Interpretability in scientific machine learning

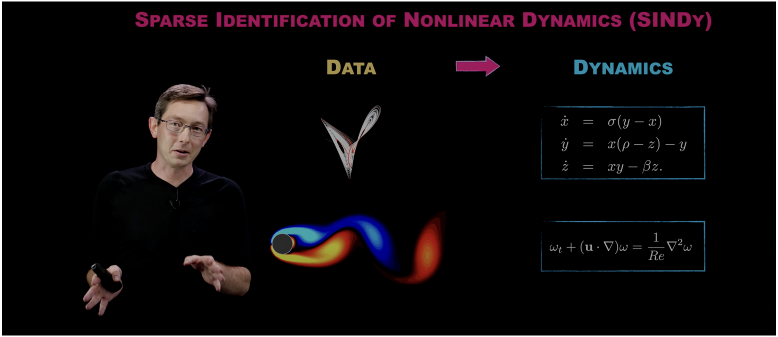
Conor Rowan [1], Alireza Doostan [1]

[1]University of Colorado Boulder, Aerospace Engineering

NREL Group

About me

- Interested in machine learning for engineering mechanics and philosophical problems in scientific research



- Sparse identification of nonlinear dynamics (SINDy) introduced in 2016 [2]
- Uses measurement data to identify governing ordinary or partial differential equations from library of candidate terms
- Recover Navier-Stokes equations [17], equations of nonlinear pendulum [3], various wave equations [19]

Background—symbolic regression

Feynman eq.	Equation	Solution time (s)	Methods used	Data needed	Solved by Eureqa	Solved w/o da	Noise tolerance
I.6.20a	$f = e^{-\theta^2/2}/\sqrt{2\pi}$	16	bf	10	no	yes	10^{-2}
I.6.20	$f = e^{-\frac{\theta^2}{2\sigma^2}}/\sqrt{2\pi\sigma^2}$	2992	ev, bf-log	10^2	no	yes	10^{-4}
I.6.20b	$f = e^{-\frac{(\theta-\theta_1)^2}{2\sigma^2}}/\sqrt{2\pi\sigma^2}$	4792	sym–, ev, bf-log	10^3	no	yes	10^{-4}
I.8.14	$d = \sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$	544	da, pf-squared	10^2	no	yes	10^{-4}
I.9.18	$F = \frac{Gm_1m_2}{(x_2-x_1)^2+(y_2-y_1)^2+(z_2-z_1)^2}$	5975	da, sym–, sym–, sep∗, pf-inv	10^6	no	yes	10^{-5}
I.10.7	$m = \frac{m_0}{\sqrt{1-\frac{v^2}{c^2}}}$	14	da, bf	10	no	yes	10^{-4}
I.11.19	$A = x_1y_1 + x_2y_2 + x_3y_3$	184	da, pf	10^2	yes	yes	10^{-3}
I.12.1	$F = \mu N_n$	12	da, bf	10	yes	yes	10^{-3}
I.12.2	$F = \frac{q_1q_2}{4\pi\epsilon r^2}$	17	da, bf	10	yes	yes	10^{-2}
I.12.4	$E_f = \frac{q_1}{4\pi\epsilon r^2}$	12	da	10	yes	yes	10^{-2}
I.12.5	$F = q_2E_f$	8	da	10	yes	yes	10^{-2}
I.12.11	$F = q(E_f + Bv\sin\theta)$	19	da, bf	10	yes	yes	10^{-3}
I.13.4	$K = \frac{1}{2}m(v^2+u^2+w^2)$	22	da, bf	10	yes	yes	10^{-4}
I.13.12	$U = Gm_1m_2(\frac{1}{r_2}-\frac{1}{r_1})$	20	da, bf	10	yes	yes	10^{-4}
I.14.3	$U = mgz$	12	da	10	yes	yes	10^{-2}
I.14.4	$U = \frac{k_{spring}x^2}{2}$	9	da	10	yes	yes	10^{-2}

- Symbolic regression uses genetic algorithms to search through large space of mathematical functions
- "AI Feynman" discovers algebraic equations from physics using noisy data [22]
- Rediscover gravitational force law from trajectory data of planets in solar system [12]

A new paradigm for science?

Miles Cranmer - The Next Great Scientific Theory is Hiding Inside a Neural Network (April 3, 2024)

[S] Simons Foundation
39.2K subscribers
Subscribe
6.2K Share Ask

ALGORITHMS

Powerful 'Machine Scientists' Distill the Laws of Physics From Raw Data

Q & A

The Physicist Working to Build Science-Literate AI

Physics

Will artificial intelligence ever discover new laws of physics?

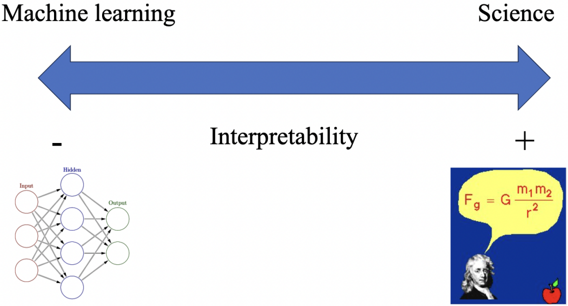
CHRIS ANDERSON SCIENCE JUN 23, 2008 12:00 PM

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

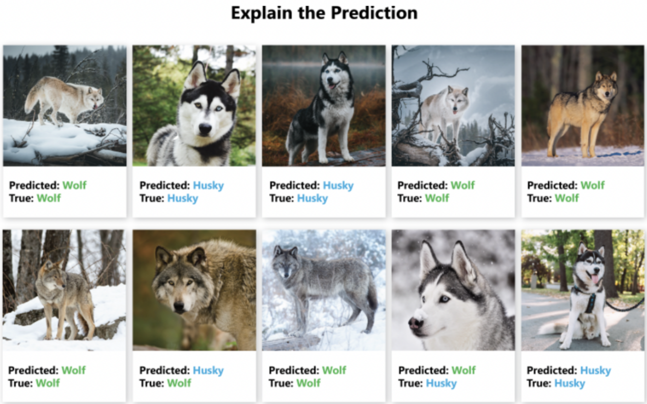
- Machine learning excels where human intuition fails in other domains—these tools suggest a new approach to scientific discovery

This promise relies on a number of assumptions...

1. There are fundamental differential equations left to discover **(not obvious)**
2. The crux of scientific discovery is fitting equations to data **(wrong)**
3. There is a difference between fitting data and true discovery; nature is parsimonious **(not obvious)**
4. Science is interpretable, traditional machine learning tools are not **(what does this mean?)**



Interpretable machine learning



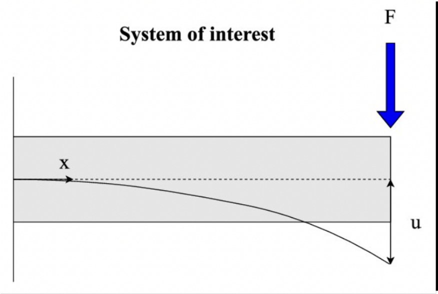
- Interpretability is considered important for safety, ethics, trust, and debugging [18]
- Here, interpretation is extracting the causal logic from a trained model
- This does not distinguish between a surrogate model and a law

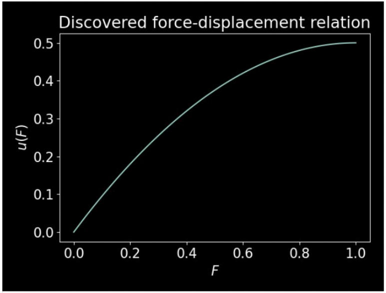
Definitions of interpretability in scientific machine learning

Author(s)	Sparsity	Transparency	Mechanism
Bongard & Lipson [1]	✓	✗	✗
Lipson & Schmidt [20]	✓	✗	✗
Brunton et al. [2]	✓	✗	✗
Champion et al. [3]	✓	✗	✗
Tripura & Chakraborty [21]	✓	✓	✗
Lu et al. [13]	✓	✓	✗
Desai & Strachan [5]	✓	✗	✓
Massonis et al. [16]	✓	✓	✓
Garbrecht et al. [8]	✓	✓	✗
Flascehl et al. [6]	✓	✓	✗
Fuhg et al. [7]	✓	✓	✗
Udrescu & Tegmark [22]	✓	✗	✗
Cranmer [4]	✓	✗	✗
Wang et al. [23]	✓	✓	✗
Makke & Chawla [15]	✓	✓	✗
Guimera et al. [9]	✓	✗	✗

- Literature review suggests three definitions: **sparsity, transparency, mechanism**
- Sparsity distinguishes between a surrogate model and a law

