STEVENS INSTITUTE of TECHNOLOGY
School of Business

Amir H Gandomi; PhD
Stevens Institute of Technology
a.h.gandomi@stevens.edu

EVOLUTIONARY COMPUTATION FOR BIG DATA MINING

iBS PAN

Dec 7, 2018

Outlines

Modelling (Data Mining)

- Genetic Programming
- Big Data

Optimization

- Techniques
- Heuristics

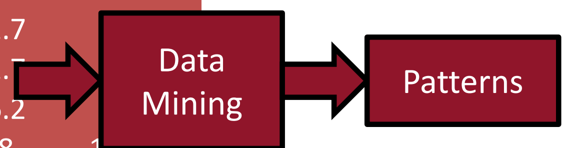
Modelling

Data Mining

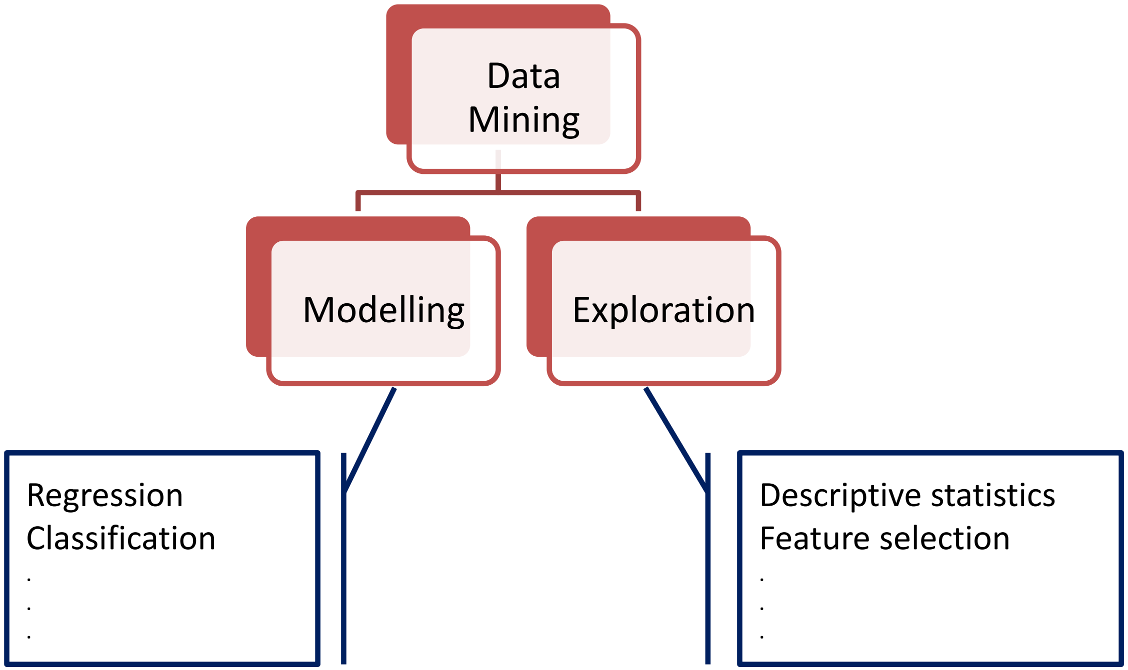
Data Mining

- Discovering patterns

ID	b	h	fcu	Mu
1	400	35	12.6	3450
2	400	35	12.6	5393
3	400	35	12.6	955
4	300	35	12.6	1935
5	300	50	42.7	[illegible]
6	300	50	42.[illegible]	[illegible]
7	200	50	46.2	[illegible]
8	200	50	48	[illegible]
9	200	80	50.8	483.6
10	100	80	50.28	684
11	100	80	49.1	1584
12	100	80	50	2168

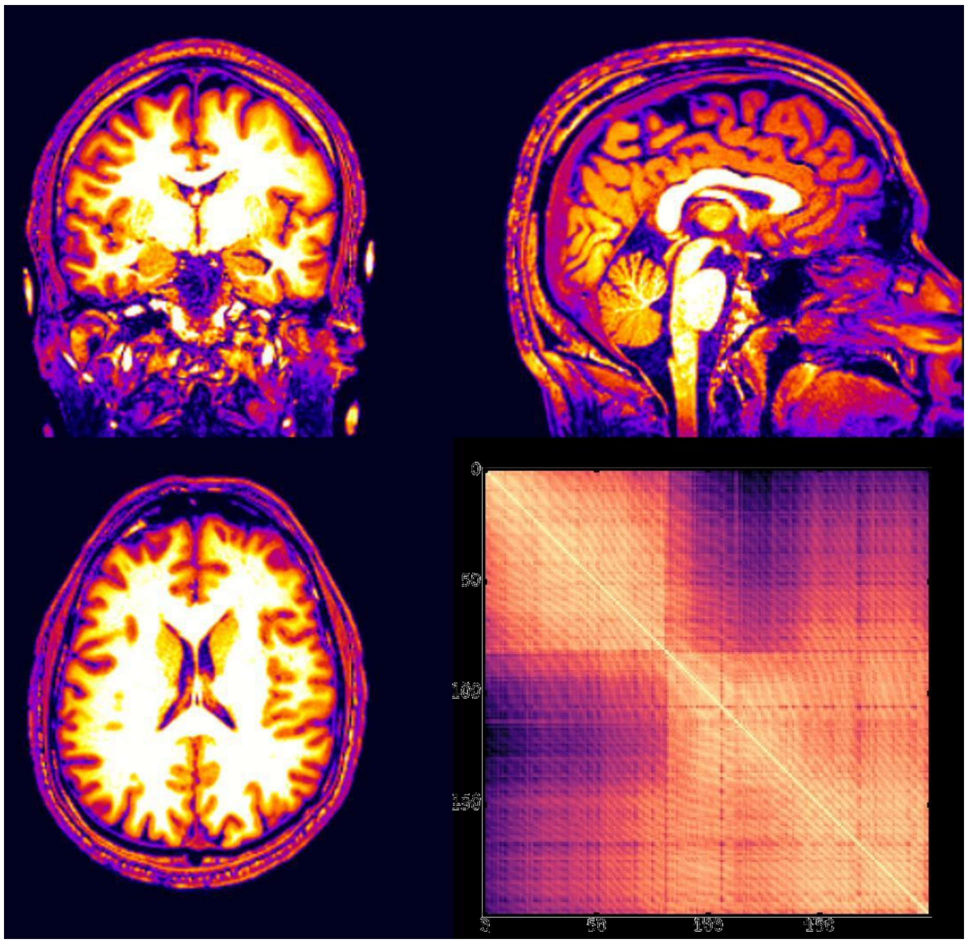


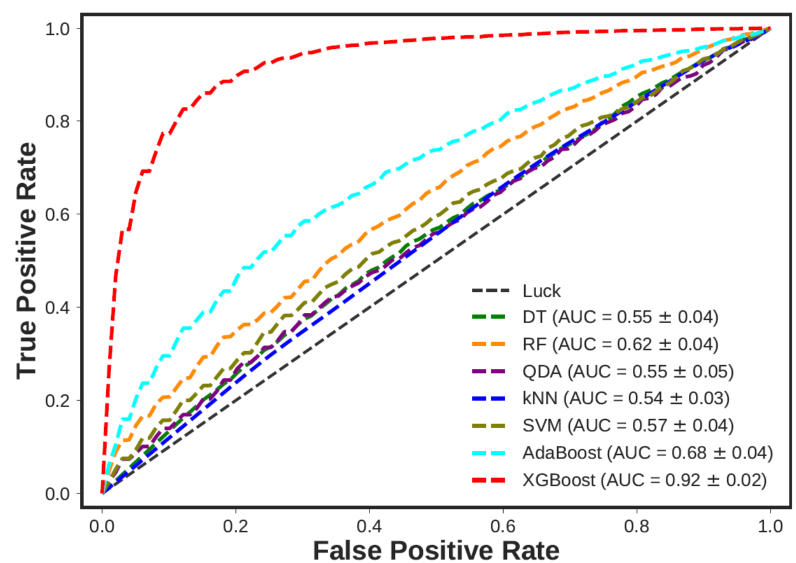
Data Mining Tools



Data Mining

After cleaning data
- Feature selection
- Model selection

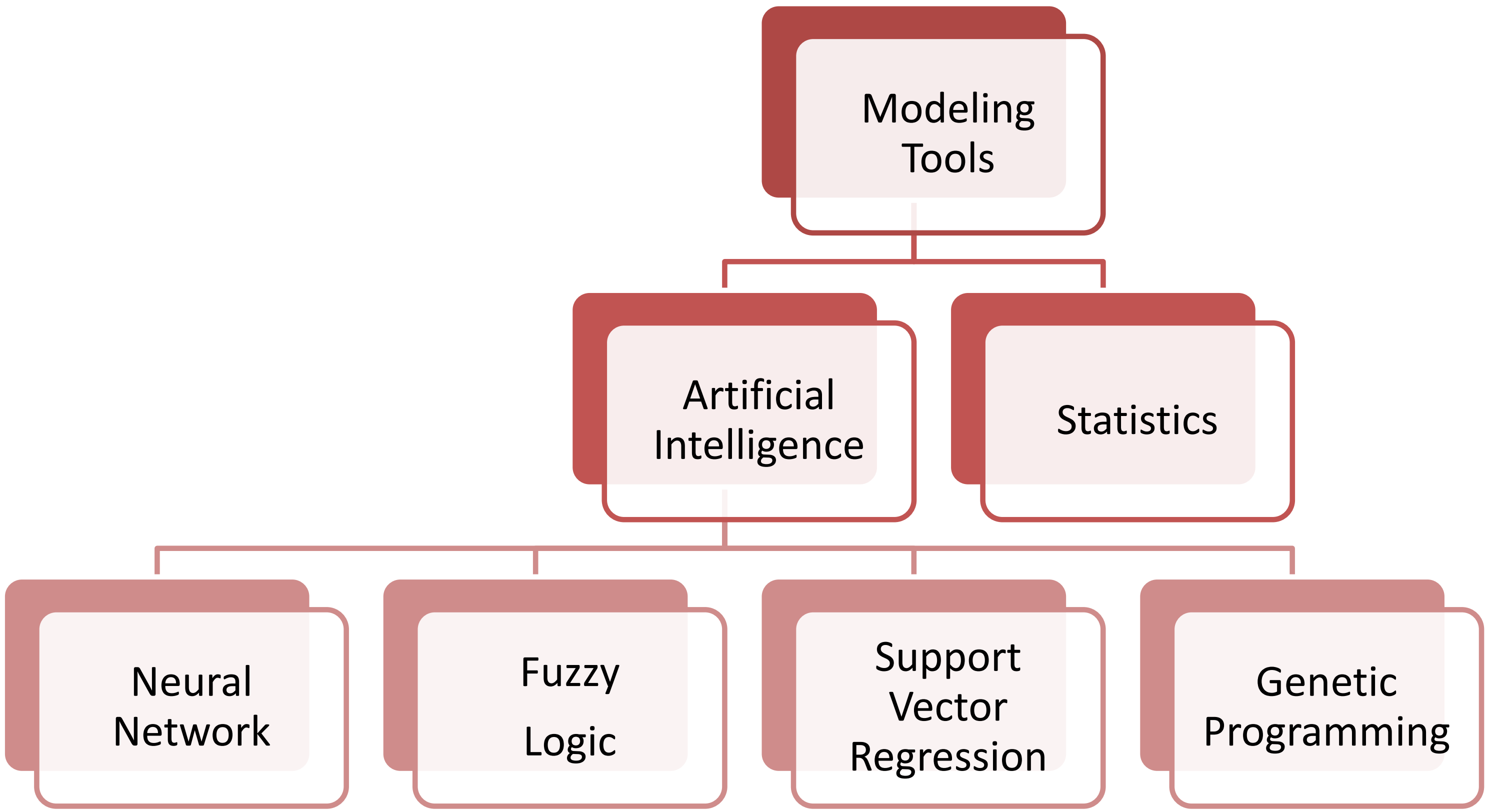


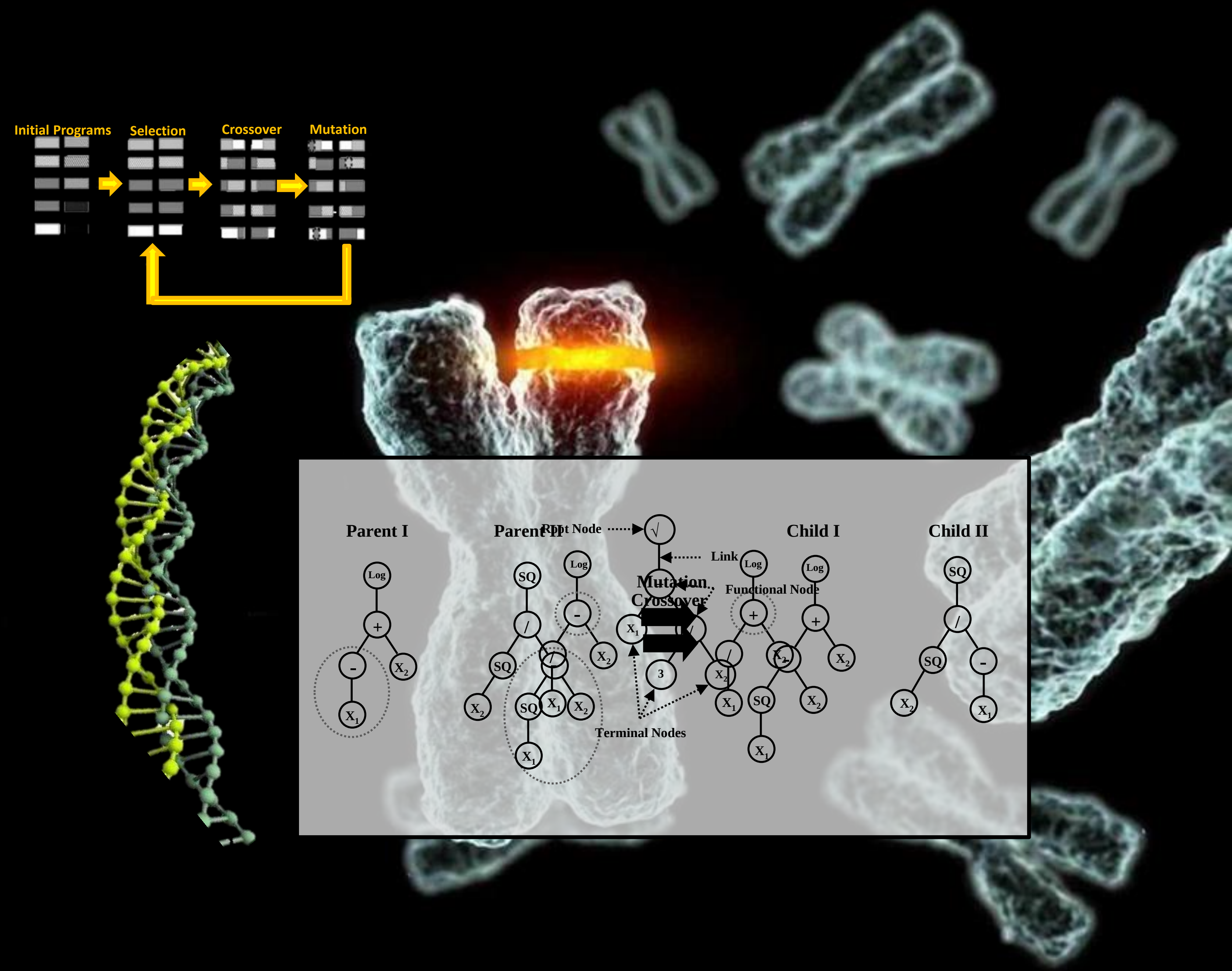


Tahmassebi, A. Gandomi A.H., et al. (2018) “Deep Learning in Medical Imaging: fMRI Big Data Analysis via Convolutional Neural Networks” Proceedings of PEARC18, ACM.

Tahmassebi, A. Gandomi A.H., et al. (2018) “Optimized Naive-Bayes and Decision Tree Approaches for fMRI Smoking Cessation Classification” Complexity, Wiley.

Data Mining Modelling Tools



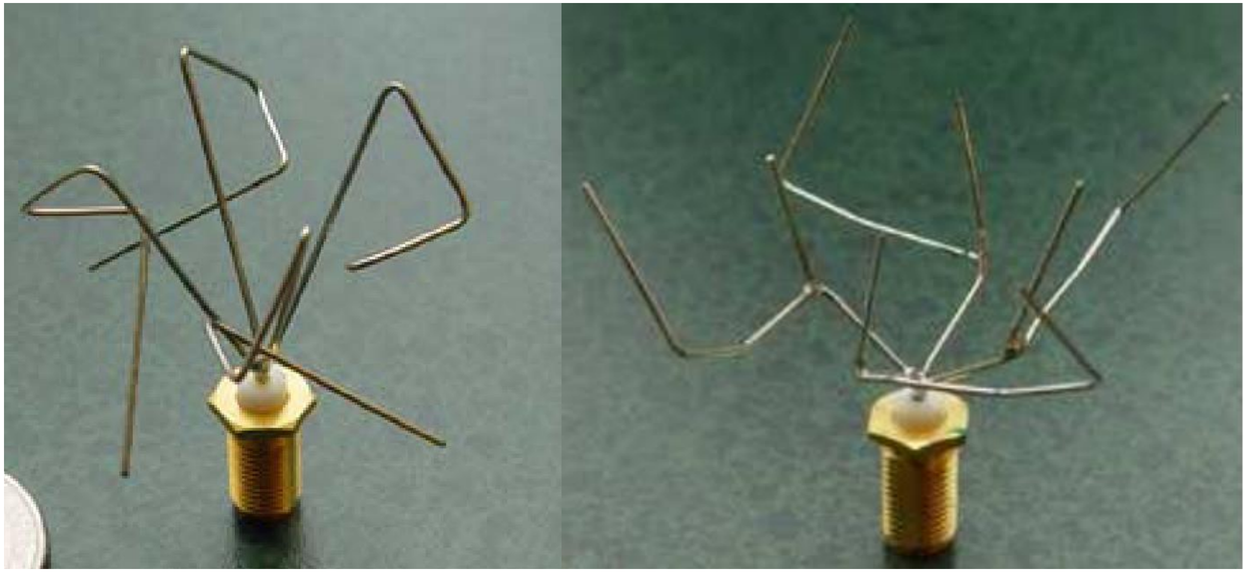

Initial Programs
Selection
Crossover
Mutation
Parent I
Parent II
Root Node
Link
Functional Node
Mutation
Crossover
Terminal Nodes
Child I
Child II
Log
SQ
√

Genetic Programming

- Pre-defined structure is not required
- Pre-processing is not required
- It can model the behaviour without any prior assumptions
- The model is relatively short and simple
- It can select the features
- It has been successfully used for formulation of several complex engineering problems

Gandomi A.H., Roke D.A., “Assessment of Artificial Neural Network and Genetic Programming as Predictive Tools.” Advanced in Engineering Software, Elsevier, 88, 63-72, 2015.

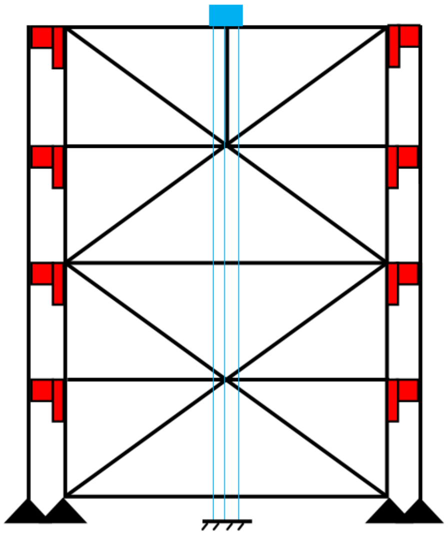
NASA Communication antennas on ST-5 mission



Genetic Programming
2006

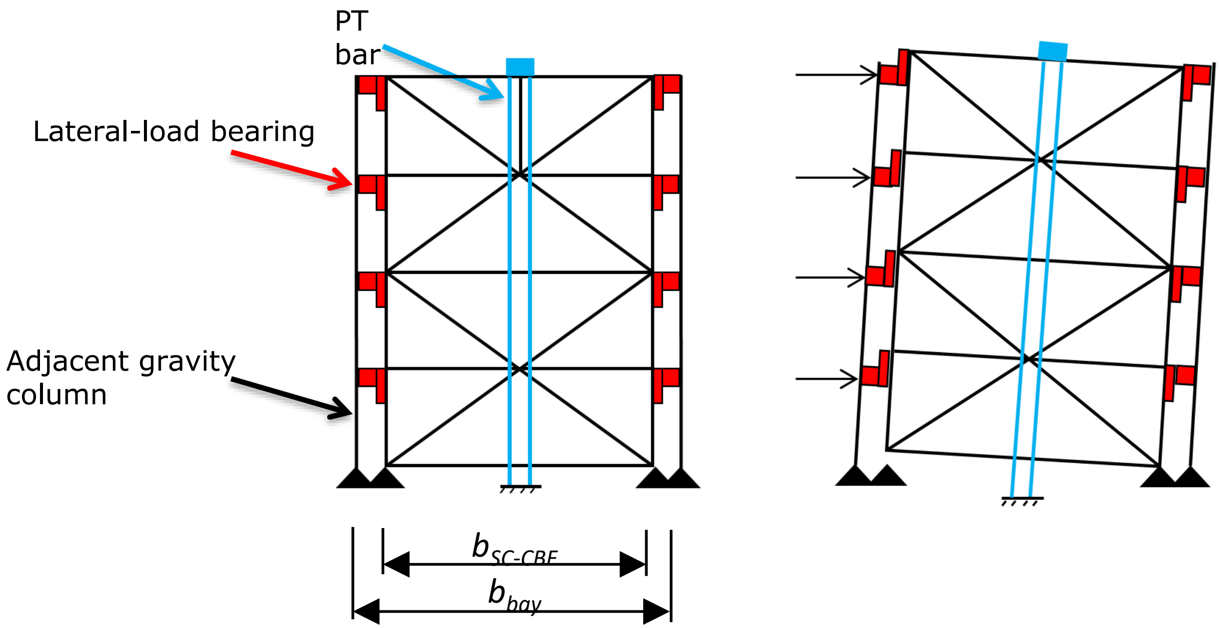
Jason D. Lohn, Gregory S. Hornby and Derek S. Linden, “Human-competitive evolved antennas”, Artificial Intelligence for Engineering Design, Analysis and Manufacturing, volume 22, issue 3, pages 235–247 (2008).

Ex. I.1: Statistical Parameters of a Structure Response

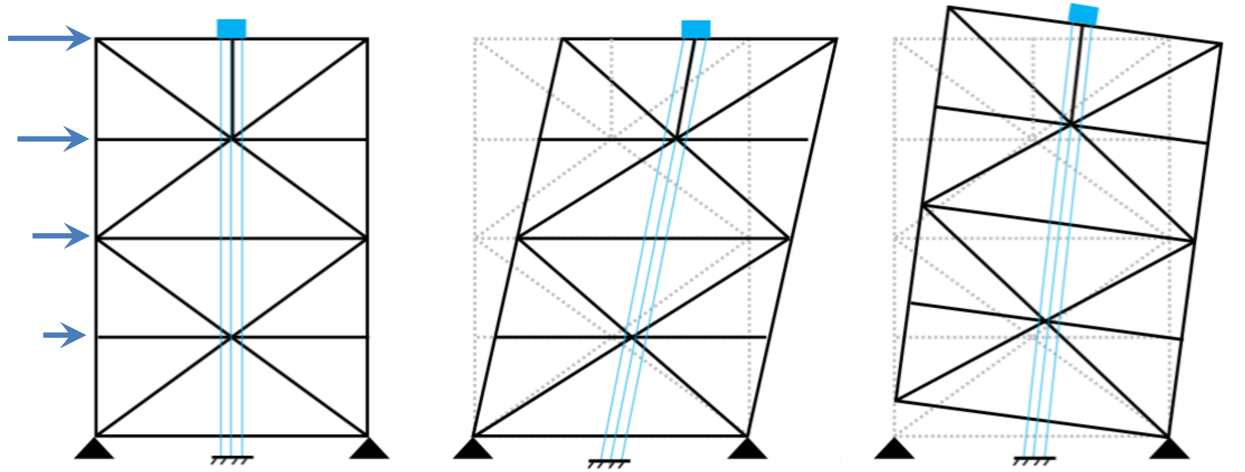


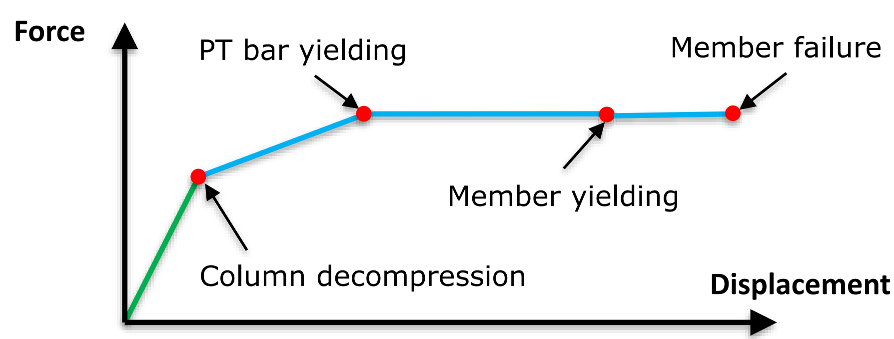
Gandomi A.H., “Seismic Response Formulation of Self-Centering Concentrically Braced Frames Using Genetic Programming” 2014 IEEE Symposium on Computational Intelligence, Orlando, FL, December 9-12, 2014.

SC-CBF Features



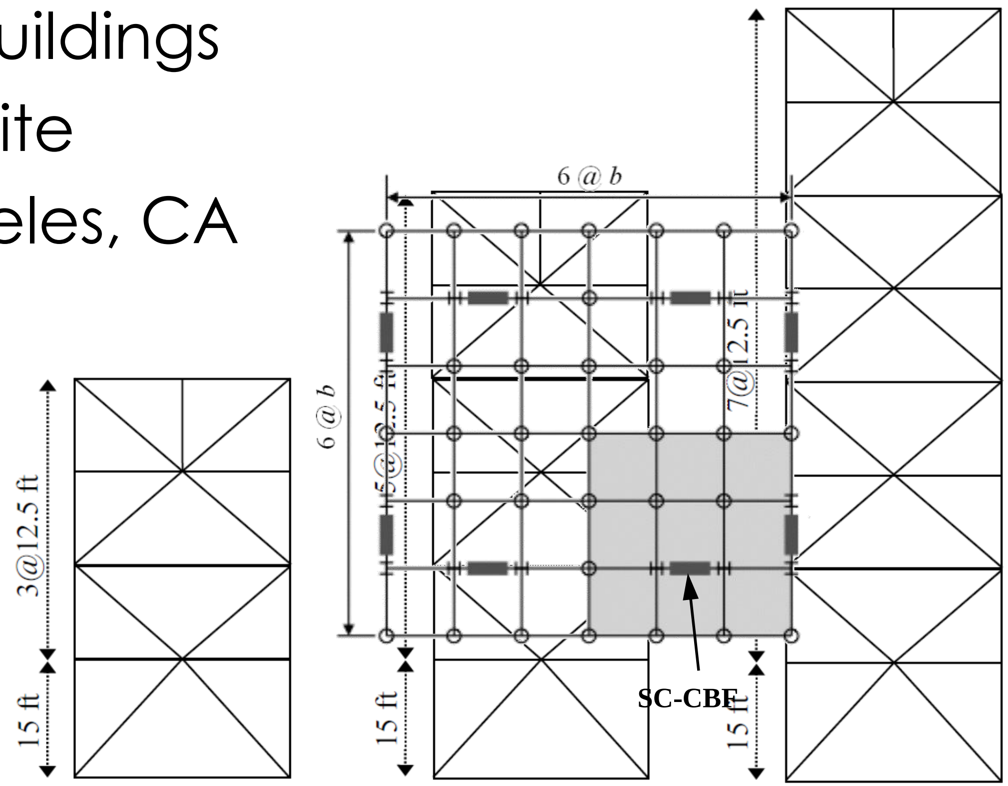
Rocking behavior of the SC-CBF





SC-CBF Parameters

- Office buildings
- Stiff soil site
- Los Angeles, CA



SC-CBF Parameters

Geometrical Parameters:

- Bay width (b)
- Building height (h)

Mechanical Parameters:

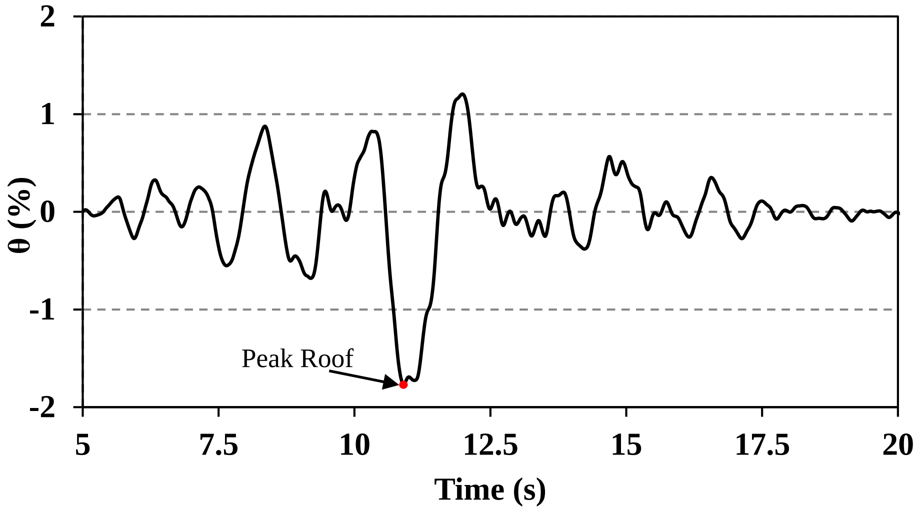
- Coefficient of friction (μ)
- Yield stress of members (Fy)

Geometrical		Mechanical	
b, ft (m)	*h, ft (m)*	*F_y, ksi (MPa)*	*μ*
22.5 (6.9)	52.5 (16)	36 (248)	0.30
30 (9.1)	77.5 (23.6)	50 (345)	0.45
40 (12.2)	102.5 (31.2)	60 (414)	0.60

Dynamic Analysis

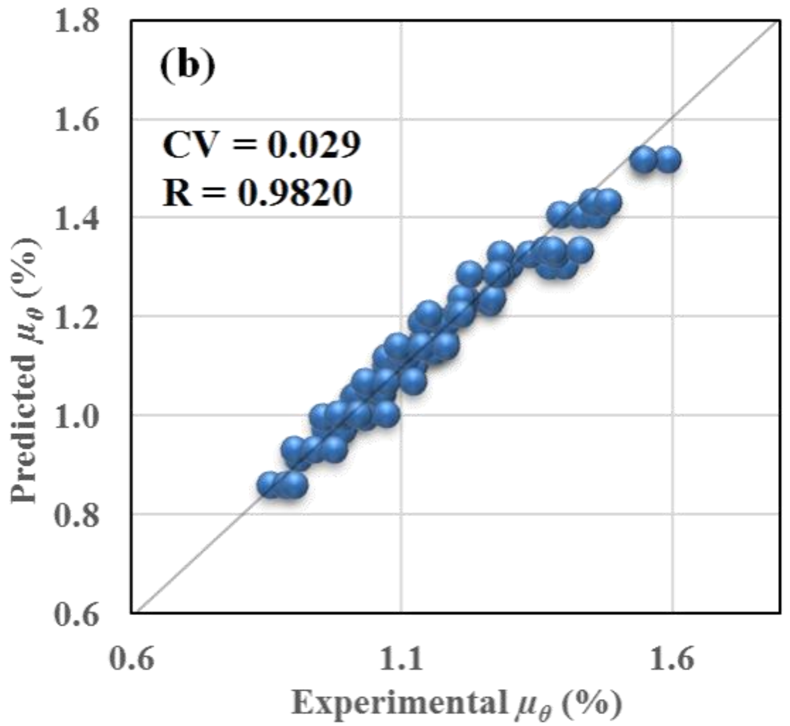
75 different SC-CBF Systems are designed and 50 different earthquake records are applied to each of them

Peak roof drifts are collected, $\theta=\Delta_{max}/H$

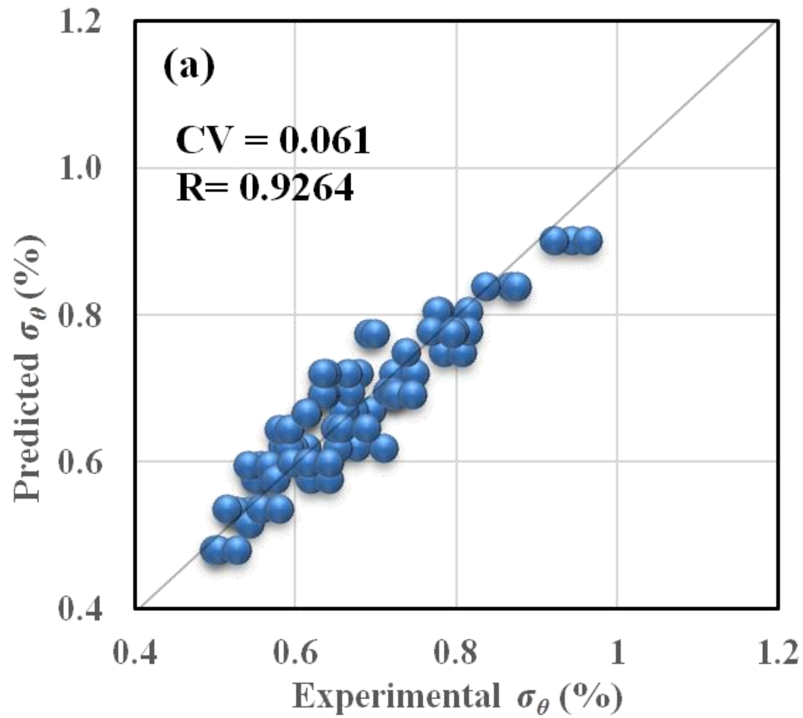


Modelling of Statistical Parameters

- μ_θ

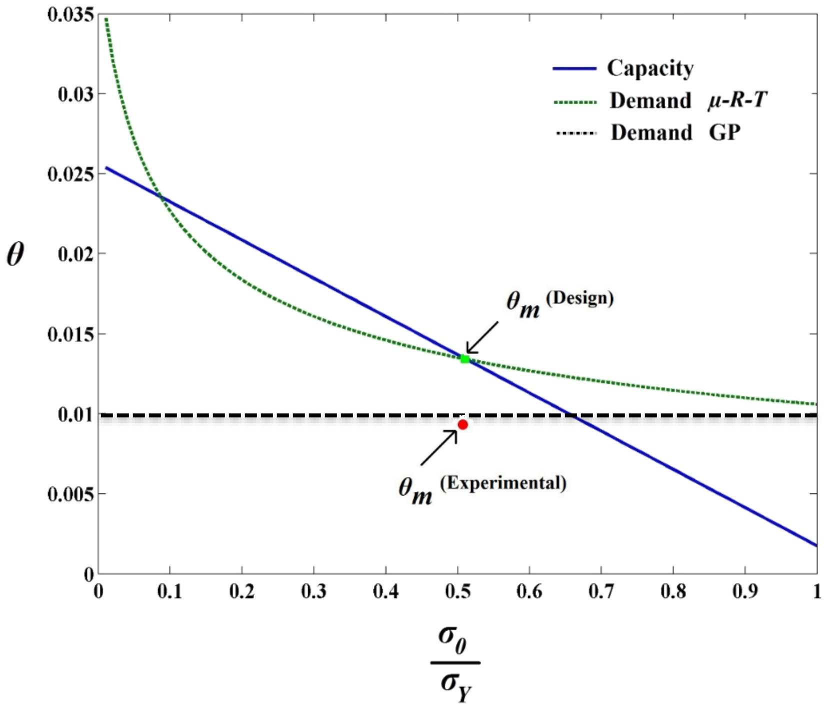


- σ_θ

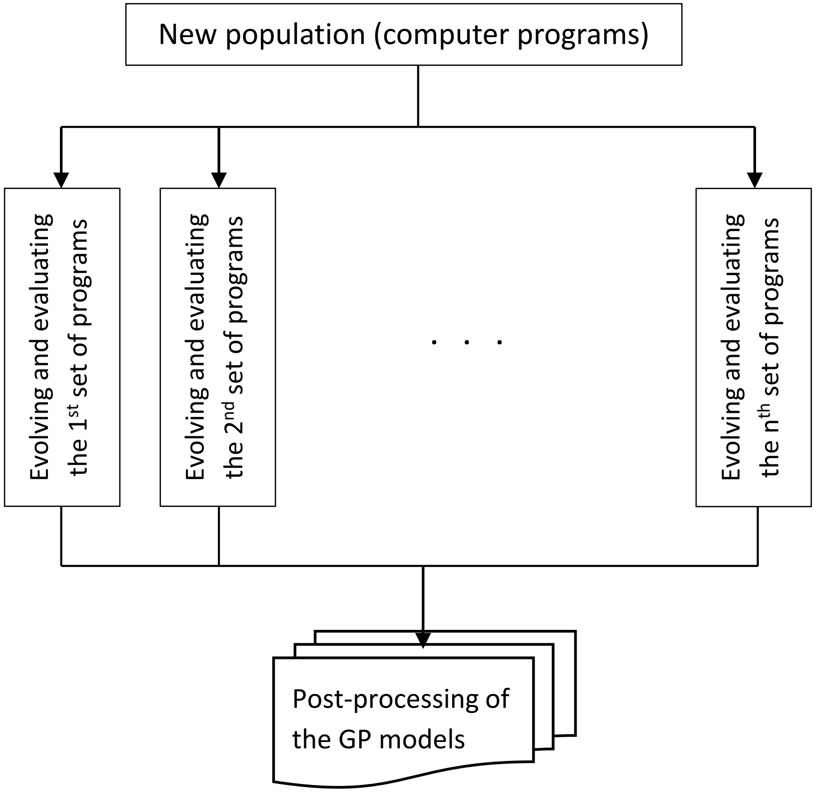


Comparison

- Median value of roof drift
  - μ-R-T (Seo 2005)
  - GP



Parallel Processing in Genetic Programming



Gandomi et al., "Genetic Programming for Experimental Big-Data Mining: A Case Study on Concrete Creep Formulation." Automation in Construction, Elsevier, 70, 89-97, 2016.

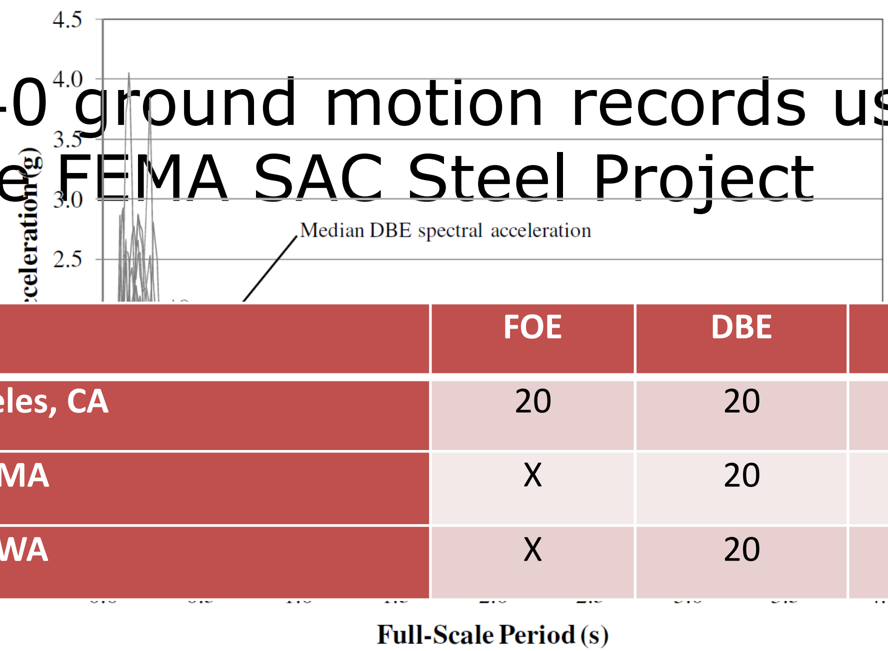
Ex. I.2: Formulation of each Record's Response

- *θ = f(Structural Design, Intensity Measures)*
  - *Structural Design*
  - *Intensity Measures:*

IM	ID	IM	ID
Elastic spectral acceleration	S_a	Cumulative absolute velocity	CAV
Elastic spectral velocity	S_v	Cumulative absolute displacement	CAD
Elastic spectral displacement	S_d	Arias intensity	I_A
Peak ground acceleration	PGA	Velocity intensity	I_v
Peak ground velocity	PGV	Root mean square acceleration	A_{rms}
Peak ground displacement	PGD	Characteristic intensity	I_c

Nonlinear Dynamic Analysis

1) 30 earthquake records in DBE level



2) 140 ground motion records used in the FEMA SAC Steel Project

Area	FOE	DBE	MCE
Los Angeles, CA	20	20	20
Boston, MA	X	20	20
Seattle, WA	X	20	20

Feature Selection: Evolutionary Coefficient

- Best correlation coefficient (R)!
- R: linear relationship

$$R_e = \frac{\sum_{i=1}^{n}\left(y_i - \overline{y_i}\right)\left(f_{j,GP}\left(x_{ij}\right) - \overline{f_{j,GP}\left(x_{ij}\right)}\right)}{\sqrt{\sum_{i=1}^{n}\left(y_i - \overline{y_i}\right)^2 \sum_{i=1}^{n}\left(f_{j,GP}\left(x_{ij}\right) - \overline{f_{j,GP}\left(x_{ij}\right)}\right)^2}}$$

- $f_{j,GP}$: Transformed and correlated x_j

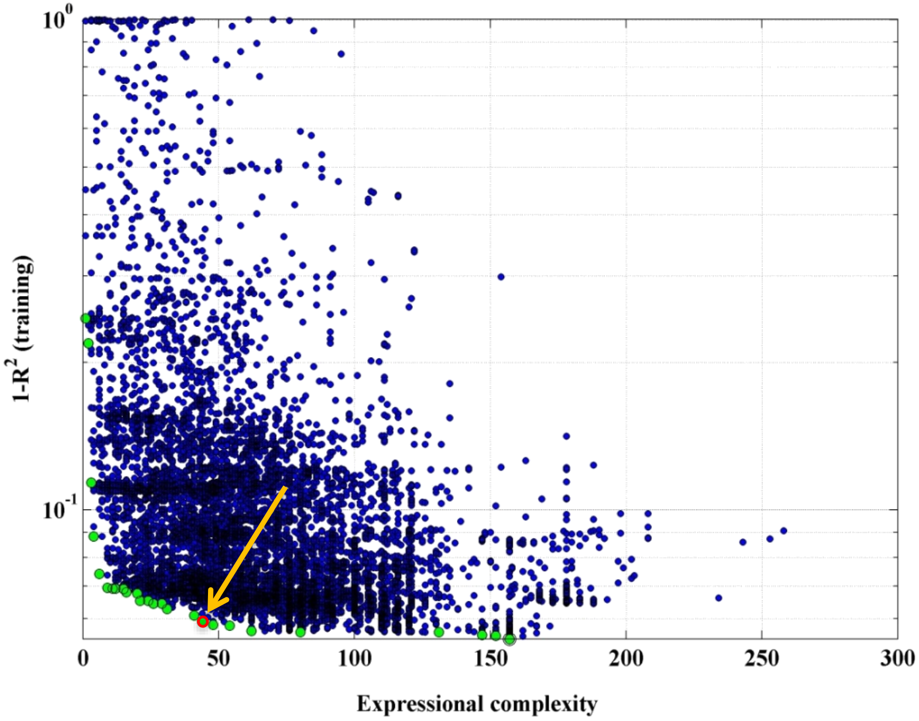
Gandomi A.H., "Seismic Response Formulation of Self-Centering Concentrically Braced Frames Using Genetic Programming" 2014 IEEE Symposium on Computational Intelligence, Orlando, FL, December 9-12, 2014.

Feature Selection: Evolutionary Coefficient

IM	ID	R_e^2	Rank
Elastic spectral acceleration	$S_a(T)$	0.7975	3
Elastic spectral acceleration	$S_a(2T)$	0.8680	2
Elastic spectral velocity	S_v	0.7938	4
Elastic spectral displacement	S_d	0.7761	5
Peak ground acceleration	PGA	0.5359	10
Peak ground velocity	PGV	0.9022	1
Peak ground displacement	PGD	0.7222	6
Cumulative absolute velocity	CAV	0.5694	11
Cumulative absolute displacement	CAD	0.6729	7
Arias intensity	I_A	0.6612	8
Velocity intensity	I_v	0.6454	9
Root mean square acceleration	A_{rms}	0.3235	13
Characteristic intensity	I_c	0.3305	12
Strong ground motion duration	T_D	0.0881	14

Formulation of each Record's Response

- *Multi-Objective Strategy*



Formulation of each Record's Response

- *Single-Objective Strategy*

$$Ln(\theta) = 1.925Ln\left|Ln\left(5.146\frac{PGV \cdot T_{n,1}}{h}\tanh\left(\frac{S_a(2T)}{g}\right)\right)\right| +$$

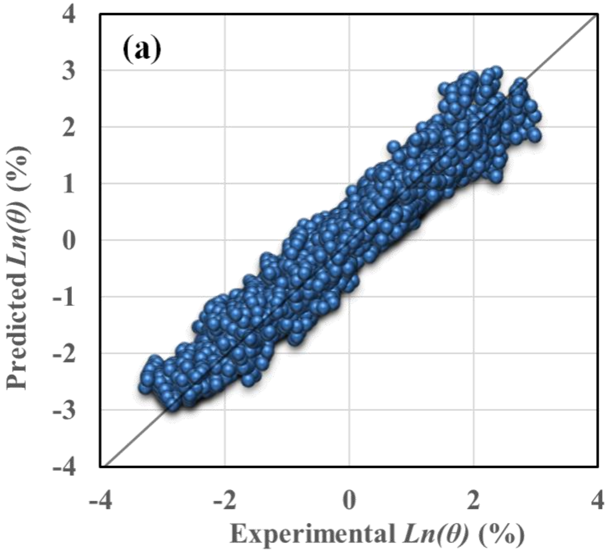
$$0.29Ln\left(\left(\frac{h}{b} + S_a(2T)\right)(1.176 - \mu)\right) - 0.37\left|Ln(S_a(T))\right| + 2.35$$

- *Multi-Objective Strategy*

$$Ln(\theta) = 25.9PGV + 0.615\ln\left|\tanh(2S_a(T))\left(S_a(2T) + \left(\frac{h}{b}\right)^2\right)\sqrt{F_y}\right| - 1.08$$

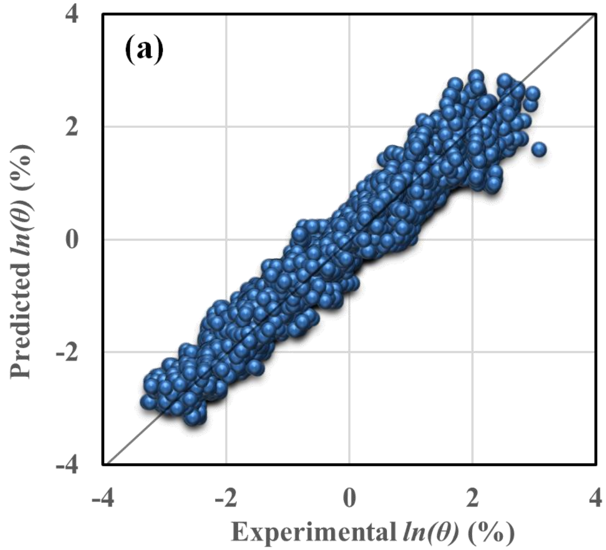
Prediction Results

Single-Objective



R= 0.9709

Multi-Objective



R=0.9700

Multi-stage genetic programming

- $f(X) = f_1(x_1) + f_2(x_2) + \cdots + f_n(x_n) + f_{int}(X) = \sum_{i=1}^{n} f_i(x_i) + f_{int}(X)$
  - $f_2(x_2) = f(X) - f_1(x_1)$
  - $f_3(x_3) = f(X) - f_1(x_1) - f_2(x_2)$

    $\vdots$

  - $f_n(x_n) = f(X) - f_1(x_1) - f_2(x_2) - \cdots - f_{n-1}(x_{n-1})$
  - $f_{int}(X) = f(X) - \sum_{i=1}^{n} f_i(x_i)$

Gandomi A.H., Alavi A.H., “Multi-Stage Genetic Programming: A New Strategy to Nonlinear System Modeling.” Information Sciences, Elsevier, 181(23): 5227-5239, 2011.

MSGP for Classification: Soil Liquefaction modelling

Stage 1:

$$F_1 = \cos((\arctan(((q_c^2 + 8.372)/(q_c - 8.372)))^2))^3$$

Stage 2:

$$F_2 = (-R_f + 1.393)/(R_f + (5.281/R_f))$$

Stage 3:

$$F_3 = \sin\left(-8.297\sigma_v'^2 + -6.012 - \sigma'_v\right)/ - 8.297$$

Stage 4:

$$F_4 == ((\arctan(\cos((\sigma_v^3)))^2)(\arctan(\cos(\sigma_v))^3))$$

Stage 5:

$$F_5 = \arctan(\arctan(\arctan(((0.0102 - 0.1011/a_{\max}) + 0.466)))$$

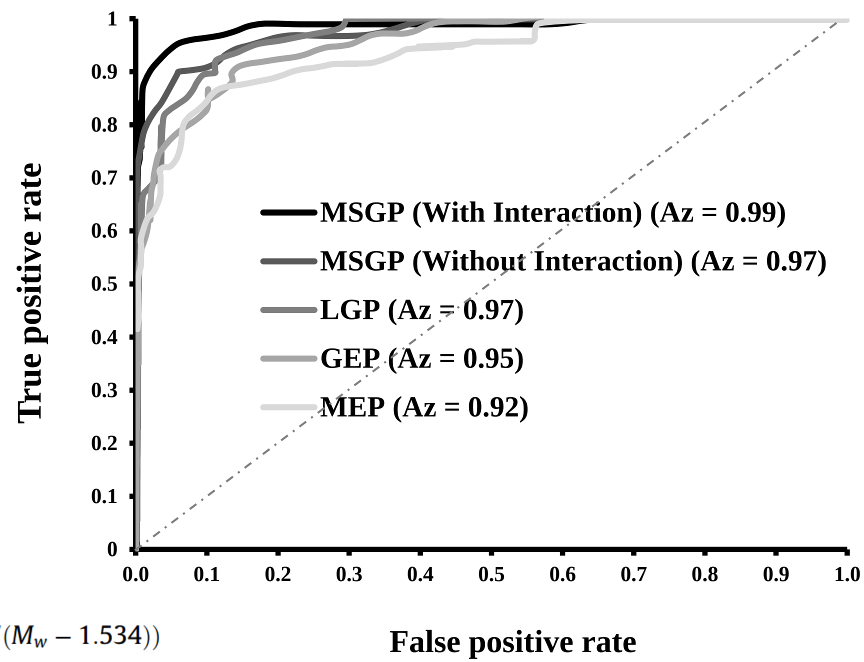
Stage 6:

$$F_6 = 0.034\sin\left(\left(\left(M_w^2\right) - (1.589 - M_w)\right)\right)$$

Stage 7 (interaction):

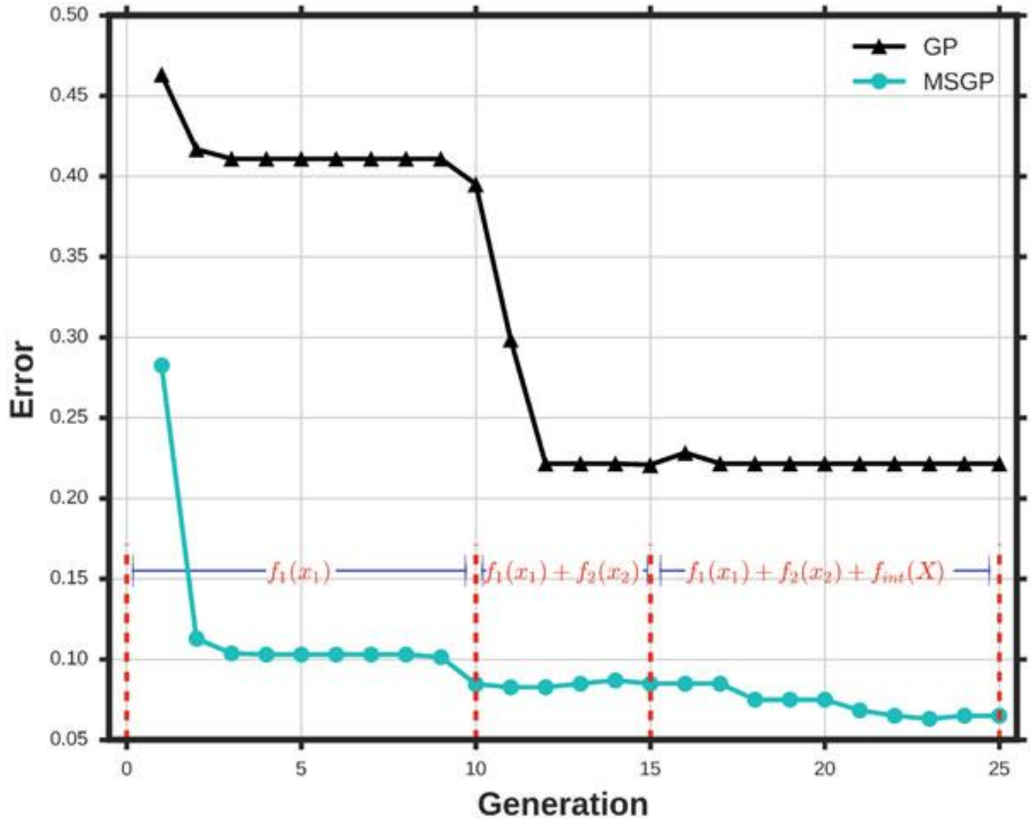
$$F_{\text{int}} = (\cos((((((\cos(M_w)a_{\max})(1.534 - M_w + 5.936))\exp(M_w))^3)/a_{\max}))/(M_w - 1.534))$$

Figure: ROC curves



Gandomi A.H., Alavi A.H., "Multi-Stage Genetic Programming: A New Strategy to Nonlinear System Modeling." Information Sciences, Elsevier, 181(23): 5227-5239, 2011.

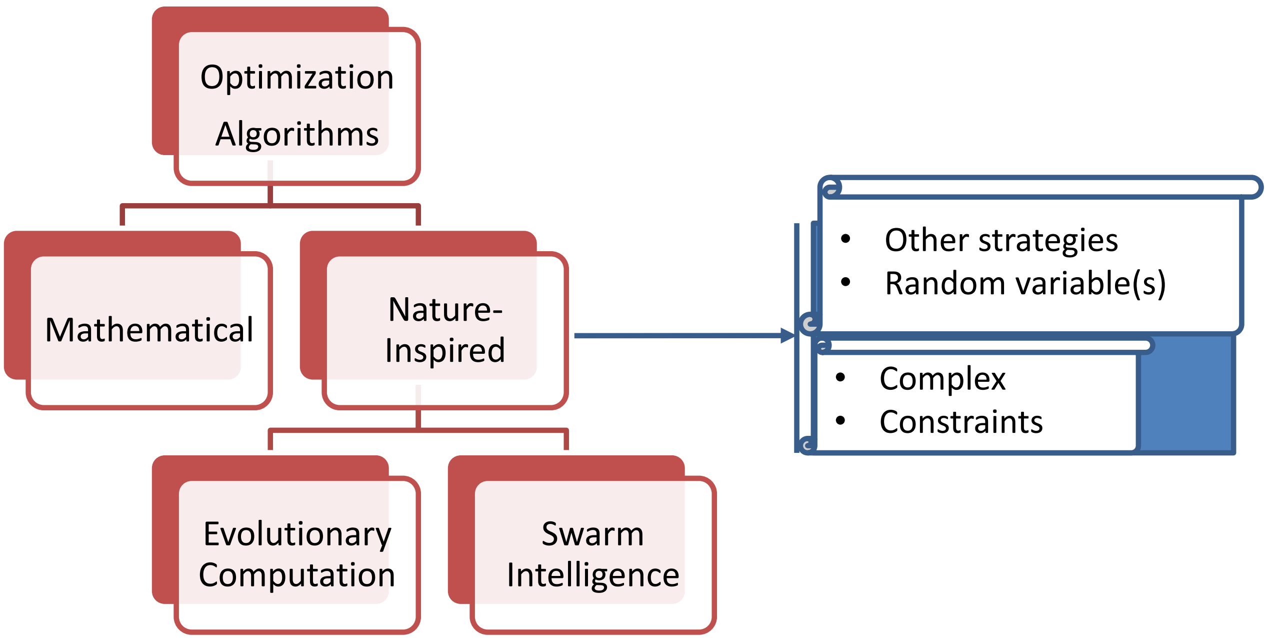
MSGP for Big Data



Tahmassebi, A. and Gandomi, A.H., 2018. Genetic programming based on error decomposition: A big data approach. In *Genetic Programming Theory and Practice XV*(pp. 135-147). Springer, Cham.

Optimization

Optimization Algorithms



Some Successful Cases: Boeing Turbine geometry of 777 GE engine

Genetic Algorithm
1998

Charles W. Petit, “Touched by nature: putting evolution to work on the assembly line.” US News & World Report, volume 125, issue 4, pages 43–45 (1998).

Some Successful Cases: Hitachi Nose cone for N700 bullet train

Genetic Algorithm
2005

Takenori Wajima, Masakazu Matsumoto and Shinichi Sekino, “Latest System Technologies for Railway Electric Cars”, Hitachi Review, 54(4), 161–168 (2005).

Nature-Inspired Algorithms

- Traditional Algorithms
  - Simulated Annealing (SA)
  - Genetic Algorithm (GA)
  - Particle Swarm Optimization (PSO)

- Recent Algorithms
  - Firefly Algorithm (FA)
  - Krill Herd Algorithm (KH)
  - Interior Search Algorithm (ISA)
  - Parameter-less Population Pyramid (P3)

Nature-Inspired Algorithms

- Traditional Algorithms
  - Simulated Annealing (SA)
  - Genetic Algorithm (GA)
  - Particle Swarm Optimization (PSO)

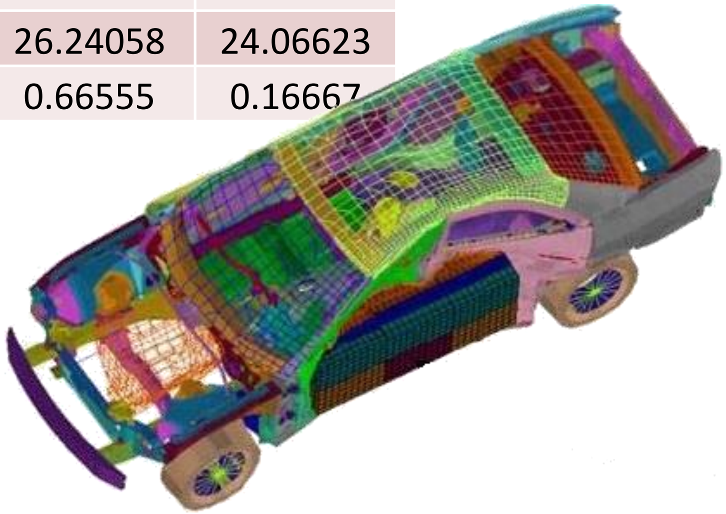
- Recent Algorithms
  - Firefly Algorithm (FA)
  - **Krill Herd Algorithm (KH)**
  - **Interior Search Algorithm (ISA)**
  - Parameter-less Population Pyramid (P3)

$$x_i^{t+1} = x_i^t + \Delta x_i$$

$$\Delta x_i = \beta_0 e^{-\gamma r^2}\left(x_j^t - x_i^t\right) + \alpha\varepsilon$$

Ex. II.1: Car side impact design

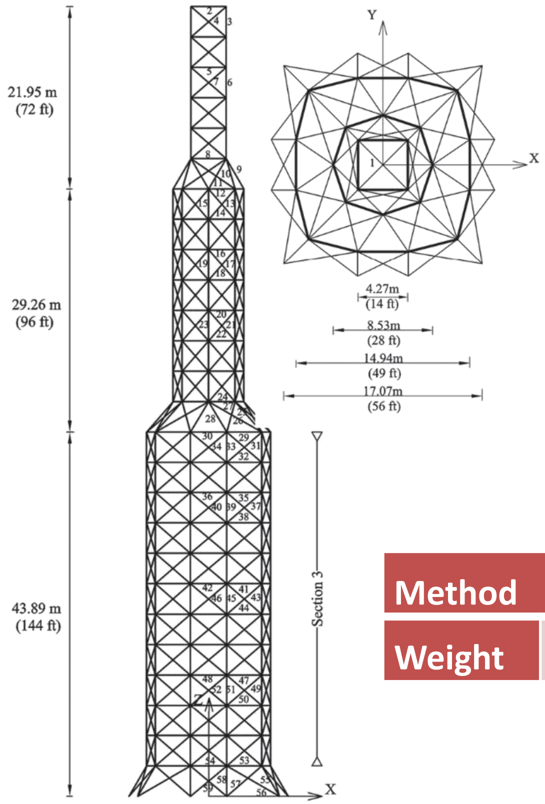
Method	PSO	GA	FA
Best Objective	22.84474	22.85653	22.84298
Mean Objective	22.89429	23.51585	22.89376
Worst objective	23.21354	26.24058	24.06623
Std. Dev.	0.15017	0.66555	0.16667

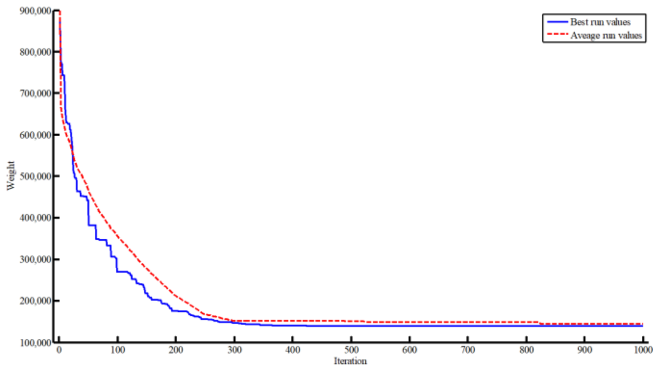


Gandomi A.H., Yang X.S., Alavi A.H., “Mixed Variable Structural Optimization Using Firefly Algorithm.” Computers and Structures, Elsevier, 89(23-24): 2325-2336, 2011.

[The 11th most popular articles in 2013 in Computational Engineering - ELSEVIER]

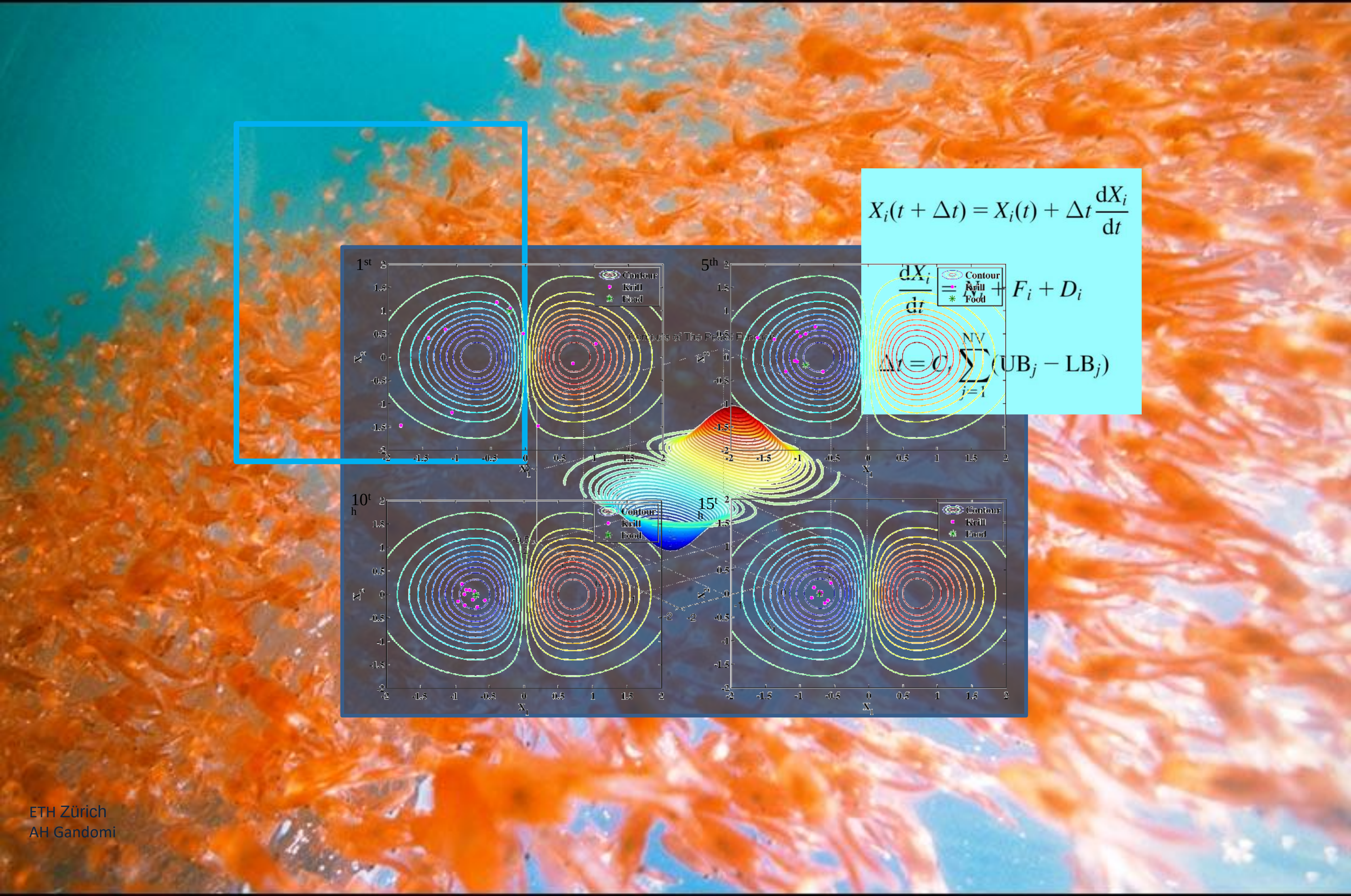
Ex. II.3: 942-bar Tower





Method		SA	ES	FA	CS
Weight	(lb)	143,436	141,241	138,878	134,120

Gandomi A.H., Talatahari S., Yang X.S., Deb S., "Design optimization of truss structures using cuckoo search algorithm." The Structural Design of Tall and Special Buildings, Wiley, 22(17), 1330–1349, 2013.
the journal's most cited article in 2012-2013



$$X_i(t + \Delta t) = X_i(t) + \Delta t \frac{dX_i}{dt}$$

$$\frac{dX_i}{dt} = N_i + F_i + D_i$$

$$\Delta t = C_t \sum_{j=1}^{NV} (UB_j - LB_j)$$

ETH Zürich
AH Gandomi

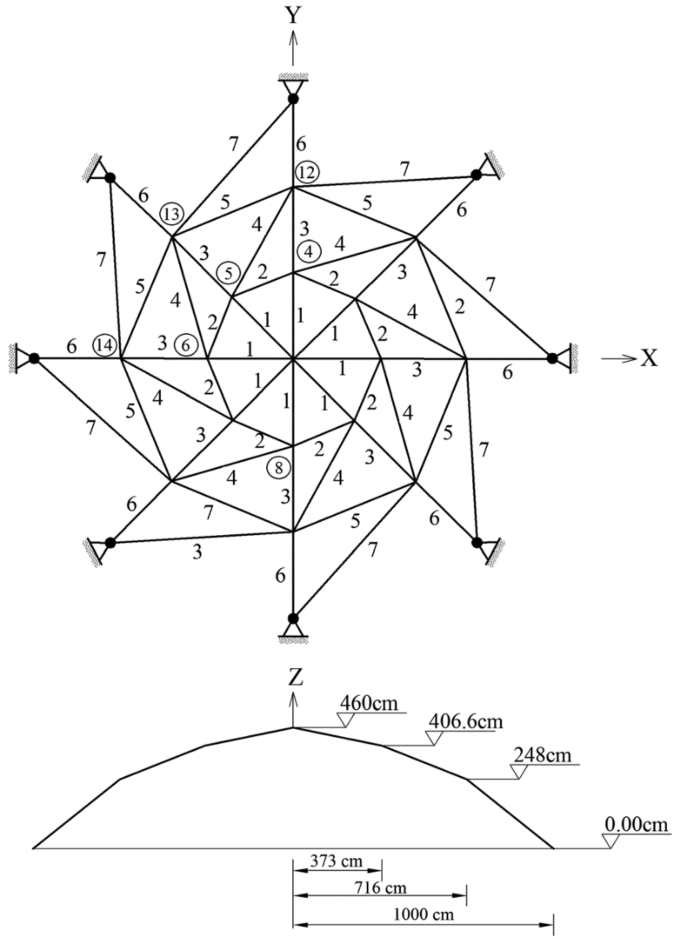
Normalized statistical results of KH algorithms and GA, ES, BBO, ACO, DE, HDE, PSO, APSO for the benchmark problems.

