

# Neural Manifold Geometry Encodes Feature Fields



Julian Yocum, Cam Allen,  
Bruno Olshausen, Stuart Russell



Center for  
Human-Compatible  
Artificial  
Intelligence

Berkeley  
UNIVERSITY OF CALIFORNIA

1

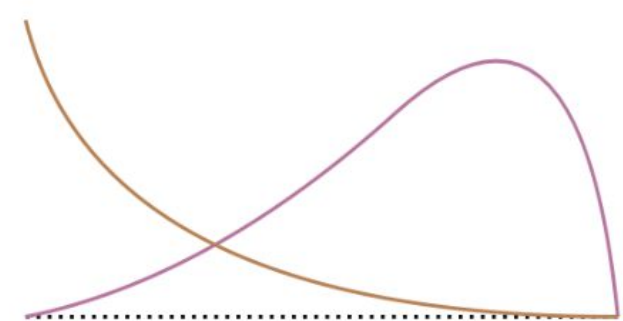
How do neural networks represent **functions over spaces**?

$$f_a(x_1, I)$$

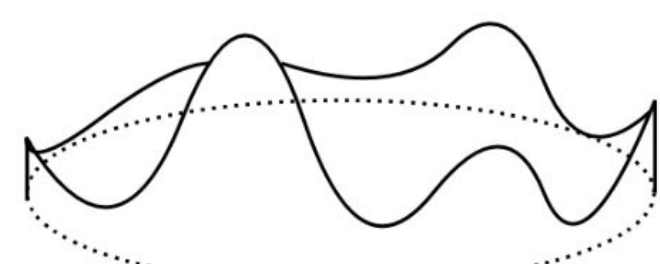
$$f_a(x_2, I)$$

$$f_b(x_3, S^1)$$

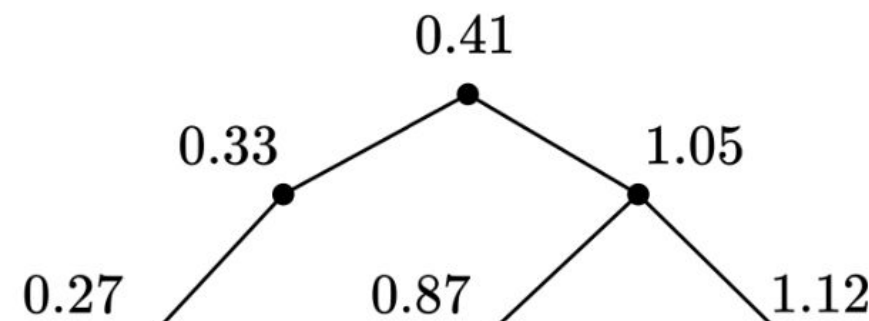
$$f_c(x_4, G)$$



$$Z = I = [0, 1]$$



$$Z = S^1$$

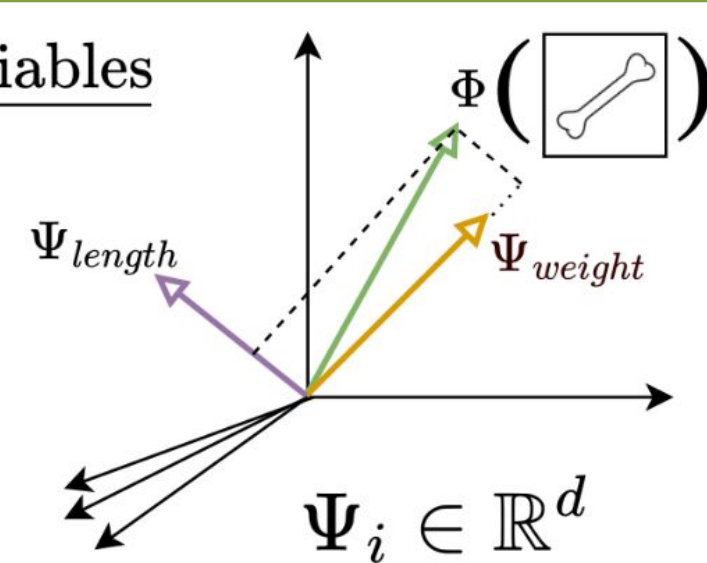


$$Z = G$$

Given a **domain space  $Z$** , a **feature field** is a distribution of functions over  $Z$ , e.g. value function, belief distribution.

While **linear probes** recover **scalars** from activations, we use **linear field probes** to recover **functions**.

