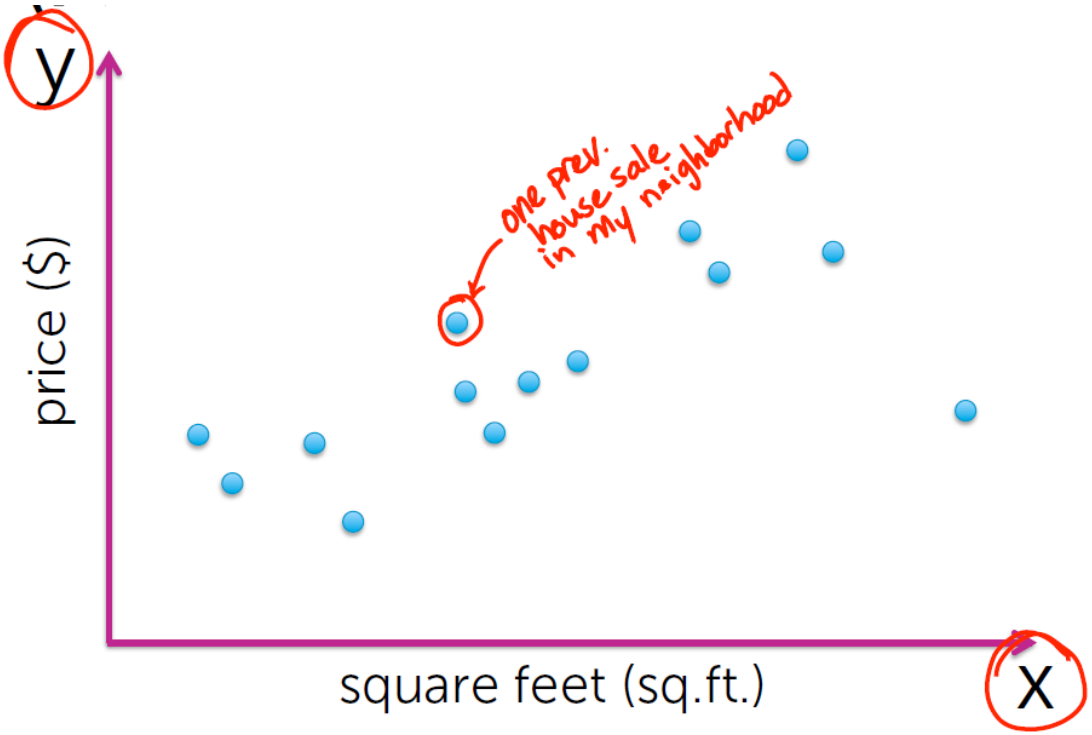
How much is my house worth

- **Predicting value of the house**

Lets look at the recent
sales in the neighborhood.
How much did they sell for?
What do that houses look like?

Naive: plot recent house sales

- **We take observations that we have and make a plot of them.**

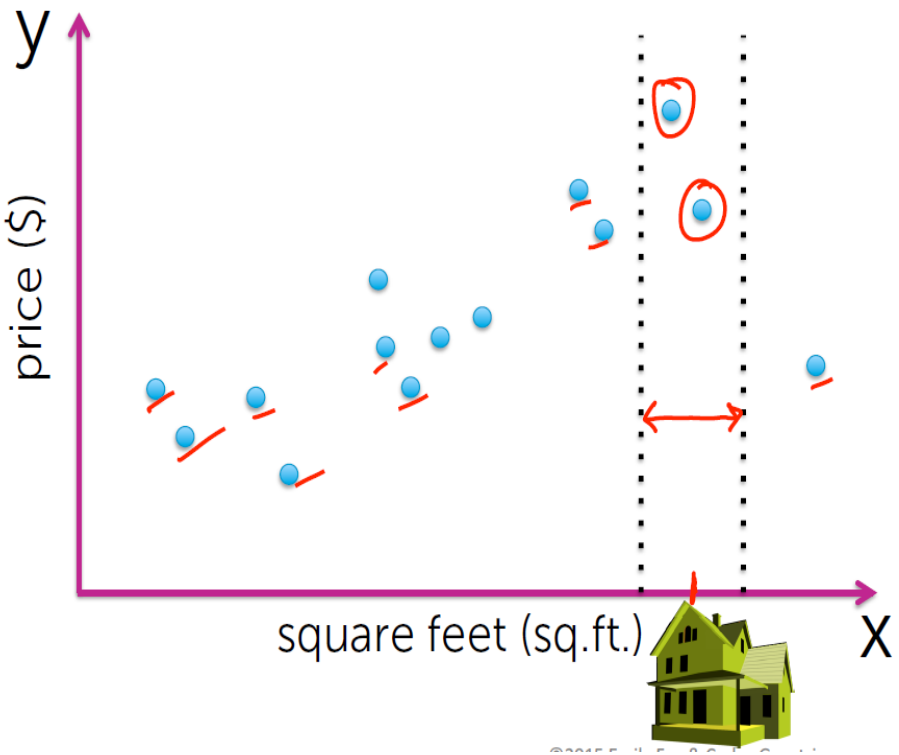


Terminology:

x – feature, covariate, or predictor

y – observation or response

Predict by prizes of similar houses

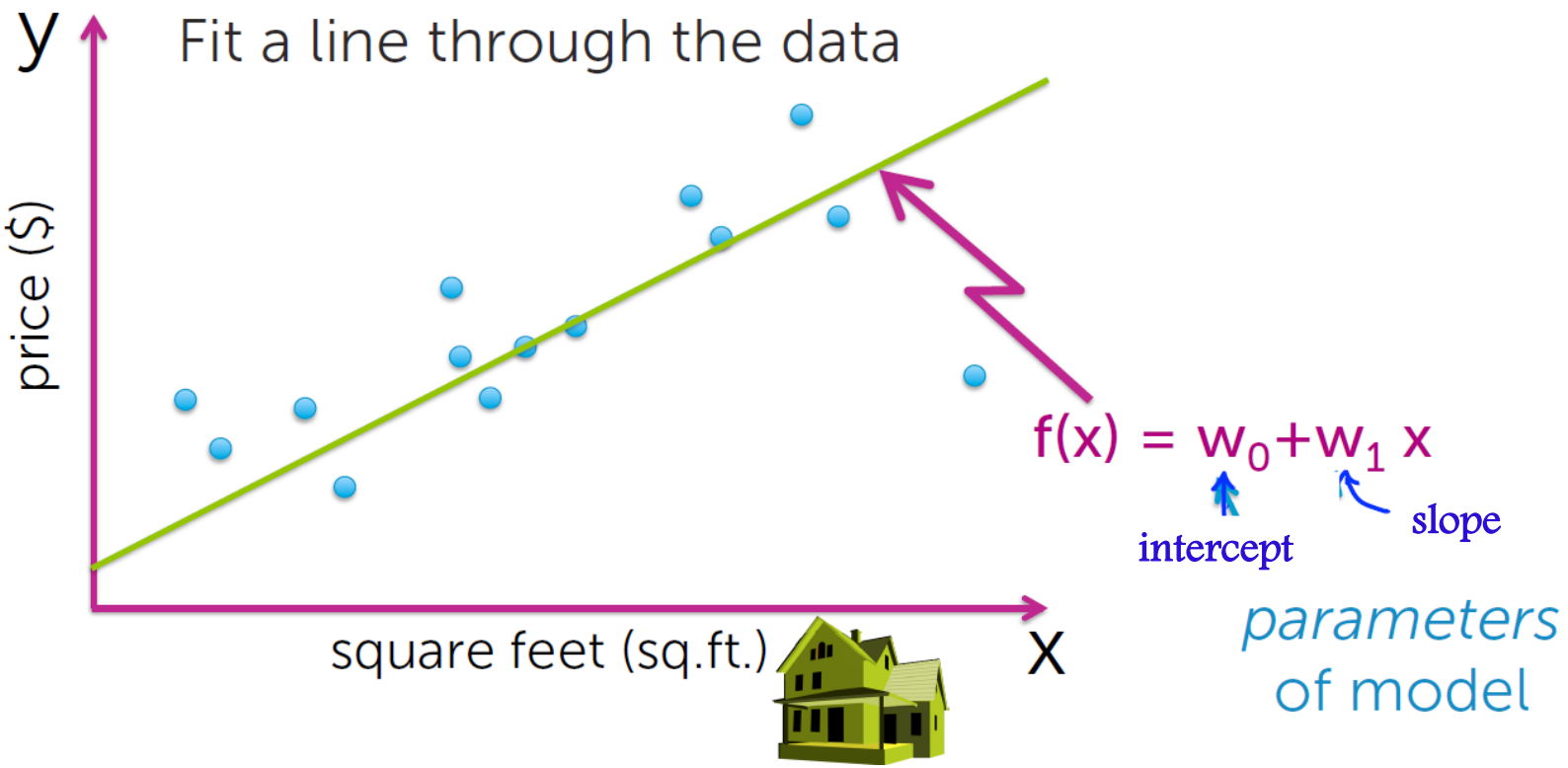


- Look at average price in range
- **Still only 2 houses!**
- Throwing out info from all other sales

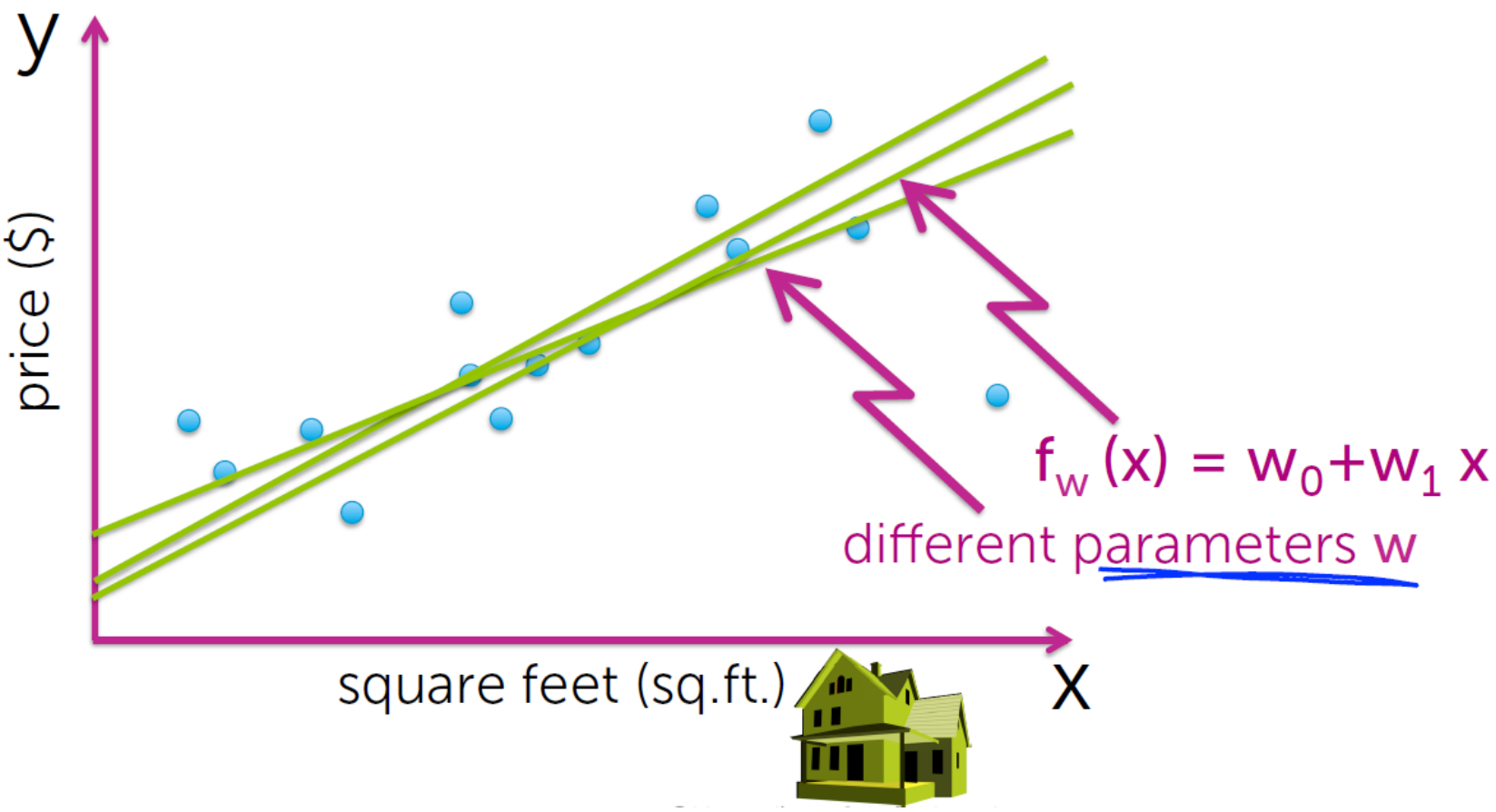
Is it really reasonably to believe that there is no information there? We would like to leverage all avaible information.

