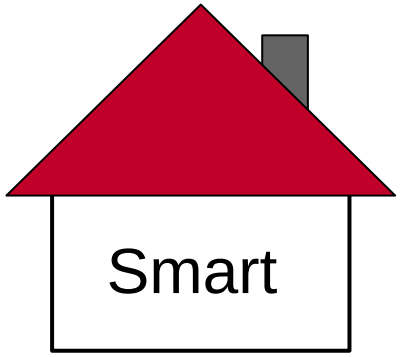
JOHANNES GUTENBERG
UNIVERSITÄT MAINZ

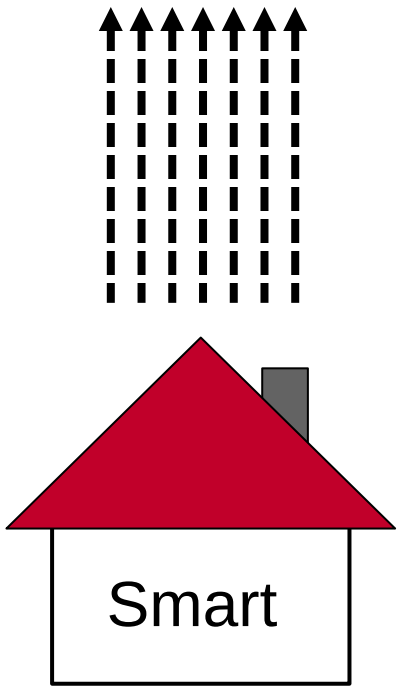
Online Density Estimation of Heterogeneous Data Streams in Higher Dimensions

Michael Geilke, Andreas Karwath, and Stefan Kramer

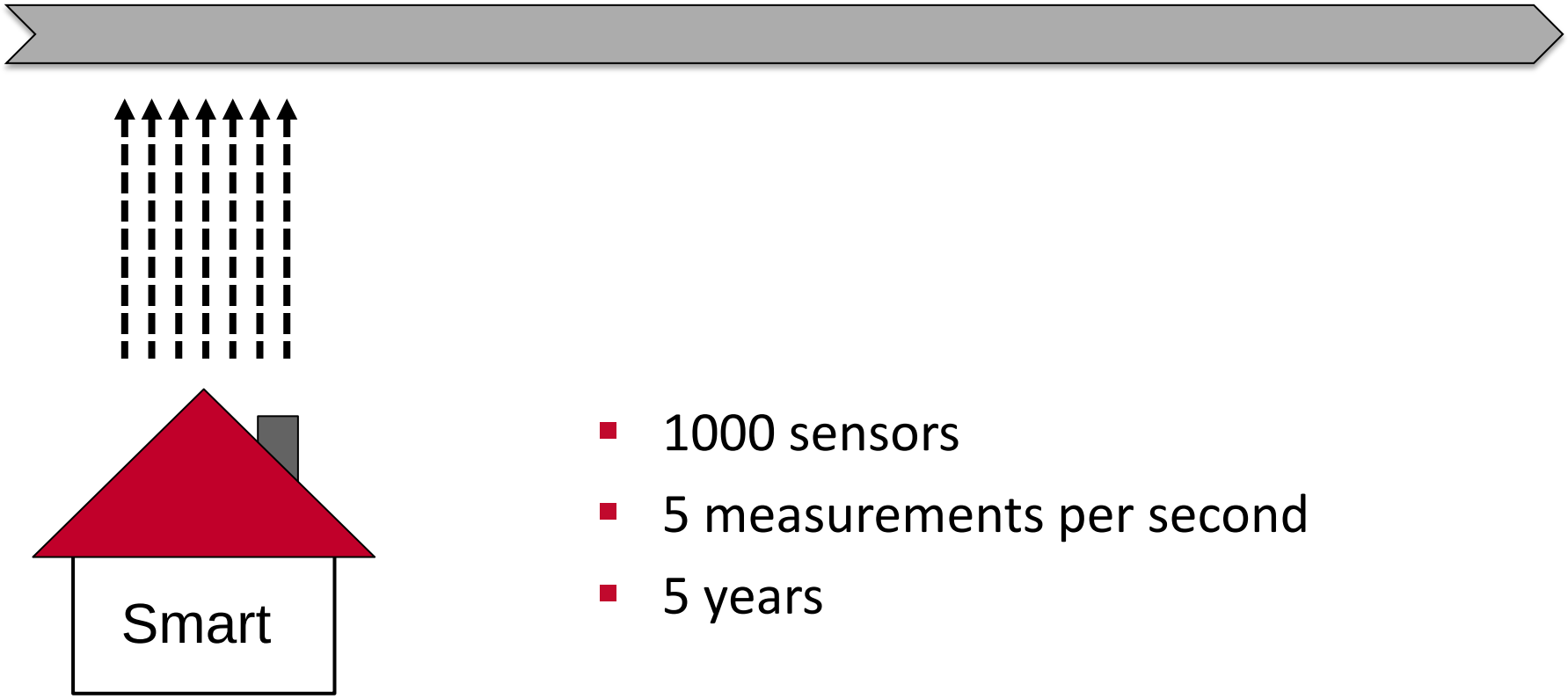
Johannes Gutenberg University Mainz, Germany

September 22, 2016


Smart


Smart

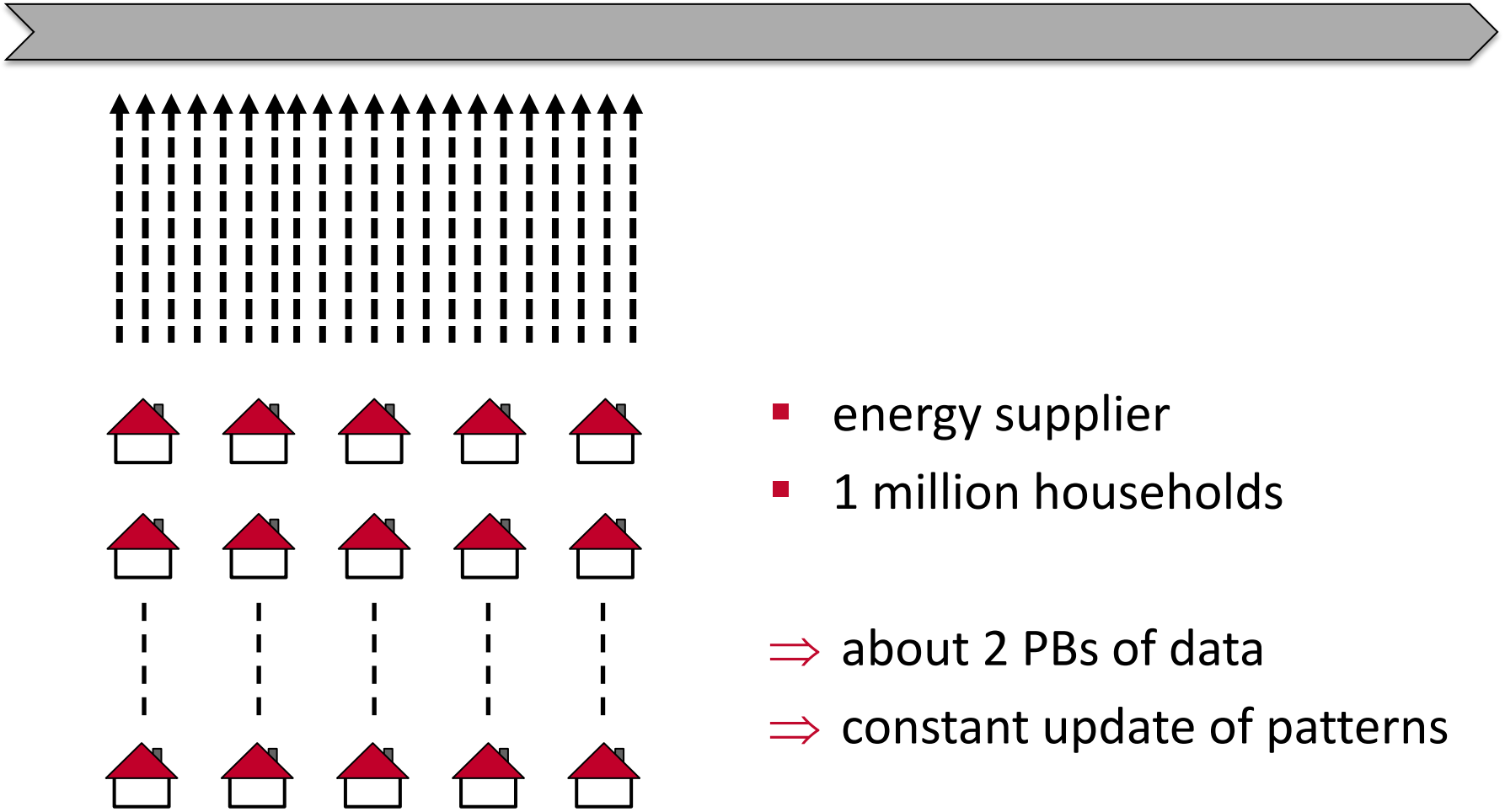

Smart



- 1000 sensors
- 5 measurements per second
- 5 years

$\Rightarrow$ more than 2 billion measurements

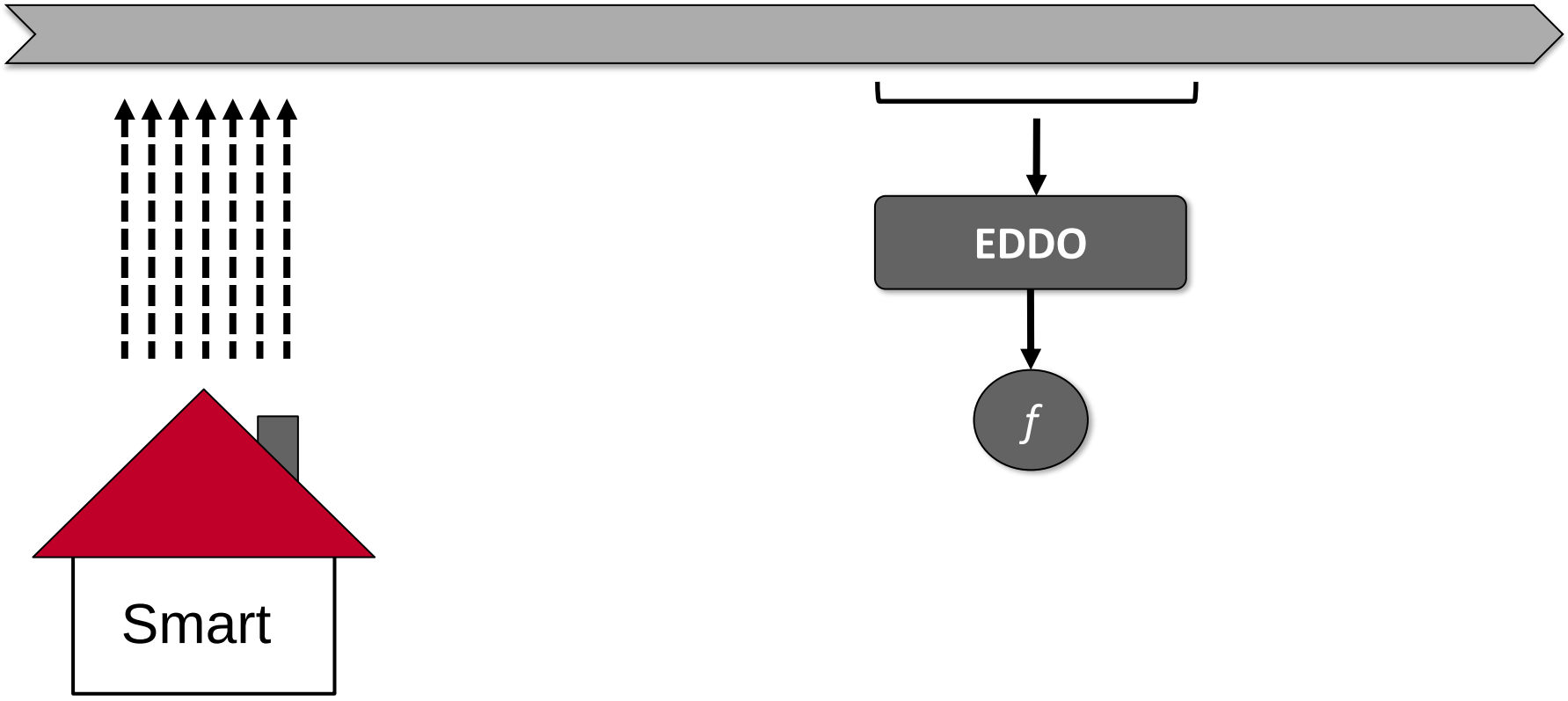
$\Rightarrow$ about 2 GBs of data

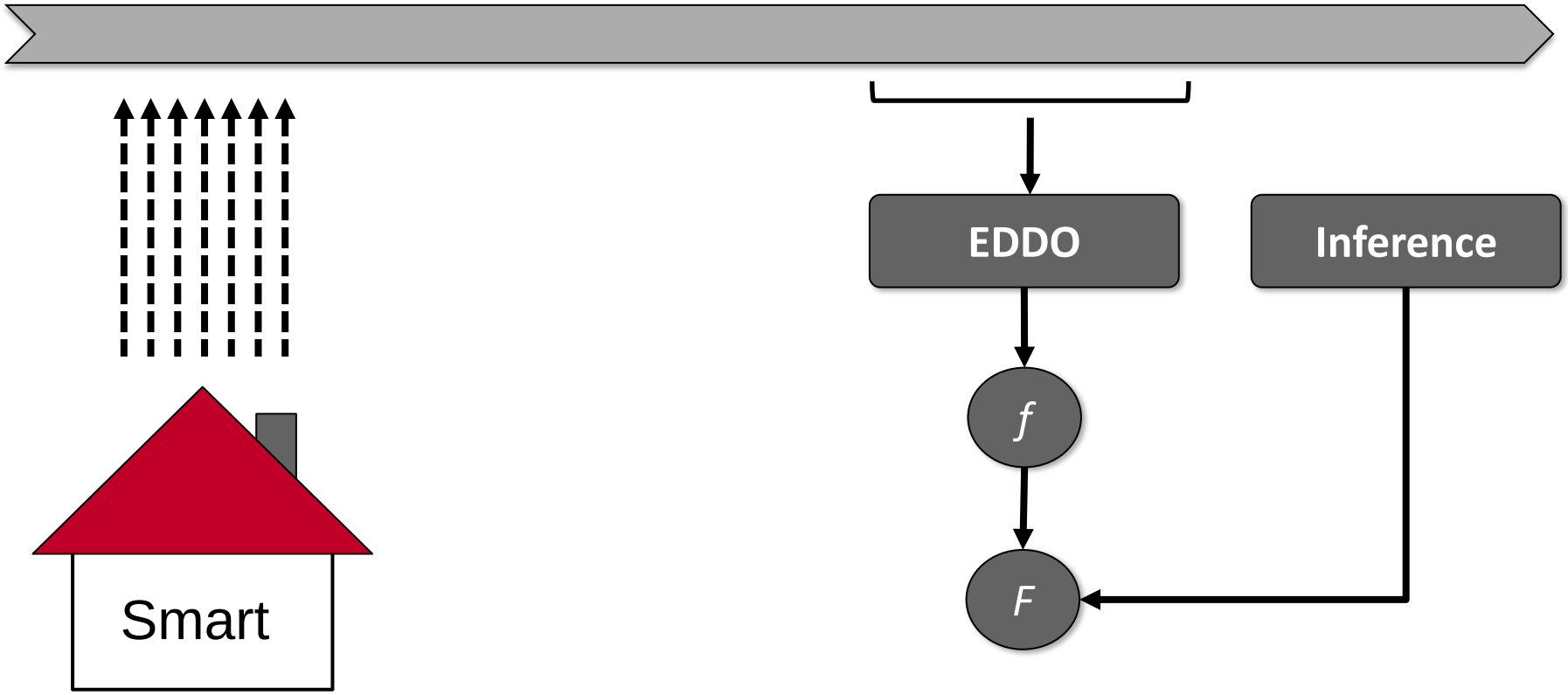


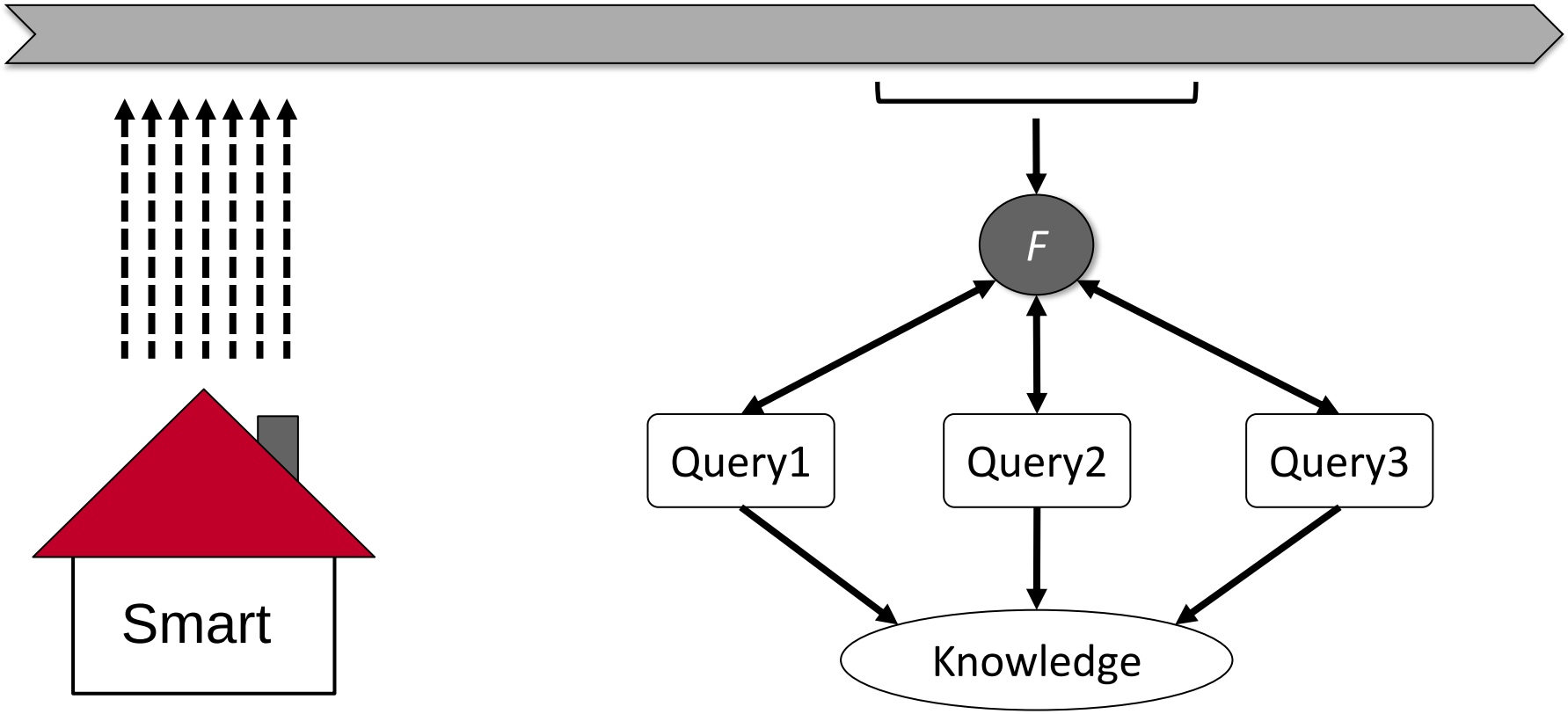
- energy supplier
- 1 million households

$\Rightarrow$ about 2 PBs of data

$\Rightarrow$ constant update of patterns


EDDO
f
Smart


EDDO
Inference
f
F
Smart

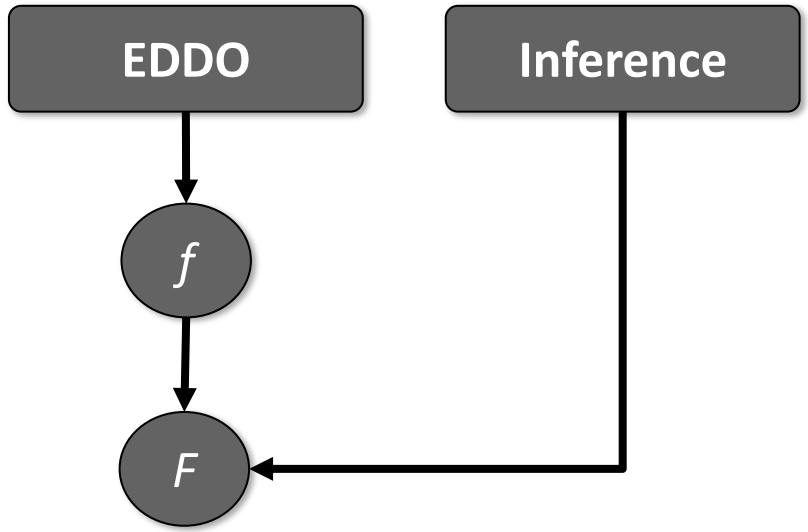


Query:

Return the probability distribution for sensors in the living room during the week days.

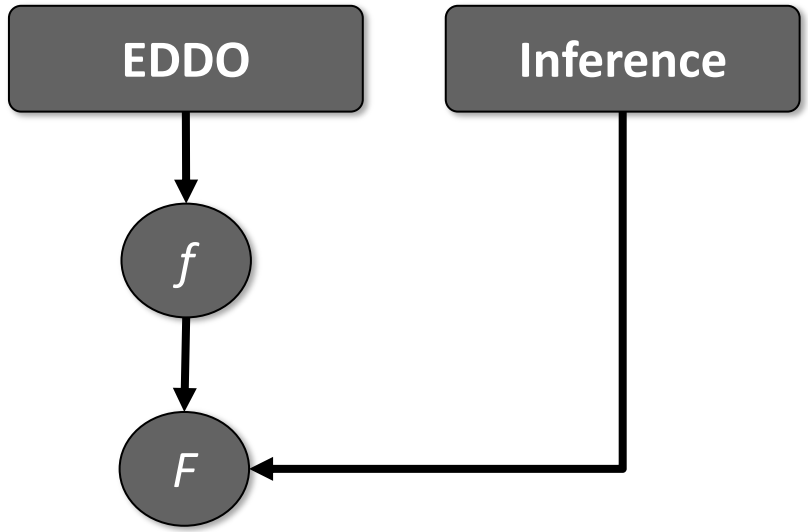
Weaknesses of EDDO

$$f(X_1, \ldots, X_n)$$



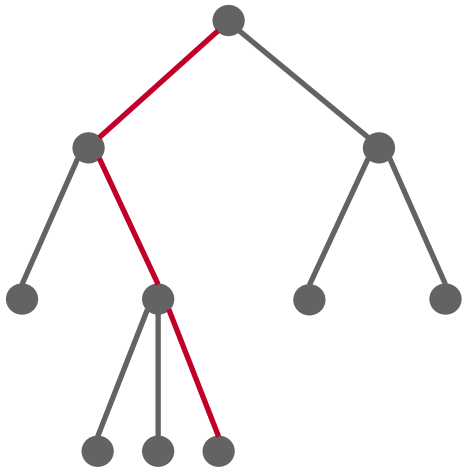
Weaknesses of EDDO

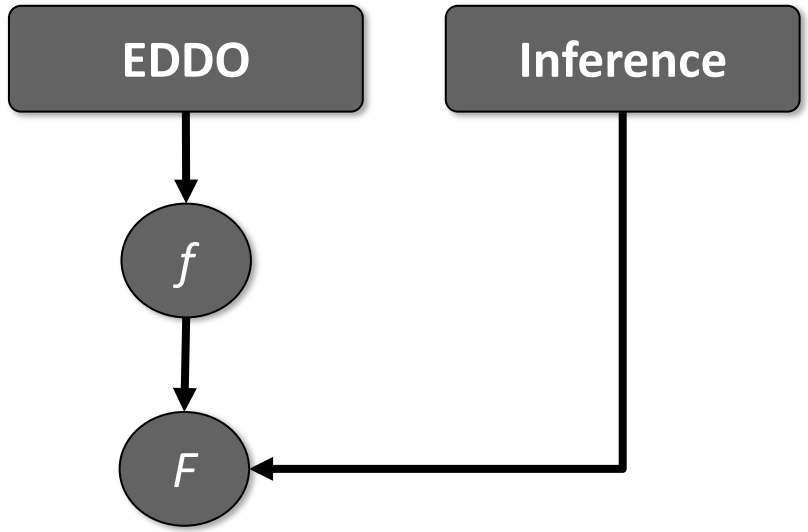
$$\mathrm{f}(X_1) \cdot \prod_{i=2}^{n} f(X_i | X_1, \ldots, X_{i-1})$$



Weaknesses of EDDO

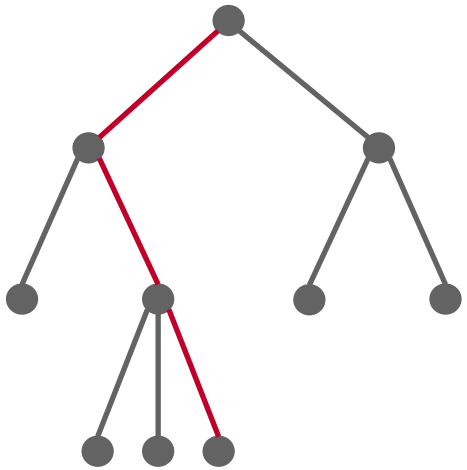
$$\mathrm{f}(X_1) \cdot \prod_{i=2}^{n} f(X_i | X_1, \ldots, X_{i-1})$$

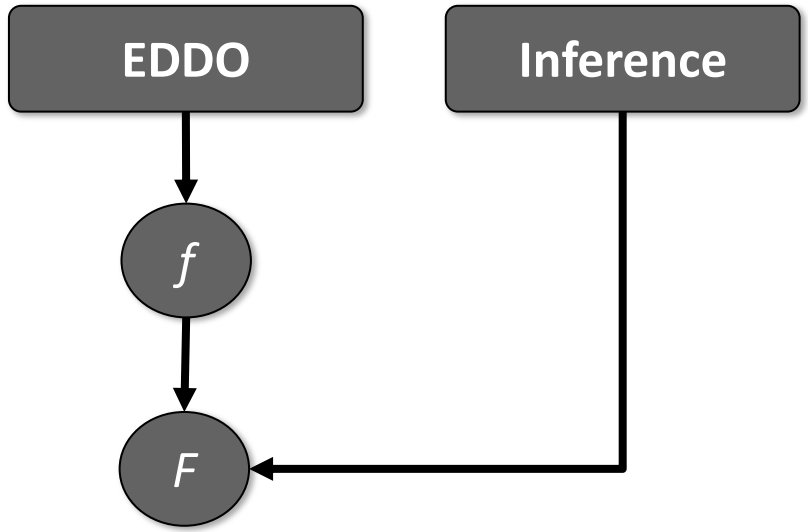




Weaknesses of EDDO

$$\mathrm{f}(X_1) \cdot \prod_{i=2}^{n} f(X_i | X_1, \ldots, X_{i-1})$$





only for discrete random variables

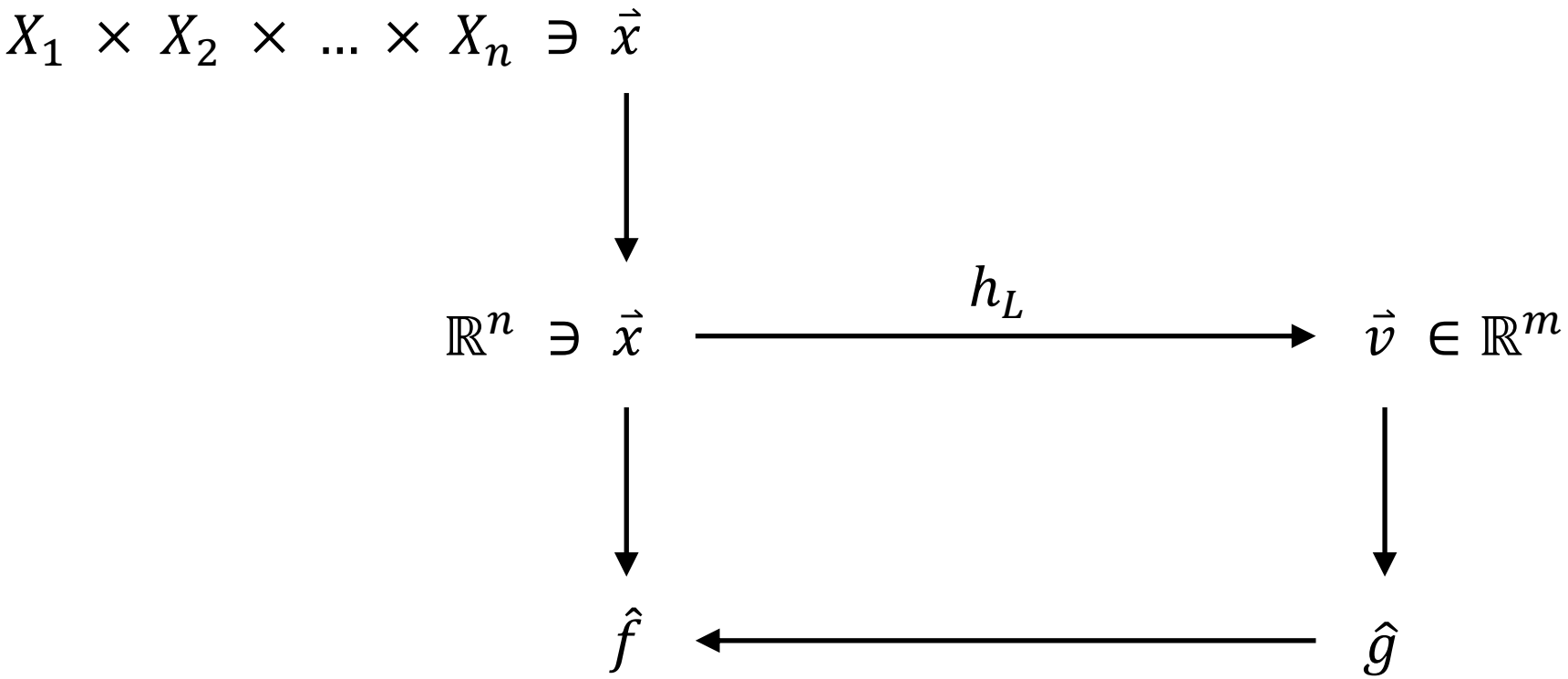
Goals

A density estimator that

- estimates joint densities from data streams
- is able to deal with heterogeneous data, and
- and works for higher dimensional data.

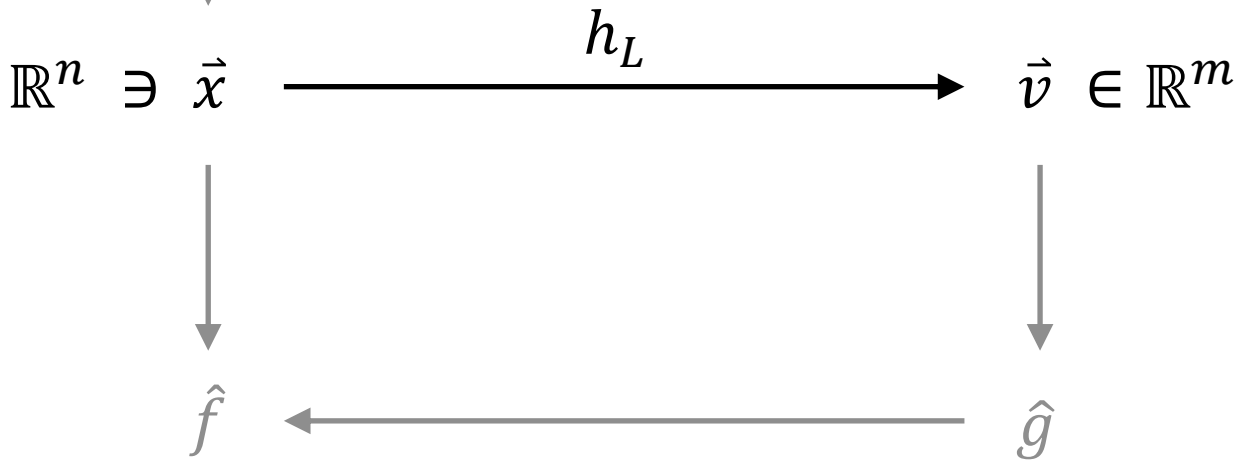
For density estimation, 100 variables is high dimensional.

Main Idea

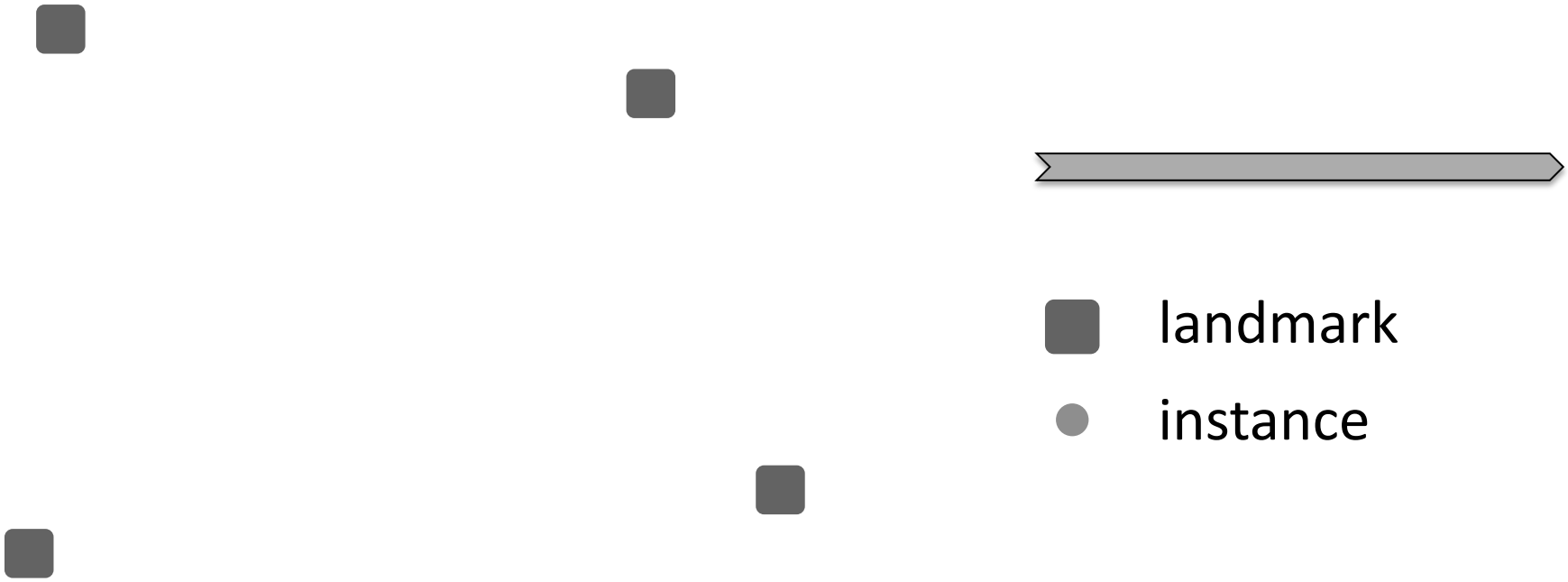

$X_1 \times X_2 \times \ldots \times X_n \ni \vec{x}$
$\mathbb{R}^n \ni \vec{x}$
h_L
$\vec{v} \in \mathbb{R}^m$
$\hat{f}$
$\hat{g}$

Main Idea

$X_1 \times X_2 \times \ldots \times X_n \ni \vec{x}$

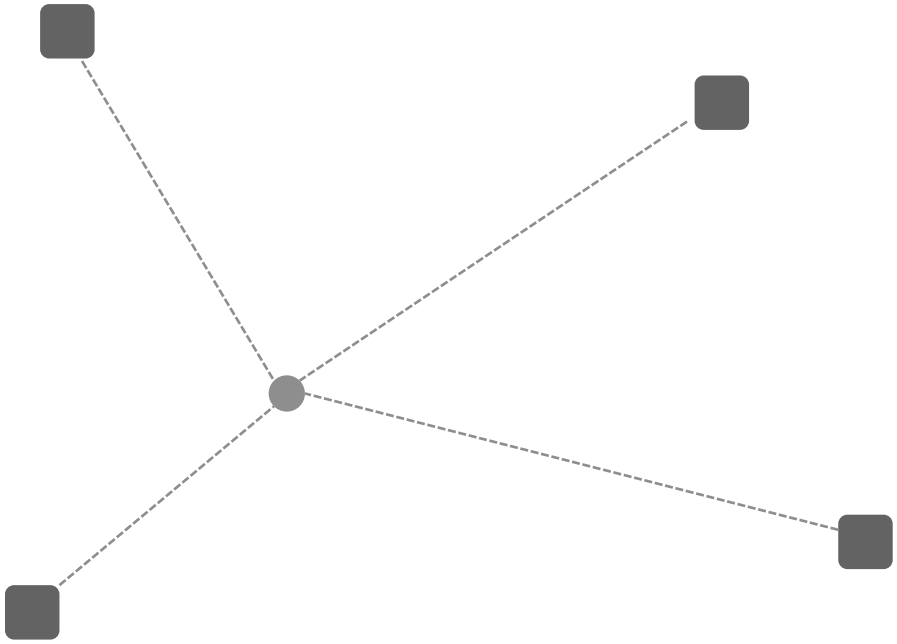


Online Density Estimation using Representatives (RED)



$(temperature = 20, humidity = 50, \dots) = \blacksquare \in \mathbb{R}^n$

Online Density Estimation using Representatives (RED)



landmark

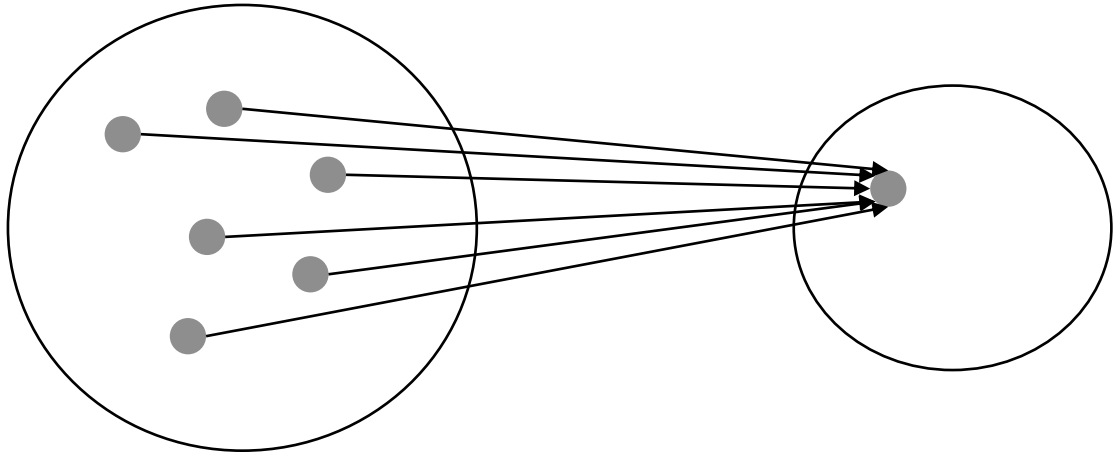
instance

$$\mathbb{R}^n \ni \vec{x} = (x_1, x_2, \ldots, x_n) \longrightarrow \vec{v} = (v_1, v_2, v_3, v_4) \in \mathbb{R}^m$$

Online Density Estimation using Representatives (RED)

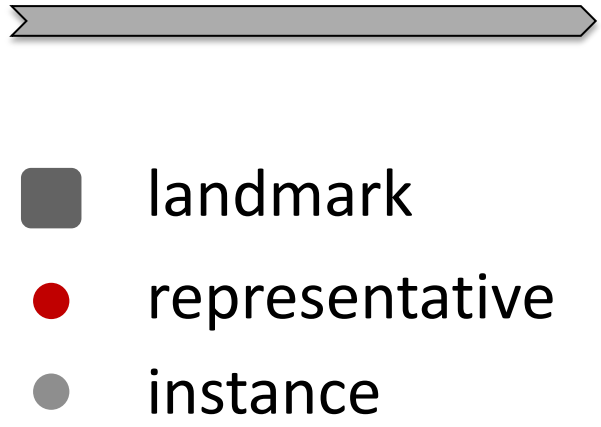
$$I = \{\vec{x} \in \mathbb{R}^n \mid h_L(\vec{x}) = \vec{v} = (v_1, v_2, v_3, v_4)\}$$

$$\hat{g}(\vec{v}) = \sum_{\vec{x} \in I} \hat{f}(\vec{x})$$



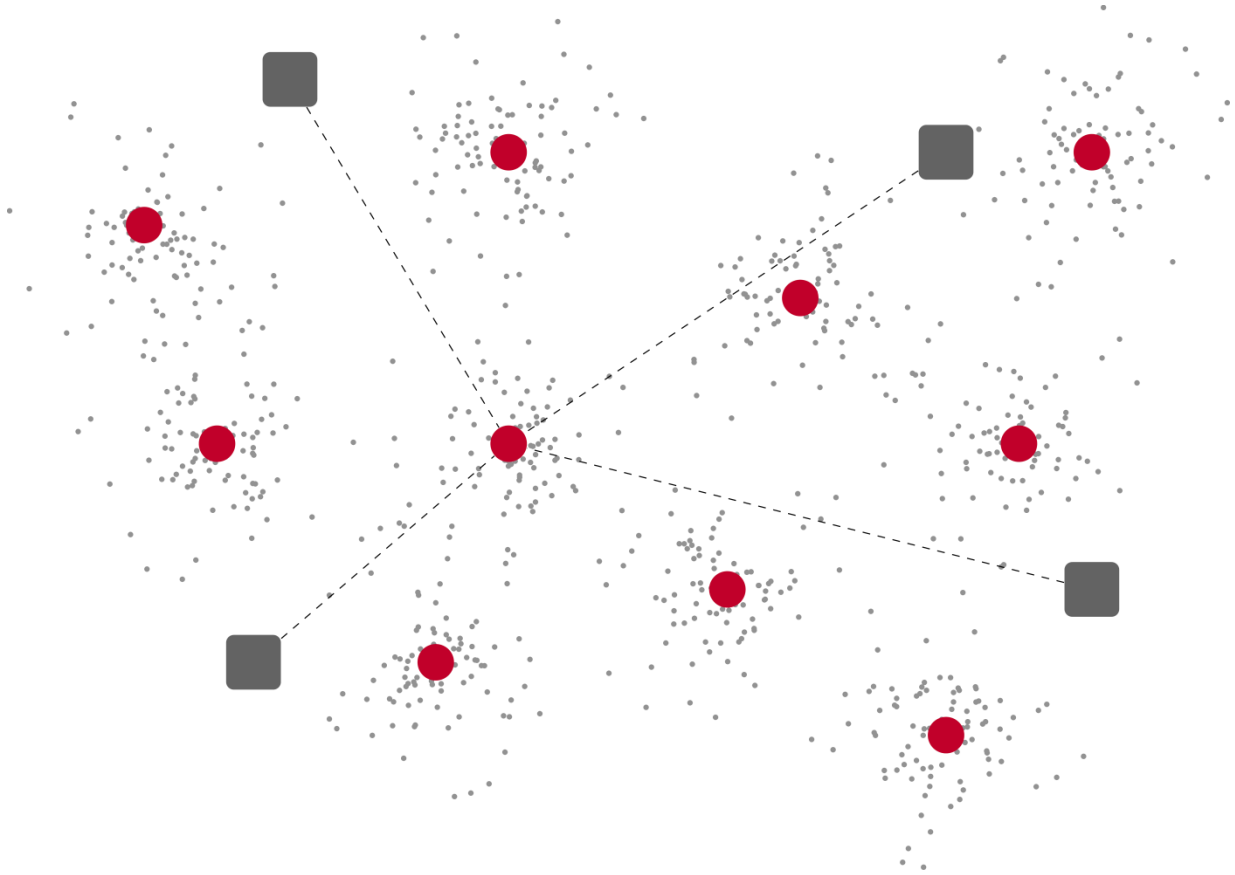
$\mathbb{R}^n$ $\mathbb{R}^m$

Online Density Estimation using Representatives (RED)



$(temperature = 10, humidity = 80, \dots) =$ ● $\in \mathbb{R}^n$

Online Density Estimation using Representatives (RED)



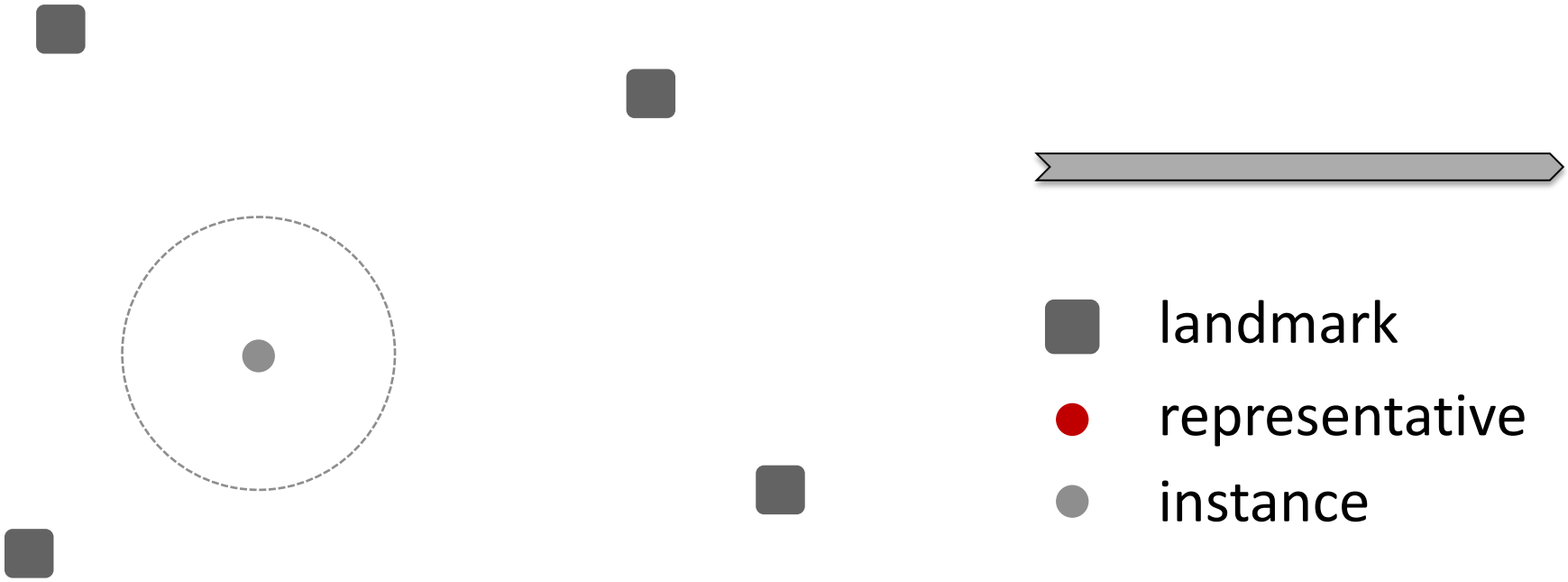

landmark


representative


instance

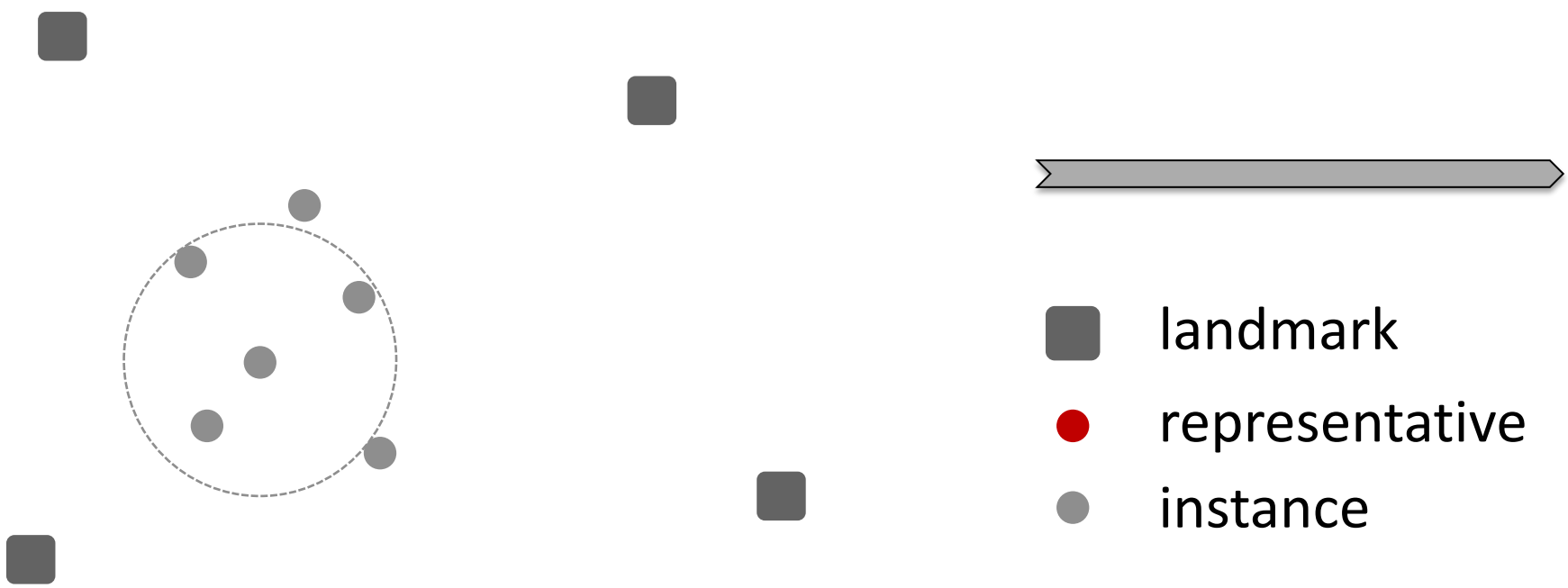
$$\{\vec{x} \in \mathbb{R}^n \mid h_L(\vec{x}) = (v_1, v_2, v_3, v_4)\}$$

Online Density Estimation using Representatives (RED)



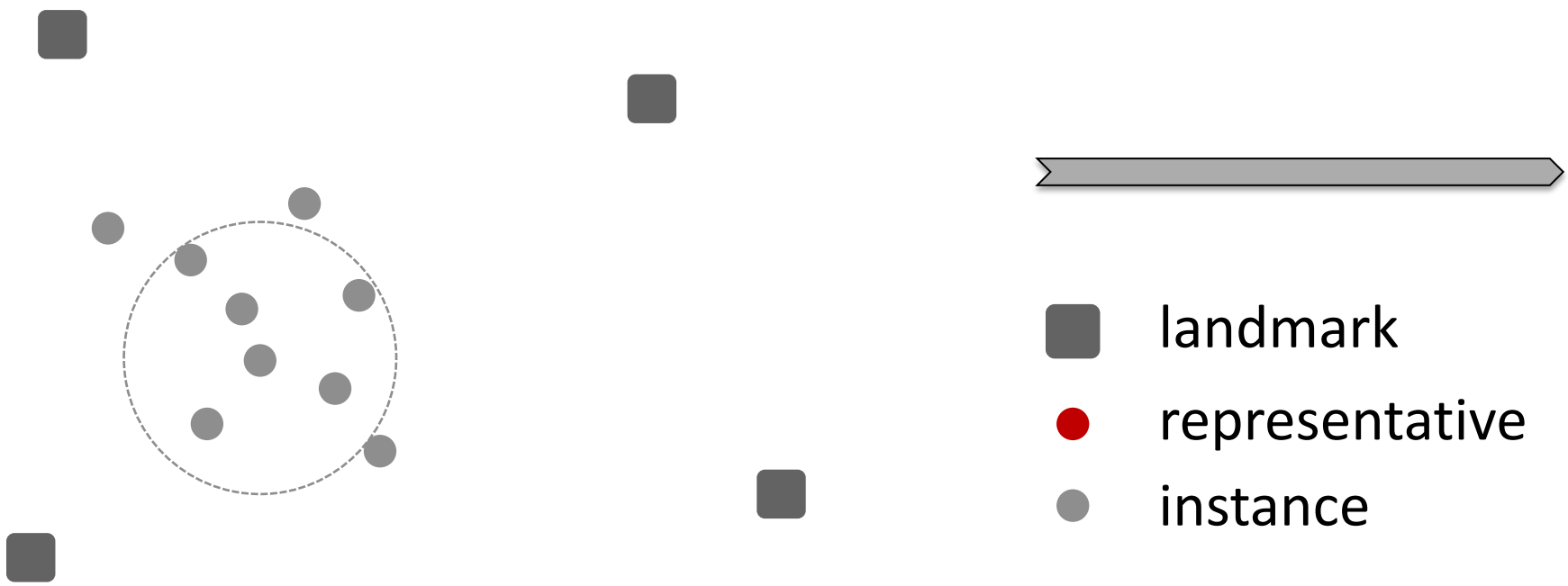
Mahalanobis distance: $\sqrt{(\vec{x} - \vec{v})^T \Sigma^{-1} (\vec{x} - \vec{v})}$

Online Density Estimation using Representatives (RED)



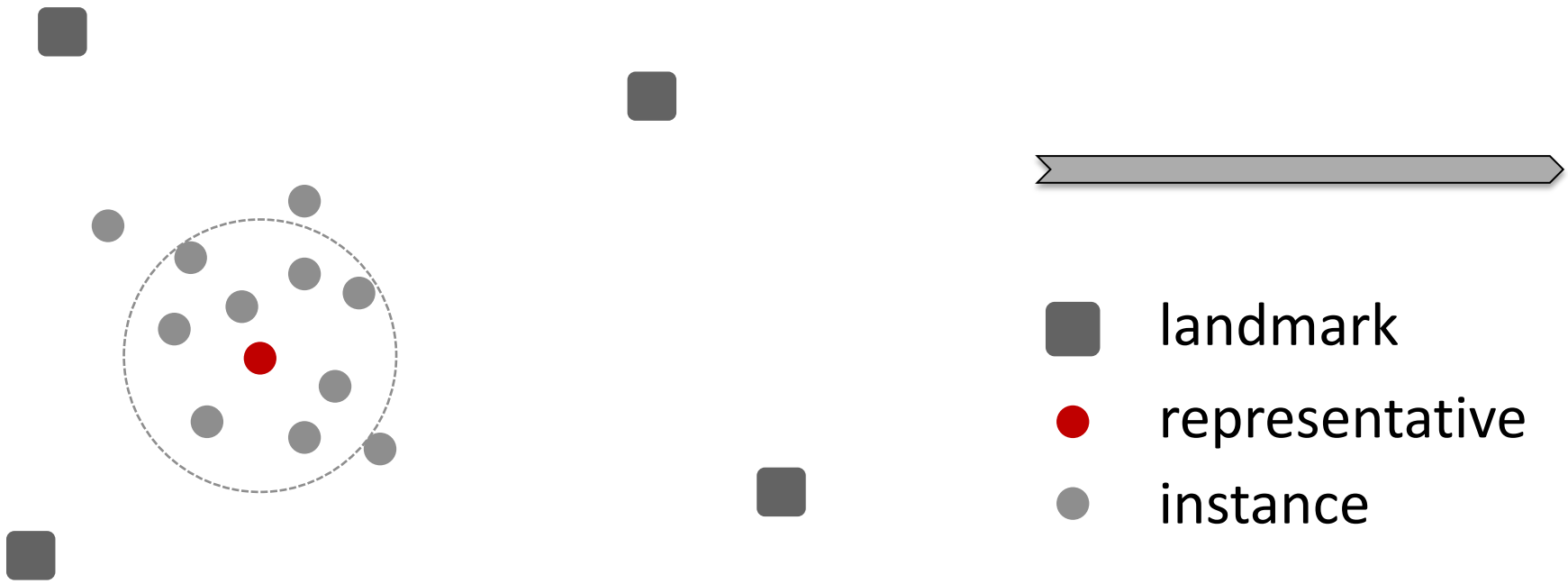
$$\vec{v} = (v_1, v_2, v_3, v_4)$$

Online Density Estimation using Representatives (RED)



$$\vec{v} = (v_1, v_2, v_3, v_4)$$

Online Density Estimation using Representatives (RED)

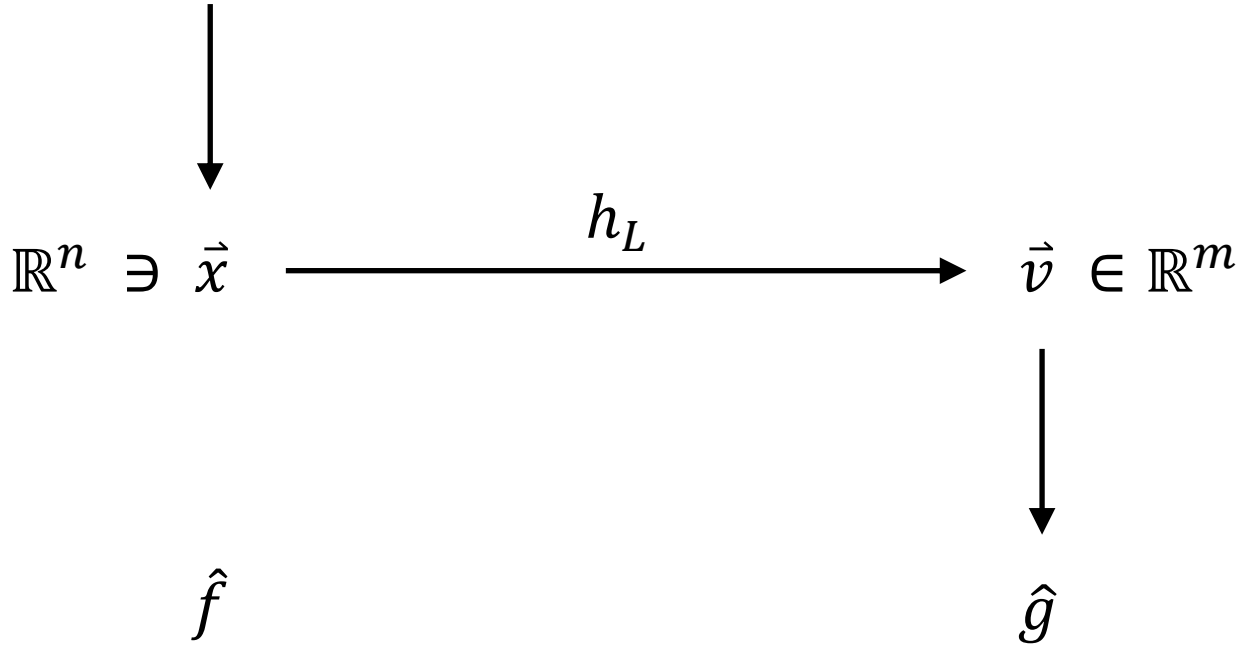


$$\vec{v} = (v_1, v_2, v_3, v_4) \longrightarrow \mathrm{f}(V_1) \cdot \prod_{i=2}^{m} f(V_i | V_1, \ldots, V_{i-1})$$

Online Density Estimation using Representatives (RED)

$X_1 \times X_2 \times \ldots \times X_n \ni \vec{x}$

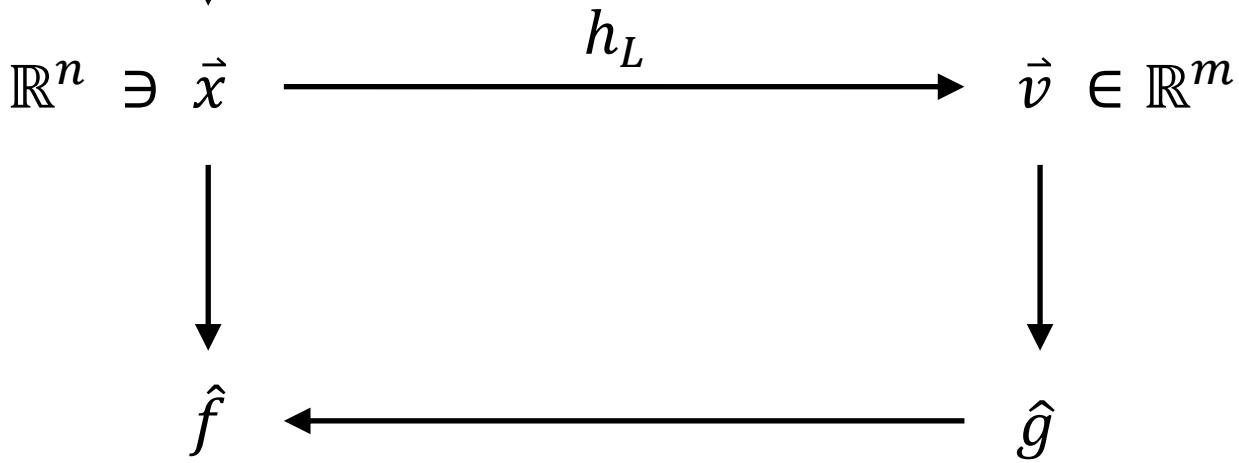
$h_L(\vec{x}) = h_L(\vec{y})$ **but** $\vec{x} \neq \vec{y}$



Online Density Estimation using Representatives (RED)

$X_1 \times X_2 \times \ldots \times X_n \ni \vec{x}$

$h_L(\vec{x}) = h_L(\vec{y})$ **but** $\vec{x} \neq \vec{y}$



$\hat{g}(\vec{v}) = \sum_{\vec{x} \in I} \hat{f}(\vec{x})$

Choice of Landmarks

Main idea:

- theoretical foundation
- landmarks are orthogonal to each other
- if $|L|$= d + 1, then consistent estimator
- back translation by system of linear equations

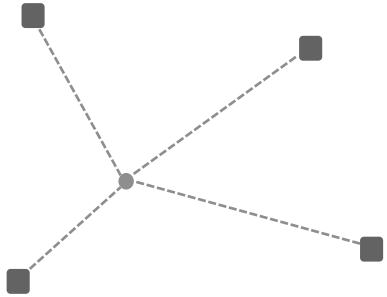
Evaluation: Parameter Setting

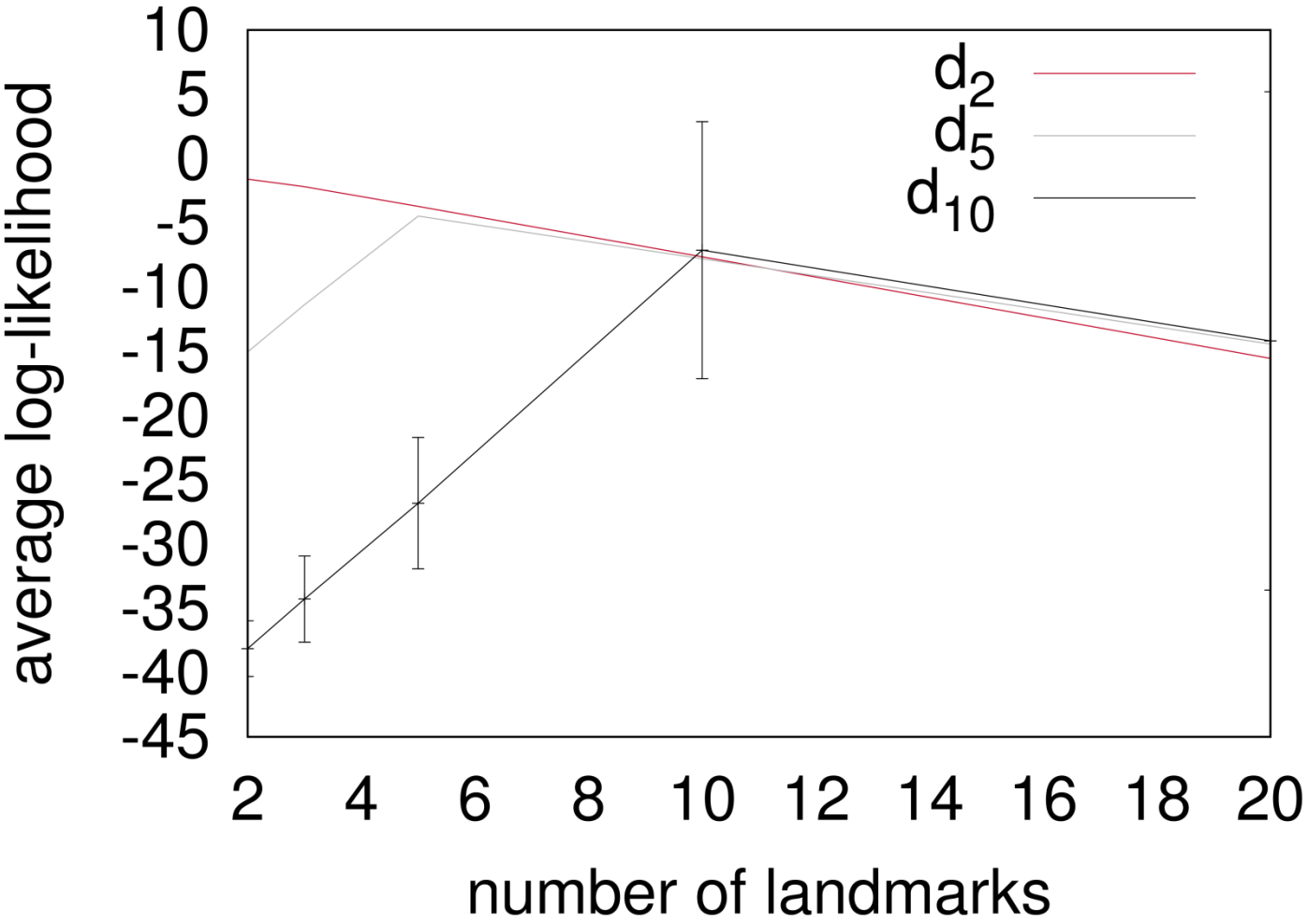
Parameters:

- $\theta_{C \to R} = 100$
- Euclidean norm
- $|L| \in \{2, 3, 5, 10, 20\}$
- $M \in \{0.1, 0.5, 1.0, 2.0, 5.0, 10.0\}$

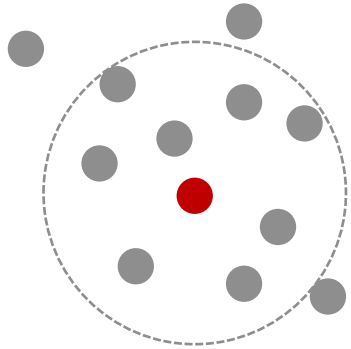
Datasets
Synthetic
Gaussian mixtures
Real-World
Covertype Electricity Letter Shuttle

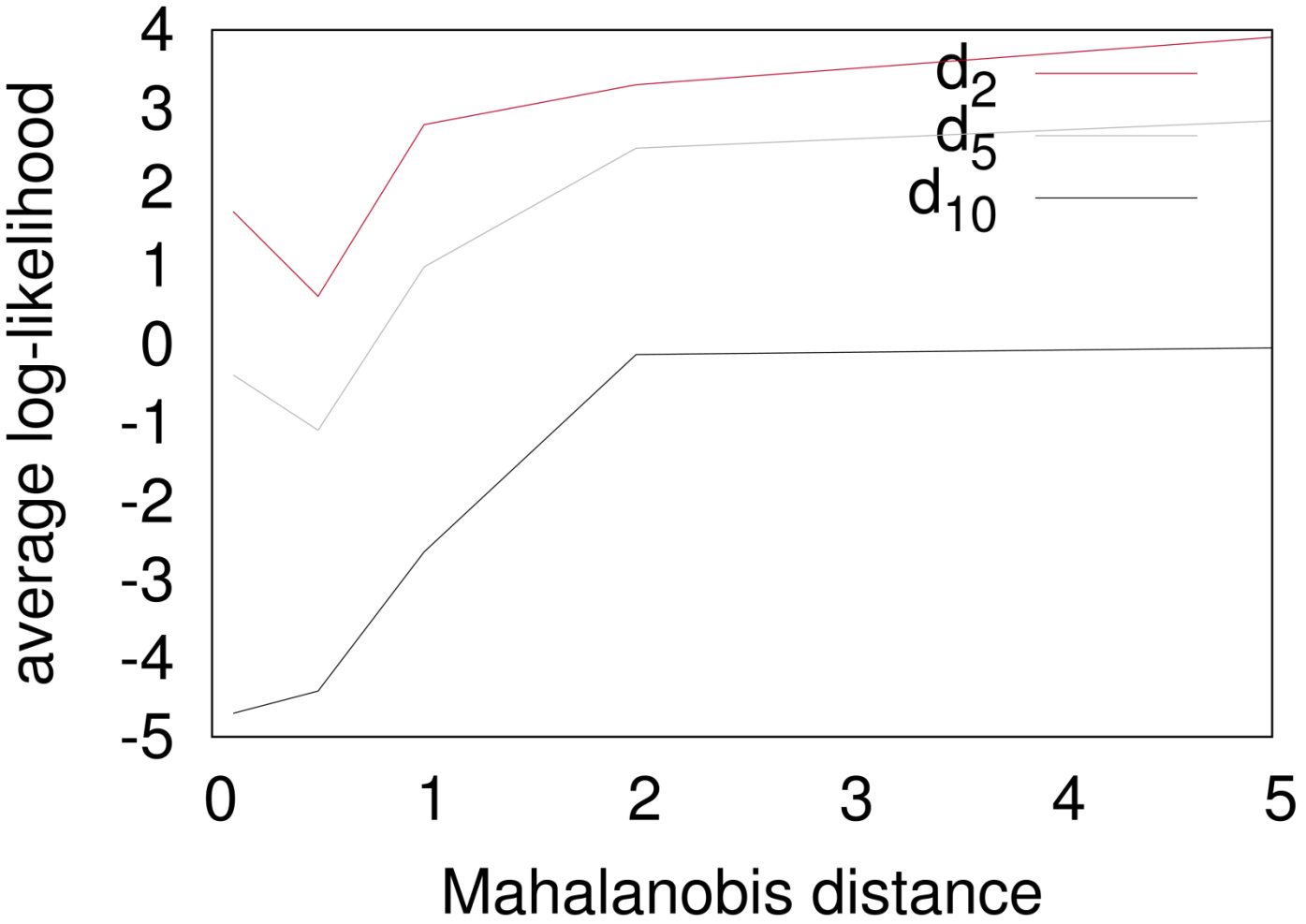
Evaluation: $|L|$



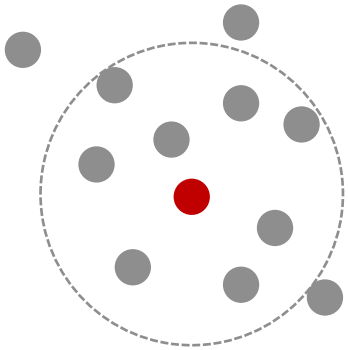


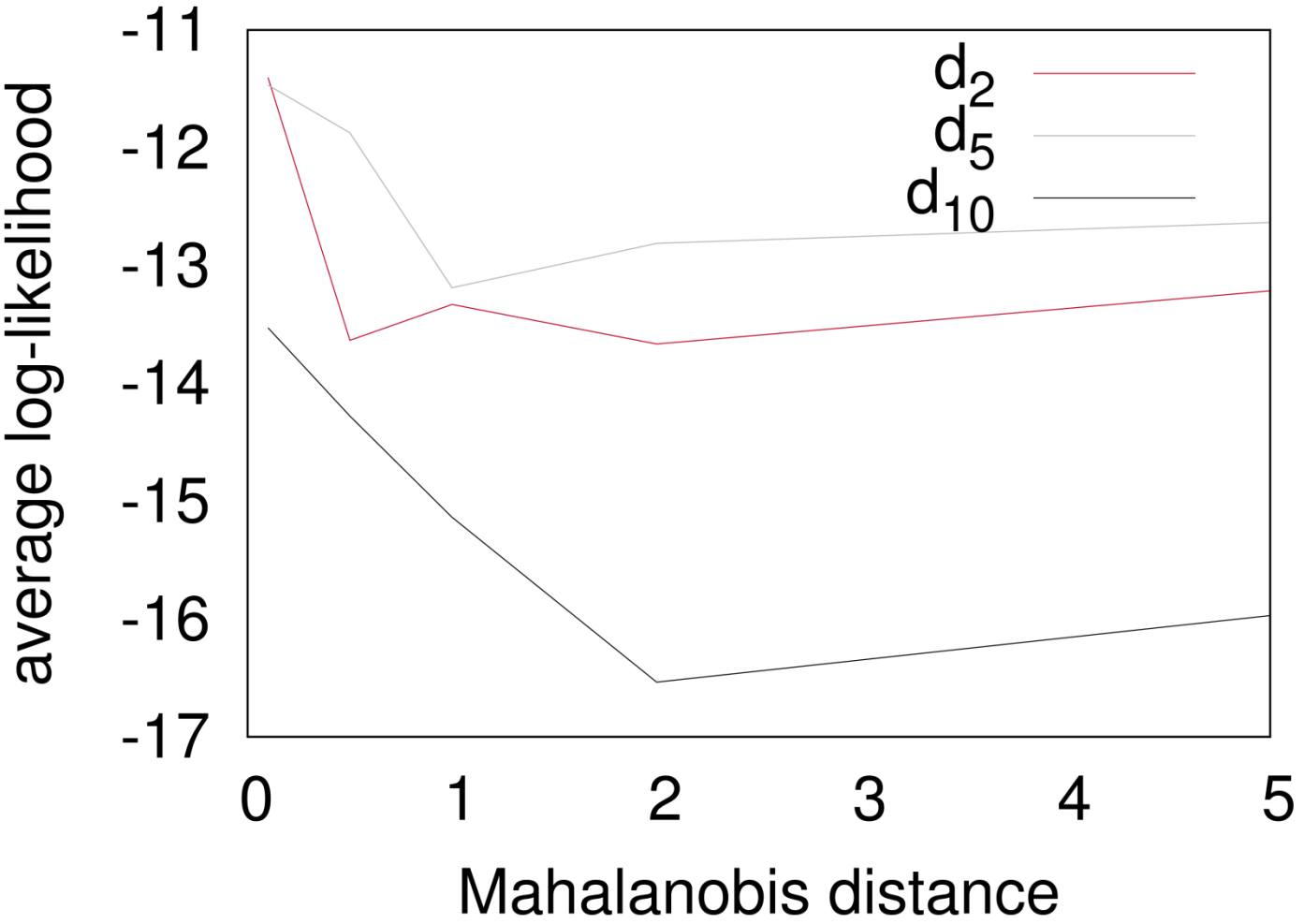
Evaluation: Mahalanobis (1 Gaussian)



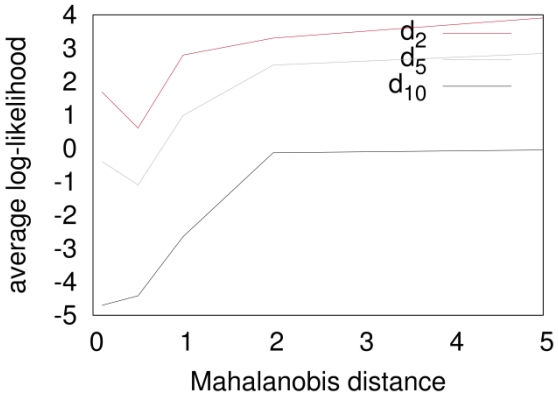


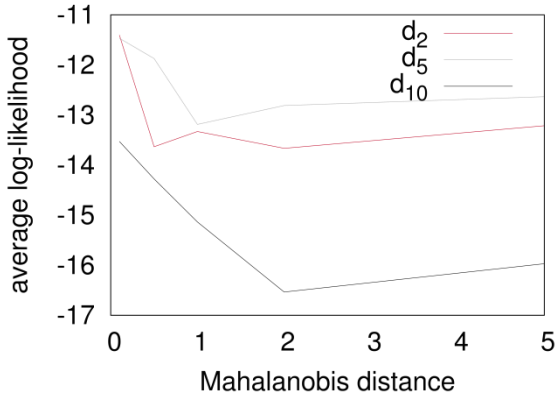
Evaluation: Mahalanobis (10 Gaussians)

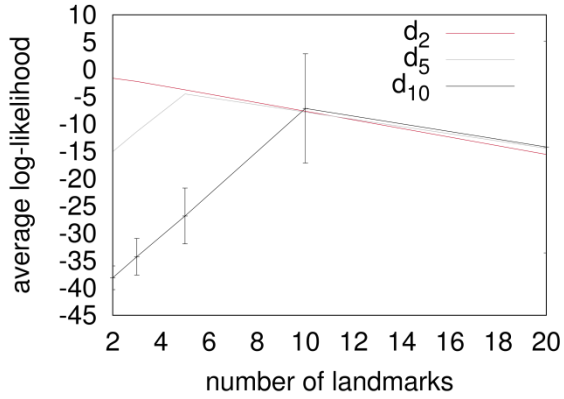




Evaluation: Parameter Setting







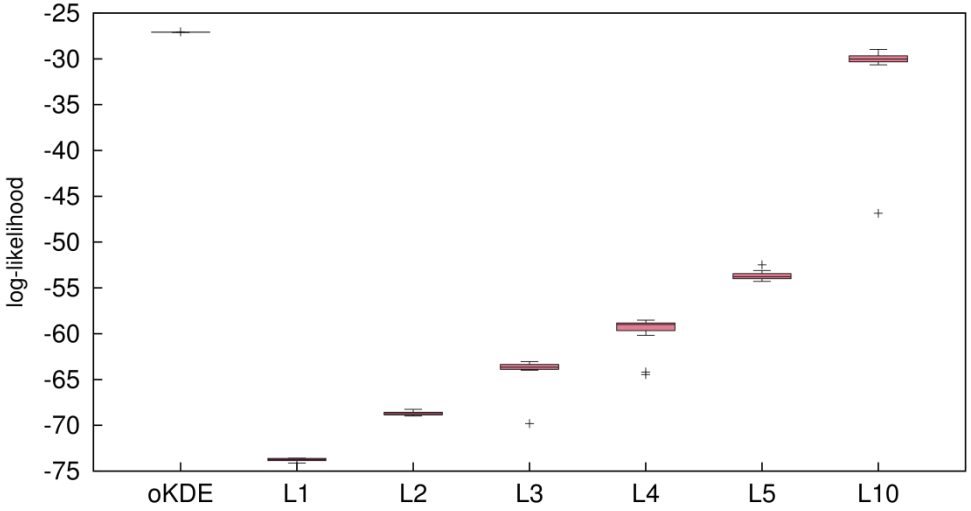
- $|L|$ depends on dimensionality of data
- small M partition the space better
- but at some point too few instances per region

Evaluation: Performance

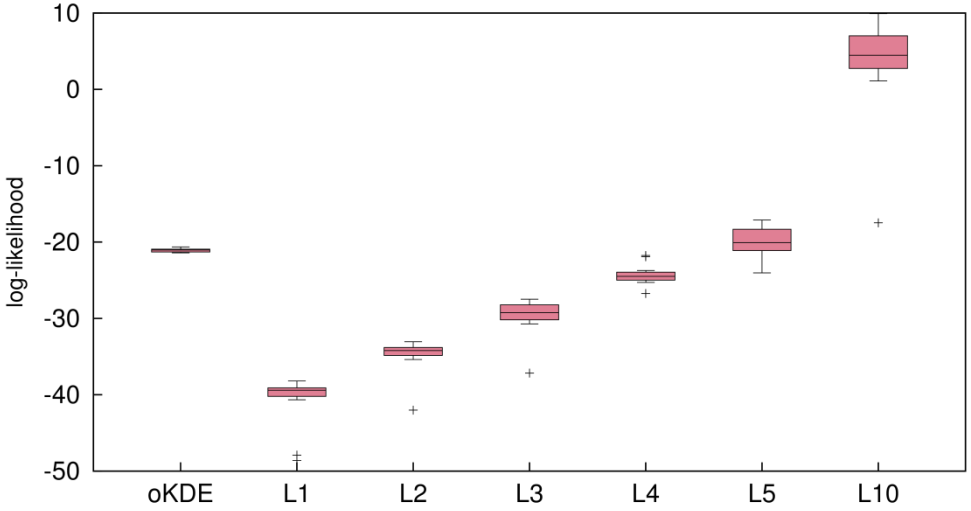
oKDE:

- online Kernel Density Estimator
- for multi-variate densities
- for continuous variables
- by Kristan et al. (2011)

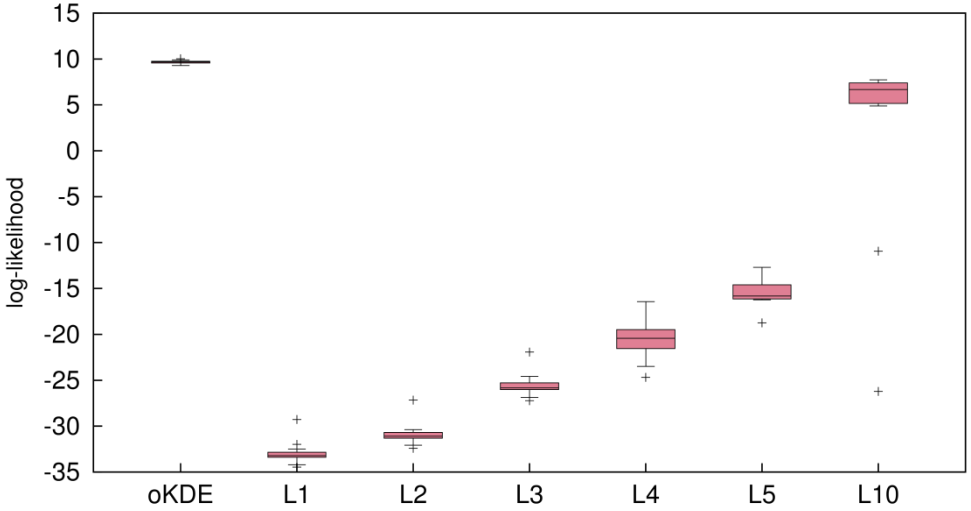
Datasets
Synthetic
Gaussian mixtures
Real-World
Covertype Electricity Letter Shuttle



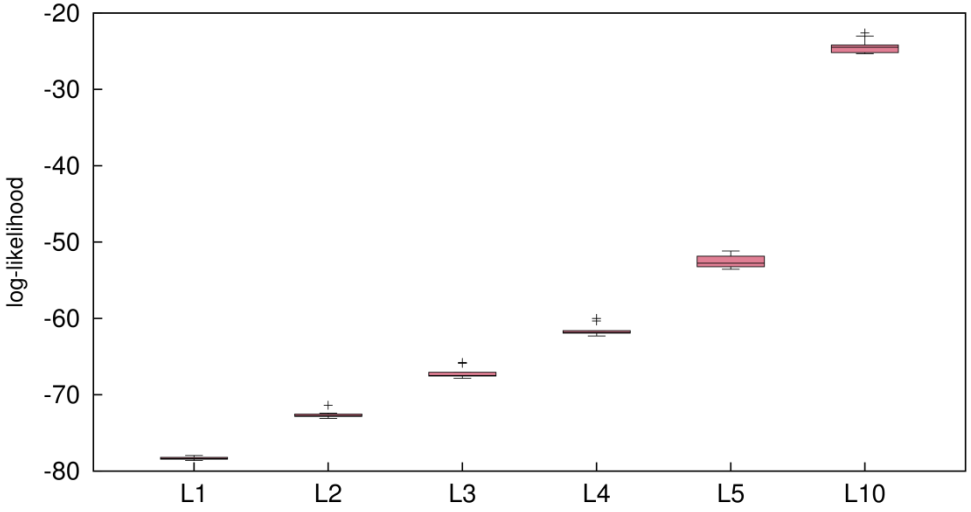
electricity (9 attributes)



shuttle (11 attributes)

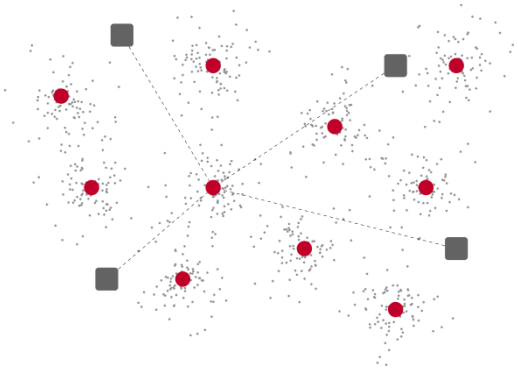


letter (17 attributes)



covertype (54 attributes)

Conclusions



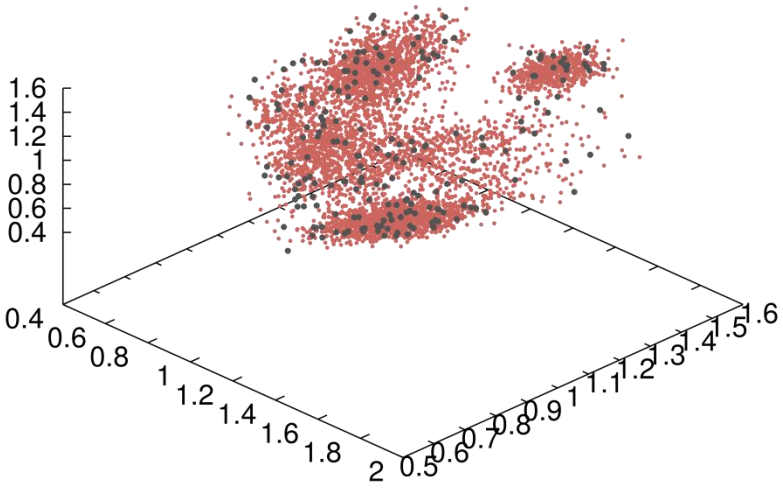
- online density estimation in higher dimensions
- heterogeneous data stream
- theoretical foundation
- comparable to the state of the art

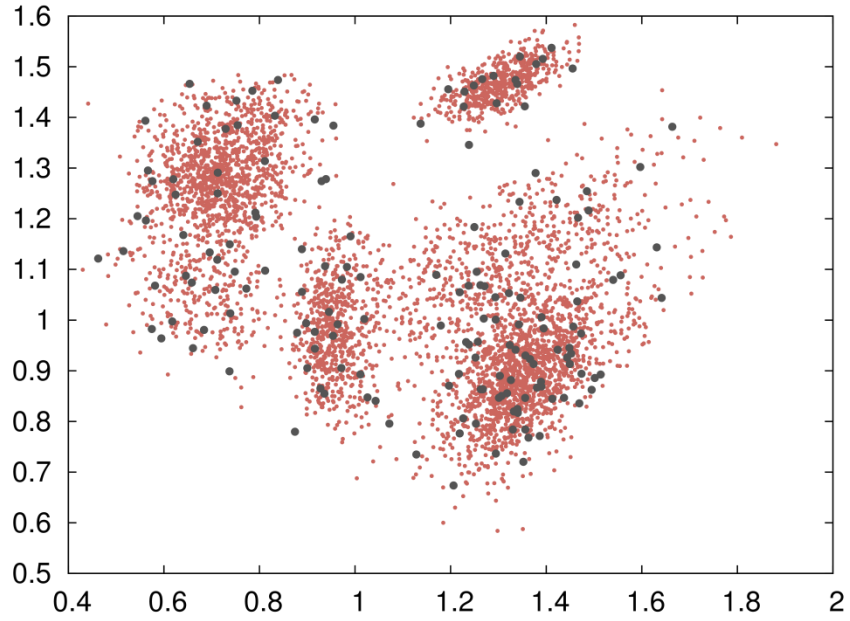
Future Work:

- new strategies for landmarks selection
- outlier detection
- detection of emerging trends

Thank you for your attention

Online Density Estimation using Representatives (RED)

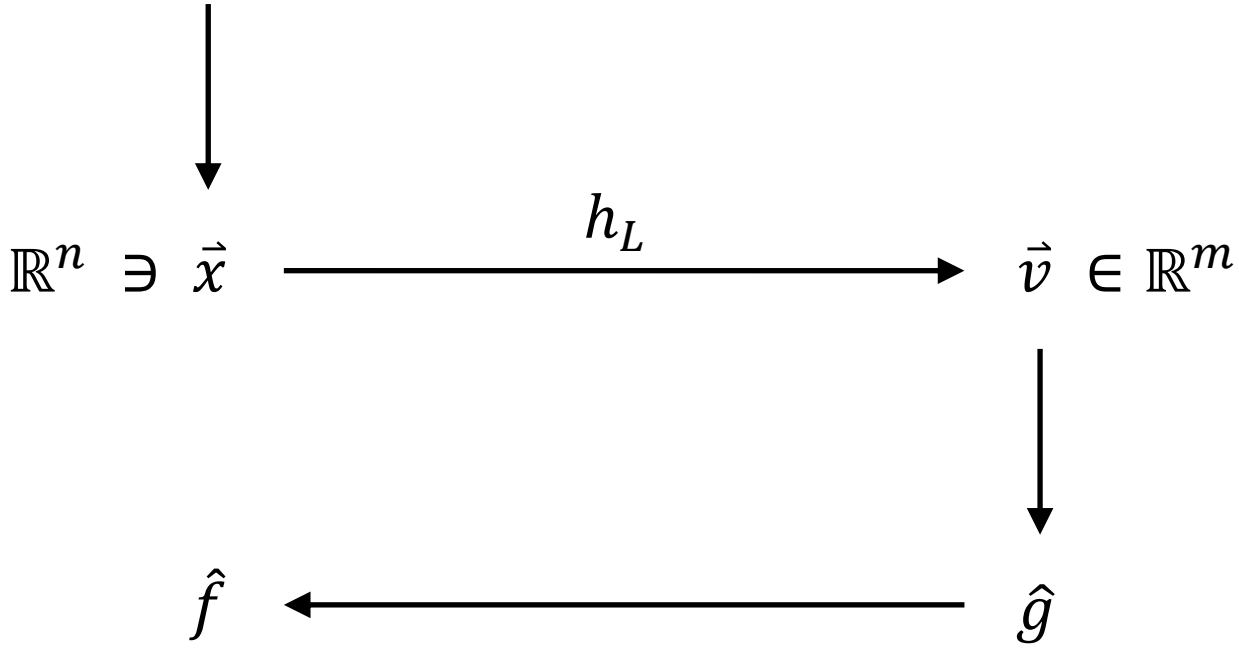




Online Density Estimation using Representatives (RED)

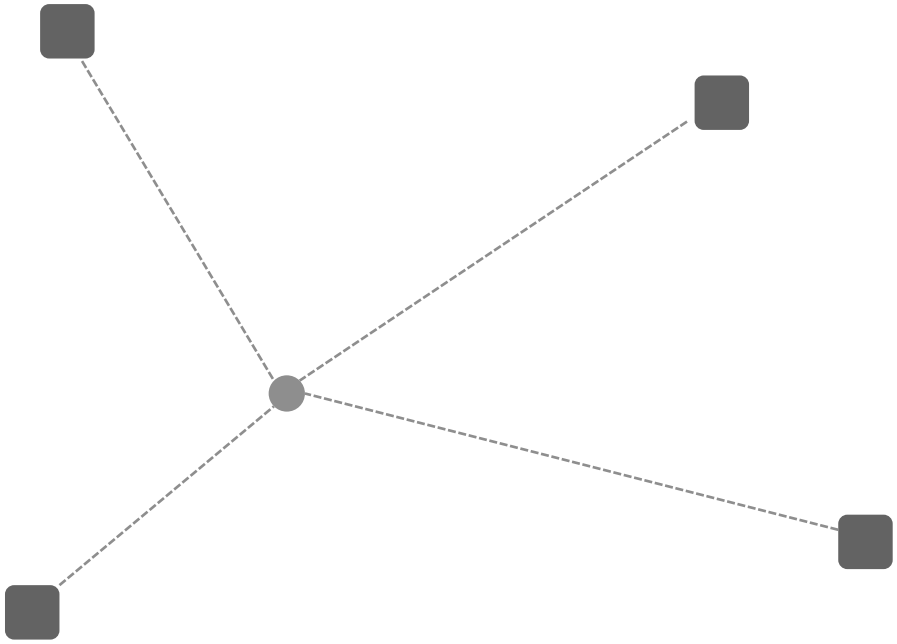
$X_1 \times X_2 \times \ldots \times X_n \ni \vec{x}$

$h_L(\vec{x}) = h_L(\vec{y})$ **but** $\vec{x} \neq \vec{y}$



$$\iiint_{-\infty}^{-\infty} \prod_{j=i}^{p} corr_j\left(x_j \middle| v_1, \ldots, v_p\right) \, dx_{i+1} dx_{i+2} \ldots d_n$$

Online Density Estimation using Representatives (RED)



landmark

instance

$$I = \{\vec{x} \in \mathbb{R}^n \mid h_L(\vec{x}) = \vec{v} = (v_1, v_2, v_3, v_4)\}$$