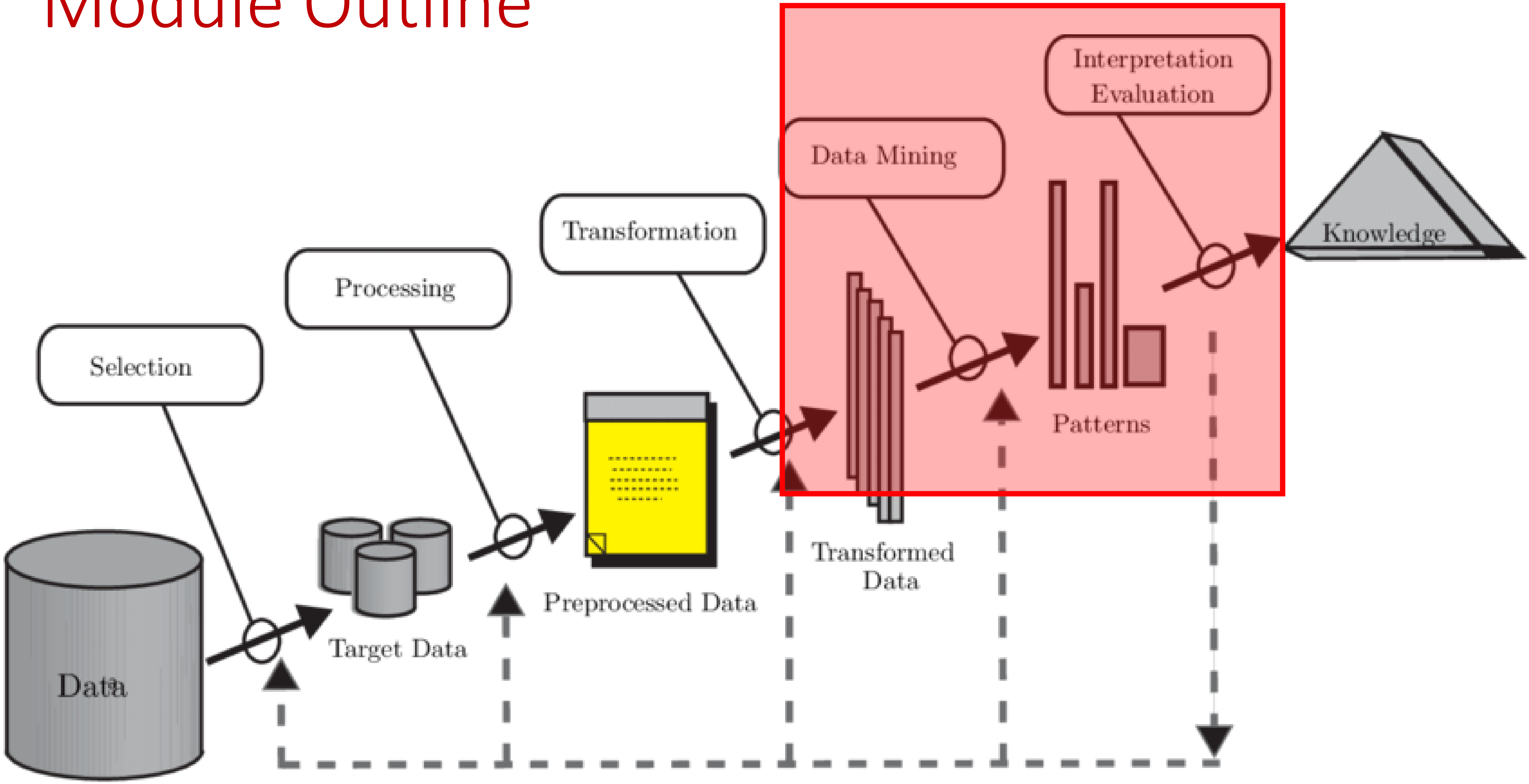
Modelling

Lecture CS1AC16

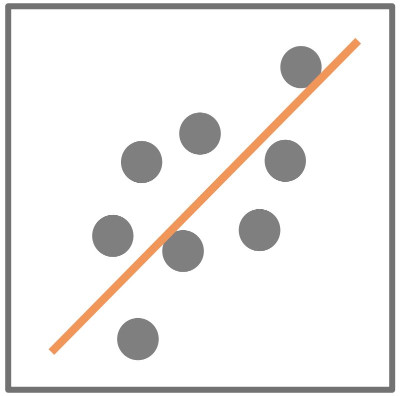
Dr Varun Ojha
University of Reading

09/03/2022

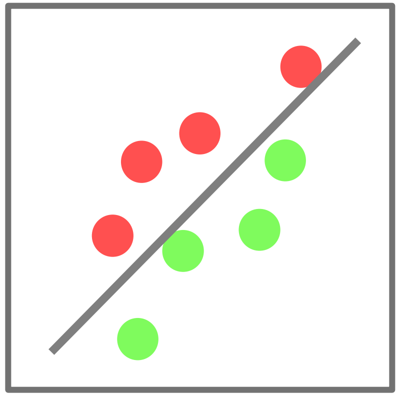
Module Outline



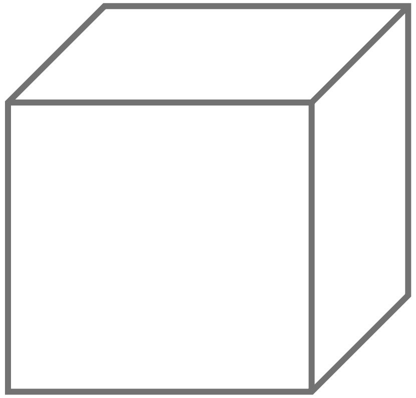
Modelling



Regression Model



Classification Model



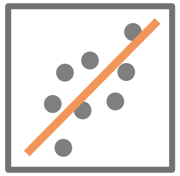
Model Evaluation

Why data modelling?

- We build **mathematical models** to help understand data

- Models can be used in two ways:
  - **Descriptive (inference)**:
    - how feature X influences outcome Y
    - how changes in X result in changes in Y

  - **Predictive:**
    - to learn the relationship between X and Y
    - to predict a value of Y for some values of X

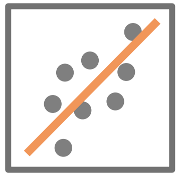
Regression and Classification

Knowing your target attribute type

	#	Inputs Attributes (Independent)		Target/Class/Output Attributes (Dependent)
		A1	**A2**	**A3**
Records	**Ex. 0**	$A1_0$	$A2_0$	$A3_0$
	Ex. 1	$A1_1$	$A2_1$	$A3_1$
	Ex. 2	$A1_2$	$A2_2$	$A3_2$
	Ex. 3	$A1_3$	$A2_3$	$A3_3$
	Ex. 4	$A1_4$	$A2_4$	$A3_4$
	Ex. 5	$A1_5$	$A2_5$	$A3_5$
	Ex. 6	$A1_6$	$A2_6$	$A3_6$
	Ex. 7	$A1_7$	$A2_7$	$A3_7$
	Ex. 8	$A1_8$	$A2_8$	$A3_8$
	Ex. 9	$A1_9$	$A2_9$	$A3_9$

A3 → Continuous **(Numerical)** labeled data → Regression

A3 → Discrete **(Categorical)** labeled data → Classification

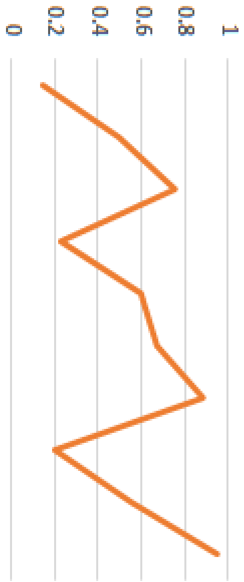


Regression and Classification

Data preparation

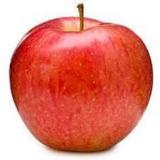
Continuous
labeled data

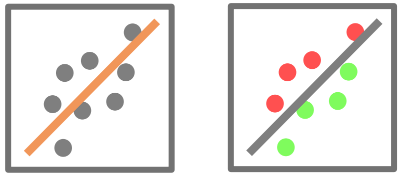
#	Inputs (Descriptor of Price) (X)		Target (Y)
	Area (m^2)	Distance(mile)	Price (£Bn)
Ex. 0	76.85	17.27	0.15
Ex. 1	76.97	19.54	0.5
Ex. 2	77.10	18.51	0.76
Ex. 3	85.28	46.09	0.23
Ex. 4	85.42	35.83	0.6
Ex. 5	88.02	2.59	0.67
Ex. 6	77.25	6.34	0.89
Ex. 7	77.49	6.98	0.2
Ex. 8	85.81	12.18	0.55
Ex. 9	98.81	2.18	9.45



Discrete
labeled data

#	Inputs (Descriptor of Fruits) (X)		Class (Y)
	Length (mm)	Weight (gm)	Fruit
Ex. 0	23.2	3.2	Apple
Ex. 1	70.9	19.5	Orange
Ex. 2	60.5	18.51	Orange
Ex. 3	24.5	4.6	Apple
Ex. 4	110.0	35.83	Orange
Ex. 5	23.8	3.7	Apple
Ex. 6	25.8	4.5	Apple
Ex. 7	24.7	4.9	Apple
Ex. 8	85.8	25.6	Orange
Ex. 9	78.8	20.33	Orange

