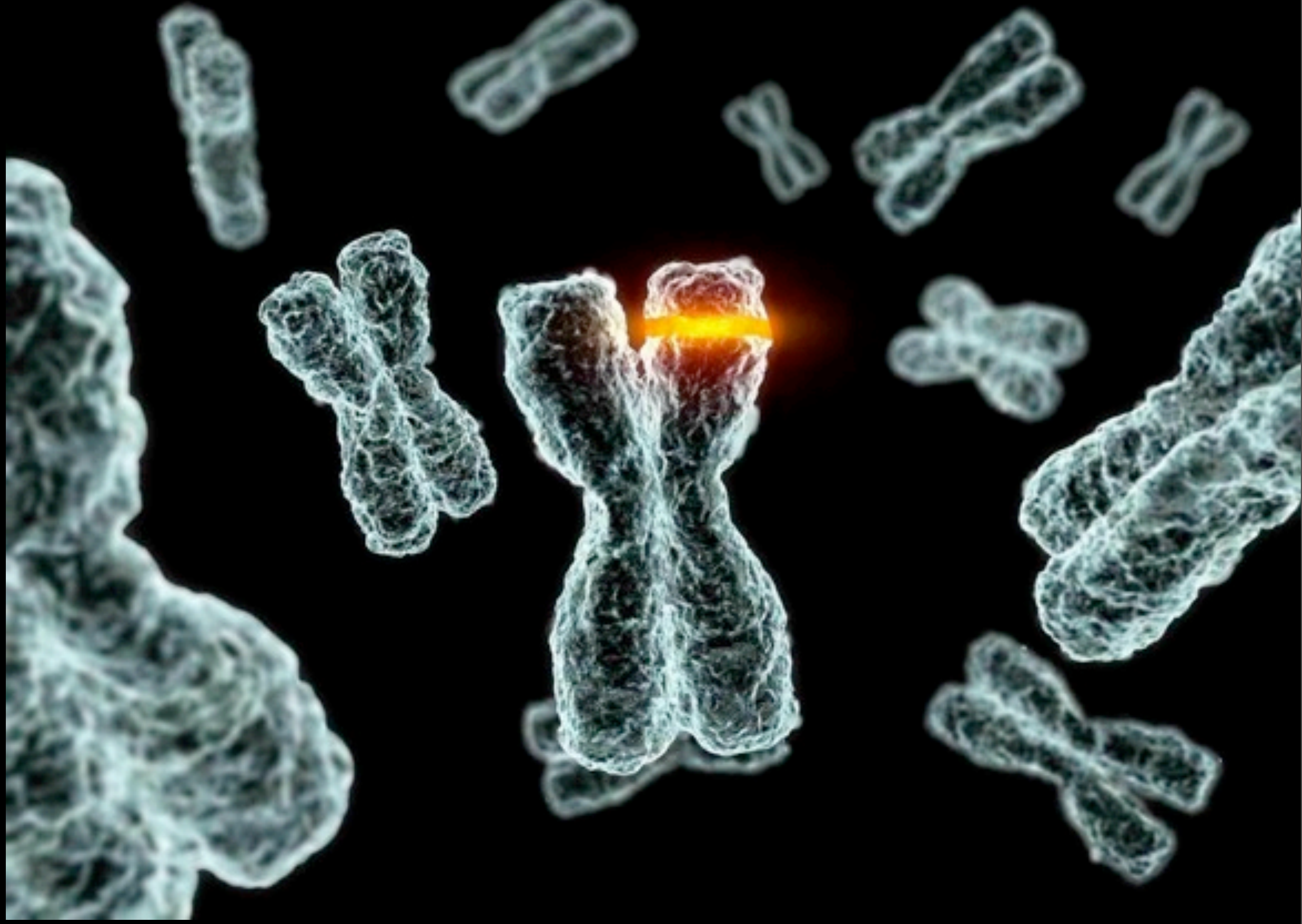
Automatic Design of Decision-Tree Induction Algorithms

Rodrigo C. Barros
Márcio P. Basgalupp
André C.P.L.F. de Carvalho
Alex A. Freitas

University of São Paulo, Brazil
Federal University of São Paulo, Brazil
University of Kent, UK



Submission

- Hyper-heuristic EA that automatically Designs Decision Tree induction algorithms (HEAD-DT)
- Currently, the first (and only) approach that automates the design of full top-down decision tree induction algorithms
- Best paper award at GECCO '12 (IGEC+S*S +SBSE)

