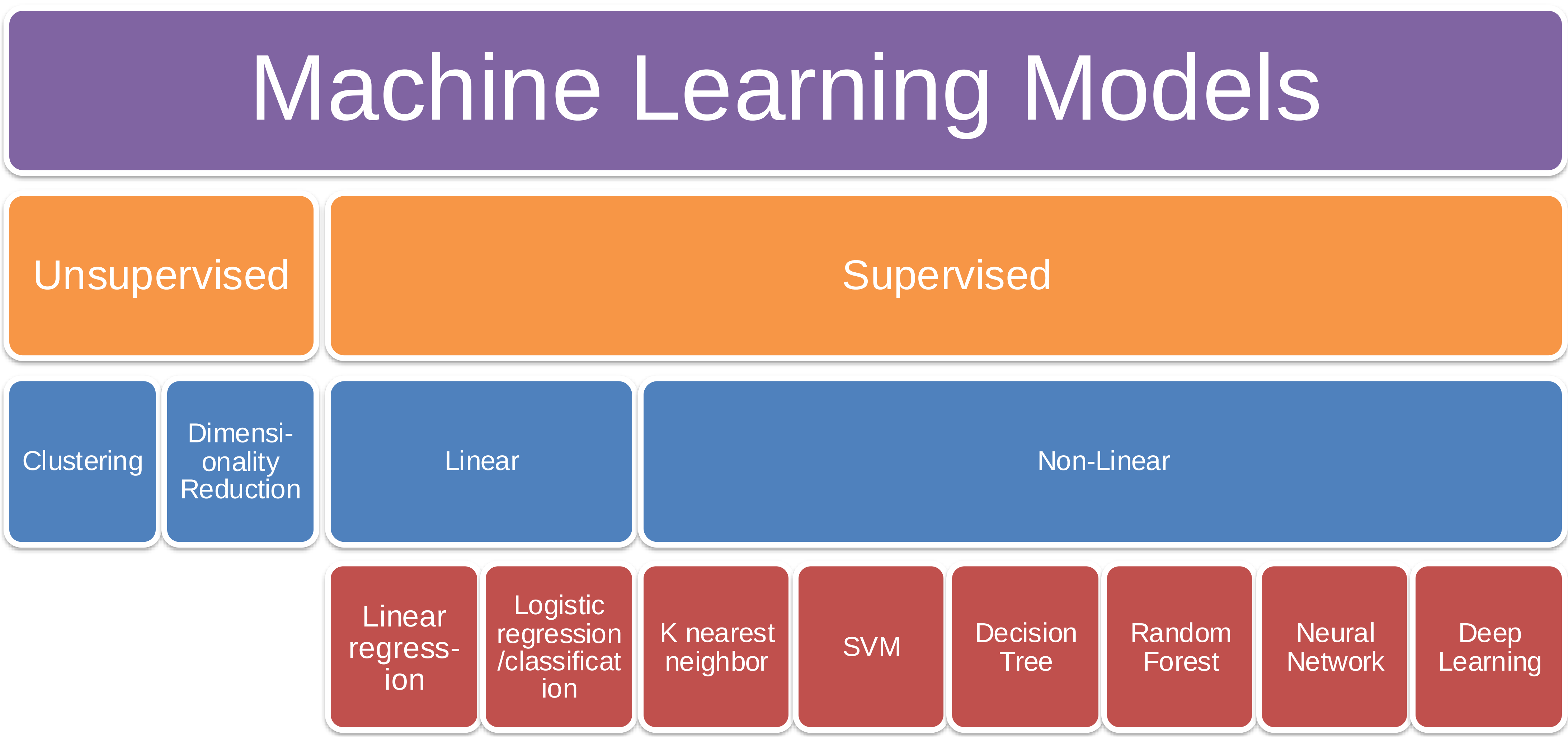
# Machine Learning – Take home messages

- **Week 11**

# Recap of the unit …



# Recap of the unit ...

- Unsupervised
  - Clustering
  - Dimensionality reduction

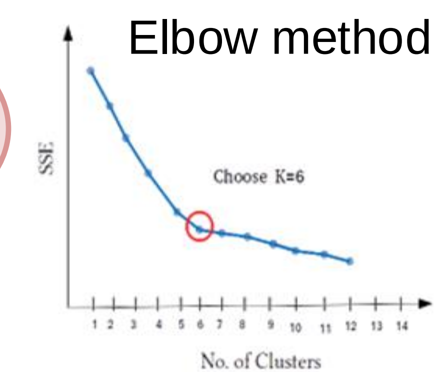

Distance measures

- Euclidean
- Mahalanobis
- Jaccard
- Cosine
- Manhattan

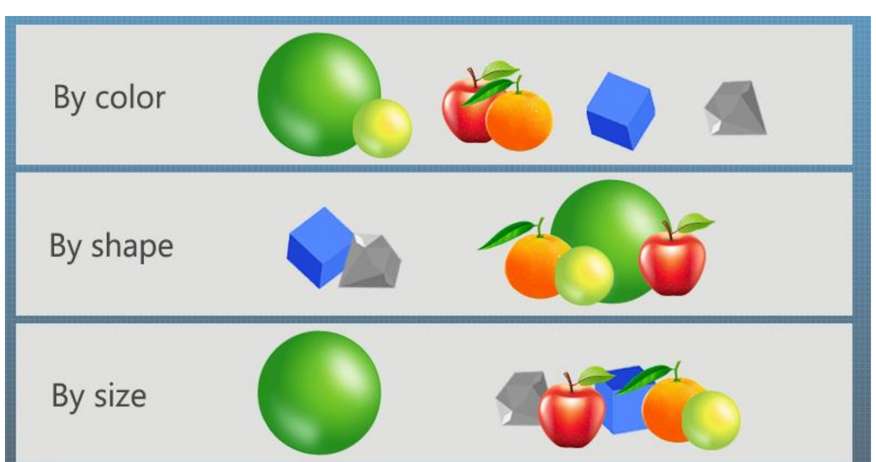
Algorithms

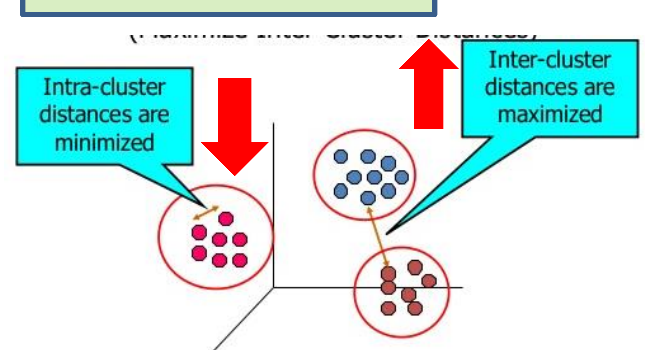
Kmeans++

KMeans




Cluster can vary based on the properties



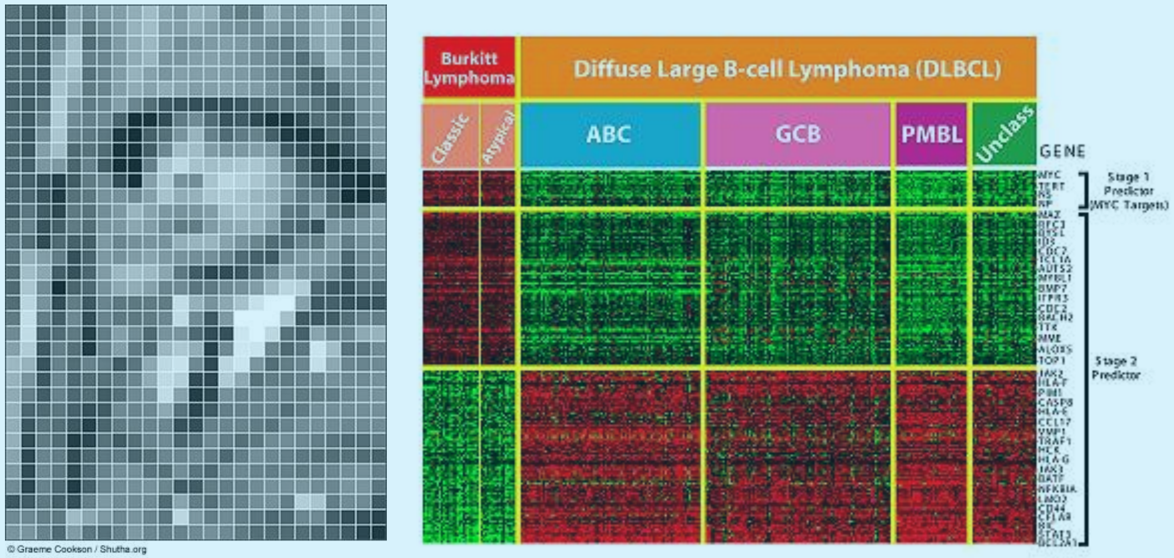

Performance measures

- Rand Index
- Purity
- Mutual Information
- Silhouette Coefficient

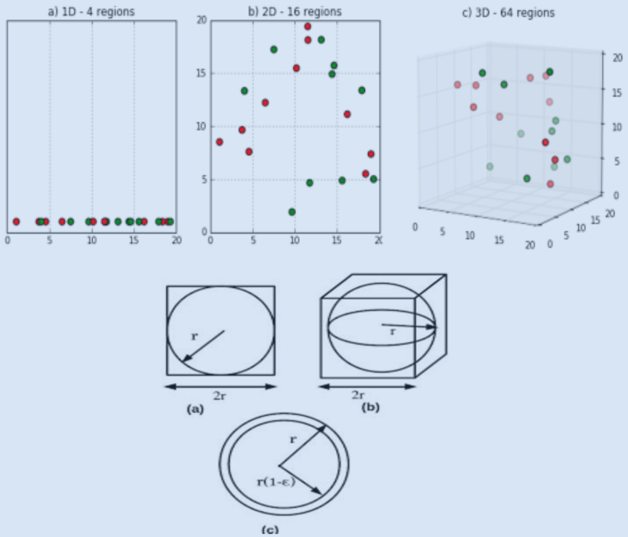
# Recap of the unit …

Unsupervised
- Clustering
- Dimensionality reduction

Example of high dimensional data




Curse of dimensionality




Principal Component Analysis

Correlated High Dimensional Data → Compute Covariance Matrix → Eigenvectors + Eigenvalues → Pick n<d eigenvectors with highest eigenvalues → Project data points to those eigenvectors → Uncorrelated Low Dimensional Data

