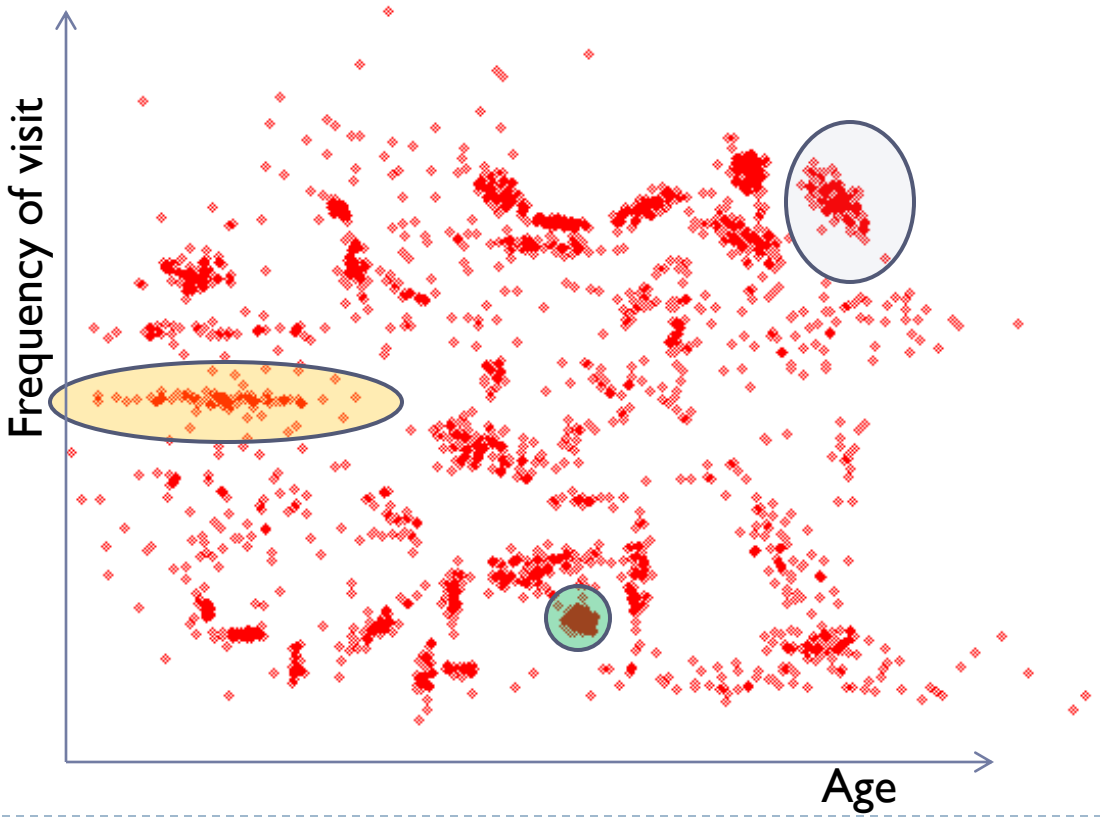
Evidence accumulation in multiobjective data clustering

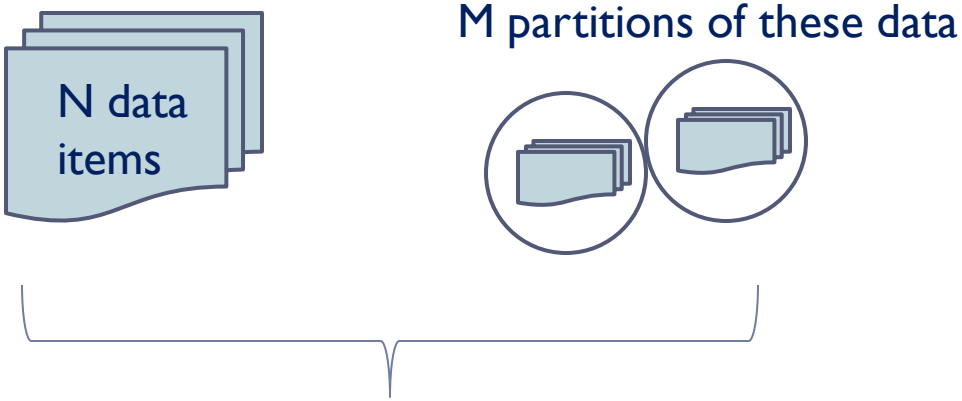
Julia Handl and Joshua Knowles
University of Manchester

Data clustering

Unsupervised classification.
Finding groups of data items that are similar "in some sense".



