Computer Discovery

machine learning in scientific domains

Inductive Bias

- Any restriction on the universe, or any "internal" ways to choose between alternatives equally called for by the environment is "inductive bias"
- Conundrum:
  - Bias is necessary for learning
  - Bias limits possibilities
- Occam's Meta Strategy
  - more compact descriptions are better.

Easy versus Hard Inductive Generalization

- Easy means
  - Finite sample
  - Finite Test
  - Given classes to find
- Hard means
  - Finite sample
  - *Infinite* test or no classes.

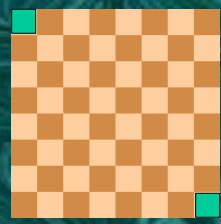
this is Chomsky's problem, sample finite for infinite generative capacity.

Supervised vs Unsupervised Learning

- Classification is GIVEN categories as input.
  - What if you only have samples, but no categorization?
  - The program has to discover a set of classes and how to divide inputs into different classes.

Where are Hard ML problems?

- Mathematical Induction and Discovery
  - What is a number?
    - the concept, not the 32 bits!
  - What is infinity?
    - Why does pi have infinite digits?
- Metacognition
  - The Mutilated Checkerboard problem
- *Scientific Discovery*
  - Observe some behavior
  - Induce an explanation of the behavior



Discovery Problems

- Theory from data
- Finding Categories or taxonomies
- Induction of Laws
- Explanation
- Prediction
- Creative discovery

How can Machine Learning help?

- Scientific theories are a form of knowledge which is acquired
- Lots of data can be reduced to simple rules, as in decision-tree learning and this is often the goal of scientific analysis
  - However, the results of scientific discovery is not in decision-tree form!
- Discovery of clusters and taxonomies using contrasting information
- Induction of quantitative theories (equations)
- Drug Discovery (profitable branch of AI)

Three Famous AI Programs

- BACON
  - Heuristic construction of equations from data
  - named after Sir Francis Bacon
- Automated Mathematician (AM)
  - heuristic search of number theory
- Eureqa
  - extraction of equations from massive data

BACON (Langley & Simon)

- Subtask of science: Fitting Theories to Data
  - Equation Creation
  - Curve Fitting
    - But not a parameterized universal method like polynomials

Bacon

- Input: Tables of experimental data
  - masses, forces, accelerations
  - length, height, area of rectangles
  - distance and period of Planets
  - Volts, Ohms, Current
  - Pressure, Temperature
- Output:
  - Equations of the "laws" which are holding

How can a program do this?

- Searching among functional relationships using heuristics to find invariant relationships
  - If X and Y are linearly related with slope S and intersect I, hypothesize a linear relation
  - (xy>0) If X increases as Y decreases and X and Y are not linearly related then define new term as product of X and Y
  - (xy > 0) IF X increases as Y increases, define a new term as ratio of X and Y
  - (XY<0) If X increases as Y increases, look at ratio
  - (XY<0) if X and Y Increase, look at product

Bacon Example

- Both are going up, nonlinearly, look at ratio…

	distance	period
Mercury	0.382	0.241
Venus	0.724	0.616
Earth	1	1
Mars	1.524	1.881
Jupiter	5.199	11.855
Saturn	9.539	29.459

Bacon Example

- Now, d/p goes down as D goes up…

	distance	period	d/p
Mercury	0.382	0.241	1.607
Venus	0.724	0.616	1.175
Earth	1	1	1
Mars	1.524	1.881	0.81
Jupiter	5.199	11.855	0.439
Saturn	9.539	29.459	0.324

Bacon Example

- D/P and DD/P move in opposite directions…

	distance	period	d/p	dd/p
Mercury	0.382	0.241	1.607	0.622
Venus	0.724	0.616	1.175	0.851
Earth	1	1	1	1
Mars	1.524	1.881	0.81	1.234
Jupiter	5.199	11.855	0.439	2.28
Saturn	9.539	29.459	0.324	3.088

Bacon Example

- We have discovered Kepler's Third Law!

	distance	period	d/p	dd/p	ddd/pp
Mercury	0.382	0.241	1.607	0.622	1
Venus	0.724	0.616	1.175	0.851	1
Earth	1	1	1	1	1
Mars	1.524	1.881	0.81	1.234	1
Jupiter	5.199	11.855	0.439	2.28	1
Saturn	9.539	29.459	0.324	3.088	1

Laws that BACON has (re)discovered

- BACON.1: Kepler's law
  D3/P2=c
- BACON.2: Ideal gas law
  PV=aNT+bN
- Coulomb's law
  FD2/Q1Q2=c

AM (Lenat 1977)

- Represented mathematical concepts in a frame systems
- Used 250 hand-made Heuristics
- Mutated small lisp programs generating sets of numbers
  - select most *interesting* concept and generate examples
  - Look for regularities and create conjectures
  - propagate through existing knowledge
- Maintained an agenda of "interestingness"
- "Discovered" many concepts in number theory

AM Implementation Concepts

- Agenda (what to try next)
- Interestingness Heuristics
- Concept Representation

AM Representation like the scripts used with CD representations.