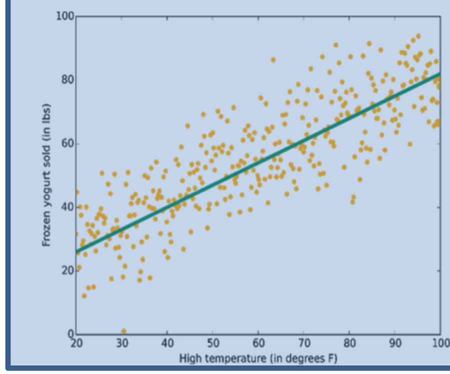
- Linear combinations of all dimensions (w>0)
- Ranked based on described variance
- Drop dimension with very low variance

- Impossible to find effect of individual attributes
- Limits importance of hand-crafted features
- Reduce clarity of domain knowledge

# Recap of the unit ...

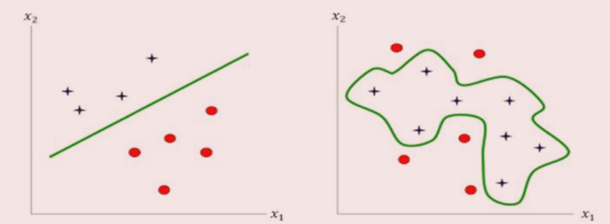
Supervised
- Linear
- Nonlinear

## Regression





## Classification



*Linear decision boundary* *Nonlinear decision boundary*

What we want to learn?

$$h : X \to Y$$

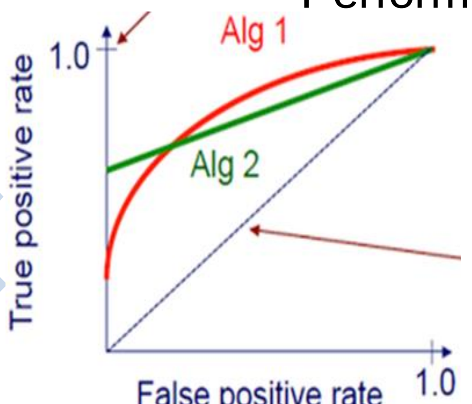
How the machine learns?

$$\min_{h \in H} \frac{1}{n} \sum_{i=1}^{n} L(y_i, h(x_i))$$

Model complexity

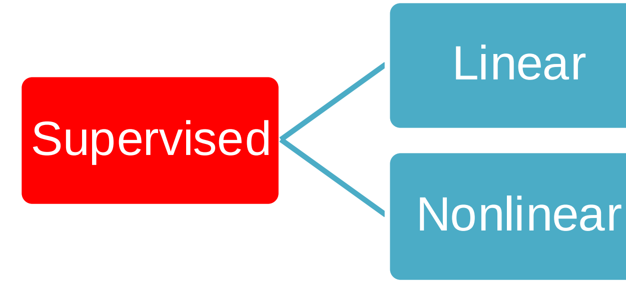
$$R_{str}(h) = R_{emp}(h) + \lambda C(h)$$

