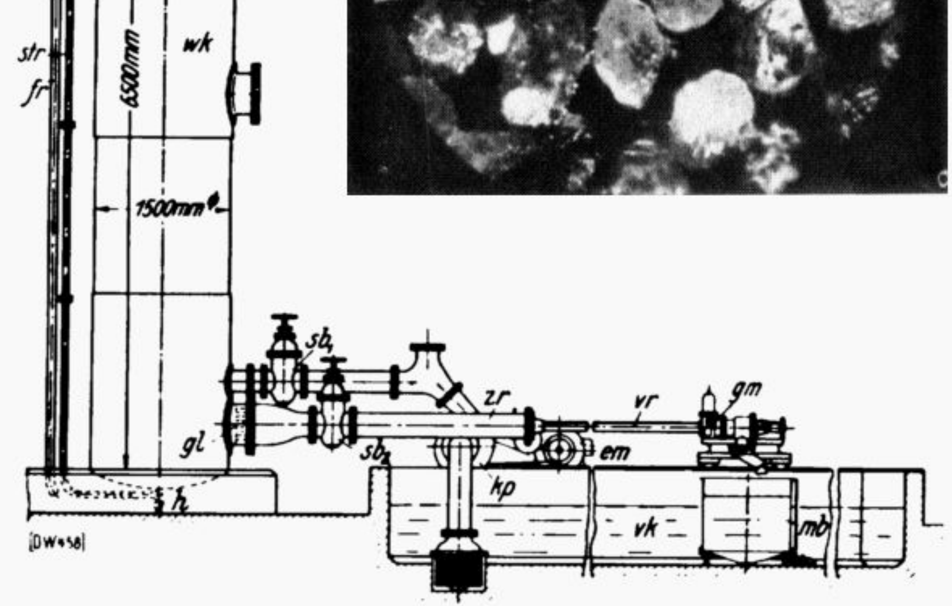
Computational scientific discovery

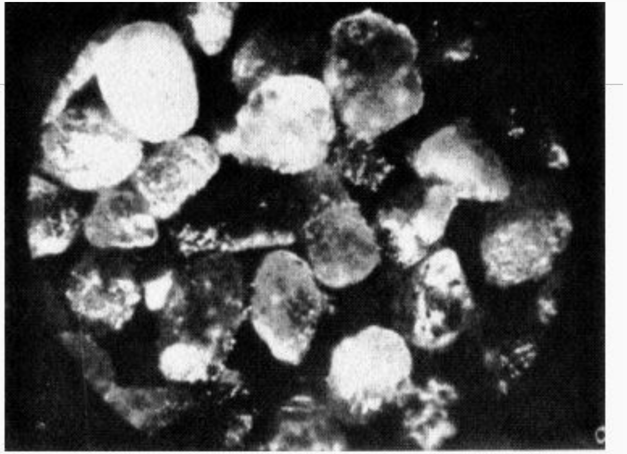
Roger Guimerà
ICREA & Univ. Rovira i Virgili, Catalonia

1st SMASH Workshop, Vipava, Slovenia
October 10th, 2024

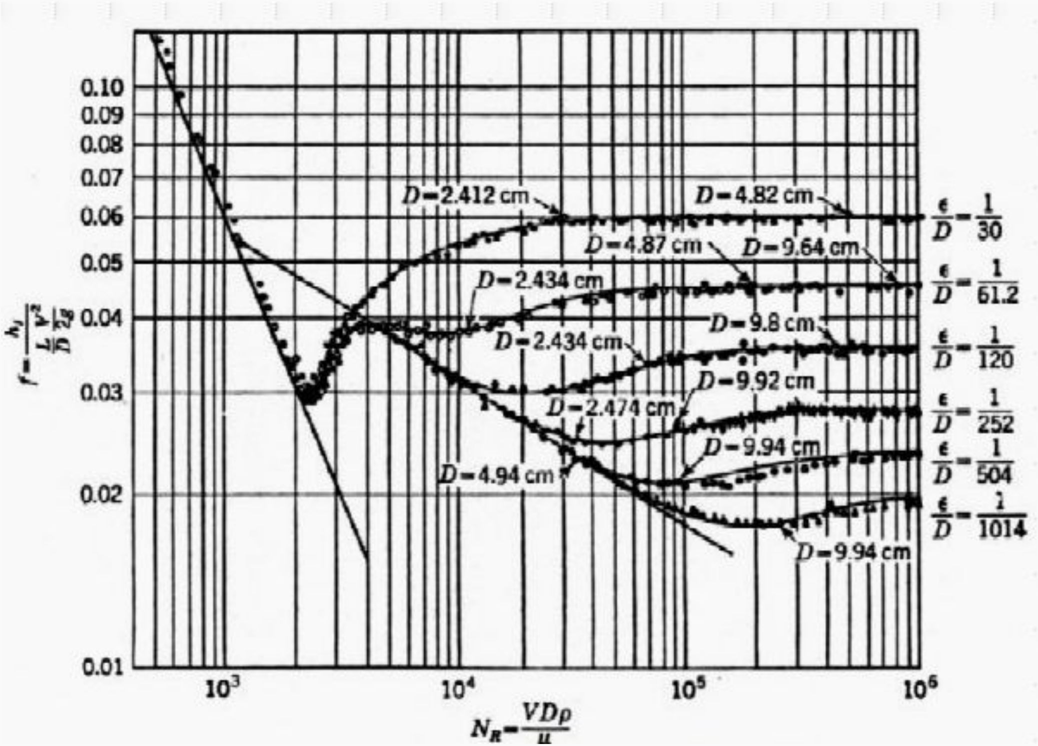
ICREA
UNIVERSITAT ROVIRA i VIRGILI

Nikuradse’s 1933 experiments about friction in rough pipes

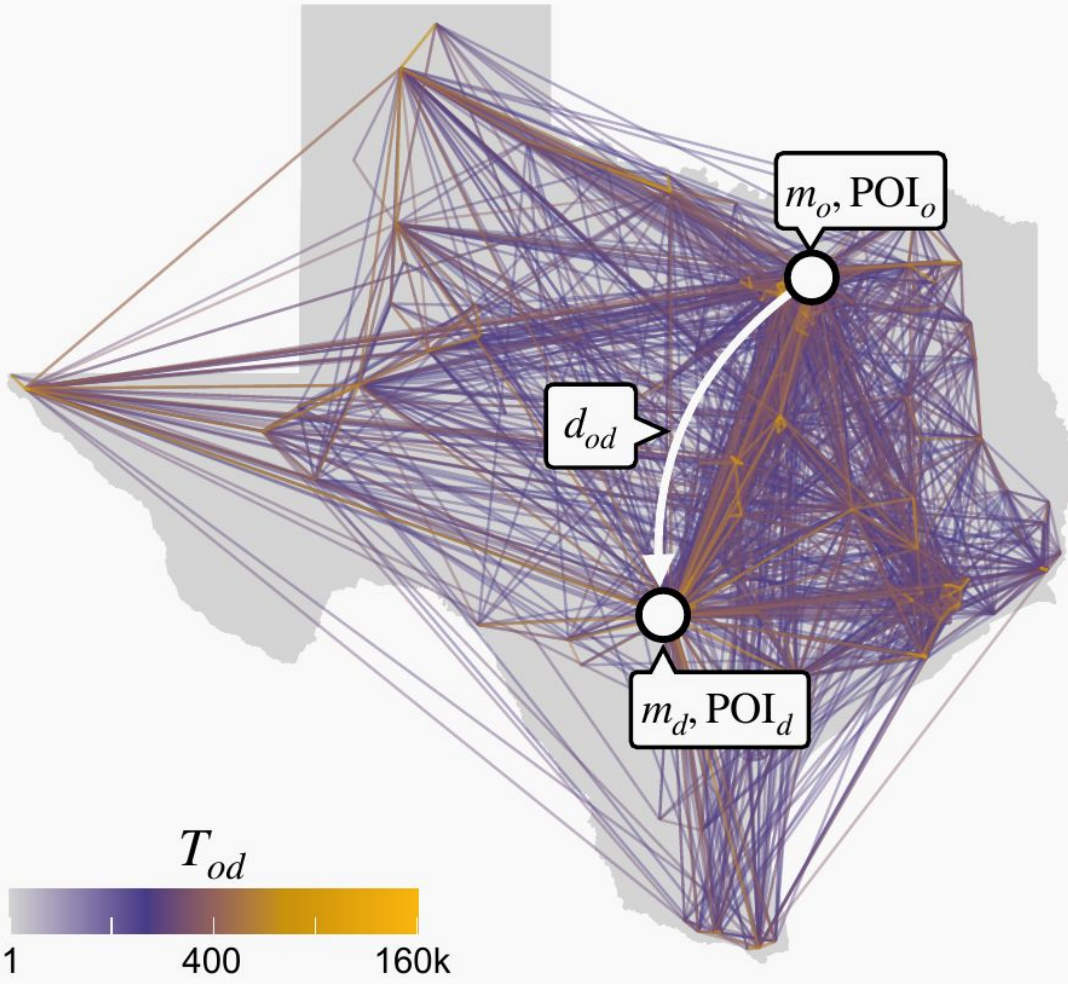




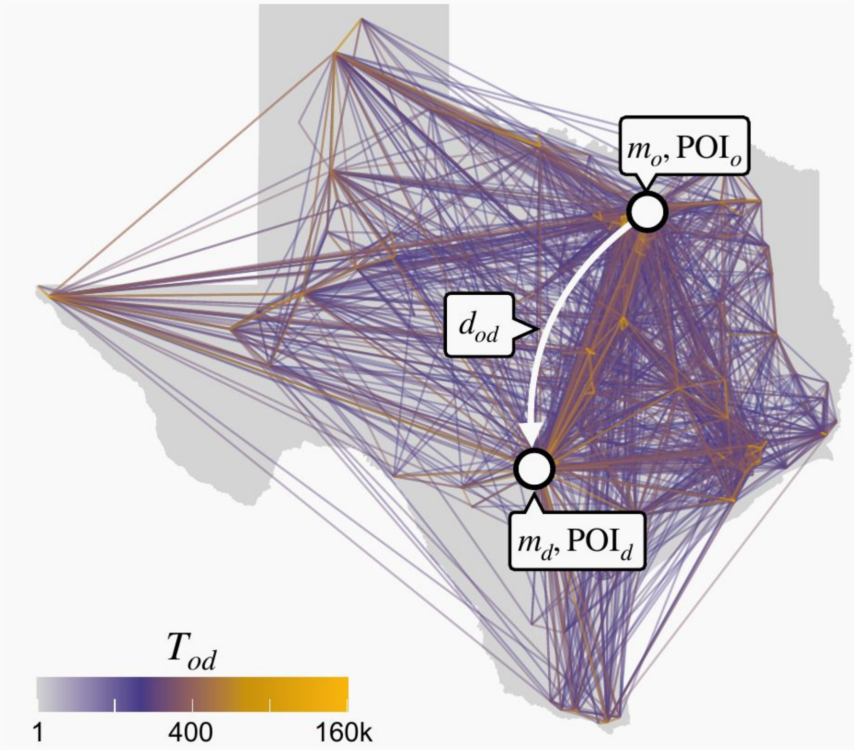
Nikuradse’s 1933 experiments about friction in rough pipes



Can we find models that predict human mobility flows?



