Week 1

Cogs 109: Data Analysis and Modeling

Fall 2018

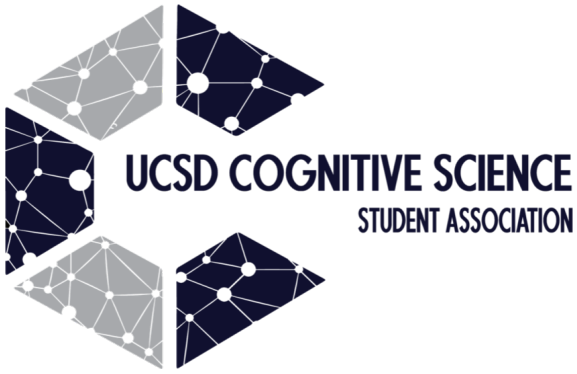
Prof. Eran Mukamel

Check out CSSA's 1st GBM!

Cognitive Science Student Association

Join us and learn more about cognitive science opportunities!!

Free food!



October 11, 6:30-8:30pm

@ PC Theater

New extra credit policy

- Extra credit: You may earn up to 2% (total) by participating as a SONA research subject, or writing a 2-page (700 words) summary of relevant research paper (1% extra credit per summary; you MUST get prior approval from a TA or Prof. Mukamel for your chosen paper. Research paper summaries must be completed before the end of Week 7). Info about SONA: http://www.psychology.ucsd.edu/undergraduate-program/undergraduate-resources/sona)

New attendance policy

- 10% - Participation in section and lectures. Attend 6/9 sections to receive full credit. Alternatively: If you are not able to attend sections due to schedule conflicts, you may choose to instead increase the weight of homework (35%), midterm (22.5%) and final exam (22.5%). At the end of the quarter, we will calculate your average using both schemes and assign you whichever grade is higher.

How do we analyze data and build a model?

$$Y = f(X) + \epsilon.$$

- Given X, Y we want to figure out f
- By estimating f, we can perform *prediction* and/or *inference*

$$\hat{Y} = \hat{f}(X) + \epsilon$$

Predicted values of Y

Inferred relationship between Y, X

Data sets are often tables

p=2 predictors and outcomes

[n x p] matrix

	A	B	C	D
	VarName1	Education	Seniority	Income3
	NUMBER	NUMBER	NUMBER	NUMBER
		Education	Seniority	Income
1		21.5862...	113.103...	99.9171...
2		18.2758...	119.310...	92.5791...
3		12.0689...	100.689...	34.6787...
4		17.0344...	187.586...	78.7028...
5		19.9310...	20	68.0099...
6		18.2758...	26.2068...	71.5044...
7		19.9310...	150.344...	87.9704...
8		21.1724...	82.0689...	79.8110...
9		20.3448...	88.2758...	90.0063...
10		10	113.103...	45.6555...
11		13.7241...	51.0344...	31.9138...
12		18.6896...	144.137...	96.2829...
13		11.6551...	20	27.9825...
14		16.6206...	94.4827...	66.6017...
15		10	187.586...	41.5319...
16		20.3448...	94.4827...	89.0007...
17		14.1379...	20	28.8163...
18		16.6206...	44.8275...	57.6816...
19		16.6206...	175.172...	70.1050...
20		20.3448...	187.586...	98.8340...
21		18.2758...	100.689...	74.7046...
22		14.5517...	137.931...	53.5321...
23		17.4482...	94.4827...	72.0789...
24		10.4137...	32.4137...	18.5706...
25		21.5862...	20	78.8057...
26		11.2413...	44.8275...	21.3885...
27		19.9310...	168.965...	90.8140...
28		11.6551...	57.2413...	22.6361...
29		12.0689...	32.4137...	17.6135...
30		17.0344...	106.896...	74.6109...

n=30 data points (observations)

X1 X2 Y

A	B	C	D
VarName1	Education	Seniority	Income3
NUMBER	NUMBER	NUMBER	NUMBER
	Education	Seniority	Income
1	21.5862...	113.103...	99.9171...
2	18.2758...	119.310...	92.5791...
3	12.0689...	100.689...	34.6787...
4	17.0344...	187.586...	78.7028...

“College” dataset (available in R)

Categorical variable:
Can take one of several discrete values (e.g. Yes/No, Male/Female, Chocolate/Vanilla/Strawberry)

Quantitative variables:
Take a continuous value (e.g. percent, size…)

		Private	Apps	Accept	Enroll	Top10perc
9	Anderson University	Yes	1216	908	423	19
0	Andrews University	Yes	1130	704	322	14
1	Angelo State University	No	3540	2001	1016	24
2	Antioch University	Yes	713	661	252	25
3	Appalachian State University	No	7313	4664	1910	20
4	Aquinas College	Yes	619	516	219	20
5	Arizona State University Main campus	No	12809	10308	3761	24
6	Arkansas College (Lyon College)	Yes	708	334	166	46
7	Arkansas Tech University	No	1734	1729	951	12
8	Assumption College	Yes	2135	1700	491	23
9	Auburn University-Main Campus	No	7548	6791	3070	25
0	Augsburg College	Yes	662	513	257	12

Parametric models

- Step 1: Assume (or propose) a specific functional form for the relationship between predictors and outcome:

$$f(X) = \beta_0 + \beta_1 X_1$$ "Simple" linear regression

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p.$$ Multiple linear regression

- Step 2: Fit the model parameters (coefficients) using the observed data:

$$Y \approx \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_p X_p.$$

- Example: $\texttt{income} \approx \beta_0 + \beta_1 \times \texttt{education} + \beta_2 \times \texttt{seniority}$.

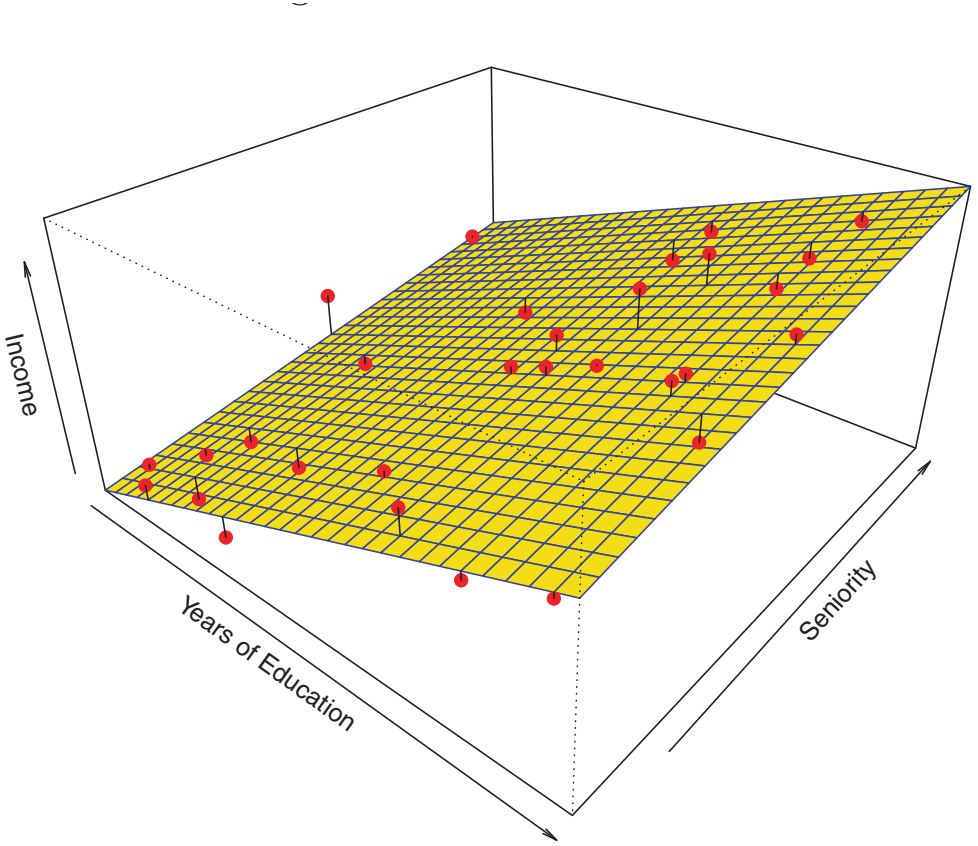


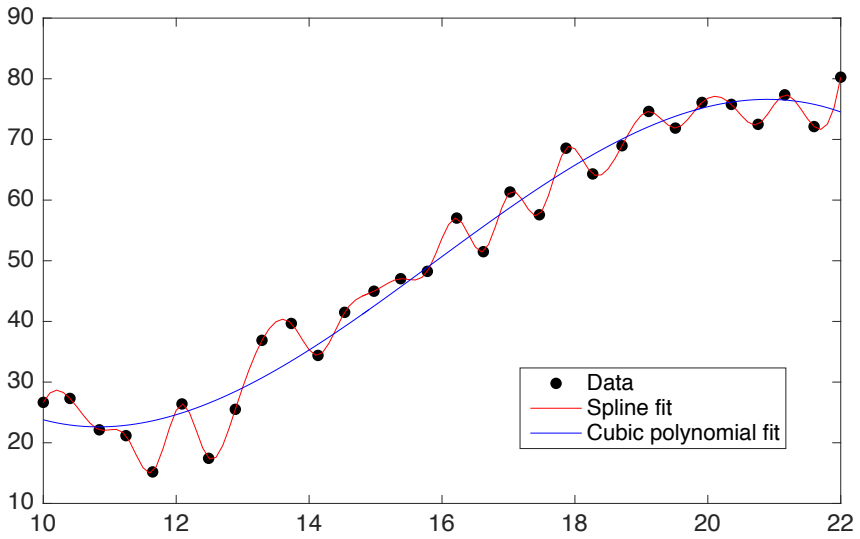
FIGURE 2.4. *A linear model fit by least squares to the* `Income` *data from Figure 2.3. The observations are shown in red, and the yellow plane indicates the least squares fit to the data.*

- `Private` : Public/private indicator
- `Apps` : Number of applications received
- `Accept` : Number of applicants accepted
- `Enroll` : Number of new students enrolled
- `Top10perc` : New students from top 10 % of high school class
- `Top25perc` : New students from top 25 % of high school class
- `F.Undergrad` : Number of full-time undergraduates
- `P.Undergrad` : Number of part-time undergraduates
- `Outstate` : Out-of-state tuition
- `Room.Board` : Room and board costs
- `Books` : Estimated book costs
- `Personal` : Estimated personal spending
- `PhD` : Percent of faculty with Ph.D.'s
- `Terminal` : Percent of faculty with terminal degree
- `S.F.Ratio` : Student/faculty ratio
- `perc.alumni` : Percent of alumni who donate
- `Expend` : Instructional expenditure per student
- `Grad.Rate` : Graduation rate

Non-parametric models

- Rather than assuming a simple functional relationship (e.g. a line), try to fit a curve (nonlinear) to the data

- Examples: Splines, K-nearest neighbors, Trees…



	Parametric models	Non-parametric models
Interpretability	Simple to interpret what each variable is doing, i.e. what effect does each predictor (X1, X2, …) have on the prediction, (Y)	No simple relationship between X and Y
Flexibility	Less flexible — e.g. linear regression can only fit lines	Very flexible — can capture many different functions f(X)

Tradeoff between flexibility and interpretability

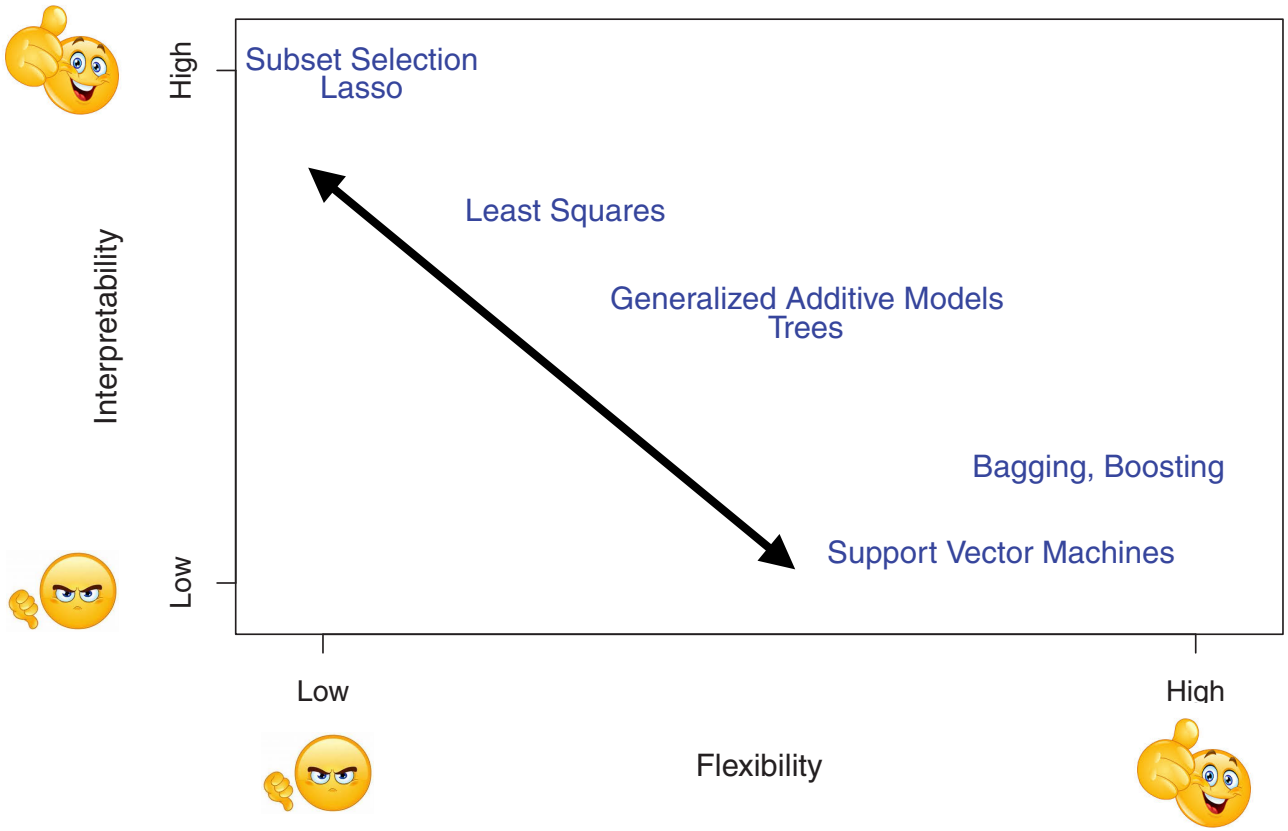


FIGURE 2.7. *A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.*

In each situation, would you care more about model flexibility or interpretability?

