Regression:

Predicting House Prices

Emily Fox & Carlos Guestrin

Machine Learning Specialization

University of Washington

Predicting house prices

預測 "連續" 的 "值"

https://www.coursera.org/learn/machine-learning/lecture/1VkCb/supervised-learning
Regression: Predict continuous valued output (ex: price)
Classification: Discrete valued output (ex: 0 or 1)

How much is my house worth?

How much is my house worth?

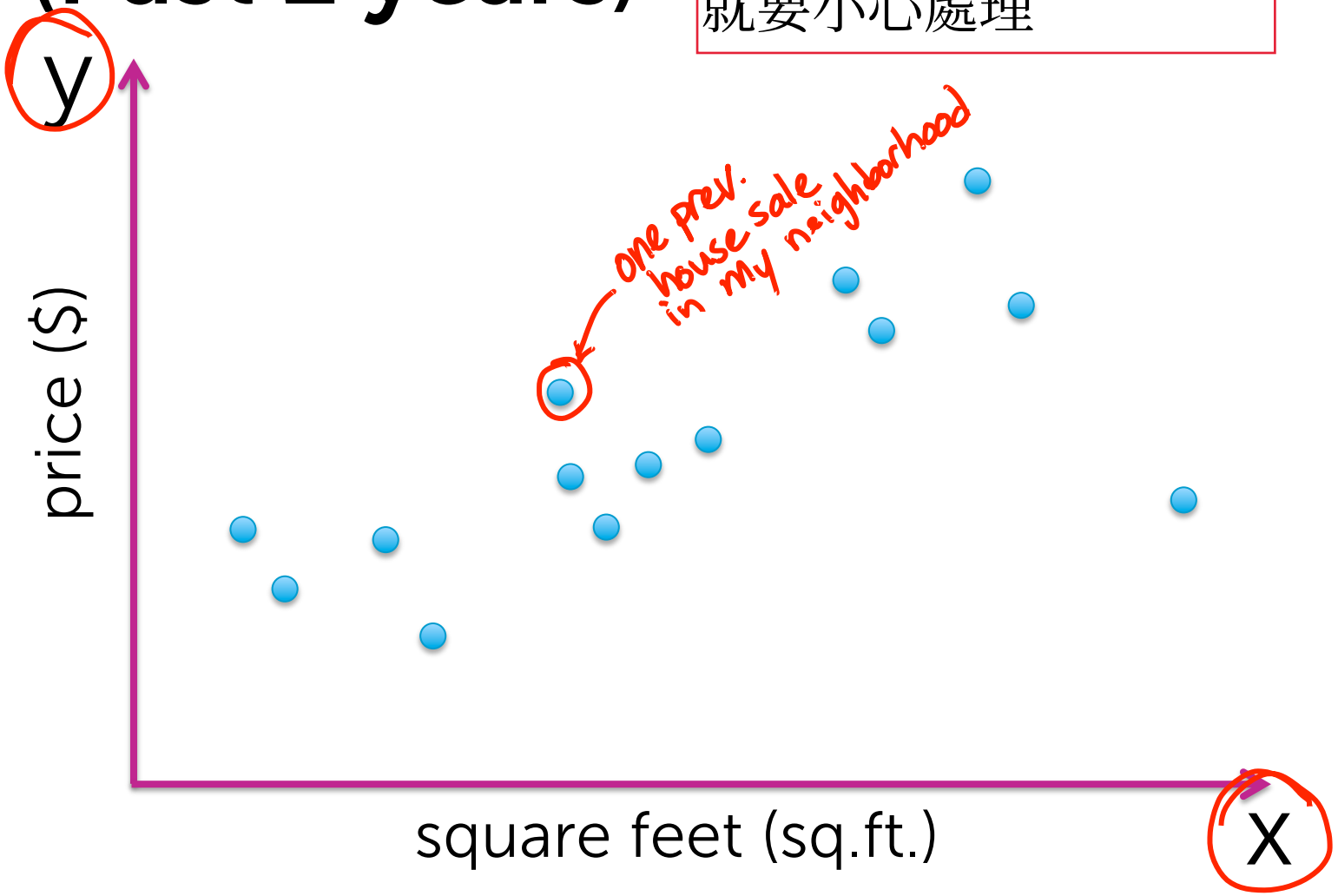
Look at recent sales in my neighborhood

- How much did they sell for?

Plot recent house sales (Past 2 years)

如果時間會影響價錢,
就要小心處理

假設：在此兩年間,
時間不影響價錢

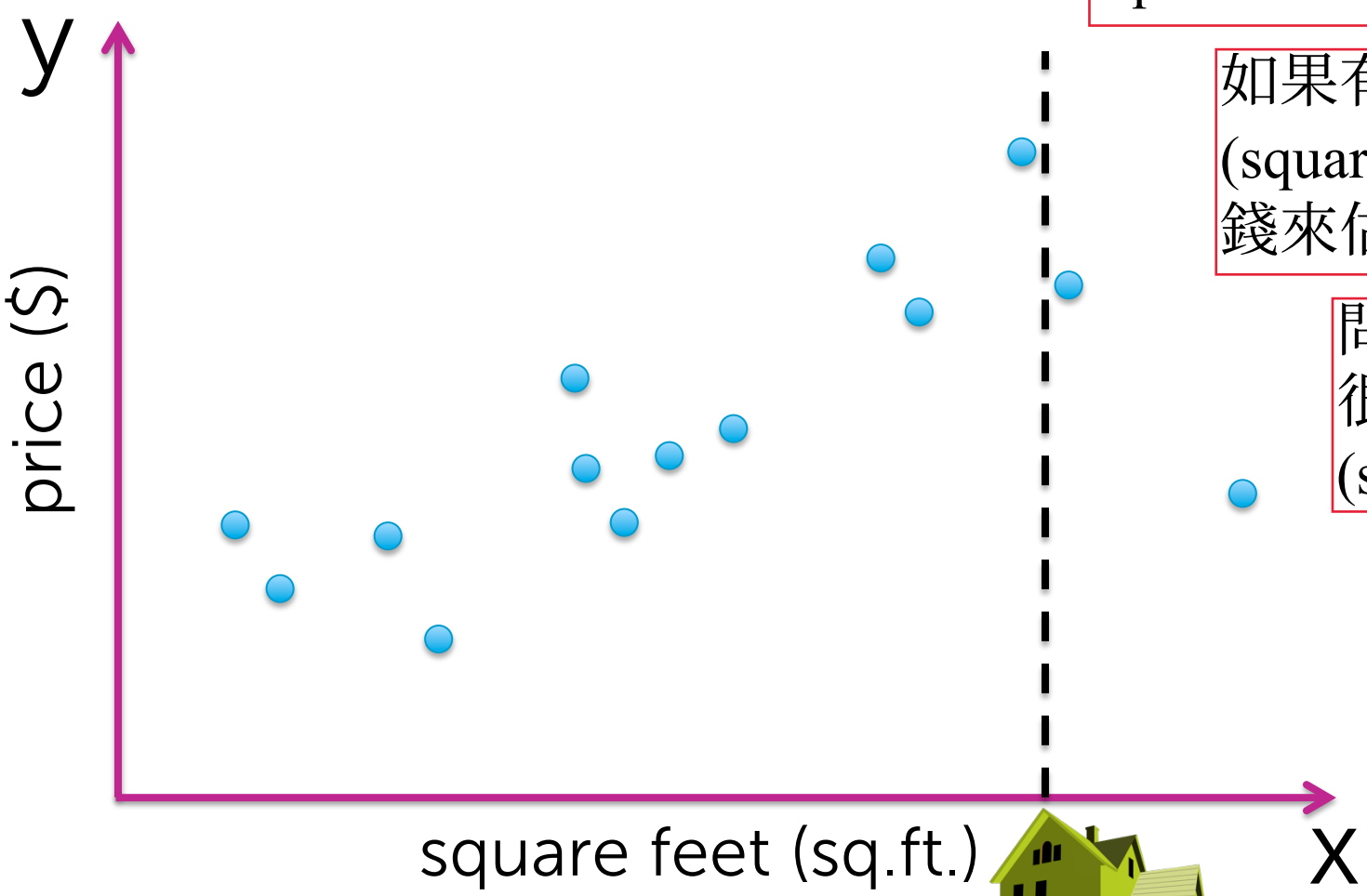


Terminology:

x – feature, covariate, or predictor

y – observation or response

Predict your house by similar houses



先假設房屋價錢只跟 square feet 有關

如果有別人的房子跟我一樣 (square feet)，則可用他的價錢來估價

問題在於，現實狀況下，很可能找不到一模一樣 (square feet) 的房子

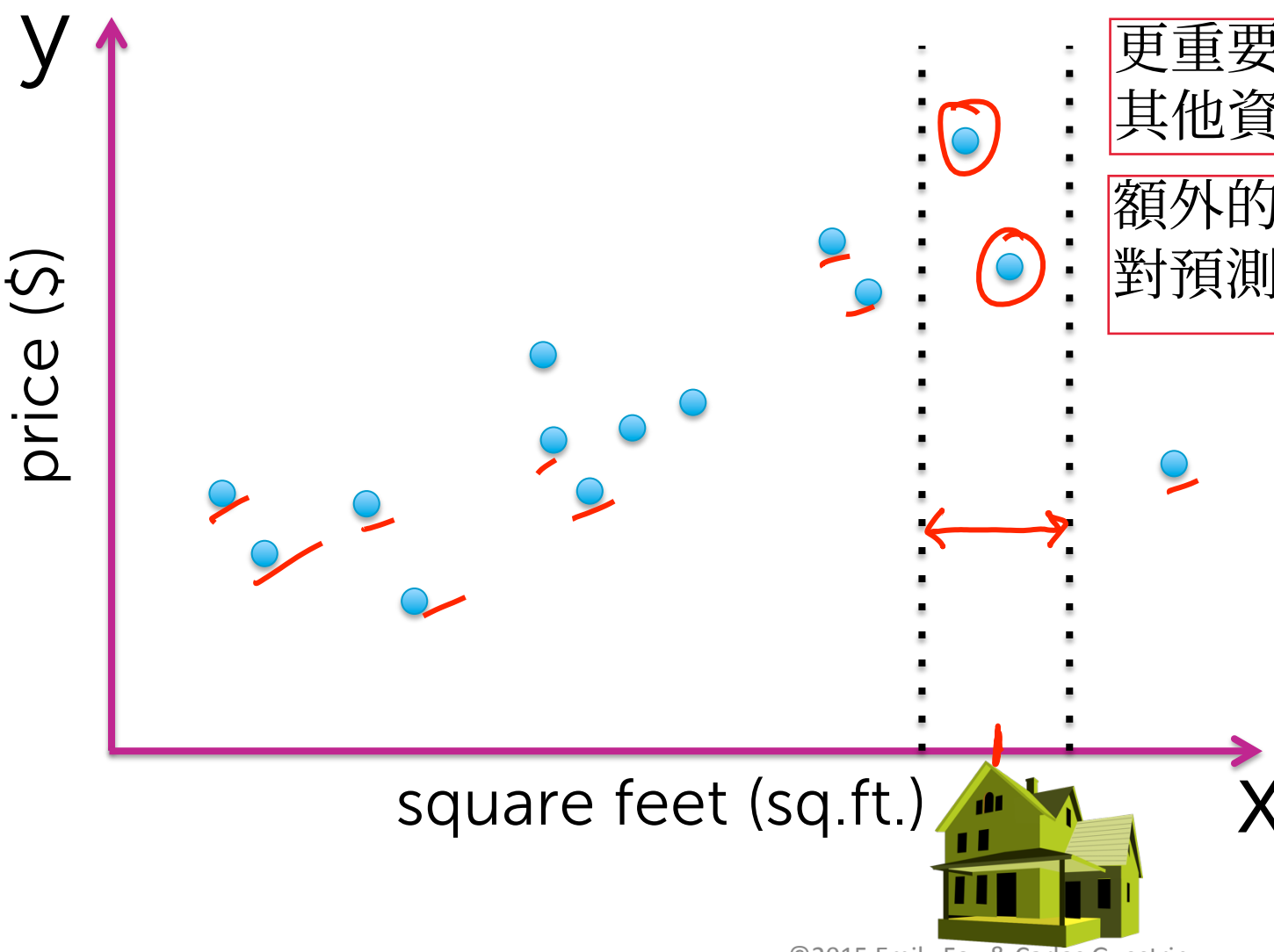
No house sold recently had *exactly* the same sq.ft.

Predict your house by similar houses

找範圍 (square feet 夠接近)？
=> 如何挑範圍？ => 有可能該範圍內的資料仍舊很少

更重要的地方在於，這樣會把其他資料的訊息丟掉

額外的假設，不夠相同的 square feet 對預測價錢沒有意義 (過度假設?)



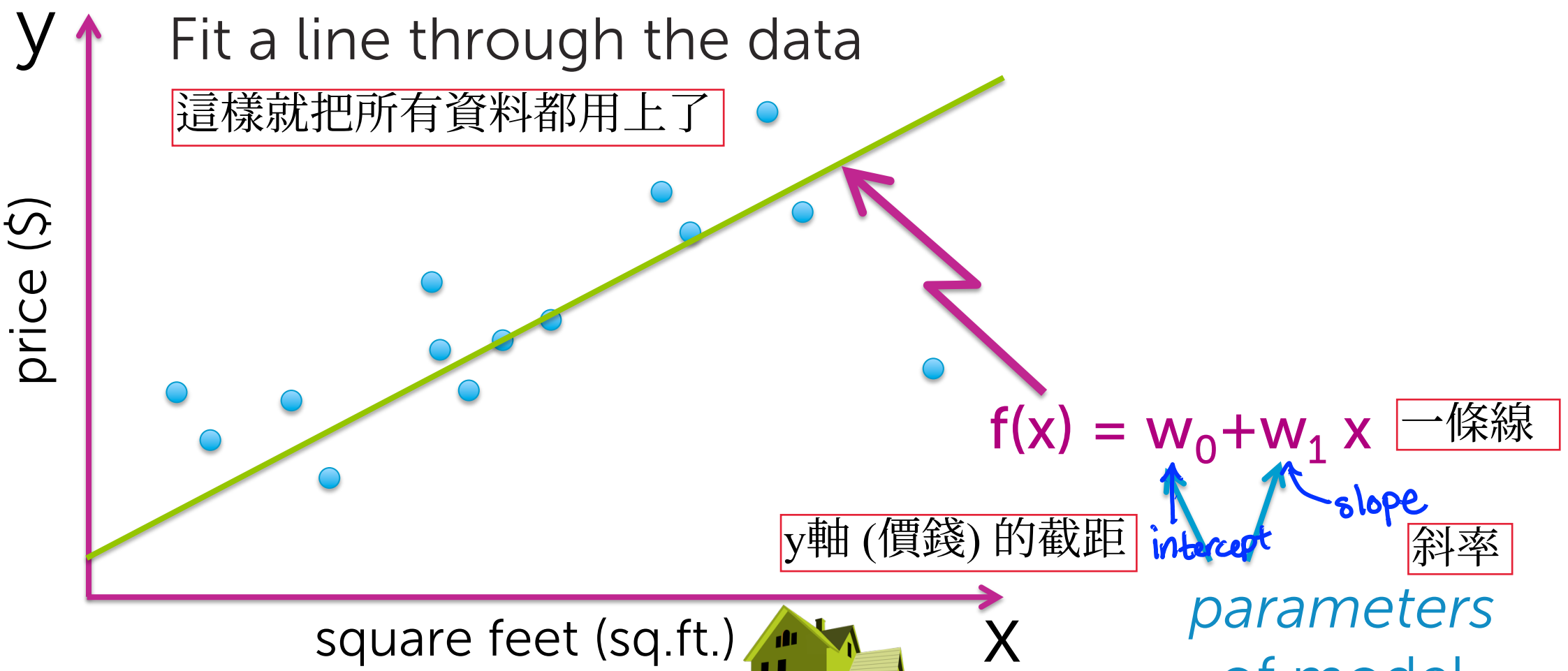
- Look at average price in range
- **Still only 2 houses!**
- Throwing out info from all other sales

額外附註：丟資料未必永遠不好

Linear regression

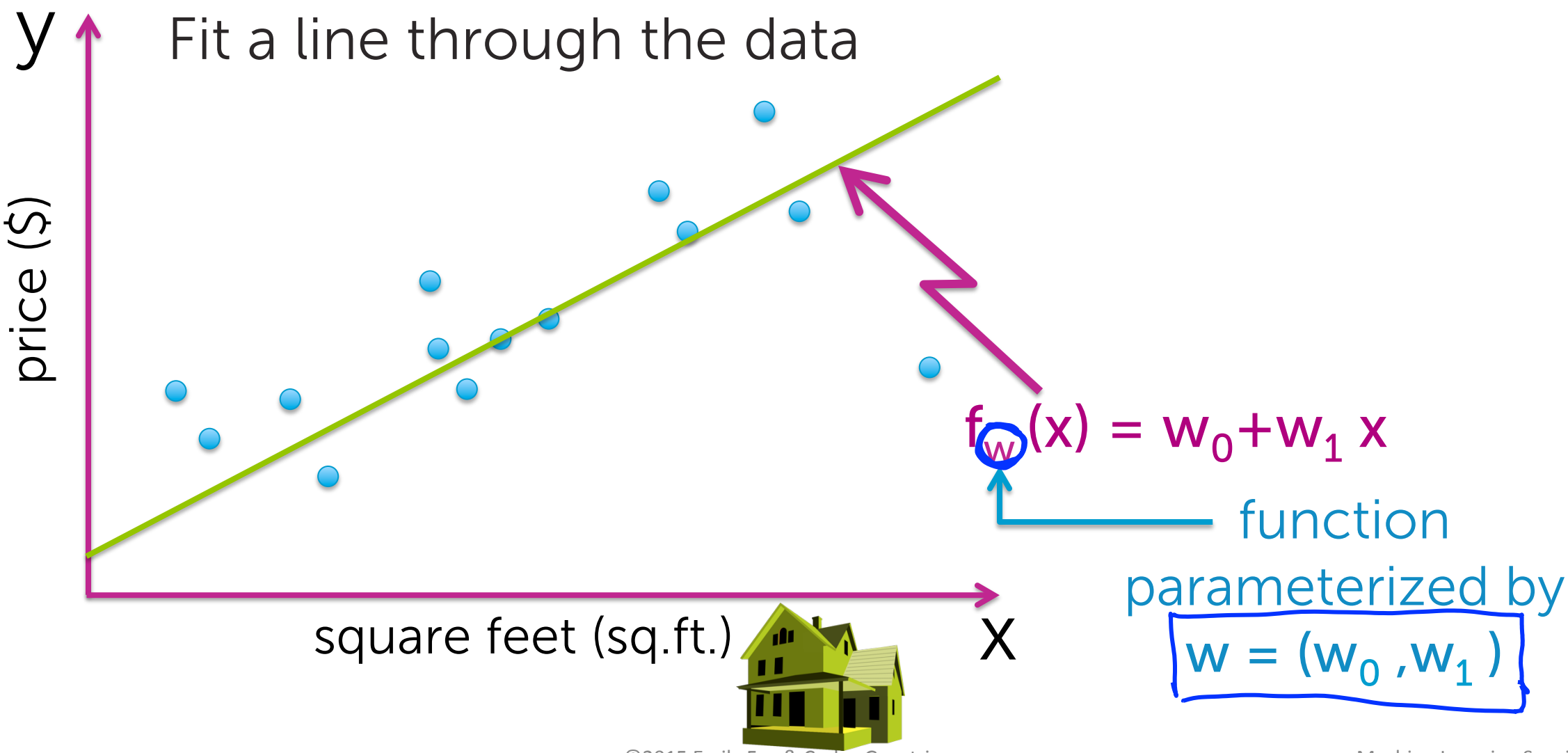
Use a **linear** regression model

假設房屋價錢跟特徵 (square feet) 是線性關係



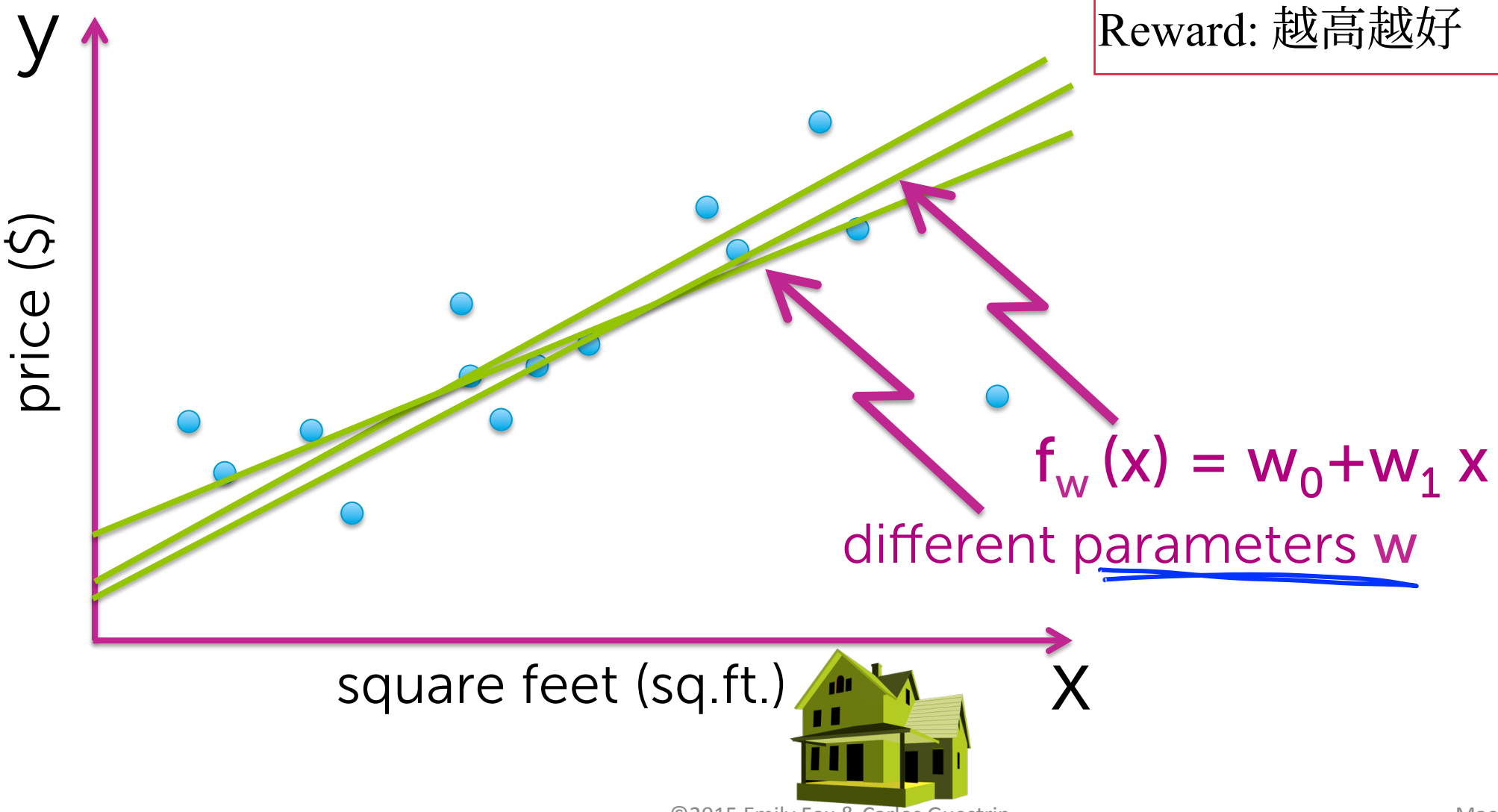
Use a **linear** regression model

參照 Andrew 的 machine learning



Which line?

