We are interested in finding the memory and compute demands of a given model.

With this we should be able to estimate if the model will fit on a given gpu/machine and how many queries per second can be supported.

1. Static method: Load only the model without the weights. This should fit in most local machines. And analyze layer by layer the parameters and the types. From here, estimate the memory and compute needed.
   1. We will start with various collection of model types, starting from image, classification and going to LLMs
   2. We need to identify and group them by type of parameters, broadly, embedding or ffn or convolutional layers.
   3. Provide summarized tabular data
2. Dynamic method: TBD