Linear regression: a model based relation

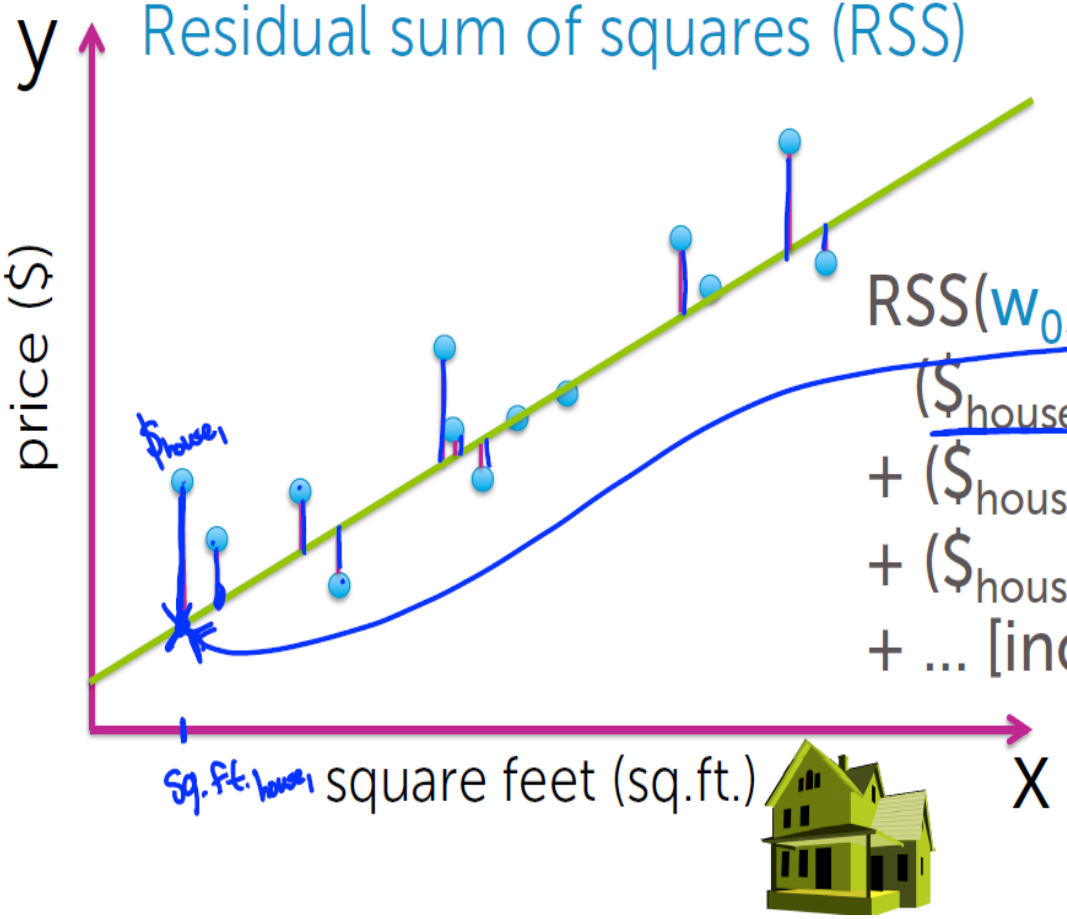
Use a linear regression model



Which line?

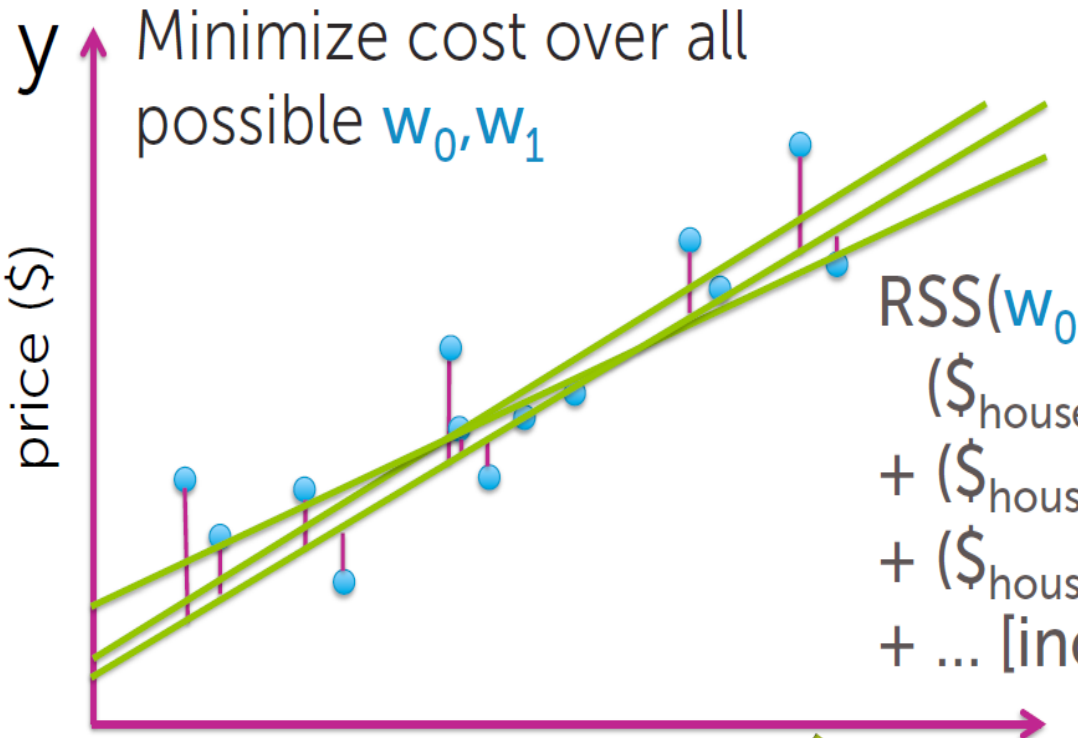


Defining a cost of a given line



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

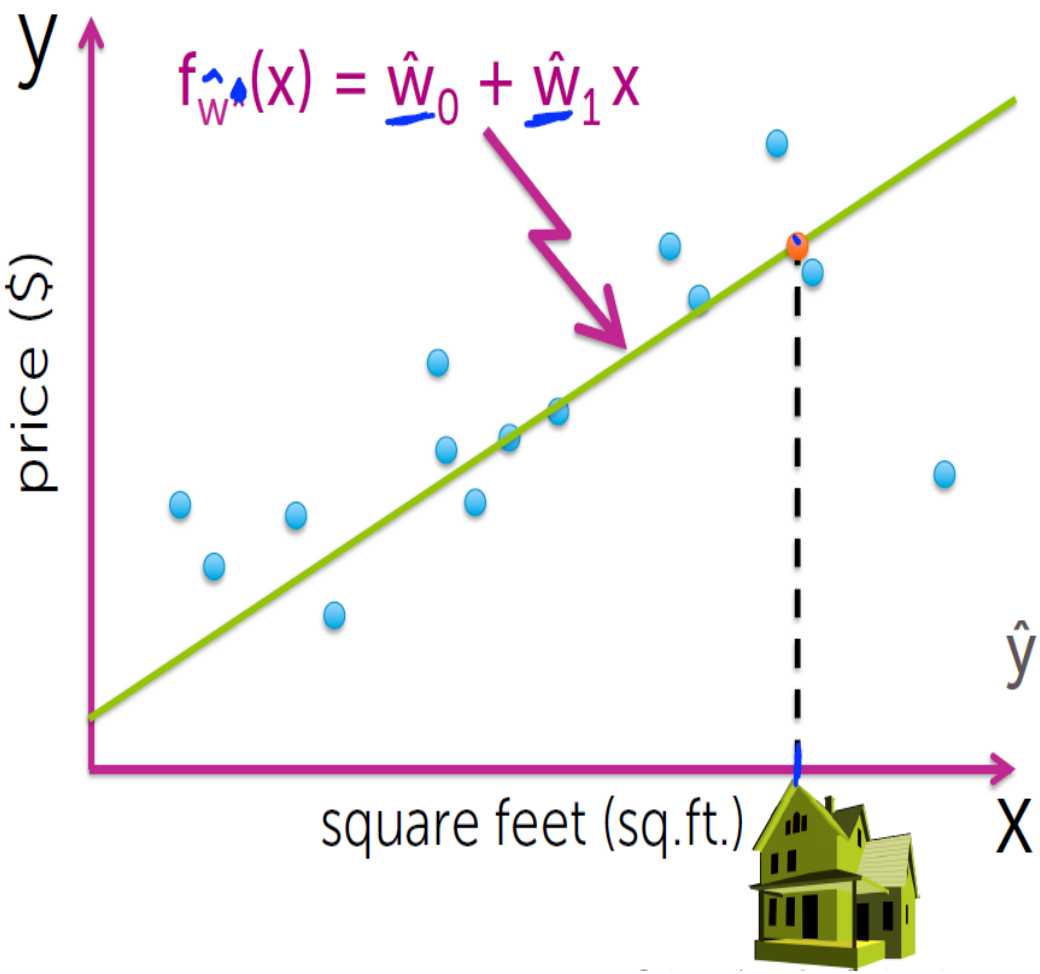
Find „best” line



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

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

Predicting your house price



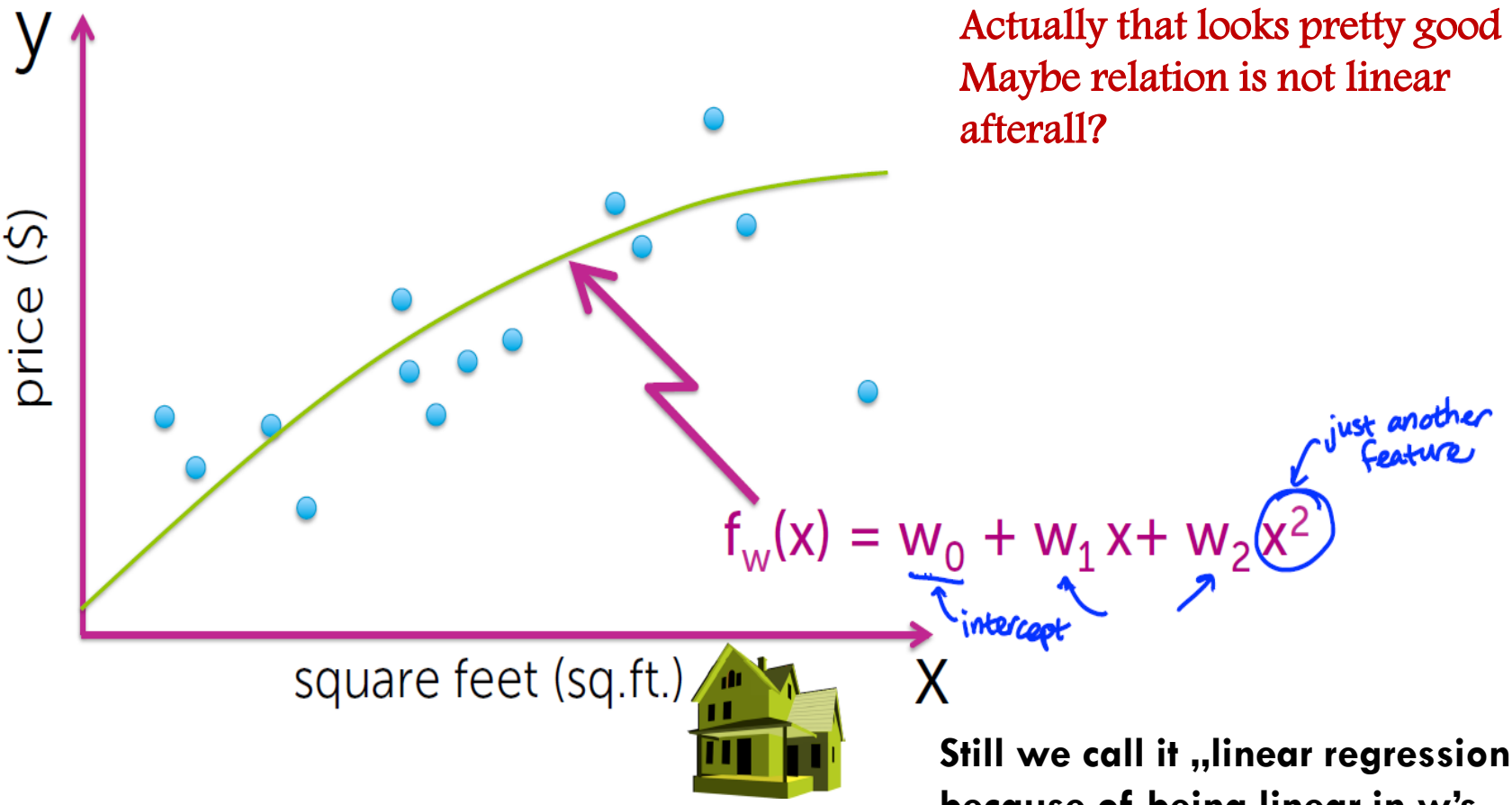
Q. What do you think?
Is it good analysis?

