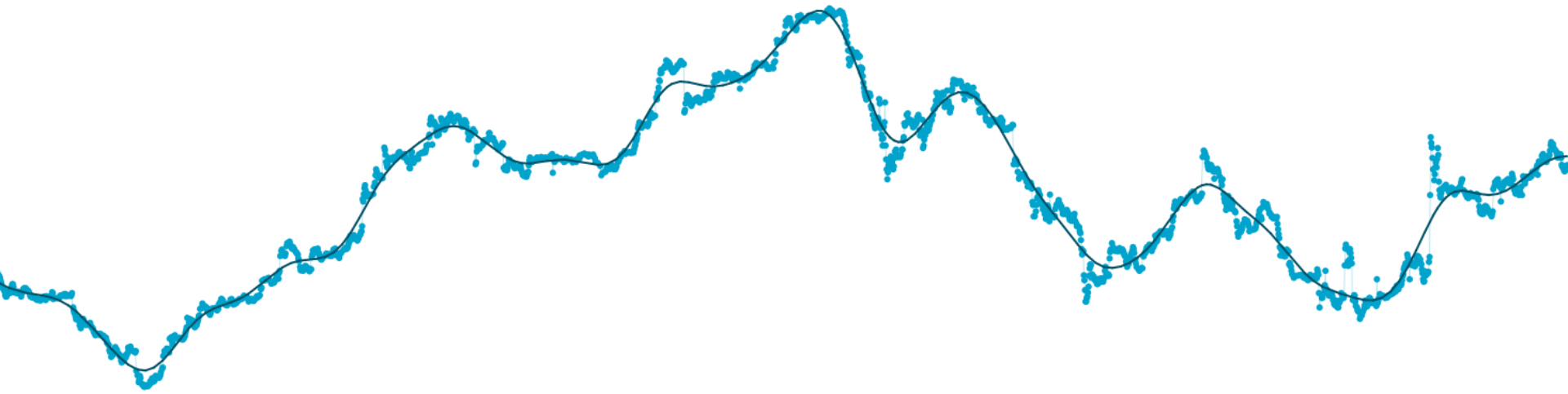
Detecting Hidden Mathematical Patterns in Raw Data

Michael Schmidt, Ph.D.



Eureqa

by nutonian

March 2, 2015

Conclusion

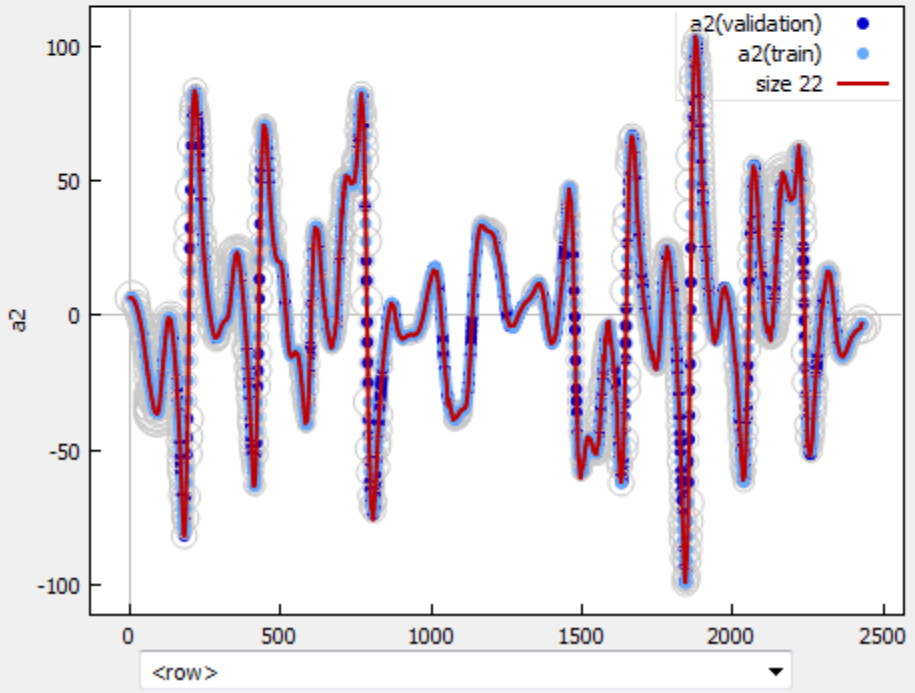
- Virtually unlimited computational power to search for patterns today
- Pattern complexity is critical for accuracy, generality, and interpretability
- Fallacy of big data – the right transformations of a limited set of signals is what actually matters
- Patterns don’t last forever (find them fast enough to act)
- Use-case: from hundreds of trade strategies, determine which are most reliable given current information

$$a_2 = v_1^{\,2}\sin(x_1 - x_2) - a_1\cos(x_2 - x_1) - 9.8\sin(x_2)$$

Best Solutions of Different Sizes:

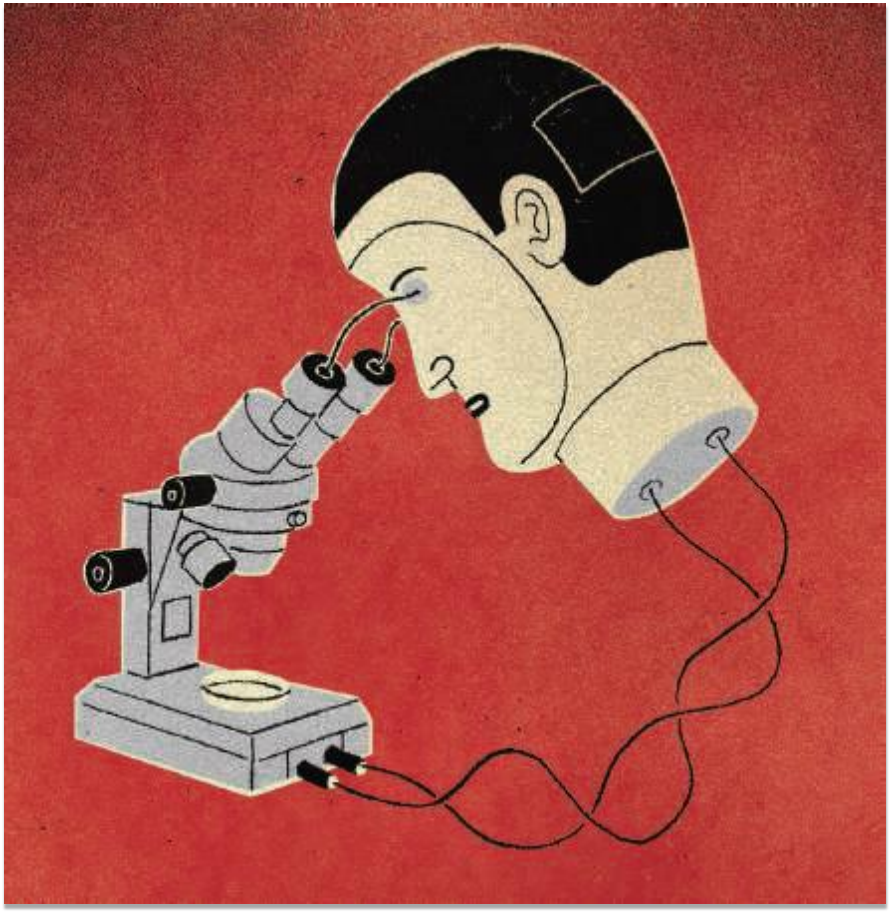
Size	Fit	Solution
13	0.197	$a_2 = -20\sin(x_2) - 1.2\,a_1\cos(x_1 - x_2)$
11	0.350	$a_2 = -(a_1\cos(x_1 - x_2)) - 0.31\,a_1$
10	0.363	$a_2 = -2.4 - v_1^{\,2}\sin(x_2 - x_1)$
9	0.400	$a_2 = 1.2\,a_1\sin(4.6 + x_2 - x_1)$
8	0.402	$a_2 = a_1\sin(4.6 + x_2 - x_1)$
7	0.484	$a_2 = a_1\sin(0.01\,v_1\,v_2)$
6	0.485	$a_2 = 0.018\,v_1\,v_2\,a_1$
5	0.492	$a_2 = 0.01\,v_1\,v_2\,a_1$
2	0.591	$a_2 = -0.86\,a_1$
1	0.999	$a_2 = 10\,v_1$

