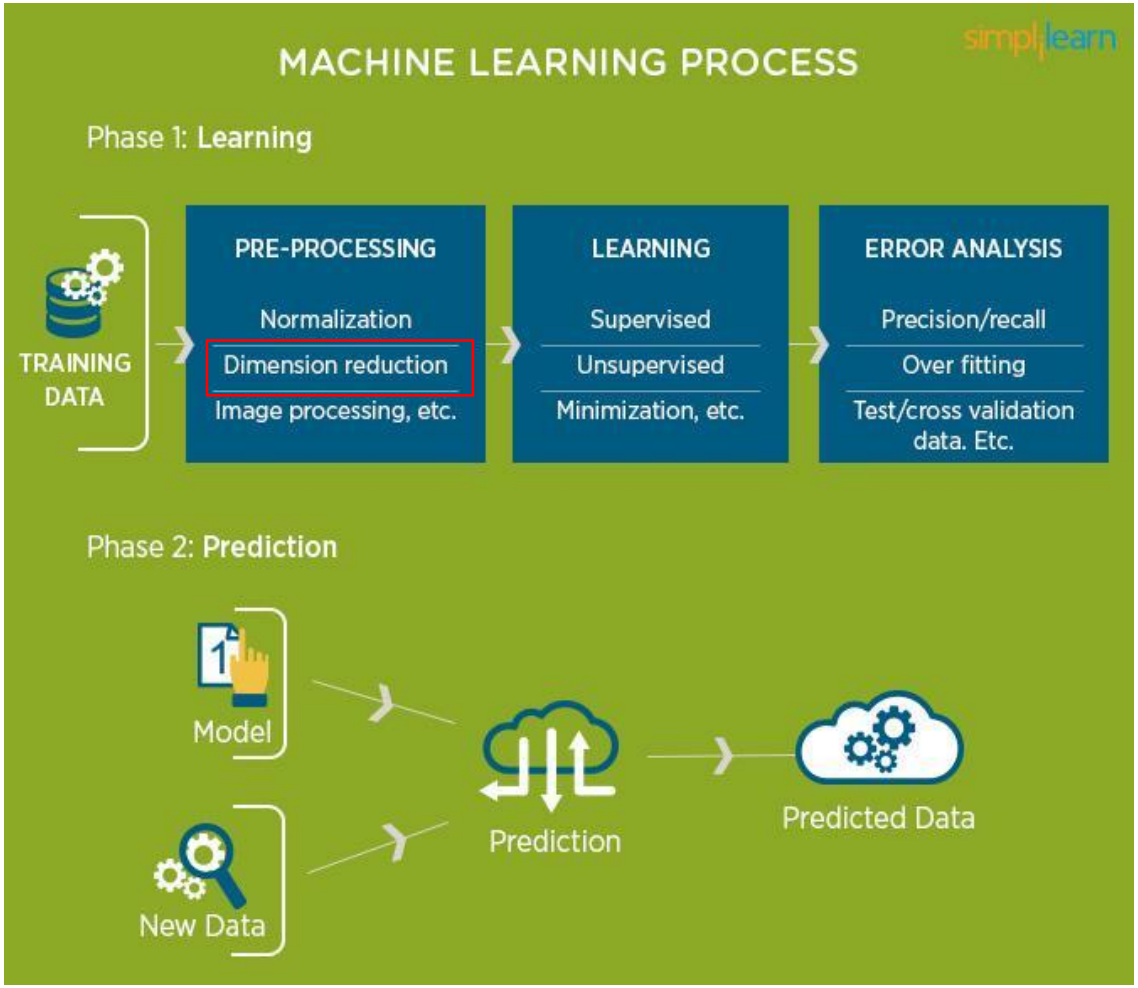
Variable Selection

강필성
고려대학교 산업경영공학부
pilsung_kang@korea.ac.kr

머신러닝 활용 데이터 분석 절차

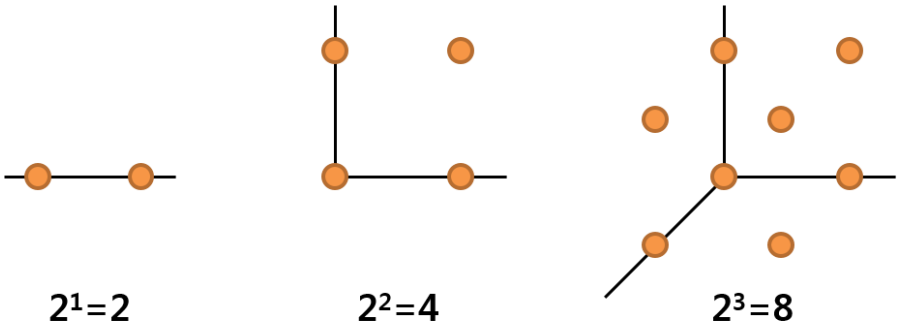
- 너무 많은 수의 변수가 존재할 경우 (FDC 센서 파라미터) 이를 줄일 필요가 있음

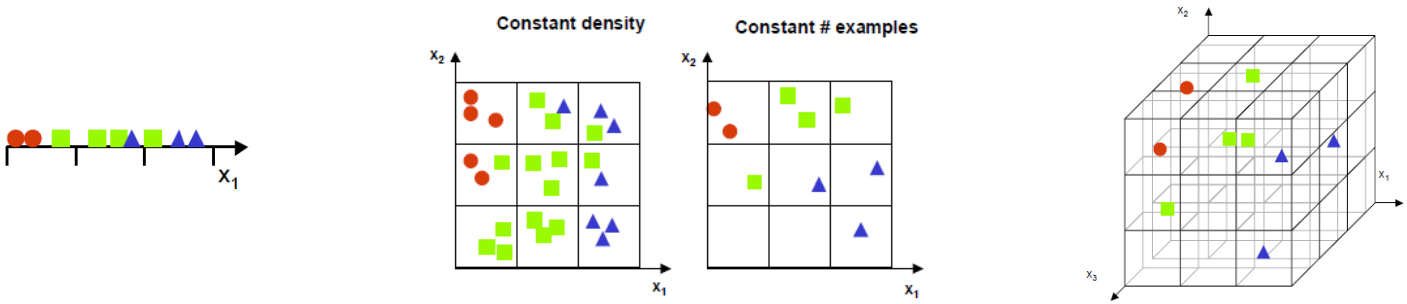


https://www.simplilearn.com/what-is-machine-learning-and-why-it-matters-article

차원축소: Dimensionality Reduction

- **차원의 저주** (Curse of Dimensionality)
  - ✓동등한 설명력을 갖기 위해서는 변수가 증가할 때 필요한 개체의 수는 기하급수적으로 증가함





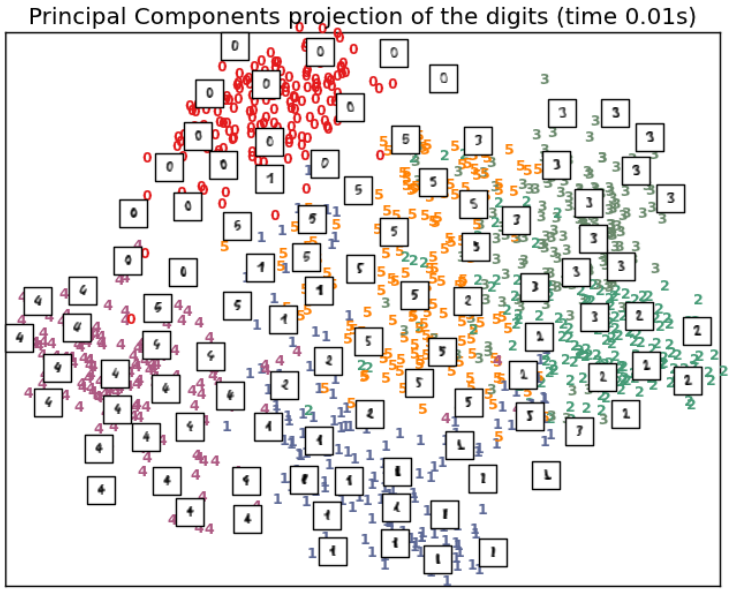
"If there are various logical ways to explain a certain phenomenon, the simplest is the best" - Occam's Razor

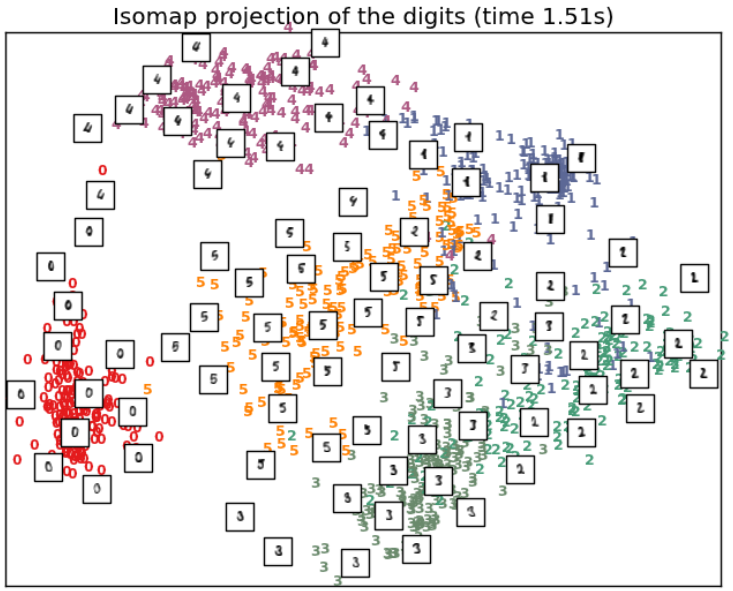
차원축소: Dimensionality Reduction

- **차원의 저주** (Curse of Dimensionality)
  - ✓ 고차원의 데이터 중에서는 실제 분석에 필요한 차원(데이터의 특성을 보존할 수 있는 내재적 차원)이 매우 낮은 경우도 존재함
    - 예시: 16 by 16 픽셀로 구성된 handwritten digit 데이터 (64차원)
    - 두 가지 차원 축소 기법을 통해 2차원으로 사영(projection) 시킨 결과

차원축소: Dimensionality Reduction

- **차원의 저주** (Curse of Dimensionality)
  - ✓ 고차원의 데이터 중에서는 실제 분석에 필요한 차원(데이터의 특성을 보존할 수 있는 내재적 차원)이 매우 낮은 경우도 존재함
    - 예시: 16 by 16 픽셀로 구성된 handwritten digit 데이터 (64차원)
    - 두 가지 차원 축소 기법을 통해 2차원으로 사영(projection) 시킨 결과