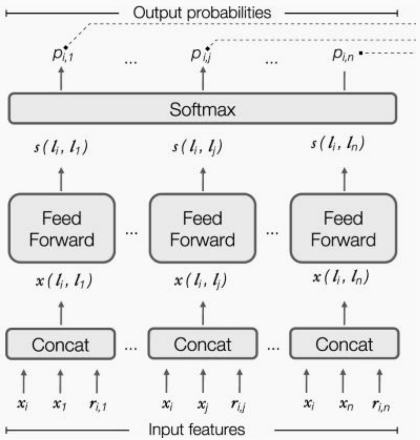
Can we find models that predict human mobility flows?



Gravity models

$$T_{od} = A\,\frac{m_o\, m_d}{d^{\alpha}}$$

"Deep gravity" model



Simini et al., *Nature Comm.* (2021)

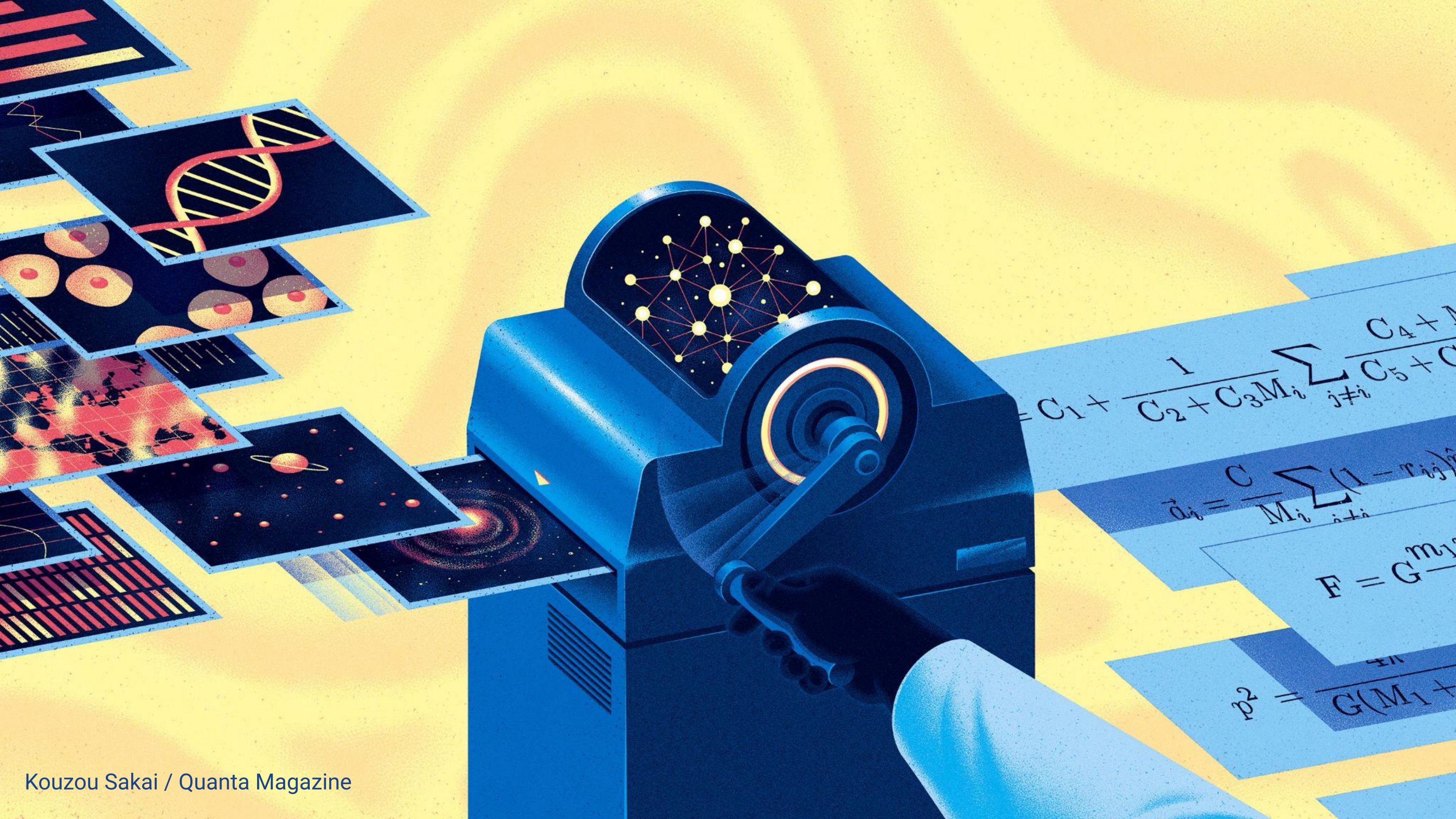
$$y = f(x, \theta)$$

Can we design a “machine scientist” that automates the task of building **closed-form mathematical models** from data?

$$y = f(x, \theta)$$

Can we design a “machine scientist” that automates the task of building **closed-form mathematical models** from data?

$$f(x) = a_0 + a_1 x \qquad f(x) = \log\left(\sin\left(\exp\left(x^{-8}\right)\right)\right)$$



Kouzou Sakai / Quanta Magazine

Desiderata

for the machine scientist

It should take arbitrary data as input and produce closed-form, interpretable mathematical models

Desiderata

for the machine scientist

It should take arbitrary data as input and produce closed-form, interpretable mathematical models

It should be able to rigorously and quantitatively ***establish the plausibility*** of a model given some data

Desiderata

for the machine scientist

It should take arbitrary data as input and produce closed-form, interpretable mathematical models

It should be able to rigorously and quantitatively ***establish the plausibility*** of a model given some data

It should be able to systematically ***explore the space*** of all possible mathematical models, so that the stationary distribution of visited models is (at least asymptotically) given by their plausibility

$y = f(x, \theta)$

$$p(f \mid \{x, y\})$$

This posterior over expressions *f* encapsulates the full probabilistic solution to the symbolic regression problem

We use probability theory to select models rigorously (aka Bayesian model selection)

Without parameters:

$$p(M|D) = \frac{p(D|M)\,p(M)}{p(D)}$$

With parameters:

We use probability theory to select models rigorously (aka Bayesian model selection)

Without parameters:

$$p(M|D) = \frac{p(D|M)\, p(M)}{p(D)}$$

With parameters:

$$p(f,\theta|D) = \frac{p(D|f,\theta)\, p(f,\theta)}{p(D)} = \frac{p(D|f,\theta)\, p(\theta|f)\, p(f)}{p(D)}$$

$$p(f|D) = \int_\Theta d\theta\; p(f,\theta|D) = \frac{1}{p(D)} \int_\Theta d\theta\; p(D|f,\theta)\, p(\theta|f) p(f)$$

integrated likelihood

But can we *act* on the probabilistic approach?

The posterior can be rewritten as

$$
\begin{aligned}
p(f|D) &= \frac{1}{p(D)} \int_\Theta d\theta \, p(D|f,\theta) \, p(\theta|f) \, p(f) \\
&= \frac{\mathrm{e}^{-\mathcal{L}(f,D)}}{p(D)}
\end{aligned}
$$

But can we *act* on the probabilistic approach?

The posterior can be rewritten as

$$p(f|D) = \frac{1}{p(D)} \int_\Theta d\theta \, p(D|f,\theta) \, p(\theta|f) \, p(f)$$
$$= \frac{\mathrm{e}^{-\mathcal{L}(f,D)}}{p(D)}$$

And the ***description length*** can be approximated as

$$\mathcal{L}(f,D) = \frac{B(f)}{2} - \log p(f)$$

BIC prior

Arguments for a probabilistic approach

Cox-type argument: Any alternative way to assign plausibilities to models must violate some of the very basic conditions in the desiderata

Dutch book-type argument: Betting on models using any alternative assignment of plausibility results in sets of bets that one would be willing to accept but that result in certain loss

Consistency argument: Any alternative that does not coincide with the probabilistic approach in the large *N* limit will **not** select the true generating model in this limit

Information theory argument: Any alternative way of selecting models will lead to models that compress the data less

Arguments for a probabilistic approach

Cox-type argument: Any alternative way to assign plausibilities to models must violate some of the very basic conditions in the desiderata

Dutch book-type argument: Betting on models using any alternative assignment of plausibility results in sets of bets that one would be willing to accept but that result in certain loss

Consistency argument: Any alternative that does not coincide with the probabilistic approach in the large N limit will **not** select the true generating model in this limit

Information theory argument: Any alternative way of selecting models will lead to models that compress the data less

Arguments for a probabilistic approach

Cox-type argument: Any alternative way to assign plausibilities to models must violate some of the very basic conditions in the desiderata

Dutch book-type argument: Betting on models using any alternative assignment of plausibility results in sets of bets that one would be willing to accept but that result in certain loss

Consistency argument: Any alternative that does not coincide with the probabilistic approach in the large *N* limit will **not** select the true generating model in this limit

Information theory argument: Any alternative way of selecting models will lead to models that compress the data less

Arguments for a probabilistic approach

Cox-type argument: Any alternative way to assign plausibilities to models must violate some of the very basic conditions in the desiderata

Dutch book-type argument: Betting on models using any alternative assignment of plausibility results in sets of bets that one would be willing to accept but that result in certain loss

Consistency argument: Any alternative that does not coincide with the probabilistic approach in the large N limit will **not** select the true generating model in this limit

Information theory argument: Any alternative way of selecting models will lead to models that compress the data less

Desiderata

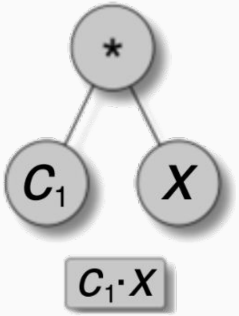
for the machine scientist

It should take arbitrary data as input and produce closed-form, interpretable mathematical models

It should be able to rigorously and quantitatively ***establish the plausibility*** of a model given some data

It should be able to systematically ***explore the space*** of all possible mathematical models, so that the stationary distribution of visited models is (at least asymptotically) given by their plausibility

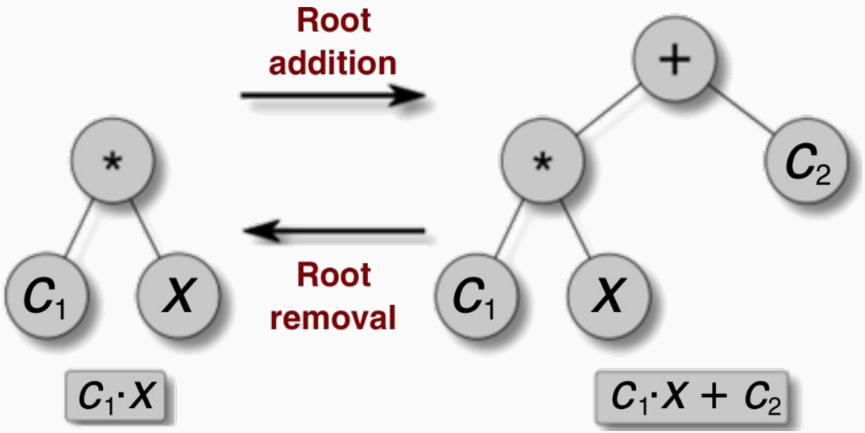
Exploring the space of models
A Metropolis-Hastings algorithm for sampling mathematical expressions



Guimera et al., *Science Advances* (2020)

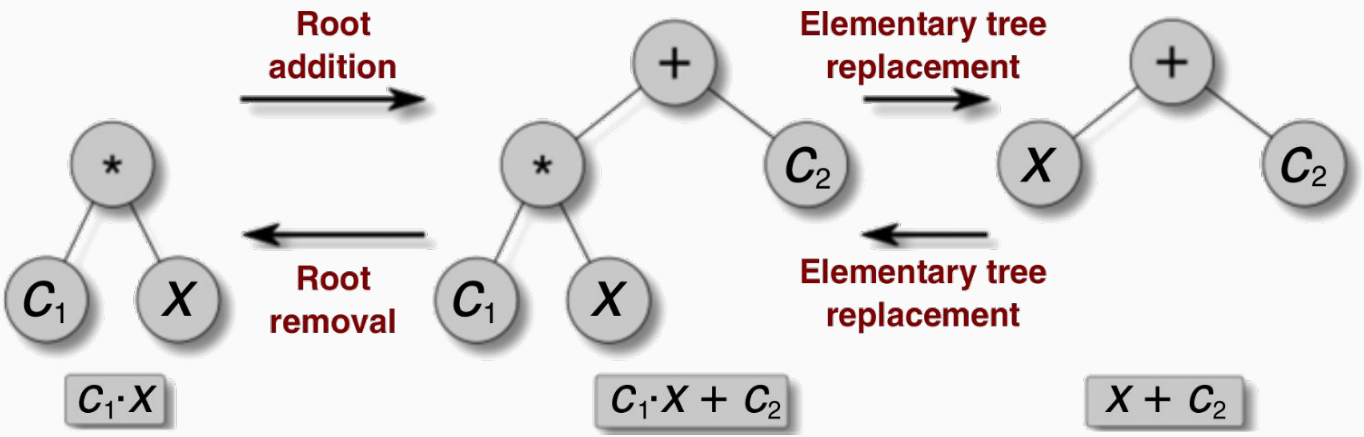
Exploring the space of models

A Metropolis-Hastings algorithm for sampling mathematical expressions



Guimera et al., *Science Advances* (2020)

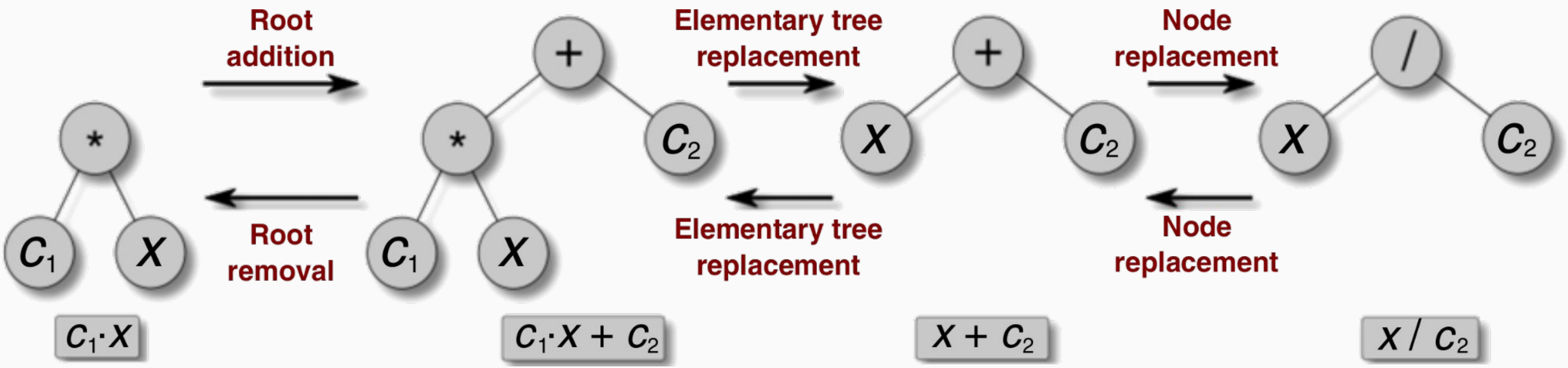
Exploring the space of models
A Metropolis-Hastings algorithm for sampling mathematical expressions



Guimera et al., *Science Advances* (2020)

Exploring the space of models

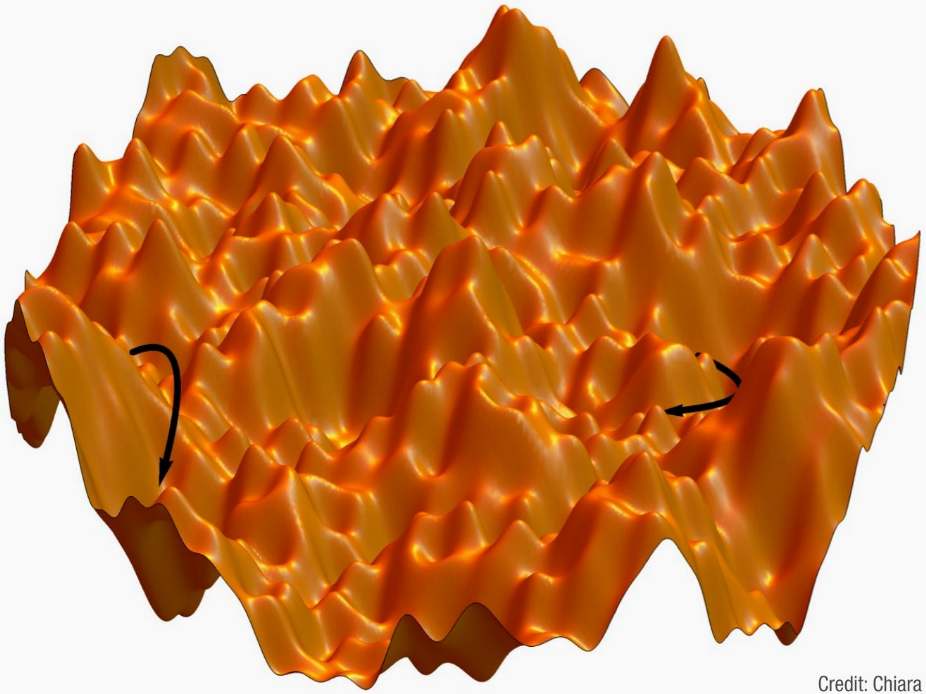
A Metropolis-Hastings algorithm for sampling mathematical expressions



Guimera et al., *Science Advances* (2020)

Exploring the space of models

A Metropolis-Hastings algorithm for sampling mathematical expressions



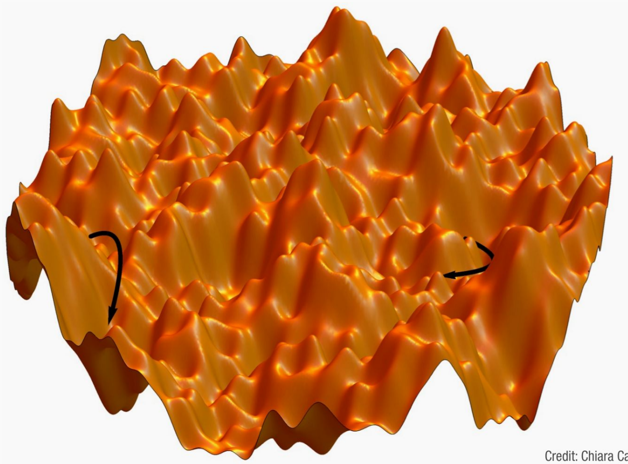
Credit: Chiara Cammarota

All in all, we have defined our Bayesian machine scientist

It establishes the plausibility of any model by means of the posterior (i.e. description length)

$$\mathcal{L}(M,D) = \frac{B(M)}{2} - \log p(M)$$

It explores the space of models and samples models from their posterior using Metropolis-Hastings



Credit: Chiara Cammarota

Standard symbolic regression vs Bayesian machine scientist

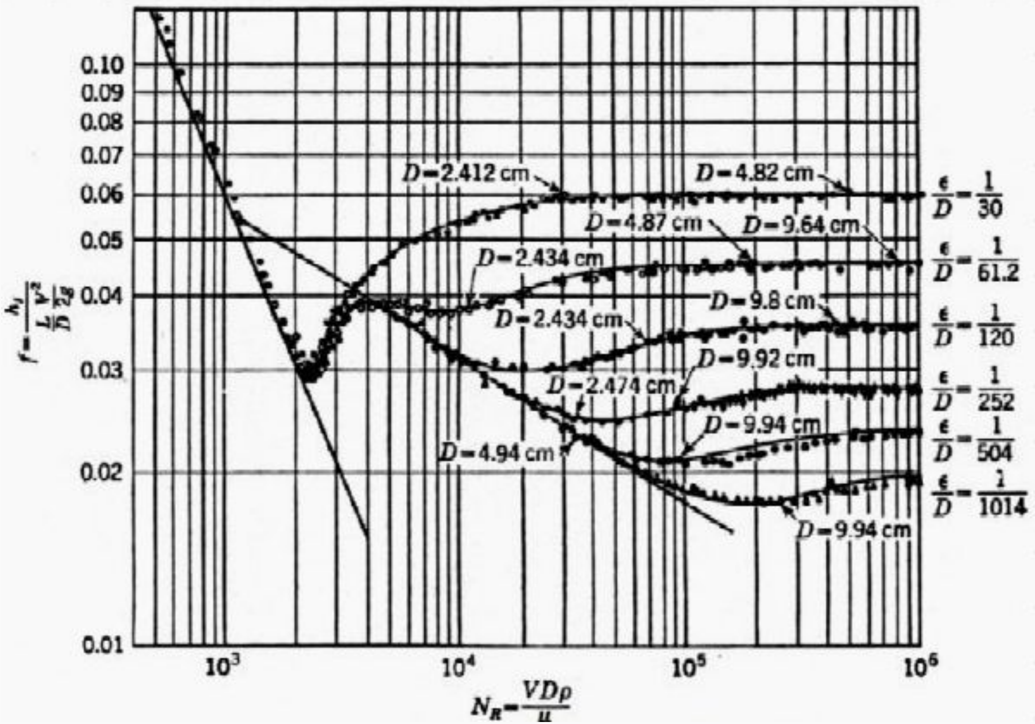
Standard symbolic regression and equation discovery	Bayesian machine scientist
Need to define goodness of fit (or loss) measure	Maximum a posterior (i.e. minimum description length) imposed by probability theory
Need to penalize model complexity heuristically	Need to specify a prior, but at least the assumptions we are making are explicit and transparent
Need to balance goodness of fit and model complexity	Goodness and complexity are balanced automatically
Heuristic exploration of the space of possible models	We sample from the posterior

So, does it work?

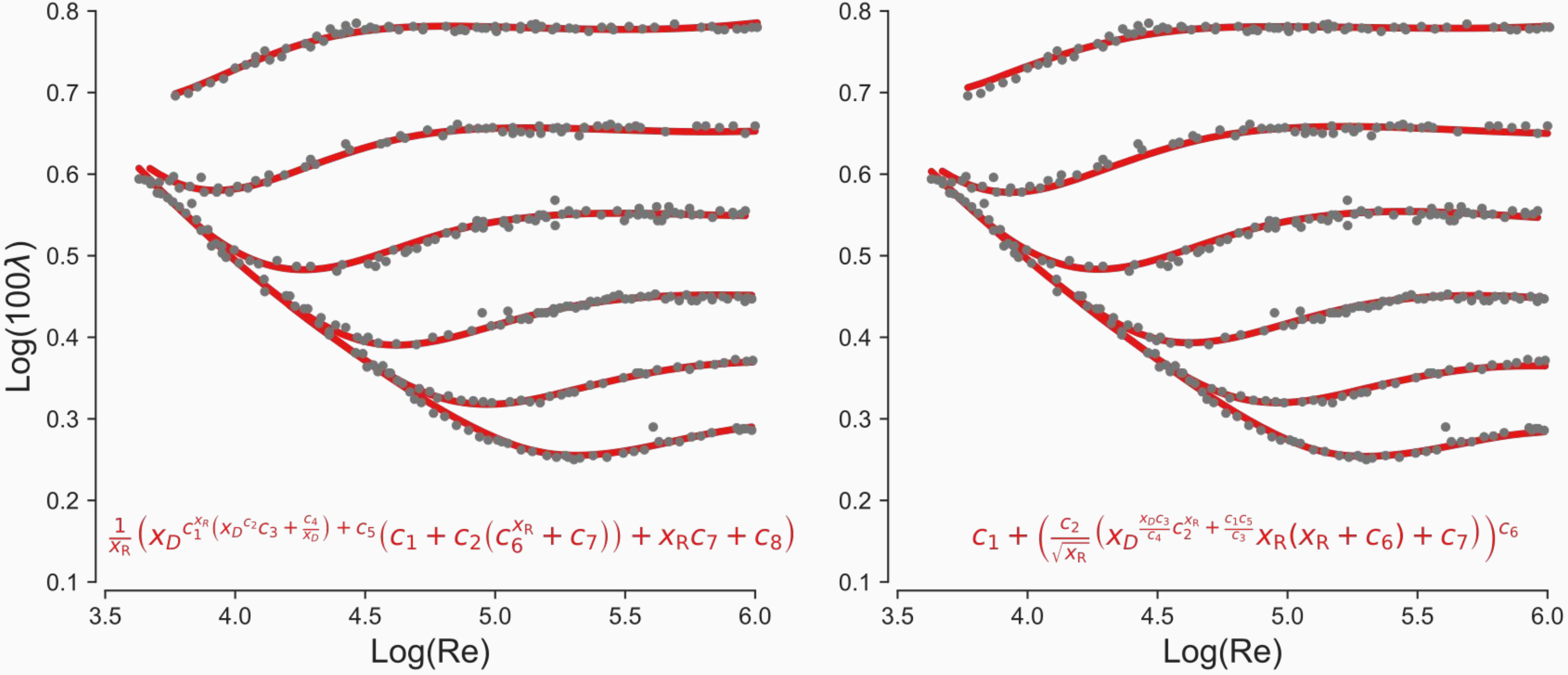
We generate synthetic data and see if the machine scientist is able to recover the correct model

Guimera et al., *Science Advances* (2020)

Nikuradse's 1933 experiments about friction in rough pipes

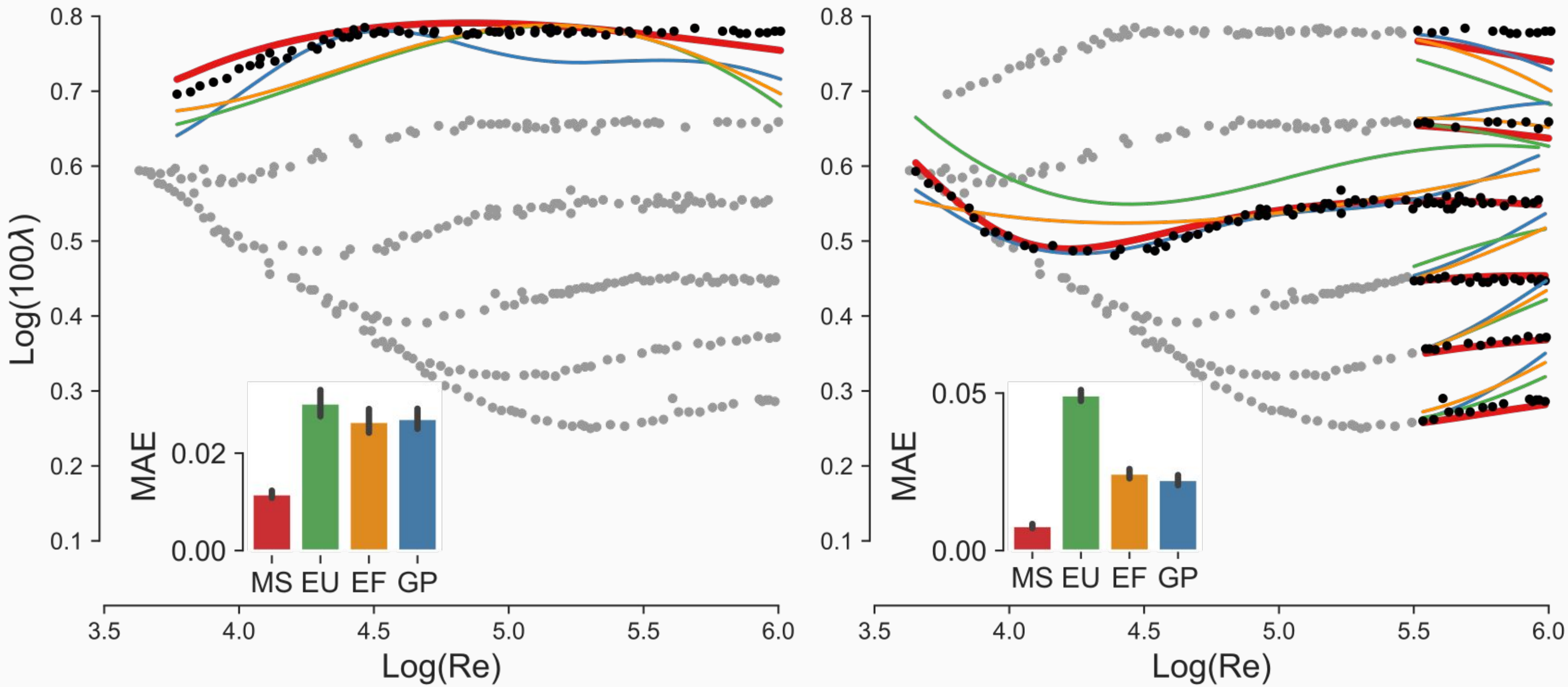


The machine scientist finds multiple expressions that describe the data well



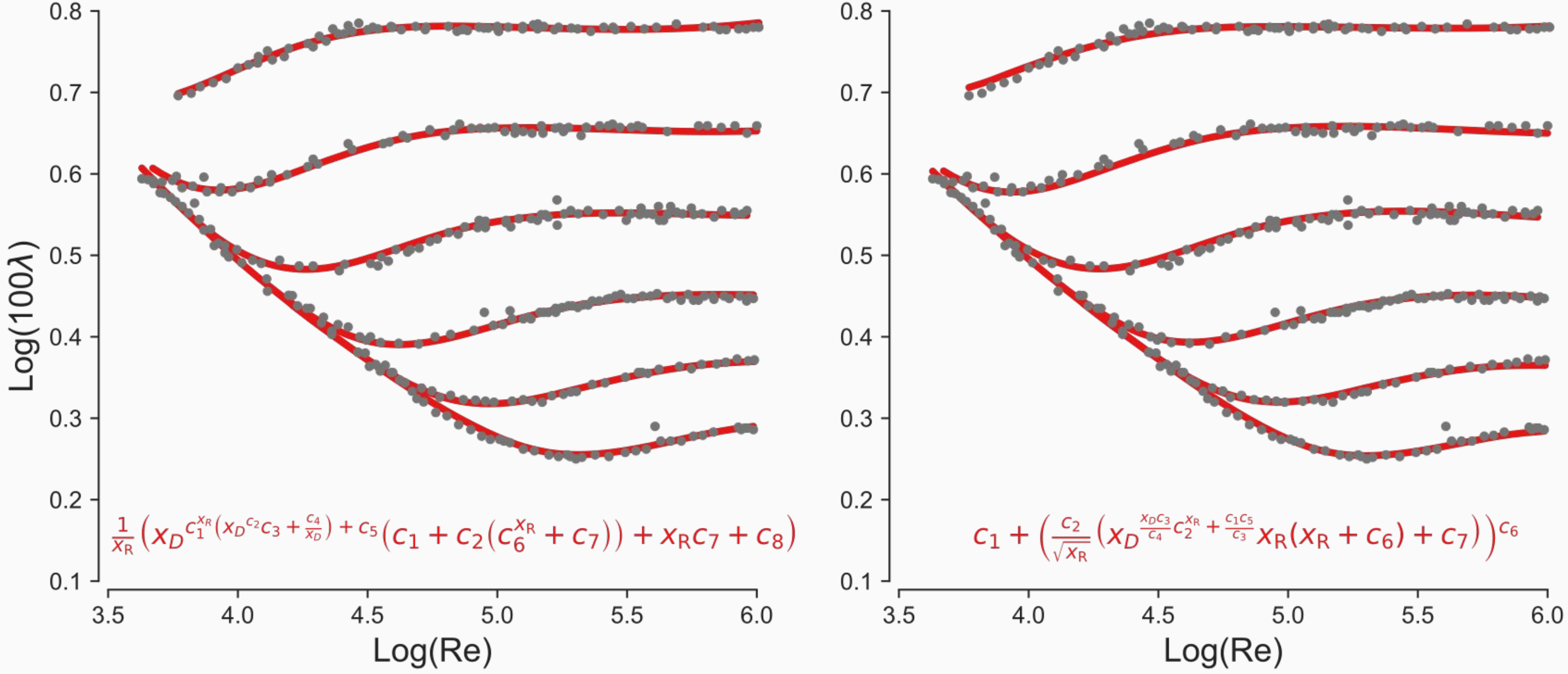
Guimera et al., *Science Advances* (2020)
Reichardt et al., *Phys. Rev. Lett* (2020)

The machine scientist is also able to make accurate predictions for unobserved data



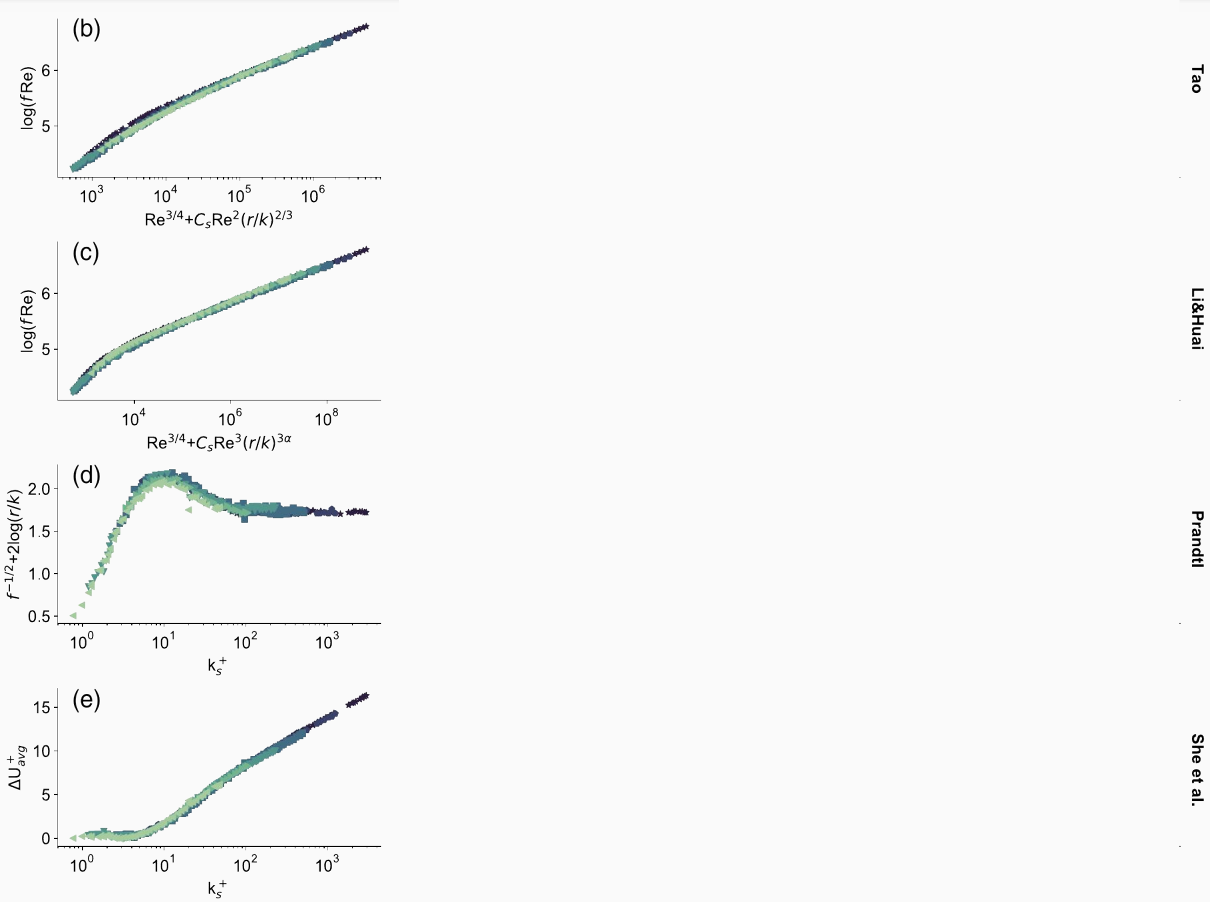
Guimera et al., *Science Advances* (2020)
Reichardt et al., *Phys. Rev. Lett* (2020)

