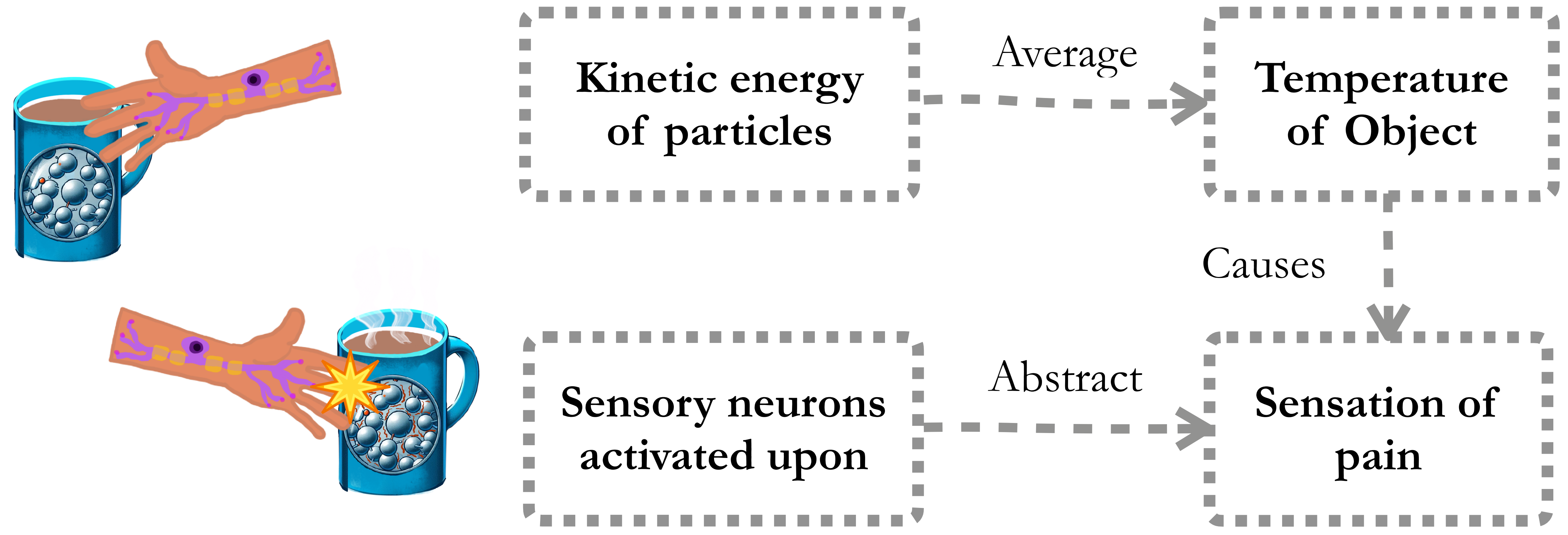
Learning macro variables using Auto encoders

Maitreyi Swaroop[1,3], Eric Elmoznino[1,2], Dhanya Sridhar[1,2]
[1]Mila, [2]Université de Montreal, [3]IIT Kharagpur

Motivation

Most causal variables that we reason over, in both science and everyday life, are coarse abstractions of low-level data.



Kinetic energy of particles → Average → **Temperature of Object**

Temperature of Object → Causes → **Sensation of pain**

Sensory neurons activated upon → Abstract → **Sensation of pain**

Related work: Causal Feature Learning

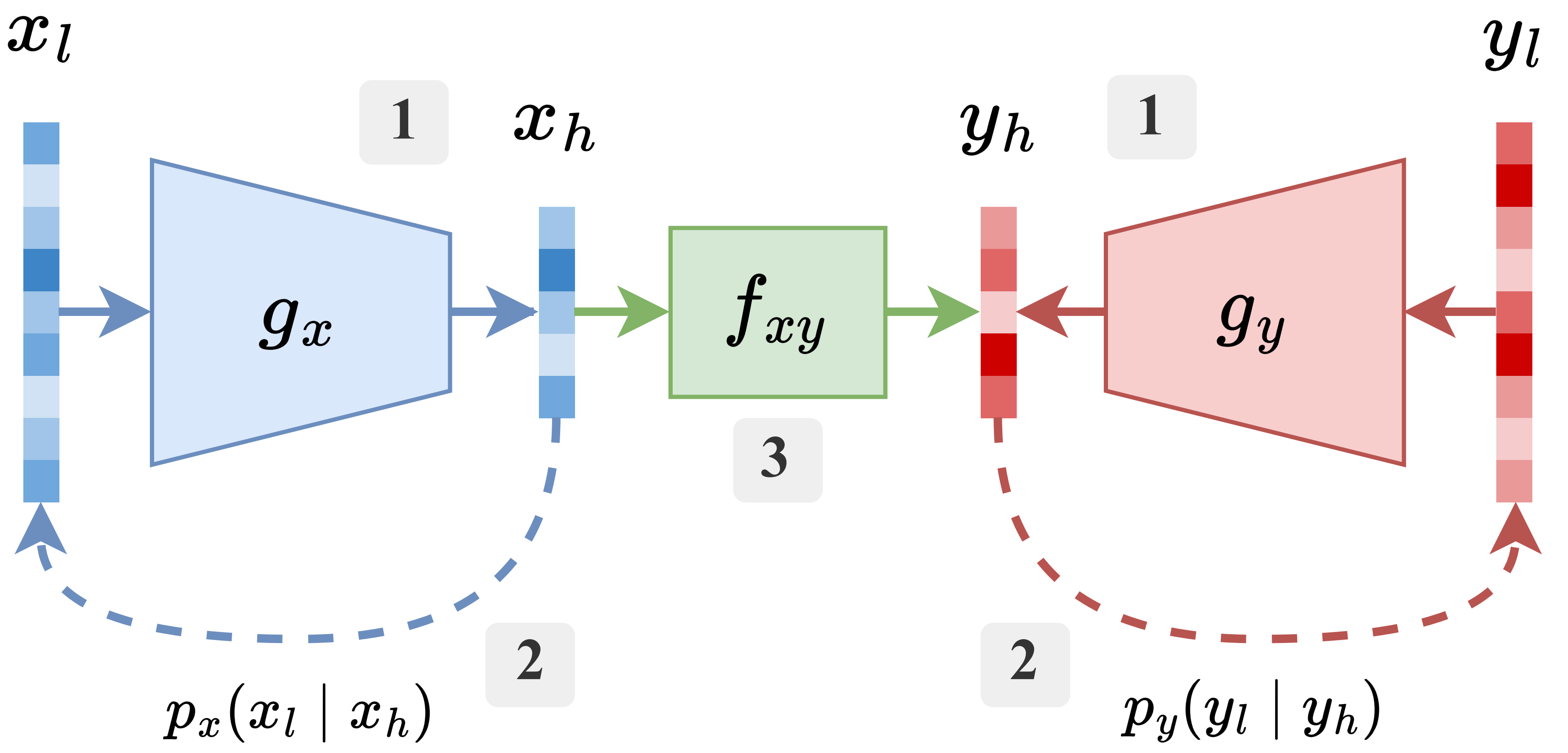
- Aggregates micro variables by defining equivalence classes (macro variables) to which they are mapped.
- Macro variables are discrete and not interpretable.

Method: DeepCFL

$$\mathscr{L} = -ELBO(g_x, p_x) - ELBO(g_y, p_y) + \lambda \frac{||f(x_h) - y_h||^2}{var(y_h)}$$