차원축소: Dimensionality Reduction

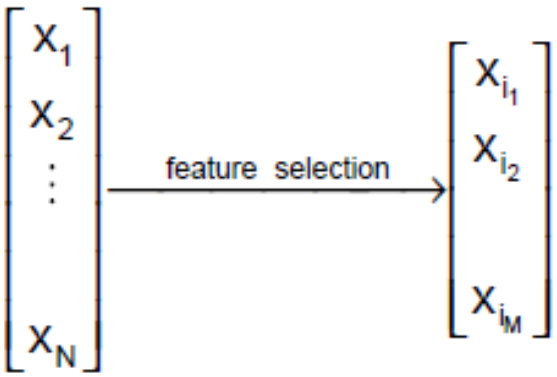
- **차원축소: 배경**
  - ✓ 이론적으로는 변수의 수가 증가할수록 모델의 성능이 향상됨 (변수간 독립성 만족 시)
  - ✓ 실제 상황에서는 변수간 독립성 가정 위배, 노이즈 존재 등으로 인해 변수의 수가 일정 수준 이상 증가하면 모델의 성능이 저하되는 경행이 있음
- **차원축소: 목적**
  - ✓ 향후 분석 과정에서 성능을 저하시키지 않는 최소한의 변수 집합 판별
- **차원축소: 효과**
  - ✓ 변수간 상관성을 제거하여 결과의 통계적 유의성 제고
  - ✓ 사후 처리(post-processing)의 단순화
  - ✓ 주요 정보를 보존한 상태에서 중복되거나 불필요한 정보만 제거
  - ✓ 고차원의 정보를 저차원으로 축소하여 시각화(visualization) 가능

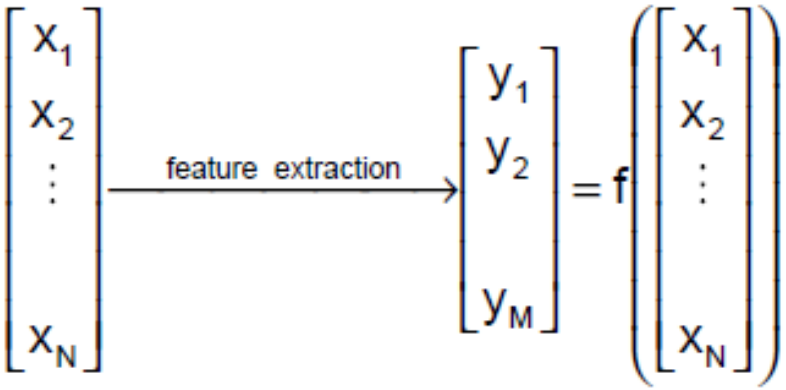
차원축소: Dimensionality Reduction

- **차원축소 방식 (모델 내부/외부 작동 기준)**
  - ✓명시적 방식 (Explicit method)
    - 학습 알고리즘 외부에 차원축소 매커니즘/모듈을 두어 독립적으로 작동
    - 교사적 차원축소 (Supervised dimensionality reduction)
      - 축소된 차원의 적합성을 검증하는데 있어 데이터마이닝 모델을 적용
      - 동일한 데이터라도 적용되는 데이터마이닝 모델에 따라 축소된 차원의 결과가 달라질 수 있음
    - 비교사적 차원축소 (Unsupervised dimensionality reduction)
      - 축소된 차원의 적합성을 검증하는데 있어 데이터마이닝 모델을 적용하지 않음
      - 특정 기법에 따른 차원축소 결과는 동일함
  - ✓묵시적 방식 (Implicit method)
    - 학습 알고리즘 내부의 목적함수에 적은 수의 변수를 찾는 것을 선호하도록 설계

차원축소: Dimensionality Reduction

- **차원축소 기법**
  - ✓변수 선택(variable/feature selection)
    - 원래의 변수 집단으로부터 유용할 것으로 판단되는 소수의 변수들을 선택
    - Filter – 변수 선택 과정과 모델 구축 과정이 독립적
    - Wrapper – 변수 선택 과정이 데이터마이닝 모델의 결과를 최적화 하는 방향으로 이루어짐
  - ✓변수 추출(variable/feature extraction)
    - 원래의 변수 집단을 보다 효율적인 적은 수의 새로운 변수 집단으로 변환
    - 데이터마이닝 모델에 독립적인 성능 지표가 추출된 변수의 효과를 측정하는 데 사용됨

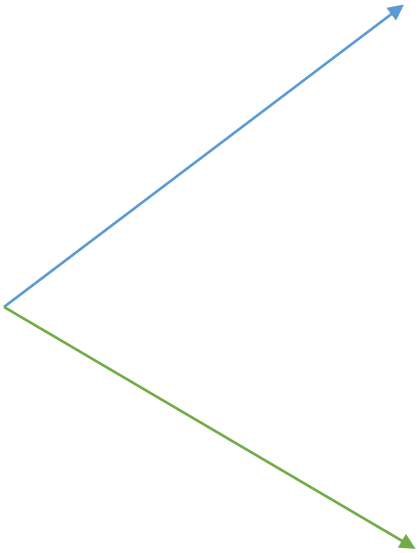




차원축소: Dimensionality Reduction

- 차원 감소 기법 (cont')
  - ✓변수 선택과 변수 추출 비교

X_1	X_2	X_3	...	X_n
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...



변수 선택

X_1	X_5	X_8
...	...	...
...	...	...
...	...	...
...	...	...
...	...	...

변수 추출

Z_1	Z_2	Z_3
...	...	...
...	...	...
...	...	...
...	...	...
...	...	...

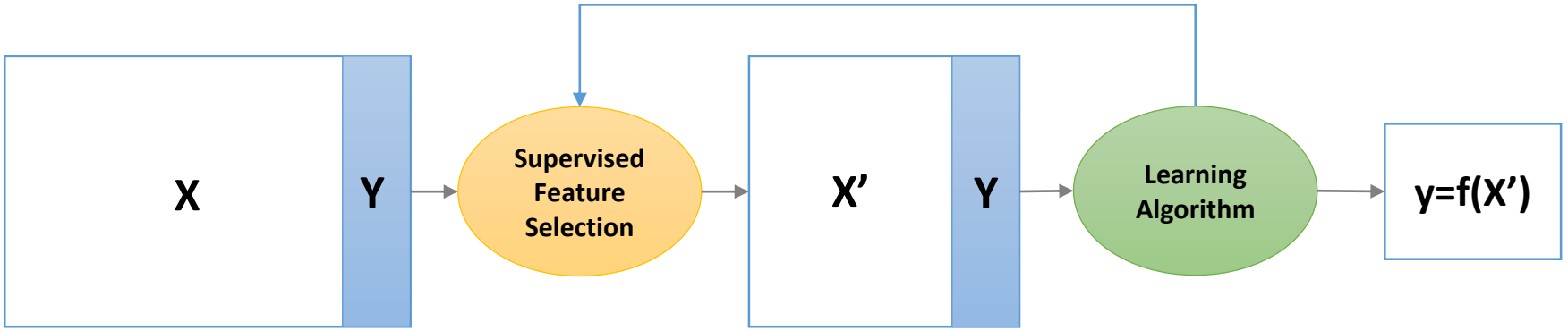
$Z_1 = X_1 + 0.2*X_2$

$Z_2 = X_3 - 2*X_5$

$Z_3 = X_4 + X_6 - X_9$

교사적 차원축소 기법

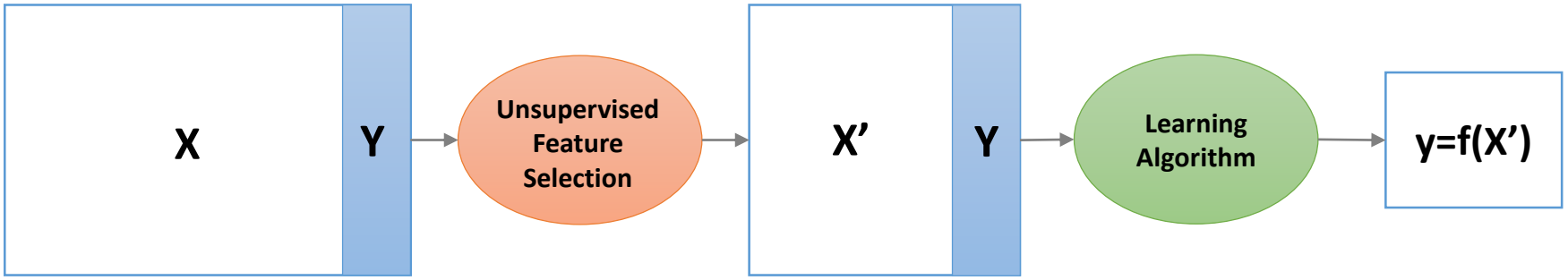
- **교사적 차원축소 기법** (Supervised feature selection)
  - ✓ d-차원의 데이터에 대하여 사용하는 모델의 성능이 최대가 되도록 하는 d'차원(d' << d)의 변수를 선택



  - ✓ 변수 선택을 하기 전, 모델 구축에 사용할 알고리즘을 먼저 선택
  - ✓ 동일한 데이터라도 모델 구축에 사용되는 알고리즘에 따라 다양한 선택 결과가 나타날 수 있음

비교사적 차원축소 기법

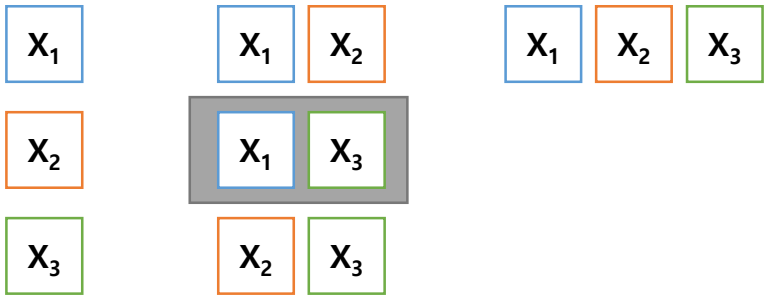
- **비교사적 차원축소 기법** (Unsupervised feature selection)
  - ✓ d-차원의 데이터에 대하여 사용하는 모델의 성능이 최대가 되도록 하는 d'차원(d' << d)의 변수를 선택



  - ✓ 변수 선택을 하기 전, 모델 구축에 사용할 알고리즘을 먼저 선택하지 않음
  - ✓ 변수 선택 결과가 모델 구축에 사용되는 알고리즘에 상관 없이 일정함