A. I am not sure that it has linear trend. Did you tried quadratic function?

Best guess of your house price:

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

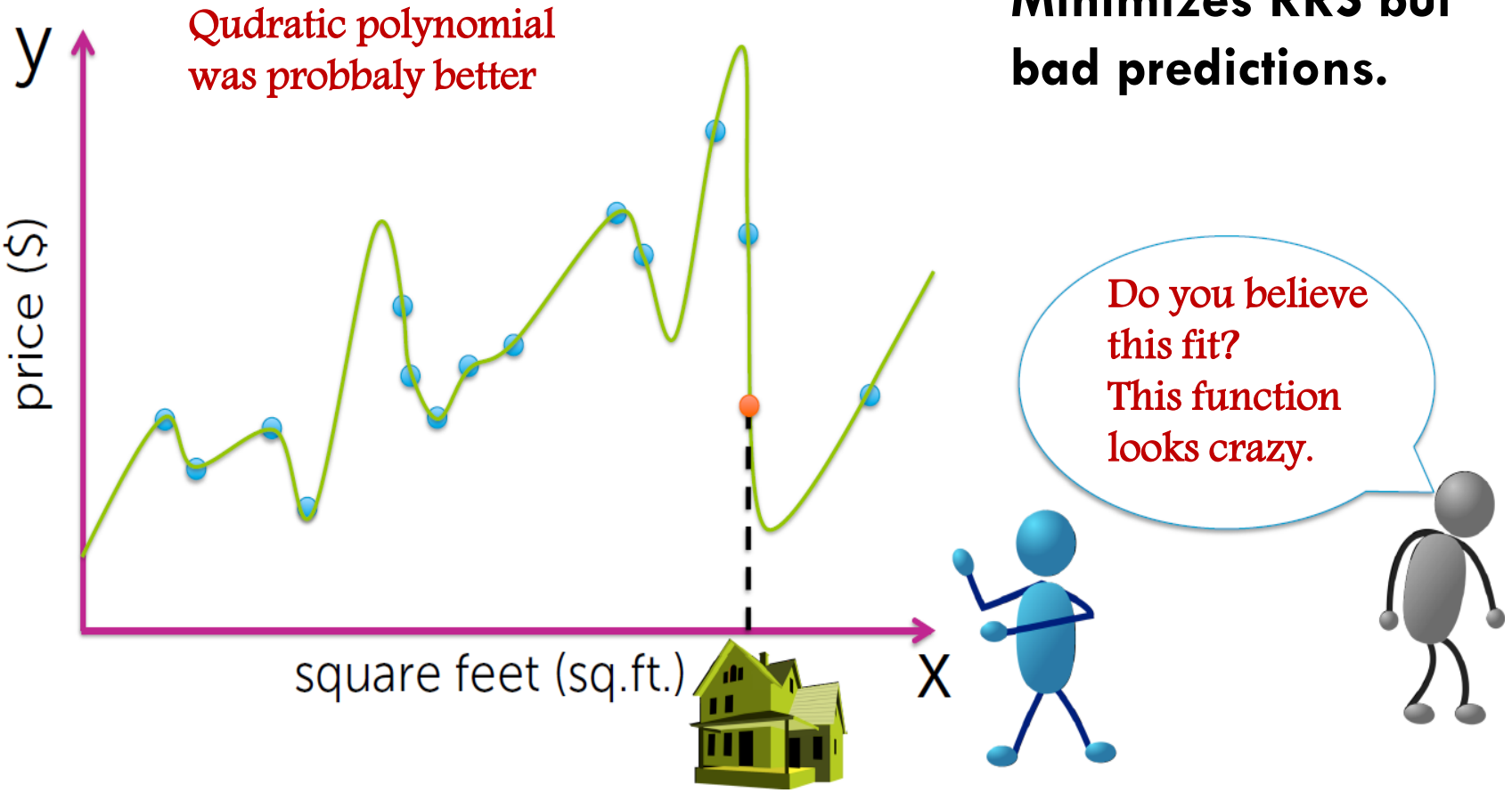
What about quadratic function?



Actually that looks pretty good
Maybe relation is not linear afterall?

Still we call it „linear regression" because of being linear in w's

Or even higher order polynomial?

