Overlap-aware speaker diarization:

New methods and ensembles

Desh Raj

CLSP Seminar
January 29, 2021

CENTER FOR LANGUAGE AND SPEECH PROCESSING

Collaborators

Zili **Huang**

Paola **Garcia**

Sanjeev **Khudanpur**

Shinji **Watanabe**

Dan **Povey**

Andreas **Stolcke**

CENTER FOR LANGUAGE AND SPEECH PROCESSING

Carnegie Mellon University School of Computer Science

xiaomi

amazon alexa

Overview

- A brief **background in diarization**:
    - The task and its applications
    - A traditional solution, and the problem of overlapping speech
- **Method**: A step towards making the traditional solution overlap-aware
- **Ensemble**: An approach for combining overlap-aware diarization systems

Background

What is speaker diarization?

Task of “who spoke when”

Input: ***recording containing multiple speakers***

Output: ***homogeneous speaker segments***

Xavier Anguera Miro et al., “Speaker diarization: A review of recent research,” IEEE Transactions on Audio, Speech, and Language Processing, 2012.

Background

Applications of Diarization

Psychotherapy and human interaction

Child language acquisition

Collaborative learning

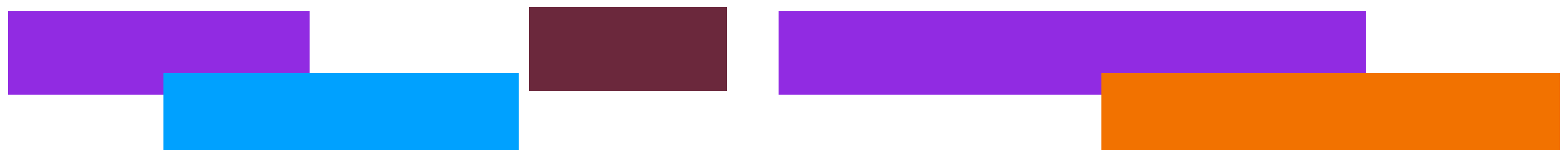
Meeting transcription

Cocktail party problem

Background

What makes Diarization difficult?

Input: *recording containing multiple speakers*



Output: *homogeneous speaker segments*

1. The recording may be very long with arbitrary silences/noise.
2. Number of speakers may be unknown.
3. Overlapping speech may be present.

Example from CHiME-6 challenge (best system achieved >30% error rate)

Watanabe, S., Mandel, M., Barker, J., & Vincent, E. (2020). CHiME-6 Challenge: Tackling Multispeaker Speech Recognition for Unsegmented Recordings. *ArXiv*.

The traditional solution

“Clustering-based” systems

- **Key idea:** formulate Diarization as a clustering problem
- Cluster small segments of audio
- Each cluster represents a distinct speaker

Basu, J., Khan, S., Roy, R., Pal, M., Basu, T., Bepari, M.S., & Basu, T.K. (2016). An overview of speaker diarization: Approaches, resources and challenges.

Tranter, S., & Reynolds, D. (2006). An overview of automatic speaker diarization systems. IEEE Transactions on Audio, Speech, and Language Processing.

Clustering-based diarization

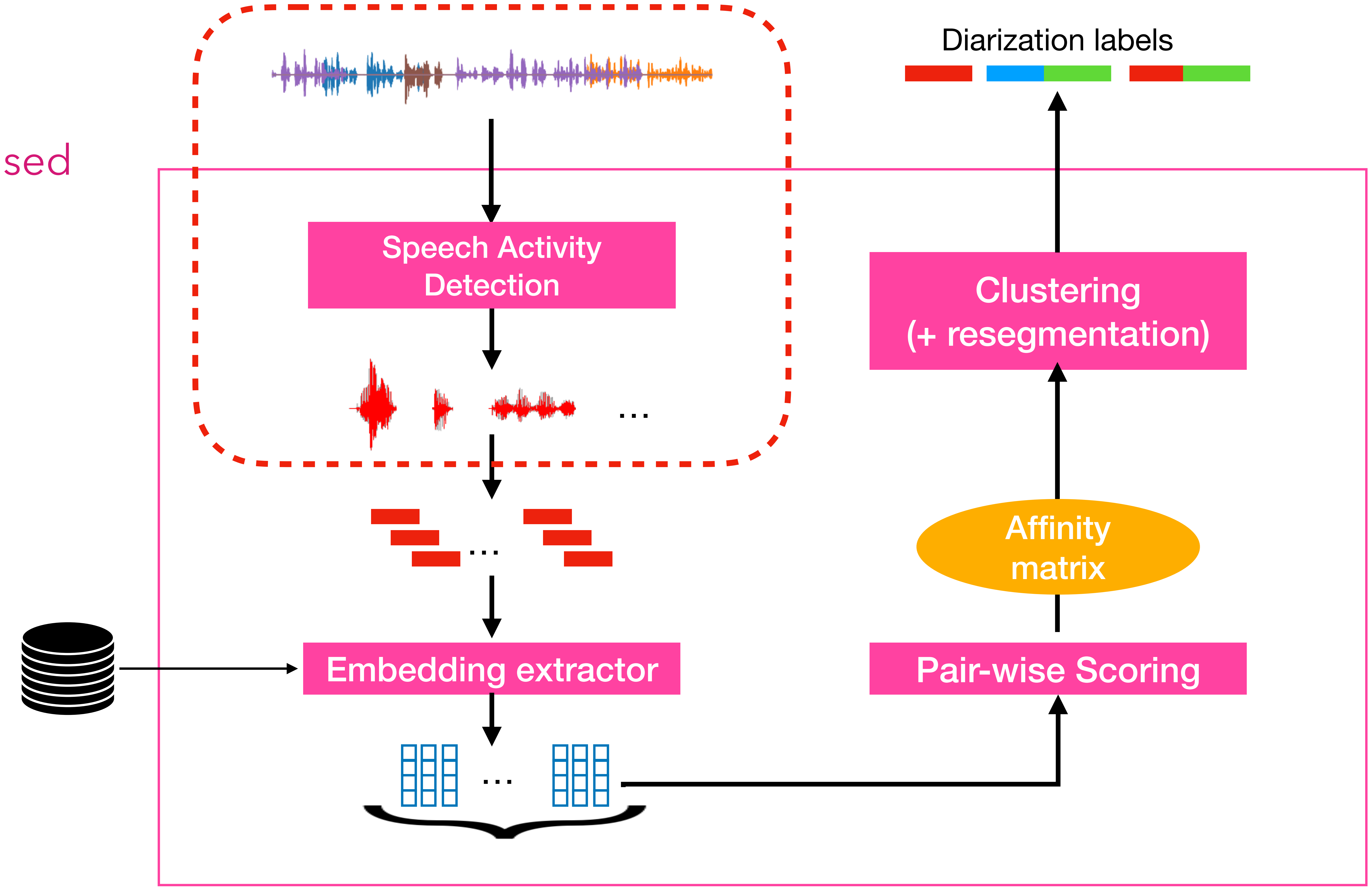
SAD extracts speech segments from recordings

Spectral energy-based

GMM-based

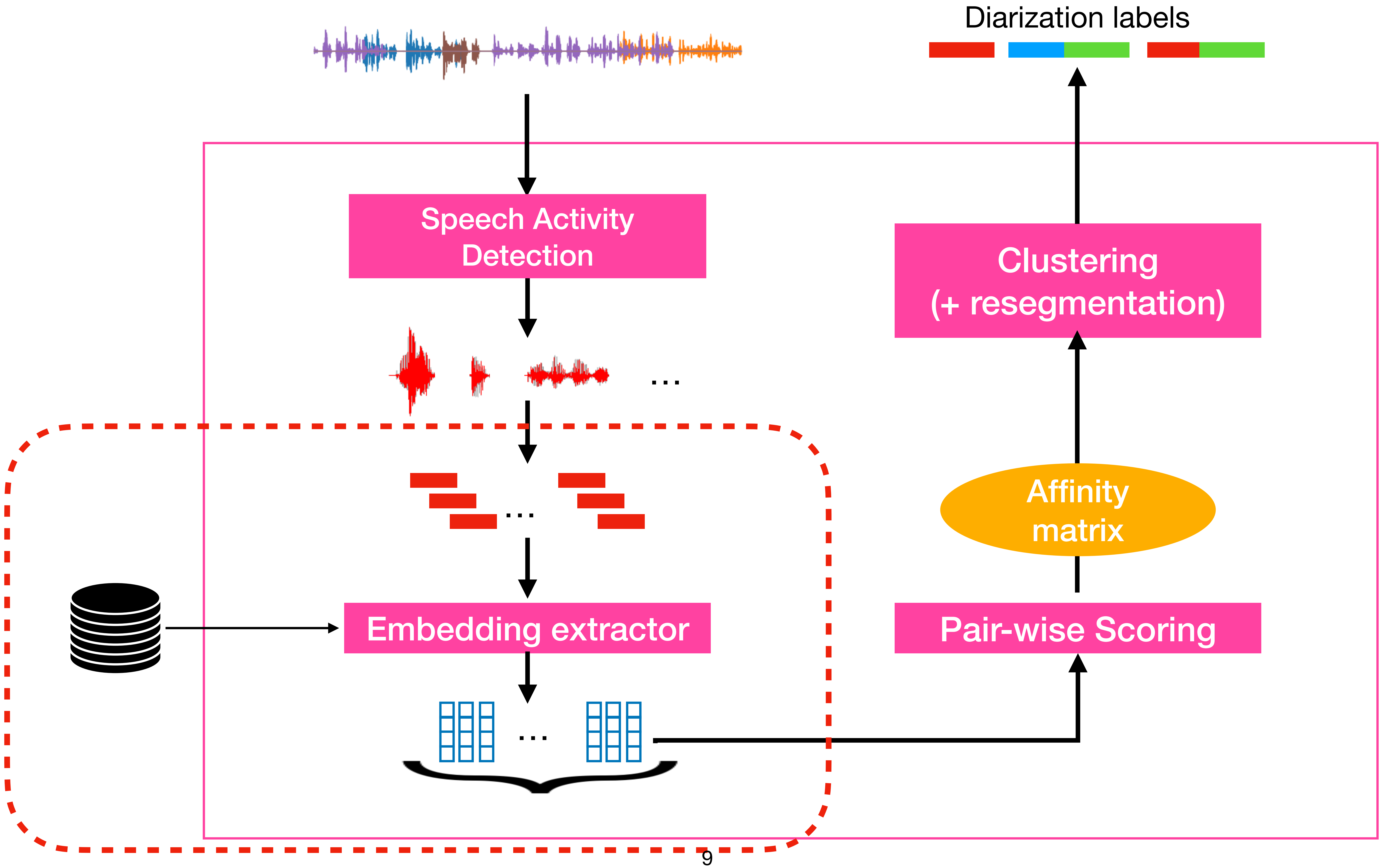
Hybrid HMM-DNN

End-to-end SAD



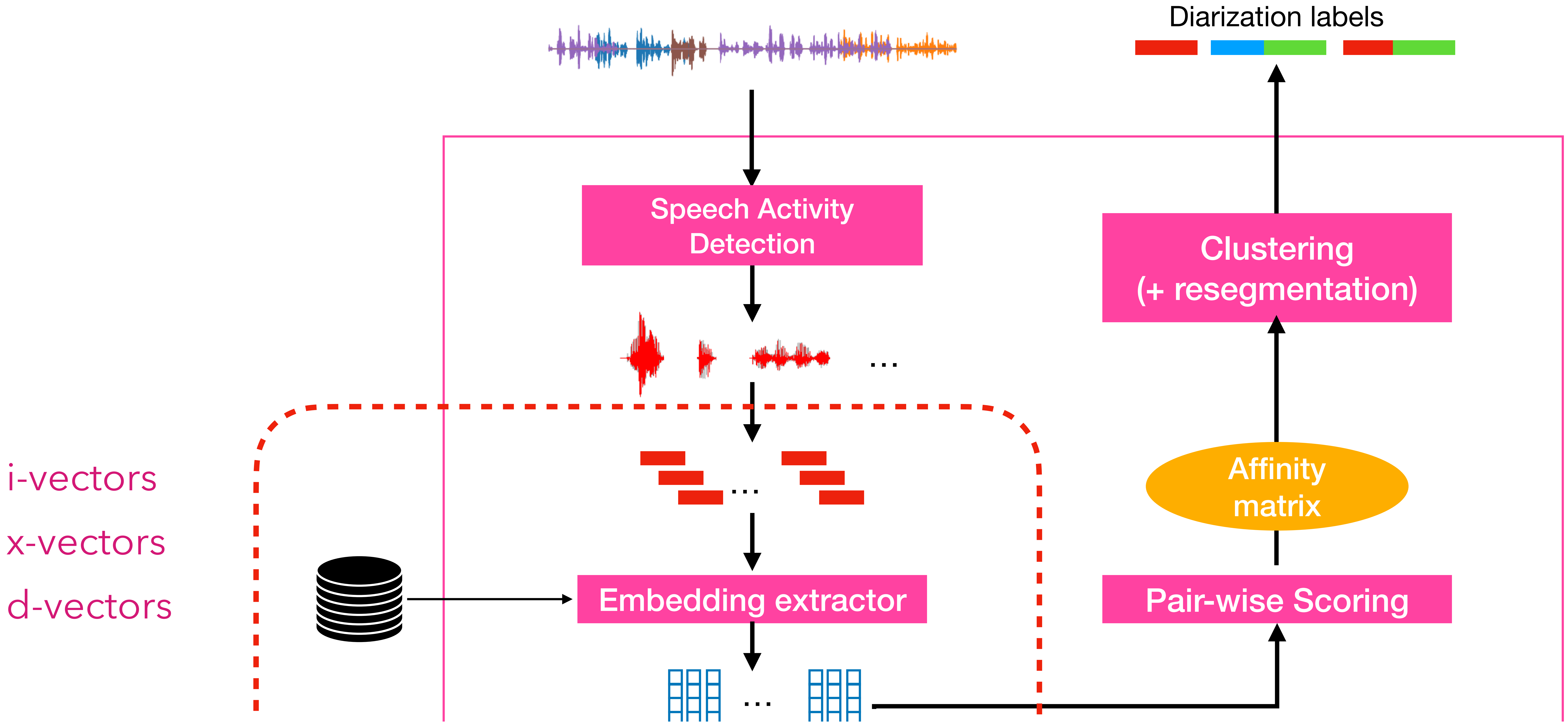
Clustering-based diarization

Embeddings extracted for small subsegments



Clustering-based diarization

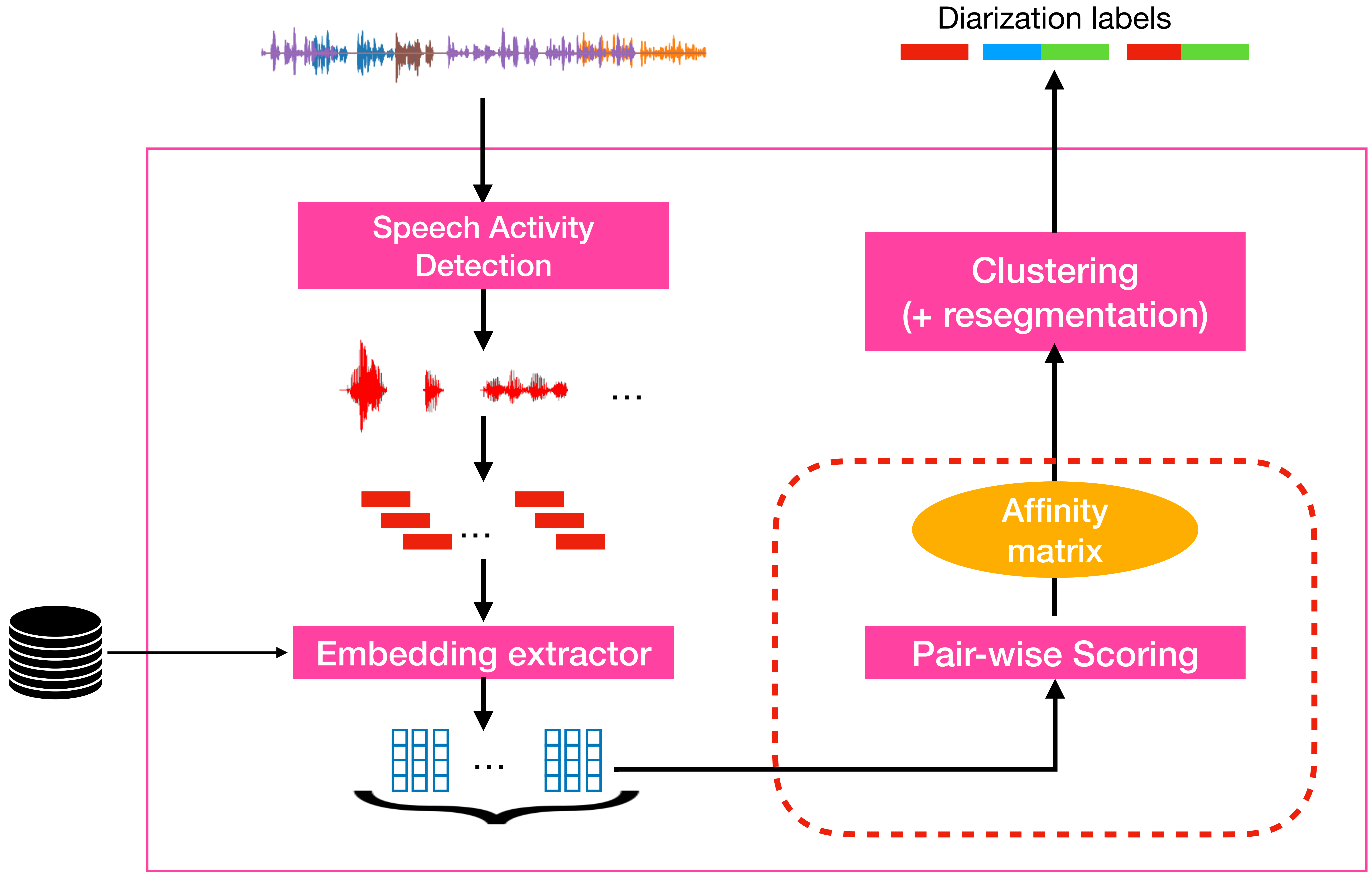
Embeddings extracted for small subsegments



Dehak, N., et al (2011). Front-End Factor Analysis for Speaker Verification. *IEEE Transactions on Audio, Speech, and Language Processing*.
Snyder, D., et al. (2018). X-Vectors: Robust DNN Embeddings for Speaker Recognition. *2018 IEEE ICASSP*.
Variani, E.,et al. (2014). Deep neural networks for small footprint text-dependent speaker verification. 2014 IEEE ICASSP.

Clustering-based diarization

Pair-wise scoring of subsegments



PLDA scoring
Cosine scoring

Sell, G., & Garcia-Romero, D. (2014). Speaker diarization with PLDA i-vector scoring and unsupervised calibration. *2014 IEEE Spoken Language Technology Workshop (SLT).*

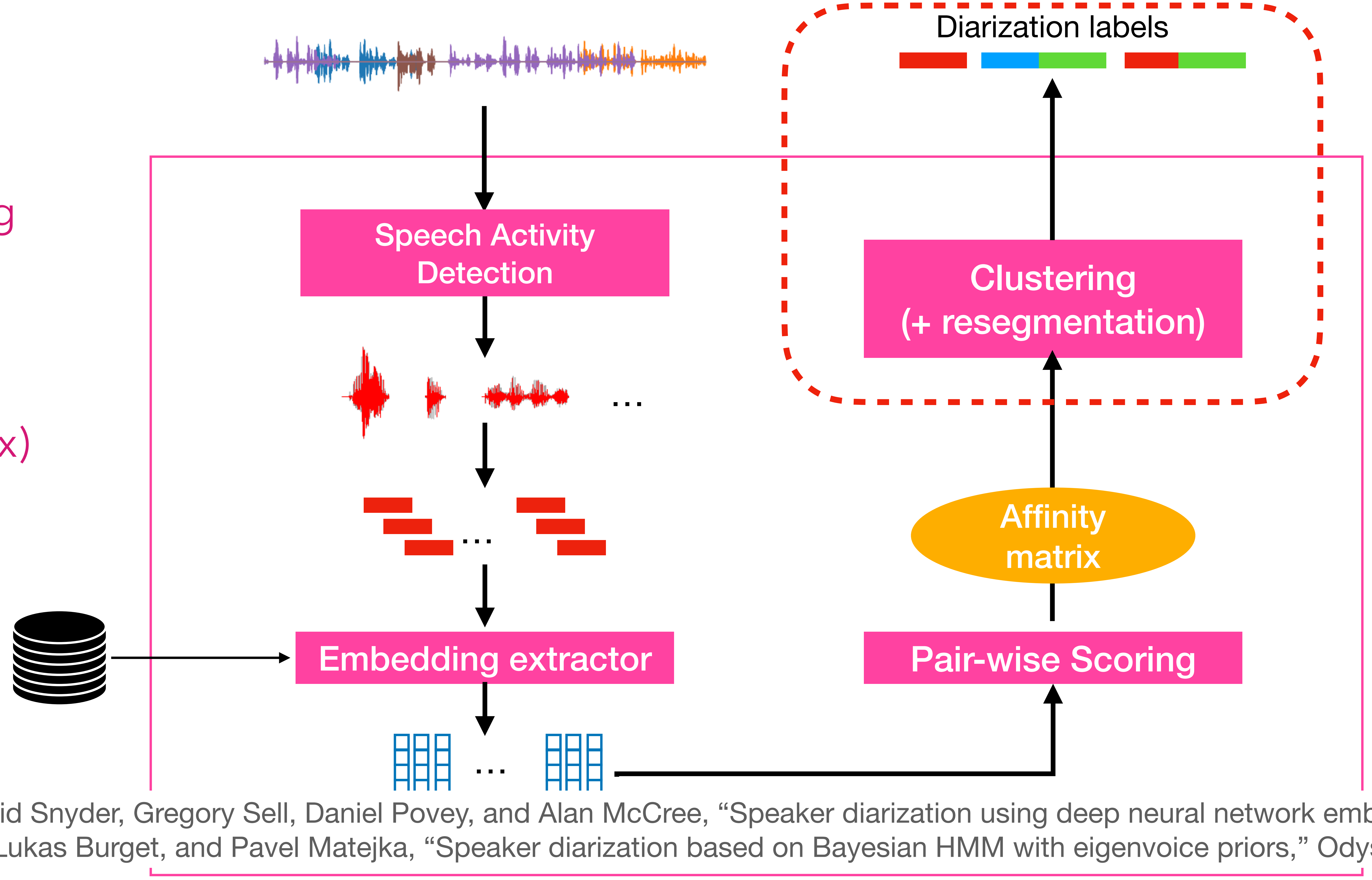
Clustering-based diarization

Clustering based on the affinity matrix, followed by optional resegmentation

Agglomerative hierarchical clustering

Spectral clustering

Variational Bayes (VBx)



Daniel Garcia-Romero, David Snyder, Gregory Sell, Daniel Povey, and Alan McCree, “Speaker diarization using deep neural network embeddings,” ICASSP 2017.
Mireia Dîez, Lukas Burget, and Pavel Matejka, “Speaker diarization based on Bayesian HMM with eigenvoice priors,” Odyssey 2018.

Clustering-based diarization

How well does it perform?

- **Winning system in DIHARD I (2018) and II (2019)**
- DIHARD contains “hard” Diarization evaluation with recordings from several domains
- But Diarization error rates (DER) still high: 37% in DIHARD I and 27% in DIHARD II

$$\text{DER} = \frac{\text{Missed speech} + \text{False alarm} + \text{Speaker error}}{\text{Total speaking time}}$$

Sell, G., et al. (2018). Diarization is Hard: Some Experiences and Lessons Learned for the JHU Team in the Inaugural DIHARD Challenge. *INTERSPEECH 2018*.
Landini, F., et al. (2020). BUT System for the Second Dihard Speech Diarization Challenge. *IEEE ICASSP 2020.*

CENTER FOR LANGUAGE AND SPEECH PROCESSING

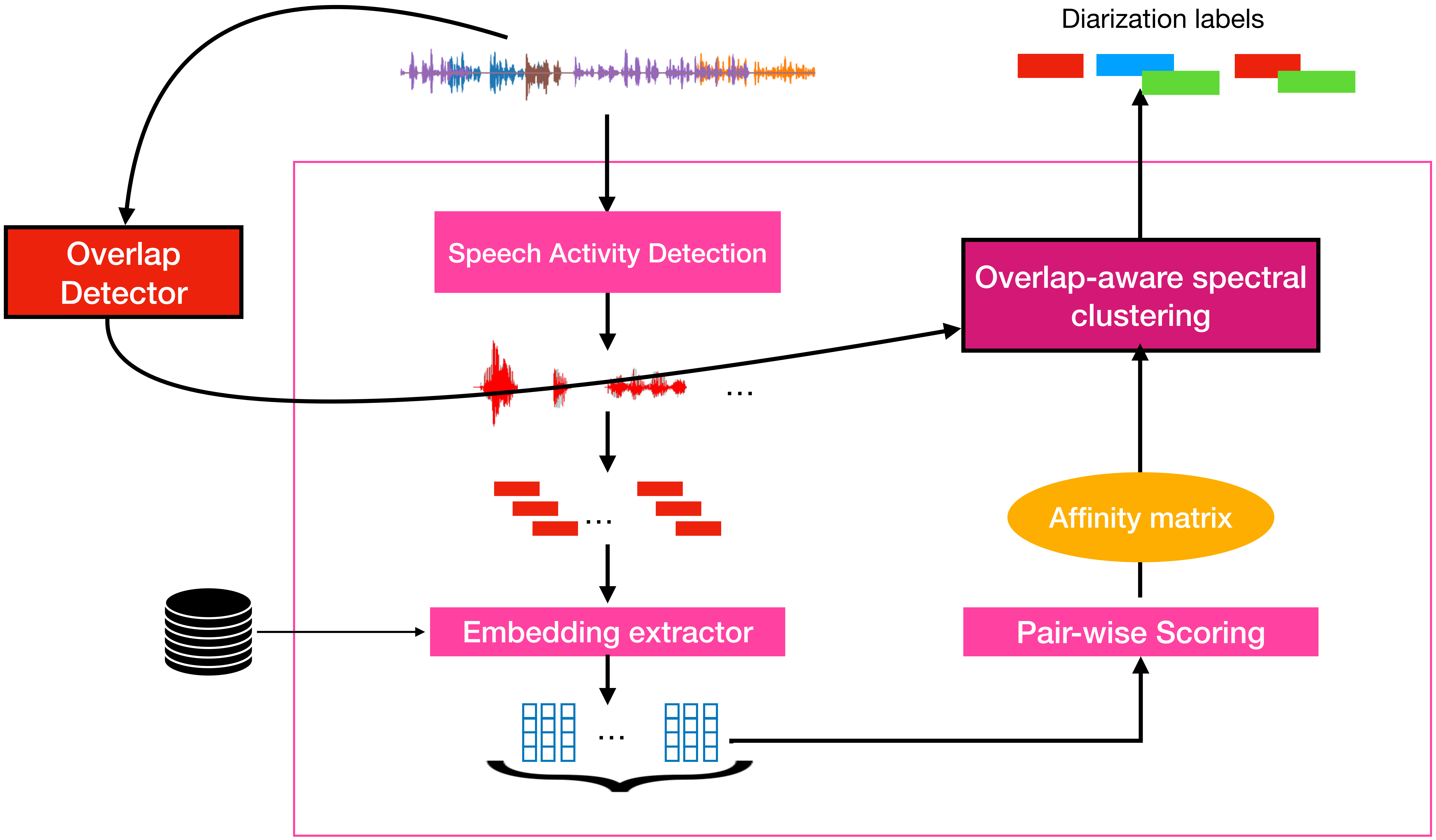
Clustering paradigm assumes single-speaker segments

So overlapping speakers are completely ignored!

> *"Roughly **8% of the absolute error** in our systems was from overlapping speech … it will likely require a **complete rethinking of the diarization process** … This is an important direction, but could not be addressed …"*
> ***- JHU team (2018)***

> *"Given the current performance of the systems, the **overlapped speech gains more relevance** … **more than 50% of the DER** in our best systems … has to be addressed in the future …"*
> ***- BUT team (2019)***

Overlap-aware spectral clustering



Raj, D., Huang, Z., & Khudanpur, S. (2021). Multi-class Spectral Clustering with Overlaps for Speaker Diarization. *IEEE SLT 2021*.

Overlap-aware spectral clustering

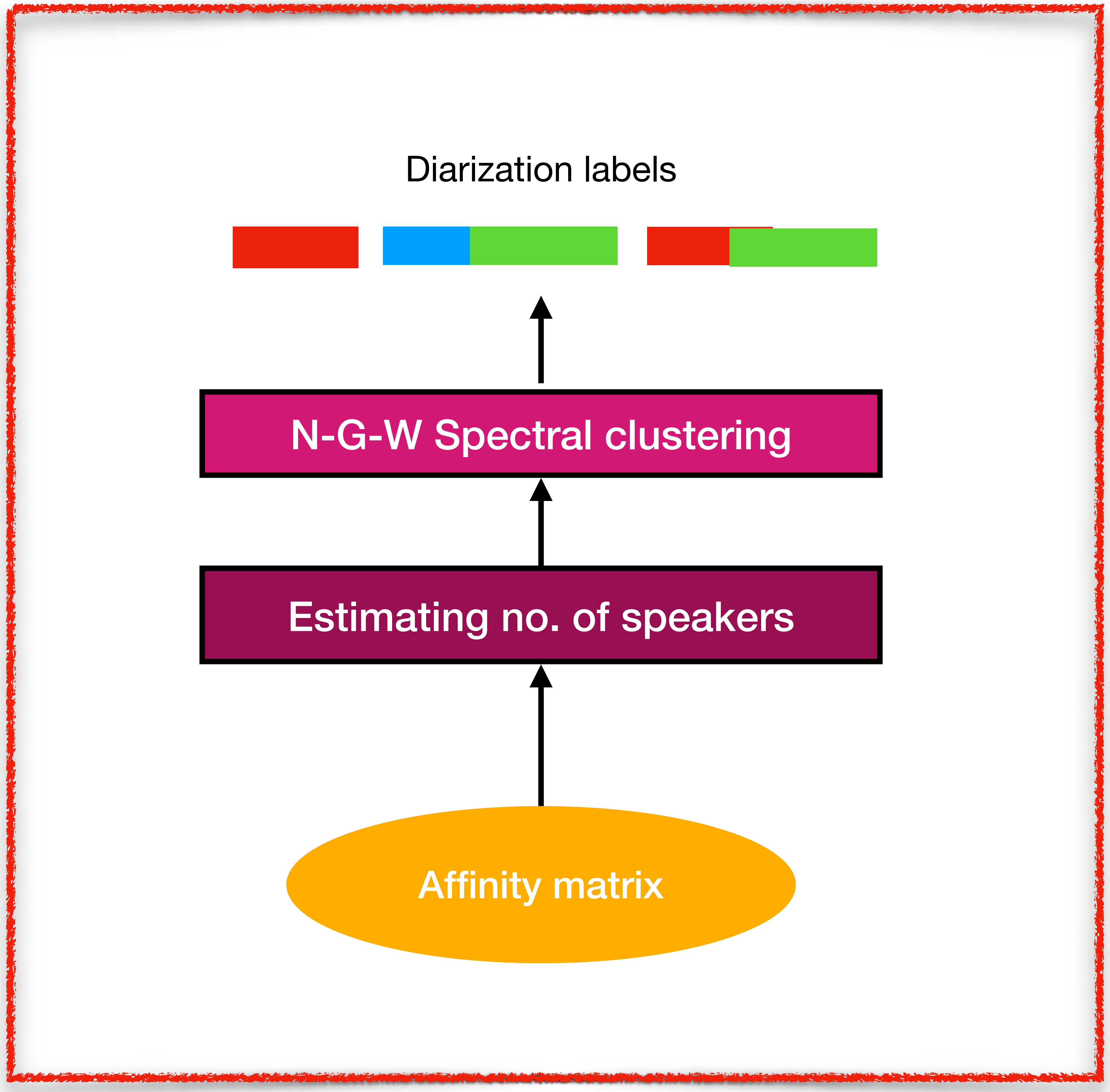
Overview of differences

Regular spectral clustering

(Ng-Jordan-Weiss algorithm):

- Estimate number of speakers (say, K)
- Compute Laplacian L of affinity matrix
- Apply K-means clustering on first K eigenvectors of L

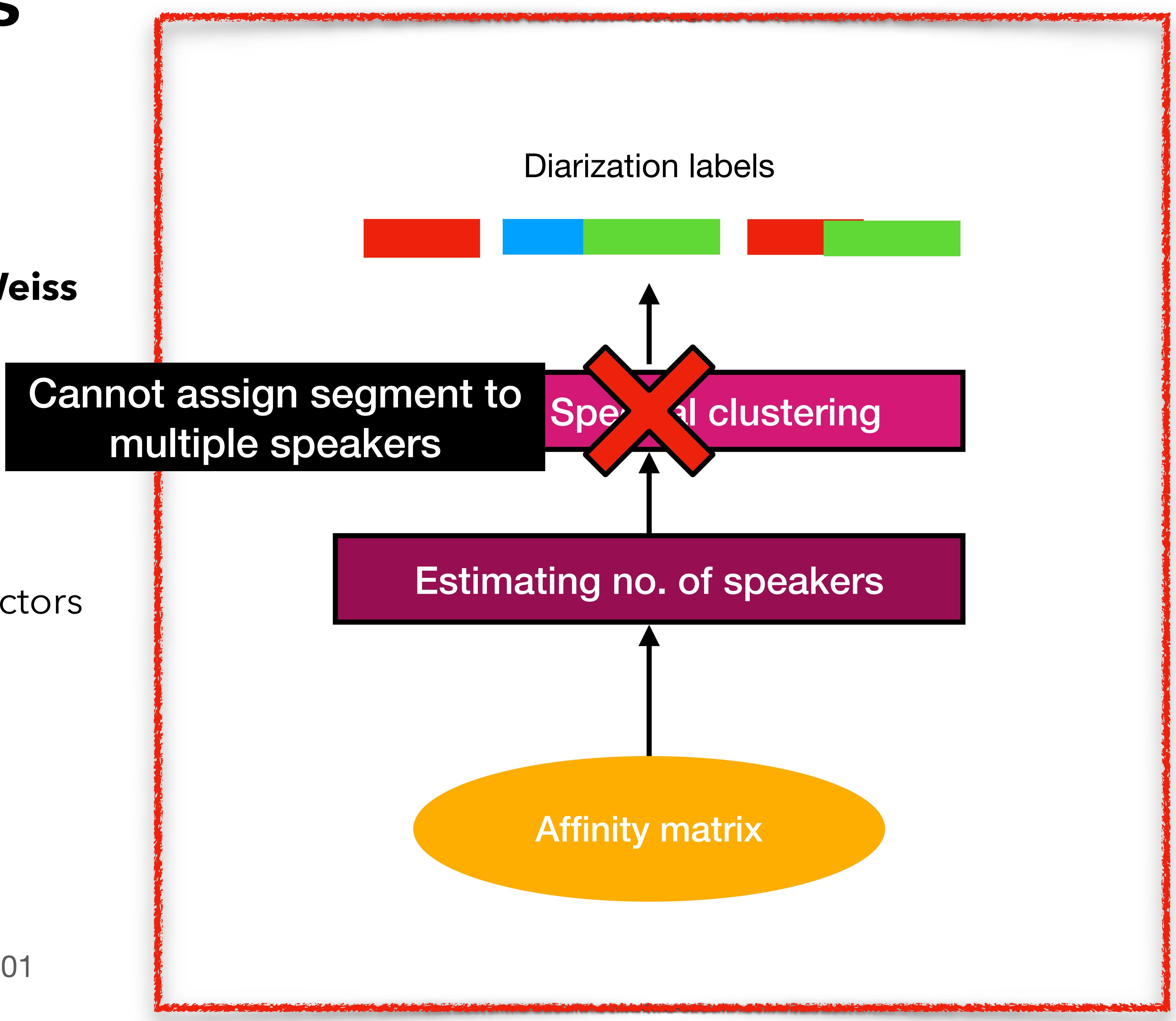
Andrew Y. Ng, Michael I. Jordan, and Yair Weiss, “On spectral clustering: Analysis and an algorithm,” NIPS, 2001



Overlap-aware spectral clustering

Overview of differences

- **Regular spectral clustering (Ng-Jordan-Weiss algorithm):**
- Estimate number of speakers (say, K)
- Compute Laplacian L of affinity matrix
- Apply K-means clustering on first K eigenvectors of L



Andrew Y. Ng, Michael I. Jordan, and Yair Weiss, “On spectral clustering: Analysis and an algorithm,” NIPS, 2001

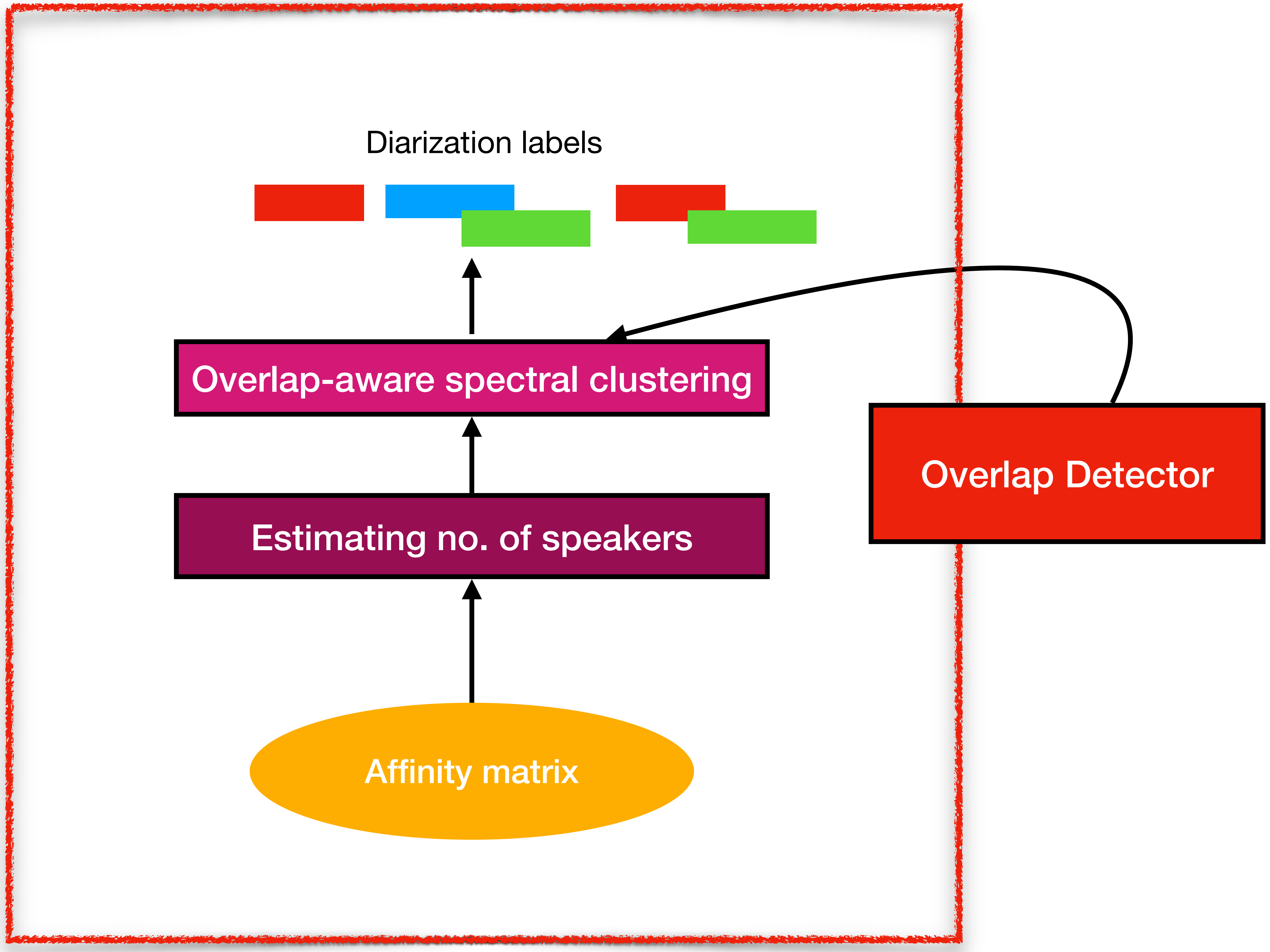
Overlap-aware spectral clustering

Overview of differences

Alternative formulation:

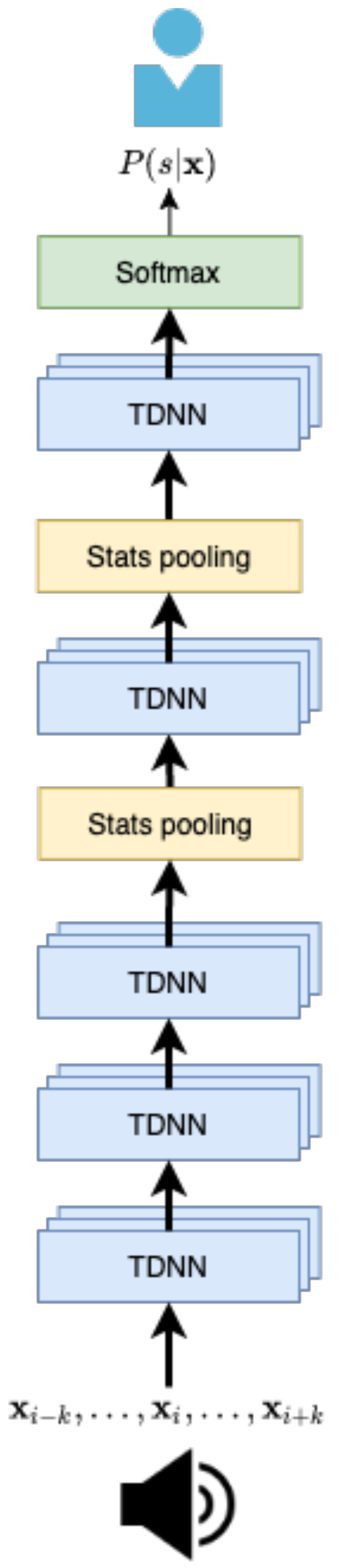
multi-class spectral clustering

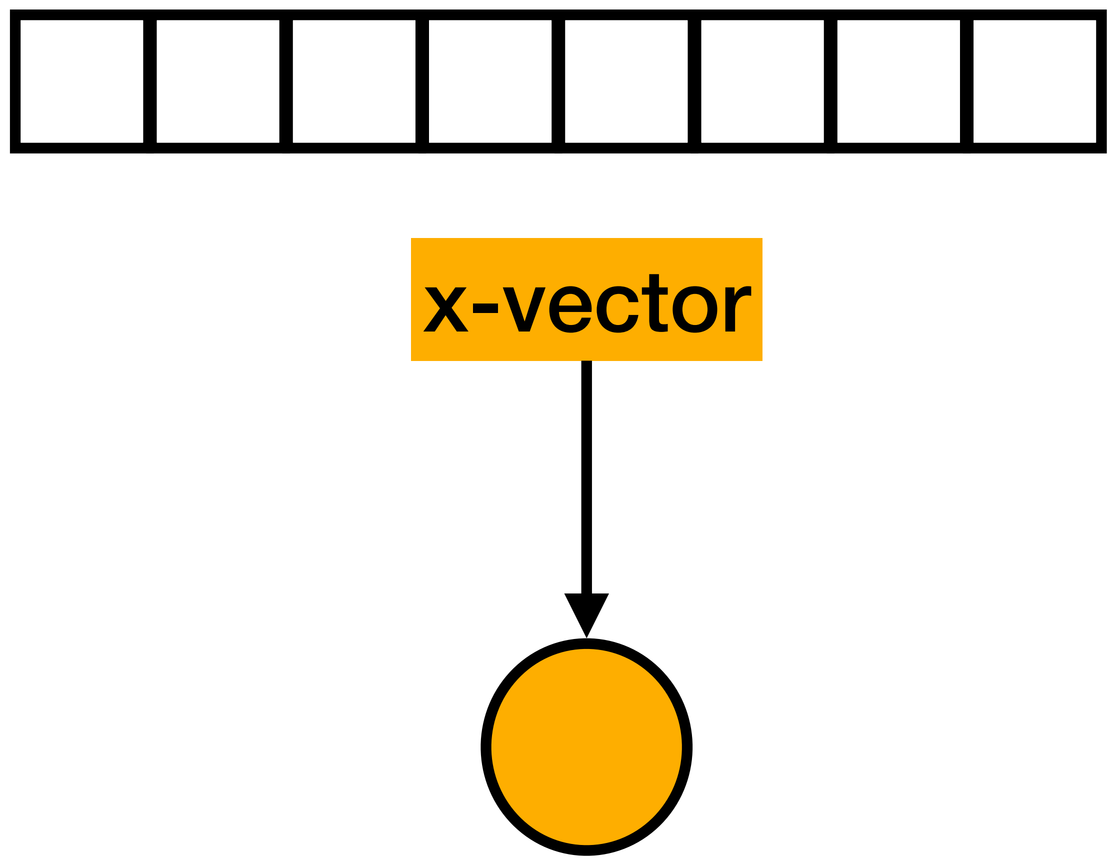
Yu, S., & Shi, J. Multiclass spectral clustering. ICCV 2003.



New formulation for spectral clustering

The basic clustering problem: a graph view



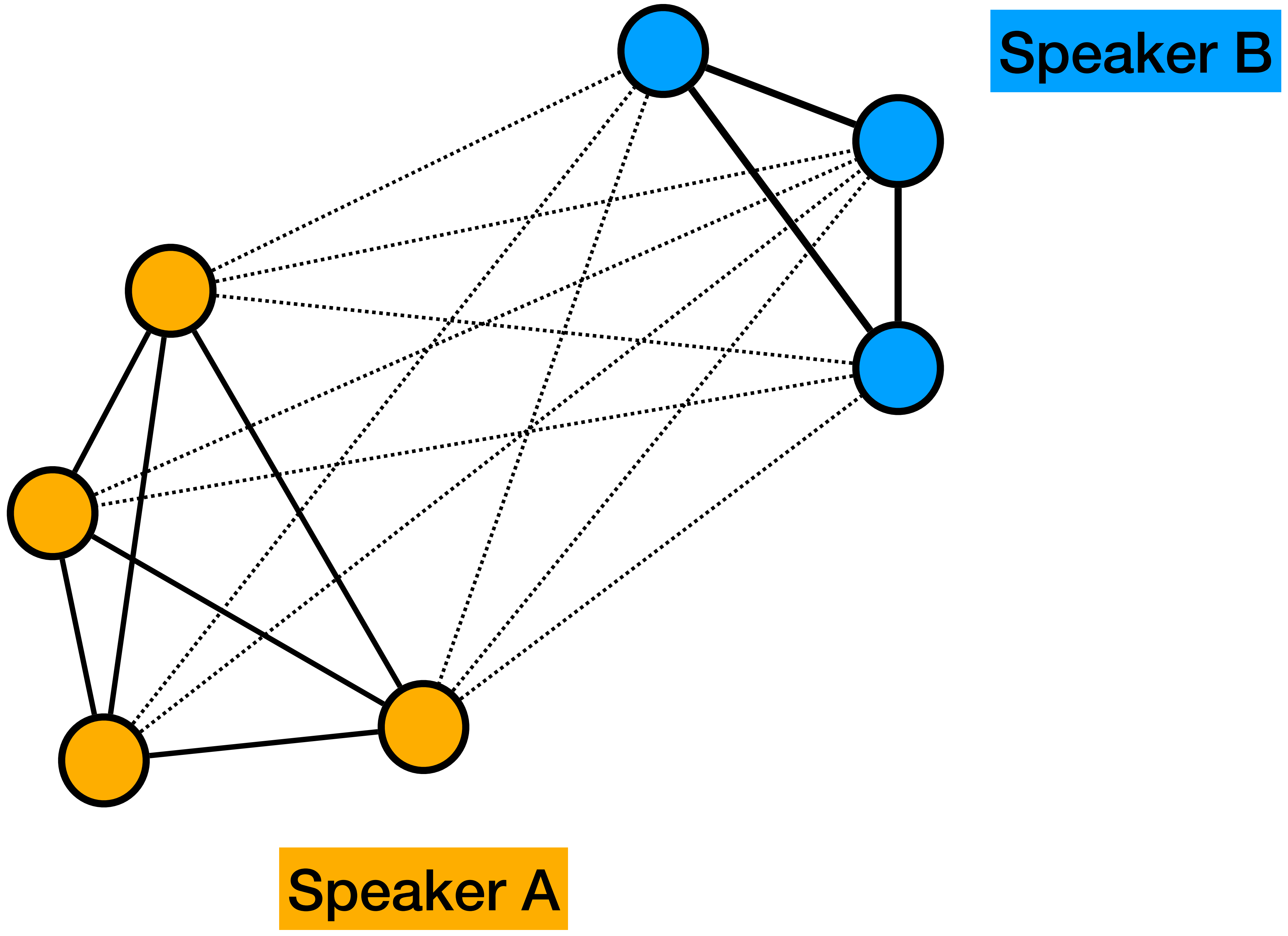


Cosine similarity

Snyder, D., et al. (2018). X-Vectors: Robust DNN Embeddings for Speaker Recognition. *2018 IEEE ICASSP*.

New formulation for spectral clustering

The basic clustering problem: a graph view

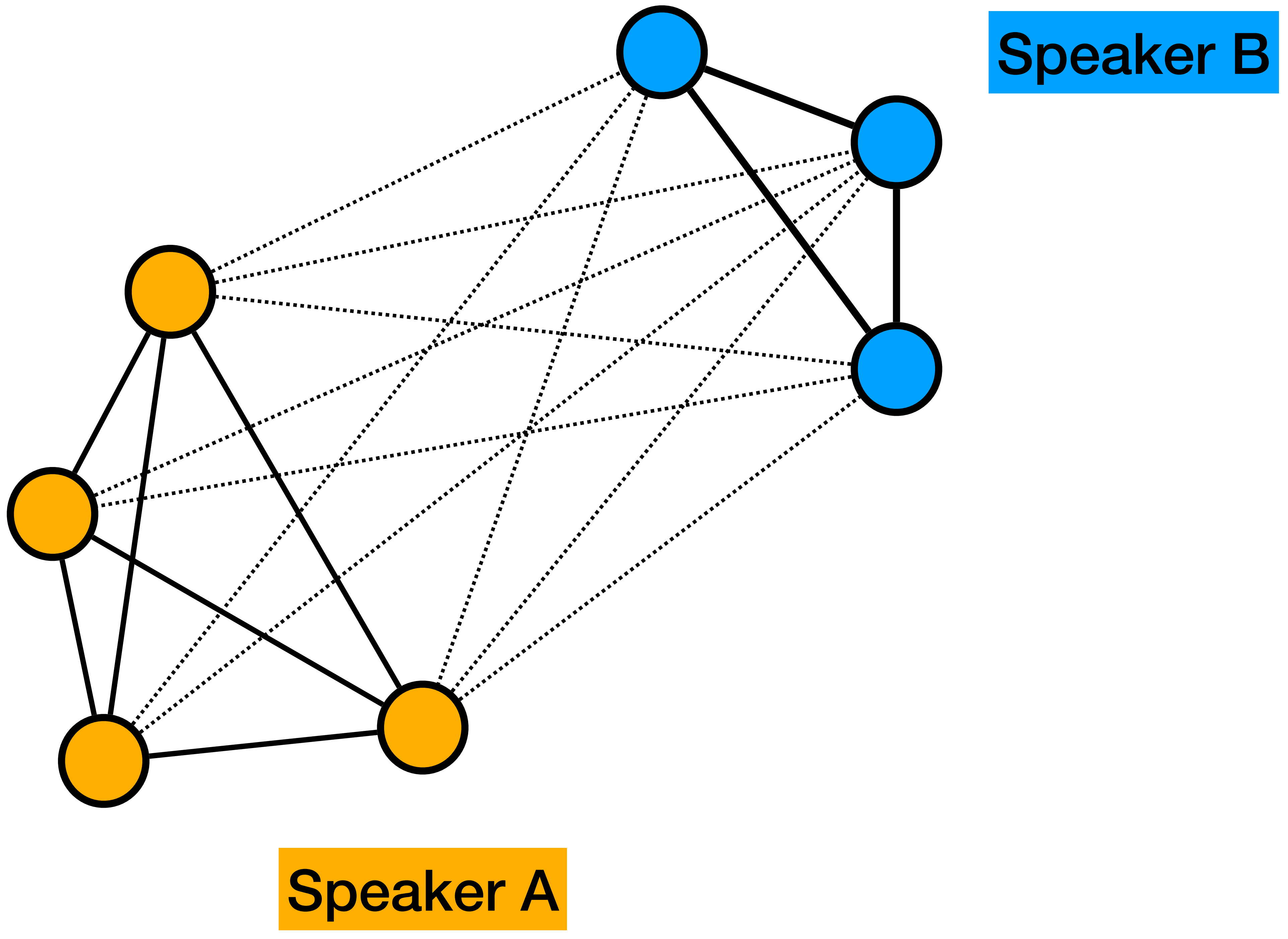


Edge weights within a group ↑

Edge weights across groups ↓

New formulation for spectral clustering

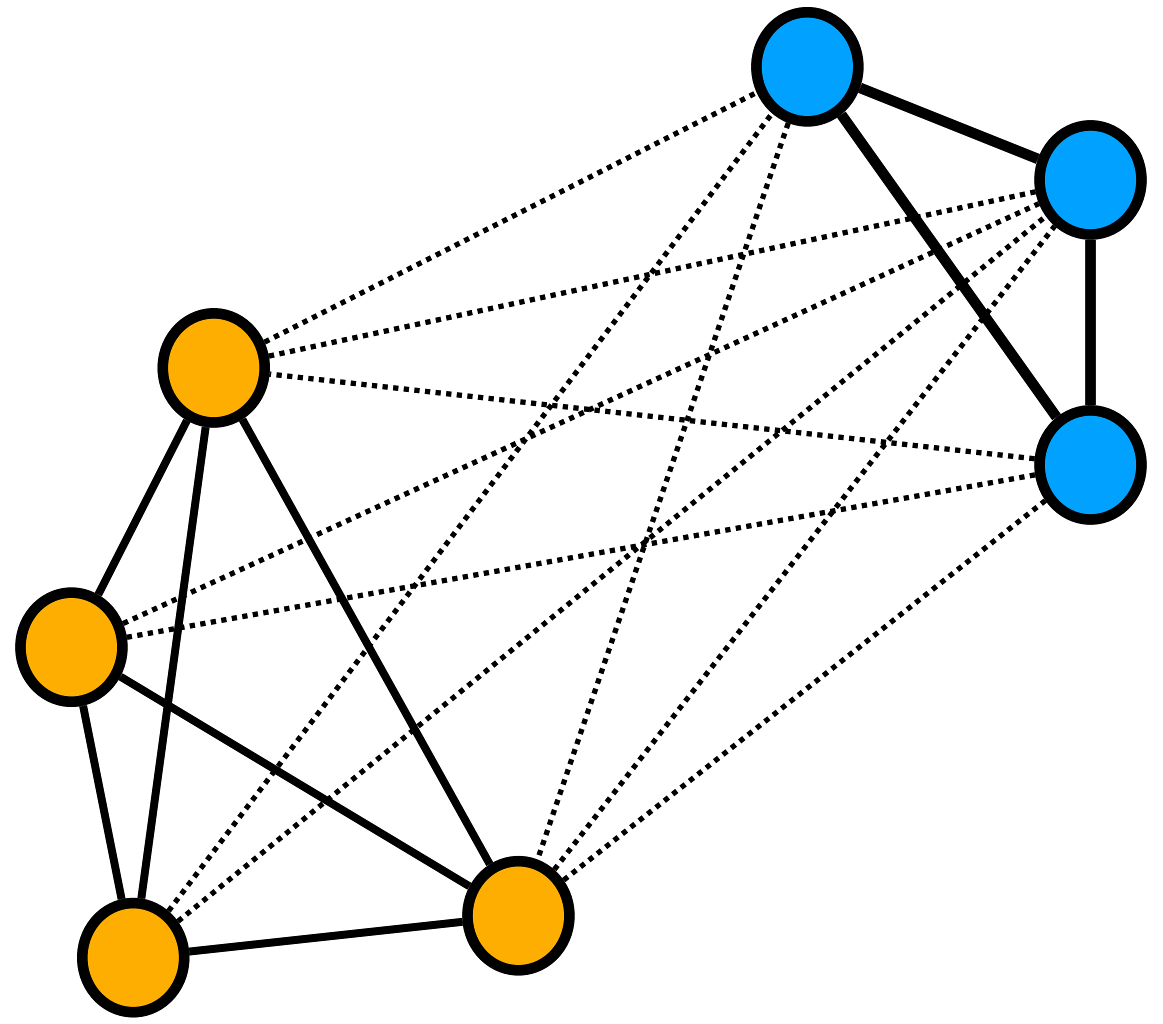
The basic clustering problem: a graph view



$$\textit{maximize} \quad \frac{\text{Edge weights within a group}}{\text{Edge weights across groups}}$$

New formulation for spectral clustering

The basic clustering problem: a graph view



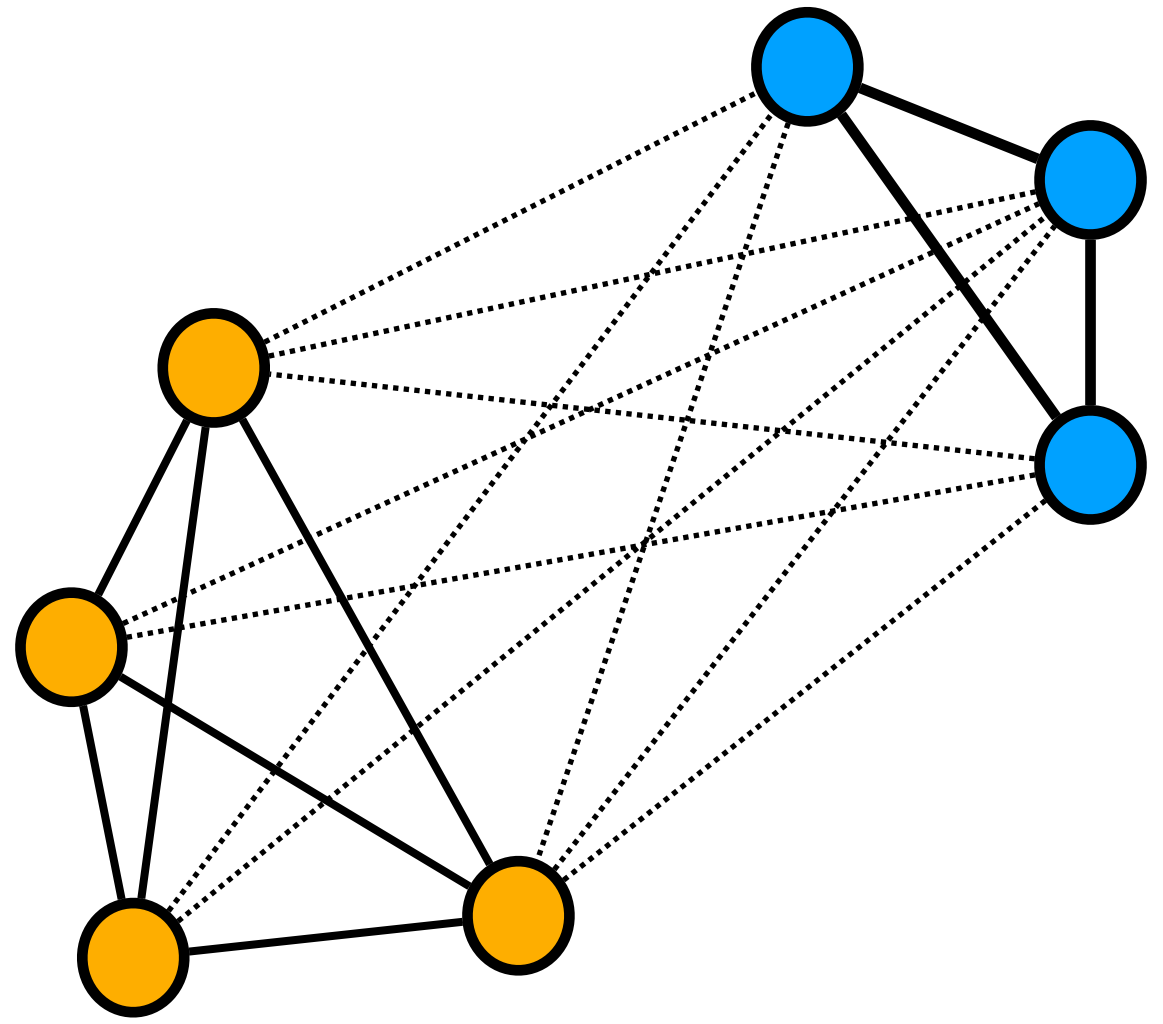
$$\textit{maximize} \quad \frac{\text{Edge weights within a group}}{\text{Edge weights across groups}}$$

$$
\begin{aligned}
\text{maximize} \quad & \epsilon(X) = \frac{1}{K}\sum_{k=1}^{K}\frac{X_k^T \mathbf{A} X_k}{X_k^T \mathbf{D} X_k} \\
\text{subject to} \quad & X \in \{0,1\}^{N\times K}, \\
& X\mathbf{1}_K = \mathbf{1}_N .
\end{aligned}
$$

K speakers, **N** segments

New formulation for spectral clustering

The basic clustering problem: a graph view



maximize $\dfrac{\text{Edge weights within a group}}{\text{Edge weights across groups}}$

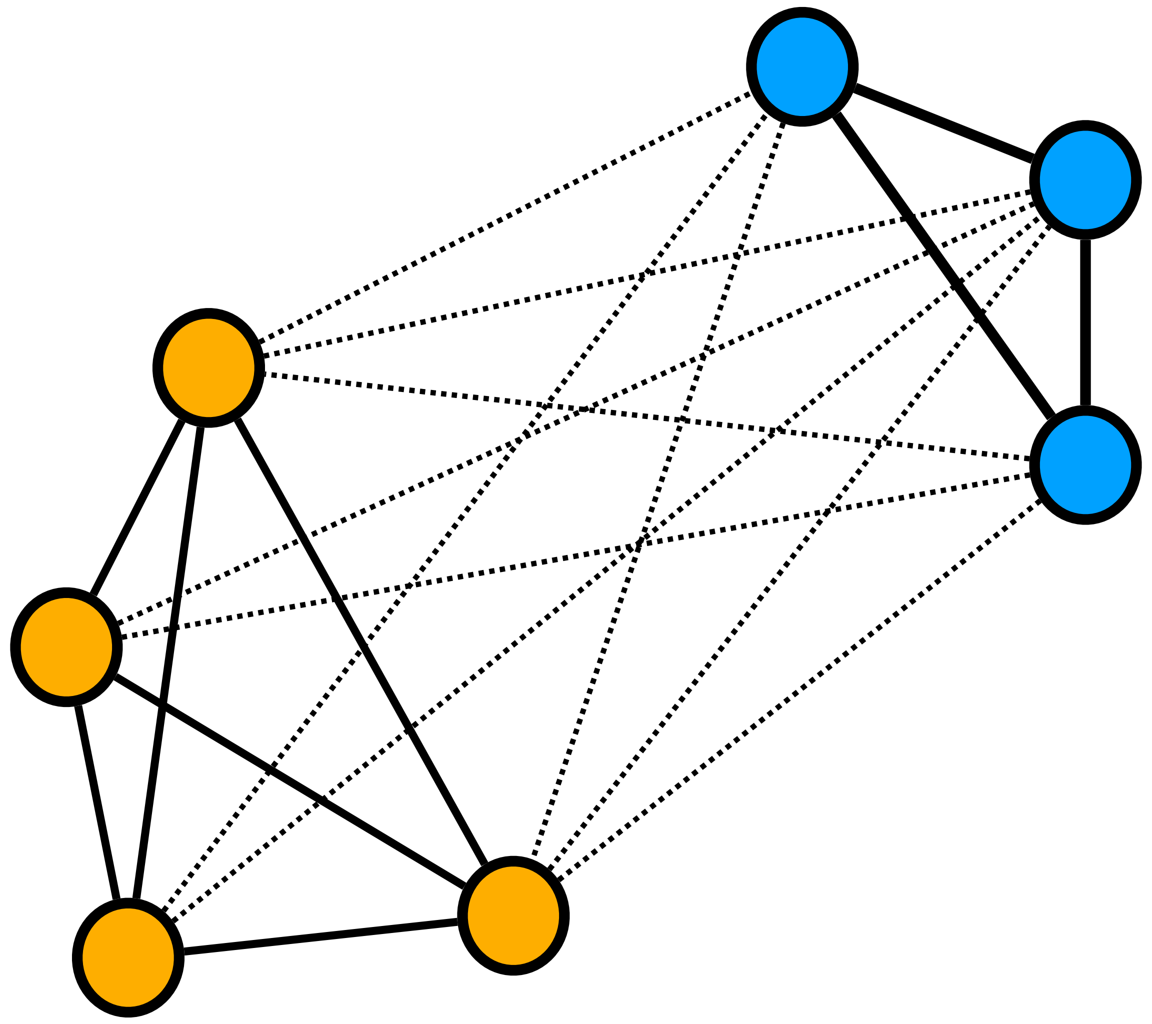
$$\begin{aligned}\text{maximize} \quad & \epsilon(X) = \frac{1}{K}\sum_{k=1}^{K}\frac{X_k^T \mathbf{A} X_k}{X_k^T \mathbf{D} X_k}\\ \text{subject to} \quad & X \in \{0,1\}^{N\times K},\\ & X\mathbf{1}_K = \mathbf{1}_N.\end{aligned}$$

Affinity matrix

Diagonal matrix containing degree of nodes

New formulation for spectral clustering

The basic clustering problem: a graph view

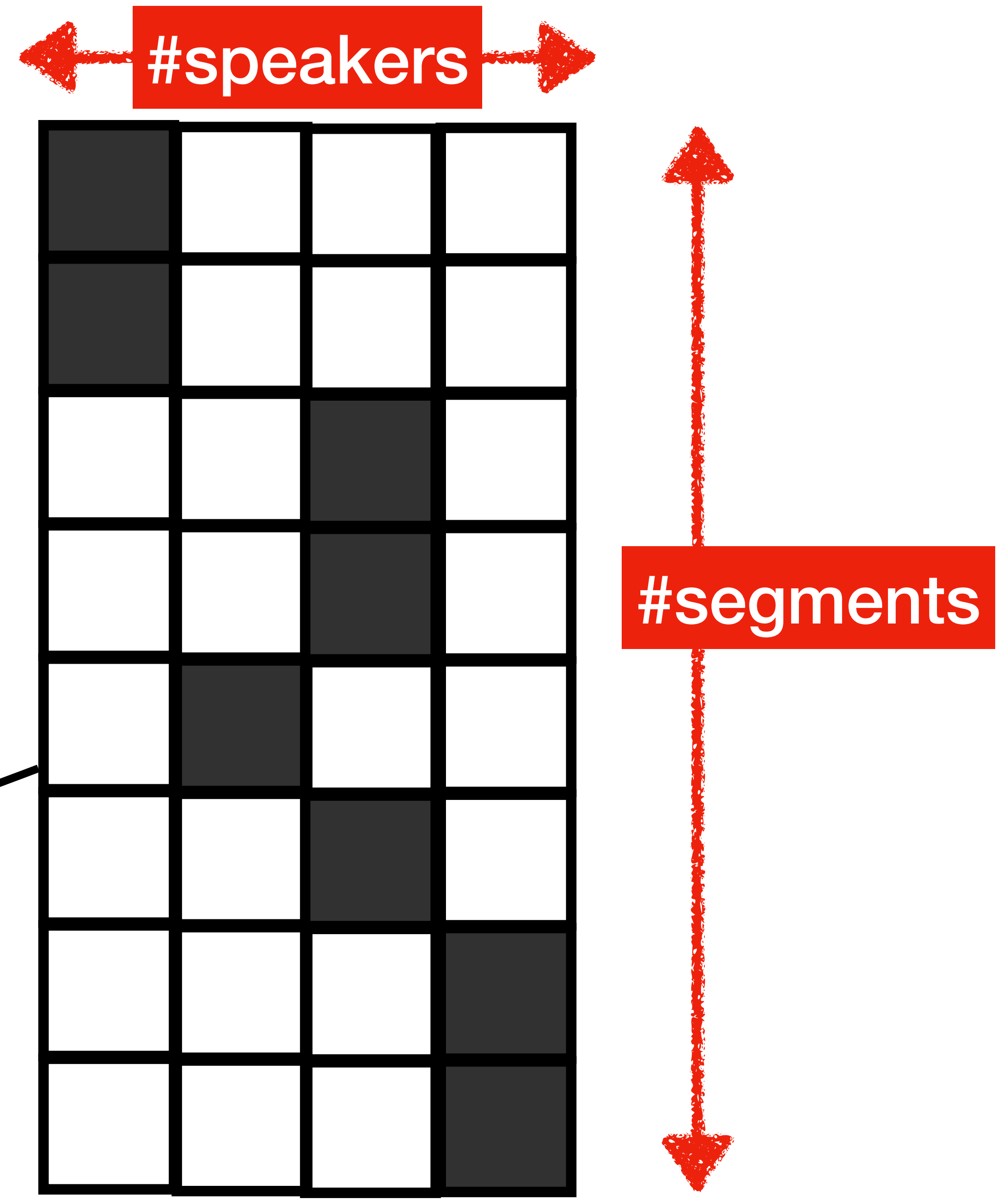


$$\text{maximize} \quad \epsilon(X) = \frac{1}{K}\sum_{k=1}^{K} \frac{X_k^T \mathbf{A} X_k}{X_k^T \mathbf{D} X_k}$$

$$\text{subject to} \quad X \in \{0,1\}^{N \times K},$$

$$X\mathbf{1}_K = \mathbf{1}_N \,.$$

Final cluster assignment matrix



New formulation for spectral clustering

This problem is NP-hard!

$$
\begin{aligned}
\text{maximize} \quad & \epsilon(X) = \frac{1}{K} \sum_{k=1}^{K} \frac{X_k^T \mathbf{A} X_k}{X_k^T \mathbf{D} X_k} \\
\text{subject to} \quad & \cancel{X \in \{0,1\}^{N \times K},} \\
& \cancel{X\mathbf{1}_K = \mathbf{1}_N}.
\end{aligned}
$$

Remove the discrete constraints to make the problem solvable

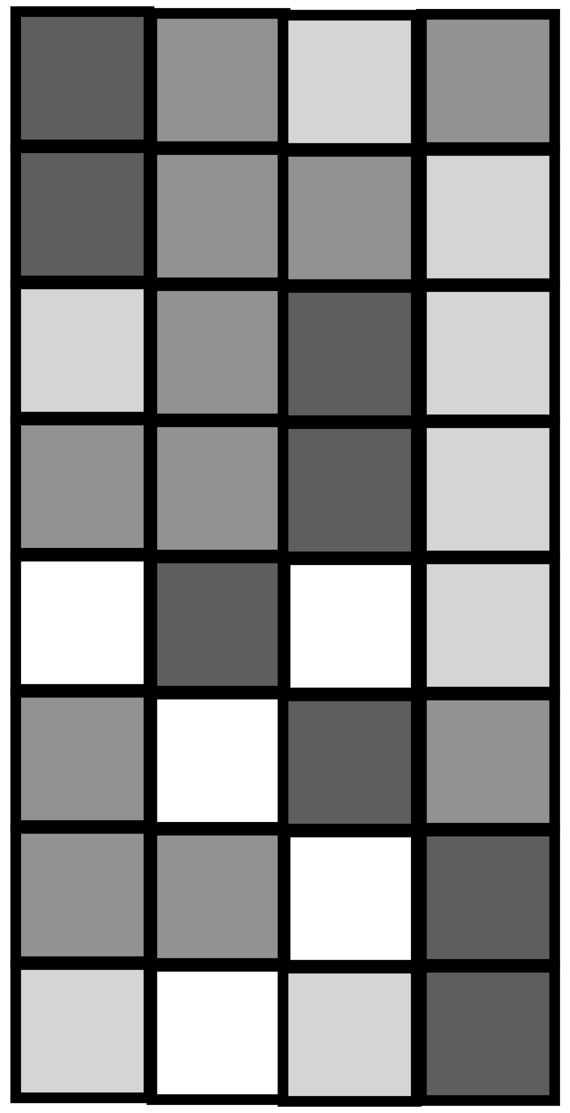
New formulation for spectral clustering

Relaxed problem has a set of solutions

$$\text{maximize} \quad \epsilon(X) = \frac{1}{K}\sum_{k=1}^{K}\frac{X_k^T \mathbf{A} X_k}{X_k^T \mathbf{D} X_k}$$

$$\text{subject to} \quad \cancel{X \in \{0,1\}^{N\times K}, \quad X\mathbf{1}_K = \mathbf{1}_N.}$$

Taking the Eigen-decomposition of $D^{-1}A$

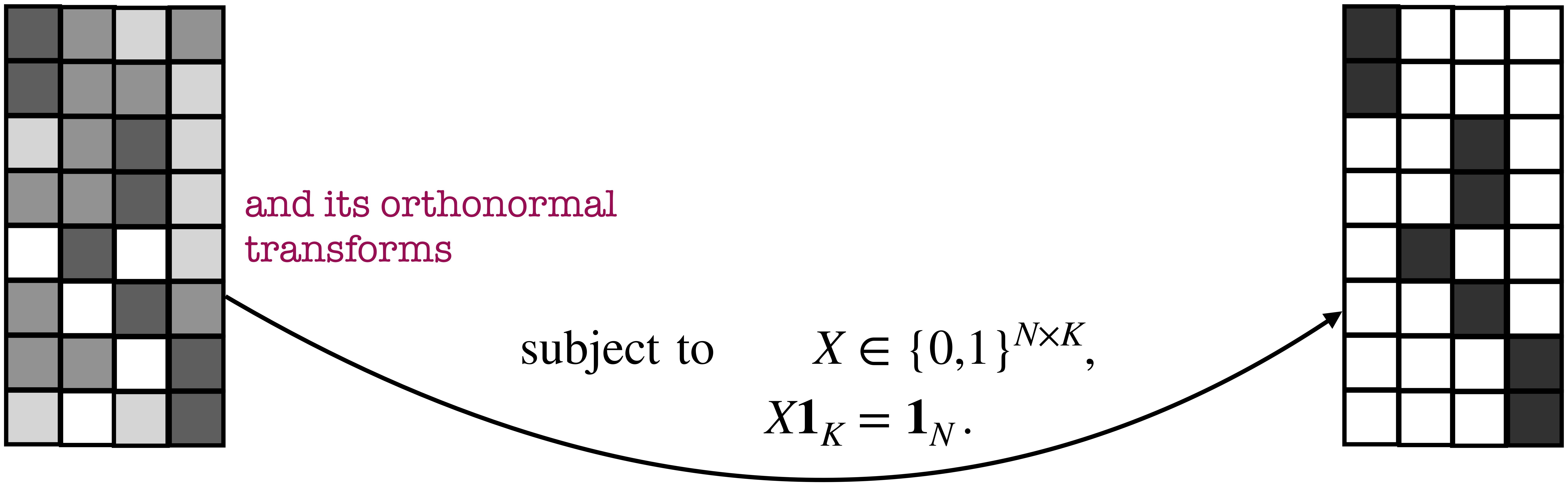


and its orthonormal transforms

Set of solutions to the **relaxed** problem

New formulation for spectral clustering

Now we need to discretize this solution!



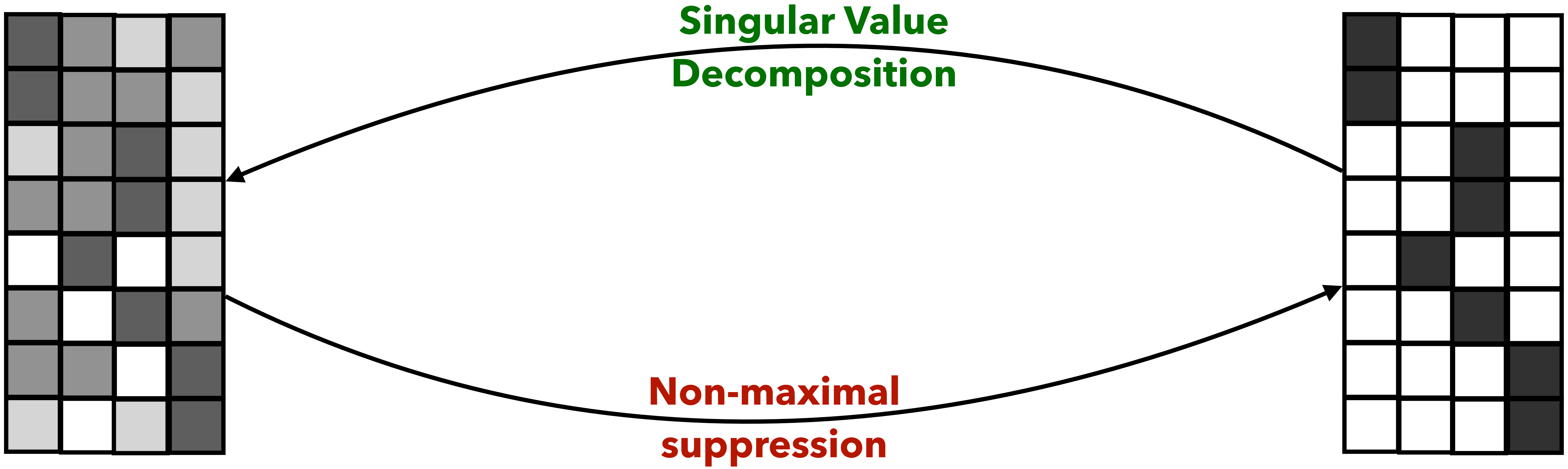
subject to $$X \in \{0,1\}^{N\times K}, \quad X\mathbf{1}_K = \mathbf{1}_N .$$

Find a matrix which is **discrete** and also close to any one of the **orthonormal transformations** of the relaxed solution

CENTER FOR LANGUAGE AND SPEECH PROCESSING

New formulation for spectral clustering

Now we need to discretize this solution!

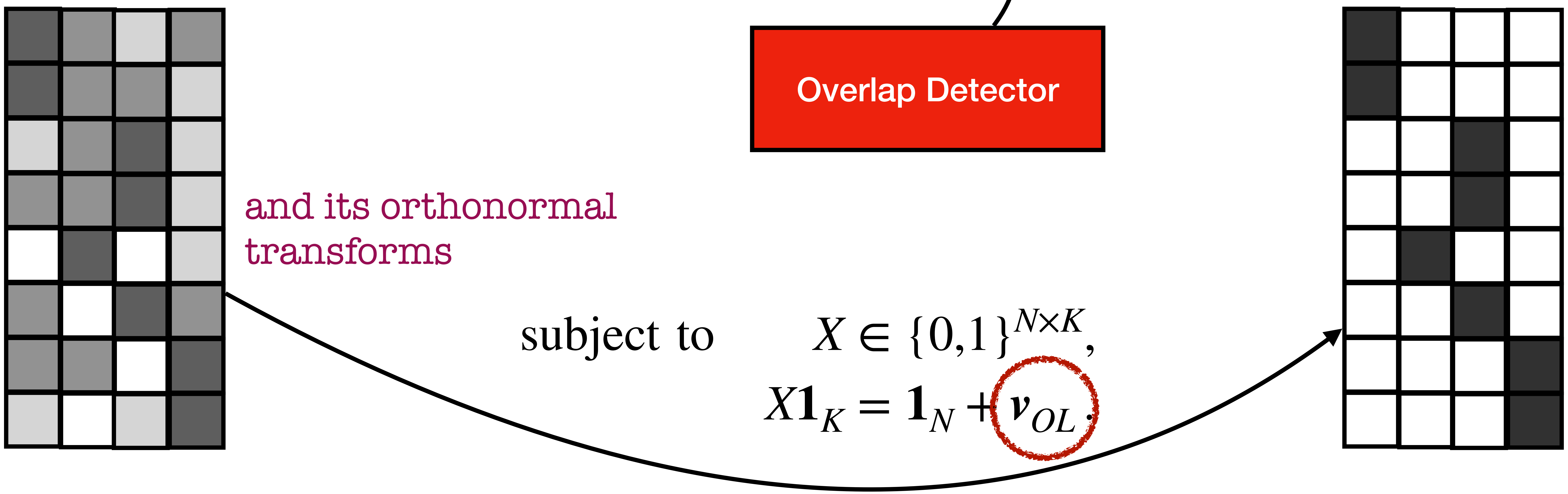


and its orthonormal transforms

Iterate until convergence

Let us now make it overlap-aware

Suppose we have $\mathbf{v}_{OL}$



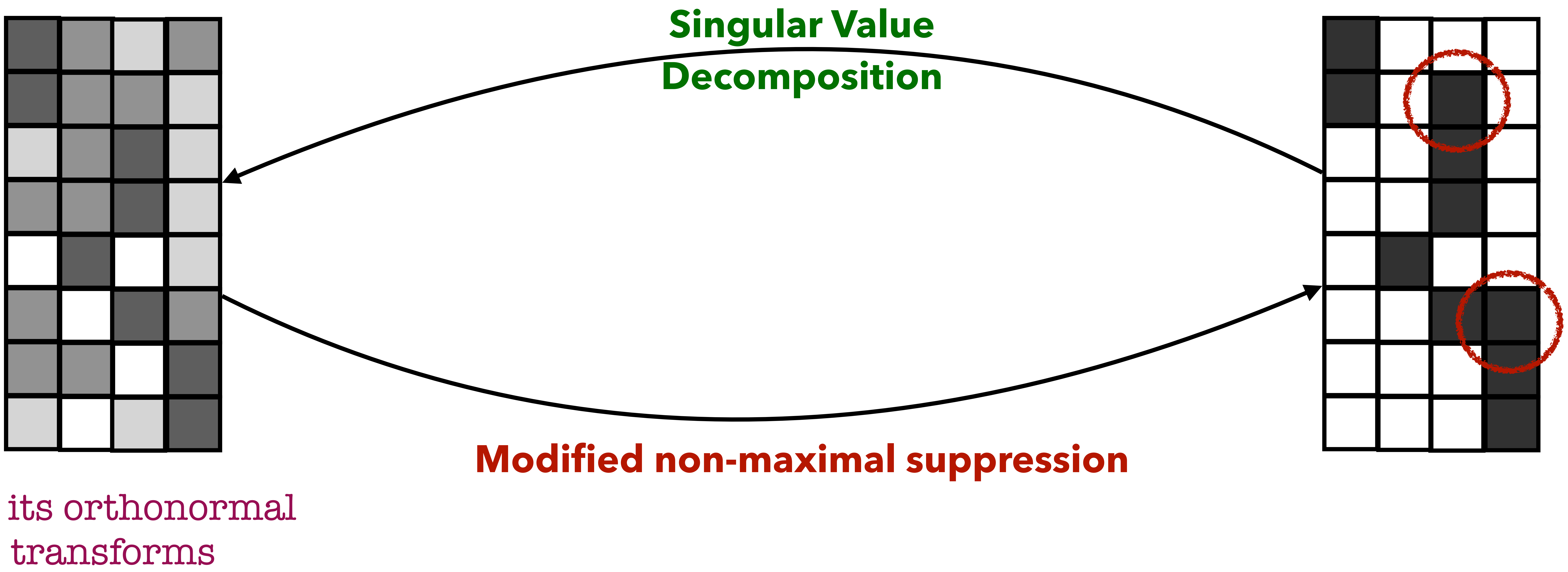
subject to $\quad X \in \{0,1\}^{N \times K},$

$$X\mathbf{1}_K = \mathbf{1}_N + \boldsymbol{v}_{OL}.$$

Discrete constraint is modified to include overlap detector output

Let us now make it overlap-aware

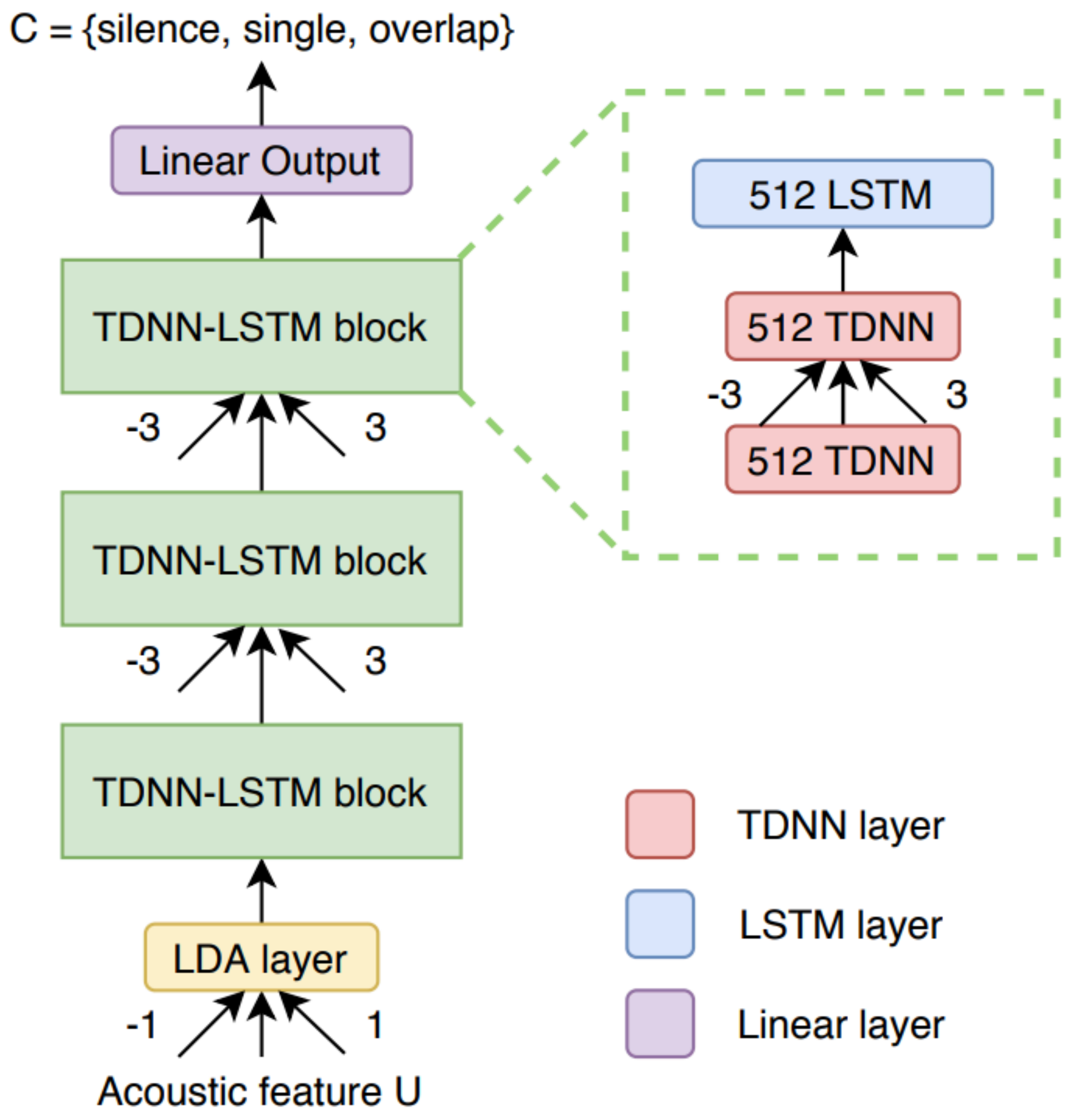
Modify non-maximal suppression to pick top 2 speakers

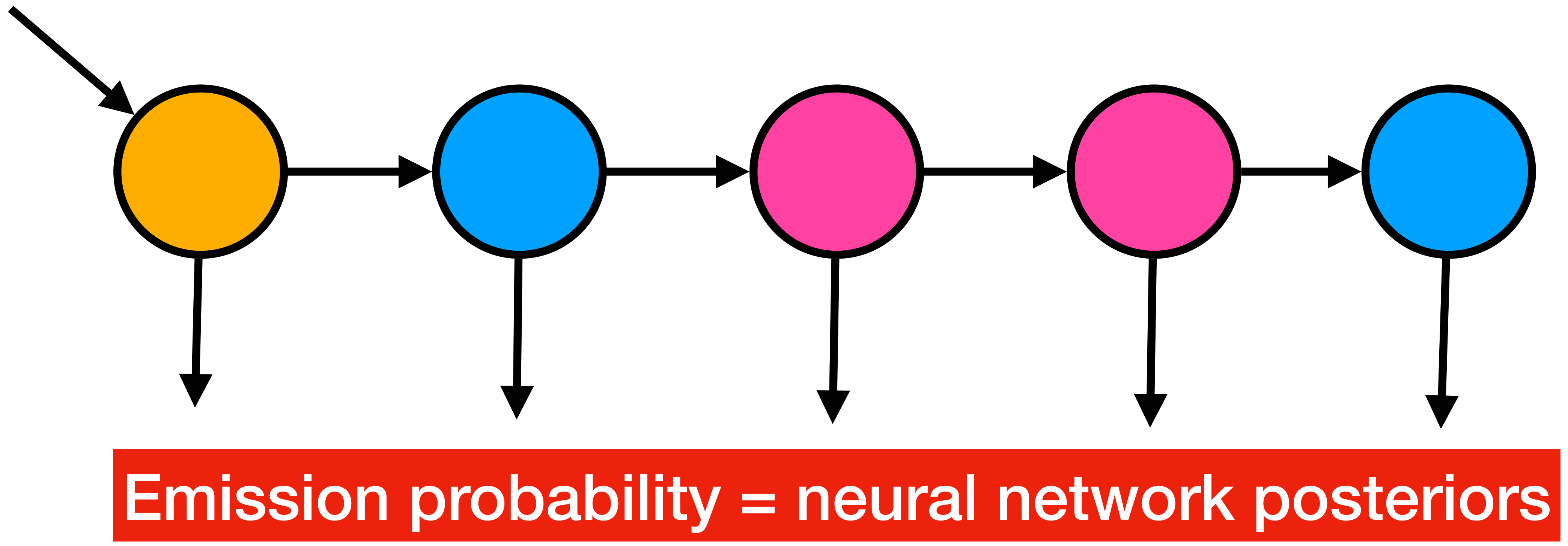


Iterate until convergence

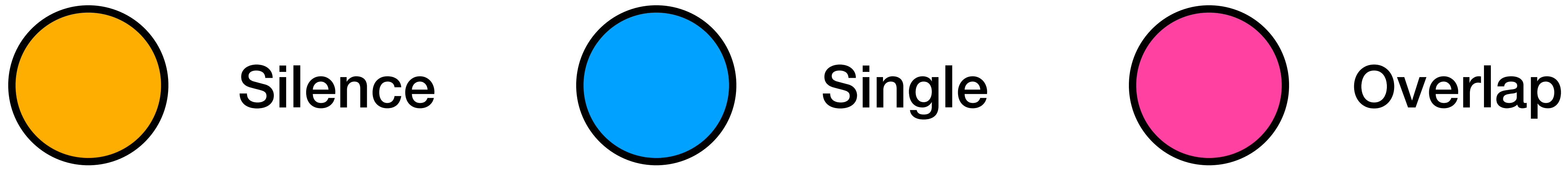
Hybrid HMM-DNN overlap detector

(Can also use other methods, e.g. end-to-end)





Viterbi decoding used for inference



Results on AMI Mix-Headset eval

12.0% relative improvement over spectral clustering baseline

System	DER
Spectral clustering	26.9
AHC	28.3
VBx	26.2
Overlap-aware SC	**24.0**

Park et al., “Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap,” IEEE Signal Processing Letters, 2020.

Garcia-Romero et al., “Speaker diarization using deep neural network embeddings,” ICASSP 2017.

Dîez et al., “Speaker diarization based on Bayesian HMM with eigenvoice priors,” Odyssey 2018.

AMI data contains **4-speaker meetings**

Results on AMI Mix-Headset eval

Comparable with other overlap-aware diarization methods

System	DER
VB-based overlap assignment	23.8
Region proposal networks	25.5
Overlap-aware SC	**24.0**

Bullock, et al., “Overlap-aware diarization: resegmentation using neural end-toend overlapped speech detection,” ICASSP 2020.

Huang et al., “Speaker diarization with region proposal network,” ICASSP 2020.

Does not require **matching training data** or **initialization** with other diarization systems.

Results: DER breakdown on AMI eval

System	Missed speech	False alarm	Speaker conf.	DER
AHC/PLDA	19.9	0.0	8.4	26.9
Spectral/cosine	19.9	0.0	7.0	28.3
VBx	19.9	0.0	6.3	26.2
VB-based overlap assignment	13.0	3.6	7.2	23.8
RPN	9.5	7.7	8.3	25.5
Overlap-aware SC	11.3	2.2	10.5	24.0

Results: DER breakdown on AMI eval

Missed speech decreases significantly

System	Missed speech	False alarm	Speaker conf.	DER
AHC/PLDA	19.9	0.0	8.4	26.9
Spectral/cosine	19.9	0.0	7.0	28.3
VBx	19.9	0.0	6.3	26.2
VB-based overlap assignment	13.0	3.6	7.2	23.8
RPN	9.5	7.7	8.3	25.5
Overlap-aware SC	11.3	2.2	10.5	24.0

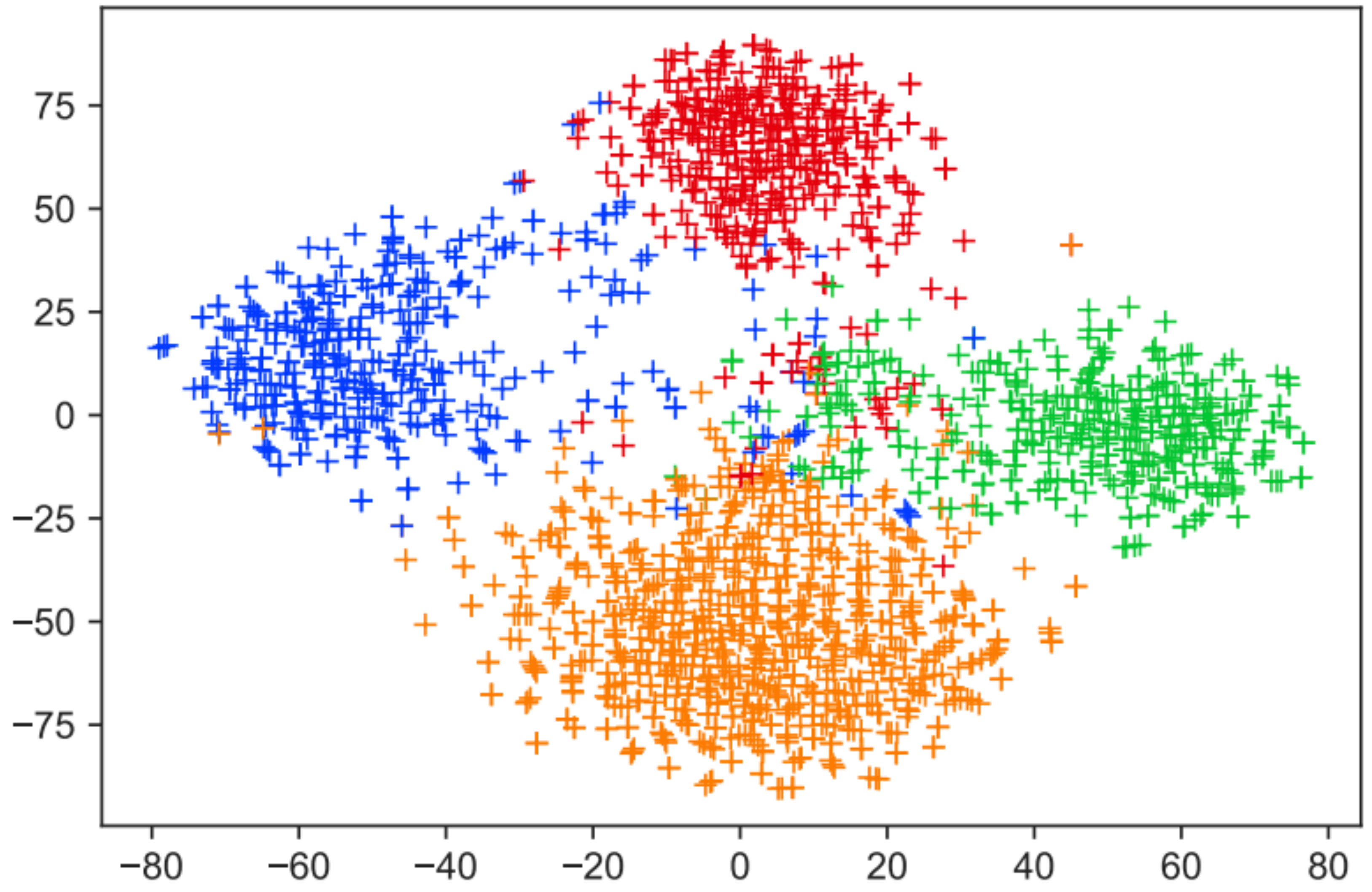
Results: DER breakdown on AMI eval

Speaker confusion increases

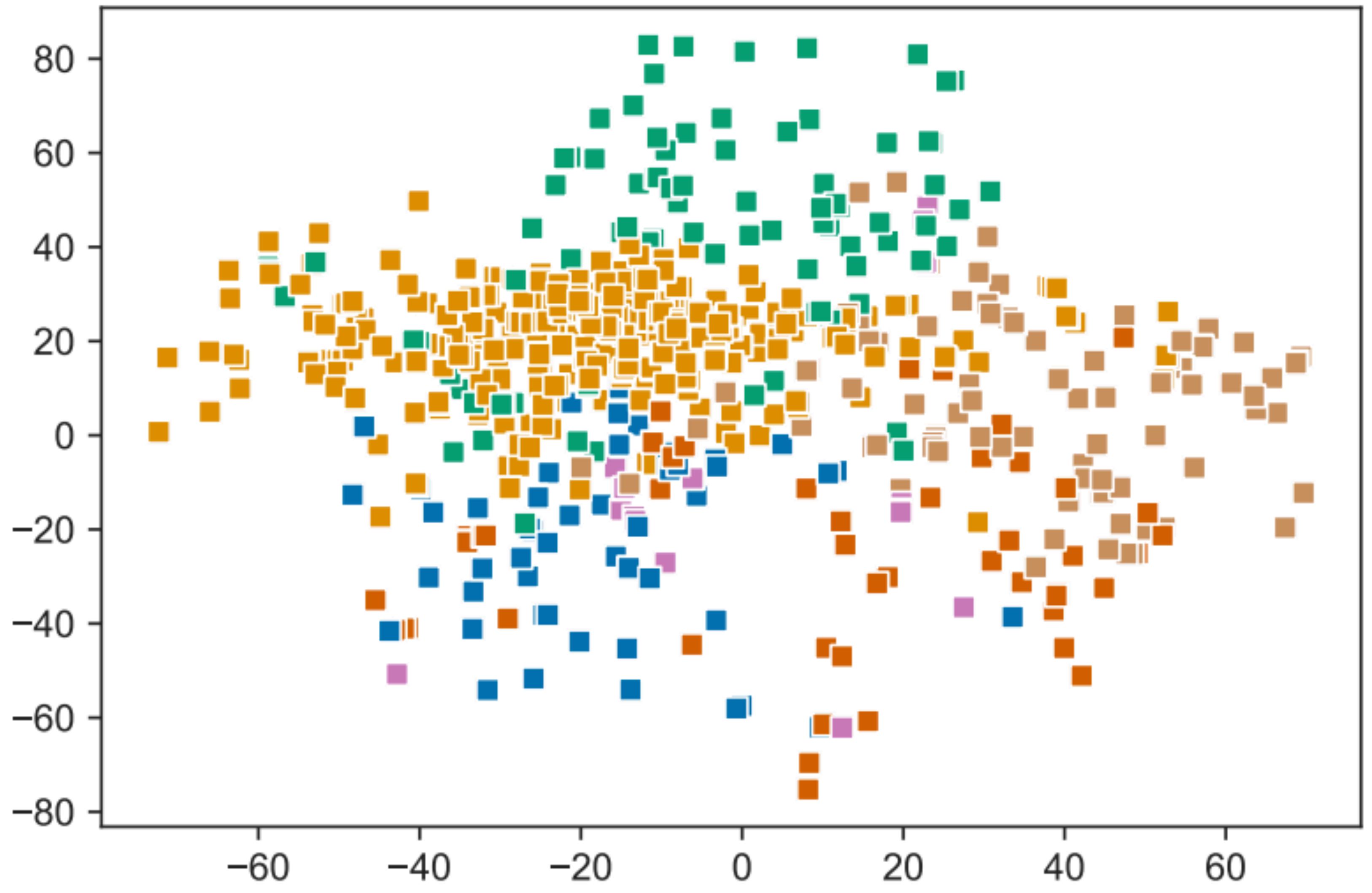
System	Missed speech	False alarm	Speaker conf.	DER
AHC/PLDA	19.9	0.0	8.4	26.9
Spectral/cosine	19.9	0.0	7.0	28.3
VBx	19.9	0.0	6.3	26.2
VB-based overlap assignment	13.0	3.6	7.2	23.8
RPN	9.5	7.7	8.3	25.5
Overlap-aware SC	11.3	2.2	10.5	24.0

CENTER FOR LANGUAGE AND SPEECH PROCESSING

Need more robust x-vector extractors



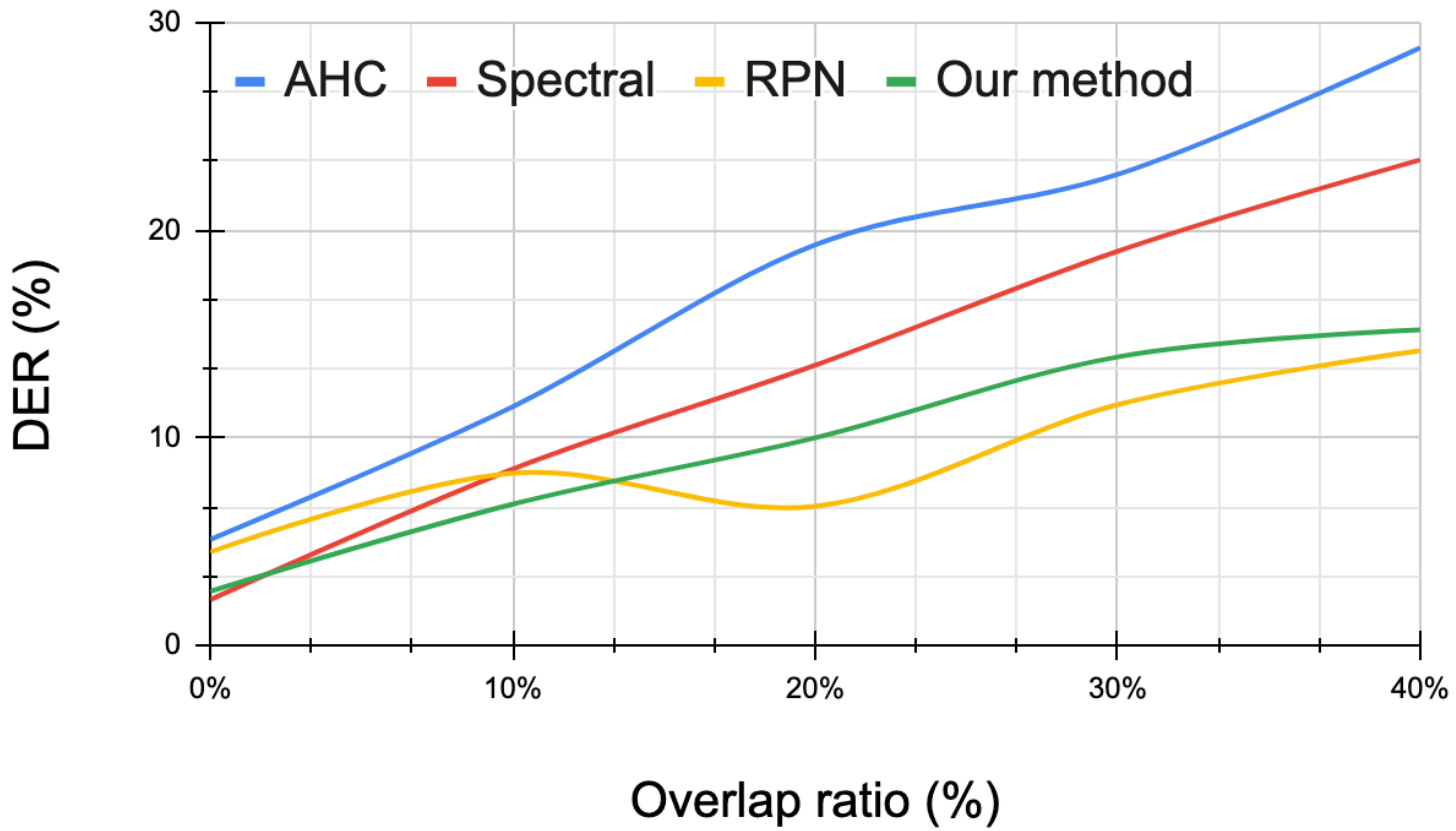
Non-overlapping segments



Overlapping segments

T-SNE plot of x-vector embeddings

More results: DER on LibriCSS



LibriCSS data contains **8-speaker meetings**

Overlap-aware Diarization

Several new methods proposed recently

Fujita et al. "**End-to-end neural diarization**: Reformulating speaker diarization as simple multi-label classification," ArXiv, 2020.

Bullock, et al., "Overlap-aware diarization: **resegmentation** using neural end-to-end overlapped speech detection," ICASSP 2020.

Huang et al., "Speaker diarization with **region proposal network**," ICASSP 2020.

Medennikov, et al. "**Target speaker voice activity detection**: a novel approach for multispeaker diarization in a dinner party scenario," Interspeech 2020.

Kinoshita, et al. **Integrating** end-to-end neural and clustering-based diarization: Getting the best of both worlds. *ArXiv, 2020*.

Machine learning tasks benefit from an ensemble of systems.

For example, ROVER is a popular combination method for ASR systems.

Jonathan G. Fiscus, “A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER),” IEEE ASRU 1997.

Problem

Why is it hard to combine diarization systems?

- System outputs may have different number of speaker estimates.
- System outputs are usually in different label space.
- There may not be agreement on whether a region contains overlap.

Solution

DOVER-Lap performs "map and vote"

- System outputs may have different number of speaker estimates.
- System outputs are usually in different label space.

Label mapping: Maximal matching algorithm based on a global cost tensor

- There may not be agreement on whether a region contains overlap.

Raj, D., García-Perera, L.P., Huang, Z., Watanabe, S., Povey, D., Stolcke, A., & Khudanpur, S. DOVER-Lap: A Method for Combining Overlap-aware Diarization Outputs. *IEEE SLT 2021*.

Solution

DOVER-Lap performs “map and vote”

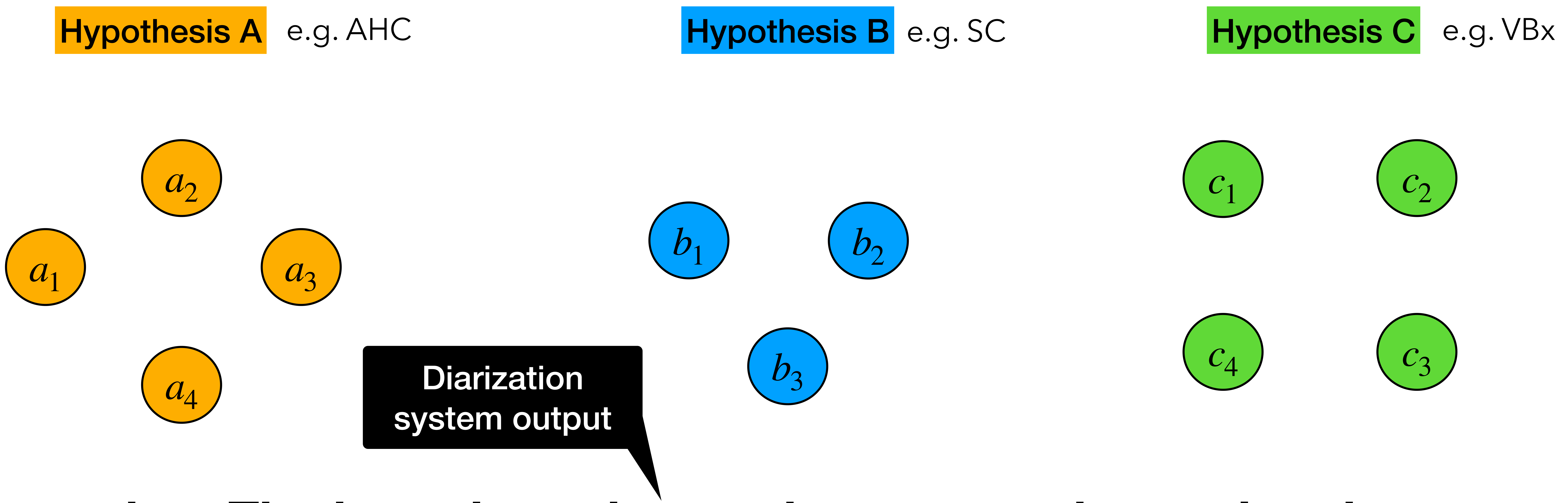
- System outputs may have different number of speaker estimates.
- System outputs are usually in different label space.
- There may not be agreement on whether a region contains overlap.

Label voting: Weighted majority voting considers speaker count in region

Raj, D., García-Perera, L.P., Huang, Z., Watanabe, S., Povey, D., Stolcke, A., & Khudanpur, S. DOVER-Lap: A Method for Combining Overlap-aware Diarization Outputs. *IEEE SLT 2021*.

DOVER-Lap extends DOVER

Diarization Output Voting Error Reduction

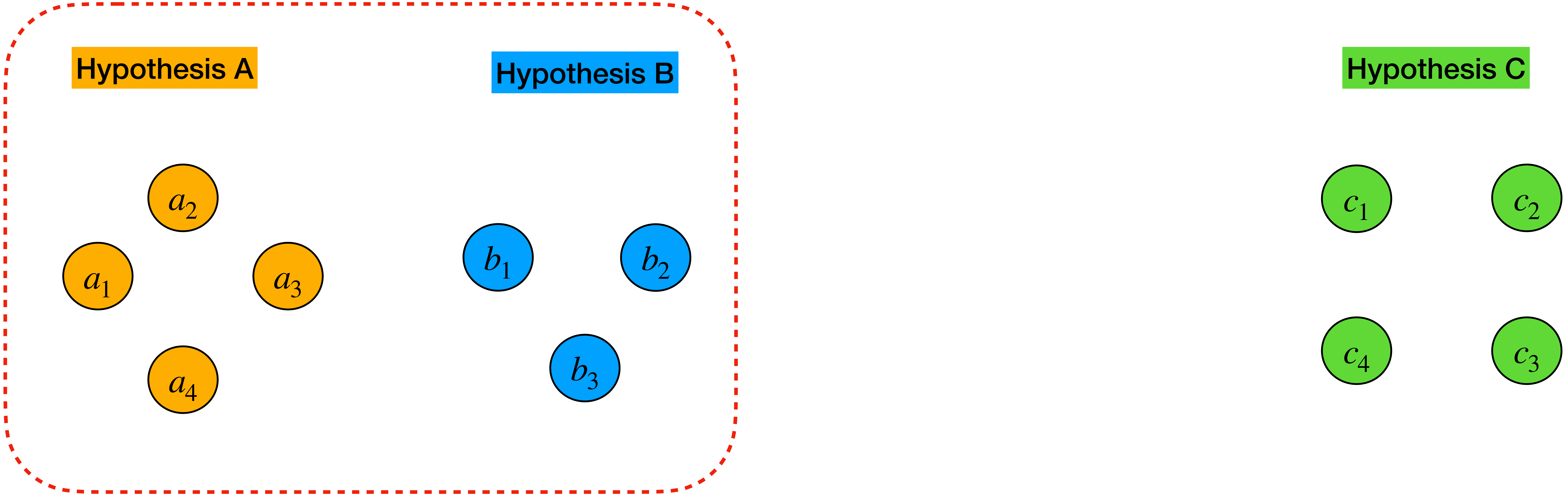


Assumption: The input hypotheses do not contain overlapping segments.

Andreas Stolcke and Takuya Yoshioka, "DOVER: A method for combining diarization outputs," IEEE ASRU 2019.

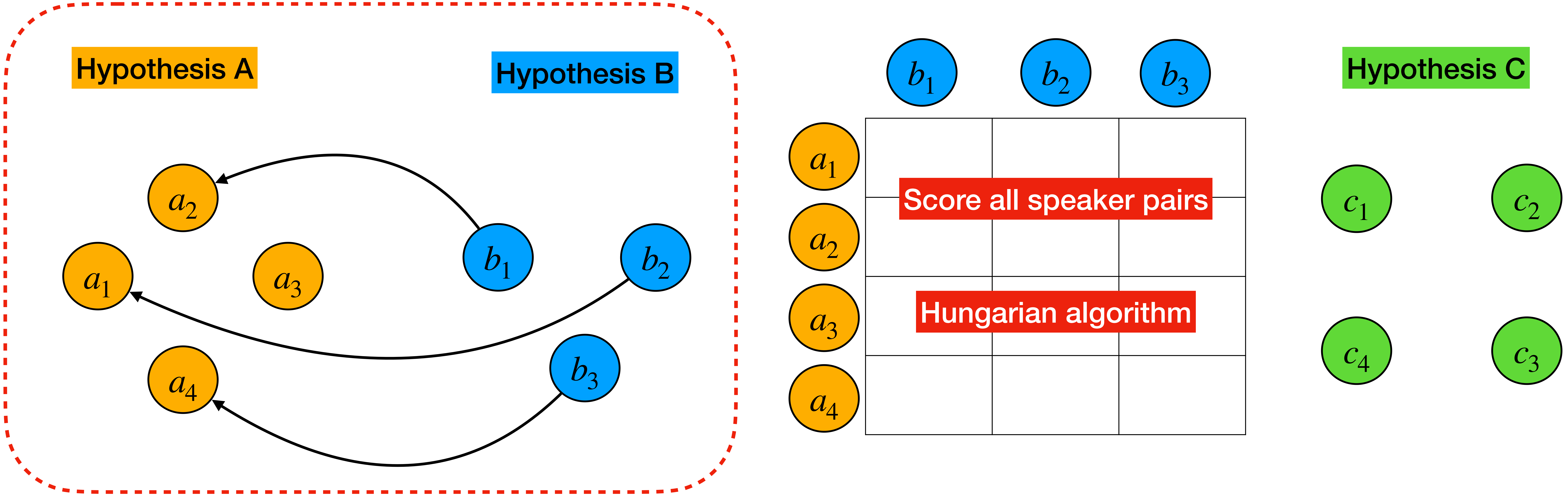
Preliminary: how DOVER works

Pair-wise incremental label mapping



Preliminary: how DOVER works

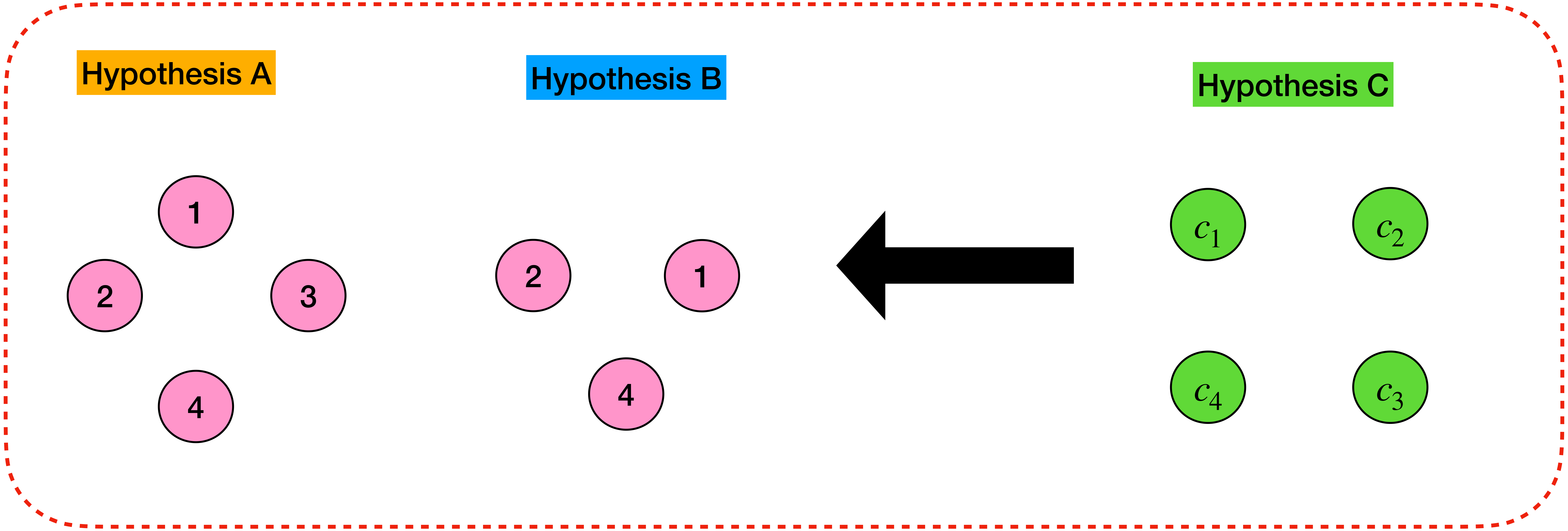
Pair-wise incremental label mapping



This is the same algorithm that is used to map hypothesis to reference for DER computation.

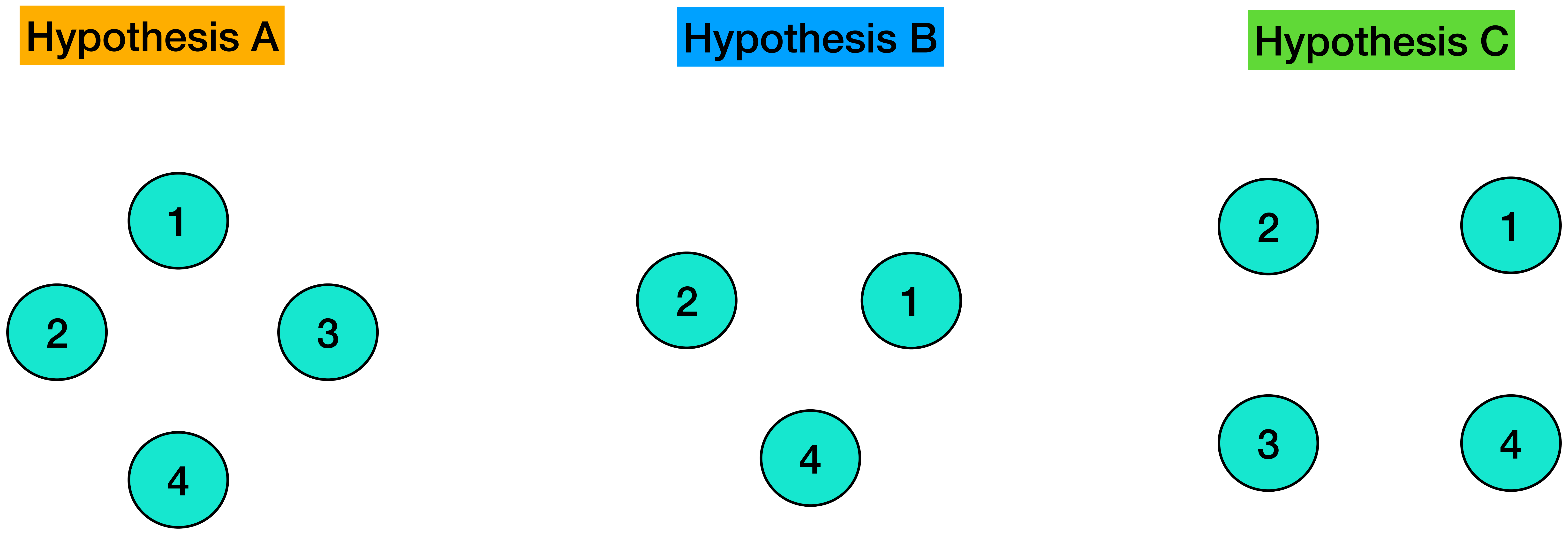
Preliminary: how DOVER works

Pair-wise incremental label mapping



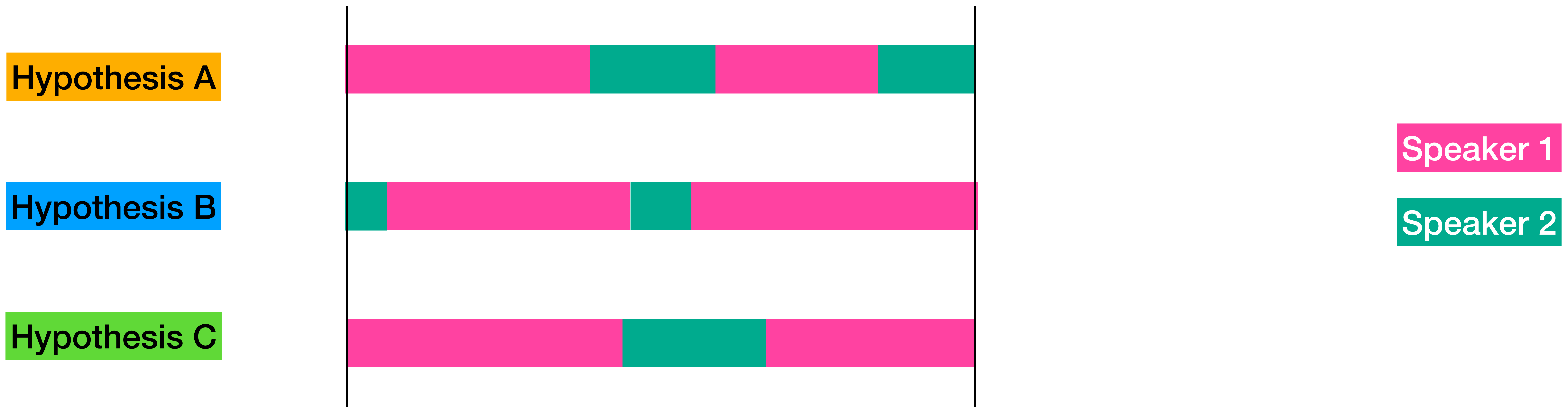
Preliminary: how DOVER works

Pair-wise incremental label mapping



Preliminary: how DOVER works

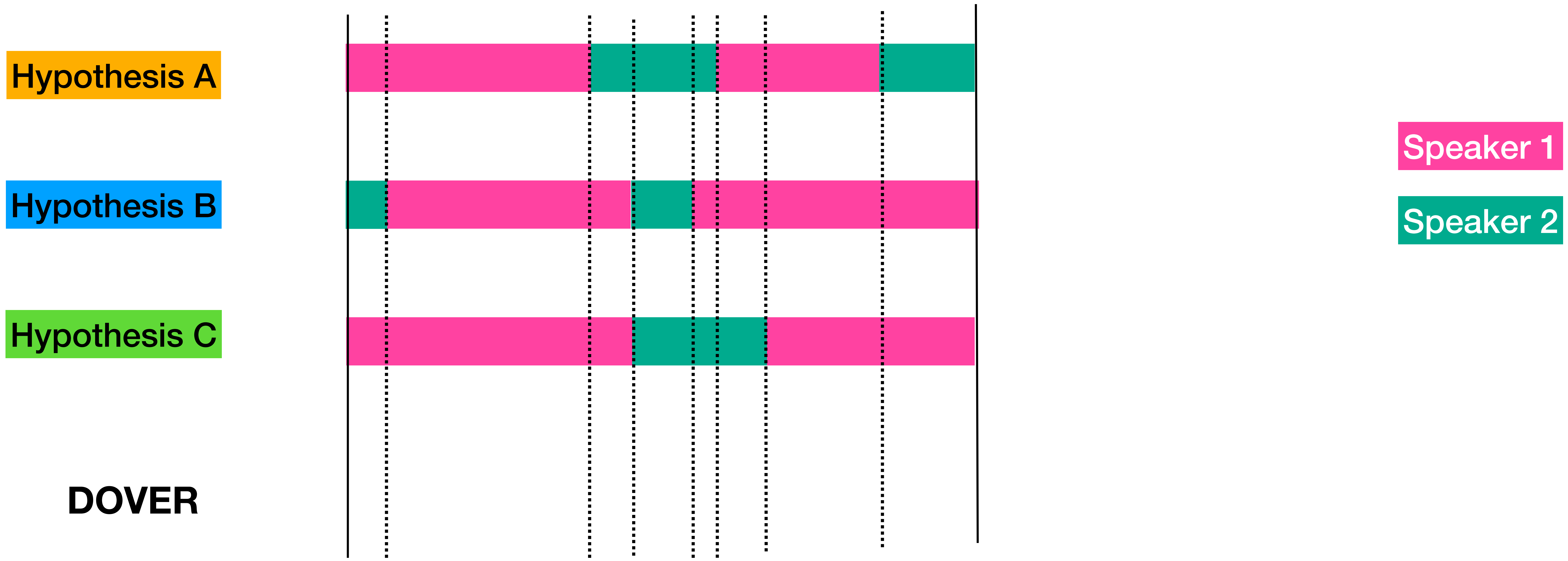
Label voting using rank-weighting



CENTER FOR LANGUAGE AND SPEECH PROCESSING

Preliminary: how DOVER works

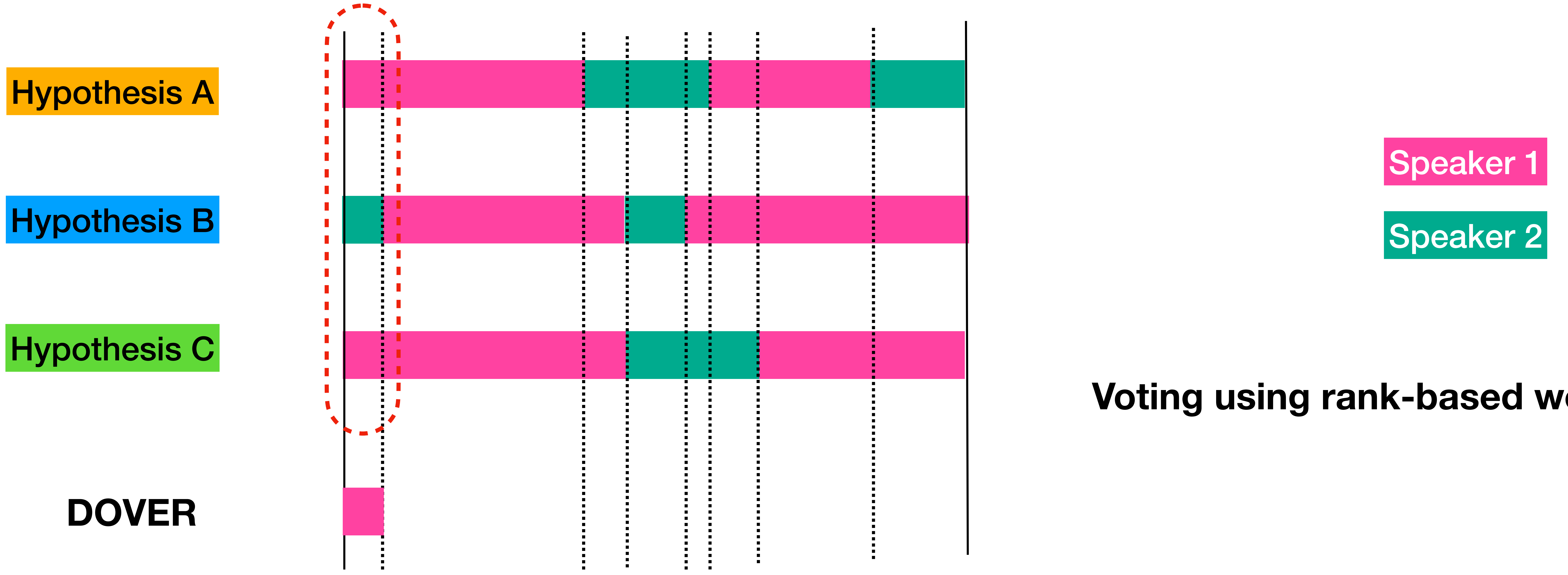
Label voting using rank-weighting



CENTER FOR LANGUAGE AND SPEECH PROCESSING

Preliminary: how DOVER works

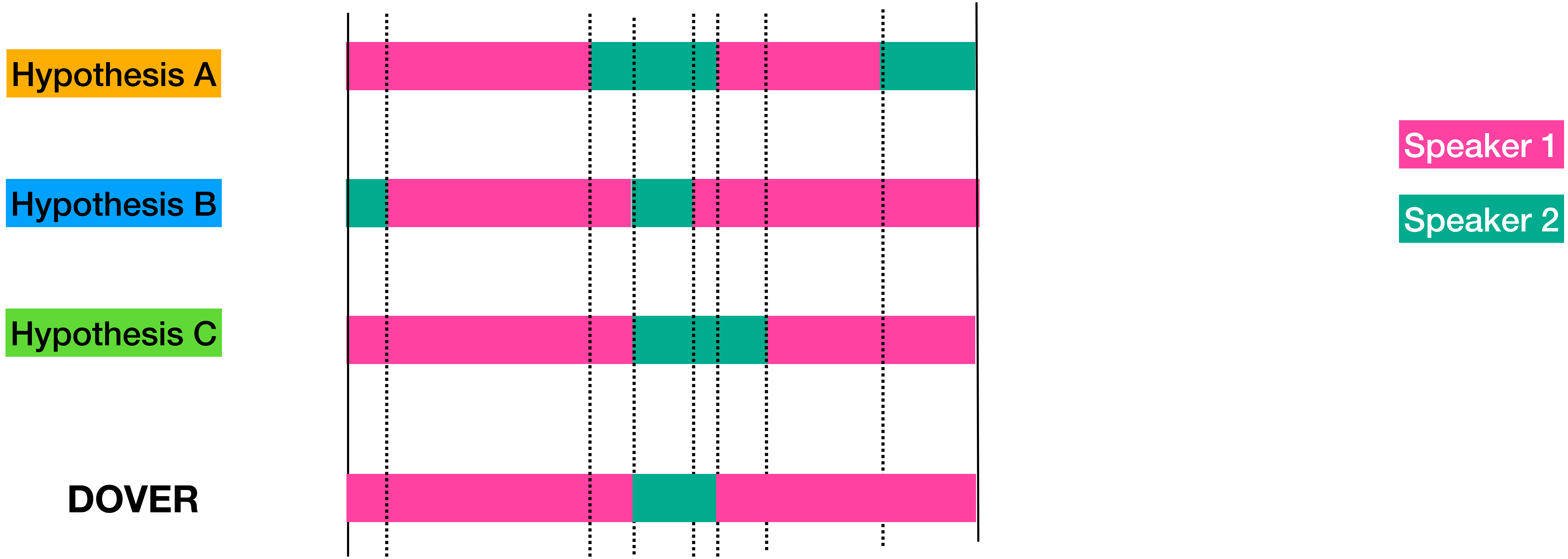
Label voting using rank-weighting



Voting using rank-based weights

Preliminary: how DOVER works

Label voting using rank-weighting

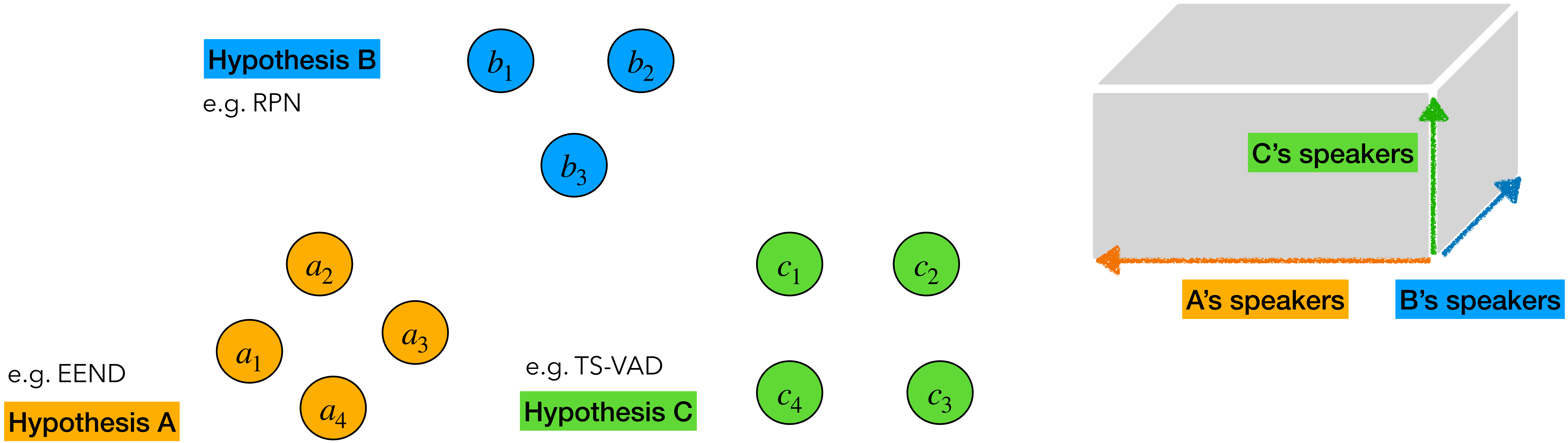


2 limitations of DOVER

1. Incremental pair-wise label assignment does not give optimal mapping
2. Voting method does not handle overlapping speaker segments

DOVER-Lap label mapping

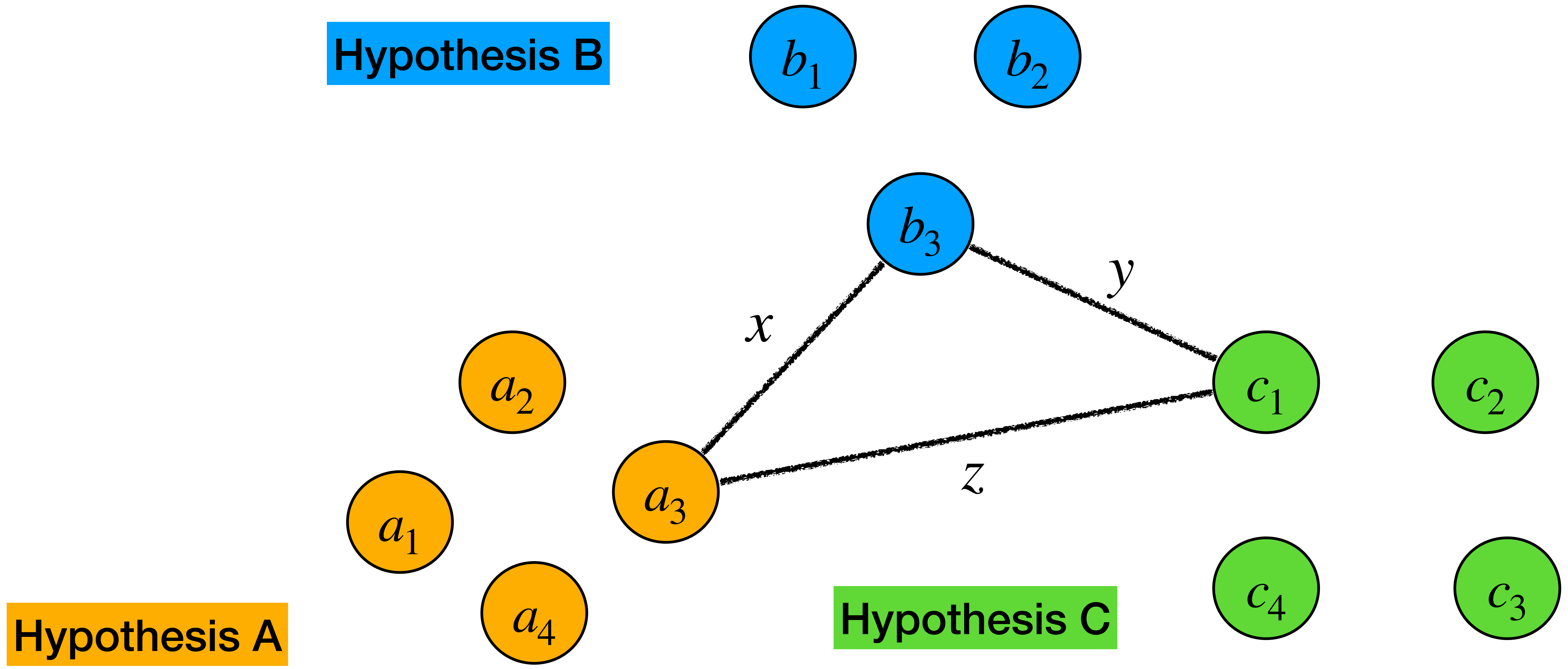
Change incremental method to global

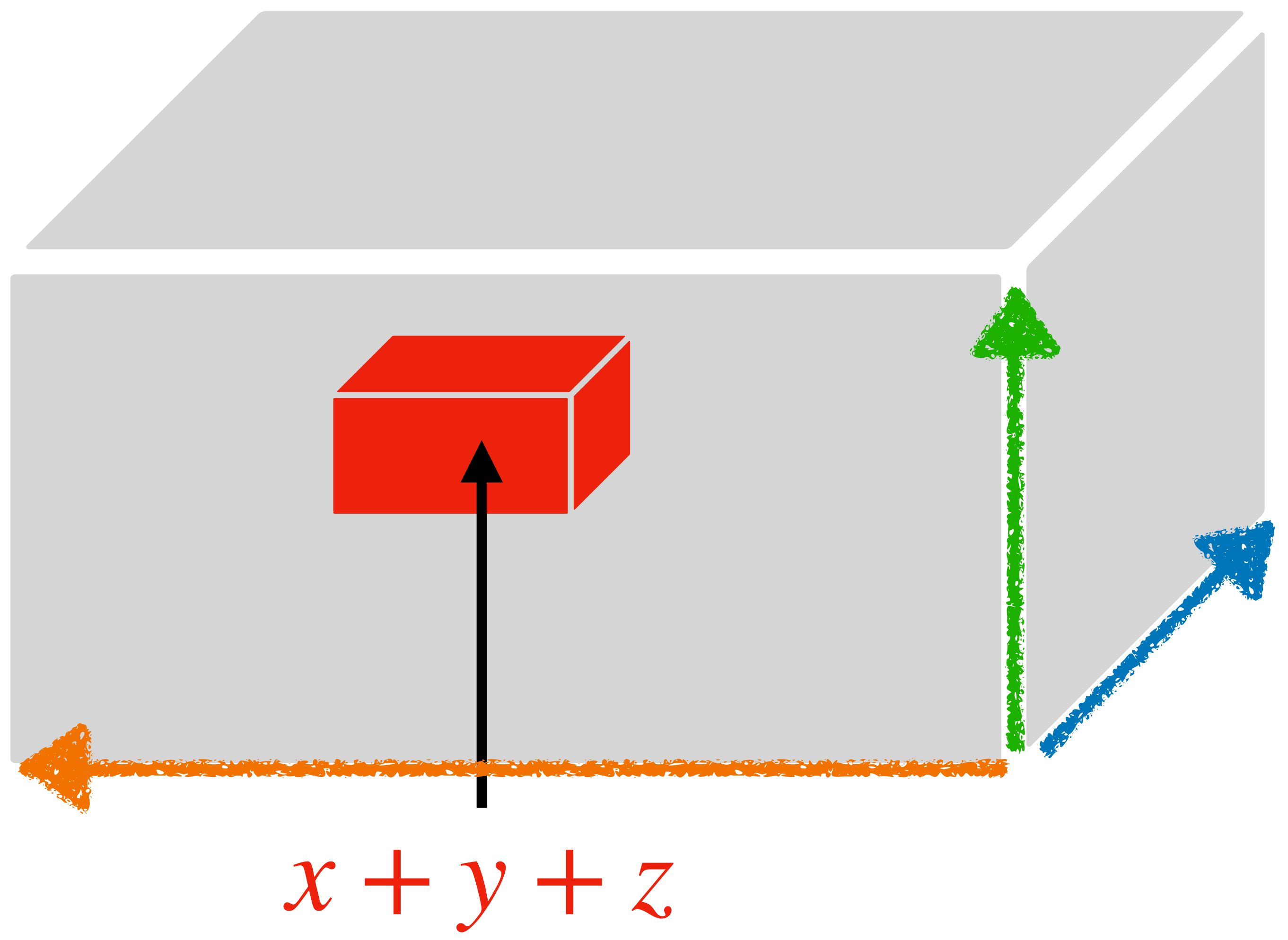


Hypotheses can contain overlapping segments.

DOVER-Lap label mapping

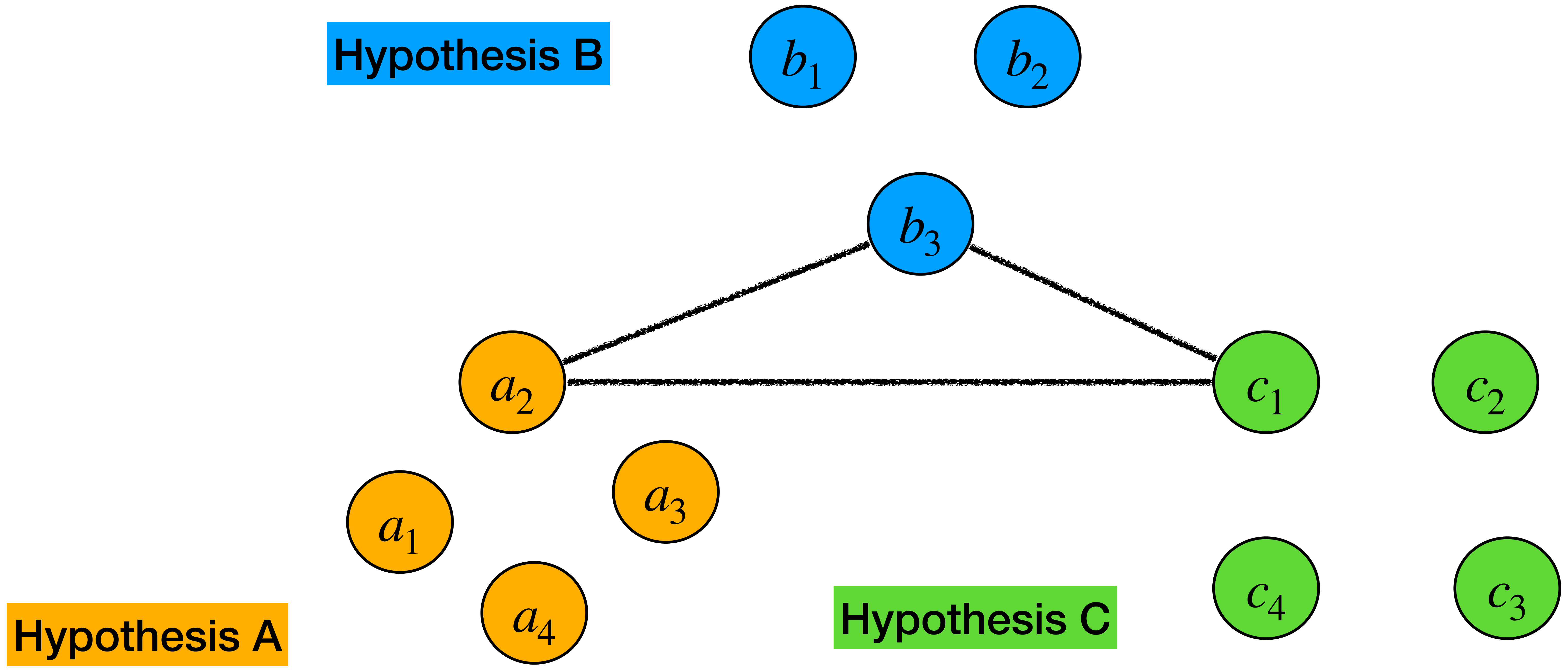
Compute "tuple costs" for all tuples

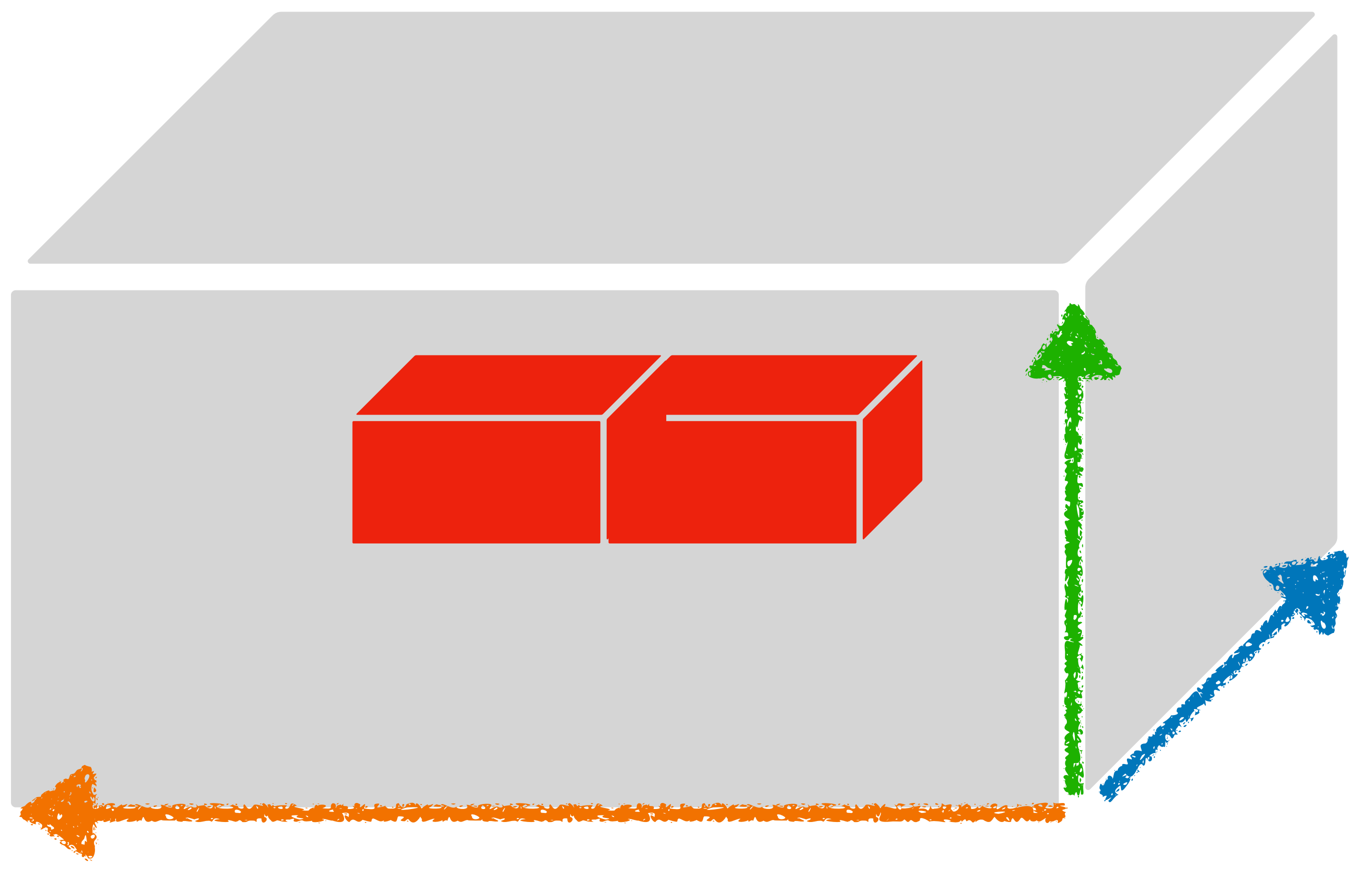




DOVER-Lap label mapping

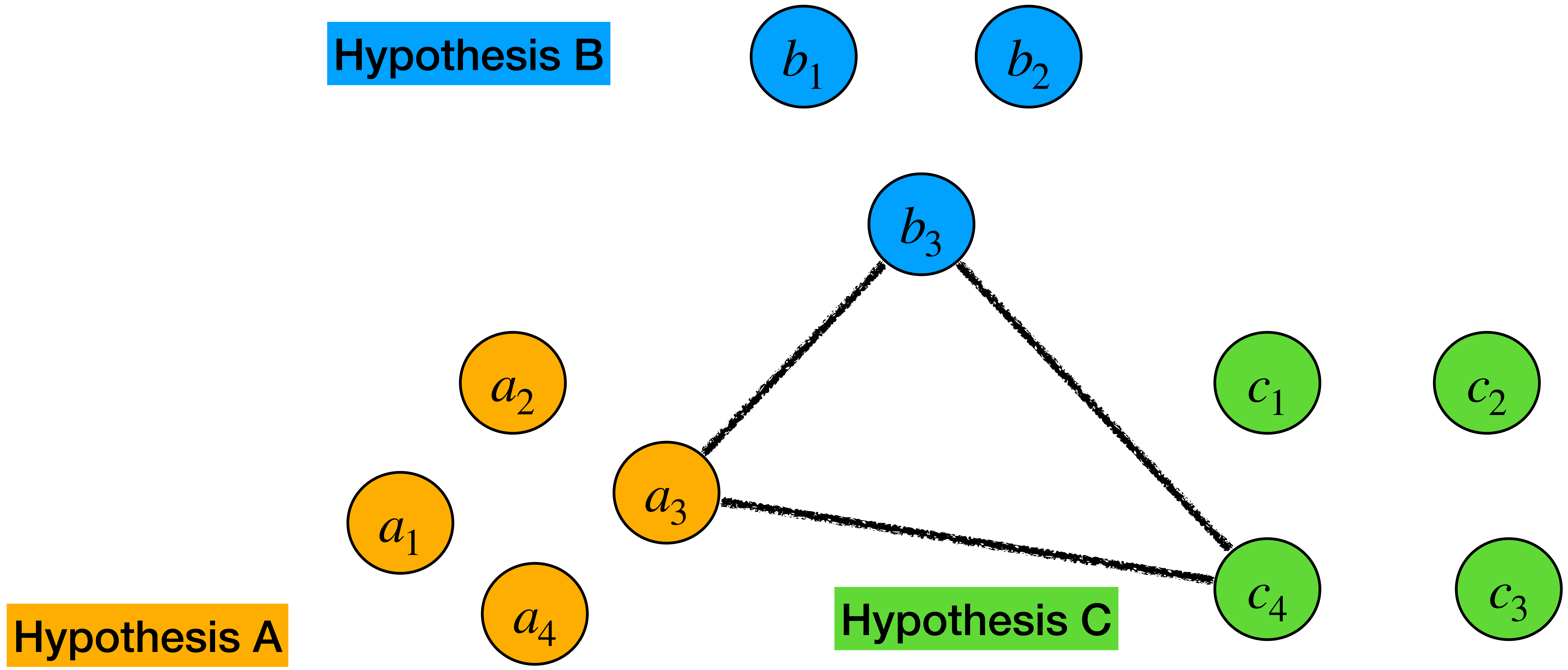
Compute “tuple costs” for all tuples

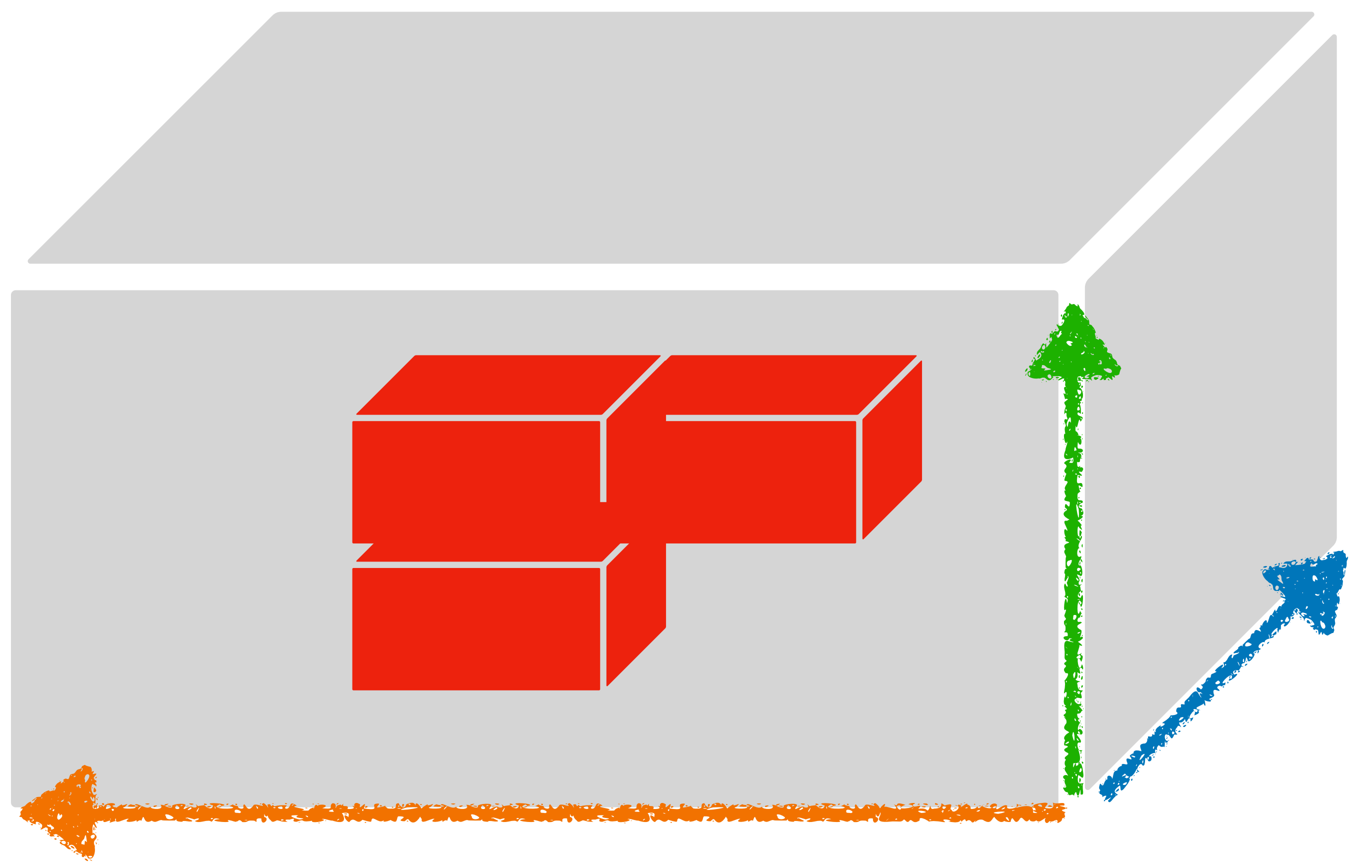




DOVER-Lap label mapping

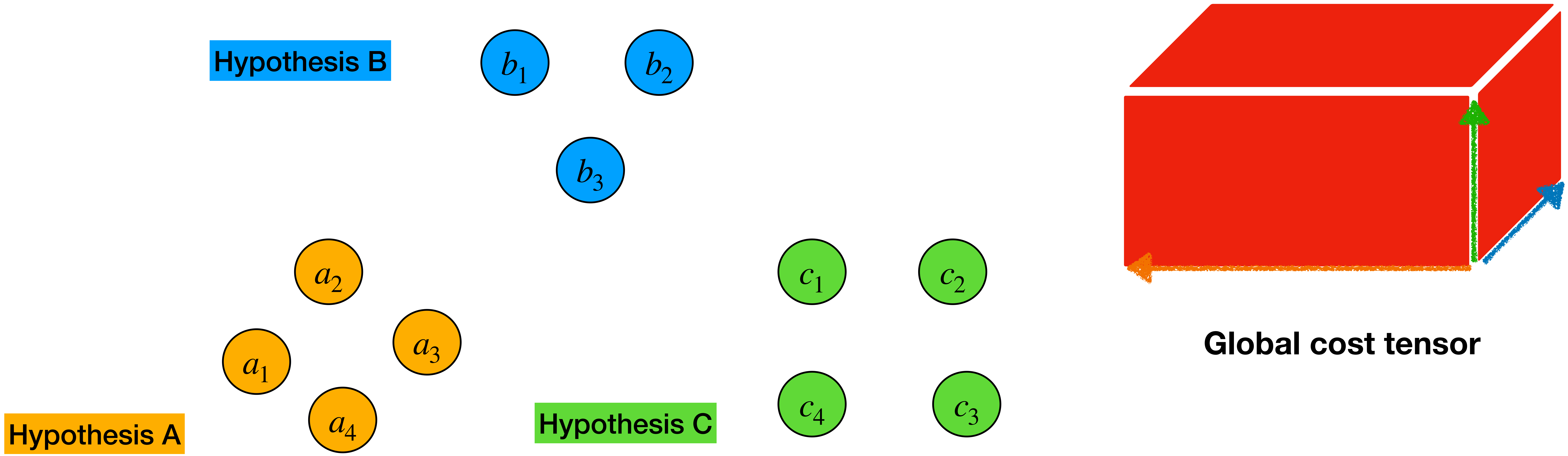
Compute “tuple costs” for all tuples





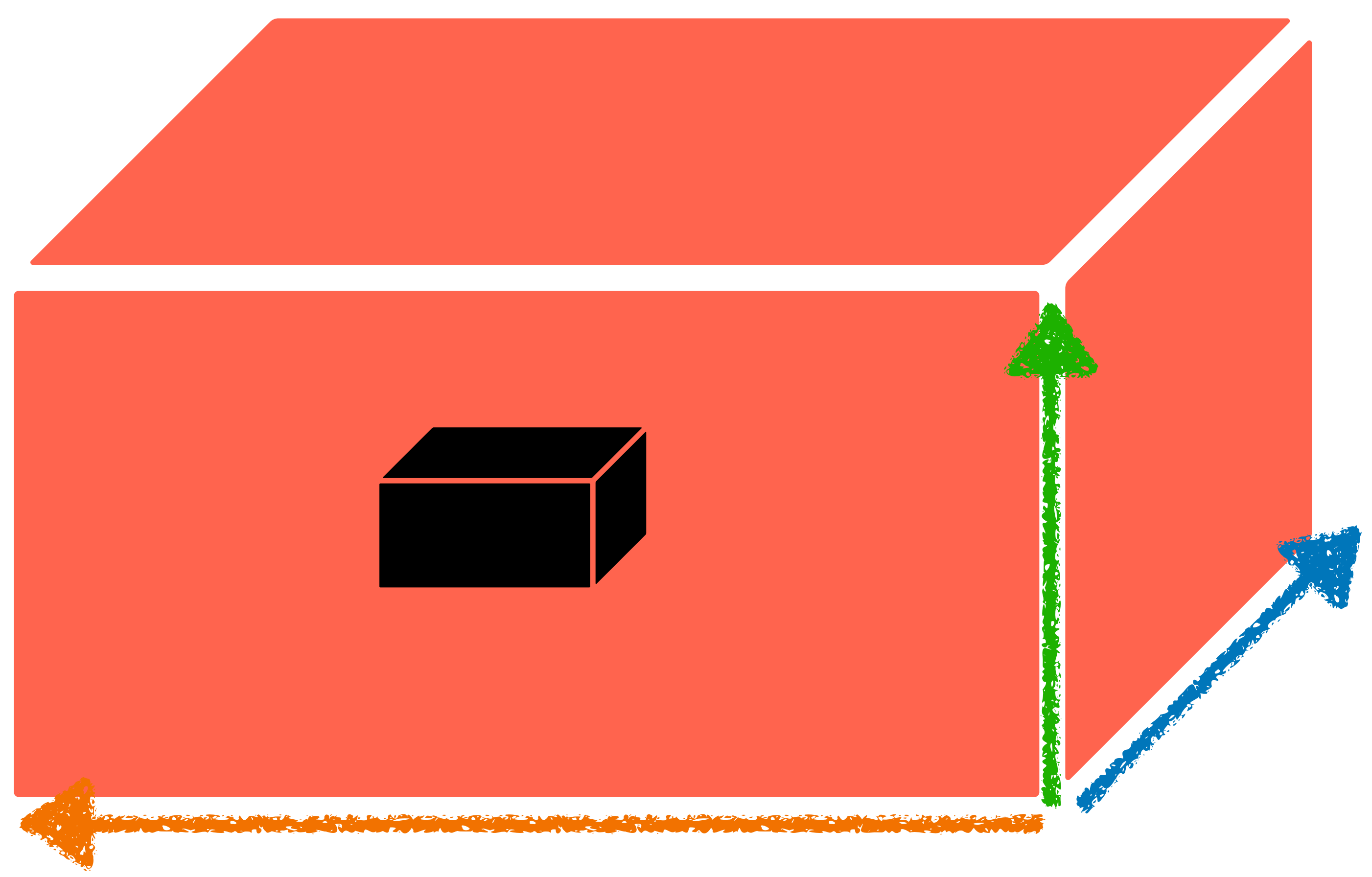
DOVER-Lap label mapping

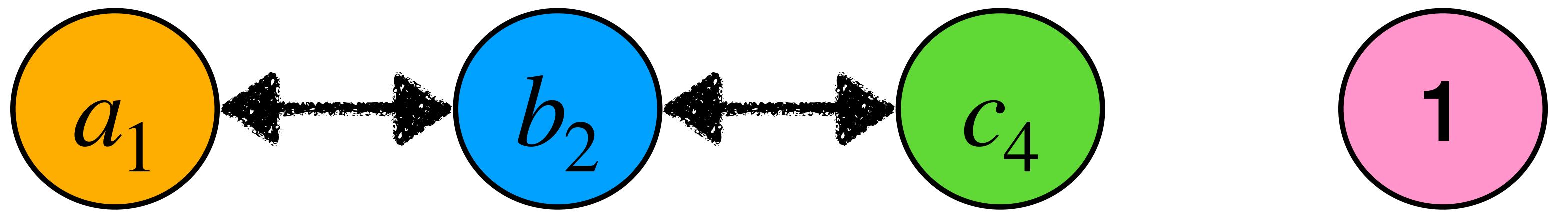
This gives us a "global" cost tensor



DOVER-Lap label mapping

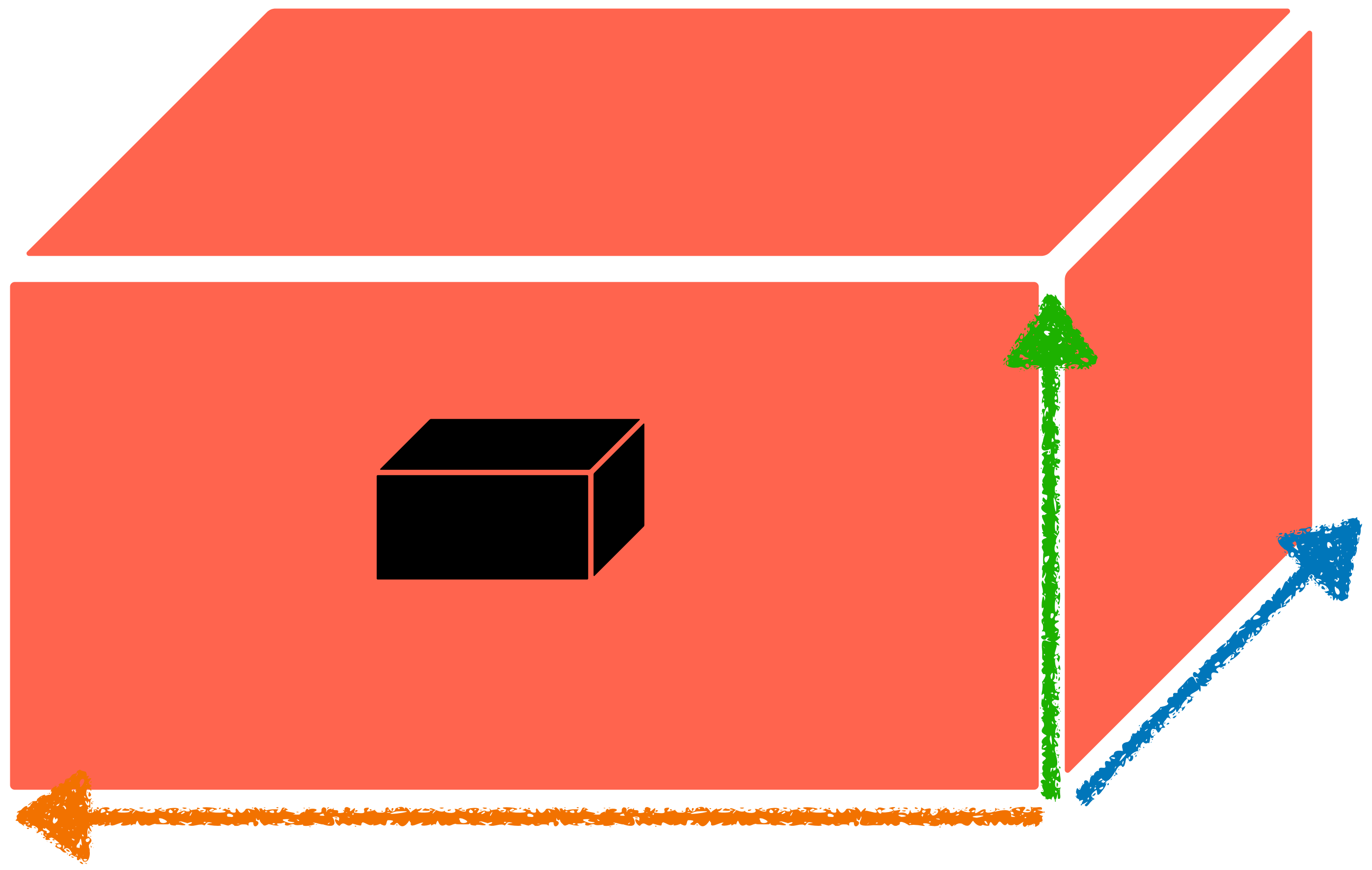
Pick tuple with the lowest cost and assign them same label

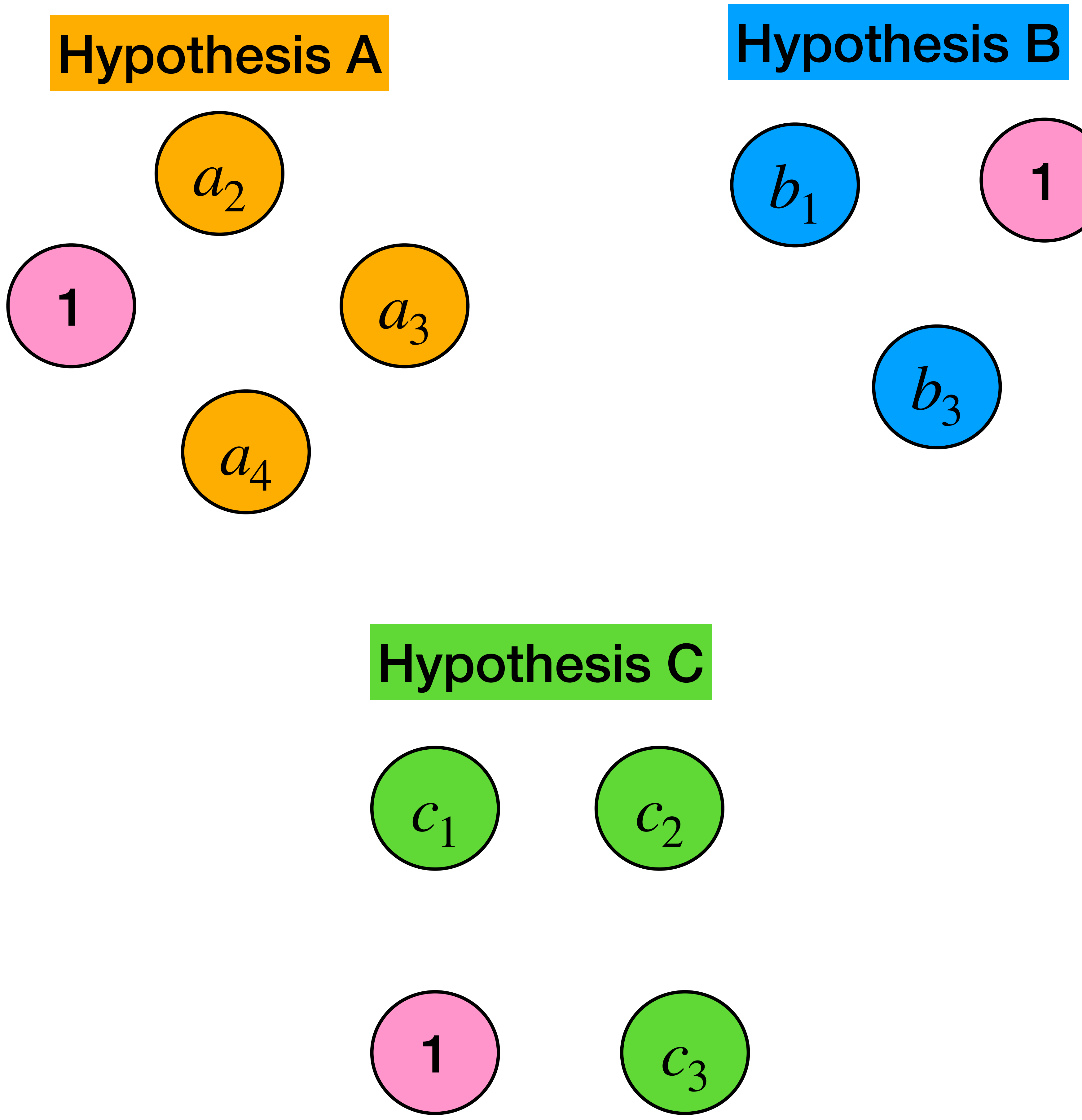




DOVER-Lap label mapping

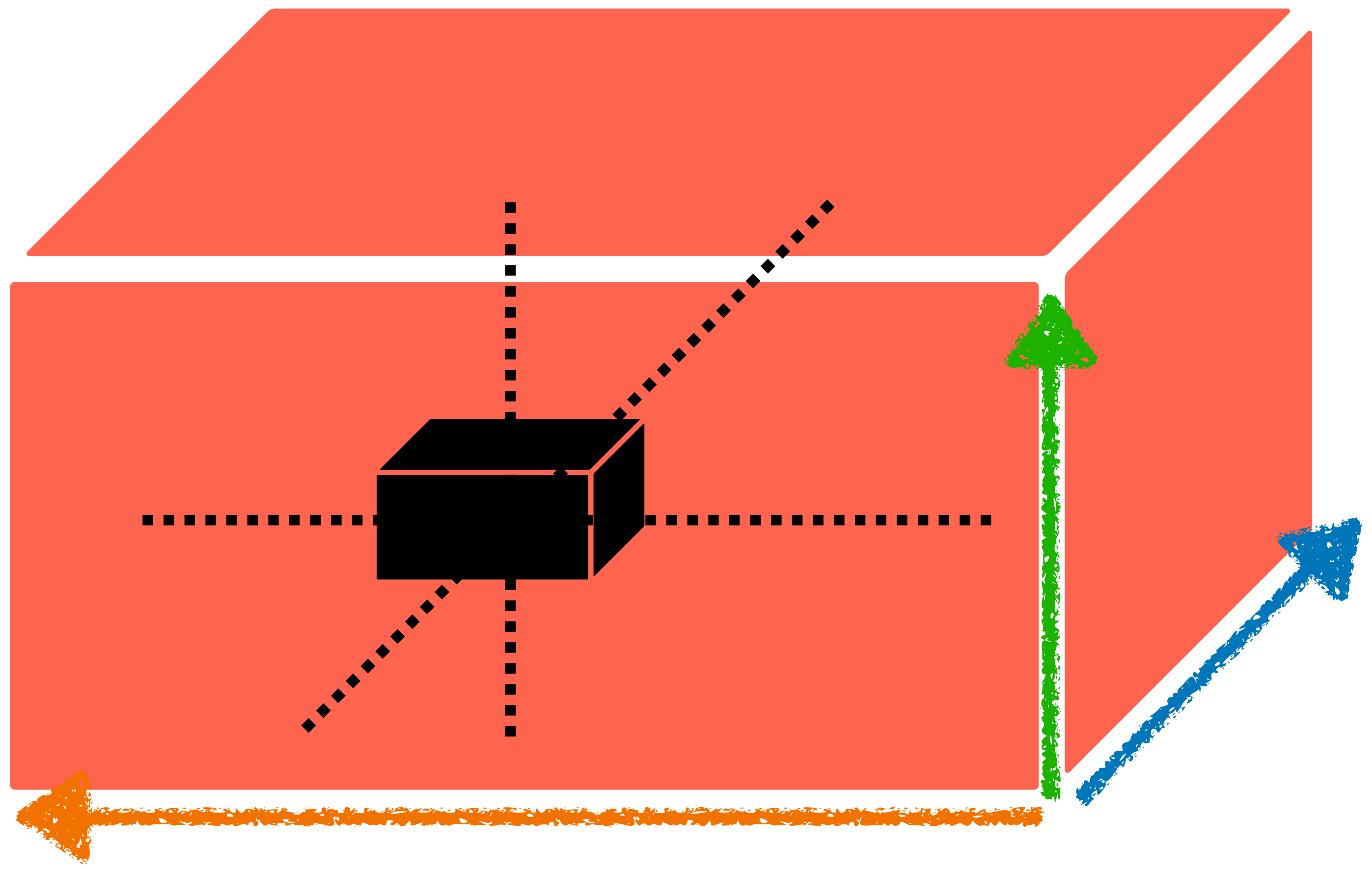
Pick tuple with the lowest cost and assign them same label





DOVER-Lap label mapping

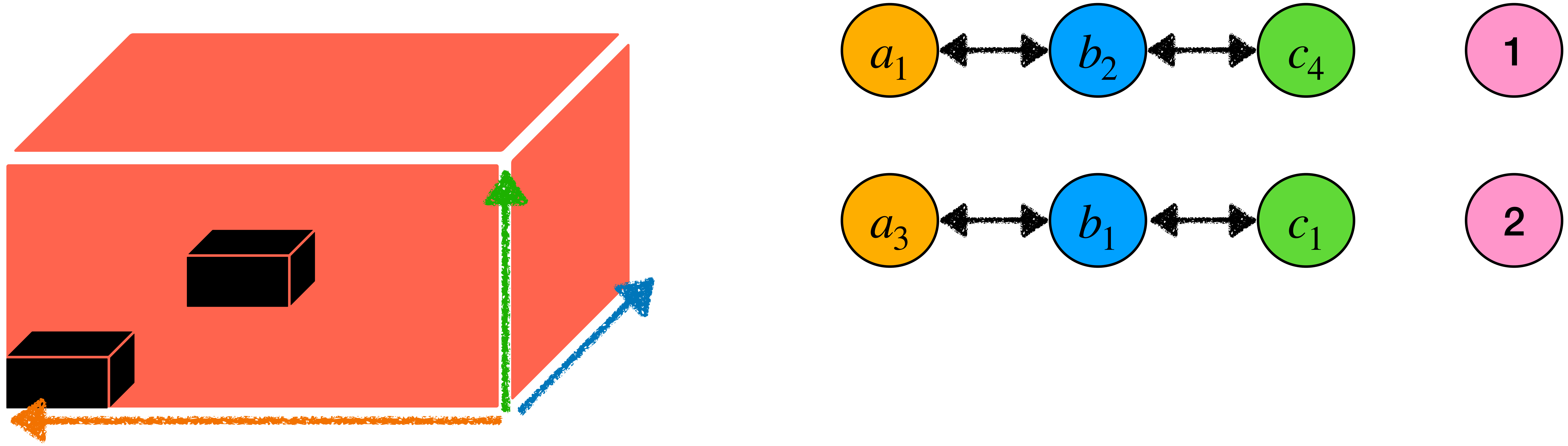
Discard all tuples containing these labels



$a_1 \longleftrightarrow b_2 \longleftrightarrow c_4$ $\quad$ 1

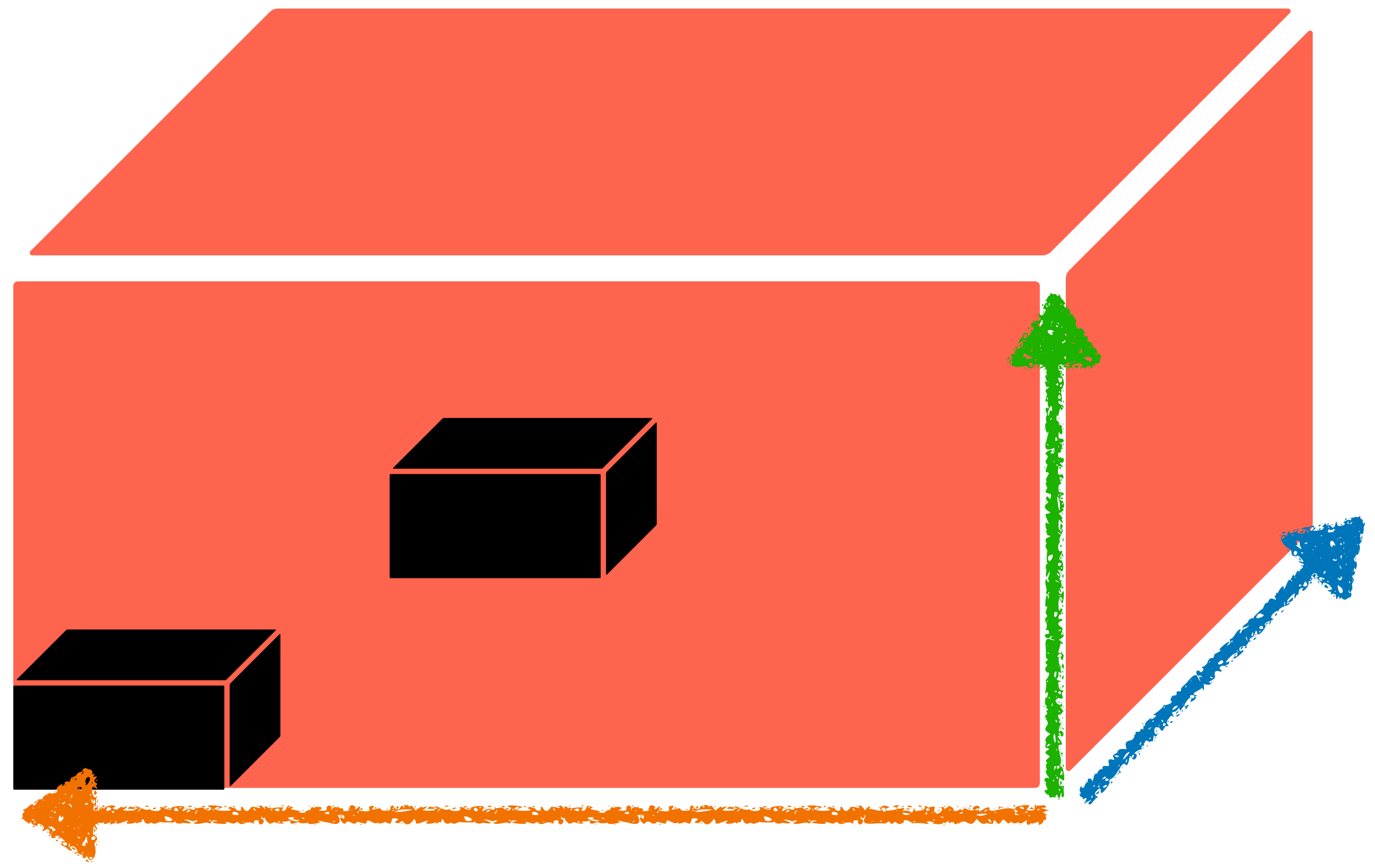
DOVER-Lap label mapping

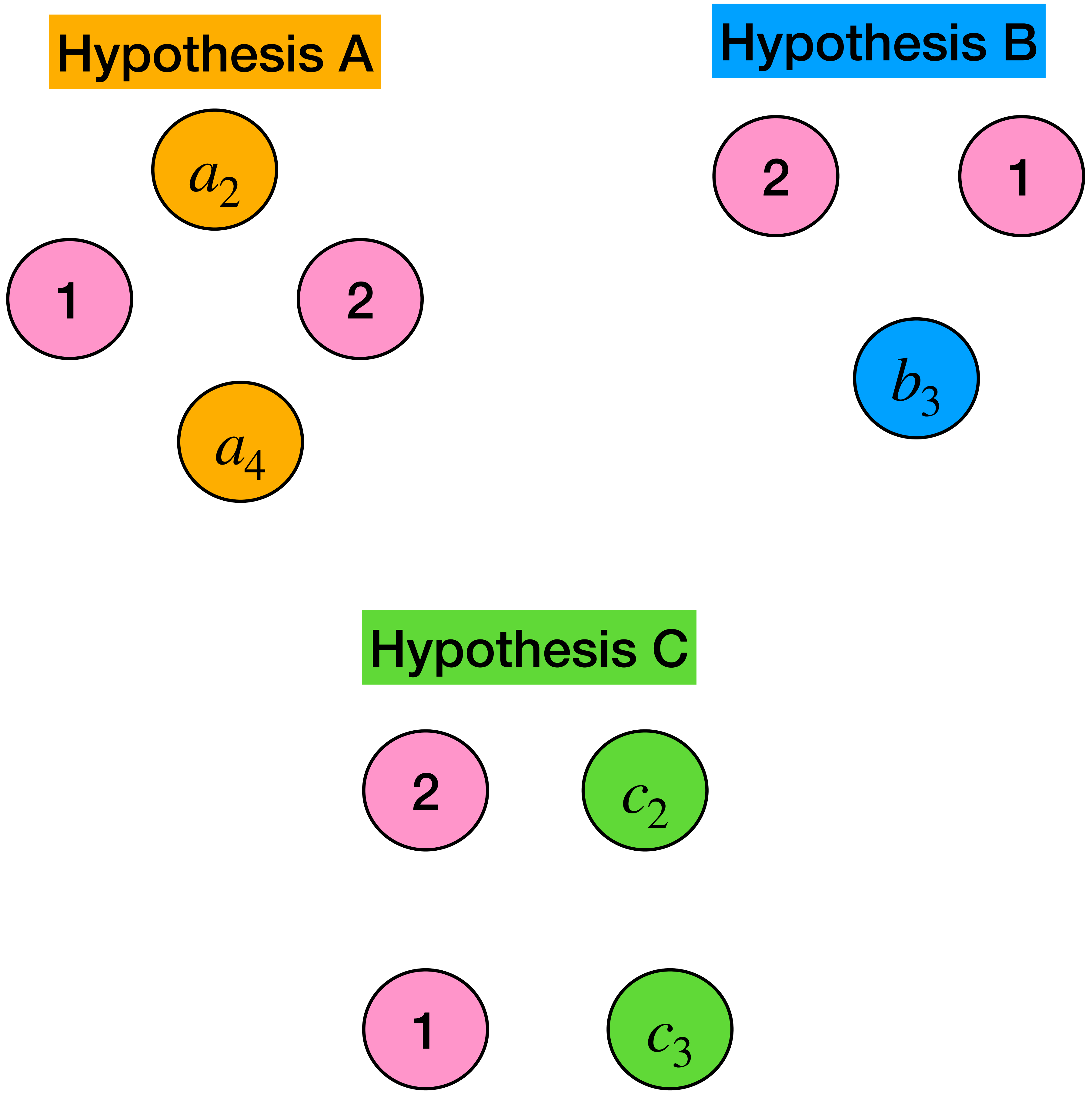
Pick tuple with lowest cost in remaining tensor



DOVER-Lap label mapping

Pick tuple with lowest cost in remaining tensor

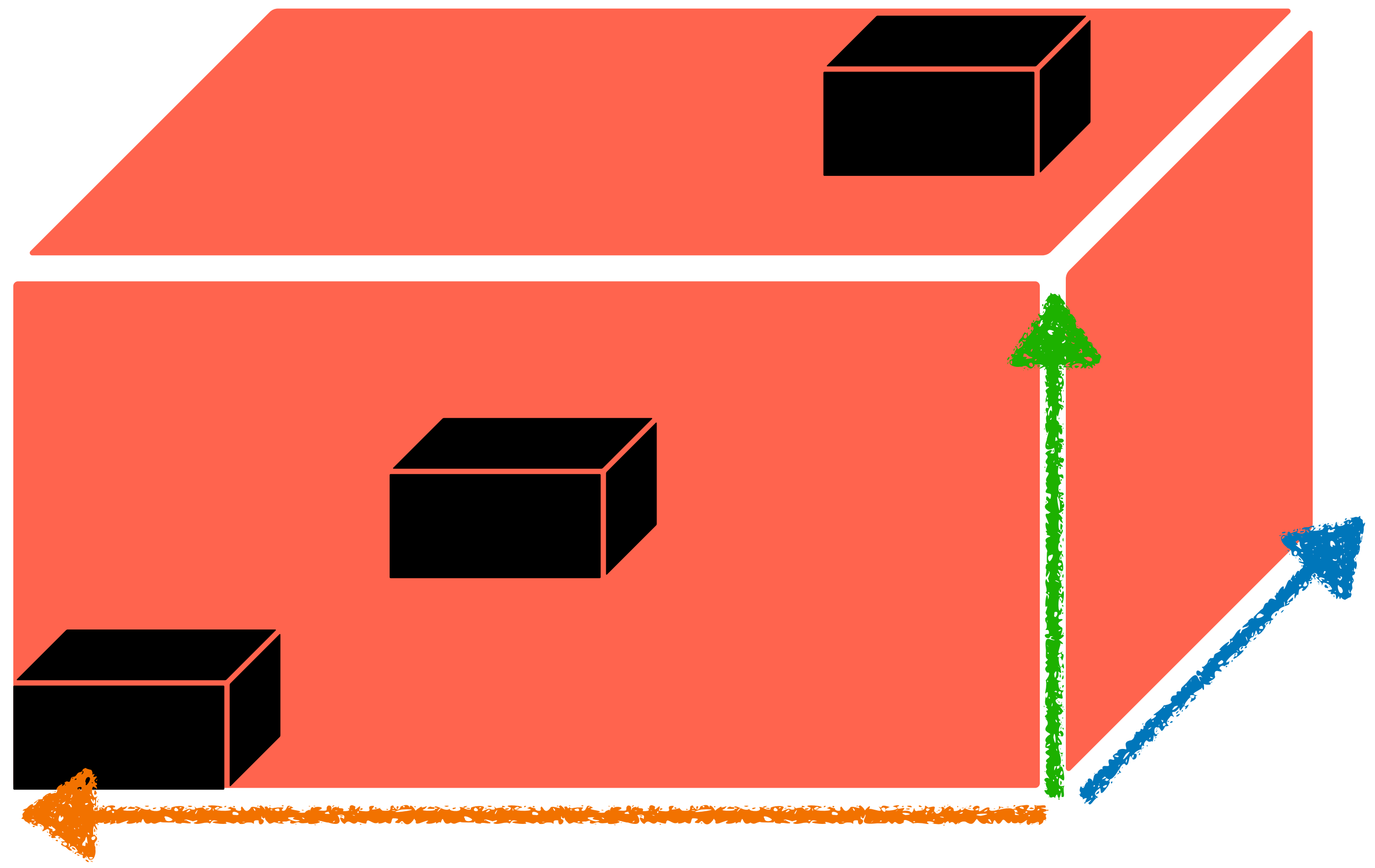


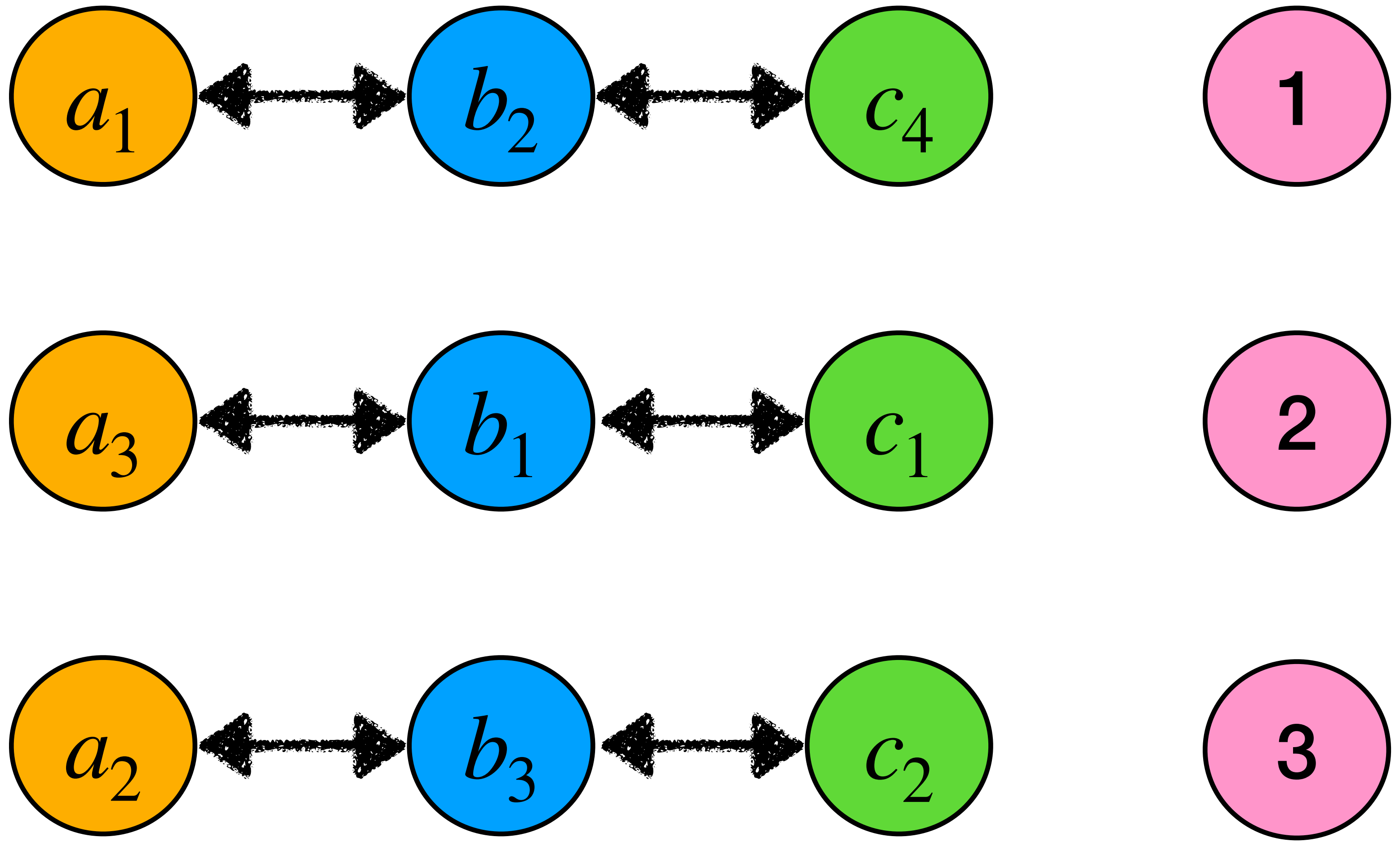


CENTER FOR LANGUAGE AND SPEECH PROCESSING

DOVER-Lap label mapping

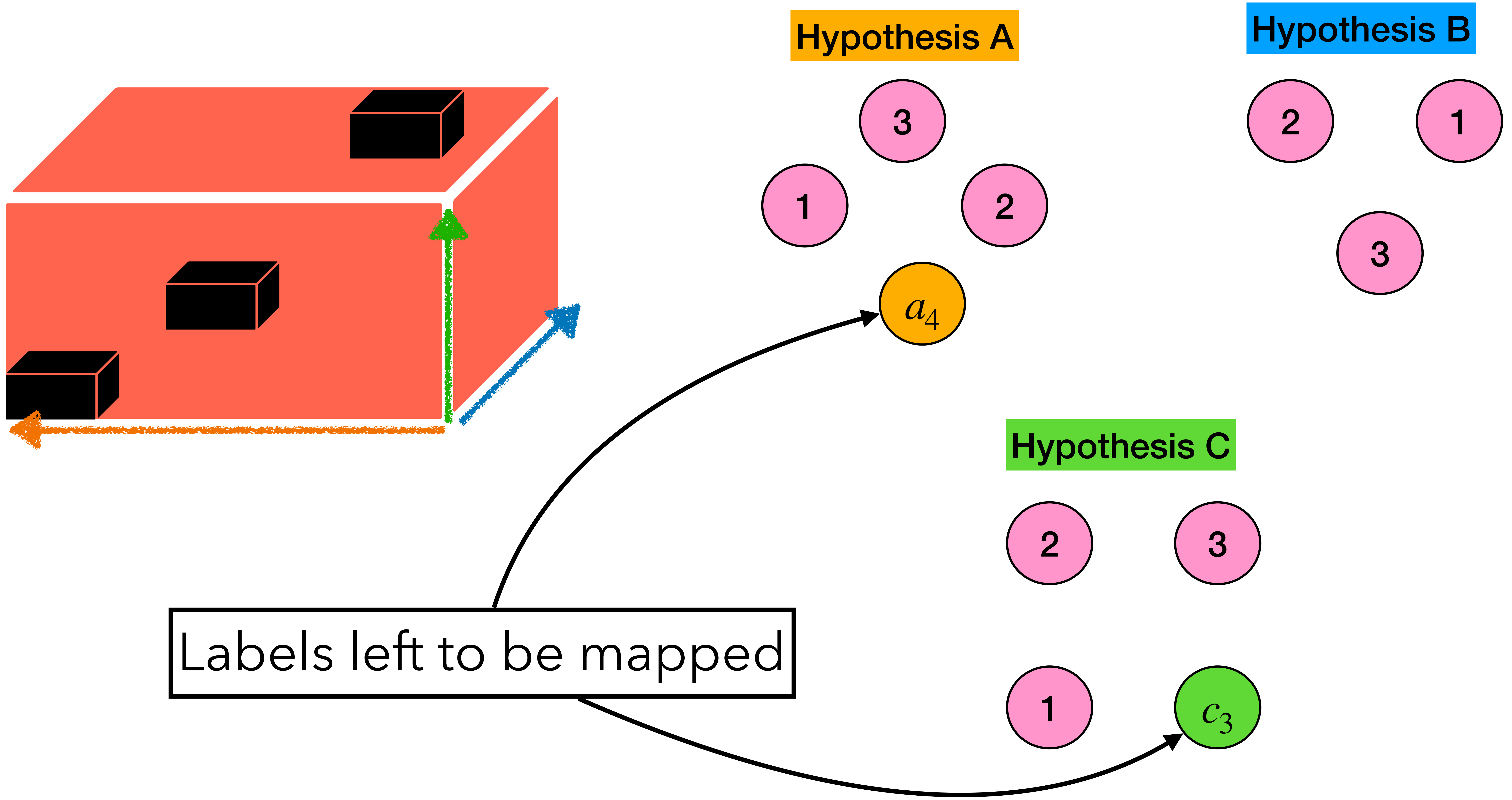
Repeat until no tuples are remaining





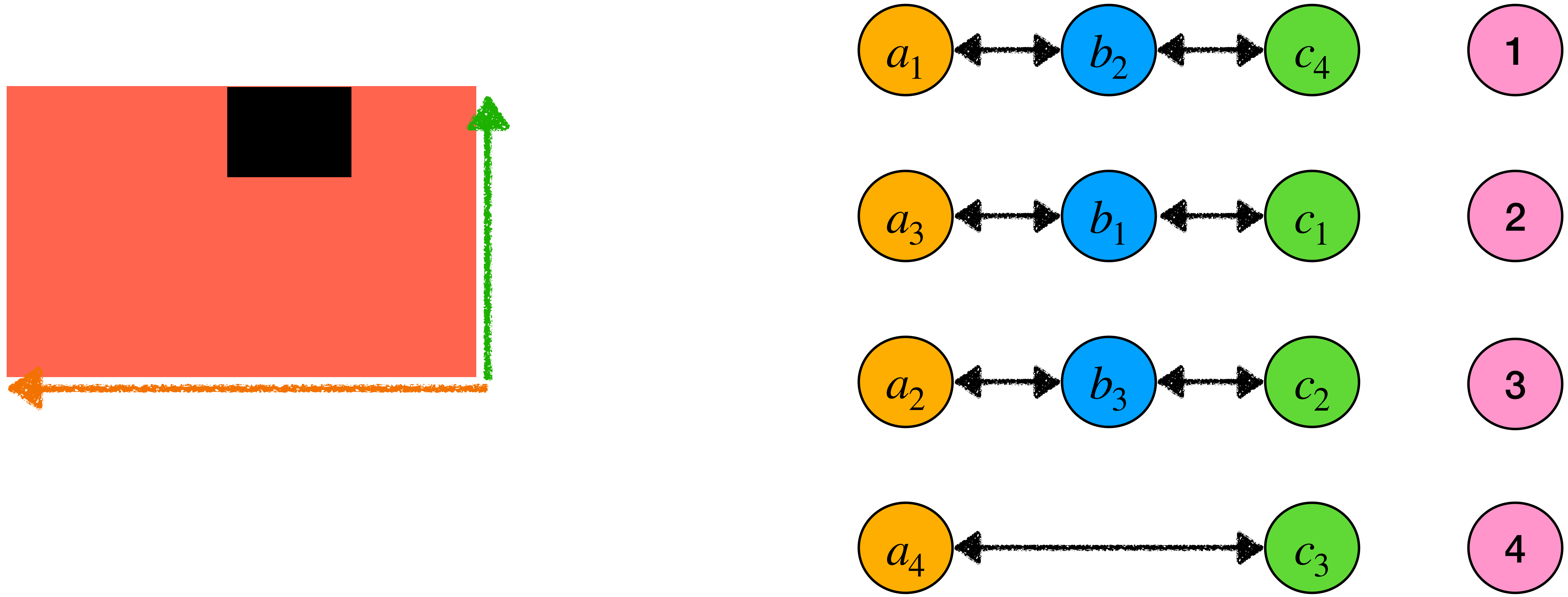
DOVER-Lap label mapping

Repeat until no tuples are remaining



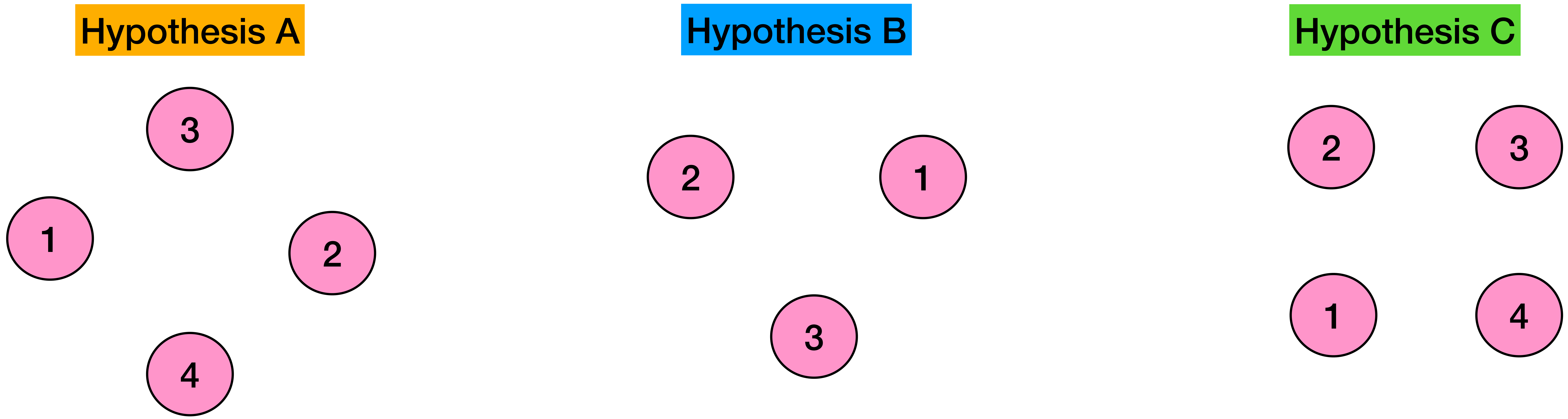
DOVER-Lap label mapping

If no tuples remaining but labels left to be mapped, remove filled dimensions and repeat



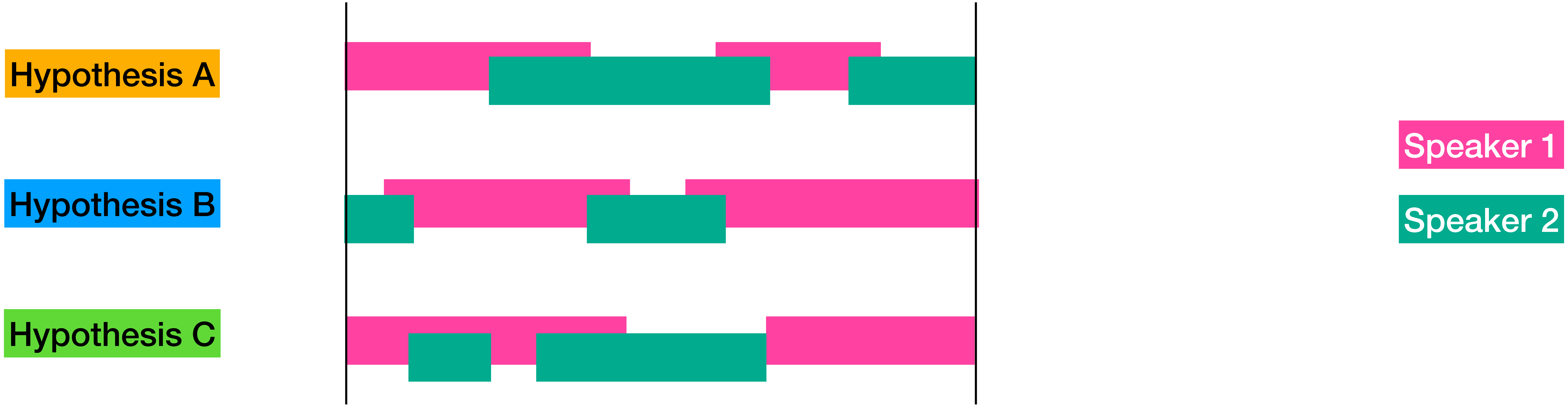
DOVER-Lap label mapping

Final mapped labels



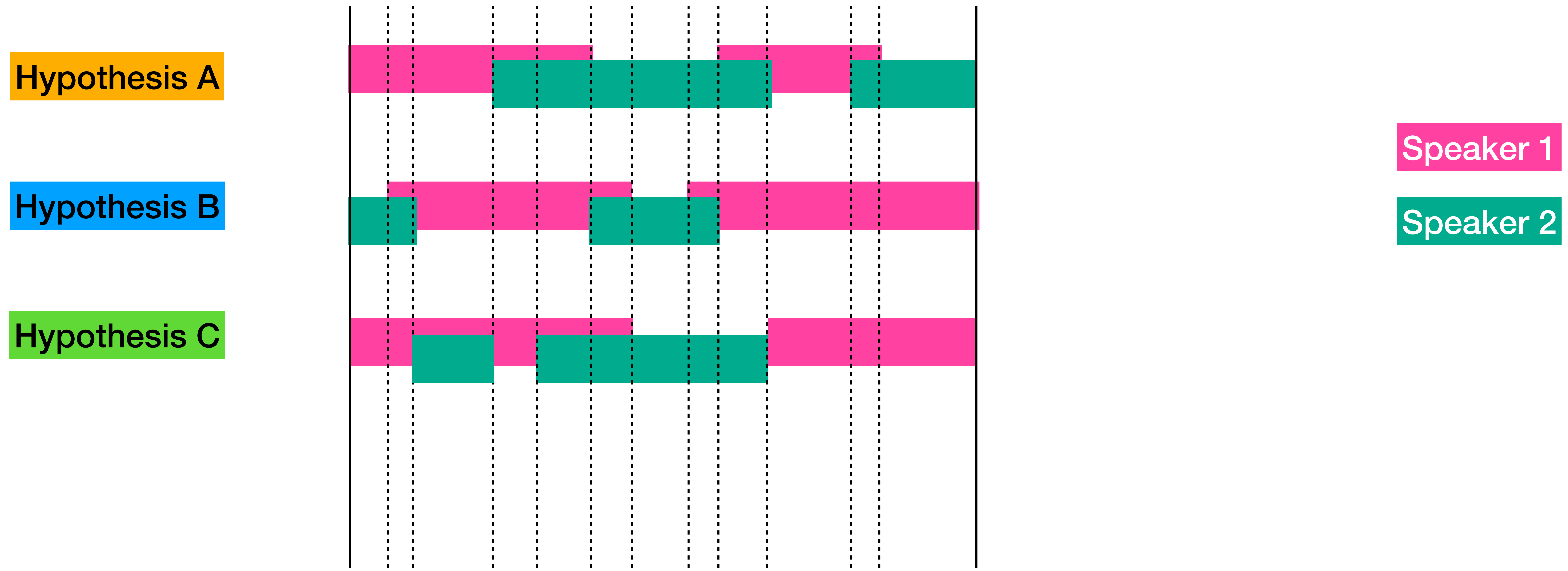
DOVER-Lap label voting

Consider 3 hypotheses from overlap-aware diarization systems



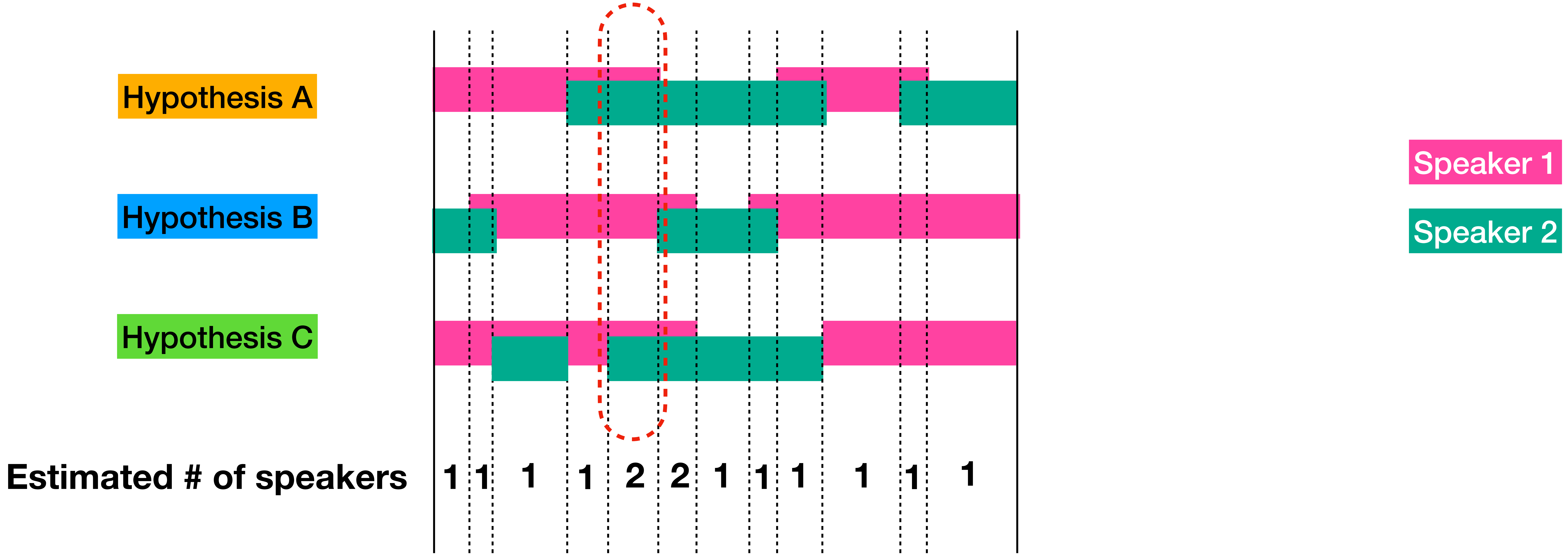
DOVER-Lap label voting

Divide into regions (similar to DOVER)



DOVER-Lap label voting

Estimate number of speakers in each region



 # speakers = weighted mean of # speakers in hypotheses

Weights -> obtained by ranking hypotheses by **total cost**

DOVER-Lap label voting

Assign highest weighted N speakers in each region

DOVER-Lap results: AMI

Effect of global label mapping algorithm

System	Spk. conf.	DER
Overlap-aware SC	10.1	23.6
VB-based overlap assignment*	9.6	21.5
Region proposal network	8.3	25.5
Average	9.3	23.5
DOVER	10.6	30.5
+ global label mapping	5.1	25.0

AMI data contains **4-speaker meetings**

DOVER-Lap results: AMI

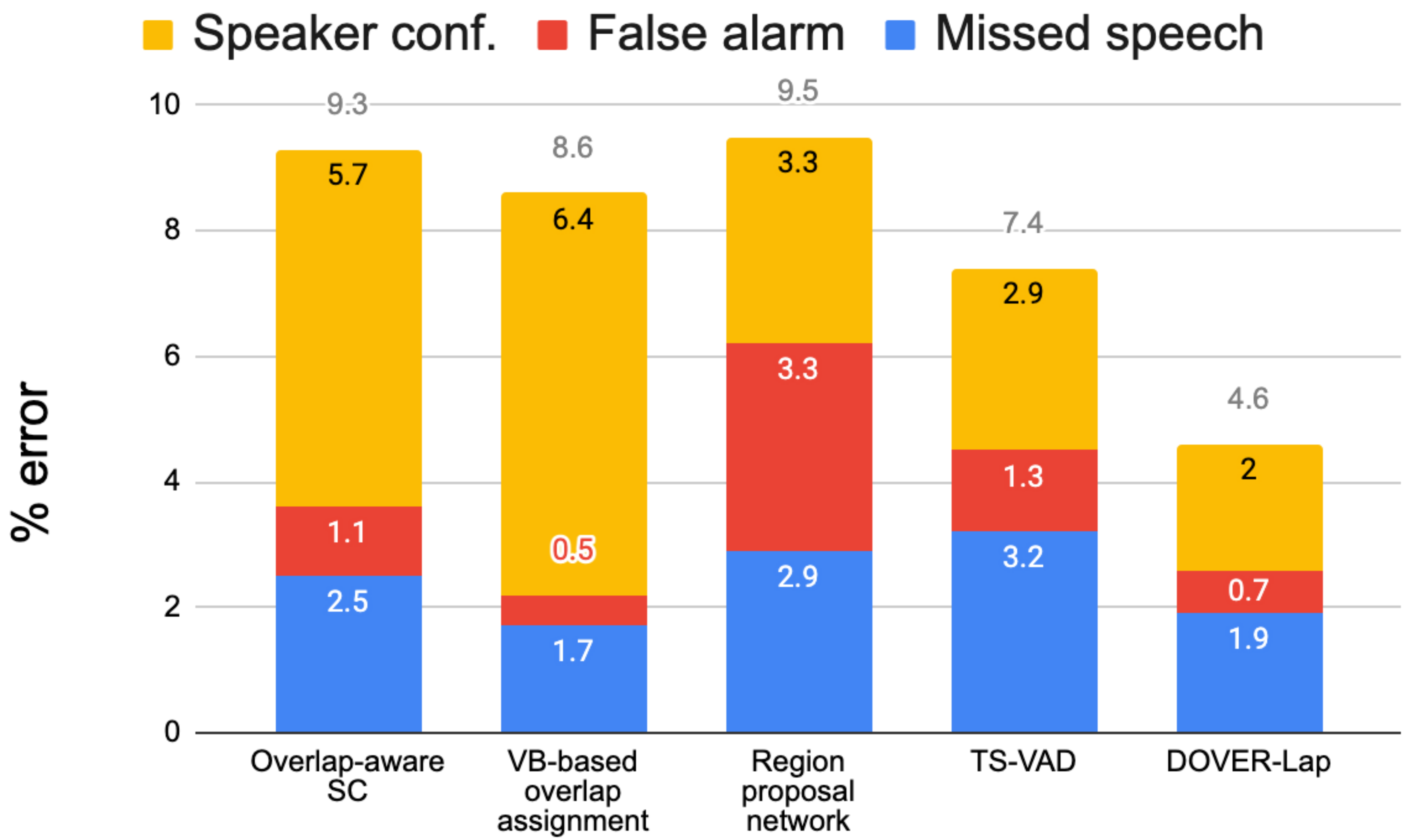
Effect of rank-weighted majority voting

CENTER FOR LANGUAGE AND SPEECH PROCESSING

System	Spk. conf.	DER
Overlap-aware SC	10.1	23.6
VB-based overlap assignment*	9.6	21.5
Region proposal network	8.3	25.5
Average	9.3	23.5
DOVER	10.6	30.5
+ global label mapping	5.1	25.0
DOVER-Lap	7.6	**20.3**

Results: Breakdown on LibriCSS

Effectively combines complementary strengths



LibriCSS data contains **8-speaker meetings**

Remember DIHARD?

Top 2 teams used DOVER-Lap for system fusion in DIHARD III

#1: USTC team combined clustering, separation-based, and TS-VAD systems

#2: Hitachi-JHU team combined VB-based and EEND-based systems

```
$ pip install dover-lap

$ dover-lap <output-rttm> <input-rttms>
```

It’s easy to use!

Summary

Diarization is a useful but difficult task.

Clustering-based systems fall short on handling overlapping speech, but small modifications inspired from mathematical insights can change this.

Ensembles work (especially for challenges). **DOVER-Lap** is a first attempt at combining overlap-aware diarization systems.

Acknowledgments

Some of the work reported here was done during **JSALT 2020** at JHU, with support from **Microsoft**, **Amazon**, and **Google**.

We thank **Maokui He** (USTC) for providing the TS-VAD diarization output on LibriCSS.

We thank **Takuya Yoshioka** (Microsoft) for providing the data simulation script that we used for training the overlap detector.

We thank **Shota Horiguchi** (Hitachi) for suggesting the modification that improved DOVER-Lap.