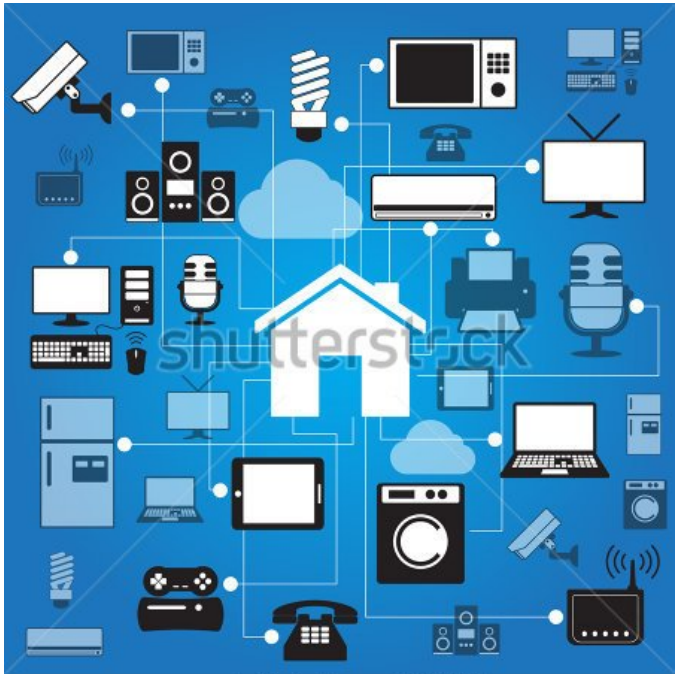
JOHANNES GUTENBERG
UNIVERSITÄT MAINZ

Privacy-Preserving Pattern Mining on Online Density Estimates

Michael Geilke and Stefan Kramer

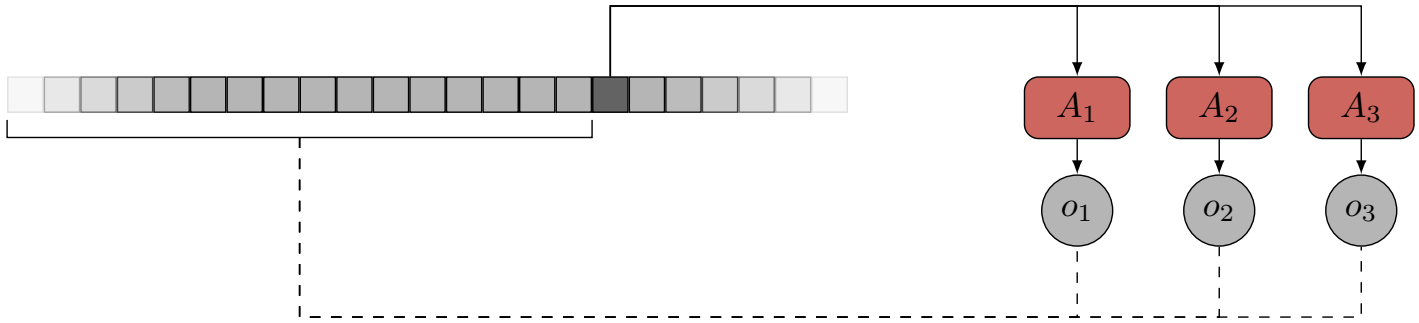
9 August 2017

Motivation

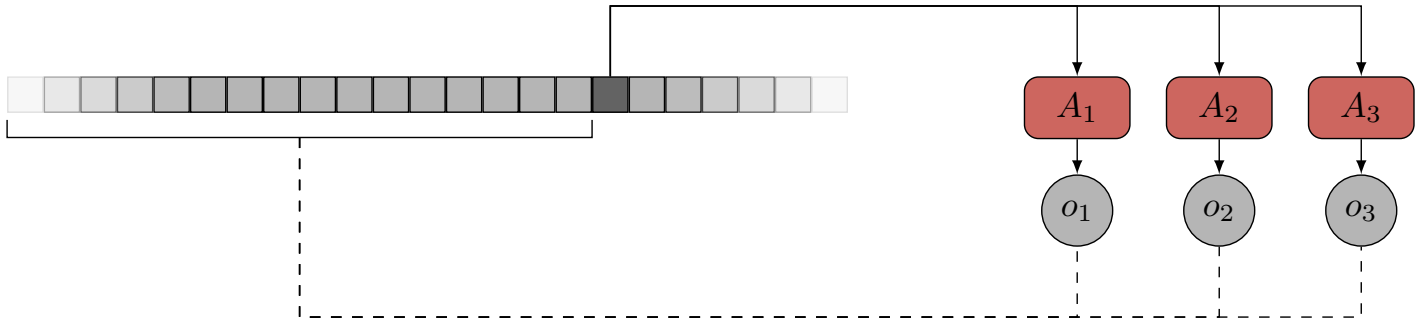


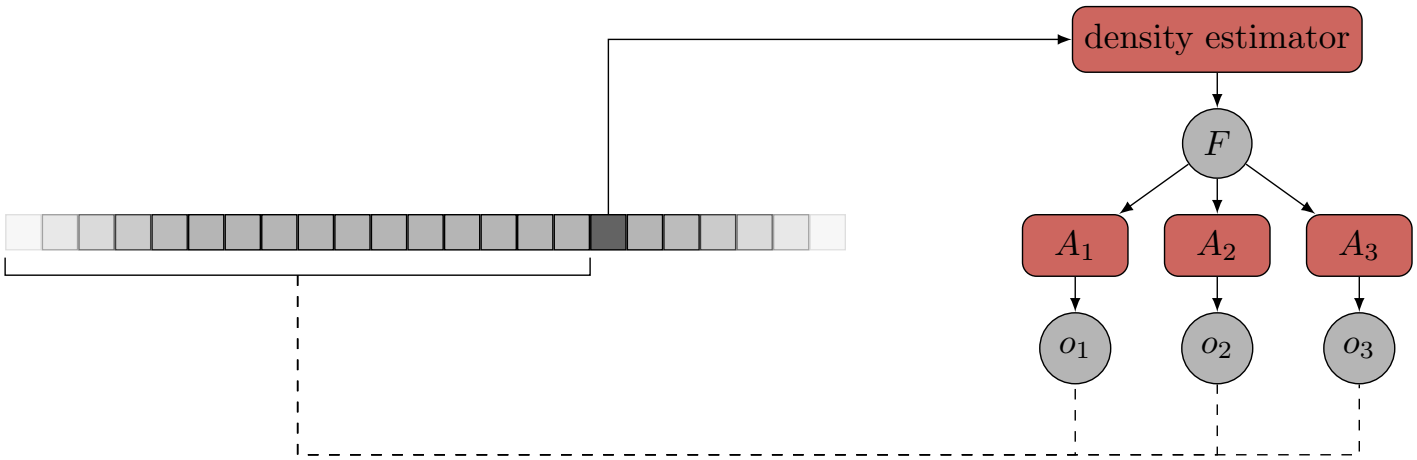
Source: www.shutterstock.com

Traditional Data Mining vs. MiDEO



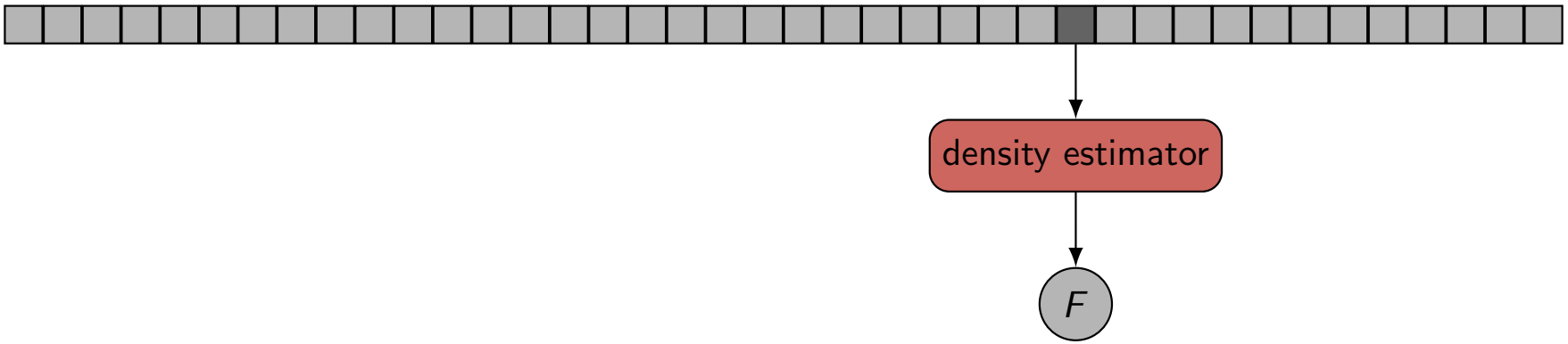
Traditional Data Mining vs. MiDEO



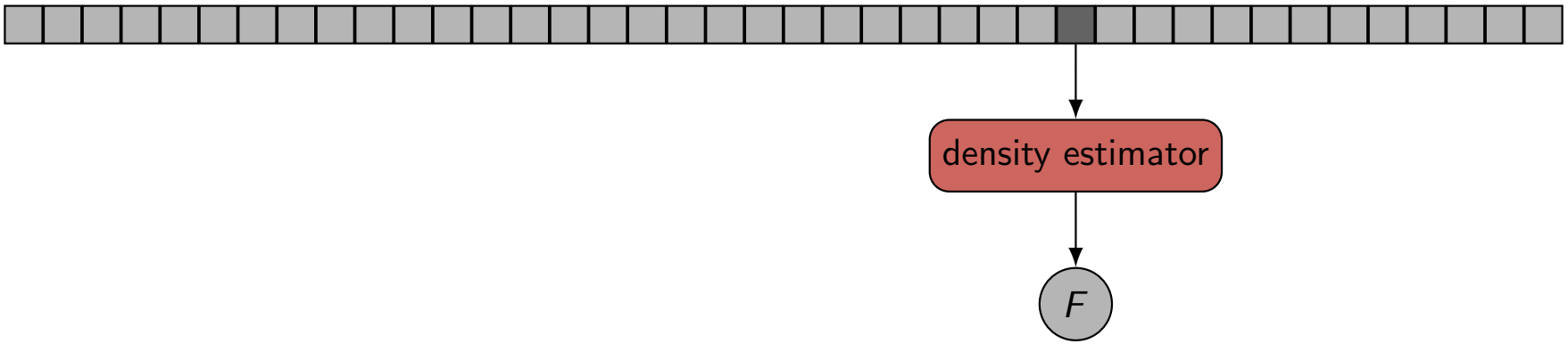


MiDEO

MiDEO

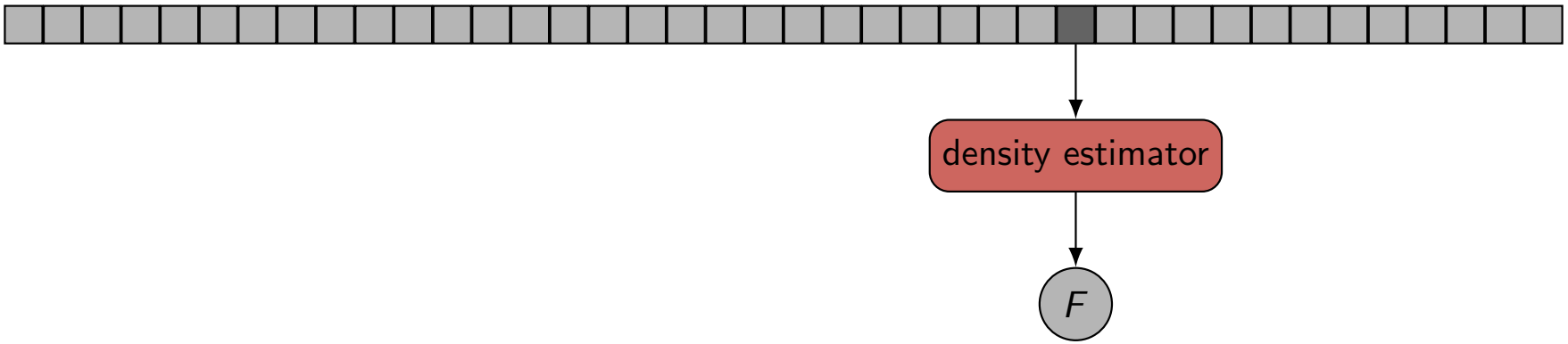


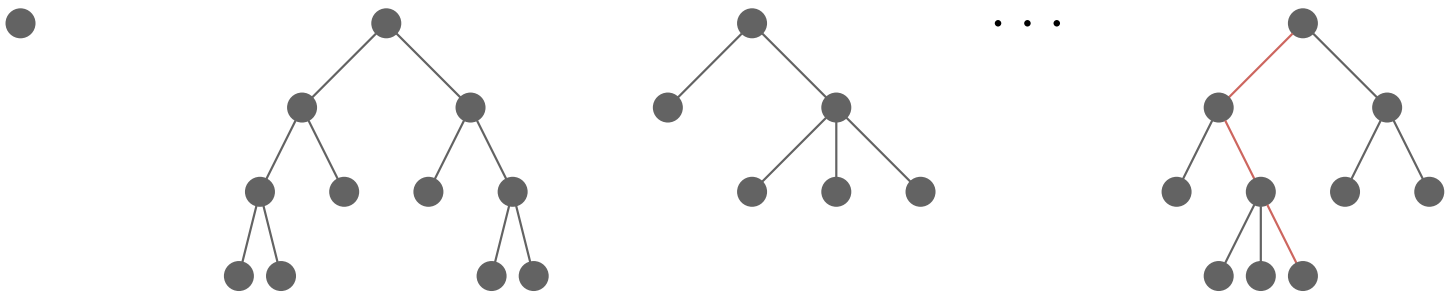
MiDEO



$$f(X_1, X_2, \ldots, X_m) = \prod_{i=1}^{m} f(X_i \mid X_1, \ldots, X_{i-1})$$

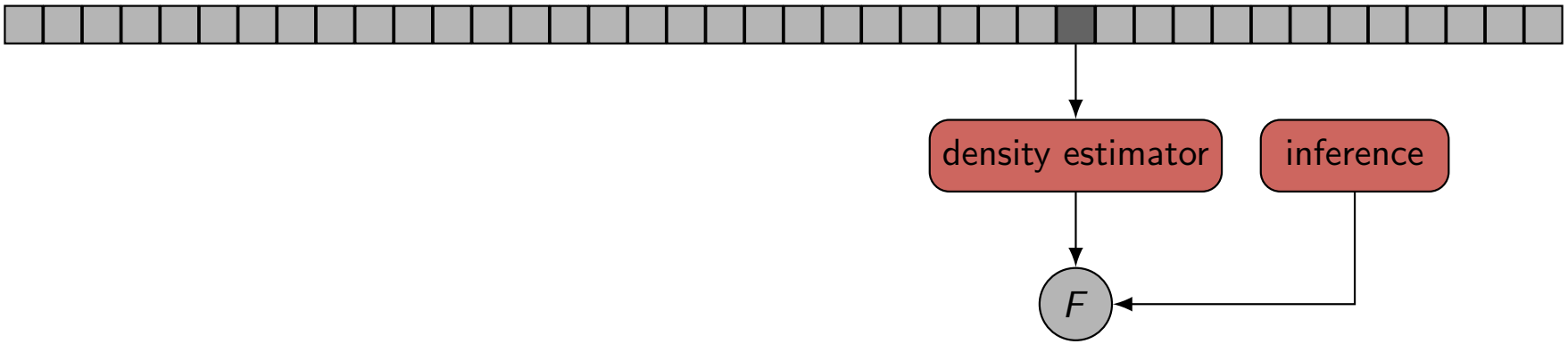
MiDEO



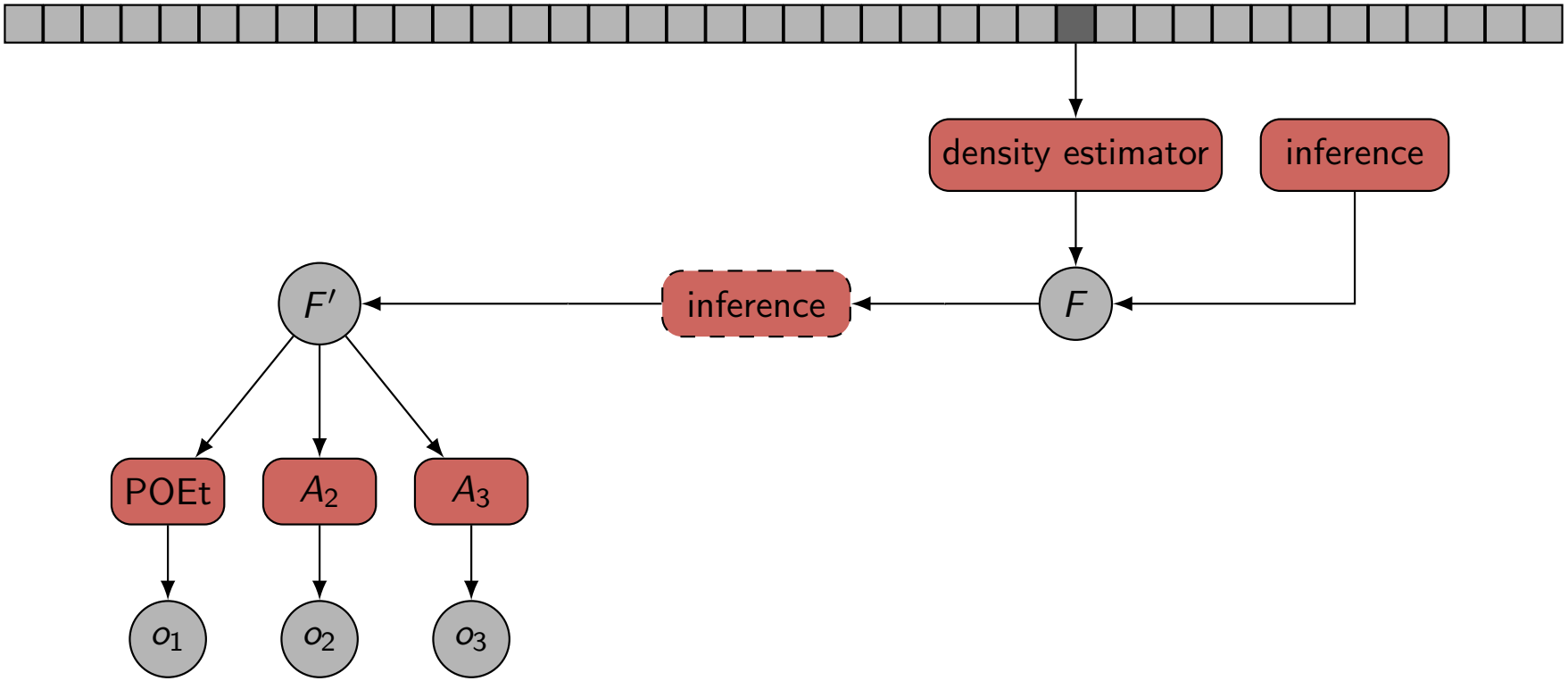