Evidence accumulation

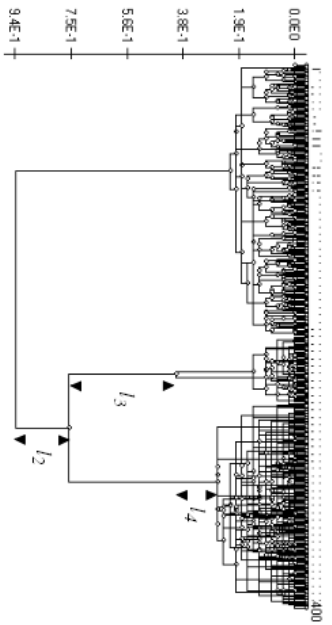


The $N \times N$ co-association matrix is then defined as $C(i, j) = \frac{m_{ij}}{M}$ where m_{ij} indicates the number of times data items i and j have been assigned to the same cluster, e.g.

$$\begin{pmatrix} 1 & \frac{m_{12}}{M} & \frac{m_{13}}{M} & \frac{m_{14}}{M} \\ \frac{m_{21}}{M} & 1 & \frac{m_{23}}{M} & \frac{m_{24}}{M} \\ \frac{m_{31}}{M} & \frac{m_{32}}{M} & 1 & \frac{m_{34}}{M} \\ \frac{m_{41}}{M} & \frac{m_{42}}{M} & \frac{m_{43}}{M} & 1 \end{pmatrix}$$

This can be used to construct a dissimilarity matrix for a further clustering step.

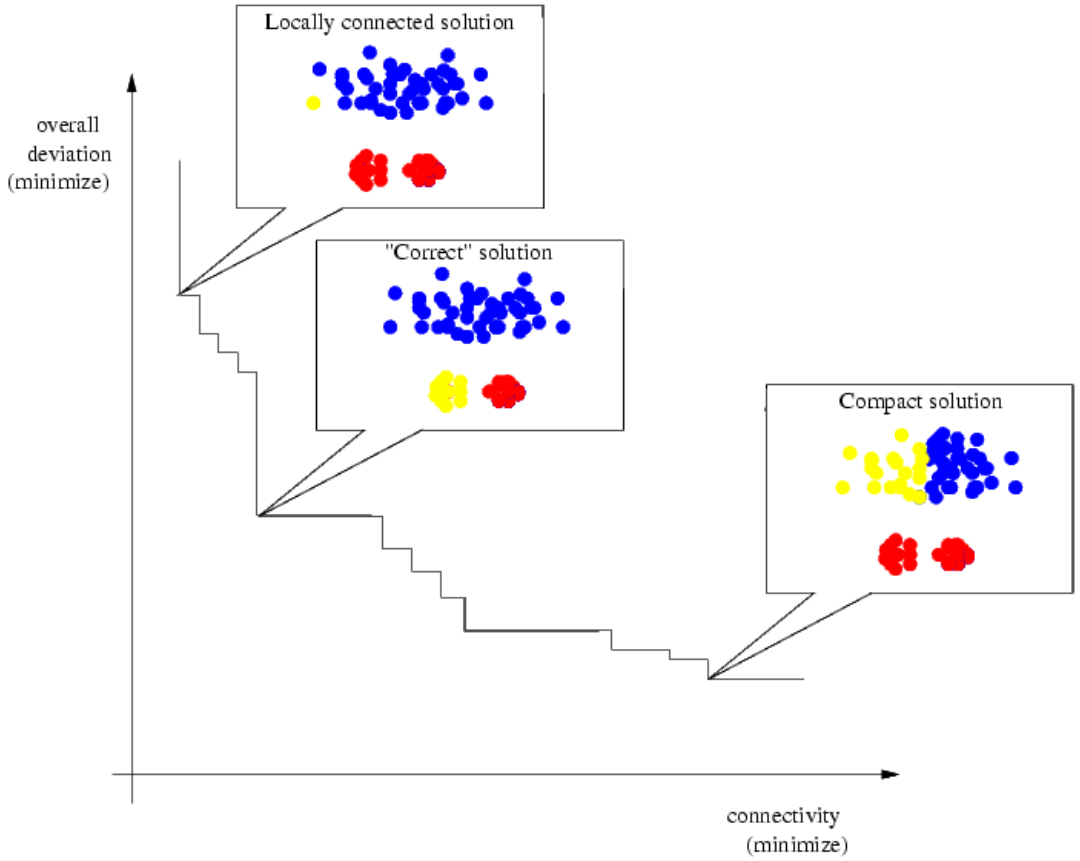
(Fred & Jain, 2005)



