Dimension reduction and transfer learning.

Mariela Sued

Universidad de San Andrés

Imagine training a model on photos of cows and camels so that it can classify them in new pictures. The model learns well — until you show it a cow in the desert. It says "camel." The model never really learned to recognize the animal; it learned to recognize the background. Cows are always on grass, camels always on sand, and the model got confused by that spurious association.

This particular problem is by now well understood, and solutions exist. But the same challenge arises across a vast range of disciplines. In medical diagnosis, a model trained at a large urban hospital may fail at a rural clinic — not because the disease changed, but because the patient population or the imaging conditions did. In many other disciplines the same issue appears: the model learned to rely on features that shift across populations rather than on the stable signal that actually matters.

These settings call for a principled methodology: one that extracts the information in the data that is genuinely predictive, while remaining blind to confounders that vary across environments. That is the problem this talk addresses.