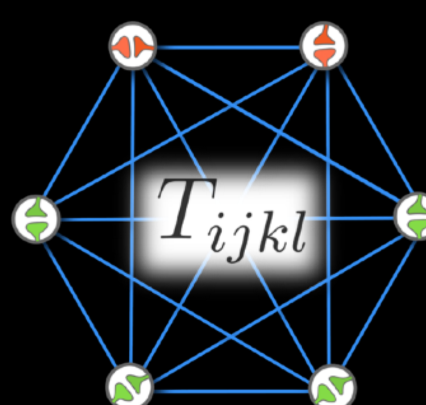
Modern Methods in Associative Memory

Dmitry Krotov *Benjamin Hoover* *Parikshit Ram*

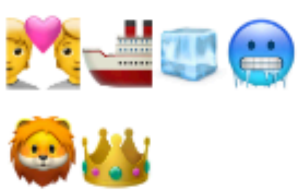
$$E = -\sum_{\mu=1}^{K} F\Big(\sum_{i=1}^{D} \xi_i^{\mu} \sigma_i\Big)$$



What is Associative Memory?

Association
Connect inputs to impose structure on a complex world

Shape+color with universal meaning

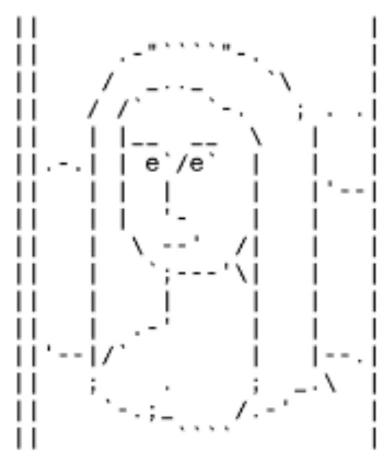


Name that movie!

What does this picture “smell” like?

Memory
Leverage association to recall missing information

Who is this?



What color is her hair?

Error Correction
Filter corruption to detect meaning behind the noise

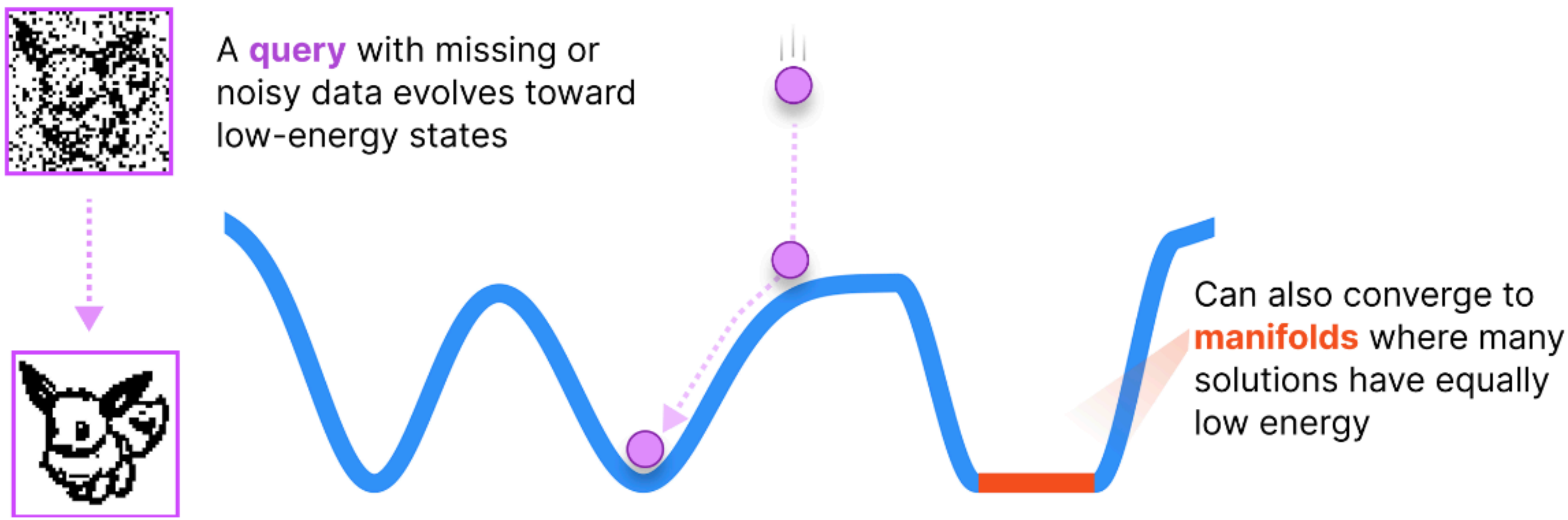
Aoccdrnig to a rscheearch sdtuy at Cmabrigde Uinervtisy, it deons't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit pobrelm.

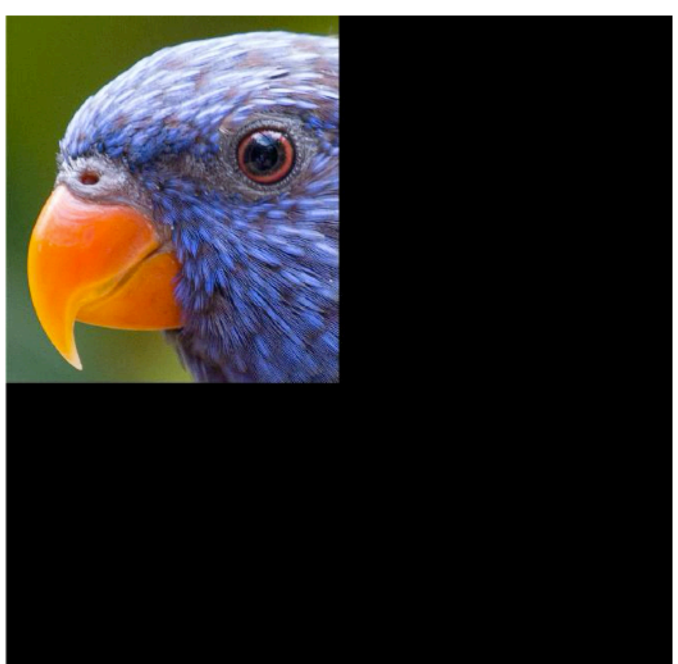
Associative Memory

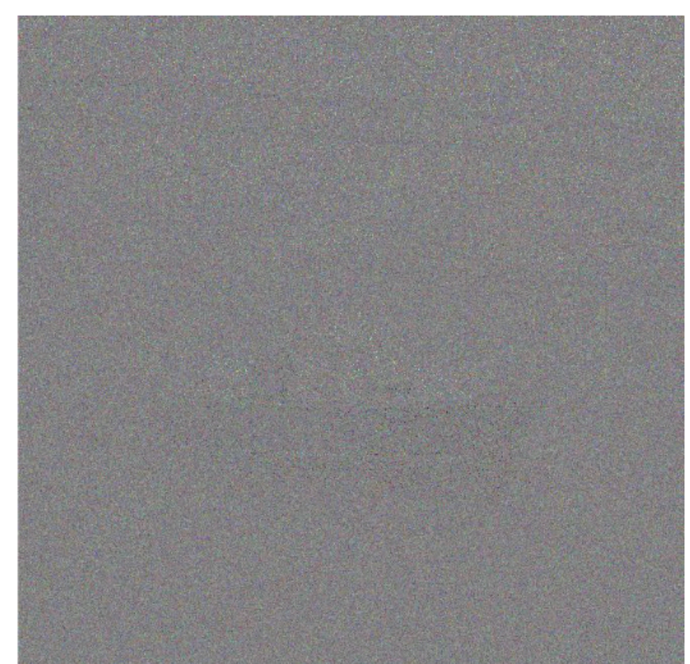
Content-addressable information storage systems capable of error correction