Multiobjective clustering

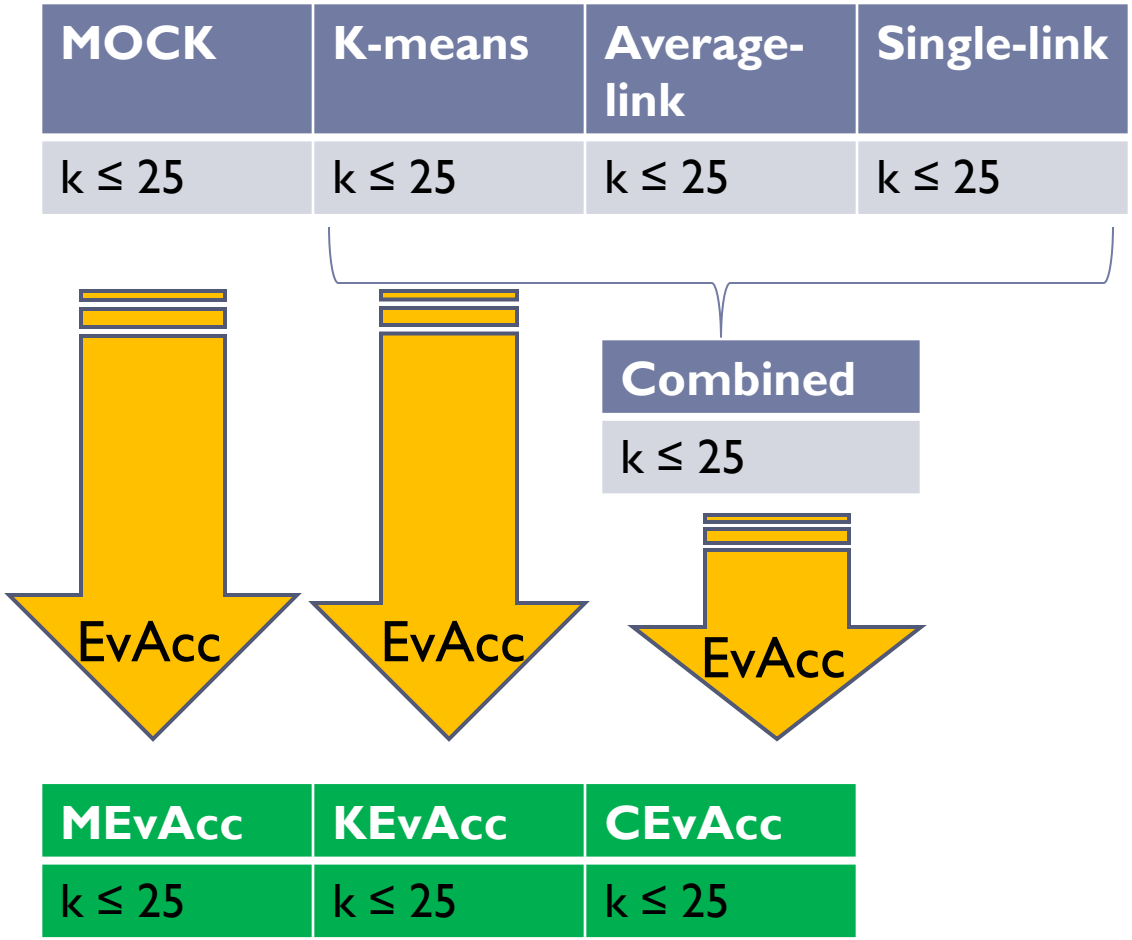
Use of an MOEA to optimize more than one clustering criterion simultaneously (MOCK).

(Handl & Knowles, 2007)

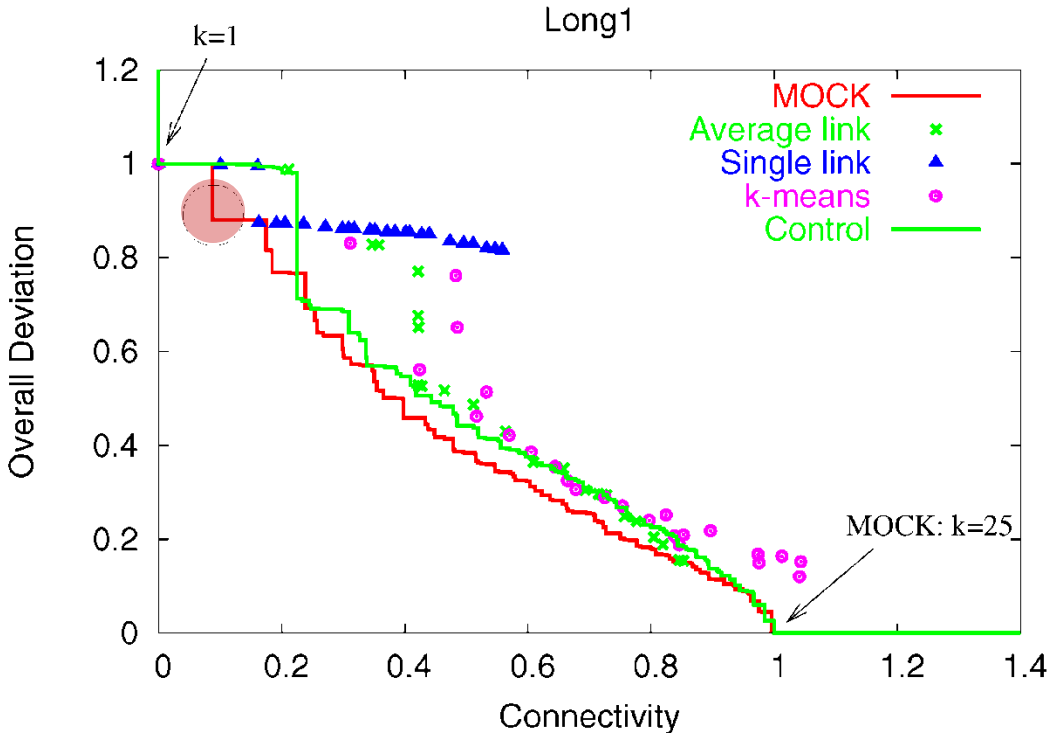


- Previously:
  - MOCK returns a set of solutions (Pareto front approximation)
  - A single preferred solution is selected from the front (model selection)
- Research question: Could we better exploit the information intrinsic to this "ensemble" of solutions?
  - Improved accuracy?
  - Support for model selection?
  - Insight into data?
- Use of Evidence Accumulation to "post-process" MOCK's clustering solutions

Methods to generate input partitions



Methods for model selection



Random control data	Minimum angle	Maximum branch length
MOCK	Any EvAcc	Any EvAccc

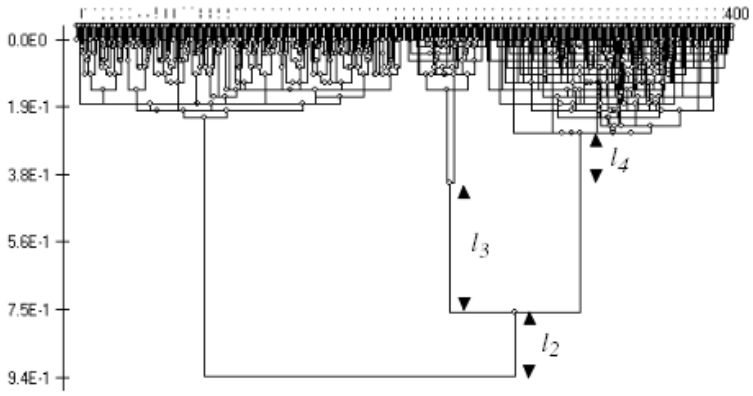
Model selection

Introduction

Research focus

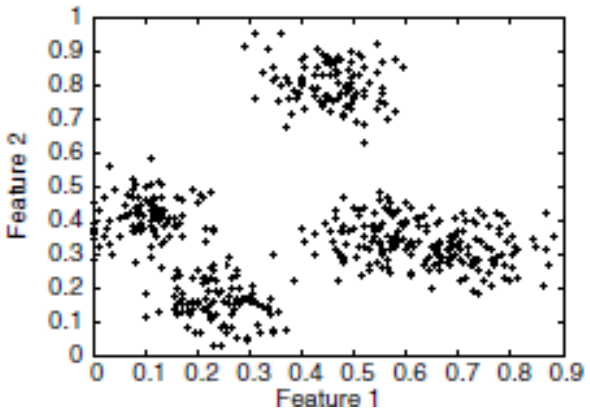
Experiments

Conclusion



(Fred & Jain, 2005)

Test suite of Gaussian clusters

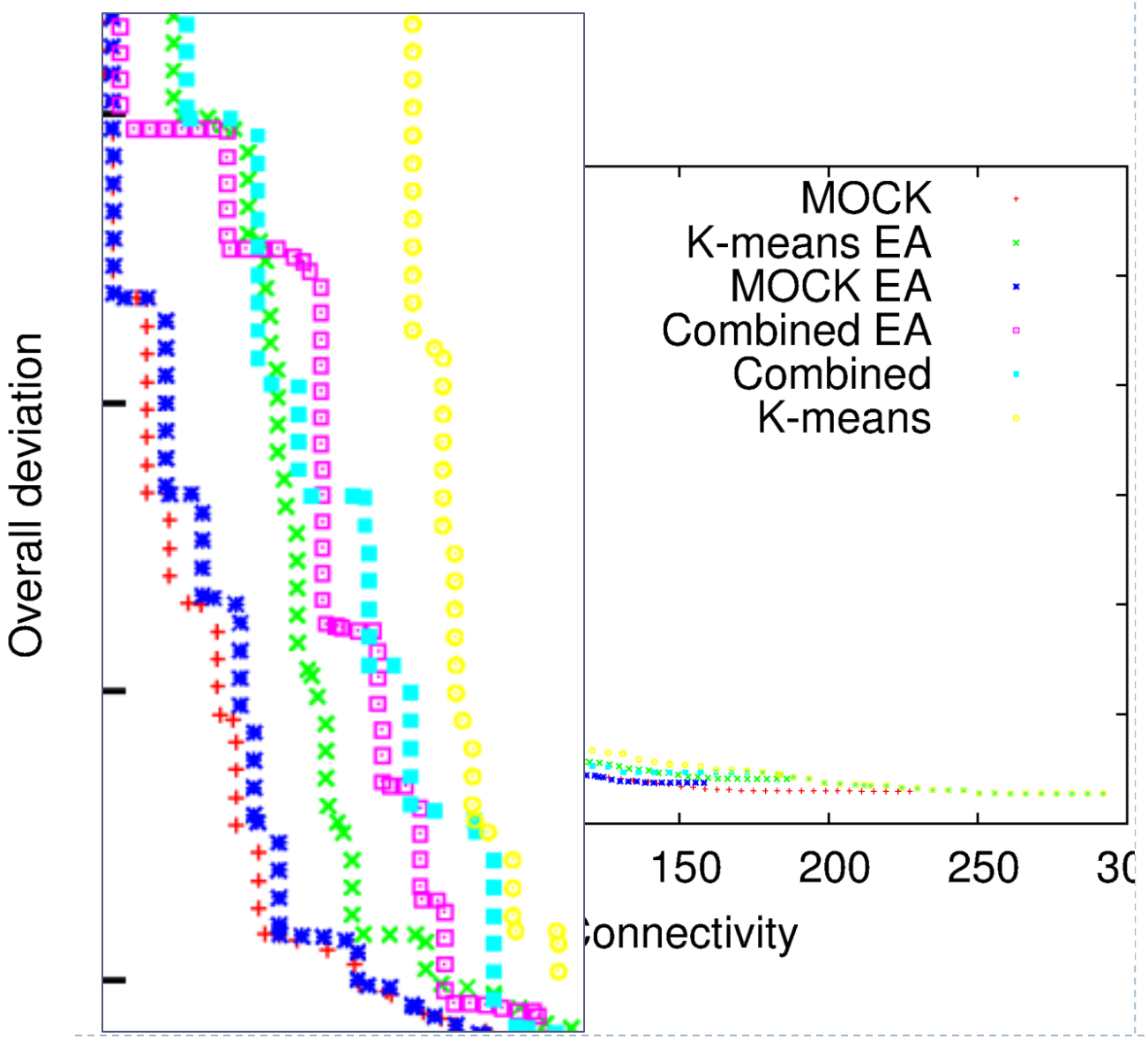


Generator available online; Parameters in paper; 21 runs per instance.

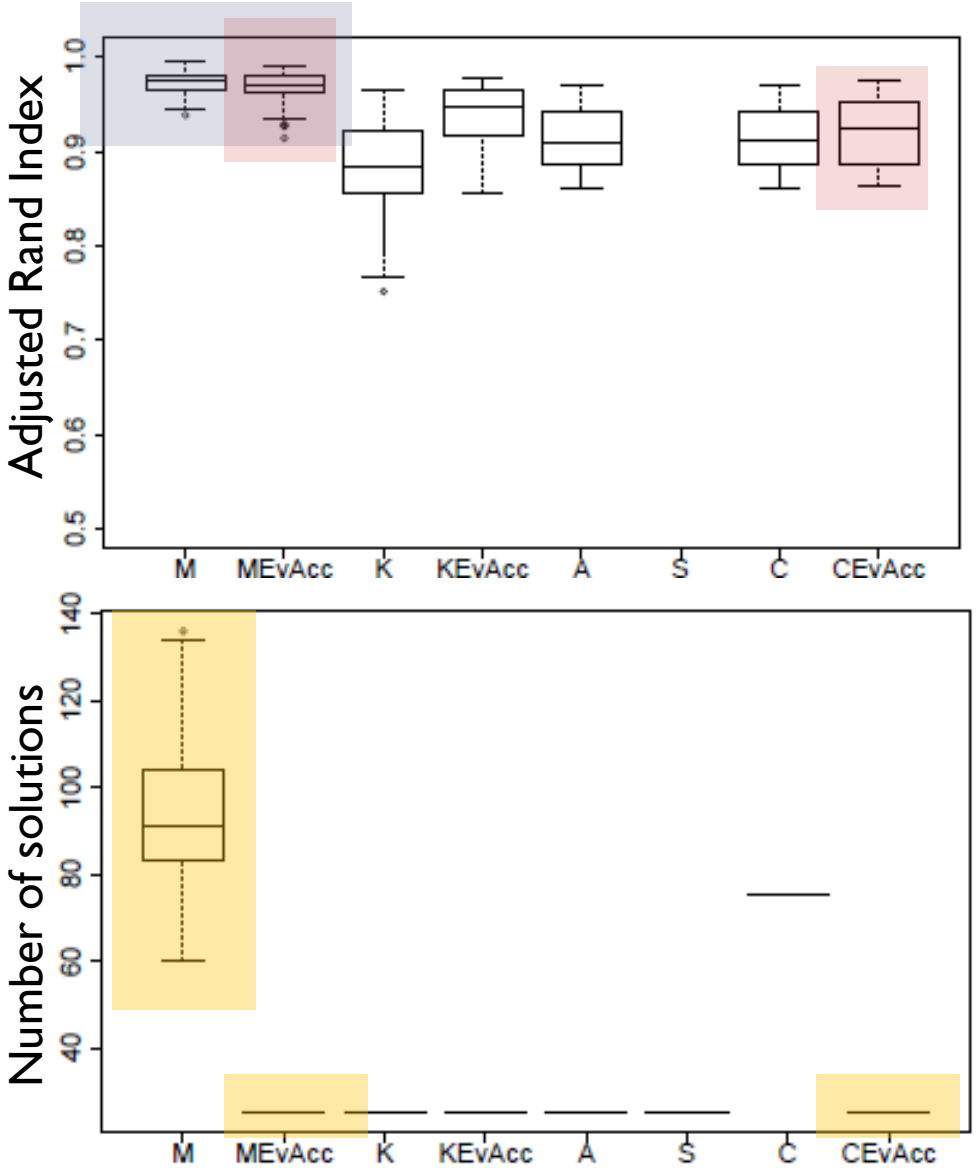
Name	Dimensionality	Number of clusters	Instances
3d-4c	3	4	10
3d-6c	3	6	10
3d-8c	3	8	10
10d-4c	10	4	10
10d-6c	10	6	10
10d-8c	10	8	10

- Performance evaluation
  - Visualization in bi-objective space (attainment fronts)
  - External validation (Adjusted Rand Index)
    - Best ***generated***
    - Best ***selected***
  - Size of solutions sets

