Clustering and Model Integration under the Wasserstein Metric

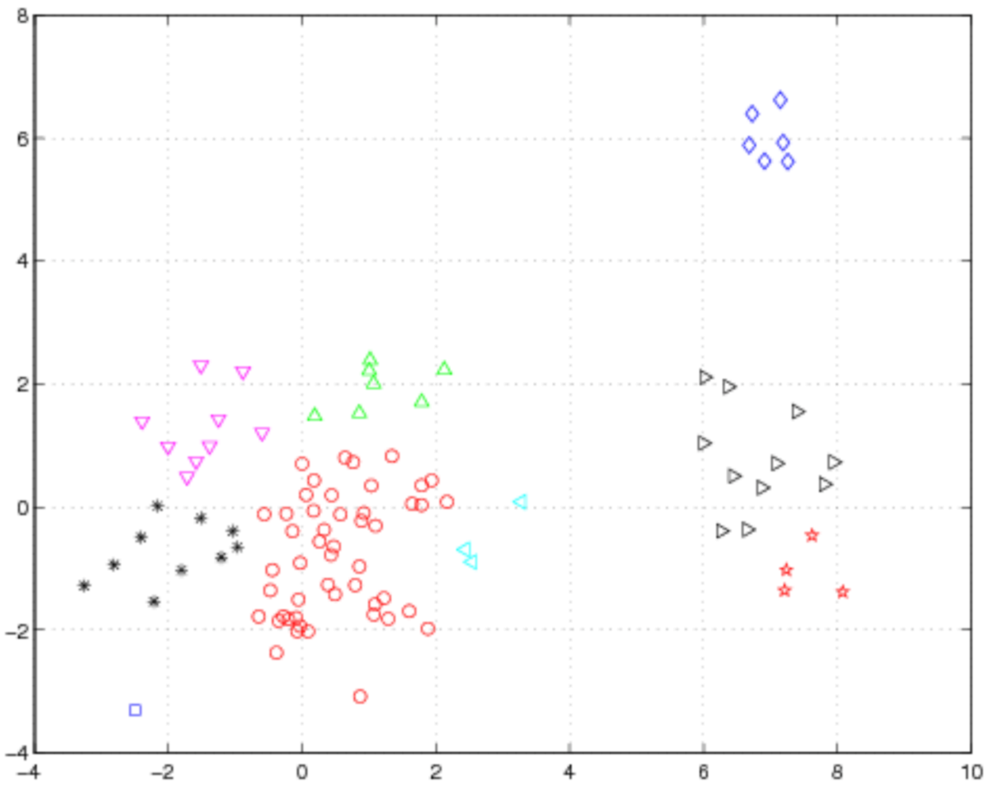
Jia Li

Department of Statistics

Penn State University

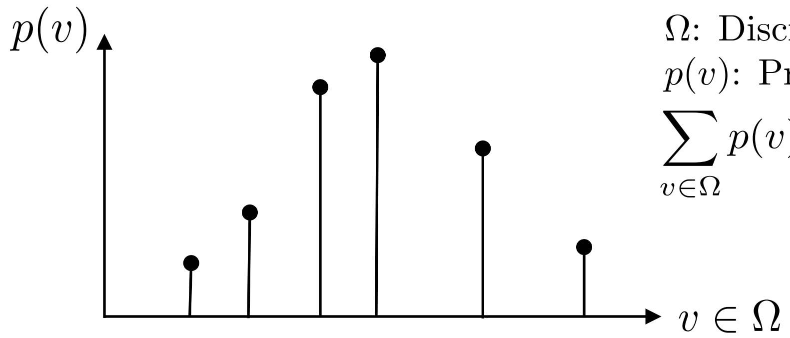
Clustering

- Data represented by vectors or pairwise distances.
- Methods
  - Top-down approaches
    - K-means
    - Statistical methods based on mixture models
  - Bottom-up approaches
    - Dendrogram clustering
  - Optimization vs. statistical modeling
- Applications
  - Exploratory study
  - Data reduction



Discrete Distributions

- Non-fixed support
- Generalize vectors
- Sparse histogram

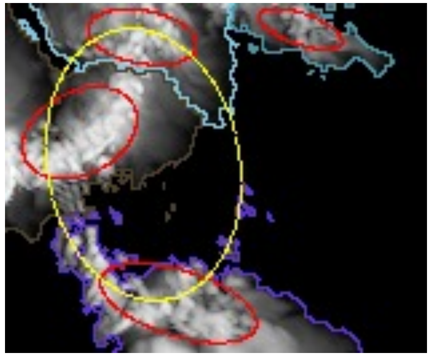


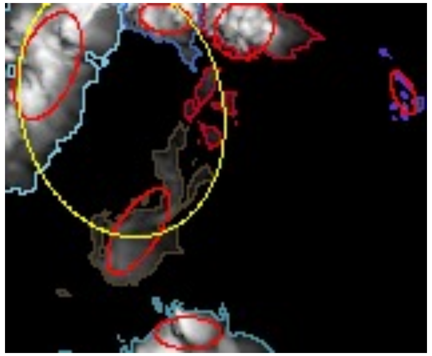
Ω: Discrete sample space
$p(v)$: Probability measure

$$\sum_{v\in\Omega} p(v) = 1$$

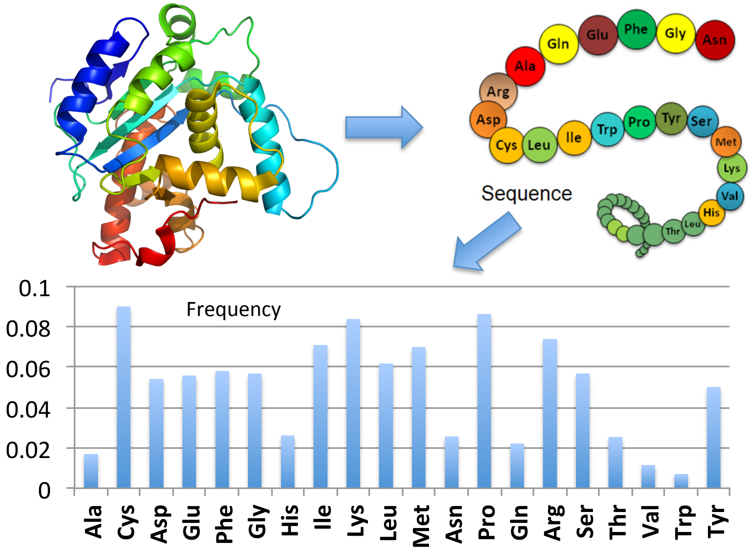
$$V = \{(v^{(1)}, p(v^{(1)})), \ldots, (v^{(t)}, p(v^{(t)}))\}$$

- Application 1: Cloud maps





- Application 2: Protein



Kantorovich-Wasserstein Metric

$$X \sim \mathcal{Q}_X, Y \sim \mathcal{Q}_Y, X, Y \in \mathcal{R}^d$$

$$D_P(\mathcal{Q}_X, \mathcal{Q}_Y) = \inf_{\mathcal{P}_{X,Y}:\mathcal{P}_X=\mathcal{Q}_X, \mathcal{P}_Y=\mathcal{Q}_Y} (E \parallel X - Y \parallel^p)^{\frac{1}{p}}$$

- Gaussian distributions:

$$D^2(\phi(\mu_1, \Sigma_1), \phi(\mu_2, \Sigma_2)) =$$
$$\|\mu_1 - \mu_2\|^2 + tr(\Sigma_1) + tr(\Sigma_2) - 2tr[(\sqrt{\Sigma_1}\Sigma_2\sqrt{\Sigma_1})^{1/2}] \, .$$

- Discrete $\mathcal{Q}_X$, $\mathcal{Q}_Y$: linear programming.

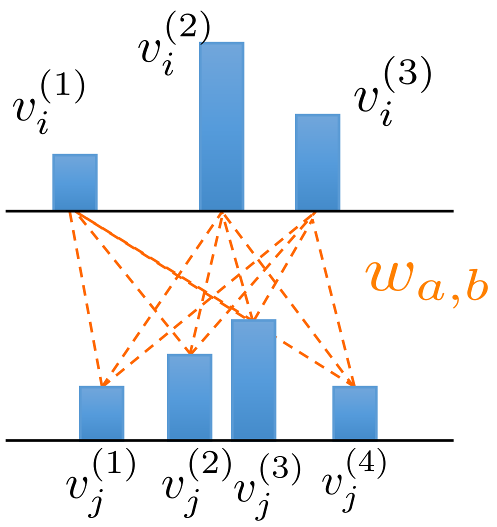
S. T. Rachev, L. Ruschendorf, *Mass Transportation Problems*, Springer, Probability and Its Applications, 1998.

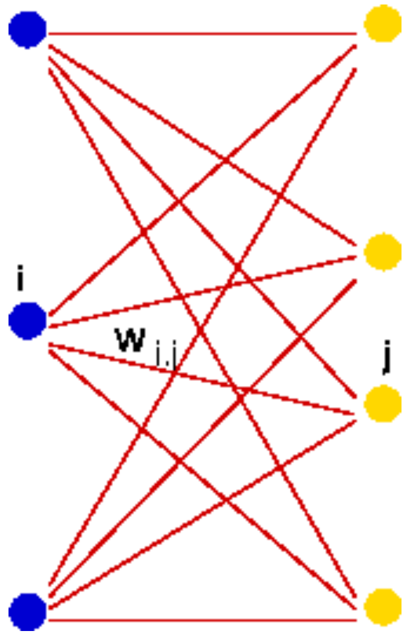
Wasserstein Distance for Discrete Distributions

- $\mathcal{P}^{(k)} = \{(x_1^{(k)}, w_1^{(k)}), (x_2^{(k)}, w_2^{(k)}), \ldots, (x_{m^{(k)}}^{(k)}, w_{m^{(k)}}^{(k)})\}$.
- The Wasserstein distance $D(\mathcal{P}^{(1)}, \mathcal{P}^{(2)})$ is

$$
\begin{aligned}
D^2(\mathcal{P}^{(1)}, \mathcal{P}^{(2)}) &= \min_{\{\pi_{i,j}\}} \sum_{i=1}^{m^{(1)}} \sum_{j=1}^{m^{(2)}} \pi_{i,j} \parallel x_i^{(1)} - x_j^{(2)} \parallel^2 \\
s.t. \quad \sum_{j=1}^{m^{(2)}} \pi_{i,j} &= w_i^{(1)}, \; i = 1, \ldots, m^{(1)}, \\
\sum_{i=1}^{m^{(1)}} \pi_{i,j} &= w_j^{(2)}, \; j = 1, \ldots, m^{(2)}, \\
\pi_{i,j} &\geq 0, \; i = 1, \ldots, m^{(1)}, j = 1, \ldots, m^{(2)}.
\end{aligned}
$$

Linear programming





Euclidean distance/cosine similarity	Wasserstein distance *(Kantorovich, 1942)*

• Fixed bins, orthogonal axes
• Pros:
    • Simple calculation
    • Abundant methodologies
• Cons:
    • Inefficient in high dimensions
    • Sensitive to vector quantization
	• Different supports
	• Pros:
	    • Directly applies to discrete distributions
	    • Good to measure sparse and high dimensional data
	• Cons:
	    • Complex computation

D2-Clustering

$B = \{\mathcal{P}^{(i)} : \mathcal{P}^{(i)} \in \Omega, i = 1, ..., n\}$

Optimize: $A = \{\mathcal{Q}^{(i)} : \mathcal{Q}^{(i)} \in \Omega, i = 1, ..., \bar{n}\}$, and cluster assignment $c(i) \in \{1, 2, ..., \bar{n}\}$, $i = 1, ..., n$.

Optimization Criterion

$$L(B, A^*, c^*) = \min_{A} \min_{c} \sum_{i=1}^{n} D^2(\mathcal{P}^{(i)}, \mathcal{Q}^{(c(i))})$$

K-means $\longrightarrow$ D2-Clustering

(vectors) $\longrightarrow$ (bags of weighted vectors)

D2-Clustering Algorithm

1. For every instance i, set

$$c(i) = \arg\min_{j=1,\ldots,\bar{n}} D^2(\mathcal{P}^{(i)}, \mathcal{Q}^{(j)})\,.$$

2. For each cluster j, $\mathcal{C}_j = \{i : c(i) = j\}$, $j = 1, \ldots, \bar{n}$, solve

$$\mathcal{Q}^{(j)} = \arg\min_{\mathcal{Q}\in\Omega} \sum_{i\in\mathcal{C}_j} D^2(\mathcal{P}^{(i)}, \mathcal{Q}) \quad \longleftarrow \textit{challenging}$$

Wasserstein Centroid Problem

Given a set of distributions $\{\mathcal{P}^{(1)}, \ldots, \mathcal{P}^{(N)}\}$, solve the centroid $\mathcal{Q} = \{(w_1, x_1), \ldots, (w_m, x_m)\}$, such that

$$\min_{\mathcal{Q}} \frac{1}{N} \sum_{k=1}^{N} D^2(\mathcal{Q}, \mathcal{P}^{(k)}) = \min_{\mathbf{x},\mathbf{w}} \sum_{k=1}^{N} \min_{\Pi^{(k)}} \sum_{i \in \mathcal{I}', j \in \mathcal{I}_k} \pi_{i,j}^{(k)} \|x_i - x_j^{(k)}\|^2$$

