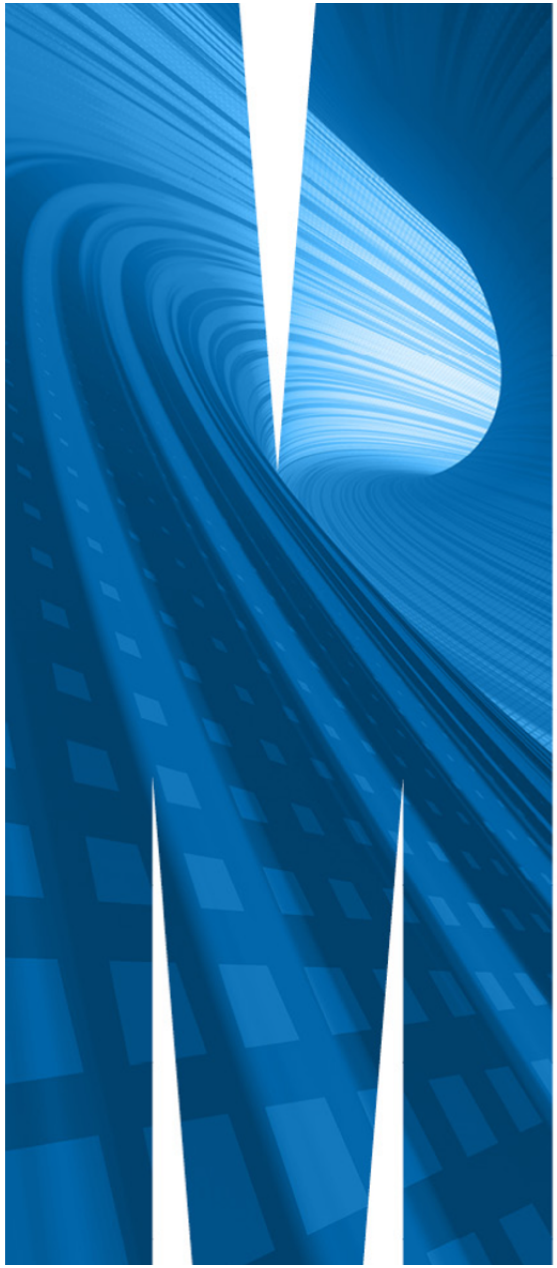
MONASH University

# ETC3250/5250: Introduction to Machine Learning

## Conceptual framework



Lecturer: *Professor Di Cook*

Department of Econometrics and Business Statistics

ETC3250.Clayton-x@monash.edu

Week 1b

# Two purposes

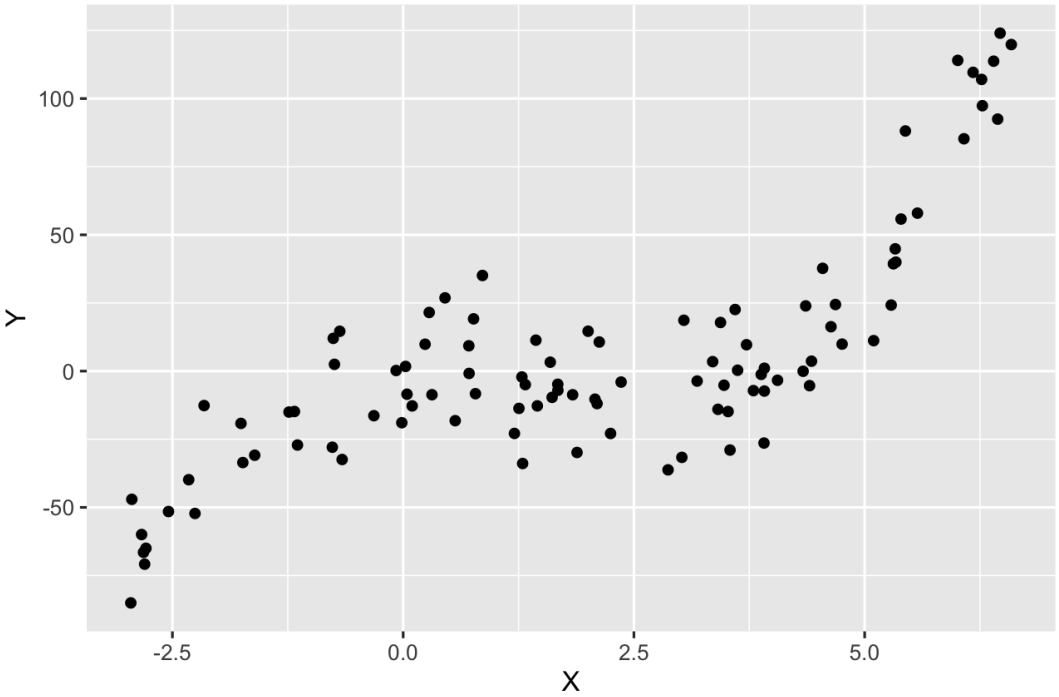
## Prediction and inference (or understanding)

# Supervised learning

We assume that there is a relationship between $Y$ and $\mathbf{X}$ that can be written as

$$Y = f(\mathbf{X}) + \varepsilon$$

where function $f(\mathbf{X})$ is fixed but unknown, and $\varepsilon$ is independent of $\mathbf{X}$ and has mean 0.




*Note: Here, I've simulated data so I know what $f$ is.*

# Why estimate $f$?

In many situations, $\mathbf{X}$ may be readily available, but $Y$ ... might be hard to collect. SO, we would like to be able to use $\mathbf{X}$ to predict new values of $Y$.
We might not so much be concerned about whether $f$ is easy to understand, just that we are confident that its going to do a good job of predicting new values.

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

where ^ reflects what we estimate, the unknown function to be, and the estimated response value.

# Why estimate $f$?

In many situations, $\mathbf{X}$ may be readily available, but $Y$ might be hard to collect. SO, we would like to be able to use $\mathbf{X}$ to predict new values of $Y$.
We might not so much be concerned about whether $f$ is easy to understand, just that we are confident that its going to do a good job of predicting new values.

