

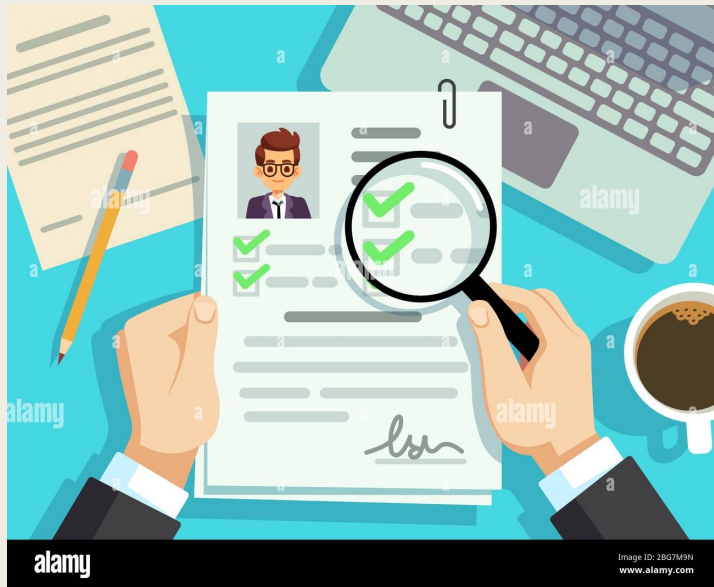
Fair Representation Learning through Implicit Path Alignment

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



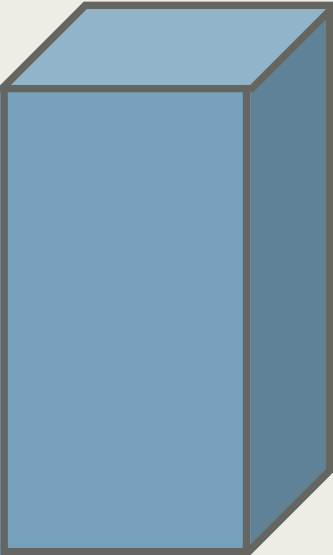
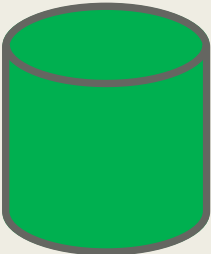
Health risk assessment

Algorithmic bias

The screenshot shows the top portion of a Science journal article page. At the top left is the Science logo and a hamburger menu icon. To the right are navigation links: 'Current Issue', 'First release papers', 'Archive', and 'About' with a dropdown arrow. A 'Submit manuscript' button is in the top right corner. Below the navigation is a 'RESEARCH ARTICLE' label and social media sharing icons for Facebook, Twitter, LinkedIn, Reddit, WeChat, and Email. The main title is 'Dissecting racial bias in an algorithm used to manage the health of populations'. Below the title are the authors: 'ZIAD OBERMEYER', 'BRIAN POWERS', 'CHRISTINE VOGELI, AND', 'SENDHIL MULLAINATHAN', and a link for 'Authors Info & Affiliations'. Below the authors is the journal information: 'SCIENCE • 25 Oct 2019 • Vol 366, Issue 6464 • pp. 447-453 • DOI: 10.1126/science.aax2342'. At the bottom of this section are download and citation counts: '14,249' downloads and '576' citations. To the right are icons for a notification bell, a bookmark, and a quote, followed by a red 'GET ACCESS' button. Below this is a grey box with the text 'Racial bias in health algorithms'.

Intelligent Health

Algorithmic fairness



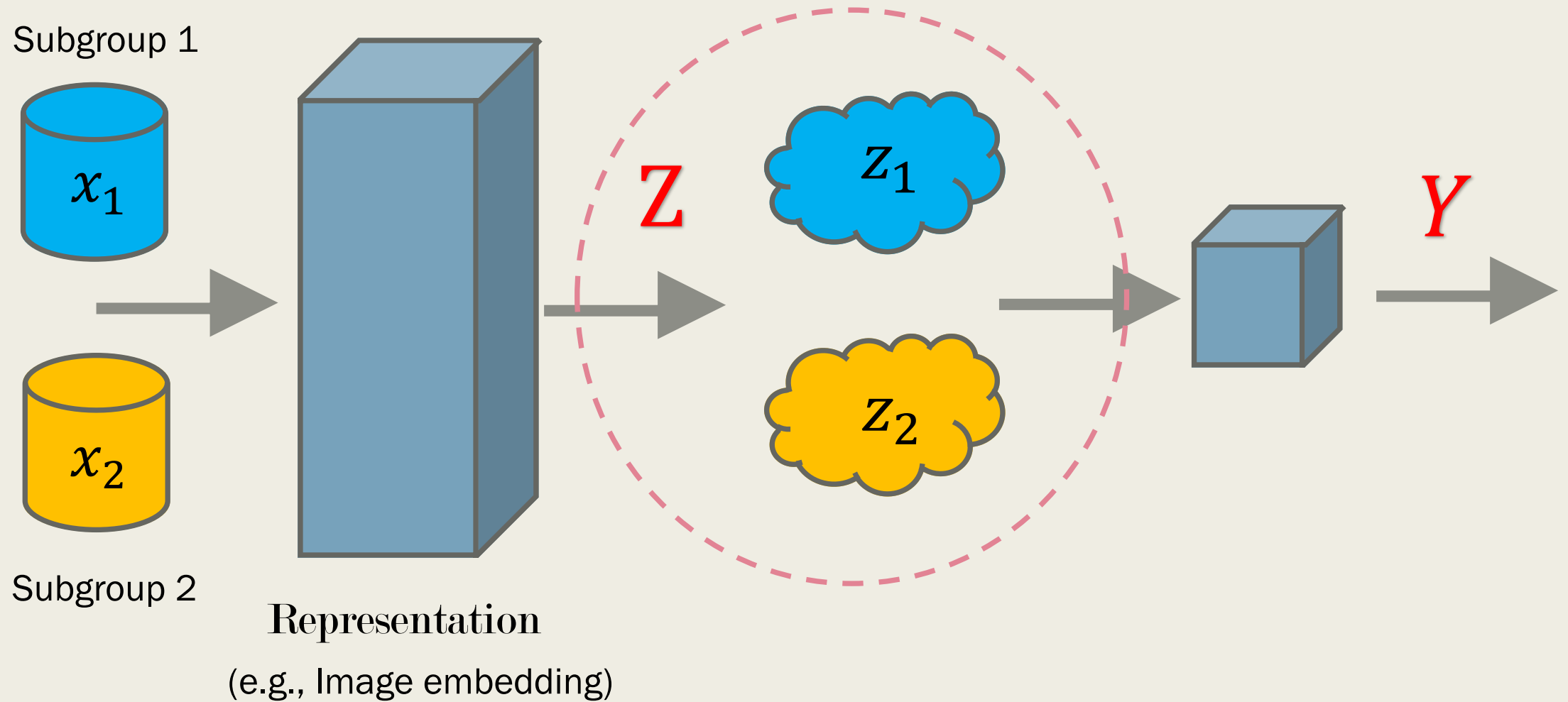
No discrimination on different demographic

