INTRODUCTION TO DATA SCIENCE

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

15/10/2019

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**
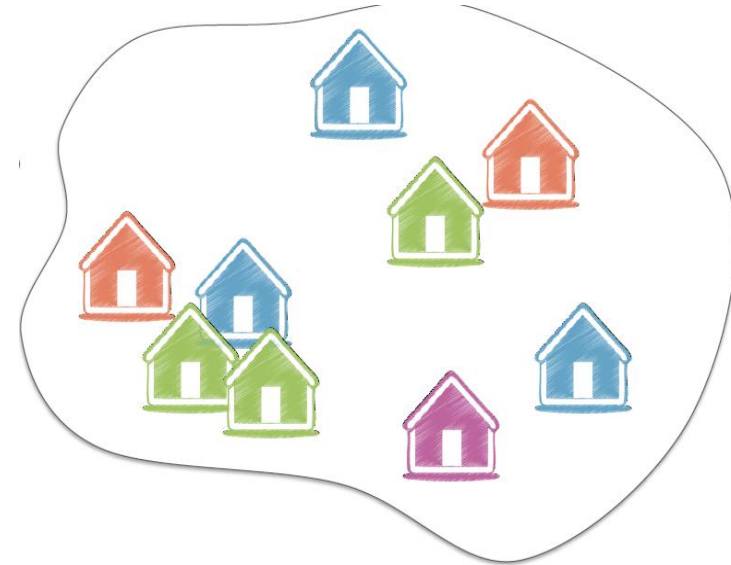   - **Nearest neighbor and kernel regression**

15/10/2019

How much is my house worth

- **Predicting value of the house**
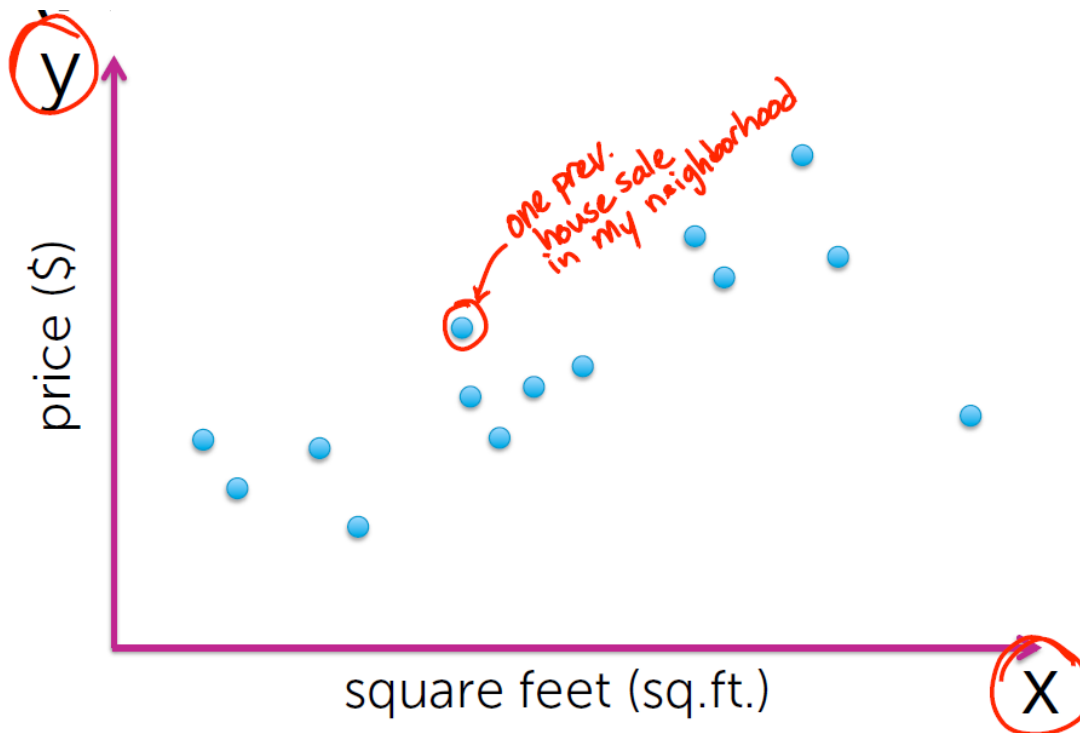
How much
is worth?

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

15/10/2019

Naive: plot recent house sales

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

price ($)

y

one prev. house sale in my neighborhood
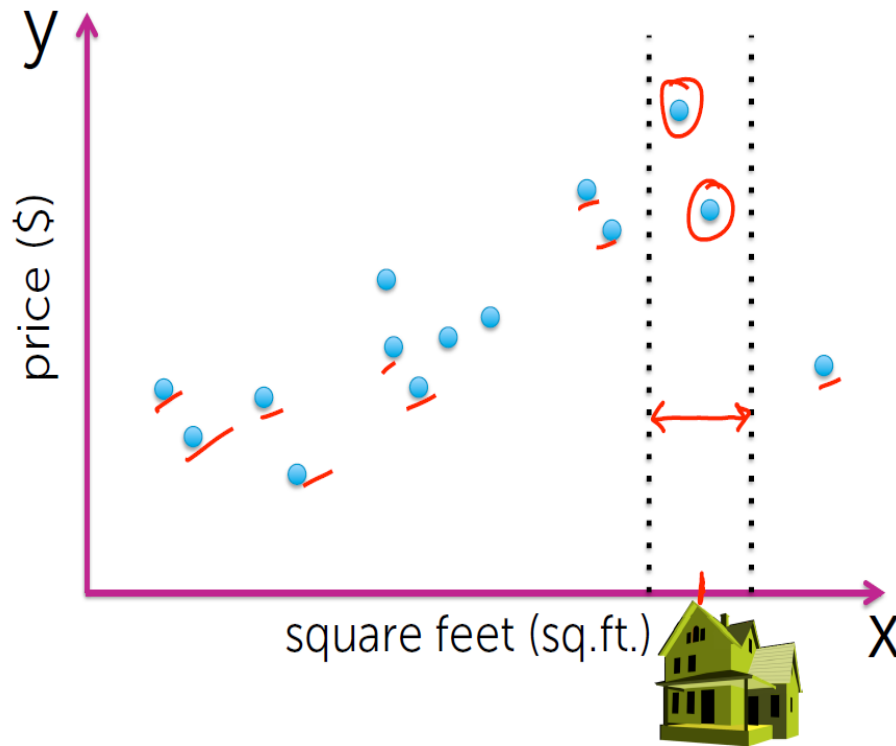
square feet (sq.ft.)

