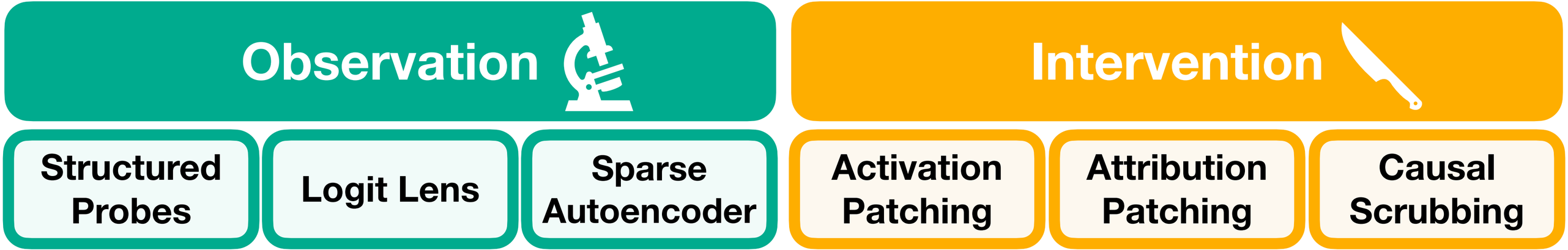
Observation
Structured Probes
Logit Lens
Sparse Autoencoder
Intervention
Activation Patching
Attribution Patching
Causal Scrubbing