Energy-based Associative Memory

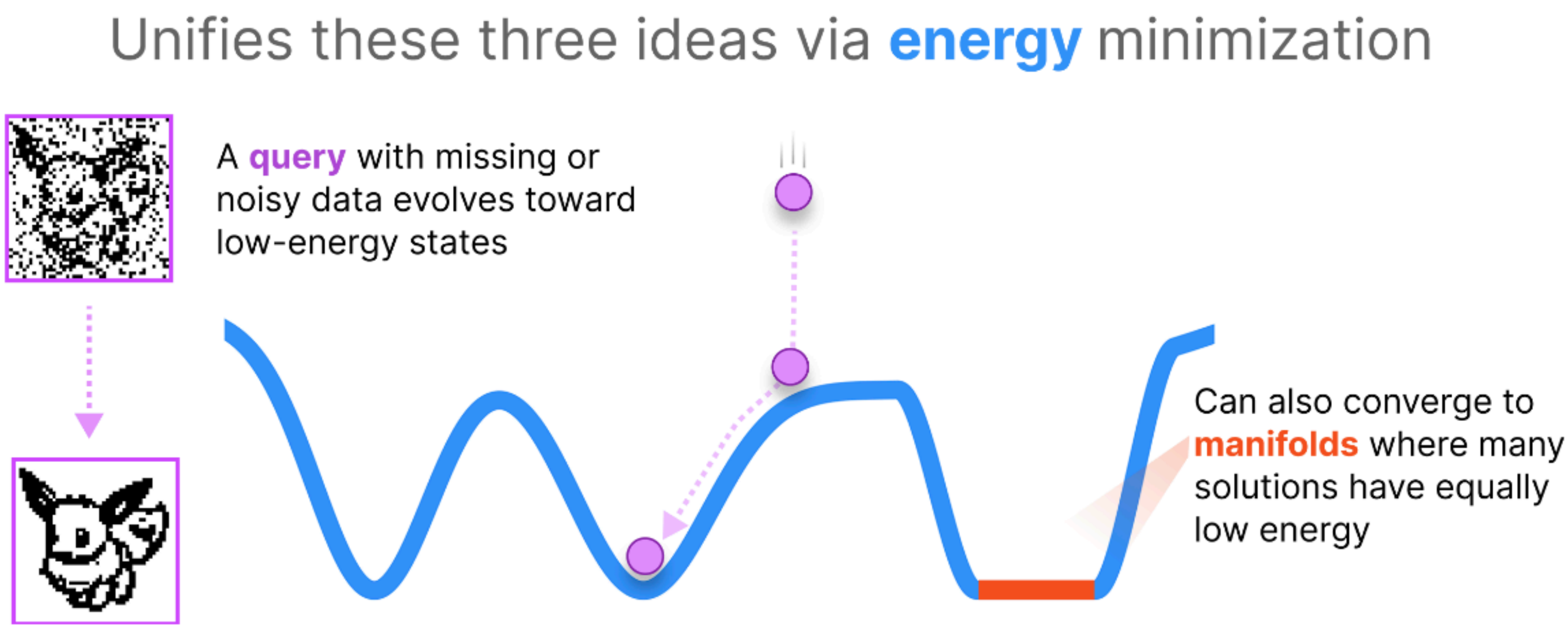
Unifies these three ideas via **energy** minimization







Energy-based Associative Memory



- Local minima are called memories.
- Non-linear dynamics of energy decent - the process of memory recall.
- Association happens through this non-linear dynamics between the state at t=0 and the final state at convergence.

Hopfield Network

$$E = -\sum_{\mu=1}^{K}\Big(\sum_{i=1}^{D}\xi_i^\mu\sigma_i\Big)^2 = -\sum_{i,j=1}^{D}\sigma_i T_{ij}\sigma_j$$

$$T_{ij} = \sum_{\mu=1}^{K}\xi_i^\mu\xi_j^\mu$$

$\sigma_i \in \{\pm 1\}$ - dynamical variables (neurons)

ξ_i^μ - memorized patterns

D - number of neurons

K - number of memories

$$K^{\max} \approx 0.14D$$

Dense Associative Memory

$$E = -\sum_{\mu=1}^{K}F\Big(\sum_{i=1}^{D}\xi_i^\mu\sigma_i\Big) =$$

$$= -\sum_{i_1,i_2,\ldots,i_n}^{D} T_{i_1 i_2 \ldots i_n}\sigma_{i_1}\sigma_{i_2}\ldots\sigma_{i_n}$$

$$T_{i_1 i_2 \ldots i_n} = \sum_{\mu=1}^{K}\xi_{i_1}^\mu\xi_{i_2}^\mu\ldots\xi_{i_n}^\mu$$

$F(x) = x^n$ - separation function

$$K^{\max} \approx \alpha_n D^{n-1}$$

$$K^{\max} \approx 2^{\frac{D}{2}}$$

Update rule for energy decent

$$\sigma_i^{(t+1)} = \underset{b\in\{-1,1\}}{\operatorname{argmin}} \left[E\left(\sigma_i = b, \sigma_{j\neq i} = \sigma_j^{(t)} \right) \right]$$

$$\sigma_i^{(t+1)} = Sign\Big[\sum_{\mu=1}^{K} \Big(F\Big(\xi_i^\mu + \sum_{j\neq i} \xi_j^\mu \sigma_j^{(t)} \Big) - F\Big(-\xi_i^\mu + \sum_{j\neq i} \xi_j^\mu \sigma_j^{(t)} \Big) \Big) \Big]$$

$$F\Big(\xi_i^\mu + \sum_{j\neq i}^{D} \xi_j^\mu \sigma_j^{(t)} \Big) \approx F\Big(\sum_{j\neq i}^{D} \xi_j^\mu \sigma_j^{(t)} \Big) + \xi_i^\mu \frac{dF}{dx}\Big|_{x=\sum_{j\neq i} \xi_j^\mu \sigma_j^{(t)}}$$

+ small higher order terms

$$\sigma_i^{(t+1)} = Sign\Big[\sum_{\mu=1}^{K} \xi_i^\mu f\Big(\sum_{j\neq i} \xi_j^\mu \sigma_j^{(t)} \Big) \Big] \qquad \text{where} \quad f(x) = \frac{dF}{dx}$$

How many memories can we store?

Imagine that memories are random binary vectors

$$\xi_i^\mu = \begin{cases} +1, & \text{with probability } \frac{1}{2} \\ -1, & \text{with probability } \frac{1}{2} \end{cases}$$

$$\sigma_i^{(t+1)} = Sign\Big[\sum_{\mu=1}^{K} \xi_i^\mu f\Big(\sum_{j \neq i} \xi_j^\mu \sigma_j^{(t)}\Big)\Big]$$

We will initialize the network in the state $\sigma_i^{(0)} = \xi_i^1$

$$\sigma_i^{(t+1)} = \text{Sign}\Big[\xi_i^1 \ f\Big(\sum_{j \neq i}^{D} \xi_j^1 \ \xi_j^1\Big) + \sum_{\mu=2}^{K} \xi_i^\mu \ f\Big(\sum_{j \neq i}^{D} \xi_j^\mu \ \xi_j^1\Big)\Big]$$

$$= \text{Sign}\Big[\underbrace{\xi_i^1 \ f\Big(D-1\Big)}_{\text{signal}} + \underbrace{\sum_{\mu=2}^{K} \xi_i^\mu \ f\Big(\sum_{j \neq i}^{D} \xi_j^\mu \ \xi_j^1\Big)}_{\text{noise}}\Big]$$

How many memories can we store?

