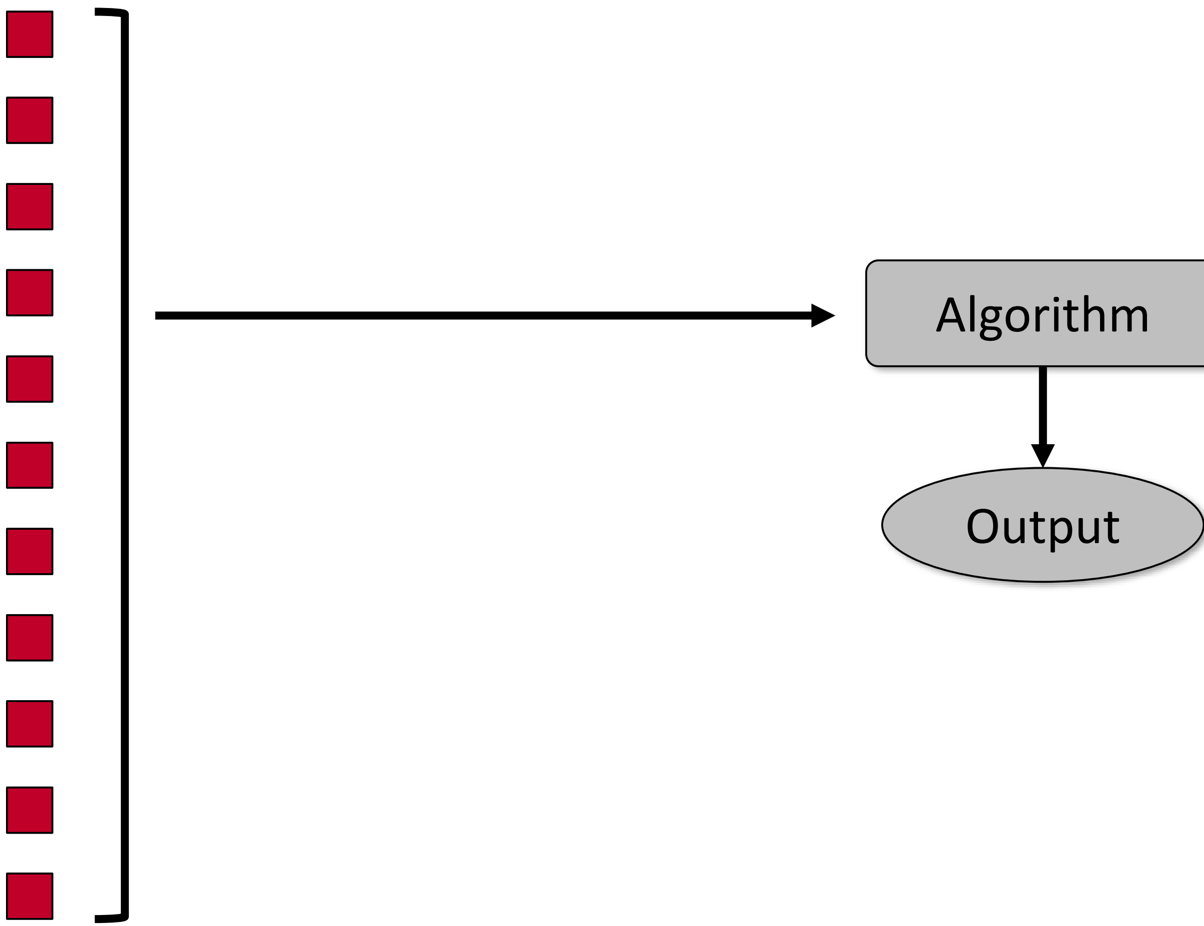
A Probabilistic Condensed Representation of Data for Stream Mining

Michael Geilke, Andreas Karwath, and Stefan Kramer

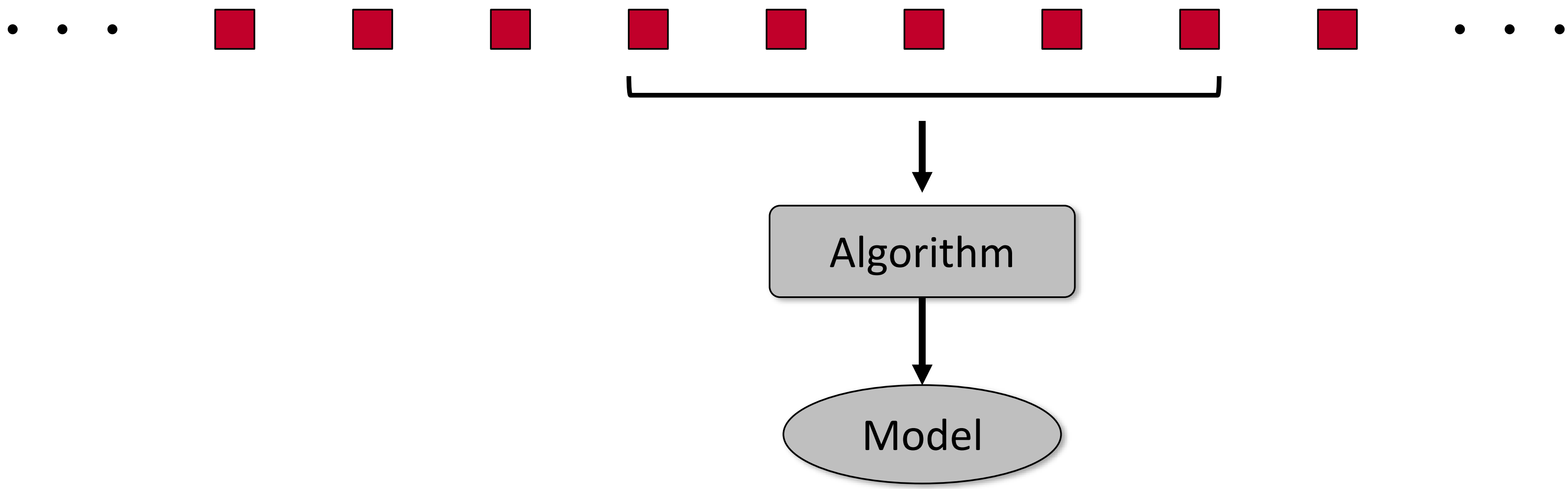
Johannes Gutenberg-Universität Mainz, Germany

