Sec 2:

End-to-End

Current state

Input representations

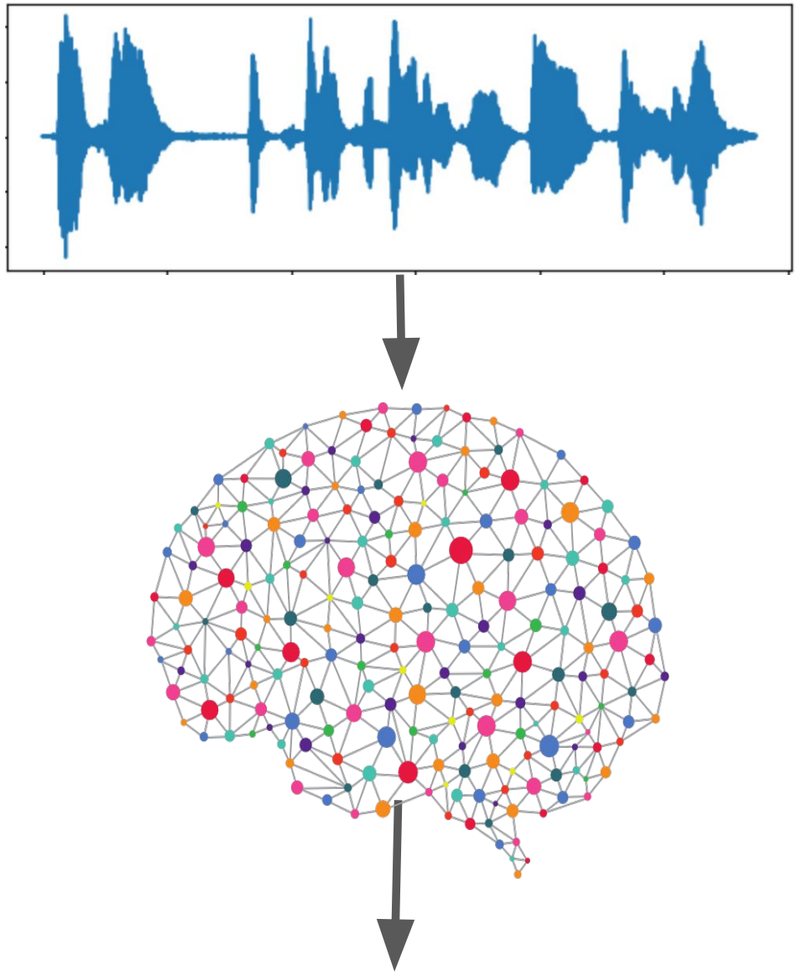
Architecture modifications

Output representations

Sec 2.1

Current state

End-to-end SLT (Bérard et al., 2016; Weiss et al., 2017)



What a wonderful tutorial!

Definition of end-to-end approach

IWSLT 2020 (Ansari et al., 2020)

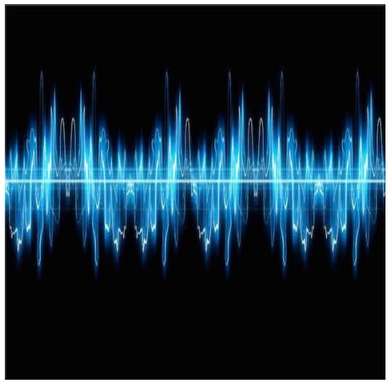
End-to-end model:

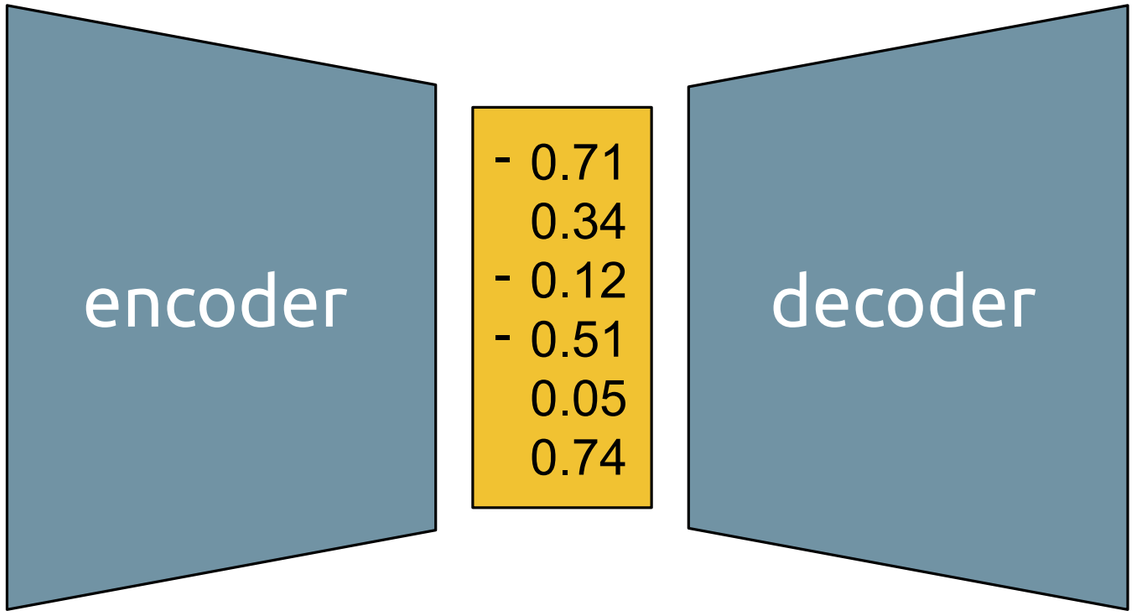
- No intermediate discrete representations (transcripts like in cascade or multiple hypotheses like in rover technique)
- All parameters/parts that are used during decoding need to be trained on the end2end task (may also be trained on other tasks → multitasking ok, LM rescoring is not ok)

