Evolving Scientific Discovery by Unifying Data And Background Knowledge With AI-Hilbert

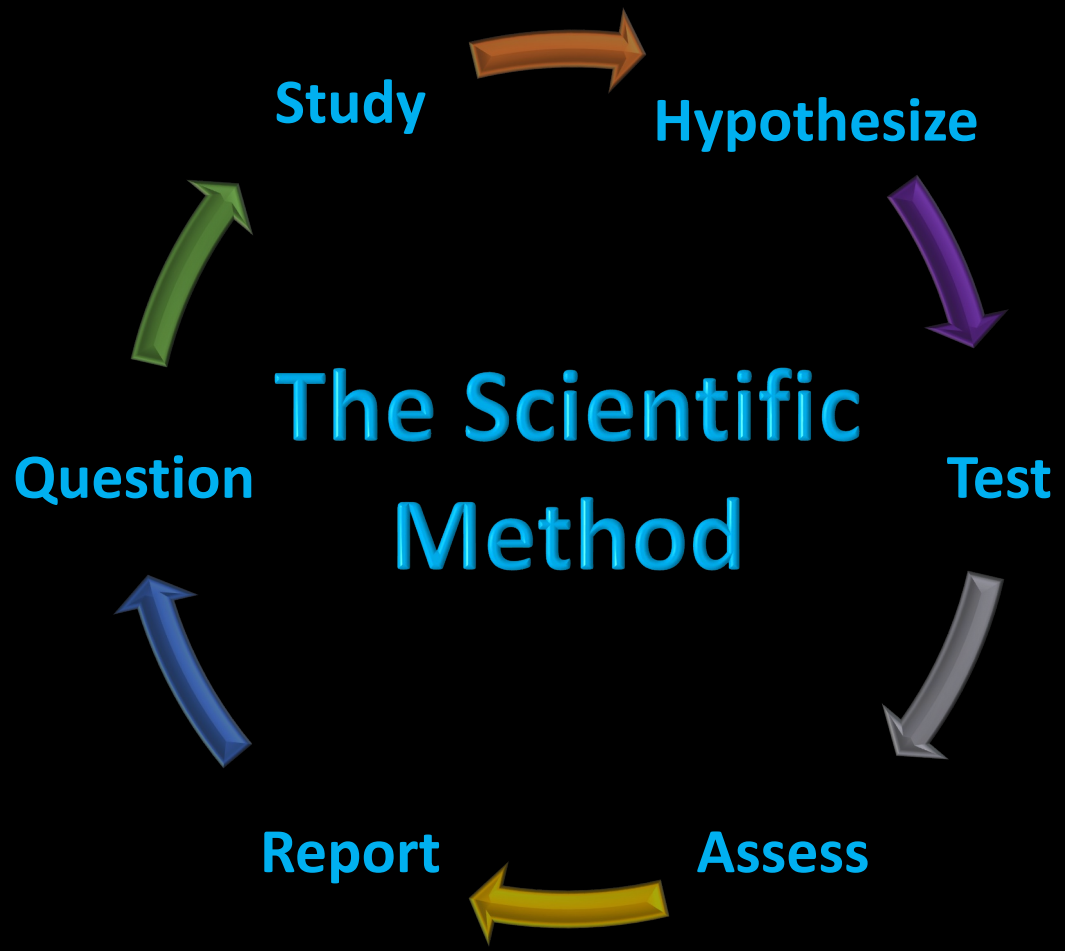
Ryan Cory-Wright (ICBS, I-X)*, Bachir El Khadir (Two Sigma), Cristina Cornelio (Samsung AI), Sanjeeb Dash (IBM), Lior Horesh (IBM)