변수 선택 기반의 차원 축소 방법론

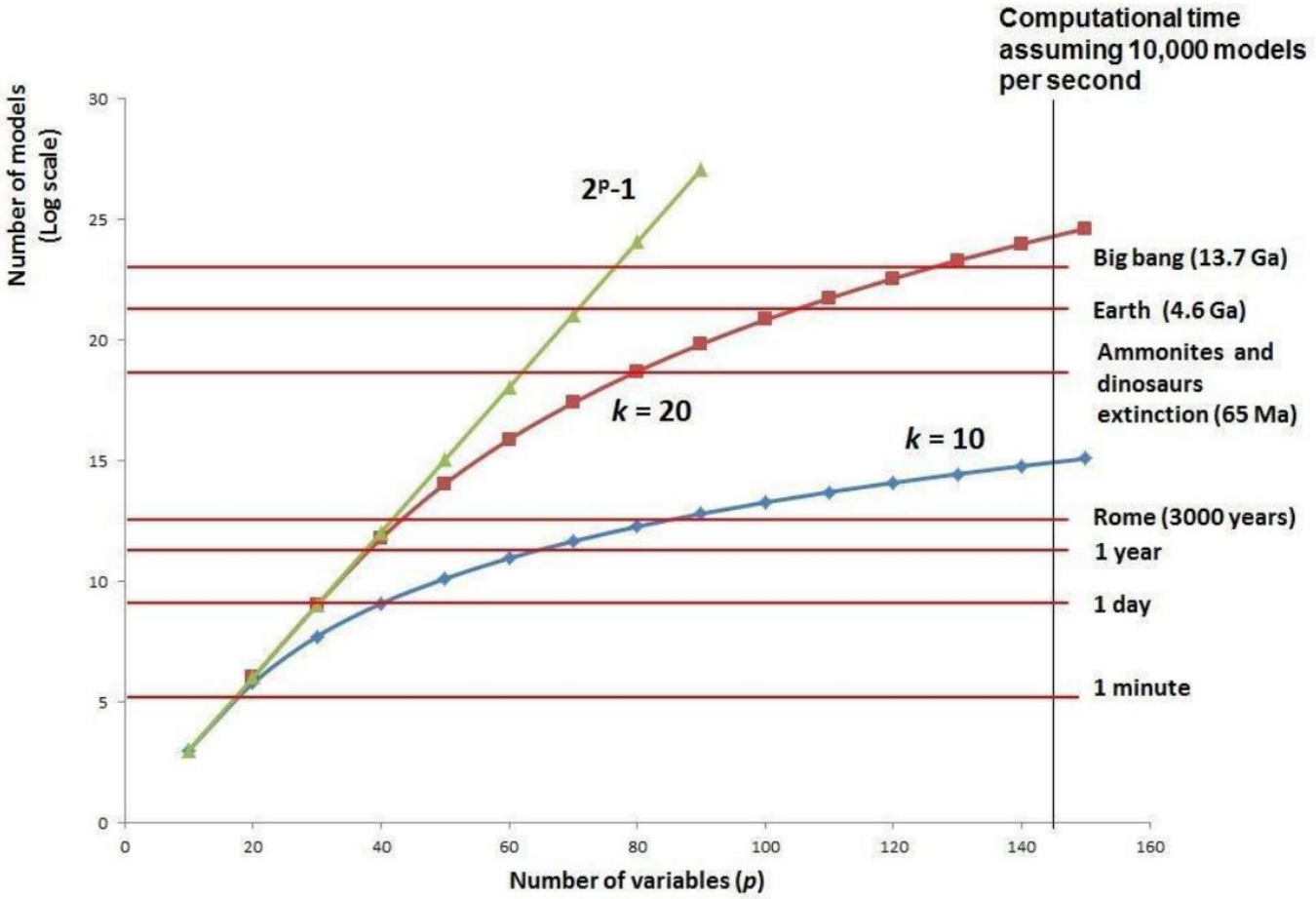
- **전역 탐색** (Exhaustive search)
  - ✓가능한 모든 경우의 조합에 대해 모델을 구축한 뒤 최적의 변수 조합을 찾는 방식
    - 예: 3개의 변수가 존재하는 경우 x_1 x_2 x_3
    - 총 여섯 가지의 가능한 변수 조합 존재



  - ✓변수 선택을 위한 모델 평가 기준
    - 선형 모형: Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), 수정 R-제곱합, Mallow's C_p 등
    - 비선형 모형: 검증 데이터에 대한 예측 정확도

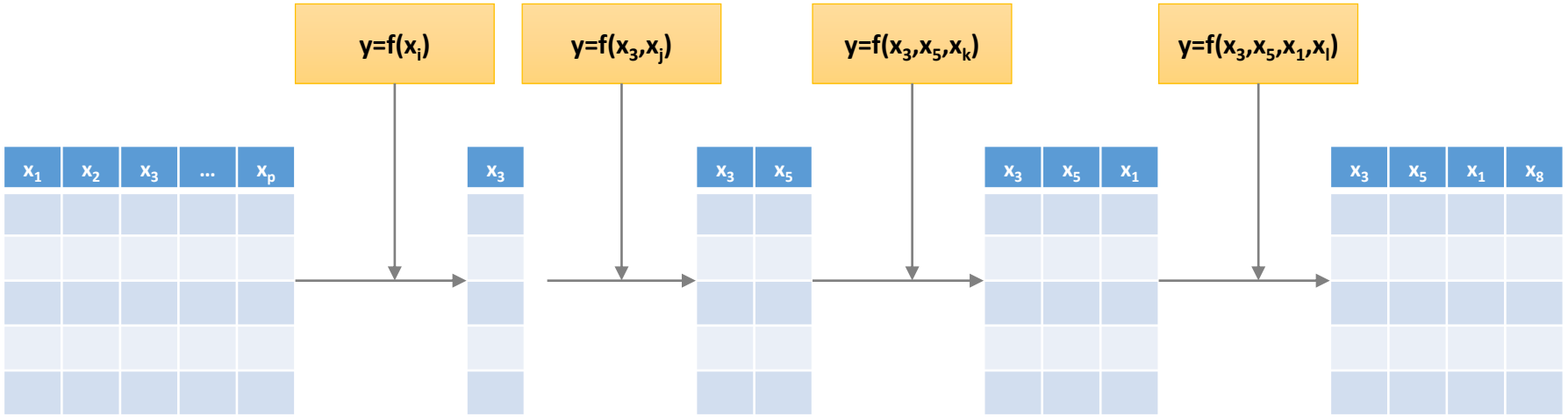
변수 선택 기반의 차원 축소 방법론

- **전역 탐색 (Exhaustive search)은 왜 섣불리 사용하면 안될까?**
  - ✓ 1초에 10,000개의 모델을 평가할 수 있는 컴퓨터를 활용할 경우 변수 선택에 소요되는 시간



전진 선택법: Forward Selection

- **전진 선택법**
  - ✓ 설명변수가 하나도 없는 모델에서부터 시작하여 가장 유의미한 변수를 하나씩 추가해 나가는 방법 (회귀분석 모델의 F-통계량 사용)
  - ✓ 한번 선택된 변수는 제거되지 않음 (변수의 숫자는 단조 증가)
  - ✓ 전진 선택법 예시



Forward Selection

- Forward Selection example
  - ✓ Forward Selection in the multiple linear regression
  - ✓ 8 input variables

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1, \quad R^2_{adj} = 0.48$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2, \quad R^2_{adj} = 0.56$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_3 x_3, \quad R^2_{adj} = 0.51$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_4 x_4, \quad R^2_{adj} = 0.50$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_5 x_5, \quad R^2_{adj} = 0.38$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_6 x_6, \quad R^2_{adj} = 0.32$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_7 x_7, \quad R^2_{adj} = 0.50$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_8 x_8, \quad R^2_{adj} = 0.19$$

Forward Selection

- Forward Selection example
  - ✓ Forward Selection in the multiple linear regression
  - ✓ 8 input variables

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2 + \hat{\beta}_1 x_1, \quad R^2_{adj} = 0.60$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3, \quad R^2_{adj} = 0.64$$

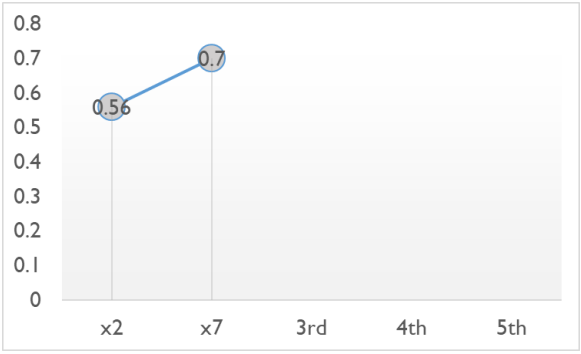
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2 + \hat{\beta}_4 x_4, \quad R^2_{adj} = 0.58$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2 + \hat{\beta}_5 x_5, \quad R^2_{adj} = 0.61$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2 + \hat{\beta}_6 x_6, \quad R^2_{adj} = 0.57$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2 + \hat{\beta}_7 x_7, \quad R^2_{adj} = 0.70$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2 + \hat{\beta}_8 x_8, \quad R^2_{adj} = 0.56$$



Forward Selection

- Forward Selection example
  - ✓ Forward Selection in the multiple linear regression
  - ✓ 8 input variables

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2 + \hat{\beta}_7 x_7 + \hat{\beta}_1 x_1, \quad R^2_{adj} = 0.71$$

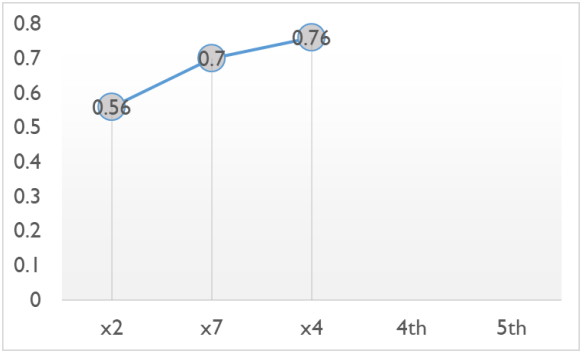
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2 + \hat{\beta}_7 x_7 + \hat{\beta}_3 x_3, \quad R^2_{adj} = 0.72$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2 + \hat{\beta}_7 x_7 + \hat{\beta}_4 x_4, \quad R^2_{adj} = 0.76$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2 + \hat{\beta}_7 x_7 + \hat{\beta}_5 x_5, \quad R^2_{adj} = 0.73$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2 + \hat{\beta}_7 x_7 + \hat{\beta}_6 x_6, \quad R^2_{adj} = 0.69$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2 + \hat{\beta}_7 x_7 + \hat{\beta}_8 x_8, \quad R^2_{adj} = 0.70$$



Forward Selection

- Forward Selection example
  - ✓ Forward Selection in the multiple linear regression
  - ✓ 8 input variables

