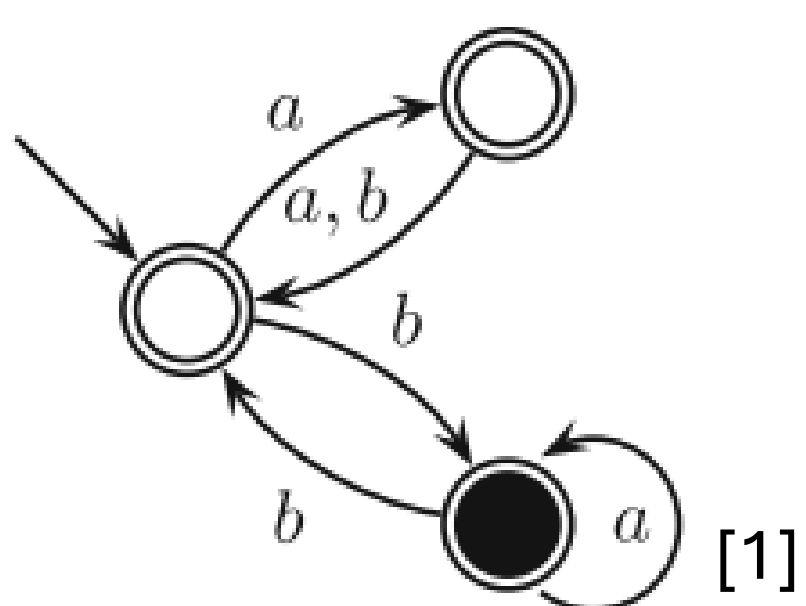


1. Deterministic Finite Automata (DFA)

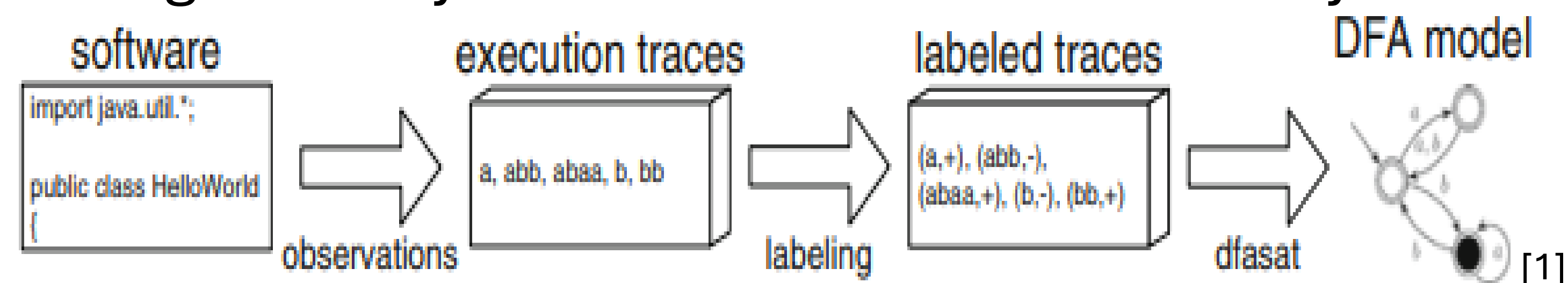
DFA – Set of states, transitions, alphabet -> A model that recognizes a certain regular language.



DFAs can act as surrogate models for software systems. However, the process of creating and maintaining them is costly and inefficient, and is usually omitted during software development.

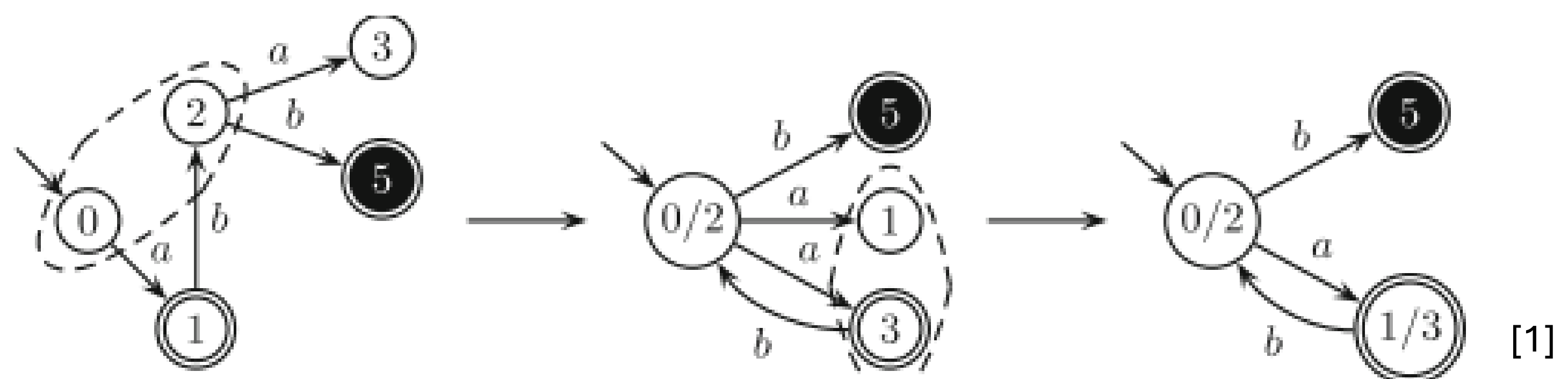
2. DFA Identification

Software Model Synthesis – models are learned through already known execution traces of a system.

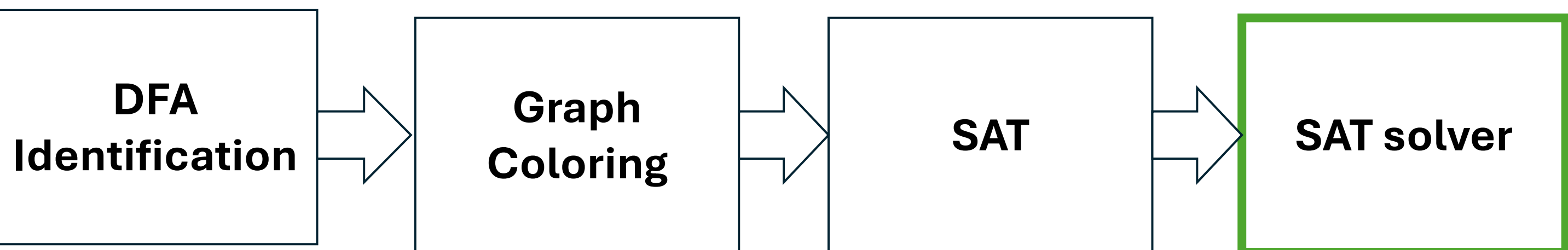


DFA Identification is the problem of finding a smallest DFA, consistent with a set of given labeled examples. Approaches:

- **Heuristic approach:** Methods such as the Evidence-driven state merging (**EDSM**) algorithm



- **Optimal methods:** Methods for exact DFA minimization such as **reduction to SAT**.



3. Research Question

To what extent and in what ways does **partially learning a model heuristically and then applying exact minimization** - based on the decisions already made by the heuristic - affect test-performance in DFA identification?

4. Experimental setup

Use **FlexFringe** (software for running heuristic and exact methods) for DFA identification experiment.

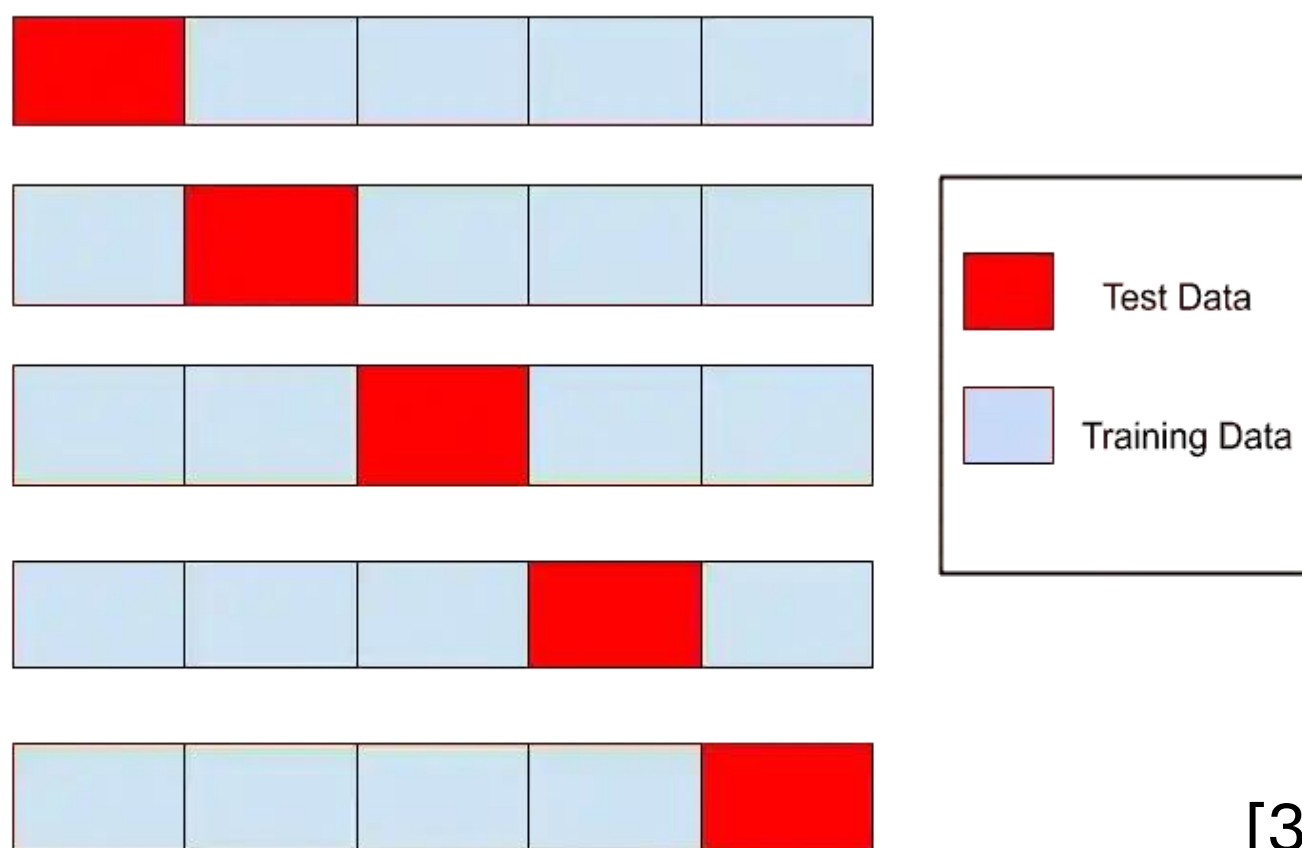
Take inspiration from the **STAMINA competition** setup:

- **BCR scores** for measuring test performance
- Datasets with different difficulties

Σ	Sparsity			
	100%	50%	25%	12.5%
2	0.99 (1)	0.95 (1)	0.67 (3)	0.66 (3)
5	0.97 (1)	0.78 (2)	0.59 (4)	0.52 (4)
10	0.93 (1)	0.64 (3)	0.51 (4)	0.5 (4)
20	0.91 (1)	0.63 (3)	0.54 (4)	0.51 (4)
50	0.81 (2)	0.64 (3)	0.57 (4)	0.5 (4)

Test the hybrid (heuristic, then exact) approach on:

- STAMINA competition training sets with **5-fold cross validation**



Independent variable:

- When to switch from heuristic to optimal

Dependent variable:

- Resulting model (size, test-performance)

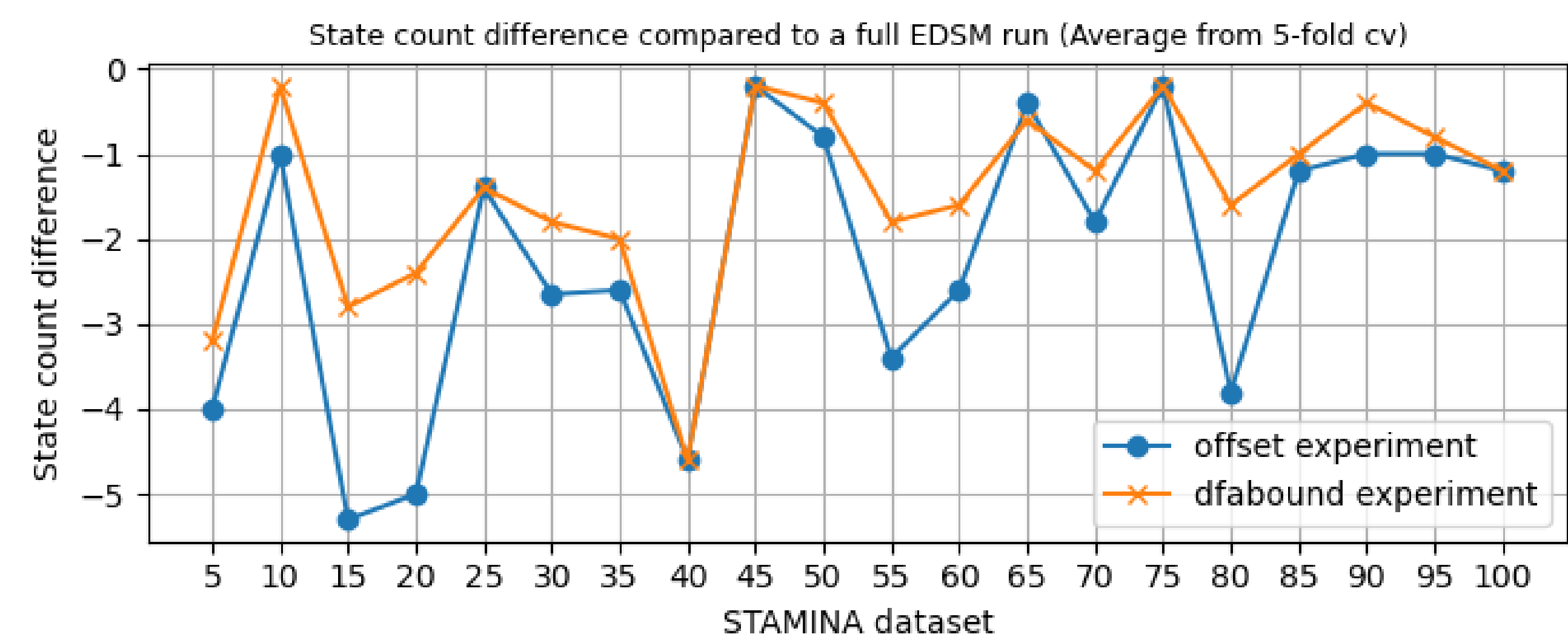
Implementations of the hybrid approach:

- Binary search on the size (**DFA bound**) of the partial automaton before we switch
- Binary search on the **SAT offset** after fixing the point at which we switch

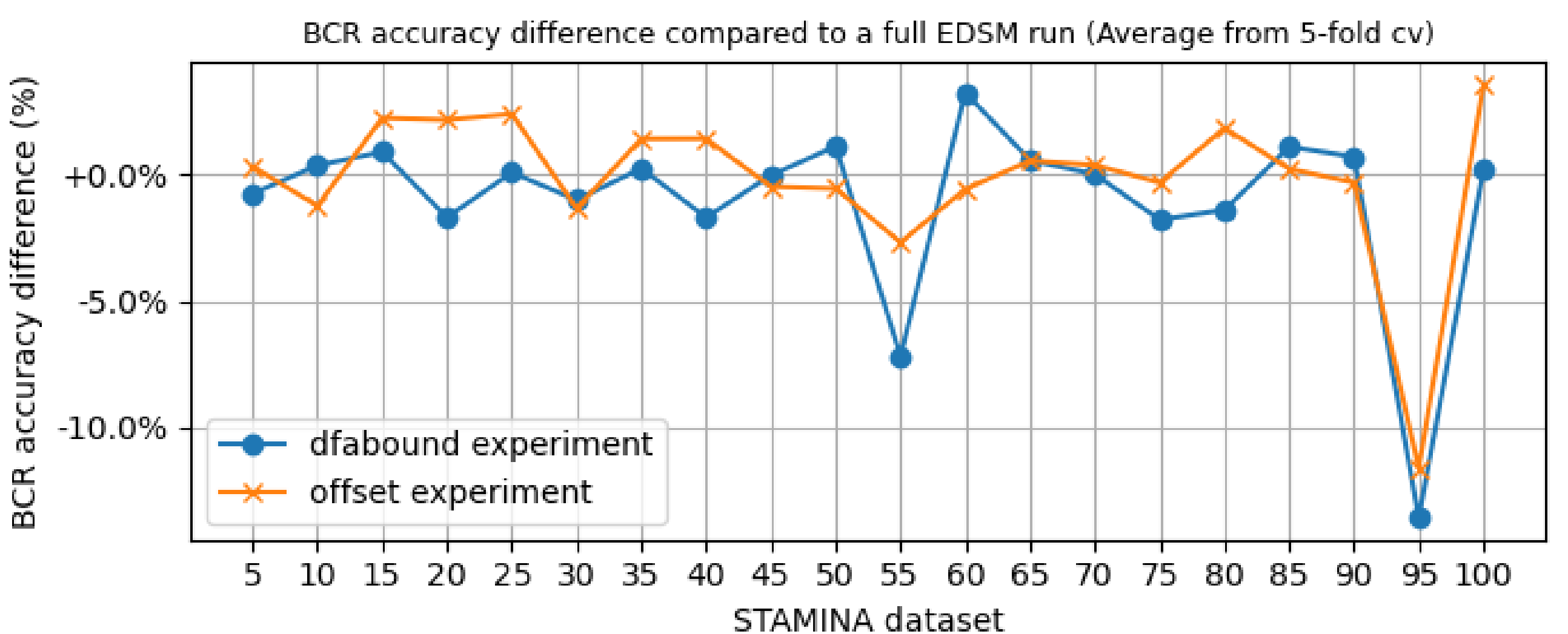
5. Results

Result differences compared to pure EDSM:

- In terms of **state counts**



- In terms of **BCR scores**



Correlations with dataset characteristics:

- Sparser datasets lead to **bigger** differences with EDSM (in terms of state count)
- Larger alphabets lead to **smaller** differences with EDSM (in terms of state counts)

6. Conclusions and future work

- Exact methods make for a bigger part of the hybrid approach => generally smaller models
- Hybrid approach identifies smaller DFAs, but that does not consistently improve accuracy
- Generate more and different datasets
- Try stopping the greedy merging earlier (would require higher computational effort)
- Implement the entire DFASAT pipeline