Toward an AI physicist for unsupervised learning

Tailin Wu, Max Tegmark 2018, arxiv:1810.10525

Motivation – warm up

0.0 s

Calculate the 13th root of following number:

7066437381674286102234008830240157375704233170702632731269721516000395709065419973141914549389684111

→Result: 47941071

→Human world record (2010): 11.8 seconds

→My laptop on average: 15 nanoseconds

Outline

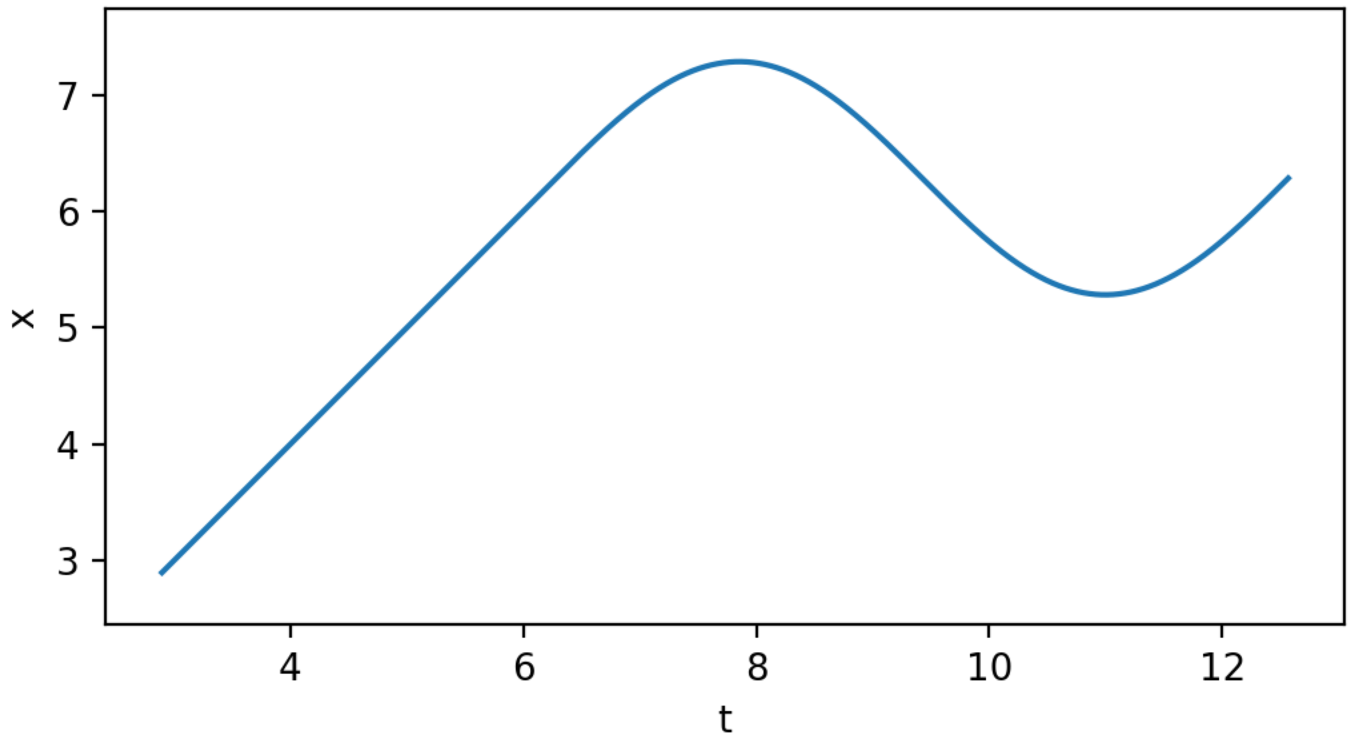
- What is a theory?
- AI physicist architecture
  - Divide and conquer
  - Occam’s razor
  - Unification
  - Life long learning
- Experiments
- Conclusion/Discussion

What is a theory?

- Definition:
  - Theory $\mathcal{T}$ is 2-tupel $(\mathbf{f}, c)$
  - Input $\mathbf{x}_t = (\mathbf{y}_{t-T}, \ldots, \mathbf{y}_{t-1}), \quad \mathbf{y}_i \in \mathbb{R}^d$
  - $\mathbf{f}$ is prediction function $\longrightarrow \mathbf{y}_t$ (3 layer NN with linear activation function)
  - c is domain classifier (3 layer NN with leakyReLU activation function)

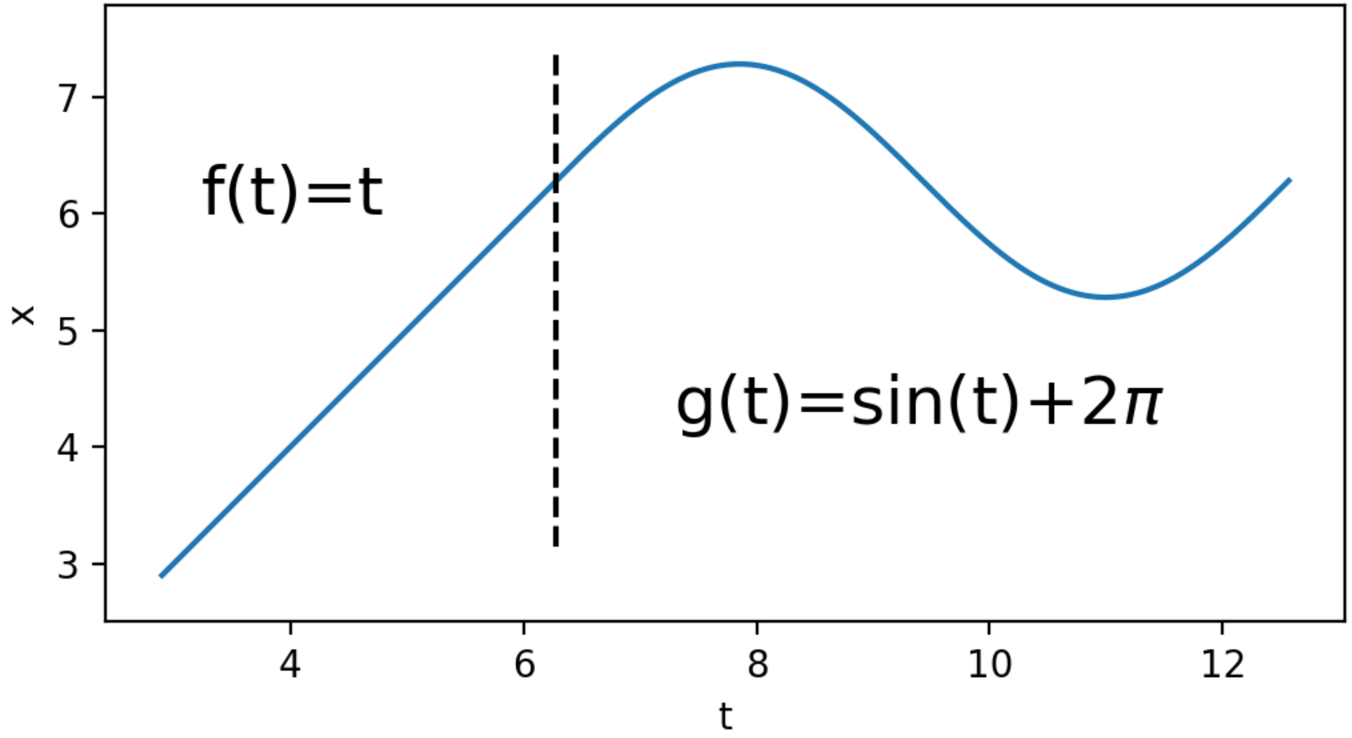
What is a theory?

