

# Discrete Codes from Continuous Substrates: A Simulation of Noise-Induced Symmetry Breaking

Ian Todd

*University of Sydney*

itod2305@uni.sydney.edu.au

November 27, 2025

## Abstract

We present a minimal simulation demonstrating that discrete symbolic codes emerge spontaneously when continuous data passes through a noisy bandwidth-limited channel. The simulation reveals that neither the dimensional bottleneck nor channel noise alone produces discretization—both ingredients are required for the topological phase transition from continuous to discrete representation. We describe the simulation architecture, present control experiments isolating the causal factors, and discuss implications for categorical perception, neural coding, and the emergence of digital information in biological systems. The simulation provides a concrete, reproducible demonstration of how “symbols emerge from substrates” through information-theoretic constraints.

## 1 Introduction

How do discrete symbols emerge from continuous physical substrates? This question spans cognitive science (how do brains form discrete concepts?), linguistics (how did discrete language emerge?), and origin-of-life research (how did digital genetic information arise from analog chemistry?).

We approach this question computationally, constructing a minimal simulation that exhibits spontaneous discretization. The goal is not to model any specific biological system, but to identify the *sufficient conditions* for code formation—the minimal ingredients that, when combined, inevitably produce discrete representations.

Our central finding is that two ingredients are required:

1. **A dimensional bottleneck:** Information must pass through a channel with fewer degrees of freedom than the input
2. **Channel noise:** The transmission is corrupted by stochastic perturbations

Neither alone produces discretization. The bottleneck without noise preserves continuous structure. Noise without a bottleneck leaves room for continuous codes. But together, they force the system into a regime where only discrete codes can reliably transmit information.

This paper describes what we simulated, how we simulated it, and what it might mean.

## 2 What We Simulated

### 2.1 The Input: A Continuous Manifold

We need a continuous input space with clear topology. We chose a ring—a 1-dimensional closed manifold—embedded in high-dimensional space. Concretely:

```
# Generate 2000 points uniformly on a circle
theta = random(0, 2*pi, n_samples=2000)
ring_2d = [cos(theta), sin(theta)]

# Embed in 512 dimensions via random projection
projection = random_orthonormal(2, 512)
data_512d = ring_2d @ projection + noise(0.05)
```

The input is *continuous*—points can take any value along the ring. There are no discrete categories in the data itself. We use color (hue) as a visual metaphor: the ring represents a continuous color wheel embedded in a high-dimensional feature space.

For *evaluation only*, we partition the ring into 6 equal sectors (like rainbow colors) and ask: do the learned codes preserve this structure? This is measured by the Adjusted Rand Index (ARI) between k-means clusters of learned codes and ground-truth sectors.

## 2.2 The Channel: Encoder-Noise-Decoder

The core architecture is a sender-channel-receiver system:

```
# Encoder: compress 512D to k dimensions
code = encoder(input) # code is k-dimensional

# Channel: add Gaussian noise
noisy_code = code + noise(sigma)

# Decoder: reconstruct 512D from noisy code
output = decoder(noisy_code)

# Training objective: minimize reconstruction error
loss = mean_squared_error(input, output)
```

The encoder and decoder are neural networks (3-layer MLPs with ReLU activations). The channel dimension  $k$  and noise level  $\sigma$  are the experimental variables.

The key insight is that the encoder must learn codes that are *robust to noise*. If two nearby codes get confused after noise corruption, the decoder cannot distinguish them, and reconstruction suffers. The only solution is to space codes apart—creating discrete clusters with “safety margins.”

## 2.3 The Metrics

We measure two quantities:

**Semantic Preservation (ARI):** Do learned codes preserve the ring’s structure? We cluster codes via k-means and compare to 6-sector ground truth.  $\text{ARI} = 1$  means perfect preservation;  $\text{ARI} = 0$  means random.

