DERIVABLE SCIENTIFIC DISCOVERY

CORNELIO CRISTINA - SAMSUNG AI

IN COLLABORATION WITH IBM RESEARCH: *SANJEEB DASH, VERNON AUSTEL, TYLER R. JOSEPHSON, JOAO GONCALVES, KENNETH CLARKSON, NIMROD MEGIDDO, BACHIR EL KHADIR, LIOR HORESH*

COLLOQUIUM PR[AI]RIE - OCTOBER 19, 2023

DERIVABLE SCIENTIFIC DISCOVERY

GOAL:
Discovering meaningful laws of nature in symbolic form from experimental data

Extraction of formulas that fit the data:

- *NN and statistical regression:*
    - good for discovery of patterns and relations in data
    - drawback: "black-box" models
- *Standard regression:*
    - the functional form is given, discovery = parameter fitting
- *Symbolic regression:*
    - the functional form is *not* given but is instead composed from the data
    - models are more "interpretable" and require less data

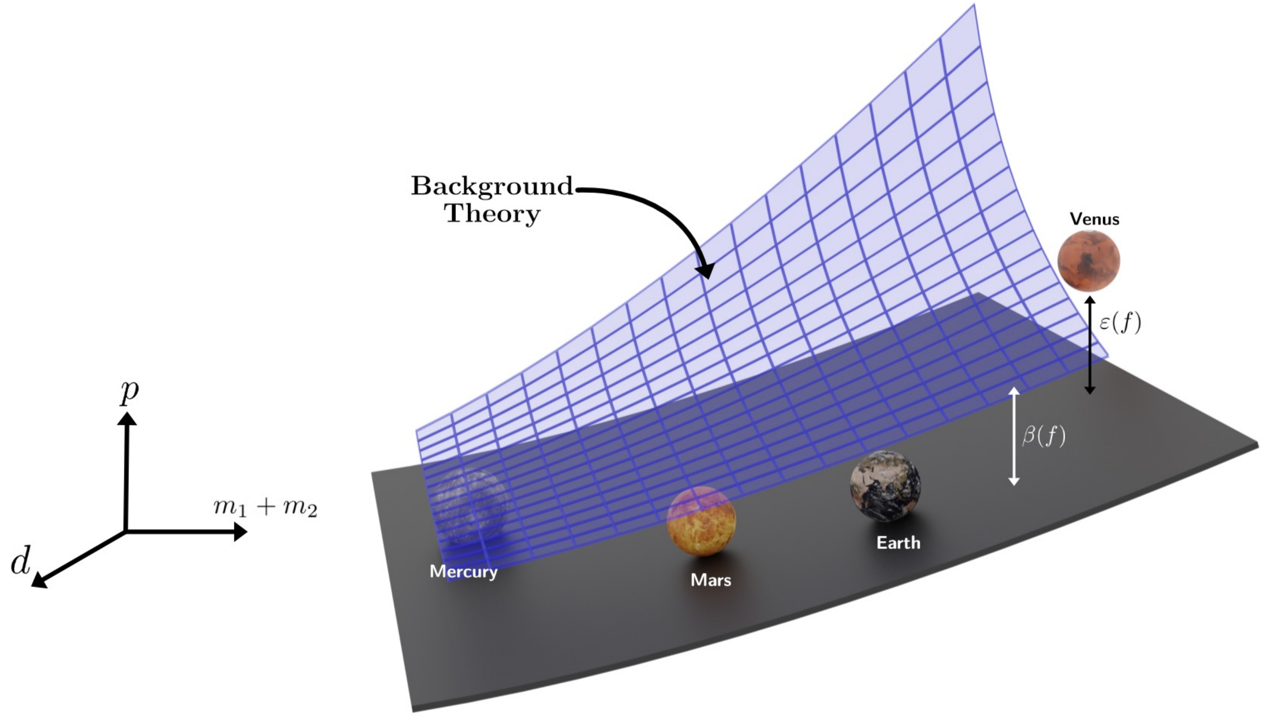
Discovery of scientifically meaningful formulas:

- Many expressions can be extracted for a given dataset, but not all are **consistent** with the known **background theory**
- Models that are derivable, and not merely empirically accurate, are appealing because they are arguably correct, predictive, and insightful

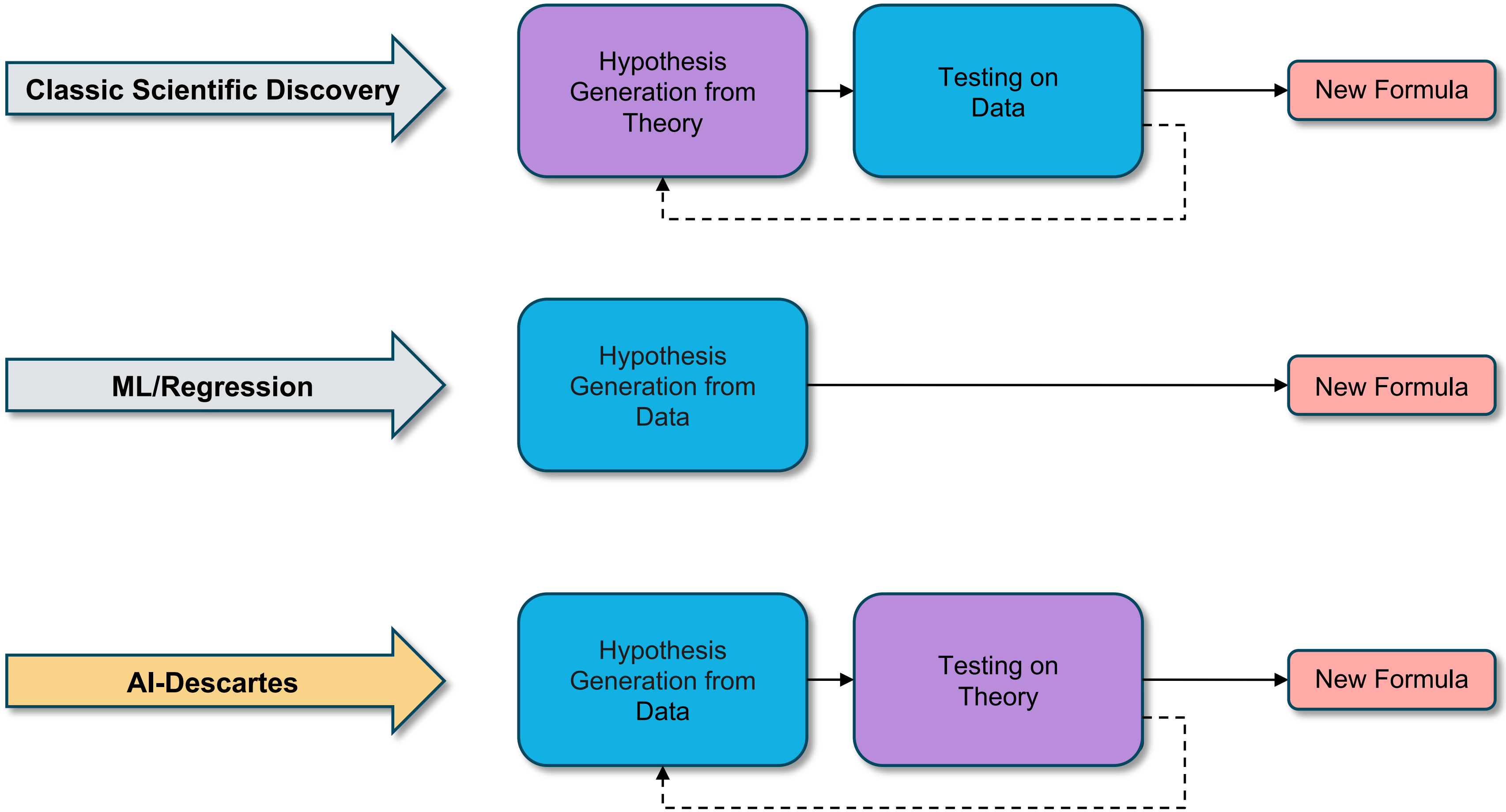
DERIVABLE SCIENTIFIC DISCOVERY

IDEA: unification of explicit *symbolic model extraction* from numerical data with *formal reasoning*

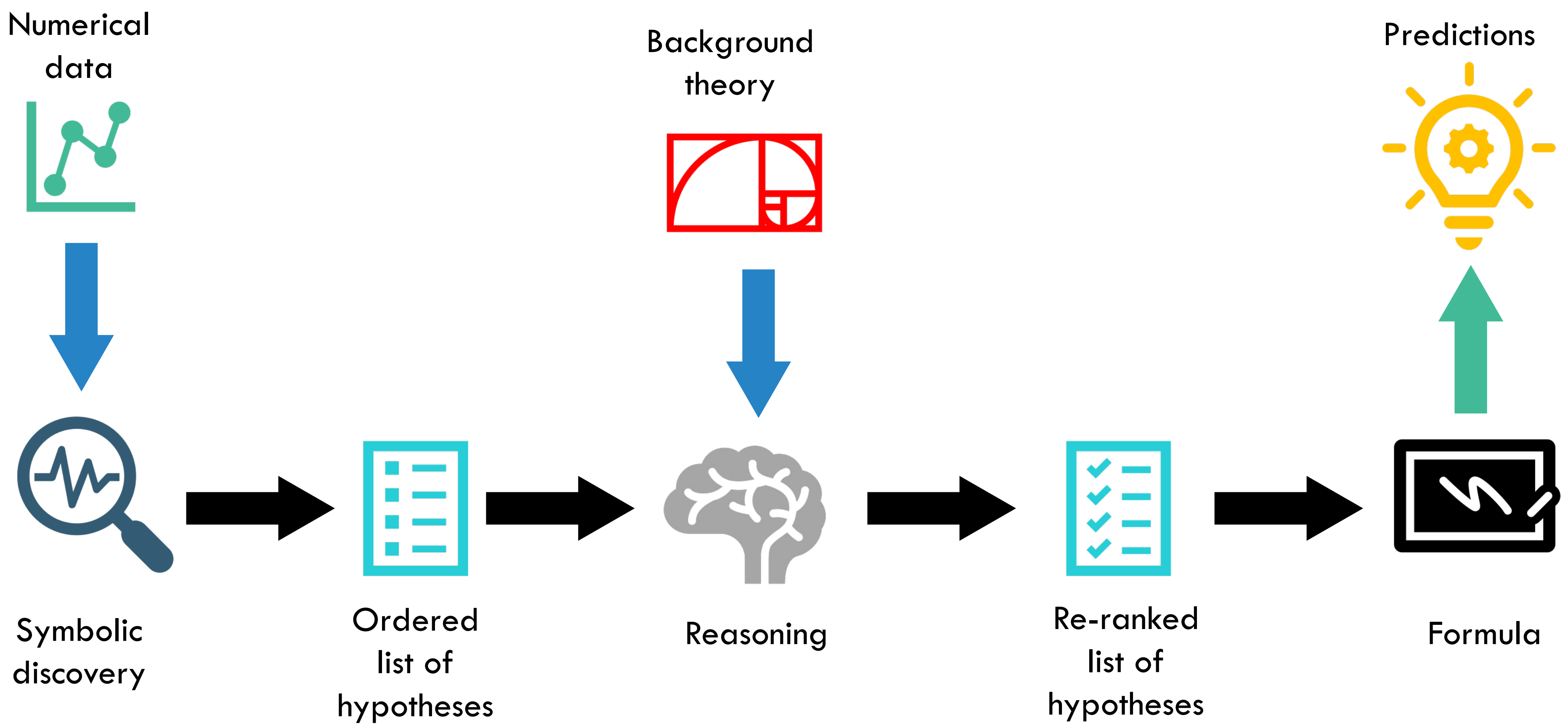
- *Verification capabilities:* providing a formal proof of the derivability of a formula produced from the data
- When not derivable: providing measures that indicate how close a formula is to a derivable formula.



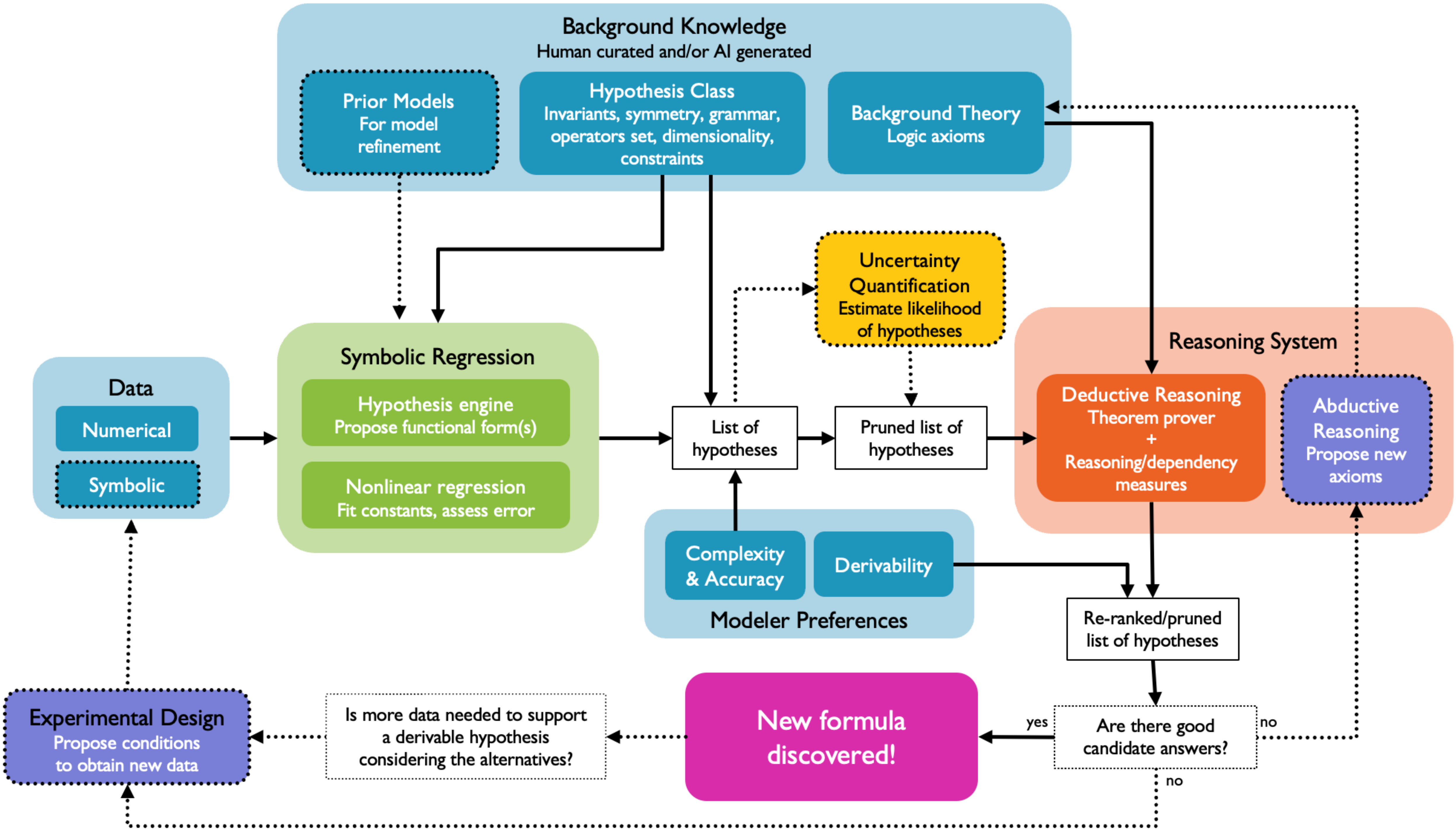
SCIENTIFIC METHODS



System overview



Given observational data, associated with a process or phenomenon, the goal is to discover an interpretable, universal, mathematical model in a symbolic form.


Background Knowledge
Human curated and/or AI generated
Prior Models
For model refinement
Hypothesis Class
Invariants, symmetry, grammar, operators set, dimensionality, constraints
Background Theory
Logic axioms
Uncertainty Quantification
Estimate likelihood of hypotheses
Reasoning System
Data
Numerical
Symbolic
Symbolic Regression
Hypothesis engine
Propose functional form(s)
Nonlinear regression
Fit constants, assess error
List of hypotheses
Pruned list of hypotheses
Deductive Reasoning
Theorem prover
+
Reasoning/dependency measures
Abductive Reasoning
Propose new axioms
Complexity & Accuracy
Derivability
Modeler Preferences
Re-ranked/pruned list of hypotheses
Experimental Design
Propose conditions to obtain new data
Is more data needed to support a derivable hypothesis considering the alternatives?
New formula discovered!
yes
Are there good candidate answers?
no
no