x

Terminology:

x – feature, covariate, or predictor

y – observation or response

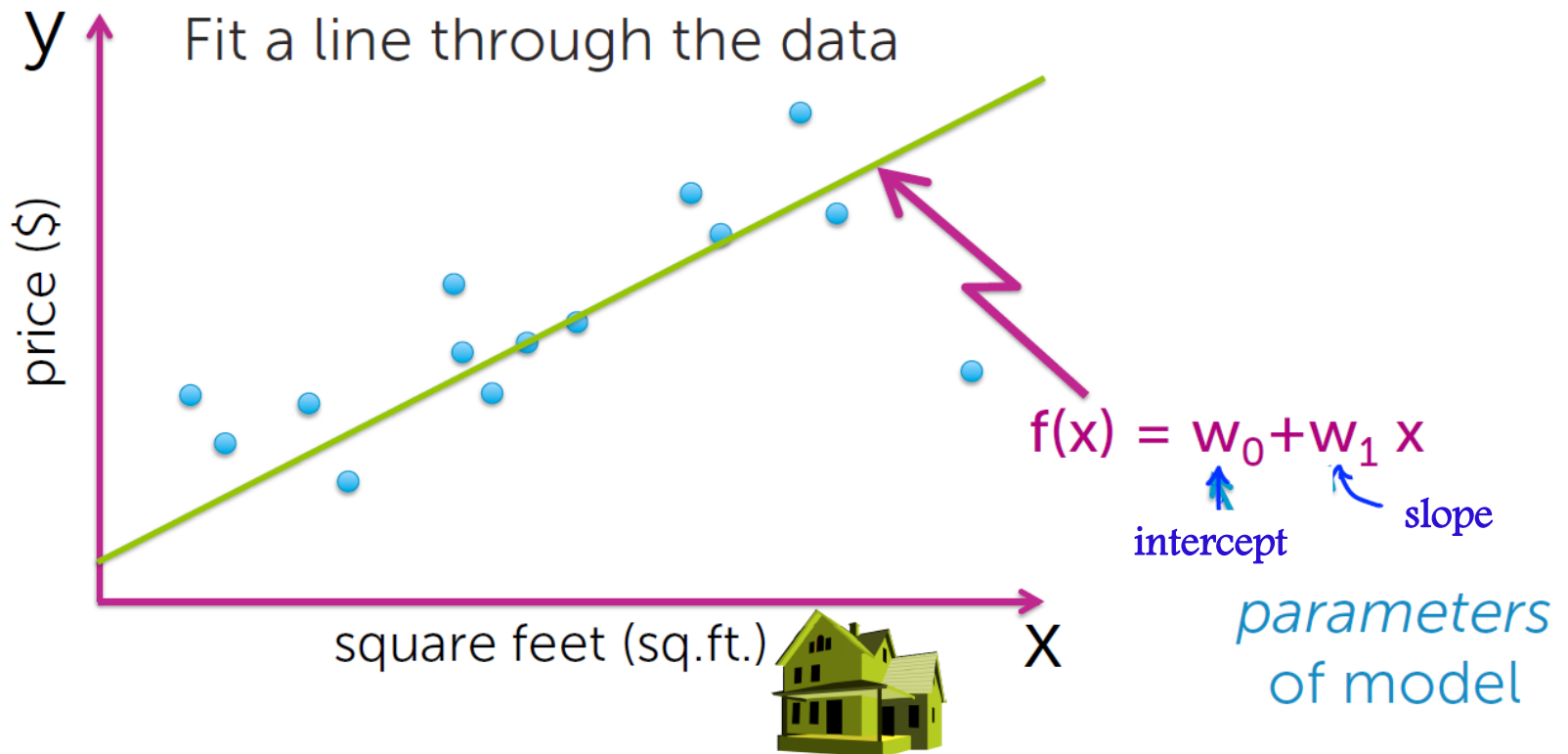15/10/2019

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.

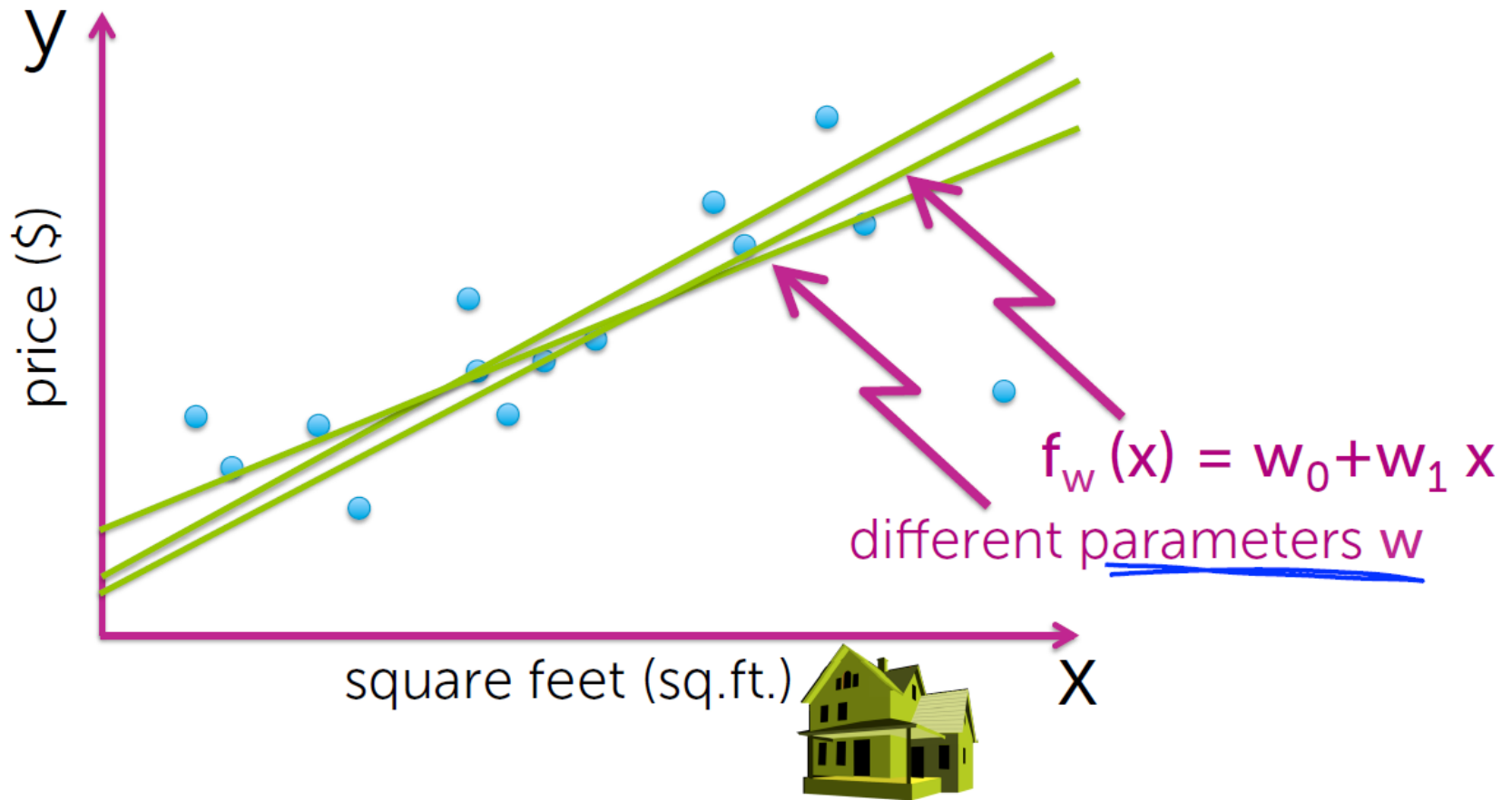15/10/2019

Use a linear regression model

y — price ($)

Fit a line through the data

$f(x) = w_0 + w_1 x$

intercept   slope

parameters of model

square feet (sq.ft.)   x

10

Which line?

$$f_w(x) = w_0 + w_1 x$$

different parameters w

15/10/2019

Defining a cost of a given line
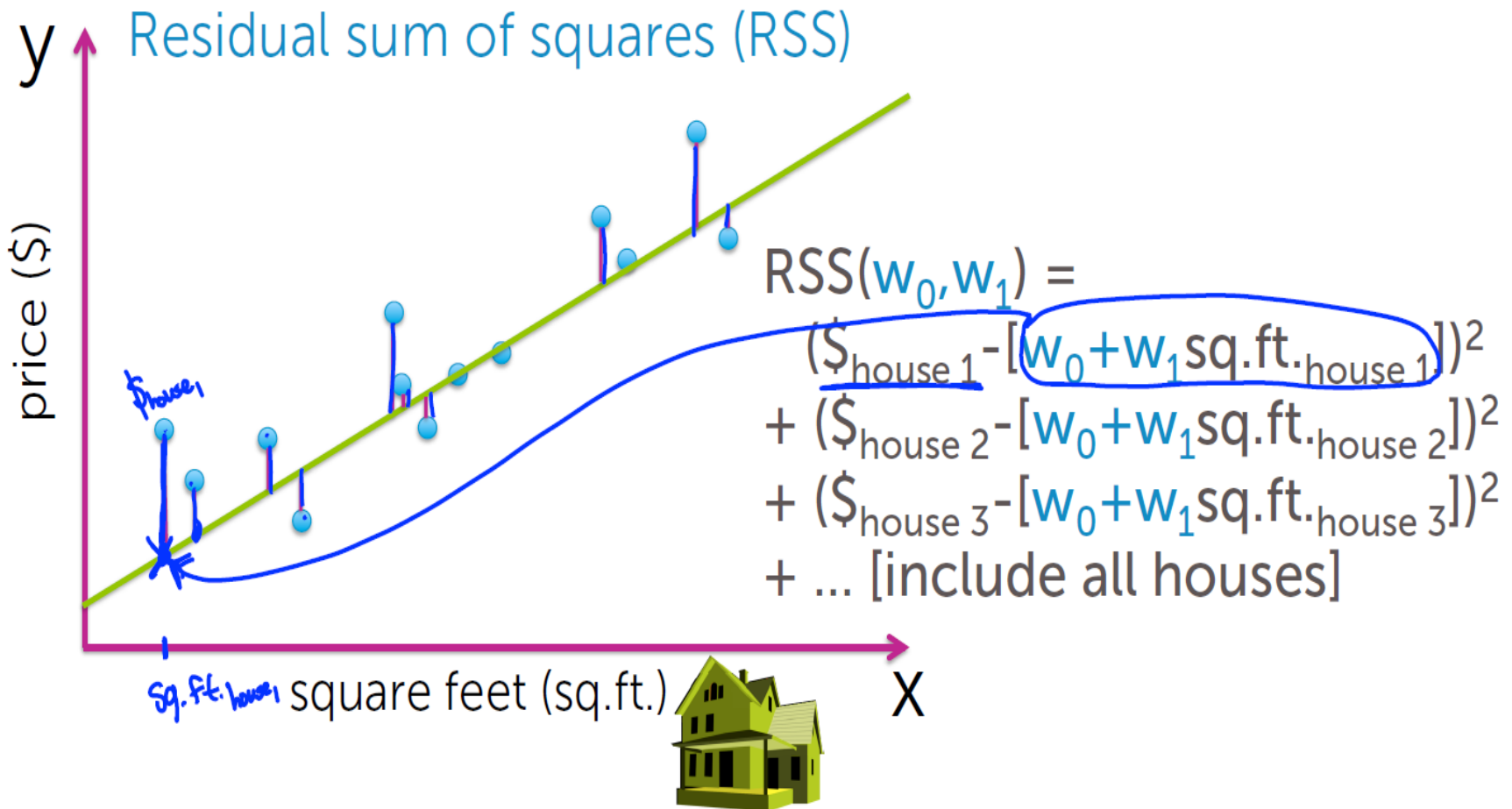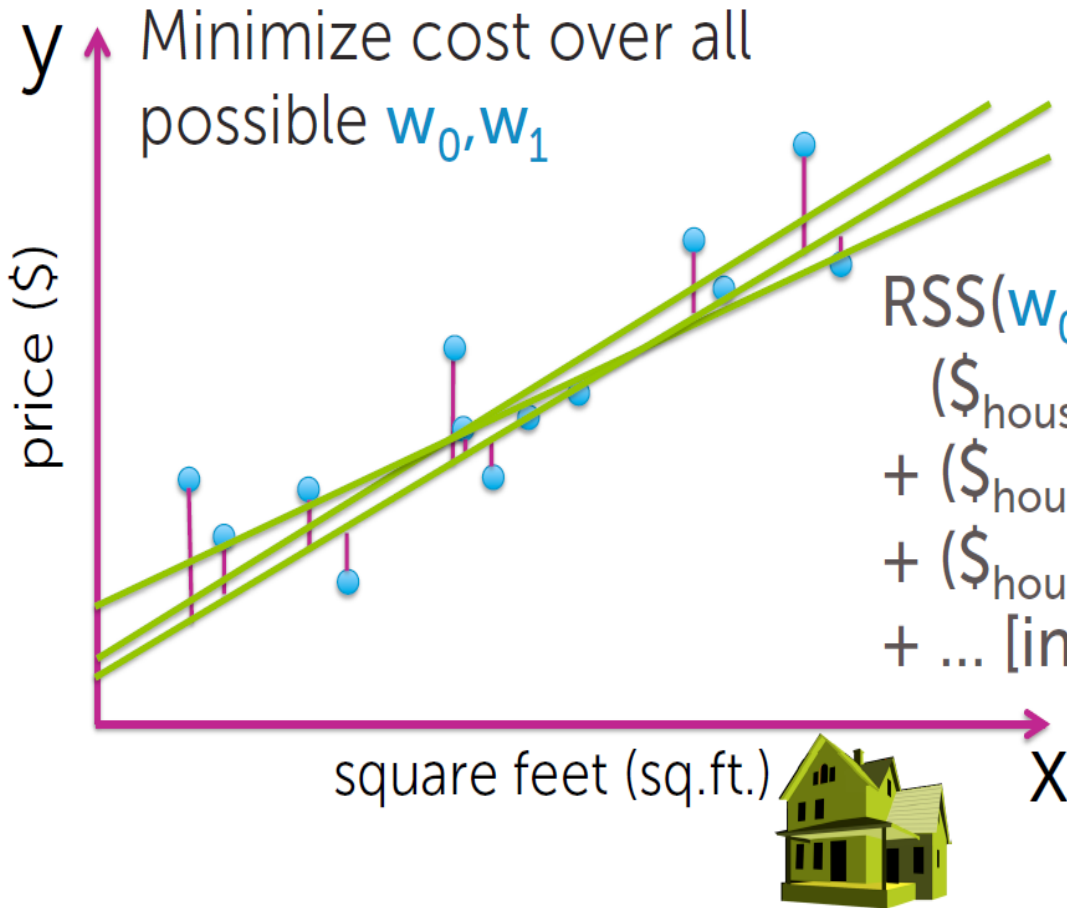
Residual sum of squares (RSS)

y

price ($)

