Multimodal Data Analysis

Lecture for ISEA Training Program

Wei Ai <aiwei@umd.edu>
University of Maryland

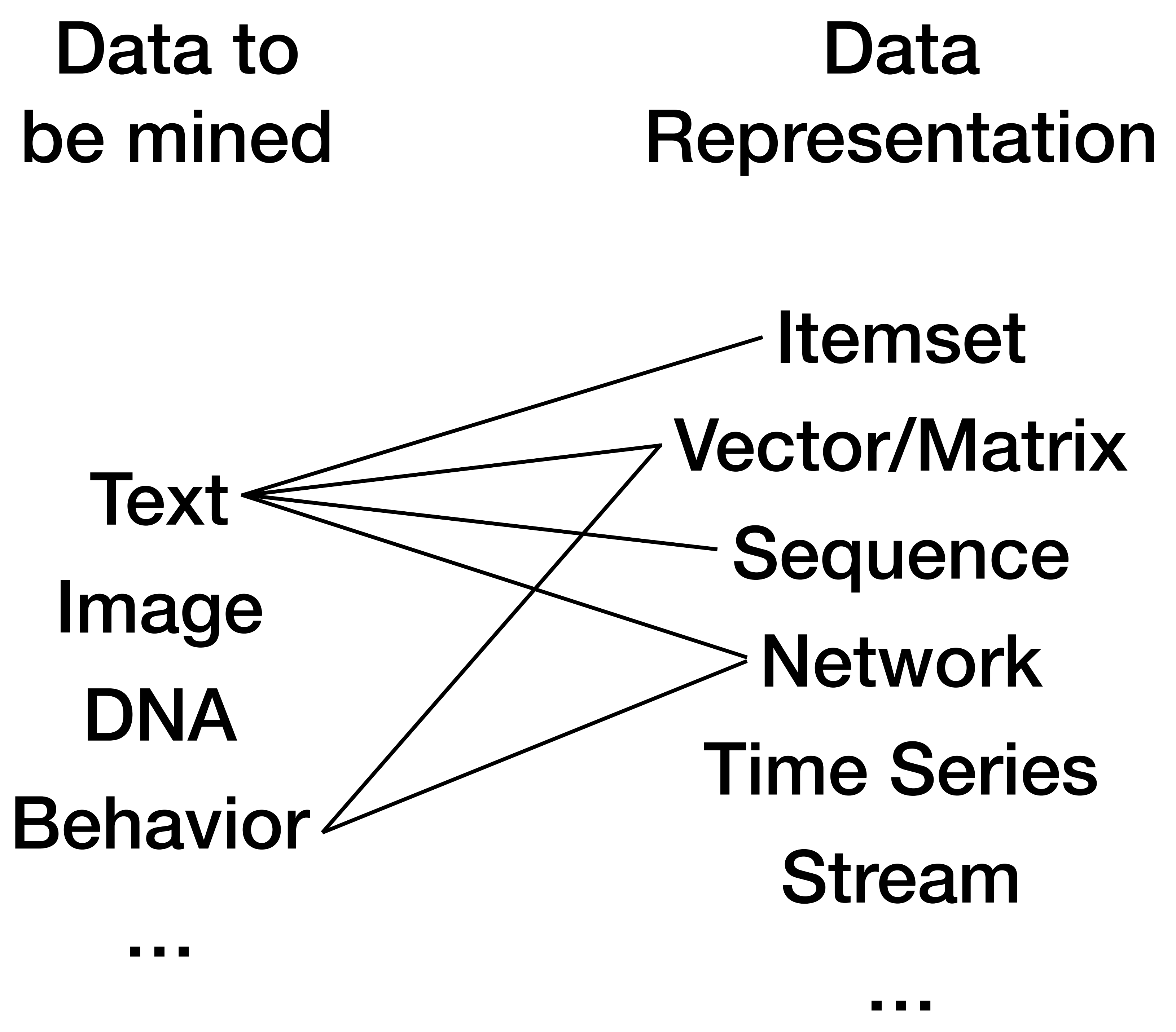
Let’s take a step back…

- What do data mining looks like

Data Mining: Knowledge Discovery from Data

Data to be mined	Data Representation	Basic Functionalities	Knowledges (Outcome)	Applications

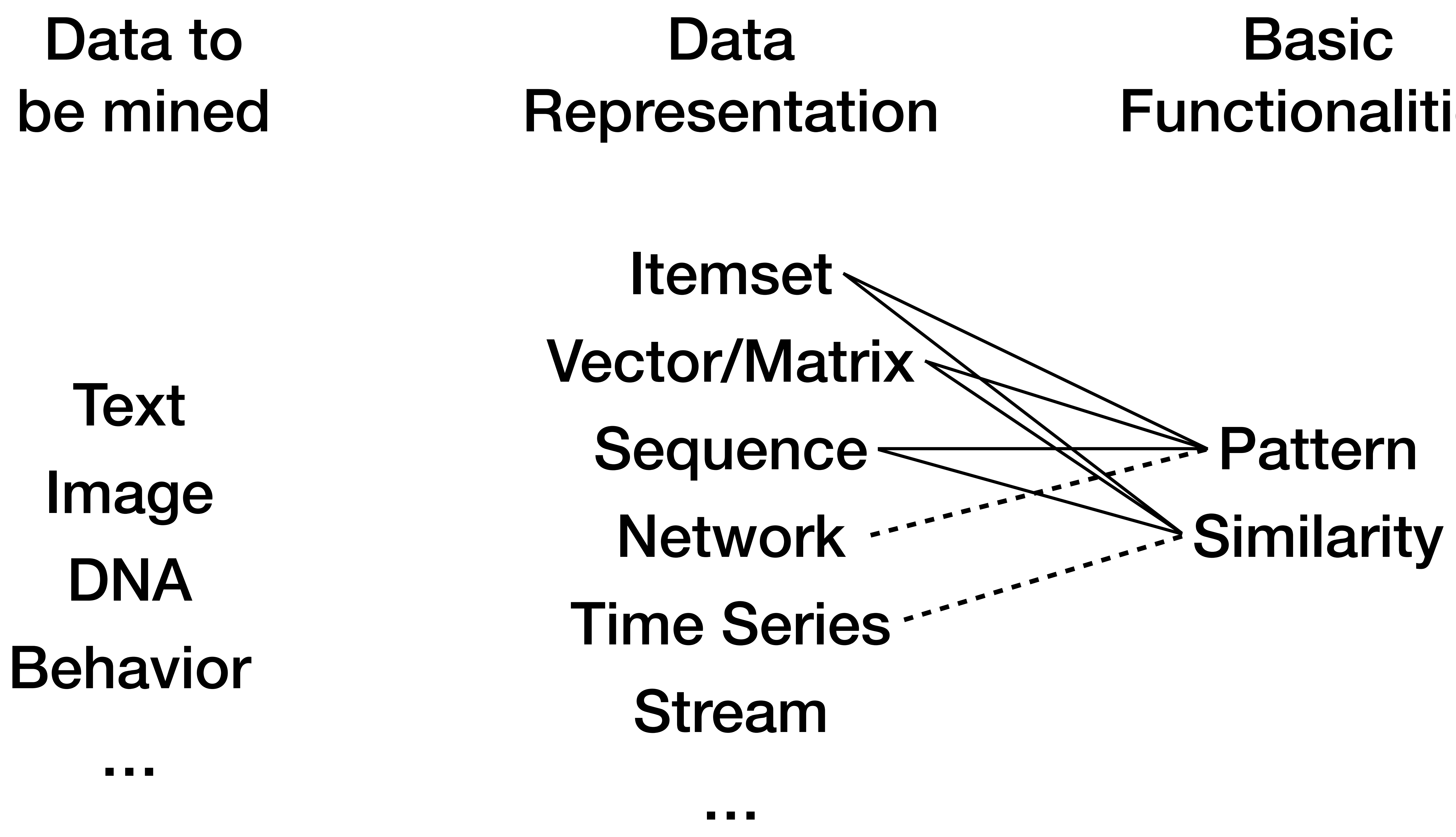
Data Mining: Knowledge Discovery from Data



- Text -> Itemset:
  - {“to”, “be”, “or”, “not”, …}
- Text -> Vector:
  - <1, 0, 0, 1>
- Behavior -> Vector:
  - User-Product Rating Matrix
- Behavior -> Network:
  - Twitter “Following” network

- Questions to keep in mind:
  - What is the granularity of the analysis
  - What counts as an observation
  - What does each “row” represent in your DataFrame

Data Mining: Knowledge Discovery from Data

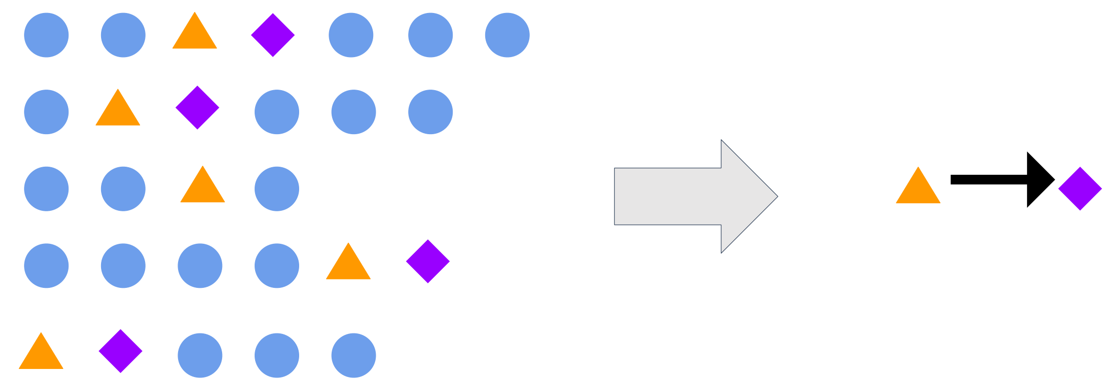


- Itemset + Patterns:
  - Frequent Pattern Mining
- Itemset + Similarity:
  - Jaccard Similarity
- Vector + Similarity:
  - Dot Product
  - Manhattan/Euclidean distance
  - Cosine Similarity
- Sequence + Similarity:
  - Edit distance
  - Shingling
- Network + Pattern:
  - ...

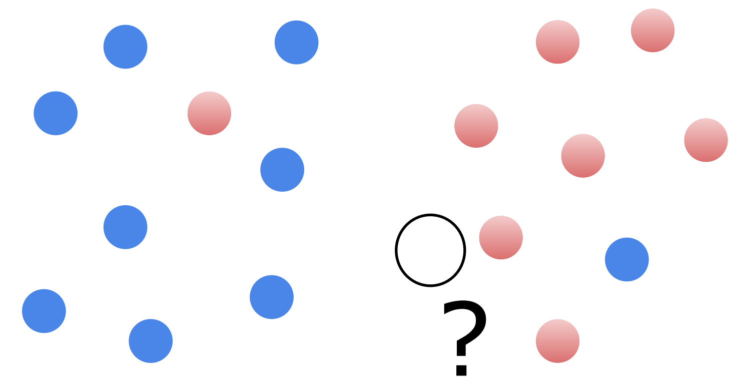
Data Mining: Knowledge Discovery from Data

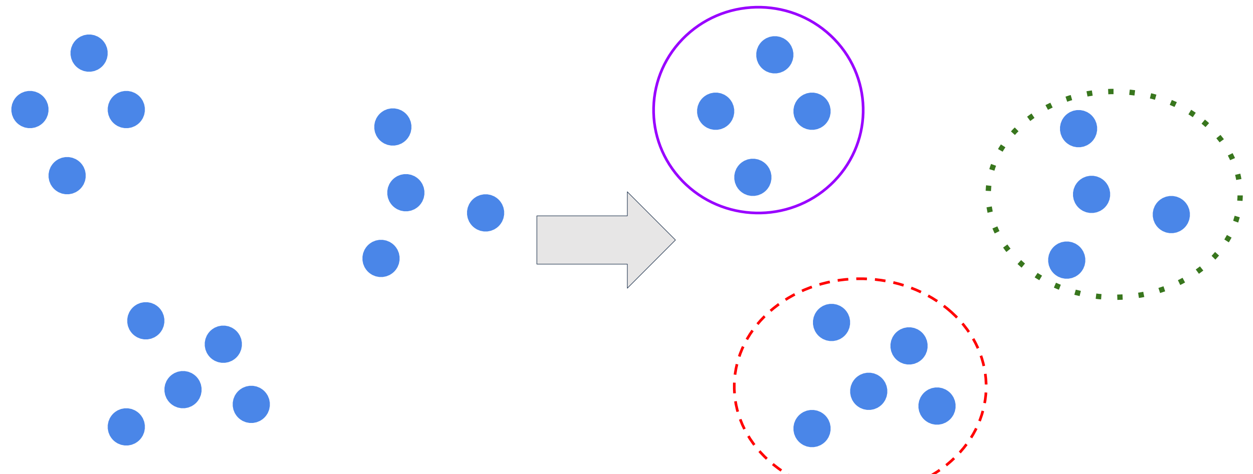
Data to be mined	Data Representation	Basic Functionalities	Knowledges (Outcome)
Text	Itemset	Pattern	Association
Image	Vector/Matrix	Similarity	Classification
DNA	Sequence		Clustering
Behavior	Network		Ranking
…	Time Series		Outlier Detection
	Stream		…
	…		

Association

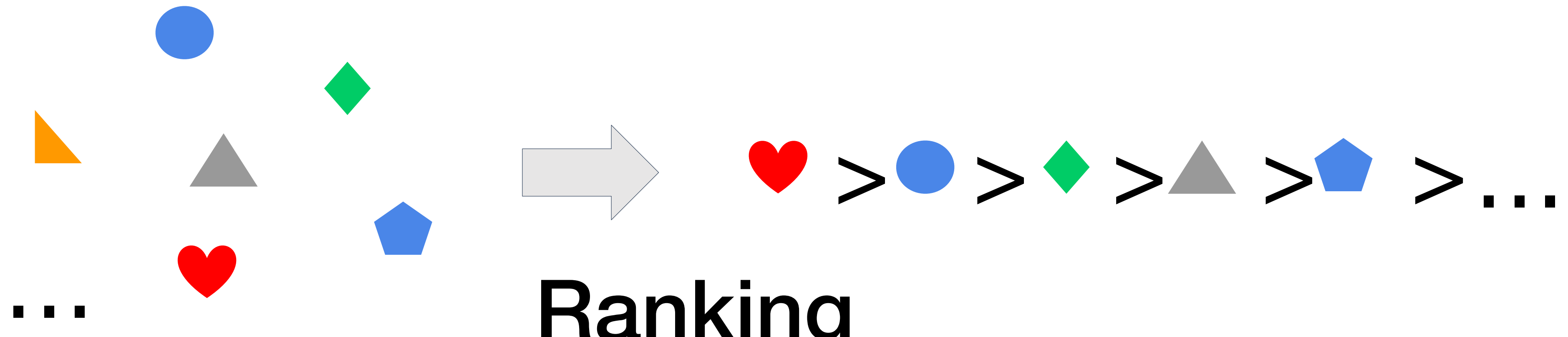


Classification





Clustering



Ranking

Data Mining: Knowledge Discovery from Data

Data to be mined	Data Representation	Basic Functionalities	Knowledges (Outcome)	Applications
Text	Itemset	Pattern	Association	Recommender System
Image	Vector/Matrix	Similarity	Classification	Fraud Detection
DNA	Sequence		Clustering	Web Search
Behavior	Network		Ranking	…
…	Time Series		Outlier Detection	
	Stream		…	
	…			

Example: Amazon “Frequently Bought Together”

Data to be mined

Text
Image
DNA
Behavior
…

Data Representation

Itemset
Vector/Matrix
Sequence
Network
Time Series
Stream
…

Basic Functionalities

Pattern
Similarity

Knowledges (Outcome)

Association
Classification
Clustering
Ranking
Outlier Detection
…

Applications

Recommender System
Fraud Detection
Web Search
…

Example: Netflix Movie Recommender System

Data to be mined

Text
Image
DNA
Behavior
...

Data Representation

Itemset
Vector/Matrix
Sequence
Network
Time Series
Stream
...

Basic Functionalities

Pattern
Similarity

Knowledges (Outcome)

Association
Classification
Clustering
Ranking
Outlier Detection
...

Applications

Recommender System
Fraud Detection
Web Search
...

Example: Text Retrieval

Data to be mined	Data Representation	Basic Functionalities	Knowledges (Outcome)	Applications
Text	Itemset	Pattern	Association	Recommender System
Image	Vector/Matrix	Similarity	Classification	Fraud Detection
DNA	Sequence		Clustering	Web Search
Behavior	Network		Ranking	…
…	Time Series		Outlier Detection	
	Stream		…	
	…			

Multimodal Data are Just Another Data Genre

Data to be mined

- Text
- Image
- DNA
- Behavior
- …
- Audio
- Video

Data Representation

- Itemset
- Vector/Matrix
- Sequence
- Network
- Time Series
- Stream
- …

Basic Functionalities

- Pattern
- Similarity

Knowledges (Outcome)

- Association
- Classification
- Clustering
- Ranking
- Outlier Detection
- …

Applications

- Recommender System
- Fraud Detection
- Web Search
- …

Multimodal Data are Just Another Data Genre

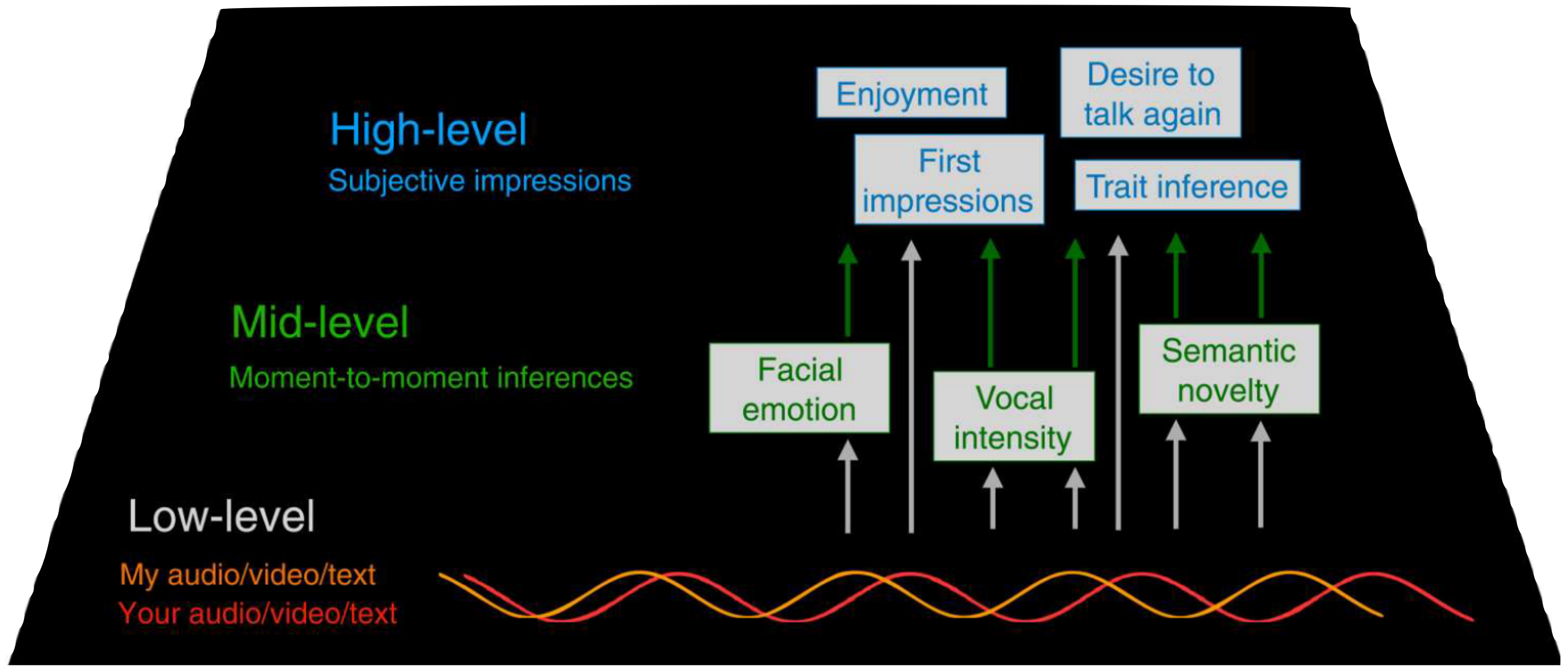
- We can still abstract multimodal data as the data representation that we are familiar with.
- It allows us to build features bottom-up, and can be naturally followed with
  - Exploratory analysis
  - Visualization
  - Correlation/prediction analysis
- For example:
  - Extract the *loudness* of one's speech at every 0.02 second, and make it a *time series*.
  - Classify one's *facial expression* with emotion labels, and *vectorize* its distribution.
  - Identify *nodding yes* and *shaking no* (head movement) as an event sequence.

The CANDOR Dataset

- 1656 unscripted conversations over video chat, recorded in 2020
- Available data
  - Pre- and Post-conversation survey
    - Perceptions of their conversation partners, their feelings about the overall conversation, their personality, etc..
  - Video recording
    - Computationally extracted features
    - Modalities: Text, Audio, and Video

Structured as an "vertically integrated" framework

- "low-level" mechanical features of conversation,
  - e.g. turn-taking.
- "mid-level" information streams,
  - e.g. semantic exchange, psycholinguistic markers, and emotion expressions,
- "high-level" judgments reported after conversation,
  - people's enjoyment and the impressions they formed of their conversation partners.



Low-level features

- Closest to raw signals in the audio, video, and text of a conversation recording
- Often vary on a nearly continuous time scale.
- Require some degree of inference:
  - Extracting vocal markers with signal processing (pitch, volume/loudness, etc.)
  - Automated transcription with Automated speech recognition (ASR)
  - Head movement, eye gaze, etc.
- Concrete, specific, and objective properties from which higher-order inferences are derived.

High-level features

- Subjective judgments about their conversations.
- Formed on a coarse time scale (e.g. conversation level)
- Reflected in the survey responses:
  - measures of liking, enjoyment, and conversational flow,
  - evaluations of one’s partner’s social status, intelligence, and personality.
  - …
- The broadest range of information and context are incorporated in making such judgements, and thus distinguished as high-level inferences.

The rich mid-level features

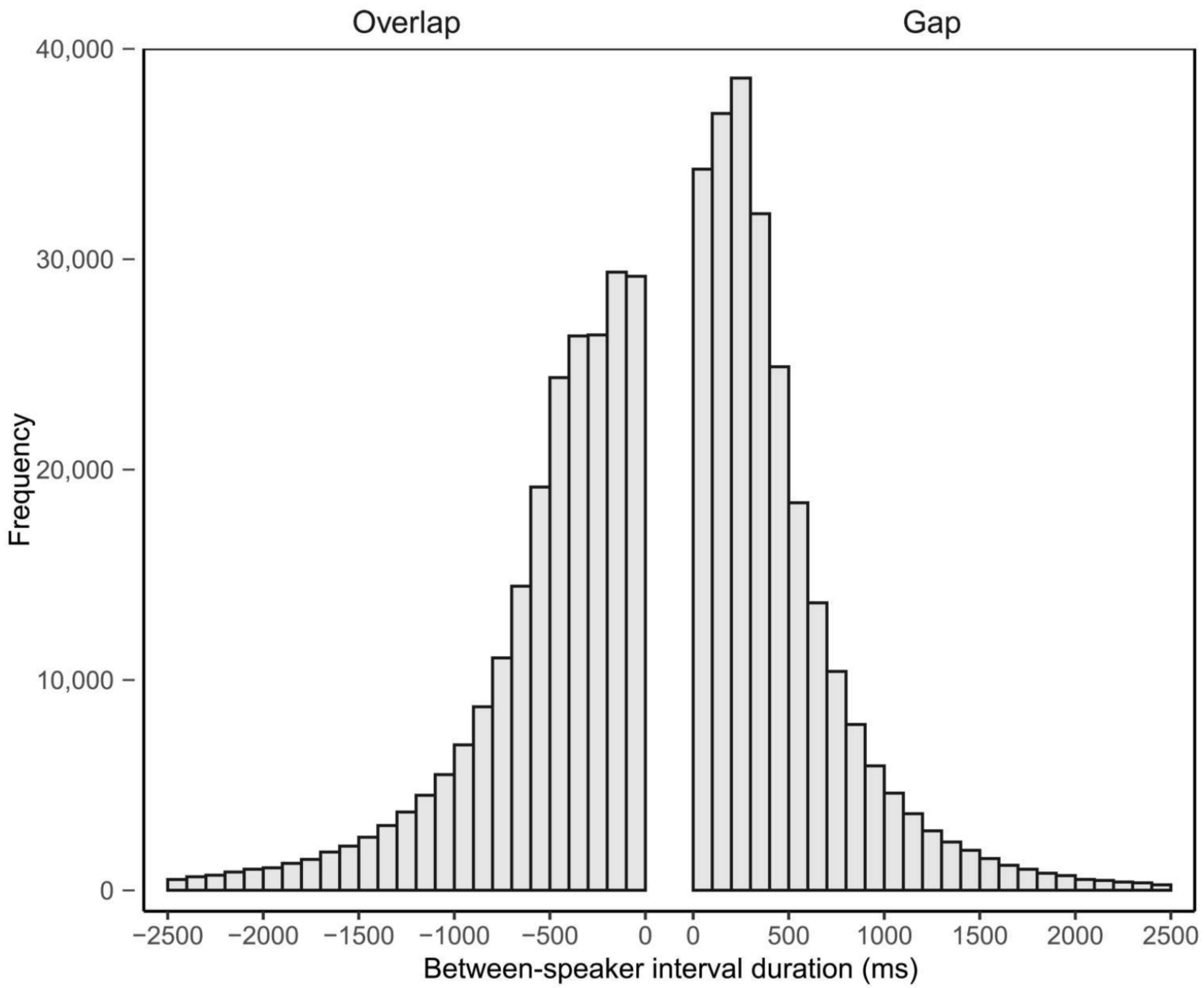
- numerous features related to subjective perceptions of the interaction that typically vary on an intermediate time scale (e.g. turn level)
- Usually computed using algorithmic tools that were trained to attend to specific aspects of speech, sound, and movement to infer a psychological content, for example:
  - happy facial expression,
  - increasing intensity in one’s voice,
  - a timely change of subject.
- Characterized by their use of a narrower scope of context and antecedent reference.
  - *“A hitch in the voice, a sad glance away; all these signals, essential for shared understanding between humans, will go unnoticed by a machine that knows only language.”*

The Turn-Taking System (low-level)

- Turn exchange: the way people manage to pass the floor back and forth in an orderly and efficient manner;
- Turn duration: how long speakers talk before they turn over the floor;
- Back-channeling: the active engagement that listeners display while speakers are talking
  - “mhm,” “yeah,” and “exactly”

Turn exchange

- Median between-speaker turn interval was +80 ms and was distributed approximately normally.
- Consistent with previous literature.
- Takeaway: reproduce findings in existing literature as sanity check!



Turn duration — What do we mean by “turn”?

Fatima: Hello, my name is Fatima. I’m from Egypt.

Eduardo: Hi. Nice to meet you.

Audiophile

Hello,	
	Hi.
my name is Fatima. I’m	
	Nice
from	
	to
Egypt.	
	meet you.

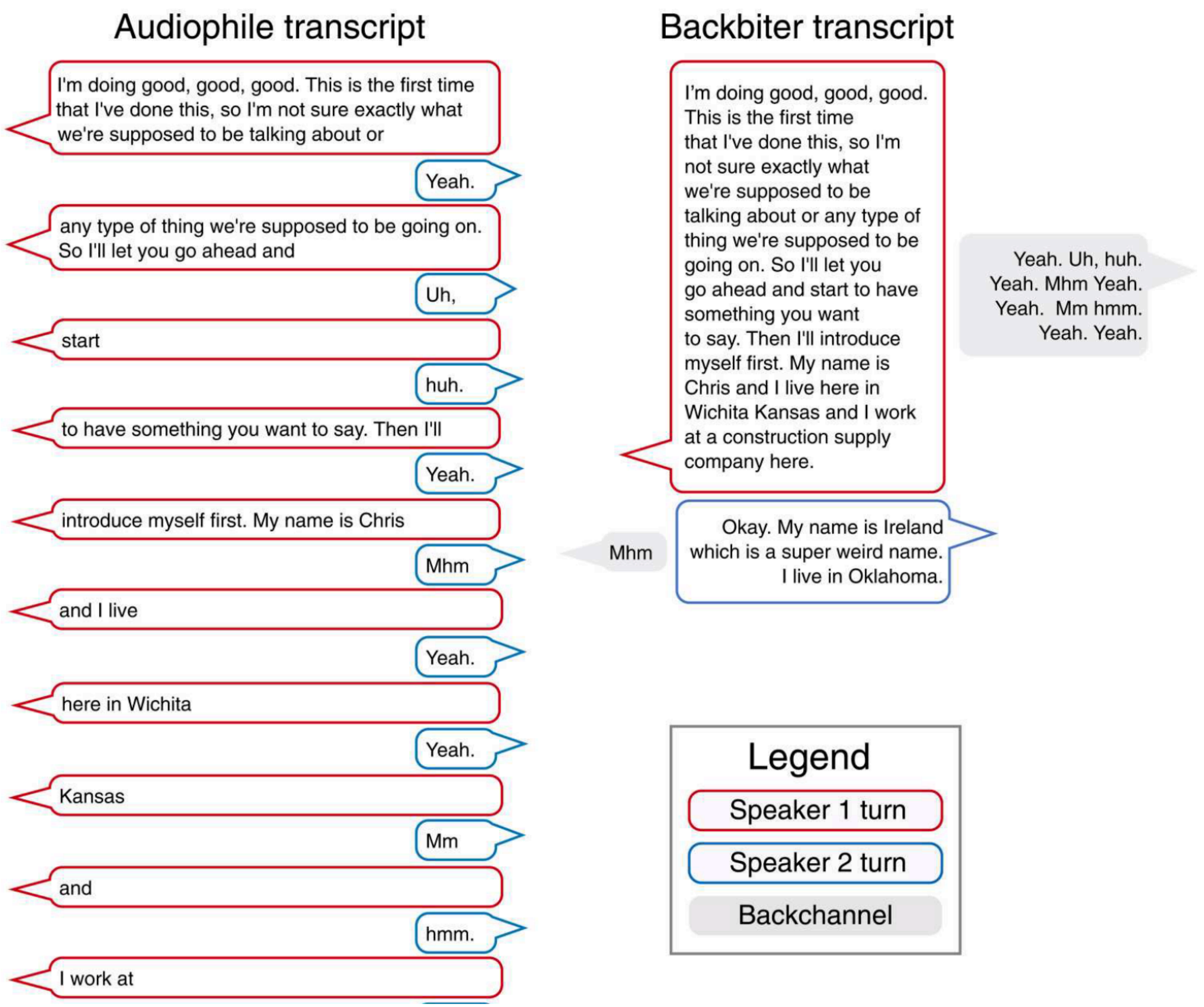
Cliffhanger

Hello, my name is Fatima.	
	Hi.
I’m from Egypt.	
	Nice to meet you.

Table 1. Basic comparison of Audiophile and Cliffhanger turn models. Speakers’ mean and median turn durations were four to five times greater for the cliffhanger turn model compared to those of Audiophile. This suggests more broadly that analytic decisions about transcript segmentation will play a key role in the empirical investigation of speaking duration, an understudied topic.

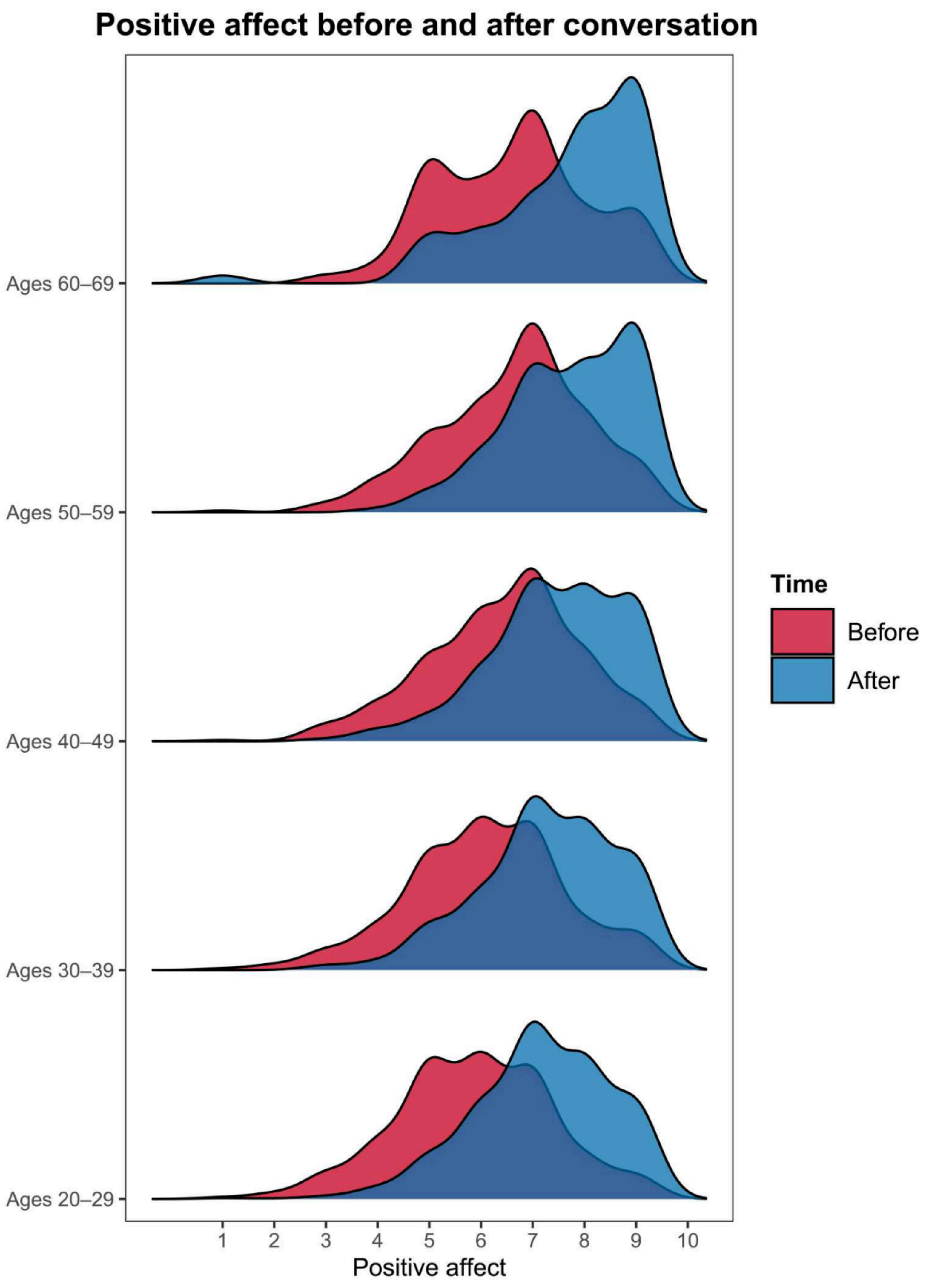
Model	Turn duration		Number of words		Average turns
	Mean	Median	Mean	Median	
Audiophile	2.22	0.92	6.40	2	440.70
Cliffhanger	8.52	5.81	17.81	9	159.41

Back-channelling



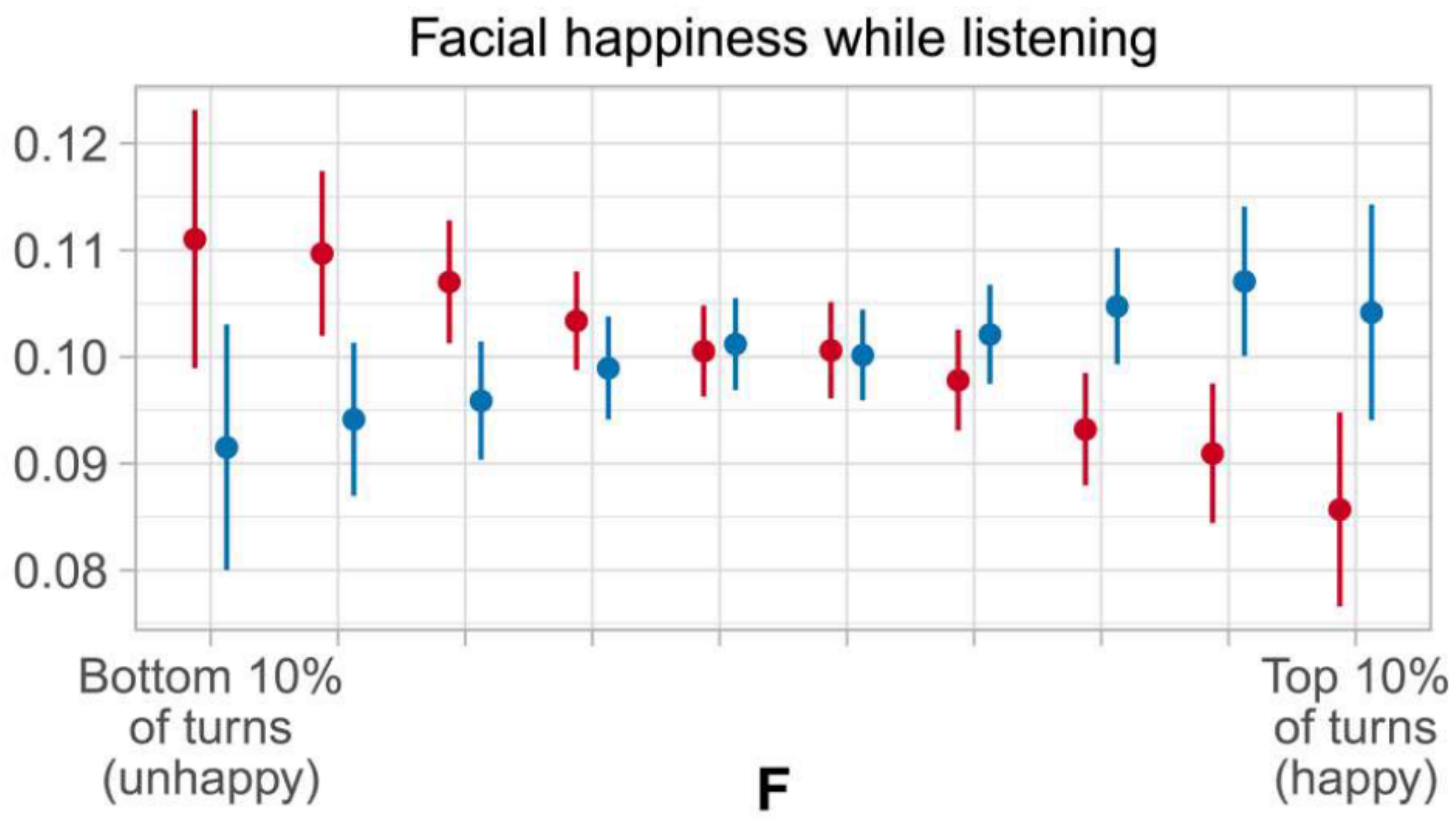
Conversation and well-being (high-level)

- Respondents were asked to report their mood immediately before (red) and after (blue) their conversation.



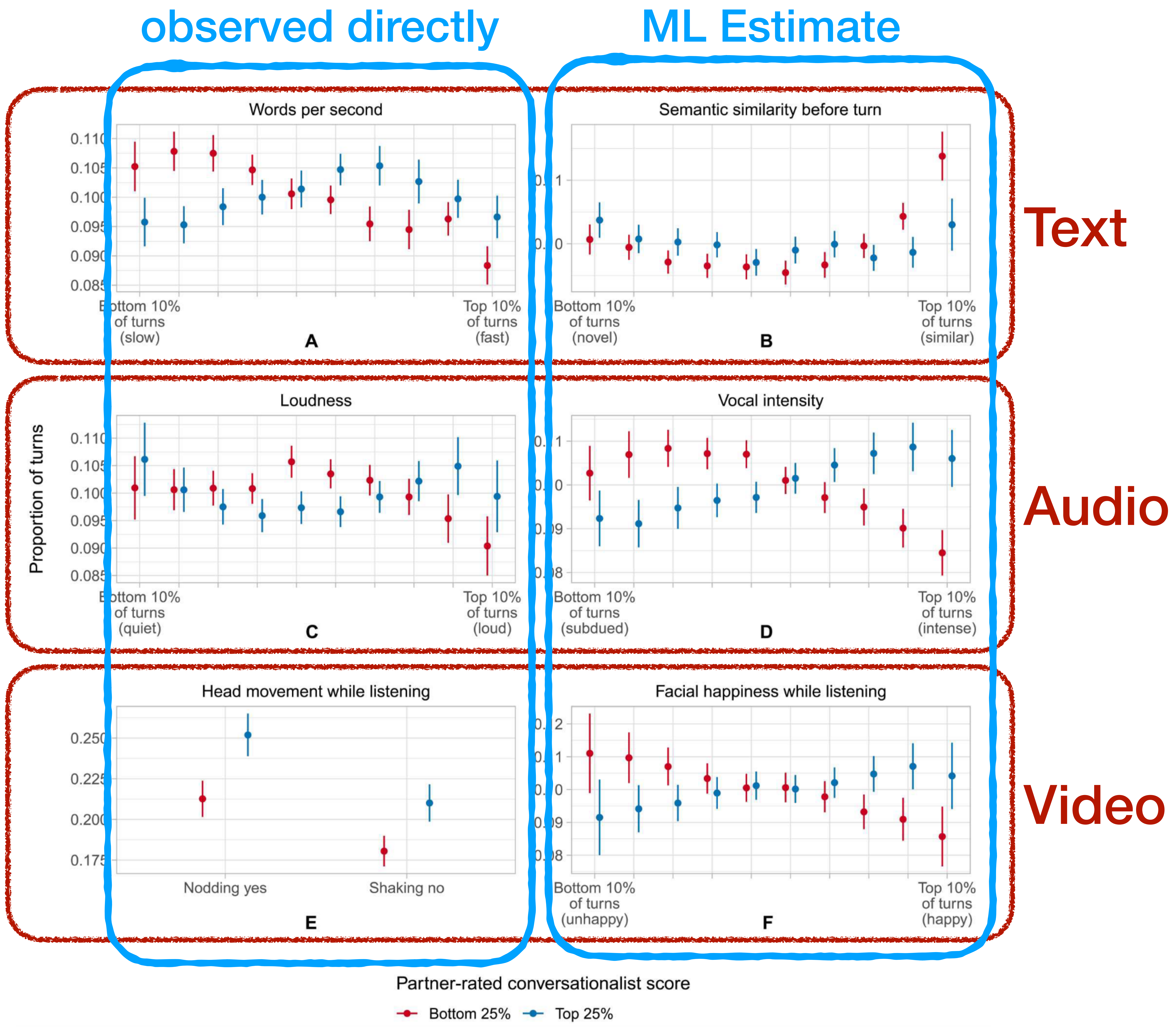
Now is the fun part!

- *We explore this rich "middle layer" of interaction and associate it with high-level impressions by exploring an open question in conversation research: What distinguishes a good conversationalist?*
- E.g. good conversationalists exhibited significantly more facial happiness expressions while listening.



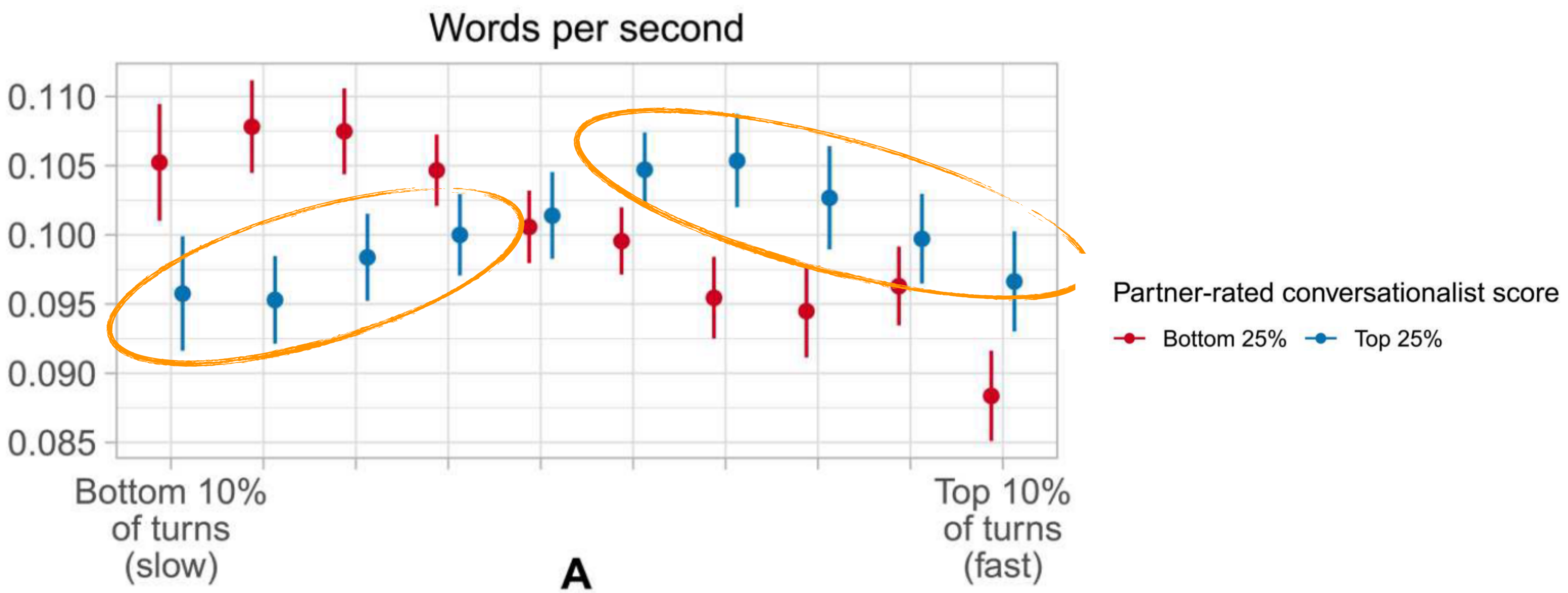
Behavior Patterns of Good and Bad Conversationalists (fig. 8)

- Visualized with frequency plots, as opposed to linear analysis.



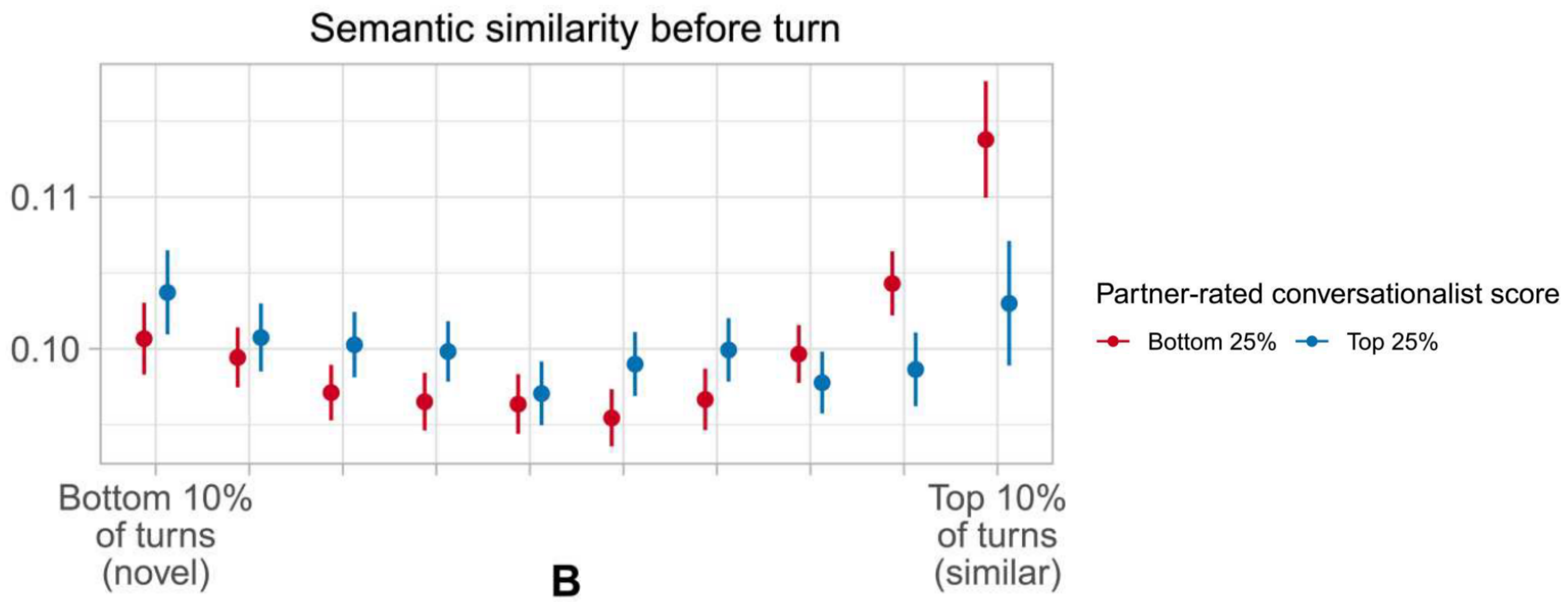
Speech rate (word per minute)

- Good conversationalists spent more of their turns speaking quickly.
- Bad conversationalists spent a greater proportion of turns speaking slowly.



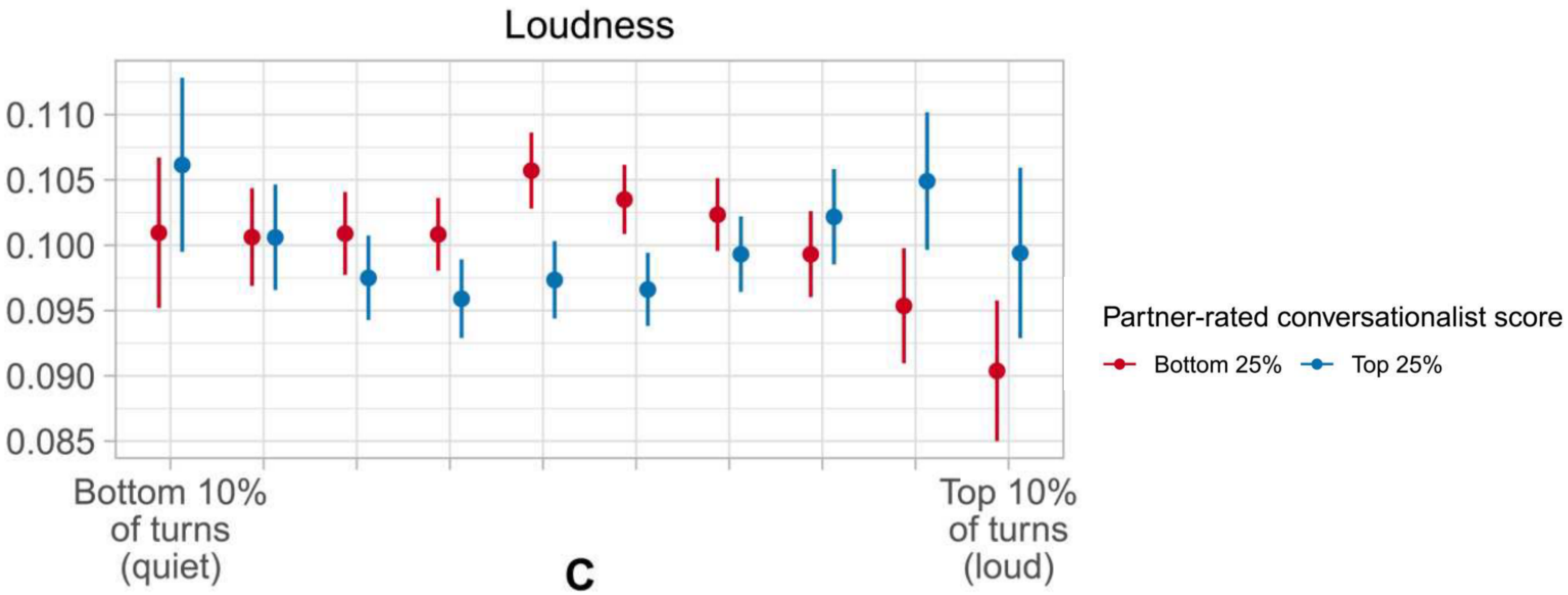
Semantic similarity/novelty

- Use Text Embedding as vector representation of the spoken content.
- Use cosine similarity of the embedding vectors (this turn vs. previous turn) as a measurement of similarity/novelty.
- Good and bad conversationalists differ significantly.
- But good conversationalists also do not always add more novelty.



Loudness

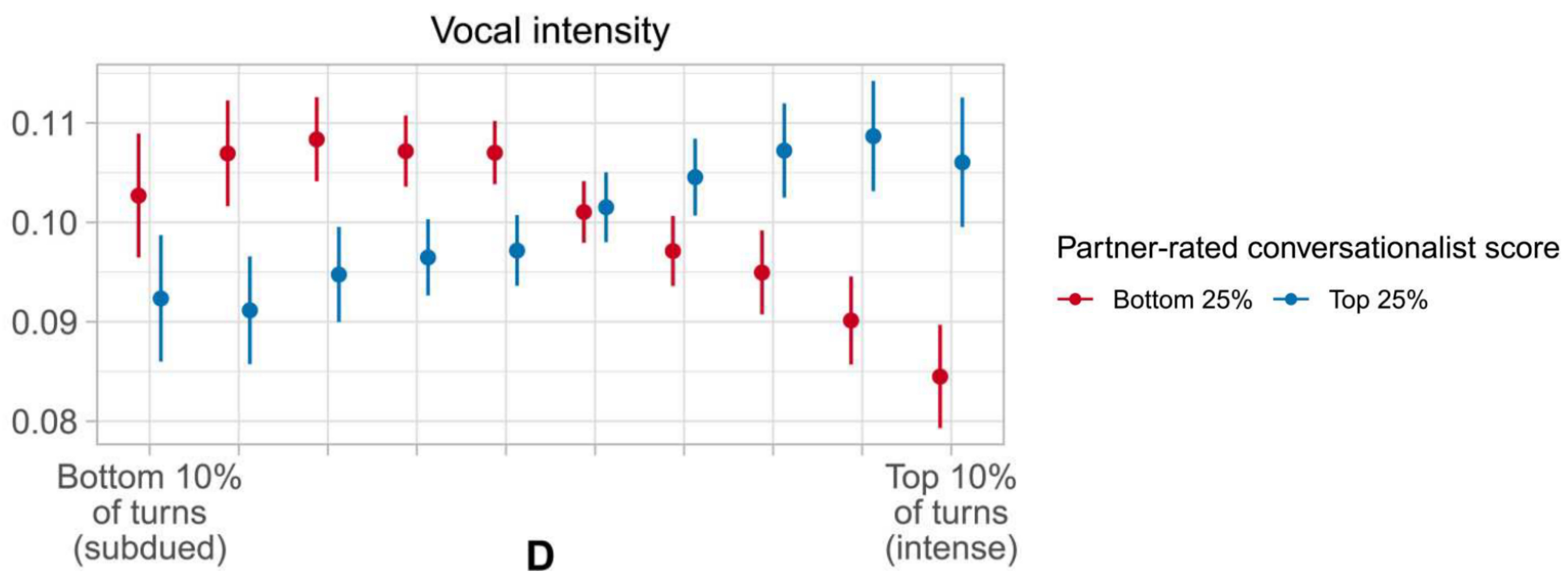
- No significant difference in the mean turn loudness.
- Bad conversationalists take more turns of medium loudness,
- Good conversationalists take more turns with either lower or higher average loudness



C

Vocal intensity

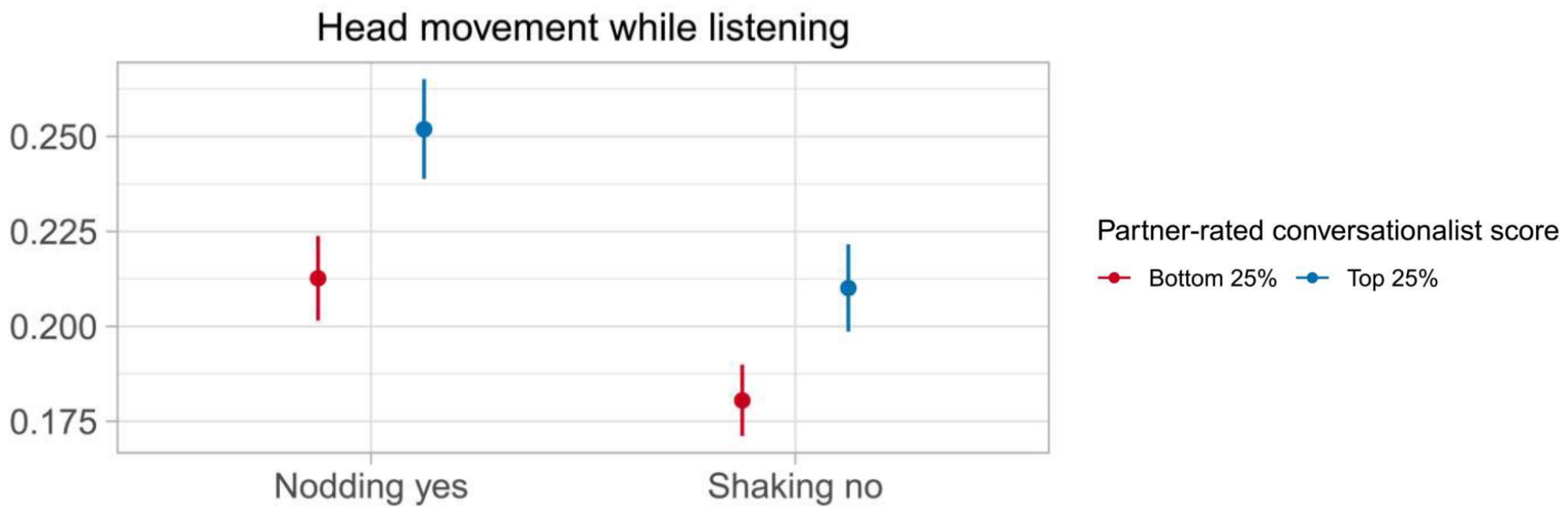
- A vocal intensity classifier is trained with the RAVDESS dataset.
  - RAVDESS: recordings of trained actors who were prompted to read simple statements with either "normal" or "high" emotional intensity (used as labels).
  - Features include summary statistics of common prosodic features, such as
    - mean, max, and SD of fundamental frequency (F0), volume (log energy), etc.
  - The classifier is then applied to each turn in the CANDOR dataset
- people rated as good conversationalists spoke with greater intensity than bad conversationalists



Head Movement

Question: Is there better way to recognize nod and shake?

- Facial recognition followed with heuristic rules
  - Check if over a 2-s period, (i) at least 10% of a participant's face (ii) crossed its beginning position at least twice.
  - If yes, record "nod" if it occurs along vertical axis, record "shake" for horizontal.
- Good conversationalists were significantly more engaged not only in nodding "yes", but also "shaking" no.

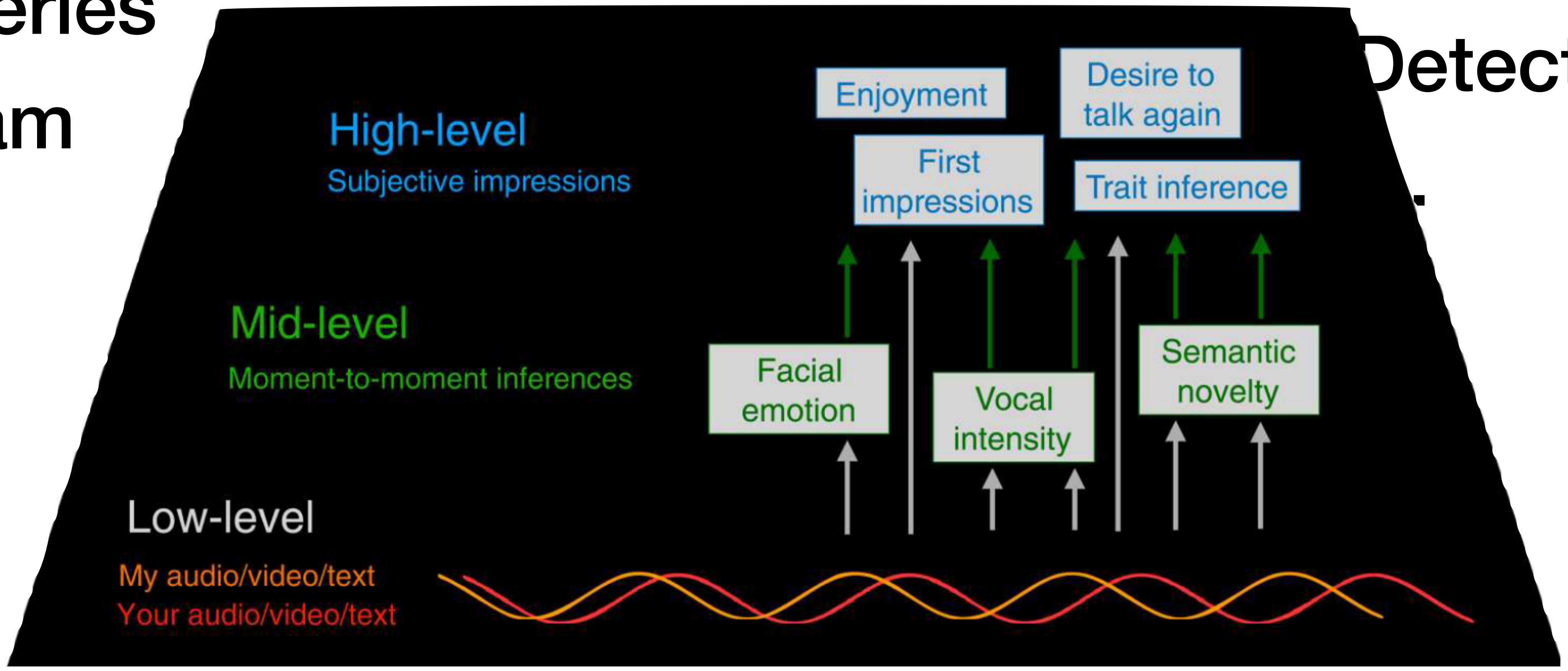


The rest of the CANDOR paper discusses

- A qualitative glance of the topics in the conversation.
- Practical considerations and limitations of the dataset.
- A detailed Supplementary Materials

Low-,Mid-,High-level features meets data mining pillars

Data to be mined	Data Representation	Basic Functionalities	Knowledges (Outcome)	Applications
Text	Itemset	Pattern	Association	Recommender System
Image	Vector/Matrix	Similarity	Classification	Fraud Detection
DNA	Sequence		Clustering	Web Search
Behavior	Network		Ranking	...
...	Time Series		Detection	
Audio	Stream			
Video	...			



My opinion

Data to be mined	Data Representation	Basic Functionalities	Knowledges (Outcome)	Applications

Itemset

Association

Vector/Matrix

Low-level feature that requires understanding of the data genre

The rich mid-level features requires creative ways to use existing data mining algorithms, with the guidance of domain knowledge.

High-level feature that relates to the outcome being pursued

Time Series

Behavior

Outlier Detection

Stream

…

…

Audio

Video

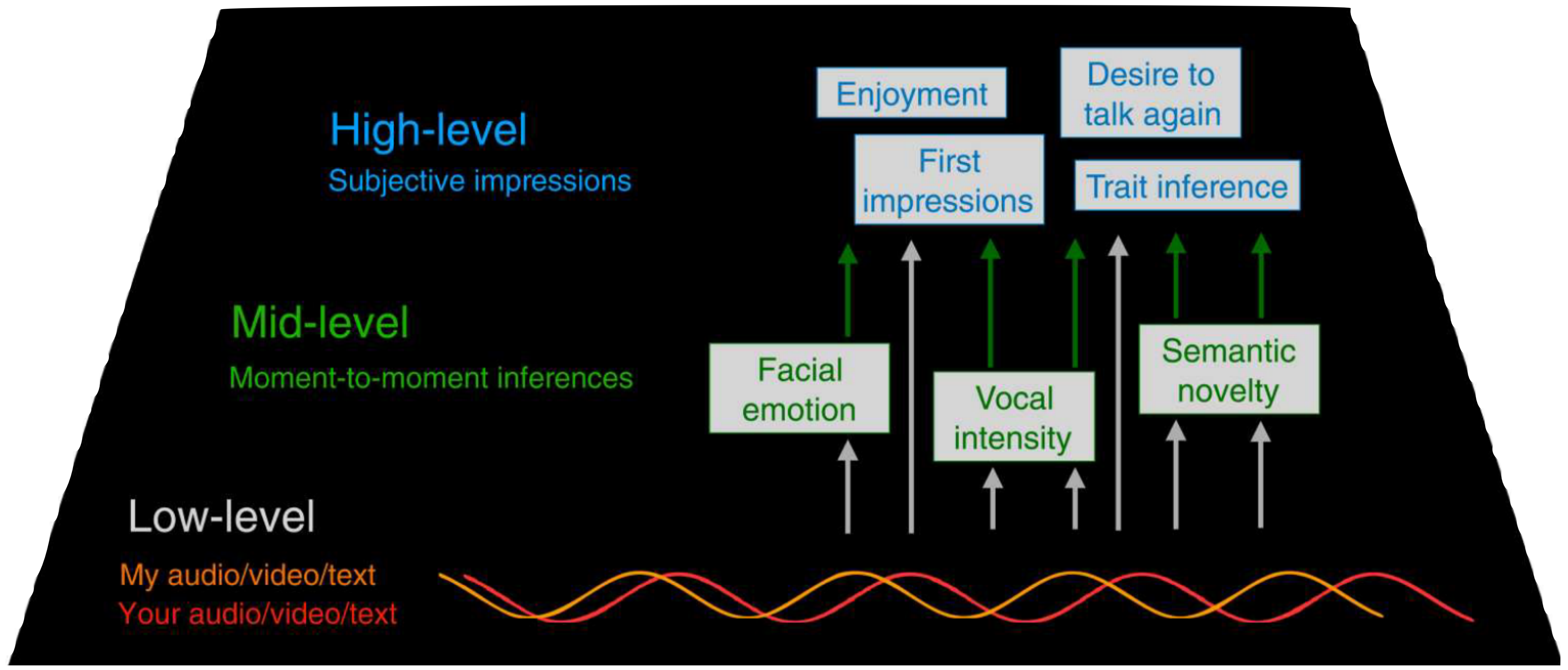
…

What you should know

- Multimodal data is a new data genre that fits into the data mining pipeline.
- Roughly categorized as low-, mid-, high-level features.
- Mid-level features offers the richest opportunities.
- Conversation analysis is centered on the turn-taking mechanism.
  - Data mining techniques and domain knowledge should go hands-in-hands.

Discussion: From Conversation to Classroom

- What is equivalent to a “turn” in classroom?
- What would be the low-, mid-, and high-level features in the classroom setting?
- What can or cannot be applied?



Demo: Audio Feature Extraction with OpenSmile

Assignment - Option 1

- Reproduce a emotion intensity classifier by following the approach described in the CANDOR paper
  - Using the RAVDESS dataset, and use “emotional intensity” as the label.
  - Compute a series of acoustic summary statistics—mean, maximum, and standard deviation for fundamental frequency, log energy, and voiced and unvoiced duration—as the features.
  - Train a logistic regression model (w/ or w/o feature normalization and regularization)

Assignment - Option 2

- Analyze how acoustic features contribute to emotion recognition, including:
  - Using the RAVDESS dataset, and use "emotional intensity" as the label.
  - Fundamental frequency (F0) as a measurement of pitch
  - Intensity/amplitude measurements (related to perceived loudness)
  - Frequency perturbation (jitter) and amplitude perturbation (shimmer)
  - Additional features such as spectral measures (e.g., MFCCs, formants) and voice quality metric
- Easier than Option 1 as it does not require training models.

Bonus Slides

How does LLM change the game?

Multimodal Foundation Models

Data to be mined	Data Representation	Basic Functionalities	Knowledges (Outcome)	Applications
Text	Itemset	Similarity	Association	Recommender System
Image	Vector/Matrix		Classification	Fraud Detection
DNA	Network		Clustering	Web Search
Behavior	Time Series		Ranking	...
...	Stream		Outlier Detection	
Audio	...		...	
Video				

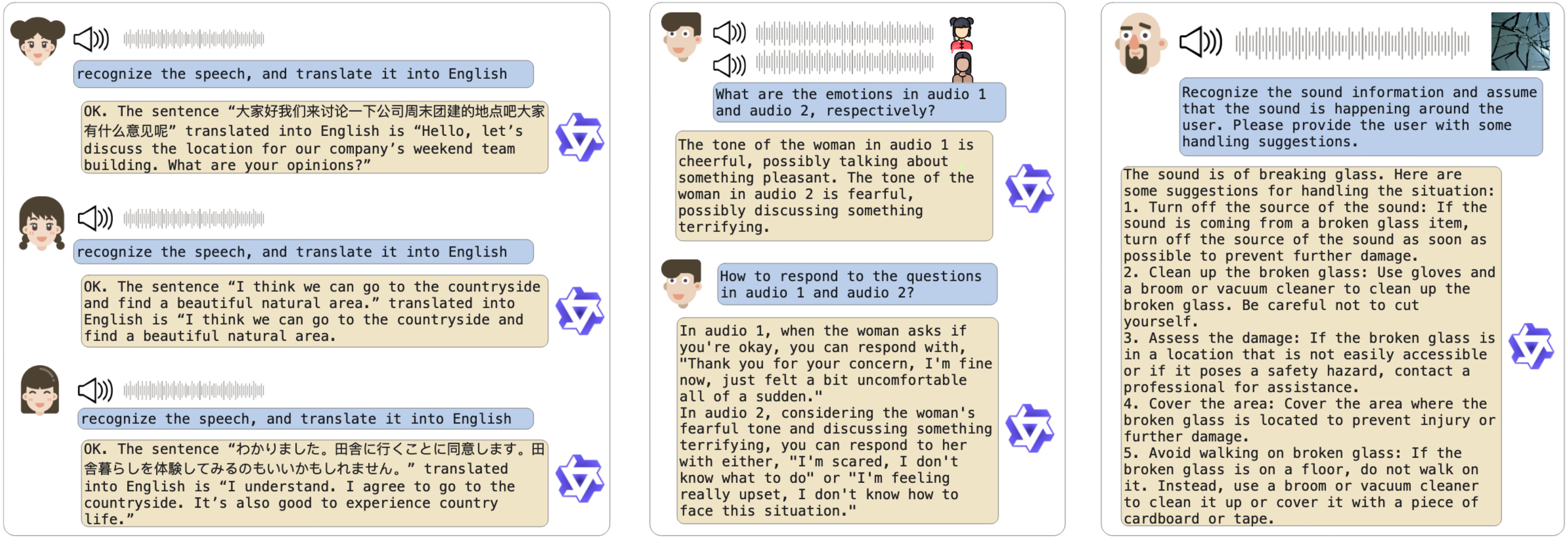
End-to-end approach leading to a pre-trained model with the capacity of in-context learning

we no longer care about the data representation and basic functionalities, we pre-trained the model and expect that it can perform (generalize) well on various task, through either fine-tuning or in-context learning

(Large) Audio-Language-Model

- Audio processing is challenging due to data scarcity, but language is abundant.
- ALM can learn a shared latent space between audio-language modalities.
- In the post-LLM era, LALMs integrate audio encoders with LLMs, unlocking deeper audio perception and reasoning.
  - Capable of captioning (ASR), reasoning, and open-ended QA.
  - Examples:
    - Gemini, GPT-4o,
    - [open access] Qwen-audio, Qwen-audio-chat, Qwen-audio-instruct,
    - [open source] GAMA, SALMONN, GAMA, etc.

Capability of Large Audio Language Models

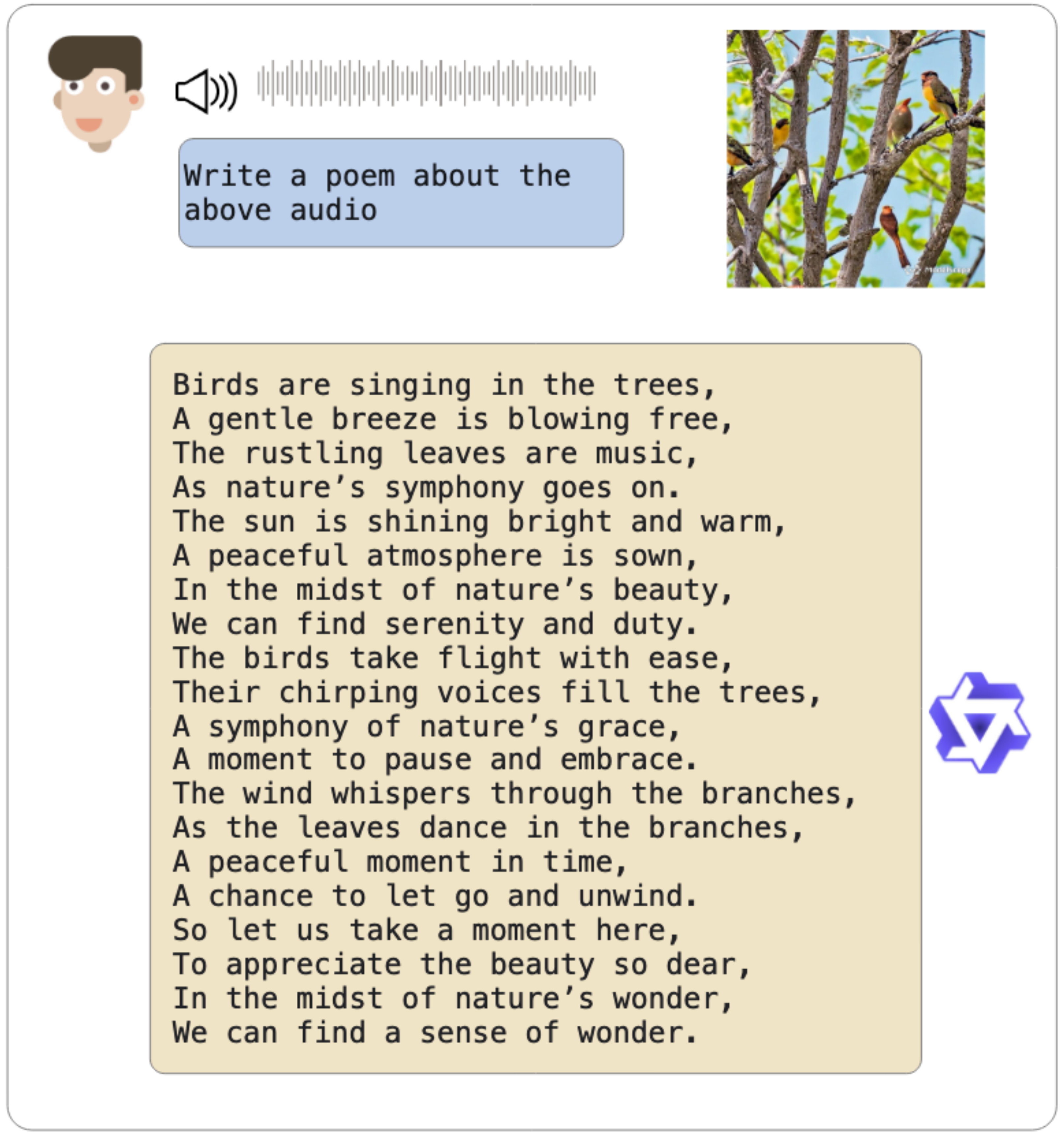


(a) Multilingual ASR and Translation

(b) Multiple Audio Analysis

(c) Sound Understanding and Reasoning

Capability of Large Audio Language Models



(d) Audio-Motivated Creative Writing

(e) Music Appreciation

(f) Speech Editing with SWRT

Assignment - Option 3 [challenging]

- Evaluate audio-capable foundation models for emotion intensity recognition
  - Using the RAVDESS dataset, and use "emotional intensity" as the label.
  - Implement and evaluate one or more audio foundation models such as Whisper and Qwen-Audio
  - For Whisper, you will need to
    - Use Whisper's pretrained model as a feature extractor,
    - Extract embeddings from one of Whisper's encoder layers,
    - Add a classification head on top of the Whisper embeddings,
    - Fine-tune to predict emotion intensity levels
  - For Qwen-Audio, you will need to
    - Practice deploying the Qwen-Audio in a local environment, write prompt to obtain the predicted emotion intensity.