Summarizing Event Sequences with Generalized Sequential Patterns

Joscha Cüppers
Jilles Vreeken

Cüppers & Vreeken Below the Surface: Summarizing Event Sequences with Generalized Sequential Patterns.
In: Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), ACM, 2023

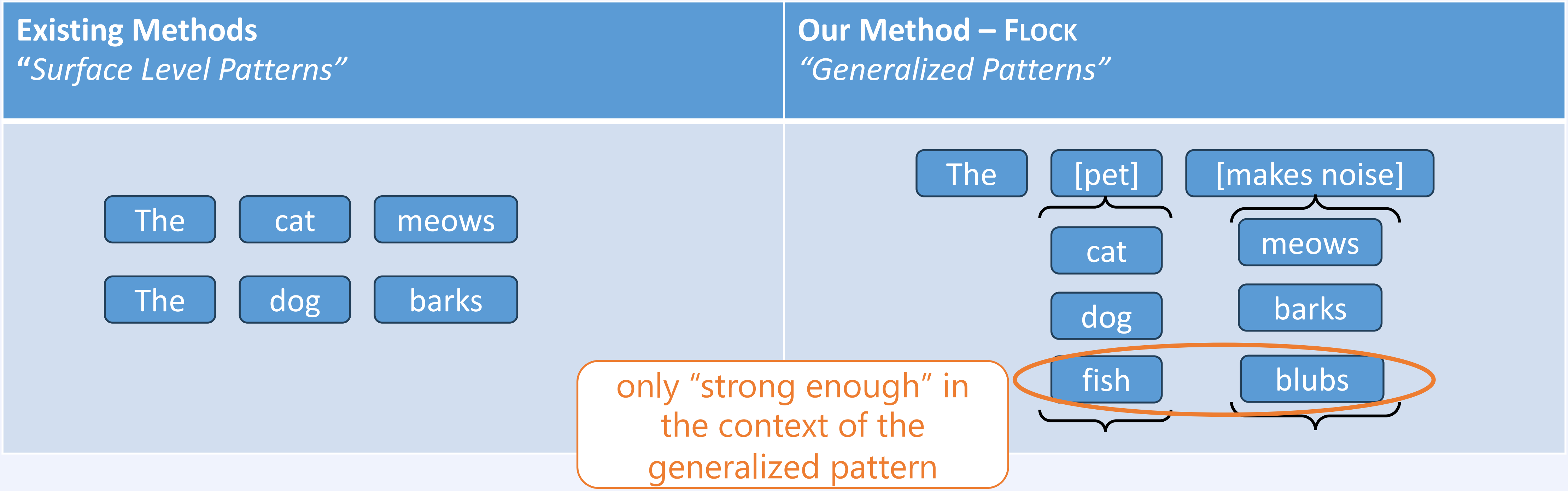
CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

Problem

Given only a set of event sequences, D.

Goal, report a set of **generalized events** and a set P of **generalized patterns**, that succinctly summarize D.

Generalized Pattern?

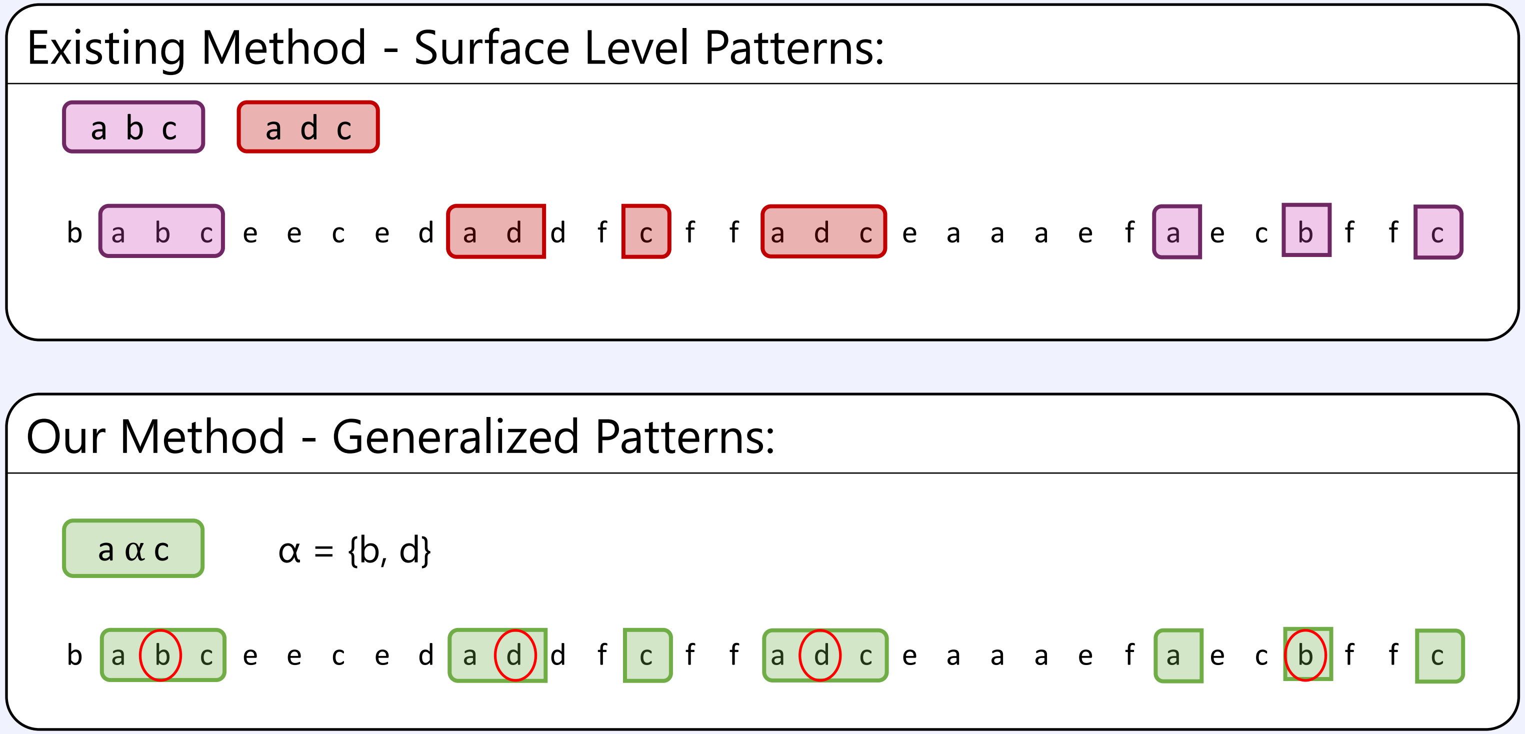


Existing Methods
“Surface Level Patterns”

