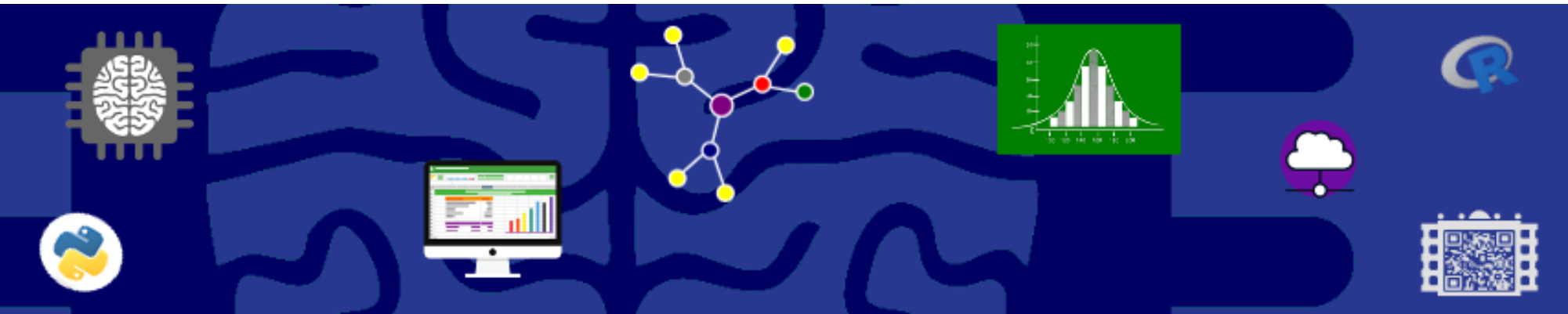
Introduction to Data Science

Winter Semester 2018/19

Oliver Ernst

TU Chemnitz, Fakultät für Mathematik, Professur Numerische Mathematik

Lecture Slides



Contents I

Contents II

Contents III

Contents

Contents

- Given a data set containing the sales numbers for a given product in 200 markets, allocate an advertising budget across the three media channels **TV**, **radio** and **newspaper**.
- The sales numbers for each medium are available for different advertising budget values.
- We will try to model the dependence of sales on advertising budgets.
- Terminology:

X_1 : TV budget
X_2 : radio budget
X_3 : newpaper budget

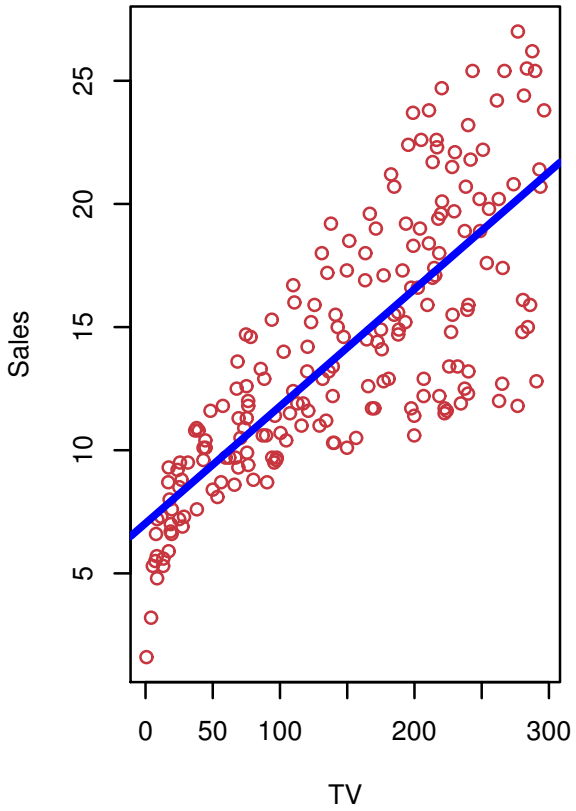
} input variables, predictors, independent variables, variables, features

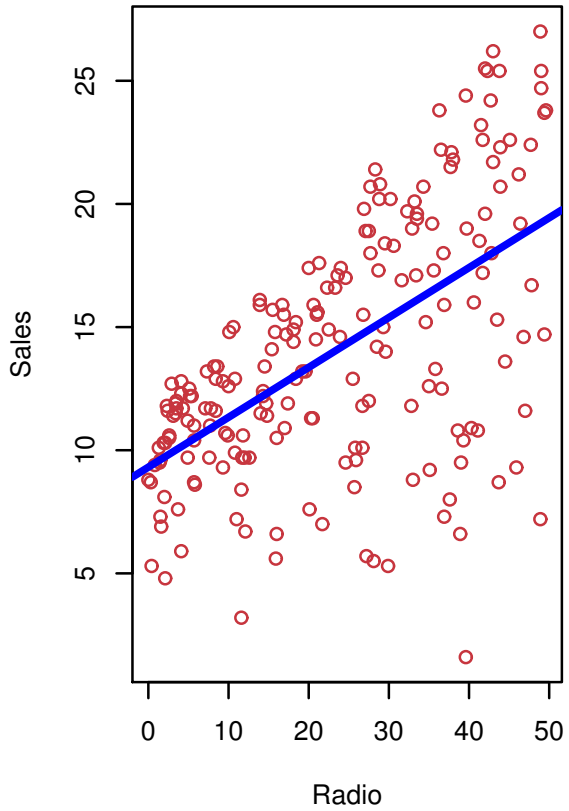
Y : sales

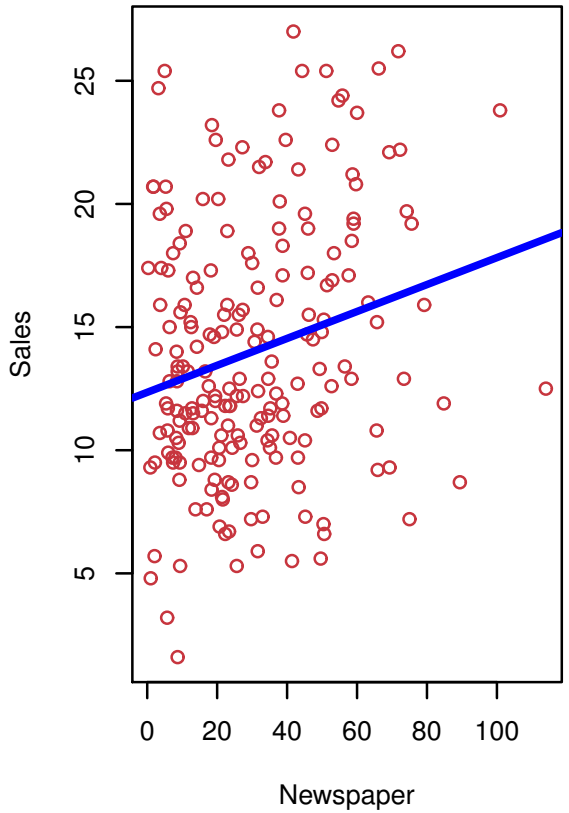
response, dependent variables.

Learning Theory

Example: Advertising channels







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

$X = (X_1, \ldots, X_p)$, $p = \#$ predictors,

ε : random error term, $\mathbf{E}[\varepsilon] = 0$, (2.1)

f : systematic information X provides about Y.

Learning Theory

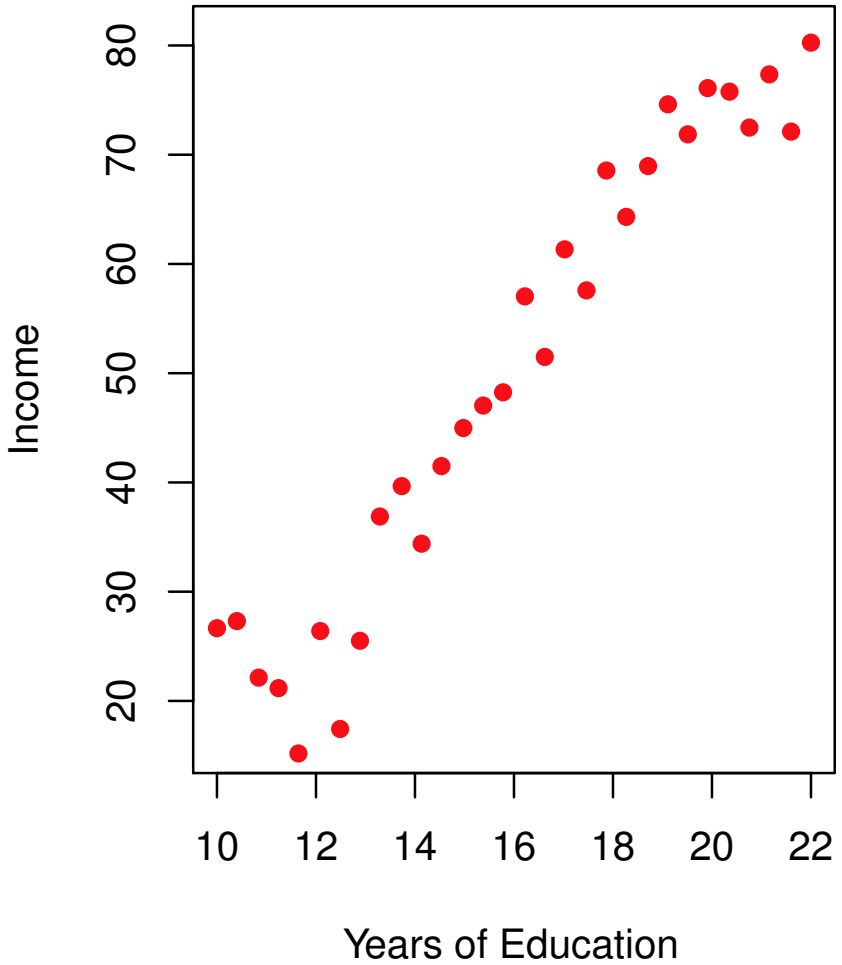
Example: Income

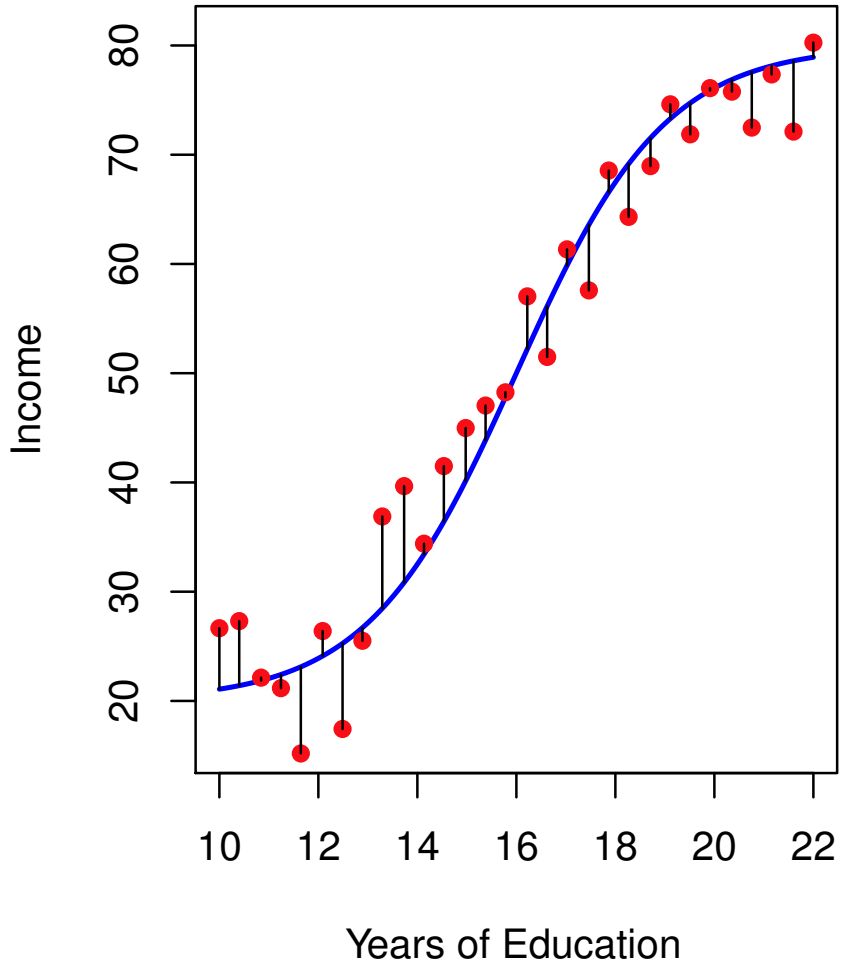
- Given data set showing **income** against **years of education** for 30 people.
- Objective: determine function f relating **income** as response to **years of education** as predictor.
- f generally unknown, must be estimated from the data.
- Here: data simulated, so f available.
- In another data set, **income** is given with respect to two input variables: **years of education** and **seniority**.

