

INTRODUCTION TO DATA SCIENCE

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

9/10/2018

WFAiS UJ, Informatyka Stosowana
II stopień studiów

Regression for predictions

2

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

How much is my house worth

3

□ Predicting value of the house



How much
is worth?

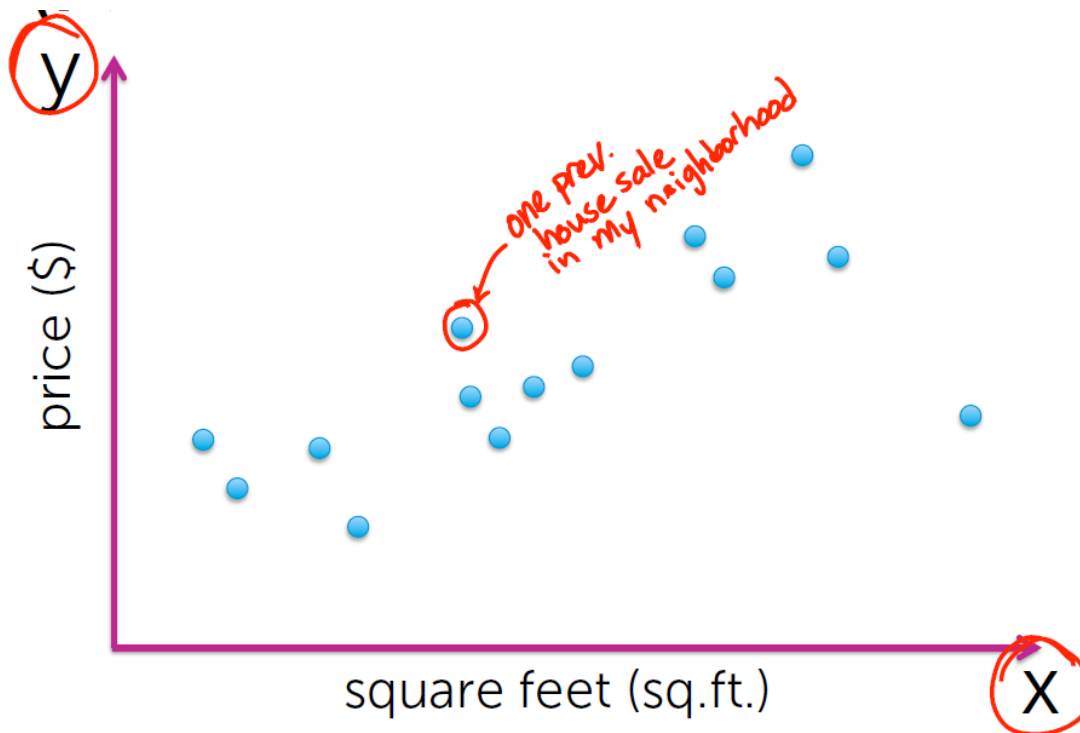


Lets look at the recent
sales in the neighborhood.
How much did they sell for?
What do that houses look like?

Naive: plot recent house sales

4

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



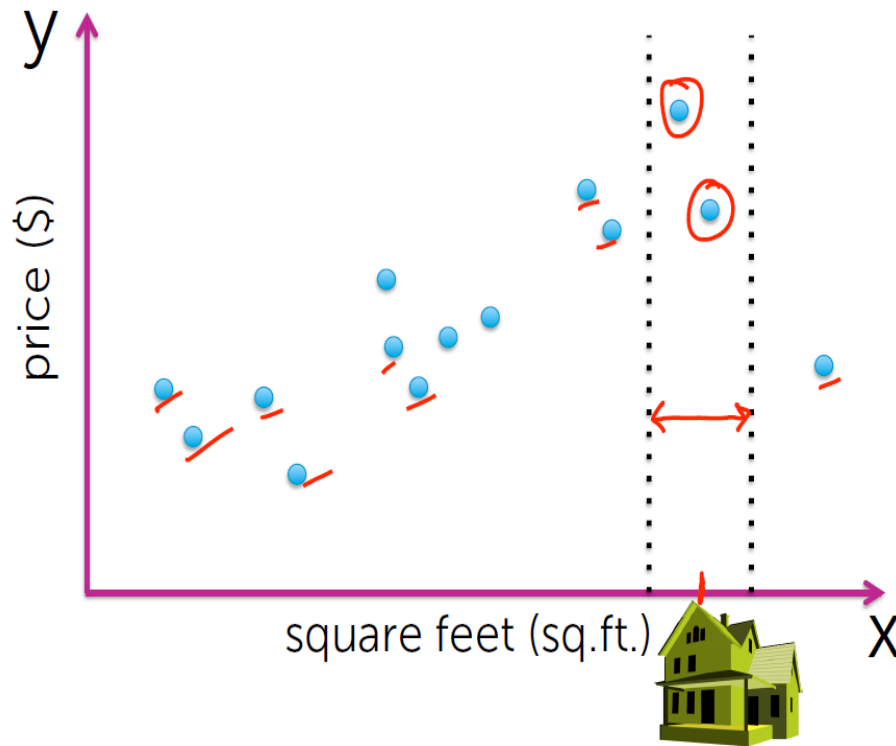
Terminology:

x – feature, covariate, or predictor

y – observation or response

Predict by prizes of similar houses

5



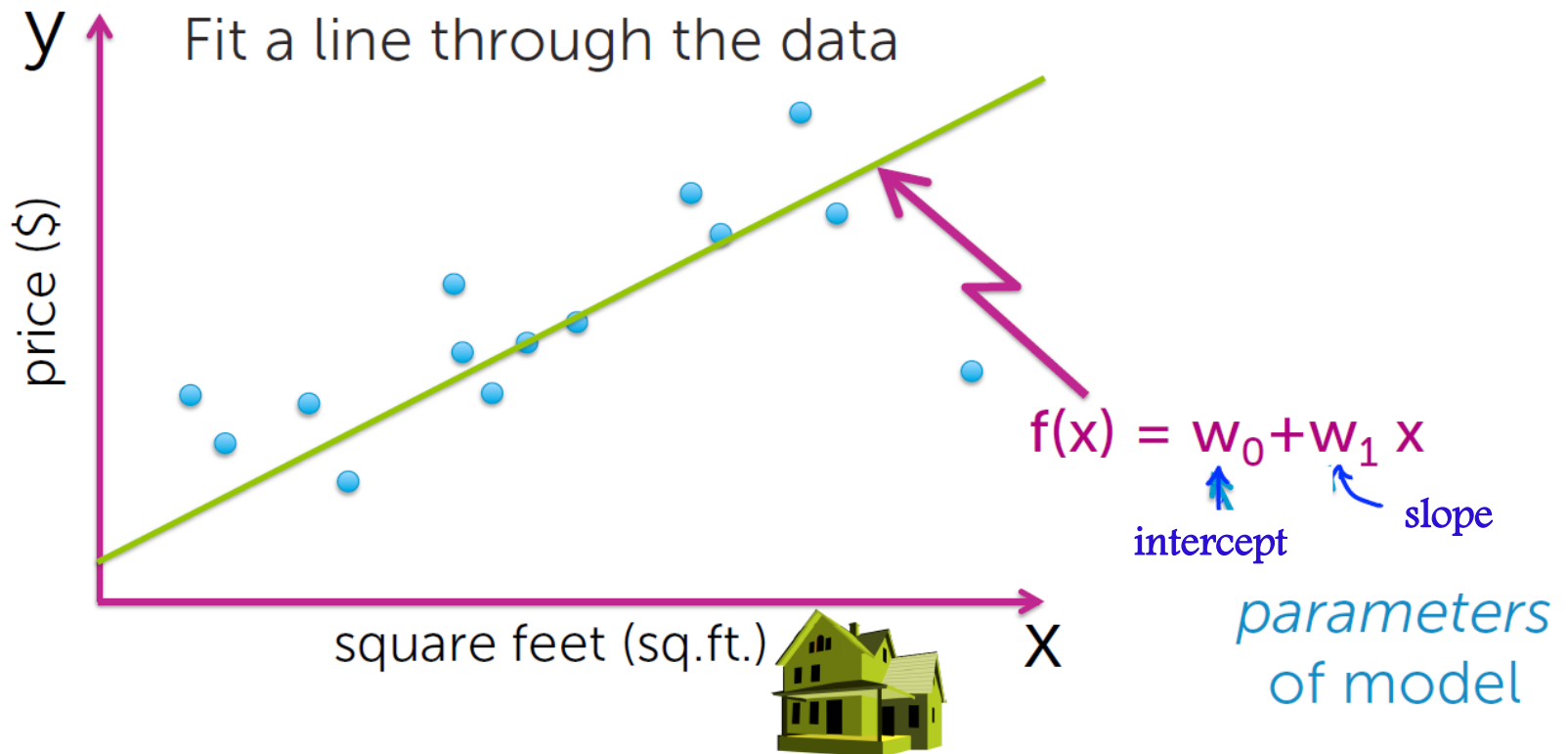
- Look at average price in range
- **Still only 2 houses!**
- Throwing out info from all other sales

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

Linear regression: a model based relation

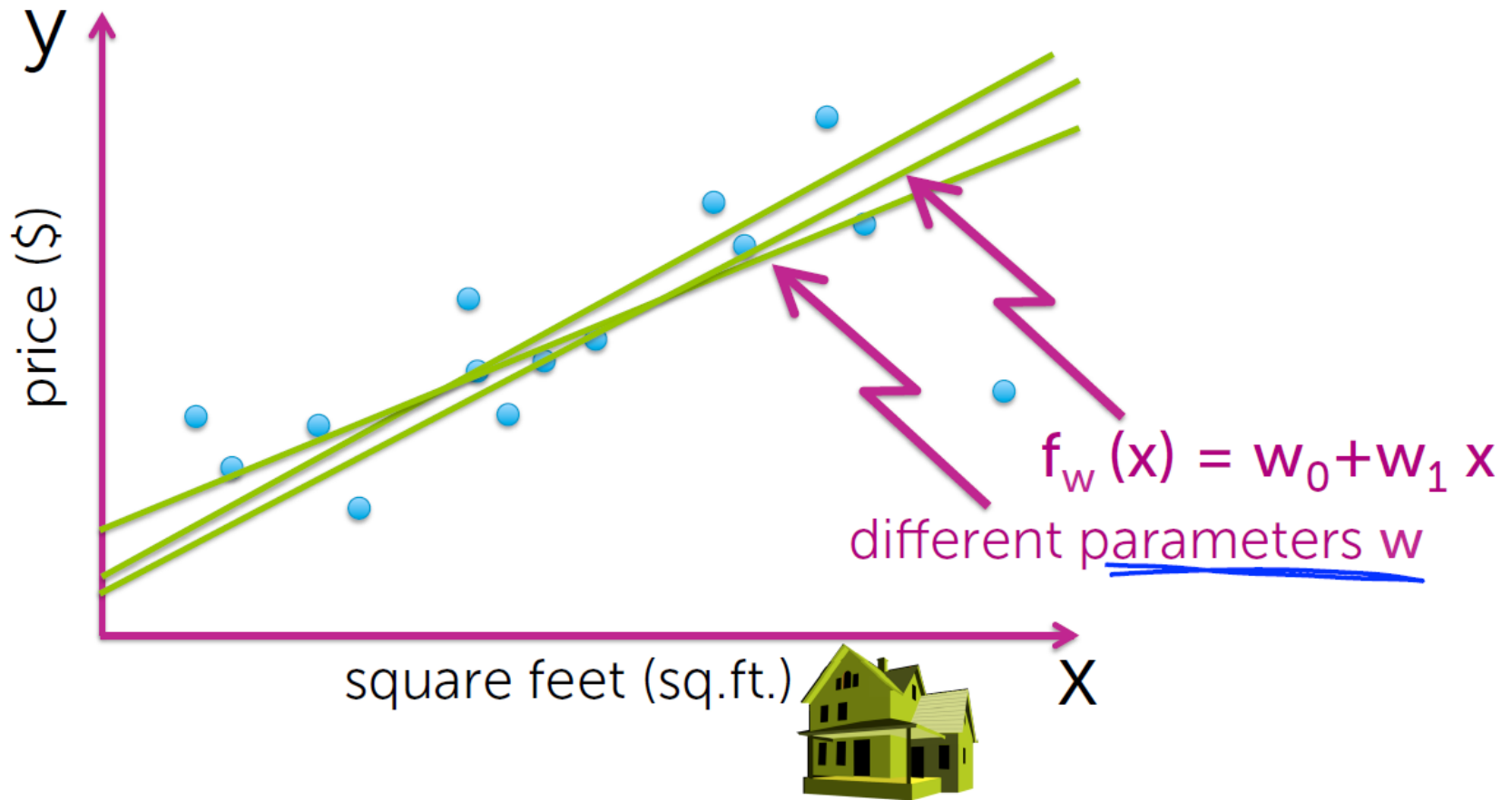
6

Use a linear regression model



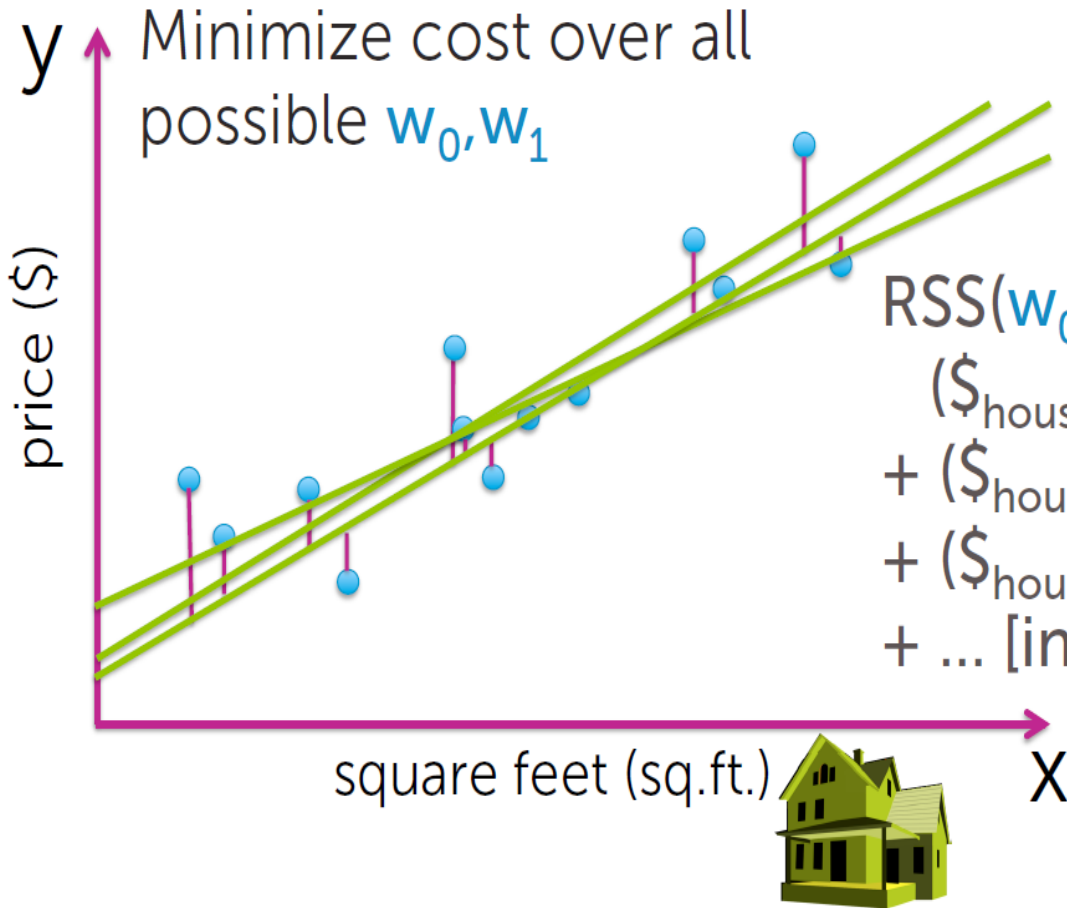
Which line?

7



Find „best” line

9



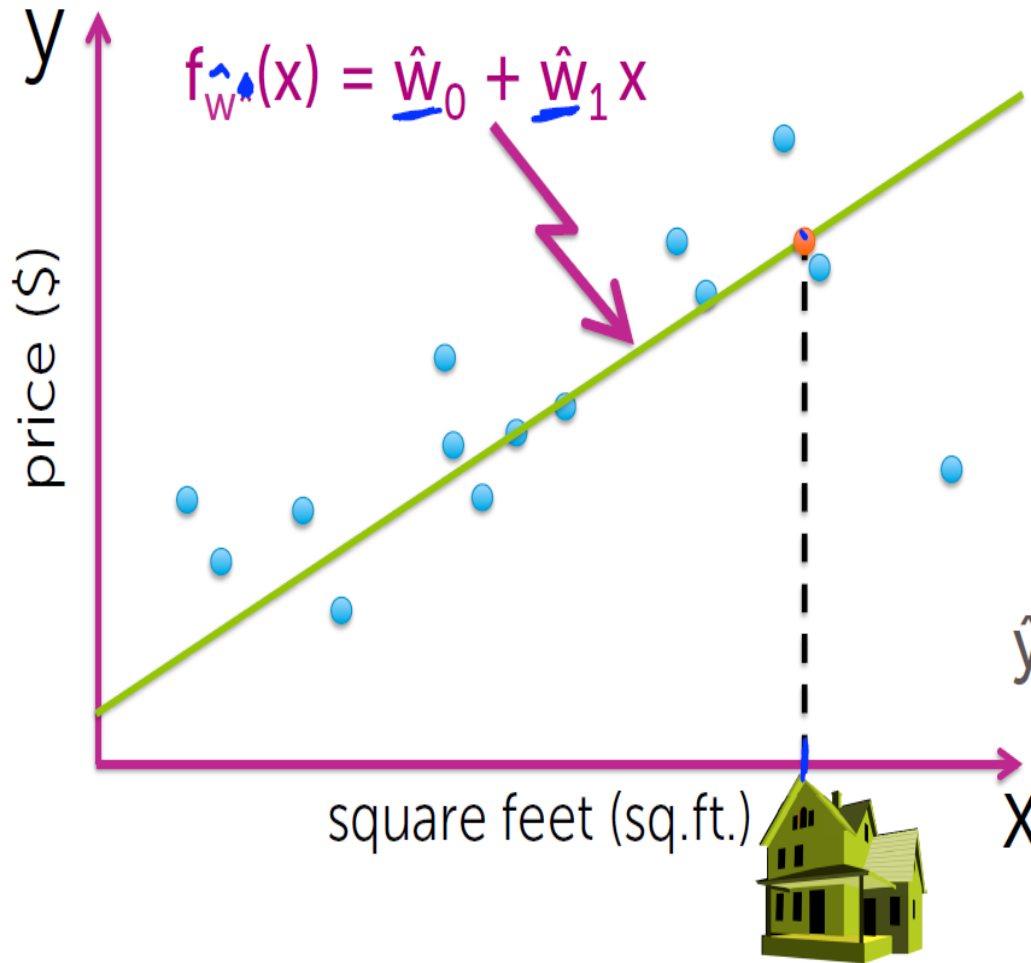
$$\begin{aligned} \text{RSS}(w_0, w_1) = & (\$_{\text{house 1}} - [w_0 + w_1 \text{sq.ft.}_{\text{house 1}}])^2 \\ & + (\$_{\text{house 2}} - [w_0 + w_1 \text{sq.ft.}_{\text{house 2}}])^2 \\ & + (\$_{\text{house 3}} - [w_0 + w_1 \text{sq.ft.}_{\text{house 3}}])^2 \\ & + \dots \text{ [include all houses]}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