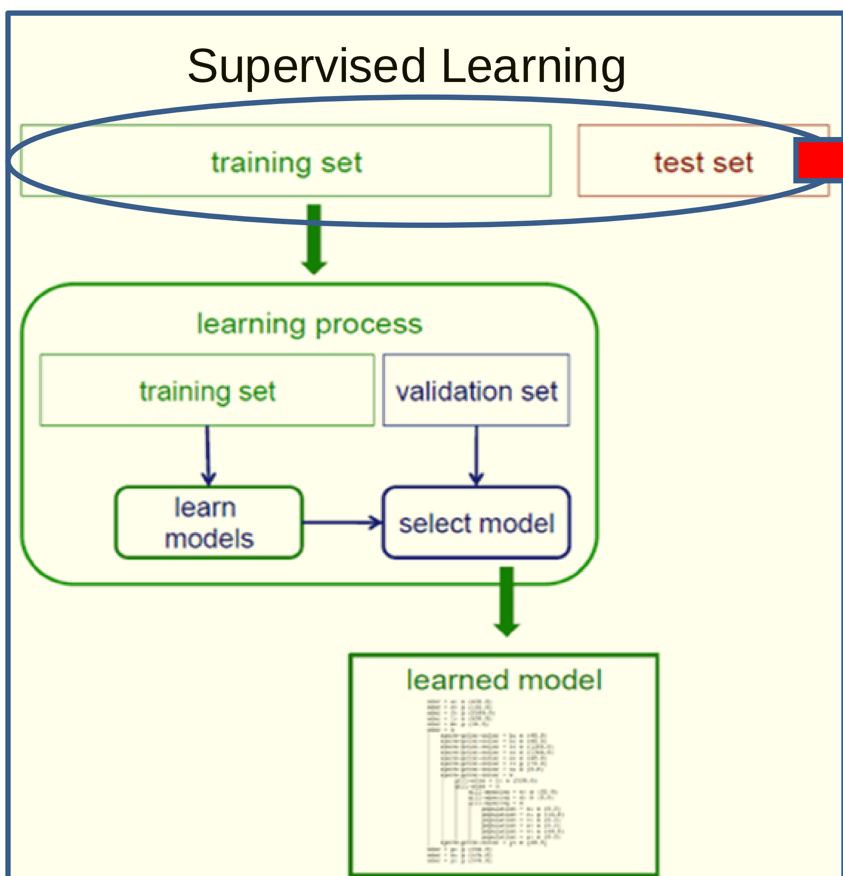
## Performance Measurements

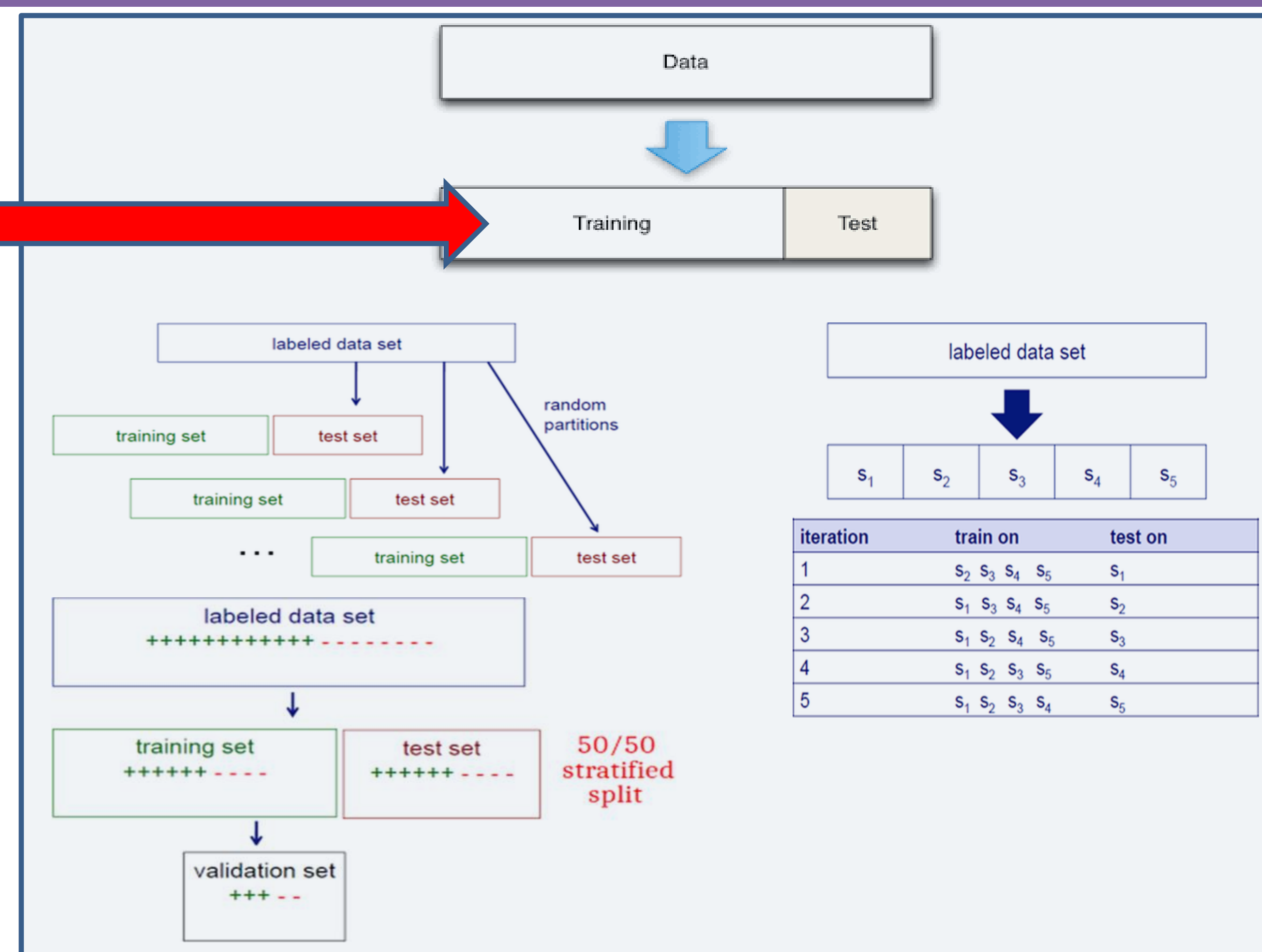



| | | Predicted Class: Positive | Predicted Class: Negative | |
|---|---|---|---|---|
| Actual Class | Positive | True Positive (TP) | False Negative (FN) **Type II Error** | **Sensitivity** $\frac{TP}{(TP+FN)}$ |
| | Negative | False Positive (FP) **Type I Error** | True Negative (TN) | **Specificity** $\frac{TN}{(TN+FP)}$ |
| | | **Precision** $\frac{TP}{(TP+FP)}$ | **Negative Predictive Value** $\frac{TN}{(TN+FN)}$ | **Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$ |

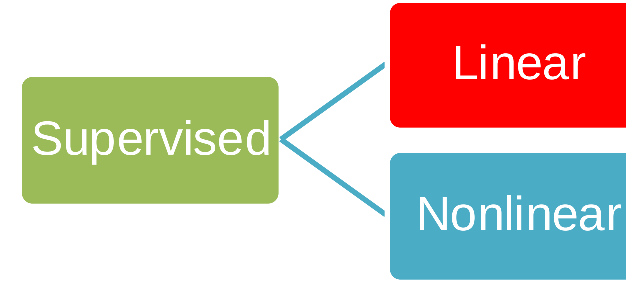
# Recap of the unit …

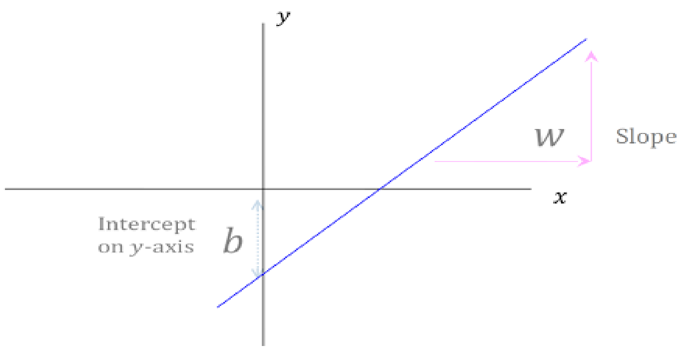


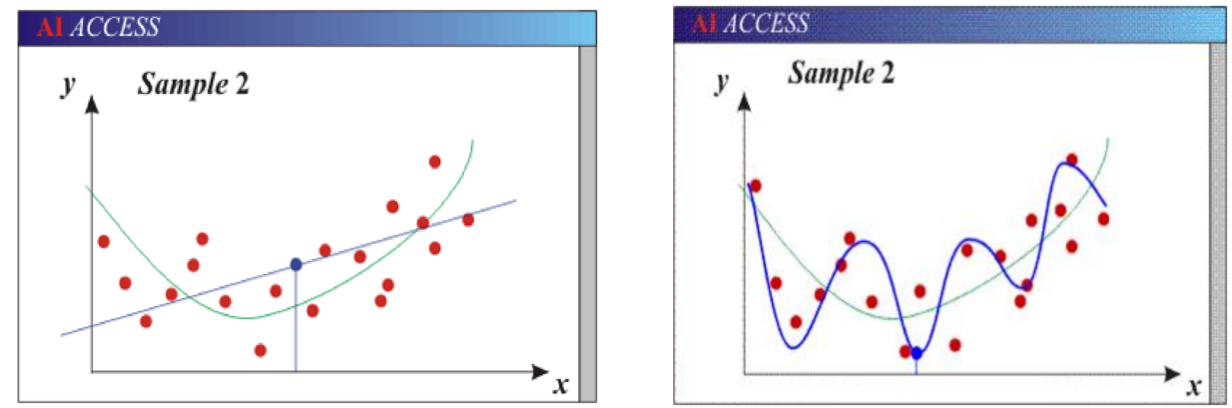
# Recap of the unit ...




$$L(y_i, \mathbf{x}_i^T\mathbf{w}) = (y_i - \mathbf{x}_i^T\mathbf{w})^2$$

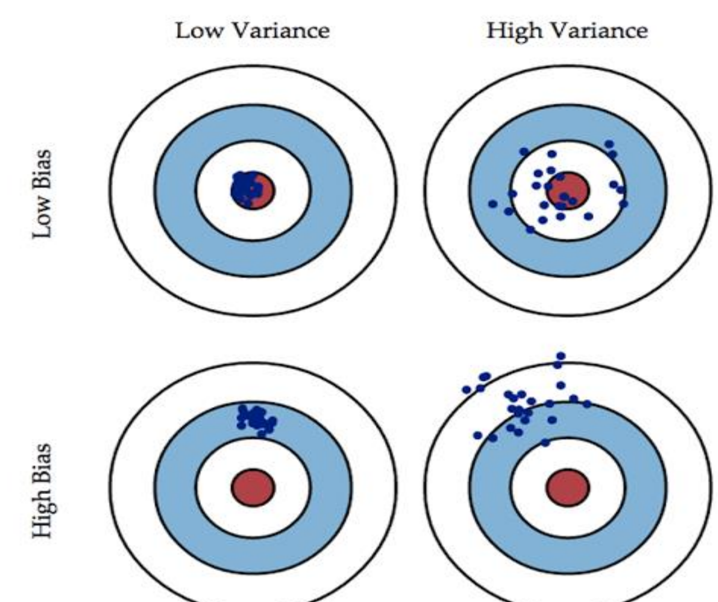
$$L(y_i, \mathbf{x}_i^T\mathbf{w}) = log(1 + exp(-y_i\mathbf{x}_i^T\mathbf{w}))$$



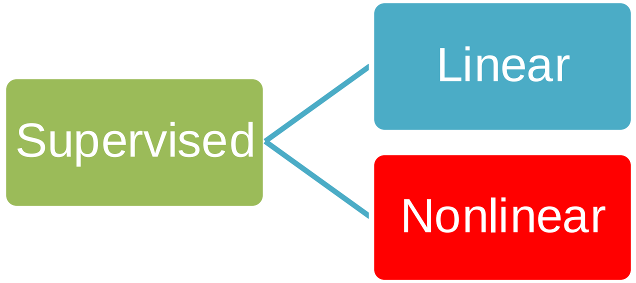

$$\min_w \frac{1}{n}\sum_i L(y_i, \mathbf{x}_i^T\mathbf{w}) + \lambda_1||\mathbf{w}||_1$$

$$\min_w \frac{1}{n}\sum_i L(y_i, \mathbf{x}_i^T\mathbf{w}) + \lambda_2||\mathbf{w}||_2^2$$