Imagine that memories are random binary vectors

$$\xi_i^\mu = \begin{cases} +1, & \text{with probability } \frac{1}{2} \\ -1, & \text{with probability } \frac{1}{2} \end{cases}$$

$$\langle\ \xi_i^\mu\ \rangle = 0, \qquad \langle\ \xi_i^\mu\ \xi_j^\nu\ \rangle = \delta^{\mu\nu}\delta_{ij}$$

$$\sigma_i^{(t+1)} = Sign\Big[\sum_{\mu=1}^{K} \xi_i^\mu f\Big(\sum_{j\neq i} \xi_j^\mu \sigma_j^{(t)}\Big)\Big]$$

We will initialize the network in the state $\sigma_i^{(0)} = \xi_i^1$

$$\sigma_i^{(t+1)} = \text{Sign}\Big[\xi_i^1\ f\Big(\sum_{j\neq i}^{D} \xi_j^1\ \xi_j^1\Big) + \sum_{\mu=2}^{K} \xi_i^\mu\ f\Big(\sum_{j\neq i}^{D} \xi_j^\mu\ \xi_j^1\Big)\Big]$$

$$= \text{Sign}\Big[\underbrace{\xi_i^1\ f\Big(D-1\Big)}_{\text{signal}} + \underbrace{\sum_{\mu=2}^{K} \xi_i^\mu\ f\Big(\sum_{j\neq i}^{D} \xi_j^\mu\ \xi_j^1\Big)}_{\text{noise}}\Big] \stackrel{?}{=} \xi_i^1$$

initial state is stable

Information storage capacity

Our noise is a Gaussian random variable with zero mean and variance Σ^2

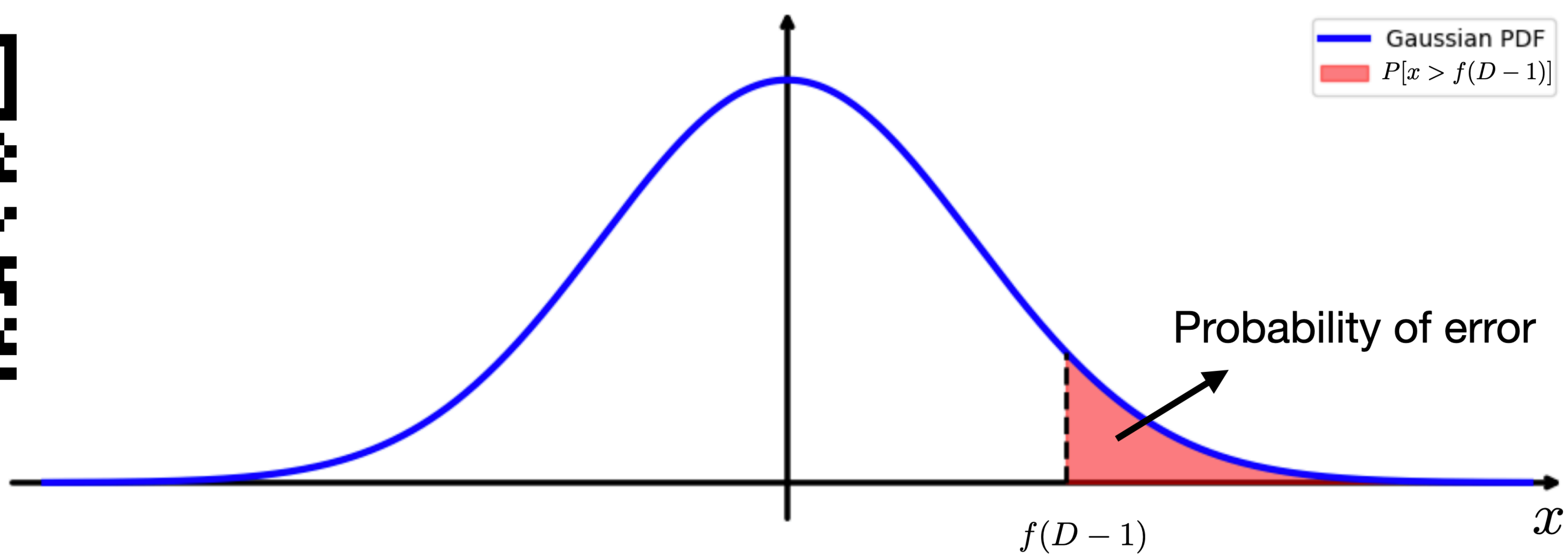
$$\langle \text{ noise } \rangle = 0$$

$$\Sigma^2 = \langle \text{ noise}^2 \rangle = (2n-3)!!\,K\,D^{n-1}$$

$$E = -\sum_{\mu=1}^{K} F\Big(\sum_{i=1}^{D} \xi_i^{\mu} \sigma_i\Big)$$

$$F(x) = x^n$$

$$\text{Sign}\Big[\underbrace{\xi_i^1 \; f\big(D-1\big)}_{\text{signal}} + \underbrace{\sum_{\mu=2}^{K} \xi_i^{\mu} \; f\Big(\sum_{j\neq i}^{D} \xi_j^{\mu} \; \xi_j^1\Big)}_{\text{noise}}\Big] \stackrel{?}{=} \xi_i^1$$



Information storage capacity

Let's compute the probability of error - bit flip

$$\mathrm{Sign}\Big[\underbrace{\xi_i^1\ f\big(D-1\big)}_{\text{signal}} + \underbrace{\sum_{\mu=2}^{K}\xi_i^\mu\ f\Big(\sum_{j\neq i}^{D}\xi_j^\mu\ \xi_j^1\Big)}_{\text{noise}}\Big] \stackrel{?}{=} \xi_i^1$$

$$P(\text{error}) = \int_{f(D-1)}^{\infty} \frac{dx}{\sqrt{2\pi\Sigma^2}} e^{-\frac{x^2}{2\Sigma^2}} = \int_{\frac{f(D-1)}{\Sigma}}^{\infty} \frac{dy}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} = g\Big(\frac{f(D-1)}{\Sigma}\Big) < 1\%$$

$$f(D-1) > \alpha\Sigma$$

$$\Sigma^2 = \langle\ \text{noise}^2\ \rangle = (2n-3)!!KD^{n-1}$$

$$F(x) = x^n$$

$$K < K^{\max} = \frac{1}{\alpha^2(2n-3)!!}D^{n-1}$$

Information storage capacity

$$K < K^{\max} = \frac{1}{\alpha^2(2n-3)!!} D^{n-1}$$

Classical Hopfield network n=2

$$K^{\max} \sim D$$

$$E = -\frac{1}{2}\sum_{\mu=1}^{K}\Big(\sum_{i=1}^{D}\xi_i^{\mu}\sigma_i\Big)^2 = -\frac{1}{2}\sum_{i,j=1}^{D}\sigma_i T_{ij}\sigma_j, \quad \text{where} \quad T_{ij} = \sum_{\mu=1}^{K}\xi_i^{\mu}\xi_j^{\mu}$$

