# Defining REPRESENTATIONS

Given:

- a neural network trained to predict $\hat{y}_i \in \mathcal{Y}$ given $x_i \in \mathcal{X}$,
- and a CUT $C$ of that network
  - (a set of neurons s.t. every path from input to output has to intersect it),

a REPRESENTATION is the mapping from $\mathcal{X}$ to $\mathcal{H}$, where

- $\mathcal{H}$ is the vector space of observed activations of neurons in $C$ (in some arbitrary fixed order).