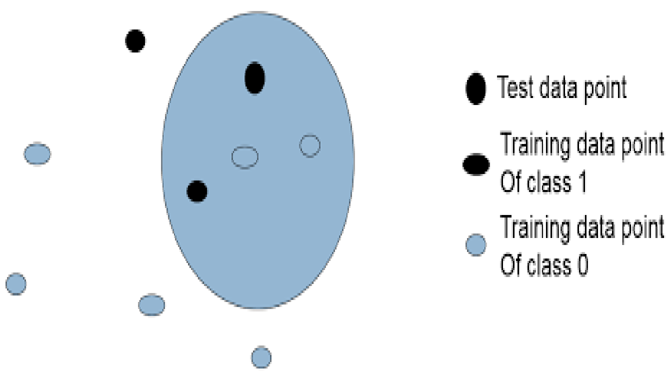
$$\min_w \frac{1}{n}\sum_i L(y_i, \mathbf{x}_i^T\mathbf{w}) + \lambda_1||\mathbf{w}||_1 + \lambda_2||\mathbf{w}||_2^2$$



# Recap of the unit …

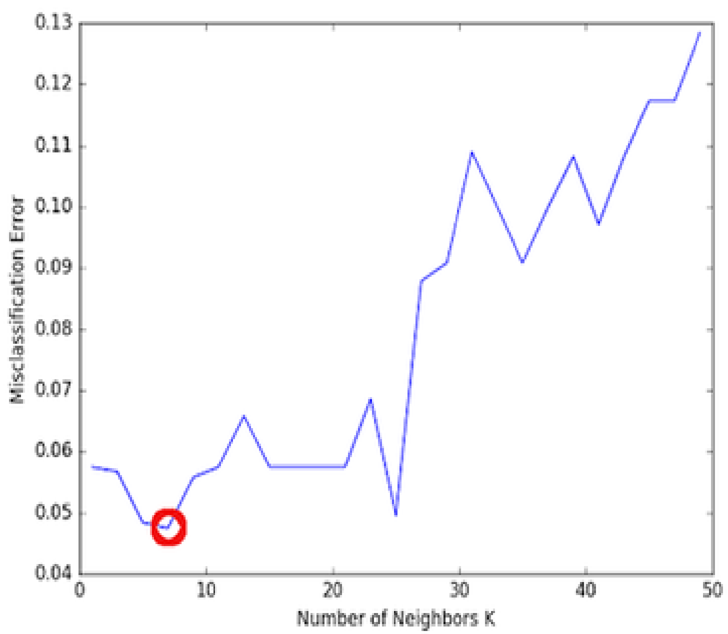



K Nearest Neighbor

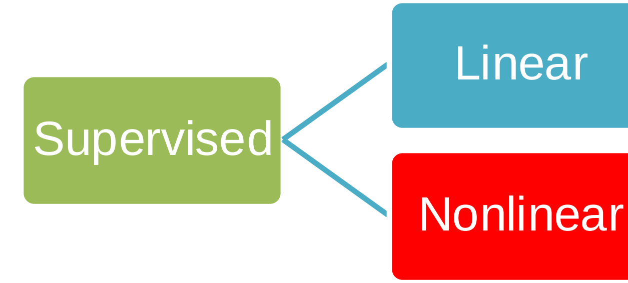



- Small values of K
  - Restrains the region of a given prediction
  - Forces classifier to be more focused on the close regions and neighbours
    - This will result in a low bias and high variance
- Higher values of K
  - Asking for more information from distant training points
  - Smoother decision boundaries
    - Lower variance but increases bias

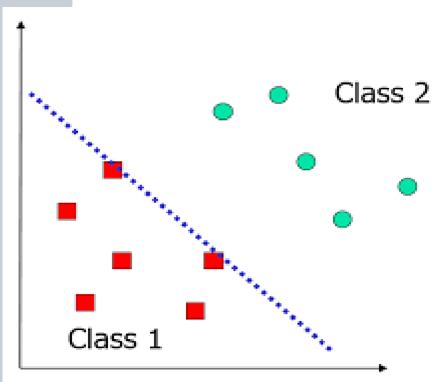
- Finding the best K
  - There is no rule of thumb in selecting $K_{max}$ since it depends on your desired rate of exploration for K
  - A simple and handy method
    - Cross-validation to partition your data into test and training samples
    - Evaluate model with different ranges of K values
    - The misclassification error can be used as a measurement of performance

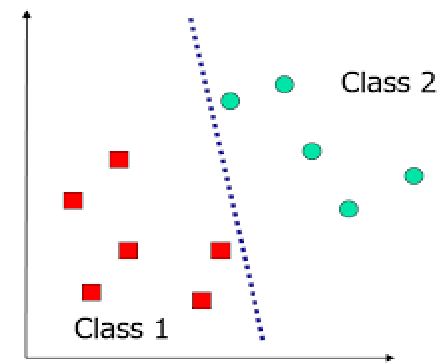


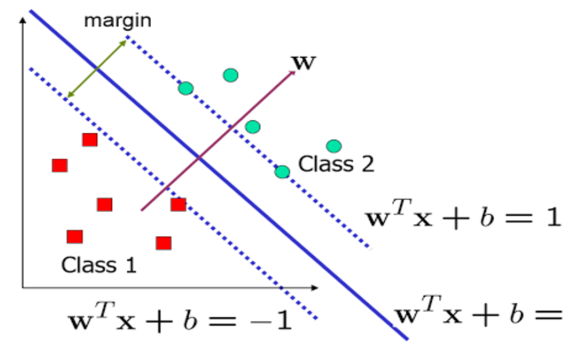
# Recap of the unit …

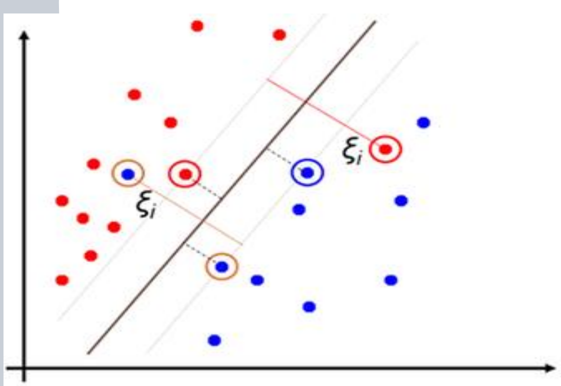



$$\min_\alpha \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{i=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$
$$\text{subject to: } \sum_{i=1}^{n} \alpha_i y_i = 0 \;\; and \;\; \alpha_i \geq 0 \;\; for\ all\ i$$

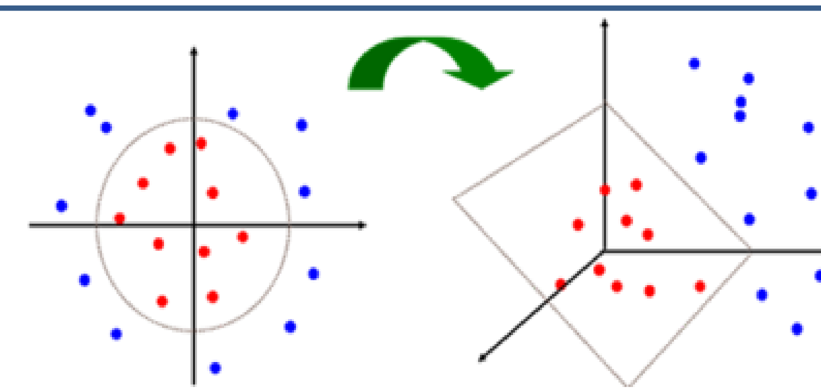



- Nonlinearly separable data
- Kernel

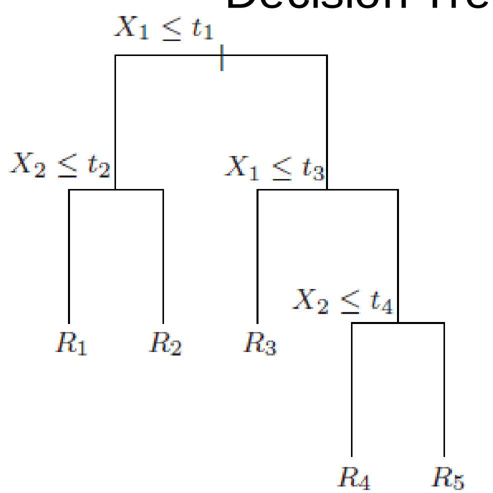


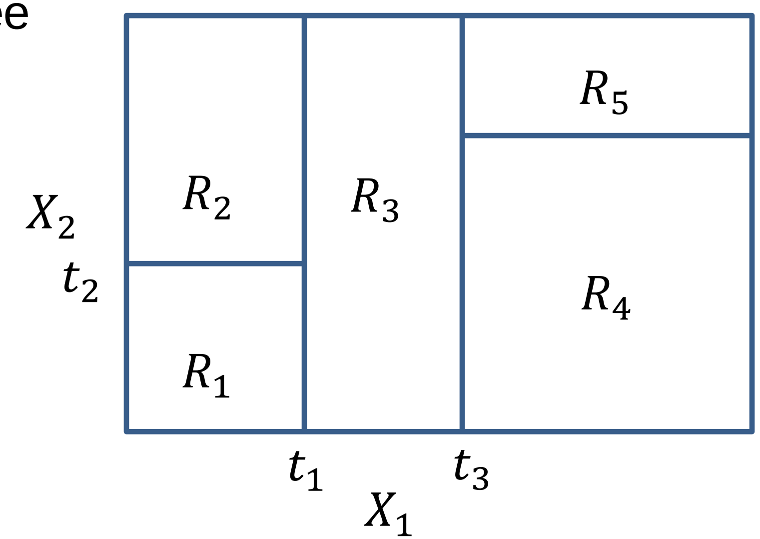
$$\min_\alpha \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{i=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \Leftrightarrow \boxed{K(X_i, X_j)}$$
$$Subject\ to: \sum_{i=1}^{n} \alpha_i y_i = 0 \;\; and \;\; 0 \leq \alpha_i \leq C \;\; \forall i$$

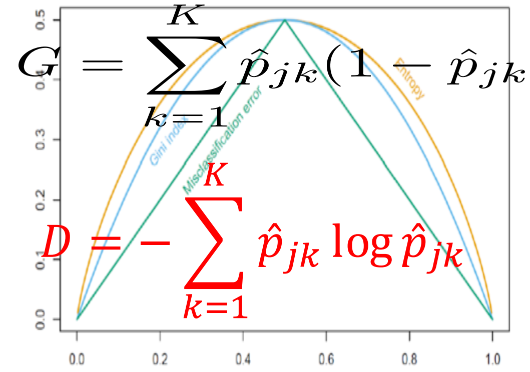
# Recap of the unit ...