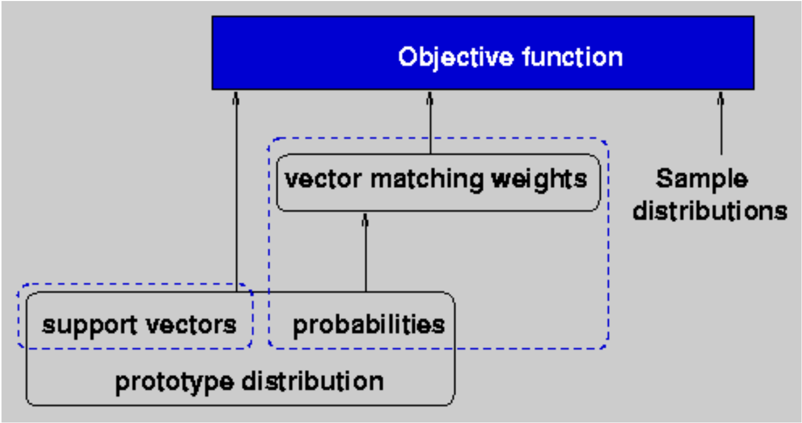
1. $\mathbf{x} = (x_1, \ldots, x_m) \in \mathbb{R}_{d \times m}$, $\mathbf{w} = (w_1, \ldots, w_m) \in \mathbb{R}_m^+$.
2. $\mathbf{x}^{(k)} = (x_1^{(k)}, \ldots, x_{m^{(k)}}^{(k)}) \in \mathbb{R}_{d \times m^{(k)}}$, $k = 1, \ldots, N$.
3. $\Pi^{(k)} = (\pi_{i,j}^{(k)}) \in \mathbb{R}_{m \times m^{(k)}}^+$, $k = 1, \ldots, N$.
4. $\mathbf{X} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}) \in \mathbb{R}_{d \times n}$, where $n = \sum_{k=1}^{N} m^{(k)}$.
5. $\Pi = (\Pi^{(1)}, \ldots, \Pi^{(N)}) \in \mathbb{R}_{m \times n}$.
6. Index set $\mathcal{I}^c = \{1, \ldots, N\}$, $\mathcal{I}_k = \{1, \ldots, m^{(k)}\}$, for $k \in \mathcal{I}^c$, and $\mathcal{I}' = \{1, \ldots, m\}$.

With $\mathbf{w}$ and Π fixed,

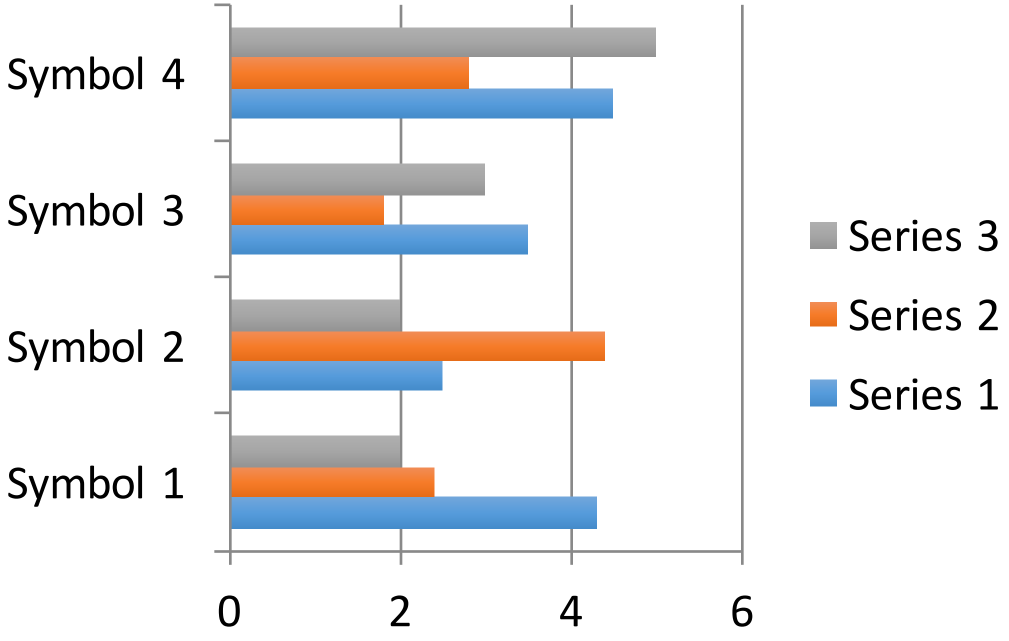
$$x_i := \frac{1}{N w_i} \sum_{k=1}^{N} \sum_{j=1}^{m^{(k)}} \pi_{i,j}^{(k)} x_j^{(k)}, \quad i \in \mathcal{I}', \tag{1}$$

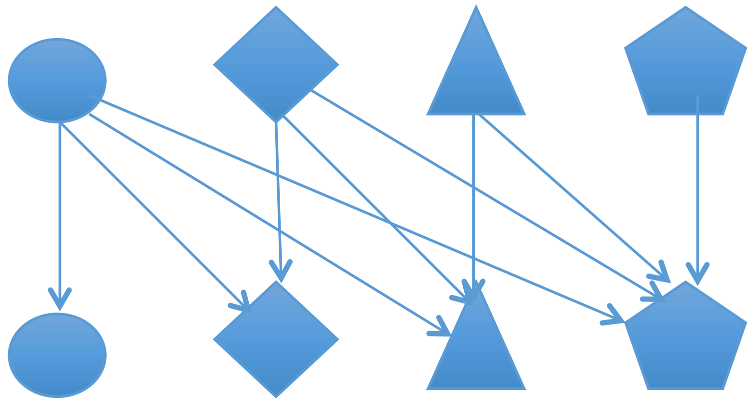
With fixed $\mathbf{x}$, updating $\mathbf{w}$ and Π is a large-scale LP:

$$\begin{aligned}
\min_{\Pi, \mathbf{w}} \quad & \textstyle\sum_{k=1}^{N} \sum_{i \in \mathcal{I}', j \in \mathcal{I}_k} \pi_{i,j}^{(k)} \|x_i - x_j^{(k)}\|^2 , \qquad (2) \\
s.t. \quad & \textstyle\sum_{l=1}^{m^{(k)}} \pi_{i,l}^{(k)} = w_i , \quad \forall k \in \mathcal{I}^c, \ i \in \mathcal{I}' , \\
& \textstyle\sum_{l=1}^{m} \pi_{l,j}^{(k)} = w_j^{(k)} , \quad \forall k \in \mathcal{I}^c, \ j \in \mathcal{I}_k , \\
& \textstyle\sum_{l=1}^{m} w_l = 1, \quad w_i \geqslant 0, \\
& \pi_{i,j}^{(k)} \geqslant 0 , \quad \forall k \in \mathcal{I}^c, \ i \in \mathcal{I}', \ j \in \mathcal{I}_k .
\end{aligned}$$



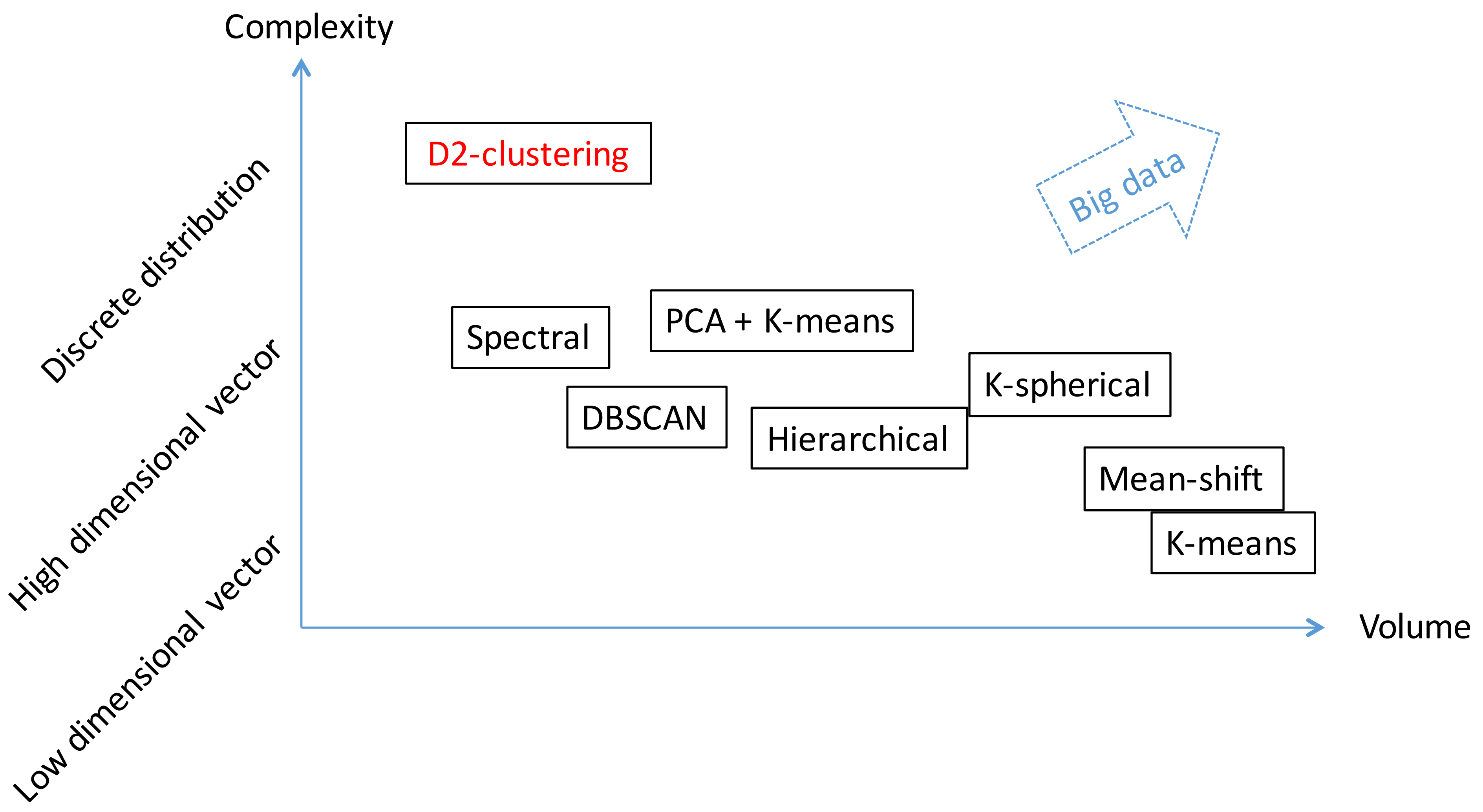
Extension to Symbolic Data





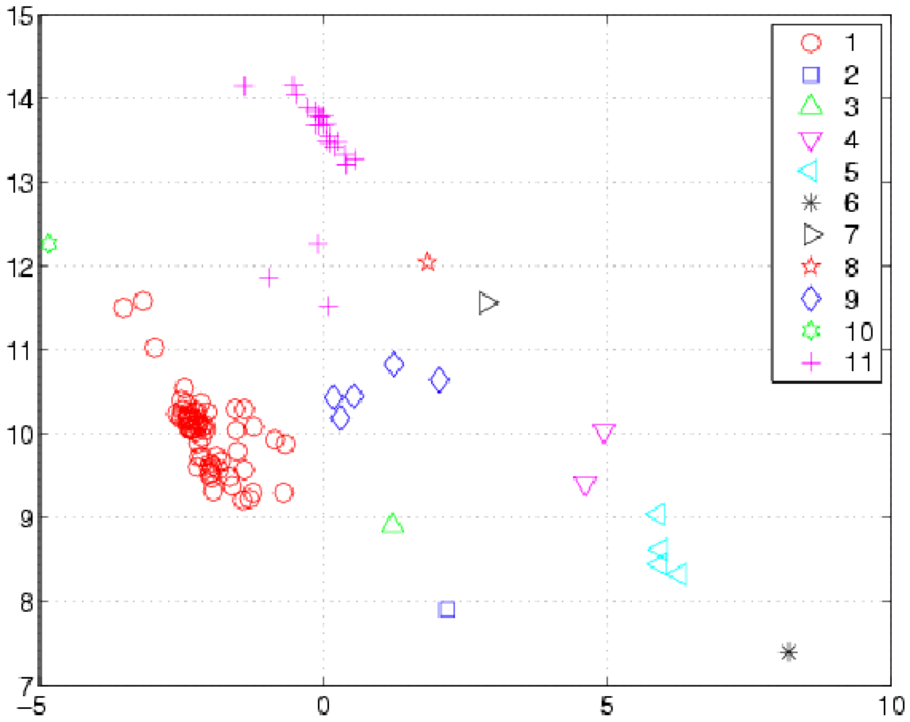
Pairwise distances between symbols are accounted for.

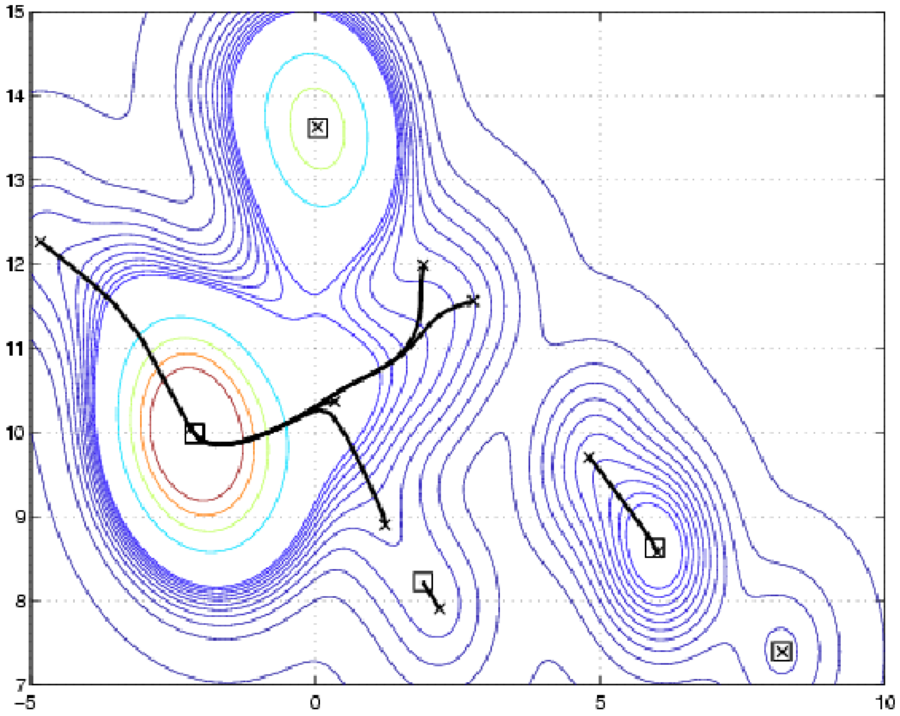
Existing Clustering Approaches in Big Data



Modal EM for Clustering: HMAC

- Finite mixture model $f(x) = \sum_{k=1}^{K} a_k f_k(x) = \sum_{k=1}^{K} a_k \phi(x \mid \mu_k, \Sigma_k)$



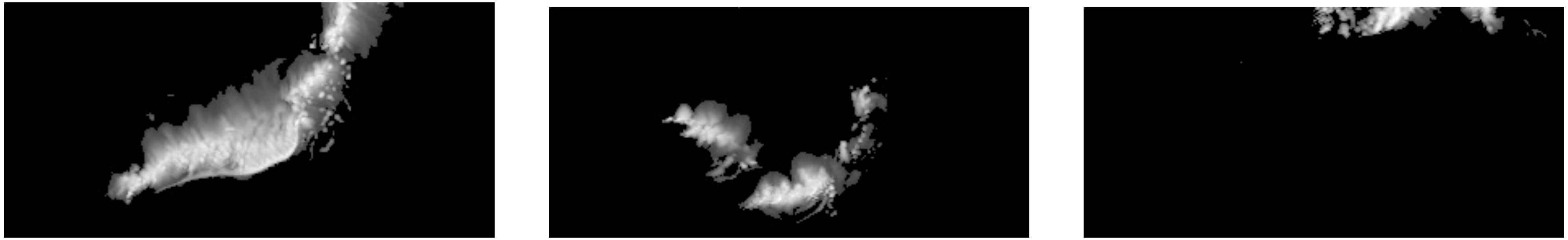


J. Li, S. Ray, B. G. Lindsay, "A nonparametric statistical approach to clustering via mode identification," **Journal of Machine Learning Research,** 8(8):1687-1723, 2007.

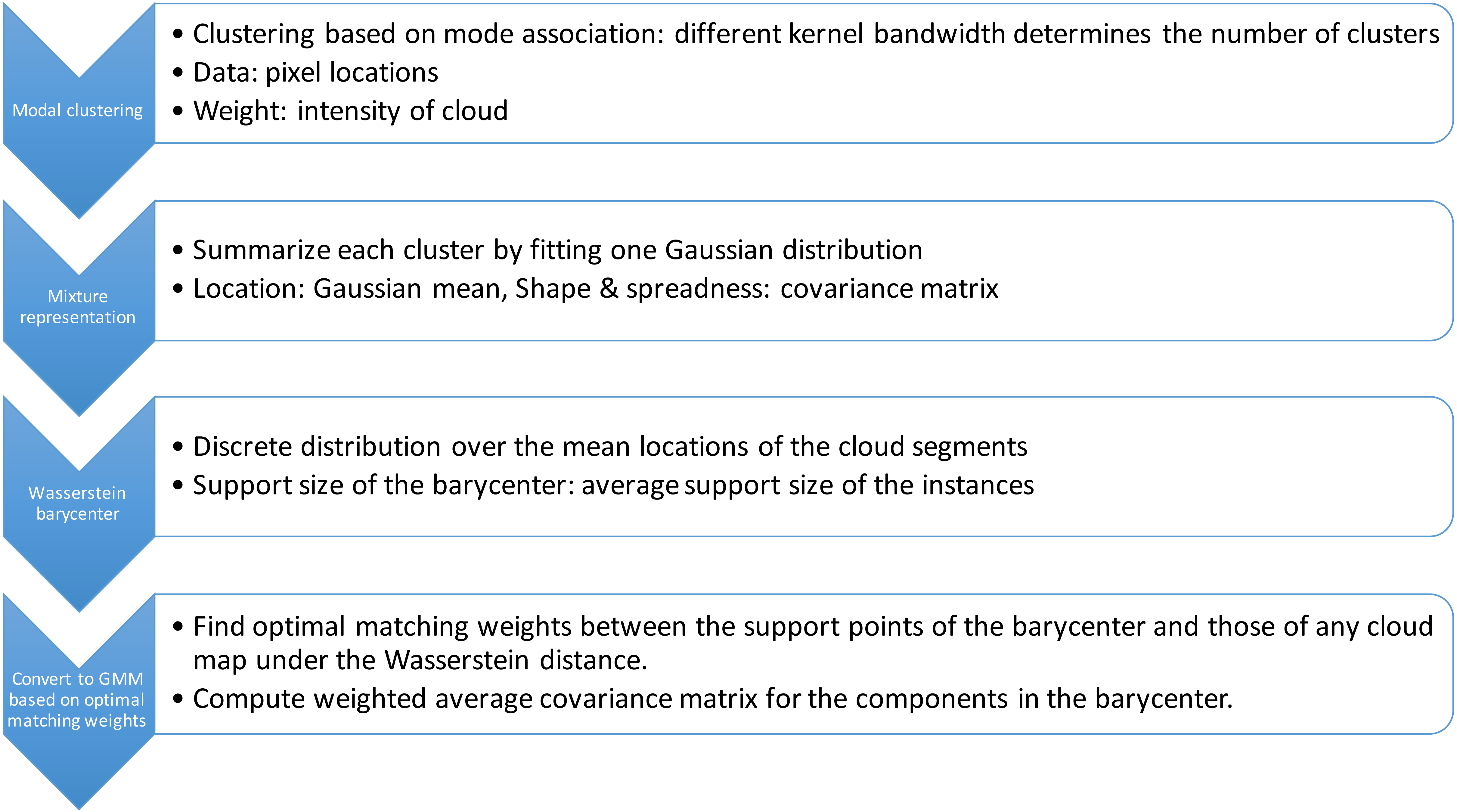
Cloud Map Synthesis

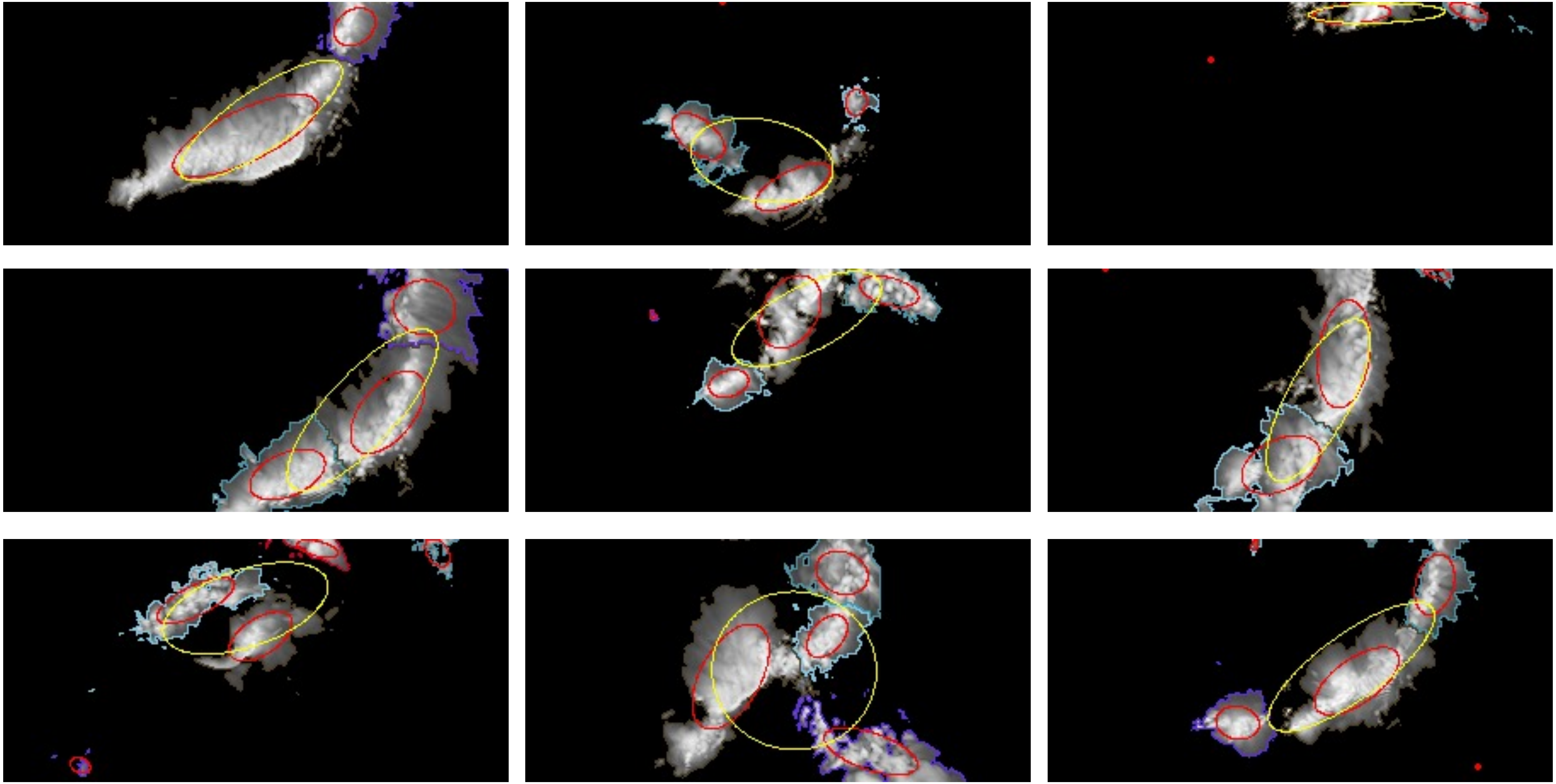
- Different models generate a set of cloud maps.
- How to synthesize the cloud maps?

Set 1

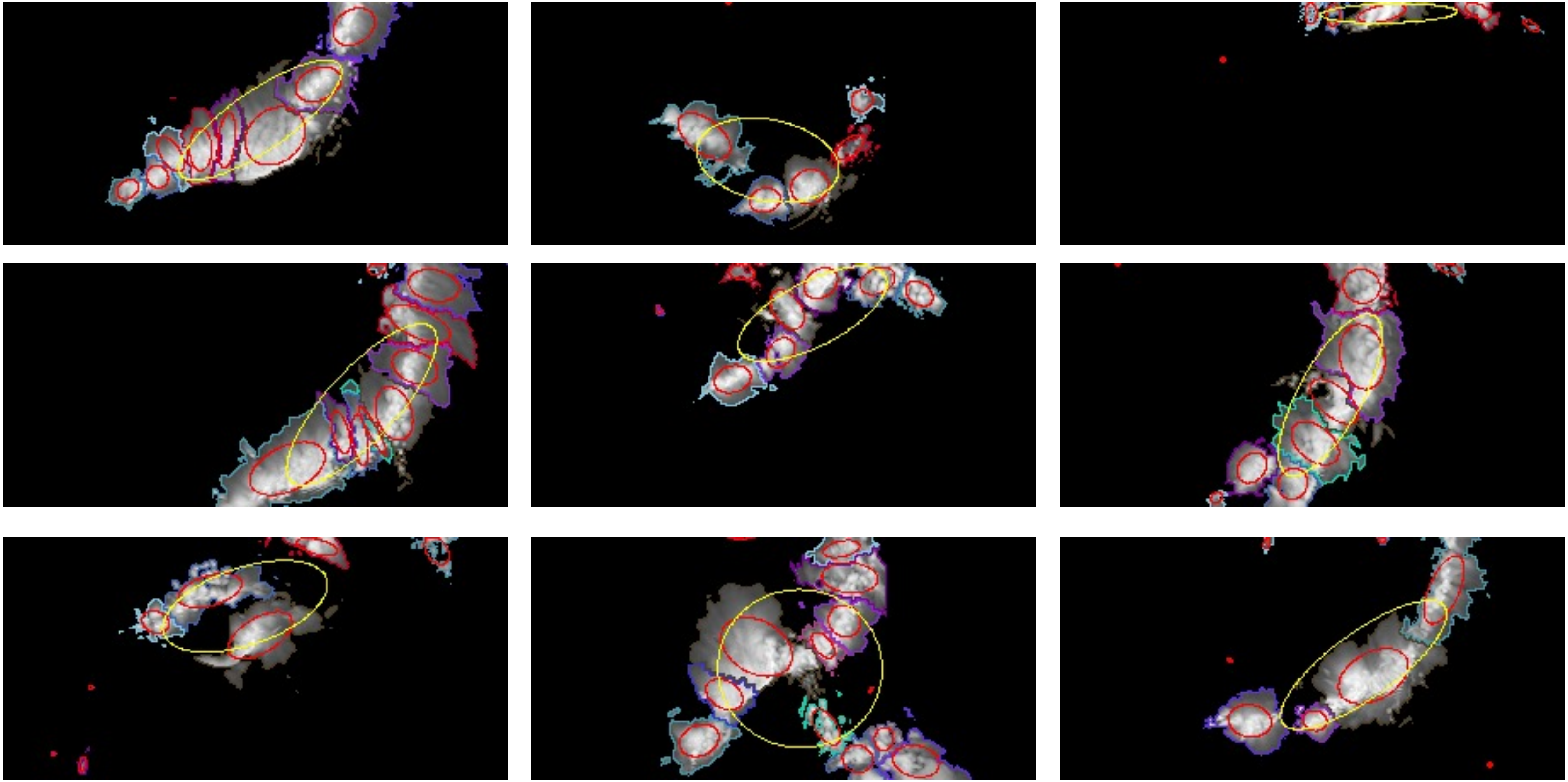


Set 2

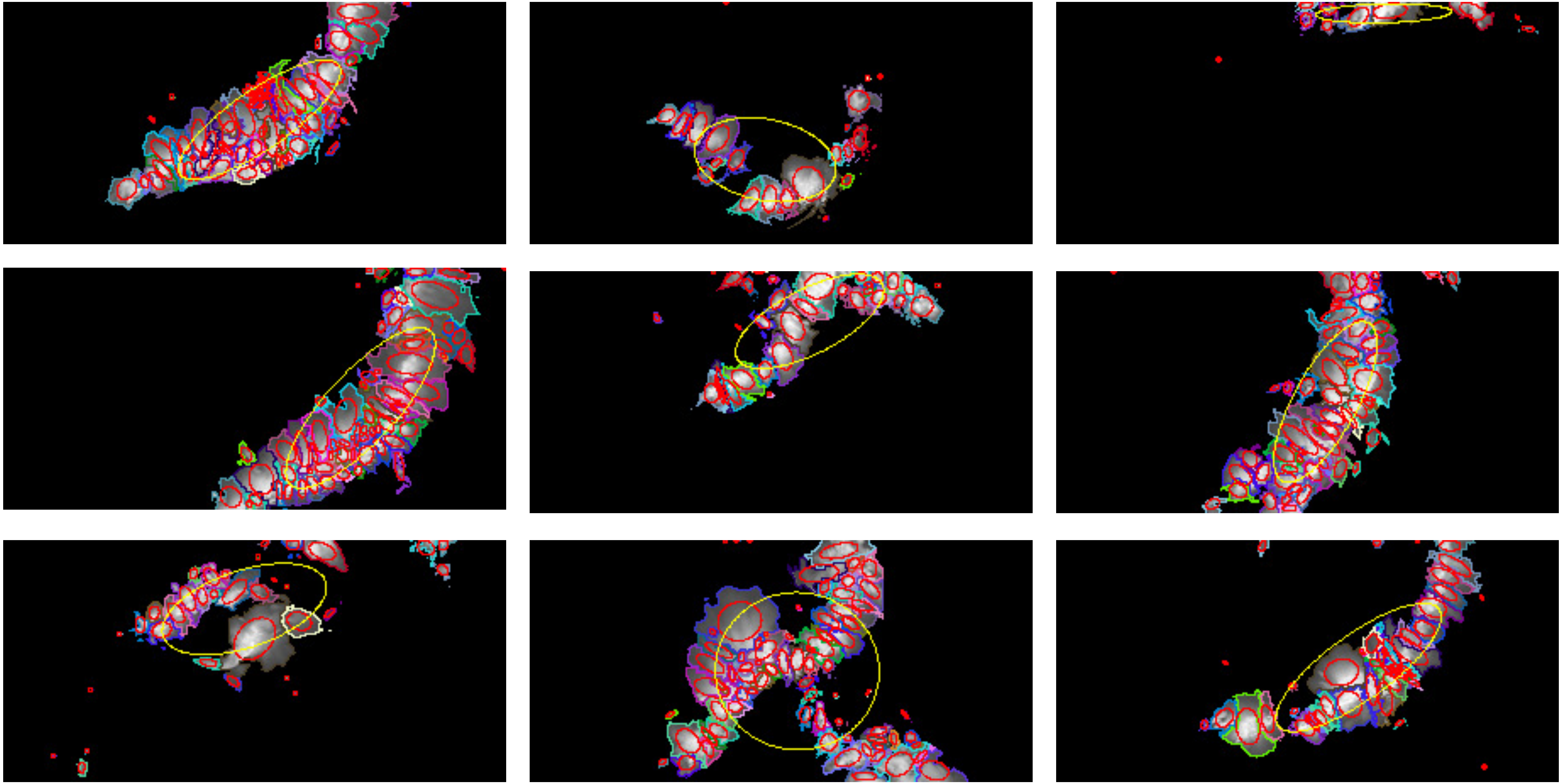

Modal clustering
• Clustering based on mode association: different kernel bandwidth determines the number of clusters
• Data: pixel locations
• Weight: intensity of cloud
Mixture representation
• Summarize each cluster by fitting one Gaussian distribution
• Location: Gaussian mean, Shape & spreadness: covariance matrix
Wasserstein barycenter
• Discrete distribution over the mean locations of the cloud segments
• Support size of the barycenter: average support size of the instances
Convert to GMM based on optimal matching weights
• Find optimal matching weights between the support points of the barycenter and those of any cloud map under the Wasserstein distance.
• Compute weighted average covariance matrix for the components in the barycenter.

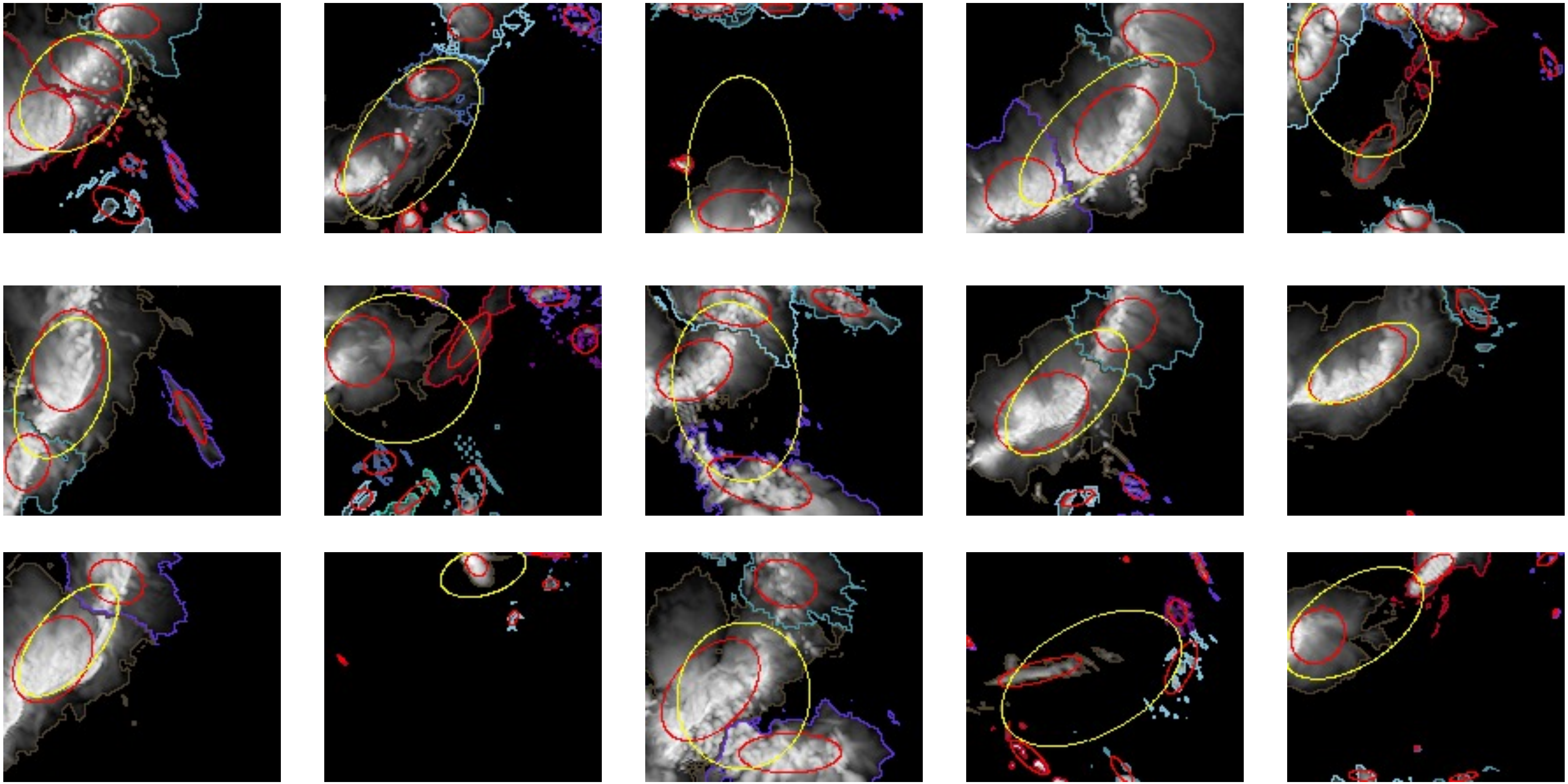


Summary based on HMAC: relatively aggressive merging



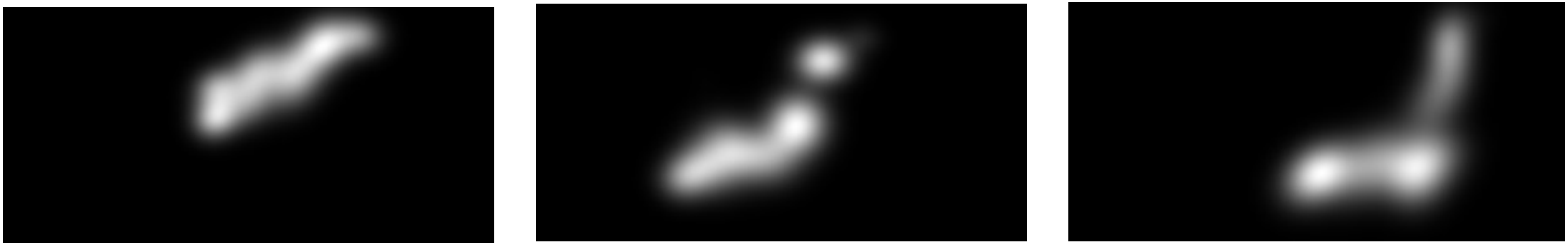
Summary based on HMAC: less merging



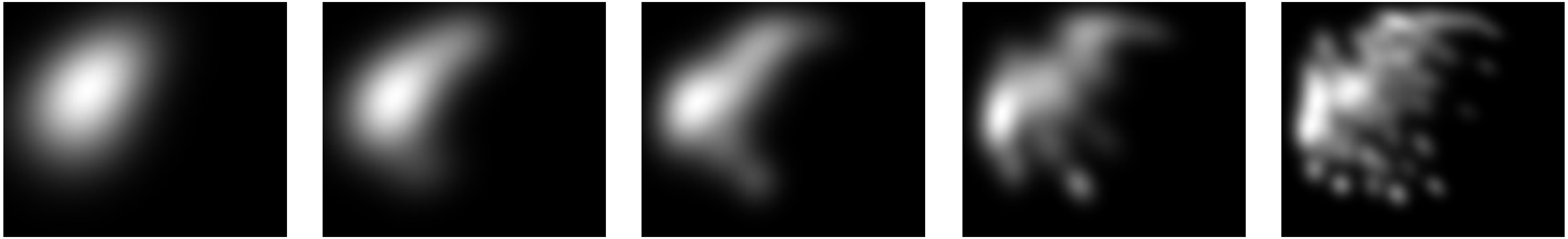


Barycenter over 41 images: #support points 3, 5, 8, 15, 45

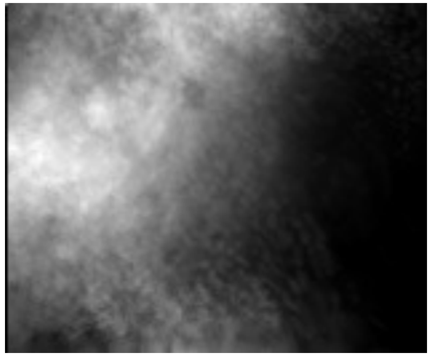
Average image



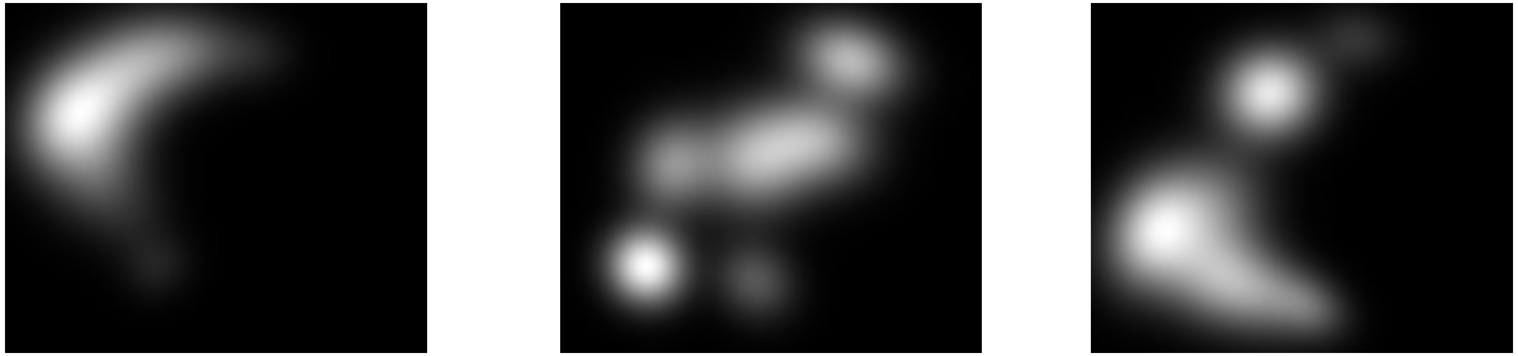
D2-clustering into 3 clusters, barycenters #support points: 8



Barycenter over 41 images: #support points 3, 5, 9, 15, 50



Average image



D2-clustering into 3 clusters
3 Barycenters, #support points: 9

References

- J. Li, J. Z. Wang, "Real-time computerized annotation of pictures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(6):985-1002, 2008. (*Proc. ACMMM 2006*)
- J. Ye, J. Li, "Scaling up discrete distribution clustering using ADMM," *Proc. IEEE Int. Conf. Image Processing*, pp. 5267-71, Paris, France, October 2014.
- Y. Zhang, J. Z. Wang, J. Li, "Parallel massive clustering of discrete distributions," *ACM Transactions on Multimedia Computing, Communications and Applications*, 11(4):1-24, 2015.
- J. Ye, P. Wu, J. Z. Wang, J. Li, "Accelerated Discrete Distribution Clustering under Wasserstein Distance," manuscript, 2015.

Acknowledgment

- NSF DMS-0705210, NSF CCF-0936948