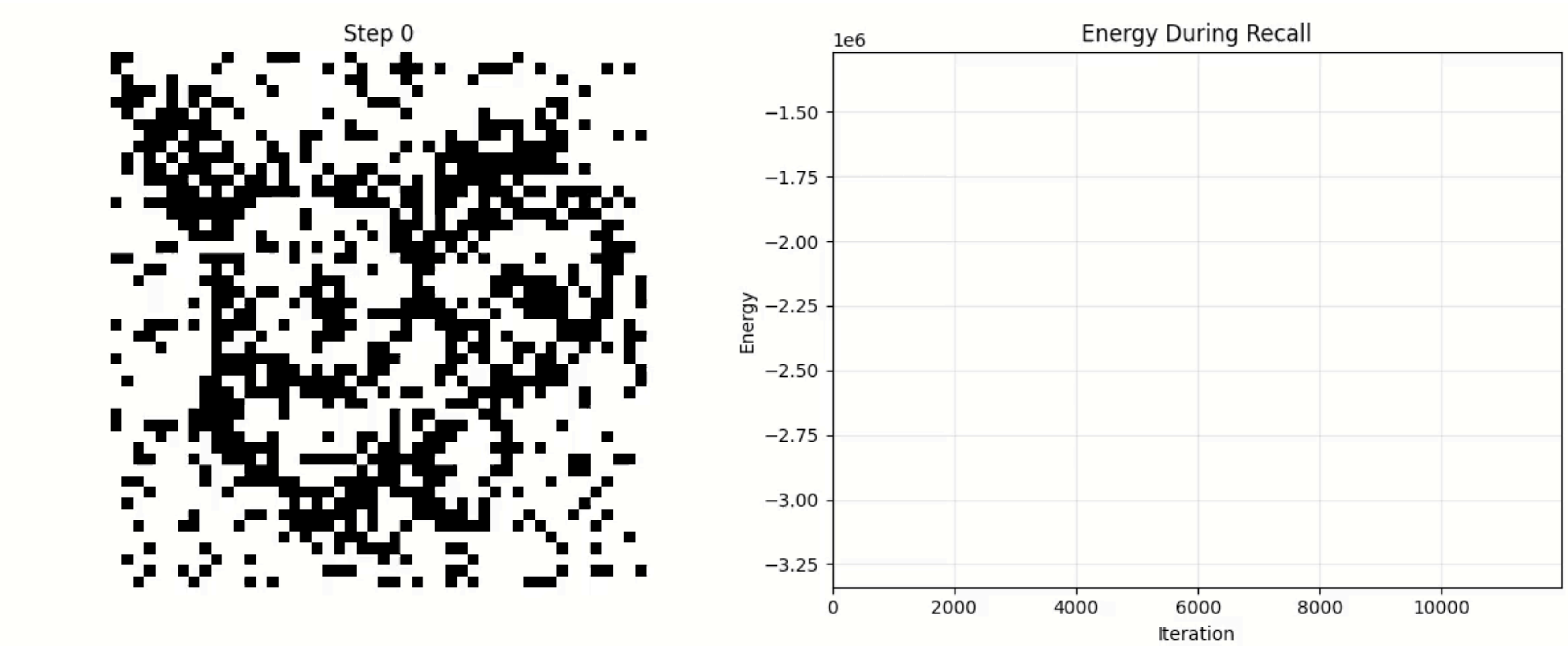
Dense Associative Memory with n=3

$$K^{\max} \sim D^2$$

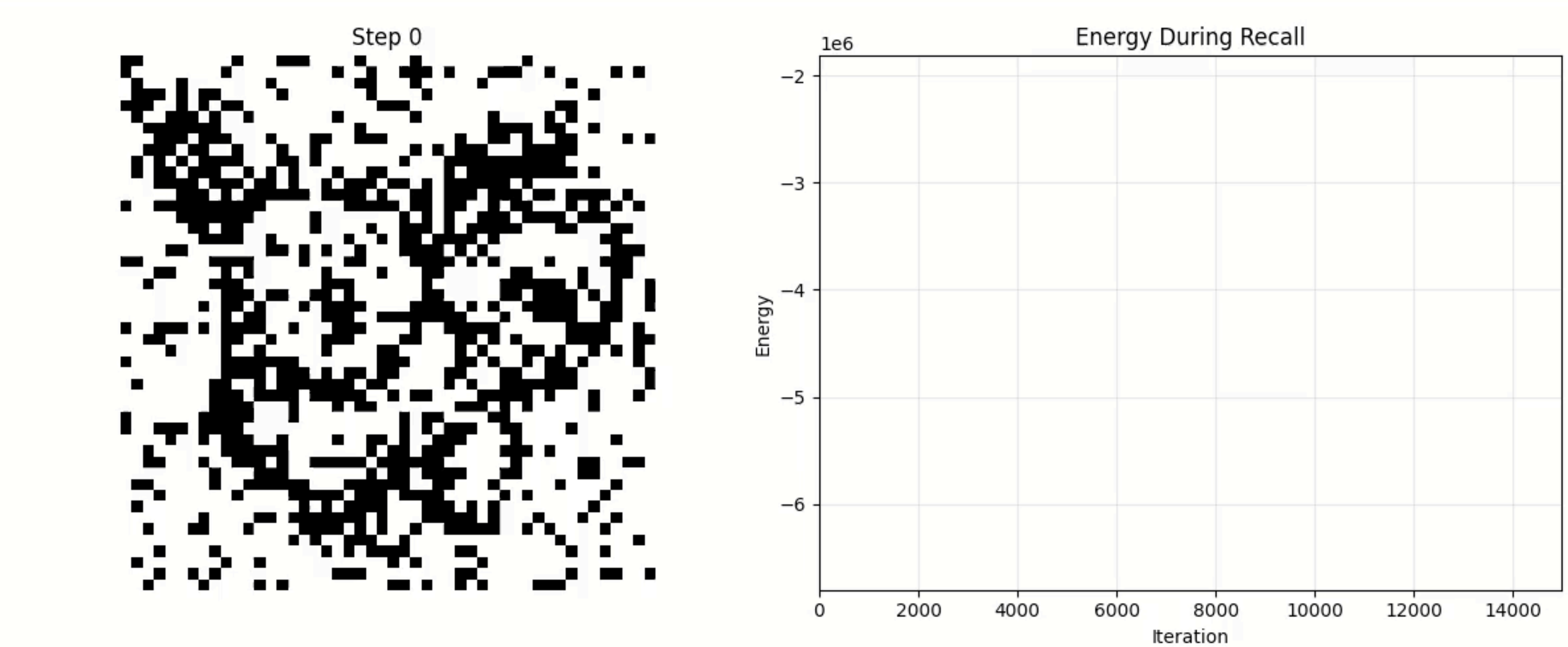
$$E = -\frac{1}{3}\sum_{\mu=1}^{K}\Big(\sum_{i=1}^{D}\xi_i^{\mu}\sigma_i\Big)^3 = -\frac{1}{3}\sum_{i,j,k=1}^{D}T_{ijk}\sigma_i\sigma_j\sigma_k, \quad \text{where} \quad T_{ijk} = \sum_{\mu=1}^{K}\xi_i^{\mu}\xi_j^{\mu}\xi_k^{\mu}$$