The machine scientist finds multiple expressions that describe the data well



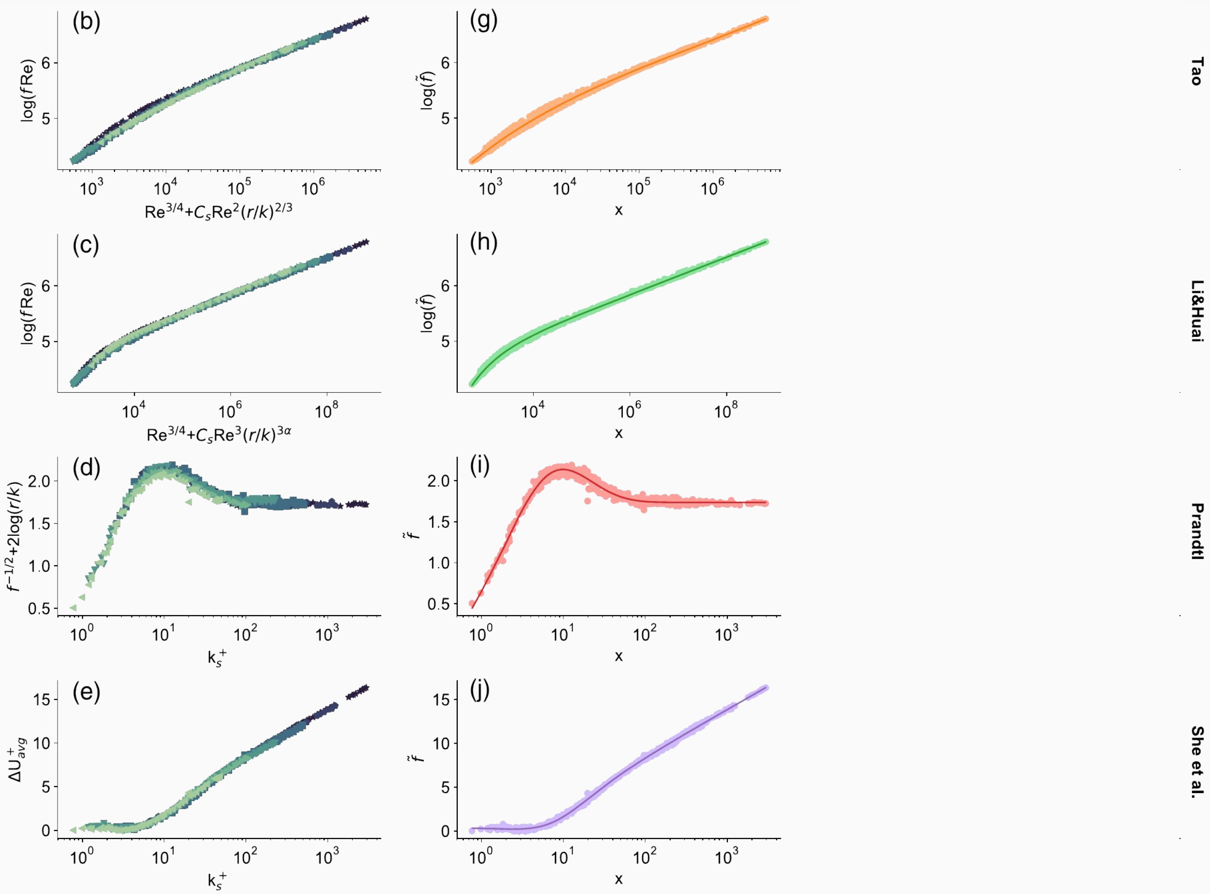
Guimera et al., *Science Advances* (2020)
Reichardt et al., *Phys. Rev. Lett* (2020)

When combined with physical knowledge, the machine scientist provides insight



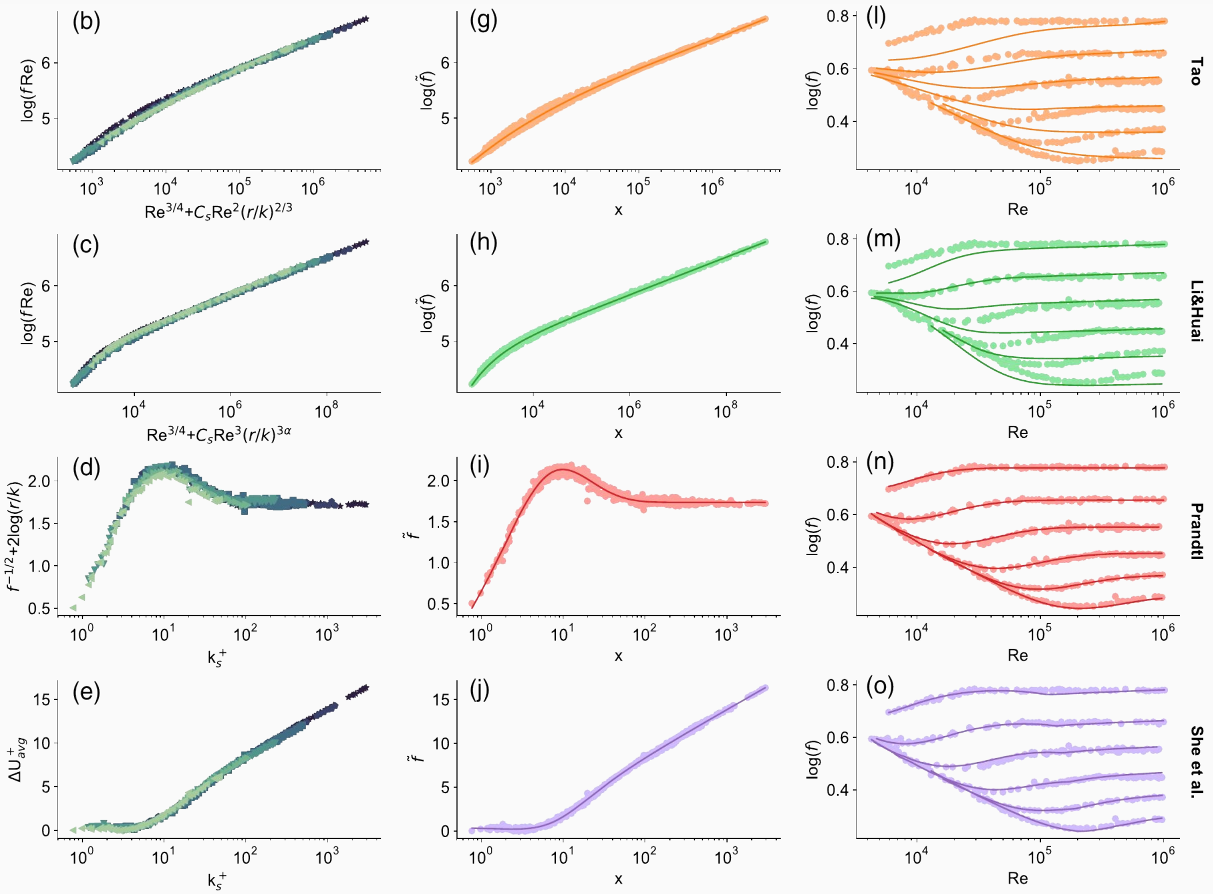
Reichardt et al., *Phys. Rev. Lett* (2020)

When combined with physical knowledge, the machine scientist provides insight



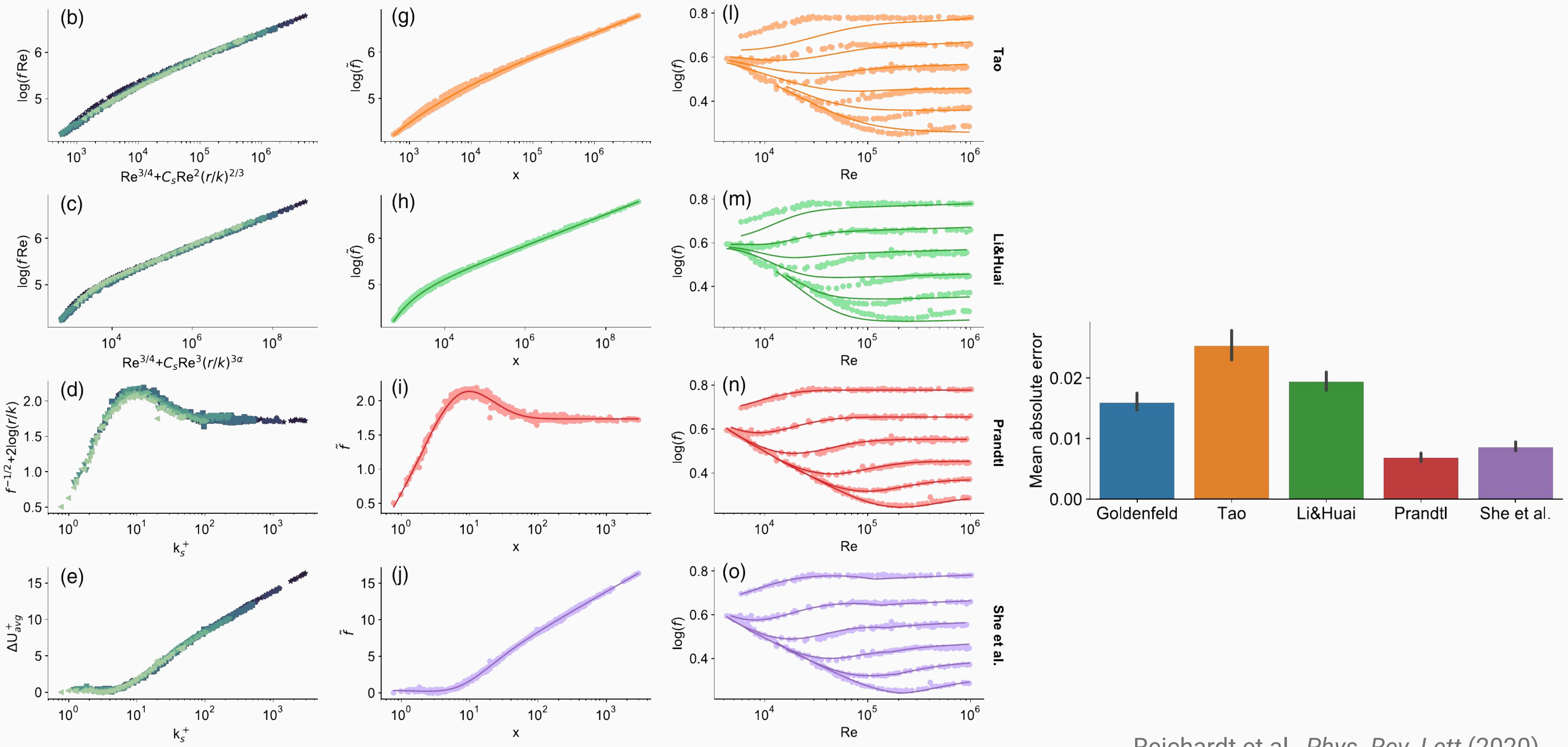
Reichardt et al., *Phys. Rev. Lett* (2020)

