Getting to Know the Unknown Unknowns

Sanjar Karaev, Pauli Miettinen, Jilles Vreeken

Vancouver, May 1, 2015

Do We Know What We Don't Know?

- There are known knowns
  - there are things we know we know
- There are known unknowns
  - we know there are some things we do not know
- But there are also unknown unknowns
  - the ones we don't know we don't know

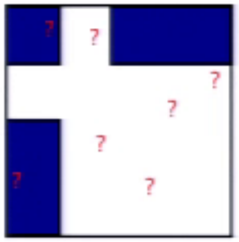
Do We Know What We Don't Know?

- There are known knowns
  - there are things we know we know
- There are known unknowns
  - we know there are some things we do not know
- But there are also unknown unknowns
  - the ones we don't know we don't know

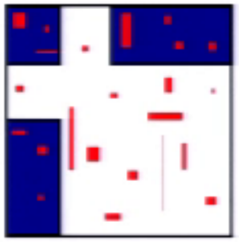
Where Is Data Mining?

- Known knowns
  - noise-free data

- Known unknowns
  - missing values



- Unknown unknowns
  - values (possibly) flipped due to noise



Unknown Unknowns: a Closer Look

- False positives $(0 \to 1)$ and false negatives $(1 \to 0)$ are often not equally likely
- E.g. some 0s might be due to a lack of observation

- Example data
  - columns are locations
  - rows are animal species
  - 1 if present, 0 if not

Unknown Unknowns: a Closer Look

- False positives $(0 \to 1)$ and false negatives $(1 \to 0)$ are often not equally likely
- E.g. some 0s might be due to a lack of observation

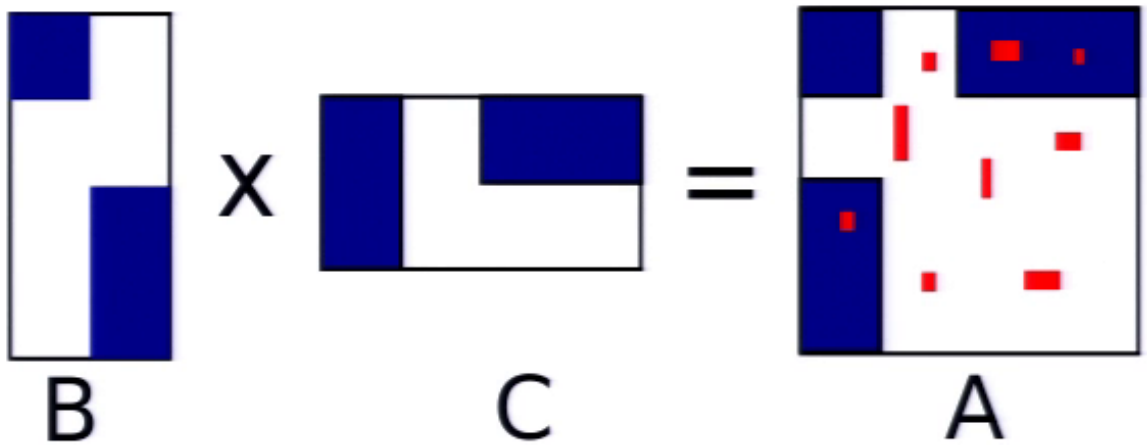
- Example data
  - columns are locations
  - rows are animal species
  - 1 if present, 0 if not

In This Talk

- Represent the data as a union of noisy patterns
- `Nassau`: a new algorithm for BMF
  - minimizes the description length
  - uses MDL to find the rank (number of patterns)
  - dynamically corrects its previous mistakes when new information is found

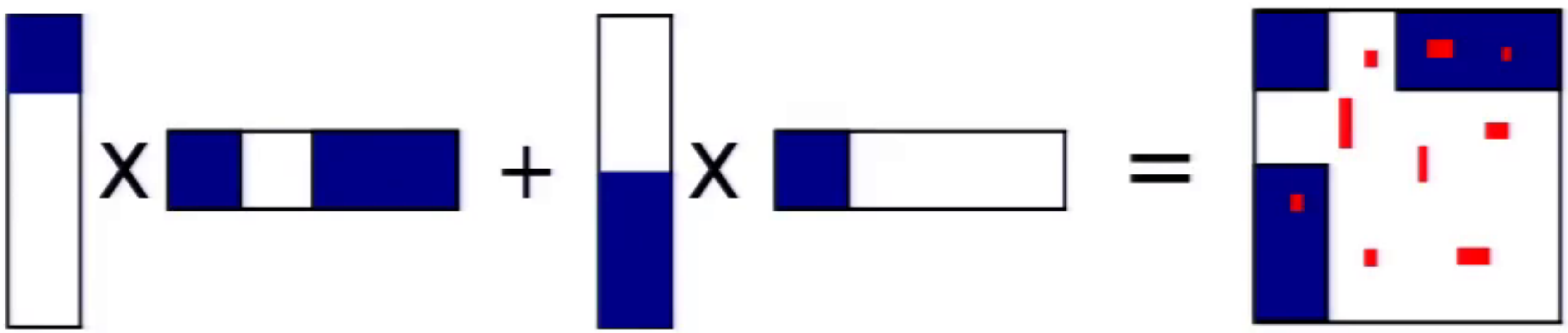
Summarizing Noisy Patterns Using BMF

- Binary data with noise
- Decompose A into a Boolean product of low rank factors $+$ noise



Summarizing Noisy Patterns Using BMF

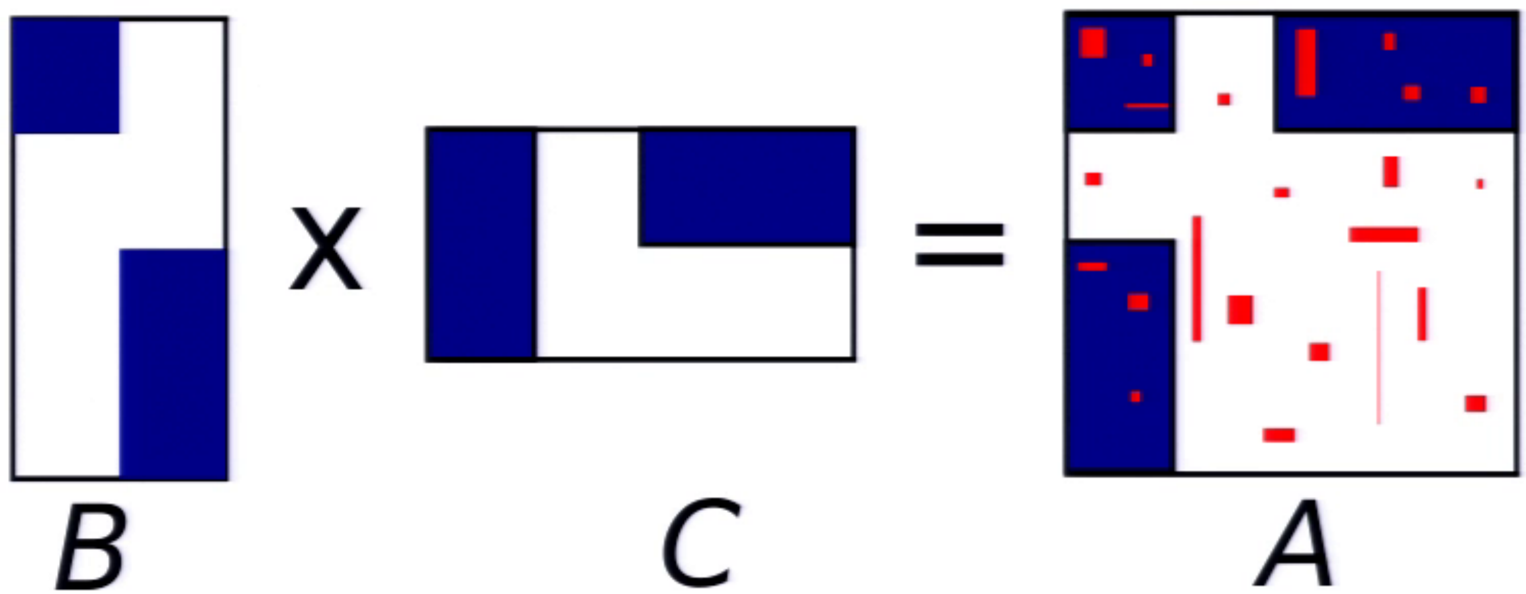
- Binary data with noise
- Decompose A into a Boolean product of low rank factors $+$ noise