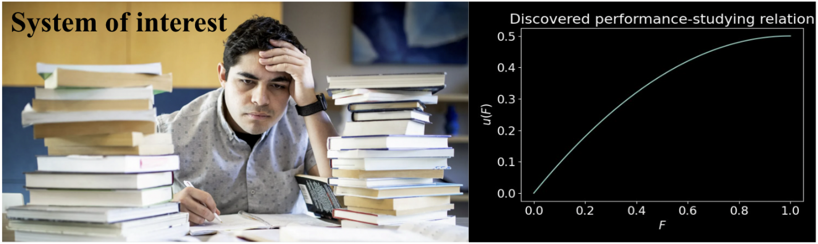
Does sparsity guarantee interpretability?





- Consider hypothetical data on force-displacement relation of cantilevered beam, obtain the model $u = \lambda_1 F - \lambda_2 F^2$
- First term is linear elastic response, second term captures geometric nonlinearity
- This interpretation requires a lot of prior knowledge...

Sparsity without prior knowledge

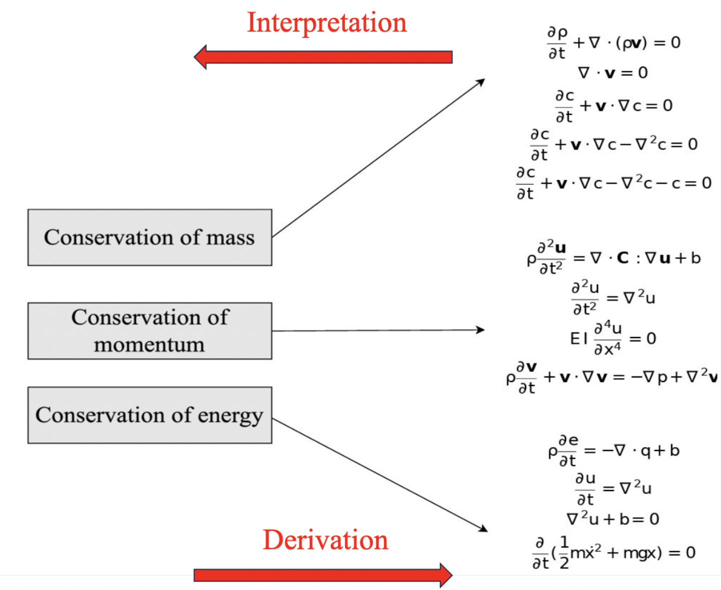


- F = number of hours spent studying for a test, u = performance on the test, discover from data the relation $u = \lambda_1 F - \lambda_2 F^2$
- Not clear what this (sparse) equation tells us about the "system of interest"
- Can define interpretation to mean whatever we want, but sparsity does not capture the common sense notion of interpretation here

The failure of sparsity

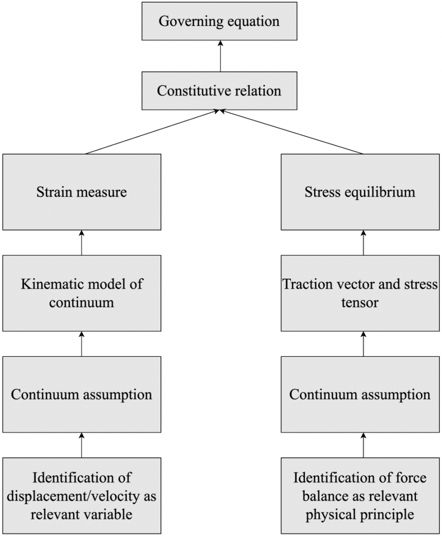
- Claim: interpretation in science happens at the level of mechanisms, not equations—interpretation is an answer to a "why" question
- There is no algorithm to back out mechanisms from sparse equations (think: statistical mechanics)
- Addressing "why" questions requires appeal to some kind of primitive/axiom
- Explanation/interpretation is a matter of pulling back to something more fundamental
- Primitives in science are empirical laws with universal applicability (conservation of mass, momentum, energy, charge, etc.) [10]
- Fundamental laws are the vehicle for interpretation, and not themselves interpretable [14]

Laws and interpretation



The case of solid mechanics

- Consider governing equation of stress equilibrium $\nabla \cdot \boldsymbol{\sigma} + \mathbf{b}(\mathbf{x}) = \mathbf{0}$



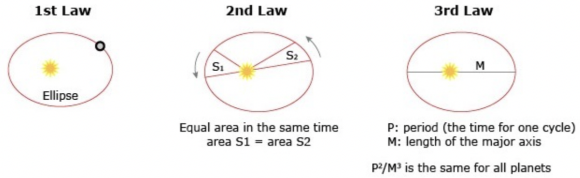
Amended definition of interpretability

Definition

A learned model is interpretable when it can be derived from fundamental physical principles or it represents an empirical component of a model derived from fundamental physical principles

- This definition prevents "great scientific theories" from being interpretable
- Limits interpretation to Kuhnian normal science [11] in which prior knowledge is abundant
- Sparsity does not guarantee interpretation, but it leaves the door open for it

The example of Kepler's laws



- Many works cite Kepler's laws as a paradigmatic case of interpretable scientific discovery [4, 22, 3, 2]
- Kepler's three laws are 1) planets move in elliptical orbits with the sun as one of the two foci, 2) at all positions of a planet's orbit, a line drawn from a planet to the sun sweeps out equal areas over a given unit of time, 3) the square of a planet's orbital period is proportional to the cube of the semi-major axis of the orbit
- Kepler's laws only become interpretable as a consequence of Newton's law of gravitation

Conservation of mass?

- We wish to discover the space-time dynamics of a scalar quantity $c(x,t)$:

$$\frac{\partial c}{\partial t} = \mathcal{N}\left(c, \frac{\partial c}{\partial x}, \frac{\partial^2 c}{\partial x^2}, \dots ; \boldsymbol{\lambda}\right)$$

- Use your favorite method to obtain:

$$\frac{\partial c}{\partial t} + \lambda_1 \frac{\partial c}{\partial x} - \lambda_2 \frac{\partial^2 c}{\partial x^2} - \lambda_3 c + \lambda_4 \left|\frac{\partial c}{\partial x}\right| c = 0$$

- Assign meaning to the terms in order to gain insight into the discovered equation:

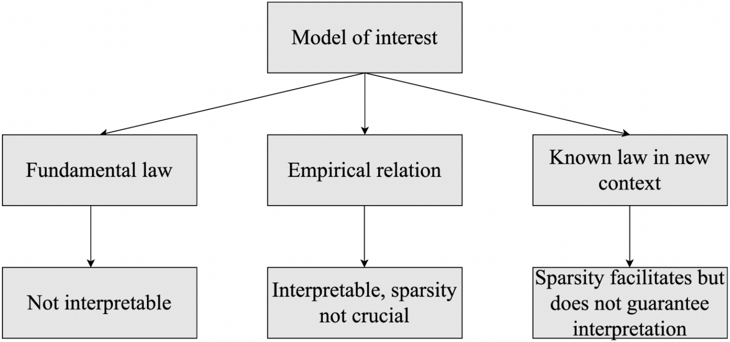
$$\underbrace{\frac{\partial c}{\partial t}}_{\text{time evolution}} + \underbrace{\lambda_1 \frac{\partial c}{\partial x}}_{\text{advection}} - \underbrace{\lambda_2 \frac{\partial^2 c}{\partial x^2}}_{\text{diffusion}} - \underbrace{\lambda_3 c}_{\text{reaction}} + \underbrace{\lambda_4 \left|\frac{\partial c}{\partial x}\right| c}_{?} = 0$$

- The unfamiliar term is only interpretable if it can be connected to a mechanism of mass transport

Conclusion

- Sparse equations have few parameters and few parameter models have the ability to generalize—there is definitely something fundamental about sparsity in science
- But generalization and interpretability are distinct properties of a model
- Truly novel discoveries are not interpretable, interpretable discoveries are not truly novel
- Data-driven models *are* a new paradigm for doing science and claims about them should be informed by the history & philosophy of science
- Many questions remain about the prospects of these tools for scientific discovery...

Thanks for listening! Questions/discussion?



Rowan, C., and Doostan, A., "On the definition and importance of interpretability in scientific machine learning," Preprint, 2025.

References I

Josh Bongard and Hod Lipson.
Automated reverse engineering of nonlinear dynamical systems.
Proceedings of the National Academy of Sciences of the United States of America, 104(24):9943–9948, June 2007.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz.
Discovering governing equations from data by sparse identification of nonlinear dynamical systems.
Proceedings of the National Academy of Sciences, 113(15):3932–3937, April 2016.
Publisher: Proceedings of the National Academy of Sciences.

References II

Kathleen Champion, Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton.
Data-driven discovery of coordinates and governing equations.
Proceedings of the National Academy of Sciences, 116(45):22445–22451, November 2019.
Publisher: Proceedings of the National Academy of Sciences.

Miles Cranmer.
Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl, May 2023.
arXiv:2305.01582 [astro-ph].

Saaketh Desai and Alejandro Strachan.
Parsimonious neural networks learn interpretable physical laws.
Scientific Reports, 11(1):12761, June 2021.
Publisher: Nature Publishing Group.

References III

Moritz Flaschel, Siddhant Kumar, and Laura De Lorenzis.
Unsupervised discovery of interpretable hyperelastic constitutive laws.
Computer Methods in Applied Mechanics and Engineering, 381:113852, August 2021.

Jan N. Fuhg, Reese E. Jones, and Nikolaos Bouklas.
Extreme sparsification of physics-augmented neural networks for interpretable model discovery in mechanics, October 2023.
arXiv:2310.03652 [cs].

Karl Garbrecht, Miguel Aguilo, Allen Sanderson, Anthony Rollett, Robert M. Kirby, and Jacob Hochhalter.
Interpretable Machine Learning for Texture-Dependent Constitutive Models with Automatic Code Generation for Topological Optimization.

References IV

Integrating Materials and Manufacturing Innovation, 10(3):373–392, September 2021.

Roger Guimera, Ignasi Reichardt, Antoni Aguilar-Mogas, Francesco A. Massucci, Manuel Miranda, Jordi Pallares, and Marta Sales-Pardo.
A Bayesian machine scientist to aid in the solution of challenging scientific problems.
Science Advances, 6(5):eaav6971, January 2020.
arXiv:2004.12157 [cs].

Philip Kitcher.
EXPLANATORY UNIFICATION.
Philosophy of Science, 1981.

Thomas S. Kuhn.
The Structure of Scientific Revolutions: 50th Anniversary Edition.
University of Chicago Press, Chicago, IL, April 2012.

References V

Pablo Lemos, Niall Jeffrey, Miles Cranmer, Shirley Ho, and Peter Battaglia.
Rediscovering orbital mechanics with machine learning, February 2022.

arXiv:2202.02306 [astro-ph].

Peter Y. Lu, Joan Ariño Bernad, and Marin Soljačić.
Discovering sparse interpretable dynamics from partial observations.
Communications Physics, 5(1):1–7, August 2022.
Publisher: Nature Publishing Group.

Ernst Mach.
On the Economical Nature of Physical Inquiry.
In Thomas J. McCormack, editor, *Popular Scientific Lectures*, Cambridge Library Collection - Physical Sciences, pages 186–213. Cambridge University Press, Cambridge, 2014.

References VI

Nour Makke and Sanjay Chawla.
Interpretable scientific discovery with symbolic regression: a review.
Artificial Intelligence Review, 57(1):2, January 2024.

Gemma Massonis, Alejandro F. Villaverde, and Julio R. Banga.
Distilling identifiable and interpretable dynamic models from biological data.
PLOS Computational Biology, 19(10):e1011014, October 2023.
Publisher: Public Library of Science.

Daniel A. Messenger and David M. Bortz.
Weak SINDy For Partial Differential Equations.
Journal of Computational Physics, 443:110525, October 2021.
arXiv:2007.02848 [math].

Cristoph Molnar.
Interpretable Machine Learning, 2025.

References VII

Samuel H. Rudy, Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz.
Data-driven discovery of partial differential equations, September 2016.
arXiv:1609.06401 [nlin].

Michael Schmidt and Hod Lipson.
Distilling Free-Form Natural Laws from Experimental Data.
Science, 324(5923):81–85, April 2009.
Publisher: American Association for the Advancement of Science.

Tapas Tripura and Souvik Chakraborty.
Discovering interpretable Lagrangian of dynamical systems from data.
Computer Physics Communications, 294:108960, January 2024.

References VIII

Silviu-Marian Udrescu and Max Tegmark.
AI Feynman: a Physics-Inspired Method for Symbolic Regression, April 2020.
arXiv:1905.11481 [physics].

Yiqun Wang, Nicholas Wagner, and James M. Rondinelli.
Symbolic Regression in Materials Science.
MRS Communications, 9(3):793–805, September 2019.
arXiv:1901.04136 [cond-mat].