System overview

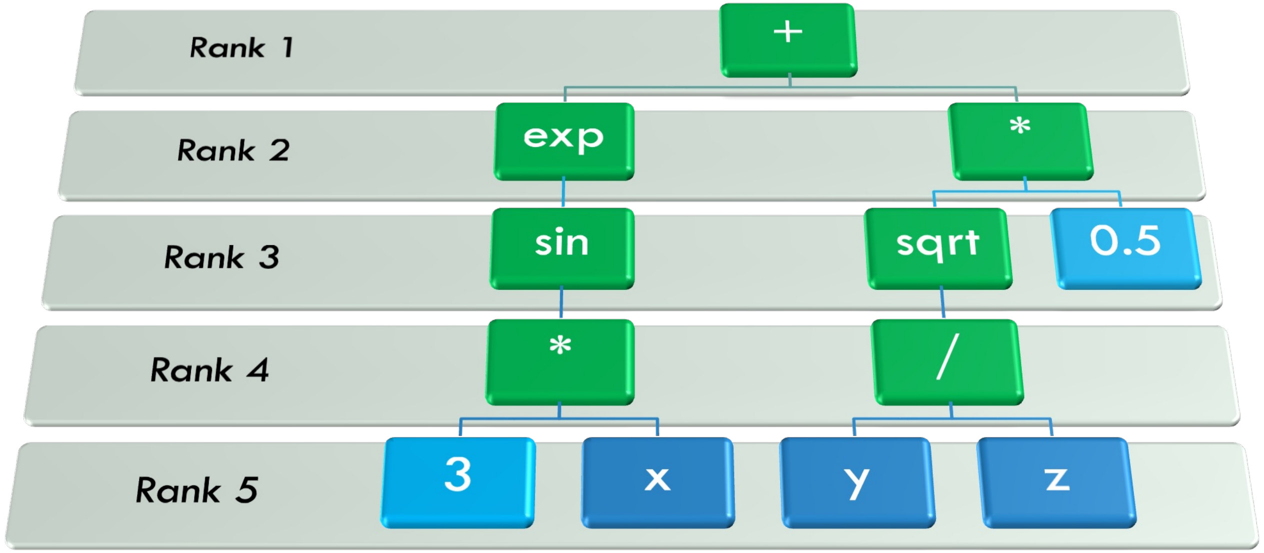
INPUT: 4-tuples $\langle B, C, D, M \rangle$

- **Background Knowledge B:** domain-specific axioms (logic formulae)
  - *Completeness* assumption: B contains all the axioms necessary to comprehensively explain the phenomena under consideration
  - *Consistency* assumption: the axioms do not contradict one another
- **A Hypothesis Class C:**
  - Grammar (set of admissible symbolic models)
  - Invariance constraints (e.g., $X + Y$ is equivalent to $Y + X$)
  - Constraints on the functional form (e.g., monotonicity)
- **Data D:** a set of *n* examples
- **Modeler Preferences M:** a set of numerical parameters (e.g., error bounds on accuracy)

SYMBOLIC REGRESSION

Define **grammar** so that every meaningful mathematical expression is a **sentence** of a **formal language** comprising:

- **Operators**
- **Variables**
- **Constants**



$$f = exp(sin(3 * x)) + 0.5 * sqrt(y/z)$$

$$f = e^{\sin(3x)} + \frac{1}{2}\sqrt{\frac{y}{z}}$$

The **grammar** is **encoded** as a set of (decision) **variables** and **constraints**

Free-form search in the space of the sentences for ones that **honors the data** while **minimizing** expression **complexity**

SYMBOLIC REGRESSION

The optimization problem can be articulated at a high level as:

$$
\begin{aligned}
&\min_{s} && D(s(x), y) && \textit{fidelity} \\
&s.t. \begin{cases} \\ \\ \\ \\ \end{cases} && C(s) \leq \tau && \textit{complexity} \\
& && parse(x) \in \mathcal{T} && \textit{structural (grammar)} \\
& && s \in \Sigma^* && \textit{symbols (primitive) choice} \\
& && s \in I && \textit{invariances} \\
& && \mathcal{N}\big(s(x)\big) \leq \delta && \textit{non-linear numerical expression}
\end{aligned}
$$

Where:

- D - error model
- $x \in \mathcal{X}$ - set of training datum
- $y \in \mathcal{Y}$ - set of targets associated with x
- Σ - set of admissible symbols
- Σ^* - set of words in the language
- $\mathcal{T}$ - a proper tree structure
- $s \in \Sigma^*$ and $\mathrm{parse}(x) \in \mathcal{T}$ implies that $\mathrm{s} \in \mathcal{L}(\mathcal{T}, \Sigma)$ (belongs to the formal language)
- C - a measure of complexity
- $\mathcal{N}$ - numerical function
- I - set of invariants

REASONING CAPABILITIES

1 – Derivability

Verify a formula from a set of axioms defining the background theory

2 – Reasoning measures

Relative/Absolute error between a formula (induced from data) and *the variable of interest* which represents a derivable formula deducible from the axioms

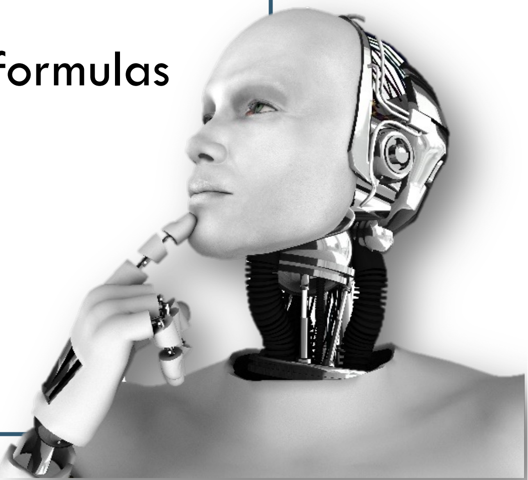
3 – Counterexample generation

Generation of new points that violates the current candidate formula, starting from the axioms

4 – Constraints pre-processing

Checking which of the candidate formulas support a set of constraints on the functional form:

- Monotonicity condition
- Conditions at the limit
- etc.



DERIVABILITY

Example:

Formula extracted from data

$$f = p/(0.709 \cdot p + 0.157)$$

Formula to prove – Direct derivability

$$(C \land A) \rightarrow p/(0.709 \cdot p + 0.157)$$

Formula to prove – Existential derivability

$$\exists c_1\ c_2\ (C \land A) \rightarrow p/(c_1 \cdot p + c_2)$$

Two types of derivability

Direct derivability:

$$(C \land A) \rightarrow f$$

Existential derivability:

$$\exists c_1 \ldots c_n\ (C \land A) \rightarrow (f\,' \land C\,')$$

- f is replaced by f ′ by introducing new existentially quantified variables for each numerical element in f.

Where:

- C = constraints
- A = axioms,
- C U A = background theory
- f = the formula we wish to prove

REASONING ERRORS

Pointwise reasoning error

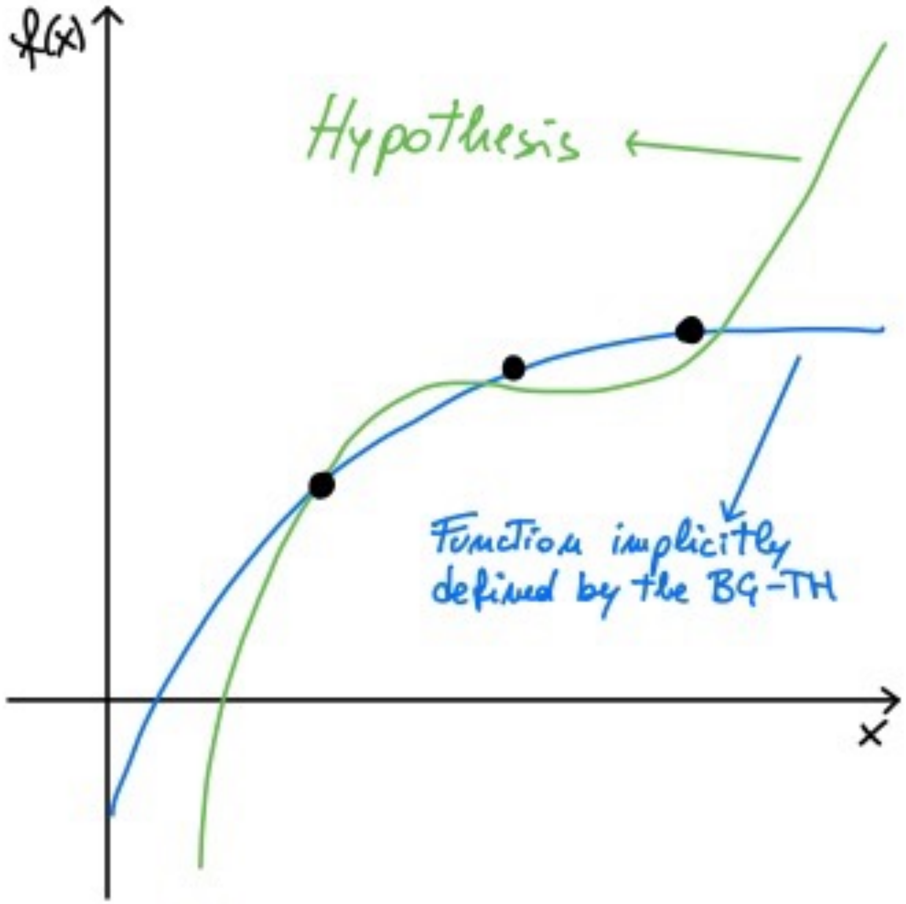
Generalized reasoning error

Dependency analysis

Distance between:

- *A formula generated from the numerical data*
- *The derivable formula that is* ***implicitly defined*** *by the axiom set*