- Statistical learning is concerned with techniques for estimating f from a data set.

Learning Theory

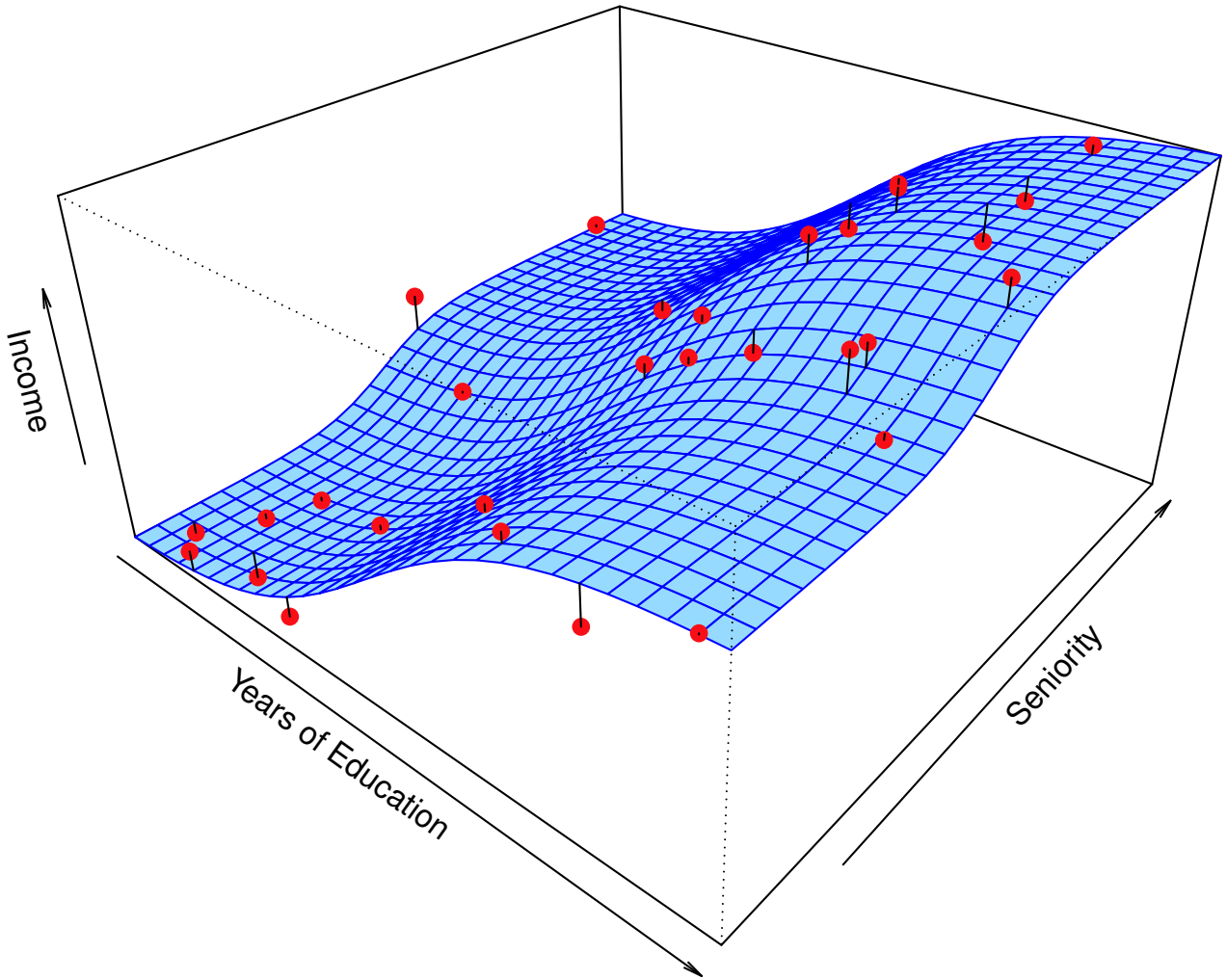
Example: Income





Learning Theory

Example: Income



Two main reasons for estimating f:
prediction
and
inference.

- Suppose inputs X readily available, but outputs Y difficult to obtain.
- Since errors average out, predict Y using

$$\hat{Y} = \hat{f}(X), \qquad \begin{aligned} &\hat{f} : \text{estimate for } f, \\ &\hat{Y} : \text{prediction for } Y = f(X). \end{aligned}$$

- Often $\hat{f}$ only available as a **black box**, i.e., a procedure for generating $\hat{Y}$ given X.

Example:

$X_1, \ldots, X_p$: characteristics of a patient's blood samples, measured in lab.
Y : patient's risk for severe adverse reaction to particular drug.

For obvious reasons, having an accurate estimate $\hat{Y} = \hat{f}(X)$ is preferable to evaluating $Y = f(X)$.

Accuracy of $\hat{Y} \approx Y$ depends on **reducible error** and **irreducible error**.

- **reducible error**: $f - \hat{f}$. Can be made smaller and smaller by employing increasingly sophisticated statistical learning techniques.
- **irreducible error**: ε. Present even for $f = \hat{f}$, cannot be predicted from X. Possible sources:
  - Additional variables Y may depend on but which are not observed/measured.
  - Unmeasurable variation.
    (E.g.: Adverse reaction may depend on manufacturing variations in drug or variations in patient's sensitivity over time.)
- Quantitative measure: **mean square error** (MSE)

$$\mathbf{E}\left[(Y - \hat{Y})^2\right] = \underbrace{[f(X) - \hat{f}(X)]^2}_{\textit{reducible}} + \underbrace{\mathbf{Var}\,\varepsilon}_{\textit{irreducible}}$$

- **Note:** irreducible error always lower bound on prediction accuracy.

Inference seeks to determine how the individual predictors $X_1, \ldots, X_p$ affect the response Y. In particular, this involves more detailed knowledge about $\hat{f}$ than simply considering it a black box.

Things to investigate:

- *Identify those predictors with the strongest effect on Y.*
  Can be a small subset of $X_1, \ldots, X_p$.
- *Determine relationship between response and each predictor.*
  Is it monotone increasing or decreasing with respect to an individual predictor? For more complex dependencies, such monotonicities can be affected by on the values of the remaining predictors.
- *Is a linear model sufficient?* Historically, most estimation methods have produced a linear (affine) function $\hat{f}$. If the true dependence of Y on X is more complicated, a linear model may not be accurate enough.

Learning Theory

Prediction example: direct-marketing campaign

- Company plans a direct-marketing campaign, wishes to identify individuals who would respond positively to a mailing.
- Respone is $Y \in \{\text{positive},\ \text{negative}\}$.
- Predictors X_j are demographic variables.
- Detailed relationship of response to demographic variables not of interest.
- A model which generates accurate predictions is all that is needed..

- In our first example (advertising by three media channels TV, radio and newspaper), one may also be interested in answers to
  - Which media increase sales?
  - Of those, which has the strongest positive effect?
  - At what rate do sales increase when the TV budget is raised?

- Another example: model brand of a product chosen by a customer as a function of predictor variables **price**, **store location**, **discount levels**, **competitor pricing** etc.
  Here detailed knowledge of how each variable affects outcome is of interest, e.g.
  - *What effect will changing the price of a product have on sales?*

Learning Theory

Example: combination of prediction and infrerence

- There are also **mixed situations**, involving both prediction and inference:
  Consider value of a house depending on prediction values **size**, **crime rate**, **zoning**, **distance from a river/ocean**, **air quality**, **schools**, **income level of community**, ...
  - *How much does an ocean view increase the value of a house?* (inference)
  - *Is this house over- or undervalued?* (prediction)

- **Note:** the two objectives may competing.
  Linear models allow for easier inference, but may not be accurate enough for given prediction goal.
  More sophisticated (highly nonlinear) approaches may yield high prediction accuracy, but the models they produce are often difficult to interpret.

Denote by

n : number of available data observations. ("**training data**")

x_{ij} : value of j-th predictor in i-th observation

$$i = 1, \ldots, n;\ j = 1, \ldots, p$$

y_i : value of response variable in i-th observation

Then training data consists of predictor-response pairs

$$\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}, \qquad \boldsymbol{x}_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix}.$$

Goal is estimating function $\hat{f}$ such that $Y \approx \hat{f}(X)$ for all observations (X, Y).

Two basic approaches: **parametric** and **non-parametric**.

Two-step model-based approach

1. Assume specific functional form for f, popular example is the **linear model**

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p. \tag{2.2}$$

   Estimation of function f now consists only in determining values of the $p+1$ **parameters** $\beta_0, \beta_1, \ldots, \beta_p$. (huge simplification)