1. Building a model to predict a child's weight gain based on components of their diet

2. Building a model to predict whether a person will default on their car loan

Supervised vs. unsupervised learning

- Supervised: Learn from examples.

Supervised vs. unsupervised learning

- Unsupervised: No “labels” or instructions are given.
- Try to learn the relationships between features.
- Example: Clustering

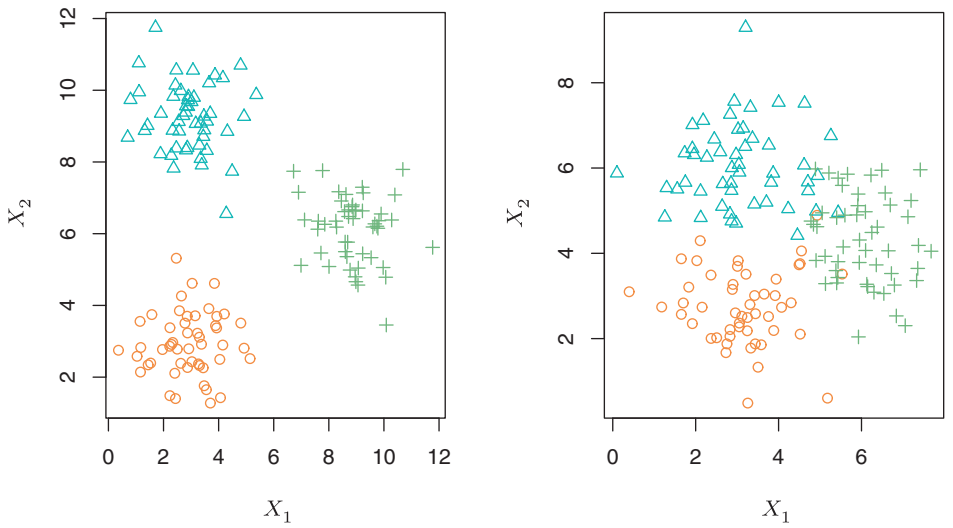
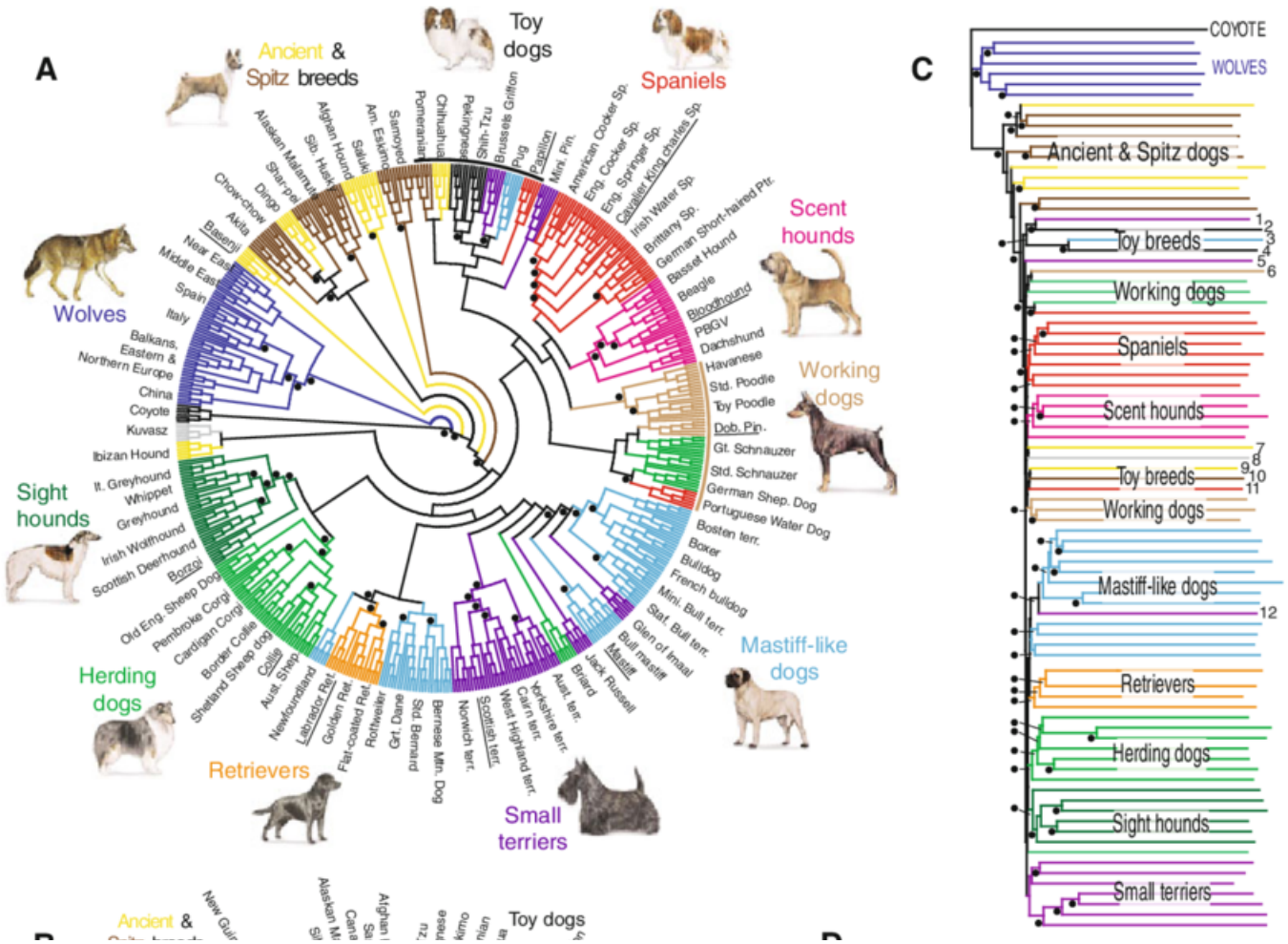


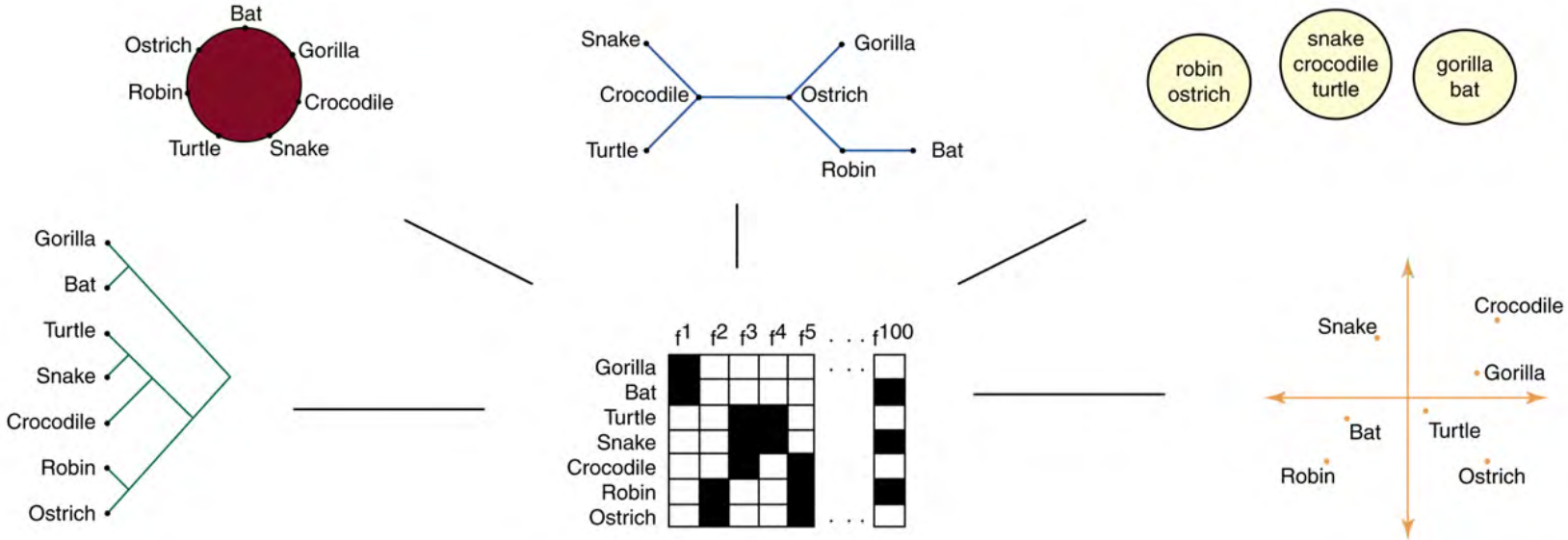
FIGURE 2.8. *A clustering data set involving three groups. Each group is shown using a different colored symbol.* Left: *The three groups are well-separated. In this setting, a clustering approach should successfully identify the three groups.* Right: *There is some overlap among the groups. Now the clustering task is more challenging.*

Unsupervised clustering of dog breeds based on genome similarity



Wayne et al. (2009). Evolutionary genomics of the dog and dog-like carnivores. Journal of Veterinary Behavior-clinical Applications and Research - J VET BEHAV-CLIN APPL RES. 4. 71-71. 10.1016/j.jveb.2008.09.045.

Clustering species



Data matrix

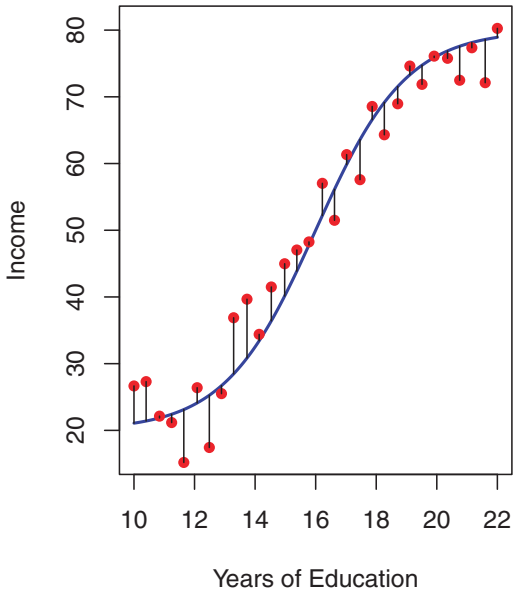
- Data: a matrix of features for each species, e.g. Has wings? Lays eggs? Sheds its skin?
- Multiple forms of clustering discover different types of relationships

1. Griffiths, T. L., Chater, N., Kemp, C., Perfors, A. & Tenenbaum, J. B. Probabilistic models of cognition: exploring representations and inductive biases. Trends Cogn. Sci. 14, 357–364 (2010).

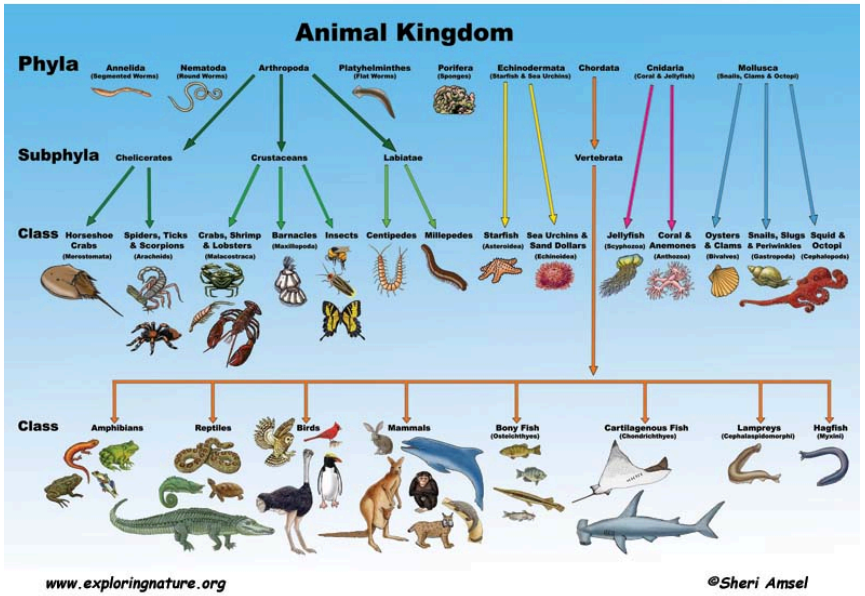
Regression vs. Classification

Quantitative prediction →

$Y \in \Re$



Features (X) could be quantitative, categorical, or a combination of both



Categorical prediction

Each item (organism) is assigned to one of a discrete set of categories (e.g. phylum or species)

How do we fit models?

- Need a way to measure (quantify) model quality
- Once we have a measure of quality, we can choose the “best” model (highest quality)
- A very common measure of quality is **mean squared error (MSE):**

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2,$$

- Properties:
  Non-negative
  MSE = 0 if y=f(x) for all examples
  Every example makes an equal contribution

Two types of error

- Training mean squared error (MSE): How accurately does our model fit the *known (training) examples*?

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2,$$

- Testing MSE: How accurately does the model predict *new (testing) data?*

$$\text{Ave}(\hat{f}(x_0) - y_0)^2,$$

- We usually care about testing MSE, not training MSE

- However, measuring training MSE may be difficult/expensive

Review: Dichotomies so far

- Prediction and Inference
- Parametric and Non-parametric models
- Supervised and Unsupervised learning
- Regression and Classification
- Training and Testing Error
- Flexible/Complex and Inflexible/Simple models
- **Next:** Bias and Variance

Can we assume training MSE ≈ testing MSE?

