Multi-Objective Particle Swarm Optimization for Feature Selection in Credit Scoring

Nikita Kozodoi[1,2] Stefan Lessmann[1]

nikita.kozodoi@hu-berlin.de

[1] Humboldt University of Berlin

[2] Monedo

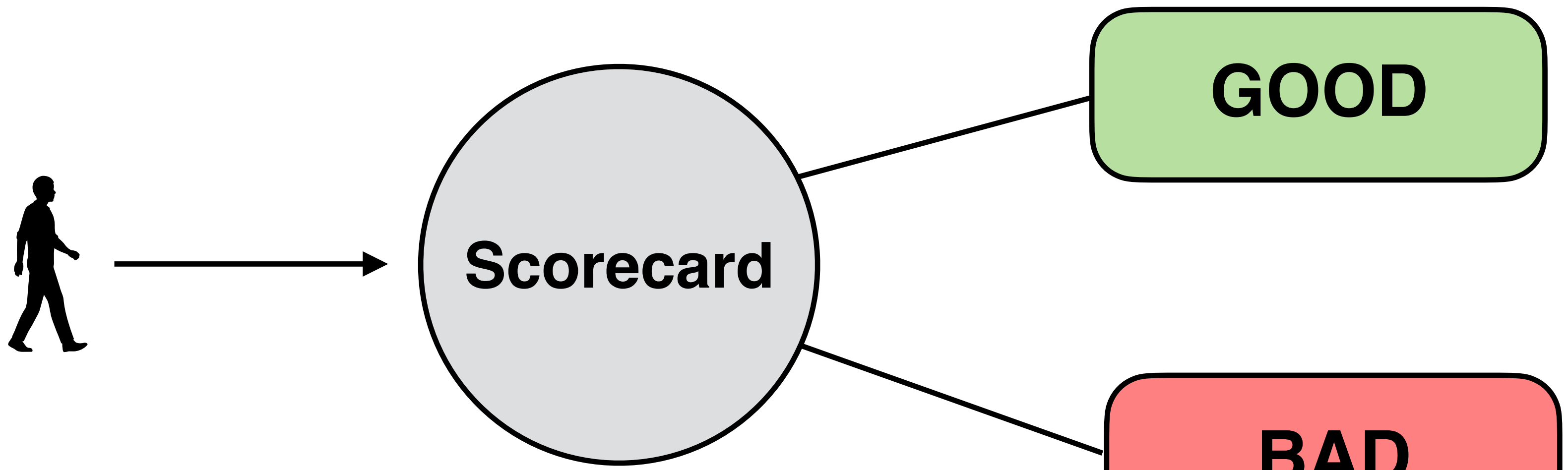
Presentation Outline

1. Problem Setting
2. Proposed Feature Selection Framework
3. Experimental Design
4. Empirical Results
5. Conclusions

Background

Credit scoring:

- the use of **statistical models** to support decision-making in the **retail credit sector** *(Crook et al. 2007)*
- classification task of distinguishing **BAD** and **GOOD** loans
- **scorecard** — model that estimates probability of default



Background

Credit scoring:

- the use of **statistical models** to support decision-making in the **retail credit sector** *(Crook et al. 2007)*
- classification task of distinguishing **BAD** and **GOOD** loans
- **scorecard** — model that estimates probability of default

Common data sources:

- application data
- credit bureau data
- transaction history
- geographical data
- social media
- …

Background

Credit scoring:

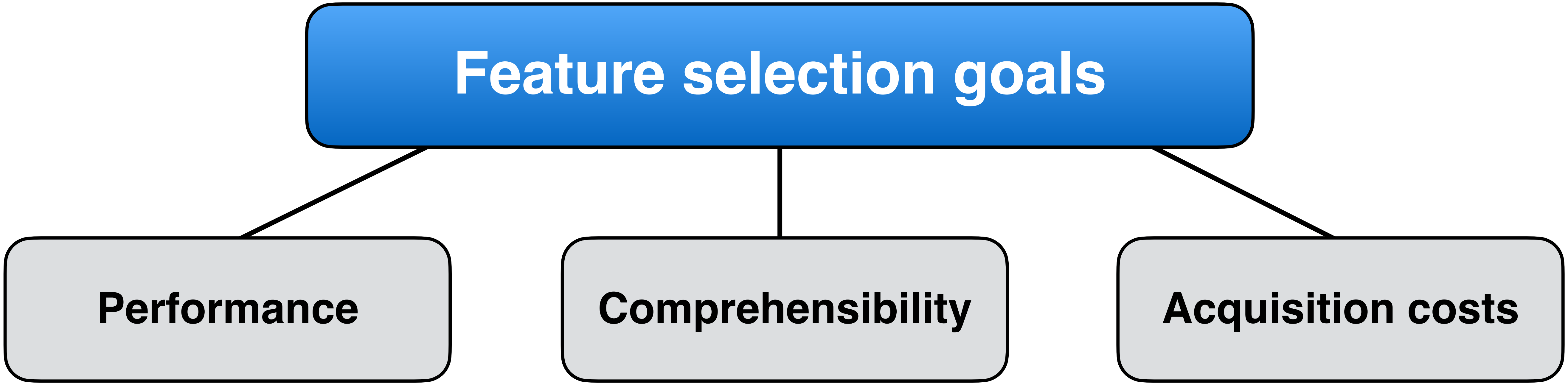
- the use of **statistical models** to support decision-making in the **retail credit sector** *(Crook et al. 2007)*
- classification task of distinguishing **BAD** and **GOOD** loans
- **scorecard** — model that estimates probability of default

Common data sources:

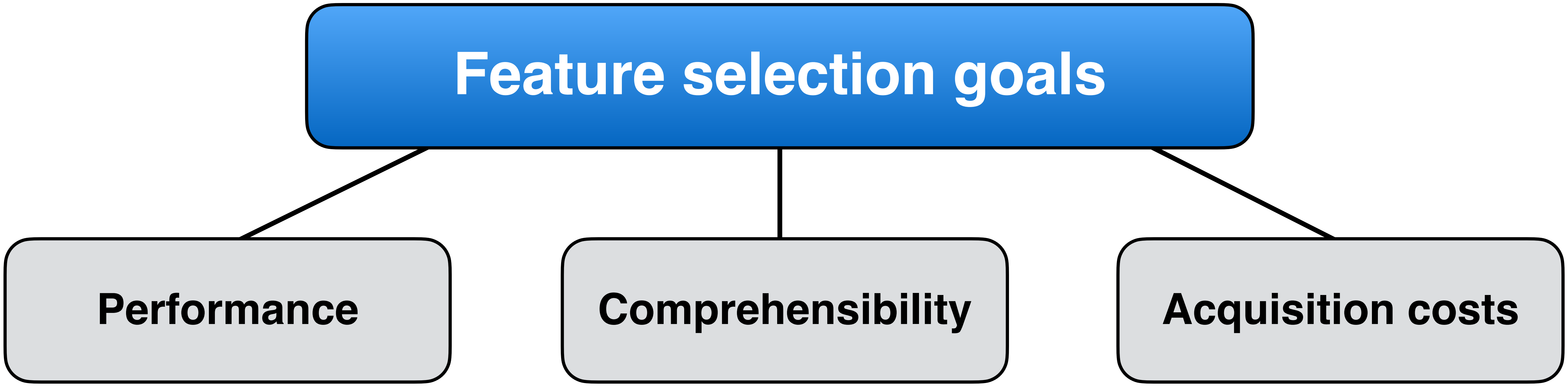
- application data
- credit bureau data
- transaction history
- geographical data
- social media
- …

High dimensionality emphasizes importance of feature selection

Feature Selection in Credit Scoring: Objectives



Feature Selection in Credit Scoring: Objectives



- **scorecard accuracy** affects profit
- usually measured in **AUC** (ranking loan applicants)

Feature Selection in Credit Scoring: Objectives

Feature selection goals

- **Performance**
- **Comprehensibility**
- **Acquisition costs**

Performance:

- **scorecard accuracy** affects profit
- usually measured in **AUC** (ranking loan applicants)

Comprehensibility:

- regulations enforce **comprehensible** models
- data **storage costs**

Feature Selection in Credit Scoring: Objectives

Feature selection goals

Performance

- **scorecard accuracy** affects profit
- usually measured in **AUC** (ranking loan applicants)

Comprehensibility

- regulations enforce **comprehensible** models
- data **storage costs**

Acquisition costs

- features are usually **purchased in groups**
- providers have different **payment options**
- this **decorrelates** feature number and their cost

Feature Selection in Credit Scoring: Objectives

Feature selection goals

Performance

- **scorecard accuracy** affects profit
- usually measured in **AUC** (ranking loan applicants)

Comprehensibility

- regulations enforce **comprehensible** models
- data **storage costs**

Acquisition costs

- features are usually **purchased in groups**
- providers have different **payment options**
- this **decorrelates** feature number and their cost

It is important to account for three distinct objectives

Multi-Objective Feature Selection

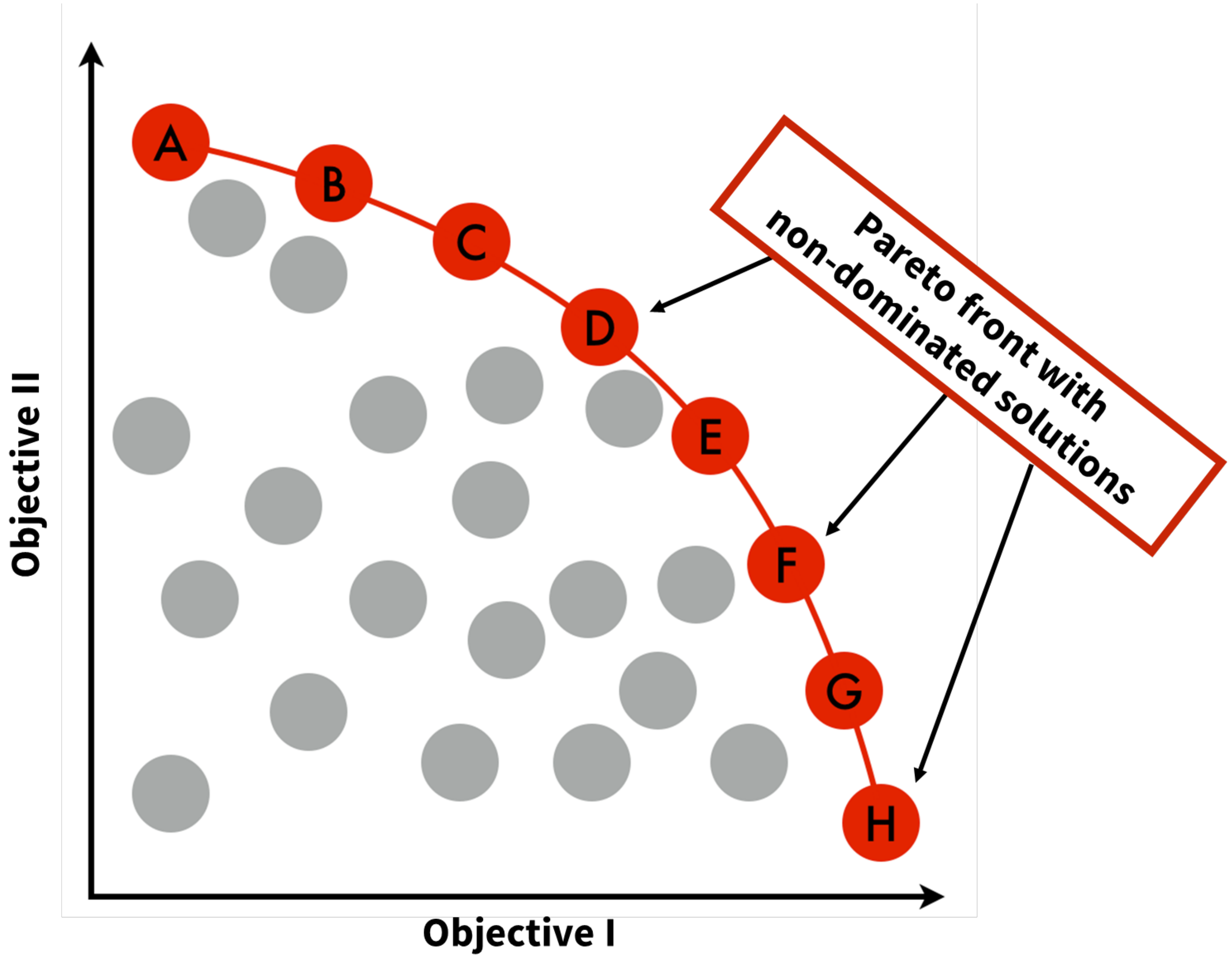
Multi-objective methods:

- **Weighting** multiple objective into one
- Introducing a **budget constraint**
- Optimizing multiple **distinct objectives**

Multi-Objective Feature Selection

Multi-objective methods:

- **Weighting** multiple objective into one
- Introducing a **budget constraint**
- Optimizing multiple **distinct objectives**



Multi-Objective Feature Selection

Search algorithms:

- Genetic algorithms (GA)
  - **NSGA-II** — well-known optimization algorithm *(Hambdani et al. 2007)*
  - **NSGA-III** — handles issues with many objectives *(Bidgoli et al. 2019)*
- Particle swarm optimization (PSO)
  - **outperforms GAs** in optimization tasks *(Zhu et al. 2017)*

Multi-Objective Feature Selection

Search algorithms:

- Genetic algorithms (GA)
  - **NSGA-II** — well-known optimization algorithm *(Hambdani et al. 2007)*
  - **NSGA-III** — handles issues with many objectives *(Bidgoli et al. 2019)*
- Particle swarm optimization (PSO)
  - **outperforms GAs** in optimization tasks *(Zhu et al. 2017)*

Credit scoring applications:

- **SVM-based** feature selection *(Maldonado et al. 2015; 2017)*
  - optimizes performance and feature costs
  - can only be used with SVMs
- **NSGA-II** *(Kozodoi et al. 2019)*
  - two objectives: number of features and model performance

Proposed Feature Selection Framework

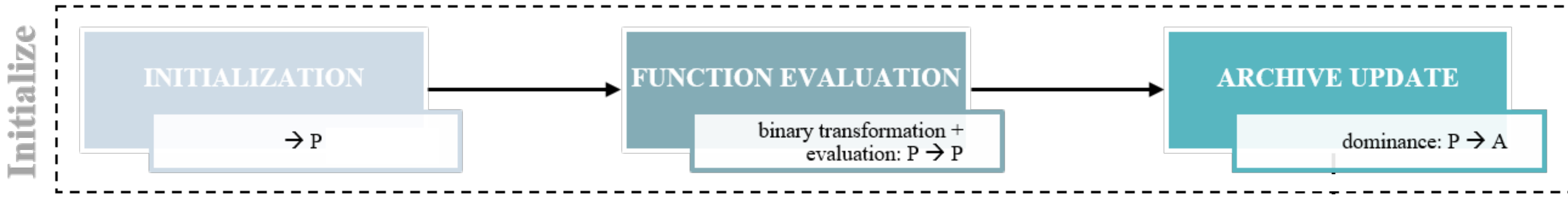
Objectives:

- AUC of the scorecard
- number of selected features
- data acquisitions costs

Feature search:

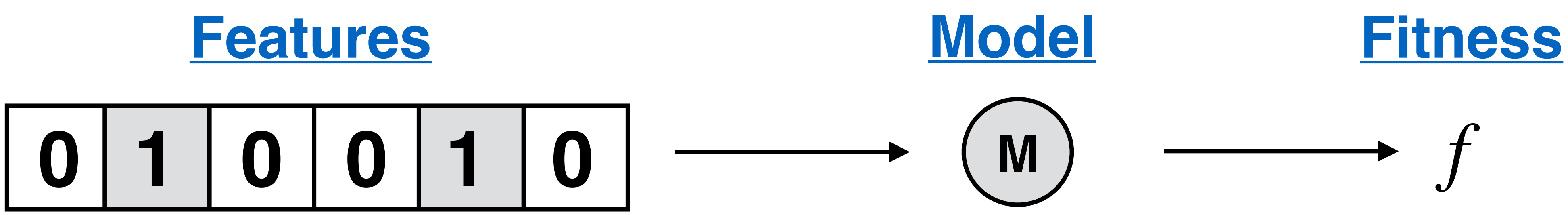
- adapting a PSO-based algorithm to improve feature search

Feature Selection with AgMOPSO

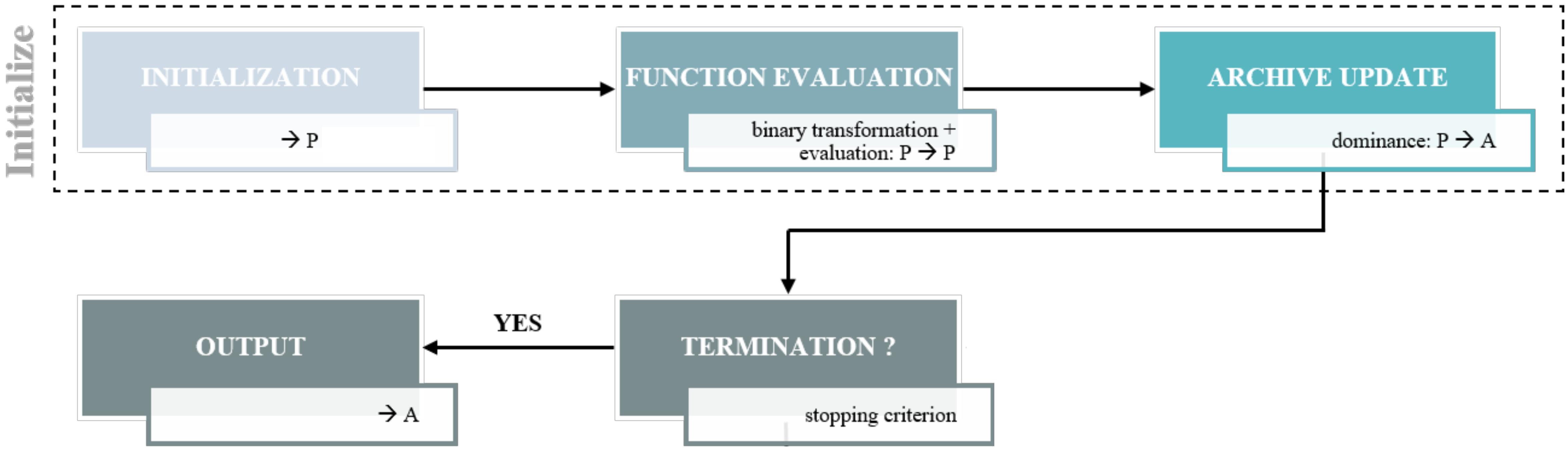


Swarm of particles

Training a scorecard



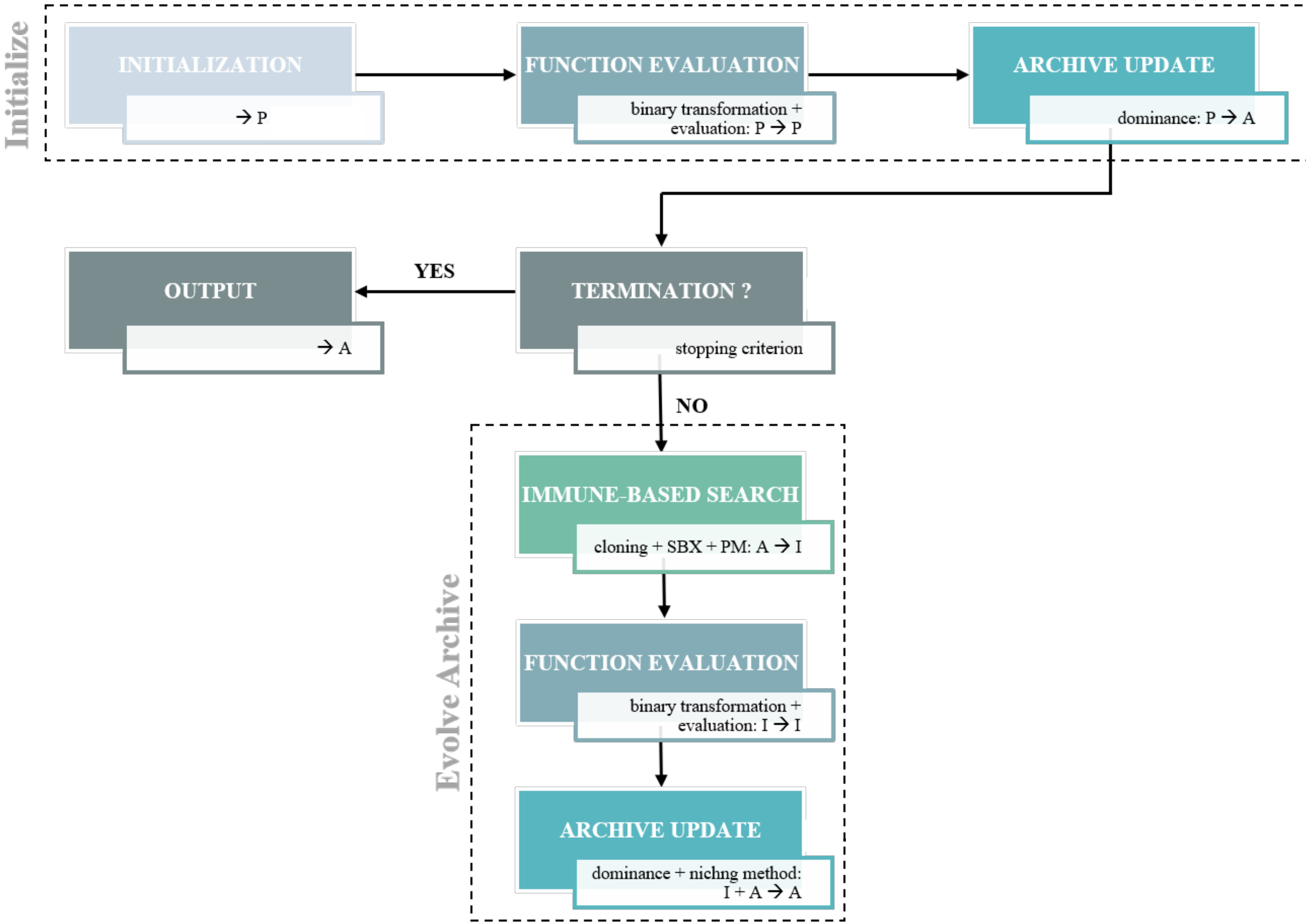
Feature Selection with AgMOPSO



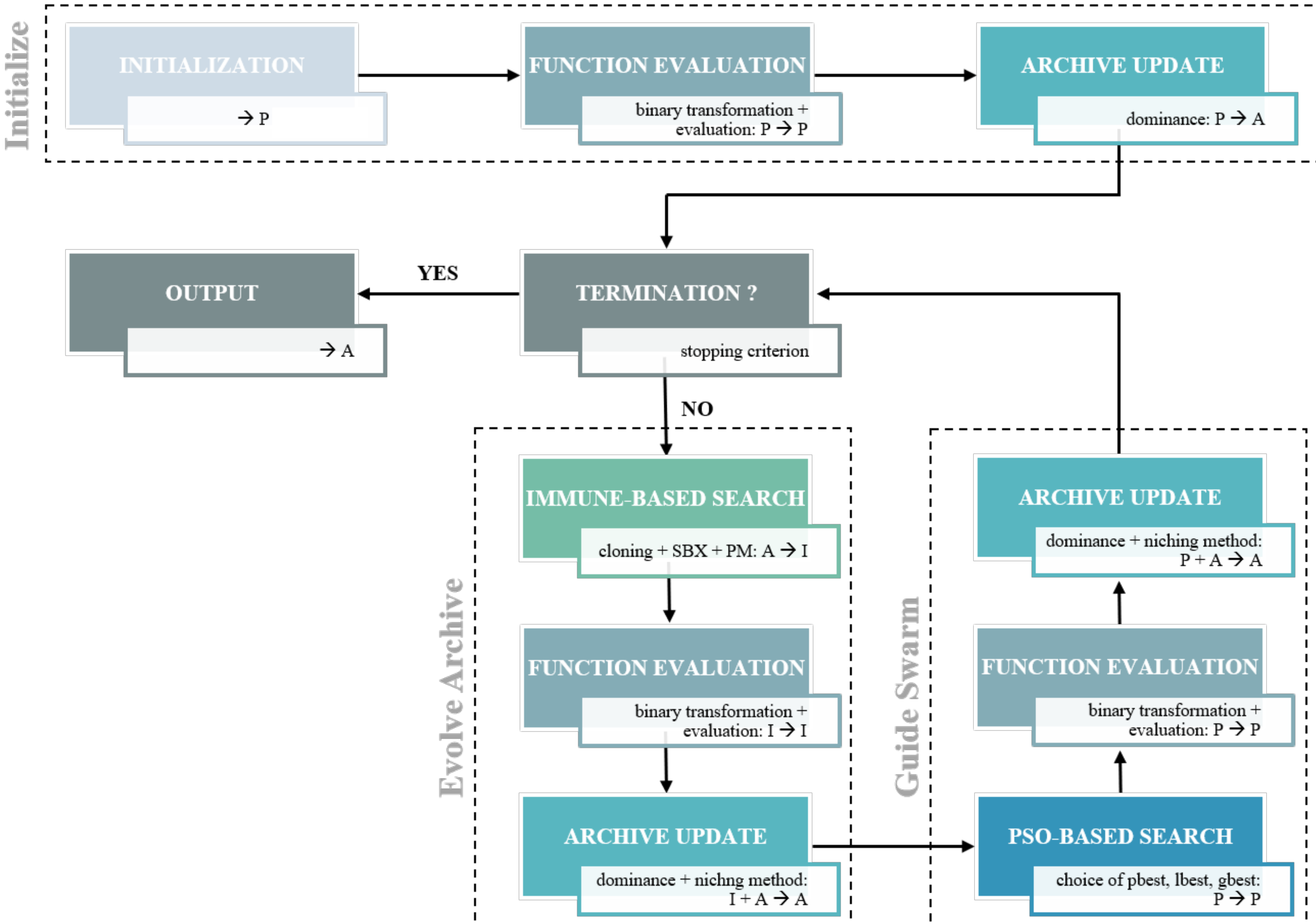
Max. number of generations

Pareto frontier

Feature Selection with AgMOPSO



Feature Selection with AgMOPSO



Credit Scoring Data Sets

Data Label	Sample Size	No. Features	Default Rate
australian	690	42	.44
german	1,000	61	.30
thomas	1,125	28	.26
hmeq	5,960	20	.20
cashbus	15,000	1.308	.10
lending club	43,344	206	.07
packdd	50,000	373	.26
paipaidai	60,000	1.934	.07
gmsc	150,000	68	.07

Experimental Setup

1. Simulate data acquisition costs

- **continuous features:** draw from Uniform distribution
- **categorical features:** group-based cost for dummies

2. Data partitioning

- **training (70%):** feature selection within 4-fold CV
- **holdout (30%):** performance evaluation

3. Benchmarks

- AgMOPSO
- NSGA-II
- NSGA-III

} multi-objective methods

- Full model with all features

Results: Performance

Algorithm	ONVG	TSC	SPC	SPR	HV
NSGA-II	1.86	1.97	2.37	**1.49**	1.98
NSGA-III	2.31	1.99	**1.45**	2.48	2.23
AgMOPSO	**1.61**	**1.79**	2.13	1.78	**1.74**

Cardinality

Convergence

Diversity

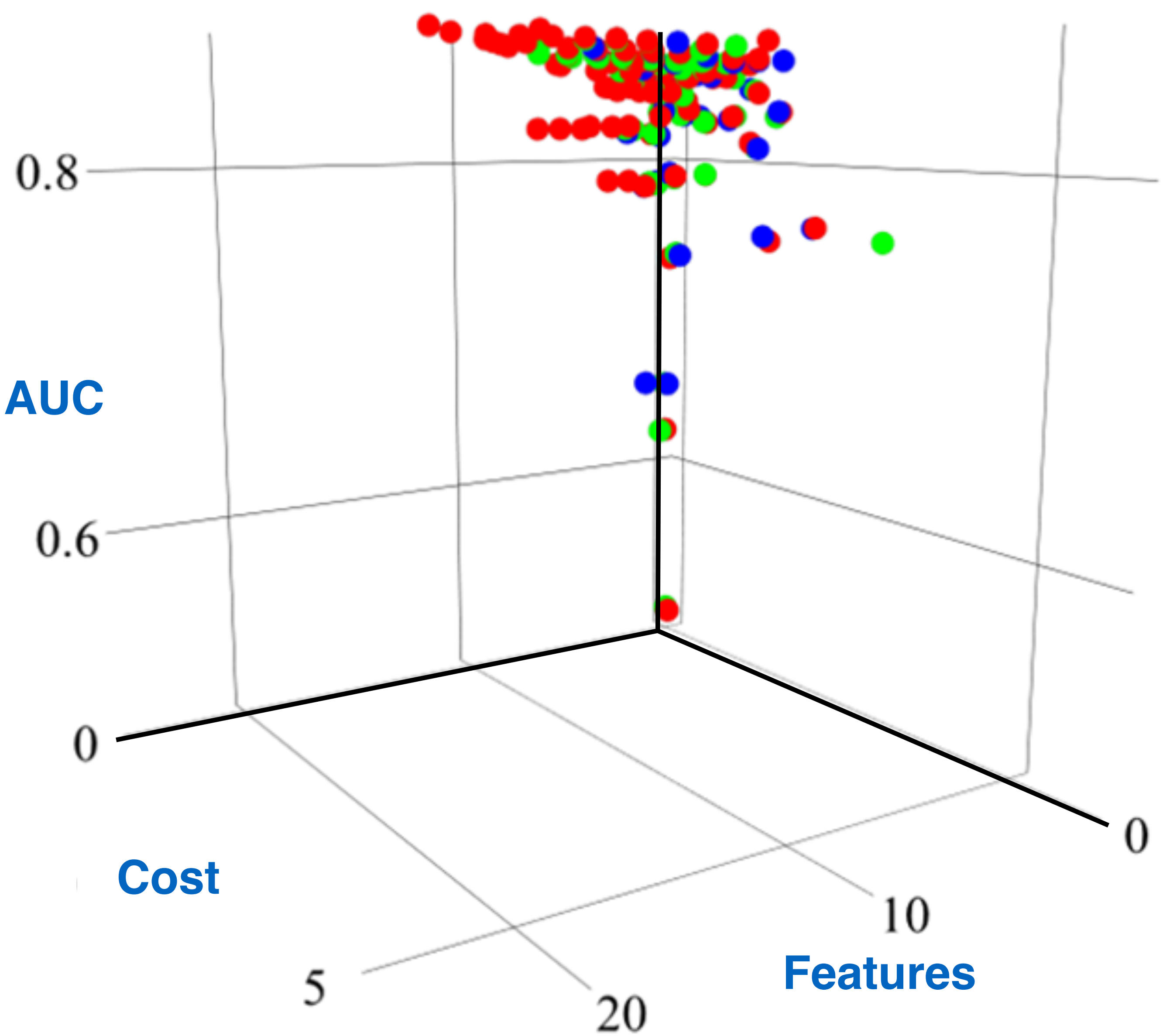
Overall performance

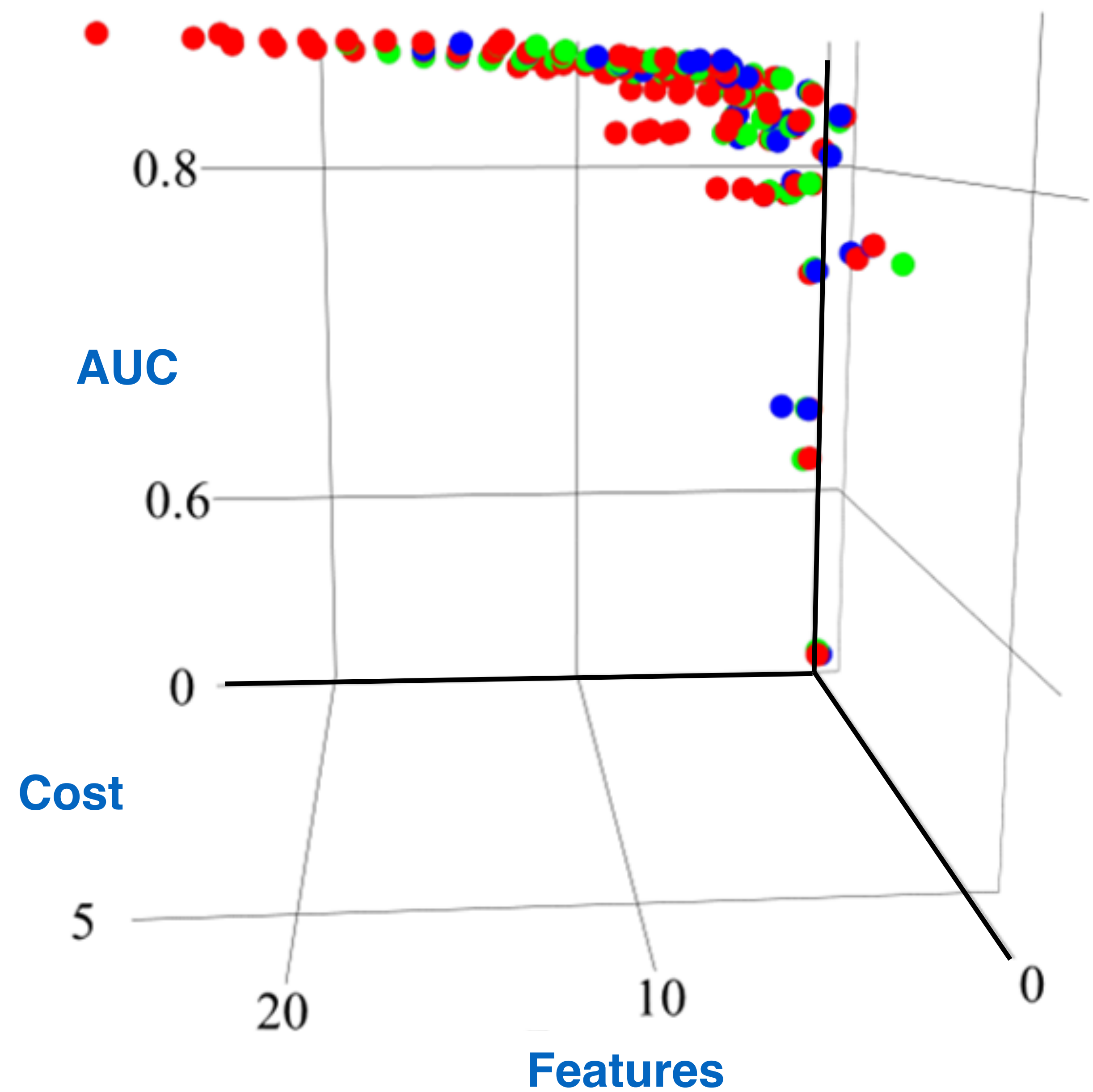
Results: Performance

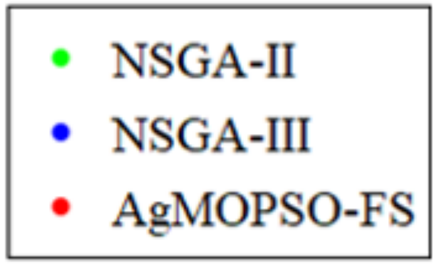
Algorithm	ONVG	TSC	SPC	SPR	HV	AUC	NF	DAC
NSGA-II	1.86	1.97	2.37	**1.49**	1.98	2.33	2.33	2.44
NSGA-III	2.31	1.99	**1.45**	2.48	2.23	2.33	**1.22**	**1.11**
AgMOPSO	**1.61**	**1.79**	2.13	1.78	**1.74**	**1.67**	2.44	2.44
Full Model	—					3.67	4.00	4.00

AgMOPSO evolves solutions in the region with high AUC better than competitors

Results: Pareto Frontiers







Example Pareto frontier on *gsmc*

Summary & Questions

1. **Problem setting**
   - **conflicting goals** of feature selection in credit scoring
   - purchasing data **decorrelates number and cost of features**
2. **New feature selection framework**
   - **optimizes three objectives:** AUC, number of features, feature costs
   - uses **PSO algorithm** for feature search
3. **Experiments on real-world data sets**
   - **competitive performance** compared to other multi-objective methods
   - efficiently explores search space with **high AUC**