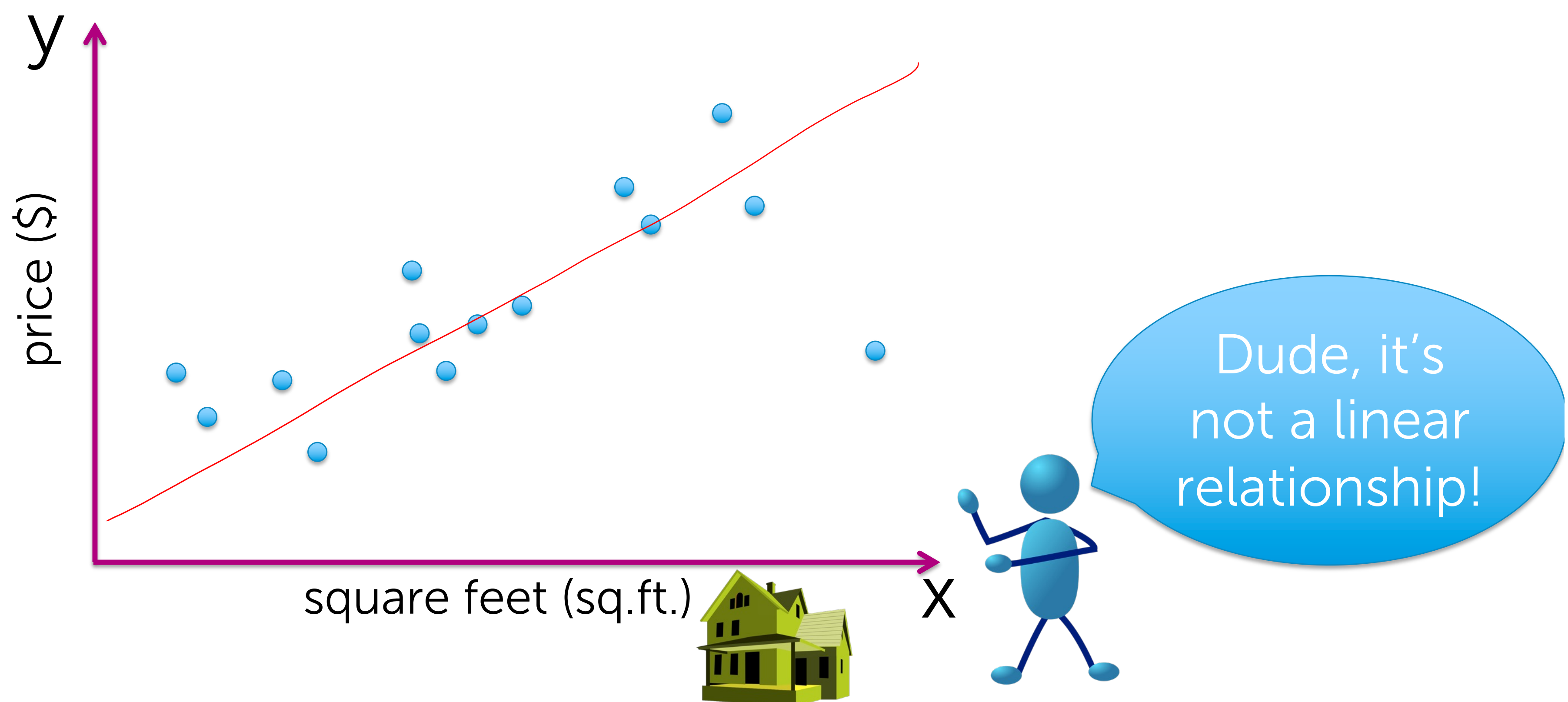
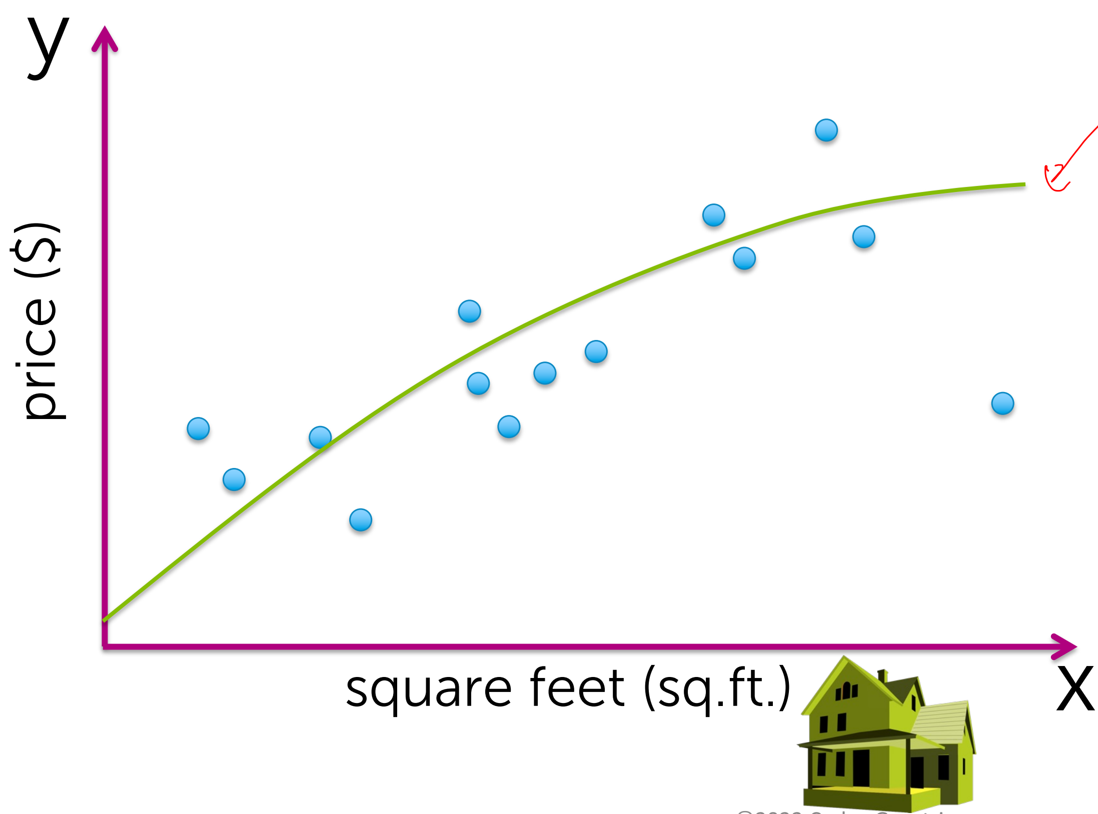
Bias-Variance Tradeoff

CS229: Machine Learning
Carlos Guestrin
Stanford University
Slides include content developed by and co-developed with Emily Fox

Fit data with a line or … ?

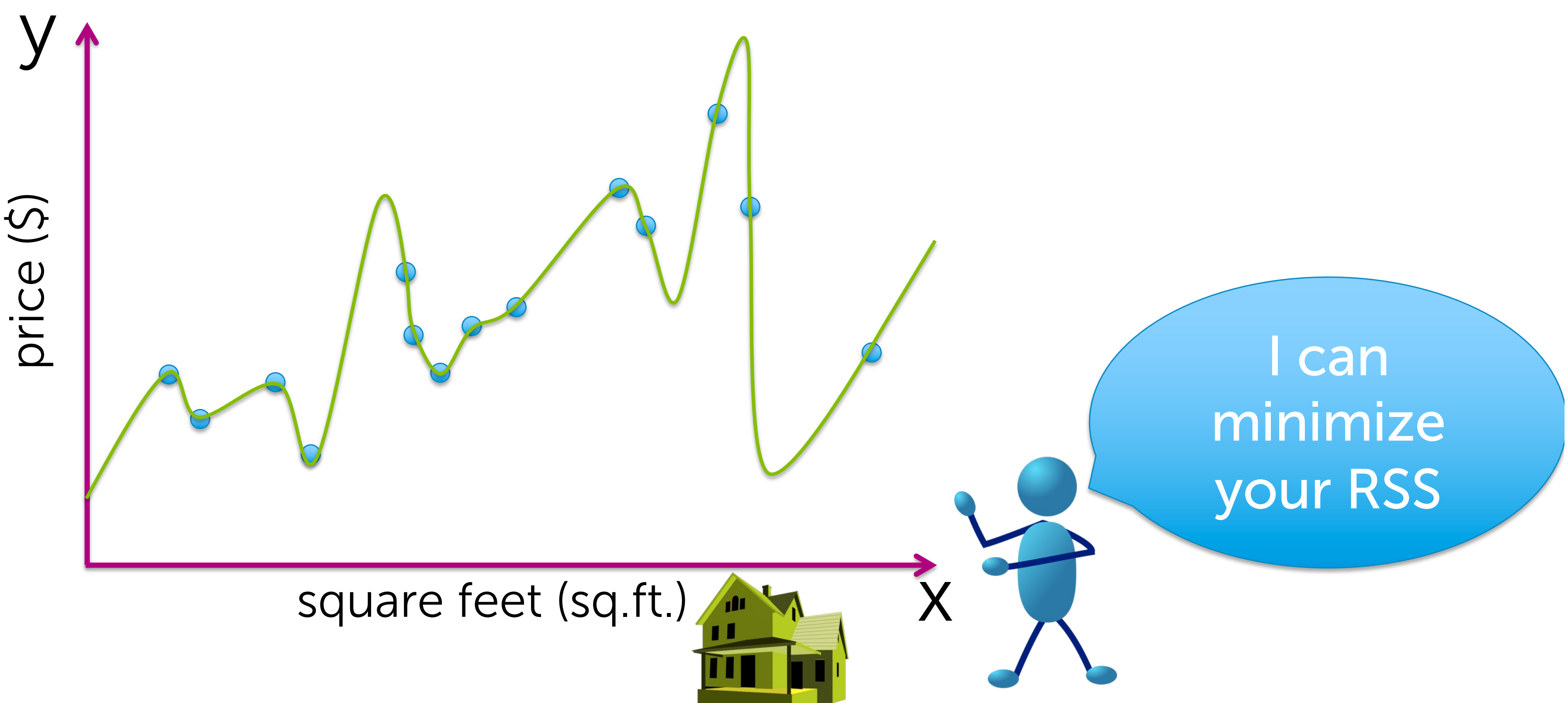


What about a quadratic function?

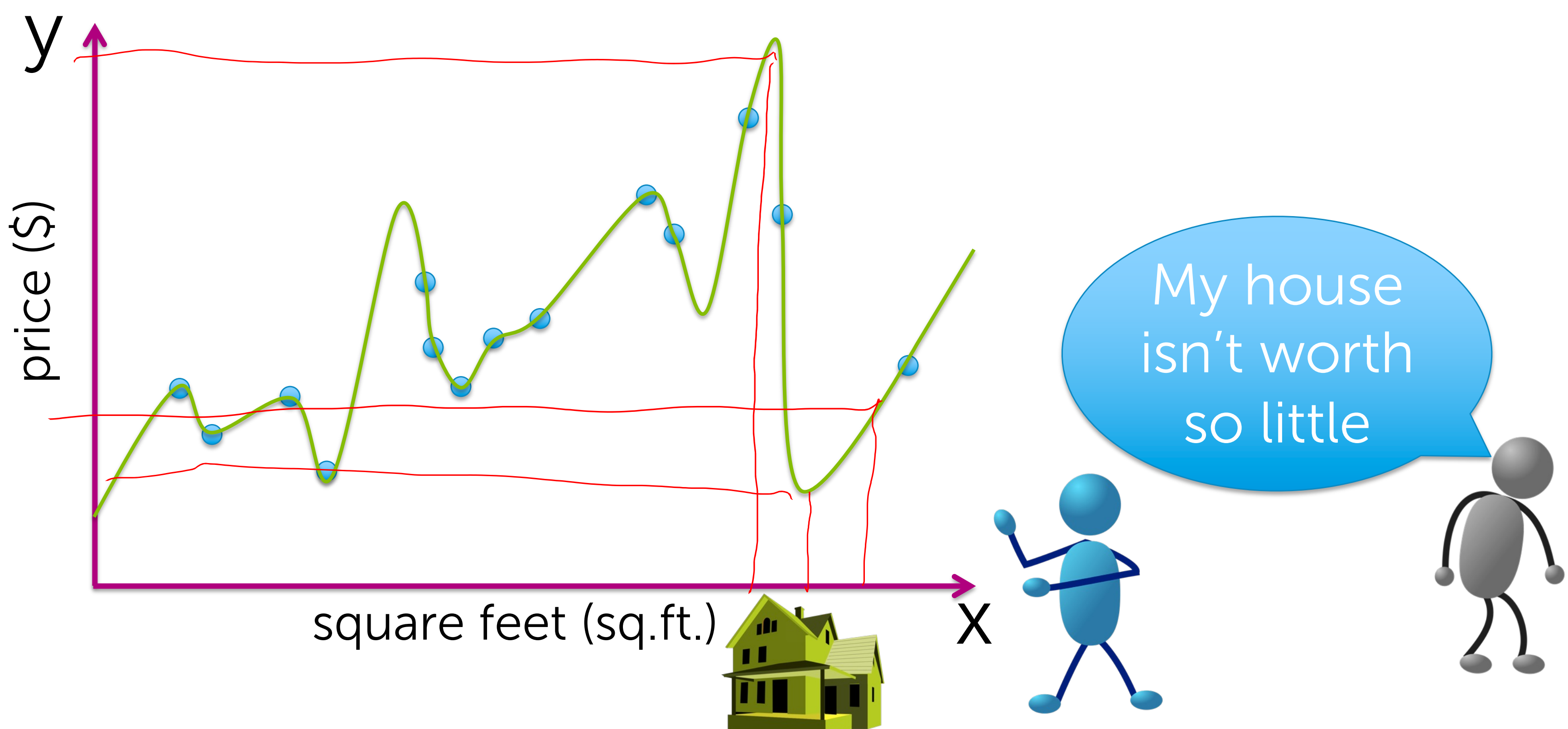


right fit?

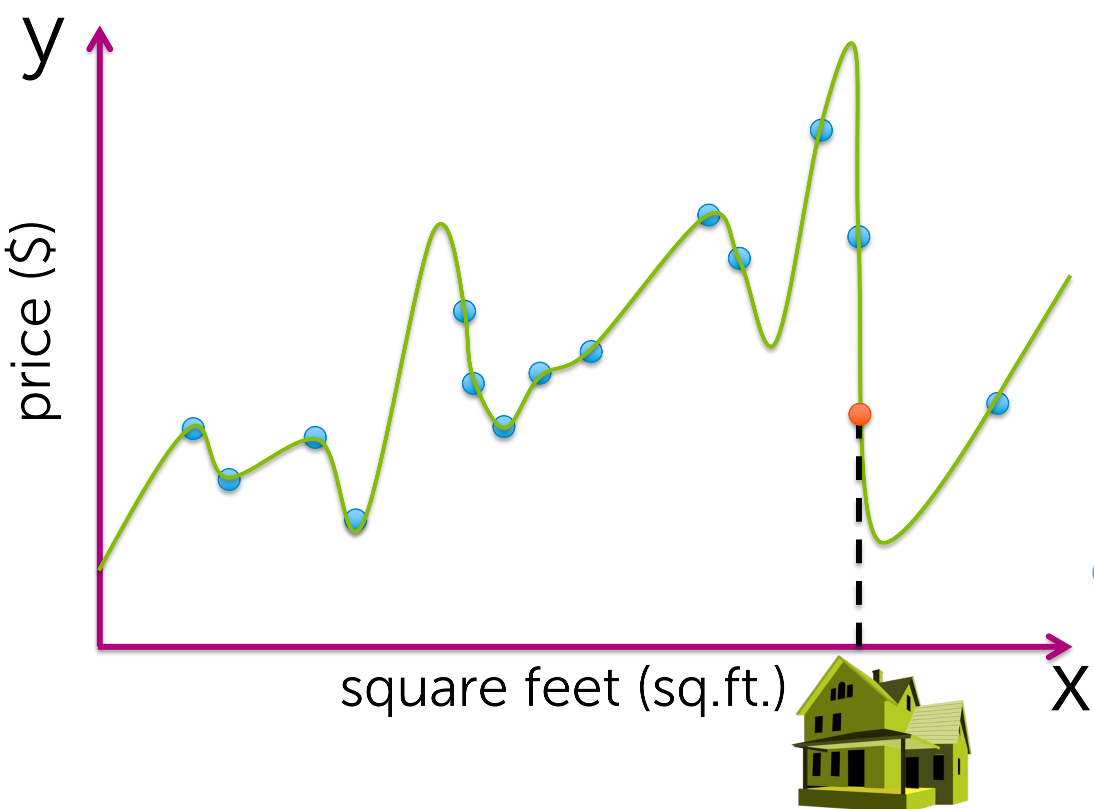
Even higher order polynomial



Do you believe this fit?



Do you believe this fit?



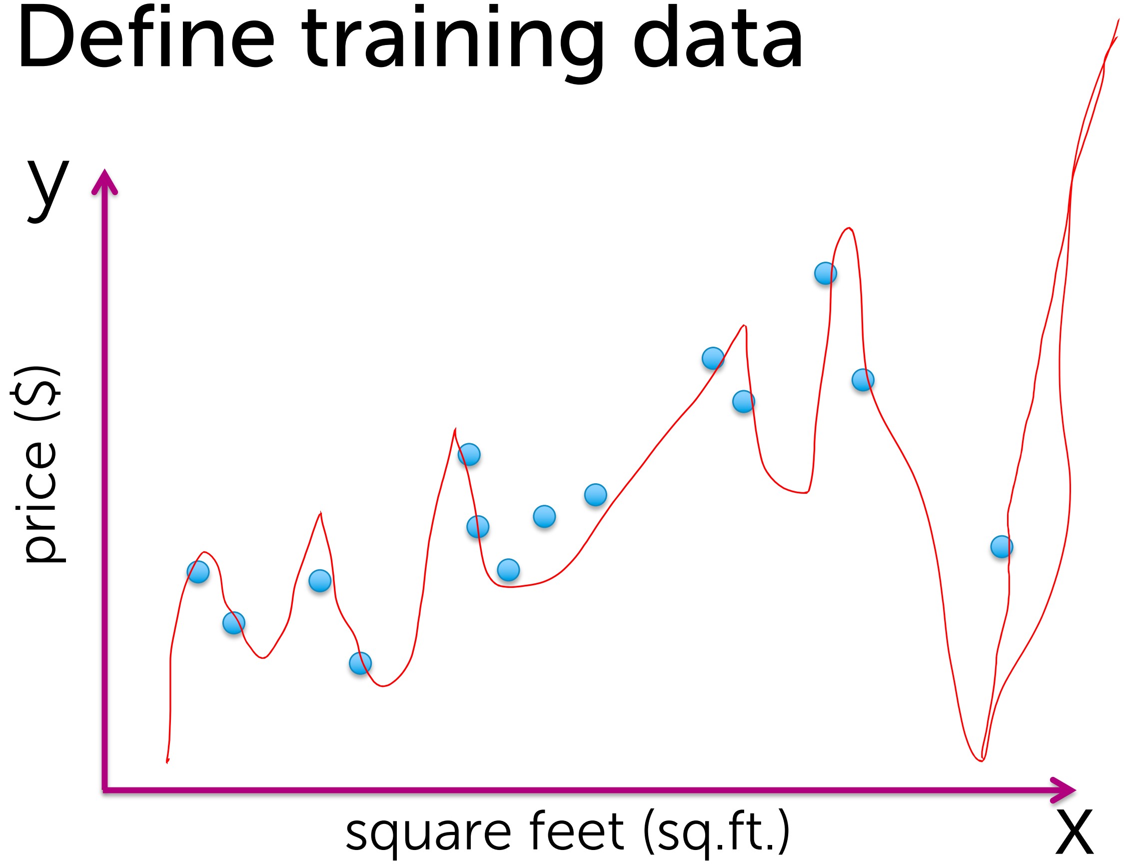
Minimizes RSS,
but bad predictions

"Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful." George Box, 1987.

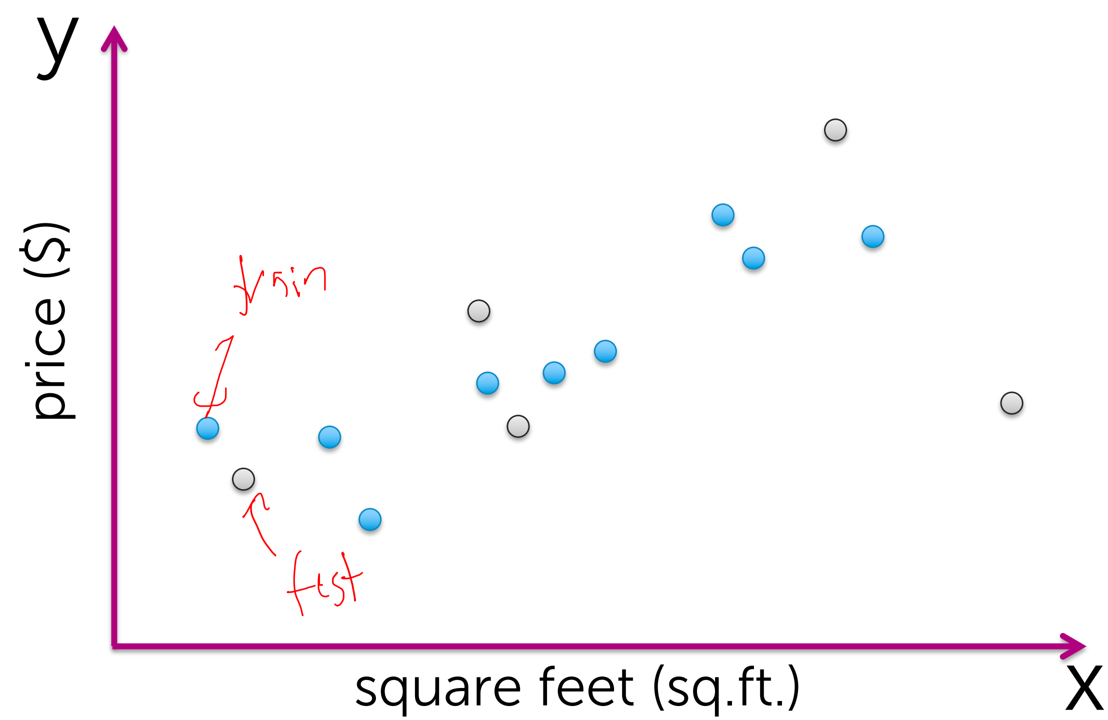
Assessing the loss

Assessing the loss
Part 1: Training error

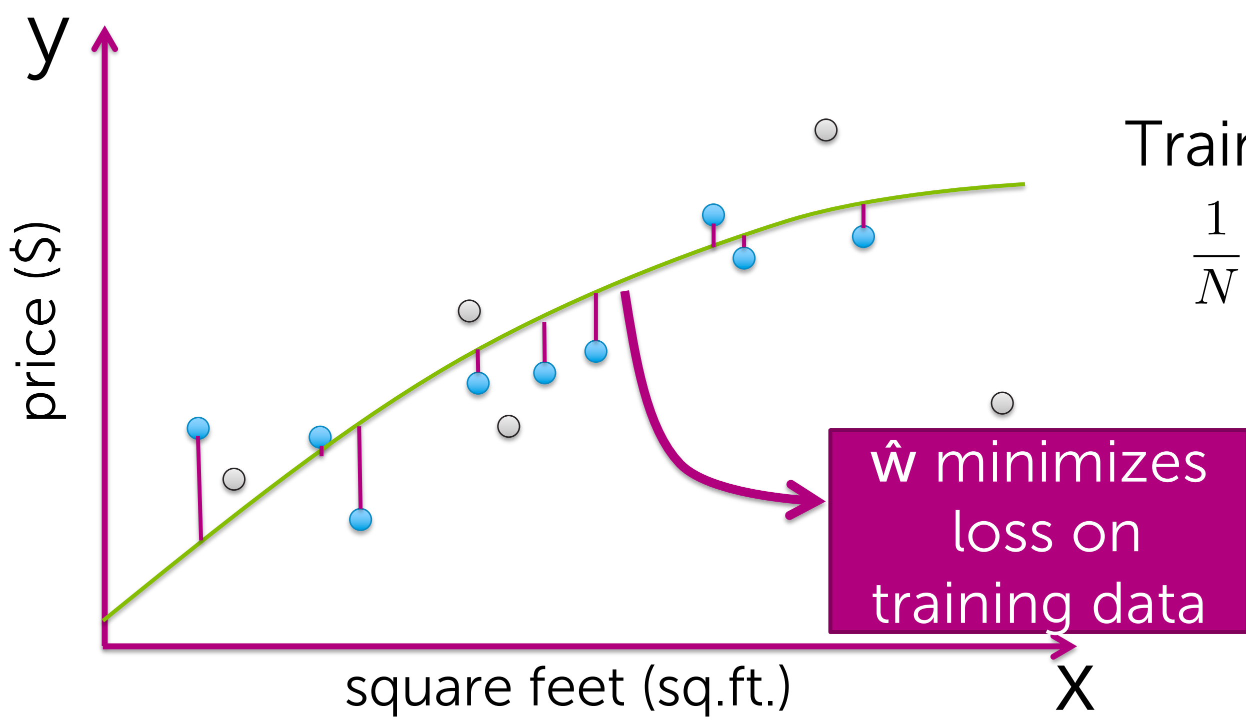
Define training data



Define training data



Example:
Fit quadratic to minimize RSS

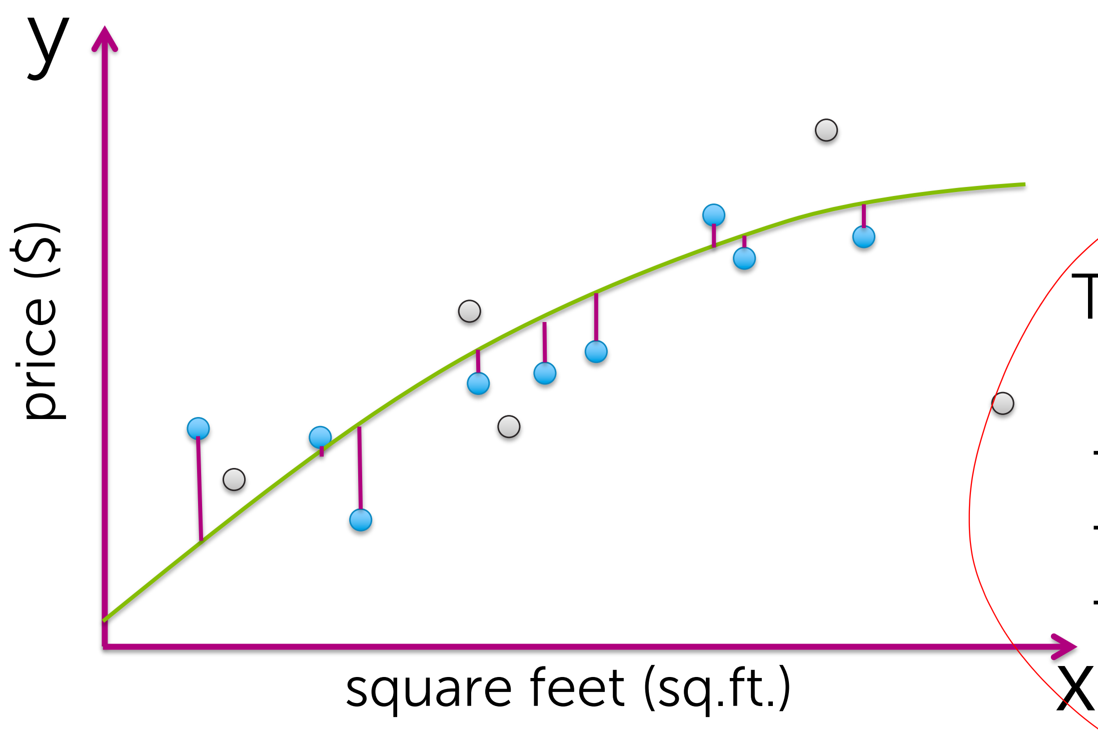


Training error ($\hat{w}$) =

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - f_{\hat{w}}(\mathbf{x}_i))^2$$

Example:

Use squared error loss $(y-f_{\hat{w}}(x))^2$

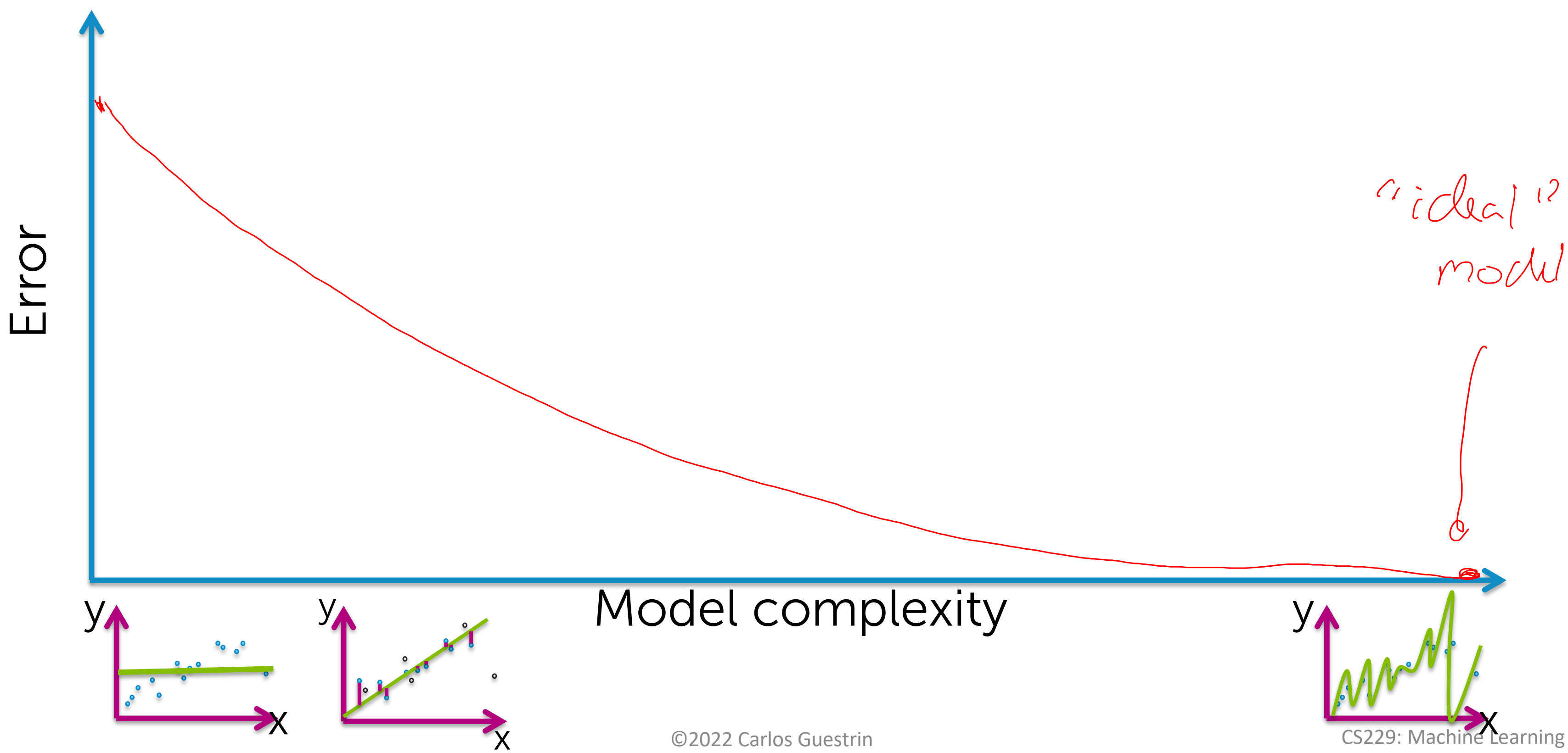


Training error $(\hat{w}) = 1/N *$

$$[(\$_{train\ 1}-f_{\hat{w}}(sq.ft._{train\ 1}))^2$$
$$+ (\$_{train\ 2}-f_{\hat{w}}(sq.ft._{train\ 2}))^2$$
$$+ (\$_{train\ 3}-f_{\hat{w}}(sq.ft._{train\ 3}))^2$$

\+ ... include all training houses]

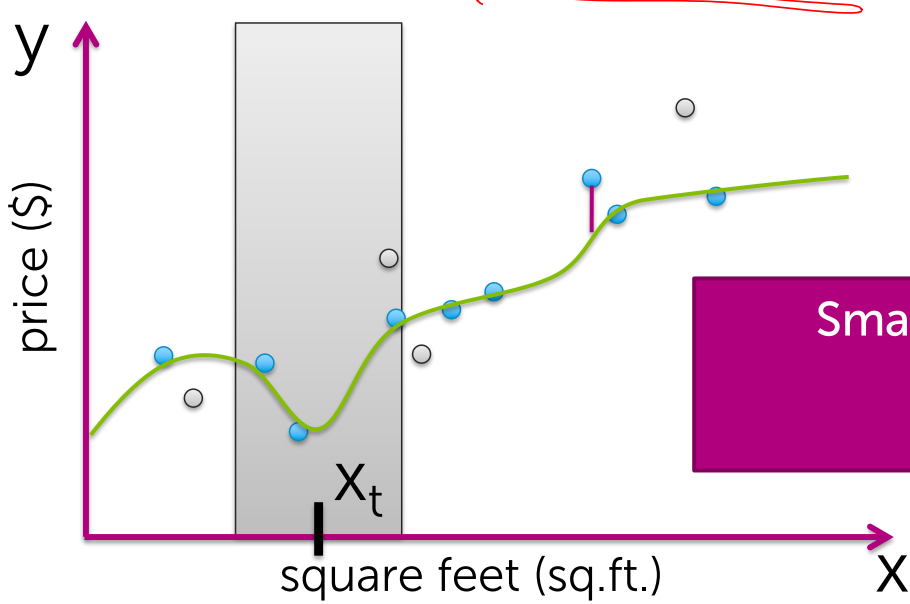
Training error vs. model complexity



Is training error a good measure of predictive performance?

Issue:

Training error is overly optimistic…$\hat{w}$ was fit to training data



Small training error ≠ good predictions
(unless training data includes everything you might ever see)

Assessing the loss
Part 2: Generalization (true) error

Generalization error

Really want estimate of loss over all possible ( ,$) pairs

Lots of houses
in neighborhood,
but not in dataset

Generalization error definition

Really want estimate of loss over all possible (, $) pairs

Formally:

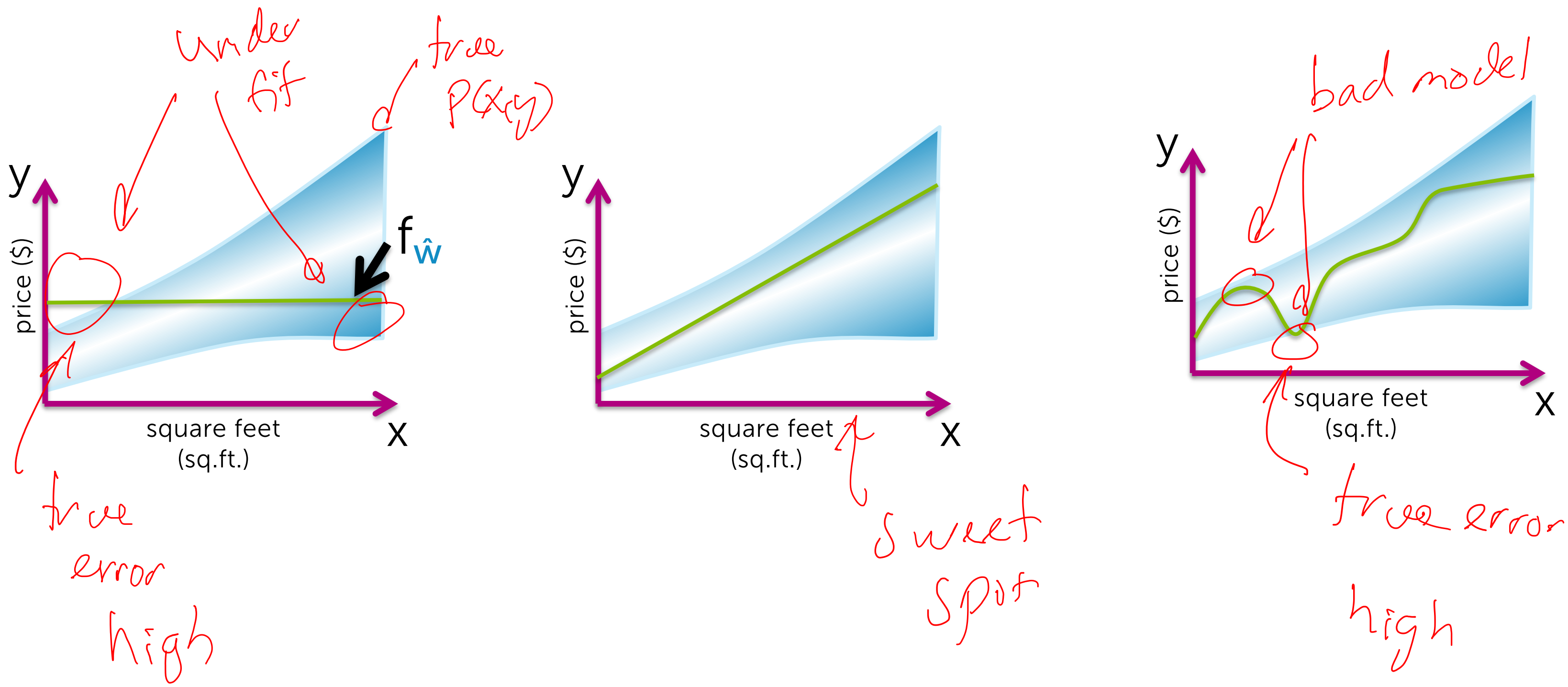
average over all possible
(**x**,y) pairs weighted by
how likely each is

$$\text{generalization error} = E_{\mathbf{x},y}[L(y, f_{\hat{\mathbf{w}}}(\mathbf{x}))]$$

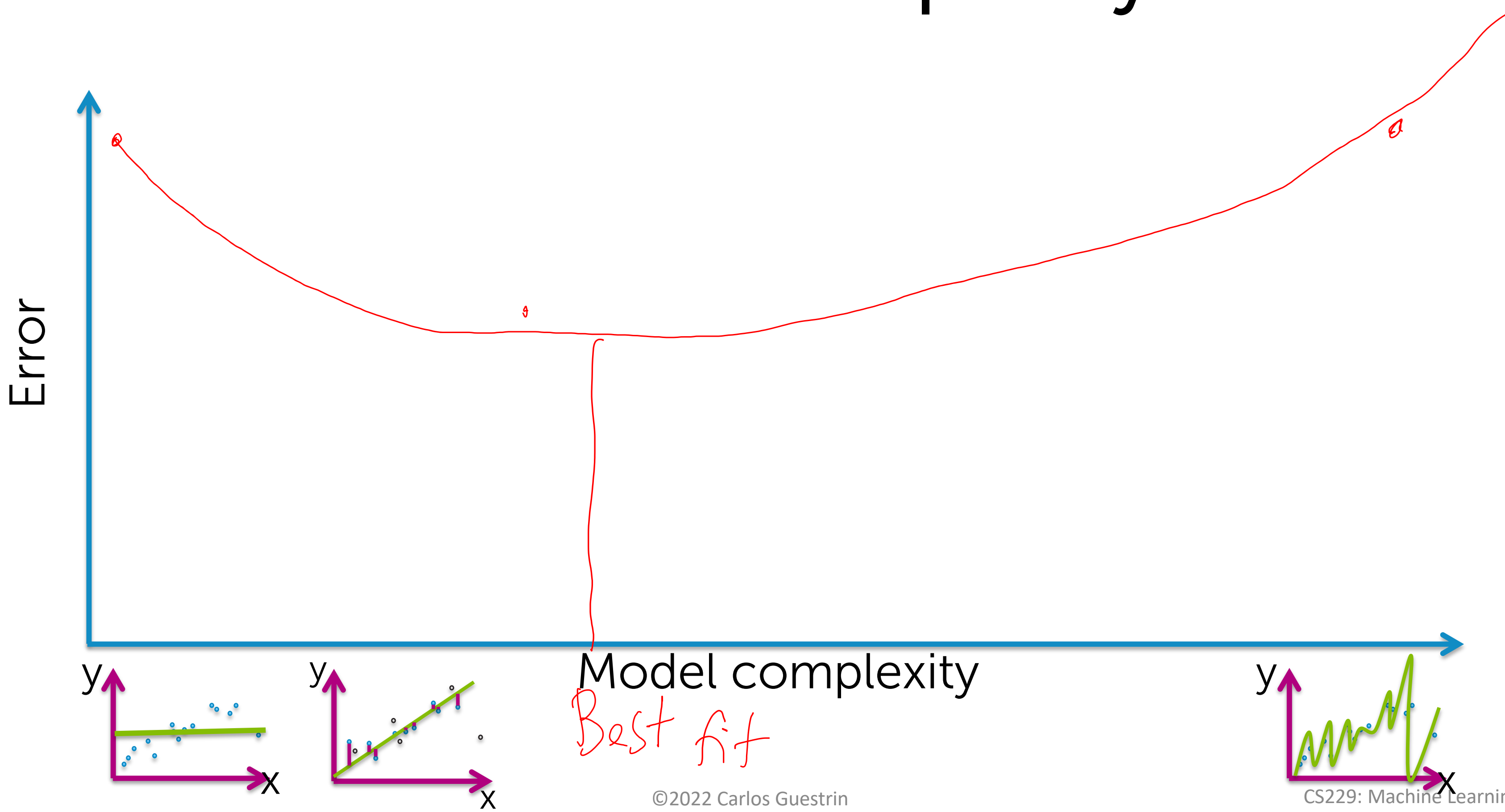
↳ true prob. p(x,y)

fit using training data

Generalization error vs. model complexity



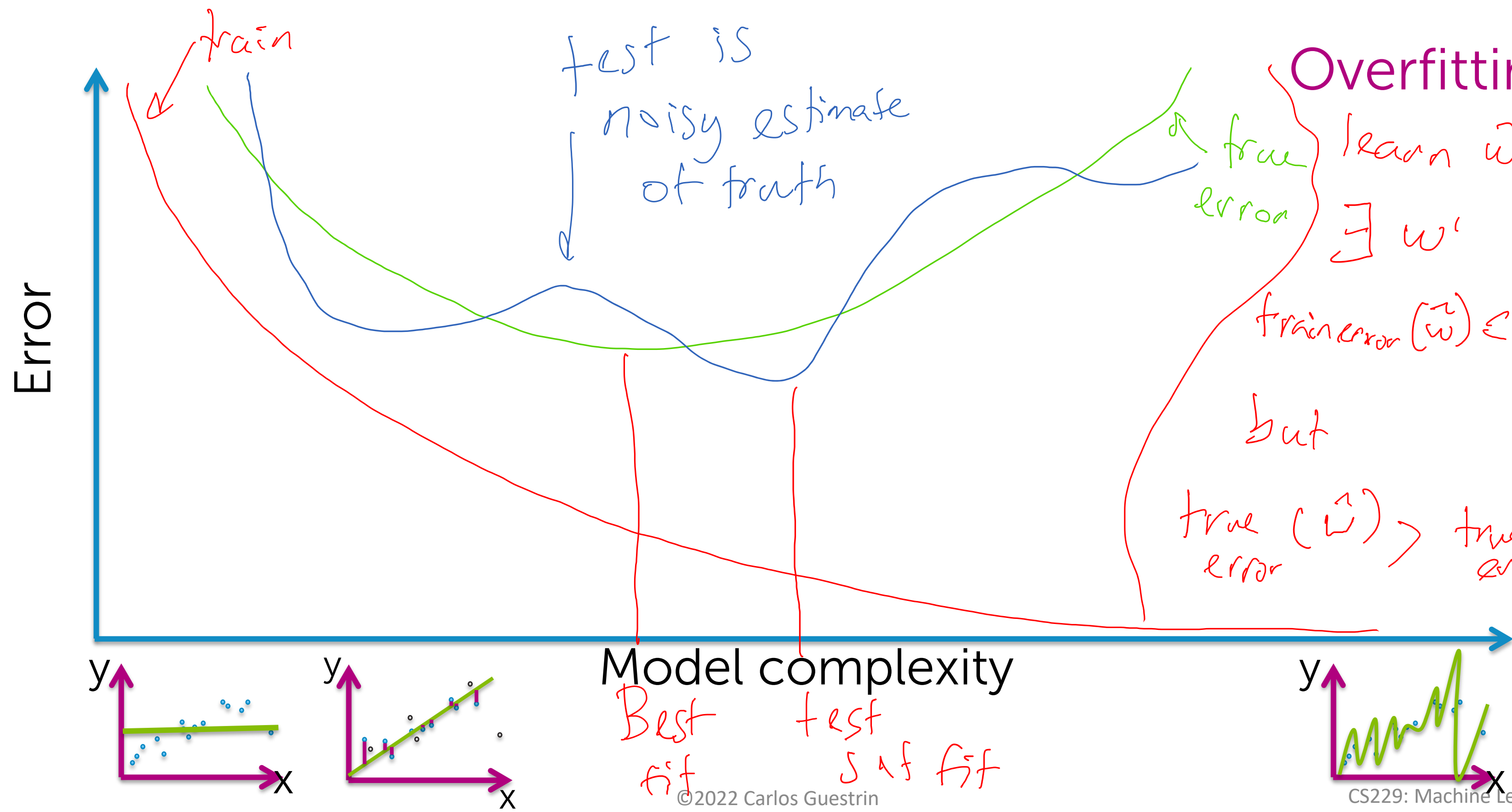
True error vs. model complexity



Assessing the loss
Part 3: Test error

↙ Comes from $p(x,y)$

Training, true, test error vs. model complexity



Overfitting if:

learn $\hat{w}$,

$\exists\ w'$

$\text{train error}(\hat{w}) < \text{train error}(w')$

but

$\text{true error}(\hat{w}) > \text{true error}(w')$

3 sources of error + the bias-variance tradeoff

3 sources of error

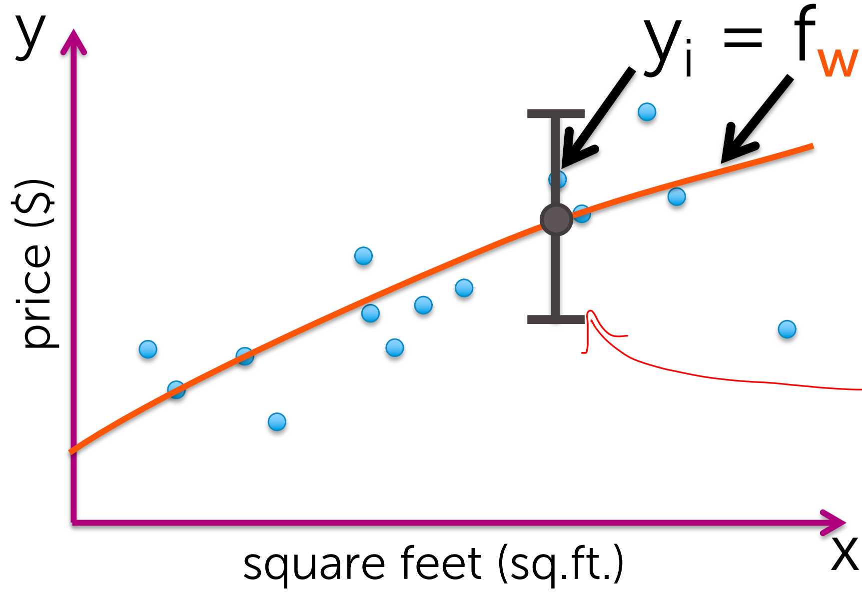
In forming predictions, there are 3 sources of error:

1. Noise
2. Bias ← how well can model fit data on avg.
3. Variance ← how much model would change if I change dataset samples

Data inherently noisy

$$y_i = f_{w(true)}(x_i) + \varepsilon_i$$

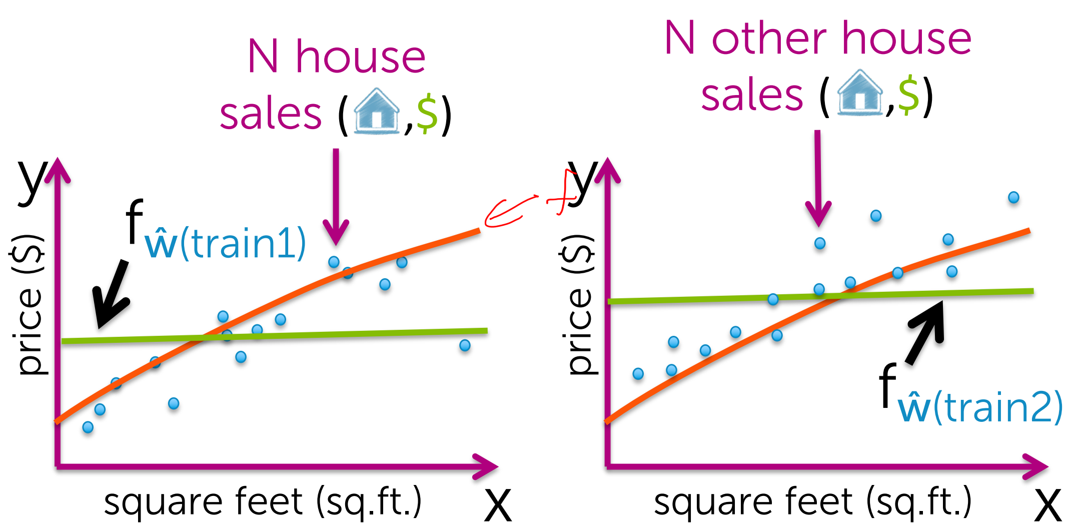
independent zero mean noise



unavoidable loss

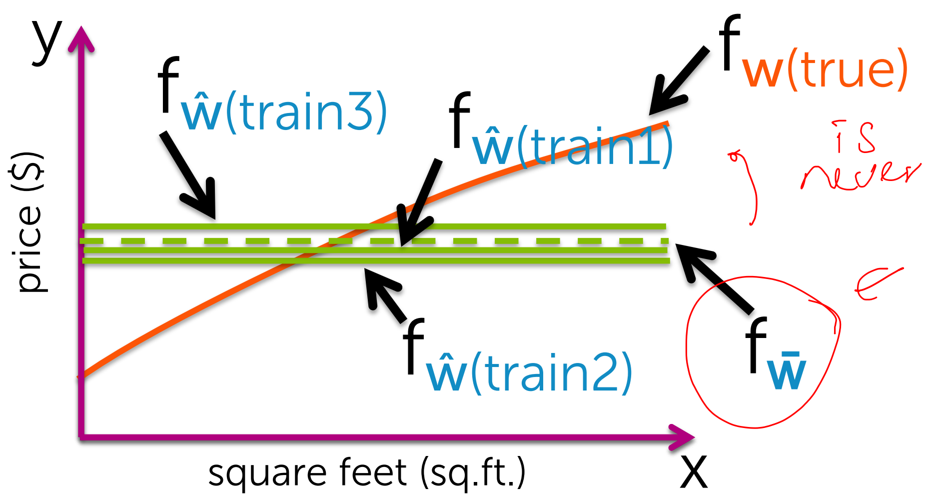
Bias contribution

Suppose we fit a constant function



Bias contribution

Over all possible size N training sets,
what do I expect my fit to be?

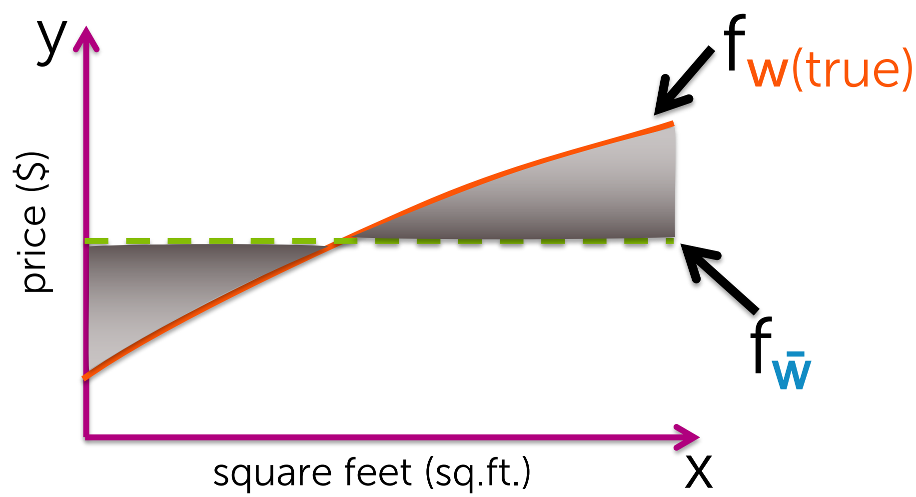


is never flexible enough ⇒ high bias

← avg fit

Bias contribution

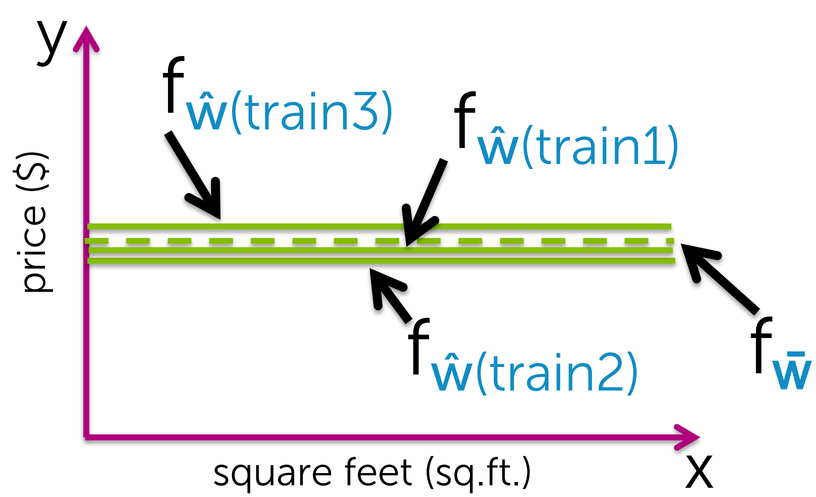
$\text{Bias}(x) = f_{w(true)}(x) - f_{\bar{w}}(x)$ ← Is our approach flexible enough to capture $f_{w(true)}$?
If not, error in predictions.



low complexity
→
high bias

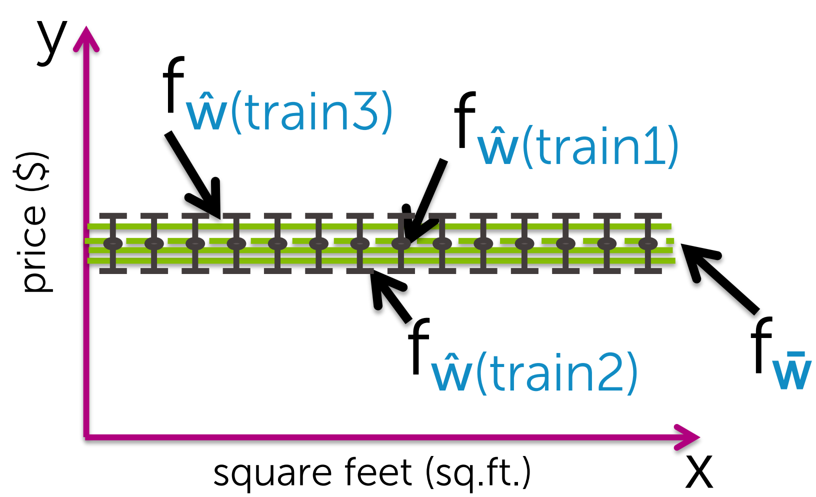
Variance contribution

How much do specific fits vary from the expected fit?



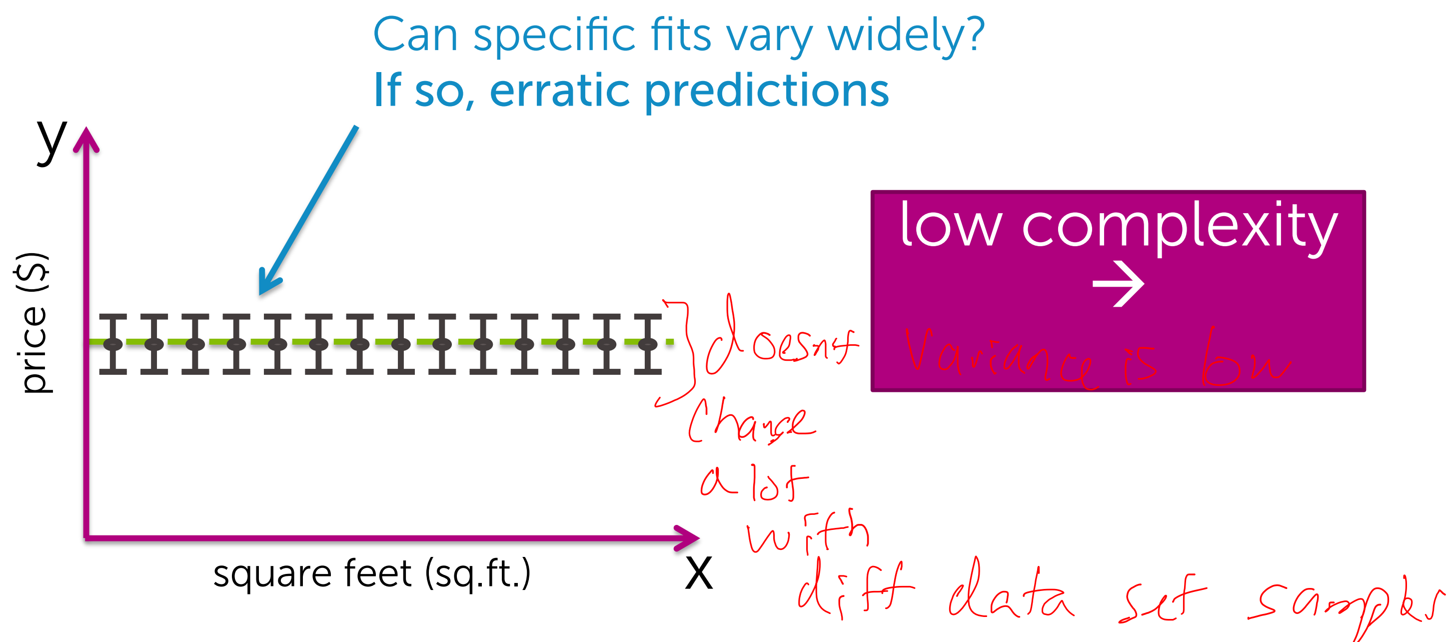
Variance contribution

How much do specific fits vary from the expected fit?



Variance contribution

How much do specific fits vary from the expected fit?

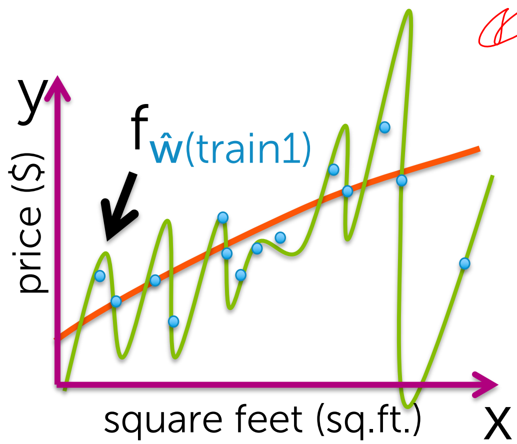


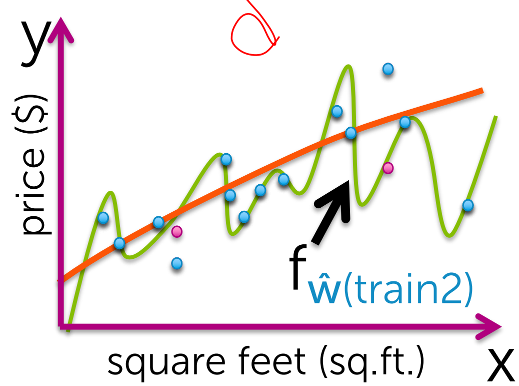
low complexity
→
Variance is low

Variance of high-complexity models

Assume we fit a high-order polynomial

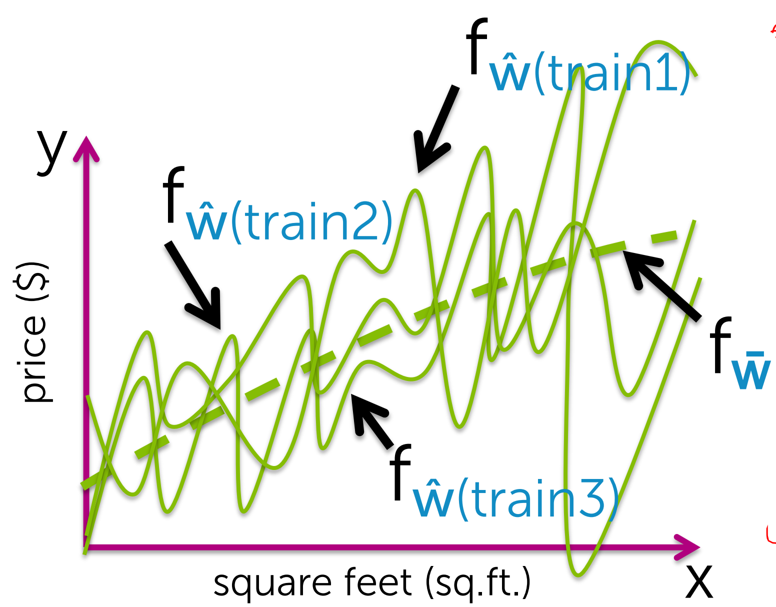
$f_{\hat{w}}$ changes a lot with diff samples of data





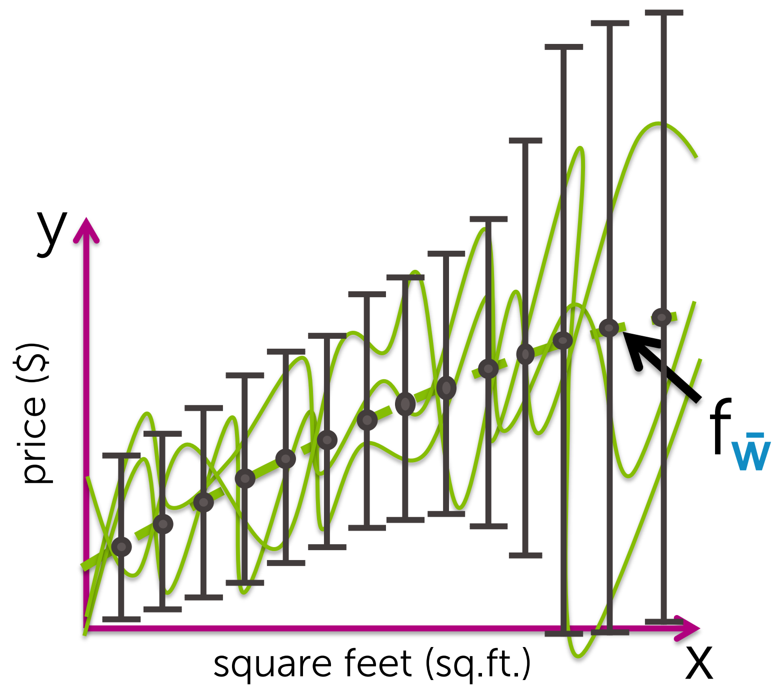
Variance of high-complexity models

Suppose we fit a high-order polynomial



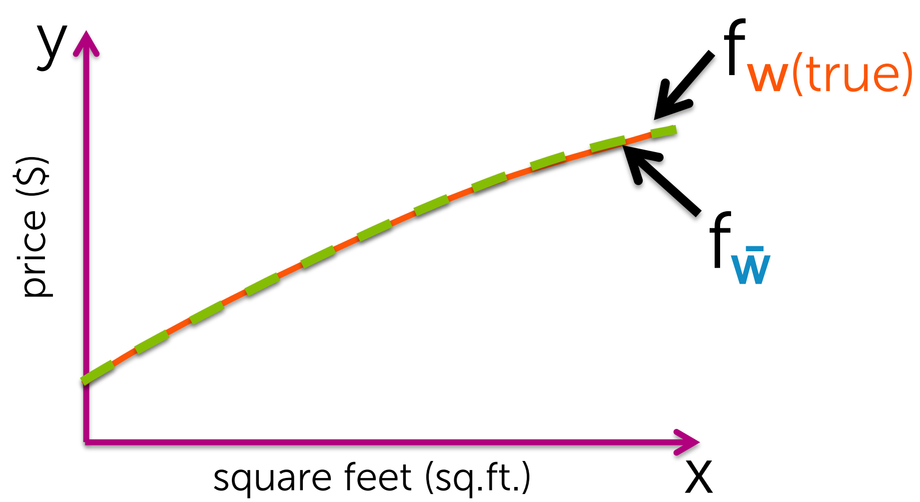
high variance

Variance of high-complexity models



high complexity
→
high variance

Bias of high-complexity models



high complexity
→
low bias

Sum of 3 sources of error

True error =

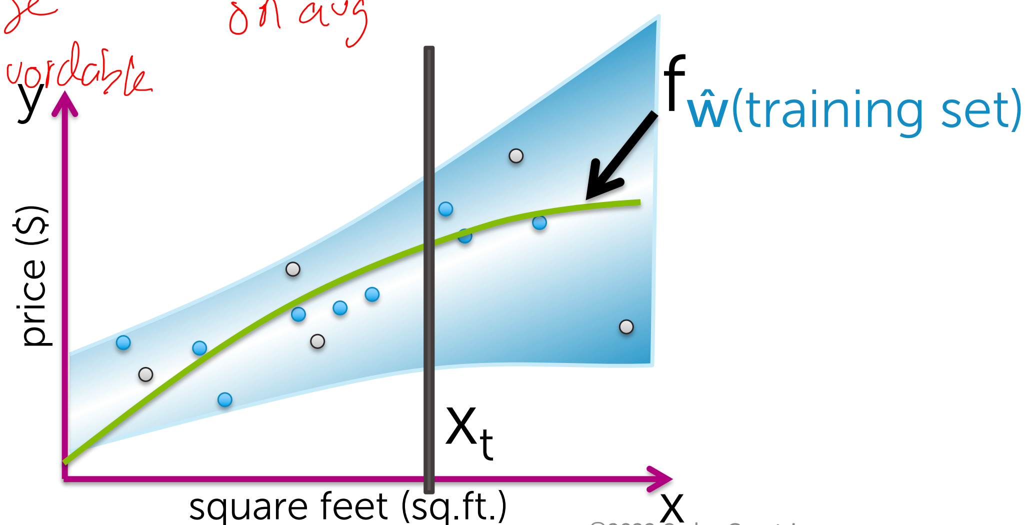
Average squared error at $\mathbf{x}_t$

$$= \sigma^2 + [\text{bias}(f_{\hat{w}}(\mathbf{x}_t))]^2 + \text{var}(f_{\hat{w}}(\mathbf{x}_t))$$

