The goal of this section will be to motivate these ideas and unpack them in detail.

It's worth noting that many of the ideas in this section have close connections to ideas in other lines of interpretability research (especially disentanglement), neuroscience (distributed representations, population codes, etc), compressed sensing, and many other lines of work. This section will focus on articulating our perspective on the problem. We'll discuss these other lines of work in detail in Related Work.

Empirical Phenomena

When we talk about "features" and how they're represented, this is ultimately theory building around several observed empirical phenomena. Before describing how we conceptualize those results, we'll simply describe some of the major results motivating our thinking:

- **Word Embeddings** - A famous result by *Mikolov et al.* [8] found that word embeddings appear to have directions which correspond to semantic properties, allowing for embedding arithmetic vectors such as `V("king") - V("man") + V("woman") = V("queen")` (*but see* [9]).
- **Latent Spaces** - Similar "vector arithmetic" and interpretable direction results have also been found for generative adversarial networks (e.g. [10]).
- **Interpretable Neurons** - There is a significant body of results finding neurons which appear to be interpretable (*in RNNs* [11, 12]; *in CNNs* [13, 14]; *in GANs* [15]), activating in response to some understandable property. This work has faced some skepticism [16, 17]. In response, several papers have aimed to give extremely detailed accounts of a few specific neurons, in the hope of dispositively establishing examples of neurons which truly detect some understandable property (notably Cammarata *et al.* [6], but also [18, 19]).
- **Universality** - Many analogous neurons responding to the same properties can be found across networks [20, 1, 18].
- **Polysemantic Neurons** - At the same time, there are also many neurons which appear to not respond to an interpretable property of the input, and in particular, many *polysemantic neurons* which appear to respond to unrelated mixtures of inputs [21].

As a result, we tend to think of neural network representations as being composed of *features* which are *represented as directions*. We'll unpack this idea in the following sections.

What are Features?

Our use of the term "feature" is motivated by the interpretable properties of the input we observe neurons (or word embedding directions) responding to. There's a rich variety of such observed properties![2] We'd like to use the term "feature" to encompass all these properties.

But even with that motivation, it turns out to be quite challenging to create a satisfactory definition of a feature. Rather than offer a single definition we're confident about, we consider three potential working definitions: