Intro to Machine Learning

Lecture 1

Adil Khan

Objectives

- Importance of and reasons for machine learning
- What is learning (a very simple examples)
- Different types of learning
- Predictors and response variables
- Regression and classification
- Goals of learning
- Parametric and non-parametric models
- Assessing the quality of learning

Your First Day at Job!!!

Classifying emails as “Spam” or “Not Spam”

Can you do simple if/else?

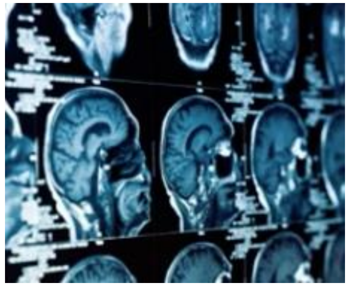
Classifying emails as “Spam” or “Not Spam”

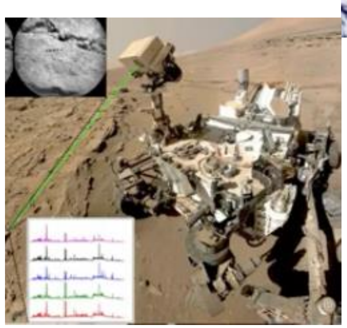
What you need is … Machine Learning

Classifying emails as “Spam” or “Not Spam”

Applications of Machine Learning

NETFLIX





What do People Think About ML?

"A breakthrough in machine learning would be worth ten Microsofts"

– (Bill Gates, Chairman, Microsoft)

"Machine learning is the next Internet"

– (Tony Tether, Director, DARPA)

"Machine learning is the hot new thing"

– (John Hennessy, President, Stanford)

"Web rankings today are mostly a matter of machine learning"

– (Prabhakar Raghavan, Dir. Research, Yahoo)

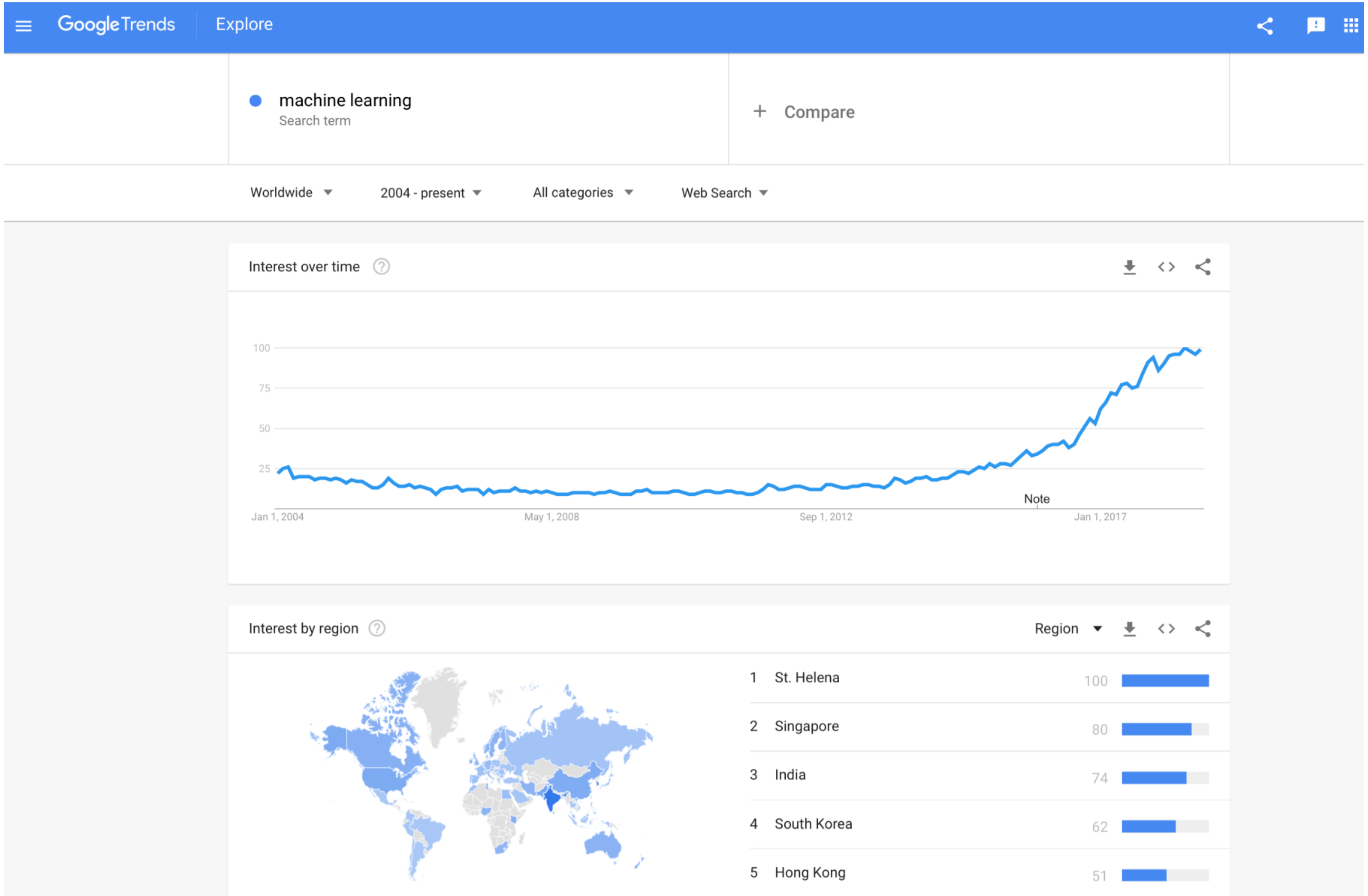
"Machine learning is going to result in a real revolution"

– (Greg Papadopoulos, CTO, Sun)

"Machine learning is today's discontinuity"

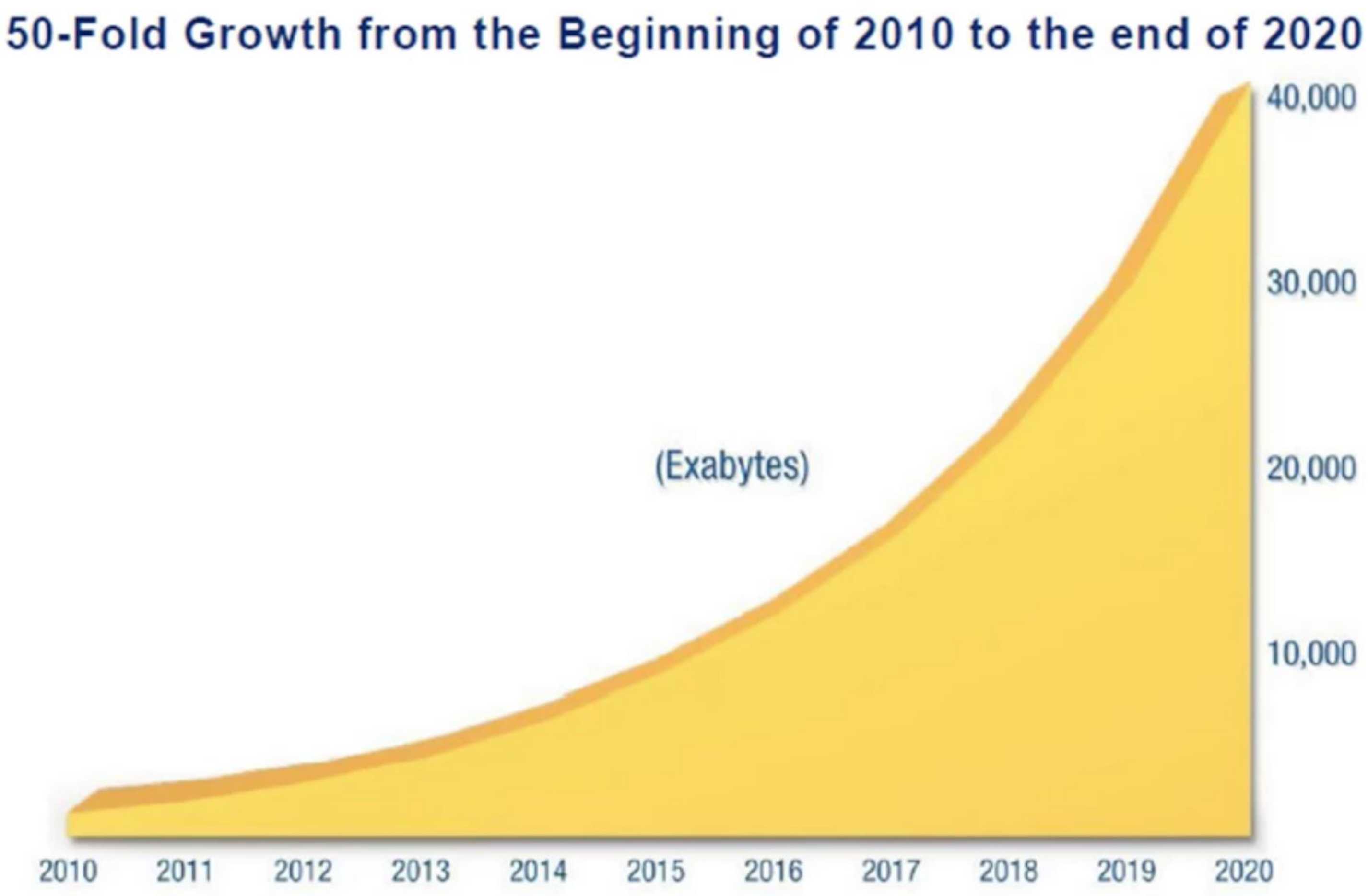
– (Jerry Yang, CEO, Yahoo)

