Model selection for deep audio source separation via clustering analysis

Alisa Liu, Prem Seetharaman, Bryan Pardo

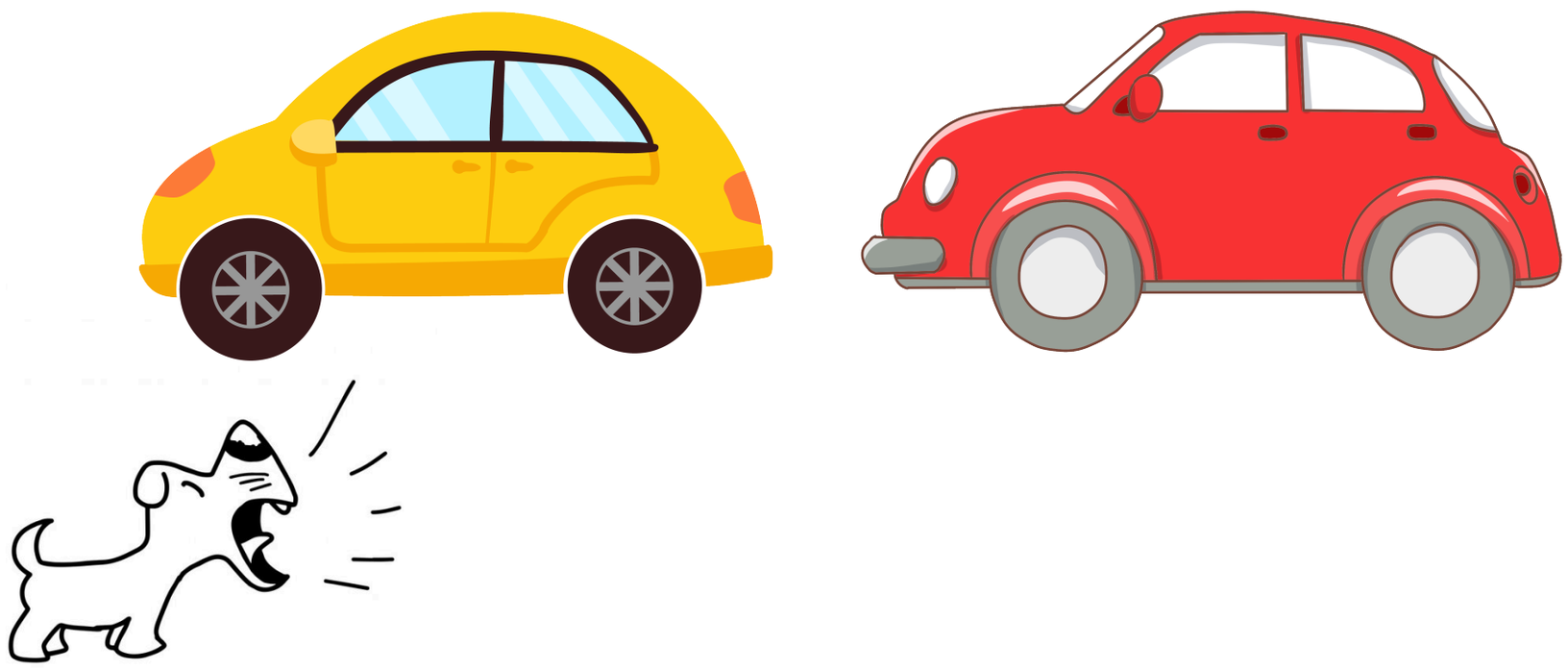
DCASE 2020

Northwestern University

Audio source separation

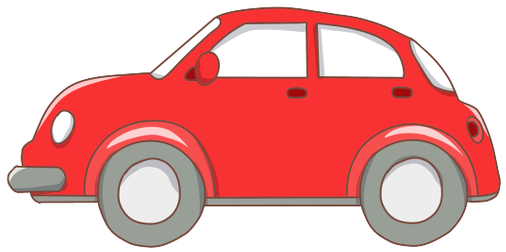
We can automatically distinguish different sound sources in an auditory scene…

Audio source separation



We can automatically distinguish different sound sources in an auditory scene…

Audio source separation



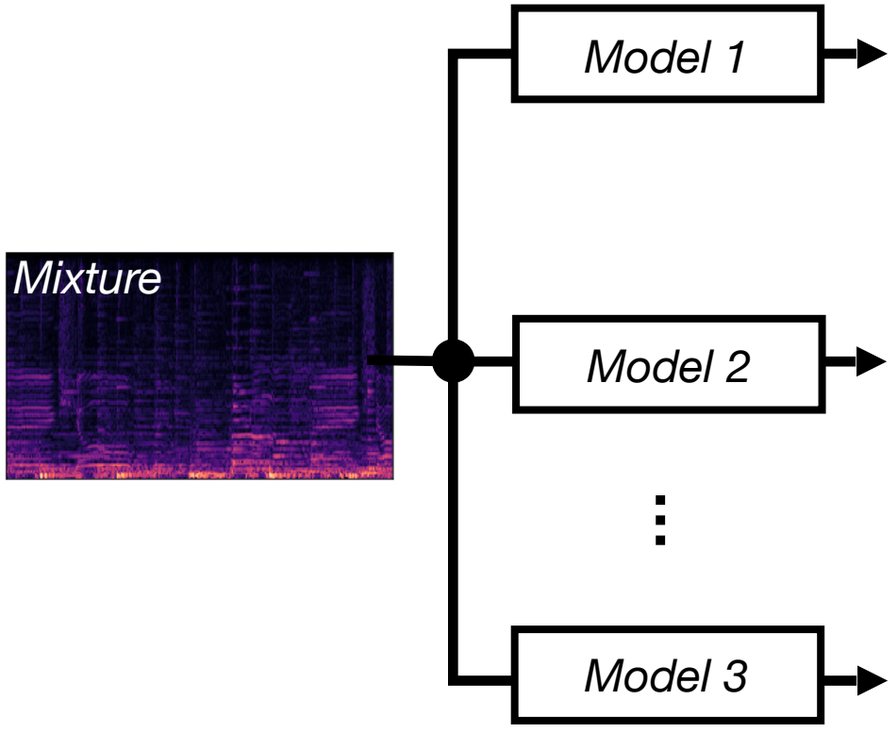
… Moreover, we can easily handle different types of auditory environments!

The problem

Source separation models are trained to separate on domain-specific data, and do not generalize across domains

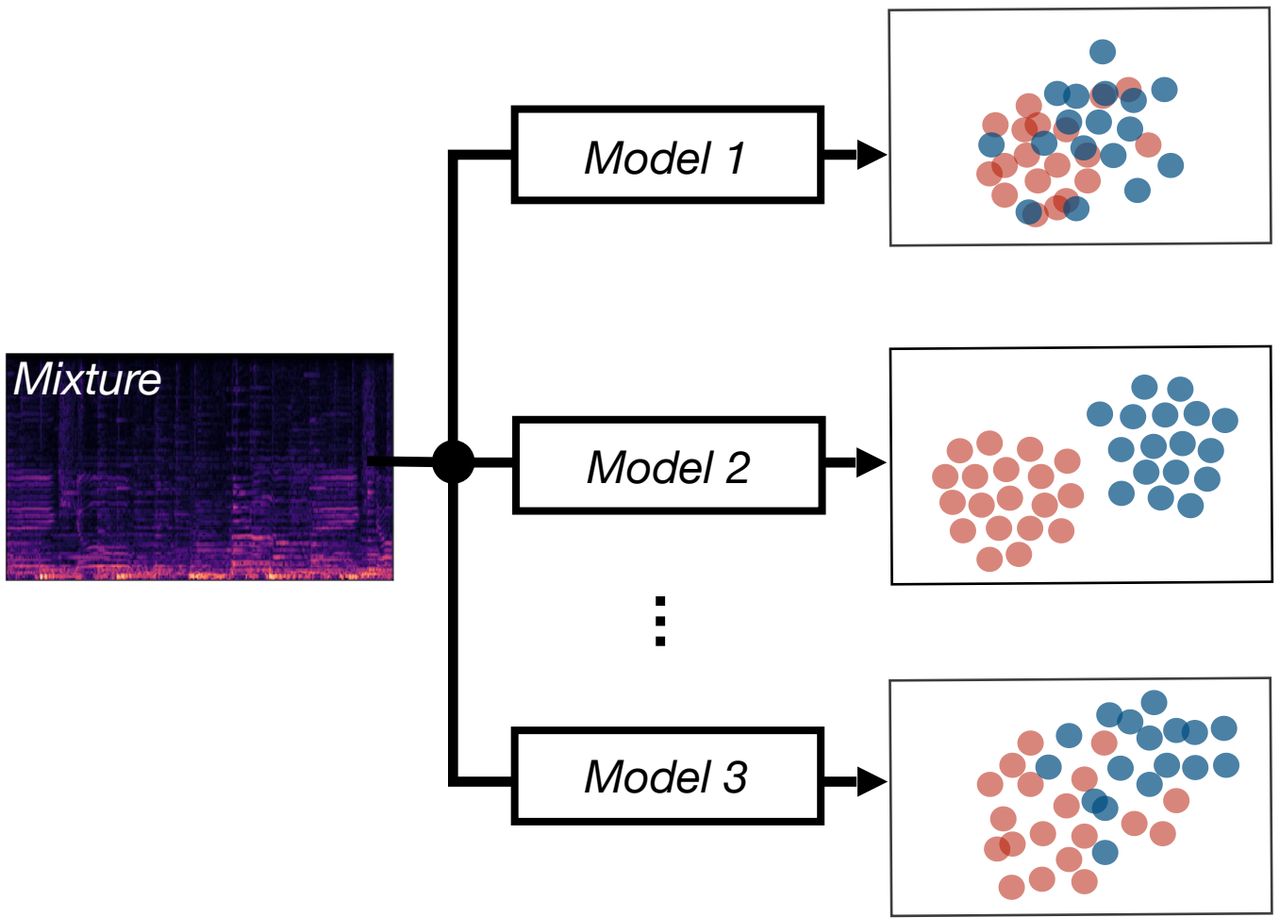
Motivating question

Given an **audio mixture** whose source domain is unknown, can we **automatically select** the best model for the mixture?



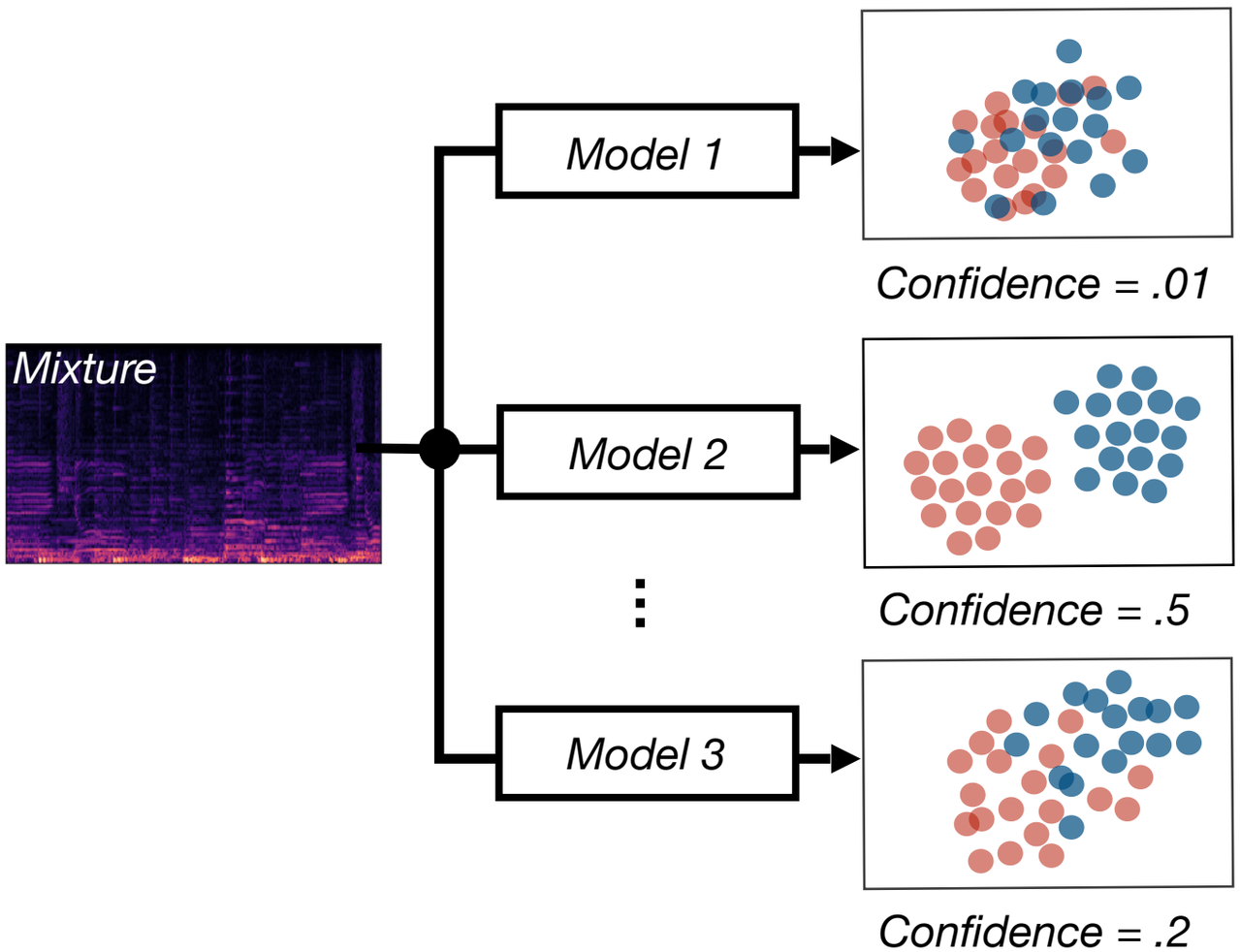
Solution: a confidence measure

We develop a confidence measure for systems that perform clustering-based separation…



