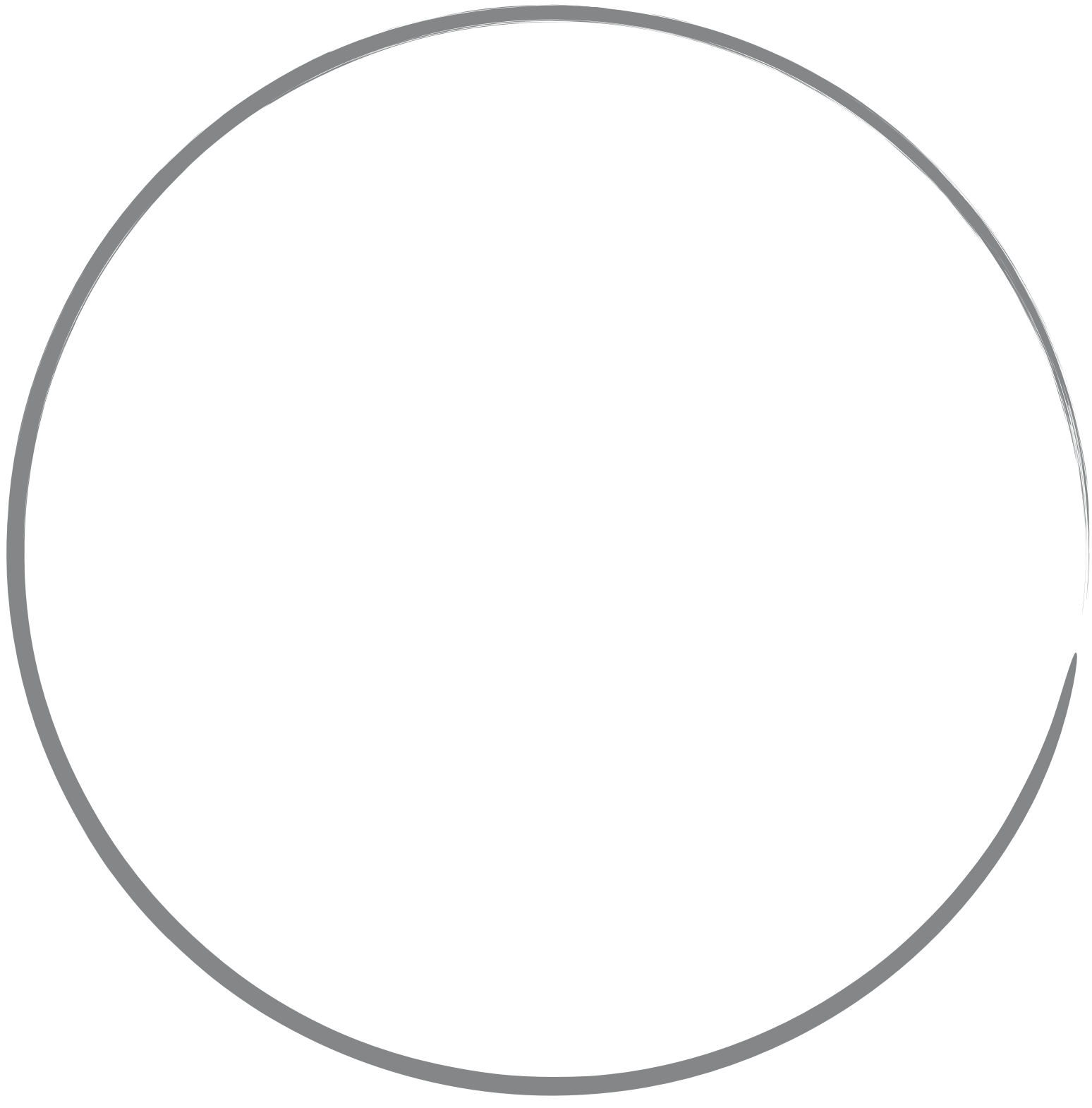


Post-hoc Interpretability

Noisy explanations

Not robust

Similar inputs often yield (very) different explanations

AM + Jaakkola, WHI'18

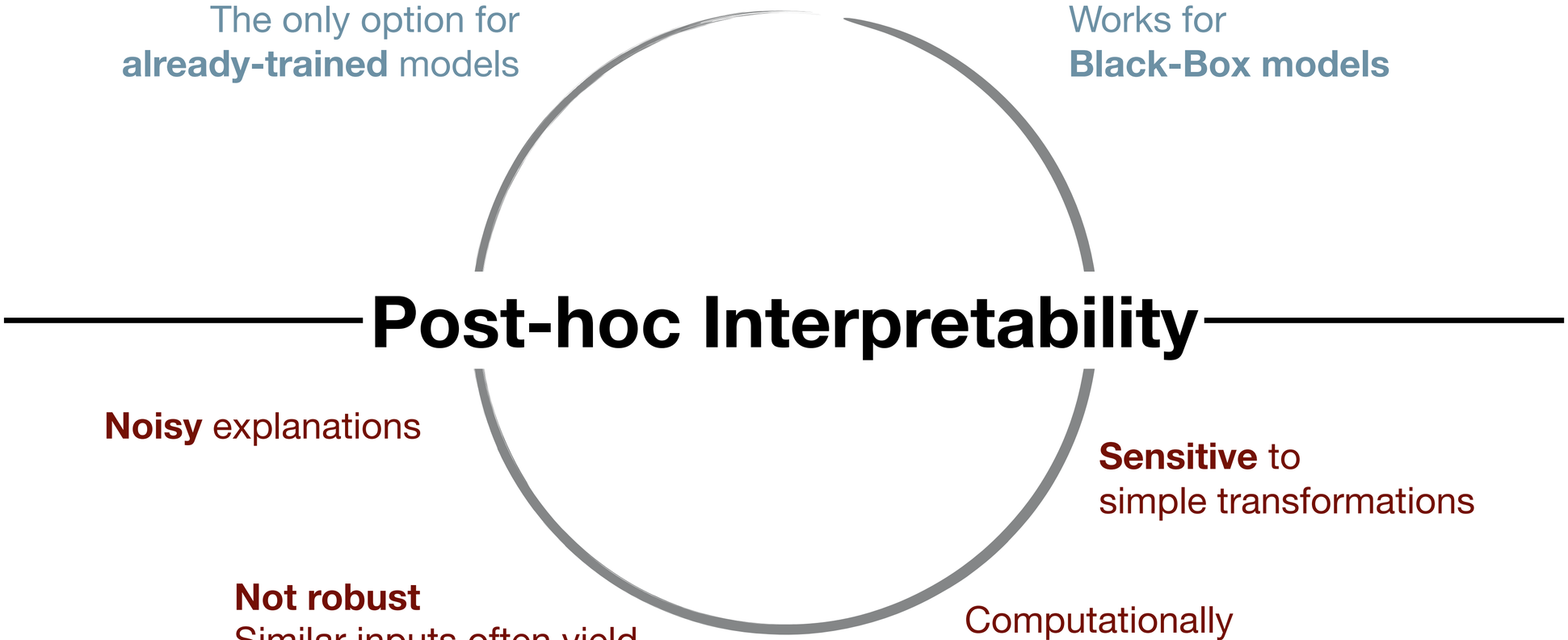
Sensitive to simple transformations

Computationally **expensive**

Works for

Black-Box models

The only option for
already-trained models


The only option for
already-trained models
Works for
Black-Box models
Post-hoc Interpretability
Noisy explanations
Sensitive to
simple transformations
Not robust
Similar inputs often yield
(very) different explanations
Computationally
expensive
AM + Jaakkola, WHI'18

Beyond Post-Hoc Explanations