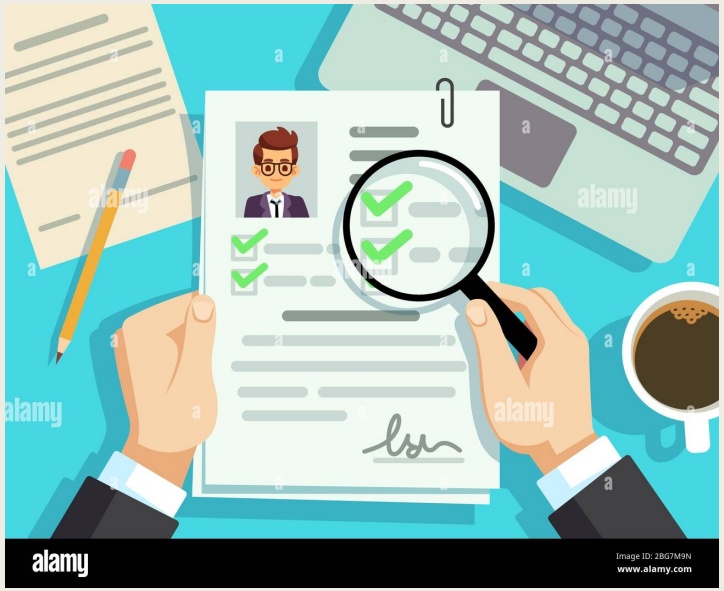
Fair Representation Learning through Implicit Path Alignment

Changjian Shui, Qi Chen, Jiaqi Li, Boyu Wang, Christian Gagné

ICML | 2022

https://cjshui.github.io/pages/inv_fair.html

Machine learning in sociotechnical system

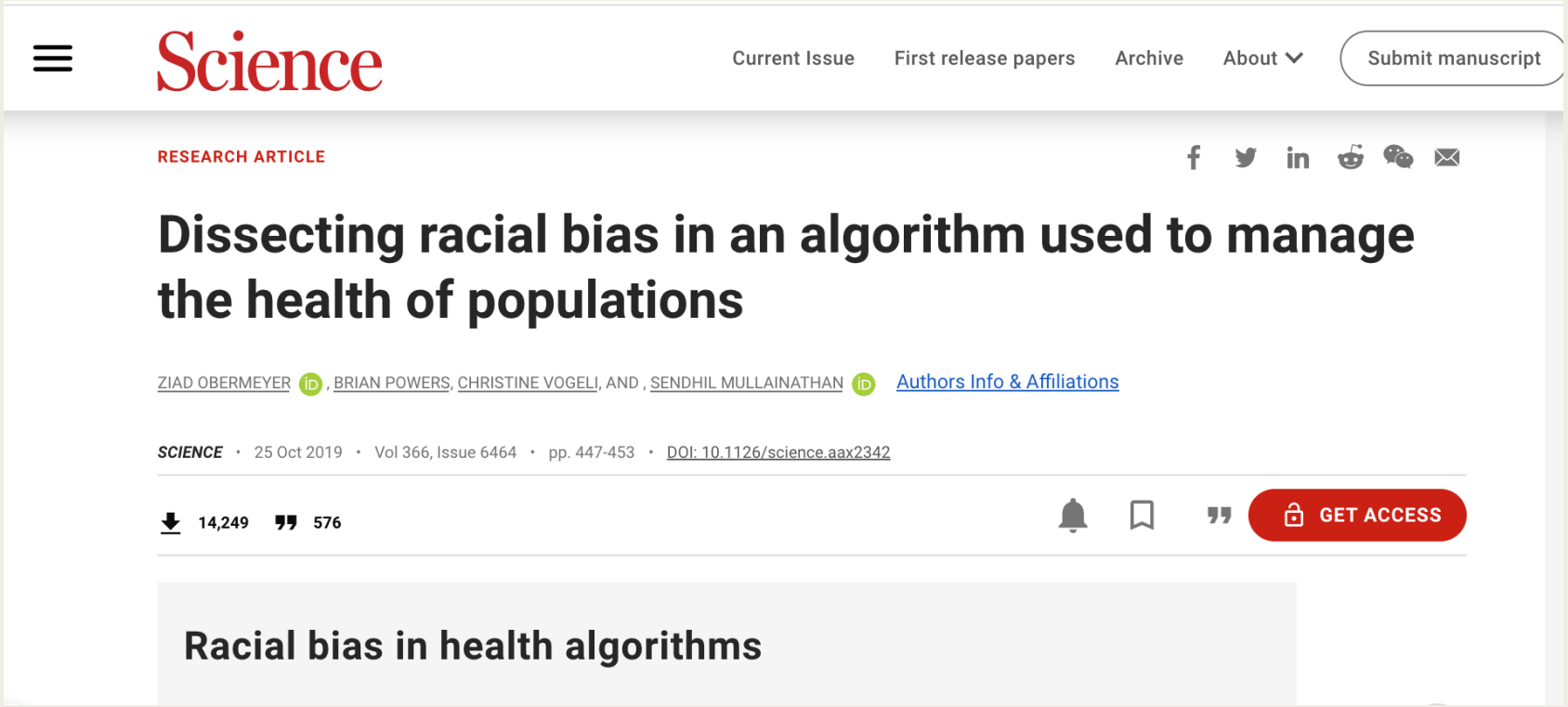


Candidate evaluations for job positions

Health risk assessment

Source: https://www.vecteezy.com/vector-art/2207767-health-risk-assessment-blue-gradient-concept-icon

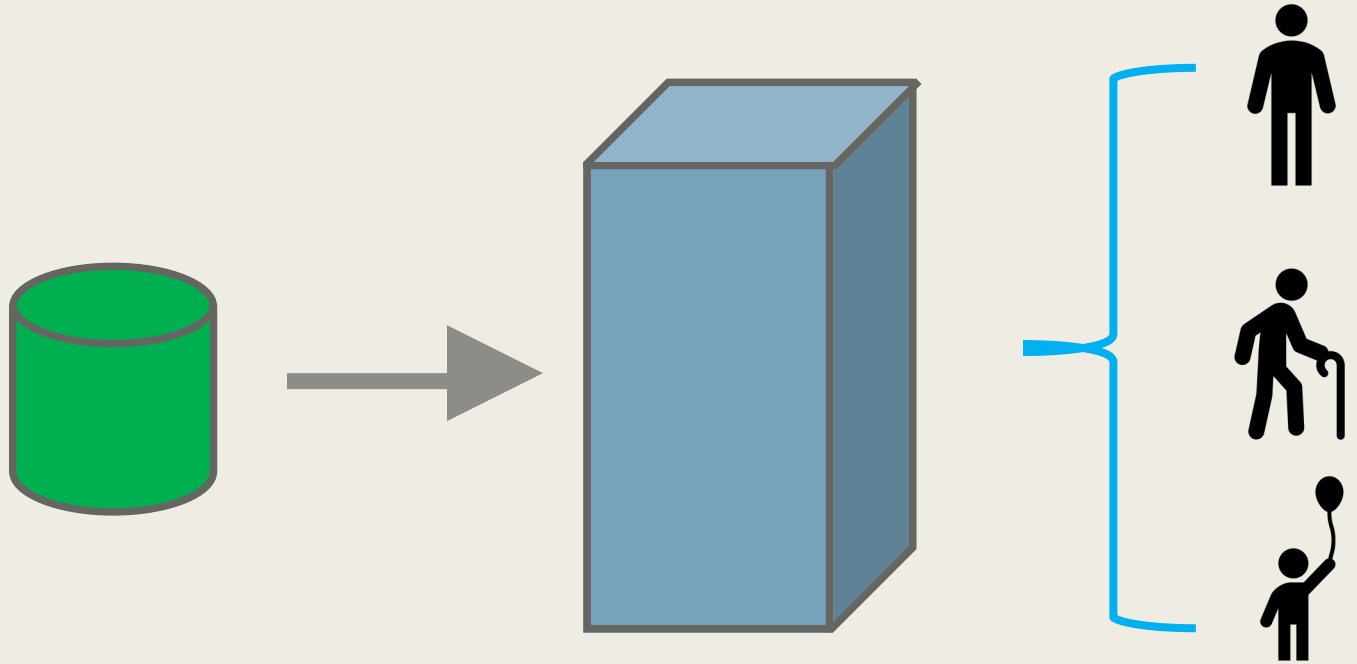
Algorithmic bias



Intelligent Health

Obermeyer et al., 366 *Science* 447 (2019)