Solution Fit Plot (default)



“Computer Program Discovers Laws of Physics”

–New York Times



Eureqa in *Nature News*

MIT Technology Review

Forbes

Discovery CHANNEL

npr

WIRED

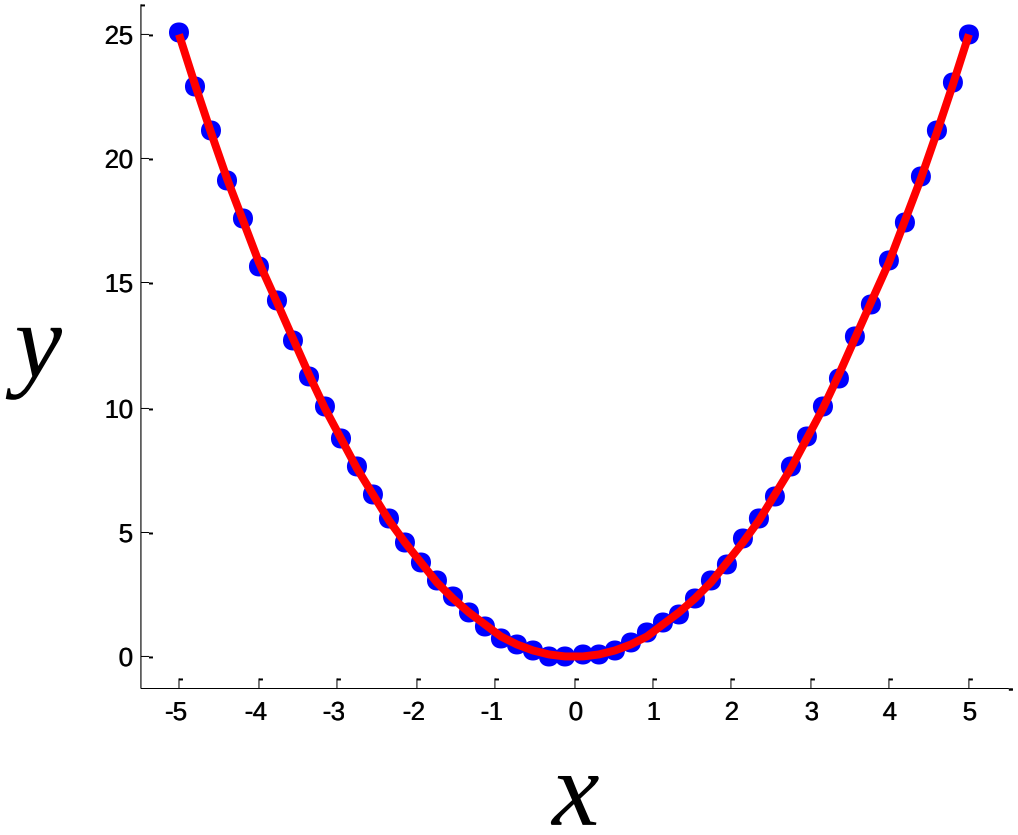
2013 The Boston Globe GLOBE

The New York Times

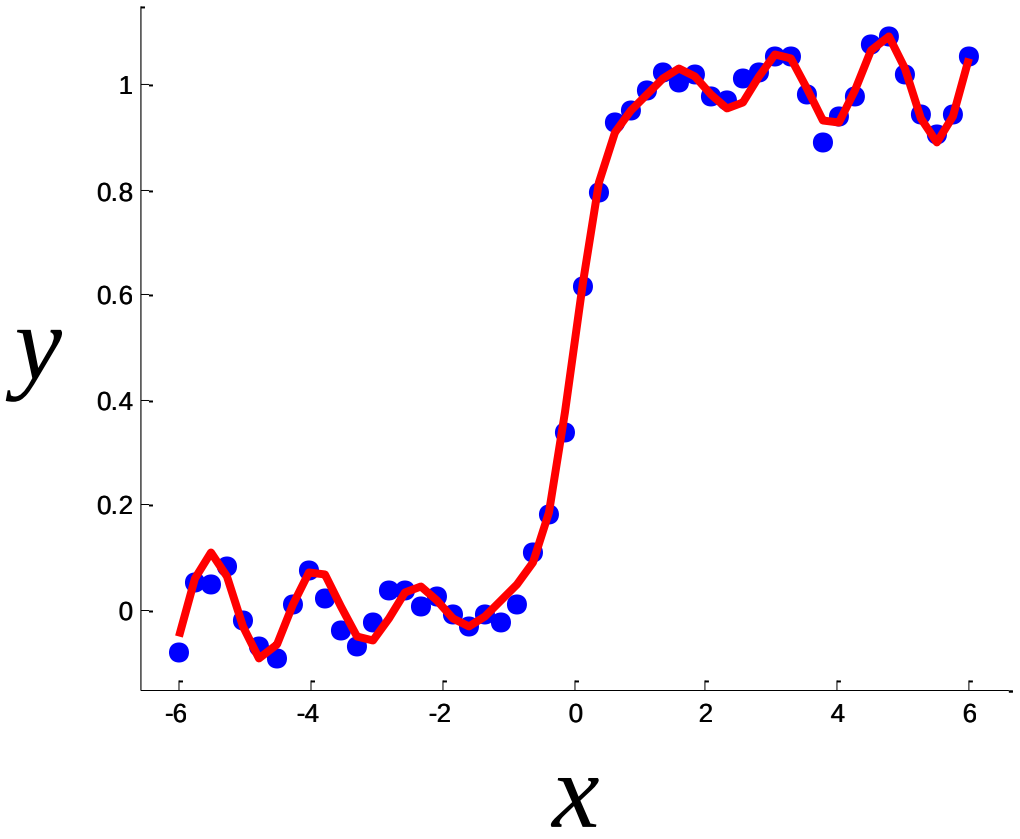
Computing Functional Relationships From Data

Video link

$$y = x^2$$

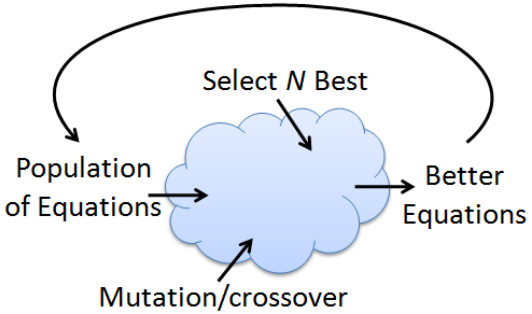


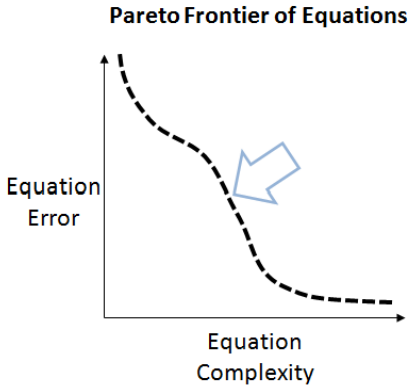
$$y = 0.02\, x \cos(4\, x) + 1/(1 + \exp(-4\, x))$$

