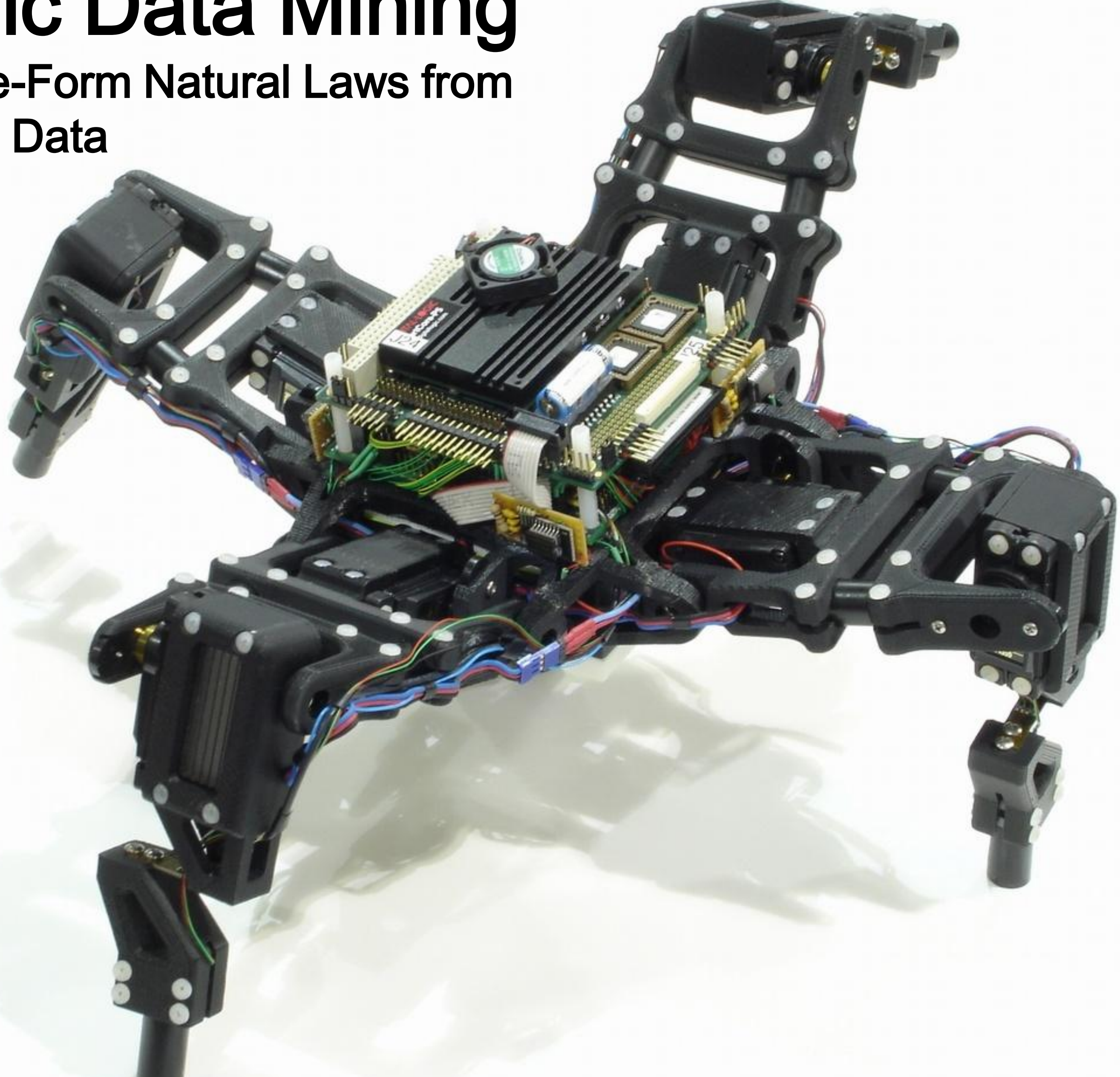
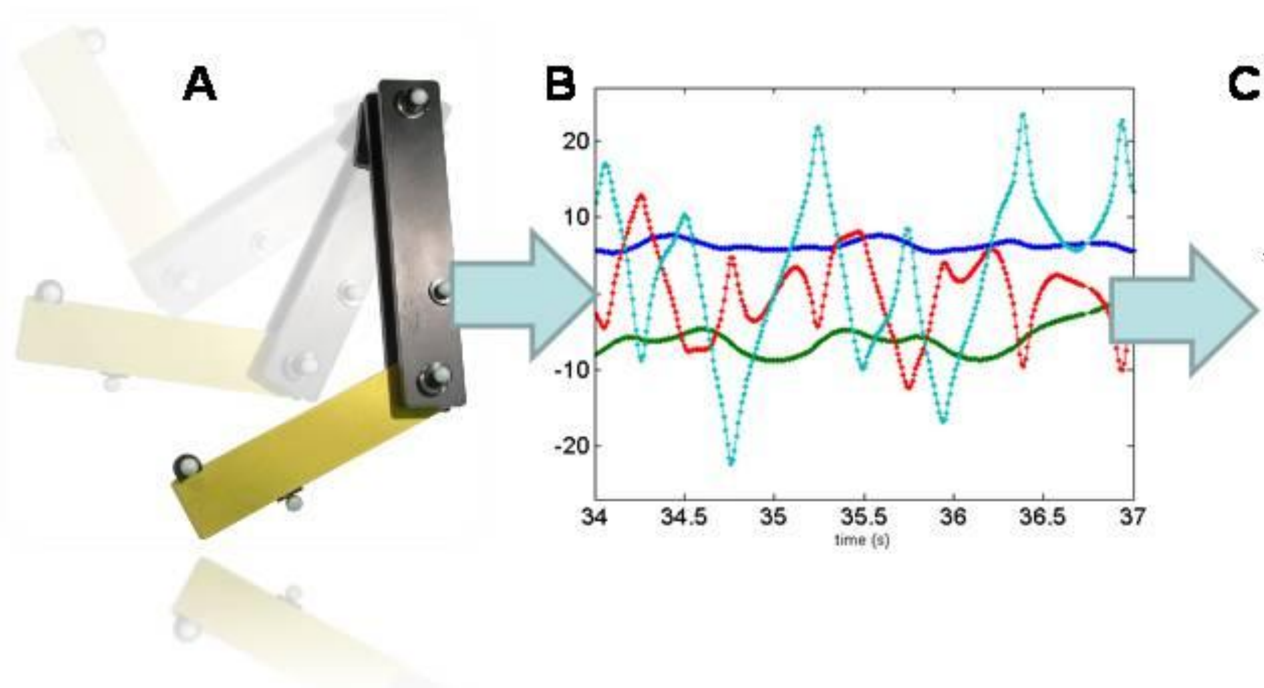
Scientific Data Mining

Distilling Free-Form Natural Laws from Experimental Data



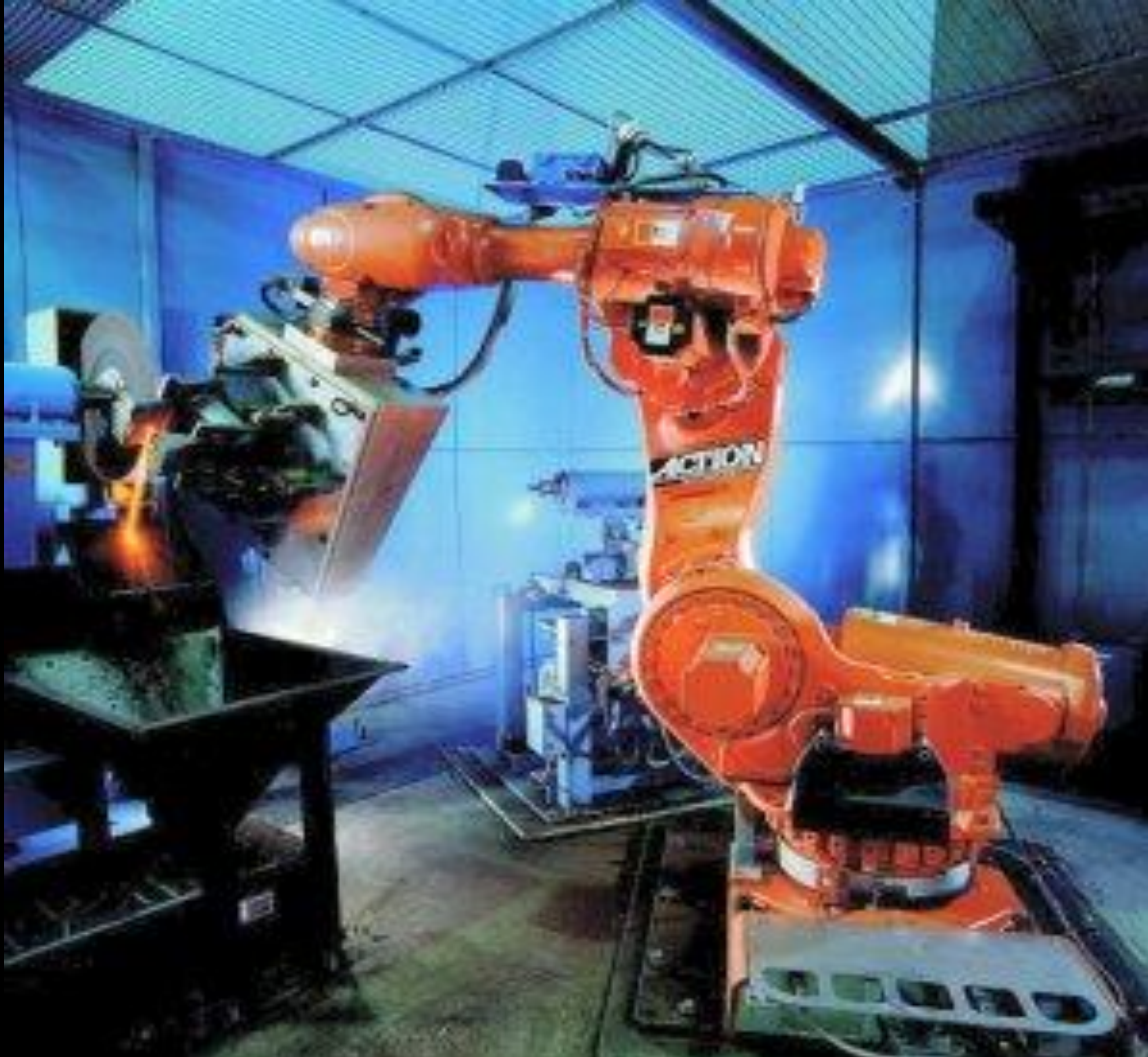
Hod Lipson, Cornell University

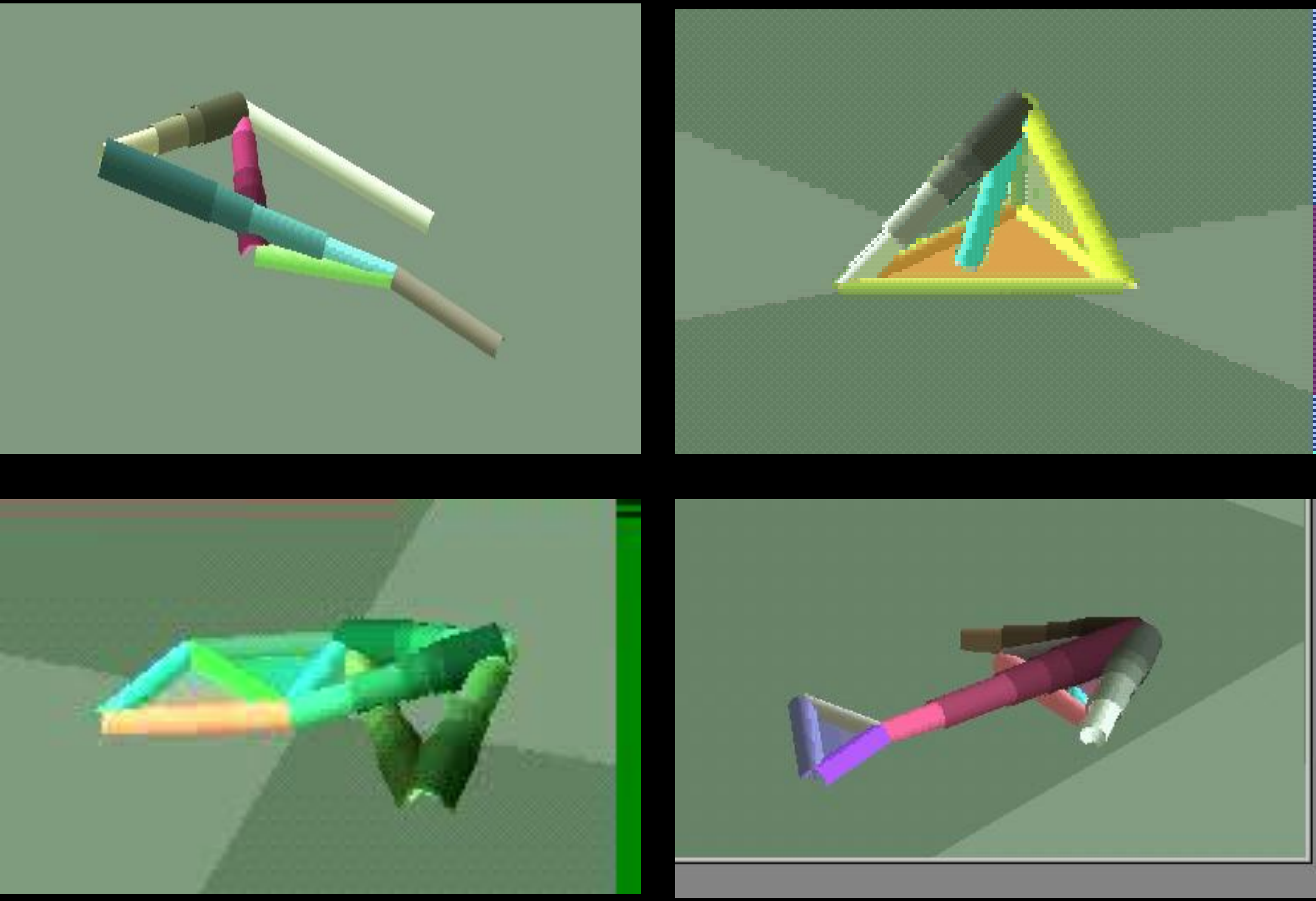


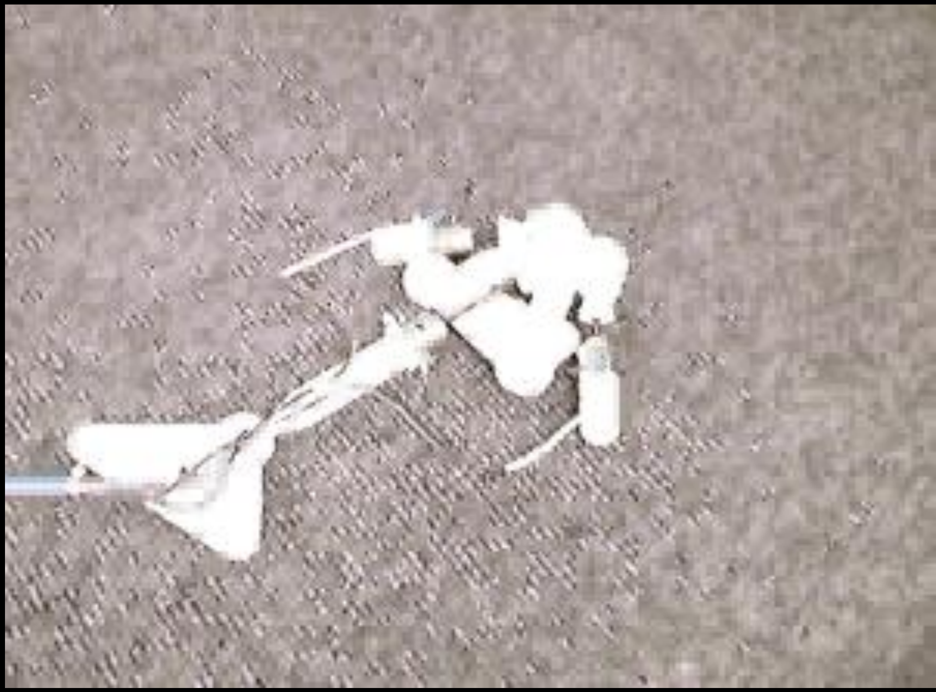
C

Detected Invariance:

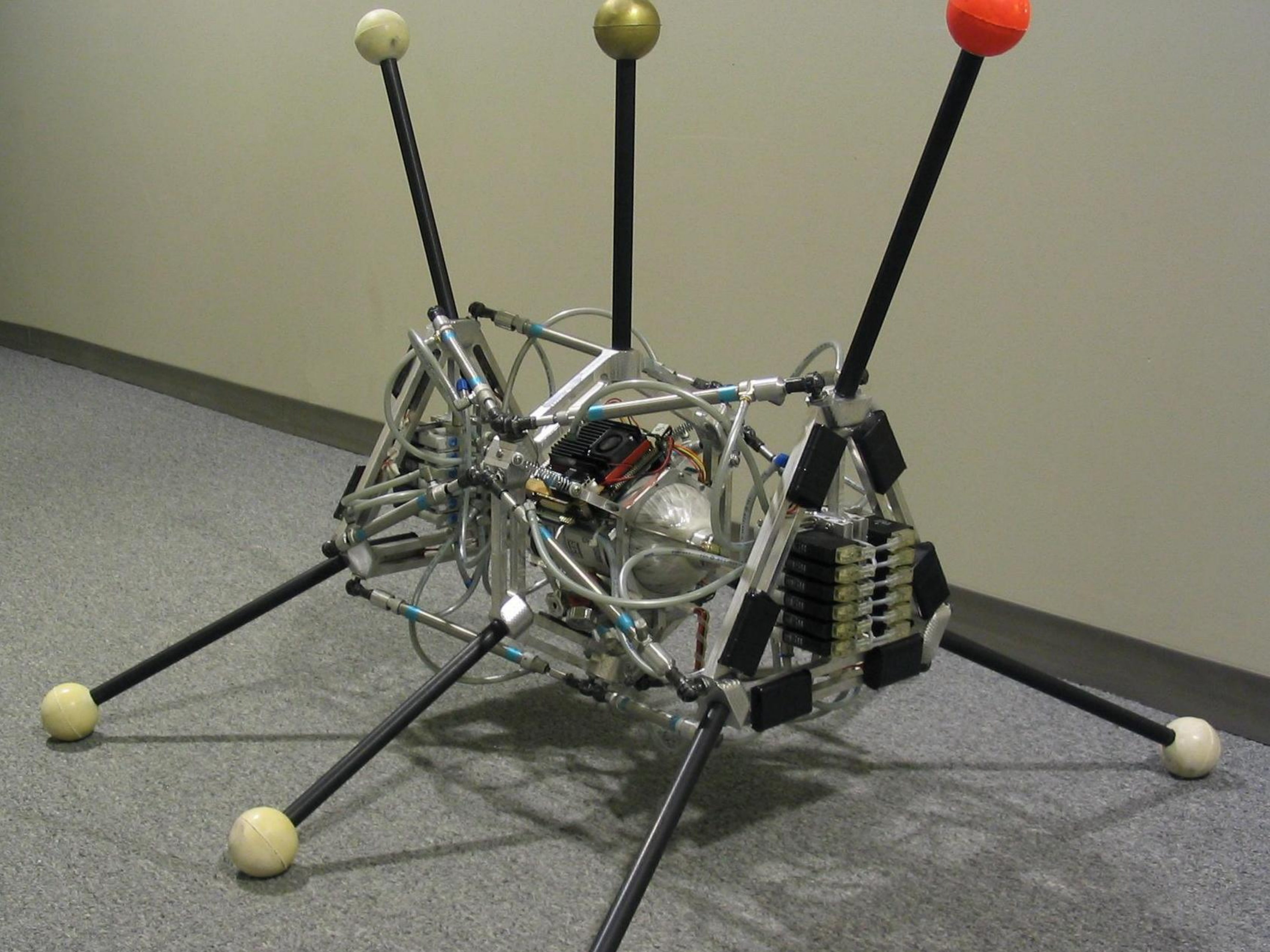
$$L_1^{\,2}(m_1+m_2)\boldsymbol{\omega_1}^{\mathbf{2}} + m_2L_2^{\,2}\boldsymbol{\omega_2}^{\mathbf{2}} + m_2L_1L_2\boldsymbol{\omega_1\omega_2}\cos(\boldsymbol{\theta_1}-\boldsymbol{\theta_2}) - 19.6L_1(m_1+m_2)\cos\boldsymbol{\theta_1} - 19.6m_2L_2\cos\boldsymbol{\theta_2}$$

