

AI 算法 & 算力 & 体系结构进步



ZOMI

Talk Overview

1. AI 系统概述

- AI 历史，现状与发展
- 算法与体系结构的进步
- AI 系统的组成与生态
- 大模型对AI系统的挑战

2. AI 芯片

3. AI 编译器

4. AI 推理引擎

5. AI 开发框架

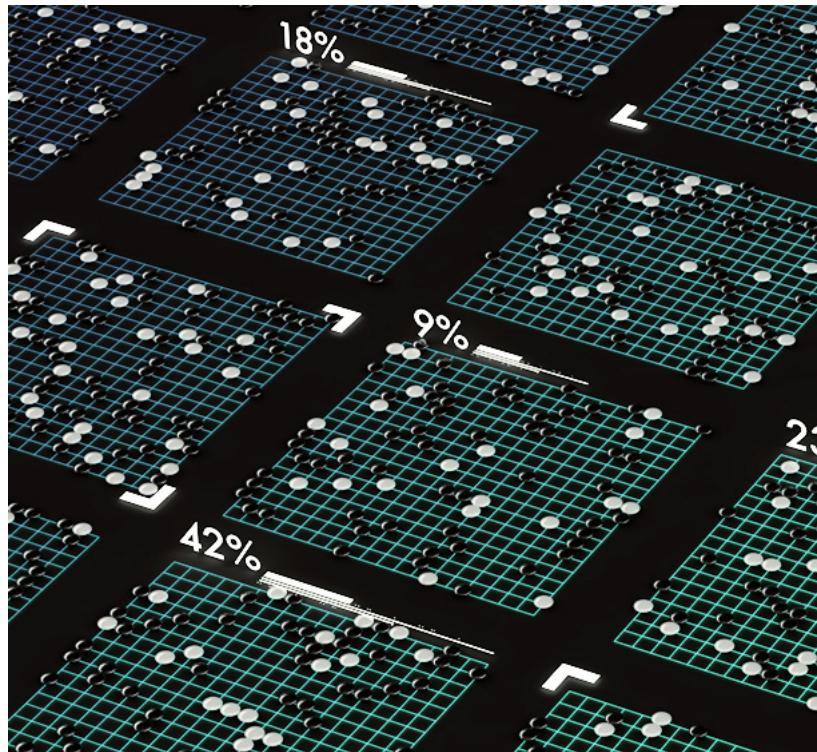
6. 异构集群调度与管理

7. AI 大模型

Talk Overview

- **AI 发展、算法、体系结构**
 1. AI algorithm, model status - AI 算法现状
 2. AI algorithm, model trends - AI 算法趋势
 3. Algorithms and Architecture - 算法与体系结构进步

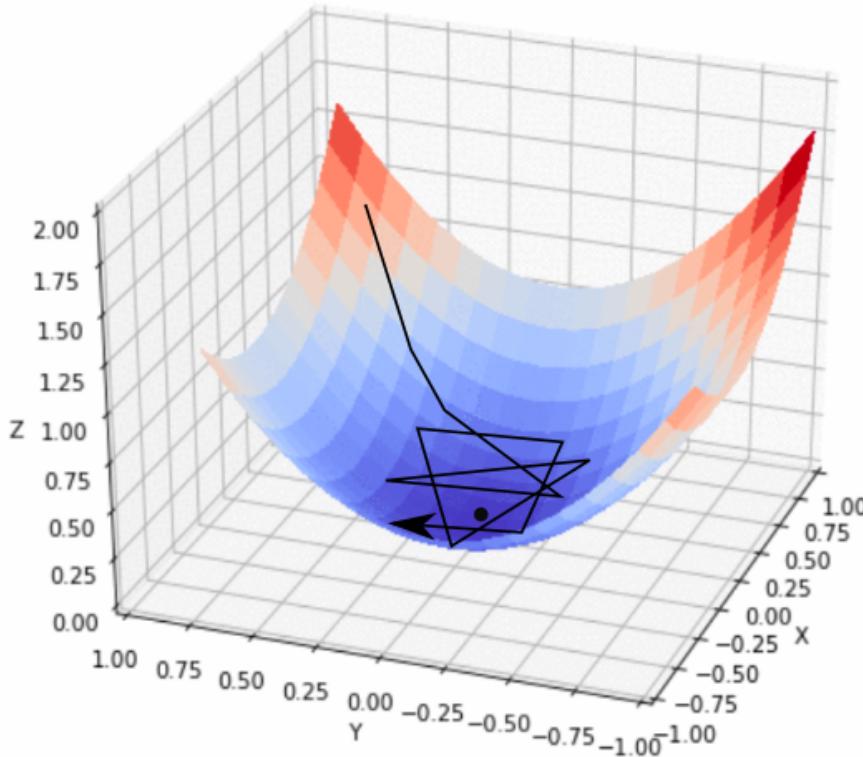
AlphaGO 在计算



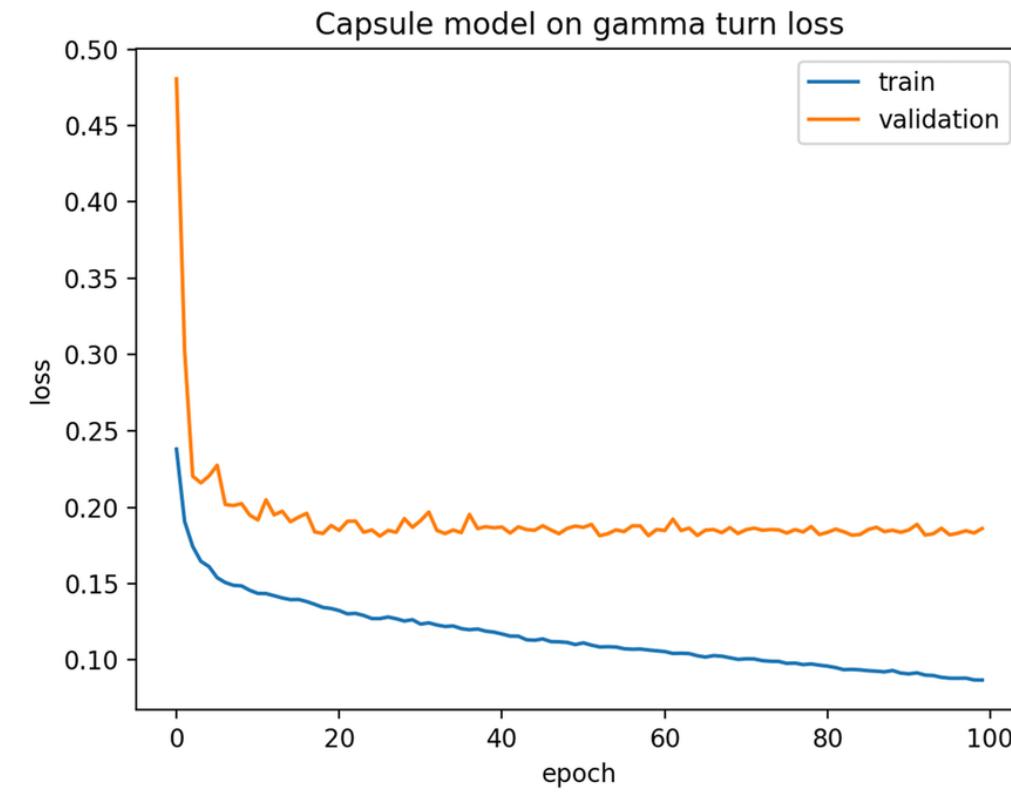
人机围棋比赛现场



梯度下降算法寻找数据鞍点



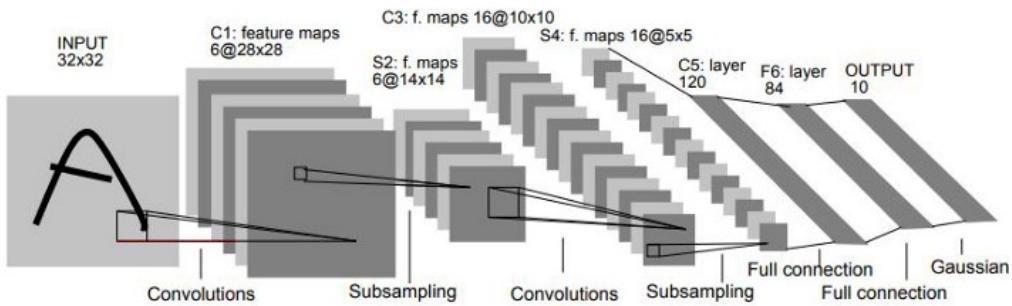
损失值在下降



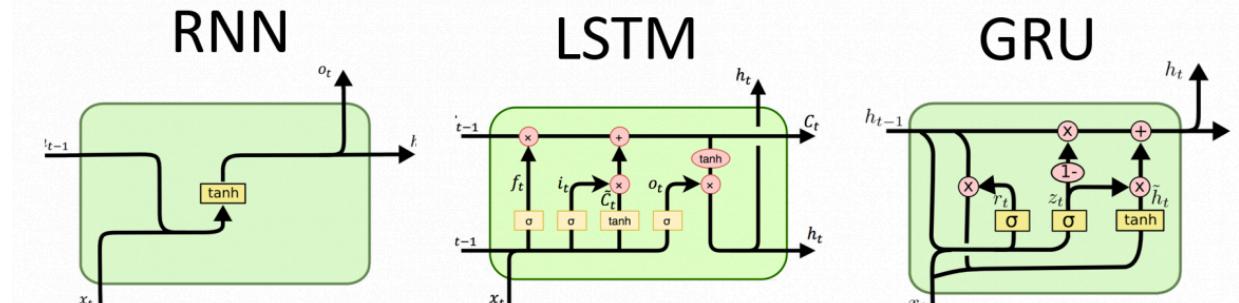
1. AI 算法模型现状

基本模型结构类型

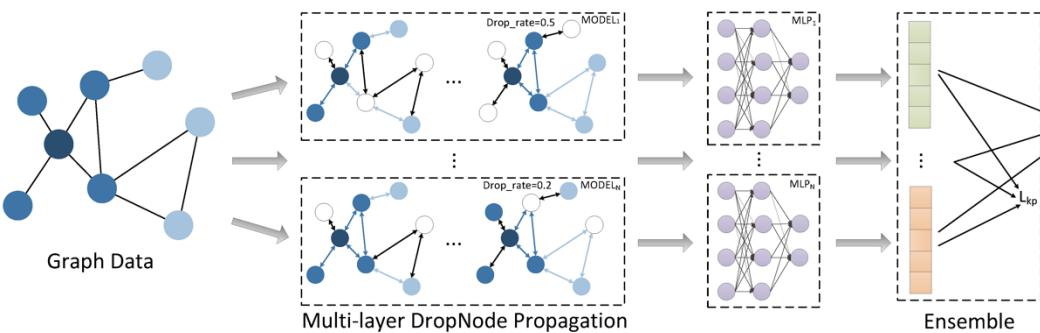
CNN



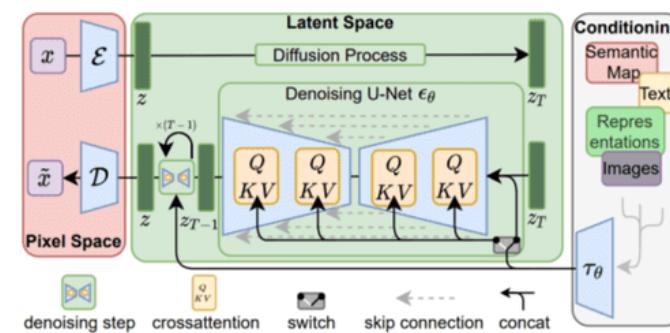
RNN



GNN



Diffusion



```
Algorithm 1 Training
1: repeat
2:    $x_0 \sim q(x_0)$ 
3:    $t \sim \text{Uniform}(\{1, \dots, T\})$ 
4:    $\epsilon \sim \mathcal{N}(0, 1)$ 
5:   Take gradient descent step on

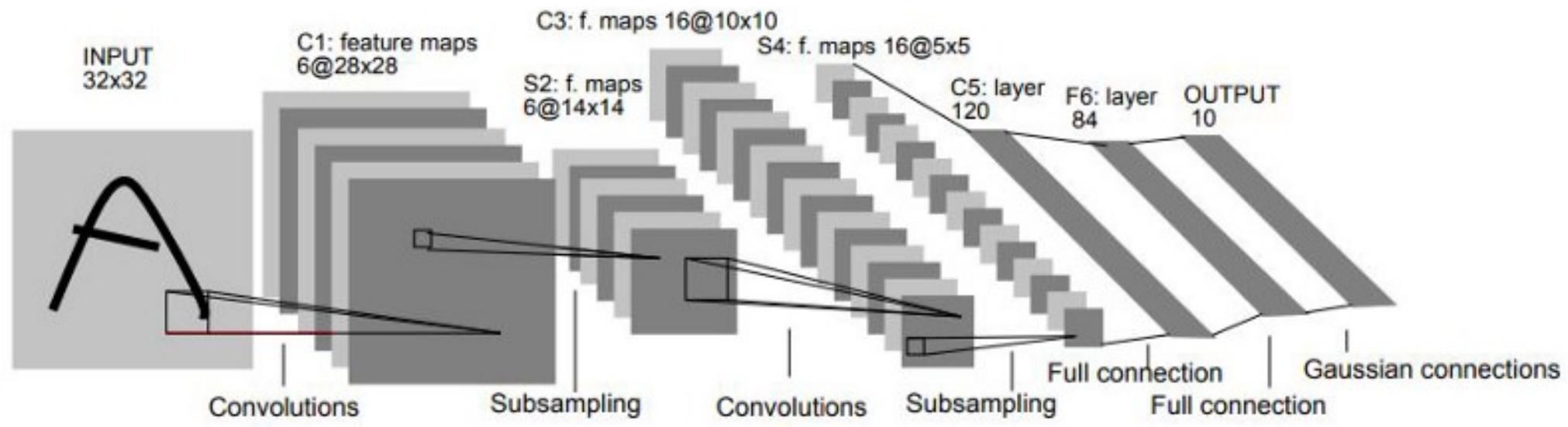
$$\nabla_\theta \|\epsilon - \epsilon_\theta(\sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon, t)\|^2$$

6: until converged
```

```
Algorithm 2 Sampling
1:  $x_T \sim \mathcal{N}(0, I)$ 
2: for  $t = T, \dots, 1$  do
3:    $z \sim \mathcal{N}(0, I)$  if  $t > 1$ , else  $z = 0$ 
4:    $x_{t-1} = \frac{1}{\sqrt{\alpha_t}} (x_t - \frac{1-\alpha_t}{\sqrt{1-\alpha_t}} \epsilon_\theta(x_t, t)) + \sigma_t z$ 
5: end for
6: return  $x_0$ 
```

基本模型结构类型

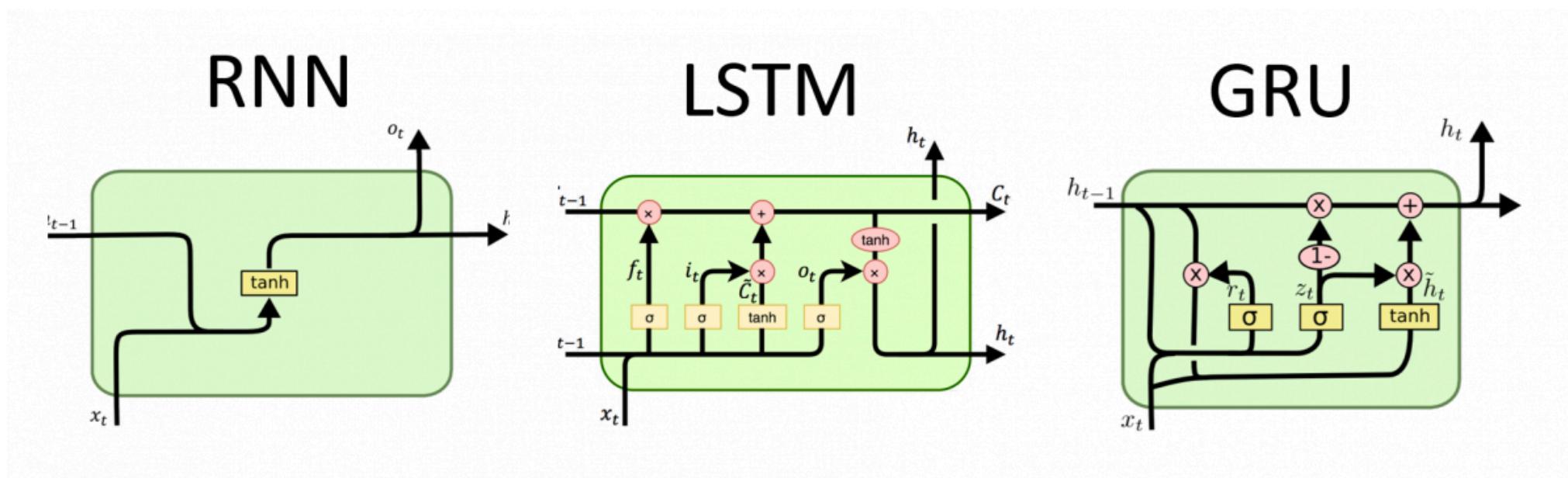
- 卷积神经网络（ Convolutional Neural Network ）
 - 以卷积层（ Convolution Layer ）, 池化层（ Pooling Layer ）, 全连接层（ Fully Connected Layer ）等算子（ Operator ）组合形成，并在 CV 领域取得明显效果和广泛应用的模型结构。



基本模型结构类型

- 循环神经网络 (Recurrent Neural Network)

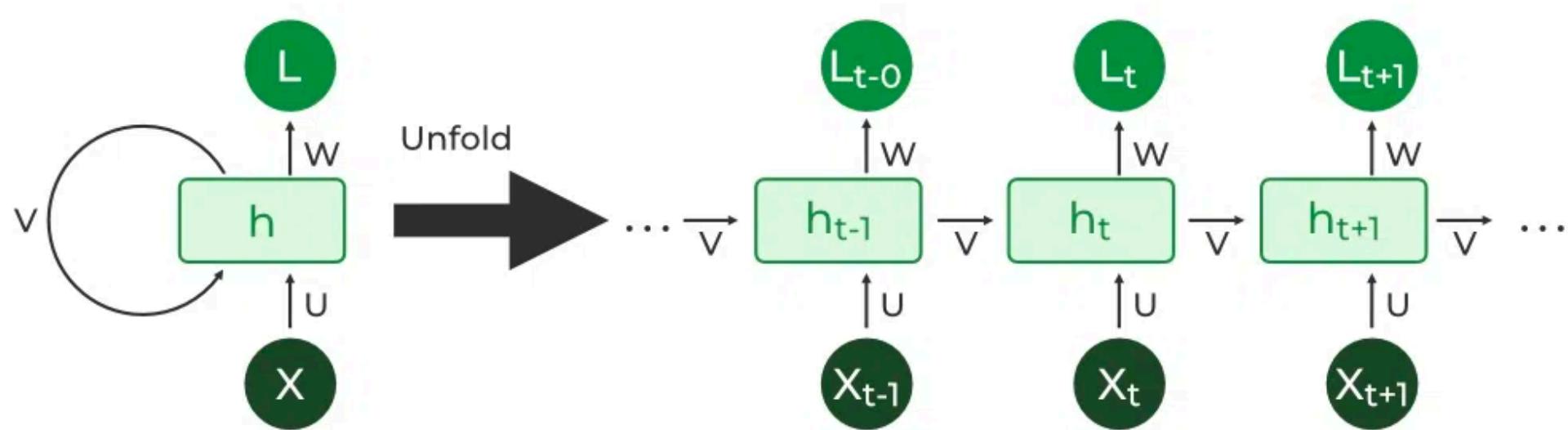
- 以循环神经网络，长短时记忆 (LSTM) 等基本单元组合形成的适合时序数据预测（例如，自然语言处理，语音识别，监控时序数据等）模型结构。



基本模型结构类型

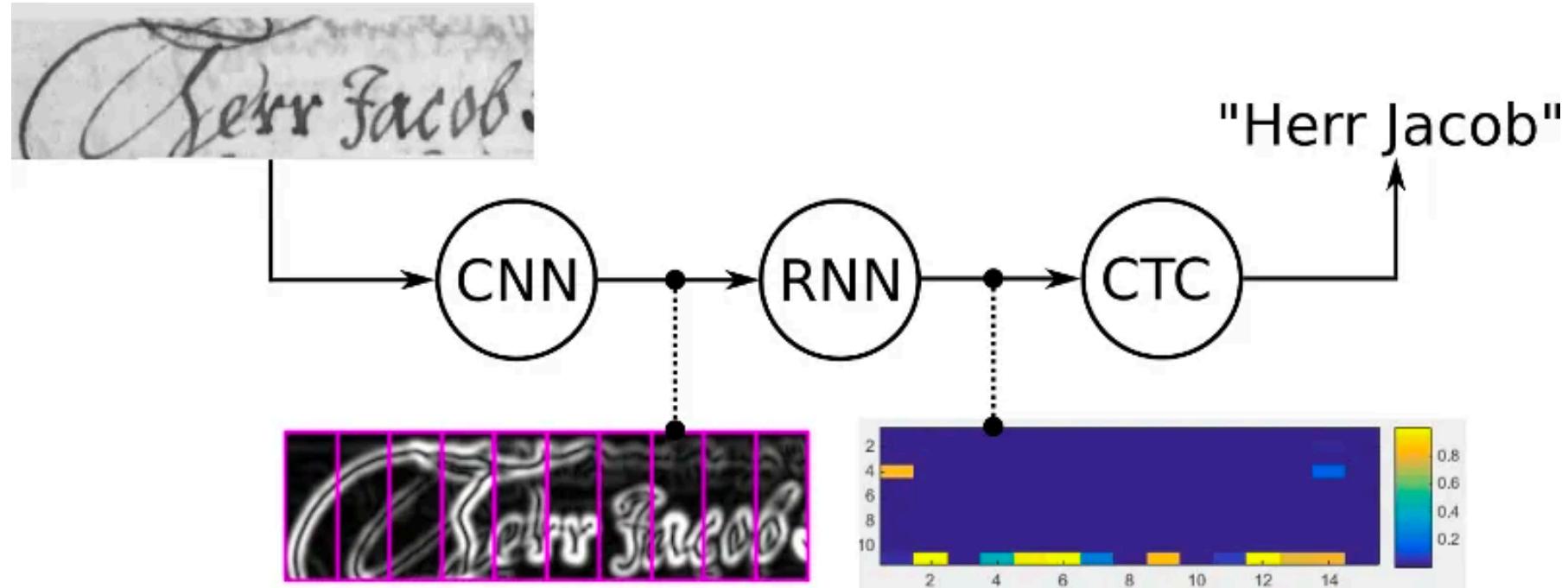
- 循环神经网络 (Recurrent Neural Network)