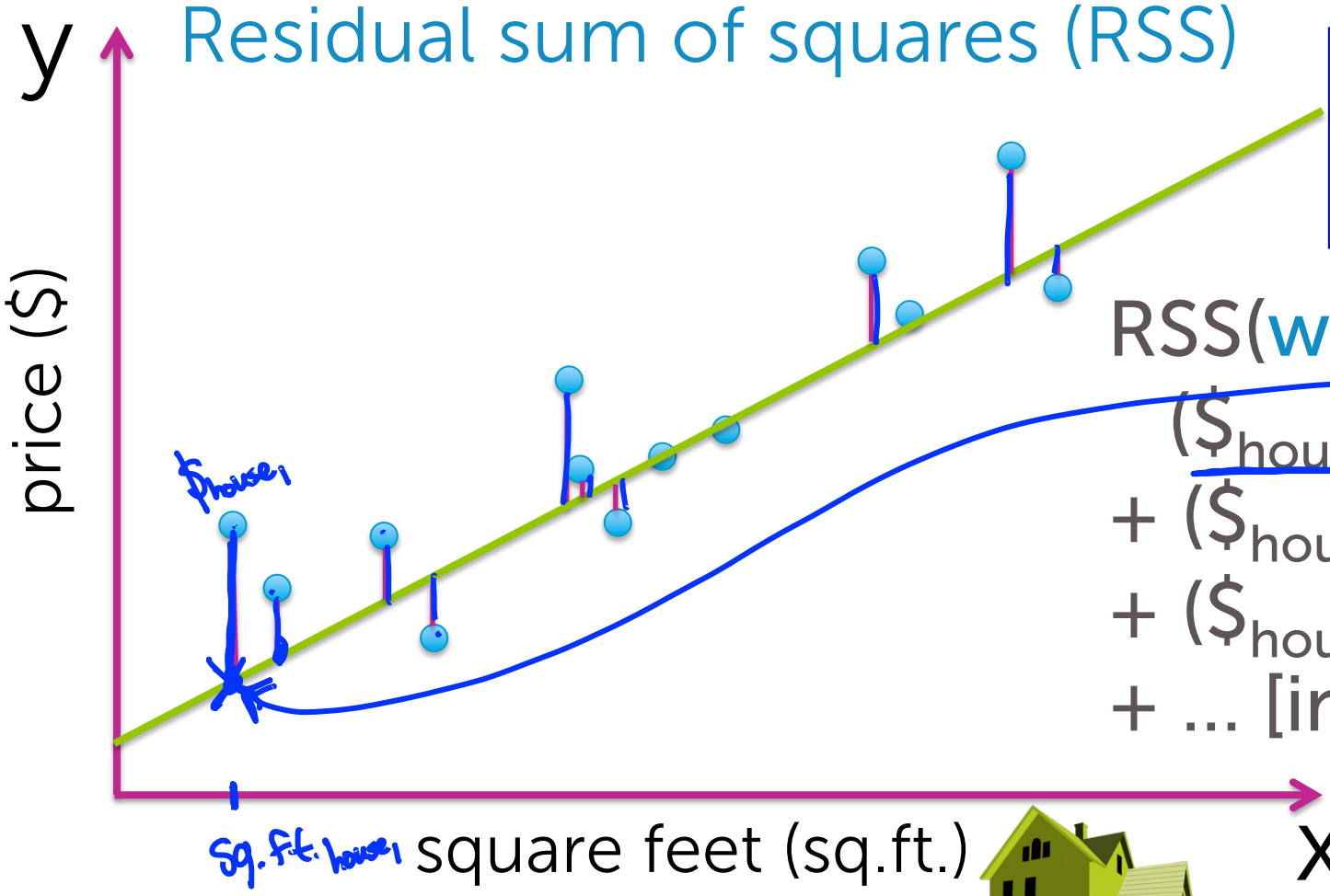
需要定義一個挑選參數的準則
(決定哪條線比較好)
Cost: 越低越好
Reward: 越高越好



"Cost" of using a given line

參照 Andrew 的 machine learning

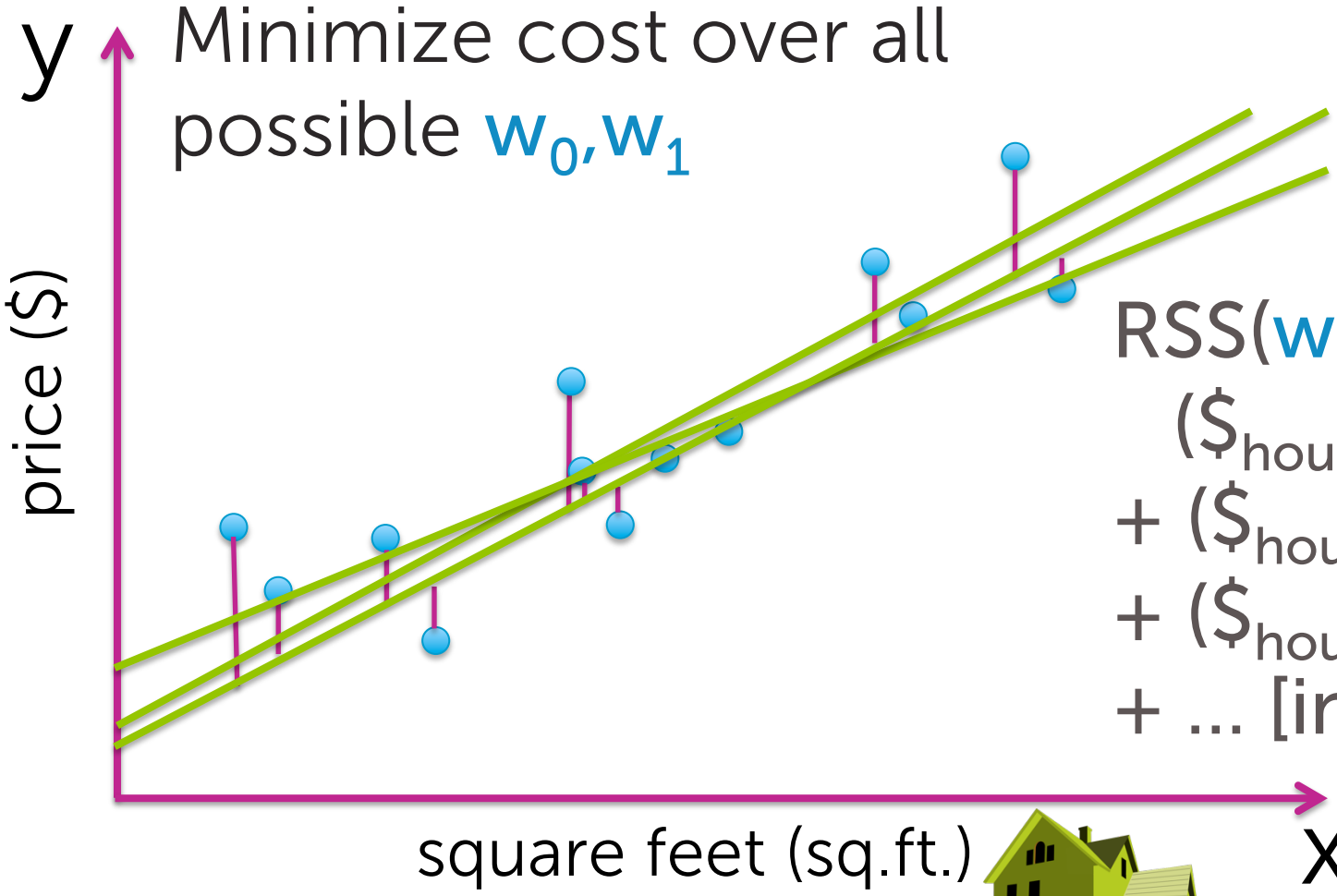
Why RSS?
跟 Andrew 用的差了"/ 2m"
使用絕對值? 更多次方?