- A can now be seen as a sum of rank-1 matrices (blocks)

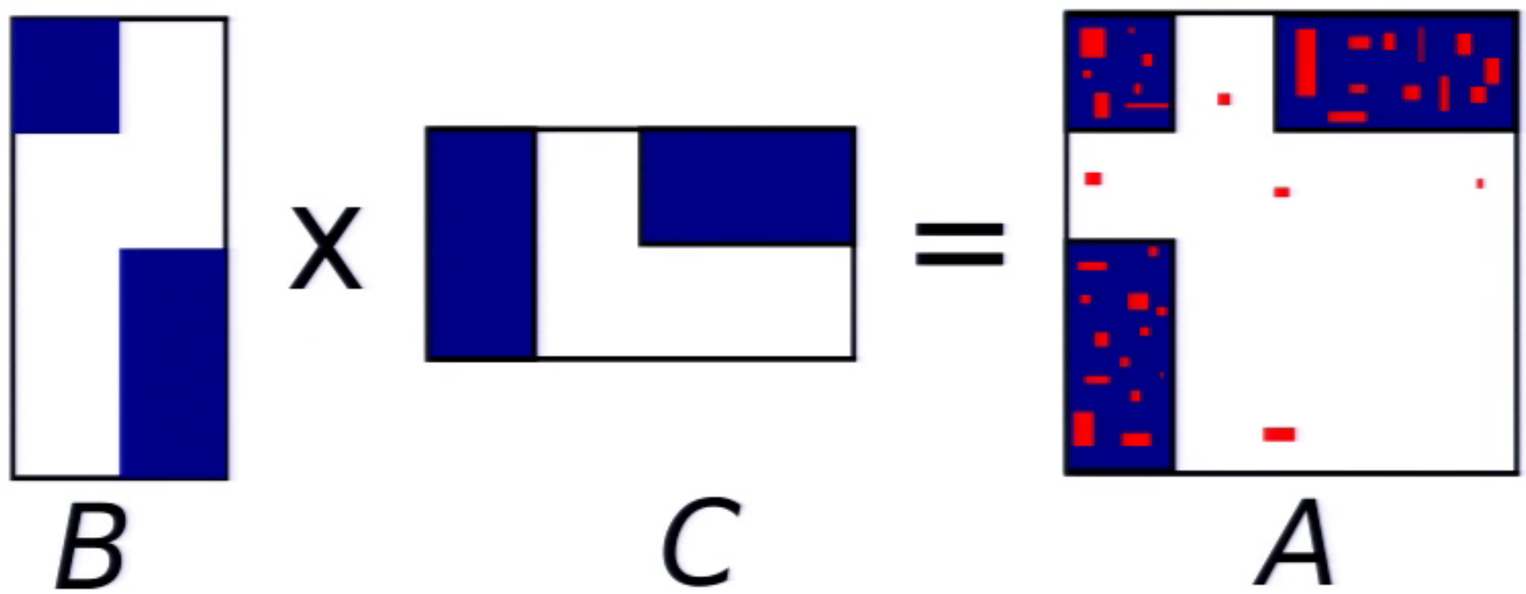
Different Kinds of Noise

- In ideal world
  - use 0/1 error as the cost
  - treat additive $(0 \to 1)$ and destructive $(1 \to 0)$ noise equally



Different Kinds of Noise

- In ideal world
  - use 0/1 error as the cost
  - treat additive $(0 \to 1)$ and destructive $(1 \to 0)$ noise equally
- In real world
  - additive and destructive noise are likely to be imbalanced
  - need to find the right ratio



BMF with MDL

- Objectives
  - isolate flipped elements (unknown unknowns)
  - find optimal rank for the data
- Minimum description length (MDL)
  - patterns in the data can be used to compress it
  - $\Rightarrow$ the more we compressed the data, the more we learned about it
- Encoding BMF
  - $A = B \circ C + E$, where E is the error matrix
  - total description length $L(A, B, C) = L(B) + L(C) + L(E)$

Algorithm

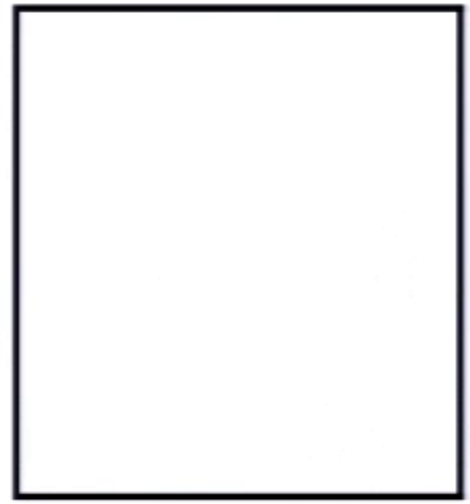
- `Nassau`: a new BMF algorithm
- directly optimizes the description length
  - helps to deal with the imbalance between different types of noise
- nonhierarchical
  - rank-k decomposition doesn't have to be a part of rank-$(k+1)$ decomposition
  - helps to fix earlier mistakes

Algorithm: Factor Updates Example

- Data

- Factors

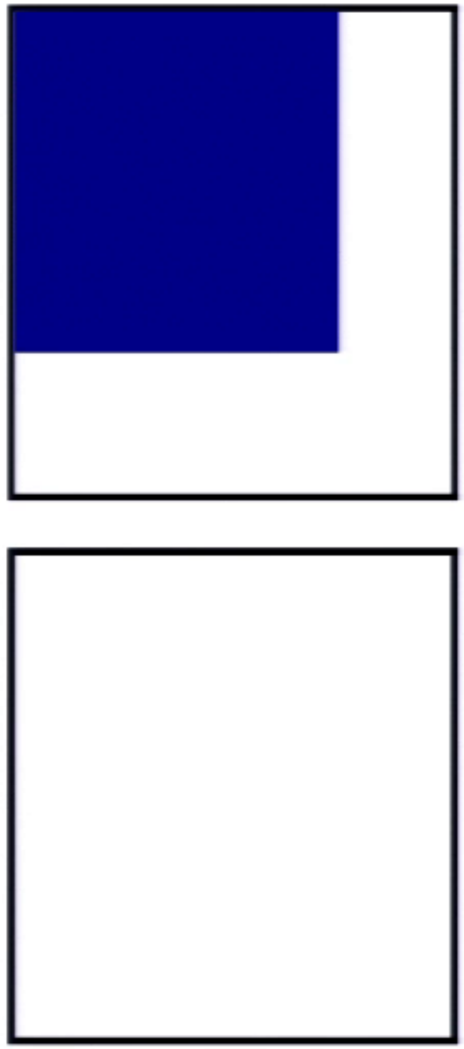
- Result



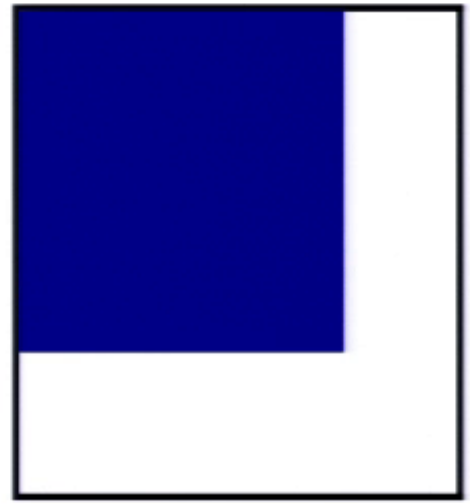
Algorithm: Factor Updates Example

- Data

- Factors



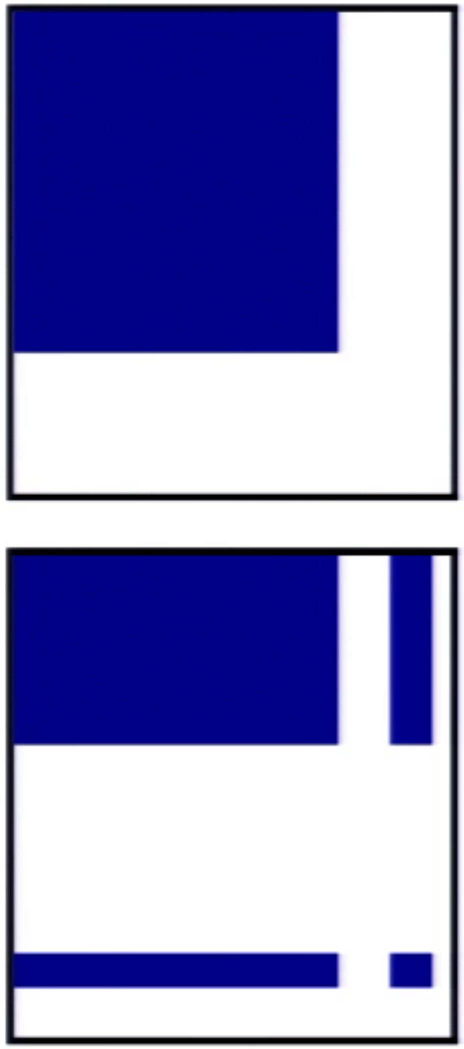
- Result



Algorithm: Factor Updates Example

- Data

- Factors

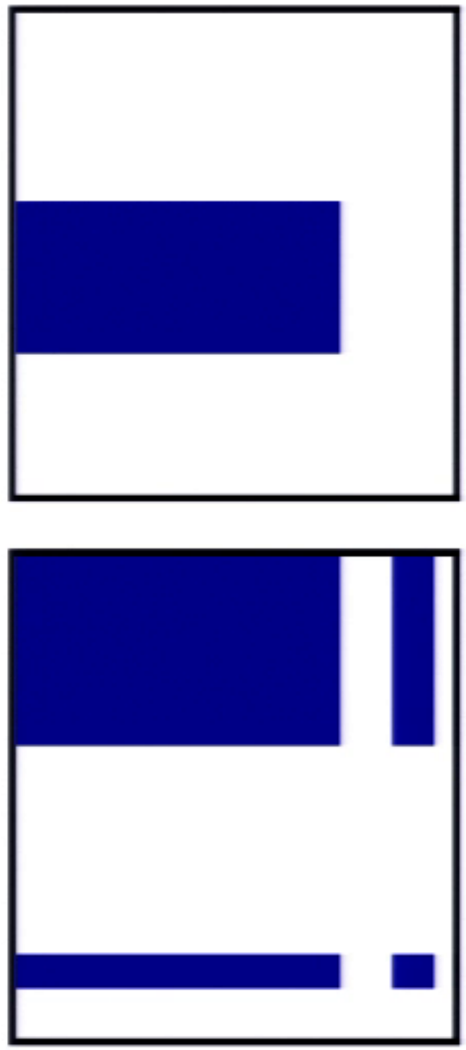


- Result

Algorithm: Factor Updates Example

- Data

- Factors

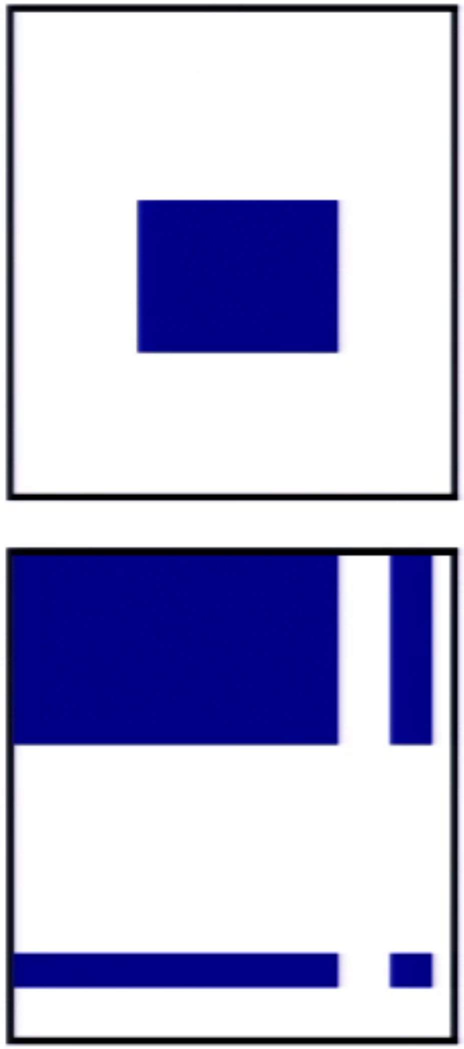


- Result

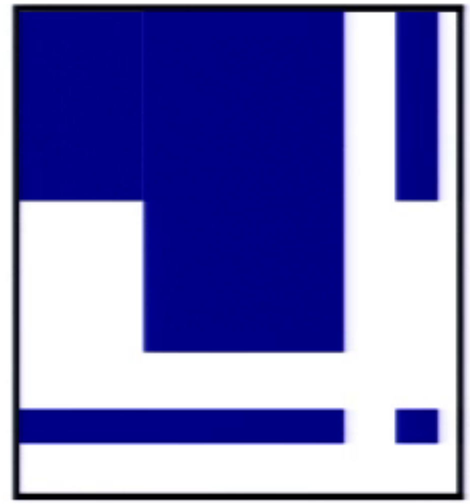
Algorithm: Factor Updates Example

- Data

- Factors

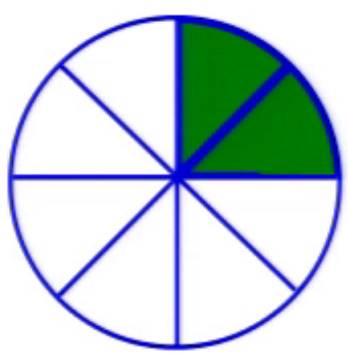


- Result



Algorithm: Summary

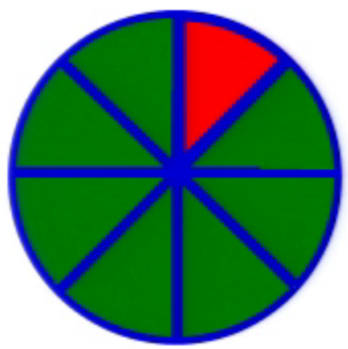
- Nassau
  - optimize the description length
  - add new block while the cost improves
  - update previous blocks in a cyclic fashion



- Stage 1: adding new blocks
- Stage 2: Cyclic updates

Algorithm: Summary

- Nassau
  - optimize the description length
  - add new block while the cost improves
  - update previous blocks in a cyclic fashion



- Stage 1: adding new blocks
- Stage 2: Cyclic updates

Results

- Compression ratio in % of the original description length
  - smaller numbers mean better compression

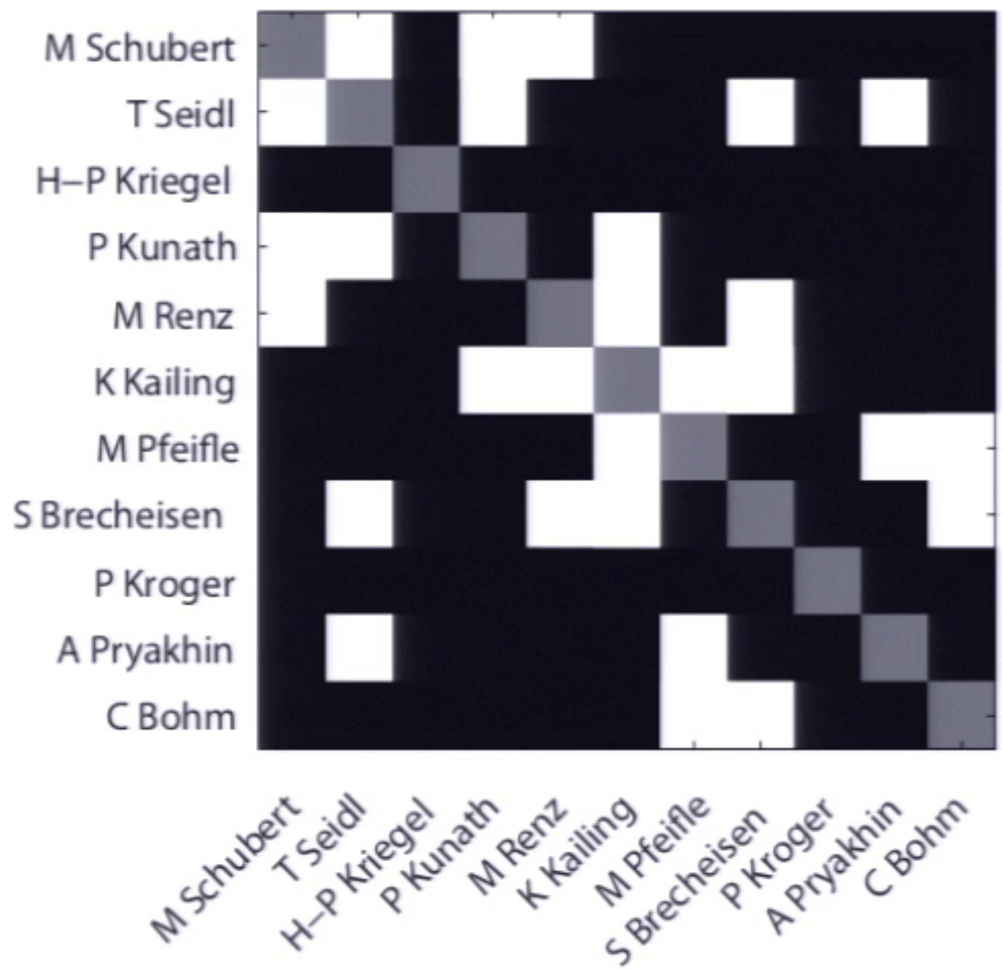
	Nassau		Panda+ [1]		Asso [2]	
Dataset	$L\%$	k	$L\%$	k	$L\%$	k
4News	93.1	12	**92.7**	5	93.6	17
DBLP co-auth.	**94.1**	60	95.9	11	95.8	178
Dialect	**42.0**	30	57.3	17	48.8	37
DNA Amp.	**43.6**	100	63.4	20	49.8	58
Mammals	**54.5**	29	66.8	8	64.6	17
Mushroom	72.6	4	63.6	15	**50.6**	59
Paleo	**89.7**	15	91.2	3	91.4	19

[1] C. Lucchese, S. Orlando, and R. Perego. A unifying framework for mining approximate top-k binary patterns. 2014

[2] P. Miettinen, T. Mielikäinen, A. Gionis, G. Das, and H. Mannila. The discrete basis problem. 2008

DBLP

- Example submatrices of *DBLP-coauth* selected by Nassau
- size 2345-by-2345, 60 factors



European Mammals

- Distribution of European mammals across different locations
- Pictured are the first four factors obtained by algorithms
- size 2670-by-194, 29 factors found