Worldwide Trends



Why This Interest?

Data Growth!



Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

What is Learning?
A Simple Classification Example

- Good and Bad Bananas

A Simple Model

Banana is good if it has

this much yellow in skin +

this much sweet taste + this much squishiness

A Simple Linear Model

Banana is good if it has

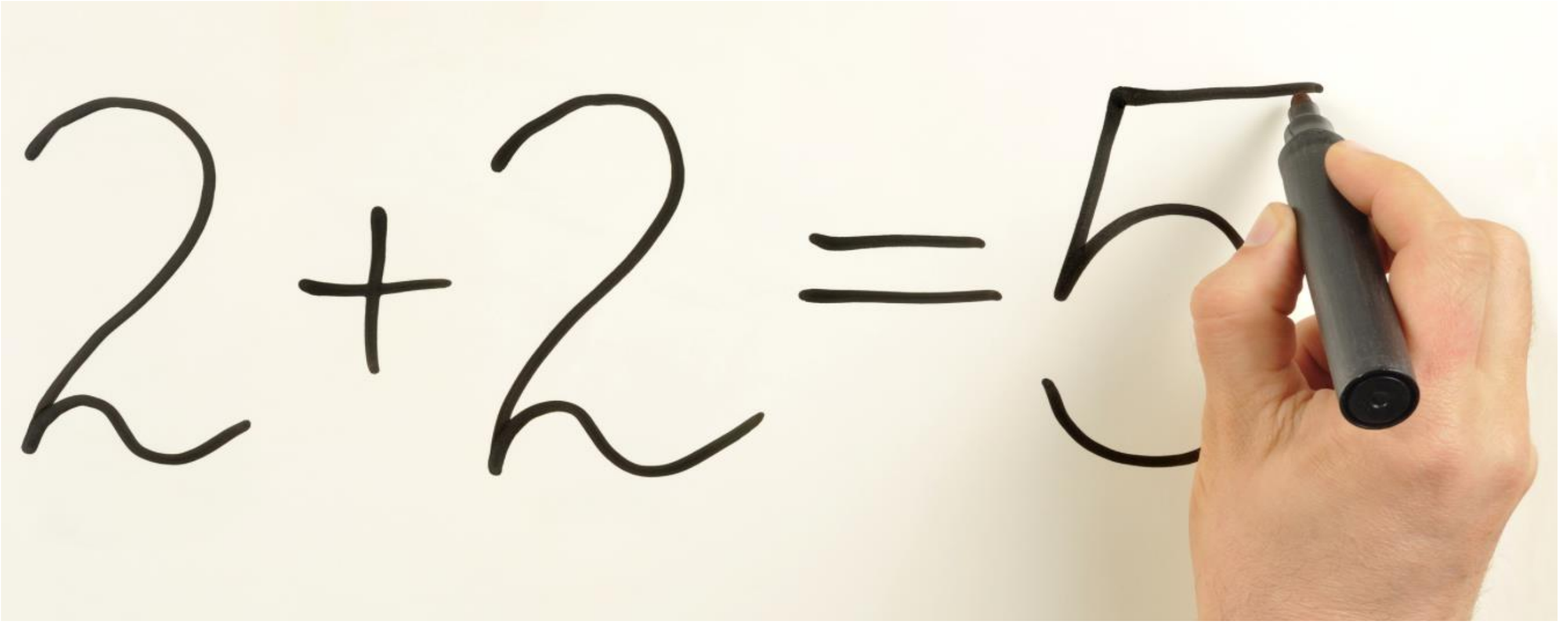
this much yellow in skin + this much sweet taste + this much squishiness

$$Model \approx (w_{skin}, w_{sweet}, w_{squishy})$$

Machine Learning is about learning the right values of these weights!

Model Learning

How do humans learn?



Model Learning

That's it:

1. Start with Random values,
2. Make predictions, and see if you made a mistake
3. Use mistake to guide you in the **right direction** with **right amount**

Supervised Learning

Take-Away Lesson

- Machine Learning (for supervised classification) involves

  1. choosing (or learning) features/descriptors
  2. And searching the weights for those features (or their combinations)

- Search is guided by mistakes

Question

- I used the term “Supervised Learning”, are there other kinds of learning, too?

Types of Learning

- Supervised
- Unsupervised
- Semi-Unsupervised

Supervised Learning

Label	Rμ	Gμ	Bμ	Rσ	Gσ	Bσ
Cat	57.61	41.36	132.44	158.33	149.86	93.33
Cat	120.23	121.59	181.43	145.58	69.13	116.91
Cat	124.15	193.35	65.77	23.63	193.74	162.70
Dog	100.28	163.82	104.81	19.62	117.07	21.11
Dog	177.43	22.31	149.49	197.41	18.99	187.78
Dog	149.73	87.17	187.97	50.27	87.15	36.65

Unsupervised Learning

Rμ	Gμ	Bμ	Rσ	Gσ	Bσ
57.61	41.36	132.44	158.33	149.86	93.33
120.23	121.59	181.43	145.58	69.13	116.91
124.15	193.35	65.77	23.63	193.74	162.70
100.28	163.82	104.81	19.62	117.07	21.11
177.43	22.31	149.49	197.41	18.99	187.78
149.73	87.17	187.97	50.27	87.15	36.65

Semi-supervised Learning

Label	Rμ	Gμ	Bμ	Rσ	Gσ	Bσ
Cat	57.61	41.36	132.44	158.33	149.86	93.33
?	120.23	121.59	181.43	145.58	69.13	116.91
?	124.15	193.35	65.77	23.63	193.74	162.70
Dog	100.28	163.82	104.81	19.62	117.07	21.11
?	177.43	22.31	149.49	197.41	18.99	187.78
Dog	149.73	87.17	187.97	50.27	87.15	36.65

Supervised Learning

Input	Output	Application
Home Features	Price	Real Estate
Ad, User info	Click ad? (0/1)	Online Advertising
Image	Object (1, …, 1000)	Photo Tagging
Audio	Text Transcript	Speech Recognition
English	Chinese	Machine Translation
Image, Radar Info	Position of other cars	Autonomous Driving

A Simple Linear Model

Banana is good if it has

this much yellow in skin + this much sweet taste + this much squishiness

$$Model \approx (w_{skin}, w_{sweet}, w_{squishy})$$

Machine Learning is all about learning the right values of these weights!