Other definitions are possible depending on the application

end-to-end speech translation (e2e)

Spanish Audio

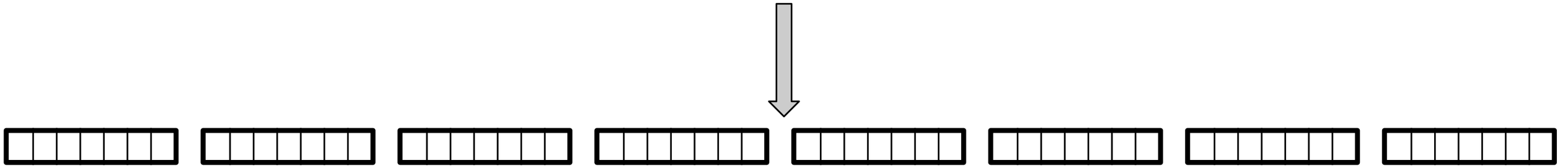




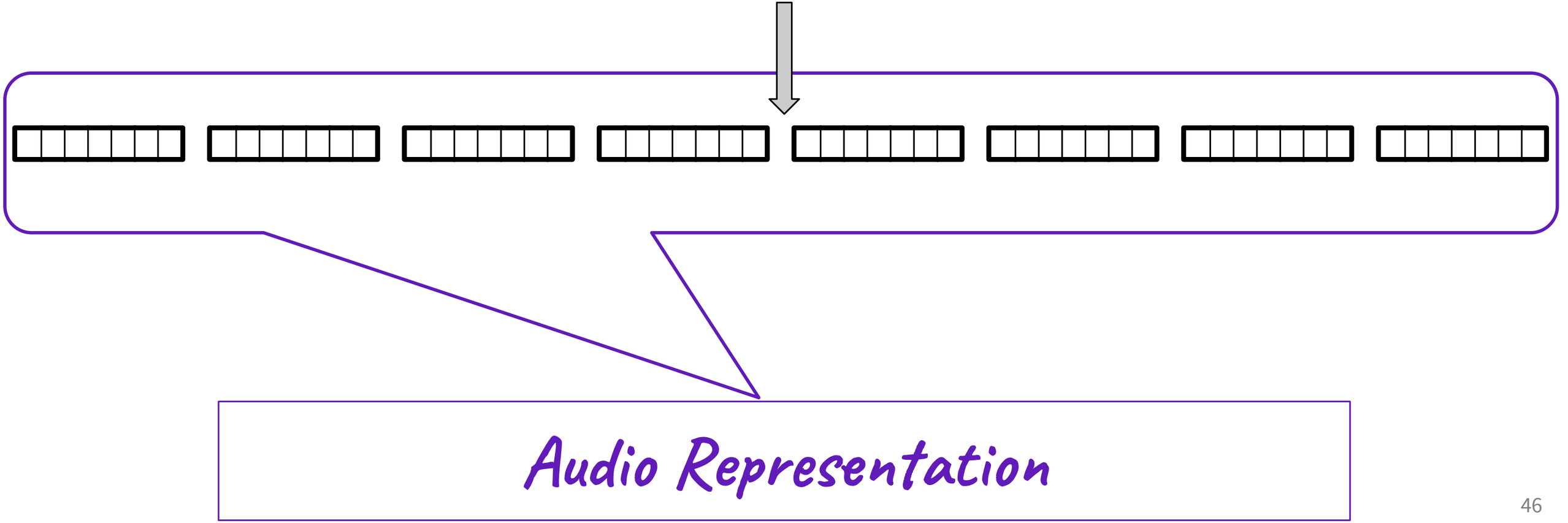
English
Translated text

What a wonderful tutorial!

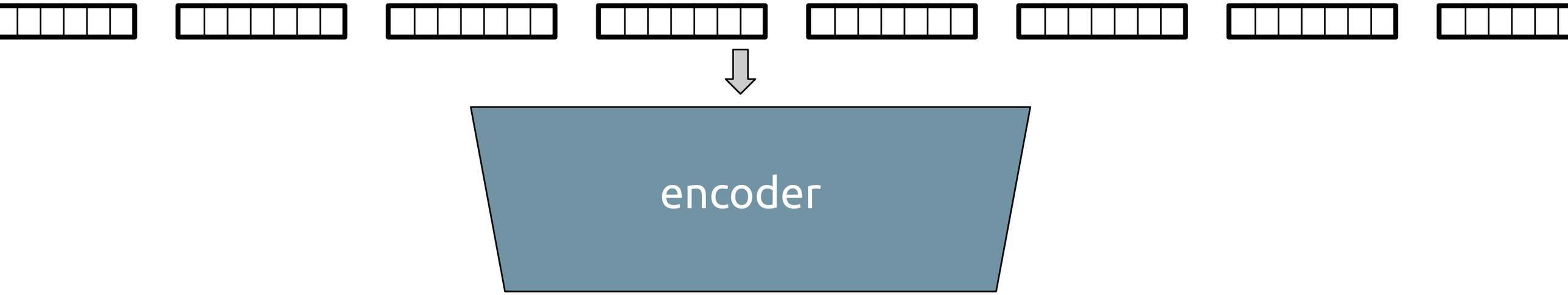
end-to-end speech translation (e2e)