Algorithmic fairness



No discrimination on different demographic

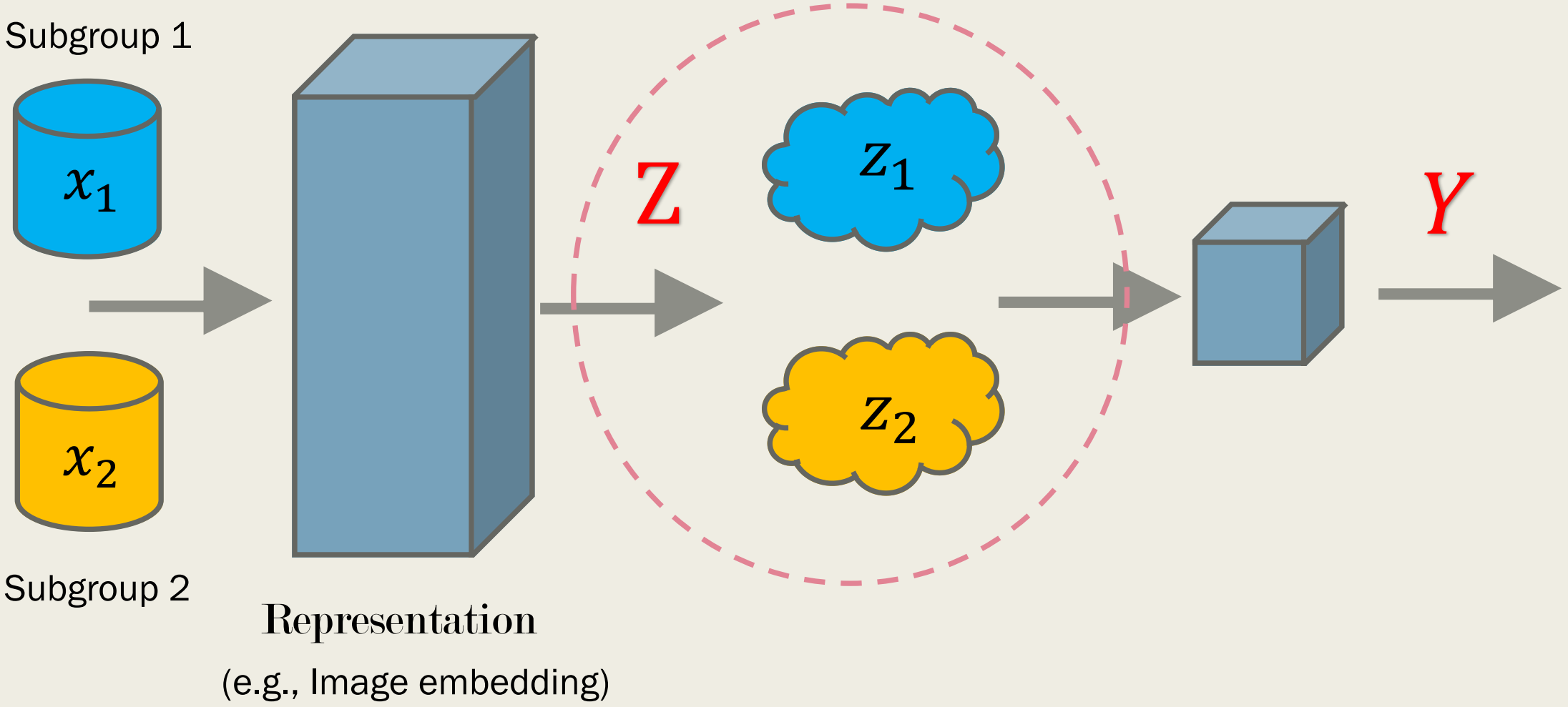
Date Preprocessing