Macro variable desiderata

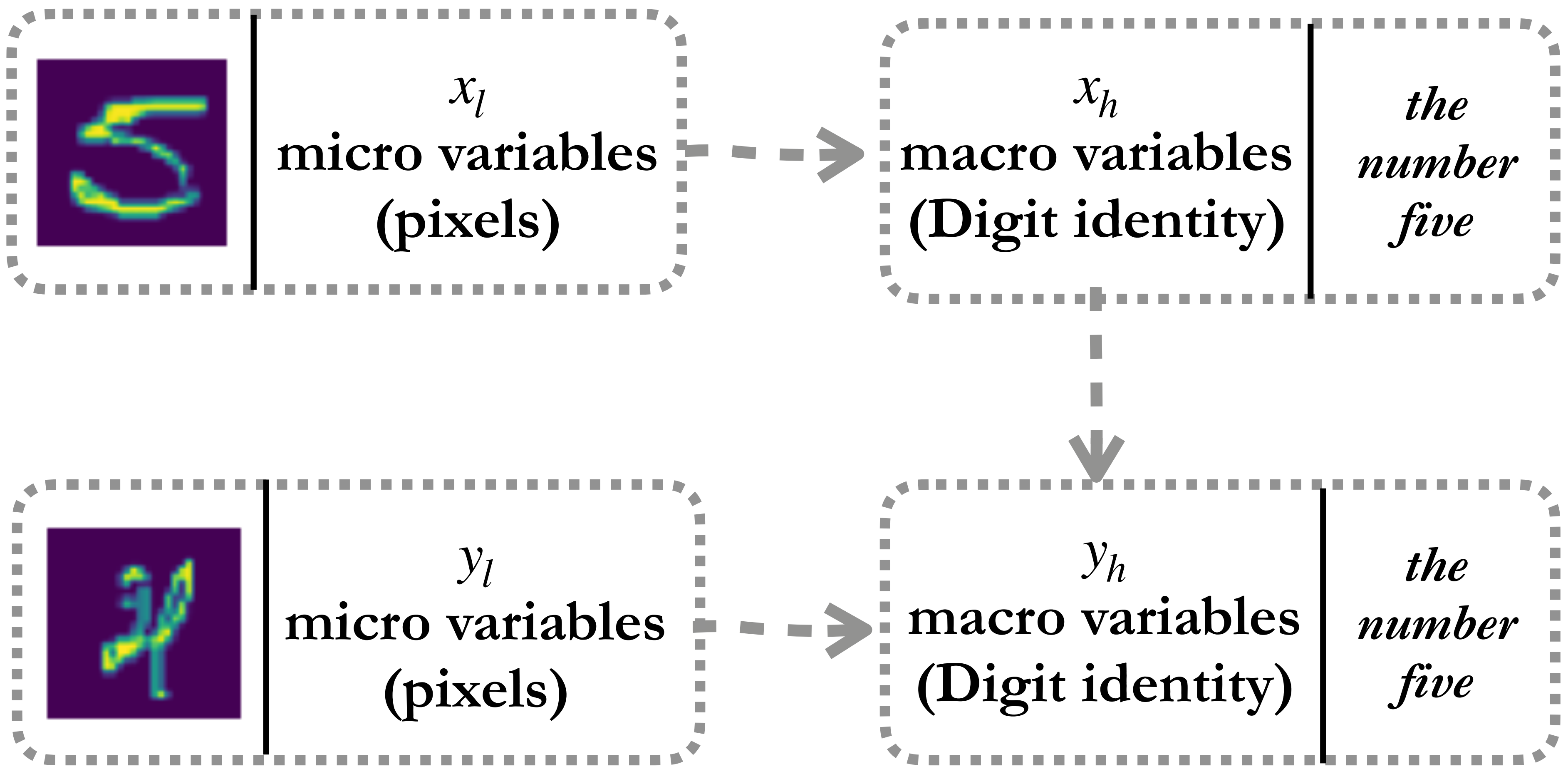
1. Macro variables are simpler than their micro variables
2. Macro variables share MI with their micro variables
3. A simple mechanism relates macro variables