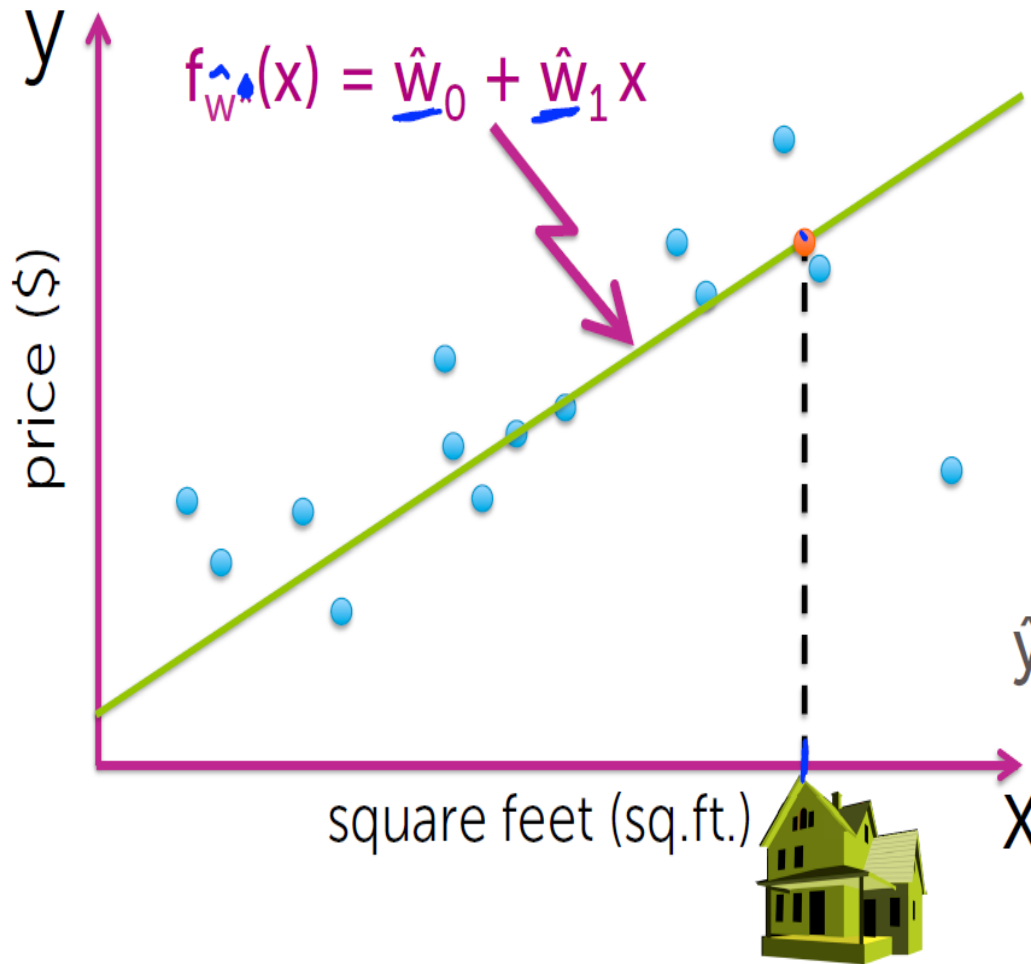$\text{RSS}(w_0, w_1) =$
$(\$_{\text{house 1}} - [w_0 + w_1 \text{sq.ft.}_{\text{house 1}}])^2$
$+ (\$_{\text{house 2}} - [w_0 + w_1 \text{sq.ft.}_{\text{house 2}}])^2$
$+ (\$_{\text{house 3}} - [w_0 + w_1 \text{sq.ft.}_{\text{house 3}}])^2$
$+ \ldots [\text{include all houses}]$

square feet (sq.ft.)

X

15/10/2019

Find „best" line

Minimize cost over all possible w_0, w_1

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

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

15/10/2019

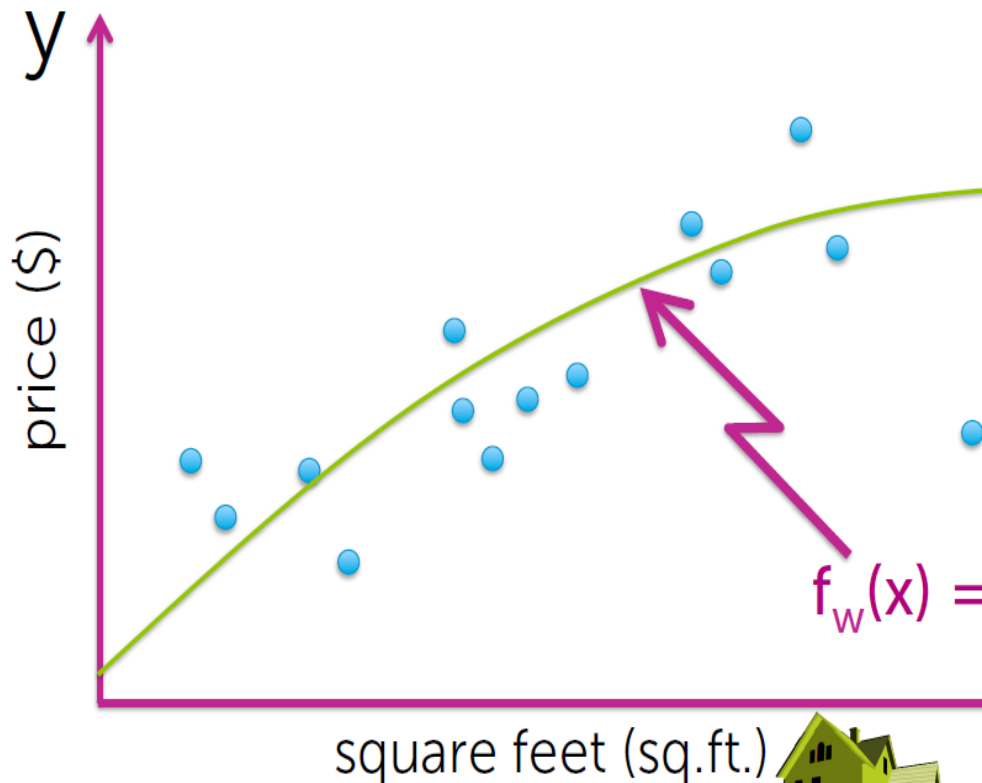# Predicting your house price

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

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

$$f_{\hat{W}}(x) = \hat{w}_0 + \hat{w}_1 x$$

Best guess of your house price:

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

price ($)

square feet (sq.ft.)

15

What about quadratic function?

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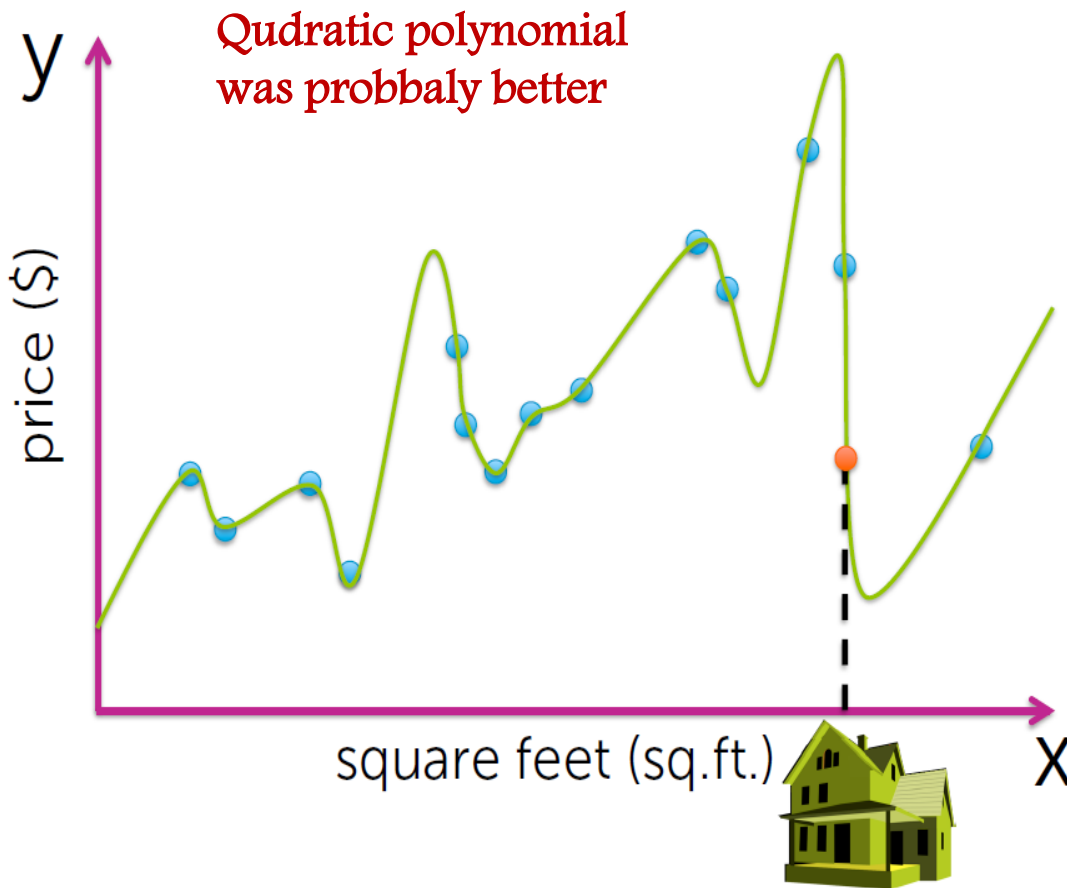afterall?

just another
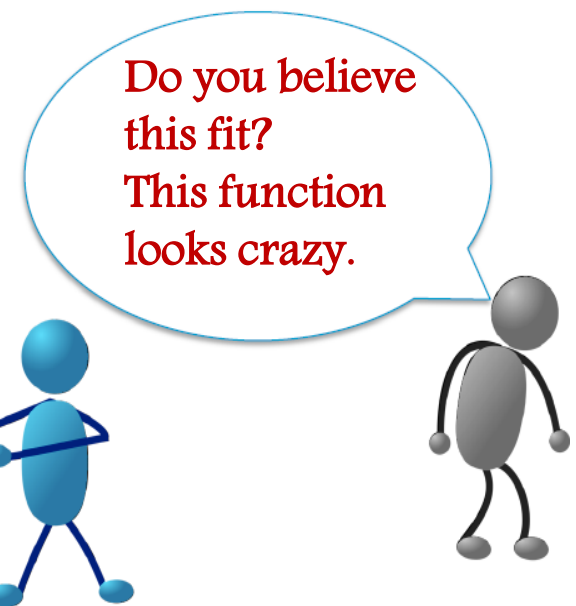feature

$$f_w(x) = w_0 + w_1 x + w_2 x^2$$

intercept

square feet (sq.ft.)

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

15/10/2019

Or even higher order polynomial?

Qudratic polynomial was probbaly better

Minimizes RRS but bad predictions.

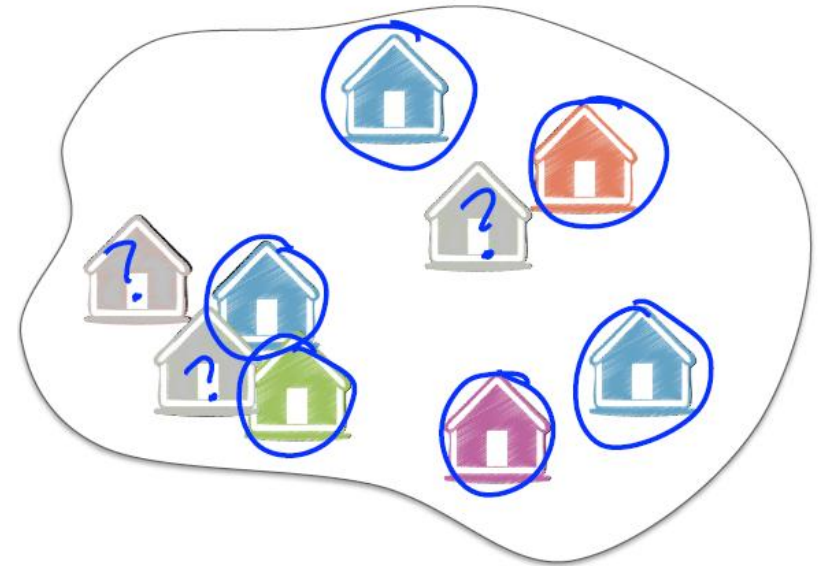Do you believe this fit?
This function looks crazy.

