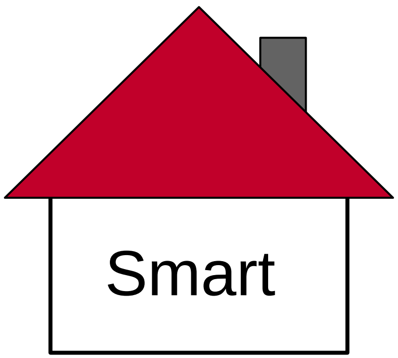
JOHANNES GUTENBERG UNIVERSITÄT MAINZ

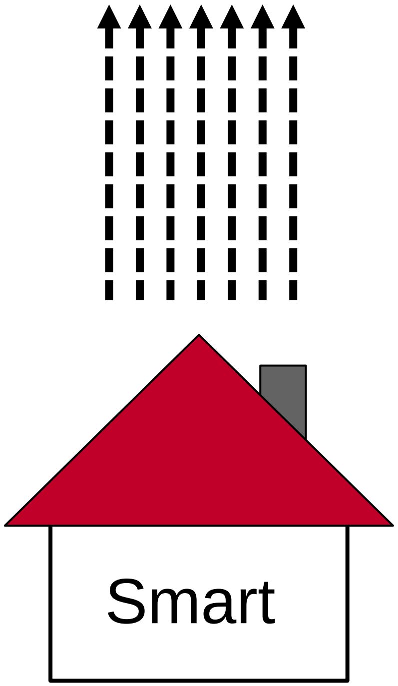
Modeling Recurrent Distributions in Streams using Possible Worlds

Michael Geilke, Andreas Karwath, and Stefan Kramer

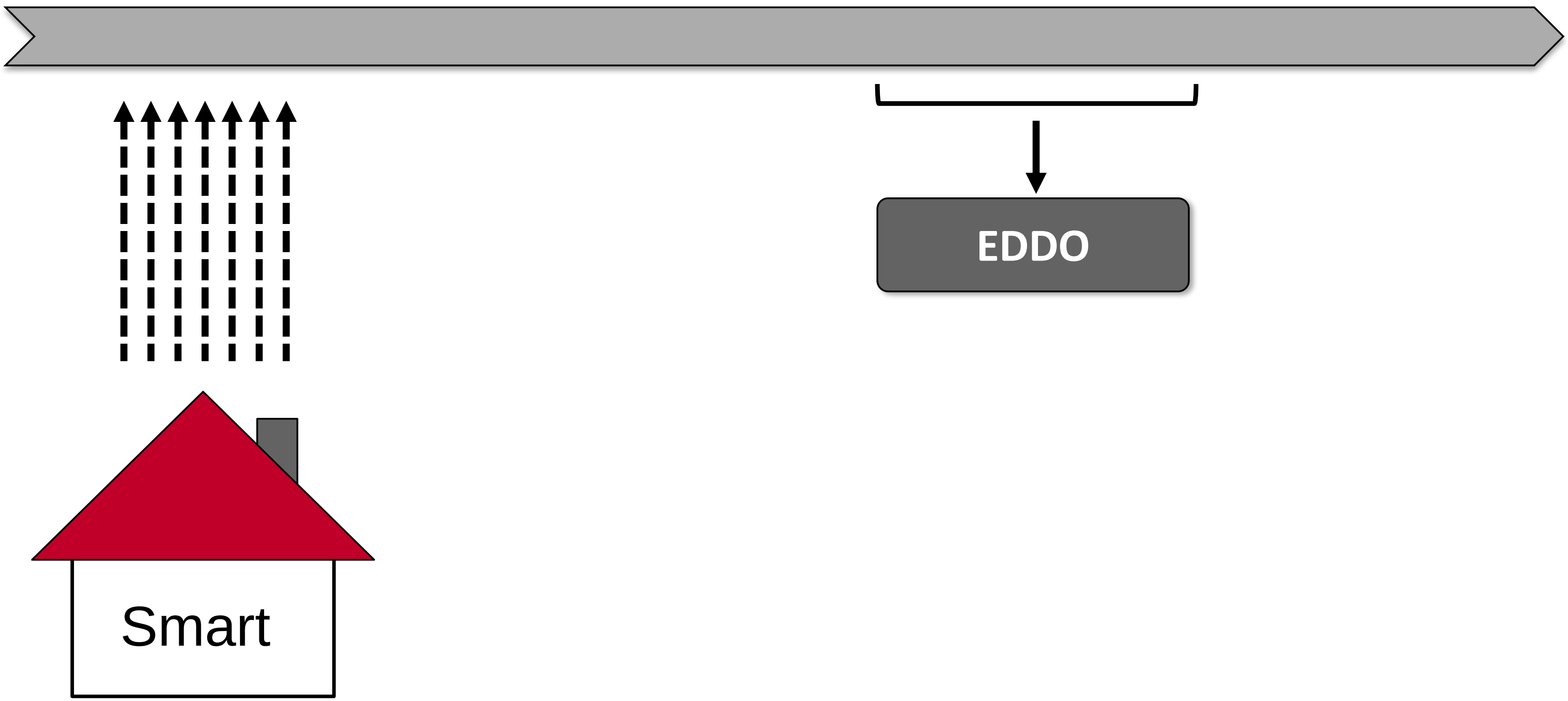
Johannes Gutenberg-Universität Mainz, Germany

