

Neural Manifold Geometry Encodes Feature Fields



Julian Yocum, Cam Allen,
Bruno Olshausen, Stuart Russell



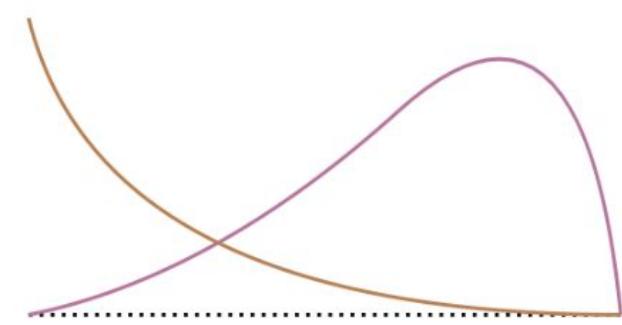
Center for
Human-Compatible
Artificial
Intelligence

Berkeley
UNIVERSITY OF CALIFORNIA

1

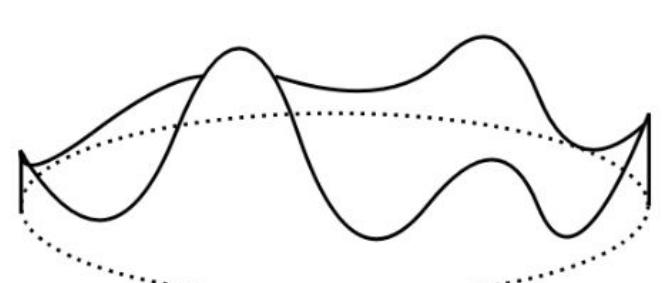
How do neural networks represent **functions over spaces**?