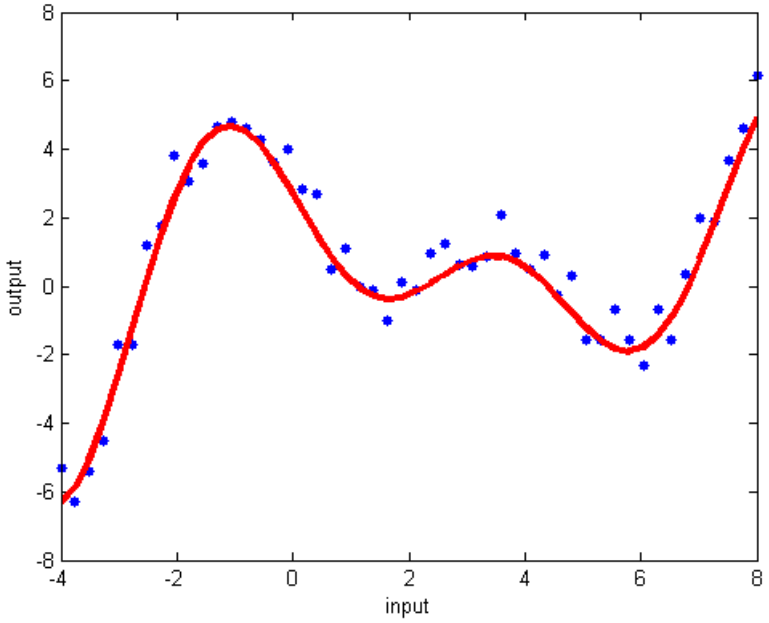


Evolutionary Search for Structure/Transformations

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





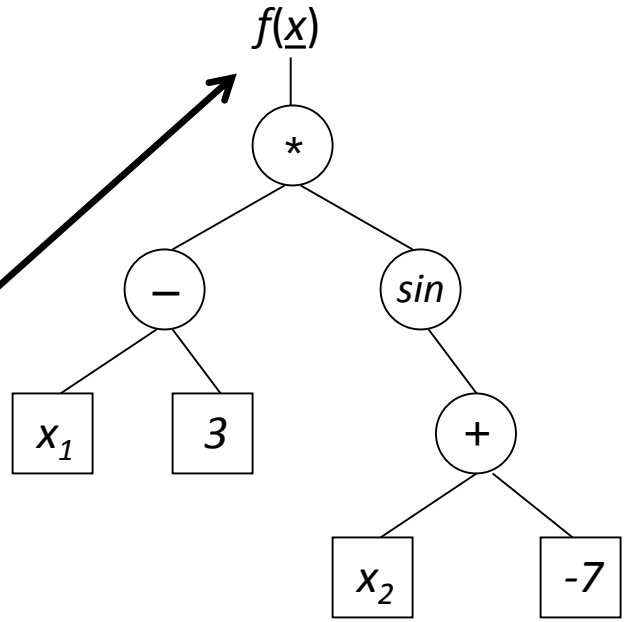


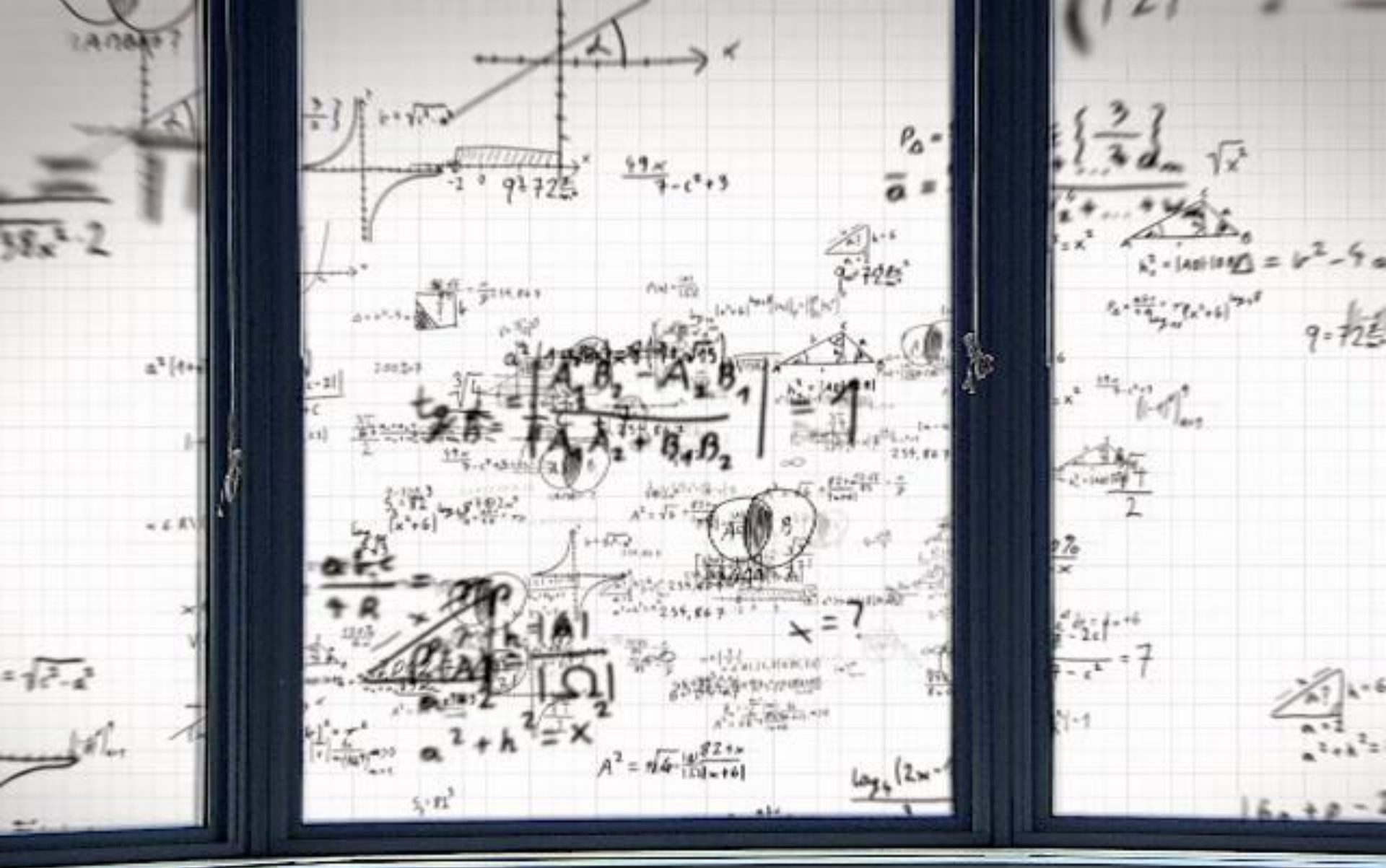
$\sin(x_2)$

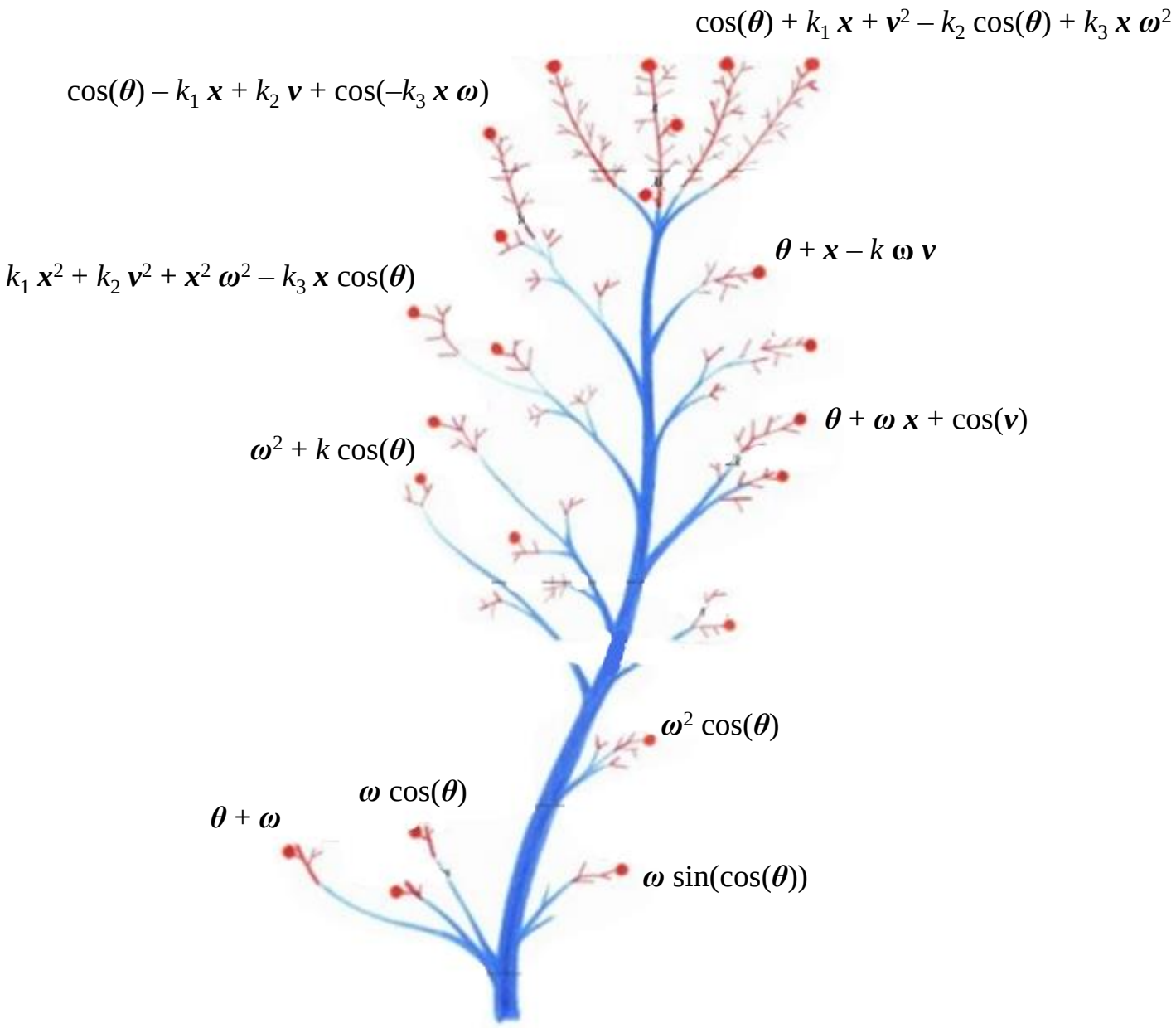
$x_1 * \sin(x_2)$

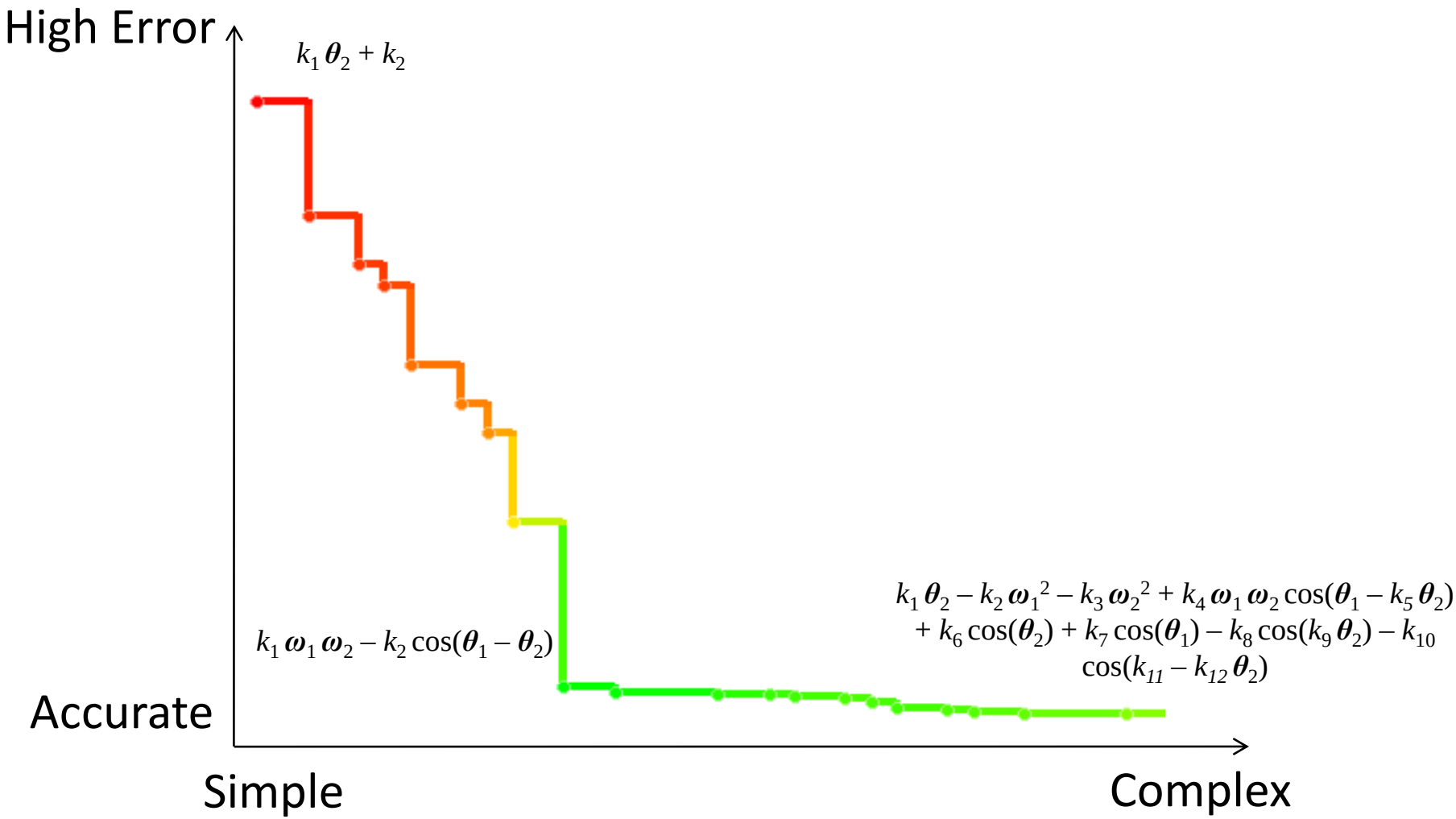
$(x_1 - 3) * \sin(x_2)$

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






$\cos(\boldsymbol{\theta}) + k_1\,\boldsymbol{x} + \boldsymbol{v}^2 - k_2\cos(\boldsymbol{\theta}) + k_3\,\boldsymbol{x}\,\boldsymbol{\omega}^2$
$\cos(\boldsymbol{\theta}) - k_1\,\boldsymbol{x} + k_2\,\boldsymbol{v} + \cos(-k_3\,\boldsymbol{x}\,\boldsymbol{\omega})$
$\boldsymbol{\theta} + \boldsymbol{x} - k\,\boldsymbol{\omega}\,\boldsymbol{v}$
$k_1\,\boldsymbol{x}^2 + k_2\,\boldsymbol{v}^2 + \boldsymbol{x}^2\,\boldsymbol{\omega}^2 - k_3\,\boldsymbol{x}\cos(\boldsymbol{\theta})$
$\boldsymbol{\theta} + \boldsymbol{\omega}\,\boldsymbol{x} + \cos(\boldsymbol{v})$
$\boldsymbol{\omega}^2 + k\cos(\boldsymbol{\theta})$
$\boldsymbol{\omega}^2\cos(\boldsymbol{\theta})$
$\boldsymbol{\omega}\cos(\boldsymbol{\theta})$
$\boldsymbol{\theta} + \boldsymbol{\omega}$
$\boldsymbol{\omega}\sin(\cos(\boldsymbol{\theta}))$

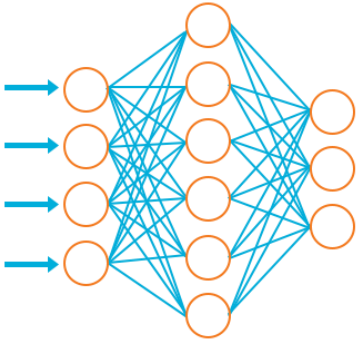

High Error
$k_1\,\theta_2 + k_2$
$k_1\,\omega_1\,\omega_2 - k_2\cos(\theta_1 - \theta_2)$
$k_1\,\theta_2 - k_2\,{\omega_1}^2 - k_3\,{\omega_2}^2 + k_4\,\omega_1\,\omega_2\cos(\theta_1 - k_5\,\theta_2) + k_6\cos(\theta_2) + k_7\cos(\theta_1) - k_8\cos(k_9\,\theta_2) - k_{10}\cos(k_{11} - k_{12}\,\theta_2)$
Accurate
Simple
Complex