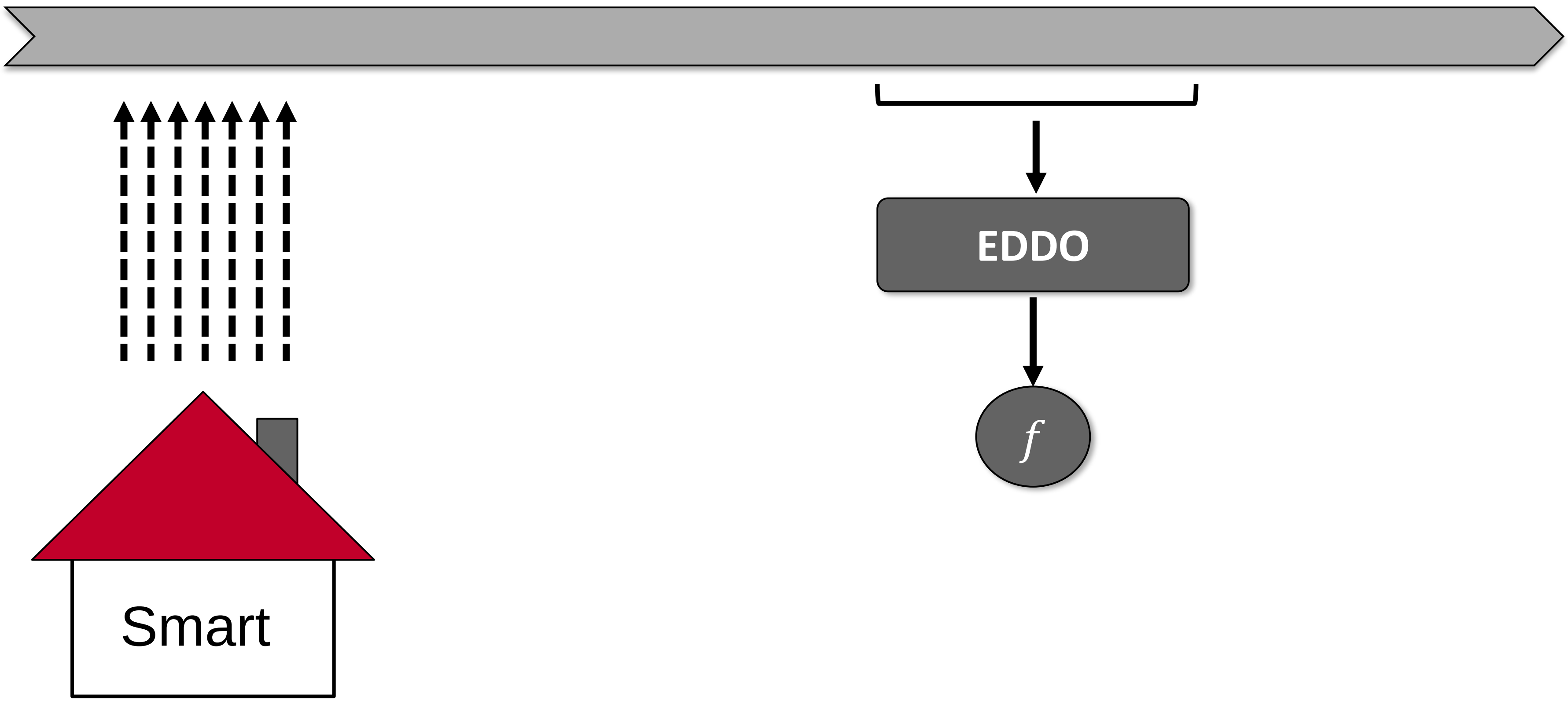
October 20, 2015

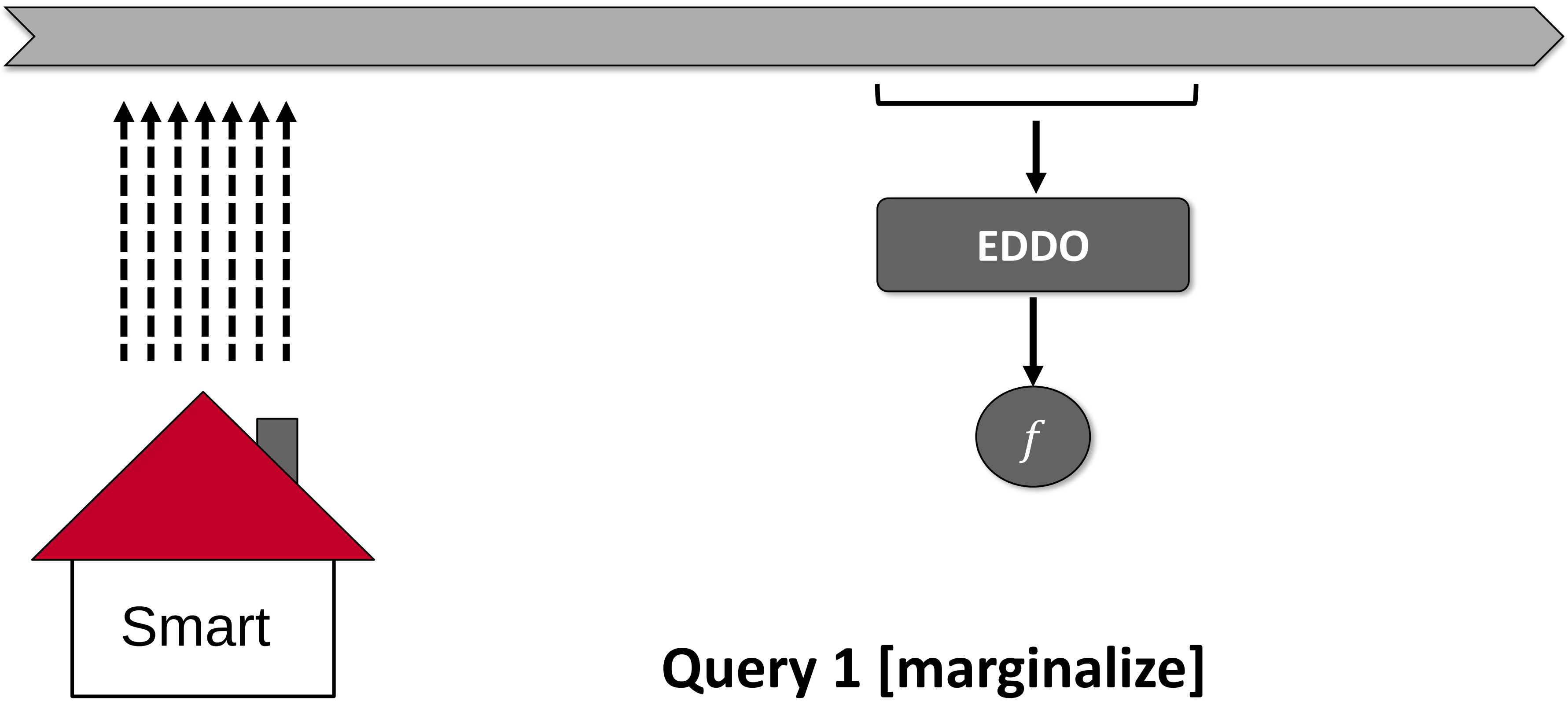

Smart


Smart


Smart

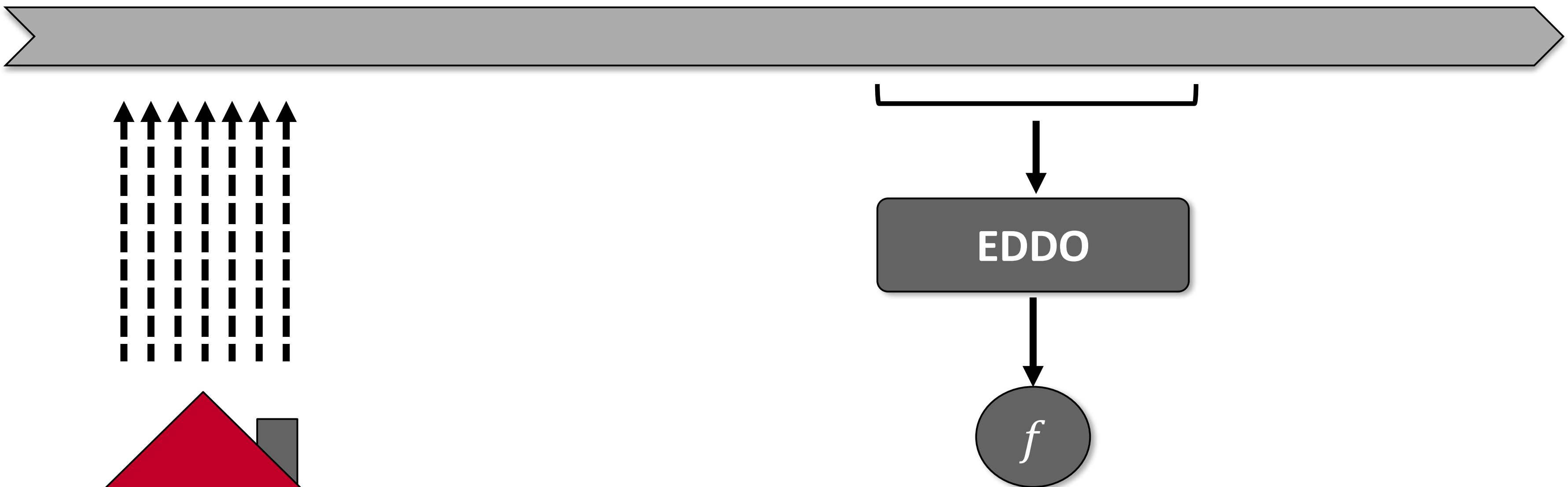

EDDO
Smart

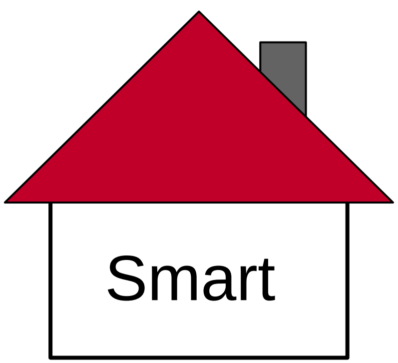

Smart
EDDO
f



Query 1 [marginalize]

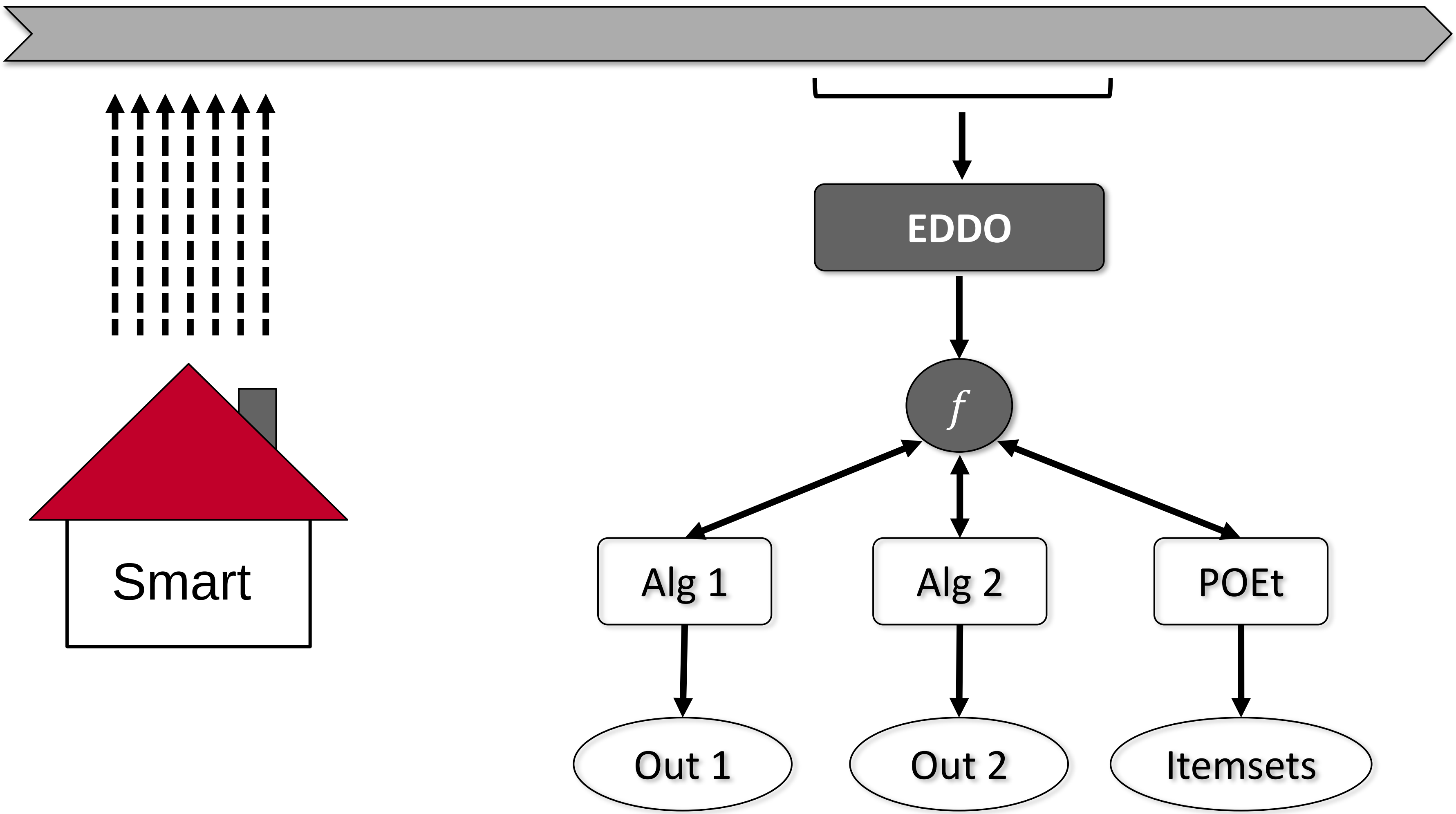
What is the data distribution of the sensors in the living room?



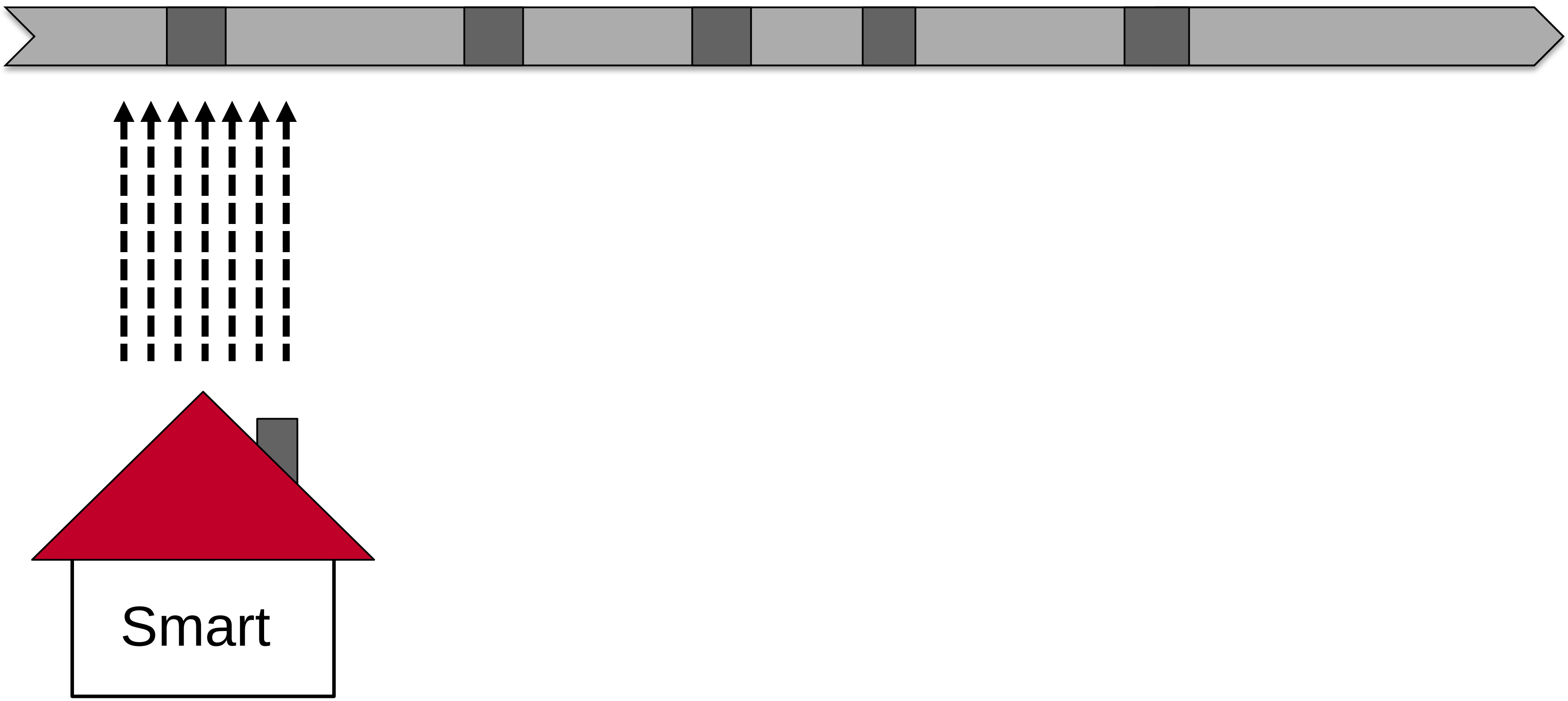


Query 2 [setting hard evidence]

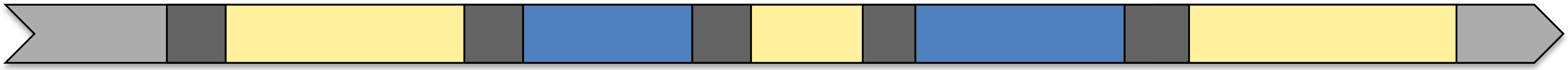
Two residents are in the living room.
What is the probability that they watch TV?

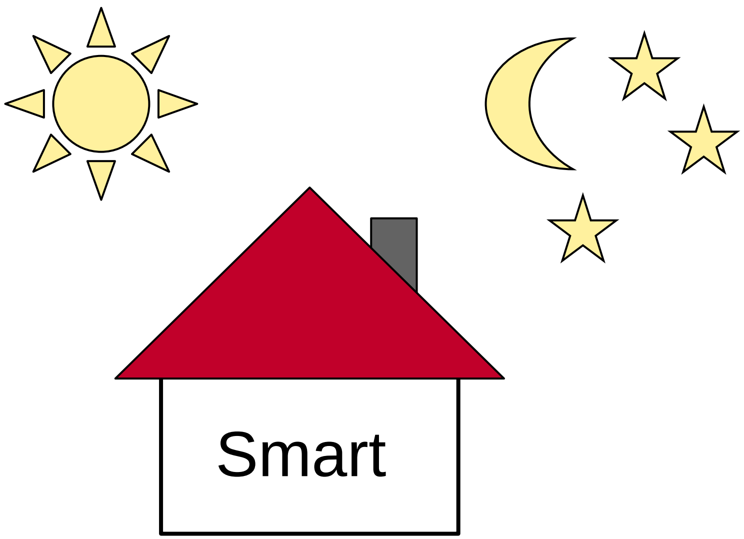

EDDO
f
Smart
Alg 1
Alg 2
POEt
Out 1
Out 2
Itemsets


Smart


Smart

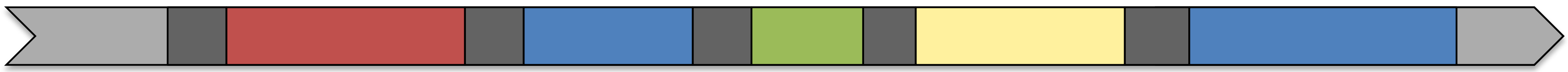

EDDO
f
Smart

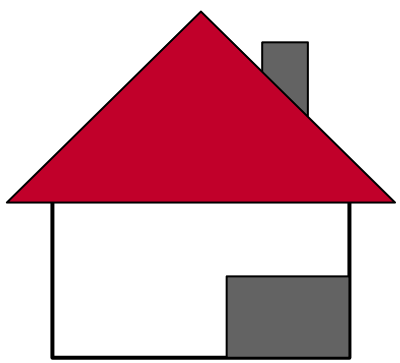




Recurrences

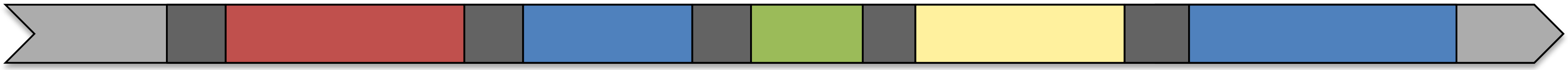
- day and night
- working days and weekends
- seasons





Recurrences

- pattern could be more complex
- may only affect a part of the house



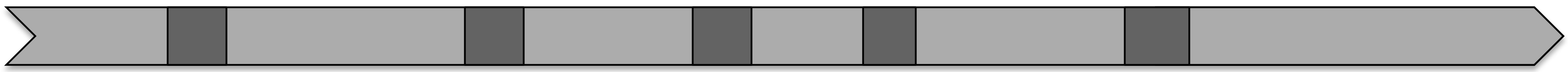
Goal: a representation that

- is constantly updated
- is representing current and historical data distributions,
- is able to represent recurrences
- provides a query mechanism

Tasks for proposed method

1. recognize regions of drift
2. represent density of data stream segments
3. identify recurrences on the density level
4. identify recurrences between parts of different densities

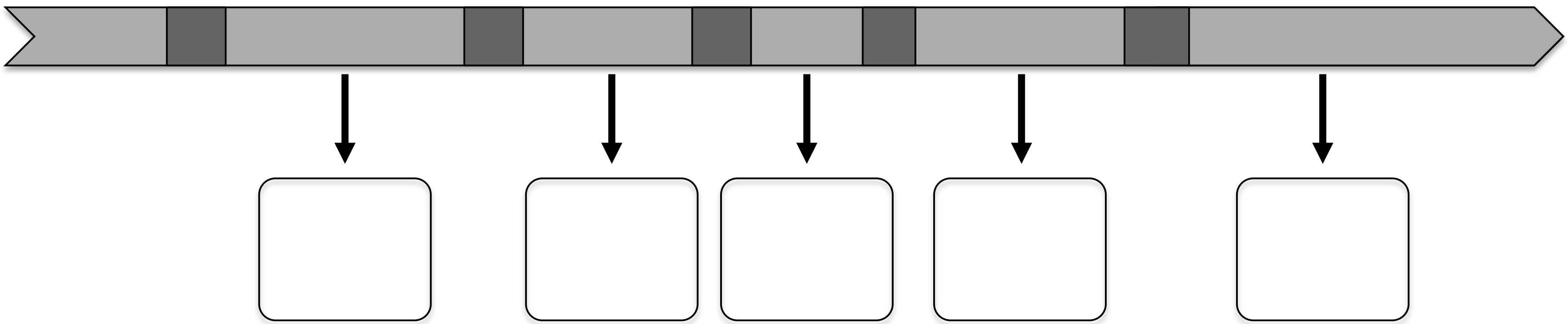
do all of that in an online fashion



Tasks for proposed method

1. recognize regions of drift
2. represent density of data stream segments
3. identify recurrences on the density level
4. identify recurrences between parts of different densities

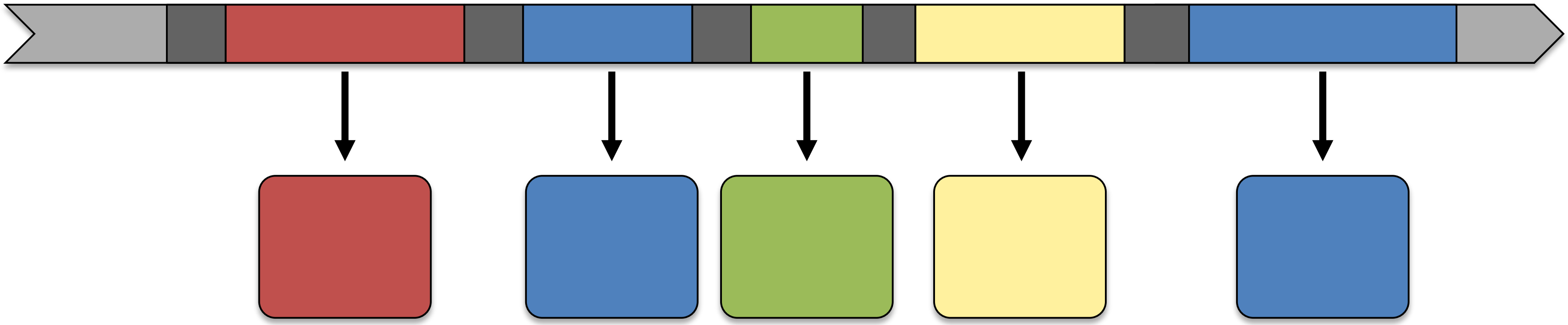
do all of that in an online fashion



Tasks for proposed method

1. recognize regions of drift
2. represent density of data stream segments
3. identify recurrences on the density level
4. identify recurrences between parts of different densities

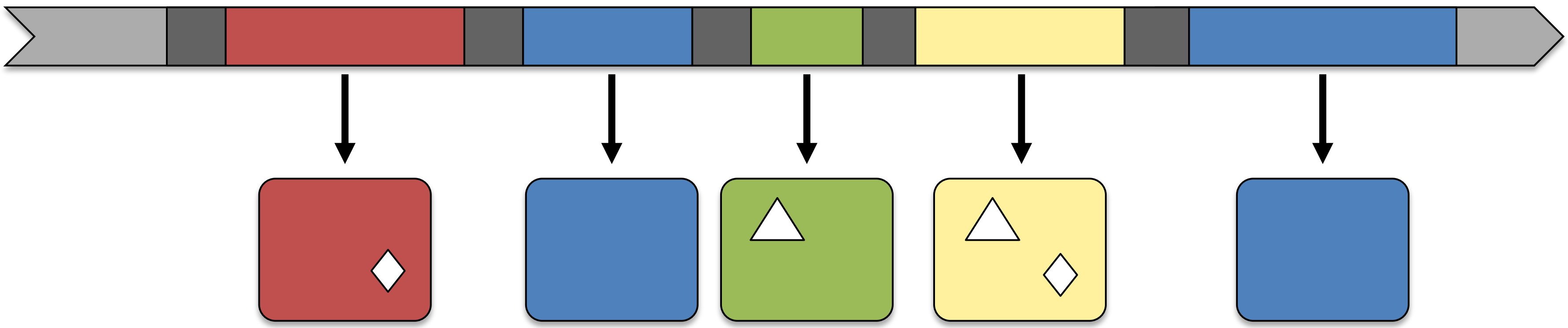
do all of that in an online fashion



Tasks for proposed method

1. recognize regions of drift
2. represent density of data stream segments
3. identify recurrences on the density level
4. identify recurrences between parts of different densities

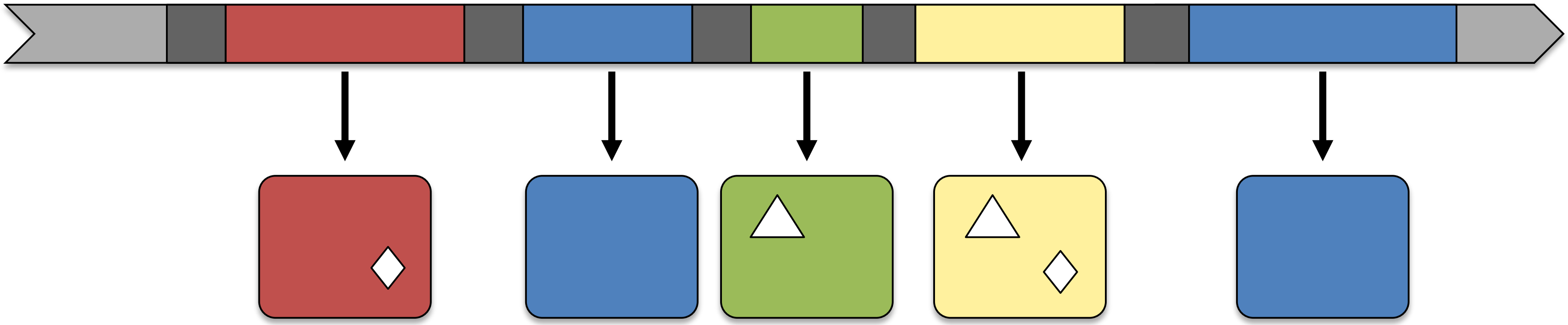
do all of that in an online fashion



Tasks for proposed method

1. recognize regions of drift
2. represent density of data stream segments
3. identify recurrences on the density level
4. identify recurrences between parts of different densities

do all of that in an online fashion

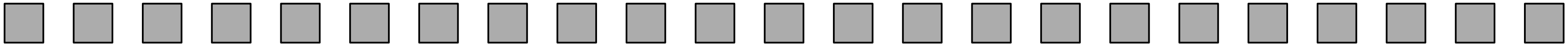


Tasks for proposed method

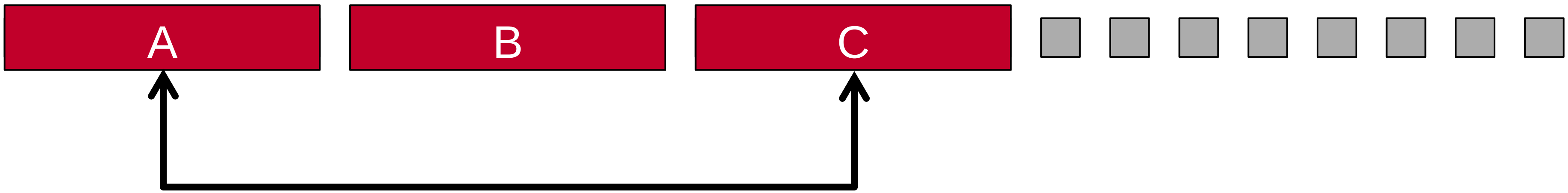
1. recognize regions of drift
2. represent density of data stream segments
3. identify recurrences on the density level
4. identify recurrences between parts of different densities

do all of that in an online fashion

Recognize Regions of Drift



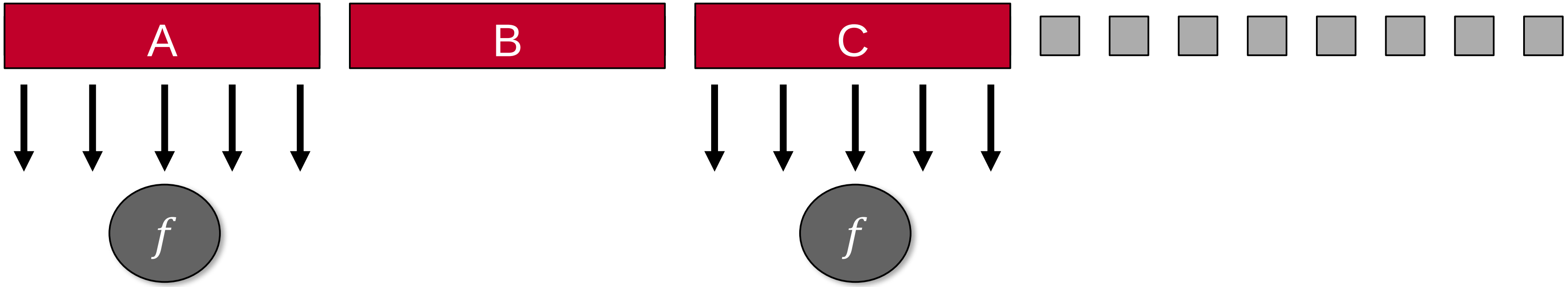
Recognize Regions of Drift



Window-based approach

- extension of an approach by Dries and Rückert

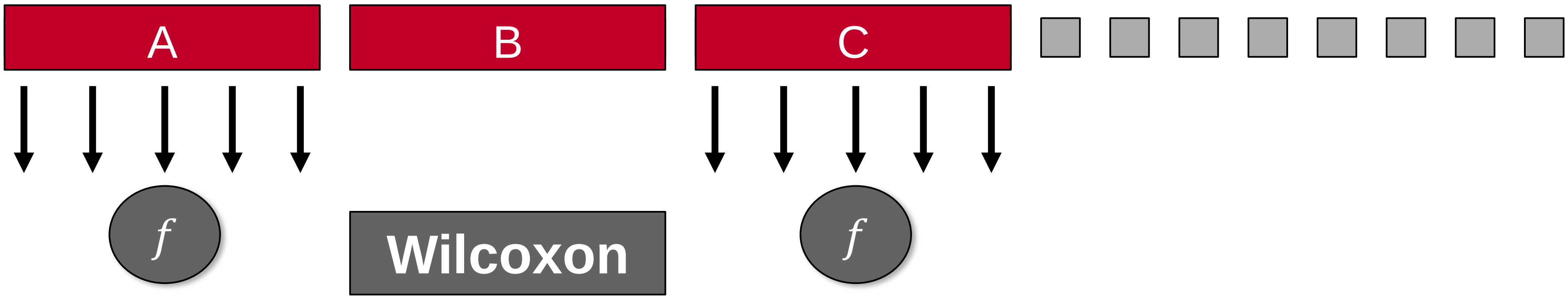
Recognize Regions of Drift



Window-based approach

- extension of an approach by Dries and Rückert
- compute density values with current estimate *f*

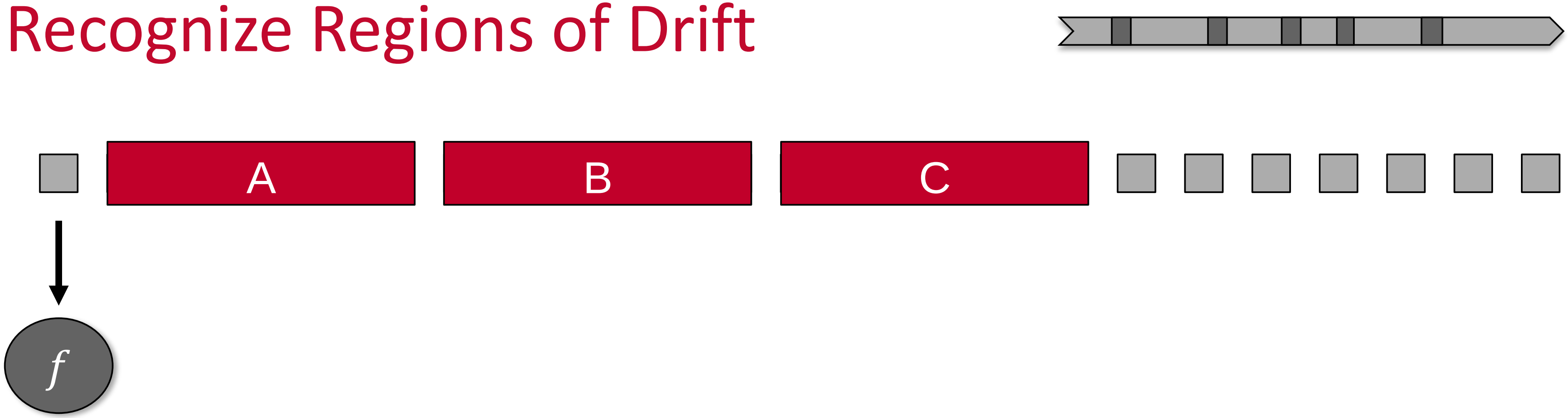
Recognize Regions of Drift



Window-based approach

- extension of an approach by Dries and Rückert
- compute density values with current estimate *f*
- perform drift detection with Wilcoxon rank-sum test

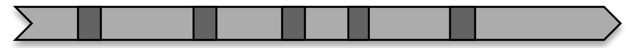
Recognize Regions of Drift

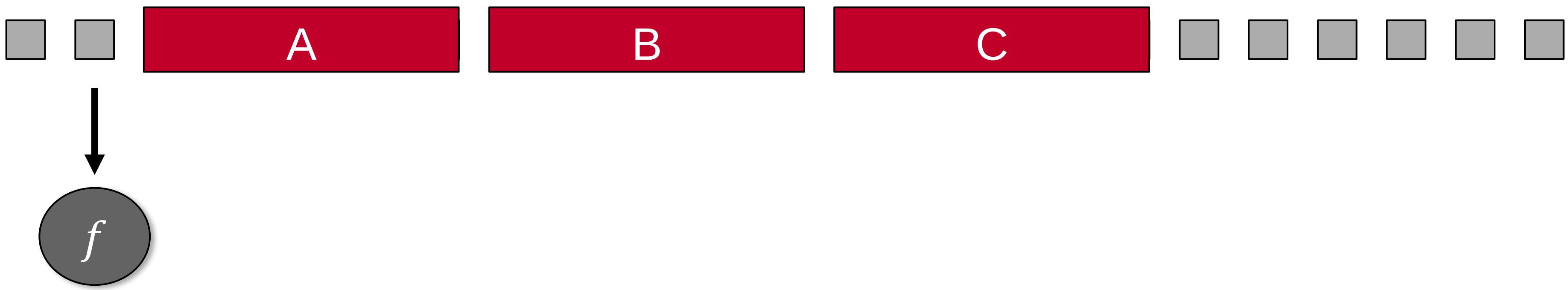


Window-based approach

- extension of an approach by Dries and Rückert
- compute density values with current estimate *f*
- perform drift detection with Wilcoxon rank-sum test
- update *f* with clean instances only

Recognize Regions of Drift

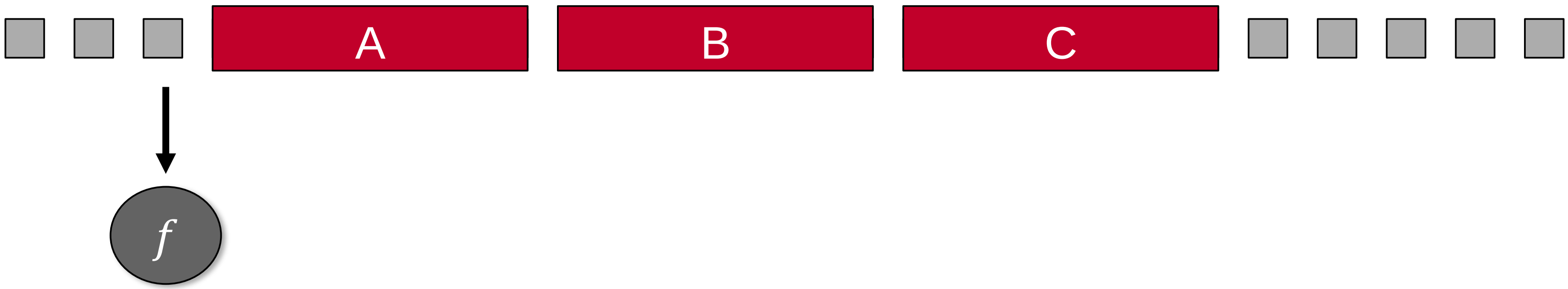




Window-based approach

- extension of an approach by Dries and Rückert
- compute density values with current estimate *f*
- perform drift detection with Wilcoxon rank-sum test
- update *f* with clean instances only

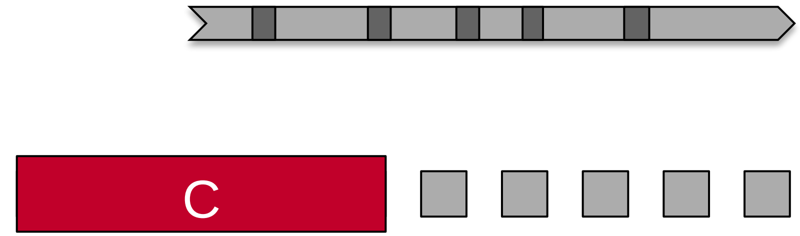
Recognize Regions of Drift



Window-based approach

- extension of an approach by Dries and Rückert
- compute density values with current estimate *f*
- perform drift detection with Wilcoxon rank-sum test
- update *f* with clean instances only

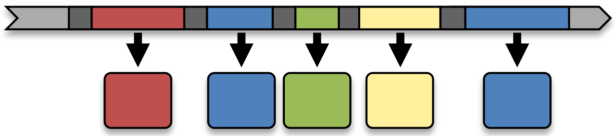
Recognize Regions of Drift



Window-based approach

- extension of an approach by Dries and Rückert
- compute density values with current estimate f
- perform drift detection with Wilcoxon rank-sum test
- update f with clean instances only

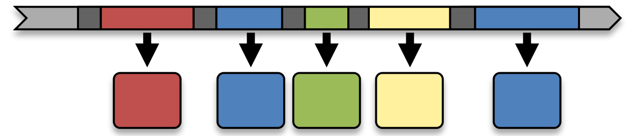
Recurrences of Densities



C

Recurrent or new?

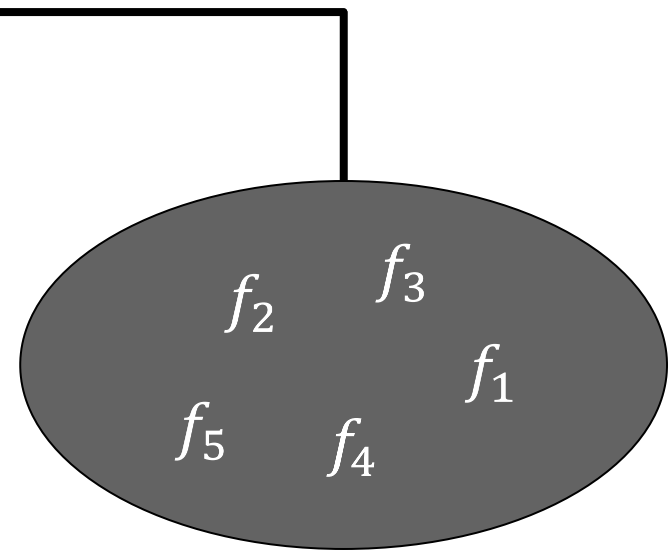
Recurrences of Densities



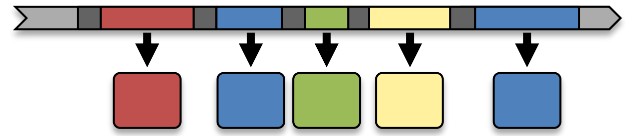
C

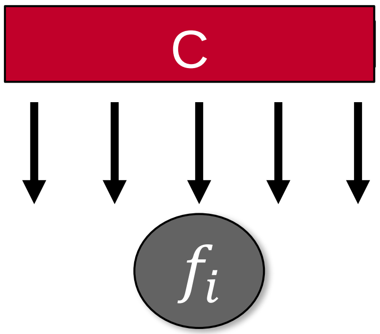
Recurrent or new?

- compare with pool of existing density estimates
- use statistical test we proposed earlier
- reactivate estimate if one is found
- initialize a new one otherwise



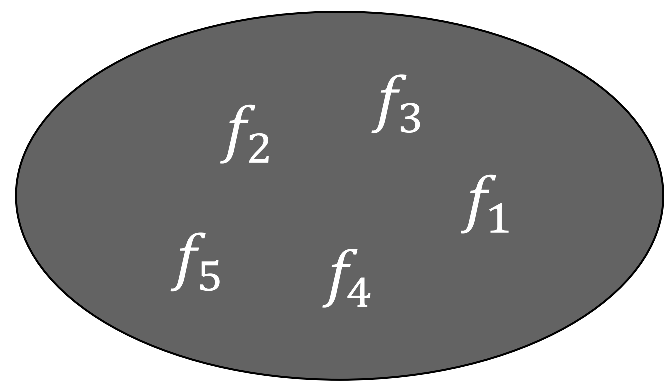
Recurrences of Densities



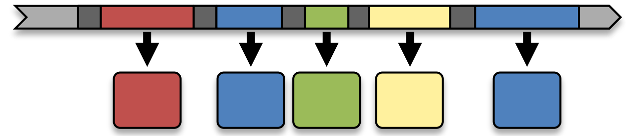


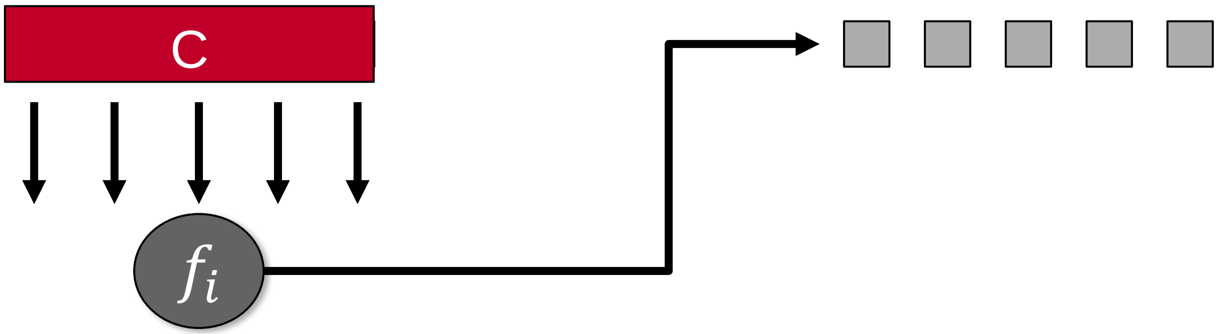
Recurrent or new?

- compare with pool of existing density estimates
- use statistical test we proposed earlier
- reactivate estimate if one is found
- initialize a new one otherwise



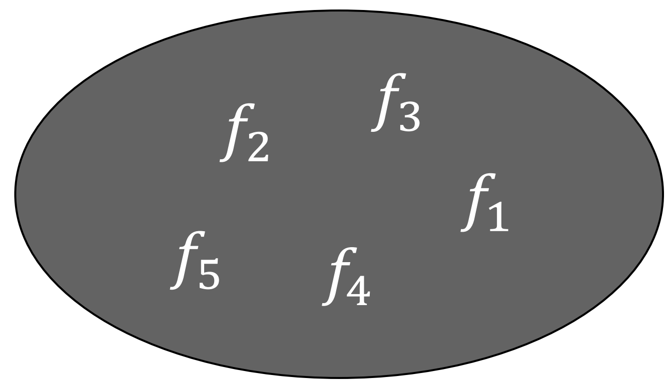
Recurrences of Densities



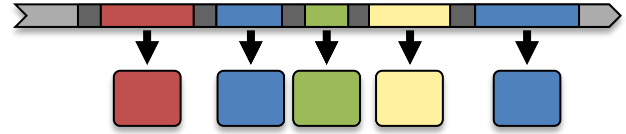


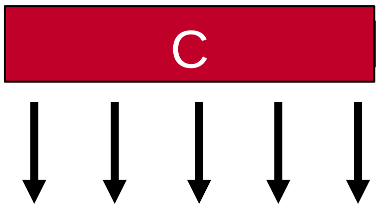
Recurrent or new?

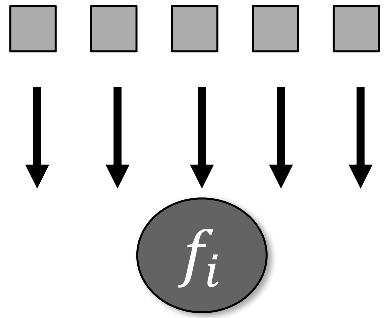
- compare with pool of existing density estimates
- use statistical test we proposed earlier
- reactivate estimate if one is found
- initialize a new one otherwise



Recurrences of Densities

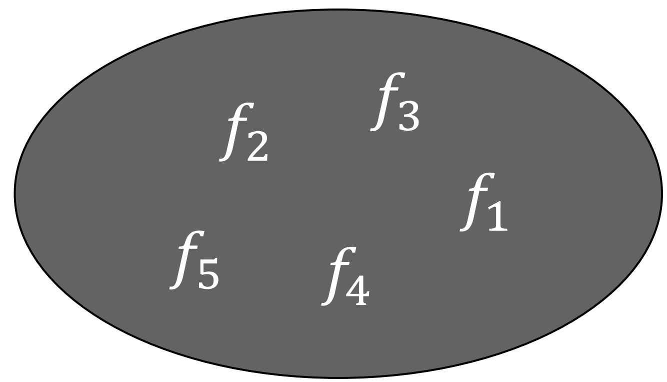




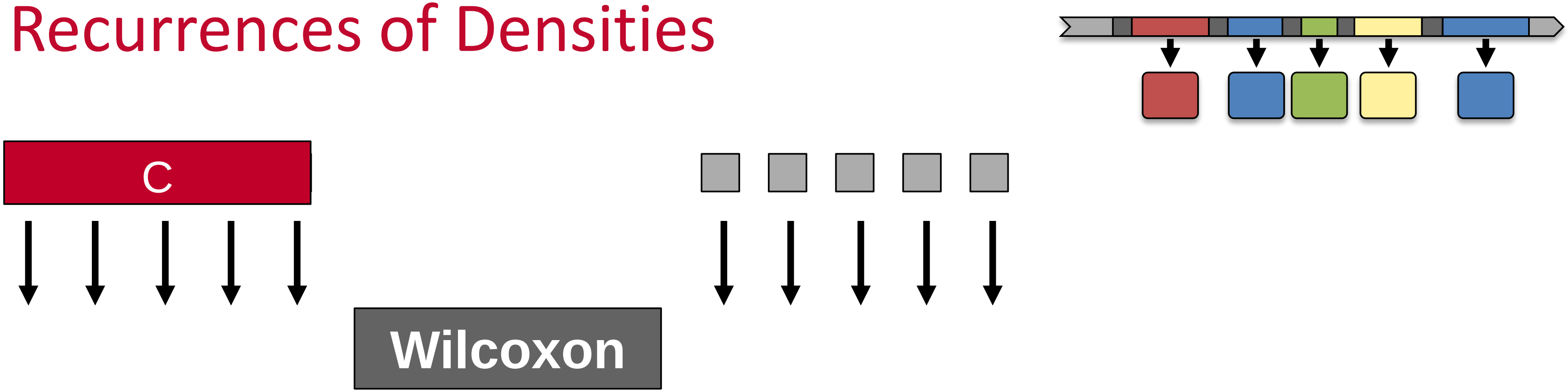


Recurrent or new?

- compare with pool of existing density estimates
- use statistical test we proposed earlier
- reactivate estimate if one is found
- initialize a new one otherwise

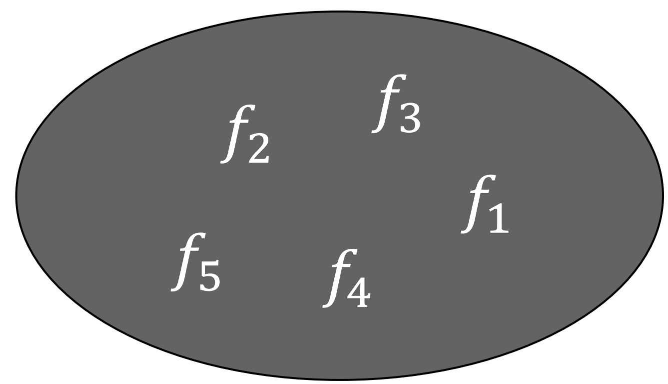


Recurrences of Densities

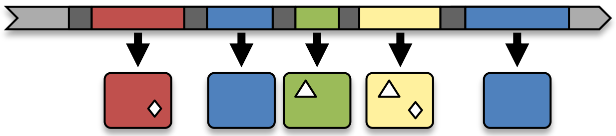


Recurrent or new?

- compare with pool of existing density estimates
- use statistical test we proposed earlier
- reactivate estimate if one is found
- initialize a new one otherwise



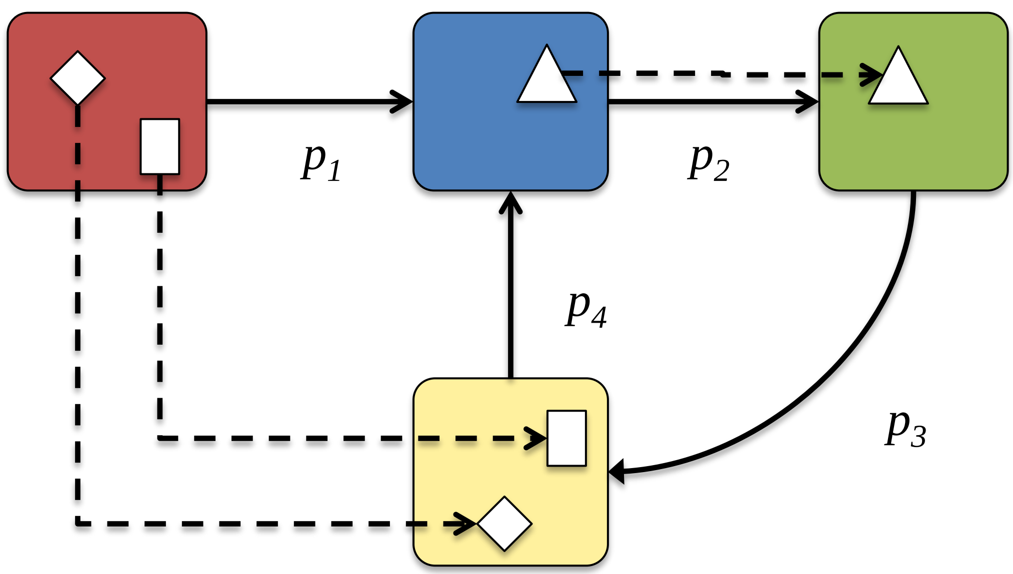
Recurrences of Density Parts



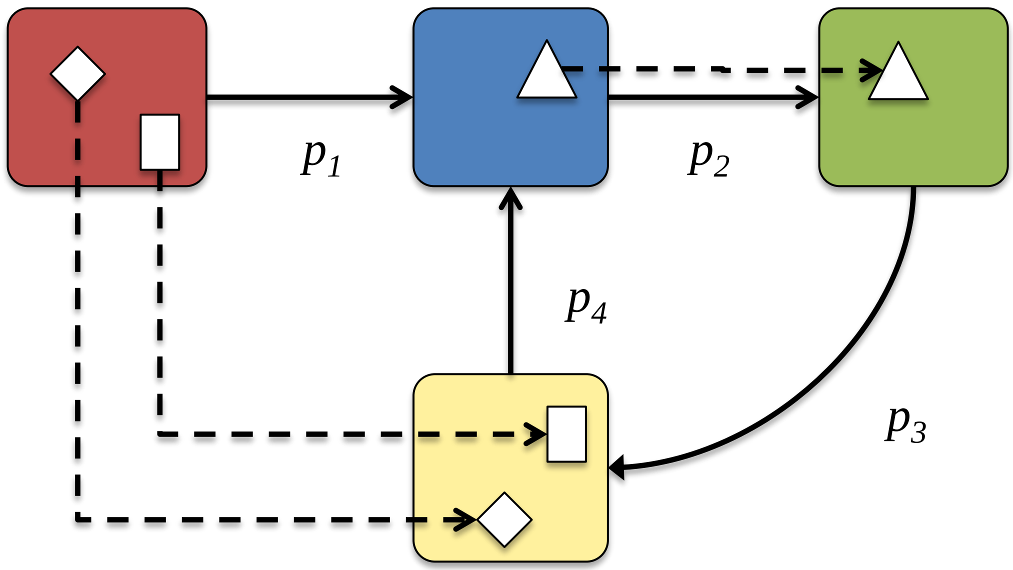
Introduction of modules

$$f(X_1, X_2, \ldots, X_8) = f_1(X_1, X_3, X_8) \cdot f_2(X_2, X_4, X_5) \cdot f_3(X_6) \cdot f_4(X_7)$$

If the f_i cannot be decomposed any further, then f_1, f_2, f_3, f_4 are called the modules of f.

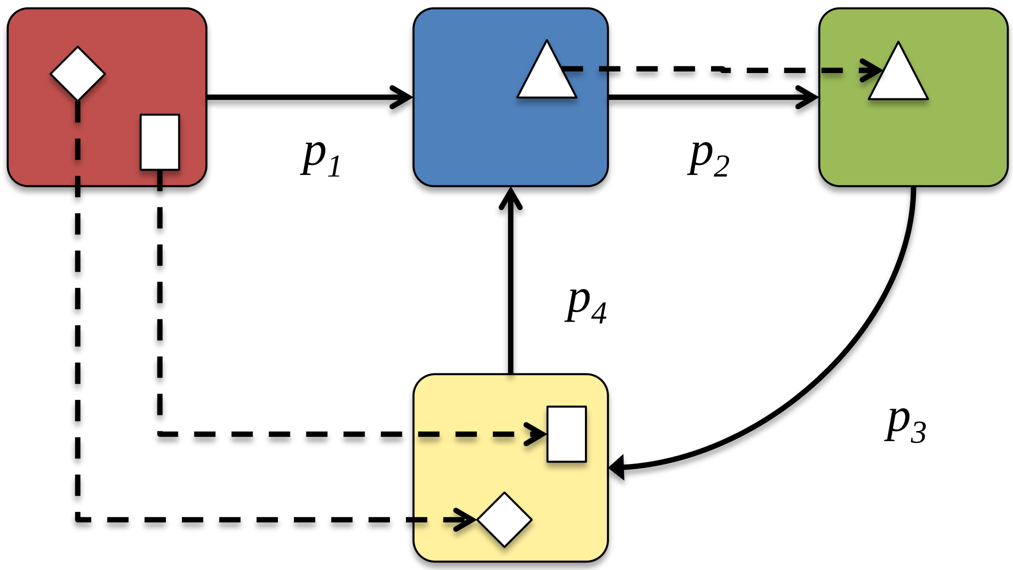

p_1
p_2
p_4
p_3

Query Mechanism



- probabilistic extension of possible worlds semantics
- requires density estimators supporting inference tasks

Query Mechanism



- probabilistic extension of possible worlds semantics
- requires density estimators supporting inference tasks

Query 3 [over multiple worlds]

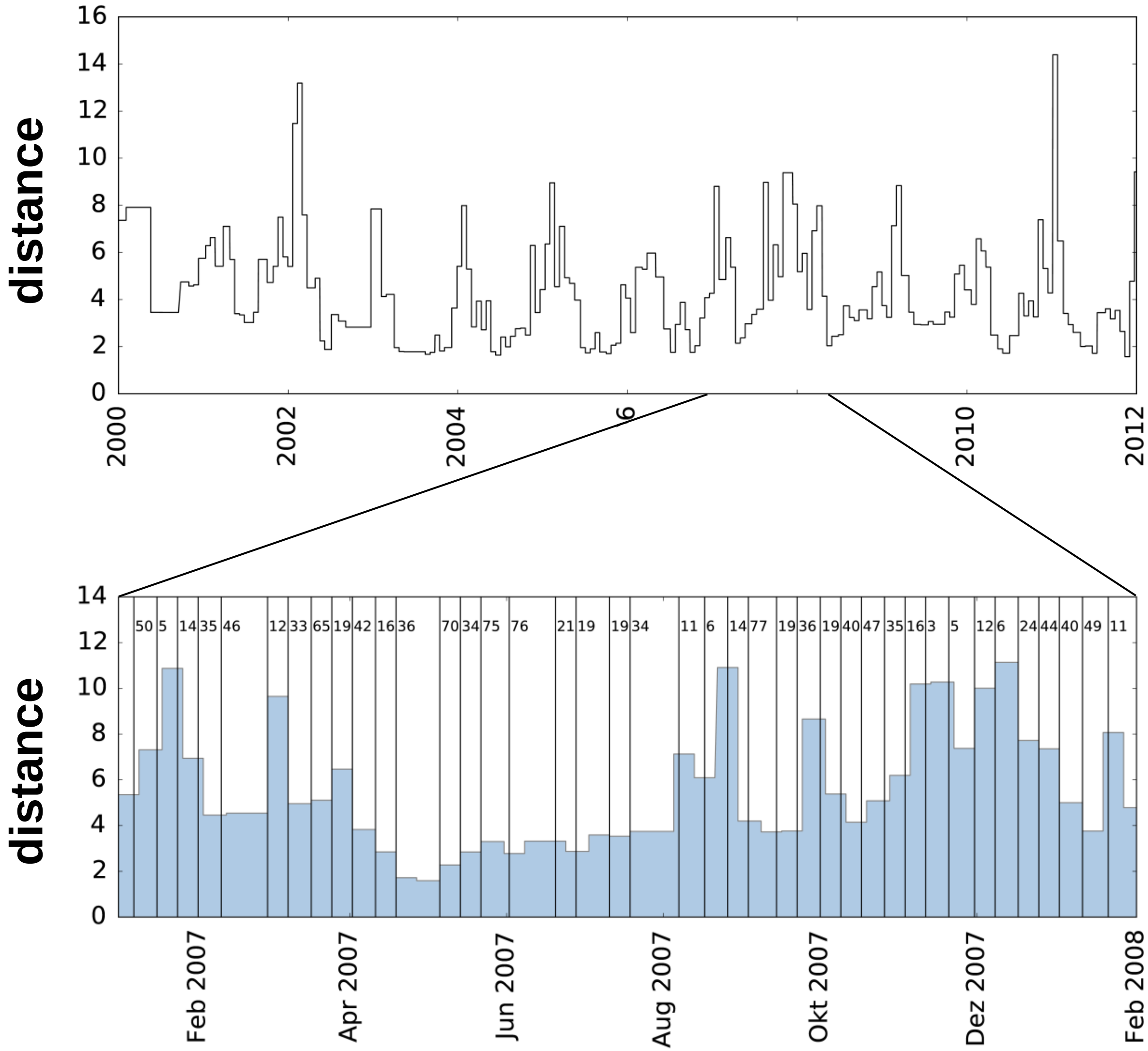
Given world *W*, what is the probability that the resident will switch on the light in the office room?

Evaluation: Modules

- evaluation on
  synthetic and real-world datasets
- without modules performance is better in many cases, but only slightly
- more explicit representation that enables detection of recurrences

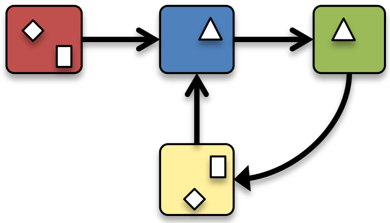
Datasets
Synthetic
Bayesian networks with different numbers of nodes, different numbers of instances different numbers of variable groups
Real-World
Electricity Shuttle Waterlevel Covertype

Evaluation: Recurrences



Recurrences	
Densities	Modules
416	1259
100%	78%

Conclusions and Future Work



- framework to model recurrent densities and recurrent parts of the densities
- online estimator
- extension of possible worlds semantics for query mechanism

Future Work:

- more sophisticated modeling of density parts (conditional)
- recycling of modules
- implementation of query mechanism

Thank you for your attention