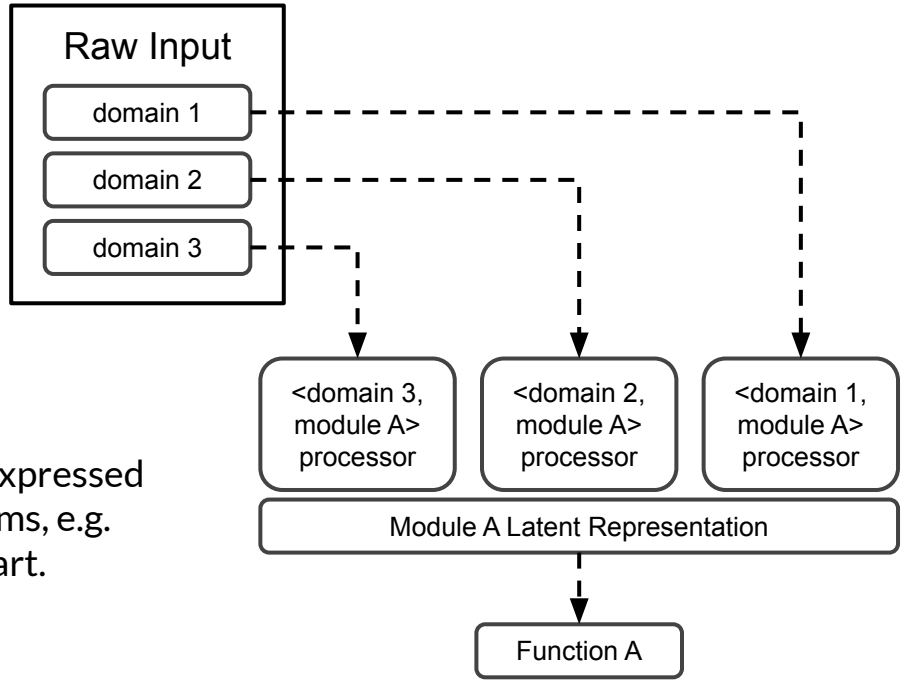
Meaning-Preserving Continual Learning v1

Meaning

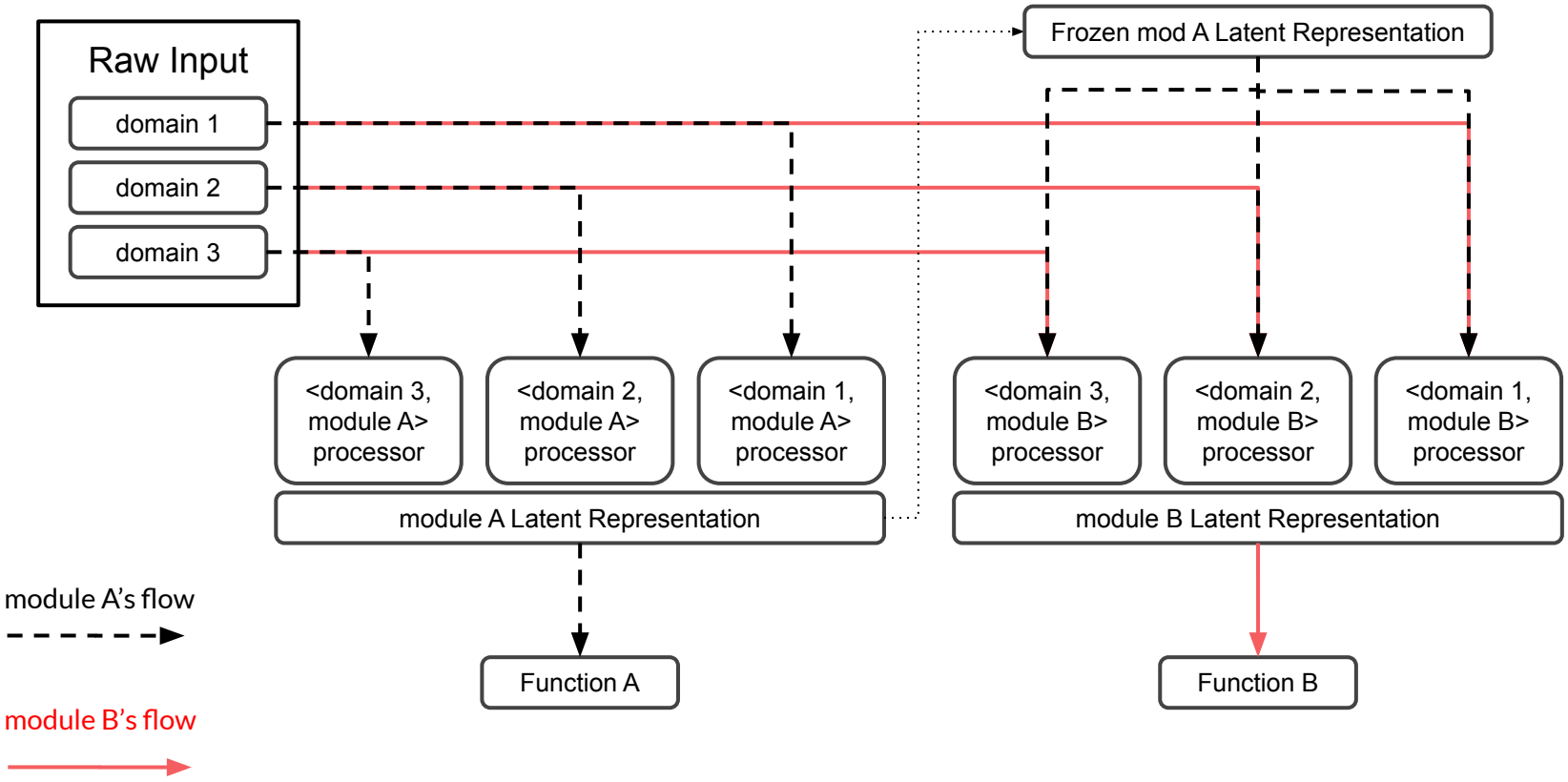


Module A's function is expressed in subjective human terms, e.g. telling cats and dogs apart.

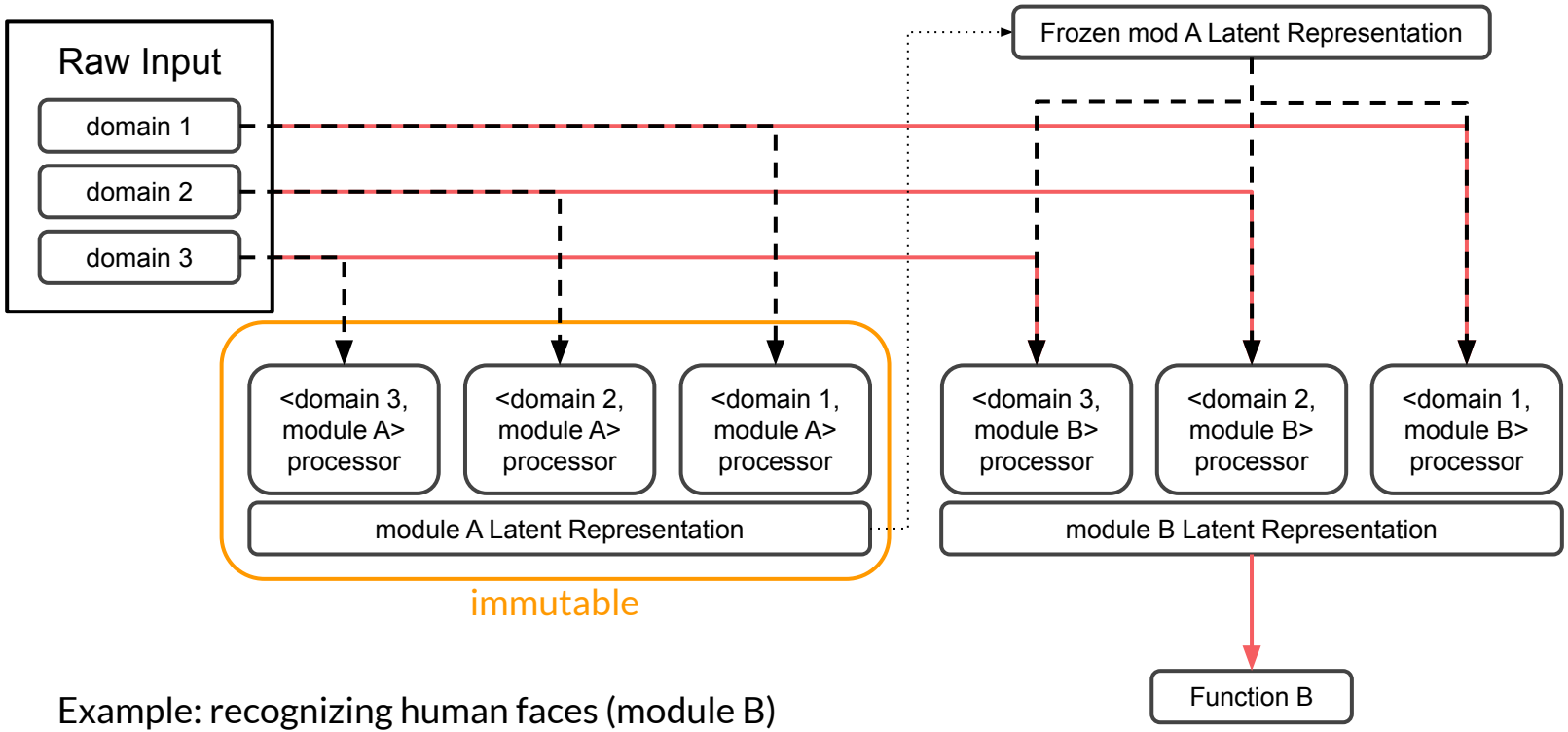
For *HumanMeaning(ModuleA) = Meaning(Module A's Latent Representation)* to hold true, module A's output classes must be accurately predicted* across many domains/contexts. If there aren't enough domains, there is no guarantee that *Meaning(Module A's Latent Representation)* aligns with *HumanMeaning(moduleA)*.

*prediction of classes or numerical values, or rewards from motor goals.

Two-module scenario

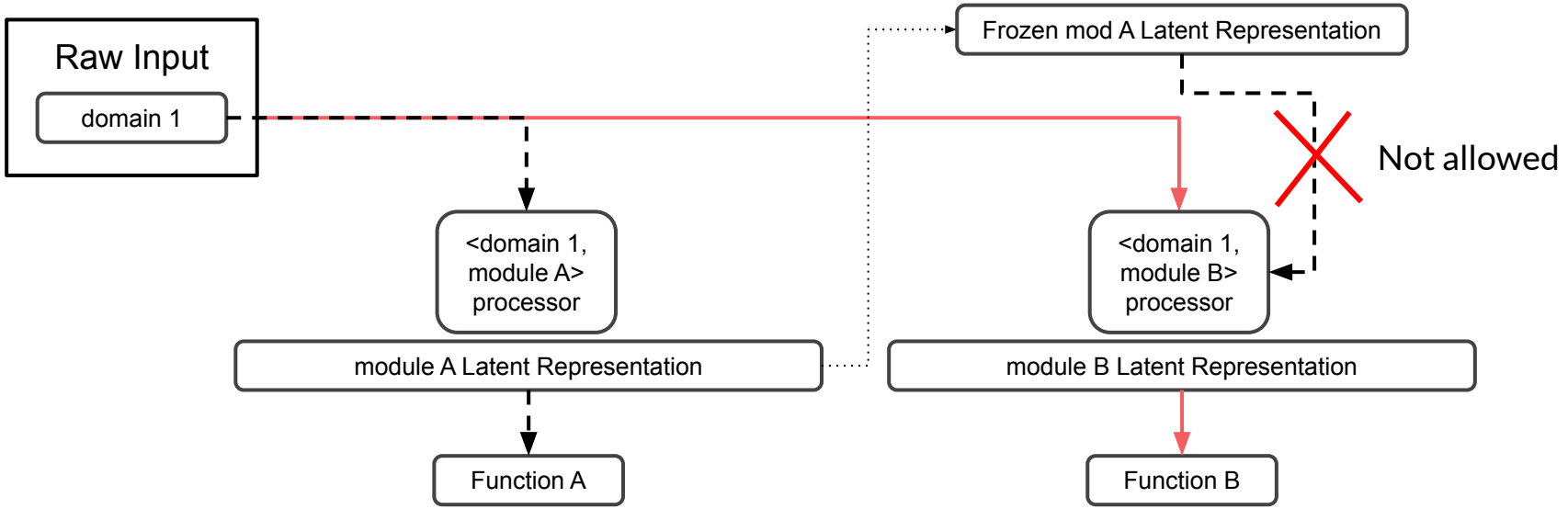


Rule 1: module A is frozen when training module B



Example: recognizing human faces (module B) cannot interfere with the function of telling cats and dogs apart (function A).

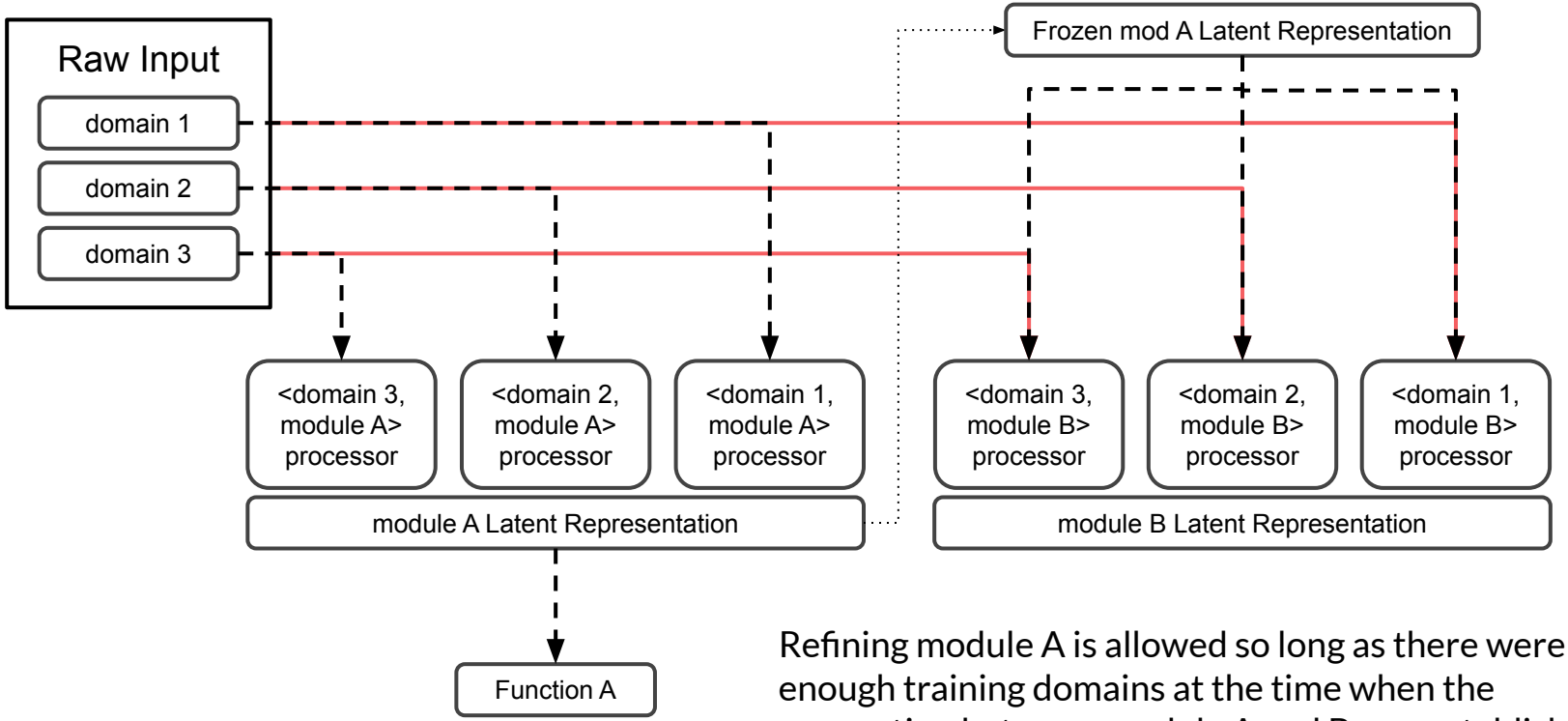
Rule 2: module B cannot utilize module A if training domains were scarce



If there is a misalignment due to the lack of training domains, i.e. *HumanMeaning(moduleA) != Meaning(module A's Latent Representation)*, then the system might find a correlation between module A and module B that doesn't exist in human reality.

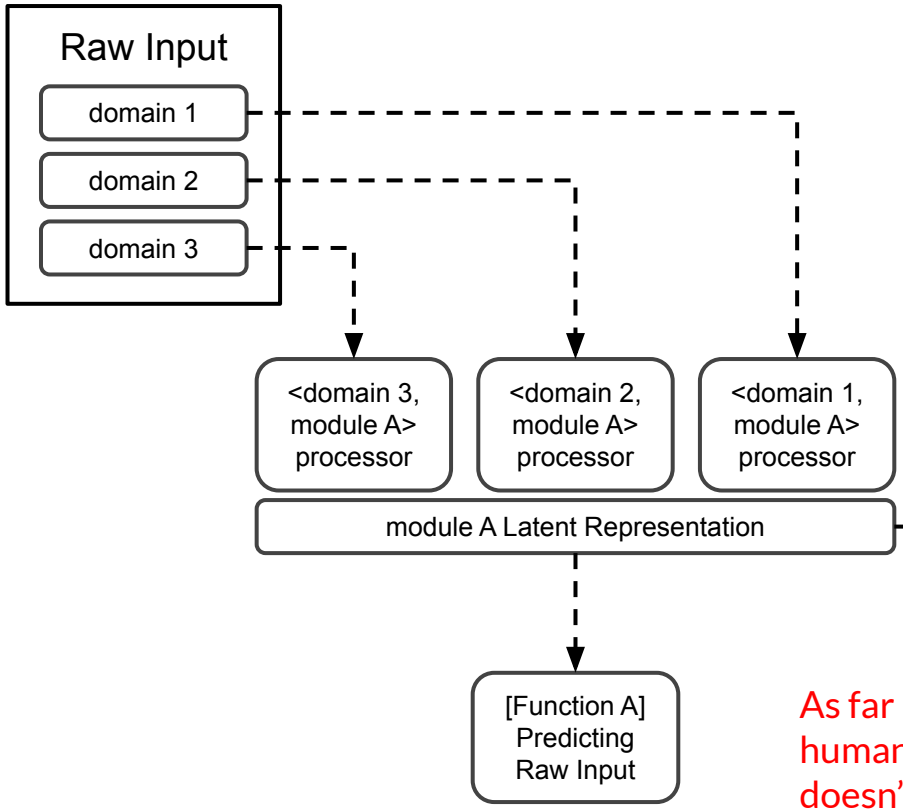
Example: without this rule, the system might mistakenly connect cats (module A) to arctic foxes (module B) if white cats were the only kind of cats seen by the system. If a connection between module A and B were to be drawn, nothing would stop module A from interfering with module B in a destructive way.

Rule 3: module A is allowed to interfere with module B if it doesn't break Rule 2



Refining module A is allowed so long as there were enough training domains at the time when the connection between module A and B was established.

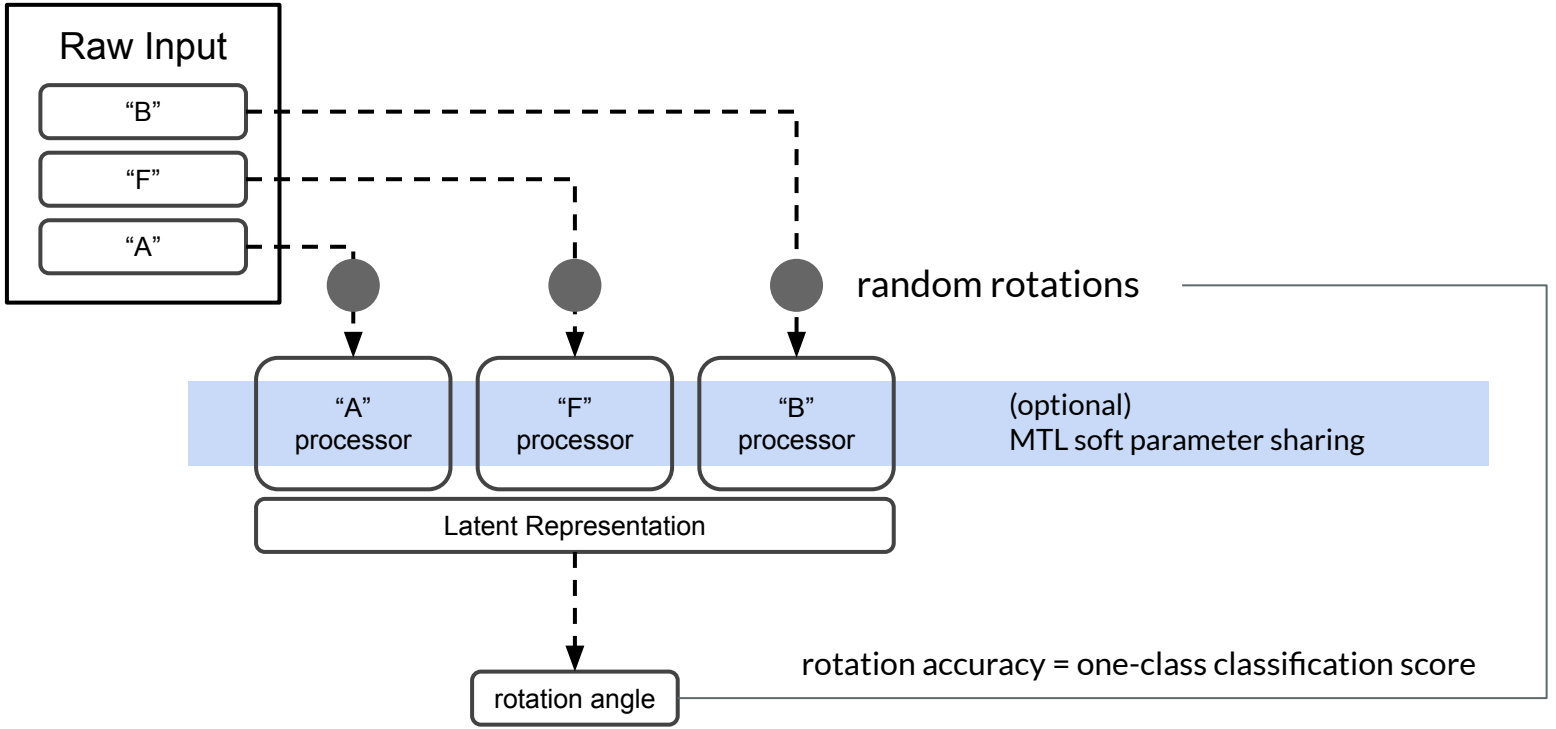
Rule 4: modules must interpret the raw input in subjective terms



Autoencoding is an example of useless function.

As far as making predictions in the human world goes, this representation doesn't carry more information than the raw input itself.

One-module scenario: EMNIST



This works because “rotation” is a subjective concept whose meaning is not carried by the input alone. This is why the system needs a custom processor for each letter, and this is why it is an adequate function for detecting discrepancies.

Example of degenerate module: if the function is to denoise the images or to count the number of black pixels, it can be done without knowing the letter, thus it doesn’t help us predict the letter class.