Date
Preprocessing

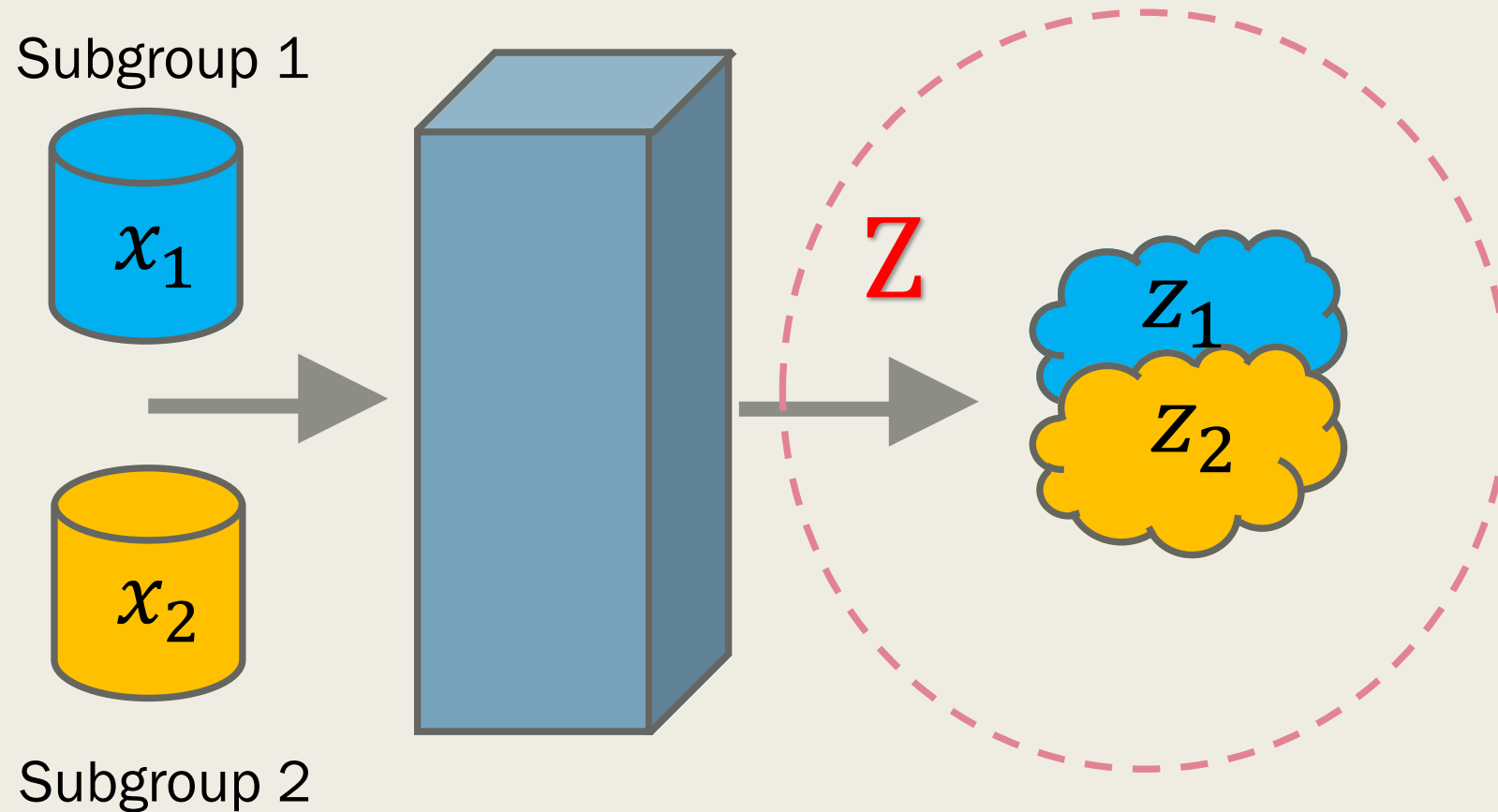
Fair constraints
during the training
(Ours)

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

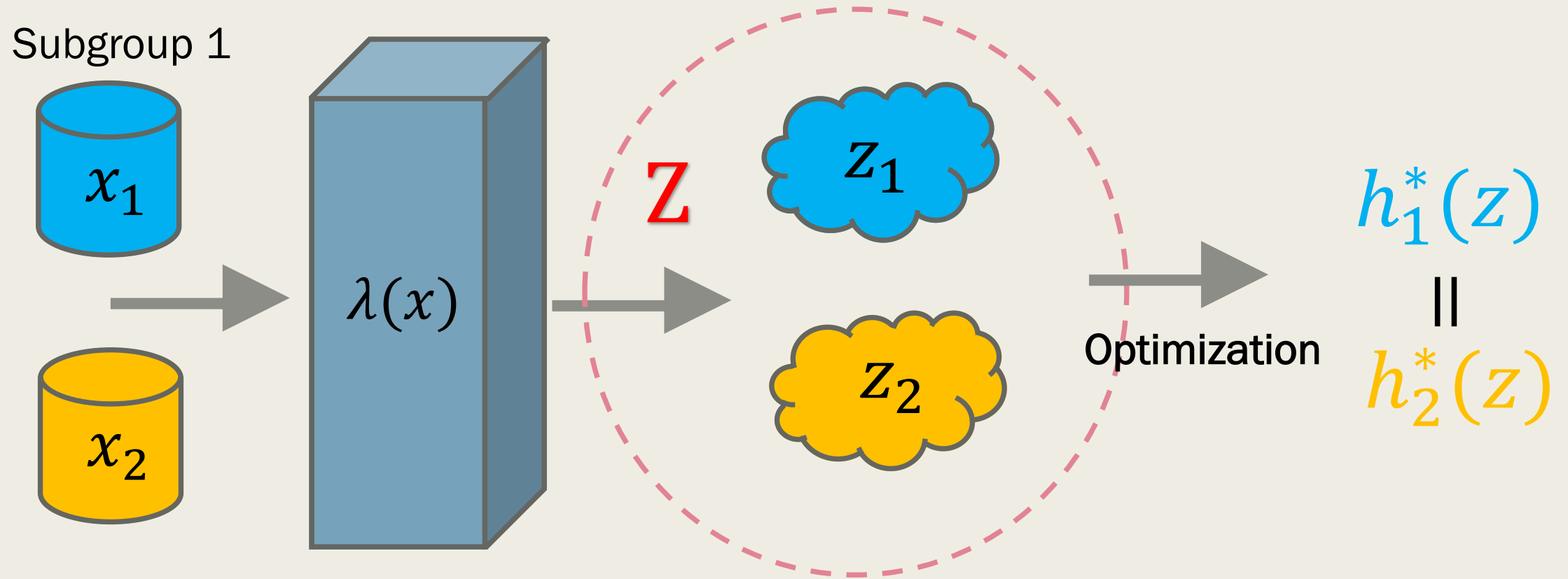
Sufficiency rule

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

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

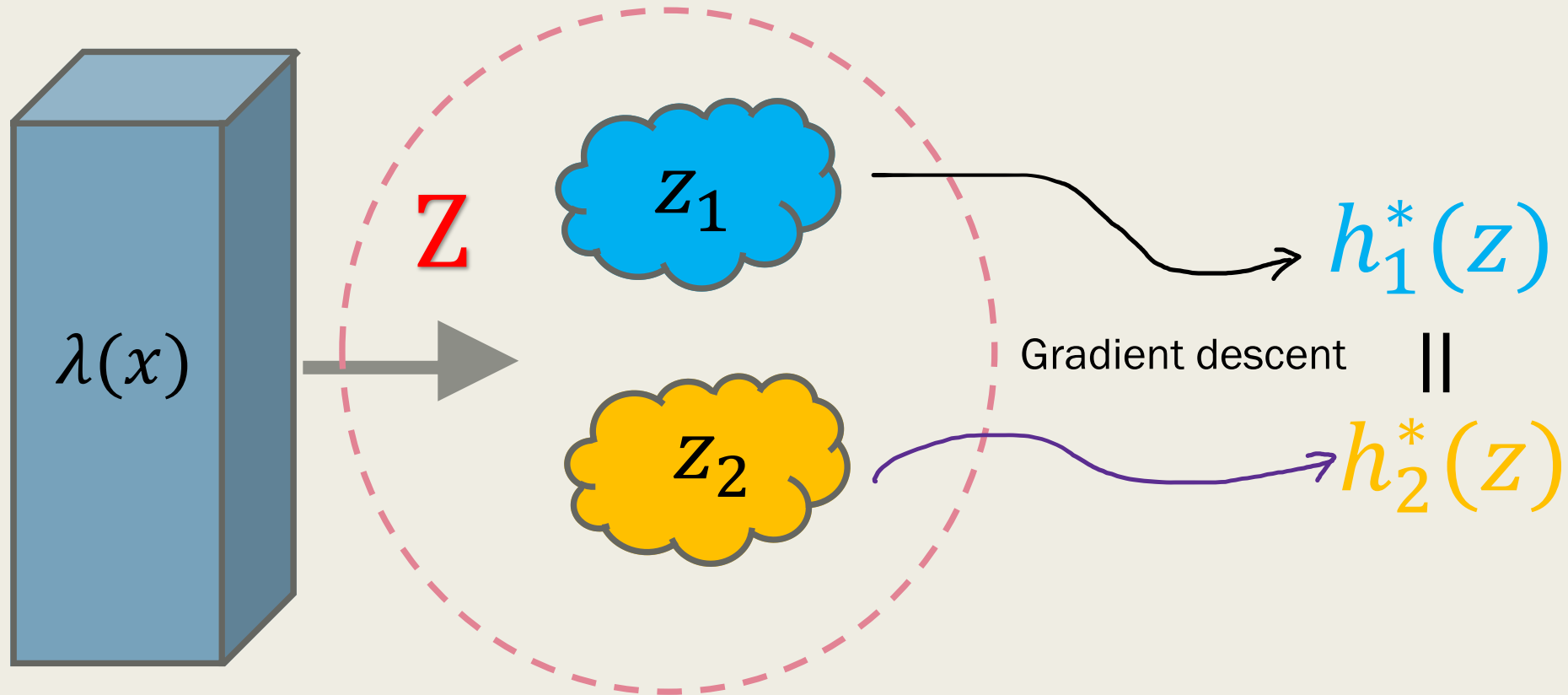
- **Not compatible** with other popular fair notions. (e.g., *demographic parity, equalized odds*)