Model Learning

That's it:

1. Start with Random values,
2. Make predictions, and see if you made a mistake
3. Use mistake to guide you in the **right direction**

Loss Function

Cost function

Supervised Learning

Example Dataset

Dependent Variable (Response)

Country	Age	Salary	Purchased
France	44	72000	No
Spain	27	48000	Yes
Germany	30	54000	No
Spain	38	61000	No
Germany	40		Yes
France	35	58000	Yes
Spain		52000	No
France	48	79000	Yes
Germany	50	83000	No
France	37	67000	Yes

Independent Variables (Predictors)

Example Datasets

Country	Age	Salary	Purchased
France	44	72000	No
Spain	27	48000	Yes
Germany	30	54000	No
Spain	38	61000	No
Germany	40		Yes
France	35	58000	Yes
Spain		52000	No
France	48	79000	Yes
Germany	50	83000	No
France	37	67000	Yes

YearsExperience	Salary
1.1	39343
1.3	46205
1.5	37731
2	43525
2.2	39891
2.9	56642
3	60150
3.2	54445
3.2	64445

Regression and Classification

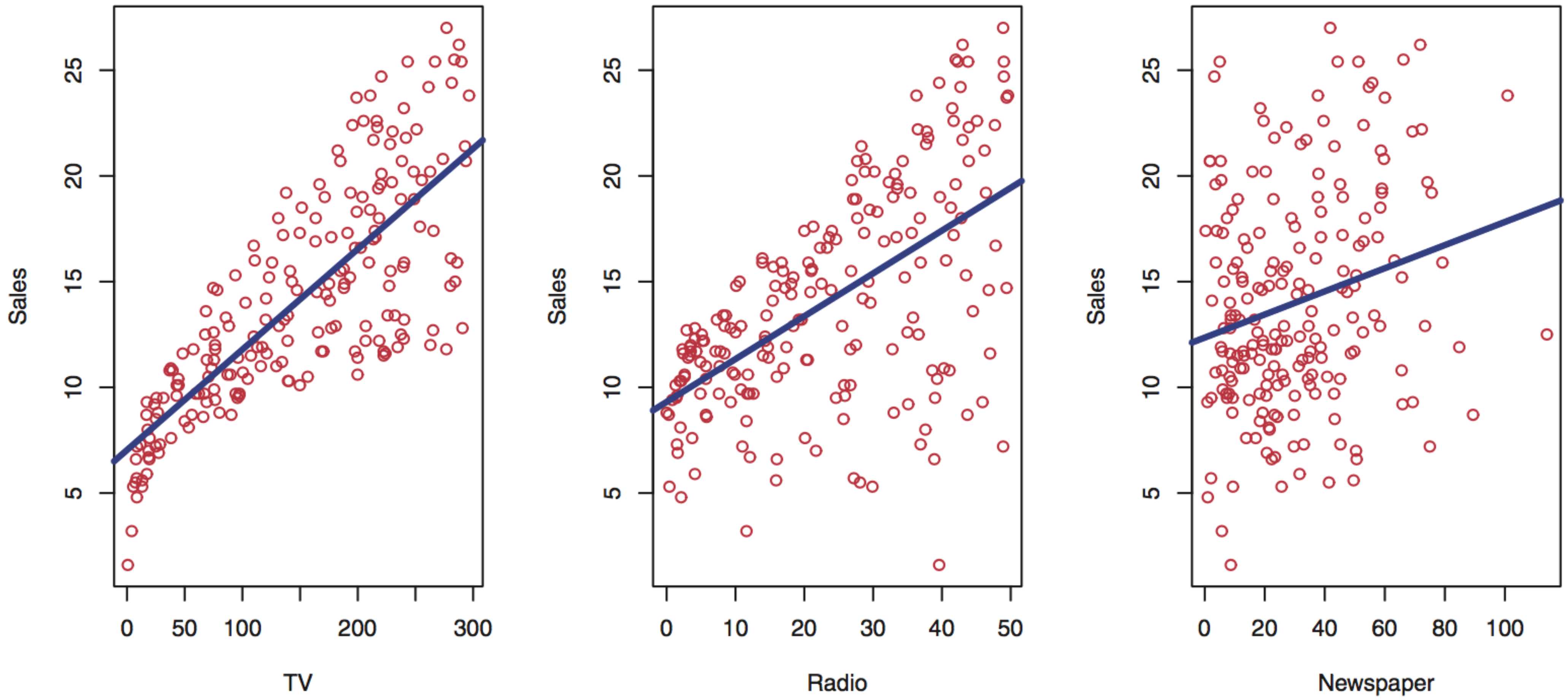
Country	Age	Salary	Purchased
France	44	72000	No
Spain	27	48000	Yes
Germany	30	54000	No
Spain	38	61000	No
Germany	40		Yes
France	35	58000	Yes
Spain		52000	No
France	48	79000	Yes
Germany	50	83000	No
France	37	67000	Yes

Classification

YearsExperience	Salary
1.1	39343
1.3	46205
1.5	37731
2	43525
2.2	39891
2.9	56642
3	60150
3.2	54445
3.2	64445

Regression

Regression



Sales
25
20
15
10
5
0
50
100
200
300
TV
Sales
25
20
15
10
5
0
10
20
30
40
50
Radio
Sales
25
20
15
10
5
0
20
40
60
80
100
Newspaper