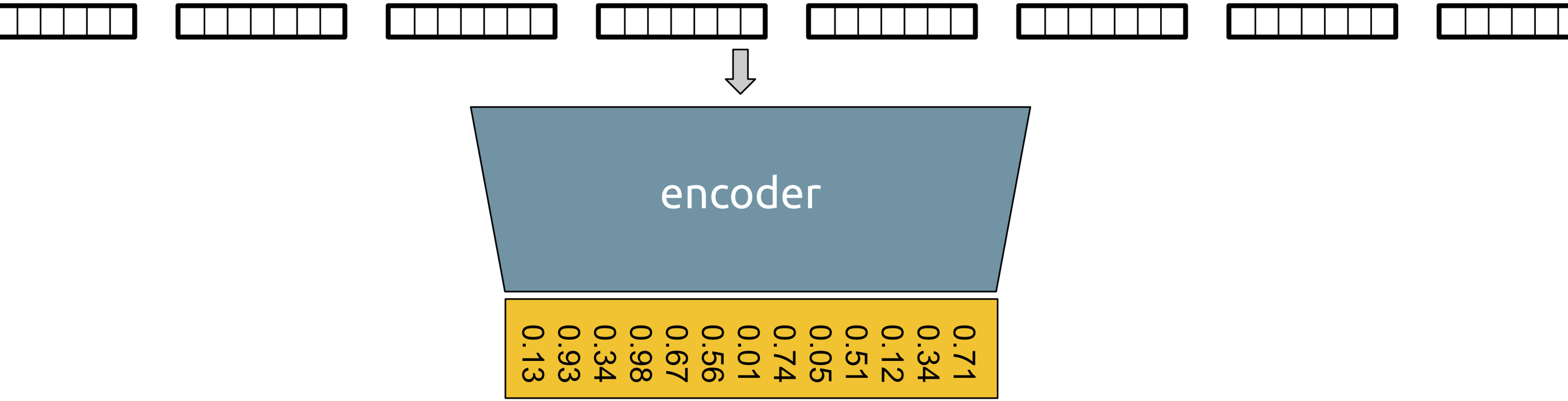
end-to-end speech translation (e2e)



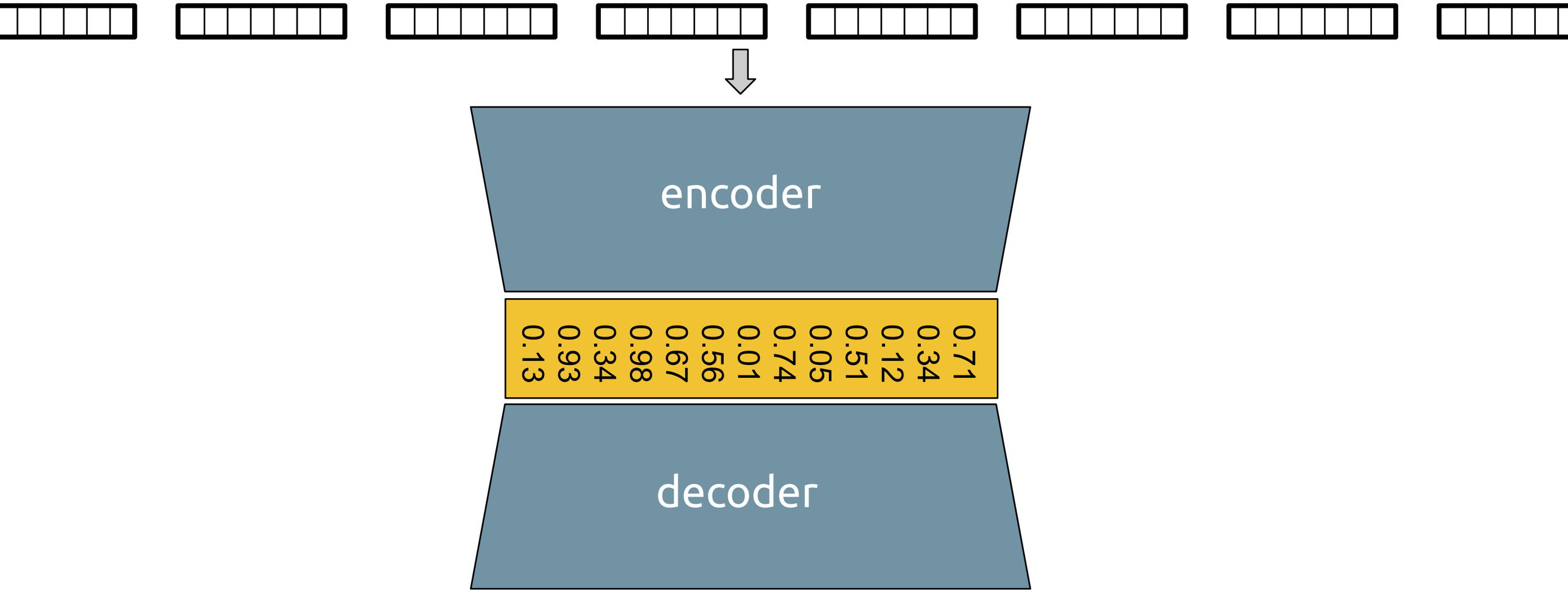
end-to-end speech translation (e2e)