f_{xy} **should be *simple***

Symbolic function
Linear transform
Shallow neural net
Sparsity regularizer
…

Empirical Studies

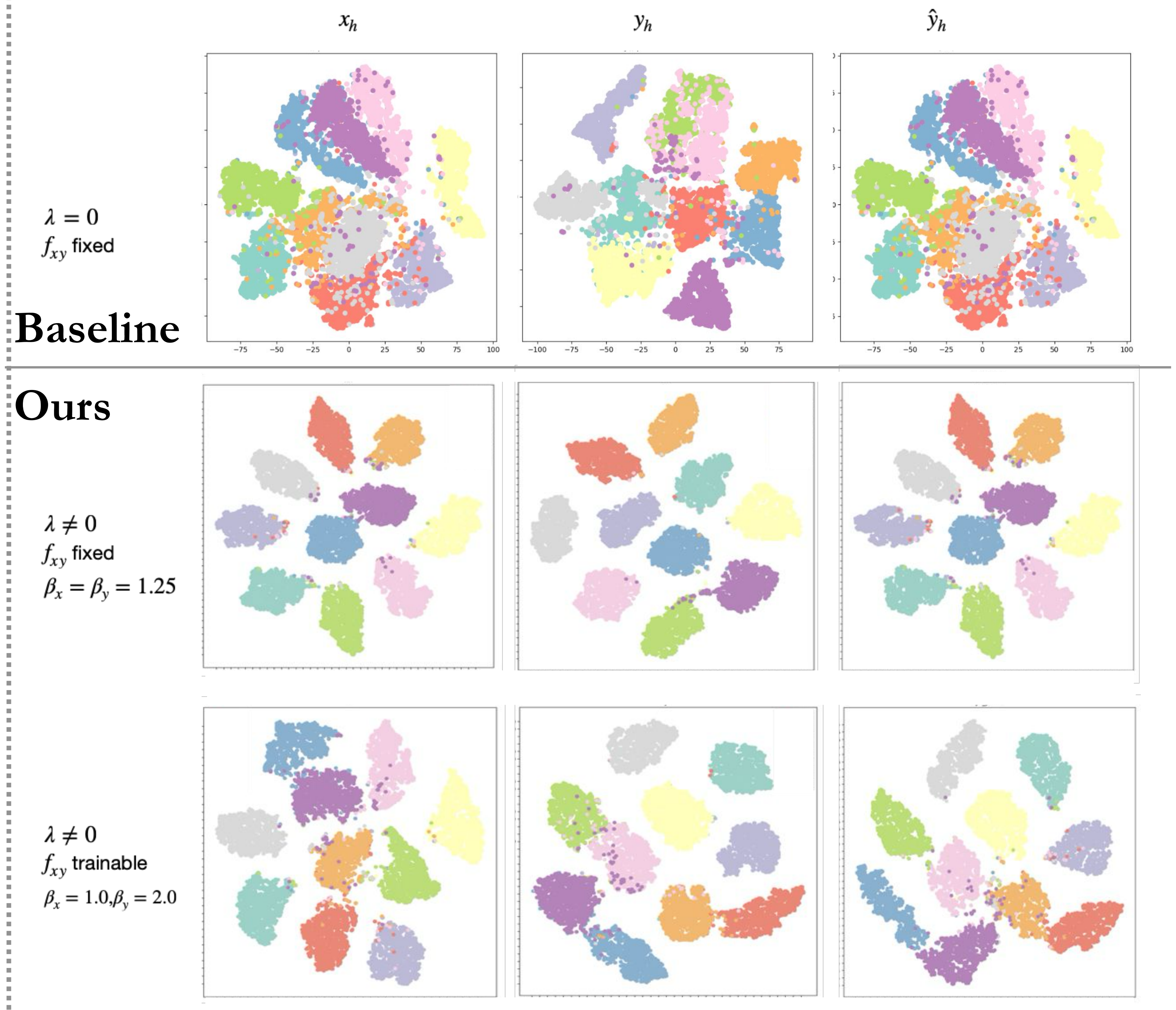


x_l micro variables (pixels) → x_h macro variables (Digit identity) — *the number five*

y_l micro variables (pixels) → y_h macro variables (Digit identity) — *the number five*

Observations

Since the correct macro variables are the digit identities, a metric of DeepCFL's performance is *how well the different classes of digits are clustered*


x_h
y_h
$\hat{y}_h$
$\lambda = 0$
f_{xy} fixed
Baseline
Ours
$\lambda \neq 0$
f_{xy} fixed
$\beta_x = \beta_y = 1.25$
$\lambda \neq 0$
f_{xy} trainable
$\beta_x = 1.0, \beta_y = 2.0$