Note: *The derivable formula is defined by the variable of interest is not given explicitly, but only* implicitly defined in the background theory



REASONING ERRORS

Pointwise reasoning error

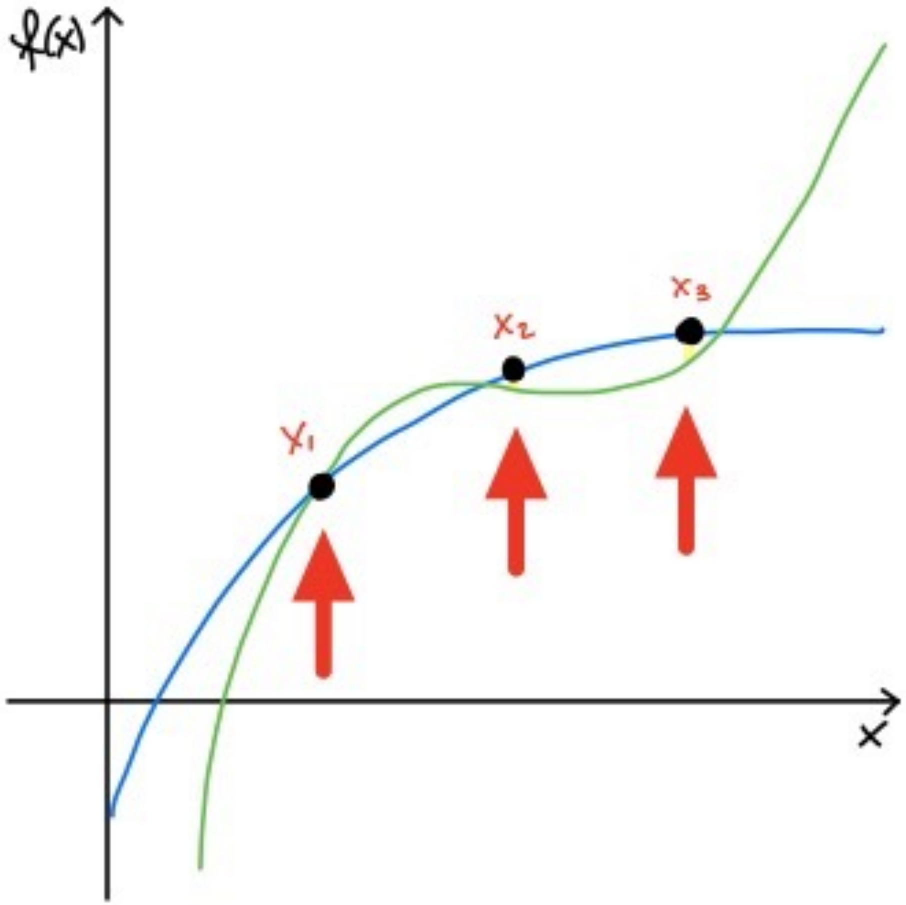
Generalized reasoning error

Dependency analysis

Pointwise reasoning error

- measured by the l_2 or l_∞ norm (holds for other norms as well)
- applied to the differences between the values of the numerically-derived formula and a derivable formula at the points in the dataset.

$$\beta_2^r = \sqrt{\sum_{i=1}^{m}\left(\frac{f(\mathbf{X}^i) - f_{\mathscr{B}}(\mathbf{X}^i)}{f_{\mathscr{B}}(\mathbf{X}^i)}\right)^2}$$



REASONING ERRORS

Pointwise reasoning error

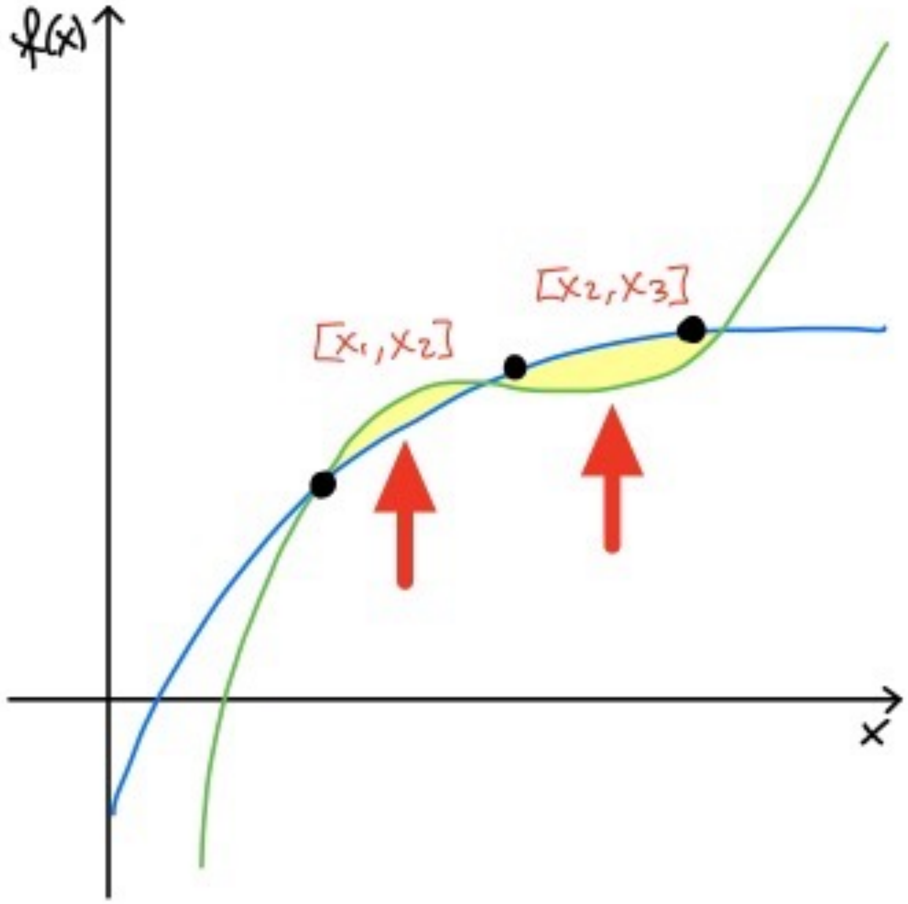
Generalized reasoning error

Dependency analysis

Generalization reasoning error

- Consider not only the specific datapoints but the interval where specific data points lie in
- Evaluate how much a formula generalizes between the data points

$$\beta_{\infty}^{r} = \max_{1 \le i \le m} \left\{ \frac{|f(\mathbf{X}^i) - f_{\mathscr{B}}(\mathbf{X}^i)|}{|f_{\mathscr{B}}(\mathbf{X}^i)|} \right\}$$



REASONING ERRORS

Pointwise reasoning error

Generalized reasoning error

Dependency analysis

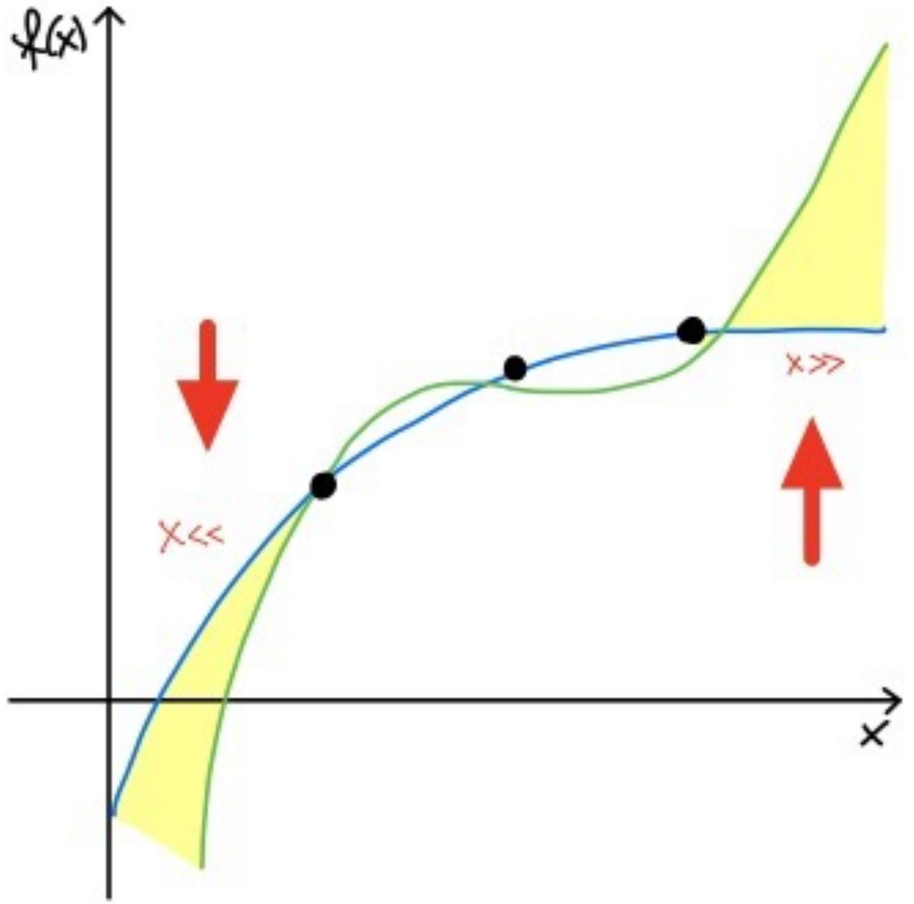
Variable dependence

Extension of the intervals beyond the order of magnitude of the data points in the dataset.

- Check if a formula generalizes even outside the space defined by the dataset

Done by:

- increasing (or diminishing) the interval end (or start) point by one order of magnitude for a variable at a time.



SHOW CASES

Kepler's third law of planetary motion

Langmuir's adsorption equation

Einstein's time-dilation formula

Challenges:

- **Real data** (noise)
- **Few data points** (~10 points)
- **Kepler**: The masses involved are of very different magnitudes.
- **Langmuir**: Background theory contains material-dependent coefficients
- **Einstein**: Different background theories: Newtonian and relativistic

EXPERIMENTS SETUP

Symbolic regression

BARON as MINLP solver

- Supported operators:
  - $+, -, \times, /$, exp, log
- Supported L-tree depth = 4 (~7 parsing tree)

Reasoning

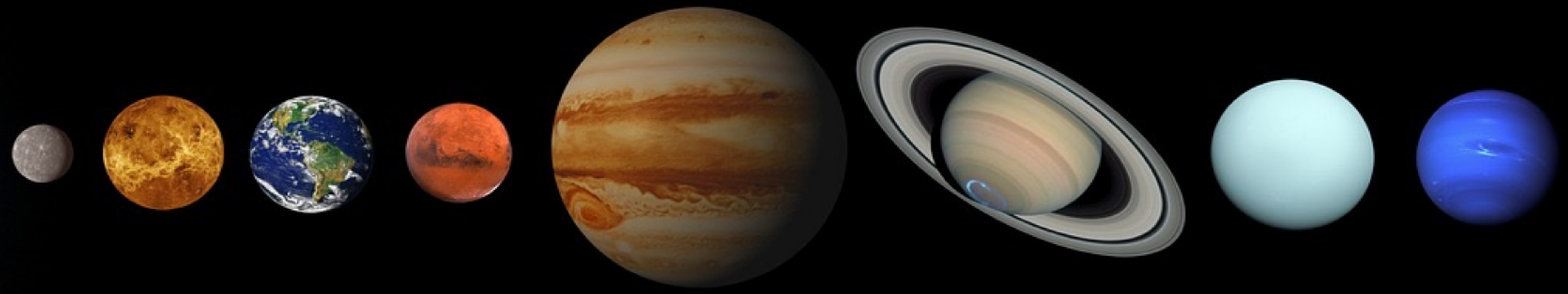
KeYmaera as reasoning tool

- ATP for hybrid systems, which combines different types of reasoning: deductive, real-algebraic, and computer-algebraic reasoning.
- has an underlying CAD system

Mathematica for certain types of analysis of symbolic expressions

- e.g. constraints checking

KEPLER'S THIRD LAW OF PLANETARY MOTION



Kepler's law captures the relationship between the distance between two bodies and their orbital periods

$$p = \sqrt{\frac{4\pi^2 d^3}{G(m_1 + m_2)}}$$

- p is the orbital period;
- G is the gravitational constant;
- m_1 and m_2 are the masses of the two bodies (e.g., the sun and a planet in the solar system)

KEPLER'S THIRD LAW OF PLANETARY MOTION

 Solar system	$\sqrt{0.1319d^3}$	Reasoning
	$(0.03765d^3 + d^2)/(2+d)$	
	$\sqrt{0.1316(d^3+d)}$	
 Exoplanets	$\sqrt{0.1319d^3/m_1}$	
	$\sqrt{m_1^2 m_2^3/d + 0.1319d^3/m_1}$	
	$\sqrt{(1-0.7362m_1)d^3/2}$	
 Binary stars	$1/(d^2m_1^2) + 1/(dm_2^2) - m_1^3m_2^2 + \sqrt{0.4787d^3/m_2 + d^2m_2^2}$	
	$\left(\sqrt{d^3} + m_1^3 m_2/\sqrt{d}\right)/\sqrt{m_1+m_2}$	
	$\boxed{\sqrt{d^3/(0.9967m_1+m_2)}}$	

KEPLER'S THIRD LAW OF PLANETARY MOTION

Background theory

K1. center of mass definition	K1. $m_1 * d_1 = m_2 * d_2$
K2. distance between bodies	K2. $d = d_1 + d_2$
K3. gravitational force	K3. $F_g = \frac{G m_1 m_2}{d^2}$
K4. centrifugal force	K4. $F_c = m_2 d_2 w^2$
K5. force balance	K5. $F_g = F_c$
K6. period definition	K6. $p = \frac{2\pi}{w}$
K7. non-negativity constraints	K7. $m_1 > 0,\ m_2 > 0,\ p > 0,\ d_1 > 0,\ d_2 > 0$.