Fair constraints during the training

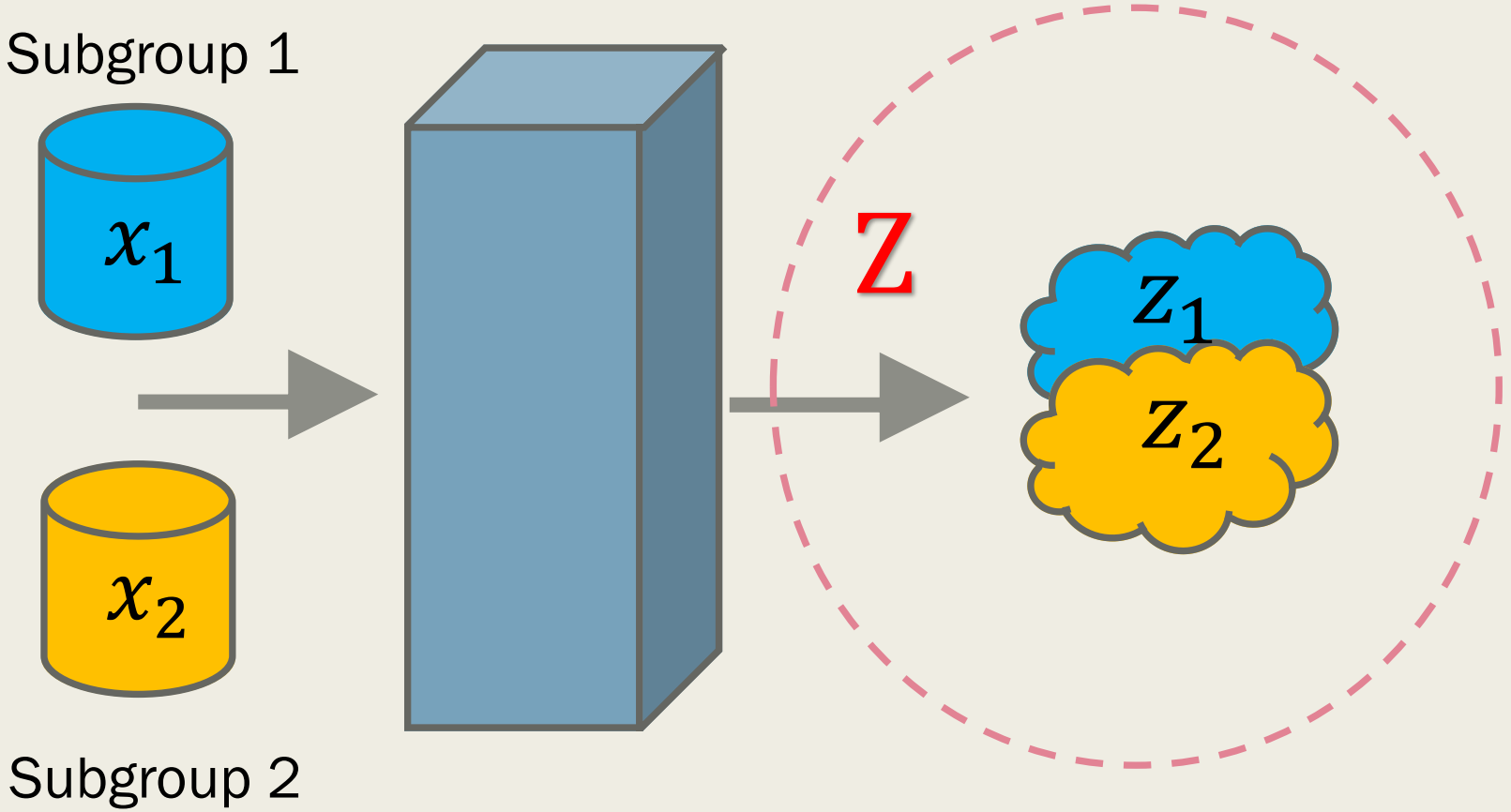
(Ours)

Modify results after training

Learning fair representation (high-level)