Data Science Today

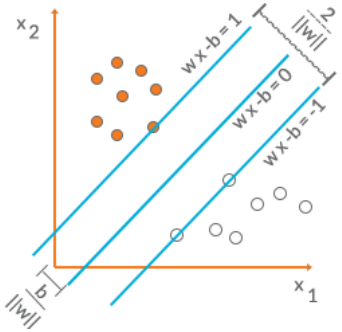
- Too Technically Focused
- Limited Human Capital and Data Science Talent

Difficult to grow analytic capacity

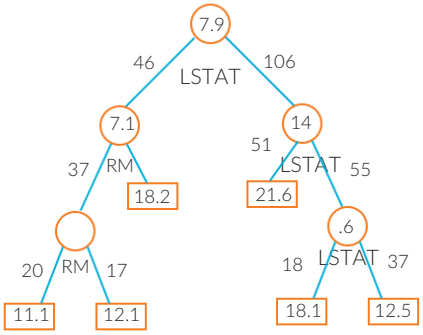
Many Methods, Where to Begin?



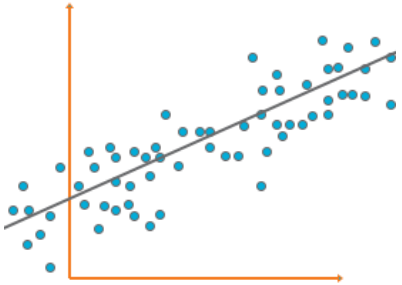
Output:
Linear Node 0
w0 = 0.0466…
w1 = 0.3708…
w2 = 0.1197…



Output:
w0 = +0.0466 (symboling)
w1 = -0.3708 (loss)
w2 = +0.1197 (fuel-type)
w3 = +0.0099 (num doors)

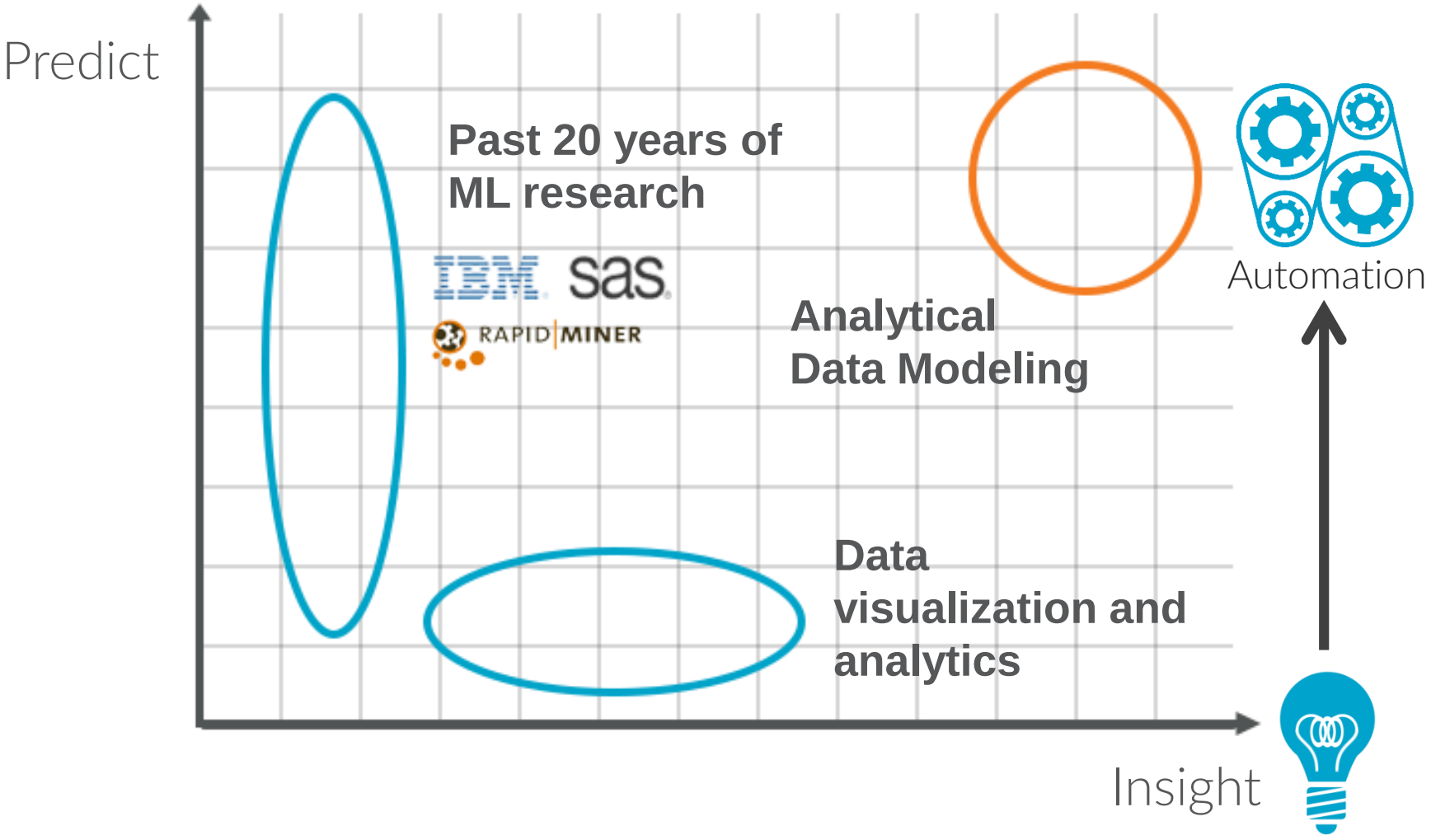


Output:
example < 71.5
| example <60 : 47.3
| example >=60
| | (widgets) > .5
| | | price < 8188.5
| | | | price < 5271
| | | |



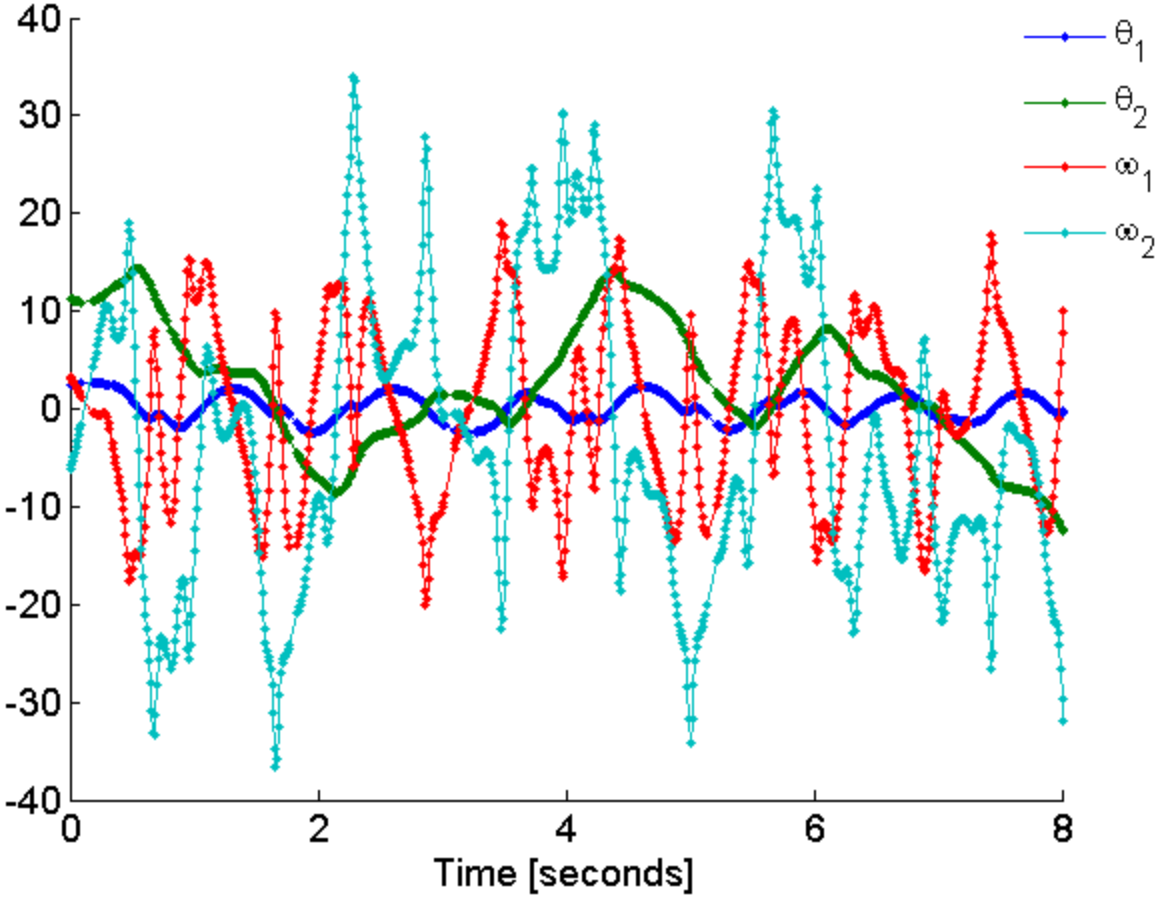
Output:
Y = 1.8(fuel-gas)+1.8(fuel-diesel)
-1.332(aspiration)- …

Data Science Today



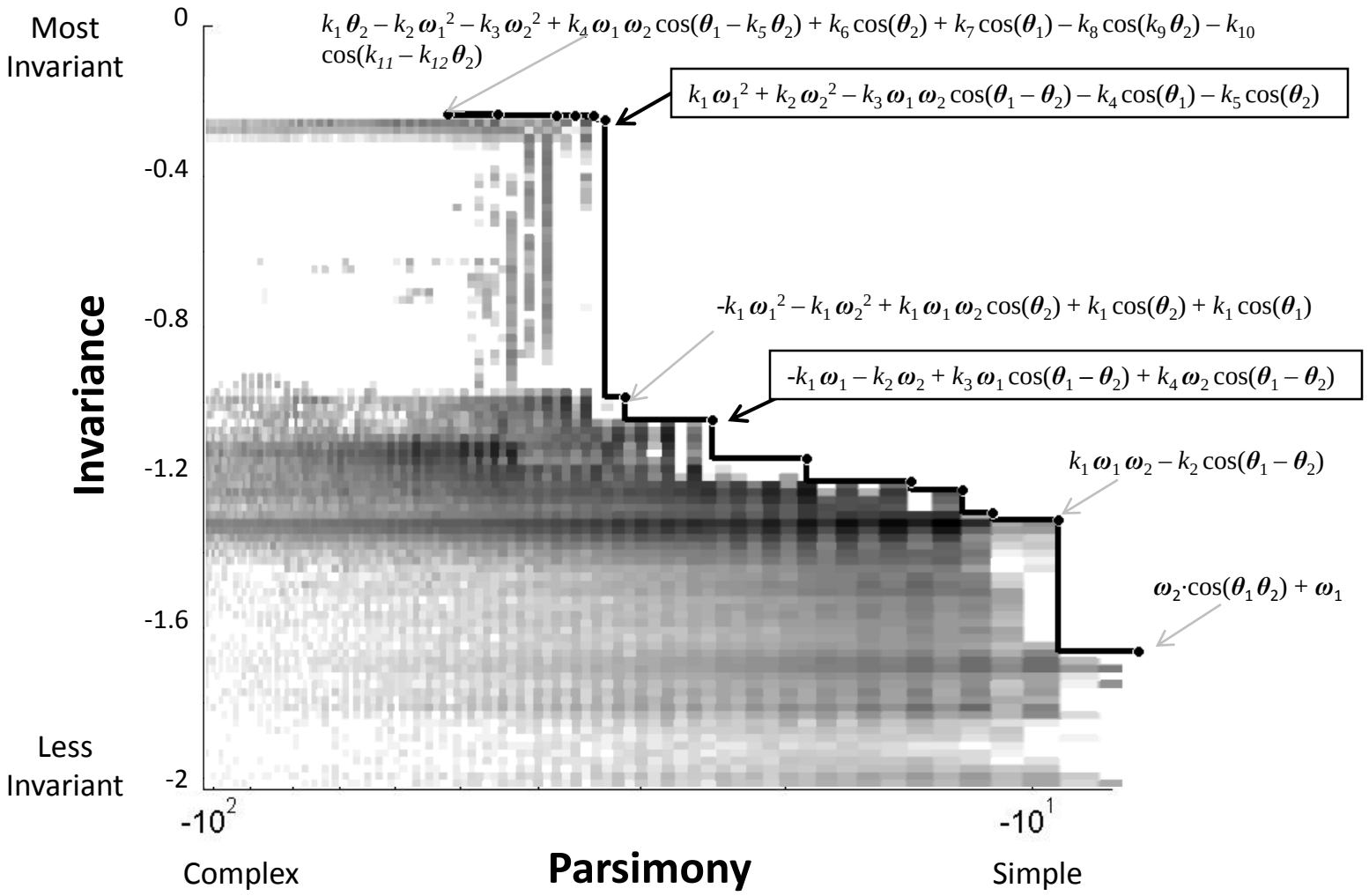
Conservation laws → Deeper insight

Signal from Noise

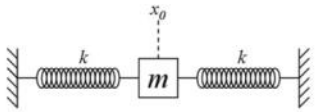
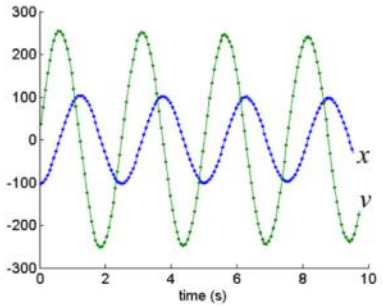
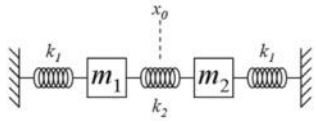
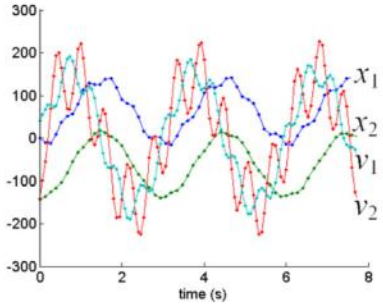
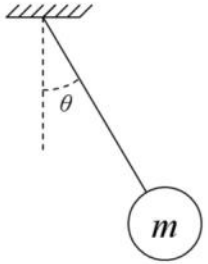
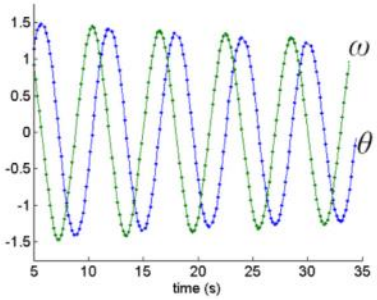
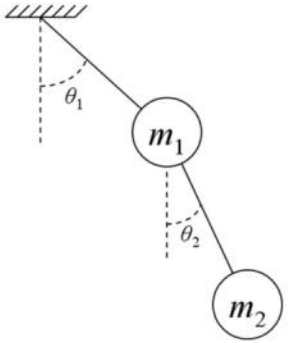
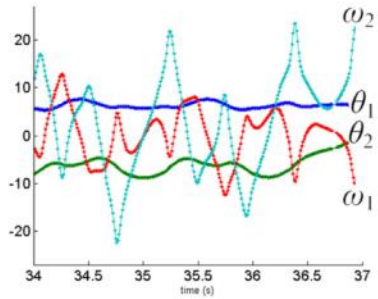


Schmidt M., Lipson H. (2009) "Distilling Free-Form Natural Laws from Experimental Data," Science, Vol. 324, no. 5923, pp. 81 - 85.