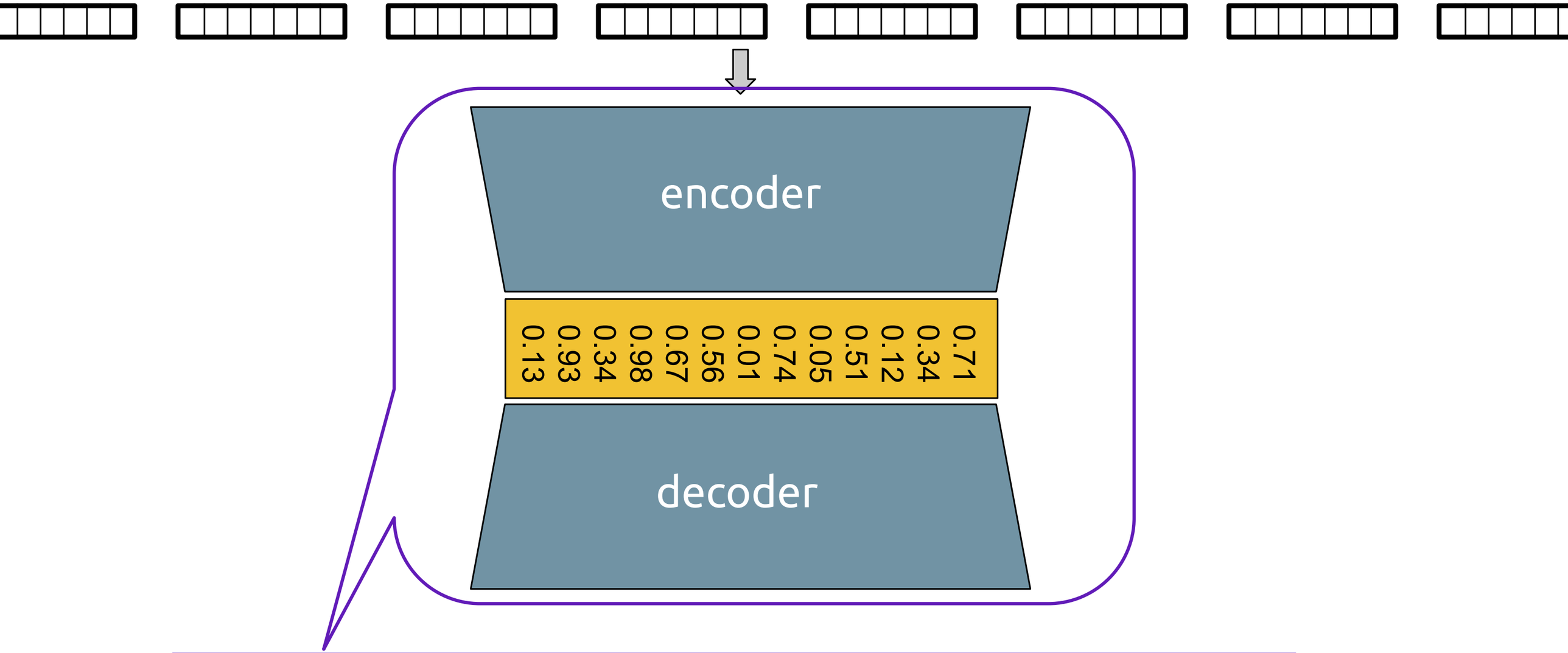
end-to-end speech translation (e2e)



end-to-end speech translation (e2e)

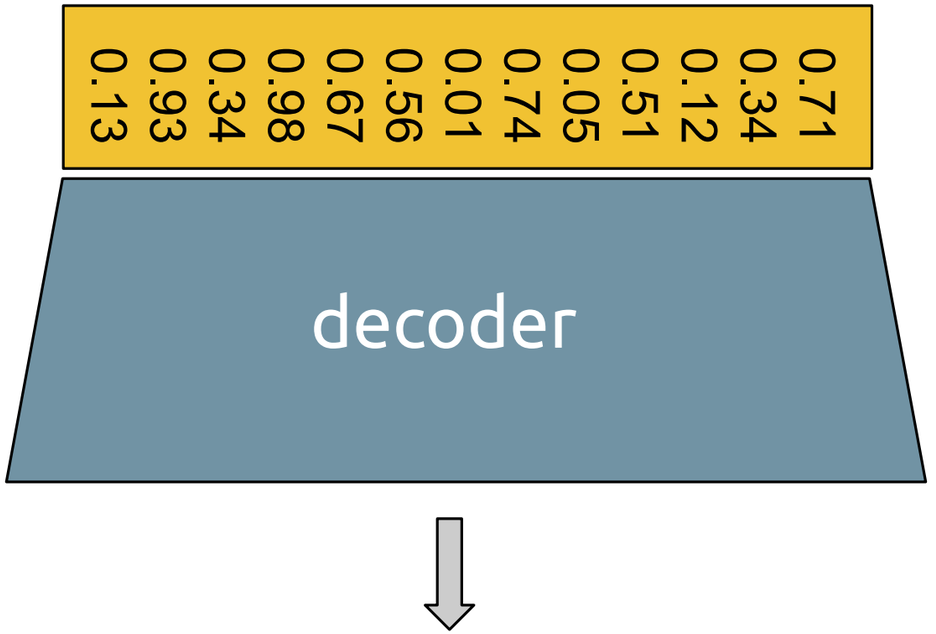


end-to-end speech translation (e2e)



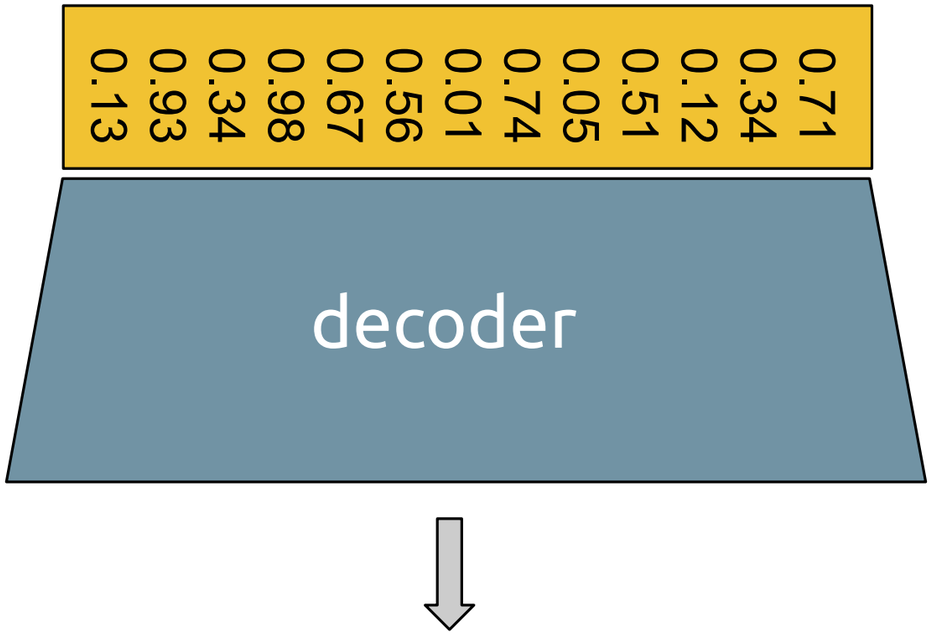
System Architectures

end-to-end speech translation (e2e)



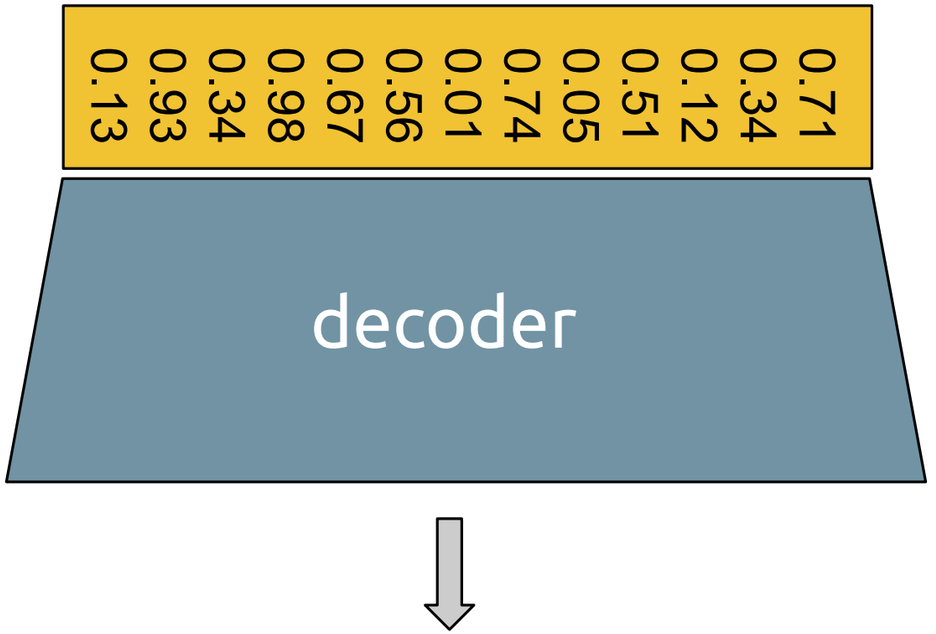
W h a t <space> a <space> w o n d e r f u l <space> t u t o r i a l !

end-to-end speech translation (e2e)



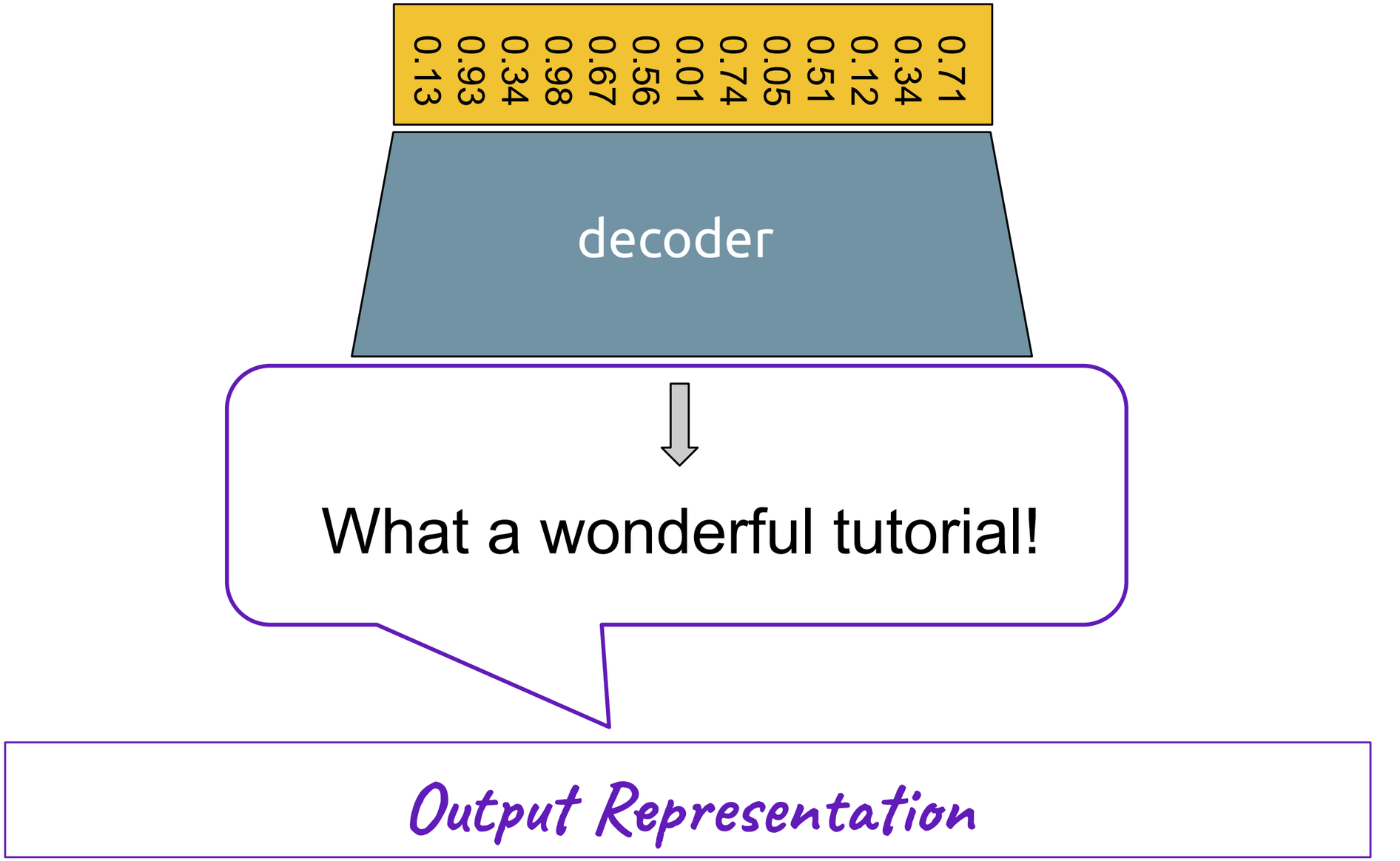
Wh @at a w @on @der @fu @l tut @or @ial!

end-to-end speech translation (e2e)

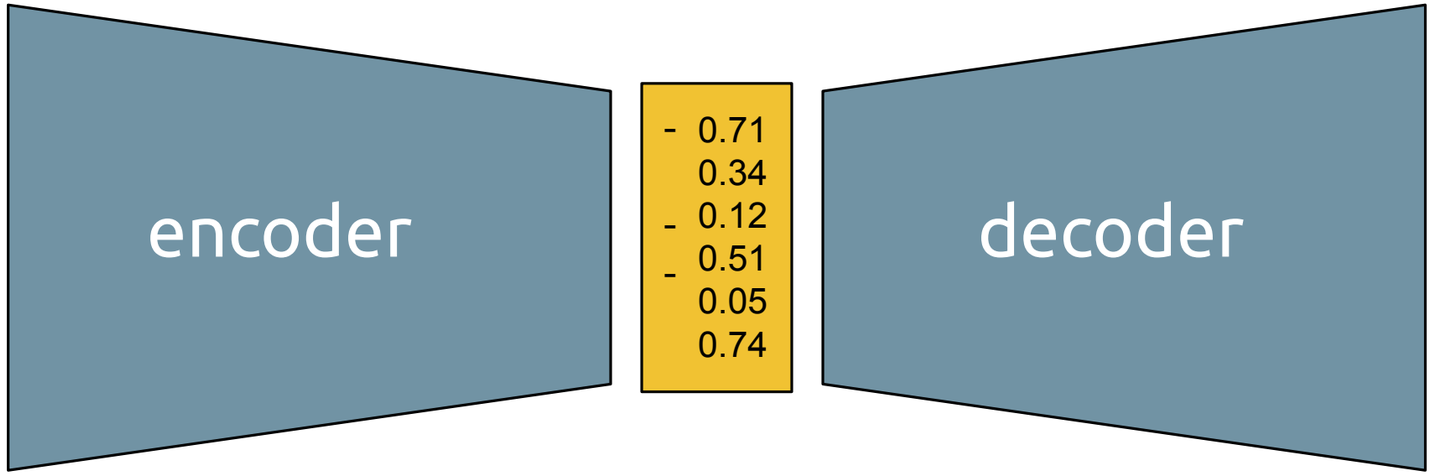


What a wonderful tutorial!

end-to-end speech translation (e2e)



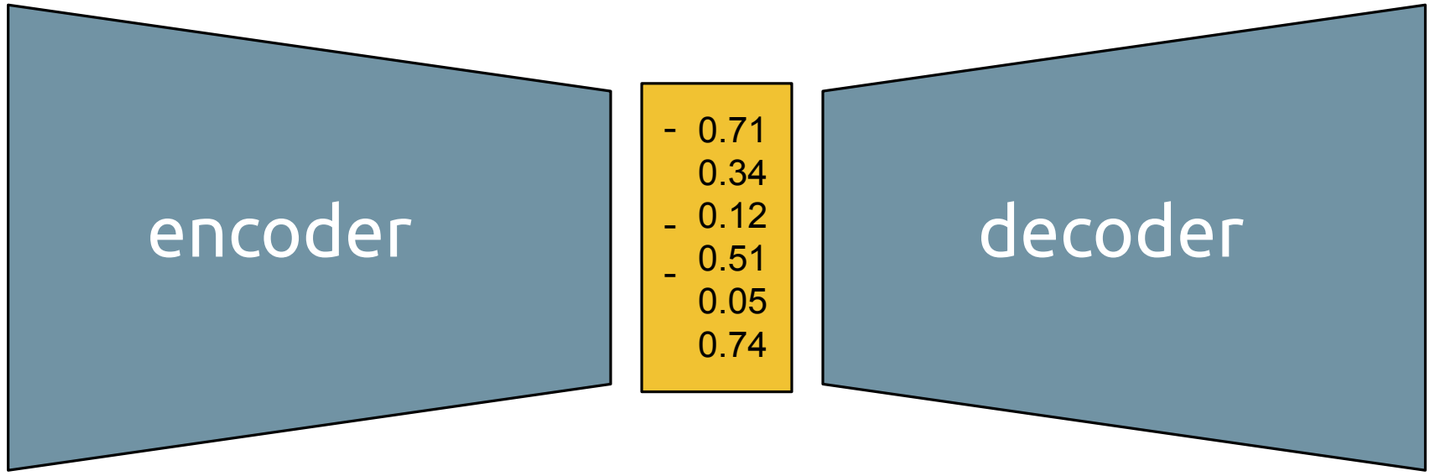
Sequence-to-Sequence Model



Pros:

- Direct access to the audio during translation
- No error propagation
- One system to maintain

Sequence-to-Sequence Model



Pros:

- Direct access to the audio during translation
- No error propagation
- One system to maintain

Cons:

- Less consolidated technology
- Scarcity of training data
- Non-monotonic alignments audio-text

Cascade vs End-to-End Systems

Cascade

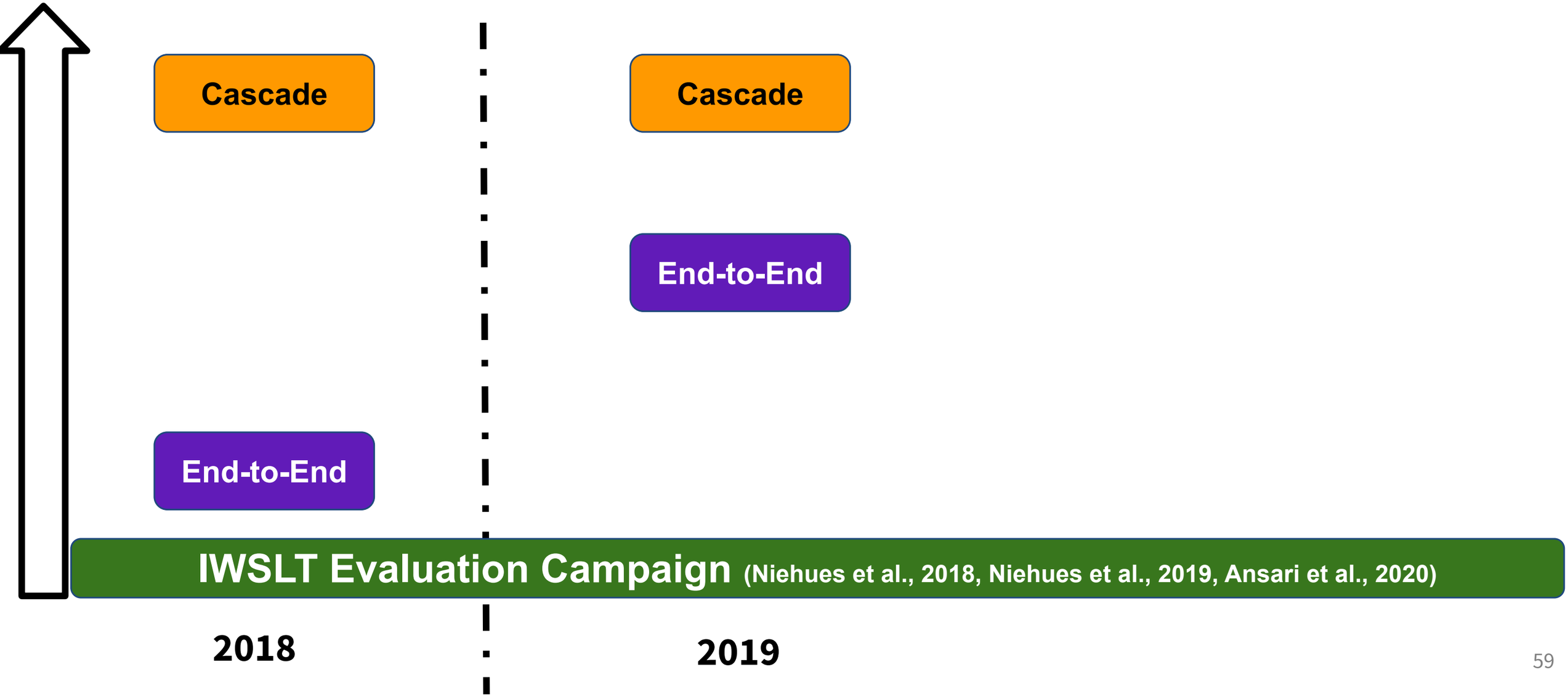
- ✓ Large corpora for ASR and MT
- ✓ Less complex tasks
- ❌ Error propagation
- ❌ Information loss
- ❌ Higher latency

End-to-End

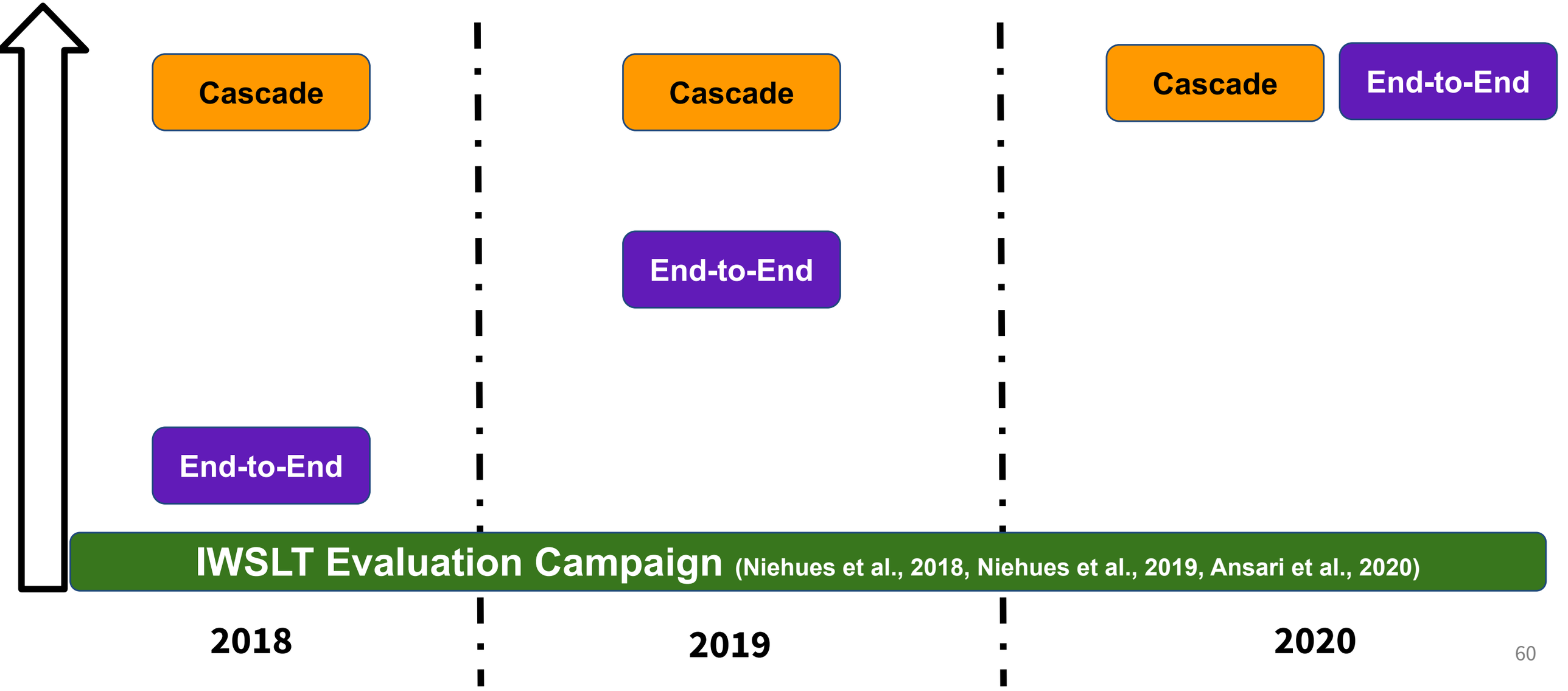
- ✓ Access to all audio information
- ✓ Reduced latency
- ✓ Easier management
- ❌ Small corpora
- ❌ More complex task

Cascade vs End-to-End Systems

Cascade vs End-to-End Systems



Cascade vs End-to-End Systems



Cascade vs End-to-End Systems

Most of the papers (Weiss et al., 2017, Jia et al., 2019, Di Gangi et al., 2019) about end-to-end SLT system mention the following advantages over the cascade:

- No error propagation:

  End-to-end naturally avoids compounding errors between the ASR and MT systems.

Cascade vs End-to-End Systems

Most of the papers (Weiss et al., 2017, Jia et al., 2019, Di Gangi et al., 2019) about end-to-end SLT system mention the following advantages over the cascade:

- No error propagation:

  End-to-end naturally avoids compounding errors between the ASR and MT systems

- Direct access to the audio:

  End-to-end better manipulates paralinguistic and non-linguistic information during translation

The correctness of these statements taken for granted

Cascade vs End-to-End Systems

Key questions:

Is it true that end-to-end avoids error propagation?

To what extent does accessing the audio help? How? When?

Cascade vs End-to-End Systems

Key questions:

Is it true that end-to-end avoids error propagation?

To what extent does accessing the audio help? How? When?

No answers in this tutorial!

No error propagation

Open issues:

- Overall translation quality is not enough to measure the reduction of error prop.

No error propagation

Open issues:

- Overall translation quality is not enough to measure the reduction of error prop.
- For a direct comparison of the Cascade and e2e, the intermediate representations cannot be used (transcript vs. null)

No error propagation

Open issues:

- Overall translation quality is not enough to measure the reduction of error prop.
- For a direct comparison of the Cascade and e2e, the intermediate representations cannot be used (transcript vs. null)
- Difficult to disentangle the impact of various components in e2e (two tasks collapsed into one)

No error propagation

Open issues:

- Overall translation quality is not enough to measure the reduction of error prop.
- For a direct comparison of the Cascade and e2e, the intermediate representations cannot be used (transcript vs. null)
- Difficult to disentangle the impact of various components in e2e (two tasks collapsed into one)
- Not a consolidated architecture in end-to-end technology

No error propagation

Open issues:

- Overall translation quality is not enough to measure the reduction of error prop.
- For a direct comparison of the Cascade and e2e, the intermediate representations cannot be used (transcript vs. null)
- Difficult to disentangle the impact of various components in e2e (two tasks collapsed into one)
- Not a consolidated architecture in end-to-end technology

Possible opening:

Sperber et al., (2019) consider the encoder output as an intermediate representation and pose the attention on the presence of errors in it

Direct access to the audio

Open issues:

- Better encoder technology results in better translation performance (not enough)

Direct access to the audio

Open issues:

- Better encoder technology results in better translation performance (not enough)
- Not clear what aspects of the audio can help (e.g. prosody, emotions, tone, pauses)

Direct access to the audio

Open issues:

- Better encoder technology results in better translation performance (not enough)
- Not clear what aspects of the audio can help (e.g. prosody, emotions, tone, pauses)
- Audio understanding capability can only be analyzed in the final translation (no transcripts)

Direct access to the audio

Open issues:

- Better encoder technology results in better translation performance (not enough)
- Not clear what aspects of the audio can help (e.g. prosody, emotions, tone, pauses)
- Audio understanding capability can only be analyzed in the final translation (no transcripts)
- Lack of *ad hoc* test sets to measure the impact of prosody, emotions, ...

Direct access to the audio

Open issues:

- Better encoder technology results in better translation performance (not enough)
- Not clear what aspects of the audio can help (e.g. prosody, emotions, tone, pauses)
- Audio understanding capability can only be analyzed in the final translation (no transcripts)
- Lack of *ad hoc* test sets to measure the impact of prosody, emotions, ...

Possible openings:

Karakanta et al. (2020): the direct access to the audio pauses improves subtitles' quality

Gaido et al. (2020): vocal characteristics can guide e2e systems in modeling gender (but opens ethical issues!)