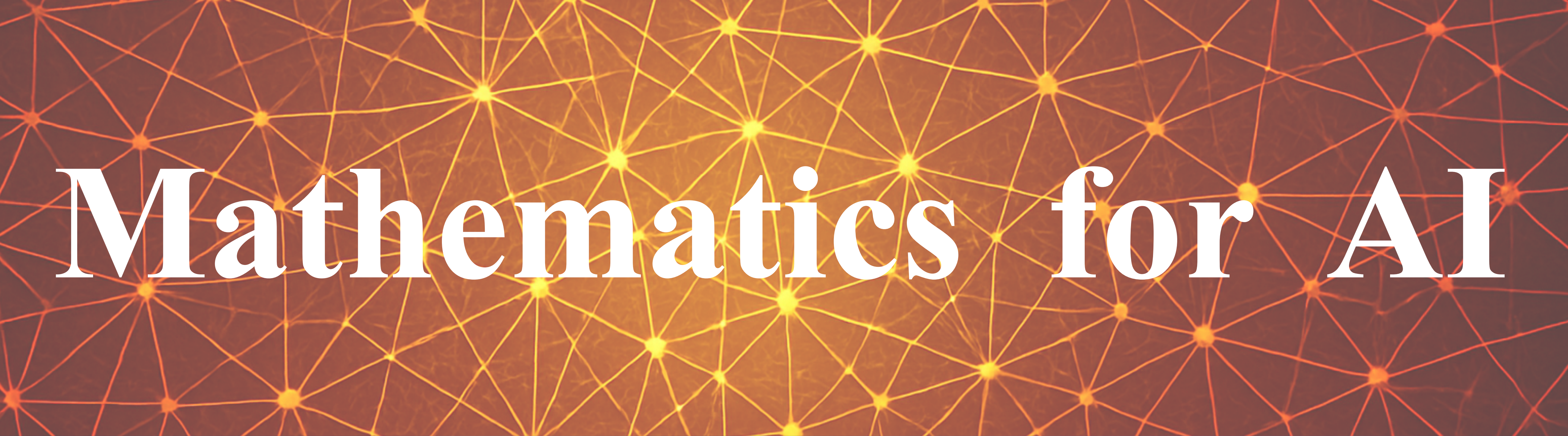


AI & Mathematics

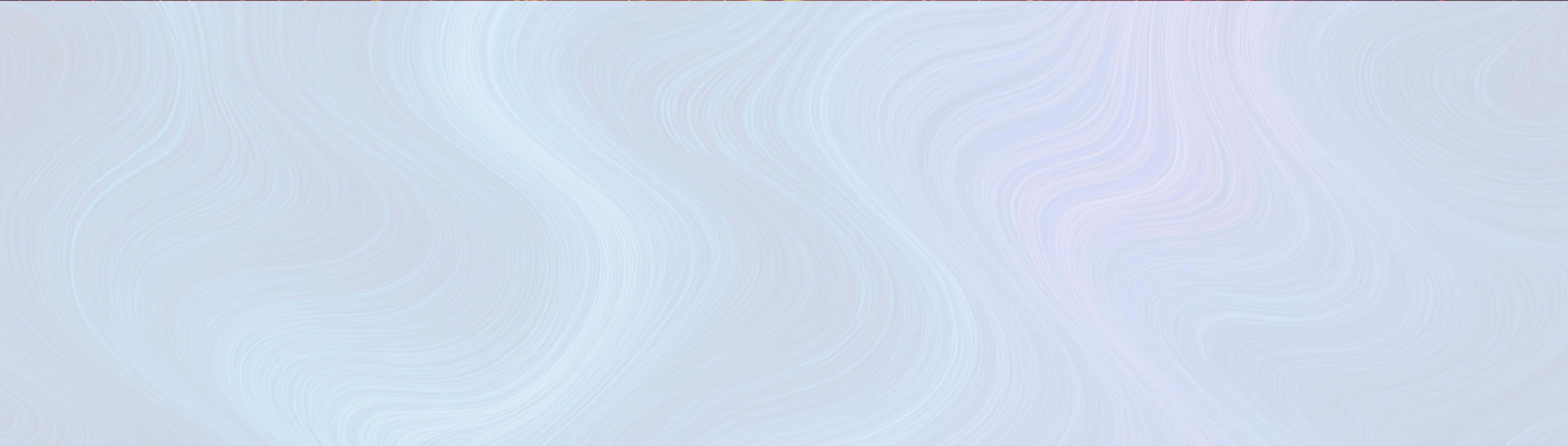


Gabriel Peyré



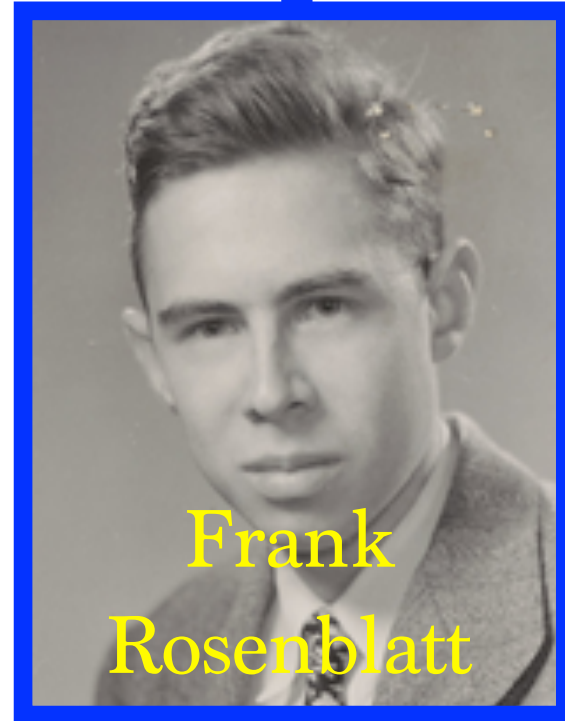


Mathematics for AI



Maths & AI

Perceptron



1957

1986



Backprop

1998



Convets

2011

Adam



2014



AlexNet

ResNets



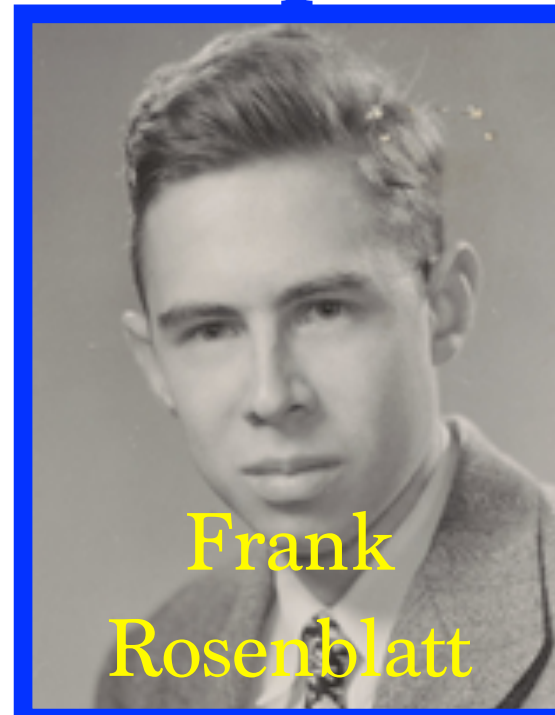
2017



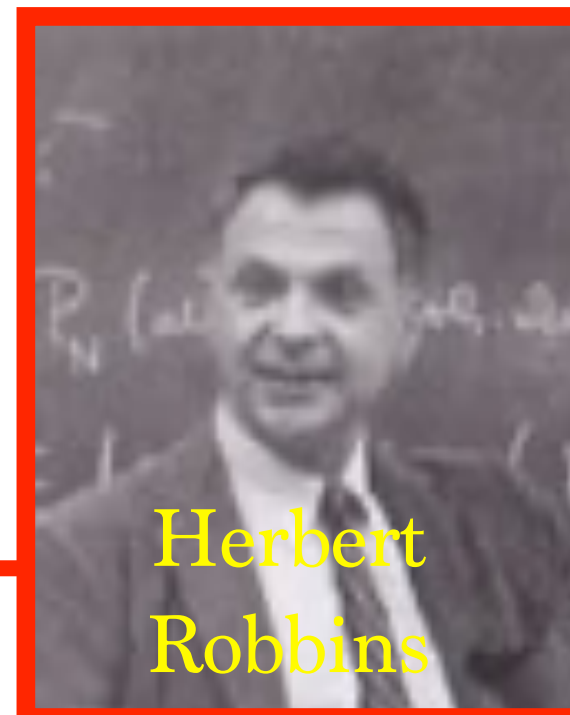
Transformers

Maths & AI

Perceptron Universality



SGD



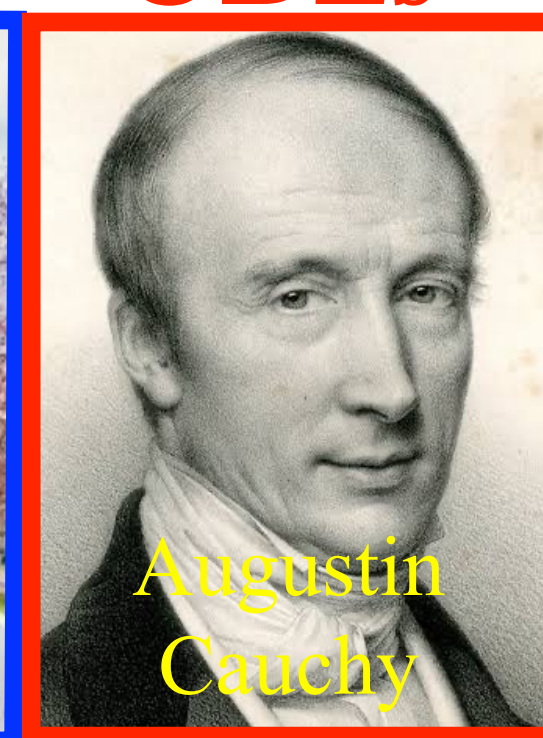
Adam



ResNets



ODEs



1957

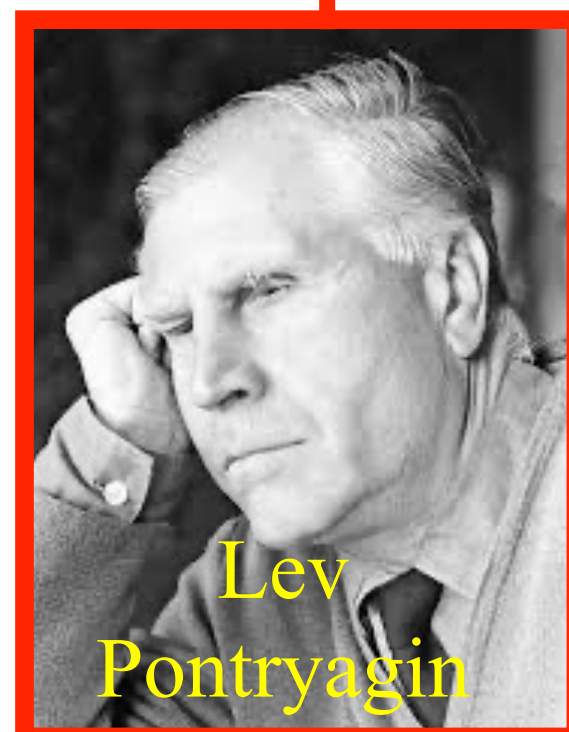
1986

1998

2011

2014

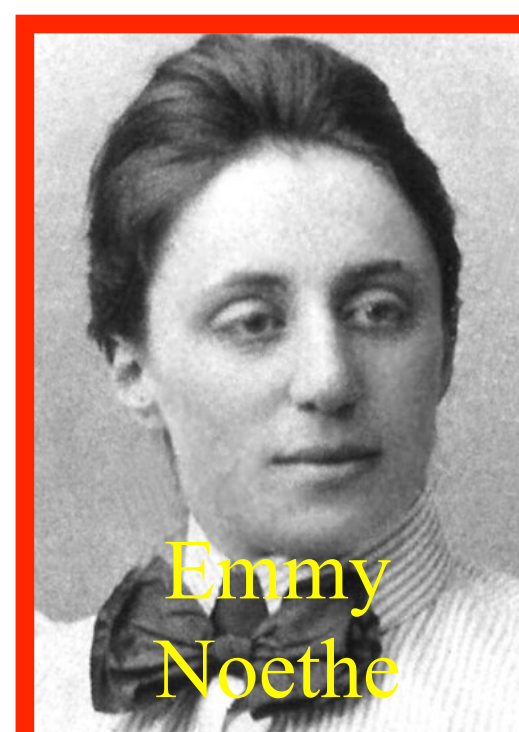
2017



Adjoint



Backprop



Invariances



Convets



Wavelets



AlexNet

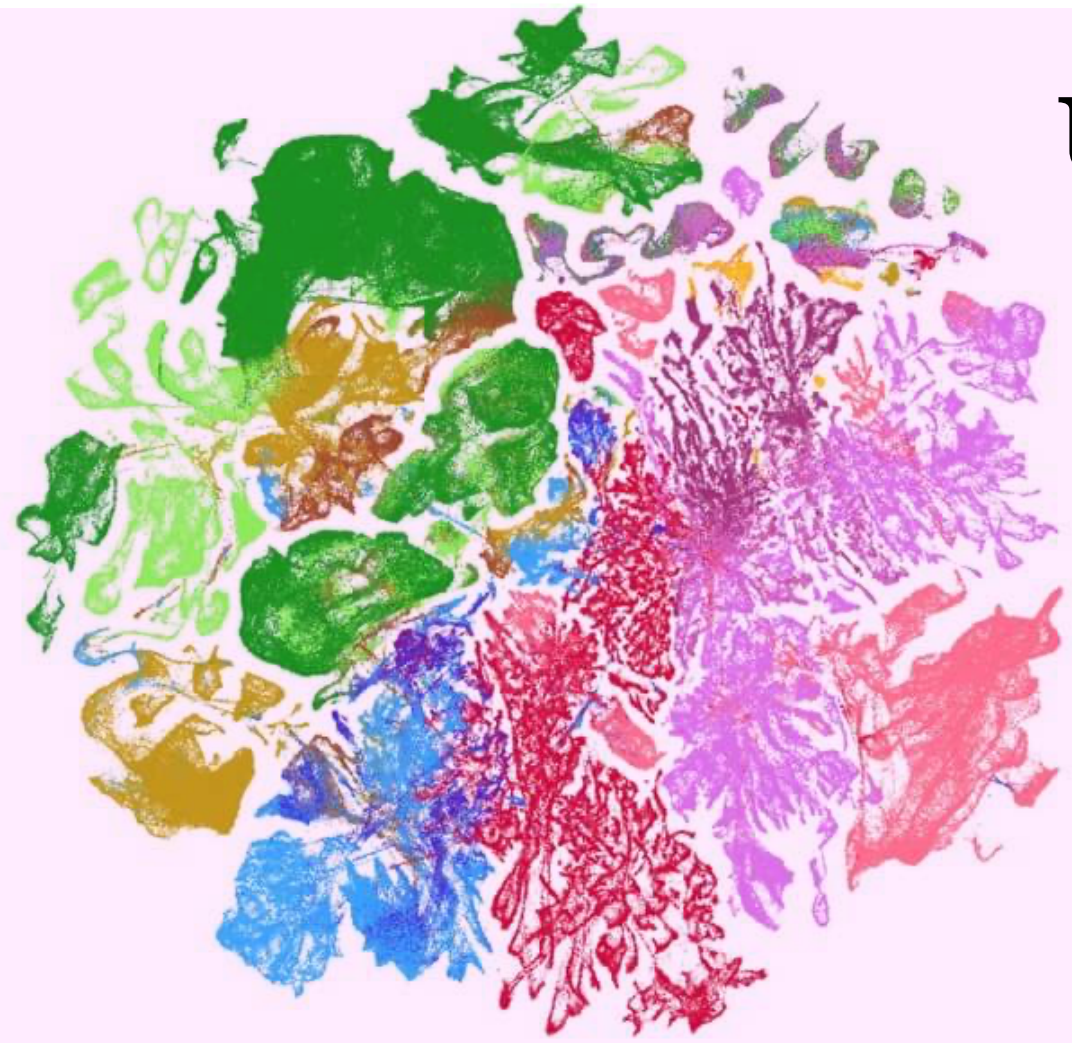


Transformers



PDEs

From Supervised to Generative Learning

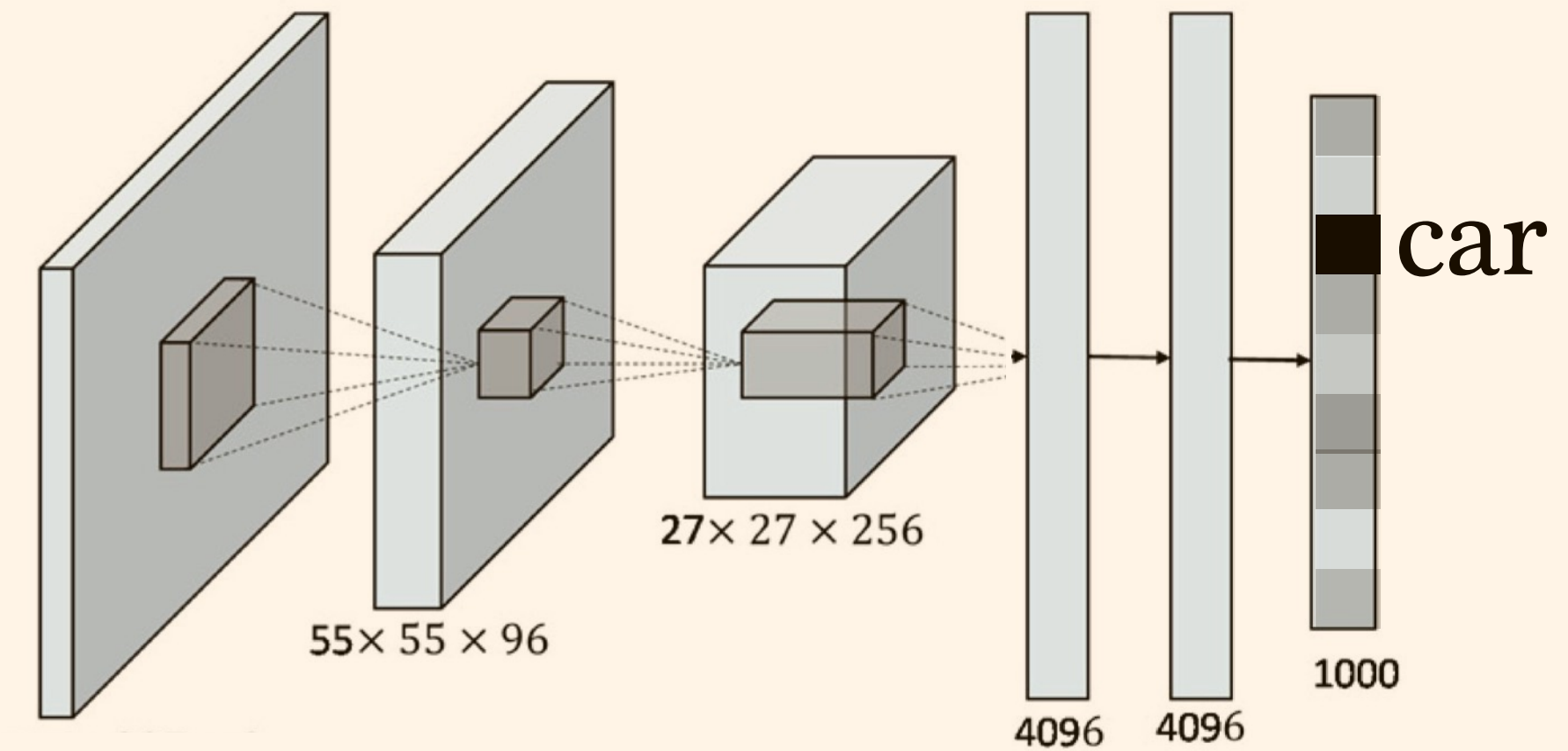


Unsupervised learning

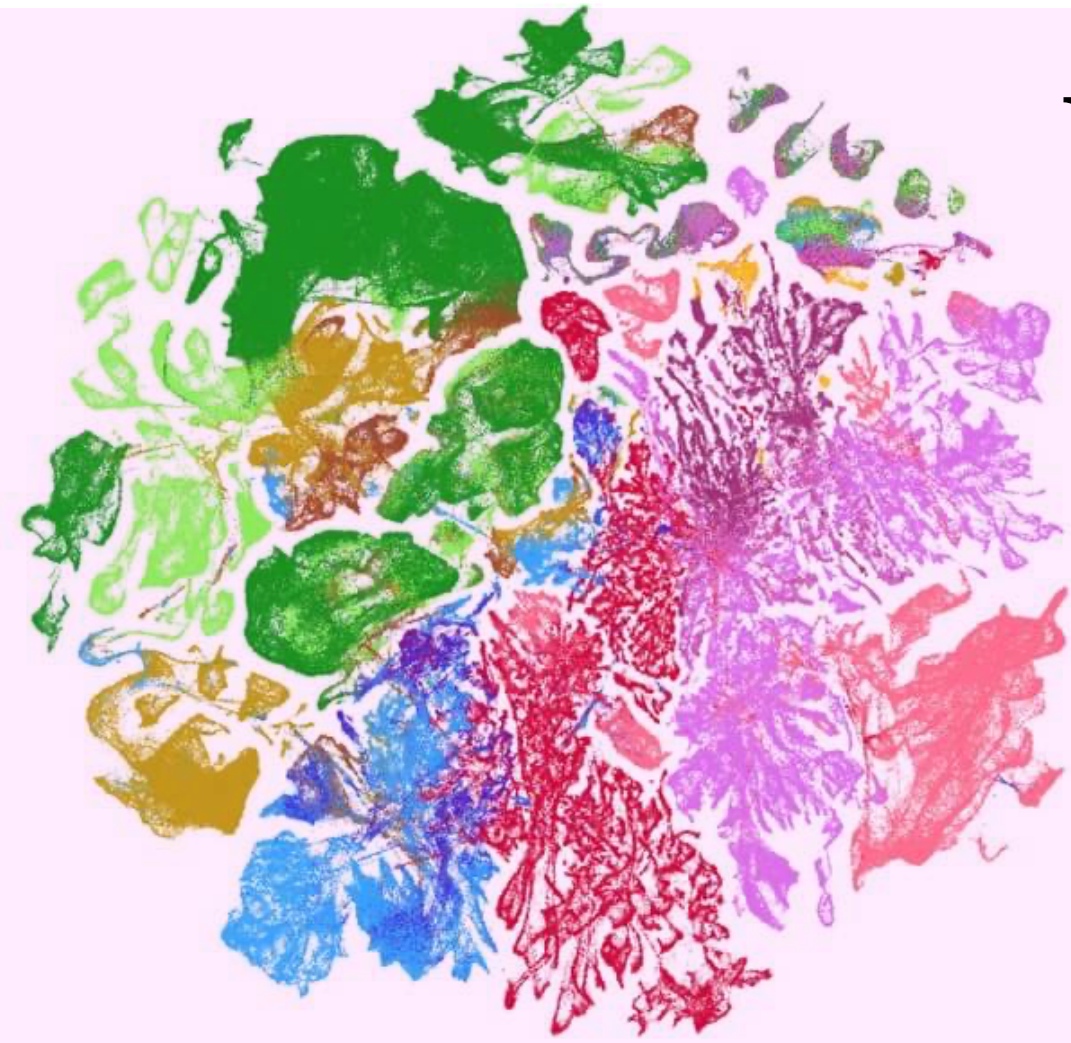
UMAP vizualization and clustering of 4M mouse brain cells [Yao Z. et al. 2023]



Supervised learning

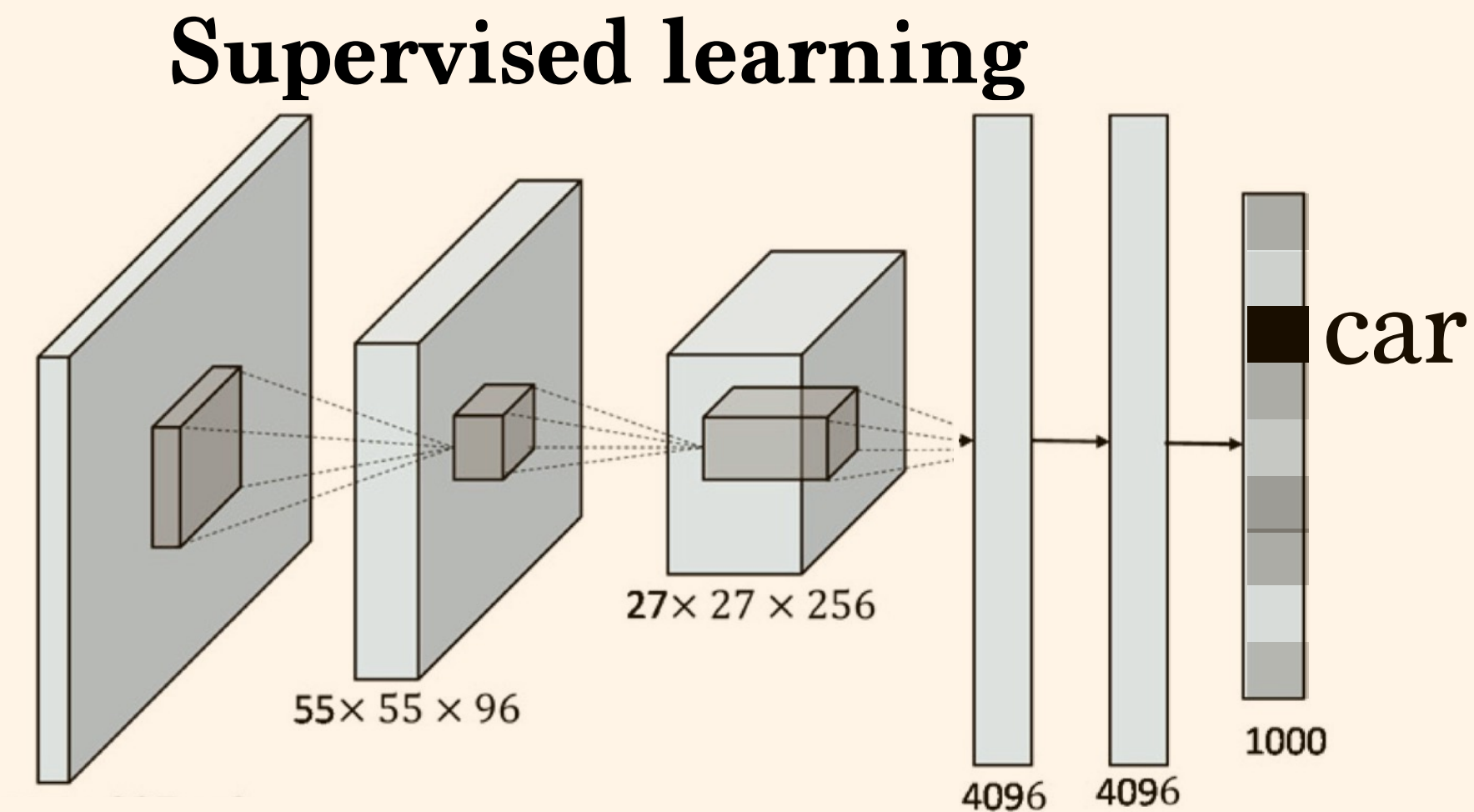


From Supervised to Generative Learning

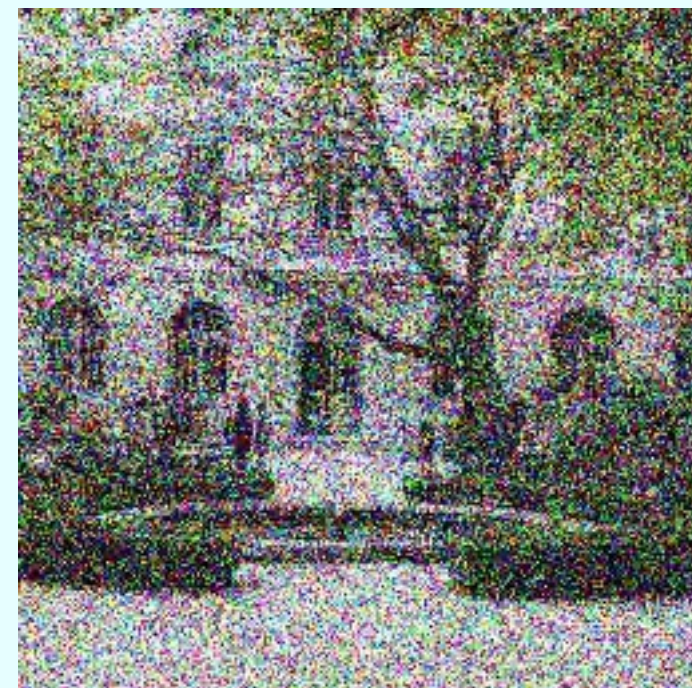


Unsupervised learning

UMAP vizualization and clustering of 4M mouse brain cells [Yao Z. et al. 2023]



Generative IA: Self supervised learning



Add noise

Denoise

L'École normale supérieure6, appelée aussi «de la rue d'Ulm», «Normale Sup'», est l'une des institutions universitaires et de recherche les plus prestigieuses et les plus sélectives de France, spécialisée en lettres et en sciences.

L'École normale supérieure6, appelée aussi «de la rue d'Ulm», «Normale Sup'», est l'une des institutions universitaires et de recherche les plus prestigieuses et les plus sélectives de France, spécialisée en lettres et en sciences.

L'École normale supérieure6, appelée aussi «de la rue d'Ulm», «Normale Sup'», est l'une des institutions universitaires et de recherche les plus prestigieuses et les plus sélectives de France, spécialisée en lettres et en sciences.

Masking

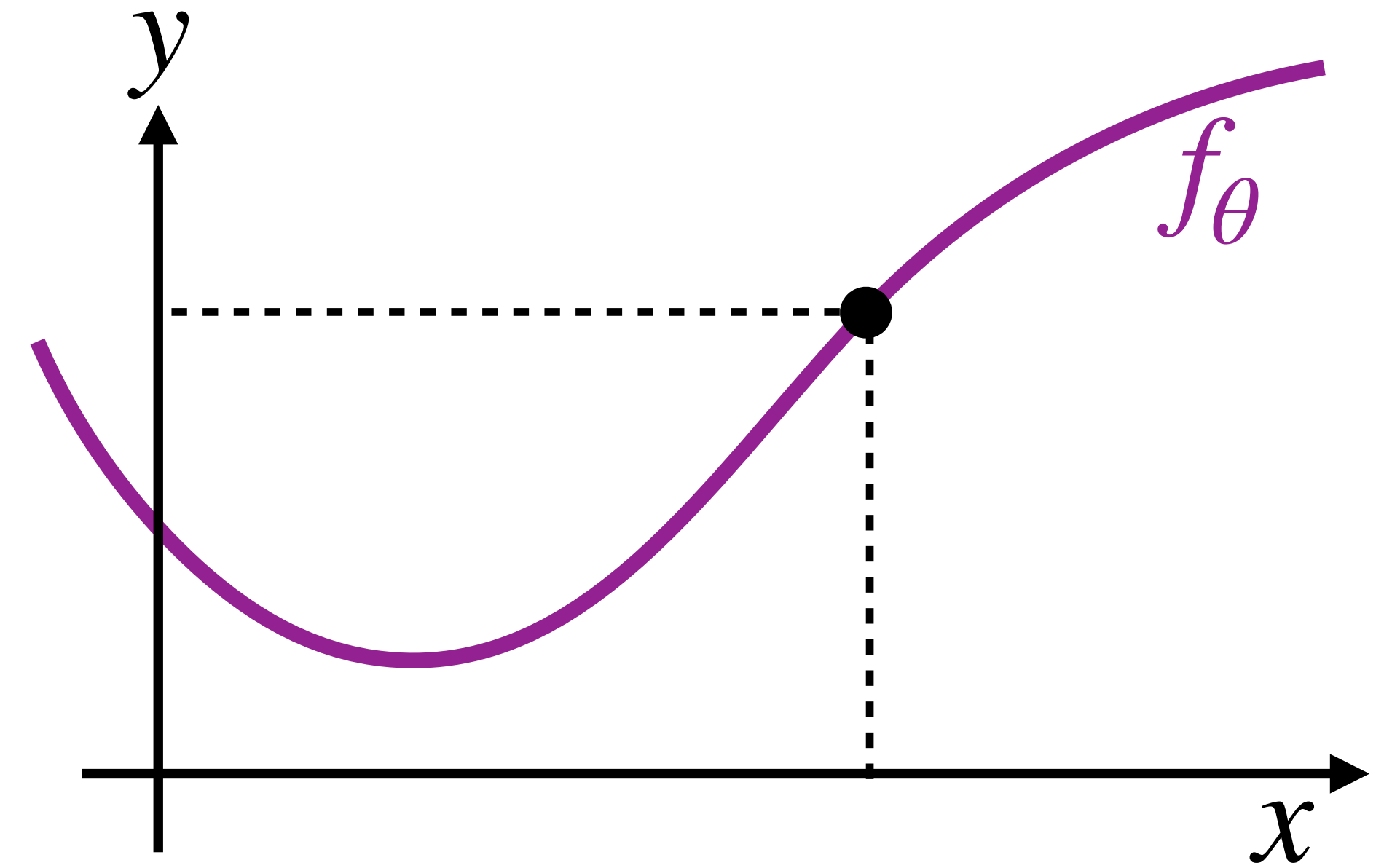
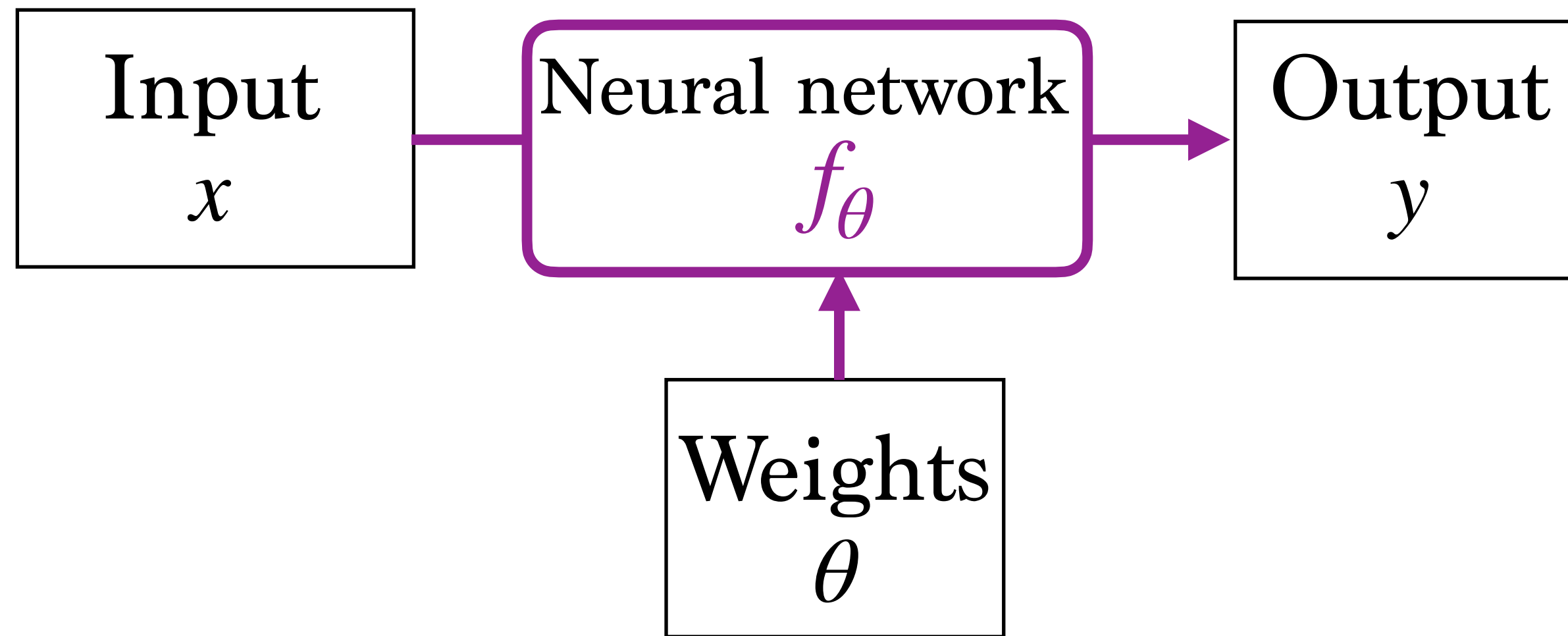
Next token prediction

DALL·E 2

stability.ai



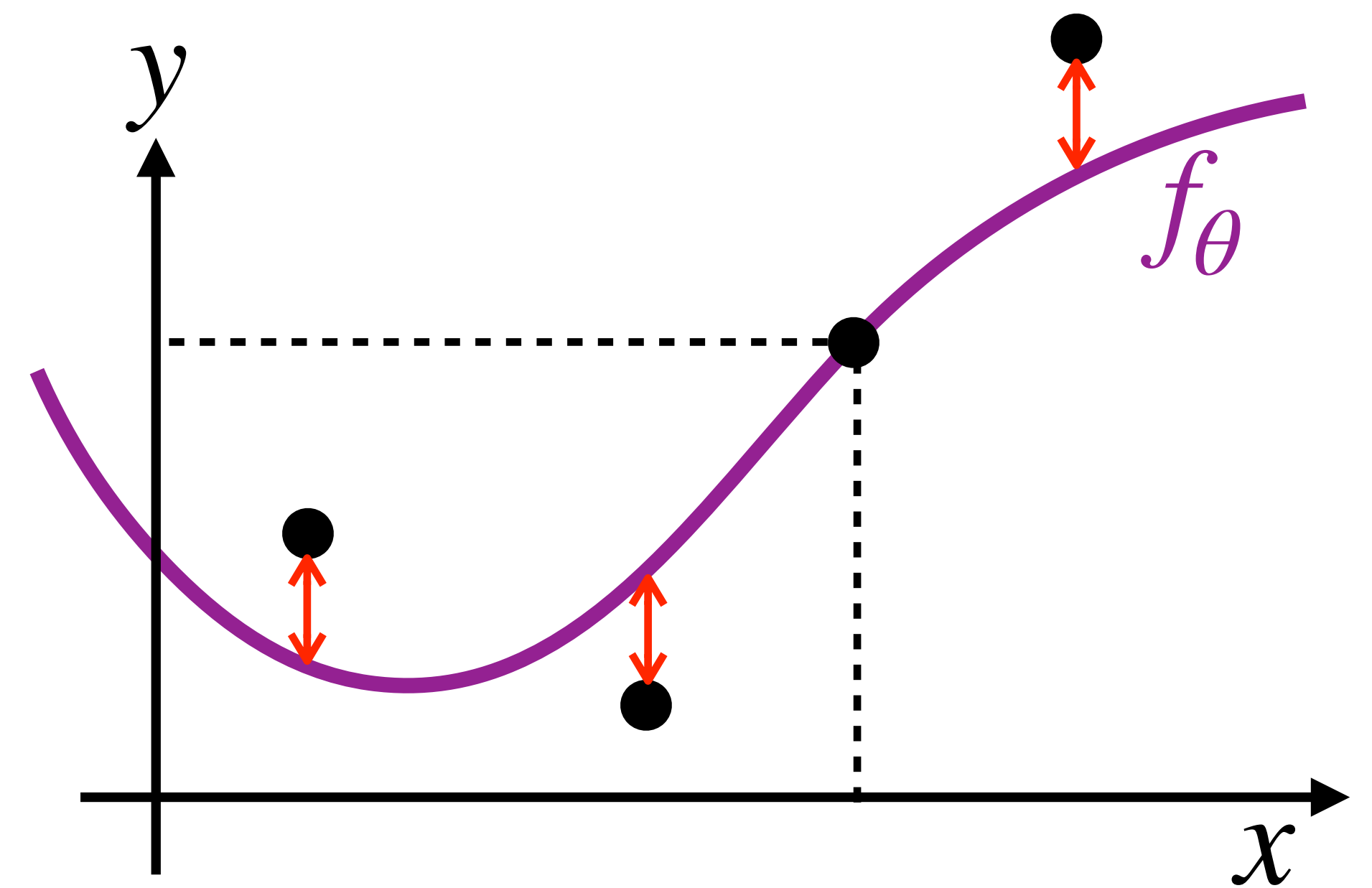
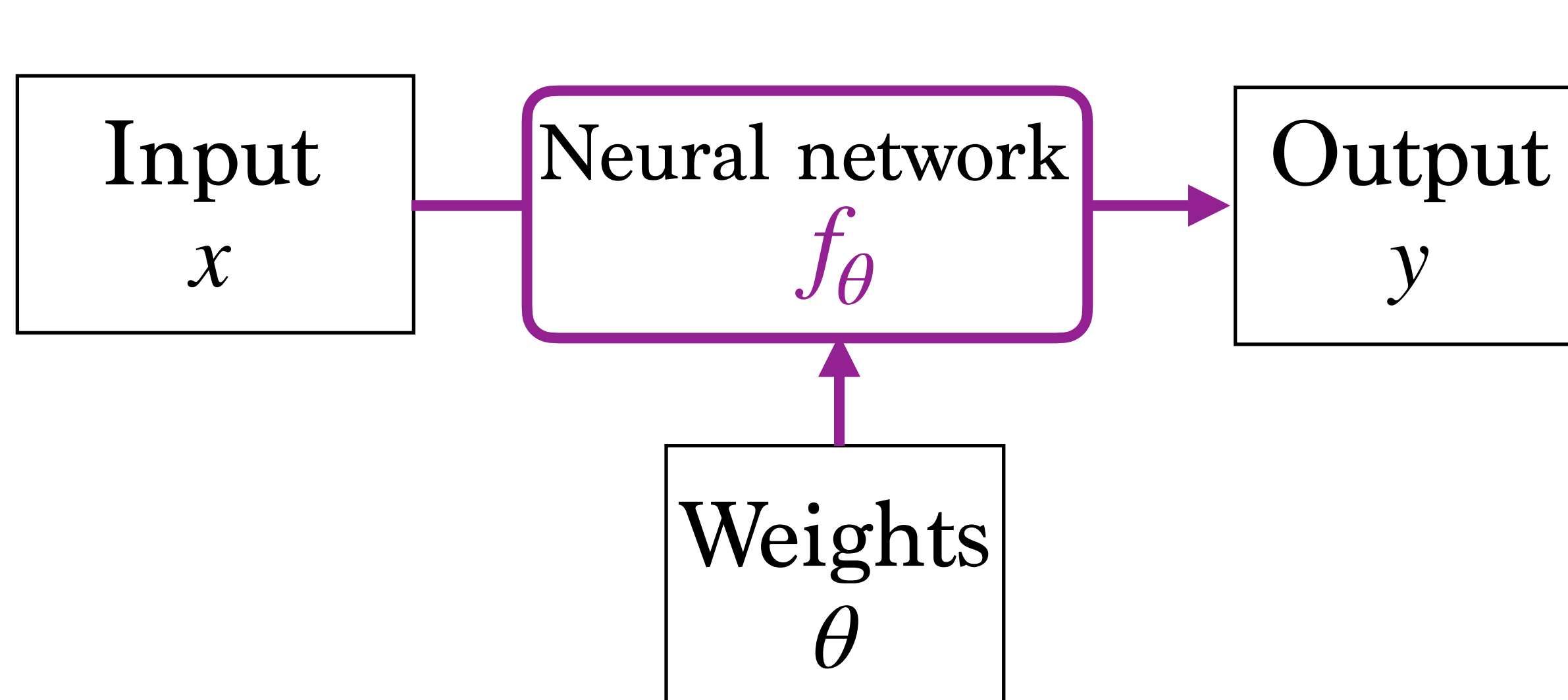
Supervised Learning via Optimization



Dataset $(x_i, y_i)_i$

Goal: $y_i \approx f_\theta(x_i)$

Supervised Learning via Optimization

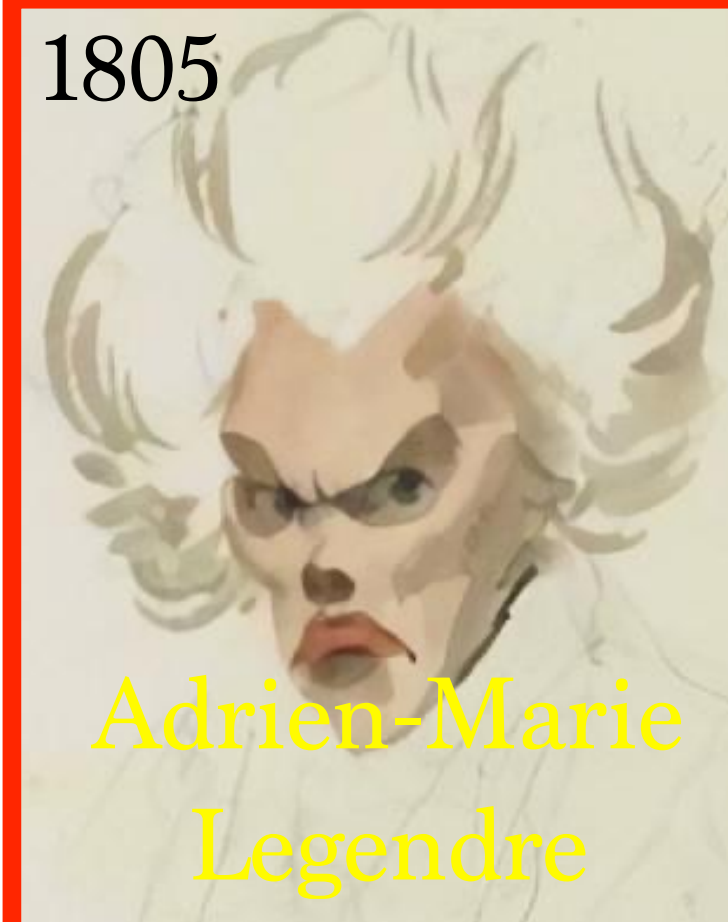


Dataset $(x_i, y_i)_i$

Goal: $y_i \approx f_\theta(x_i)$

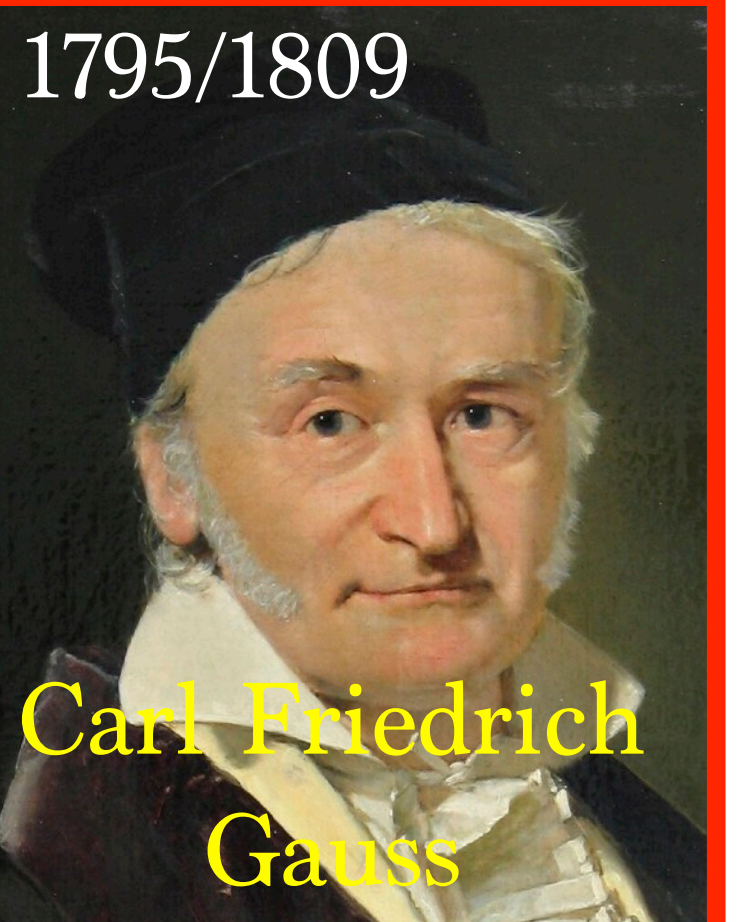
Learning: minimize $E(\theta) := \sum_i \text{Error}(f_\theta(x_i), y_i)$

1805



Adrien-Marie
Legendre

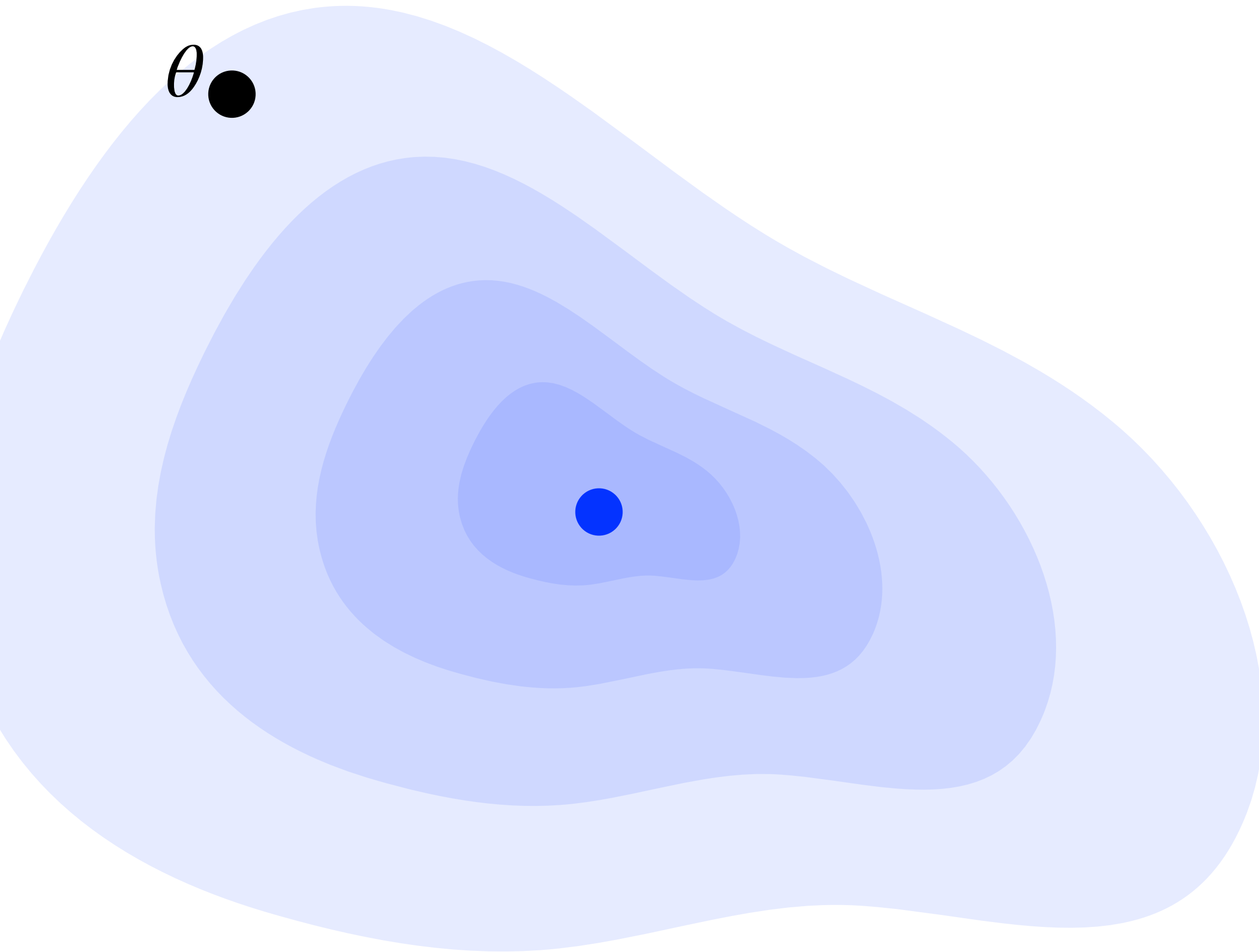
1795/1809



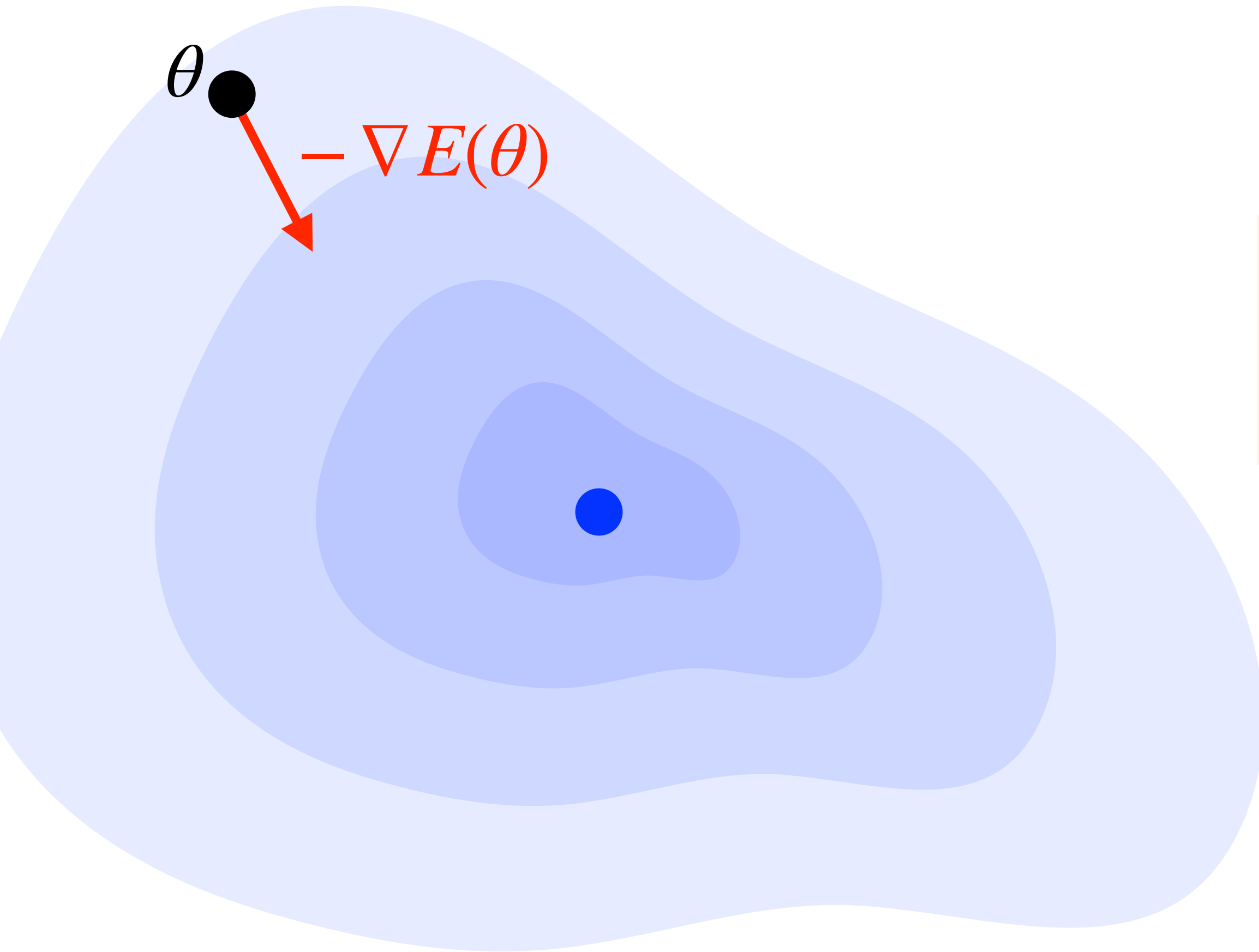
Carl Friedrich
Gauss

Learning using gradient descent

Minimize: $E(\theta) := \sum_i \text{Error}(f_{\theta}(x_i), y_i)$



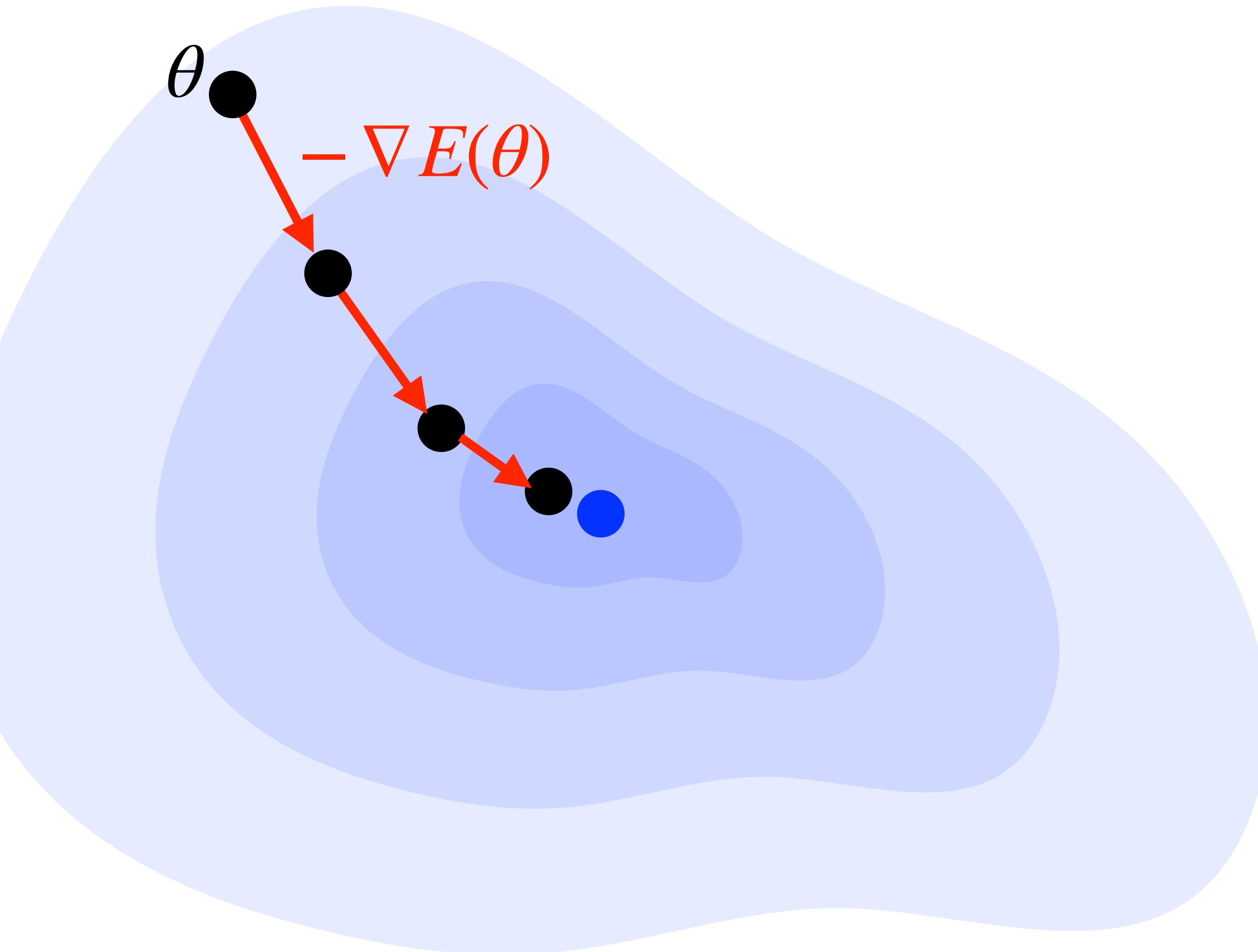
Learning using gradient descent



Minimize: $E(\theta) := \sum_i \text{Error}(f_\theta(x_i), y_i)$

Gradient: $\nabla E(\theta) = \left(\frac{\partial E}{\partial \theta_1}(\theta), \frac{\partial E}{\partial \theta_2}(\theta), \dots \right)$

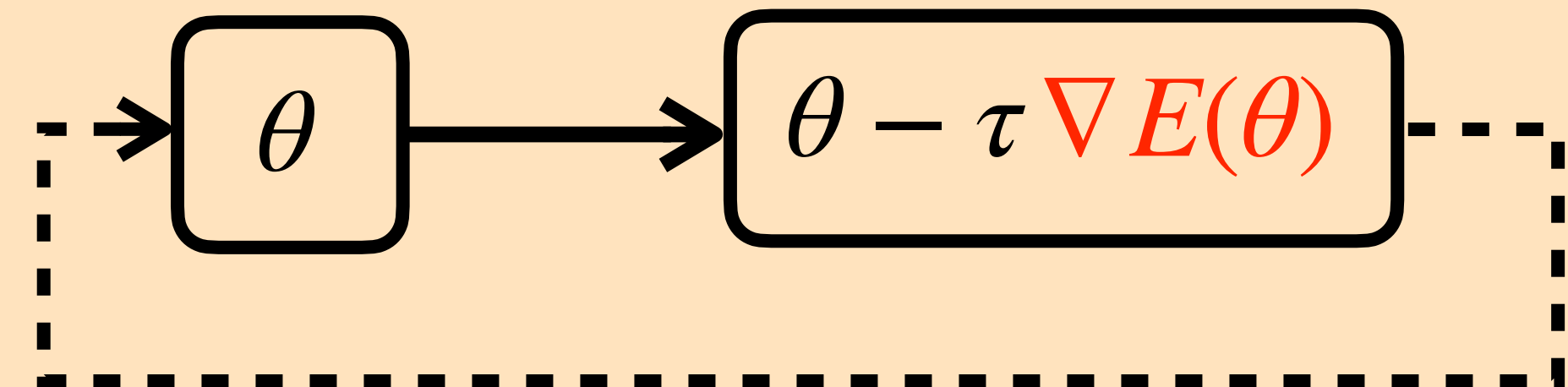
Learning using gradient descent



Minimize: $E(\theta) := \sum_i \text{Error}(f_\theta(x_i), y_i)$

Gradient: $\nabla E(\theta) = \left(\frac{\partial E}{\partial \theta_1}(\theta), \frac{\partial E}{\partial \theta_2}(\theta), \dots \right)$

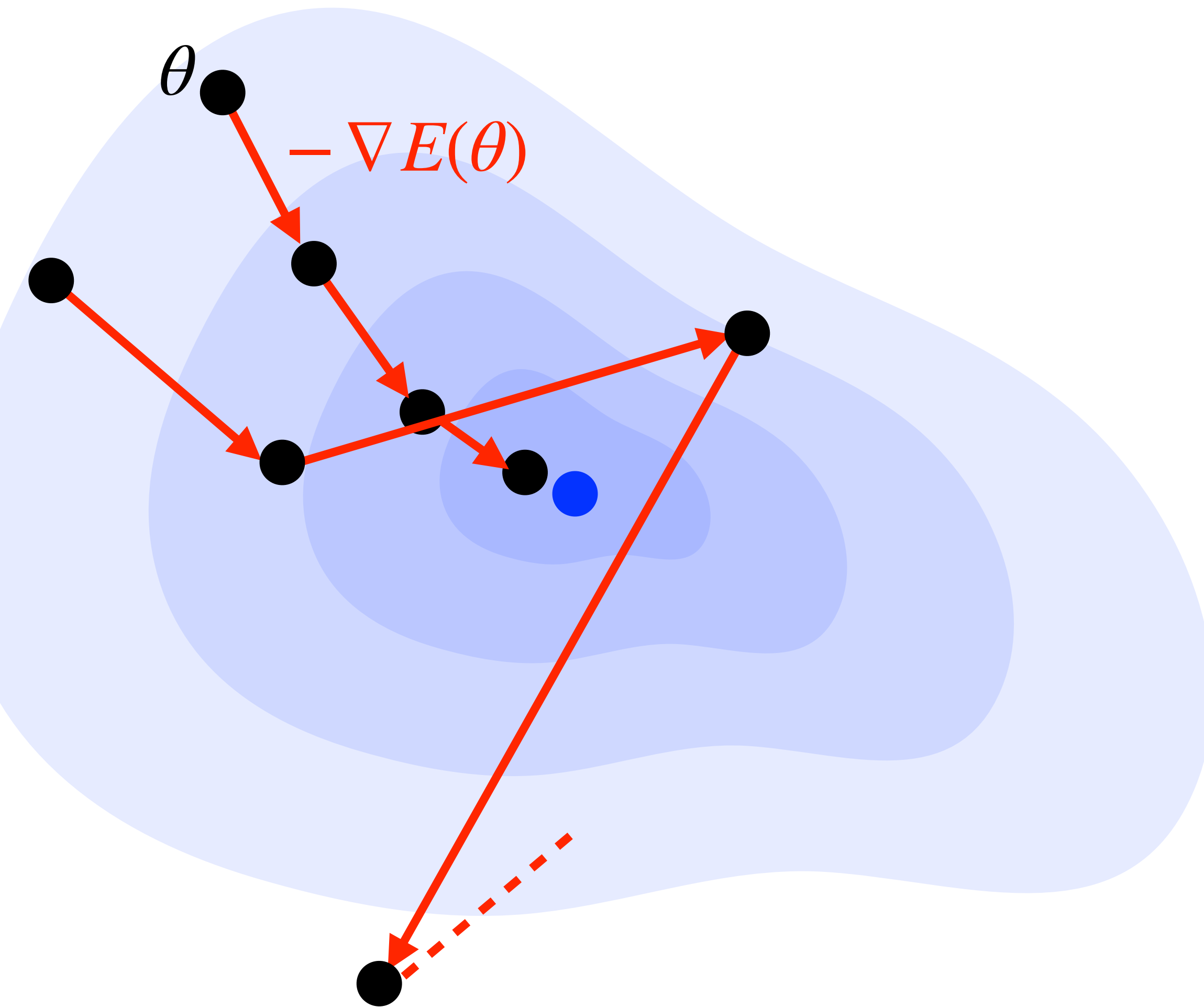
Steepest descent:



Stochastic



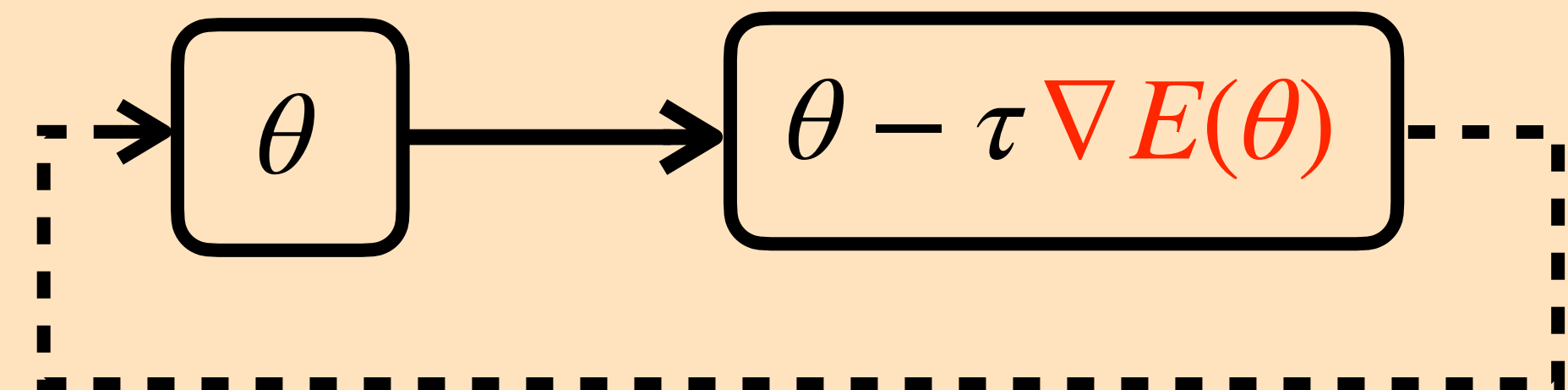
Learning using gradient descent



Minimize: $E(\theta) := \sum_i \text{Error}(f_\theta(x_i), y_i)$

Gradient: $\nabla E(\theta) = \left(\frac{\partial E}{\partial \theta_1}(\theta), \frac{\partial E}{\partial \theta_2}(\theta), \dots \right)$

Steepest descent:



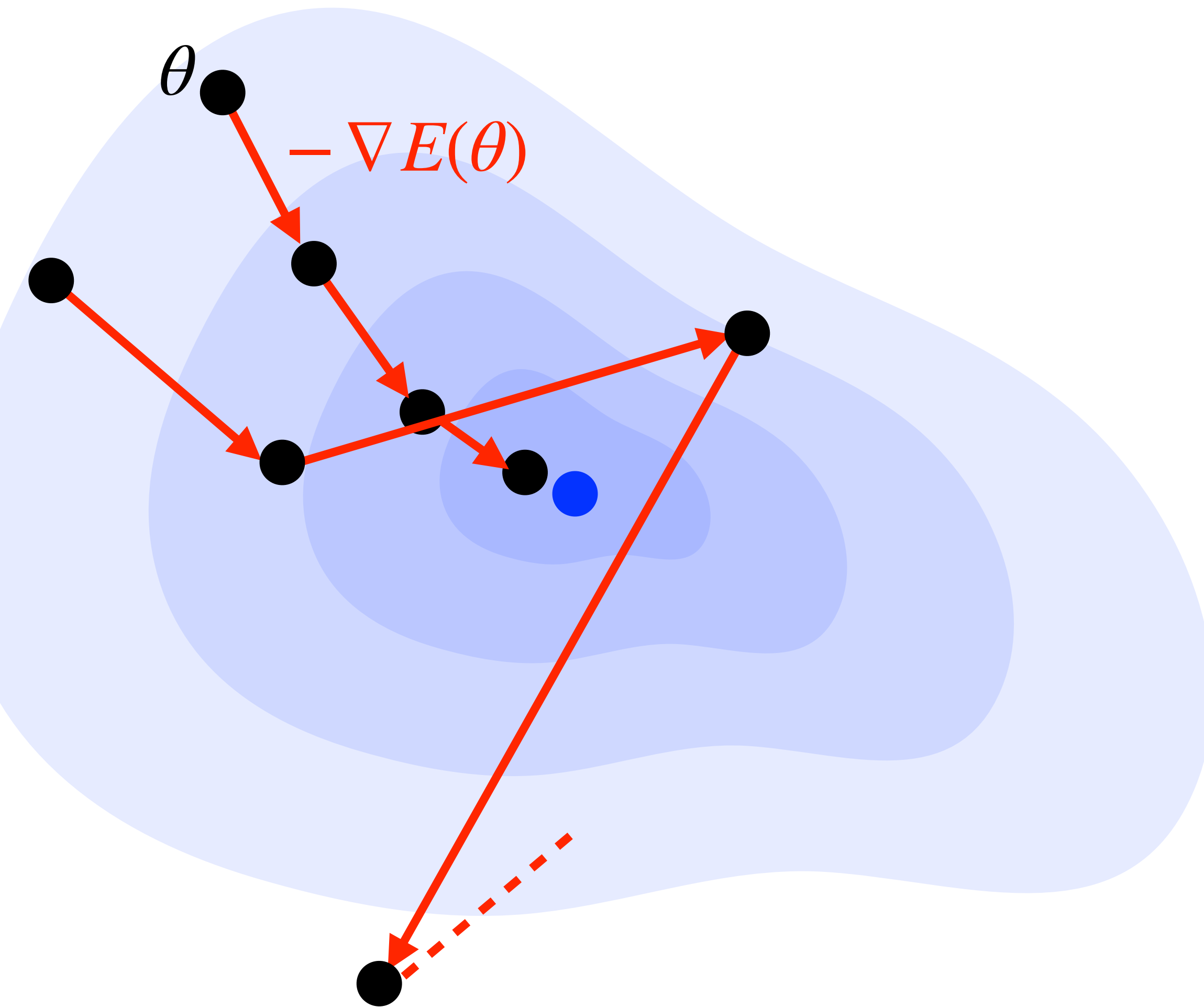
Open problems

Step size τ selection

Stochastic



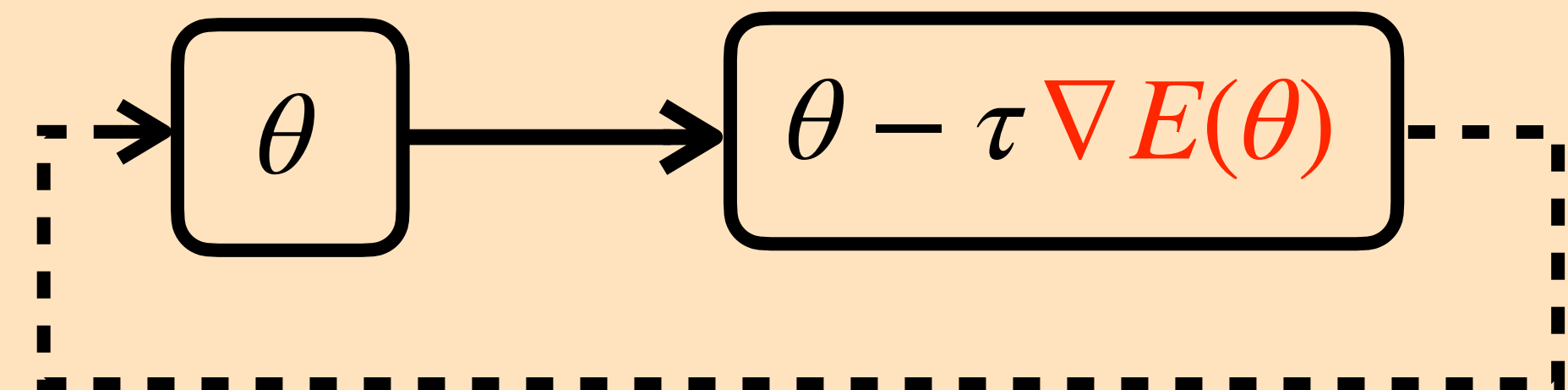
Learning using gradient descent



Minimize: $E(\theta) := \sum_i \text{Error}(f_\theta(x_i), y_i)$

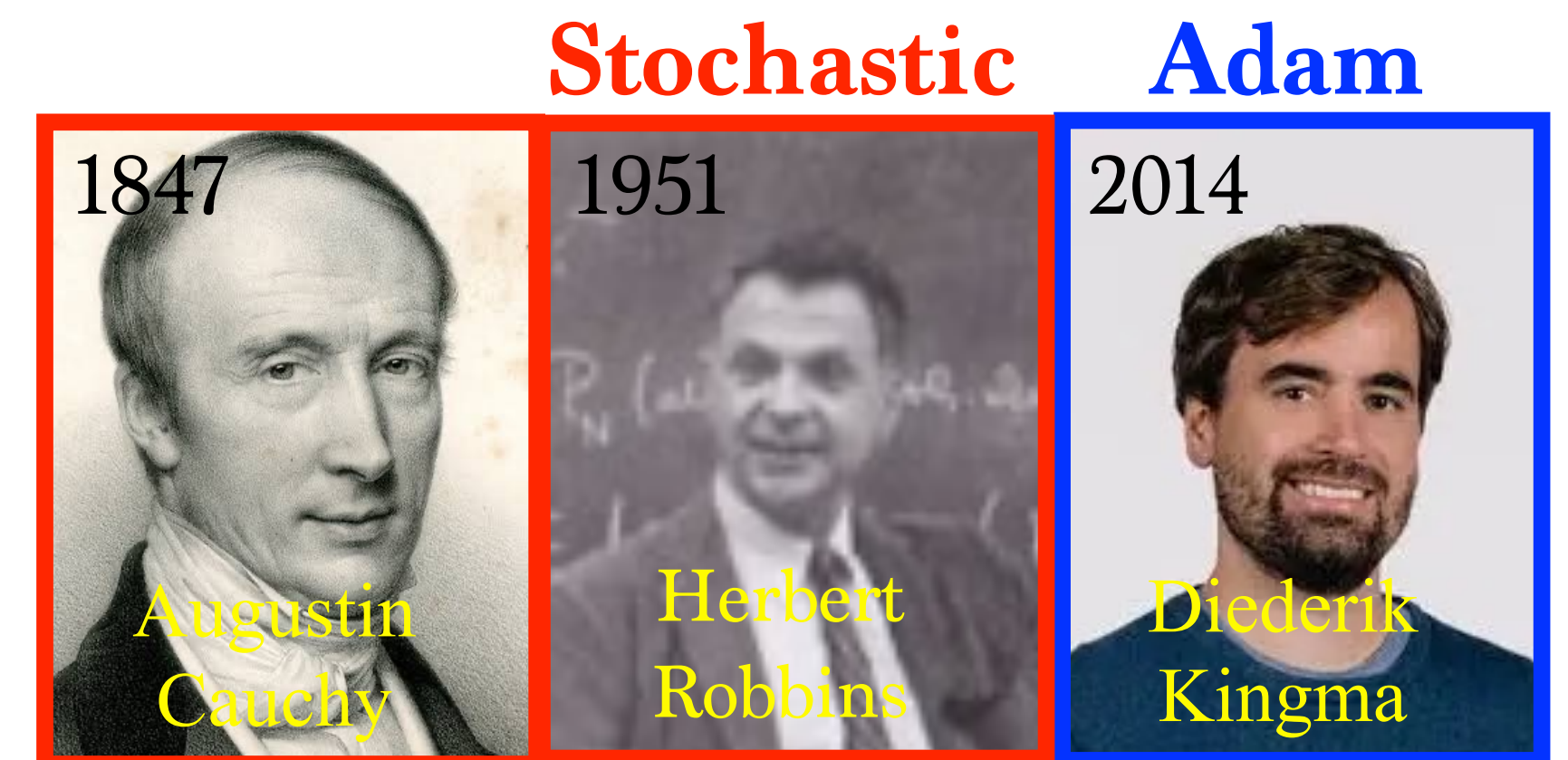
Gradient: $\nabla E(\theta) = \left(\frac{\partial E}{\partial \theta_1}(\theta), \frac{\partial E}{\partial \theta_2}(\theta), \dots \right)$

Steepest descent:



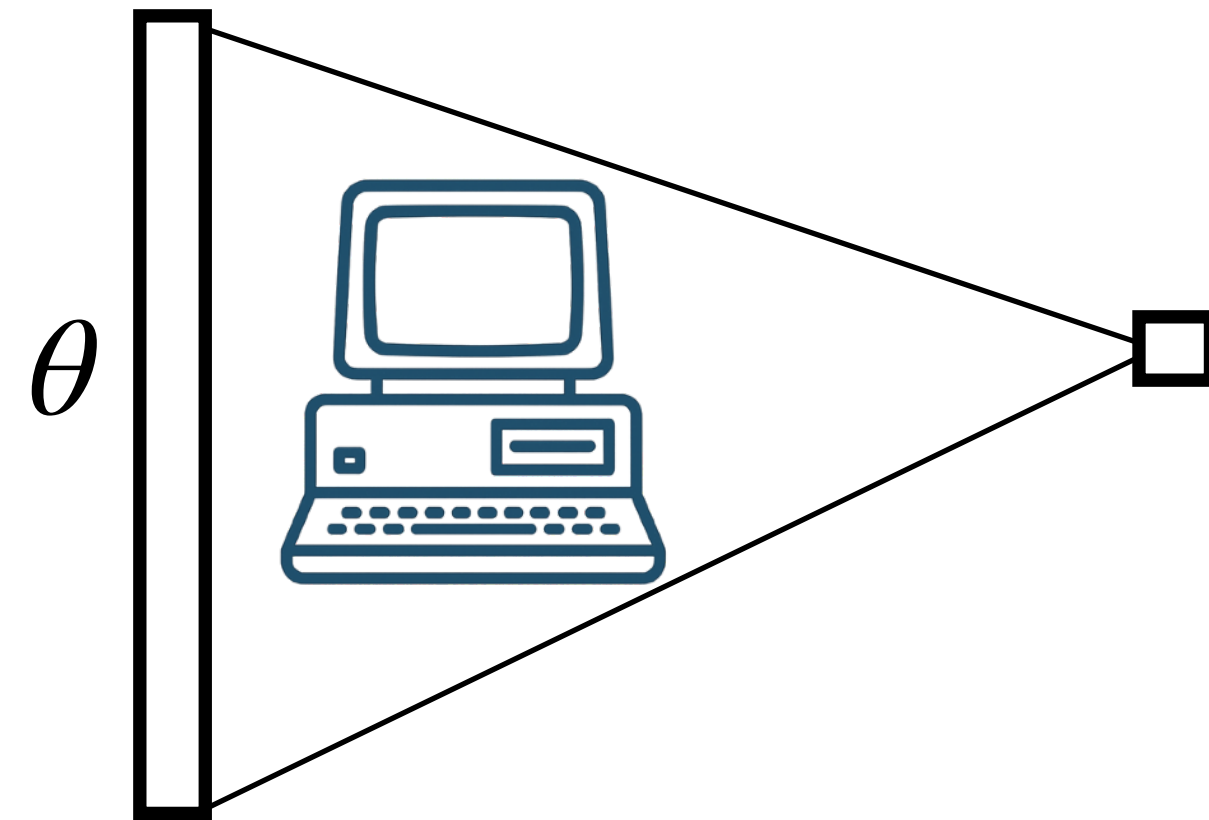
Open problems

- Step size τ selection
- Understanding Adam



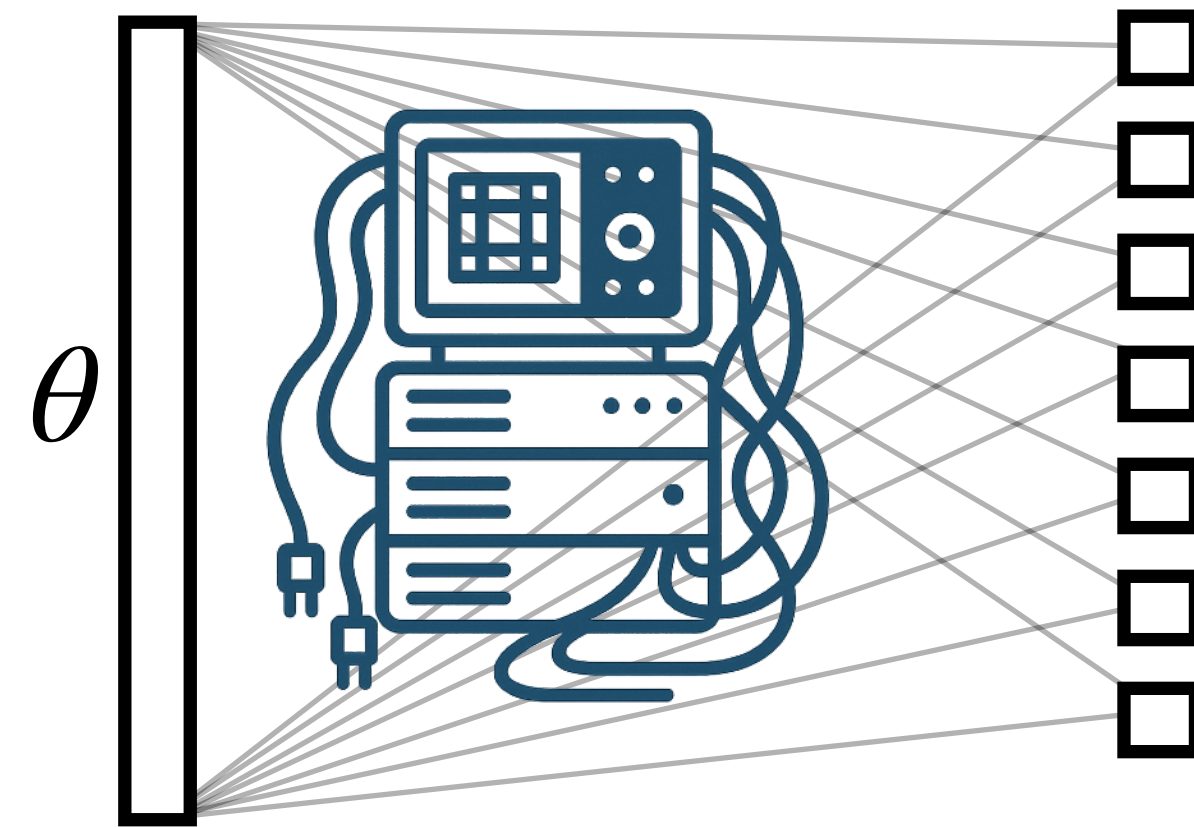
The (non)-Complexity of Gradient Computation

Energy:



$$E(\theta) := \sum_i \text{Error}(f_{\theta}(x_i), y_i)$$

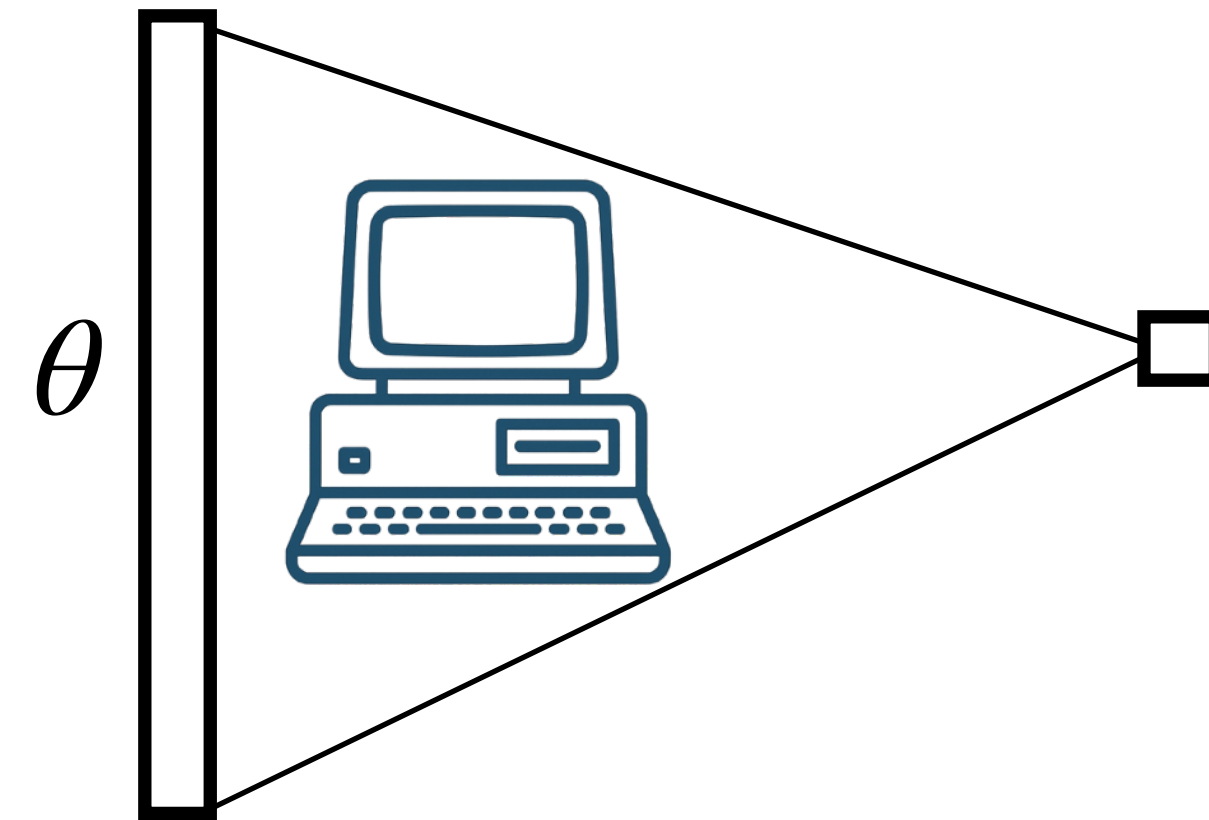
Gradient:



$$\nabla E(\theta) := \left(\frac{\partial E}{\partial \theta_1}(\theta), \frac{\partial E}{\partial \theta_2}(\theta), \dots \right)$$

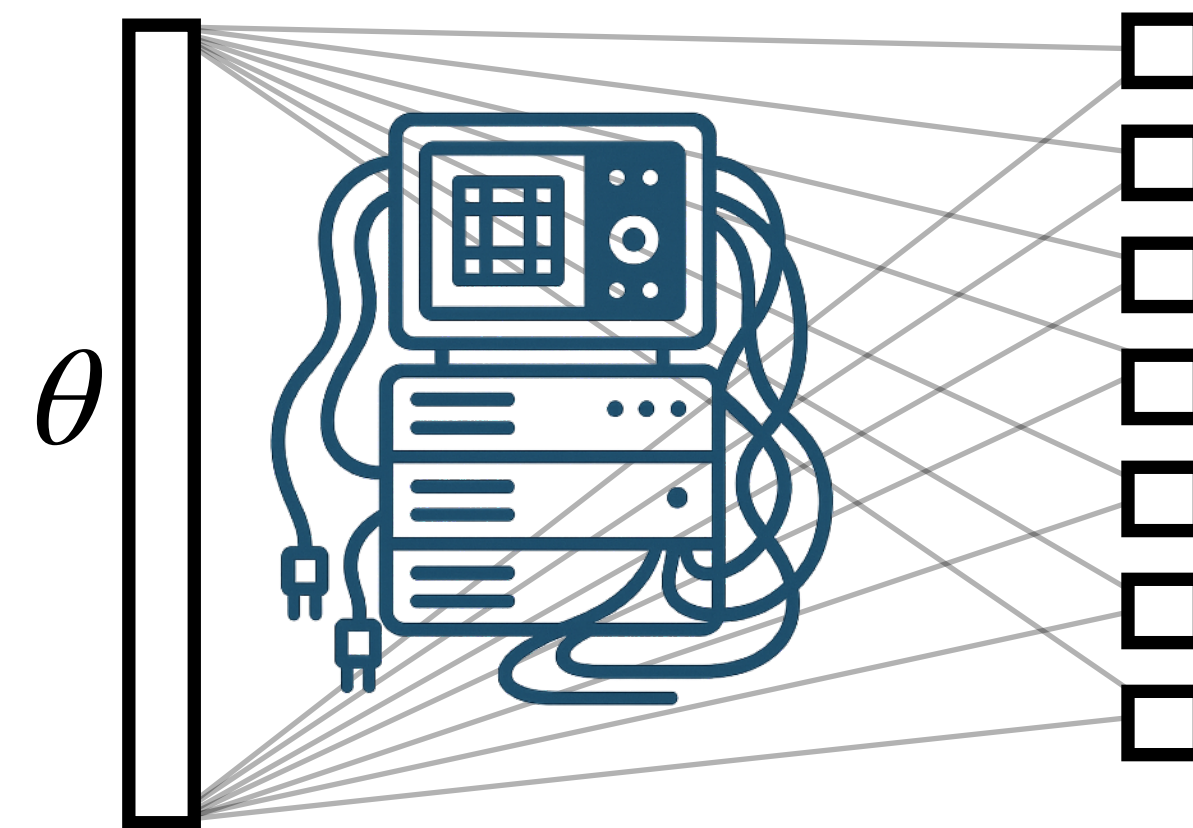
The (non)-Complexity of Gradient Computation

Energy:



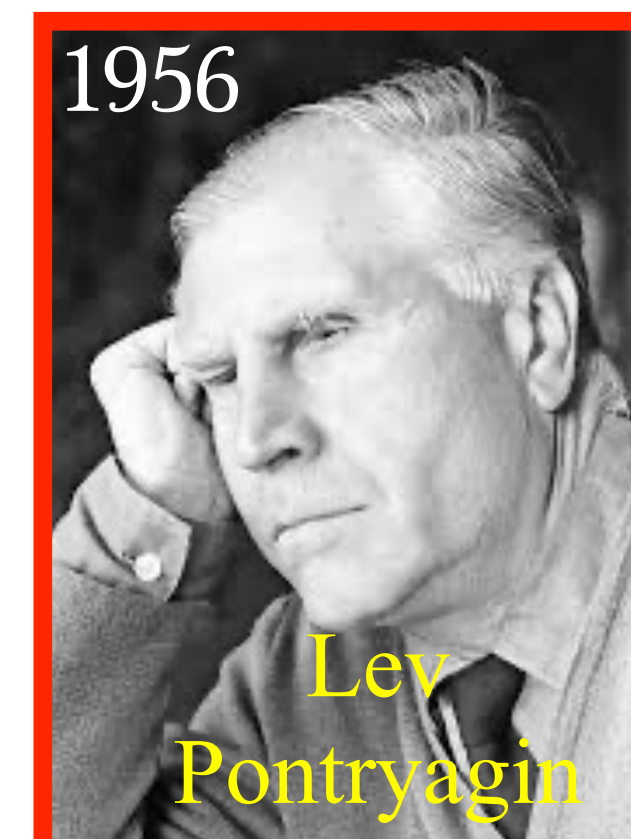
$$E(\theta) := \sum_i \text{Error}(f_{\theta}(x_i), y_i)$$

Gradient:



$$\nabla E(\theta) := \left(\frac{\partial E}{\partial \theta_1}(\theta), \frac{\partial E}{\partial \theta_2}(\theta), \dots \right)$$

Theorem: $\nabla E(\theta)$ is computed with the same amount of time as $E(\theta)$ by **backpropagation**.



Lev
Pontryagin



Jacques-
Louis Lions

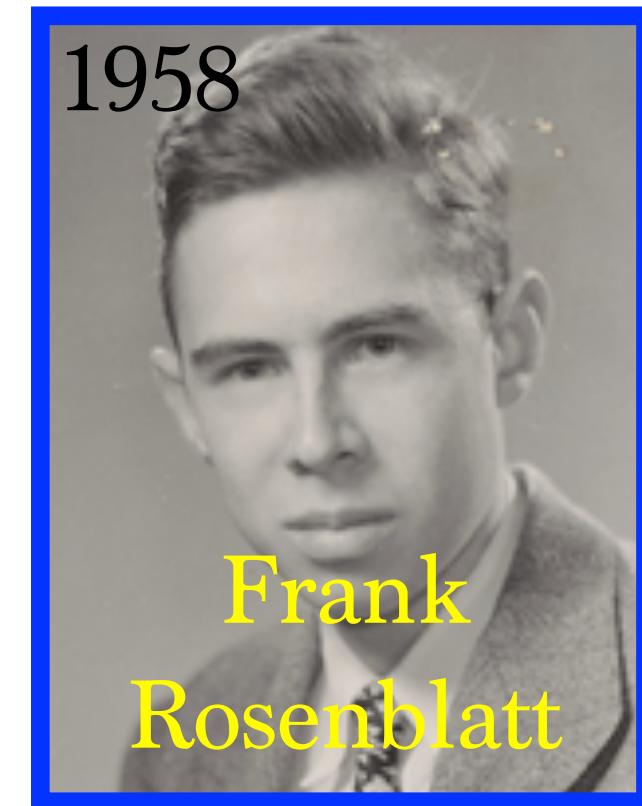
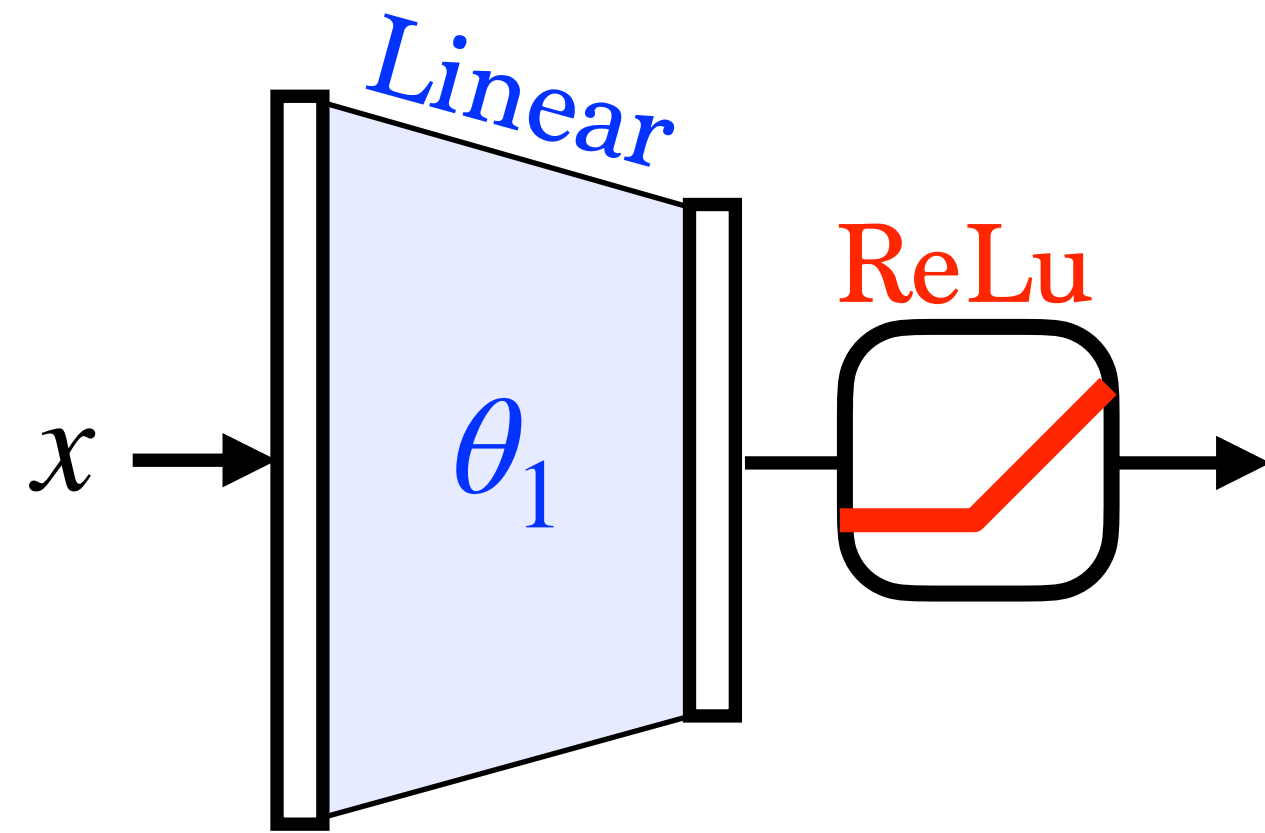


Seppo
Linnainmaa

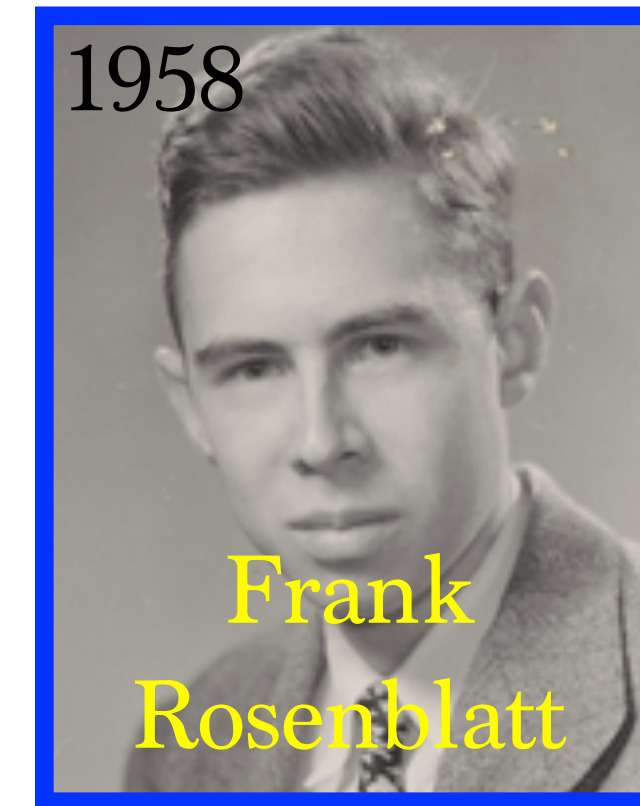
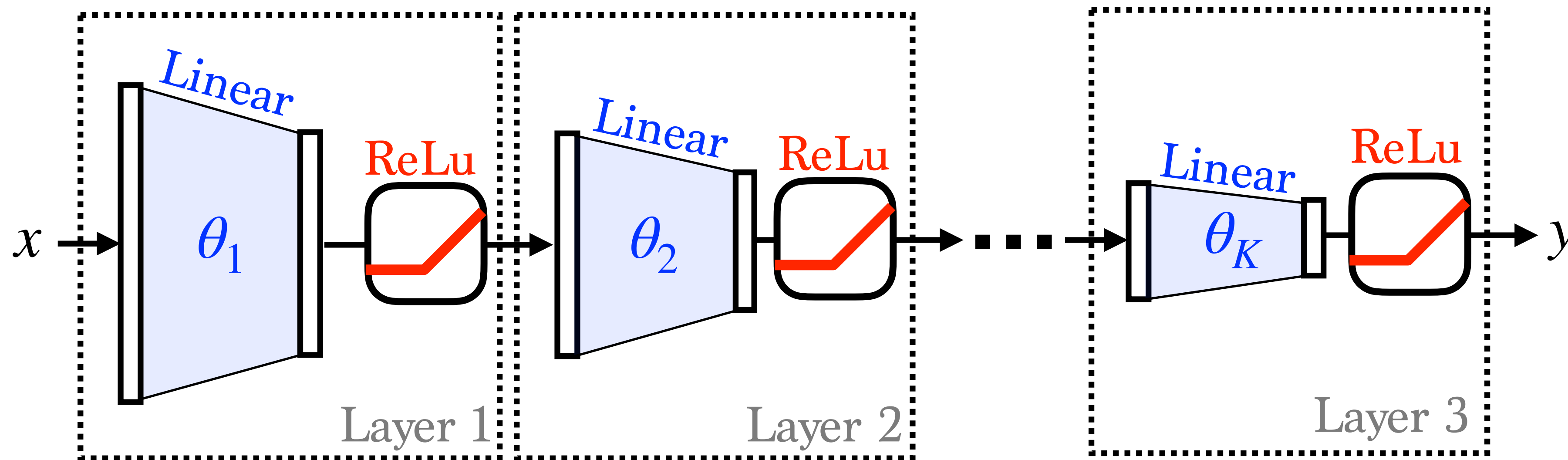
Adjoint method

Back propagation

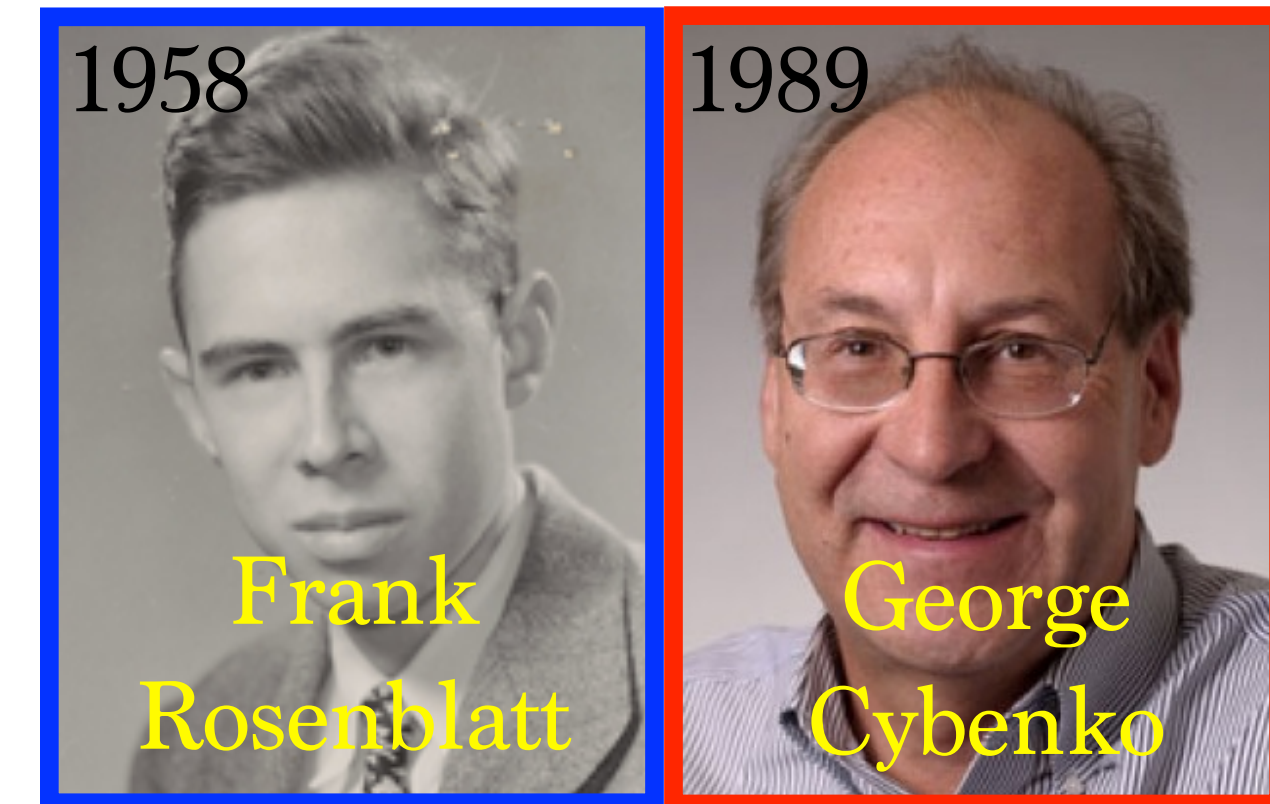
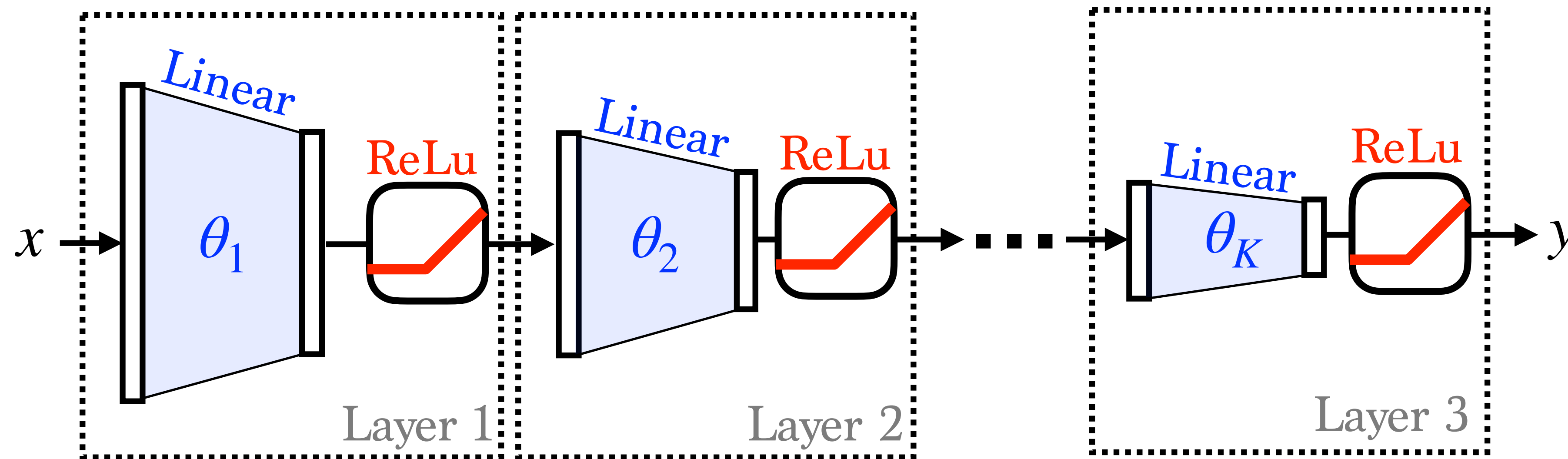
Perceptrons and Universality



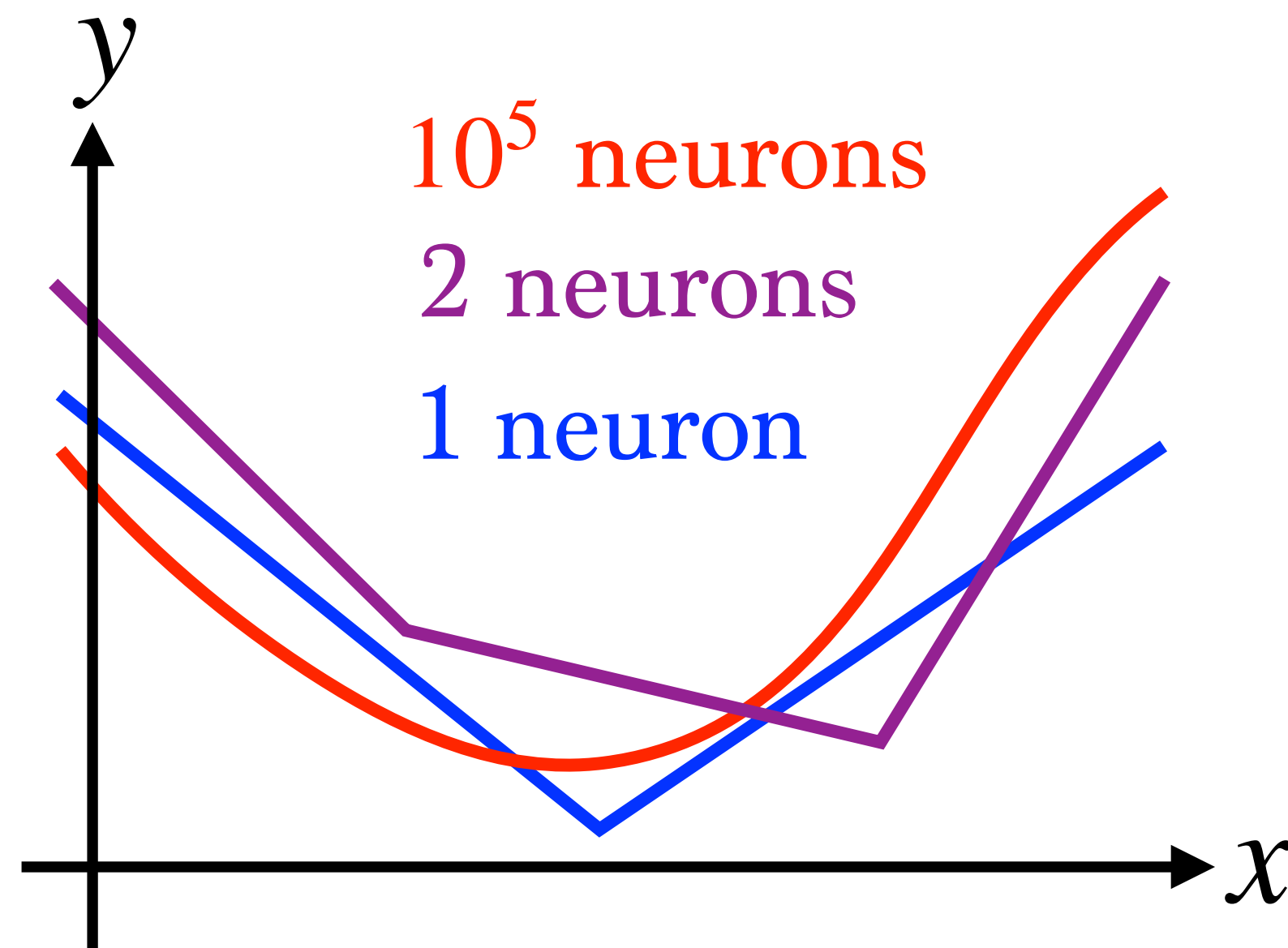
Perceptrons and Universality



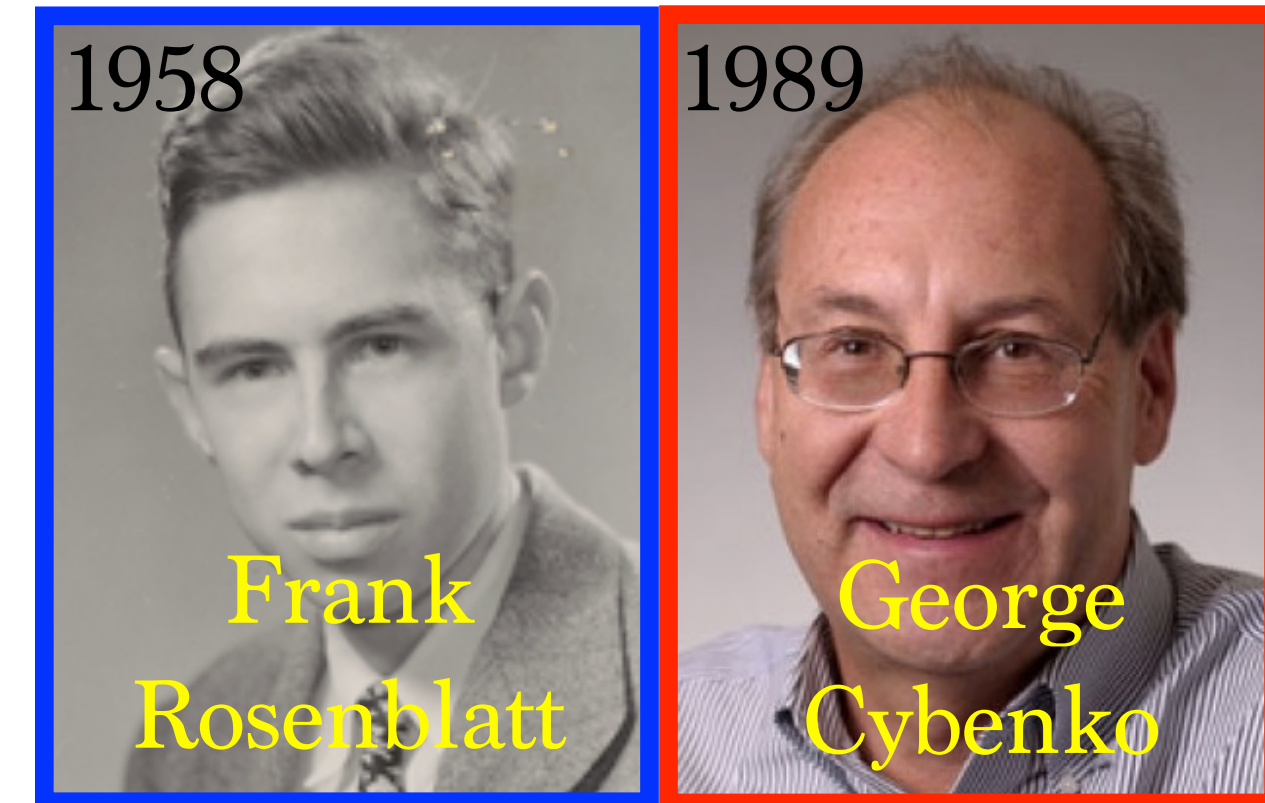
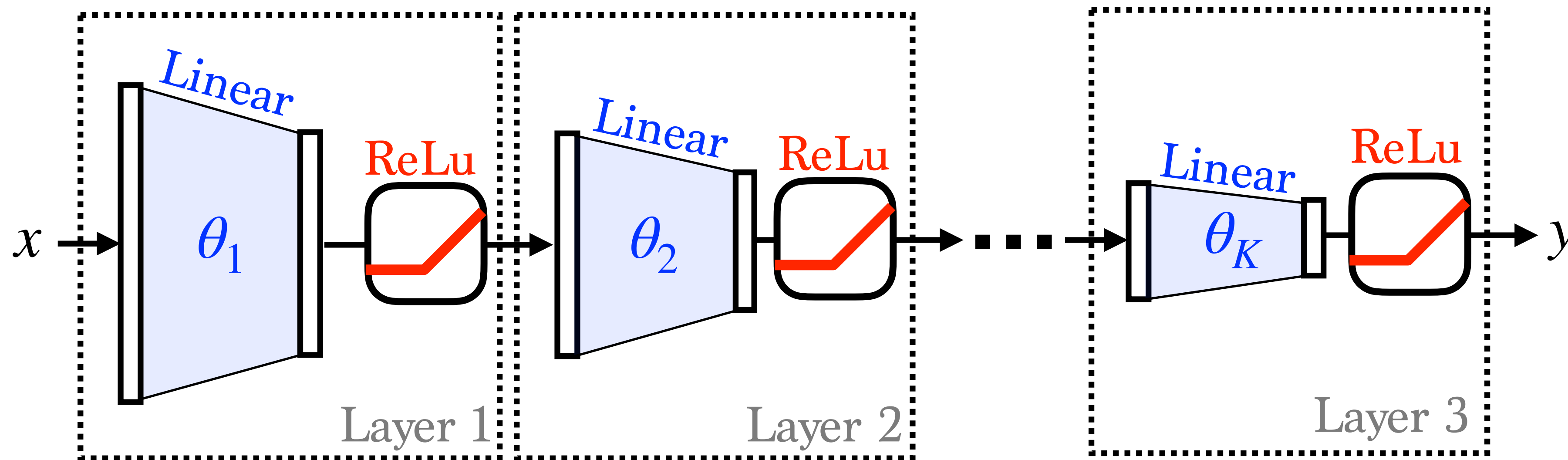
Perceptrons and Universality



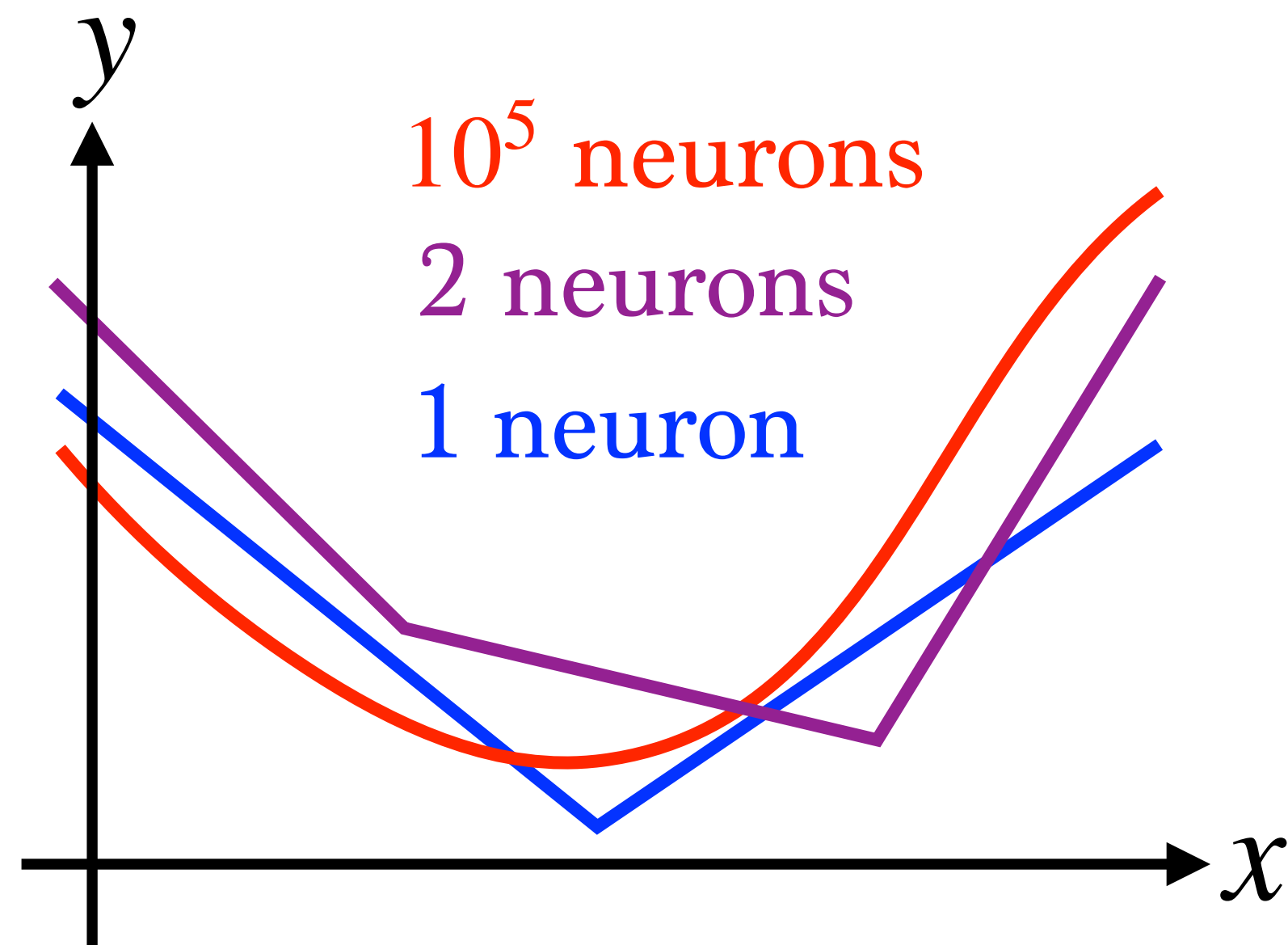
Theorem: $K = 2$ layers and enough neurons can approximate any continuous function.



Perceptrons and Universality



Theorem: $K = 2$ layers and enough neurons can approximate any continuous function.



Open problems

Role of depth?

Convergence of gradient descent



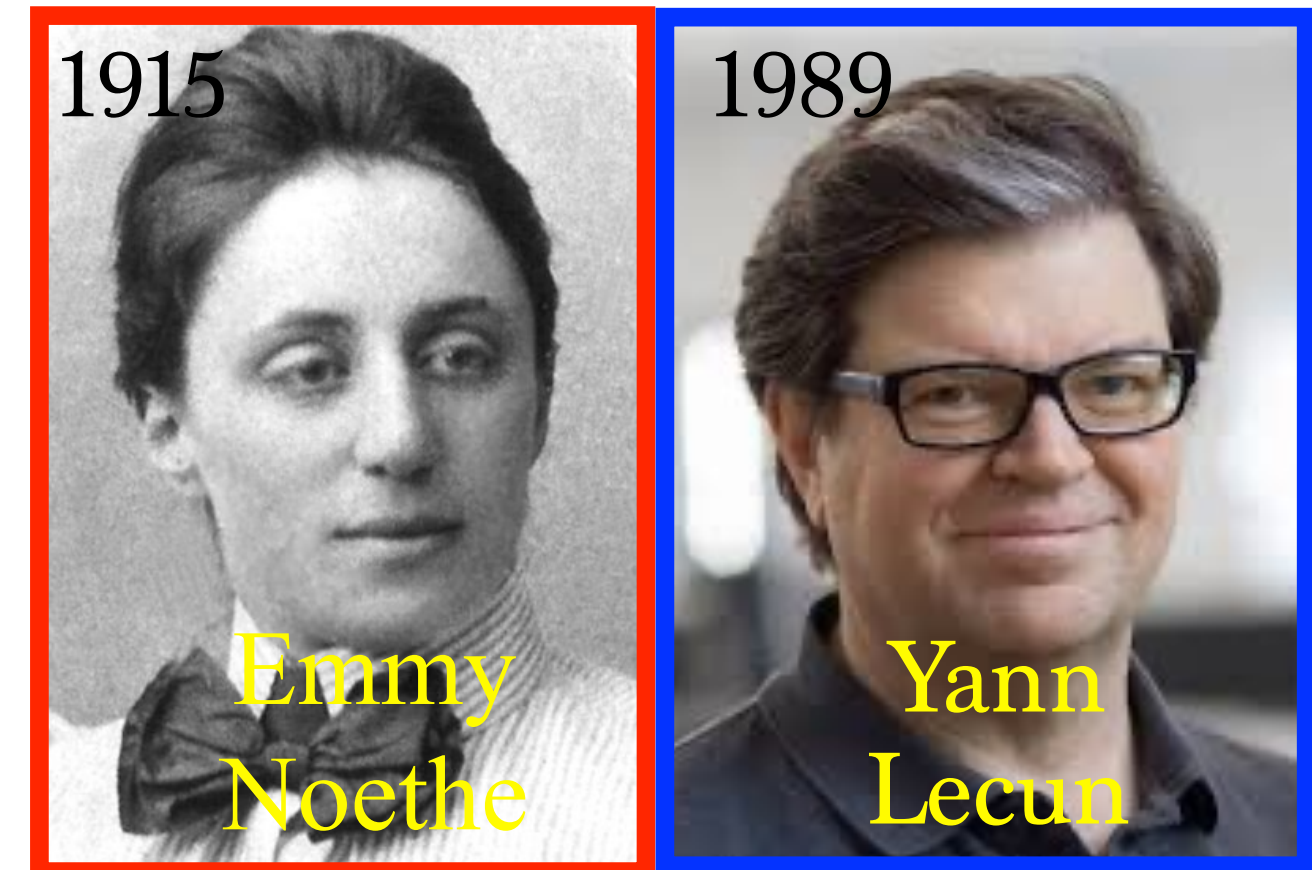
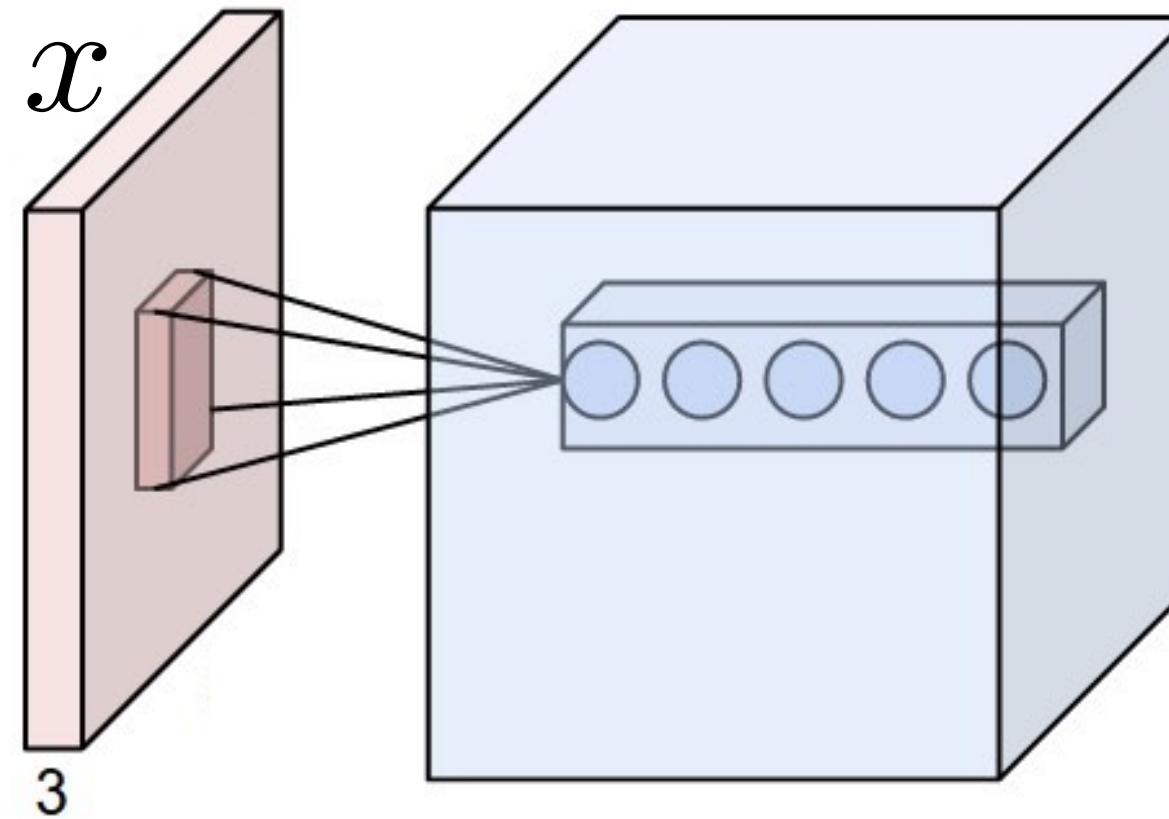
Invariances and Convolutional Networks

Invariance of data



weight sharing

Images \rightarrow convolution



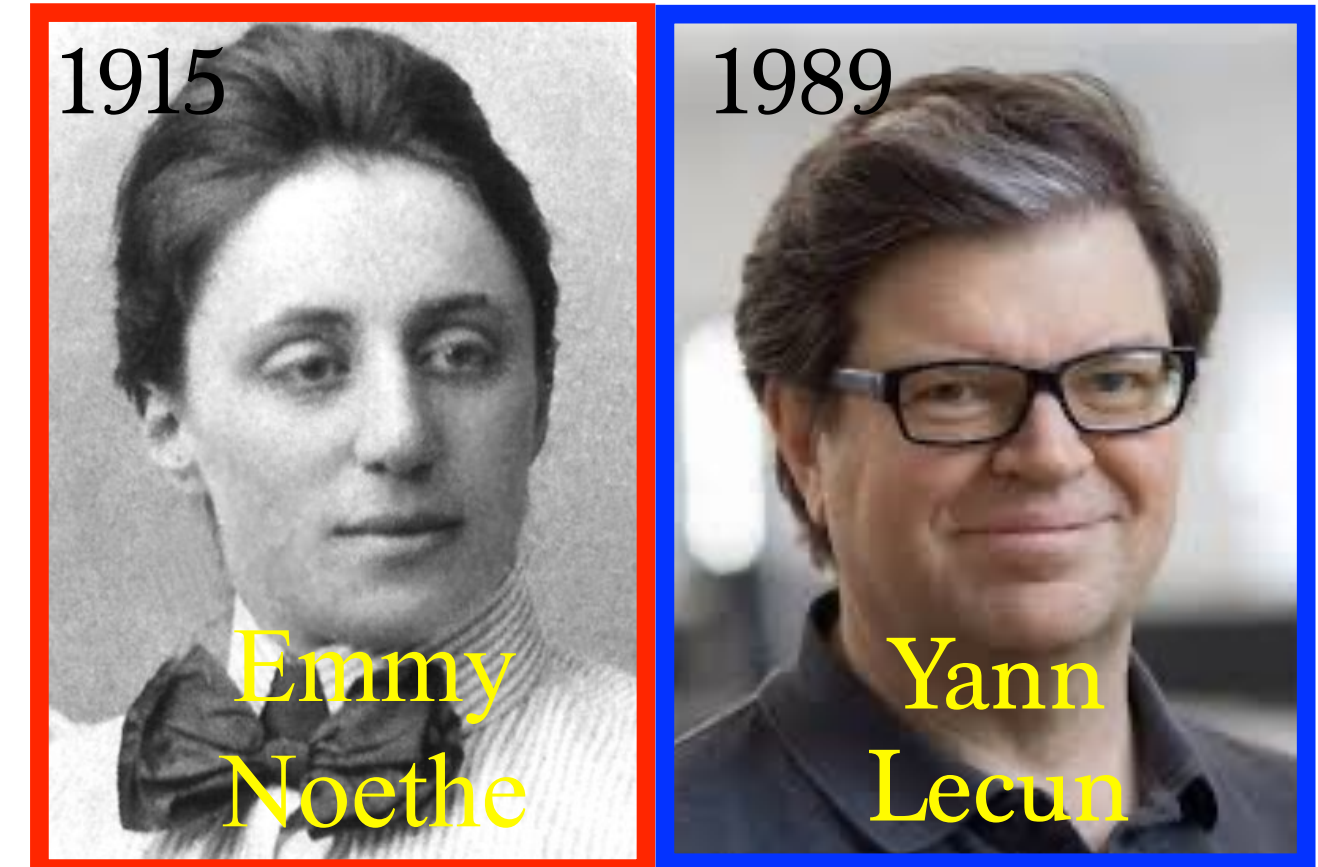
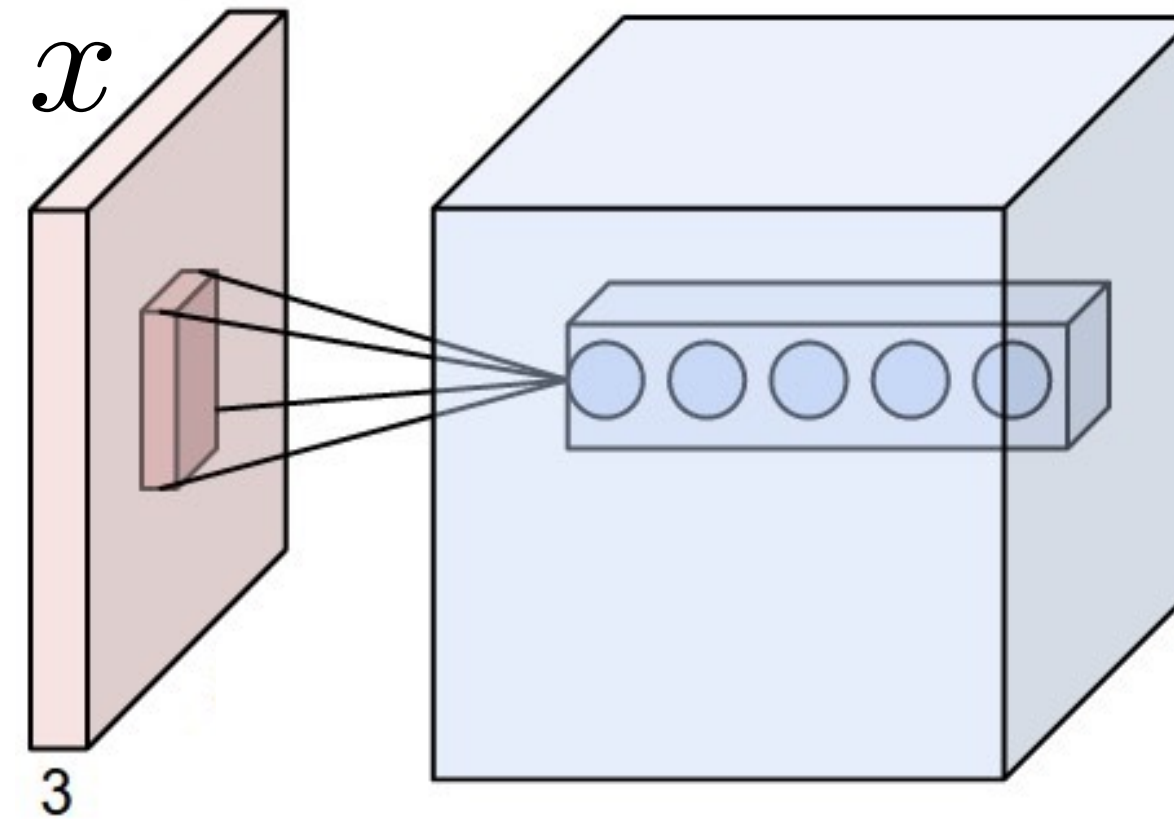
Invariances and Convolutional Networks

Invariance of data

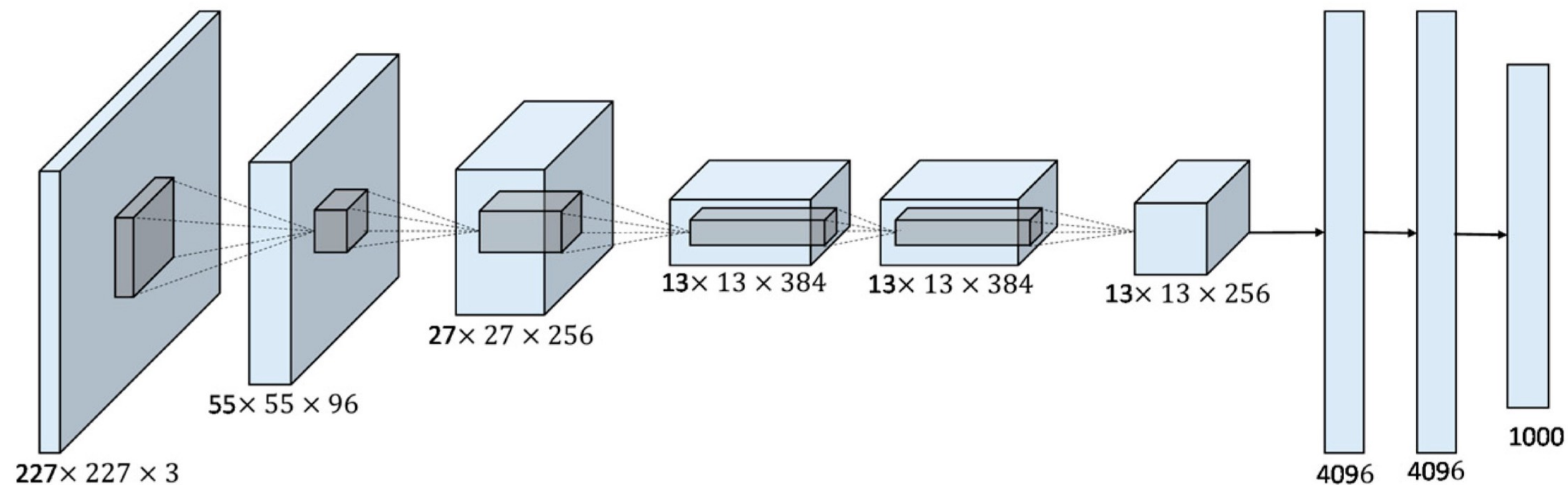


weight sharing

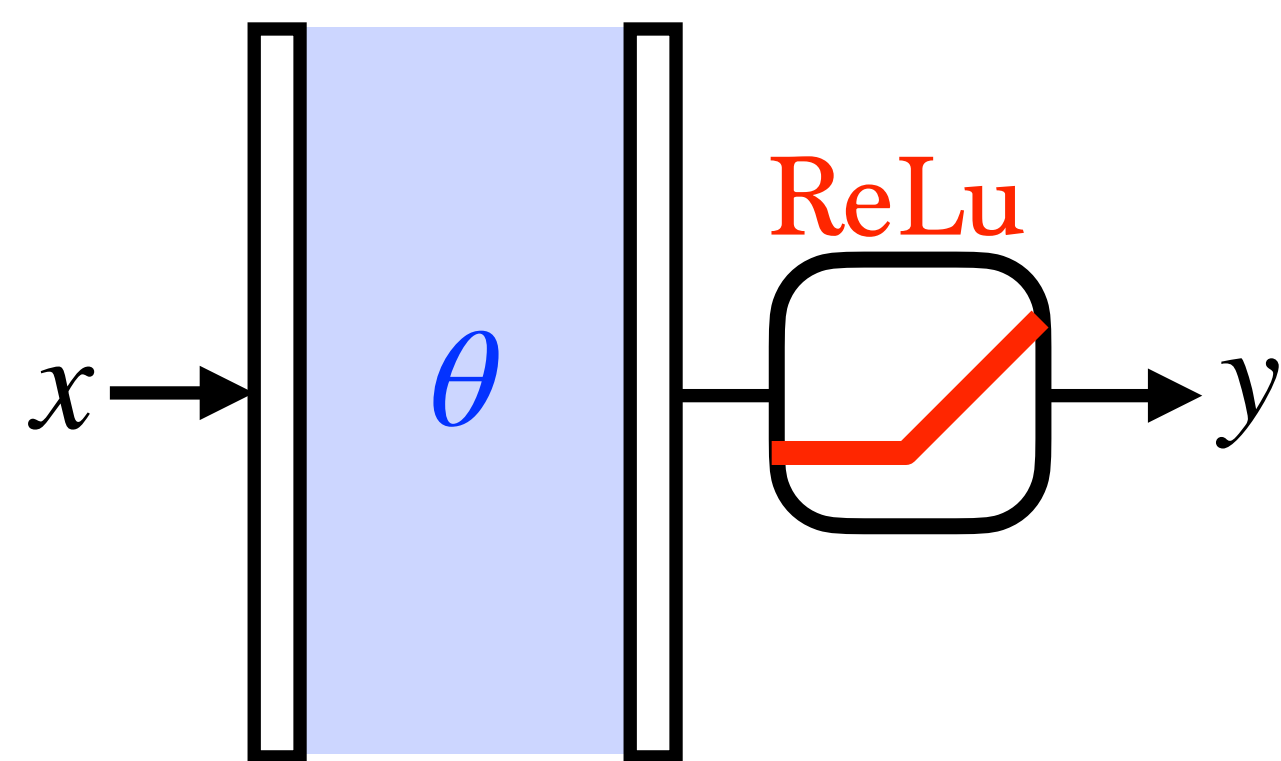
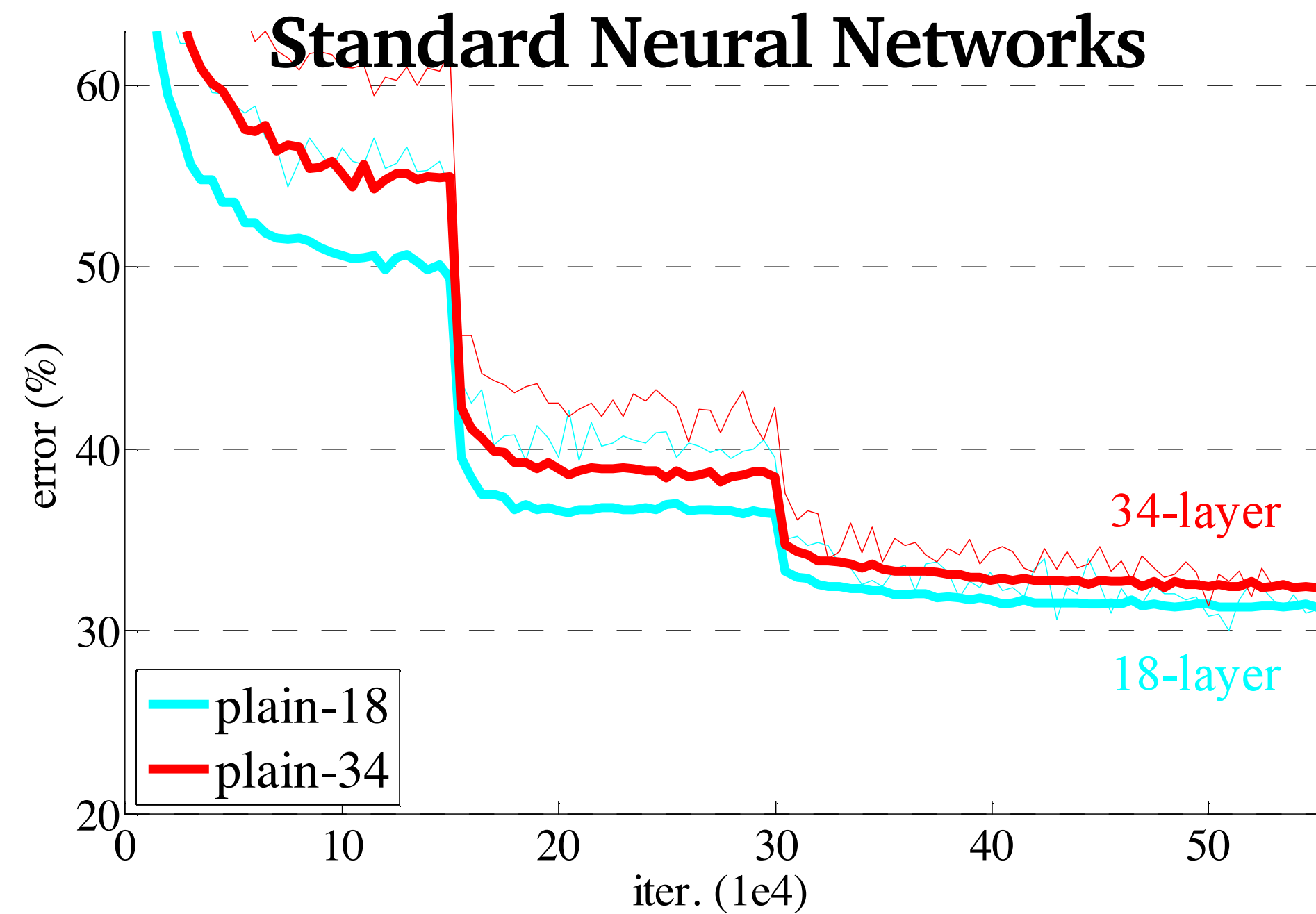
Images \rightarrow convolution



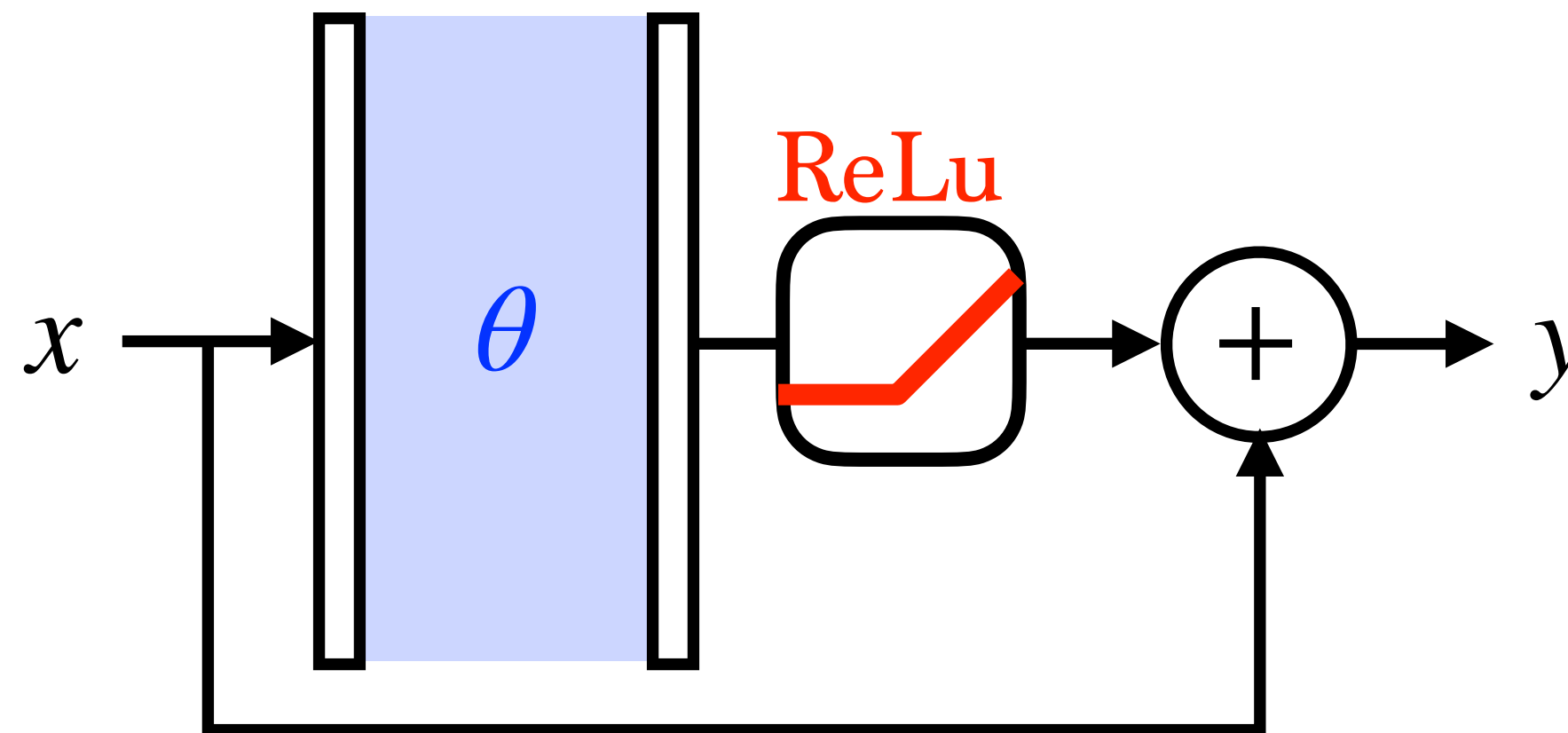
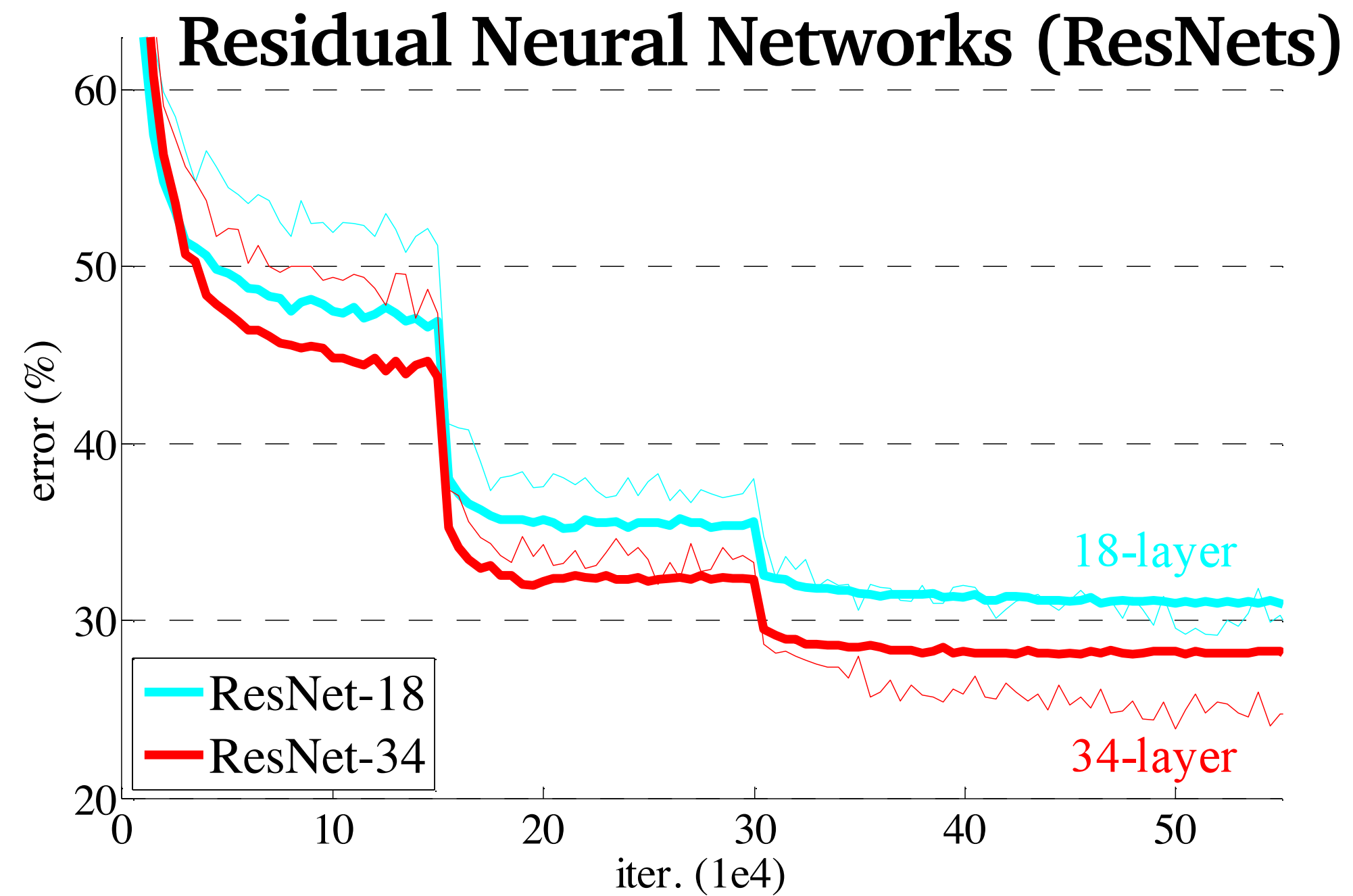
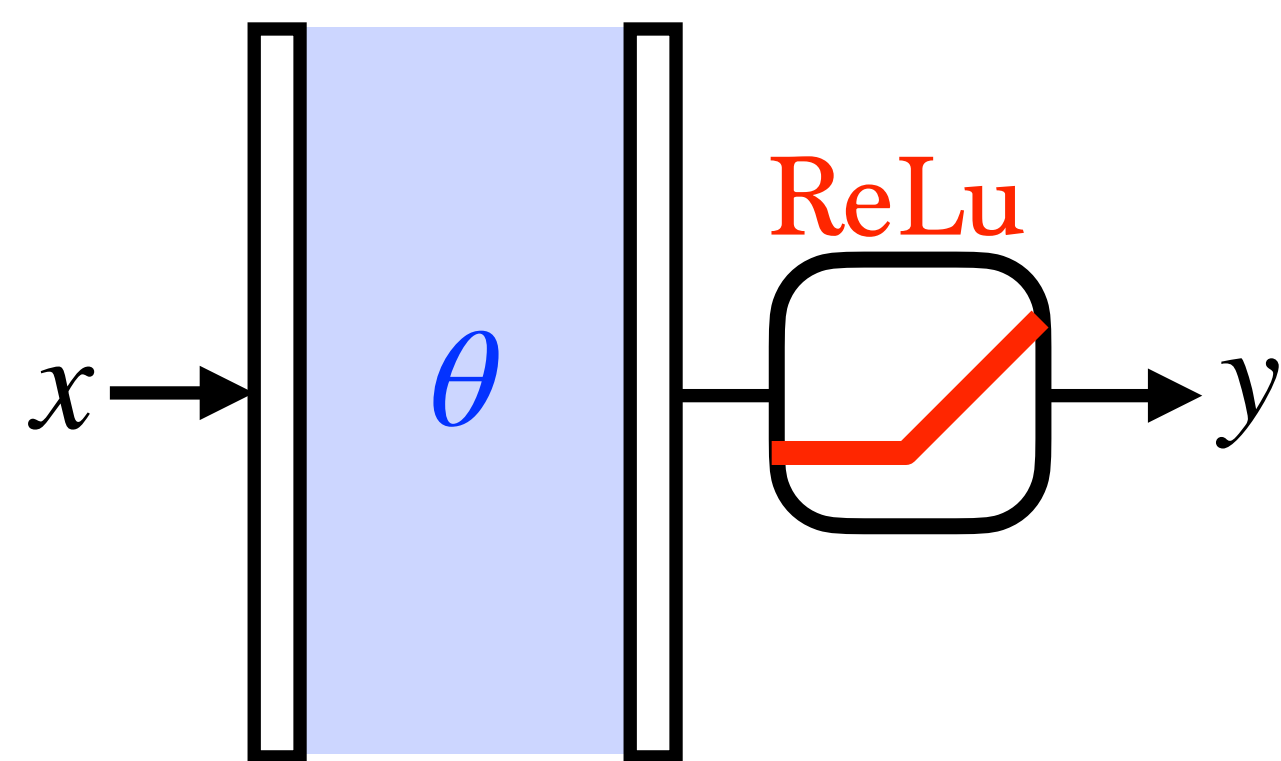
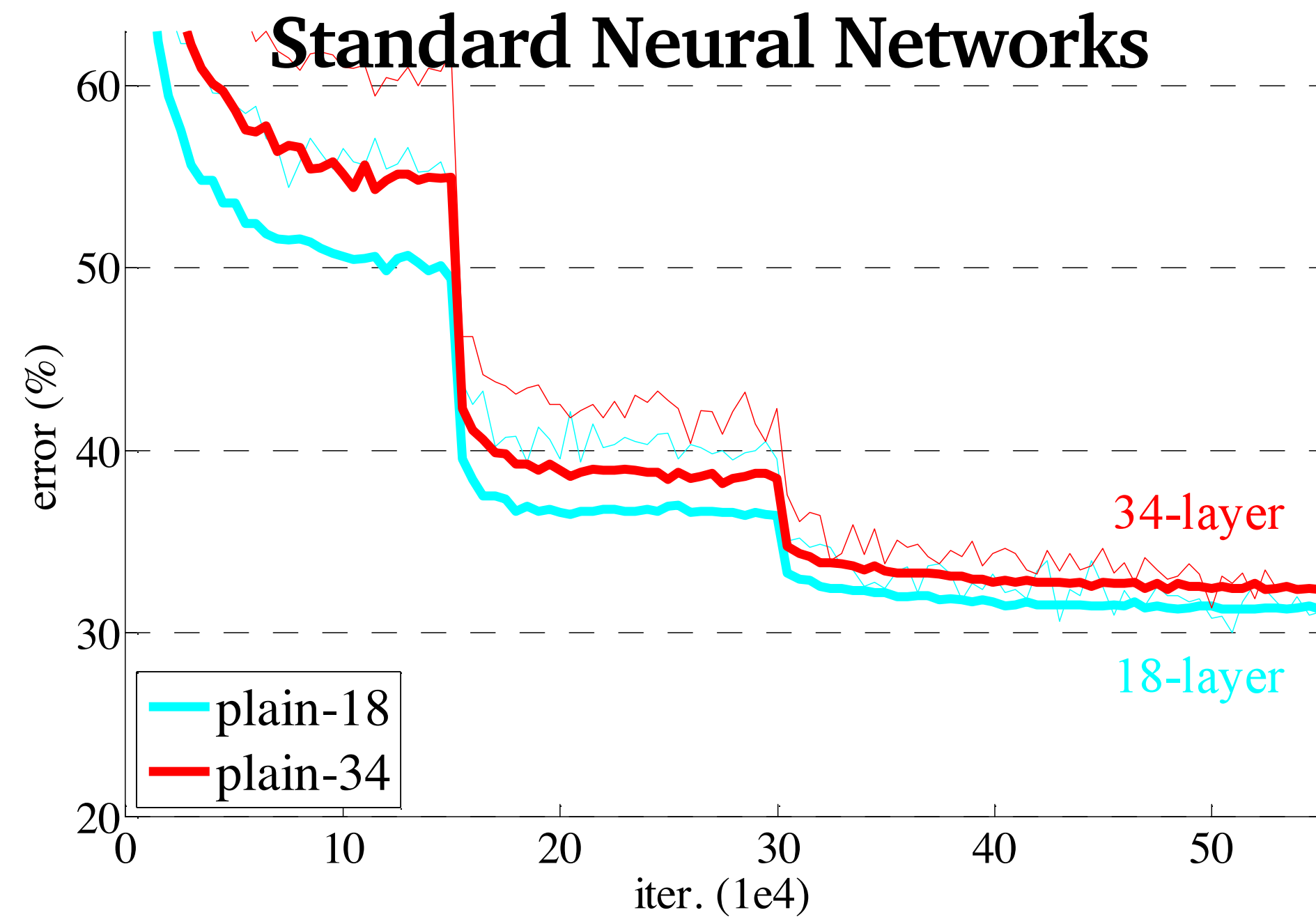
AlexNet



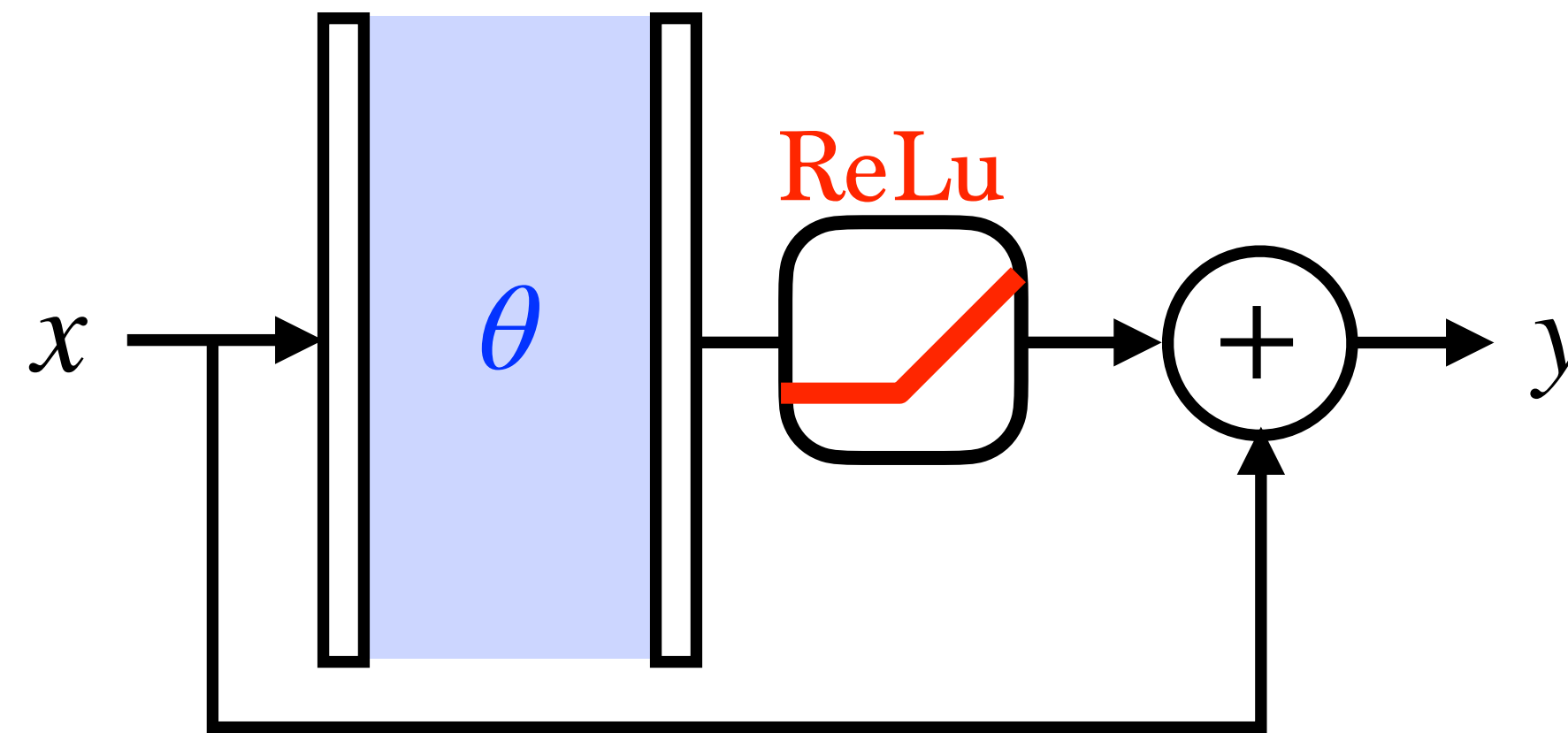
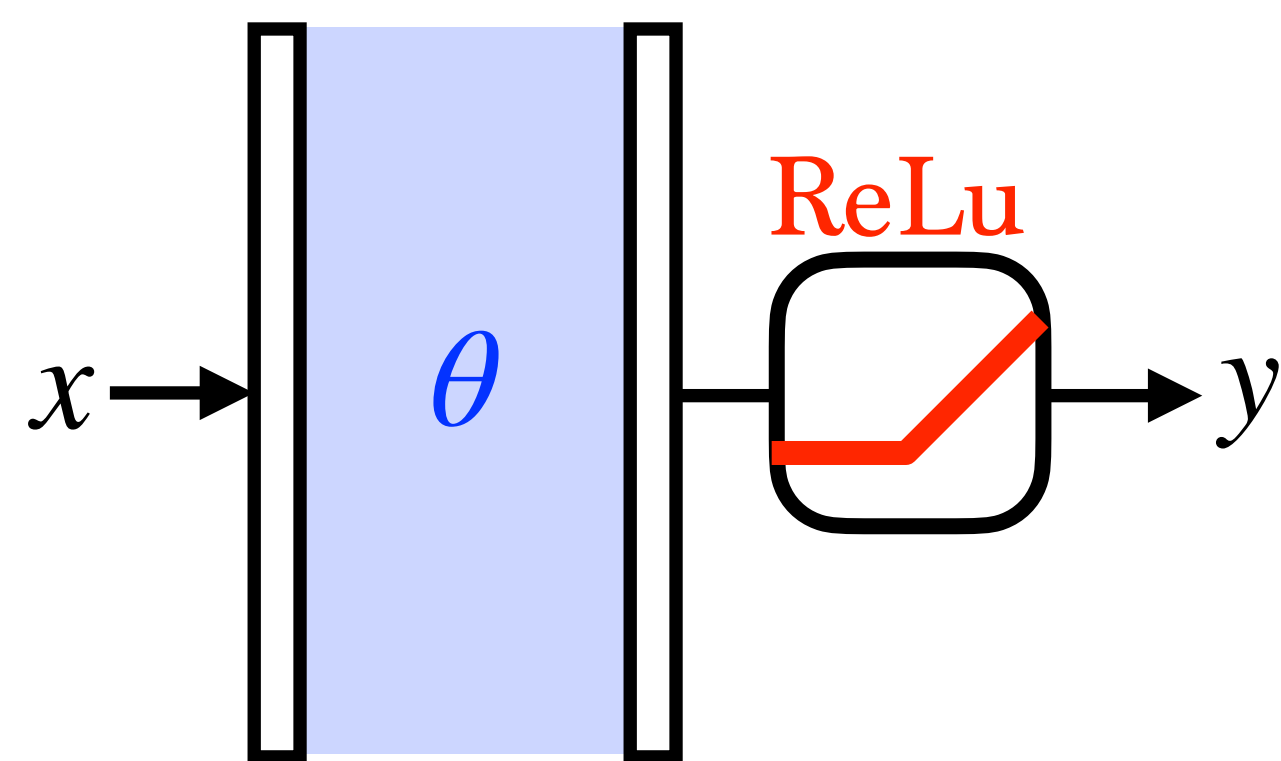
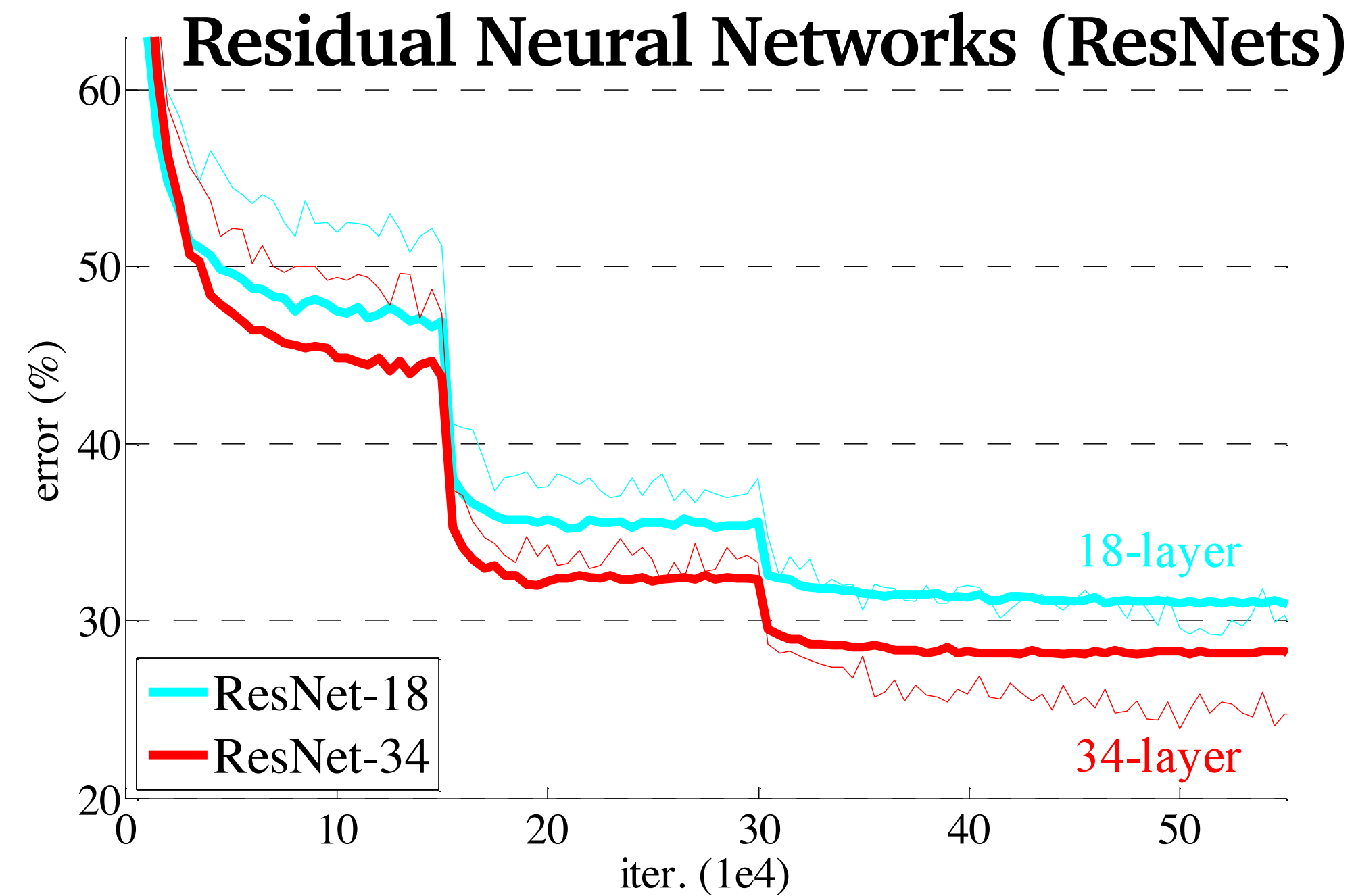
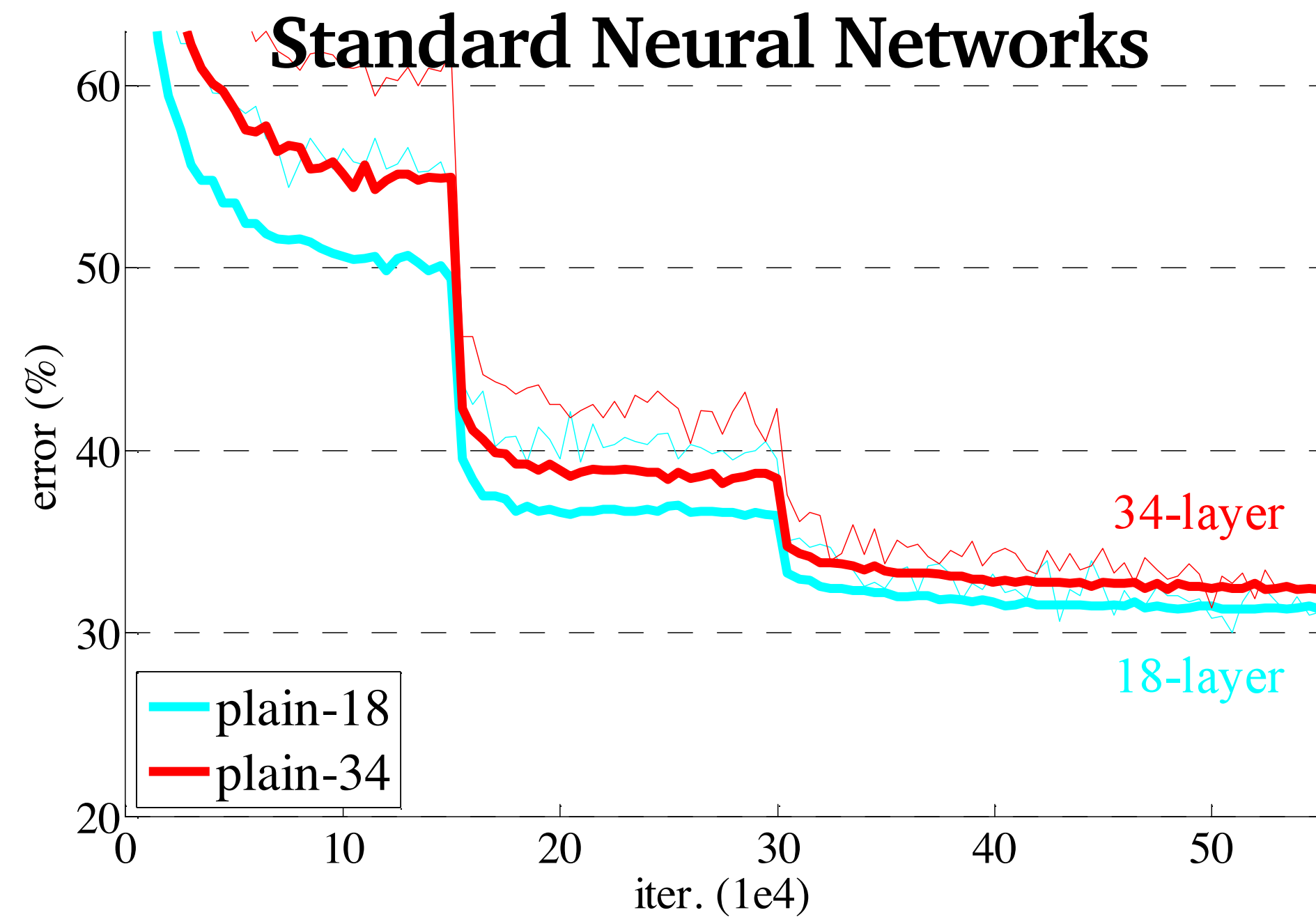
The deeper, the better



The deeper, the better



The deeper, the better



Open problems

Convergence of gradient descent for very deep ResNet.

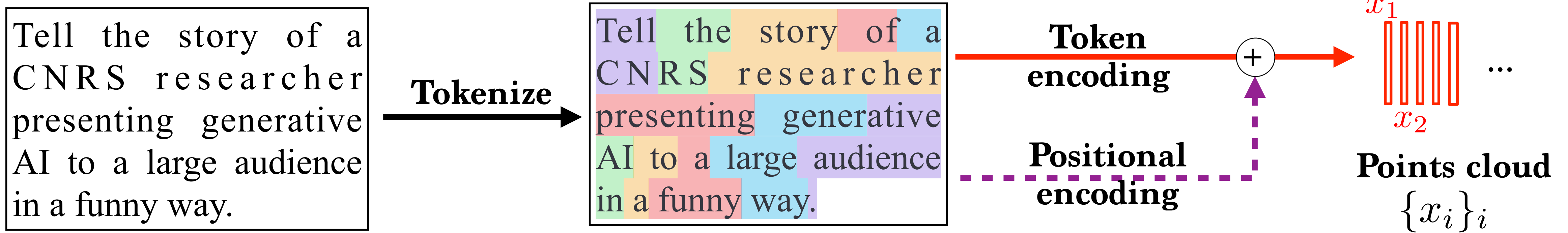


Infinite depth

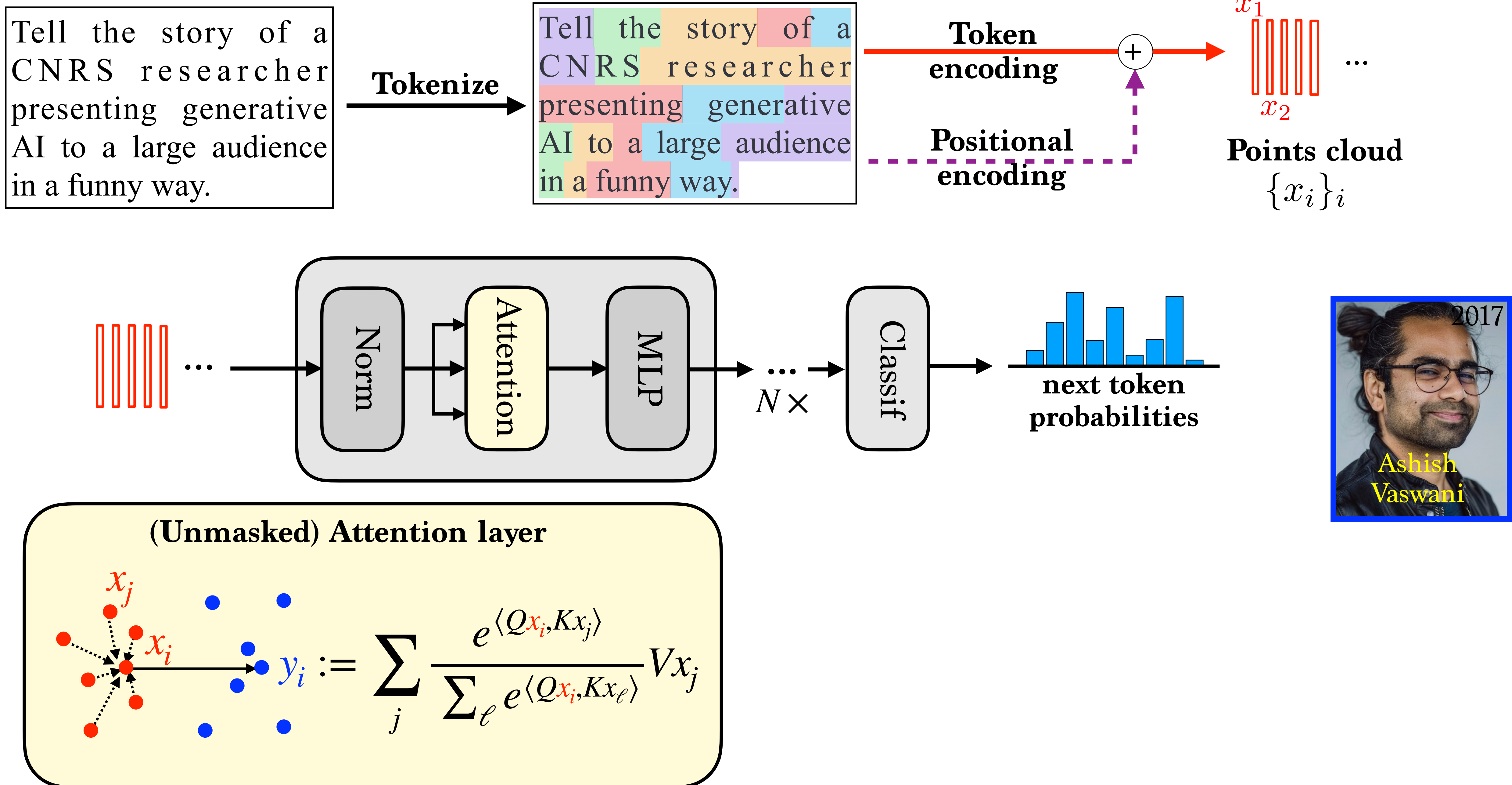


Differential equations

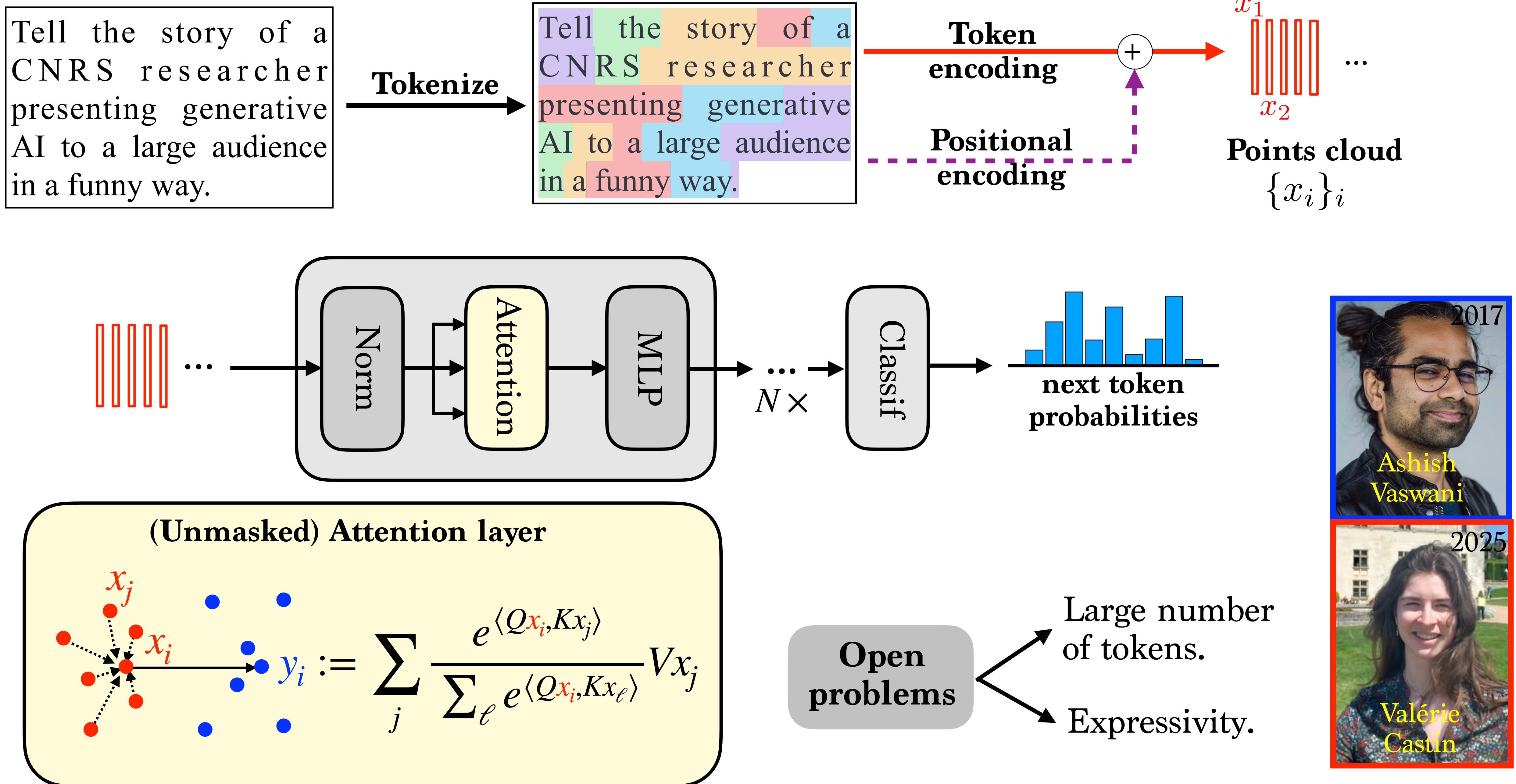
Transformers and attention mechanism



Transformers and attention mechanism



Transformers and attention mechanism





AI for Mathematics

Generative AI: text vs. images

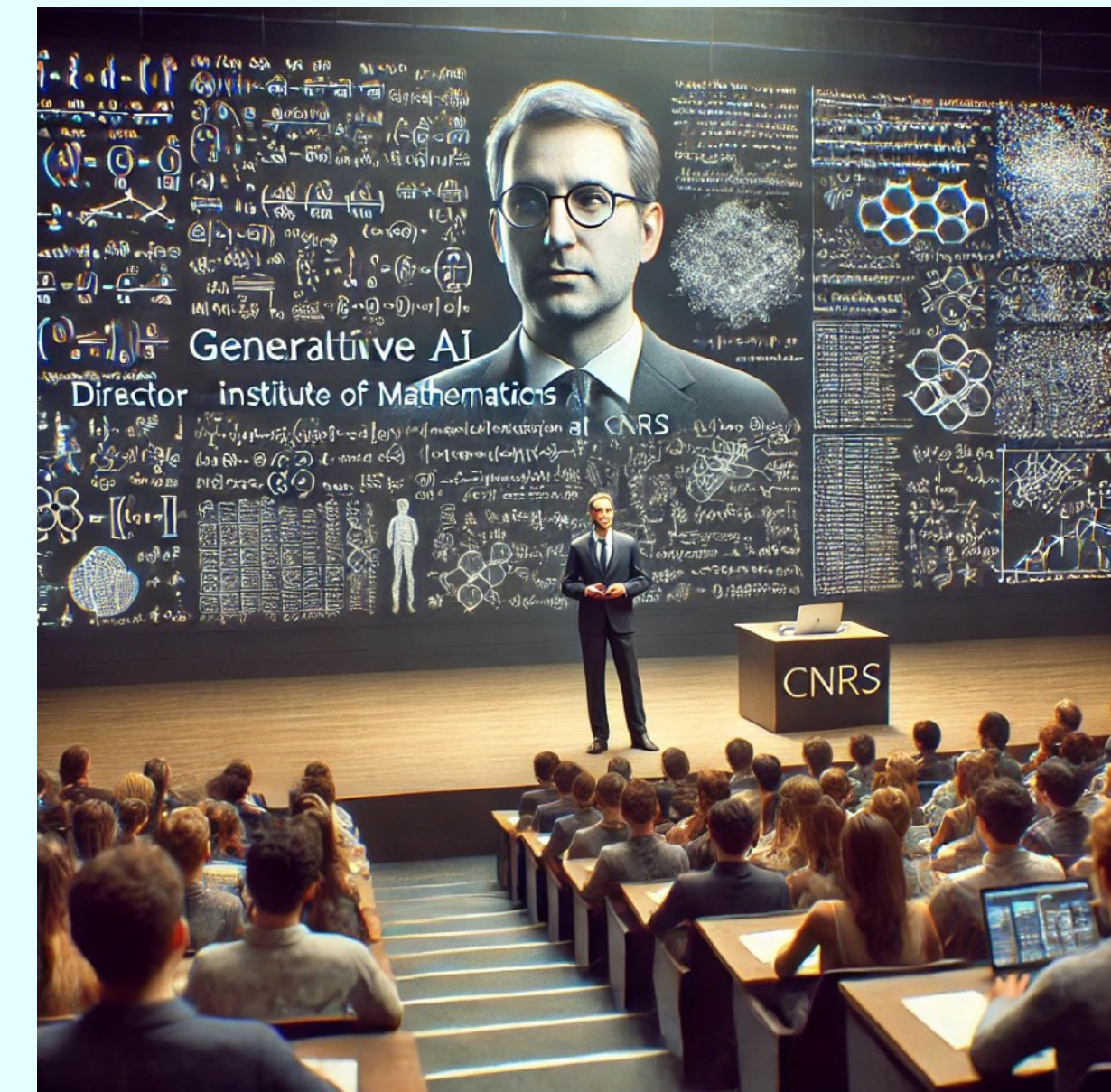
Tell the story of a CNRS researcher presenting generative AI to a large audience in a funny way.

LE CHAT
MISTRAL

Dr. Martin Lefèvre, the kind of CNRS researcher who refers to debugging as “therapy,” was scheduled to present generative AI to a huge audience—hundreds of people, two drones for some reason, and a guy in the front row eating chips directly out of his backpack. Martin strolled on stage, slightly sweaty, holding a USB stick like it was the One Ring. “Ladies and gentlemen,” he began, “generative AI is revolutionizing research, art, and occasionally my grocery list.” ...

Depict a CNRS mathematics researcher presenting generative AI in front of a large audience.

DALL·E 2



Generative AI: text vs. images

Tell the story of a CNRS researcher presenting generative AI to a large audience in a funny way.

LE CHAT
MISTRAL

Dr. Martin Lefèvre, the kind of CNRS researcher who refers to debugging as “therapy,” was scheduled to present generative AI to a huge audience—hundreds of people, two drones for some reason, and a guy in the front row eating chips directly out of his backpack. Martin strolled on stage, slightly sweaty, holding a USB stick like it was the One Ring. “Ladies and gentlemen,” he began, “generative AI is revolutionizing research, art, and occasionally my grocery list.” ...

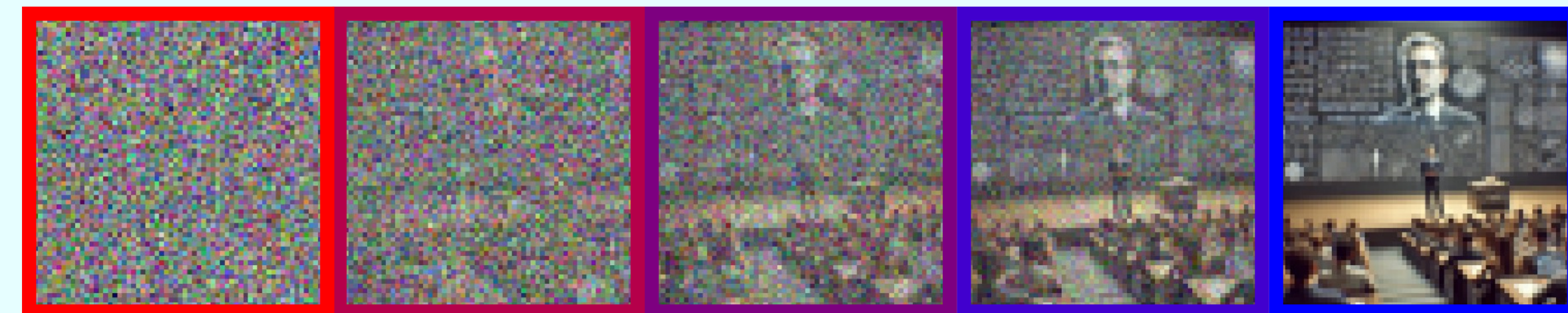
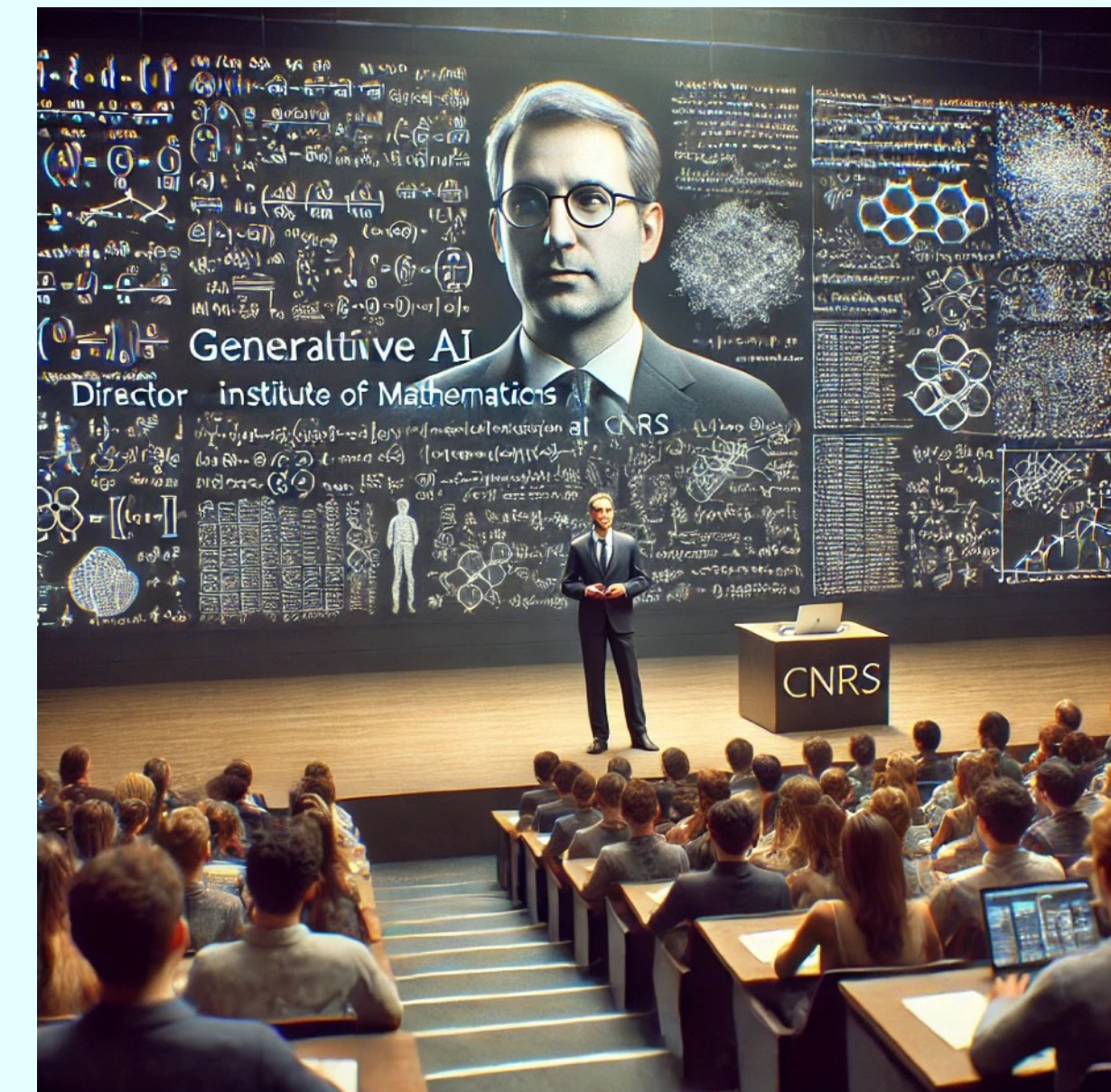
Dr. Martin Lefèvre, the kind of
CNRS researcher who refers
to debugging as

Generation: auto-regressive.

Pre-training: next token prediction.

Depict a CNRS mathematics researcher presenting generative AI in front of a large audience.

DALL·E 2



Generation: dynamic transport.

Pre-training: denoising.

LLM Workflow for Mathematics



LLMs:

- Pre-trained for data generation ... but use to solve (unseen?) problems.
- Change the workflow of mathematical research.
- Can be fine-tuned easily (LORA).

Let $(x_i)_{i=1}^n$ be iid $N(0,1)$, compute the asymptotics in n of the std of $A_n := \#(x_i > 0) / (n \cdot \max(x_i) + \#(x_i < 0) / n \cdot \min(x_i))$.

hence

$$\text{Var}(C_n) = \frac{2 \log n}{n} (1 + o(1)).$$

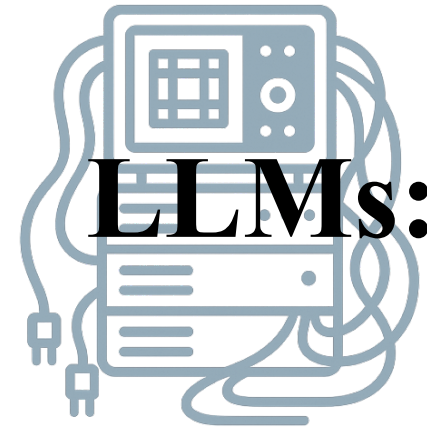
Since $\frac{2 \log n}{n} = o((\log n)^{-1})$, $\text{Var}(C_n)$ and the cross-term $\text{Cov}(B_n, C_n)$ are negligible compared with $\text{Var}(B_n)$.

4. Asymptotic variance and standard deviation

$$\text{Var}(A_n) = \frac{\pi^2}{24 \log n} (1 + o(1)), \quad \text{sd}(A_n) = \frac{\pi}{2\sqrt{6} \log n} (1 + o(1)).$$

The sole leading source of randomness is the mismatch between the two extreme order statistics; fluctuations from the sign count fade faster by a factor of $n^{-1/2}$.

LLM Workflow for Mathematics



LLMs:

- Pre-trained for data generation ... but use to solve (unseen?) problems.
- Change the workflow of mathematical research.
- Can be fine-tuned easily (LORA).

Let $(x_i)_{i=1}^n$ be iid $N(0,1)$, compute the asymptotics in n of the std of $A_n := \frac{\#\{x_i > 0\}}{n \max(x_i) + \#\{x_i < 0\}/n \min(x_i)}$.

Write this into a clean latex file with `begin{proposition}` and `begin{proof}`.

Write down a .ipynb notebook performing monte-carlo simulation to show the ratio between empirically computed std and the obtained theoretical formula.

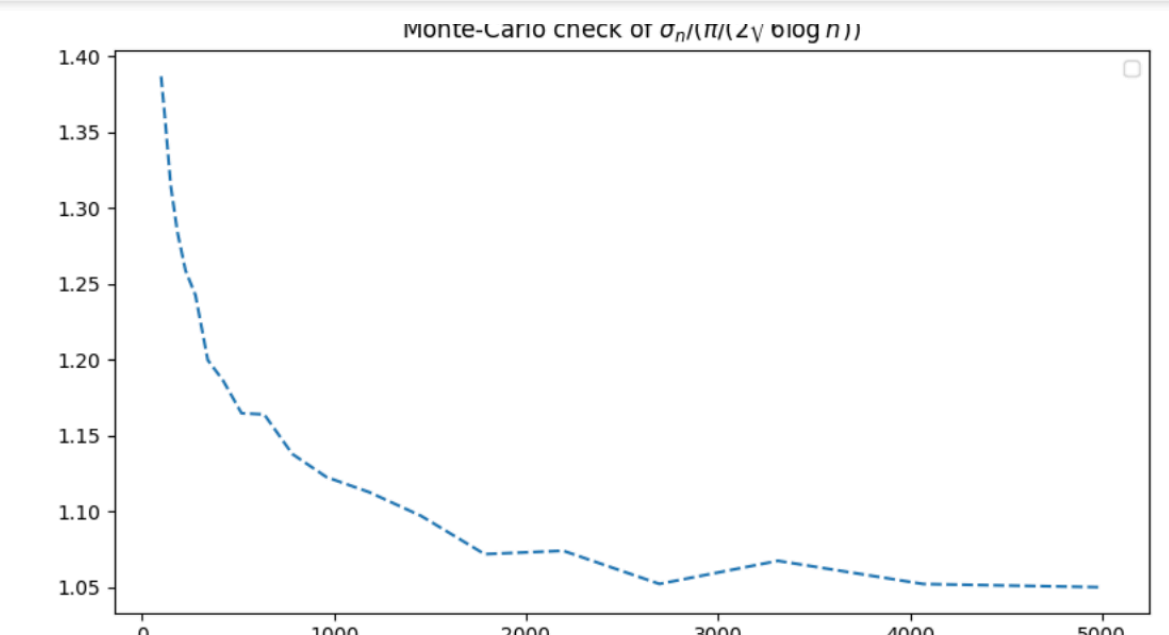
✓ Monte-Carlo verification of the asymptotic standard deviation

This notebook estimates the standard deviation (σ_n) of the Gaussian functional

$$A_n = \frac{\#\{X_i > 0\}}{n} \max_i X_i + \frac{\#\{X_i < 0\}}{n} \min_i X_i,$$

and compares the estimates to the theoretical prediction $\frac{\pi}{2\sqrt{6} \log n}$.

```
import numpy as np
import matplotlib.pyplot as plt
def sigma_n(n: int, m: int = 5000, rng=np.random.default_rng()):
    X = rng.standard_normal(size=(m, n))
    pos = X > 0, n_pos = pos.sum(axis=1), n_neg = n - n_pos
    max_vals = X.max(axis=1), min_vals = X.min(axis=1)
    A = (n_pos / n) * max_vals + (n_neg / n) * min_vals
    return A.std(ddof=0)
n_values = np.geomspace(100, 5_000, 20, dtype=int)
sims = 20_000 # Monte-Carlo replicates per n
rng = np.random.default_rng(seed=123456)
sigma_hat = np.array([sigma_n(n, sims, rng) for n in n_values])
asympt = np.pi / (2 * np.sqrt(6) * np.sqrt(np.log(n_values)))
```



Fluctuations of an Extreme-Weighted Sign Statistic

June 16, 2025

Abstract

We prove that the variance of a statistic combining the empirical sign ratio with the sample extremes of n i.i.d. $\mathcal{N}(0,1)$ observations decays like $\pi^2/(24 \log n)$, so its standard deviation is asymptotically $\pi/(2\sqrt{6} \log n)$.

Proposition 1. Let $(X_i)_{i=1}^n$ be i.i.d. $\mathcal{N}(0,1)$ variables and set

$$A_n = \frac{\#\{i : X_i > 0\}}{n} \max_{1 \leq i \leq n} X_i + \frac{\#\{i : X_i < 0\}}{n} \min_{1 \leq i \leq n} X_i.$$

Then, as $n \rightarrow \infty$,

$$\text{Var}(A_n) = \frac{\pi^2}{24 \log n} (1 + o(1)), \quad \text{sd}(A_n) = \frac{\pi}{2\sqrt{6} \log n} (1 + o(1)).$$

Proof. Let $N_+ = \#\{i : X_i > 0\}$, $N_- = n - N_+$, $p_n = N_+/n$, $M_+ = \max_{1 \leq i \leq n} X_i$, and $M_- = \min_{1 \leq i \leq n} X_i$. Then $A_n = p_n M_+ + (1 - p_n) M_-$.

hence

$$\text{Var}(C_n) = \frac{2 \log n}{n} (1 + o(1)).$$

Since $\frac{2 \log n}{n} = o((\log n)^{-1})$, $\text{Var}(C_n)$ and the cross-term $\text{Cov}(B_n, C_n)$ are negligible compared

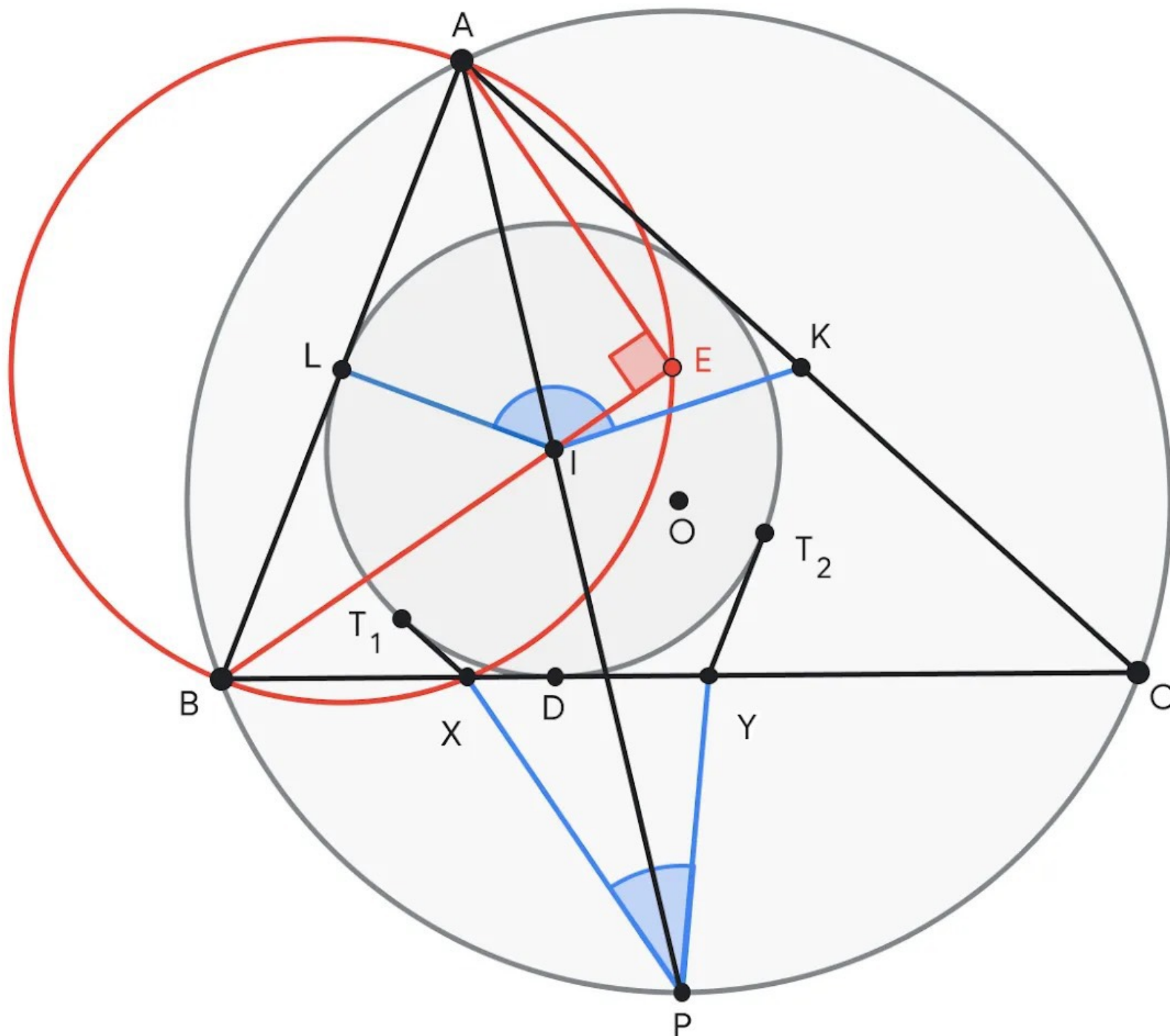
4. Asymptotic variance and standard deviation

$$\text{Var}(A_n) = \frac{\pi^2}{24 \log n} (1 + o(1)), \quad \text{sd}(A_n) = \frac{\pi}{2\sqrt{6} \log n} (1 + o(1))$$

The sole leading source of randomness is the mismatch between the two extreme order statistics; fluctuations from the sign count fade faster by a factor of $n^{-1/2}$.

IA for Mathematics

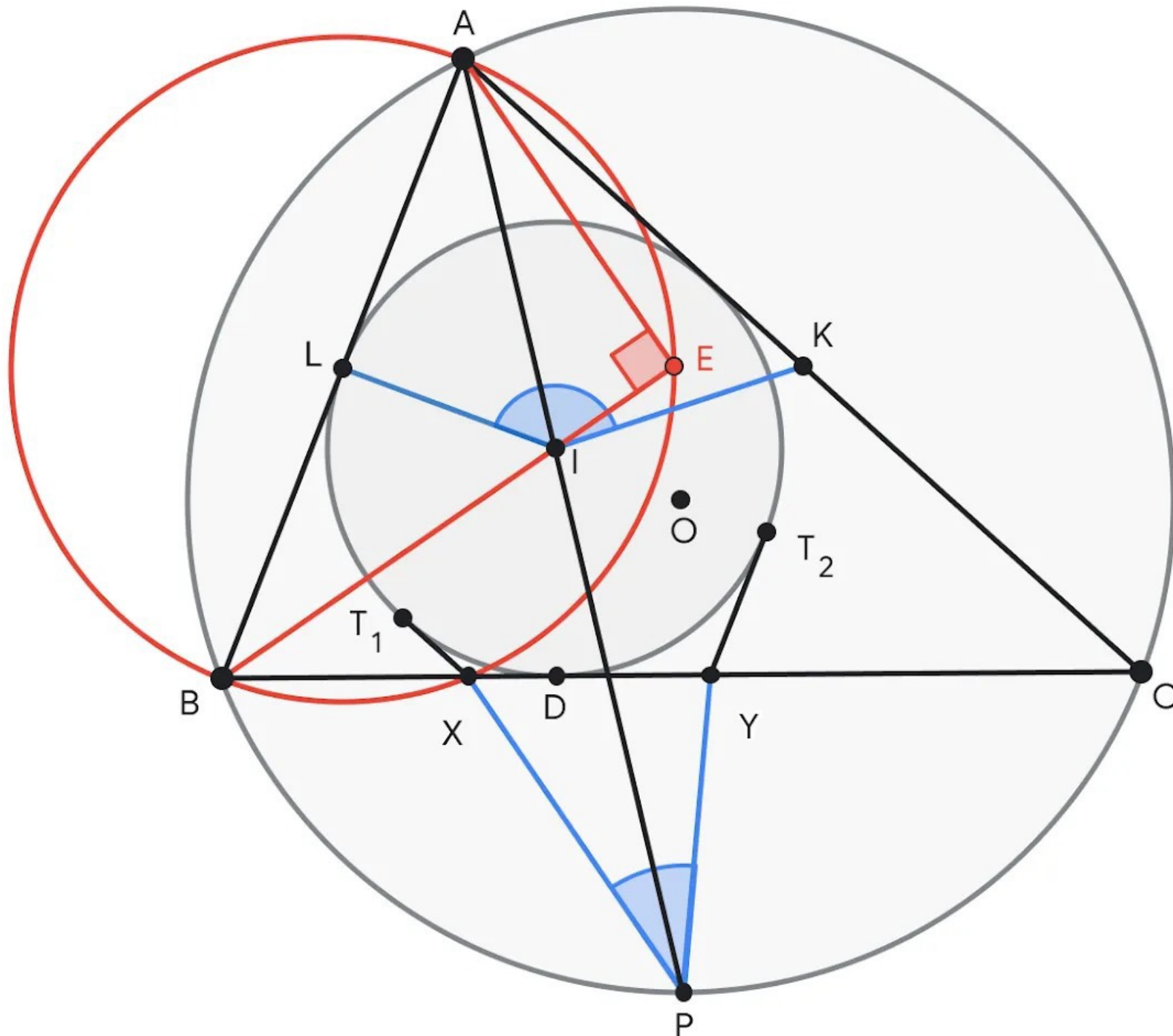
AlphaProof :
silver medal level at
the Olympiad.



Prove that $\angle KIL + \angle XPY = 180^\circ$

IA for Mathematics

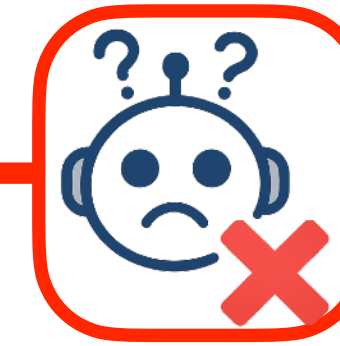
AlphaProof:
silver medal level at
the Olympiad.



Prove that $\angle KIL + \angle XPY = 180^\circ$

Prompt:

What is the 100th
term of the
arithmetic sequence
6, 10, 14, 18, ...?



Answer: 412



Pattern:
each term + 4

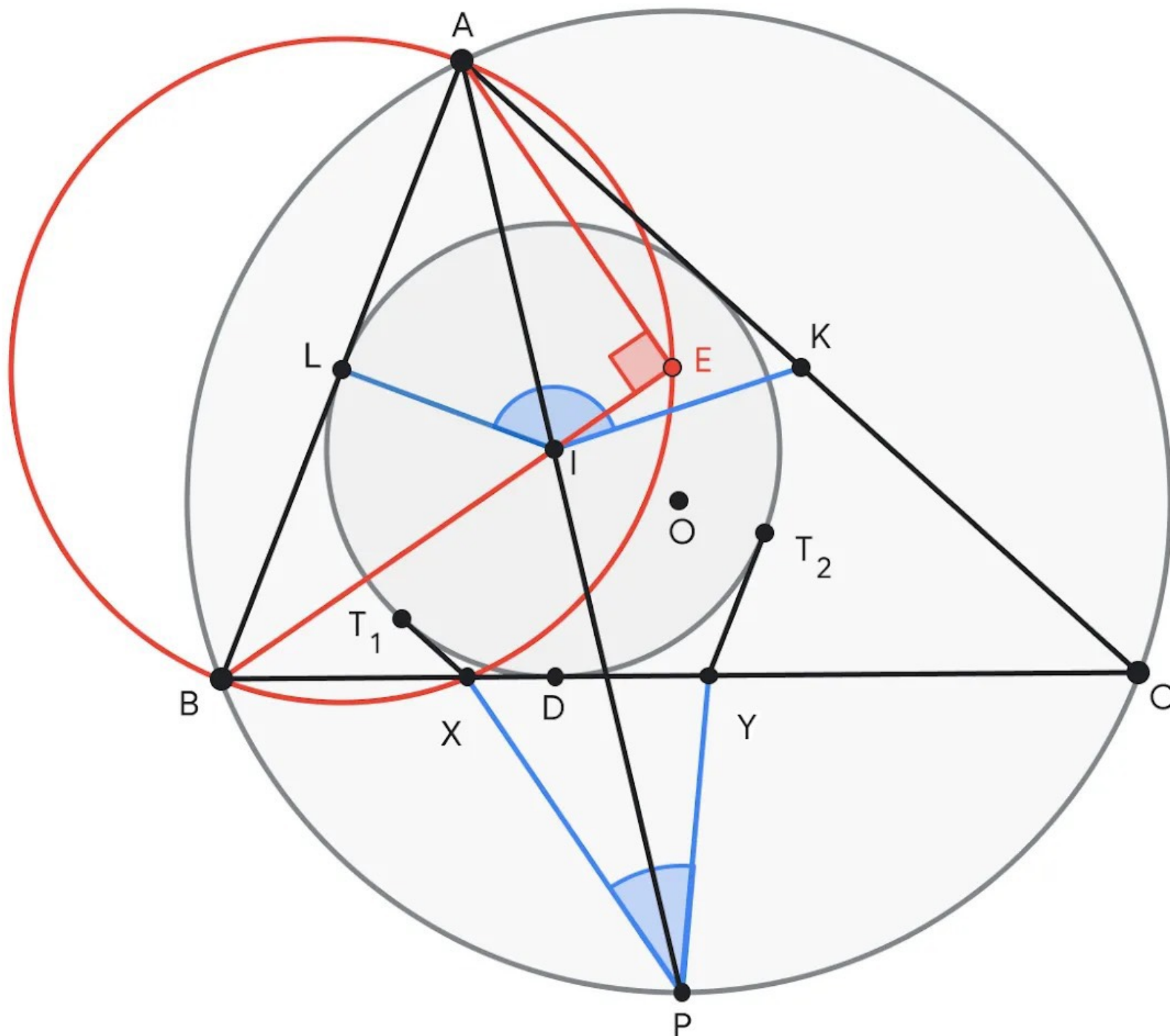
Rule:
 $a_n = 6 + (n-1) \cdot 4$

Answer: $n=100$,
 $6 + 99 \cdot 4 = 402$

Mathematics $\xrightarrow[\text{models}]{\text{Reasoning}}$ Billions \$
industry

IA for Mathematics

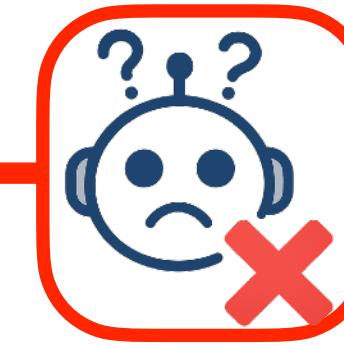
AlphaProof:
silver medal level at
the Olympiad.



Prove that $\angle KIL + \angle XPY = 180^\circ$

Prompt:

What is the 100th
term of the
arithmetic sequence
6, 10, 14, 18, ...?



Answer: 412



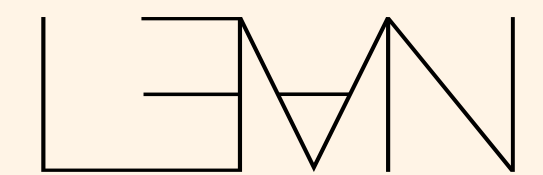
Pattern:
each term + 4

Rule:
 $a_n = 6 + (n-1) \cdot 4$

Answer: $n=100$,
 $6 + 99 \cdot 4 = 402$

Mathematics $\xrightarrow{\text{Reasoning models}}$ Billions \$ industry

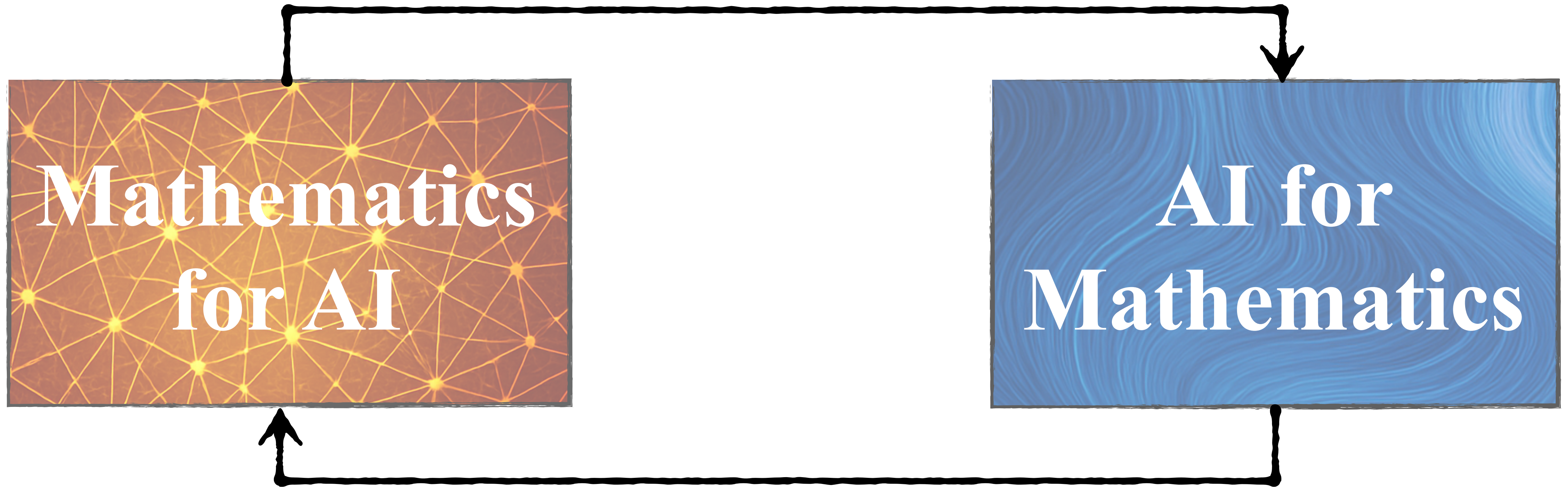
Integrating formal proof languages



The future of $\begin{cases} \text{the mathematicians?} \\ \text{the teacher?} \end{cases}$



Conclusion



Math concepts are the heart of AI

Theory is key to replace transformers.

Are LLMs interpolating or reasoning?

Math reasoning: industrial shift.

LLM as a assistant for mathematician.

Formal vs informal reasoning