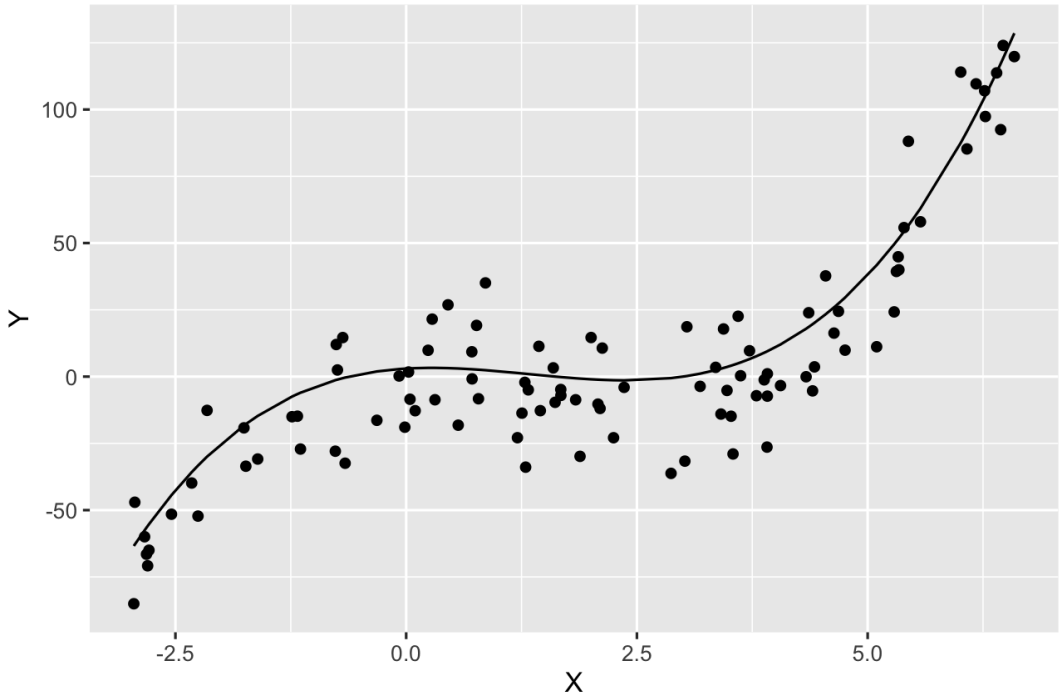
$$\hat{Y} = \hat{f}(\mathbf{X})$$

where ^ reflects what we estimate, the unknown function to be, and the estimated response value.

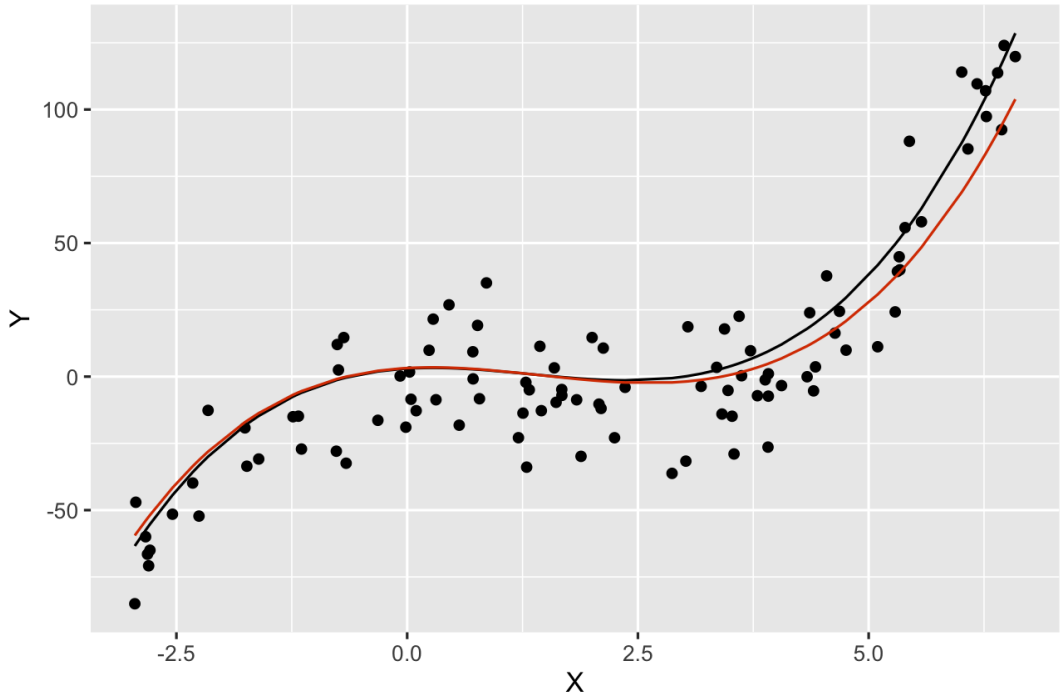
The accuracy of $\hat{Y}$ as a prediction for $Y$ depends on what we will call reducible AND irreducible error. We can write this as

$$\begin{aligned} E(Y - \hat{Y})^2 &= E(f(\mathbf{X}) + \varepsilon - \hat{f}(\mathbf{X}))^2 \\ &= \underbrace{E(f(\mathbf{X}) - \hat{f}(\mathbf{X}))^2}_{reducible} + \underbrace{\mathrm{Var}(\varepsilon)}_{irreducible} \end{aligned}$$

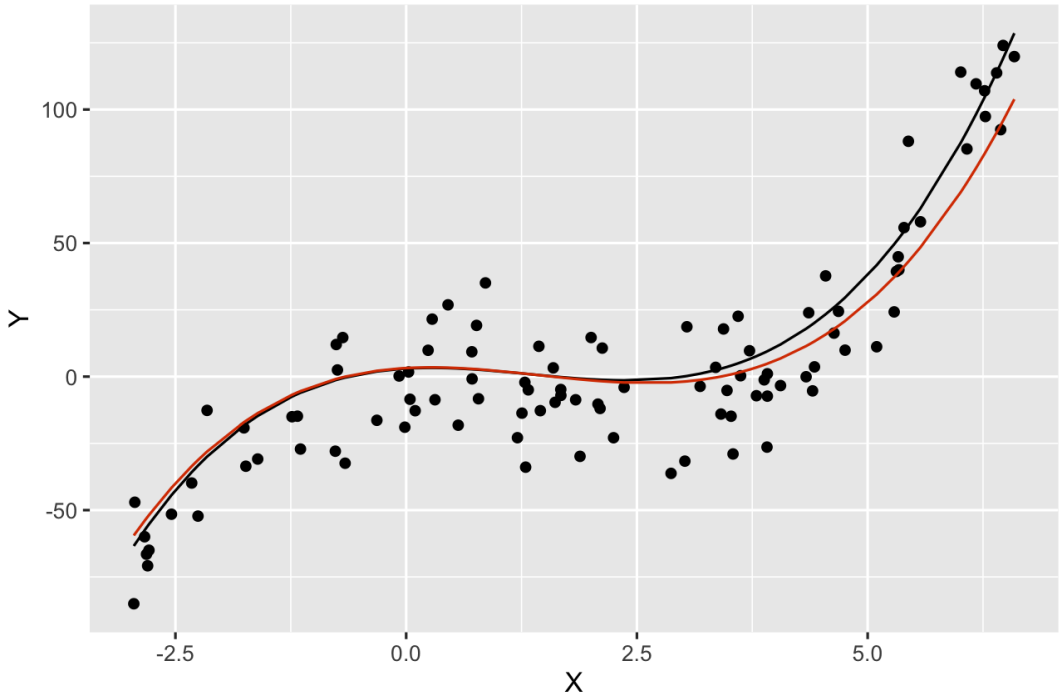
where $E(Y - \hat{Y})^2$ represents the average or expected value of the squared difference between the observed and predicted response, and $\mathrm{Var}(\varepsilon)$ represents the variance of the error.