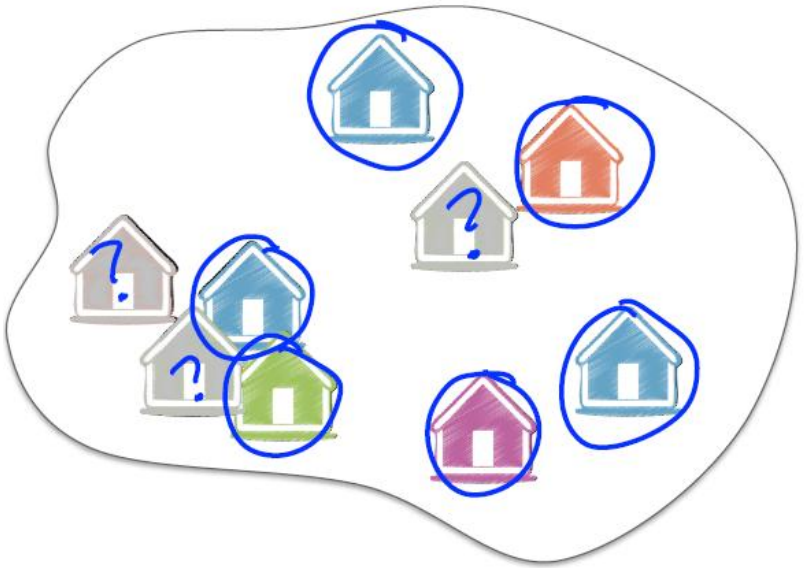


How to choose model order/complexity

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

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

We have to work with the data that we have

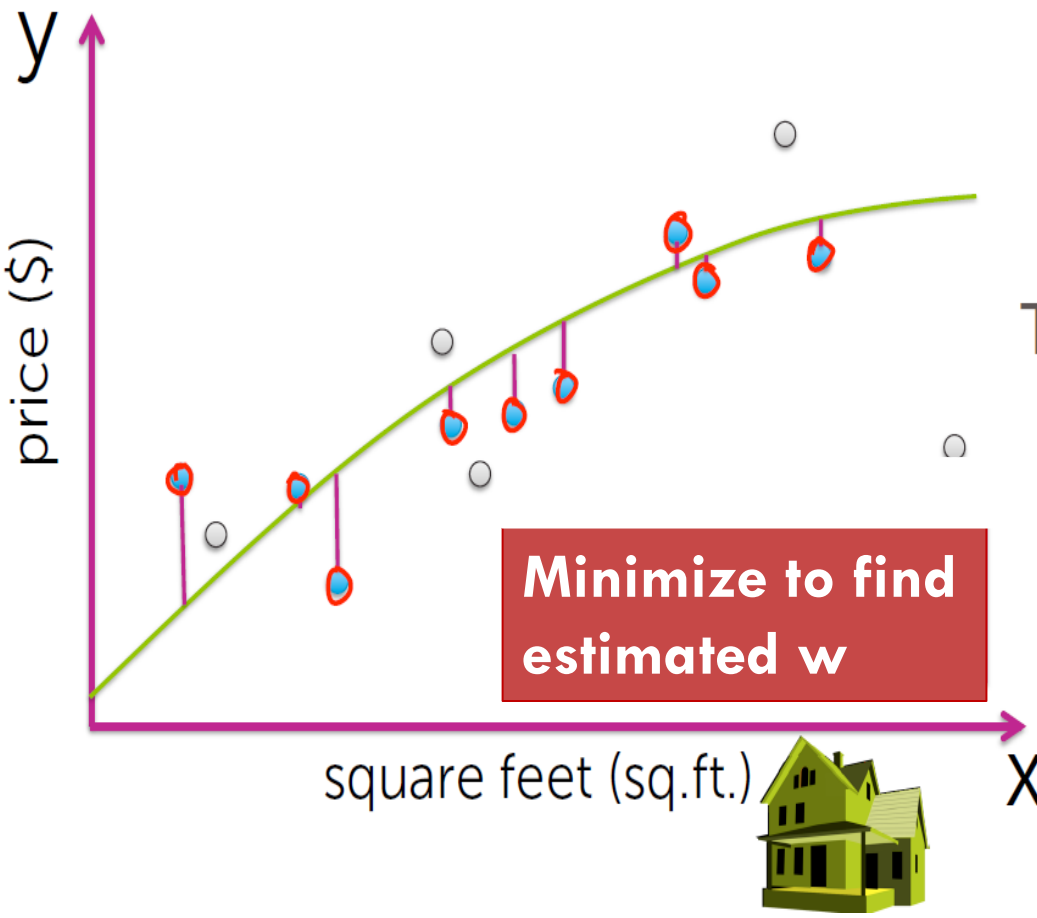


Training/test split

– training set

– test set

Training error



Training error (w) =

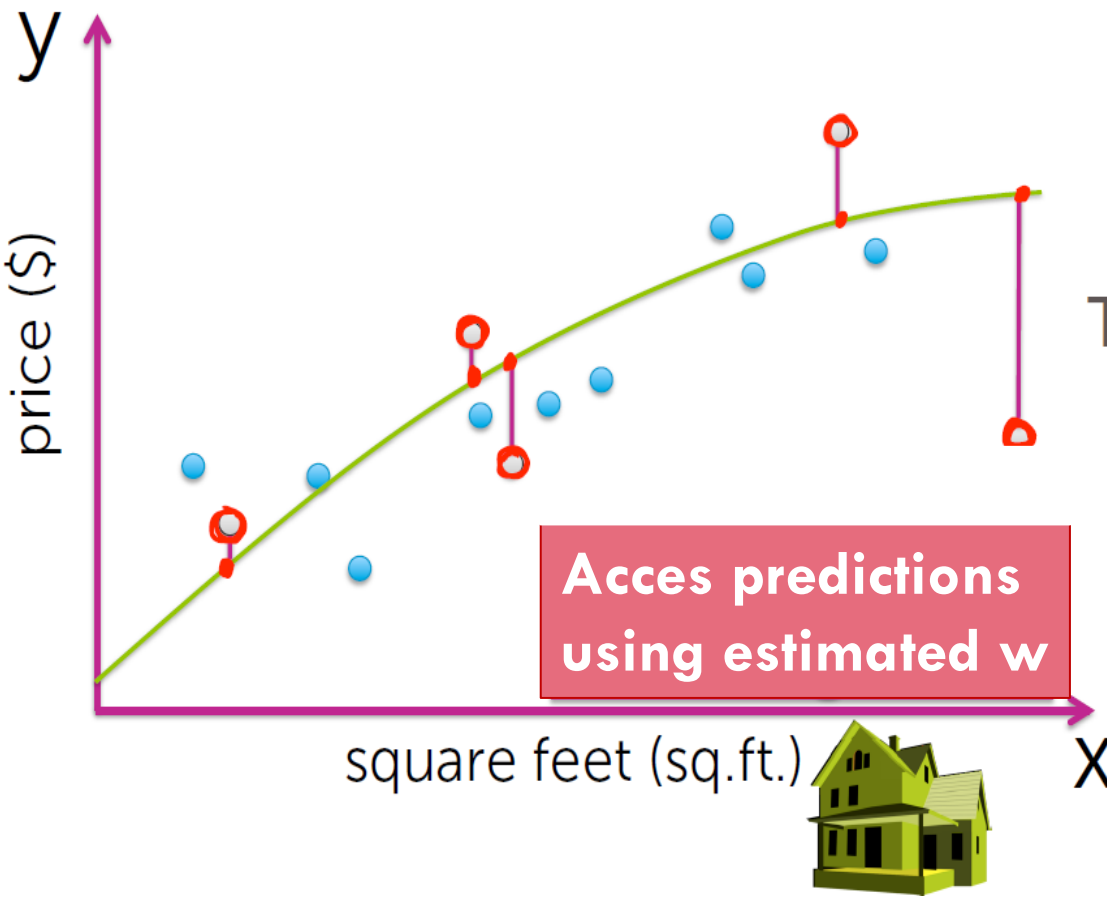
$(\$_{train\ 1} - f_w(sq.ft._{train\ 1}))^2$