When combined with physical knowledge, the machine scientist provides insight



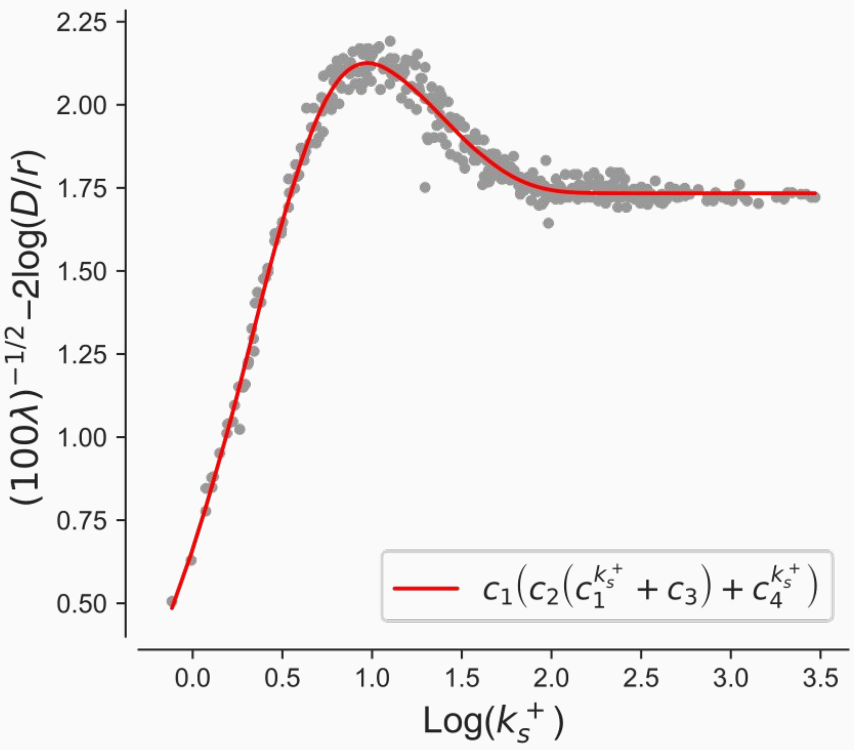
Reichardt et al., *Phys. Rev. Lett* (2020)

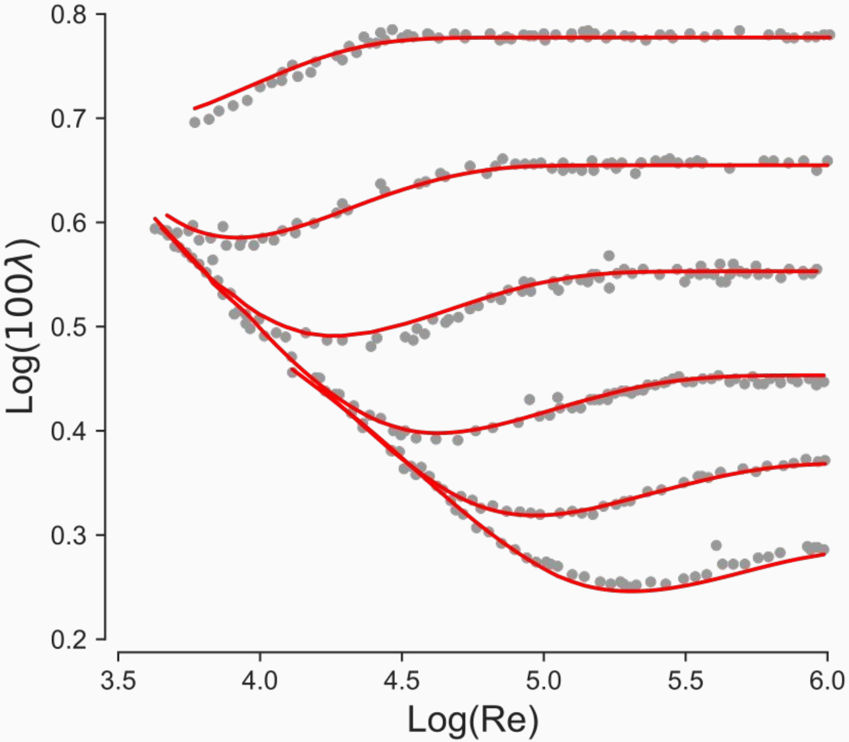
When combined with physical knowledge, the machine scientist provides insight



Reichardt et al., *Phys. Rev. Lett* (2020)

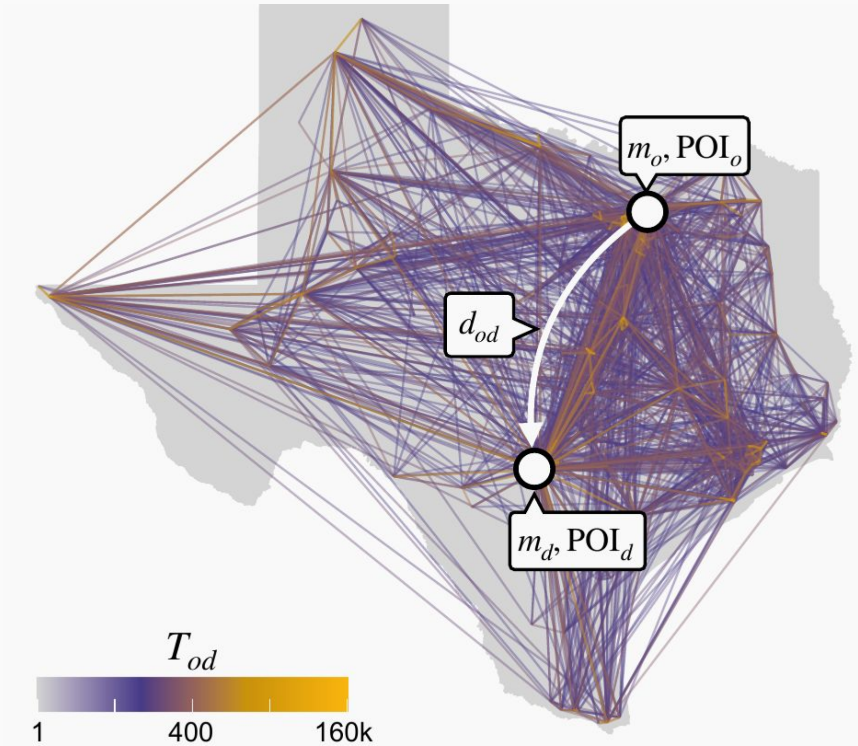
When combined with physical knowledge, the machine scientist provides insight





Guimera et al., *Science Advances* (2020)
Reichardt et al., *Phys. Rev. Lett* (2020)

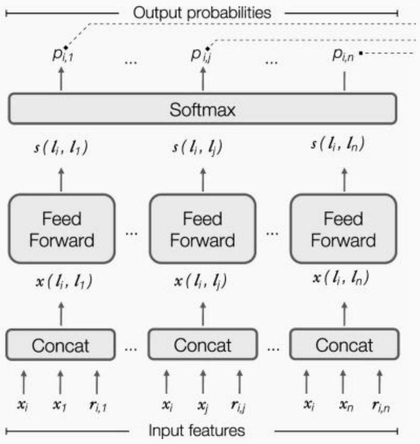
Can we find models that predict human mobility flows?



Gravity models

$$T_{od} = A \frac{m_o m_d}{d^\alpha}$$

"Deep gravity" model



Simini et al., *Nature Comm.* (2021)

The models identified by the Bayesian machine scientist are *gravity-like*

A

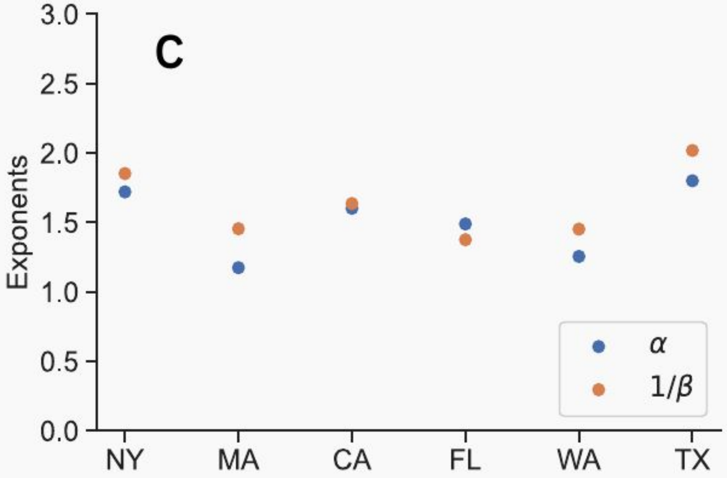
$$\log T_{od} = A\left(1 + \frac{B((m_d + C)(m_o + D))^{\beta}}{d}\right)^{\xi}$$

	A	B($/10^{-6}$)	C ($/10^{2}$)	D ($/10^{4}$)	ξ	β
New York	4.27	441	1.76	1.51	0.26	0.54
Massachusetts	6.79	9.15	144	11.0	0.28	0.69
California	21.43	20.2	92.0	34.8	0.50	0.61
Florida	2.66	6.87	231	2.26	0.33	0.73
Washington	3.68	17.9	64.2	4.09	0.24	0.69
Texas	4.10	1240	0.612	1.79	0.30	0.50

B

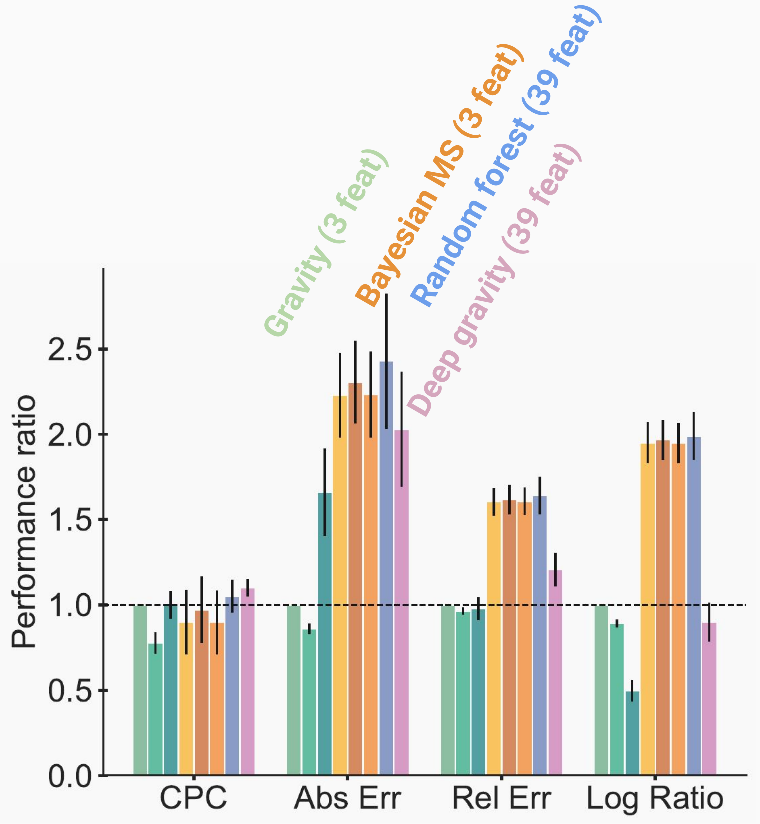
$$\log T_{od} = \log\left(A\left(\frac{B(m_d m_o + C m_d + D)}{d^{\alpha}} + 1\right)^{\gamma}\right)$$