Invariance indicates Fairness



1. Invariant predictions on Z_1, Z_2 -> no discriminations
2. Different invariance criteria -> different fair notions

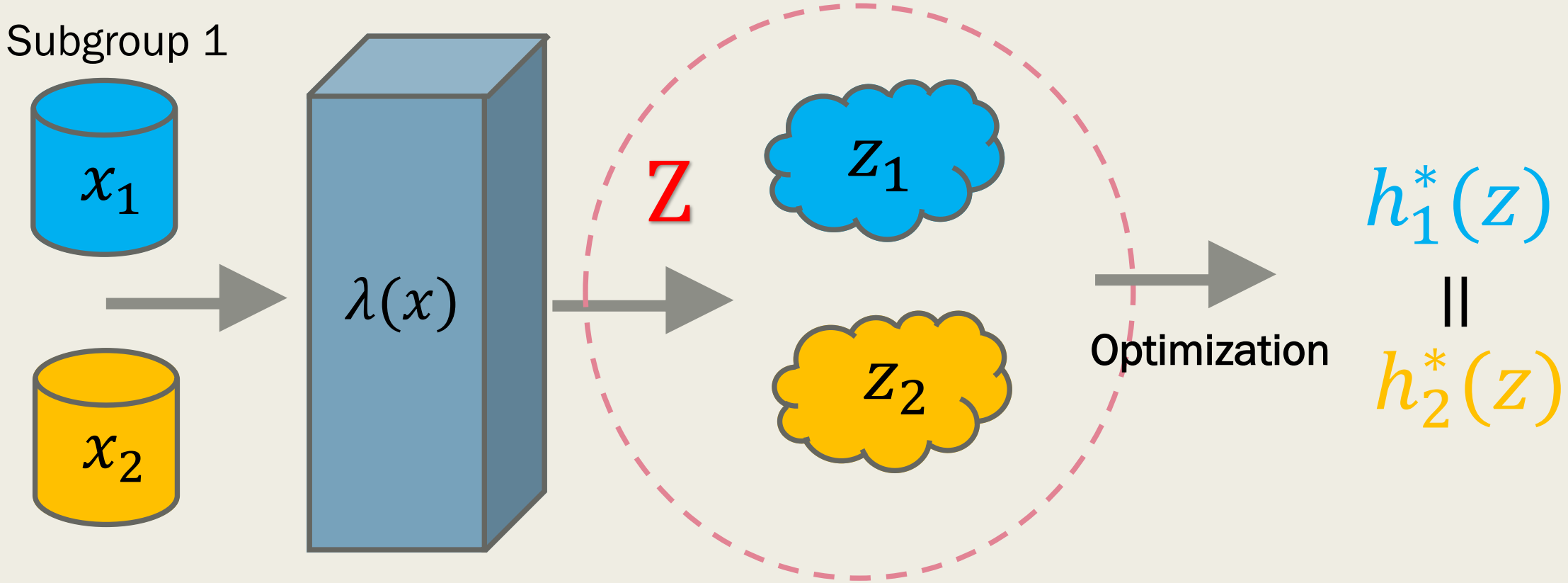
Sufficiency rule

- Sufficiency rule: given the same predicted output $\hat{Y} = y$, the identical true output.

$$E_1\left[Y \middle| \hat{Y} = y\right] = E_2\left[Y \middle| \hat{Y} = y\right]$$

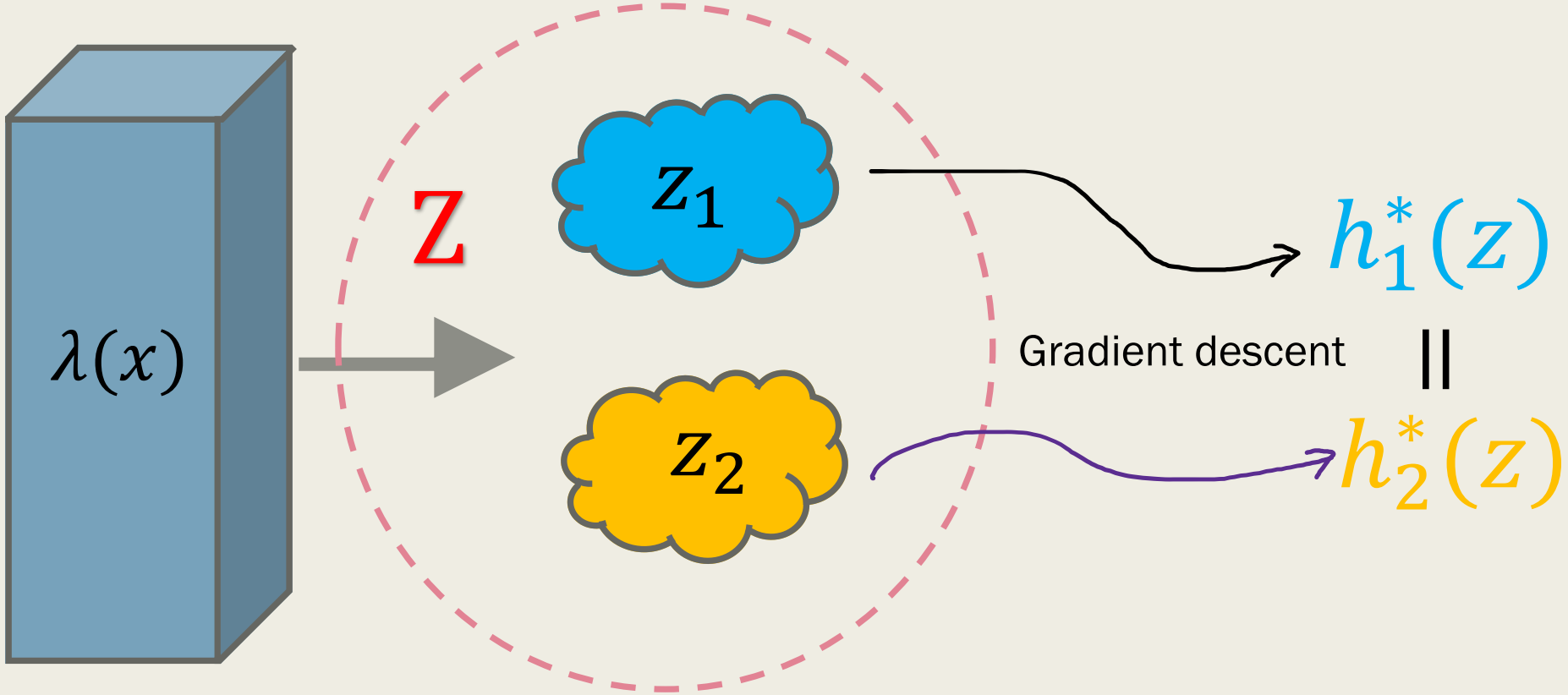
- Not compatible with other popular fair notions. (e.g., *demographic parity, equalized odds*)