↓

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

Predicting your house price

10



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

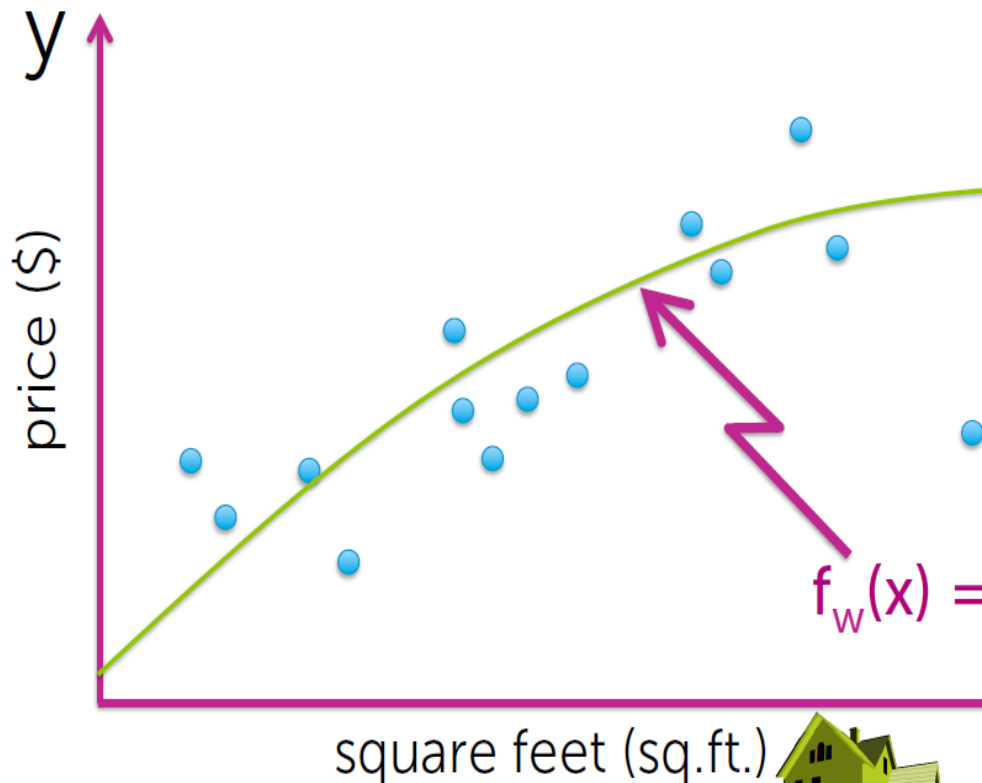
A. I am not sure that it has
linear trend. Did you tried
quadratic function?

Best guess of your
house price:

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

What about quadratic function?

11



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

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

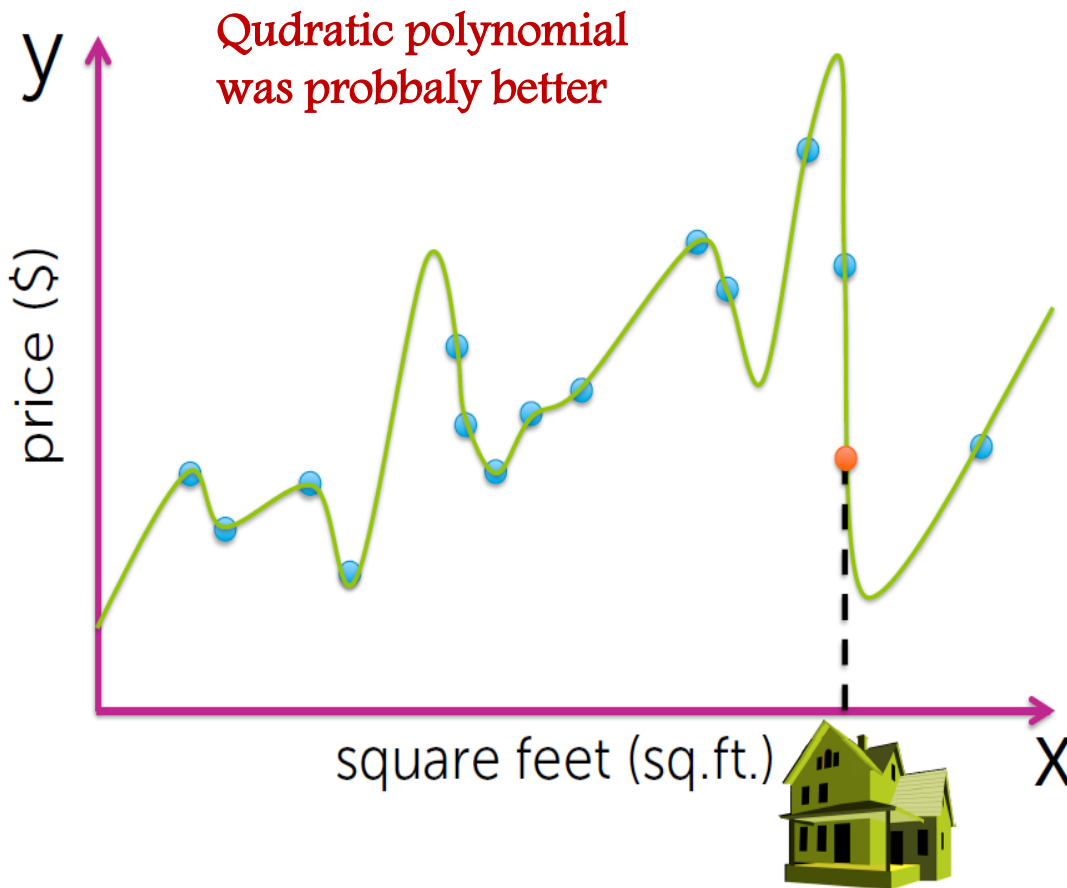
intercept

just another
feature

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

Or even higher order polynomial?