- Example:



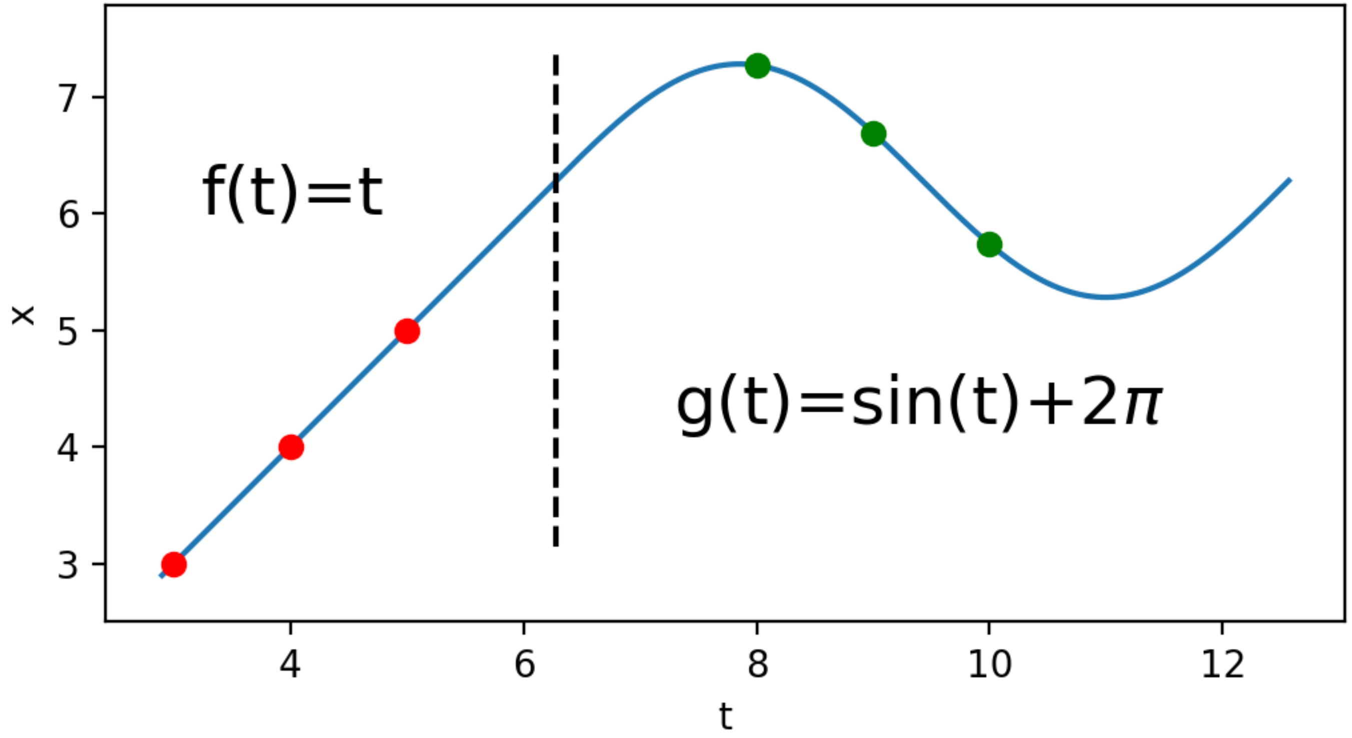
What is a theory?

- Example:



What is a theory?

- Example:



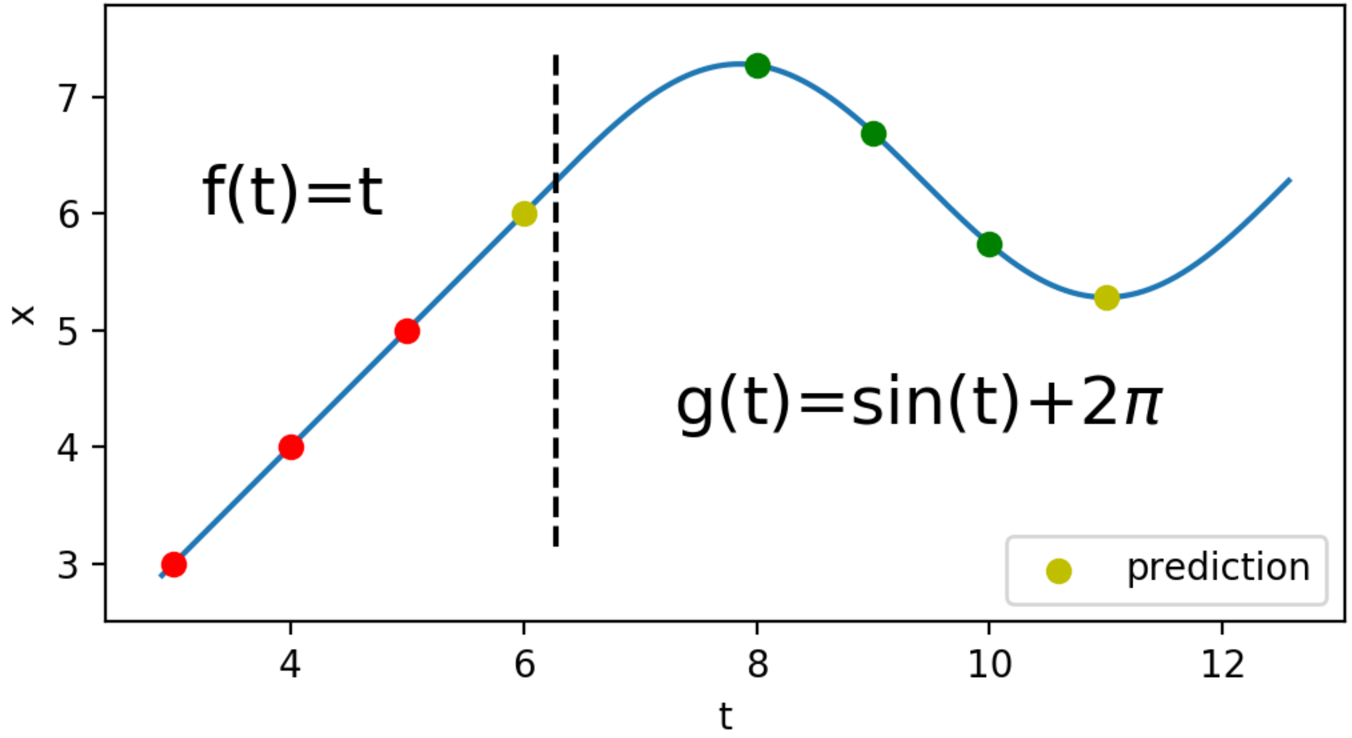
Data:

$$\mathbf{x}_t = (3.0, 4.0, 5.0) \xrightarrow{c} \mathbf{f}(\mathbf{x}_t)$$

$$\mathbf{x}_t = (7.3, 6.7, 5.7) \xrightarrow{c} \mathbf{g}(\mathbf{x}_t)$$

What is a theory?

- Example:



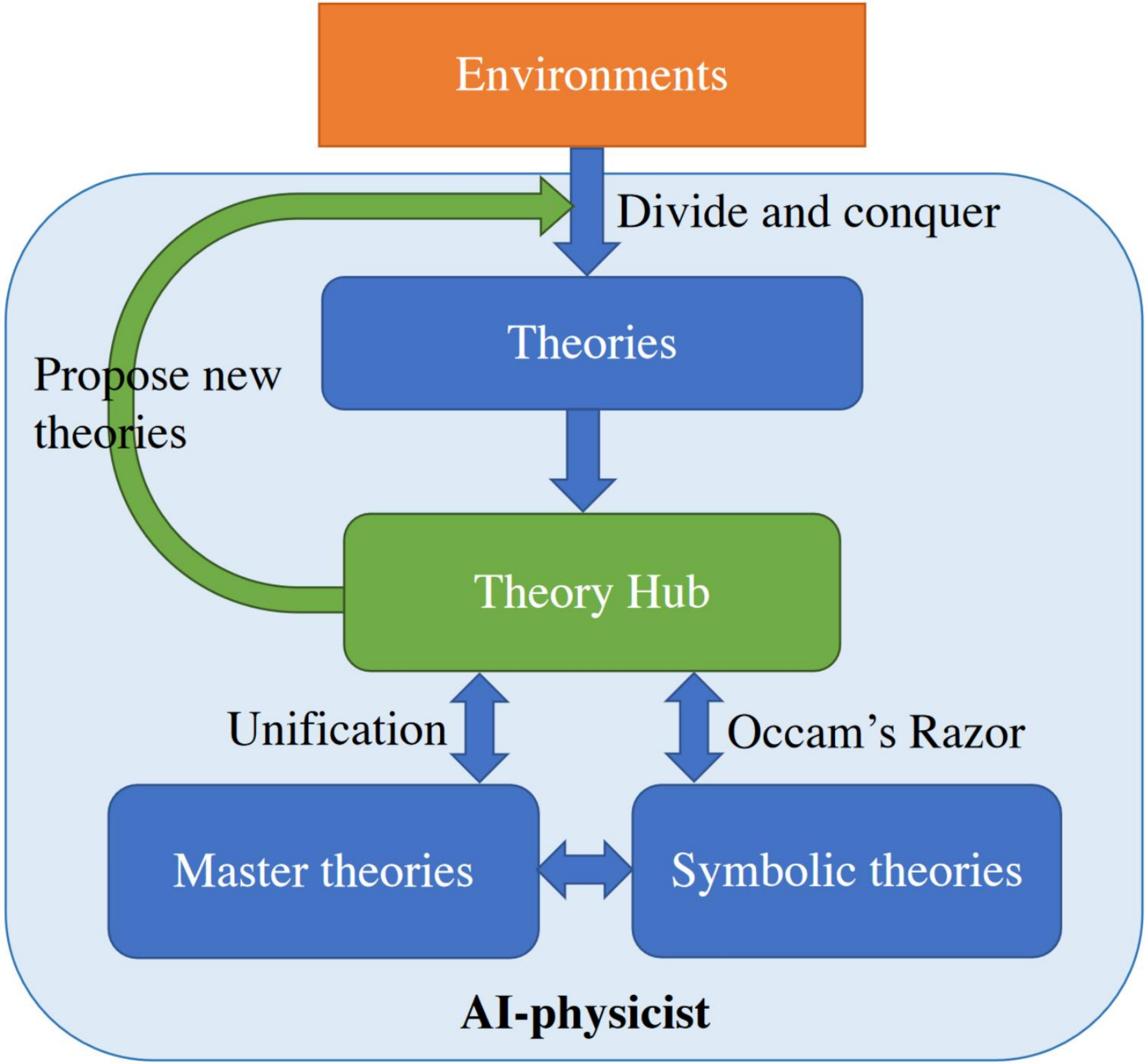
Data:

$$\mathbf{x}_t = (3.0, 4.0, 5.0) \xrightarrow{c} \mathbf{f}(\mathbf{x}_t) = 6.0$$

$$\mathbf{x}_t = (7.3, 6.7, 5.7) \xrightarrow{c} \mathbf{g}(\mathbf{x}_t) = 5.3$$

Architecture

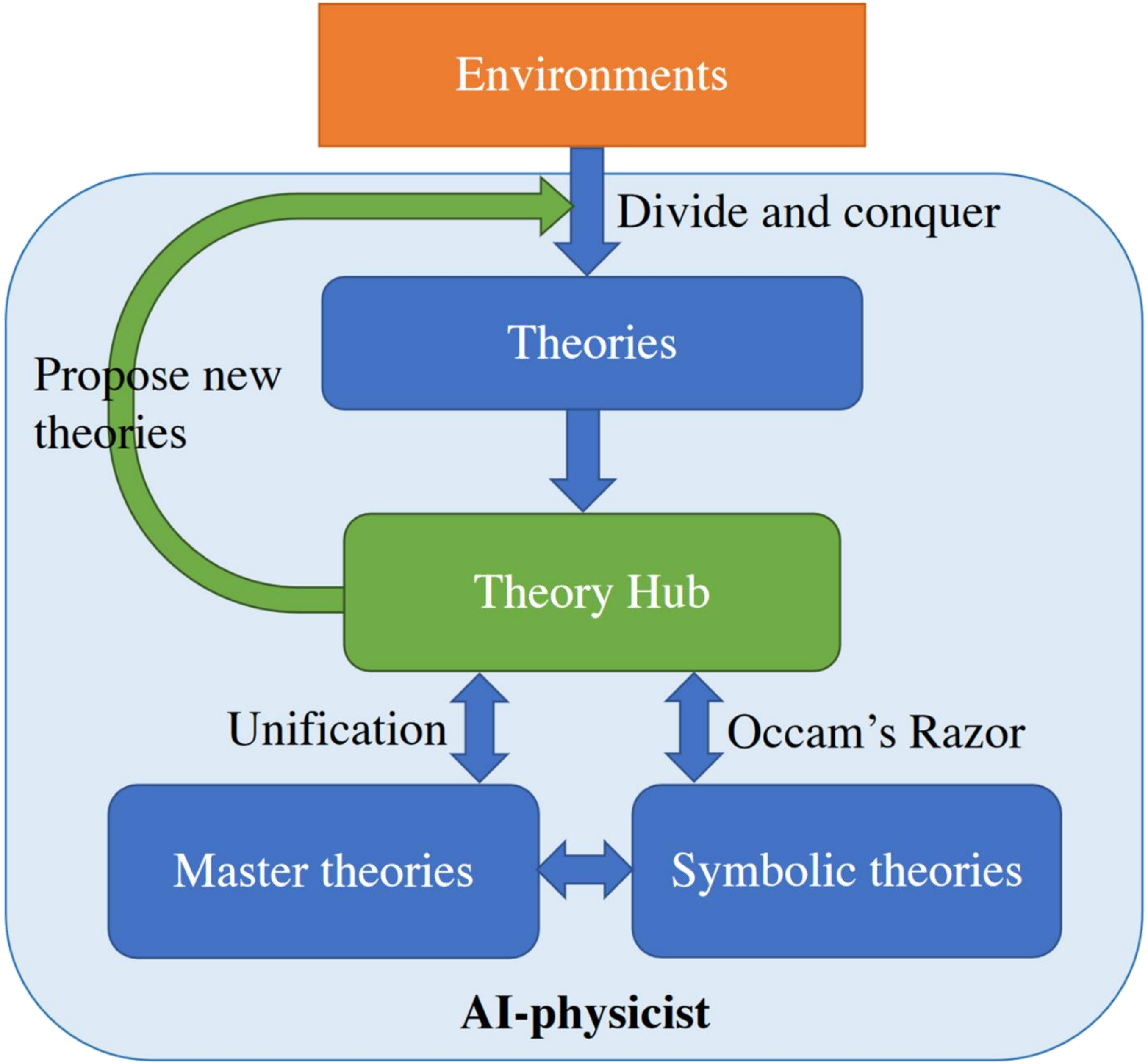
- Divide-and-conquer
- Lifelong learning
- Occam’s Razor
- Unification



Source: [1]

Architecture

- AI physicist leaves many small models applicable in different domains
- Not approximate predictions but near exact predictions with complete intelligibility



Source: [1]

Differentiable divide and conquer (DDAC)

- Idea: Not predict everything in one model but only parts of the data

  → DDAC learns prediction functions and corresponding domain classifiers

- Each prediction function will specialize in its domain due to:
  - Generalized mean loss
  - Bisected training

Differentiable divide and conquer (DDAC)

Algorithmic idea:

- Initialize random theories
- Iteratively train prediction functions and domain classifier
  - Loss function:

$$\mathcal{L}_\gamma = \sum_t \left(\frac{1}{M} \sum_{i=1}^{M} \ell[\mathbf{f}_i(\mathbf{x}_t), \mathbf{y}_t]^\gamma \right)^{1/\gamma}$$

- M: Number of theories
- ℓ: description length loss ($\ell[\mathbf{f}_i(\mathbf{x}_t), \mathbf{y}_i] = \frac{1}{2} log_2 \left(1 + \left(\frac{\mathbf{f}_i(\mathbf{x}_t) - \mathbf{y}_t}{\epsilon}\right)^2\right)$)
- γ: parameter ($\gamma = -1$)
- $\mathbf{f}$: prediction function
- $\mathbf{x}_t$: input data
- $\mathbf{y}_t$: label

Differentiable divide and conquer (DDAC)

Training:

- $\mathbf{f}_\theta = (\mathbf{f}_1, \ldots, \mathbf{f}_M),\ \mathbf{c}_\phi = (c_1, \ldots, c_M)$
- Gradient descent on $\mathbf{f}_\theta$:

$$\mathbf{g}_\mathbf{f} \leftarrow \nabla_\theta \mathcal{L}[\mathcal{T}, D, \ell]$$

- Gradient descent on $\mathbf{c}_\phi$:

$$b_t \leftarrow \arg\min_i \ell[\mathbf{f}_i(\mathbf{x}_t), \mathbf{y}_t] \forall t$$

$$\mathbf{g}_\mathbf{c} \leftarrow \nabla_\phi \sum_{(\mathbf{x}_t, \cdot) \in D} \mathrm{CrossEntropy}[\mathrm{softmax}(\mathbf{c}_\phi(\mathbf{x}_t)), b_t]$$

Occam's Razor – The simpler the better!