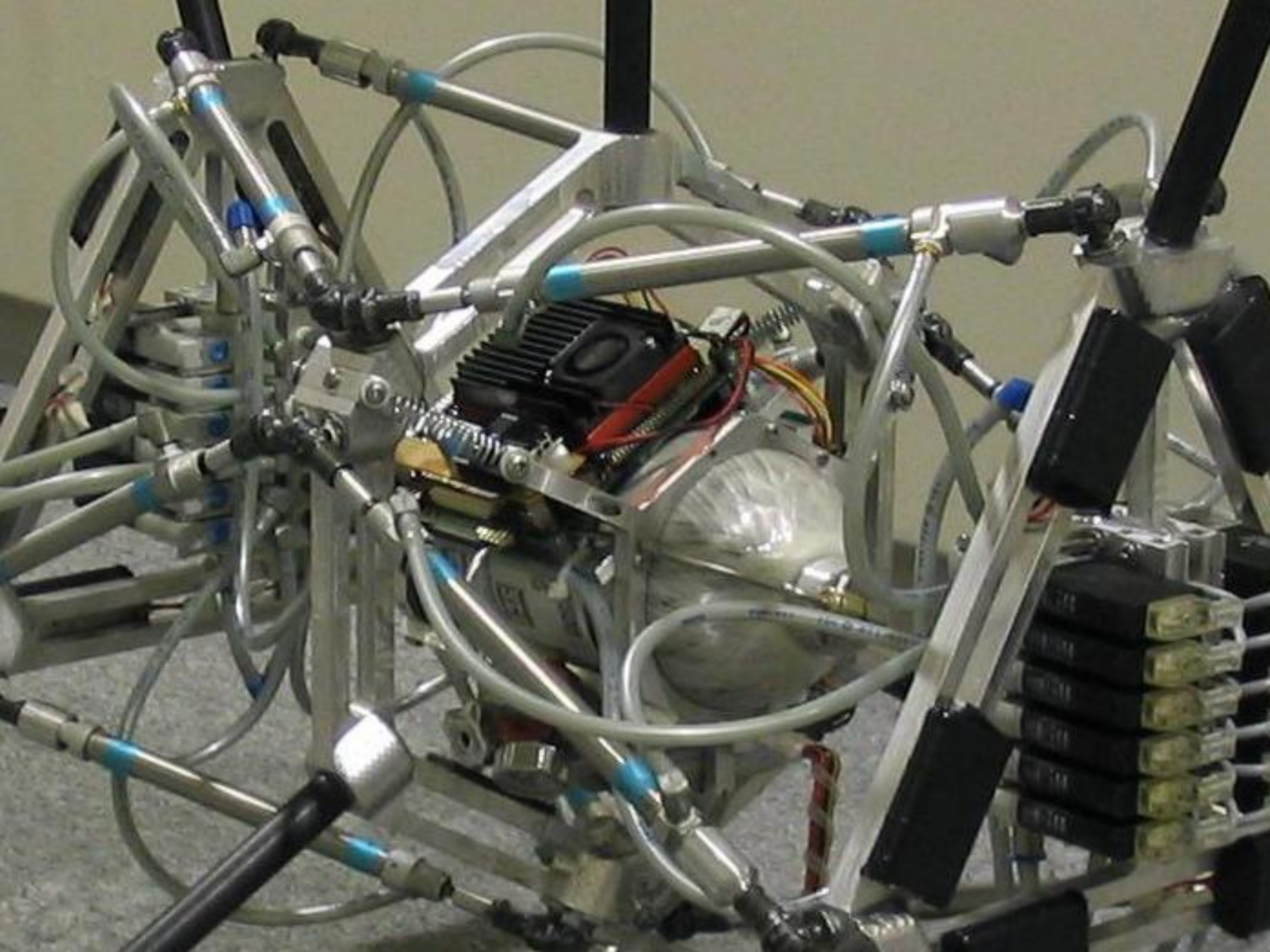


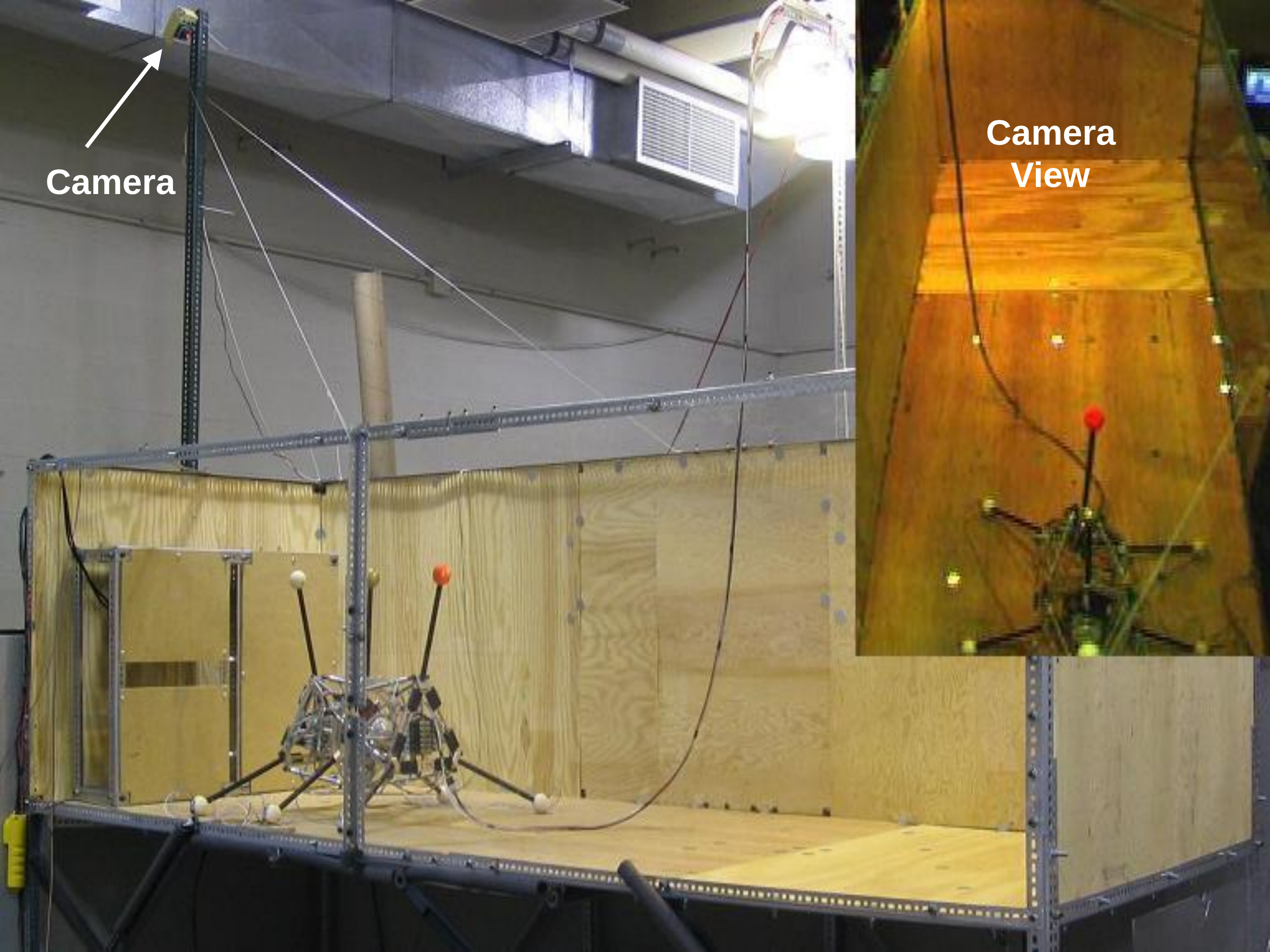




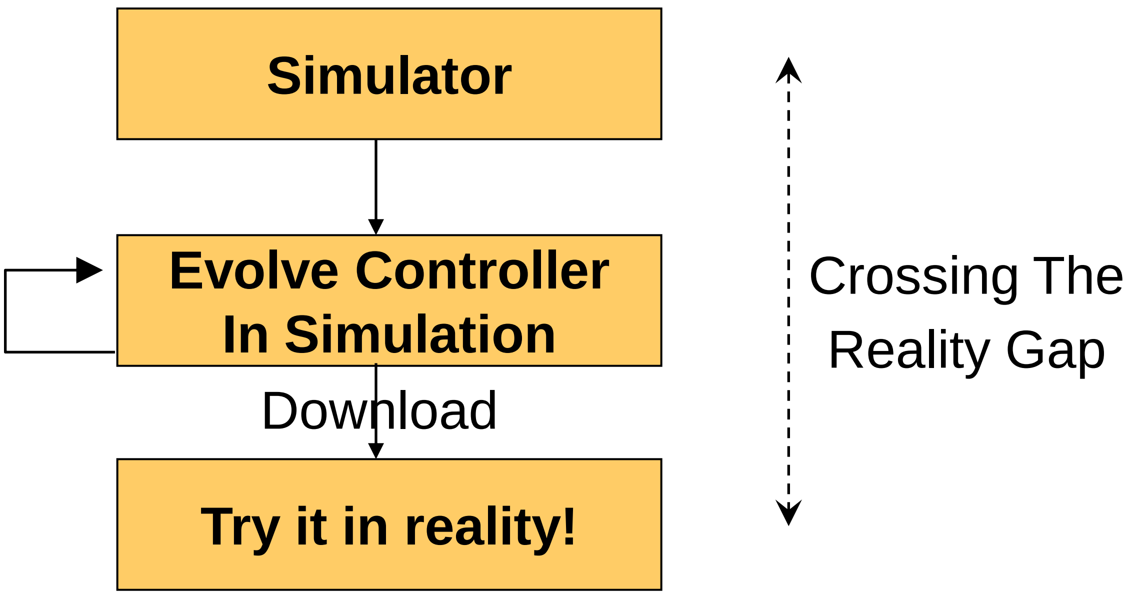
Lipson & Pollack, Nature 406, 2000



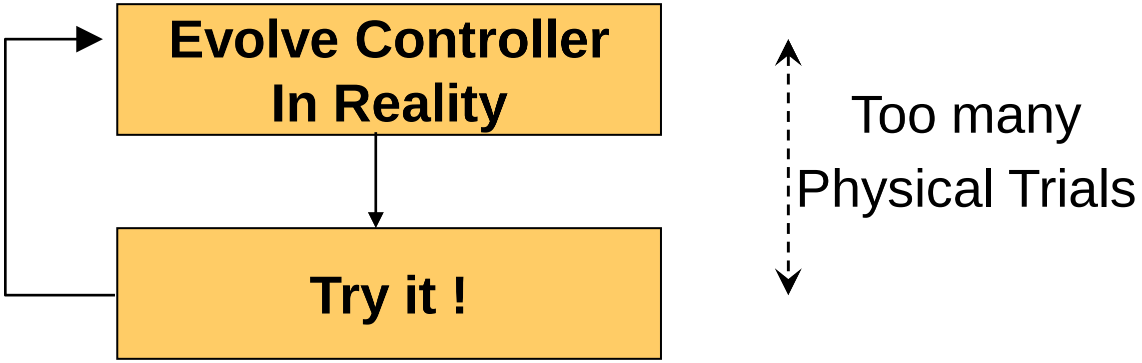



Camera
Camera
View

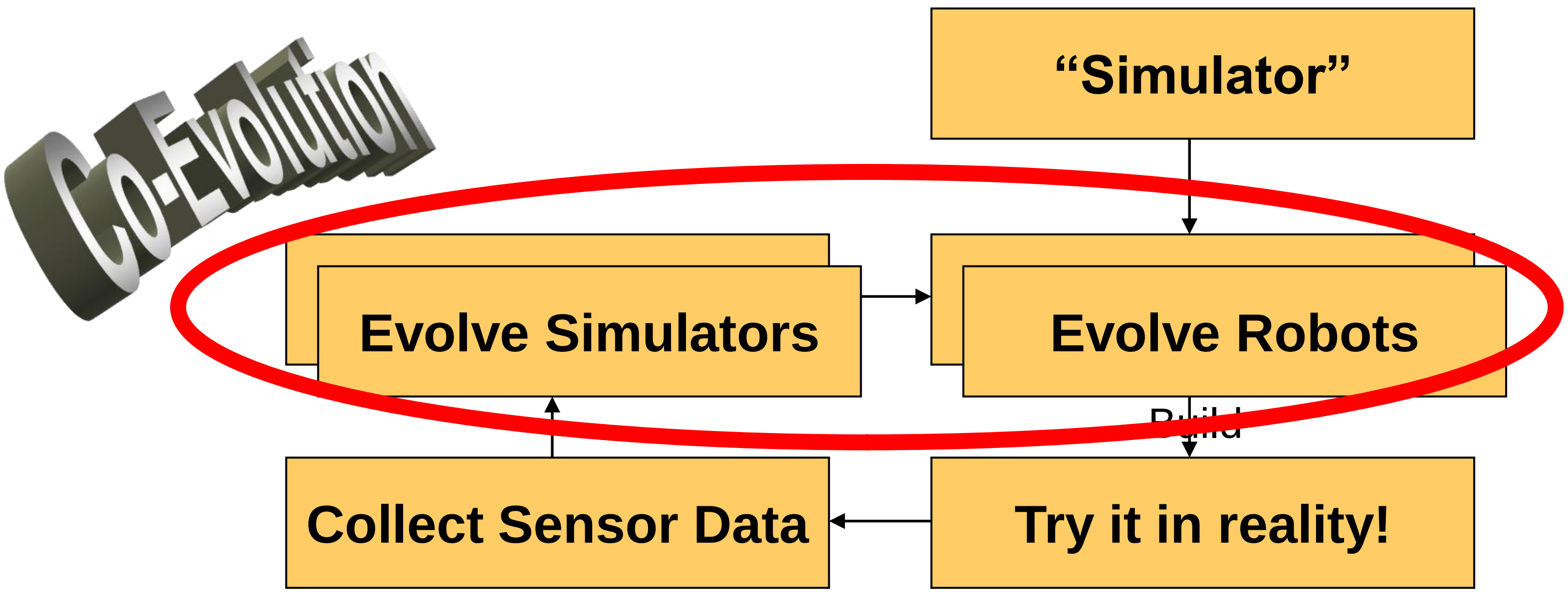
Adapting in simulation

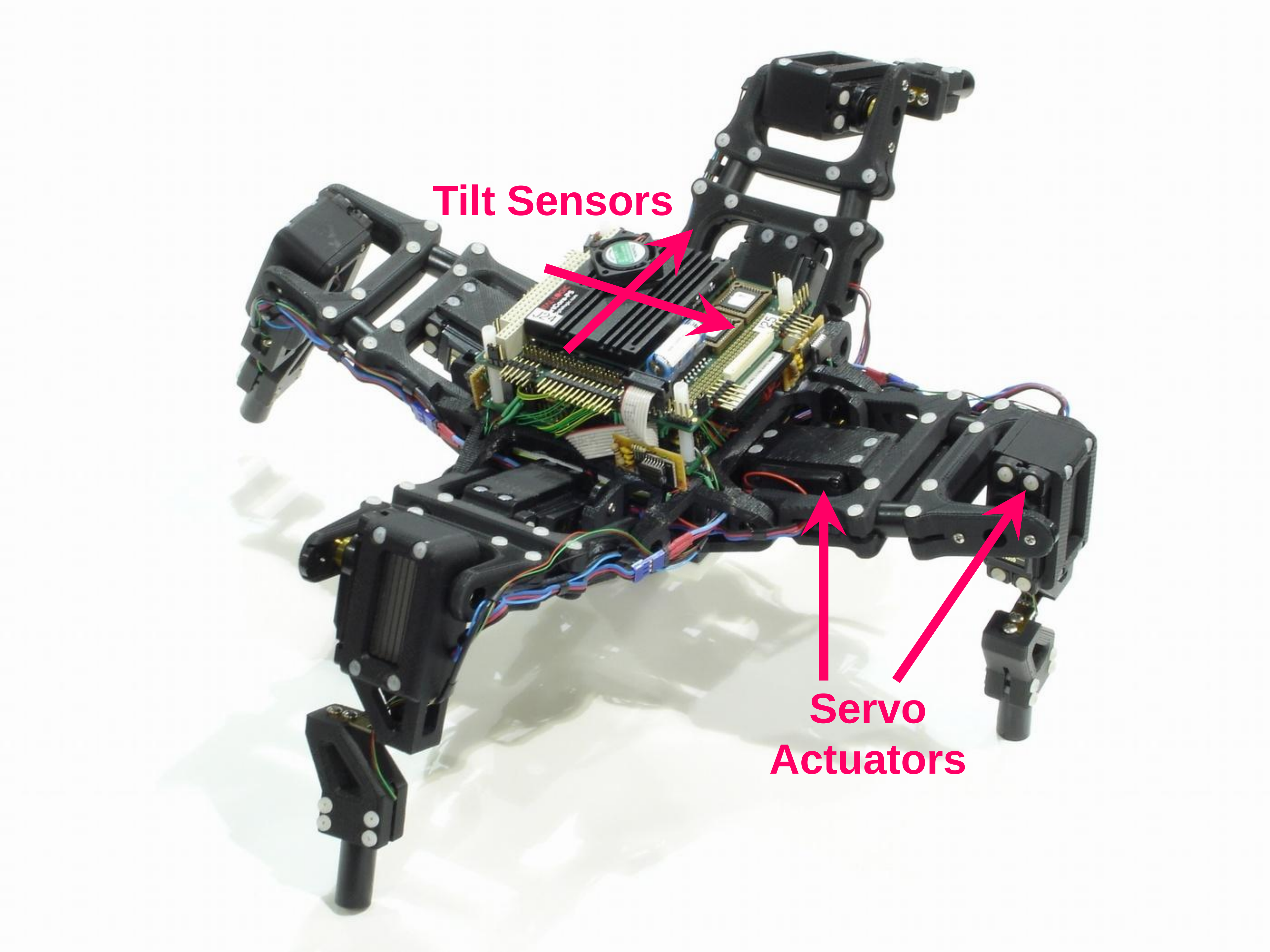


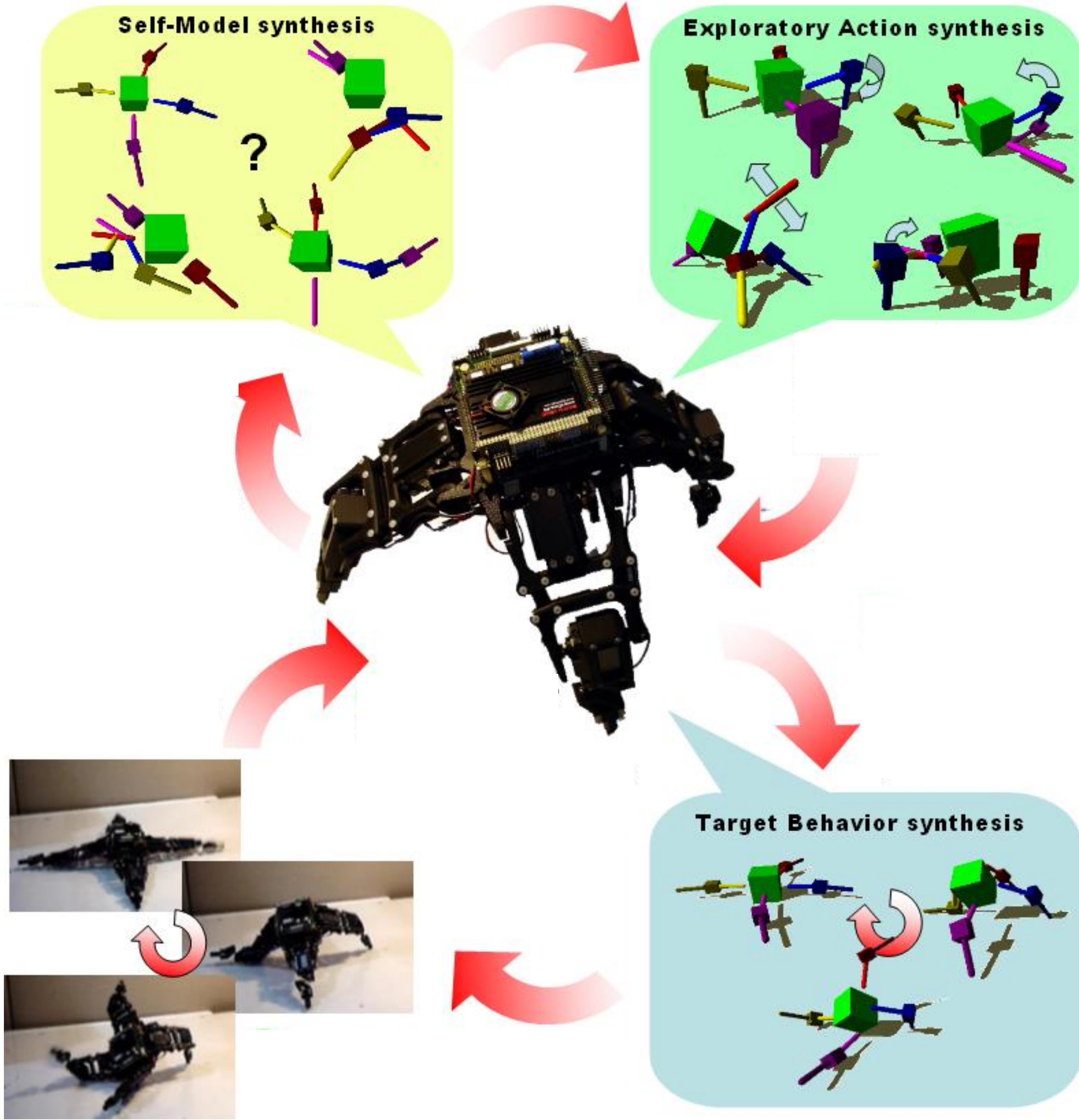
Adapting in reality



Simulation & Reality

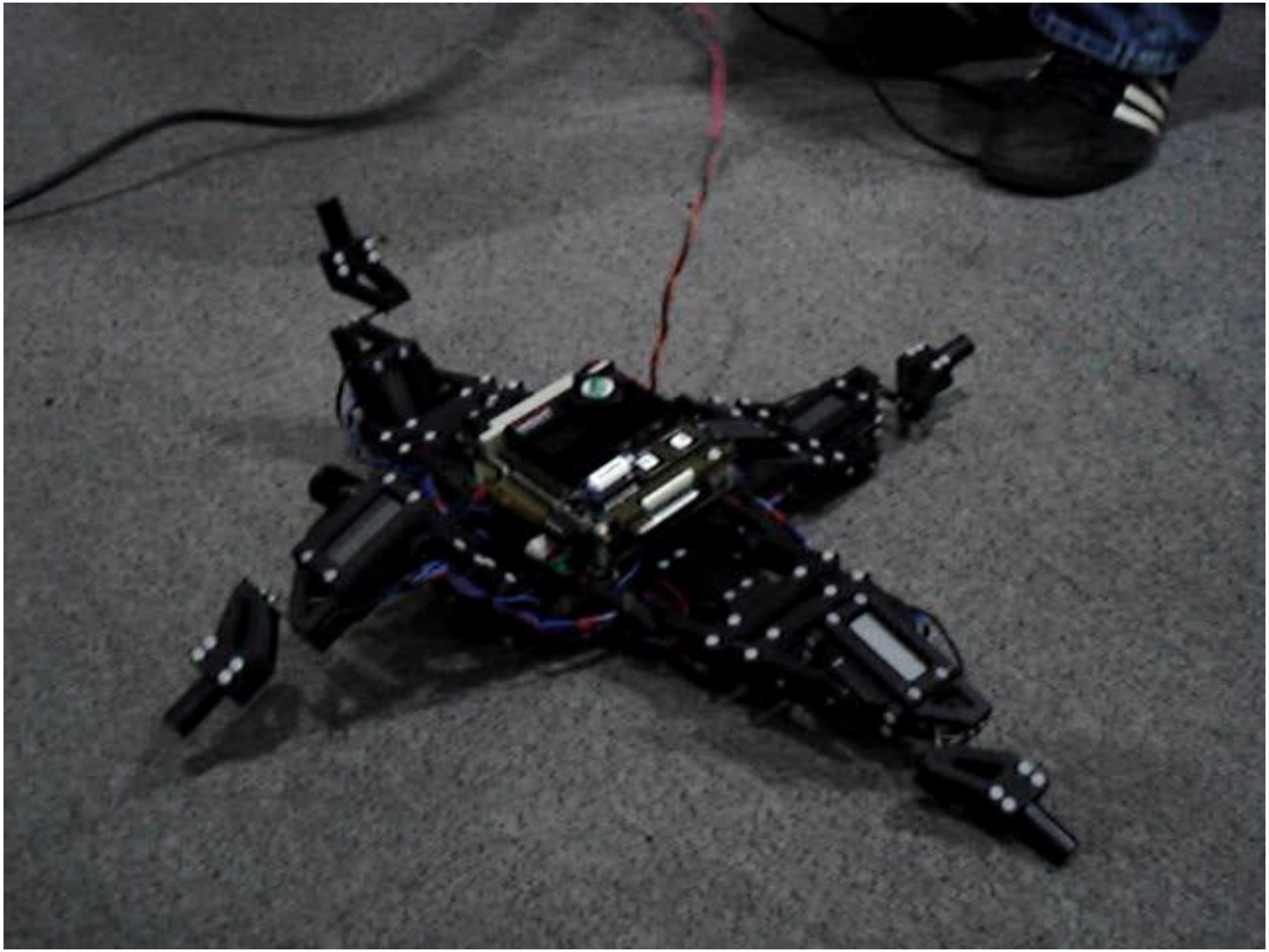



Tilt Sensors
Servo
Actuators

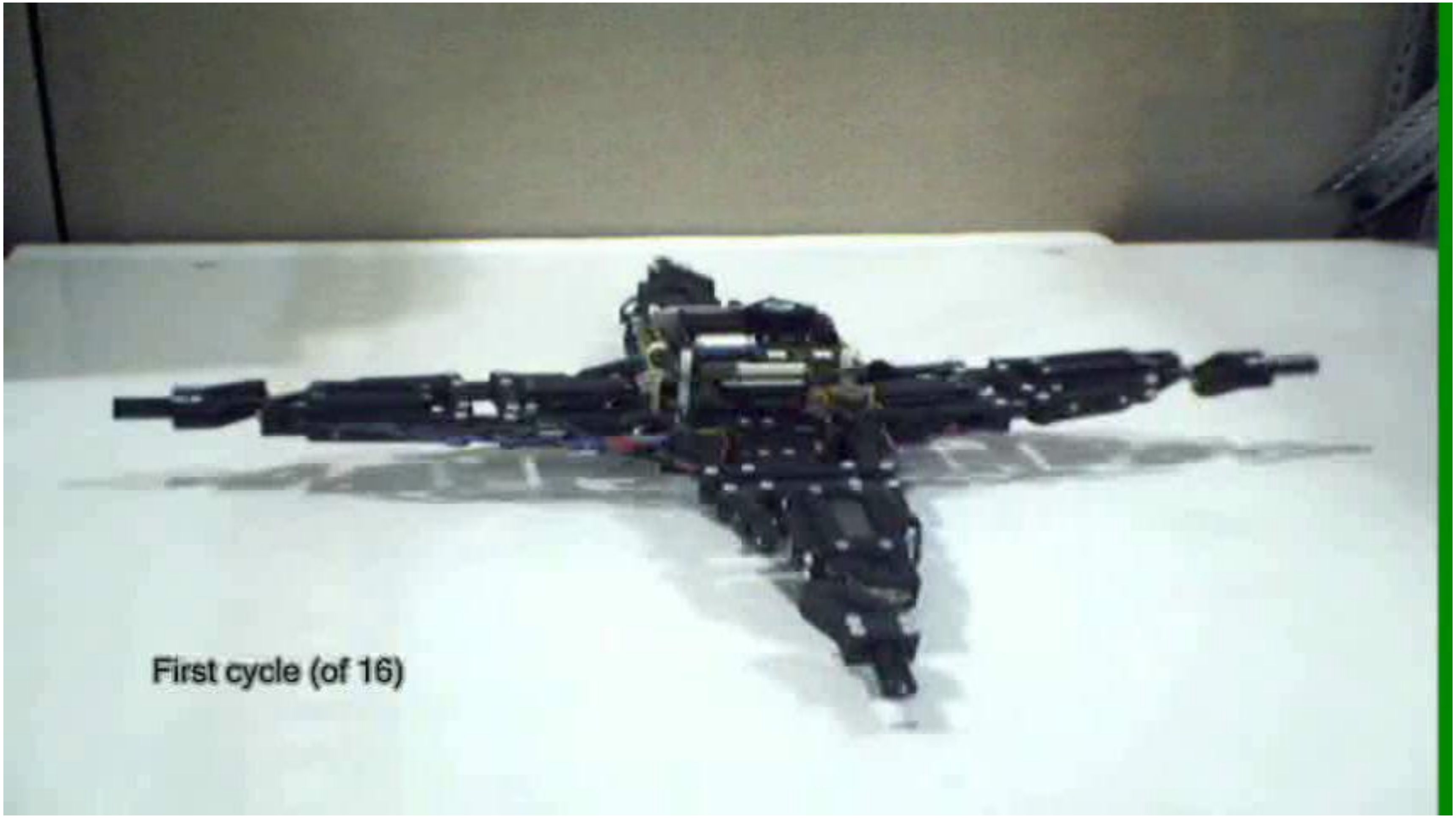


Self-Model synthesis
?
Exploratory Action synthesis
Target Behavior synthesis

Morphological Estimation



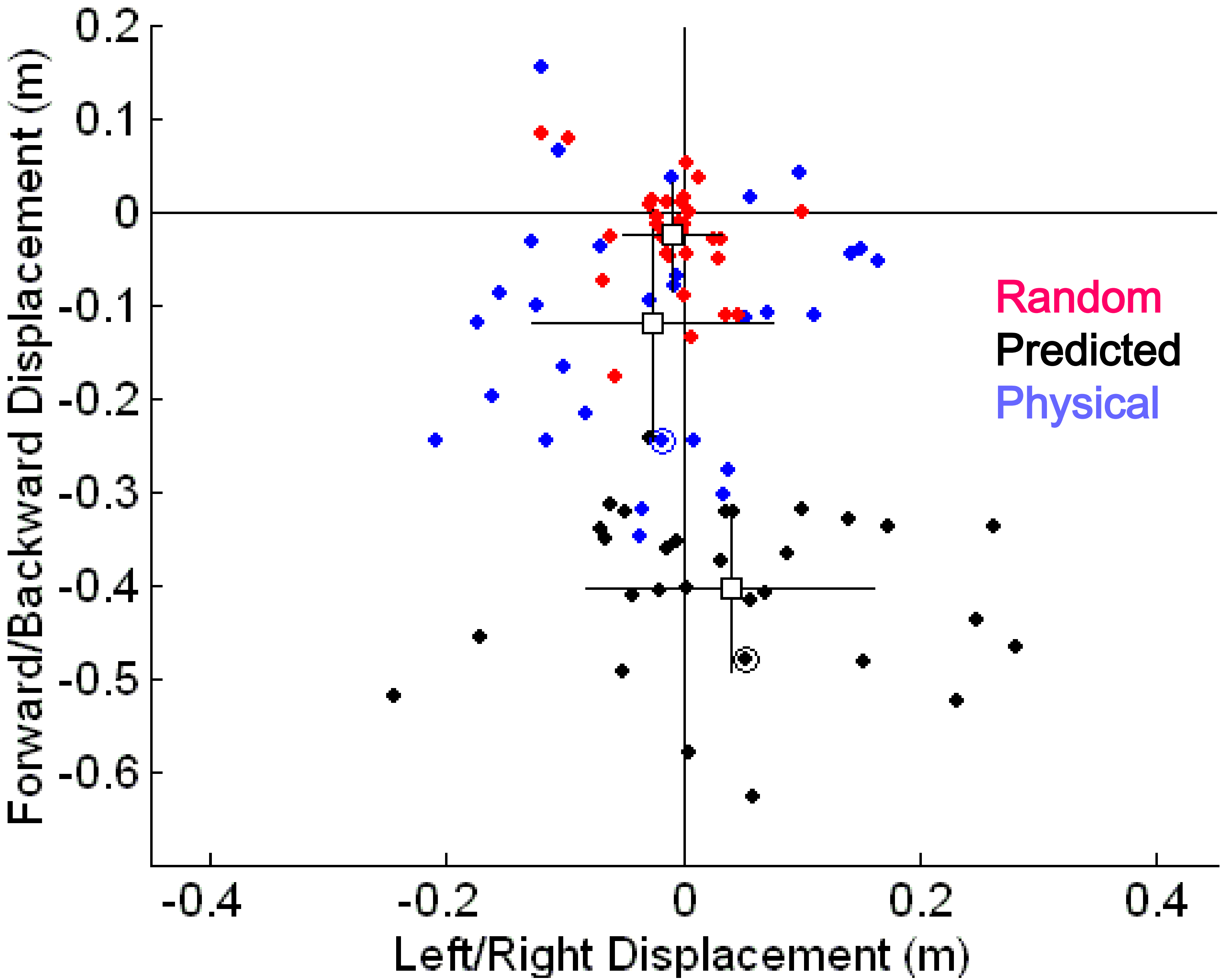
Emergent Self-Model

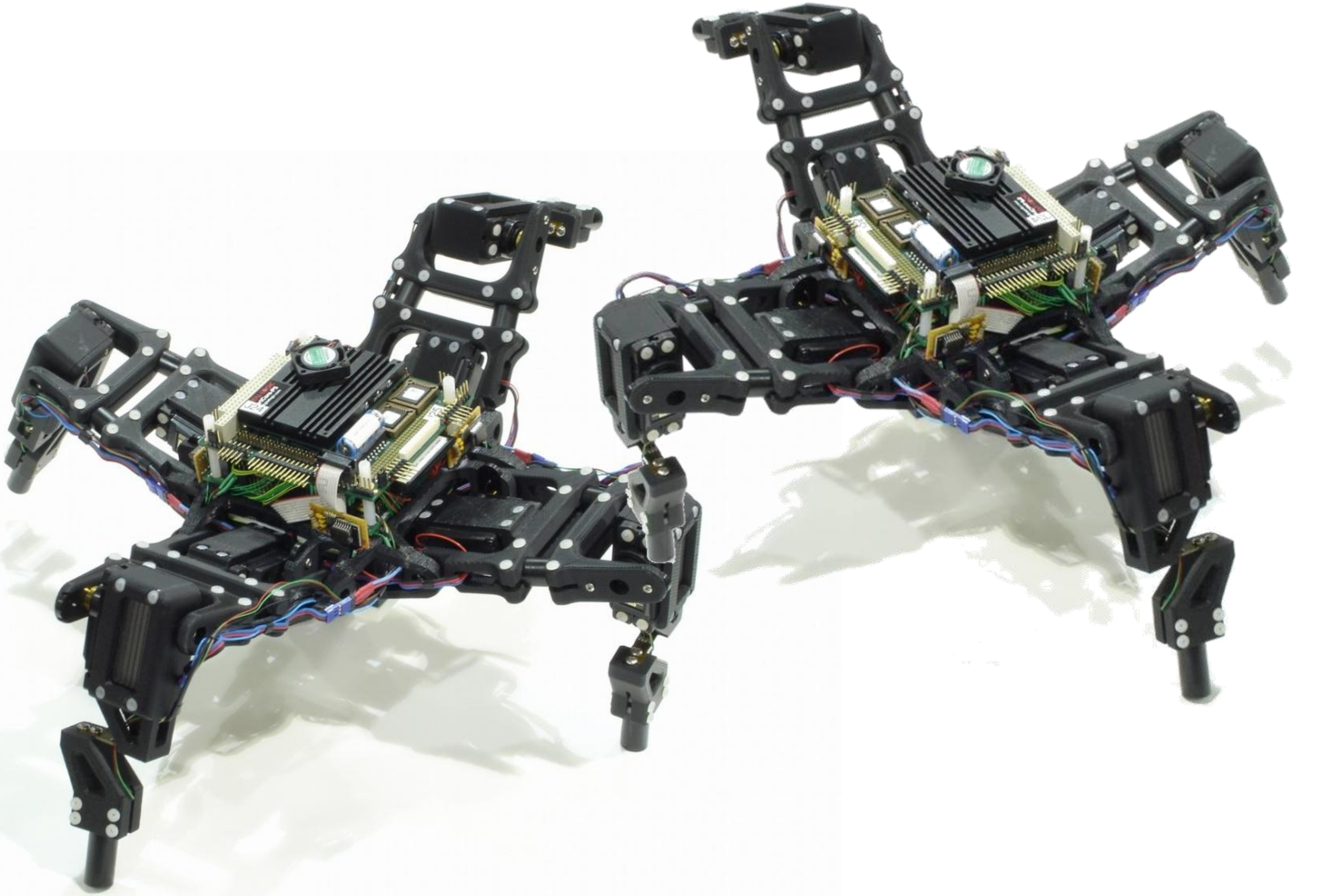


With Josh Bongard and Victor Zykov, Science 2006

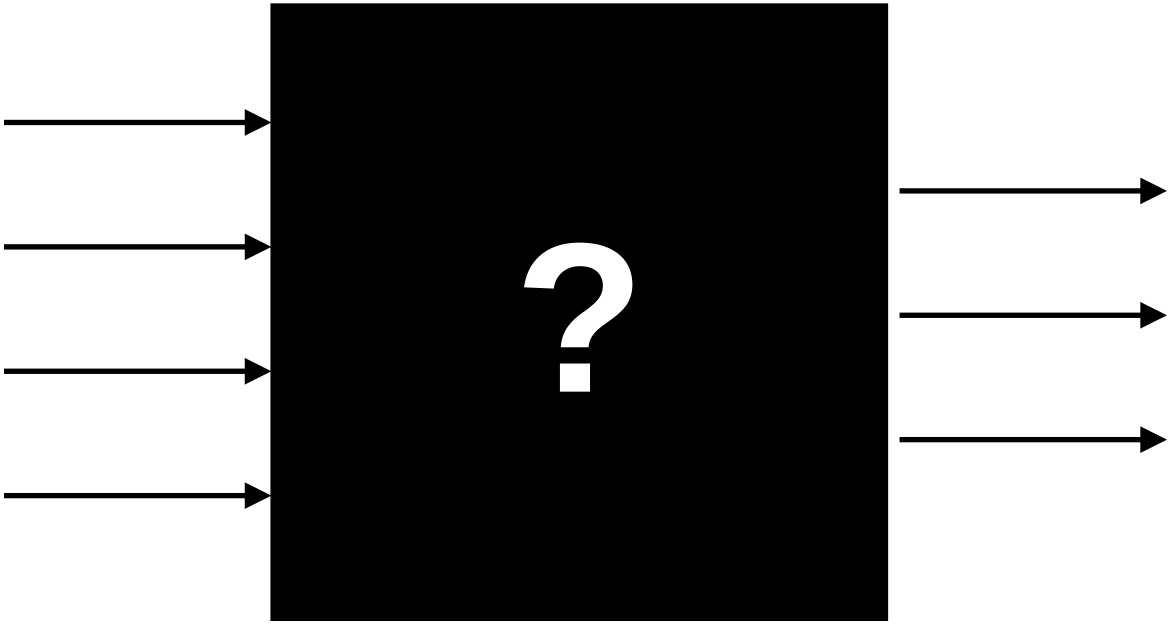
Damage Recovery

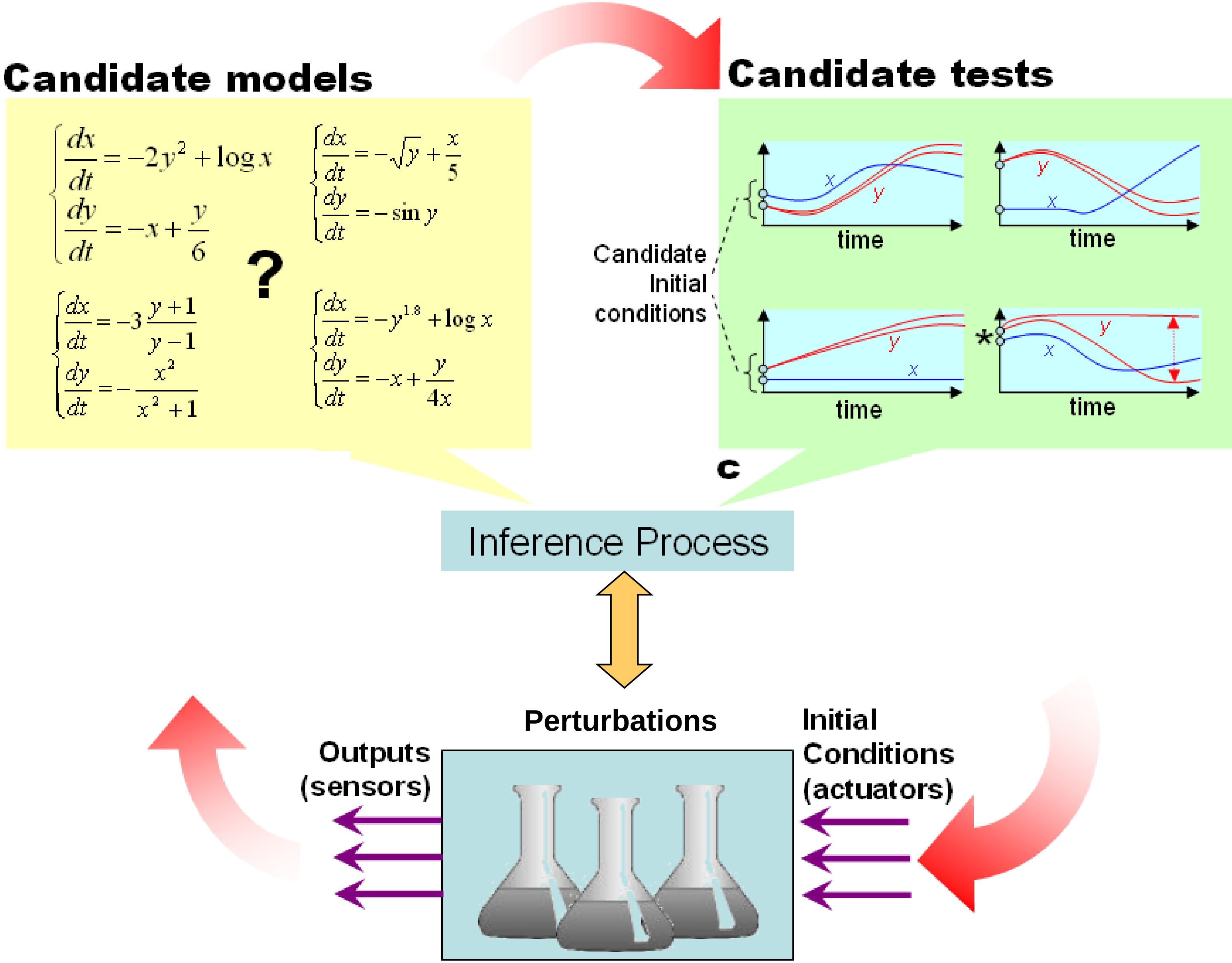
With Josh Bongard and Victor Zykov, Science 2006


0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
Forward/Backward Displacement (m)
-0.4
-0.2
0
0.2
0.4
Left/Right Displacement (m)
Random
Predicted
Physical



System Identification



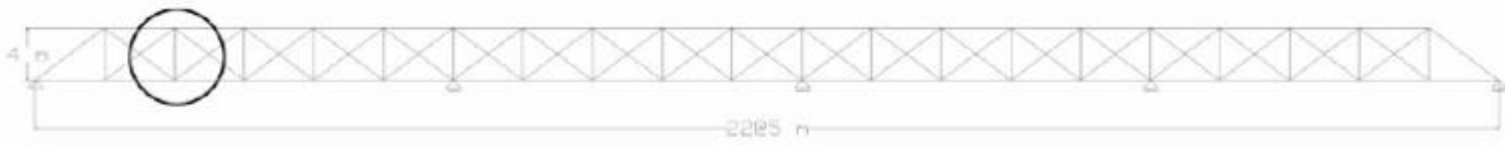


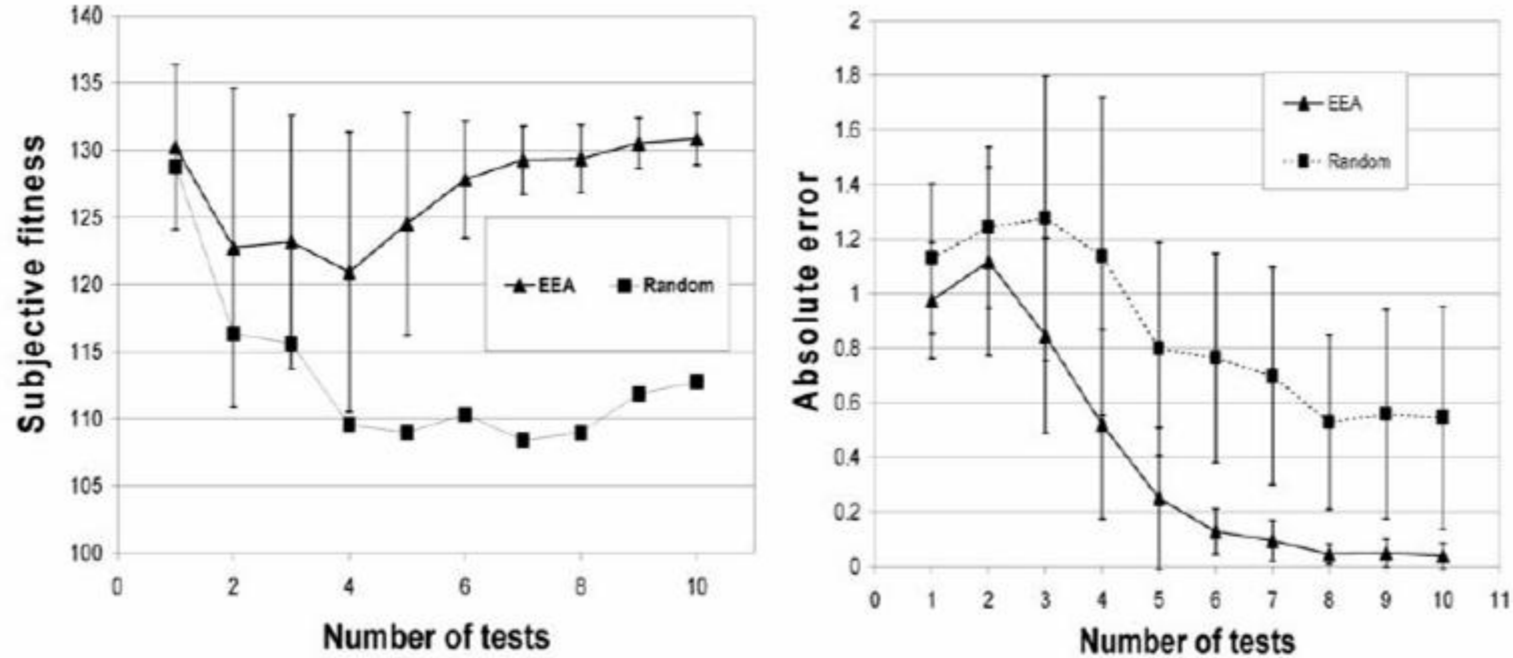
Candidate models
$$\begin{cases}\frac{dx}{dt} = -2y^2 + \log x \\ \frac{dy}{dt} = -x + \frac{y}{6}\end{cases}$$
$$\begin{cases}\frac{dx}{dt} = -\sqrt{y} + \frac{x}{5} \\ \frac{dy}{dt} = -\sin y\end{cases}$$
?
$$\begin{cases}\frac{dx}{dt} = -3\frac{y+1}{y-1} \\ \frac{dy}{dt} = -\frac{x^2}{x^2+1}\end{cases}$$
$$\begin{cases}\frac{dx}{dt} = -y^{1.8} + \log x \\ \frac{dy}{dt} = -x + \frac{y}{4x}\end{cases}$$
Candidate tests
Candidate Initial conditions
x
y
time
time