Preprint: arXiv 2308.09474, in revision at Nature Comms

* Imperial College Business School and Imperial-X

Website: ryancorywright.github.io

Motivation: The Scientific Method

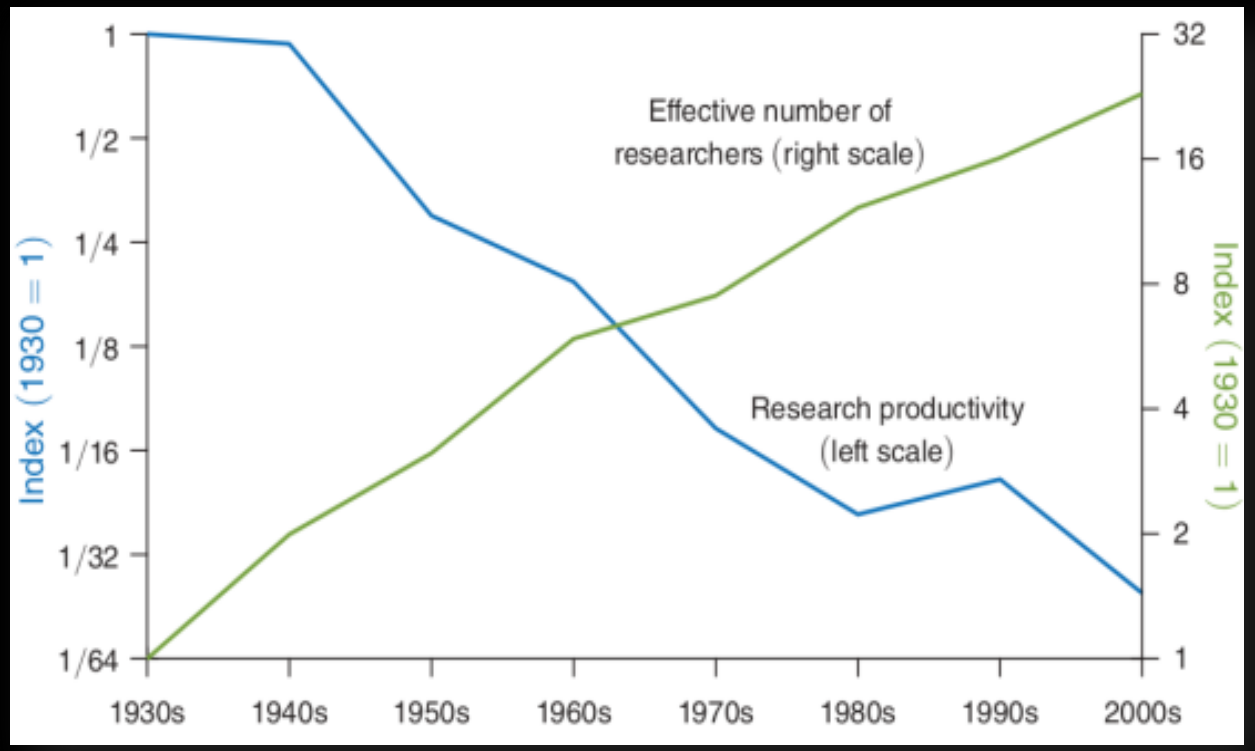


- The scientific method has been transformative for mankind's economic growth
- A principled practice for scientists to make substantiated discoveries

Is The Scientific Method Still Fit For Purpose?

Many authors: No!

Emergence of discoveries and contribution to economic growth stagnating relative to capital invested

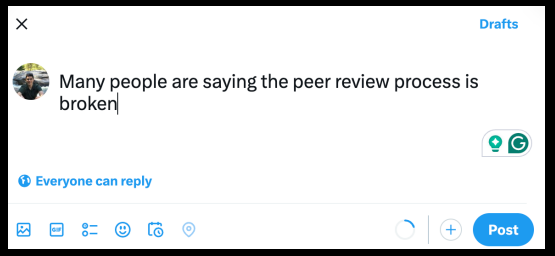


Economic Growth = Research Productivity × Numbers of Researchers

"It is now more challenging for first-rate physicists to make second-rate discoveries than it was previously for second-rate physicists to make first-rate discoveries"

Paul Dirac
1902 - 1984

"The number of researchers required today to achieve the famous doubling of computer chip density is more than 18 times larger than the number required in the early 1970s."-Bloom et al



- me, each time I read a report from Reviewer Two

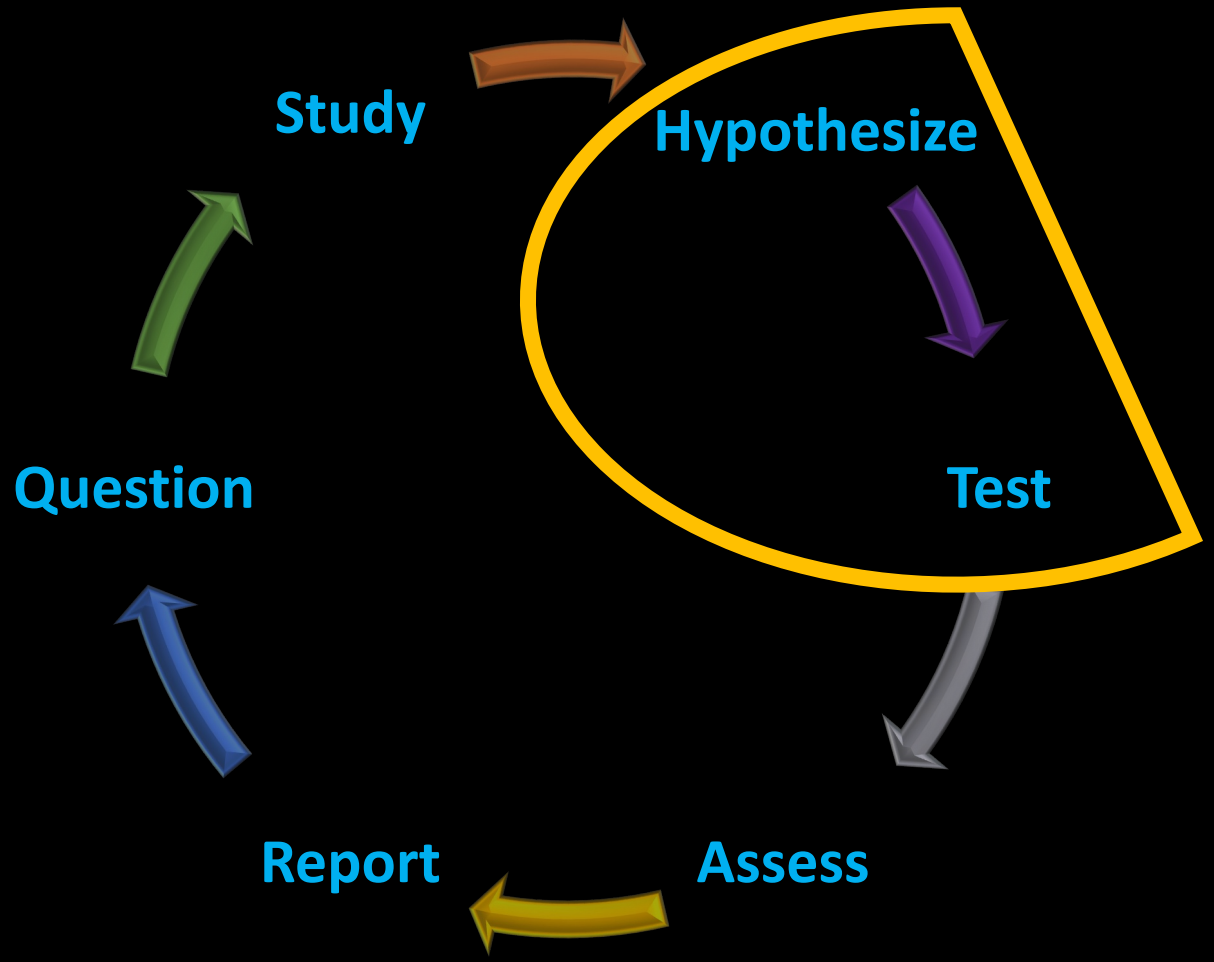
Bloom, Jones, Van Reenan & Webb. Are Good Ideas Getting Harder to Find?, American Economic Review, 2020
Dirac. Directions in physics. lectures delivered during a visit to Australia and New Zealand, August/September 1975

Overarching Research Question:
Can We Evolve Beyond The Scientific Method?

Too Hard!

…. For Now (Stay Tuned)

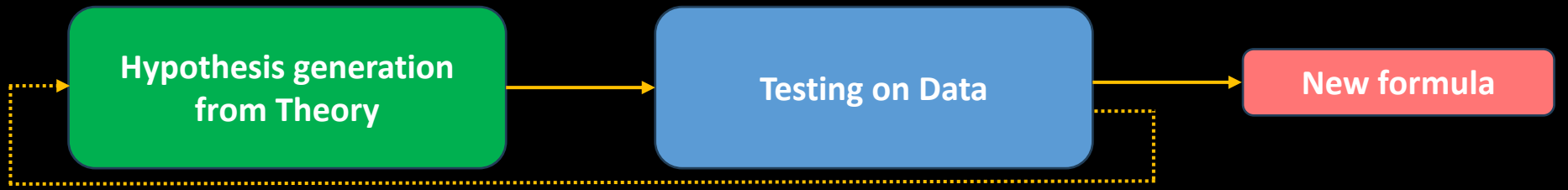
Overall Problem Too Hard: Focus on Evolving Discovery Phase



Existing Discovery Paradigms

Hypothesis Generation And Testing Paradigms

1. Classical Scientific Discovery



2. Data Driven Modeling (e.g., AI-Feynman)

- **Link input** and **output** data via a **generic functional form**
- **Functional form** is chosen to be **computationally exploitable**
- Can be effective when:

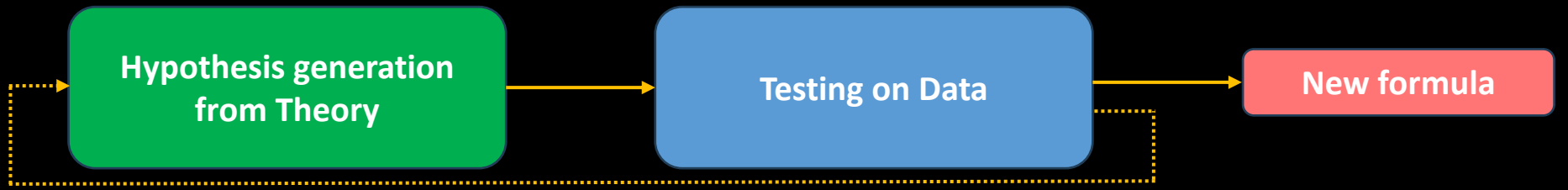
  Large volume of input-output **data** pairs is accessible

  There is **no / little domain knowledge** regarding the process

- May **fail** to **fully generalize** the **underlying behavior**
- **Low interpretability**

Hypothesis Generation And Testing Paradigms

1. Classical Scientific Discovery



2. ML / Regression

Can We Combine Data And Theory?

3. Combining Data and Theory via AI Descartes

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.

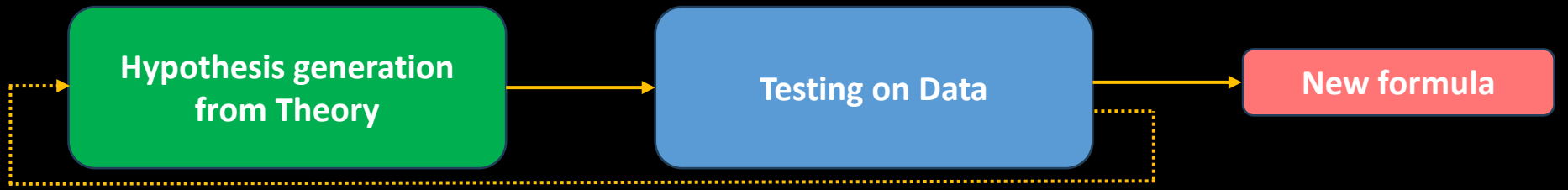
Main criticism of AI-Descartes:

Background theory not fully integrated with data in symbolic regression solver

Cornelio, Dash, Austel, Josephson, Goncalves, Clarkson, Megiddo, El Khadir, Horesh. Combining data and theory for derivable scientific discovery with AI-Descartes. Nature Comm, 2023

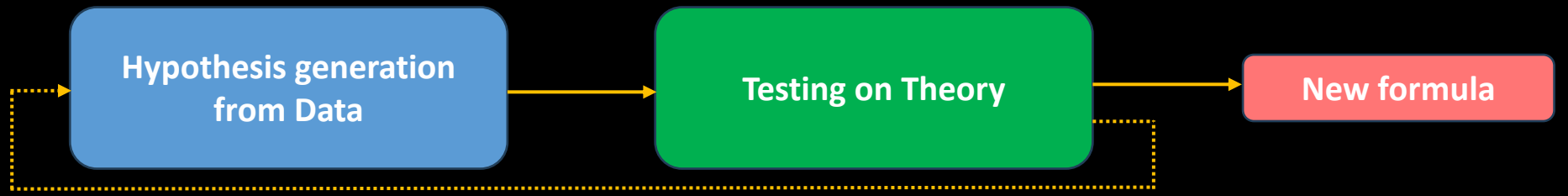
Hypothesis Generation And Testing Paradigms

1. Classical Scientific Discovery



2. ML / Regression

3. AI - Descartes

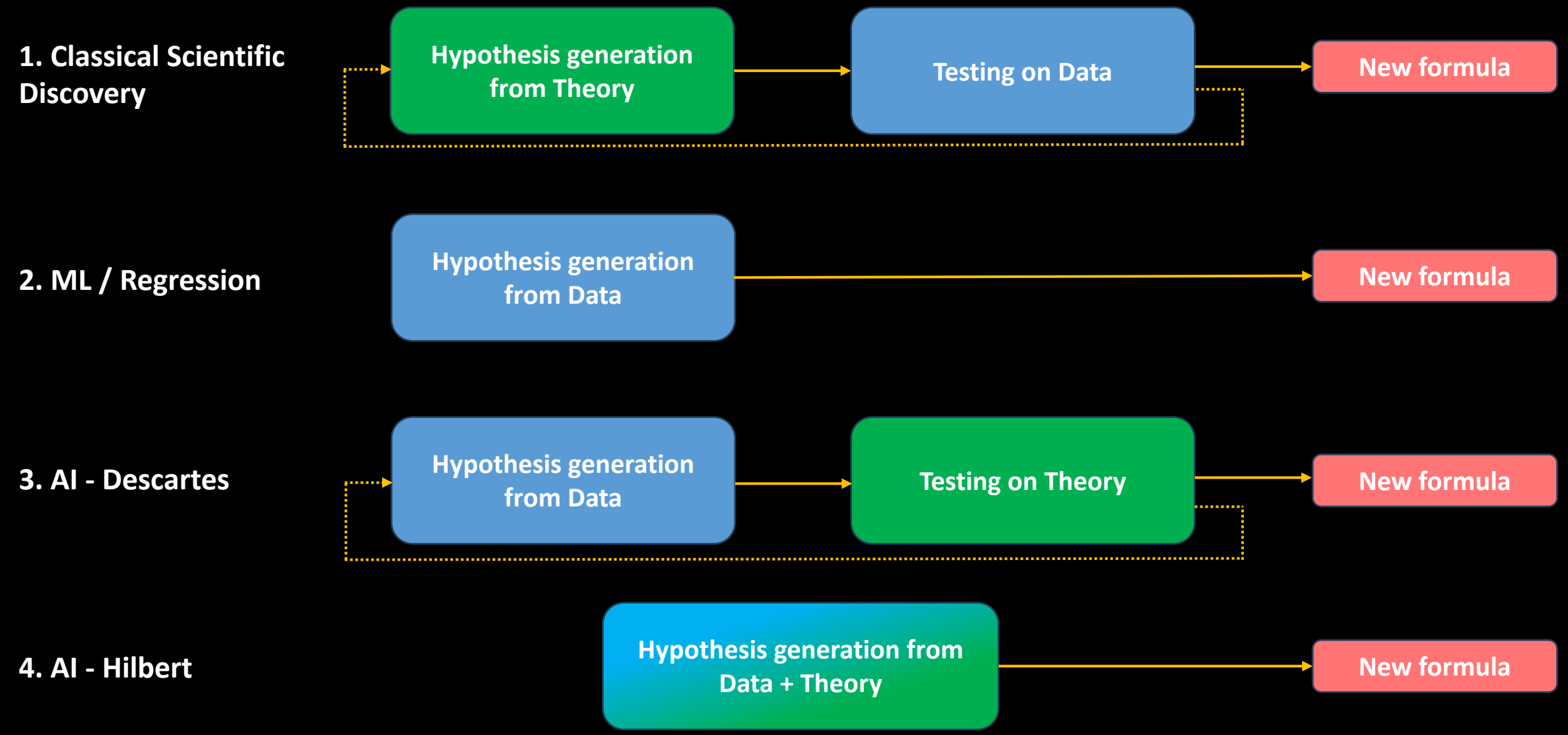


Can we do Even Better?

"Why don't we just use sum-of-squares optimization to do everything at once?"
- Bachir, a few months before I joined

— Me, around when I joined this project

Using Sum-of-Squares Optimization, we Will Show..



A Unified Framework for Scientific Discovery

Computational and mathematical framework to discover a symbolic mathematical model $f(x_1, x_2, \ldots, x_n) = 0$, where $\deg(f(x_1)) \geq 1$
that **simultaneously** fits numerical data and is compatible with background knowledge
where x_1 dependent variable, $x_2, \ldots, x_n$ independent variables, we may exclude some $x_i's$ from f.
Note: we seek implicit functions of x_1 so we can model ratios of polynomials, roots etc.

Concretely, a multi-objective optimization problem that aims to minimize the following criteria:

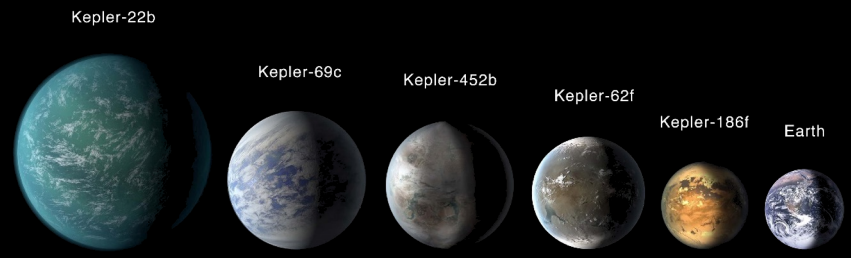
- Compatibility with background knowledge: *background knowledge compatibility* $L_T(f)$
- Fidelity to numerical data: *data fidelity* $L_D(f)$
- Formula complexity: *complexity loss* $L_C(f)$

Data Fidelity

Data is given by a table of numerical values

$$\mathrm{X}_1 = (x_1^1, x_2^1, \ldots, x_n^1)$$
$$\vdots$$
$$\mathrm{X}_\mathrm{N} = (x_1^N, x_2^N, \ldots, x_n^N)$$

planet	p	m1	m2	d
Mercury	0.234481	1.0	0.055	0.3871
Venus	0.456604	1.0	0.815	0.7233
Earth	0.707107	1.0	1.000	1.0000
Mars	1.787621	1.0	0.107	1.5237
Jupiter	0.664962	1.0	317.800	5.2044
Saturn	3.025033	1.0	95.159	9.5826
Uranus	21.375006	1.0	14.536	19.2184
Neptune	38.707732	1.0	17.147	30.0700



The fit of a mathematical model f to the data is measured with a "data fidelity" function $L_D(f)$

- ℓ_2 loss function

$$L_D(f) = \sum_{i=1,\ldots,N} f(X_i)^2$$

- ℓ_∞ loss function

$$L_D(f) = \max_{i=1,\ldots,N} |f(X_i)|$$

Model Complexity

- Given a polynomial model

$$f(x_1, \dots, x_n) = f_1 x_1^d + f_2 x_1^{d-1} x_2 + \cdots + f_N x_n^d$$

- Potential ***complexity loss*** functions:
  - $L_C(f) = \deg(f)$
  - $L_C(f) = \sum_i |f_i|^p$ with $p \geq 0$, i.e., ℓ_p norm of the coefficients of f
  - $L_C(f) = \#\{i \,|f_i \neq 0\}$, i.e., number of nonzero coefficients of f

Distance to Background Knowledge Via Putinar's Positivestellensatz

- Consider background knowledge expressed via basic (semi)algebraic sets

$$\mathcal{G} := \{\boldsymbol{x} \in \mathbb{R}^n : \ g_1(\boldsymbol{x}) \geq 0, \ldots, g_m(\boldsymbol{x}) \geq 0\},$$
$$\mathcal{H} := \{\boldsymbol{x} \in \mathbb{R}^n : \ h_1(\boldsymbol{x}) = 0, \ldots, h_n(\boldsymbol{x}) = 0\},$$

  where $g_i, h_j \in \mathbb{R}[x]_n$, and suppose that $\mathcal{G}$ satisfies the Archimedean property, i.e., there exists an $R > 0$ and $\alpha_0, \ldots \alpha_n \in \Sigma[\boldsymbol{x}]_n$ such that

$$R - \sum_{i=1}^{n} x_i^2 = \alpha_0(\boldsymbol{x}) + \sum_{i=1}^{m} \alpha_i(\boldsymbol{x}) g_i(\boldsymbol{x})$$

- Then, for any $f \in \mathbb{R}[x]_{\boldsymbol{n},2d}$ the implication

$$\boldsymbol{x} \in \mathcal{G} \cap \mathcal{H} \implies f(\boldsymbol{x}) \geq 0$$

  holds if and only if there exist Sum Of Squares (SOS) polynomials $\alpha_0, \ldots, \alpha_m \in \Sigma[\boldsymbol{x}]_{n,2d}$, and real polynomials $\beta_1, \ldots, \beta_n \in \mathbb{R}[\boldsymbol{x}]_{n,2d}$ such that

$$f(x) = \alpha_0(x) + \sum_{i=1}^{m} \alpha_i(\boldsymbol{x}) g_i(\boldsymbol{x}) + \sum_{j=1}^{n} \beta_j(\boldsymbol{x}) h_j(\boldsymbol{x})$$

Putinar, Positive Polynomials on Compact Semi-Algebraic Sets. Indiana University Mathematics Journal, 1993

DISTANCE TO (INCOMPLETE) BACKGROUND KNOWLEDGE

- Consider the basic (semi)algebraic sets which encode background knowledge

$$\mathcal{G} := \{\boldsymbol{x} \in \mathbb{R}^n : \ g_1(\boldsymbol{x}) \geq 0, \ldots, g_m(\boldsymbol{x}) \geq 0\},$$
$$\mathcal{H} := \{\boldsymbol{x} \in \mathbb{R}^n : \ h_1(\boldsymbol{x}) = 0, \ldots, h_n(\boldsymbol{x}) = 0\},$$

we minimize distance from derived polynomial $q(x)$ to background theory. Two cases w.l.o.g.:

- 1. Minimize distance to closest Psatz certificate:

$$d^c(f, \mathcal{G} \cap \mathcal{H}) := \min_{\substack{\alpha_0,\ldots,\alpha_m \in \Sigma_{n,2d}[\boldsymbol{x}], \\ \beta_1,\ldots,\beta_n \in \mathbb{R}_{n,2d}}} \left\| \text{Coefficients}\left(f - \alpha_0 - \sum_{i=1}^{m} \alpha_i g_i - \sum_{j=1}^{n} \beta_j h_j \right) \right\|_2$$