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

Find "best" line

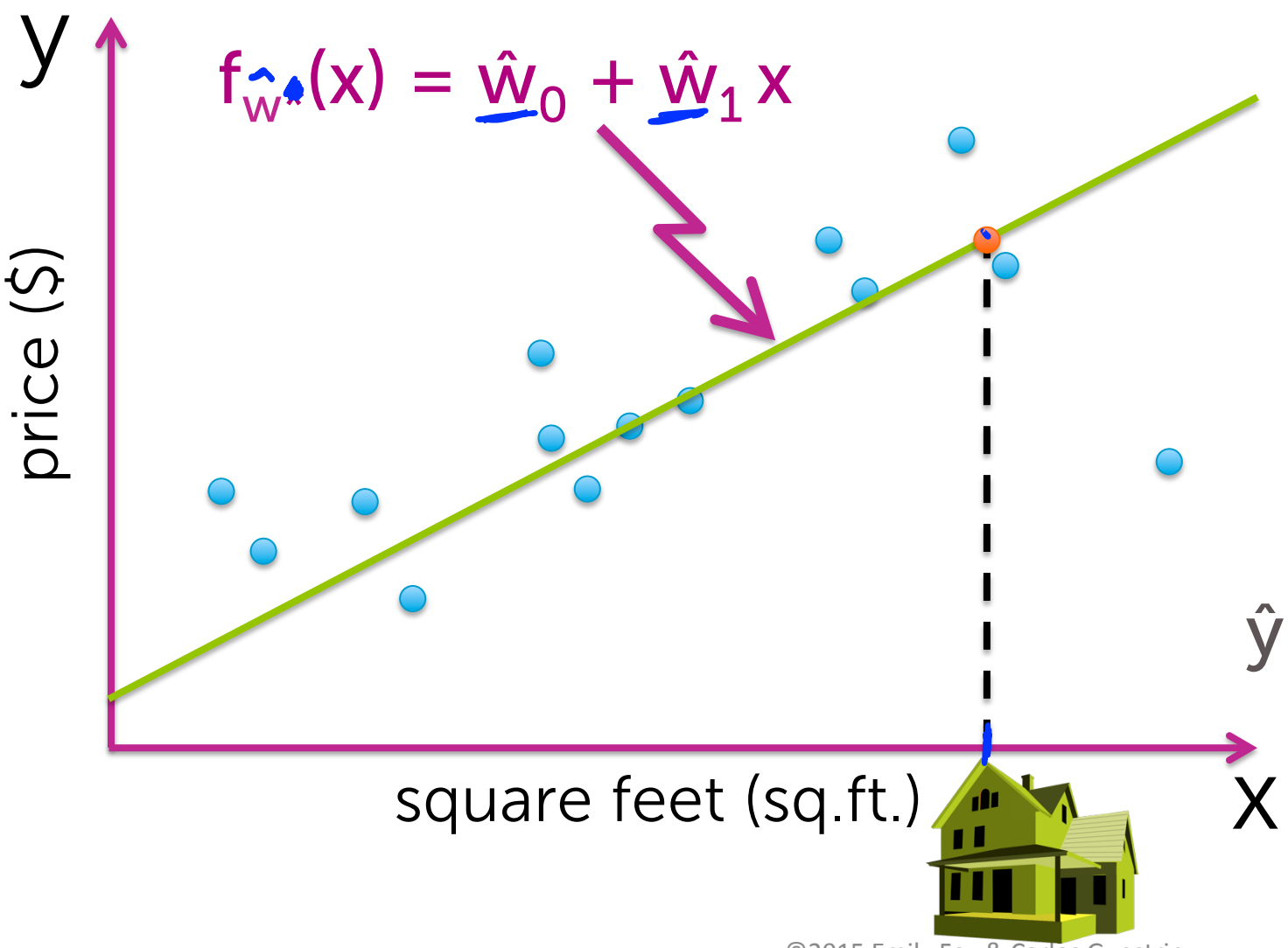
在此，"best" 即 lowest cost



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

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

Predicting your house price

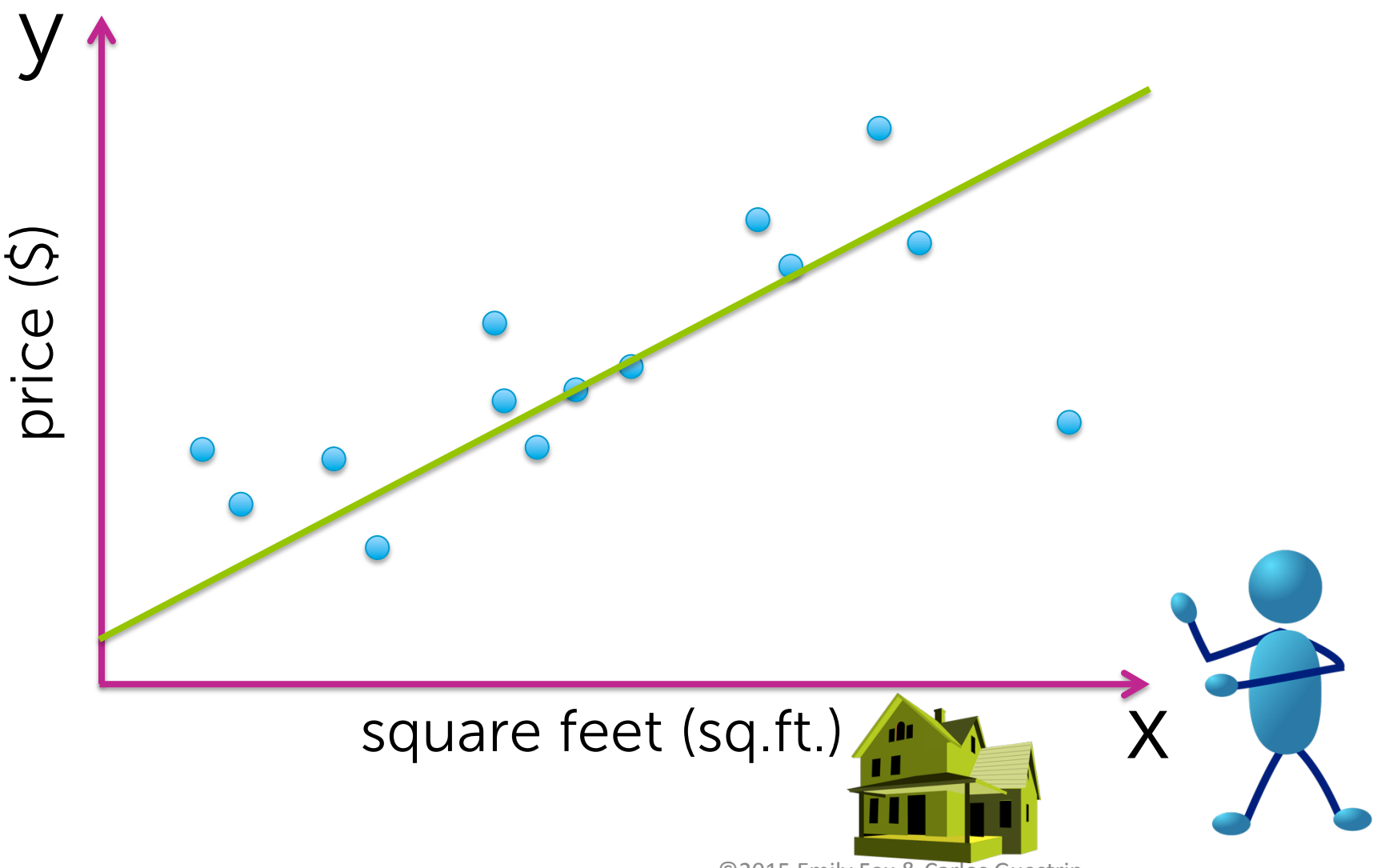


Best guess of your house price:

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

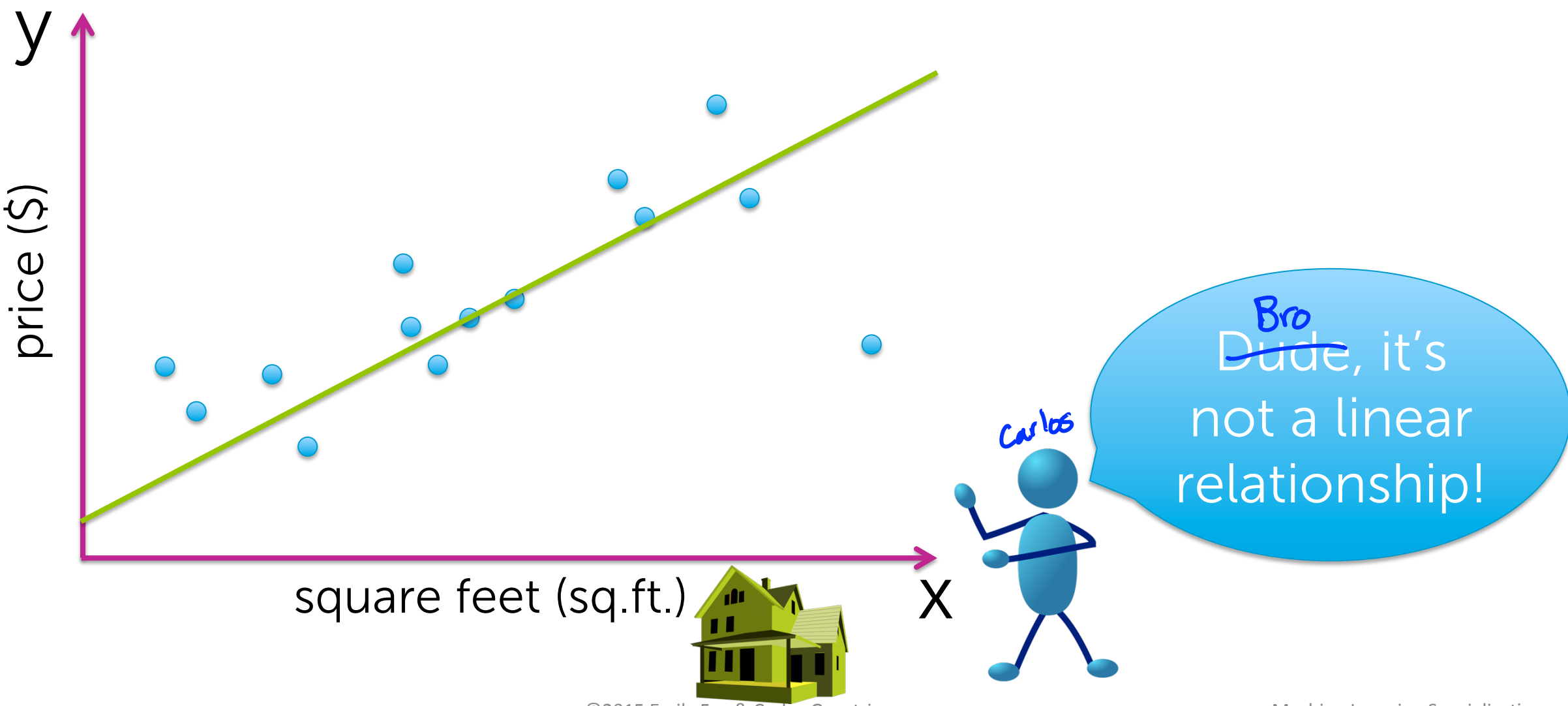
Adding higher order effects

Fit data with a line or … ?

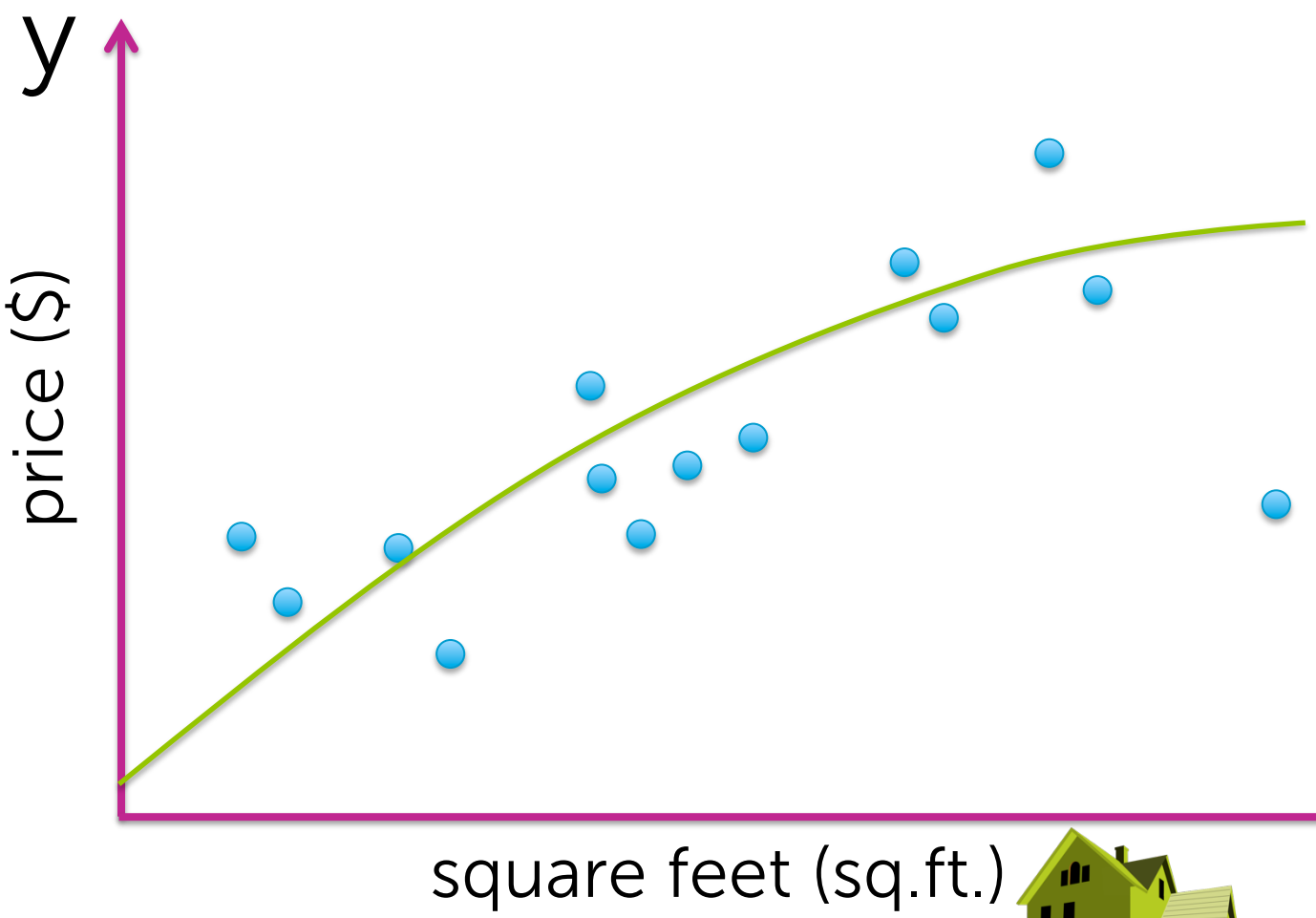


You show your friend your analysis

Fit data with a line or … ?



What about a quadratic function?



直線或許不好，所以考慮二次曲線(?)