- 以循环神经网络，长短时记忆（LSTM）等基本单元组合形成的适合时序数据预测（例如，自然语言处理，语音识别，监控时序数据等）模型结构。



基本模型结构类型

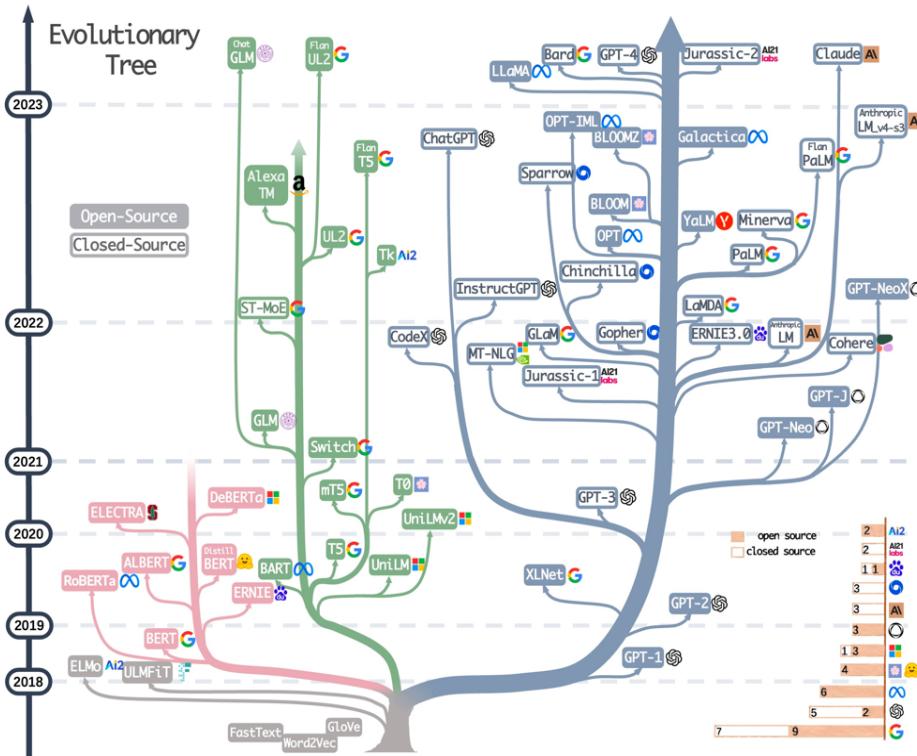
- 混合结构
 - 组合卷积神经网络和循环神经网络，进而解决如光学字符识别（OCR）等复杂应用场景的预测任务。



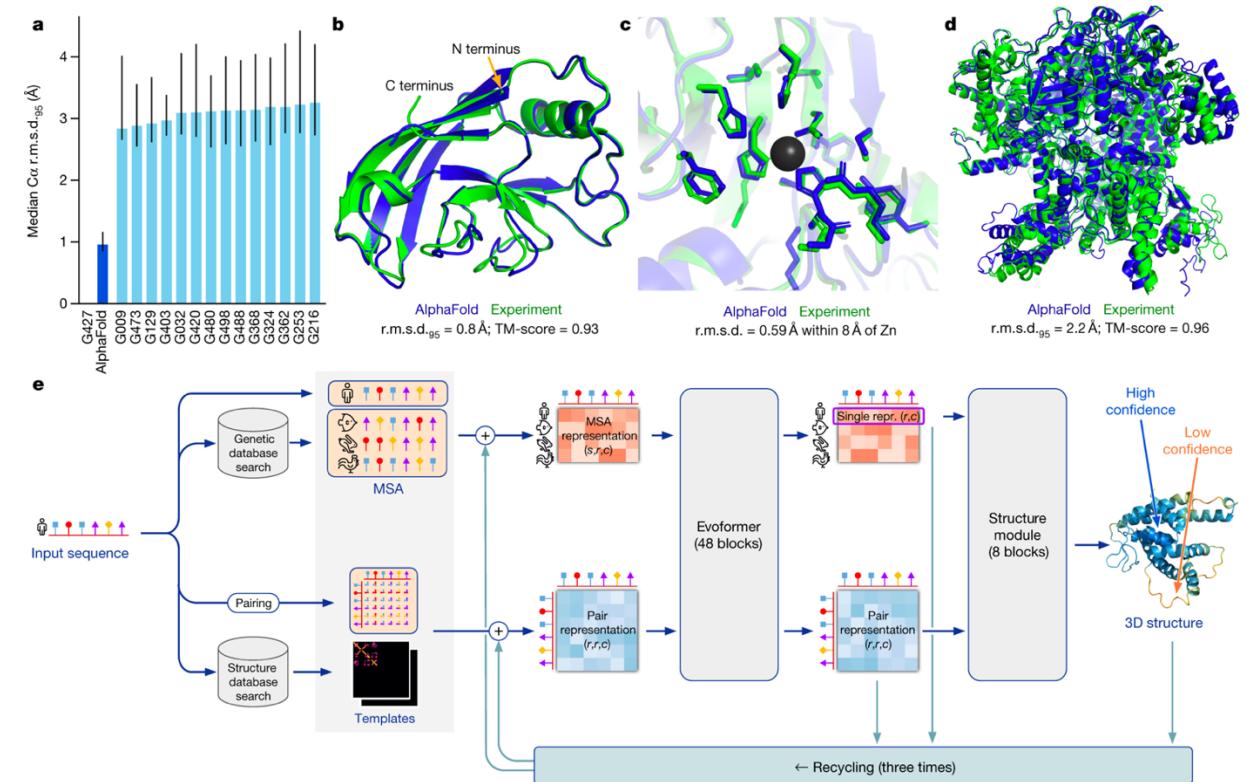
2. AI 算法趋势

发展趋势

大模型 / 稀疏结构



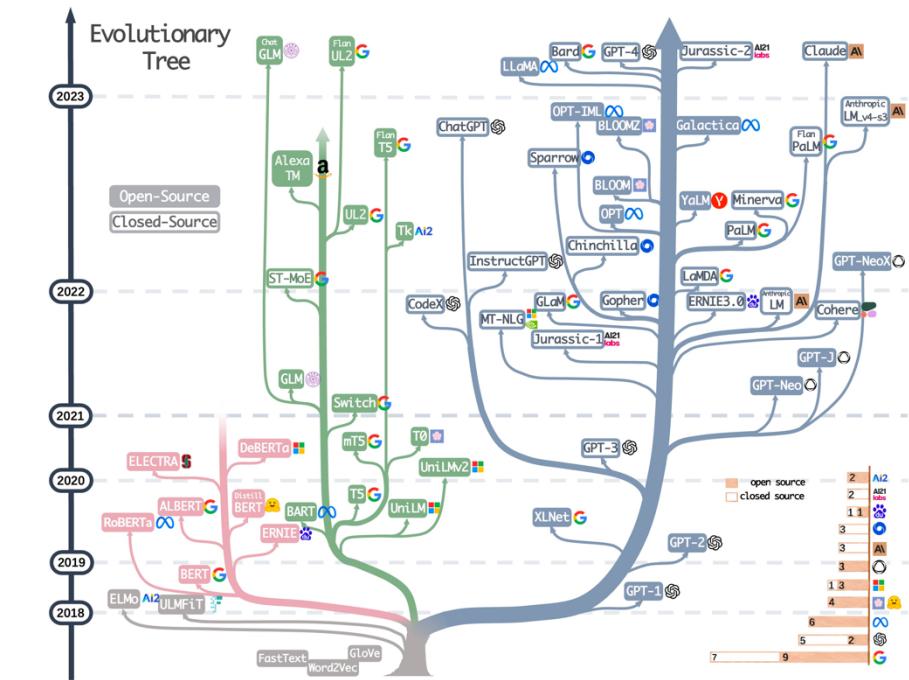
搜索空间 / 训练方式 / AI4SCI



更大的模型：大模型

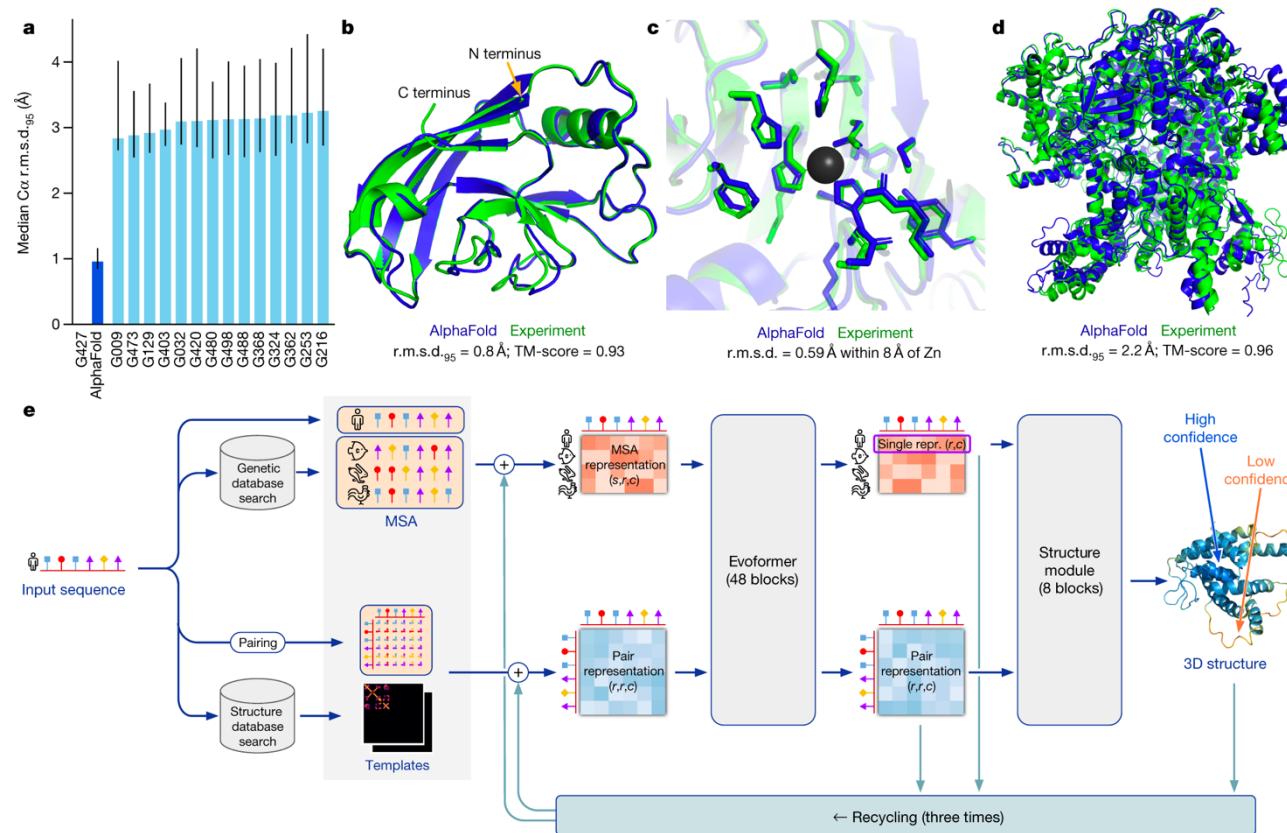
- 大模型是指具有大量参数和复杂结构的机器学习模型。这些模型可以应用于处理大规模的数据和复杂的问题。

Training compute (FLOPs) of milestone Machine Learning systems over time



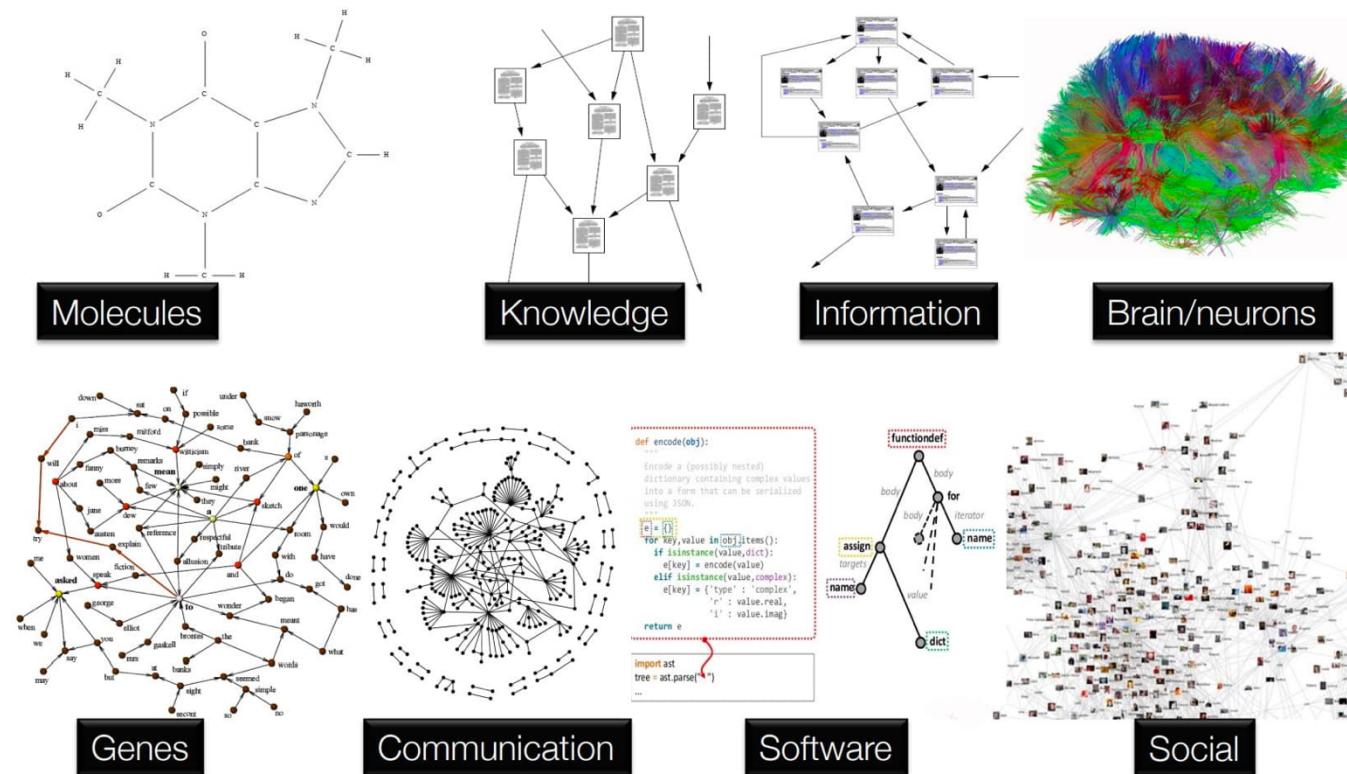
灵活结构和建模能力：AI + 科学计算

- AI for Science指的是运用AI算法从海量数据中快速找出潜在规律，科学家可从AI预测的规律中甄别出正确的结果。



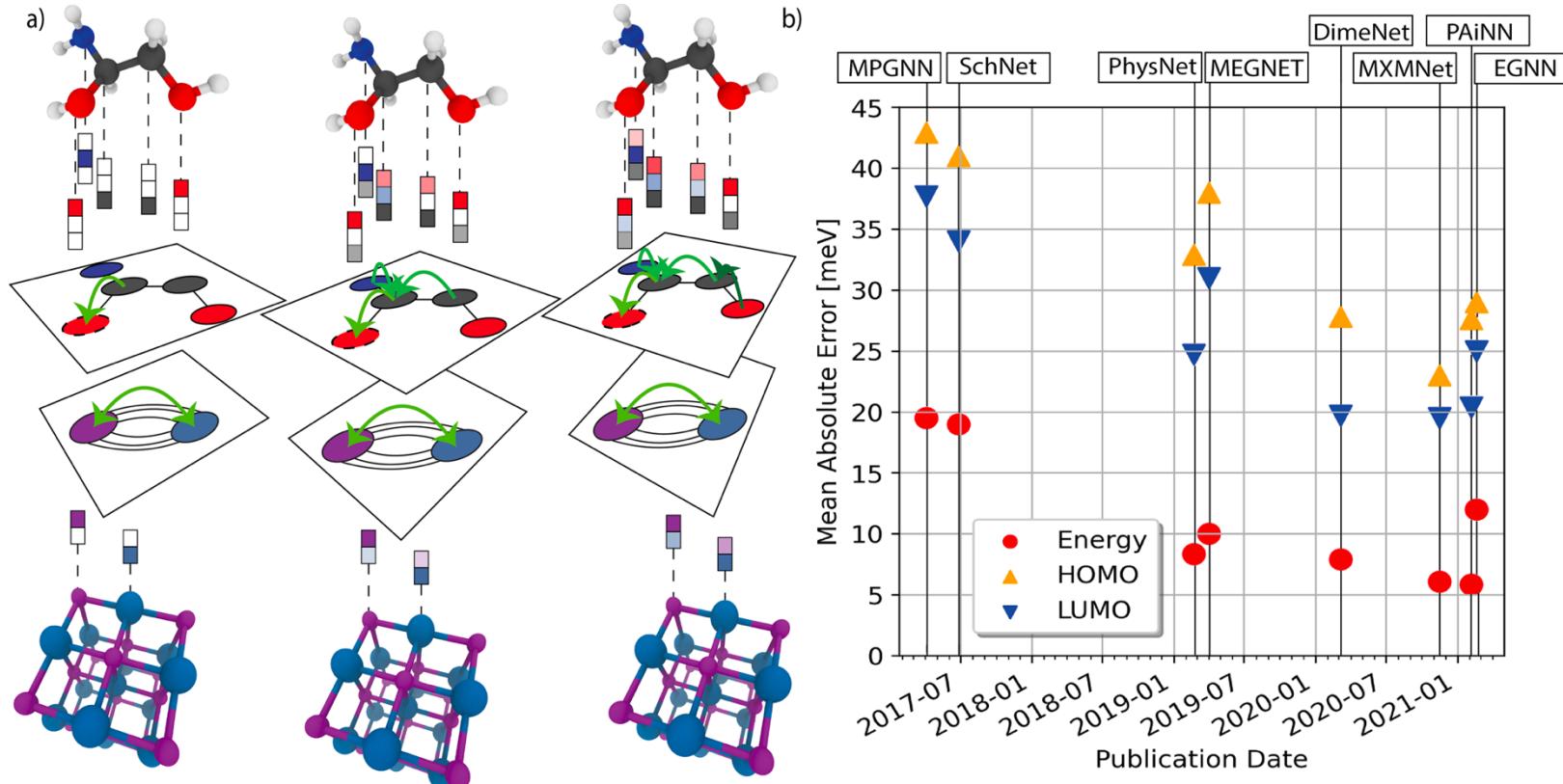
灵活结构和建模能力：图神经网络

- 图神经网络等网络不断抽象多样且灵活的数据结构，应对更为复杂的建模需求。进而衍生了新的算子（例如：图卷积等）与计算框架（例如：图神经网络框架等）。



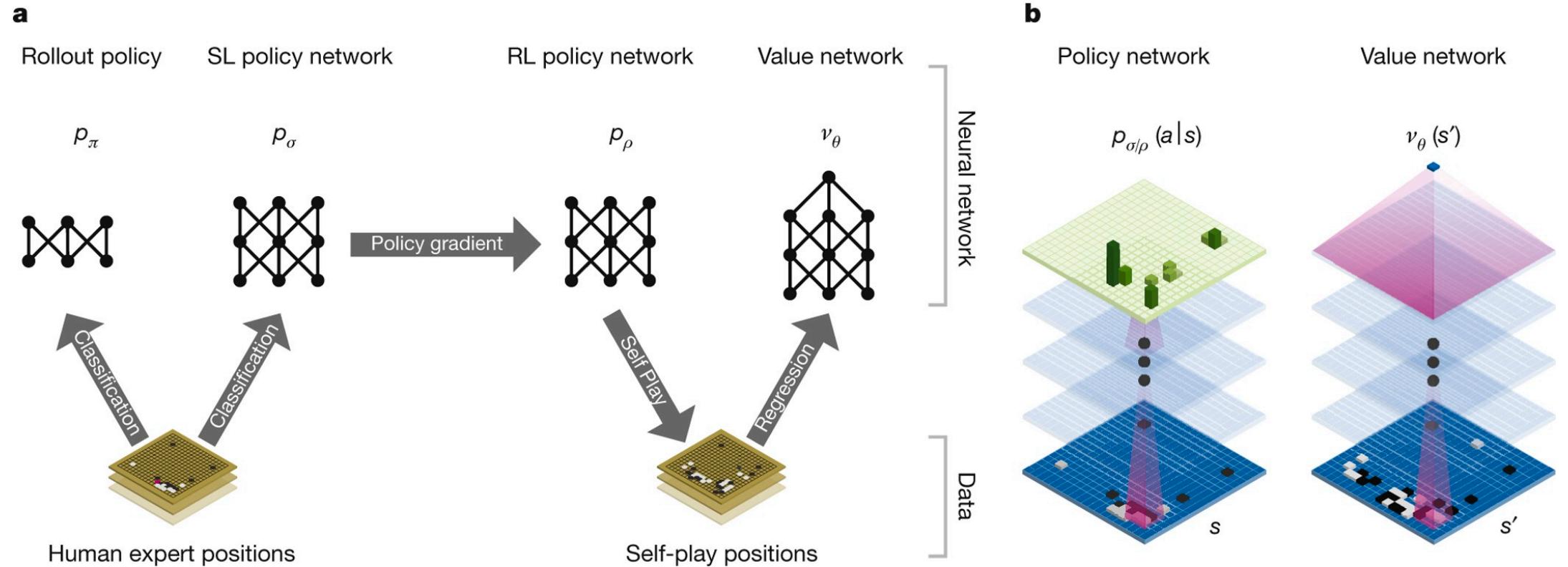
灵活结构和建模能力：图神经网络

- 图神经网络等网络不断抽象多样且灵活的数据结构，应对更为复杂的建模需求。进而衍生了新的算子（例如：图卷积等）与计算框架（例如：图神经网络框架等）。



更多样的训练方式：深度强化学习

- 深度强化学习将深度学习的感知能力和强化学习的决策能力相结合，可以直根据输入的图像进行控制，是一种更接近人类思维方式的人工智能方法。

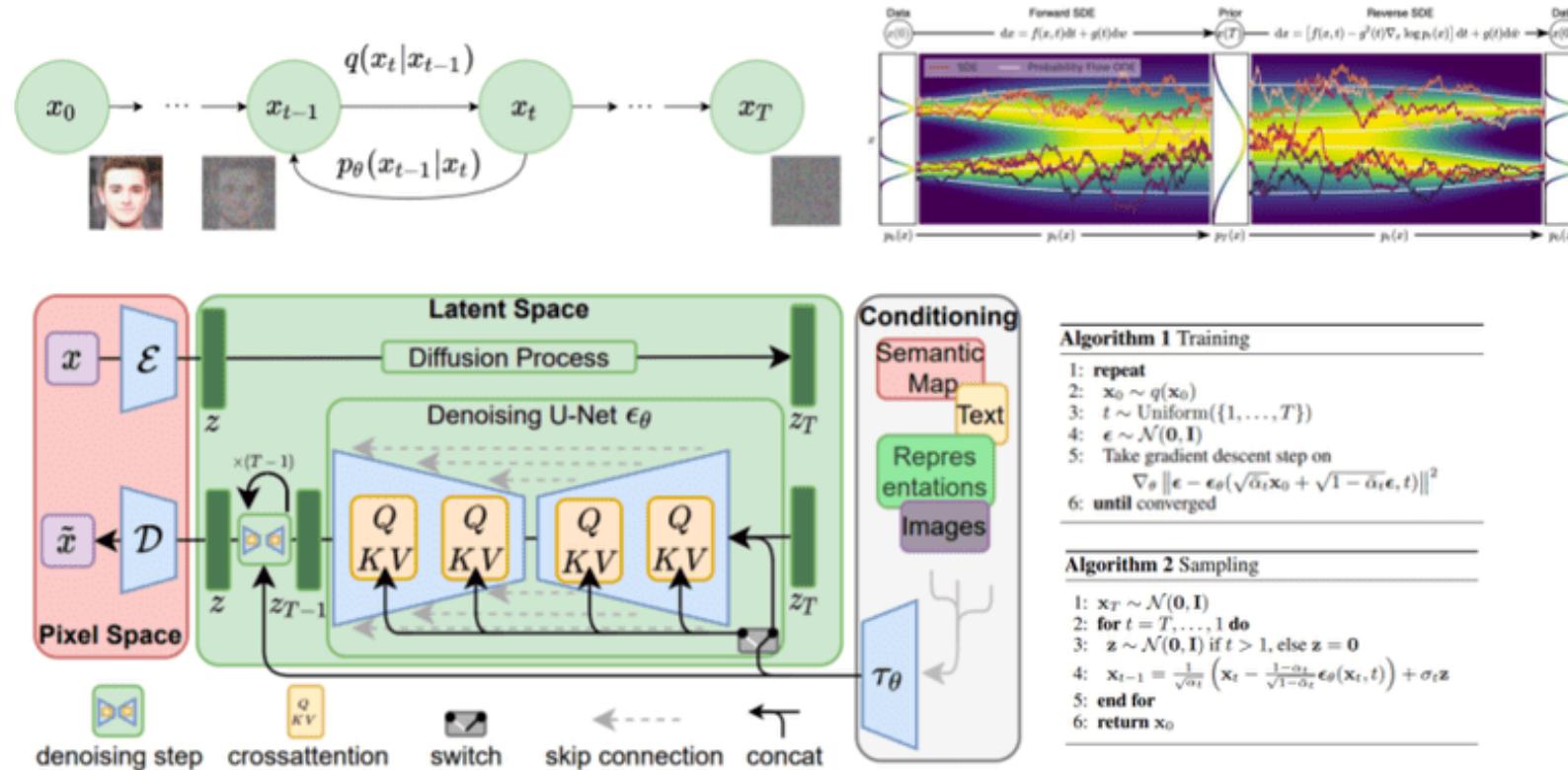


更多样的训练方式：生成式AI



更多样的训练方式：生成式AI

- 生成式AI是指使用AI来创作新内容，例如文本、图片、音乐、音频和视频。生成式AI基于可以执行多任务处理和执行开箱即用任务。



Still curious about AI ?

- 卷 AI 算法好有意思呢？
- 卷 AI 系统保饭碗呢？



Question?

1. 满足 AI 爆发三大因素是不是还有其他算法？
2. 为啥子其他算法 🔥 不起来？



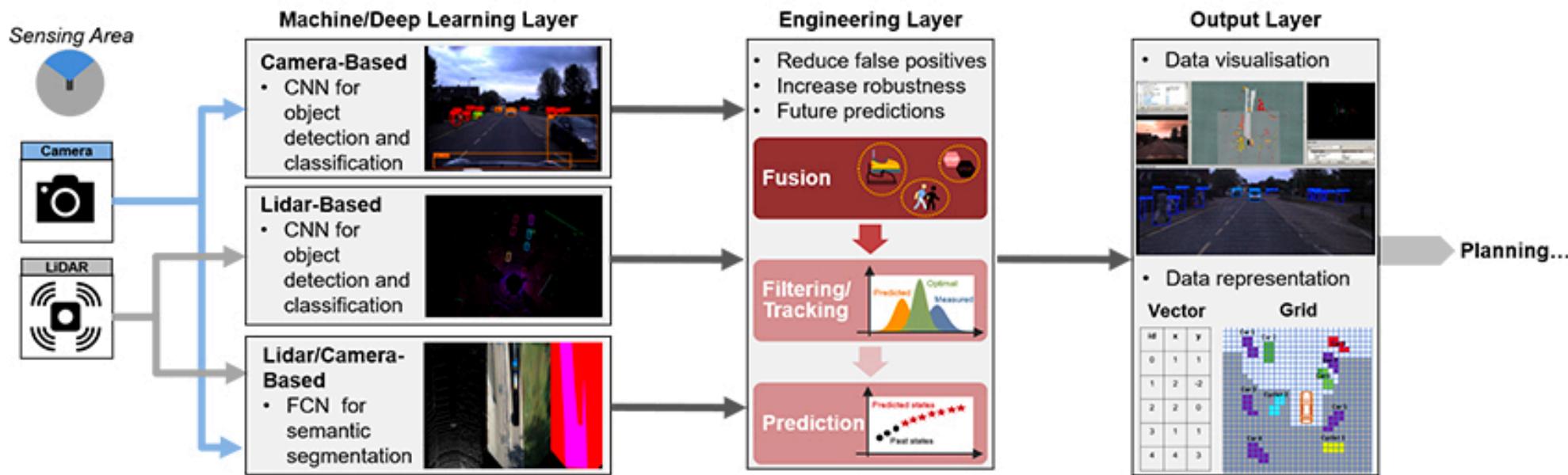
**催生这轮人工智能热潮的原因有三个重要因素：
大数据的积累、机器学习尤其是深度学习算法、
超大规模的计算能力支撑都取得了突破性进展。**

3. 大数据涌现



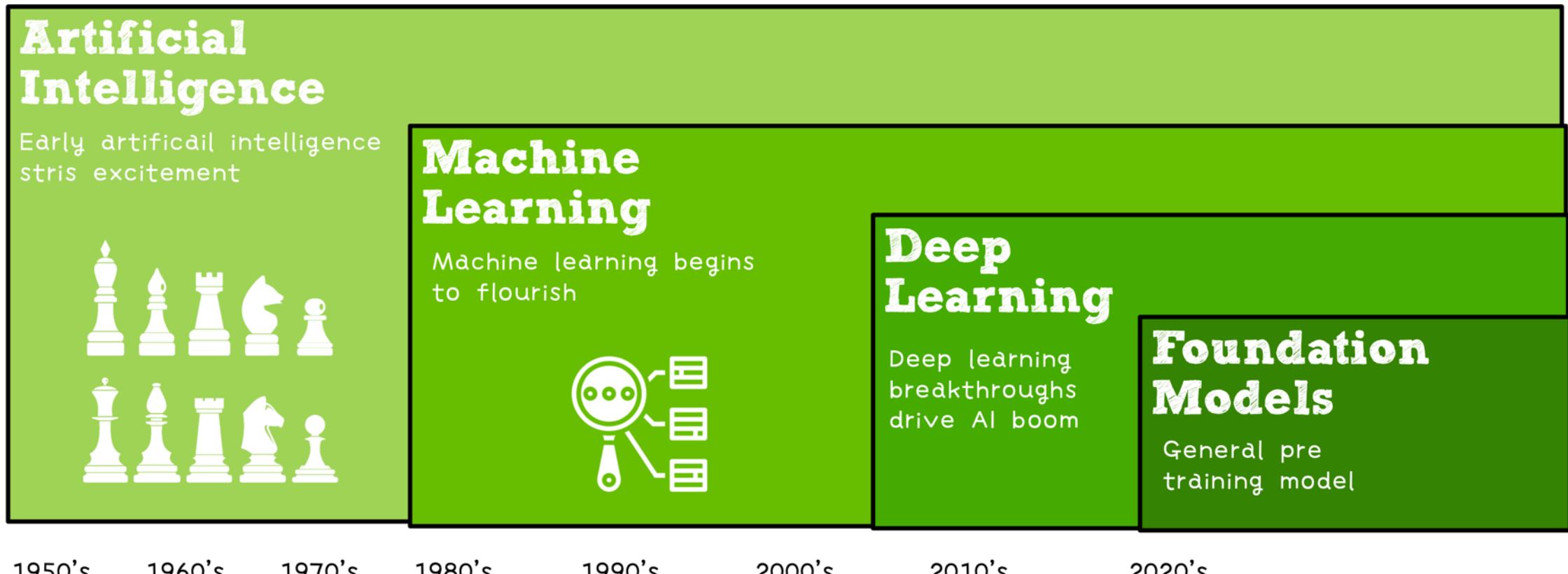
数据驱动 (Data Driven)

- 不管是有监督学习还是无监督学习，人工智能算法是数据驱动 (Data Driven) 的方式解决问题，从数据中不断学习出规律和模型，进而完成预测任务。



数据集原始积累

- 互联网和移动互联网的兴起，让企业越来越方便收集和存储数据；



海量数据为 AI 奠基

1. 推动深度学习算法不断在指定任务上产生更高的准确度与更低的误差。产生了针对深度学习的系统与硬件发展的用户基础，应用落地场景驱动力和研发资源投入。
2. 海量的数据集让单机越来越难以完成深度学习模型的训练，进而产生了分布式训练和平台的需求，让传统的机器学习库不能满足相应的需求。
3. 多样的数据格式和任务，产生了模型结构的复杂性，驱动框架或针对深度学习的程序语言需要有更灵活的表达能力对问题进行表达与映射。

4. AI算法的进步

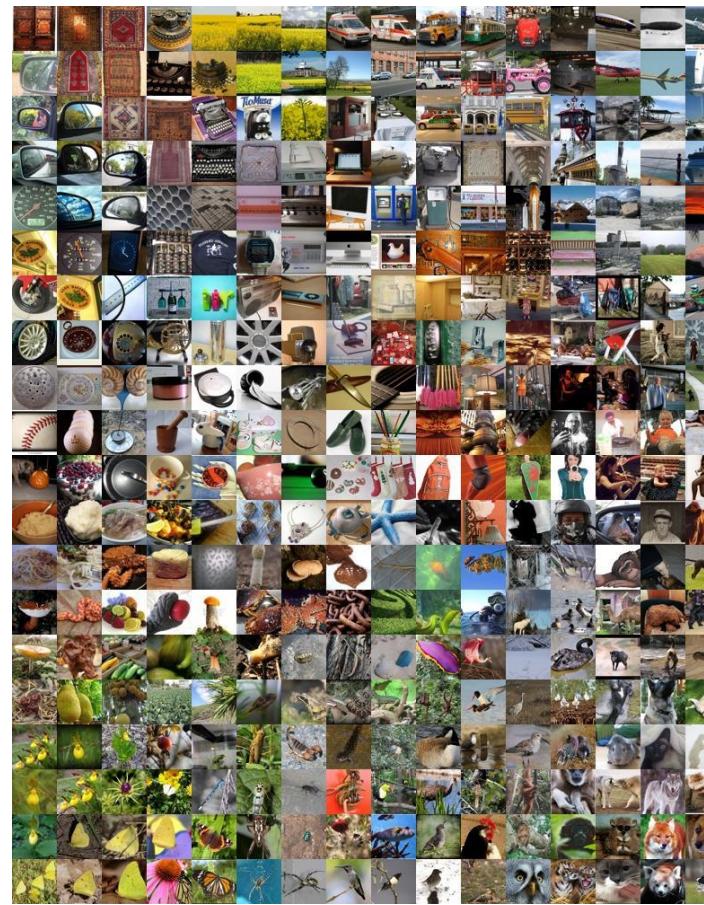


数据

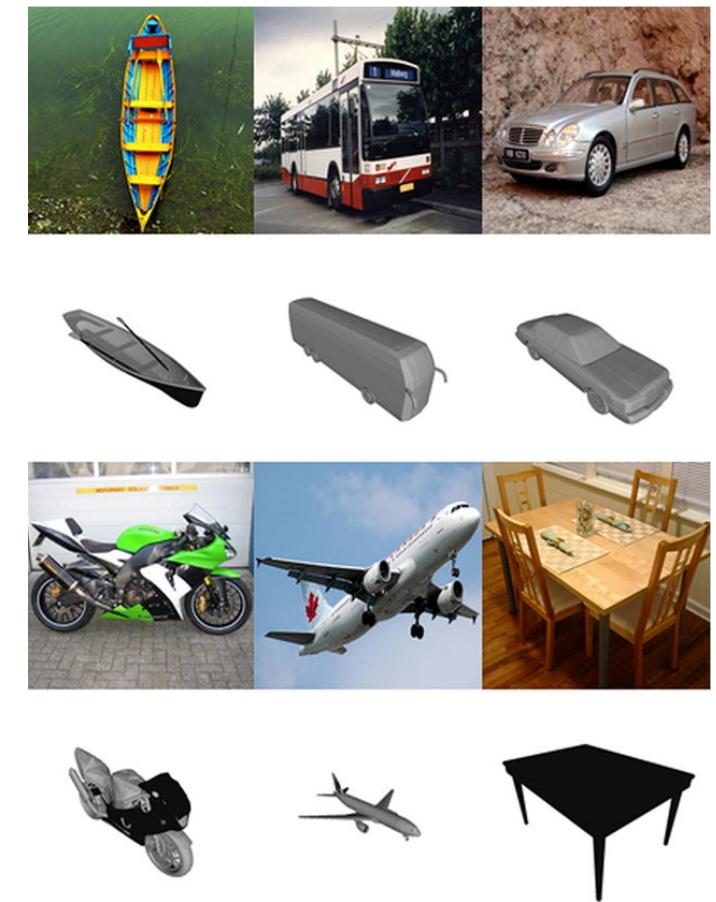
MNIST



ImageNet



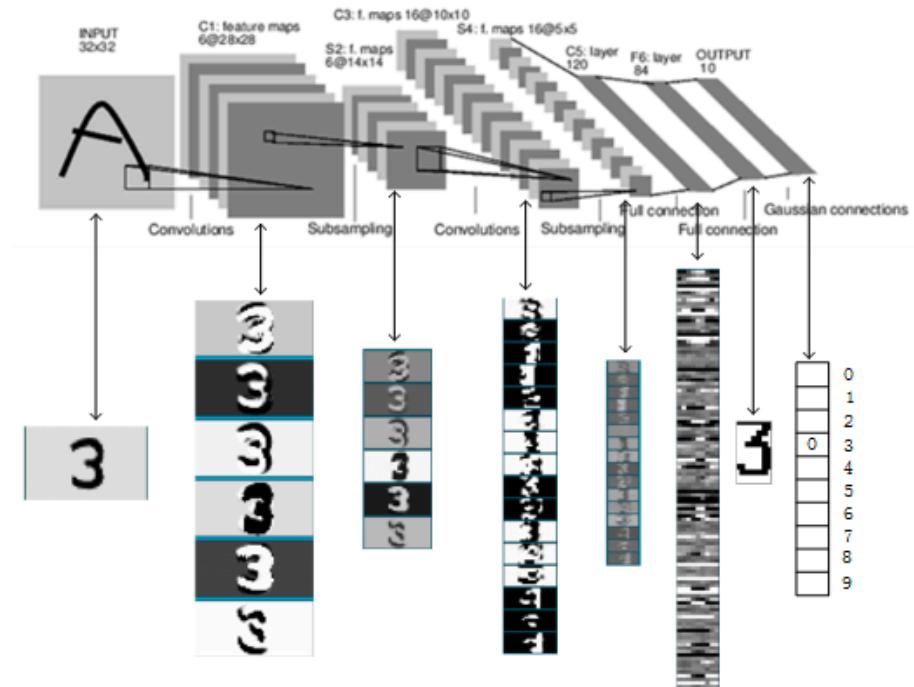
PASCAL3D+



MNIST 数据集 + LeNet5 模型

- 1998 年，一个简单的卷积神经网络可以取得和 SVM 取得的最好效果接近。
- 2012 年，CNN 可以将错误率降低到 0.23%，此结果已经可以和人类所达错误率 0.2% 非常接近。

0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 1 1 1 1 1 1 1 1 1 1 1 1 1
2 2 2 2 2 2 2 2 2 2 2 2 2 2
3 3 3 3 3 3 3 3 3 3 3 3 3 3
4 4 4 4 4 4 4 4 4 4 4 4 4 4
5 5 5 5 5 5 5 5 5 5 5 5 5 5
6 6 6 6 6 6 6 6 6 6 6 6 6 6
7 7 7 7 7 7 7 7 7 7 7 7 7 7
8 8 8 8 8 8 8 8 8 8 8 8 8 8
9 9 9 9 9 9 9 9 9 9 9 9 9 9

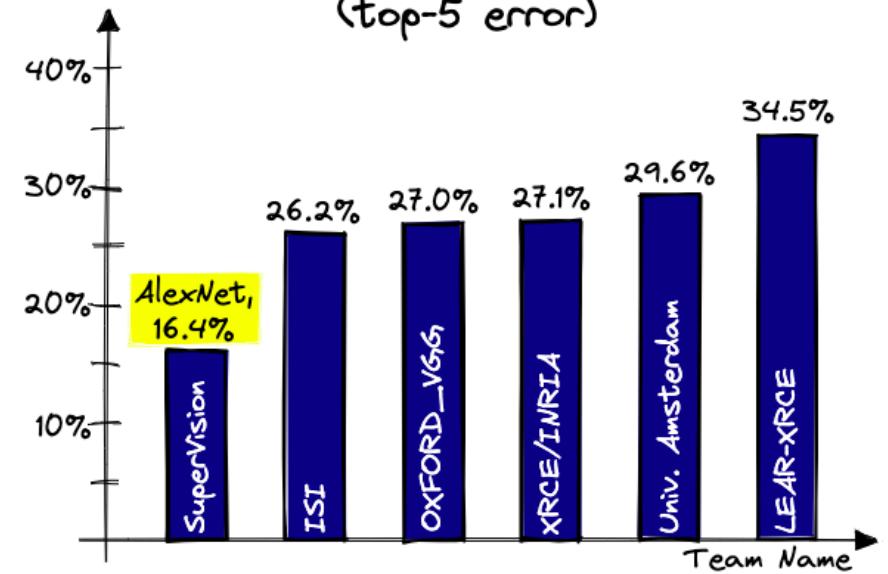


ImageNet 数据集 + AlexNet 模型

- 1998 年的 Lenet5 到 2012 年 AlexNet，不仅效果提升，模型持续增大变深，同时引入了 GPU 训练，新的层（ReLU 等）。

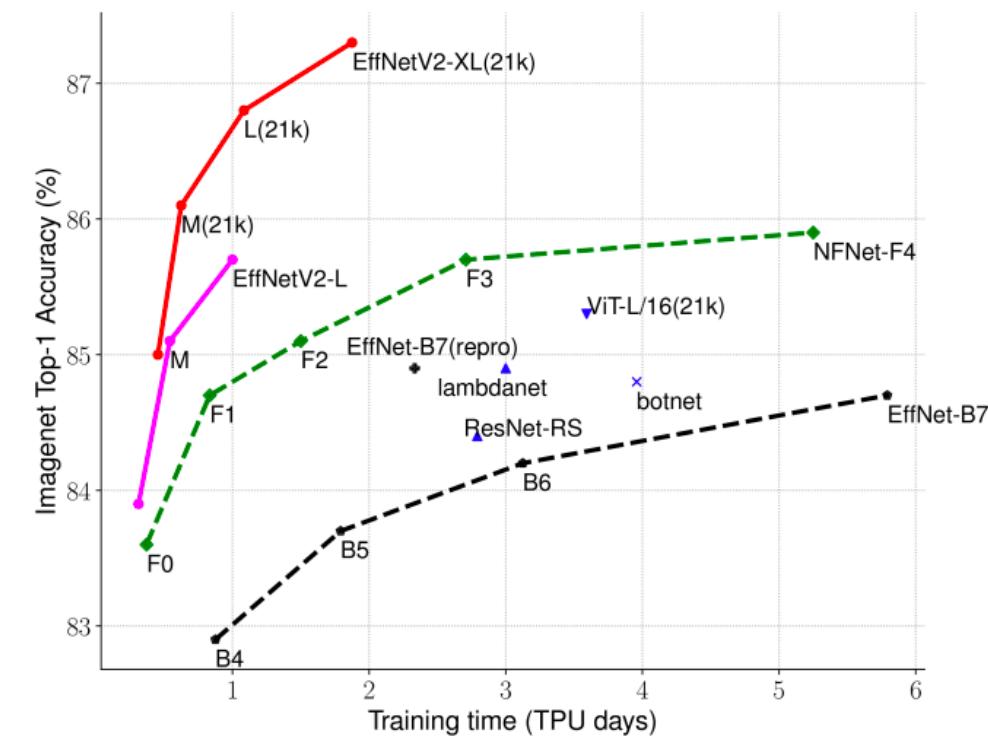
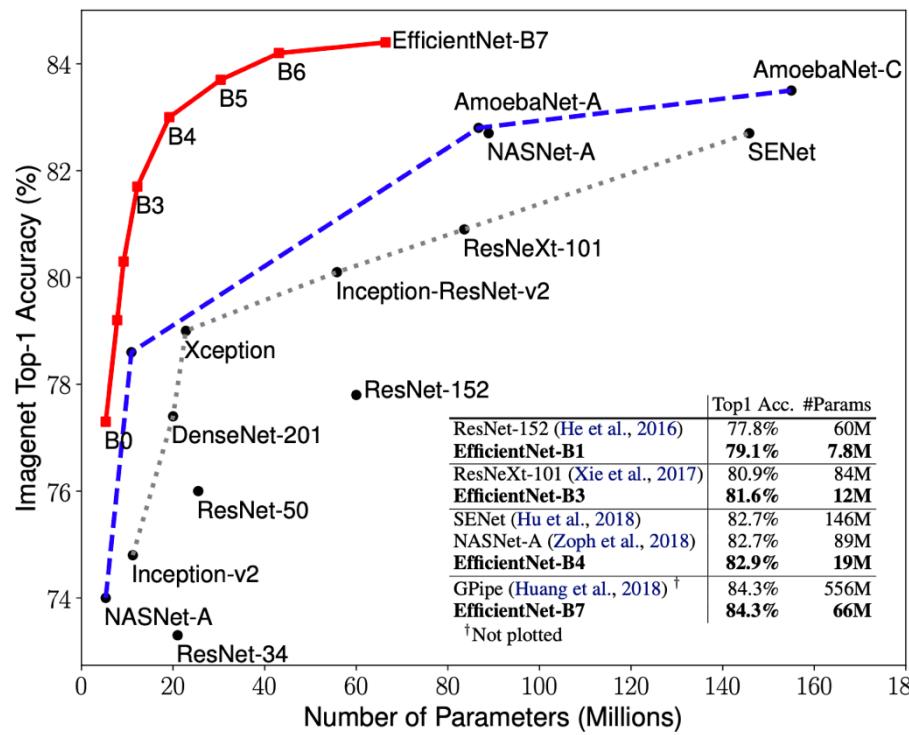


2012 ImageNet Challenge
(top-5 error)



深度学习在 ImageNet 精度持续突破

- 2015 年 Inception，模型进一步复杂新的层被提出，错误率进一步降低到 6.7%。到 2015 年 ResNet 模型层数进一步加深，甚至达到上百层。到 2019 年 NAS，模型设计朝自动化方式设计。



算法的主要演进点

- **创新的激活函数和模型层**：ReLU、GReLU、Batch Normalization、Group Normalization 等；
- **更复杂更深的网络结构**：ResNet50、ResNet101、ResNet150等；
- **更好的训练技巧**：正则化（Regularization）、初始化（Initialization）、学习方法（Learning）等。

其他领域算法突破：NLP

- 2019年，在斯坦福大学举办的 SQuAD (Stanford Question Answering Dataset) 和 CoQA (Convitational Question Answering) 挑战赛中，微软亚洲研究院 (MSRA) 的 NLP 团队通过多阶段 (Multi-Stage)，多任务 (Multi-Task) 学习的方式取得第一。

Passage Sentence

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity.

Question

What causes precipitation to fall?

Answer Candidate

gravity

- Path from passage sentence words (that also occur in question) to answer



- Combined with path from wh-word to question word.



其他领域算法突破：Speech

- 2016 年，微软研究院（ MSR ）提出的 Combined 模型系统的在 NIST 2000 数据集上错误率为 6.2 % ，超越之前报告的基准测试结果。



其他领域算法突破：Reinforcement

- 2016年，Google DeepMind 研发的 AlphaGo 在围棋比赛中以 4:1 的高分击败了世界大师级冠军李世石。OpenAI 训练出了名为 OpenAI Five 的 Dota 2 游戏智能体，击败 Dota 2 世界冠军战队，这是首个击败电子竞技游戏世界冠军的人工智能系统。

Task domain	DM Control Suite / Real World RL Suite	DM Locomotion Humanoid	DM Locomotion Rodent	Atari 2600
Action space	continuous	continuous	continuous	discrete
Observation space	state	pixels	pixels	pixels
Exploration difficulty	low to moderate	high	moderate	moderate
Dynamics	deterministic / stochastic	deterministic	deterministic	stochastic



数据、算法对 AI 系统的挑战

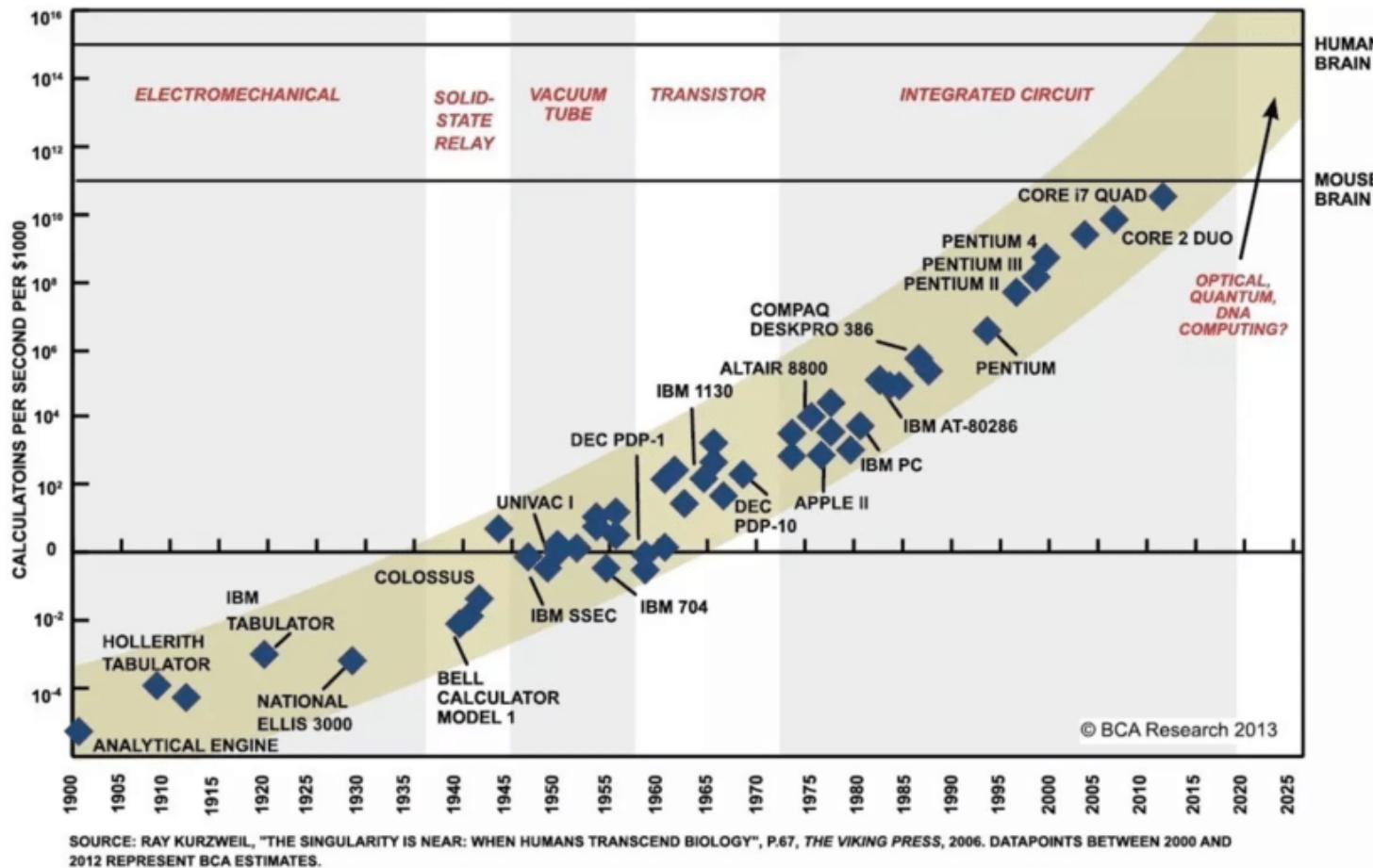
- 除了数据本身不断的沉淀，算法研究员和工程师不断设计新的算法和模型提升预测效果，不断取得突破性进展。
- 新算法和模型结构需要 AI 框架提供编程的表达力和灵活性，对硬件执行的系统优化有可能会改变现有的计算架构，进而产生了针对 AI 系统的软硬件架构变革的挑战。

5. 计算体系的进步

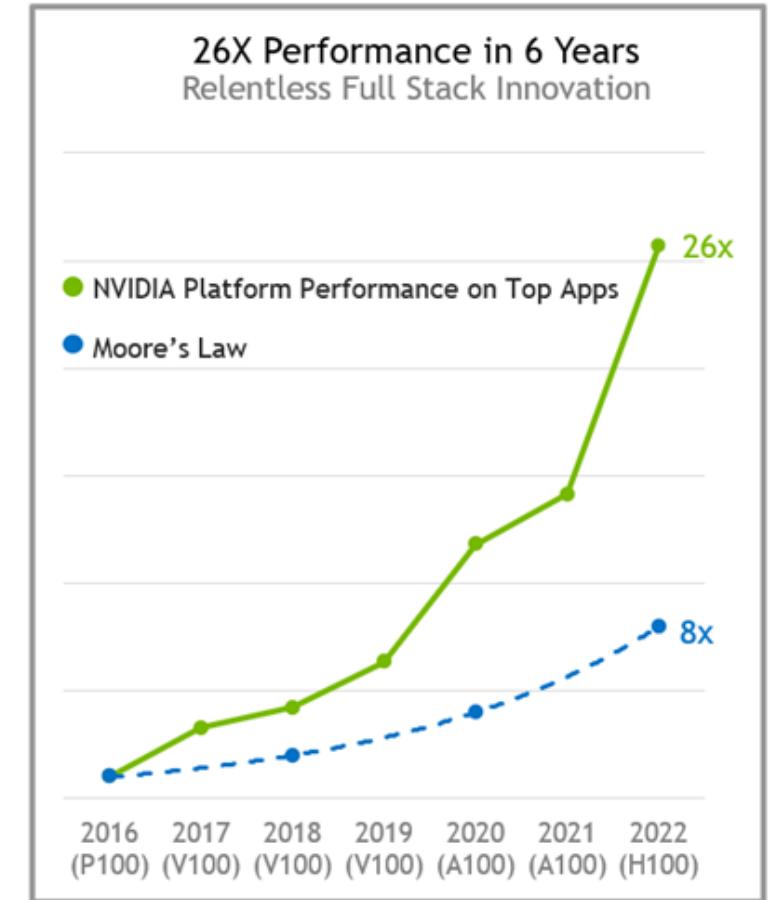


算力提升

CPU

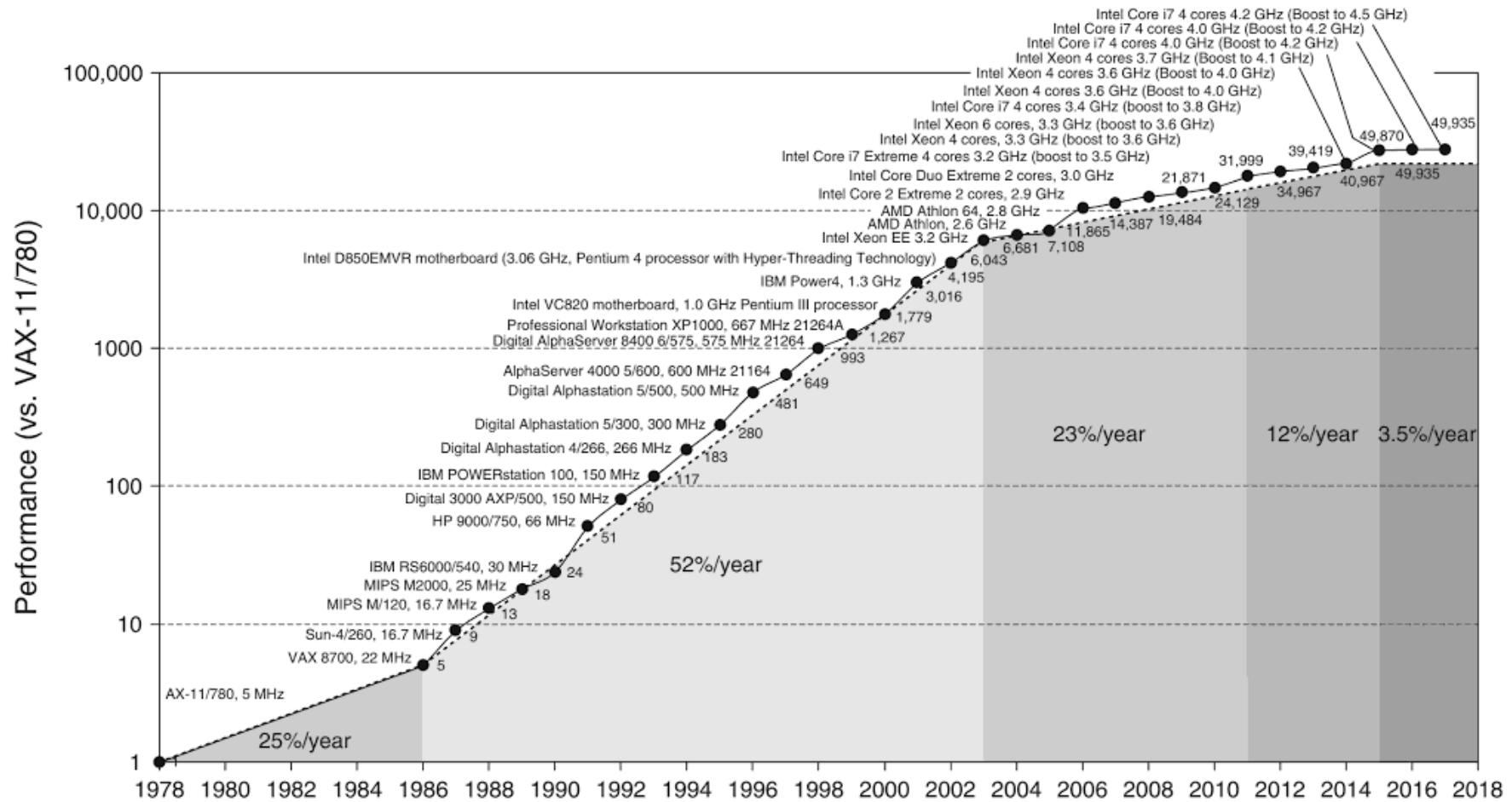


GPU



摩尔定律

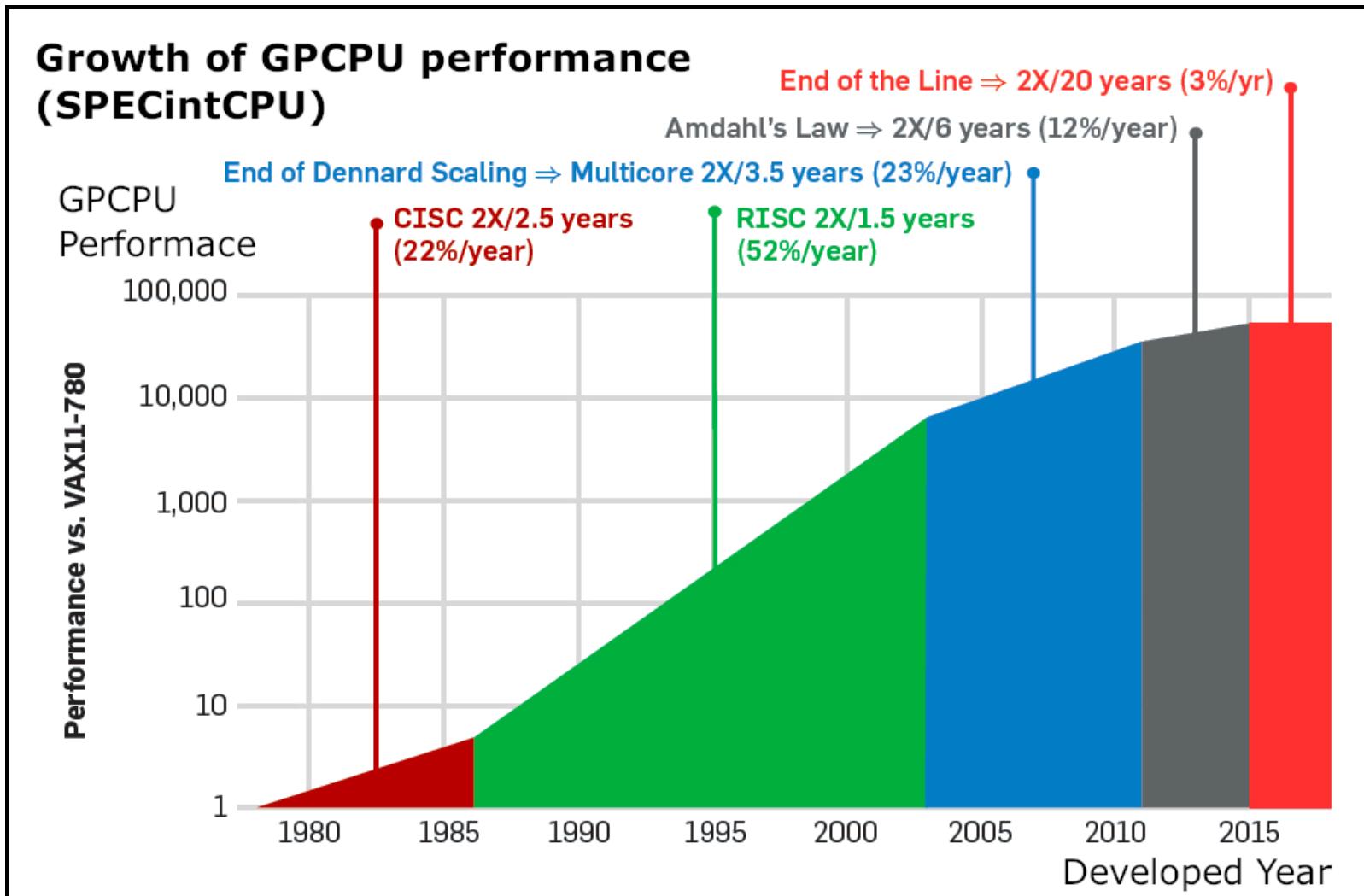
- 从 1960 年以来，计算机性能增长主要来自摩尔定律
- 到二十世纪初大概增长了 10^8 倍。
- 摩尔定律停滞，性能增长逐渐放缓。单纯靠工艺进步，无法满足各种应用对性能要求。



CPU 性能提升的五个阶段

1. CISC阶段。上世纪80年代，x86架构为代表的CISC架构开启了CPU性能快速提升的时代，CPU性能每年提升约25%（图中22%数据有误），大约3年性能可以翻倍。
2. RISC阶段。CISC指令系统越来越复杂，而RISC证明了“越精简，越高效”。随着RISC架构的CPU开始流行，性能每年可以达到52%，性能翻倍只需要18个月。
3. 多核阶段。单核CPU的性能提升越来越困难，通过集成更多CPU核并行的方式来进一步提升性能。这一时期，每年性能提升可以到23%，性能翻倍需要3.5年。
4. 多核性能递减阶段。随着CPU核的数量越来越多，阿姆达尔定律证明了处理器数量的增加带来的收益会逐渐递减。这一时期，CPU性能提升每年只有12%，性能翻倍需要6年。
5. 性能提升瓶颈阶段。不管是从架构/微架构设计、工艺、多核并行等各种手段都用尽的时候，CPU整体的性能提升达到了一个瓶颈。从2015年之后，CPU性能每年提升只有3%，要想性能翻倍，需要20年。

CPU 性能提升的五个阶段



GPU 性能提升

- 1995 年后开始为特殊应用定制专用芯片，通过消除通用处理器中冗余的功能部分，来进一步提高对特定应用的计算性能。比如，图形图像处理器 GPU 就对图像类算法做专用加速。后来出现 GPGPU，对适合于抽象为单指令流多数据流（ SIMD ）的并行算法与工作负载能起到加速效果。



DSA 专用硬件

- 对深度学习模型中计算模式进行抽象，转换为矩阵乘法或非线性变换，根据专用负载特点进一步定制流水线化执行的硬件，减少访存提升计算密度，提高专用应用性能。
- 深度学习负载本身在算法层常常应用的稀疏性和量化等加速手段也逐渐被硬件厂商定制到专用加速器中，在专用计算领域进一步协同优化加速。

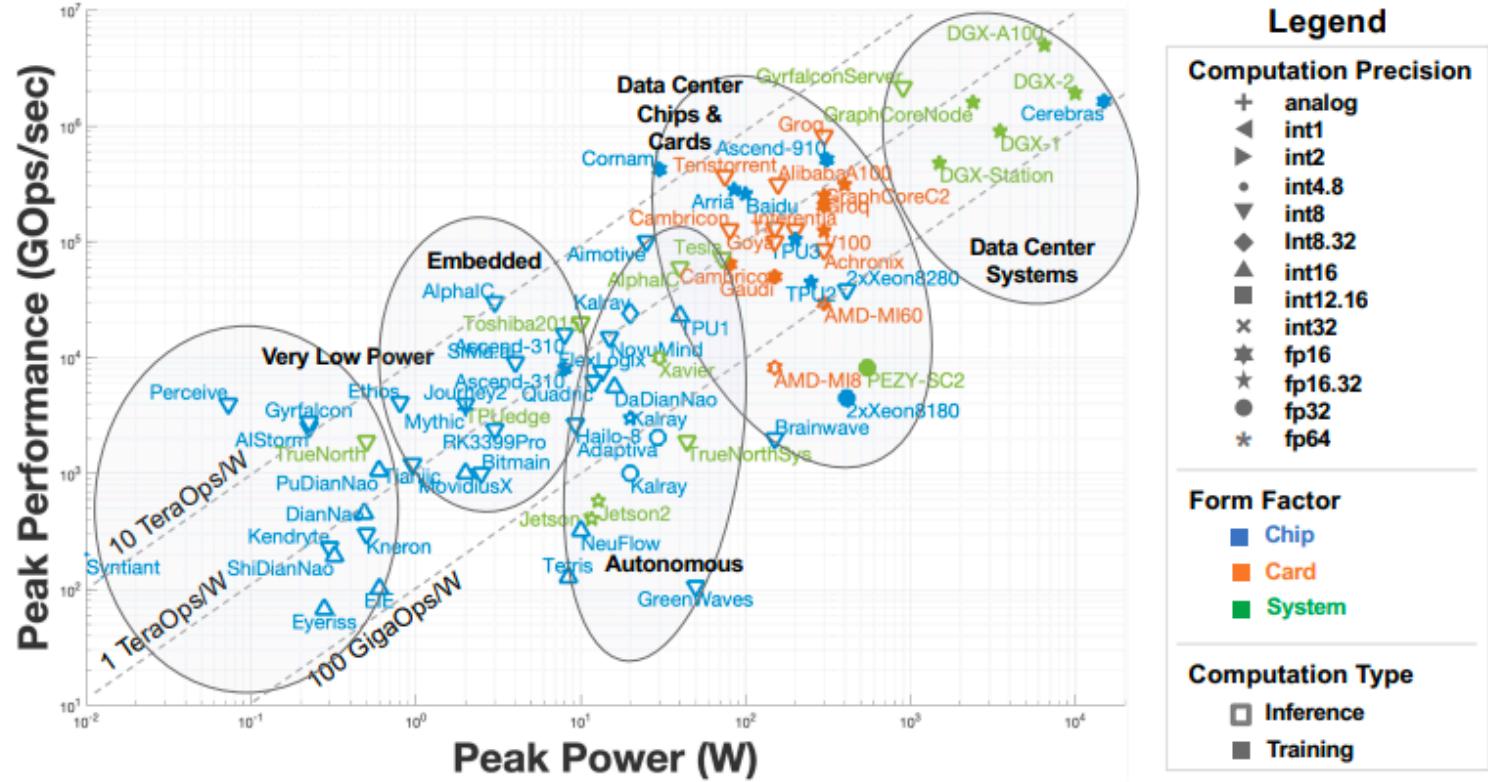


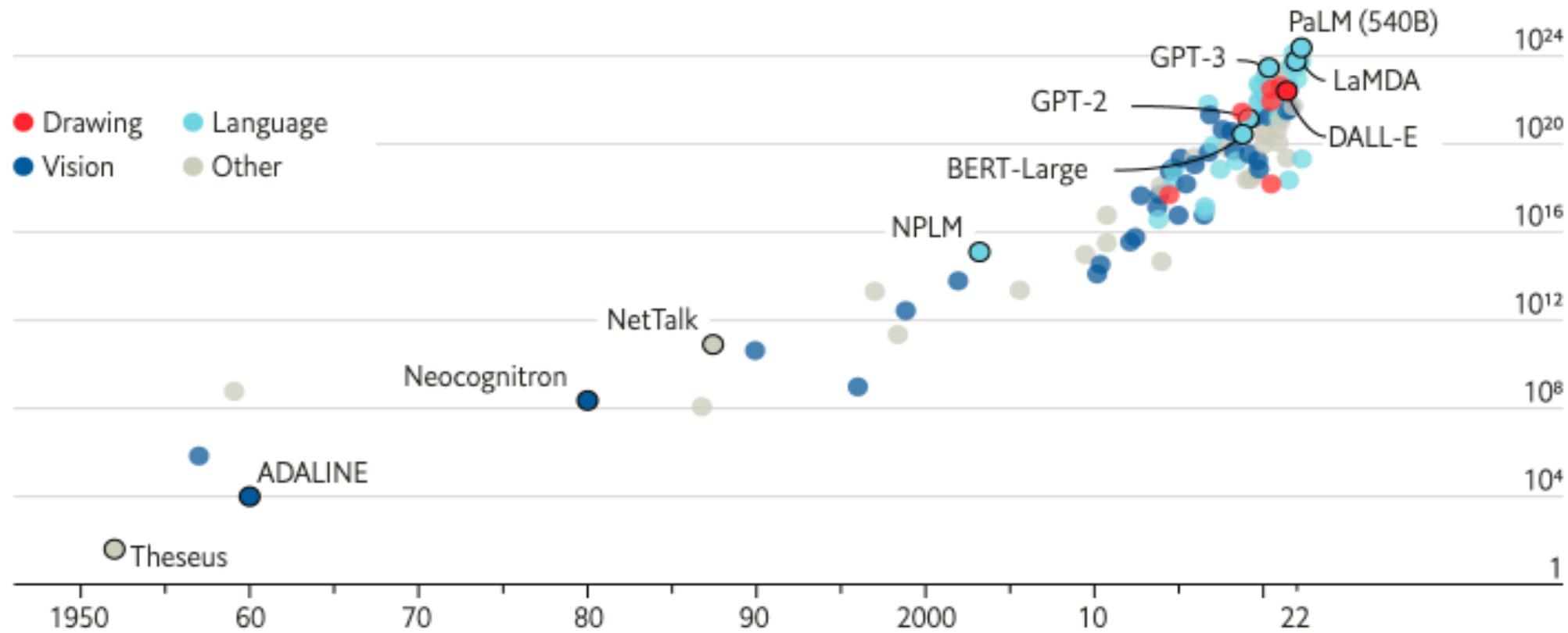
Fig. 2. Peak performance vs. power scatter plot of publicly announced AI accelerators and processors.

模型对算力的需求

The blessings of scale

AI training runs, estimated computing resources used

Floating-point operations, selected systems, by type, log scale



Sources: "Compute trends across three eras of machine learning", by J. Sevilla et al., arXiv, 2022; Our World in Data

算力瓶颈之后

- 算力还是可能在短期内成为瓶颈，那么人工智能系统的性能下一代的出路在哪？
- 在后面的 AI 系统中会深入地介绍，除单独芯片不断迭代进行性能放大（Scale Up），系统工程师不断设计更好的分布式计算系统，将计算并行开来达到向外扩展（Scale Out），同时发掘 AI 作业特点，系统上软硬件协同设计，进一步提升计算效率和性能。

Still curious about AI System ?

- AI 爆发三大因素：数据、算法、算力，对AI 系统带来哪些挑战和冲击？
- 需要配套哪些 AI 系统的体系结构来支撑 AI 行业快速发展？





Thank you

把AI系统带入每个开发者、每个家庭、
每个组织，构建万物互联的智能世界

Bring AI System to every person, home and
organization for a fully connected,
intelligent world.

Copyright © 2023 XXX Technologies Co., Ltd.
All Rights Reserved.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. XXX may change the information at any time without notice.



Course chenzomi12.github.io

GitHub github.com/chenzomi12/DeepLearningSystem