**Effective Dimensionality (Participation Ratio):**

$$\text{PR} = \frac{(\sum_i \lambda_i)^2}{\sum_i \lambda_i^2} \quad (1)$$

where  $\lambda_i$  are eigenvalues of the code covariance matrix. For data spread evenly across  $d$  dimensions,  $\text{PR} \approx d$ . For data collapsed to a line,  $\text{PR} \approx 1$ .

## 3 The Control Experiment

Our main result comes from a simple control experiment. We ask: what *causes* the discretization?

Table 1: Control experiment: both bottleneck AND noise required for clustering.

Condition	Bottleneck ( $k$ )	Noise ( $\sigma$ )	Result
Bottleneck only	2	0	Continuous ring
Noise only	32	0.5	Continuous ring
<b>Both</b>	<b>2</b>	<b>0.5</b>	<b>Discrete clusters</b>

Figure 1 shows this visually. The top row tells a linear story:

1. **The Input:** A perfect continuous ring (color = position on ring)
2. **Bottleneck Only:** Compressed to  $k = 2$  with no noise—ring topology preserved
3. **Bottleneck + Noise:** Compressed to  $k = 2$  with  $\sigma = 0.5$ —discrete clusters emerge

The bottom row shows that increasing  $k$  (with noise held constant) restores continuous representation. At  $k = 32$ , there is enough “room” in the channel for the ring to survive despite noise.

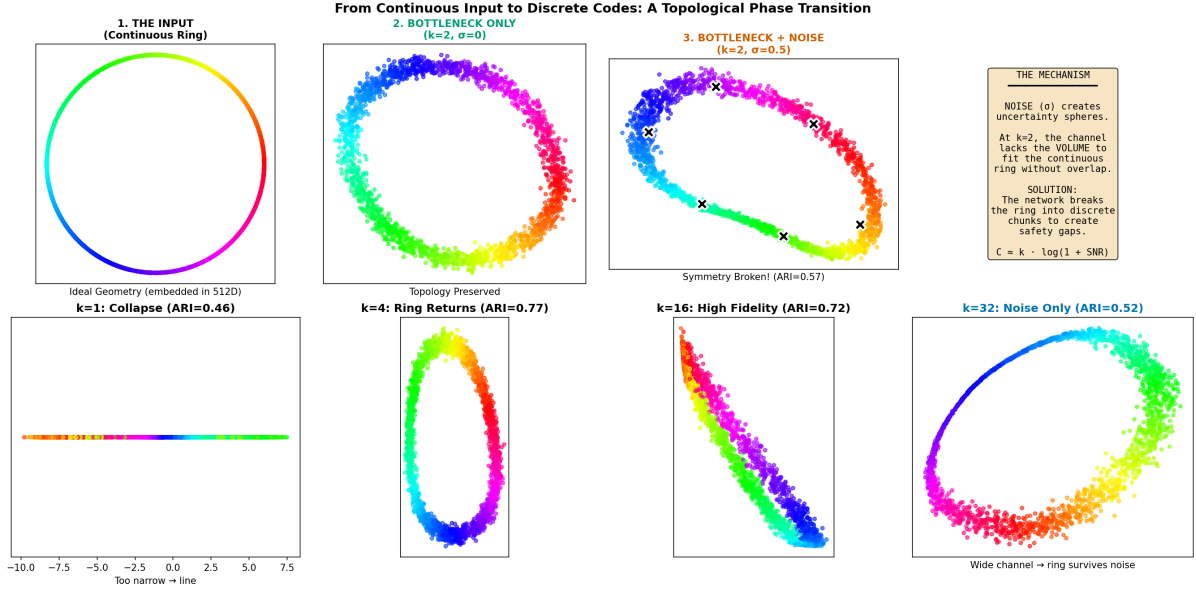


Figure 1: **Visual proof of the two-ingredient hypothesis.** Top row: (1) Continuous input ring, (2) bottleneck alone preserves topology, (3) bottleneck + noise  $\rightarrow$  discrete clusters. Black X’s = learned centroids. Bottom row: As channel dimension increases, ring structure returns despite noise.

