# ITCH: Information-theoretic Cluster Hierarchies

## Contributions

ITCH – a new hierarchical clustering approach on top of the idea of data compression which yields numerous advantages:

❖ All single clusters as well as their hierarchical arrangement are guaranteed to be **meaningful**.

❖ Each cluster is represented by an **intuitive description** of its content in form of a Gaussian PDF.

❖ **Outliers** are assigned to the root of the cluster hierarchy or to an appropriate inner node, depending on the degree of outlierness.

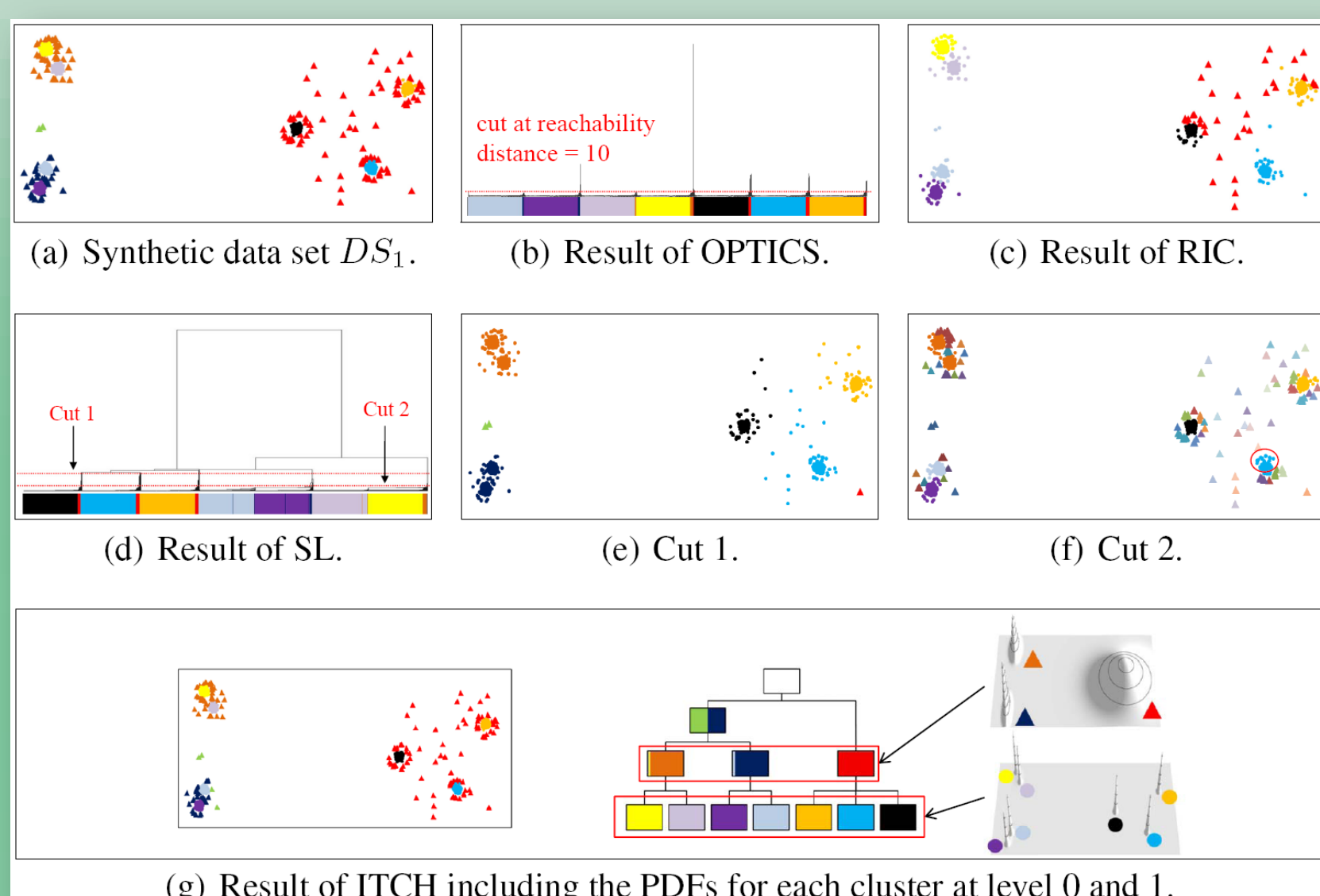❖ ITCH is **fully automatic** as no difficult parameter settings are necessary.
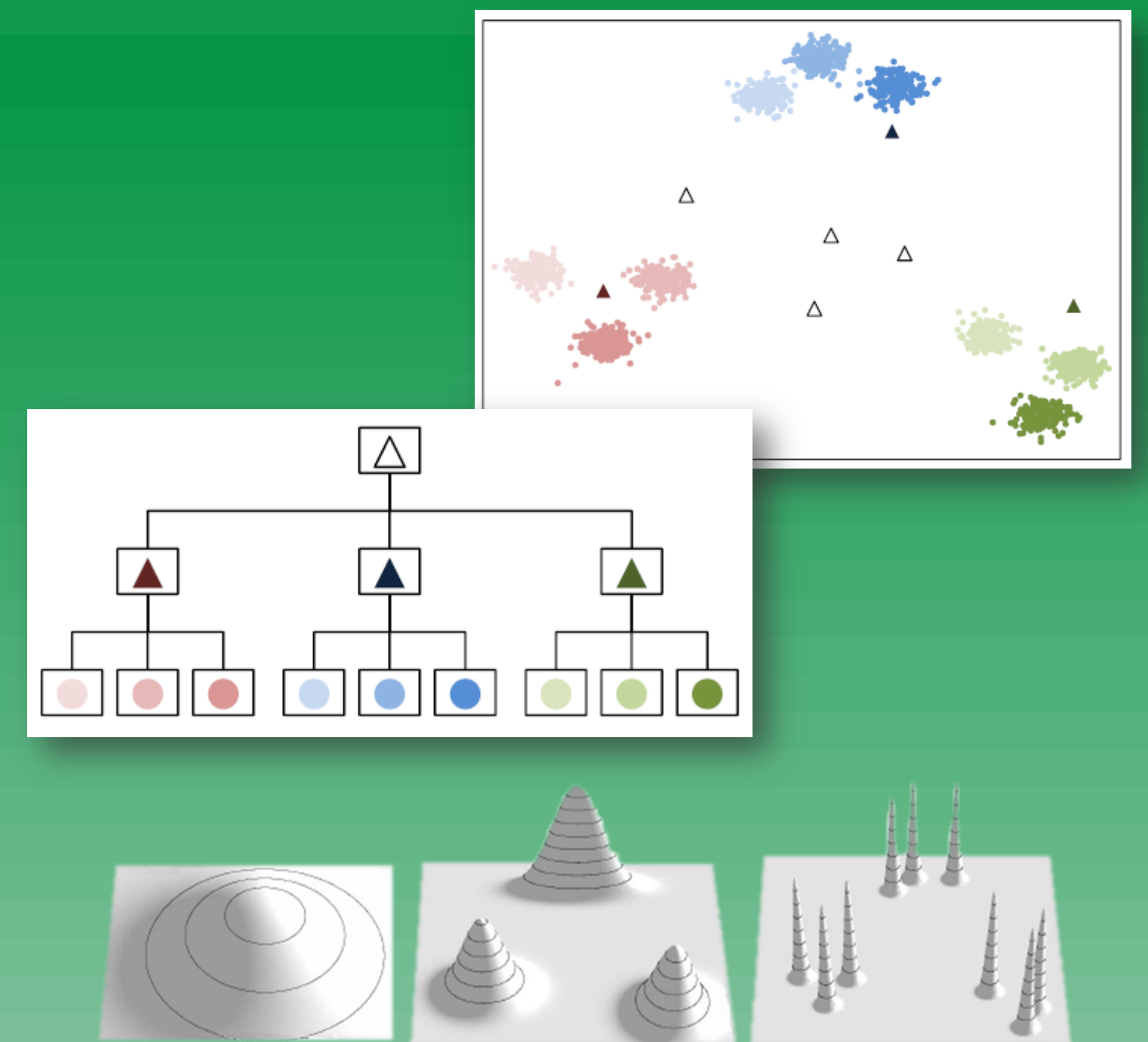


## hMDL: A MDL principle for hierarchical cluster structures
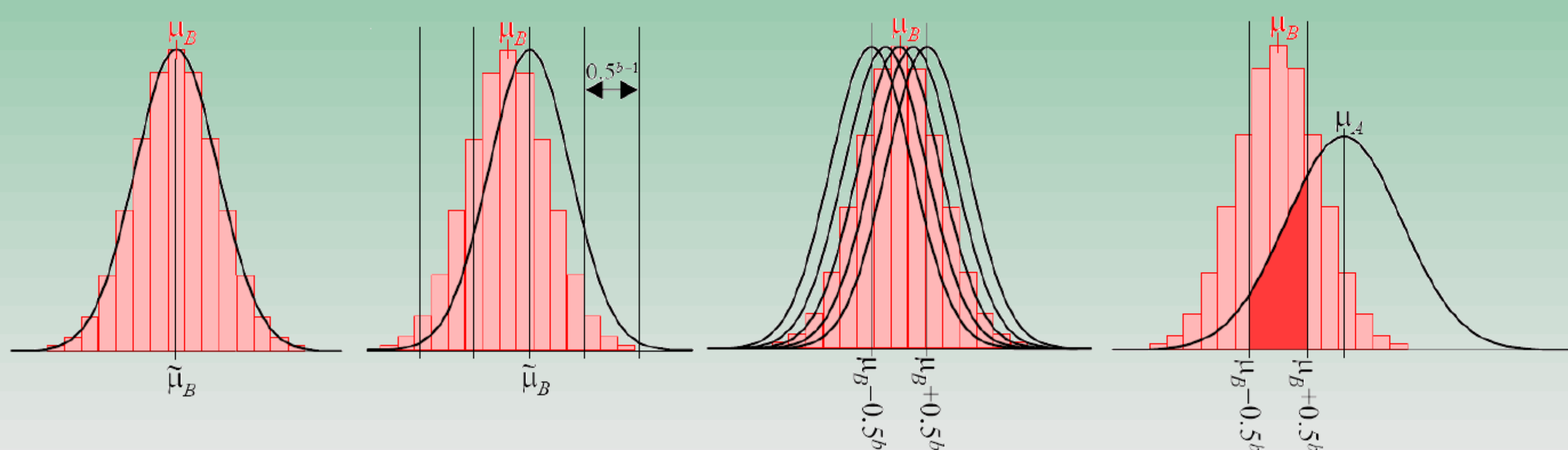


$$C_{\text{data}} = -\sum_{C \in \mathcal{N}} \left( nW_C \log_2 W_C + \sum_{x \in C;\, 1 \leq j \leq d} \log_2 N(\mu_{C,j}, \sigma^2_{C,j}, x_j) \right)$$

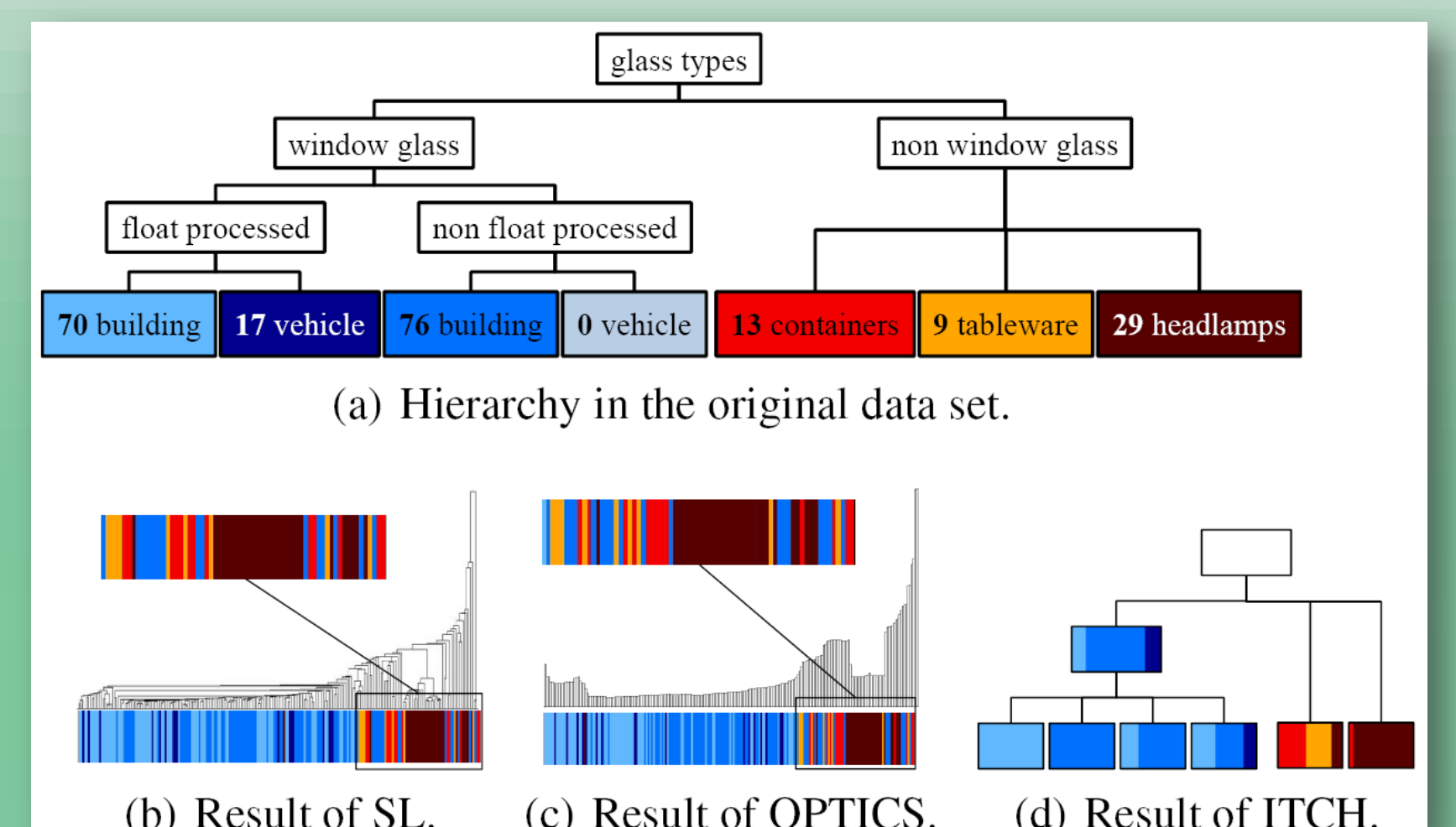$$hMDL = \sum_{C \in \mathcal{N}} \left( C_{\text{hMDL}}(C) - nW_C \log_2(W_C) - \log_2\left( \sum_{X \sqsubseteq \text{parent of } C} W_X \right) \right)$$
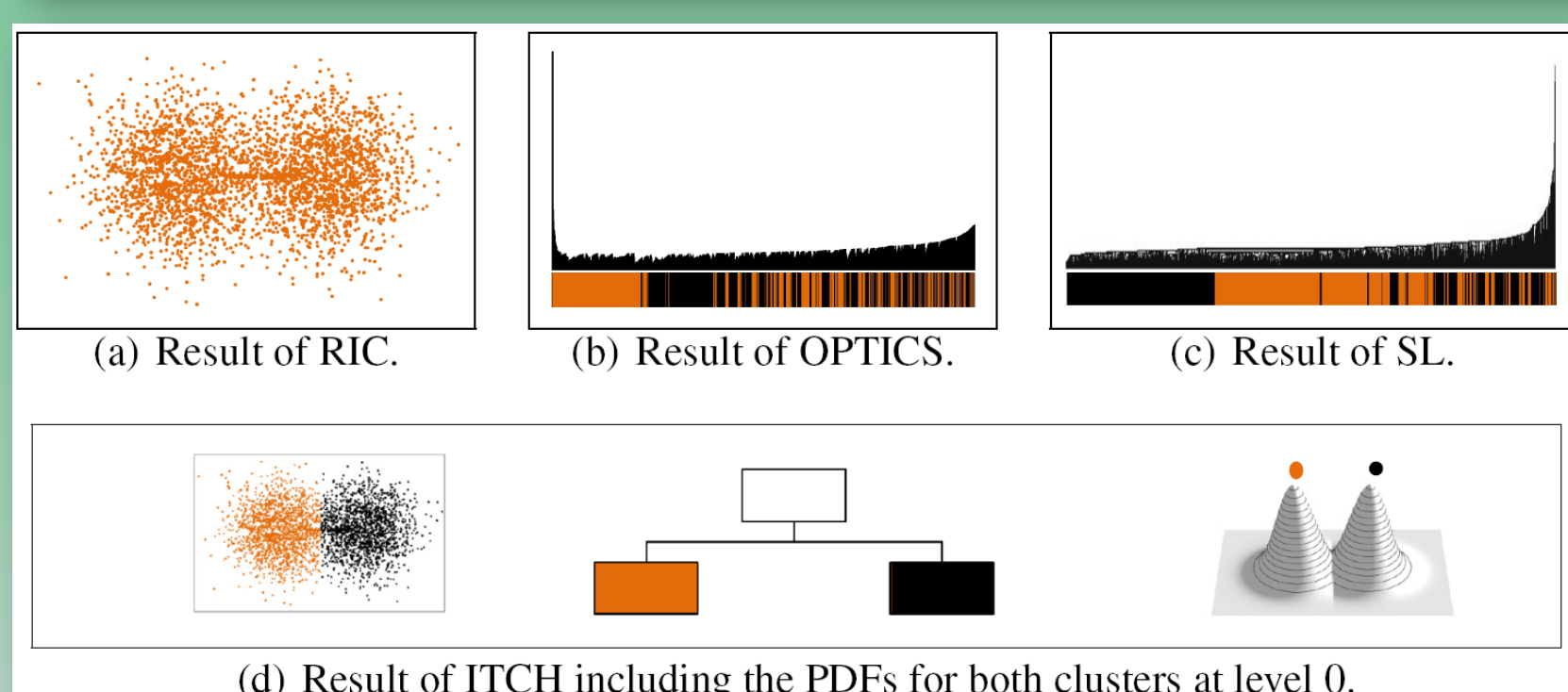
## Experimental Evaluation



(a) Synthetic data set $DS_1$.    (b) Result of OPTICS.    (c) Result of RIC.

(d) Result of SL.    (e) Cut 1.    (f) Cut 2.

(g) Result of ITCH including the PDFs for each cluster at level 0 and 1.

*Synthetic dataset DS$_1$*

|  | Precision | Recall |
|---|---|---|
| OPTICS | 94.8% | 95.4% |
| RIC | **94.9%** | 92.2% |
| SL (Cut 1) | 70.0% | 85.9% |
| SL (Cut 2) | 8.5% | 9.0% |
| ITCH | 93.8% | **97.5%** |



(a) Result of RIC.    (b) Result of OPTICS.    (c) Result of SL.

(d) Result of ITCH including the PDFs for both clusters at level 0.

*Synthetic dataset DS$_2$*

|  | Precision | Recall |
|---|---|---|
| RIC | 50.0% | 100.0% |
| ITCH | 99.2% | 99.7% |



(a) Hierarchy in the original data set.

| glass types | | |
|---|---|---|

window glass — non window glass

float processed — non float processed

| 70 building | 17 vehicle | 76 building | 0 vehicle | 13 containers | 9 tableware | 29 headlamps |

(b) Result of SL.    (c) Result of OPTICS.    (d) Result of ITCH.

*„Glass"dataset*



(a) Result of SL.    (b) Result of OPTICS.    (c) Result of ITCH.
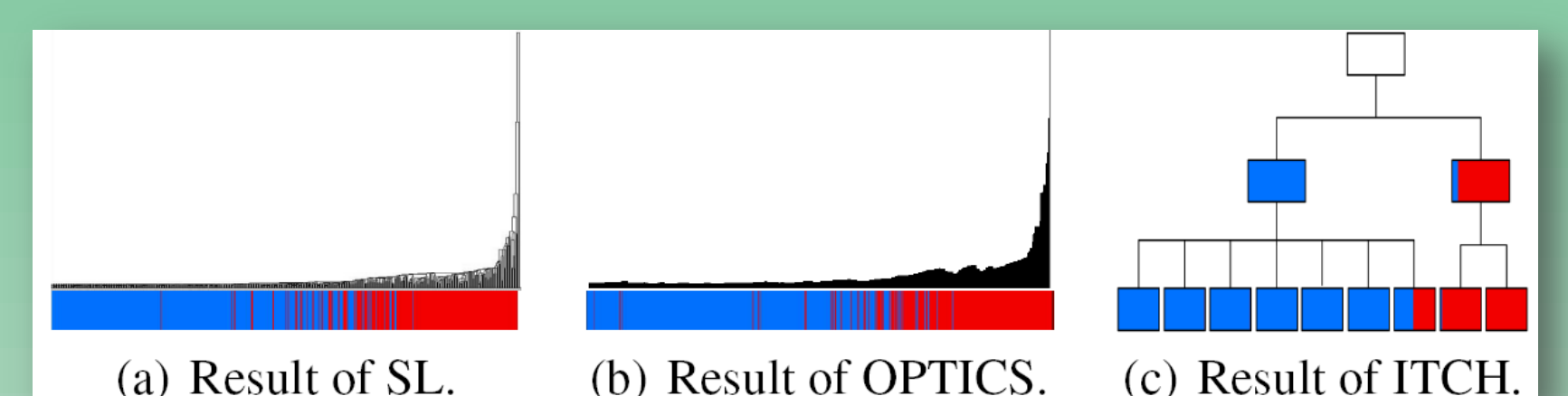
*„Cancer"dataset*

C. Böhm[1], F. Fiedler[1], A. Oswald[1], C. Plant[2], B. Wackersreuther[1], P. Wackersreuther[1]

Database Systems

[1] Department of Computer Science
Database Systems Group

[2] Florida State University