Invariance for sufficiency



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

Optimization viewpoint

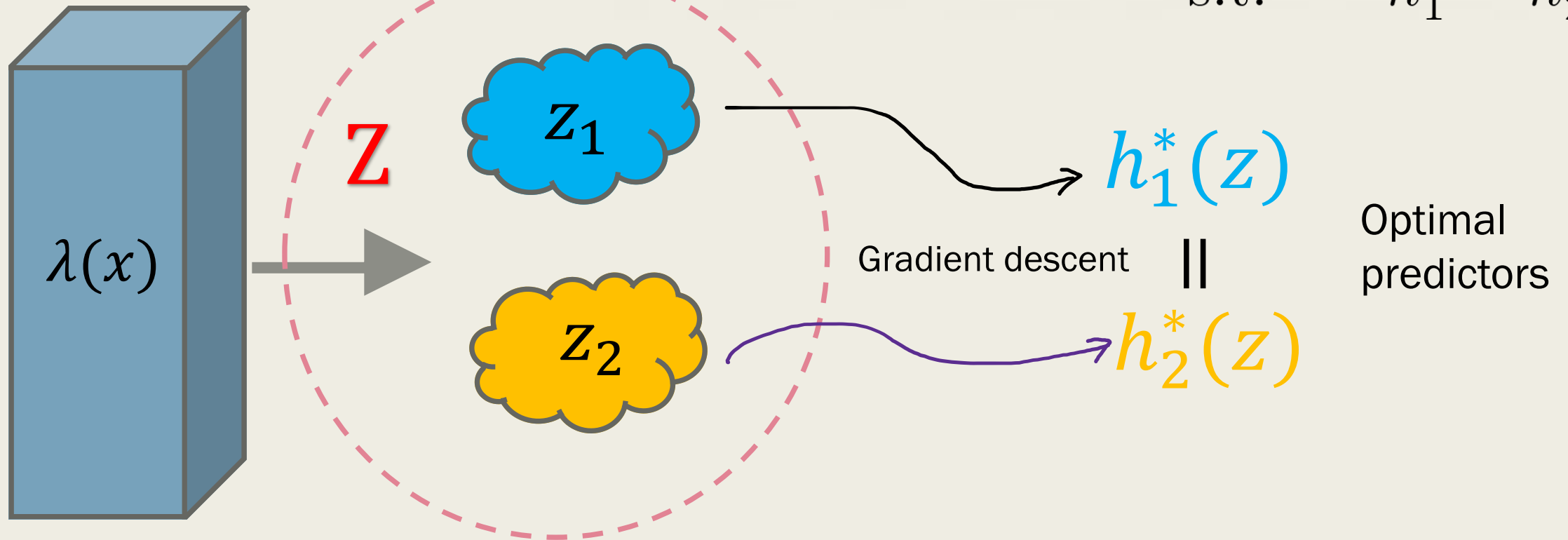


- Adjusting (optimizing) representation to ensure optimal invariant predictor on Z .
- Representation viewed as prior information (or hyper-parameter)

Formulating as bi-level optimization

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

$$\text{s.t.} \quad h_1^* = h_2^*,$$



Efficient optimization through implicit theorem.

Check details in the paper

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