Mathematically

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$$

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

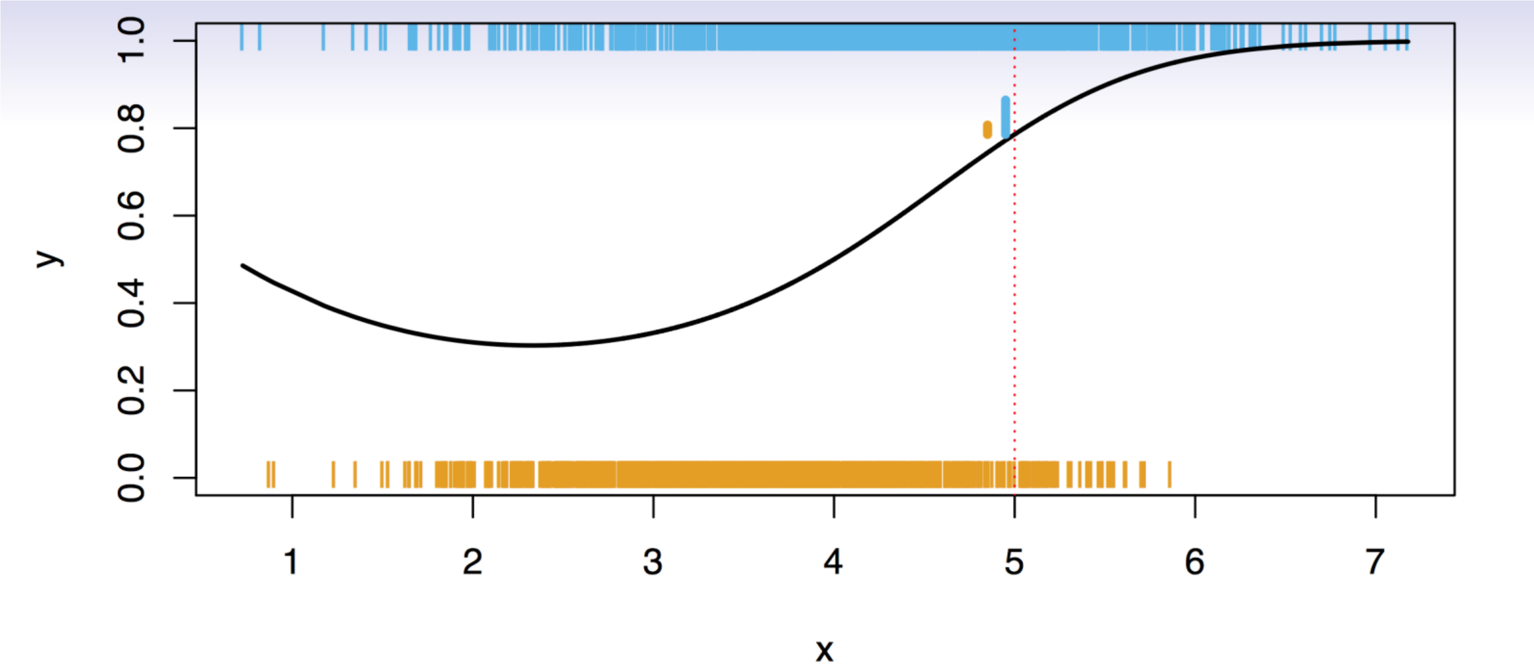
Goal of learning

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$$

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

$$\widehat{f} \approx f$$

Classification



Mathematically

$$X = (X_1)$$

$$Y = C(X) + \epsilon$$

Goal of learning

$$X = (X_1)$$

$$Y = C(X) + \epsilon$$

$$\widehat{C} \approx C$$

Why estimate f or C?

- Prediction
  - In many situations, a set of inputs X are readily available, but the output Y cannot be easily obtained. In this setting, since the error term averages to zero, we can predict Y using

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

- Inference
  - Which predictors are associated with the response?
  - What is the relationship between the response and each predictor
  - Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?
  - Etc.

How do we estimate?

1. Parametric methods

2. Non-parametric methods

Parametric methods (1)

- First, we make an assumption about the functional form, or shape, of f. For example

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

- This is a linear model.

- The problem of estimating f is greatly simplified. Instead of having to estimate an entirely arbitrary p-dimensional function f, one only needs to estimate the $p + 1$ coefficients

Parametric methods (2)

- After a model has been selected, we need a procedure that uses the training data to fit or train the model

- For example, ordinary least squares method, gradient descent, etc.

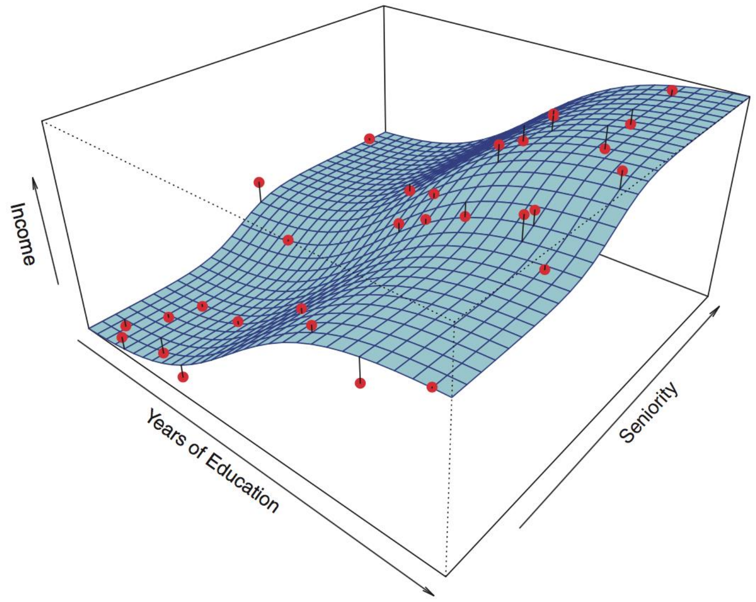
Parametric methods (3)

- Thus parametric approach reduces the problem of estimating f down to one of estimating a set of parameters.

- Assuming a parametric form for f simplifies the problem of estimating f because it is generally much easier to estimate a set of parameters

- The potential disadvantage of a parametric approach is that the model we choose will usually not match the true unknown form of f

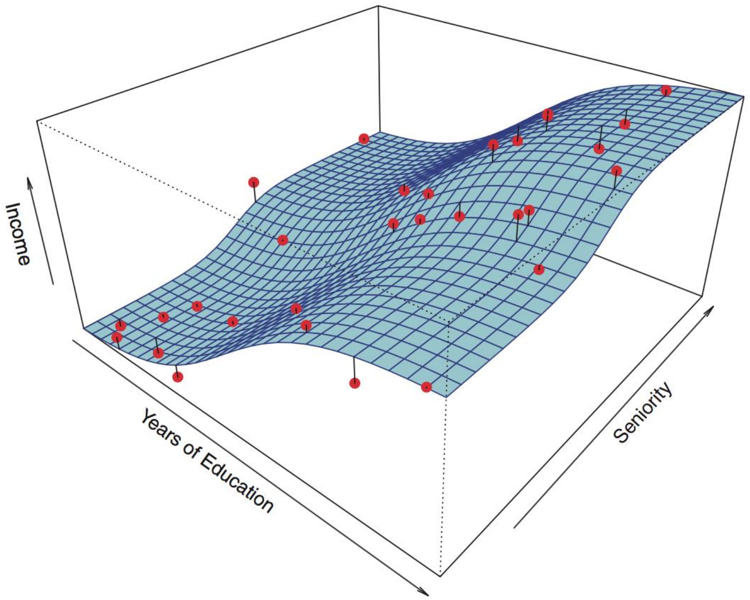
Example (1)



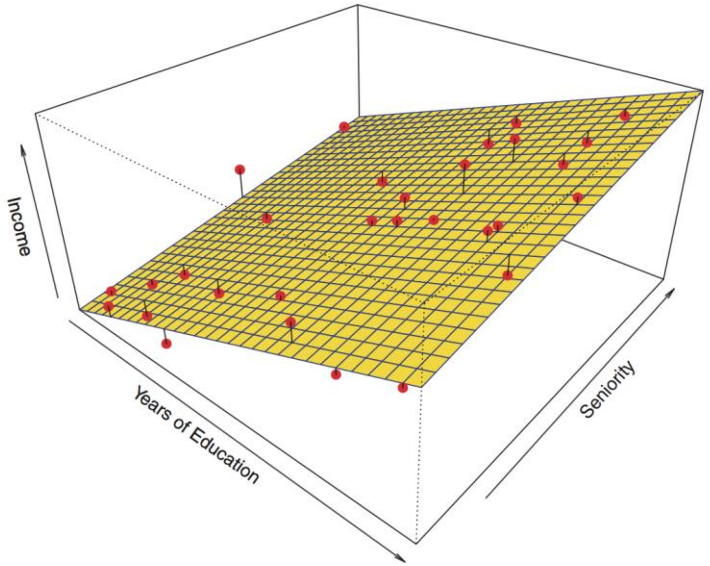
(a)

The plot displays income as a function of years of education and seniority in the Income data set. The blue surface represents the true underlying relationship between income and years of education and seniority, which is known since the data are simulated. The red dots indicate the observed values of these quantities for 30 individuals.

Example (2)



(a)



(b)

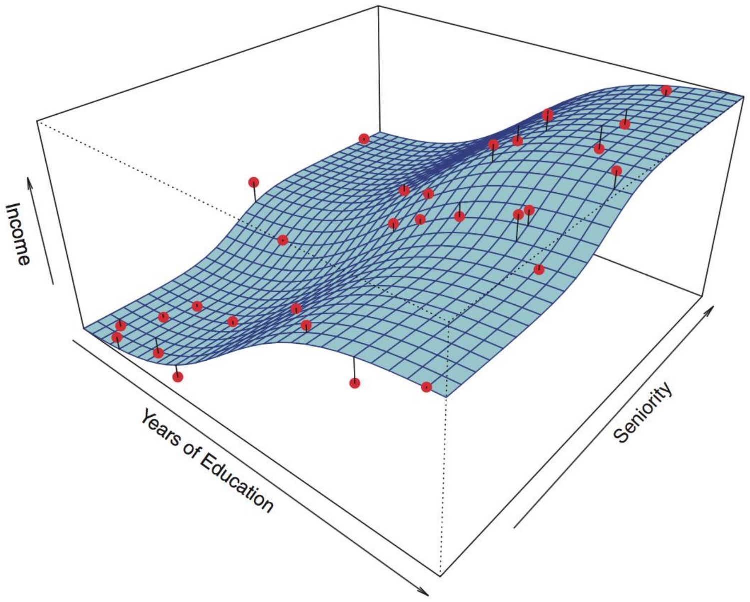
A linear model fit by least squares to the Income data from (a). The observations are shown in red, and the yellow plane indicates the least squares fit to the data.

Non-parametric methods (1)

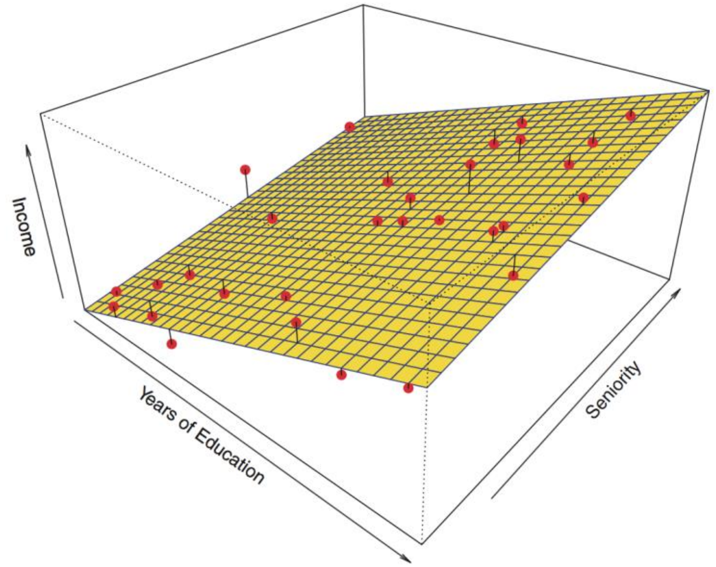
- Non-parametric methods do not make explicit assumptions about the functional form of f.

- Instead they seek an estimate of f that gets as close to the data points as possible without being too rough or wiggly

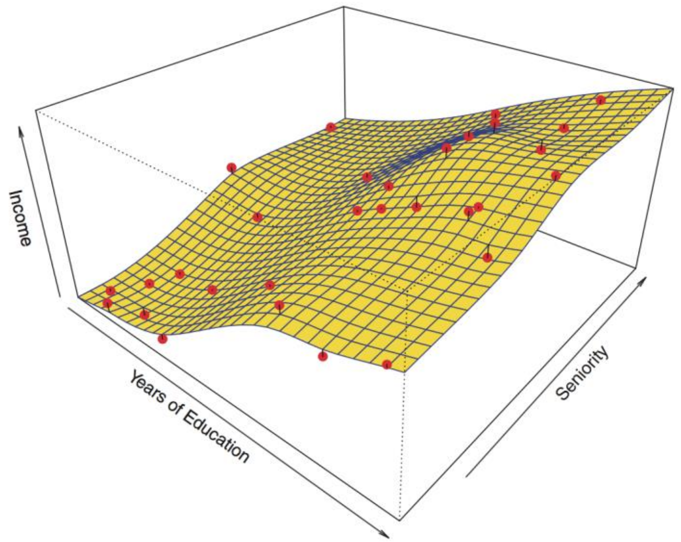
Example (3)



(a)



(b)



(c)

A smooth thin-plate spline fit to the Income data from (a) is shown in yellow; the observations are displayed in red.

Non-parametric methods (2)

- But non-parametric approaches do suffer from a major disadvantage:
  - since they do not reduce the problem of estimating f to a small number of parameters, a very large number of observations (far more than is typically needed for a parametric approach) is required in order to obtain an accurate estimate for f.

Assessing the quality of learning (1)

- Let $T_r = \{x_i, y_i\}_1^N$ be the training data use to estimate $\hat{f}(x)$. To assess the quality of estimate, we can compute

$$\mathrm{MSE}_{\mathsf{Tr}} = \mathrm{Ave}_{i \in \mathsf{Tr}}[y_i - \hat{f}(x_i)]^2$$

- But this is not a reliable way, why?

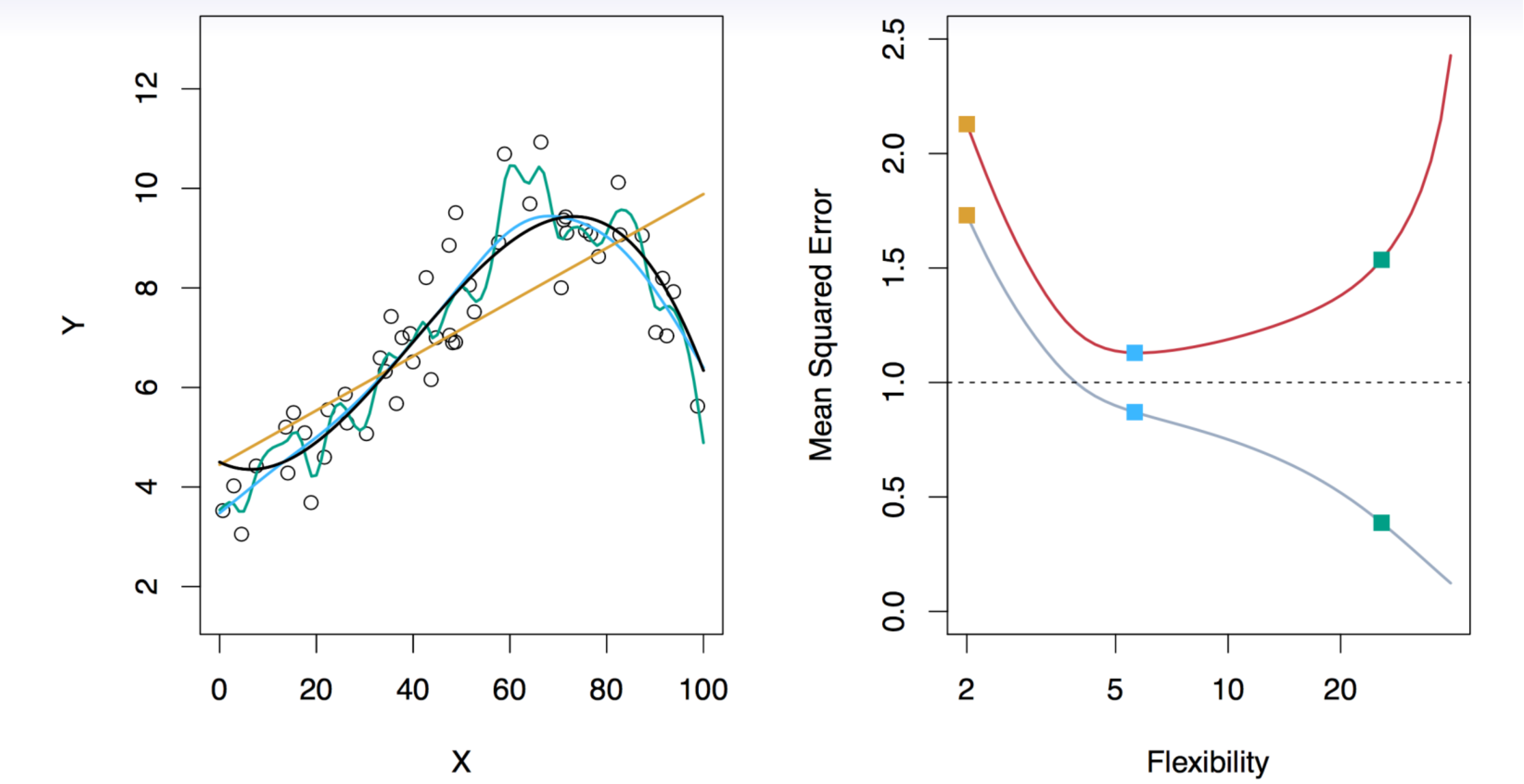
Assessing the quality of learning (2)

- Thus, if possible, we should try to use the test data $T_e = \{x_i, y_i\}_1^M$

$$\mathrm{MSE}_{\mathsf{Te}} = \mathrm{Ave}_{i \in \mathsf{Te}}[y_i - \hat{f}(x_i)]^2$$

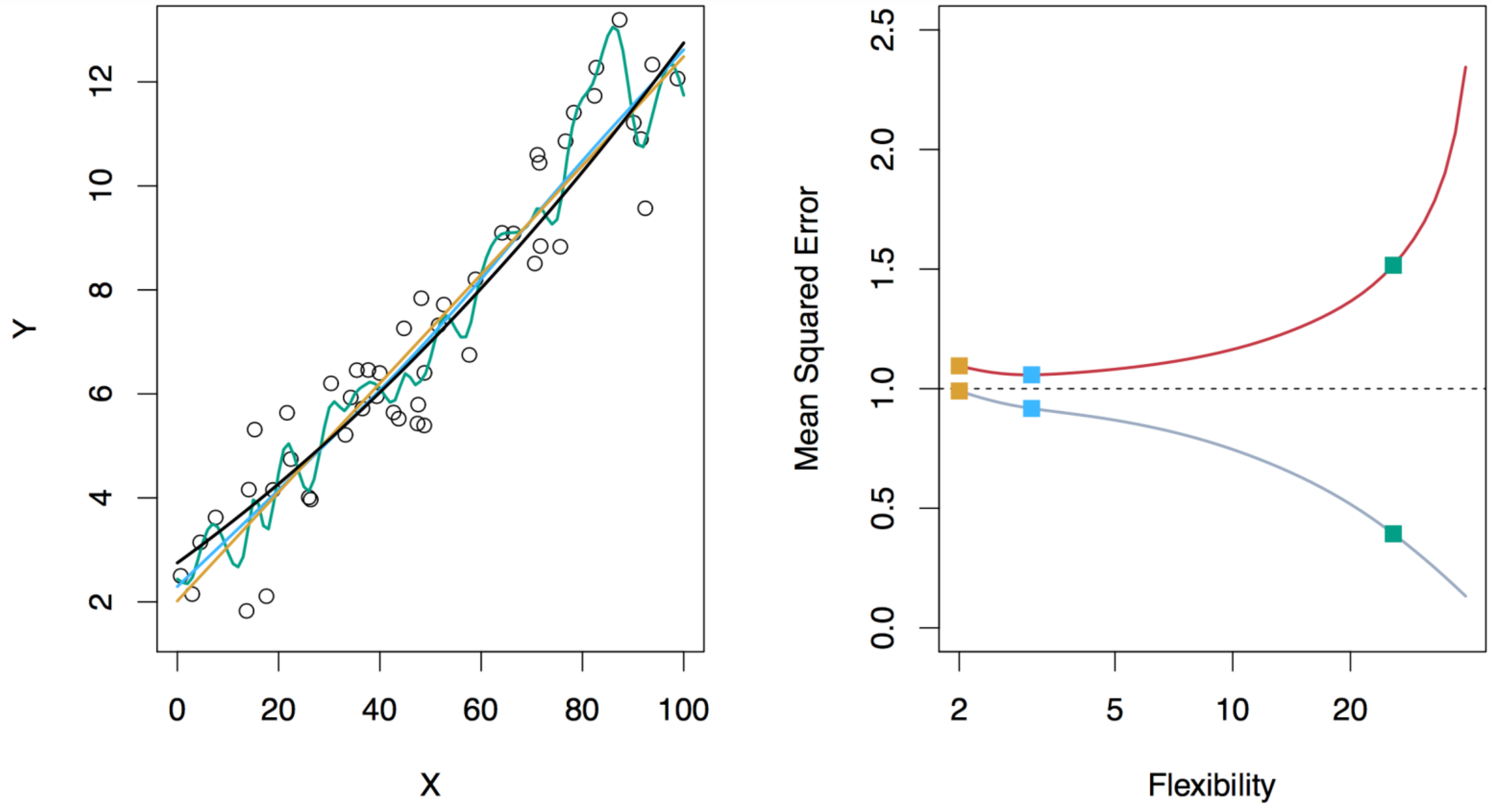
- But this is not a reliable way, why?

Example (1)



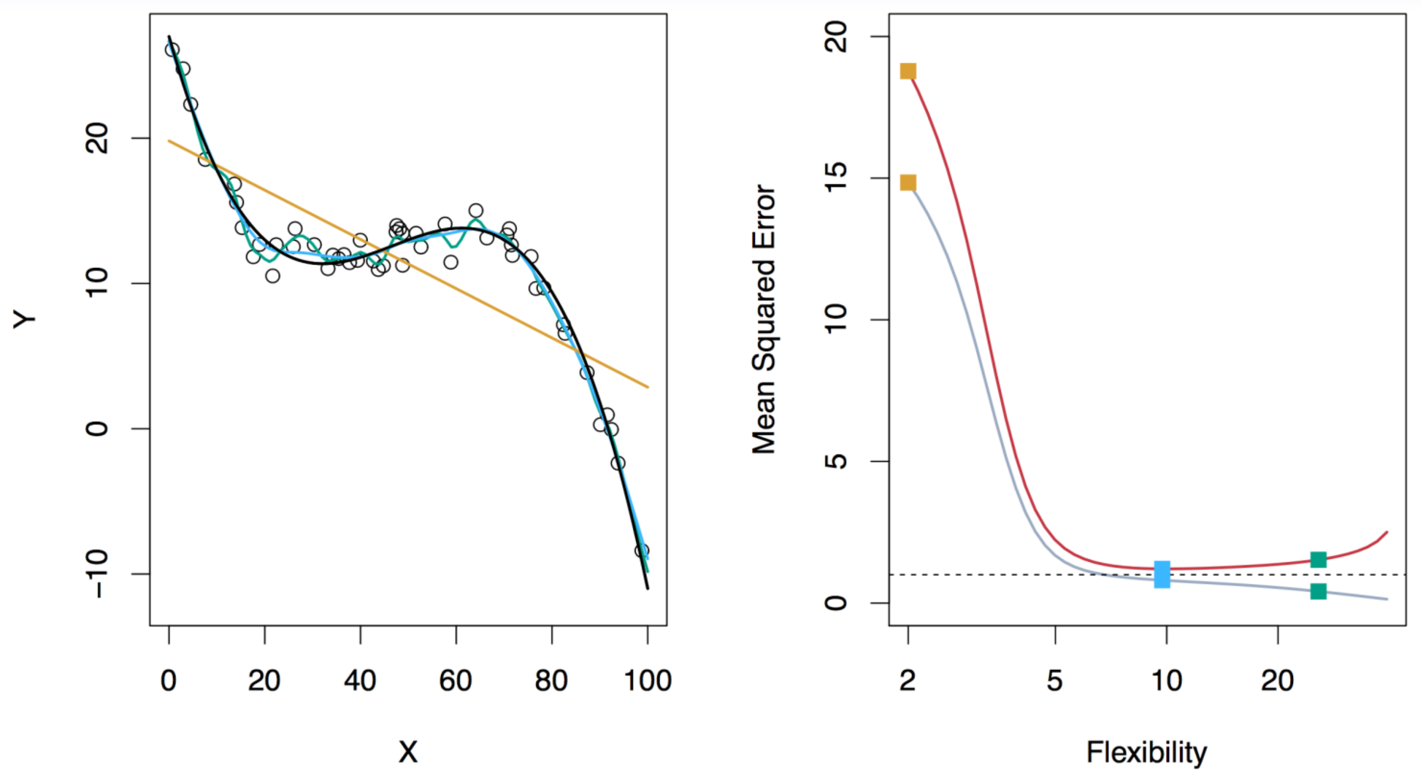
Black curve is truth. Red curve on right is MSE_{Te}, grey curve is MSE_{Tr}. Orange, blue and green curves/squares correspond to fits of different flexibility.