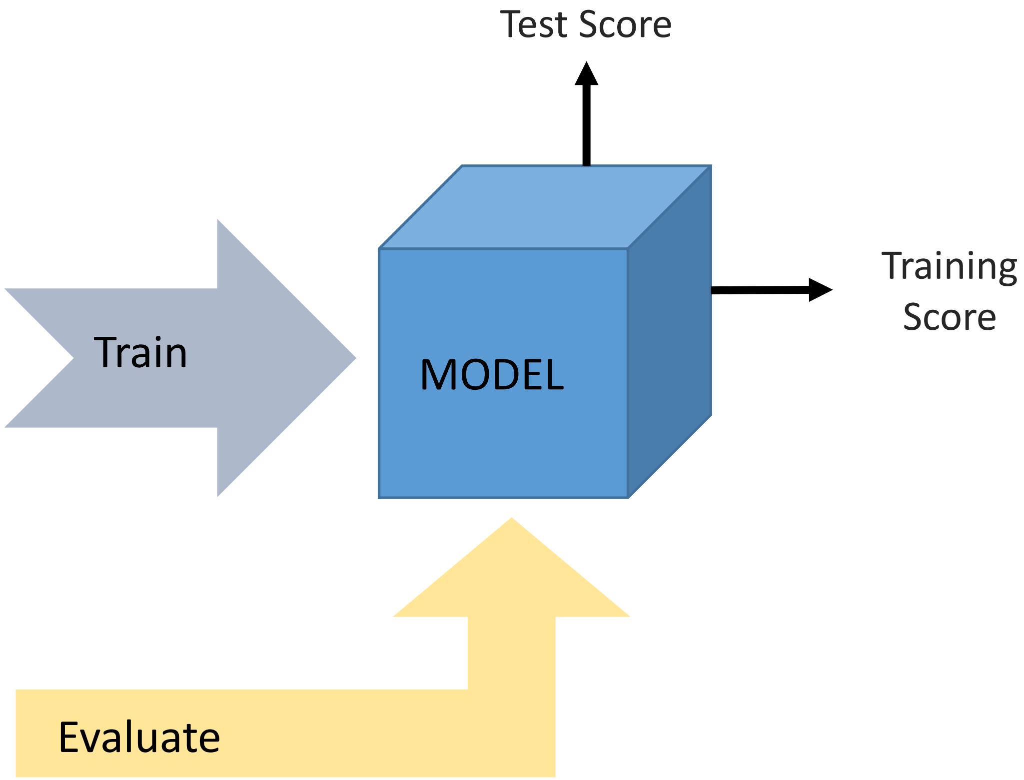




Regression and Classification

The goal of Regression and Classification is to produce a model

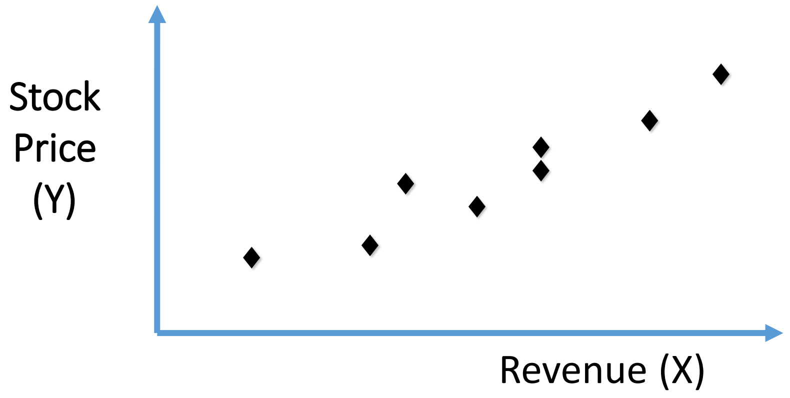
	#	X Inputs Attributes (Independent)		Y Target/Class Attributes Dependent
		A1	**A2**	**A3**
Training Records (Set)	**Ex. 0**	$A1_0$	$A2_0$	$A3_0$
	Ex. 1	$A1_1$	$A2_1$	$A3_1$
	Ex. 2	$A1_2$	$A2_2$	$A3_2$
	Ex. 3	$A1_3$	$A2_3$	$A3_3$
	Ex. 4	$A1_4$	$A2_4$	$A3_4$
	Ex. 5	$A1_5$	$A2_5$	$A3_5$
Test Records (Set)	**Ex. 6**	$A1_6$	$A2_6$	$A3_6$
	Ex. 7	$A1_7$	$A2_7$	$A3_7$
	Ex. 8	$A1_8$	$A2_8$	$A3_8$
	Ex. 9	$A1_9$	$A2_9$	$A3_9$

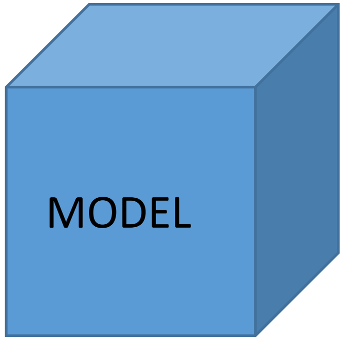


Regression Models

Table of paired data values (x, y)

X (Revenue)	Y (Stock price)
3.4	0.7
4.7	9.5
5.2	1.1
…	…



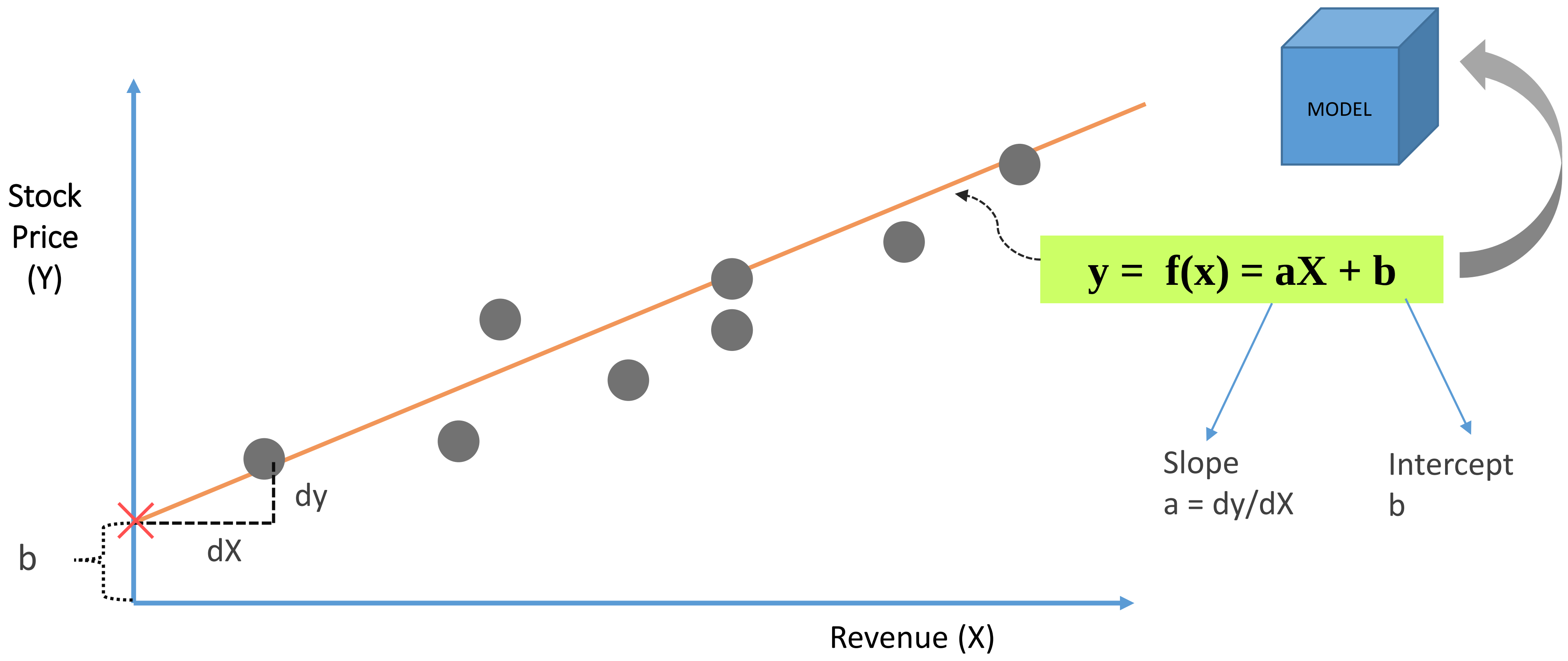


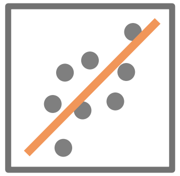
Create: *a* model

- ✓ to describe the data and
- ✓ to predict the outcome value **y** given a new instance **x**

