Monosemanticity spectrum (decoder.layer3) Monosemanticity
O O O O O
O O O O O 1500 2000 2500 Neuron index