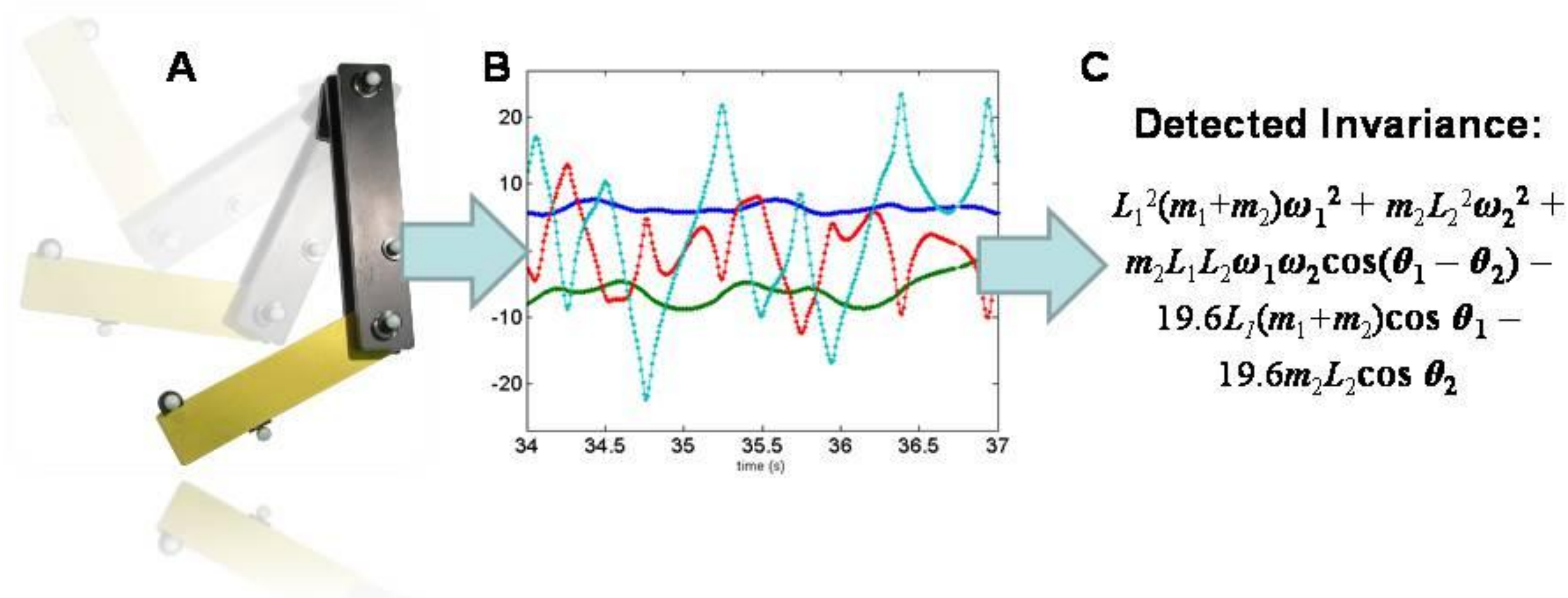
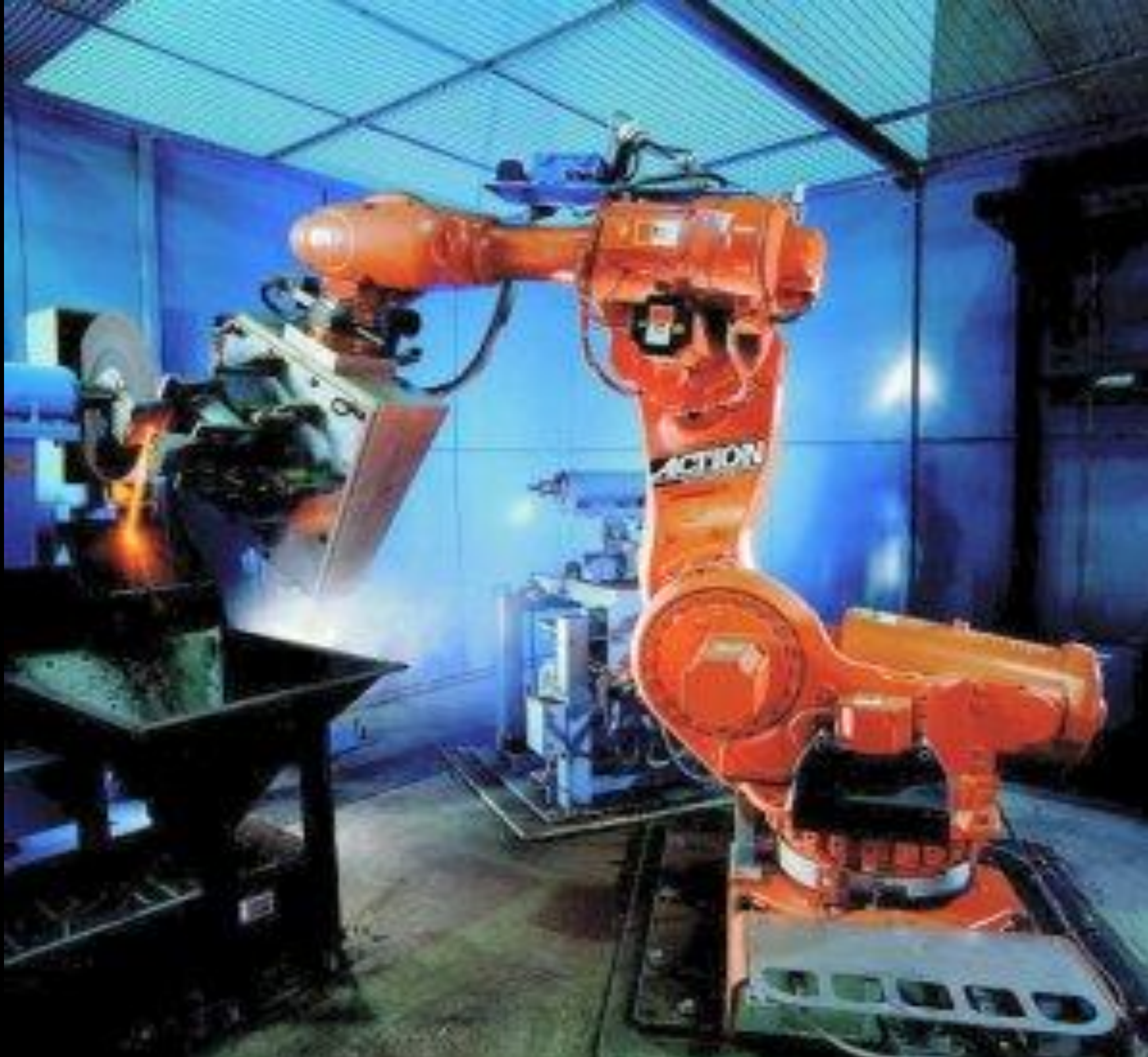
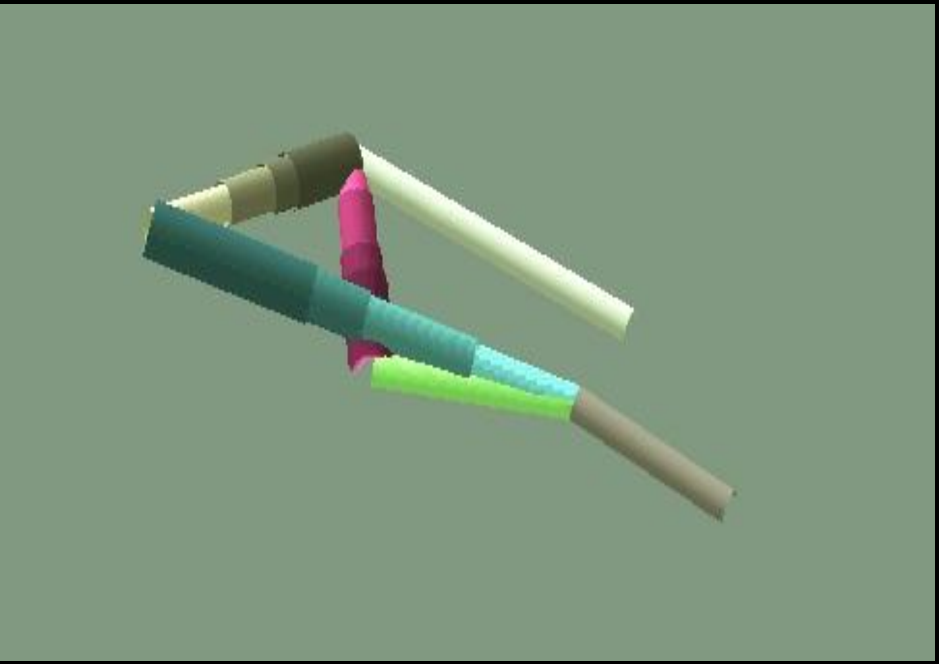
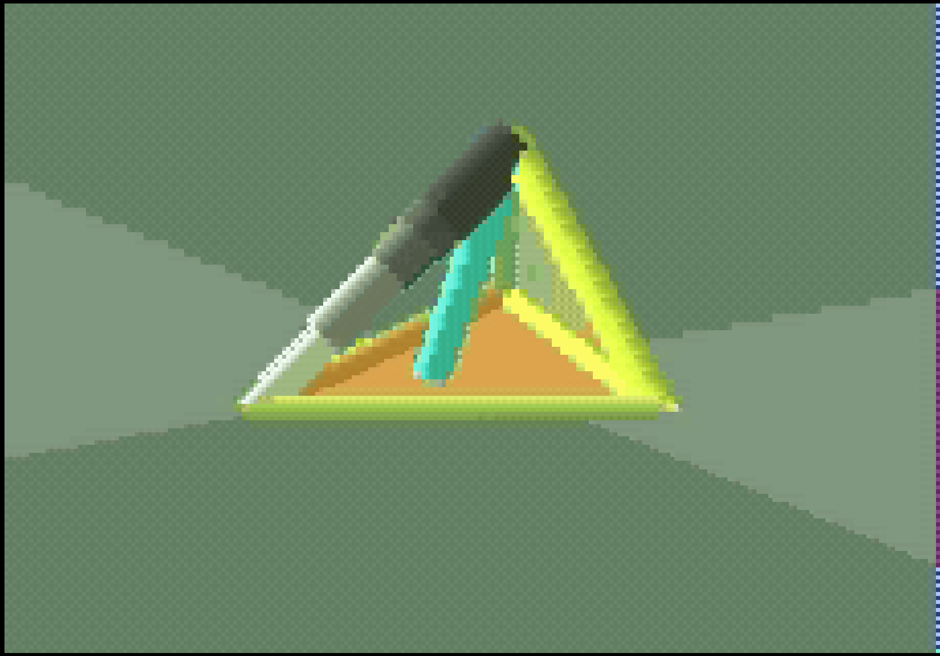
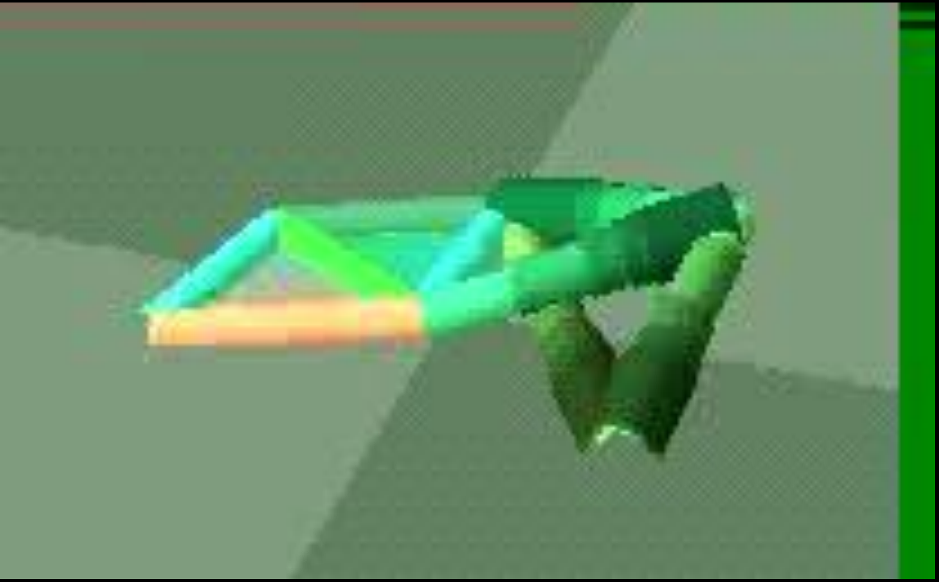
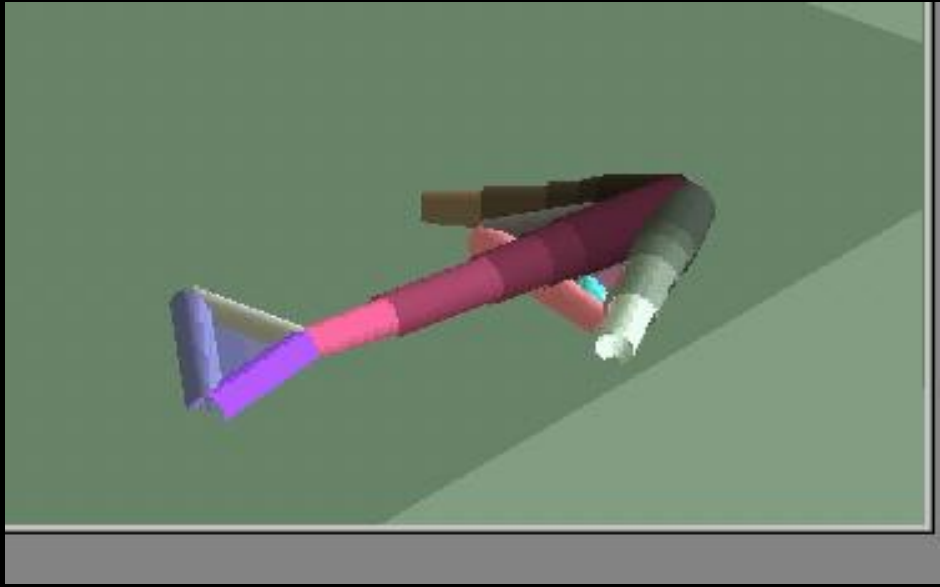
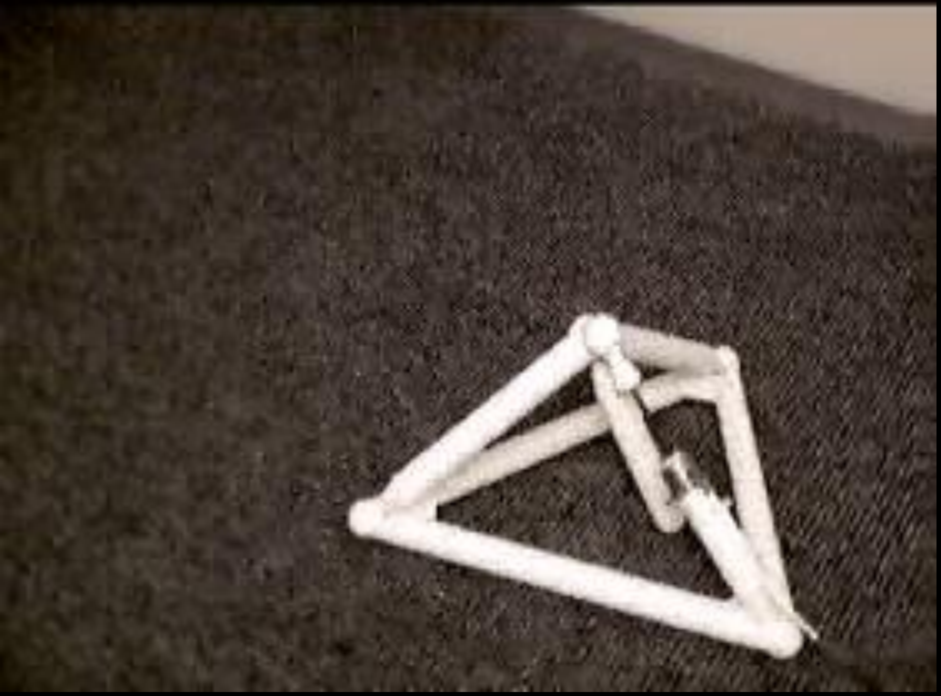
Mining Experimental Data for Dynamical Invariants

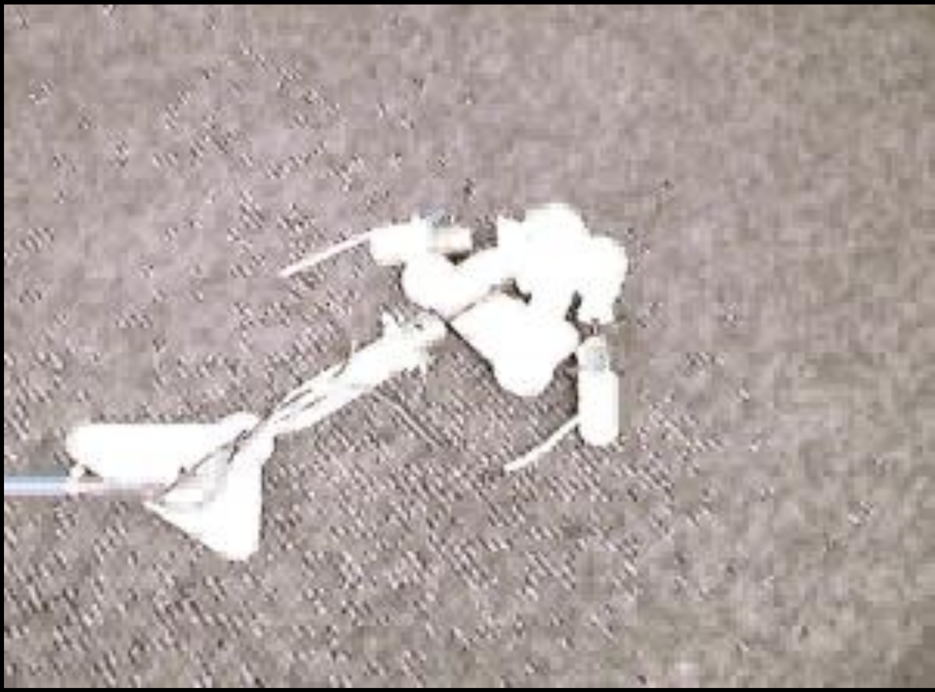


Hod Lipson, Cornell University

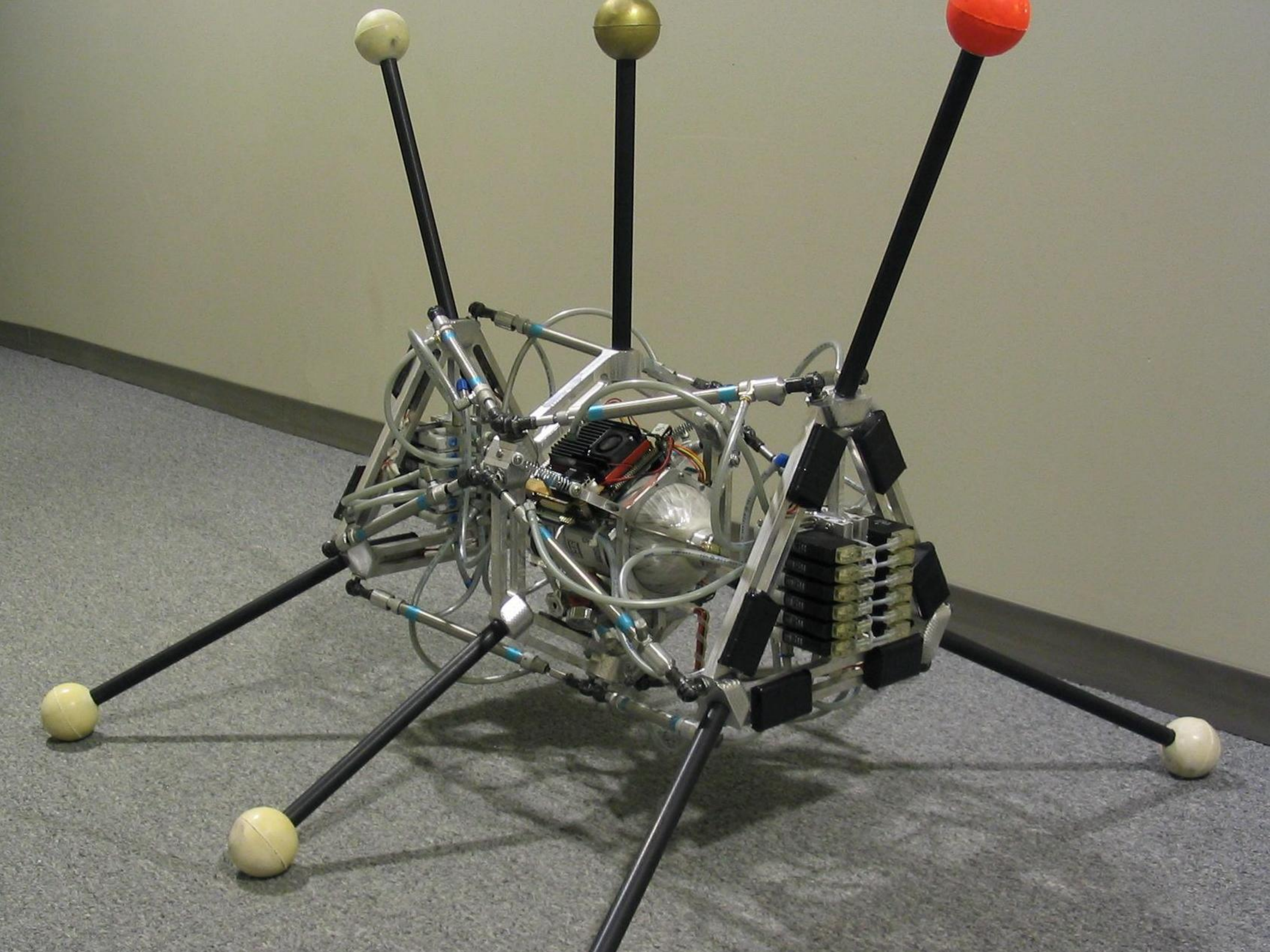


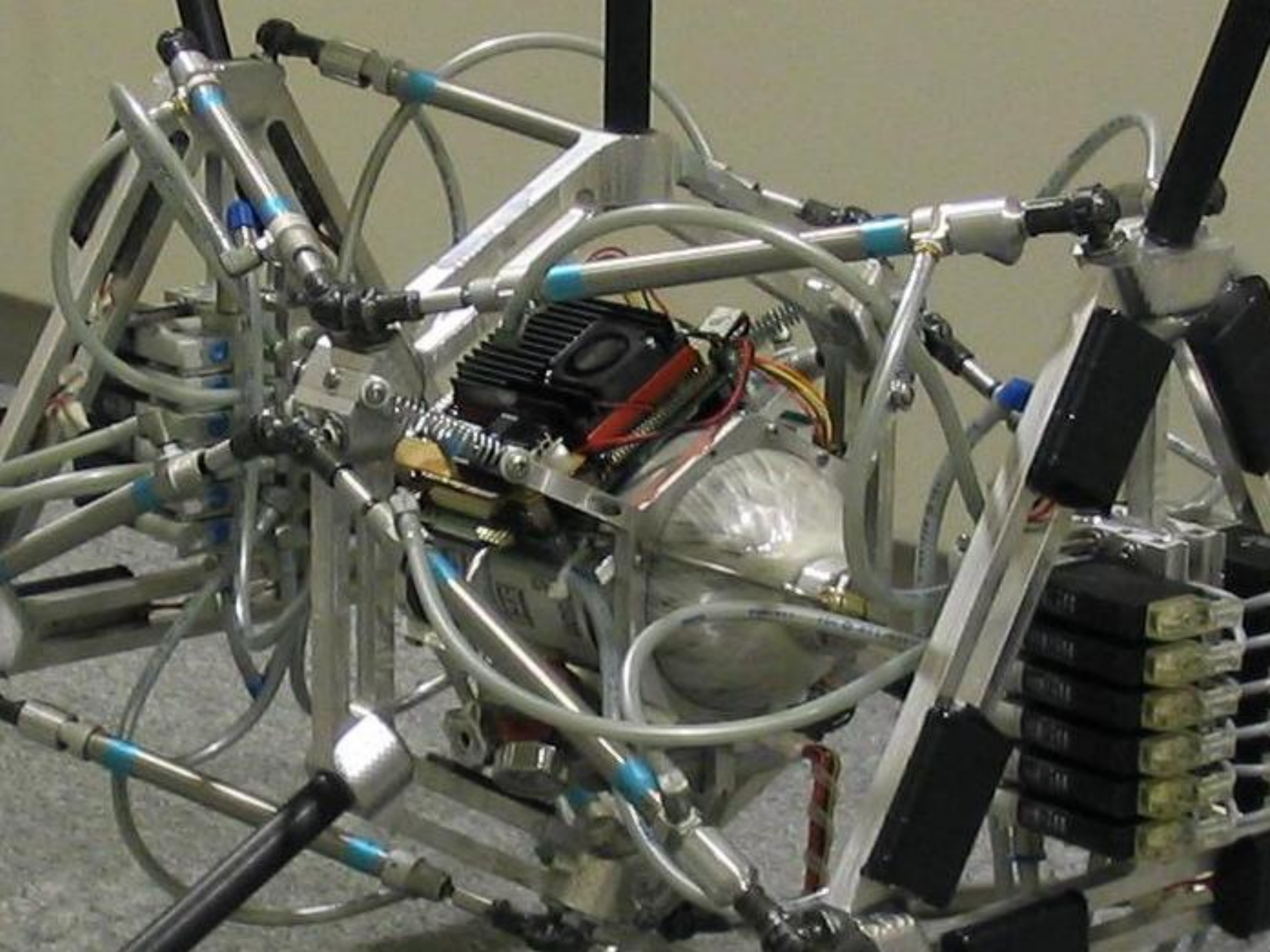





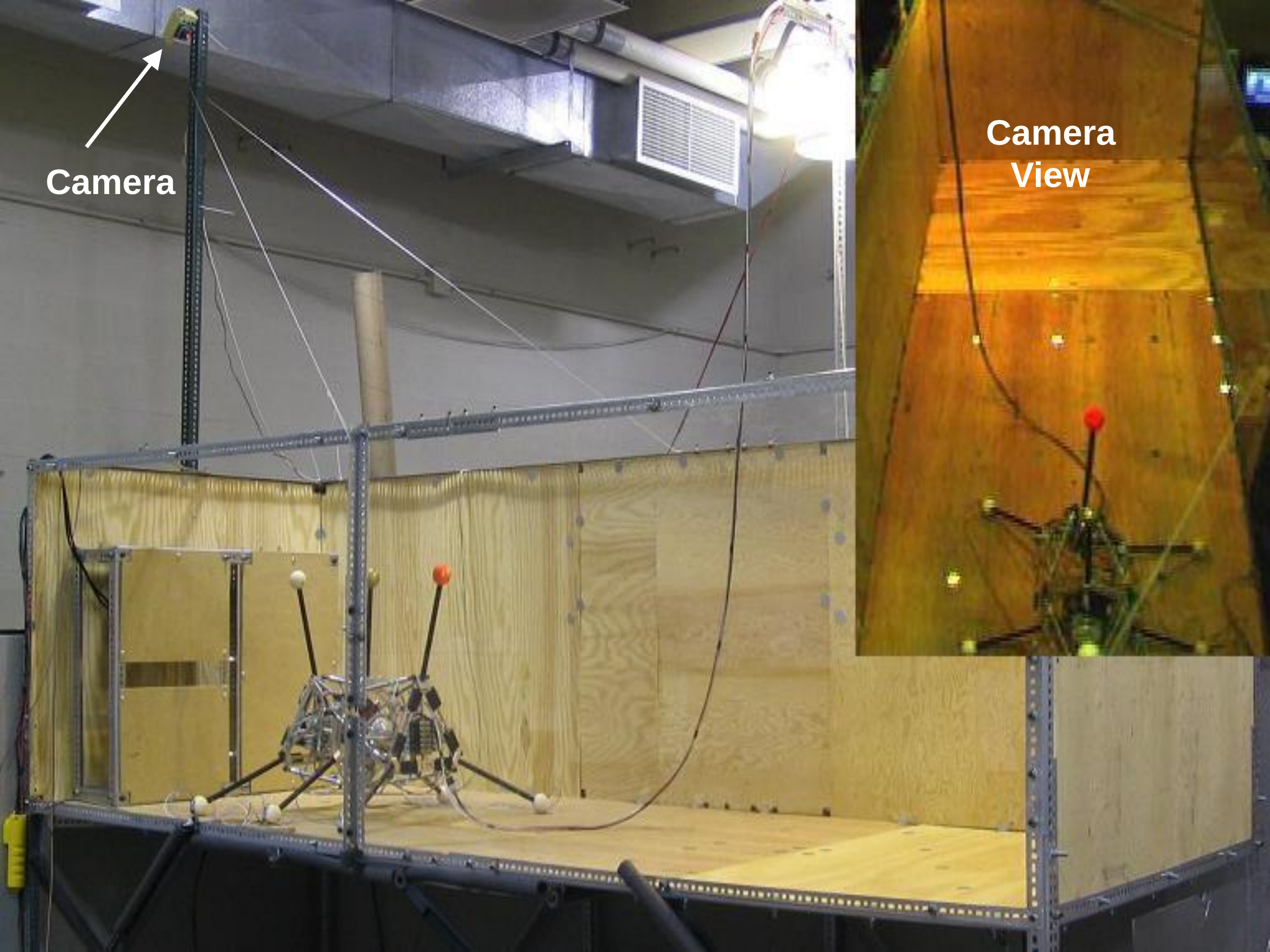




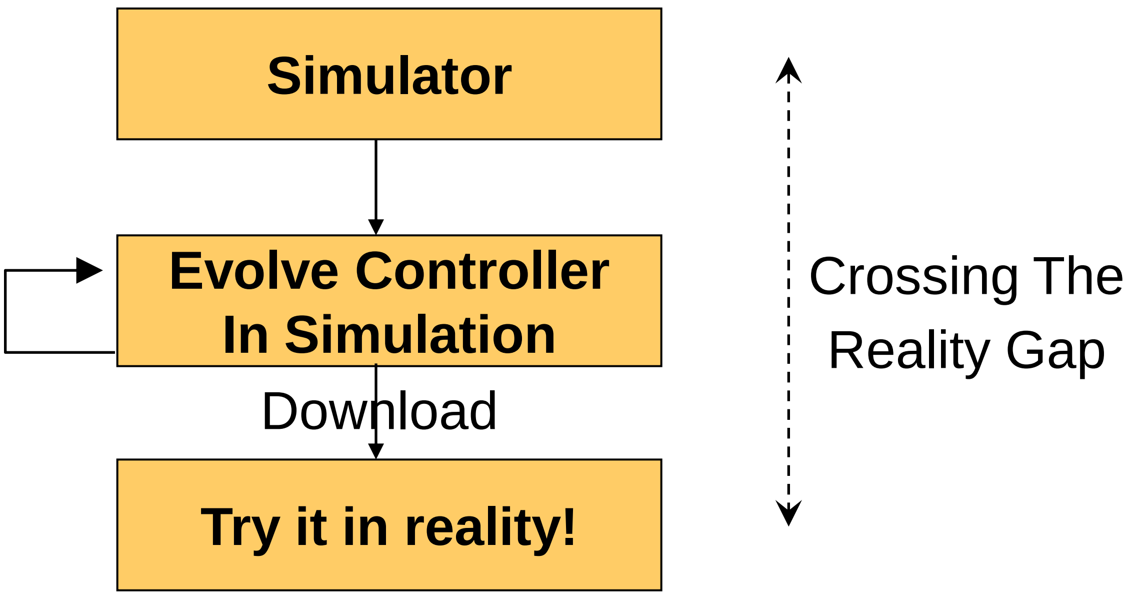
Lipson & Pollack, Nature 406, 2000



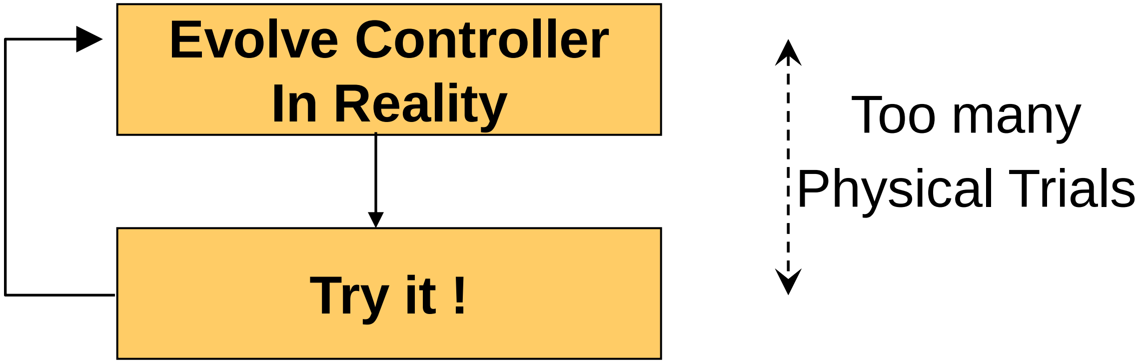



Camera
Camera
View

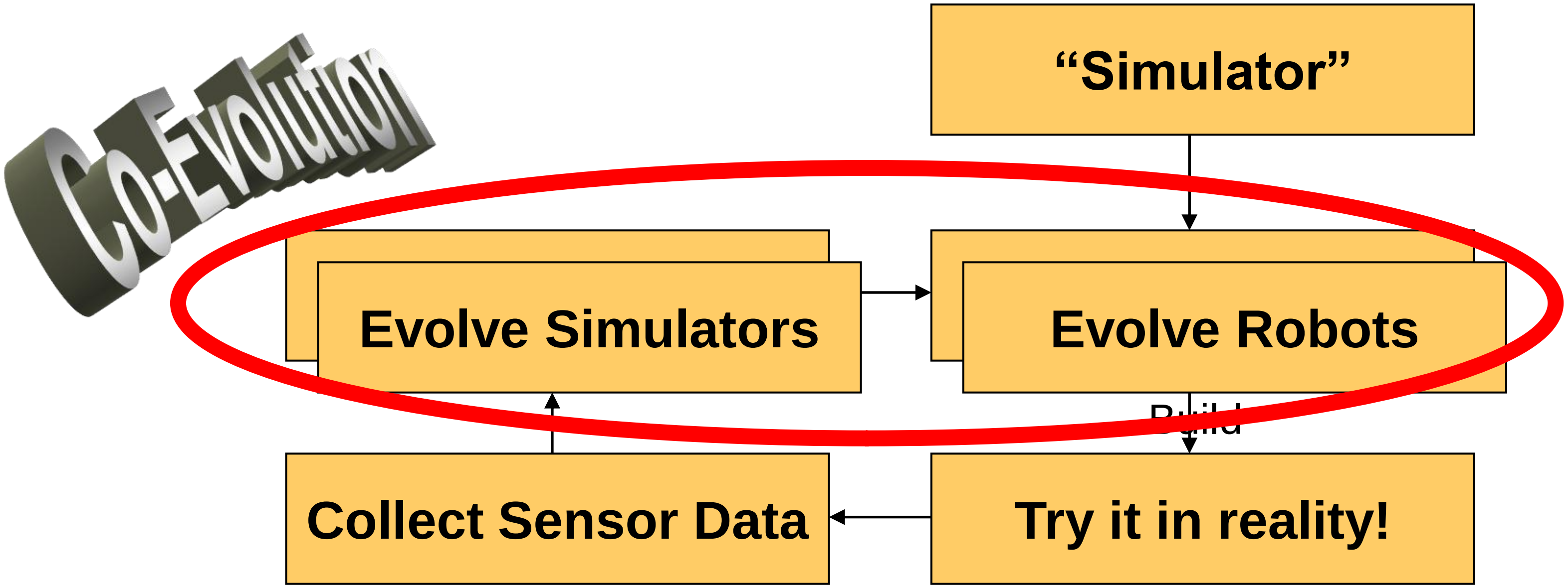
Adapting in simulation

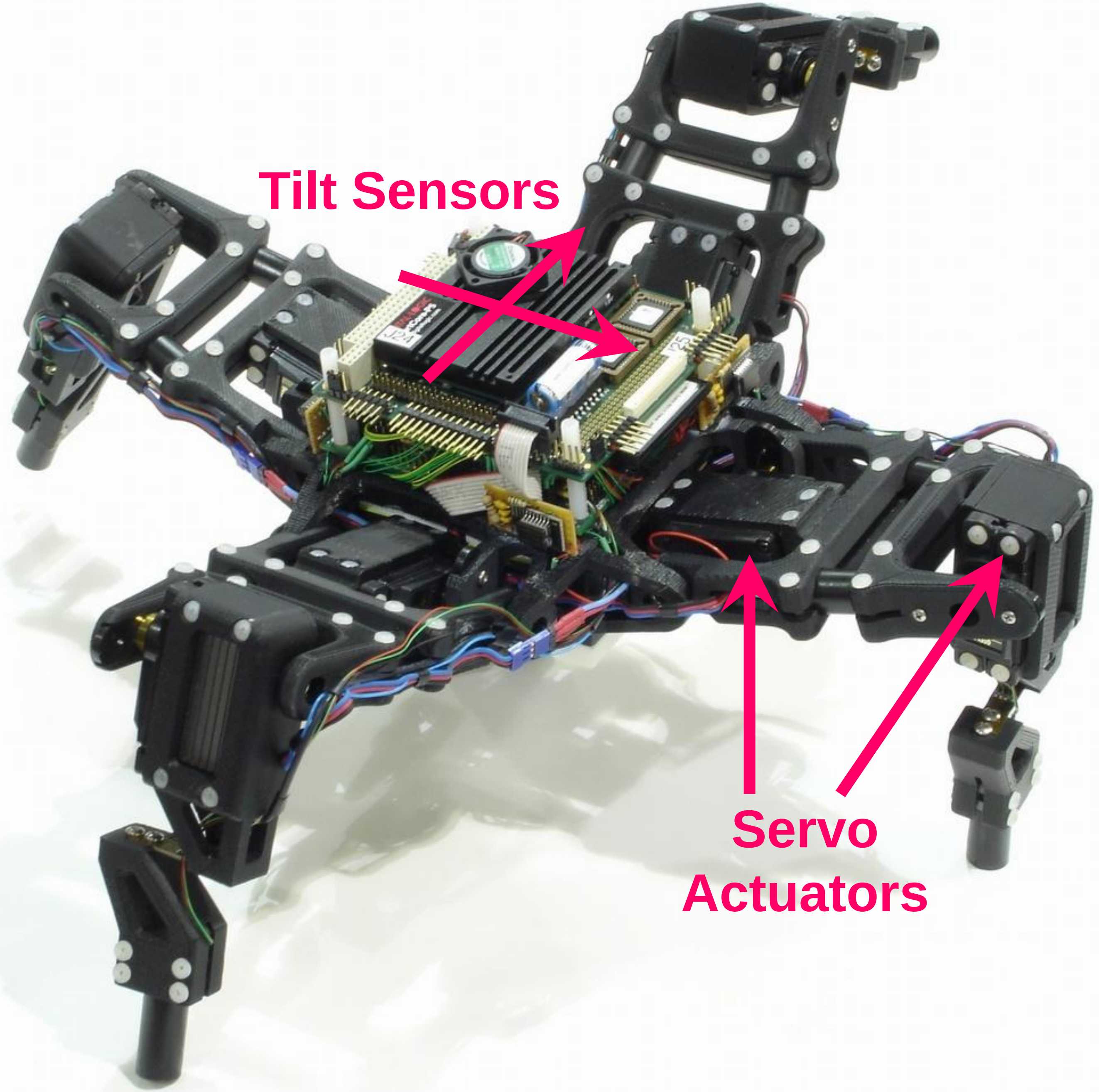


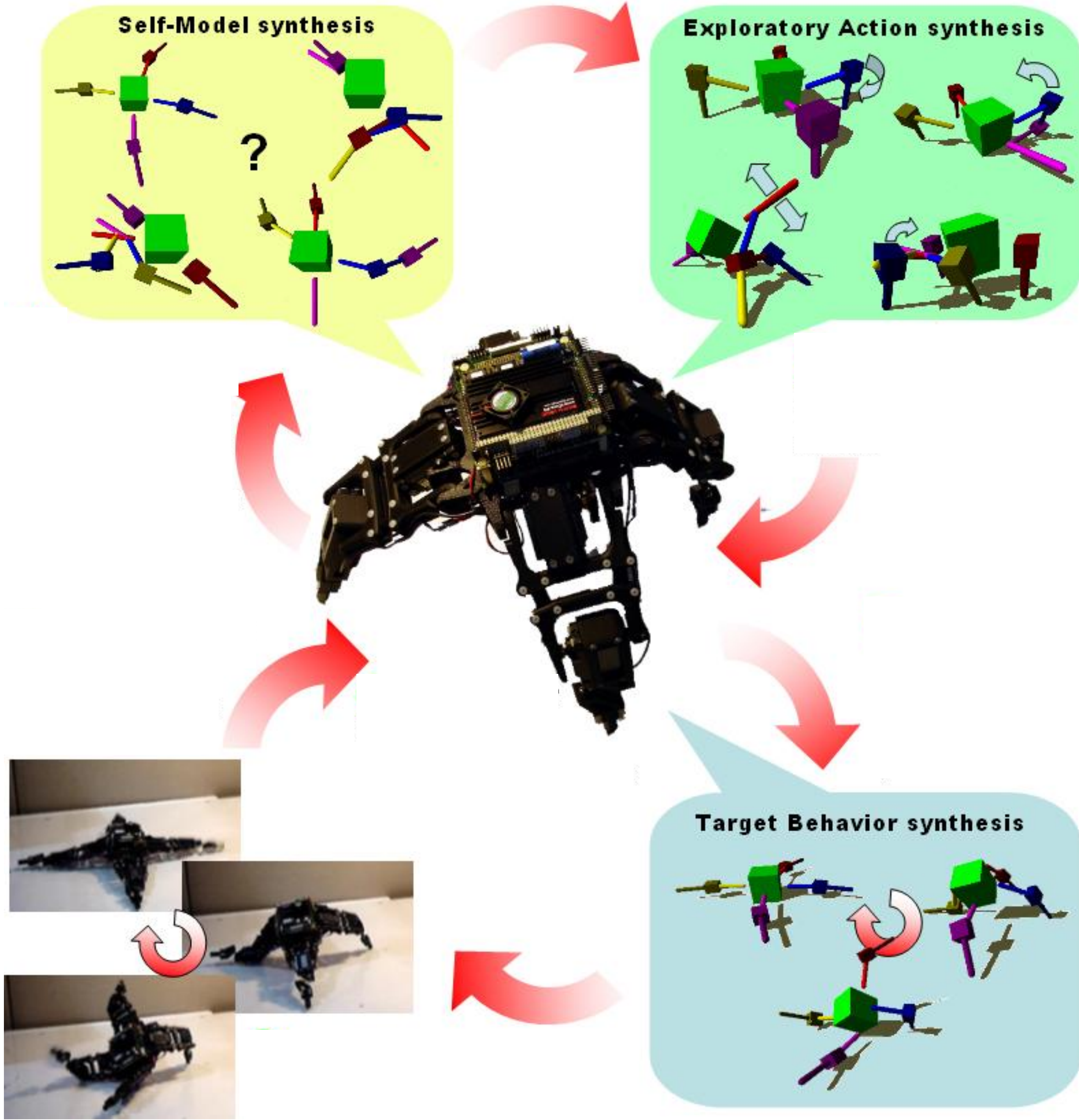
Adapting in reality



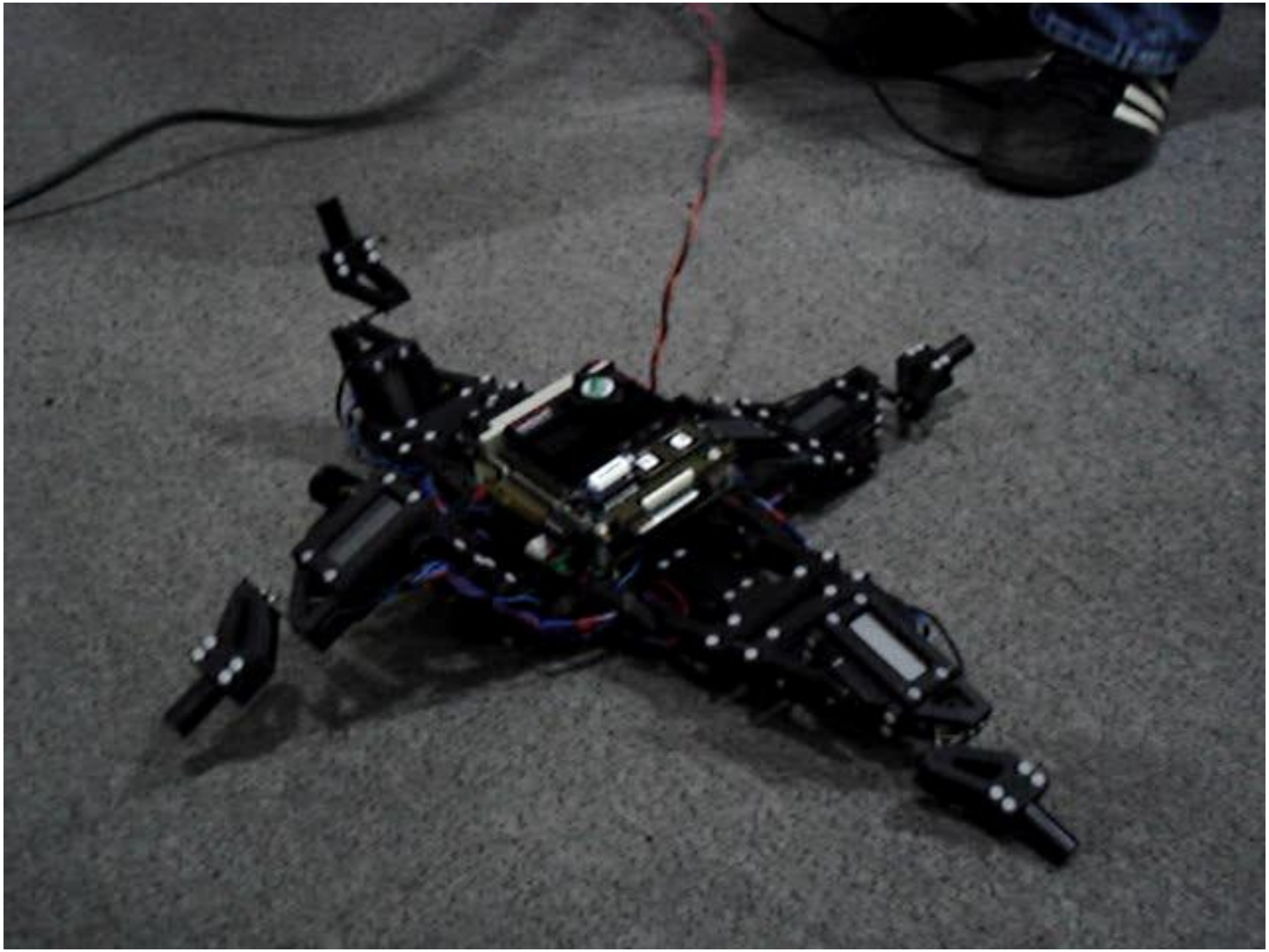
Simulation & Reality



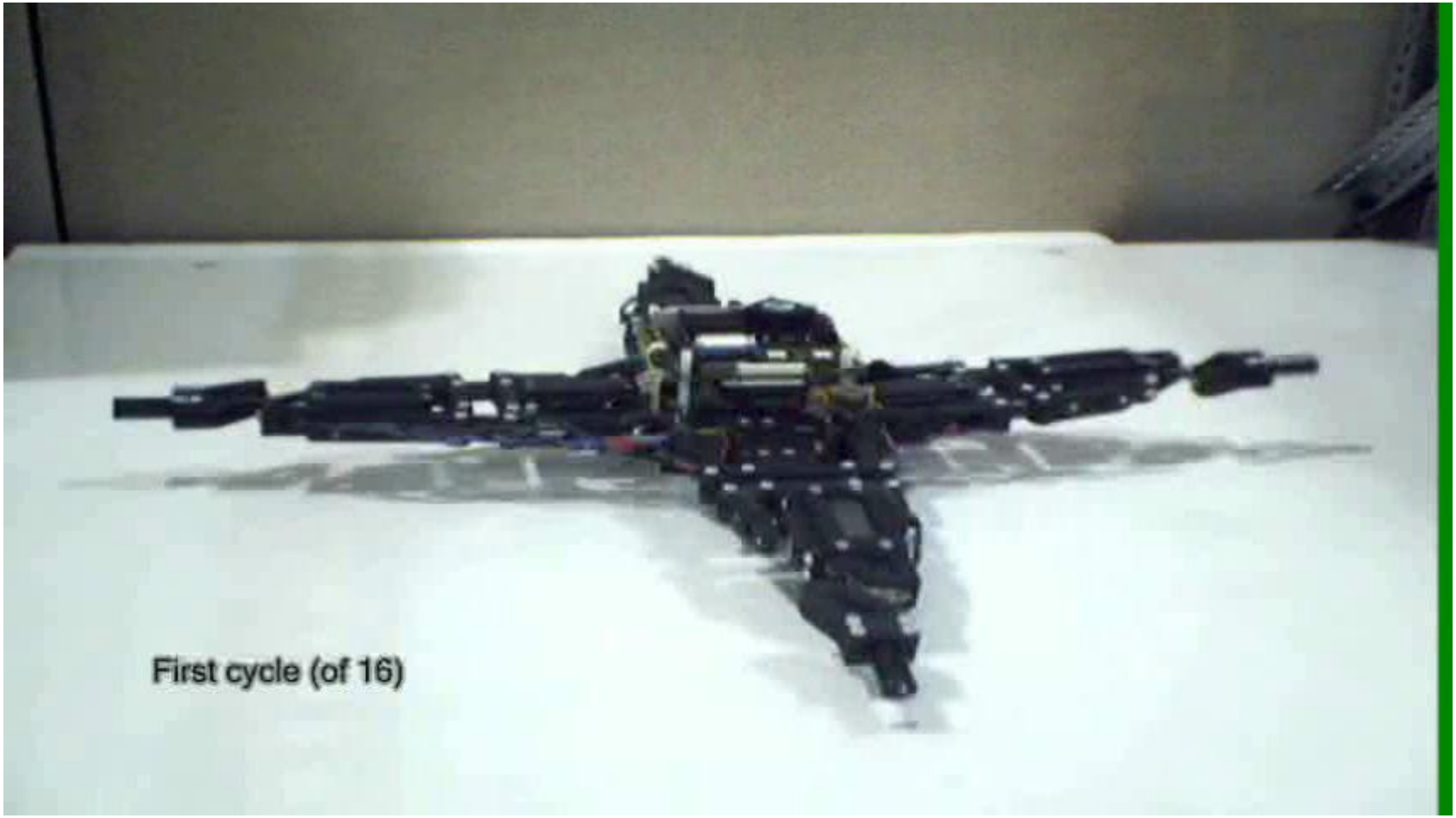

Tilt Sensors
Servo
Actuators


Self-Model synthesis
?
Exploratory Action synthesis
Target Behavior synthesis

Morphological Estimation



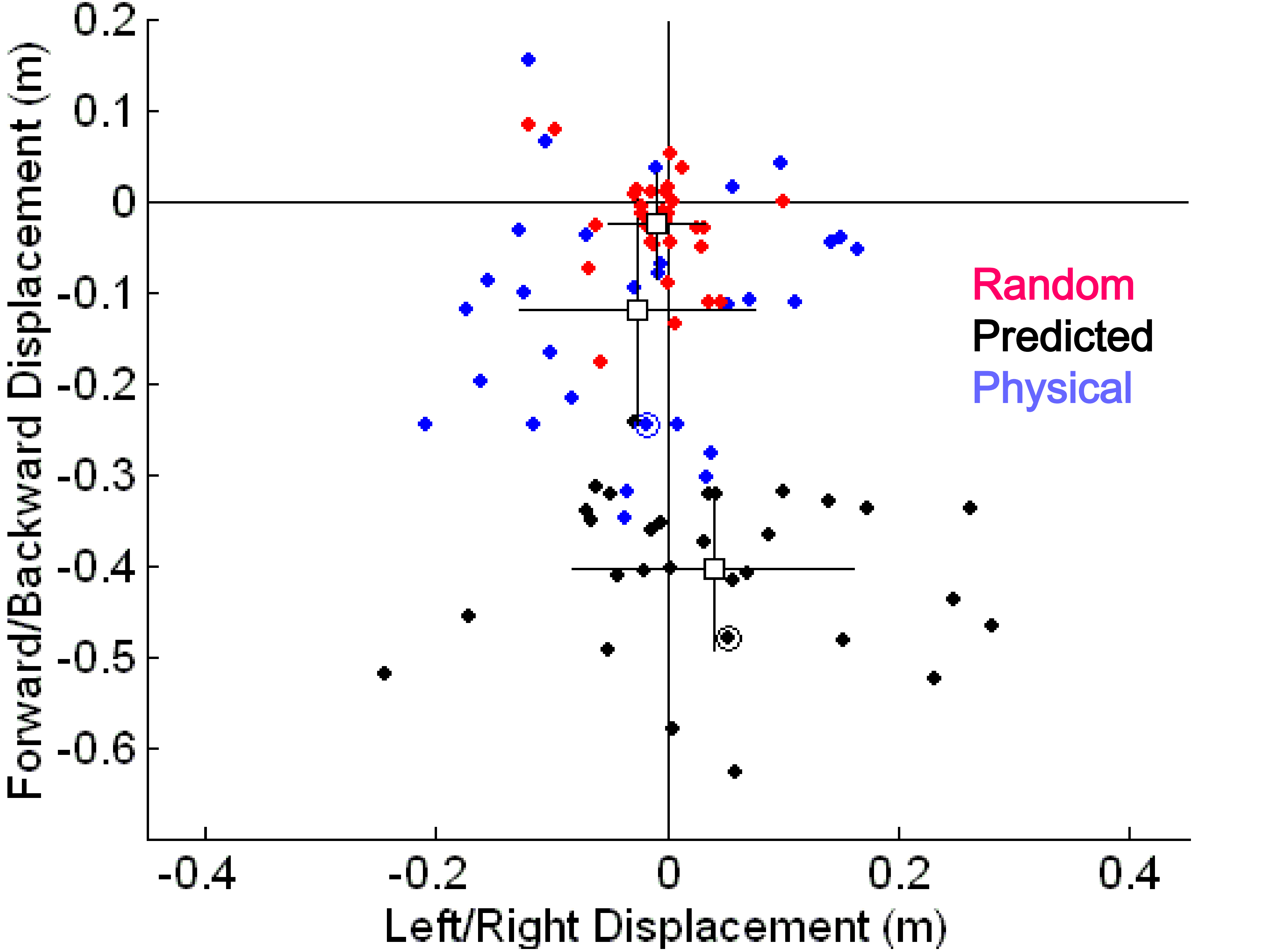
Emergent Self-Model



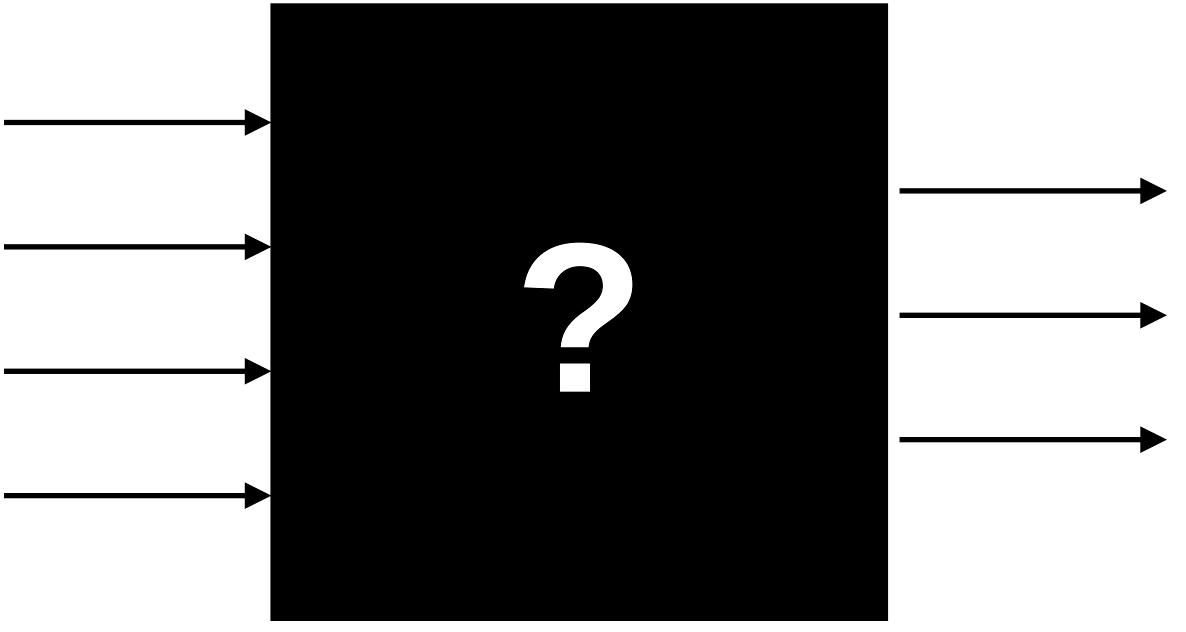
With Josh Bongard and Victor Zykov, Science 2006

Damage Recovery

With Josh Bongard and Victor Zykov, Science 2006


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

System Identification



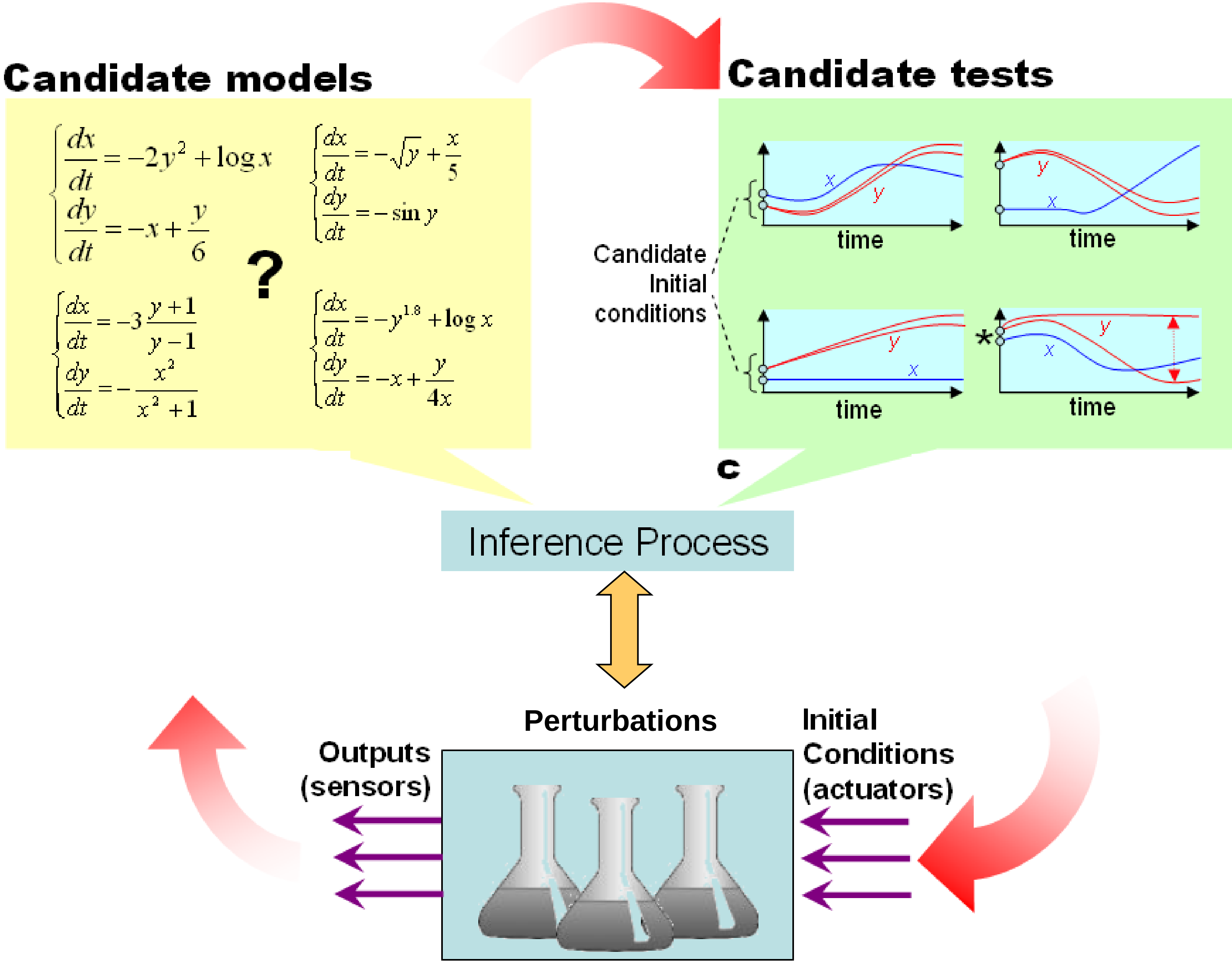
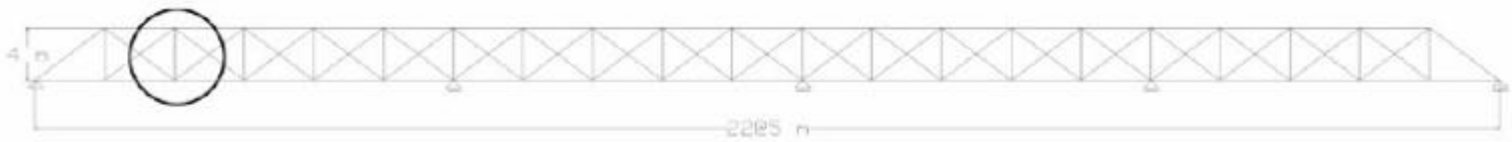
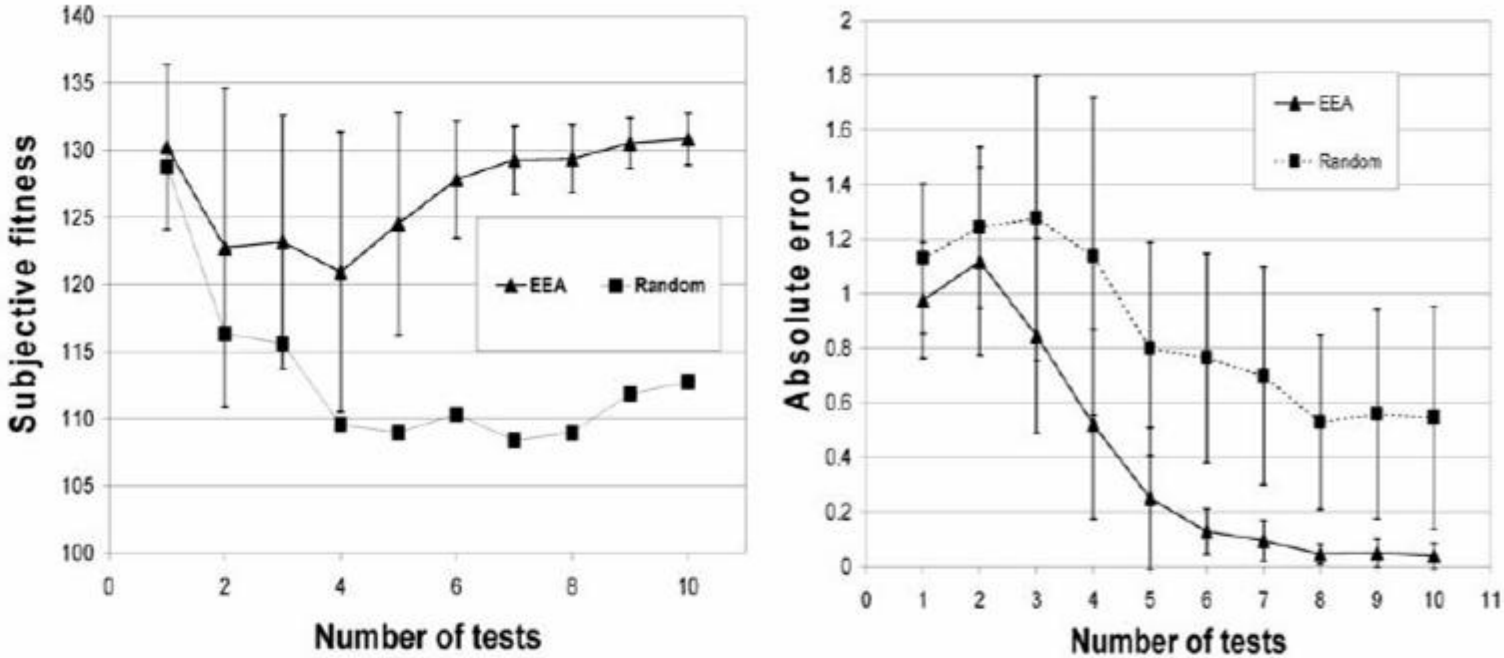

Candidate models
?
Candidate tests
Candidate
Initial
conditions
time
time
time
time
c
Inference Process
Perturbations
Outputs
(sensors)
Initial
Conditions
(actuators)

Photo: Floris van Breugel

Photo: Floris van Breugel

Static ID: Damage Diagnosis

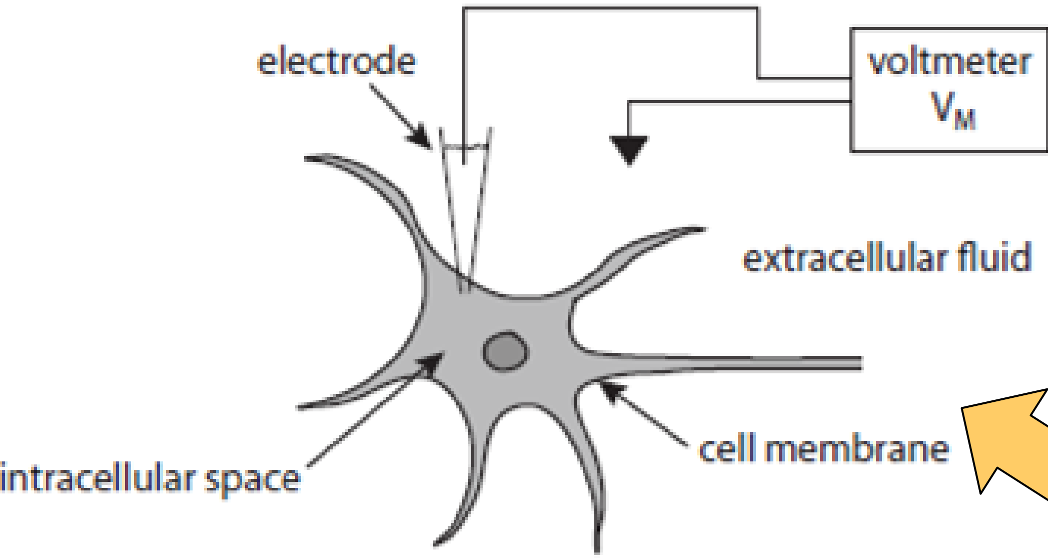


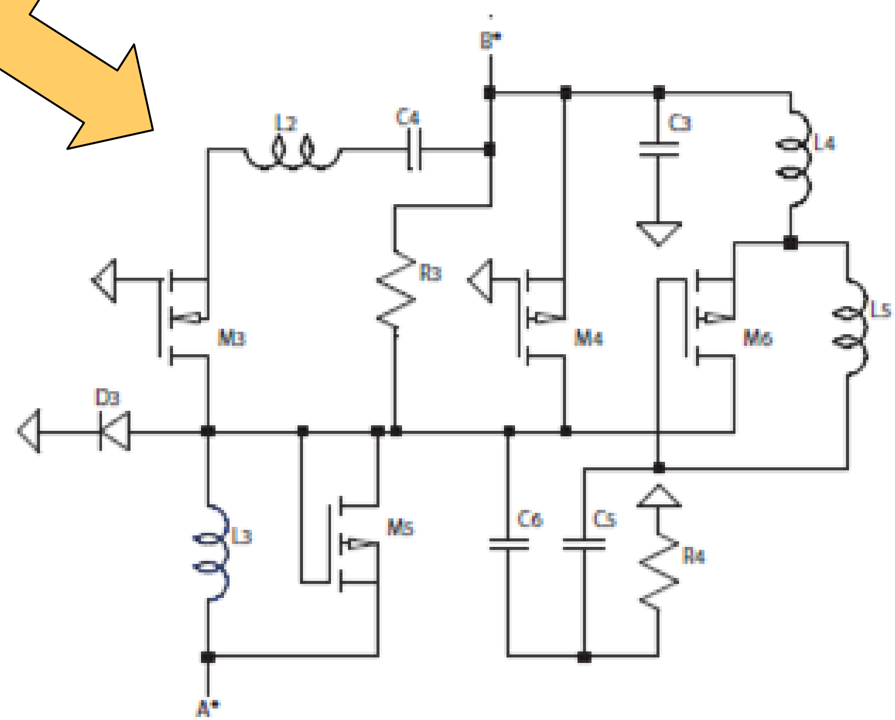


(a) (b)

With Wilkins Aquino

Circuit Building Blocks

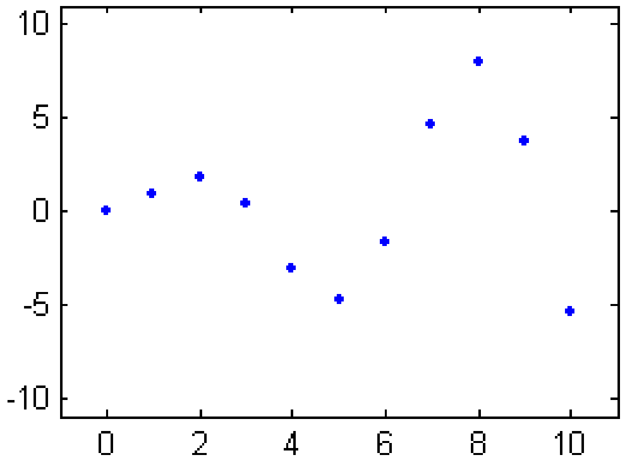


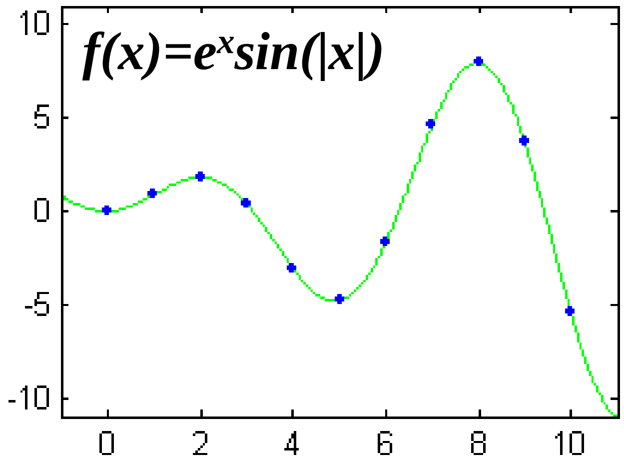


Ted Cornforth

Symbolic Regression

What function describes this data?

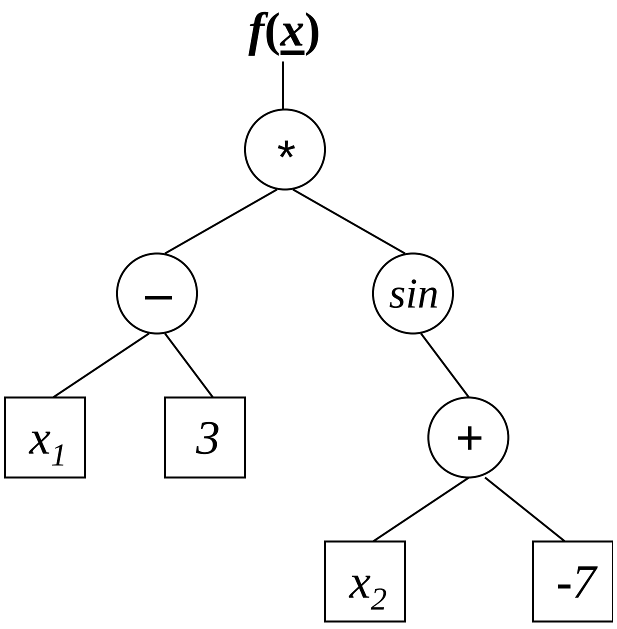


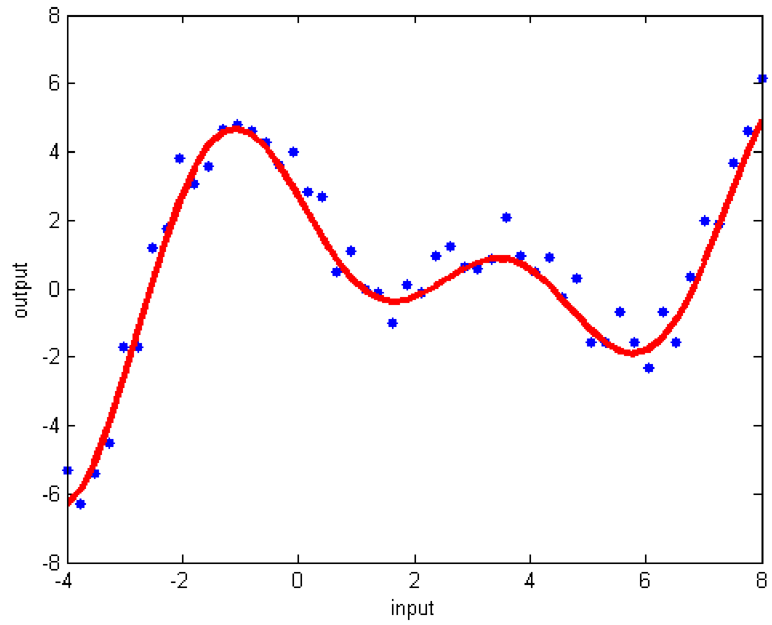


John Koza, 1992

Encoding Equations

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





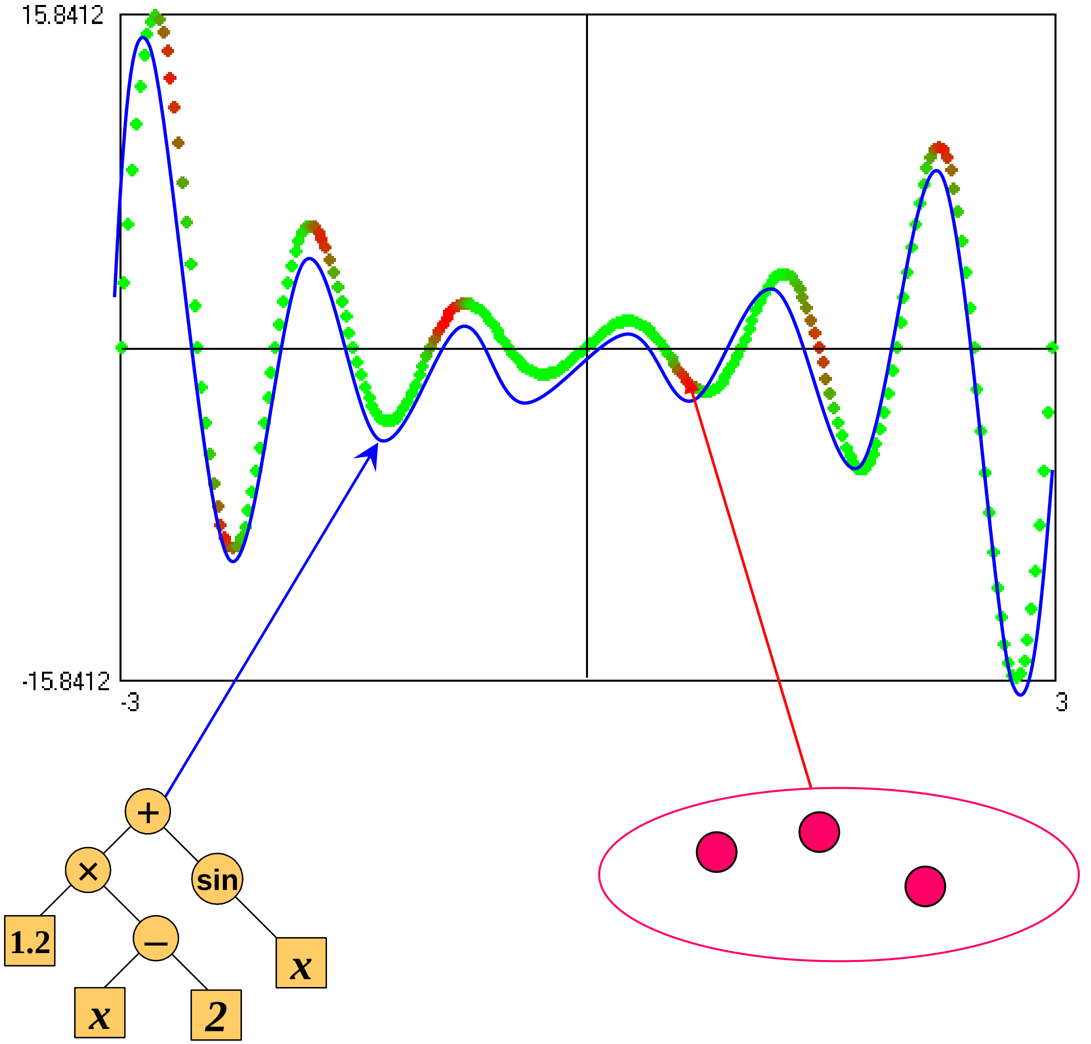
sin(x_2)

x_1*sin(x_2)

(x_1 – 3)*sin(x_2)

(x_1 – 3)*sin(-7 + x_2)

John Koza, 1992



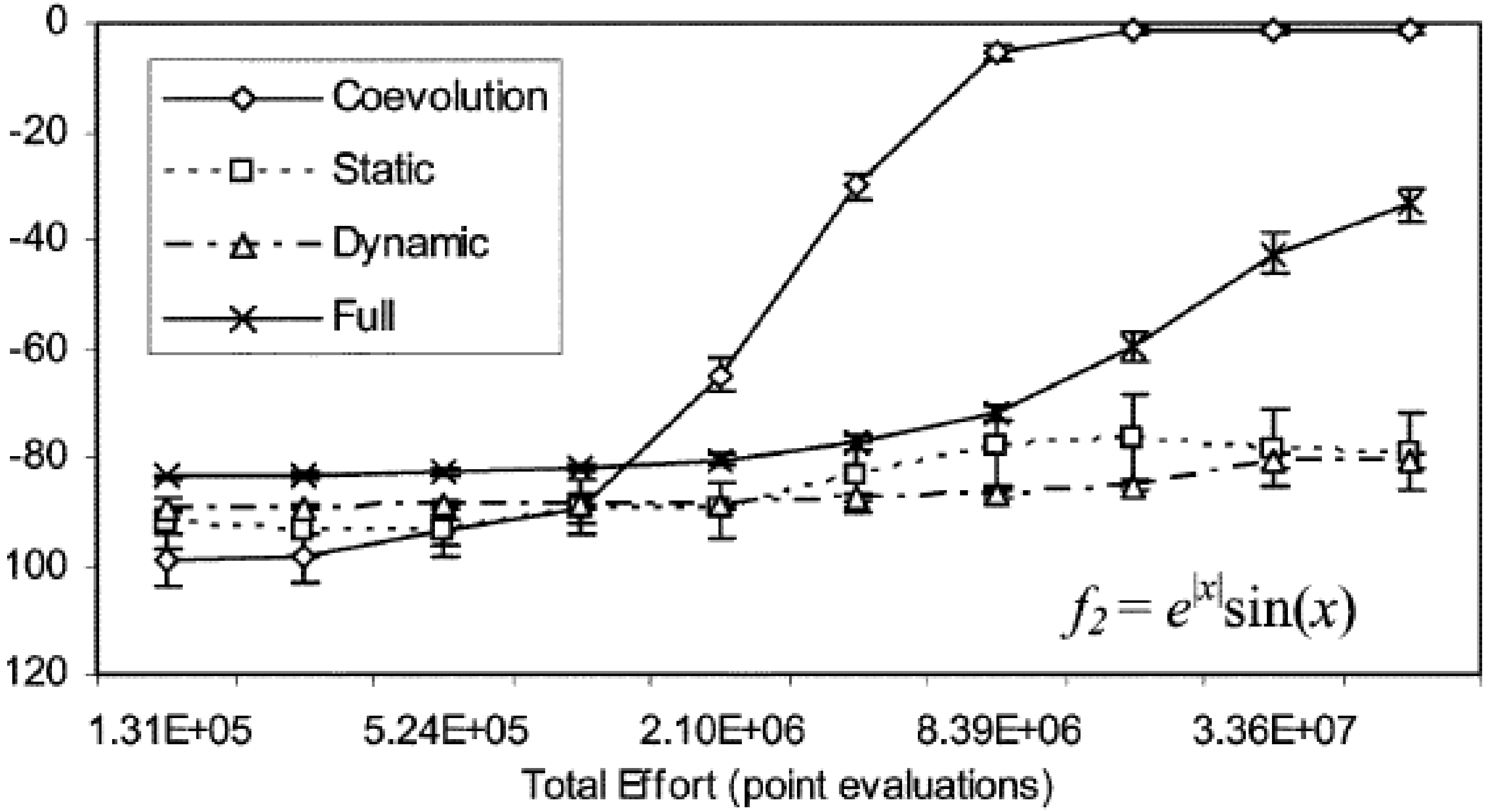
Models: Expression trees
Subject to mutation and selection

Experiments: Data-points
Subject to mutation and selection

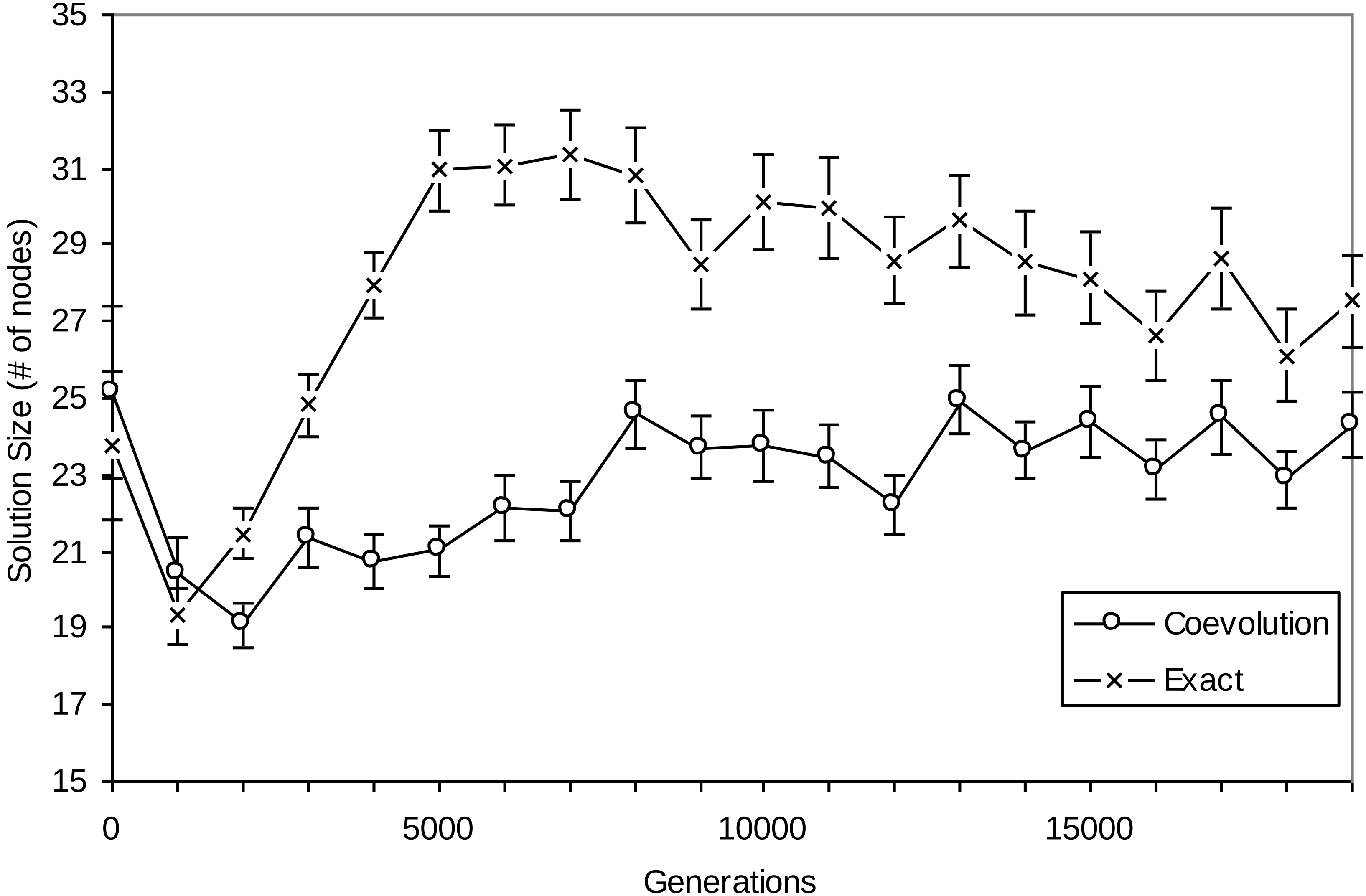
`{const,+,-,*,/,sin,cos,exp,log,abs}`

Michael D. Schmidt, Hod Lipson (2006)

Solution Accuracy



Solution Complexity



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:

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

Weizsäcker's Formula:

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

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

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

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

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

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

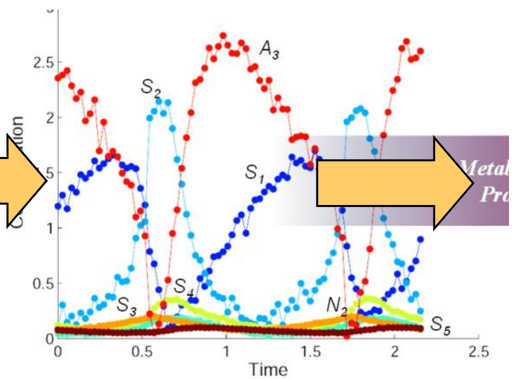
$$\frac{dS_4}{dt} = \underbrace{16 * A_2 * S_3}_{v_3} - \underbrace{100 * N_2 * S_4}_{-v_4}$$

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

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

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

Original Equations



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

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

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

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

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

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

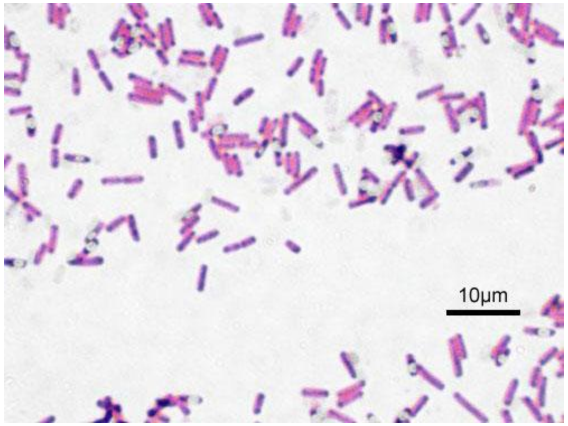
$$\frac{dS_5}{dt} = \underbrace{-1.29626 * S_5}_{\varphi J}$$

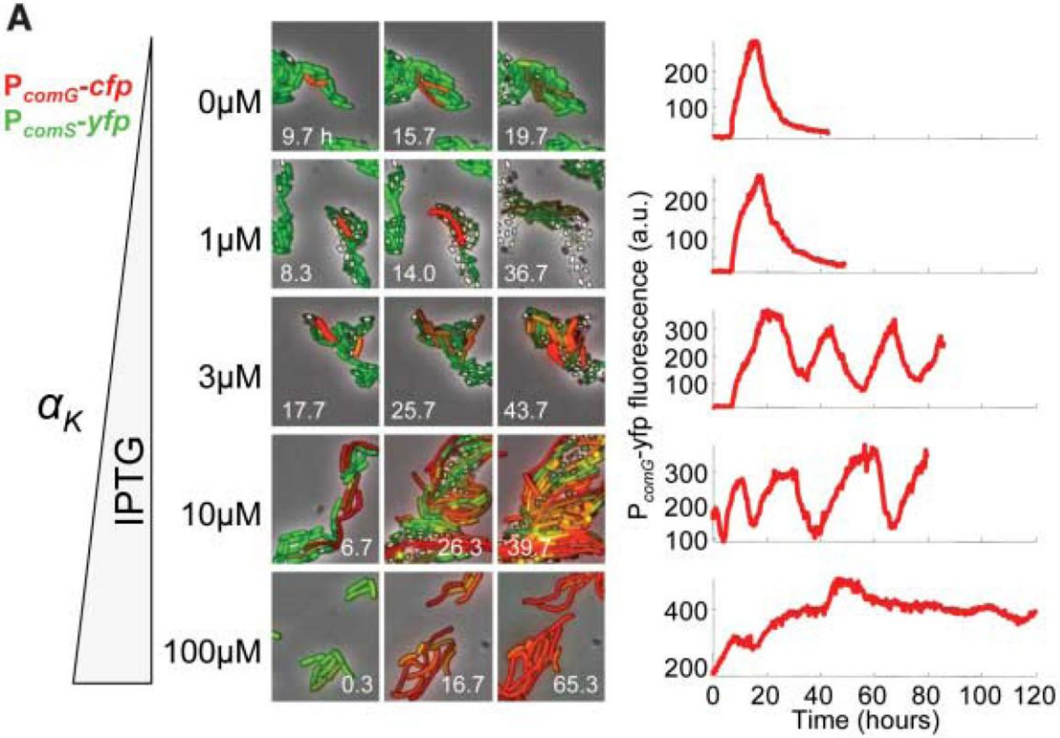
Inferred Equations

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

Wet Data, Unknown System

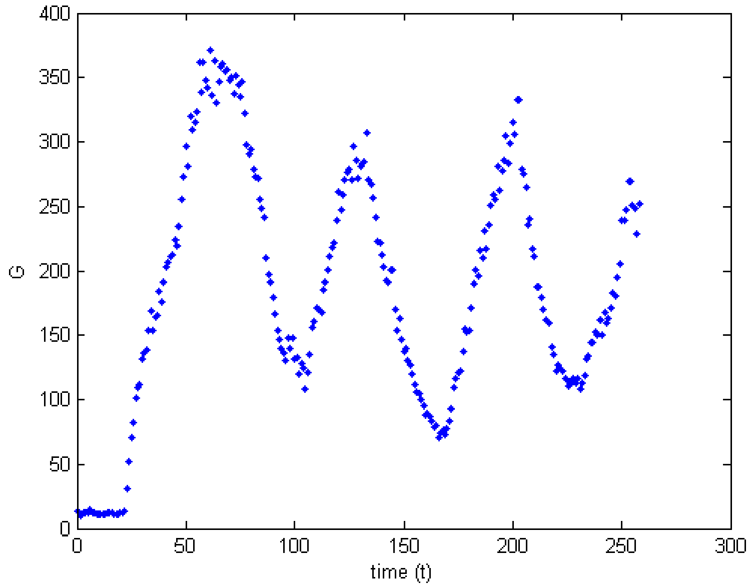
Bacillus Bacteria

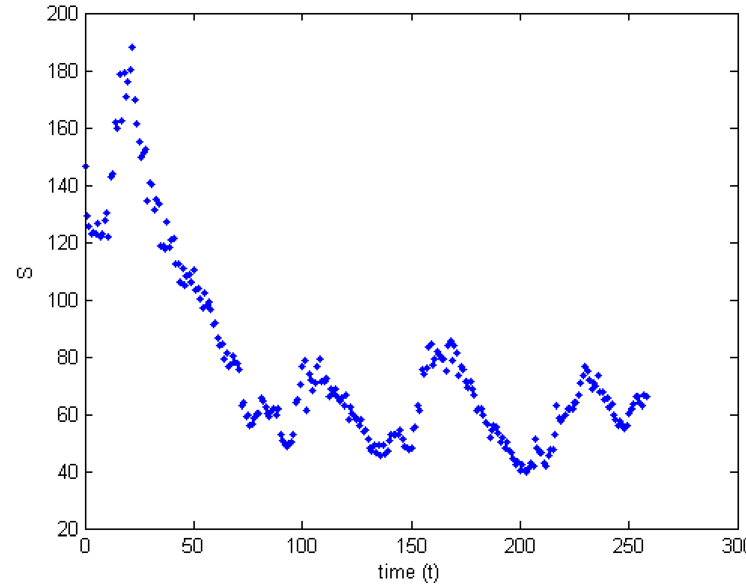




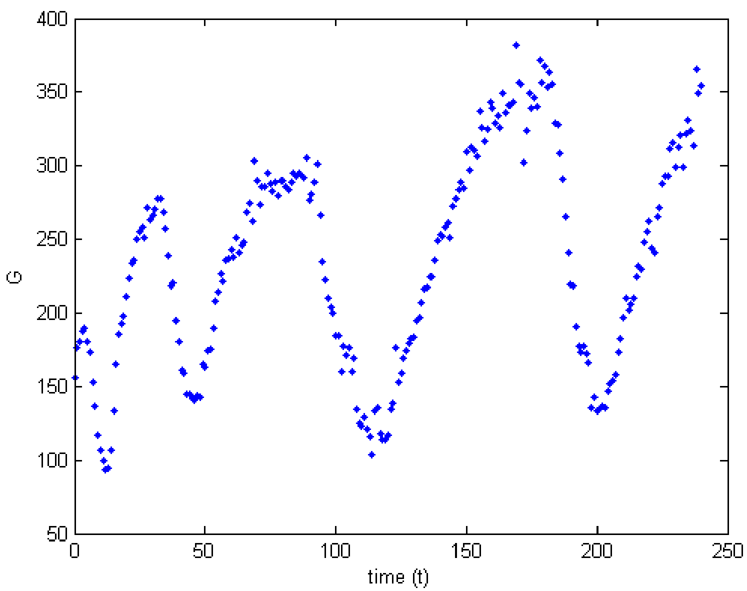
With Michael Schmidt (Cornell) and Gurol Suel (UT Southwestern)

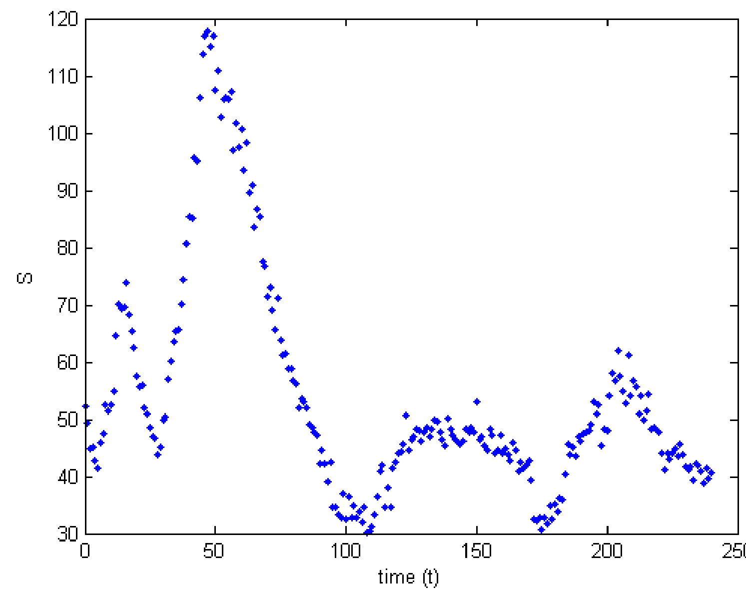
Cell #1





Cell #2

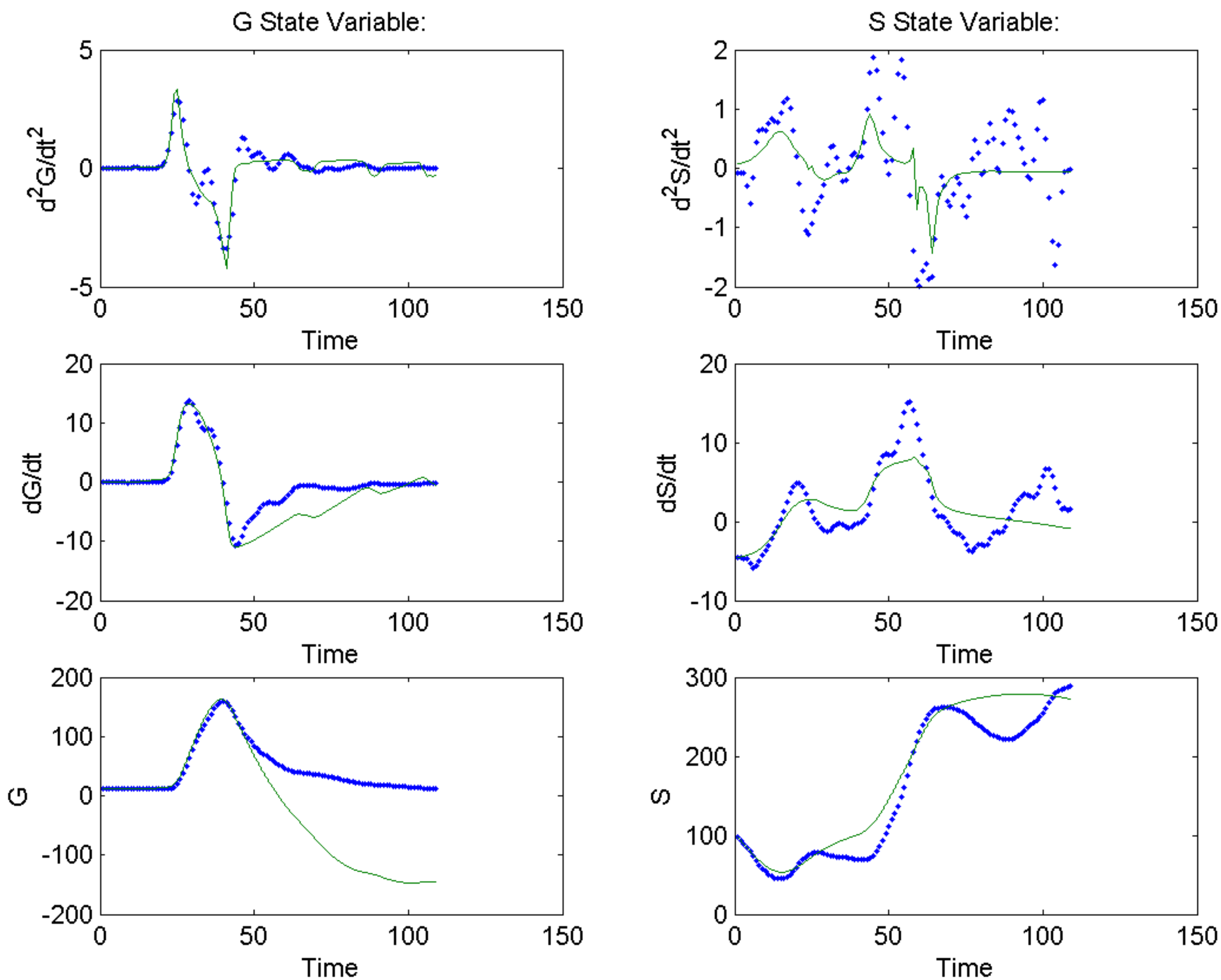




Cell #3-60 …

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

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



Blue Dots = data points, Green Line = regressed fit

Symbolic Regression Inferred *Time-Delay* Model:

$$\frac{d\boldsymbol{K}}{d\boldsymbol{t}} = \alpha_{\boldsymbol{K}} + \frac{\beta_{\boldsymbol{K}} + \delta_{\boldsymbol{K}} \boldsymbol{S}_{\boldsymbol{t}-\boldsymbol{t}_1}}{\boldsymbol{K}_{\boldsymbol{t}-\boldsymbol{t}_2}}$$

$$\frac{d\boldsymbol{S}}{d\boldsymbol{t}} = \alpha_{\boldsymbol{S}} + \frac{\beta_{\boldsymbol{S}} + \delta_{\boldsymbol{S}} \boldsymbol{G}_{\boldsymbol{t}-\boldsymbol{t}_3}}{\boldsymbol{S}_{\boldsymbol{t}-\boldsymbol{t}_4}}$$

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

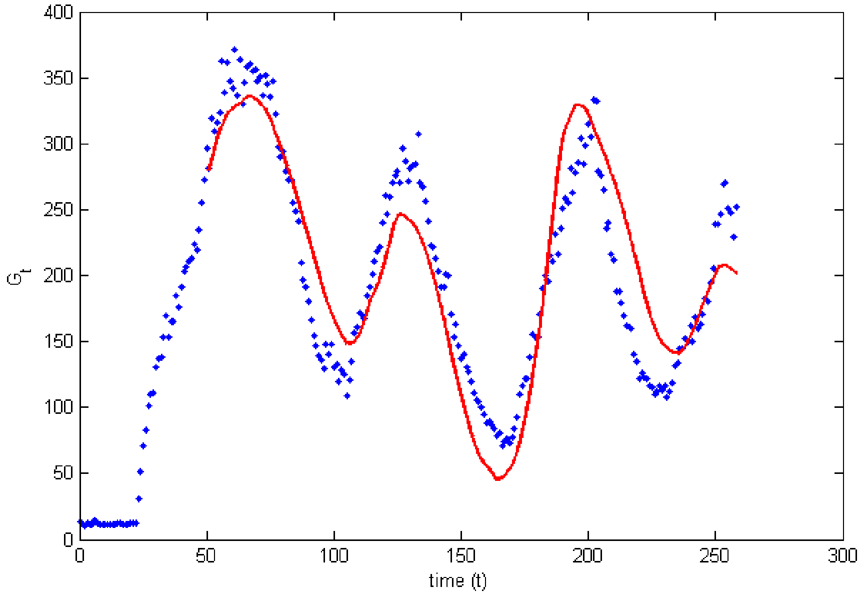
$$\frac{d\boldsymbol{K}}{d\boldsymbol{t}} = \alpha_{\boldsymbol{k}} + \frac{\beta_{\boldsymbol{K}} \boldsymbol{K}^{\boldsymbol{n}}}{\boldsymbol{k}_0^{\boldsymbol{n}} + \boldsymbol{K}^{\boldsymbol{n}}} - \frac{\delta_{\boldsymbol{K}} \boldsymbol{K}}{1 + \boldsymbol{K}/\Gamma_{\boldsymbol{K}} + \boldsymbol{S}/\Gamma_{\boldsymbol{S}}}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

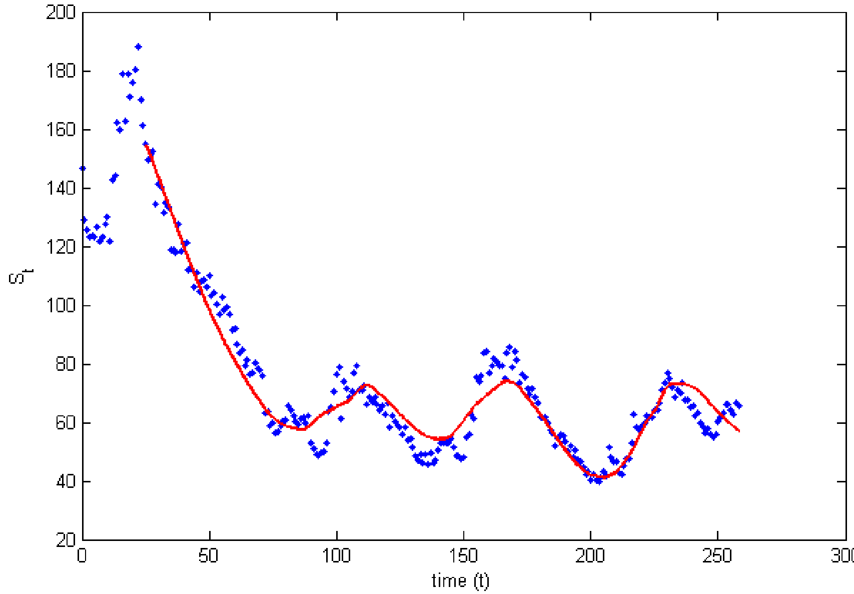
$$\frac{d\boldsymbol{S}}{d\boldsymbol{t}} = \alpha_{\boldsymbol{S}} + \frac{\beta_{\boldsymbol{S}}}{1 + \left(\boldsymbol{K}/\boldsymbol{k}_1\right)^{\boldsymbol{p}}} - \frac{\delta_{\boldsymbol{k}} \boldsymbol{S}}{1 + \boldsymbol{K}/\Gamma_{\boldsymbol{K}} + \boldsymbol{S}/\Gamma_{\boldsymbol{S}}} -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