- How to characterize simplicity? → Description length!
- Finding the minimum description length is a hard problem!
- Therefore Heuristic for description length:

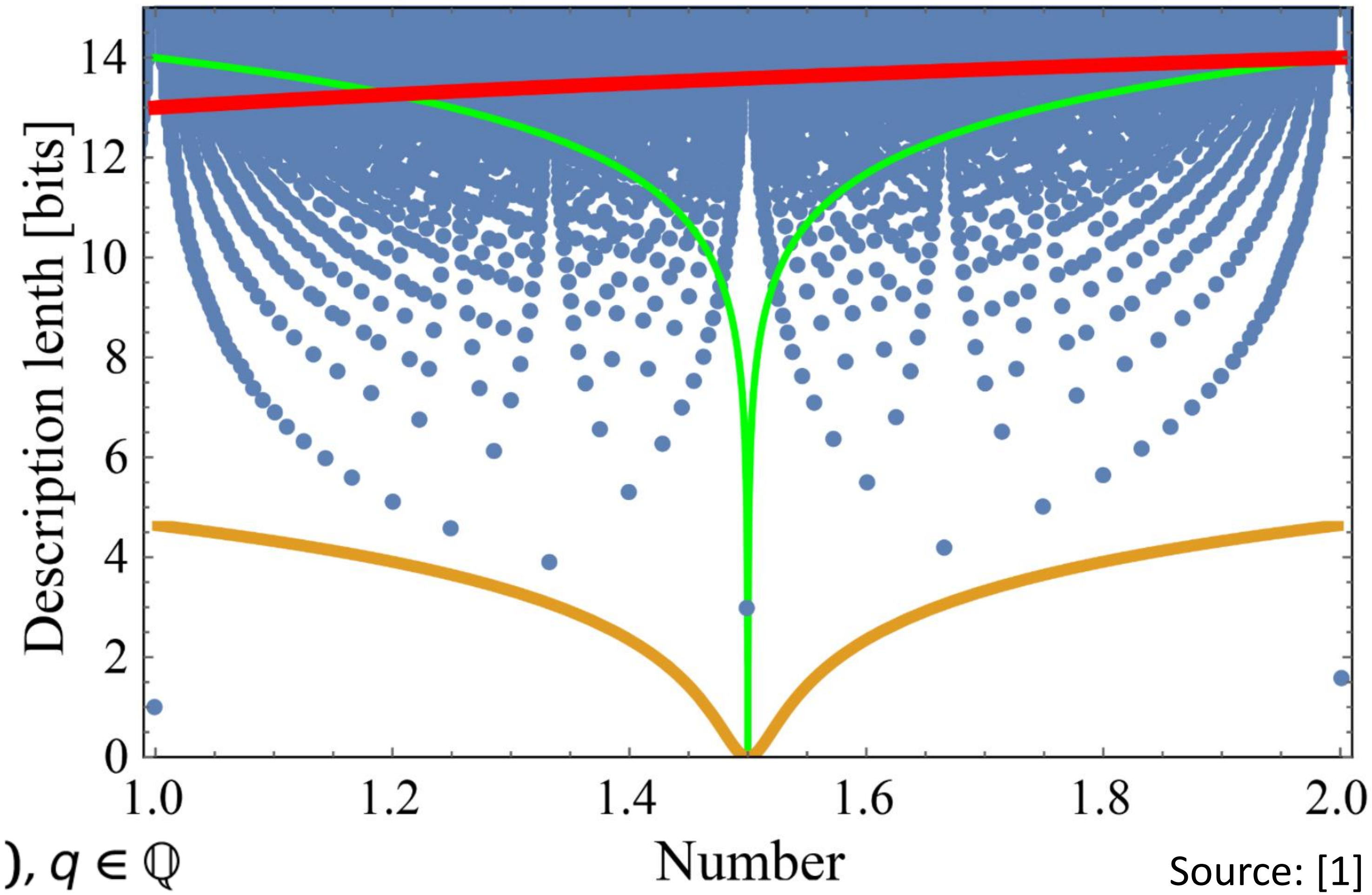
$$\mathrm{DL}(n) = \log_2(n), n \in \mathbb{N}$$

$$\mathrm{DL}(m) = \log_2(1 + |m|), m \in \mathbb{Z}$$

$$\mathrm{DL}(q = \frac{m}{n}) = \log_2((1 + |m|)n), q \in \mathbb{Q}$$

$$\mathrm{DL}(r) = \log_+\left(\frac{r}{\epsilon}\right), r \in \mathbb{R} \text{ and } \log_+(x) = \frac{1}{2}\log_2(1 + x^2)$$

Occam's Razor



Source: [1]

$$DL\left(q = \frac{m}{n}\right) = \log_2((1+m)n),\ q \in \mathbb{Q}$$

$$DL(r) = \log_+\left(\frac{r}{\epsilon}\right),\ r \in \mathbb{R} \text{ and } \log_+(x) = \frac{1}{2}\log_2(1+x^2)$$