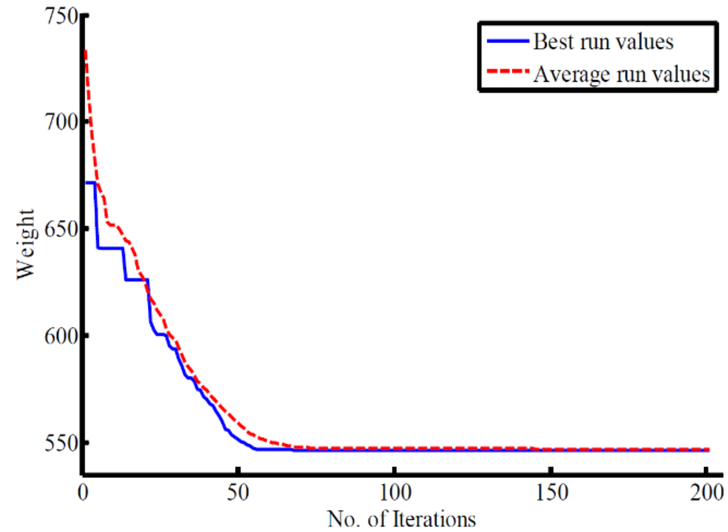
ID	KH I	KH II	KH III	KH IV	GA	ES	BBO	ACO	DE	HDE	PSO	APSO
F1	0.995	1.000	0.987	0.999	0.964	0.000	0.989	0.395	0.838	0.998	0.390	1.000
F2	0.989	1.000	0.983	1.000	0.000	0.859	0.906	0.556	0.484	1.000	0.973	1.000
F3	1.000	1.000	1.000	1.000	0.000	0.306	0.967	0.325	0.780	0.997	0.891	1.000
F4	0.968	1.000	0.933	0.969	0.734	0.858	0.940	0.808	0.909	0.301	0.000	0.798
F5	0.865	0.934	0.596	0.892	0.905	0.000	1.000	0.697	0.925	0.958	0.998	0.943
F6	0.320	1.000	0.654	0.999	0.973	0.993	0.990	0.986	0.984	0.000	1.000	0.452
F7	0.989	0.995	0.969	0.995	0.683	0.000	1.000	0.831	0.902	1.000	0.933	1.000
F8	0.690	0.852	0.499	0.783	0.176	0.000	1.000	0.660	0.498	0.999	0.198	0.481
F9	0.721	0.918	0.679	0.853	0.976	0.000	1.000	0.479	0.823	1.000	0.999	1.000
F10	0.999	1.000	0.990	1.000	1.000	0.000	1.000	0.989	0.911	1.000	1.000	1.000
F11	1.000	1.000	1.000	1.000	0.666	0.000	0.666	0.666	0.666	1.000	0.666	1.000
F12	1.000	1.000	1.000	1.000	0.993	0.000	1.000	0.997	1.000	1.000	1.000	1.000
F13	1.000	1.000	1.000	1.000	0.797	0.249	0.969	0.722	0.965	1.000	0.000	1.000
F14	1.000	1.000	1.000	1.000	0.968	0.000	0.995	0.744	0.960	1.000	0.974	1.000
F15	1.000	1.000	1.000	1.000	0.000	0.180	0.740	0.520	0.616	1.000	0.667	1.000
F16	0.990	1.000	1.000	1.000	0.209	0.000	0.534	0.299	0.124	1.000	0.282	1.000
F17	0.966	1.000	1.000	1.000	0.315	0.421	0.974	0.000	0.257	1.000	0.885	0.999
F18	1.000	1.000	1.000	1.000	0.984	0.000	1.000	0.822	0.333	0.999	1.000	1.000
F19	0.999	1.000	0.999	0.998	0.993	0.000	0.999	0.877	0.997	0.983	0.999	1.000
F20	1.000	1.000	1.000	1.000	0.991	0.000	1.000	0.998	0.998	0.993	0.999	1.000
$\sum$	18.491	19.697	18.288	19.487	13.325	3.866	18.669	13.373	14.971	18.229	14.854	18.673
Rank	5	1	6	2	11	12	4	10	8	7	9	3

- **KH I : without any genetic operators (KH I);**
- **KH II : with crossover operator (KH II);**
- **KH III : with mutation operator (KH III); and**
- **KH IV: with crossover and mutation operators (KH IV).**

Gandomi A.H., Alavi A.H., “Krill herd: A new bio-inspired optimization algorithm” Commun Nonlinear Sci Numer Simulat 17 (2012) 4831–4845. the journal’s hottest article in 2012, 2013 and 2014

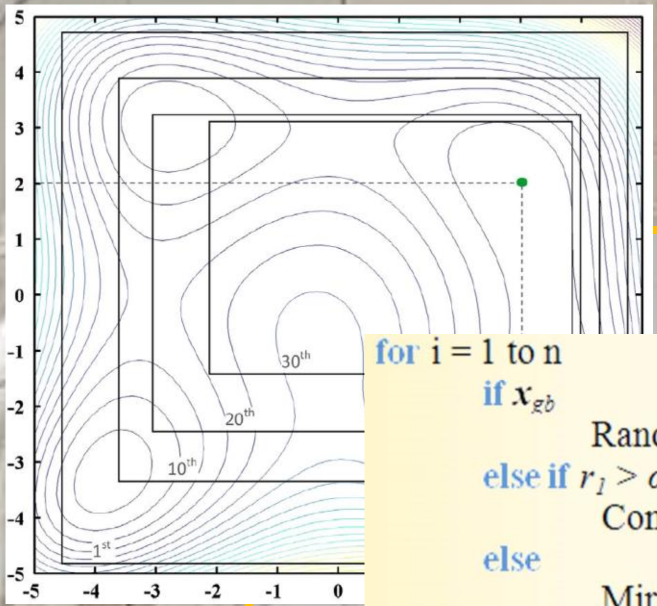
Ex. II.1: Spatial Dome

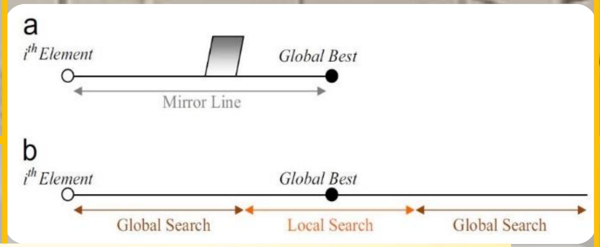




Method	SA	GA	KH
Best	5,461.24	5,948.41	5,436.35
Mean	5,472.32	6,895.06	5,460.09
Std. Dev.	63.51	2,020.46	17.98

Gandomi A.H., Talatahari S., Tadbiri F., Alavi A.H., "Krill herd algorithm for optimum design of truss structures" *International Journal of Bio-Inspired Computation,* 5(5), 281-288, 2013. [Invited Paper]





for i = 1 to n

  if x_{gb}

    Random Walk

  else if $r_1 > \alpha$

    Composition Optimization

  else

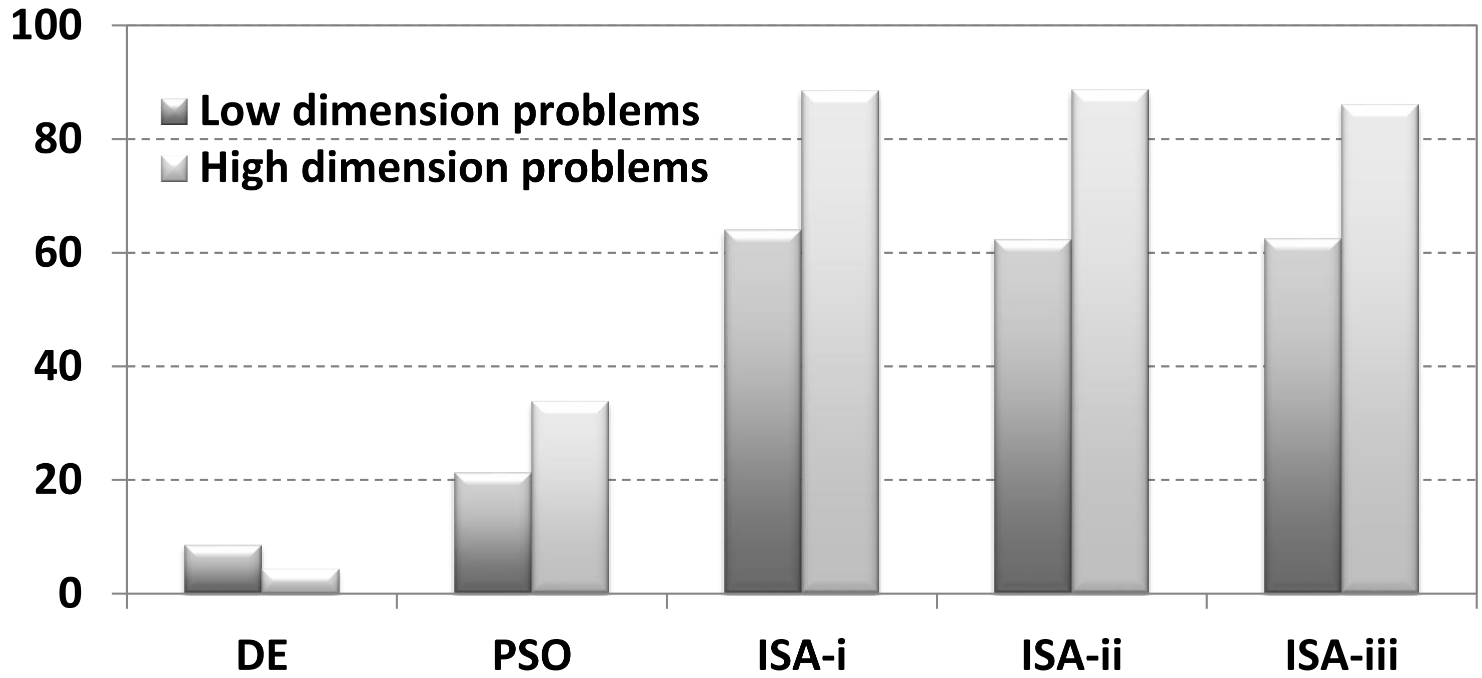
    Mirror Search

  end if

Check the boundaries except for decomposition elements

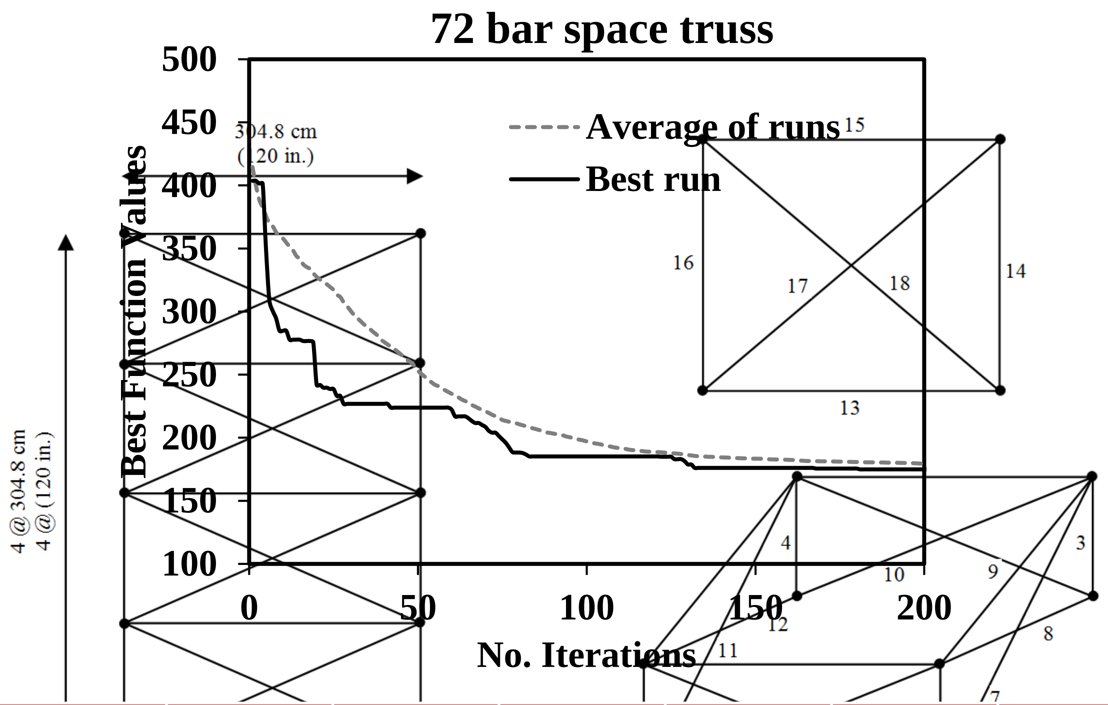
end for

Convergence Rate (%)



Gandomi A.H. “Interior Search Algorithm (ISA): A Novel Approach for Global Optimization.” ISA Transactions, Elsevier, 53(4), 1168–1183, 2014. hot article in 2014

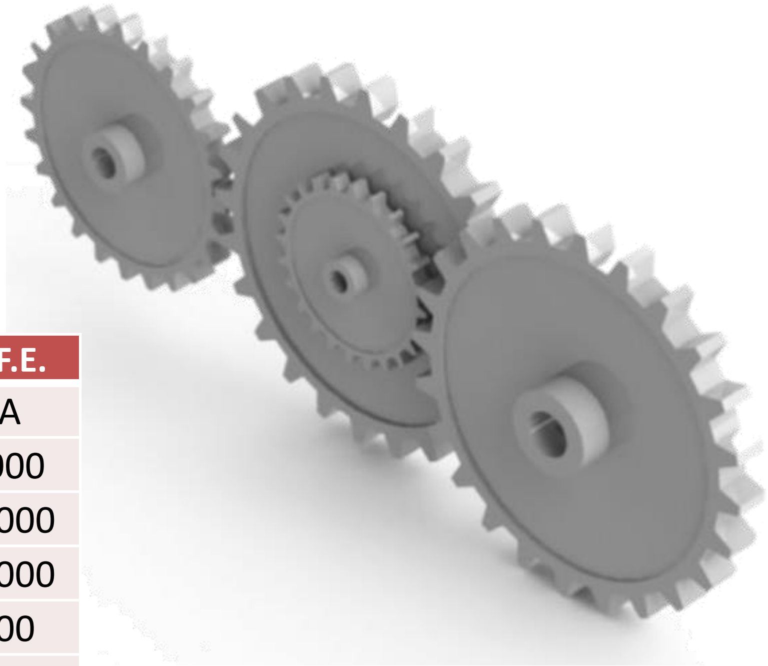
Ex. II.2: 72-bar Truss Structure



Methods	GA	PSO	HPSO	GSO	MHS	ISA
Weight (kg)	181.72	494.32	176.41	438.89	175.97	174.88
No. F.E.	60,000	50,000	50,000	50,000	30,000	10,000

Gandomi A.H. "Interior Search Algorithm (ISA): A Novel Approach for Global Optimization." ISA Transactions, Elsevier, 53(4), 1168–1183, 2014. [hot article in 2014]