	A	B ($/10^{-9}$)	C ($/10^{4}$)	D ($/10^{8}$)	α	γ
New York	86.96	289	1.02	1.03	1.72	0.97
Massachusetts	68.08	8.50	5.28	27.8	1.17	1.78
California	105.7	27.3	2.43	5.90	1.60	2.02
Florida	58.14	99.2	2.07	4.29	1.49	1.33
Washington	89.07	33.7	2.47	6.10	1.26	1.41
Texas	75.94	278	1.85	3.43	1.80	1.16



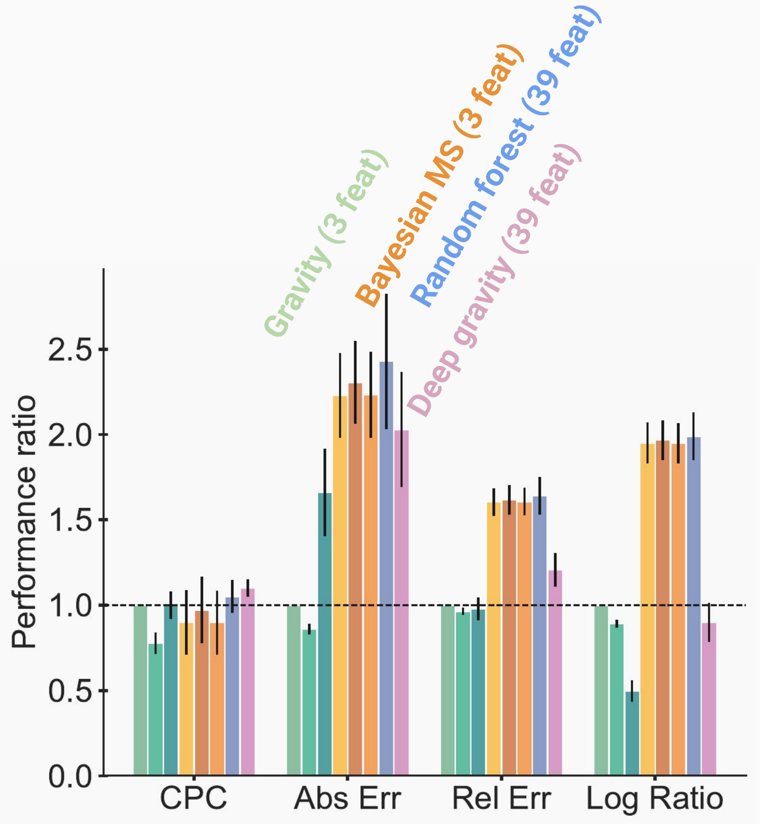
Cabanas-Tirapu et al., *submitted* (2024)

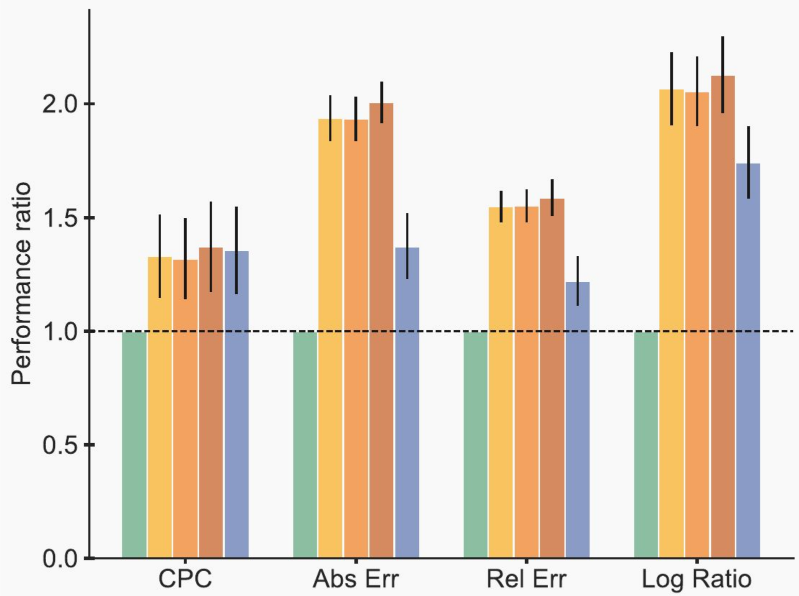
These gravity-like models are as predictive as “black box” machine learning approaches…



Cabanas-Tirapu et al., *submitted* (2024)

These gravity-like models are as predictive as “black box” machine learning approaches… and extrapolate significantly better





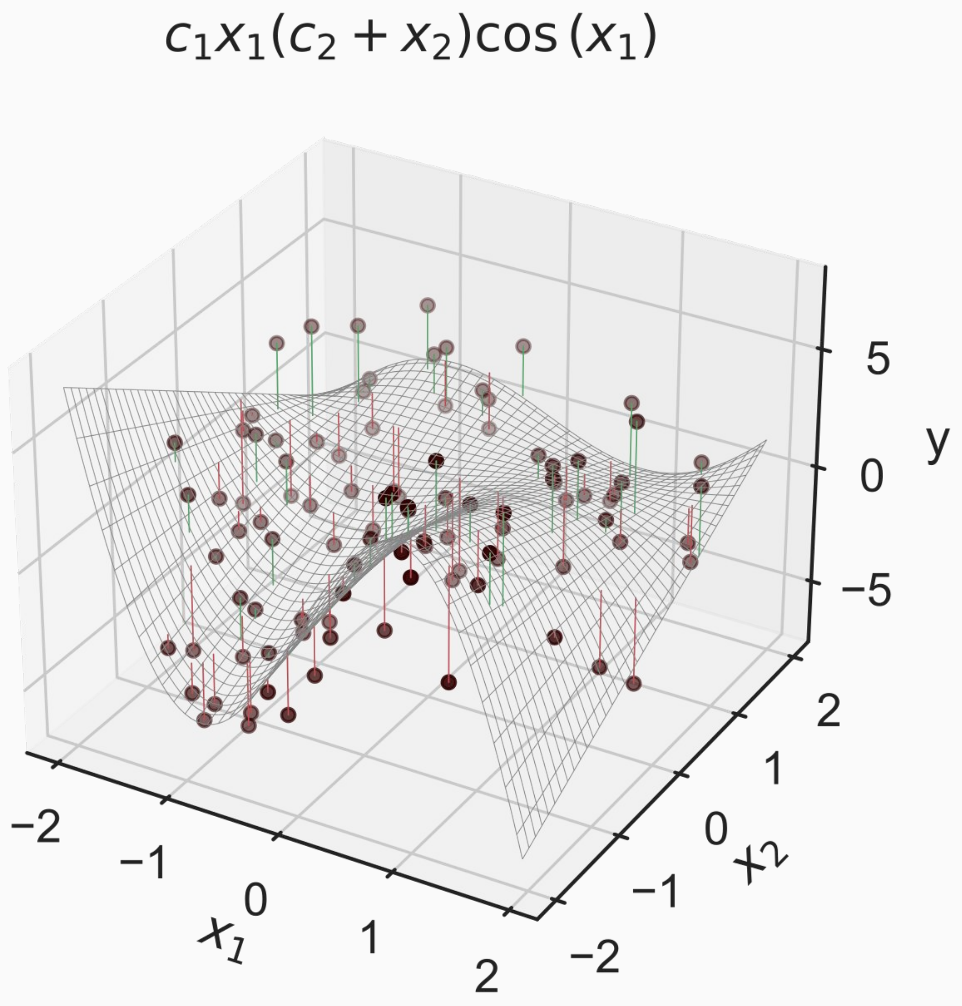
Cabanas-Tirapu et al., *submitted* (2024)

Is it always possible to learn the true generating model?

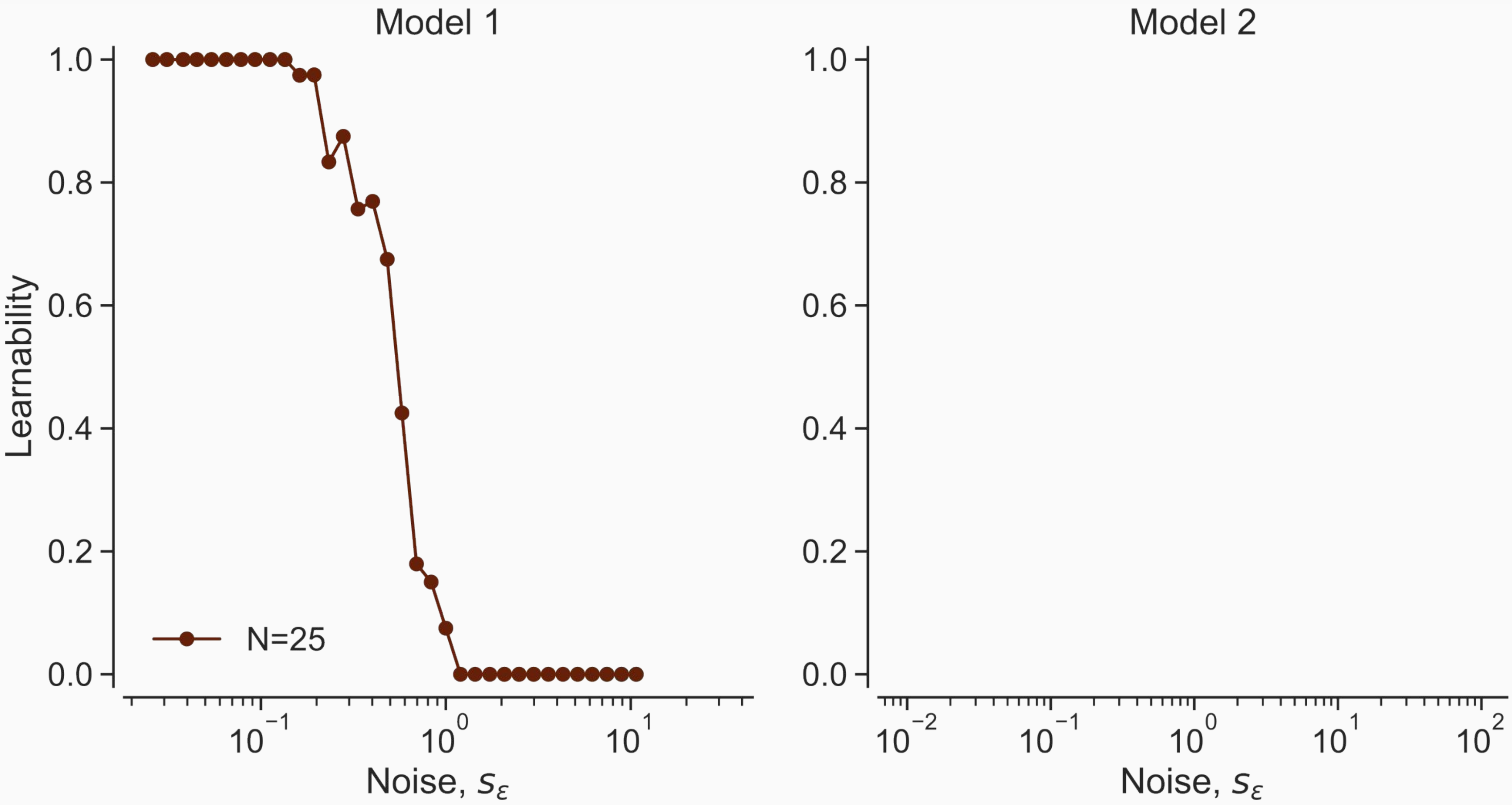
Intuition

With *enough information*, we should be able to recover the true generating model