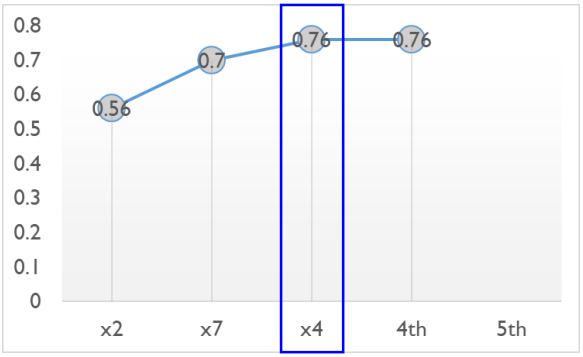
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2 + \hat{\beta}_7 x_7 + \hat{\beta}_4 x_4 + \hat{\beta}_1 x_1, \quad R^2_{adj} = 0.76$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2 + \hat{\beta}_7 x_7 + \hat{\beta}_4 x_4 + \hat{\beta}_3 x_3, \quad R^2_{adj} = 0.76$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2 + \hat{\beta}_7 x_7 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5, \quad R^2_{adj} = 0.75$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2 + \hat{\beta}_7 x_7 + \hat{\beta}_4 x_4 + \hat{\beta}_6 x_6, \quad R^2_{adj} = 0.76$$

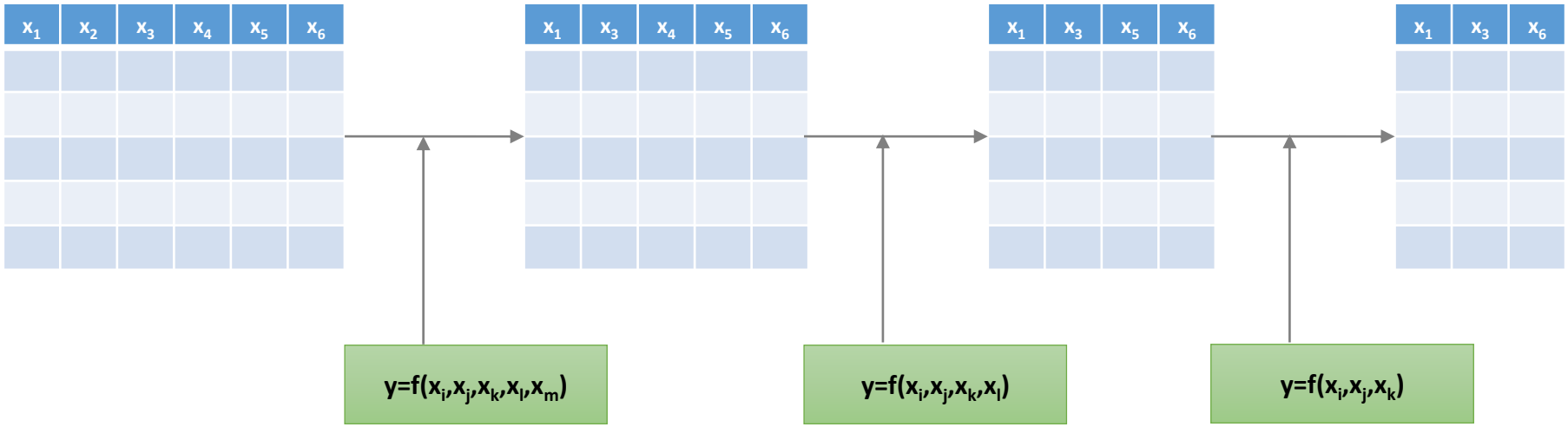
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2 + \hat{\beta}_7 x_7 + \hat{\beta}_4 x_4 + \hat{\beta}_8 x_8, \quad R^2_{adj} = 0.75$$



✓ Final model: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2 + \hat{\beta}_7 x_7 + \hat{\beta}_4 x_4, \quad R^2_{adj} = 0.76$

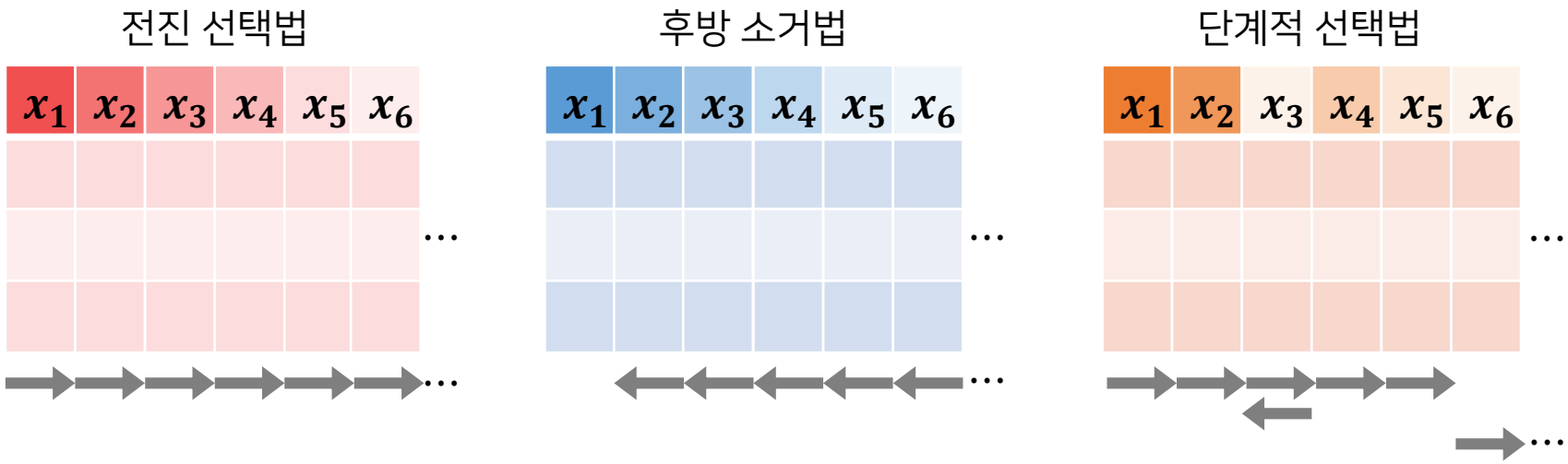
후방 소거법: Backward Elimination

- **후진 소거법**
  - ✓ 모든 변수를 사용하여 구축한 모델에서 유의미하지 않은 변수를 하나씩 제거해 나가는 방법
  - ✓ 한번 제거된 변수는 다시 선택될 가능성이 없음 (변수의 숫자는 단조 증가)
  - ✓ 후진 소거법 예시



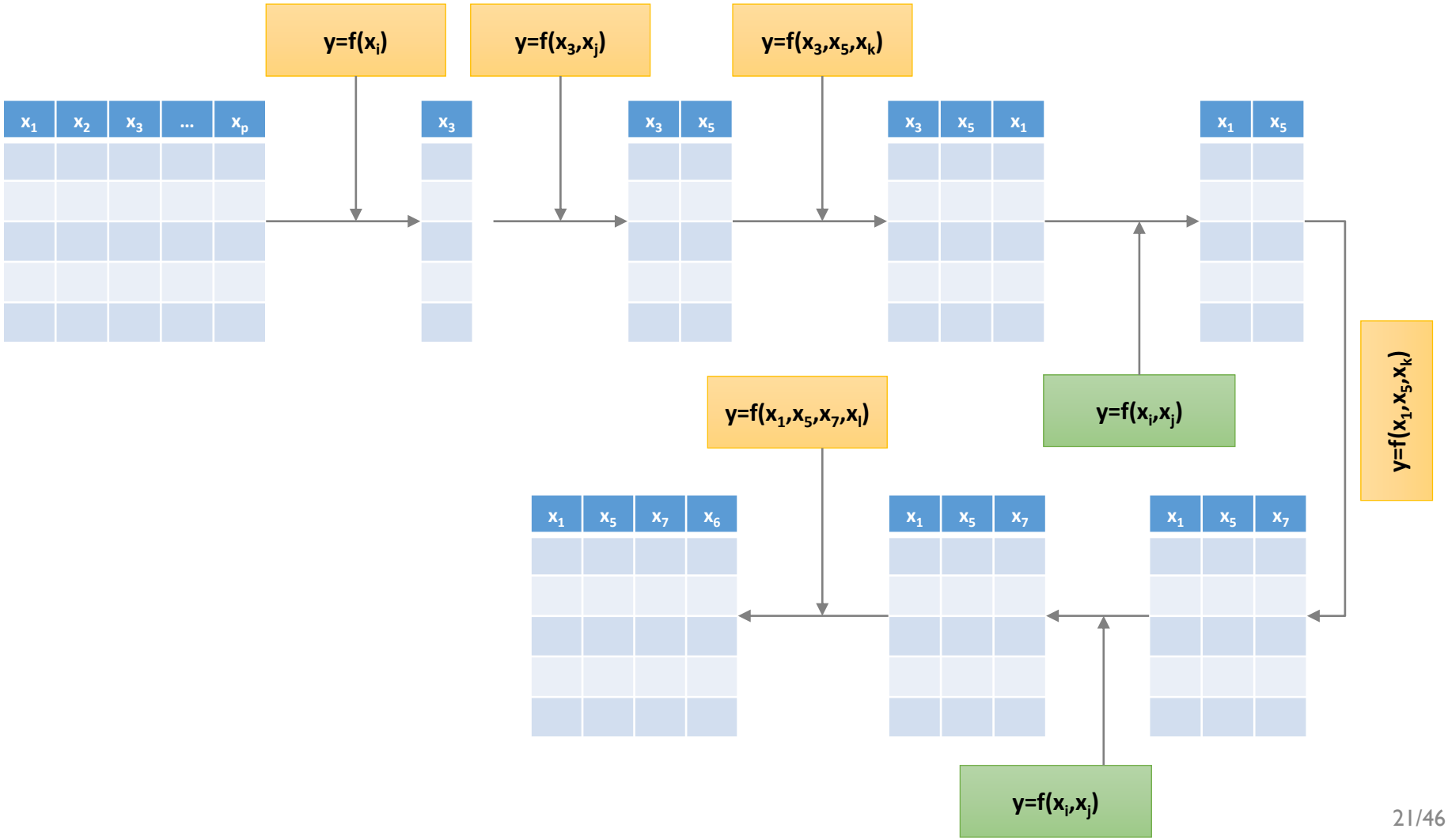
단계적 선택법: Stepwise Selection

- **단계적 선택법**
  - ✓ 설명변수가 하나도 없는 모델에서부터 시작하여 전진선택법과 후진소거법을 번갈아가며 수행
  - ✓ 전진선택법 및 후진소거법에 비해 시간을 오래 걸리나 보다 우수한 예측 성능을 나타내는 변수 집합을 찾아낼 가능성이 높음
  - ✓ 한번 선택되거나 제거된 변수라도 다시 선택/제거될 가능성이 있음
  - ✓ 변수의 수는 초기에는 일반적으로 증가하나 중반 이후에는 증가와 감소를 반복