2. **Train** or **fit** the chosen model to the data, i.e., choose parameters $\{\beta_j\}_{j=0}^{p}$ in order that (here for linear model (2.2))

$$f(X) \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p.$$

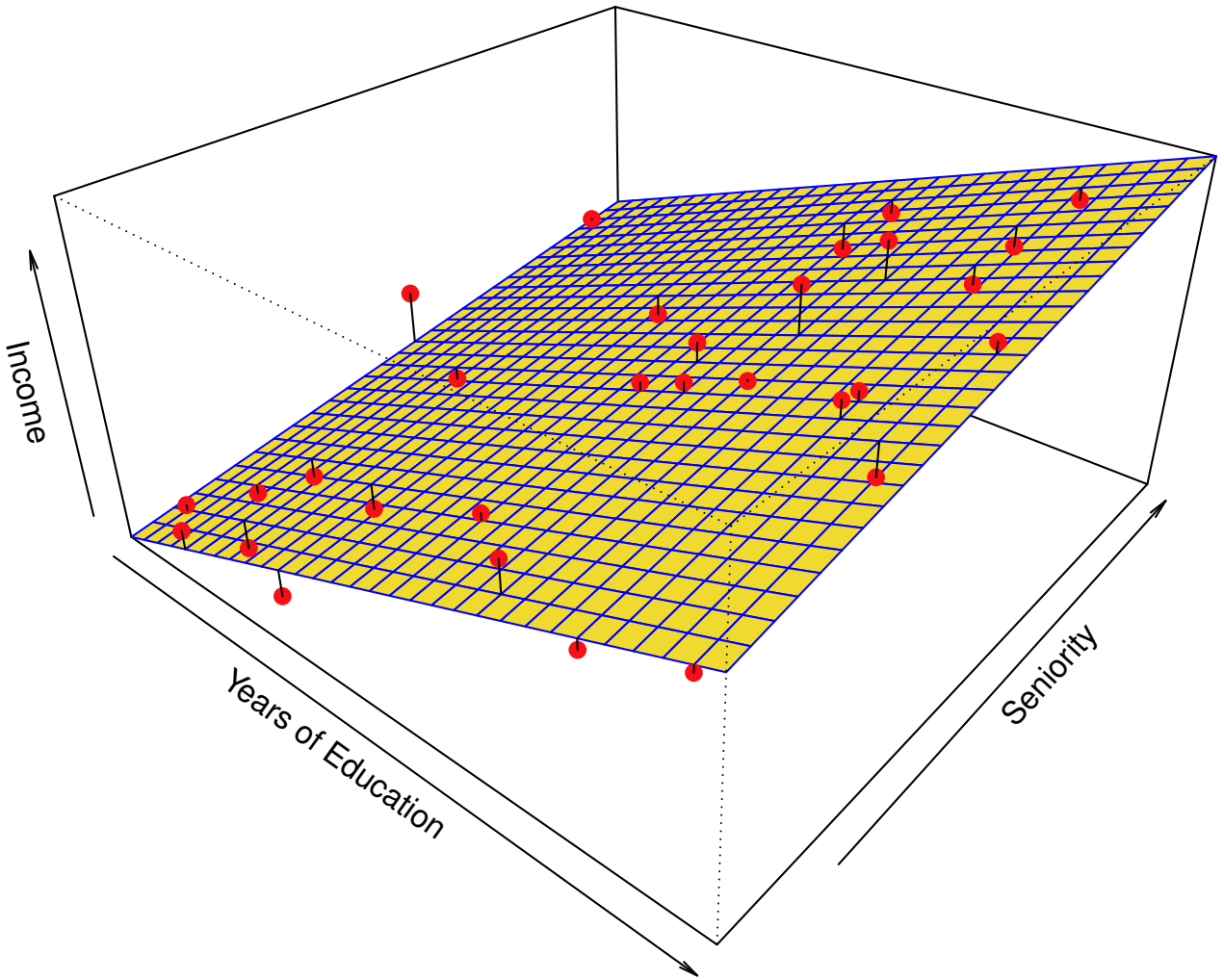
   Most common fitting technique: (ordinary) least squares, but many other techniques exist.

Problem of estimating f reduced to estimating a finite number of parameters.

Fundamental difficulty:

- Simplification comes at expense of strong restriction on type of dependence.
- For a bad choice, model cannot match the data well.
- More **flexible** models can better adapt to given data, but will generally involve more parameters to be estimated.
- Moreover, even if we are willing to fit our data extremely well with a flexible model, we may be adapting the model only to the fluctuations due to the random error contained in the data ("fitting the noise").
  In this case, our model will not **generalize** well, i.e., have a low prediction value for new data.
  This phenomenon is called **overfitting**.

Linear model: $\textbf{income} \approx \beta_0 + \beta_1 \times \textbf{education} + \beta_2 \times \textbf{seniority}$.



Learning Theory

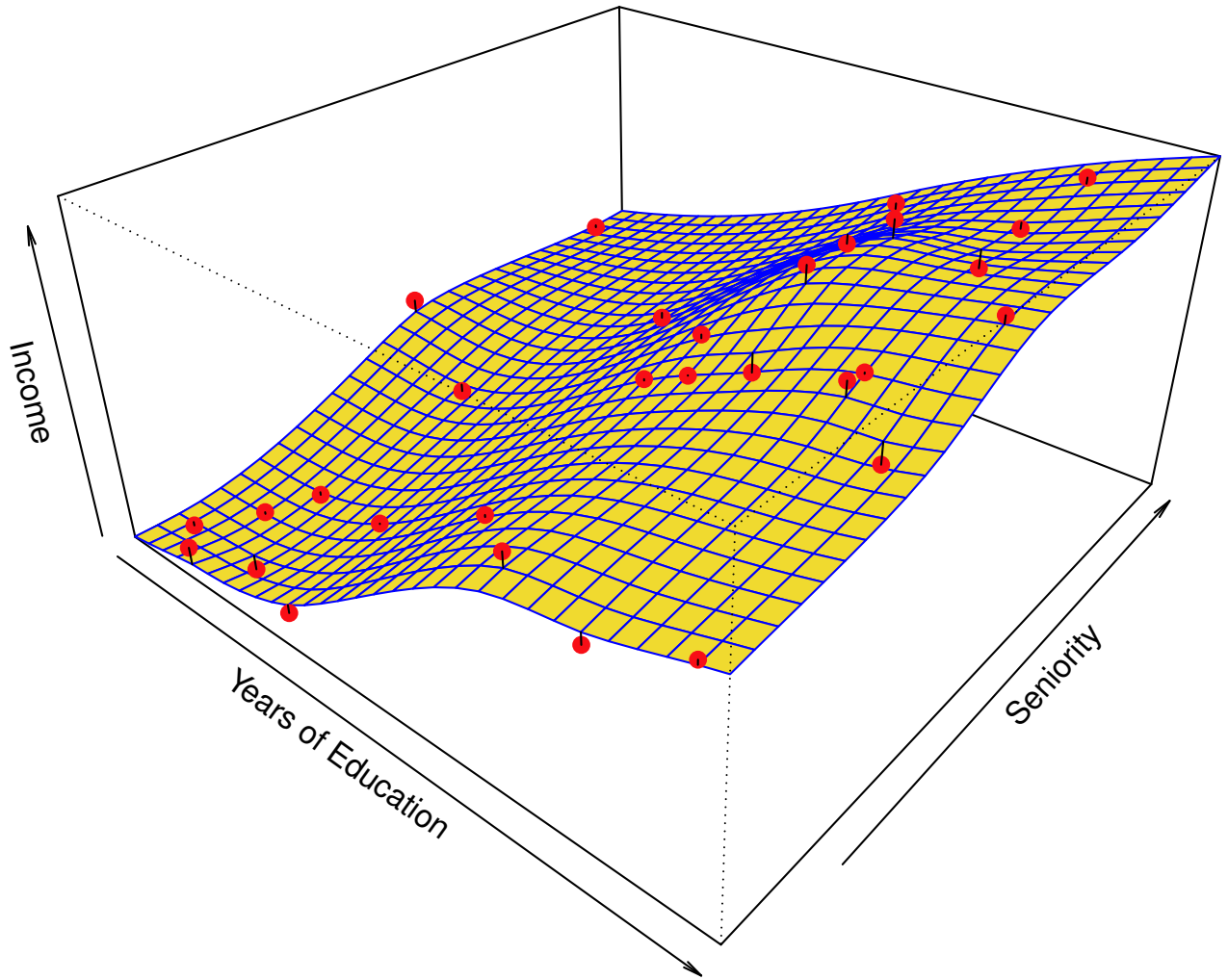
Non-parametric methods

- Non-parametric methods make no a priori assumptions on the functional form of f.
- Instead, they try to achieve as close an approximation to f as possible without being too rough or too oscillatory.

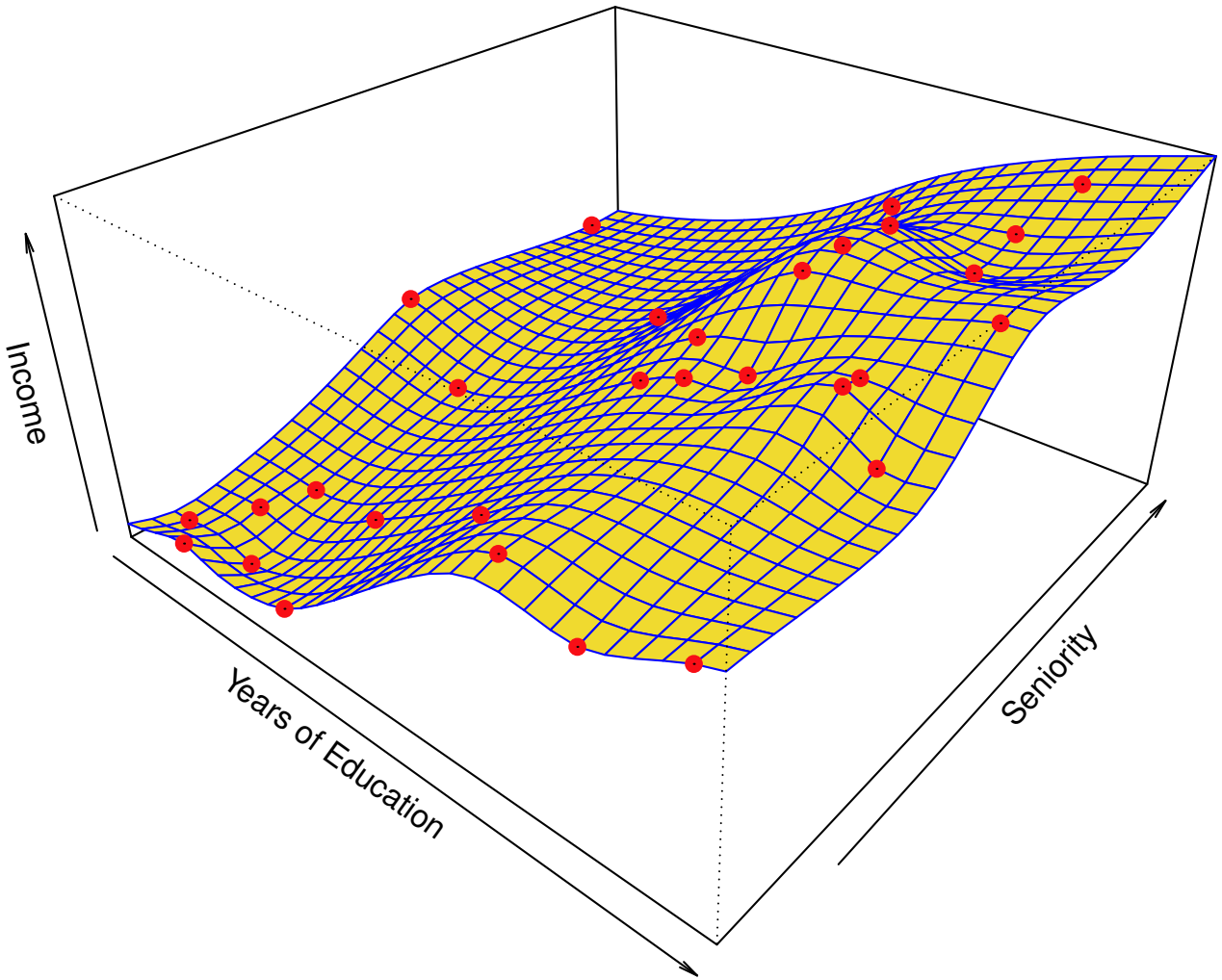
\+ Bad a priori assumption can't limit approximation accuracy.

\- Far more observations necessary than for parametric methods.

Smooth **thin-plate spline** model (later):



Rough **thin-plate spline** model:



Near perfect fit.
Are we overfitting?.

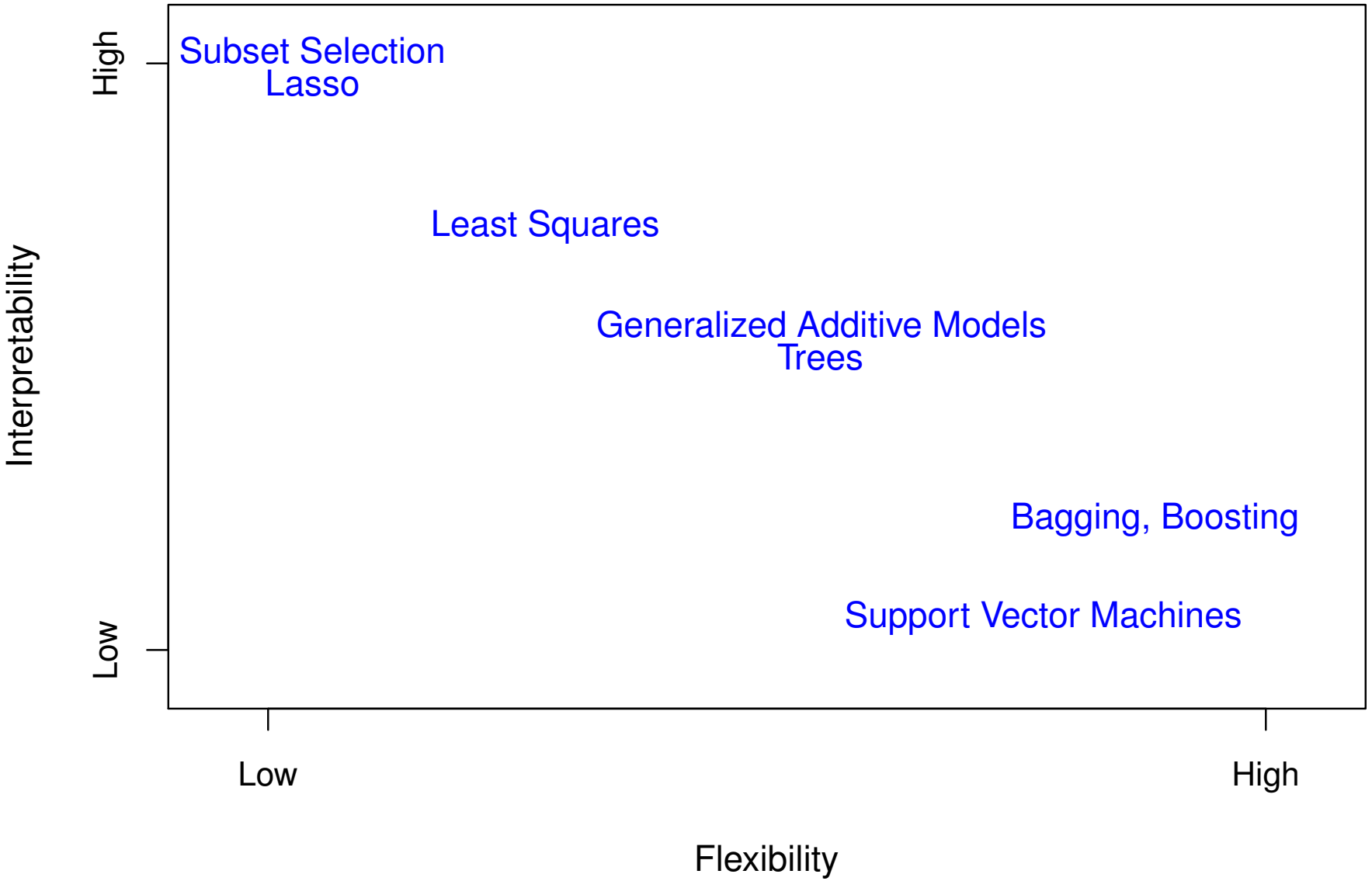
Learning Theory

Tradeoff: prediction accuracy vs. model interpretability

- Less flexible/more restrictive models can only produce a small range of shapes for f. E.g.: linear regression always provides *linear* approximation to f.
- More flexible methods (e.g. thin-plate splines) offer larger variety of function shapes.
- Advantage of restrictive methods:
  - \+ For inference, restrictive models much more interpretable.
  - Linear least-squares easy to interpret.
  - **Lasso**: linear model, different way of selecting coefficients, sets some to zero.
    More restrictive than least squares, but also more interpretable..
  - **Generalized additive models** (GAMs): extend model by certain nonlinear relationships.
    More flexible, less easy to interpret.
  - **Bagging**, **boosting**, **support-vector machines**: fully nonlinear methods, very flexible, very difficult to interpret.

Learning Theory

Tradeoff: prediction accuracy vs. model interpretability



Up to now: observation pairs $(\boldsymbol{x}_i, y_i), i = 1, \ldots, n$.

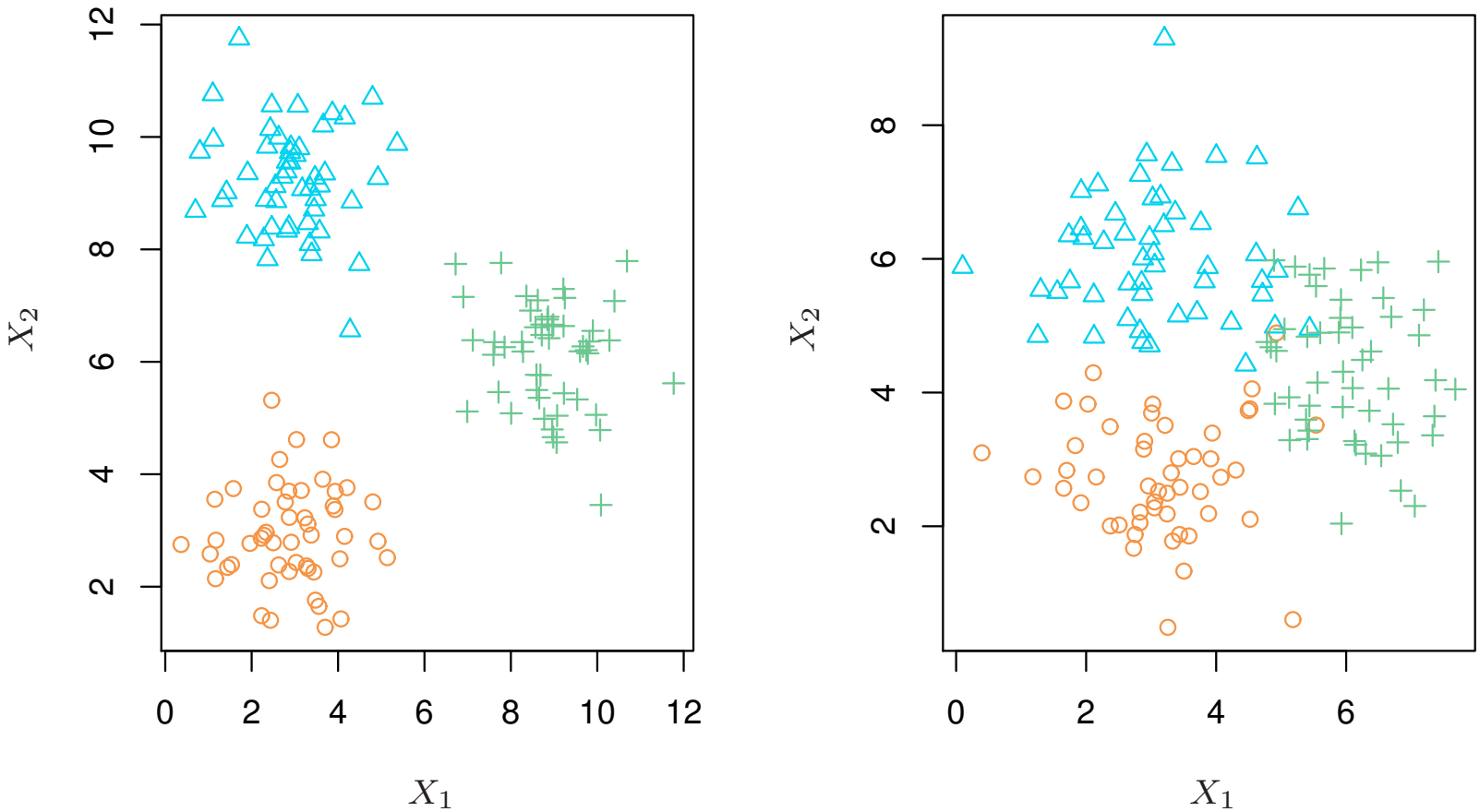
Seek model $\hat{f}$ such that $Y \approx \hat{f}(X)$ for all observations. This is called **supervised learning**.

In **unsupervised learning** only predictor variables X are observed, but no associated responses Y.

- No fitting is possible (nothing to fit to); we are, in a sense, working blind.
- Less ambitious goal: discover relationships between observations, draw conclusions for predictor variables.
- **Cluster analysis** (clustering): statistical learning tool to ascertain whether observations $\{\boldsymbol{x}_i\}_{i=1}^n$ fall into (more or less) distinct groups.
- **Example:** market segment analysis, observe multiple characteristics of potential customers (zip code, family income, shopping habits). Possible groups: big spenders, low spenders.
  In the absence of spending pattern data, clustering may reveal whether potential big spenders may be distinguished by the available data.

Learning Theory

Example: clustering



$n = 150$ observations of two variables X_1 and X_2, each belonging to one of three groups (colored for better distinction). Left: well-separated clusters, easily identified. Right: some overlap between groups, more challenging. Some observations will likely be mis-classified.

Note:

- Clustering more challenging in $p > 2$ dimensions, e.g. there are $p(p-1)/2$ possible scatterplots to look at. Automated methods needed.
- **Semi-supervised learning**: Only $m < n$ observations come with responses. (Responses could be very expensive to obtain compared to the predictor observations).
  Goal: incorporate both types of observations in an optimal way.

Another useful distinction is between **continuous** and **discrete** prediction and response variables.

- Continuous or **quantitative** variables – such as a person's height, age, income, the price of a house or stock – typically take on values in the real numbers.
- Discrete or **qualitative** variables – a person's gender, whether or not an event occurs, a cancer diagnosis – take on values in a in one of a finite number of different **classes** or categories.
- Problems with a quantitative response variable are typically referred to as **regression problems**, those with a qualitative response as **classification problems**.
- The distinction is not always sharp, e.g. logistic regression is used for (two-valued) qualitative responses (it estimates class probabilities).

Contents

Assessing Model Accuracy

Mean squared error

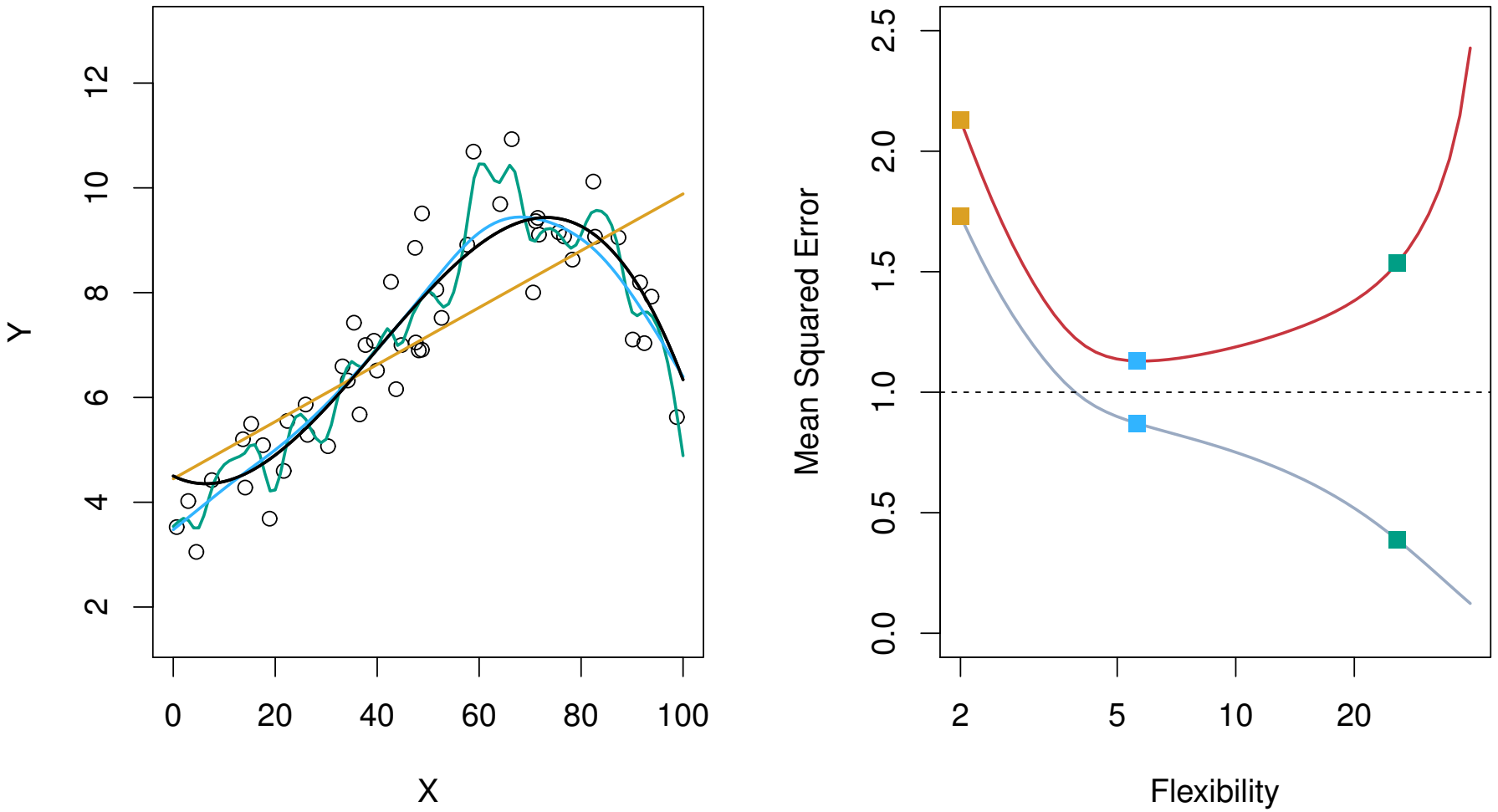
Most common error metric in regression: **mean squared error (MSE)**:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2. \tag{2.3}$$

- When applied to training data: **training MSE**.
- More interesting (particularly for prediction): **test MSE** resulting from data not used to train (fit) the model $\hat{f}$.
- If a **test data** set is available in addition to the training data, different learning (fitting) methods can be compared with respect to their test MSE values.
- In the absence of a test data set, choosing a learning method based solely on the training MSE can be deceptive.

Assessing Model Accuracy

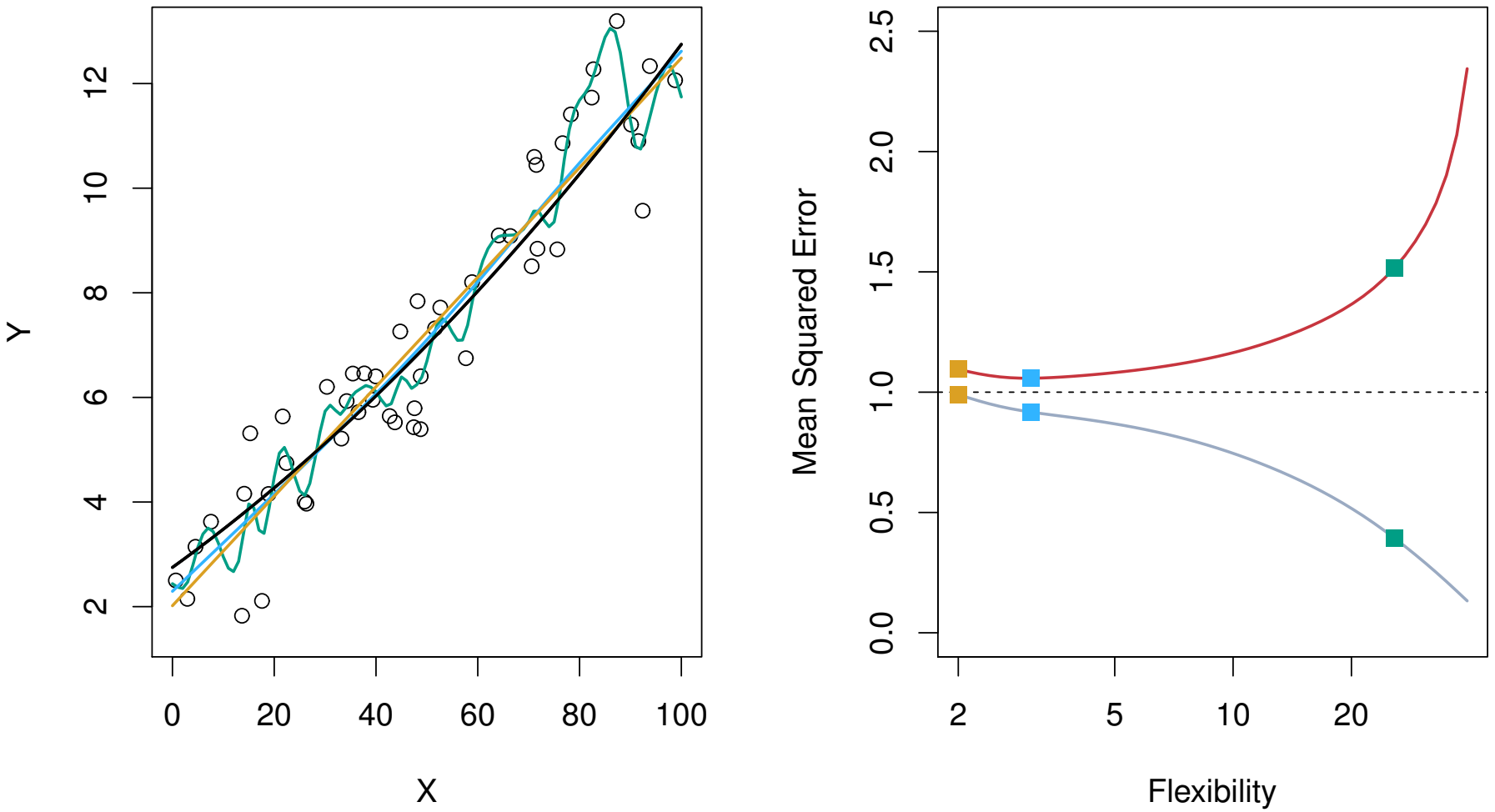
Example: smoothing spline models



Left: Observations from model (2.1), true f in black, estimates in orange, blue, green.
Right: Average MSE for training data (gray), test data (red) vs. flexibiliy parameter.
Horizontal dashed line denotes $\mathbf{Var}\,\varepsilon$.

Assessing Model Accuracy

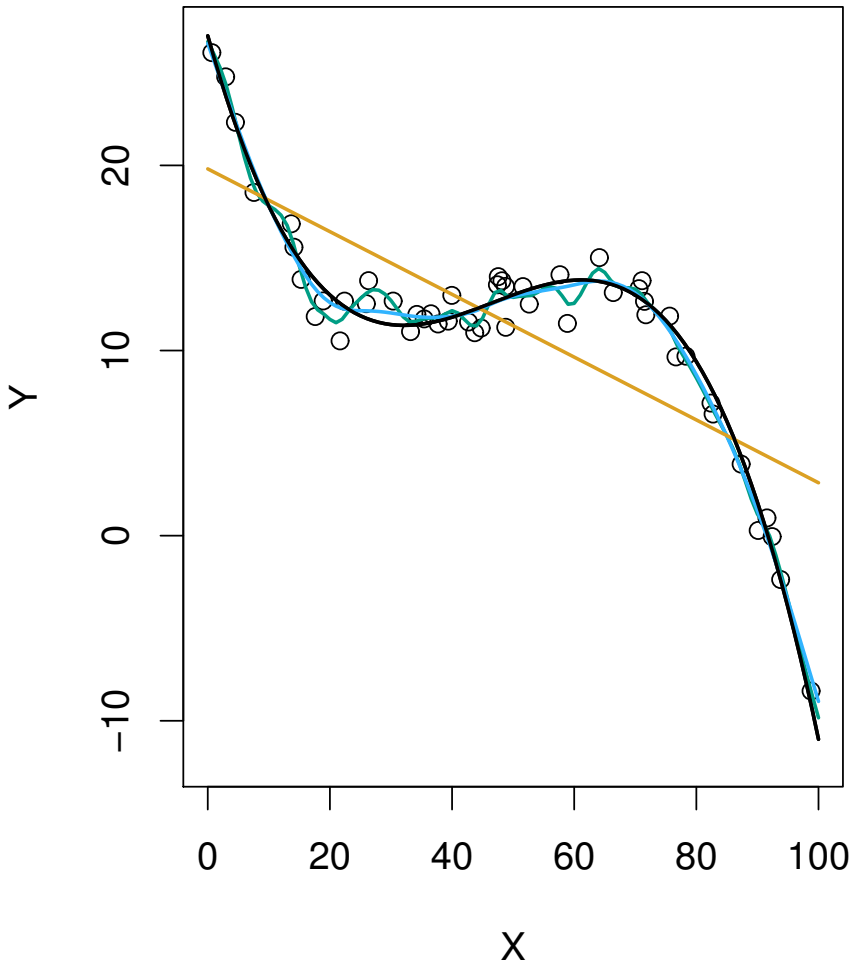
Example: smoothing spline models

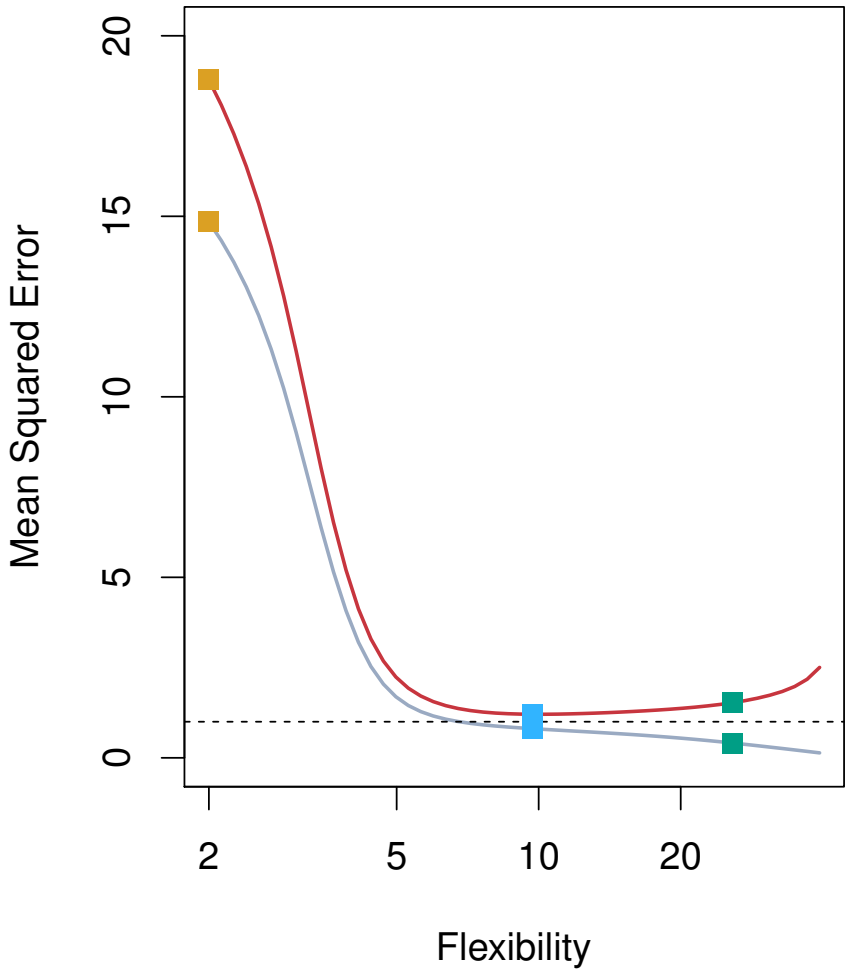


Same plots as on previous figure, but with a true model that is nearly linear. Initial estimate (with few degrees of freedon) already quite accurate.

