# A theoretical analysis of optimal and heuristic methods for DFA learning

How much less data is necessary to identify the correct model when exact minimal methods are used as compared to heuristics?

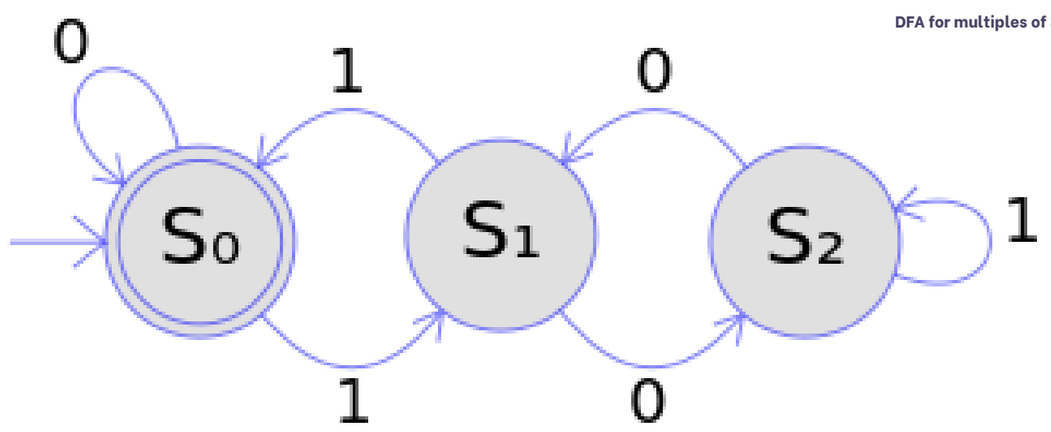

**Author**
Horia Radu - horiaradu@tudelft.nl

**Supervisin Team**
Dr. Sicco Verwer, Simon Dieck

**Affiliations**
EEMCS, TU Delft


## Introduction

Automata learning aims to infer a deterministic finite automaton (DFA) that models system behavior from sequence data. These DFAs serve as interpretable surrogates for software analysis. Guided by Occam's Razor, minimal DFAs are preferred for their simplicity and clarity.
Heuristic methods like EDSM and Alergia approximate minimal DFAs, while exact methods (e.g., SAT-based) guarantee them. This research asks: Do exact methods require less data to learn correct models?.




## Objective

We explore:

- Do exact methods require less data than heuristics?
- Is the other way around true?
- When and how much less?
- Can this be proven mathematically?

## Methodology

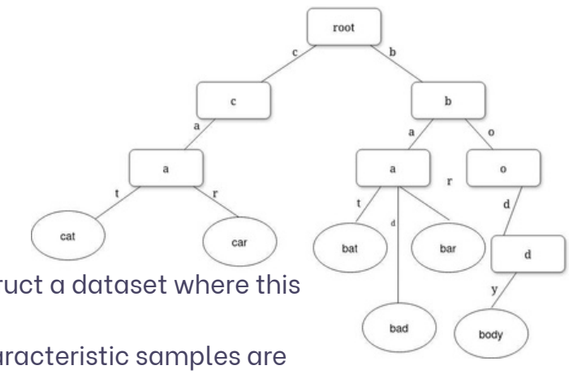
- Brainstorming
- Reading literature
- Experiments to test hypotheses
- Discussions with supervisors and peers
- Coffee

## Results

- L = a → heuristic can be just as good
- L =(abc)+ → heuristic can outperform unless properly defined
- Same dataset, different DFAs → heuristic always possibly outperforms (from a certain POV)
- Plain EDSM always at least as good
- BlueFringe as efficient as optimal

## Final Proof

A proof that shows how one always needs the same amount of data to achieve a DFA of certain size that recognizes an input for both the BlueFringe framework and optimal methods.




Key points:

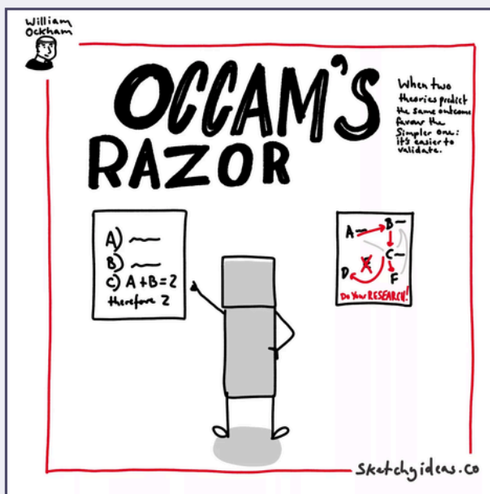
- Equivalence classes
- Show that it is impossible to construct a dataset where this is not the case
- Only interested in finding if the characteristic samples are equal in size as by definition the optimal method is at least as small for this definition

This shows that data efficiency for learning DFAs is method independent.

## Myhill Nerode Theorem:

The following three statements are equivalent

1. The set L ϵ $\sum^*$ is accepted by a FSA
2. L is the union of some of the equivalence classes of a right invariant equivalence relation of finite index.
3. Let equivalence relation $R_L$ be defined by :
   $xR_Ly$ iff for all z in $\sum^*$ xz is in L exactly when yz is in L.
   Then $R_L$ is of finite index.




## Conclusion

- Heuristics can be on the same level as optimal methods
- Some heuristics can outperform optimal methods
- BlueFringe as data efficient as the optimal method under some restrictions
- In the end no rule that applies universally, a very simple answer

Future research will likely have to revolve around expected / average performance of heuristics vs optimal methods.

## Related Literature


1. Angluin, D.: Learning regular sets from queries and counterexamples. Information and computation
75(2), 87–106 (1987)
2. Gold, E.M.: Language identification in the limit. Information and Control 10(5), 447–474 (1967). https://doi.org/https://doi.org/10.1016/S0019-9958(67)91165-5, https://www.sciencedirect.com/science/article/pii/S0019995867911655
3. Lang, K.J., Pearlmutter, B.A., Price, R.A.: Results of the abbadingo one DFA learning competition and
a new evidence-driven state merging algorithm. In: International Colloquium on Grammatical Inference.
pp. 1–12. Springer (1998)
4. Nerode, A., Sauer, B.P.: Fundamental Concepts in the Theory of Systems. ASTIA Document, Wright Air Development Center, Air Research and Development Command, United States Air Force (1957), https://books.google.nl/books?id=QjZwlSLU4rAC
5. Verwer, S., Heule, M.J.: Exact DFA identification using SAT solvers. In: Grammatical Inference: Theo-
retical Results and Applications: 10th International Colloquium, ICGI 2010, Valencia, Spain, September
13–16, 2010. Proceedings 10. pp. 66–79. Springer (2010