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

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

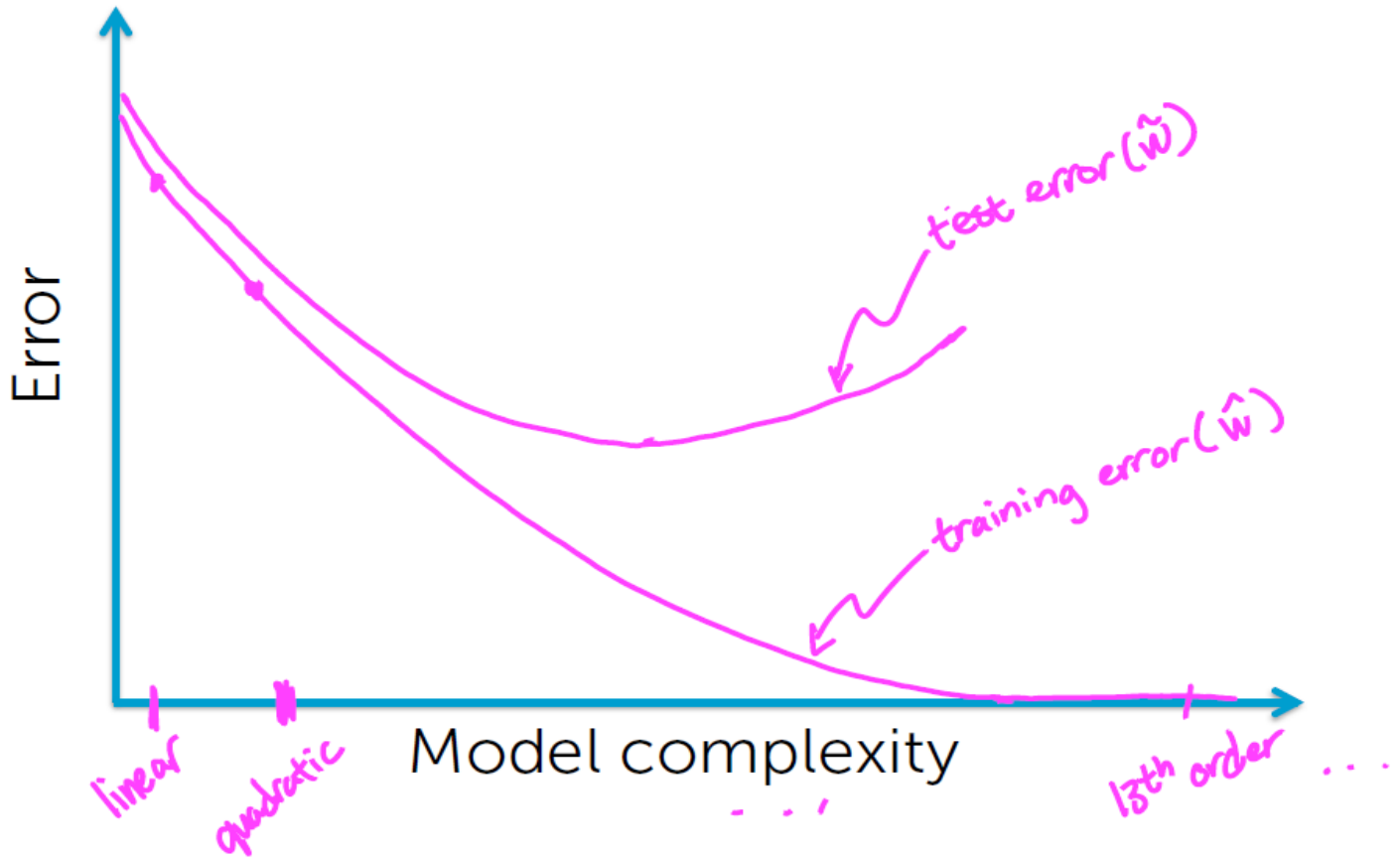
+ ... [include all training houses]