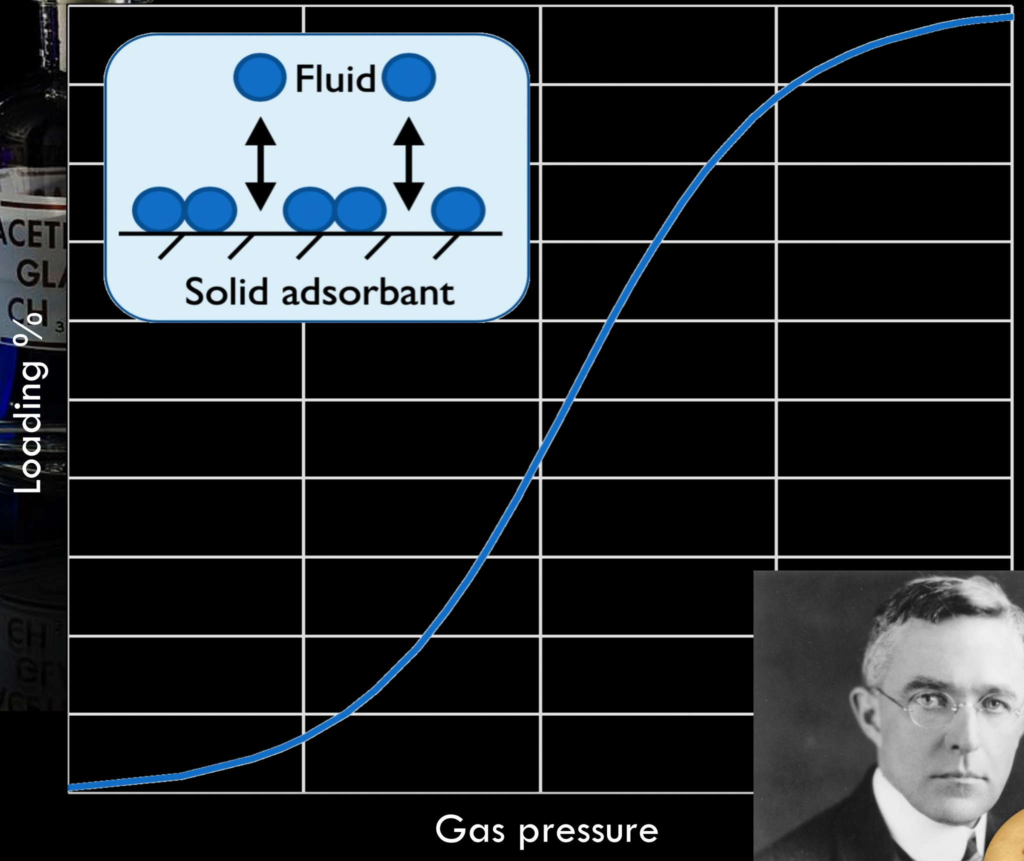
KEPLER'S THIRD LAW OF PLANETARY MOTION

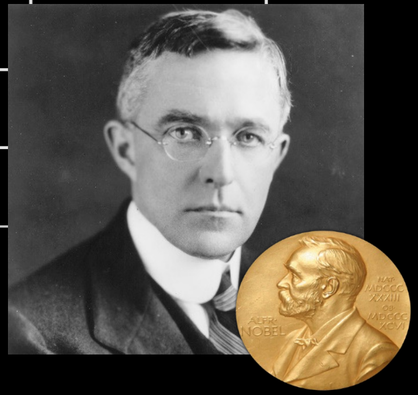
1	2	3	4	5	6	7	8	9	10
Dataset	Candidate formula $p =$	numerical error ε_2^r	ε_∞^r	point. reas. err. β_2^r	β_∞^r	gen. reas. error $\beta_{\infty,S}^r$	dependencies m_1	m_2	d
solar	$\boxed{\sqrt{0.1319 \cdot d^3}}$	.01291	.006412	.0146	.0052	.0052	0	0	1
	$\sqrt{0.1316 * (d^3 + d)}$	1.9348	1.7498	1.9385	1.7533	1.7559	0	0	0
	$(0.03765 d^3 + d^2)/(2 + d)$	.3102	.2766	.3095	.2758	.2758	0	0	0
exoplanet	$\boxed{\sqrt{0.1319 d^3 / m_1}}$	.08446	.08192	.02310	.0052	.0052	0	0	1
	$\sqrt{m_1^2 m_2^3 / d + 0.1319\, d^3 / m_1}$	.1988	.1636	.1320	.1097	> 550	0	0	0
	$\sqrt{(1 - .7362 m_1) d^3 / 2}$	1.2246	.4697	1.2418	.4686	.4686	0	0	1
binary stars	$1/(d^2 m_1^2) + 1/(d m_2^2) - m_1^3 m_2^2 + + \sqrt{.4787 d^3 / m_2 + d^2 m_2^2}$	.002291	.001467	.0059	.0050	timeout	0	0	0
	$(\sqrt{d^3} + m_1^3 m_2 / \sqrt{d}) / \sqrt{m_1 + m_2}$	.003221	.003071	.0038	.0031	timeout	0	0	0
	$\boxed{\sqrt{d^3 / (0.9967 m_1 + m_2)}}$	.005815	.005337	.0014	.0008	.0020	1	1	1

LANGMUIR'S ADSORPTION EQUATION

The Langmuir adsorption equation describes a chemical process in which gas molecules contact a surface, and relates the **loading on the surface** to the **pressure of the gas.**

$$\frac{q}{q_{max}} = \frac{K_a \cdot p}{1 + Ka \cdot p}$$





- ***p** is the pressure of the gas*
- **q** is loading *q* on the surface
- **q_{max}** is the maximum loading
- ***K_a** is the adsorption strength*

LANGMUIR'S ADSORPTION EQUATION

Background theory

L1.	Site balance:	$S_0 = S + S_a$
L2.	Adsorption rate model:	$r_{ads} = k_{ads} \cdot p \cdot S$
L3.	Desorption rate model:	$r_{des} = k_{des} \cdot S_a$
L4.	Equilibrium assumption:	$r_{ads} = r_{des}$
L5.	Mass balance on q	$q = S_a$