단계적 선택법: Stepwise Selection

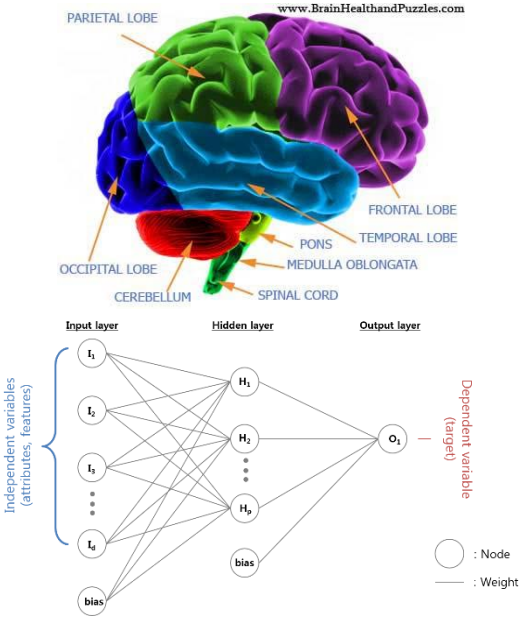
- 단계적 선택법 예시



유전 알고리즘: Genetic Algorithm

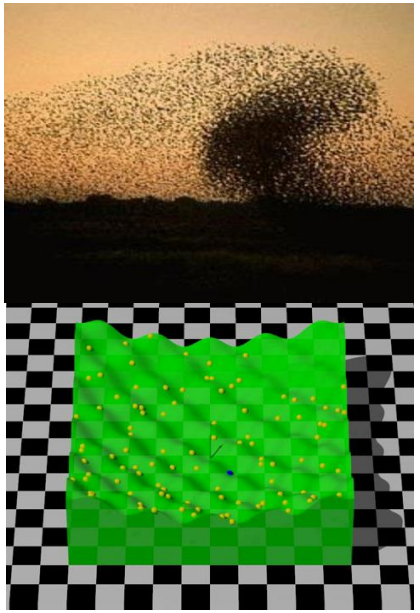
- Meta-Heuristic Approach
  - ✓ Solve a complex problem by doing trials and errors efficiently
  - ✓ Among the optimization algorithms, many of them mimic the way of a natural system works

Artificial Neural Networks



Ant Colony Algorithm

Particle Swarm Optimization



유전 알고리즘: Genetic Algorithm

- Meta-Heuristic Approach

AlphaStar: An Evolutionary Computation Perspective

Kai Arulkumaran
Imperial College London
London, United Kingdom
ka709@ic.ac.uk

Antoine Cully
Imperial College London
London, United Kingdom
a.cully@imperial.ac.uk

Julian Togelius
New York University
New York City, NY, United States
julian@togelius.com

ABSTRACT

In January 2019, DeepMind revealed AlphaStar to the world—the first artificial intelligence (AI) system to beat a professional player at the game of StarCraft II—representing a milestone in the progress of AI. AlphaStar draws on many areas of AI research, including deep learning, reinforcement learning, game theory, and evolutionary computation (EC). In this paper we analyze AlphaStar primarily through the lens of EC, presenting a new look at the system and relating it to many concepts in the field. We highlight some of its most interesting aspects—the use of Lamarckian evolution, competitive co-evolution, and quality diversity. In doing so, we hope to provide a bridge between the wider EC community and one of the most significant AI systems developed in recent times.

CCS CONCEPTS

• **Computing methodologies** → **Multi-agent reinforcement learning**; **Neural networks**; *Bio-inspired approaches*;

KEYWORDS

Lamarckian evolution, co-evolution, quality diversity

beat a grandmaster at StarCraft (SC), a real-time strategy game. Both the original game, and its sequel SC II, have several properties that make it considerably more challenging than even Go: real-time play, partial observability, no single dominant strategy, complex rules that make it hard to build a fast forward model, and a particularly large and varied action space.

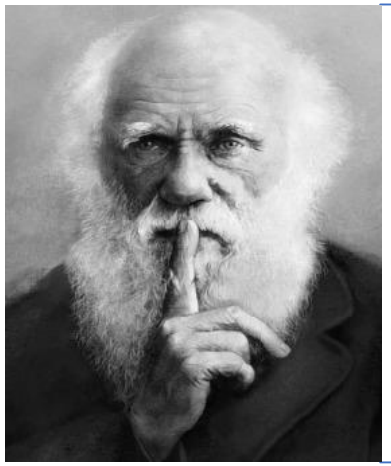
DeepMind recently took a considerable step towards this grand challenge with AlphaStar, a neural-network-based AI system that was able to beat a professional SC II player in December 2018 [20]. This system, like its predecessor AlphaGo, was initially trained using imitation learning to mimic human play, and then improved through a combination of reinforcement learning (RL) and self-play. At this point the algorithms diverge, as AlphaStar utilises population-based training (PBT) [9] to explicitly keep a population of agents that train against each other [8]. This part of the training process was built upon multi-agent RL and game-theoretic perspectives [2, 10], but the very notion of a population is central to evolutionary computation (EC), and hence we can examine AlphaStar through this lens as well[1].

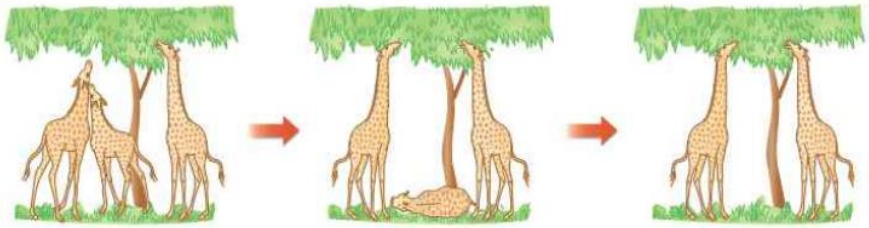
2 COMPONENTS

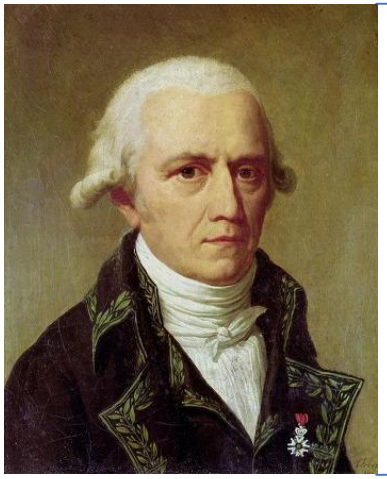
Arulkumaran, K., Cully, A., & Togelius, J. (2019). Alphastar: An evolutionary computation perspective. *arXiv preprint arXiv:1902.01724*.

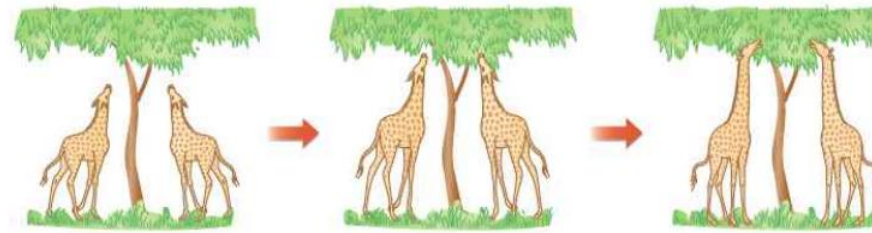
유전 알고리즘: Genetic Algorithm

- 자연선택설 vs. 용불용설



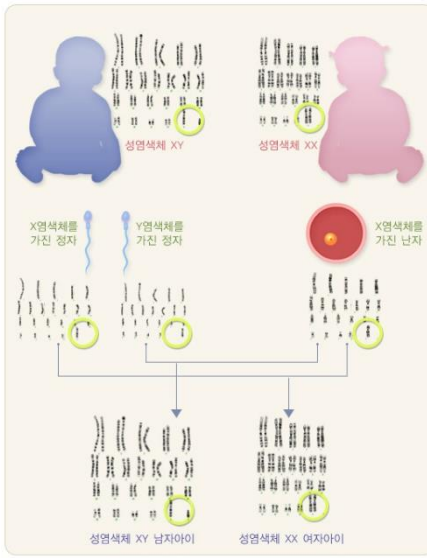


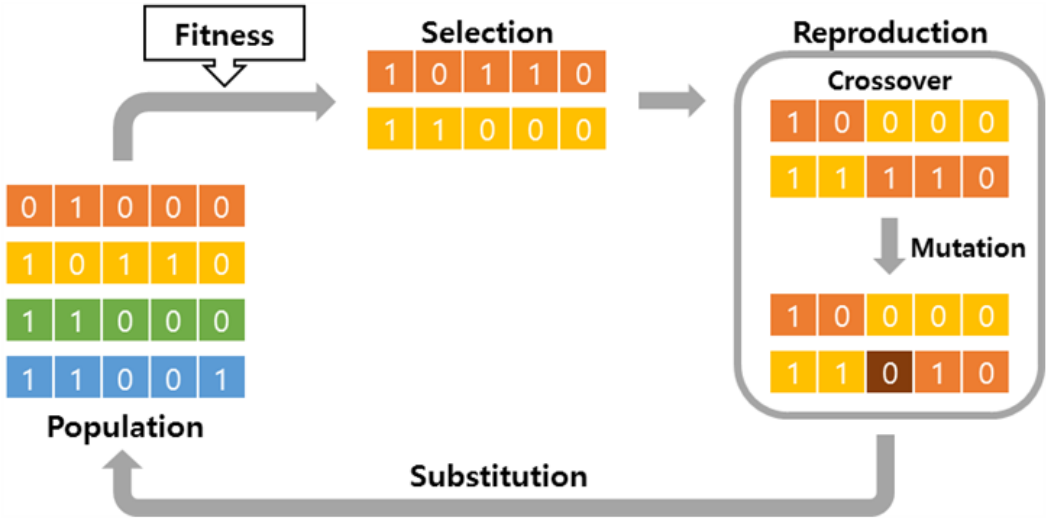


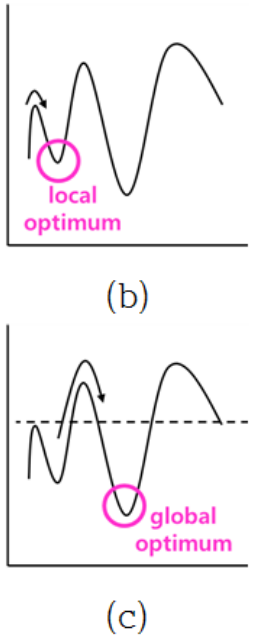


유전 알고리즘: Genetic Algorithm

- 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







유전 알고리즘: Genetic Algorithm

- Genetic Algorithm for Feature Selection

Step1:
염색체 초기화 및 파라미터 설정 (Initiation)

