# Neural Manifold Geometry Encodes Feature Fields

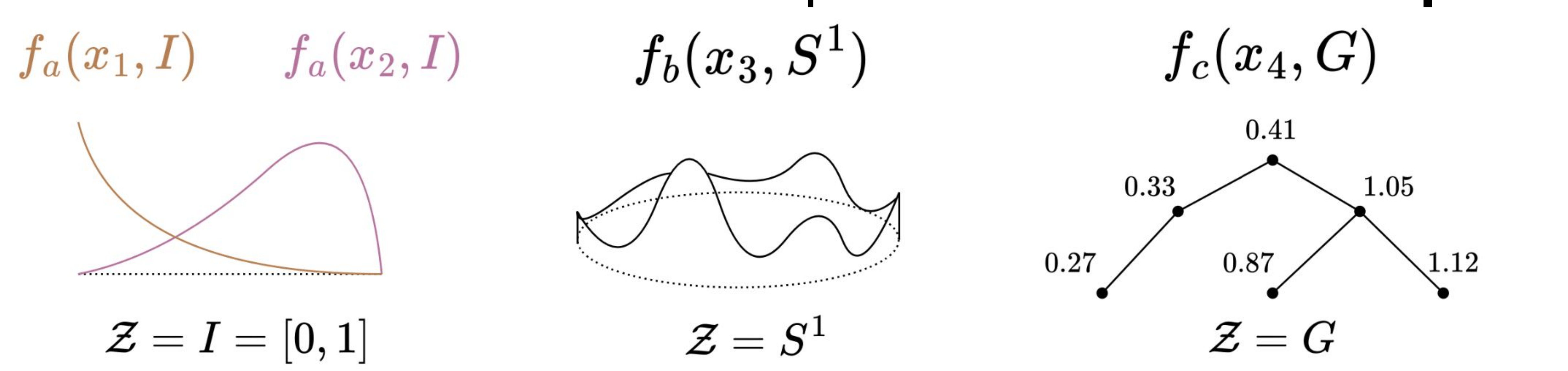
**Julian Yocum**, Cam Allen,
Bruno Olshausen, Stuart Russell

CHAI Center for Human-Compatible Artificial Intelligence

Berkeley UNIVERSITY OF CALIFORNIA

## 1

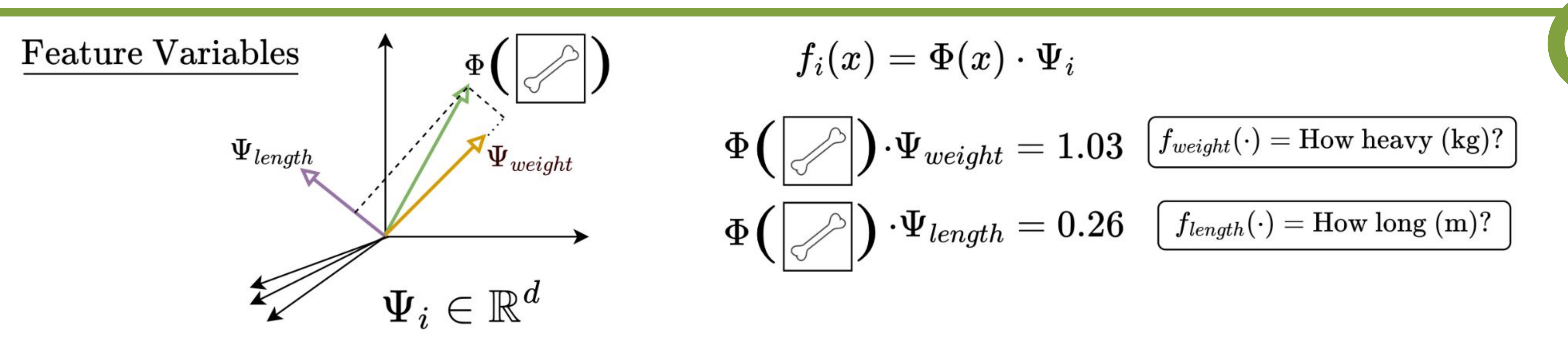
How do neural networks represent **functions** over **spaces**?