Line indicates **true** $f$. This is the best possible because all that is left is irreducible error.

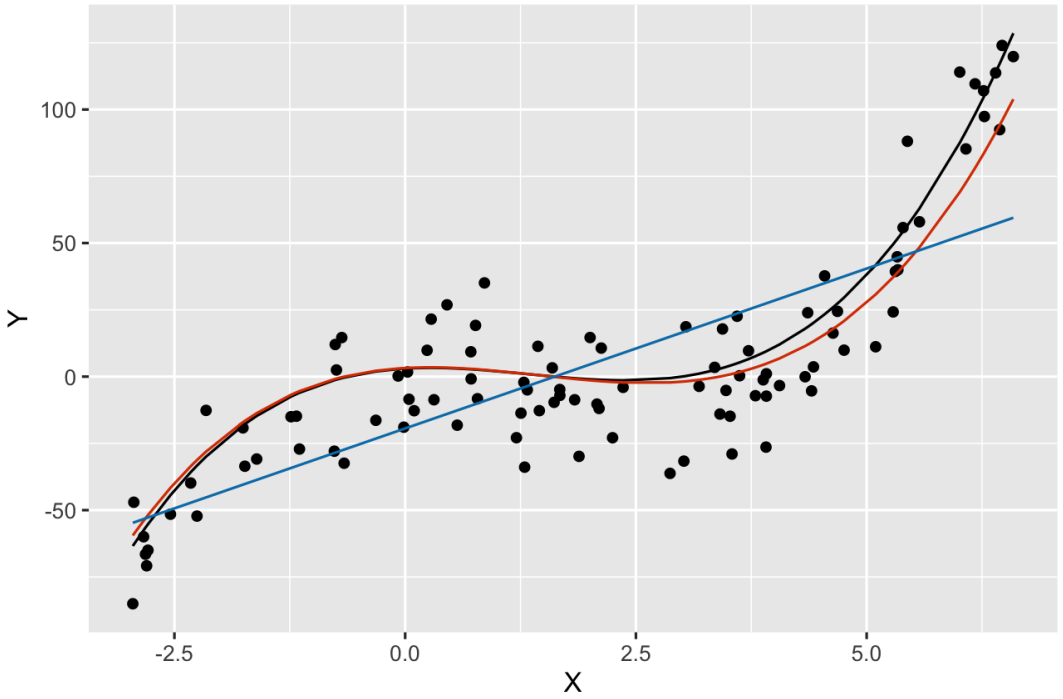



Orange line indicates an estimated model ($\hat{f}$). This could be improved.




Orange line indicates an estimated model ($\hat{f}$). This could be improved.

**Suppose we used a much simpler model, a linear model.**




Blue line indicates a simpler estimated model ($\hat{f}$). There is a lot of room to improve this.

# Remember

reducible is what we can improve on by producing the best model.

irreducible there is some random fluctuation from one sample to the next which is not systematic.

The goal is that the predictions from the model are accurate for future samples.

# Inference (understanding)

# Inference

We would like to understand the way that $Y$ is related to $\mathbf{X}$.

- 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?

**It is important here not to treat $f$ as a black box.**

Ideally, good prediction also allows for good inference and understanding.

# How do we estimate $f$?

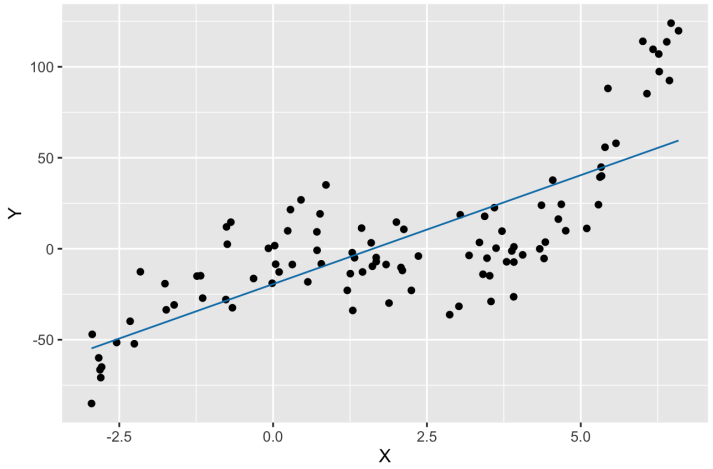
- Parametric methods:
  - Assume that the model takes a specific form
  - Fitting then is a matter of estimating the parameters of the model
  - Generally considered to be less flexible
  - If assumptions are wrong, performance likely to be poor
- Non-parametric methods:
  - No specific assumptions
  - Allow the data to specify the model form, without being too rough or wiggly
  - More flexible
  - Generally needs more observations

# Parametric models

**When you force your data to fit the constraints of your model**

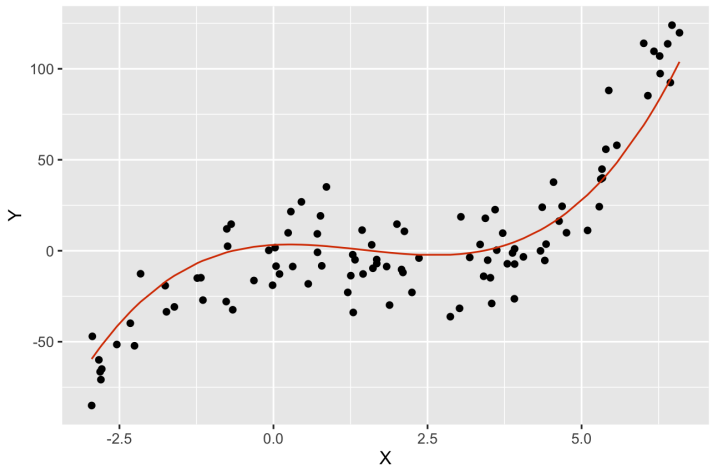
A linear regression model:

$$f(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$



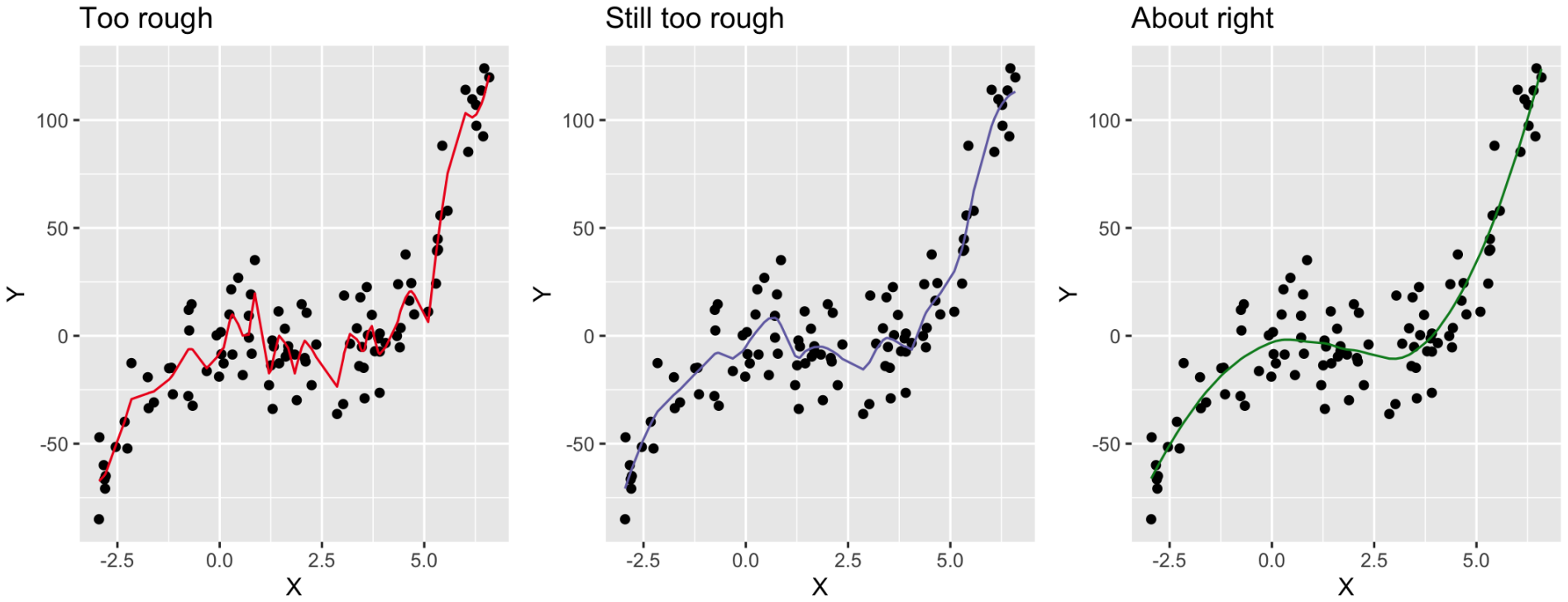

And nonlinear regression model:

$$f(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \ldots$$



# Non-parametric models

Example: Local polynomial regression, called loess. Fit a linear model to many small subsets of the data.

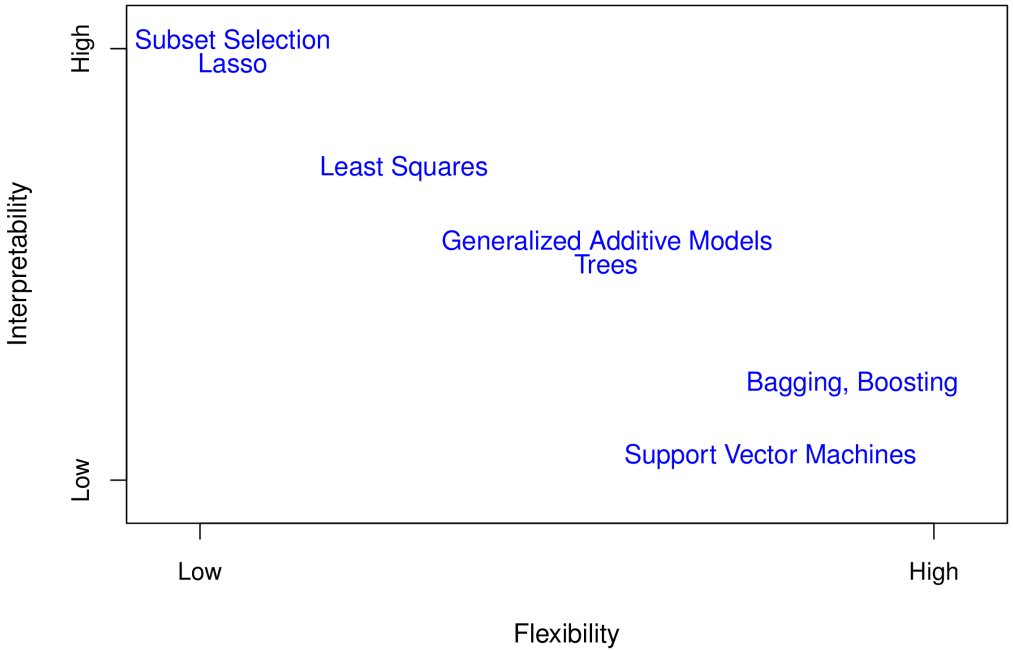



A more general approach is called $k$-nearest neighbours, $\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in n_k(x)} y_i$.

# Trade-off between predictive accuracy and model interpretability

A summary of common models and how they tend to lie in terms of predictive accuracy vs interpretability.

Chapter2/2.7.pdf



# Assessing model accuracy

# Assessing model accuracy

Training data: the set of observations used to train or teach the method to estimate $f$.

Suppose we have a regression model $y = f(\mathbf{x}) + \varepsilon$. Estimate $\hat{f}$ from some training data, $Tr = \{(y_i, \mathbf{x}_i)\}_{i=1}^n$.

The most common measure of accuracy is the training Mean Squared Error (MSE)

$$MSE_{Tr} = \underset{i \in Tr}{\text{Ave}}[y_i - \hat{f}(\mathbf{x}_i)]^2 = \frac{1}{n}\sum_{i=1}^{n}[(y_i - \hat{f}(\mathbf{x}_i)]^2$$