Example (2)



Here the truth is smoother, so the smoother fit and linear model do really well.

Example (3)



Here the truth is wiggly and the noise is low, so the more flexible fits do the best.

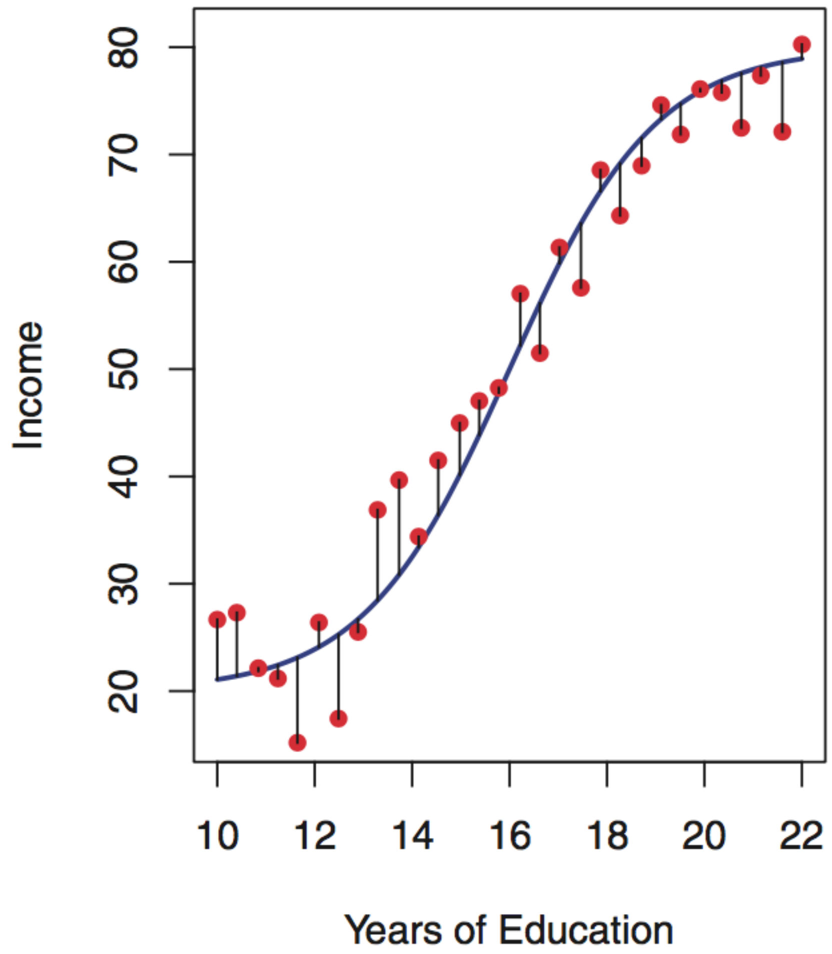
Bias-variance Tradeoff (1)

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

Where (x_0, y_0) is a test observation

Typically as the flexibility of fˆ increases, its variance increases, and its bias decreases. So choosing the flexibility based on average test error amounts to a bias-variance trade-off.

$Var(\epsilon)$



- The red dots are the observed values of income (in tens of thousands of dollars) and years of education for 30 individuals.

- The blue curve represents the true underlying relationship between income and years of education, which is generally unknown (but is known in this case because the data were simulated).

- The black lines represent the error associated with each observation. Note that some errors are positive (if an observation lies above the blue curve) and some are negative (if an observation lies below the curve). Overall, these errors have approximately mean zero.

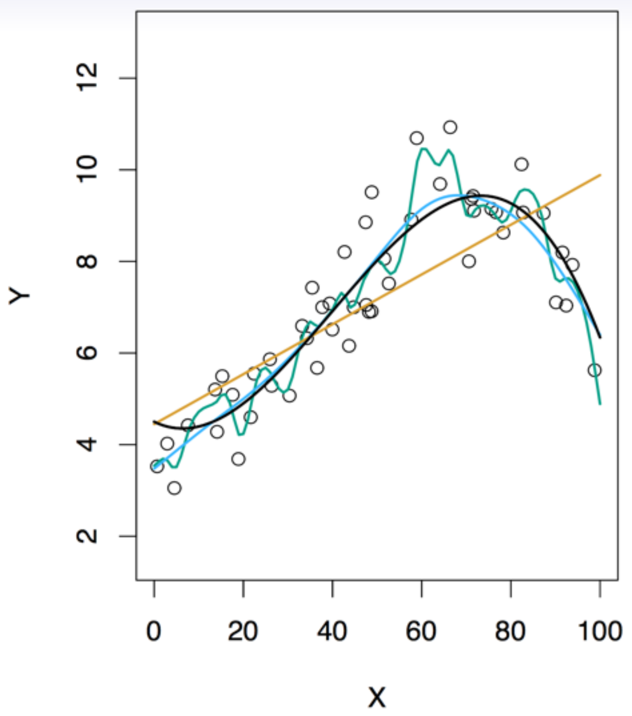
Bias-variance Tradeoff (2)

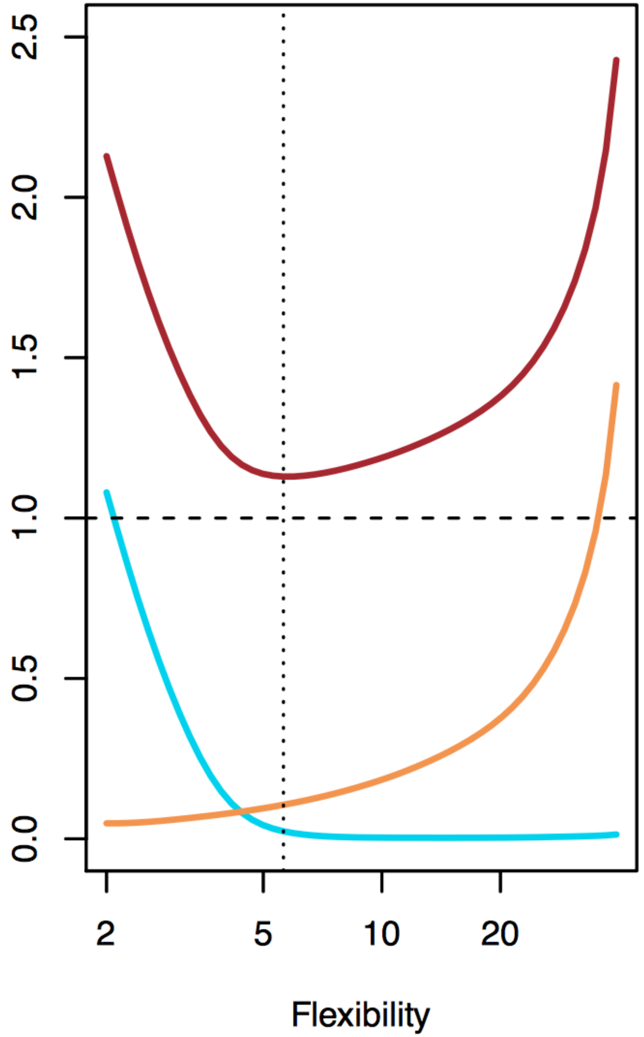
$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon).$$