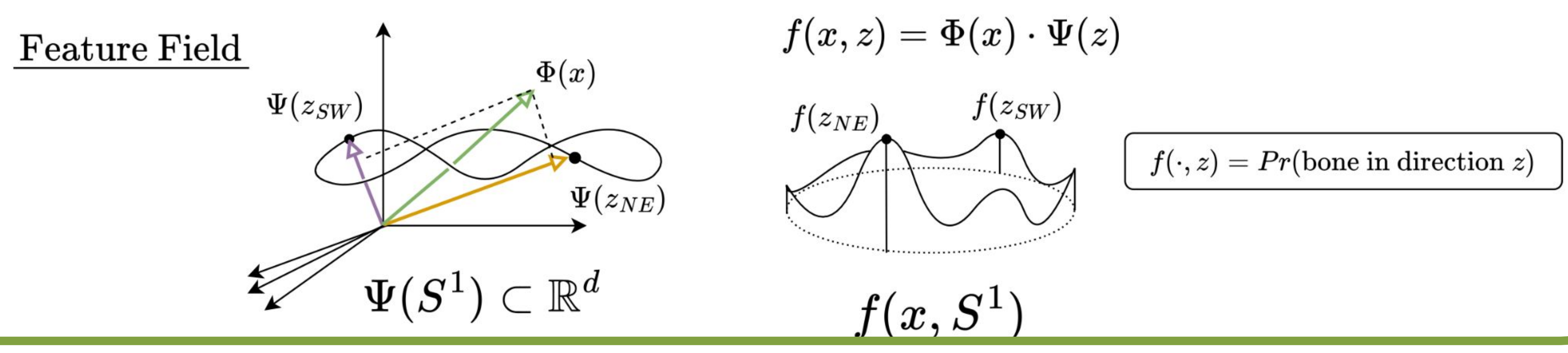



Given a ***domain space* Z**, a ***feature field*** is a distribution of functions over **Z**, e.g. value function, belief distribution.

## 2

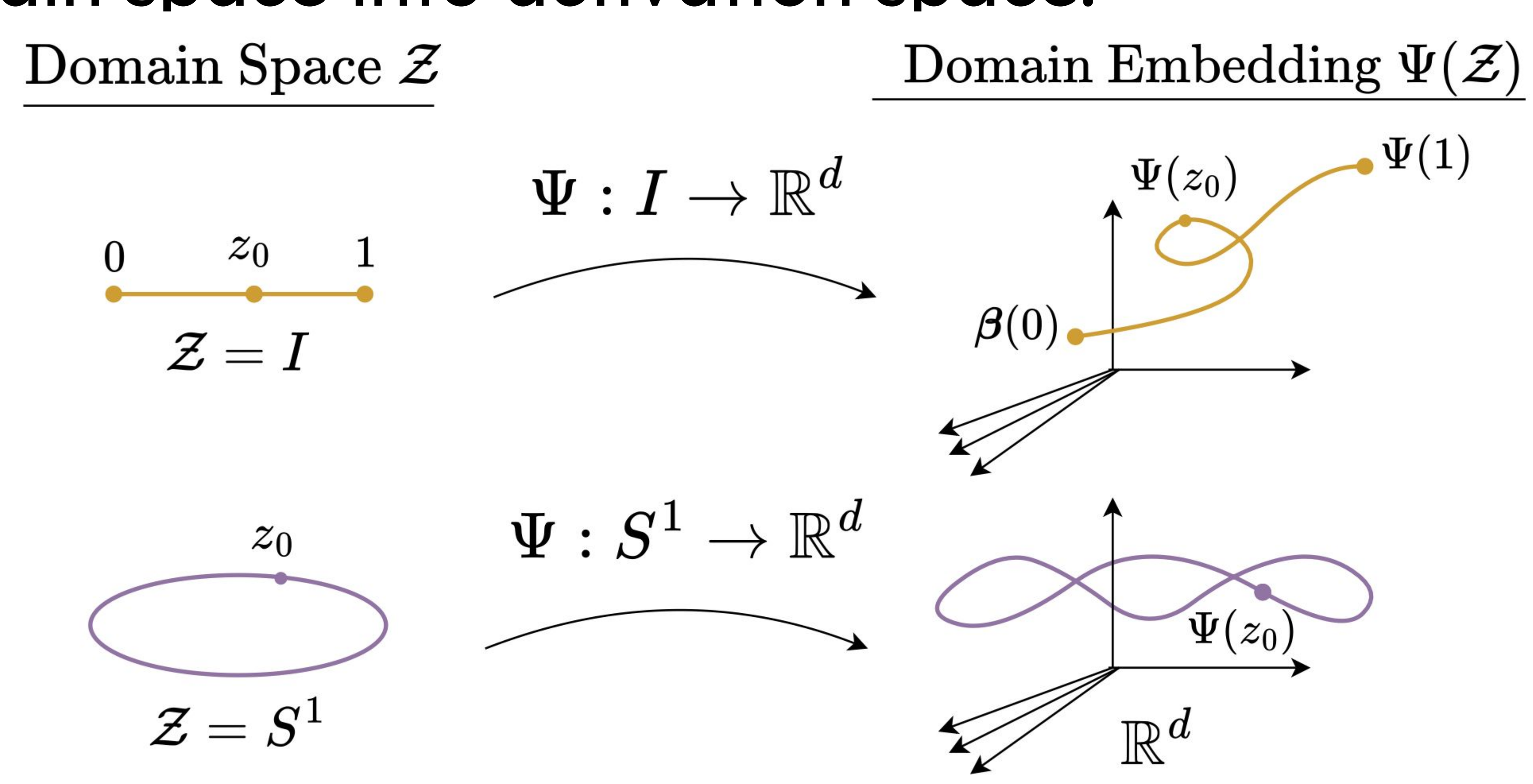
While **linear probes** recover **scalars** from activations, we use ***linear field probes*** to recover **functions.**