time
time
*
c
Inference Process
Perturbations
Outputs (sensors)
Initial Conditions (actuators)

Photo: Floris van Breugel

Photo: Floris van Breugel

Static ID: Damage Diagnosis





(a) (b)

With Wilkins Aquino

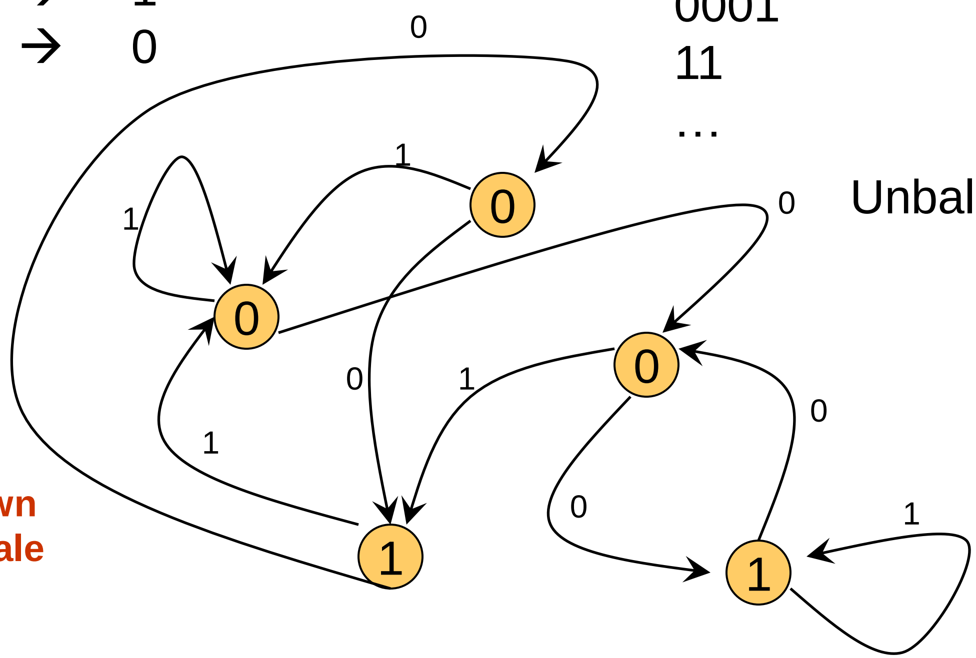
Discrete Dynamics Inference

100010101101	→	1
0101010	→	0
00111010	→	1
101010	→	0
0011101	→	1
0001	→	1
11	→	0
…		

Balanced DFA

100010101101	→	0
0101010	→	0
00111010	→	0
101010	→	0
0011101	→	1
0001	→	0
11	→	0
…		

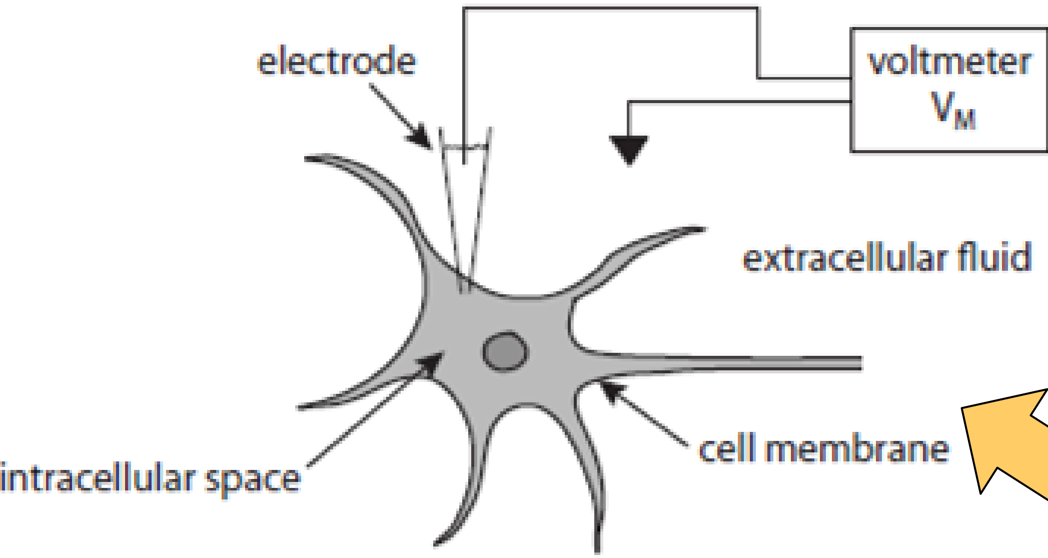
Unbalanced DFA

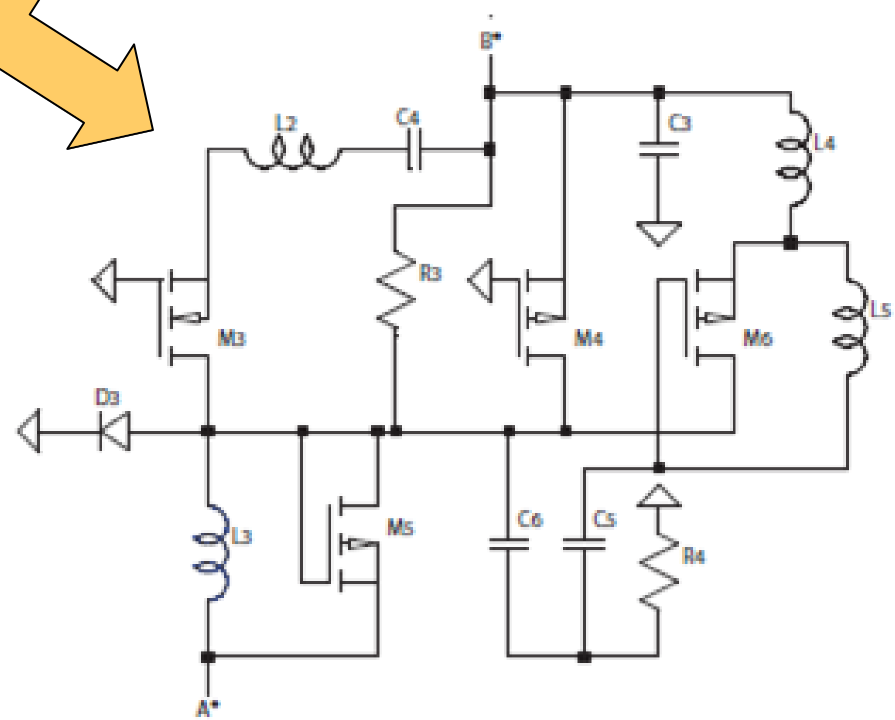


Outperforms best known algorithms on large scale unbalanced DFAs

Bongard J. C, Lipson H., (2005) “"Active Coevolutionary Learning of Deterministic Finite Automata”, Journal of Machine Learning research (JMLR), Vol. 6 No. 10, pp. 1651-1678

Circuit Building Blocks

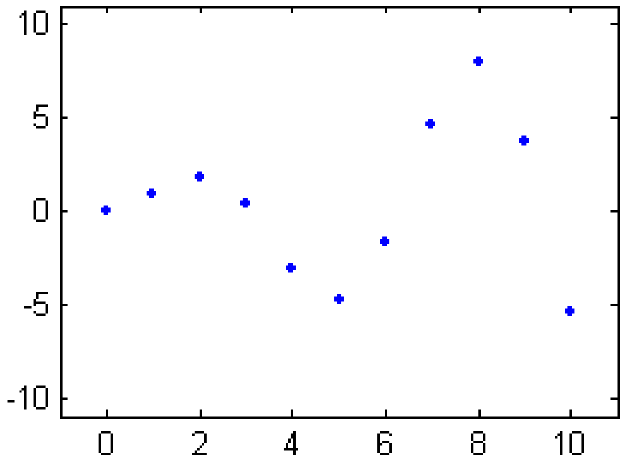


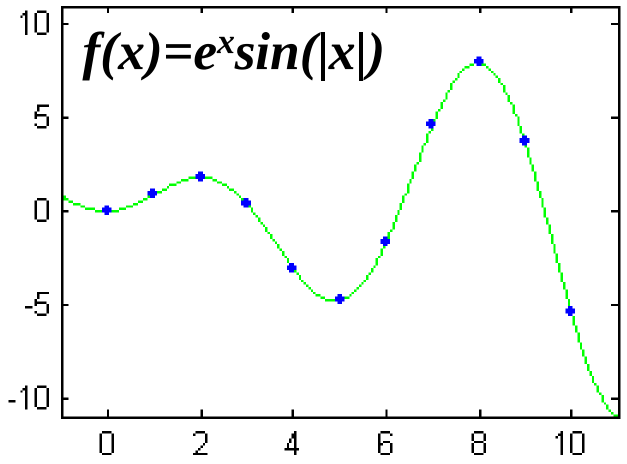


Ted Cornforth

Symbolic Regression

What function describes this data?

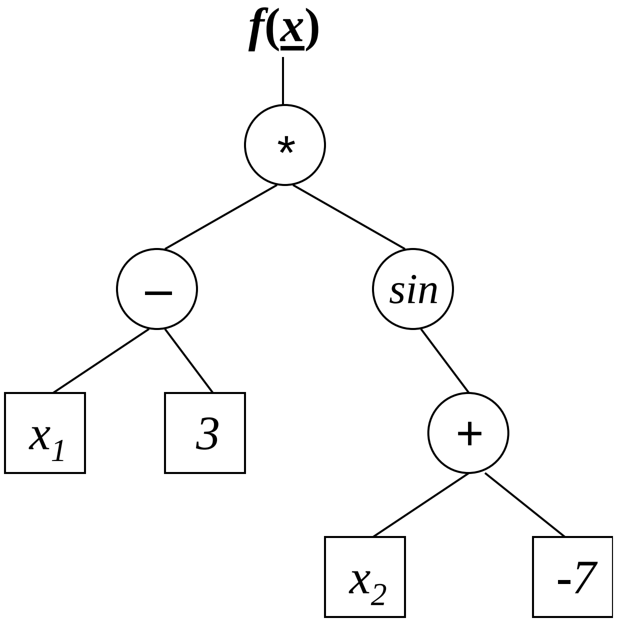


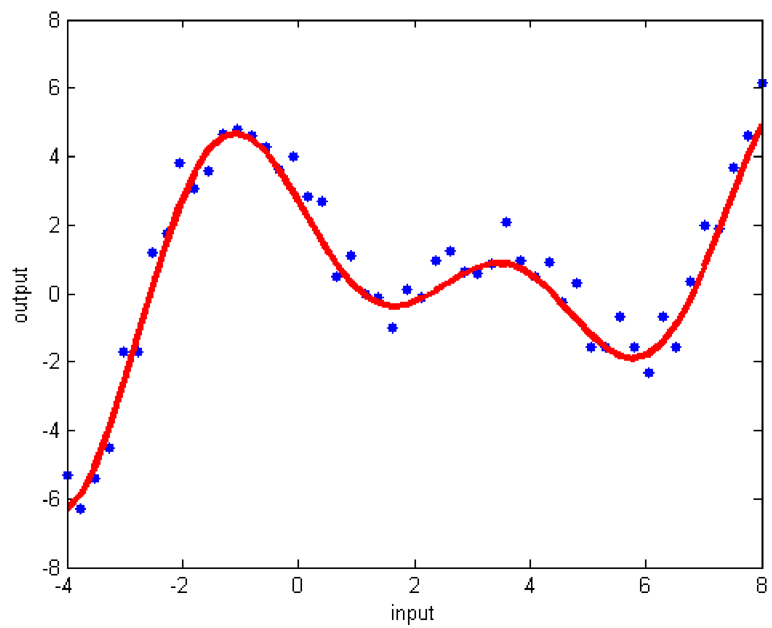


John Koza, 1992

Encoding Equations

Building Blocks: + - * / sin cos exp log … etc





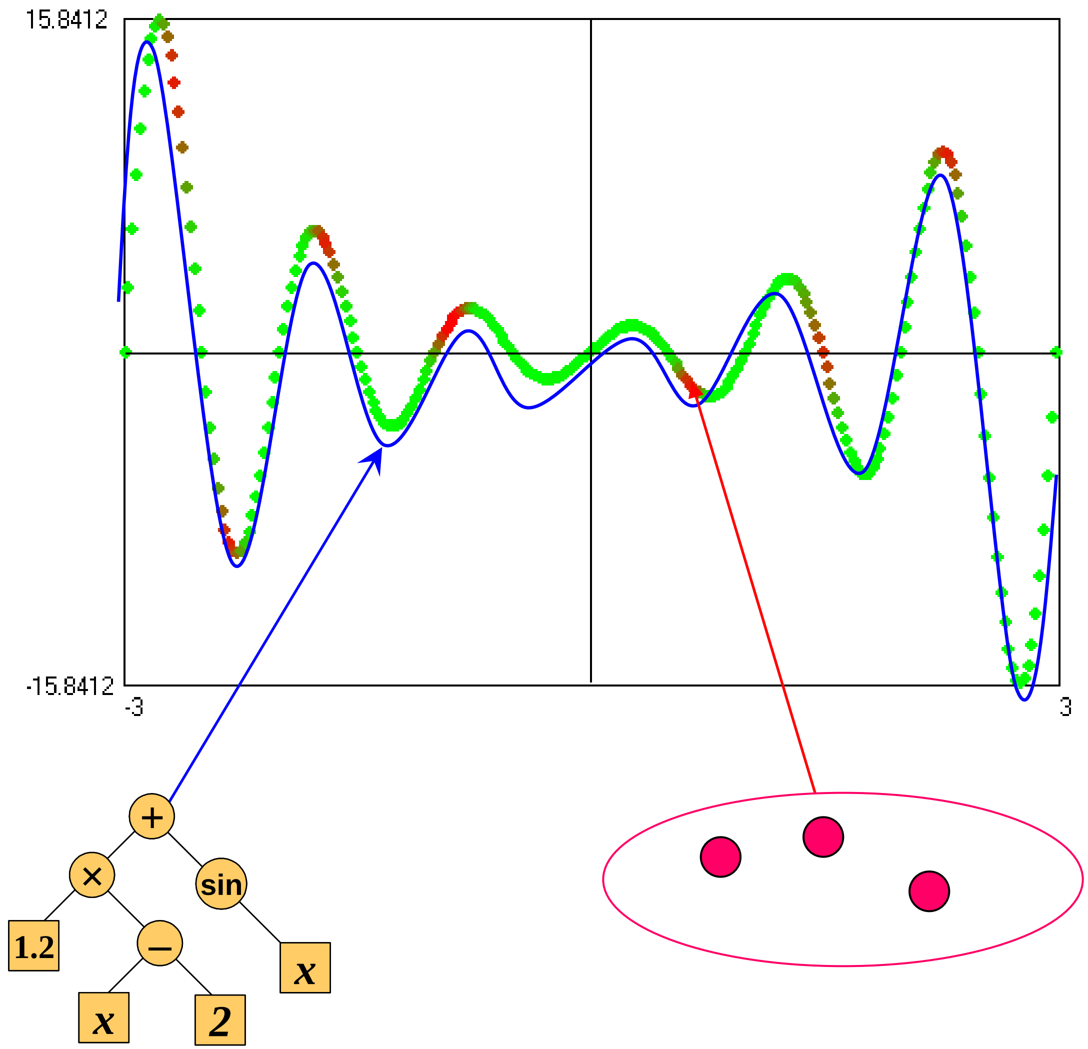
$\sin(x_2)$

$x_1 * \sin(x_2)$

$(x_1 - 3) * \sin(x_2)$

$(x_1 - 3) * \sin(-7 + x_2)$

John Koza, 1992



Models: Expression trees
Subject to mutation and selection

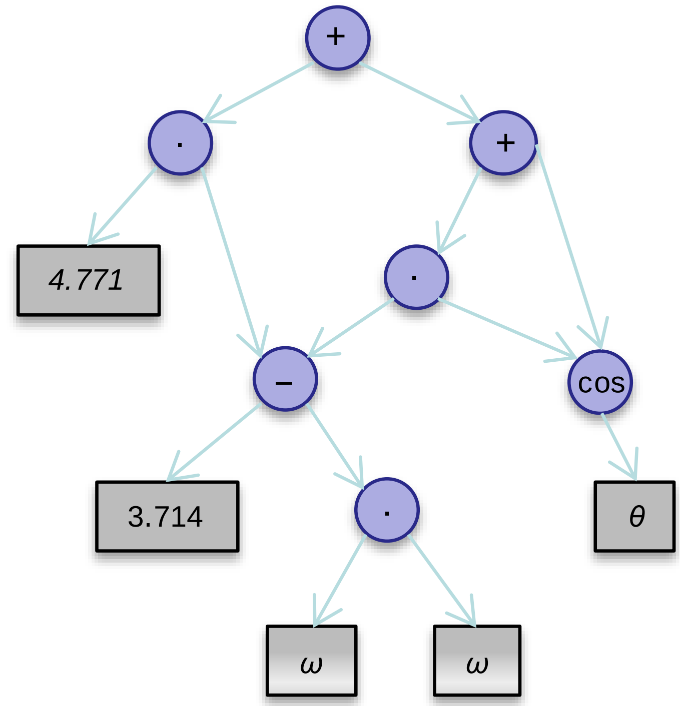
Experiments: Data-points
Subject to mutation and selection

`{const,+,-,*,/,sin,cos,exp,log,abs}`

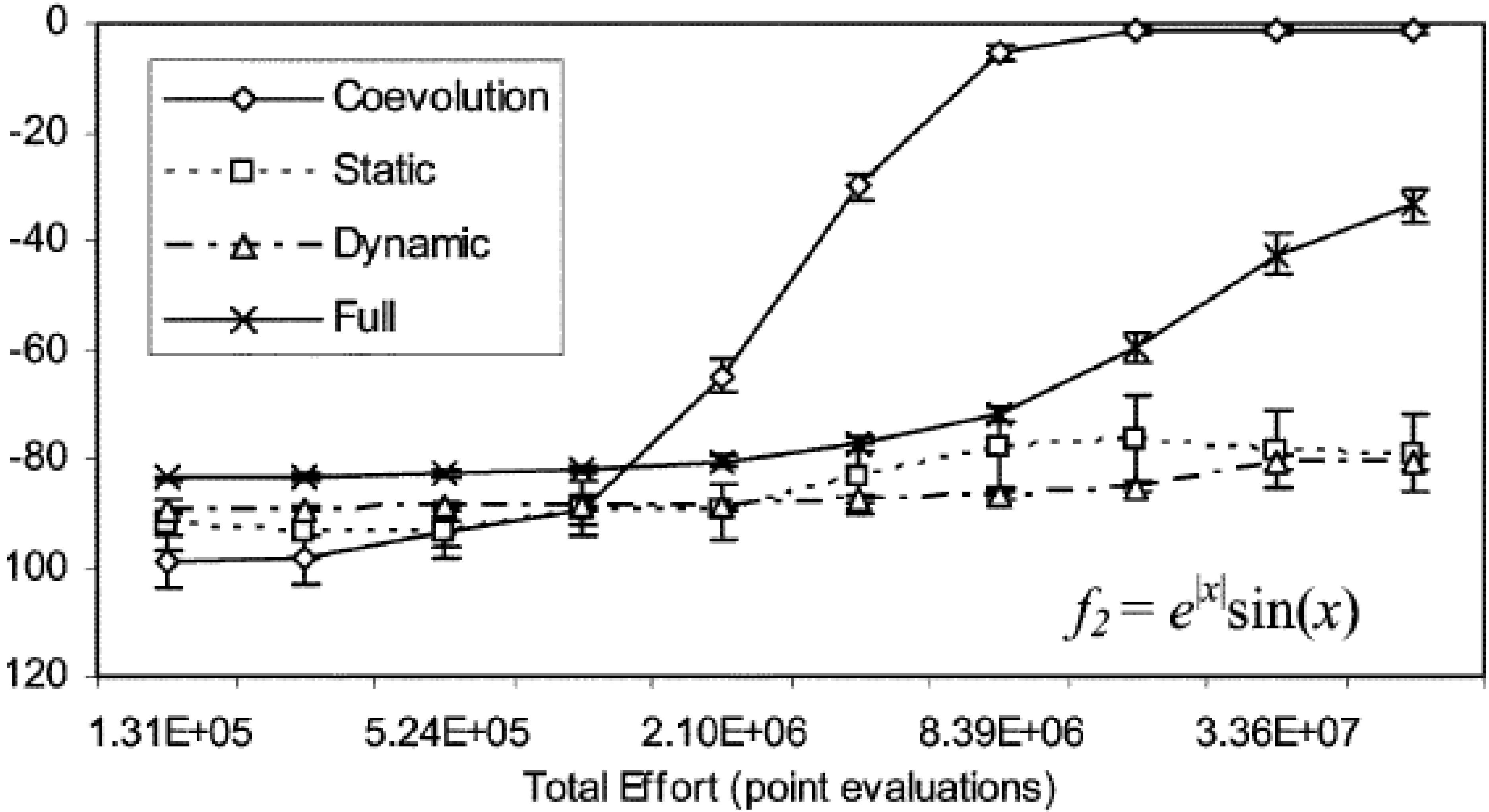
Michael D. Schmidt, Hod Lipson (2006)

$$f(\theta,\omega) = 4.771\cdot(3.714-\omega^2) + \cos(\theta) + (3.714-\omega^2)\cdot\cos(\theta)$$