Invariance for sufficiency



Adjust representation $\lambda(x)$ to ensure identical **optimal predictors** of subgroups.

Optimization viewpoint

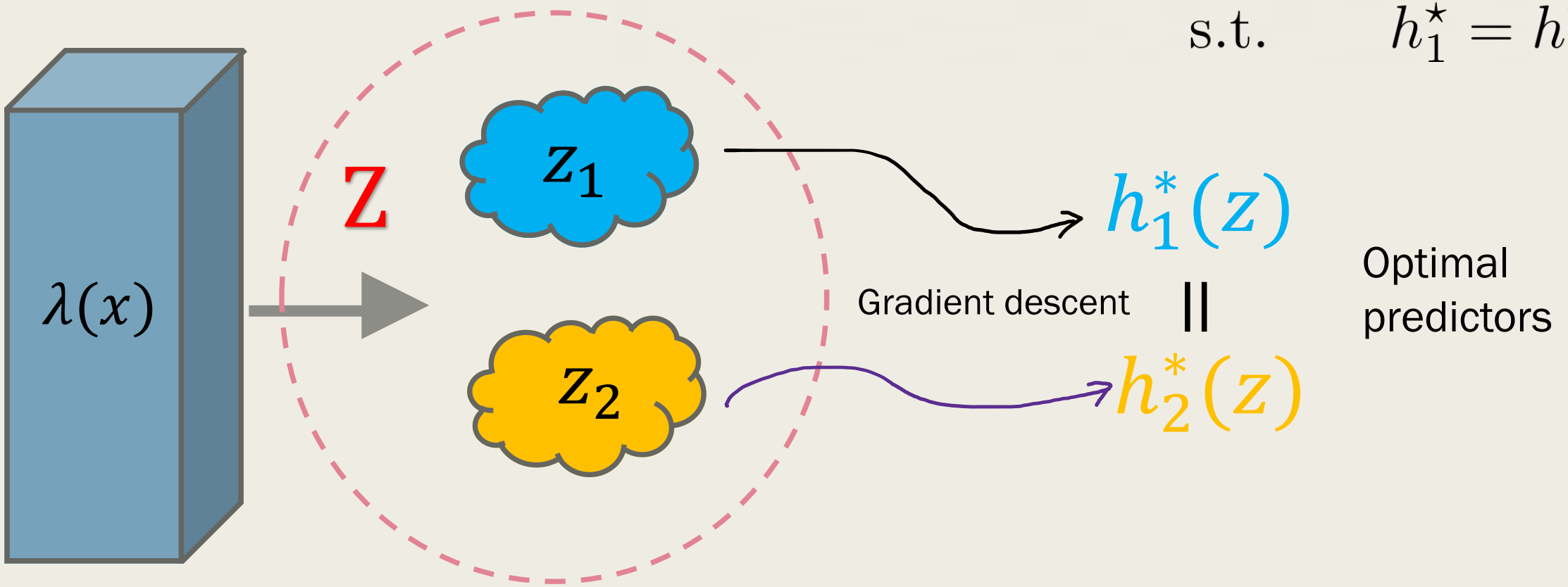


- Adjusting (optimizing) representation to ensure optimal invariant predictor on Z.
- Representation viewed as prior information (or hyper-parameter)

Formulating as bi-level optimization

$$\min_{\lambda} \mathcal{L}_1(h_1^\star \circ \lambda(x), y) + \mathcal{L}_2(h_2^\star \circ \lambda(x), y)$$

$$\text{s.t.} \qquad h_1^\star = h_2^\star,$$



Efficient optimization through **implicit theorem.**

Check details in the paper

- Code https://github.com/cjshui/fair-path