Median attainment fronts (3d-8c-no0)



Best generated & Solution Set Size



(10d-8c)

Results

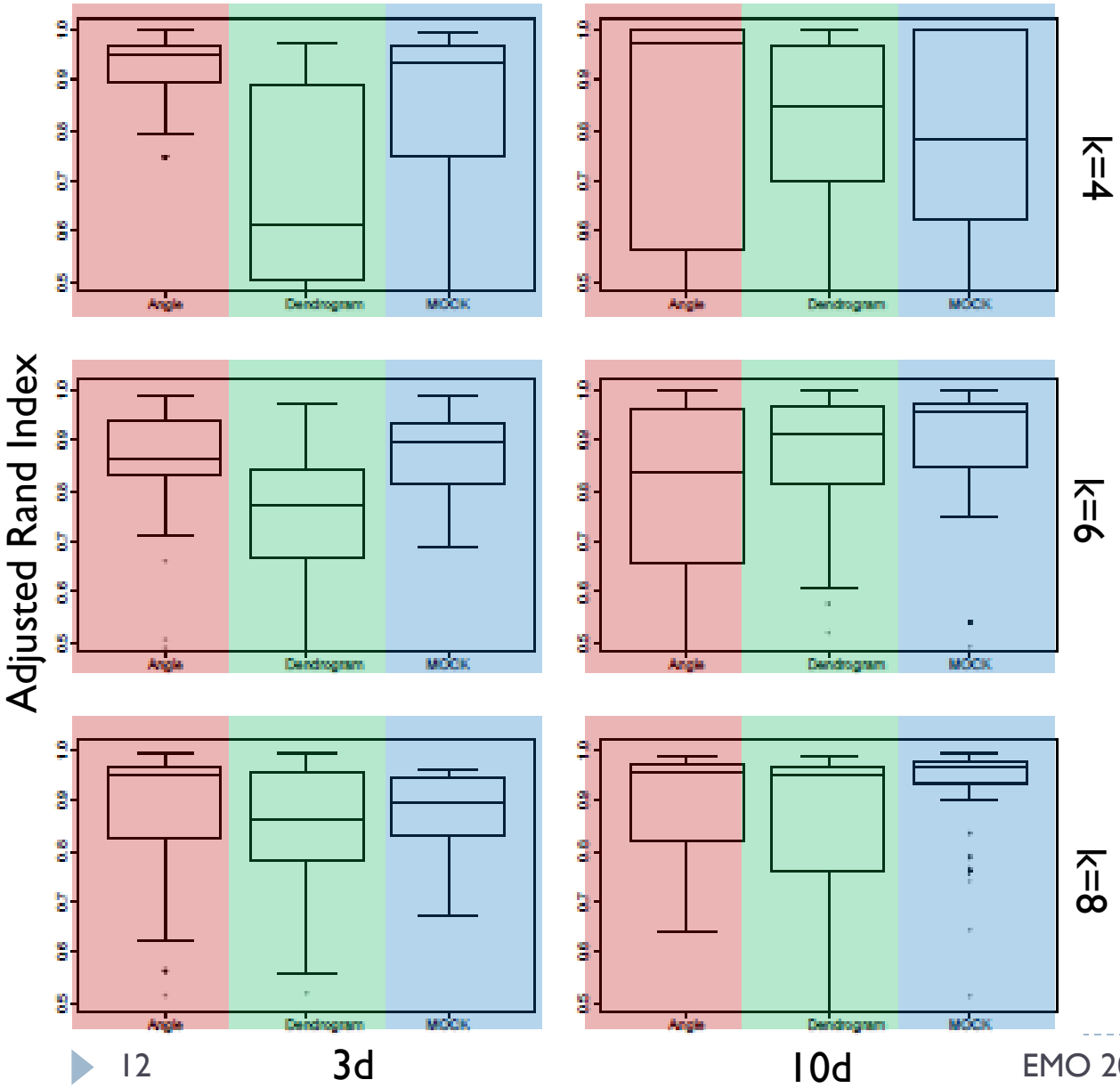
Introduction

Research focus

Experiments

Conclusion

- Best selected



Results

Introduction

Research focus

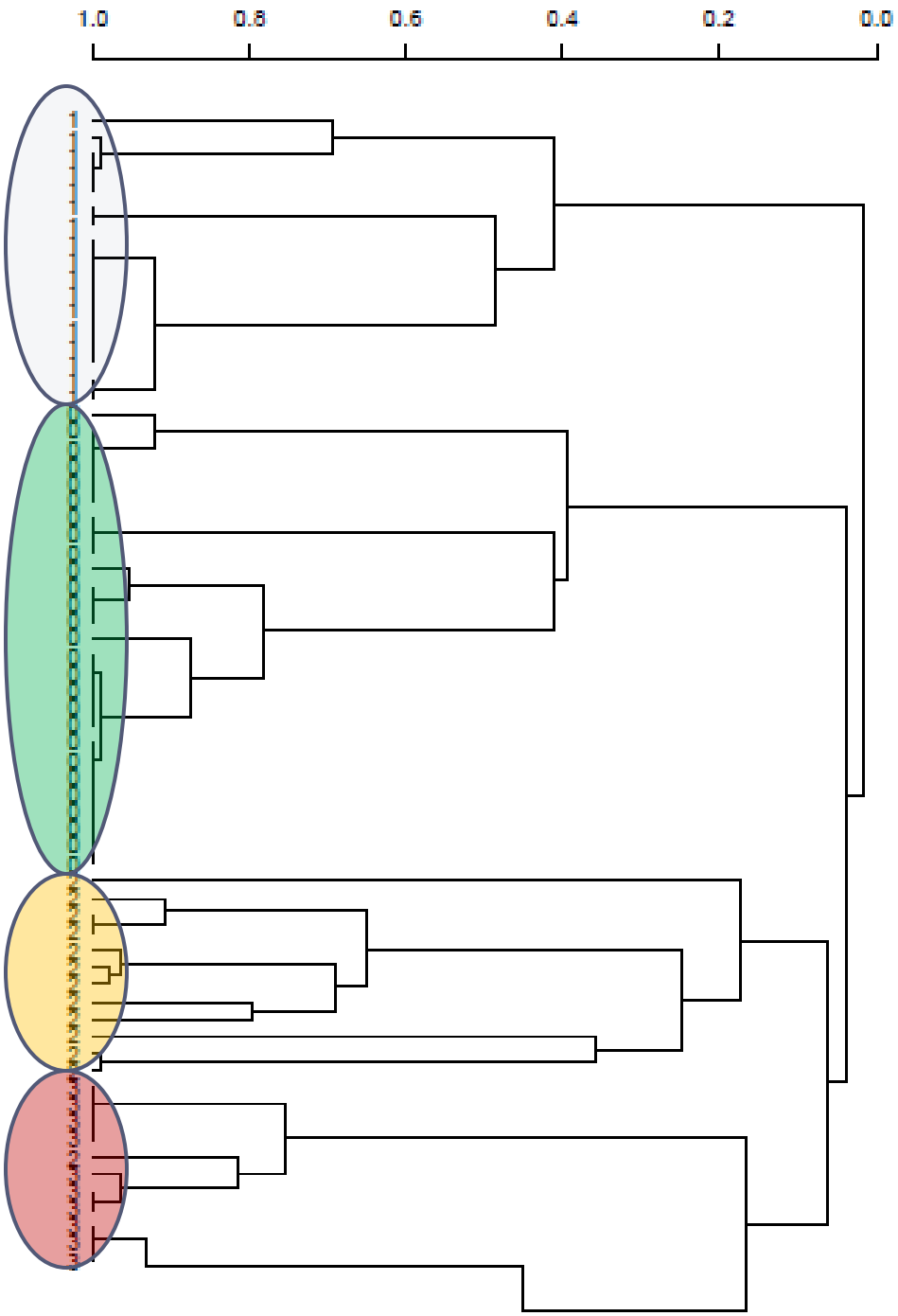
Experiments

Conclusion

Legend:
Minimum angle
Maximum branch length
Random control data

Key findings

- MOCK with Evidence Accumulation:
  - Small decrease in average optimization performance and external validity
  - Significant reduction in size of solution sets
- Different inputs to Evidence Accumulation:
  - Best performance for clustering input from MOCK
  - To some extend, Evidence Accumulation appears to "implicitly optimize" MOCK's objectives
- Solution selection:
  - Inconsistent performance (potential for hybridization?)



Future work

Introduction

Research focus

Experiments

Conclusion