- No!
- If we rely only on training MSE, we may **overfit** the data
- Overfitting: When the model captures noise in the data, rather than the true relationship

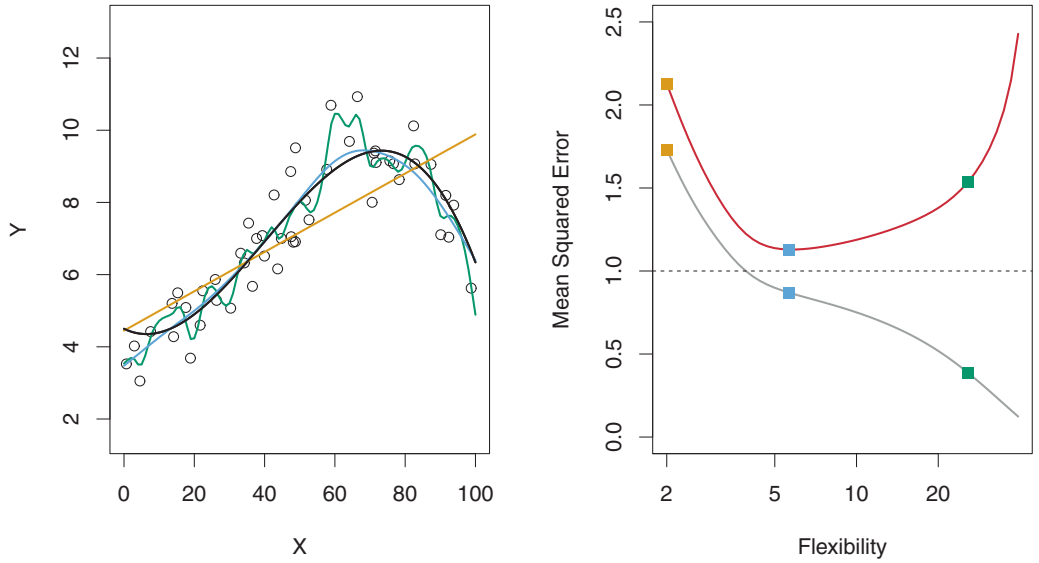
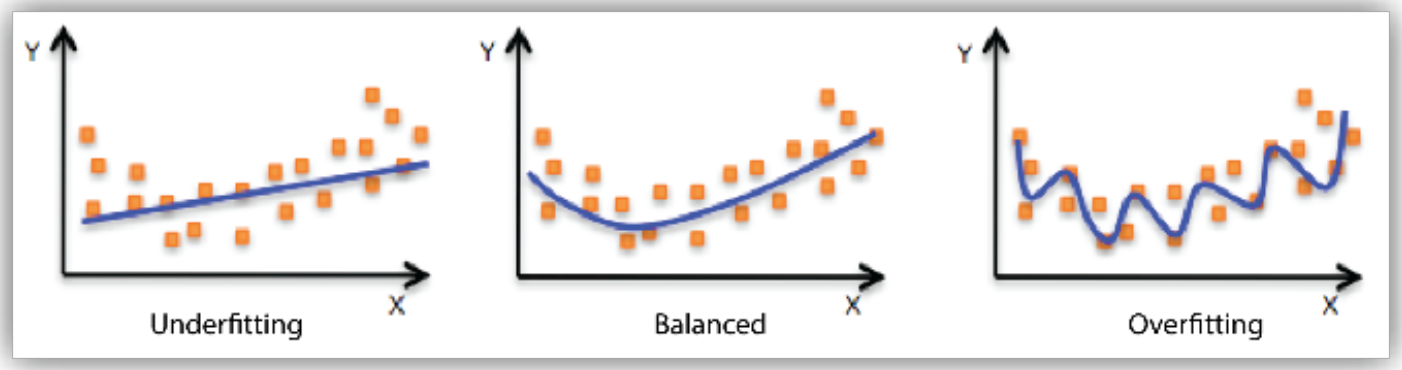
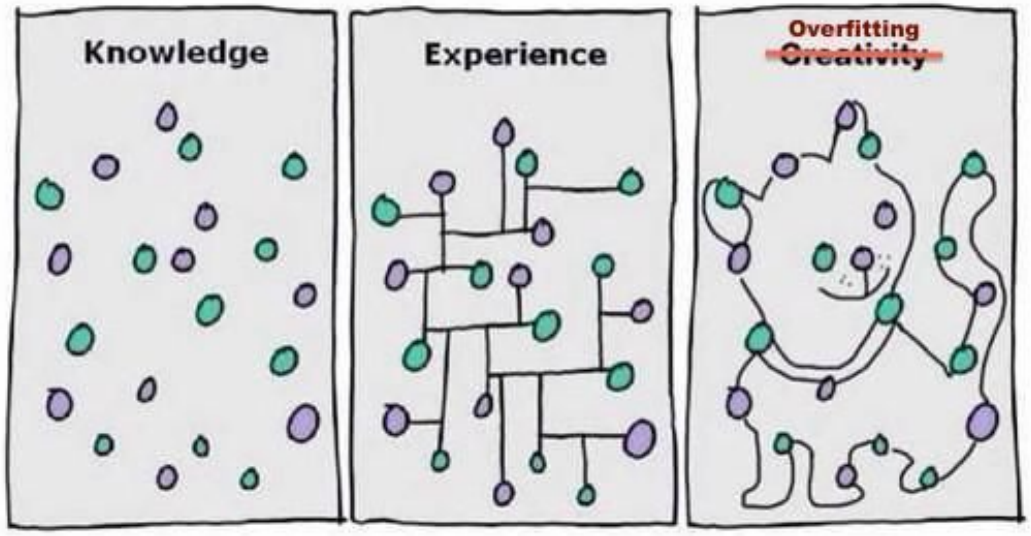


FIGURE 2.9. Left: *Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves).* Right: *Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.*

$$Y = f(X) + \epsilon.$$

Overfitting





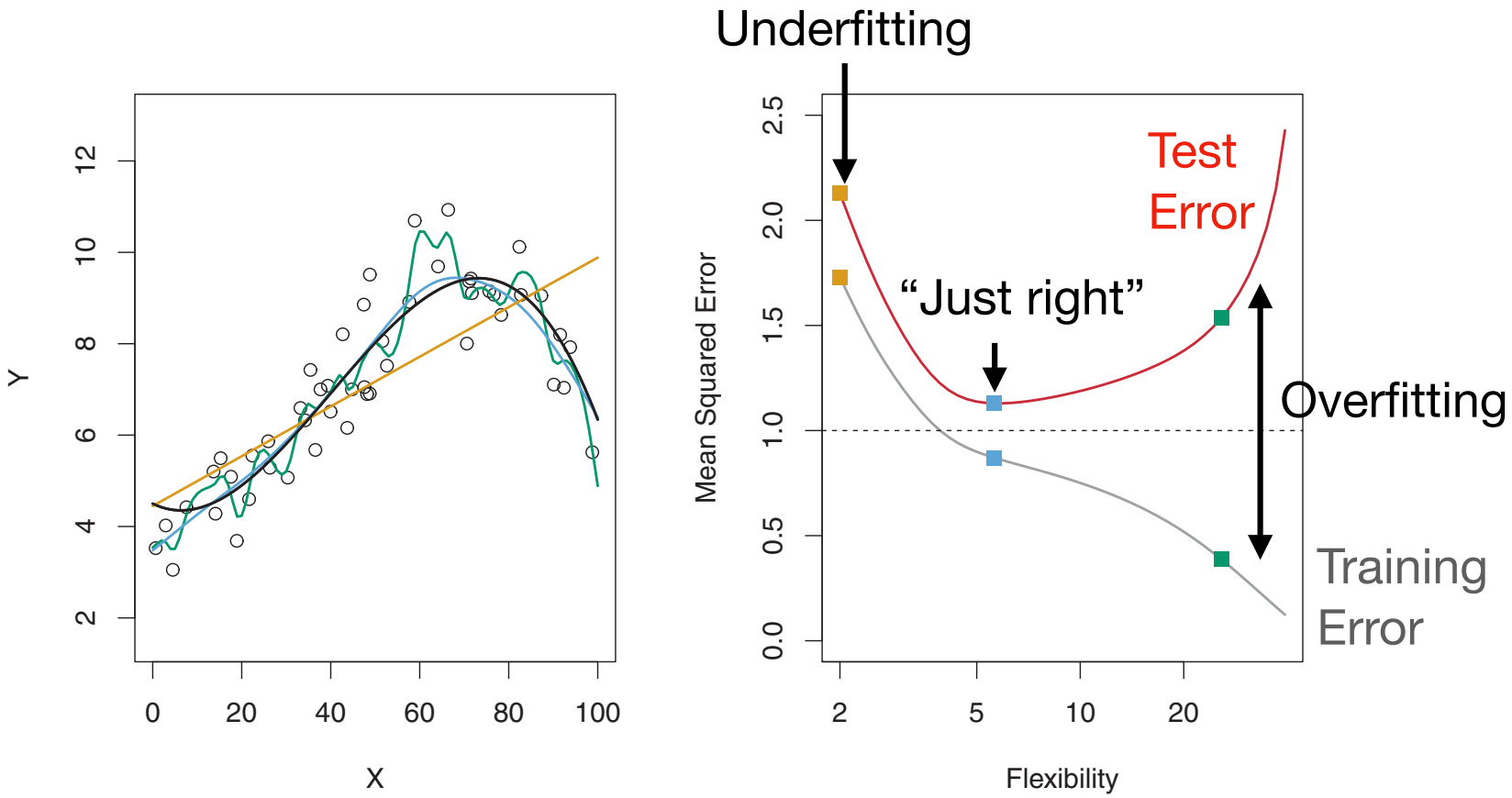
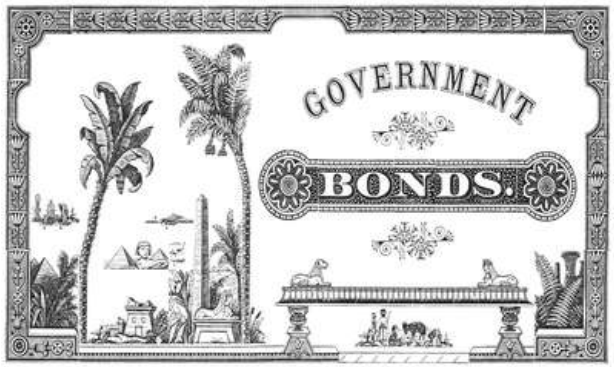


FIGURE 2.9. Left: *Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves).* Right: *Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.*

Match each picture with the corresponding strategy

- Overfitting
- Underfitting



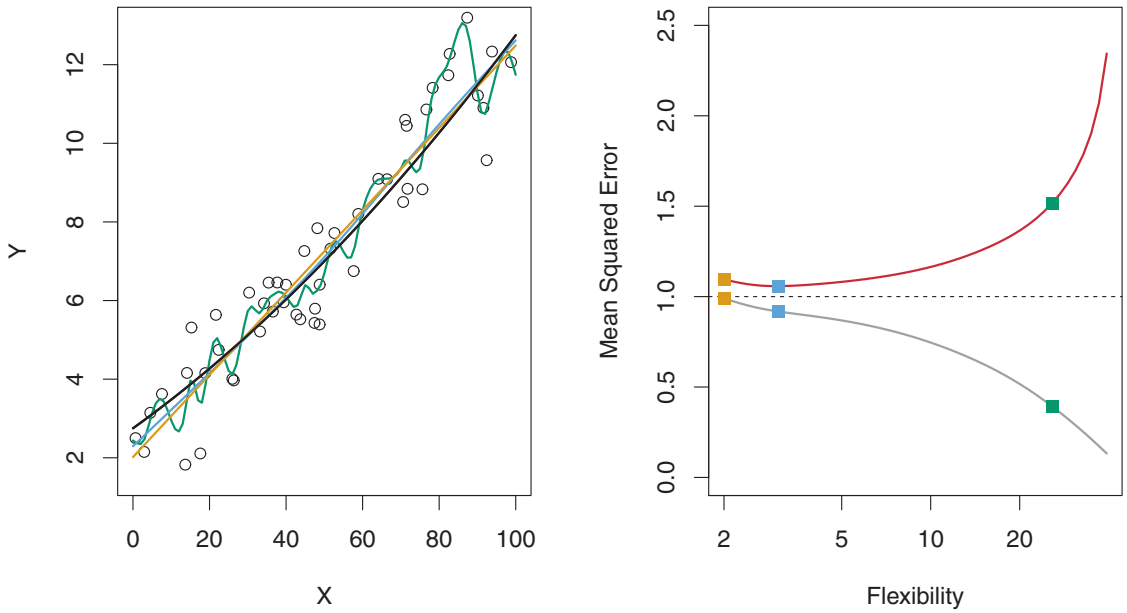


FIGURE 2.10. *Details are as in Figure 2.9, using a different true f that is much closer to linear. In this setting, linear regression provides a very good fit to the data.*

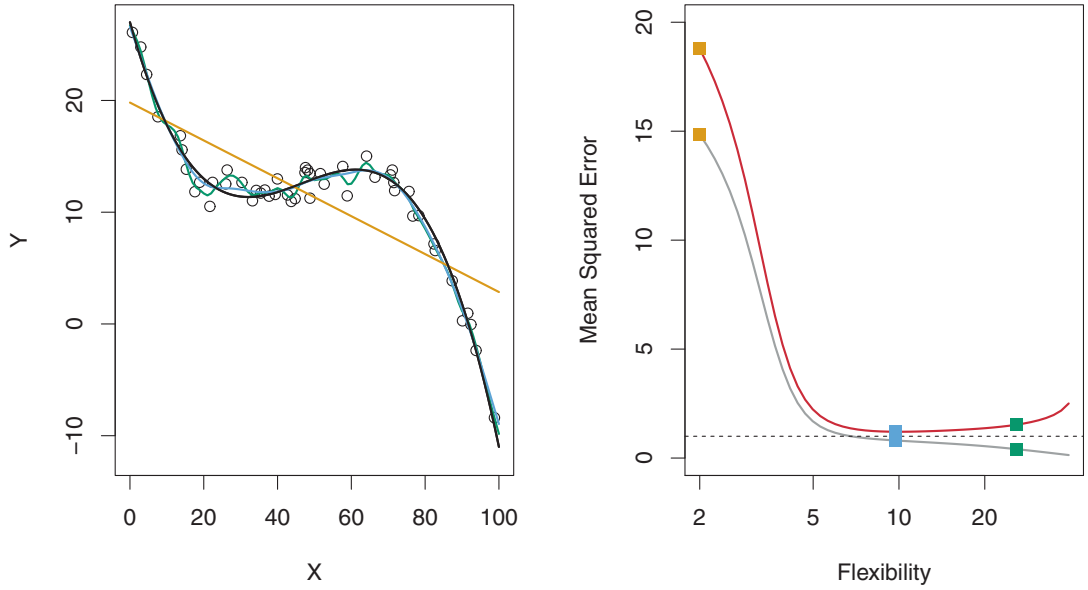


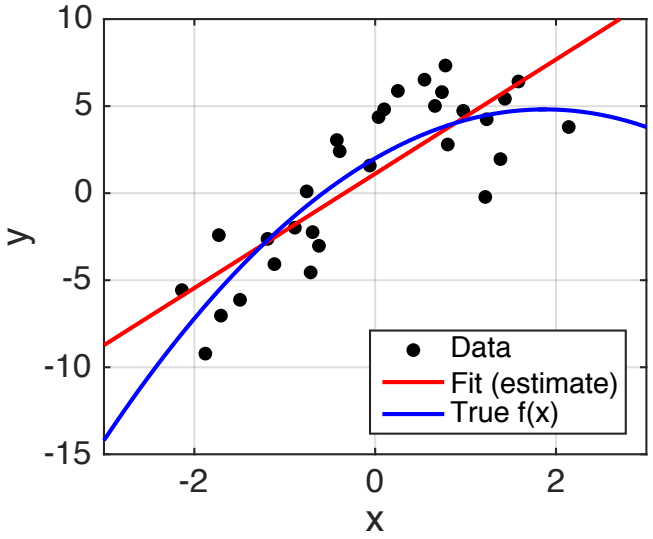
FIGURE 2.11. *Details are as in Figure 2.9, using a different f that is far from linear. In this setting, linear regression provides a very poor fit to the data.*