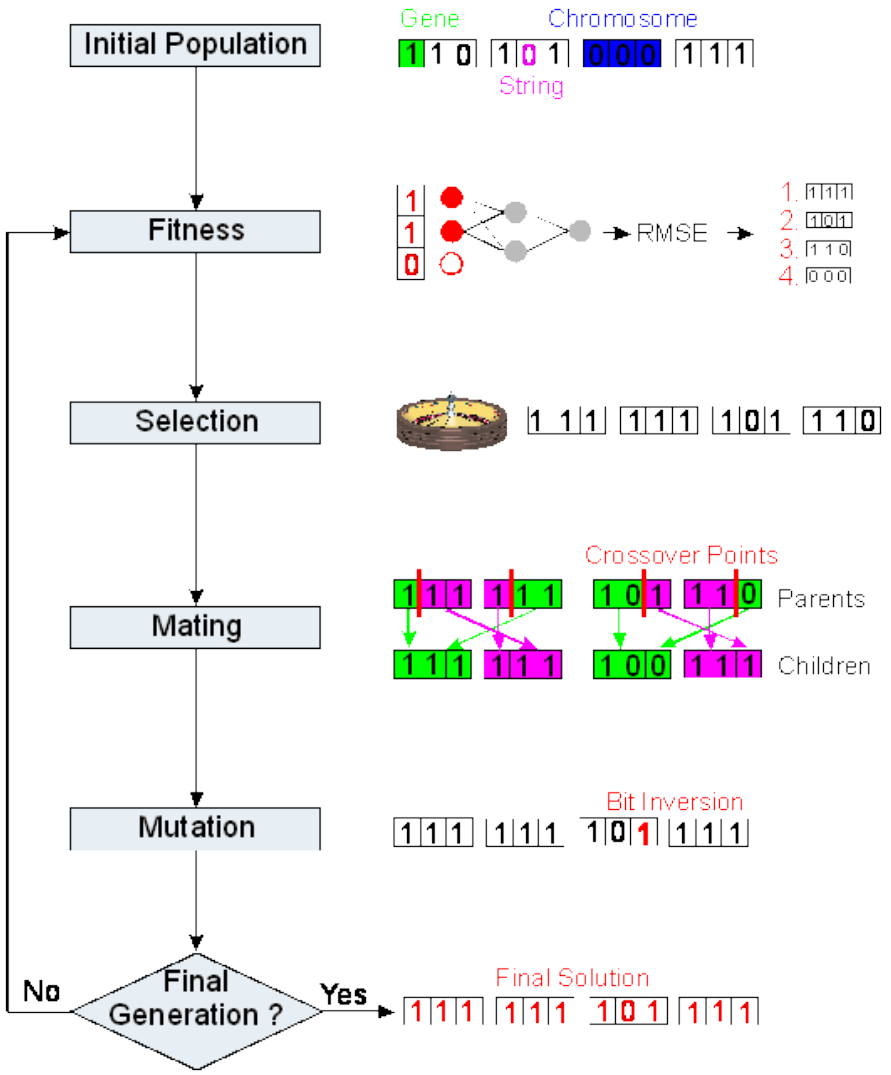
Step2:
각 염색체 선택 변수별 모델 학습

Step3:
각 염색체 적합도 평가 (Fitness evaluation)

Step4:
우수 염색체 선택 (Selection)

Step5:
다음 세대 염색체 생성 (Crossover & Mutation)

Step6:
최종 변수 집합 선택



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

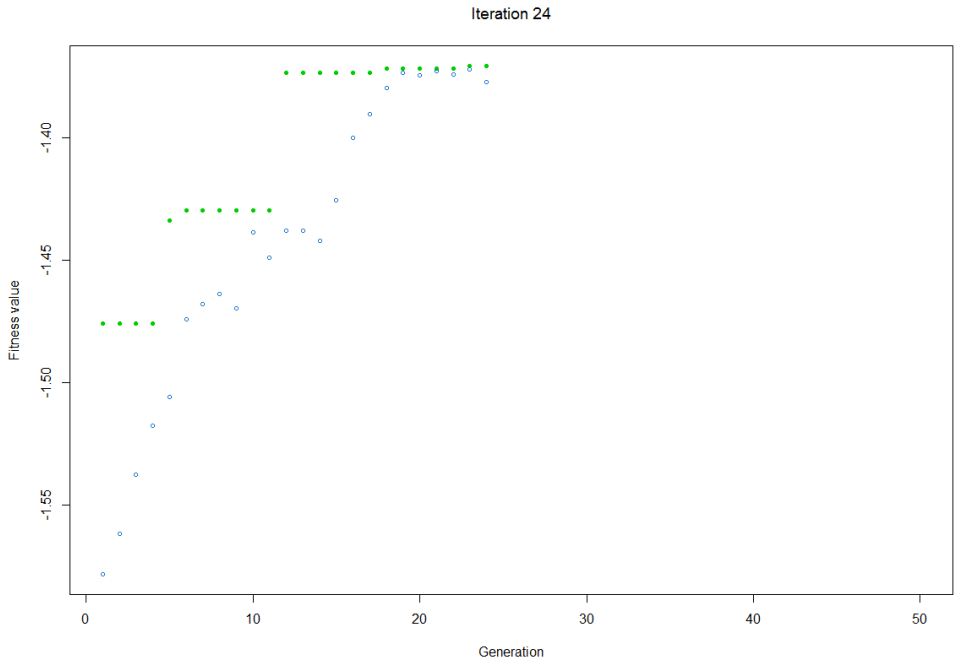
Gene

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 1: Initialization

- Example: Population Initialization
  - ✓ Random number generation for each gene
  - ✓ Convert the random numbers to binary values (cut-off = 0.5 in this example)

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
Chromosome 1	0.16	0.74	0.90	0.96	0.19	0.70	0.31	0.23	0.83	0.62
Chromosome 2	0.71	0.75	0.82	0.83	0.41	0.97	0.14	0.40	0.89	0.66
Chromosome 3	0.36	0.56	0.40	0.08	0.80	0.32	0.44	0.27	0.34	0.31
Chromosome 4	0.30	0.26	0.83	0.71	0.70	0.74	0.11	0.57	0.81	0.01
Chromosome 5	0.27	0.29	0.67	0.54	0.38	0.09	0.07	0.90	0.00	0.53
Chromosome 6	0.70	0.59	0.62	0.18	0.36	0.02	0.45	0.24	0.40	0.06
Chromosome 7	0.43	0.49	0.34	0.36	0.26	0.67	0.09	0.68	0.29	0.44
Chromosome 8	0.10	0.96	0.20	0.16	0.65	0.07	0.73	0.14	0.91	0.77

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
Chromosome 1	0	1	1	1	0	1	0	0	1	1
Chromosome 2	1	1	1	1	0	1	0	0	1	1
Chromosome 3	0	1	0	0	1	0	0	0	0	0
Chromosome 4	0	0	1	1	1	1	0	1	1	0
Chromosome 5	0	0	1	1	0	0	0	1	0	1
Chromosome 6	1	1	1	0	0	0	0	0	0	0
Chromosome 7	0	0	0	0	0	1	0	1	0	0
Chromosome 8	0	1	0	0	1	0	1	0	1	1

GA Step 1: Initialization

- Example: Train the model
  - ✓ Assume that the model is a multivariate linear regression (MLR)

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
Chromosome 1	0	1	1	1	0	1	0	0	1	1
Chromosome 2	1	1	1	1	0	1	0	0	1	1
Chromosome 3	0	1	0	0	1	0	0	0	0	0
Chromosome 4	0	0	1	1	1	1	0	1	1	0
Chromosome 5	0	0	1	1	0	0	0	1	0	1
Chromosome 6	1	1	1	0	0	0	0	0	0	0
Chromosome 7	0	0	0	0	0	1	0	1	0	0
Chromosome 8	0	1	0	0	1	0	1	0	1	1

  - For chromosome 1, fit the MLR model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_6 x_6 + \hat{\beta}_9 x_9 + \hat{\beta}_{10} x_{10}$$

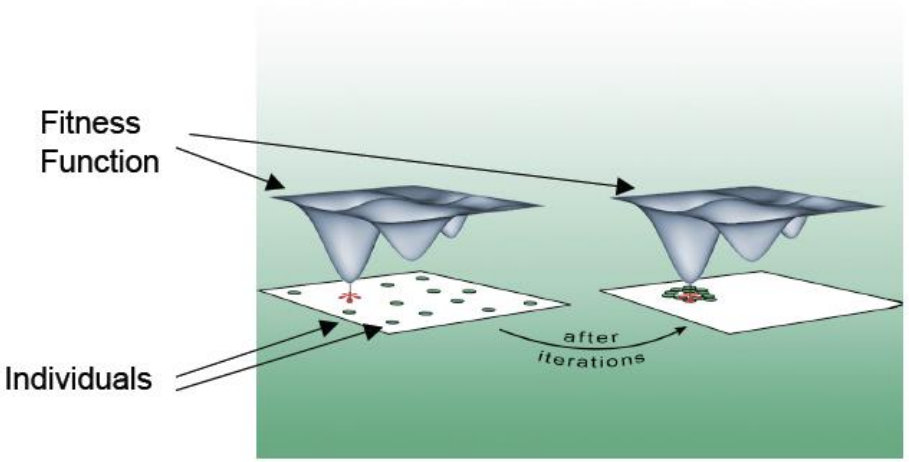
  - For chromosome 2, fit the MLR model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_6 x_6 + \hat{\beta}_9 x_9 + \hat{\beta}_{10} x_{10}$$

  - And so on…

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 R^2
    - Akaike information criterion (AIC)
    - Bayesian information criterion (BIC)



GA Step 3: Fitness Evaluation

- Example: Fitness Function

	Adj R^2	Rank	Weight
Chromosome 1	0.75	2	0.177
Chromosome 2	0.78	1	0.184
Chromosome 3	0.50	5	0.118
Chromosome 4	0.65	3	0.154
Chromosome 5	0.40	6	0.095
Chromosome 6	0.35	7	0.083
Chromosome 7	0.25	8	0.059
Chromosome 8	0.55	4	0.130

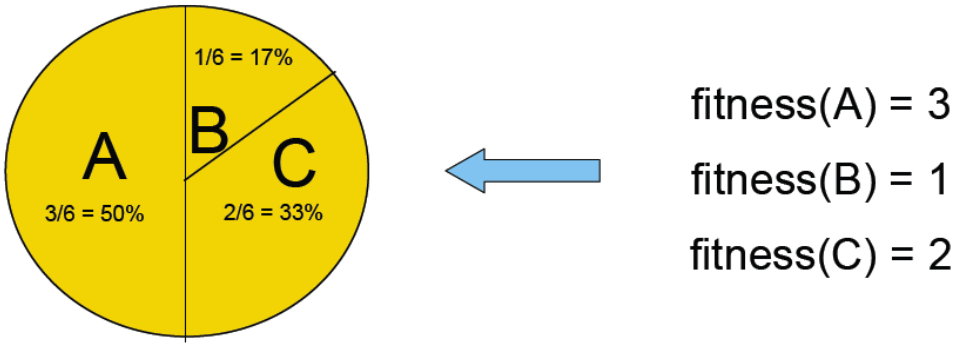
✓ Adjusted R^2 is used for the fitness function

✓ Each chromosome's weight is its Adj. R^2 divided by the sum of Adj. R^2s of all chromosomes

- Used for probabilistic chromosome selection

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



GA Step 4: Selection

- Example: Selection
  - ✓ Case 1: Deterministic selection

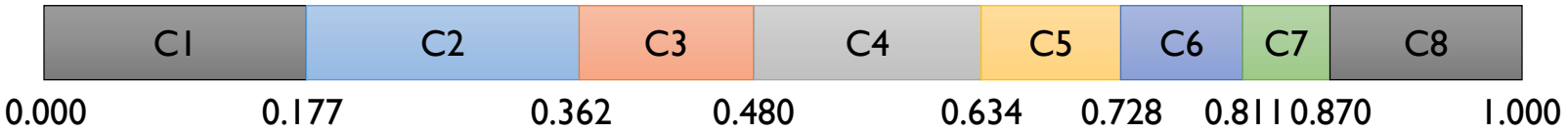
	Adj R^2	Rank	Weight
Chromosome 1	0.75	2	0.177
Chromosome 2	0.78	1	0.184
Chromosome 3	0.50	5	0.118
Chromosome 4	0.65	3	0.154
Chromosome 5	0.40	6	0.095
Chromosome 6	0.35	7	0.083
Chromosome 7	0.25	8	0.059
Chromosome 8	0.55	4	0.130

  - ✓ Only the chromosomes 1, 2, 4, and 8 can be a parent for the next generation

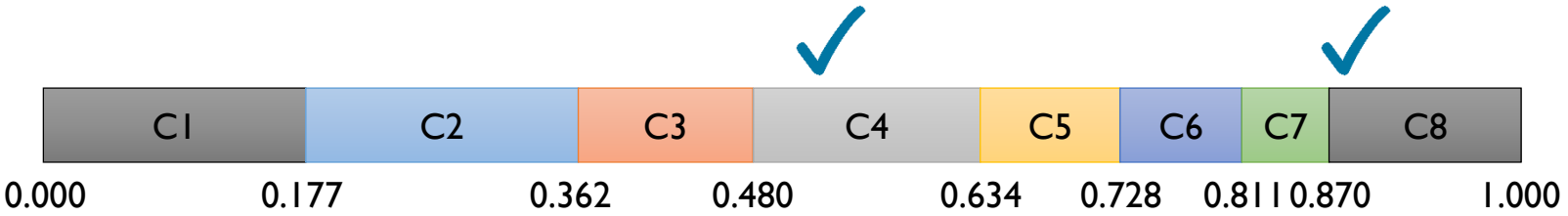
	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
Chromosome 1	0	1	1	1	0	1	0	0	1	1
Chromosome 2	1	1	1	1	0	1	0	0	1	1
Chromosome 4	0	0	1	1	1	1	0	1	1	0
Chromosome 8	0	1	0	0	1	0	1	0	1	1

GA Step 4: Selection

- Example: Selection
  - ✓ Case 2: Probabilistic selection
    - Cumulative weight



    - For the first set of parents
      - Random number initialization: (0.881, 0.499)

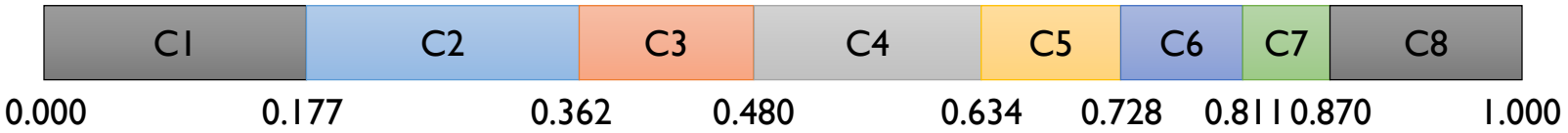


- Choose the chromosome 4 and 8

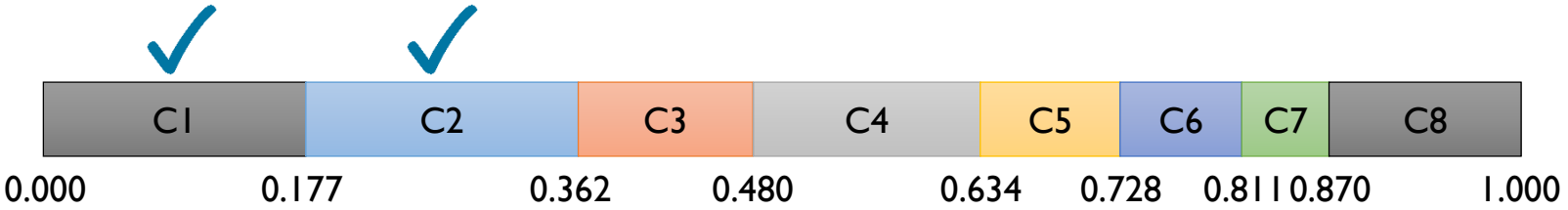
	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
Chromosome 4	0	0	1	1	1	1	0	1	1	0
Chromosome 8	0	1	0	0	1	0	1	0	1	1

GA Step 4: Selection

- Example: Selection
  - ✓ Case 2: Probabilistic selection
    - Cumulative weight



    - For the second set of parents
      - Random number initialization: (0.098, 0.252)



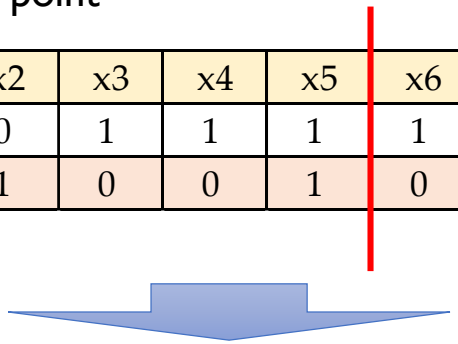
      - Choose the chromosome 1 and 2

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
Chromosome 1	0	1	1	1	0	1	0	0	1	1
Chromosome 2	1	1	1	1	0	1	0	0	1	1

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)
    - Example with 1 crossover point

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
Chromosome 4	0	0	1	1	1	1	0	1	1	0
Chromosome 8	0	1	0	0	1	0	1	0	1	1



	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
Child 1	0	1	0	0	1	1	0	1	1	0
Child 2	0	0	1	1	1	0	1	0	1	1

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)
    - Example with 2 crossover point

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
Chromosome 4	0	0	1	1	1	1	0	1	1	0
Chromosome 8	0	1	0	0	1	0	1	0	1	1

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
Child 1	0	1	0	1	1	1	0	0	1	1
Child 2	0	0	1	0	1	0	1	1	1	0

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)
    - Example with 10 crossover point

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
Chromosome 4	0	0	1	1	1	1	0	1	1	0
Chromosome 8	0	1	0	0	1	0	1	0	1	1
Random number	0.94	0.89	0.27	0.76	0.54	0.5	0.56	0.08	0	0.46

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
Child 1	0	1	1	0	1	0	1	1	1	0
Child 2	0	0	0	1	1	1	0	0	1	1

GA Step 5: Crossover & Mutation

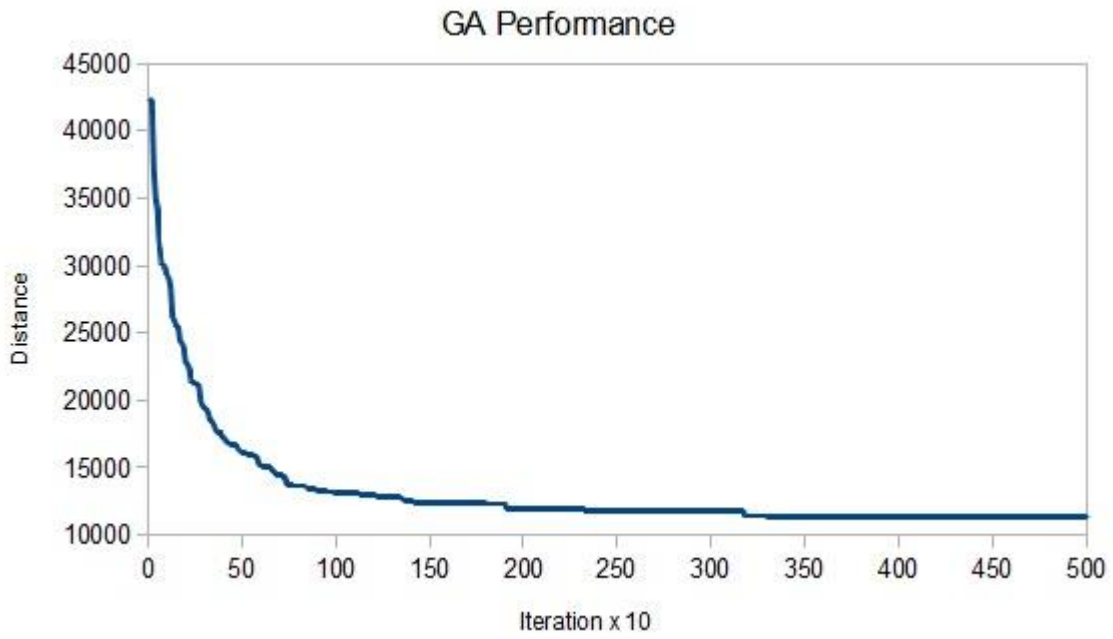
- Mutation
  - ✓ Genetic operator used to maintain diversity from one generation of a population of chromosomes to the next
  - ✓ Alters one or more gene values in a chromosome from its initial state, which result in entirely new gene values being added to the gene pool
  - ✓ By mutation, the current 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)

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
Child 1	0	1	1	0	1	0	1	1	1	0
Random number	0.43	0.35	0.71	0.54	0.62	0.73	0.71	0.92	0.95	0.91
Child 2	0	0	0	1	1	1	0	0	1	1
Random number	0.91	0.03	0.22	0.96	0.32	0.73	0.43	0.32	0.01	0.04

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
Child 1	0	1	1	0	1	0	1	1	1	0
Child 2	0	0	0	1	1	1	0	0	**0**	1