$$f_a(x_1, I) \quad f_a(x_2, I)$$



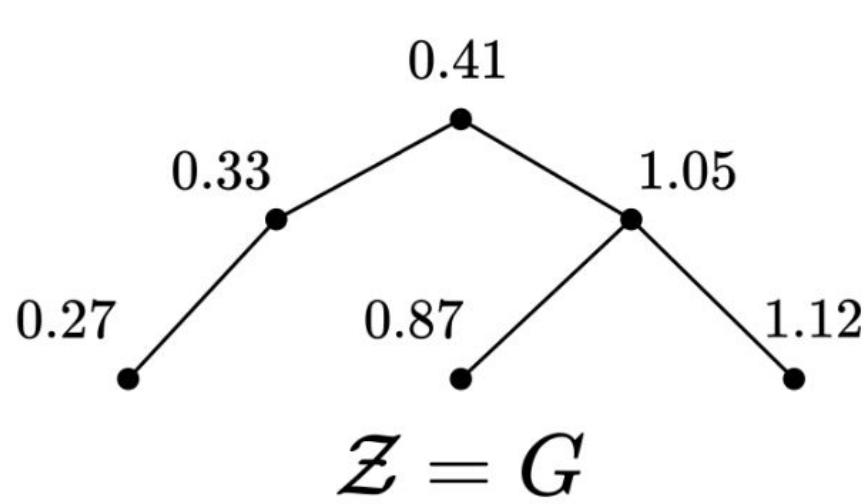
$$Z = I = [0, 1]$$

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



$$Z = S^1$$

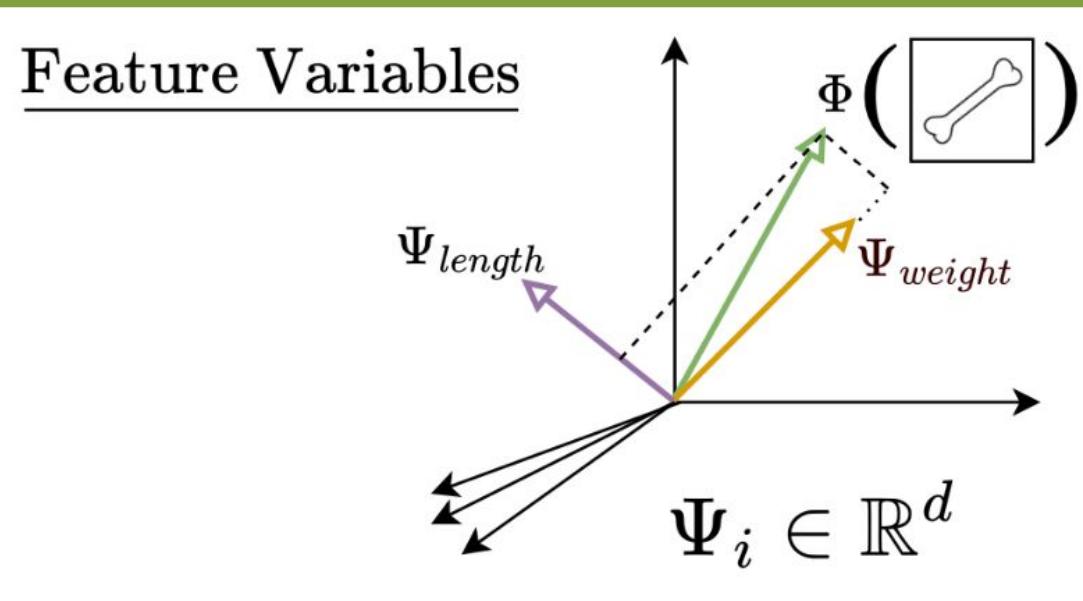
$$f_c(x_4, G)$$



$$Z = G$$

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

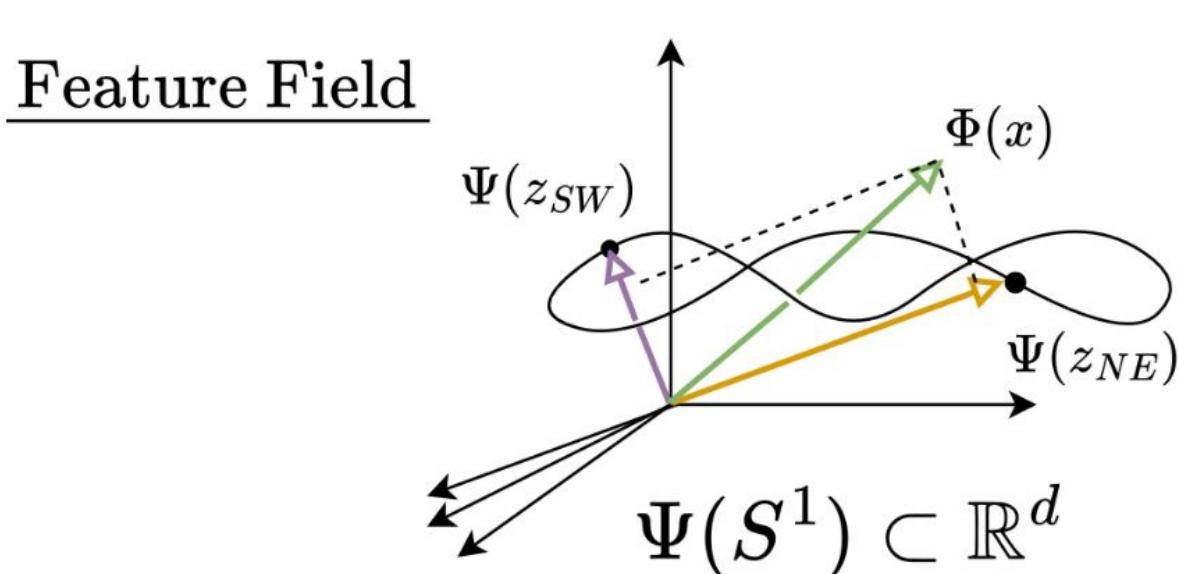
While **linear probes** recover **scalars** from activations, we use **linear field probes** to recover **functions**.