Where (x_0, y_0) is a test observation

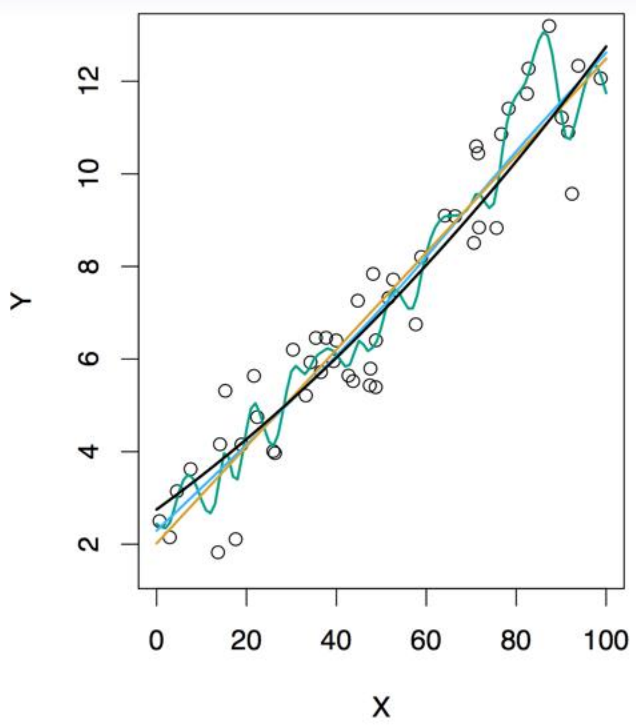
Typically as the flexibility of fˆ increases, its variance increases, and its bias decreases. So choosing the flexibility based on average test error amounts to a bias-variance trade-off.

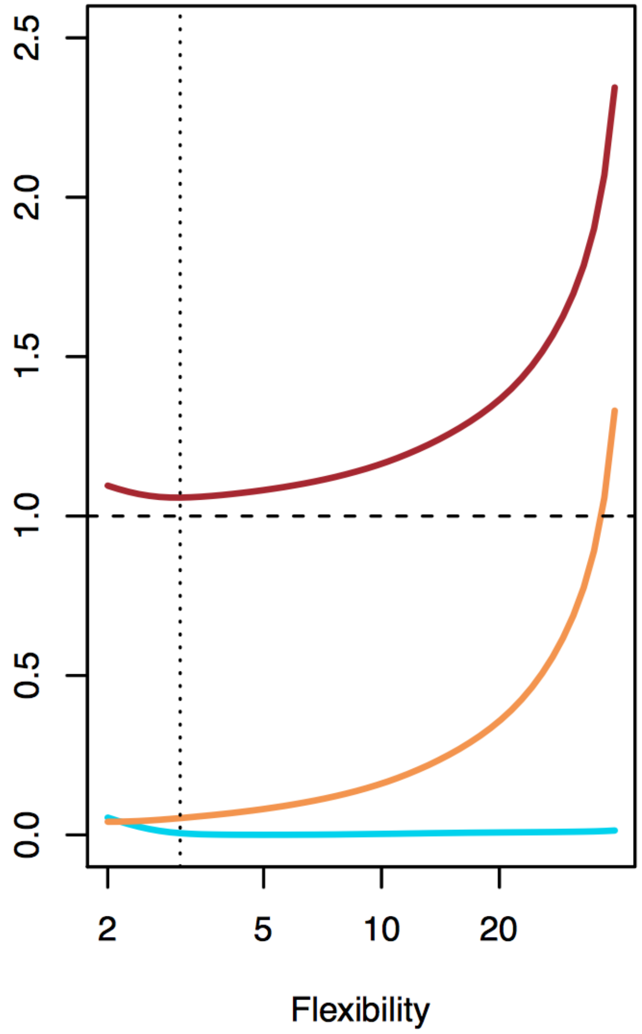
Example (1) Bias-variance Tradeoff



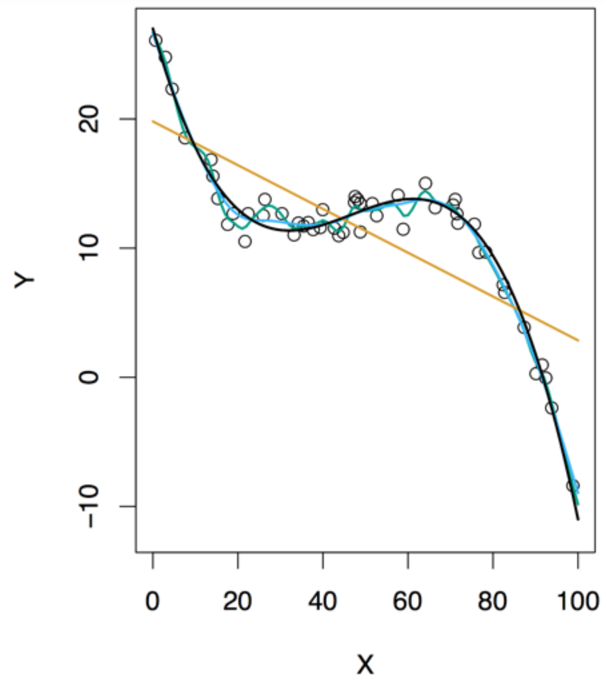


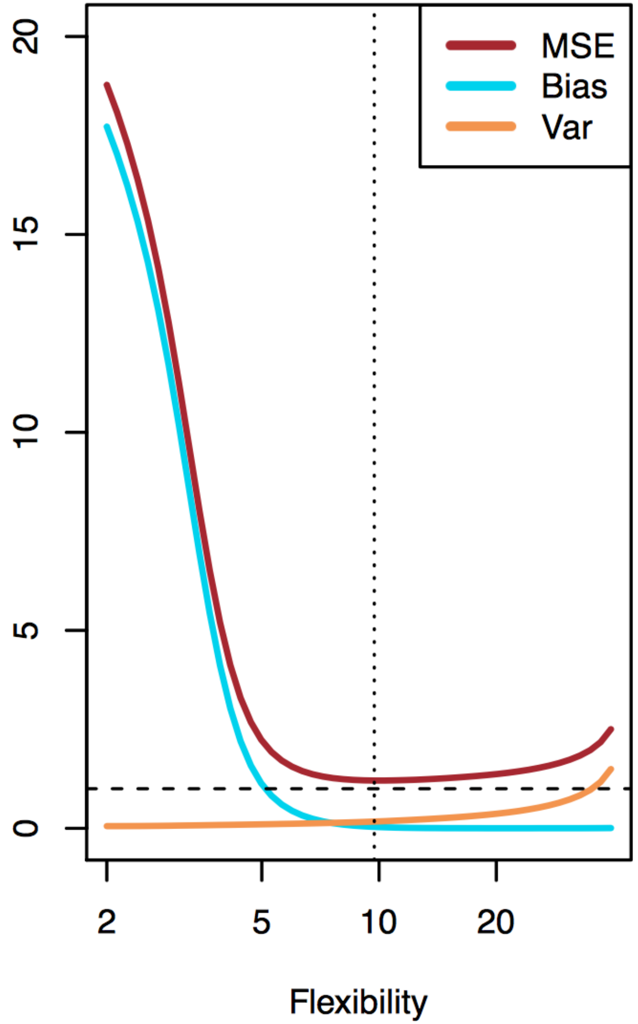
Example (2) Bias-variance Tradeoff





Example (3) Bias-variance Tradeoff





Did we achieve today's objectives?

- Importance of and reasons for machine learning
- What is learning (a very simple examples)
- Different types of learning
- Predictors and response variables
- Regression and classification
- Goals of learning
- Parametric and non-parametric models
- Assessing the quality of learning