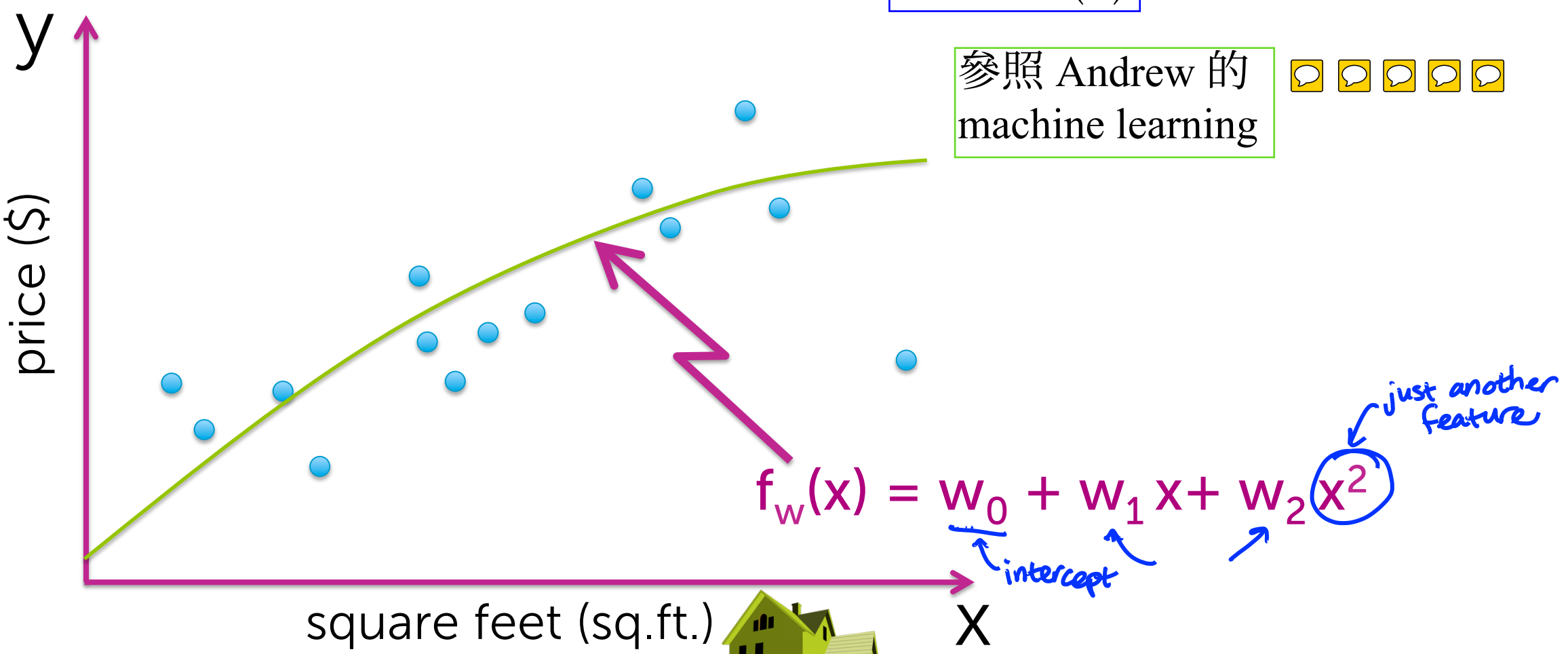
為何覺得線性關係不好?(怎麼知道的)

Dude, it's not a linear relationship!

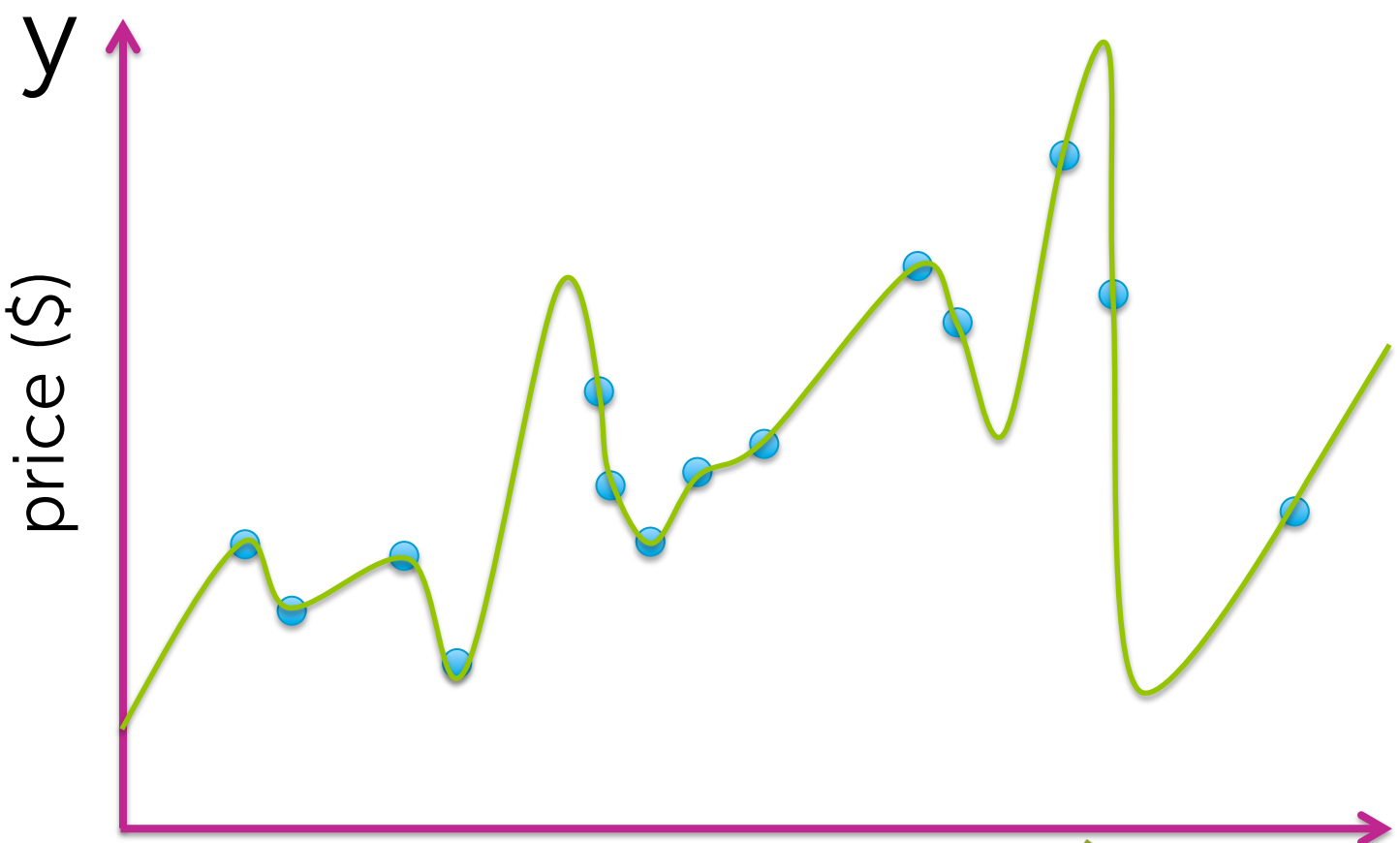
What about a quadratic function?

二次曲線(?)

參照 Andrew 的 machine learning



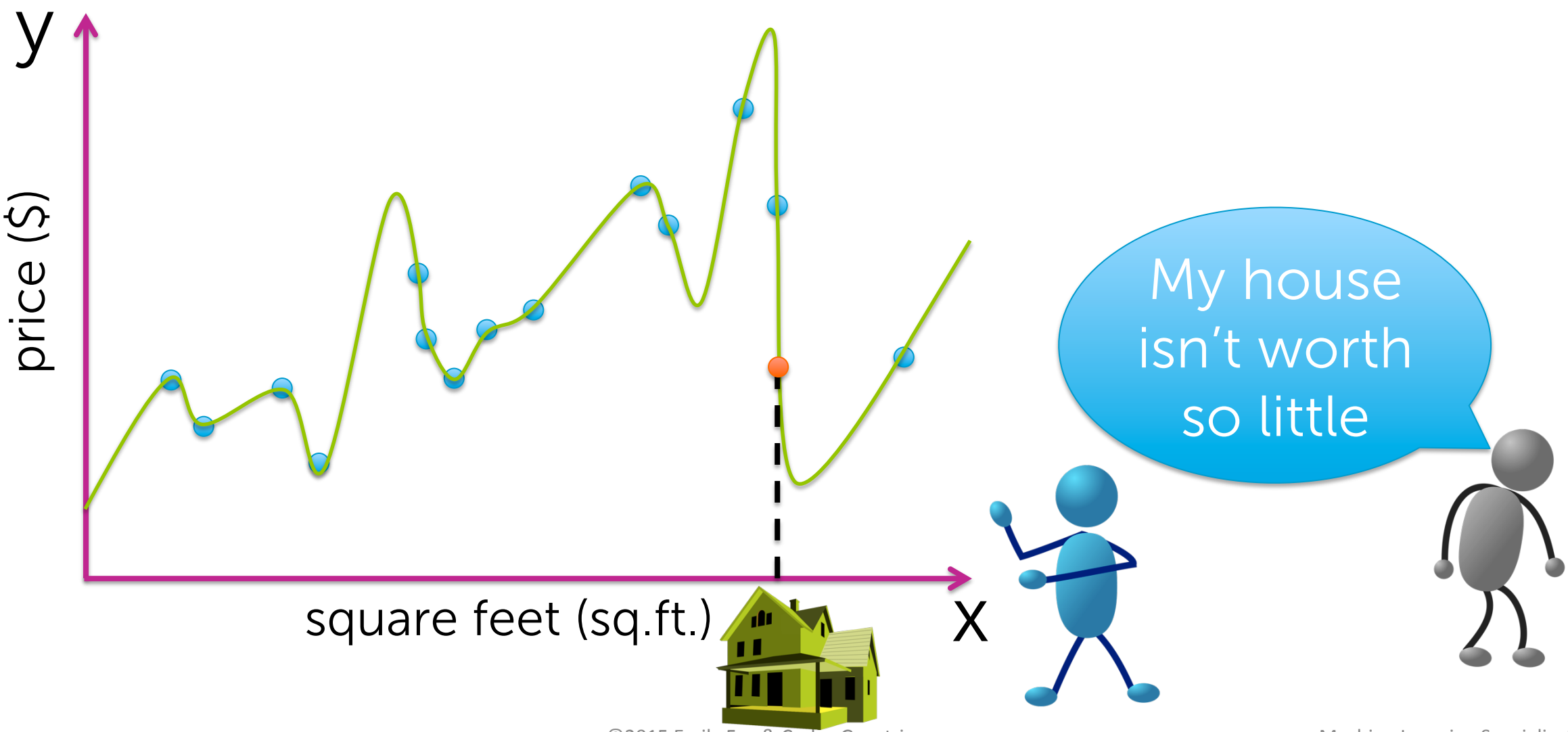
Even higher order polynomial



I can minimize your RSS

Do you believe this fit?