Withheld Test Set #1 Fit

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

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

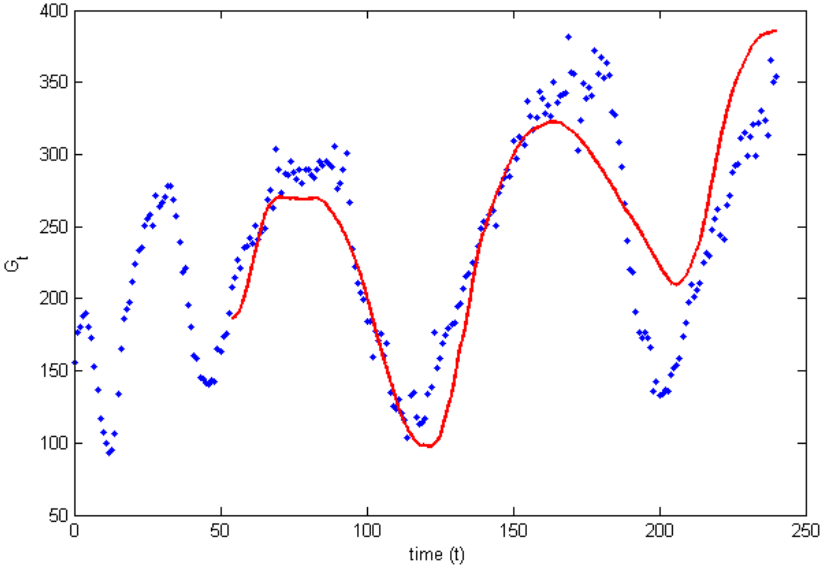


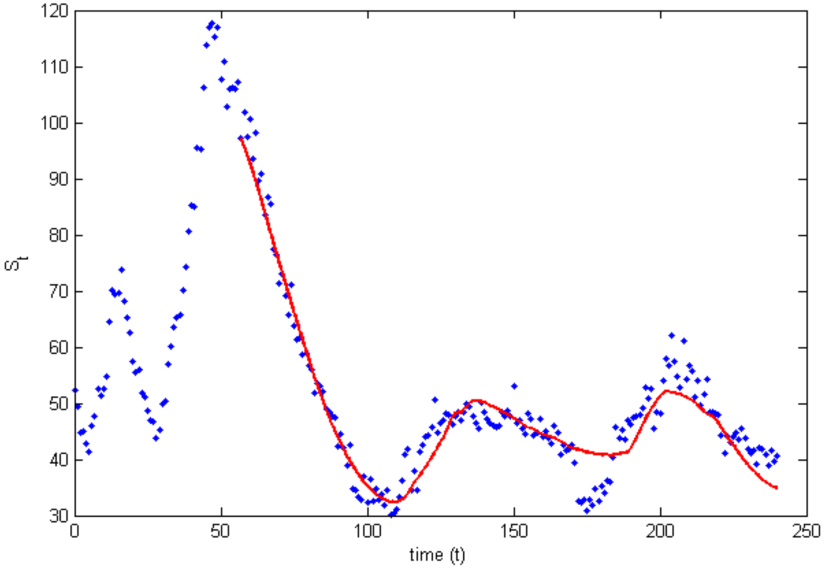


Withheld Test Set #2 Fit

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

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

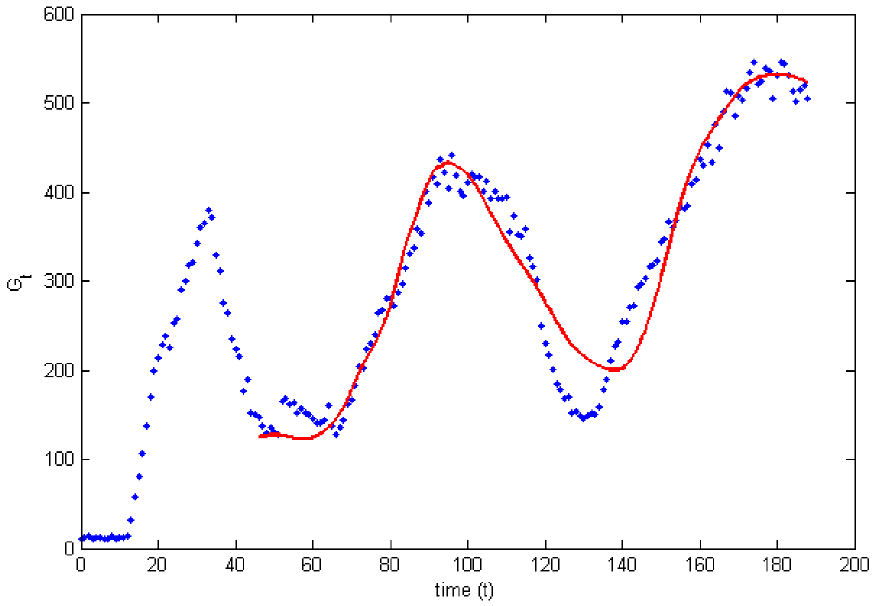


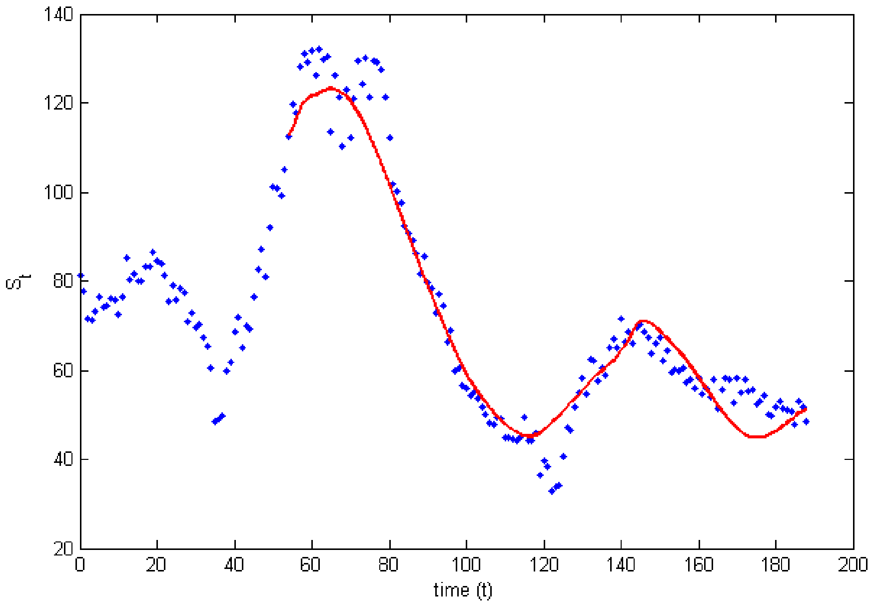


Withheld Test Set #3 Fit

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

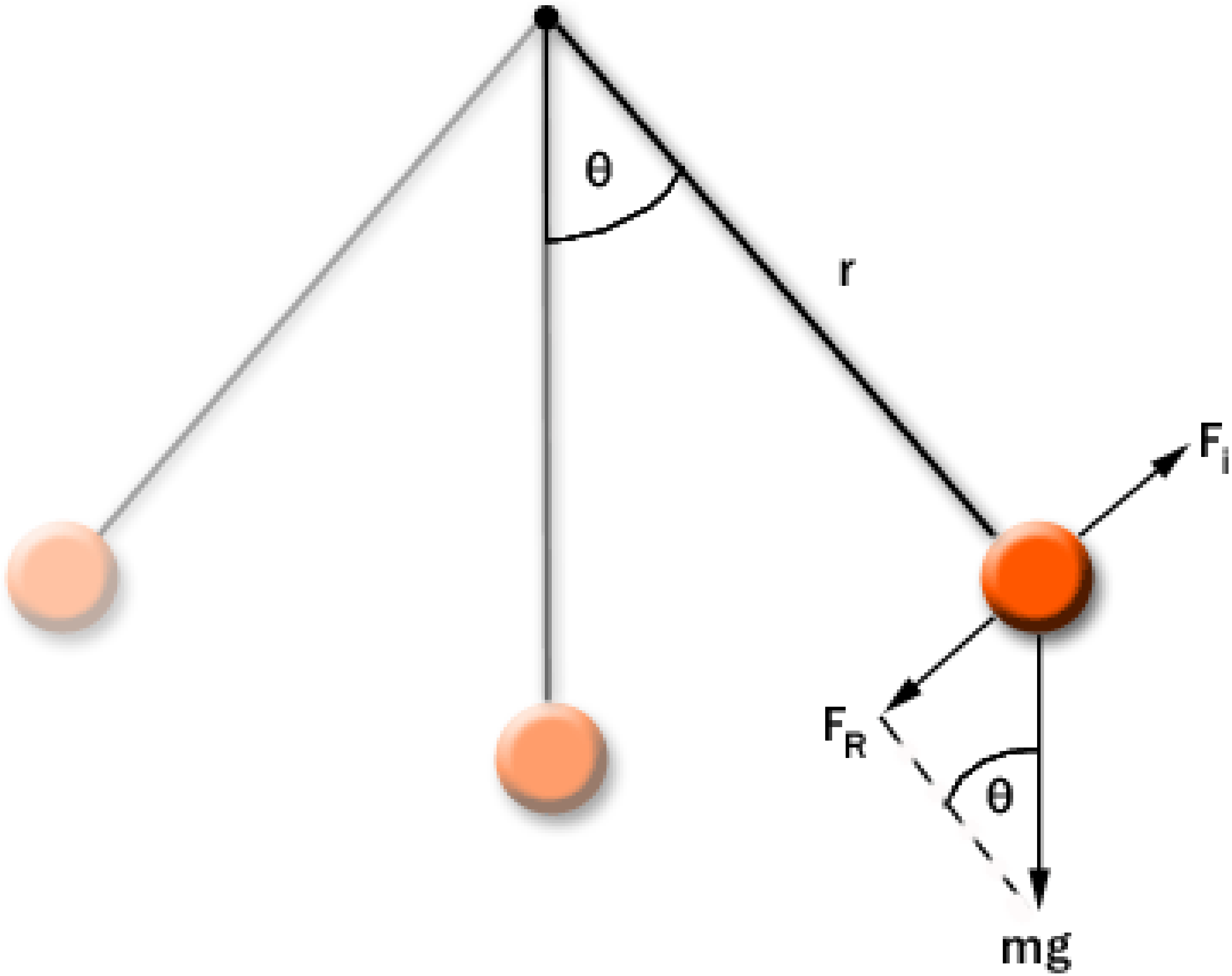
$$\frac{dS_t}{dt} = \frac{295.426 - 0.2965 \cdot G_{t-54}}{S_{t-20}} - 3.871$$





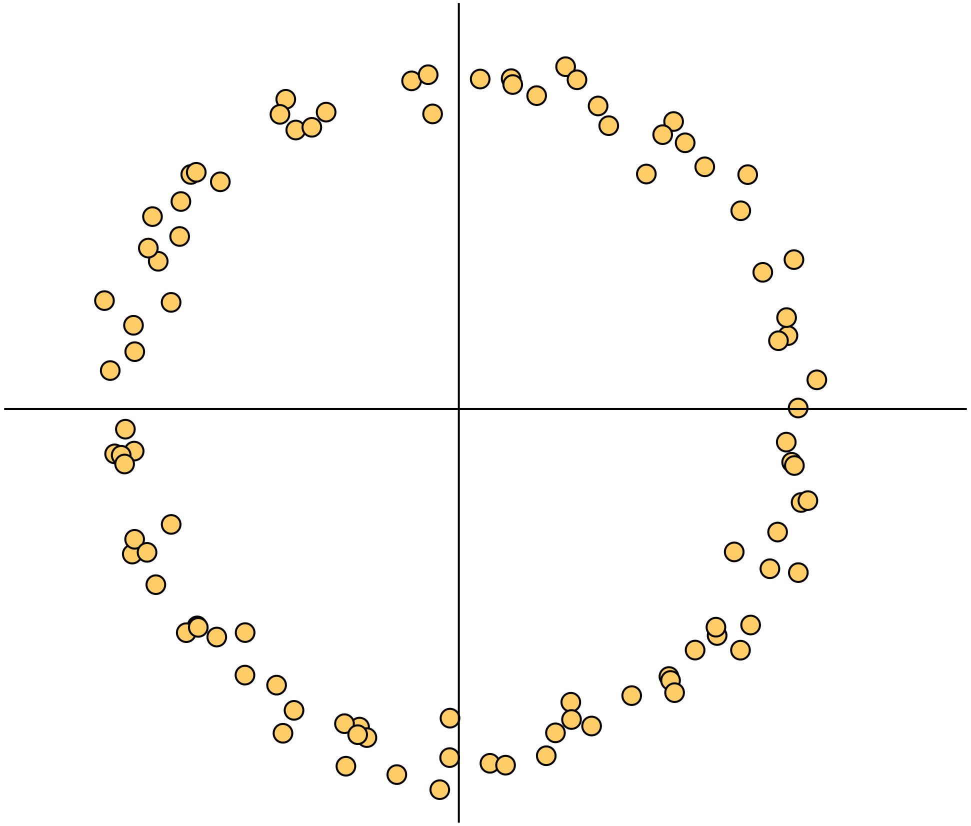
Looking For Invariants

Data Mining

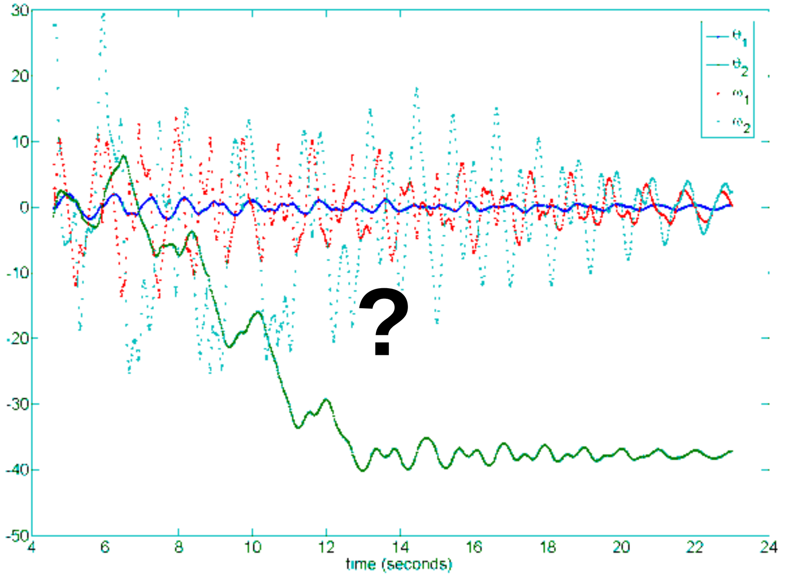

θ
r
F_i
F_R
θ
mg

Invariants = Unsupervised Learning

Mining for invariants in data



$$x^2+y^2-R^2 \approx 0$$



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:

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

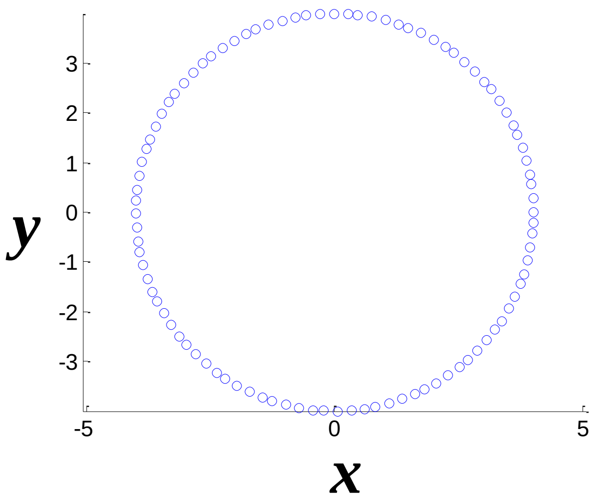
From Equation:

Calculate predicted partial derivatives Symbolically:

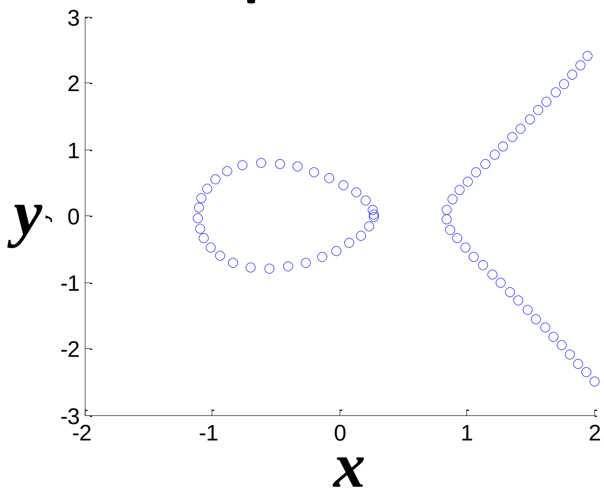
$$\frac{\delta f}{\delta x} \Big/ \frac{\delta f}{\delta y} \longrightarrow \quad \frac{\delta x'}{\delta y'}, \quad \frac{\delta y'}{\delta x'}, \quad \ldots$$

Experiments

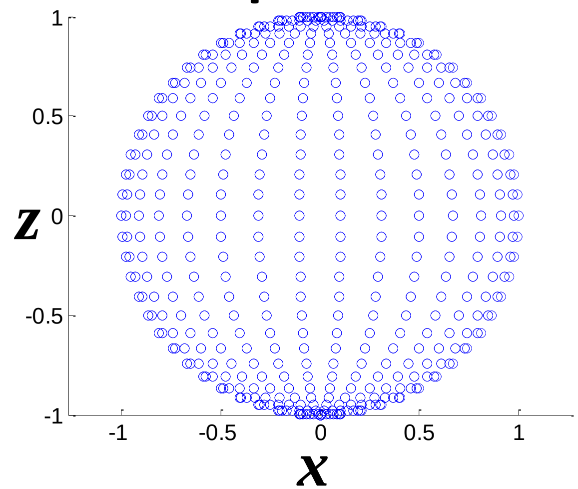
Circle



Elliptic Curve



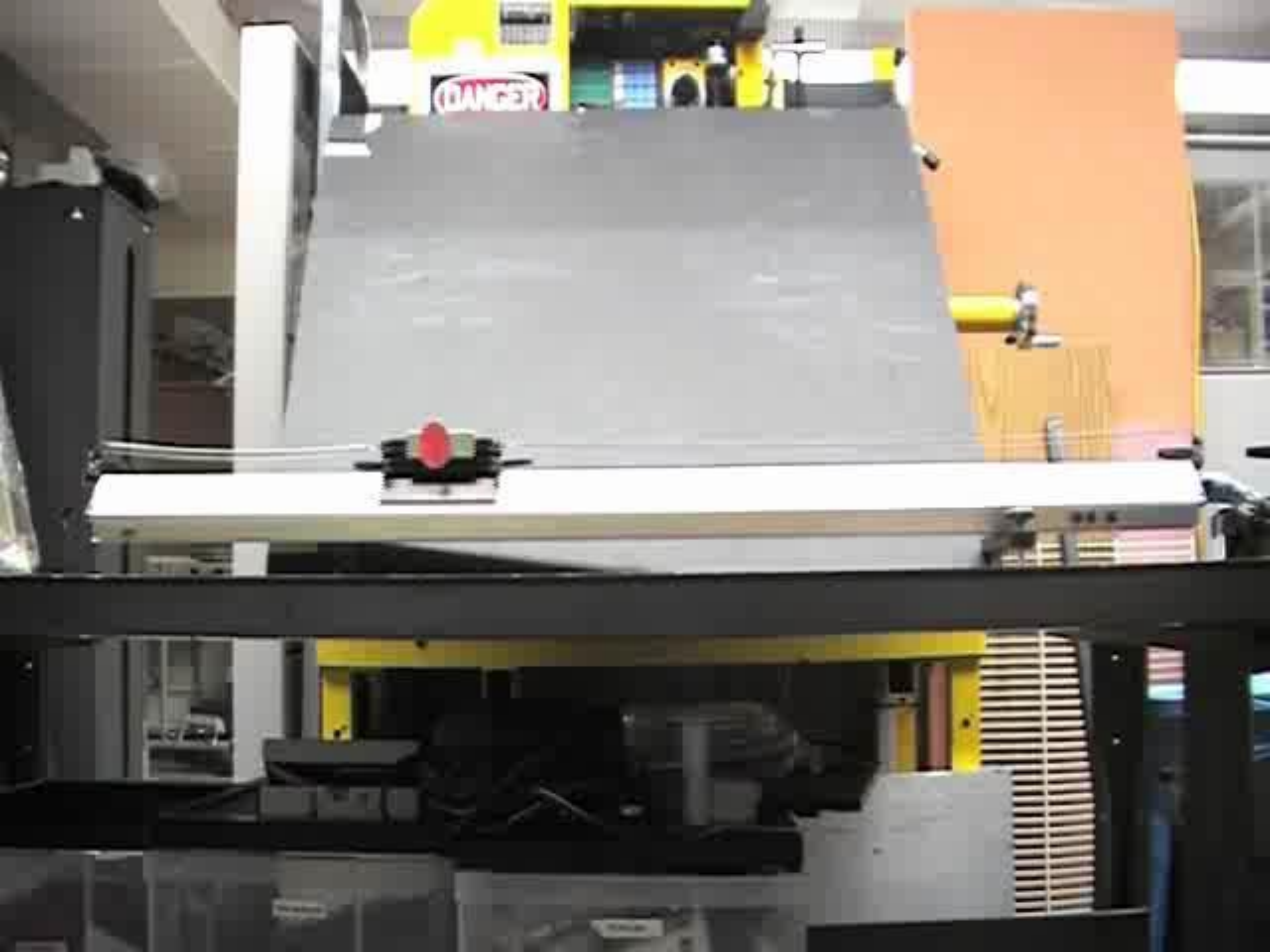
Sphere

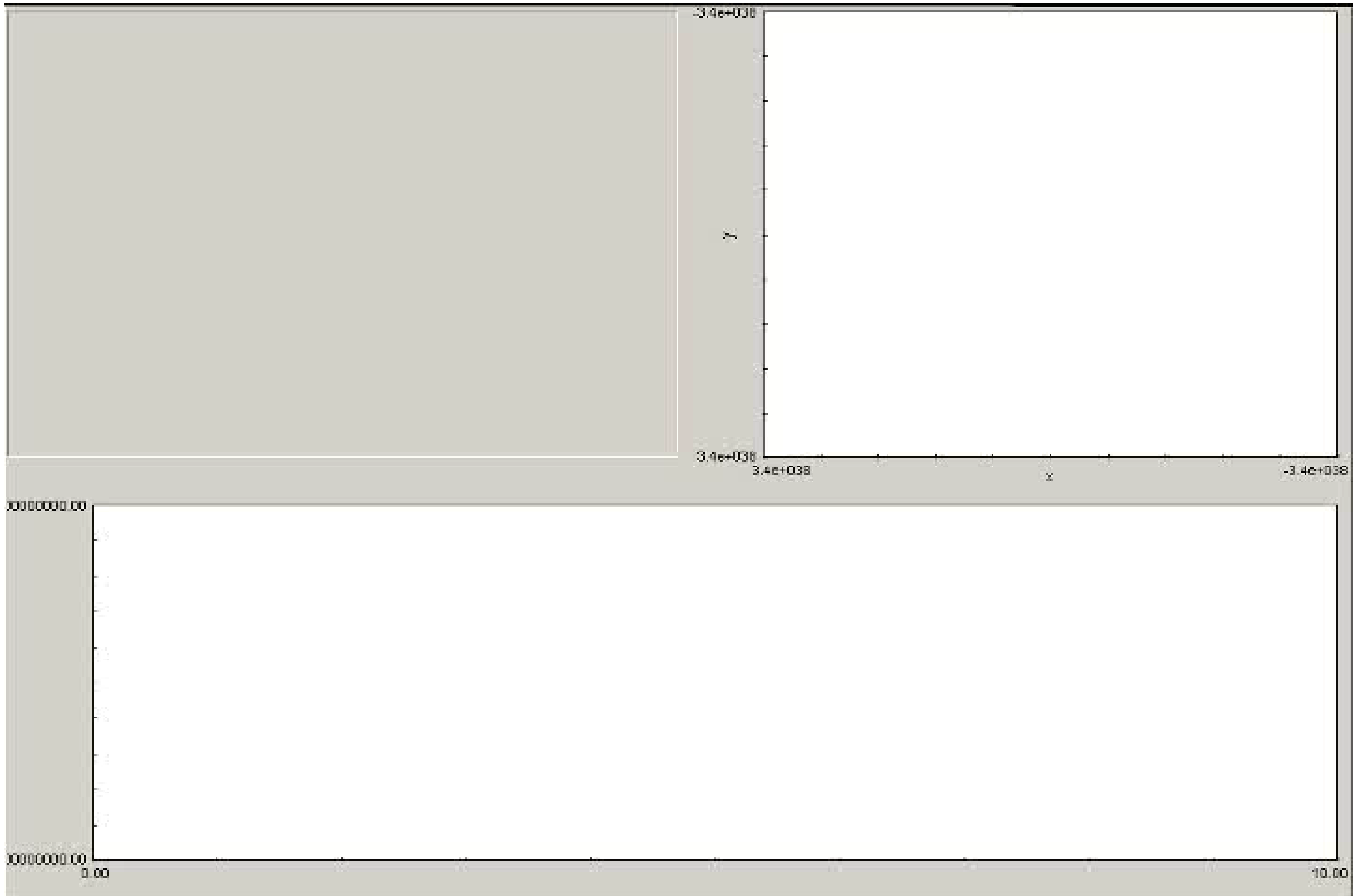


$$x^2 + y^2 - 16 = 0$$

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

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

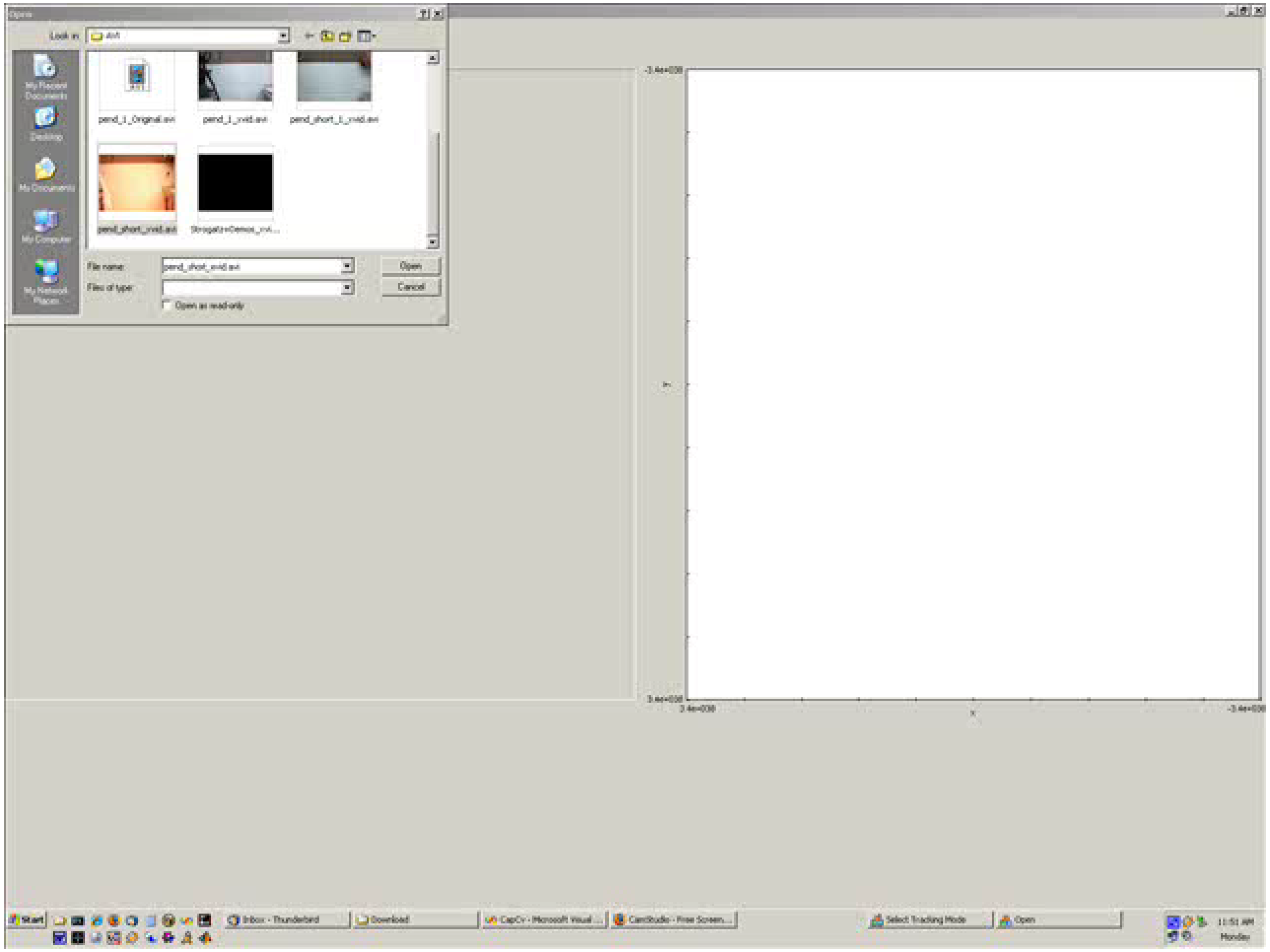

DANGER



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

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

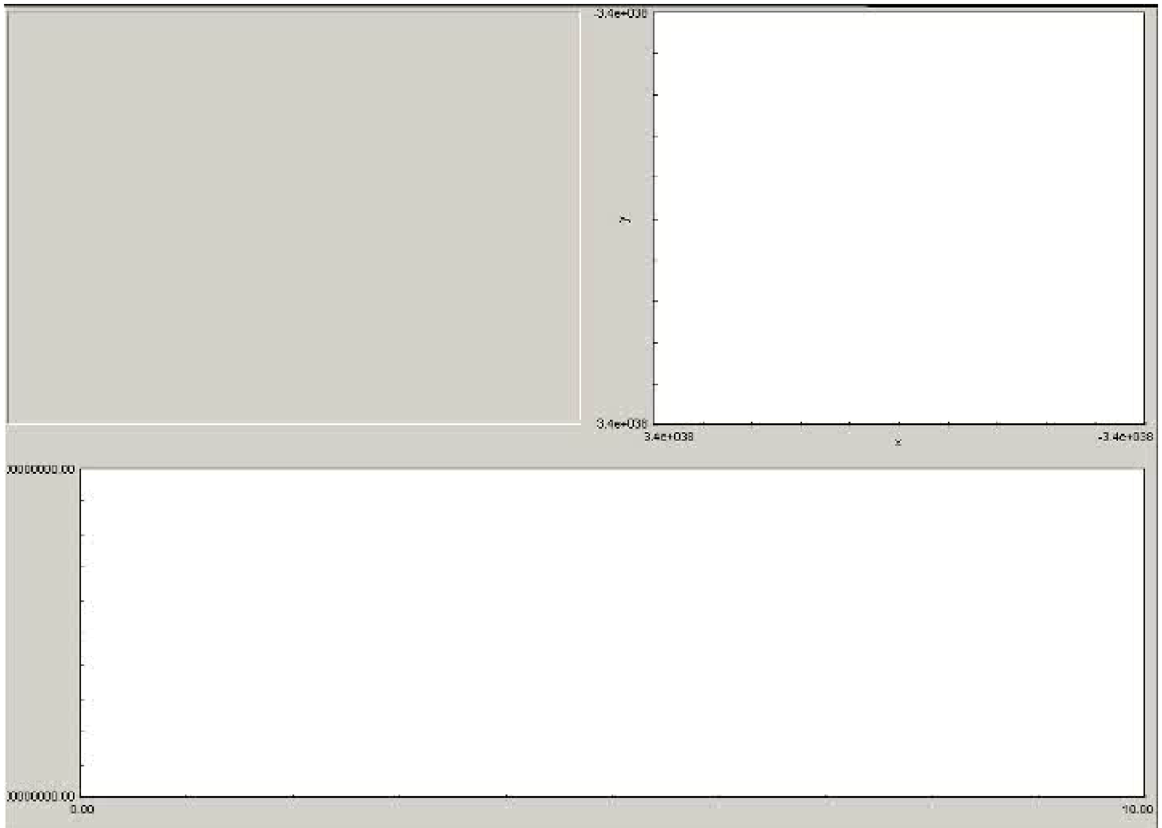
- Coefficients may have different scales and offsets each run



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

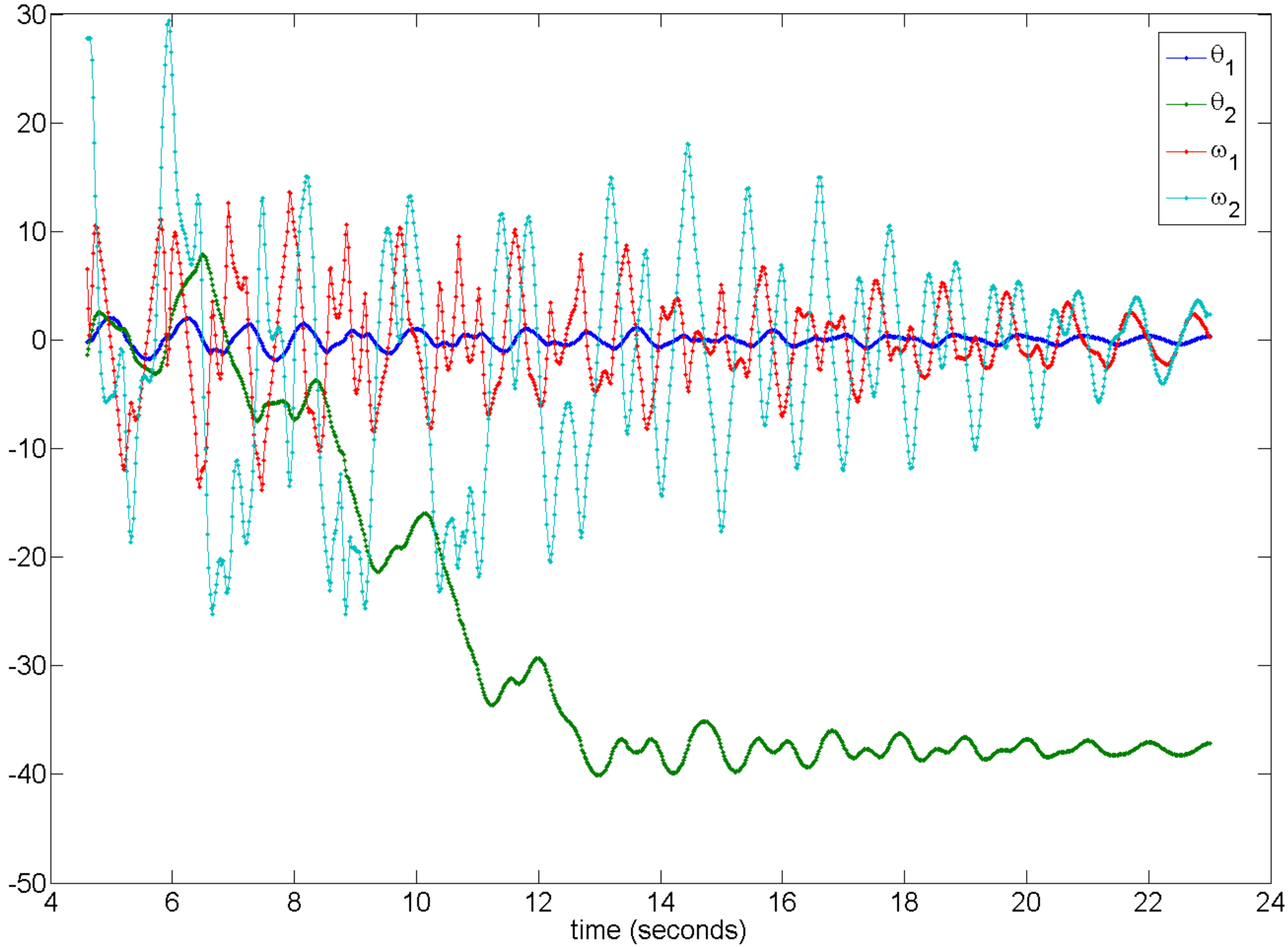
$$\mathbf{L} = 3.52768 * \left(\frac{d\theta}{dt}\right)^2 - 9.43429 * \cos(\theta)$$

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)
θ1
θ2
ω1
ω2

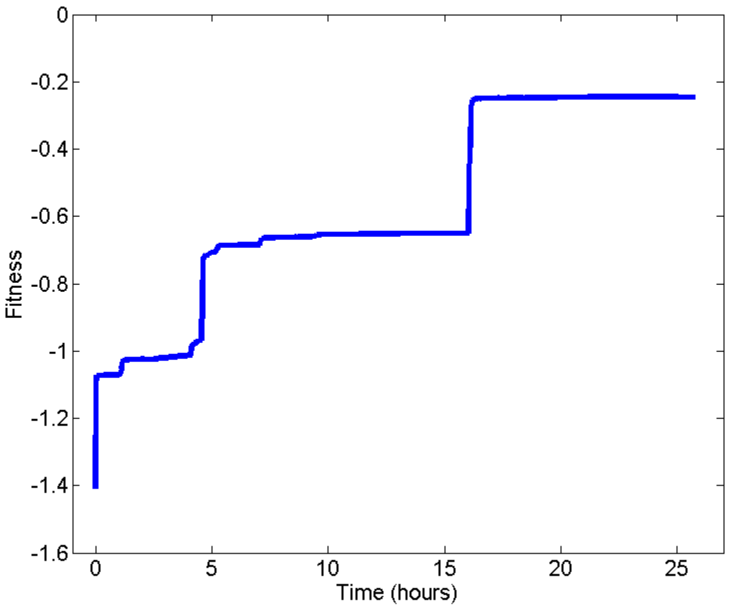
Solution found for the real double pendulum:

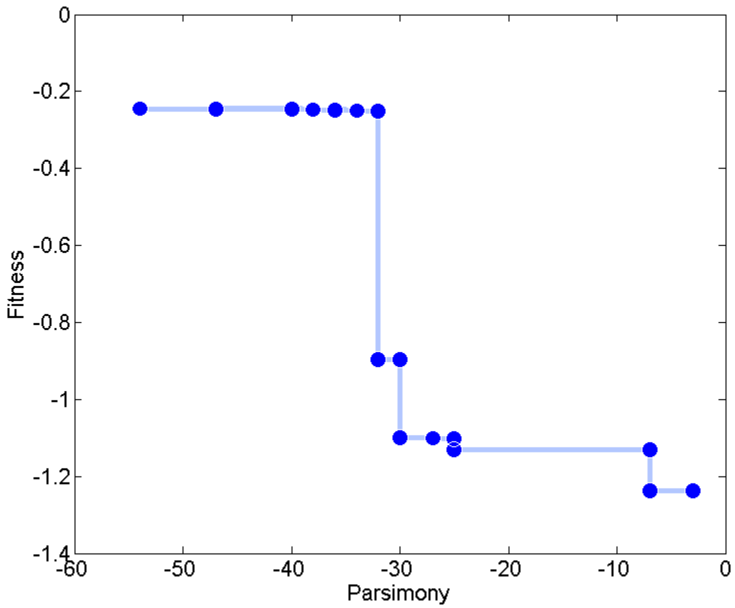
$$L = 434.4822 \cdot \omega_1^2 - 20449.6973 \cdot \cos(\theta_2) + 139.6004 \cdot \omega_2^2 - 53876.0938 \cdot \cos(\theta_1) - 358.1621 \cdot w1 \cdot w2 \cdot \cos(\theta_1 - 1.0001 \cdot \theta_2)$$

$$L = \omega_1^2 + 0.3213 \cdot \omega_2^2 + 0.8243 \cdot \omega_1 \cdot \omega_2 \cdot \cos(\theta_1 - \theta_2) + 124.0007 \cdot \cos(\theta_1) + 47.0668 \cdot \cos(\theta_2)$$

$$L = \omega_1^2 + \alpha \cdot \omega_2^2 + \beta \cdot \omega_1 \cdot \omega_2 \cdot \cos(\theta_1 - \theta_2) + \gamma \cdot \cos(\theta_1) + \lambda \cdot \cos(\theta_2)$$

$$\alpha = 0.3213$$
$$\beta = 0.8243$$
$$\gamma = 124.0007$$
$$\lambda = 47.0668$$





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

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

$k_1/k_1 = 1$

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

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

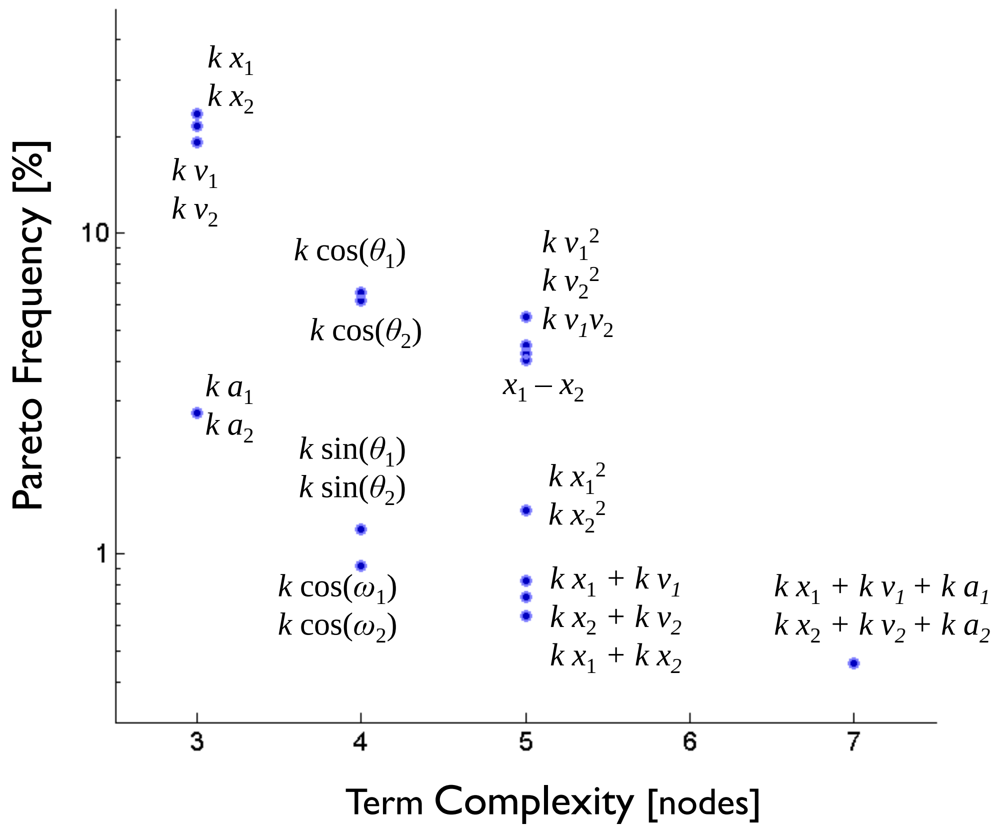
$k_4/k_1 = 19.6/L_1$

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

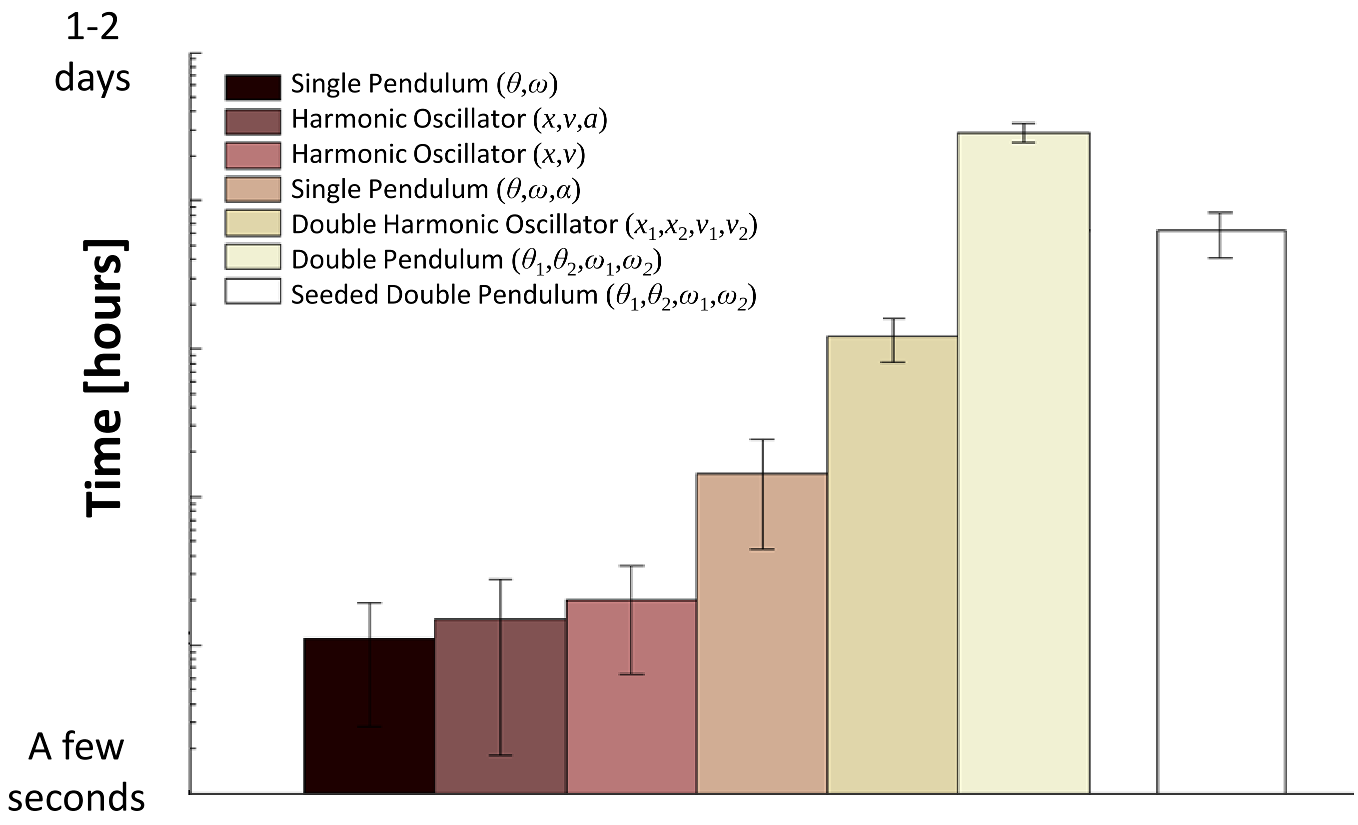
$$L_1^2(m_1+m_2)\omega_1^2 + m_2L_2^2\omega_2^2 + 2\cdot m_2L_1L_2\omega_1\omega_2\cos(\theta_1-\theta_2)$$
$$-19.6\cdot L_1(m_1+m_2)\cos\theta_1 - 19.6 m_2L_2\cos\theta_2$$

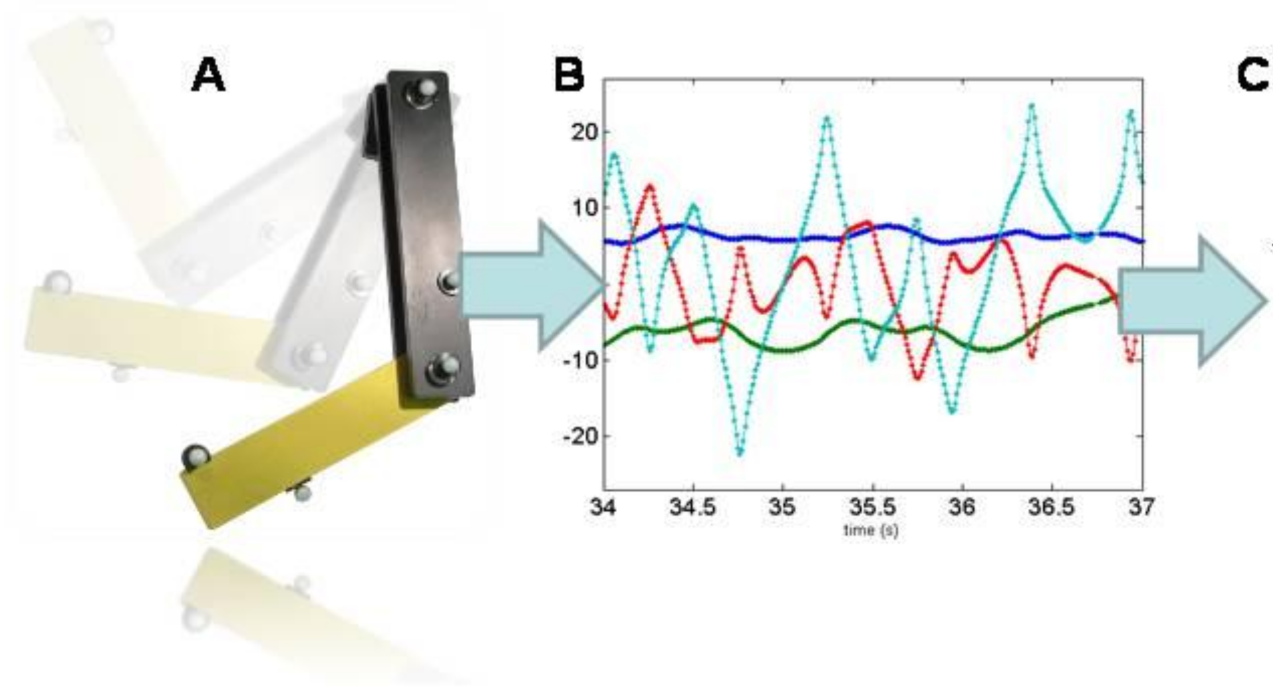
Alphabet

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

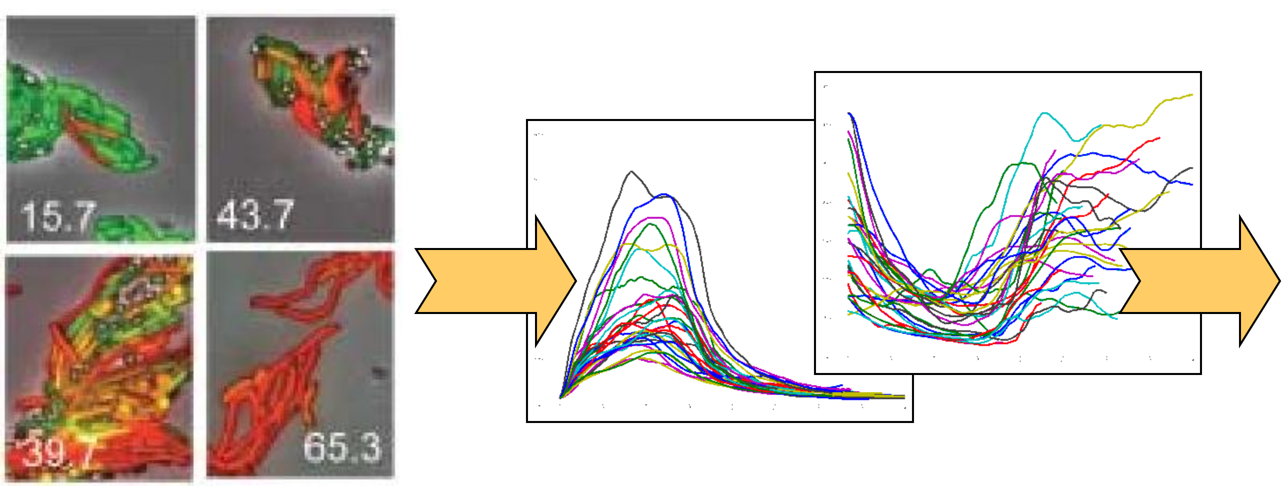




C

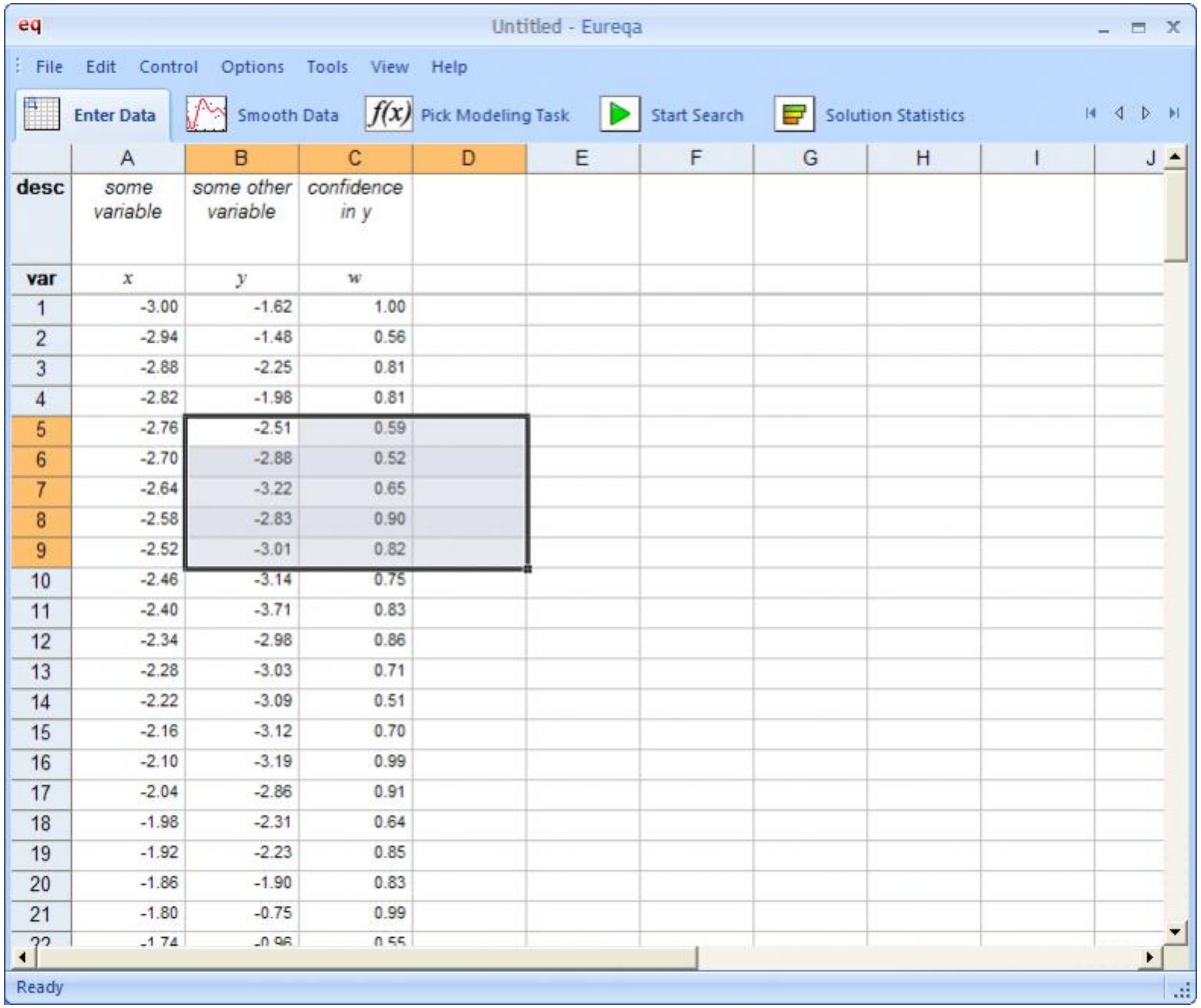
Detected Invariance:

$$L_1^2(m_1+m_2)\omega_1^2 + m_2 L_2^2\omega_2^2 + m_2 L_1 L_2 \omega_1 \omega_2 \cos(\theta_1 - \theta_2) - 19.6 L_1 (m_1+m_2)\cos\theta_1 - 19.6 m_2 L_2 \cos\theta_2$$



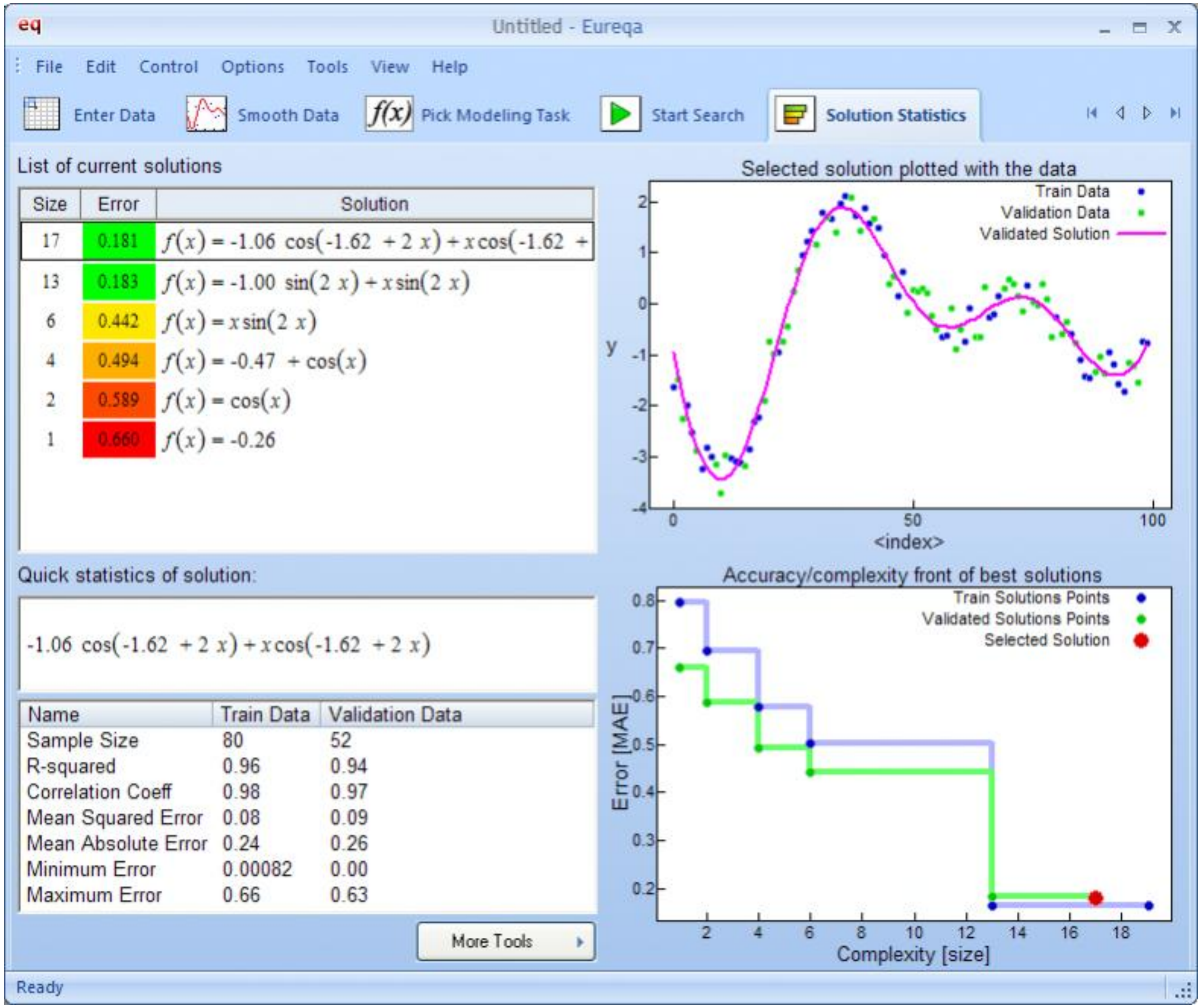
$$\frac{d\mathbf{K}}{dt} = a_K + \frac{b_K + c_K \mathbf{S}_{t-t_1}}{\mathbf{K}_{t-t_2}}$$

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



	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							

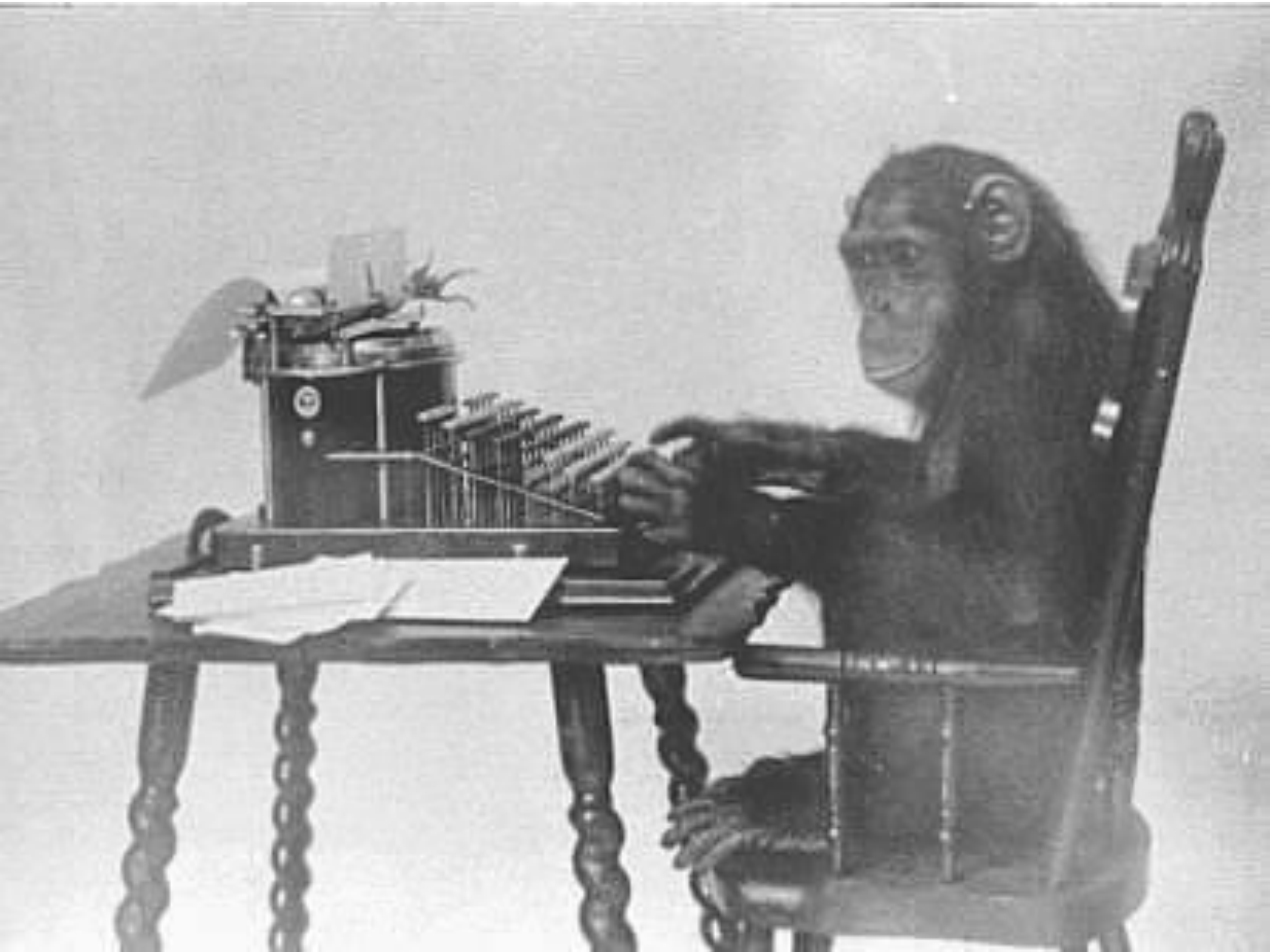
Eureqa



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…



Concluding Remarks

Wired 16.07

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

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

Scalability

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

Conclusions

- **Analysis by Synthesis**
  - explaining an observation by generation
    - E.g. Generate predictive model of self
  - Good for prediction, understanding, classification, simulation, control
- **Co-evolution: Active learning**
  - “intelligent” probing to extract invariants
- **Challenges**
  - Models without meaning
  - A new frontier of “Machine Teaching”

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 (4th order)*