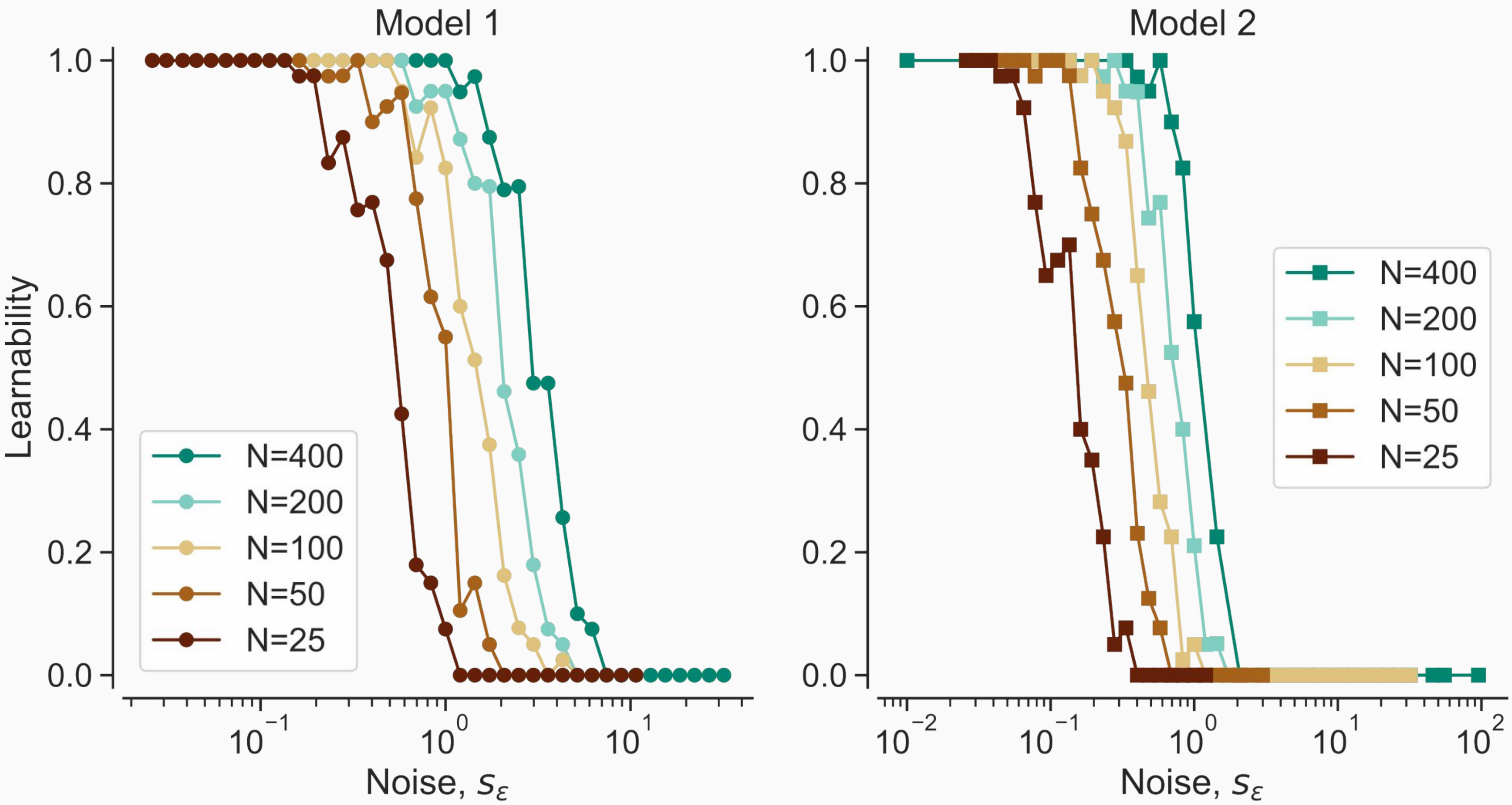
But, for a fixed number of points, if the *noise grows*, the true model will eventually become unlearnable



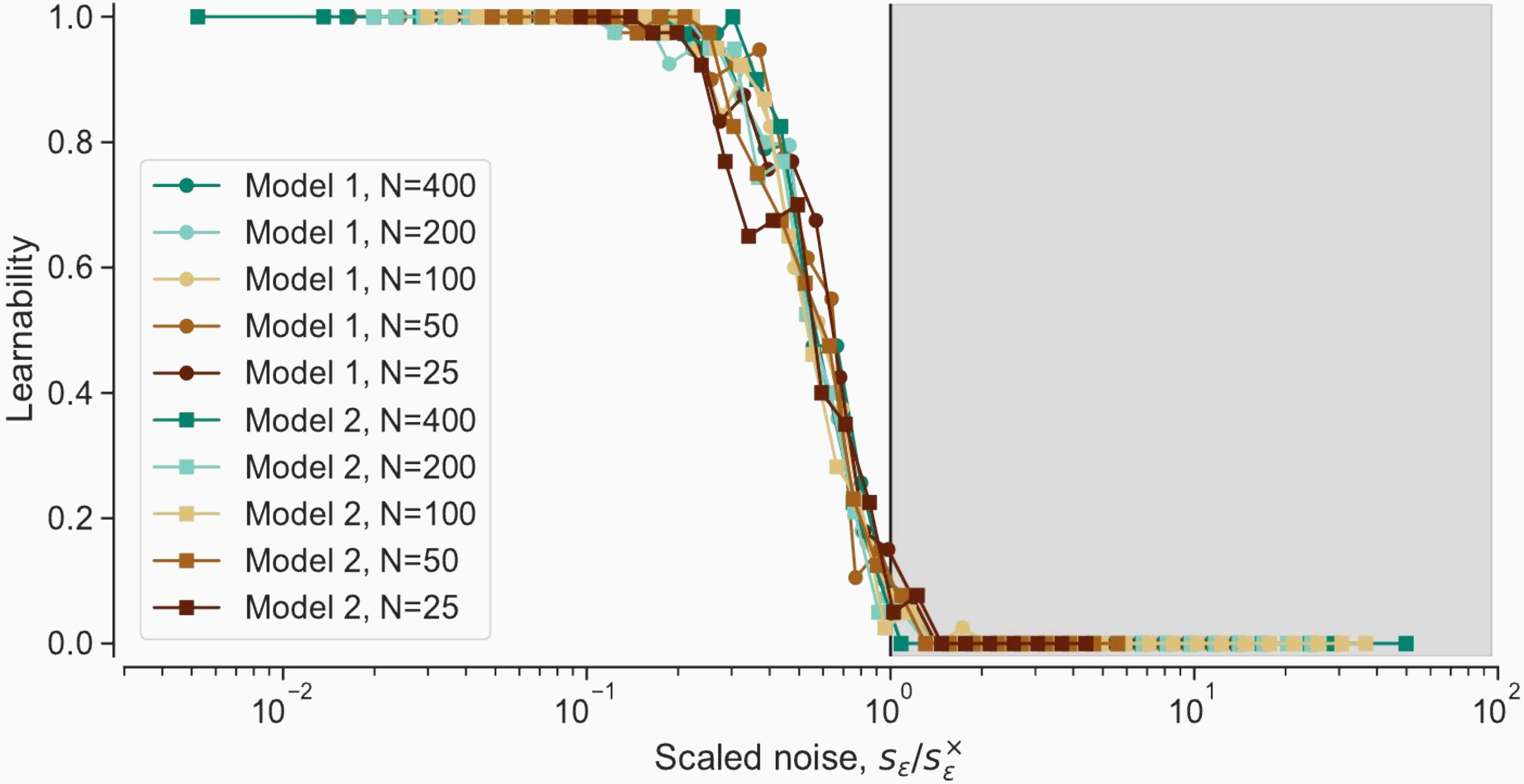
We observe a learnability transition



We observe a learnability transition



Once noise is scaled, all curves collapse: universal behavior?

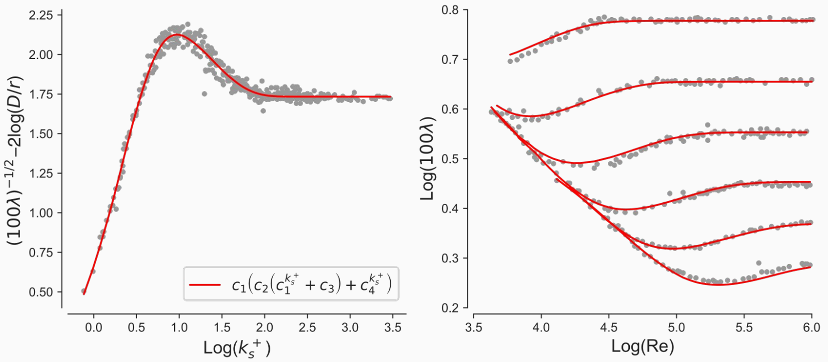


Fajardo-Fontiveros, Reichardt, et al., *Nature Comm. (2023)*

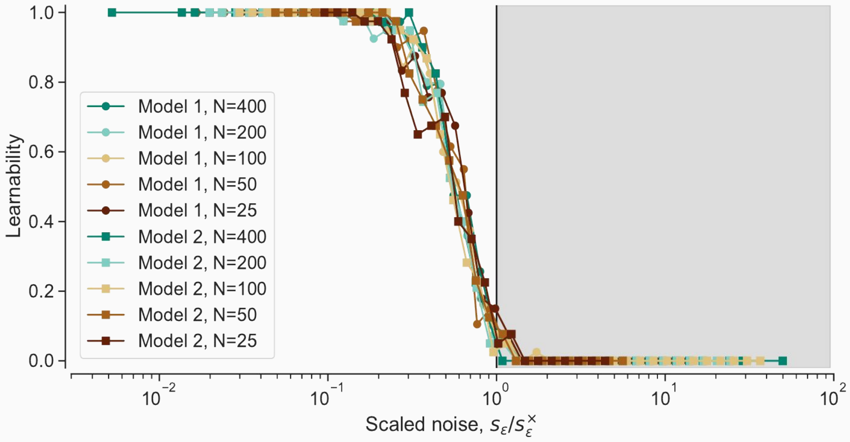
http://seeslab.info Twitter: @sees_lab

Conclusions

We can identify closed-form mathematical models from data using a Bayesian approach to symbolic regression…



…but there are fundamental and universal limits to our ability to do so



http://seeslab.info Twitter: @sees_lab

Thank you

A. Aguilar-Mogas, O. Cabanas-Tirapu, A., H. R. De Los Rios, J. Duch, O. Fajardo-Fontiveros, G. Guillén-Gosálbez, F. A. Massucci, M. Miranda, E. Moro, V. Negri, J. Pallarès, I. Reichardt, **M. Sales-Pardo**, D. Vázquez

ICREA

Universitat Rovira i Virgili

Gobierno de España — Ministerio de Ciencia e Innovación

Generalitat de Catalunya

Papers: