Backgrounds

Black-Box Models

- High accuracy & low error
- Lacks transparency in decisions

◆ ⋓

Instance-wise Feature Selection (IFS)

▸ Select instance-specific set of important features, by assigning an ***importance score*** to each feature

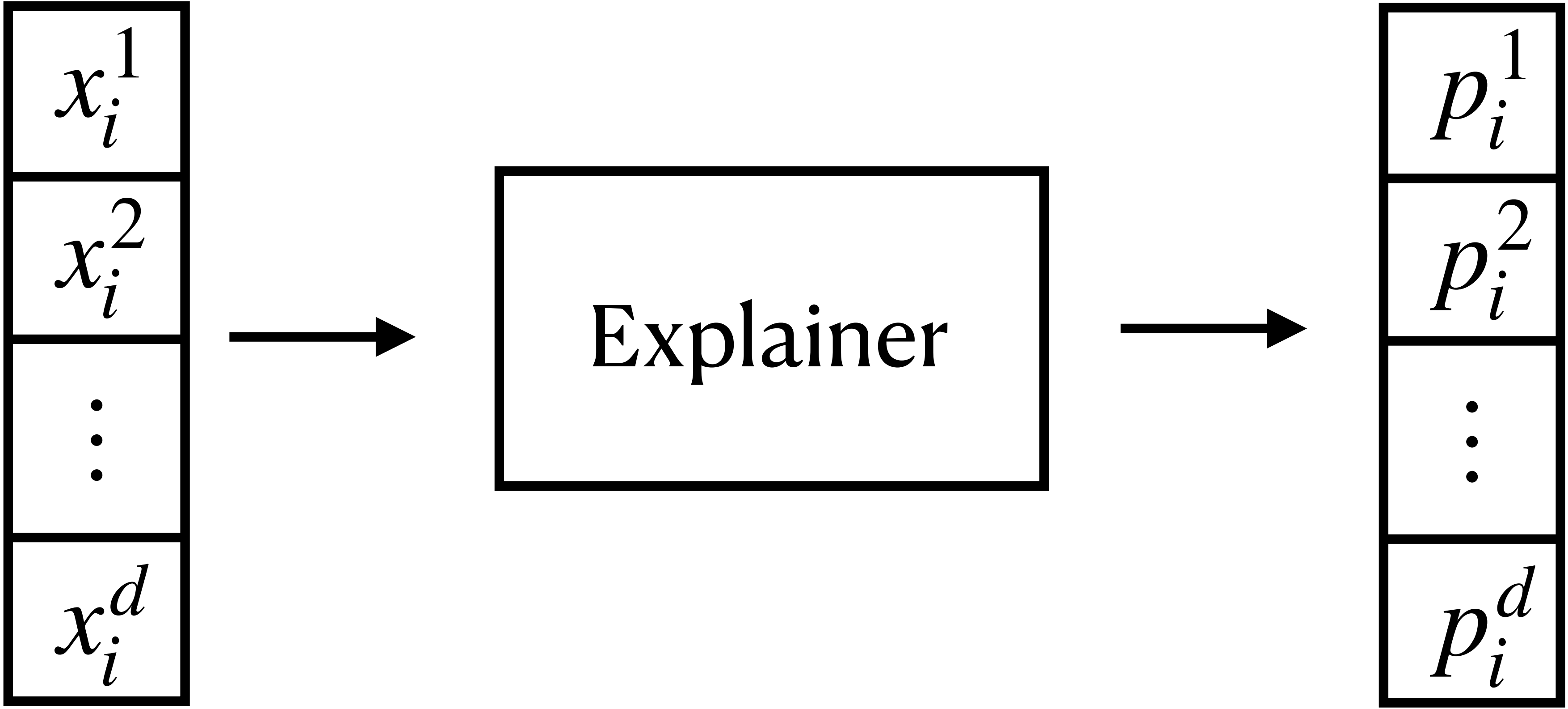
- Instances: $x_i = (x_i^1, x_i^2, \ldots, x_i^d)$
- Black-box model to explain: M

➡ Response variable: $y_i = M(x_i)$

➡ Dataset: $D = (x_i,\ y_i)_{i=1}^{n}$

Explainable AI

- **Techniques** that make decisions of AI models *understandable* and *interpretable* to humans
- Understand whys begind black-box models


x_i^1
x_i^2
$\vdots$
x_i^d
Explainer
p_i^1
p_i^2
$\vdots$
p_i^d

← Instance-specific feature inprotance

Backgrounds

Black-Box Models

- High accuracy & low error
- Lacks transparency in decisions

→

Explainable AI

- **Techniques** that make decisions of AI models *understandable* and *interpretable* to humans
- Understand whys begind black-box models

◆ ⟒

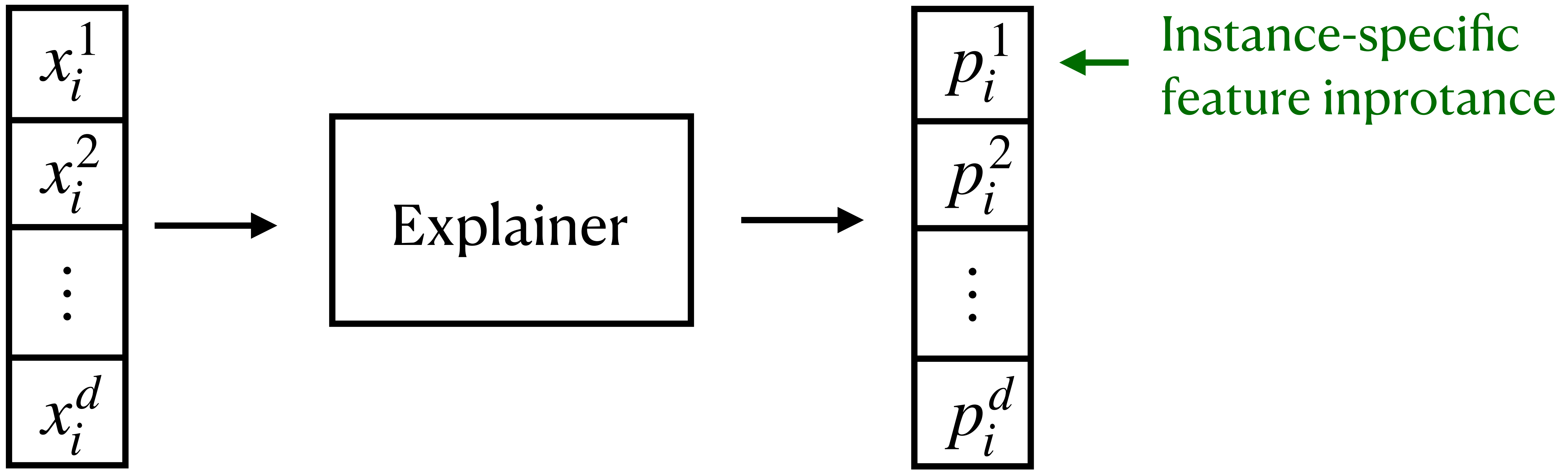
Instance-wise Feature Selection (IFS)

- Select instance-specific set of important features, by assigning an *importance score* to each feature

- Instances: $x_i = (x_i^1, x_i^2, \dots, x_i^d)$
- Black-box model to explain: M

➡ Response variable: $y_i = M(x_i)$

➡ Dataset: $D = (x_i,\ y_i)_{i=1}^n$



Instance-specific feature inprotance

Related Work

IFS Methods {

- Model-specific
- Model-agnostic