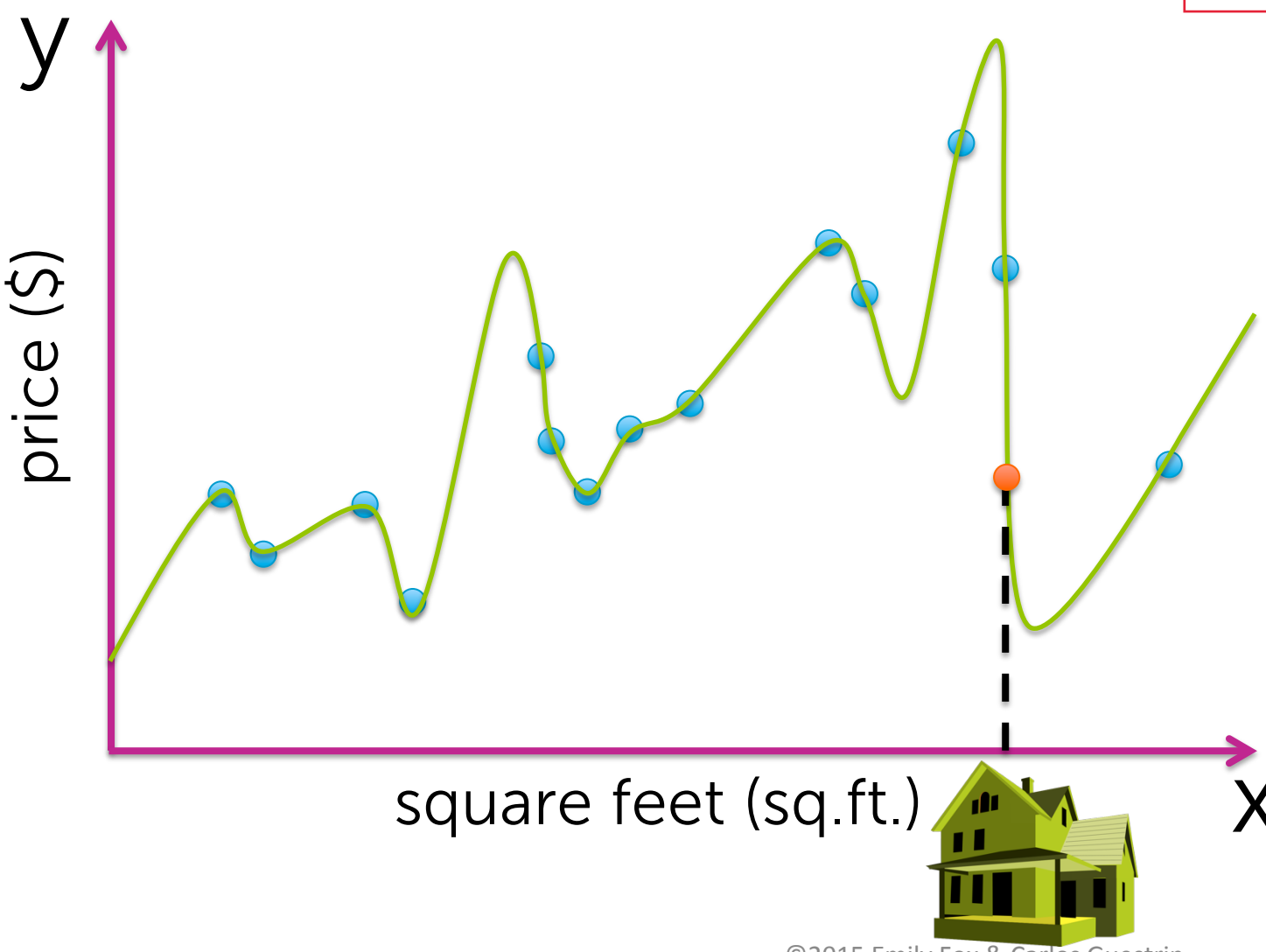
很符合 training data 並不代表 testing 時會有好的成果



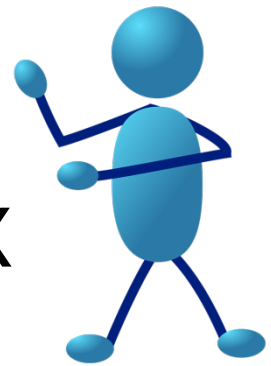
Evaluating overfitting via training/test split

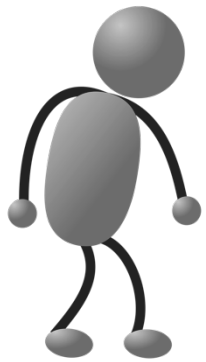
Do you believe this fit?

檢驗是否 over-fitting 可以看資料以及學出來的曲線 (or decision boundary)

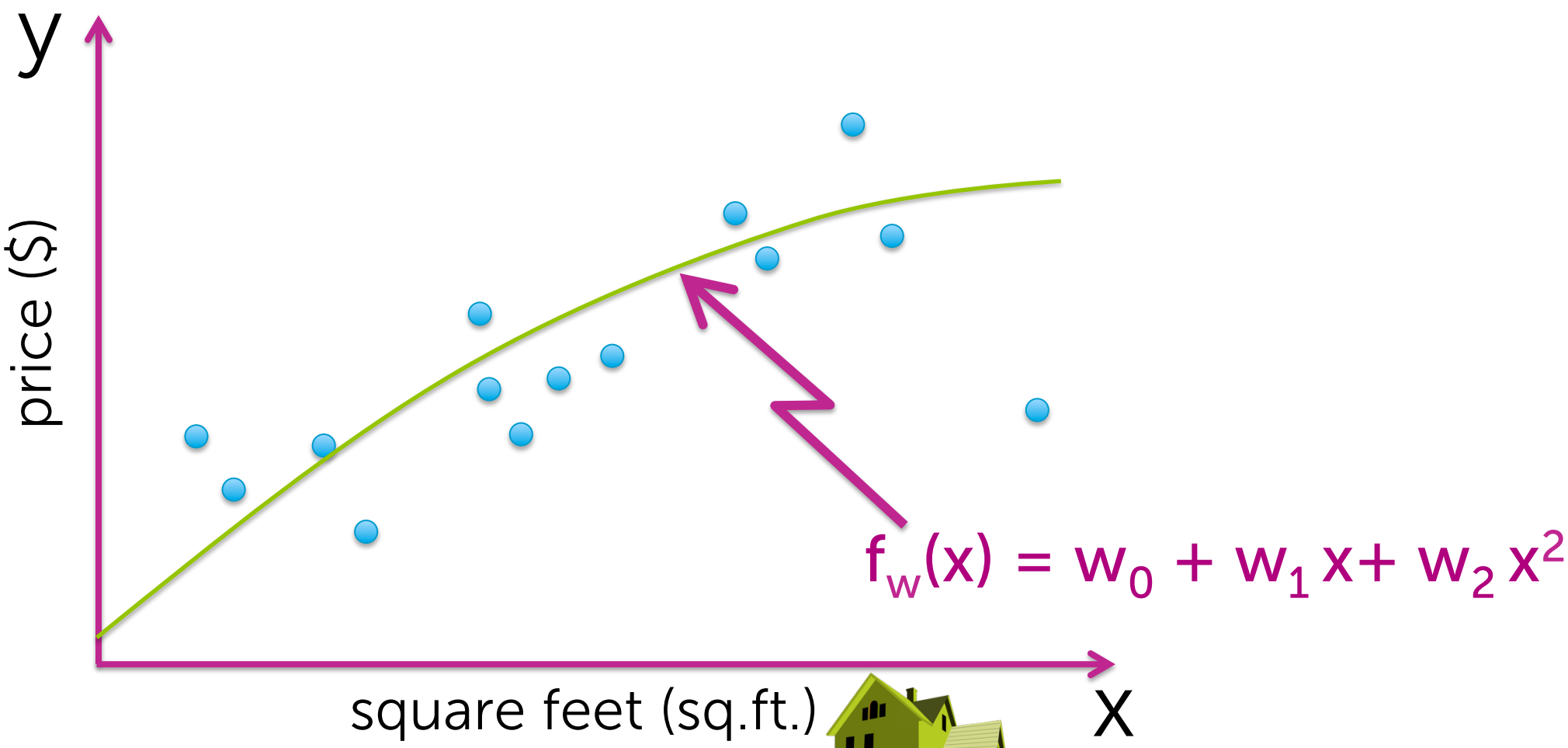


Minimizes RSS,
but bad predictions

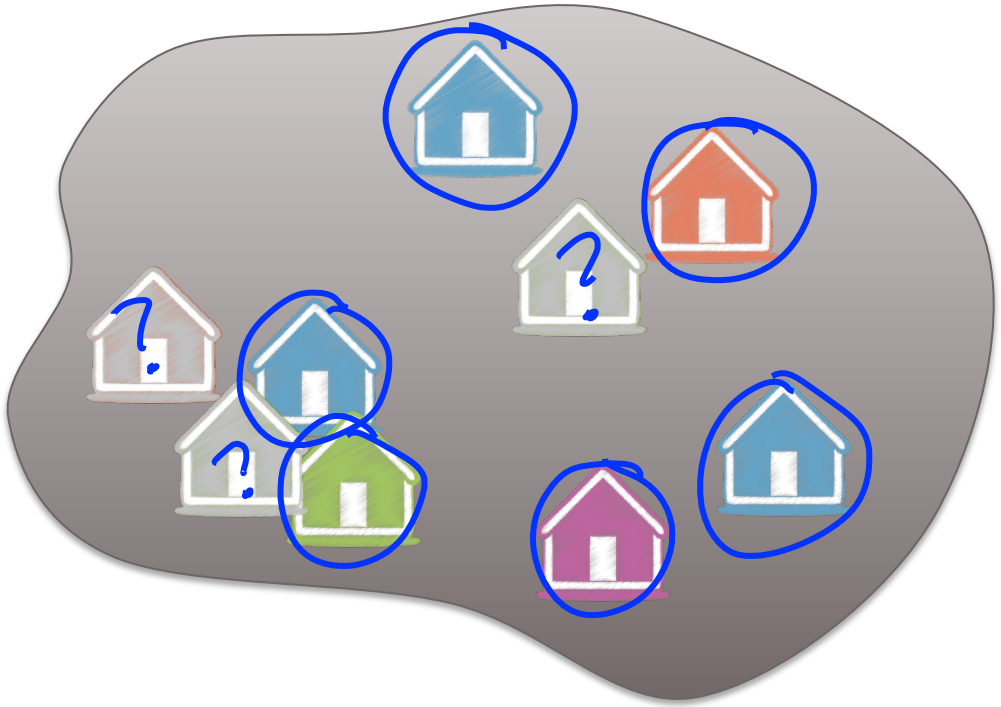




What about a quadratic function?



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

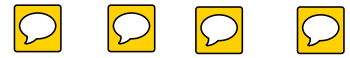
Training/test split

此觀念後續會修正，選 model order / complexity 一定不是用 test set 去決定

單純想感受 testing performance 而沒有要調 hyper-parameters 的情形下可以僅做 training / test split (ex: 70-30)

