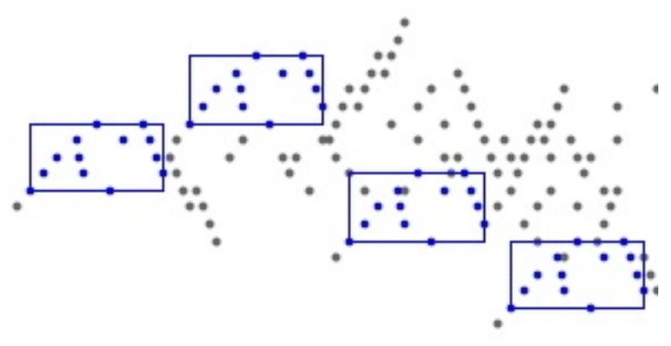
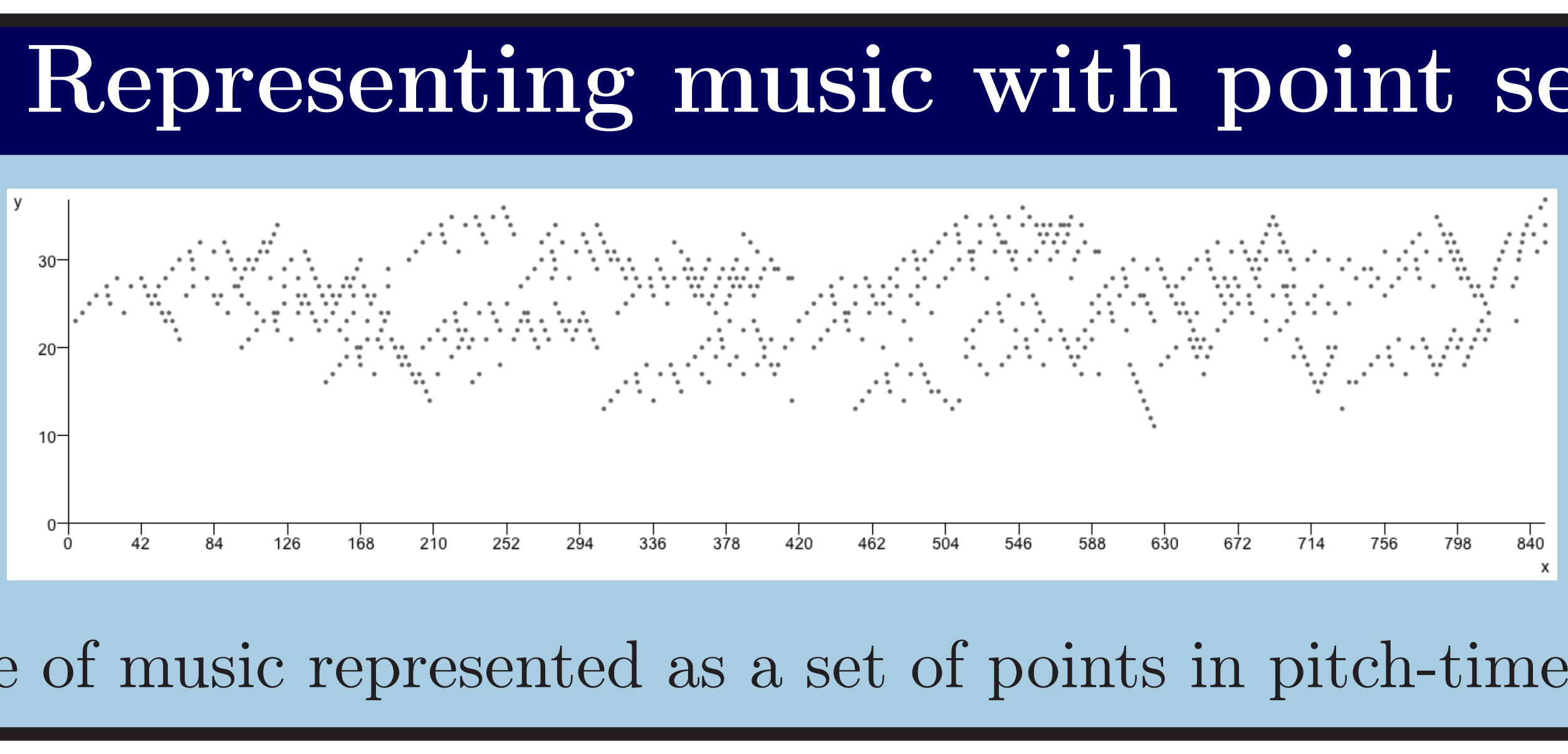


COSIATEC and SIATECCompress: Pattern Discovery by Geometric Compression

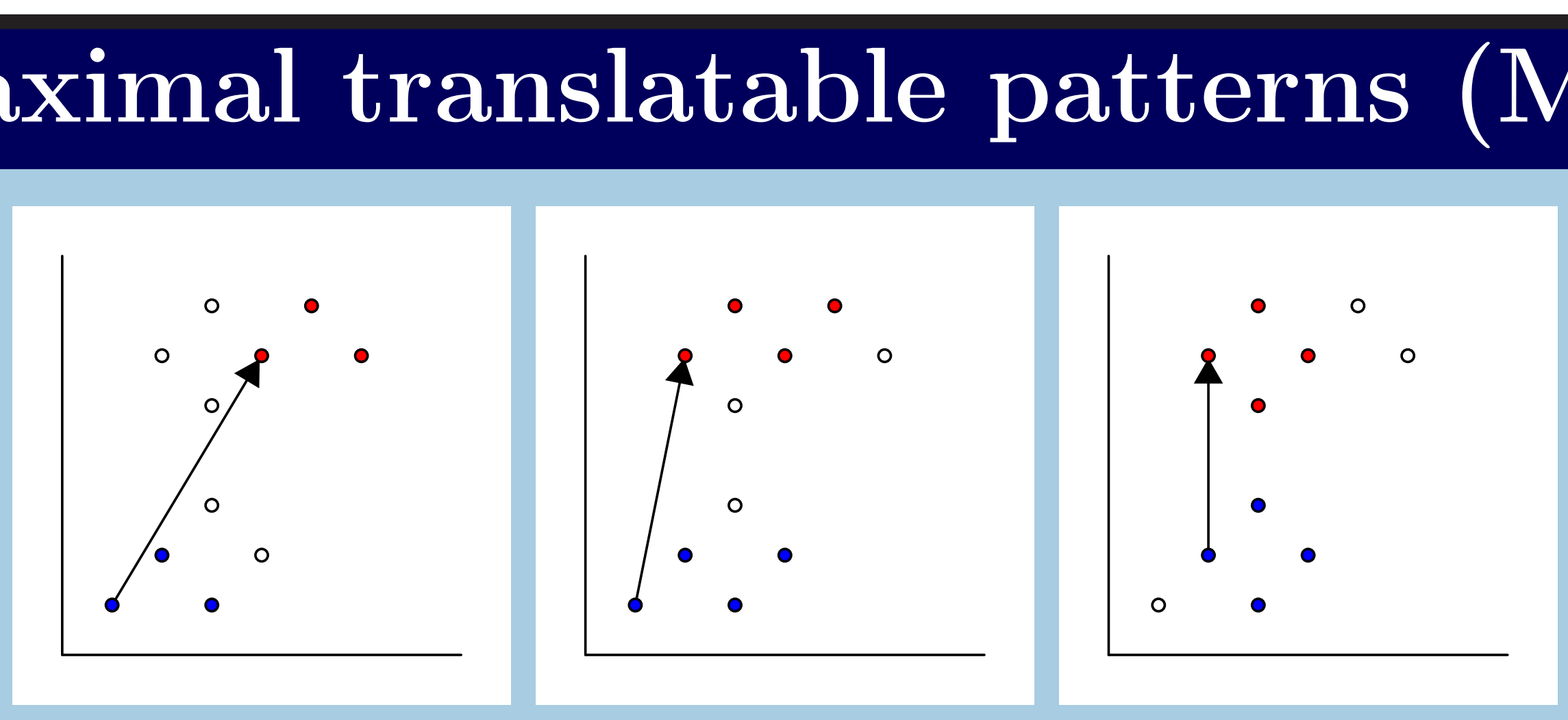
David Meredith
dave@create.aau.dk

1. Representing music with point sets



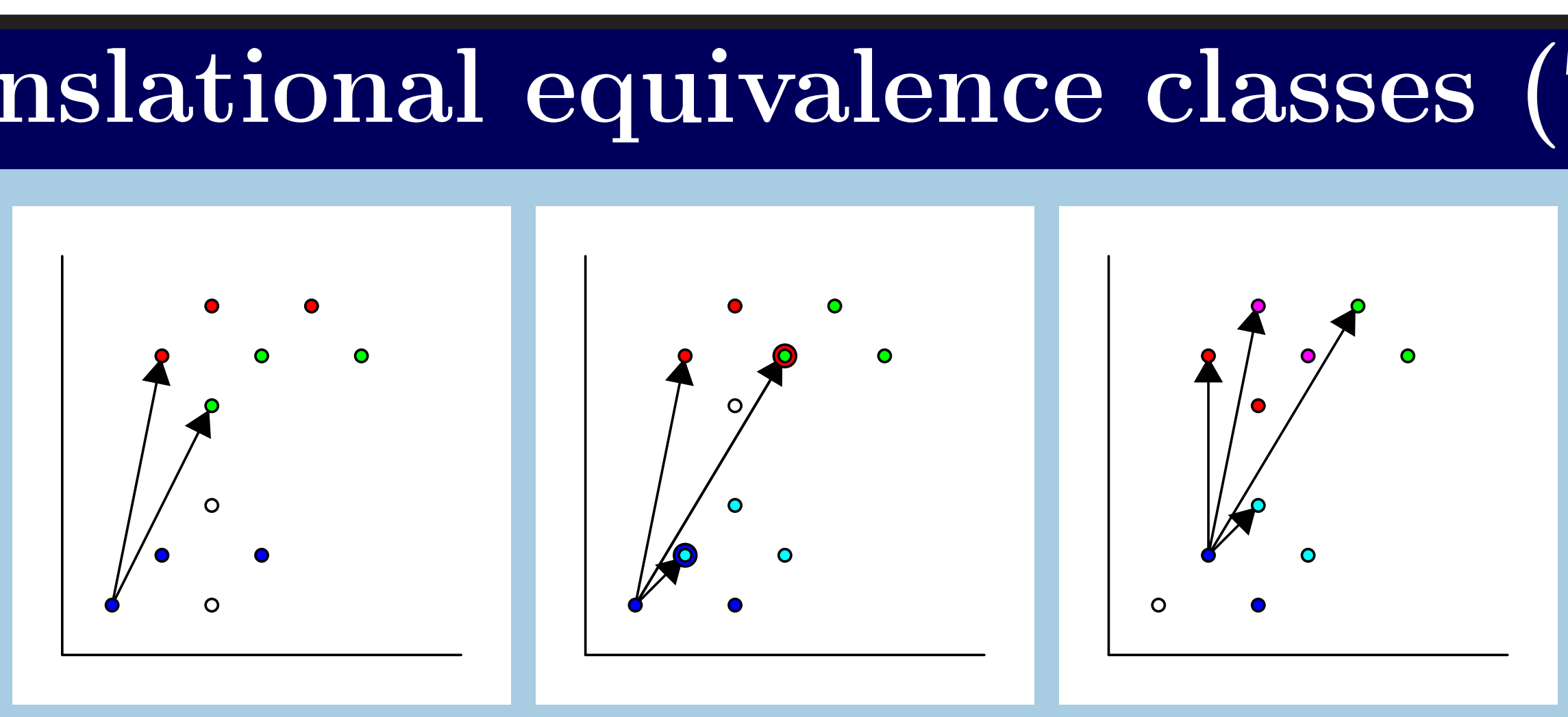
- Piece of music represented as a set of points in pitch-time space

2. Maximal translatable patterns (MTPs)



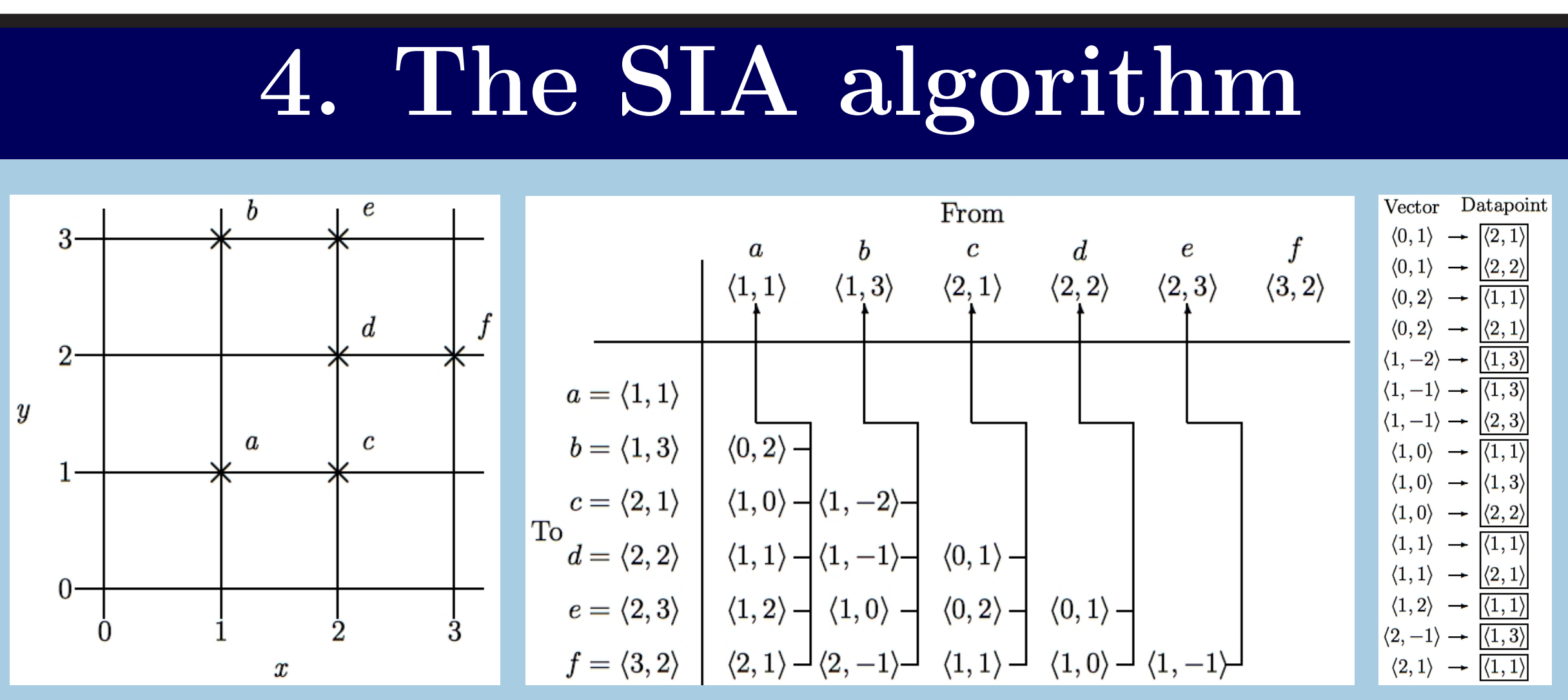
- **Maximal translatable pattern (MTP)** for a given vector in a given dataset contains all points that can be **translated** by that vector to other points in the dataset

3. Translational equivalence classes (TECs)



- **Translational equivalence class (TEC)** of a pattern contains all translationally invariant occurrences of the pattern in a given dataset
- A translational equivalence class can be compactly expressed as a $\langle$pattern, vector set$\rangle$ pair, giving the points in one occurrence of its pattern along with the vectors that map that occurrence onto all the other occurrences
- The **covered set** of a TEC is the union of all the occurrences in the TEC (e.g., the coloured points in each of the diagrams above)

4. The SIA algorithm

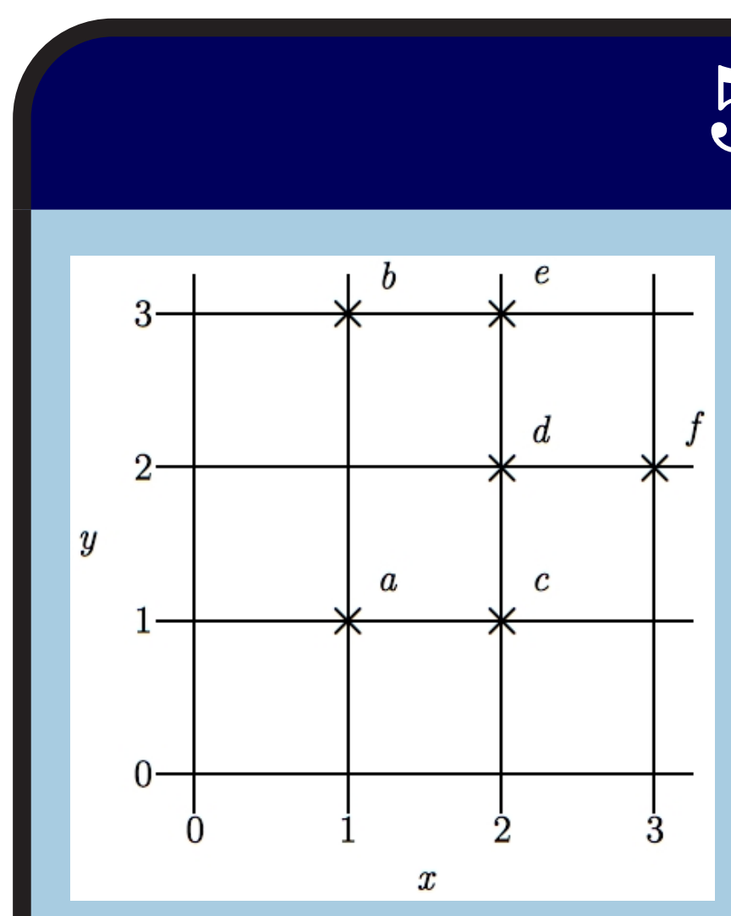


		From					
		a $\langle 1,1\rangle$	b $\langle 1,3\rangle$	c $\langle 2,1\rangle$	d $\langle 2,2\rangle$	e $\langle 2,3\rangle$	f $\langle 3,2\rangle$
	$a = \langle 1,1\rangle$						
	$b = \langle 1,3\rangle$	$\langle 0,2\rangle$					
	$c = \langle 2,1\rangle$	$\langle 1,0\rangle$	$\langle 1,-2\rangle$				
To	$d = \langle 2,2\rangle$	$\langle 1,1\rangle$	$\langle 1,-1\rangle$	$\langle 0,1\rangle$			
	$e = \langle 2,3\rangle$	$\langle 1,2\rangle$	$\langle 1,0\rangle$	$\langle 0,2\rangle$	$\langle 0,1\rangle$		
	$f = \langle 3,2\rangle$	$\langle 2,1\rangle$	$\langle 2,-1\rangle$	$\langle 1,1\rangle$	$\langle 1,0\rangle$	$\langle 1,-1\rangle$	

Vector		Datapoint
$\langle 0,1\rangle$	→	$\langle 2,1\rangle$
$\langle 0,1\rangle$	→	$\langle 2,2\rangle$
$\langle 0,2\rangle$	→	$\langle 1,1\rangle$
$\langle 0,2\rangle$	→	$\langle 2,1\rangle$
$\langle 1,-2\rangle$	→	$\langle 1,3\rangle$
$\langle 1,-1\rangle$	→	$\langle 1,3\rangle$
$\langle 1,-1\rangle$	→	$\langle 2,3\rangle$
$\langle 1,0\rangle$	→	$\langle 1,1\rangle$
$\langle 1,0\rangle$	→	$\langle 1,3\rangle$
$\langle 1,0\rangle$	→	$\langle 2,2\rangle$
$\langle 1,1\rangle$	→	$\langle 1,1\rangle$
$\langle 1,1\rangle$	→	$\langle 2,1\rangle$
$\langle 1,2\rangle$	→	$\langle 1,1\rangle$
$\langle 2,-1\rangle$	→	$\langle 1,3\rangle$
$\langle 2,1\rangle$	→	$\langle 1,1\rangle$

- **SIA (Structure Induction Algorithm)** discovers all the MTPs in a set of n, k-dimensional points in $O(kn^2 \log_2 n)$ time and $O(kn^2)$ space in the worst case
- Finds vector from each point to every lexicographically later point and stores vector in a table with a pointer back to the origin point
- $\langle$vector, point$\rangle$ pairs sorted lexicographically, then segmented at points where vector changes
- Points in each resulting segment form an MTP

5. The SIATEC algorithm



		From					
		$a = \langle 1,1\rangle$	$b = \langle 1,3\rangle$	$c = \langle 2,1\rangle$	$d = \langle 2,2\rangle$	$e = \langle 2,3\rangle$	$f = \langle 3,2\rangle$
	$a = \langle 1,1\rangle$	$\langle 0,0\rangle$	$\langle 0,-2\rangle$	$\langle -1,0\rangle$	$\langle -1,-1\rangle$	$\langle -1,-2\rangle$	$\langle -2,-1\rangle$
	$b = \langle 1,3\rangle$	$\langle 0,2\rangle$	$\langle 0,0\rangle$	$\langle -1,2\rangle$	$\langle -1,1\rangle$	$\langle -1,0\rangle$	$\langle -2,1\rangle$
	$c = \langle 2,1\rangle$	$\langle 1,0\rangle$	$\langle 1,-2\rangle$	$\langle 0,0\rangle$	$\langle 0,-1\rangle$	$\langle 0,-2\rangle$	$\langle -1,-1\rangle$
To	$d = \langle 2,2\rangle$	$\langle 1,1\rangle$	$\langle 1,-1\rangle$	$\langle 0,1\rangle$	$\langle 0,0\rangle$	$\langle 0,-1\rangle$	$\langle -1,0\rangle$
	$e = \langle 2,3\rangle$	$\langle 1,2\rangle$	$\langle 1,0\rangle$	$\langle 0,2\rangle$	$\langle 0,1\rangle$	$\langle 0,0\rangle$	$\langle -1,1\rangle$
	$f = \langle 3,2\rangle$	$\langle 2,1\rangle$	$\langle 2,-1\rangle$	$\langle 1,1\rangle$	$\langle 1,0\rangle$	$\langle 1,-1\rangle$	$\langle 0,0\rangle$

- **SIATEC (SIA+TECs)** discovers all MTP TECs in a k-dimensional dataset of size n in $O(kn^3)$ time and $O(kn^2)$ space in the worst case
- Executes SIA, but fills in complete vector table: set of vectors by which an MTP is translatable is intersection of columns in vector table headed by points in the MTP

6. The COSIATEC algorithm

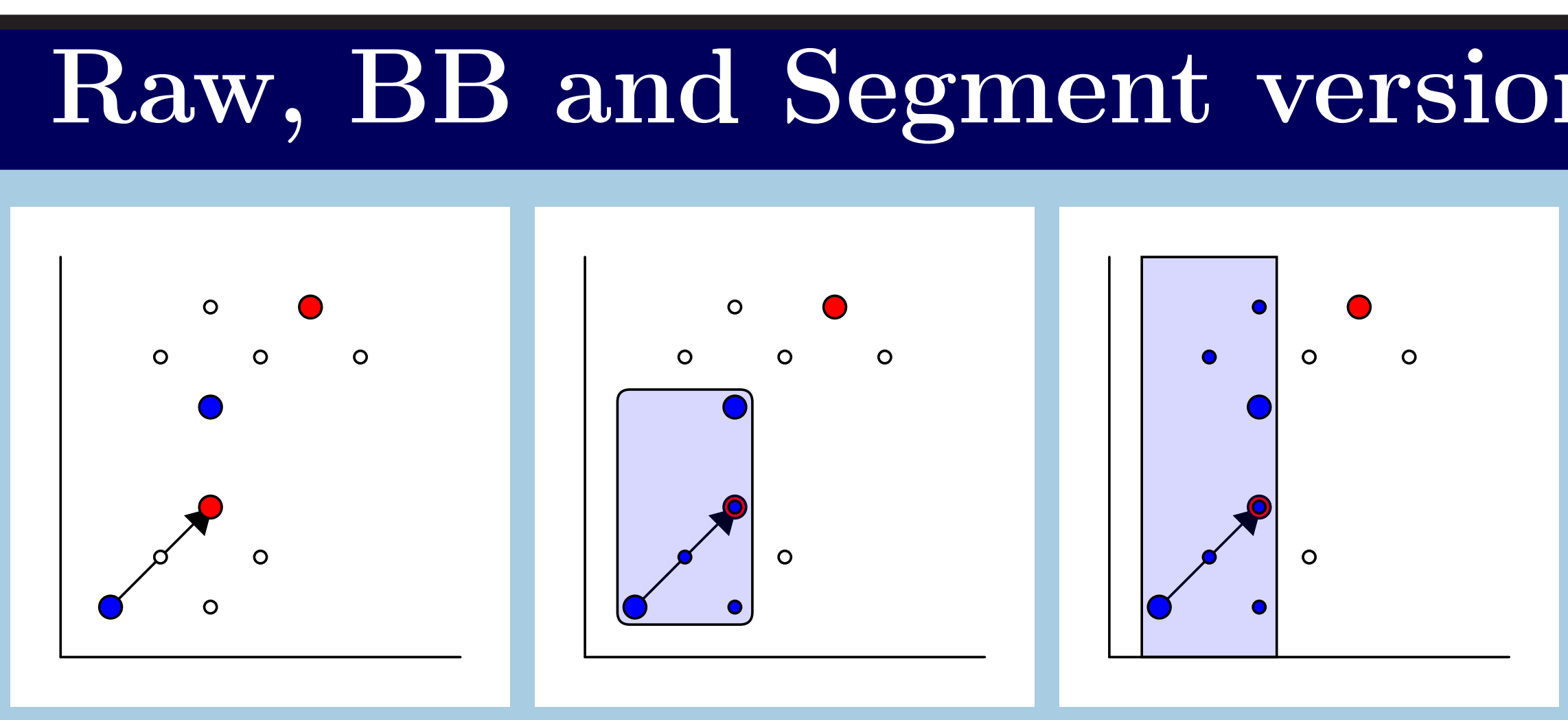
```
COSIATEC($D$)
1   $P \leftarrow$ Copy($D$)
2   $T^* \leftarrow$ nil
3   $\mathbf{T} \leftarrow \langle\rangle$
4   while $P \neq \emptyset$
5       $T^* \leftarrow$ GetBestTEC($P, D$)
6       $\mathbf{T} \leftarrow \mathbf{T} \oplus \langle T^* \rangle$
7       $P \leftarrow P \setminus$ COV($T^*$)
8   return $\mathbf{T}$
```

- **COSIATEC (COmpression with SIATEC)** partitions a dataset into the covered sets of a set of MTP TECs
- SIATEC is run on the dataset and the 'best' TEC is selected and added to the output (lines 5–6)
- The best TEC is the one that gives the best compression. Compactness is used as a tie-breaker.
- The best TEC's covered set is removed from the dataset (line 7)
- The process is repeated until there are no points left (i.e., $P = \emptyset$ in line 4)
- The output is a sequence of TECs, $\mathbf{T}$, such that the covered sets of these TECs form a partition of the input dataset
- By representing each TEC as a $\langle$pattern, vector set$\rangle$ pair, the output encoding is shorter than the explicitly encoded input dataset

7. The SIATECCompress algorithm

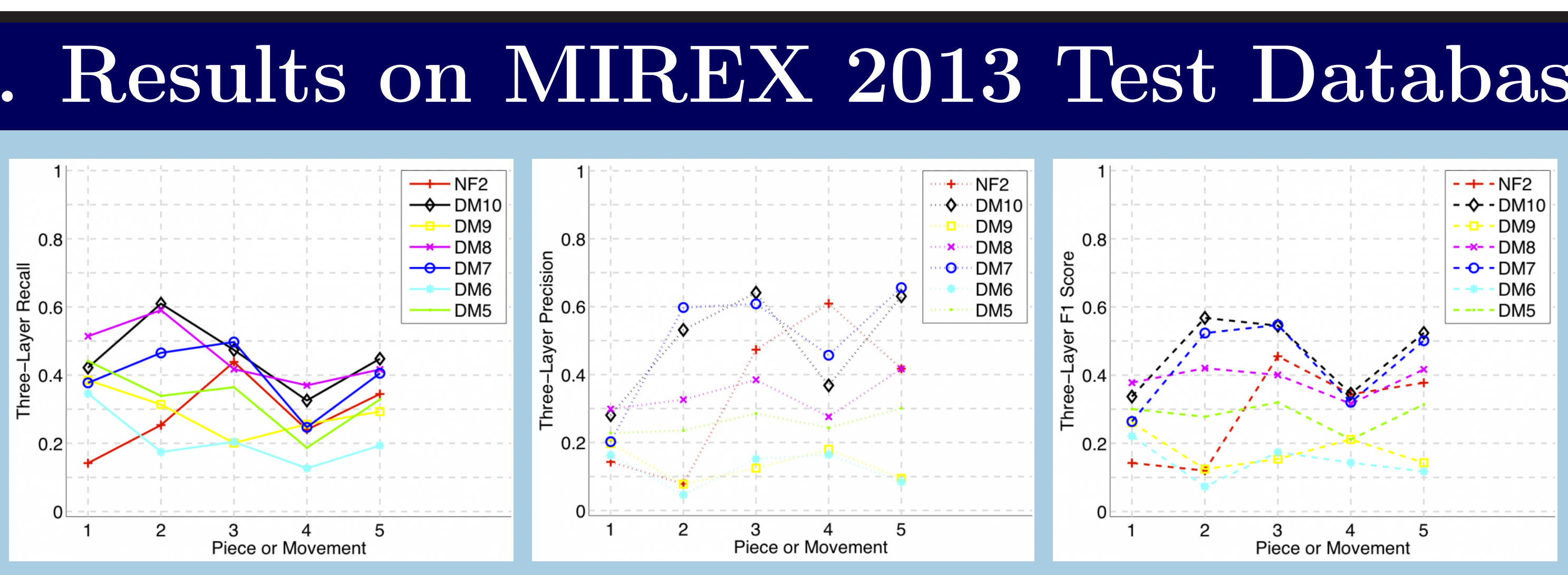
- Computes the MTP TECs using SIATEC
- Sorts TECs into decreasing order of 'quality'
- Quality of a TEC determined by compression ratio, with compactness of pattern used as a tie breaker
- Descends list of sorted TECs, choosing TECs that cover sufficient number of uncovered points
- Encoding consists of a list of TECs, in decreasing order of quality, such that their covered sets cover the input dataset
- Covered sets of TECs in output may intersect (unlike COSIATEC)
- Encoding produced by SIATECCompress typically less compressed than that produced by COSIATEC

8. Raw, BB and Segment versions



- Three versions of each algorithm (COSIATEC and SIATECCompress) submitted to MIREX 2013 competition
- Raw versions: patterns found are the raw MTP occurrences (blue points in left figure)
- BB versions: patterns found are the contents of the bounding boxes of the MTP occurrences (blue points in middle figure)
- Segment versions: patterns found are the contents of the segments spanning the MTP occurrences (blue points in right figure)

9. Results on MIREX 2013 Test Database



- DM10 (SIATECCompressSegment) generally performed best overall with respect to precision, recall and F1 measure.
- NF2 and DM10 not significantly different on establishment recall on a per-pattern basis and NF2 significantly faster than DM10.
- Full results available at `http://www.music-ir.org/mirex/wiki/2013:Discovery_of_Repeated_Themes_&_Sections_Results`