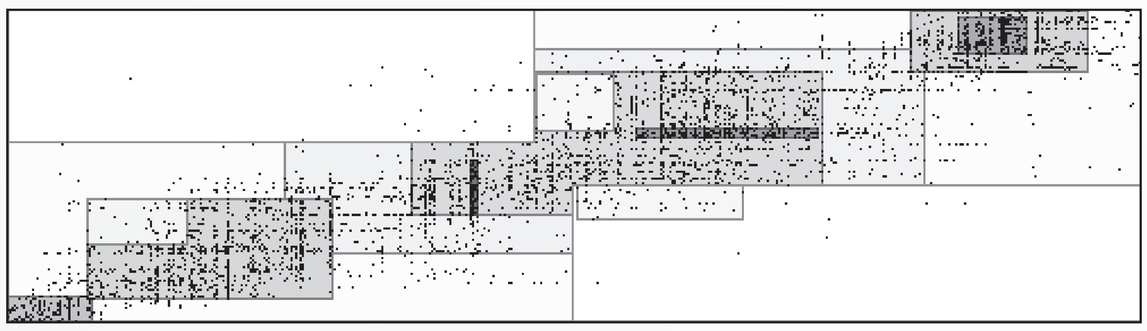
Discovering Descriptive Tile Trees
by Mining Optimal Geometric Subtiles

Nikolaj Tatti and **Jilles Vreeken**
{nikolaj.tatti, jilles.vreeken}@ua.ac.be, ADReM, University of Antwerp, Belgium

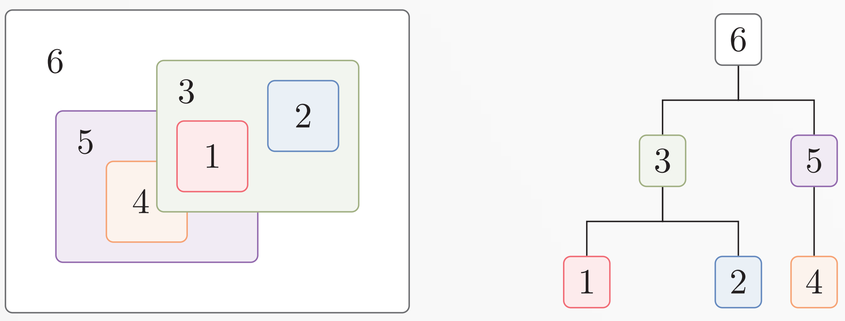
The Goal

Describe ordered binary data with a hierarchy of tiles.
Each tile identifies a significantly dense, or sparse, area of the data. We return a tree of such tiles: each child tile models an exceptional region of its parent.



Tile Trees

a tile ↔ a consecutive submatrix of the data
a child tile ↔ a consecutive submatrix of parent tile



How do we measure quality?

Each tile is modelled by a Bernoulli variable. We allow overlap: the first most specific tile describes a cell.

We score a tile tree by **MDL**; essentially the negative log-likelihood of the data, and a complexity term.

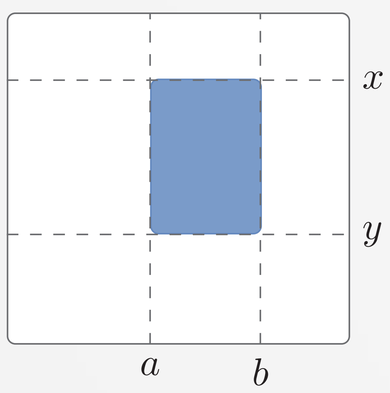
$$L(D|M) = -\sum_T p_T \log \frac{p_T}{p_T + n_T} + n_T \log \frac{n_T}{p_T + n_T}$$

How do we construct a tile tree?

Given a current tile T we find the optimal subtile P of T. Repeat for T, and for P, until MDL tells us to stop.

Finding the Optimal Subtile

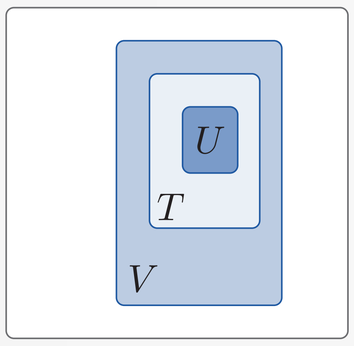
The naïve approach takes $\Theta(N^2M^2)$ time



KEY THEOREM

We can **ignore** a candidate T if

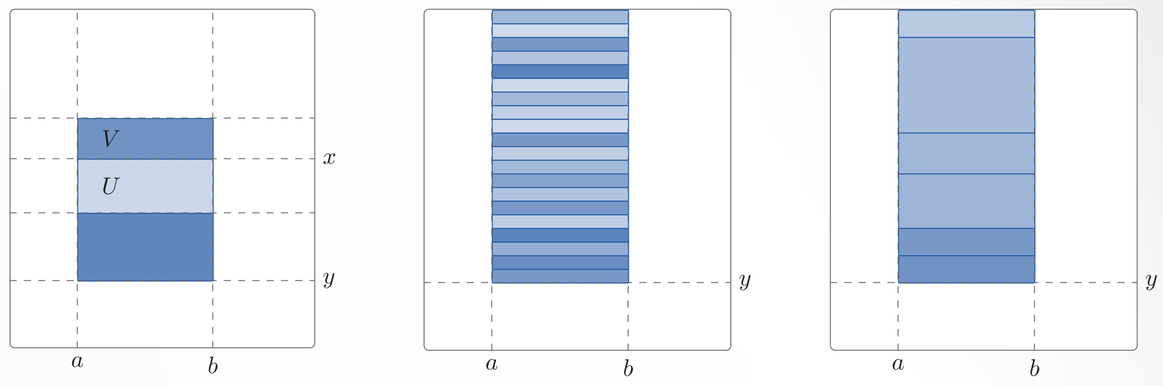
$$dens(V \setminus T) \geq dens(T \setminus U)$$



We show how to do it in only
$\Theta(NM \min(N, M))$
by ignoring suboptimal borders

Borders

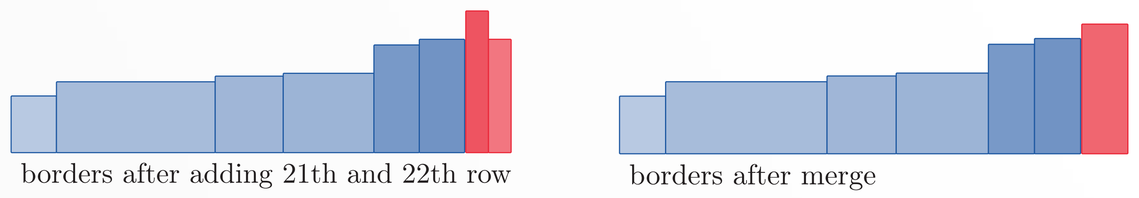
if $dens(V) \geq dens(U)$, x is not a border of y



Example here we go from 20 to only 6 borders

Updating

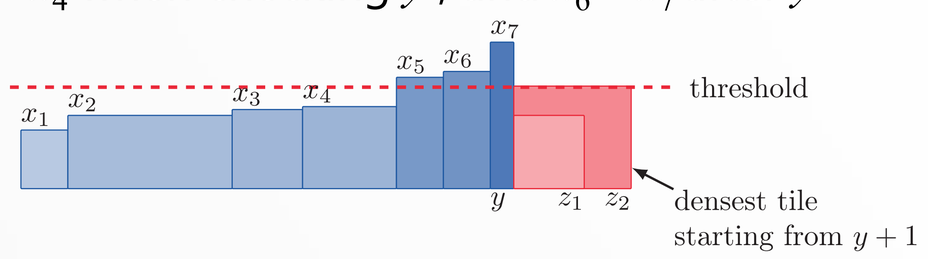
- consider rows one by one, in order of the data
- while $dens(2^{nd}$ to last tile$) \geq dens($last tile$)$
  - these tiles can be safely merged



each merge deletes a border,
hence amortized $\Theta(1)$ time

Pruning

Ignore x_1—x_4 when checking y ; and x_6—x_7 after y

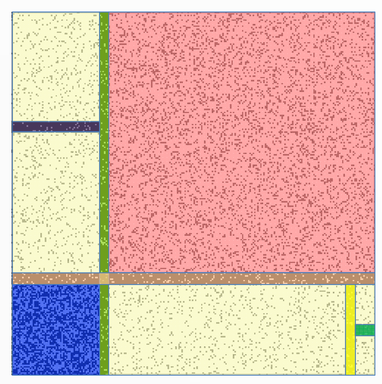


K iterations removes $K - 1$ borders,
hence amortized $\Theta(1)$ time

Experiments

We consider both synthetic and real world data

		W/ OVERLAP		
	$N \times M$	$L\%$	$\lvert T \rvert$	*time*
Composition	240 × 240	81.6	7	1m23s
Abstracts	859 × 541	89.5	14	27m54s
DNA	4590 × 391	61.6	446	625m
Mammals	2183 × 121	54.6	50	3m06s
Paleo	501 × 139	79.1	13	1m22s



Conclusions

If your data is, or can be, meaningfully ordered, take this into account when analysing the data.

- **we model** ordered binary data hierarchically
- we show how to find optimal **subtiles efficiently**
- **future work** includes
  - **richer** data/pattern types

FWO VLAANDEREN
Universiteit Antwerpen