12



**Minimizes RRS but
bad predictions.**

Do you believe
this fit?
This function
looks crazy.

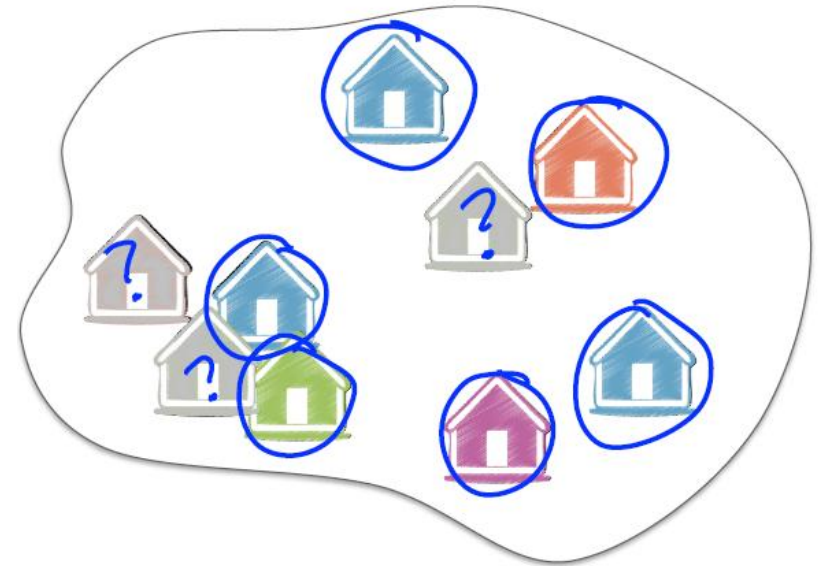


How to choose model order/complexity

13

- 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

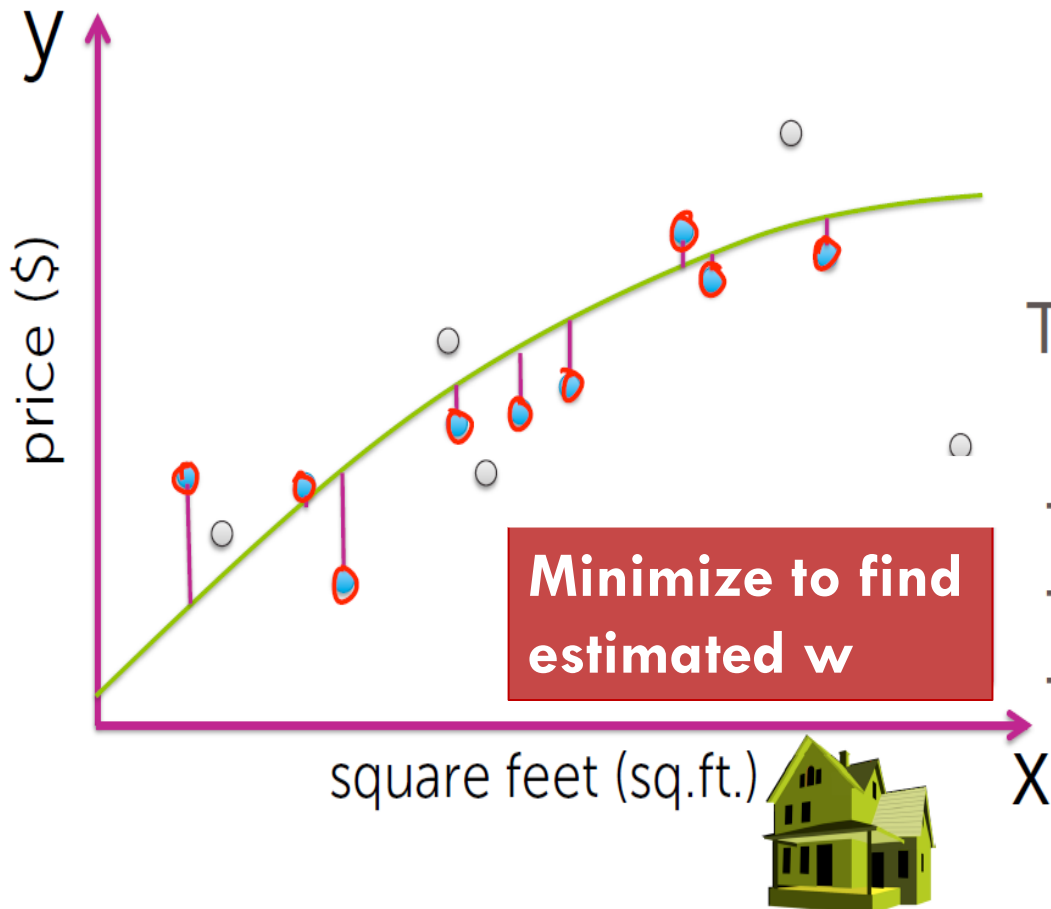
14



- training set 
- test set 

Training error

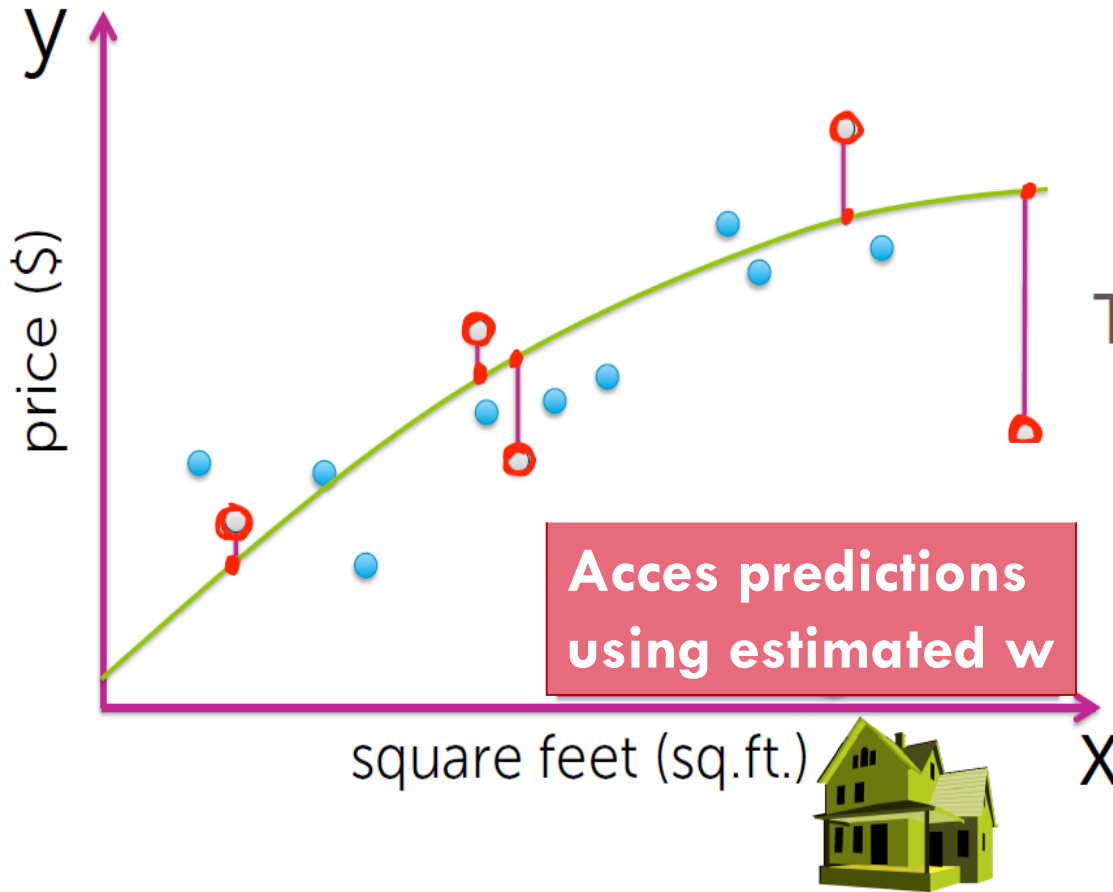
15



$$\begin{aligned} \text{Training error } (w) = & (\$_{\text{train } 1} - f_w(\text{sq.ft.}_{\text{train } 1}))^2 \\ & + (\$_{\text{train } 2} - f_w(\text{sq.ft.}_{\text{train } 2}))^2 \\ & + (\$_{\text{train } 3} - f_w(\text{sq.ft.}_{\text{train } 3}))^2 \\ & + \dots \text{ [include all} \\ & \quad \text{training houses]} \end{aligned}$$

Test error

16

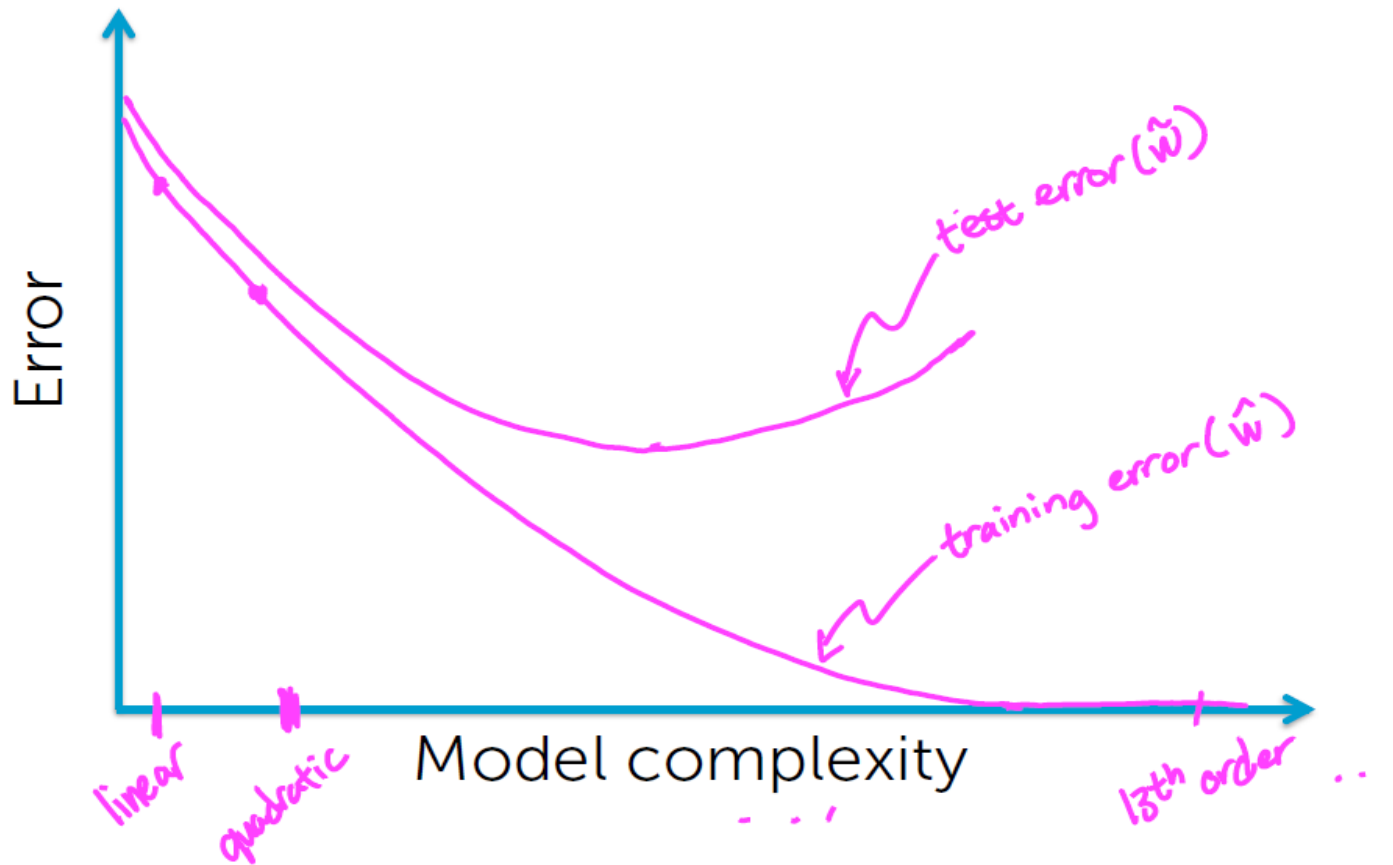


Test error \hat{w} =

$$\begin{aligned} & (\$_{\text{test } 1} - f_{\hat{w}}(\text{sq.ft.}_{\text{test } 1}))^2 \\ & + (\$_{\text{test } 2} - f_{\hat{w}}(\text{sq.ft.}_{\text{test } 2}))^2 \\ & + (\$_{\text{test } 3} - f_{\hat{w}}(\text{sq.ft.}_{\text{test } 3}))^2 \\ & + \dots \text{ [include all} \\ & \quad \text{test houses]} \end{aligned}$$

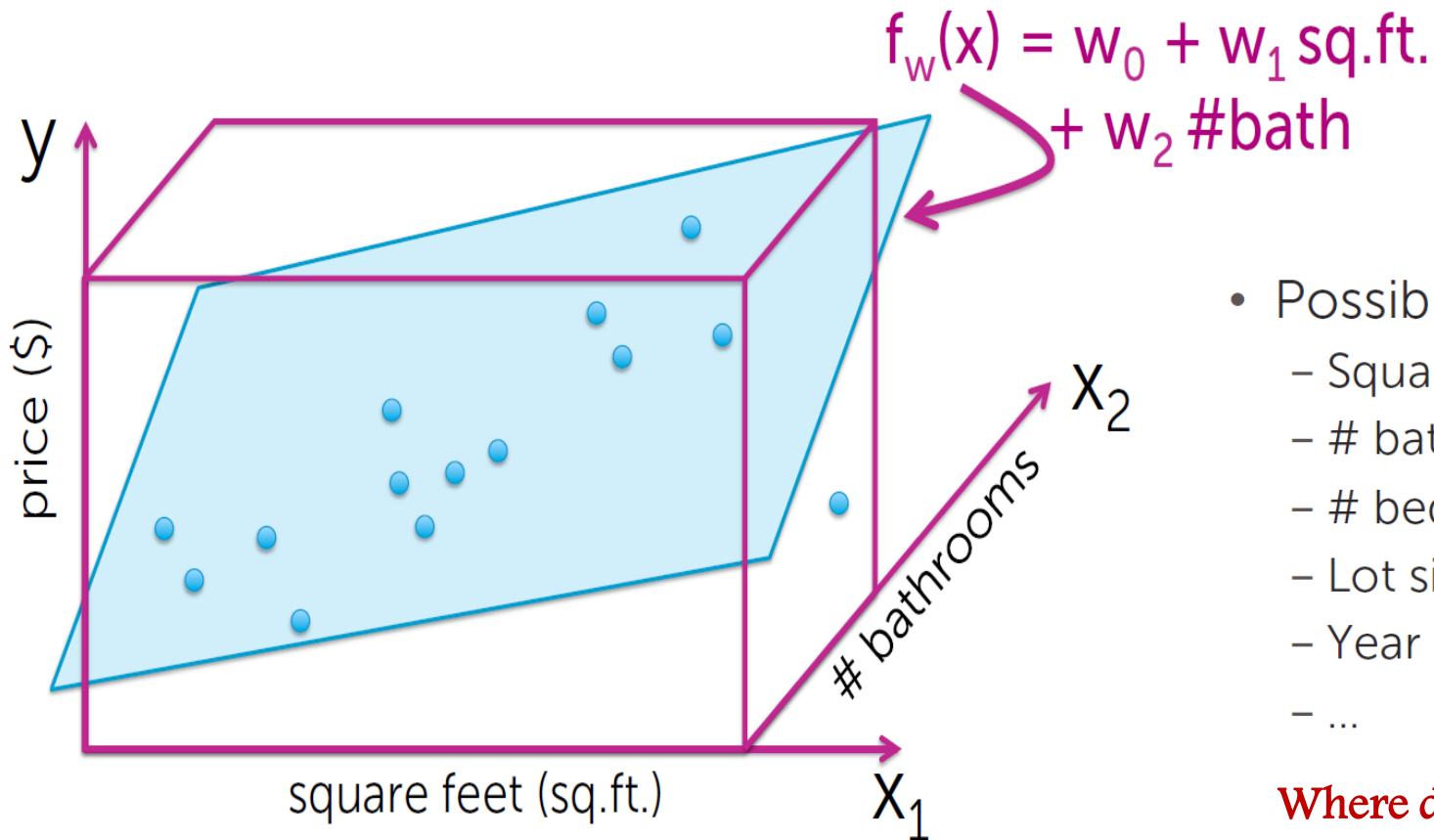
Training/test curve

17



Add more features

18

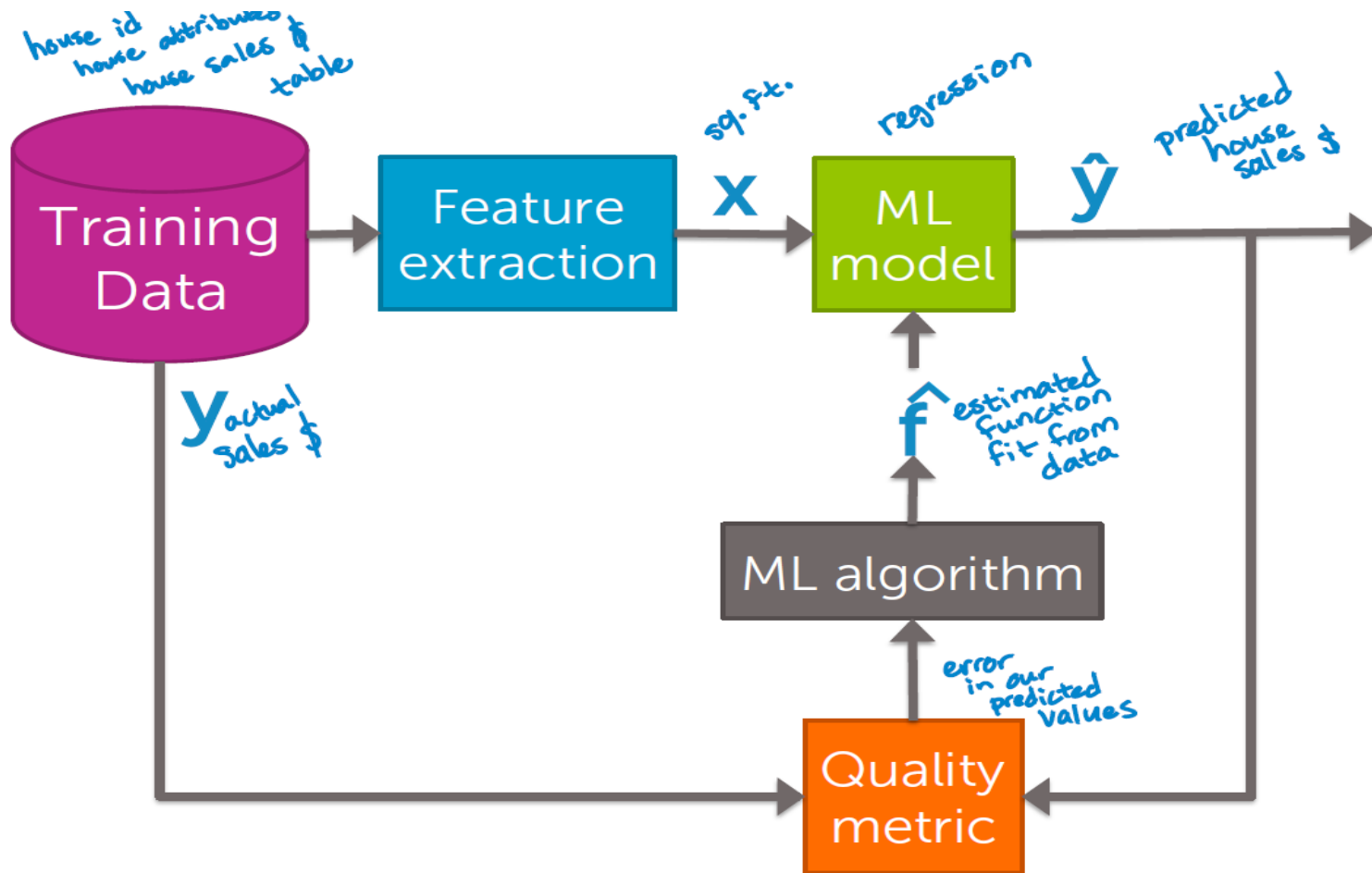


- Possible choices:
 - Square feet
 - # bathrooms
 - # bedrooms
 - Lot size
 - Year built
 - ...

Where do we stop?

Regression ML block

19

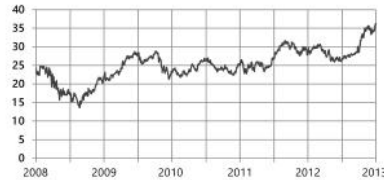


Other applications

20

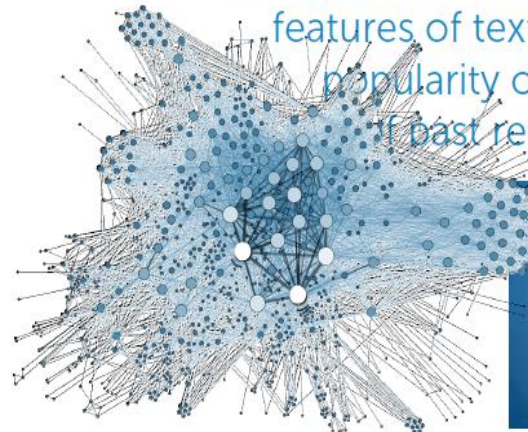
Stock predictions

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



Tweed popularity

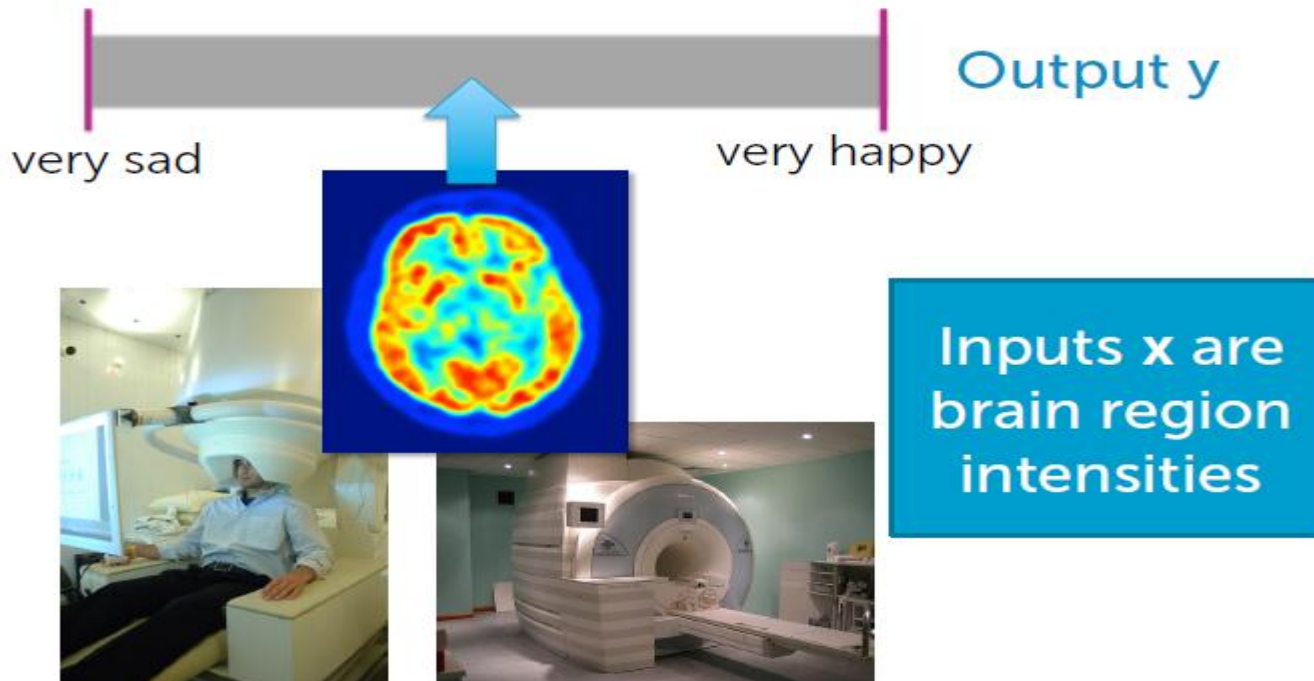
- How many people will retweet your tweet?
- Depends on # followers,
of followers of followers,
features of text tweeted,
popularity of hashtag,
of past retweets,...



Other applications

21

Reading your mind



We discussed how to

22

- 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