Our Method – Flock
“Generalized Patterns”

only “strong enough” in the context of the generalized pattern

Generalized Patterns



Existing Method - Surface Level Patterns:

a b c
a d c
b a b c e e c e d a d d f c f f a d c e a a a e f a e c b f f c

Our Method - Generalized Patterns:

a α c
α = {b, d}
b a b c e e c e d a d d f c f f a d c e a a a e f a e c b f f c

- Set of Observed Events - Ω_o e.g. $\Omega_o = \{a, b, c, \dots\}$
- Set of Generalized Events - Ω_g e.g. $\Omega_g = \{\alpha\}$
- Alphabet – $\Omega = \Omega_o \cup \Omega_g$

How do we do that?

The Minimum Description Length (MDL) principle:

given a model class $\mathcal{M}$, the best model $\boldsymbol{M} \ \in \mathcal{M}$
is that $\boldsymbol{M}$ that minimizes

$$L(D, M) = \ L(M) + L(D|\ M)$$

where:

$L(M)$ is the length of the model, in bits
$L(D \mid M)$ is the length of the data, in bits, when encoded using $\boldsymbol{M}$

Length of Model

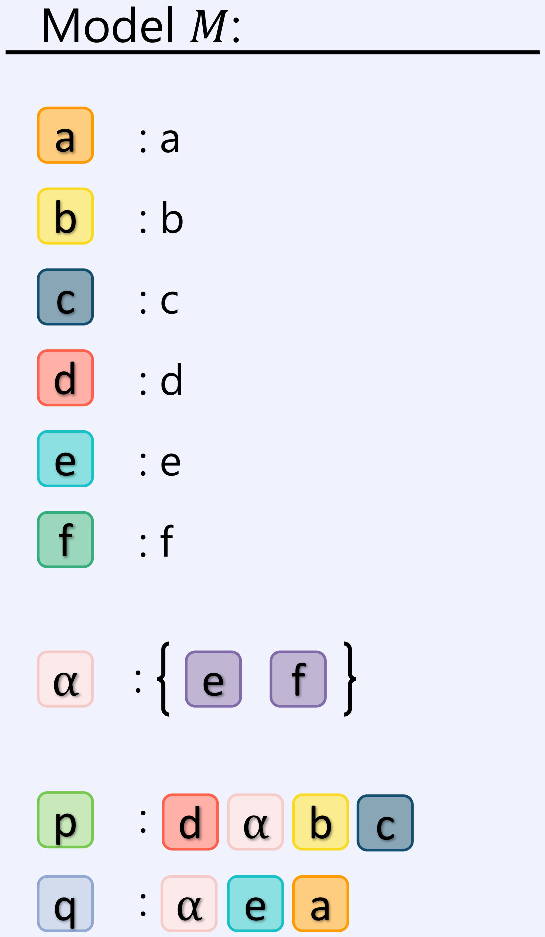
$$L(M) = L(CT) + L(\Omega_g)$$

Code Table – Pattern set and usage of each pattern

$$L(\mathrm{CT}) = \underbrace{L_{\mathbb{N}}(|P'|)}_{\text{how many patterns}} + \underbrace{L_{\mathbb{N}}(usage(P))}_{\text{usage sum over all patterns}} + \underbrace{\log \binom{usage(P) - 1}{|P| - 1}}_{\text{usage of each pattern}} + \underbrace{\sum_{p \in P'} L(p)}_{\text{encoding of patterns}}$$

Set of Generalized Events $\boldsymbol{\Omega_g}$

$$L(\Omega_g) = \sum_{e \in \Omega_g} L(e)$$



Model M:

a : a

b : b

c : c

d : d

e : e

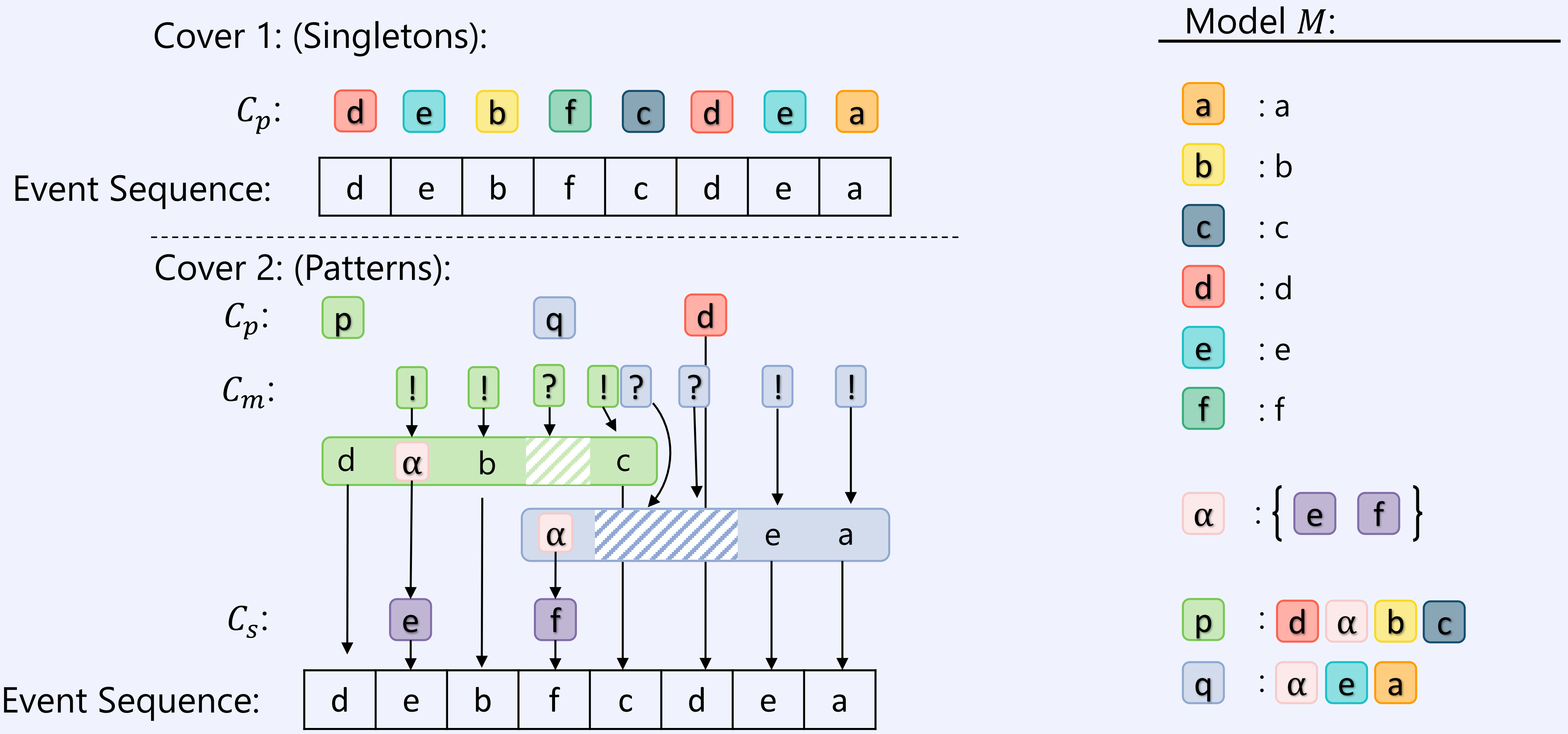
f : f

α : { e f }

p : d α b c

q : α e a

Length of Data



Cover 1: (Singletons):

C_p: d e b f c d e a

Event Sequence: d e b f c d e a

Cover 2: (Patterns):

C_p: p q d

C_m: ! ! ? ! ? ? ! !

C_s: e f

Event Sequence: d e b f c d e a

Model M:

a : a

b : b

c : c

d : d

e : e

f : f

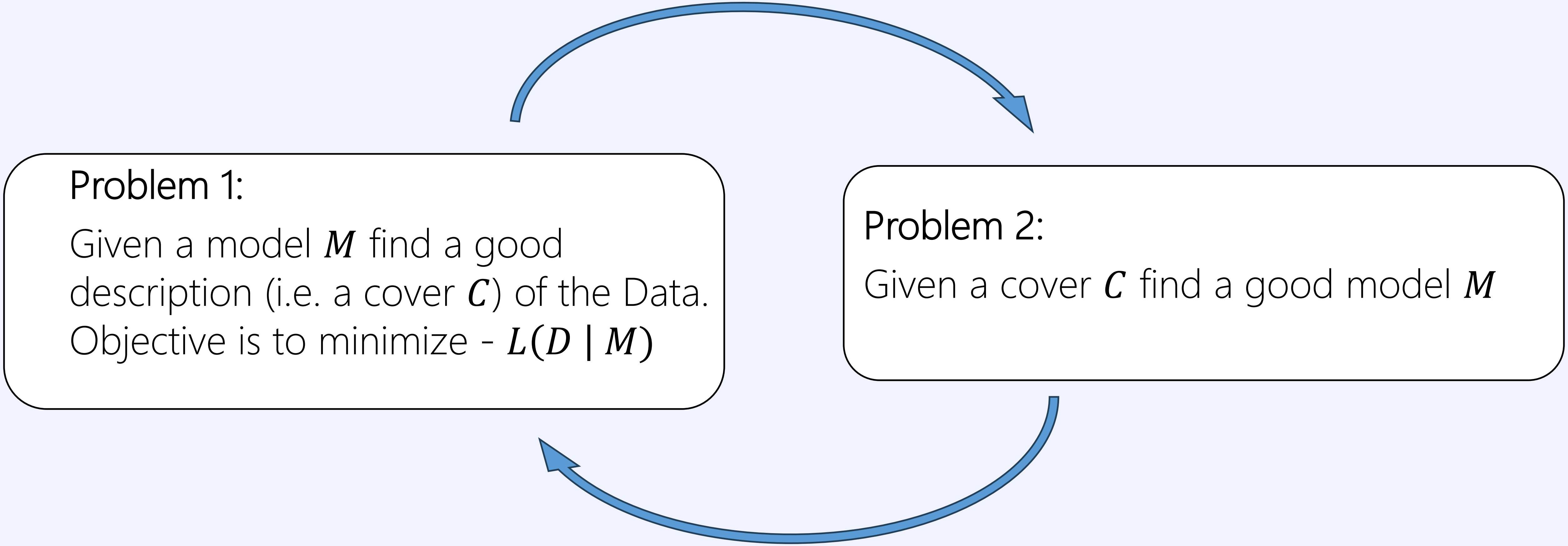
α : { e f }

p : d α b c

q : α e a

$$L(D|M) = L(C_p) + L(C_m) + L(C_s)$$

Mining Models



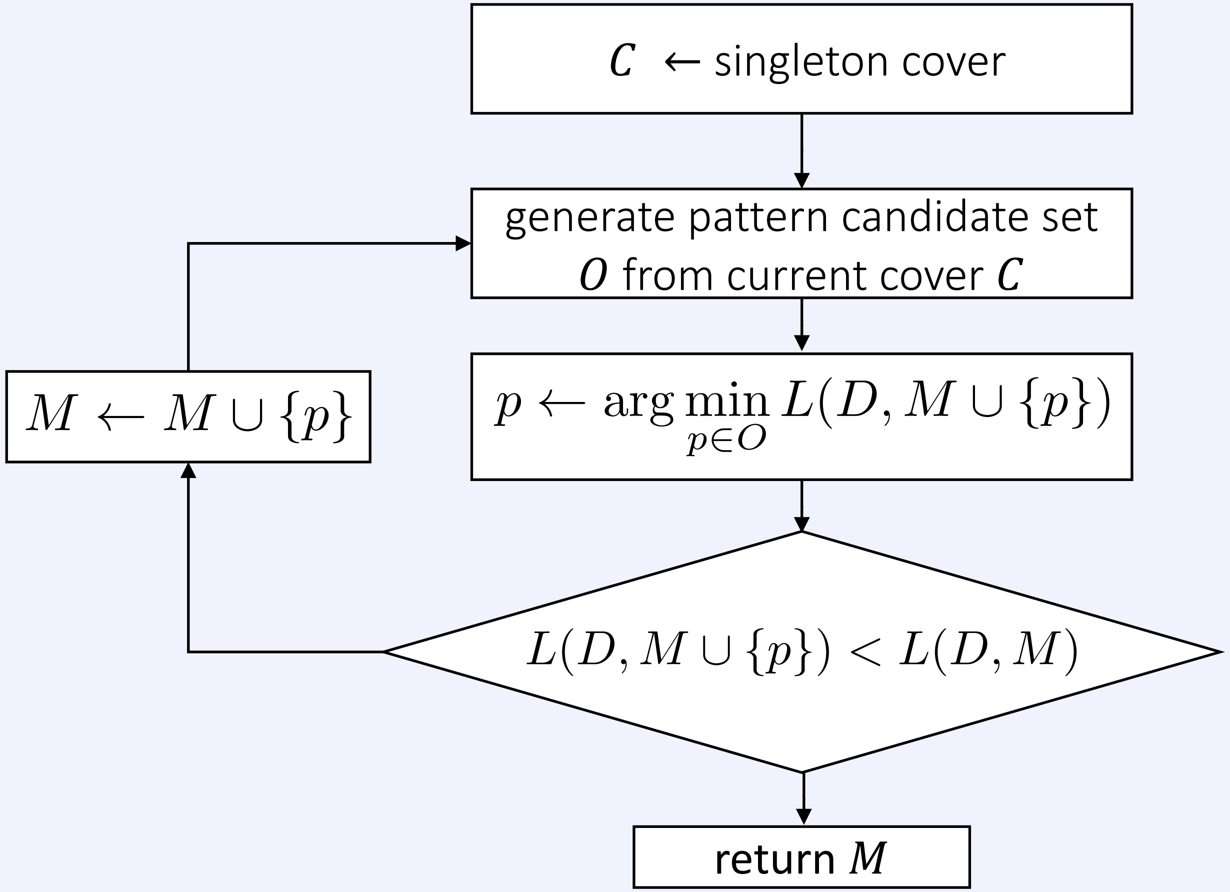
Problem 1:

Given a model M find a good description (i.e. a cover C) of the Data. Objective is to minimize - $L(D \mid M)$

Problem 2:

Given a cover C find a good model M

FLOCK Algorithm – Basic Idea



C ← singleton cover

generate pattern candidate set O from current cover C

$p \leftarrow \arg\min_{p \in O} L(D, M \cup \{p\})$

$L(D, M \cup \{p\}) < L(D, M)$

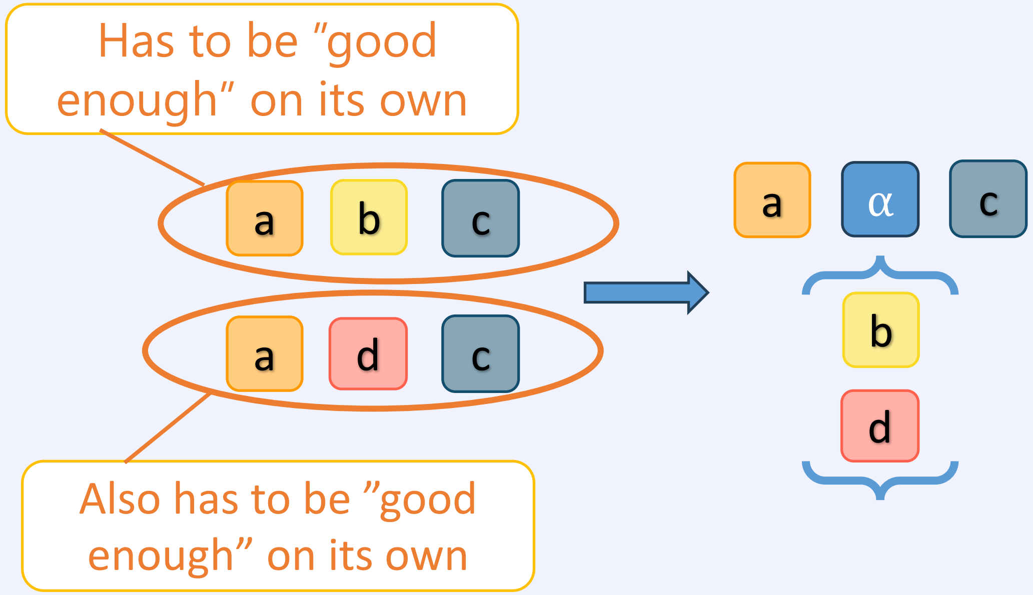
$M \leftarrow M \cup \{p\}$

return M

Discovering Generalized Events

Merge

1. mine "surface level" patterns
2. merge patterns



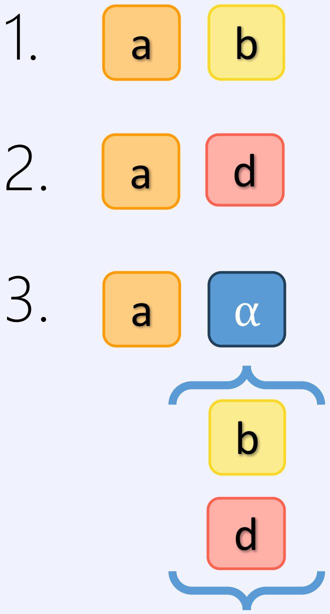
Has to be "good enough" on its own

Also has to be "good enough" on its own

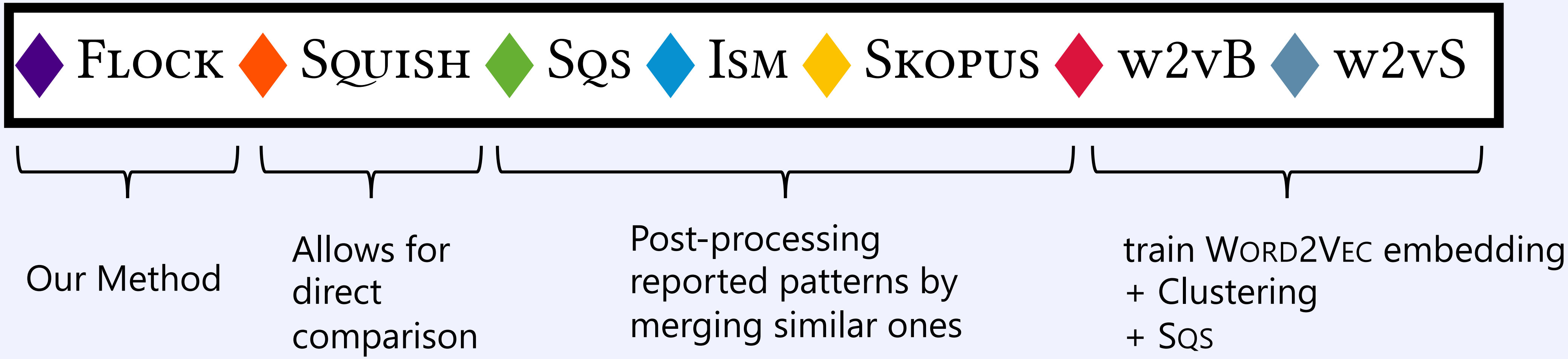
Candidate Generation

Suppose a is often followed by b and d , with similar number of gaps.

Generate Candidates:



Experiments / Related Work



◆ Flock ◆ Squish ◆ Sqs ◆ Ism ◆ Skopus ◆ w2vB ◆ w2vS

Our Method

Allows for direct comparison

Post-processing reported patterns by merging similar ones

train Word2Vec embedding
+ Clustering
+ Sqs

Evaluation – Synthetic Data

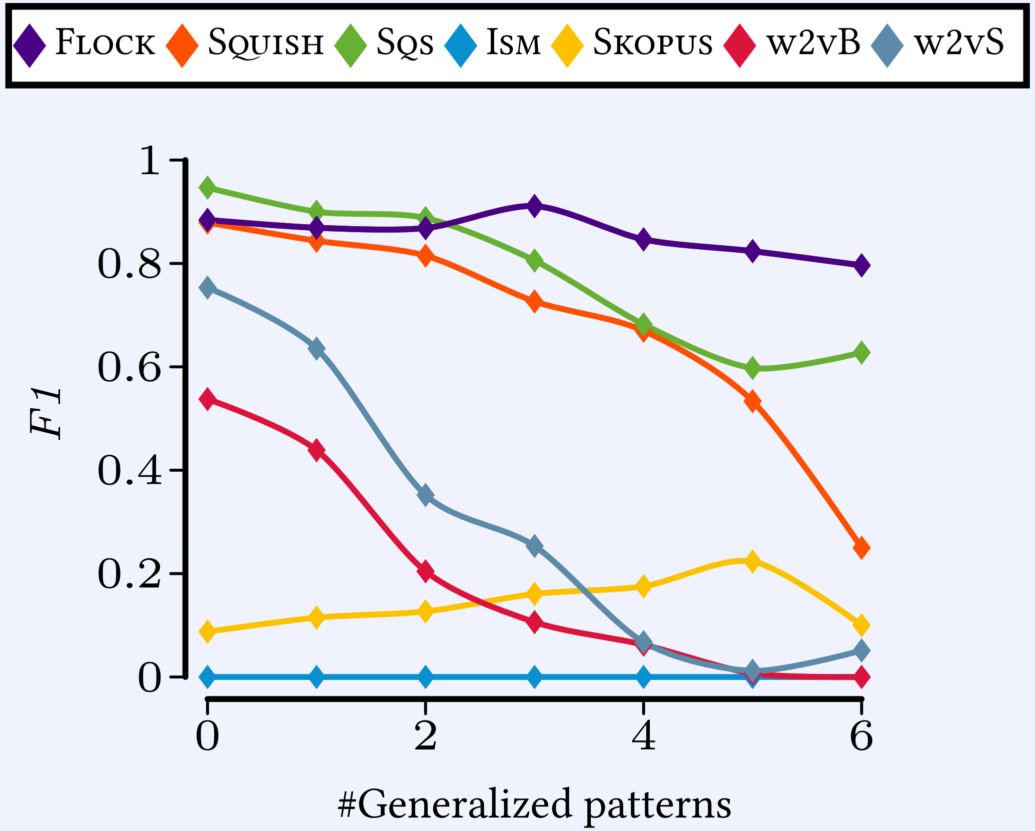
Data with known ground truth.

0 Generalized Patterns → 1 Generalized Patterns → ...

1.	a b c
2.	b d t
3.	y j i
4.	t i f
5.	y d p
6.	q h k
⋮	
30	x l d

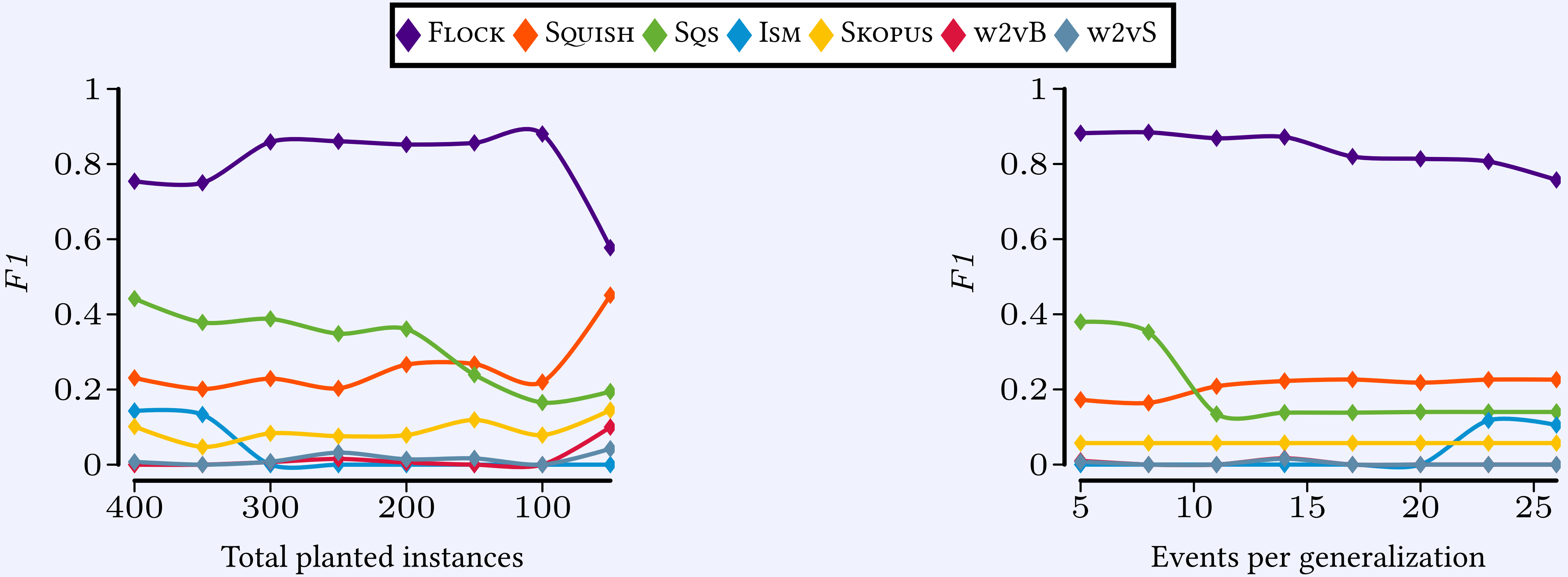
1.	a b α
2.	q h k
⋮	
26	x l d

α = {j, t, p, w, i}



Evaluation – Synthetic Data

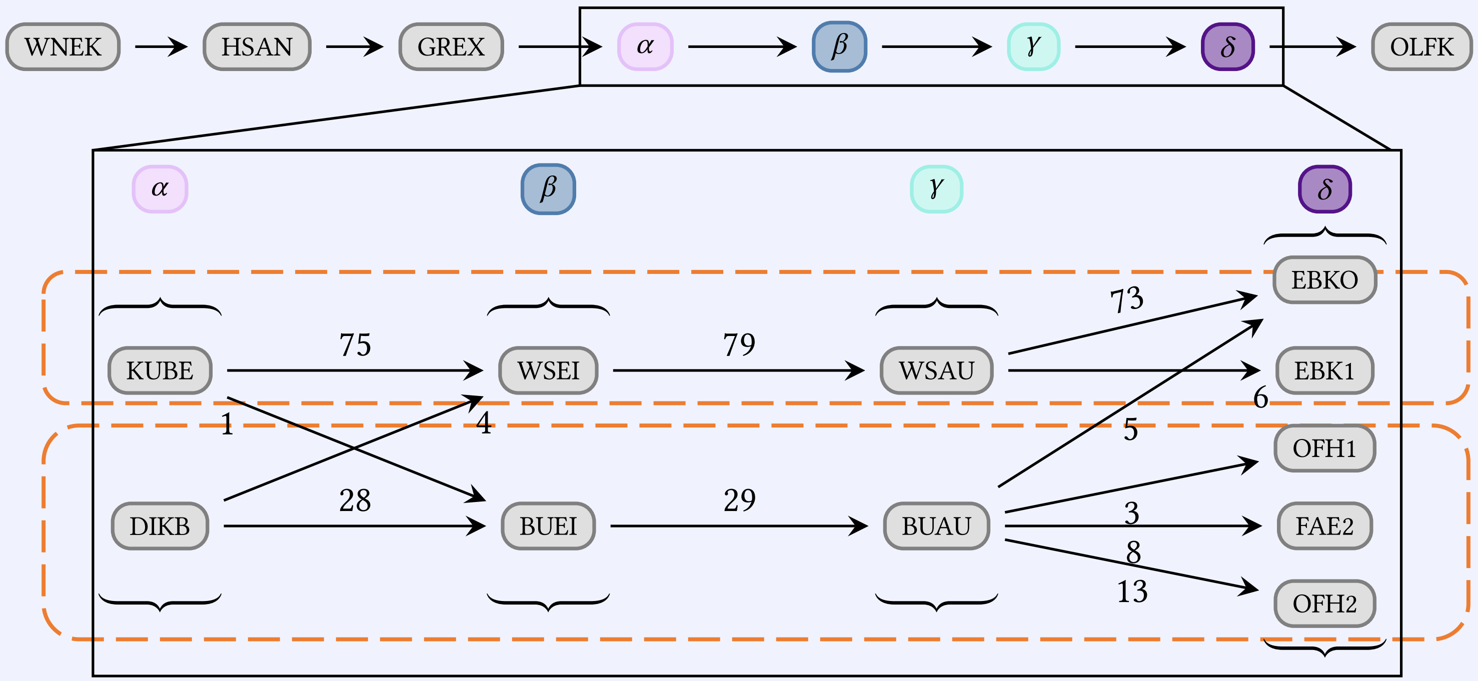
- 5 patterns
- 5 generalizations in total
- 2 generalizations per pattern



$|\alpha| = 5, \dots, |\alpha| = 25$

Real World Pattern Example

- Data: Production Log of Steel Rolling Mill



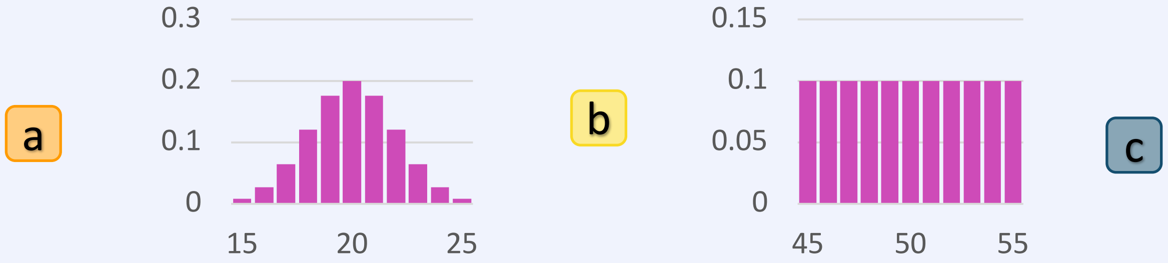
line for wide steel

line for thin steel

Image: https://c1.wallpaperflare.com/preview/82/931/800/bloom-rolling-mill-rail-hot-product.jpg

Something Else

Patterns with Predictable Inter-Event Delays



- Explicit modeling of delays between events
- Ability to model and **discover** patterns with long inter-event delays

Conclusion

We consider the problem of finding a succinct set of generalized patterns that describes the data

- Generalized pattern describe general “behavior”– not instances
- Capture infrequent instances of general patterns

Formalized the problem with the Minimum Description Length (MDL) principle

- Define model class and encoding of model
- Encoding of Data given a Model

Present greedy algorithm to mine patterns and generalized events

Evaluation shows that we can discover generalized patterns

- Recover ground truth well on synthetic data

Thank you!

We consider the problem of finding a succinct set of generalized patterns that describes the data

- Generalized pattern describe general “behavior”– not instances
- Capture infrequent instances of general patterns

Formalized the problem with the Minimum Description Length (MDL) principle

- Define model class and encoding of model
- Encoding of Data given a Model

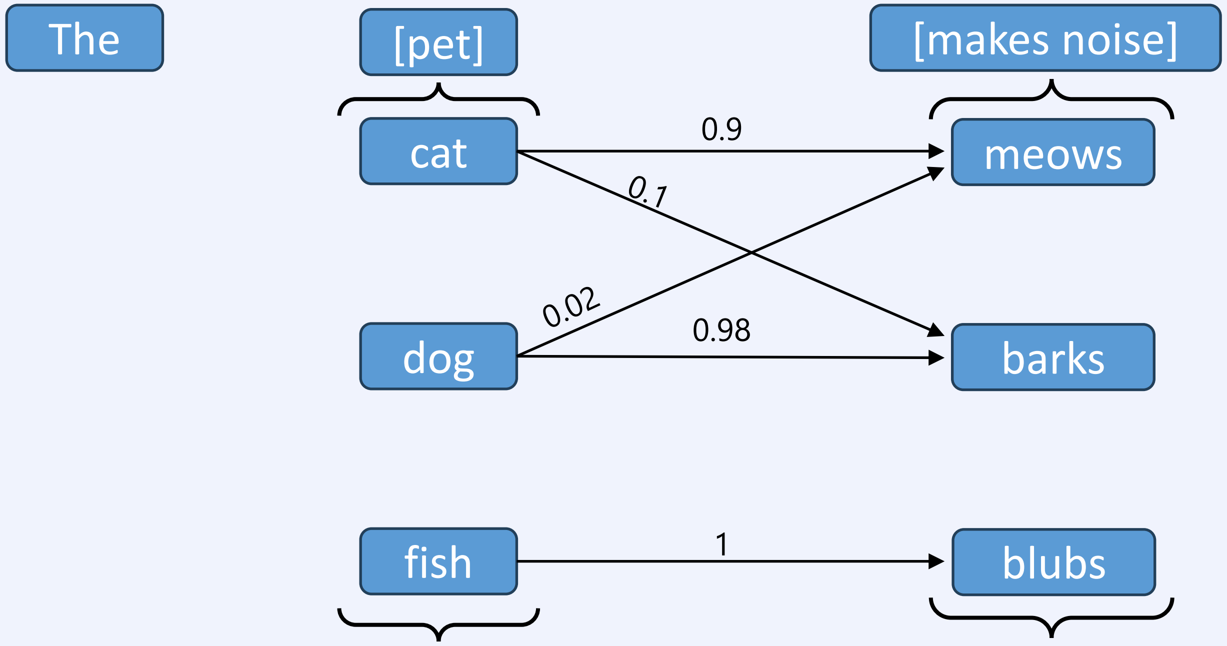
Present greedy algorithm to mine patterns and generalized events

Evaluation shows that we can discover generalized patterns

- Recover ground truth well on synthetic data

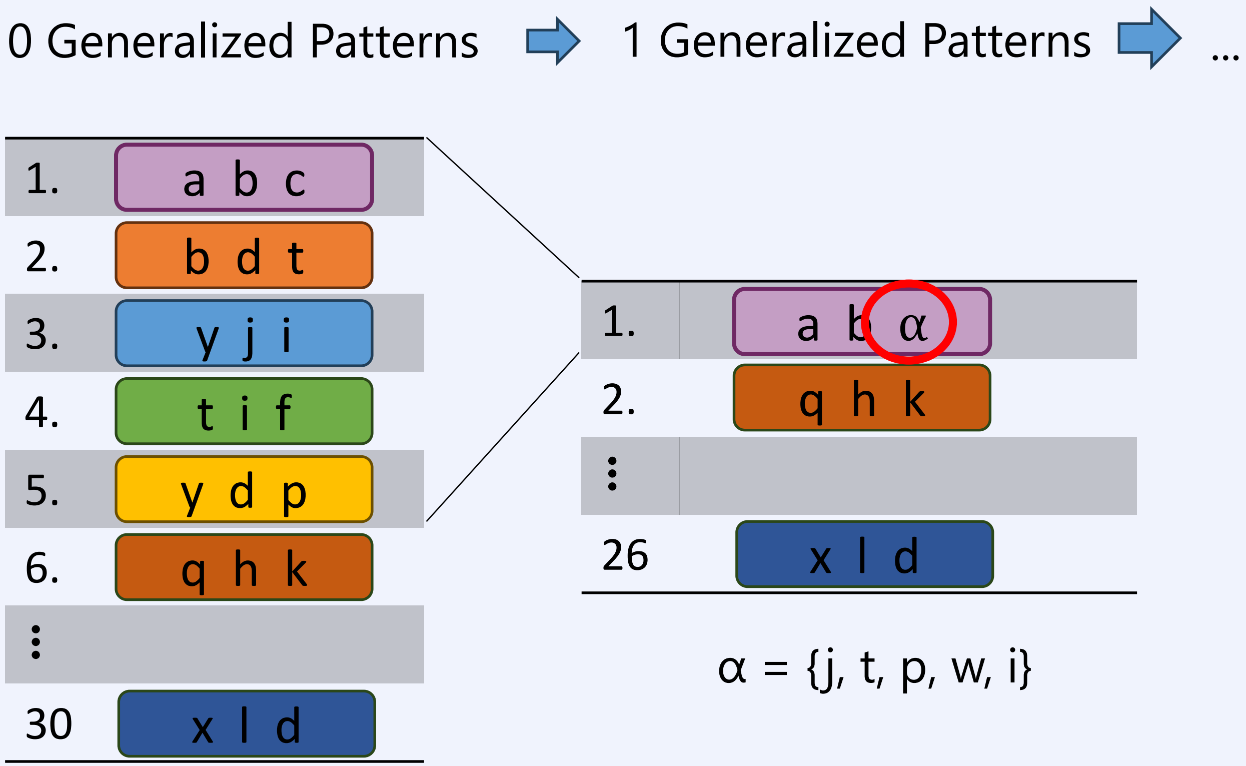
Transition probabilities / frequencies

Pattern:



Evaluation – Synthetic Data

- Data with known ground truth.

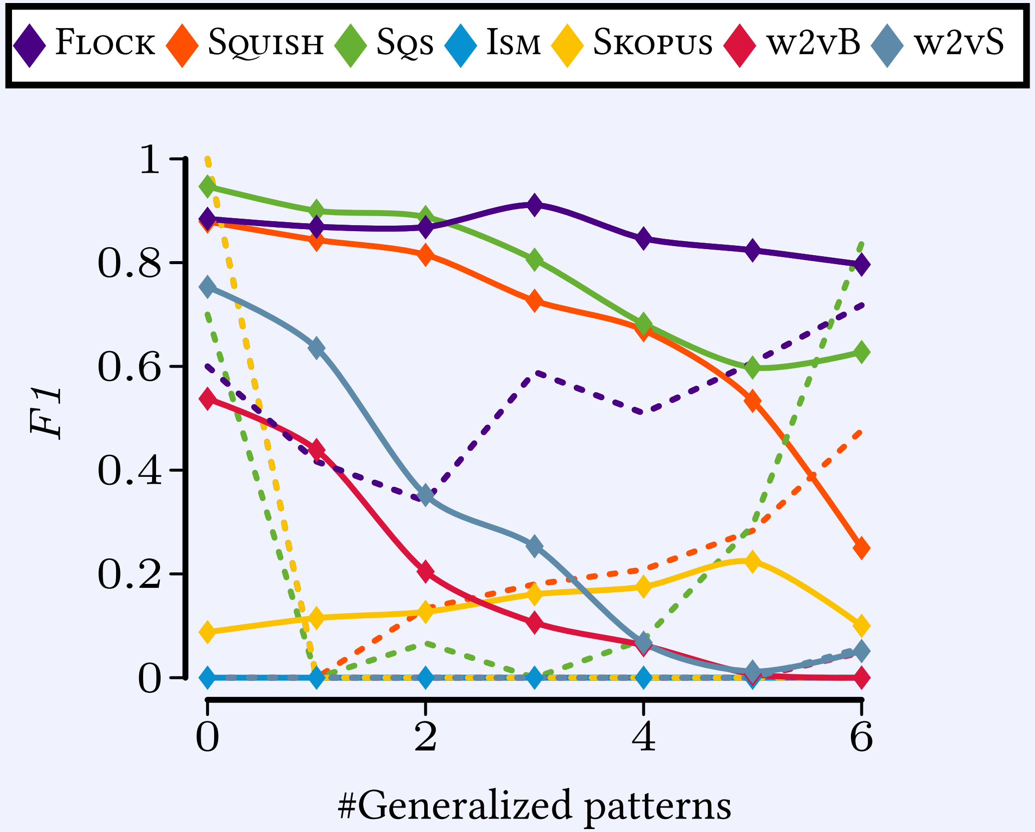


0 Generalized Patterns → 1 Generalized Patterns → ...

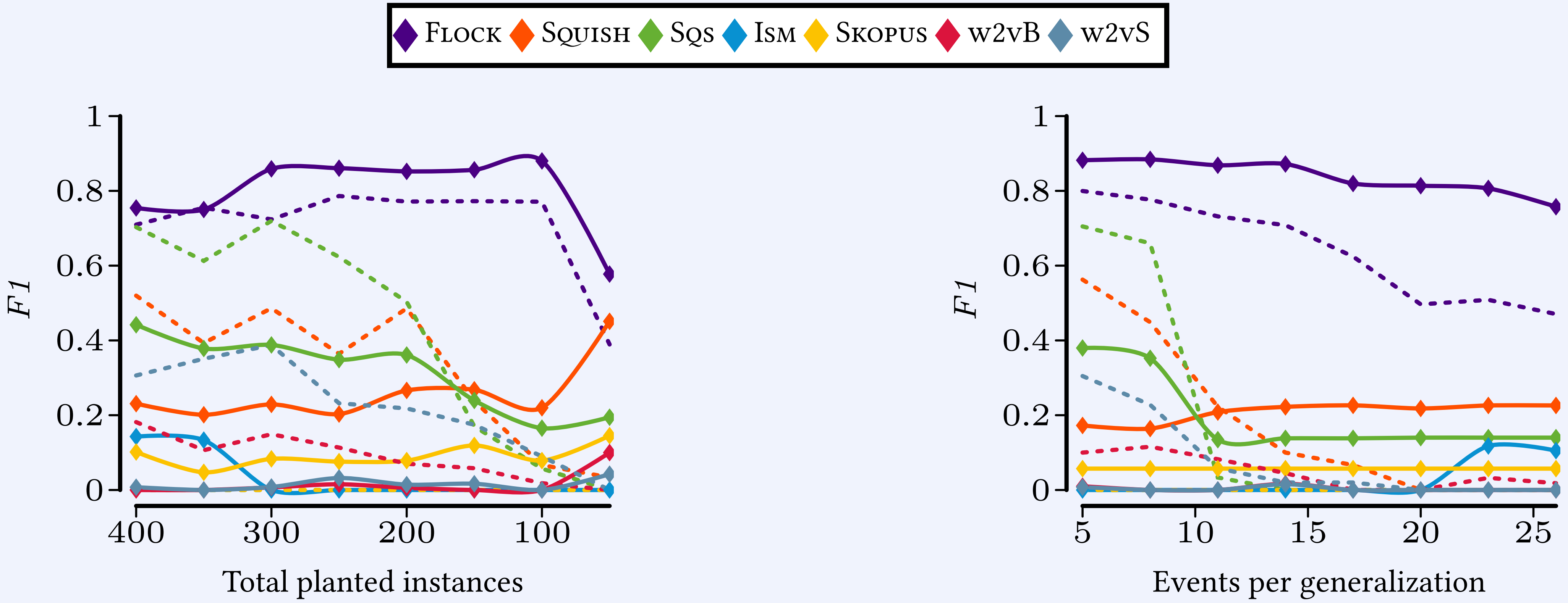
1.	a b c
2.	b d t
3.	y j i
4.	t i f
5.	y d p
6.	q h k
⋮	
30	x l d

1.	a b α
2.	q h k
⋮	
26	x l d

α = {j, t, p, w, i}



Evaluation – Synthetic Data



$|\alpha| = 5, \ldots, |\alpha| = 25$