- NAME: Prime
- ISA: set
- DEFN: Prime (x) if Z|x then z is element of {1,x}
- SPECIALIZATIONS: odd-primes
- GENERALIZATIONS: NUMBERS
- EXAMPLES
- INTERESTINGNESS: 100
- ORIGIN: 11-10-92 10:00
- SEE-ALSO: divisors-of

AM Heuristics

- A set of rules which helped to generate new concepts and to direct the search to more interesting paths
- Consider Extreme elements
  - Factor sets -> prime numbers, Squares, Maximally divisable
- Look at intersections of concepts
- Generalize and specialize concepts

Results

- Discovered "Prime numbers"
- Discovered various mathematical conjectures
- ultimately petered out into uselessness
- What is problem?
  - Heuristics need to be improved
  - bias of mathematics built into the programming language
  - Lenat knew what he wanted to happen

Inconclusivities…

- Bacon series worked on inducing other laws, but started with "nice" data : Who found the data?
- AM was revealed as a bit of a fraud
  - The "micro-lenat" measure of Bogosity!!!
  - Code was never released
  - AM was "discovering" mathematical elements intentionally placed into LISP by its designers
- Famous for creative naming, Lenat's next system was called Eurisko…

Fast Forward

- Nutonian's Eureqa
- Takes large amorphous scientific and engineering data and
- (using evolutionary algorithm)
- extracts compact equations.

SHARE

REPORT

Distilling Free-Form Natural Laws from Experimental Data

Michael Schmidt[1], Hod Lipson[2,3,*]
+ See all authors and affiliations

Science 03 Apr 2009:
Vol. 324, Issue 5923, pp. 81-85
DOI: 10.1126/science.1165893

Article | Figures & Data | Info & Metrics | eLetters | PDF

Abstract

For centuries, scientists have attempted to identify and document analytical laws that underlie physical phenomena in nature. Despite the prevalence of computing power, the process of finding natural laws and their corresponding equations has resisted automation. A key challenge to finding analytic relations automatically is defining algorithmically what makes a correlation in observed data important and insightful. We propose a principle for the identification of nontriviality. We demonstrated this approach by automatically searching motion-tracking data captured from various physical systems, ranging from simple harmonic oscillators to chaotic double-pendula. Without any prior knowledge about physics, kinematics, or geometry, the algorithm discovered Hamiltonians, Lagrangians, and other laws of geometric and momentum conservation. The discovery rate accelerated as laws found for simpler systems were used to bootstrap explanations for more complex systems, gradually uncovering the "alphabet" used to describe those systems.

View Full Text

Science
Vol 324, Issue 5923
03 April 2009

Table of Contents
Print Table of Contents
Advertising (PDF)
Classified (PDF)
Masthead (PDF)

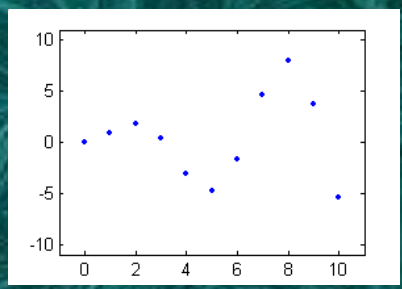
ARTICLE TOOLS
Email
Print
Alerts
Share
Download Powerpoint
Request Permissions
Citation tools

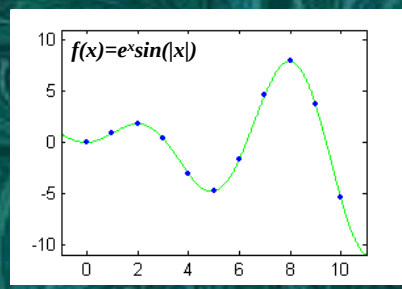
MY SAVED FOLDERS
Save to my folders
View my saved folders

