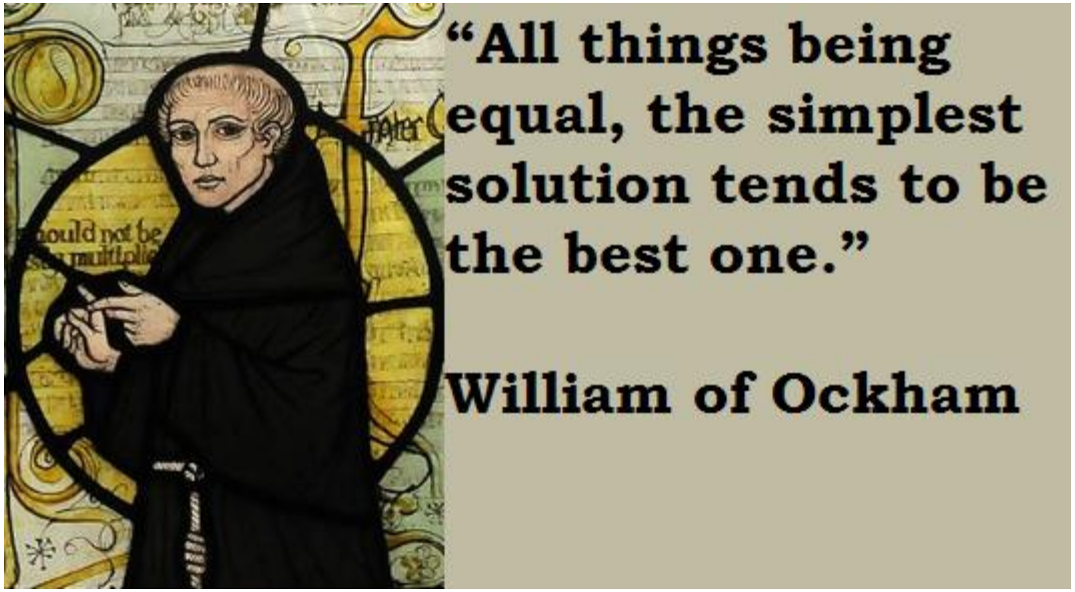


Dimensionality Reduction

Pilsung Kang
School of Industrial Management Engineering
Korea University

AGENDA

01 Dimensionality Reduction

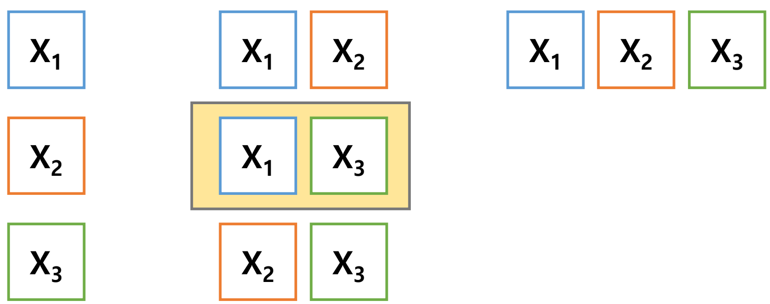
02 Variable Selection Methods

03 Shrinkage Methods

04 R Exercise

Exhaustive Search

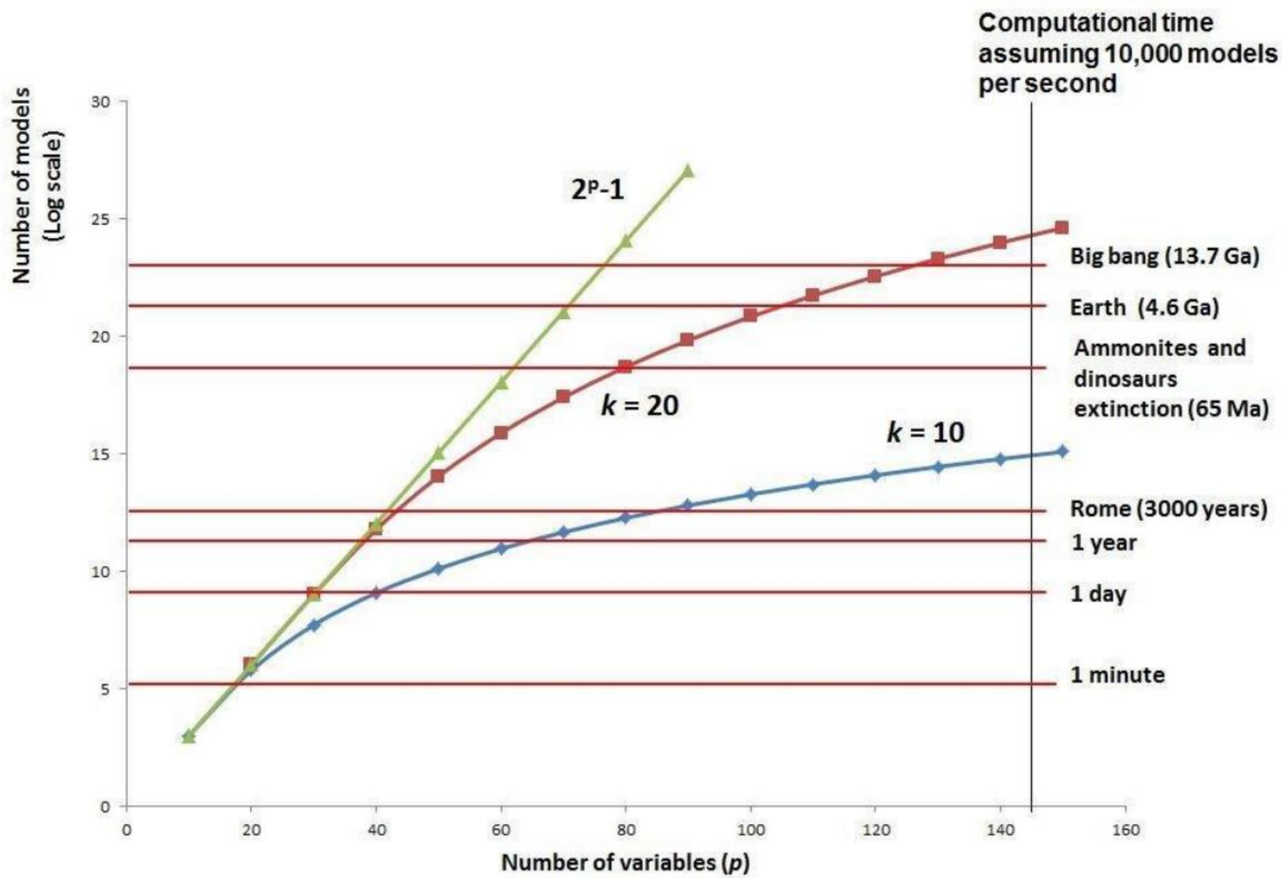
- Exhaustive search
  - ✓ Search all possible combinations
    - Ex) 3 variables X_1 X_2 X_3
    - A total of 7 possible subsets are tested