$$f(X_1, X_2, \ldots, X_m) = \prod_{i=1}^{m} f(X_i \mid X_1, \ldots, X_{i-1})$$

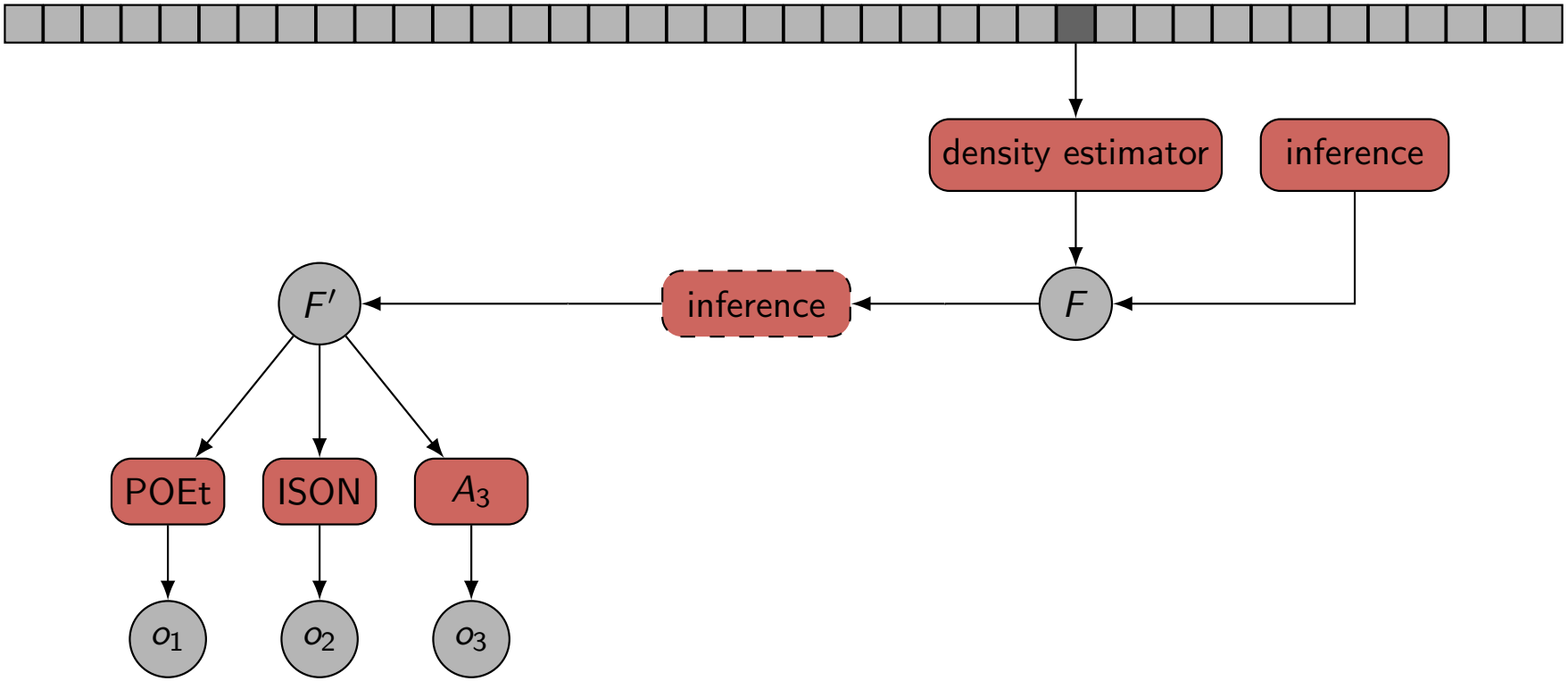
MiDEO



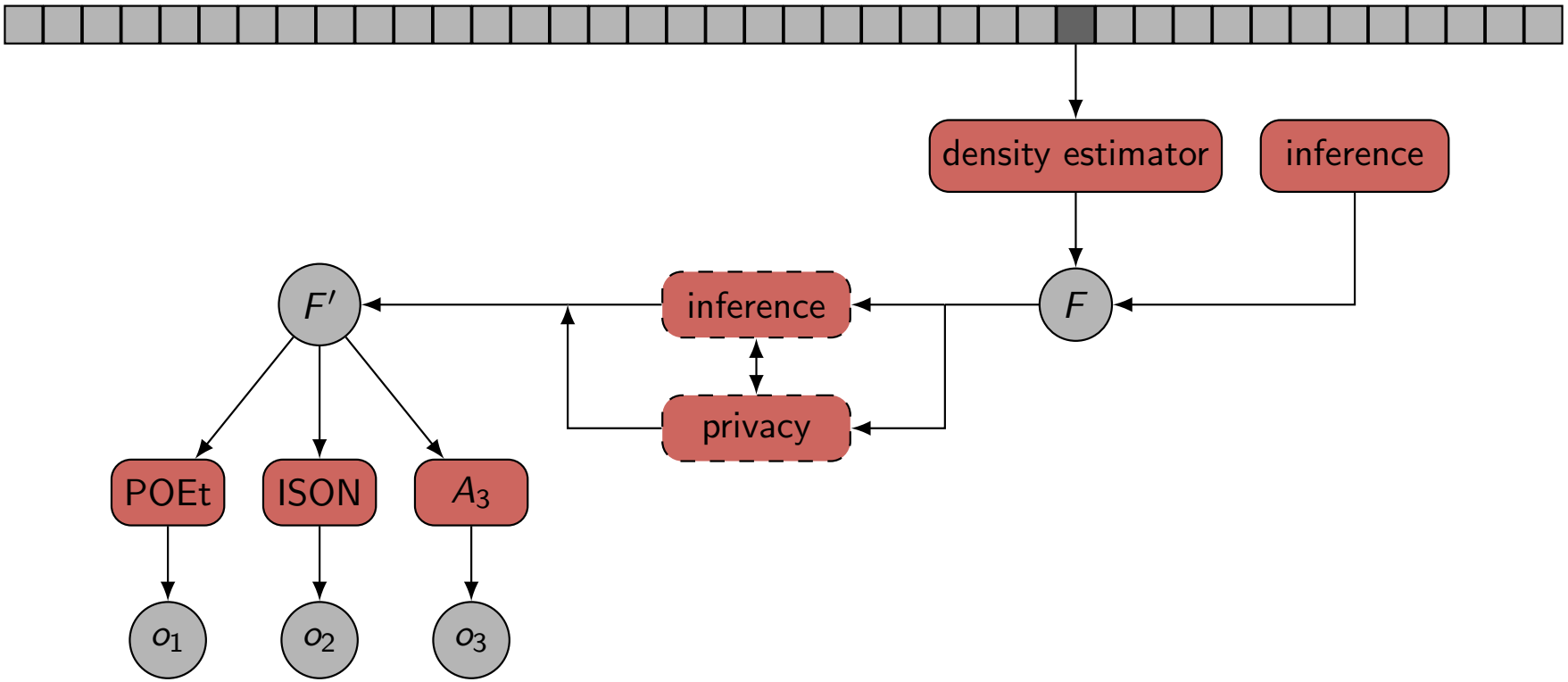
MiDEO


density estimator
inference
F
inference
F′
POEt
A2
A3
o1
o2
o3

MiDEO


density estimator
inference
F
inference
F′
POEt
ISON
A3
o1
o2
o3

MiDEO


density estimator
inference
F'
inference
F
privacy
POEt
ISON
A_3
o_1
o_2
o_3

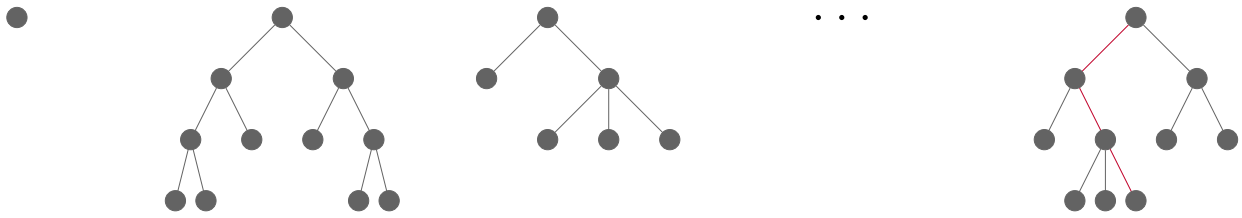
Problem Definition

Given:

- $\mathcal{X} := \{X_1, X_2, \ldots, X_m\}$
- $\vec{x}_i = \{(X_j, v_j) \mid 1 \leq j \leq m\}$
- $S := \vec{x}_1, \vec{x}_2, \ldots$
- an itemset is a subset $I \subseteq \vec{x}$
- $freq(I)$ is the relative frequency of I in S

Goal: Determine $\mathcal{F}_S = \{I \subseteq \vec{x} \mid \vec{x} \in S \wedge freq(I) \geq \theta\}$, while also preserving privacy-preserving properties such as t-closeness.

ISON



Iteration 0

$C' \leftarrow \emptyset$

for $ht \in \hat{f}_{cc}$ *and* $\hat{f}_{cc} \in \hat{f}_{ecc}$ **do**

  $\mathcal{P} \leftarrow \{node_1, \ldots, node_l \in ht \mid node_1.isRoot() \wedge node_l.isLeaf()\}$

  for $path \in \mathcal{P}$ **do**

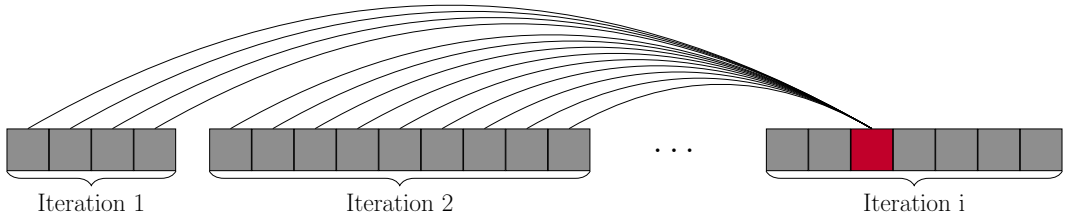
    $c \leftarrow$ turn *path* into an itemset

    `// iterate over itemsets by size and prune candidates using filter`

    $it \leftarrow subsets(c, \textbf{pruning} = \textbf{true})$

    $C' \leftarrow C' \cup \{s \mid s \in it\}$

ISON



Iteration ≥ 1

while $|C'| > 0$ **do**
  $C_{curr} \leftarrow C'$, $C' \leftarrow \emptyset$
  if $|iterations| = 1$ **then**
    $mergePartners \leftarrow iterations[1]$
  else
    $mergePartners \leftarrow iterations[1 : i]$
  for $c_1 \in mergePartners$ *and* $c_2 \in C_{curr}$ **do**
    if $filter(merge(c_1, c_2))$ **then**
      $C' \leftarrow C' \cup \{merge(c_1, c_2)\}$
  $iterations.append(C')$

Protecting Individual Entities (1)

t-closeness:

Protecting Individual Entities (1)

t-closeness:

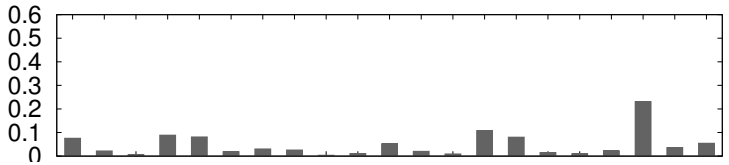
- a set of sensitive attributes $\mathcal{Q}$
  e.g., $\{age, gender, occupation\}$

Protecting Individual Entities (1)

t-closeness:

- a set of sensitive attributes $\mathcal{Q}$
  e.g., $\{age, gender, occupation\}$
- g: global distribution

global:

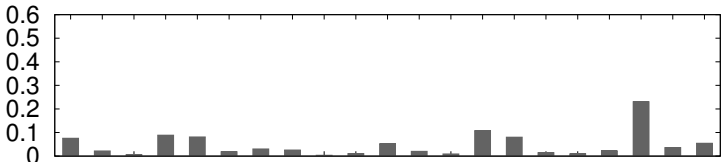


Protecting Individual Entities (1)

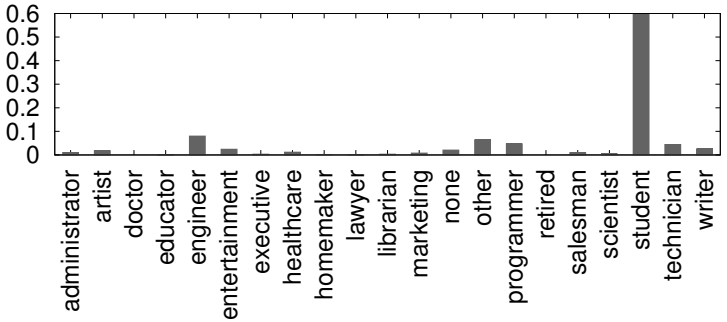
t-closeness:

- a set of sensitive attributes $\mathcal{Q}$
  e.g., $\{age, gender, occupation\}$
- g: global distribution
- d: local distribution of a $Q \in \mathcal{Q}$

global:



age 0-24:

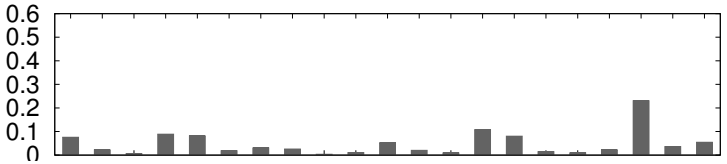


Protecting Individual Entities (1)

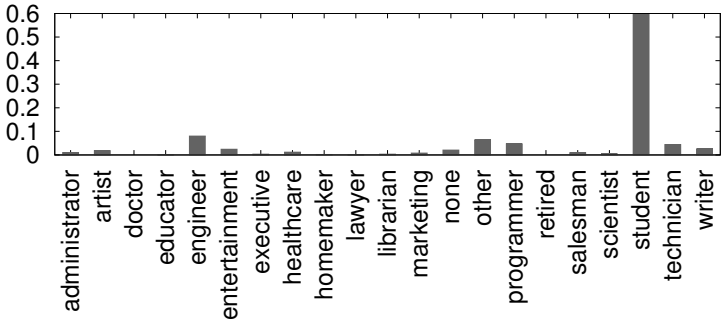
t-closeness:

- a set of sensitive attributes $\mathcal{Q}$
  e.g., $\{age, gender, occupation\}$
- g: global distribution
- d: local distribution of a $Q \in \mathcal{Q}$
- δ: maximally allowed deviation

global:



age 0-24:

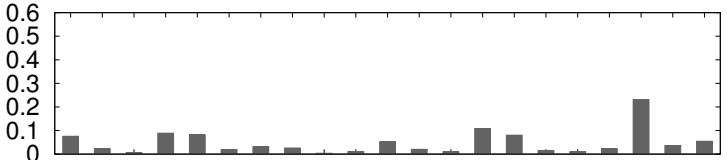


Protecting Individual Entities (1)

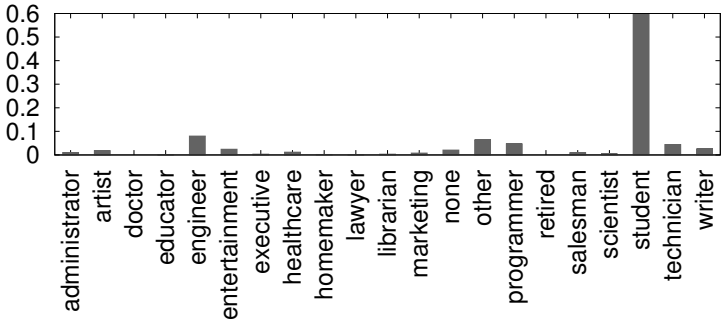
t-closeness:

- a set of sensitive attributes $\mathcal{Q}$
  e.g., $\{age, gender, occupation\}$
- g: global distribution
- d: local distribution of a $Q \in \mathcal{Q}$
- δ: maximally allowed deviation
- $\|g - d\| \leq \delta$

global:



age 0-24:



Evaluation (1)

Dataset	Shortcut	Purpose	#Variables	#Instances
$\text{IBM}_{AP\in\{2,4\}}$	ibm01	general	10	50,000
	ibm02			500,000
	ibm03			5,000,000
$\text{IBM}_{AP=4}$		running time	10	10,000,000
				20,000,000
				30,000,000
				40,000,000
				50,000,000
pokerhand	poker	discovery	11	1,025,015
skin			4	225,009
movielens-01	mov01		9	49,282
movielens-02	mov02		16	49,282
movielens-03	mov03		23	49,282

Evaluation (2)

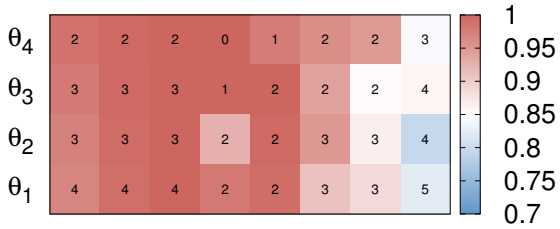
Algorithms:

- POEt
- Apriori
- Moment (window sizes: $10^2, 10^3, 10^4, 4 \cdot 10^4$)

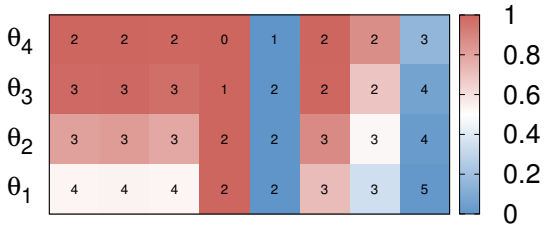
Performance Measure:

- $TPR = \frac{|\mathcal{F}_{ISON}(S) \subseteq \mathcal{F}_{Apriori}(S)|}{|\mathcal{F}_{Apriori}(S)|}$
- $FPR = \frac{|\{I \in \mathcal{F}_{ISON}(S) | I \notin \mathcal{F}_{Apriori}(S)\}|}{|\mathcal{F}_{Apriori}(S)|}$

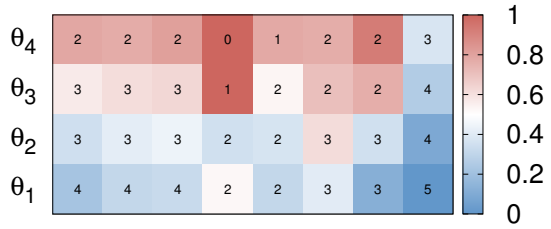
Discovery of Frequent Itemset



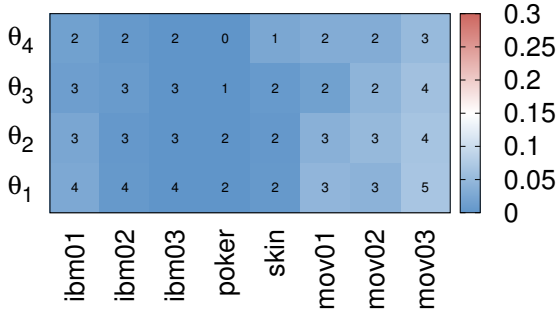
(a) ISON vs Apriori



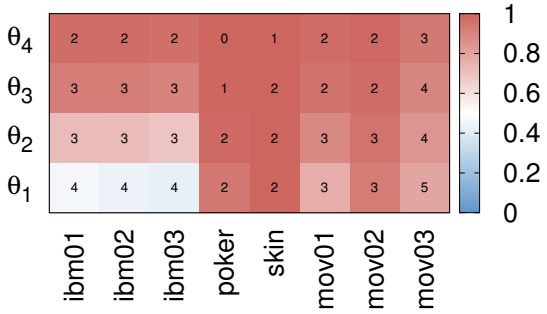
(b) POEt vs Apriori



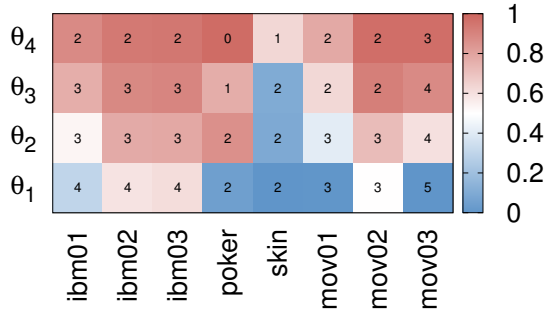
(c) Moment vs Apriori



(d) ISON vs Apriori

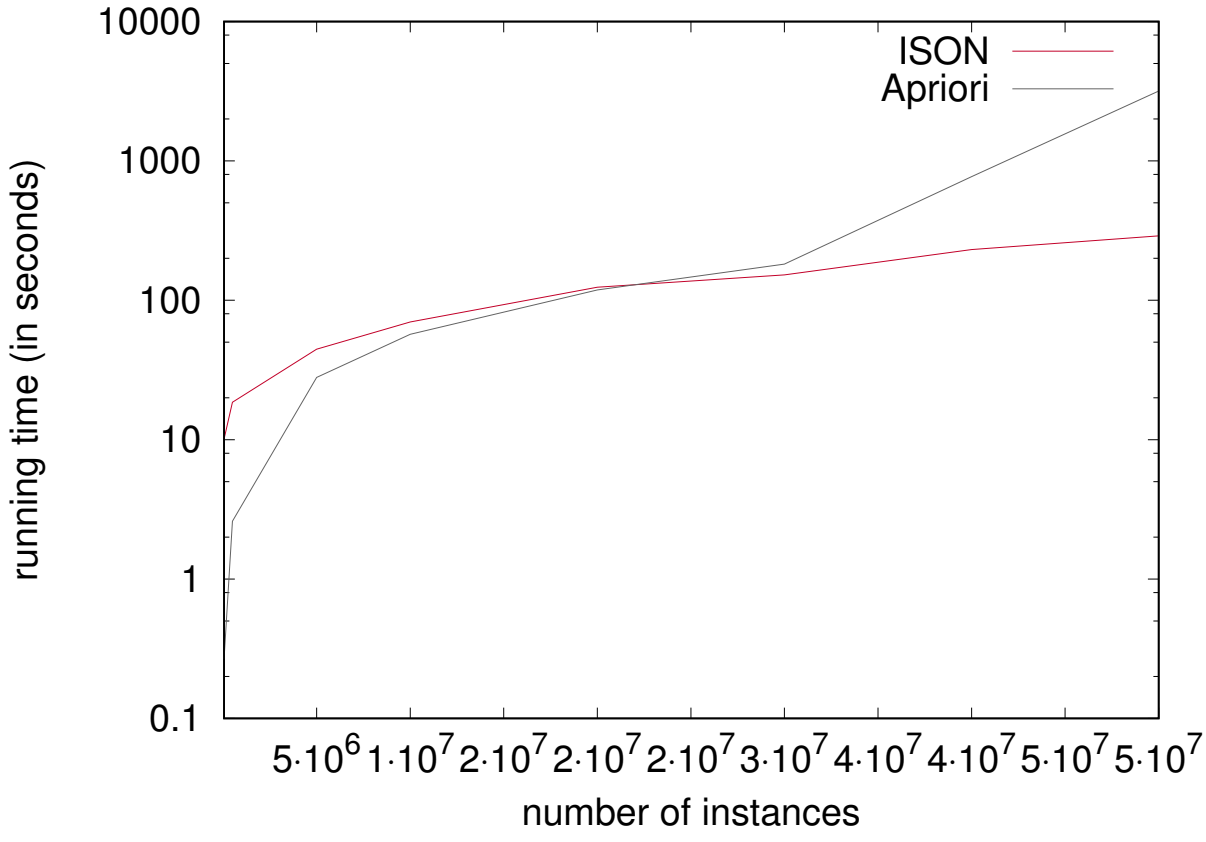


(e) POEt vs Apriori

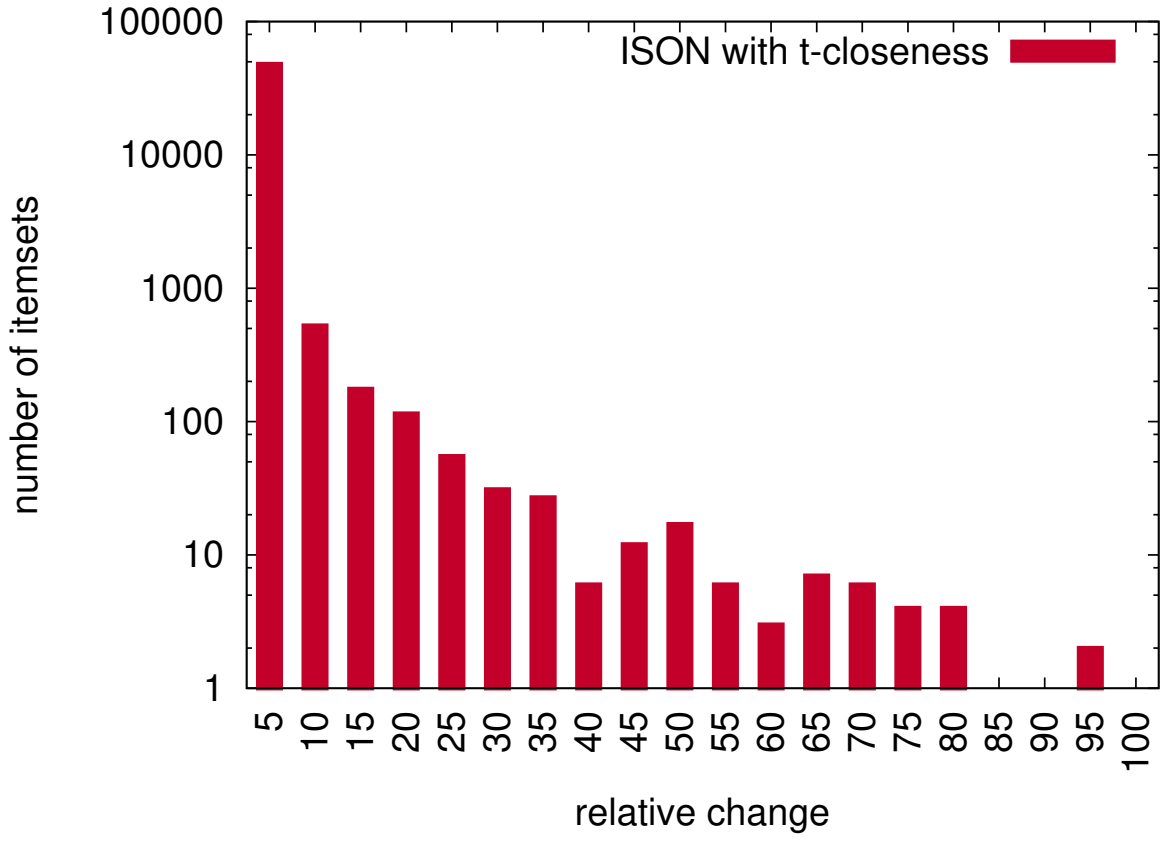


(f) Moment vs Apriori

Running Times



Effect of t-closeness Property



Conclusion

- ISON extracts frequent itemsets from the structure of a probabilistic condensed representation
- proof of correctness
- good performance on datasets with a high combination ratio
- t-closeness can be enforced by modifying the density estimates

Future Work:

- other types of itemsets
- extension to numeric attributes
- other frequency-related constraints such as class-correlations

Thank you for your attention!