15/10/2019

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

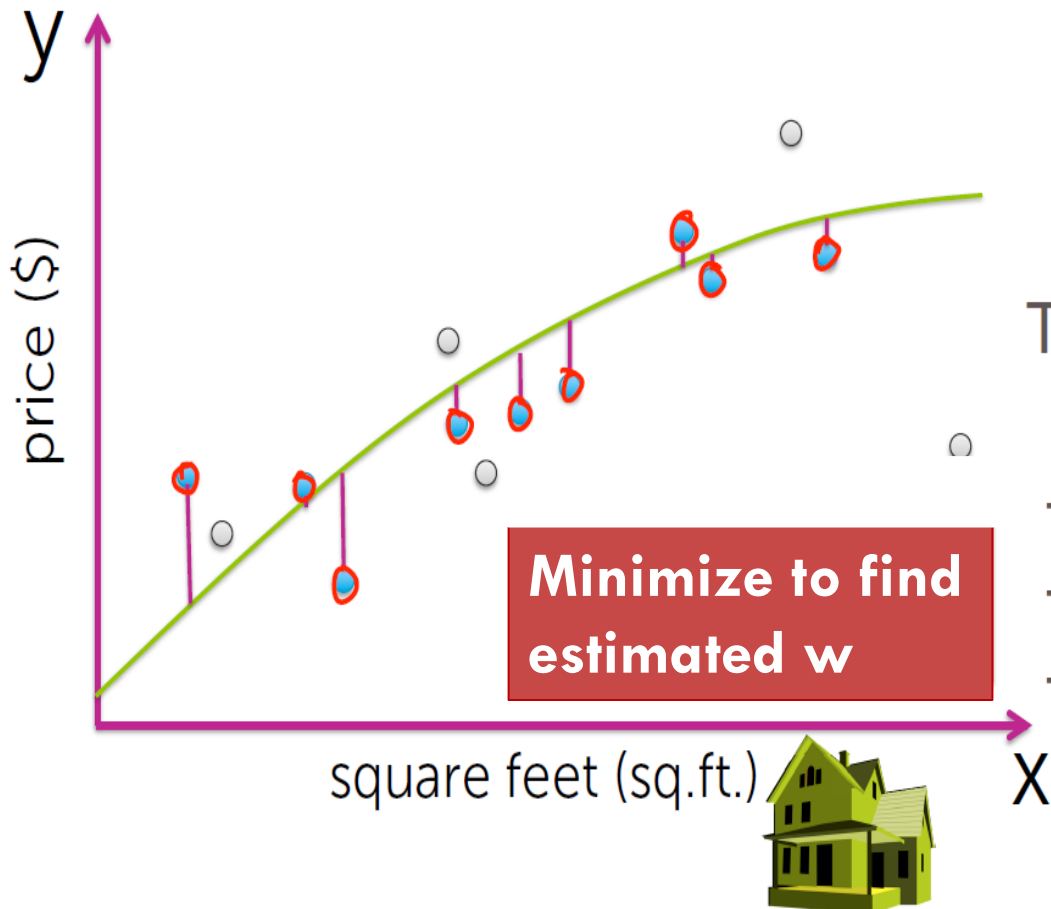We have to work with the data that we have

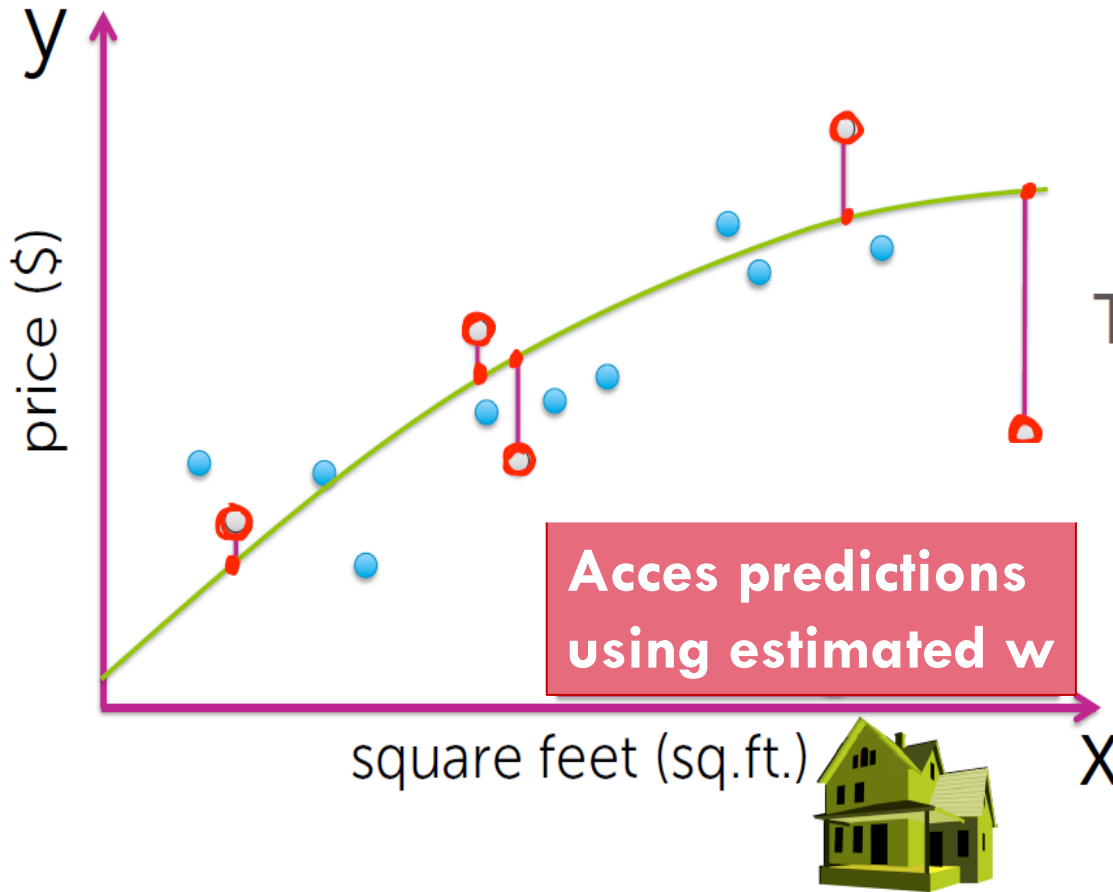# Training/test split

− 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]

Minimize to find estimated w

15/10/2019

Test error

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

Acces predictions using estimated w

y

price ($)

square feet (sq.ft.)   x

15/10/2019

Training/test curve

Error

test error($\tilde{w}$)

training error($\hat{w}$)

Model complexity

linear   quadratic   ...   13th order   ...