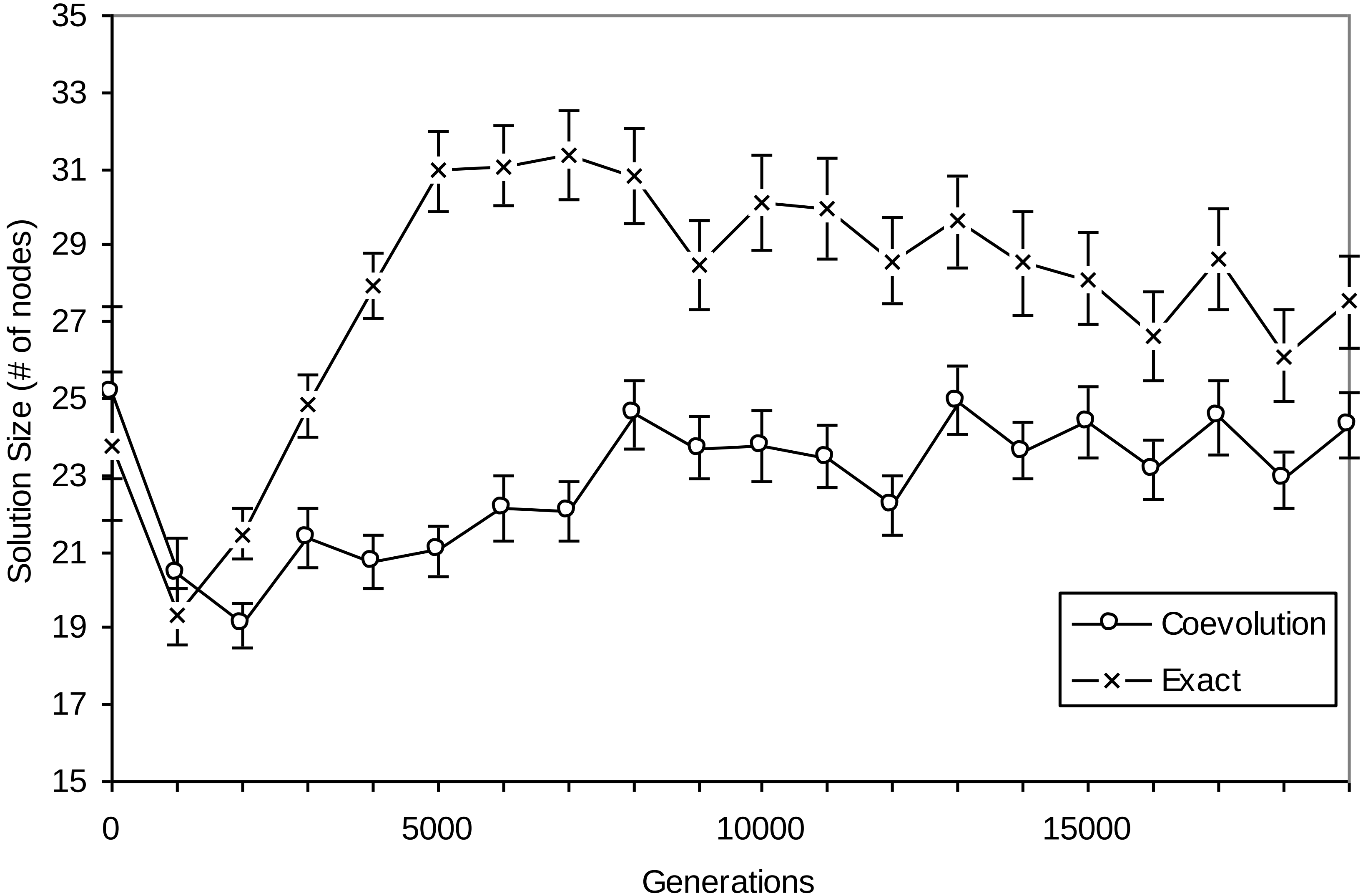
```
 (0) <- load [3.714]
 (1) <- load [ω]
 (2) <- mul  (1), (1)
 (3) <- sub  (0), (2)
 (4) <- load [θ]
 (6) <- cos  (4)
 (7) <- mul  (3), (6)
 (9) <- load [4.771]
(12) <- mul  (9), (3)
(13) <- add  (12), (6)
(15) <- add  (13), (7)
```



Solution Accuracy



Solution Complexity



GEO
LC
ORE
APM
OPM
PERC
110
69
0225
142
095

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100

Semi-empirical mass formula

Modeling the binding energy of an atomic nucleus

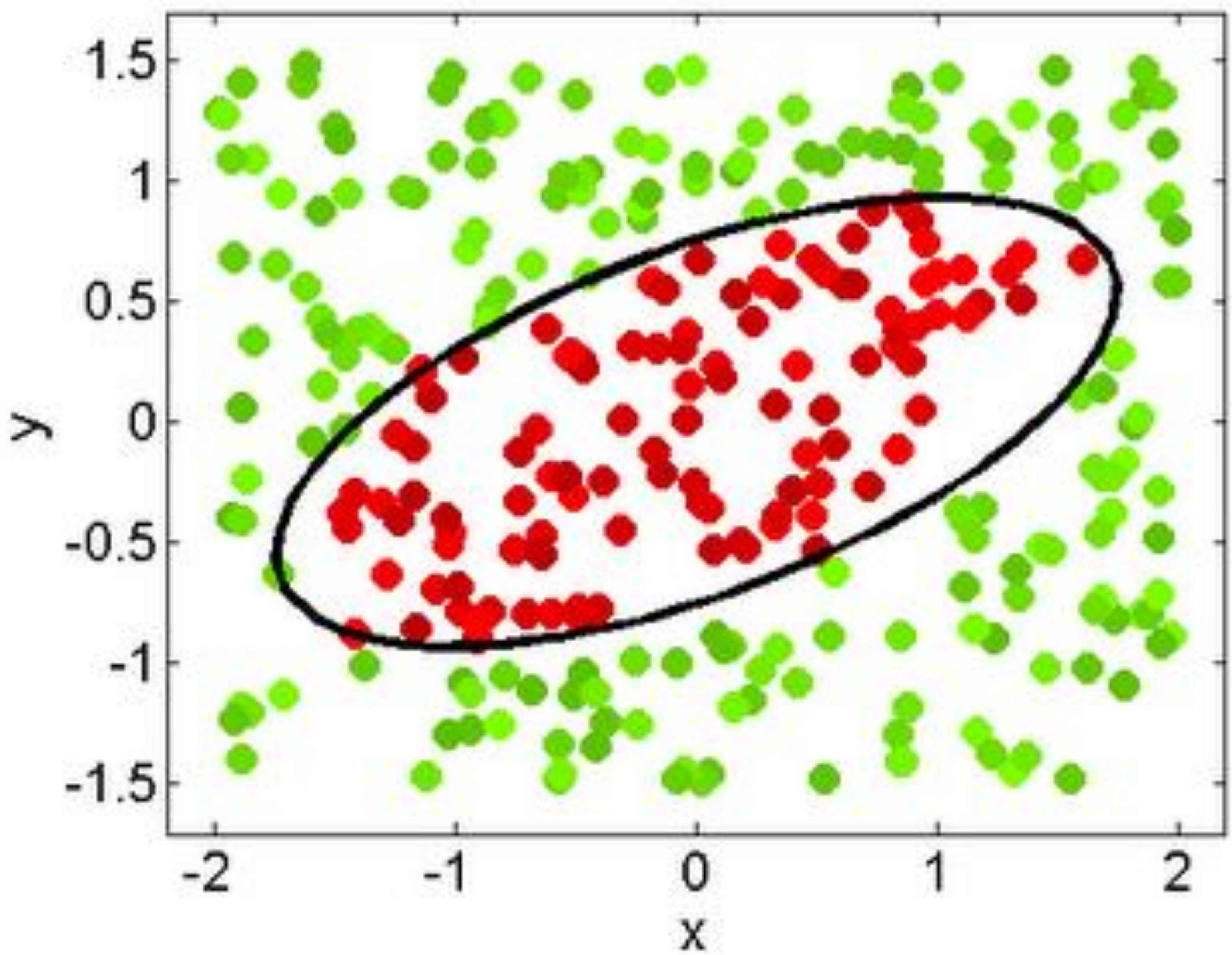
Inferred Formula:

$$E_B = 14.83 - 13.43A + 12.39A^{0.64} + \frac{0.39Z^2}{A^{0.26}} + \frac{17.29(N-Z)^2}{A} \longrightarrow \boldsymbol{R^2 = 0.99944}$$

Weizsäcker's Formula:

$$E_B = a_V A - a_S A^{2/3} - a_C \frac{Z(Z-1)}{A^{1/3}} - a_A \frac{(A-2Z)^2}{A} + \delta(A,Z) \longrightarrow \boldsymbol{R^2 = 0.999915}$$

$$\delta(A,Z) = \begin{cases} +\delta_0 & Z, N \text{ even} \\ 0 & A \text{ odd} \\ -\delta_0 & Z, N \text{ odd} \end{cases} \qquad \delta_0 = \frac{a_P}{A^{1/2}}$$


1.5
1
0.5
0
-0.5
-1
-1.5
y
-2
-1
0
1
2
x

Systems of Differential Equations

- Regress on derivative

	State Variables			*Derivatives*		
time	**x_1**	**x_2**	**…**	**dx_1/dt**	**x_2/dt**	**…**
0	3.4	-1.7	…	-2.0	8.0	…
0.1	3.2	-0.9	…	-1.0	8.0	…
0.2	3.1	-0.1	…	-4.0	1.3	…
0.3	2.7	1.2	…	-5.7	1.9	…
…	…	…	…	…	…	…

Inferring Biological Networks

$$\frac{dS_1}{dt} = \underbrace{2.5}_{J_0} - \underbrace{100\left(\frac{S_1 * A_3}{1+13.6769 * A_3^4}\right)}_{-v_1}$$

$$\frac{dS_2}{dt} = \underbrace{200\left(\frac{S_1 * A_3}{1+13.6769 * A_3^4}\right)}_{2 * v_1} - \underbrace{6 * S_2 * N_1}_{-v_2} - \underbrace{12 * S_2 * N_1}_{-v_6}$$

$$\frac{dS_3}{dt} = \underbrace{6 * S_2 * N_1}_{v_2} + \underbrace{16 * S_3 * A_2}_{v_3}$$

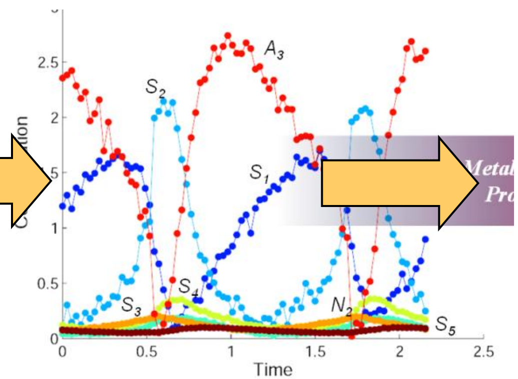
$$\frac{dS_4}{dt} = \underbrace{16 * A_2 * S_3}_{v_3} - \underbrace{100 * N_2 * S_4}_{-v_4}$$

$$\frac{dN_2}{dt} = \underbrace{6 * S_2 * N_1}_{v_2} - \underbrace{100 * N_2 * S_4}_{-v_4}$$

$$\frac{dA_3}{dt} = \underbrace{-200\left(\frac{S_1 * A_3}{1+13.6769 A_3^4}\right)}_{-2 * v_1} + \underbrace{32 * A_2 * S_3}_{2 v_3} - \underbrace{1.28 * A_3}_{-v_5}$$

$$\frac{dS_5}{dt} = \underbrace{-1.3 * S_5}_{\varphi J}$$

Original Equations



$$\frac{dS_1}{dt} = \underbrace{2.42114}_{J_0} - \underbrace{99.2721\left(\frac{S_1 * A_3}{1+13.5956 * A_3^4}\right)}_{-v_1}$$

$$\frac{dS_2}{dt} = \underbrace{199.935\left(\frac{S_1 * A_3}{1+13.6734 * A_3^4}\right)}_{2 * v_1} - \underbrace{5.99475 * S_2 * N_1}_{-v_2} - \underbrace{11.9895 * S_2 * N_1}_{-v_6}$$

$$\frac{dS_3}{dt} = \underbrace{5.99857 * S_2 * N_1}_{v_2} + \underbrace{15.99606 * S_3 * A_2}_{v_3} - \underbrace{0.01286 * S_3}_{extraneous}$$

$$\frac{dS_4}{dt} = \underbrace{15.997 * A_2 * S_3}_{v_3} - \underbrace{100.015 * N_2 * S_4}_{-v_4}$$

$$\frac{dN_2}{dt} = \underbrace{5.99857 * S_2 * N_1}_{v_2} - \underbrace{99.9963 * N_2 * S_4}_{-v_4}$$

$$\frac{dA_3}{dt} = \underbrace{-197.781\left(\frac{S_1 * A_3}{1+13.2633 A_3^4}\right)}_{-2 * v_1} + \underbrace{31.9682 * A_2 * S_3}_{2 v_3} - \underbrace{1.29659 * A_3}_{-v_5}$$

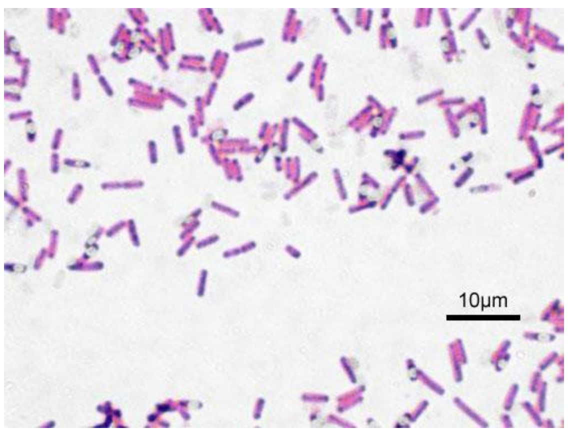
$$\frac{dS_5}{dt} = \underbrace{-1.29626 * S_5}_{\varphi J}$$

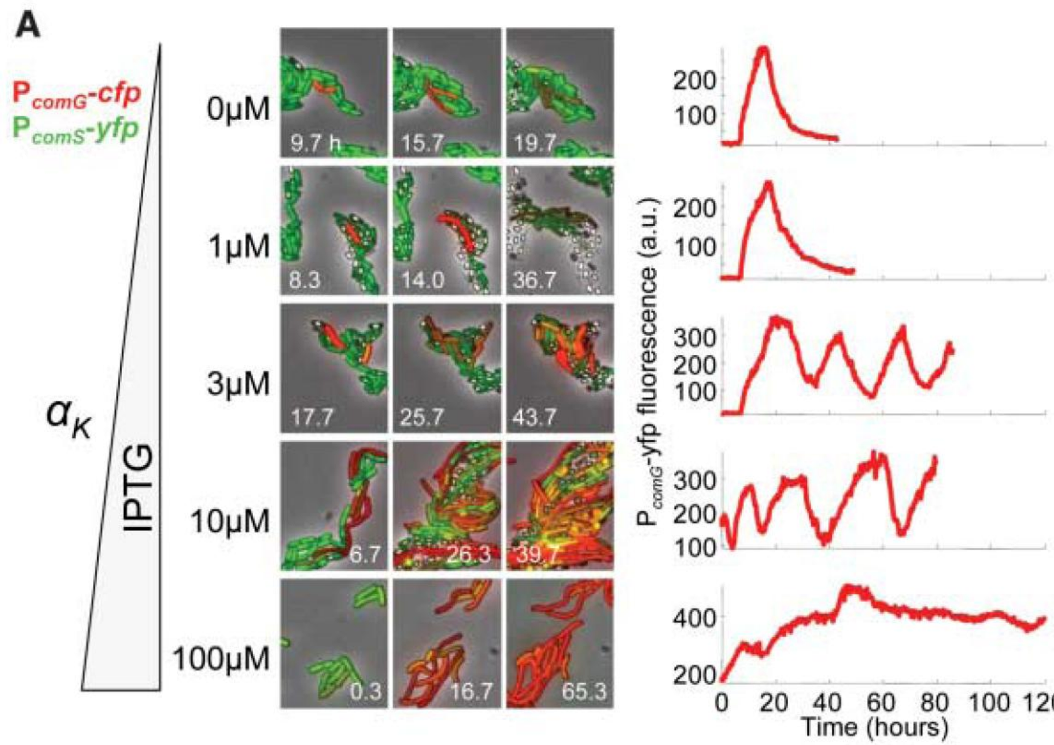
Inferred Equations

With Michael Schmidt, John Wikswo (Vanderbilt), Jerry Jenkins (CFDRC)

Wet Data, Unknown System

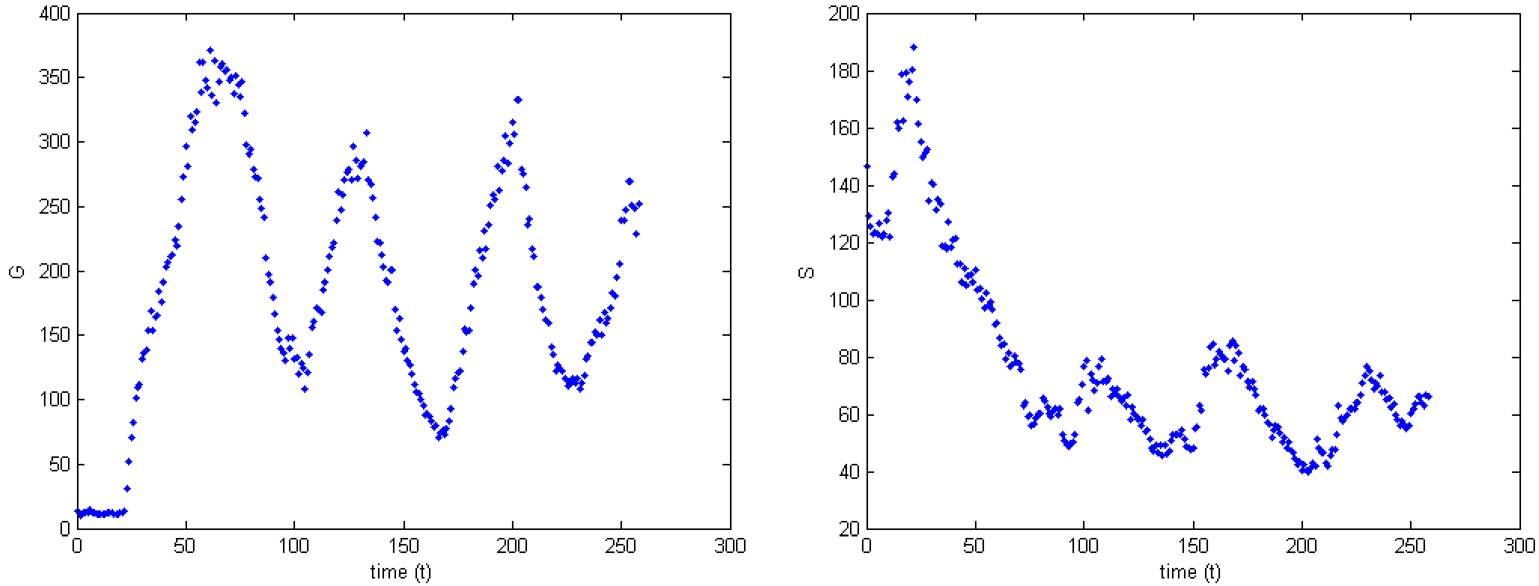
Bacillus Bacteria



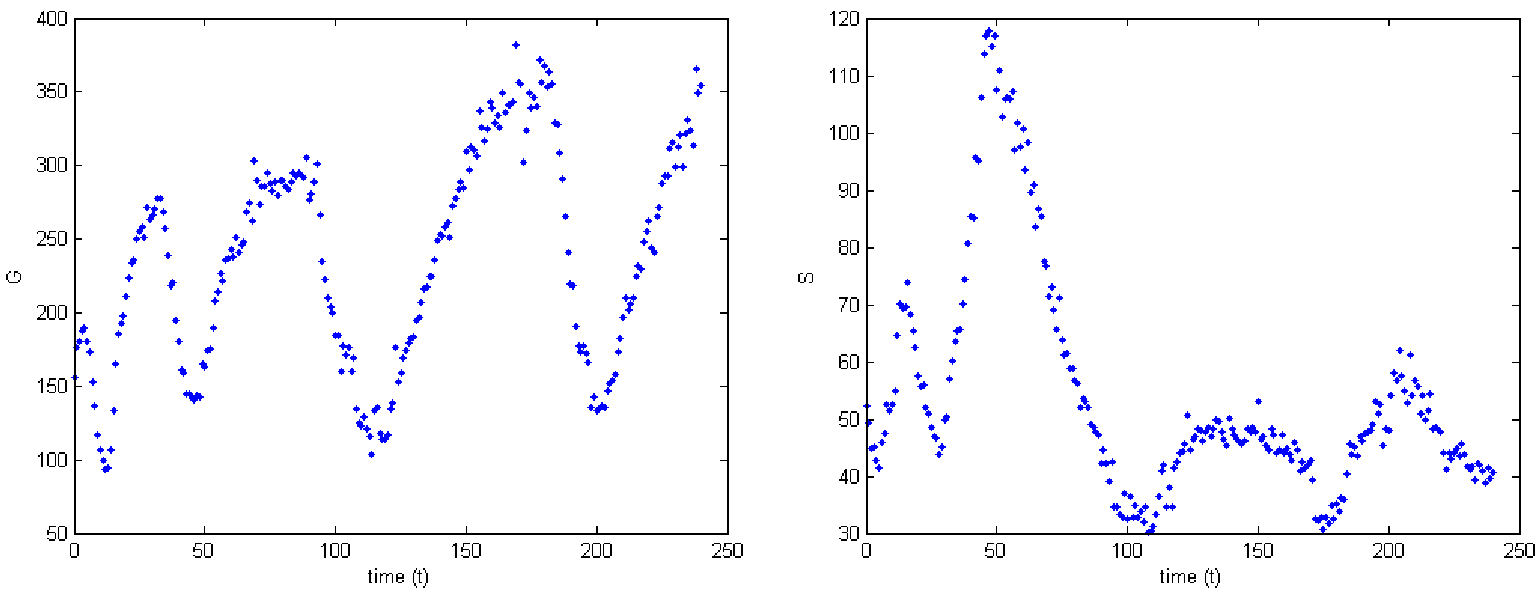


With Michael Schmidt (Cornell) and Gurol Suel (UT Southwestern)

Cell #1



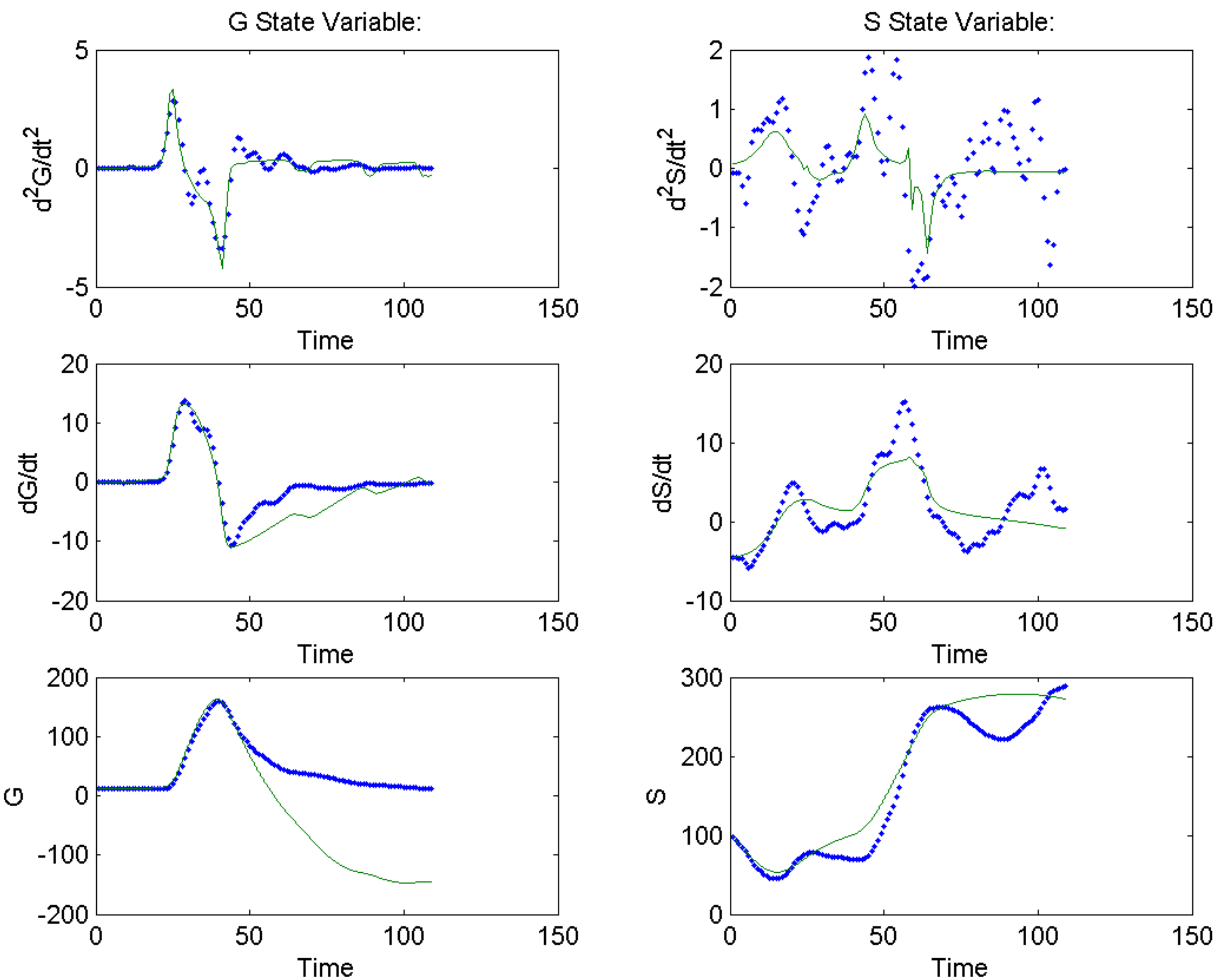
Cell #2



Cell #3-60 …

$$\frac{d^2 G}{dt^2} = 0.329708 - \frac{0.494562 \left(G - \frac{dG\, S^2}{(dG-G)^2} \right)}{19.75 + 2\, dG + e^{\frac{dG\, S}{dG-G}}}$$

$$\frac{d^2 S}{dt^2} = -0.0949 + \frac{0.511803}{-36.951 + 4\, dG + G} + \frac{13.1334 - 1.46\, dG}{-1.42633\, (18.96 - dG) + S}$$



Blue Dots = data points, Green Line = regressed fit

Symbolic Regression Inferred *Time-Delay* Model:

$$\frac{dK}{dt} = a_K + \frac{b_K + c_K S}{K}$$

$$\frac{dS}{dt} = a_S + \frac{b_S + c_S K}{S}$$

Biologist's Inferred Model: Gurol Suel, et. al., Science 2007

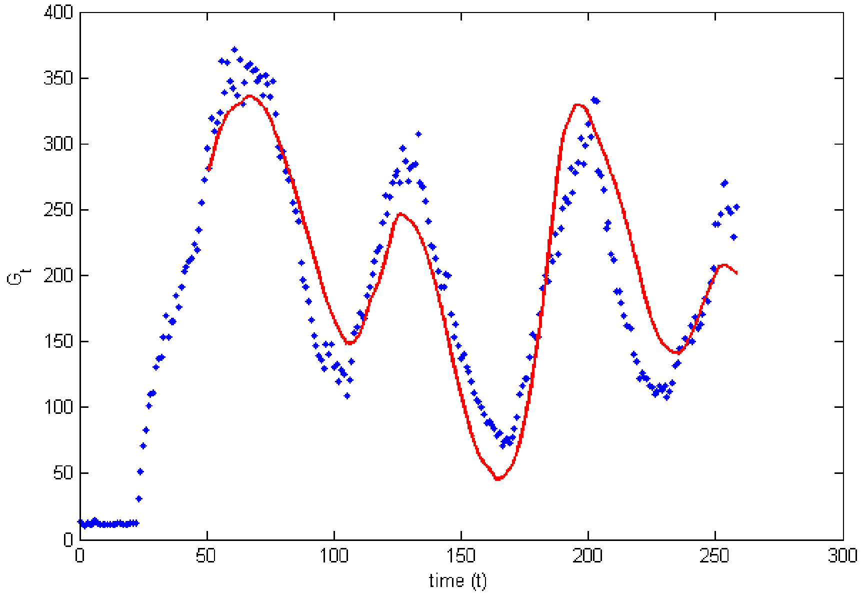
$$\frac{d\boldsymbol{K}}{d\boldsymbol{t}} = \alpha_{\boldsymbol{k}} + \frac{\beta_{\boldsymbol{K}} \boldsymbol{K}^{\boldsymbol{n}}}{\boldsymbol{k}_0^{\boldsymbol{n}} + \boldsymbol{K}^{\boldsymbol{n}}} - \frac{\delta_{\boldsymbol{K}} \boldsymbol{K}}{1 + \boldsymbol{K} / \Gamma_{\boldsymbol{K}} + \boldsymbol{S} / \Gamma_{\boldsymbol{S}}} - \lambda_{\boldsymbol{K}} \boldsymbol{K}$$

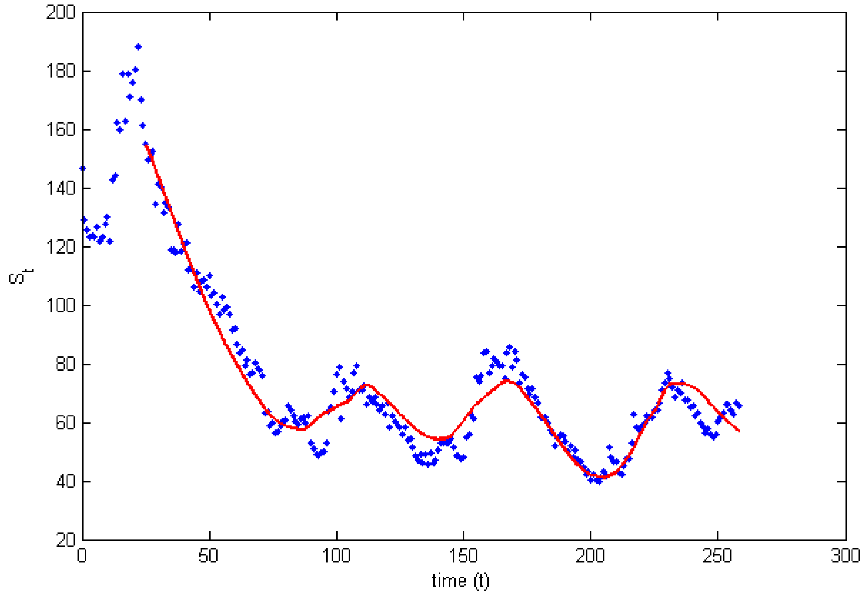
$$\frac{d\boldsymbol{S}}{d\boldsymbol{t}} = \alpha_{\boldsymbol{S}} + \frac{\beta_{\boldsymbol{S}}}{1 + \left(\boldsymbol{K} / \boldsymbol{k}_1\right)^{\boldsymbol{p}}} - \frac{\delta_{\boldsymbol{k}} \boldsymbol{S}}{1 + \boldsymbol{K} / \Gamma_{\boldsymbol{K}} + \boldsymbol{S} / \Gamma_{\boldsymbol{S}}} - \lambda_{\boldsymbol{S}} \boldsymbol{S}$$

Withheld Test Set #1 Fit

$$\frac{dG_t}{dt} = \frac{1582.0 + 17.3214 \cdot S_{t-51}}{G_{t-18}} - 16.7423$$

$$\frac{dS_t}{dt} = \frac{114.922 + 0.3019 \cdot G_{t-25}}{S_{t-15}} - 3.05$$

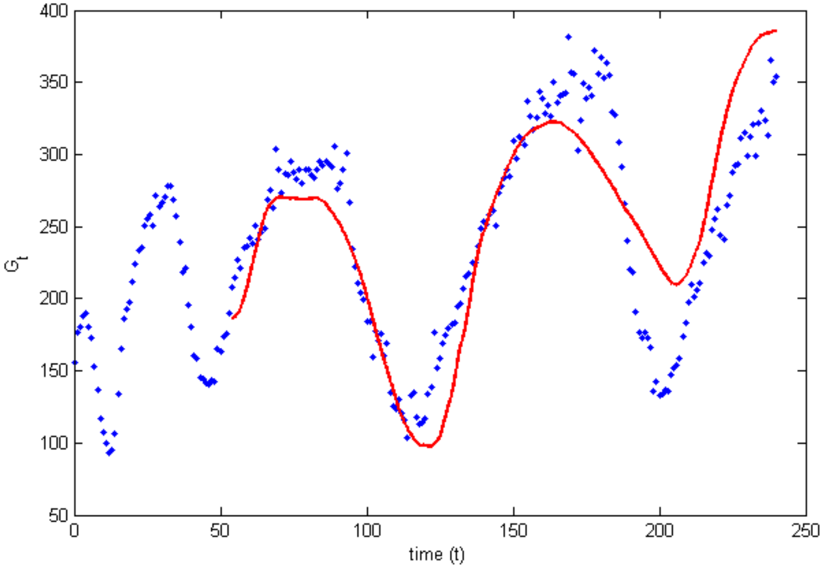


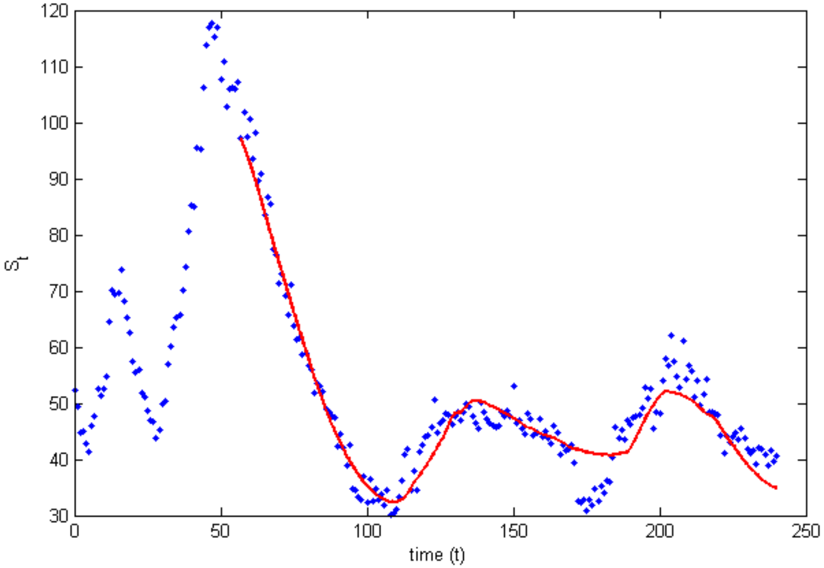


Withheld Test Set #2 Fit

$$\frac{dG_t}{dt} = \frac{3526.92 - 21.312 \cdot S_{t-54}}{G_{t-17}} - 10.1355$$

$$\frac{dS_t}{dt} = \frac{132.271 - 0.0178 \cdot G_{t-57}}{S_{t-18}} - 2.9693$$

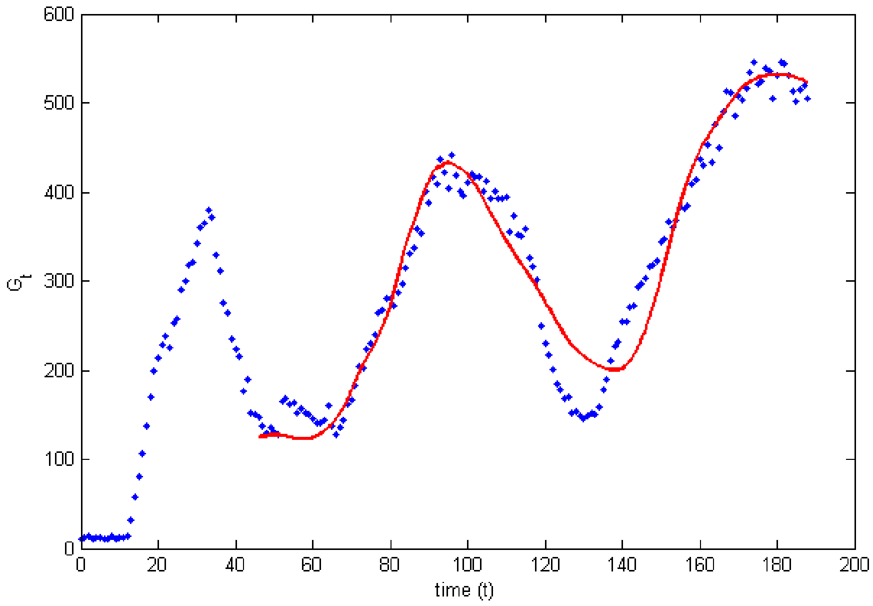


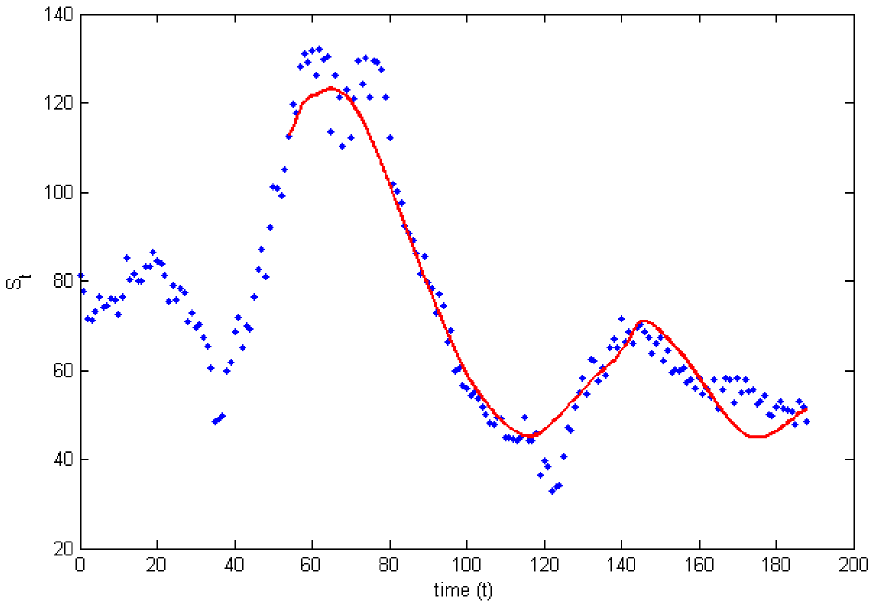


Withheld Test Set #3 Fit

$$\frac{dG_t}{dt} = \frac{5057.1 - 39.7452 \cdot S_{t-46}}{G_{t-21}} - 6.4406$$

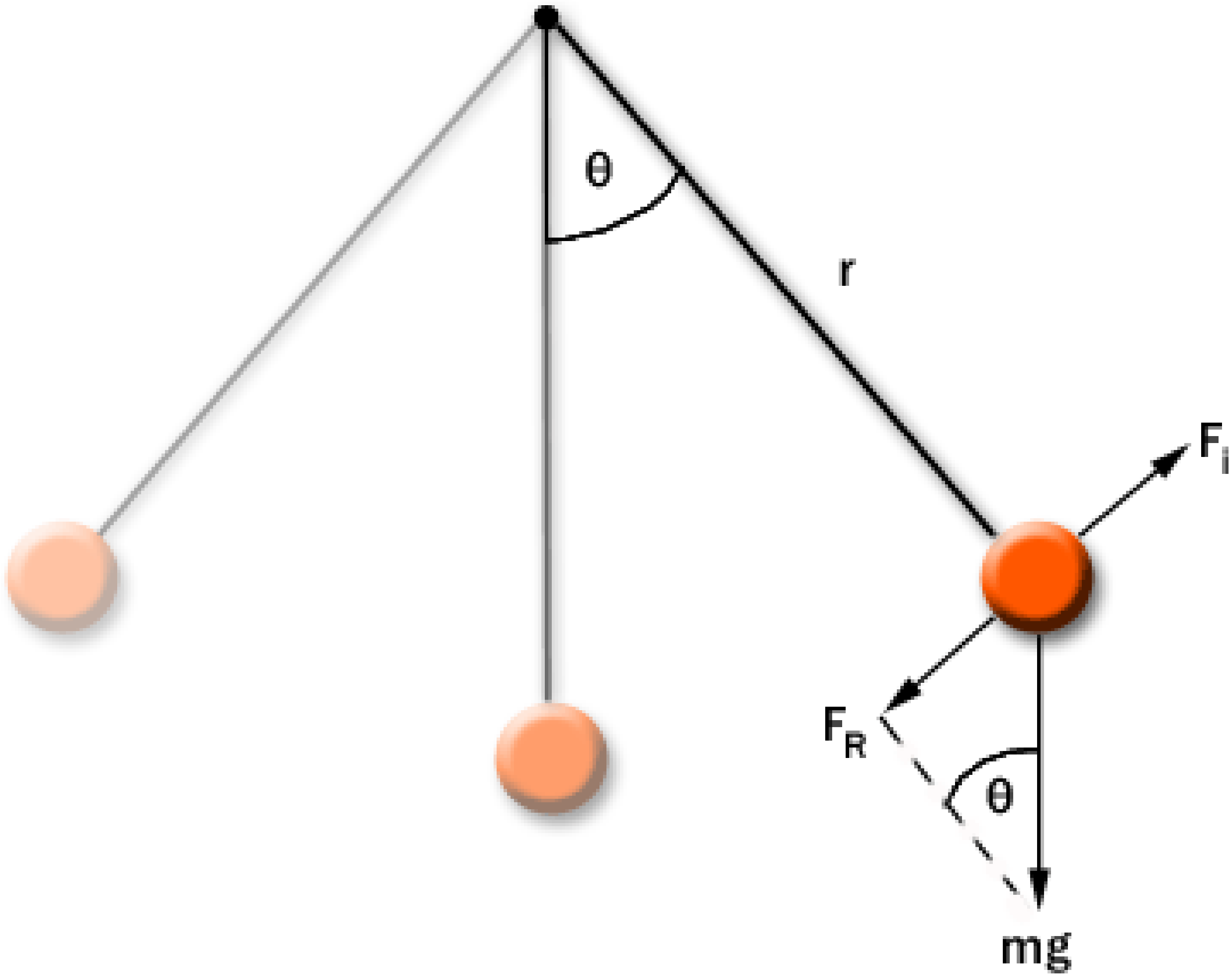
$$\frac{dS_t}{dt} = \frac{295.426 - 0.2965 \cdot G_{t-54}}{S_{t-20}} - 3.871$$

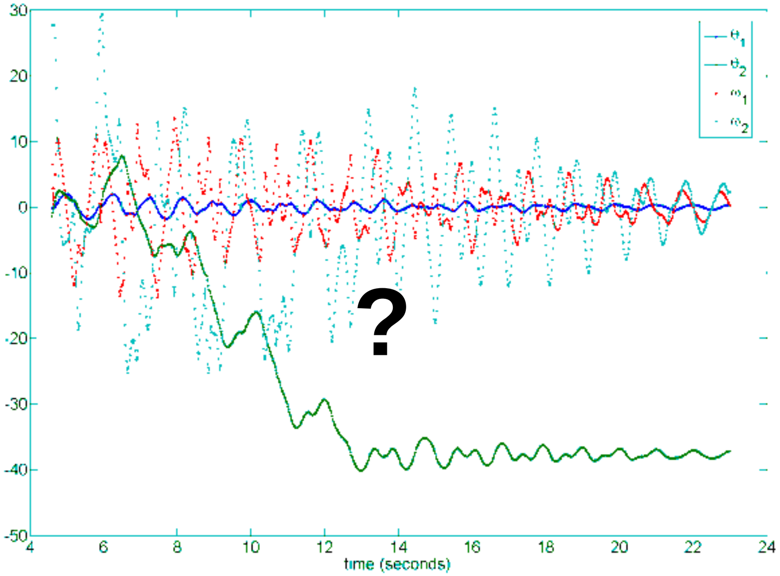




Looking For Invariants

Data Mining


θ
r
F_i
F_R
θ
mg



42

42+x-x

$42+1/(1000+x^2)$

From Data:

x	*y*	…
0.1	2.3	
0.2	4.5	
0.3	9.7	
0.4	5.1	
0.5	3.3	
0.6	1.0	
…	…	…

Calculate partial derivatives Numerically:

$$\longrightarrow \quad \frac{\delta x}{\delta y}, \quad \frac{\delta y}{\delta x}, \quad \ldots$$

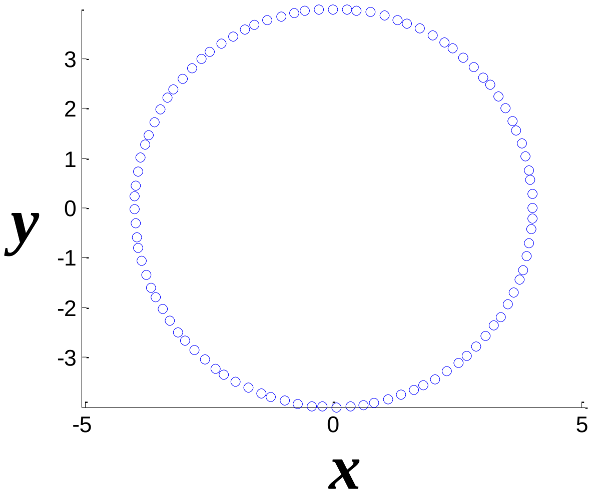
From Equation:

Calculate predicted partial derivatives Symbolically:

$$\frac{\delta f}{\delta x} \Big/ \frac{\delta f}{\delta y} \longrightarrow \quad \frac{\delta x'}{\delta y'}, \quad \frac{\delta y'}{\delta x'}, \quad \ldots$$