STAY CONNECTED TO *SCIENCE*
- Facebook
- Twitter

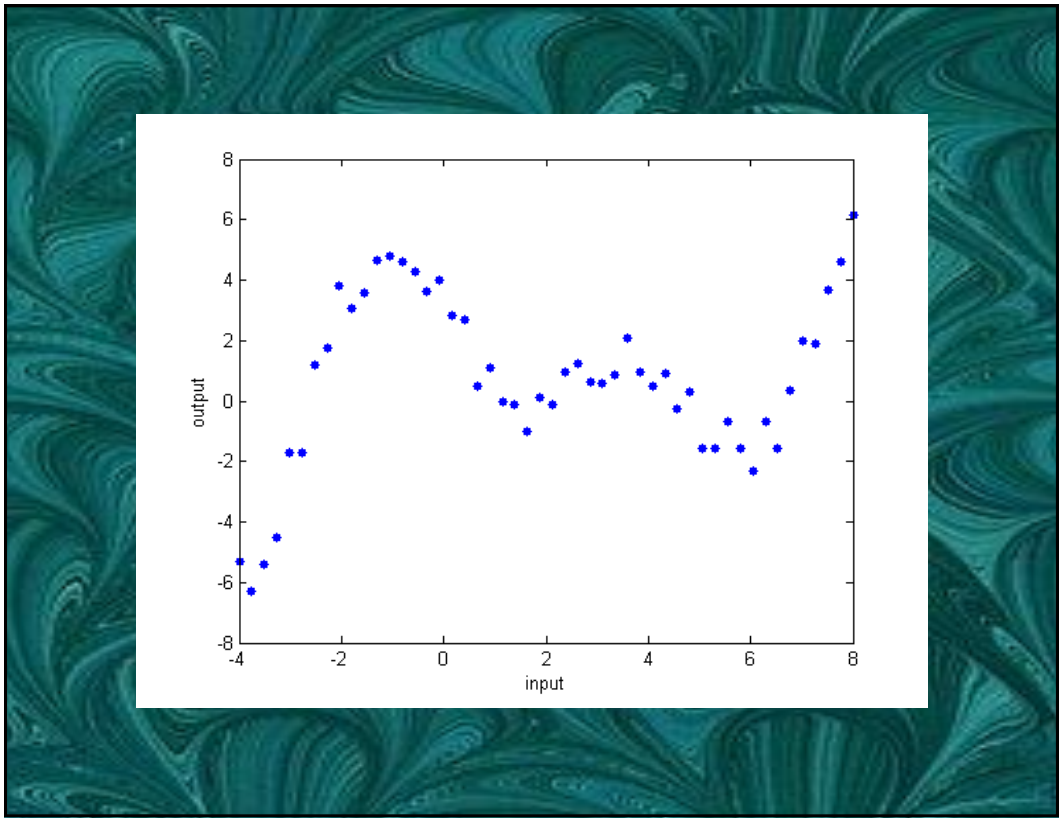
Symbolic Regression

What function describes this data?



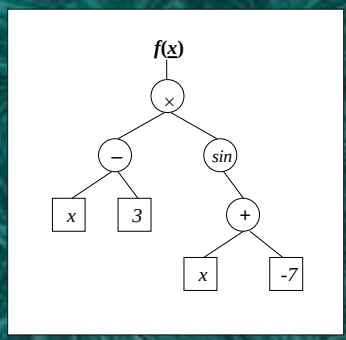


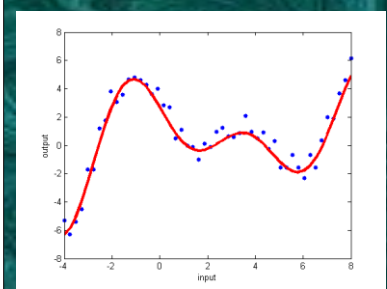
John Koza, 1992



Encoding Equations

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





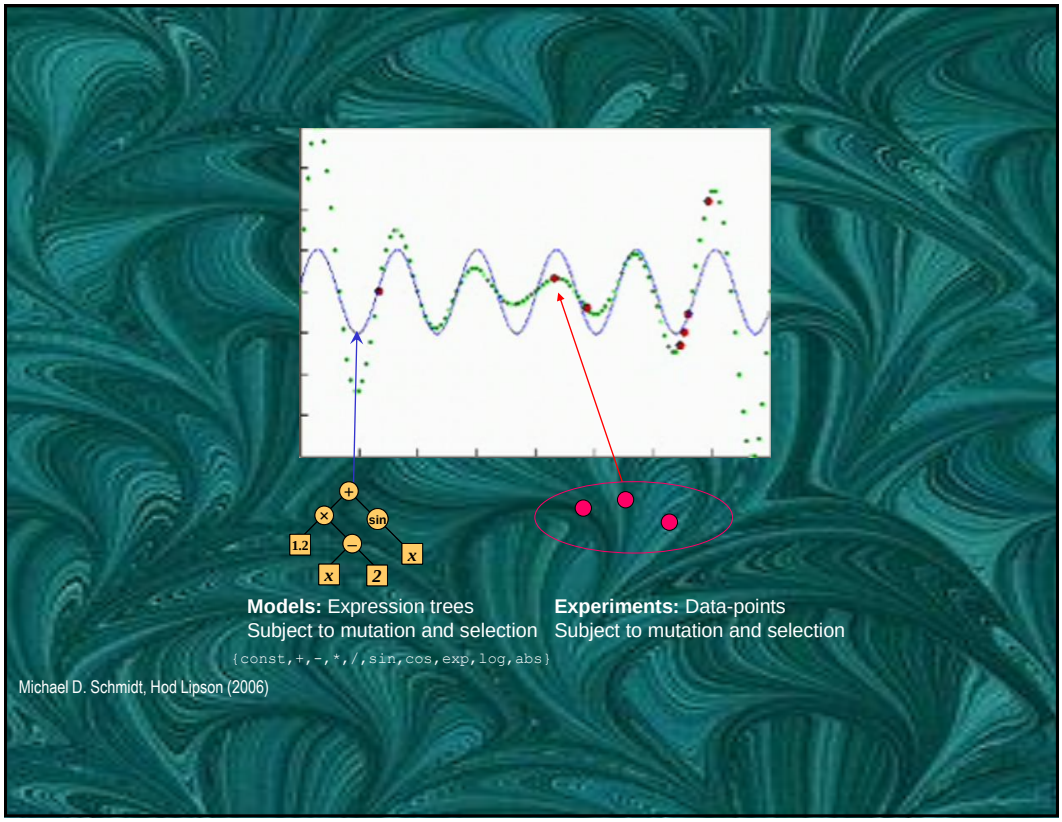
$\sin(x_2)$

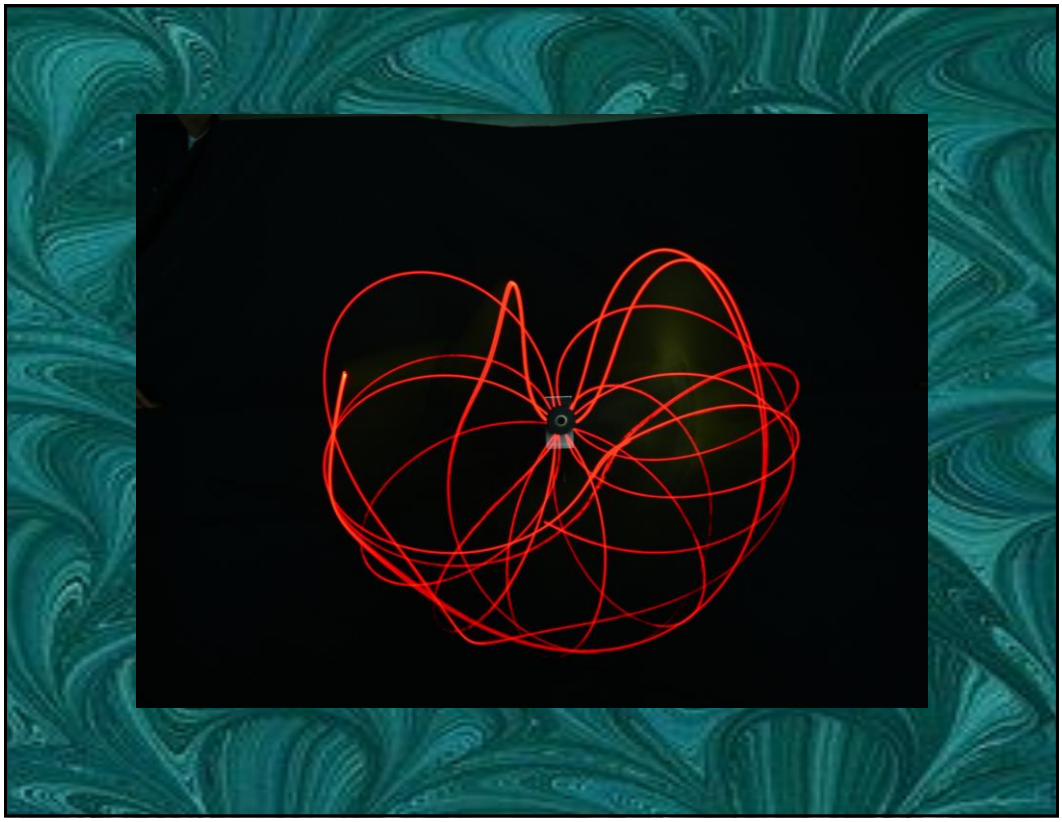
$x_1 * \sin(x_2)$

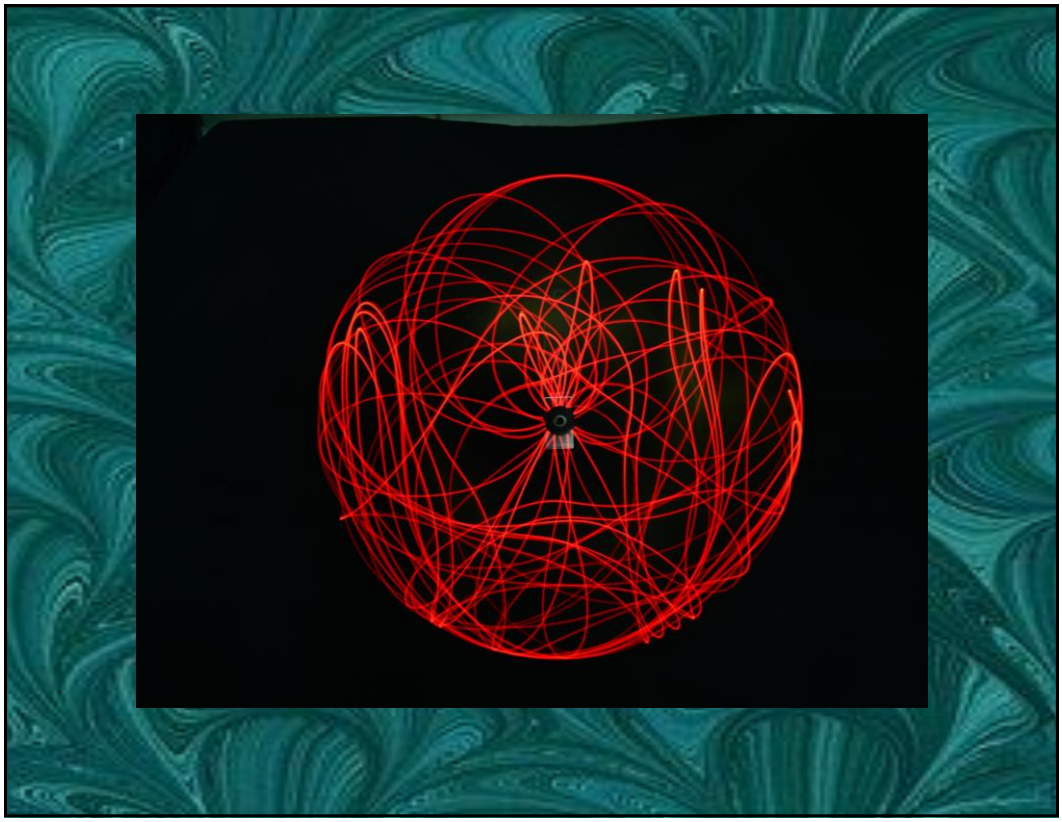
$(x_1 - 3) * \sin(x_2)$

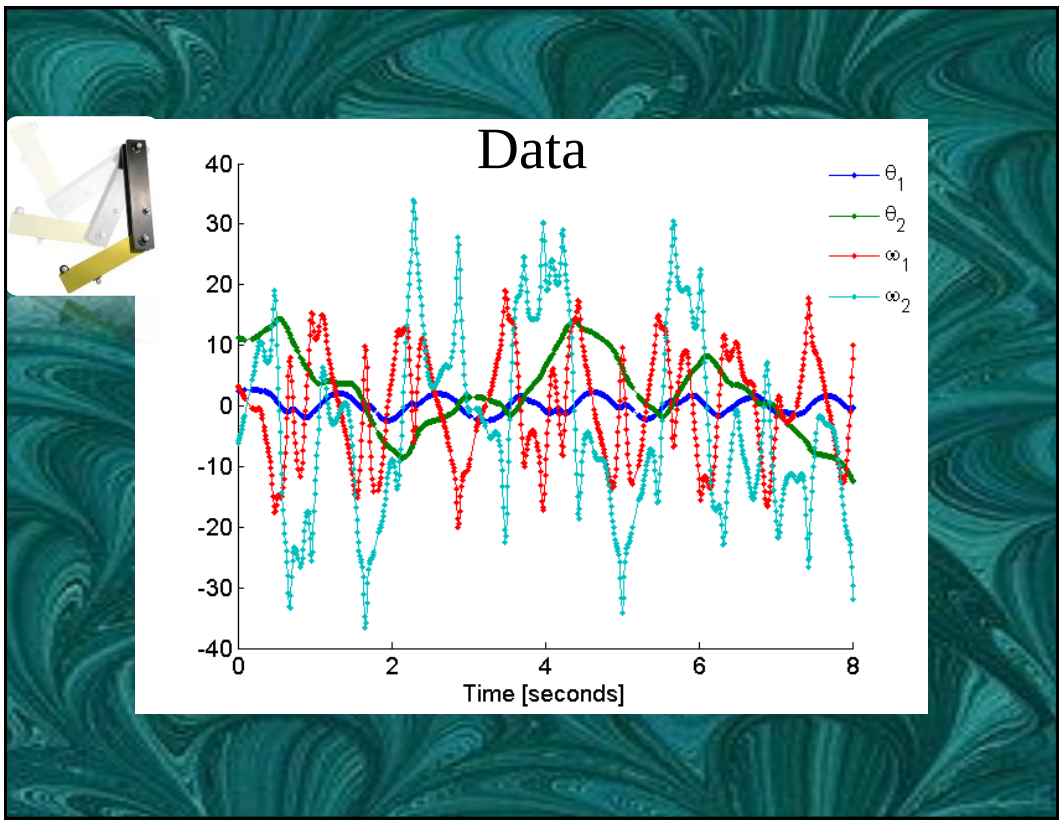
$(x_1 - 3) * \sin(-7 + x_2)$

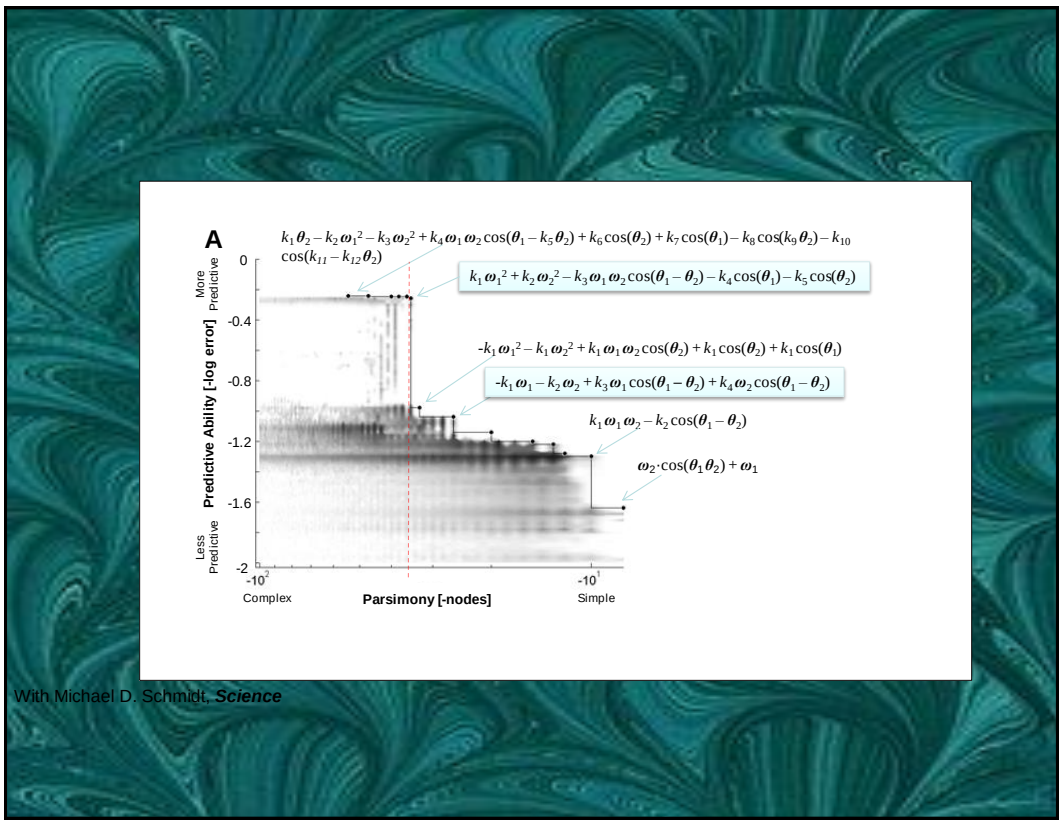
John Koza, 1992


+
×
sin
1.2
−
x
x
2
Models: Expression trees
Subject to mutation and selection
{const,+,-,*,/,sin,cos,exp,log,abs}
Experiments: Data-points
Subject to mutation and selection
Michael D. Schmidt, Hod Lipson (2006)





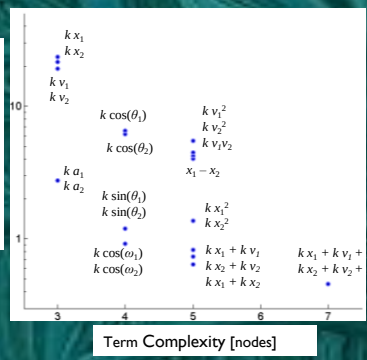

Data
40
30
20
10
0
-10
-20
-30
-40
0
2
4
6
8
Time [seconds]
θ1
θ2
ω1
ω2

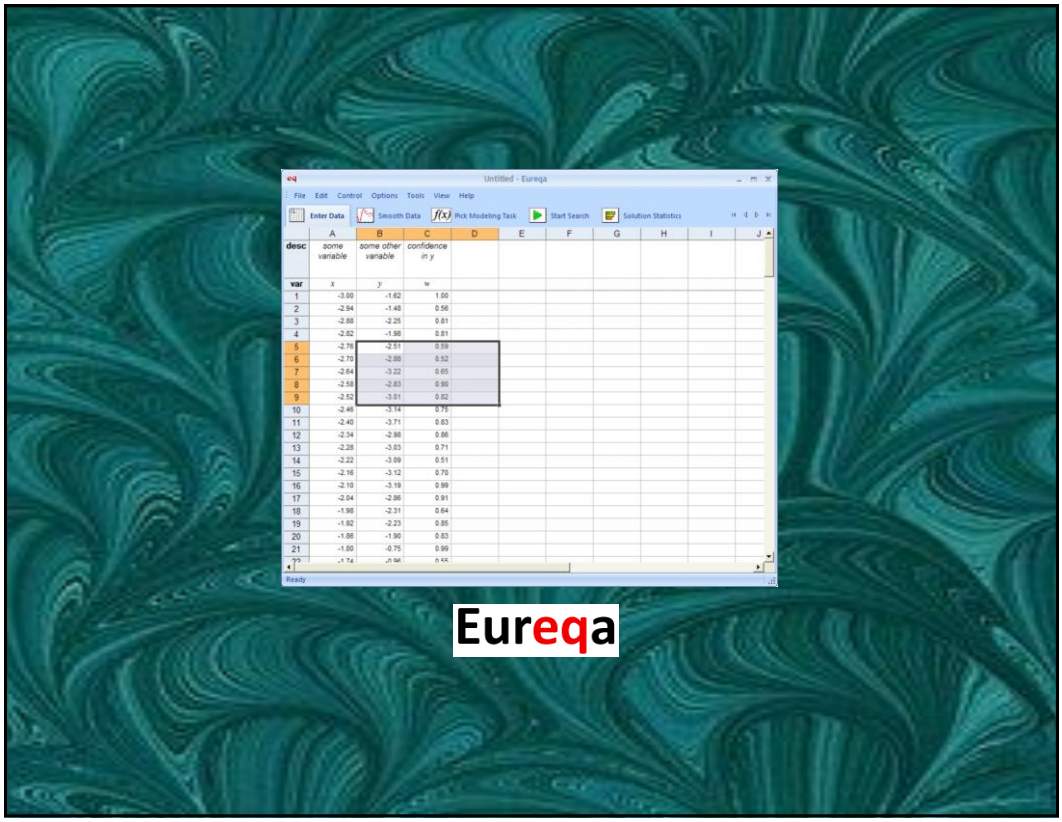


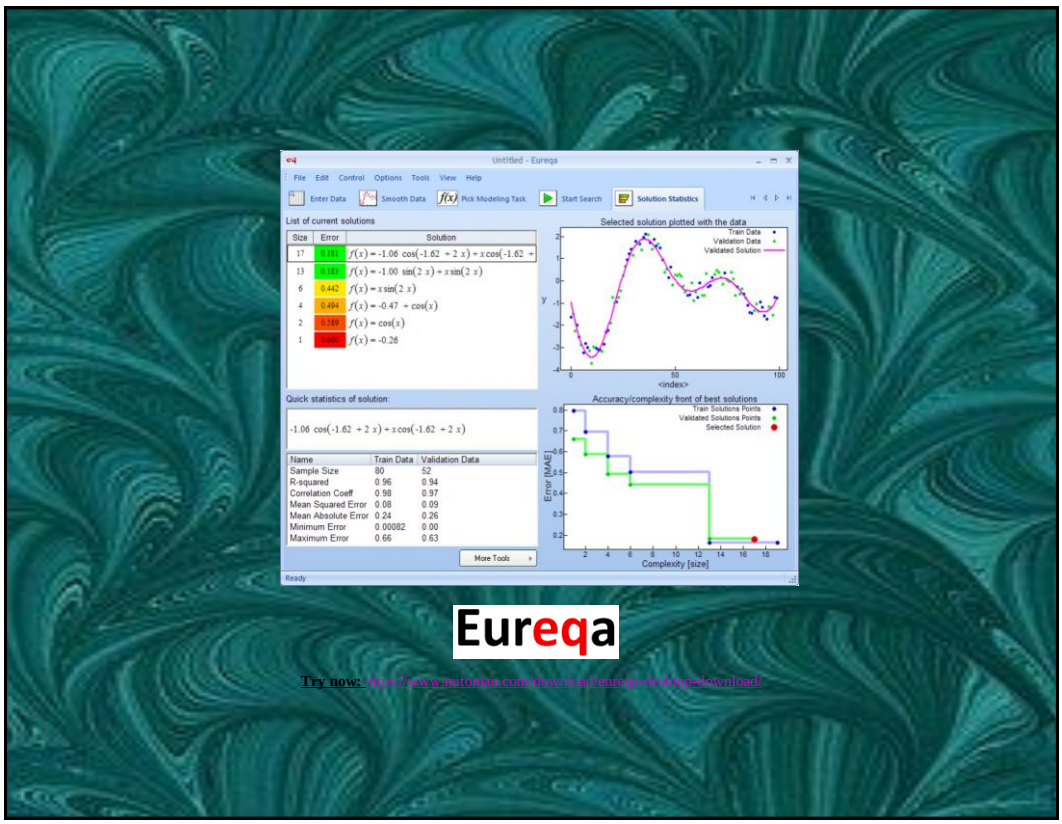
With Michael D. Schmidt, ***Science***

Recurring Motifs – New Building Blocks

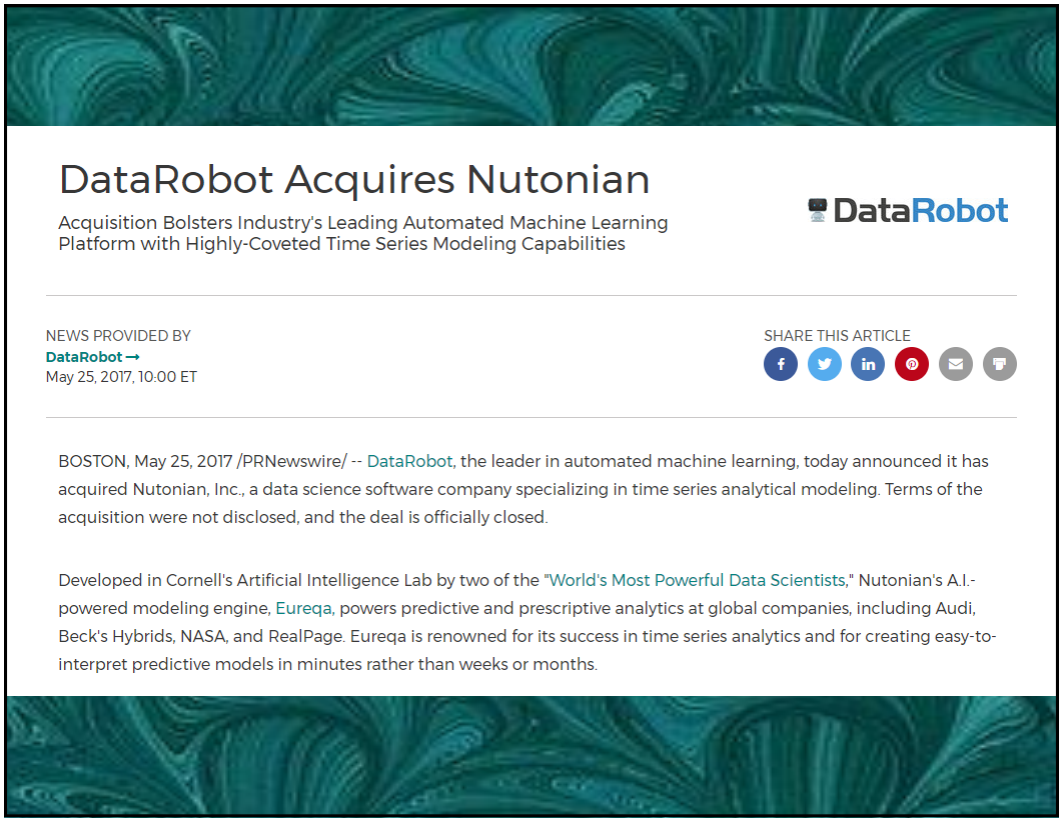
Term	Complexity	Frequency	Systems
$k\,x_2$	1	23.578	4
$k\,x_1$	1	21.6514	4
$k\,v_2$	1	21.5596	4
$k\,v_1$	1	19.1743	4
$k\cos(\theta_2)$	2	6.51376	2
$k\cos(\theta_1)$	2	6.14679	2
$k\,v_2^2$	3	5.50459	4
$k\,v_1^2$	3	4.49541	4
$(x_1 - x_2)$	3	4.22018	2
$k\,v_1 v_2$	3	4.0367	2
$k\,a_2$	1	2.75229	2
$k\,a_1$	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







Try now: https://www.nutonian.com/download/eureqa-desktop-download/



DataRobot Acquires Nutonian

Acquisition Bolsters Industry's Leading Automated Machine Learning Platform with Highly-Coveted Time Series Modeling Capabilities

DataRobot

NEWS PROVIDED BY
DataRobot →
May 25, 2017, 10:00 ET

SHARE THIS ARTICLE

BOSTON, May 25, 2017 /PRNewswire/ -- DataRobot, the leader in automated machine learning, today announced it has acquired Nutonian, Inc., a data science software company specializing in time series analytical modeling. Terms of the acquisition were not disclosed, and the deal is officially closed.

Developed in Cornell's Artificial Intelligence Lab by two of the "World's Most Powerful Data Scientists," Nutonian's A.I.-powered modeling engine, Eureqa, powers predictive and prescriptive analytics at global companies, including Audi, Beck's Hybrids, NASA, and RealPage. Eureqa is renowned for its success in time series analytics and for creating easy-to-interpret predictive models in minutes rather than weeks or months.