15/10/2019

Add more features

$$f_w(x) = w_0 + w_1 \, \text{sq.ft.} + w_2 \, \#\text{bath}$$
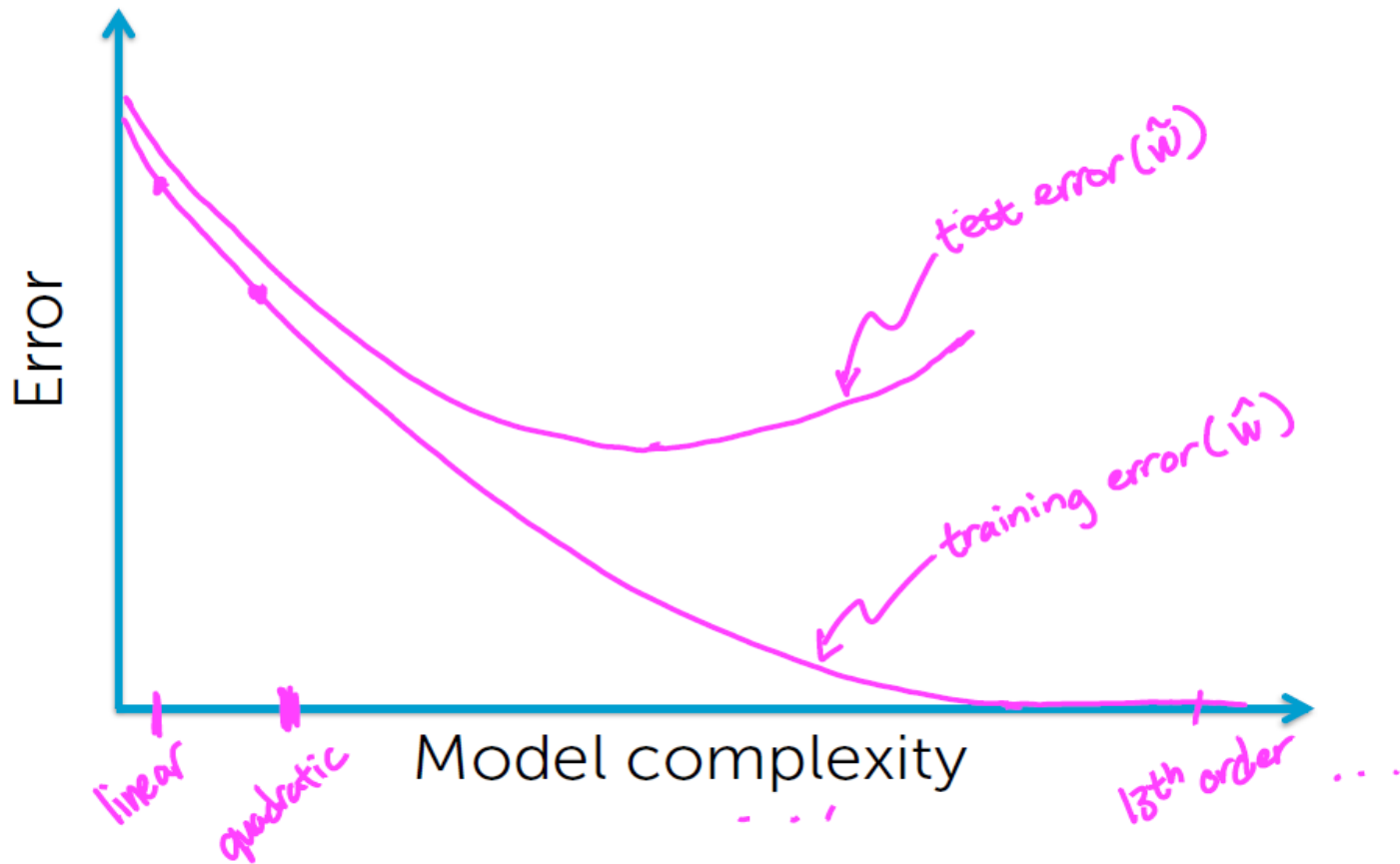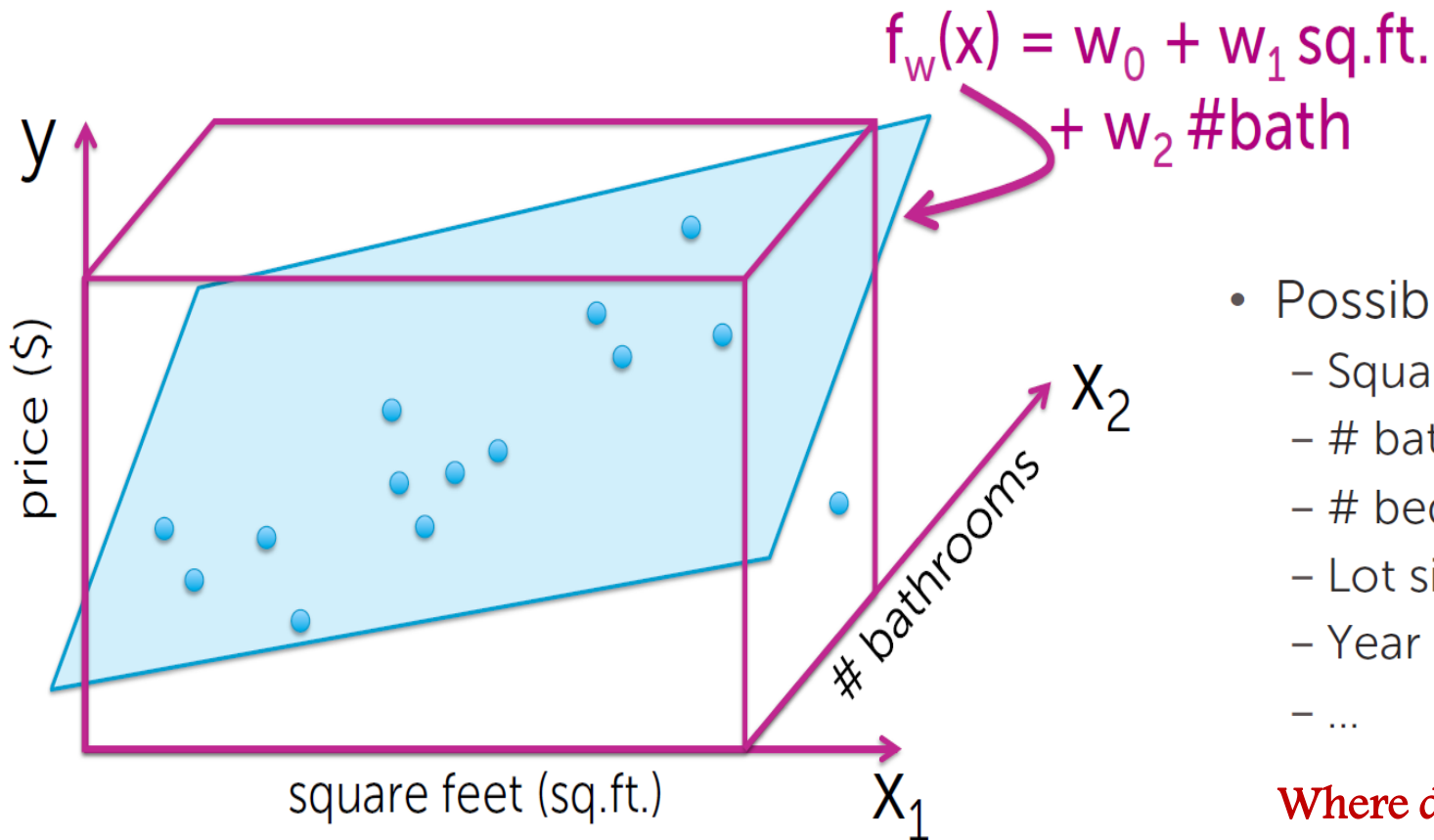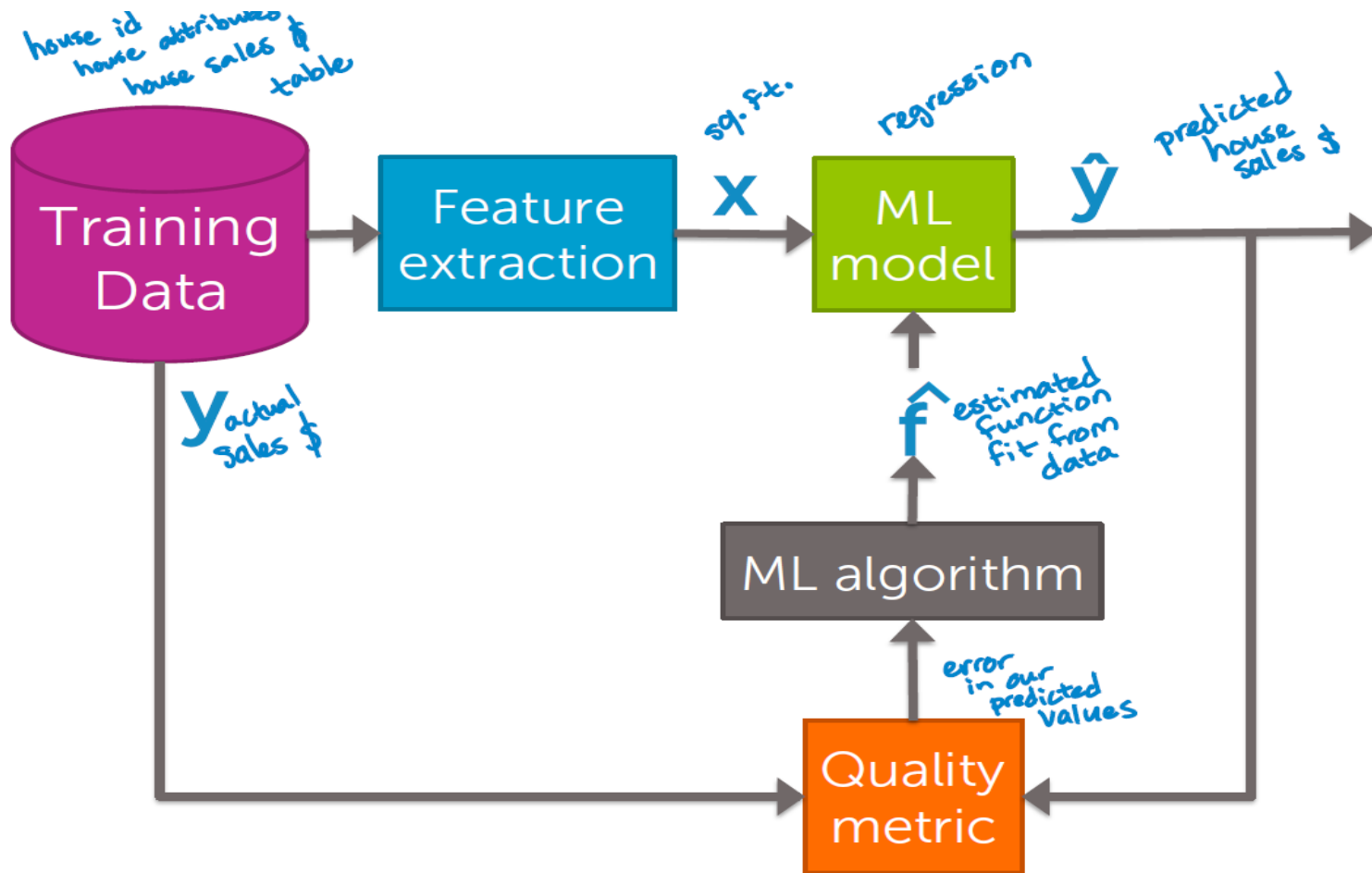
- Possible choices:
  - Square feet
  - # bathrooms
  - # bedrooms
  - Lot size
  - Year built
  - ...