## 3

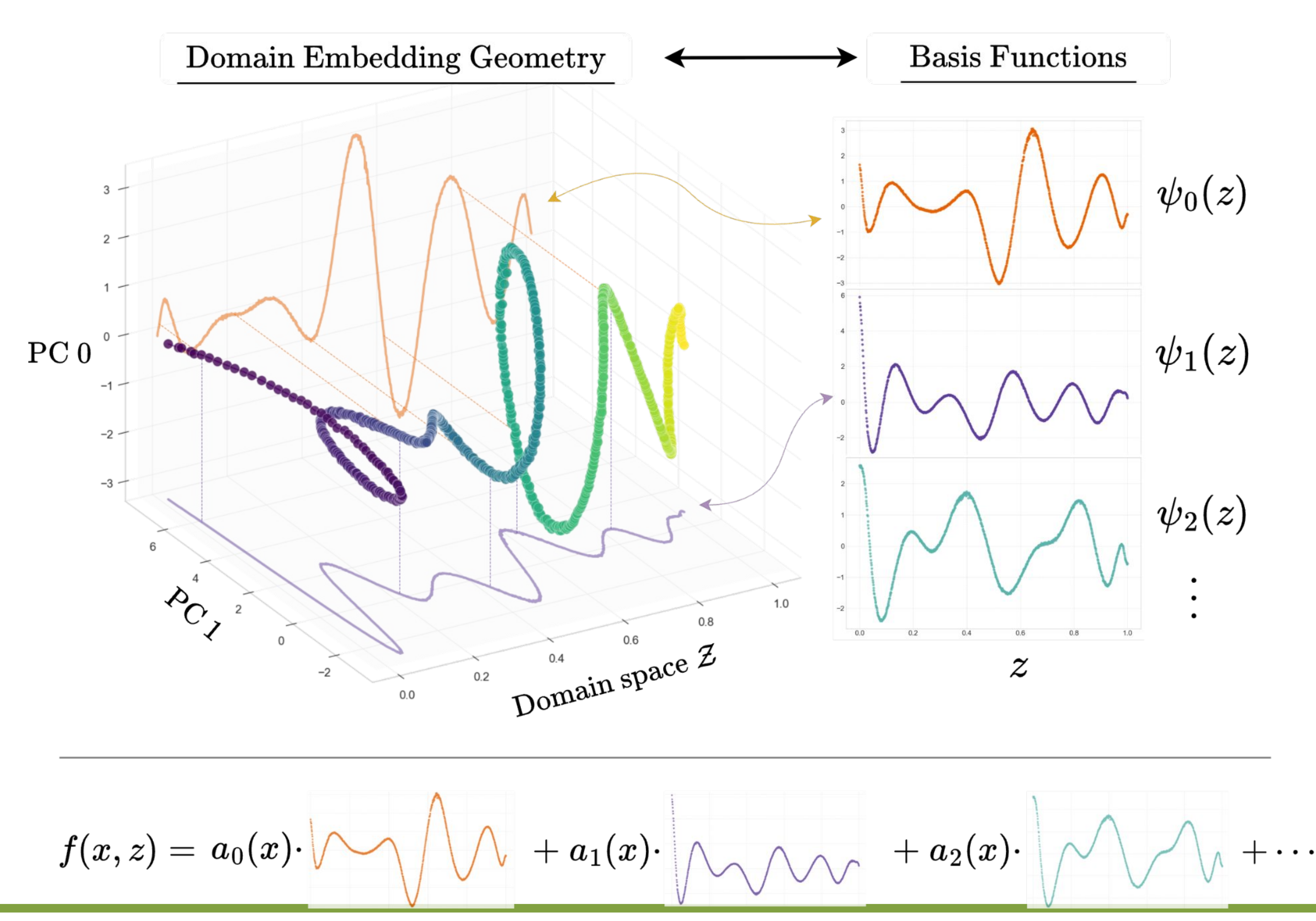
Neural networks represent a feature field by embedding its domain space into activation space.




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.

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