$\mathcal{K}$ CONSTRAINTS

C1. $f(0) = 0$

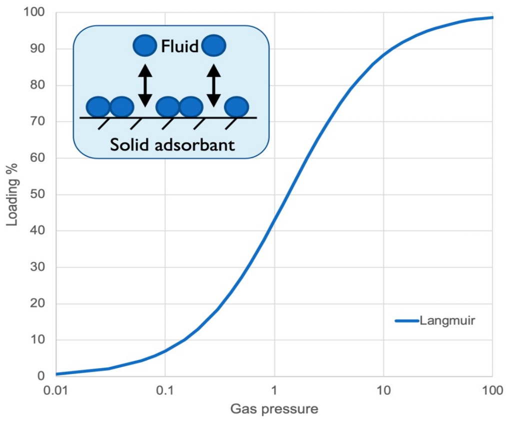
C2. $(\forall p > 0)\ (f(p) > 0)$

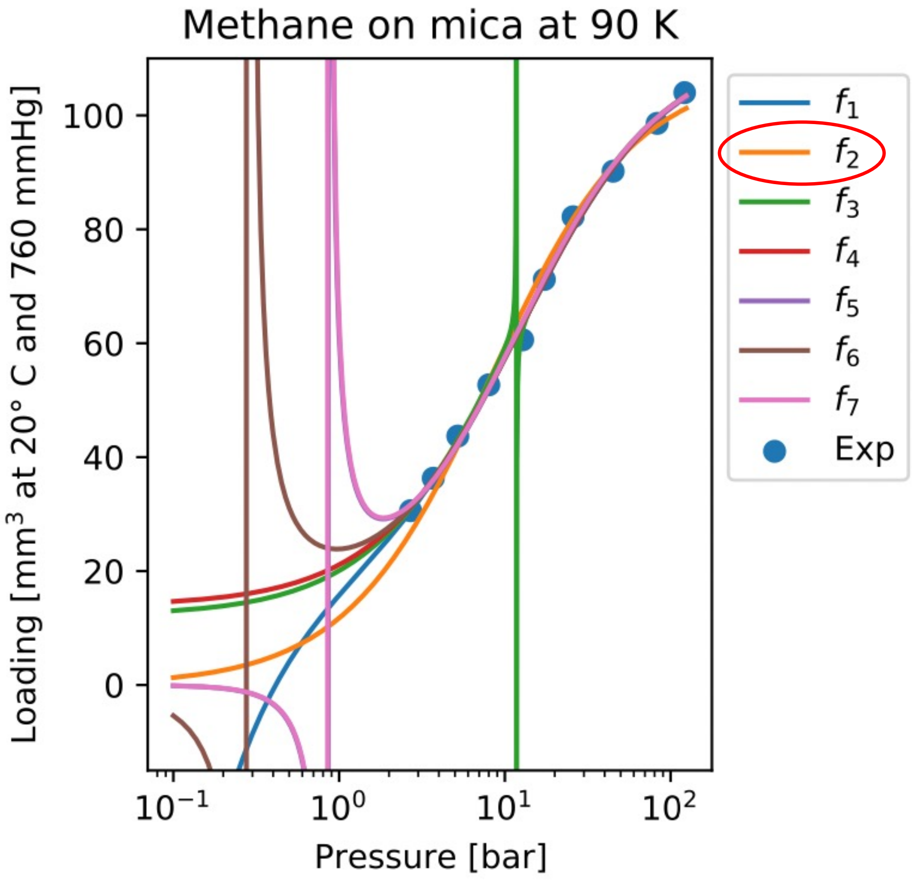
C3. $(\forall p > 0)\ (f'(p) \geq 0)$

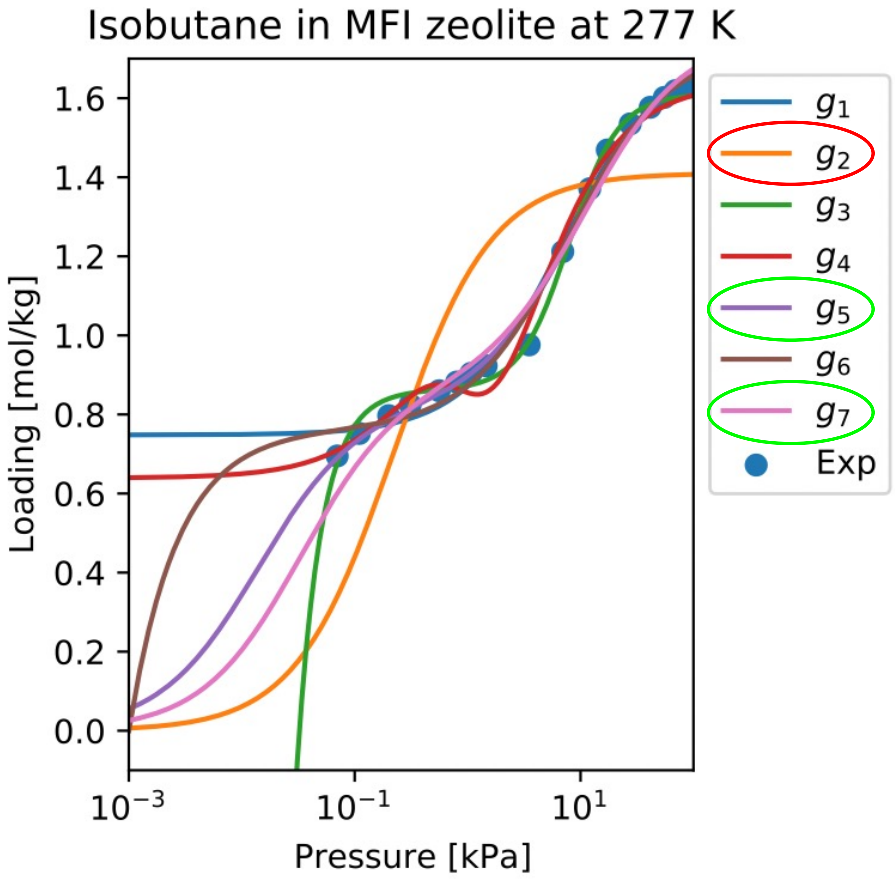
C4. $0 < \lim_{p \to 0} f'(p) < \infty$

C5. $0 < \lim_{p \to \infty} f(p) < \infty$

LANGMUIR'S ADSORPTION EQUATION



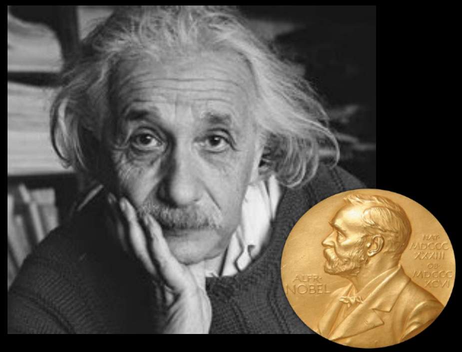


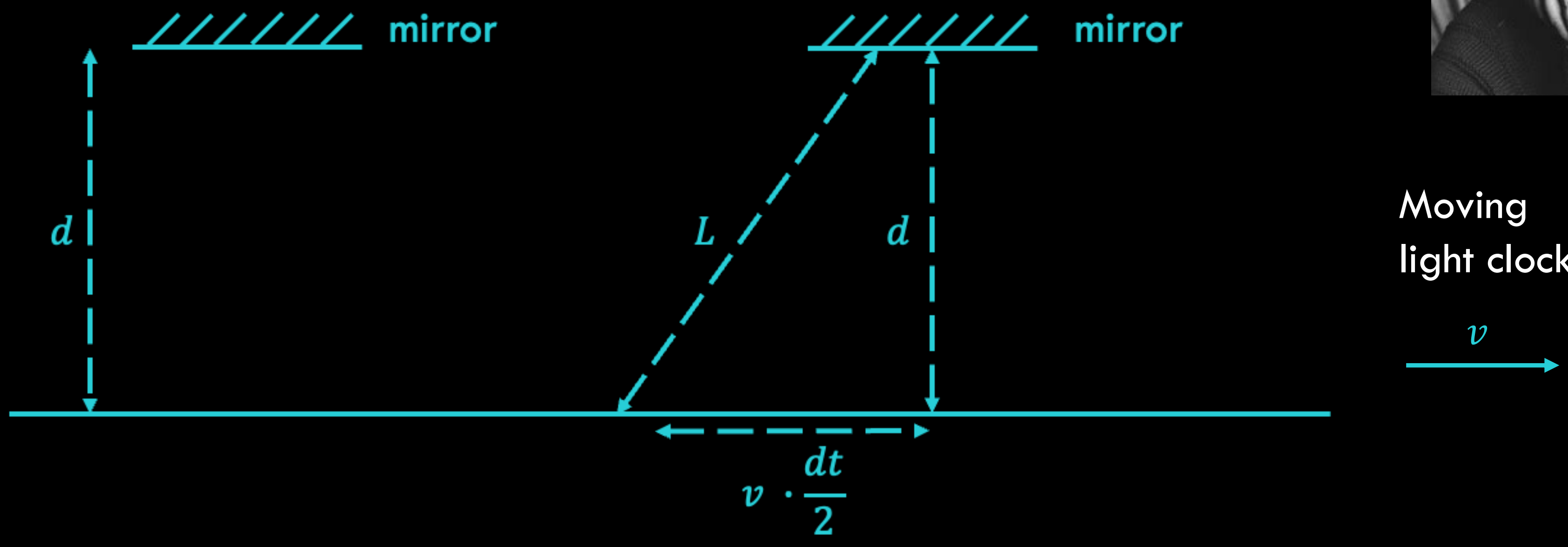


Candidate formula $q =$	
f_1 :	$(p^2+2p-1)/(.00888p^2+.118p)$
f_2 :	$p/(.00927p+.0759)$ *
f_3 :	$(p^2-10.5p-15.)/(.00892p^2-1.23)$
f_4 :	$(8.86p+13.9)/(.0787p+1)$
f_5 :	$p^2/(.00895p^2+.0934p-.0860)$
f_6 :	$(p^2+p)/(.00890p^2+.106p-.0311)$
f_7 :	$(112p^2-p)/(p^2+10.4p-9.66)$
g_1 :	$(p+3)/(.584p+4.01)$
g_2 :	$p/(.709p+.157)$
g_3 :	$(.0298p^2+1)/(.0185p^2+1.16)-.000905/p^2$
g_4 :	$1/(p^2+1)+(2.53p-1)/(1.54p+2.77)$
g_5 :	$(1.74p^2+7.61p)/(p^2+9.29p+0.129)$
g_6 :	$(.226p^2+.762p-7.62*10^{-4})/(.131p^2+p)$
g_7 :	$(4.78p^2+26.6p)/(2.71p^2+30.4p+1.)$

- f_2 and g_2 $\longrightarrow$ **provable**
- g_5 g_7 $\longrightarrow$ satisfy the **constraints**

RELATIVISTIC TIME DILATION



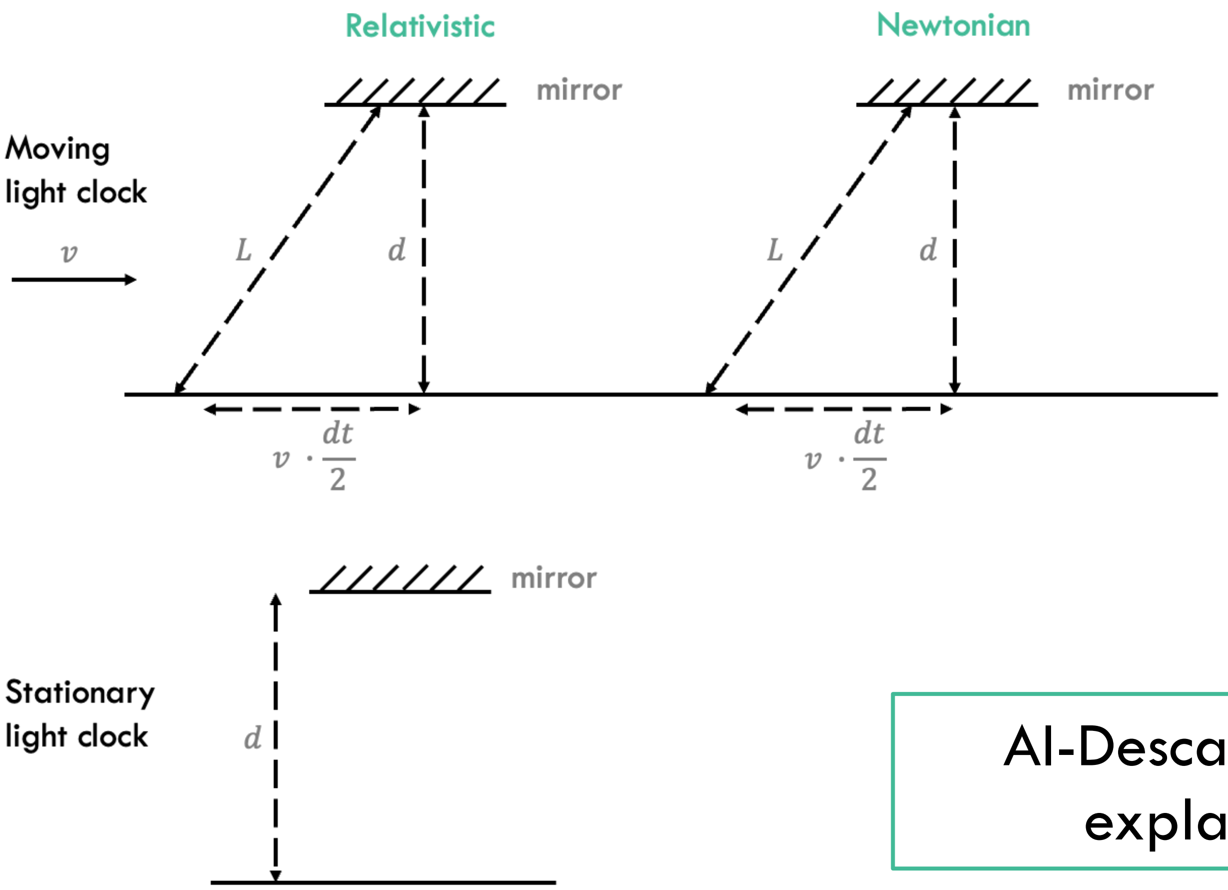


- Einstein's theory of relativity: the speed of light is constant
- Two observers in relative motion to each other will experience time differently and observe different clock frequencies

$$\frac{f - f_0}{f_0} = \sqrt{1 - \frac{v^2}{c^2}} - 1$$

Relativistic time dilation formula computes:
The frequency f for a clock moving at speed v is related to the frequency f_0 of a stationary clock by the formula

RELATIVISTIC TIME DILATION



2 Background theories

R1.	$dt_0 = 2 \cdot d/c$	R1'.	$dt_0 = 2 \cdot d/c$
R2.	$dt = 2 \cdot L/c$	R2'.	$dt = 2 \cdot L/\sqrt{v^2 + c^2}$
R3.	$L^2 = d^2 + (v \cdot dt/2)^2$	R3.	$L^2 = d^2 + (v \cdot dt/2)^2$
R4.	$f_0 = 1/dt_0$	R4'.	$f_0 = 1/dt_0$
R5.	$f = 1/dt$	R5'.	$f = 1/dt$
R6.	$df = f - f_0$	R6'.	$df = f - f_0$
R7.	$d > 0, v > 0$	R7'.	$d > 0, v > 0$
R8.	$c = 3 \times 10^8$	R8'.	$c = 3 \times 10^8$

Relativistic axioms **Newtonian axioms**

AI-Descartes can identify which theory explains the phenomenon better

RELATIVISTIC TIME DILATION

Candidate formula $y =$	Numerical Error Absolute ε_2^a	ε_∞^a	Numerical Error Relative ε_2^r	ε_∞^r	S s.t. Absolute Gen. Reas. Error $\beta_{\infty,S}^a \leq 1$	S s.t. Relative Gen. Reas. Error $\beta_{\infty,S}^r \leq 0.02$
$-0.00563v^2$	.3822	.3067	1.081	.001824	$37 \leq v \leq 115$	$37 \leq v \leq 10^8$
$\frac{v}{1+0.00689v} - v$	.3152	.2097	1.012	.006927	$37 \leq v \leq 49$	$37 \leq v \leq 38$
$-0.00537\frac{v^2\sqrt{v+v^2}}{(v-1)}$	.3027	.2299	1.254	.002147	$37 \leq v \leq 98$	$37 \leq v \leq 109$
$-0.00545\frac{v^4}{\sqrt{v^2+v^{-2}}(v-1)}$	.3238	.2531	1.131	.0009792	$37 \leq v \leq 126$	$37 \leq v \leq 10^7$

COMPARISON WITH SOTA SYSTEMS

- **AI-Feynman:** deep learning based symbolic regression algorithm.
- **TuringBot:** simulated annealing method to find expressions that fit the input data.
- **PySR:** based on regularized evolution, simulated annealing, and gradient-free optimization.
- **Bayesian Machine Scientist (BMS):** Markov chain Monte Carlo based method exploiting a prior learned from a large empirical corpus of mathematical expressions.

	AI-Descartes	AI-Feynman	PySR	BMS	TuringBot
Accuracy	**0.60**	0.41	0.49	0.48	-
Accuracy (max 2 var)	**0.87**	0.80	0.73	0.80	0.80

Accuracy on the Feynman synthetic dataset

CONCLUSION & FUTURE/ONGOING WORK

Strengths:

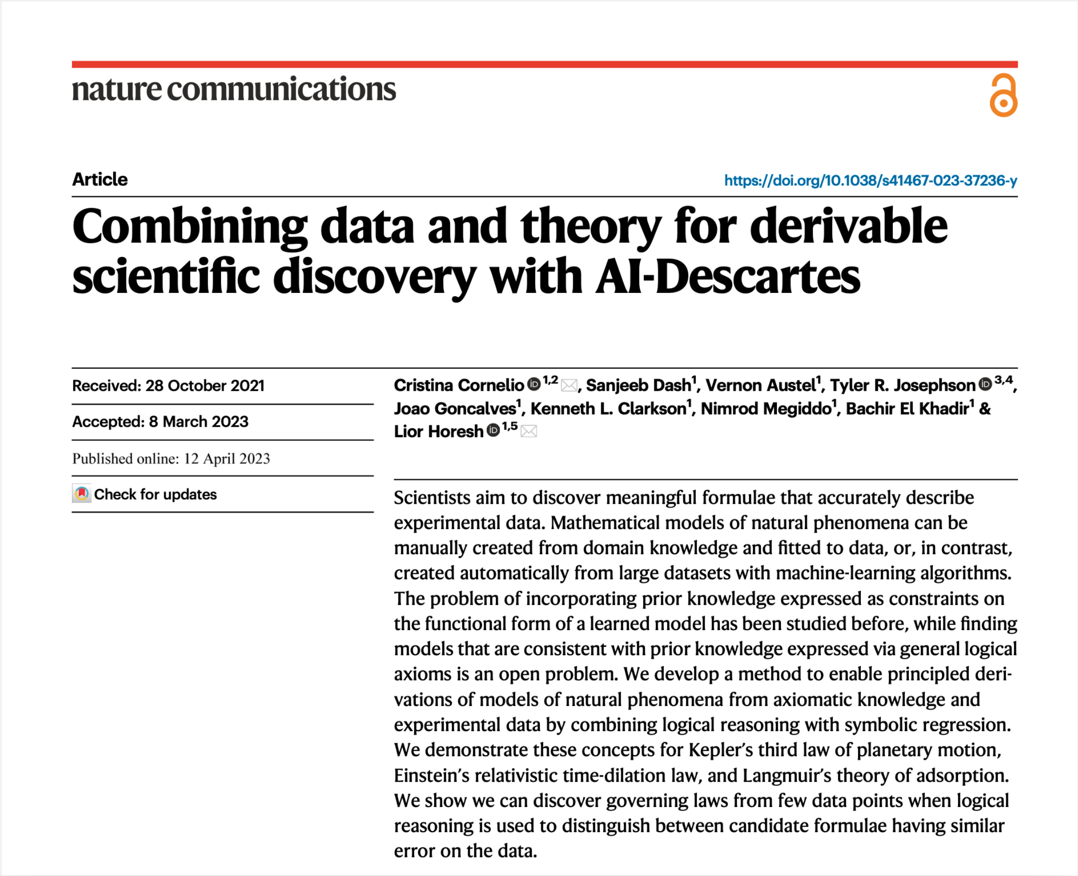
- Few data points / Real data
- Logical reasoning to distinguish the correct formula from a set of plausible formulas with similar error on the data

Limitations:

- Limitation of the tools
- Scalability to bigger formulas
- Rely on correctness & completeness of background theory

Main challenges:

- We need more real-data datasets (only synthetic datasets with non-realistic amount/type of noise)
- We need more numerical datasets with associated background theory


nature communications

Article https://doi.org/10.1038/s41467-023-37236-y

Combining data and theory for derivable scientific discovery with AI-Descartes

Received: 28 October 2021

Accepted: 8 March 2023

Published online: 12 April 2023

Check for updates

Cristina Cornelio[1,2], Sanjeeb Dash[1], Vernon Austel[1], Tyler R. Josephson[3,4], Joao Goncalves[1], Kenneth L. Clarkson[1], Nimrod Megiddo[1], Bachir El Khadir[1] & Lior Horesh[1,5]

Scientists aim to discover meaningful formulae that accurately describe experimental data. Mathematical models of natural phenomena can be manually created from domain knowledge and fitted to data, or, in contrast, created automatically from large datasets with machine-learning algorithms. The problem of incorporating prior knowledge expressed as constraints on the functional form of a learned model has been studied before, while finding models that are consistent with prior knowledge expressed via general logical axioms is an open problem. We develop a method to enable principled derivations of models of natural phenomena from axiomatic knowledge and experimental data by combining logical reasoning with symbolic regression. We demonstrate these concepts for Kepler's third law of planetary motion, Einstein's relativistic time-dilation law, and Langmuir's theory of adsorption. We show we can discover governing laws from few data points when logical reasoning is used to distinguish between candidate formulae having similar error on the data.

https://github.com/IBM/AI-Descartes https://ai-descartes.github.io

REFERENCES

Papers:

- AI Descartes: Combining Data and Theory for Derivable Scientific Discovery, Nature Communications, C. Cornelio, S. Dash, V. Austel, T. R. Josephson, J. Goncalves, K. Clarkson, N. Megiddo, B. El Khadir, and L. Horesh.
- Symbolic Regression using Mixed-Integer Nonlinear Optimization, V. Austel, C. Cornelio, S. Dash, J. Gonçalves, L. Horesh, T. Josephson, N. Megiddo.
- Bayesian Experimental Design for Symbolic Discovery, K. Clarkson, C. Cornelio, S. Dash, J. Gonçalves, L. Horesh, N. Megiddo.
- [Under submission] AI-Hilbert

Patents:

- Generative Reasoning for Symbolic Discovery, C. Cornelio, L. Horesh, V. Pestun, R. Yan.
- Symbolic Model Discovery based on a combination of Numerical Learning Methods and Reasoning, C. Cornelio, L. Horesh, A. Fokoue-Nkoutche, S. Dash.
- Experimental Design for Symbolic Model Discovery, L. Horesh, K. Clarkson, C. Cornelio, S. Magliacane.
- Background Theory-Based Method for Refinement and Evaluation of Functional Models Extracted from Numerical Data, Lior Horesh, C. Cornelio, Bachir El Khadir, Sanjeeb Dash, Joao P. Goncalves, Kenneth Lee Clarkson
- Logical and Statistical Composite Models, L. Horesh, B. El Kadir, S. Dash, K. Clarkson, C. Cornelio
- Symbolic Model Discovery Rectification, L. Horesh, C. Cornelio, S. Dash, J.P. Goncalves, K. L. Clarkson, N. Megiddo, V. Austel, B. El Khadir