  - ✓ Performance criteria for variable selection
    - Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), Adjusted R^2, Mallow's C_p, etc.

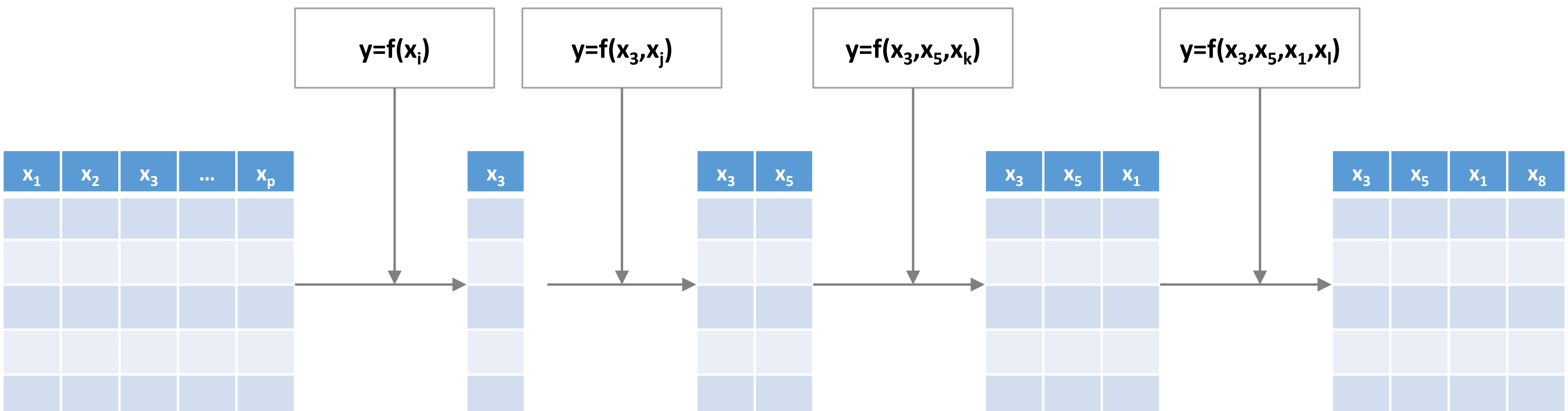
Exhaustive Search

- Exhaustive search
  - ✓ Assume that we have a computer that can evaluate 10,000 models/second



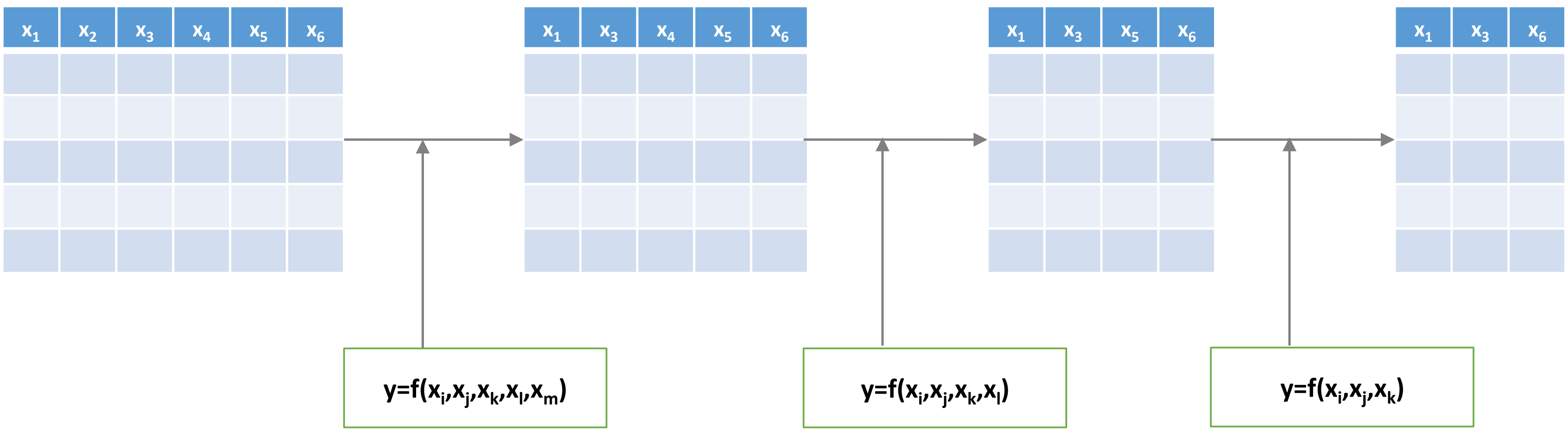
Forward Selection

- Forward selection
  - ✓ From the model with no variable, significant variables are sequentially added
  - ✓ Once the variable is selected, it will never be removed (The number of variables gradually increases)



Backward Elimination

- Backward Elimination
  - ✓ From the model with all variables, irrelevant variables are sequentially removed
  - ✓ Once a variable is removed, it will never be selected (The number of variables gradually decreases)

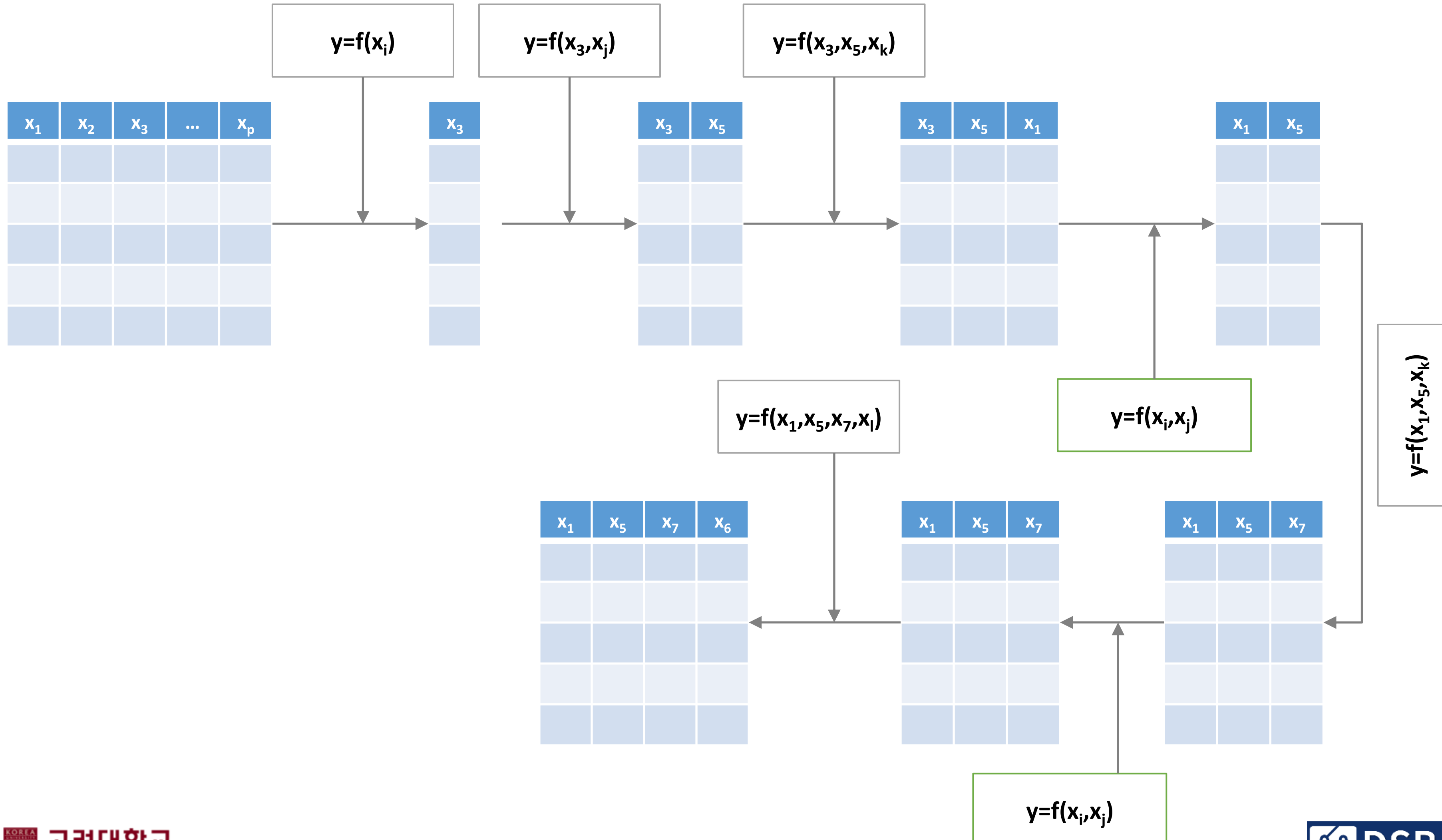


Stepwise Selection

- Stepwise Selection
  - ✓ From the model with no variable, conduct the forward selection and backward elimination alternately
  - ✓ Takes longer time than forward selection/backward elimination, but has more chances to find the optimal set of variables
  - ✓ Variables that is either selected/removed can be reconsidered for selection/removal
  - ✓ The number of variables increases in the early period, but it can either increase or decrease

Stepwise Selection

- Stepwise selection example



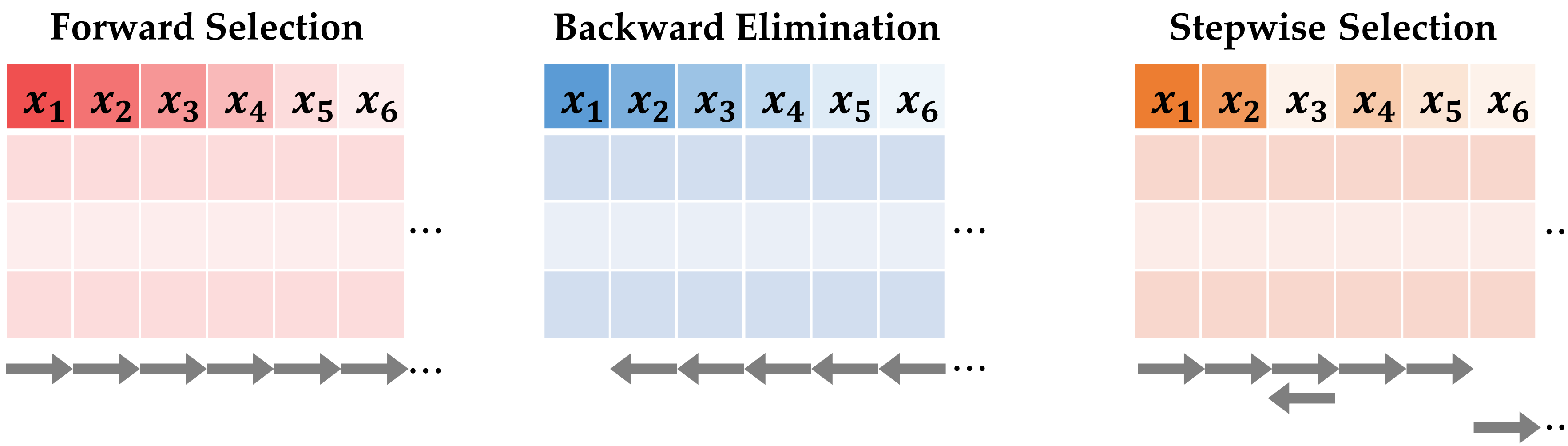
Stepwise Selection

- Stepwise Selection
  - ✓ Stepwise selection process

- ▶ Start with model with no predictors.
- ▶ Add variable with largest F-statistic (provided P less than some cut-off).
- ▶ Refit with this variable added. Recompute all F statistics for adding one of the remaining variables and add variable with largest F statistic.
- ▶ At each step after adding a variable try to eliminate any variable not significant at some level (that is, do BACKWARD elimination till that stops).
- ▶ After doing the backwards steps take another FORWARD step.
- ▶ Continue until every remaining variable is significant at cut-off level and every excluded variable is insignificant OR until variable to be added is same as last deleted variable.

Comparison among FS/BE/SS

- Illustrative Example



Performance Metrics

- Akaike Information Criteria (AIC)
  - ✓ Sum of squared error (SSE) with the number of variables as a penalty term

$$AIC = n \cdot ln\left(\frac{SSE}{n}\right) + 2k$$

- Bayesian Information Criteria (BIC)
  - ✓ SSE, number of variables, standard deviation obtained by the model with all variables

$$BIC = n \cdot ln\left(\frac{SSE}{n}\right) + \frac{2(k+2)n\sigma^2}{SSE} - \frac{2n^2\sigma^4}{SSE^2}$$

Performance Metrics

- Adjusted R^2
  - ✓ Simple R^2 increases when the number of variable increases

$$\text{Model 1}: \quad y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon$$
$$\text{Model 2}: \quad y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \ldots + \beta_{k+m} x_{k+m} \epsilon$$

$$R^2(M2) \geq R^2(M1)$$

  - ✓ Use the adjusted R^2 that account for the number of variables (k)

$$\text{Adjusted } R^2 = 1 - \left(\frac{n-1}{n-k-1}\right)(1 - R^2) = 1 - \frac{n-1}{n-k-1}\frac{SSE}{SST}$$

Genetic Algorithm

Siedlecki, W. Sklansky (1989)

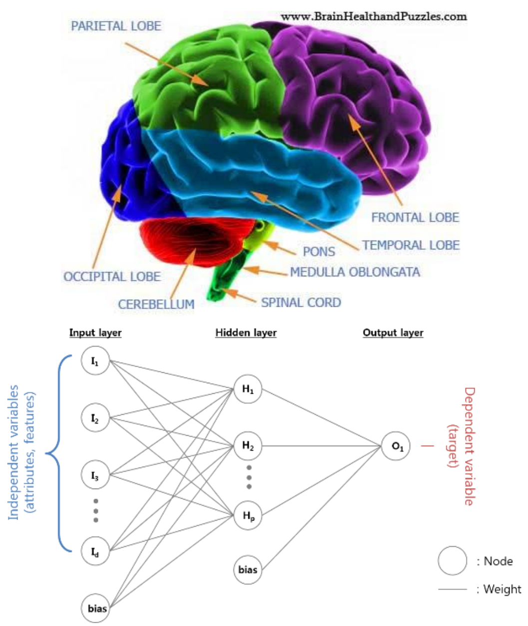
- Limitations of the previous variable selection methods
  - ✓ Exhaustive search: guarantee the optimal subset, but takes too long time (practically impossible for many tasks)
  - ✓ Local search (forward/backward/stepwise): efficient search but the search space is very limited, which leads to a low probability of finding the optimal solution

- Idea
  - ✓ Improve the performance of local searches with a little additional computational time!

Genetic Algorithm

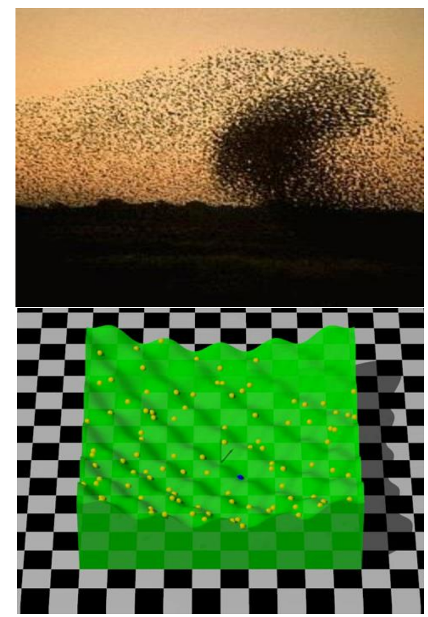
- Meta-Heuristic Approach
  - ✓ Solve a complex problem by doing trials and errors efficiently
  - ✓ Among the optimization algorithms, many of them mimic the way of a natural system works

Artificial Neural Networks



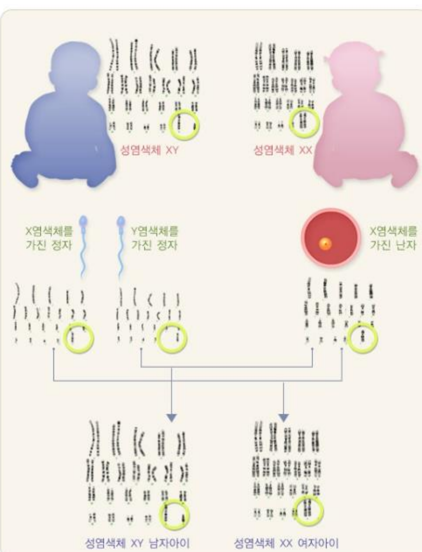
Ant Colony Algorithm

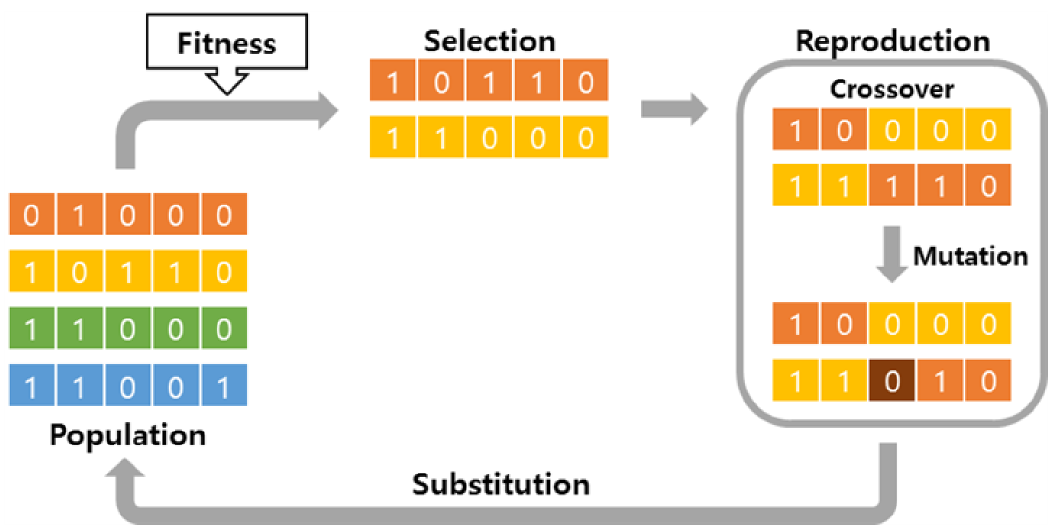
Particle Swarm Optimization



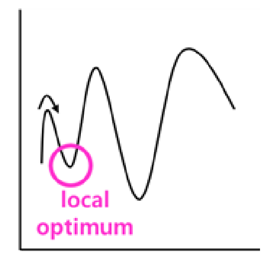
Genetic Algorithm

- An Evolutionary Algorithm that mimics the Reproduction of Creatures
  - ✓ Find a superior solutions and preserve by repeating the reproduction process
    - Selection: Select a superior solution to improve the quality
    - Crossover: Search various alternatives based on the current solutions
    - Mutation: Give a chance to escape the local optima

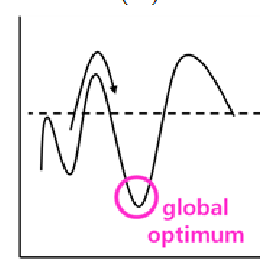




(a)



(b)



(c)

Genetic Algorithm

- Genetic Algorithm for Feature Selection

Step 1: Initialize Chromosomes

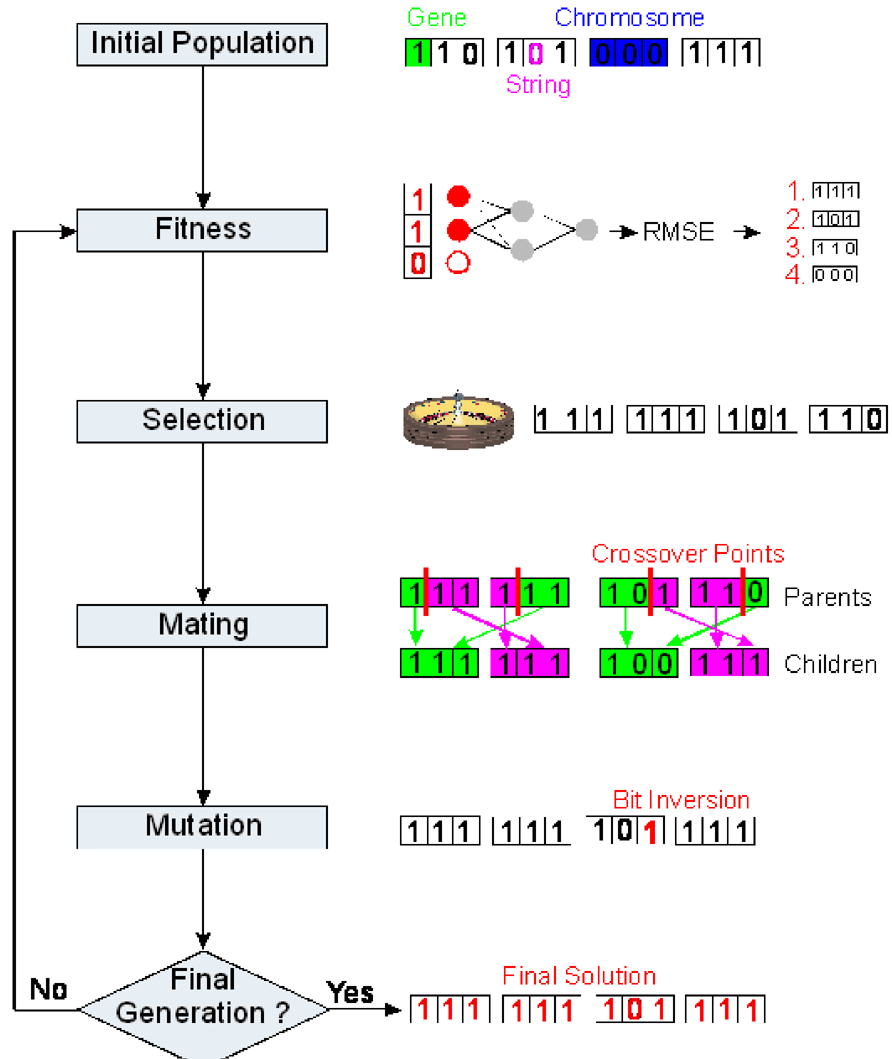
Step 2:
Model training based on chromosomes

Step 3:
Fitness evaluation

Step 4:
Selection good chromosomes

Step 5:
Create next generation: Crossover & Mutation

Step 6:
Select the final variable set



GA Step 1: Initialization

- Encoding Chromosomes
  - ✓ Genetic algorithm can be used not only for variable selection, but for a wide range of optimization problems
  - ✓ Encoding scheme can be different for different tasks
  - ✓ Binary encoding is commonly used for variable selection

Chromosome

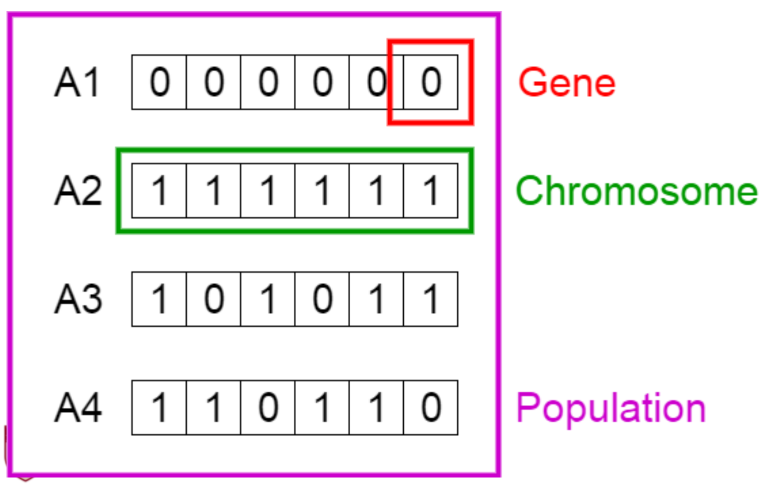
Gene

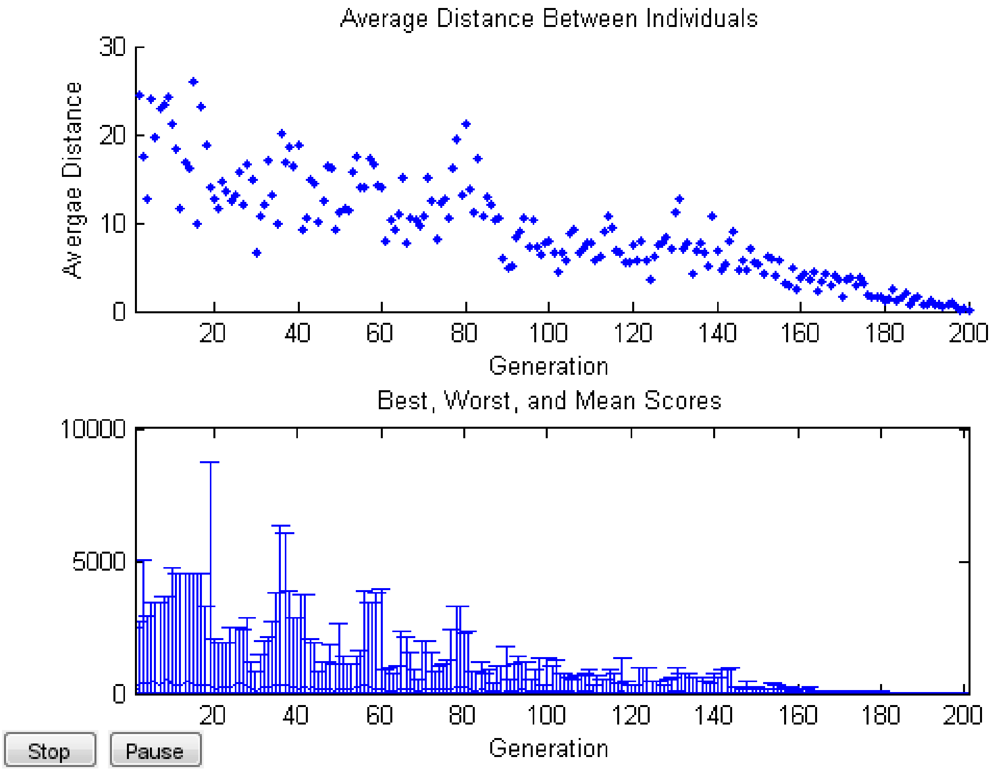
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	…	x_d
1	0	0	1	0	1	1	0	…	1

1: Use the corresponding variable in the modeling
0: Do not use the variable

GA Step 1: Initialization

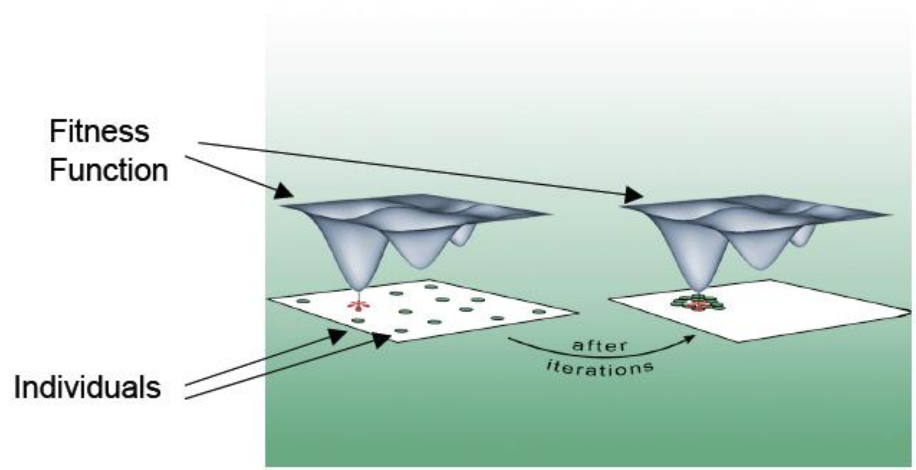
- Parameter Initialization
  - ✓ The number of chromosome (population)
  - ✓ Fitness function
  - ✓ Crossover mechanism
  - ✓ The rate of mutation
  - ✓ Stopping criteria
    - minimum fitness improvement
    - maximum iterations, etc.





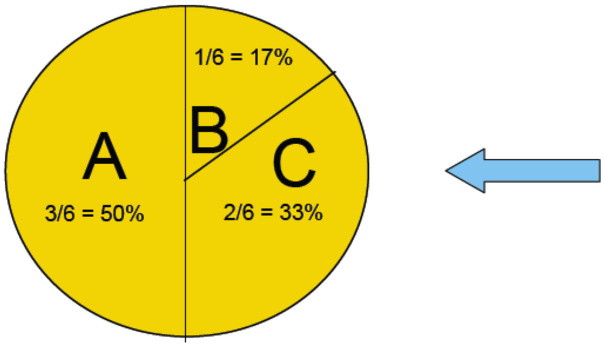
GA Step 3: Fitness Evaluation

- Fitness Function
  - ✓ A criterion that determines which chromosomes are better than others
  - ✓ In general, the higher the fitness value, the better the chromosomes
  - ✓ Common criteria that are embedded in the fitness function
    - If two chromosomes have the same fitness value, the one with fewer variables is preferred
    - If two chromosomes use the same number of variables, the one with higher predictive performance is preferred
  - ✓ In case of multiple linear regression
    - Adjusted R2
    - Akaike information criterion (AIC)
    - Bayesian information criterion (BIC)



GA Step 4: Selection

- Selection
  - ✓ Select superior chromosomes in the current population to reproduce the population of the next generation
  - ✓ Deterministic selection
    - Select only top N% of chromosomes
    - Bottom (100-N)% chromosomes are never selected
  - ✓ Probabilistic selection
    - Use the fitness value of each chromosome as the selection weight
    - All chromosomes can be selected with different probabilities



fitness(A) = 3

fitness(B) = 1

fitness(C) = 2

GA Step 5: Crossover & Mutation

- Crossover (Reproduction)
  - ✓ Two child chromosomes are produced from two parent chromosomes
  - ✓ The number of crossover points can vary from 1 to n (total number of genes)

Crossover point = 2

Crossover points = 3

Crossover points = N

Assume array: [0.35, 0.62, 0.18, 0.42, 0.83, 0.76, 0.39, 0.51, 0.36]

GA Step 5: Crossover & Mutation

- Mutation
  - ✓ Genetic operator used to maintain diversity from one generation of a population of chromosomes to the next
  - ✓ Alters one or more gene values in a chromosome from its initial state, which result in entirely new gene values being added to the gene pool
  - ✓ By mutation, the current solution can have a chance to escape from the local optima
  - ✓ A too mutation rate can increase the time to converge (0.01 can be a good choice)

Consider the two original off-springs selected for mutation.

Original offspring 1	1	1	0	1	1	1	1	0	0	0	0	1	1	1	1	0
Original offspring 2	1	1	0	1	1	0	0	1	0	0	1	1	0	1	1	0

Invert the value of the chosen gene as **0** to **1** and **1** to **0**

The Mutated Off-spring produced are :

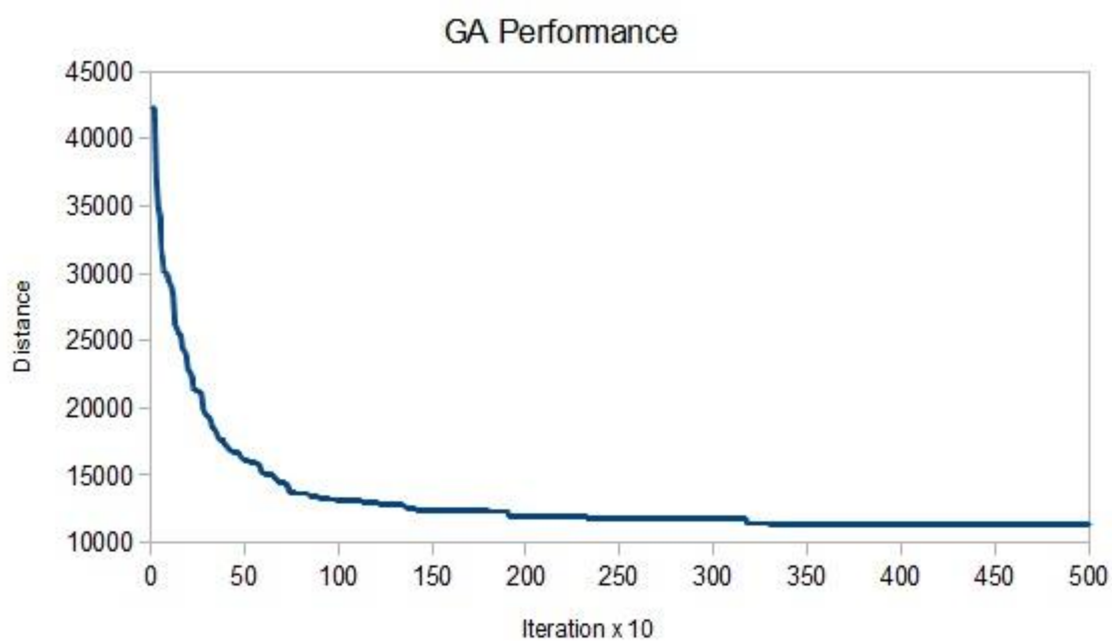
Mutated offspring 1	1	1	0	0	1	1	1	0	0	0	0	1	1	1	1	0
Mutated offspring 2	1	1	0	1	1	0	1	1	0	0	1	1	0	1	0	0

고려대학교 KOREA UNIVERSITY

DSBA Data Science & Business Analytics

GA Step 5: Find the Best Solution

- Find the best variable subset
  - ✓ Select the chromosome with the highest fitness value after the stopping criteria are satisfied.
  - ✓ Generally, significant fitness improvement occurs in the early stages, which becomes marginal after some generations




ANY
questions?