Explaining and building trust in machine learning: the local and subgroup perspectives

Eliana Pastor

Seminar @ sqIRL/IDLab, University of Antwerp

Hello!

I'm Eliana!

I'm an assistant professor at Politecnico di Torino, Italy

I work on Trustworthy and Explainable AI

My interests.. in keywords

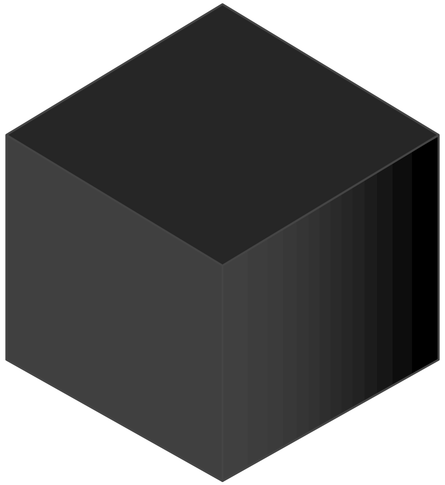
Trustworthy AI

Explainable AI, Fairness in AI, Robusteness, Debugging

- Analysis of disparities in data subgroups
- XAI for Speech & Sound
- Post-hoc XAI for tabular & text data
- KANs
- Concept-based XAI

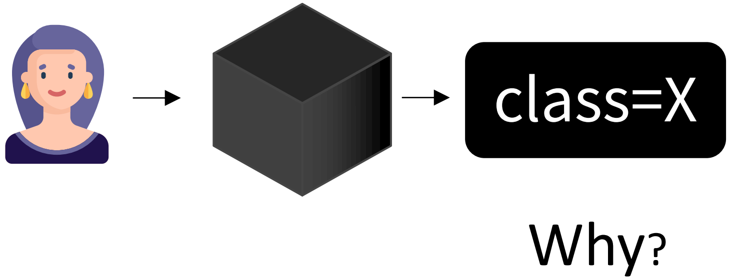
& other stuff

Our problem: open the box



From which perspective we open the black box

Local perspective



Subgroup perspective

Global perspective

Outline

Subgroup perspective

- Identification of subgroups with *divergent* classification behavior
- Divergent subgroup analysis in speech data
- Subgroup-based model comparison
- Mitigate subgroup disparities
- Interpretable subgroup drift detection

Outline

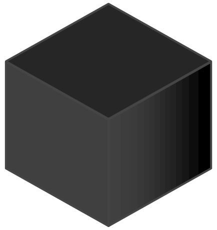
Local perspective

- Explaining prediction of speech models
- Assessing explainability methods for transformers models

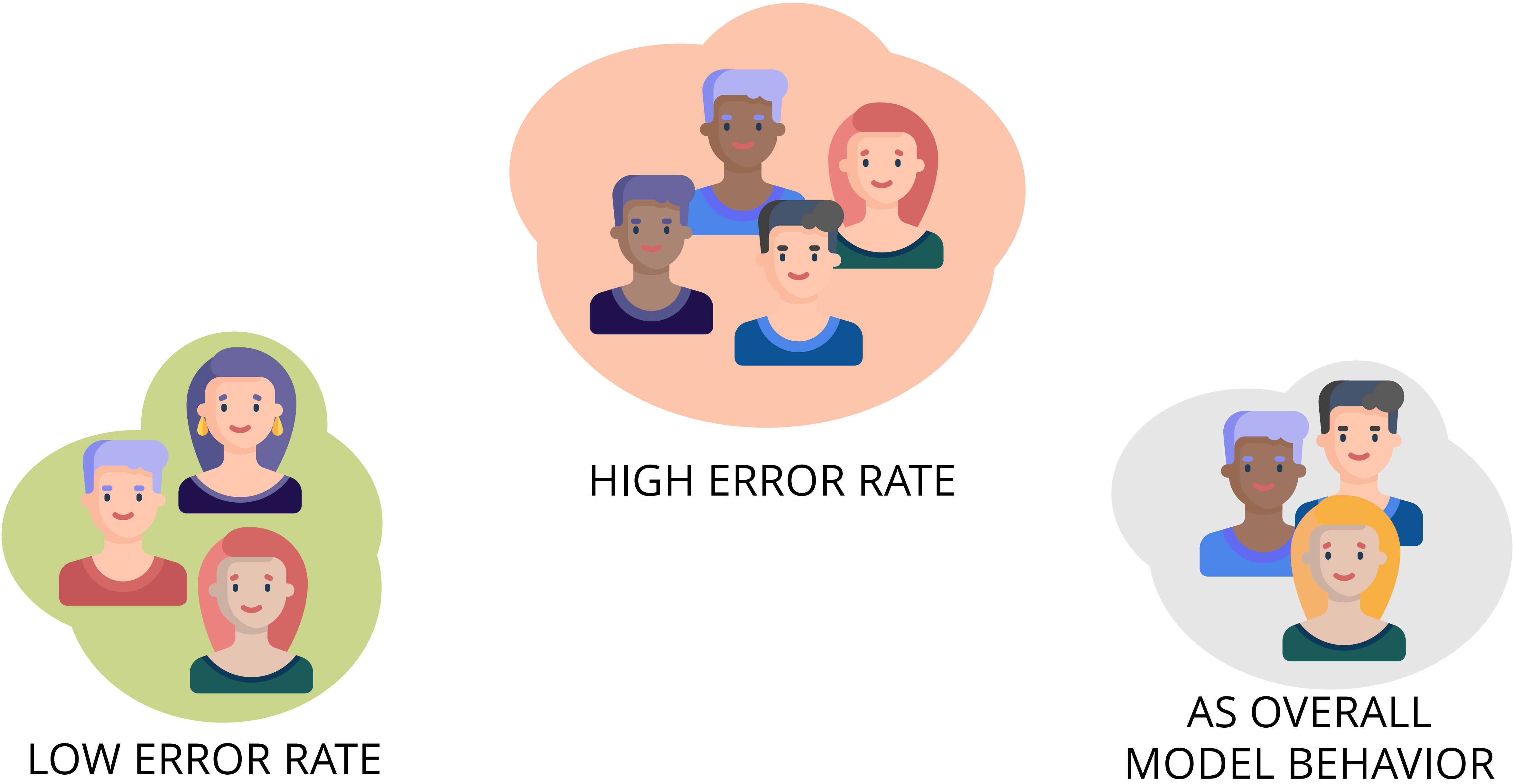
Outline

Subgroup perspective

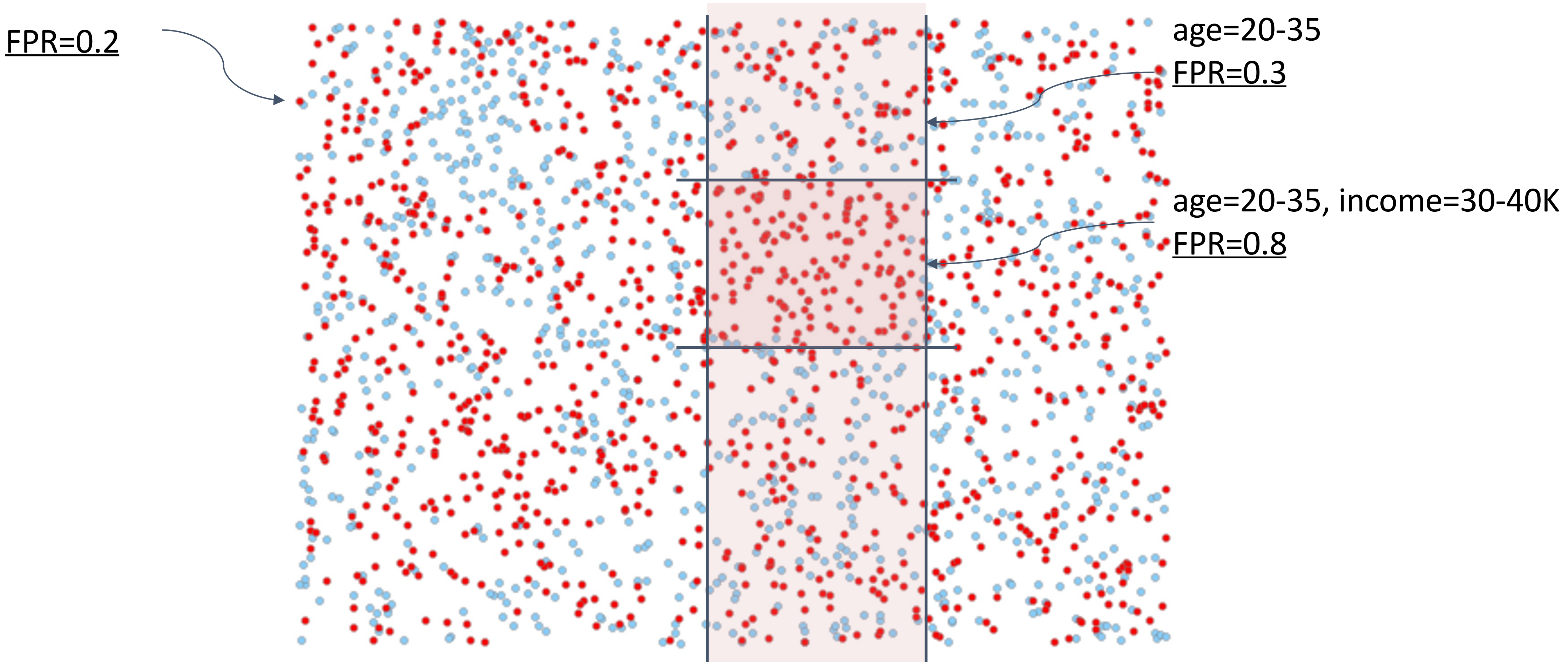
- **Identification of subgroups with *divergent* classification behavior**
- Divergent subgroup analysis in speech data
- Subgroup-based model comparison
- Mitigate subgroup disparities
- Interpretable subgroup drift detection



PERFORMANCE
X%


HIGH ERROR RATE
LOW ERROR RATE
AS OVERALL
MODEL BEHAVIOR

Motivation



Divergence of a subgroup

pattern, **interpretable**, e.g., {age=20-35, gender=female}

$$\Delta(S) = f(S) - f(D)$$

***performance* measure**

all dataset

Generic & model agnostic

- **Automatic** identification of subgroups via **frequent pattern mining**

Pastor, E, et al. " Looking for trouble: Analyzing classifier behavior via pattern divergence " ACM SIGMOD 2021

ProPublica

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Divergent subgroups - Example

COMPAS dataset. Recidivism predictions based on defendant information

Divergence Statistical significance

Subgroup

Itemset	Δ_{FPR}	support	t
age=25-45, #prior>3, race=Afr-Am, sex=Male	0.22	0.13	7.1
age=25-45, #prior>3, race=Afr-Am	0.211	0.15	7.4
age=25-45, charge=F, #prior>3, race=Afr-Am	0.202	0.11	6.2

Subgroup frequency

Contributions of items to divergence

Itemset	Δ_{FPR}
age=25-45, #prior>3, race=Afr-Am, sex=Male	0.22

What is the contribution
of each term?

Contributions of items to divergence

Team
&
Team score

↓

Subgroup I
&
Subgroup divergence

?

Contribution

↓

Contribution of an attribute=value (item) α

→

Shapley value

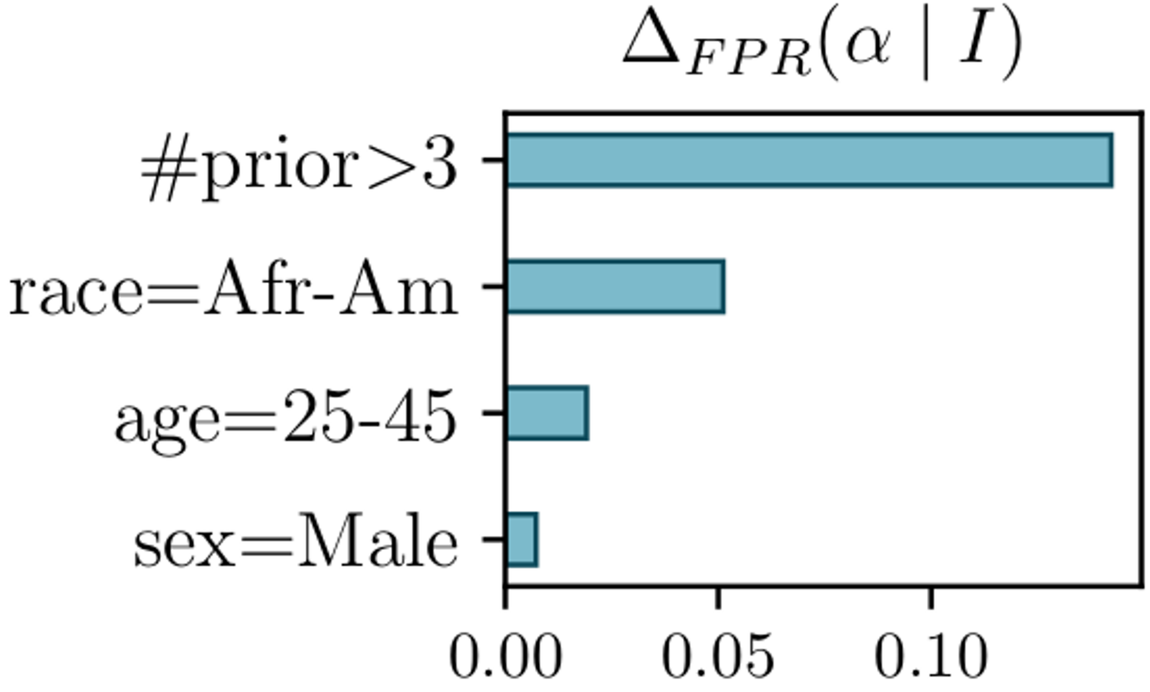
Given the score of subset of players

Contribution of item α in I:

$$\Delta(\alpha \mid I) = \sum_{J \subseteq I \setminus \{\alpha\}} \frac{|J|!(|I|-|J|-1)!}{|I|!} [\Delta(J \cup \alpha) - \Delta(J)]$$

Contributions of items to divergence

Itemset	Δ_{FPR}
age=25-45, #prior>3, race=Afr-Am, sex=Male	0.22



Divergent subgroups - Example

Itemset	Δ_{FPR}	support	t
age=25-45, #prior>3, race=Afr-Am, sex=Male	0.22	0.13	7.1
age=25-45, #prior>3, race=Afr-Am	0.211	0.15	7.4
age=25-45, charge=F, #prior>3, race=Afr-Am	0.202	0.11	6.2

Globally?

Global divergence

Global Shapley Value

A generalization of Shapley value that accounts for:

- Incompatible items (e.g. {*age<25, age>45*})
- Minimum support threshold

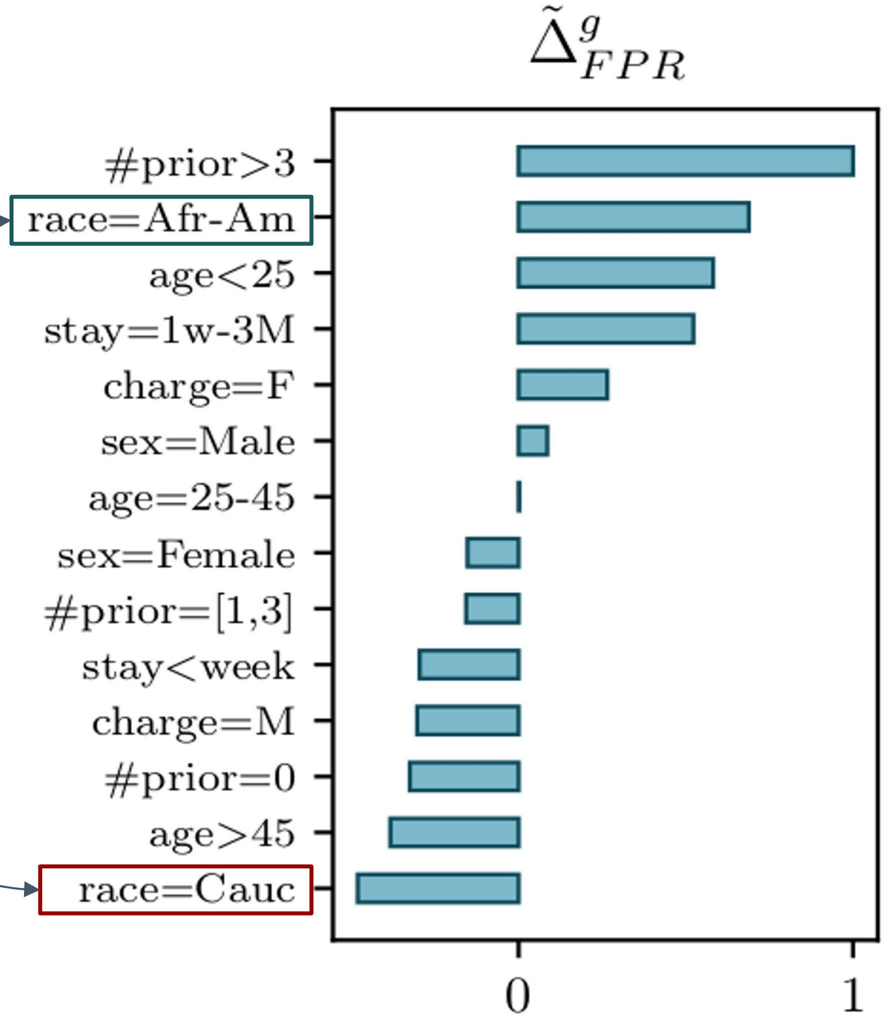
$$\widetilde{\Delta}^{g}(I,s) = \sum_{B \subseteq A \setminus \mathrm{attr}(I)} \frac{|B|!(|A|-|B|-|I|)!}{|A|! \prod_{b \in B \cup \mathrm{attr}(I)} m_b} \sum_{J : J \cup I \in \mathcal{I}^{\star}_{B \cup \mathrm{attr}(I)}} \left[\Delta(J \cup I) - \Delta(J)\right]$$

normalization factor, where m_b is # attribute values of b

set of frequent itemsets with attributes BUattr(I)

Global divergence - COMPAS

Role of ethnicity



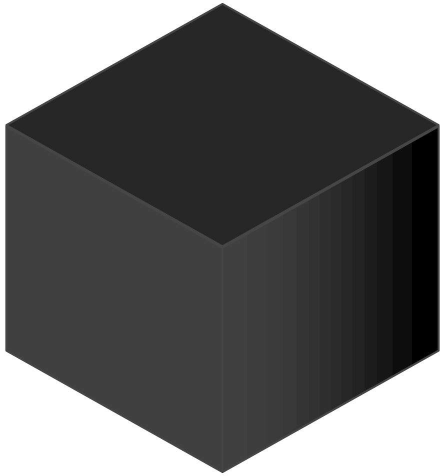
Outline

Subgroup perspective

- Identification of subgroups with *divergent* classification behavior
- **Divergent subgroup analysis in speech data**
- Subgroup-based model comparison
- Mitigate subgroup disparities
- Interpretable subgroup drift detection

Our scenario

Turn on the kitchen lights



Action: activate

Object: lights

Location: kitchen

Desidered properties of a subgroup

Interpretable

- e.g., lower performance for *young women*

Adequately represented

- Statistically and operational significant

Highlighting peculiar behavior

How to make an interpretable data grouping?

Clustering?

But… clusters of utterances are not directly interpretable

Enhance utterance with interpretable metadata

Speaker demographics

Speaking and recording conditions

Task- or dataset specific features

+

Metadata

gender=female
country=Italian
noise-level=high
speaking rate=fast
…

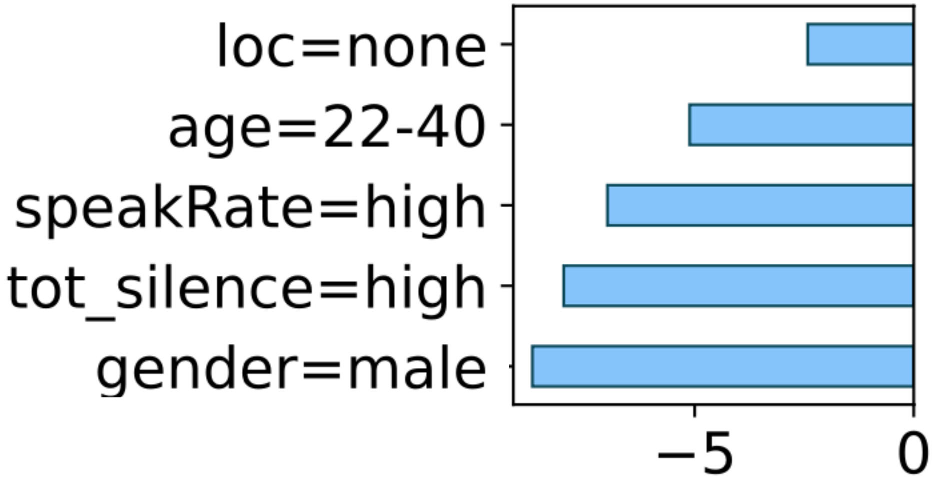
Koudounas, Alkis, et al. "Exploring subgroup performance in end-to-end speech models." IEEE ICASSP 2023

Divergent subgroup

By 31.22 less accurate!

	Subgroups	f	Δ_f
I^- :	$\boxed{\{\textit{age=22-40, gender=male, location=none, speaking rate=high, tot silence=high}\} \quad 60.50 \quad -31.22}$		
I^+ :	{*age=22-40, location=washroom, speaking rate=low, trimmed duration=high*}	100.0	8.28

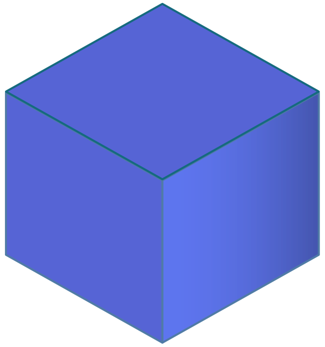
More accurate than average



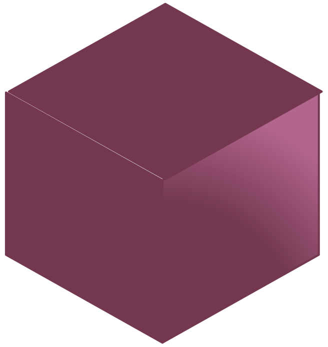
Outline

Subgroup perspective

- Identification of subgroups with *divergent* classification behavior
- Divergent subgroup analysis in speech data
- **Subgroup-based model comparison**
- Mitigate subgroup disparities
- Interpretable subgroup drift detection



Accuracy 91.72%



Accuracy 93.17%

Which model to choose?

.. most accurate..?

But on subgroups?

Inter-model performance gap

S = **pattern**, e.g., {age=20-35, gender=female}

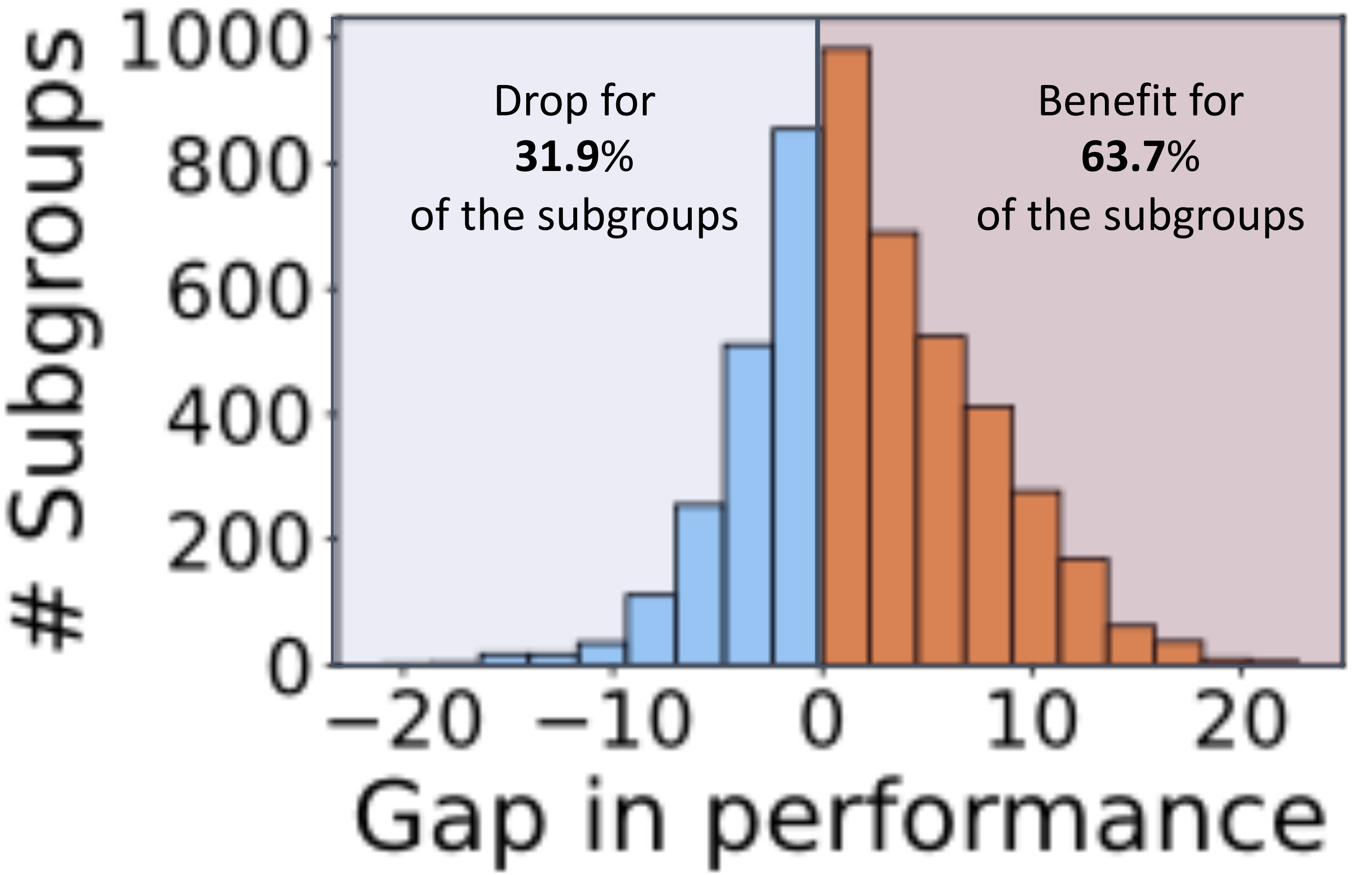
$$gap_f(S, M_1, M_2) = f(S, M_2) - f(S, M_1)$$

performance on S of model $\boldsymbol{M_2}$

performance on S of model $\boldsymbol{M_1}$

Koudounas, Alkis, et al. "Towards comprehensive subgroup performance analysis in speech models." ACM/ IEEE Transactions on Audio, Speech and Language Processing (2024).

Distribution of gain in performance



An example

Increase in performance

	Subgroups	gap_f	$f_{\text{w2v2-b}}$	$f_{\text{w2v2-l}}$
↑	{*action=increase, location=none, tot duration=low, trimmed speaking rate=low, trimmed duration=low*}	22.69	75.63	98.32
↓	{*action=activate, gender=male, speaking rate=low*}	-20.97	96.77	75.81

Drop in performance

Outline

Subgroup perspective

- Identification of subgroups with *divergent* classification behavior
- Divergent subgroup analysis in speech data
- Subgroup-based model comparison
- **Mitigate subgroup disparities**
- Interpretable subgroup drift detection

From identification to mitigation

Once we identify **divergent patterns**.. actively operate on **mitigation**

Post-processing

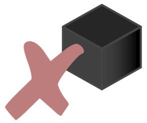
- Subgroup-guided data acquisition

In-processing

- Divergence regularization
- Subgroup-based contrastive loss
- Targeted data augmentation

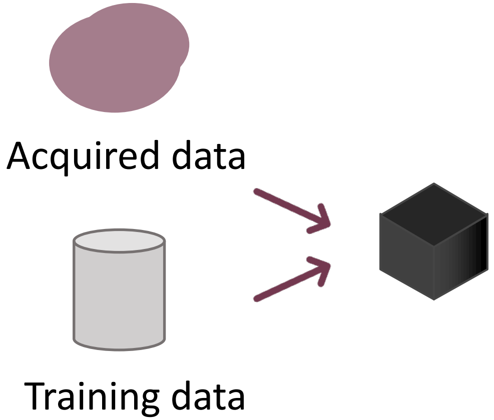
Post-processing
Subgroup-guided data acquisition

Speaking rate=high, gender=male



Step 1.
Identify the divergent patterns

Step 2.
Acquire data satisfying the patterns



Step 3.
Speech model re-training

Koudounas, Alkis, et al. "Prioritizing data acquisition for end-to-end speech model improvement." IEEE *ICASSP 2024*

In-processing
Divergence regularization

Add a divergence regularization term

$$\mathcal{L}_{\Delta} = \sum_{x_i \in D} \max_{S \in \mathbb{S}(x_i)} |\Delta(S)| \, \mathcal{L}_{CE}(y_i, \hat{y}_i)$$

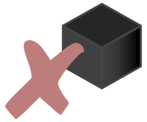
where $\mathbb{S}(x_i)$ is the set of subgroups satisfied by an instance x_i and $\mathcal{L}_{CE}$ is the cross-entropy loss,

Higher weight for samples with high-divergence

Koudounas, Alkis, et al. "Mitigating Subgroup Disparities in Speech Models: A Divergence-Aware Dual Strategy." IEEE Transactions on Audio, Speech and Language Processing (2025).

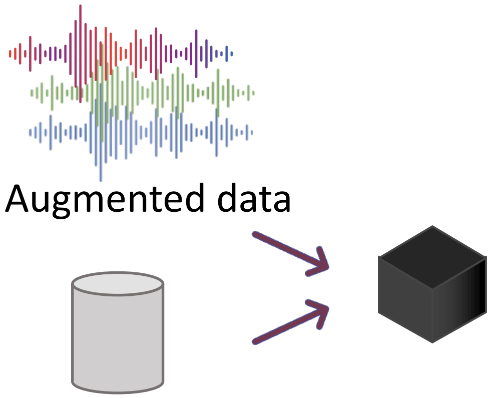
In-processing
Targeted data augmentation

Speaking rate=high, gender=male



Step 1.
Identify the divergent patterns

Step 2.
Data augmentation
Time stretching, background noise injection, reverberation, pitch shifting

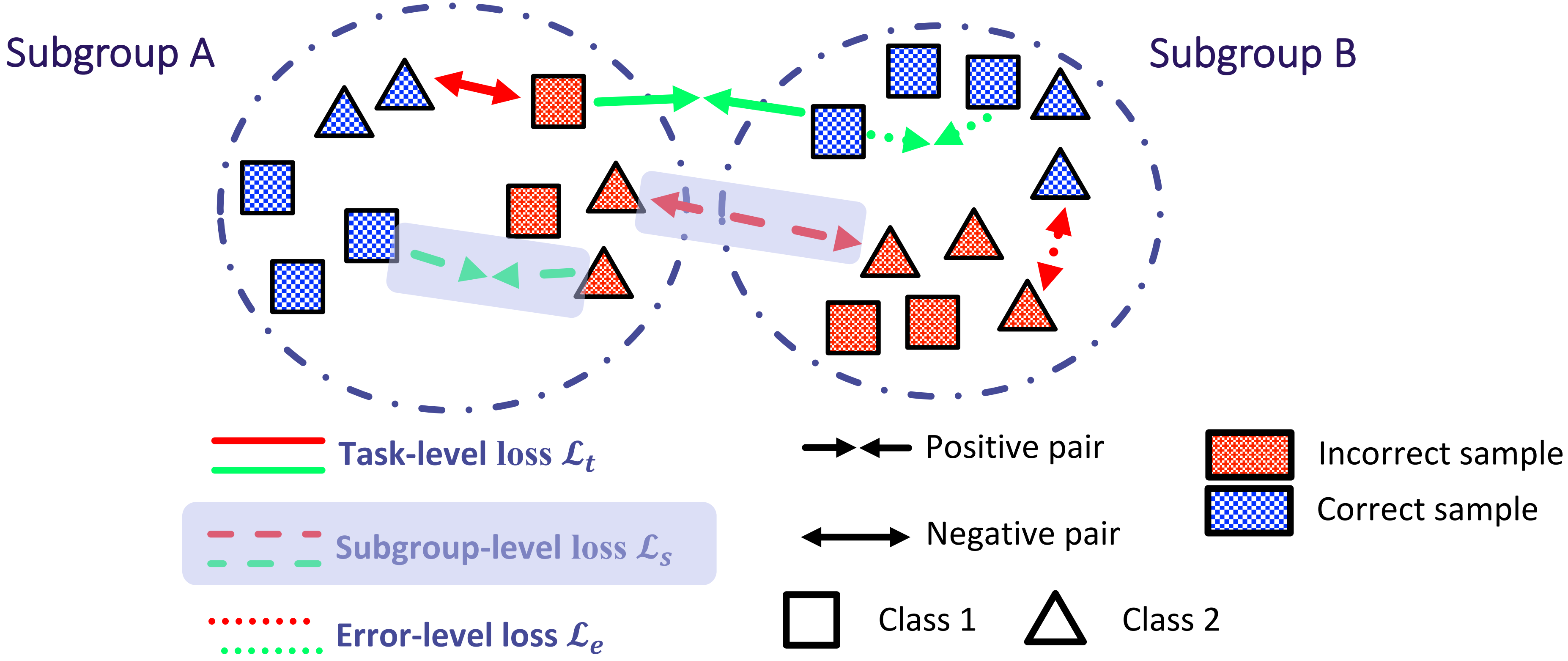


Step 3.
Speech model training

Koudounas, Alkis, et al. "Mitigating Subgroup Disparities in Speech Models: A Divergence-Aware Dual Strategy." IEEE Transactions on Audio, Speech and Language Processing (2025).

In-processing
Subgroup-based contrastive training

Three separate contrastive learning levels: **task**, **subgroup**, and **error**. At each level, we employ a multi-similarity (MS) loss to selectively contrast sample pairs based on their affinity



Koudounas, Alkis, et al. "A Contrastive Learning Approach to Mitigate Bias in Speech Models." Interspeech 2024 – Best student paper

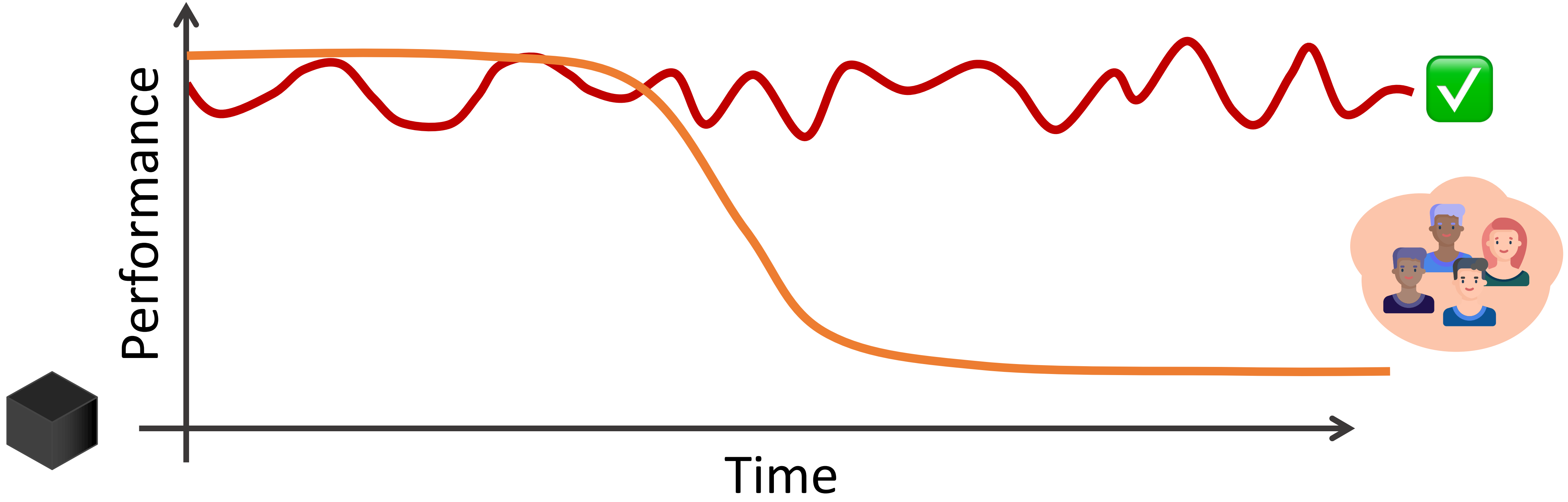
Outline

Subgroup perspective

- Identification of subgroups with *divergent* classification behavior
- Divergent subgroup analysis in speech data
- Subgroup-based model comparison
- Mitigate subgroup disparities
- **Interpretable subgroup drift detection**

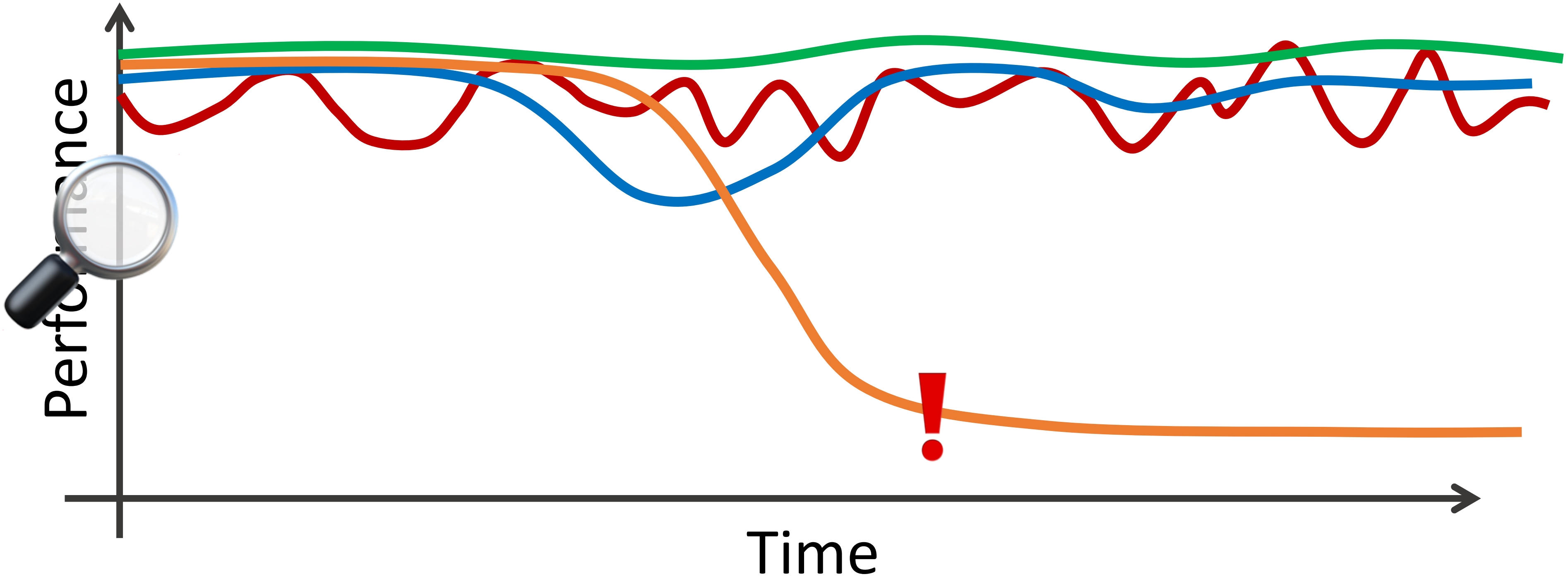
Subgroup based-drift detection

Typically, we monitor drift for overall performance



Subgroup based-drift detection

We propose an efficient algorithm to **monitorate subgroups** overtime **and detect subgroup drifts**



Giobergia, Flavio, et al. "Detecting Interpretable Subgroup Drifts." ACM KDD 2025 (to appear).

Outline

Subgroup perspective

- Identification of subgroups with *divergent* classification behavior
- Divergent subgroup analysis in speech data
- Subgroup-based model comparison
- Mitigate subgroup disparities
- Interpretable subgroup drift detection

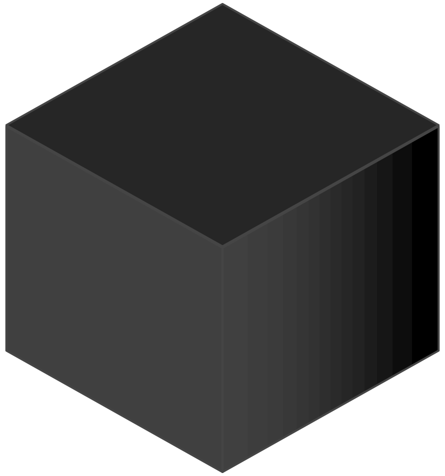
Outline

Local perspective

- **Explaining prediction of speech models**
- Assessing explainability methods for transformers models

Our scenario

Turn on the kitchen lights

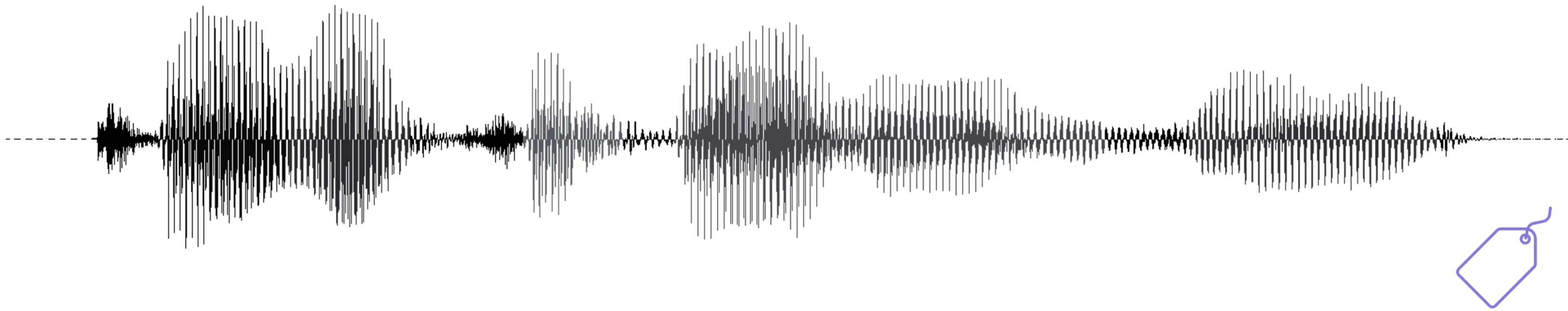


Action: activate

Object: lights

Location: kitchen

Why?



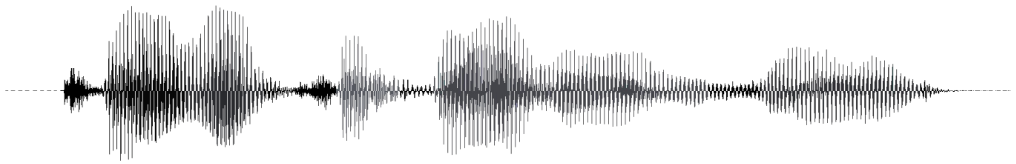
Explain the **interaction** between **utterance components** and **predictions** in a **human-understandable** manner

Pastor, Eliana, et al. " Explaining Speech Classification Models via Word-Level Audio Segments and Paralinguistic Features." EACL 2024

How do we define **interpretable representations** describing utterances?

Semantic

Spoken words

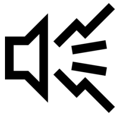


Turn up the bedroom heat

Paralinguistic

Prosody & external conditions

 Pitch

 Noise level

 Speaking rate

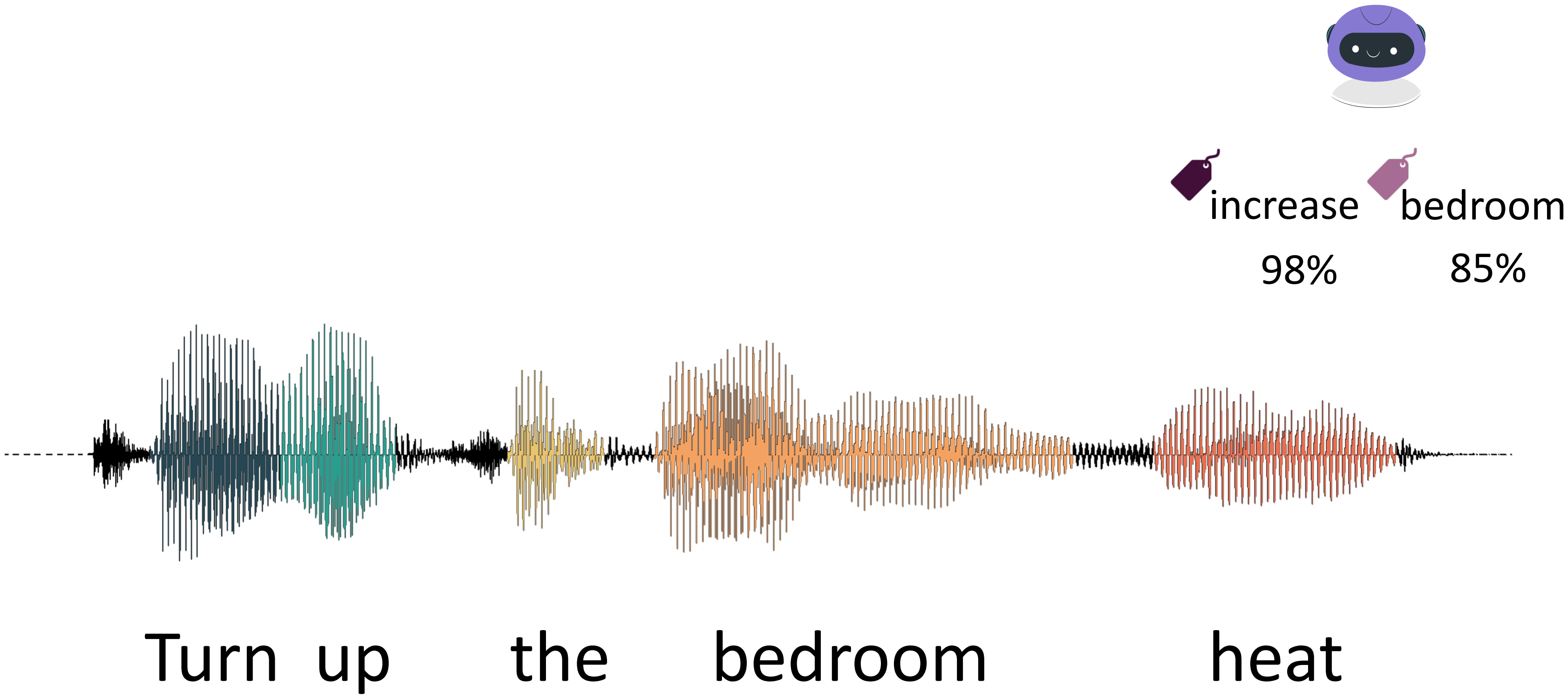
How do we **explain predictions** at the semantic and paralinguistic levels?

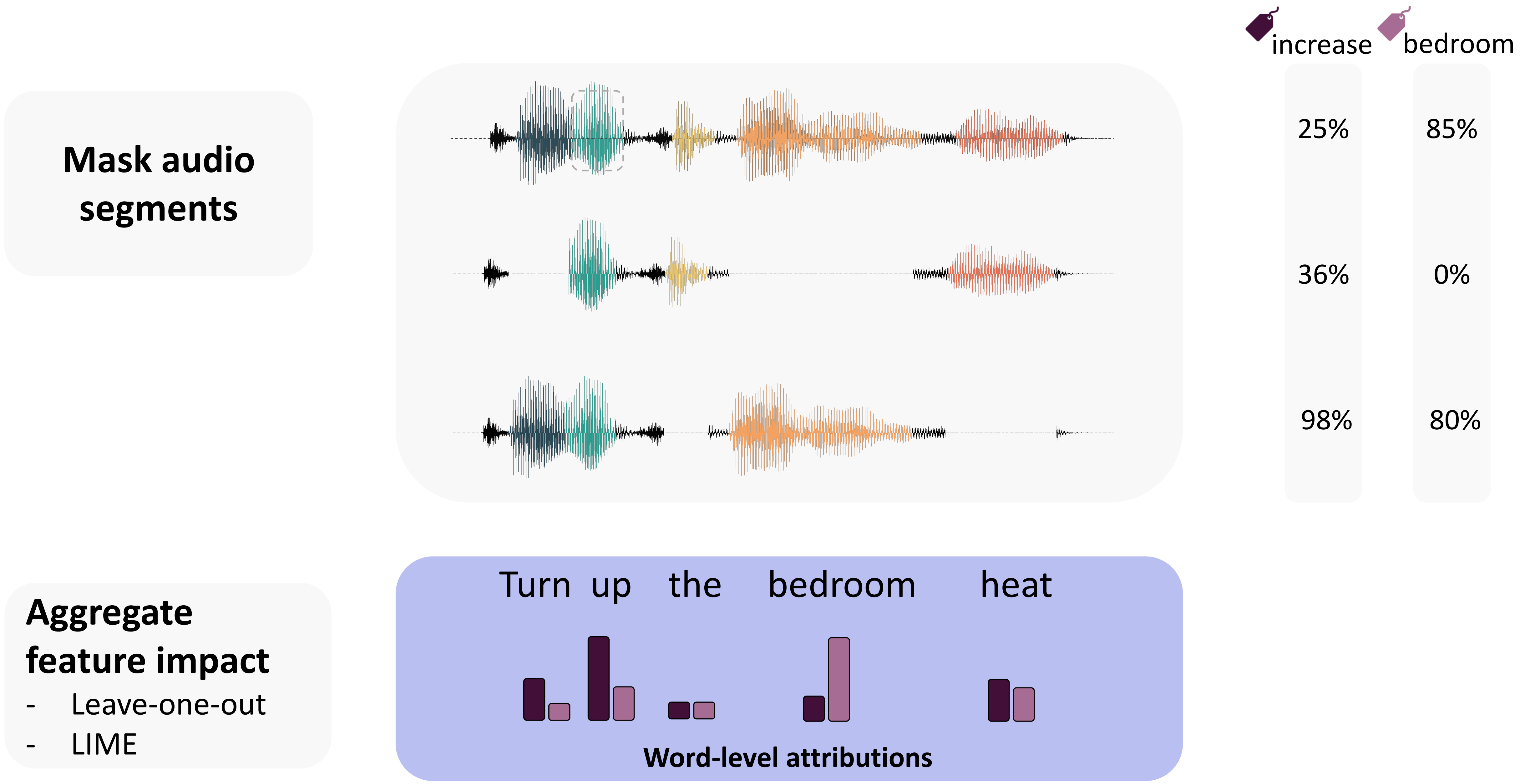
Perturbation-based approach

- Perturb the utterance based on an interpretable feature
- Measure the impact on predictions
- The greater the change, the more the model relies on this feature!

Semantic

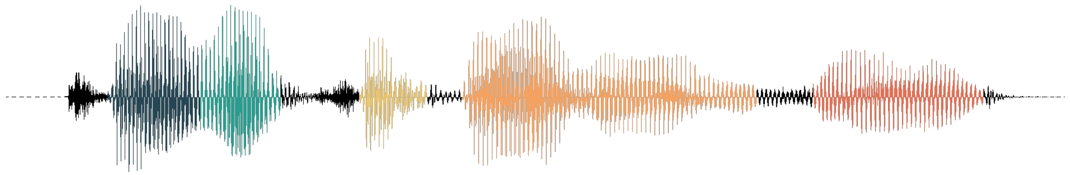
Use a word-level time alignment model





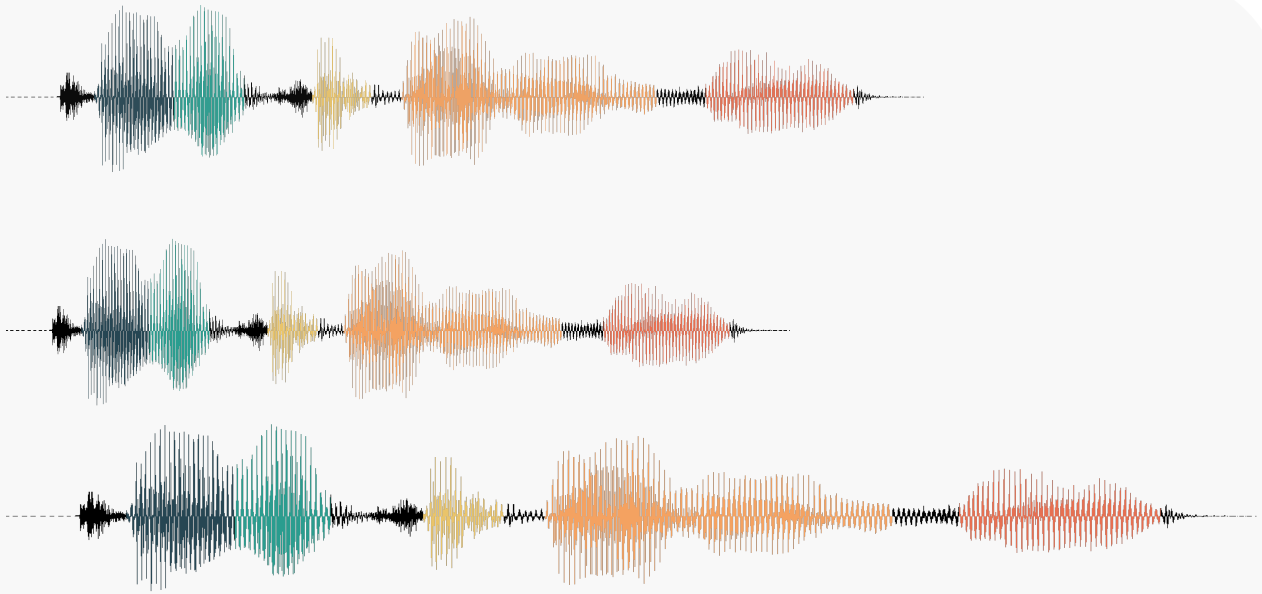
Mask audio segments
increase
bedroom
25%
85%
36%
0%
98%
80%
Aggregate feature impact
- Leave-one-out
- LIME
Turn up the bedroom heat
Word-level attributions

Paralinguistic



Paralinguistic

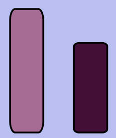
Perturb signal on paralinguistic



increase	bedroom
30%	85%
20%	70%
98%	90%

Aggregate feature impact

Time stretching



Speaking rate

Try it!

elianap/SpeechXAI

```
from speechxai import Benchmark
from transformers import Wav2Vec2ForSequenceClassification, Wav2Vec2FeatureExtractor


model = Wav2Vec2ForSequenceClassification.from_pretrained("superb/wav2vec2-base-superb-ic")

feature_extractor = Wav2Vec2FeatureExtractor.from_pretrained("superb/wav2vec2-base-superb-ic")


benchmark = Benchmark(model, feature_extractor)

explanation = benchmark.explain(audio_path=audio_path, methodology="LIME")

benchmark.show_table(explanation)
```

Outline

Local perspective

- Explaining prediction of speech models
- **Assessing explainability methods for transformers models**

Library to explain and benchmark explainers

We proposed **ferret**, Python **library for benchmarking interpretability** techniques on **Transformers** for text and speech data

F E R R E T

Attanasio, G, et al. " ferret: a Framework for Benchmarking Explainers on Transformers." EACL Demo 2023

What can you do with ferret?

Explaining individual prediction

Example. Sentiment classification – positive prediction

Token	__Great	__movie	__for	__a	__great	__nap	!
Partition SHAP	0.35	0.12	0.05	0.06	0.35	-0.00	0.05
LIME	-0.07	-0.08	0.03	-0.01	-0.24	0.17	0.06
Gradient	0.12	0.17	0.06	0.04	0.14	0.23	0.05
Gradient (x Input)	-0.11	-0.09	-0.08	0.03	0.03	0.11	-0.05
Integrated Gradient	-0.09	0.10	0.11	-0.02	0.10	0.02	-0.03
Integrated Gradient (x Input)	-0.09	-0.15	-0.17	-0.15	-0.10	-0.24	-0.10

What can you do with ferret?

Evaluate explanations

Faithfulness

How accurately the explanation reflects the inner working of the model

Plausibility

How explanations are aligned with human reasoning

Faithfulness

Token	_Great	_movie	_for	_a	_great	_nap	!
Partition SHAP	0.35	0.12	0.05	0.06	0.35	-0.00	0.05
LIME	-0.07	-0.08	0.03	-0.01	-0.24	0.17	0.06
Gradient	0.12	0.17	0.06	0.04	0.14	0.23	0.05
Gradient (x Input)	-0.11	-0.09	-0.08	0.03	0.03	0.11	-0.05
Integrated Gradient	-0.09	0.10	0.11	-0.02	0.10	0.02	-0.03
Integrated Gradient (x Input)	-0.09	-0.15	-0.17	-0.15	-0.10	-0.24	-0.10

	aopc_compr	aopc_suff	taucorr_loo
Partition SHAP	-0.13	-0.05	-0.33
LIME	-0.01	-0.20	0.24
Gradient	-0.16	-0.09	0.24
Gradient (x Input)	-0.00	-0.20	0.52
Integrated Gradient	-0.02	-0.17	0.33
Integrated Gradient (x Input)	0.00	1.00	-0.62

Plausibility

Human explanation

Token	_Great	_movie	_for	_a	_great	_nap	!
Partition SHAP	0.35	0.12	0.05	0.06	0.35	-0.00	0.05
LIME	-0.07	-0.08	0.03	-0.01	-0.24	0.17	0.06
Gradient	0.12	0.17	0.06	0.04	0.14	0.23	0.05
Gradient (x Input)	-0.11	-0.09	-0.08	0.03	0.03	0.11	-0.05
Integrated Gradient	-0.09	0.10	0.11	-0.02	0.10	0.02	-0.03
Integrated Gradient (x Input)	-0.09	-0.15	-0.17	-0.15	-0.10	-0.24	-0.10

	auprc_plau	token_f1_plau	token_iou_plau
Partition SHAP	1.00	0.50	0.33
LIME	0.14	0.00	0.00
Gradient	0.29	0.44	0.29
Gradient (x Input)	0.24	0.40	0.25
Integrated Gradient	0.22	0.33	0.20
Integrated Gradient (x Input)	0.64	0.00	0.00

Try ferret!

ferret-xai 0.4.2

```
pip install ferret-xai
```

```python
from transformers import AutoModelForSequenceClassification, AutoTokenizer
from ferret import Benchmark

name = "cardiffnlp/twitter-xlm-roberta-base-sentiment"
model = AutoModelForSequenceClassification.from_pretrained(name)
tokenizer = AutoTokenizer.from_pretrained(name)

bench = Benchmark(model, tokenizer)
explanations = bench.explain("You look stunning!", target=1)
evaluations = bench.evaluate_explanations(explanations, target=1)

bench.show_evaluation_table(evaluations)
```

Outline

Local perspective

- Explaining prediction of speech models
- Assessing explainability methods for transformers models

Joint works with

Alkis Koudounas

Elena Baralis

Luca de Alfaro

Flavio Giobergia

and all the other collegues
& collaborators!

elianap.github.io/

eliana.pastor@polito.it

eliana__pastor

@elianapastor.bsky.social

eliana-pastor