Assessing Model Accuracy

Example: smoothing spline models





Another such plot, now the true model is highly nonlinear. Maximal accuracy for training and test data not attained until many degrees of freedom employed.

Assessing Model Accuracy

Example: smoothing spline models

Recap:

- Monotone decrease of training MSE as model becomes more flexible (more degrees of freedeom, DoF) and can more flexibly follow data variation.
- Typically test MSE curve U-shaped, rises again once overfitting sets in.
- This is a fundamental property of statistical learning, regardless of data set and regardless of statistical technique being used.
- Interpretation: in overfitting, estimate is finding patterns (signal variation) where there are none.
  James et al: "When we overfit the training data, the test MSE will be very large because the supposed patterns that the method found in the training data simply don't exist in the test data."
- Overfitting: less flexible model would have yielded smaller test MSE.

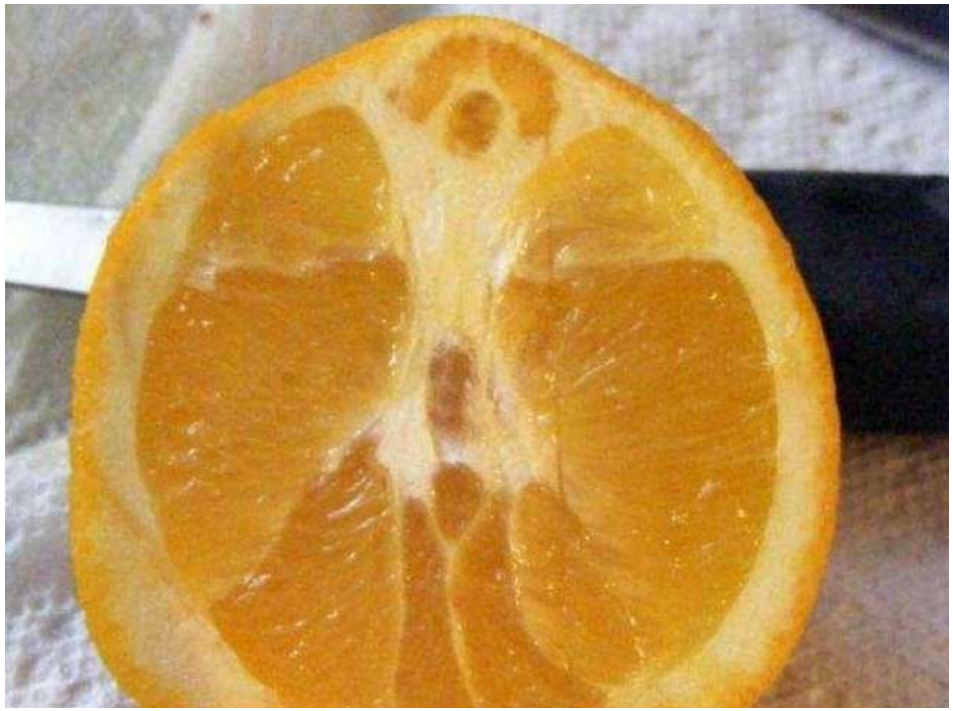
Note: Estimation methods based on minimizing MSE with respect to the DoF in the method, hence training MSE almost always less than test MSE.

Assessing Model Accuracy

Apophenia

The tendency to misclassify random events as systematic or, more generally, to see patterns where there are none, is common to human experience and known as **Apophenia**.

- It is believed to be an advantage in the process of natural selection.
- It encourages conspiracy theories.
- It is used to explain the **gambler's fallacy** in probability theory.
- In his bestselling book Thinking Fast and Low, the famous behavioral economist Kahneman calls this phenomenon the "law of small numbers".



Jesus and Mary in an orange.

Source: anorak.co.uk

Assessing Model Accuracy

Trade-off: bias vs variance

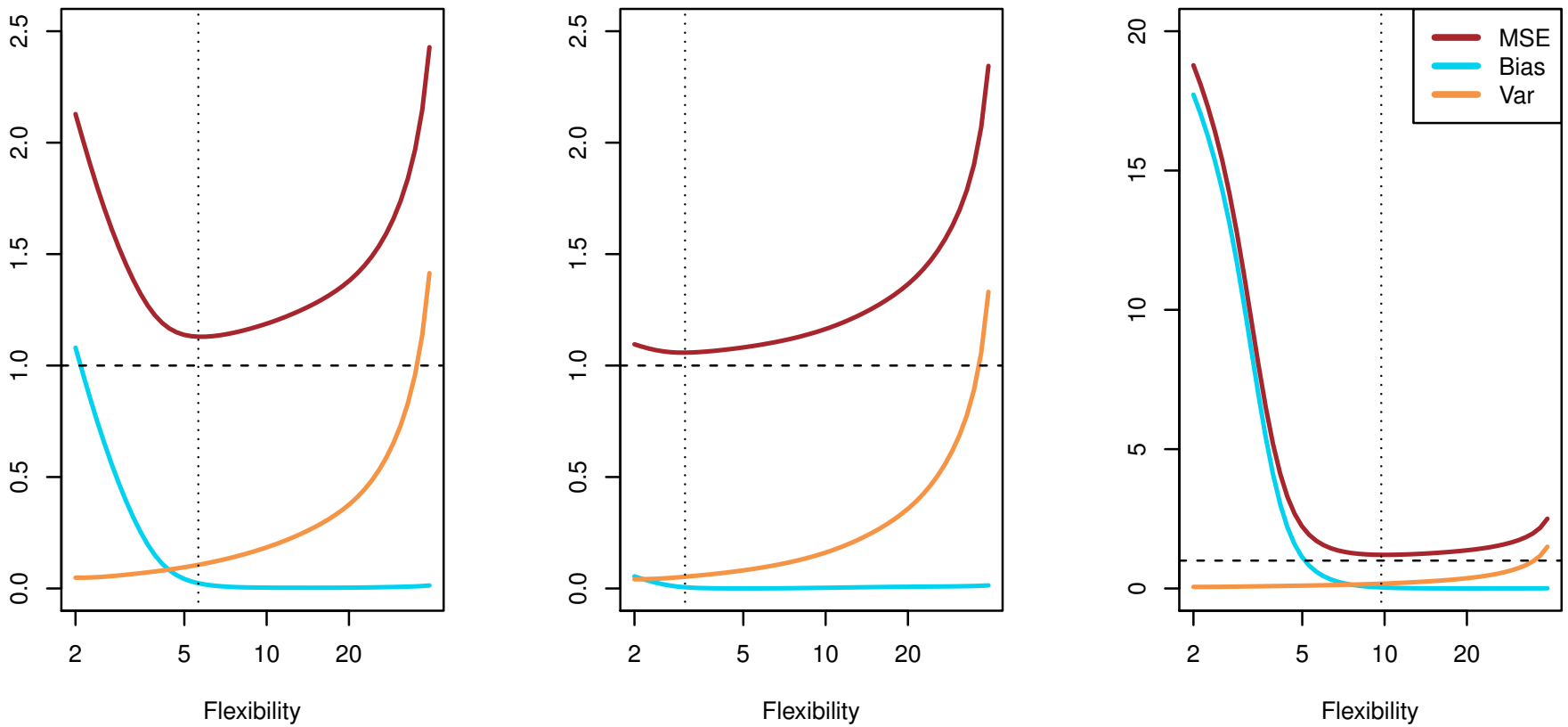
Can show: expected test MSE for new value x_0 of test data has representation

$$\mathbf{E}\left[\left(y_0 - \hat{f}(x_0)\right)^2\right] = \mathbf{Var}\, f(x_0) + [\text{Bias}\, \hat{f}(x_0)]^2 + \mathbf{Var}\, \varepsilon. \qquad (2.4)$$

- $\mathbf{E}\left[\left(y_0 - \hat{f}(x_0)\right)^2\right]$ is the expected test MSE with respect to the distribution of the predictor variable X, i.e., the average test MSE we would obtain by repeatedly estimating f using a large number of training sets, and testing each at x_0.
- (2.4) implies that a good statistical learning method needs to achieve both low bias and low variance.
- **Variance**: amount by which $\hat{f}$ would change if estimated using a different training data set. A method with high variance is sensitive to small changes in the data set.
- **Bias**: error introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model. Flexible methods have lower bias.

Assessing Model Accuracy

Trade-off: bias vs variance



Bias-variance decomposition for last 3 examples. Horizontal dashed line: $\mathbf{Var}\,\varepsilon$. Flexibility level of minimal test MSE varies due to different rates of change in bias and variance.

Assessing Model Accuracy

Bias vs. variance for classification

For qualitative (discrete) response variable Y, replace MSE with training **error rate**:

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{y_i \neq \hat{y}_i\}} \tag{2.5}$$

expressing the fraction of incorrect classifications, where

$$\hat{y}_i \text{ : predicted class label for } i\text{-th observation using } \hat{f},$$

$$\mathbb{1}_{\{y_i \neq \hat{y}_i\}} = \begin{cases} 1 & y_i \neq \hat{y}_i, \\ 0 & y_i = \hat{y}_i, \end{cases} \qquad \text{(indicator variable).}$$

As in regression setting, of more interest than training error rate (2.5) is **test error rate**, which averages classification errors $\mathbb{1}_{\{y_0 \neq \hat{y}_0\}}$ over a test set of observations (x_0, y_0) with classification prediction $\hat{y}_0$ for predictor variable x_0.

