TL;DR for Desiderata for Representation Learning by Yixin

roughly, we tried to understand what desiderata we want from representation learning in both supervised and unsupervised learning. and it turns out that the causal language can help us a lot with articulating these desiderata, which in turn enables metrics and algorithms for representation learning.

for supervised representation learning, we found that the desiderata of nonspuriousness and efficiency correspond to the concept of sufficiency and necessity of causes. so we formulate representation learning as a task of
finding necessary and sufficient causes (basically maximizing
probability of necessity and sufficiency).

it turns out that (perhaps surprisingly), with this formulation, it is possible to learn non-spurious features (of images and text) with a single dataset,
without enforcing invariance or requiring multiple datasets as most of
"causal representation learning" algorithms do. at least in the
benchmark datasets, the algorithm due to our formulation performs as
well as existing algorithms that leverage multiple environments.

for unsupervised representation learning, we specifically studied
"disentangled representation learning". we look at observable
implications of disentanglement, which turns out to be "independent
support" (i.e. the set of possible values each feature can take does
not depend on the values of other features) under a positivity
condition.

again, it turns out that, with this observation, we could evaluate and enforce disentanglement, even when the underlying ground truth features are correlated. limiting our attention to representations with compact support, we can also establish the identifiability of the disentangled representations, (again perhaps surprisingly) without leveraging auxiliary variables or weak supervision or ground truth features (as most disentanglement metrics do). this identifiability result enables disentangled representation learning, which boils down to enforcing this independent support condition.

from both examples in supervised and unsupervised representation learning, we hope that these theoretical and empirical results illustrate how causality can provide a fruitful and practically useful perspective for representation learning.