Feature Variables



$$f_i(x) = \Phi(x) \cdot \Psi_i$$

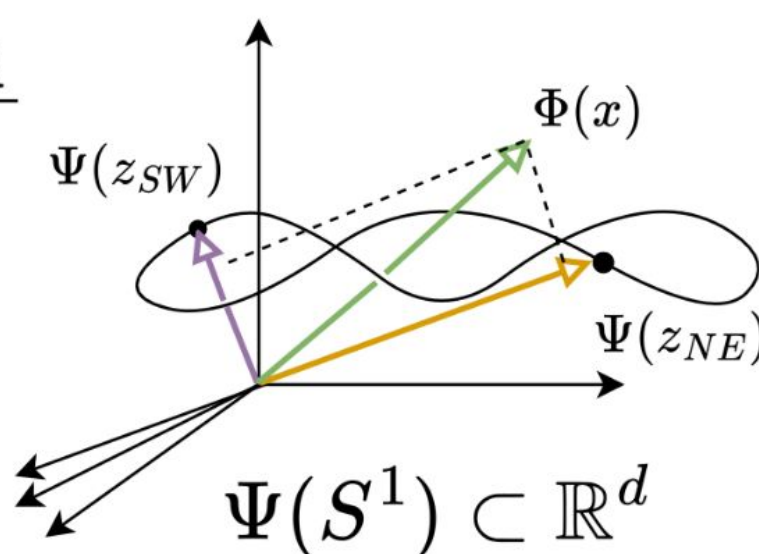
$$\Phi(\text{bone}) \cdot \Psi_{\text{weight}} = 1.03$$

$f_{\text{weight}}(\cdot) = \text{How heavy (kg)?}$

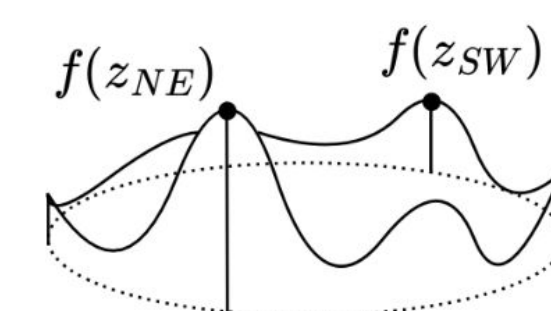
$$\Phi(\text{bone}) \cdot \Psi_{\text{length}} = 0.26$$

$f_{\text{length}}(\cdot) = \text{How long (m)?}$

Feature Field



$$f(x, z) = \Phi(x) \cdot \Psi(z)$$



$$f(x, S^1)$$

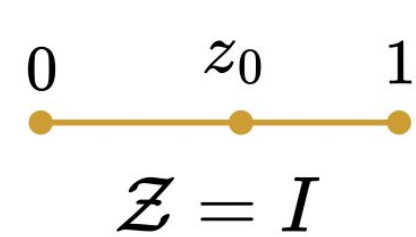
$f(\cdot, z) = \text{Pr}(\text{bone in direction } z)$

2

3

Neural networks represent a feature field by embedding its domain space into activation space.

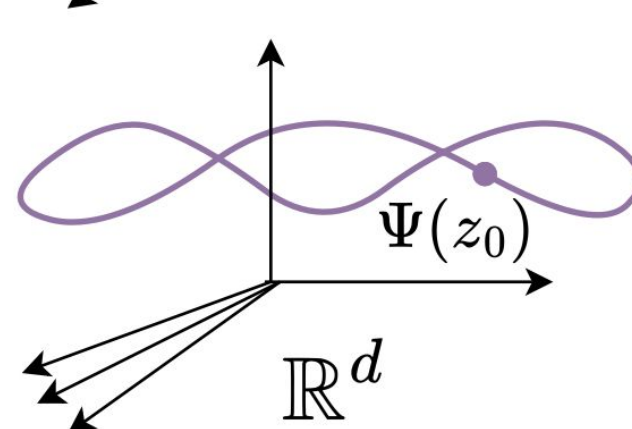
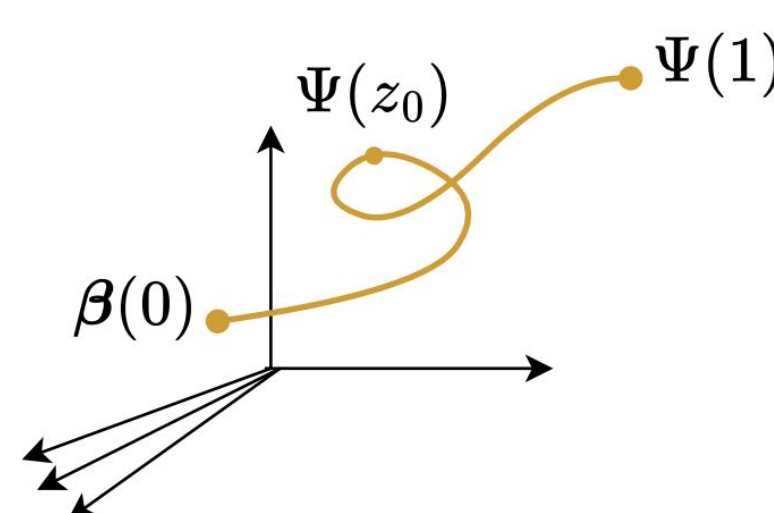
Domain Space  $Z$