Precision: $\epsilon = 2^{-14} \approx 6 * 10^{-5}$

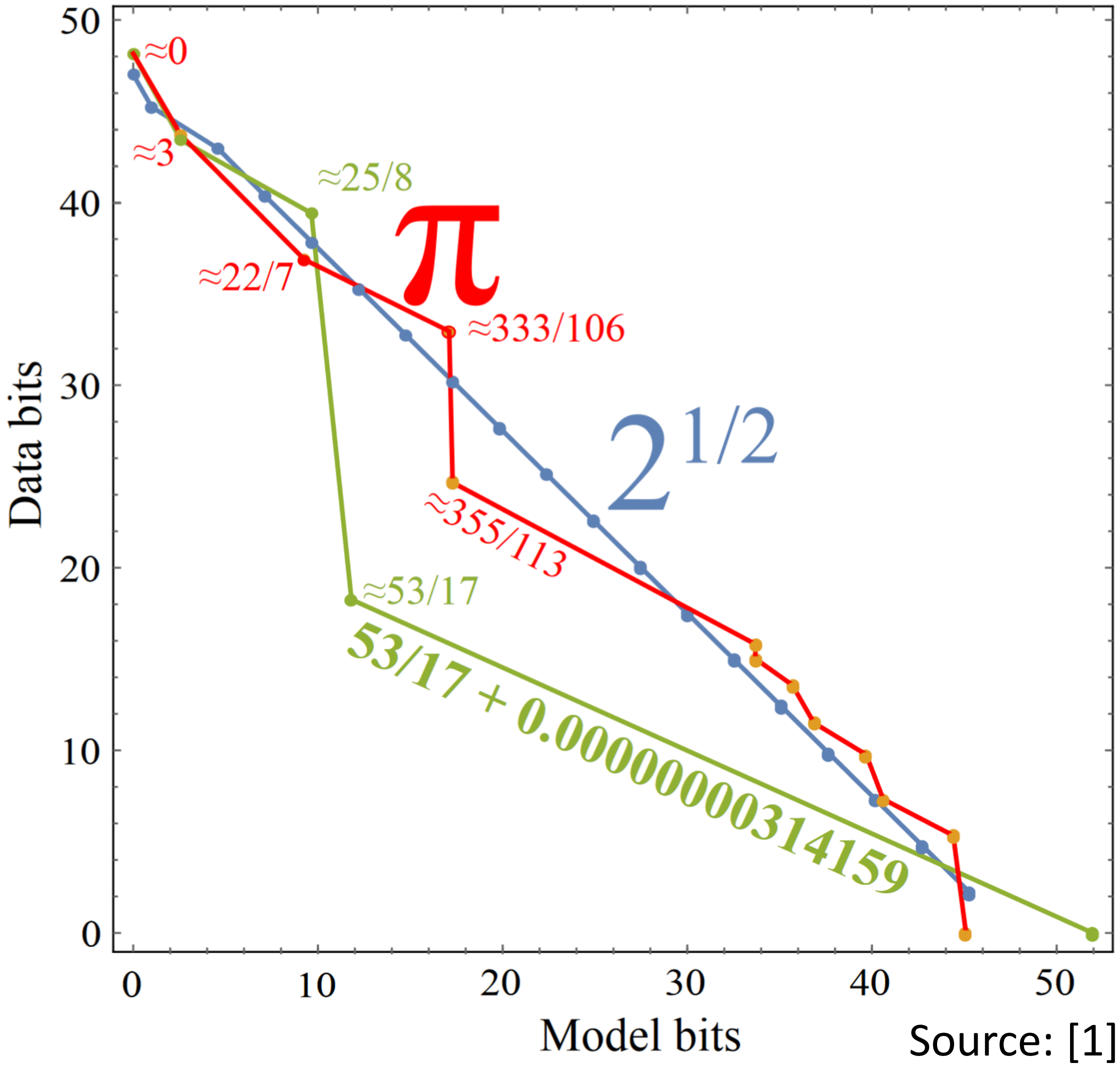
Occam's Razor

- The total description length (DL):

$$DL(\mathcal{T}, D) = DL(\mathcal{T}) + \sum_t DL(\mathbf{u}_t) \qquad \mathbf{u}_t = |\hat{\mathbf{y}}_t - \mathbf{y}_t|$$

- First term is the DL of the parameters of the model
- Second term is the DL of the prediction error
- Algorithmically:
  1. Set all parameters to real numbers to minimize $\sum_t DL(\mathbf{u}_t)$
  2. Finding close rational numbers with continued fraction expansion to minimize $DL(\mathcal{T}, D)$

Occam’s Razor



Source: [1]

Unification

- Goal: Finding underlying similarities between theories and unify them

→ Master theory $\mathscr{T} = (\mathbf{f}_\mathbf{p}, \cdot)$, varying the parameter vector $\mathbf{p} \in \mathbb{R}^n$ can generate a continuum of theories.

- Algorithmic idea:
  1. Description length of every prediction function (as symbolic function)
  2. Clustering on the theories
  3. Finding similarities and variations between the symbolic functions in ever cluster

Unification – The algorithm

Algorithm 4 AI Physicist: Theory Unification

Require Hub: theory hub
Require C: initial number of clusters

Preparation

for $(\mathbf{f}_i, c_i)$ **in Hub**.all-symbolic-theories **do**:
    $\mathrm{dl}^{(i)} \leftarrow \mathrm{DL}(\mathbf{f}_i)$
end for

Clustering

$\{S_k\} \leftarrow$Cluster $\{\mathbf{f}_i\}$ into C clusters based on $\mathrm{dl}^{(i)}$

Generalization

for S_k **in** $\{S_k\}$ **do**:
    $(\mathbf{g}_{i_k}, \mathbf{h}_{i_k}) \leftarrow$ **Canonicalize**$(\mathbf{f}_{i_k})$, $\forall \mathbf{f}_{i_k} \in S_k$
    $\mathbf{h}_k^* \leftarrow$ Mode of $\{\mathbf{h}_{i_k} | \mathbf{f}_{i_k} \in S_k\}$.
    $G_k \leftarrow \{\mathbf{g}_{i_k} | \mathbf{h}_{i_k} = \mathbf{h}_k^*\}$
    $\mathbf{g}_{\mathbf{p}_k} \leftarrow$Traverse all $\mathbf{g}_{i_k} \in G_k$ with synchronized steps, replacing the coefficient by a $\mathbf{p}_{j_k}$ when not all coefficients at the same position are identical.
    $\mathbf{f}_{\mathbf{p}_k} \leftarrow \mathrm{toPlainForm}(\mathbf{g}_{\mathbf{p}_k})$
end for
$\mathscr{T} \leftarrow \{(\mathbf{f}_{\mathbf{p}_k}, \cdot)\}$, $k = 1, 2, \ldots C$
$\mathscr{T} \leftarrow \mathrm{MergeSameForm}(\mathscr{T})$
return $\mathscr{T}$