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

$$\Phi(\text{bone icon}) \cdot \Psi_{\text{weight}} = 1.03 \quad f_{\text{weight}}(\cdot) = \text{How heavy (kg)?}$$

$$\Phi(\text{bone icon}) \cdot \Psi_{\text{length}} = 0.26 \quad f_{\text{length}}(\cdot) = \text{How long (m)?}$$



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

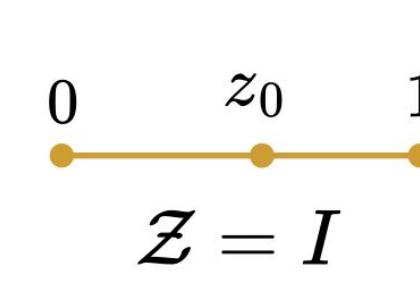
$$f(z_{NE}) \quad f(z_{SW})$$

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

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

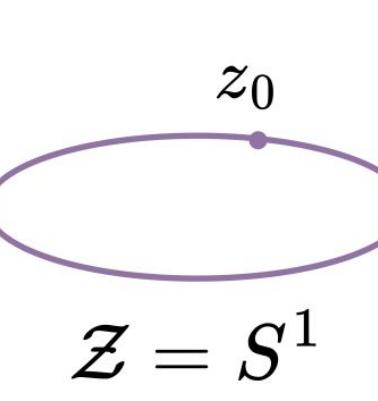
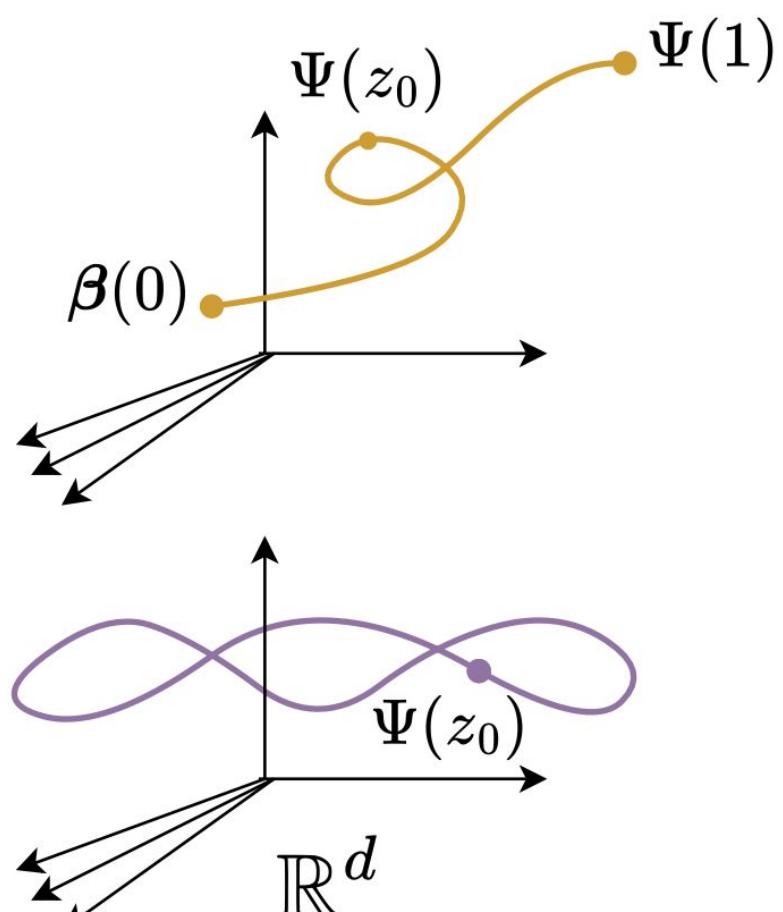
3 Neural networks represent a feature field by embedding its domain space into activation space.

Domain Space Z

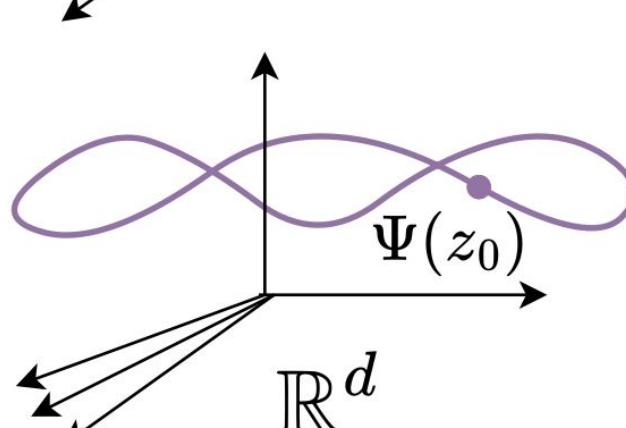


Domain Embedding $\Psi(Z)$

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

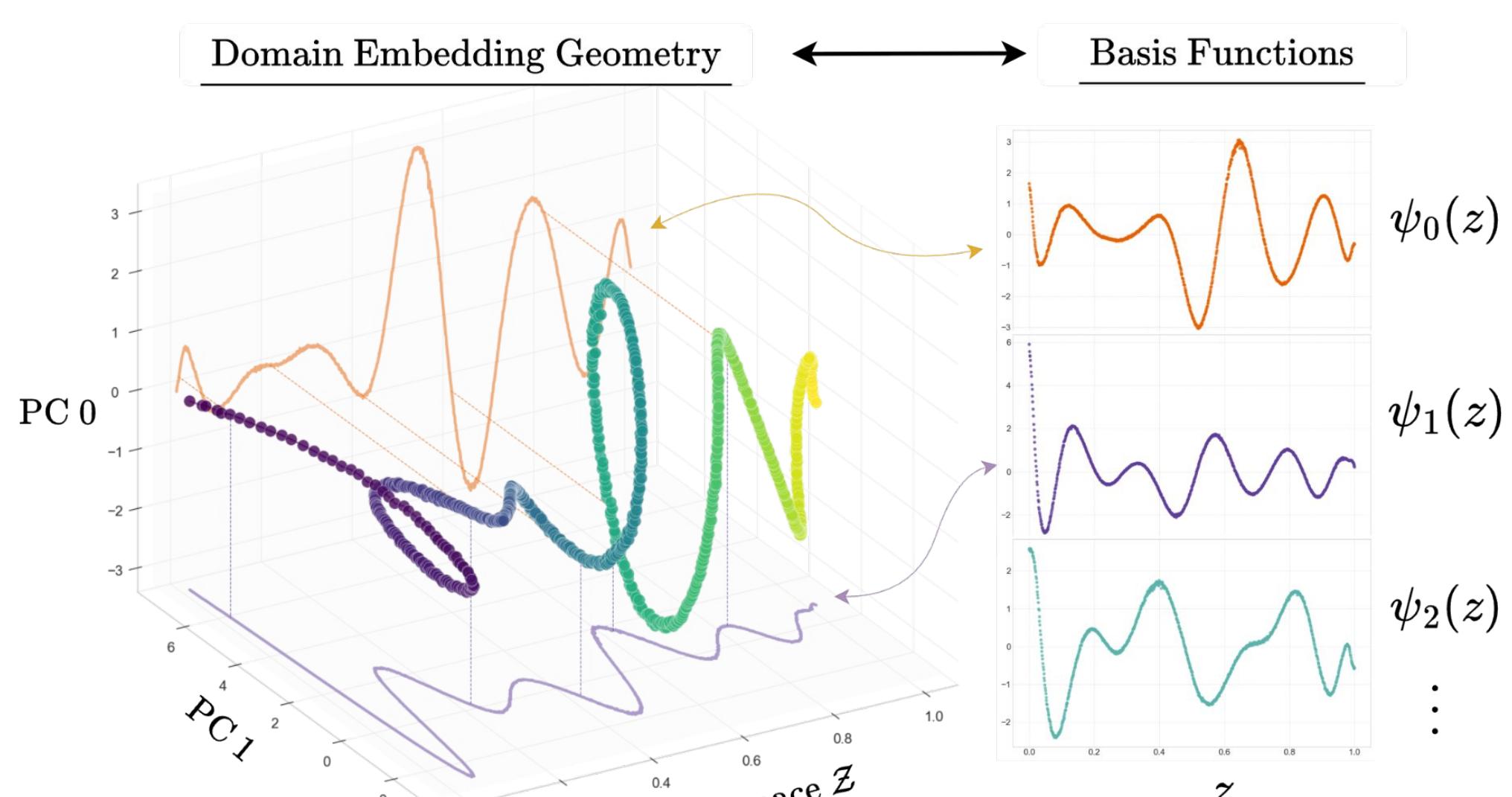


$$\Psi : S^1 \rightarrow \mathbb{R}^d$$



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

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



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

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

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