

What does a good ML theory look like?

– A physicist's perspective

Speaker: Ziming Liu (刘子鸣), MIT, June 2023

Personal website: <https://kindxiaoming.github.io/>



Ziming Liu (刘子鸣)
PhD student
@MIT and IAIFI

Biography [cv](#)

I am a third-year Physics PhD student at [MIT](#) and [IAIFI](#), advised by [Prof. Max Tegmark](#). I have interned at Microsoft Research Asia. Before that, I received my Bachelor's degree in physics from Peking University. Prior to that, my memories are sealed in my hometown, Wuhan.

I research on the intersection of artificial intelligence and physics in general, including but not limited to:

- (1) AI for physics: extracting physical insights (e.g. conservation laws and symmetries) from data, improving prediction accuracy and sampling efficiency for data analysis in physics;
- (2) Physics for AI: developing effective theories to understand the dynamics and generalization of neural networks, and building physics-inspired machine learning models.

知乎



Email: zmliu.at.mit.edu | liu_zi_ming.at.pku.edu.cn

Overview

- Theory in general
- Classical ML theories
 - PAC learning
 - Statistical Physics
- What I think is good ML theory: A physics-like ML theory?

Theory in general

What is theory?

Dictionary

Definitions from [Oxford Languages](#) · [Learn more](#)



the·o·ry



Learn to pronounce

noun

a supposition or a system of ideas intended to explain something, especially one based on general principles independent of the thing to be explained.

"Darwin's theory of evolution"

Similar:

hypothesis

thesis

conjecture

supposition

speculation

postulation



- a set of principles on which the practice of an activity is based.
"a theory of education"
- an idea used to account for a situation or justify a course of action.
"my theory would be that the place has been seriously mismanaged"

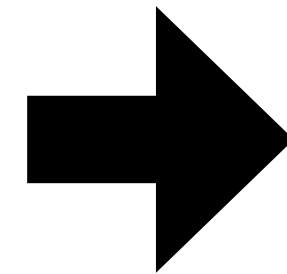
Why theory?

10000000 bits

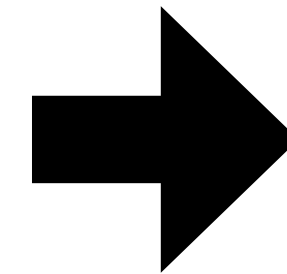
100 bits

10000000 bits

Past Observations



Theory



Predicting new observations

Information compression

Easy to store and communicate
(human brains are limited)

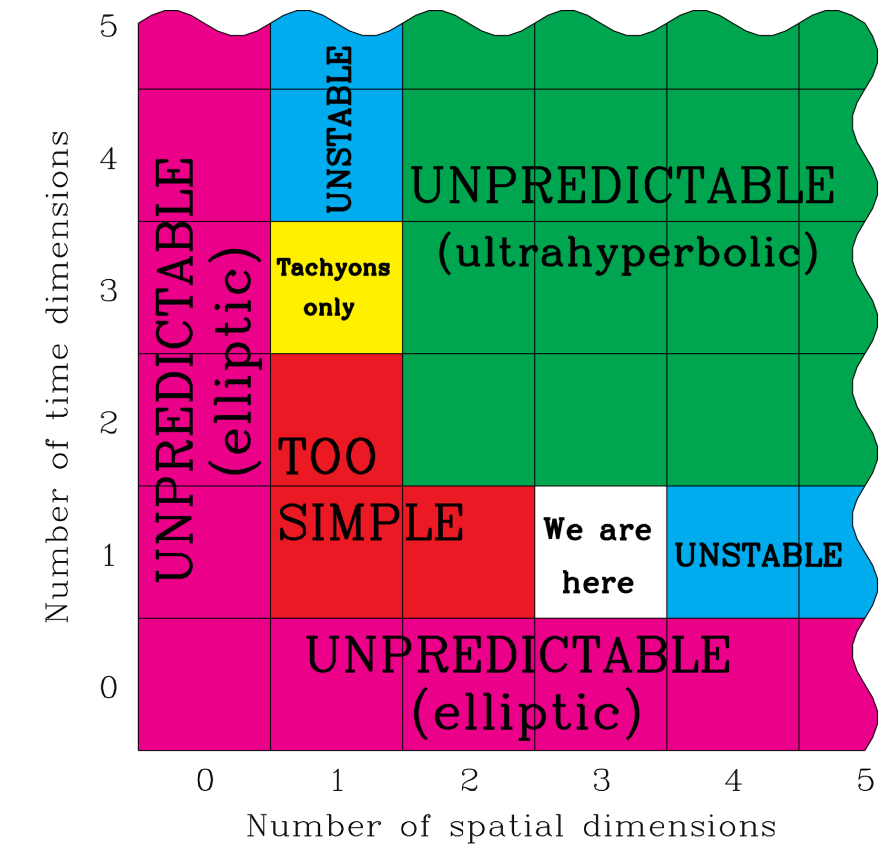
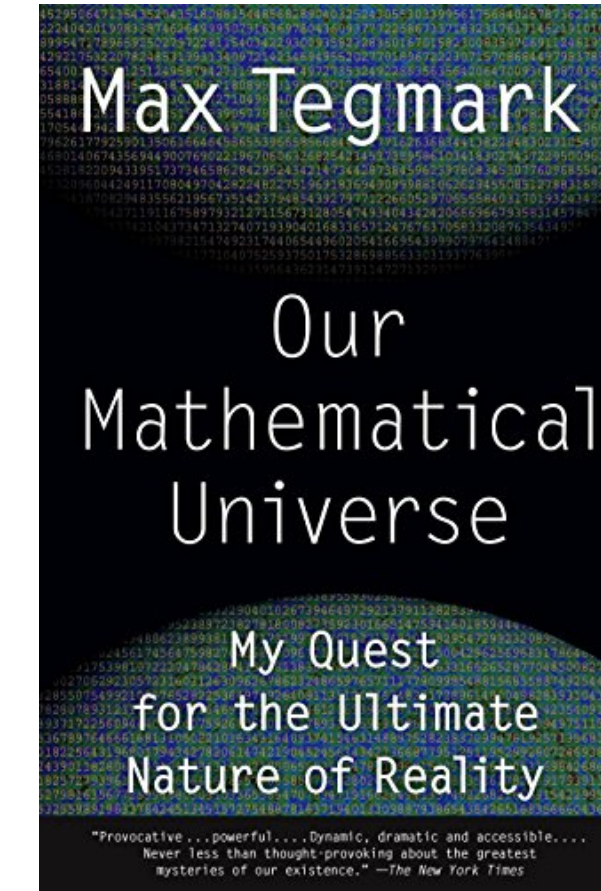
Information gain

Predicting future worlds is essential to survival



Side: Why is theory possible?

- * Unreasonable effectiveness of mathematics
- * Anthropic principle



Anthropic principle

🌐 40 languages ▾

Article [Talk](#)

[Read](#) [Edit](#) [View history](#) [Tools](#) ▾

From Wikipedia, the free encyclopedia

"Anthropic bias" redirects here. For the book by Nick Bostrom, see [Anthropic Bias \(book\)](#).

The **anthropic principle**, also known as the "observation selection effect",^[1] is the hypothesis, first proposed in 1957 by [Robert Dicke](#), that the range of possible observations that we could make about the universe is limited by the fact that observations could only happen in a universe capable of developing intelligent life in the first place.^[2] Proponents of the anthropic principle argue that it explains why this universe has the [age](#) and the [fundamental physical constants](#) necessary to accommodate conscious life, since if either had been different, we would not have been around to make observations. Anthropic reasoning is often used to deal with the notion that the universe seems to be [finely tuned for the existence of life](#).^[3]

I don't want to sound philosophical here, but my take is that:
We should not take the existence of (good) theories for granted!

What is a good theory?

16 October 1964, Volume 146, Number 3642

SCIENCE

Strong Inference

Certain systematic methods of scientific thinking may produce much more rapid progress than others.

John R. Platt

“nature” or the experimental outcome chooses—to go to the right branch or the left; at the next fork, to go left or right; and so on. There are similar branch points in a “conditional computer program,” where the next move depends on the result of the last calculation. And there is a “conditional inductive tree” or “logical tree” of this kind written out in detail in many first-year chemistry books, in the table of steps for qualitative analysis of an unknown sample, where the student is led through a real problem of consecutive inference: Add reagent A; if

new particles explicitly enough so that if they are not found the theories will fall. As the biologist W. A. H. Rushton has said (11), “A theory which cannot be mortally endangered cannot be alive.” Murray Gell-Mann and Yuval Ne’eman recently used the particle grouping which they call “The Eightfold Way” to predict a missing particle, the Omega-Minus, which was then looked for and found (12). But one alternative branch of the theory would predict a particle with one-third the usual electronic charge, and it was not found in the experiments, so this branch must be rejected.

A new theory predicts an event E to be very likely $p_{\text{new}} \approx 1$,
but old theories think that E is very unlikely $p_{\text{old}} \ll 1$.

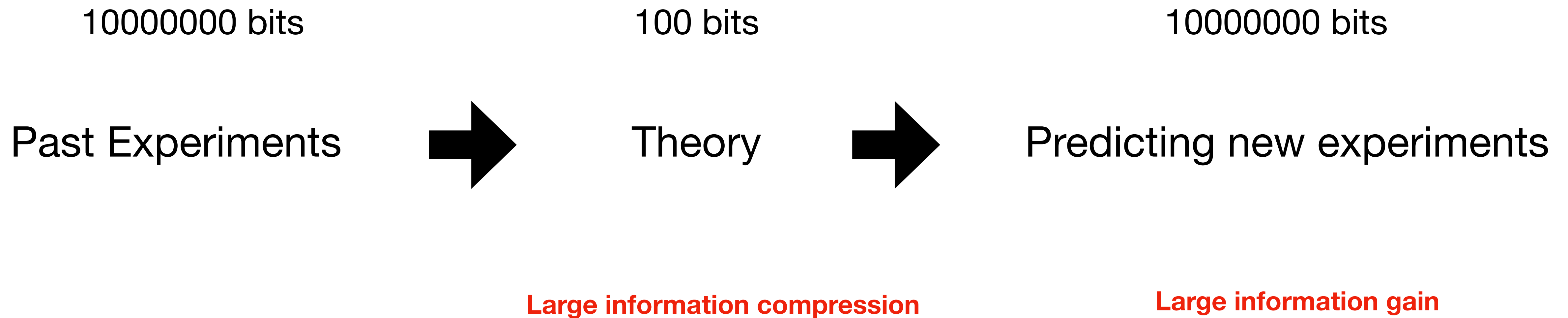
E happens!