## 4 The Phase Transition

Figure 2 shows how semantic preservation varies with channel dimension  $k$ . There are three regimes:

1.  $k = 1$ : **Topological collapse.** The ring is forced onto a line. Adjacent points on the ring may map to distant points on the line. Topology is destroyed.
2.  $k \approx k_c$ : **Critical capacity.** Discrete codes emerge that optimally preserve semantic structure. The system “discovers” that 6 well-separated clusters is the best way to use the limited channel.

3.  $k \gg k_c$ : **Continuous preservation.** With sufficient bandwidth, the ring can be transmitted continuously. Noise is present but doesn't force discretization.

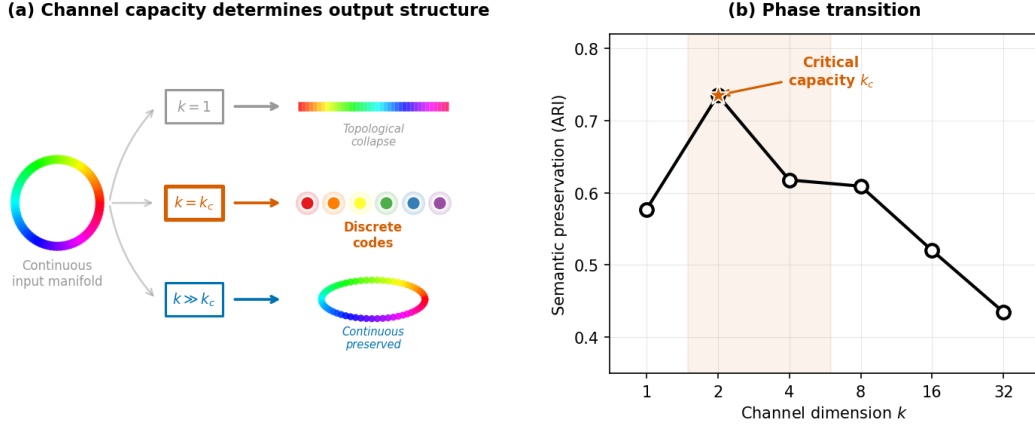


Figure 2: **Phase transition in code structure.** (a) Schematic showing three regimes. (b) Semantic preservation (ARI vs. 6-sector ground truth) as function of channel dimension  $k$ . Peak at critical capacity  $k_c \approx 2$ . Noise  $\sigma = 0.5$ .

This is a *topological phase transition*. The continuous  $O(2)$  rotational symmetry of the ring breaks into discrete  $C_6$  symmetry (6-fold rotation). The order parameter is the discreteness of the code distribution.

## 5 The Mechanism: Sphere Packing

Why does discretization occur at critical capacity? The answer is **sphere packing**.

Channel noise smears each code point into an “uncertainty sphere” of radius  $\sim \sigma$ . If two codes are closer than  $2\sigma$ , their uncertainty spheres overlap, and the decoder cannot reliably distinguish them.

The question becomes: how many non-overlapping spheres can you pack into a  $k$ -dimensional channel? This is the classical sphere-packing problem [Conway and Sloane, 1999]. The answer determines the channel capacity:

$$C \approx k \cdot \log_2(1 + \text{SNR}) \quad (2)$$

where SNR is the signal-to-noise ratio [Shannon, 1948].

For our parameters ( $k = 2$ ,  $\sigma = 0.5$ ), the capacity supports approximately 6 distinguishable codes—exactly what we observe. The network doesn't “know” about sphere packing; it discovers this constraint through gradient descent on reconstruction error.

## 6 What Gets Preserved

When compressing from 512 dimensions to  $k$  dimensions, what does the encoder choose to preserve?

Figure 3 shows that the encoder learns to preserve **high-variance directions**—essentially the top principal components of the input. The ring lives primarily in the first 2 PCs (with  $\text{PC1} \approx \cos \theta$  and  $\text{PC2} \approx \sin \theta$ ), and these are what the encoder captures.

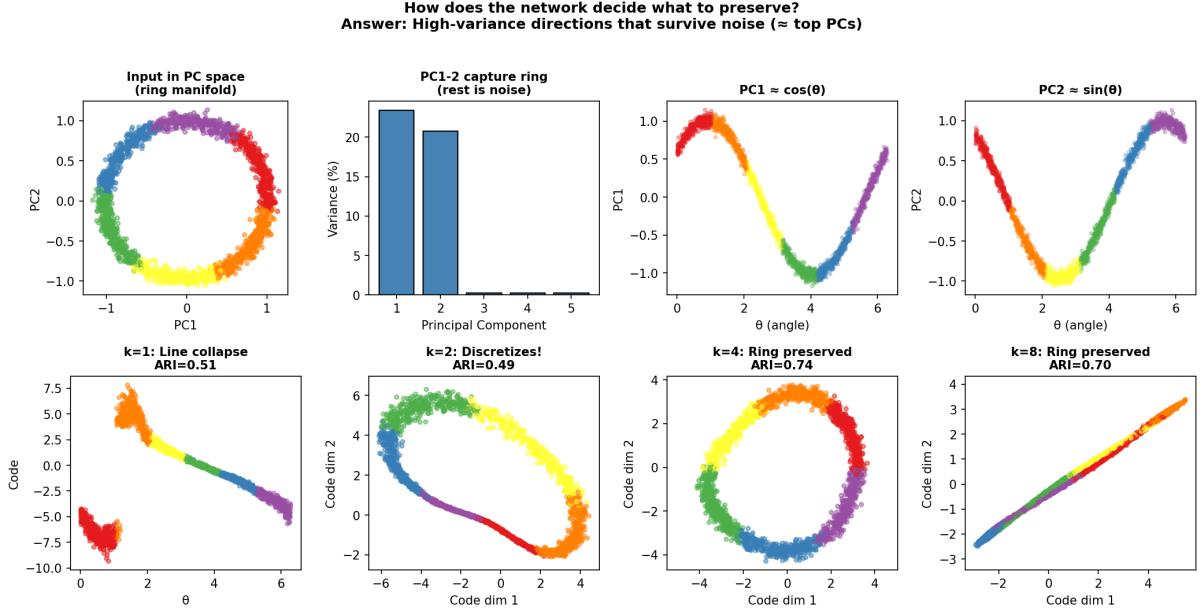


Figure 3: **What gets preserved.** Top: Input lives in first 2 PCs. PC1-2 capture the ring; higher PCs are noise. Bottom: Encoder learns to preserve these high-variance directions. At  $k = 2$  with noise, discretization occurs.

This makes intuitive sense: high-variance directions carry the most information about input structure. Low-variance directions (the remaining 510 dimensions of noise) are discarded as uninformative.

The encoder is essentially learning a *noise-robust PCA*—with the crucial addition that non-linear activations allow it to “tear” the topology when capacity is insufficient.

## 7 Extension: Stochastic Resonance

Our framework provides a geometric interpretation of **stochastic resonance** (SR)—the counterintuitive phenomenon where adding noise *improves* signal detection [Gammaitoni et al., 1998].

Consider an ambiguous signal sitting exactly on the boundary between two code clusters. This signal is effectively *one-dimensional*: it is constrained to the decision boundary between basins.

Adding noise *expands the dimensionality* of the representation. The signal can now fluctuate perpendicular to the boundary, enabling escape into the correct attractor basin.

Figure 4 demonstrates this:

- At zero noise, ambiguous signals are trapped on 1D boundaries  $\rightarrow$  poor decoding
- At intermediate noise, dimensionality expands enough to escape  $\rightarrow$  peak accuracy
- At high noise, random walk dominates  $\rightarrow$  decoding degrades

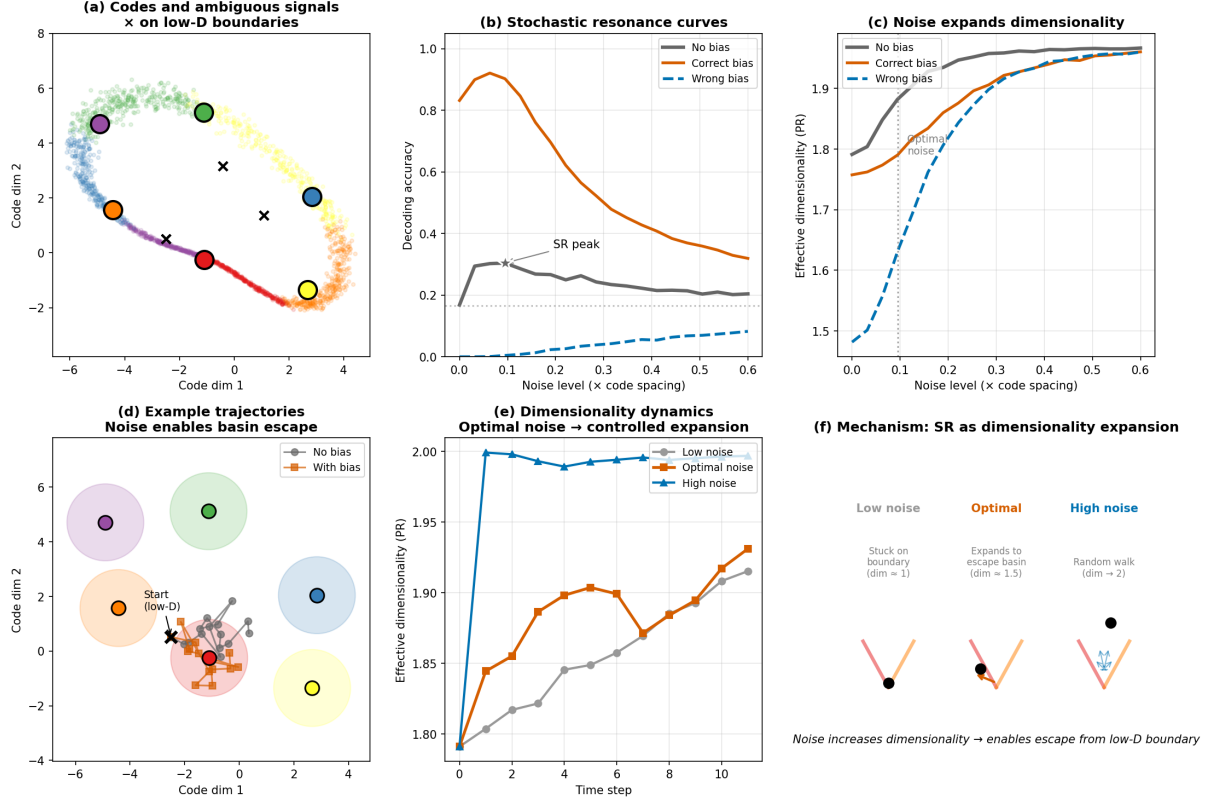


Figure 4: **Stochastic resonance as dimensionality expansion.** (a) Codes and ambiguous test signals. (b) SR curve: accuracy peaks at intermediate noise. (c) Effective dimensionality increases with noise. (d-f) Geometric interpretation.

This reframes SR from an “energy barrier” phenomenon to a **dimensionality** phenomenon: noise provides the extra degrees of freedom needed to escape low-D traps.

## 8 Implications

### 8.1 Categorical Perception

Humans perceive continuous stimuli as discrete categories—colors, phonemes, facial expressions [Harnad, 1987]. Our simulation suggests this may arise from channel capacity limits in neural processing. If sensory information must pass through bandwidth-limited pathways (due to metabolic constraints, noise, or architectural bottlenecks), discretization becomes inevitable.