$$Z = I$$

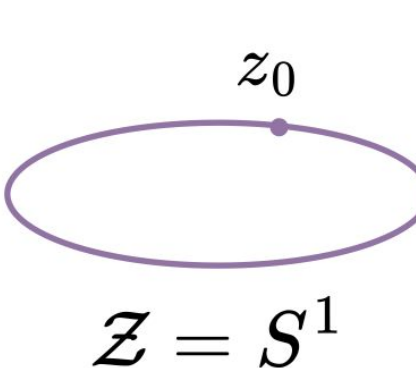
$$\Psi : I \rightarrow \mathbb{R}^d$$

Domain Embedding  $\Psi(Z)$



$$\mathbb{R}^d$$

$$\Psi : S^1 \rightarrow \mathbb{R}^d$$



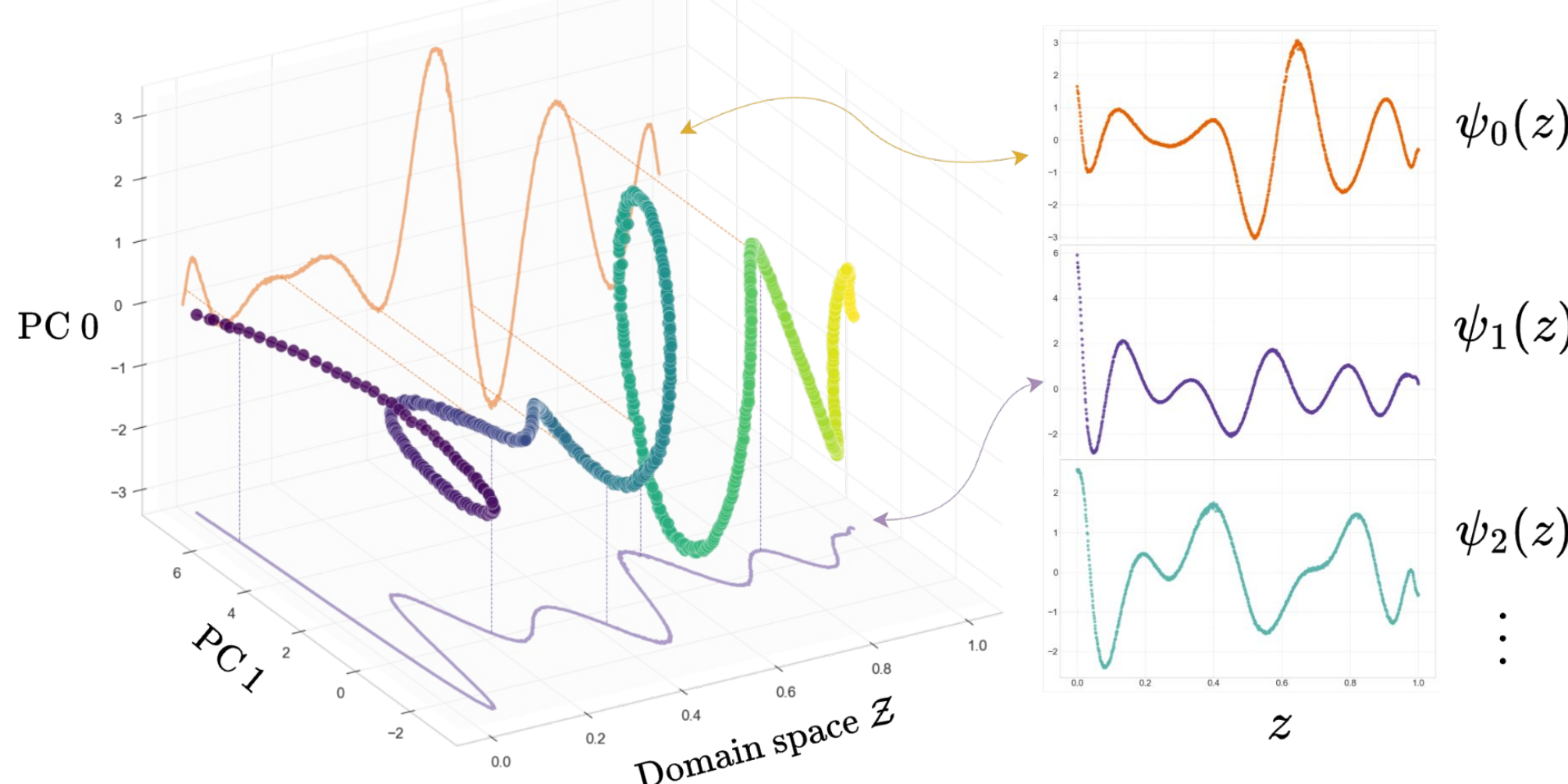
$$Z = S^1$$

This **domain embedding** preserves the topology of the latent domain space for continuous feature fields.

The geometry of the **domain embedding encodes basis functions over  $Z$** .

Domain Embedding Geometry

Basis Functions



Neural networks represent functions over  $Z$  as linear combinations of these basis functions.

$$f(\phi, x) = \sum_{i=1}^d a_i(x) \psi_i(\phi)$$

$$f(x, z) = a_0(x) \cdot \psi_0(z) + a_1(x) \cdot \psi_1(z) + a_2(x) \cdot \psi_2(z) + \dots$$

4