OBJECTIVE

- To **automatically design a new algorithm** for decision tree induction (i.e., an algorithm that models data by decision trees)
- For this automatic algorithm design, we propose an Evolutionary Algorithm (EA) system: HEAD-DT
- Note: there are many EAs that induce **decision trees** for a given data set, but HEAD-DT is very different
    - HEAD-DT creates a new **generic** decision tree induction **algorithm**, which can be used to discovery decision trees **in any classification data set**

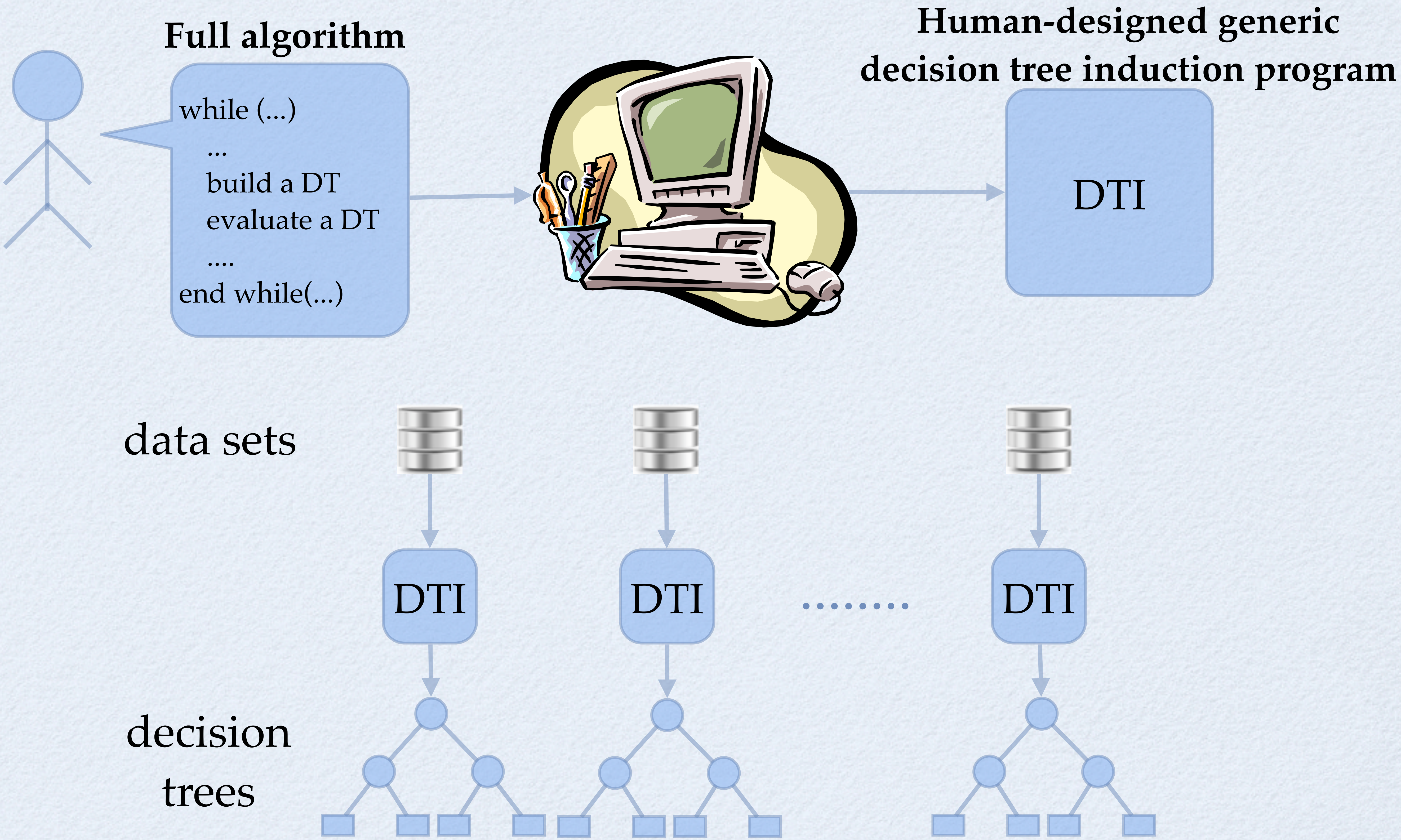
Automating the design of decision tree induction algorithms

- Timeline of the manual design of DT induction algorithms

Kass	Breiman	Quinlan	Quinlan	Breiman	...
CHAID	CART	ID3	C4.5	Random Forest	...
1980	1984	1986	1985	2001	...

- HEAD-DT replace this manual, “ad-hoc”, evolution with an automatic, “data-driven”, evolution of DT induction algorithms

Manual invention of decision tree induction algorithms



Full algorithm

```
while (...)
   ...
   build a DT
   evaluate a DT
   ....
end while(...)
```

Human-designed generic decision tree induction program

DTI

data sets

DTI

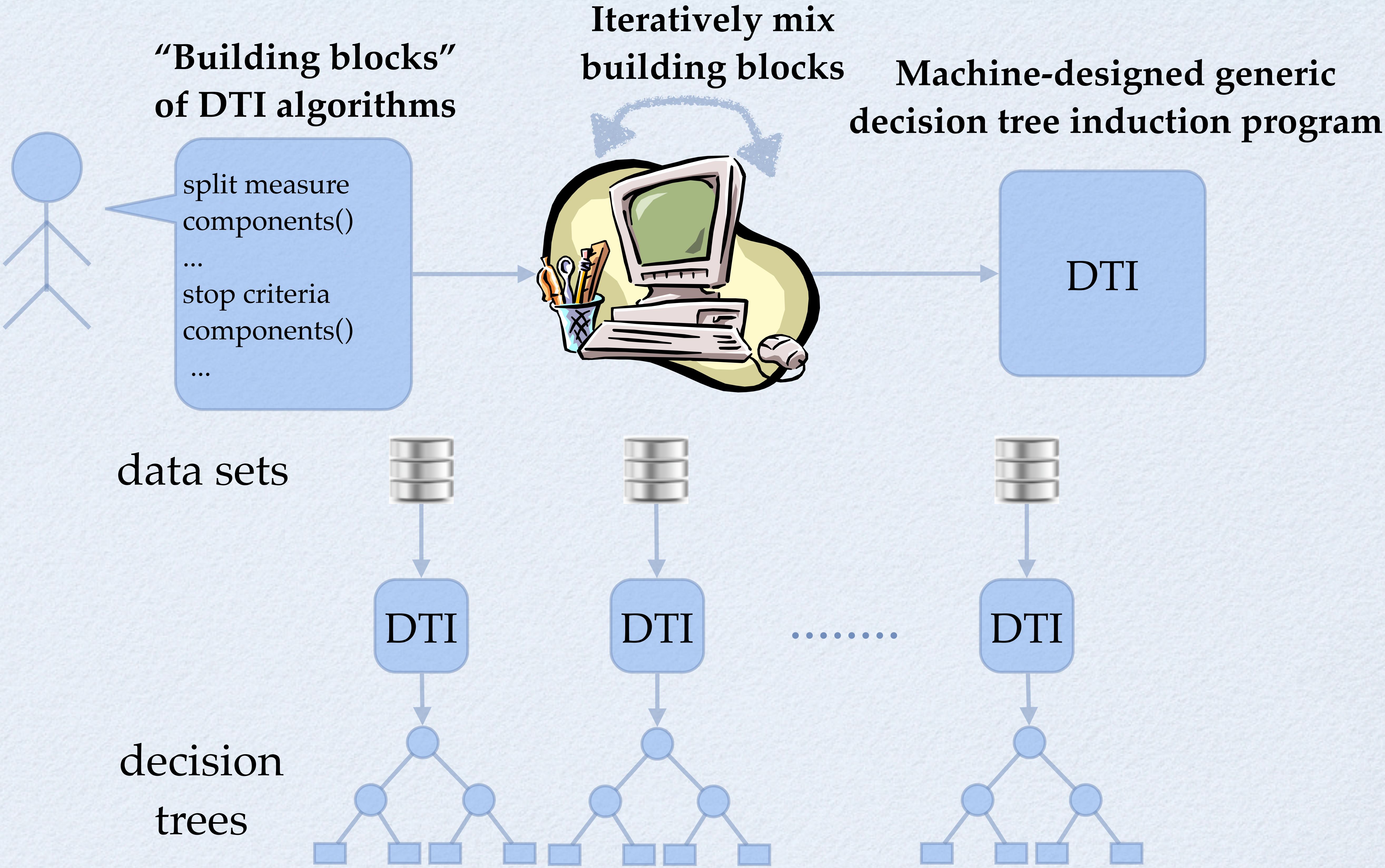
DTI

....

DTI

decision trees

Automatic invention of decision tree induction algorithms



"Building blocks" of DTI algorithms

split measure components()
...
stop criteria components()
...

Iteratively mix building blocks

Machine-designed generic decision tree induction program

DTI

data sets

DTI

DTI

....

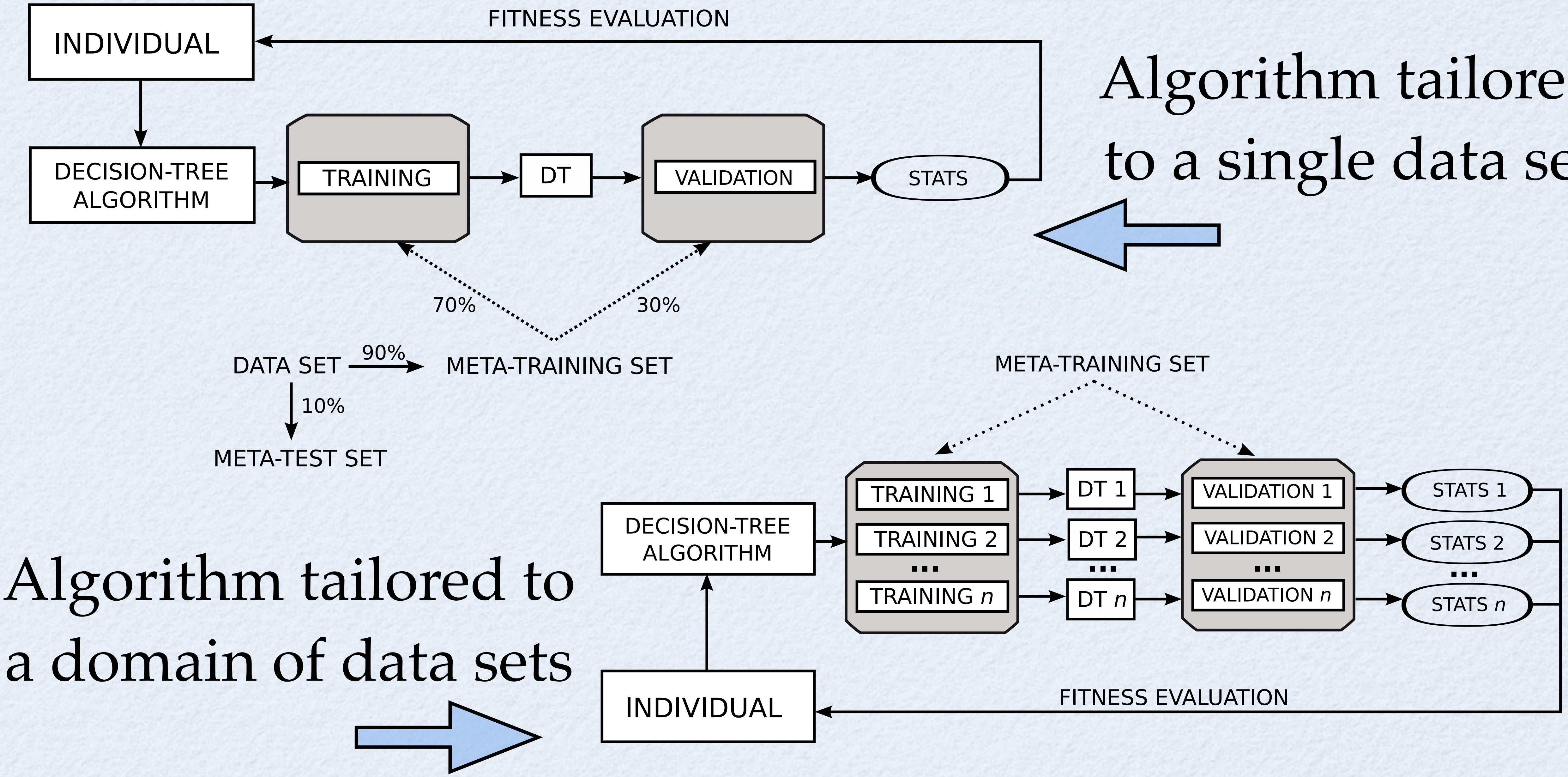
DTI

decision trees

Motivation of automating the design of data mining algorithms

- New level of automation in data mining
  - Relevant research topic for both data mining and AI in general
  - Study of differences between human-designed and machine-designed algorithms
- Avoid algorithm biases introduced by the human algorithm designer
- No classification algorithm is "the best" across all datasets
  - New machine-designed algorithm can be useful for types of data sets where human-designed algorithms have not a good predictive performance
- Focus on decision tree induction algorithms, due to the comprehensibility of the discovered model (IF-THEN rules)

Algorithm Evaluation



Algorithm tailored to a single data set

Algorithm tailored to a domain of data sets

Successful Applications

- Algorithm tailored to a single data set - public UCI data

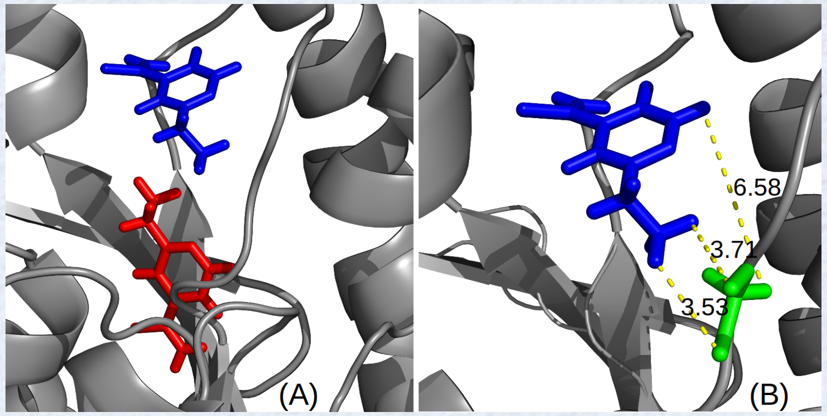
Evolutionary Computation

Summer 2013

Volume 21, Number 2

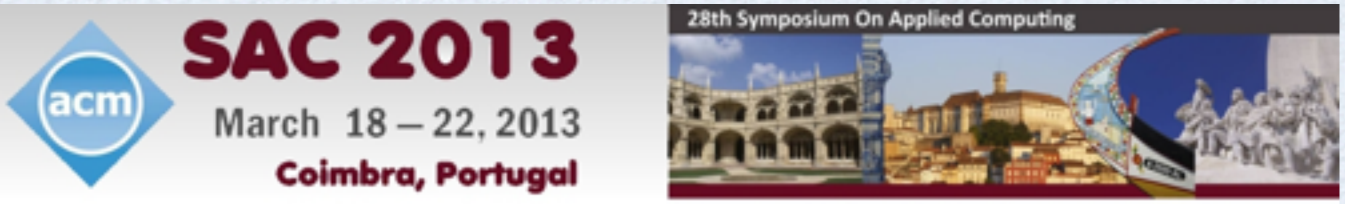
- Algorithm tailored to flexible-receptor molecular docking data

BMC Bioinformatics



Successful Applications

- Algorithm tailored to a software maintenance effort data set from HP



SAC 2013

March 18 – 22, 2013

Coimbra, Portugal

28th Symposium On Applied Computing

- Algorithm tailored to gene expression data sets - currently under review in:

IEEE Computational Intelligence Society

Transactions on Evolutionary Computation

Result is human competitive

(E) Result >= most recent human-created solutions for a long-standing problem

(F) Result >= result considered an achievement in its field at the time it was first discovered

For several data sets, HEAD-DT performed significantly better than state-of-the-art algorithms CART and C4.5, both still largely employed in both academia and industry.

> Even though enhancements have been proposed to both CART and C4.5, no top-down algorithm to date has achieved predictive results comparable to them.

Result is Human Competitive

(G) Result solves a problem of indisputable difficulty in its field

HEAD-DT is the first (and so far the only) algorithm able to automatically design a complete top-down decision tree algorithm tailored to a particular data set or to a particular domain.

It was successfully applied to challenging problems such as:

(i) prediction of flexible-receptor data,

(ii) software maintenance effort prediction, and

(iii) gene expression analysis.

SUMMARY

- HEAD-DT automatically designs complete top-down decision tree induction algorithms
- These algorithms can be tailored to particular data sets or application domains
- Successful “tailor-made” algorithms designed by HEAD-DT provide efficient solutions to several real problems

THANK YOU!

- Questions?