參照 Andrew 的 machine learning

參照 Andrew 的 deep learning

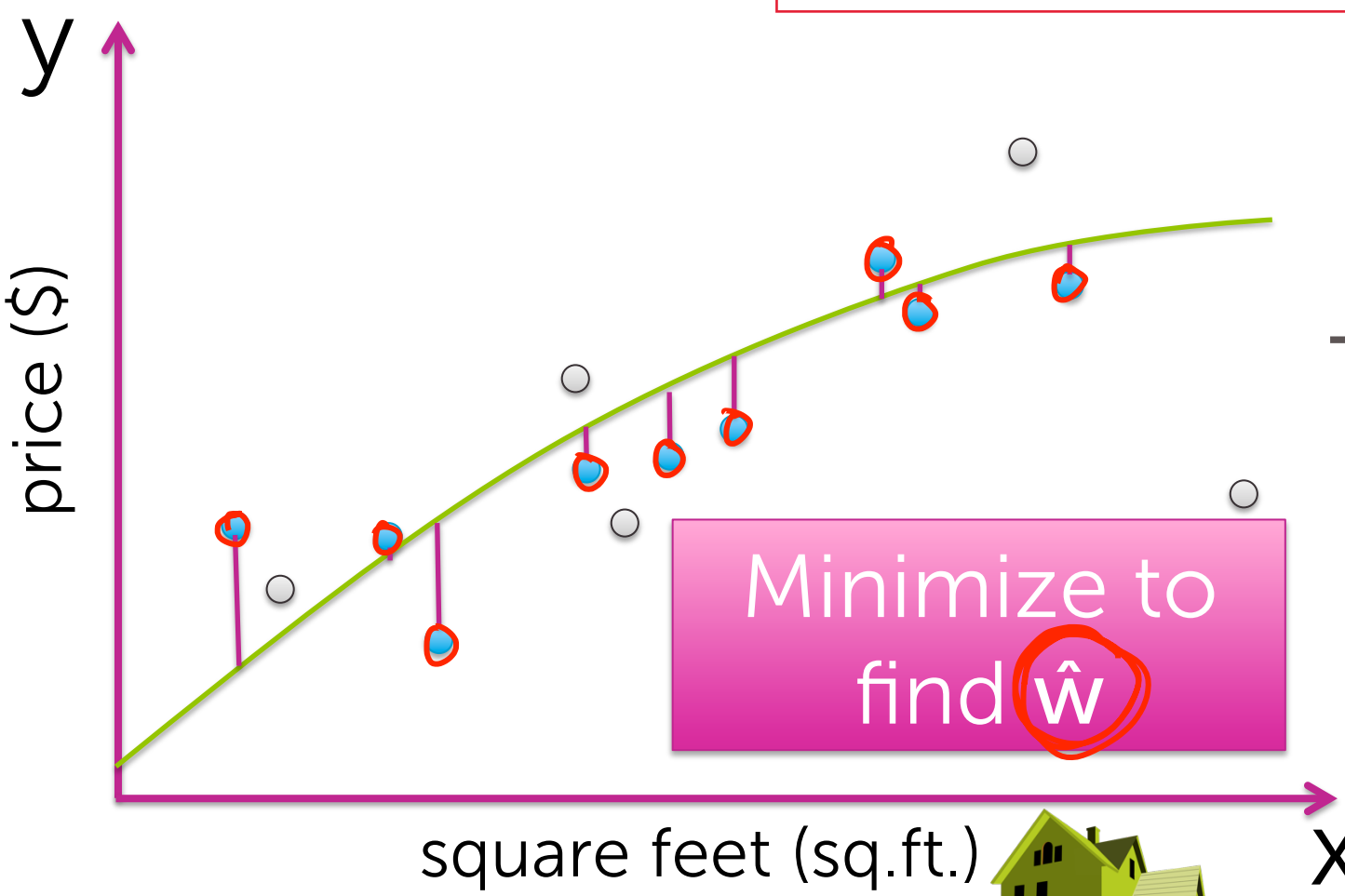


Terminology: – training set

– test set

Training error

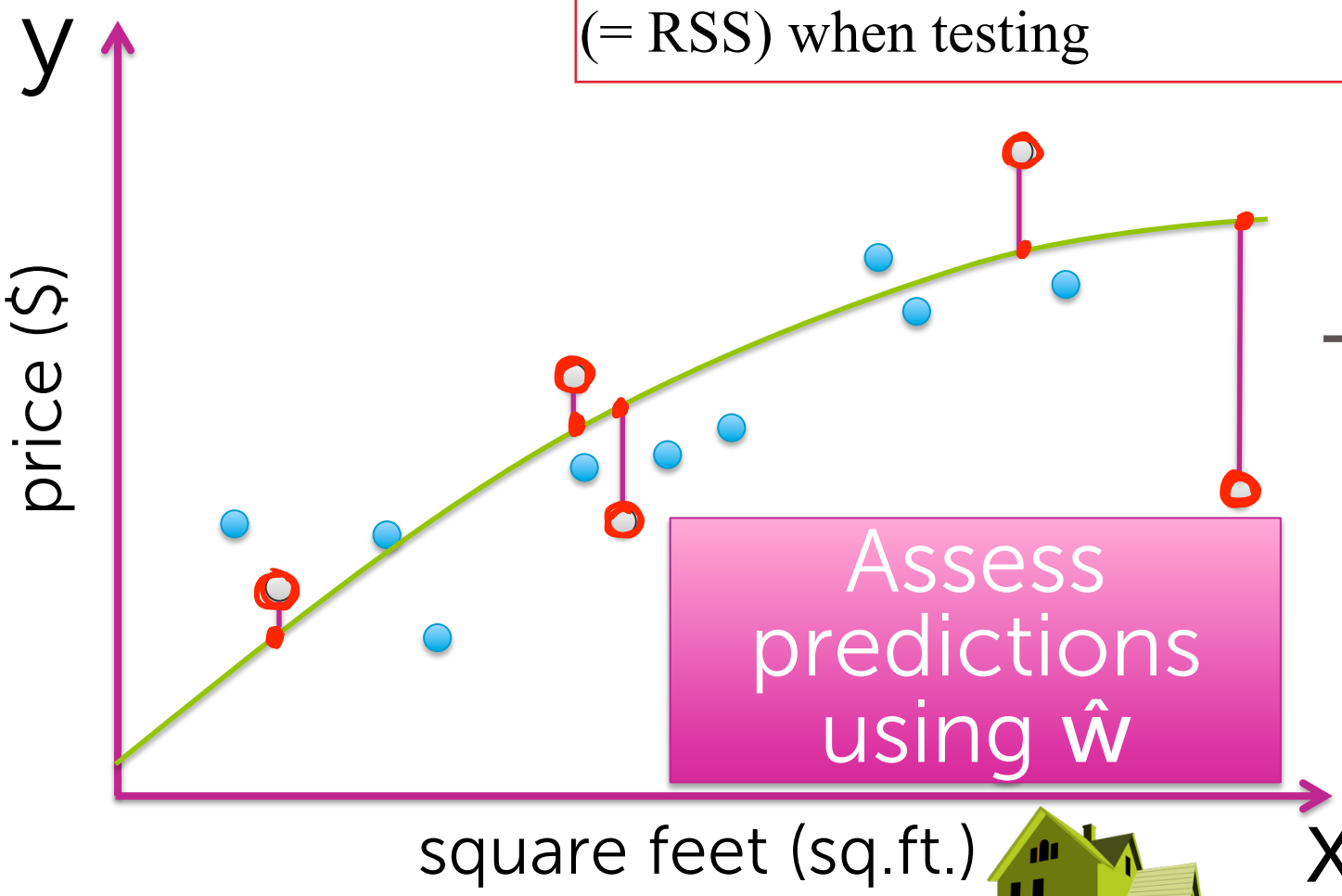
不是一般的錯誤率 (error rate)
此處更像是 training cost (= RSS)
或是定義 error function = RSS



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

不是一般的錯誤率 (error rate)
完整解釋為 RSS when testing
或是 value of the error function (= RSS) when testing

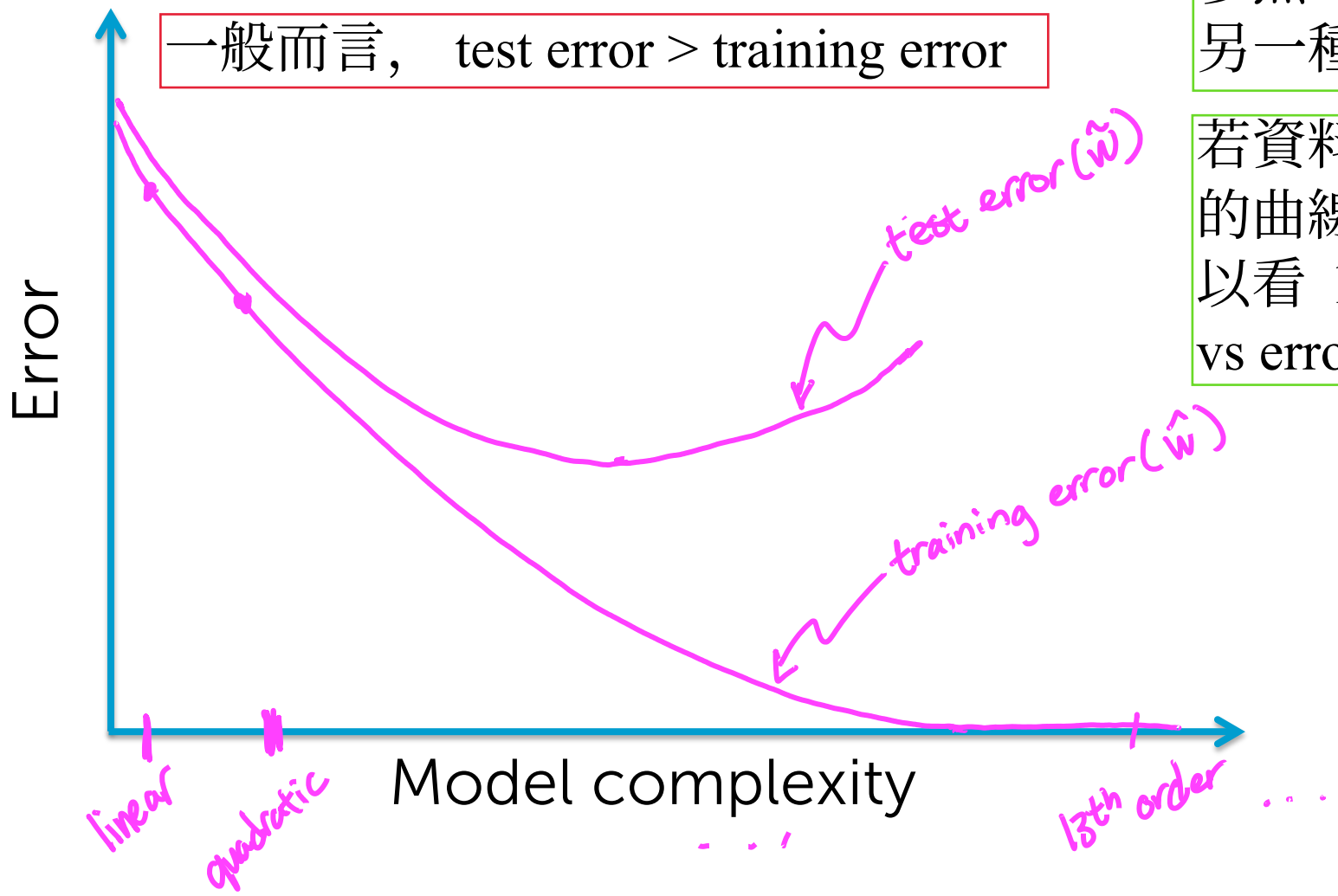


Test error ($\hat{w}$) =
$(\$_{test\ 1} - f_{\hat{w}}(sq.ft._{test\ 1}))^2$
$+ (\$_{test\ 2} - f_{\hat{w}}(sq.ft._{test\ 2}))^2$
$+ (\$_{test\ 3} - f_{\hat{w}}(sq.ft._{test\ 3}))^2$
+ … [include all test houses]

Training/Test Curves

Model complexity vs error

一般而言， test error > training error

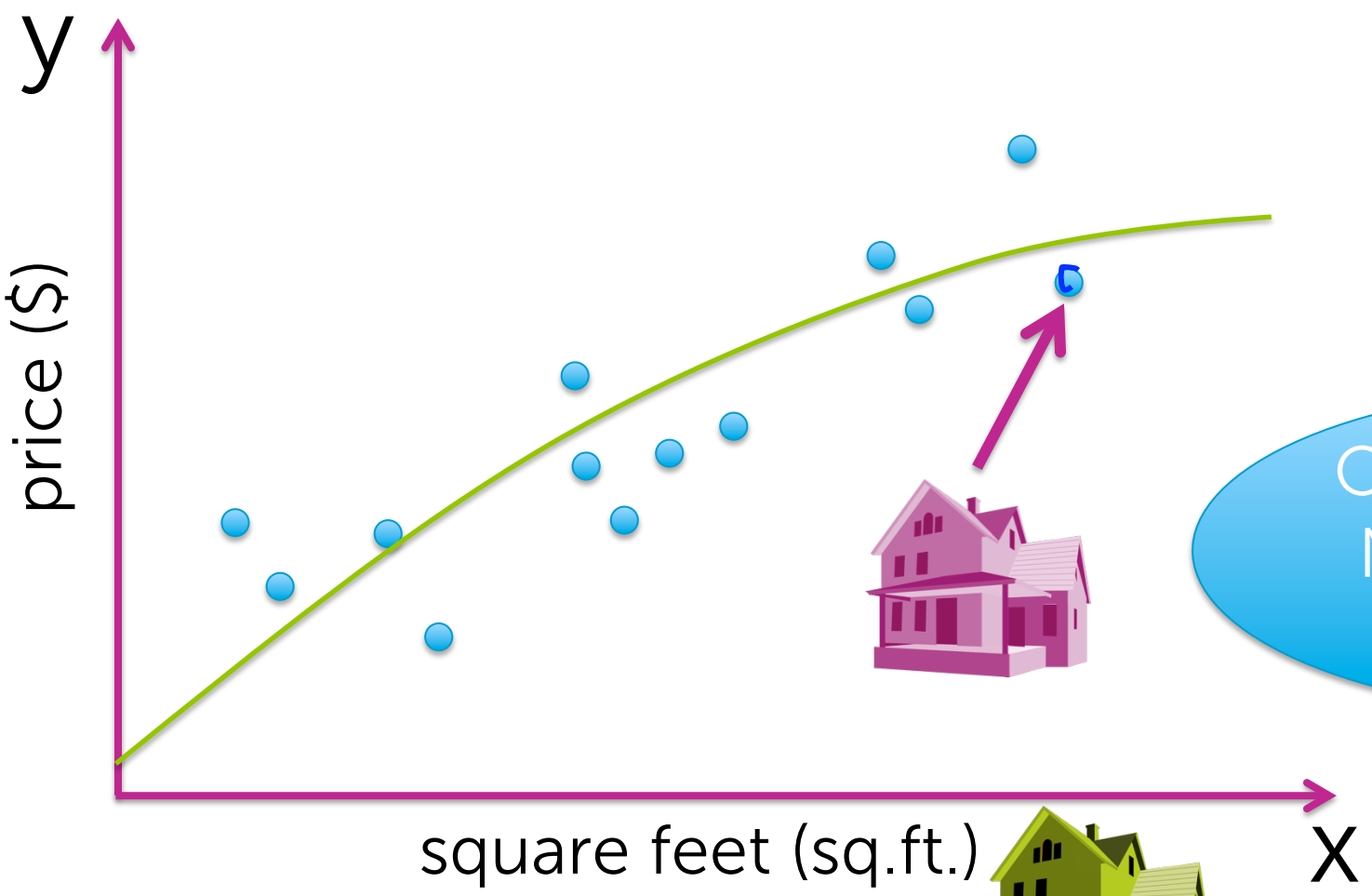


參照 Andrew 的 machine learning 另一種 curve (learning curve)

