**Goal 1:** figure out a large fraction of all the features represented as directions in the model hidden states.

**Linear Probes** To be confident that you achieved this goal *just* with linear probes, you would have to do the equivalent of

1. Listing a large fraction of *all the features that could possibly be represented*
2. For each feature on the list, train the probe (and maybe control-probes)

To the extent that the list is long, this approach is computationally expensive. In the case of gLMs or LLMs that list is plausibly so long, that this approach is intractable.

**SAEs**

* Under which conditions should you expect an SAE to achieve this goal?
  + Under which conditions should you expect the set of decoder rows to sufficiently approximate the set of feature-directions?

**Goal 2:** for a particular, well-defined, feature F, verify that F is represented as a direction in hidden model states

Hot take: in this case, almost always, linear probes >>> SAEs