### 8.2 The Magical Number Seven

Miller’s famous “ $7 \pm 2$ ” limit on working memory capacity [Miller, 1956] may reflect a sphere-packing constraint. If working memory operates through a noisy channel, the number of distinguishable items is determined by  $C \approx k \cdot \log(1 + \text{SNR})$ —not by arbitrary neural architecture.

### 8.3 Language and Symbol Emergence

Discrete tokens are fundamental to language. Our simulation suggests that discretization doesn’t require explicit symbolic machinery—it emerges automatically when information must pass through capacity-limited channels. This has implications for theories of language evolution: discrete symbols may have emerged as the *inevitable* solution to noisy communication.

## 8.4 The Origin of Digital Information

Life uses digital information (genetic sequences) despite operating in an analog chemical medium. How did this transition occur? Our framework suggests that noise + capacity constraints may have forced early self-replicating systems toward discrete coding [Walker and Davies, 2013, Deacon, 2011]. The genetic code may be an instance of noise-induced symmetry breaking.

## 8.5 Sub-Landauer Computing

Landauer’s principle states that erasing one bit of information dissipates at least  $k_B T \ln 2$  of energy [Landauer, 1961]. But our simulation shows that discrete codes can emerge *without explicit erasure*—through the interaction of compression and noise. This suggests a thermodynamic regime below Landauer’s limit where information processing occurs through constraint satisfaction rather than explicit bit operations [Bennett, 1982].

## 9 Limitations

This simulation is deliberately minimal. Limitations include:

- **Static data:** We use a fixed manifold, not dynamical or temporal structure
- **Gaussian noise:** Real channels have more complex noise statistics
- **Feedforward architecture:** Biological systems have recurrence and feedback
- **Single manifold:** The input is a simple ring; real data has more complex topology

These limitations define the scope of the current work. Extending to time-series, recurrent architectures, and complex manifolds is future work.

## 10 Reproducibility

The simulation is implemented in Python using PyTorch. All code and figures are available at [github.com/\[repository\]](https://github.com/[repository]). Key parameters:

- Input: 2000 points, 512 dimensions, ring manifold
- Encoder/Decoder: 3-layer MLP, 256 hidden units, ReLU activation
- Training: Adam optimizer, 150 epochs, MSE loss
- Critical experiment:  $k = 2$ ,  $\sigma = 0.5$

Random seeds are fixed for reproducibility. Running `python code_formation.py` generates all figures.

## 11 Conclusion

We have presented a minimal simulation demonstrating noise-induced symmetry breaking: the emergence of discrete codes from continuous substrates. The key finding is that **both** a dimensional bottleneck **and** channel noise are required—neither alone suffices.

The mechanism is sphere packing under uncertainty. When noise creates uncertainty spheres around each code point, and channel capacity is limited, the only way to maximize information transmission is to space codes discretely with safety margins.

This is not just a computational curiosity. It suggests a fundamental principle: *wherever information must pass through noisy, capacity-limited channels, discrete symbols will emerge*. This may underlie categorical perception, working memory limits, the structure of language, and even the digital nature of genetic information.

Symbols don't need to be built in. They emerge.

## References

- Charles H Bennett. The thermodynamics of computation—a review. *International Journal of Theoretical Physics*, 21(12):905–940, 1982.
- John H Conway and Neil JA Sloane. *Sphere packings, lattices and groups*, volume 290. Springer Science & Business Media, 1999.
- Terrence W Deacon. Incomplete nature: How mind emerged from matter. *WW Norton & Company*, 2011.
- Luca Gammaritoni, Peter Hänggi, Peter Jung, and Fabio Marchesoni. Stochastic resonance. *Reviews of Modern Physics*, 70(1):223, 1998.
- Stevan Harnad. Categorical perception: The groundwork of cognition. *Cambridge University Press*, 1987.
- Rolf Landauer. Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development*, 5(3):183–191, 1961.
- George A Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2):81, 1956.
- Claude E Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.
- Sara Imari Walker and Paul CW Davies. The algorithmic origins of life. *Journal of the Royal Society Interface*, 10(79):20120869, 2013.