Visualizing the Search

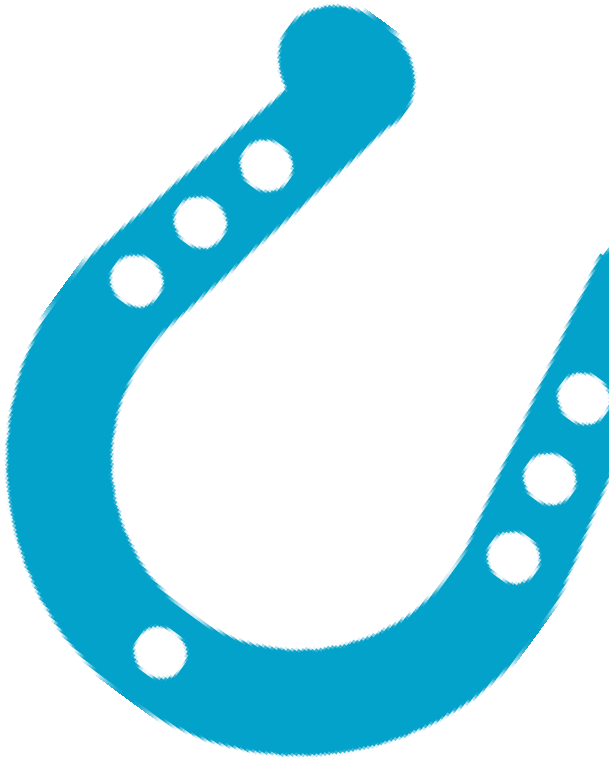


Schmidt M., Lipson H. (2009) "Distilling Free-Form Natural Laws from Experimental Data," Science, Vol. 324, no. 5923, pp. 81 - 85.

Physical System	Schematic	Experimental Data	Inferred Laws
			$114.28v^2 + 692.32x^2$ **Hamiltonian** $v^2 - 6.04x^2$ **Lagrangian** $a - 0.008v - 6.02x$ **Equation of motion**
			$-142.19x_1 - 74.65x_2 + 0.12{x_1}^2 - 1.89x_1x_2 - 1.51{x_2}^2 - 0.49{v_2}^2 + 0.41v_1v_2 - 0.082{v_1}^2$ **Lagrangian**
			$1.37\cdot\omega^2 + 3.29\cdot\cos(\theta)$ **Lagrangian** $2.71\alpha + 0.054\omega - 3.54\sin(\theta)$ **Equation of motion** $(x - 77.72)^2 + (y - 106.48)^2$ **Circular manifold**
			${\omega_1}^2 + 0.32{\omega_2}^2 - 124.13\cos(\theta_1) - 46.82\cos(\theta_2) + 0.82\omega_1\omega_2\cos(\theta_1 - \theta_2)$ **Hamiltonian**

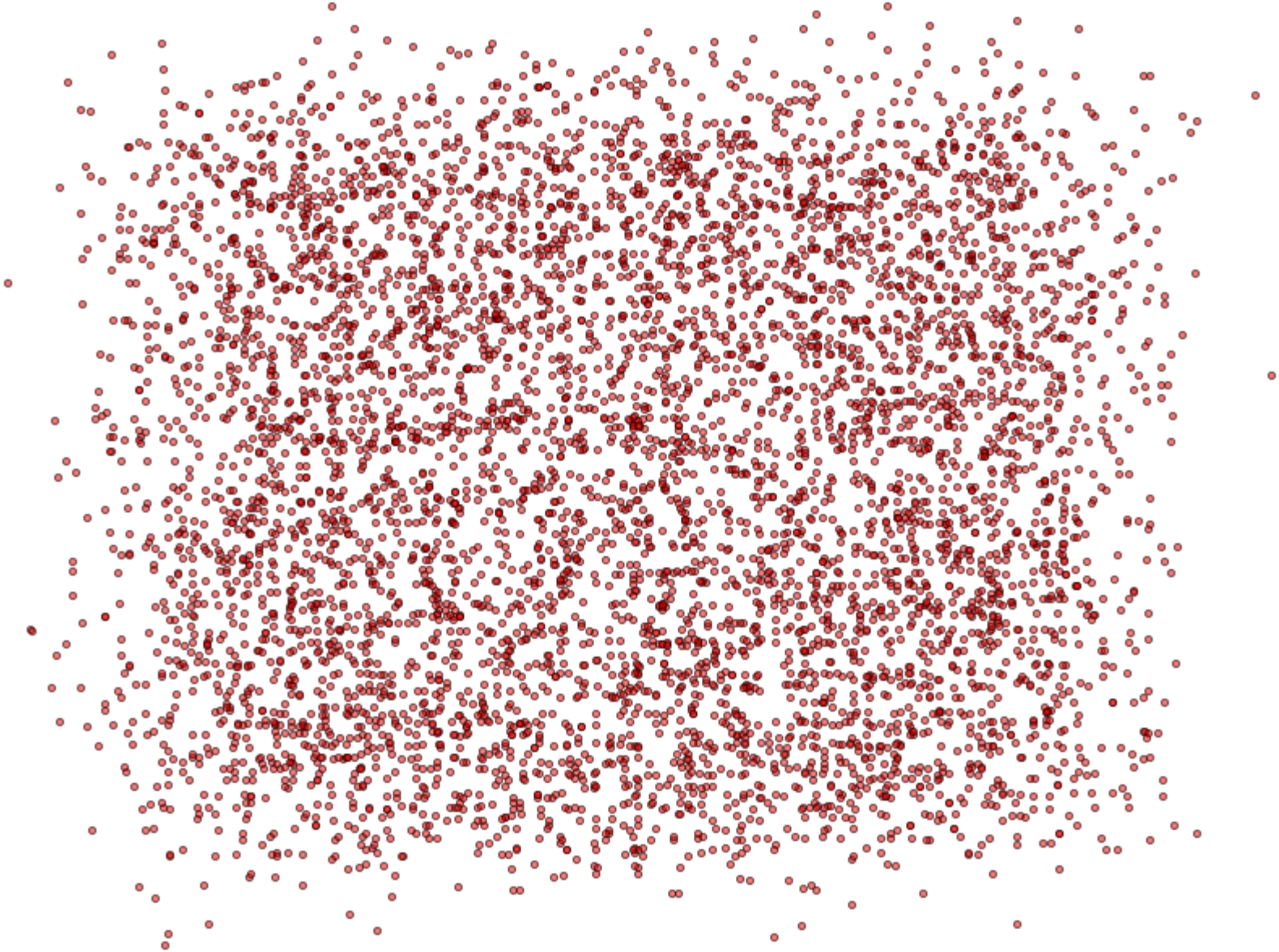
Applying Moneyball Methods

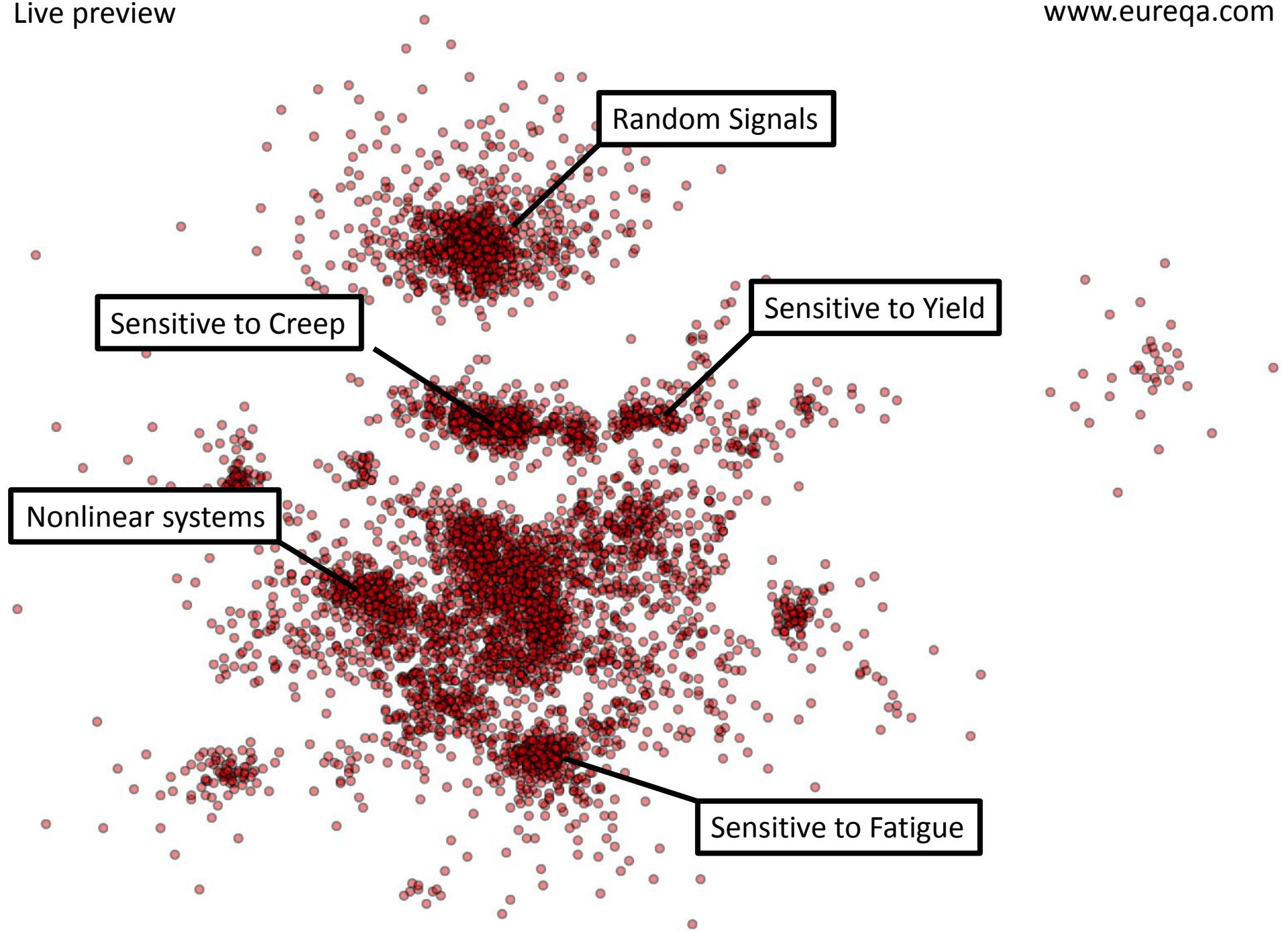
- Goals
  - Predict finish positions of the 2014 Kentucky Derby
- Data Sources
  - Brisnet.com handicapping data
  - Race past performance
- Nutonian Difference
  - Expose relationships between running style, speed, and trainer record
  - Easily explained model
- Impact
  - Predict 4 out of top 5 horses
  - Exact predictions for #1, #3, and #5



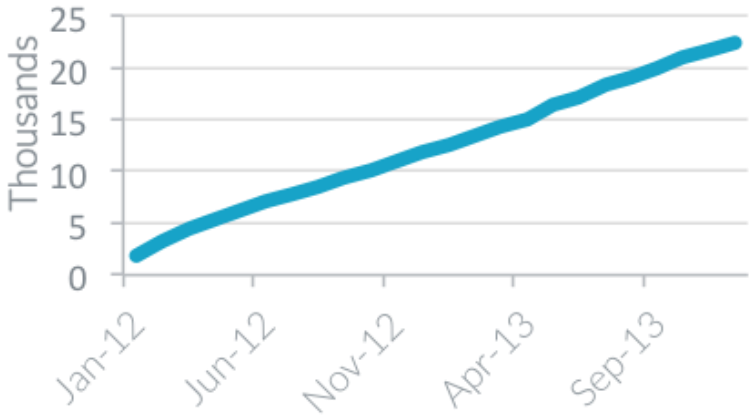
Data Scientist API / Community

- Publish algorithms for internal / external consumption
- Seamlessly integrate business and data science communities
- Scale access to rapidly advancing ecosystem
- Community rankings / recognition



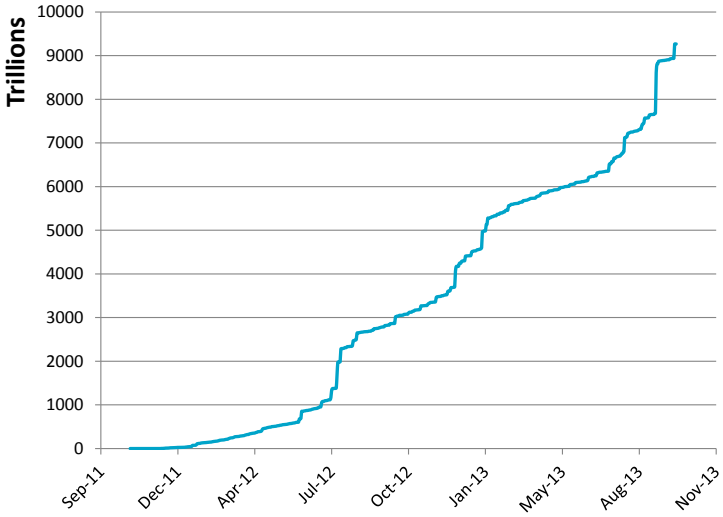

Random Signals
Sensitive to Creep
Sensitive to Yield
Nonlinear systems
Sensitive to Fatigue

Eureqa in Action



80% increase in Eureqa users over 2013

100% increase in forum users since August



Average growth of 659 Million formulas evaluated per second

156% increase in Eureqa searches over 2013

Try it on your data

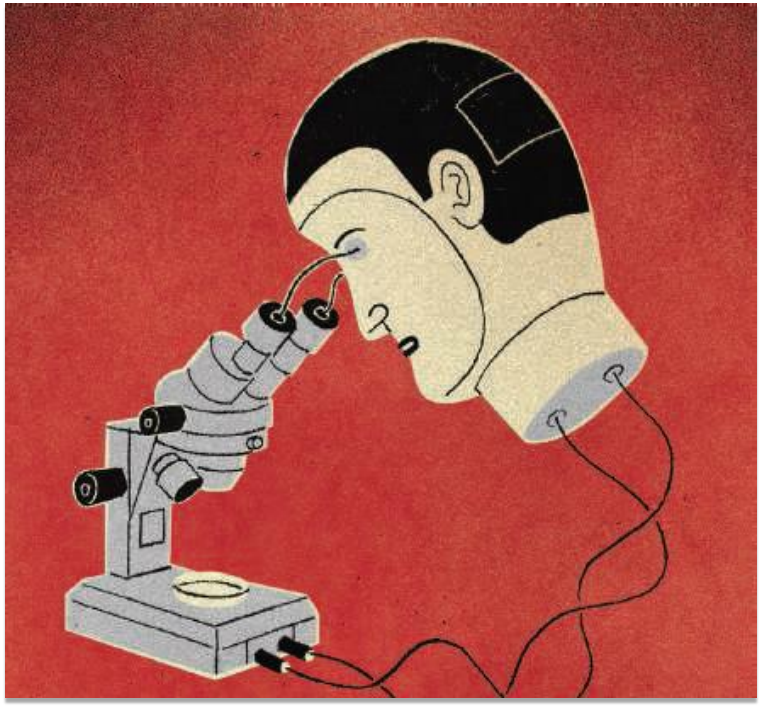
Eureqa

www.nutonian.com/download

Free on small data | Test on unlimited data

Michael Schmidt
michael@nutonian.com

nutonian



Blog: http://blog.nutonian.com

Twitter: @Nutonian