Assessing Model Accuracy

Classification: Bayes classifier

One can show (we won't) that expectation of test rate error is minimized by the **Bayes classifier**: assign to test observation with predictor vector x_0 the class j for which conditional probability

$$\mathbf{P}\left(Y = j | X = x_0\right)$$

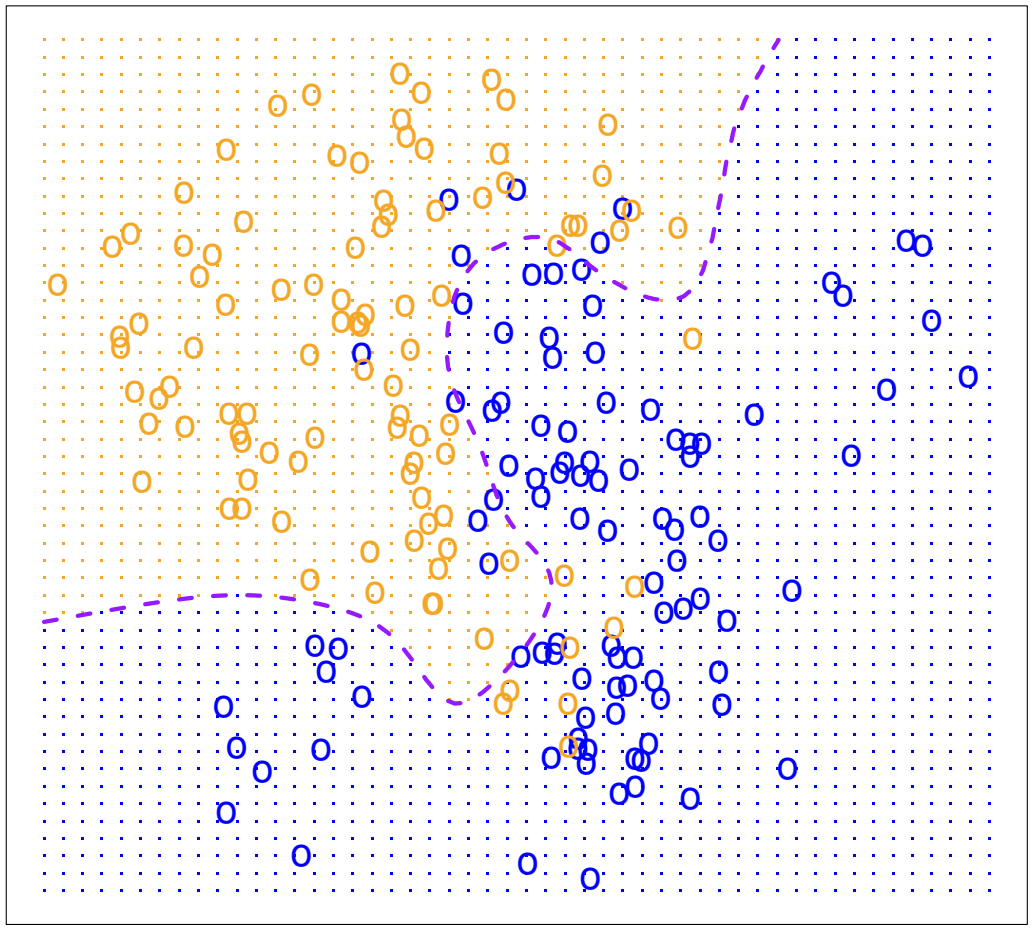
is maximized over all j.

Special case: two-class problem, i.e., $Y \in \{1, 2\}$; predict

$$\hat{y}_0 = \begin{cases} 1 & \text{if } \mathbf{P}\{Y = 1 | X = x_0\} > 0.5 \\ 2 & \text{otherwise.} \end{cases}$$

Assessing Model Accuracy

Example: Bayes classifier, 2 classes



Predictors: X_1, X_2
Response: $Y \in \{\text{orange, blue}\}$
Observations: circles

Orange shading:

$$\mathbf{P}\{Y = \text{orange}|X\} > 0.5$$

Blue shading

$$\mathbf{P}\{Y = \text{orange}|X\} < 0.5$$

(simulated data)

Dashed line: **Bayes decision boundary**

$$\mathbf{P}\{Y = \text{orange}|X\} = 0.5$$

Assessing Model Accuracy

Bayes error rate

- Bayes classifier produces lowest possible test error rate, the **Bayes error rate**.
- By definition, error rate at $X = x_0$ is

$$1 - \max_j \mathbf{P}\{Y = j|X\}.$$

- Overall Bayes error rate:

$$1 - \mathbf{E}\left[1 - \max_j \mathbf{P}\{Y = j|X\}\right],$$

  expectation with respect to distribution of X.
- Previous example: Bayes error rate is 0.1304, positive since some observations on wrong side of decision boundary, hence $\max_j \mathbf{P}\{Y = j|X = x_0\} < 1$ for some x_0.
- Bayes error rate analogous to irreducible error.

Assessing Model Accuracy

K-nearest neighbors

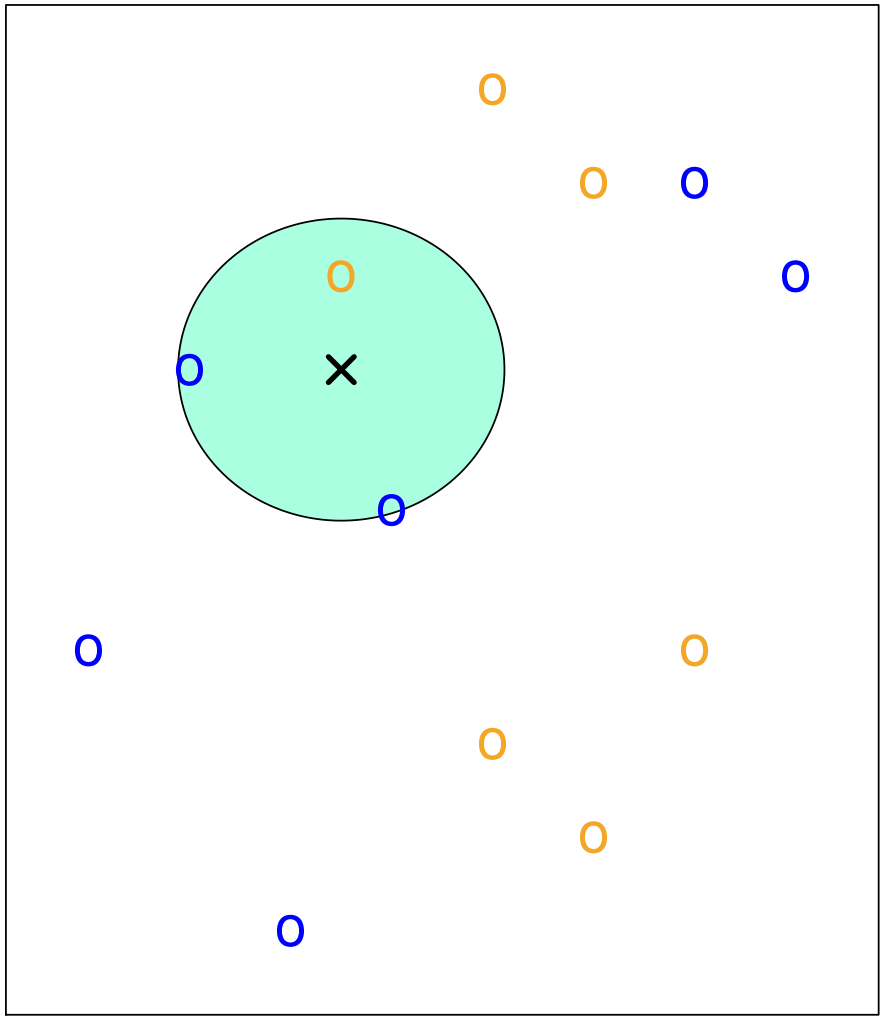
- Bayes classifier not realizable, since based on unknown conditional distribution, represents unattainable reference value.
- **K-nearest neighbors (KNN)** classifier: classify based on *estimate* of conditional distribution.
- Given $X = x_0$, denote by $\mathcal{N}_0$ the $K \in \mathbb{N}$ training set points closest to x_0 and estimate

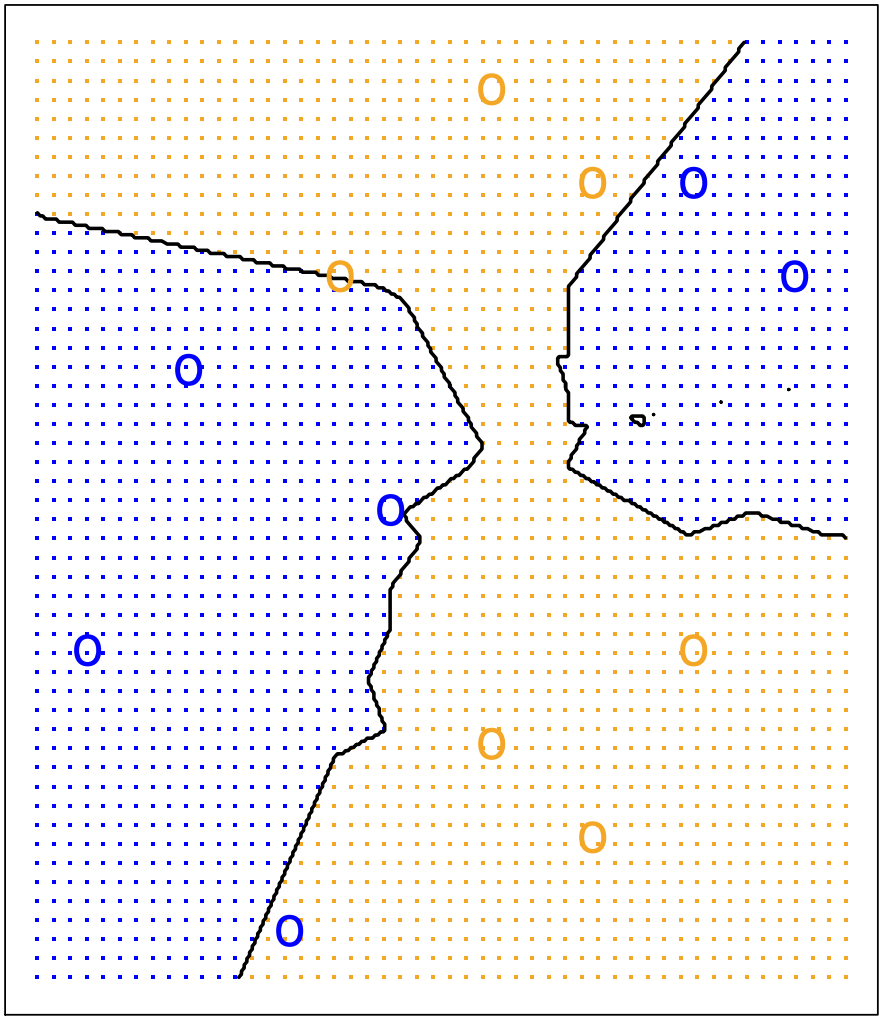
$$\mathbf{P}\{Y = j | X = x_0\} \approx \frac{1}{K} \sum_{i \in \mathcal{N}_0} \mathbb{1}_{\{y_i = j\}},$$

  i.e., by fraction of K nearest neighbors belonging to class j.
- Now proceed as in Bayes estimate with this approximate conditional distribution.