Transformation

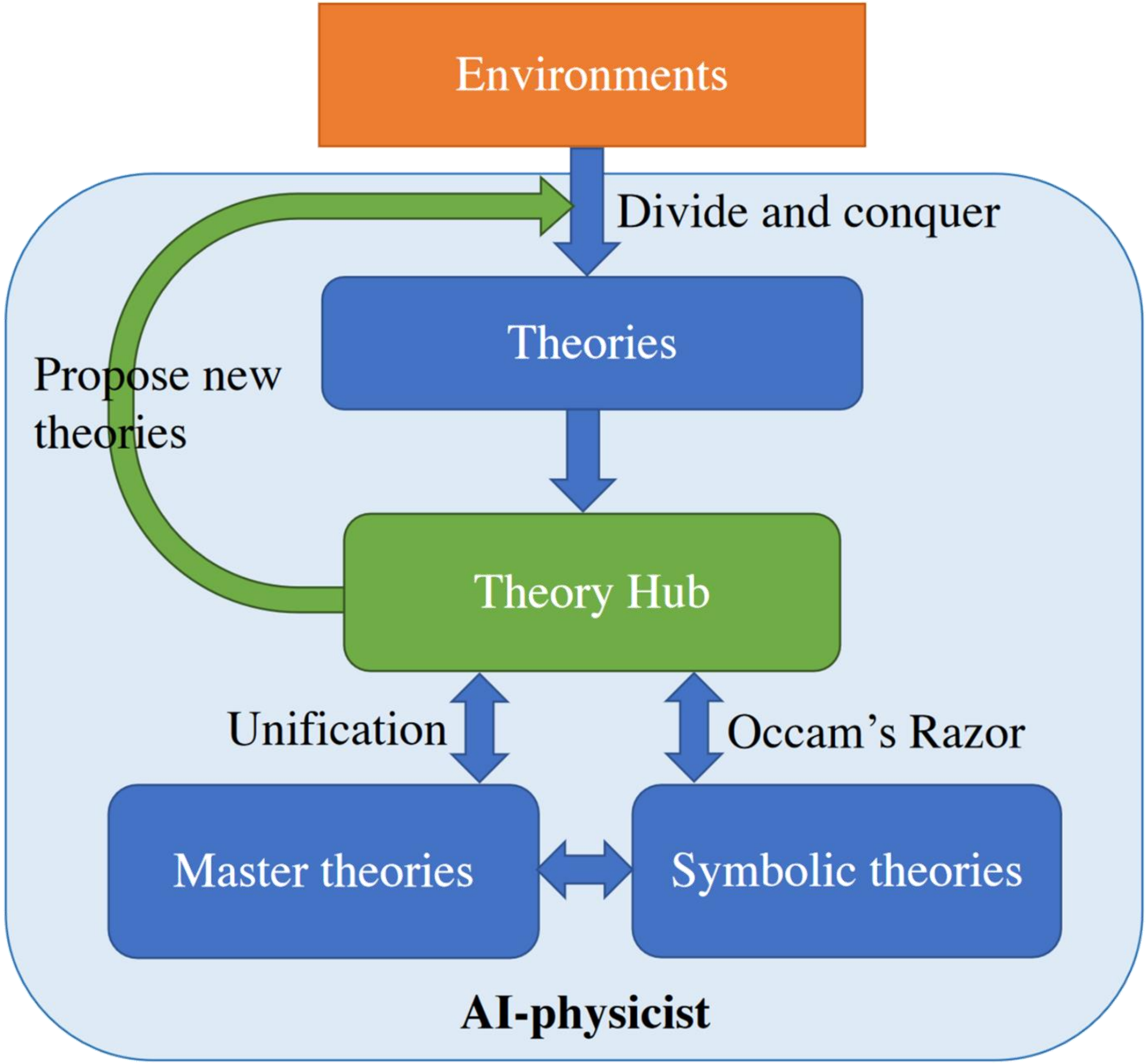
subroutine Canonicalize$(\mathbf{f}_i)$:
$\mathbf{g}_i \leftarrow \mathrm{ToTreeForm}(\mathbf{f}_i)$
$\mathbf{h}_i \leftarrow$ Replace all non-input coefficient by a symbol s
return $(\mathbf{g}_i, \mathbf{h}_i)$

Example on Blackboard ←

Source: [1]

Life long learning

- Idea: Past experiences/knowledge give us the ability to model new environment faster
- Represented in the architecture through the theory hub



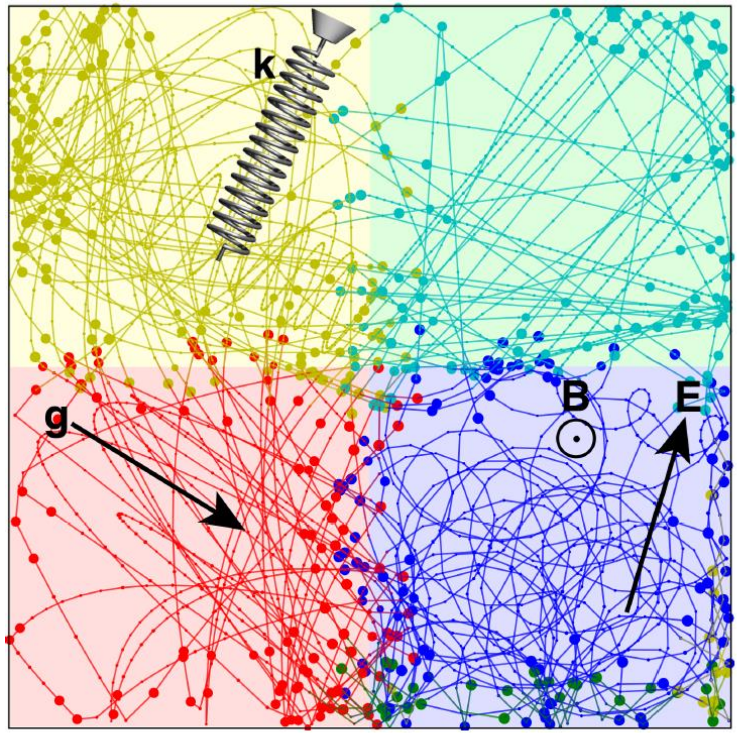
Source: [1]

The experiments

- AI physicist was tested in two randomized environments:
  1. Mystery world

     $V \propto (ax + by + c)^n$

     1. Gravity (n=1)
     2. E-field (and optionally uniform B-field) (n=1)
     3. Springs/Hooke's law (n=2)
     4. Bounce boundaries (n=∞)
  2. Charged double pendulum in two adjacent E-fields



Own visualization, data from [2]

The experiments – results

Benchmark	Baseline	Newborn	AI Physicist
$\log_{10}$ mean-squared error	-3.89	-13.95	-13.88

$$\boxed{\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \hat{\mathbf{y}}_i)^2}$$

Source: [1]

The experiments – results

Benchmark	Baseline	Newborn	AI Physicist
$\log_{10}$ mean-squared error	-3.89	-13.95	-13.88
Classification accuracy	67.56%	100.00%	100.00%

Source: [1]

The experiments – results

Benchmark	Baseline	Newborn	AI Physicist
$\log_{10}$ mean-squared error	-3.89	-13.95	-13.88
Classification accuracy	67.56%	100.00%	100.00%
Fraction of worlds solved	0.00%	90.00%	92.50%

- A domain is solved then:
  - Any rational number must be calculated exactly
  - Any irrational number in the theory must be recovered by an accuracy of 10^{-4}

Source: [1]

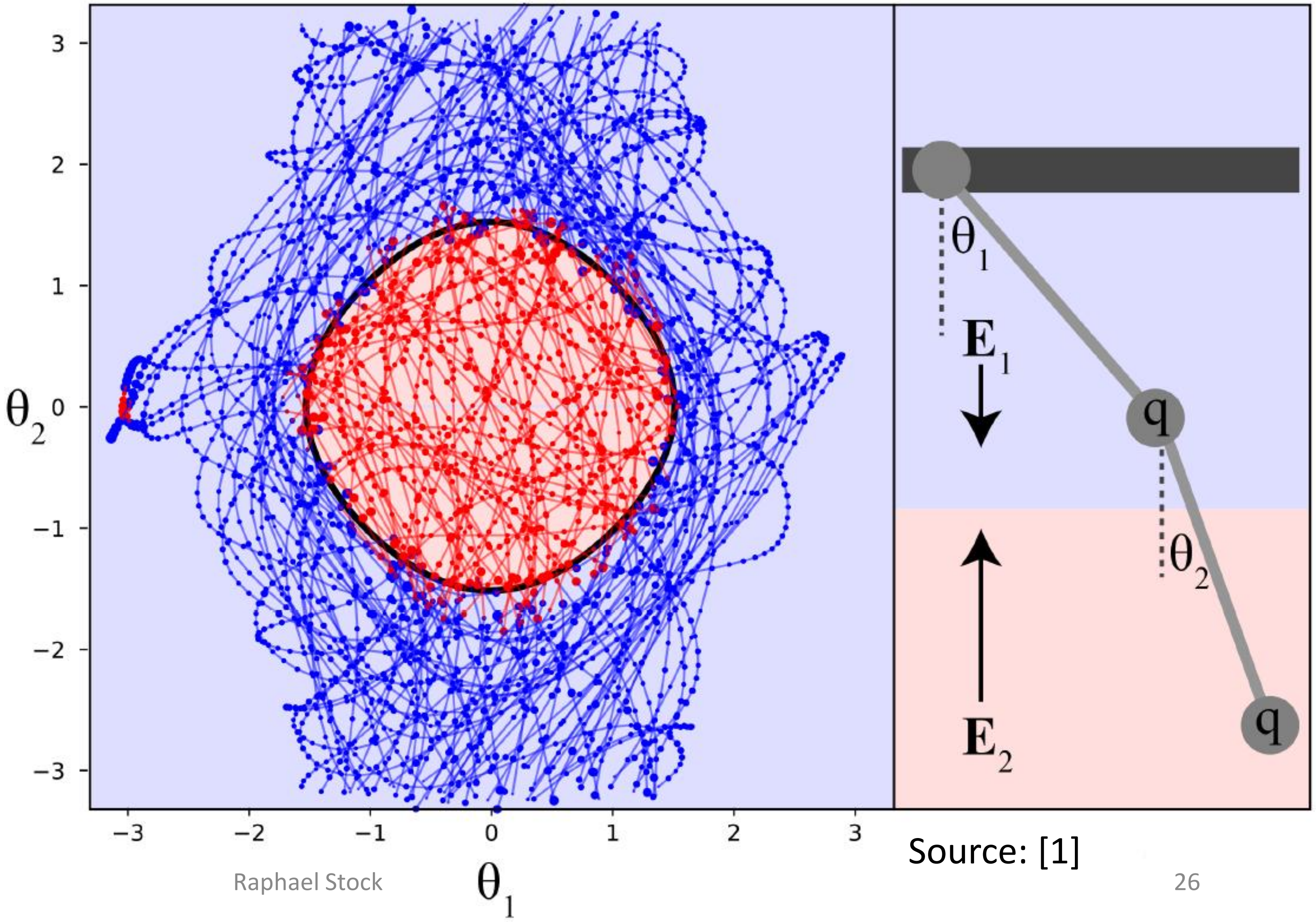
The experiments – results

Benchmark	Baseline	Newborn	AI Physicist
$\log_{10}$ mean-squared error	-3.89	-13.95	-13.88
Classification accuracy	67.56%	100.00%	100.00%
Fraction of worlds solved	0.00%	90.00%	92.50%
Description length for $\mathbf{f}$	11,338.7	198.9	198.9
Epochs until 10^{-2} MSE	95	83	15
Epochs until 10^{-4} MSE	6925	330	45
Epochs until 10^{-6} MSE	∞	5403	3895
Epochs until 10^{-8} MSE	∞	6590	5100

Source: [1]

The experiments – Charged double pendulum

- No double-pendulum was solved exactly
- Domain classification accuracies:
  - Baseline: 76.9%
  - Newborn: 96.5%



Source: [1]

Summary & Conclusion

- Approach towards AI physicist for finding EOM's, whose architecture is orientated on principles a real physicist uses
- The agent works well on simple "mystery world"
- Agent has troubles for harder problems

Discussion

- Classification accuracy is whitewashed by only considering domain centres
- Questionable benchmark with baseline neural network compensating the more complex AI physicist only by double the neurons
- AI physicist obviously outperforms on solved domains and description length because of architecture
- Mystery worlds are limited in their appearance

Thank you for your Attention!

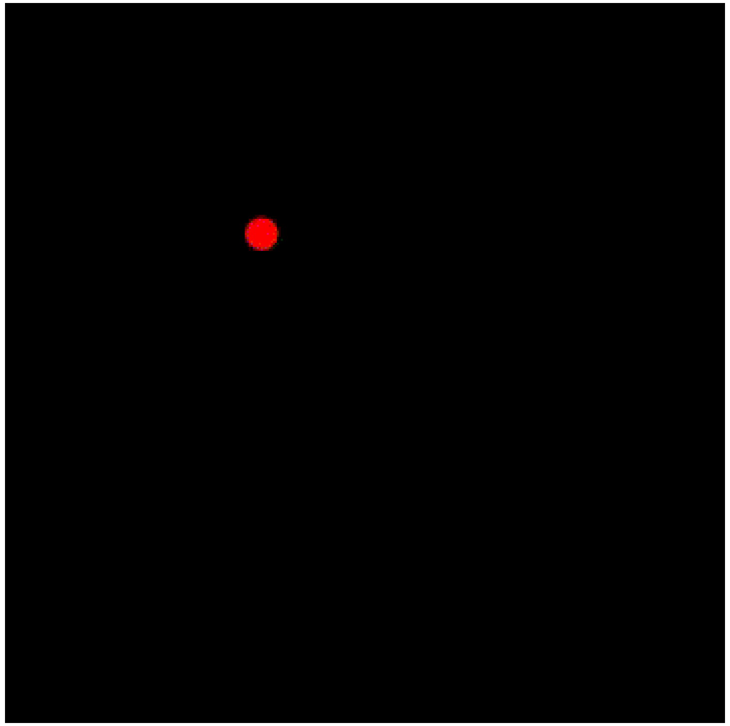
Sources

[1] Tailin Wu, Max Tegmark. Toward an AI Physicist for Unsupervised Learning

[2] https://space.mit.edu/home/tegmark/aiphysicist.html

[3] https://www.sueddeutsche.de/panorama/weltrekord-im-kopfrechnen-in-11-8-sekunden-zur-13-wurzel-1.664320

Further questions?



Source: [2]