Transformed Distribution Matching for Missing Value Imputation

He Zhao, Ke Sun, Amir Dezfouli, Edwin Bonilla

CSIRO's Data61

Data with missing values

X

				?	
		?		?	
	?				
				?	
			?		
	?				

Examples

Surveys	• A participant didn’t want to answer a question
Medical records	• A patient didn’t take a blood test
Sensor data	• Failures of sensors
…	

Many existing methods

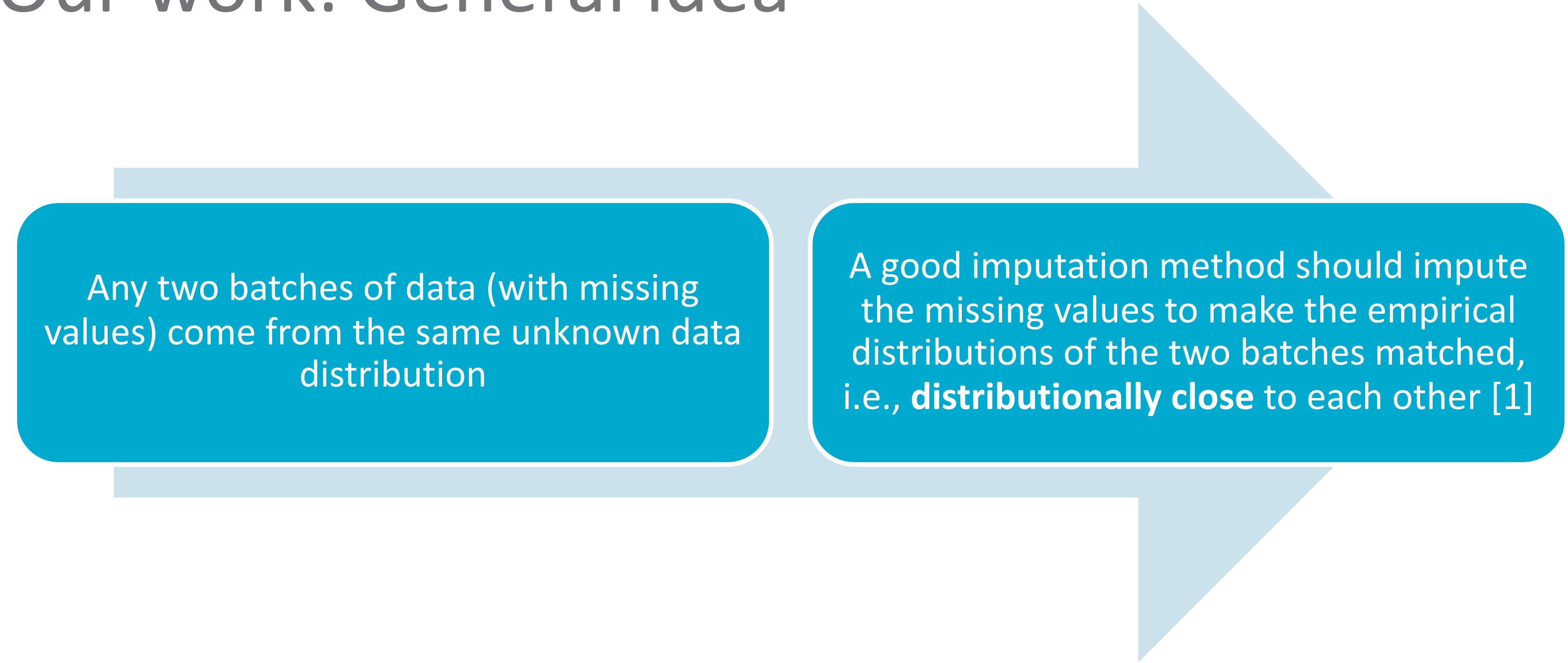
$X \sim p(x)$

				?	
		?		?	
	?				
				?	
			?		
	?				

Data distribution $p(x)$ can be hard to model

- Data has missing values
- Missing values can be generated by different (unknown) mechanisms

Our work: General idea

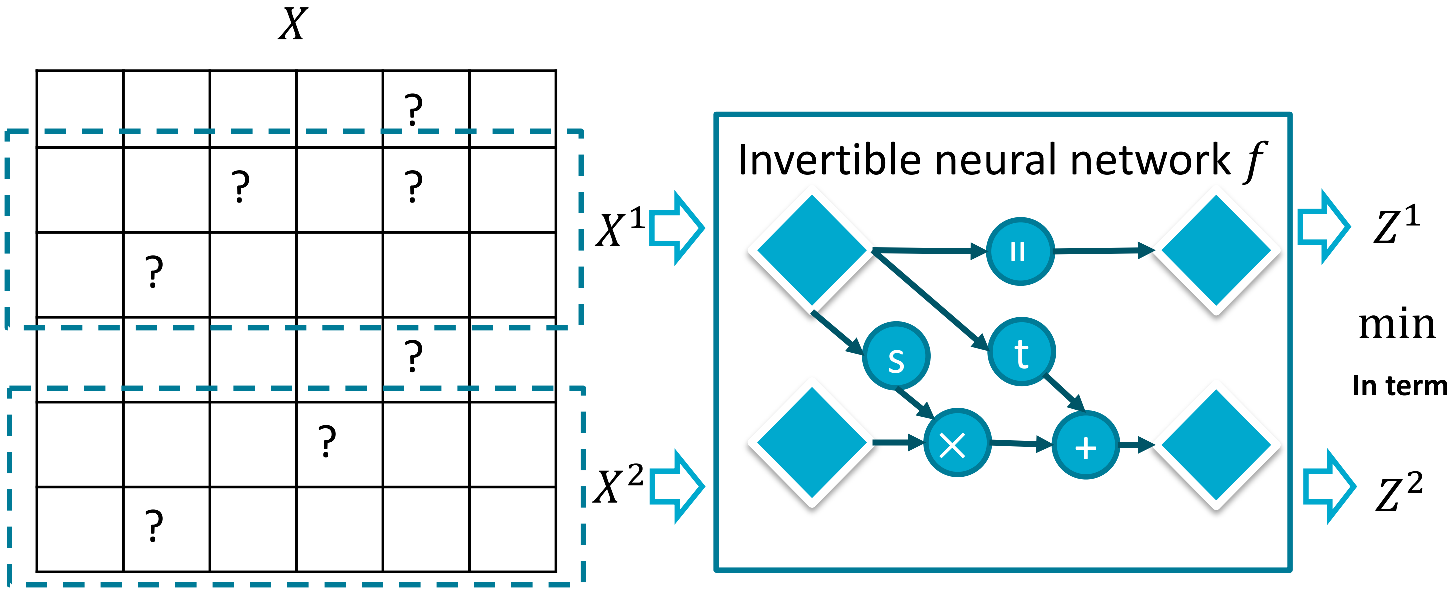


Any two batches of data (with missing values) come from the same unknown data distribution

A good imputation method should impute the missing values to make the empirical distributions of the two batches matched, i.e., **distributionally close** to each other [1]

[1] Muzellec, Boris, et al. "Missing data imputation using optimal transport." *International Conference on Machine Learning*, 2020.

Our work: Framework



X

				?	
		?		?	
	?				
				?	
			?		
	?				

$$\min W_2^2(\mu(Z^1), \mu(Z^2))$$

In terms of the missing values and f

Our work: Appealing properties

High-quality imputations for data with complex geometry

- State-of-the-art imputation performance

Effective regardless of the mechanisms of missing values

- More robust

Easy to train and less parameters to fine-tune

- More applicable in real applications

Thank you!

CSIRO