What are generative models?

Generative modeling loosely refers to building a model of data, for instance p(image), that we can sample from. This is in contrast to discriminative modeling, such as regression or classification, which tries to estimate conditional distributions such as p(class | image).

Why generative models?

Even when we're only interested in making predictions, there are practical reasons to build generative models:

- **Data efficiency and semi-supervised learning** - Generative models can reduce the amount of data required. As a simple example, building an image classifier p(class | image) requires estimating a very high-dimenisonal function, possibly requiring a lot of data, or clever assumptions. In contrast, we could model the data as being generated from some low-dimensional or sparse latent variables z, as in $p(image) = \int p(image|z)p(z)dz$. Then, to do classification, we only need to learn p(class | z), which will usually be a much simpler function. This approach also lets us take advantage of unlabeled data - also known as semi-supervised learning.
- **Model checking by sampling** - Understanding complex regression and classification models is hard - it's often not clear what these models have learned from the data and what they missed. There is a simple way to sanity-check and inspect generative models - simply sample from them, and compare the sampled data to the real data to see if anything is missing.
- **Understanding** - Generative models usually assume that each datapoint is generated from a (usually low-dimensional) latent variable. These latent variables are often interpretable, and sometimes can tell us about the hidden causes of a phenomenon. These latent variables can also sometimes let us do interesting things such as interpolating between examples