Information storage capacity

Classical Hopfield network n=2



K=2 memories stored

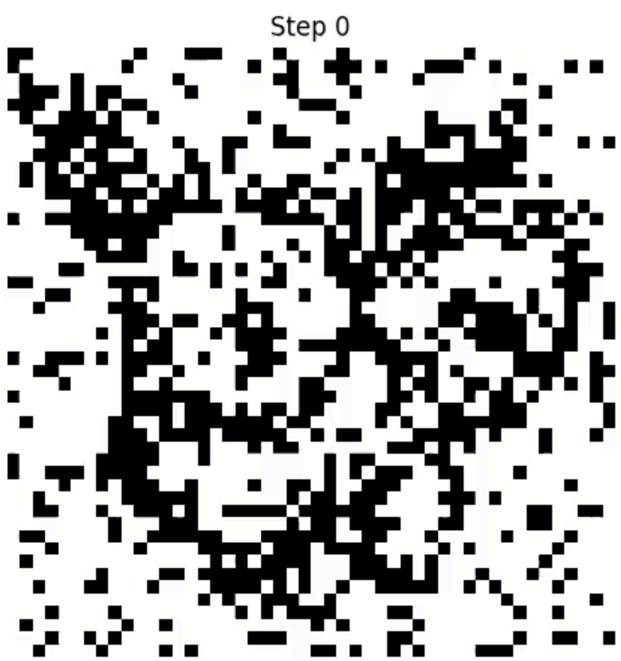


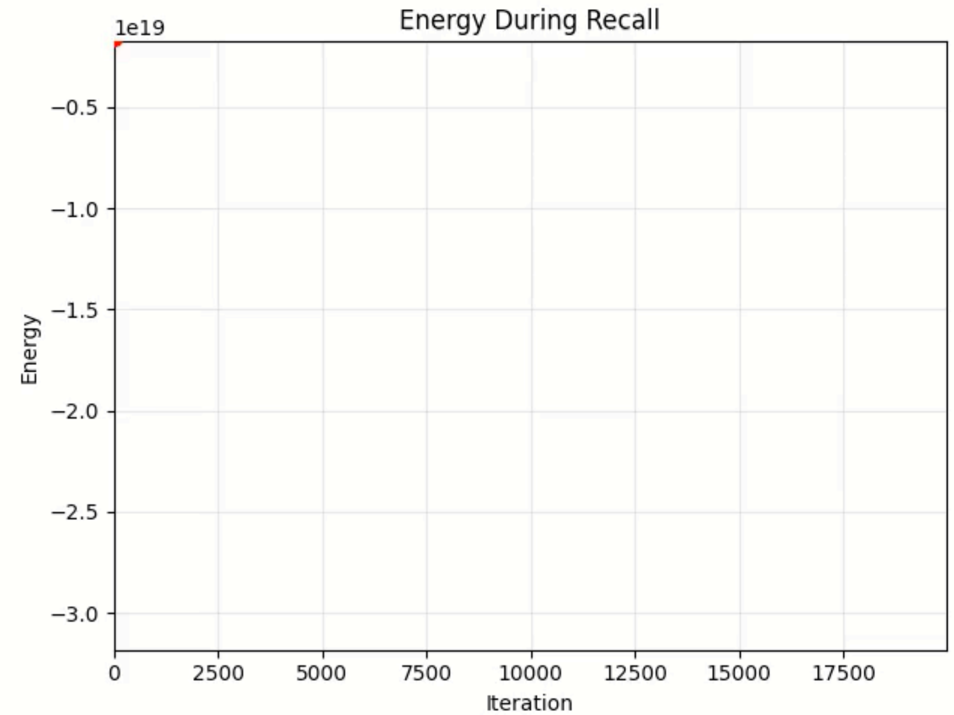
K=6 memories stored

Information storage capacity

Dense Associative Memory with n=6

$$K^{\max} \sim D^5$$





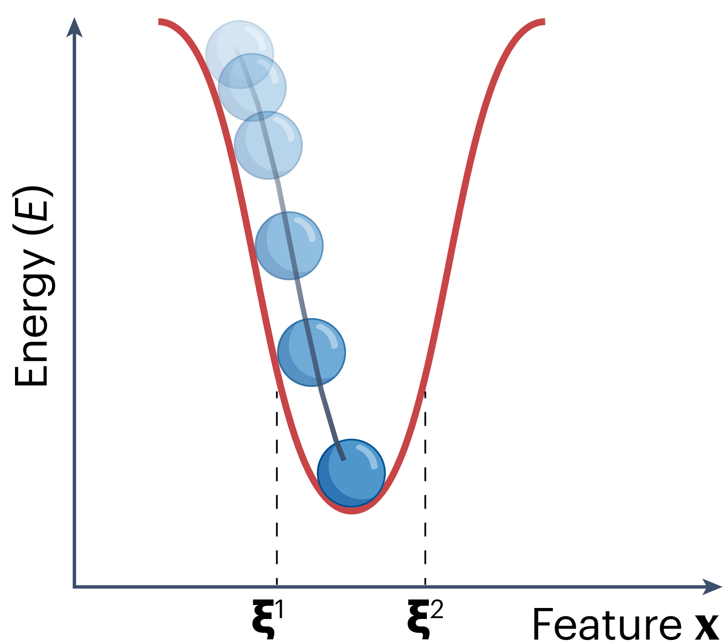
K=100 memories stored

Information storage capacity

What have we learned so far?

- The number of memories K is upper bounded.
- The Memory storage capacity heavily depends on the shape of the energy function $F(\cdot)$ and the shape of the activation function $f(\cdot)$.
- The sharper the energy peaks around memories – the larger the memory storage capacity.

Classical Hopfield network



Dense Associative Memory

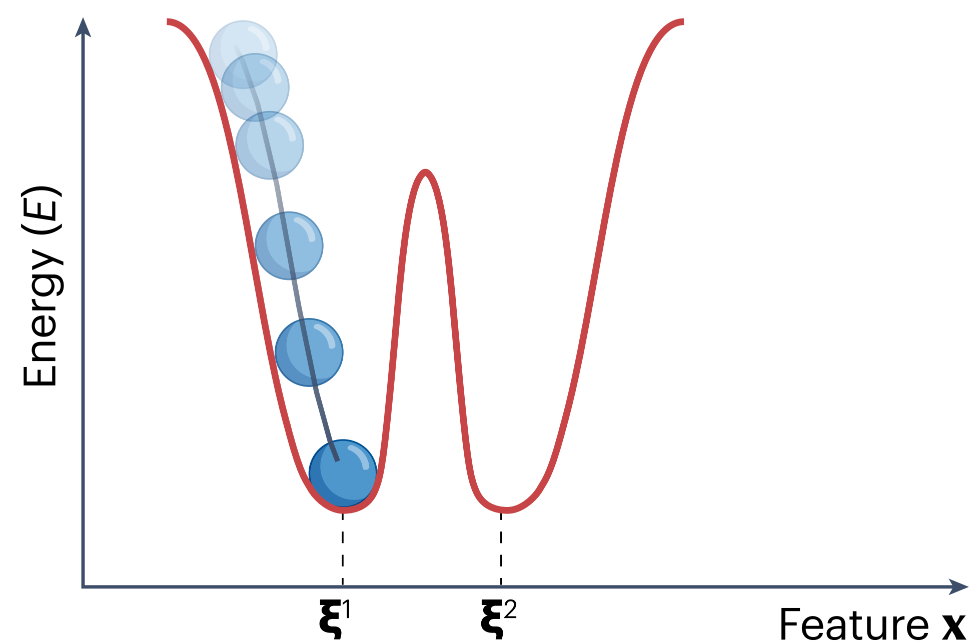


Image from “A new frontier for Hopfield networks”, Nature Reviews, 2023

General Dense Associative Memory

$$E = -Q\Big[\sum_{\mu=1}^{K} F\Big(S\big[\boldsymbol{\xi}^{\mu}, \boldsymbol{\sigma}\big]\Big)\Big]$$

$S[\boldsymbol{x}, \boldsymbol{x'}]$ - similarity function

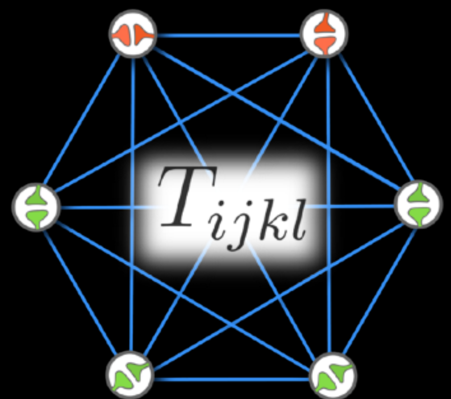
$F(\cdot)$ - separation function

$Q(\cdot)$- scalar monotone function

Failure of Memory and Generative AI

Dmitry Krotov *Benjamin Hoover* *Parikshit Ram*

$$E = -\sum_{\mu=1}^{K} F\Big(\sum_{i=1}^{D} \xi_i^\mu \sigma_i\Big)$$



When someone—especially you—reminds me of that day, I remember that it was you who told me about the murder, or at least that's how I remember it. <…> I suppose you... Or rather, I know that you came downstairs and told me that you heard it on the news. I don't know what time it happened. There, in that hole, in <Name of the Place>, it was easy to lose track of time. <…> I had already been working for quite a while and was very focused on what I was doing when you suddenly interrupted me, saying that you had heard something. I'm sure it was you who said: "The President has been killed, or rather, shot—he's been shot." Then I probably looked up and asked: "What?" And you replied: "Kennedy—he was shot." I said: "What do you mean? Where?" And you said you didn't know...

Roger Brown

James Kulik

EF Loftus, Memory., 1988

NAS, Biographical Memoirs, 1999

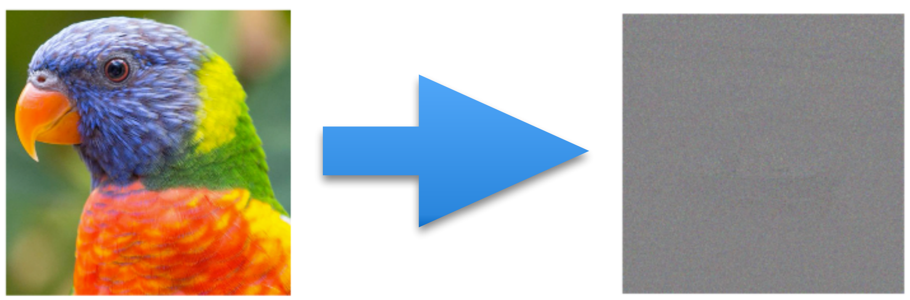
Misremembering is a failure of human memory in which multiple observed events (training data) blend together and form novel memories, which are different from any of the observed events (training data points).

Misremembering leads to creativity

Diffusion Models

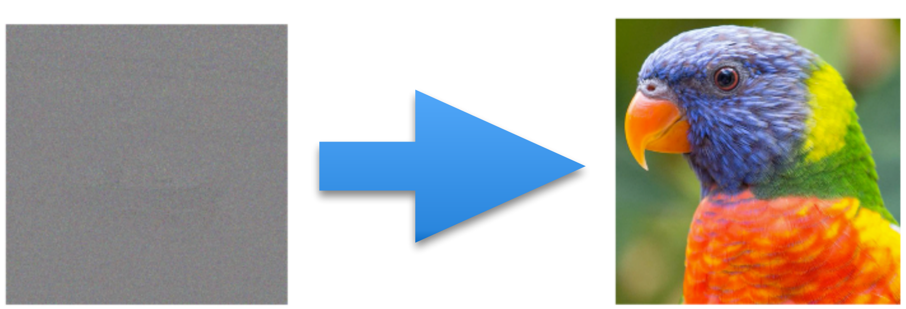
Forward process:

$$\mathrm{d}\mathbf{x}_t = g(t)\mathrm{d}\mathbf{w}_t$$



Reverse process:

$$\mathrm{d}\mathbf{x}_t = -g(t)^2\nabla_{\mathbf{x}_t}\log p_t(\mathbf{x}_t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}_t$$



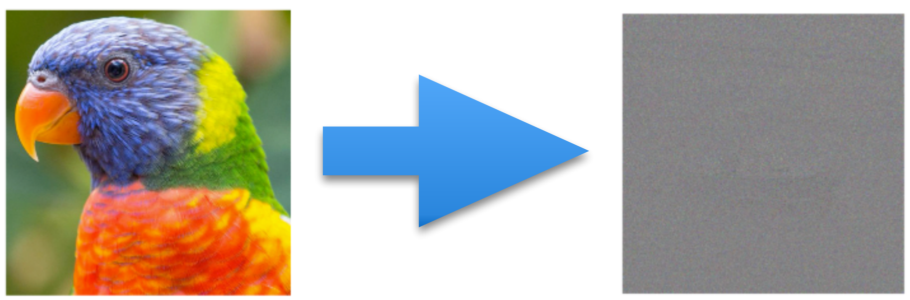
Neural network

$$s_\theta(\mathbf{x}, t) = \nabla_{\mathbf{x}_t}\log p_t(\mathbf{x}_t)$$

Diffusion Models

Forward process:

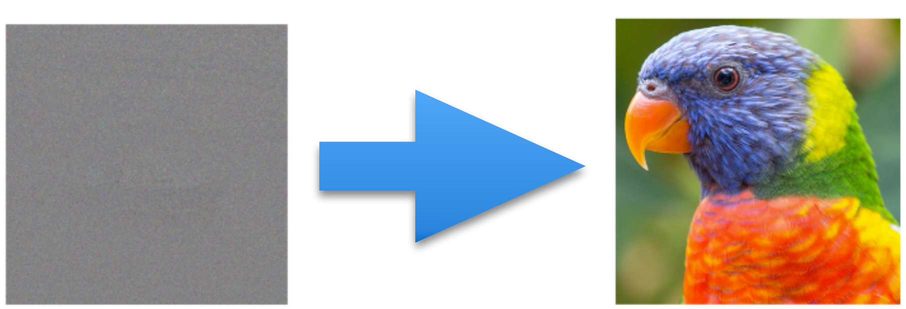
$$\mathrm{d}\mathbf{x}_t = g(t)\mathrm{d}\mathbf{w}_t$$



Training of neural network = writing information into the memory

Reverse process:

$$\mathrm{d}\mathbf{x}_t = -g(t)^2 \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}_t$$



Reverse process = attempt of memory recall

Neural network

$$s_\theta(\mathbf{x}, t) = \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$$

Diffusion Models as DenseAM

$$p(\mathbf{x}_\tau, \tau) = \mathbb{E}_{\mathbf{y}\sim\text{data}}\left[\frac{1}{(2\pi\sigma^2\tau)^{\frac{N}{2}}}\exp\Big(-\frac{\|\mathbf{x}_\tau - \mathbf{y}\|_2^2}{2\tau\sigma^2}\Big)\right]$$

$$p(\mathbf{y}) = \frac{1}{K}\sum_{\mu=1}^{K}\delta^{(N)}(\mathbf{y} - \boldsymbol{\xi}^\mu)$$

$$p(\mathbf{x}_\tau, \tau) \approx \frac{1}{K}\sum_{\mu=1}^{K}\frac{1}{(2\pi\sigma^2\tau)^{\frac{N}{2}}}\exp\Big(-\frac{\|\mathbf{x}_\tau - \boldsymbol{\xi}^\mu\|_2^2}{2\tau\sigma^2}\Big)$$

Diffusion Models as DenseAM

$$p(\mathbf{x}_\tau, \tau) = \mathbb{E}_{\mathbf{y}\sim\text{data}}\left[\frac{1}{(2\pi\sigma^2\tau)^{\frac{N}{2}}}\exp\Big(-\frac{\|\mathbf{x}_\tau - \mathbf{y}\|_2^2}{2\tau\sigma^2}\Big)\right]$$

$$p(\mathbf{y}) = \frac{1}{K}\sum_{\mu=1}^{K}\delta^{(N)}(\mathbf{y} - \boldsymbol{\xi}^\mu)$$

$$p(\mathbf{x}_\tau, \tau) \approx \frac{1}{K}\sum_{\mu=1}^{K}\frac{1}{(2\pi\sigma^2\tau)^{\frac{N}{2}}}\exp\Big(-\frac{\|\mathbf{x}_\tau - \boldsymbol{\xi}^\mu\|_2^2}{2\tau\sigma^2}\Big) \overset{\text{def}}{\equiv} \exp\Big(-\frac{E^{\text{DM}}(\mathbf{x}_\tau, \tau)}{2\tau\sigma^2}\Big)$$

Diffusion Models as DenseAM

$$p(\mathbf{x}_\tau, \tau) = \mathbb{E}_{\mathbf{y}\sim \text{data}}\Bigg[\frac{1}{(2\pi\sigma^2\tau)^{\frac{N}{2}}} \exp\Big(-\frac{\|\mathbf{x}_\tau - \mathbf{y}\|_2^2}{2\tau\sigma^2}\Big)\Bigg]$$

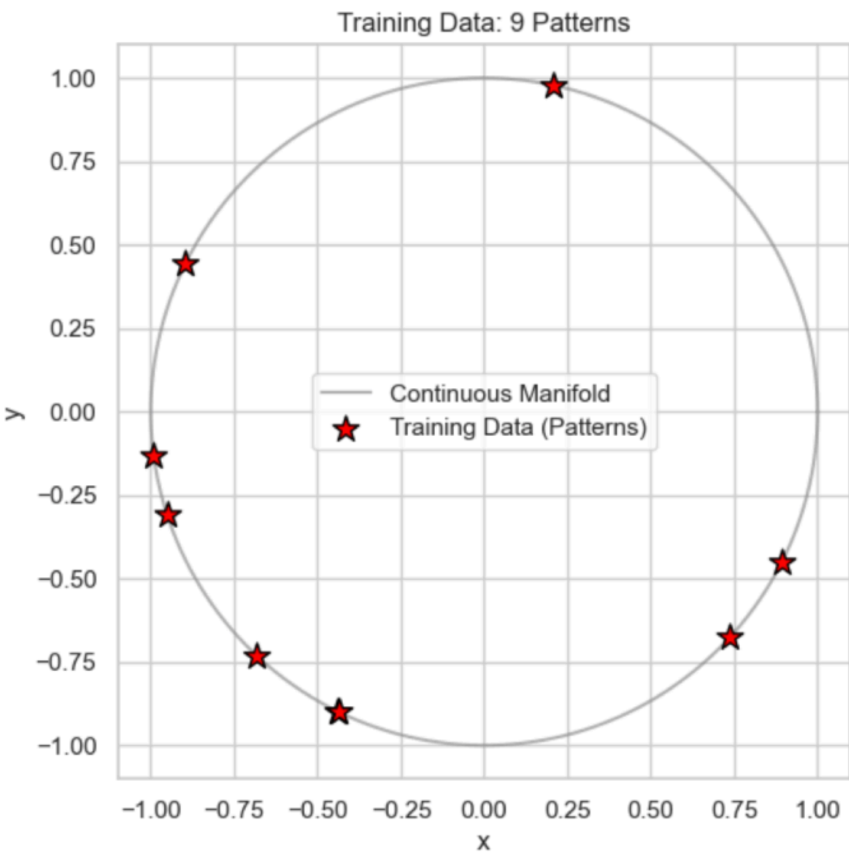
$$p(\mathbf{y}) = \frac{1}{K}\sum_{\mu=1}^{K} \delta^{(N)}(\mathbf{y} - \boldsymbol{\xi}^\mu)$$

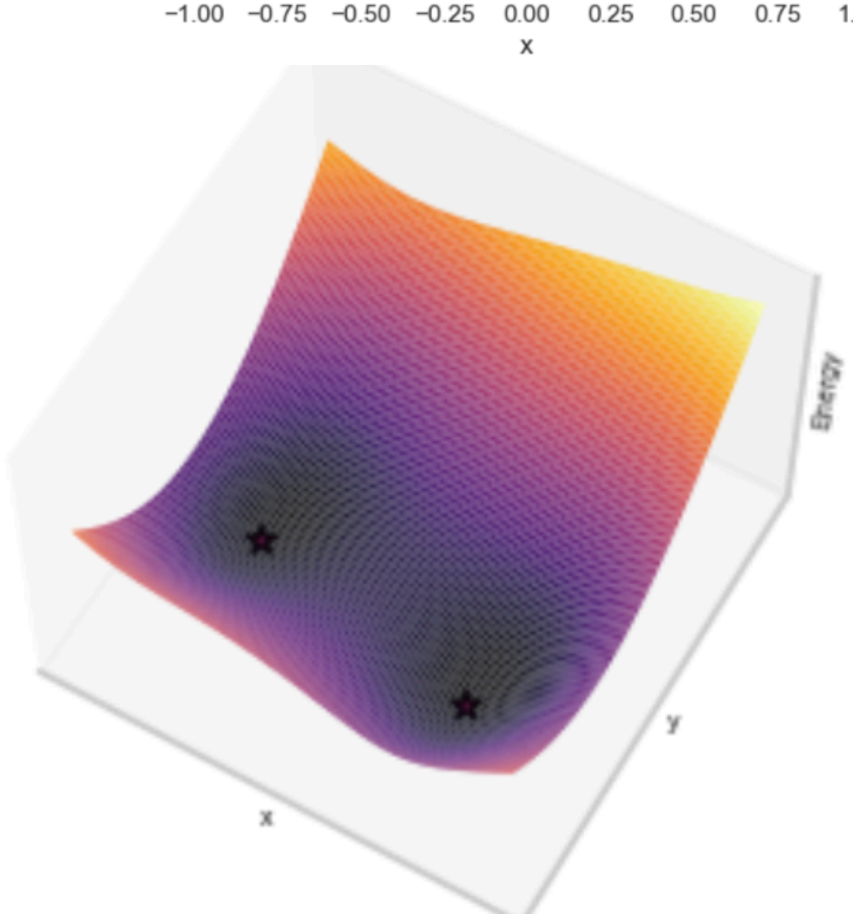
$$p(\mathbf{x}_\tau, \tau) \approx \frac{1}{K}\sum_{\mu=1}^{K} \frac{1}{(2\pi\sigma^2\tau)^{\frac{N}{2}}} \exp\Big(-\frac{\|\mathbf{x}_\tau - \boldsymbol{\xi}^\mu\|_2^2}{2\tau\sigma^2}\Big) \stackrel{\text{def}}{\equiv} \exp\Big(-\frac{E^{\text{DM}}(\mathbf{x}_\tau, \tau)}{2\tau\sigma^2}\Big)$$

$$E^{\text{DM}}(\mathbf{x}_\tau, \tau) = -2\tau\sigma^2 \log\Bigg[\sum_{\mu=1}^{K} \exp\Big(-\frac{\|\mathbf{x}_\tau - \boldsymbol{\xi}^\mu\|_2^2}{2\tau\sigma^2}\Big)\Bigg]$$

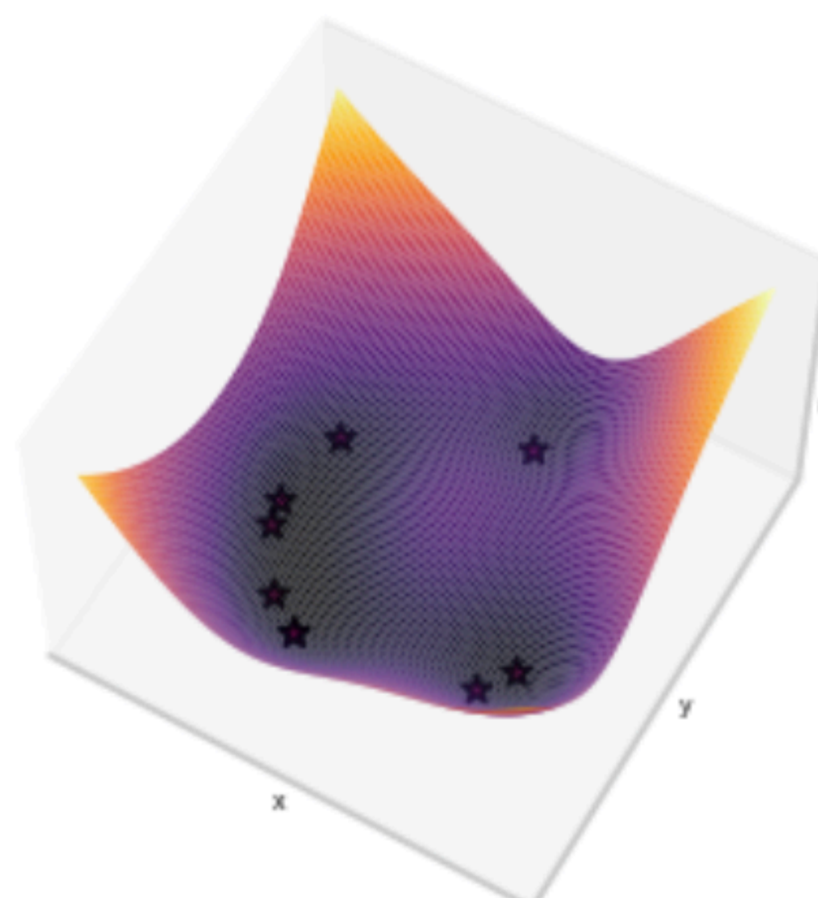
$$E^{\text{AM}}(\mathbf{x}) = -\beta^{-1} \log\Bigg[\sum_{\mu=1}^{K} \exp\Big(-\beta\|\mathbf{x} - \boldsymbol{\xi}^\mu\|_2^2\Big)\Bigg]$$

Diffusion in 2D as an Associative Memory

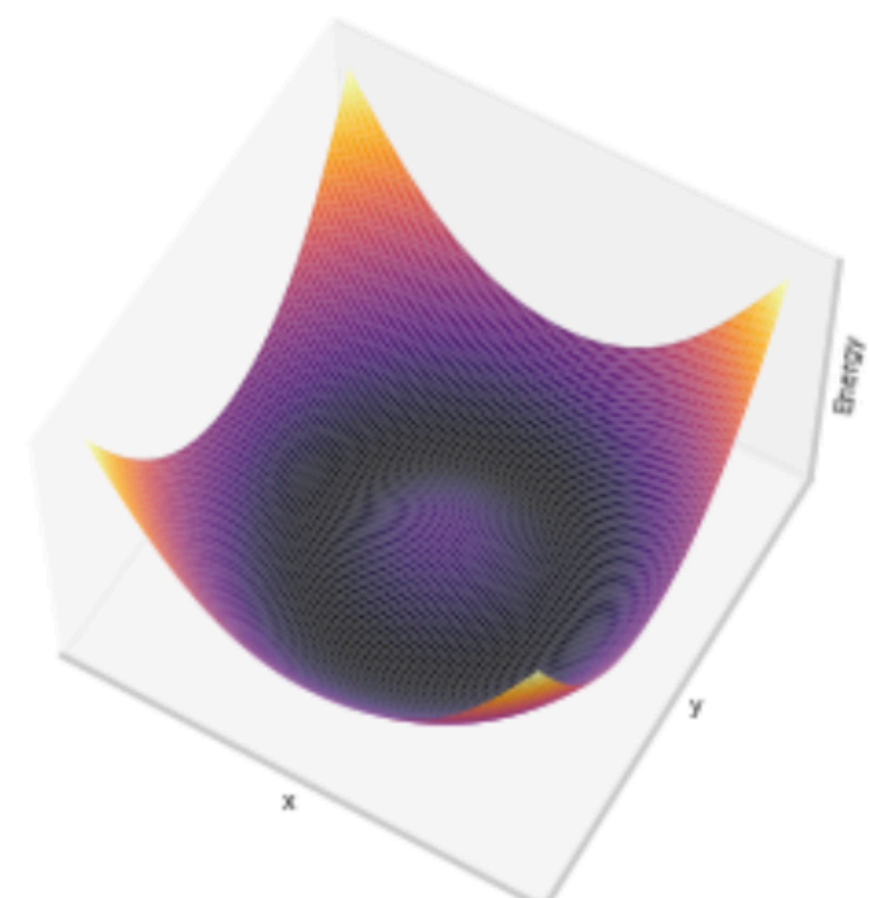




K = 2

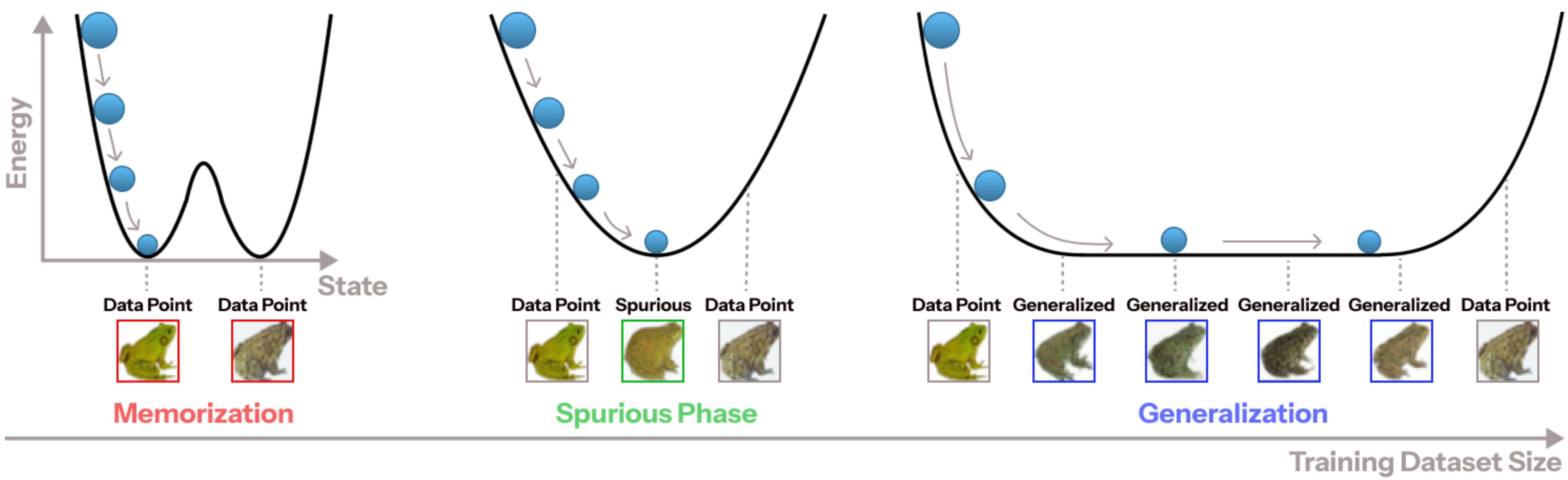


K = 9



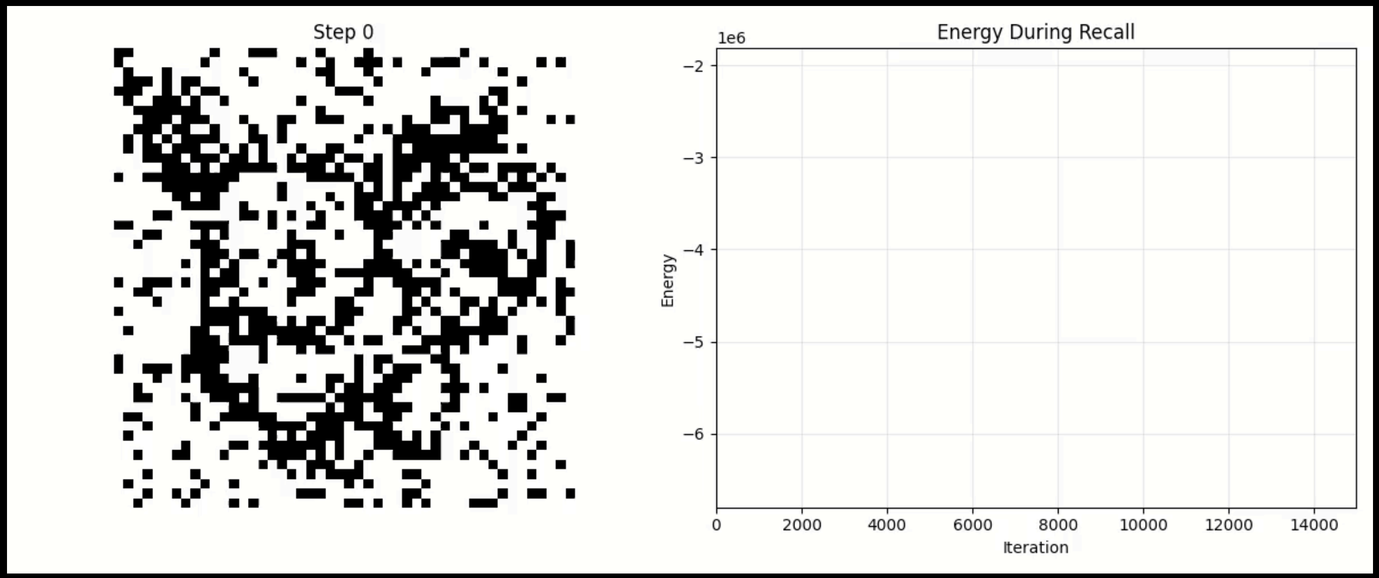
K = 1000

AM-based description of DM predicts existence of spurious states



B.Pham, et al.,Memorization to Generalization: Emergence of Diffusion Models from Associative Memory Networks, 2025

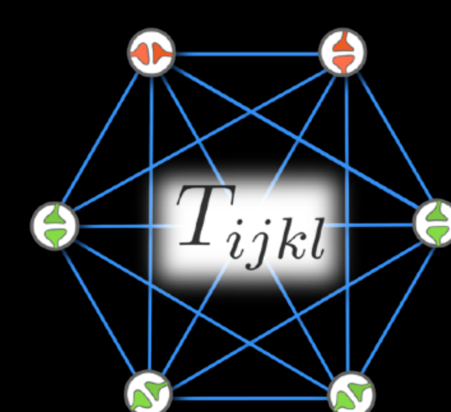
Diffusion models are Dense Associative Memories above the critical memory storage capacity



Modern Methods in Associative Memory

Dmitry Krotov   *Benjamin Hoover*   *Parikshit Ram*

$$E = -\sum_{\mu=1}^{K} F\Big(\sum_{i=1}^{D} \xi_i^{\mu} \sigma_i\Big)$$



Diffusion models are energy-based Associative Memories: neural network encodes the gradient of the energy

Conclusions

- Dense Associative Memory perspective on DMs is a useful theoretical tool.
- Spurious states in DMs are real. They represent a new phase among generated samples that has been completely overlooked by the mainstream CS community.
- Misremembering can be mathematically conceptualized as a formation of spurious states.
- Emergence of spurious states is the earliest sign of creativity in DMs.

$$E = -\sum_{\mu=1}^{K} F\Big(\sum_{i=1}^{N} \xi_i^{\mu} \sigma_i\Big)$$