Ex. II.3: Gear Train Design

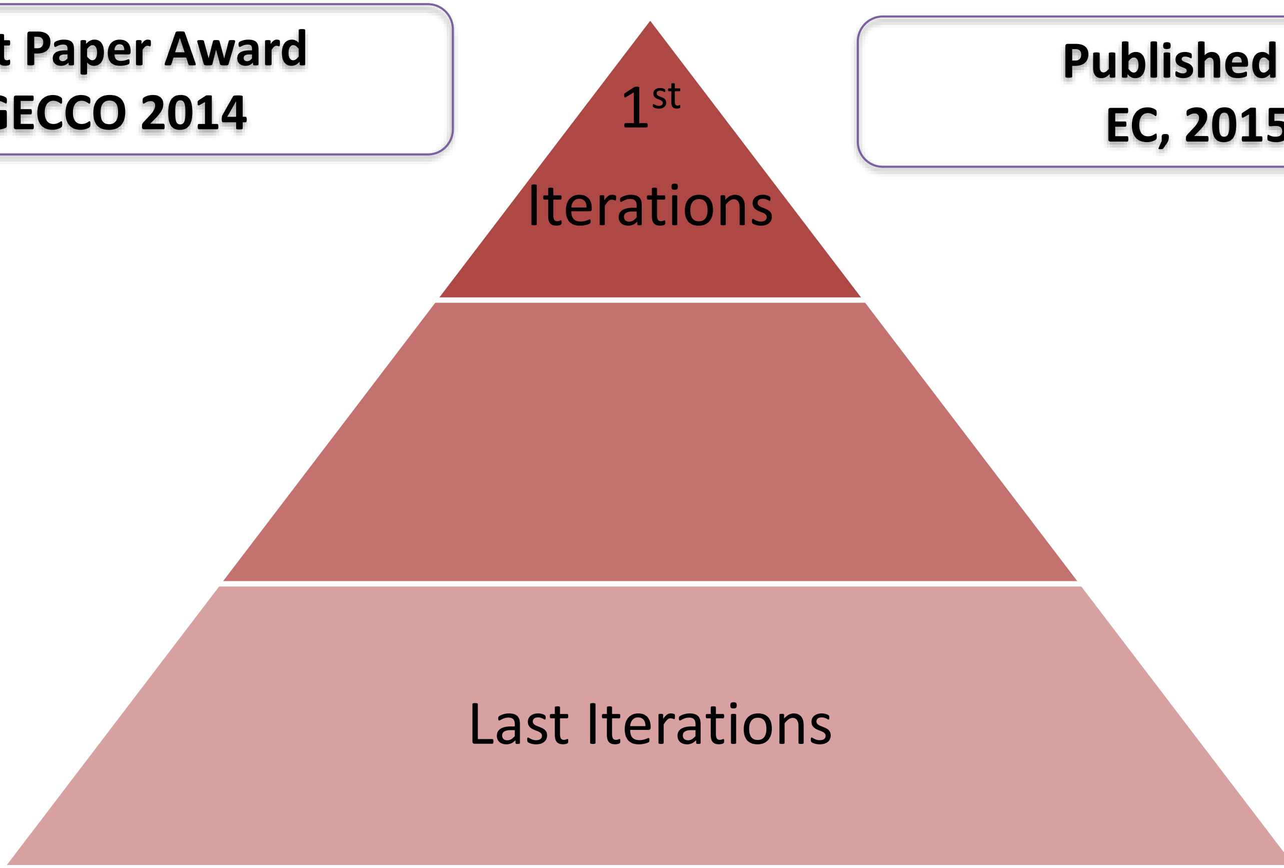


Method	f_{min}	No. F.E.
SA	2.36×10^{-9}	N/A
GA	2.33×10^{-7}	10,000
FLGA	2.701×10^{-12}	100,000
PSO	2.701×10^{-12}	100,000
CPSO	2.701×10^{-12}	2,000
ISA	2.701×10^{-12}	120

Gandomi A.H., Roke D.A., "Engineering Optimization using Interior Search Algorithm" 2014 IEEE Symposium on Computational Intelligence, Orlando, FL, December 9-12, 2014.

Parameter-less Population Pyramid (P3)

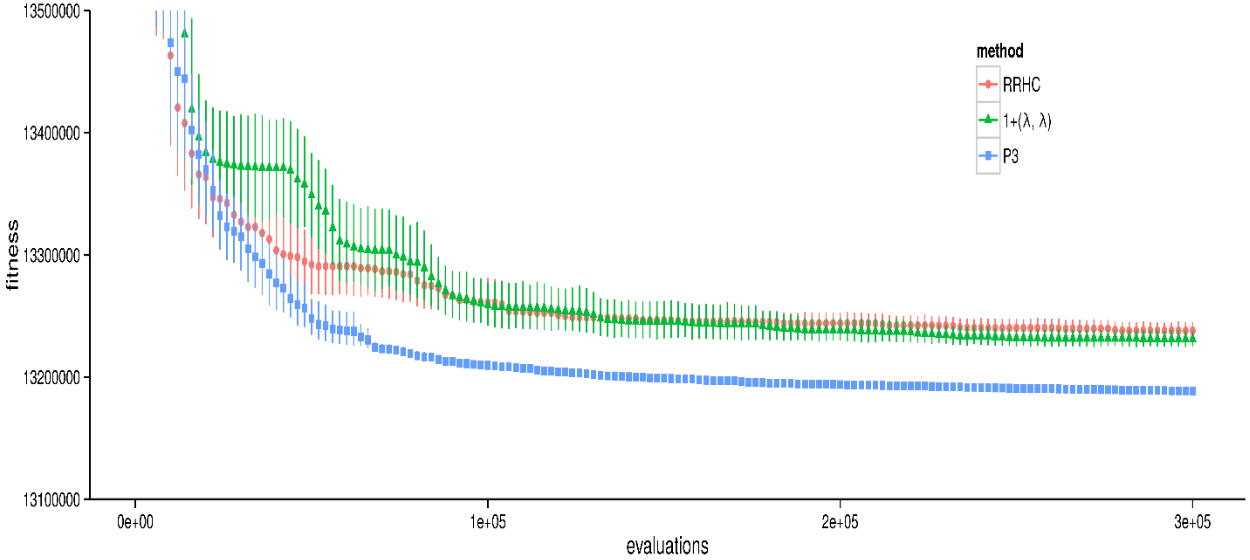
Best Paper Award
GECCO 2014

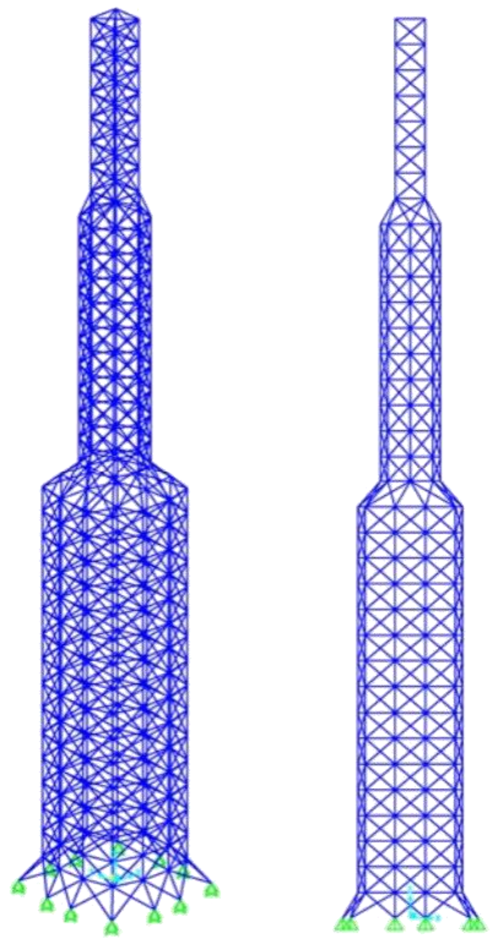


Published at
EC, 2015

Ex. II.4: 35-Storey Space Tower

1262 members and 936 degrees of freedom





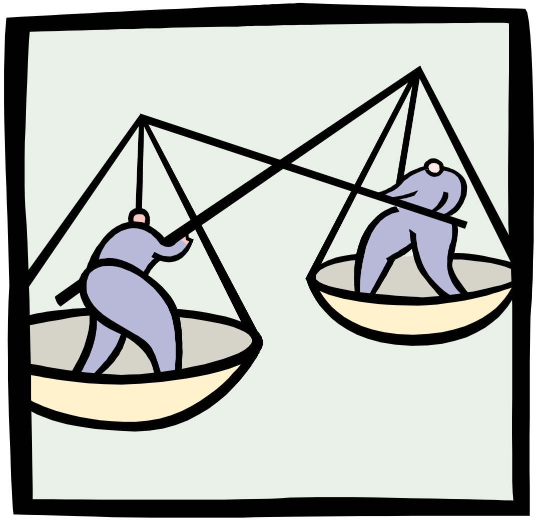
Gandomi, A. H., & Goldman, B. W. Parameter-less population pyramid for large-scale tower optimization. Expert Systems with Applications, 96, 175-184, 2018.

Customization in Optimization

Nature-Inspired Algorithms

- PROs
  – Derivative-free
  – Global
  -- Flexible

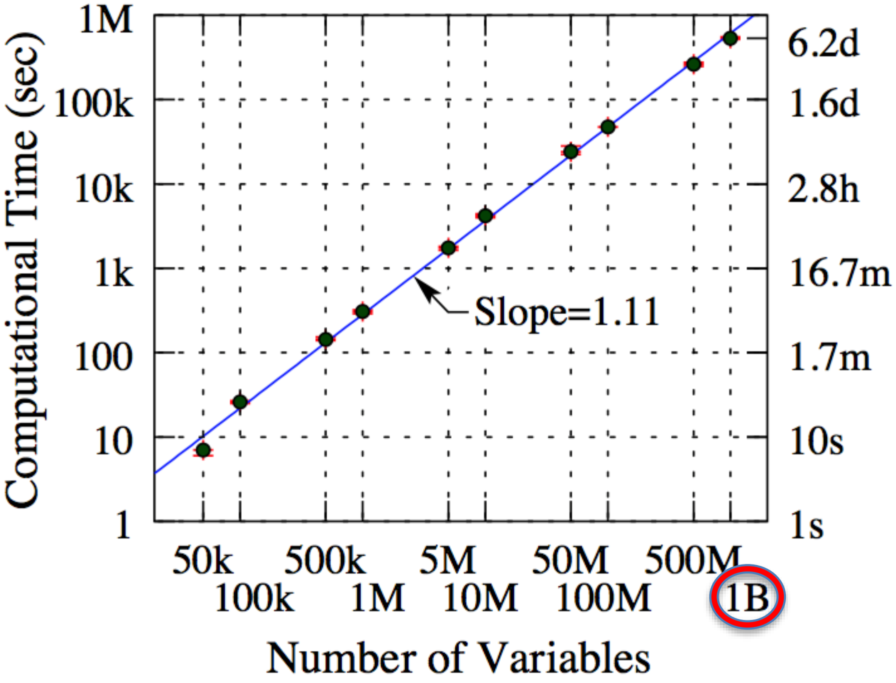
- CONs
  – Slow



- Heuristic
  - More efficient
    - Convergence
    - Speed

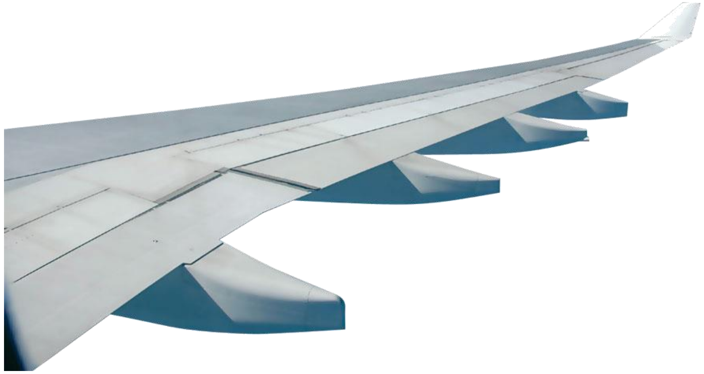
Billion Variable Problem Solved using EAs

- Multi-knapsack problem is NP-hard
- Discrete Variables
- Pop. Size: 60 for all problems
- How much is a Billion?
  - 4GBytes for a solution, 240GB RAM for a population



Deb, K., and C. Myburgh. "A population-based fast algorithm for a billion-dimensional resource allocation problem with integer variables." *European Journal of Operational Research* 261, no. 2 (2017): 460-474.

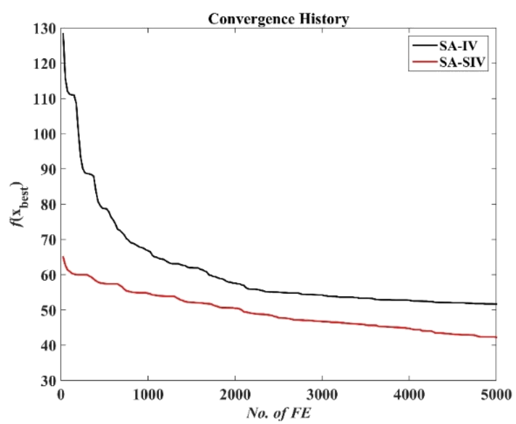
Semi-Independent Variables (SIV)

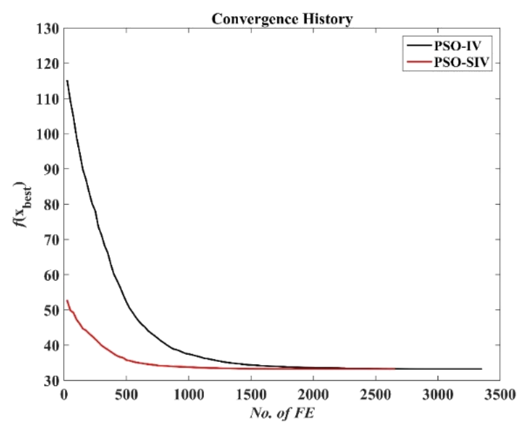


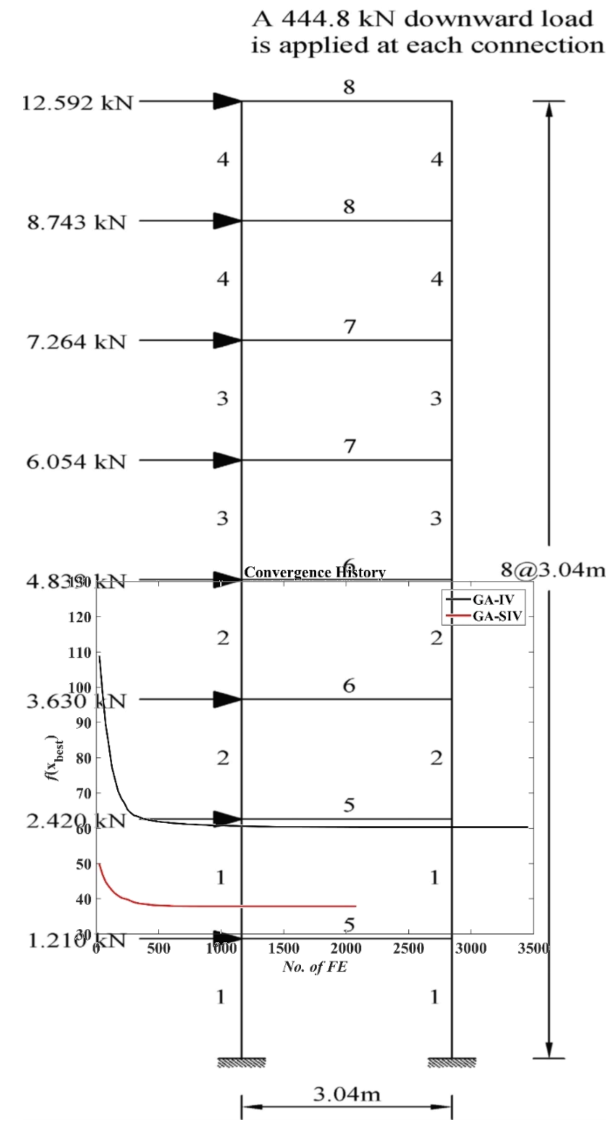
Gandomi, A.H., Deb, K., Averill, R.C., Rahnamayan, S. and Omidvar, M.N., 2018. Using Semi-independent Variables to Enhance Optimization Search. *Expert Systems with Applications*. 120, 279-297, 2019.