November 1, 2014

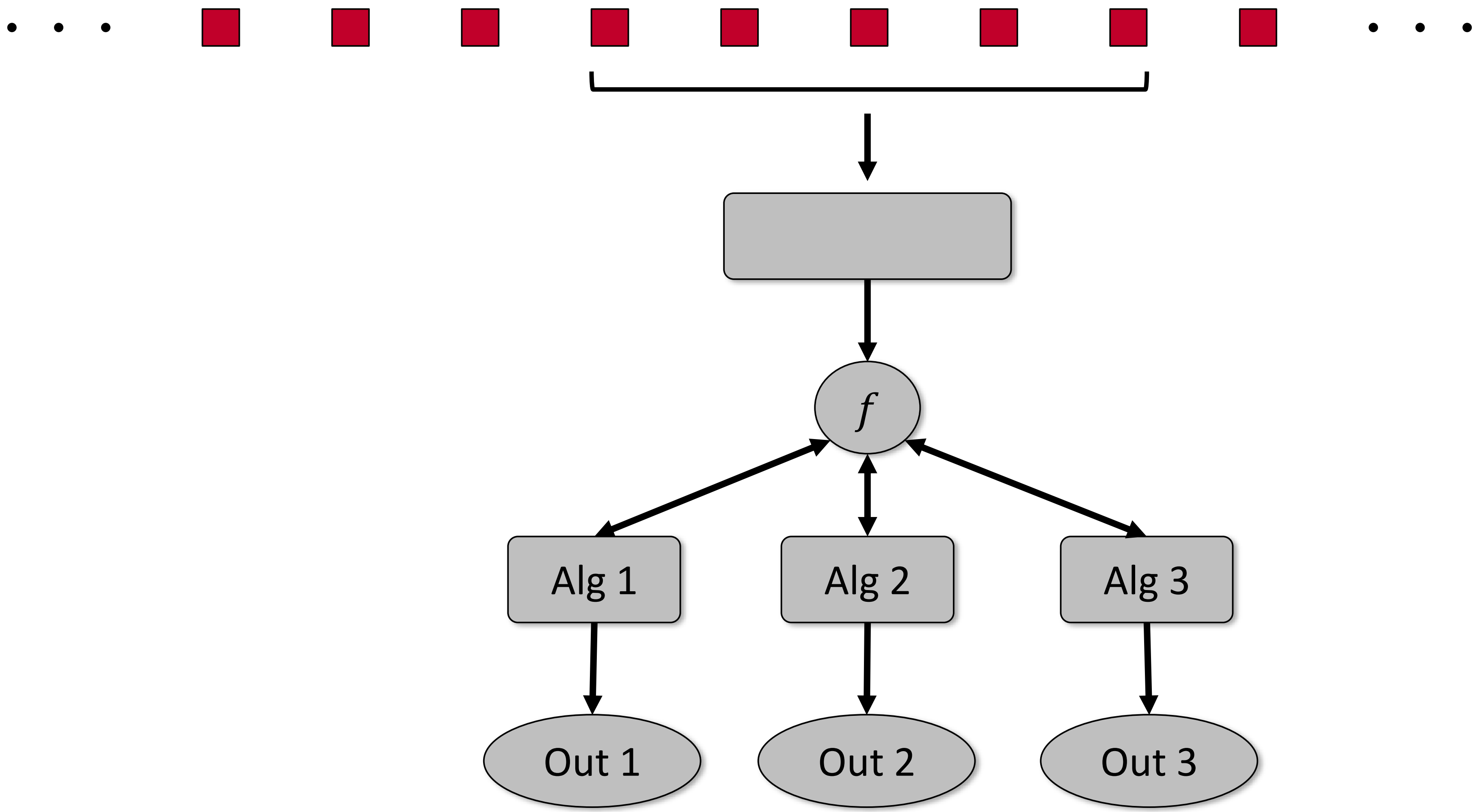
Batch Learning



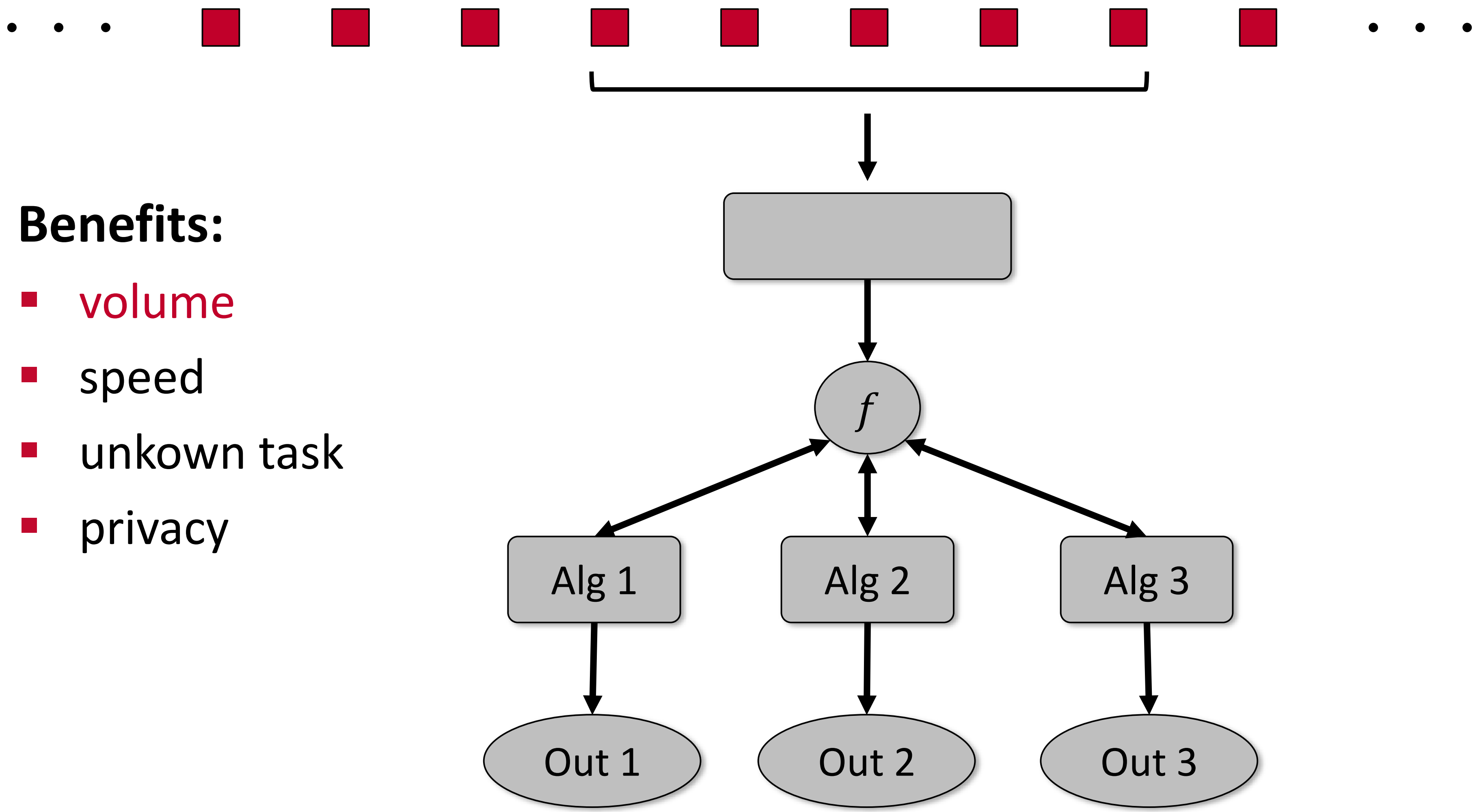
Streams



Condensed Representation



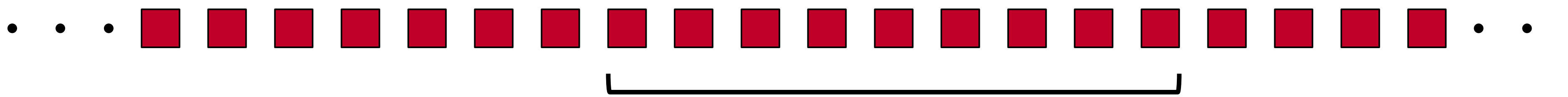
Benefits



Benefits:

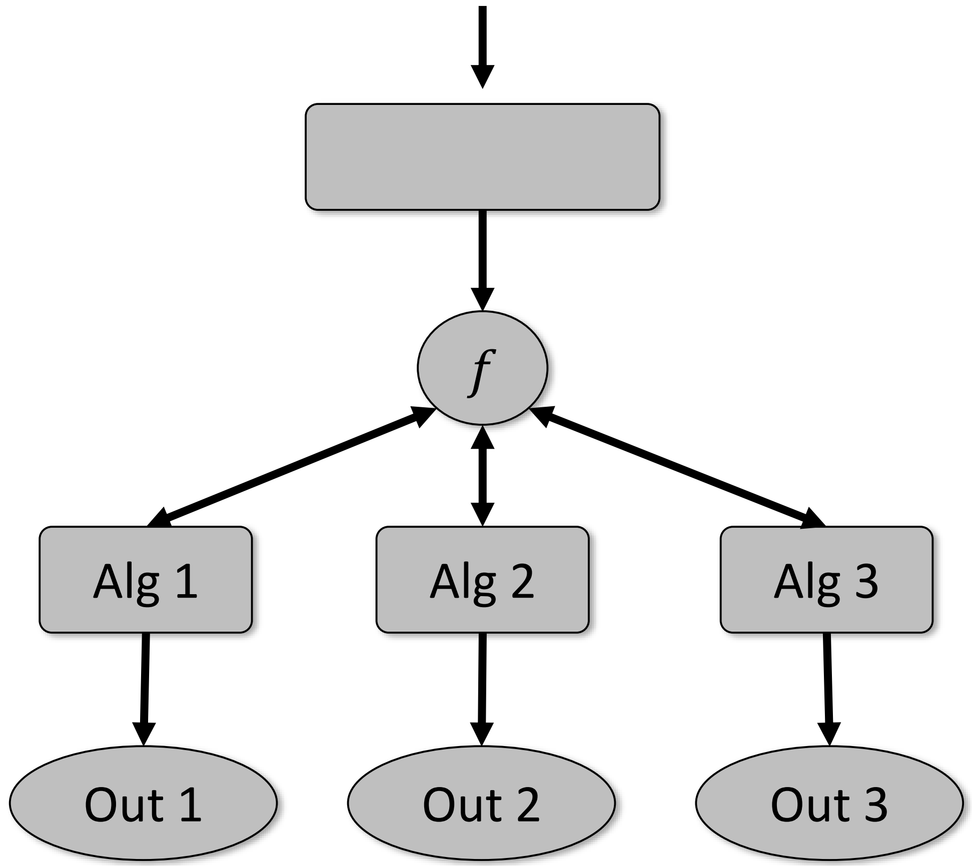
- volume
- speed
- unkown task
- privacy

Benefits

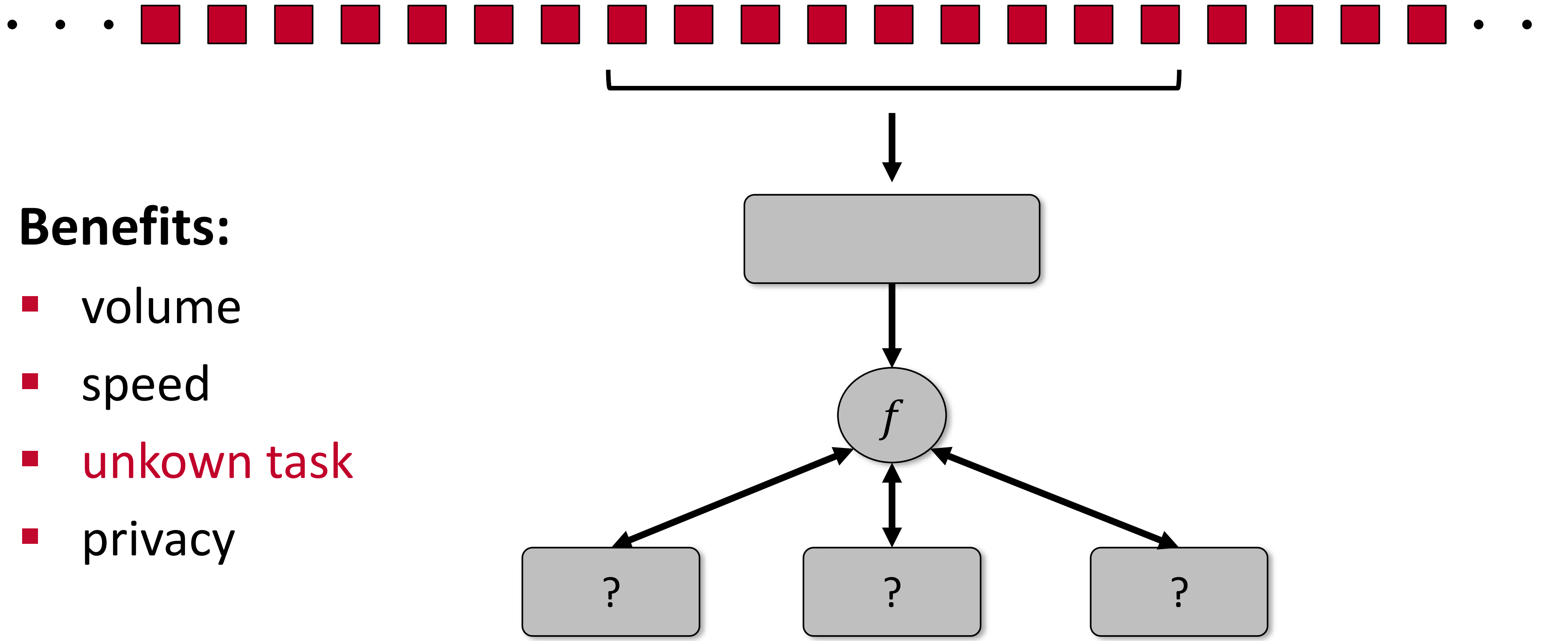


Benefits:

- volume
- speed
- unkown task
- privacy



Benefits



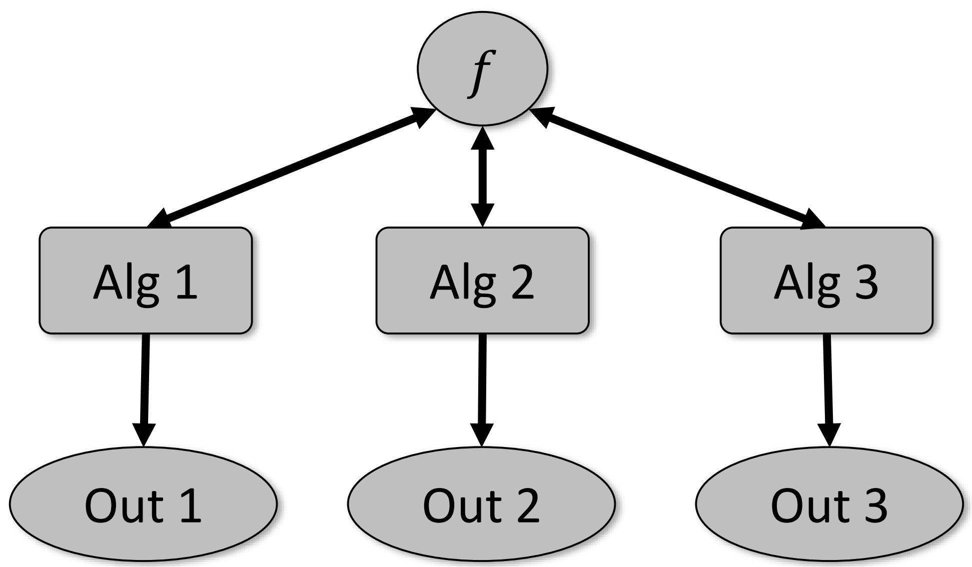
Benefits:

- volume
- speed
- unkown task
- privacy

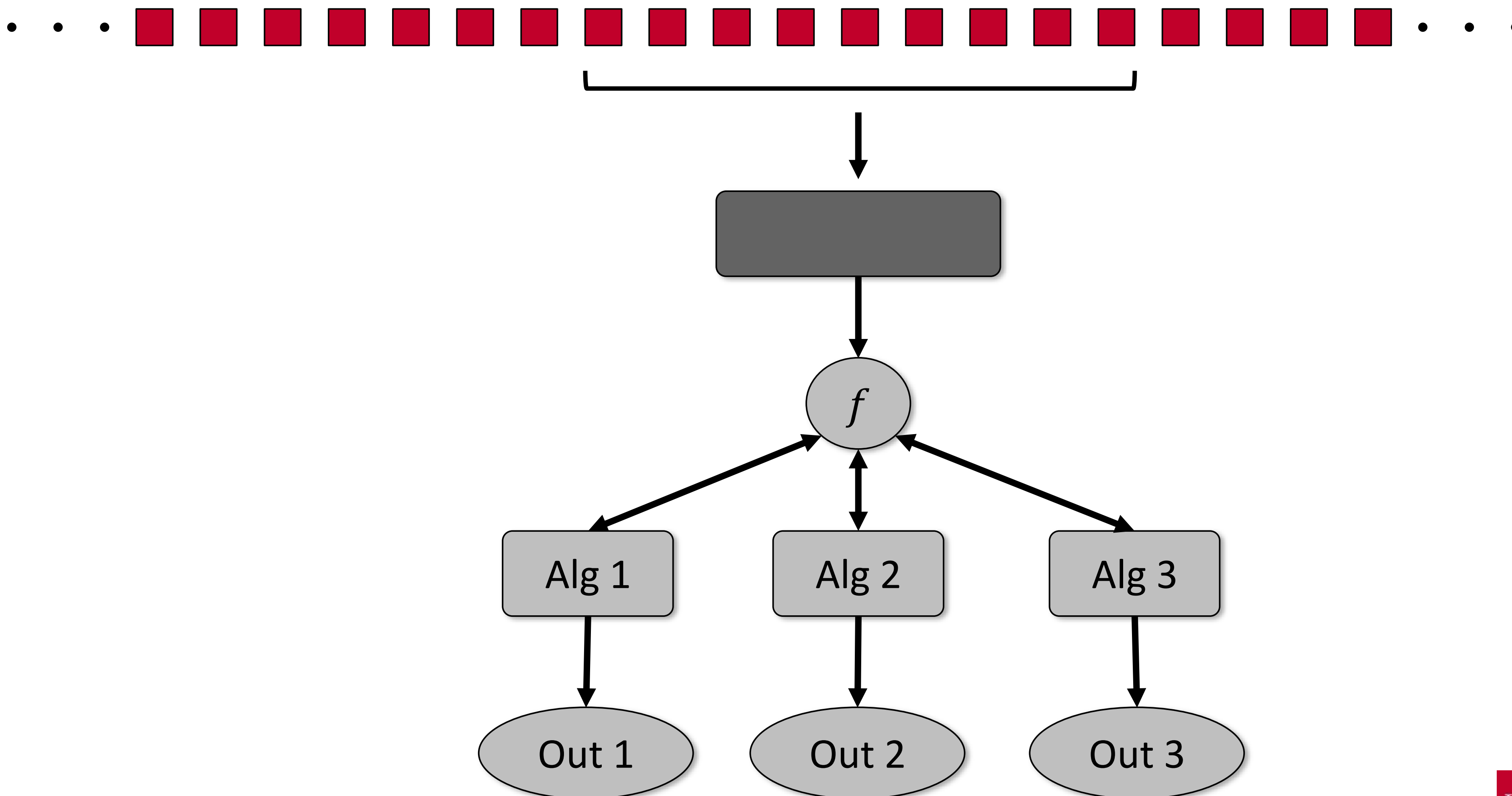
Benefits

Benefits:

- volume
- speed
- unkown task
- privacy



Condensed Representation



EDDO (Estimation of Discrete Densities Online)

Applying the product rule to $f(X_1, X_2, \ldots, X_n)$ yields

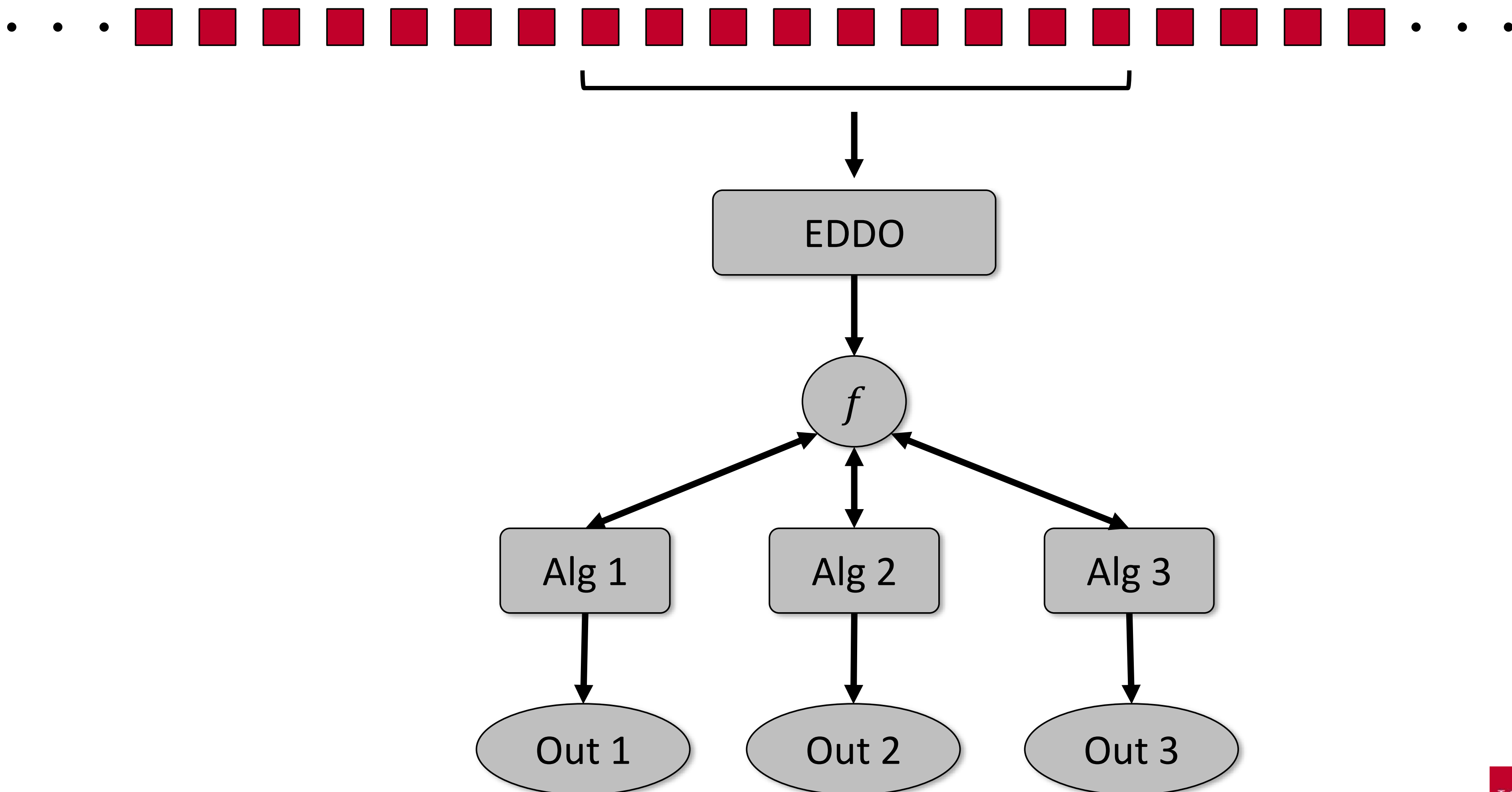
$$f_1(X_1) \cdot f_2(X_2 \mid X_1) \cdot \; \ldots \; \cdot f_n(X_n \mid X_1, X_2, \ldots, X_{n-1})$$

Classifier

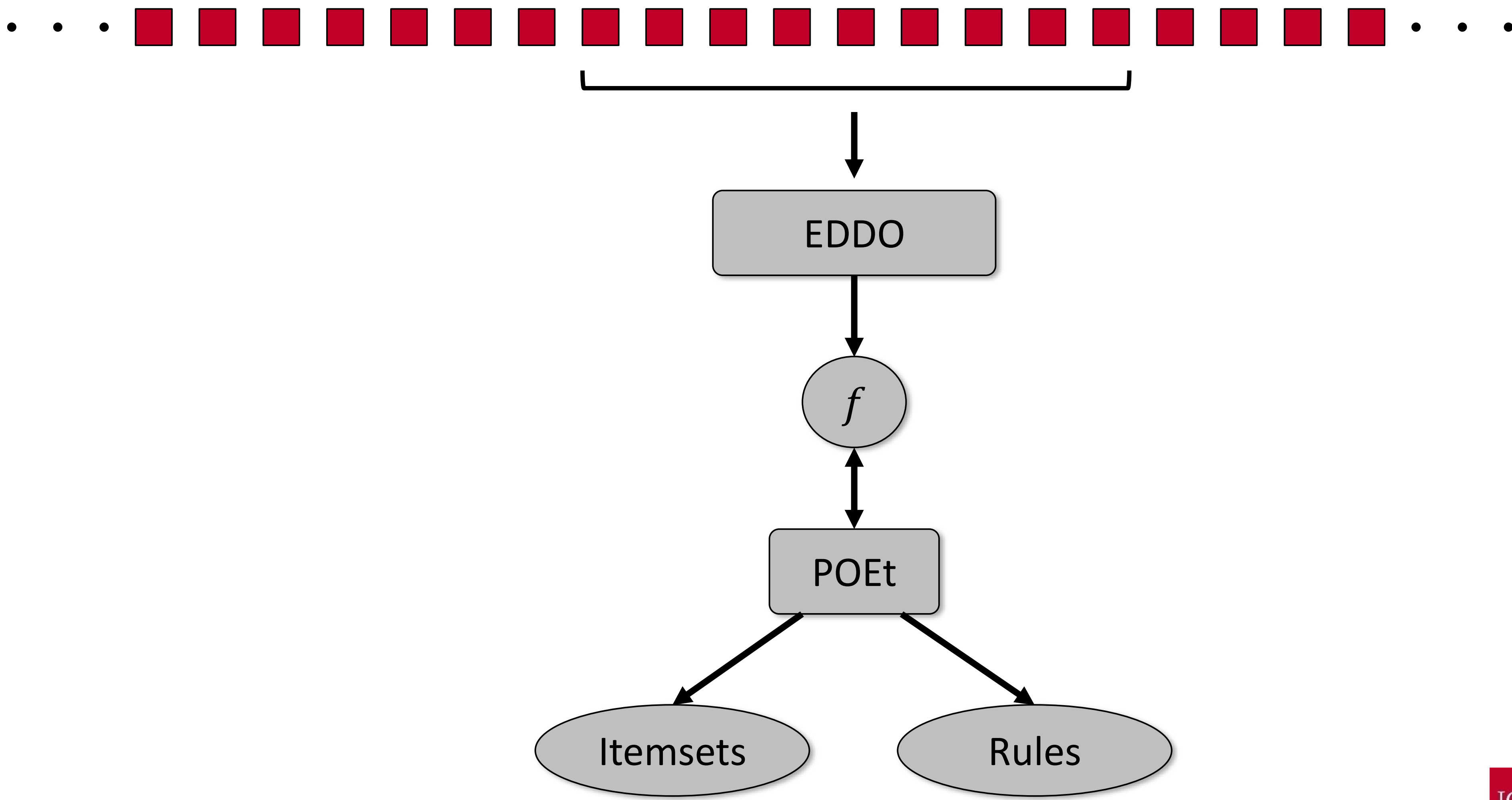
Majority class	for $f_1(X_1)$
Hoeffding trees	for $f_i(X_i \mid X_1, X_2, \ldots, X_{i-1})$

Both enable the estimation in an online fashion.

MiDEO (Mining Density Estimates inferred Online)



Pattern Mining



Setting

Itemsets $(X_4, v_3), (X_9, v_1), (X_1, v_5)$

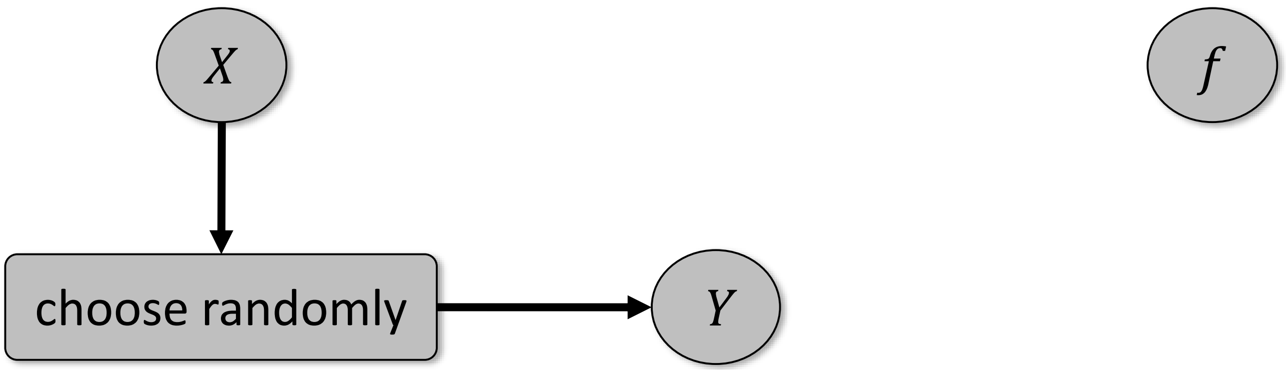
Association rules $(X_4, v_3), (X_9, v_1) \Rightarrow (X_1, v_5)$

Measure of interestingness

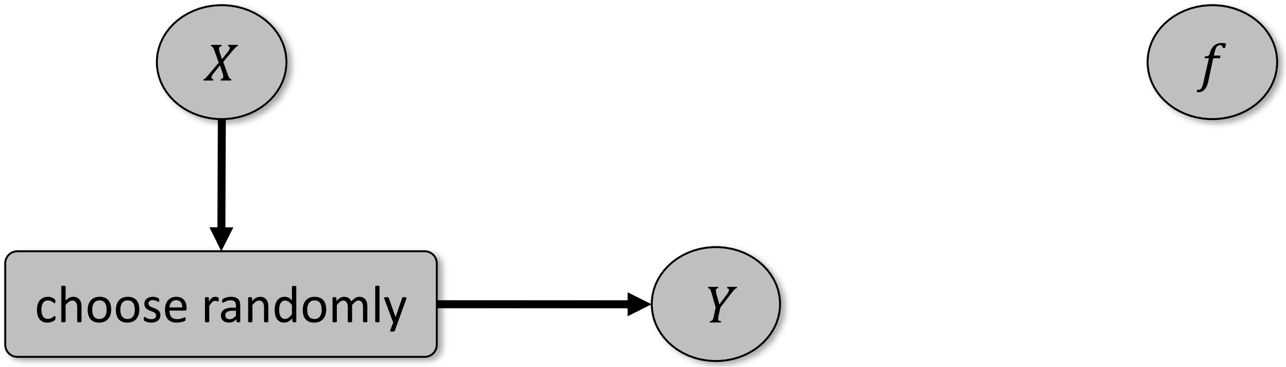
- minimum support threshold
- confidence $f((X_1, v_5)|(X_4, v_3), (X_9, v_1))$

POEt – generating itemsets

POEt – generating itemsets

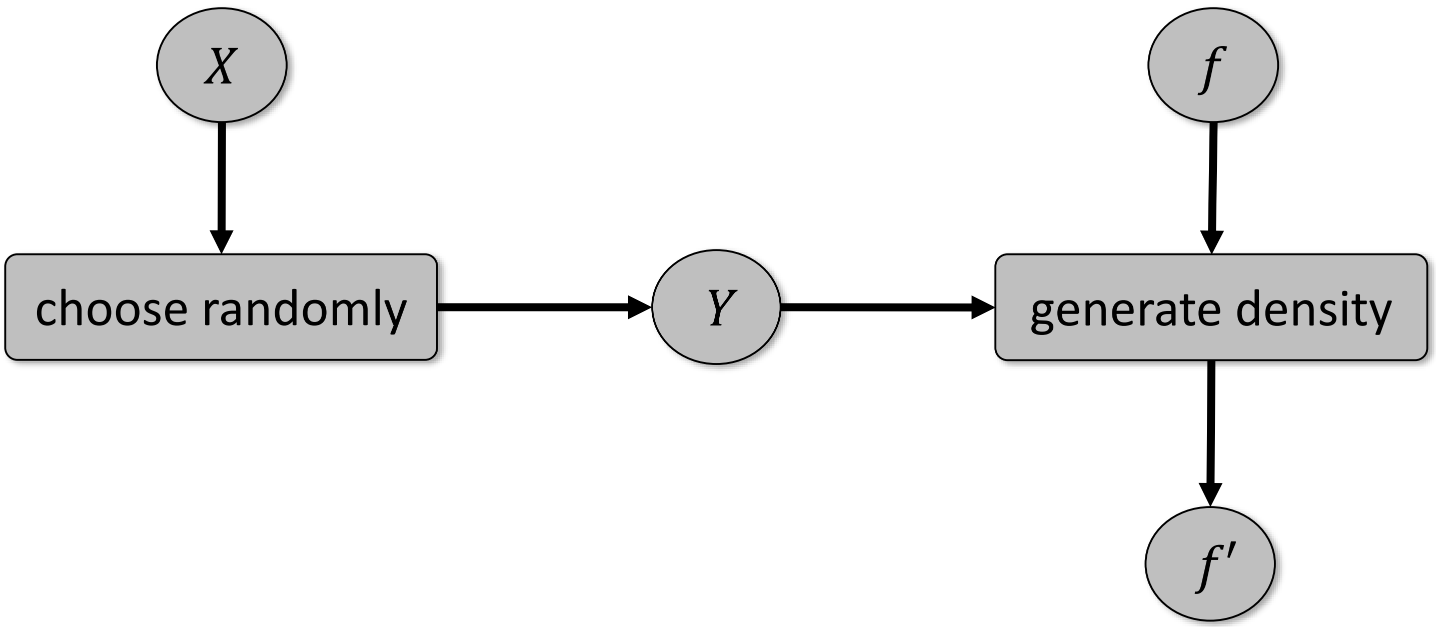


POEt – generating itemsets

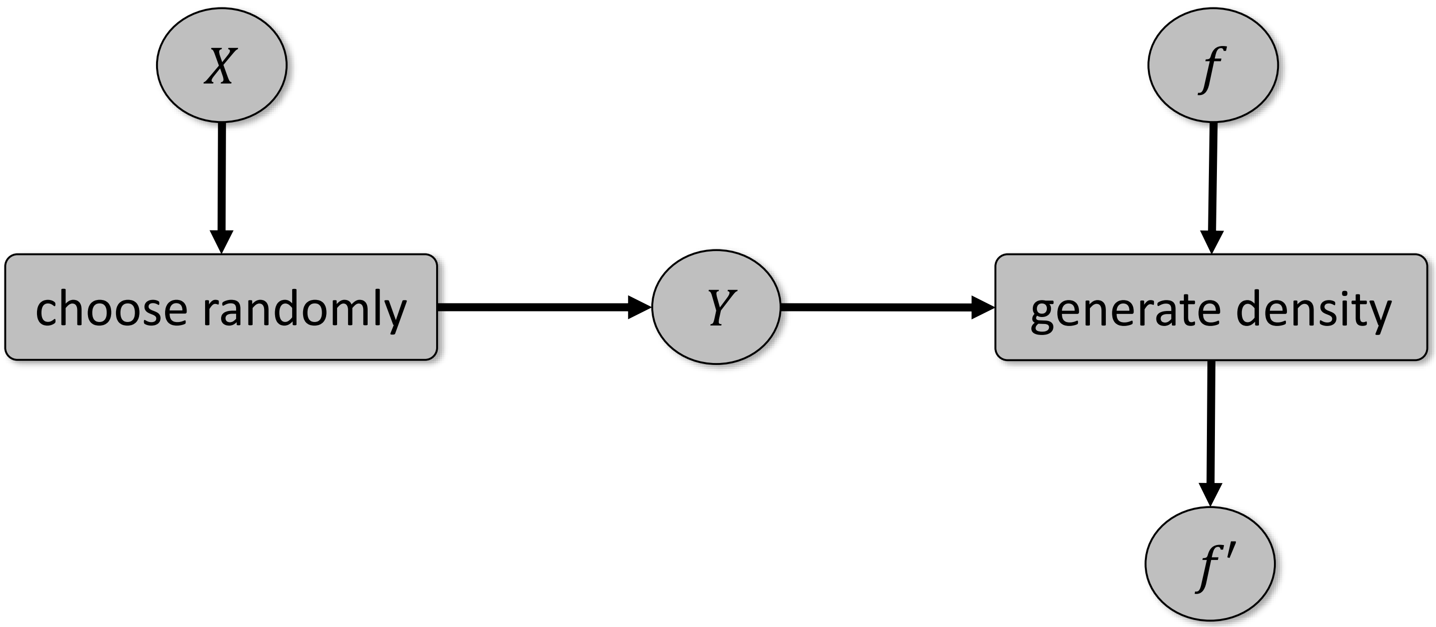


- Geometric distribution for size
- Uniformly at random for the elements

POEt – generating itemsets

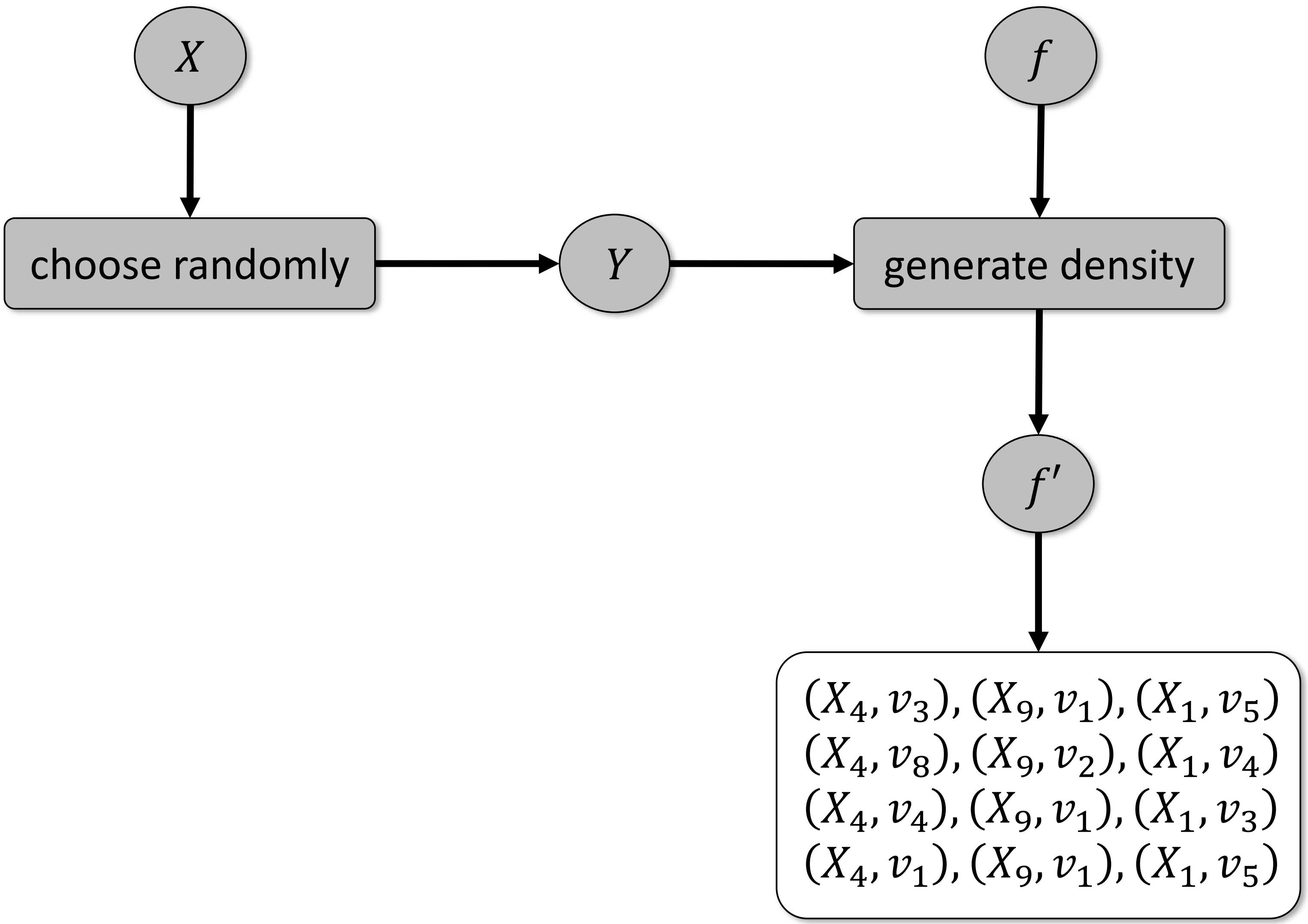


POEt – generating itemsets

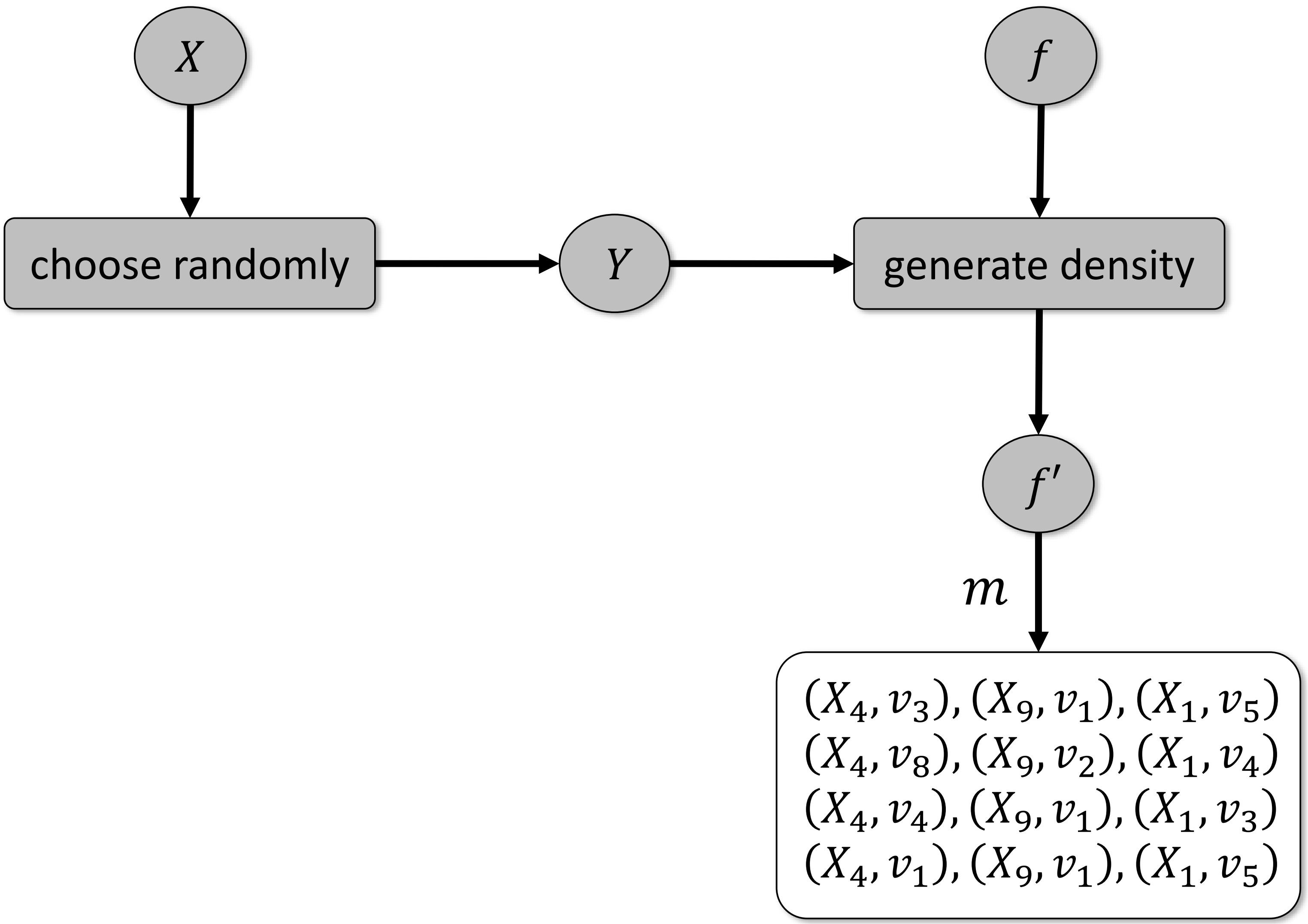


- Chernoff bounds
- train estimator

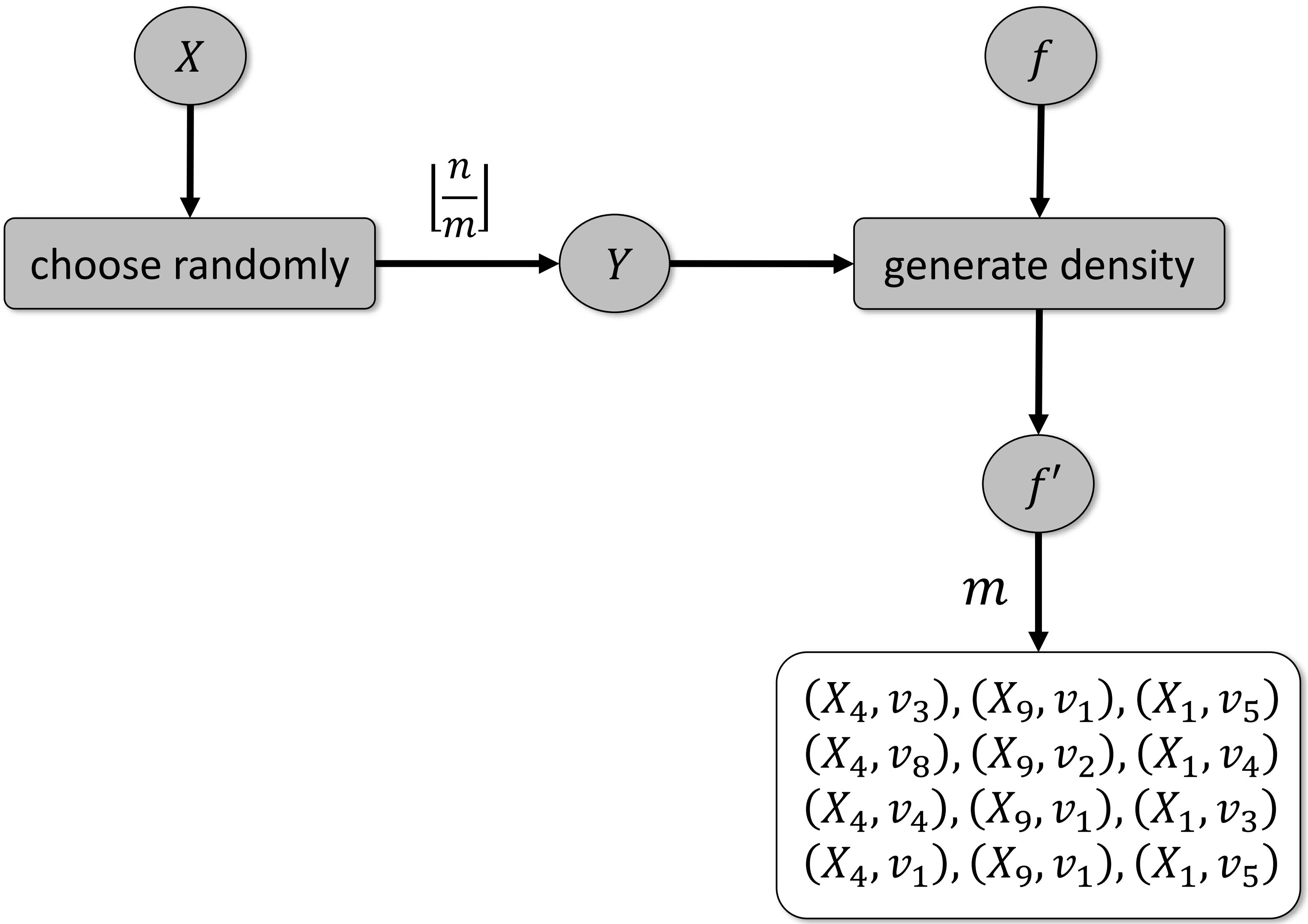
POEt – generating itemsets



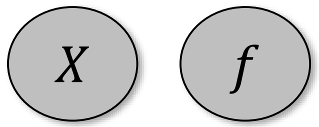
POEt – generating itemsets



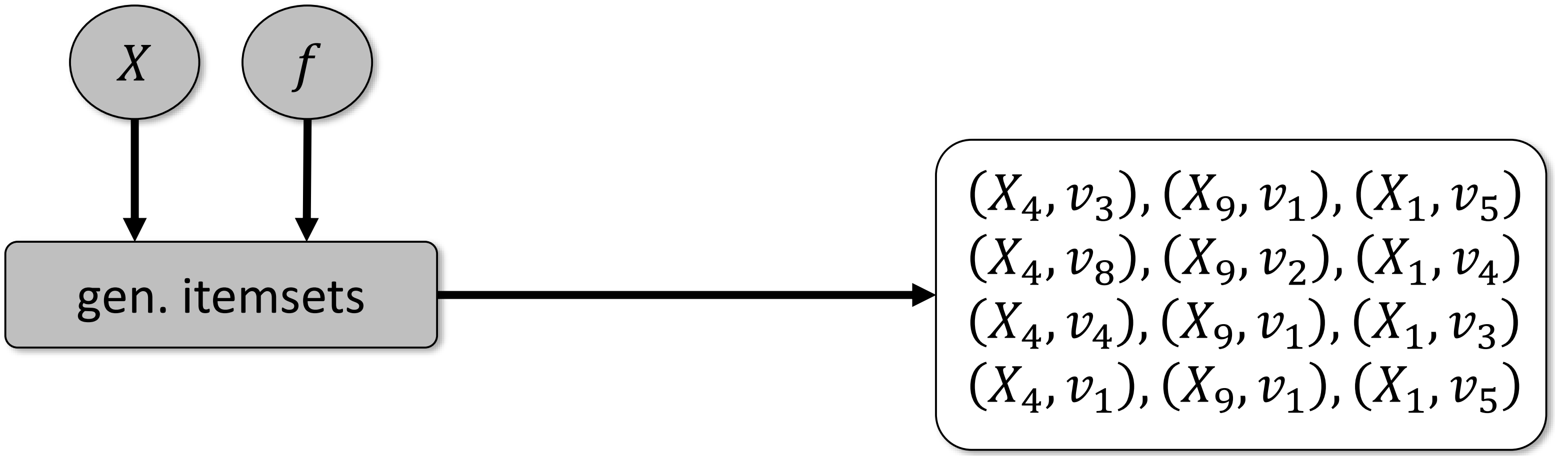
POEt – generating itemsets



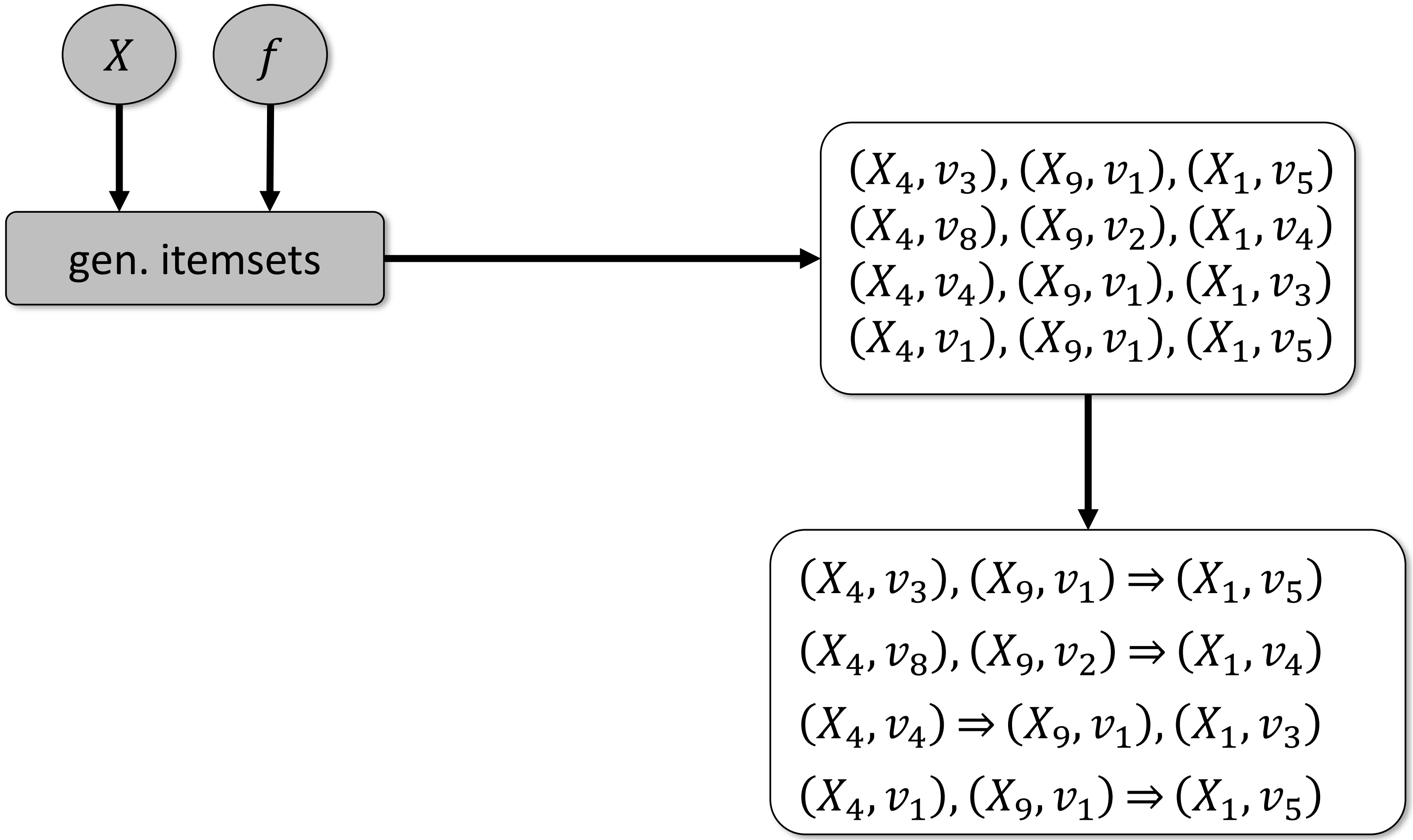
POEt – generating association rules



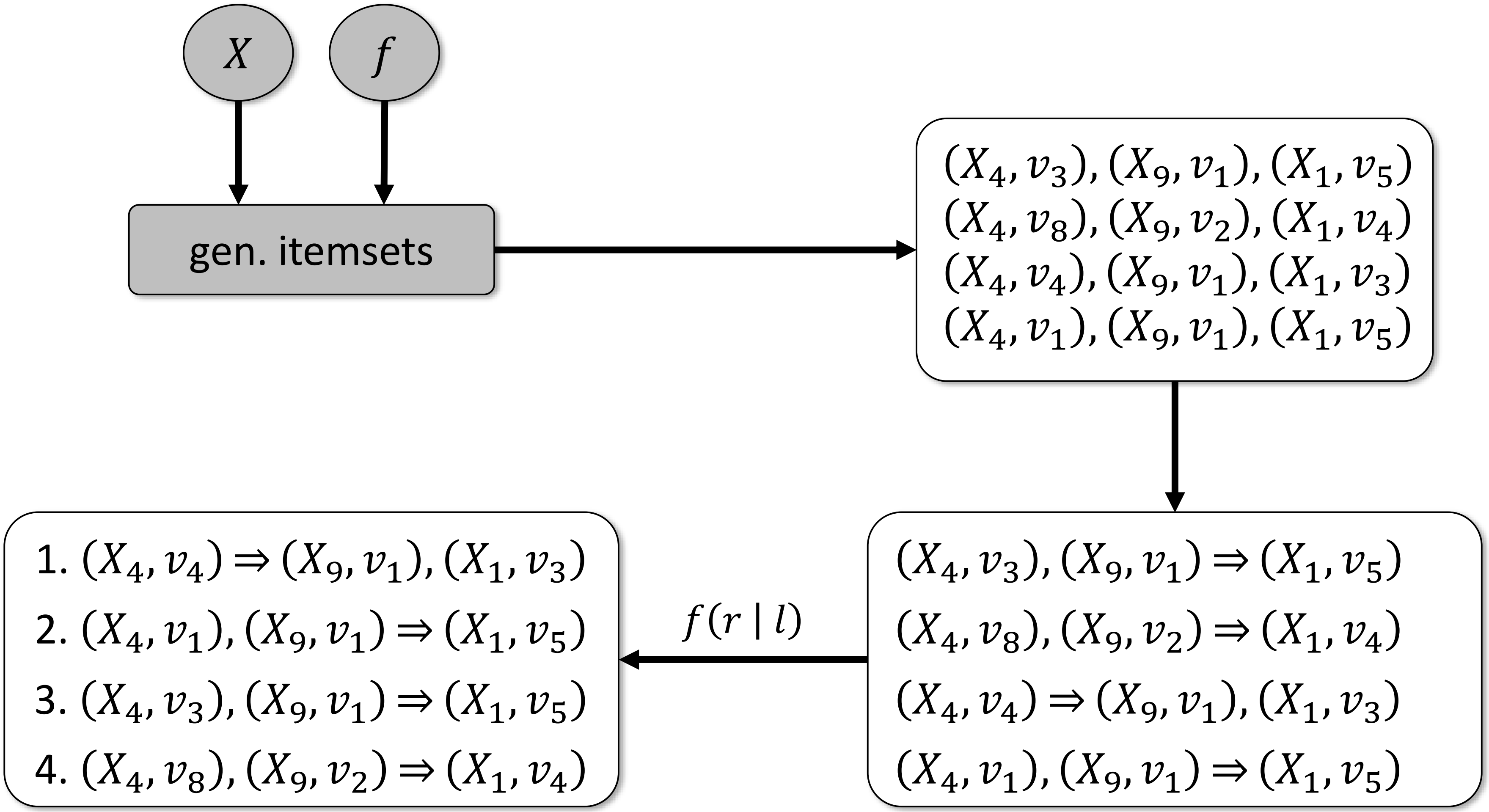
POEt – generating association rules



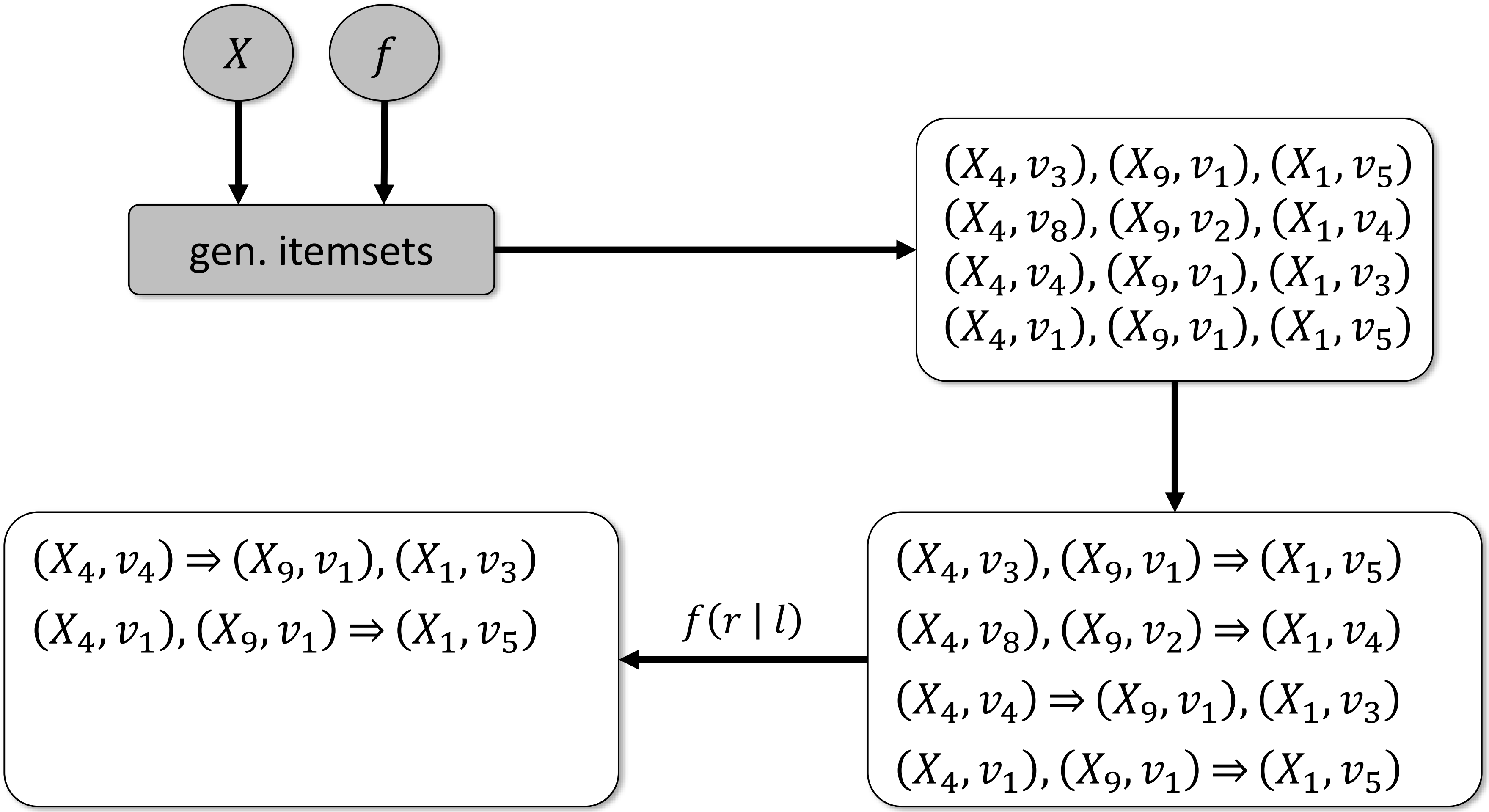
POEt – generating association rules



POEt – generating association rules



POEt – generating association rules



Evaluation

Datasets

Dataset	Instances	Attributes
IBM dataset generator	100,000	100
Bayesian networks	100,000	10
MovieLens	49,282	23

Compared to

- Apriori
- Moment

Performance measure

- percentaged overlap

$$\frac{|I_1 \cap \ I_2|}{|I_2|}$$

JG|U

Itemsets (1)

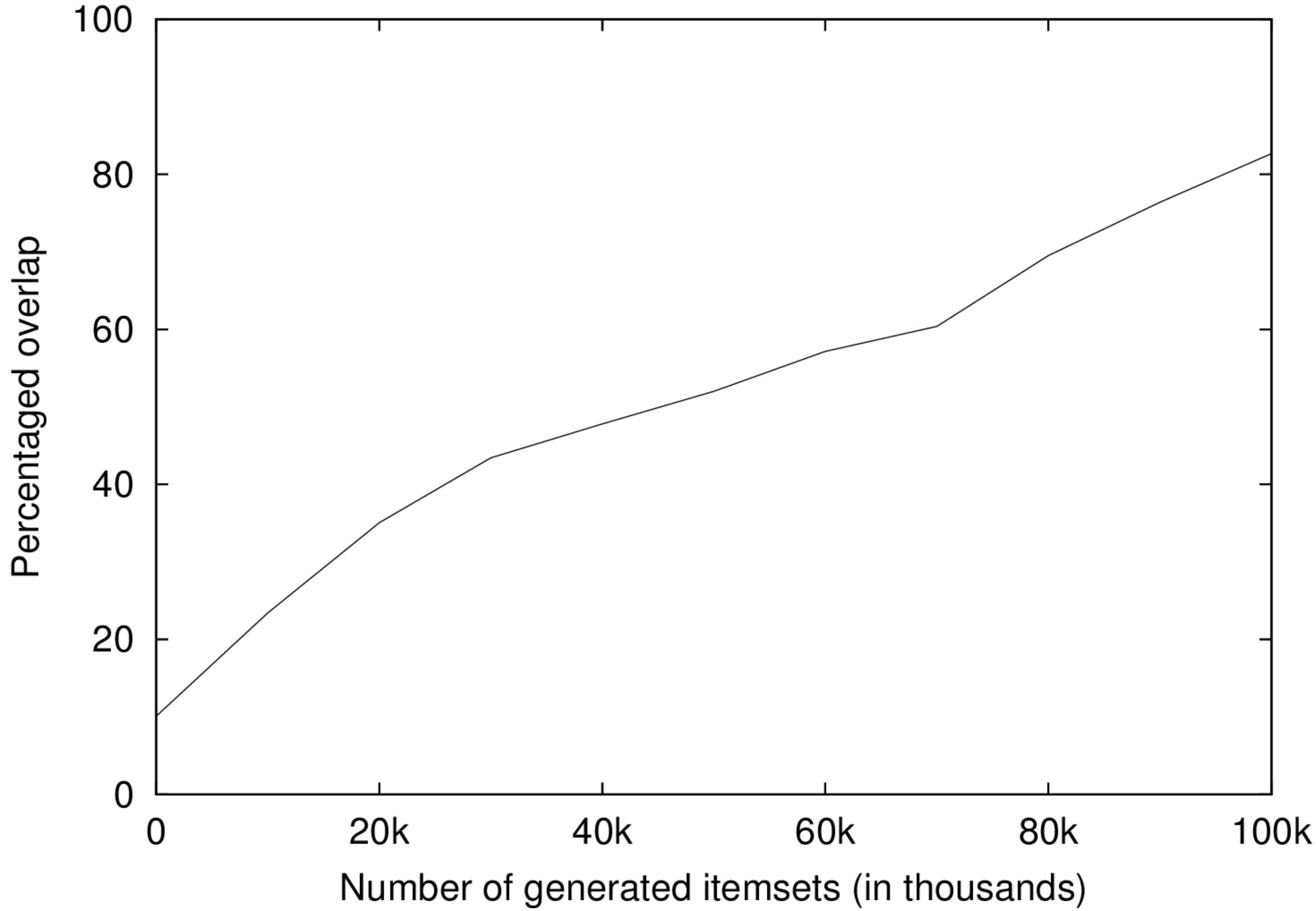
Dataset	Algorithm	Support		
		5%	10%	25%
IBM dataset generator	Apriori	0.002	0.002	0.006
	Moment	0.001	0.000	0.001
Bayesian networks	Apriori	0.384	0.487	0.524
	Moment	0.101	0.195	0.415
MovieLens	Apriori	0.133	0.111	0.333
	Moment	0.143	0.111	0.143

Evaluation - Itemsets (1)

Dataset	Algorithm	Support		
		5%	10%	25%
IBM dataset generator	Apriori	0.002	0.002	0.006
	Moment	0.001	0.000	0.001
Bayesian networks	Apriori	0.384	0.487	0.524
	Moment	0.101	0.195	0.415
MovieLens	Apriori	0.133	0.111	0.333
	Moment	0.143	0.111	0.143

$$(X_{gender}, male), (X_{thriller}, true), (X_{comedy}, false)$$

Evaluation - Itemsets (2)

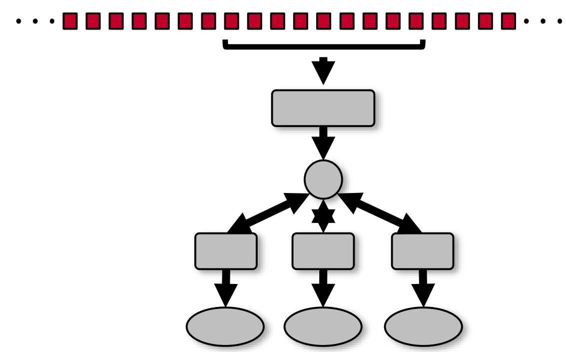


Evaluation - Association rules

Dataset	Confidence		
	0%	25%	50%
IBM dataset generator	0.000	0.000	0.000
Bayesian networks	0.389	0.345	0.210
MovieLens	0.098	0.093	0.100

Conclusions and Future Work

- framework for algorithms operating on density estimates
- a probabilistic condensed representation of data
- pattern mining on condensed representation



Future Work:

- more accurate itemset and association rule mining
- fast inference algorithm for speed-ups
- other algorithms that perform traditional data mining tasks on online density estimates

Thank you for your attention