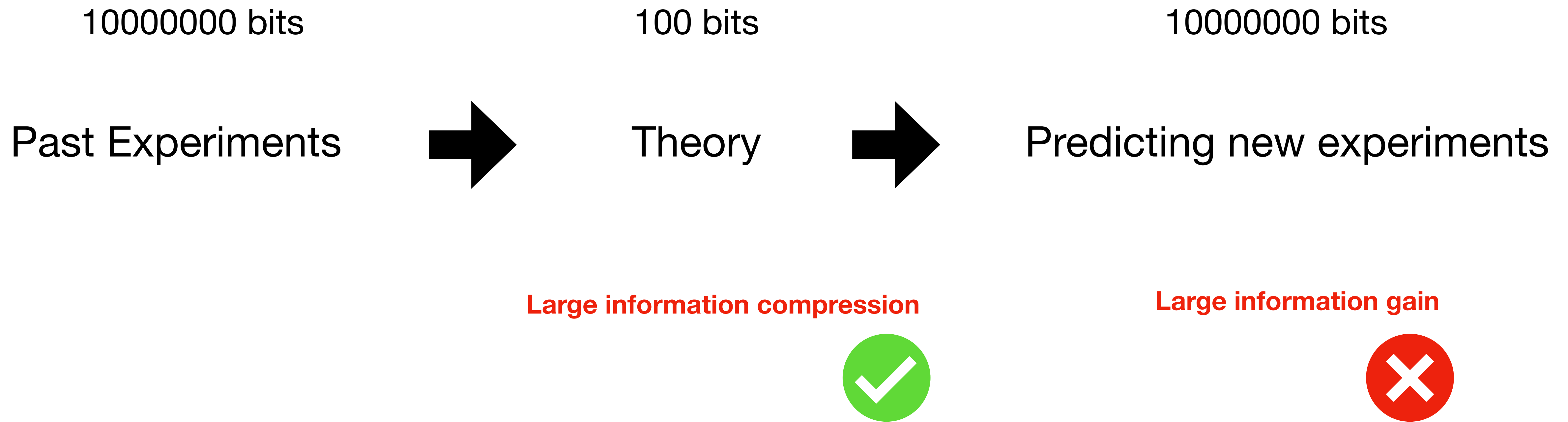
Information gain (surprisal): $\log(p_{\text{new}}/p_{\text{old}})$

Also **information compression** (8 mesons unified by 1 group)

What is a good theory?

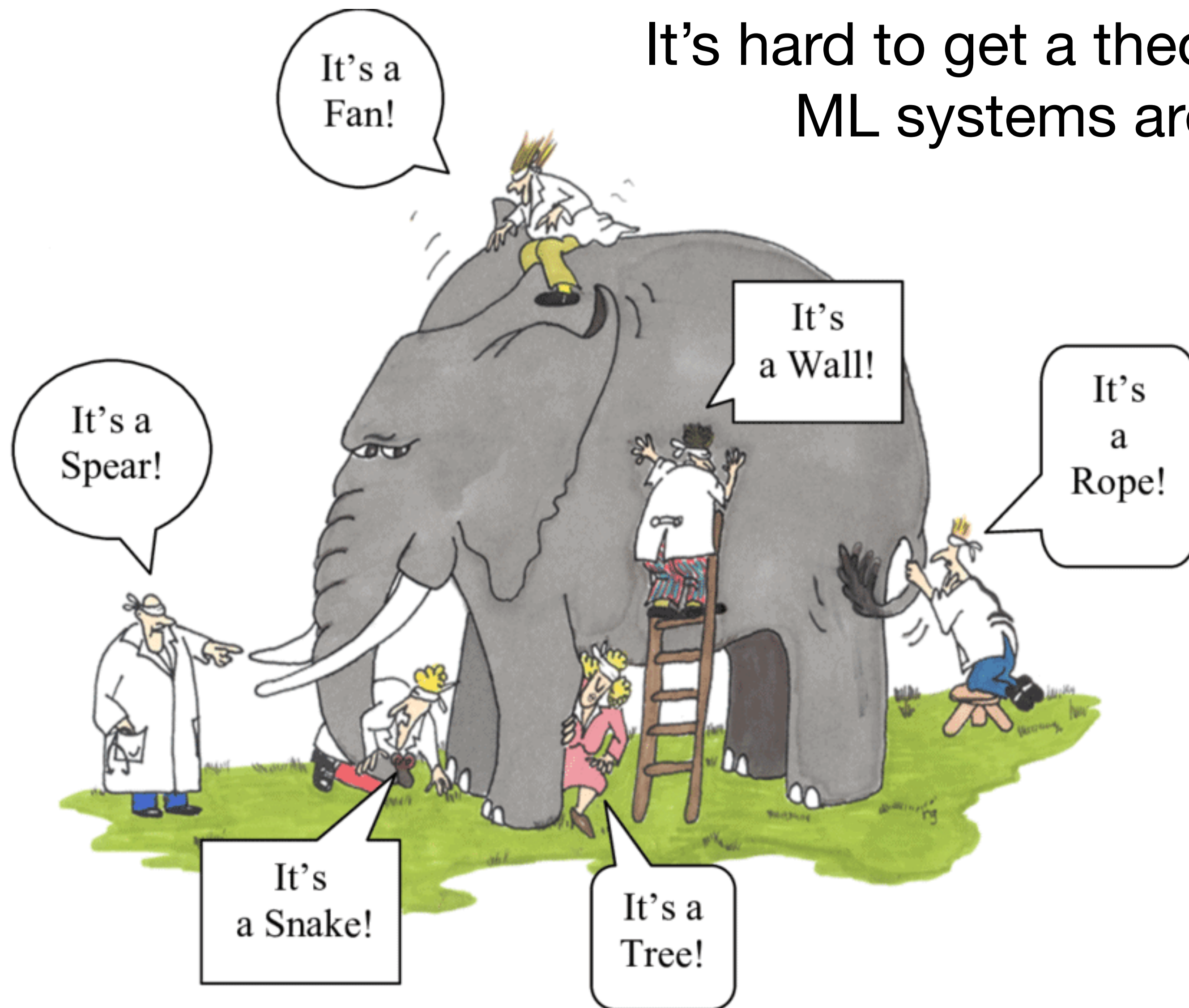


Are classical ML theories good?



Why is ML theory hard?

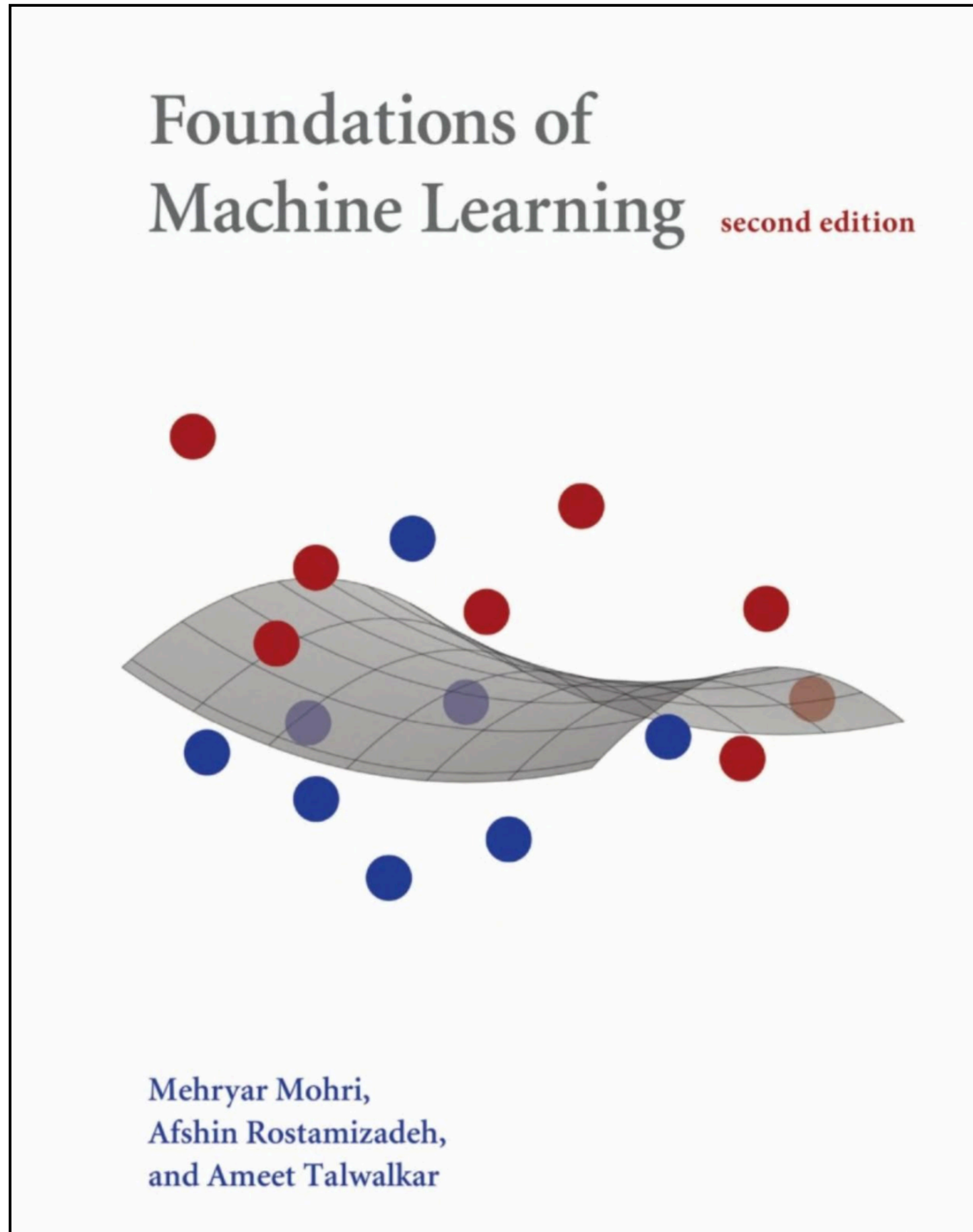
It's hard to get a theory for any type of complex system!
ML systems are of course complex systems!



Classical ML theories

Statistics (PAC), Physics (stat mech)

Probably Approximately Correct (PAC) Learning



Probably approximately correct learning

Article [Talk](#)

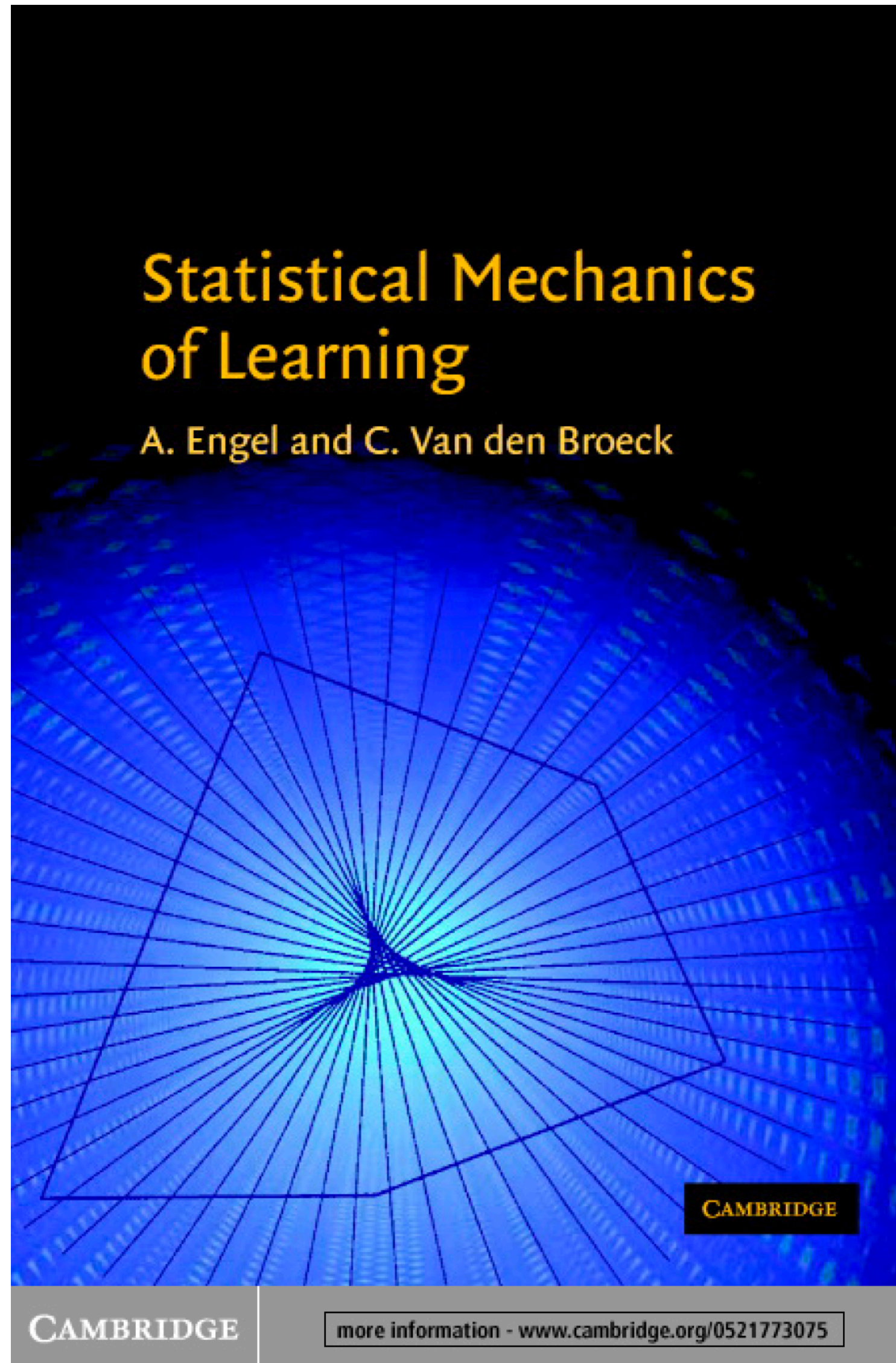
From Wikipedia, the free encyclopedia

In [computational learning theory](#), **probably approximately correct (PAC) learning** is a framework for mathematical analysis of [machine learning](#). It was proposed in 1984 by [Leslie Valiant](#).^[1]

In this framework, the learner receives samples and must select a generalization function (called the *hypothesis*) from a certain class of possible functions. The goal is that, with high probability (the "probably" part), the selected function will have low [generalization error](#) (the "approximately correct" part). The learner must be able to learn the concept given any arbitrary approximation ratio, probability of success, or [distribution of the samples](#).

Statistical Mechanics

Correspond to ensemble average in stat mech



In order to implement the machinery of statistical mechanics for the analysis of learning problems one hence has to determine typical values of interesting quantities such as the generalization error. It is, however, in general rather difficult to calculate the most probable values since this requires one more or less to calculate the complete probability distribution. Fortunately, for some quantities the most probable value coincides with the *average* value, which is much more easily accessible analytically. If additionally the variance of the probability distribution tends to zero in the thermodynamic limit such a quantity is called *self-averaging* since the probability for a value different from its average tends to zero in the thermodynamic limit. It is very important to always remember, however, that *not all* interesting quantities are automatically self-averaging.⁵ We will therefore find that the *identification* of the self-averaging quantities is the first and rather crucial step in the statistical mechanics analysis of a learning problem.

To summarize, the mathematical analysis of learning from examples requires a proper treatment of the various probabilistic elements essential for such a problem. Statistical mechanics mainly considers the special scenario of a teacher and student neural network and aims at producing *exact* results for the *typical* learning behaviour. This becomes possible by considering the thermodynamic limit, in which both the number of adjustable couplings of the student network and the number of examples in the training set diverge. After identifying the self-averaging quantities of the problem the typical performance is characterized by calculating their averages over the relevant probability distributions.

Example: axis-aligned rectangle

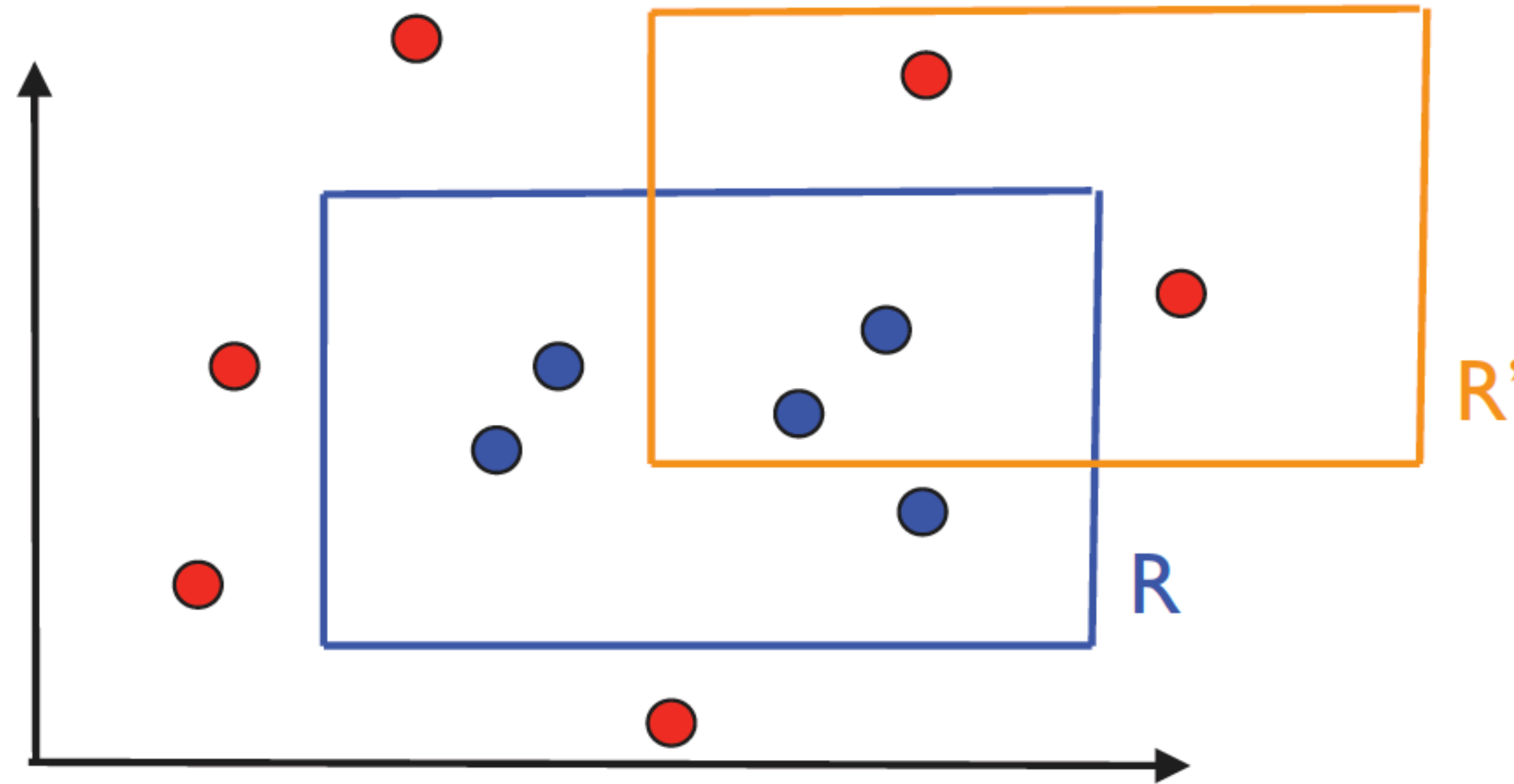
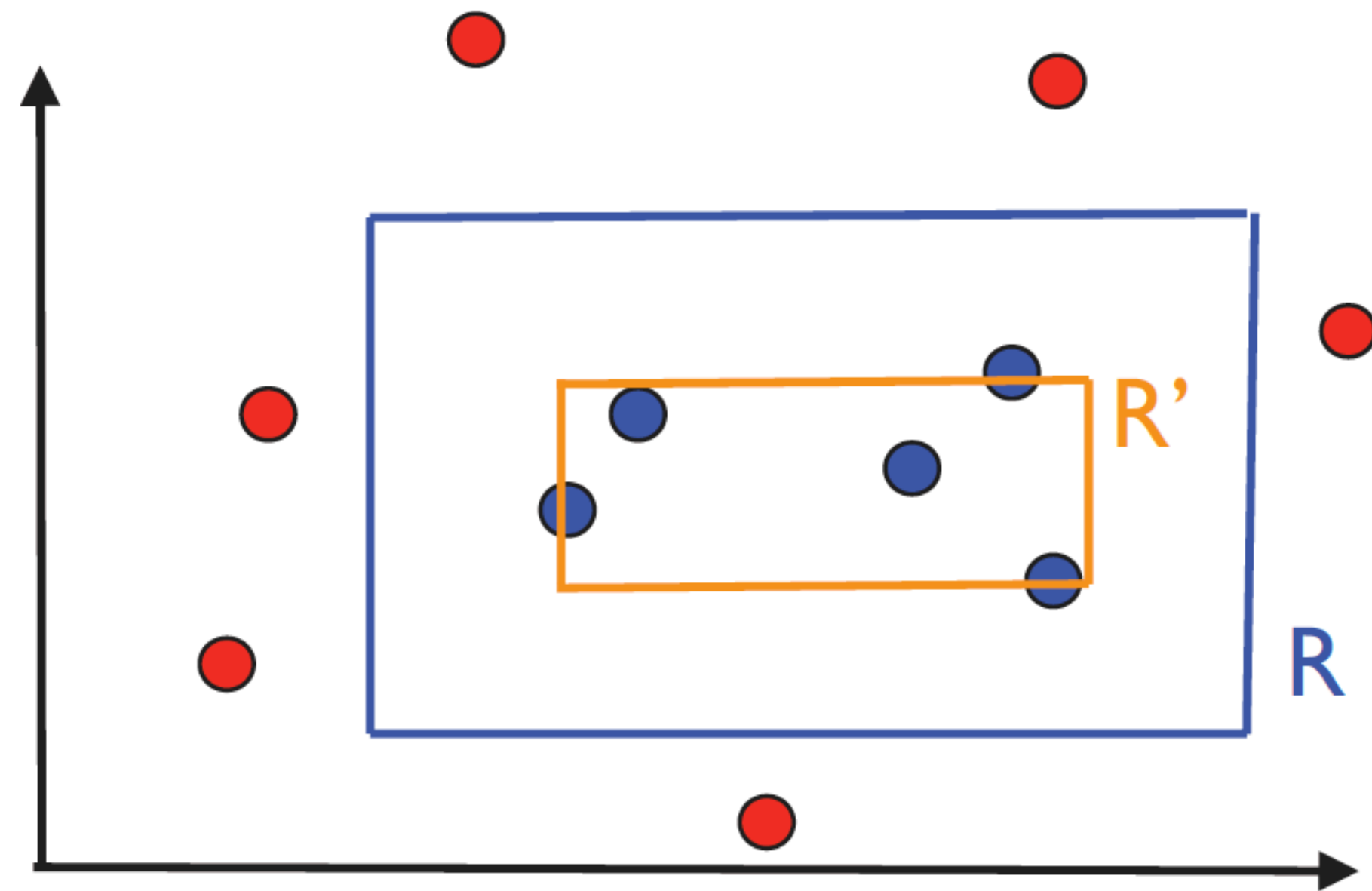


Figure 2.1 Target concept R and possible hypothesis R' . Circles represent training instances. A blue circle is a point labeled with 1, since it falls within the rectangle R . Others are red and labeled with 0.

Example (PAC)

Care about worst case



$$R(R_S) \leq \frac{4}{m} \log \frac{4}{\delta}.$$

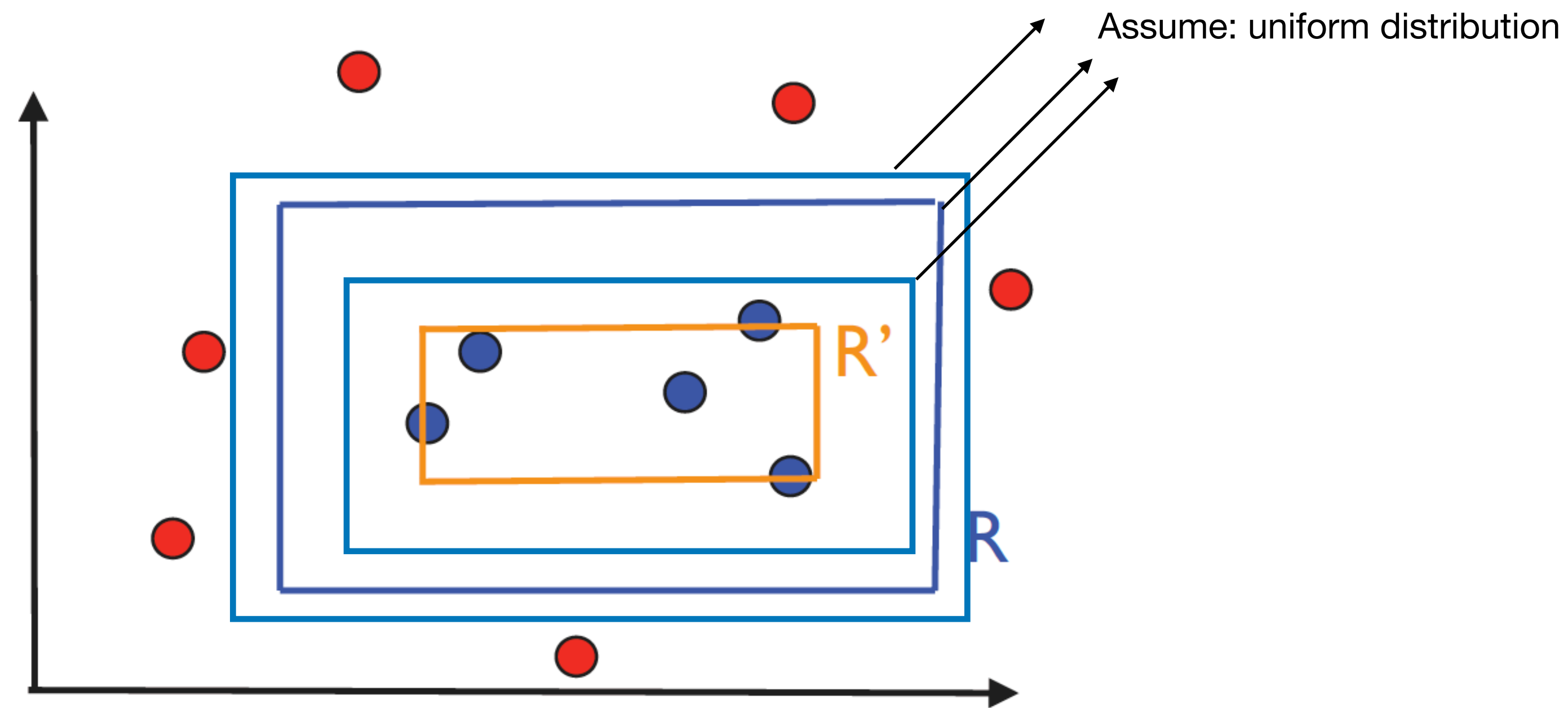
error

number of sample

error probability

Example (Stat Mech)

Care about global/typical/average behaviour



$$R \sim 1/m$$

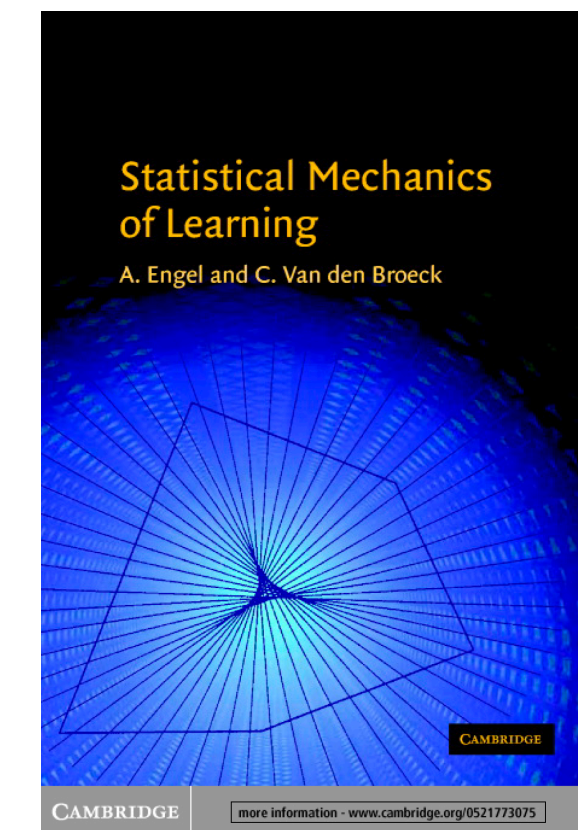
error

number of sample

PAC vs statistical physic: different

We will elucidate some aspects of the *PAC* approach in chapter 10. Here we just note that it allows one to derive very general *bounds* for the performance of learning scenarios by studying *worst case* situations. These worst cases generally include the worst possible choice of the student vector for a given value of the training error, the worst choice of the target rule and the worst realization of the training set. It is hence quite possible that the results obtained are over-pessimistic and may not characterize the average or most probable behaviour.

The theoretical description of learning from examples outlined in the present book is based on concepts different from *PAC* learning. Contrary to mathematical statistics, statistical mechanics tries to describe the *typical* behaviour *exactly* rather than to *bound* the worst case. In statistical mechanics *typical* means not just *most probable* but in addition it is required that the probability for situations different from the typical one can be made arbitrarily small. This remarkable property is achieved by what is called the *thermodynamic limit*. In this limit the number N of degrees of freedom tends to infinity, and the success of statistical mechanics rests on the fact that the probability distributions of the relevant quantities become sharply peaked around their maximal values in this limit.

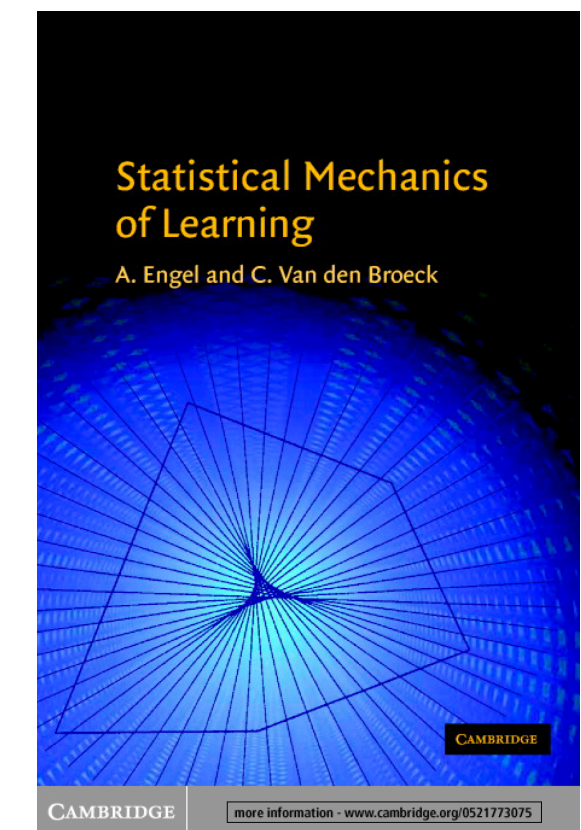


PAC vs statistical physic: same

To summarize, the mathematical analysis of learning from examples requires a proper treatment of the various probabilistic elements essential for such a problem. Statistical mechanics mainly considers the special scenario of a teacher and student neural network and aims at producing *exact* results for the *typical* learning behaviour. This becomes possible by considering the thermodynamic limit, in which both the number of adjustable couplings of the student network and the number of examples in the training set diverge. After identifying the self-averaging quantities of the problem the typical performance is characterized by calculating their averages over the relevant probability distributions.





Both PAC and Stat mech assume to know:

- (1) hypothesis set.
- (2) ground truth hypothesis.
- (3) how to select a hypothesis.



Limitation of PAC/Stat Mech Learning

Both PAC and Stat mech learning assume to know:

- (1) hypothesis set. In deep learning, depending on neural architectures. /
- (2) ground truth hypothesis. However, in deep learning, data/algorithm structures are unclear. 
- (3) how to select a hypothesis. However, in deep learning, inductive biases are unclear. 

Prof. Yang Yuan's view

<https://zhuankan.zhihu.com/p/634193692>

AI还需要理论么？



袁洋

1,774 人赞同了该文章

最近，有几个朋友以不同的方式跟我表达了同一个意思：**过去十年，机器学习理论没有给AI的发展带来任何帮助，它只是个理论圈自娱自乐的玩具。** *In the past ten years, ML theory did not bring any help to AI development. It is simply a self-entertaining toy of ML theorists.*

这个说法当然存在夸张成分，但是它背后的含义却让人难以辩驳。如果我们把AI顶会过去10年90%以上的机器学习理论论文都删去，恐怕几乎不会影响AI过去10年的发展，也几乎不会影响OpenAI推出ChatGPT。更直白一点说，那些AI大佬们恐怕根本没有时间去阅读机器学习理论论文，因为他们要忙着做真正能够推动AI前进的事情。

AI还需要理论么？

我认为，并不是AI不需要理论，而是我们之前做理论的方式有问题。

Prof. Yang Yuan's view

<https://zhuanlan.zhihu.com/p/634193692>

Engineering-like theory

登山式理论 (工程式理论)

登山式理论像登山一样，总是充满挑战，让人热血沸腾。这一类理论工作特点是目标清晰，就好像珠穆朗玛峰顶一样。当我们站在山底，目标很明确，就是要想方设法、不惜一切爬到山顶。不过，到山顶的路有千万条，我们只需要找到一条最合适的路就可以了。虽然说，我们迈出的每一步都是数学推导，但是当山峰很高很陡的时候，我们也很难快速地找到一条可行之路。这个时候，有两类工具是比较常用的：

1. 制定规划。登山之路过于漫长，我们可以找到几个重要的节点，把登山之路拆分成几个不同的阶段，每次处理其中的一个阶段。这样，把一个复杂的问题拆分成很多简单的问题，往往就会容易很多。
2. 加假设。一旦制定了明确的登顶目标，理论分析的难度就容易受到影响。很多地方不是靠制定规划就可以解决的，有的步骤你不得不用一些工具，比如绳索、直升机、木板等等才能过去。这样的工具在机器学习理论圈就是加假设。比如，我们可以假设输入 x 服从高斯分布，可以假设目标函数是光滑的，lipschitz的等等。

在很多机器学习理论的论文中，假设的选取是核心艺术。如果假设太强了，比如我们使用了传送器直接传到了山顶，那么整个登山路线显得索然无味。如果假设太弱了，比如我们连绳索都不准用，那么就会发现爬来爬去爬不到山顶。问题是，这些用于登顶的假设，在实际中往往是不完全成立的；或者说，就算成立，可能也只覆盖了一个很小的部分，不能够真正用于解释和分析实际的AI算法。

举个例子，我们观察到了LayerNorm在实际算法中效果很好，于是我们决定把LN的分析当做我们的山顶去攀登。可是，实际的数据分布到底长什么样子？我们可能不得不假设数据服从高斯分布。损失函数满足什么性质？我们可能需要假设它是光滑的。网络结构是什么样子？我们可能需要假设它是一个两层或者三层的网络，因为网络层数一多分析起来就非常困难。优化算法的步长是多少？我们可能需要假设它非常小，这样优化的过程在一个小小的邻域中才便于分析。这些假设就像是登山运动员的工具包里形形色色的工具，要清晰理解它们的用途并不容易，把它们组合起来完成登顶的任务更是一种壮举。但是，真实的训练过程往往和这些假设有一定差距：机器学习理论工作所攀登的山峰，更像是作者精心设计的理想山峰，而不是AI科学家日常真正遇到的那些。

过去十年，AI领域蓬勃发展，各种概念层出不穷。理论学家为了理解一个概念或算法，制定了很高的登山目标；但是限于工具的能力，又不得不加上各种假设助力登顶。最后，很多结论南辕北辙，得不到理论圈外部的认可，我认为这和登山式理论的研究范式是脱不开关系的。

To understand a concept or an algorithm, ML theorists have made goals that are hard to reach. Due to limits of available tools, they had to add all sorts of assumptions to reach the goals in the sky. As a result, many conclusions make no sense, and are not embraced by ML researchers outside theory.

Theory-like engineering

铺路式理论 (理论式工程)

如果说登山式理论目标明确，一切都是围绕登顶；那铺路式理论则更加佛系，完全是好奇心驱动。我把它叫做“铺路式”，可能会有一些歧义：听起来这样的理论仍然有目标要完成，毕竟铺路也是一项工程。我想澄清的是，使用“铺路”这个词，我更想强调它是从某个点出发，向四周蔓延，是一种自然而然的过程。如果我们看到了一个小池塘，就修一条到小池塘的路；如果我们看到了一个小山坡，就修一条绕开它的路。总之，修路的目标就是以修路的方式对这个世界进行四处探索，忠实地、不加假设或粉饰地去理解世界。这样的路一开始修得很慢，但是会越来越快，因为在数学的世界里，一切已有的结论都可以成为未来结论的基础；这样的路也修得很扎实，因为从头到尾都在描述世界的真实，所以修一步算一步——只要人们对这个世界有兴趣，就会想要来看看已经修好的路。

有很多数学大师有过类似的观点，我不过是拾人牙慧，换了个比方。例如，

柯西：在纯数学的领域里，似乎没有实际的物理现象来印证，也没有自然界的事物可说明，但那是数学家遥遥望见的应许之地。理论数学家不是一个发现者，而是这个应许之地的报导者。
格罗滕迪克：人们永远不应该试图证明那些并非几乎显而易见的事情。
格罗滕迪克：我脑海中浮现出的类比就像是把坚果浸入某种软化液体中。你会不时地擦拭，以便液体更好地渗透进去，其他时候则是让时间流逝。经过数周甚至数月，外壳变得更加灵活，当时间成熟时，手的力量就足够了，壳就像完美熟透的牛油果一样打开！几周前，我有了另一个形象。未知的事物在我看来就像是一片土地或者坚硬的白垩，抵抗着渗透……海水无声无息地缓缓推进，似乎没有什么发生，没有任何东西移动，水太远了，你几乎听不见它的声音……但最终，它包围了那个抵抗的物质。

小平邦彦讲的故事则更加引人入胜：

现在数学的研究对象一般都非常抽象，实例也十分抽象，让人难以理解。所以依靠具体事实归纳来猜想定理的方式，在大多数情况下已经难以适用。目前的情况下，关于发现新定理的思考实验方式，我本人也是不得而知。如果将精力都花费在思索新的思考方式上，恐怕难有所得。实际上很多时候无论如何思考都得不到相应的结果。这样看的话，是否可以说数学研究是一份极其困难的工作呢？不过这倒也未必。有时候感觉自己什么也没做，那些应当思考的事情却很自然地呈现在眼前，研究工作也得以顺利推进。
夏目漱石在《梦十夜》中对运庆（注：日本镰仓时代的高僧，雕刻技艺十分精湛）雕刻金剛手菩萨像的描述，充分表现了这种感受。这部分内容引用如下：
运庆在金剛手菩萨的粗眉上端一寸处横向凿刻，手中的凿刀忽而竖立，转而上而下凿去。凿刀被敲入坚硬的木头中，厚厚的木屑应声飞落，再仔细一看，金剛手菩萨怒意盈盈的鼻翼轮廓已清晰呈现。运庆的运刀方式无拘无束，雕琢过程中丝毫没有任何迟疑。
“他的手法真如行云流水，凿刀所到之处，居然都自然地雕琢出了内心所想的眉毛、鼻子样子。”我感慨至极，不禁自言自语道。
结果，方才那位年轻男子回应道：
“什么呀，那可不是凿刻出的眉毛、鼻子，而是眉毛、鼻子本来就埋藏在木头中，他只是用锤子和凿子将其呈现出来。就像从泥土中挖出石头一样，当然不会出现偏差。”
在这种时刻，我常常感到世间没有比数学更容易的学科了。如果遇到一些学生在犹豫将来是否从事数学方面的工作，我就会想建议他们“一定要选数学，因为再没有比数学更容易的学科了”。

这些故事自然有趣，但是如果没有亲身体验，恐怕云里雾里，不知所云。我想，铺路式科研最重要的一点就是它没有预设的目标，不会为了某个目标而强行加入假设；它更在意研究对象的真实性，以平常心忠实地记录。现代的纯数研究，大多都是遵从这一思想向前推进的。

Theory-like engineering does not have a pre-set goal. It won't add in unrealistic assumptions just to reach certain goal. Rather, it cares more about reality of the objects at study, documenting everything with a normal heart. Most of pure math research are along this way.

Physics-like ML theory

A physics-like theory, 铺路式理论

Note:

When I say “physics-like”,

I don't necessarily mean technical tools in physics research or physical phenomenon, but rather a mindset that physicists adopt to **approach** our **physical reality**.

Physicists' mindset:

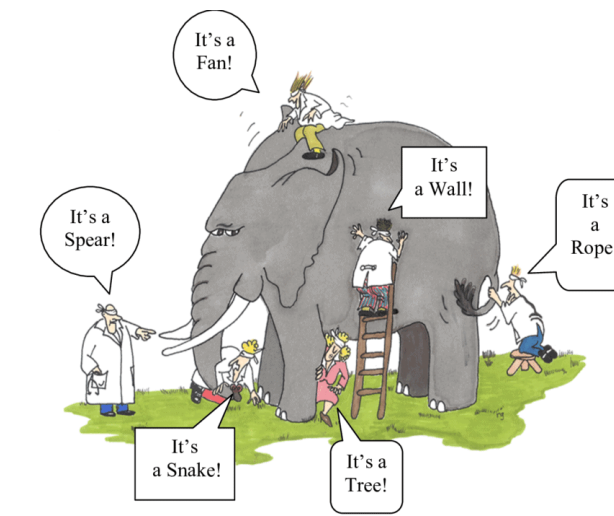
- (0) put an emphasis on reality (build theories driven by experiments/observations)
- (1) identify useful/relevant degrees of freedom (while ignoring other details)
- (2) view the world dynamically
- (3) appreciate mental pictures more than mathematical rigour

Questions for physics-like ML theory

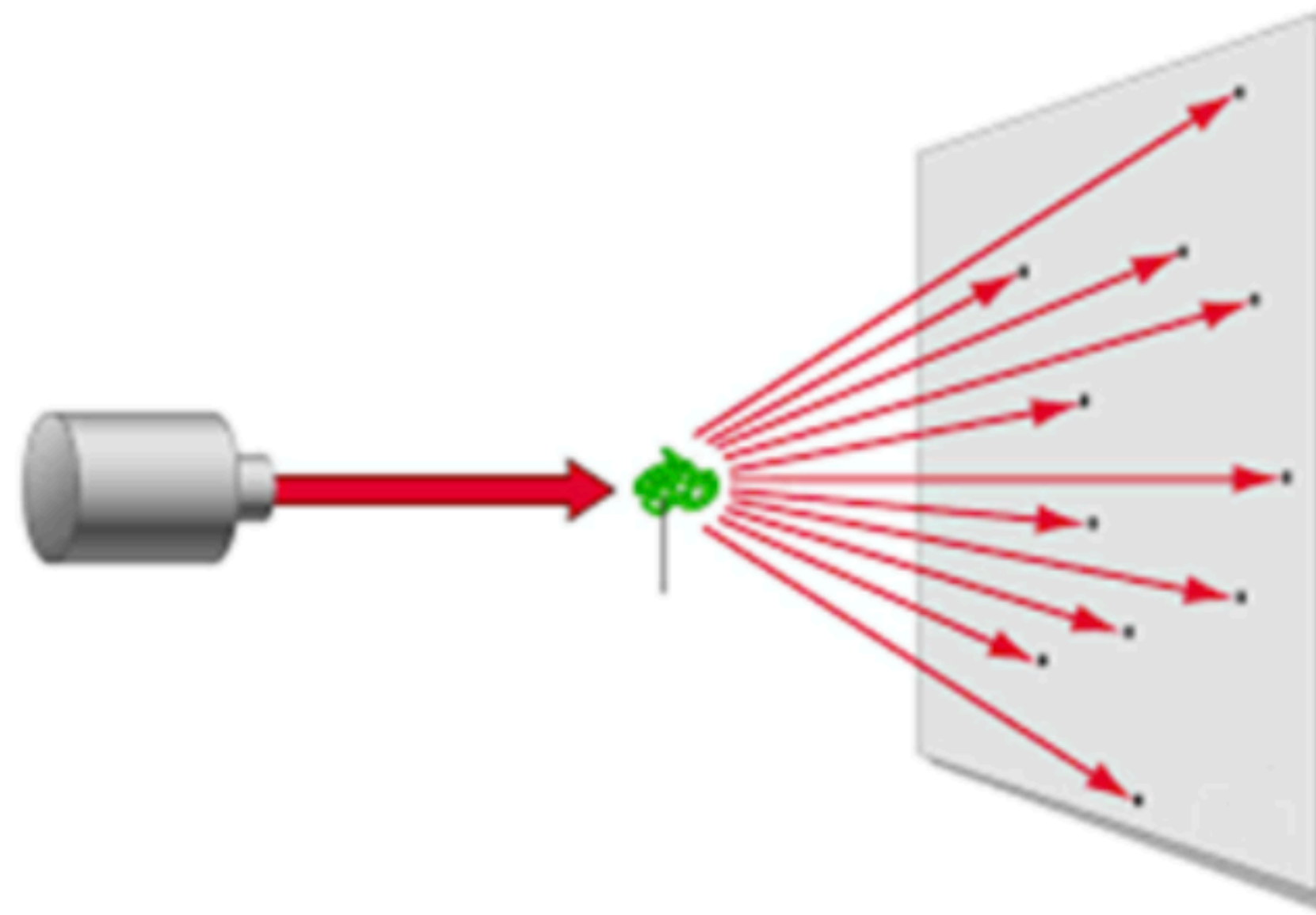
Q1: What is **reality** in ML?

Q2: How do we **approach** the reality?

Q1: What is reality in ML?



- * Architectures
- * Optimisation
- * Regularisation
- * Task & Data



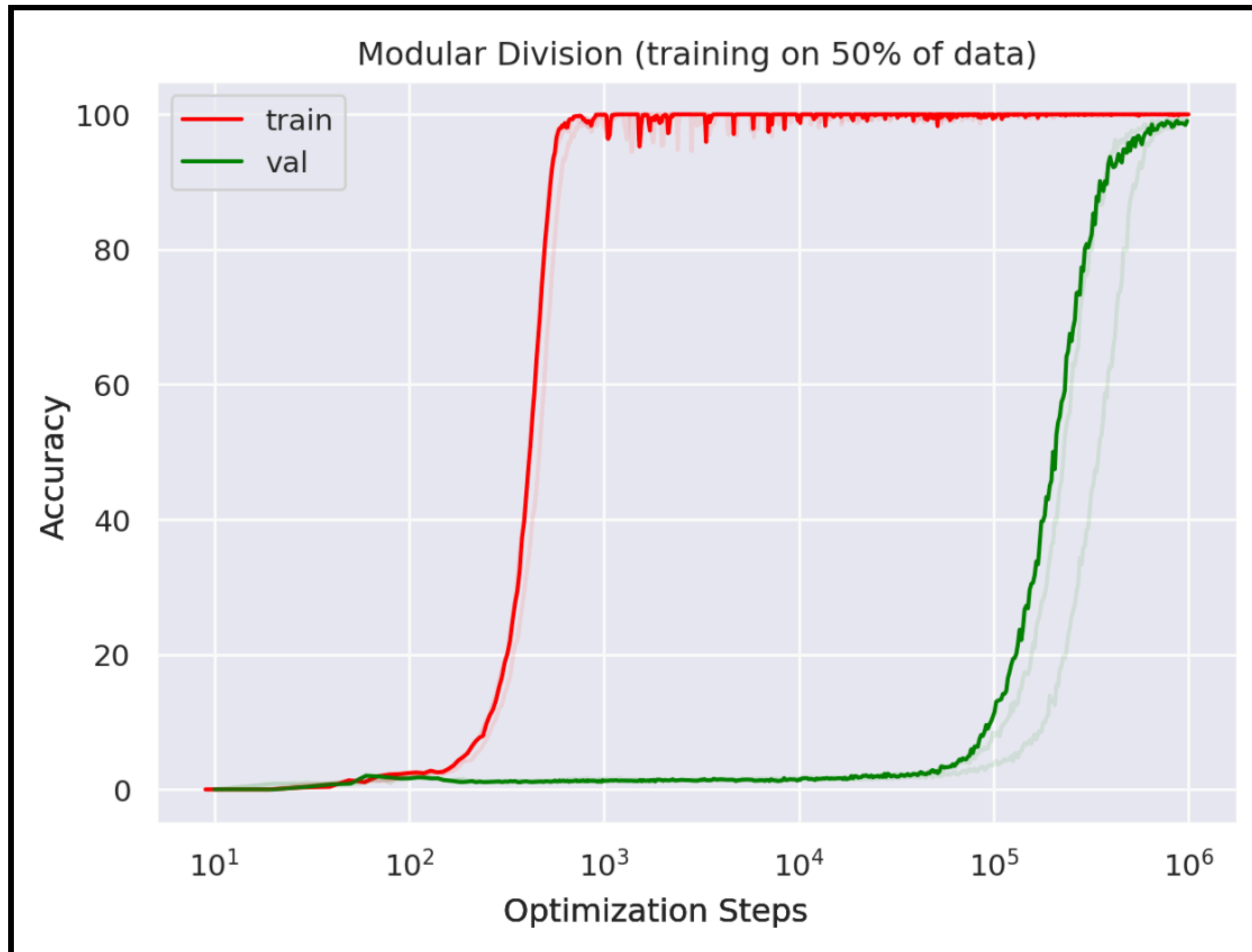
- * Double descent
- * Grokking
- * Neural Scaling Laws
- * Emergent abilities
- * Edge of Stability
- * Optimizer inductive biases
- * Neural collapse
- * Information bottleneck
- * Effective energy descent
- * Modularity
- * Loss spike
- * Condensation
- * Linear separability

Q2: How do we approach reality?

Physicists:

- (0) put an emphasis on reality (build theories driven by experiments/observations)
- (1) identify useful/relevant degrees of freedom (while ignoring other details)
- (2) view the world dynamically
- (3) appreciate mental pictures more than mathematical rigour

Example: Grokking



Example: Grokking

$$\boxed{a} \circ \boxed{b} = \boxed{c}$$

★	a	b	c	d	e
a	a	d	?	c	d
b	c	d	d	a	c
c	?	e	d	b	d
d	a	?	?	b	c
e	b	b	c	?	a

From **Figure 1** of "Grokking: Generalization beyond overfitting on small algorithmic datasets." by *Power et al.*

Example: Grokking

Split the table into
train & **val** datasets

★	a	b	c	d	e
a	a	d	?	c	d
b	c	d	d	a	c
c	?	e	d	b	d
d	a	?	?	b	c
e	b	b	c	?	a

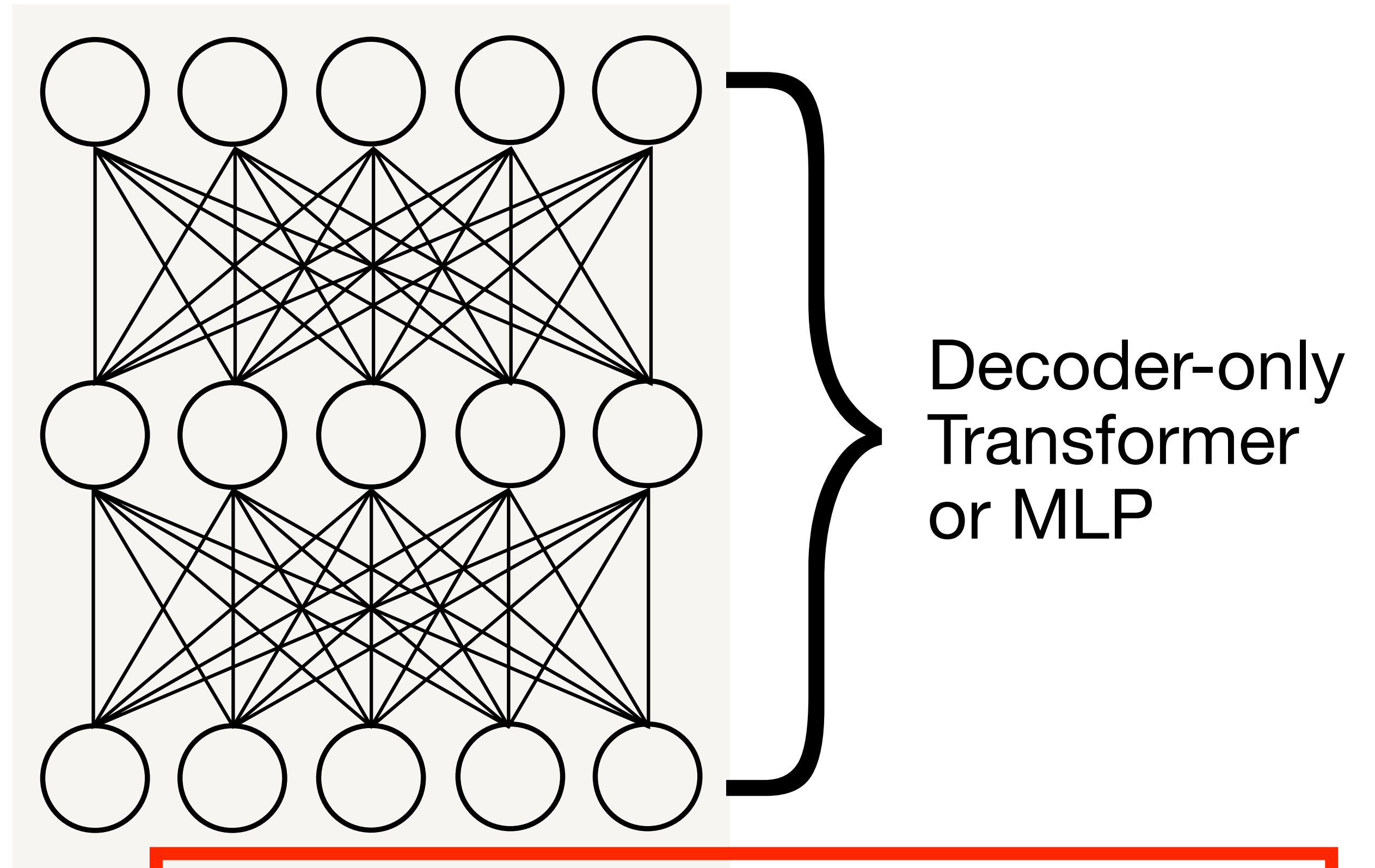
From **Figure 1** of "Grokking: Generalization beyond overfitting on small algorithmic datasets." by *Power et al.*

Example: Grokking

Task: learn a binary operation

$$a + b \pmod{p} = c$$

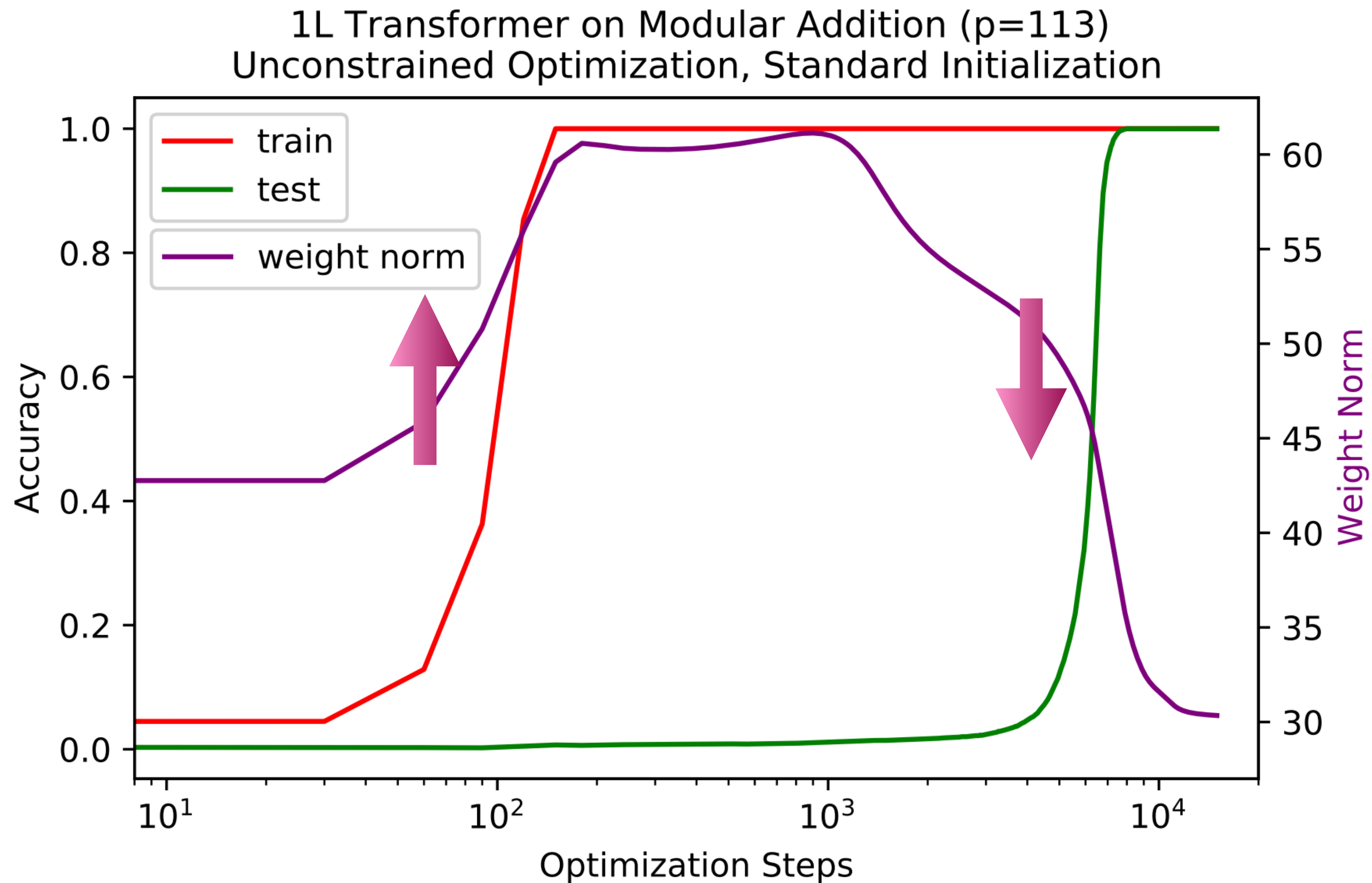
Logits for a, b, c, ...



a **b**

← Trainable Embeddings

Strategy 1: identify relevant variables



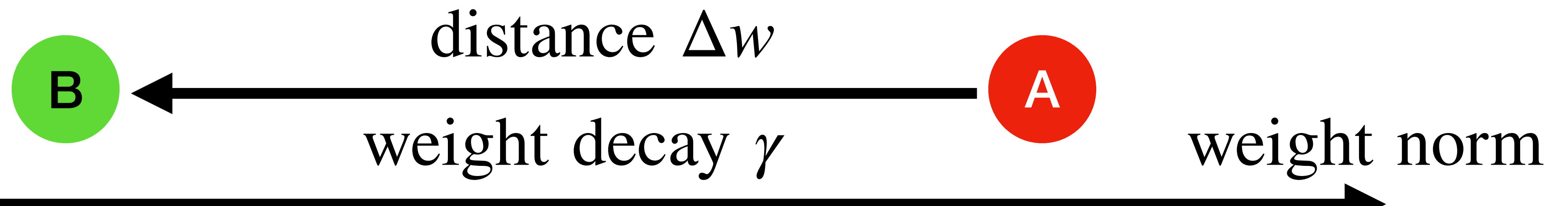
Strategy 2: view the world dynamically



The time to travel from city A to city B is $t = \frac{d}{v} \propto v^{-1}$

Model B: generalisation circuit,
small weight norm

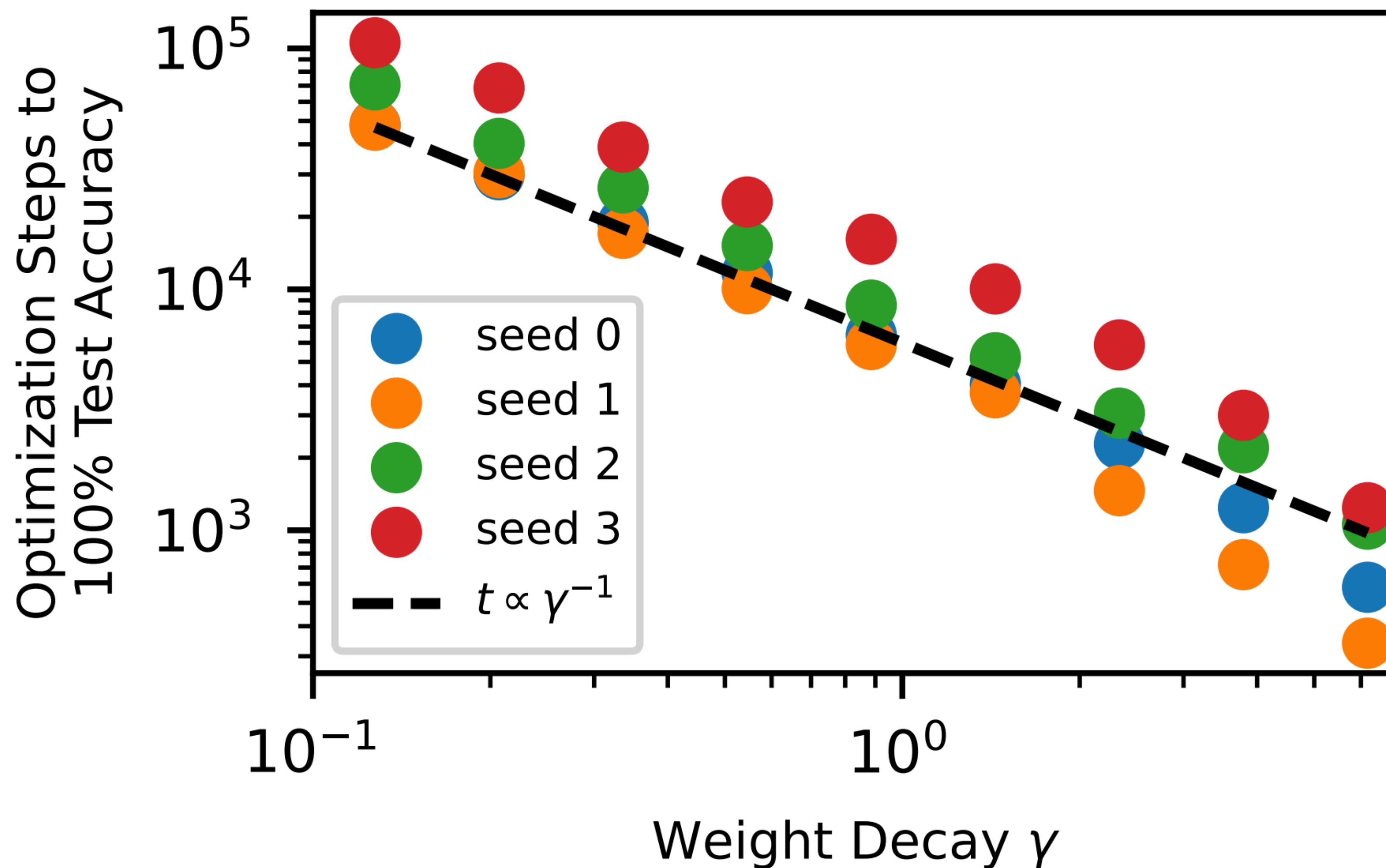
Model A: memorization circuit,
large weight norm



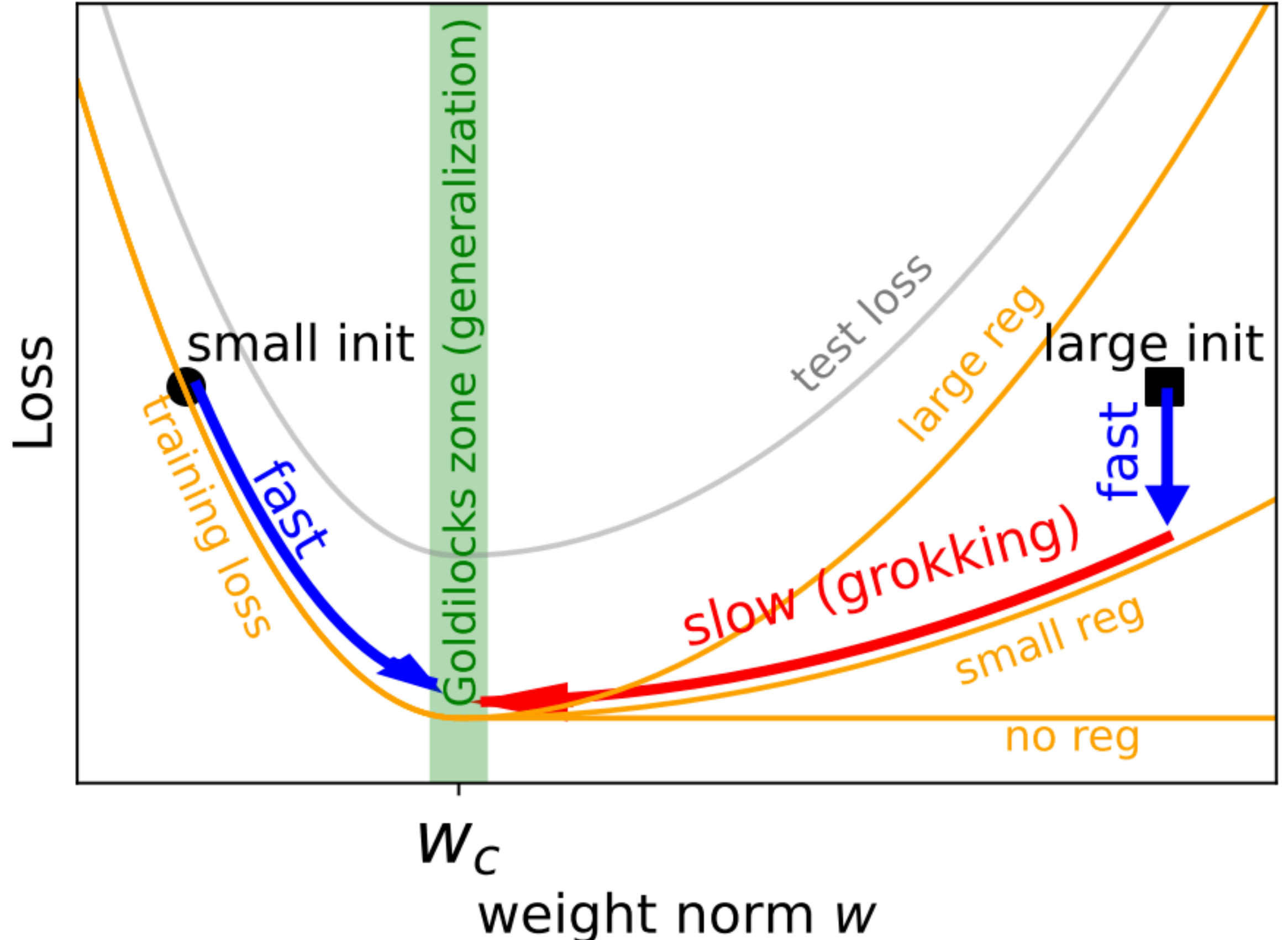
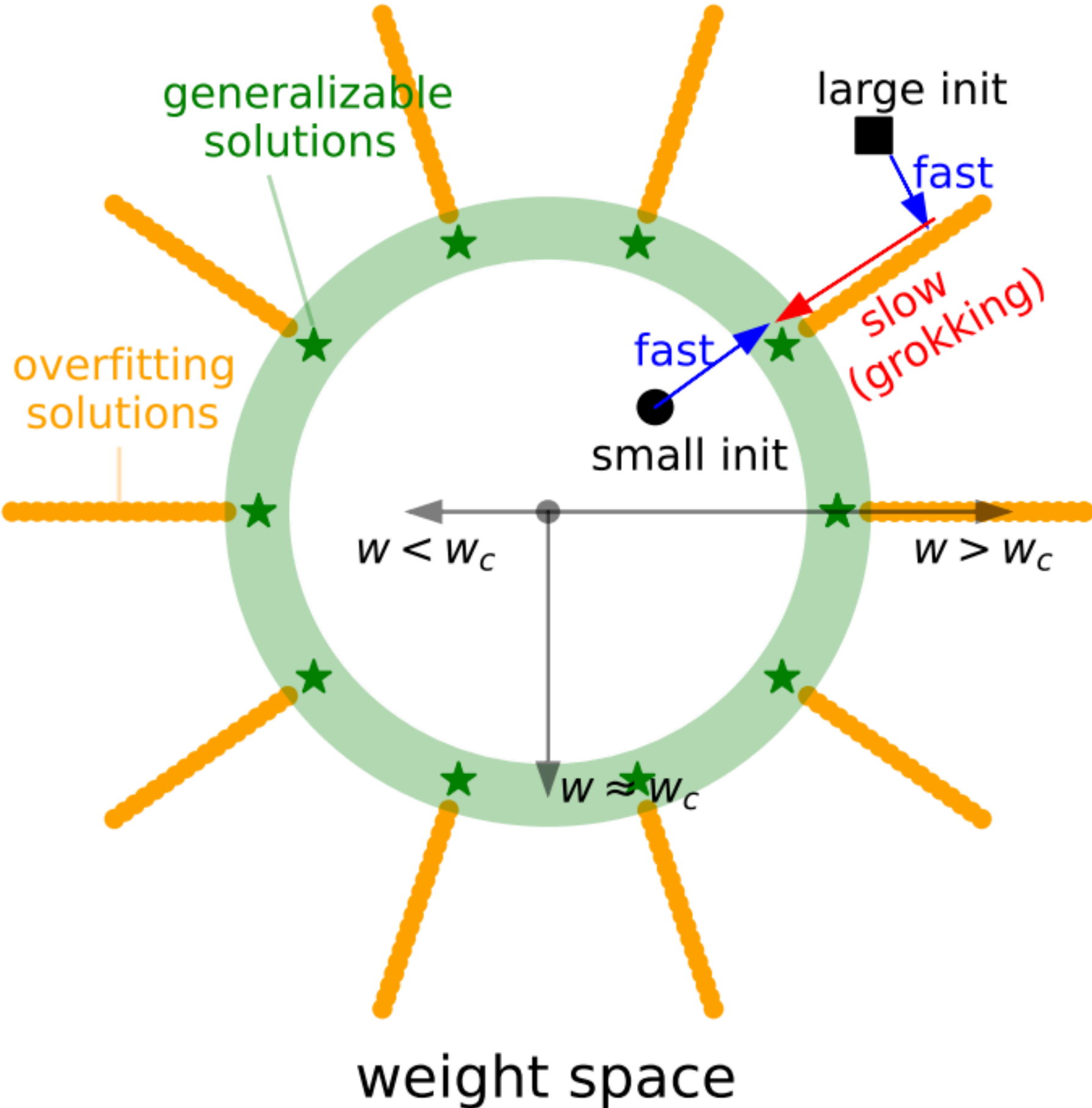
The time to travel from model A to model B is $t = \frac{\Delta \log w}{\gamma} \propto \gamma^{-1}$

Strategy 2: view the world dynamically

1L Transformer on Mod Addition | $\alpha = 1.0$



Strategy 3: Forming mental pictures



A physics-like theory, 铺路式理论

Note:

When I say “physics-like”,

I don't necessarily mean **technical tools in physics research** or **physical phenomenon**, but they are also useful!

Technical tools and/or physical concepts

- (1) Phase transition. NN behaviour depending on control parameters.
- (2) Renormalization: how to do coarse graining. NN macroscopic behaviour emergent from microscopic variables.
- (3) NN training as dynamical systems. Fast-slow dynamics, adiabatic approximation.
- (4) NN as a bulk of matter/complex systems: response function, hysteresis etc.
- (5) Modularity. Decompose a large system into a few weakly-coupled systems.
- (6) Mean field theory and quasi-particles.

Summary: Core questions for AI theory researchers

Q1: What is **reality** in AI?

Q2: How do we **approach** the reality?

(Q3: And build something useful?)