Frame Design

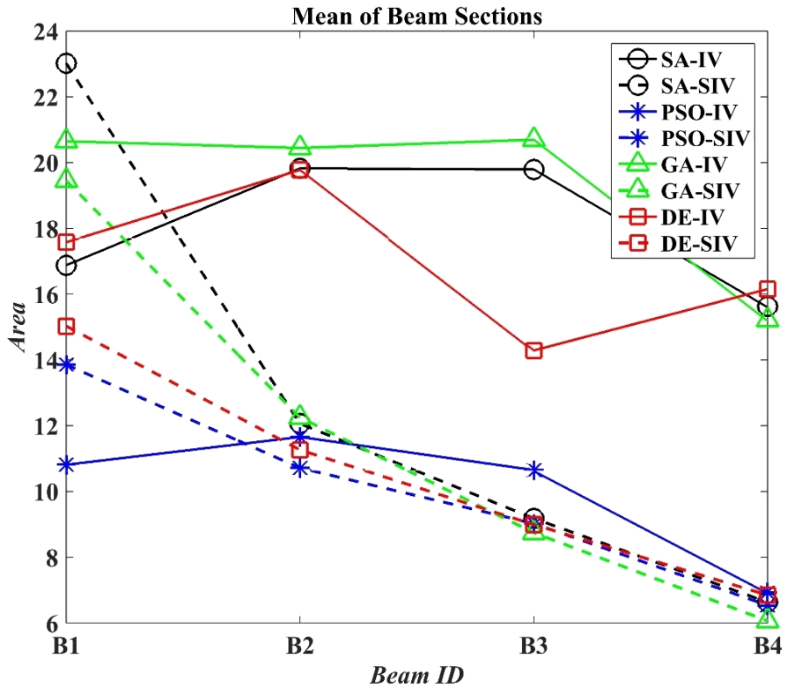
- $W = \sum_{i=1}^{NG} \rho \left(\sum_{j=1}^{NE_i} l_{i,j} \right) A_i$
- IV: $A_i \in \{W_1, \ldots, W_{267}\}$
- SIV: $A_1, A_5 \in \{W_1, \ldots, W_{267}\}$

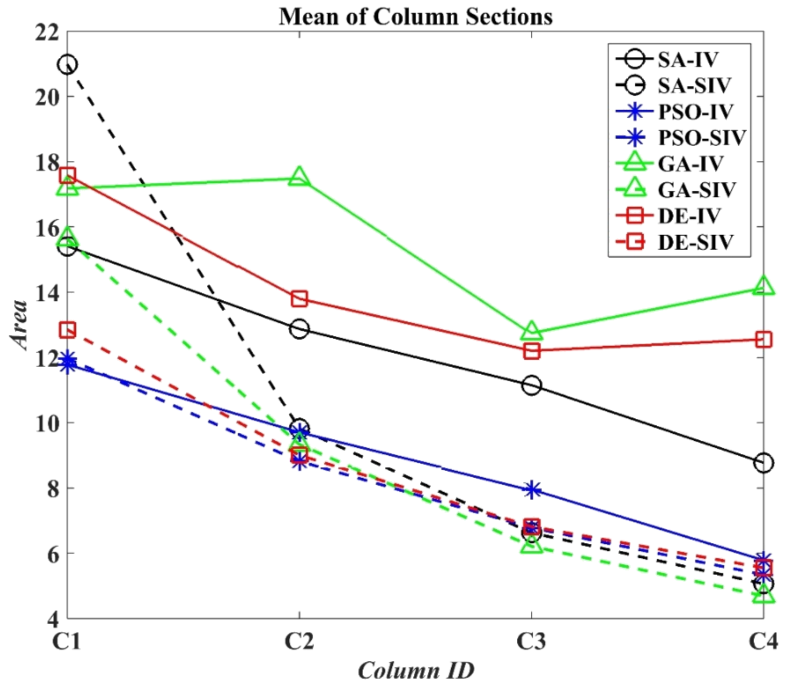






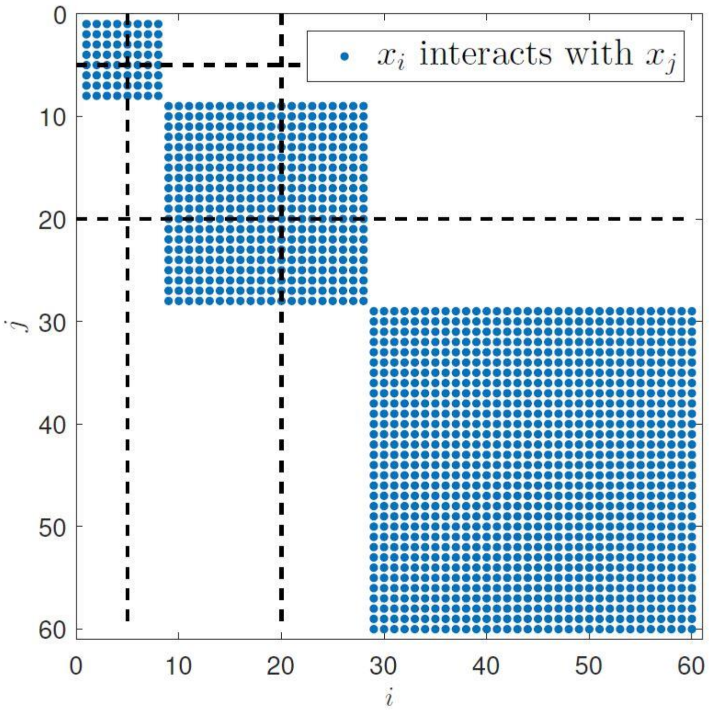
Final Solution

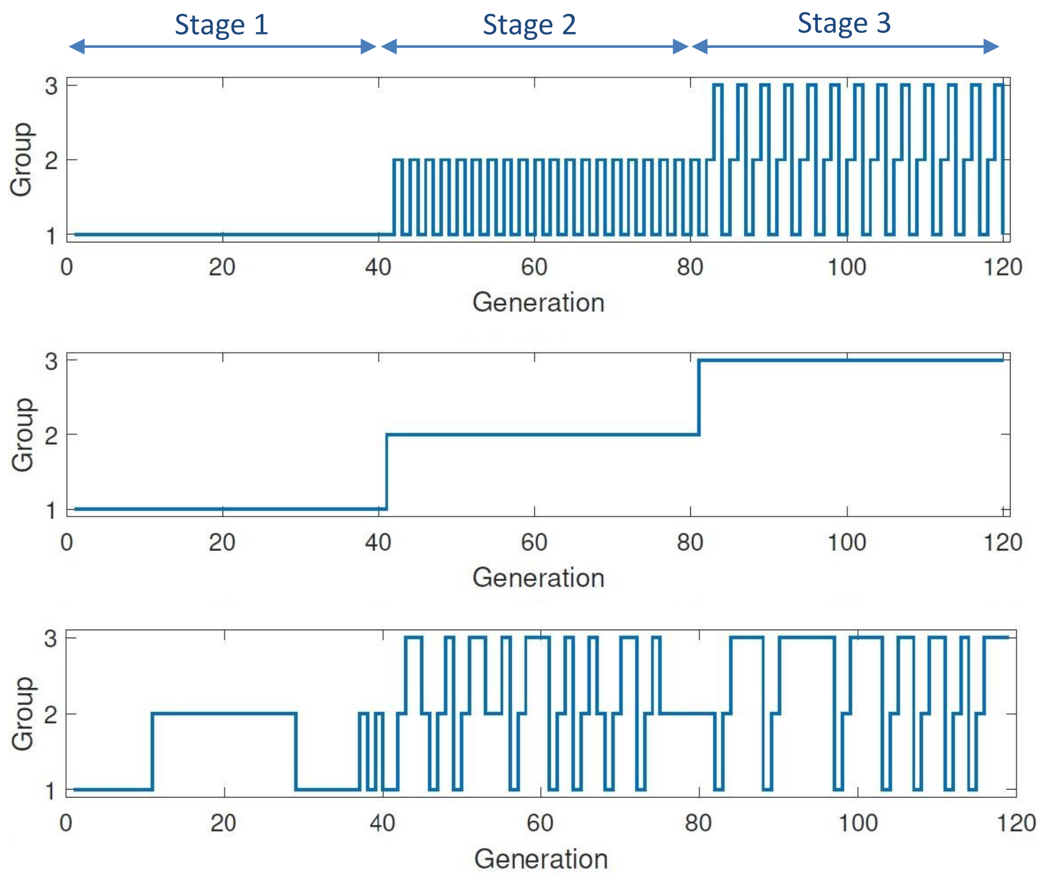




Incremental Optimization Problems -1

Increments:
from 5 → 20 → 60 variables

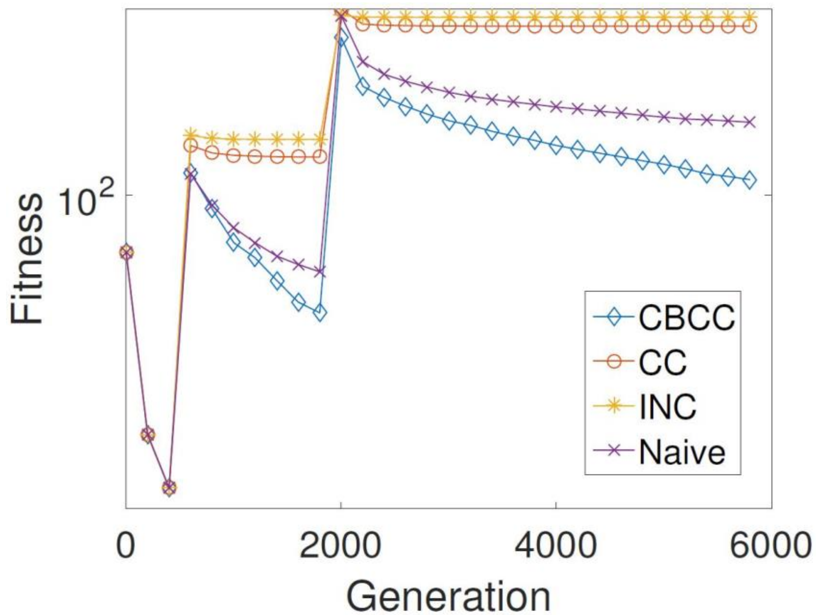


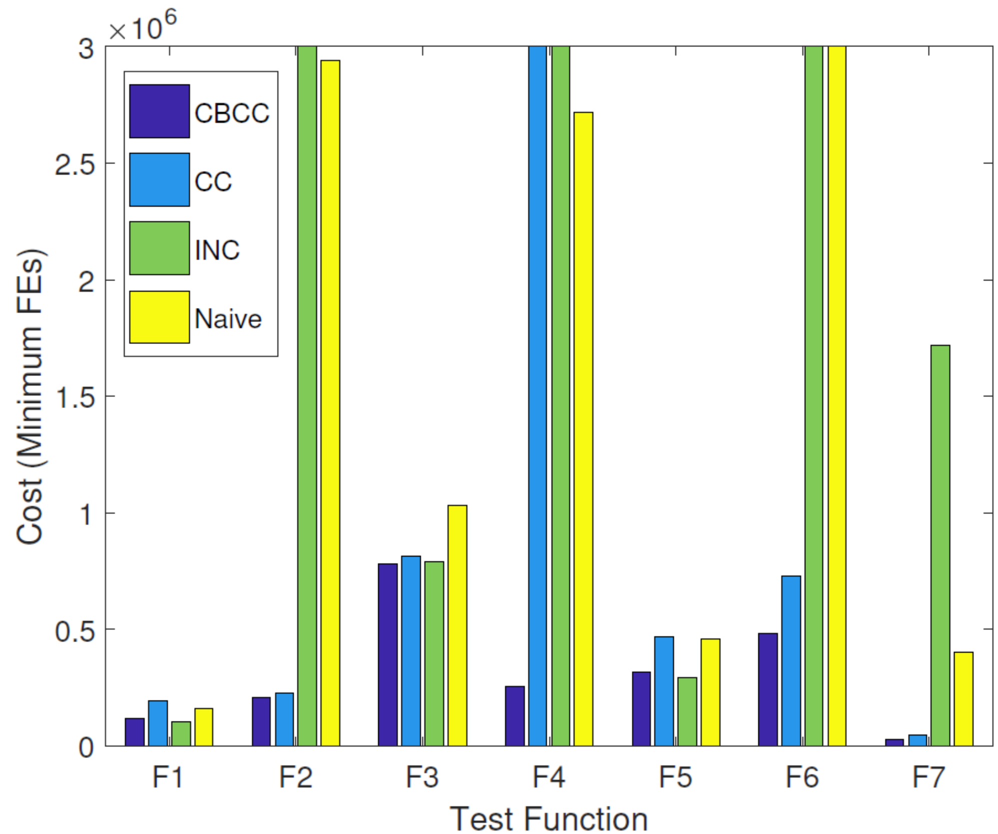


Cheng, Omidvar, Gandomi, et al. Solving Incremental Optimization Problems via Cooperative Coevolution. IEEE Transactions on Evolutionary Computation, in press. DOI: 10.1109/TEVC.2018.2883599

Incremental Optimization Problems -2

- CBCC (proposed)





Cheng, Omidvar, Gandomi, et al. Solving Incremental Optimization Problems via Cooperative Coevolution. IEEE Transactions on Evolutionary Computation, in press. DOI: 10.1109/TEVC.2018.2883599

Conclusion

- Although nature-inspired algorithms are simple, they are useful for solving complex Eng. Problems.
- Each evolutionary algorithm has its own advantages, For example:
  - FA: multi-modal optimization problems
  - ISA: discrete optimization problems
  - P3: large-scale discrete optimization problem
- Best algorithm(s) should be found for each Problem
- Heuristics can be used within nature-inspired algorithms
- Customization can enhance the optimization process

STEVENS
INSTITUTE *of* TECHNOLOGY
THE INNOVATION UNIVERSITY®

stevens.edu

Thank You

STEVENS INSTITUTE of TECHNOLOGY

School of Business

stevens.edu

Amir H Gandomi; PhD

Assistant Professor of Analytics & Information Systems

a.h.gandomi@stevens.edu