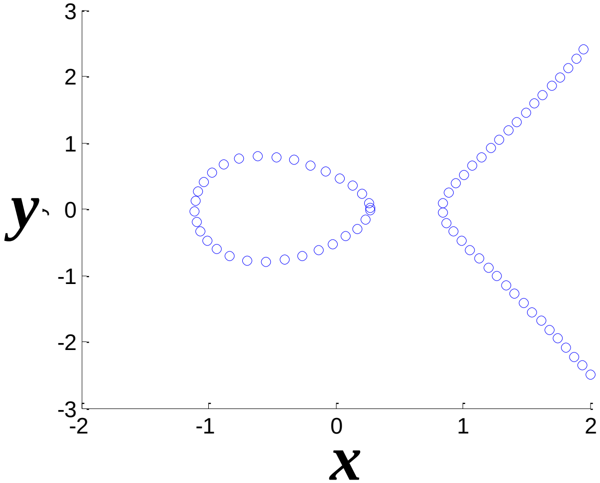
Experiments

Circle



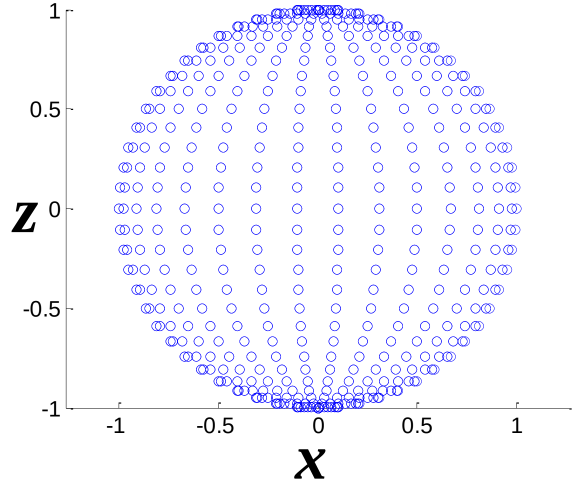
$$x^2 + y^2 - 16 = 0$$

Elliptic Curve

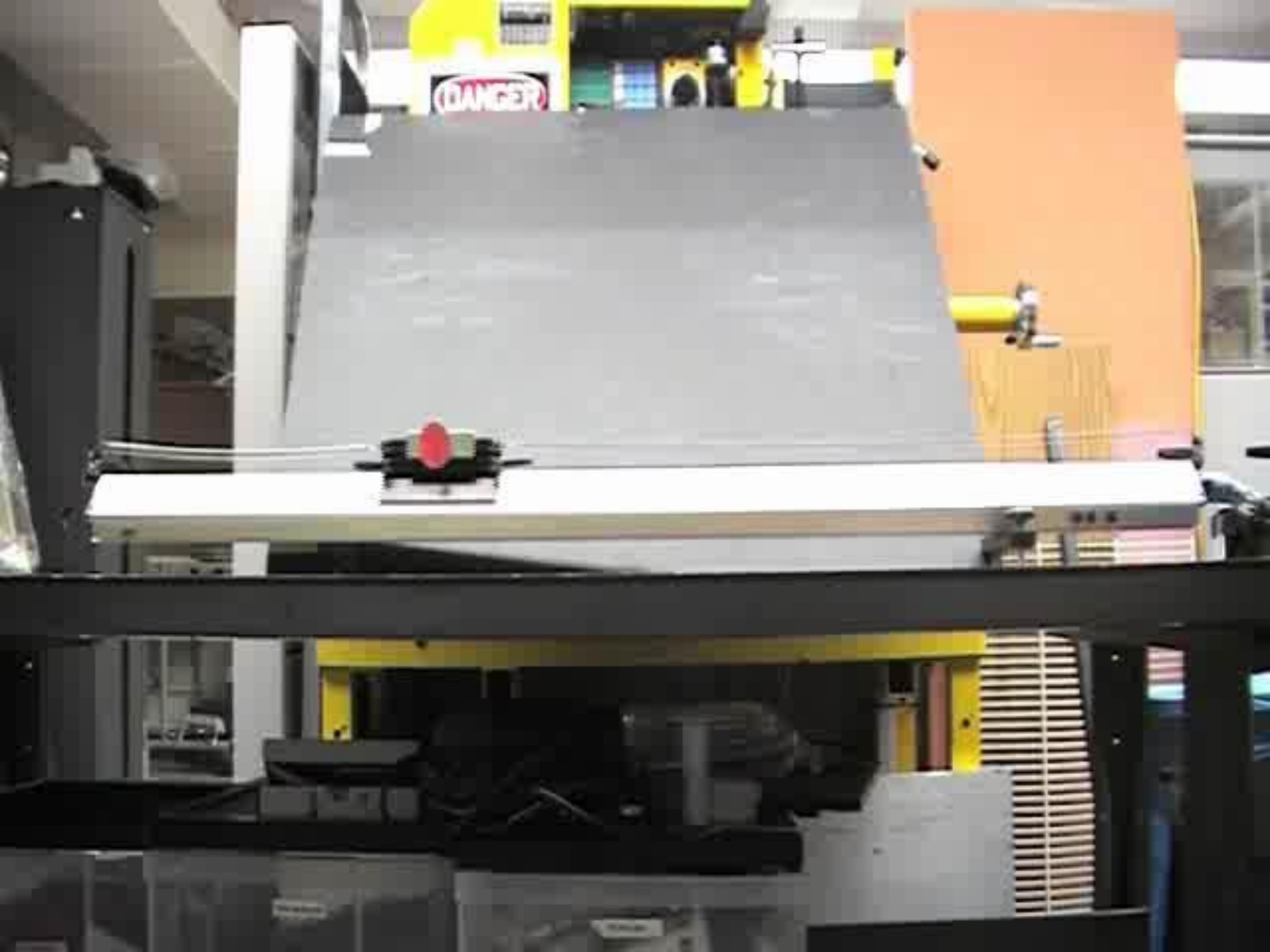


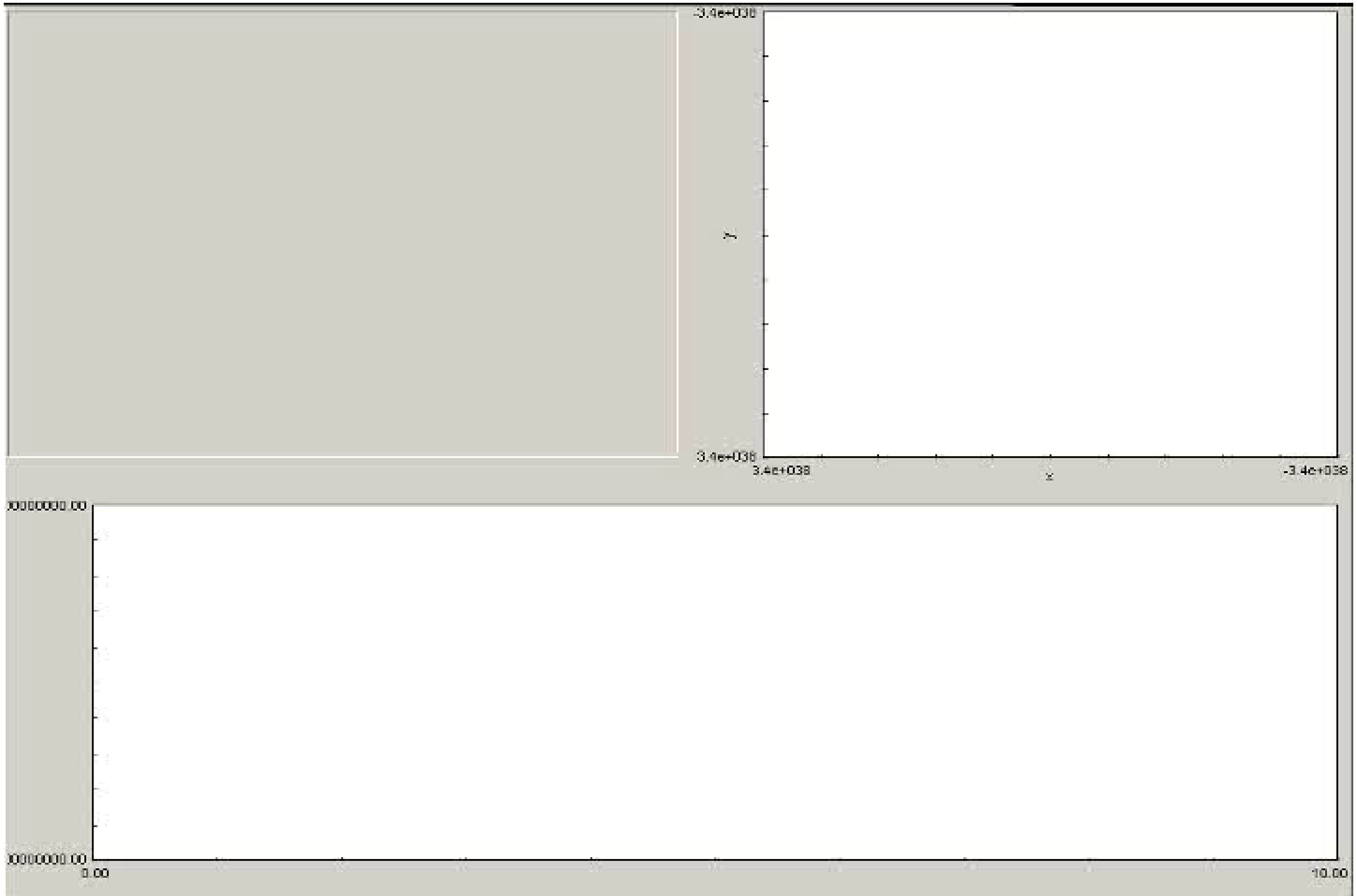
$$x^3 + x - y^2 - 1.5 = 0$$

Sphere



$$x^2 + y^2 + z^2 - 1 = 0$$

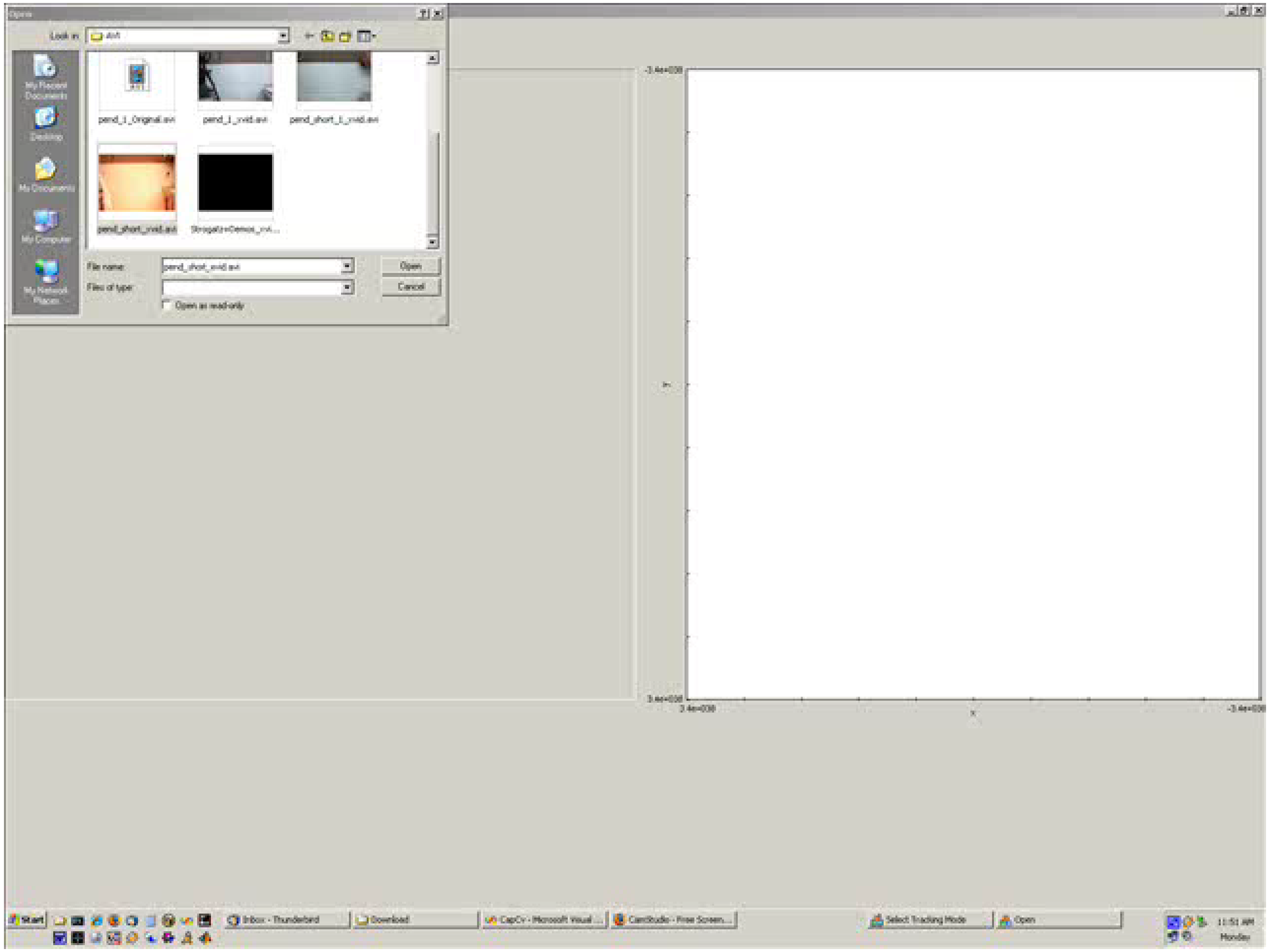

DANGER



$$H = 114.28 * \left(\frac{dx}{dt} \right)^2 + 692.322 * x^2$$

$$L = 61.591 * \left(\frac{dx}{dt} \right)^2 - 369.495 * x^2$$

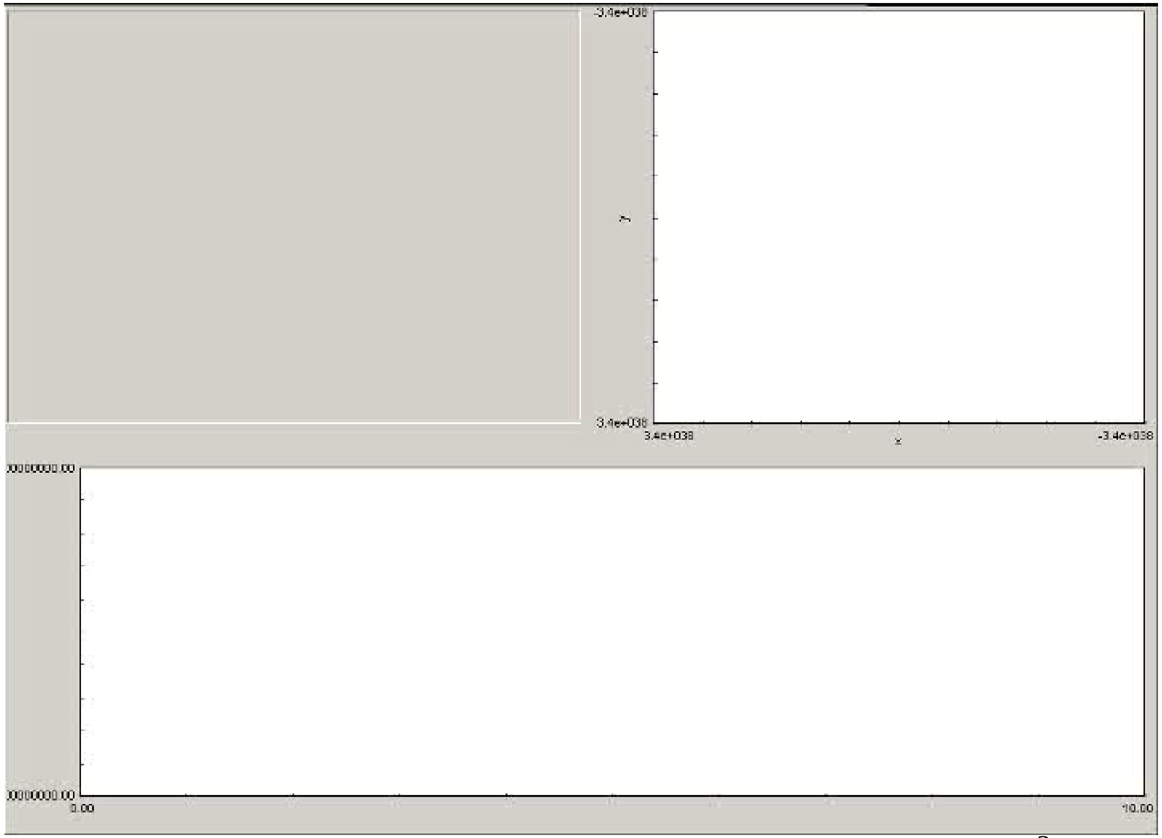
• Coefficients may have different scales and offsets each run



$$\mathbf{H} = \left(\frac{d\theta}{dt}\right)^2 + 2.42847 * \cos(\theta)$$

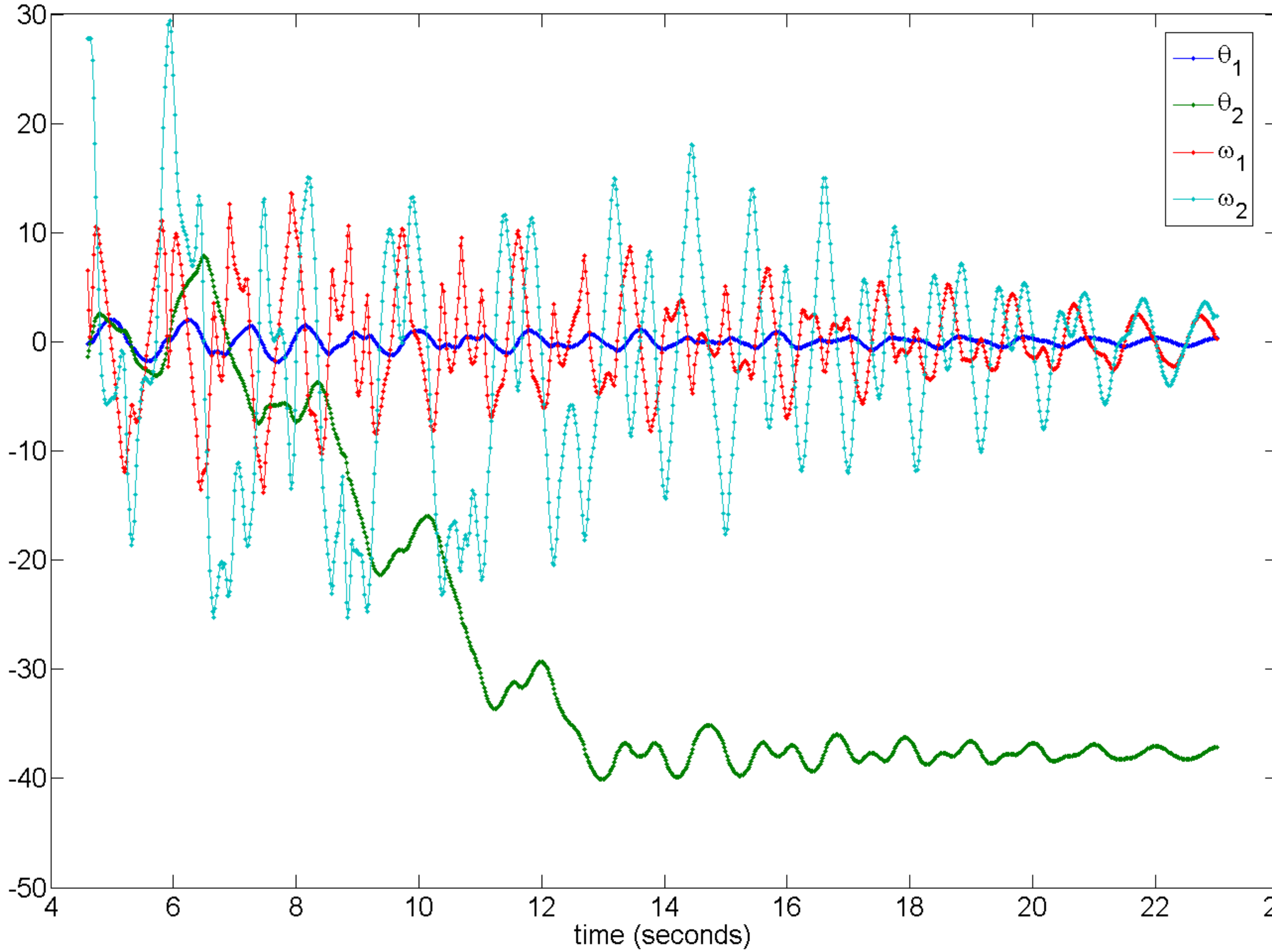
$$\mathbf{L} = 3.52768 * \left(\frac{d\theta}{dt}\right)^2 - 9.43429 * \cos(\theta)$$

Double Linear Oscillator

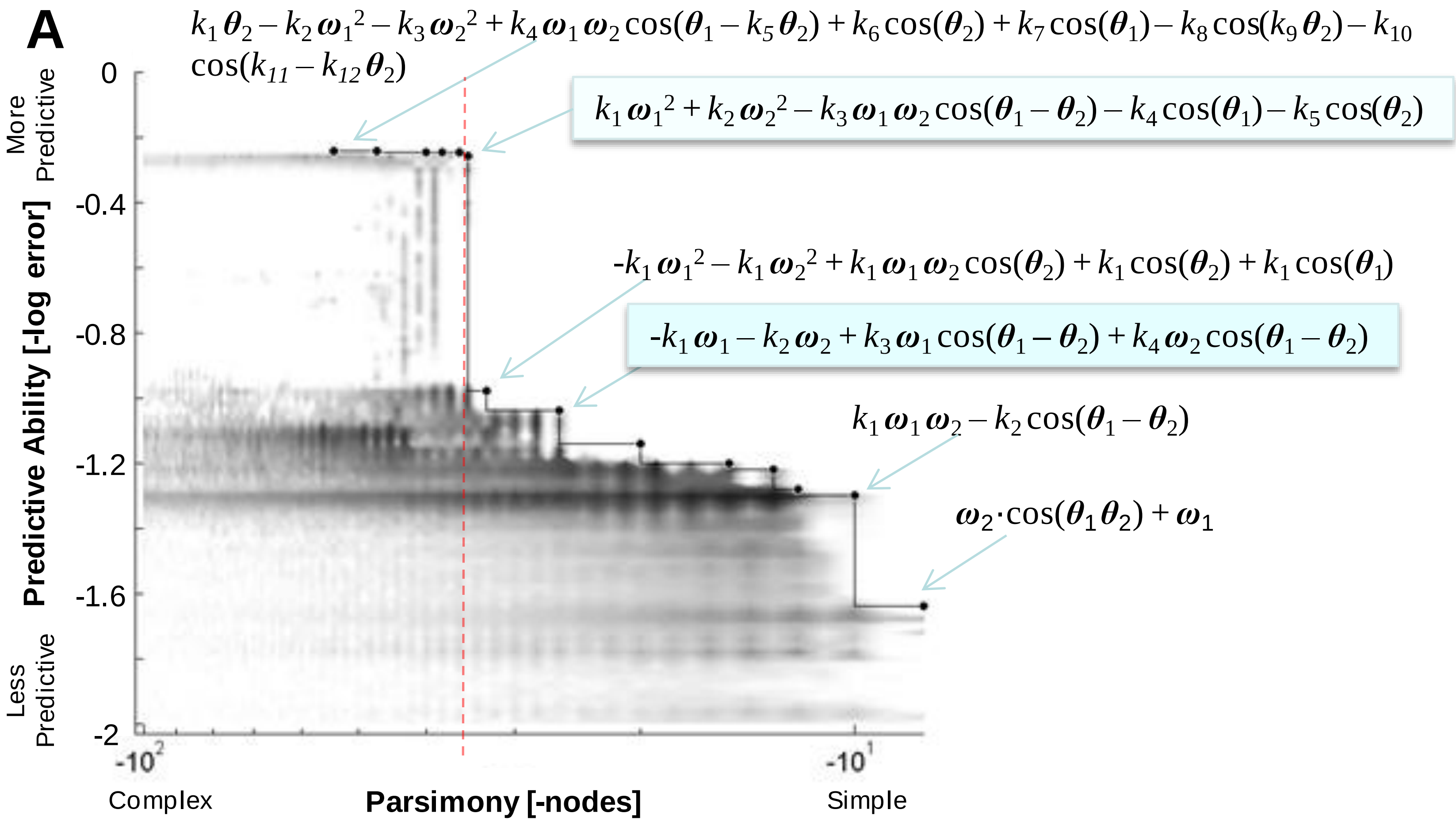


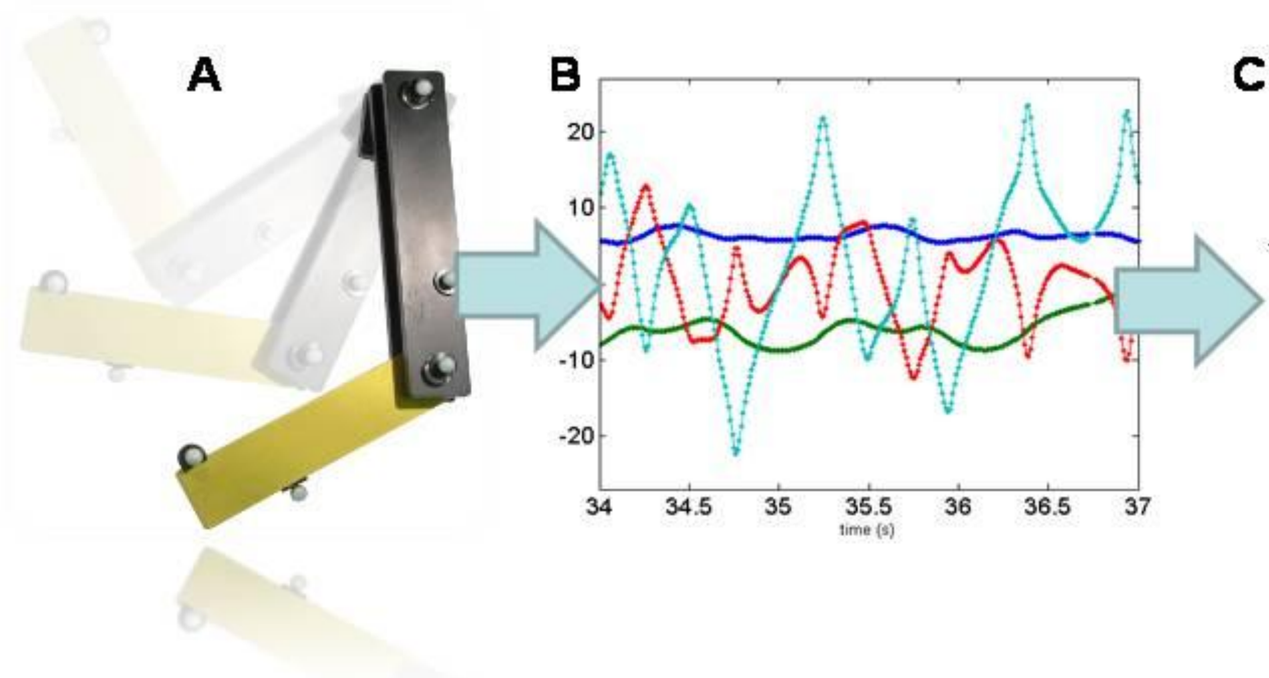
$$H = -14.691 * x_1^2 - 15.551 * x_2^2 - 21.676 * x_1 x_2 + 8.3808 * \left(\frac{dx_2}{dt}\right)^2 + 2.6046 * \left(\frac{dx_1}{dt}\right)^2$$

would be plus for Lagrangian



30
20
10
0
-10
-20
-30
-40
-50
4
6
8
10
12
14
16
18
20
22
24
time (seconds)
θ_1
θ_2
ω_1
ω_2

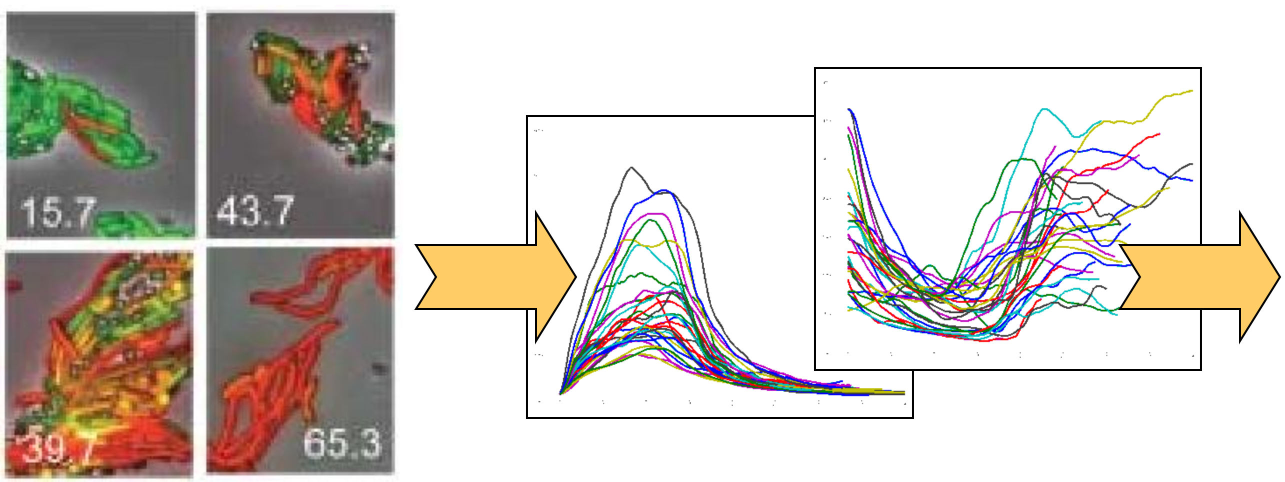

A
$k_1\theta_2 - k_2\omega_1^2 - k_3\omega_2^2 + k_4\omega_1\omega_2\cos(\theta_1 - k_5\theta_2) + k_6\cos(\theta_2) + k_7\cos(\theta_1) - k_8\cos(k_9\theta_2) - k_{10}\cos(k_{11} - k_{12}\theta_2)$
$k_1\omega_1^2 + k_2\omega_2^2 - k_3\omega_1\omega_2\cos(\theta_1 - \theta_2) - k_4\cos(\theta_1) - k_5\cos(\theta_2)$
$-k_1\omega_1^2 - k_1\omega_2^2 + k_1\omega_1\omega_2\cos(\theta_2) + k_1\cos(\theta_2) + k_1\cos(\theta_1)$
$-k_1\omega_1 - k_2\omega_2 + k_3\omega_1\cos(\theta_1 - \theta_2) + k_4\omega_2\cos(\theta_1 - \theta_2)$
$k_1\omega_1\omega_2 - k_2\cos(\theta_1 - \theta_2)$
$\omega_2\cdot\cos(\theta_1\theta_2) + \omega_1$
More Predictive
Less Predictive
Predictive Ability [-log error]
0
-0.4
-0.8
-1.2
-1.6
-2
-10^2
-10^1
Complex
Parsimony [-nodes]
Simple



C

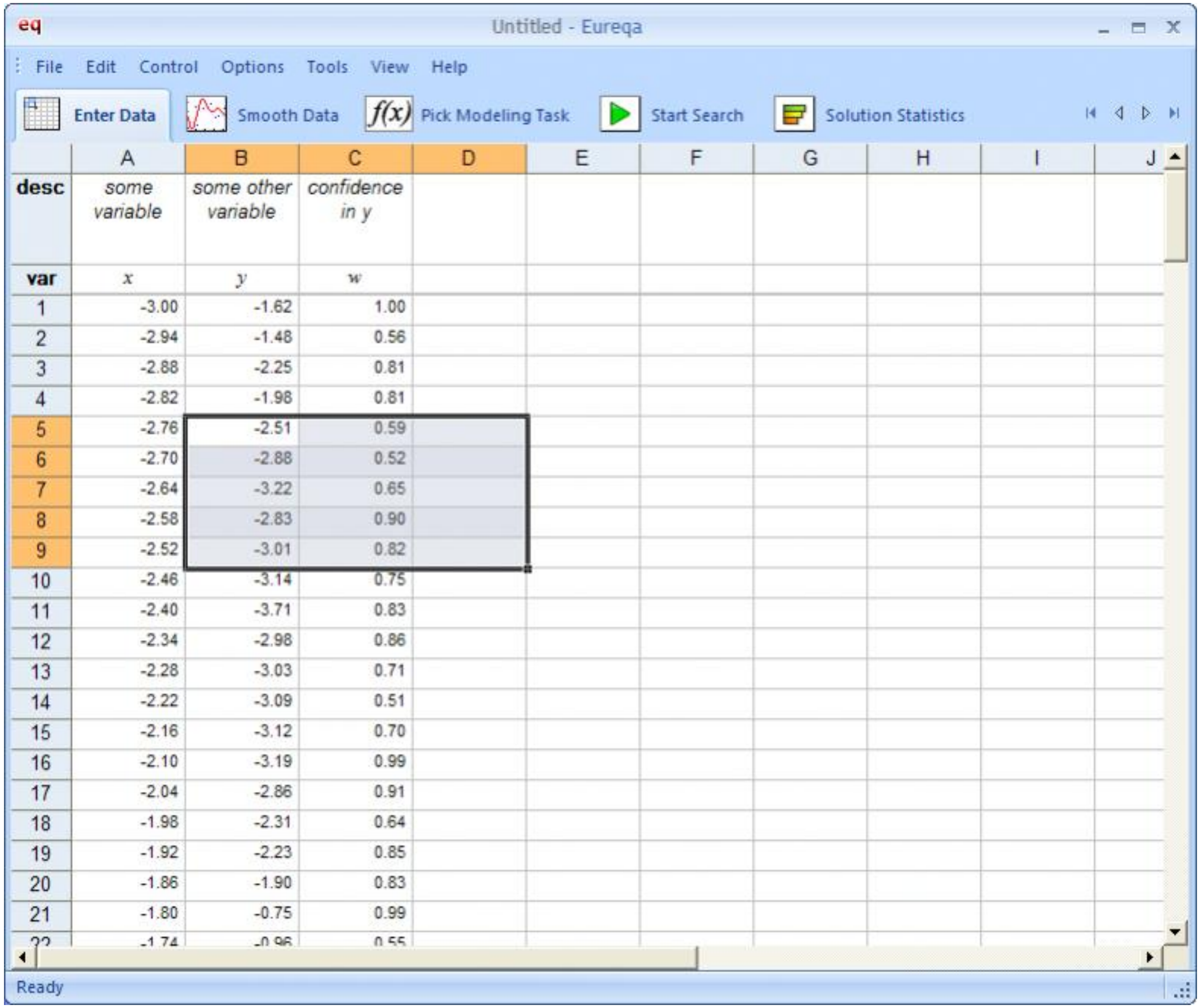
Detected Invariance:

$$L_1^{\,2}(m_1+m_2)\omega_1^{\,2} + m_2L_2^{\,2}\omega_2^{\,2} + m_2L_1L_2\omega_1\omega_2\cos(\theta_1-\theta_2) - 19.6L_1(m_1+m_2)\cos\theta_1 - 19.6m_2L_2\cos\theta_2$$



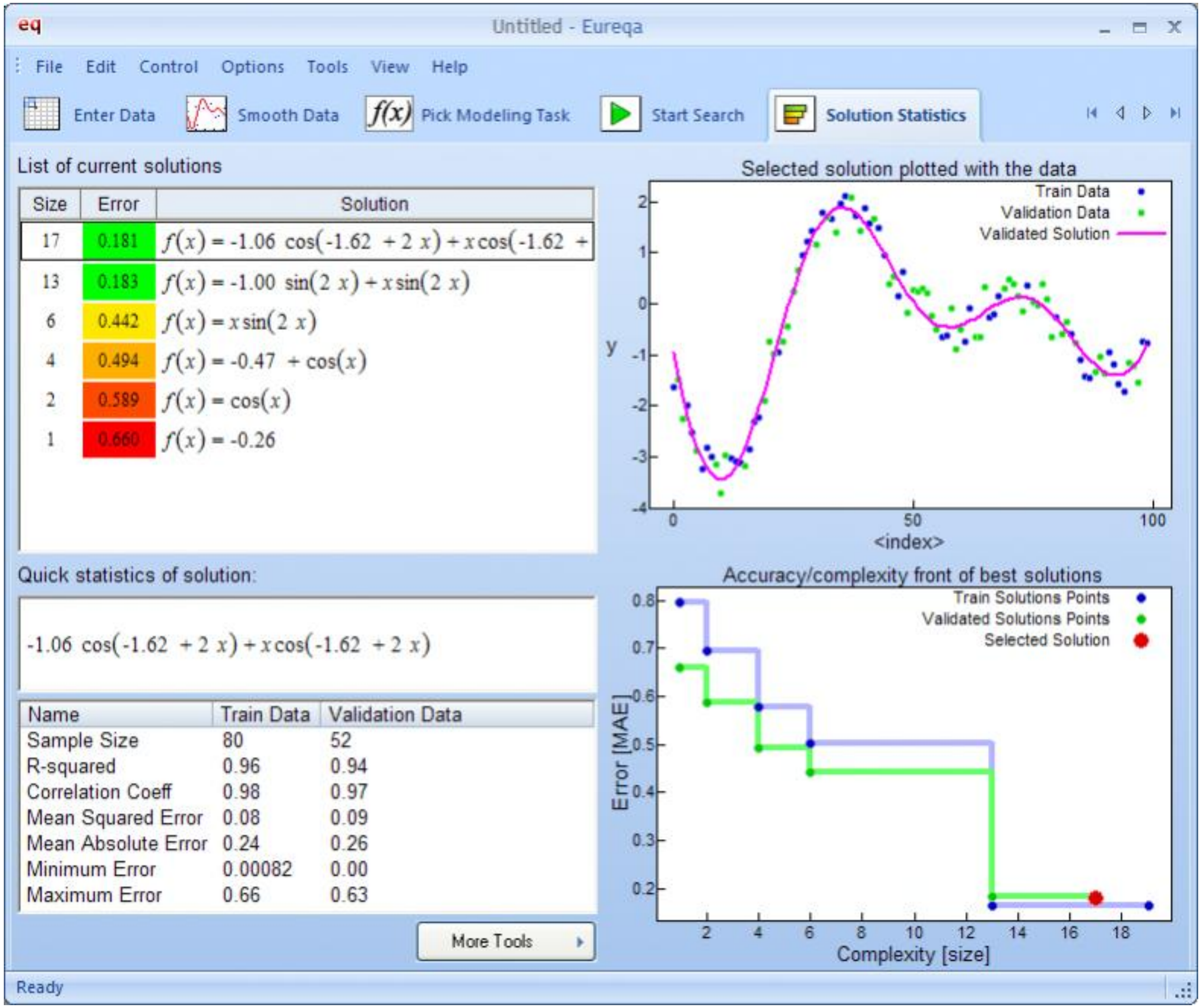
$$\frac{d\boldsymbol{K}}{dt} = a_K + \frac{b_K + c_K\boldsymbol{S}_{t-t_1}}{\boldsymbol{K}_{t-t_2}}$$

$$\frac{d\boldsymbol{S}}{dt} = a_S + \frac{b_S + c_S\boldsymbol{K}_{t-t_3}}{\boldsymbol{S}_{t-t_4}}$$



	A	B	C	D	E	F	G	H	I	J
desc	some variable	some other variable	confidence in y							
var	x	y	w							
1	-3.00	-1.62	1.00							
2	-2.94	-1.48	0.56							
3	-2.88	-2.25	0.81							
4	-2.82	-1.98	0.81							
5	-2.76	-2.51	0.59							
6	-2.70	-2.88	0.52							
7	-2.64	-3.22	0.65							
8	-2.58	-2.83	0.90							
9	-2.52	-3.01	0.82							
10	-2.46	-3.14	0.75							
11	-2.40	-3.71	0.83							
12	-2.34	-2.98	0.86							
13	-2.28	-3.03	0.71							
14	-2.22	-3.09	0.51							
15	-2.16	-3.12	0.70							
16	-2.10	-3.19	0.99							
17	-2.04	-2.86	0.91							
18	-1.98	-2.31	0.64							
19	-1.92	-2.23	0.85							
20	-1.86	-1.90	0.83							
21	-1.80	-0.75	0.99							
22	-1.74	-0.96	0.55							

Eureqa


Untitled - Eureqa
File Edit Control Options Tools View Help
Enter Data
Smooth Data
Pick Modeling Task
Start Search
Solution Statistics
List of current solutions
Size Error Solution
17 0.181 f(x) = -1.06 cos(-1.62 + 2 x) + x cos(-1.62 +
13 0.183 f(x) = -1.00 sin(2 x) + x sin(2 x)
6 0.442 f(x) = x sin(2 x)
4 0.494 f(x) = -0.47 + cos(x)
2 0.589 f(x) = cos(x)
1 0.660 f(x) = -0.26
Quick statistics of solution:
-1.06 cos(-1.62 + 2 x) + x cos(-1.62 + 2 x)
Name Train Data Validation Data
Sample Size 80 52
R-squared 0.96 0.94
Correlation Coeff 0.98 0.97
Mean Squared Error 0.08 0.09
Mean Absolute Error 0.24 0.26
Minimum Error 0.00082 0.00
Maximum Error 0.66 0.63
More Tools
Selected solution plotted with the data
Train Data
Validation Data
Validated Solution
y
<index>
Accuracy/complexity front of best solutions
Train Solutions Points
Validated Solutions Points
Selected Solution
Error [MAE]
Complexity [size]
Ready