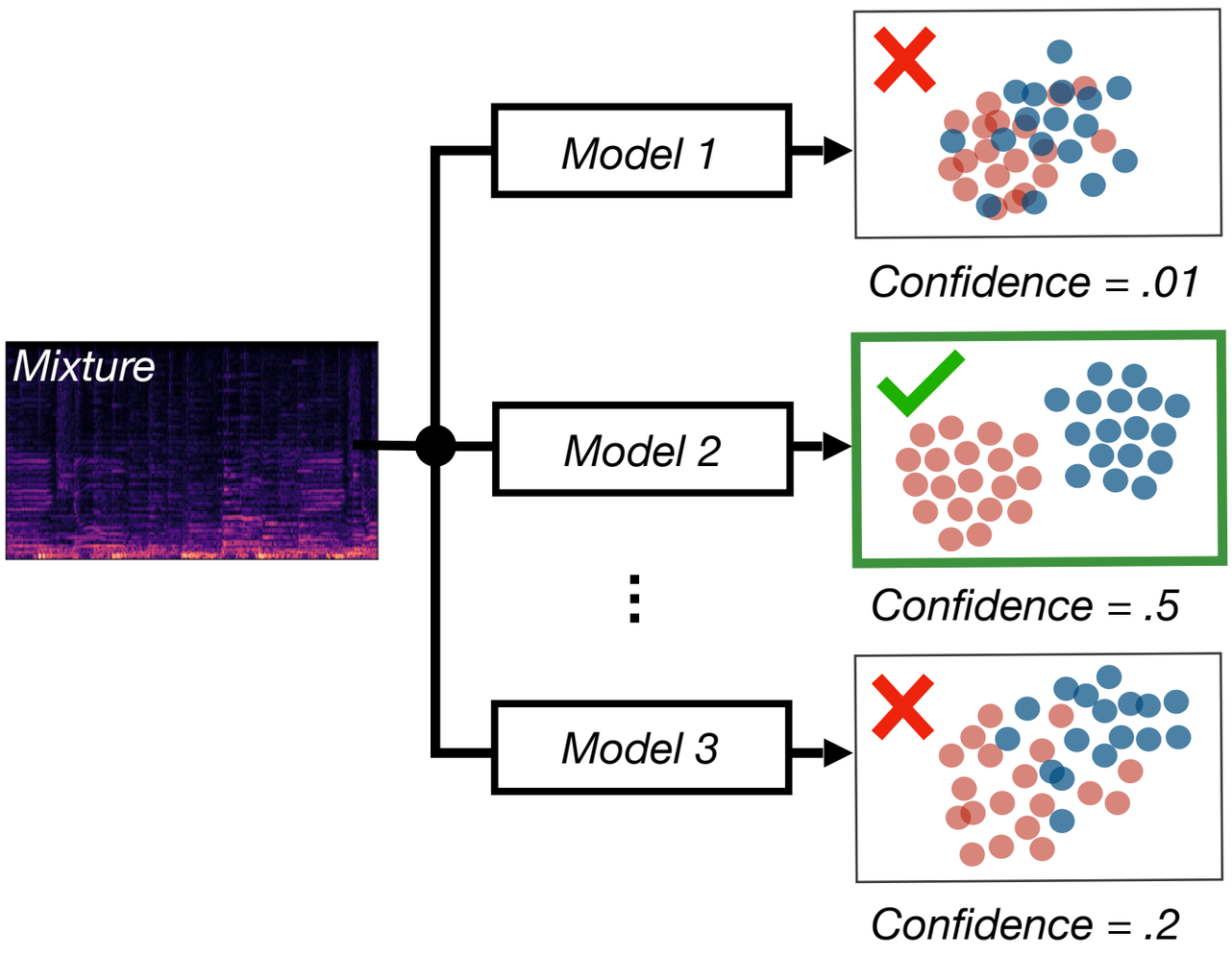
Solution: a confidence measure

We develop a confidence measure for systems that perform clustering-based separation…



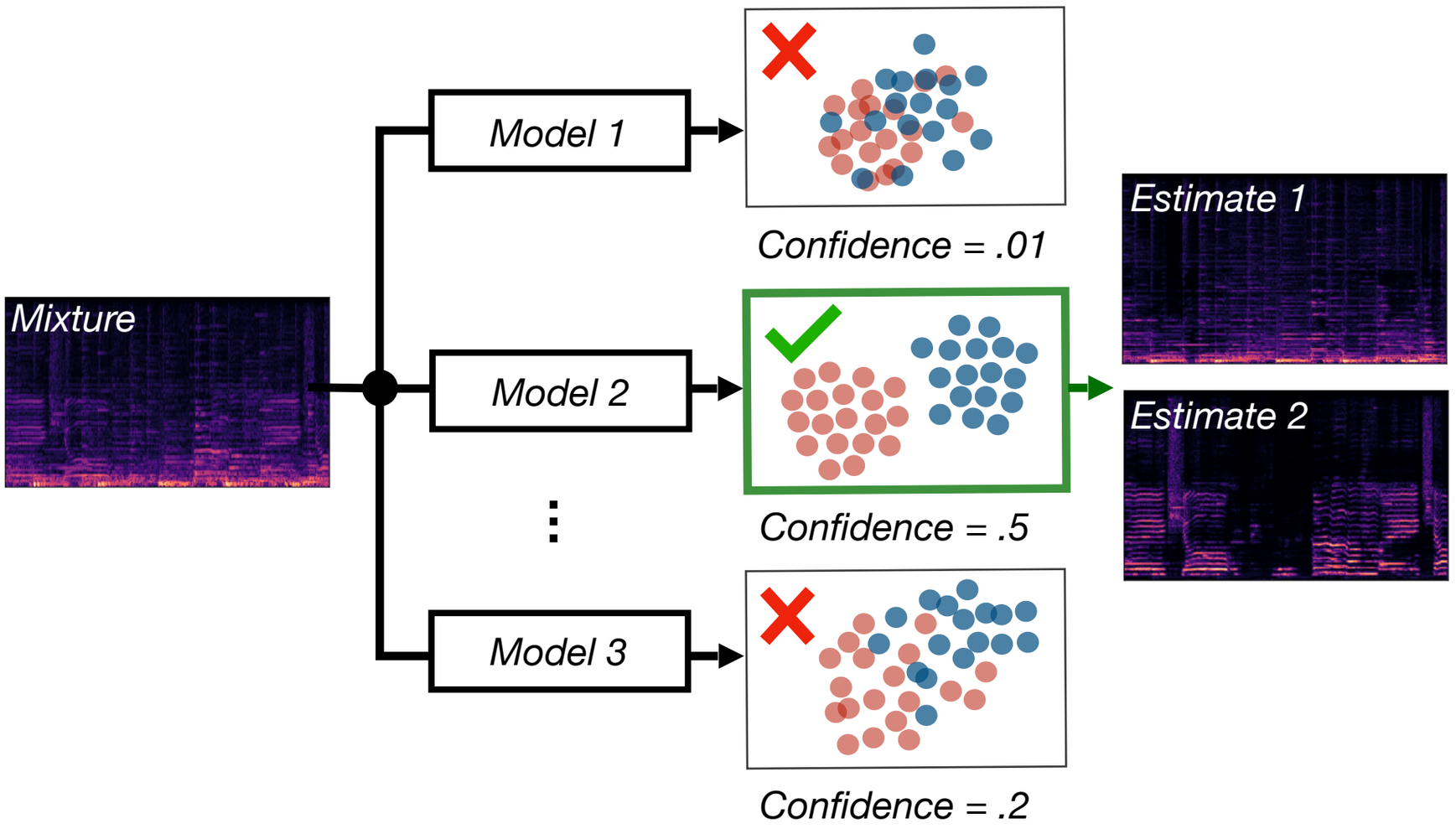
Solution: a confidence measure

We develop a confidence measure for systems that perform clustering-based separation…

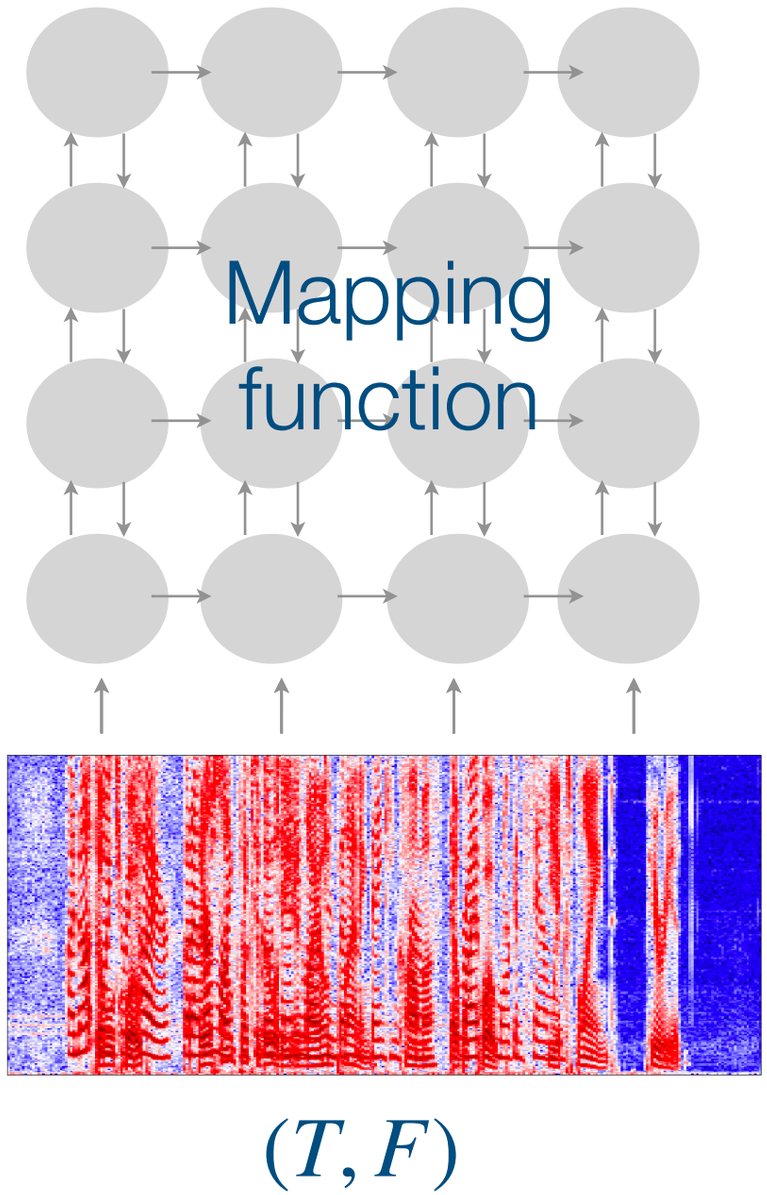


Solution: a confidence measure

We develop a confidence measure for systems that perform clustering-based separation… to automatically select the model output with the best predicted separation quality

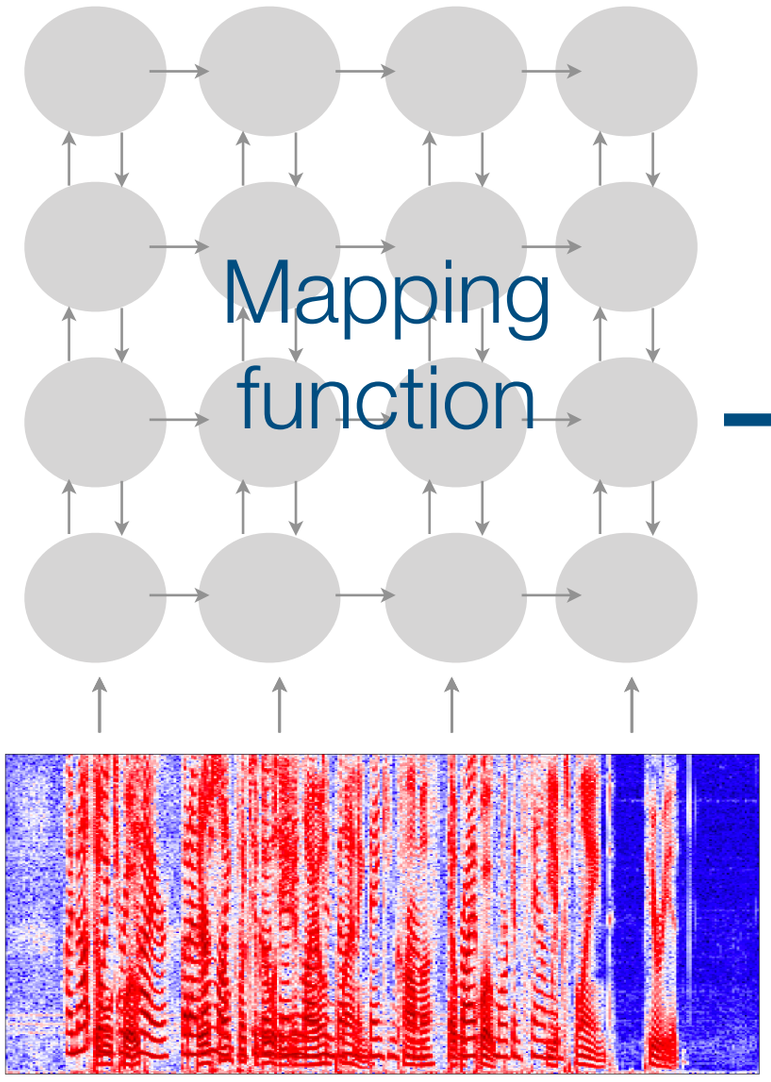


Clustering-based separation

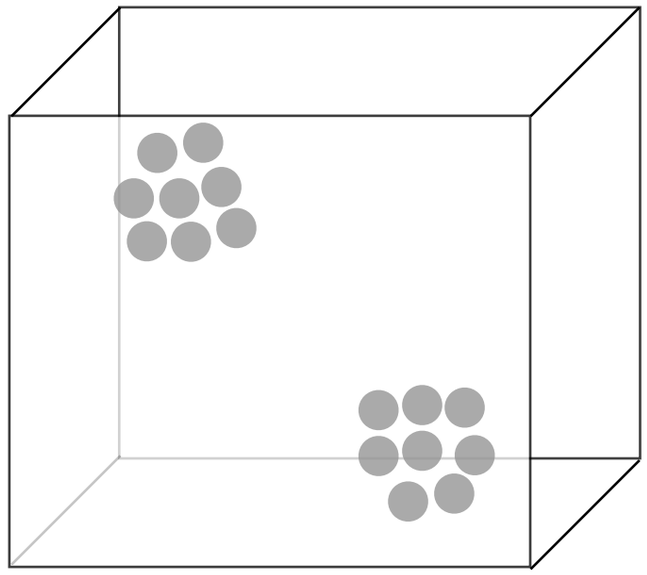


Map each time-frequency point…

Clustering-based separation

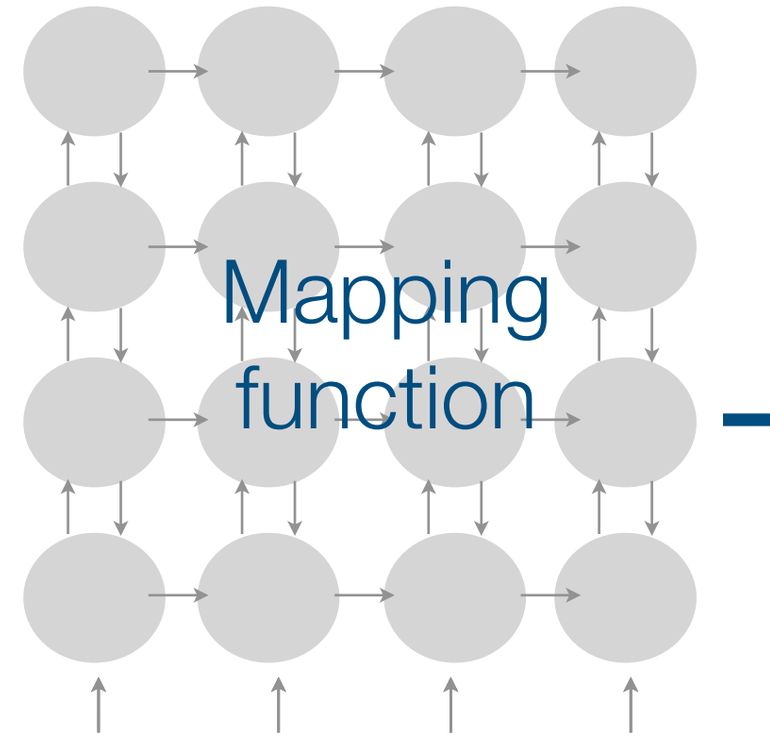


… to points in an embedding space…

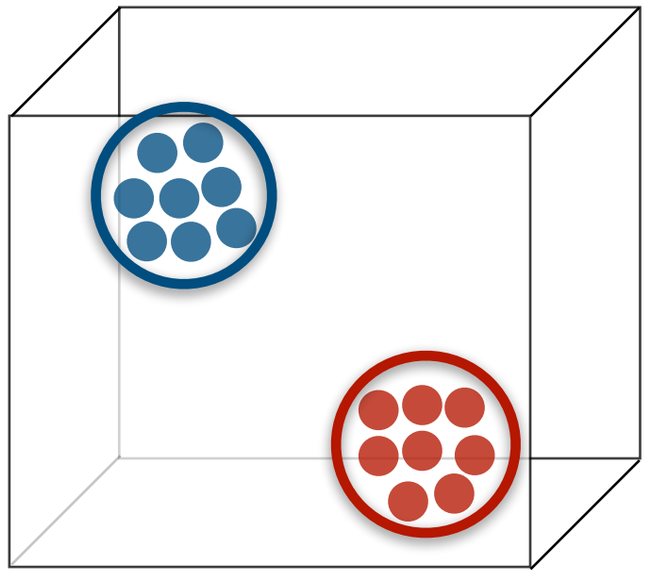


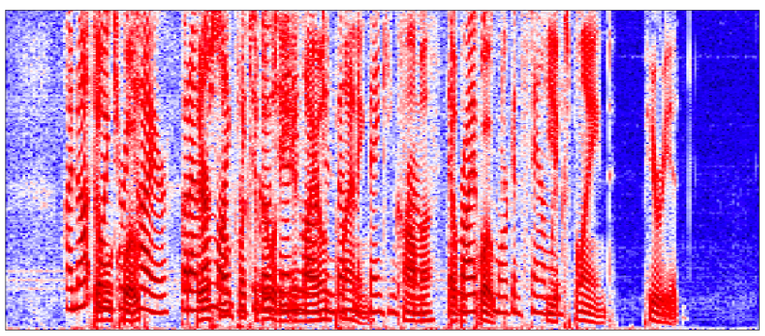
(T, F) → (T, F, K)

Clustering-based separation



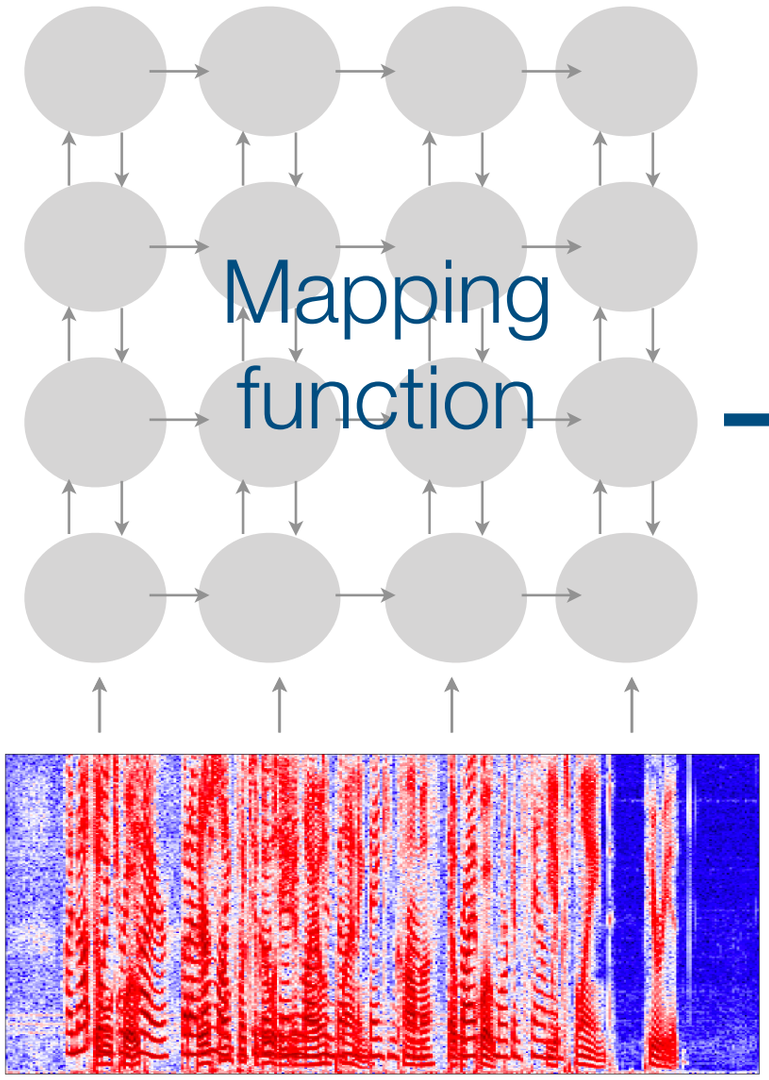
… so we can cluster the embeddings…



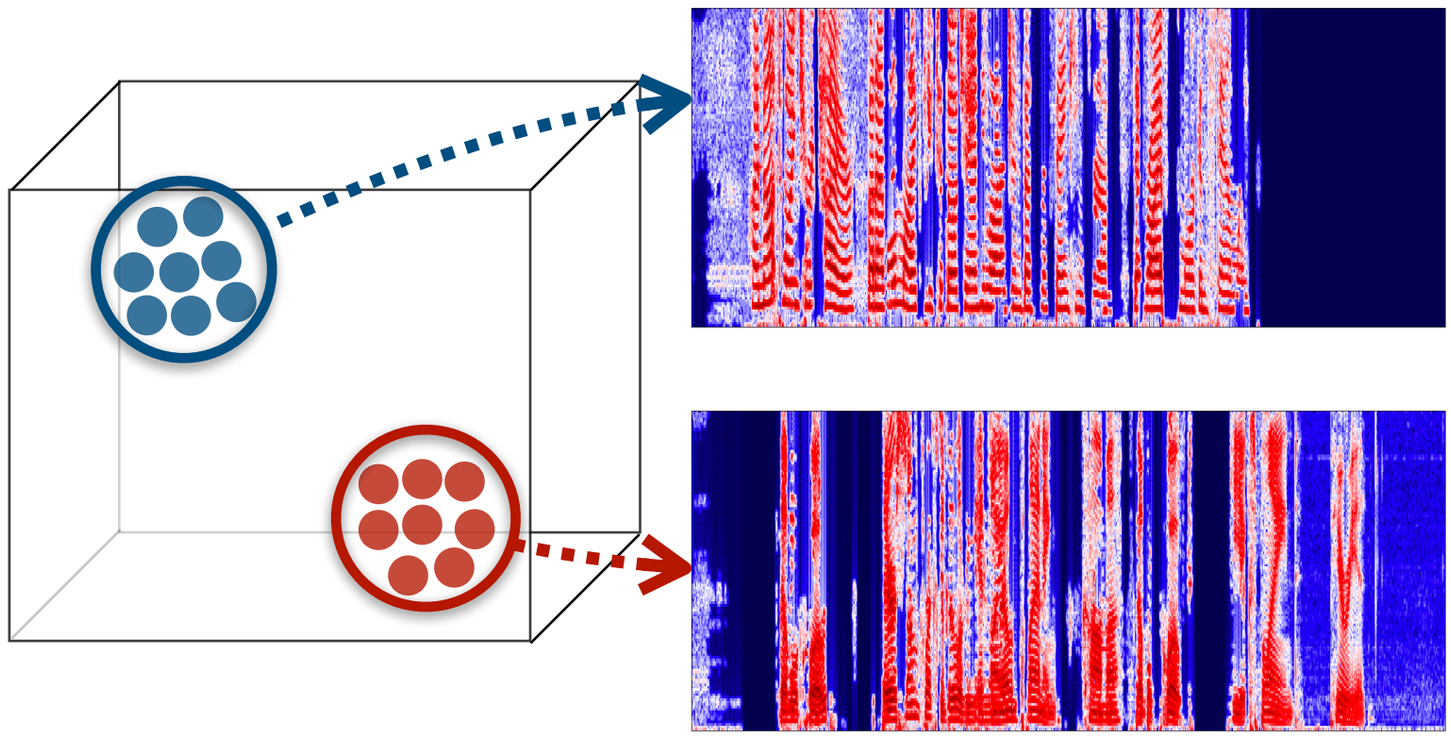


(T, F) → (T, F, K)

Clustering-based separation

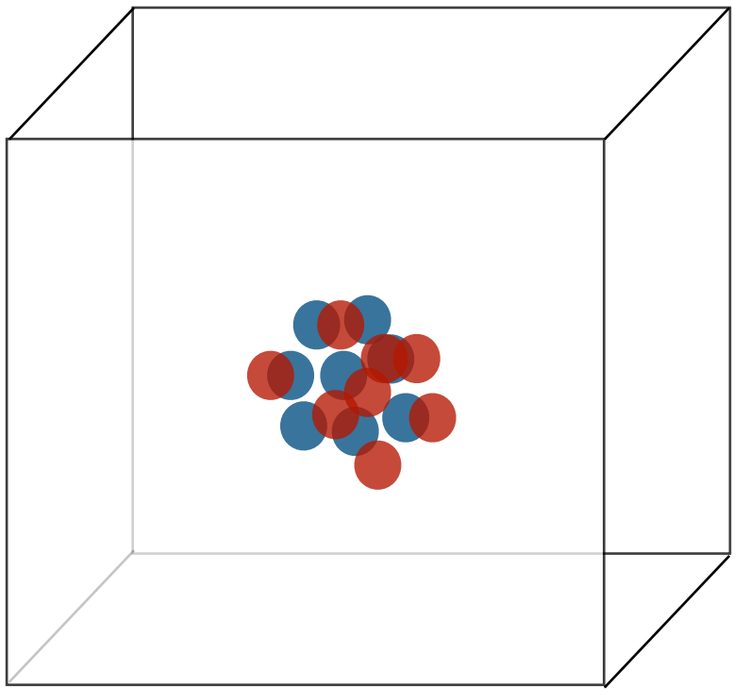


… to separate sources

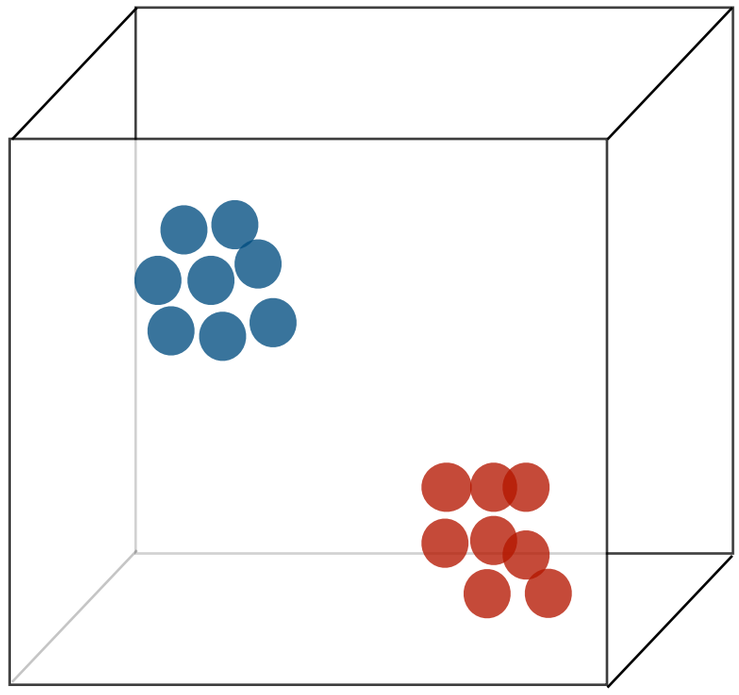


(T, F) → (T, F, K) → Get sources

Intuition for confidence measure



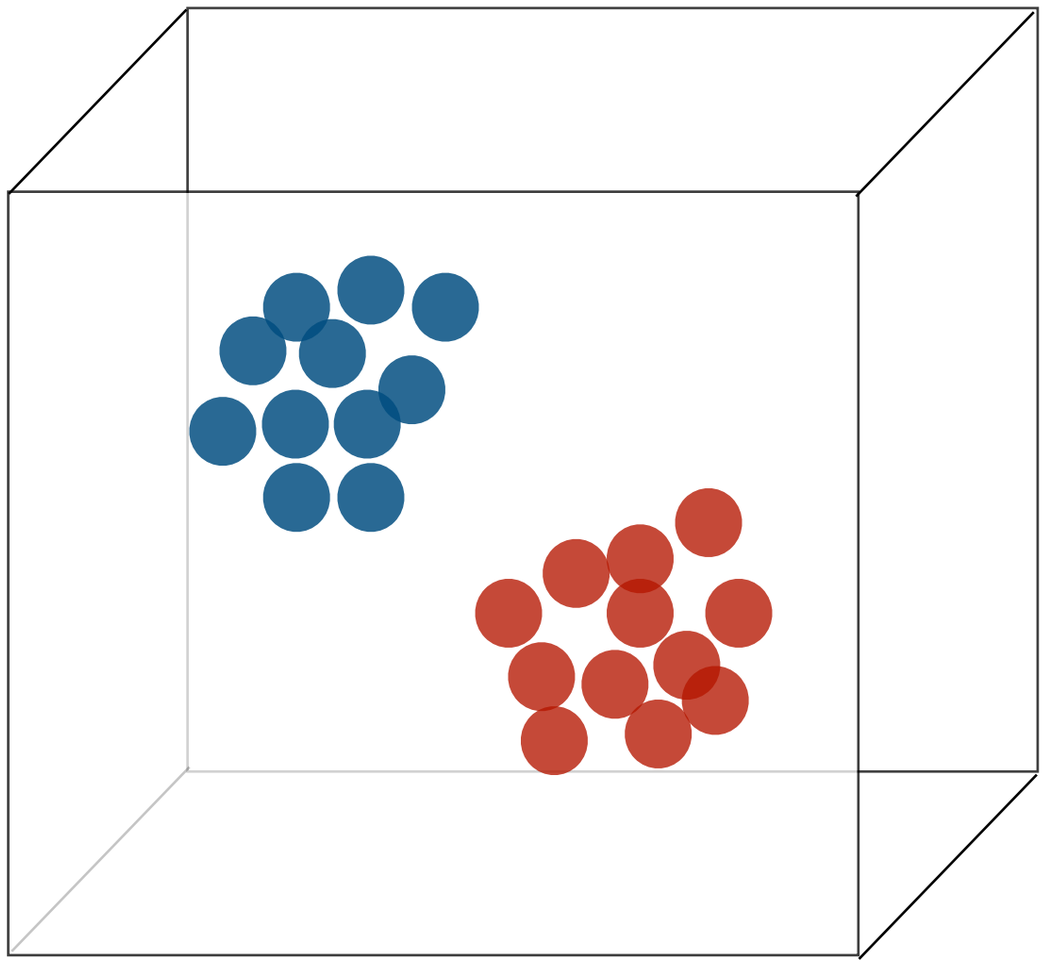
Low confidence case



High confidence case

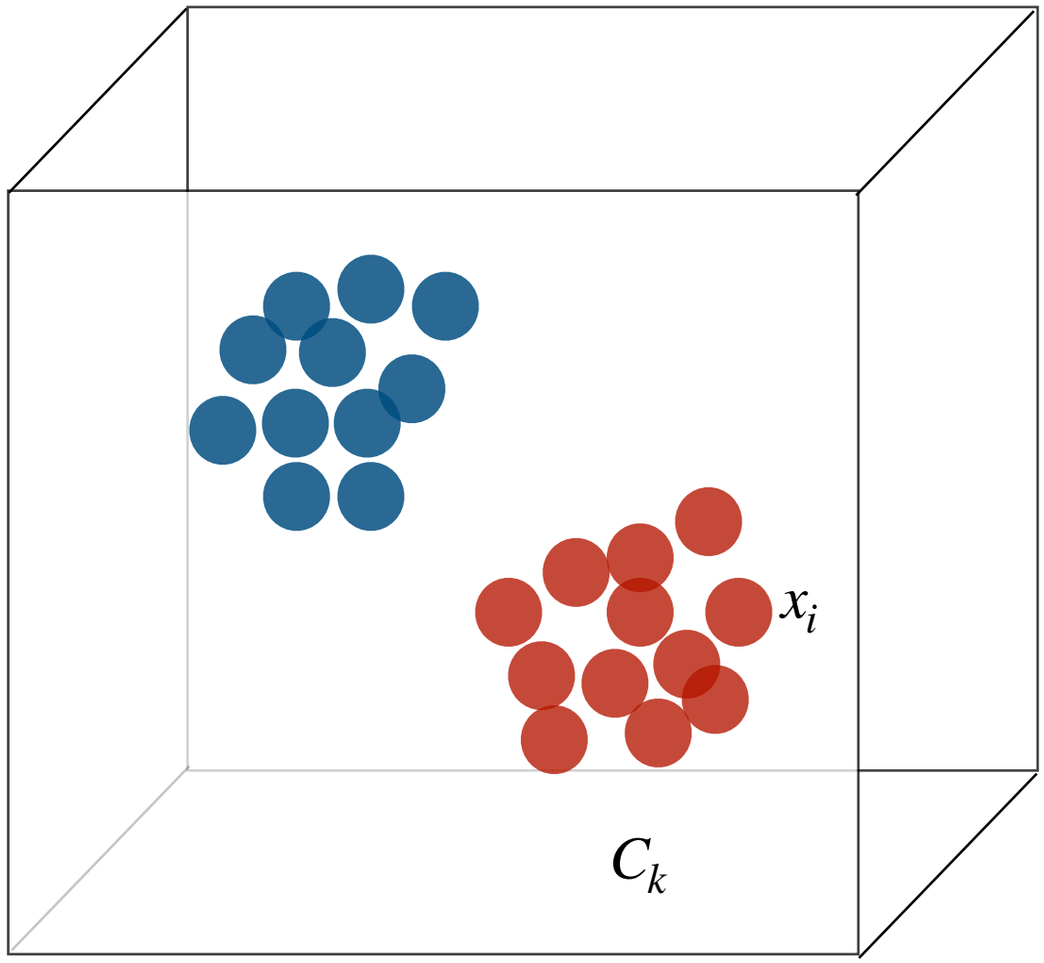
Key insight: The distribution of embedded TF points is predictive of the performance of the algorithm

1) Silhouette score



Clusters should be tight and far from each other

1) Silhouette score

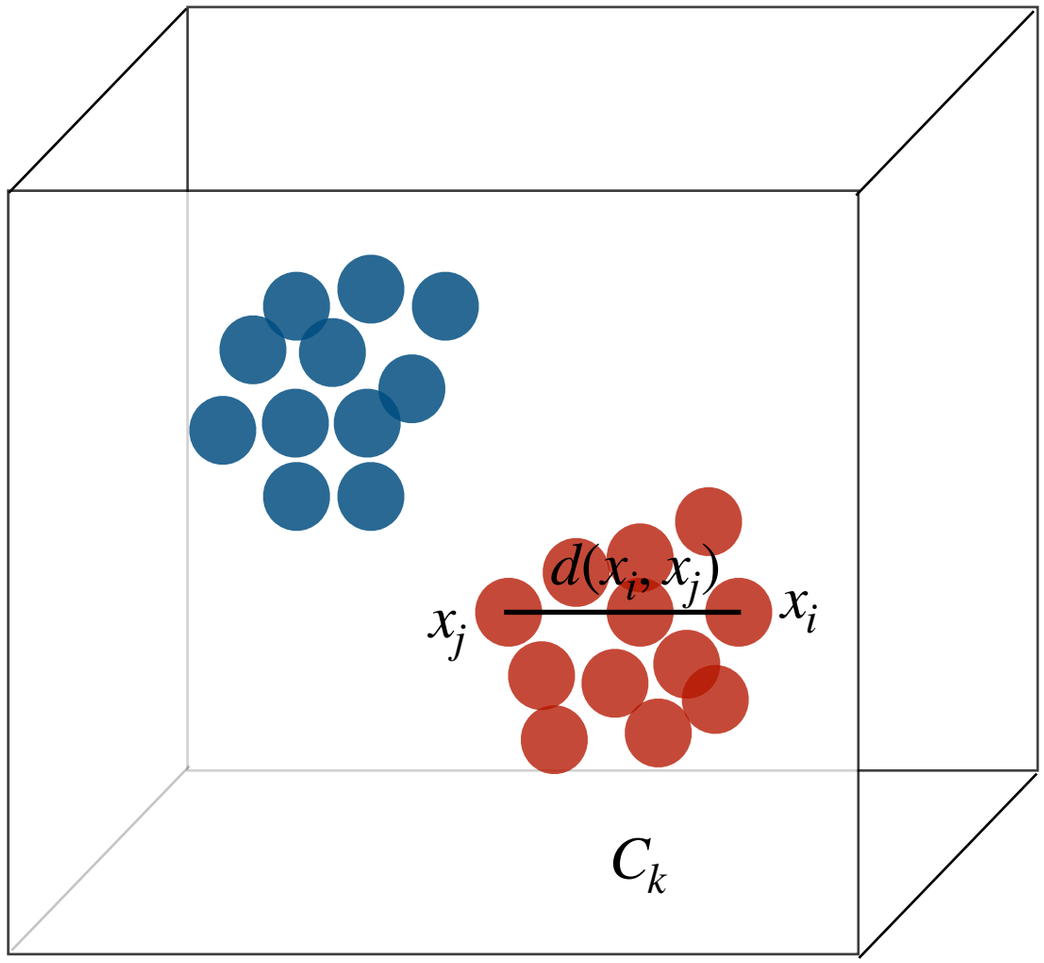


For a point x_i in cluster C_k

Intracluster distance

Clusters should be **tight** and far from each other

1) Silhouette score

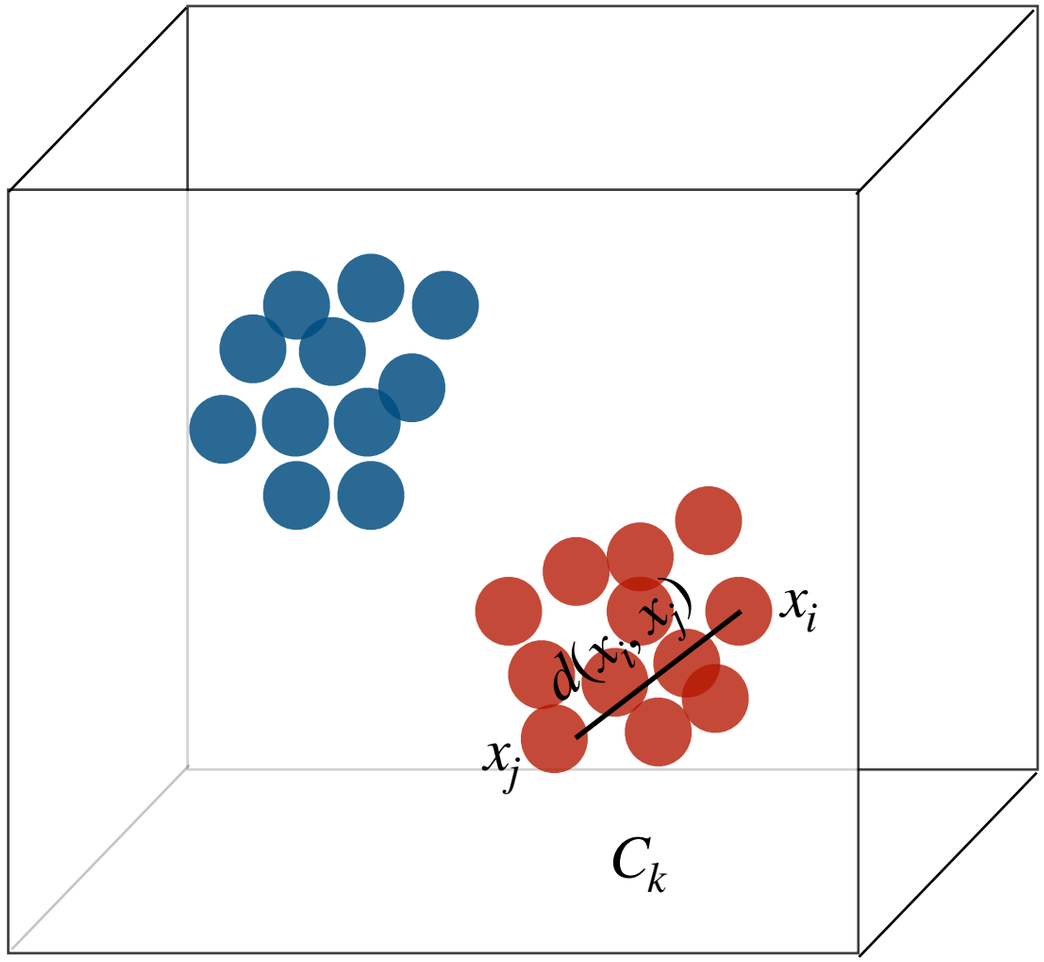


For a point x_i in cluster C_k

Intracluster distance

Clusters should be **tight** and far from each other

1) Silhouette score

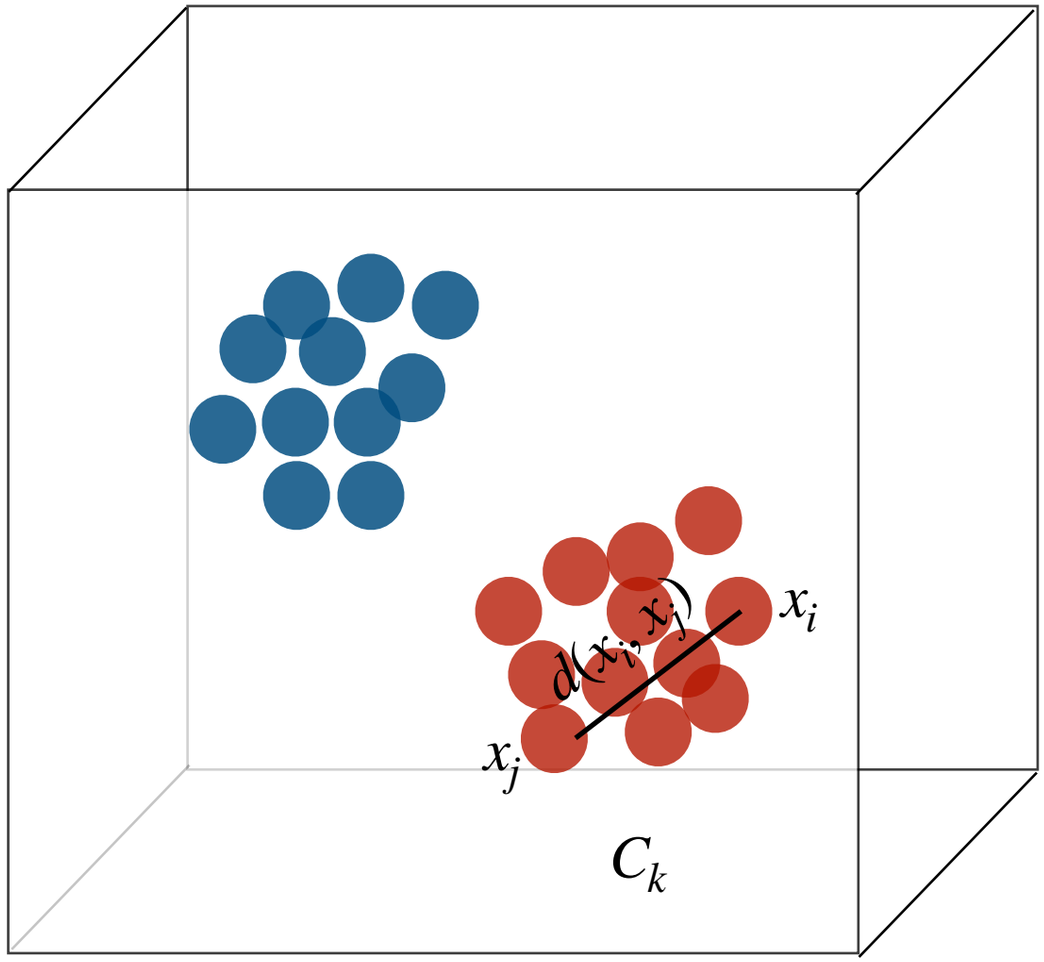


For a point x_i in cluster C_k

Intracluster distance

Clusters should be **tight** and far from each other

1) Silhouette score



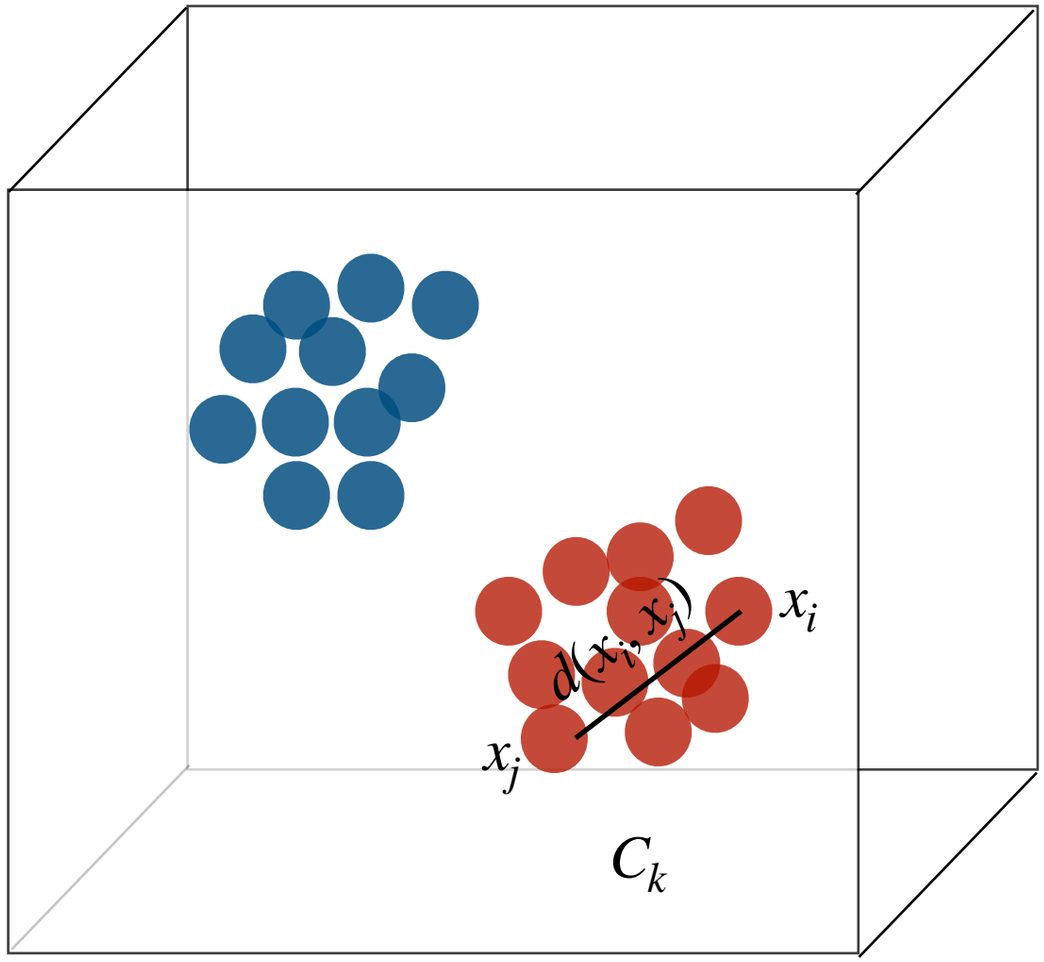
For a point x_i in cluster C_k

Intracluster distance

$$a(x_i) =$$

Clusters should be **tight** and far from each other

1) Silhouette score



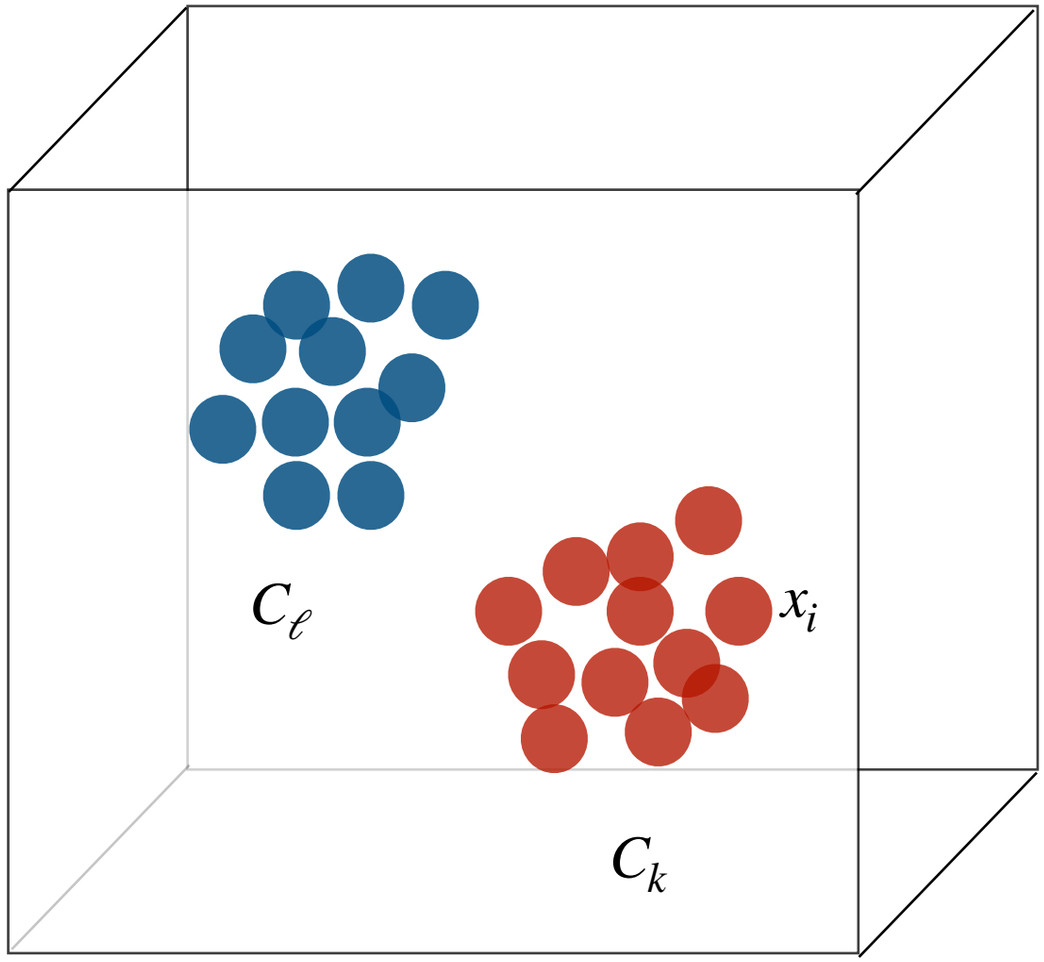
For a point x_i in cluster C_k

Intracluster distance

$$a(x_i) = \frac{1}{|C_k| - 1} \sum_{\substack{x_j \in C_k, \\ x_i \neq x_j}} d(x_i, x_j)$$

Clusters should be **tight** and far from each other

1) Silhouette score



For a point x_i in cluster C_k

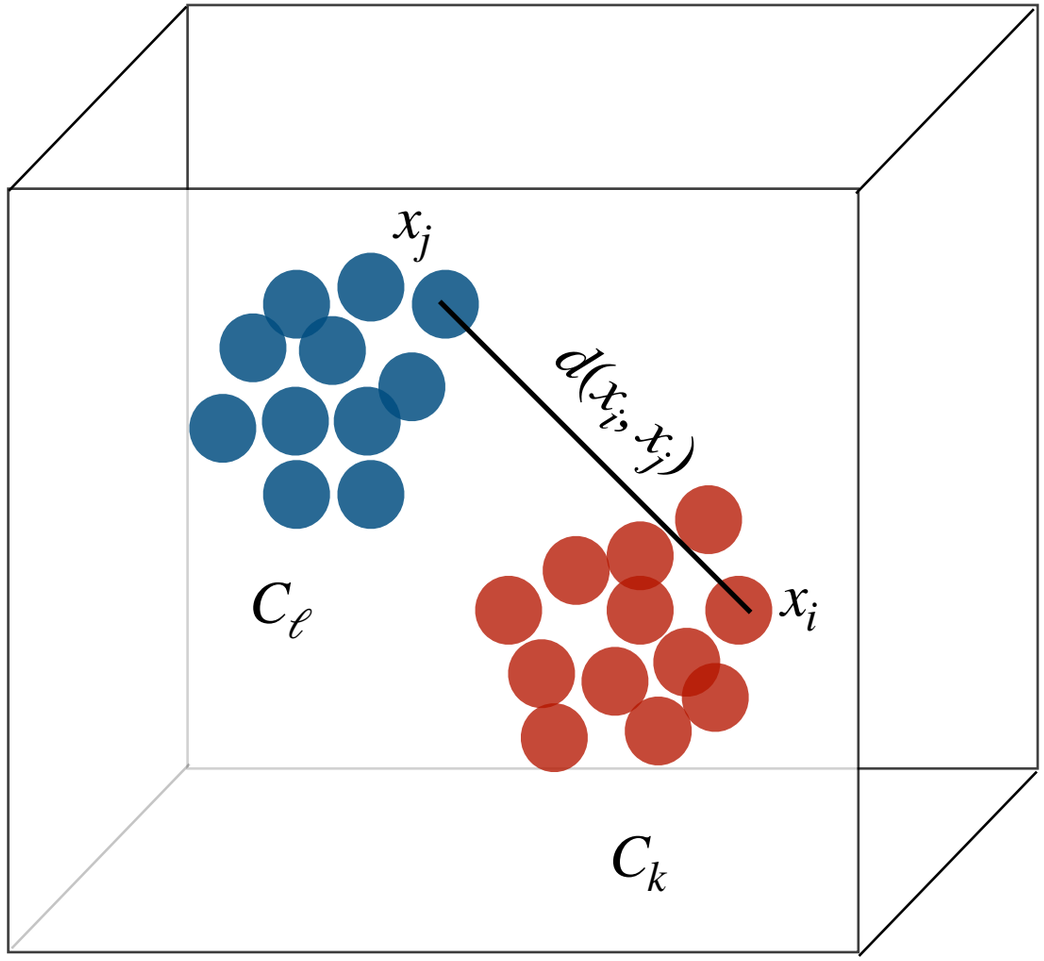
Intracluster distance

$$a(x_i) = \frac{1}{|C_k| - 1} \sum_{\substack{x_j \in C_k, \\ x_i \neq x_j}} d(x_i, x_j)$$

Intercluster distance

Clusters should be tight and **far** from each other

1) Silhouette score



For a point x_i in cluster C_k

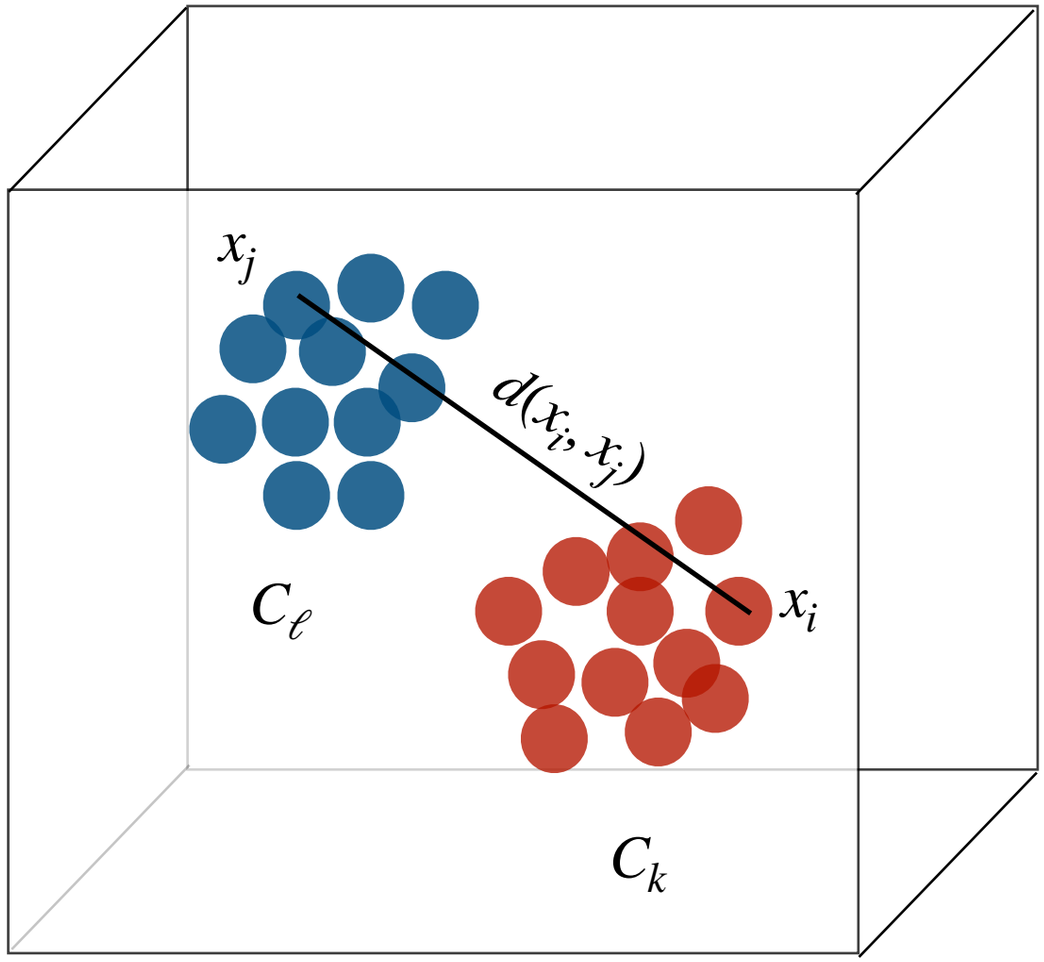
Intracluster distance

$$a(x_i) = \frac{1}{|C_k| - 1} \sum_{\substack{x_j \in C_k, \\ x_i \neq x_j}} d(x_i, x_j)$$

Intercluster distance

Clusters should be tight and **far** from each other

1) Silhouette score



For a point x_i in cluster C_k

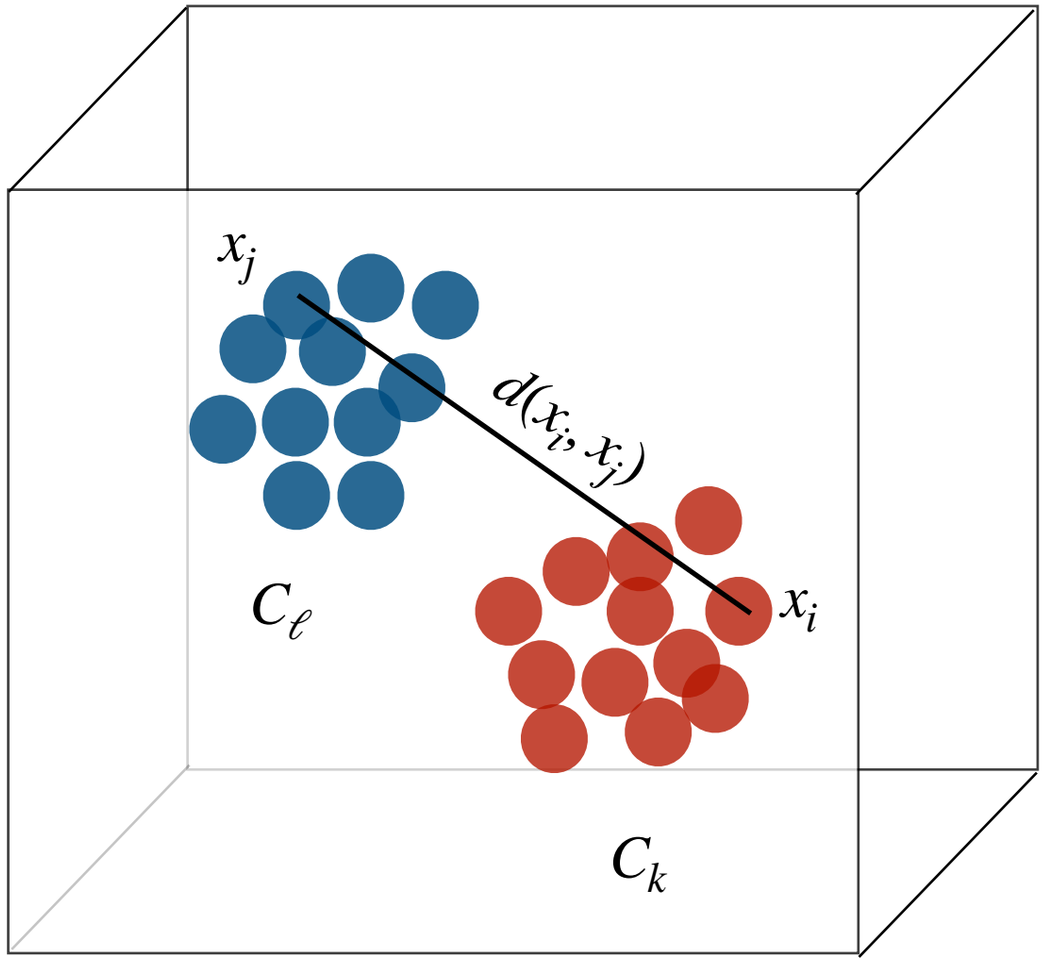
Intracluster distance

$$a(x_i) = \frac{1}{|C_k| - 1} \sum_{\substack{x_j \in C_k, \\ x_i \neq x_j}} d(x_i, x_j)$$

Intercluster distance

Clusters should be tight and **far** from each other

1) Silhouette score



For a point x_i in cluster C_k

Intracluster distance

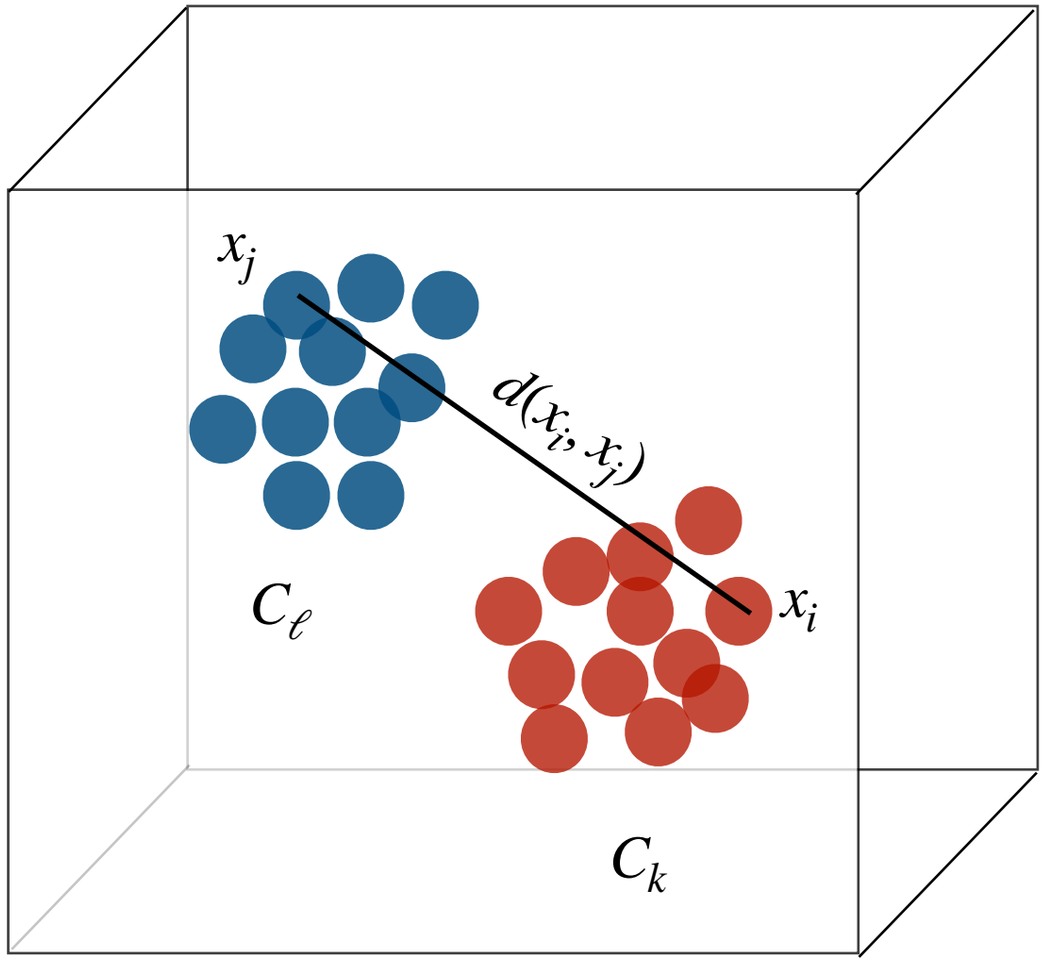
$$a(x_i) = \frac{1}{|C_k| - 1} \sum_{\substack{x_j \in C_k, \\ x_i \neq x_j}} d(x_i, x_j)$$

Intercluster distance

$$b(x_i) =$$

Clusters should be tight and **far** from each other

1) Silhouette score



For a point x_i in cluster C_k

Intracluster distance

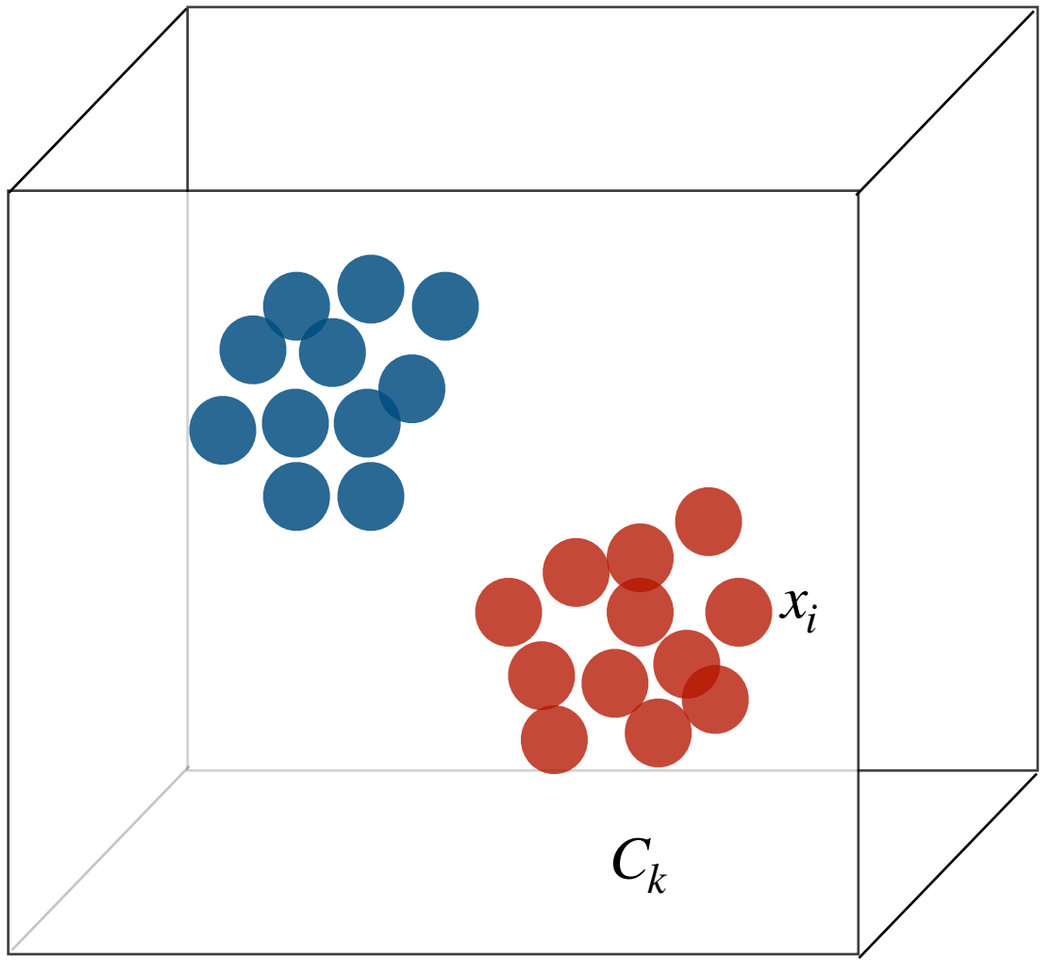
$$a(x_i) = \frac{1}{|C_k| - 1} \sum_{\substack{x_j \in C_k, \\ x_i \neq x_j}} d(x_i, x_j)$$

Intercluster distance

$$b(x_i) = \frac{1}{|C_\ell|} \sum_{x_j \in C_\ell} d(x_i, x_j)$$

Clusters should be tight and **far** from each other

1) Silhouette score



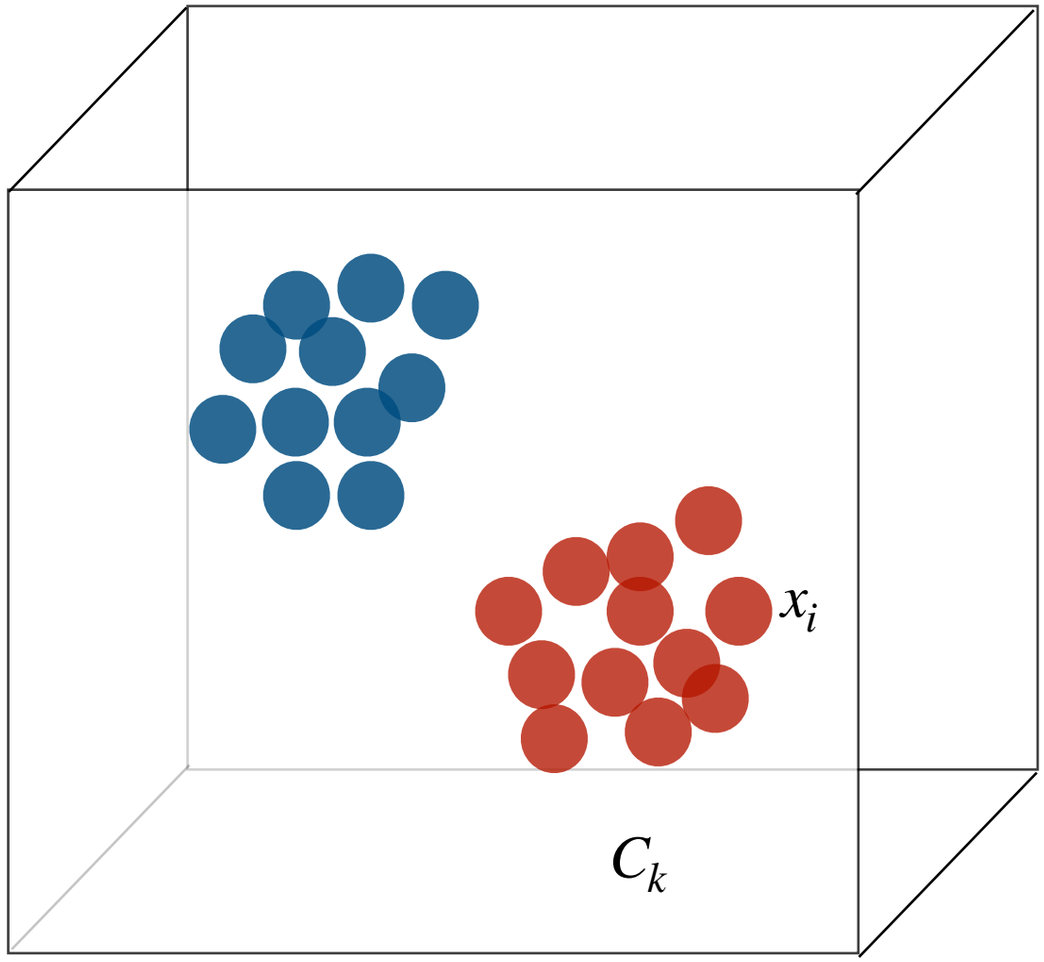
For a point x_i in cluster C_k

Combine and scale:

$$s(x_i) = \frac{b(x_i) - a(x_i)}{\max\left\{a(x_i), b(x_i)\right\}}$$

Clusters should be tight and far from each other

1) Silhouette score



For a point x_i in cluster C_k

Combine and scale:

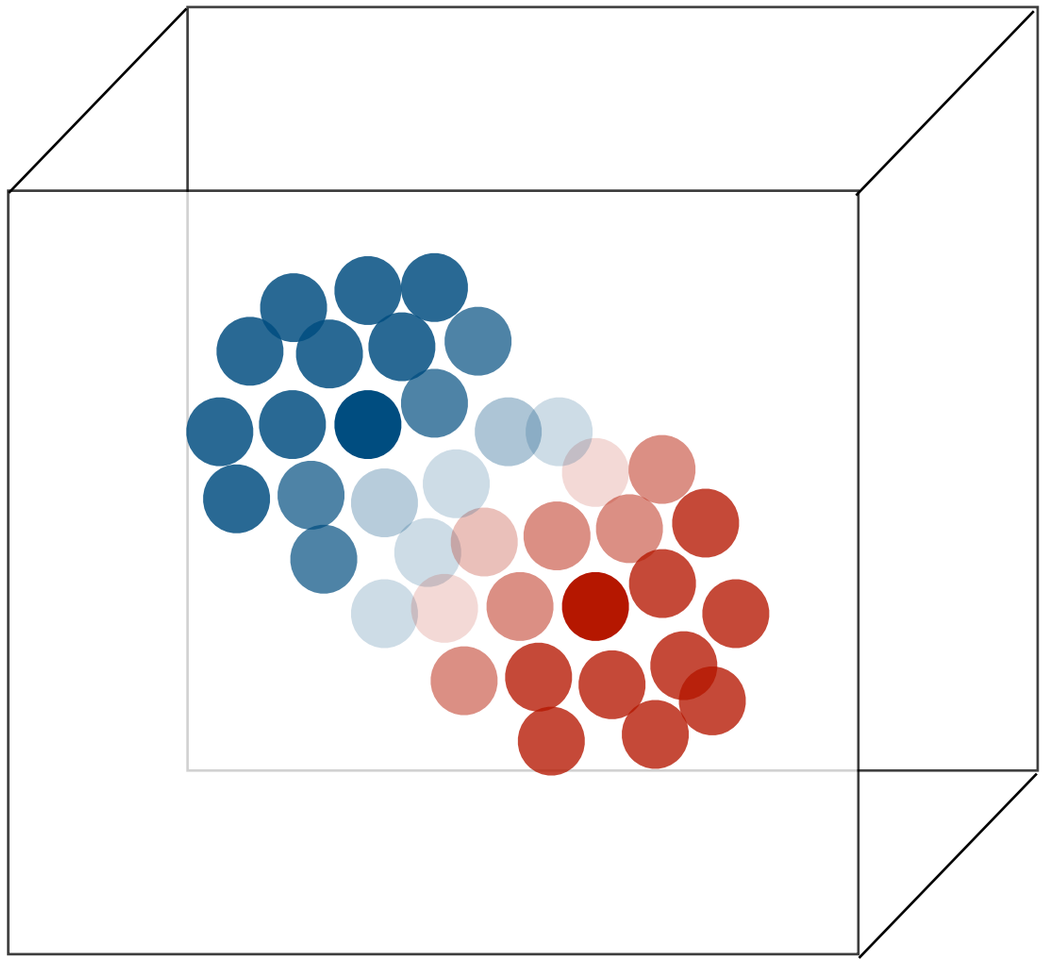
$$s(x_i) = \frac{b(x_i) - a(x_i)}{\max\left\{a(x_i), b(x_i)\right\}}$$

Average across a sample of 1000 TF bins in the top 1% loudest bins

$$S(X) = \sum_{x_i \in X} s(x_i)$$

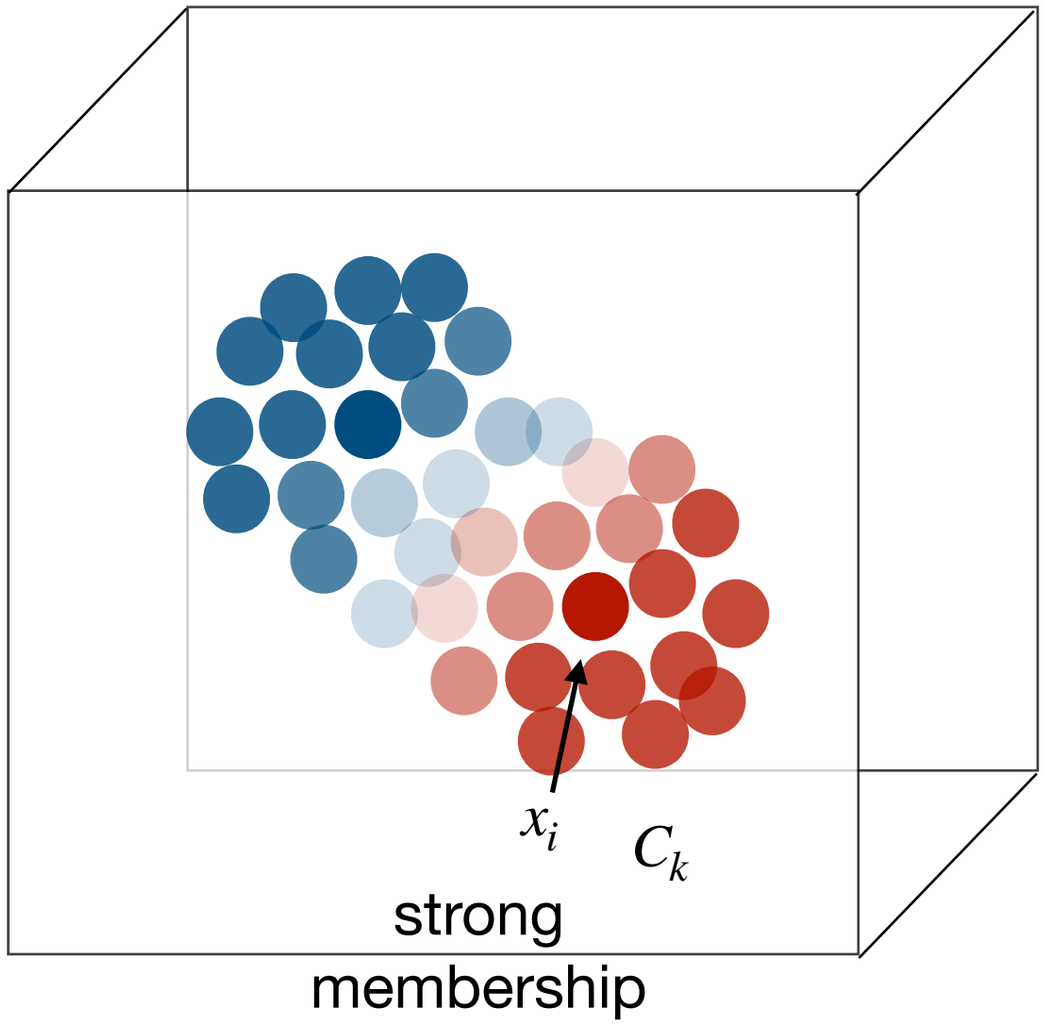
Clusters should be tight and far from each other

2) Posterior strength



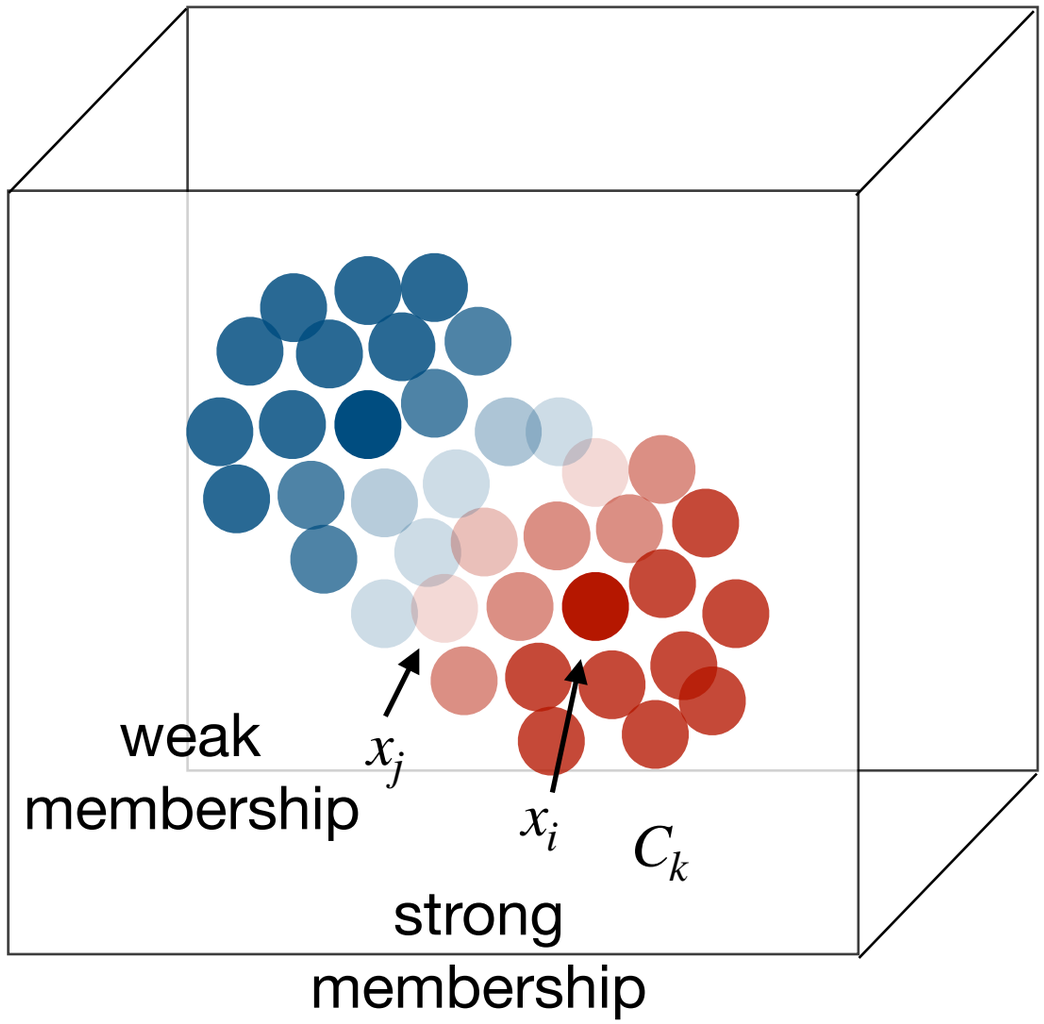
Points should have **strong membership** to their assigned cluster

2) Posterior strength



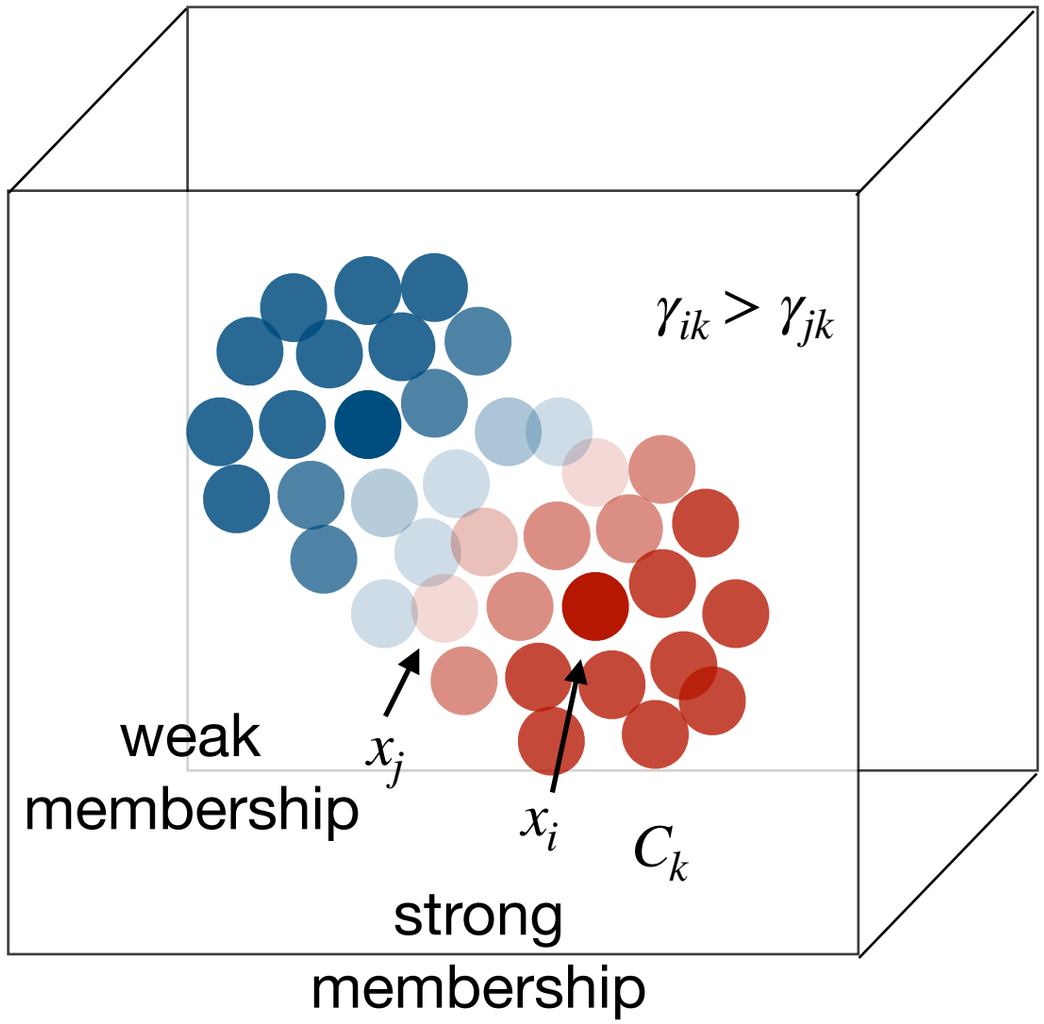
Points should have **strong membership** to their assigned cluster

2) Posterior strength



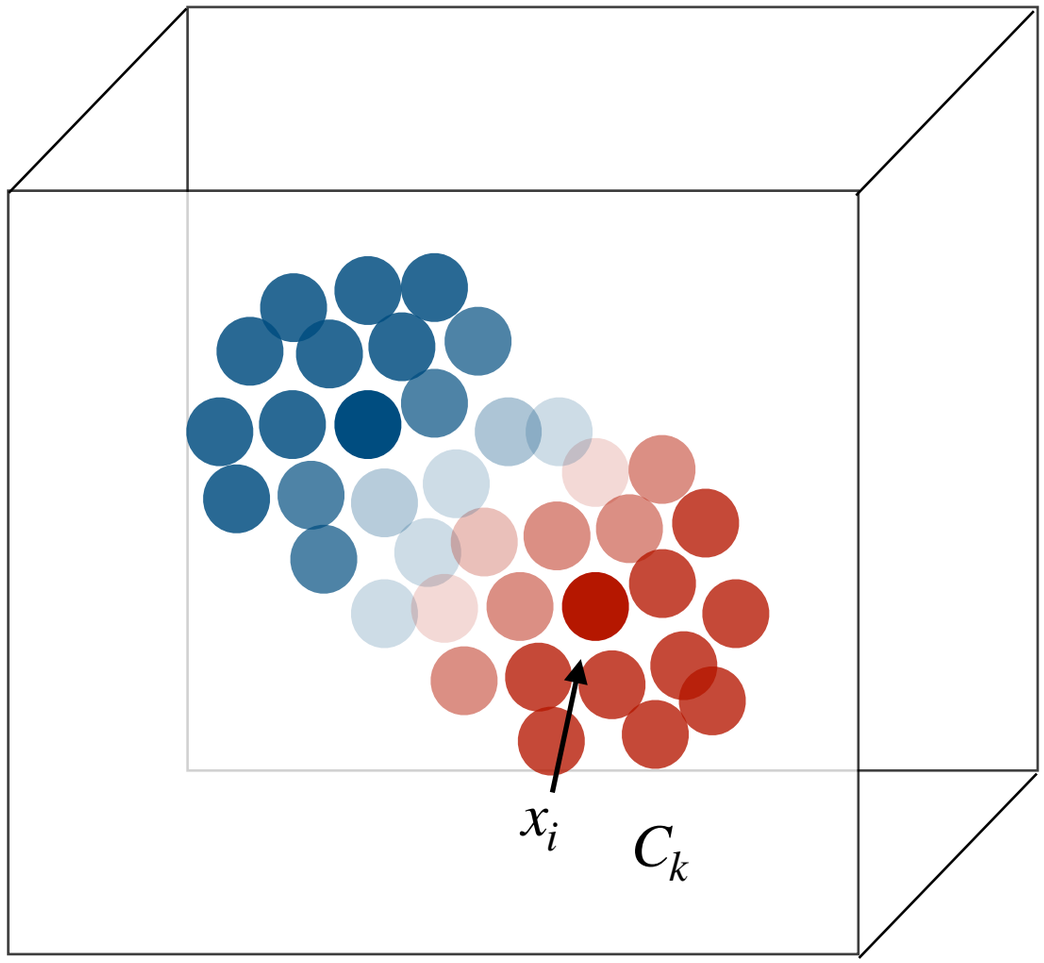
Points should have **strong membership** to their assigned cluster

2) Posterior strength



Points should have **strong membership** to their assigned cluster

2) Posterior strength

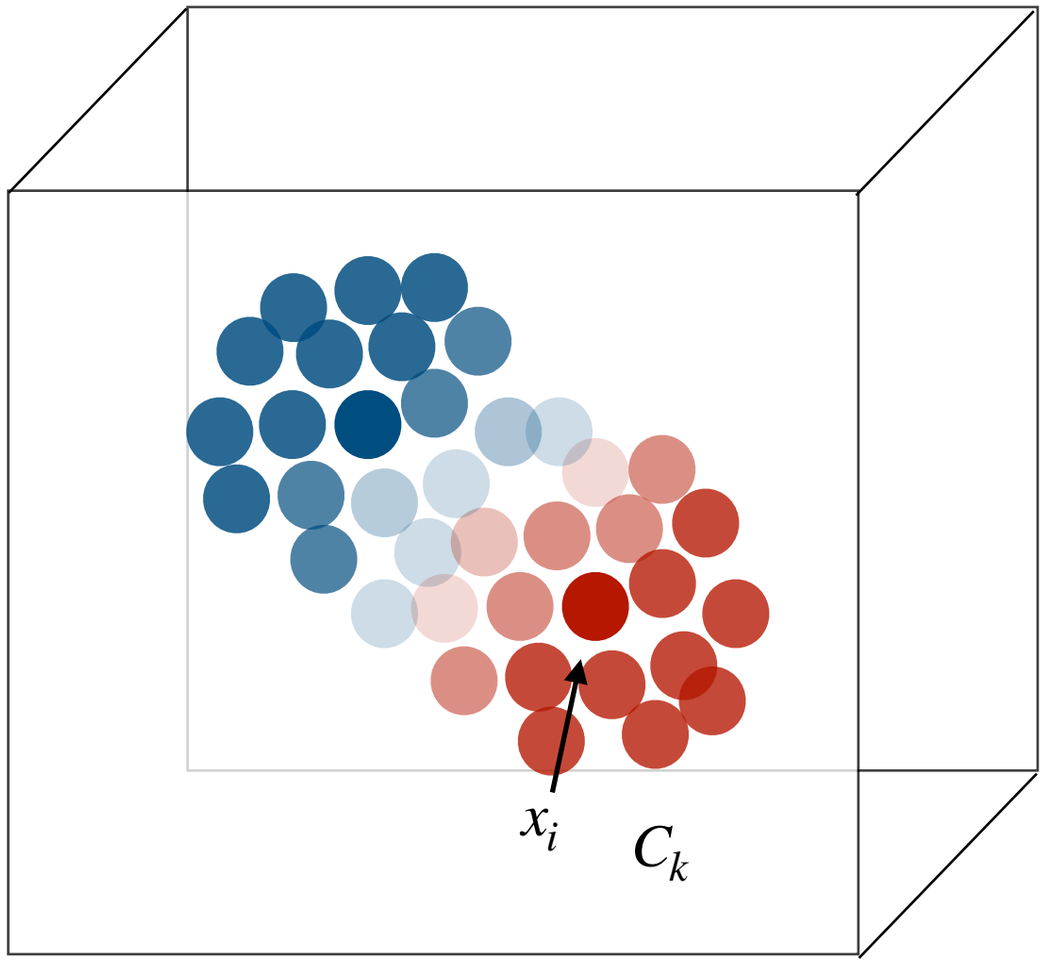


For a point x_i in cluster C_k

$p(x_i) =$

Points should have **strong membership** to their assigned cluster

2) Posterior strength

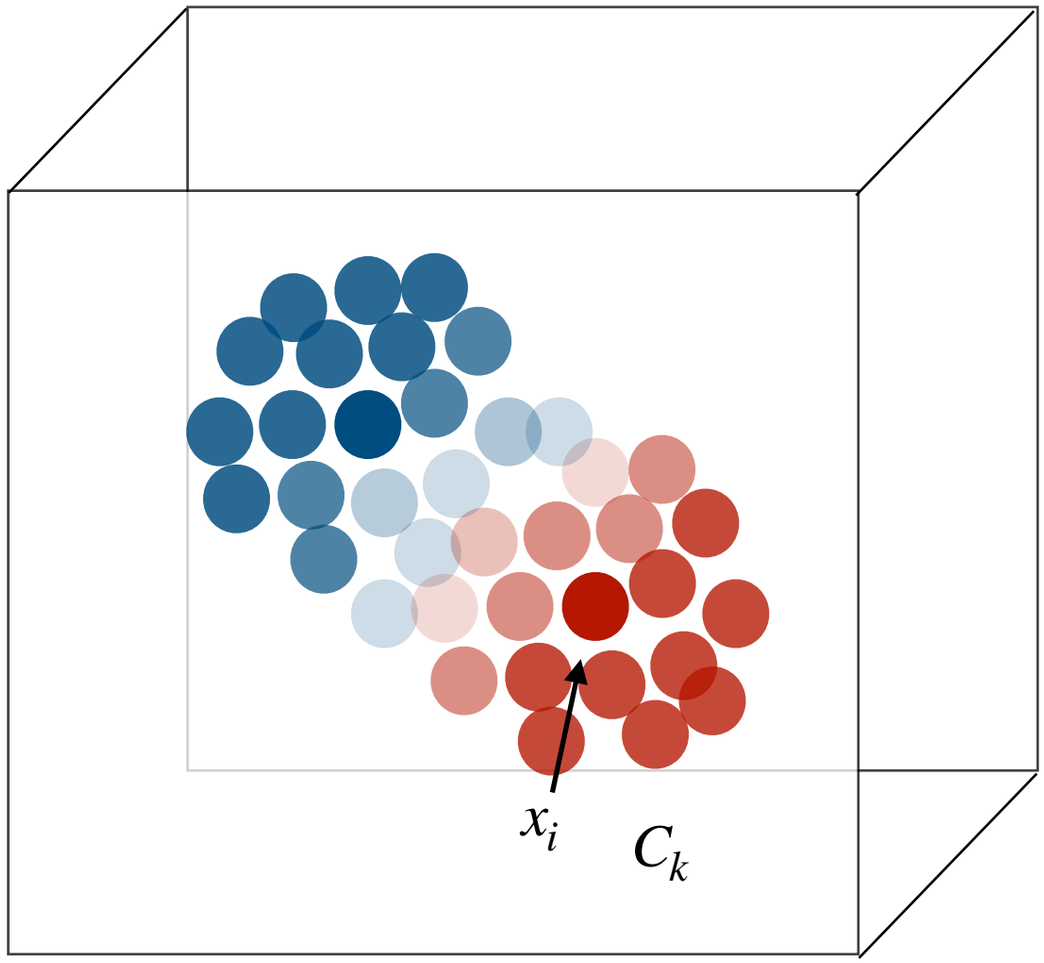


For a point x_i in cluster C_k

$$p(x_i) = \frac{K\left(\max_{k\in[0,\ldots,K]} \gamma_{ik}\right) - 1}{K - 1}$$

Points should have **strong membership** to their assigned cluster

2) Posterior strength



For a point x_i in cluster C_k

$$p(x_i) = \frac{K\left(\max_{k\in[0,\ldots,K]} \gamma_{ik}\right) - 1}{K - 1}$$

Average across the top 1% loudest bins

$$P(X) = \sum_{x_i \in X} p(x_i)$$

Points should have **strong membership** to their assigned cluster

Confidence measure

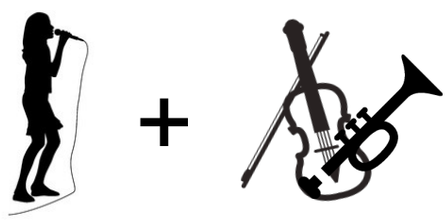
Combine silhouette score and posterior strength in a **product** to obtain the overall confidence measure for the mixture:

confidence measure

$$C(X) = \underbrace{S(X)}_{\text{silhouette score}}\underbrace{P(X)}_{\text{posterior strength}}$$

Experimental Design

Music

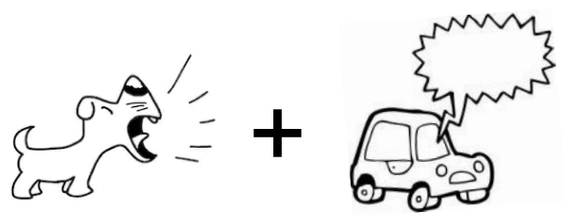


MusDB

Speech

Wall Street Journal

Environmental sounds

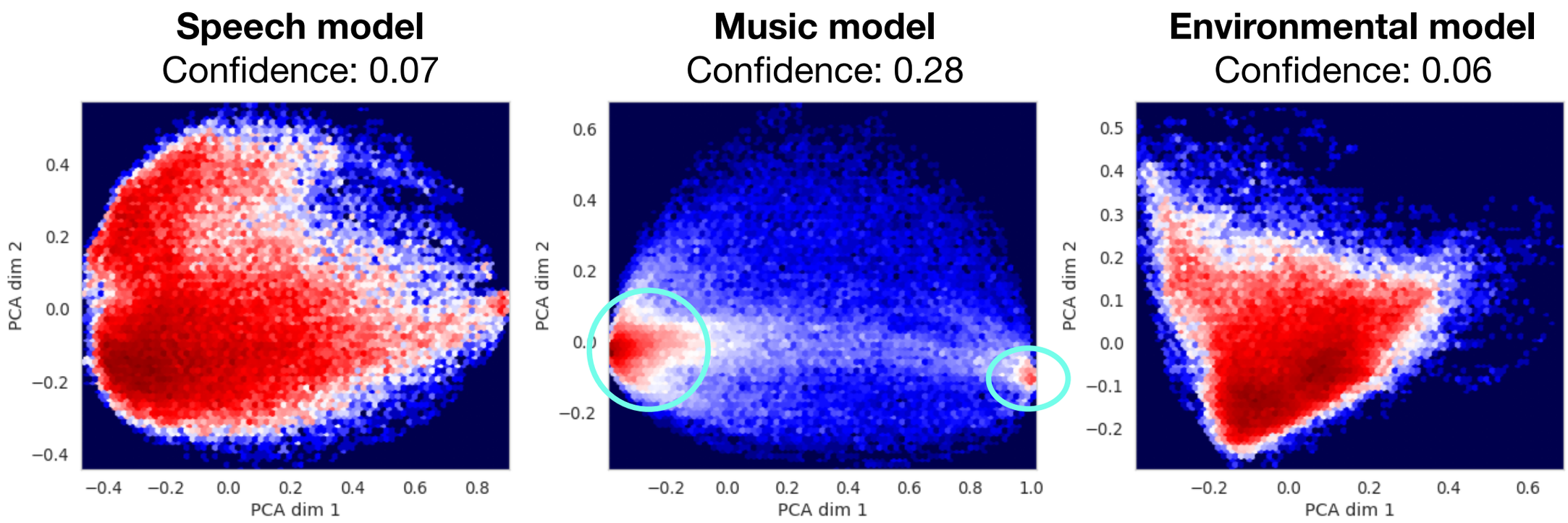


UrbanSound8k

For each domain, train a model with the deep clustering objective

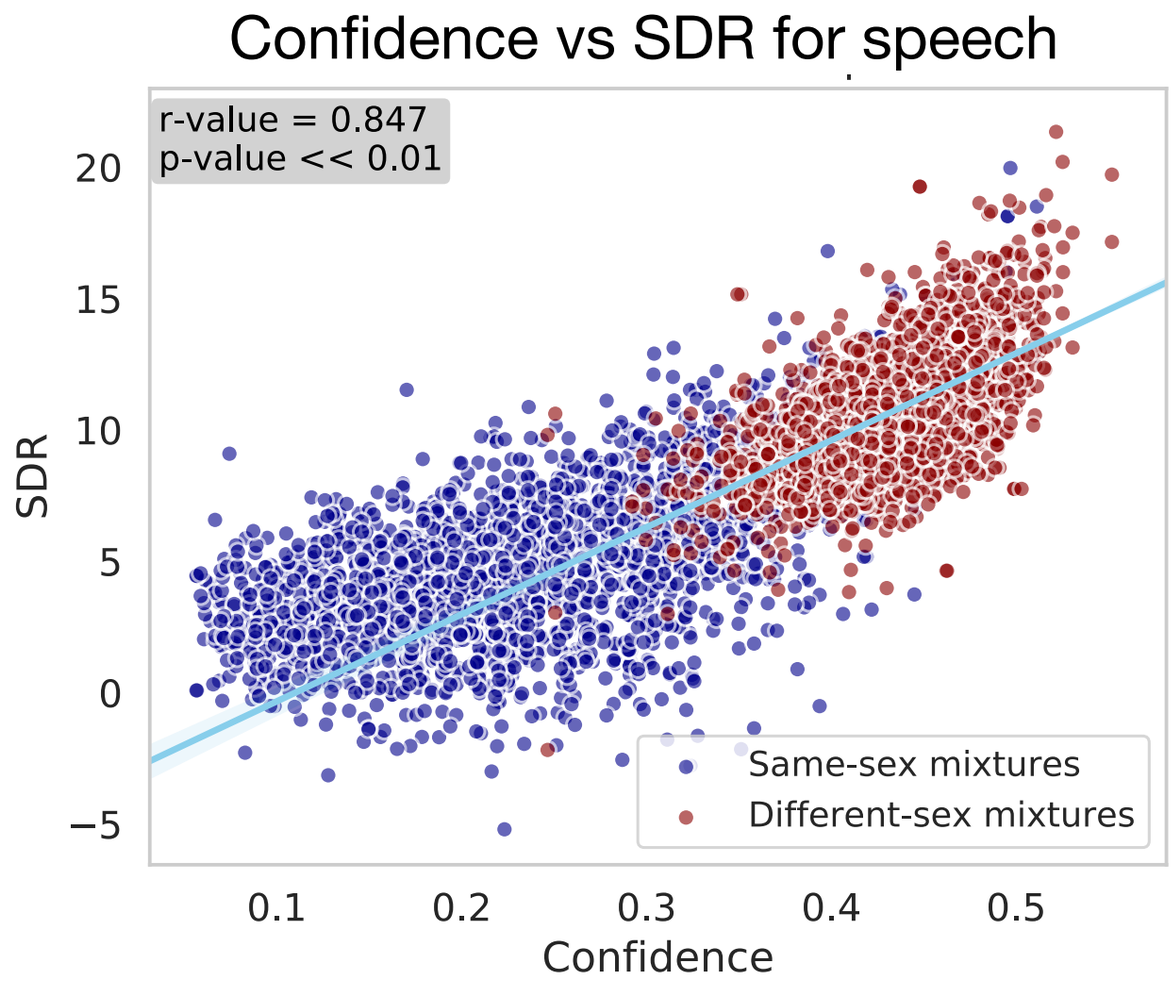
- 2 BLSTM layers
- 300 hidden units in both directions

2D embedding visualization of a music mixture



Evaluation of confidence measure

Strong correlations between the confidence measure and ground-truth separation quality



Confidence-mediated ensemble

Compare different methods of choosing the appropriate model given an audio mixture

Approach	Speech	Music	Environ.
Oracle ensemble			
Confidence ensemble			
Random ensemble			
Speech model			
Music model			
Environ. model			

Confidence-mediated ensemble

Compare different methods of choosing the appropriate model given an audio mixture

Approach	Speech	Music	Environ.
Oracle ensemble			
Confidence ensemble			
Random ensemble			
Speech model			
Music model			
Environ. model			

Confidence-mediated ensemble

Compare different methods of choosing the appropriate model given an audio mixture

Approach	Speech	Music	Environ.
Oracle ensemble			
Confidence ensemble			
Random ensemble			
Speech model			
Music model			
Environ. model			

Confidence-mediated ensemble

Compare different methods of choosing the appropriate model given an audio mixture

Approach	Speech	Music	Environ.
Oracle ensemble	8.3	6.5	12.2
Confidence ensemble			
Random ensemble			
Speech model			
Music model			
Environ. model			

Oracle: select model with best ground-truth performance

Confidence-mediated ensemble

Compare different methods of choosing the appropriate model given an audio mixture

Approach	Speech	Music	Environ.
Oracle ensemble	8.3	6.5	12.2
Confidence ensemble	7.6	6.4	10.5
Random ensemble			
Speech model			
Music model			
Environ. model			

Confidence: select model with highest confidence

Confidence-mediated ensemble

Compare different methods of choosing the appropriate model given an audio mixture

Approach	Speech	Music	Environ.
Oracle ensemble	8.3	6.5	12.2
Confidence ensemble	7.6	6.4	10.5
Random ensemble	4.8	4.2	2.8
Speech model			
Music model			
Environ. model			

Random: select model randomly with equal probability

Confidence-mediated ensemble

Compare different methods of choosing the appropriate model given an audio mixture

Approach	Speech	Music	Environ.
Oracle ensemble	8.3	6.5	12.2
Confidence ensemble	7.6	6.4	10.5
Random ensemble	4.8	4.2	2.8
Speech model	8.2	2.0	3.0
Music model	1.4	6.5	2.5
Environ. model	2.1	1.7	11.9

Domain-specific models: applied without switching

Confidence-mediated ensemble

Compare different methods of choosing the appropriate model given an audio mixture

Approach	Speech	Music	Environ.
Oracle ensemble	8.3	6.5	12.2
Confidence ensemble	7.6	6.4	10.5
Random ensemble	4.8	4.2	2.8
Speech model	8.2	2.0	3.0
Music model	1.4	6.5	2.5
Environ. model	2.1	1.7	11.9

Domain-specific models: applied without switching

Confidence-mediated ensemble

Compare different methods of choosing the appropriate model given an audio mixture

Approach	Speech	Music	Environ.
Oracle ensemble	8.3	6.5	12.2
Confidence ensemble	7.6	6.4	10.5
Random ensemble	4.8	4.2	2.8
Speech model	8.2	2.0	3.0
Music model	1.4	6.5	2.5
Environ. model	2.1	1.7	11.9

Domain-specific models: applied without switching

Confidence-mediated ensemble

Compare different methods of choosing the appropriate model given an audio mixture

Approach	Speech	Music	Environ.
Oracle ensemble	8.3	6.5	12.2
Confidence ensemble	7.6	6.4	10.5
Random ensemble	4.8	4.2	2.8
Speech model	8.2	2.0	3.0
Music model	1.4	6.5	2.5
Environ. model	2.1	1.7	11.9

Confidence-mediated ensemble

Compare different methods of choosing the appropriate model given an audio mixture

Approach	Speech	Music	Environ.
Oracle ensemble	8.3	6.5	12.2
Confidence ensemble	7.6	6.4	10.5
Random ensemble	4.8	4.2	2.8
Speech model	8.2	2.0	3.0
Music model	1.4	6.5	2.5
Environ. model	2.1	1.7	11.9

Confidence-mediated ensemble

Compare different methods of choosing the appropriate model given an audio mixture

Approach	Speech	Music	Environ.
Oracle ensemble	8.3	6.5	12.2
Confidence ensemble	7.6	6.4	10.5
Random ensemble	4.8	4.2	2.8
Speech model	8.2	2.0	3.0
Music model	1.4	6.5	2.5
Environ. model	2.1	1.7	11.9

Conclusion

- Confidence measure effectively *estimates the performance* of clustering-based source separation algorithms
- Apply the confidence measure to effectively *select the appropriate model* for a given mixture