Test error

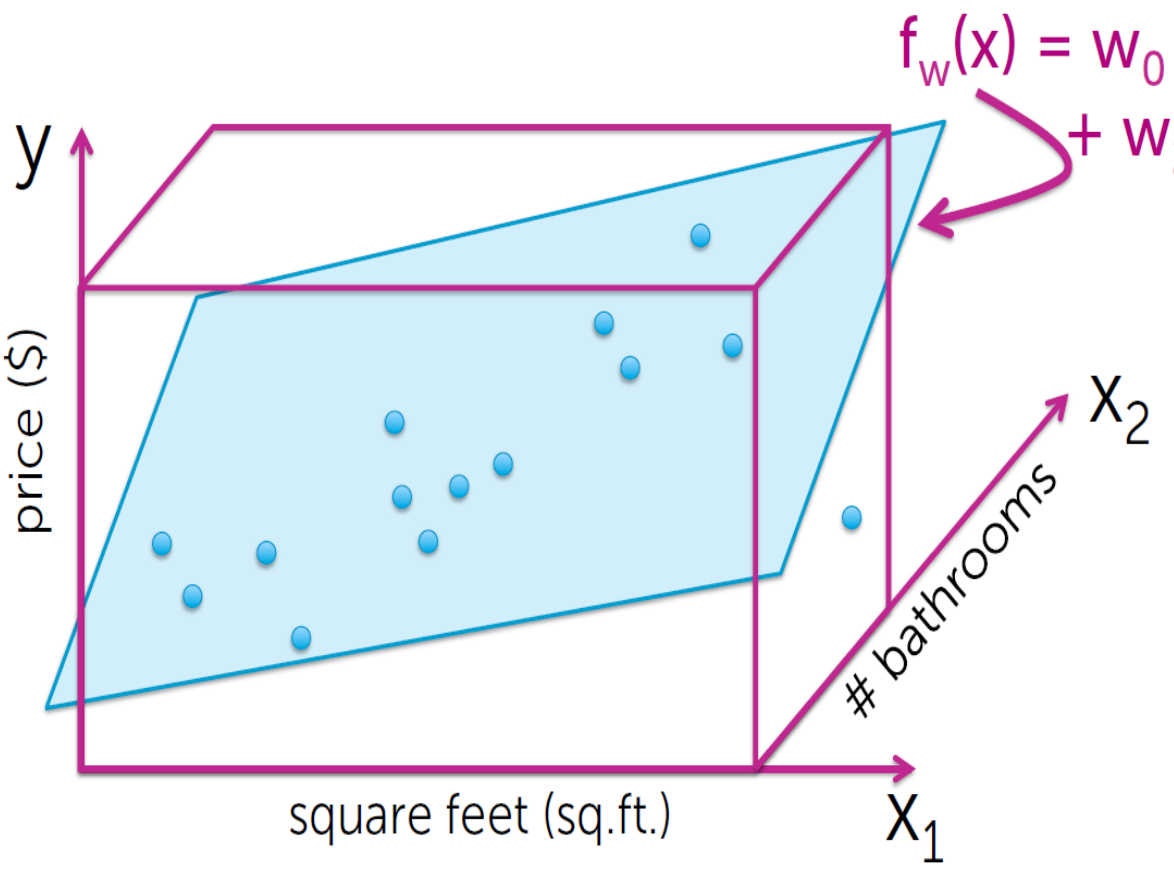


Test error ($\hat{w}$) =
$(\$_{test\ 1} - f_{\hat{w}}(sq.ft._{test\ 1}))^2$
$+ (\$_{test\ 2} - f_{\hat{w}}(sq.ft._{test\ 2}))^2$
$+ (\$_{test\ 3} - f_{\hat{w}}(sq.ft._{test\ 3}))^2$
+ ... [include all test houses]

Training/test curve



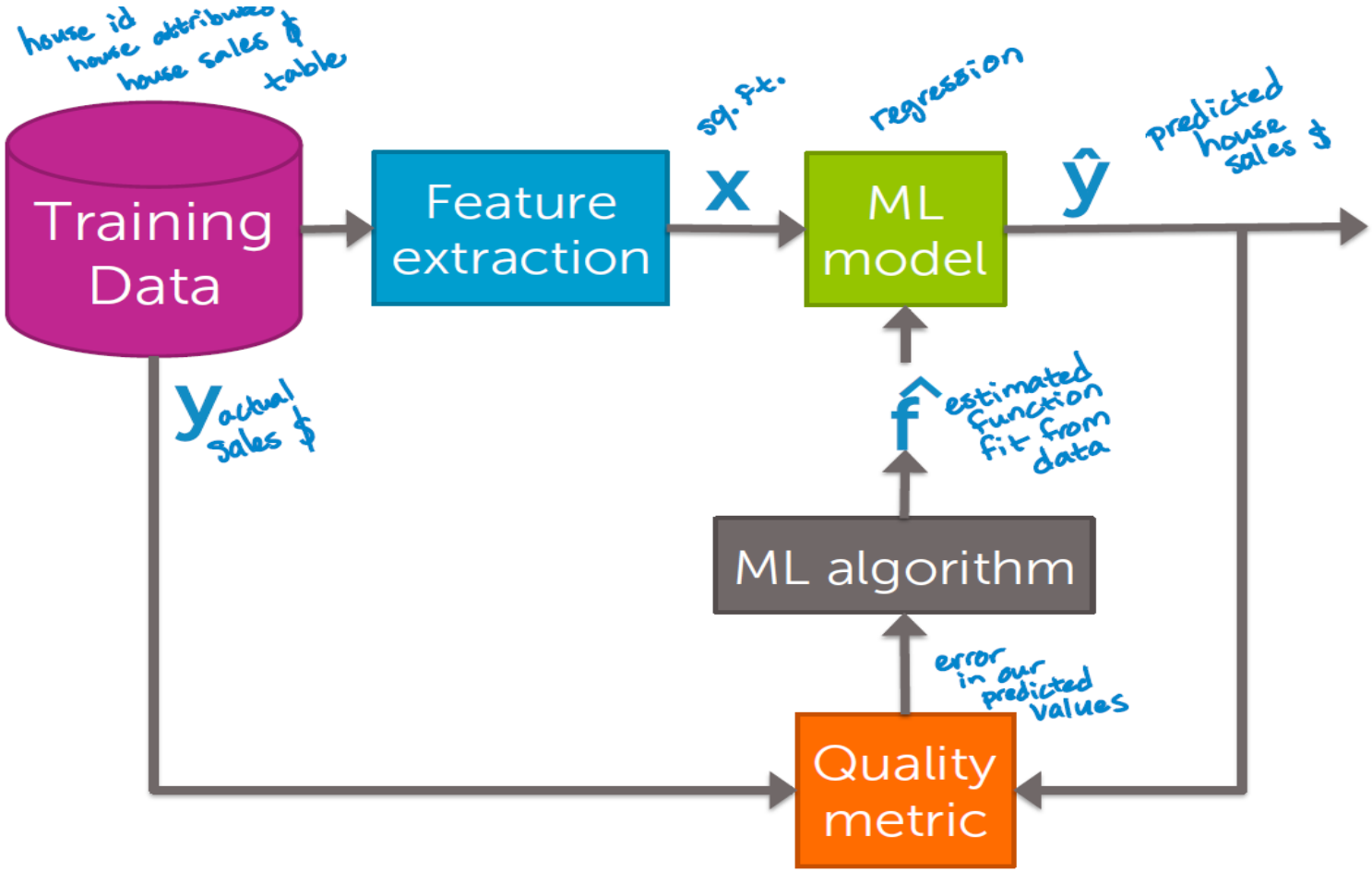
Add more features



- Possible choices:
  - Square feet
  - # bathrooms
  - # bedrooms
  - Lot size
  - Year built
  - …

Where do we stop?

Regression ML block



We will discuss how to

- 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