Supervised
- Linear
- Nonlinear

Decision Tree



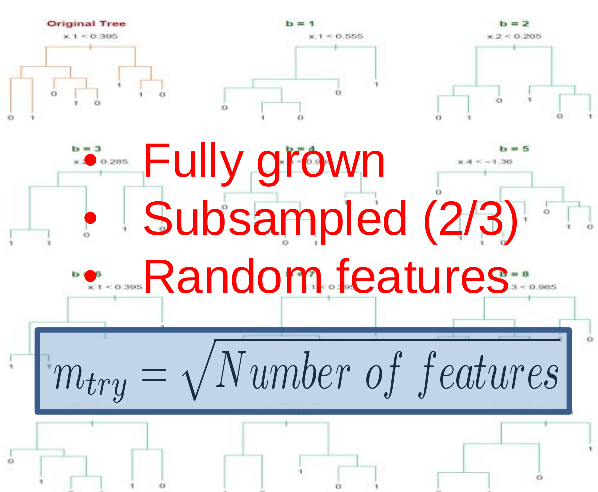




$$G = \sum_{k=1}^{K} \hat{p}_{jk}(1 - \hat{p}_{jk})$$

$$D = -\sum_{k=1}^{K} \hat{p}_{jk} \log \hat{p}_{jk}$$

- Simple
- Explainable
- Overfitting

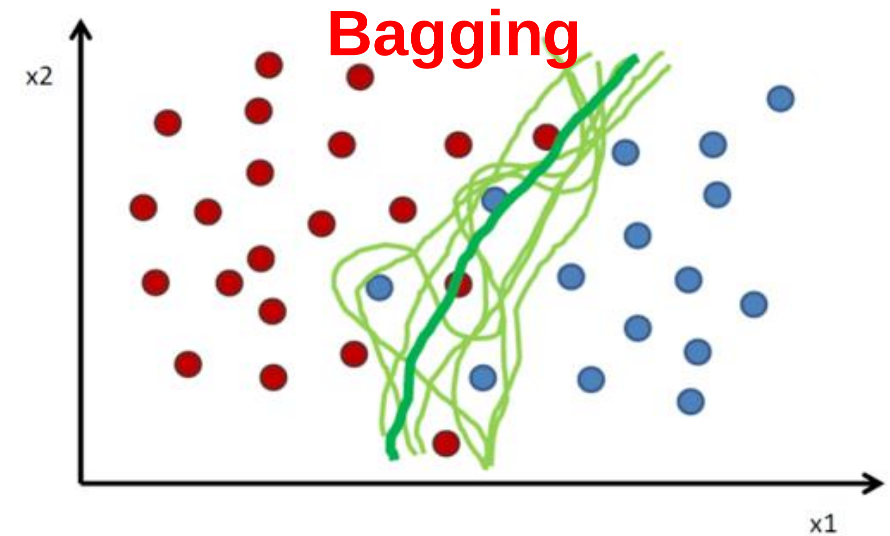
- Pruning
  - Pre-pruning
  - Post-pruning




- Fully grown
- Subsampled (2/3)
- Random features

$$m_{try} = \sqrt{Number\ of\ features}$$

**Power of collective knowledge**




**Bagging**

# Recap of the unit ...

Supervised
- Linear
- Nonlinear

| | |
|---|---|
| Hyperparameter | Parameter |
| Performance measurement | Data splitting |
| Overfitting | Model generalisation |

# Recap of the unit …

Supervised
- Linear
- Nonlinear

## Design questions

### Which ML model to use?

- Sample size
- Class balance
- Dimensionality of the data
- Application environment
- Realtime or asynchronous

### How to optimize the model?

- Identify hyperparameters (especially sensitive ones)
- Do not try to use any rule of thumb
- Do not use the test data in the model selection process
- Select the best model based on cross-validation results
- Generate multiple test performance and report the aggregated results

# eVALUate is n

This is your chance

Confidential (we take this very seriously)

Any device, anytime, anywhere

Your opportunity to
impact student experience

Important part of
Deakin's learning and teaching

DEAKIN UNIVERSITY

From: Monday 23rd September 2023
Ends: Sunday 27th October 2023

# Giving feedback on eVALUate

DEAKIN UNIVERSITY

## DO

~ Be polite and respectful

~ Be human and considerate

~ Comment on specific issues

~ Be objective

~ Focus on the issue not personal traits

~ Aim for balance about what was helpful and what you would like to see improve.

## DON'T

! Make it personal

! Be judgemental and insulting

! Use derogatory, sexist or racist language

! Go on a massive rant

! Be a troll

# Thank You.