Bias-Variance Tradeoff

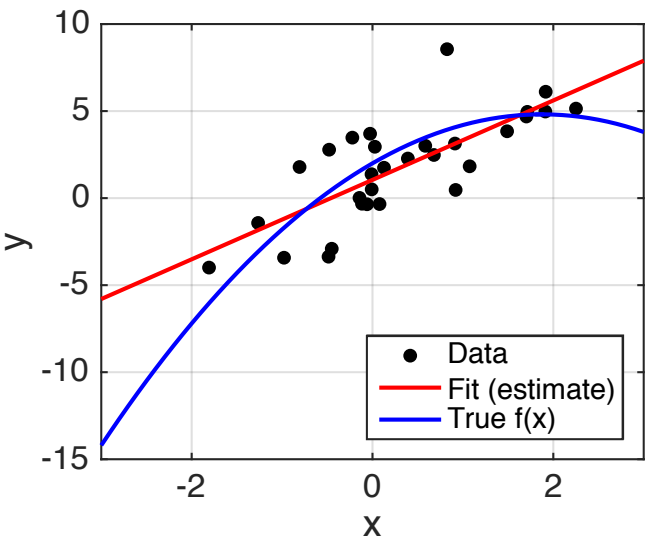
$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon).$$

- Total error is a combination of 3 parts:
  - Variance: How much does the model vary due to random training samples?
  - (Squared) Bias: How far off would the model be, even if we had plenty of data (infinite samples)?
  - Noise ("irreducible" error): No model can capture this part.

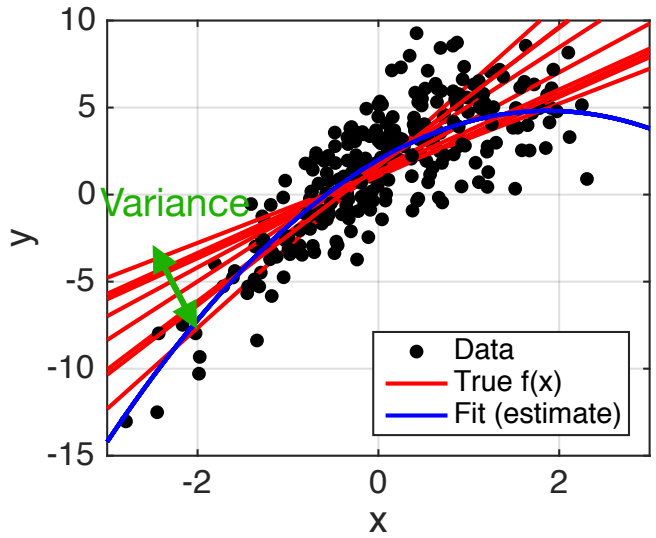
Sample A
N=30 data points



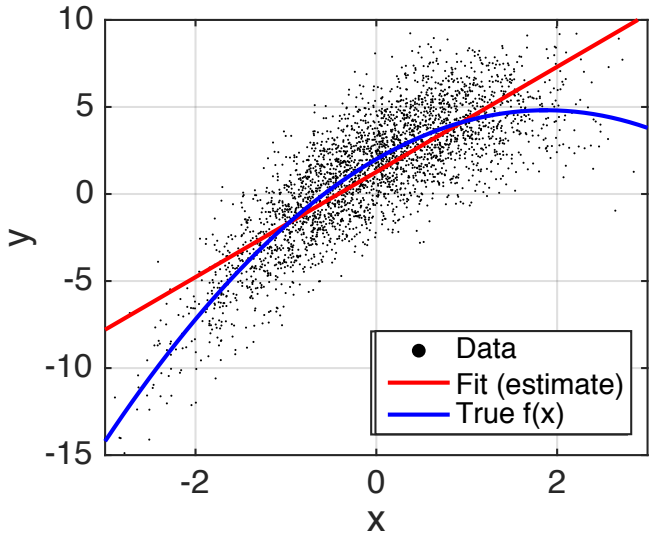
Sample B
N=30 data points



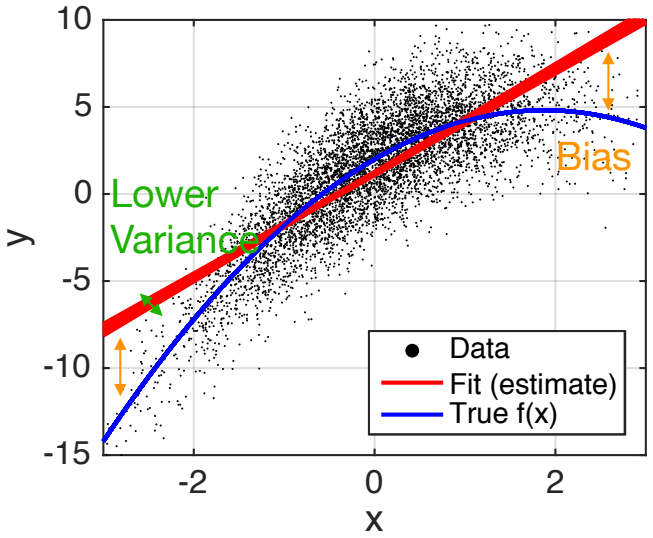
Many samples,
N=30 data points each

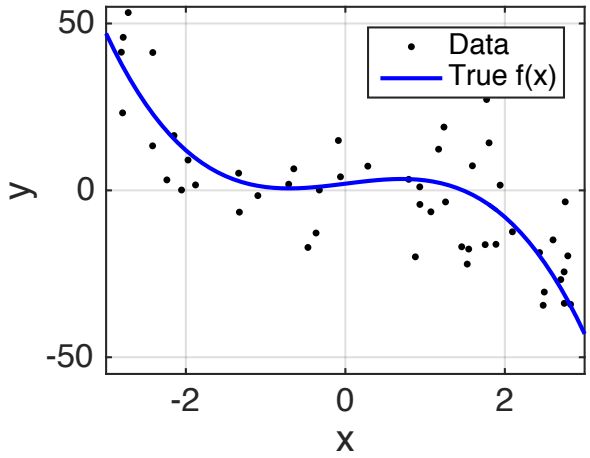


Sample C
N=3000 data points



Many sample,
N=3000 data points each

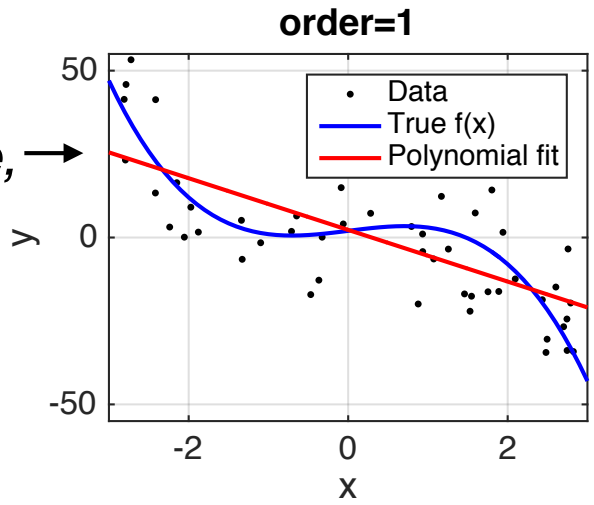


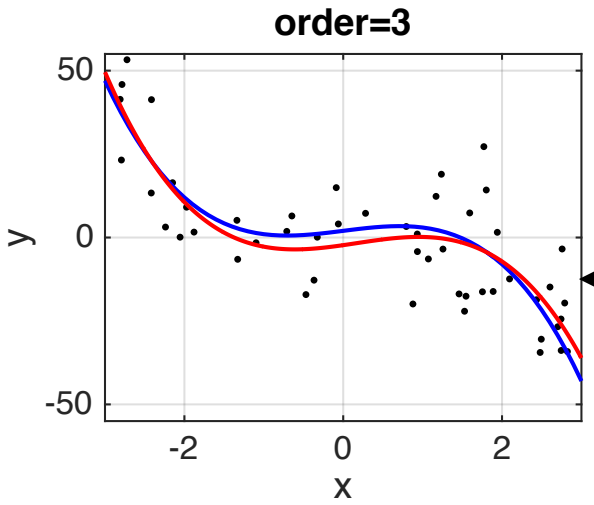

50
0
-50
y
-2
0
2
x
Data
True f(x)

$y = \beta_0 + \beta_1 x$

$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$

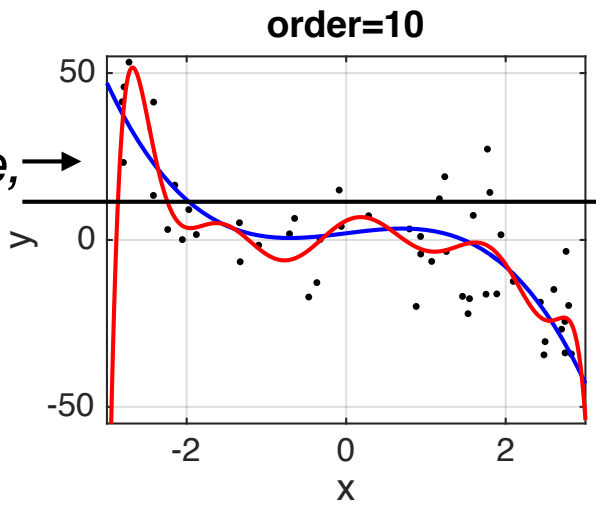
Too simple/inflexible, Model has high <u>bias</u>

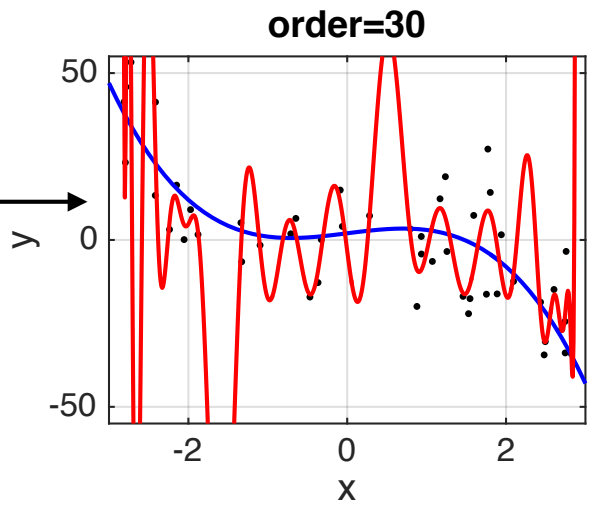




Just right, model has balance of bias and variance

Too complex/flexible, Model has high <u>variance</u>

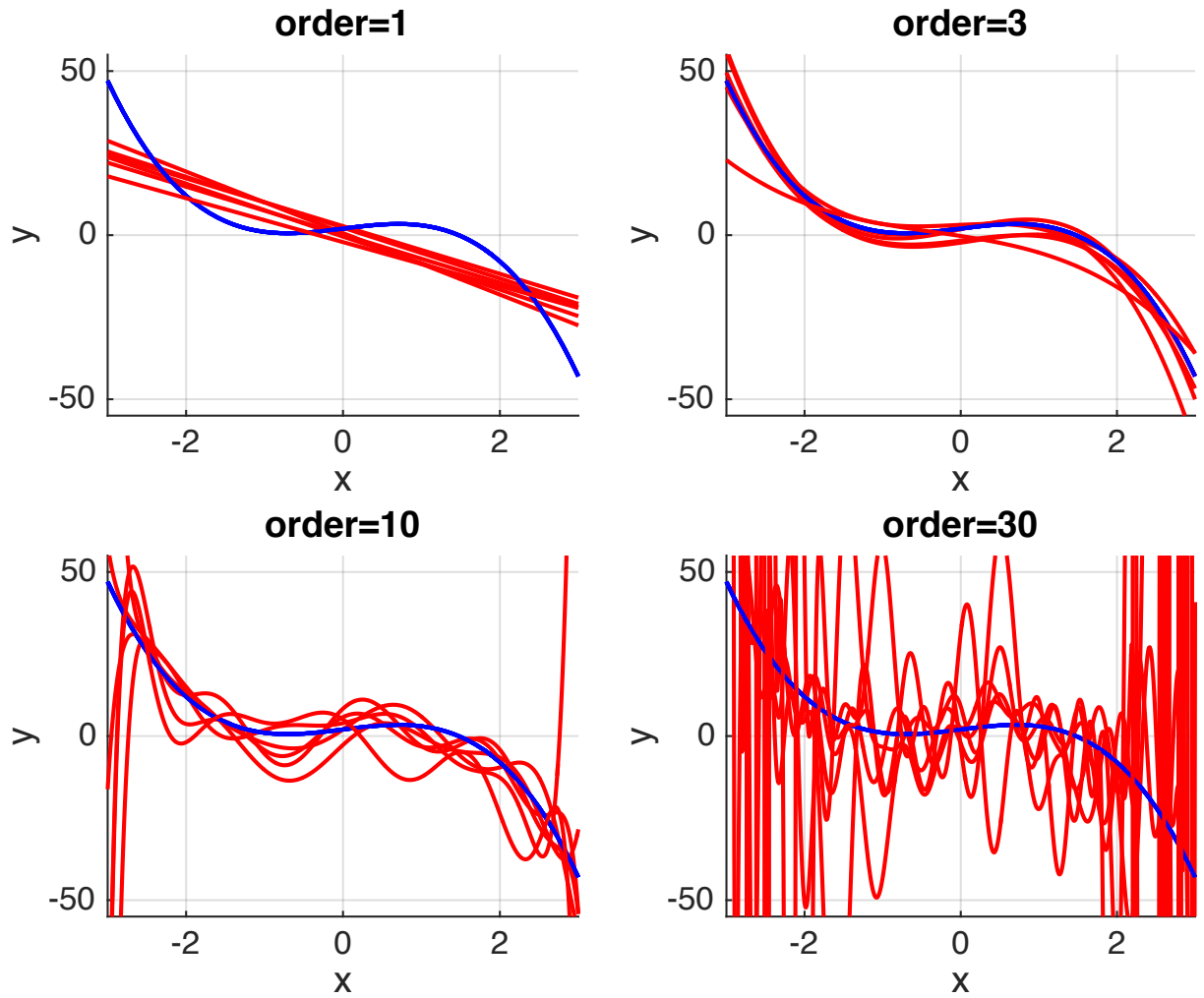




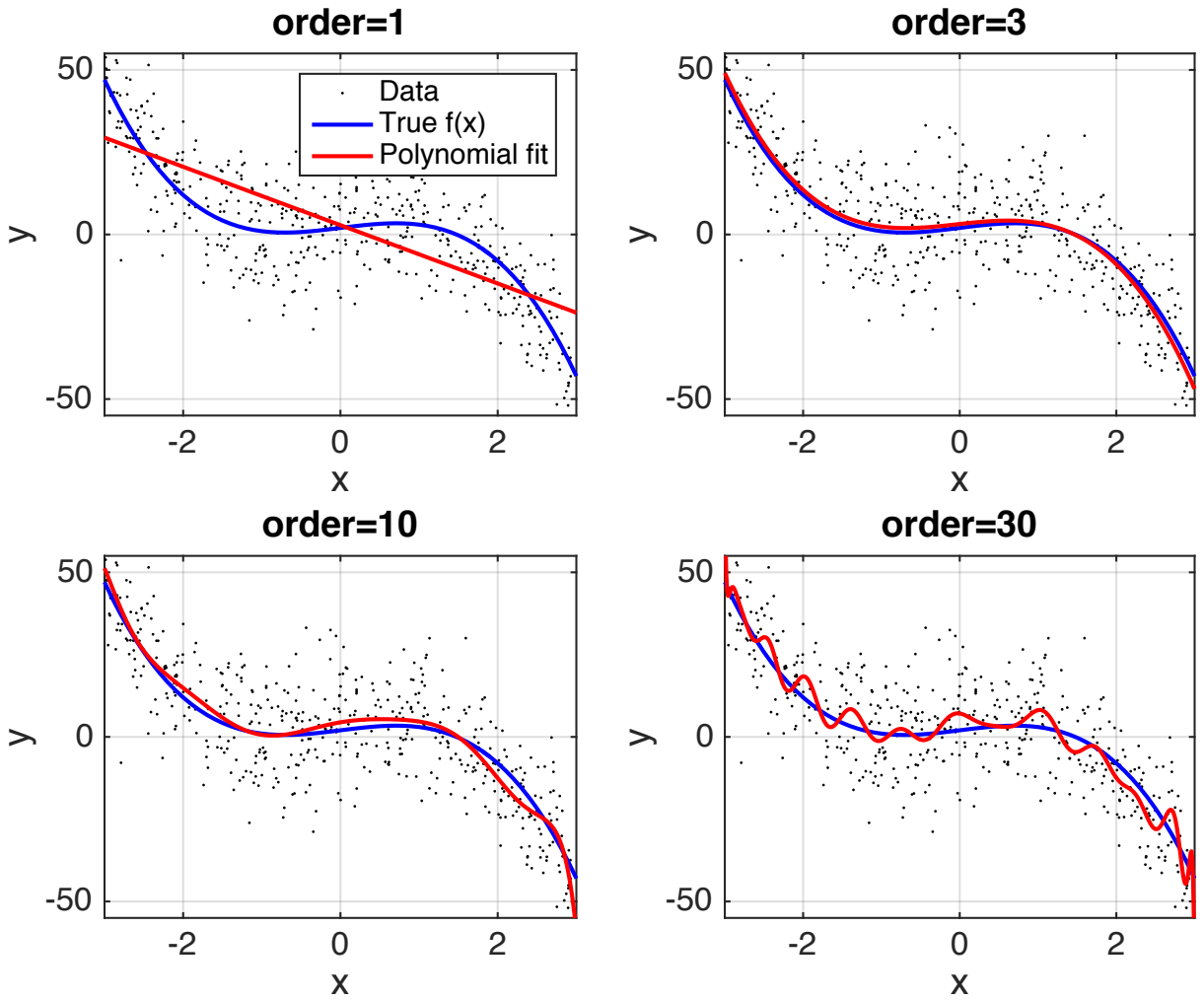
$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots + \beta_{10} x^{10}$

Polynomial fits with increasing order show the problem of overfitting

Variance: Wild differences with each new data sample



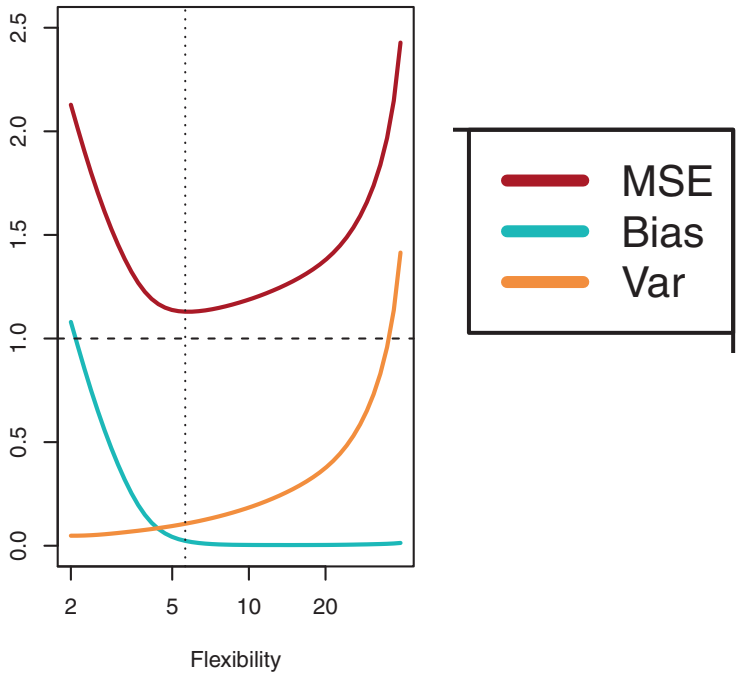
Bias: Error remains even when n is very large



Bias and Variance as a function of model flexibility

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon).$$

- Variance: Lower for simple, inflexible models.
- (Squared) Bias: Lower for complex, flexible models.
- Total MSE is a U-shaped function of flexibility
- Optimal flexibility is a balance, with low variance and low bias



Classification

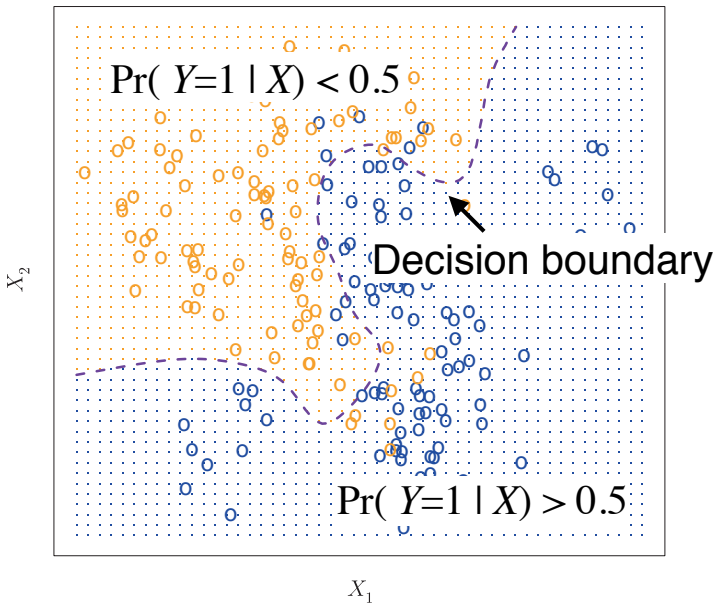
- For classification, we can't use MSE (mean squared error) because the output is a *category*, not a number
- Instead, use the **error rate:** $\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i).$
    - The "indicator function"
      $I(A)$ = 1 if A is true,
      = 0 if A is false

Optimal (Bayes) classifier

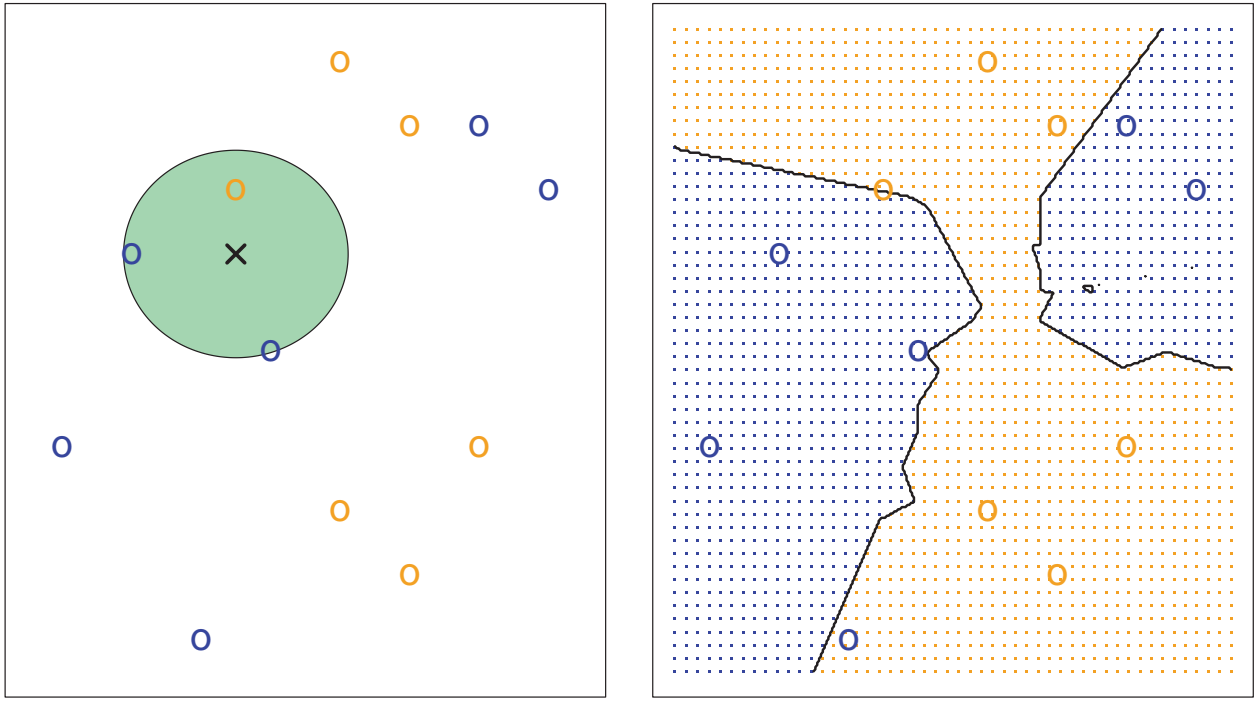
- Bayes studied probability and formulated *Bayes' theorem*
- In modern data analysis, Bayesian algorithms use calculate outcome probabilities to make optimal choices
- For a classifier, this means we should choose the class with the highest probability for each data point

$$\Pr(Y = j|X = x_0)$$

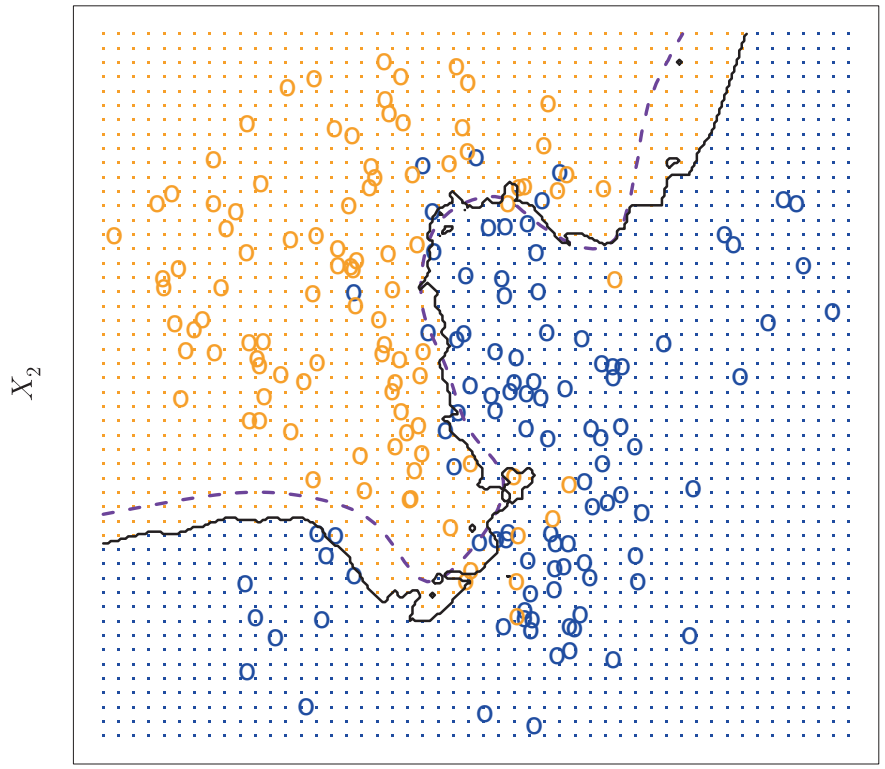
Reverend Thomas Bayes
1701-1761



Example of a classifier: K-nearest neighbors

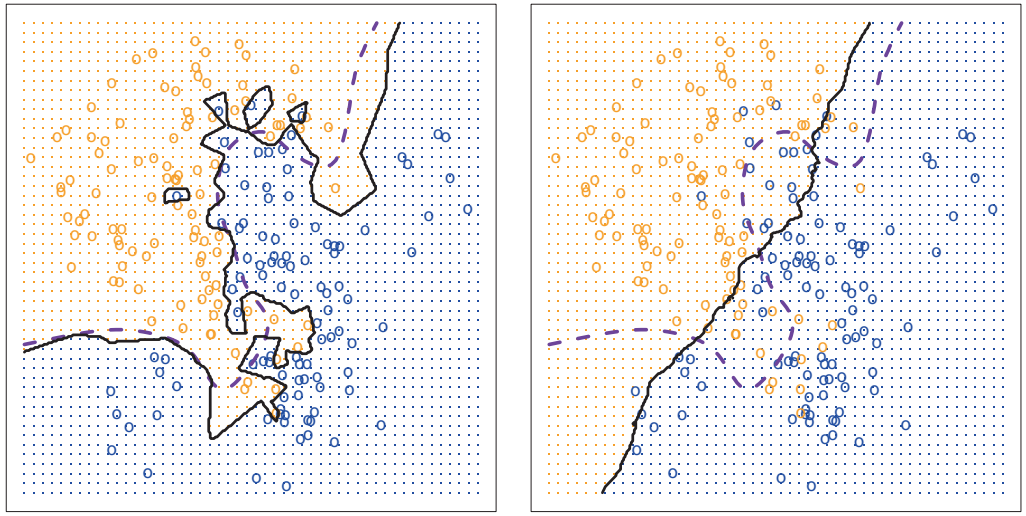



Zillow
Buy
Rent
Sell
Mortgages
Agent finder
Home design
More
san diego ca
Listing Type
Any Price
0+ Beds
Home Type
More
Don't miss out!
New homes are getting added all the time. Save your search and be the first to know.
Get Started
More Map
$1.13M
$3.93M
$3.56M
$3.29M
$3.11M
$2.84M
$1.1M
$1.24M
Del Mar Mesa Rd
$934K
$2.92M
$3.45M
$860K
$923K
$3.50M
$881K
$4.18M
$1.7M
$1.2M
$12K
Del Mar Mesa Rd
$3.72M
$2.58M


KNN: K=10
X_2

Flexibility and classifiers

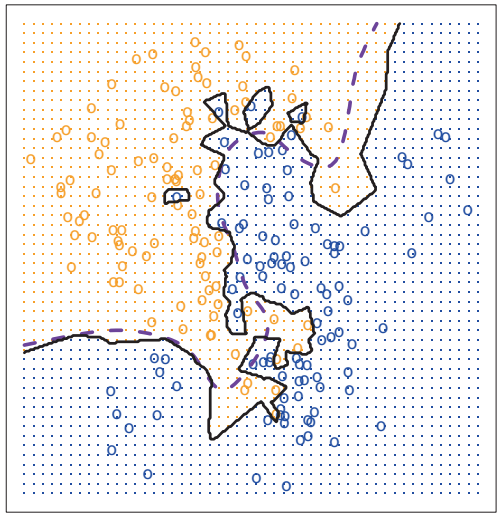
- For each of the two classifiers in the figure:
- Which one is K=1, and which one K=100?
- Which is more flexible/more complex?
- Which has higher bias?
- Whic has higher variance?



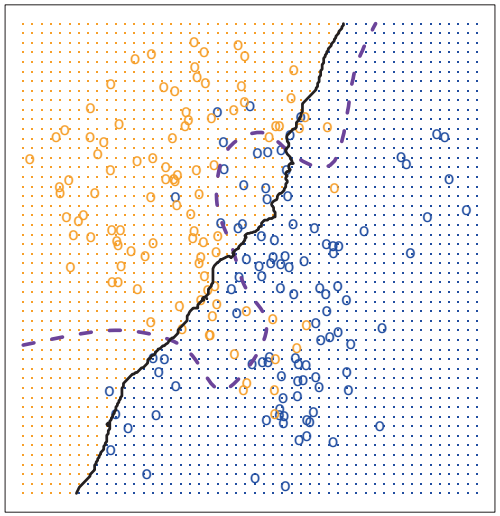
Flexibility and classifiers

- Which of the two classifiers below (K=1 or K=100) is:
  - More flexible/more complex?
    - **K=1 is more flexible; it can fit very complex decision boundaries**
  - Higher bias?
    - **K=100 has higher bias. It imposes a smooth decision boundary**
  - Higher variance?
    - **K=1 has higher variance. The decision boundary is very sensitive to the exact location of each data point**

KNN: K=1



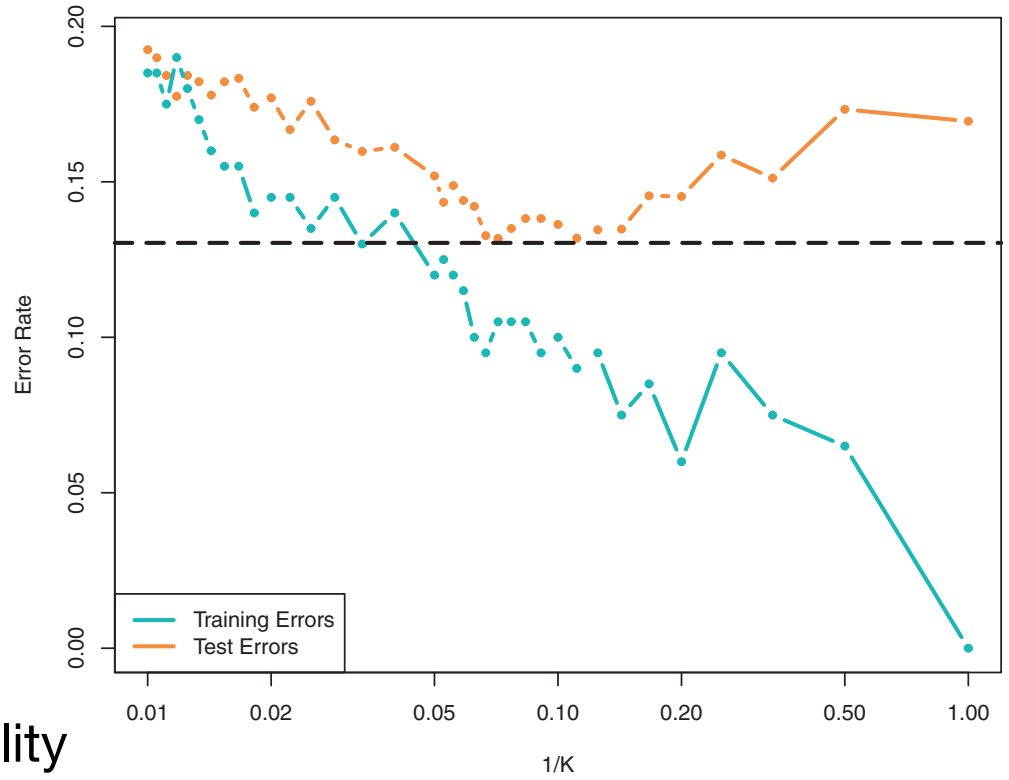
KNN: K=100



How can we control the flexibility of a KNN classifier?

- A flexible classifier can have a complex (nonlinear, highly curved) decision boundary.
- A less flexible classifier would have a smoother decision boundary.
- Small values of K: High flexibility/complex decision boundary

Choosing the right value of K to balance bias and variance

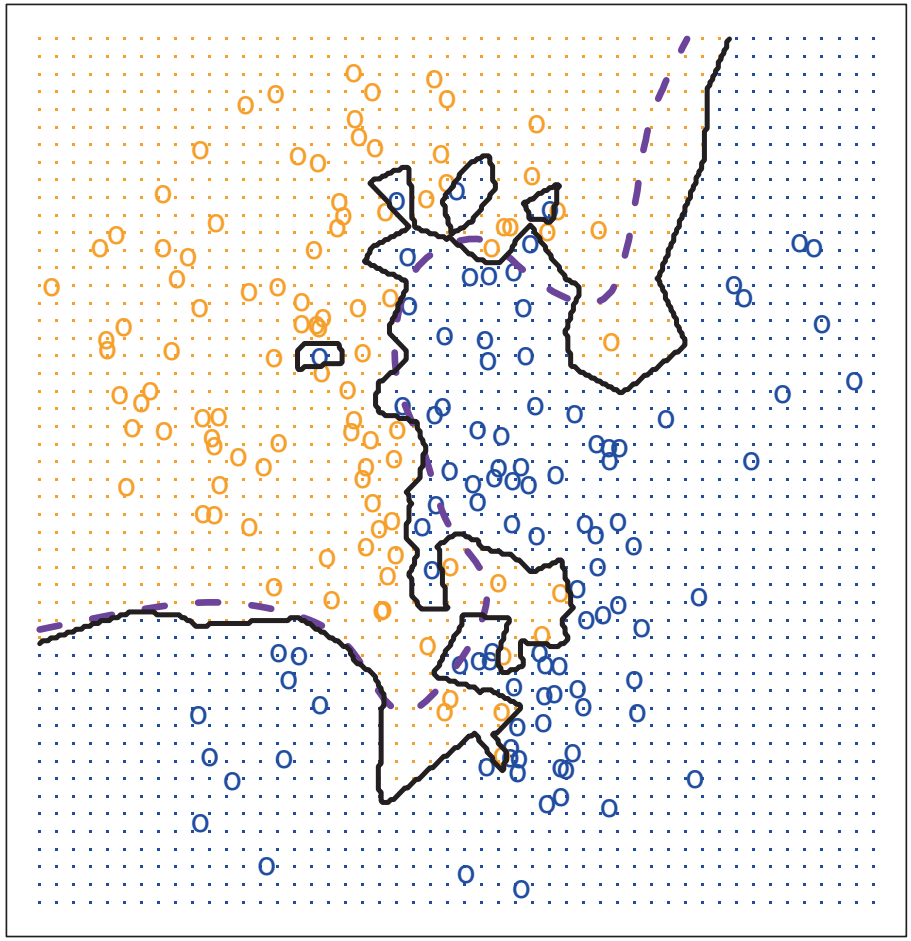


Large K,
Low flexibility

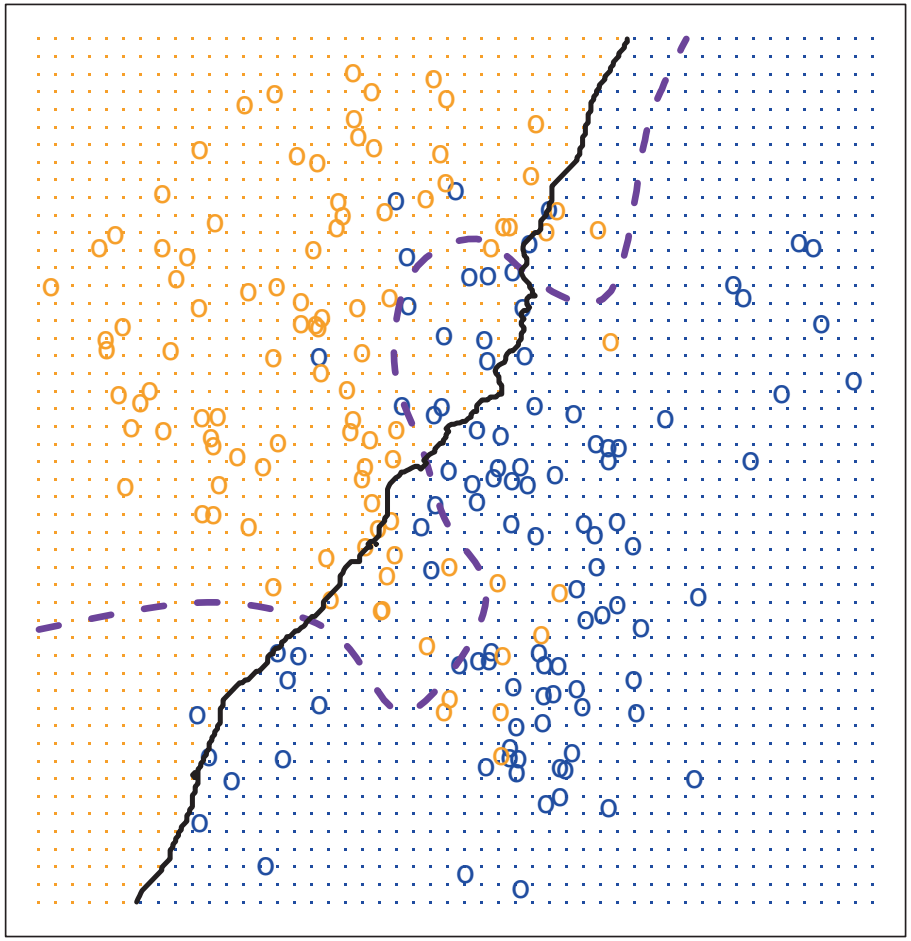
Small K,
Very flexible model

FIGURE 2.17. *The KNN training error rate (blue, 200 observations) and test error rate (orange, 5,000 observations) on the data from Figure 2.13, as the level of flexibility (assessed using $1/K$) increases, or equivalently as the number of neighbors K decreases. The black dashed line indicates the Bayes error rate. The jumpiness of the curves is due to the small size of the training data set.*

KNN: K=1

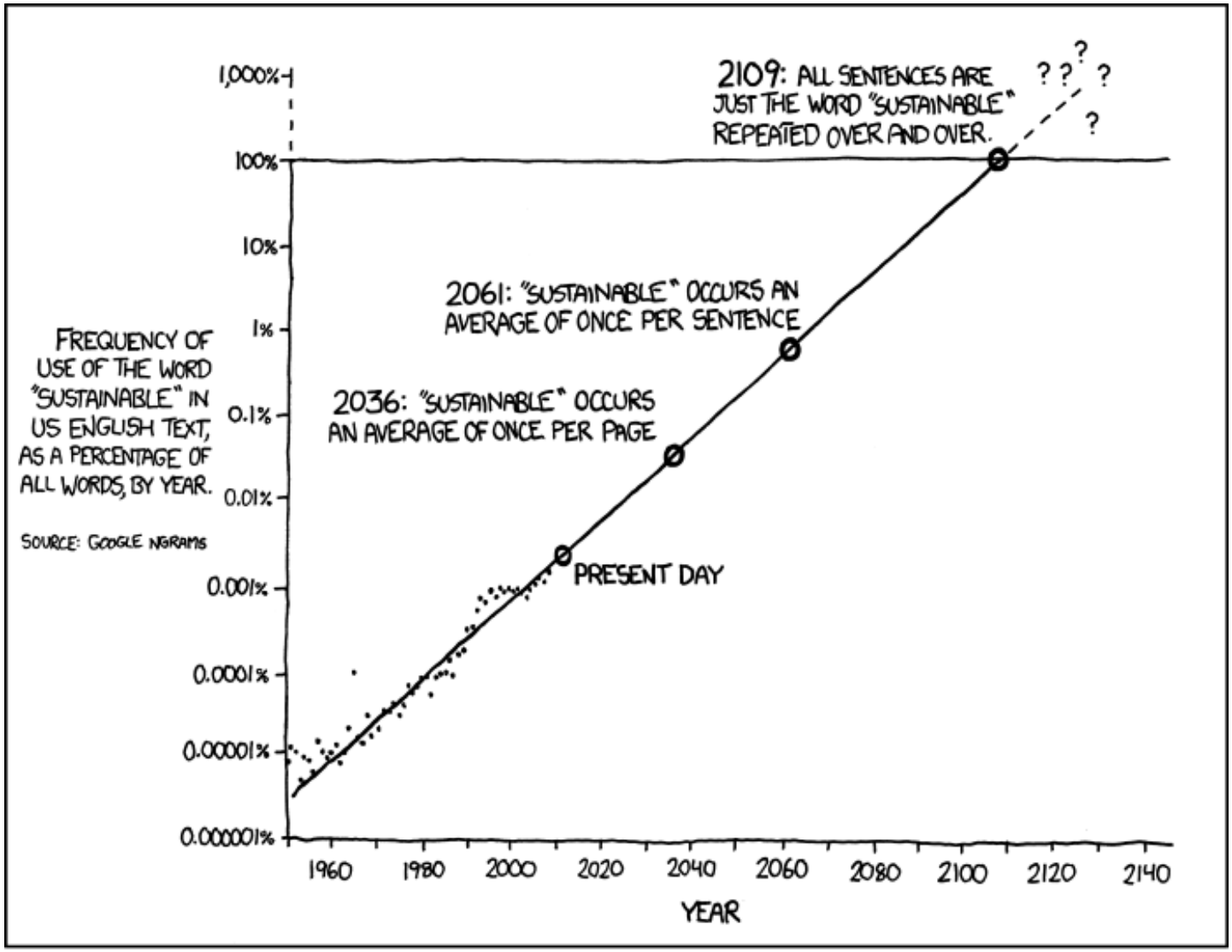


KNN: K=100



Linear Regression

- Linear regression answers:
  - Is there a relationship between Y and X1, X2, …? [Inference]
  - What is the relationship between Y and X1, X2…? Positive or negative correlation? Weak or strong? [Inference]
  - Given a new value of X, what do I expect Y will be? [Prediction]
  - How accurate will our prediction be? [Prediction]
  - Are there interactions between some of the prediction variables? [Inference]



THE WORD "SUSTAINABLE" IS UNSUSTAINABLE.

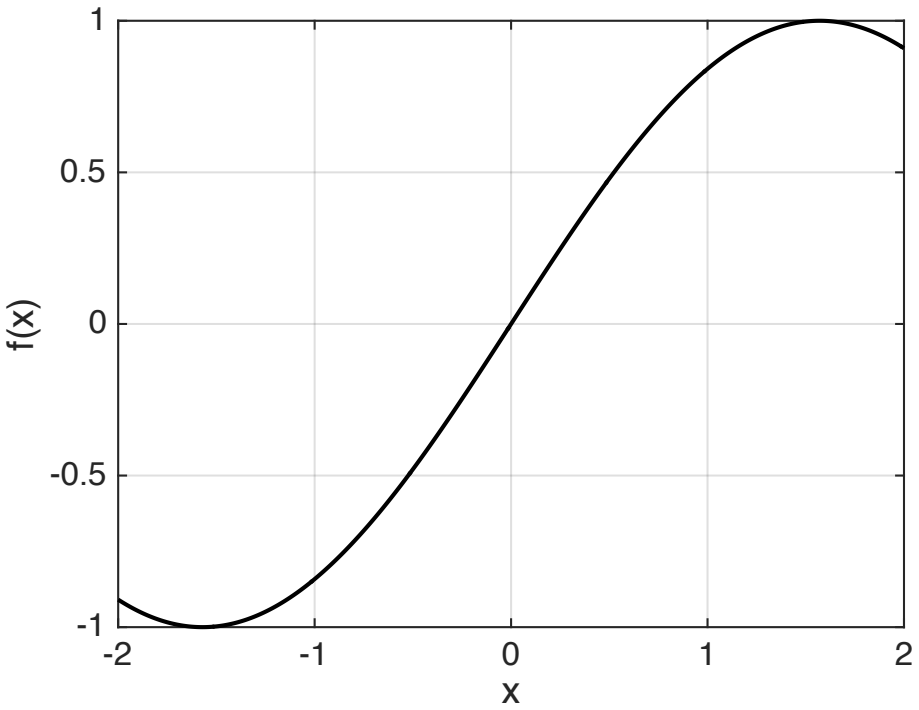
Why *linear*?

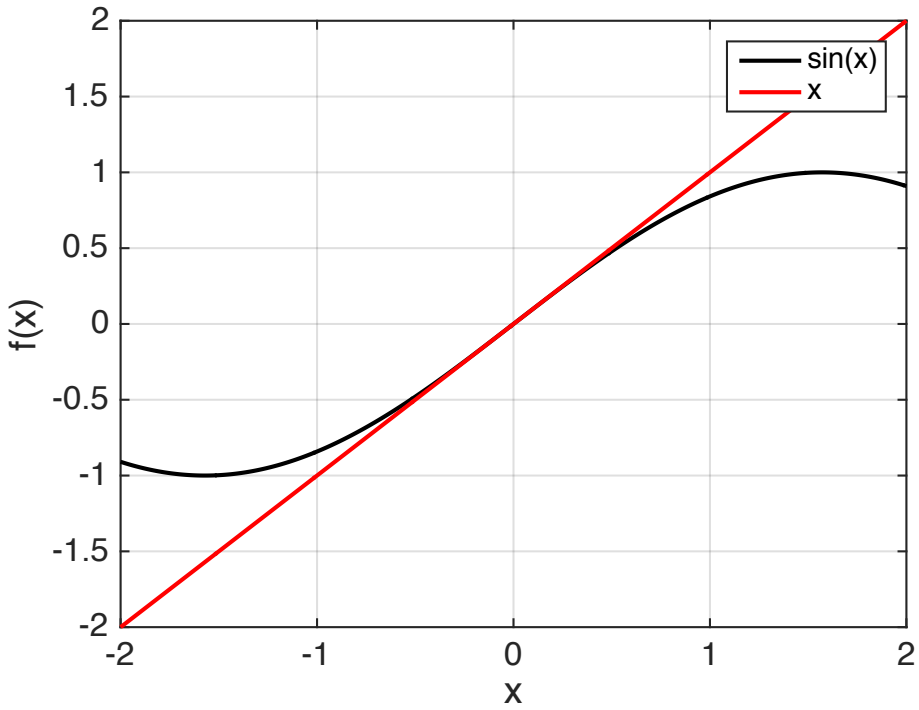
- Real world correlations are (usually) not exactly linear
- Real world correlations are (often) approximately linear
- Recall **Taylor Series**: Any function f(x) can be approximated by a polynomial:

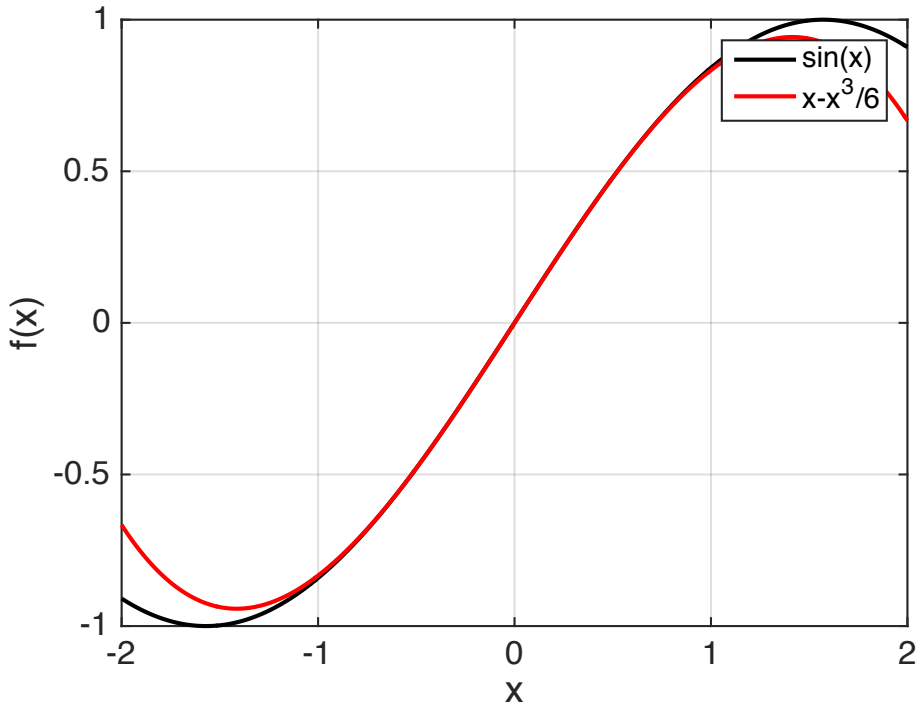
$$f(x) \approx f(0) + f'(0)\,x + \tfrac{1}{2}f''(0)\,x^2 + \tfrac{1}{6}\,f'''(0)\,x^3 + \ldots$$

$$f(x) \approx a + bx + cx^2 + dx^3 + \ldots$$

$$f(x) \approx \beta_0 + \beta_1 x + \ldots$$


1
0.5
0
-0.5
-1
f(x)
-2
-1
0
1
2
x


2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
f(x)
-2
-1
0
1
2
x
sin(x)
x


1
0.5
0
-0.5
-1
f(x)
-2
-1
0
1
2
x
sin(x)
x-x3/6

Why *linear*?

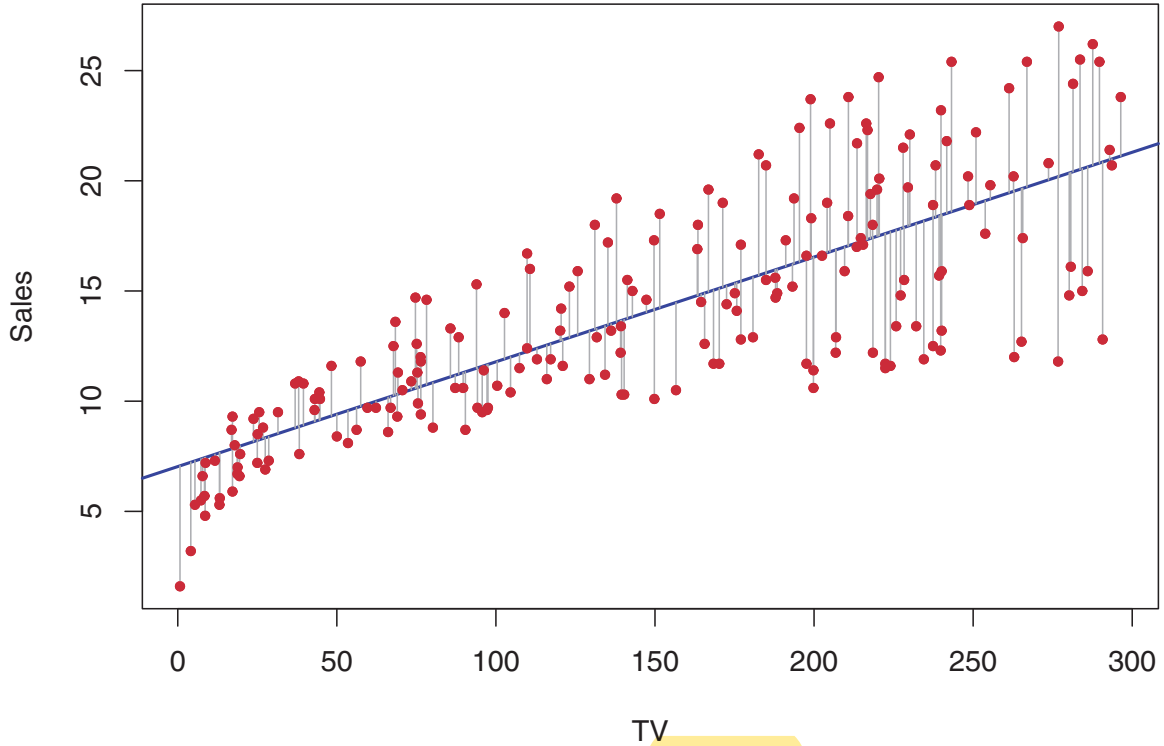
- In practice, we often start our exploration of a data set using simple linear models because they are easy to fit and easy to interpret/understand.
- If a linear model is a good fit to the data and makes accurate predictions, then there's no need for a more complex/flexible models
- If not, we can try more complex models, e.g. non-linear regression (coming up in week 7)
- Non-linear models use the linear regression as a basic building block — so you should master the linear regression first

Simple Linear Regression (1 predictor)

$$Y \approx \beta_0 + \beta_1 X.$$

$$\texttt{sales} \approx \beta_0 + \beta_1 \times \texttt{TV}.$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

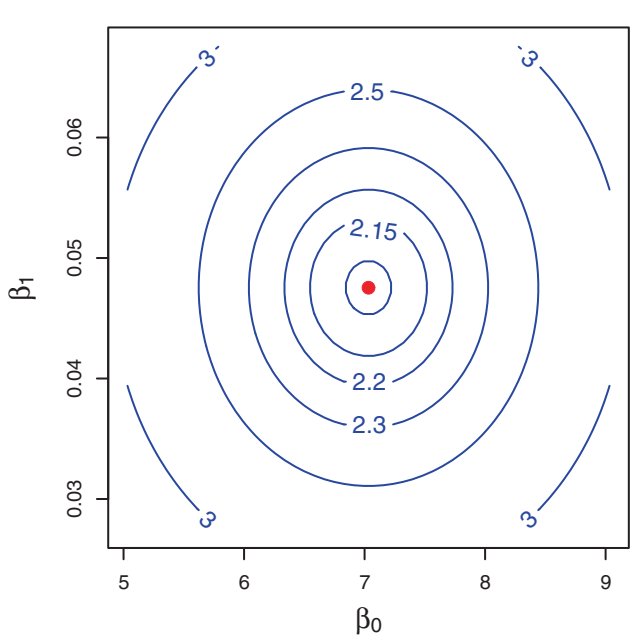


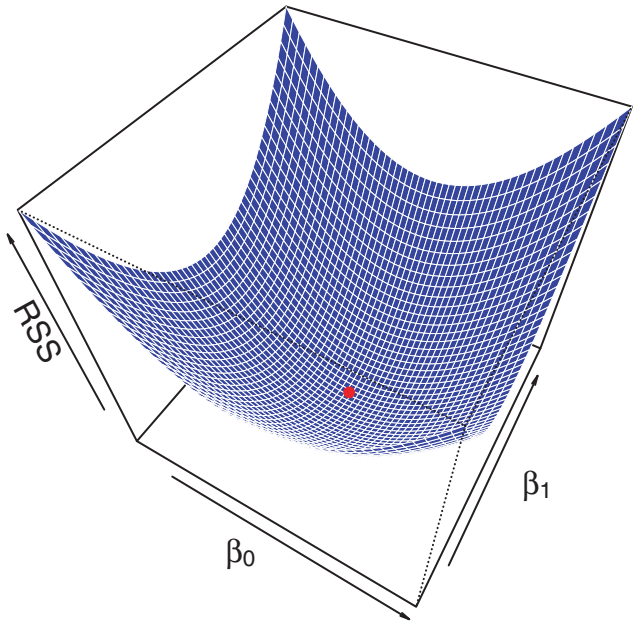
To find good estimates of beta0 and beta1, we need to minimize the *residual sum of squares (RSS)*. This is called ***Least Squares Regression***

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2, \quad = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

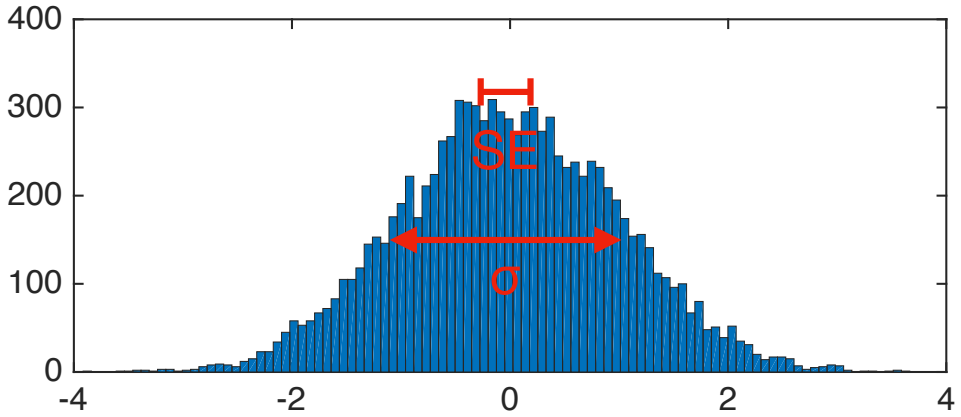
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$





How good is our estimate?

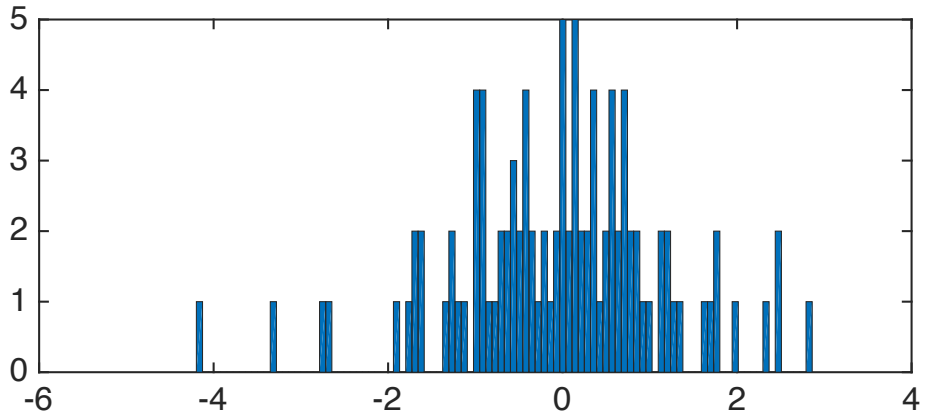
- Recall the difference between standard deviation (SD) and standard error (SE)
    - SD (σ) is the width of a distribution of observations
    - SE is the uncertainty of our estimate of a parameter based on a a set of *n* observations



$$\mathrm{Var}(\hat{\mu}) = \mathrm{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n},$$

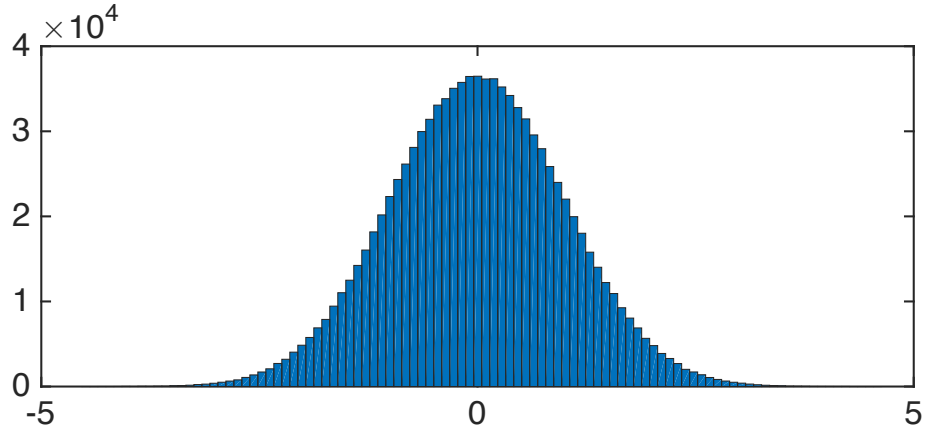
Example: Same SD, different SE

$\sigma = 1$



n=100

SE = 0.1



n = 1,000,000

SE = 10^{-3}

Standard error for regression parameters

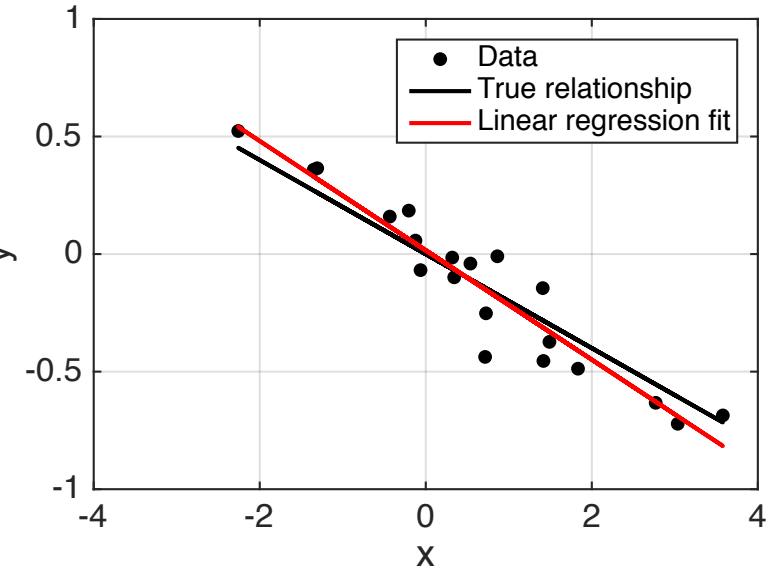
- Depends on: σ, n, and also the distribution of x values

- When x values are more spread out, we have greater leverage for estimating parameters, so the SE is lower.

$$\sigma^2 = \mathrm{Var}(\epsilon)$$

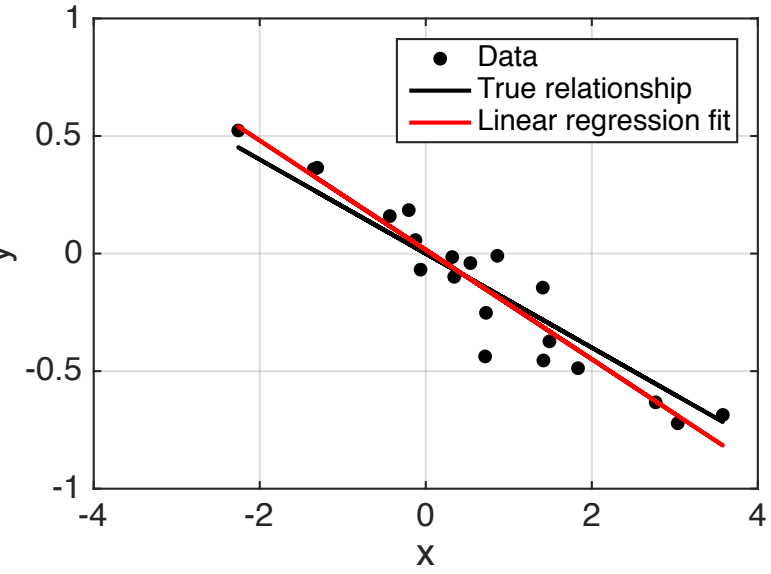
$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right], \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

n = 20, σ = 0.1



	Estimate	SE	tStat	pValue
(Intercept)	0.011361	0.025885	0.4389	0.66596
x1	0.16244	0.026655	6.0942	9.2986e-06

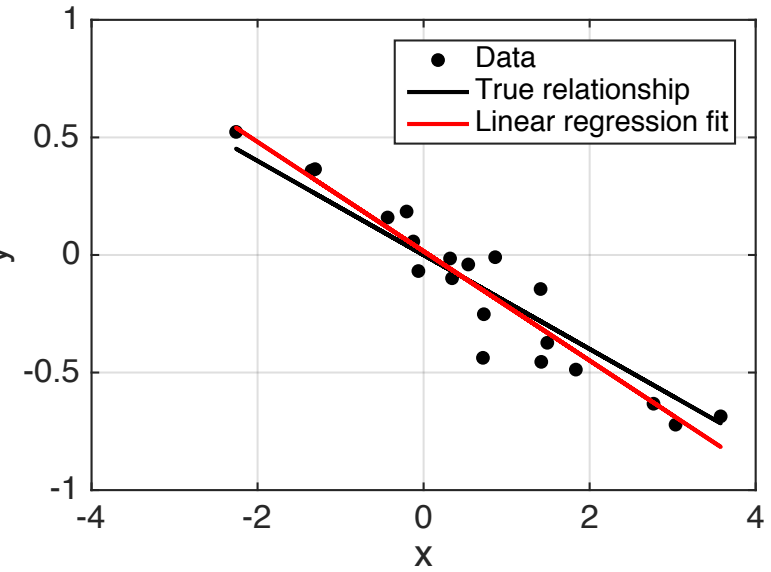
n = 200, σ = 0.1



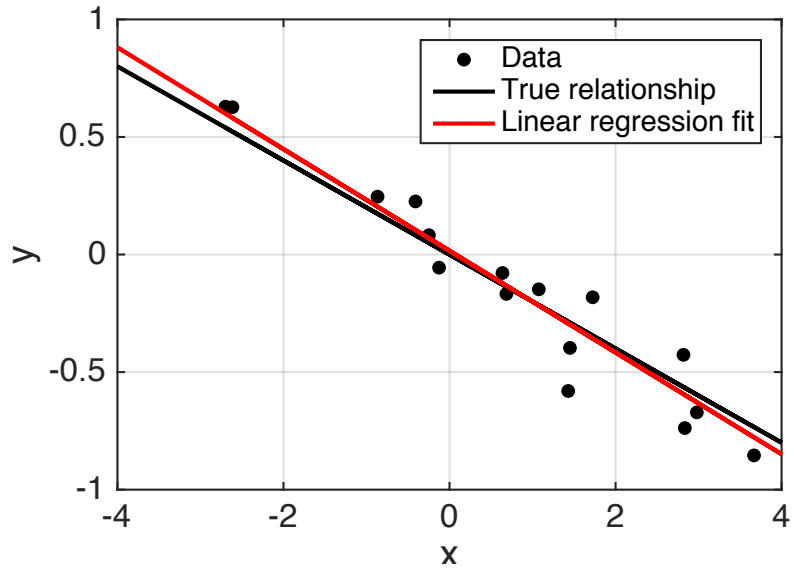
	Estimate	SE	tStat	pValue
(Intercept)	-0.0088005	0.006932	-1.2695	0.20574
x1	-0.19906	0.006384	-31.182	2.521e-78

Effect of “leverage”

n = 20, σ = 0.1



	Estimate	SE	tStat	pValue
(Intercept)	0.011361	0.025885	0.4389	0.66596
x1	0.16244	0.026655	6.0942	9.2986e-06



	Estimate	SE	tStat	pValue
(Intercept)	0.015844	0.028063	0.5646	0.57931
x1	-0.21624	0.0088359	-24.473	2.8799e-15

Confidence intervals

- The regression coefficients are our best guess about the true values of the parameters
- The confidence interval defines a range of values that could be consistent with the data
- 95% confidence interval is given by ± 2 SE:

$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1).$$

- Note: This is based on some assumptions, but it is a good approximation as long as n is not too small