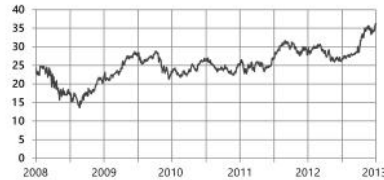**Where do we stop?**

15/10/2019

Regression ML block
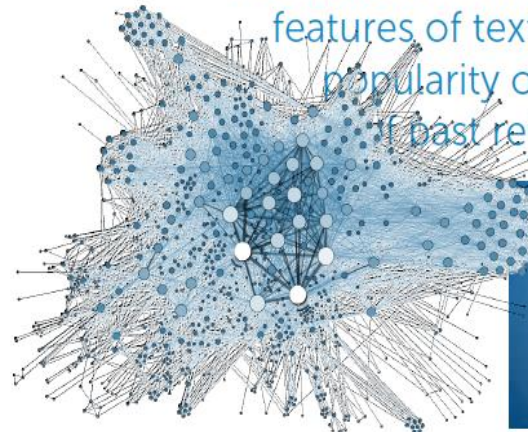
Other applications

Stock predictions

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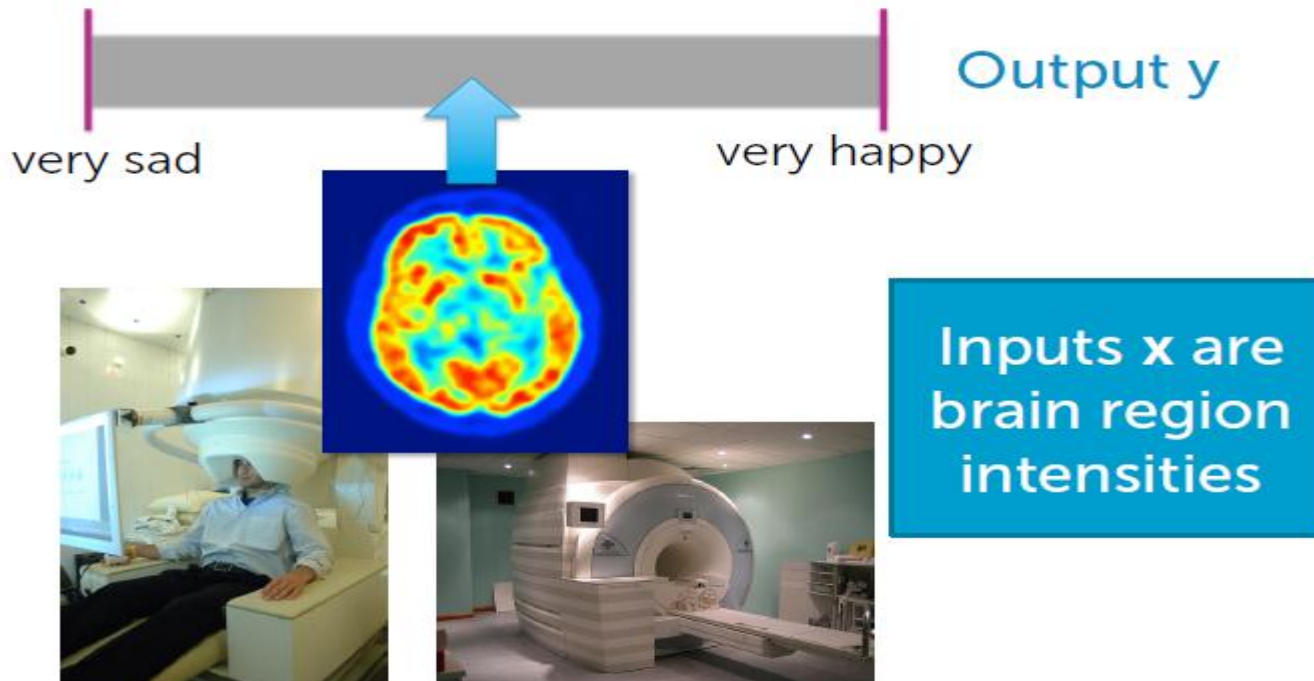 - News events
  - Related commodities

Tweed popularity

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

15/10/2019

Other applications

Reading your mind

Output y

very sad   very happy

Inputs x are brain region intensities

15/10/2019

We discussed 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

15/10/2019