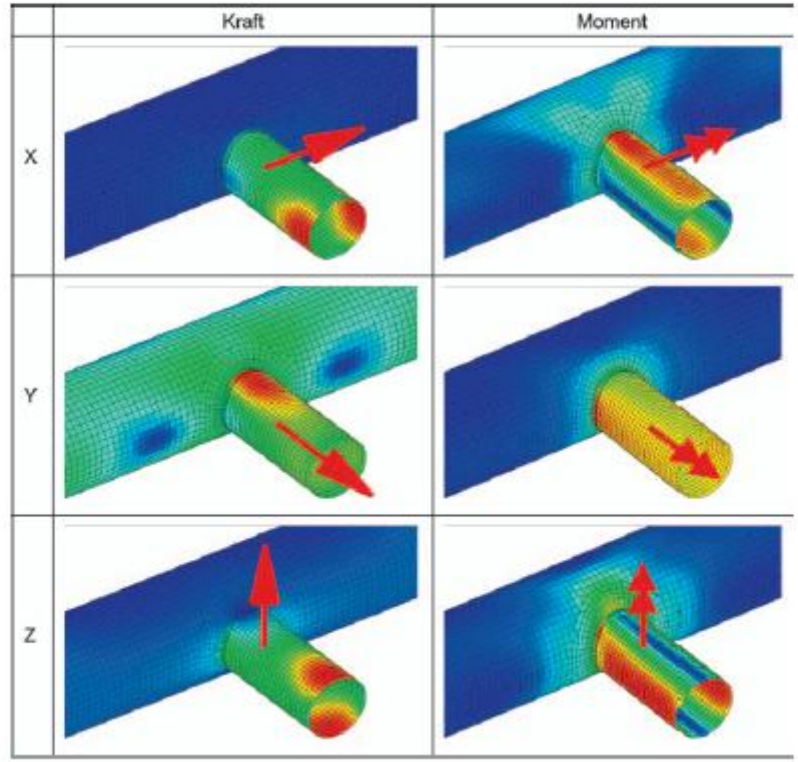
Eureqa

Run…

CITIZEN
SCIENCE


RADON
DETECTOR
R.P.I.I. Tel: 353 1 269 77 66
RADON
DETECTOR
R.P.I.I. Tel: 353 1 269 77 66

	Bauteil	Formel	n_1	n_2	n_3
C_W	Rohr	$C_W = n_1 (\beta + \gamma)^{n_2} \cdot \left(\tau + \frac{n_3 \tau \cdot \beta}{n_4 - \beta}\right)^{n_5}$	1,3034	0,9145	−0,168
	Nocken	$C_W = n_1 (\tau \cdot \gamma + n_2 \tau^3 \cdot \beta^3 \gamma + n_3 \tau \cdot \beta \cdot \gamma)^{n_4}$	2,149	−0,1871	−0,523
C_L	Rohr	$C_L = \tau (\gamma + n_1 \beta)^{n_2}$	1,729	0,5269	
	Nocken	$C_L = \exp((\tau \cdot \gamma)^{n_1} - \beta^{\beta})$	0,237		
C_N	Rohr	$C_N = n_1 \tau \beta \gamma + n_2 \tau \beta^2 \gamma$	2,233	−1,789	
	Nocken	$C_N = \beta + n_1 \beta \gamma + n_2 \tau \beta \gamma - \tau \beta^2 \gamma + n_3 \beta^2 \gamma$	0,486	1,2526	−0,388
C_T	Rohr	$C_T = n_1 t + n_2 b^4$	0,903	0,268	
	Nocken	$C_T = n_1 \exp\left(\frac{n_2 \tau + n_2 \beta}{1 + n_4 \gamma}\right)$	0,910	0,2223	0,4737

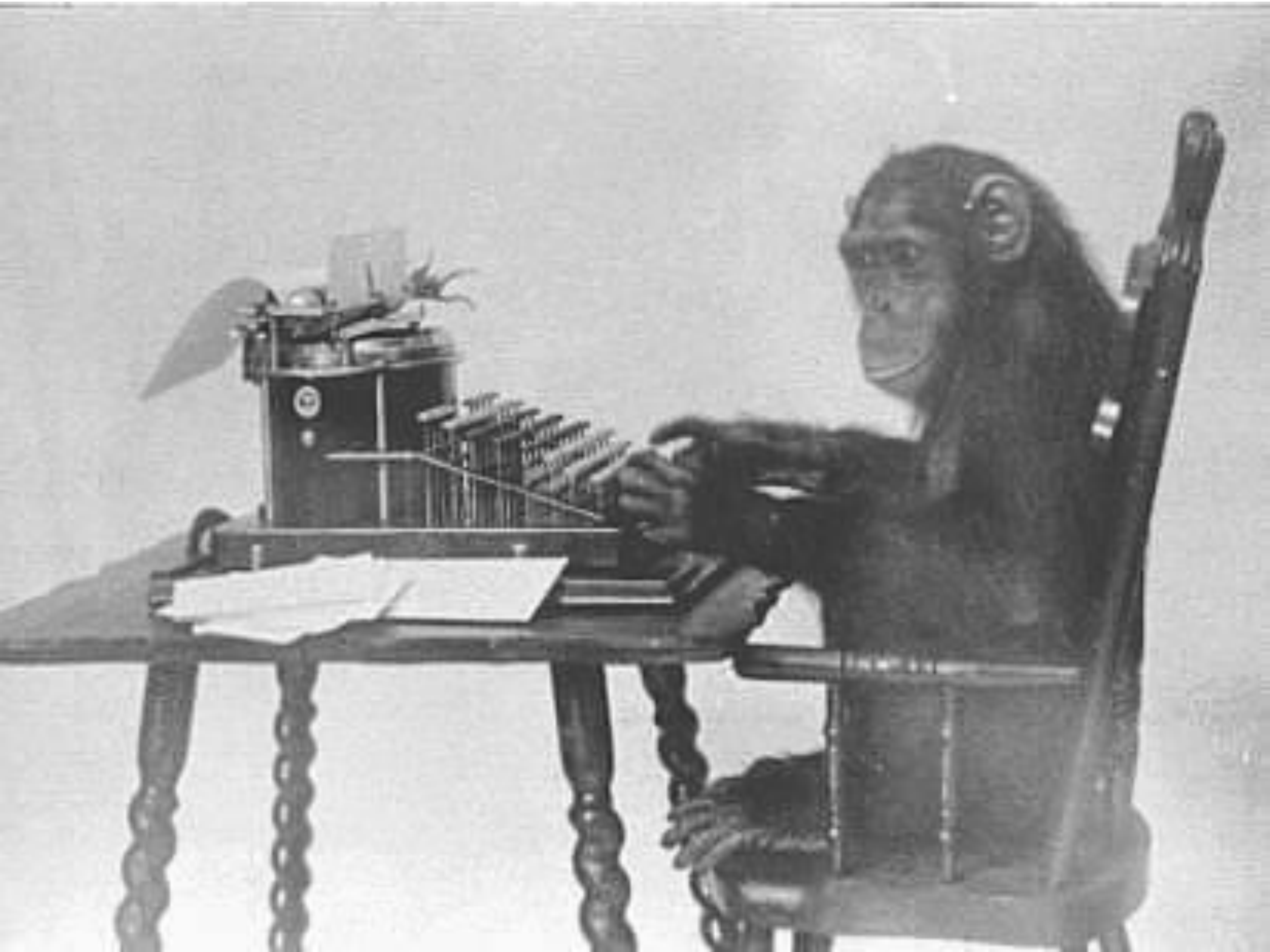


	Bauteil	Formel	n_1	n_2	n_3		
B_W	Rohr	$B_W = \exp(n_1 \beta / \gamma + n_2) \cdot \tau^{n_3} \cdot \gamma^{n_4}$	3,5363	−1,4796	0,9551		
	Nocken	$B_W = n_1 (\gamma - \tau)^{(n_2 \tau^{n_3})}$	0,6354	0,5072	0,26126		
B_L	Rohr	$B_L = \tau \cdot \exp((n_1/\gamma)^{n_2} + n_1 \cdot (\beta \cdot \gamma)^{n_4})$	1,6629	−0,07132	−2,645	−0,7284	
	Nocken	$B_L = n_1 \cdot \exp(n_2 \tau \gamma + n_3 \tau \beta^{n_4} + n_5 \beta)$	1,0141	0,02131	−0,1538	−0,7421	−0,04504
B_N	Rohr	$B_N = n_1 \cdot \exp(n_2 \beta) \cdot (n_3 \tau \gamma)^{n_4 \exp(n_5/\beta)}$	5,2163	0,16729	0,01469	0,7907	0,10408
	Nocken	$B_N = n_1 \exp\left(\frac{n_2 \cdot \tau \cdot \beta + n_3 \cdot \tau / \gamma}{(\beta \cdot \tau + (n_4 / \gamma)^{n_5})}\right)$	1,0523	1,7175	−7,1793	7,8903	1,4792
B_T	Rohr	$B_T = n_1 \cdot \tau \cdot \exp\left(\frac{n_2}{\gamma \cdot \beta^{\tau} \cdot (\tau \cdot \beta^2 - 1) - \beta \cdot \gamma \cdot (1 + n_3 \tau \cdot \beta)}\right)$	1,0646	2,008	0,02014		
	Nocken	$B_T = n_1 \left(\frac{\beta \cdot \gamma^{n_2 \beta}}{\exp(n_3 \tau)}\right)^{(1/(\gamma + n_4 \beta/\tau))}$	0,9369	2,594	2,0855	2,4033	

hungsfaktoren für die Rundnocken in der EN 13480-3 Kapitel11 für Membranspannungen

Table 3: Proposal for new stress intensification factor formulars for circular attachments in the EN 13480-3 clause 11 for membrane stresses.

Von Thomas Hermanowski, Dr. Andreas Rick, Dr. Jochen Weber



Scalability

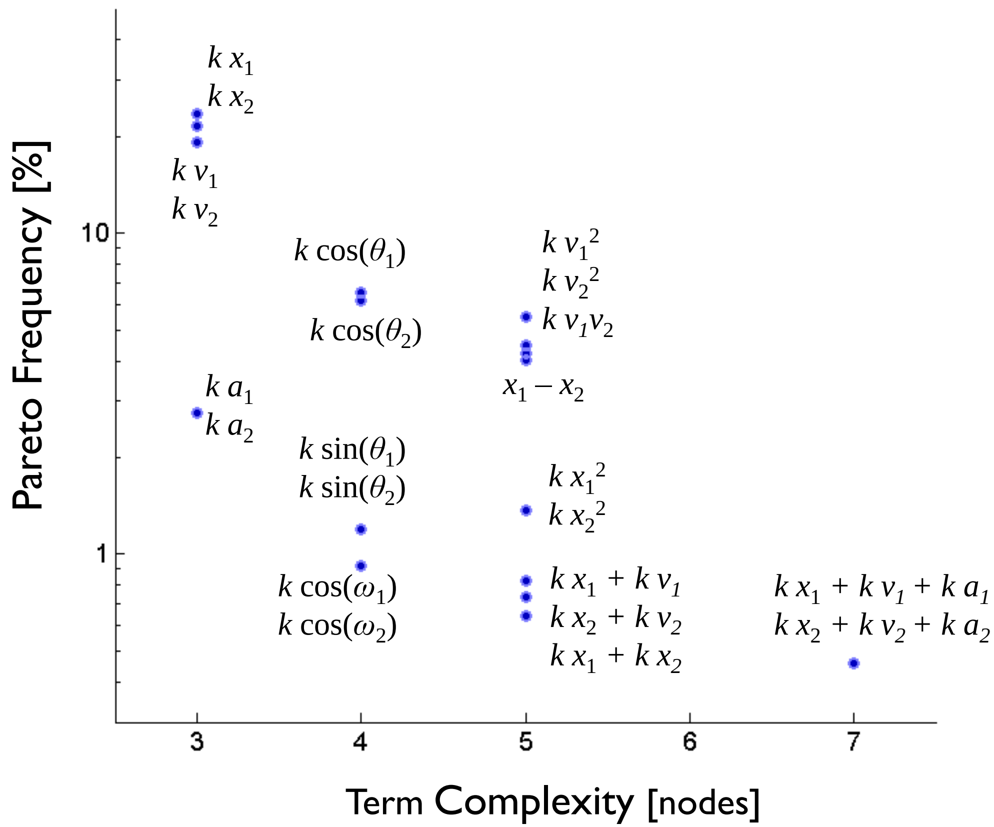
- Complexity
- Noise
- Hidden (unobservable) variables
- Justification

Approximations

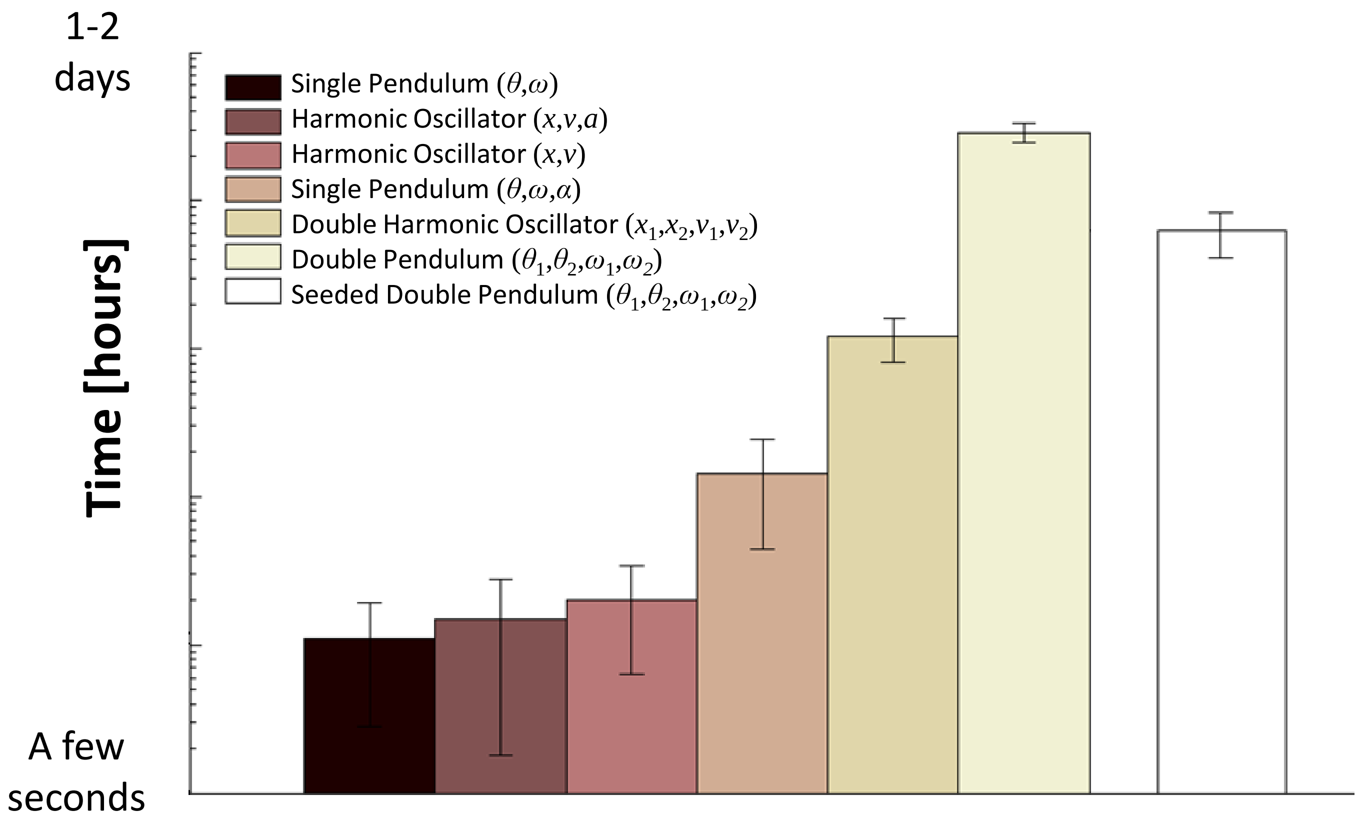
Building Blocks	Detected Pendulum Law	Approximation
*, +, –, cos(), sin()	$\omega^2 - 19.6 \cdot \cos(\theta)$	*Exact Solution*
*, +, –, sin()	$\omega^2 - 19.5999 \cdot \sin(-1.57079 + \theta)$	*Trigonometric identity*
*, +, –	$\omega^2 + 9.7108 \cdot \theta^2 - 0.7042 \cdot \theta^4$	*Taylor series expansion (4^{th} order)*

Alphabet

Term	Complexity	Frequency	Systems
$k\,x_2$	1	23.578	4
$k\,x_1$	1	21.6514	4
$k\,v_2$	1	21.5596	4
$k\,v_1$	1	19.1743	4
$k\cos(\theta_2)$	2	6.51376	2
$k\cos(\theta_1)$	2	6.14679	2
$k\,v_2^2$	3	5.50459	4
$k\,v_1^2$	3	4.49541	4
$(x_1 - x_2)$	3	4.22018	2
$k\,v_1 v_2$	3	4.0367	2
$k\,a_2$	1	2.75229	2
$k\,a_1$	1	2.75229	2
$k\,x_2^2$	3	1.37615	2
$k\sin(\theta_2)$	2	1.19266	2
$k\sin(\theta_1)$	2	1.19266	2
$k\cos(\theta_1)$	2	0.917431	2
$k\cos(\theta_2)$	2	0.917431	2
$k\,x_1^2$	3	0.825688	2
$k\,x_1^2$	3	0.825688	2
$k\,x_1 + k\,v_1$	3	0.733945	3
$k\,x_2 + k\,v_2$	3	0.733945	3
$k\,x_1 + k\,x_2$	3	0.642202	2
$k\,x_1 + k\,v_1 - k\,a_1$	5	0.458716	2
$k\,x_2 + k\,v_2 - k\,a_2$	5	0.458716	2

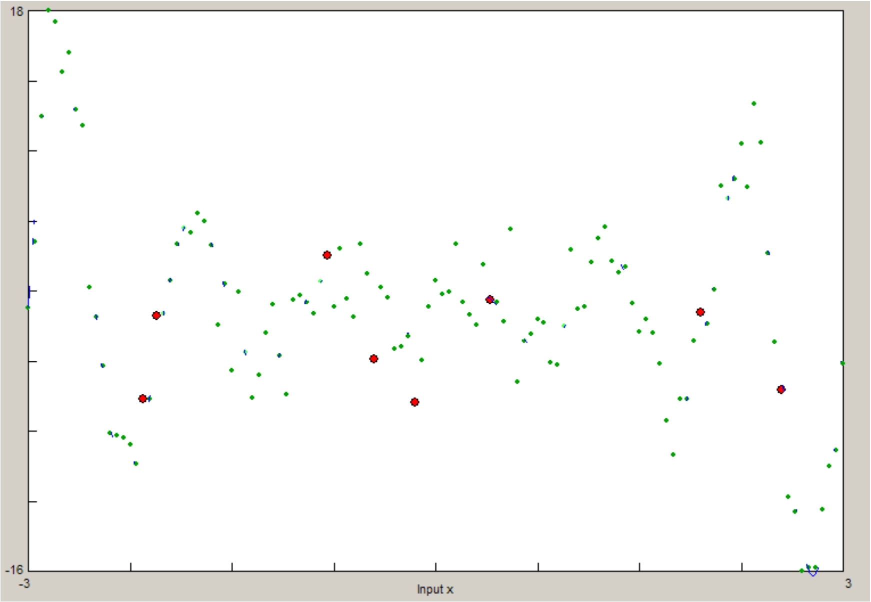


Time to Regress

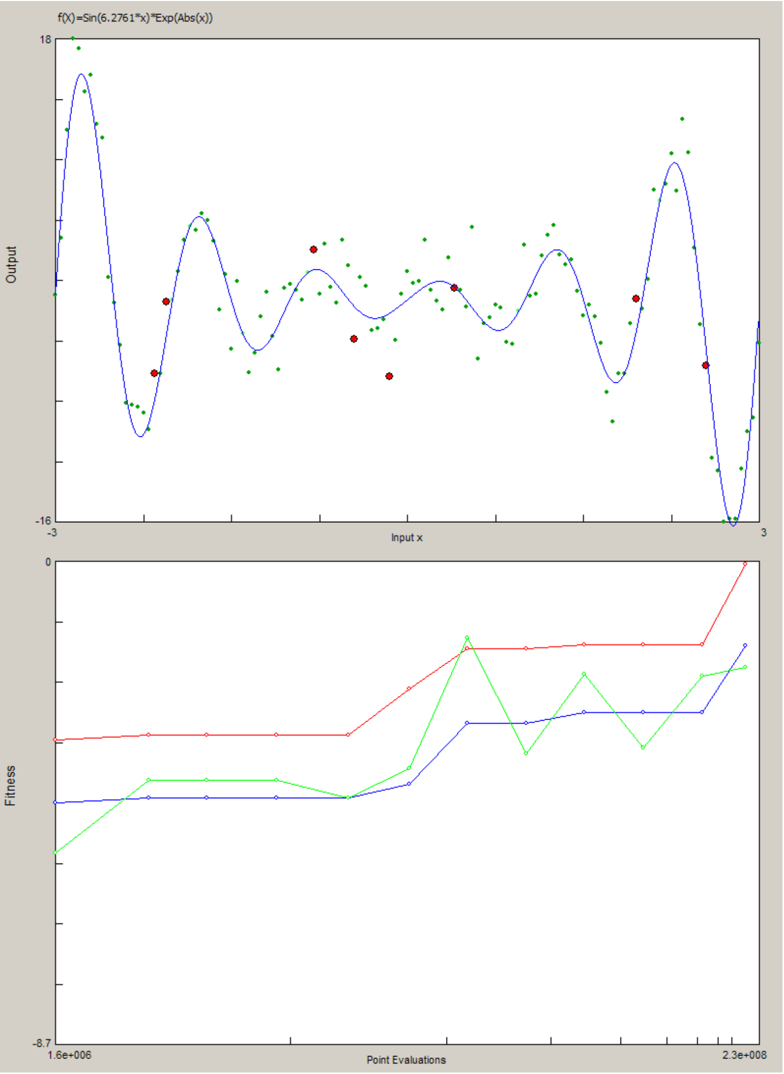


STOCHASTIC MODELS

10% Noise

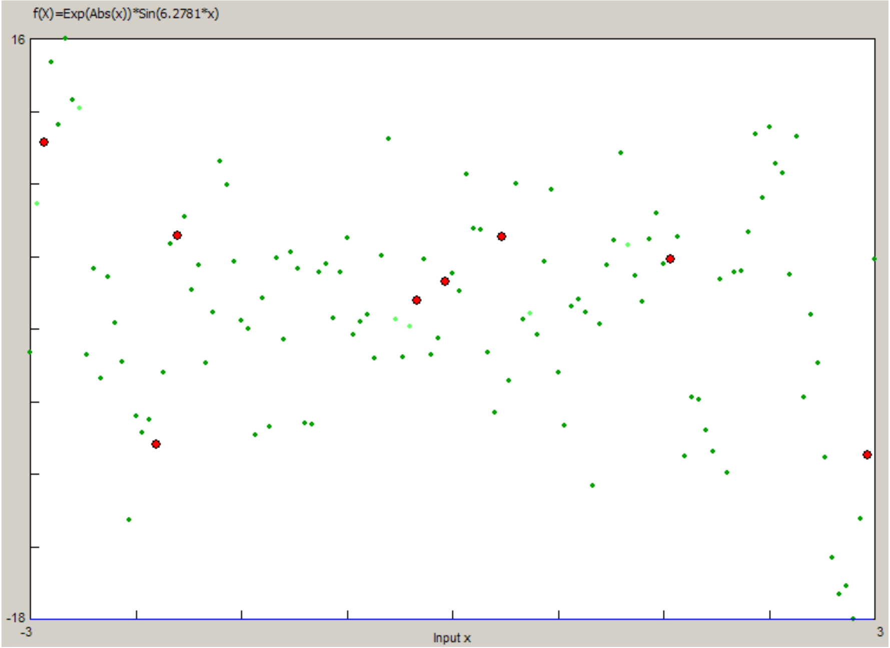


18
-16
-3
3
Input x

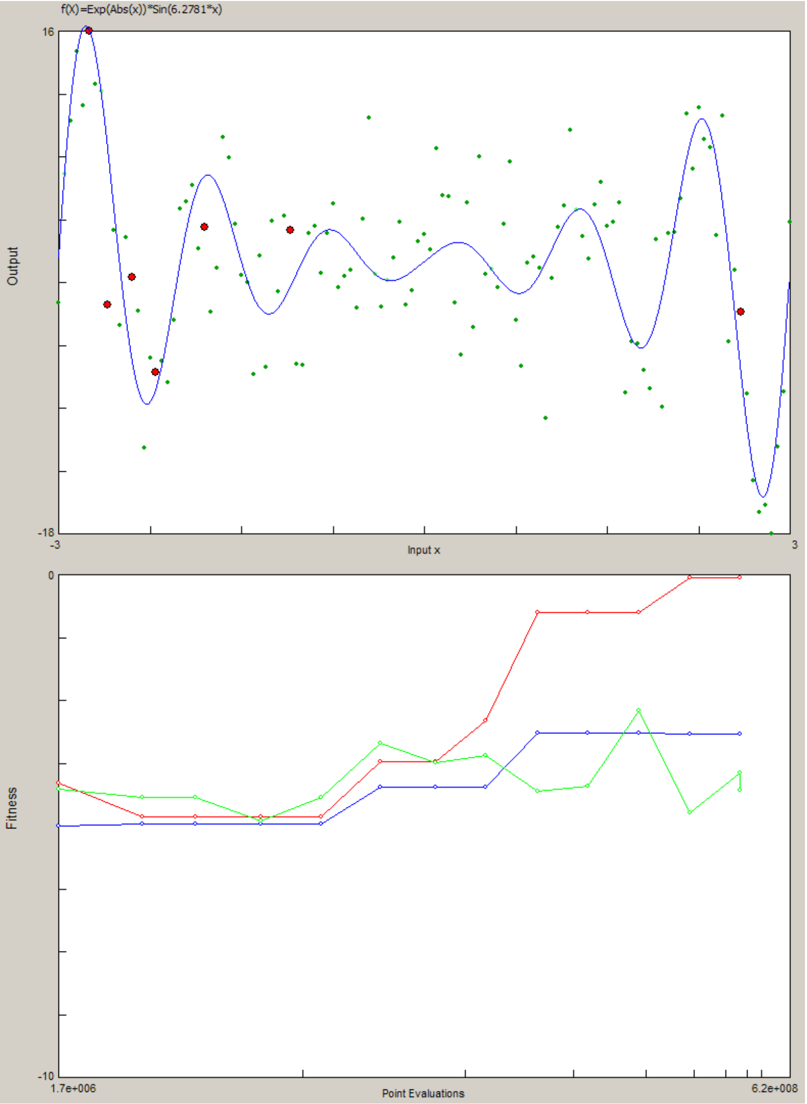


f(X)=Sin(6.2761*x)*Exp(Abs(x))
18
-16
Output
-3
3
Input x
0
-8.7
Fitness
1.6e+006
2.3e+008
Point Evaluations