若資料維度過高無法一併畫出學得的曲線 (or decision boundary)， 可以看 learning curve (training set size vs error)

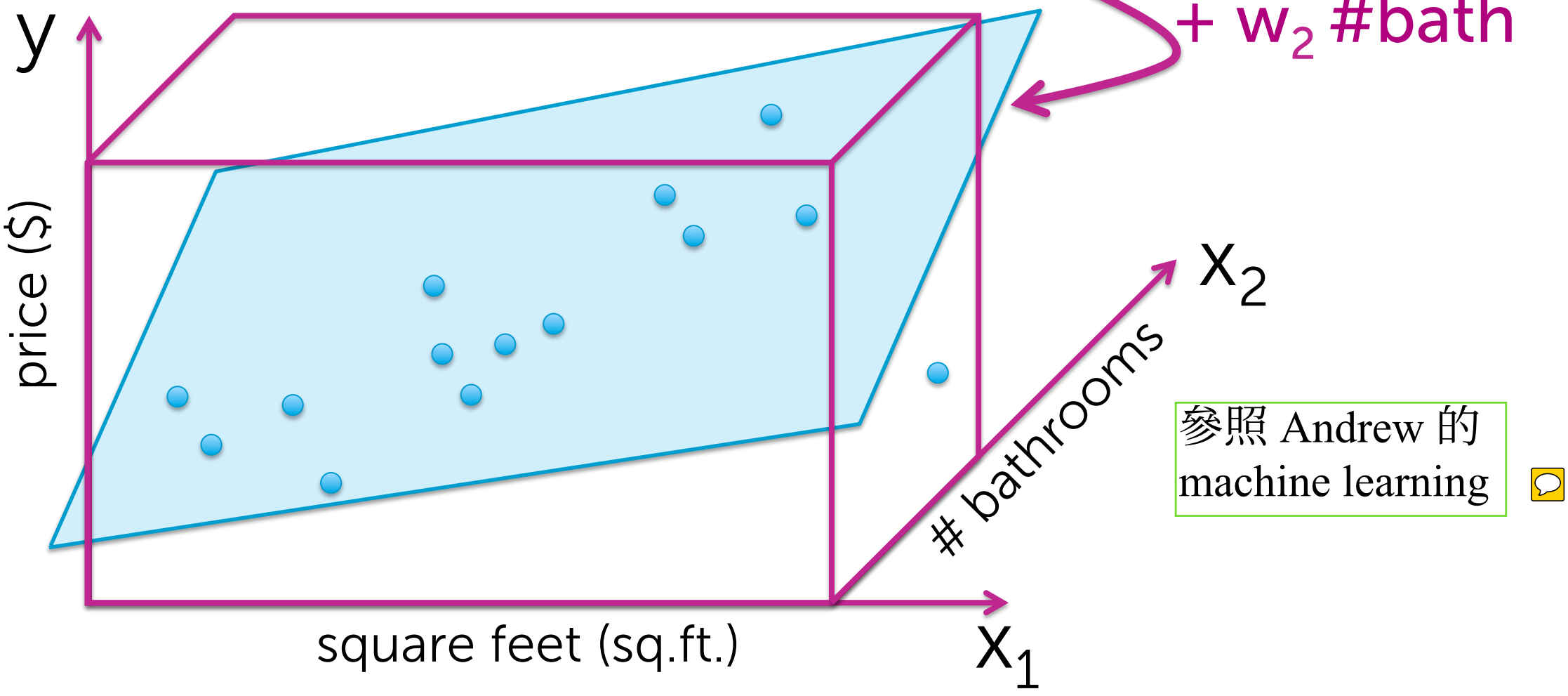
Adding other features

Predictions just based on house size



Add more features

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



參照 Andrew 的 machine learning

How many features to use?

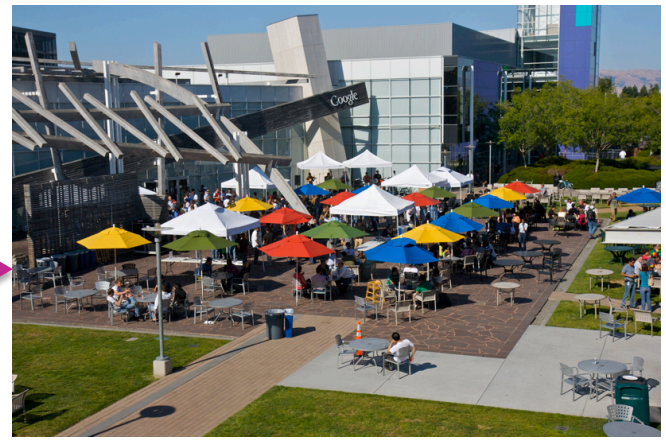
- Possible choices:
  - Square feet
  - # bathrooms
  - # bedrooms
  - Lot size
  - Year built
  - …
- **See Regression Course!**

Other regression examples

後續沒有 note

Salary after ML specialization

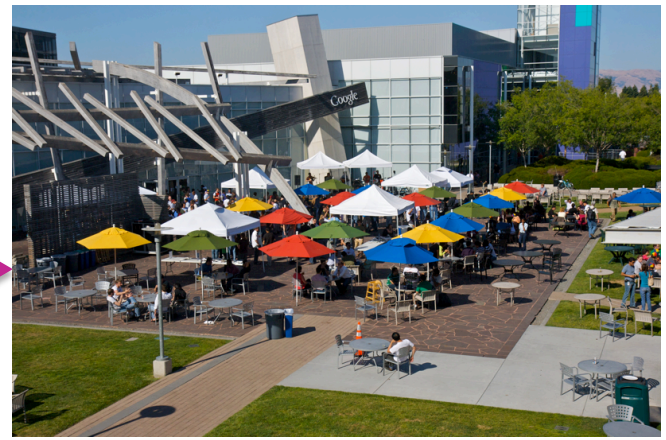
hard work



- 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

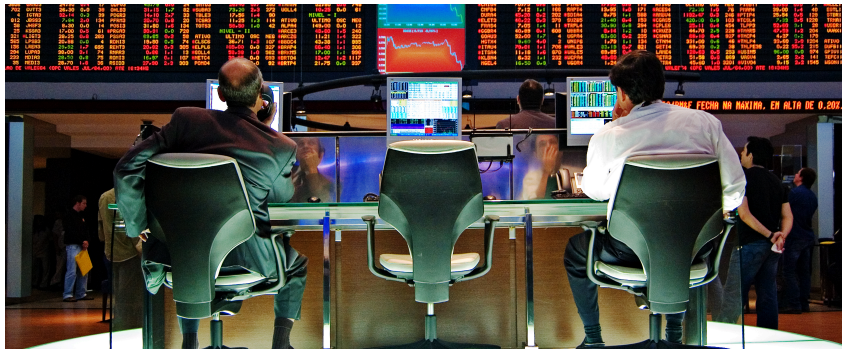


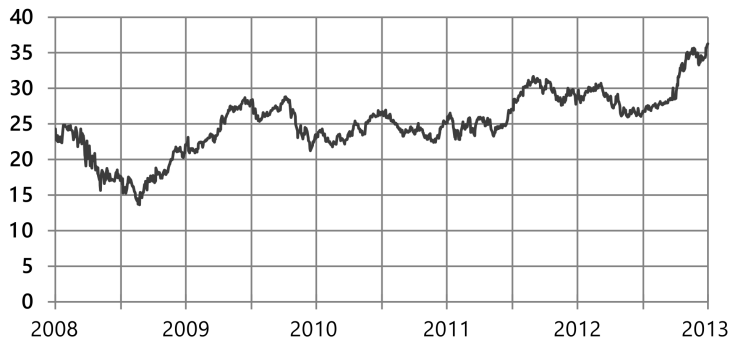
$$\hat{y} = \hat{w}_0 + \hat{w}_1 \text{ performance} + \hat{w}_2 \text{ capstone} + \hat{w}_3 \text{ forum}$$

informed by other students who completed specialization

Stock prediction

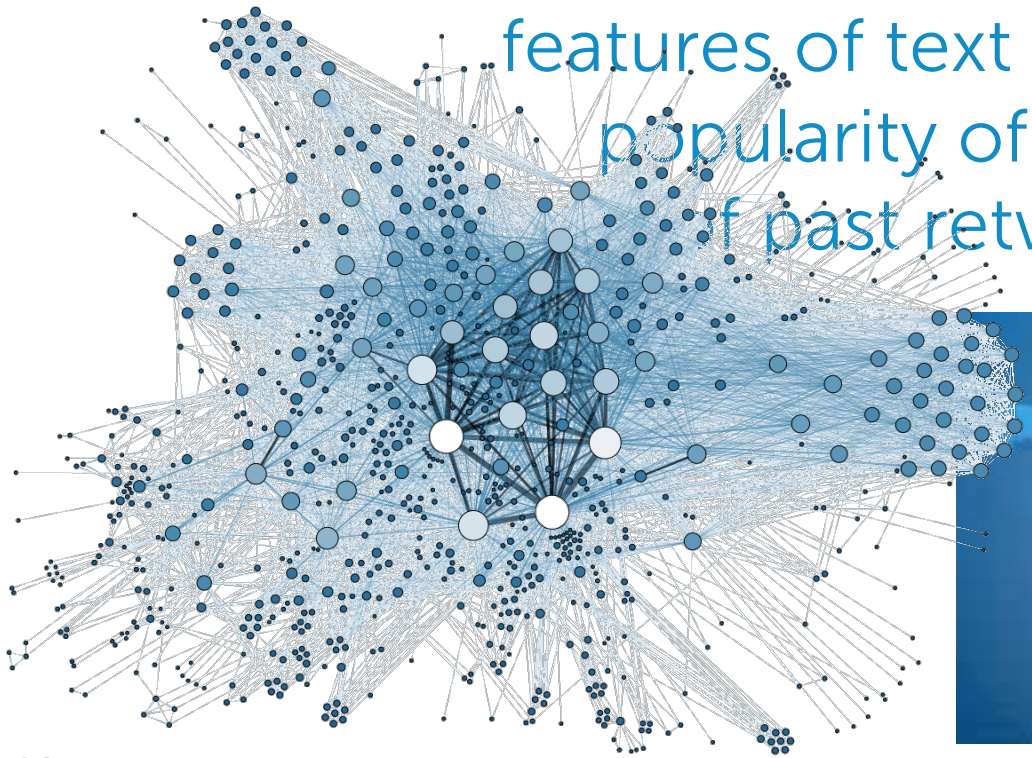
- Predict the price of a stock
- Depends on
  - Recent history of stock price
  - News events
  - Related commodities





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