Assessing Model Accuracy

Example: KNN, K=3



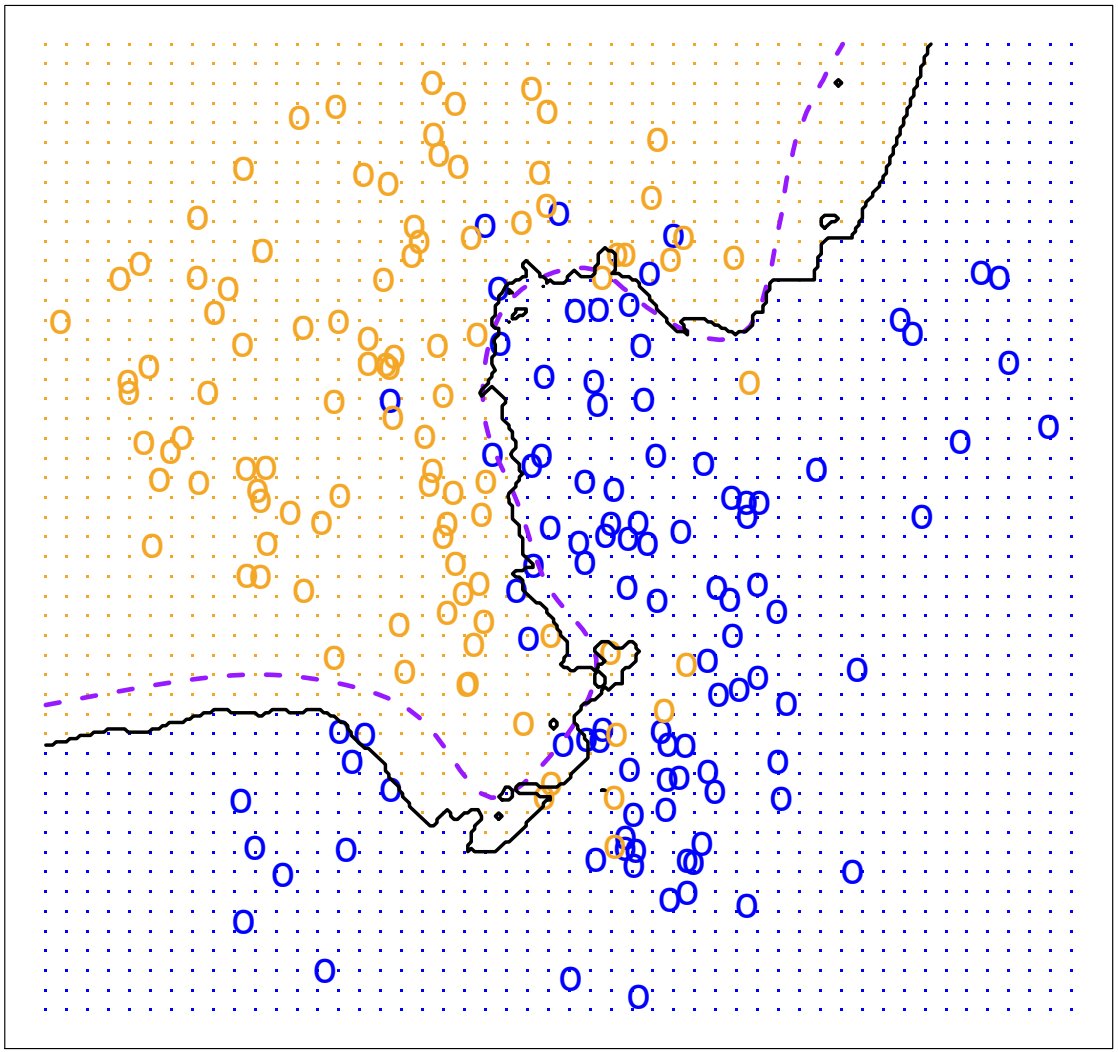


Left: $\times$: x_0; green circle: $\mathcal{N}_0$

Right: KNN applied to all shaded points, resulting decision boundary

Assessing Model Accuracy

Example: KNN, K=10



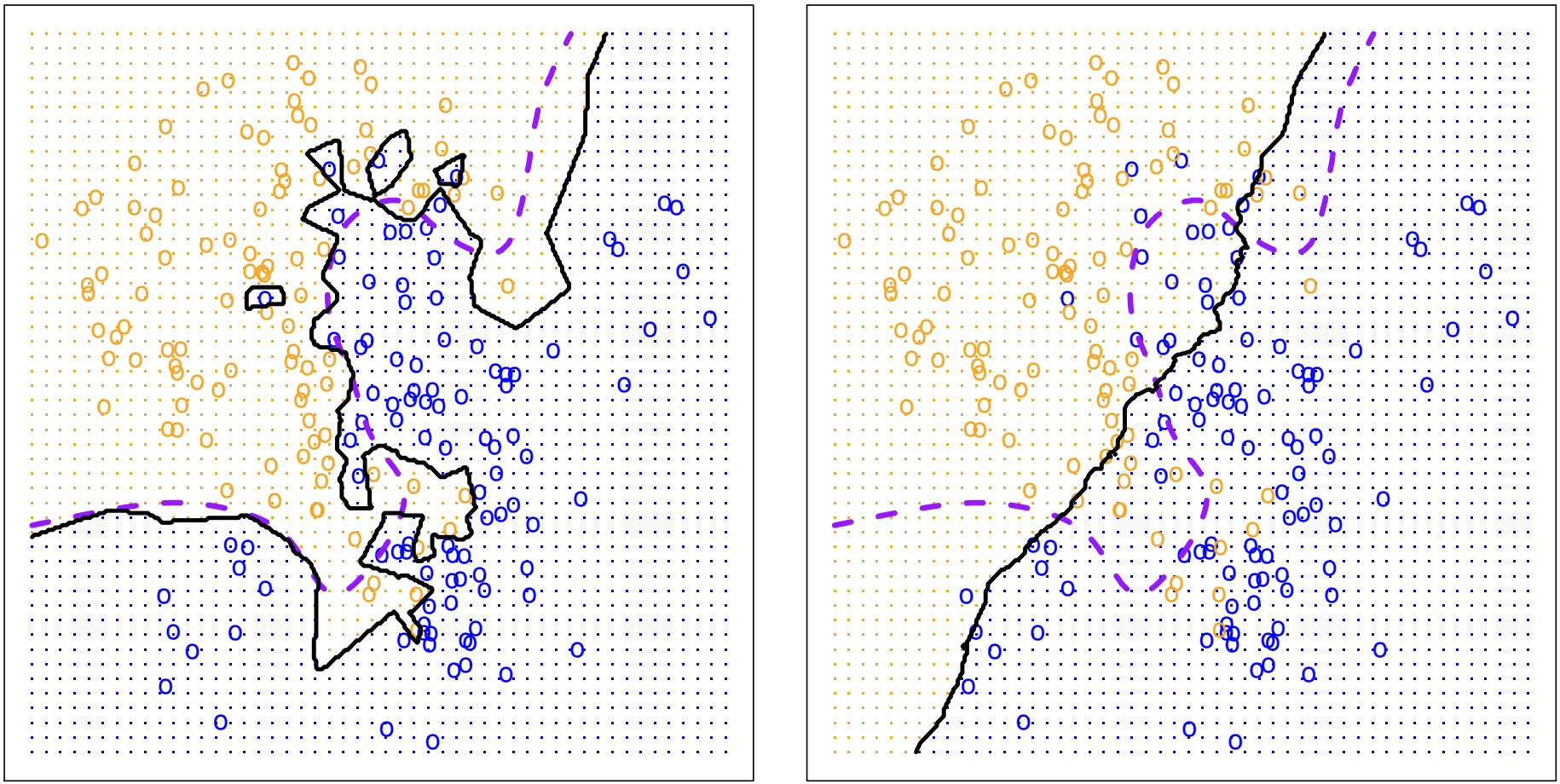
KNN decision boundary for $K = 10$ applied to data set from Slide 62.

Test error rates:
Bayes: 0.1304
KNN: 0.1363.

Assessing Model Accuracy

Example: KNN, K=1,100

KNN applied to data set from Slide 62:

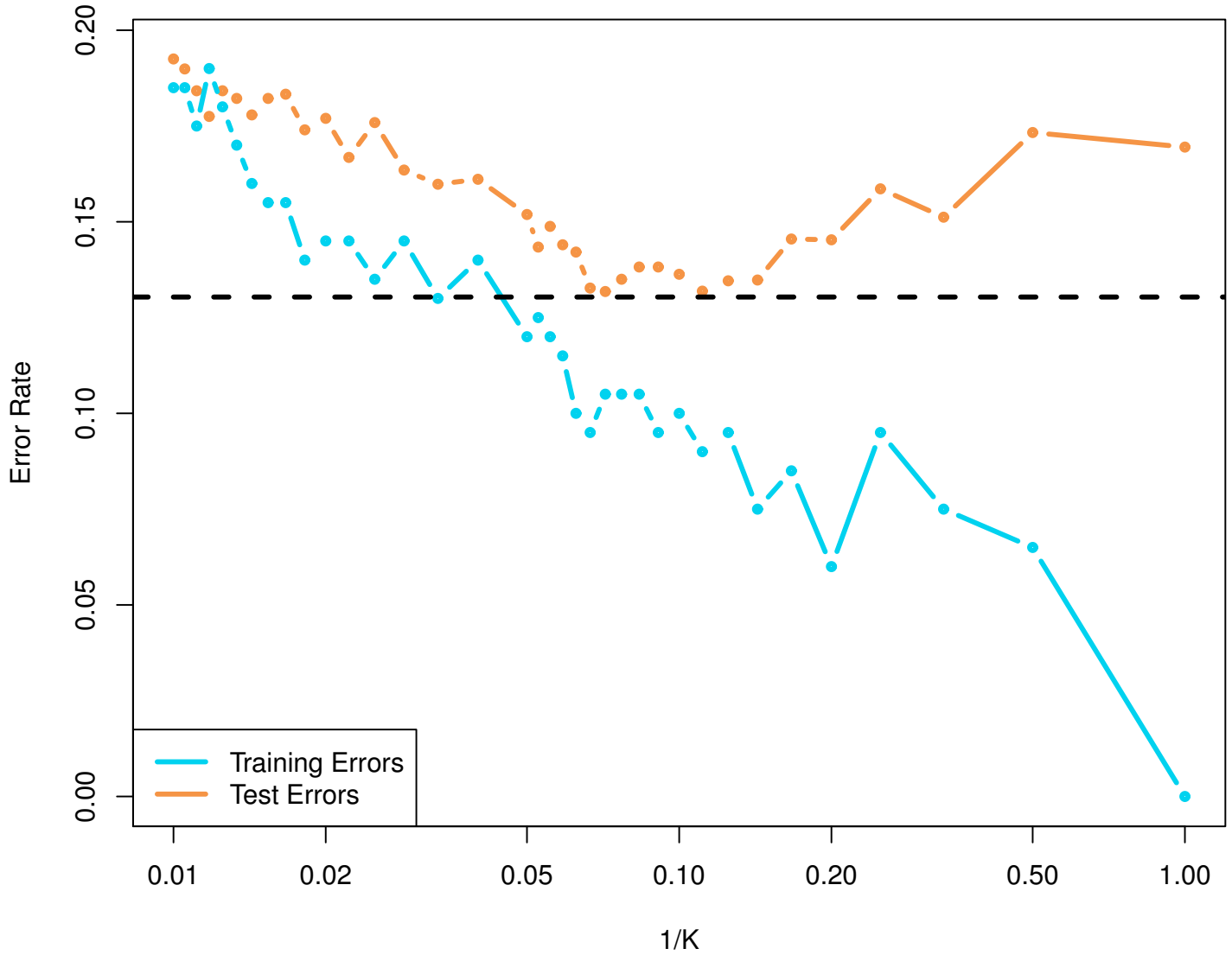


Left: $K = 1$, high variance; test error rate 0.1695.

Right: $K = 100$, high bias; test error rate 01925.

Assessing Model Accuracy

Example: KNN error rates against K



Training and test errors of KNN classification for same data plotted against $1/K$.