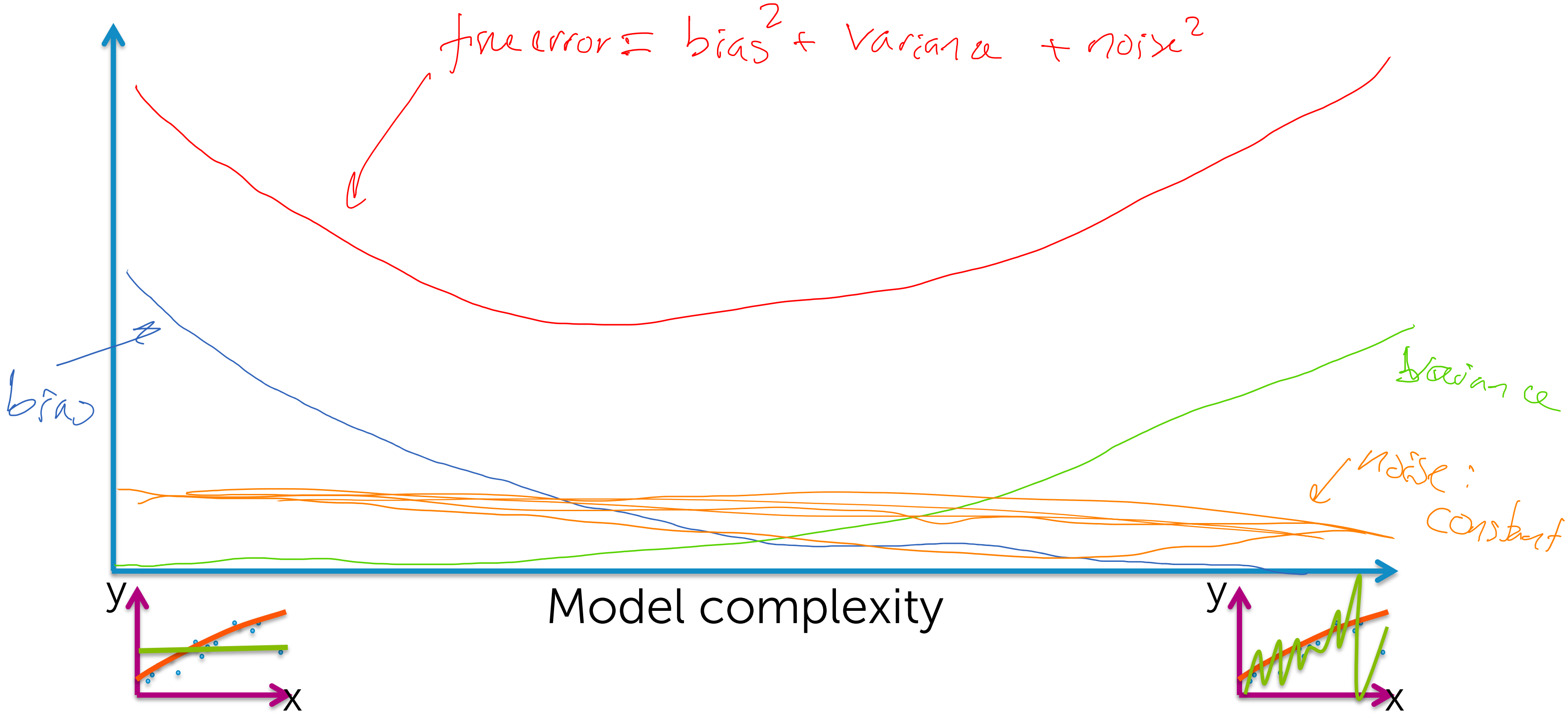
noise unavoidable

fit on avg

variance: variability predictions across datasets

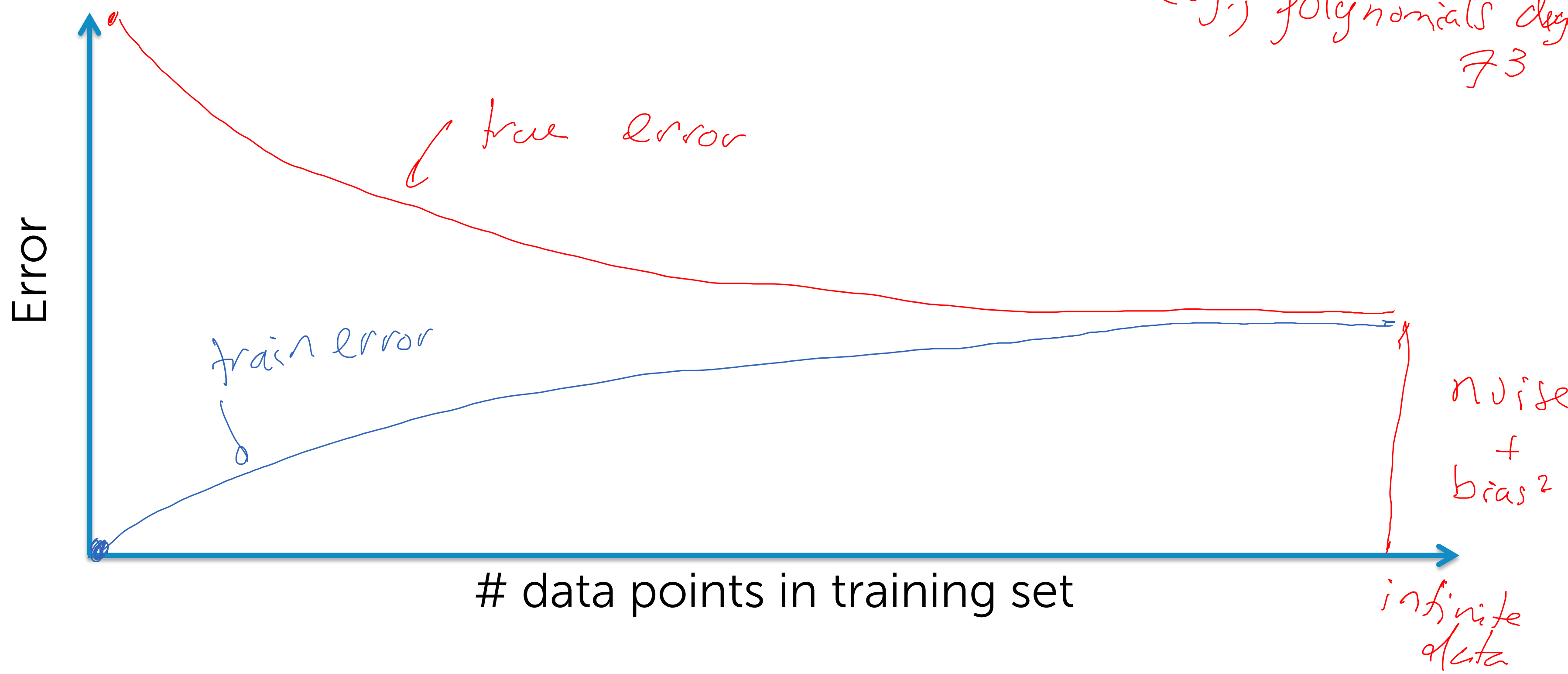


Bias-variance tradeoff



Error vs. amount of data

for a fixed model complexity.
e.g., polynomials degree 73



Why 3 sources of error?
A formal derivation

Deriving expected prediction error

Expected prediction error

$= E_{\text{train}}$ [generalization error of $\hat{\mathbf{w}}(\text{train})$]

$= E_{\text{train}} \left[E_{\mathbf{x},y} [L(y, f_{\hat{\mathbf{w}}(\text{train})}(\mathbf{x}))] \right]$

1. Look at specific $\mathbf{x}_t$
2. Consider $L(y, f_{\hat{\mathbf{w}}}(\mathbf{x})) = (y - f_{\hat{\mathbf{w}}}(\mathbf{x}))^2$

Expected prediction error at $\mathbf{x}_t$

$= E_{\text{train}, y_t} \left[(y_t - f_{\hat{\mathbf{w}}(\text{train})}(\mathbf{x}_t))^2 \right]$

Simplifying Notation

- Expected prediction error at $\mathbf{x}_t$

  $= E_{\text{train}, y_t}\left[(y_t - f_{\hat{w}(\text{train})}(\mathbf{x}_t))^2\right]$

- Simple (and abusive ☺) notation:
  - $y_t \rightarrow y$
  - $f_{w(\text{true})}(\mathbf{x}_t) \rightarrow f \leftarrow$ true
  - $f_{\hat{w}(\text{train})}(\mathbf{x}_t) \rightarrow \hat{f} \leftarrow$ learned
  - $E_{\text{train}}\left[f_{\hat{w}(\text{train})}(\mathbf{x}_t)\right] = f_{\bar{w}}(\mathbf{x}_t) \rightarrow \bar{f}$ learned on avg.

Deriving expected prediction error

$$E[ab] = E[a]\,E[b]$$

independent

Expected prediction error at $\mathbf{x}_t$

$$= E_{\text{train},y_t}\left[(y_t - f_{\hat{w}(\text{train})}(\mathbf{x}_t))^2\right] = E_{\text{train}}[(y - \hat{f})^2] =$$

$$= E_{\text{train}}\left[((y - f) + (f - \hat{f}))^2\right]$$

$$= \underbrace{E[(y-f)^2]}_{\text{by definition: noise } \sigma^2} + 2\,\underbrace{E_{\text{train}}[(y-f)(f-\hat{f})]}_{\cancel{E_{\text{train}}[(y-f)]}\; E_{\text{train}}[(f-\hat{f})]} + E_{\text{train}}[(f-\hat{f})^2]$$

0: zero mean noise

$\triangleq$ MSE$(\hat{f})$

mean squared error

$$= \sigma^2 + \text{MSE}(\hat{f})$$

Equating MSE with bias and variance

$$\mathrm{MSE}[f_{\hat{w}(\mathrm{train})}(\mathbf{x}_t)]$$

$$= E_{\mathrm{train}}[(f - \hat{f})^2]$$

$$= E_{\mathrm{train}}[((f - \bar{f}) + (\bar{f} - \hat{f}))^2]$$

$$= \underbrace{E_{\mathrm{train}}[(f-\bar{f})^2]}_{\text{by defn. } \mathrm{bias}^2} + \cancel{2 E_{\mathrm{train}}[(f-\bar{f})(\bar{f}-\hat{f})]} + E_{\mathrm{train}}[(\overbrace{\bar{f}}^{\text{mean}} - \overbrace{\hat{f}}^{\text{random var.}})^2]$$

$$\cancel{2 E_{\mathrm{train}}[(f-\bar{f})]\, \underbrace{E_{\mathrm{train}}[\bar{f}-\hat{f}]}_{E[\bar{f}] - E[\hat{f}]}} \quad 0$$

$$= \mathrm{Var}(\hat{f})$$

$$= \mathrm{bias}^2 + \mathrm{Var}$$

Putting it all together

Expected prediction error at $\mathbf{x}_t$

$$= \sigma^2 + \text{MSE}\left[f_{\hat{w}}(\mathbf{x}_t)\right]$$

$$= \sigma^2 + [\text{bias}(f_{\hat{w}}(\mathbf{x}_t))]^2 + \text{var}(f_{\hat{w}}(\mathbf{x}_t))$$

model fit

flexibility of the model

3 sources of error

Summary of bias-variance tradeoff

What you can do now…

- Contrast relationship between model complexity and train, true and test loss
- Compute training and test error given a loss function for different model complexities
- List and interpret the 3 sources of avg. prediction error
  - Irreducible error, bias, and variance