30% Noise

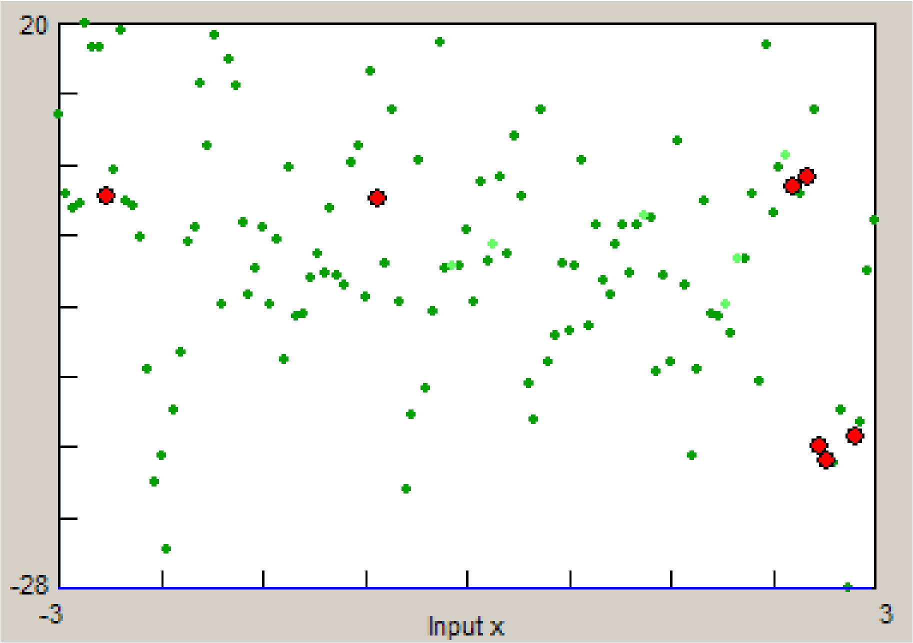


f(X)=Exp(Abs(x))*Sin(6.2781*x)
16
-18
-3
3
Input x

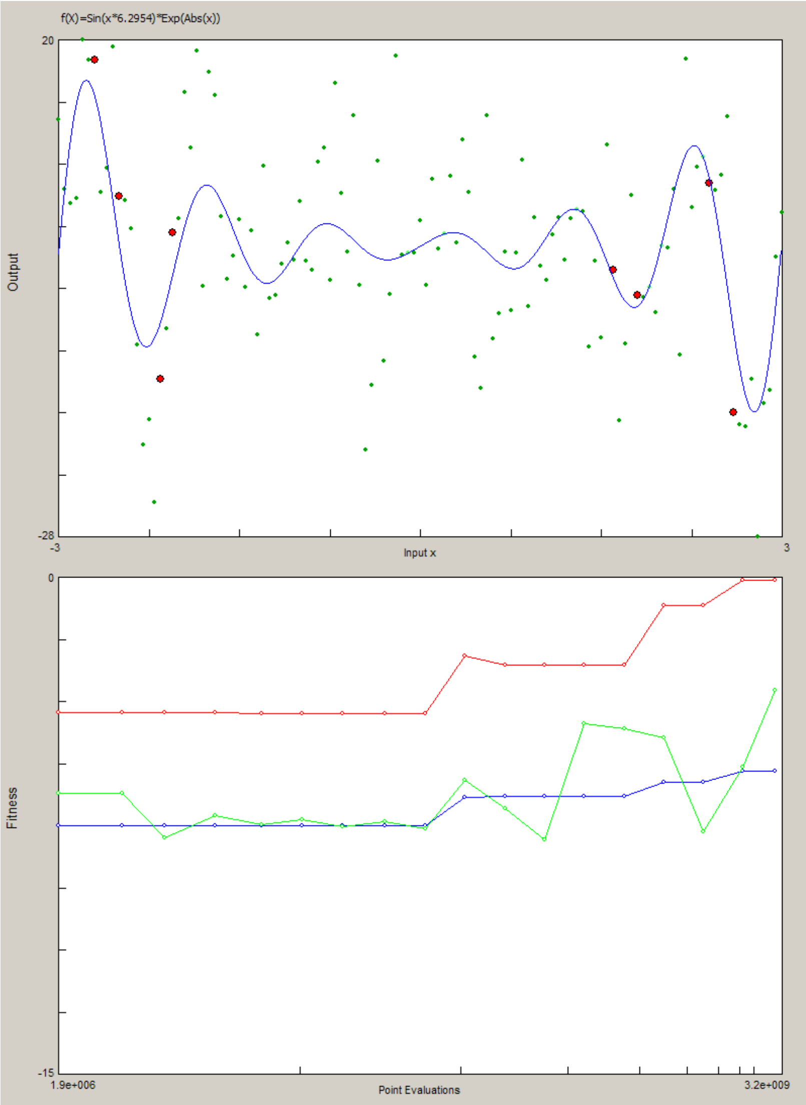


f(X)=Exp(Abs(x))*Sin(6.2781*x)
16
-18
-3
3
Output
Input x
0
-10
Fitness
1.7e+006
6.2e+008
Point Evaluations

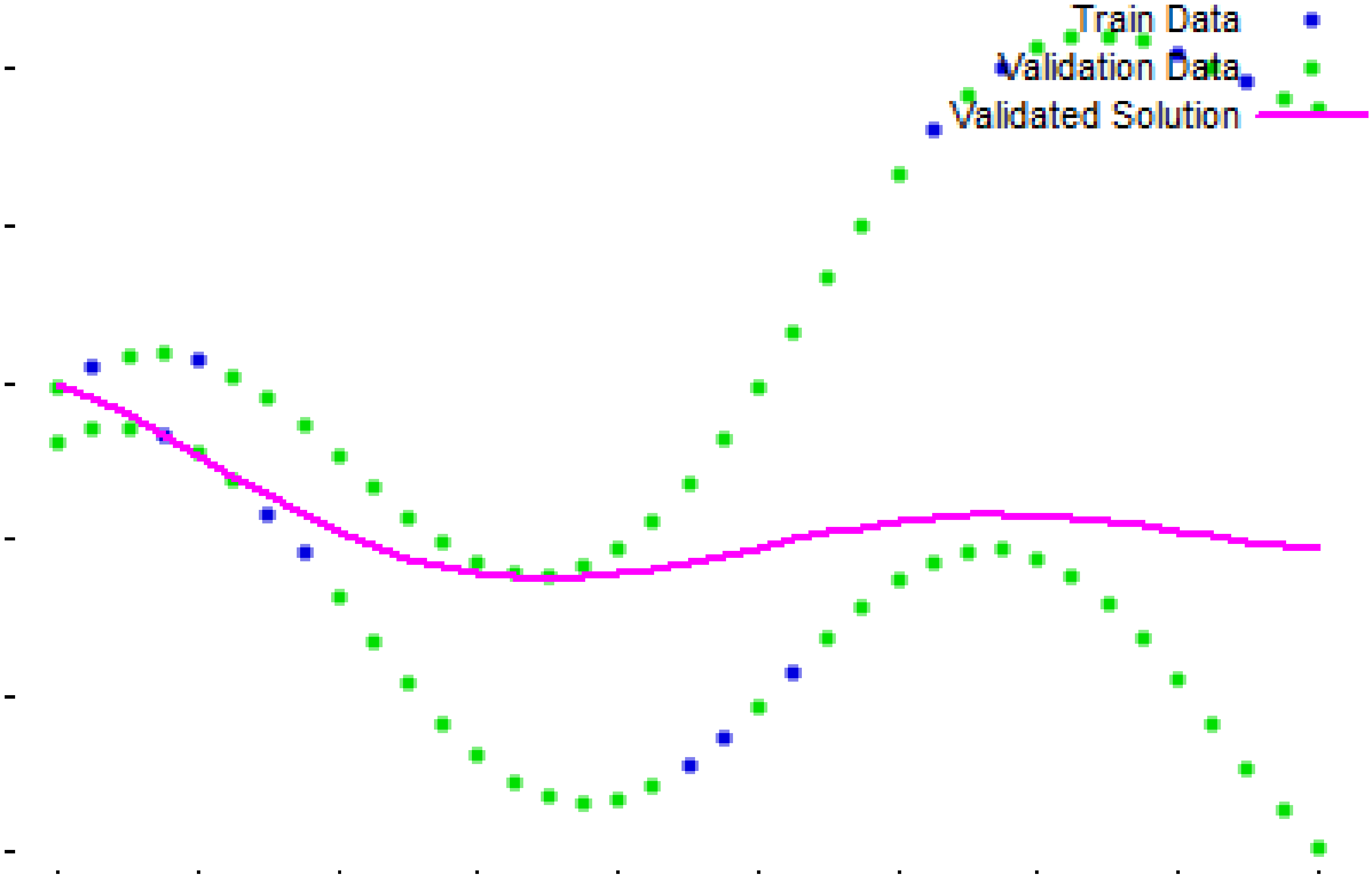
70% Noise



20
-28
-3
Input x
3



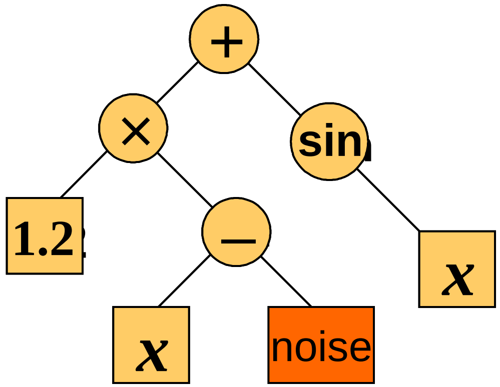
f(X)=Sin(x*6.2954)*Exp(Abs(x))
20
Output
-28
-3
Input x
3
0
Fitness
-15
1.9e+006
Point Evaluations
3.2e+009

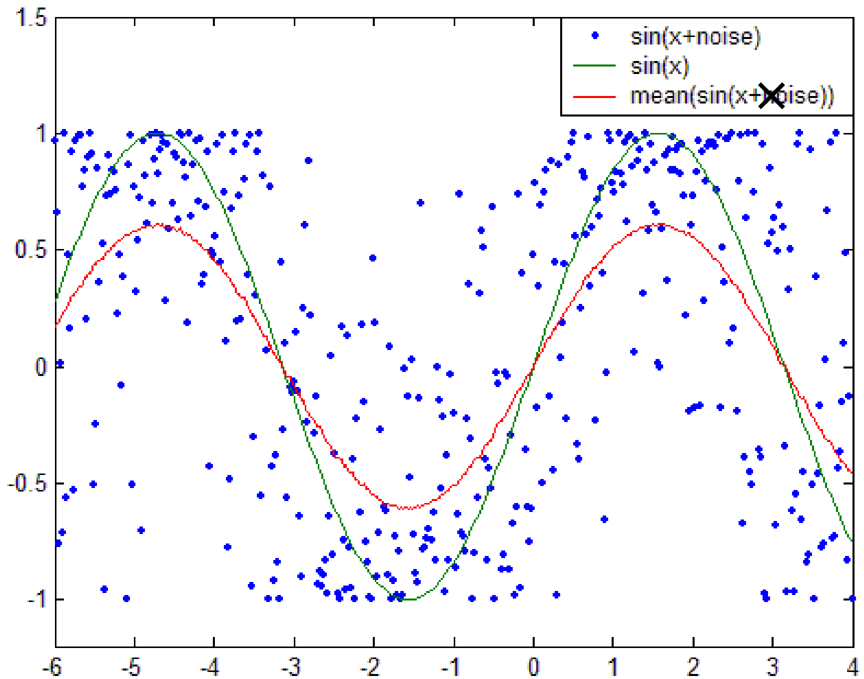


Train Data
Validation Data
Validated Solution

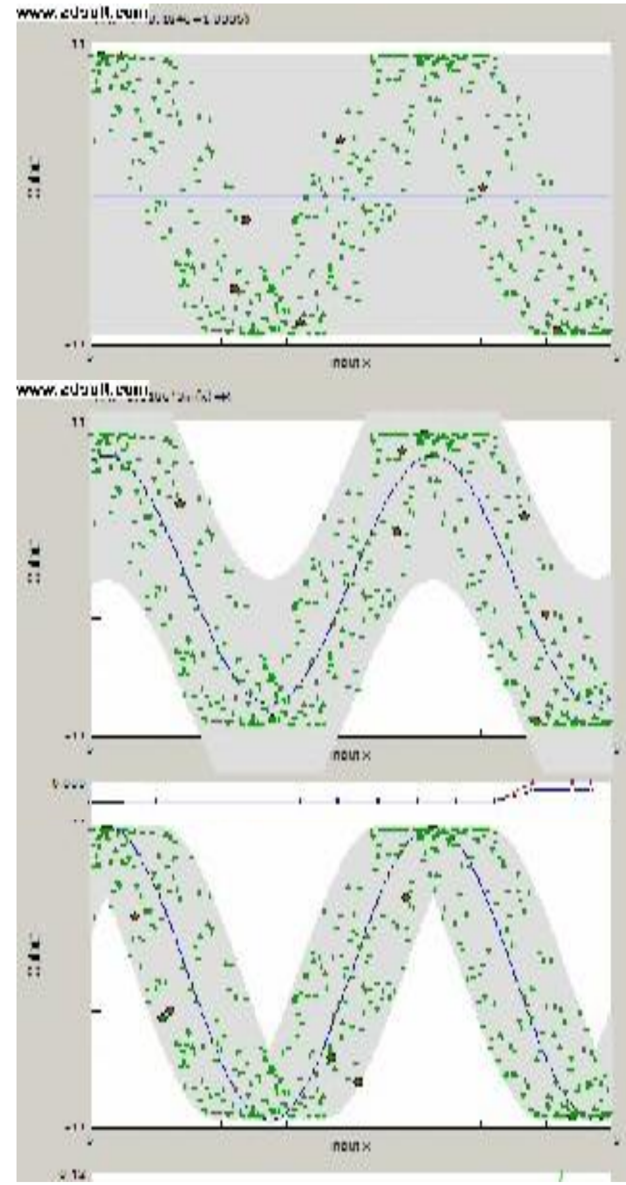
Regressing Stochastic elements

Add a "noise" building block

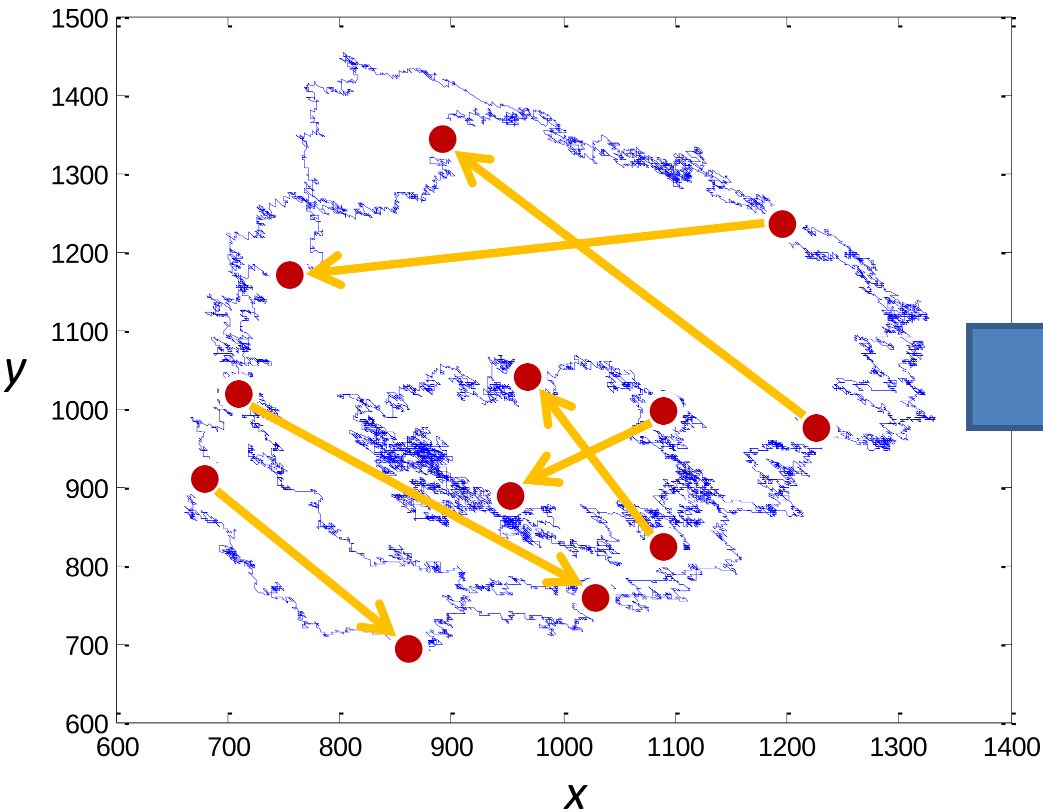




$f(x) = \sin\,(x + \boldsymbol{noise})$



Periodic Samples of a Stochastic System



Maximum Likelihood Stochastic Model

$$x \xrightarrow{10} 2x$$

$$x + y \xrightarrow{0.1} 2y$$

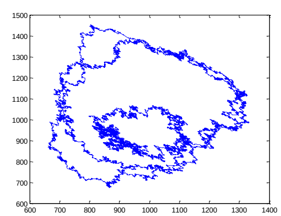
$$y \xrightarrow{10} \emptyset$$

Simulation…

Likelihood Fitness

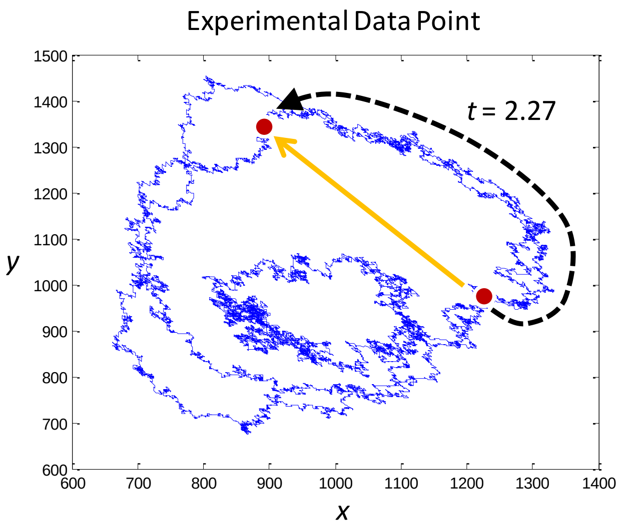
Find a model that maximizes the probability of seeing this data

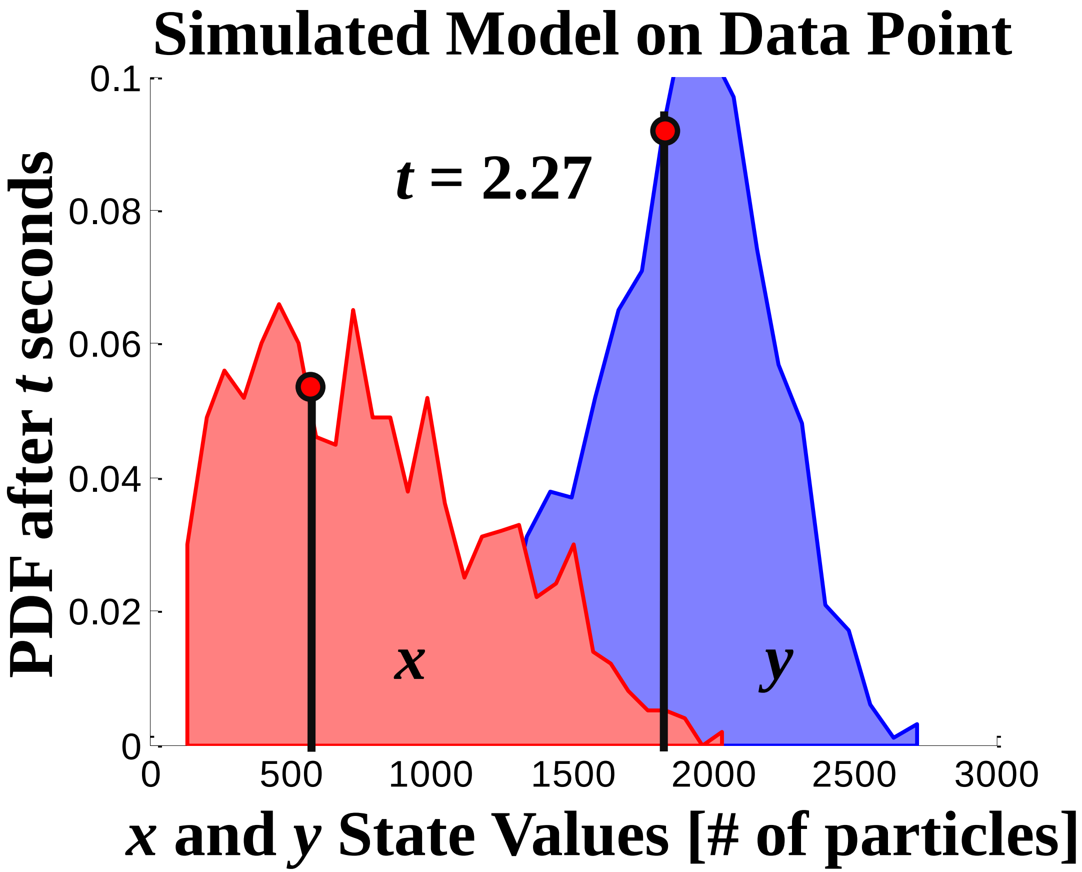
P[

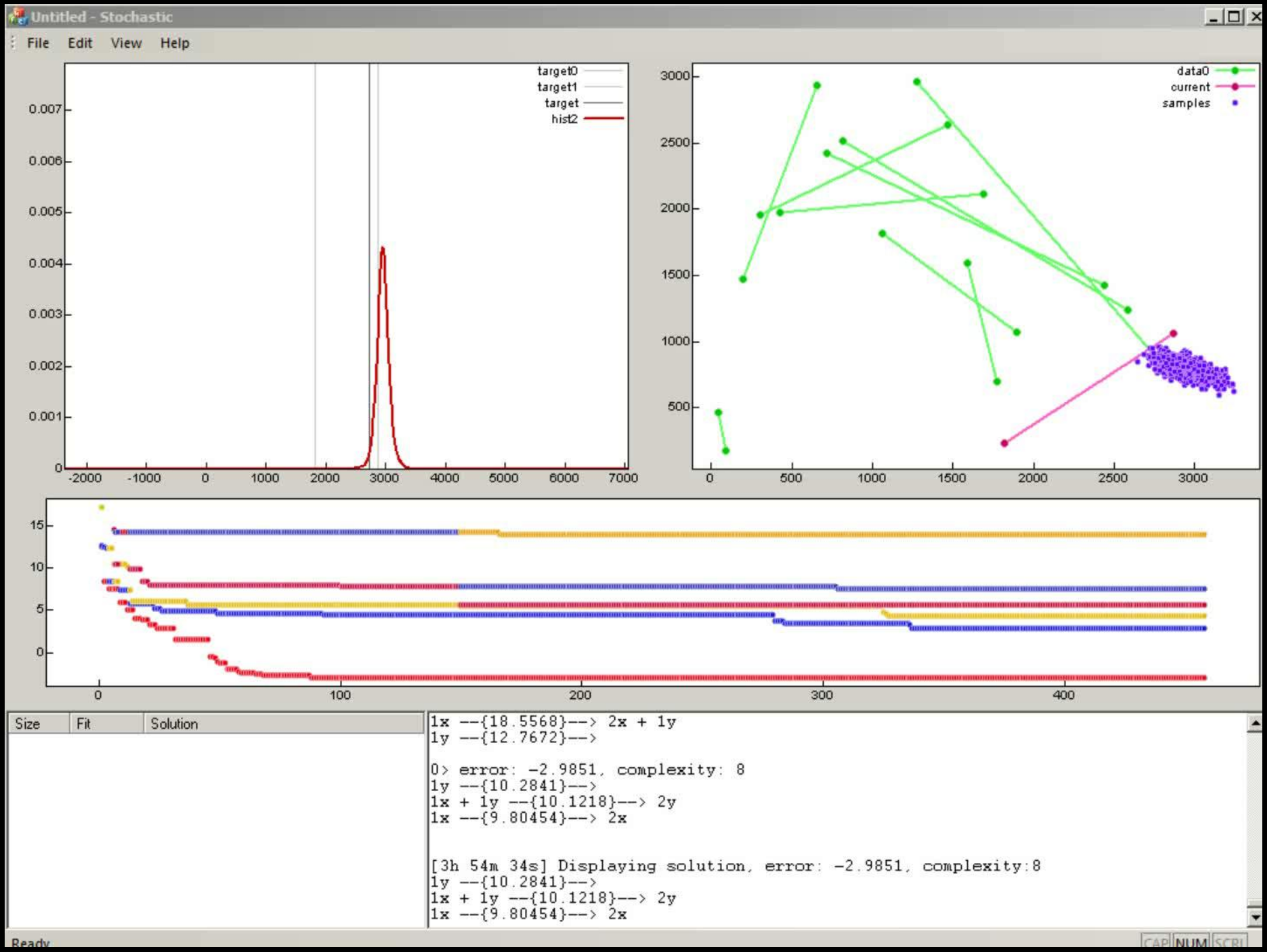


$$\Big| \; m = \left\{ \begin{array}{rcl} y_1 & \xrightarrow{c_1} & 2\,y_1 \\ y_1 + y_2 & \xrightarrow{c_2} & 2\,y_2 \\ y_2 & \xrightarrow{c_3} & \varnothing \end{array} \right. \Bigg]$$

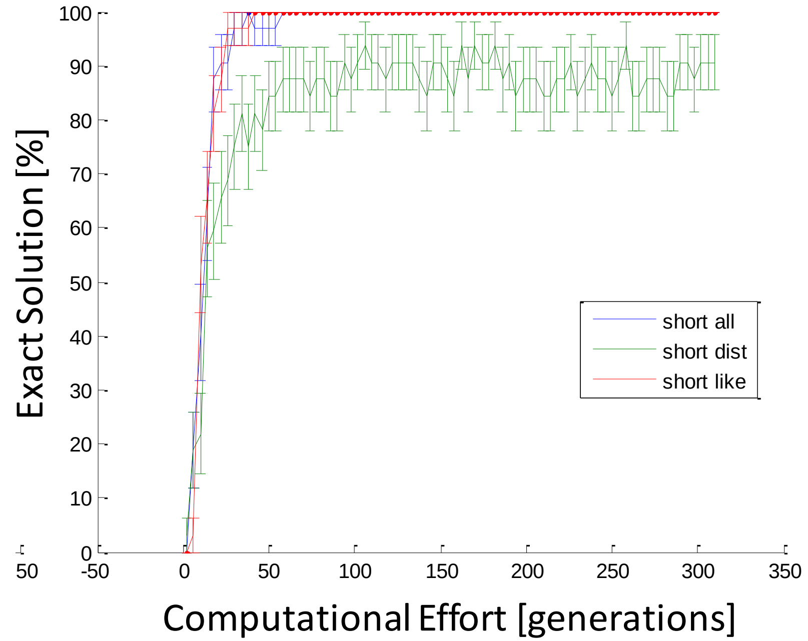
Sample the Timespan



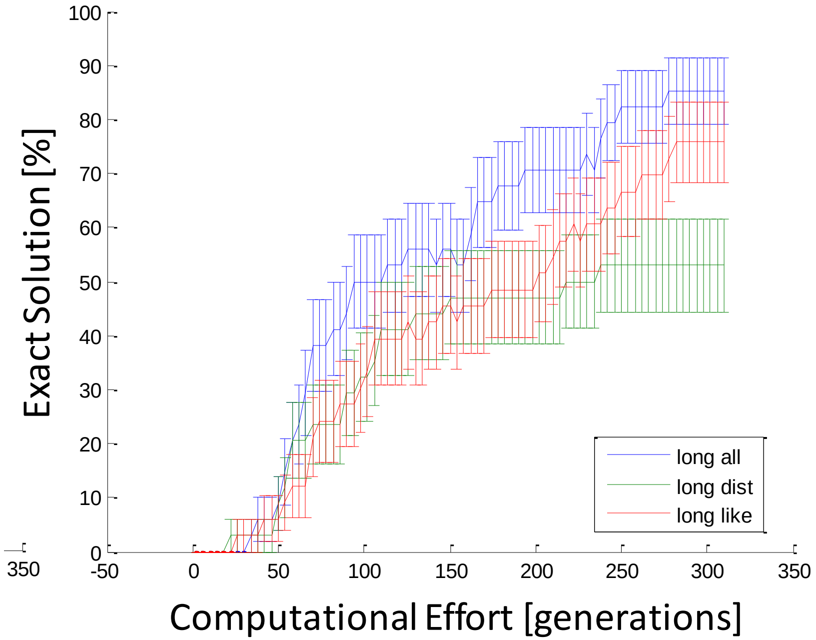



Untitled - Stochastic
File Edit View Help
target0
target1
target
hist2
data0
current
samples
Size
Fit
Solution
1x --{18.5568}--> 2x + 1y
1y --{12.7672}-->
0> error: -2.9851, complexity: 8
1y --{10.2841}-->
1x + 1y --{10.1218}--> 2y
1x --{9.80454}--> 2x
[3h 54m 34s] Displaying solution, error: -2.9851, complexity:8
1y --{10.2841}-->
1x + 1y --{10.1218}--> 2y
1x --{9.80454}--> 2x
Ready
CAP NUM SCRL

Short Time Gaps in Experimental Data:



Long Time Gaps in Experimental Data:



Concluding Remarks

Wired 16.07

WIRED

Chris Anderson

“Correlation is enough. Faced with massive data, [the Scientific Method] is becoming obsolete. We can stop looking for models.”

The data deluge accelerates our ability to hypothesize, model, and test.

The New York Times

Theoretical physicists are not yet obsolete, but scientists have taken steps toward replacing themselves

The end of insight

I am worried that we have enjoyed a brief window in human history where we could actually understand things, but that period may be coming to an end.

-- Steve Strogatz

NSF

BRIDGING THE GAP
1958 DARPA 2008
50 YEARS
POWERED BY IDEAS

NATIONAL INSTITUTES OF HEALTH

NASA

DEFENSE THREAT REDUCTION AGENCY
UNITED STATES OF AMERICA

$$k_1\omega_1^2 + k_2\omega_2^2 + k_3\omega_1\omega_2\cos\left(\theta_1-\theta_2\right) - k_4\cos\theta_1 - k_5\cos\theta_2$$

m_1, m_2, L_1, L_2

m_1, m_2, L_1, L_2

m_1, m_2, L_1, L_2

m_1, m_2, L_1, L_2

m_1, m_2, L_1, L_2

m_1, m_2, L_1, L_2

$k_1/k_1 = 1$

$k_2/k_1 = m_2L_2{}^2/(m_1L_1{}^2 + m_2L_1{}^2)$

$k_3/k_1 = 2.00055m_2L_2/(m_1L_1 + m_2L_1)$

$k_4/k_1 = 19.6/L_1$

$k_5/k_1 = 19.6\cdot m_2L_2/(m_2L_1{}^2 + m_1L_1{}^2)$

$$L_1^2(m_1+m_2)\omega_1^2 + m_2L_2^2\omega_2^2 + 2\cdot m_2L_1L_2\omega_1\omega_2\cos\left(\theta_1-\theta_2\right)$$
$$-19.6\cdot L_1(m_1+m_2)\cos\theta_1 - 19.6m_2L_2\cos\theta_2$$