DISTANCE TO (INCONSISTENT) BACKGROUND KNOWLEDGE

- Consider the basic (semi)algebraic sets which encode background knowledge

$$\mathcal{G} := \{\boldsymbol{x} \in \mathbb{R}^n : \ g_1(\boldsymbol{x}) \geq 0, \ldots, g_m(\boldsymbol{x}) \geq 0\},$$
$$\mathcal{H} := \{\boldsymbol{x} \in \mathbb{R}^n : \ h_1(\boldsymbol{x}) = 0, \ldots, h_n(\boldsymbol{x}) = 0\},$$

we minimize distance from derived polynomial $q(x)$ to background theory. Two cases w.l.o.g.:

2. Minimize distance to Psatz certificate with $\leq k$ polynomials from background theory

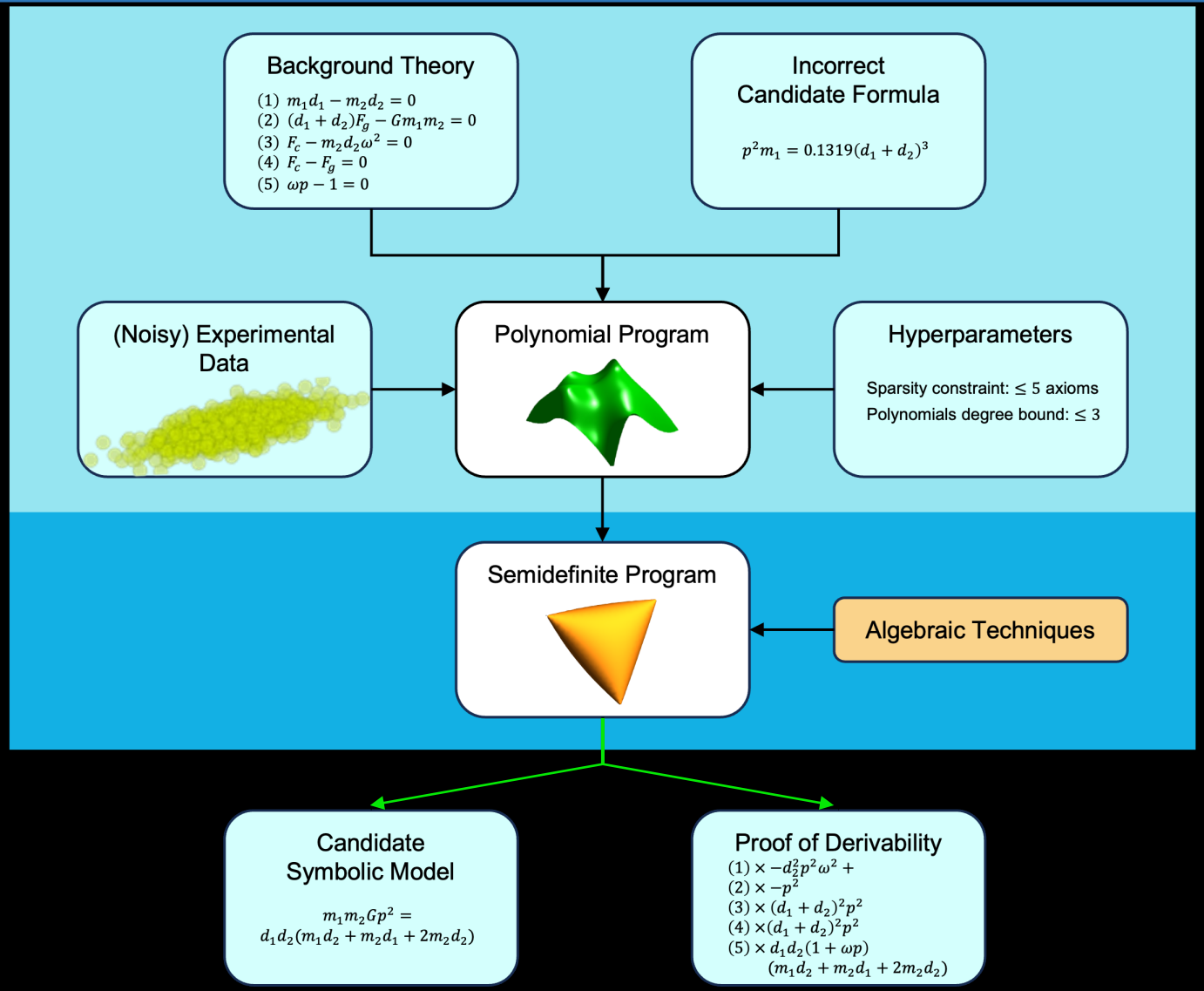
Implicit hypothesis: at least *k* polynomials in background theory are correct

$$d^c(f, G \cap \mathcal{H}) := \min \quad \left\| \text{Coefficients}\left(f - \alpha_0 - \sum_{i=1}^{m} \alpha_i g_i - \sum_{j=1}^{n} \beta_j h_j \right) \right\|_2$$

$$\text{s.t.} \quad \alpha_i = 0 \text{ if } z_i = 0, \forall i \in \{0, \ldots, m\}$$
$$\beta_j = 0 \text{ if } y_j = 0, \ \forall j \in \{1, \ldots n\}$$
$$\sum_{i=0}^{m} z_i + \sum_{j=1}^{n} y_j \leq k, \ \alpha_0, \ldots, \alpha_m \in \Sigma_{n,2d}[\boldsymbol{x}]$$
$$z_0 \ldots z_m \in \{0,1\}, \ \beta_1, \ldots, \beta_n \in \mathbb{R}_{n,2d}, y_1, \ldots y_n \in \{0,1\}$$

Let’s Look at Some Examples

1. Kepler's Law of Planetary Motion



Cory-Wright, El-Khadir, Cornelio, Dash, Horesh. AI Hilbert: From Data and Background Knowledge to Automated Scientific Discovery, https://arxiv.org/abs/2308.09474, 2023

2. Relativity: Radiation Gravitational Wave Power

- Assume that two objects orbit a constant distance of *r* away from each other

Kepler's previously derived third law of planetary motion

$$\omega^2 r^3 - G(m_1 + m_2) = 0$$

Gravitational power of a wave (wavelength $\gg$ source dimensions)

$$5(m_1 + m_2)^2 c^5 P + G \operatorname{tr}\left(\frac{d^3}{dt^3}\left(m_1 m_2 r^2 \begin{pmatrix} x^2 - \frac{1}{3} & xy & 0 \\ xy & y^2 - \frac{1}{3} & 0 \\ 0 & 0 & -\frac{1}{3} \end{pmatrix}\right)^2\right) = 0$$

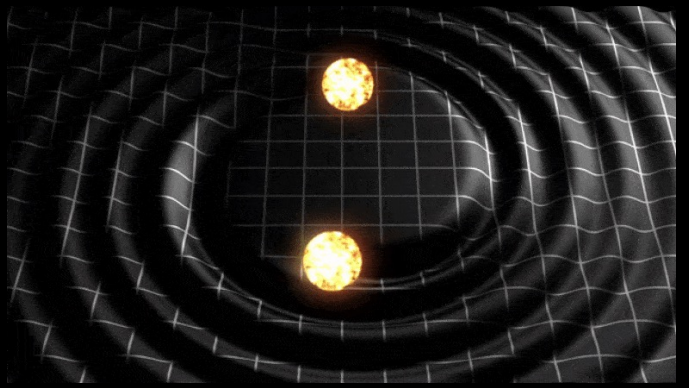
Quadrupole moment tensor

$$x^2 + y^2 = 1$$

where *P* is the (average) power of the wave, $G = 6.6743 \times 10^{-11} m^3 kg^{-1} s^{-2}$ is the universal gravitational constant, *c* is the speed of light, m_1, m_2 are masses of the objects

- AI - Hilbert correctly derives:

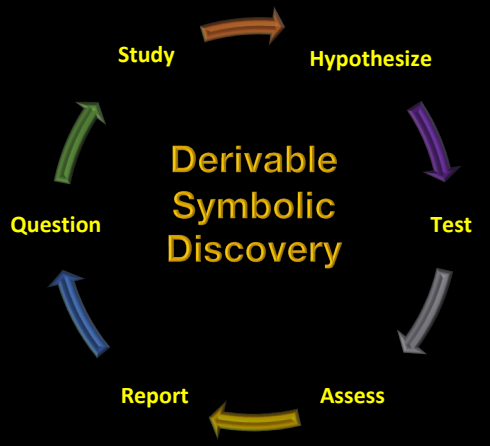
$$P = -\frac{32G^4}{5c^5 r^5}(m_1 m_2)^2(m_1 + m_2)$$



Wrapping Up

Can we Evolve the Scientific Method?

- Accelerated scientific discovery while fostering models that are:
  - Universal
  - Derivable (and interpretable)
  - Small data demand
  - Noisy data and imperfect knowledge
  - Simultaneously and consistently account for data and knowledge



- An adjunct to human scientists through the phases of the discovery cycle

- A step towards making discoveries outside the realms of either data-driven or classical approaches alone

- Goal for next paper: use SOS to make new discoveries
  - “Are these discoveries in the room with us now?”-Reviewer Two

THANK YOU