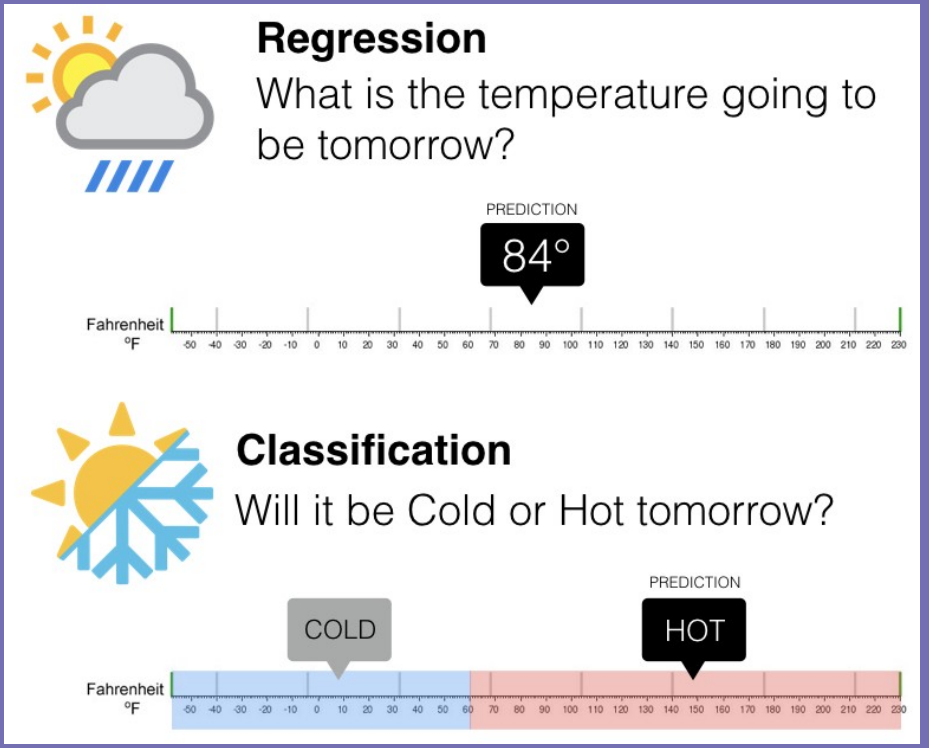
# Assessing model accuracy

Test data: the set of observations reserved to compute accuracy of the model for new data.

A better measure of accuracy is obtained by using the test data, denoted as $Te = \{(y_i, \mathbf{x}_i)\}_{i=1}^{m}$, Test Mean Squared Error

$$MSE_{Te} = \underset{j \in Te}{\text{Ave}}[y_j - \hat{f}(\mathbf{x}_j)]^2 = \frac{1}{m}\sum_{j=1}^{m}[(y_j - \hat{f}(\mathbf{x}_j)]^2$$

# Regression vs classification




Source: Taylor Fogarty

# Assessing model accuracy for classification

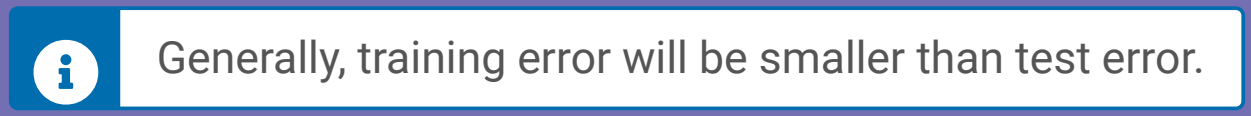
To indicate the categorical response, we will use $\hat{C}$ instead of $\hat{f}$.

Compute $\hat{C}$ from some training data, $Tr = \{(y_i, \mathbf{x}_i)\}_{i=1}^{n}$. In place of MSE, we now use the error rate (fraction of misclassifications) to get the Training Error Rate

$$\text{Error rate}_{Tr} = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{C}(\mathbf{x}_i))$$

And again a better estimate of future accuracy is obtained using test data $Te = \{(y_i, \mathbf{x}_i)\}_{i=1}^{m}$ to get the Test Error Rate

$$\text{Error rate}_{Te} = \frac{1}{m}\sum_{j=1}^{m} I(y_j \neq \hat{C}(\mathbf{x}_j))$$




Because the training data is used to fit the model, by design the error will be small relative to the error when the model is used on new data.

# Bias-variance trade-off

There are two competing forces that govern the choice of learning method: bias and variance.

Bias is the error that is introduced by modeling a complicated problem by a simpler problem.

- For example, linear regression assumes a linear relationship and perhaps the relationship is not exactly linear.
- In general, but not always, the more flexible a method is, the less bias it will have.

This site has a lovely explanation.

# Bias-variance trade-off

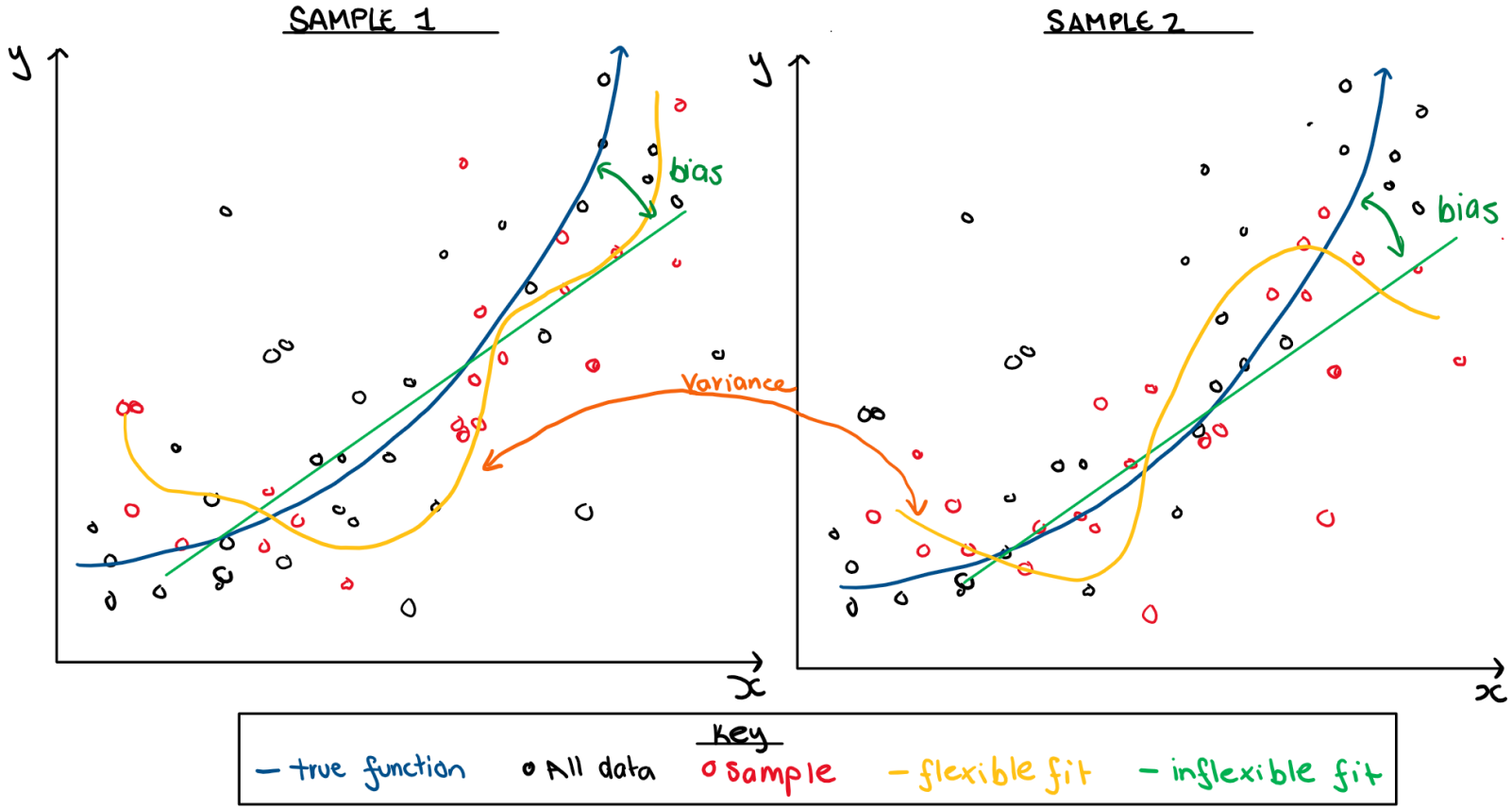
There are two competing forces that govern the choice of learning method: bias and variance.

Variance refers to how much your estimate would change if you had different training data. Its measuring how much your model depends on the data you have, to the neglect of future data.

- In general, the more flexible a method is, the more variance it has.
- The size of the training data has an impact on the variance.

# Flexibility, bias and variance

This blog post by Harriet Mason, former ETC3250 student has a lovely explanation of the trade-off in flexibility and effect on bias and variance.



# MSE decomposition into bias and variance

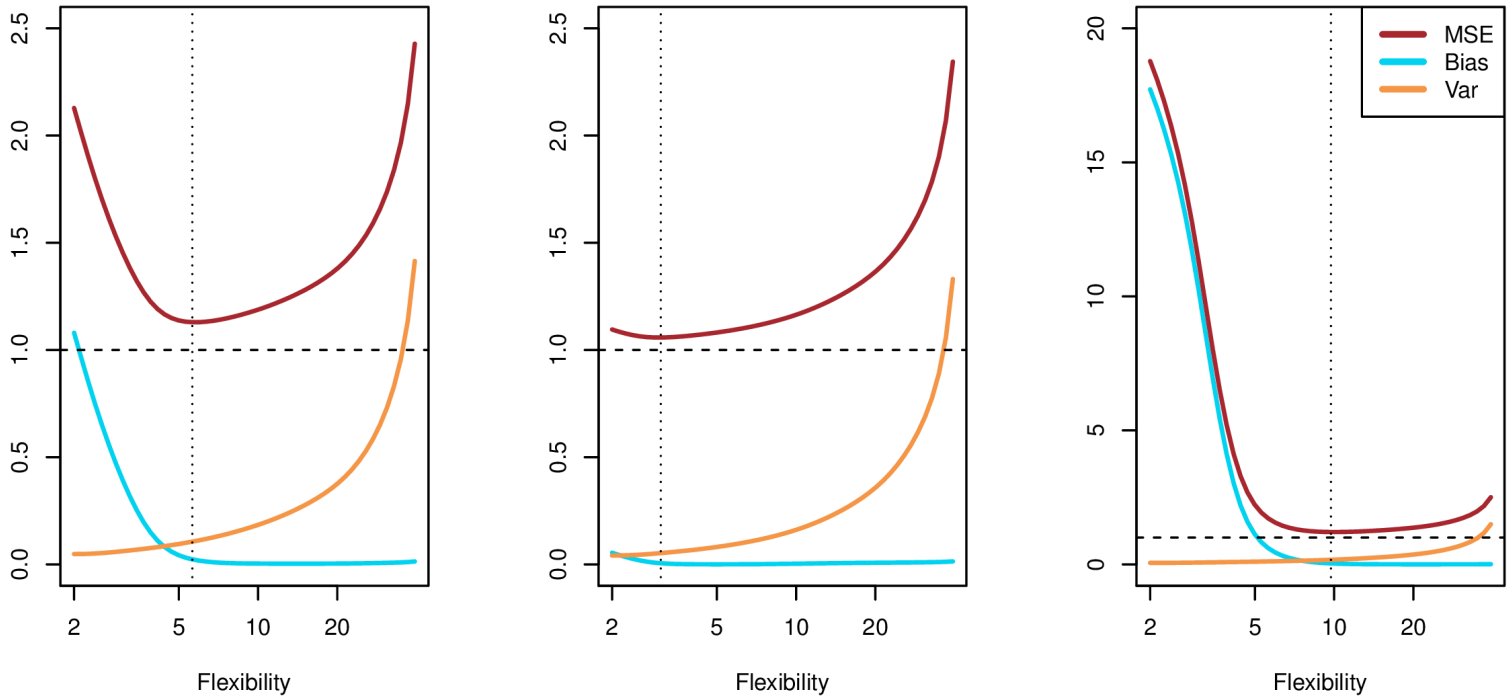
The expected **test** MSE for a new $y_0$ at a new observation, called $\mathbf{x}_0$, will be equal to

$$E[(y_0 - \hat{f}(\mathbf{x}_0))^2] = [\text{Bias}(\hat{f}(\mathbf{x}_0))]^2 + \text{Var}(\hat{f}(\mathbf{x}_0)) + \text{Var}(\varepsilon)$$

Test MSE = Bias $^2$ + Variance + Irreducible variance

- The expectation averages over the variability of $Y$ as well as the variability in the training data.
- As the flexibility of $\hat{f}$ increases, its variance increases and its bias decreases.
- The decide on the best model, from a range with different flexibility, you choose based on average test MSE at the bias-variance trade-off, where both are minimised.

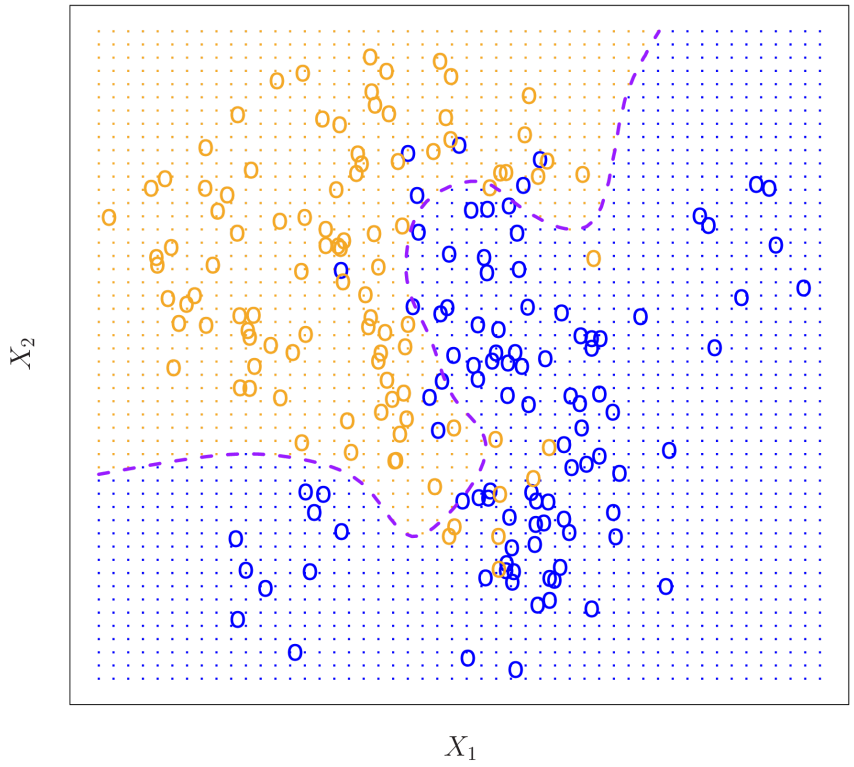
# Bias-variance tradeoff




Squared bias, variance, Var(ε) (dashed line), and test MSE for the three data sets shown earlier. The vertical dotted line indicates the flexibility level corresponding to the smallest test MSE.

(Chapter2/2.12.pdf)

# A case study using nearest neighbours classification




Colour indicates true class of each observation.

Dashed line indicates true boundary.

(Chapter2/2.13.pdf)

# K Nearest Neighbours (KNN)

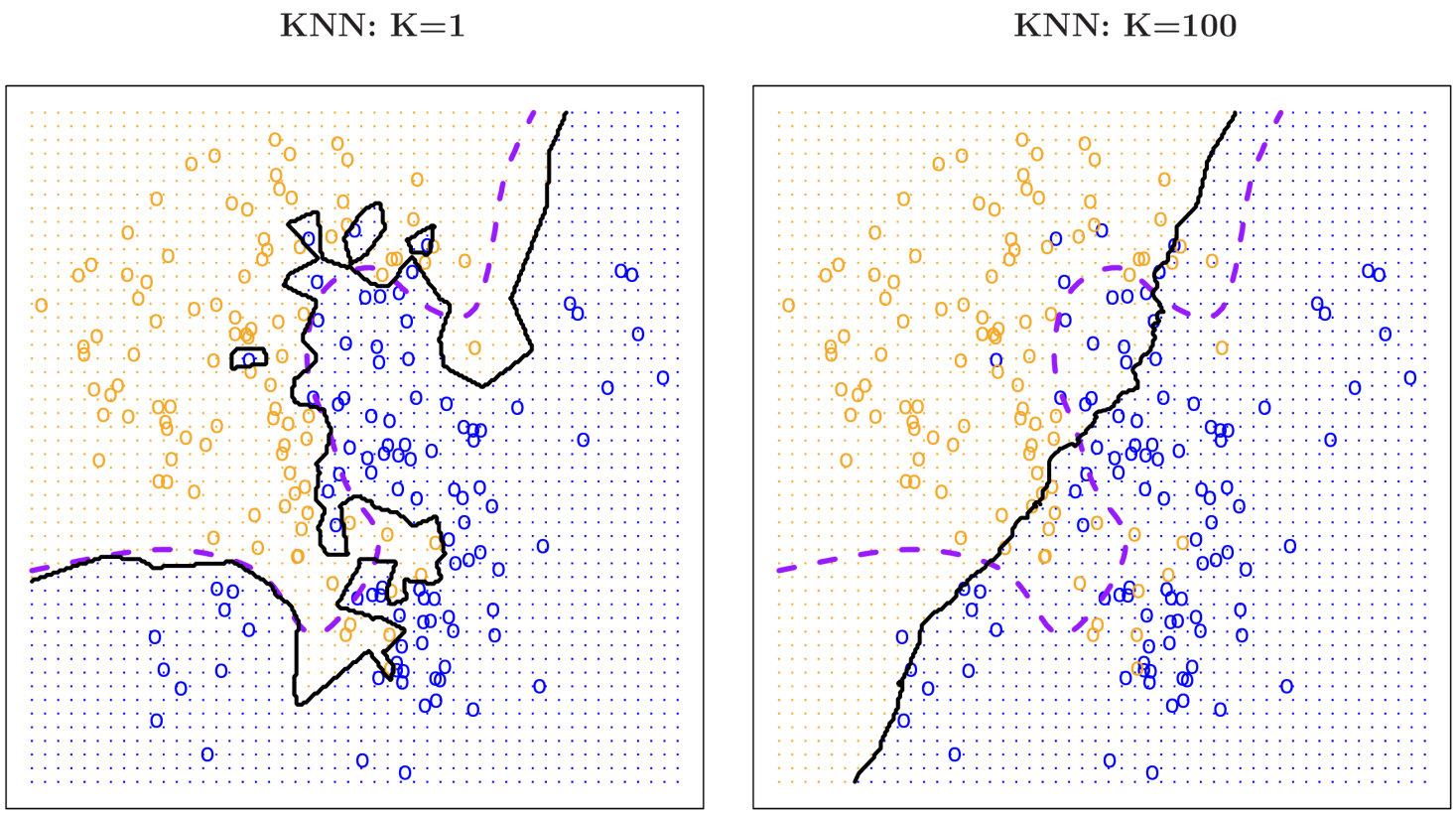
One of the simplest classifiers. Given a test observation $\mathbf{x}_0$,

- Find the $K$ nearest points to $\mathbf{x}_0$ in the training data, call this $\mathcal{N}_0$.
- Estimate conditional probabilities

$$P(Y = C_j \mid \mathbf{X} = \mathbf{x}_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = C_j).$$

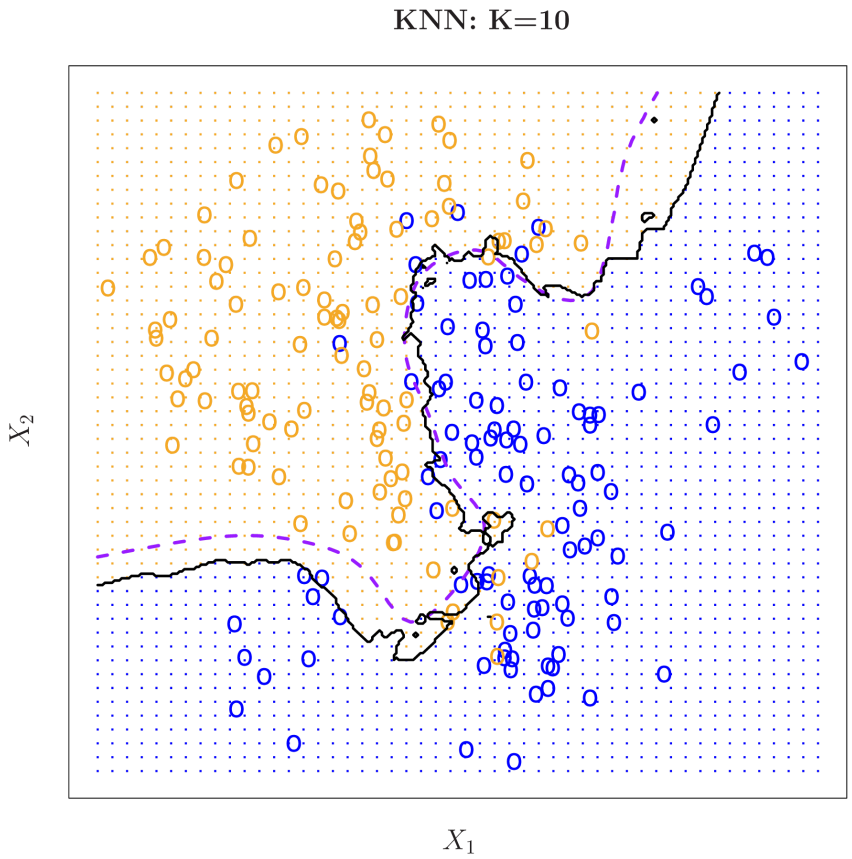
- Classify $\mathbf{x}_0$ to class with largest probability.

# KNN: too flexible and not enough



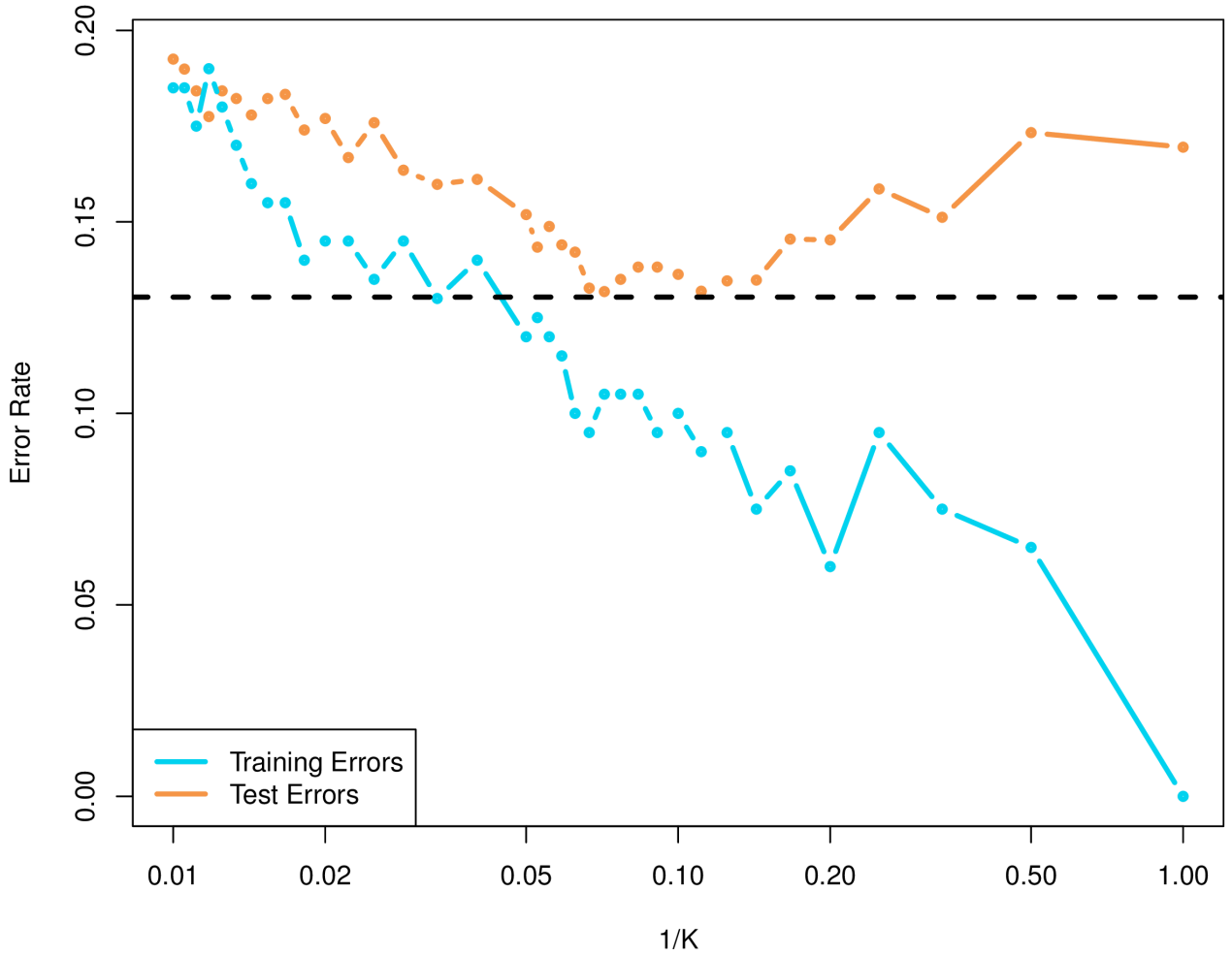

(Chapter2/2.16.pdf)

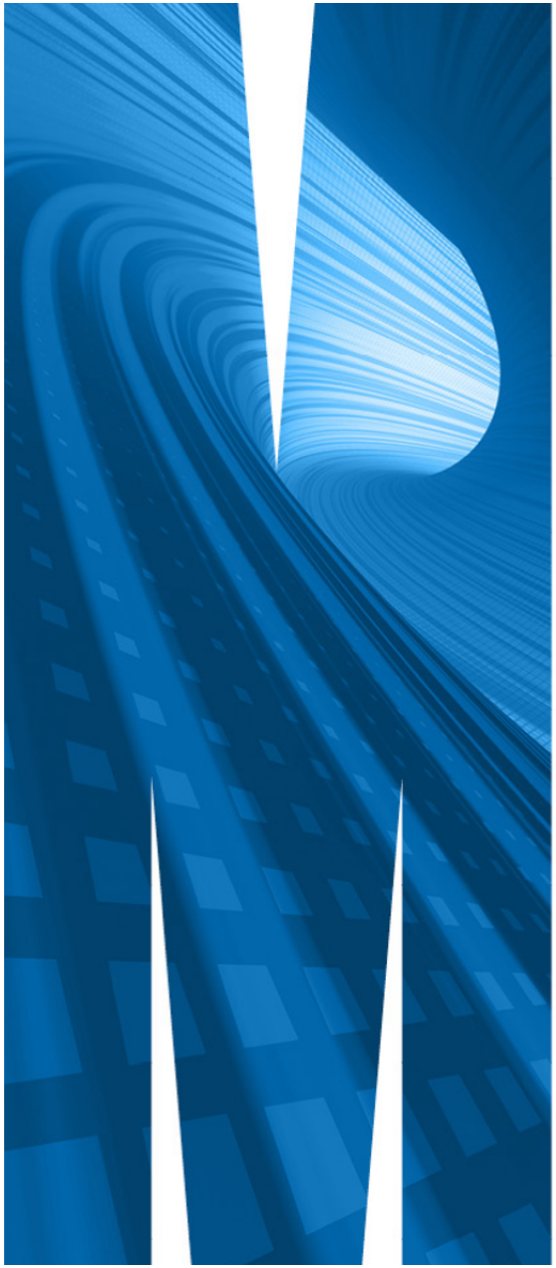
# KNN: about as good as possible




(Chapter2/2.15.pdf)

# KNN: bias variance trade-off, using training and test error







Lecturer: *Professor Di Cook*

Department of Econometrics and Business Statistics

ETC3250.Clayton-x@monash.edu

Week 1b