Regression Models

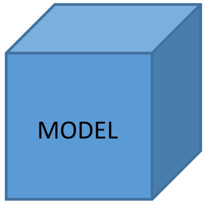
Regression Model Construction



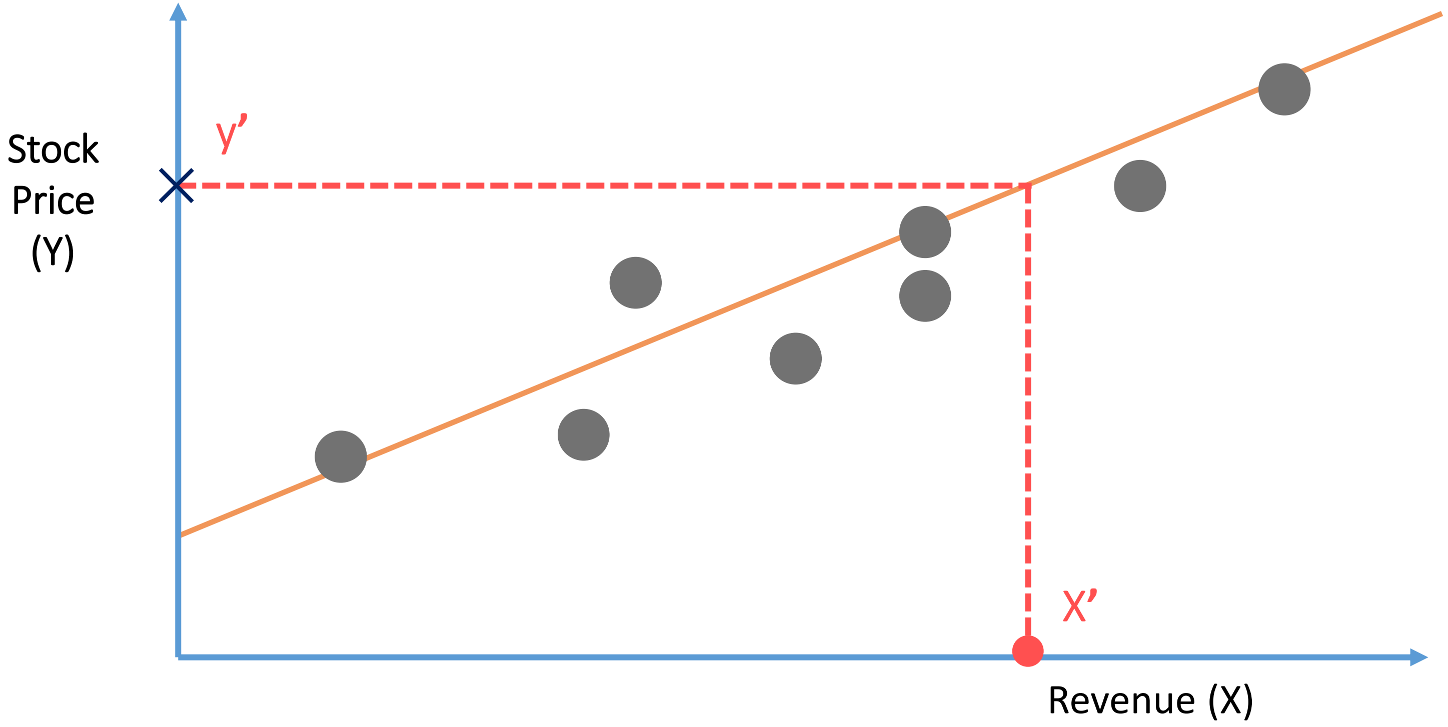


Regression Models

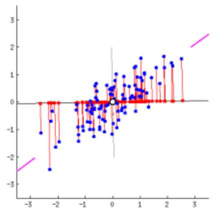
Prediction from the model (for new point X')



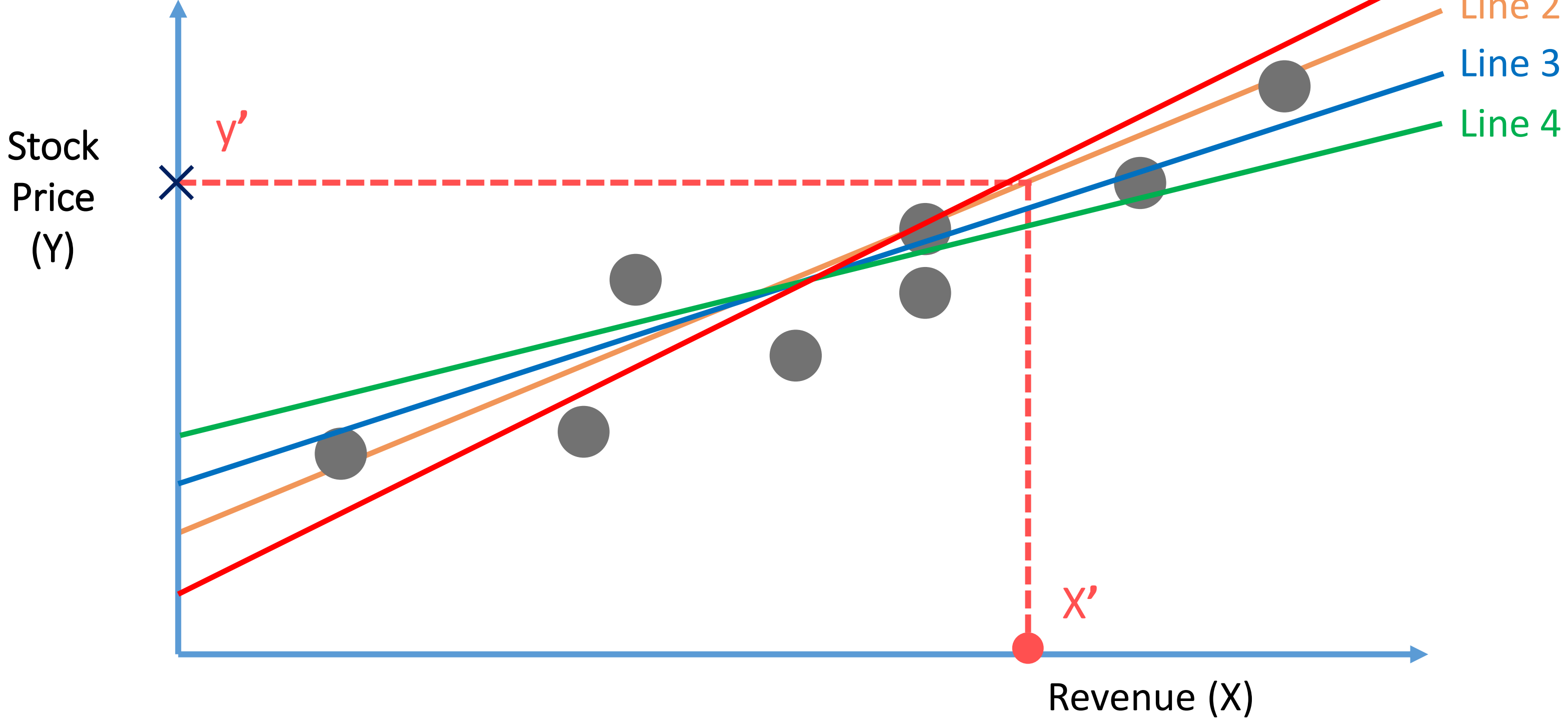
$$y' = f(X) = aX' + b$$



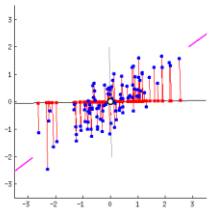
Regression: Which is the best line?

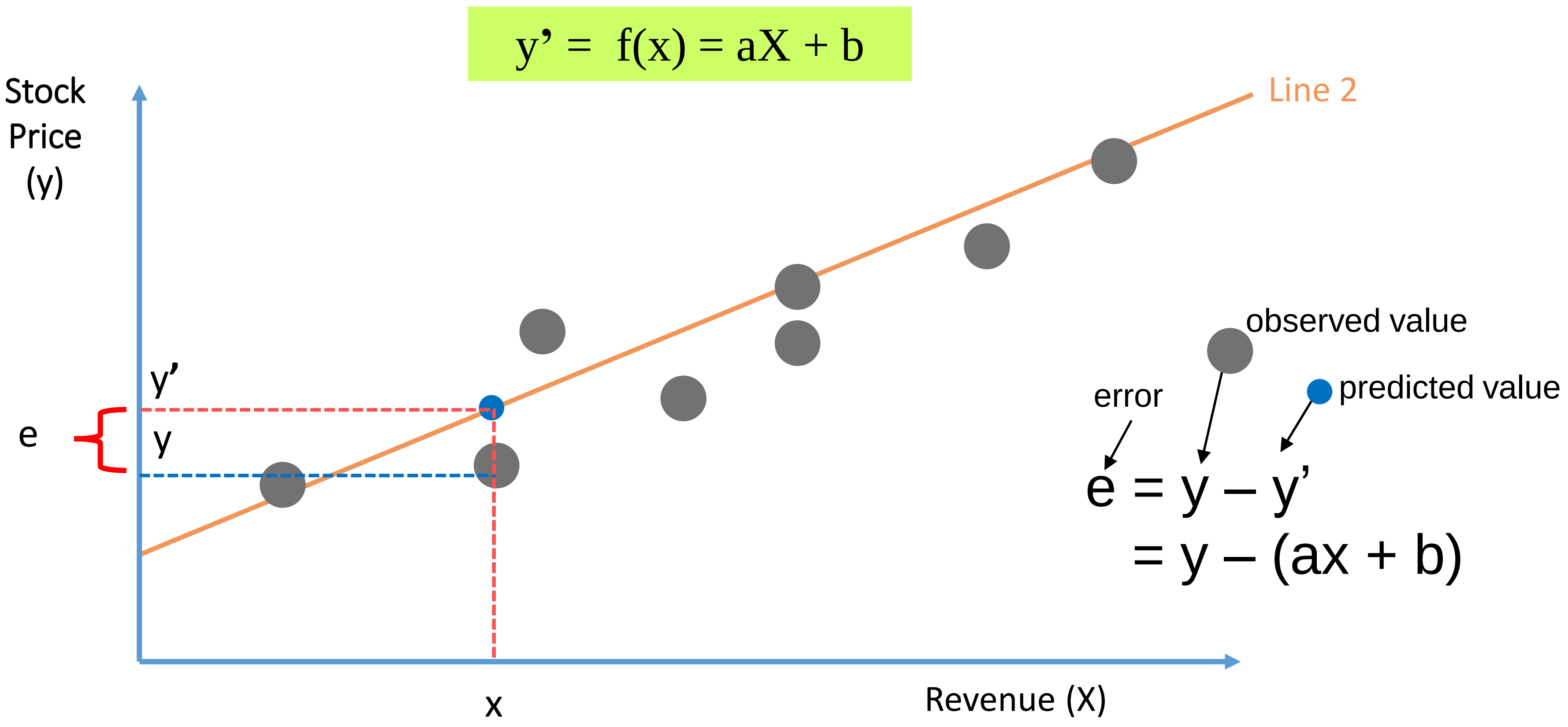


$$y' = f(x) = aX' + b$$



Regression: Finding the “**best fit**” line?

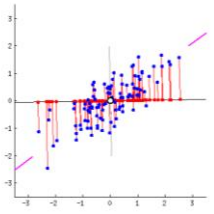


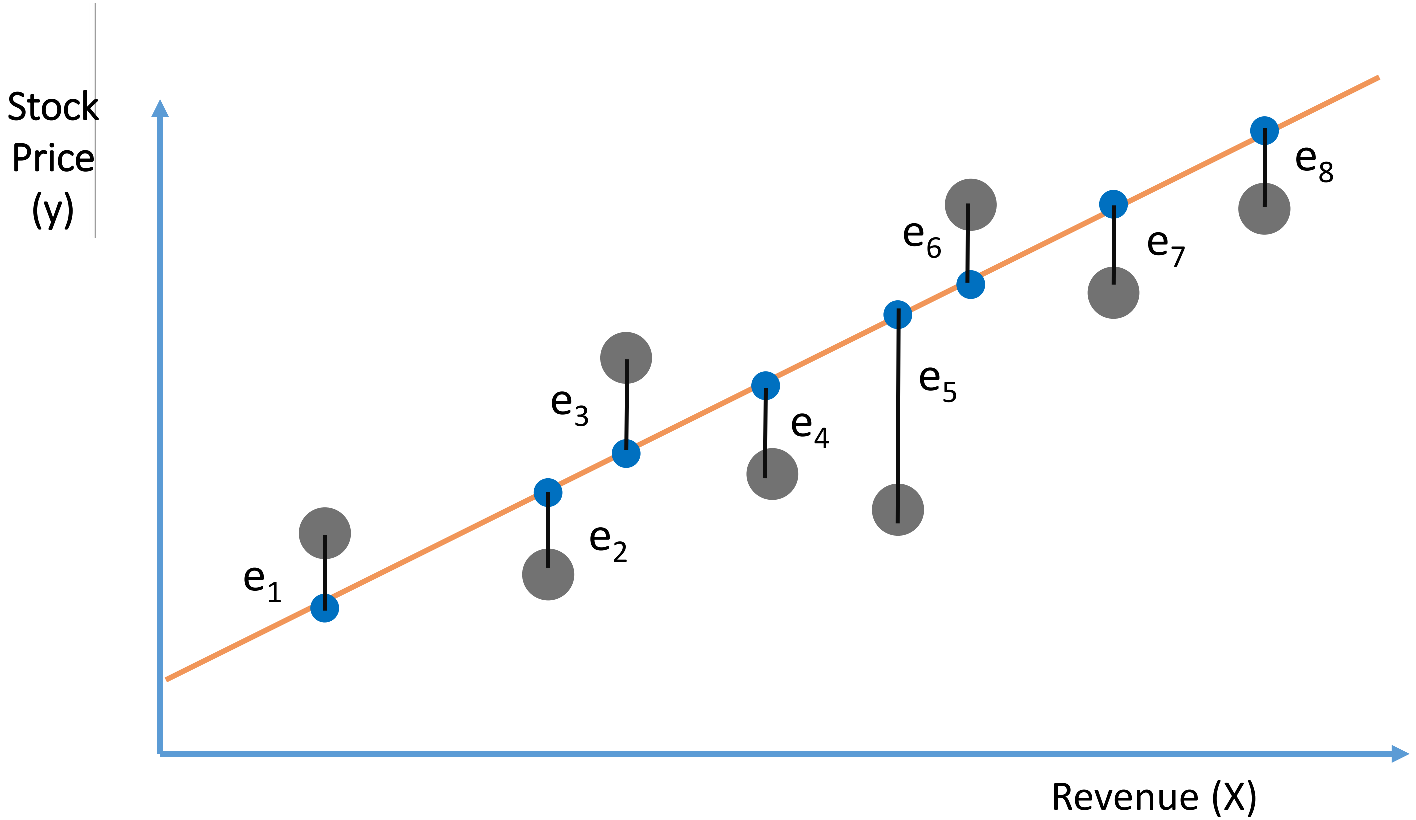


$e = y - y'$

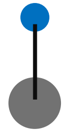
Regression: Least Square

Minimisation of Squared Error





- ✓ Best Fit
- ✓ Find the line (parameters a and b of a line equation) that minimize the norm of the *y* errors
- ✓ (sum of the squares)



Error

$e_i = y_i - y_i'$

$$\|e\|^2 = \sum_i (y_i - y_i')^2$$

Linear Regression Example

Dataset of an advertising company sales

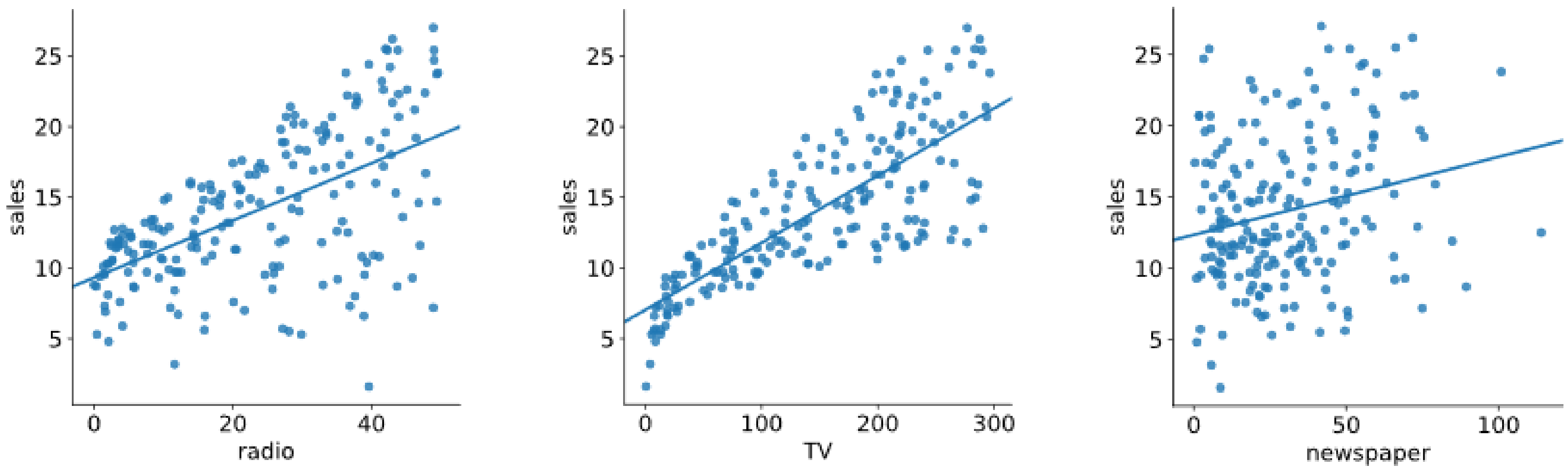
TV	radio	newspaper	sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9

- We are interested in
  - learning **how the different advertising budgets affect sales**.
  - predicting **how changing advertising budgets will affect sales**

Linear Regression Example

- **Three Linear regression models**

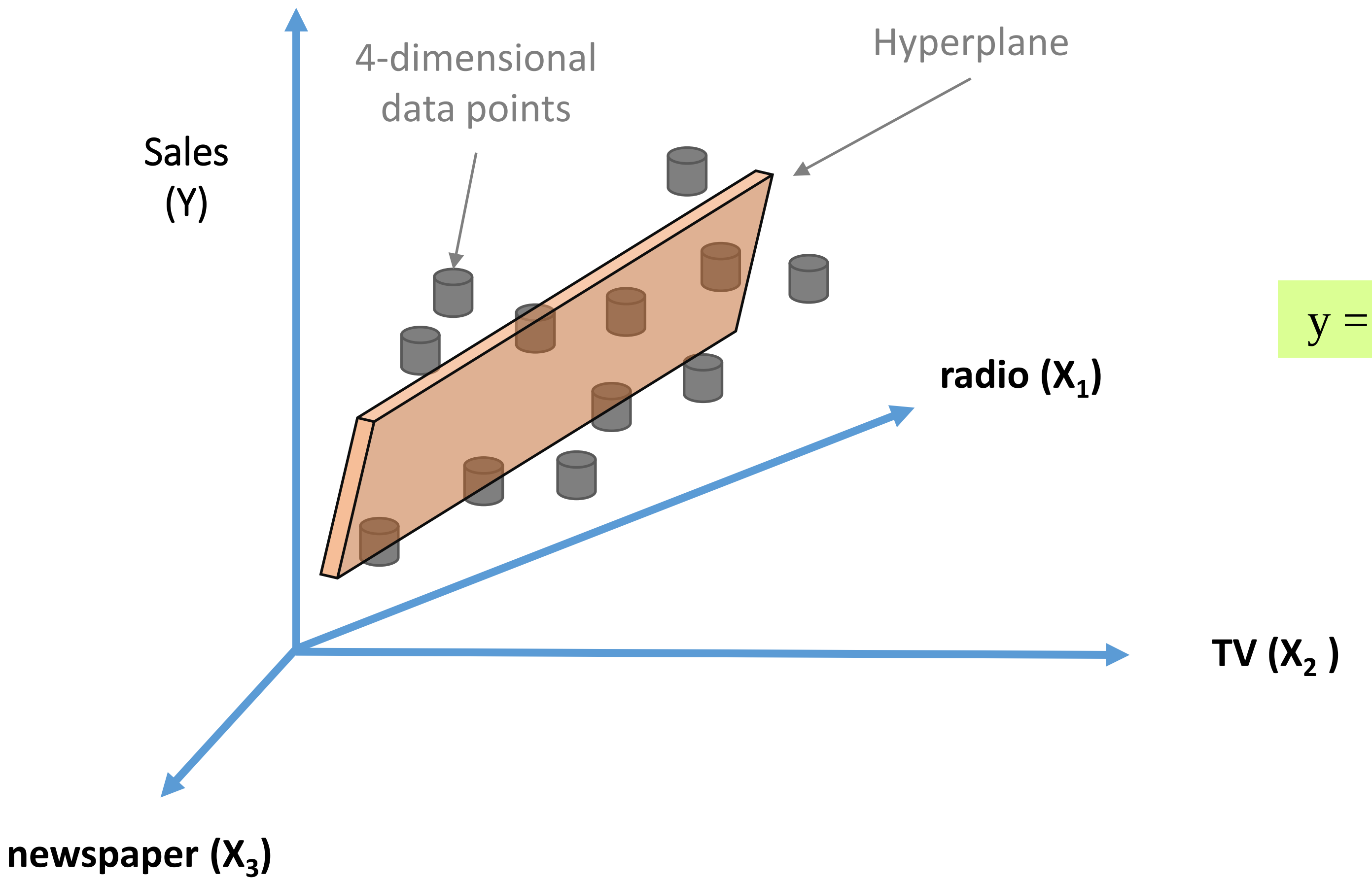
  Multivariate data

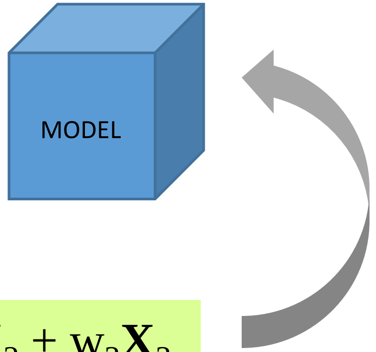


relationship between each of the three different budgets and sales

Multivariate Linear Regression Models

Taking all variables into account





$$y = f(x) = w_0 + w_1\mathbf{X_1} + w_2\mathbf{X}_2 + w_3\mathbf{X}_3$$

Slope
$w_1 = dy/dX$
$w_2 = dy/dZ$

Intercept
w_0

Least Square Estimation

We can sum up multivariate linear regression as

Prediction — input data — coefficient (scaling factor) — noise in data

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

The Model

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

n is the number of examples

p is the number of inputs

Using R2 to score a regression model

- Mean Squared Error (MSE) minimisation is a measure to find the best fit

- **R2 instead is a measure of the sensitivity of the predictions**
  - It describes *the fraction of the variance explained* by the regression model

$$R^2 = \frac{\sum_i (y_i - \overline{y})^2 - (y_i - \hat{y})^2}{\sum_i (y_i - \overline{y})^2}$$

- Where $\bar{y}$ is the mean, $\hat{y}$ predicted value of y

- R2 = 0.61 for TV advertising data means
  - 61% of the variation in the sales figure can be explained by the TV marketing budget.

KNIME example – Iris data

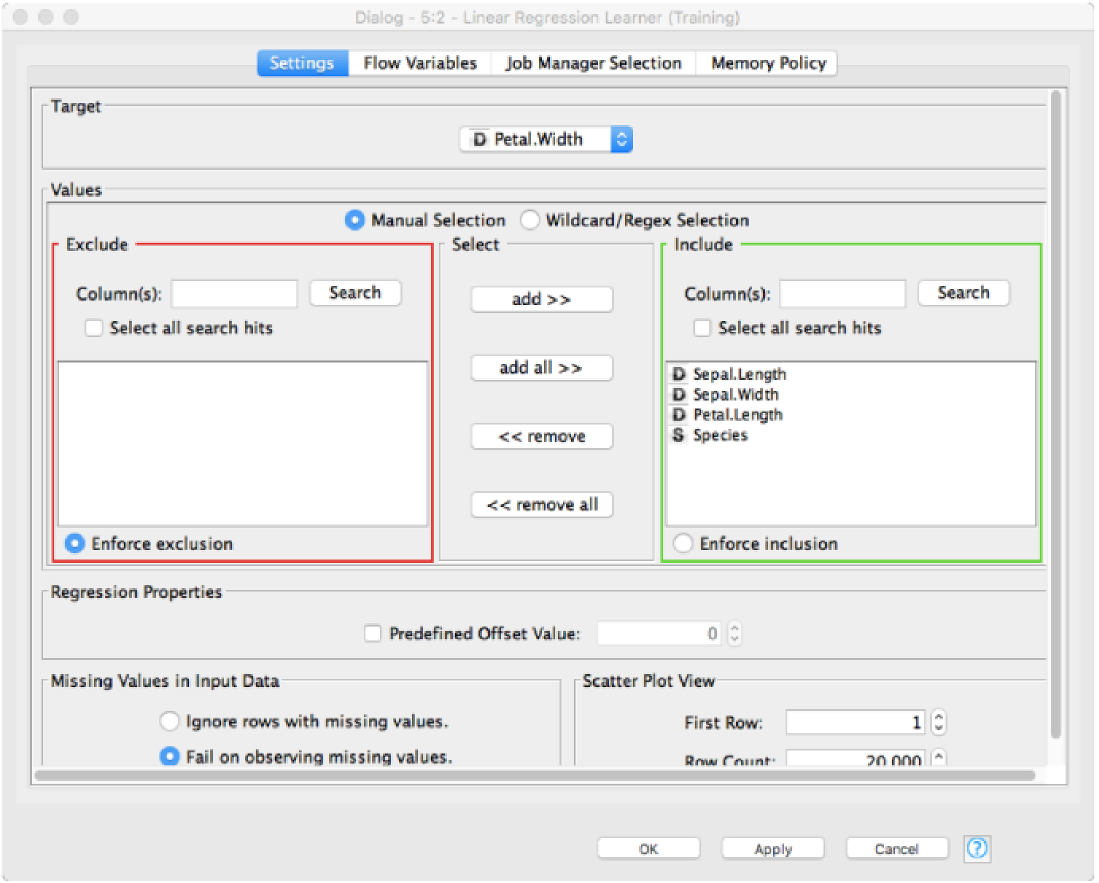
- Measurements of flowers of different species of iris plant.

Target Column (→ Petal.Width)

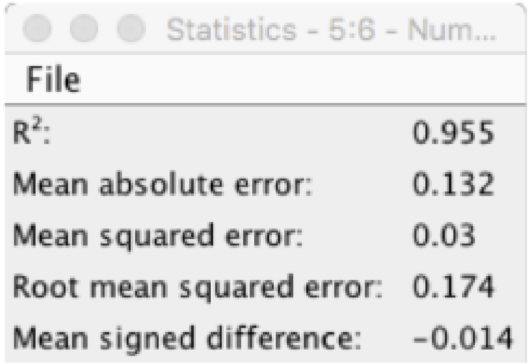
Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa

Regression with Iris Data

- Using linear regression to predict **Petal Width** using the remaining variables



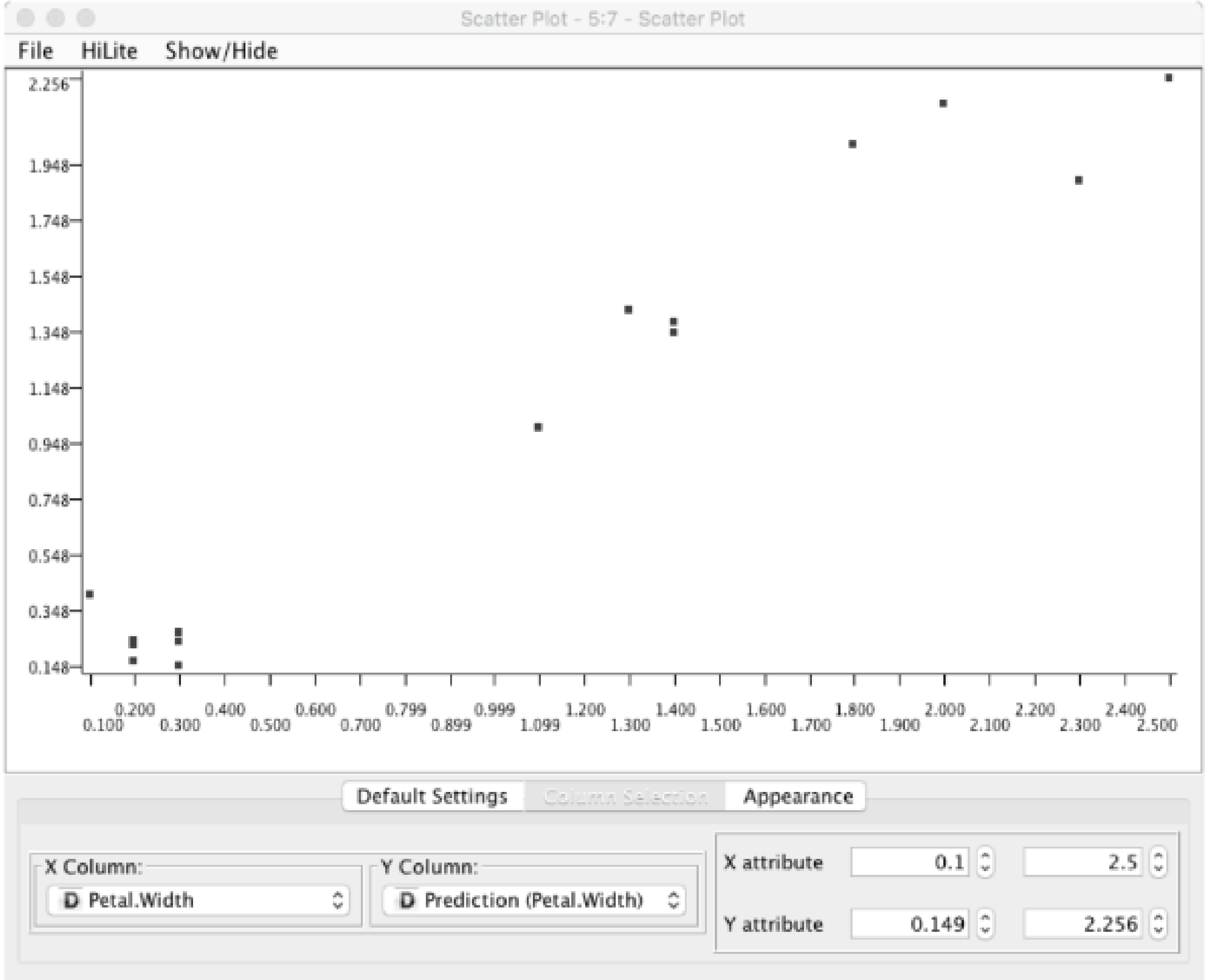
Iris Data Regression: Results



Statistics - 5:6 - Num...

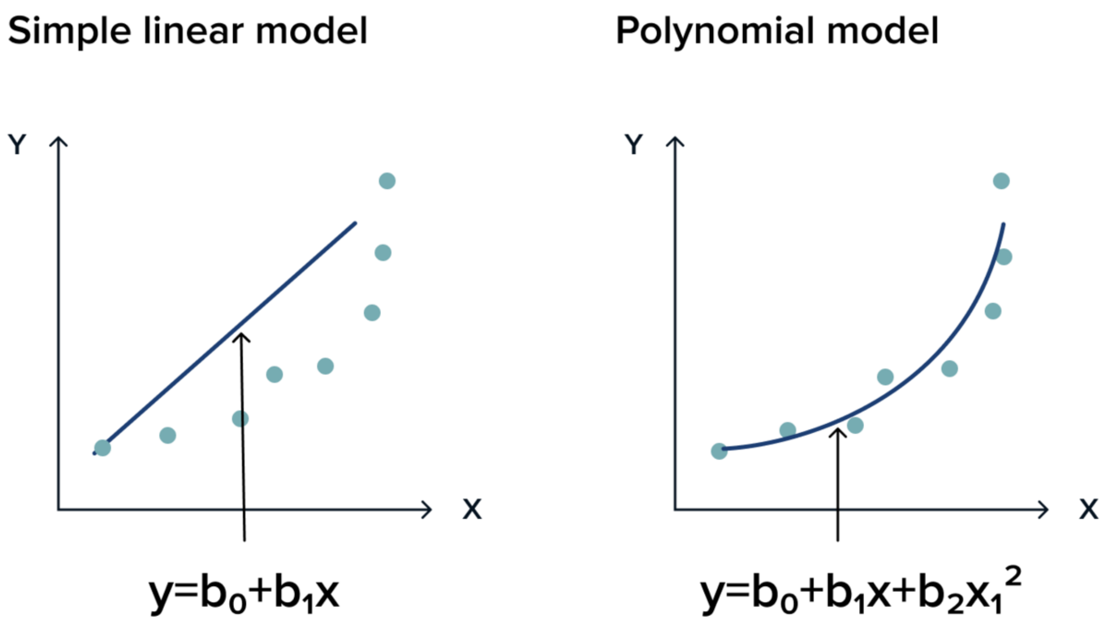
File

R^2:	0.955
Mean absolute error:	0.132
Mean squared error:	0.03
Root mean squared error:	0.174
Mean signed difference:	-0.014



Model Problems: Overfitting

- complex a model makes predictions perfectly on training data, i.e., it **overfits** the data.
- this is not a good model as it does not **generalise** well and performs poorly when new data

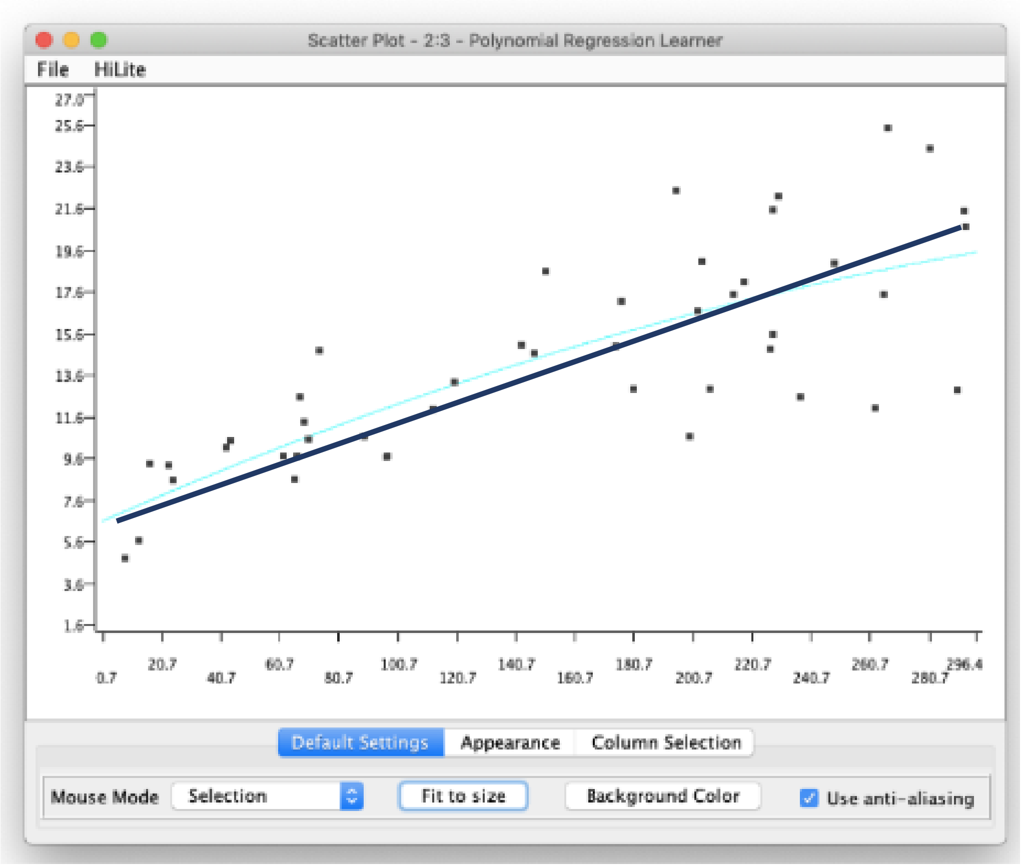


Simple Linear Model (underfit)

- Fit a simple linear equation to the data

- Mean Squared Error (MSE) of the **simplest linear model** on training data is 12.8

- MSE on validation data is **9.51**

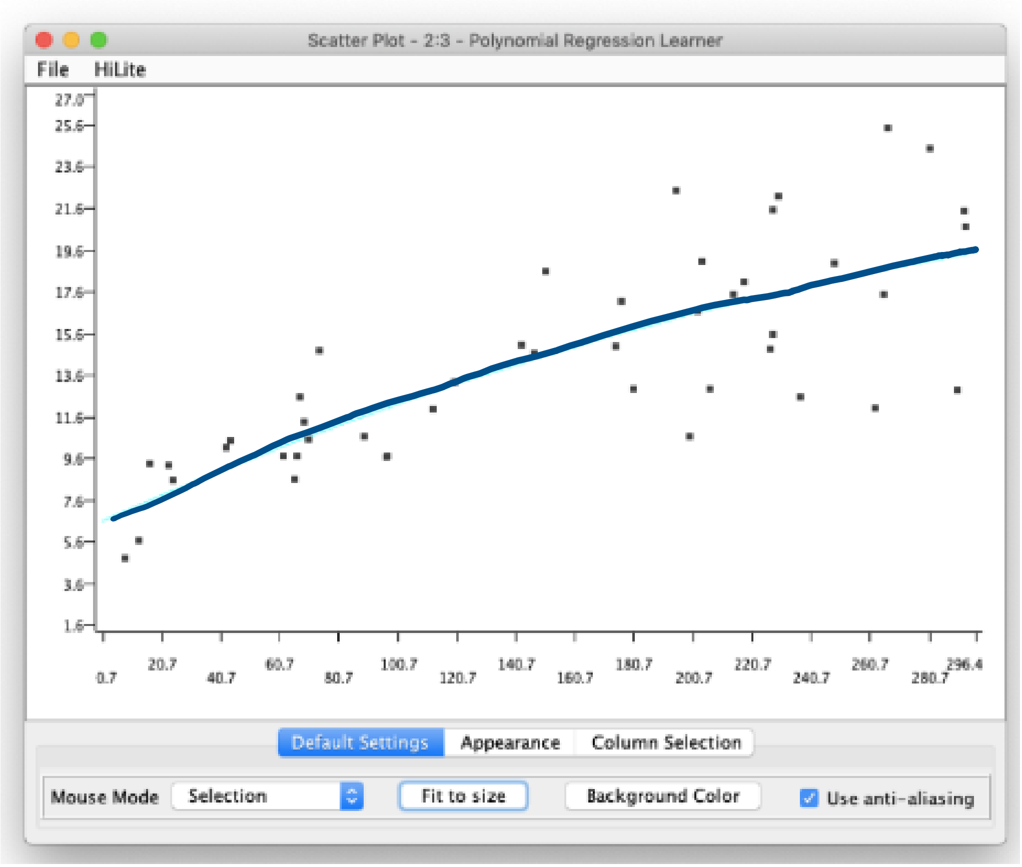


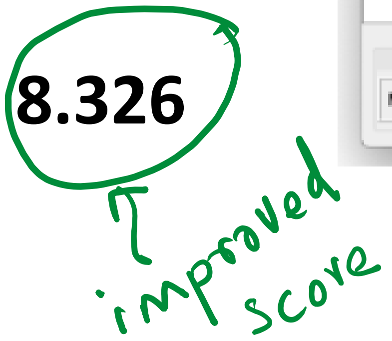
Polynomial fitting (good fit)

- The quadratic polynomial equation improves the regression fit over the linear regression model

- Mean Squared Error (MSE) of a **simple model** on training data is 9.21.

- MSE on validation data is **8.326**

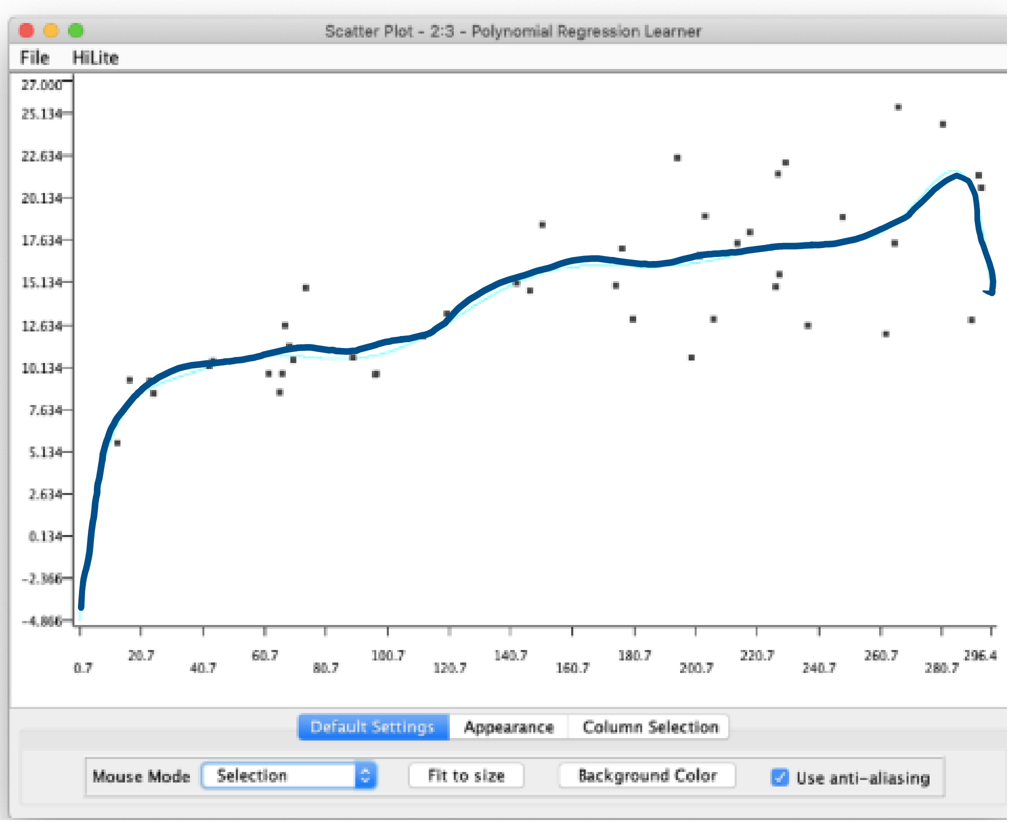




Extreme Polynomial fitting (overfit)

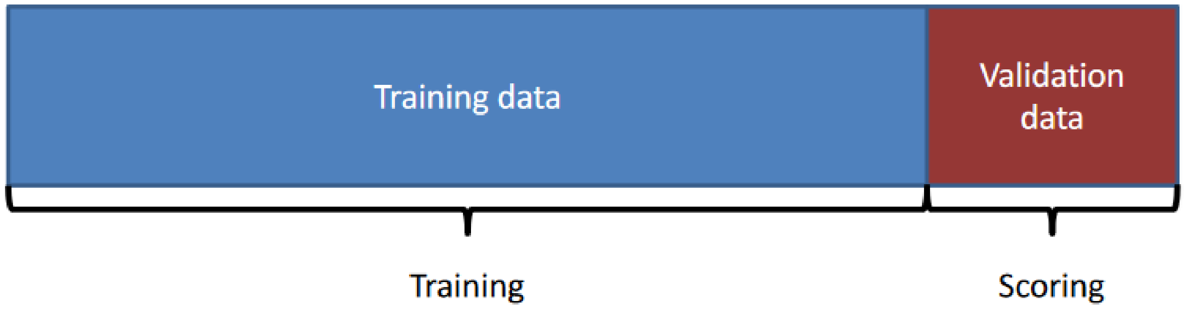
- Mean Squared Error (MSE) of a **complex model** on training data is 5.34. ← good training

- When used to make predictions on the validation data, MSE is **11.9** ← worse



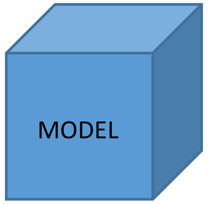
Sampling Bias (avoiding Holdout)

- Holdout partitioning splits the data **only once** into
  - training and
  - validation data



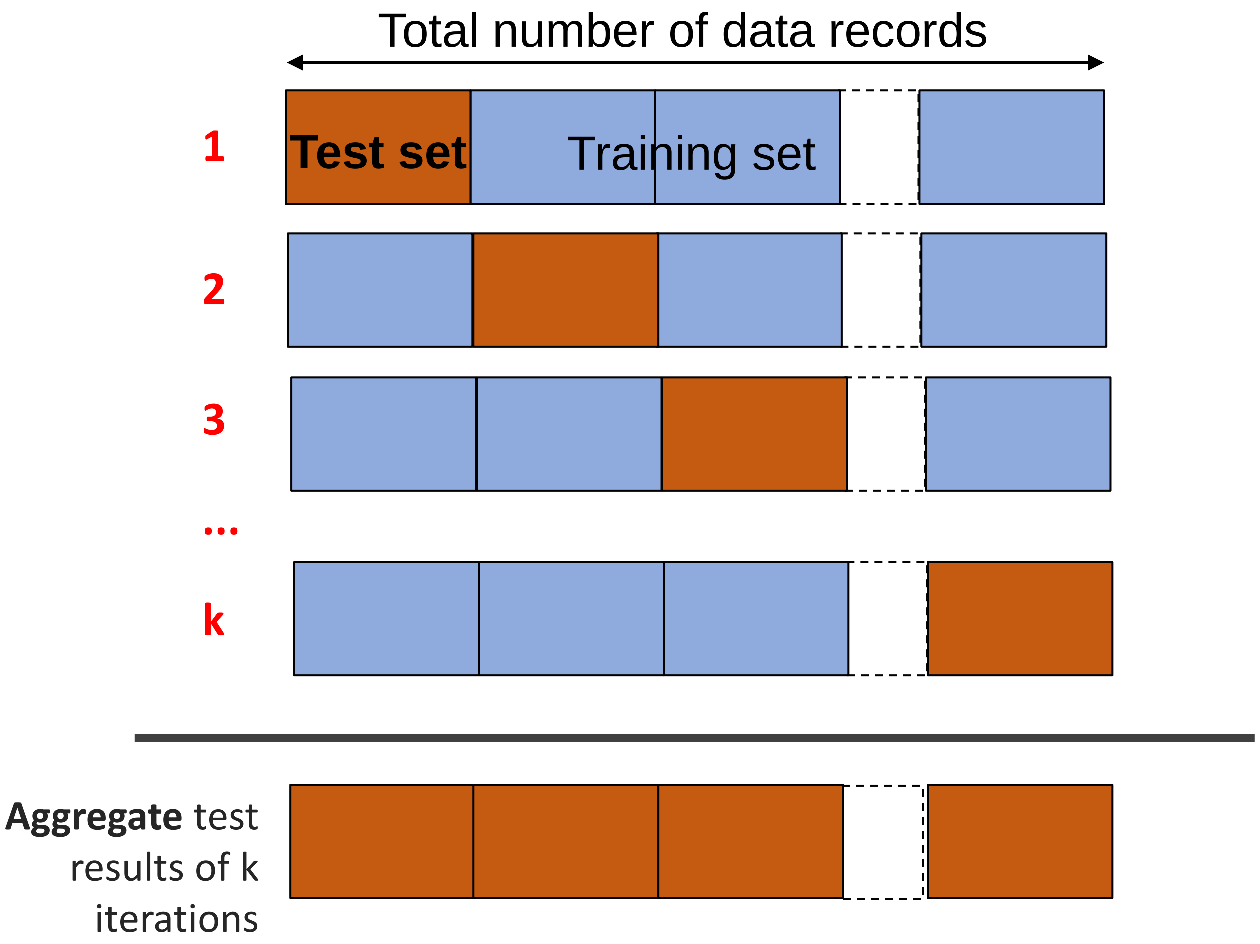
- If the split has been a **lucky split,** then your model will perform well

- **Sampling bias** is common in imbalanced classification problems



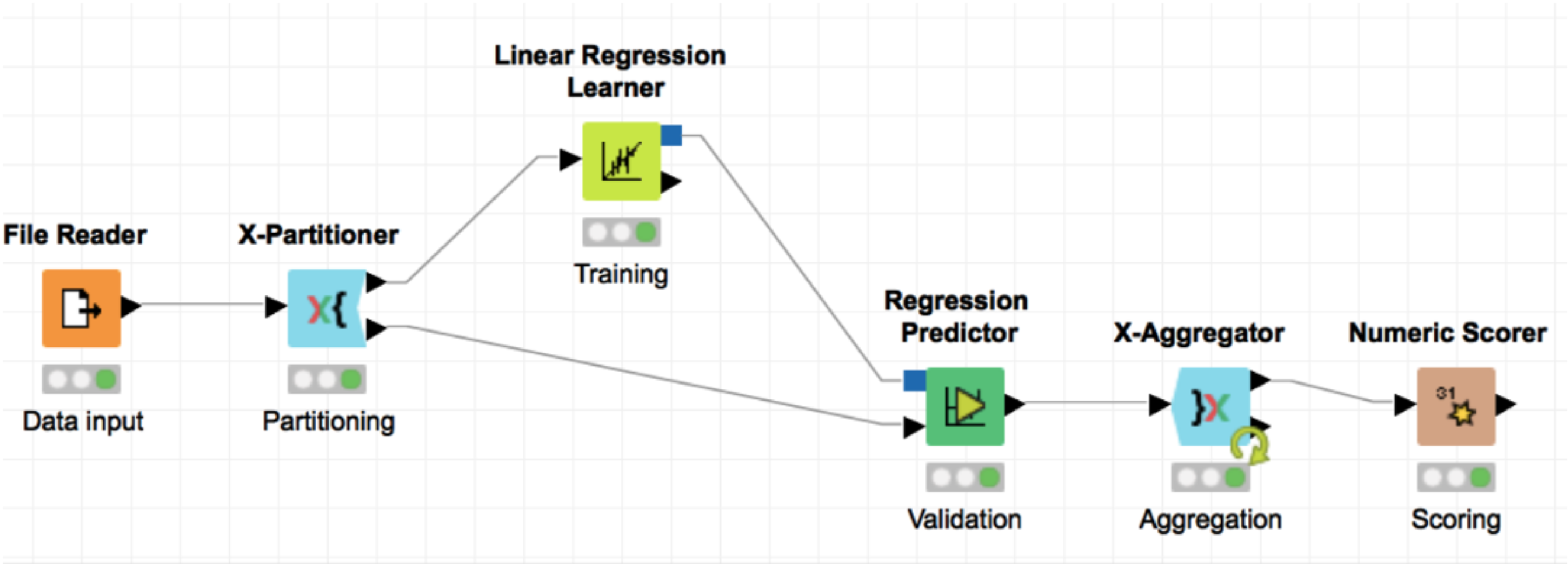
Cross Validation

- **k-fold cross-validation**
- At each fold:
  - Training set
    - **k-1 folds** are used to train a model
  - Test set
    - **1-fold** is used to score a trained model



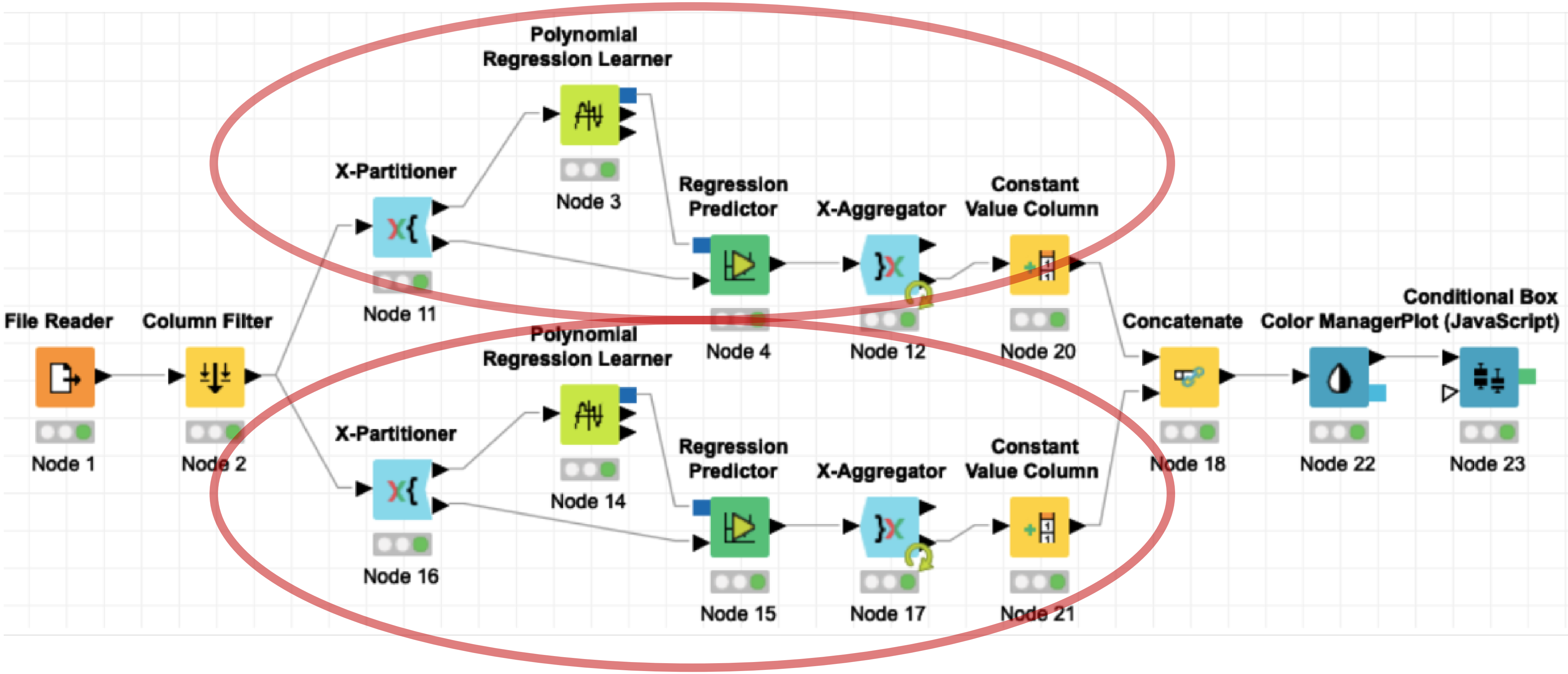
Cross validation in KNIME

- Use **X-Partitioner** split the data
- Use **X-Aggregator** to collect the predictions from each split



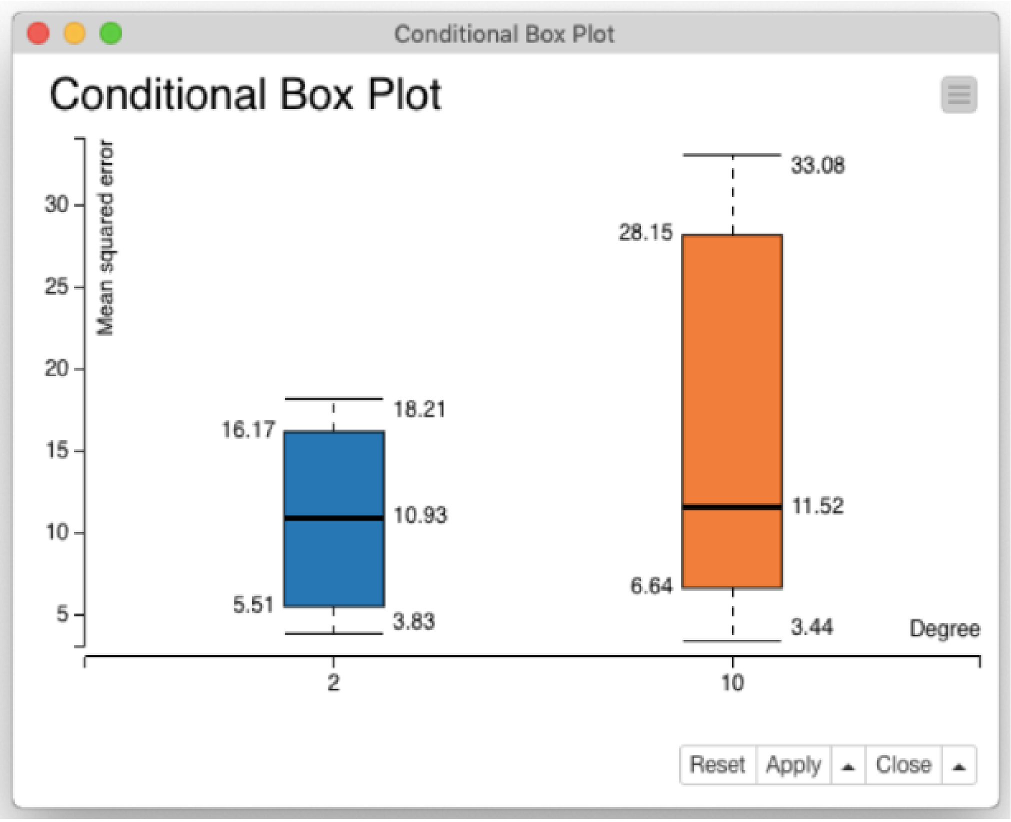
Comparing Models in KNIME

- We can create two or more models with cross-validation
- We can help us compare models to **find the least overfitting model**



Model Comparison Output

- A single partition (holdout) of the data
  - A simple model may give a high error
  - A complex model may give a small error
- A 10-fold cross-validation
  - models produce 10 different scores
  - A model that has a lower variance of scores is the least overfitting model



Conclusions

1. Modelling is used for prediction and inference
2. Classification models assign data points to a discrete class
3. Regression models predict the value of a dependent variable for each data point.
4. Data is partitioned into training and validation data.
5. Model scores depend on how the data is divided into training and validation subsets
6. The best way to avoid overfitting is by using cross-validation

Next Two Weeks

1. Practice Sessions in G56 (Polly Vacher Building)

2. Try KNIME exercises before coming to the lecture

3. Last week you have a **MCQ Blackboard Test**