DOVER-Lap:

A method for combining overlap-aware diarization outputs

Desh Raj, Paola Garcia, Zili Huang, Shinji Watanabe, Daniel Povey, Andreas Stolcke, Sanjeev Khudanpur

Center for Language and Speech Processing, Johns Hopkins University
Xiaomi Corp., Beijing
Amazon Alexa Speech

Motivation

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.

CENTER FOR LANGUAGE AND SPEECH PROCESSING

Motivation

What is speaker diarization?

Task of “who spoke when”

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

Number of speakers may be unknown

Overlapping speech may be present

Output: ***homogeneous speaker segments***

SLT 2021

CENTER FOR LANGUAGE AND SPEECH PROCESSING

Motivation

Existing methods for diarization

Conventional methods

Spectral clustering (SC)
Agglomerative hierarchical clustering (AHC)
Variational Bayes (VBx)

- Clustering of small segment embeddings, such as i-vectors or x-vectors
- Optionally include overlap assignment

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.

Latane Bullock, Hervé Bredin, and L. Paola García-Perera, “Overlap-aware diarization: resegmentation using neural end-toend overlapped speech detection,” ICASSP 2020.

Motivation

Existing methods for diarization

New methods

Region proposal networks (RPN)

End-to-end neural diarization (EEND)

Target speaker voice activity detection (TS-VAD)

- Supervised training based systems, trained to directly predict segments.
- Includes overlap assignment by design

Zili Huang, Shinji Watanabe, Yusuke Fujita, Paola García, Yiwen Shao, Daniel Povey, and Sanjeev Khudanpur, “Speaker diarization with region proposal network,” ICASSP 2020.

Yusuke Fujita, Shinji Watanabe, Shota Horiguchi, Yawen Xue, and Kenji Nagamatsu, “End-to-end neural diarization: Reformulating speaker diarization as simple multi-label classification,” ArXiv.

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

SLT 2021

CENTER FOR LANGUAGE AND SPEECH PROCESSING

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?

- Systems 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”

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

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

Solution

DOVER-Lap performs “map and vote”

- Systems outputs may have different number of speaker estimates.
- System outputs may be 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

Result
on LibriCSS eval set

System	DER
Overlap-aware SC	9.3
VB-based overlap assignment	8.6
Region proposal network	9.5
TS-VAD	7.4
Combination using DOVER-Lap	**5.4**

Raj et al., “Multi-class spectral clustering with overlaps for speaker diarization,” IEEE SLT 2021.

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

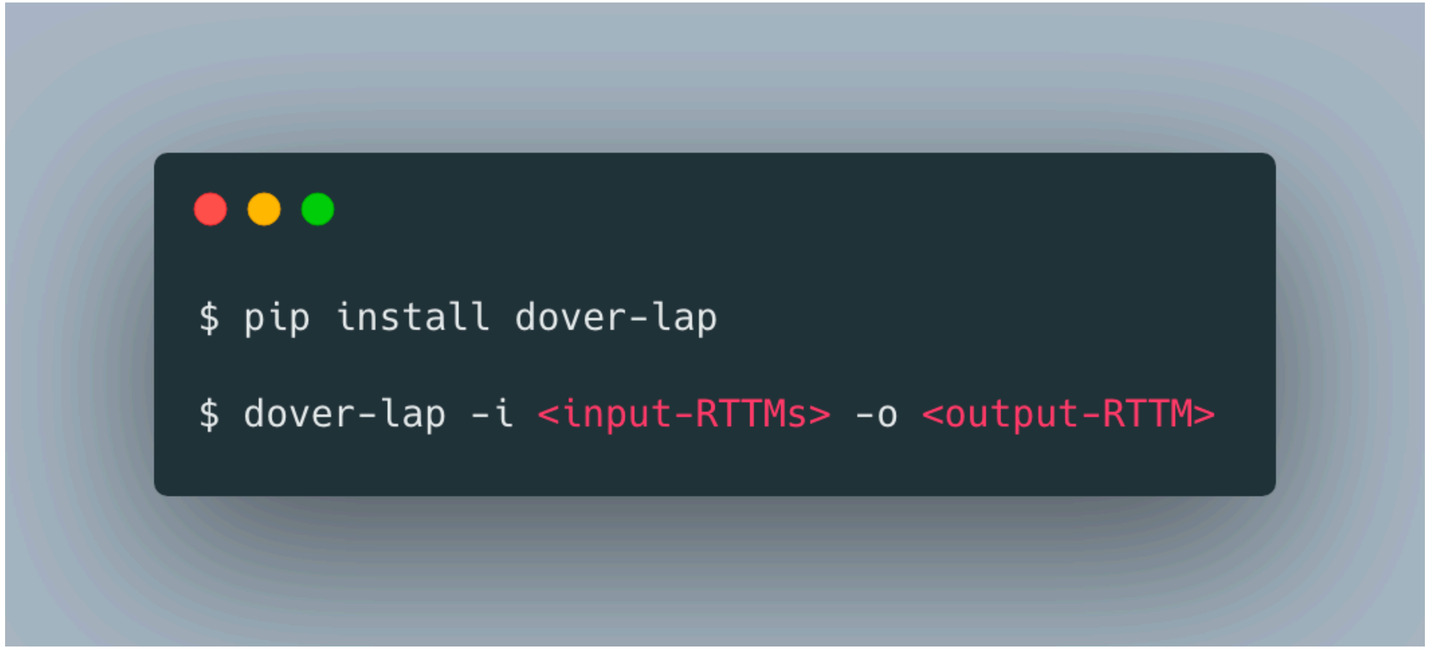
Huang et al., “Speaker diarization with region proposal network,” ICASSP 2020.

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

SLT 2021

CENTER FOR LANGUAGE AND SPEECH PROCESSING

How to use DOVER-Lap?



https://github.com/desh2608/dover-lap

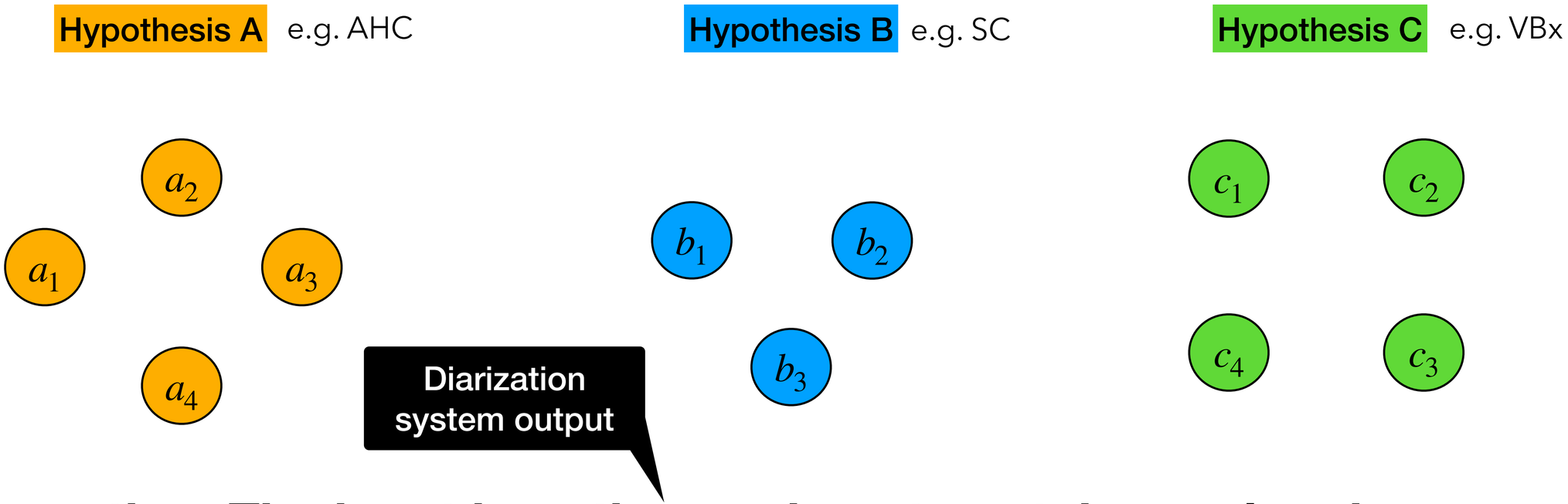
End of Highlight

Overview

- The DOVER-Lap algorithm
  - Preliminary: DOVER
  - How to map labels to a common space?
  - Overlap-aware majority voting
- Extended Results
  - Effect of global label mapping
  - System combination results (AMI and LibriCSS)

Preliminary: how DOVER works

Diarization Output Voting Error Reduction



Assumption: The input hypotheses do not contain overlapping segments.

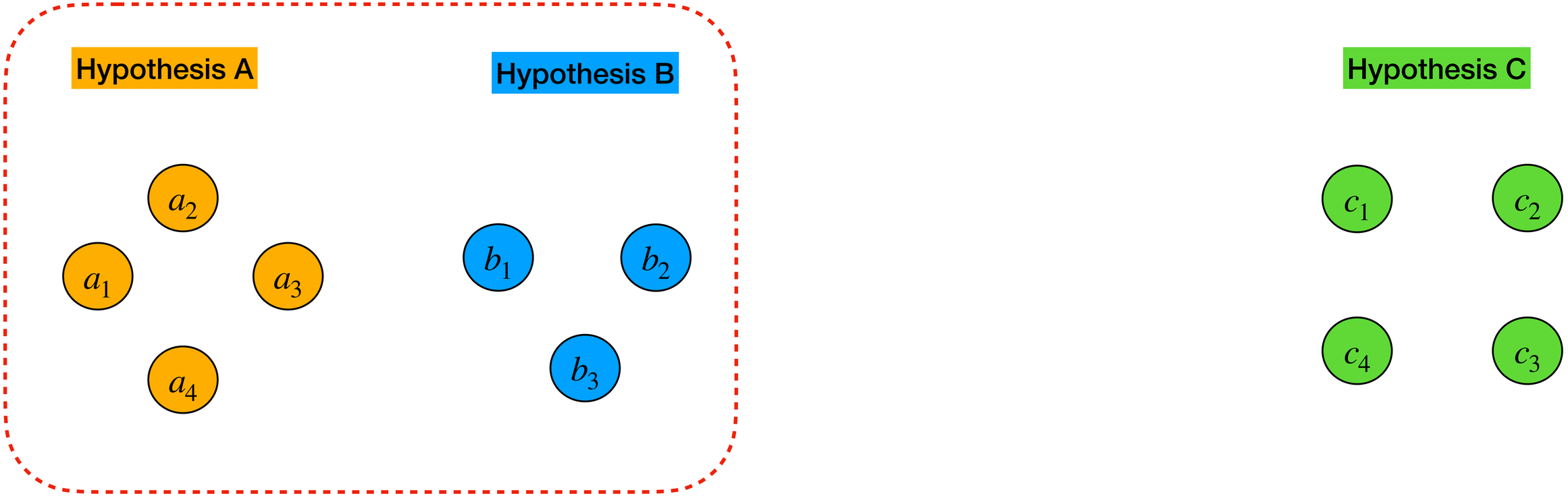
SLT 2021

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

CENTER FOR LANGUAGE AND SPEECH PROCESSING

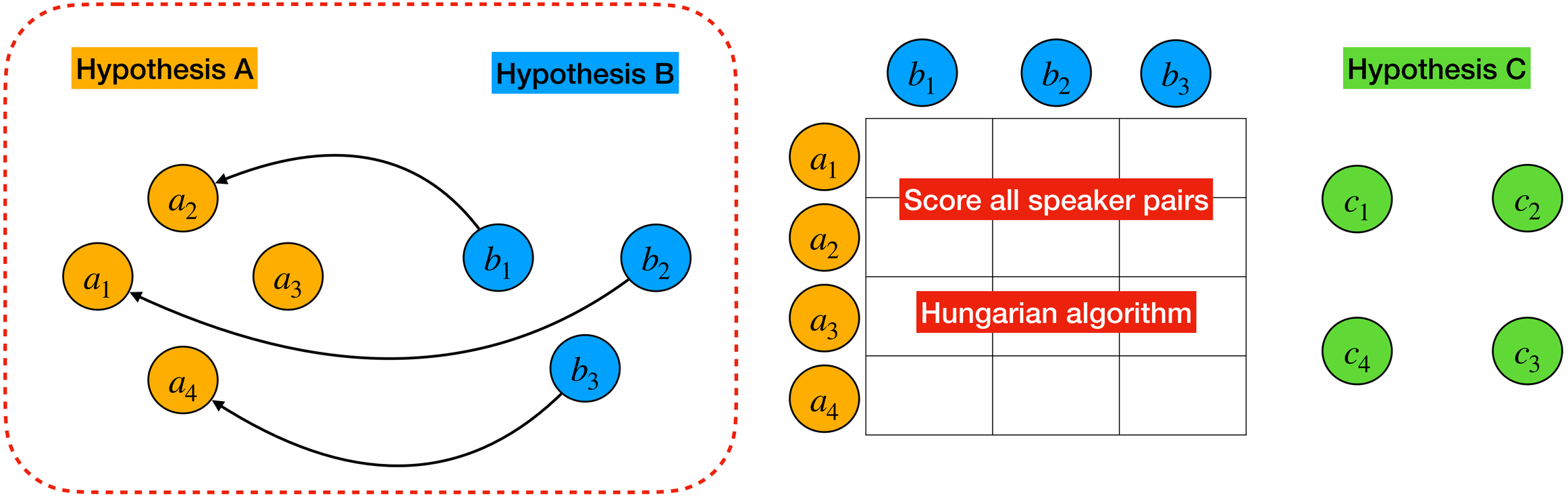
Preliminary: how DOVER works

Pair-wise incremental label mapping



Preliminary: how DOVER works

Pair-wise incremental label mapping



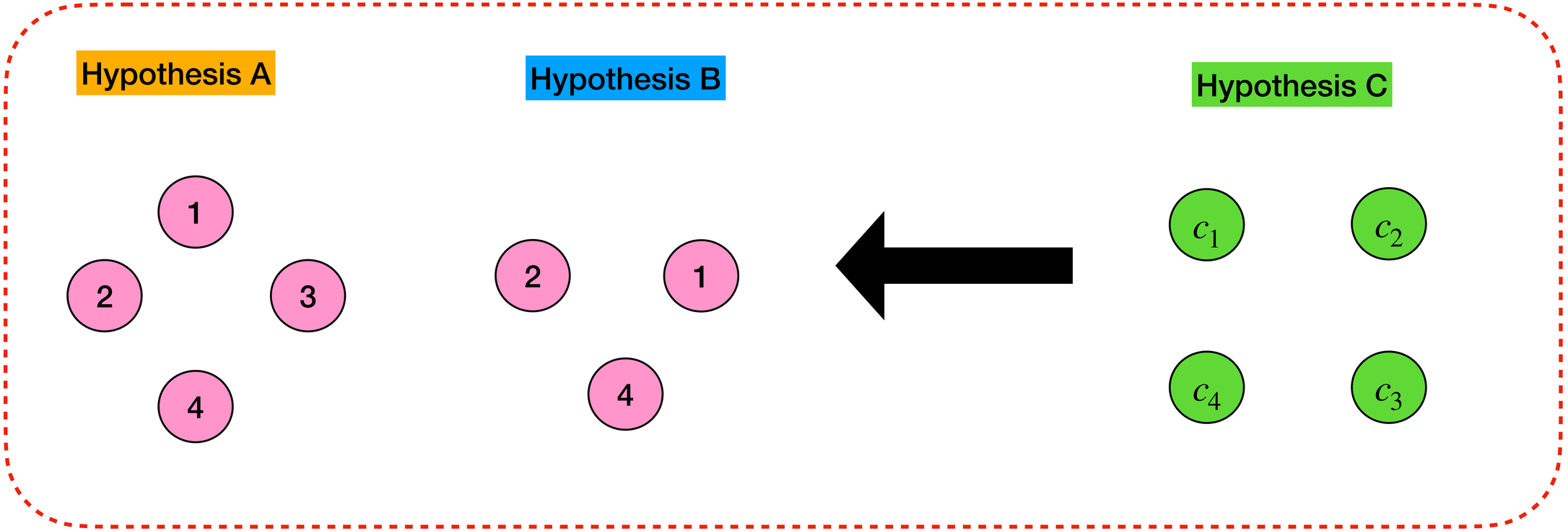
This is the same algorithm that is used to map hypothesis to reference for DER computation.

SLT 2021

CENTER FOR LANGUAGE AND SPEECH PROCESSING

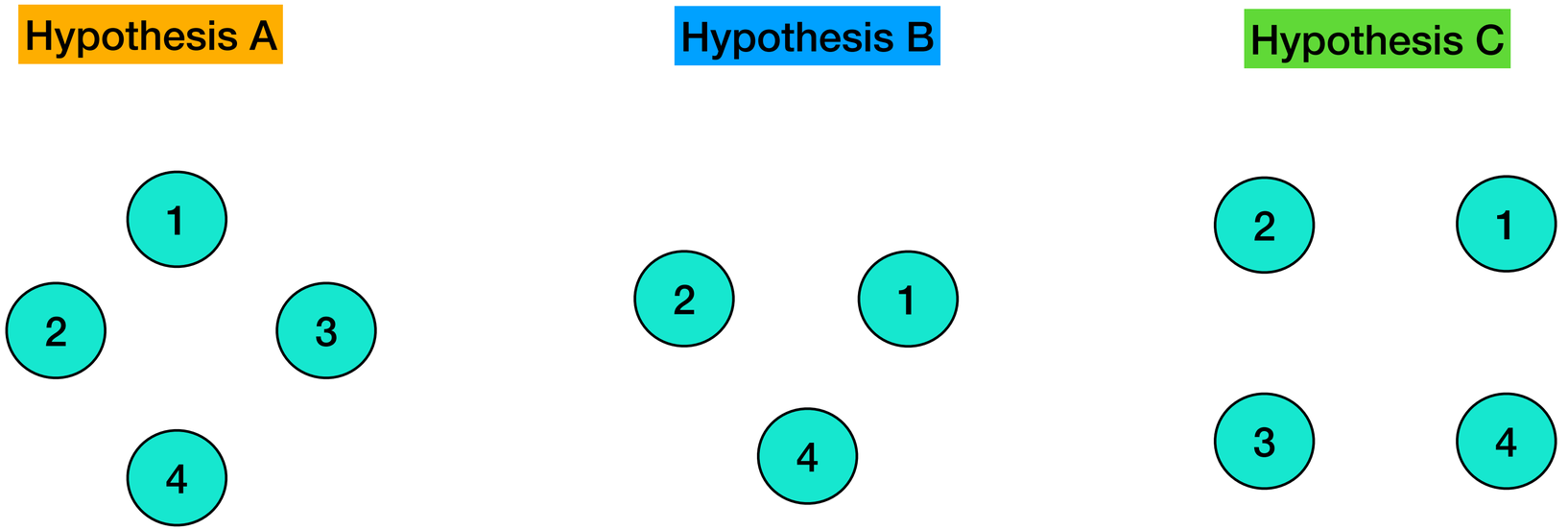
Preliminary: how DOVER works

Pair-wise incremental label mapping



Preliminary: how DOVER works

Pair-wise incremental label mapping



Preliminary: how DOVER works

Pair-wise incremental label mapping

Hypothesis A

Hypothesis B

Hypothesis C

How to choose starting anchor?

Method 1 (centroid selection): Rank all the hypothesis based on average DER to all other hypothesis. Choose the top-ranked as anchor.

Method 2: Run N times, once with each hypothesis as anchor and finally average all.

a_4 a_4 a_3 a_4

SLT 2021

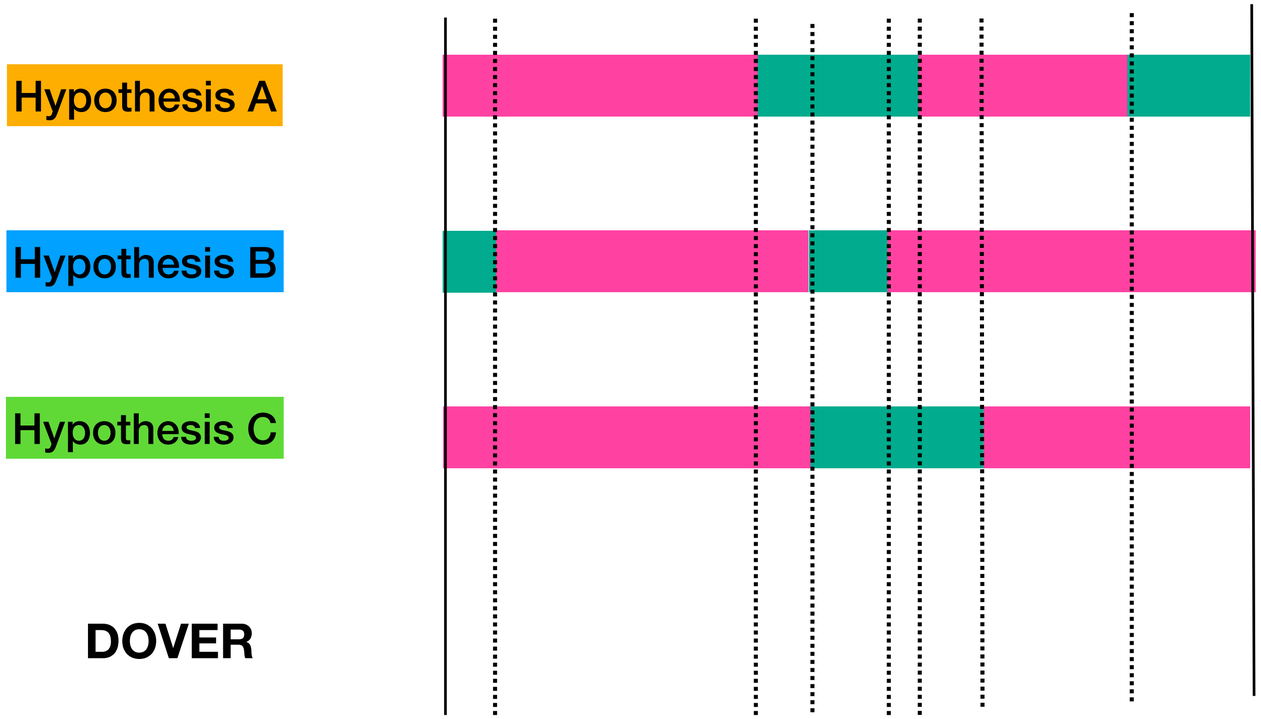
CENTER FOR LANGUAGE AND SPEECH PROCESSING

Preliminary: how DOVER works

Label voting using rank-weighting

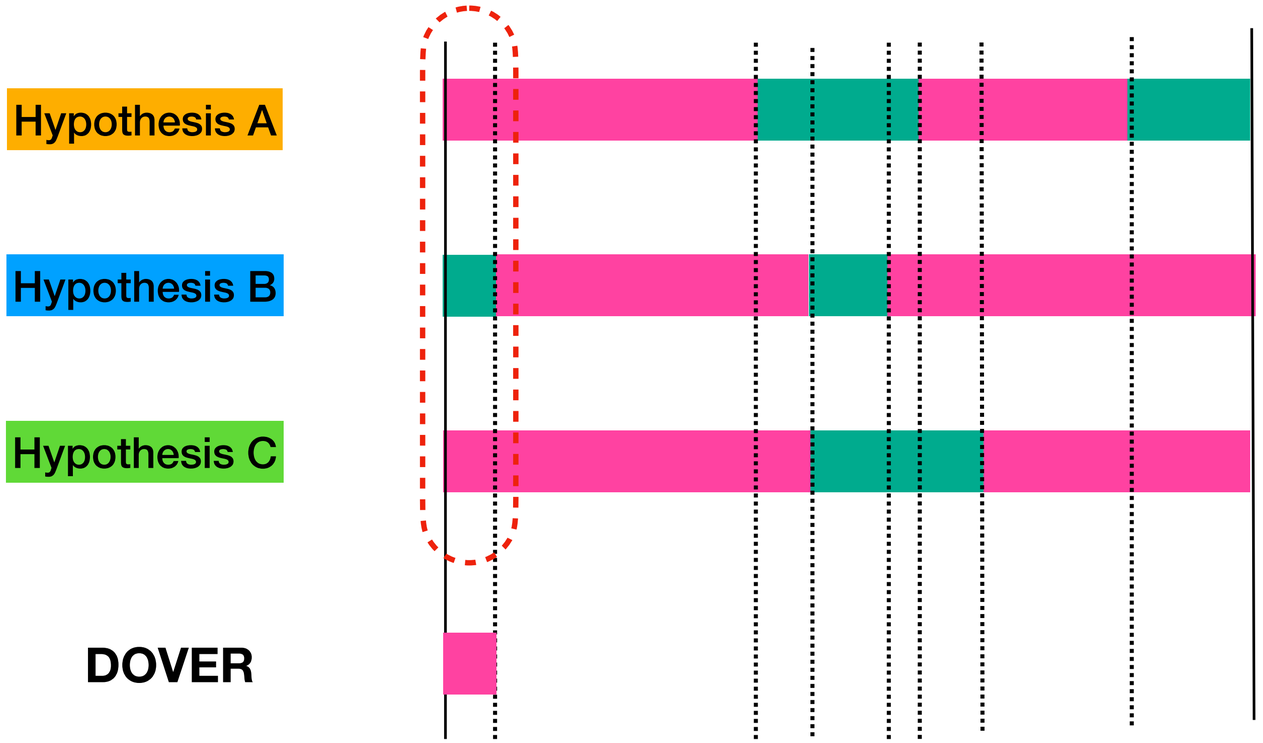
Preliminary: how DOVER works

Label voting using rank-weighting



Preliminary: how DOVER works

Label voting using rank-weighting



Speaker 1

Speaker 2

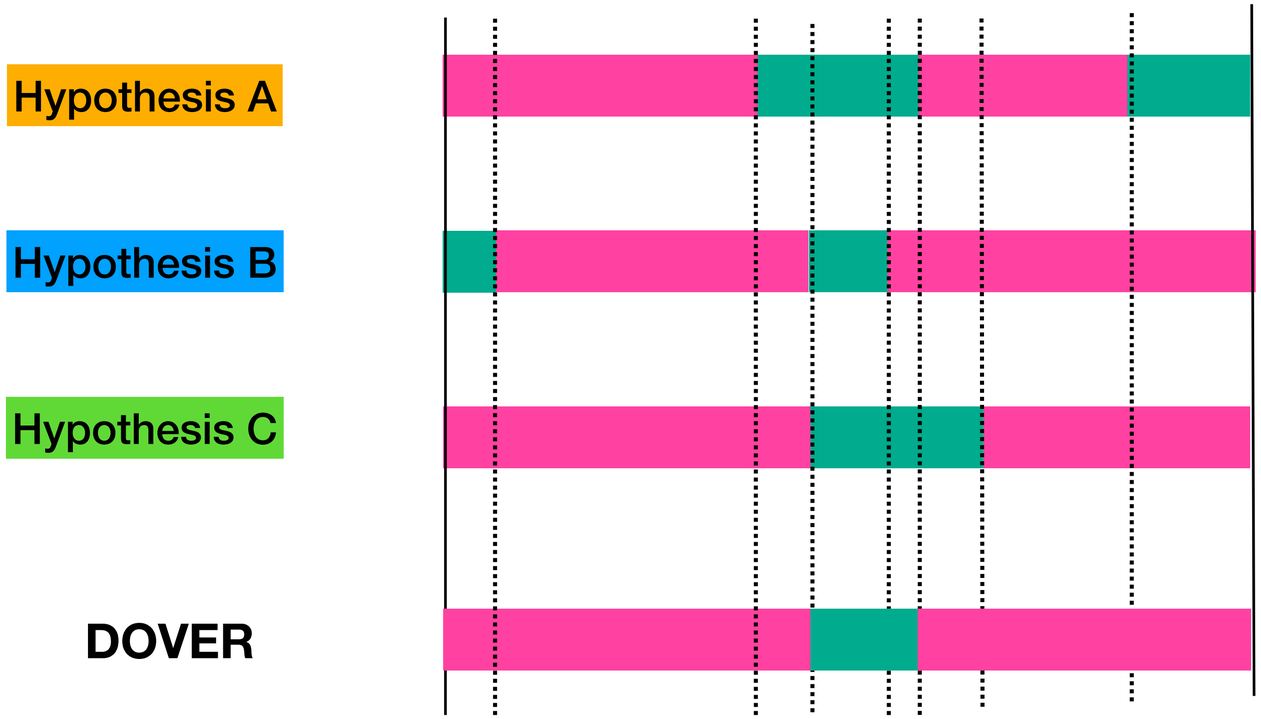
Voting using rank-based weights

SLT 2021

CENTER FOR LANGUAGE AND SPEECH PROCESSING

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

SLT 2021

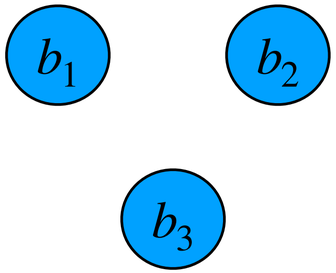
CENTER FOR LANGUAGE AND SPEECH PROCESSING

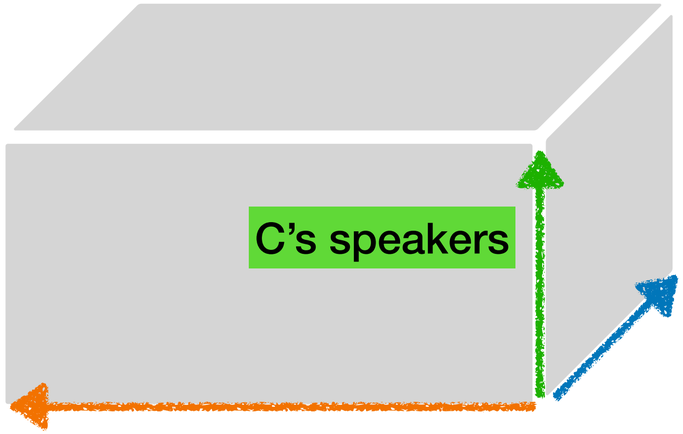
DOVER-Lap label mapping

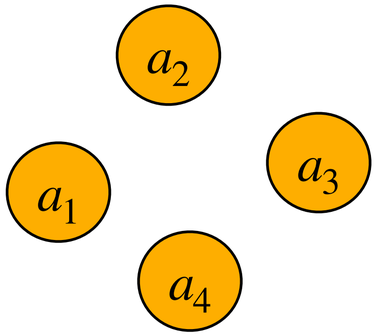
Change incremental method to global

Hypothesis B

e.g. RPN

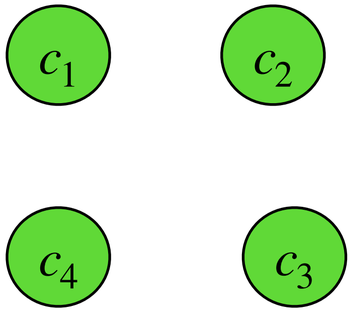






e.g. EEND

Hypothesis A



e.g. TS-VAD

Hypothesis C

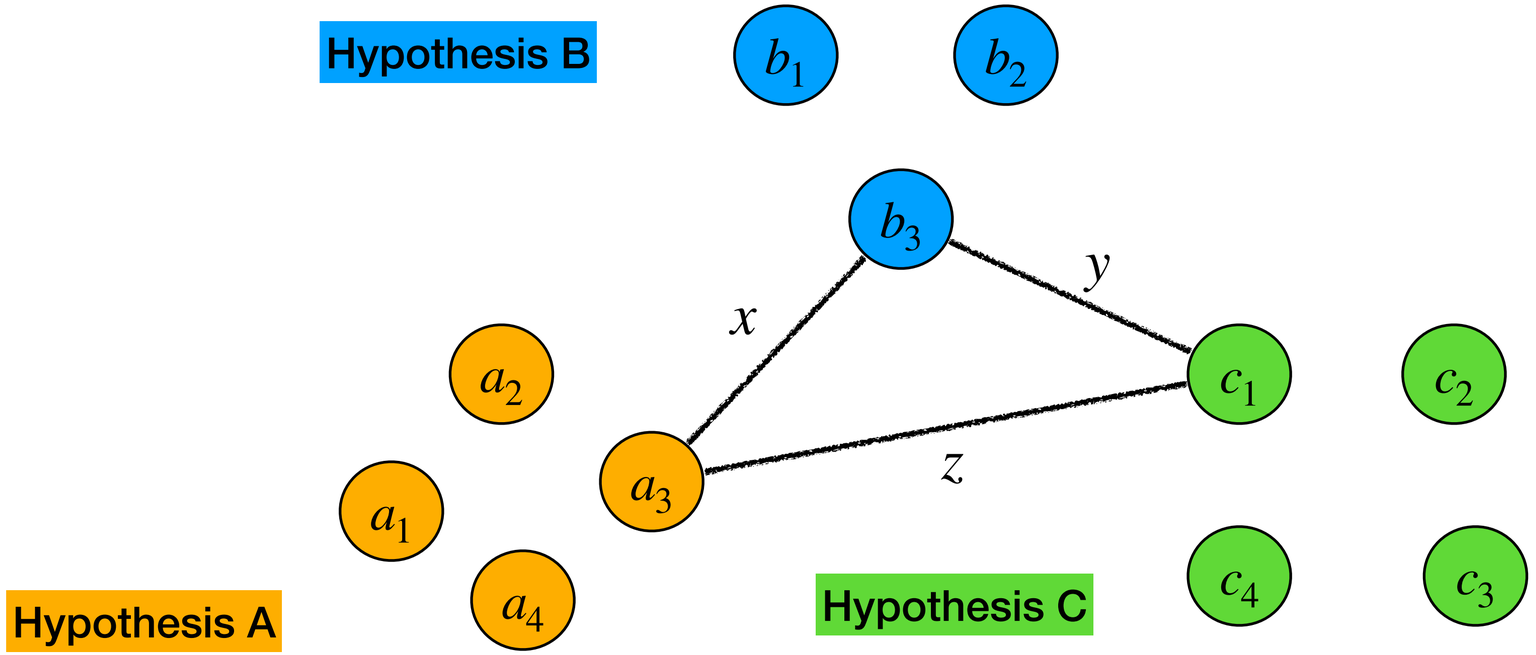
Hypotheses can contain overlapping segments.

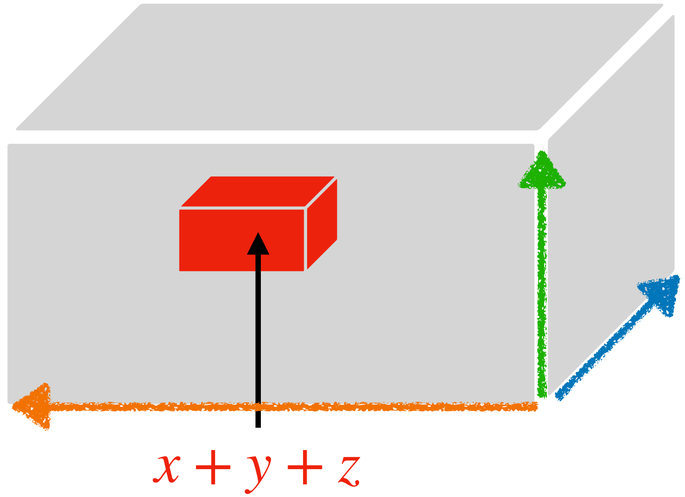
SLT 2021

CENTER FOR LANGUAGE AND SPEECH PROCESSING

DOVER-Lap label mapping

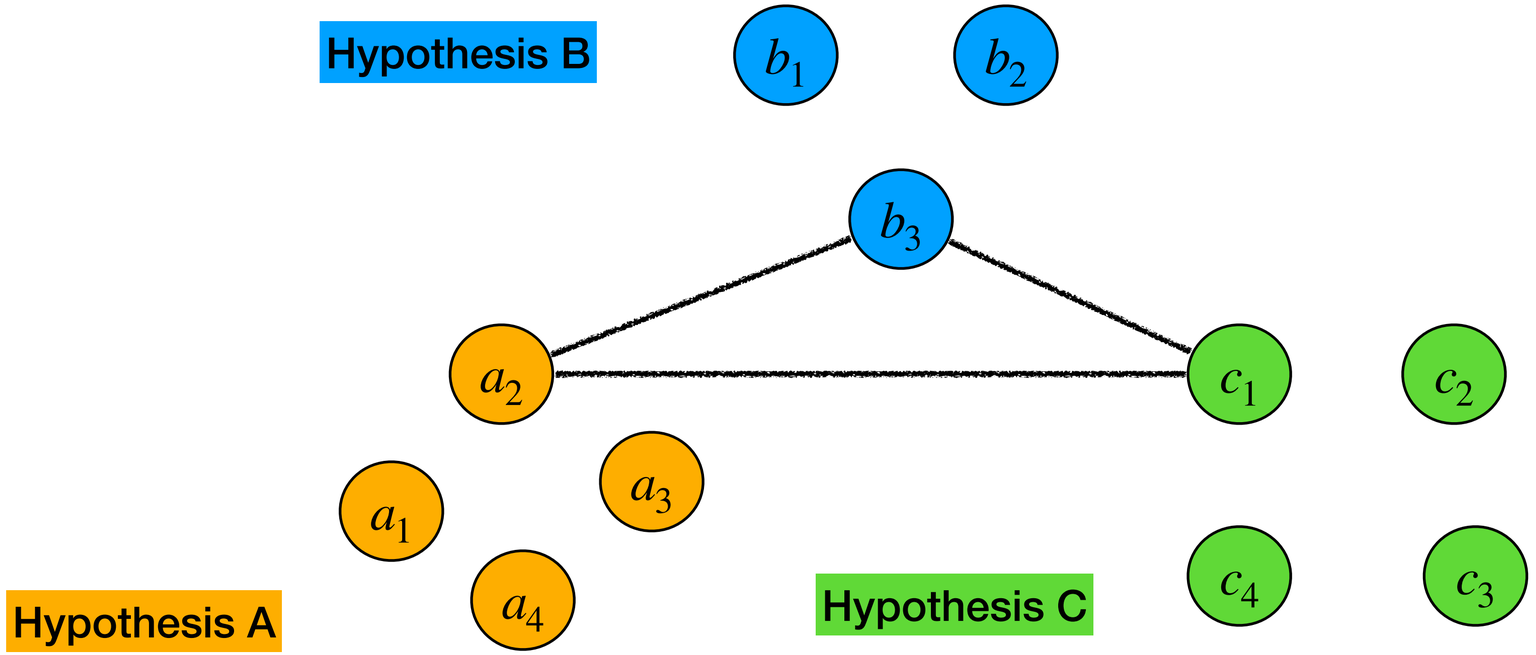
Compute “tuple costs” for all tuples

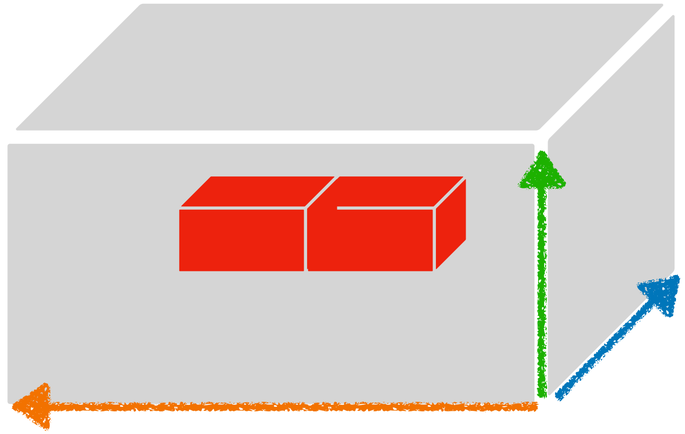




DOVER-Lap label mapping

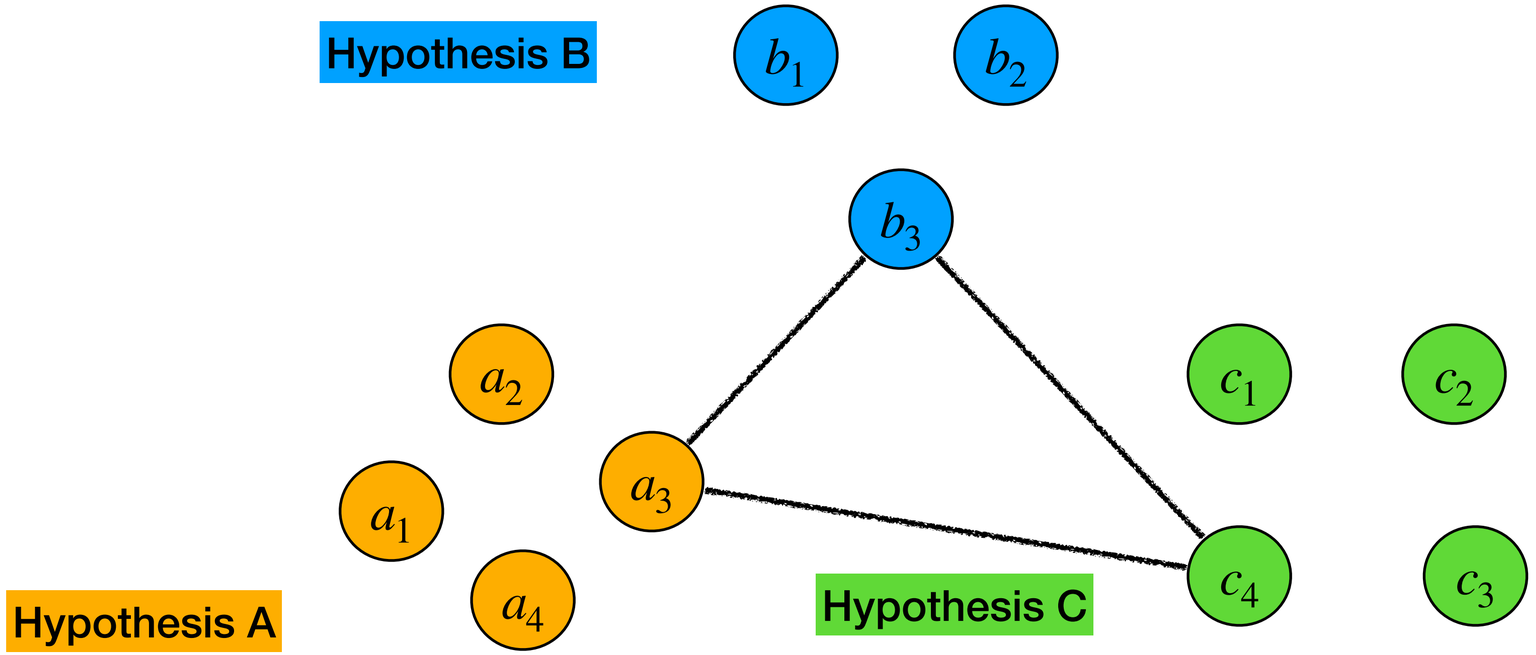
Compute “tuple costs” for all tuples

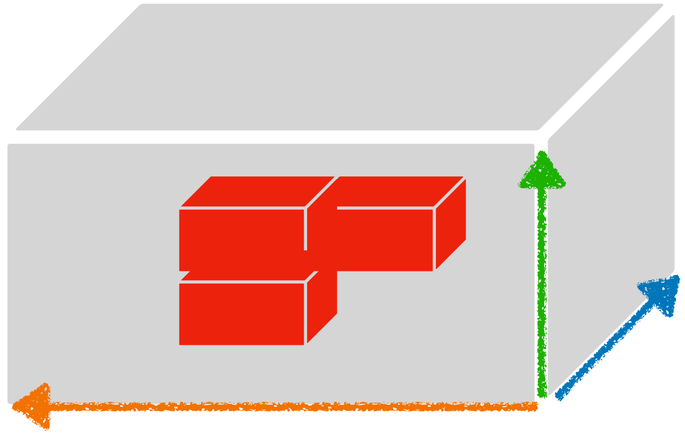




DOVER-Lap label mapping

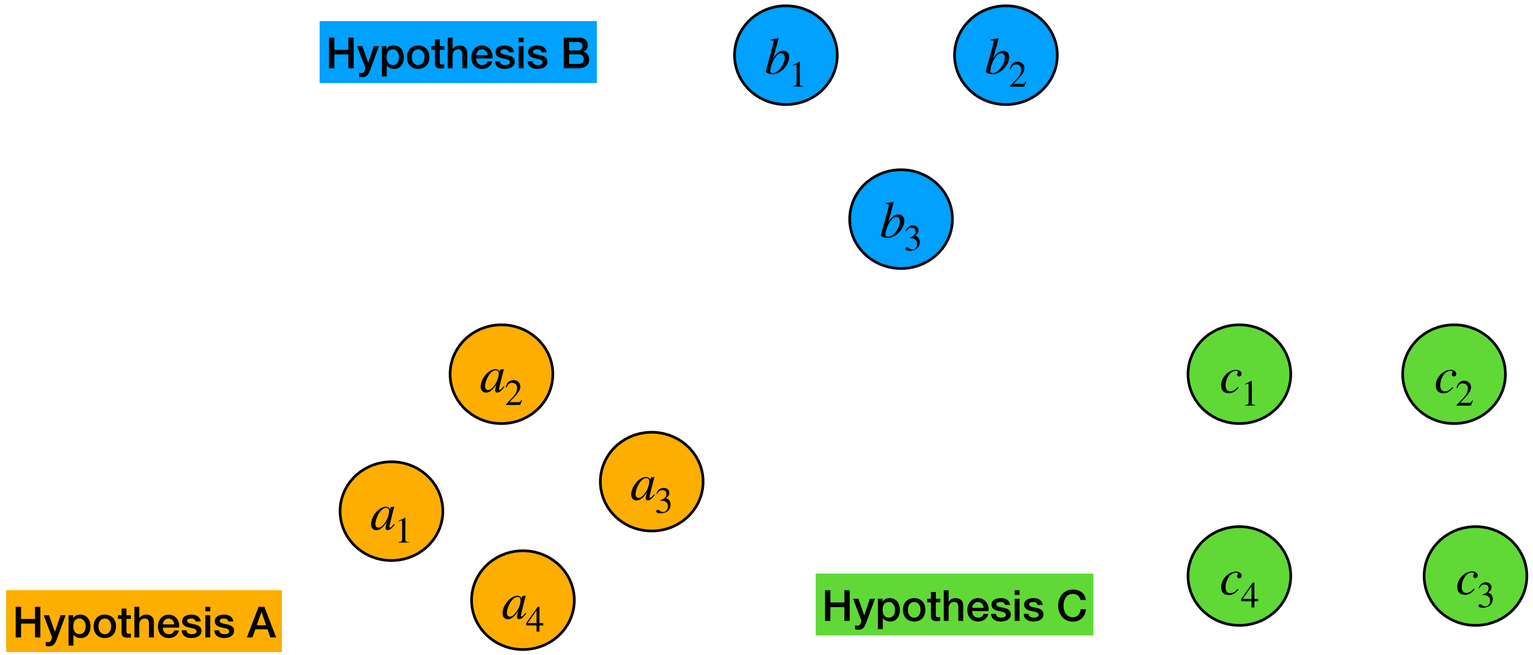
Compute “tuple costs” for all tuples





DOVER-Lap label mapping

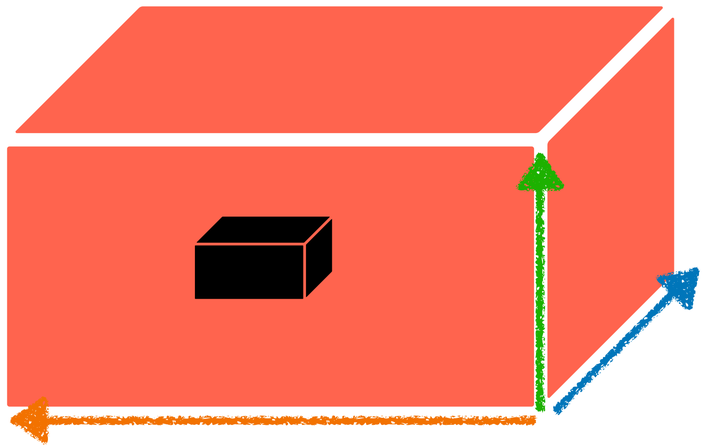
This gives us a “global” cost tensor

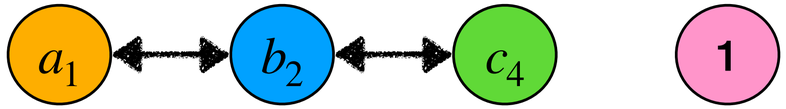


Global cost tensor

DOVER-Lap label mapping

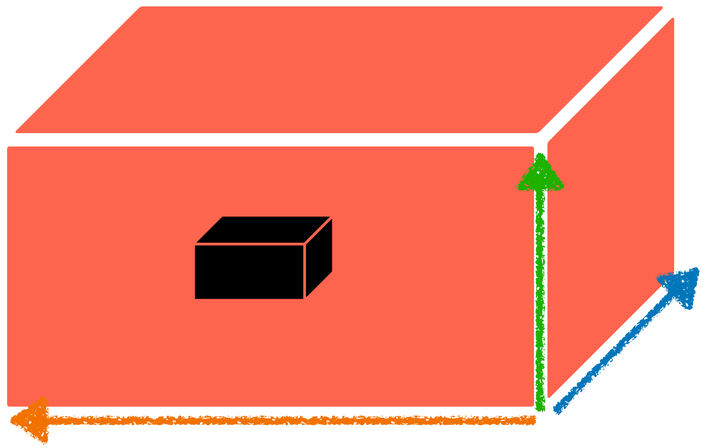
Pick tuple with the lowest cost and assign them same label

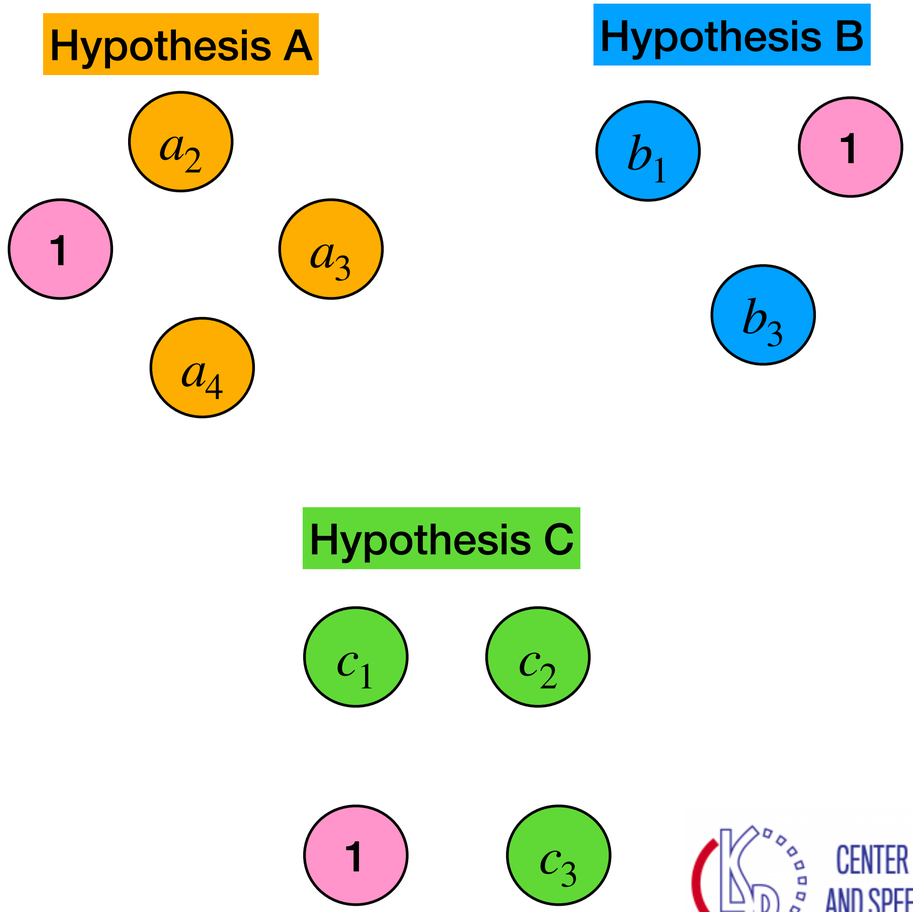




DOVER-Lap label mapping

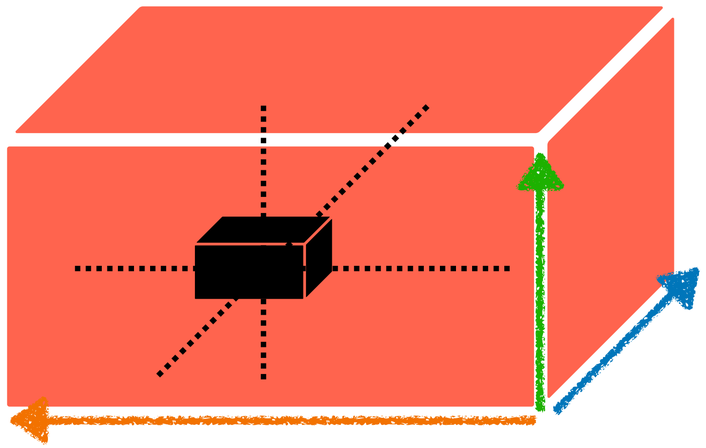
Pick tuple with the lowest cost and assign them same label

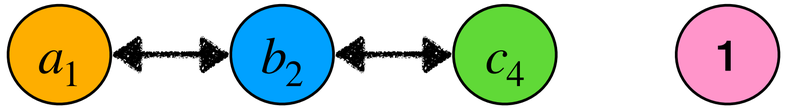




DOVER-Lap label mapping

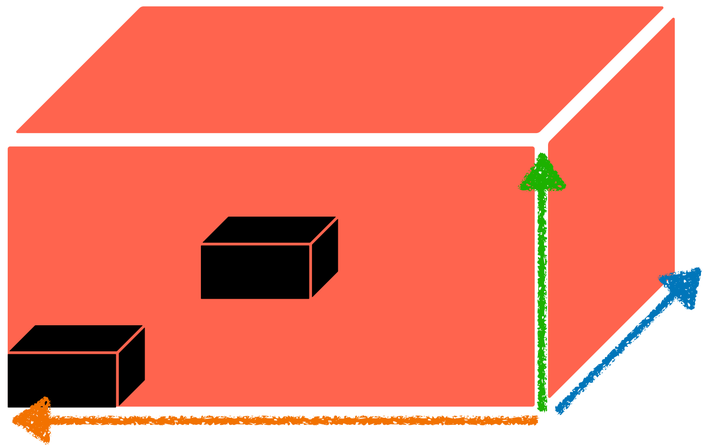
Discard all tuples containing these labels

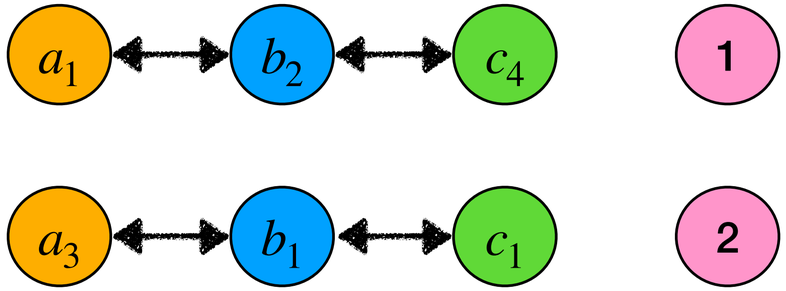




DOVER-Lap label mapping

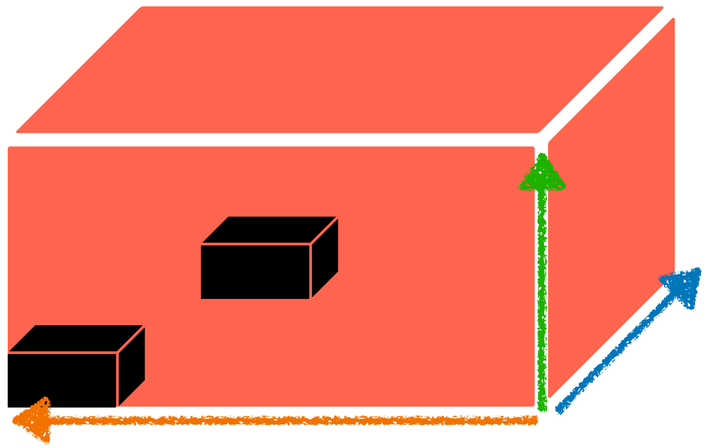
Pick tuple with lowest cost in remaining tensor

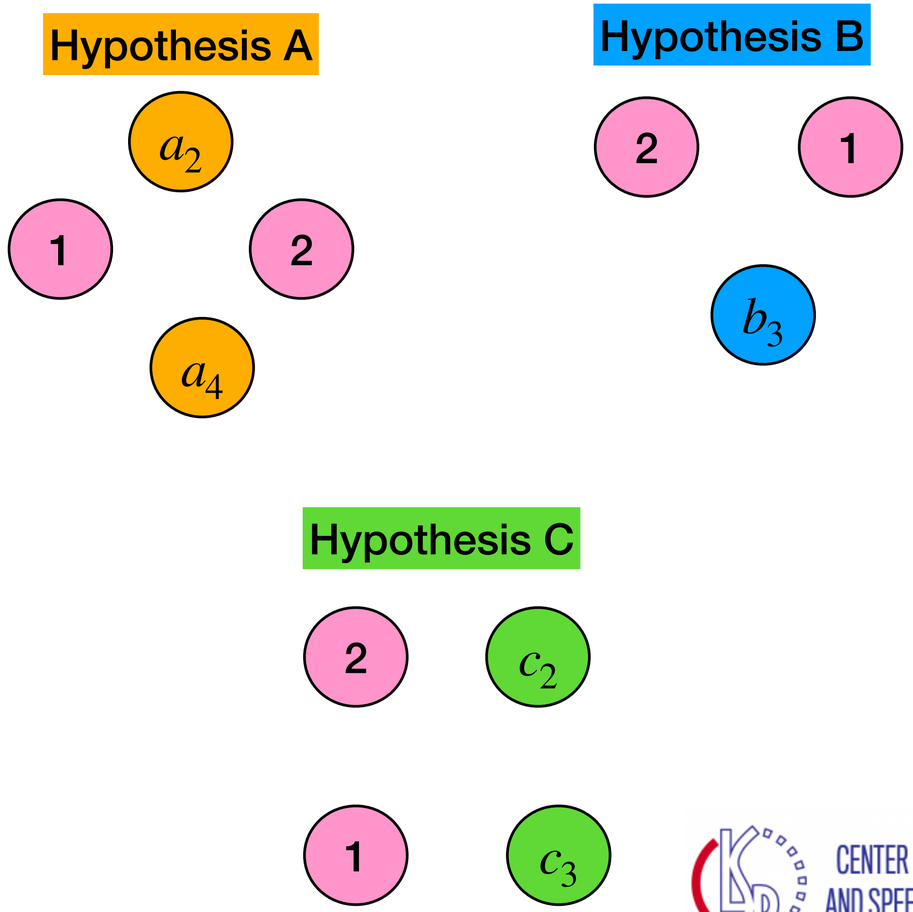




DOVER-Lap label mapping

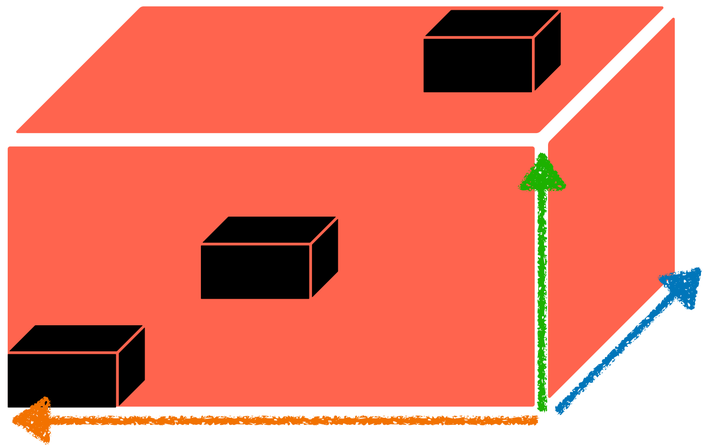
Pick tuple with lowest cost in remaining tensor

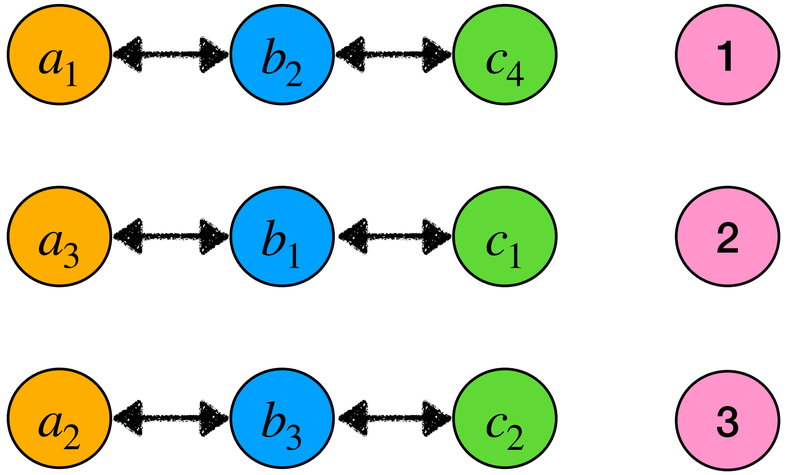




DOVER-Lap label mapping

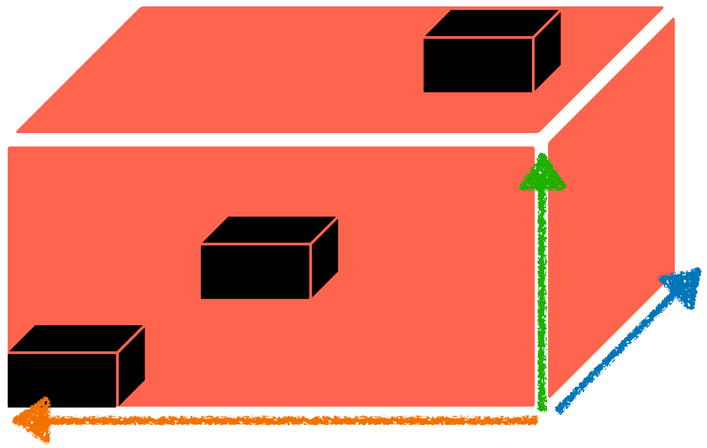
Repeat until no tuples are remaining

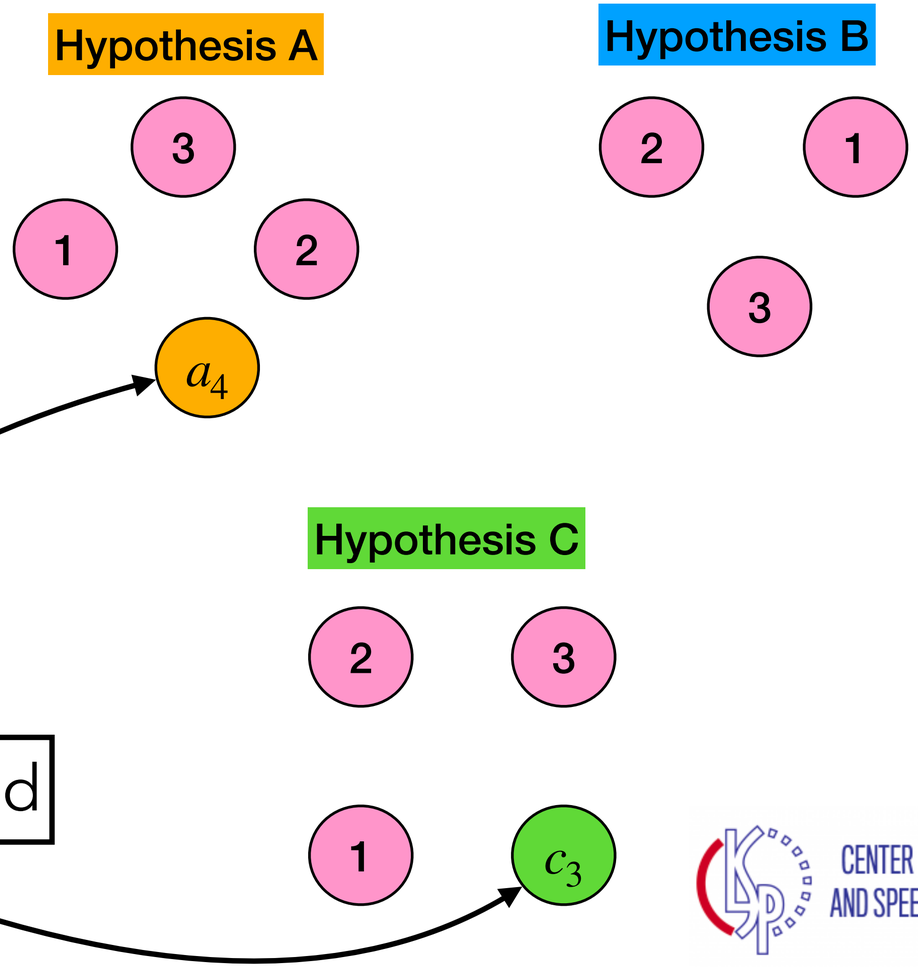




DOVER-Lap label mapping

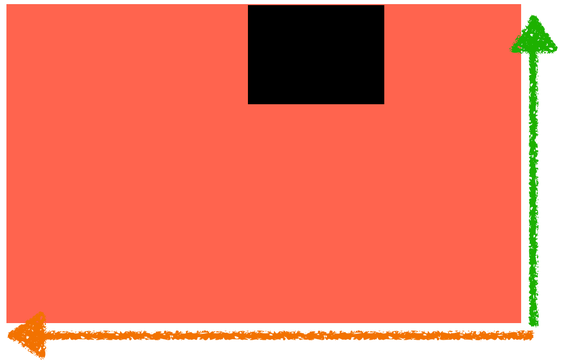
Repeat until no tuples are remaining

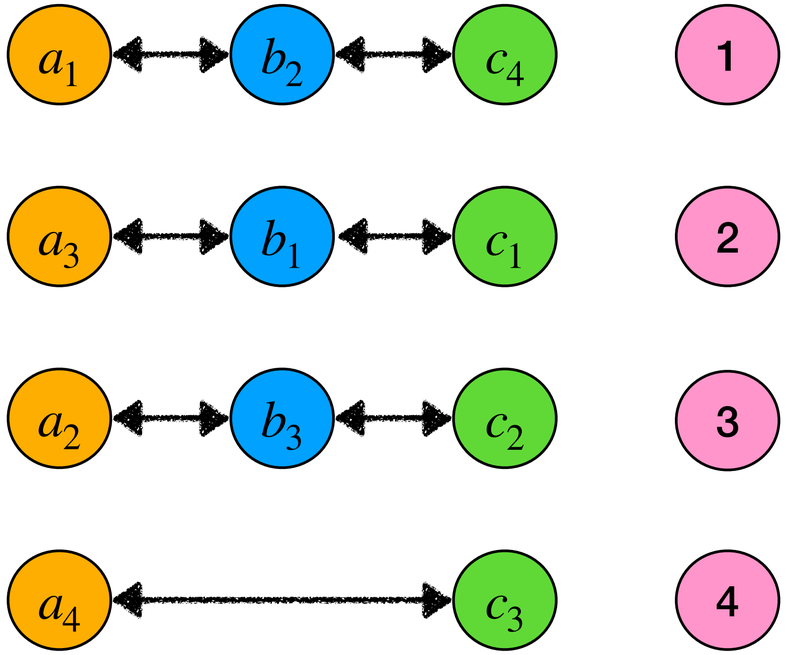




DOVER-Lap label mapping

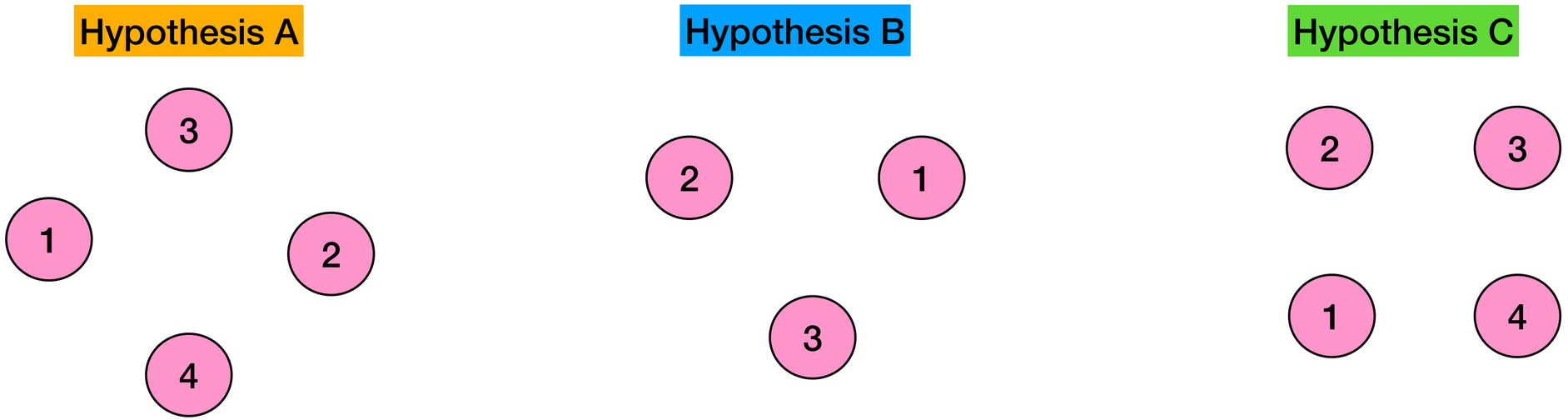
If no tuples remaining but labels left to be mapped, remove filled dimensions and repeat





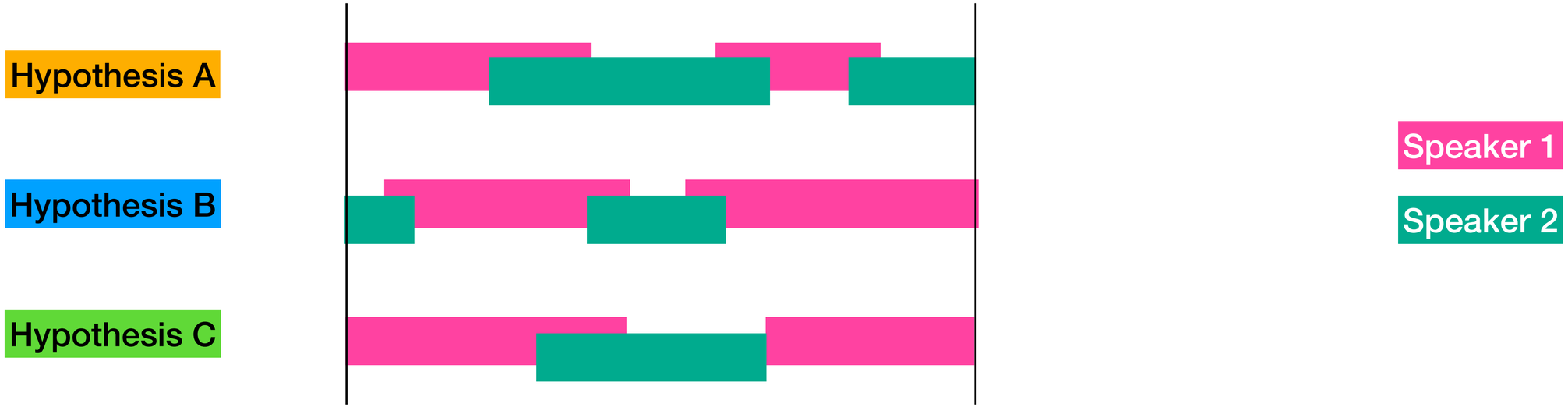
DOVER-Lap label mapping

Final mapped labels



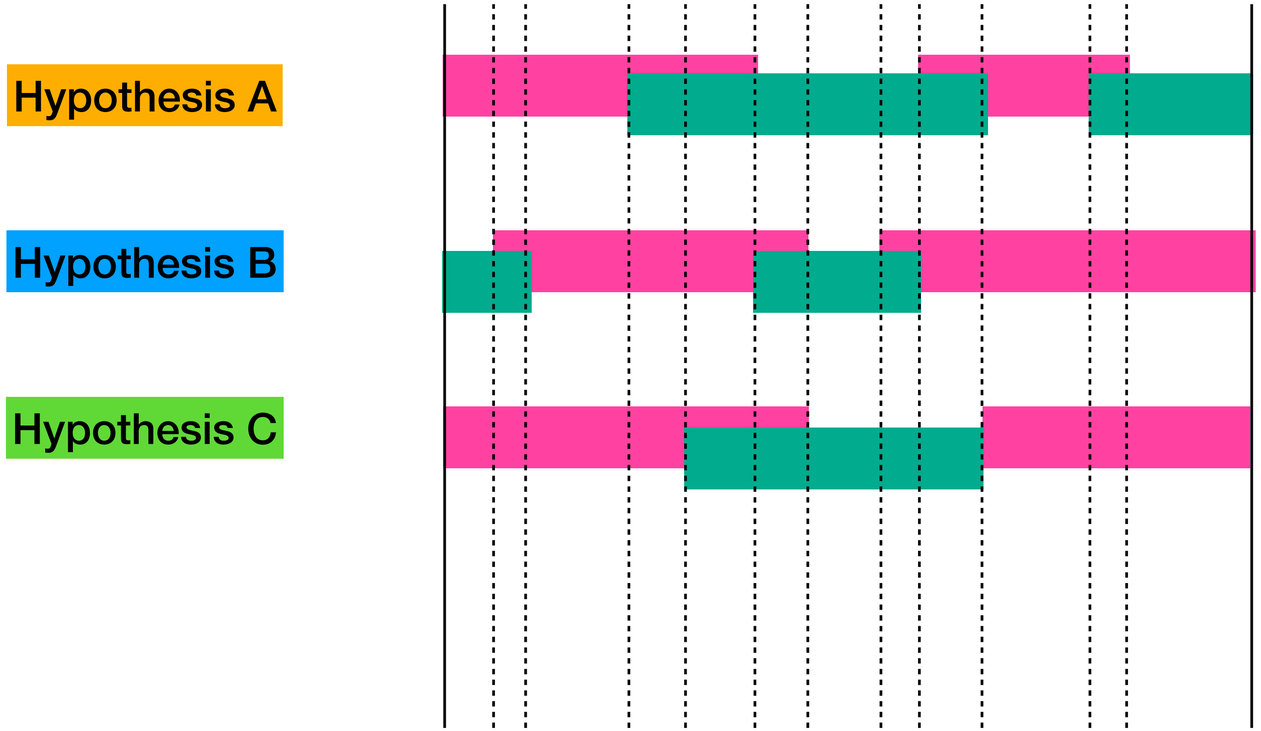
DOVER-Lap label voting

Consider 3 hypotheses from overlap-aware diarization systems



DOVER-Lap label voting

Divide into regions (similar to DOVER)



Speaker 1

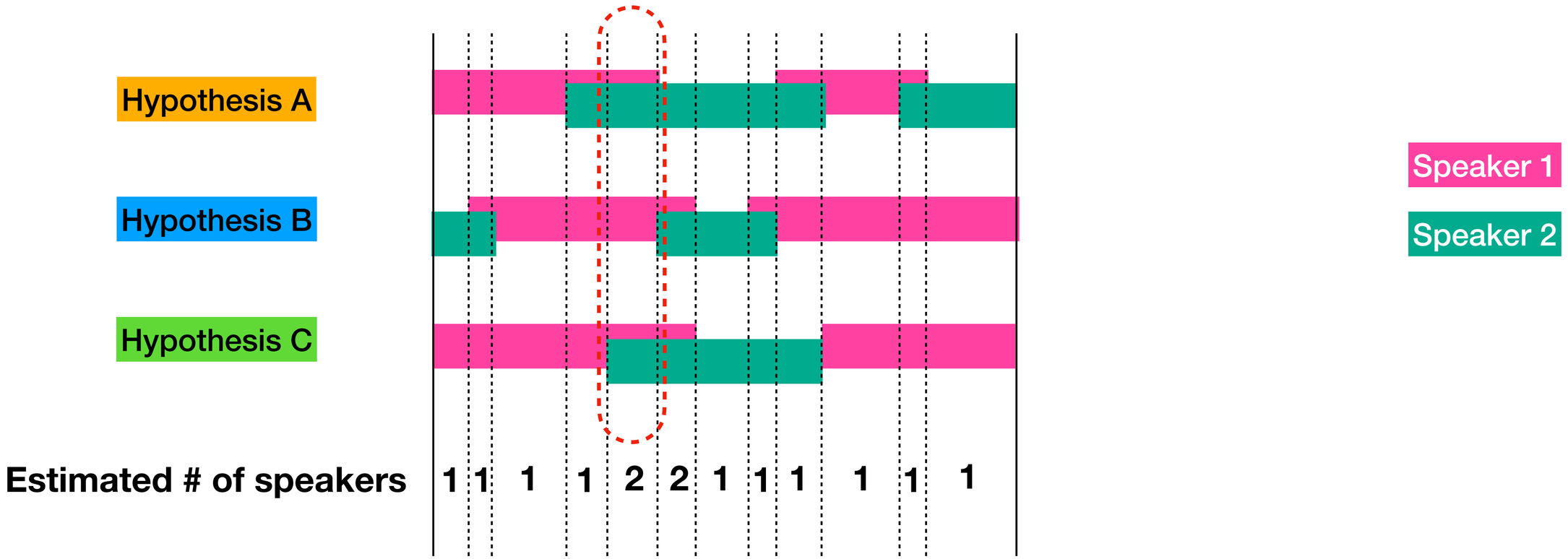
Speaker 2

SLT 2021

CENTER FOR LANGUAGE AND SPEECH PROCESSING

DOVER-Lap label voting

Estimate number of speakers in each region



 # speakers = weighted mean of # speakers in hypotheses

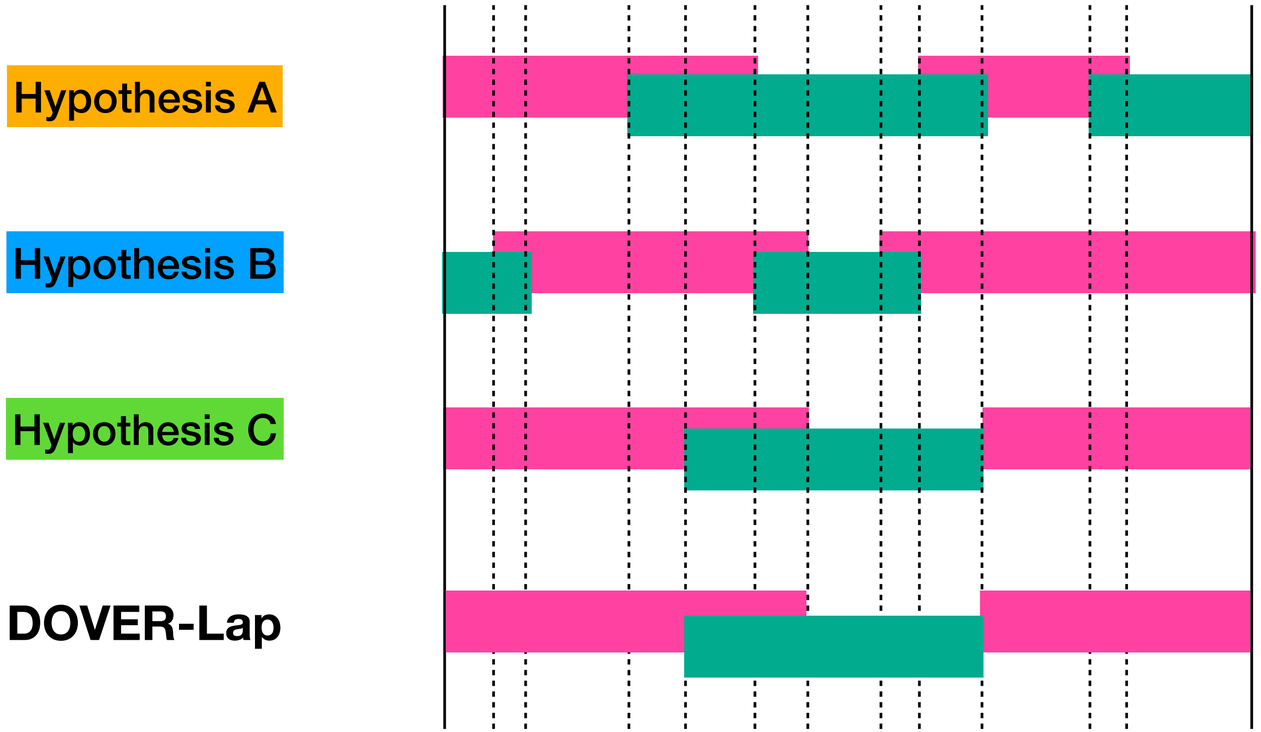
Weights -> obtained by ranking hypotheses by **total cost**

SLT 2021

CENTER FOR LANGUAGE AND SPEECH PROCESSING

DOVER-Lap label voting

Assign highest weighted N speakers in each region



Speaker 1

Speaker 2

SLT 2021

CENTER FOR LANGUAGE AND SPEECH PROCESSING

Overview

- **The DOVER-Lap algorithm**
  - Preliminary: DOVER
  - How to map labels to a common space?
  - Overlap-aware majority voting
- **Extended Results**
  - Effect of global label mapping
  - System combination results (AMI and LibriCSS)

Results: AMI dev

Effect of global label mapping algorithm

System	Spk. conf.	DER
Overlap-aware SC	12.8	24.5
VB-based overlap assignment	12.3	22.0
Region proposal network	29.8	35.3
Average	18.3	27.3
DOVER	20.4	36.5
+ global label mapping	7.7	26.0

SLT 2021

CENTER FOR LANGUAGE AND SPEECH PROCESSING

Results: AMI dev

Effect of rank-weighted majority voting

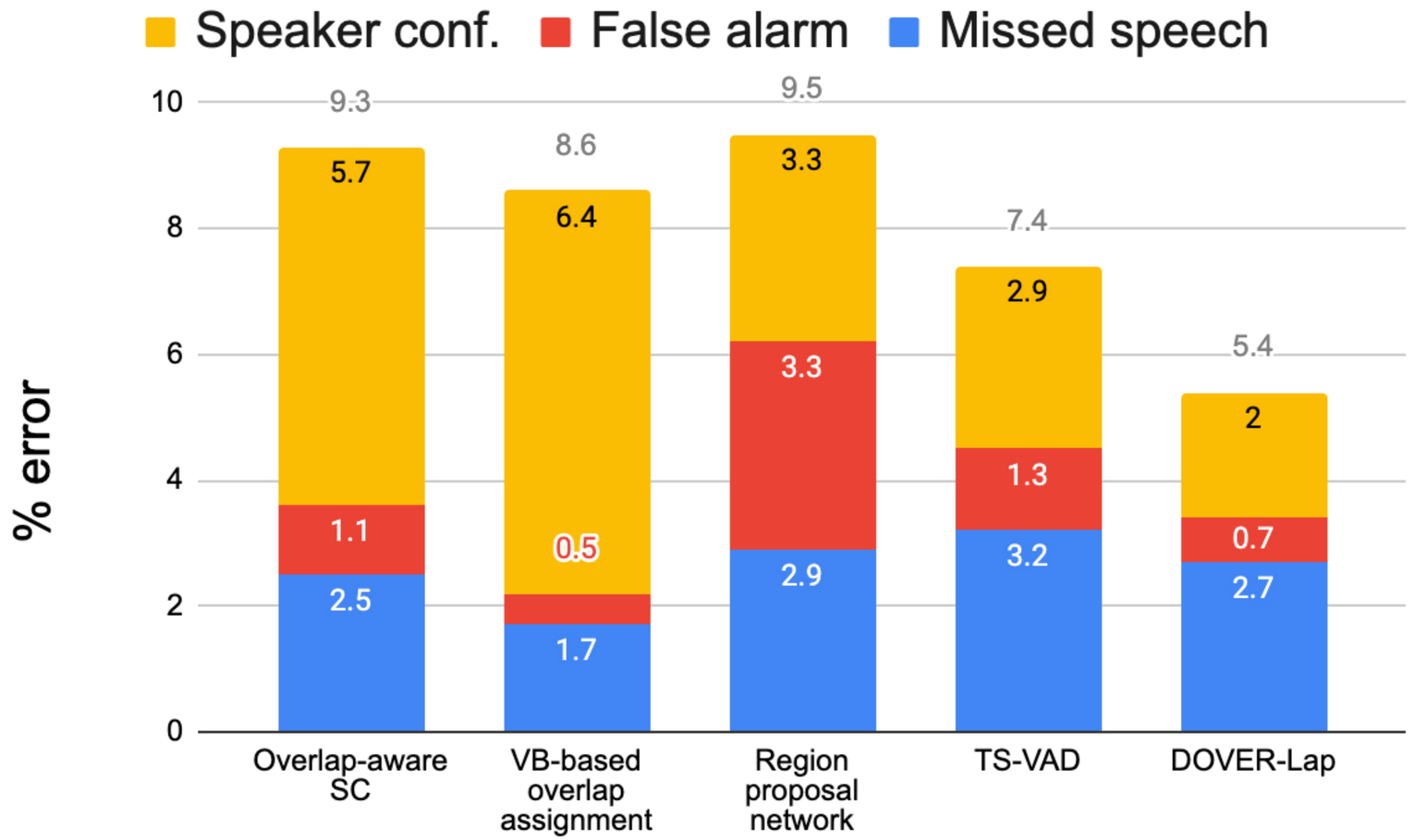
System	Spk. conf.	DER
Overlap-aware SC	12.8	24.5
VB-based overlap assignment	12.3	22.0
Region proposal network	29.8	35.3
Average	18.3	27.3
DOVER	20.4	36.5
+ global label mapping	7.7	26.0
DOVER-Lap	10.8	**21.6**

SLT 2021

CENTER FOR LANGUAGE AND SPEECH PROCESSING

Results: Breakdown on LibriCSS eval

Effectively combines complementary strengths



Limitations and Future Work

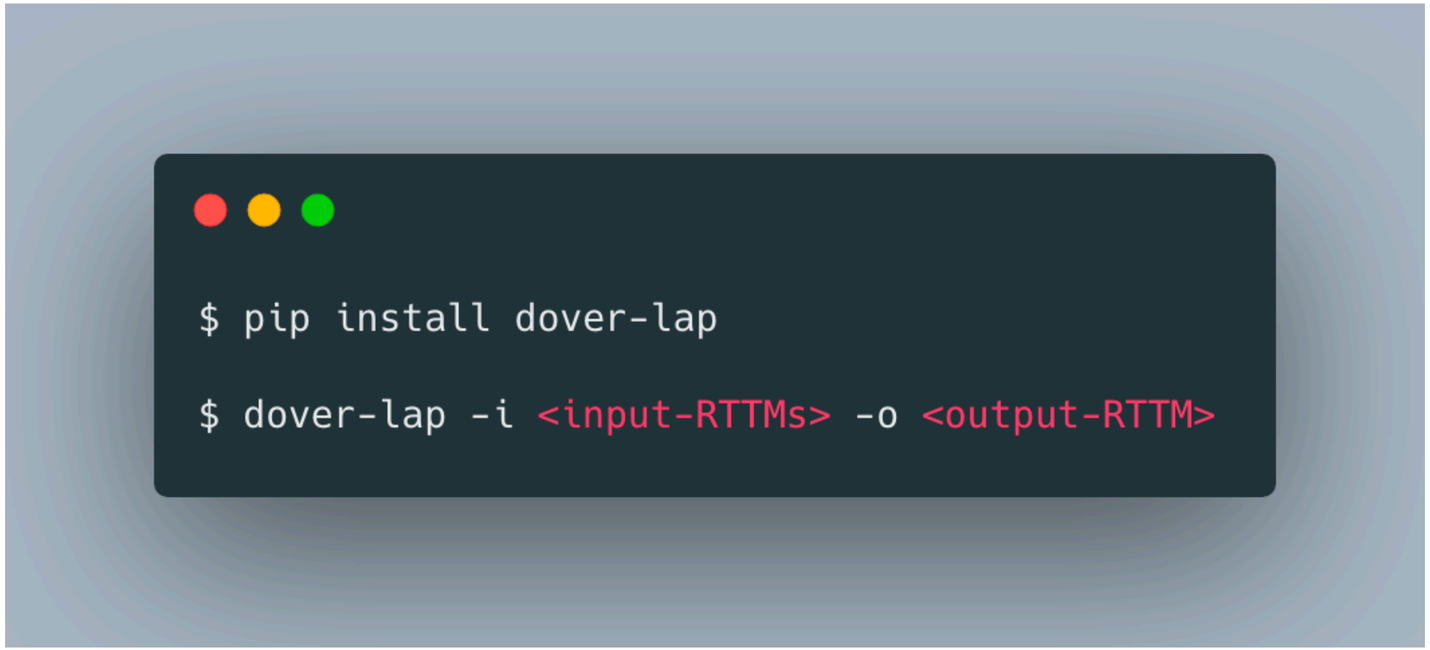
Cannot effectively combine mixed-type hypotheses (e.g. 2 with overlaps and 1 without) -> How to solve this problem?

Greedy search used to get maximal matching from cost tensor -> Can be improved?

SLT 2021

CENTER FOR LANGUAGE AND SPEECH PROCESSING

Try it out on your DIHARD systems!



https://github.com/desh2608/dover-lap

SLT 2021

CENTER FOR LANGUAGE AND SPEECH PROCESSING

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 for providing the TS-VAD diarization output on LibriCSS.

We also thank the anonymous reviewers for their useful feedback.