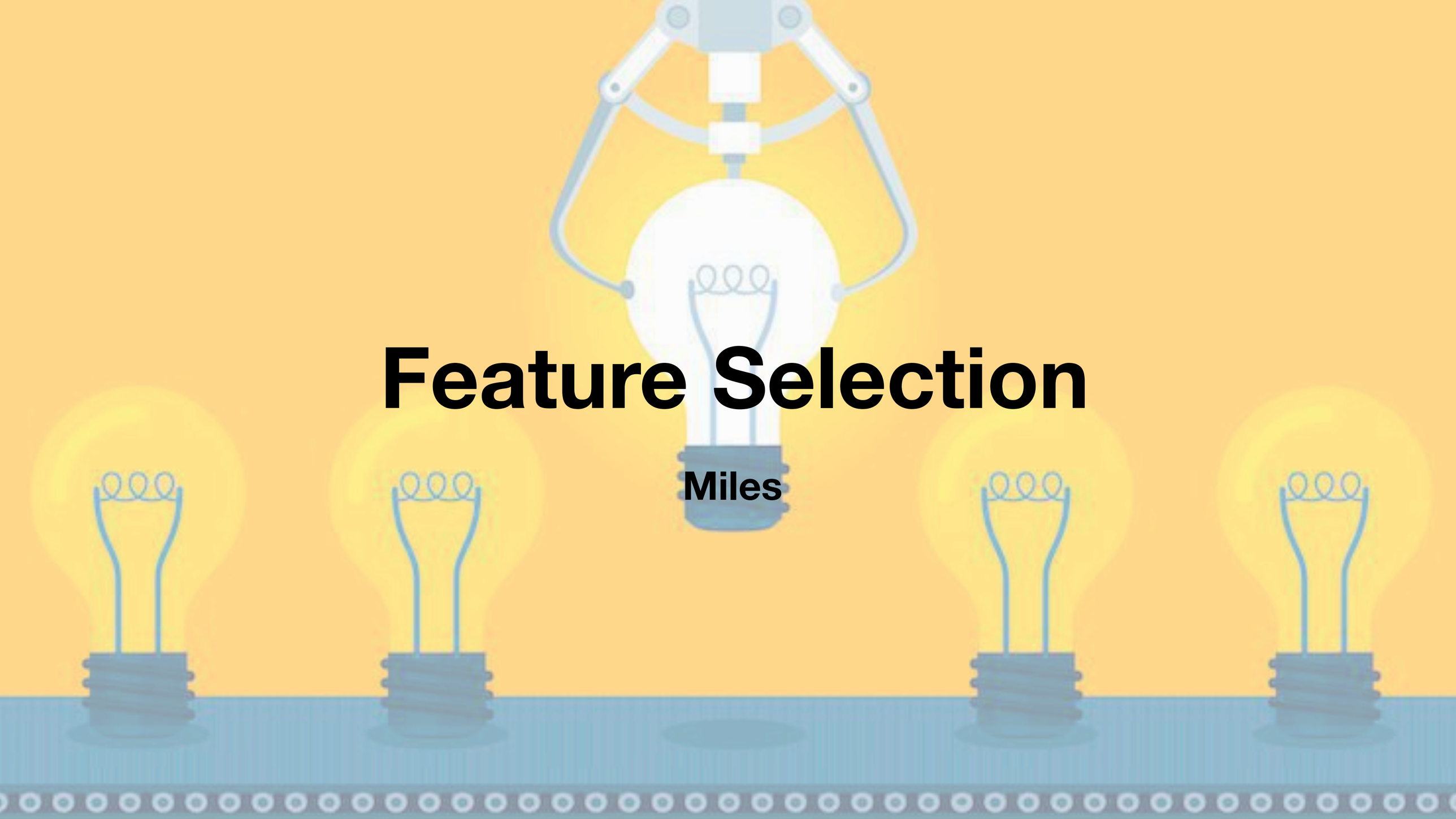


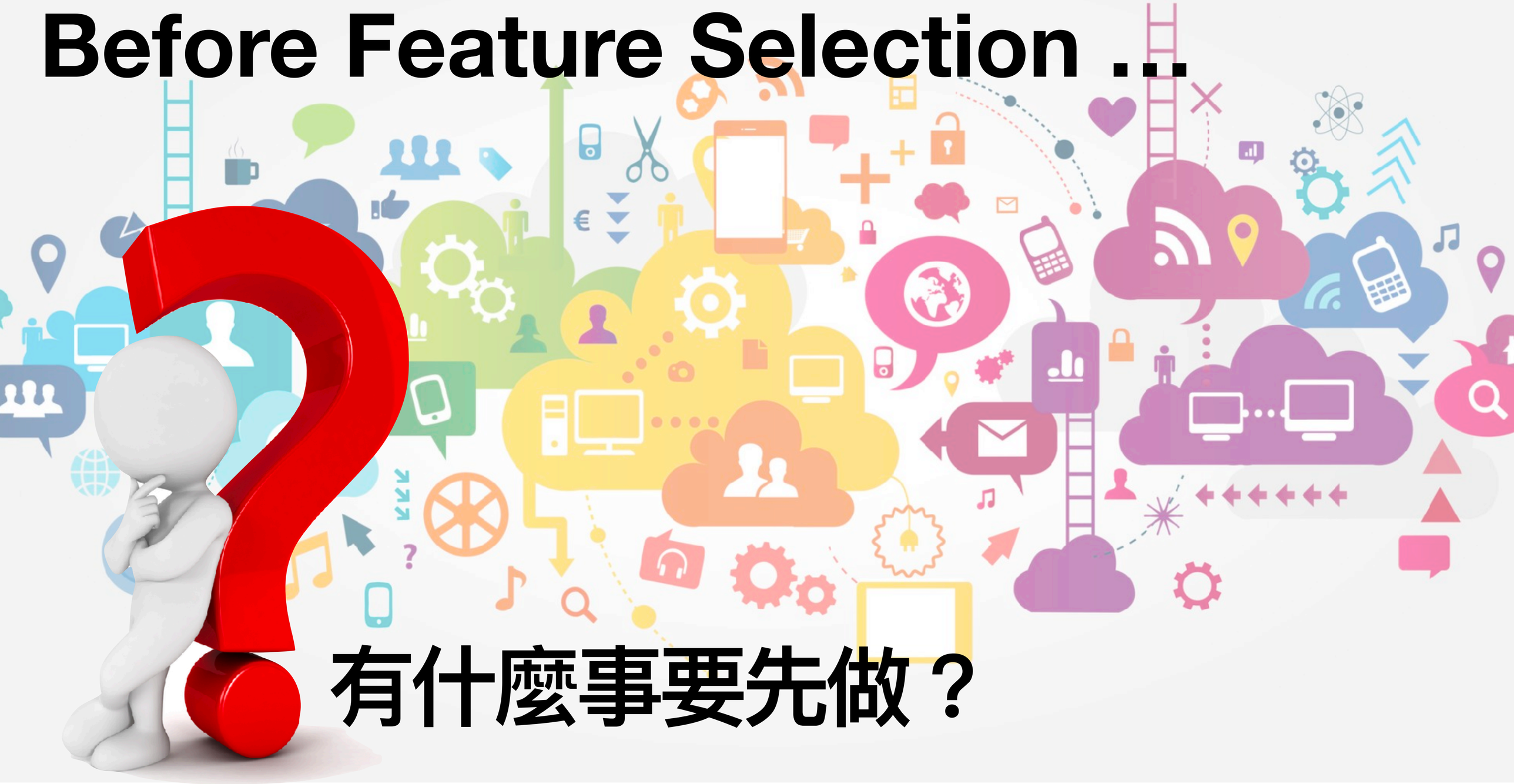
Feature Selection

Miles

WHY Do We Need Feature Selection?

- 建立模型的過程**更有效率**
- 改善**模型的品質**:吵雜或多餘的資料會使有意義的資料模式更難發現
- 改善通用性、**降低過擬合**

Before Feature Selection ...



有什麼事要先做？

確認需要回答的問題

會不會辦長榮聯名卡？

是否有旅平險需求？

有沒有信貸需求？

是否有開證券戶需求？

是否為高風險客戶？

針對特定的問題，對描述一個事物所需的信息

進行篩選幾乎是必不可少的過程

Feature Selection

數據包含許多**冗餘 或無關** 的特徵，因而移除這些特徵並不會導致丟失信息

- 統計**特徵與目標的相關性**
- 計算**特徵對目標的影響力**
- 留下與目標最相近的特徵，使判斷準確率能夠提升。

How?

Complete Feature Set

Generation

Subset

Evaluation

Goodness of the Subset

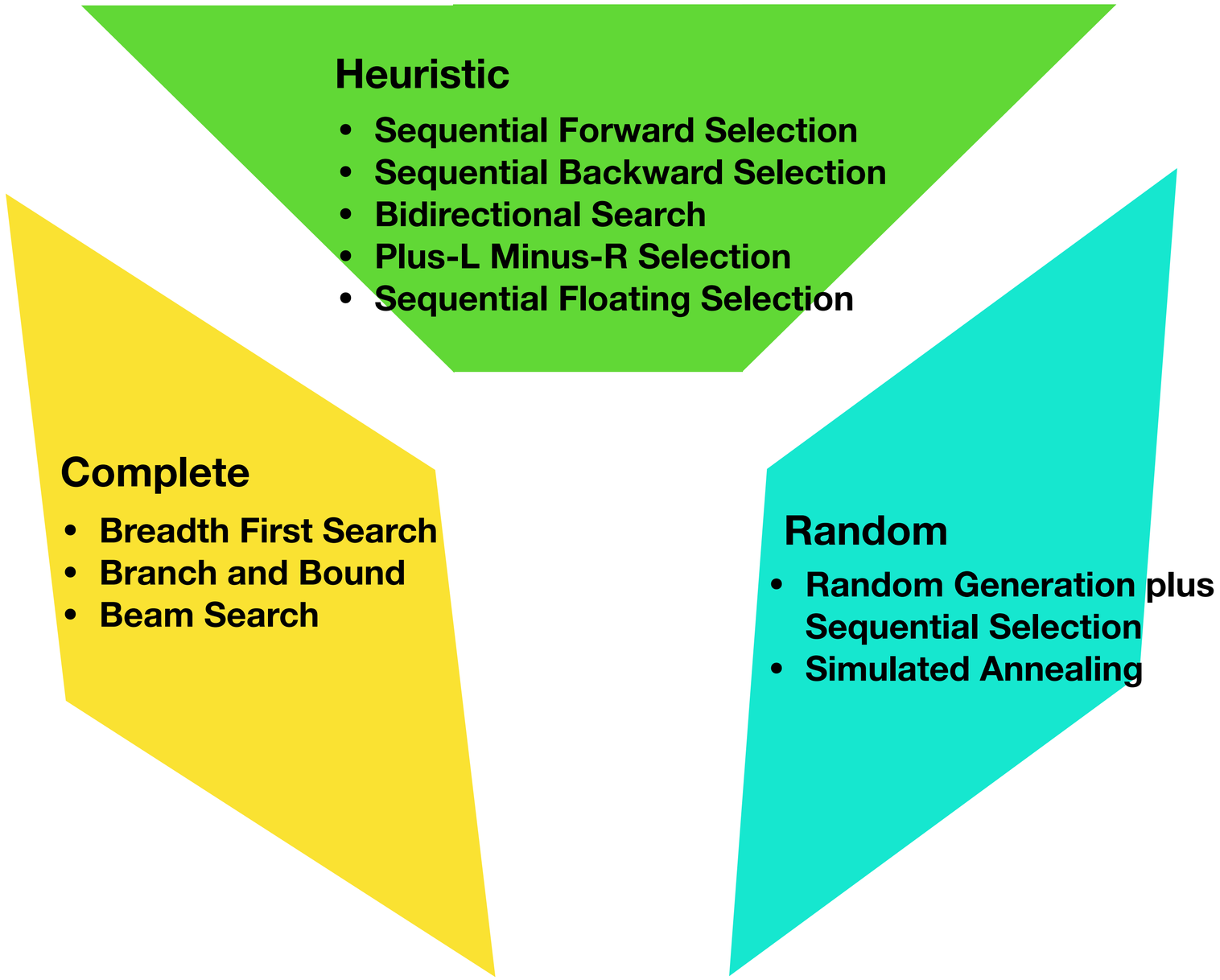
Stopping Criterion

No

Yes

Validation

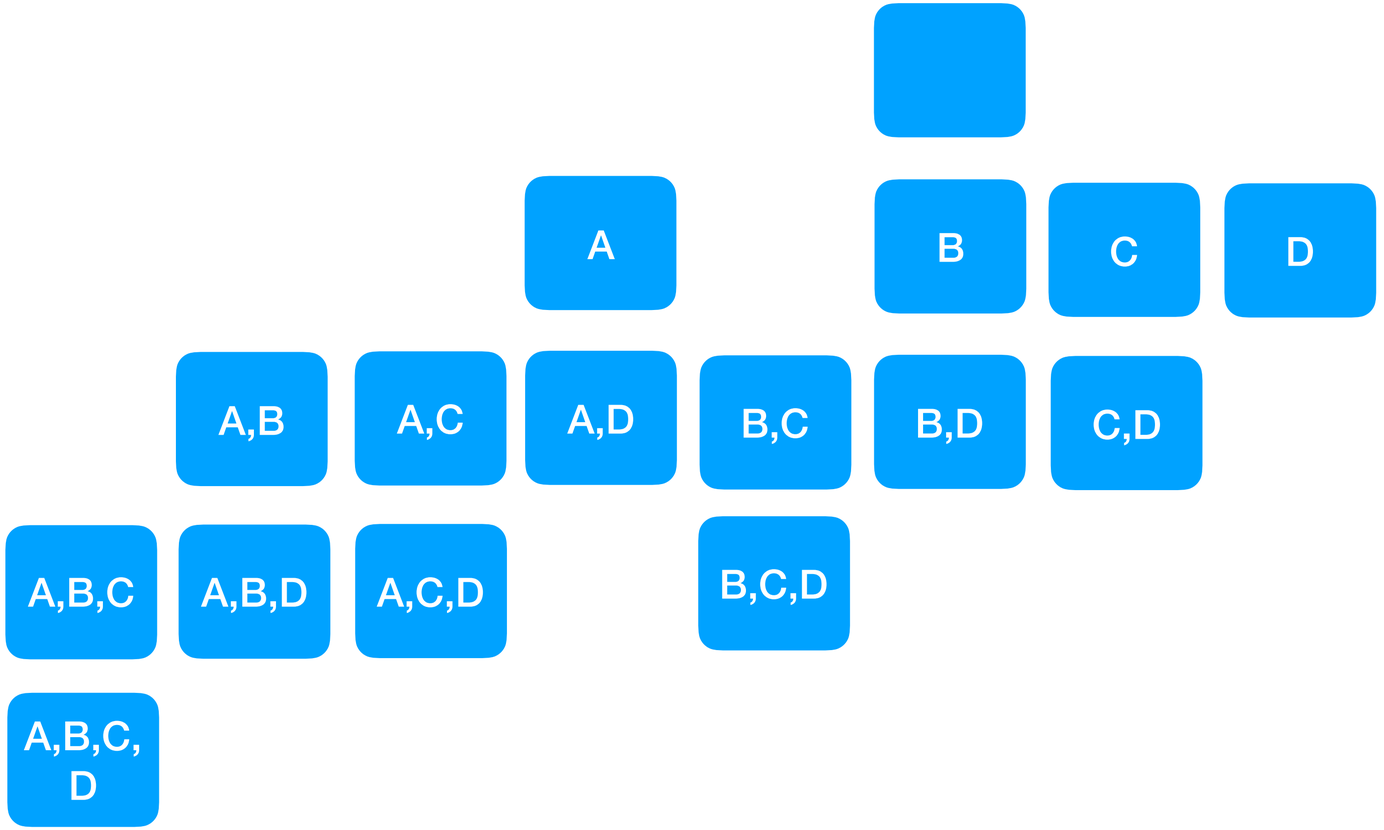
Generation


Heuristic
• Sequential Forward Selection
• Sequential Backward Selection
• Bidirectional Search
• Plus-L Minus-R Selection
• Sequential Floating Selection
Complete
• Breadth First Search
• Branch and Bound
• Beam Search
Random
• Random Generation plus Sequential Selection
• Simulated Annealing

Complete

Breadth First Search

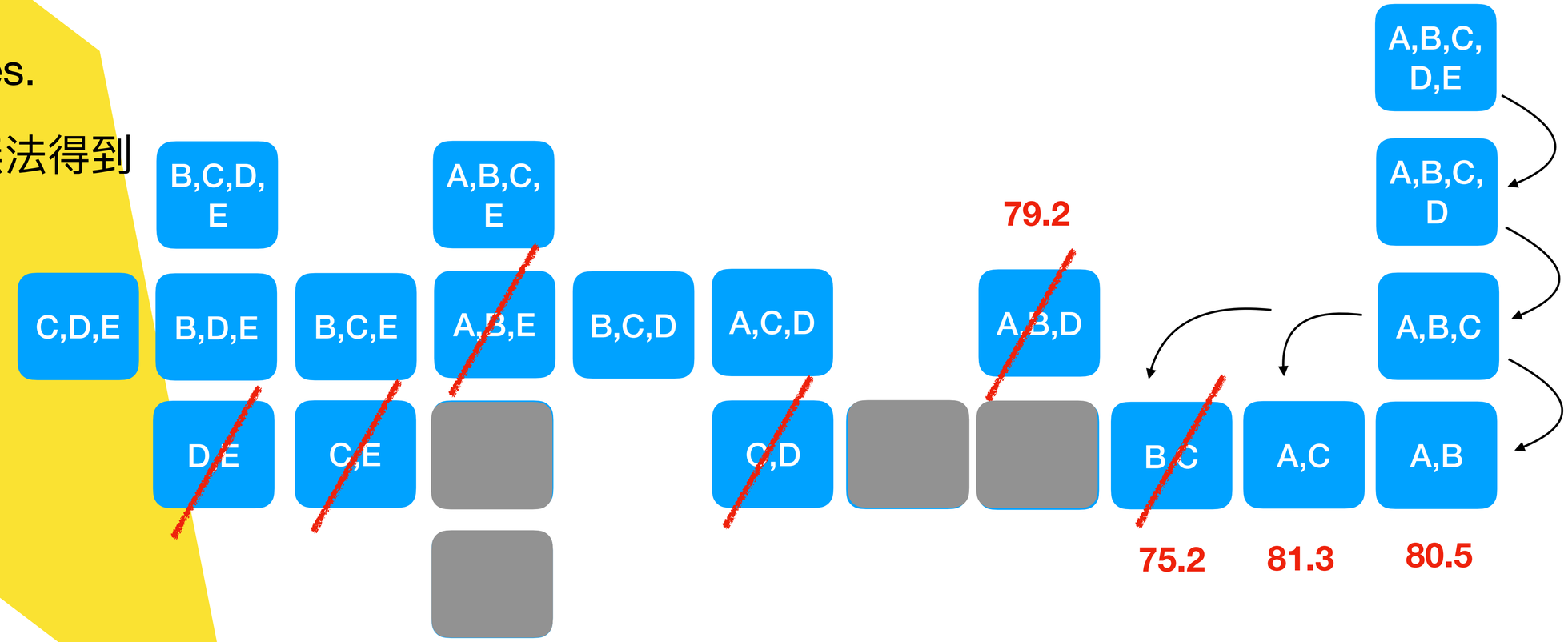
- 列舉所有特徵組合



Complete

Branch and Bound

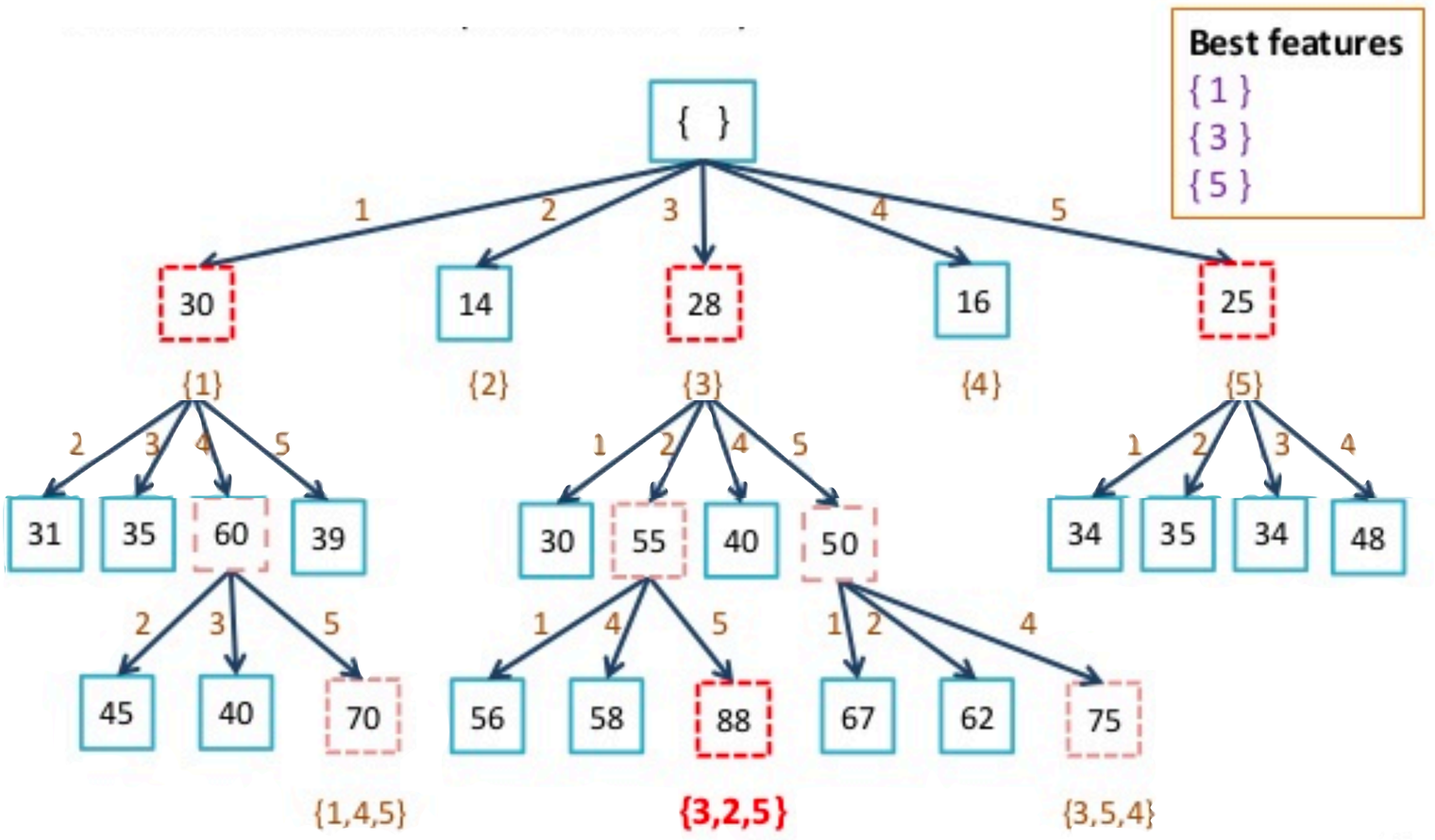
- Set the best **K** features.
- **砍掉分支**，若判斷無法得到更好的結果。



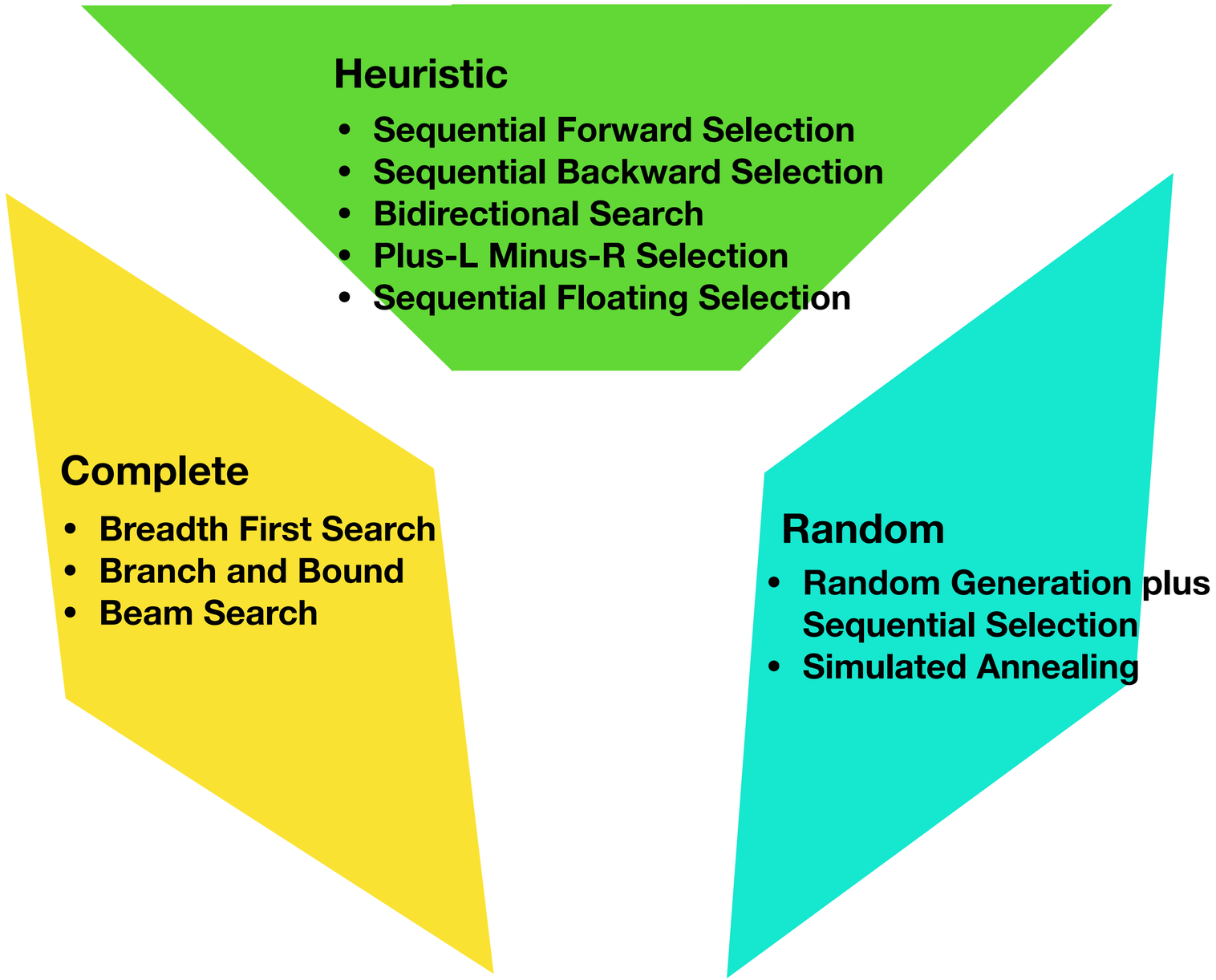
Complete

Beam Search

- Choose a value for **Beam width(K).**
- **Cut of value** can be checked before select.



Generation


Heuristic
• Sequential Forward Selection
• Sequential Backward Selection
• Bidirectional Search
• Plus-L Minus-R Selection
• Sequential Floating Selection
Complete
• Breadth First Search
• Branch and Bound
• Beam Search
Random
• Random Generation plus Sequential Selection
• Simulated Annealing

Heuristic

- **Sequential Forward Selection**

  {1}, {2}, {3}, {4}, {3,1}, {3,2}, {3,4}, {3,4,1}, {3,4,2}, {3,4,1,2}

- **Sequential Backward Selection**

  {1, 2, 3, 4}, {2, 3, 4}, {1, 3, 4}, {1, 2, 4}, {1, 2, 3}, {2, 3}, {1, 3}, {1, 4}, {3}, {1}

- **Bidirectional Search**

  使用SFS從空集開始，同時使用SBS從全集開始搜索，當兩者搜索到一個**相同的特徵子集**C時停止搜索。

Heuristic

- **Plus-L Minus-R Selection**

該算法有兩種形式:

<1> 算法從空集開始，每輪**先加入L個特徵**，然後從中**去除R個特徵**，使得評價函數值最優。(L> R)

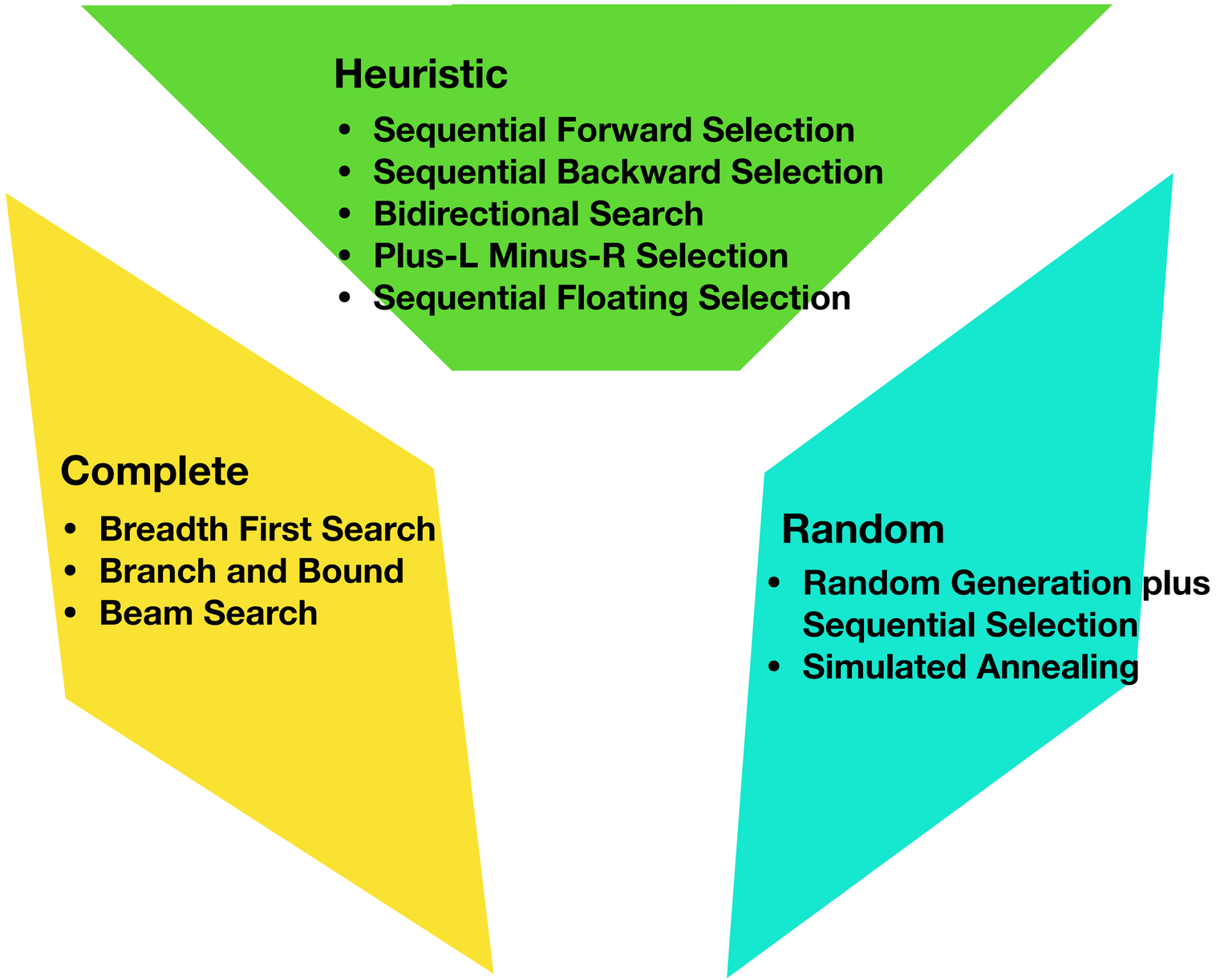
<2> 算法從全集開始，每輪**先去除R個特徵**，然後**加入L個特徵**，使得評價函數值最優。(L< R)

- **Sequential Floating Selection**

由增L去R選擇算法發展而來，該算法與增L去R選擇算法的不同之處在於：

L與R不是固定的，而是**浮動**的。

Generation


Heuristic
• Sequential Forward Selection
• Sequential Backward Selection
• Bidirectional Search
• Plus-L Minus-R Selection
• Sequential Floating Selection
Complete
• Breadth First Search
• Branch and Bound
• Beam Search
Random
• Random Generation plus Sequential Selection
• Simulated Annealing

Random

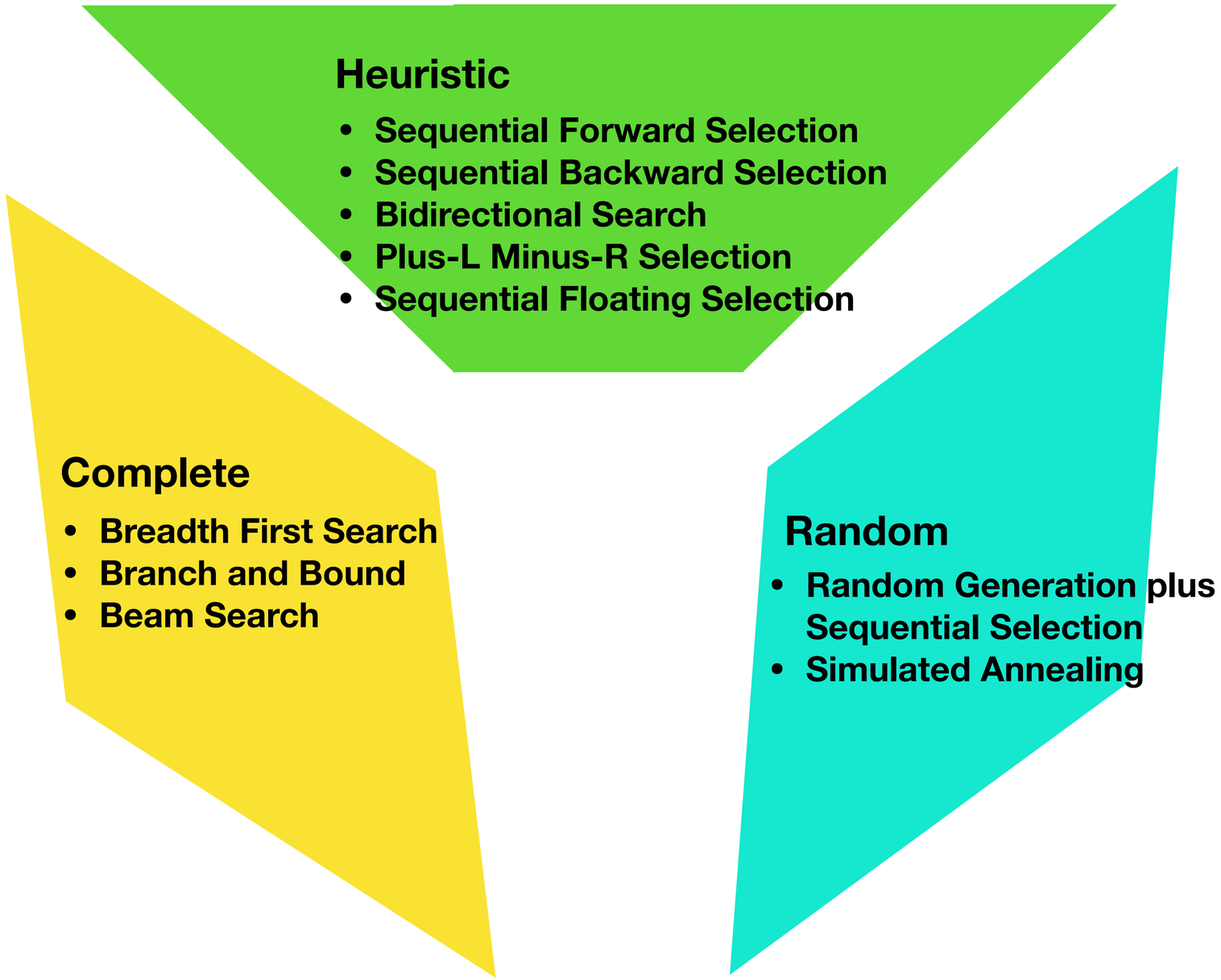
- **Random Generation plus Sequential Selection**

  隨機產生一個特徵子集，然後在該子集上執行SFS與SBS算法。

- **Simulated Annealing**

  以**一定的概率**来接受一個比當前解要差的解，因此**有可能**会跳出這個局部的最優解，達到全局的最優解

Generation


Heuristic
• Sequential Forward Selection
• Sequential Backward Selection
• Bidirectional Search
• Plus-L Minus-R Selection
• Sequential Floating Selection
Complete
• Breadth First Search
• Branch and Bound
• Beam Search
Random
• Random Generation plus Sequential Selection
• Simulated Annealing

Heuristic

- 不考慮特徵之間的相關性
- Total Feature數量多時適用，節省時間
- 易陷入局部最佳解

Complete

- 考慮特徵之間的相關性，從而能更好地找到最優集合
- 耗費時間長
- Total Feature數量少時適用

Random

- 用於跳出局部最優值
- 依賴於隨機因素，有實驗結果難以重現

How?

Complete Feature Set

Generation

Subset

Evaluation

Goodness of the Subset

Stopping Criterion

No

Yes

Validation

Evaluation

- **Correlation**

  好的特征子集所包含的特征應該是與**分類**的相關度**較高**，而**特征間**相關度**較低**

  常使用線性相關係數(correlation coefficient) 來衡量向量之間線性相關度。

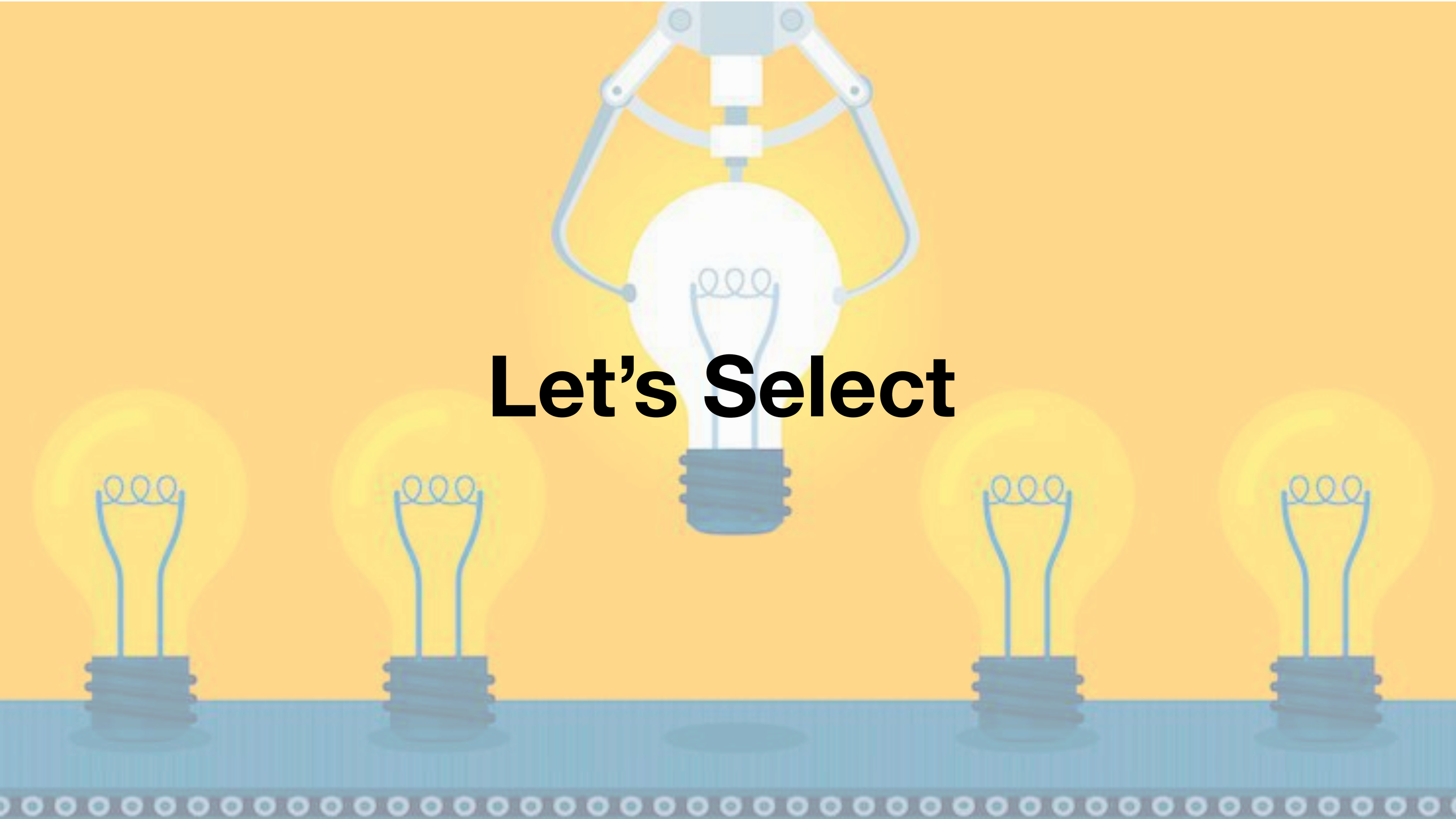
- **Distance Metrics**

  好的特徵子集應該使得屬於**同一類的樣本**距離儘可能**近**，屬於**不同類的樣本**之間的距離儘可能**遠**。

  常用的距離度量包括歐氏距離、曼哈頓距離等

- **Classifier Error Rate**

  使用特定的分類器，用給定的特徵子集對樣本集進行分類，用**分類的精度**來衡量特徵子集的好壞。


Let’s Select