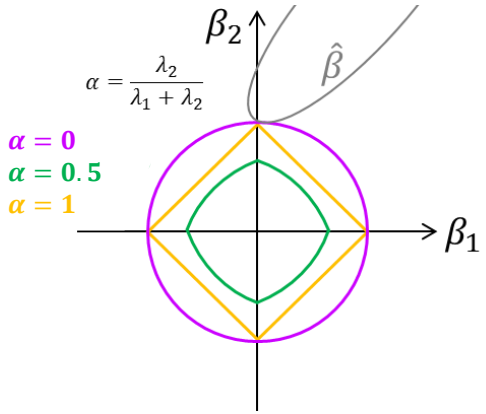
GA Step 5: Find the Best Solution

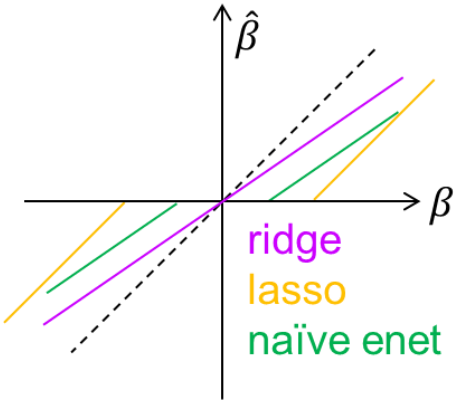
- Find the best variable subset
  - ✓ Select the chromosome with the highest fitness value after the stopping criteria are satisfied.
  - ✓ Generally, significant fitness improvement occurs in the early stages, which becomes marginal after some generations



Empirical Study

- Compare four variable selection and three shrinkage methods in linear regression
  - ✓ Variable selection: forward selection, backward elimination, stepwise selection, genetic algorithms
  - ✓ Shrinkage methods: ridge regression, lasso regression, elastic net





Ridge	$\hat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta}} \lvert \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \rvert^2 + \lambda_1 \lvert \boldsymbol{\beta} \rvert^2$	**shrinkage**
Lasso	$\hat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta}} \lvert \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \rvert^2 + \lambda_2 \lvert \boldsymbol{\beta} \rvert^1$	**shrinkage, variable selection**
Elastic net	$\hat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta}} \lvert \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \rvert^2 + \lambda_2 \lvert \boldsymbol{\beta} \rvert^1 + \lambda_1 \lvert \boldsymbol{\beta} \rvert^2$	**shrinkage, variable selection, grouping effect**

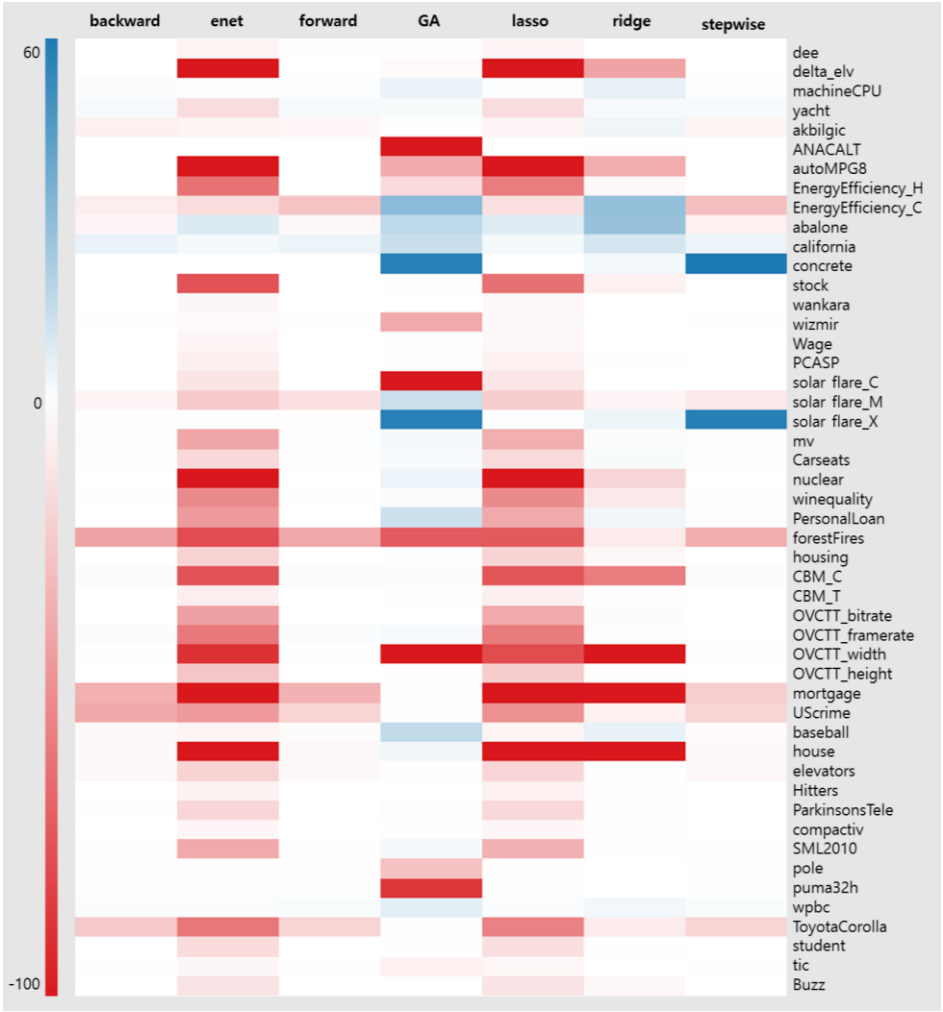
Empirical Study

- Data sets: 49 benchmark datasets

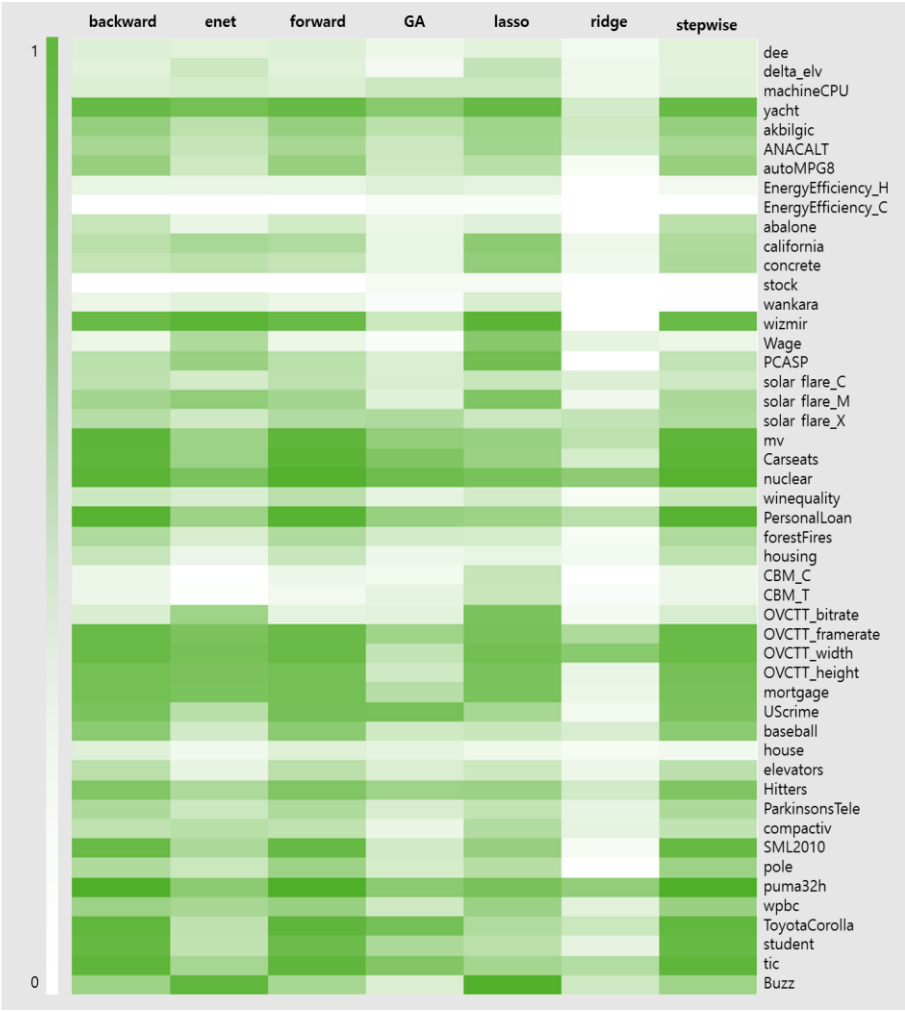
Dataset	source	records	variables	Dataset	source	records	variables
abalone	KEEL	4,177	9	OVCTT_bitrate	UCI	68,784	16
akbilgic	UCI	536	8	OVCTT_framerate	UCI	68,784	16
ANACALT	KEEL	4,052	8	OVCTT_height	UCI	68,784	16
autoMPG8	KEEL	392	8	OVCTT_width	UCI	68,784	16
baseball	KEEL	336	17	ParkinsonsTele	UCI	5,875	21
Buzz	UCI	28,179	95	PersonalLoan	etc.	2,500	13
california	KEEL	20,640	9	PCASP	UCI	45,730	10
Carseats	R	400	11	pole	KEEL	14,998	27
CBM_C	UCI	11,934	15	puma32h	KEEL	4,124	33
CBM_T	UCI	11,934	15	SML2010	UCI	4,137	24
compactiv	KEEL	8,192	22	solar flare_C	UCI	323	11
concrete	KEEL	1,030	9	solar flare_M	UCI	323	11
dee	KEEL	365	7	solar flare_X	UCI	323	11
delta_elv	KEEL	9,517	7	stock	KEEL	950	10
elevators	KEEL	16,599	19	student	UCI	382	51
EnergyEfficiency_C	UCI	768	9	tic	KEEL	9,822	86
EnergyEfficiency_H	UCI	768	9	ToyotaCorolla	etc.	1,436	34
forestFires	KEEL	517	13	UScrime	R	47	16
Hitters	R	263	20	Wage	R	3,000	10
house	KEEL	22,784	17	wankara	KEEL	1,609	10
housing	UCI	506	14	winequality	UCI	6,497	12
machineCPU	KEEL	209	7	wizmir	KEEL	1,461	10
mortgage	KEEL	1,049	16	wpbc	UCI	194	34
mv	KEEL	40,768	11	yacht	UCI	308	7
nuclear	R	32	11				

Empirical Study

Error Rate Improvement



Variable Reduction Ratio



Empirical Study

- Rankings in terms of
  - ✓ (1) Error rate improvement
  - ✓ (2) Variable reduction rate
  - ✓ (3) Computational efficiency

Variable selection technique	Error rate improvement	Variable reduction rate	Computational efficiency
Forward	5	4	1
Backward	4	3	2
Stepwise	3	2	6
GA	1	6	7
Ridge	2	7	5
LASSO	7	1	3
Enet	6	5	4


ANY
questions?