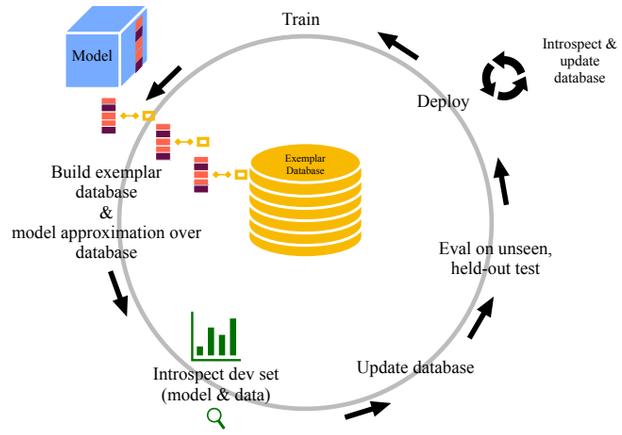
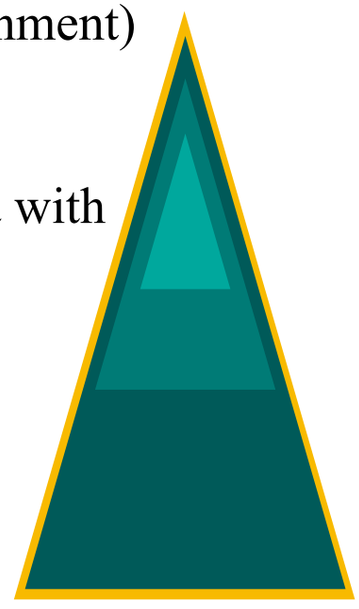
Decomposable Model Approximations for Data-Mediated AI (DMA^2)

- We seek the following (tightly coupled) characteristics ("Data-Mediated AI"), which are not typically associated with the deep neural networks out-of-the-box:
  - Class-conditional feature detection, relating global-level predictions to local-level predictions (and vice-versa)
  - An explicit connection to the training data for a given prediction, at a given unit of analysis
  - Out-of-domain detection and uncertainty estimates
  - A degree of updatability/adaptability (without re-training the full network)



- Methods to add these characteristics to deep networks (via "Decomposable Model Approximations"):
  - *Horizontal* (across the input), *vertical* (across the support set), and *diagonal* (sequence alignment) model decompositions
  - Exemplar auditing: Dense representation matching analysis and constraints against the data with known labels
  - An abstain+update/adapt paradigm made possible by the approximations (orthogonal to updatability properties the original model itself may have, such as via retrieval)



Neural network interpretability (and deployment) is an interactive, human-in-the-loop prediction task.