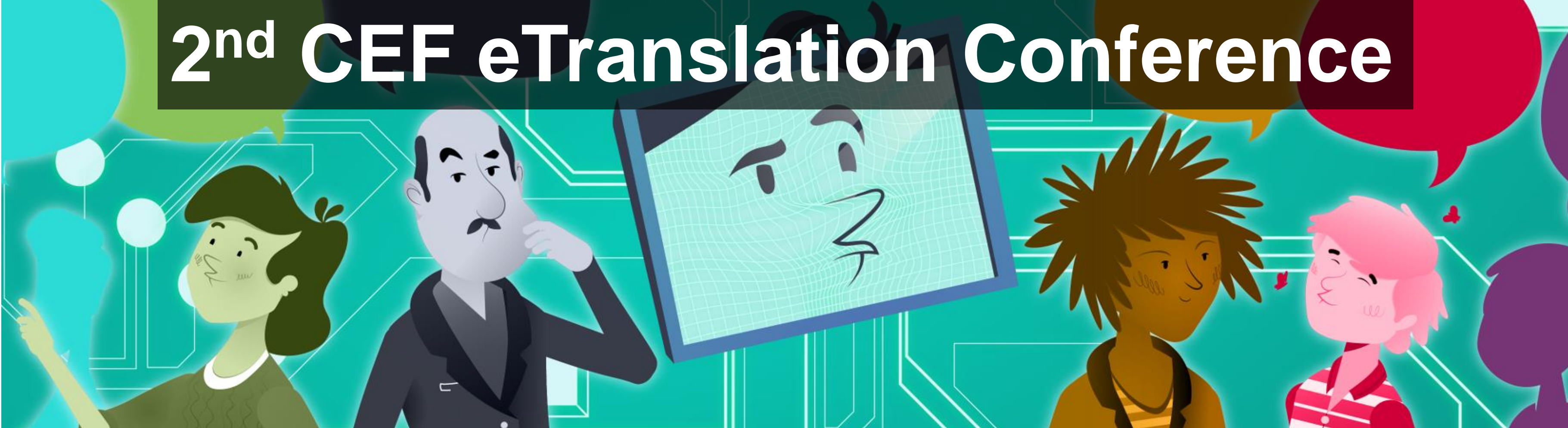


2nd CEF eTranslation Conference



Text-to-speech & Speech recognition

Kris Demuynck

 **umec**

IDLab
INTERNET & DATA LAB


**UNIVERSITEIT
GENT**

2nd CEF eTranslation Conference

Text-to-speech & Speech recognition

(technology for speech to speech translation)

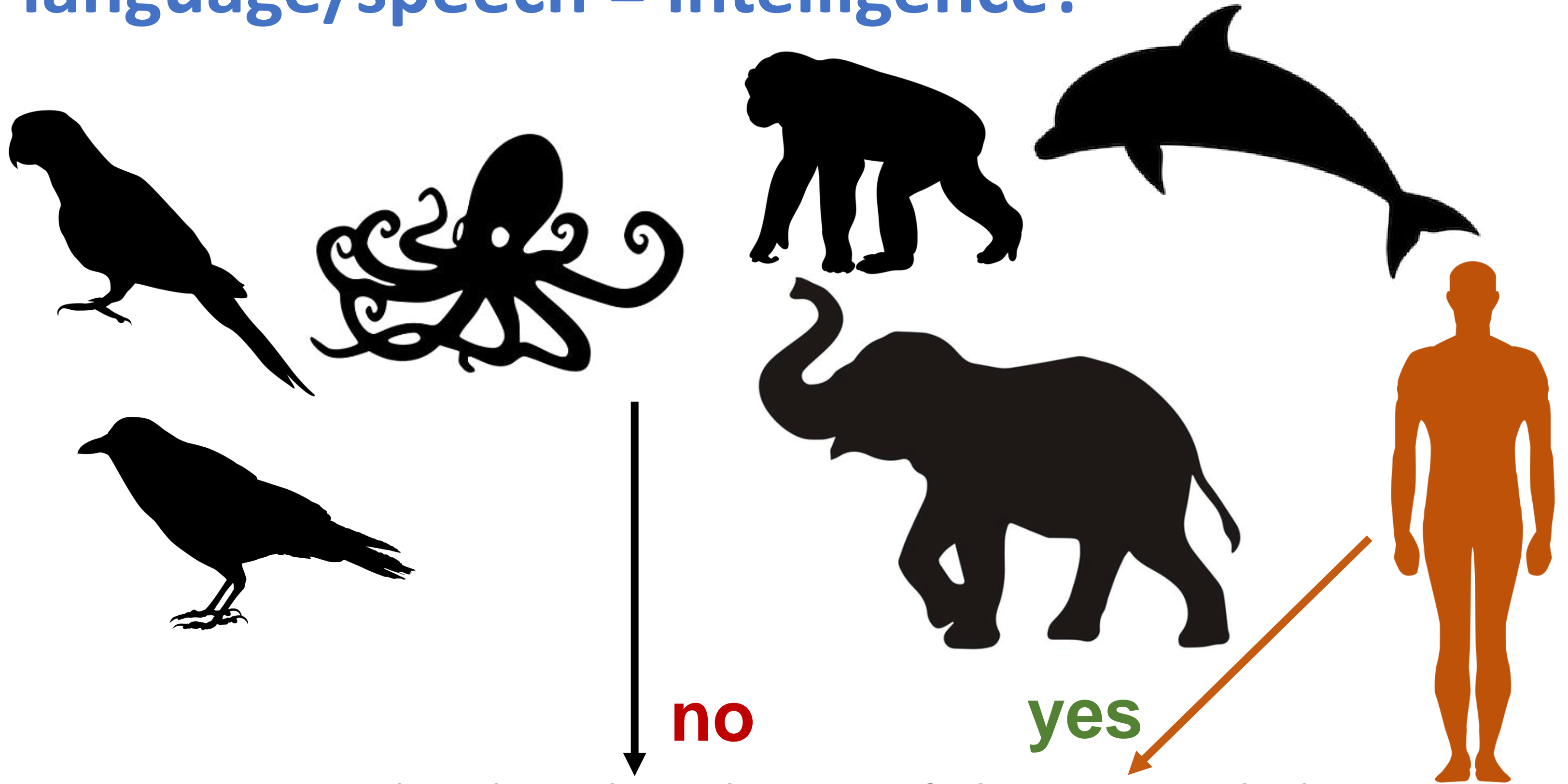
Aim

high-level understanding of the technology and its limitations

Topics

- ***Intelligence: human intelligence vs artificial intelligence***
- ***How do computers cope with spoken and written language***
- ***Challenges in speech→text, text→speech and speech→speech***

language/speech = intelligence?



conveying complex thoughts; the use of abstract symbols

**Taal is de sleutel tot echte artificiële
intelligentie**

***Language is the key to true artificial
intelligence (strong AI)***

(Walter Daelemans, De Tijd, Sept 18, 2018)

Speech & language by humans ...

Seems easy ... but we tend to forget it took years of daily practice to master this.



*start: passive
listening*



*3 years kinder garden:
short sentences*



*6 year primary school:
mastering complex sentences*



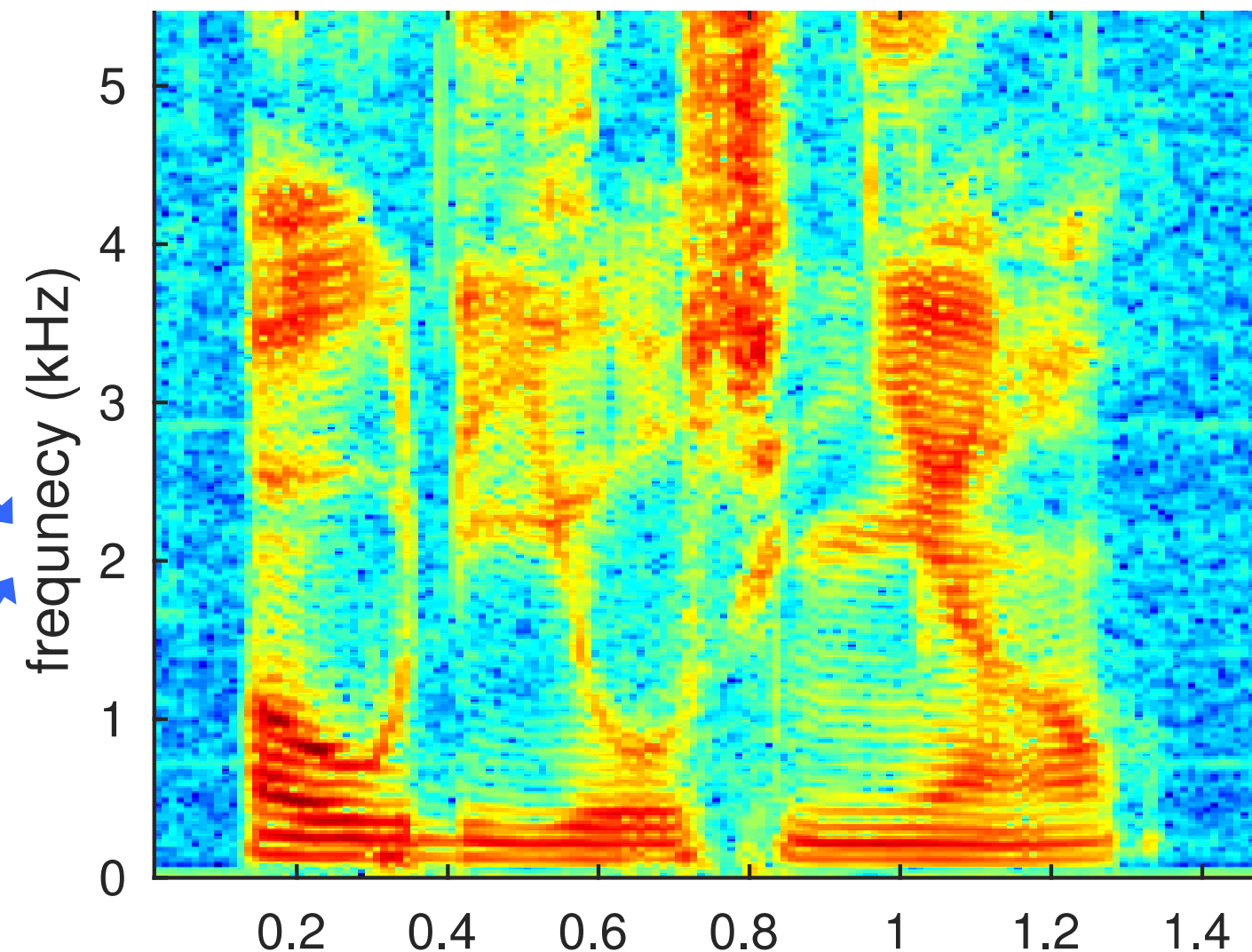
*6 years of secondary school:
extend vocabulary, world
knowledge, other languages*

Speech is difficult: an analogy

humans



similar to handwriting recognition



computers

The quick brown fox jumps over the lazy dog.

Coping with speech: an analogy

note: speech does not have pauses (blanks) between words

'She said that she would dance with me if I brought her red roses,' cried the young student, 'but in all my garden there is no red rose.'

From her nest in the holm-oak tree the nightingale heard him, and she looked out through the leaves and wondered.

According to a state Council decision, China gave a boost up to promote the nine-year compulsory education system by abolishing tuitions and other incidental fees for urban students from the coming autumn term. The decision was taken at the State Council meeting presided over by the Prime Minister, Mr. Wen Jiabao. It was decided in the meeting that the

well articulated, standard (dialect free) speech; note: still lots of variation

Coping with speech: an analogy

speech with heavy dialect

1200 2001 December
Yes
reached on my way to
the in the morning
unpleasant perhaps
when very possible
of the day. In the morning the
in the morning
then

speech in noise

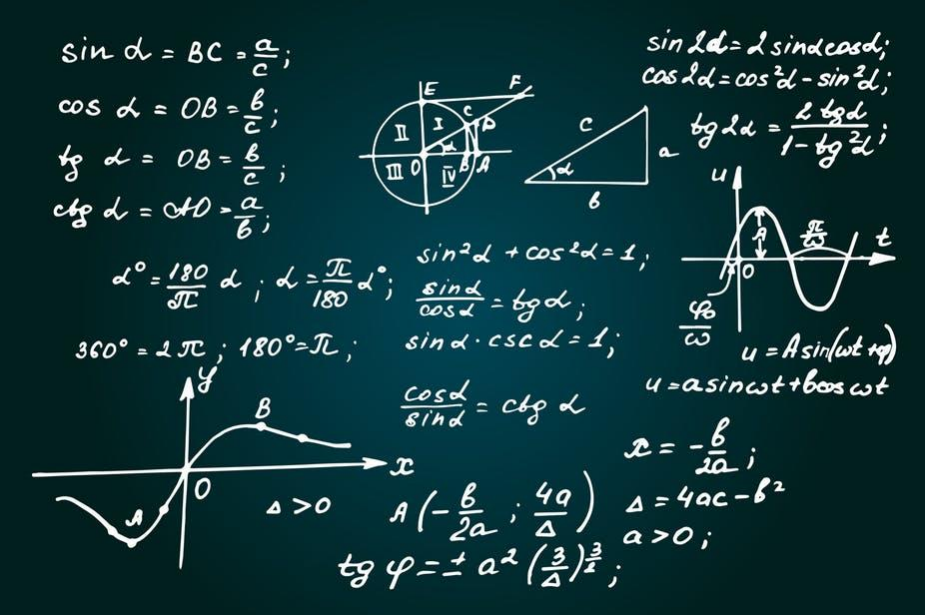
Dear father
I received A letter
from Charles D Knight and
was glad to hear from you to hear
the family was as well as they ask
for help and health I just
Come to the railroad station
I have been A helping road

spontaneous speech

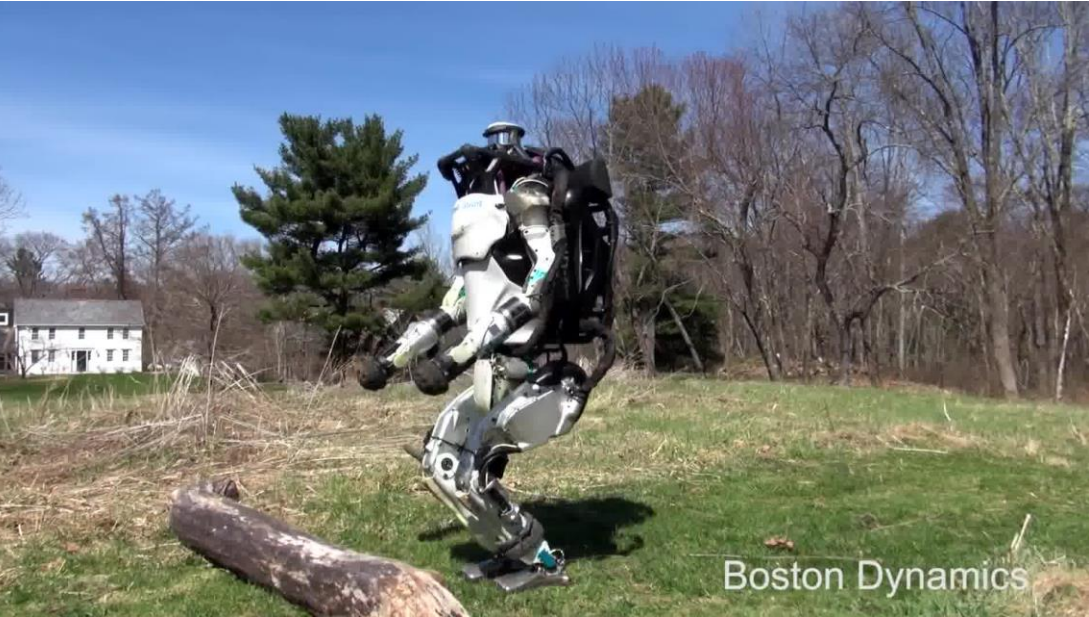
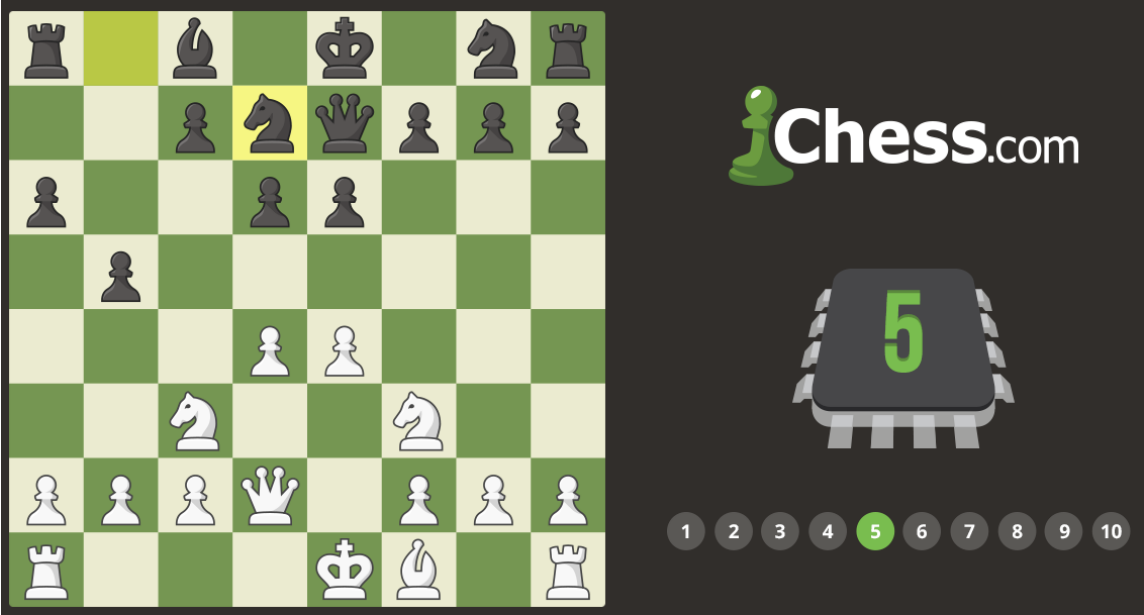
"Let's not go," behind said, sounding exhausted
"Hah!" Barton said quietly, as if ~~talking to himself~~ making a note to
Everyone headed for the exit while Brandon lingered behind. Something entrance
him about the mirrors... he couldn't put his finger on it. It was too bizarre
to make sense of.
As he stared down into the ~~murky~~ ^{dark} reflections, he ~~then~~ ^{saw} a ~~sharp~~ ^{glint} of
red gleam near the ceiling. ~~Then~~ ^{He} ~~looked~~ ^{reached} ~~it~~ ^{the} ~~spot~~ ^{glint} turning
into a deeper, more lurid shade. He jerked the flashlight up and ~~zoomed in on~~
the corresponding spot above his head, ~~which~~ ^{which} ~~down slightly.~~ ^{down slightly.} ~~But then~~
Nothing ~~was~~ ^{was} ~~there~~ ^{there} ~~but~~ ^{but} the strong, wispy remnants of.

Artificial intelligence: Moravec's paradigm

what is considered difficult for humans, tends to be easy for computers,
what seems easy to humans tends to be very challenging for computers



difficult for humans



difficult for computers



Artificial intelligence: Moravec's paradigm

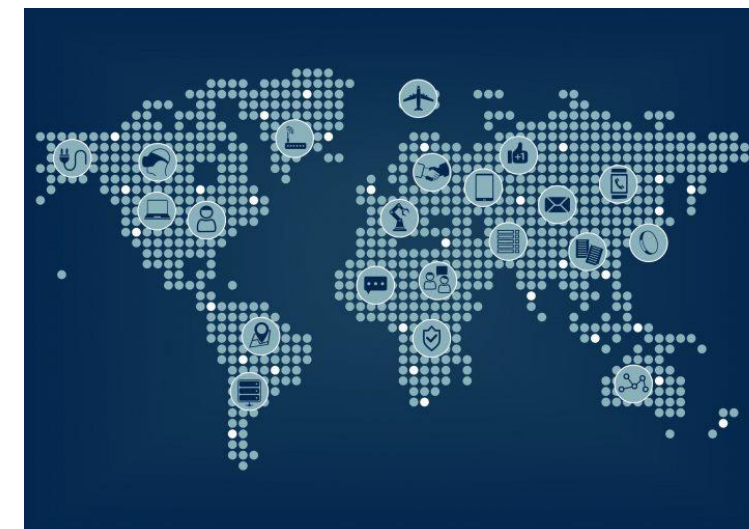
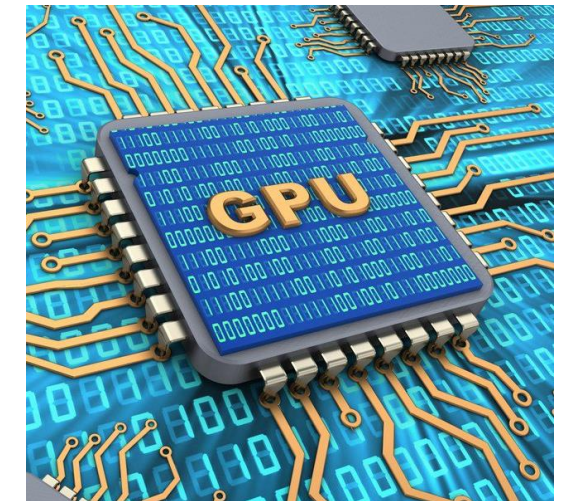
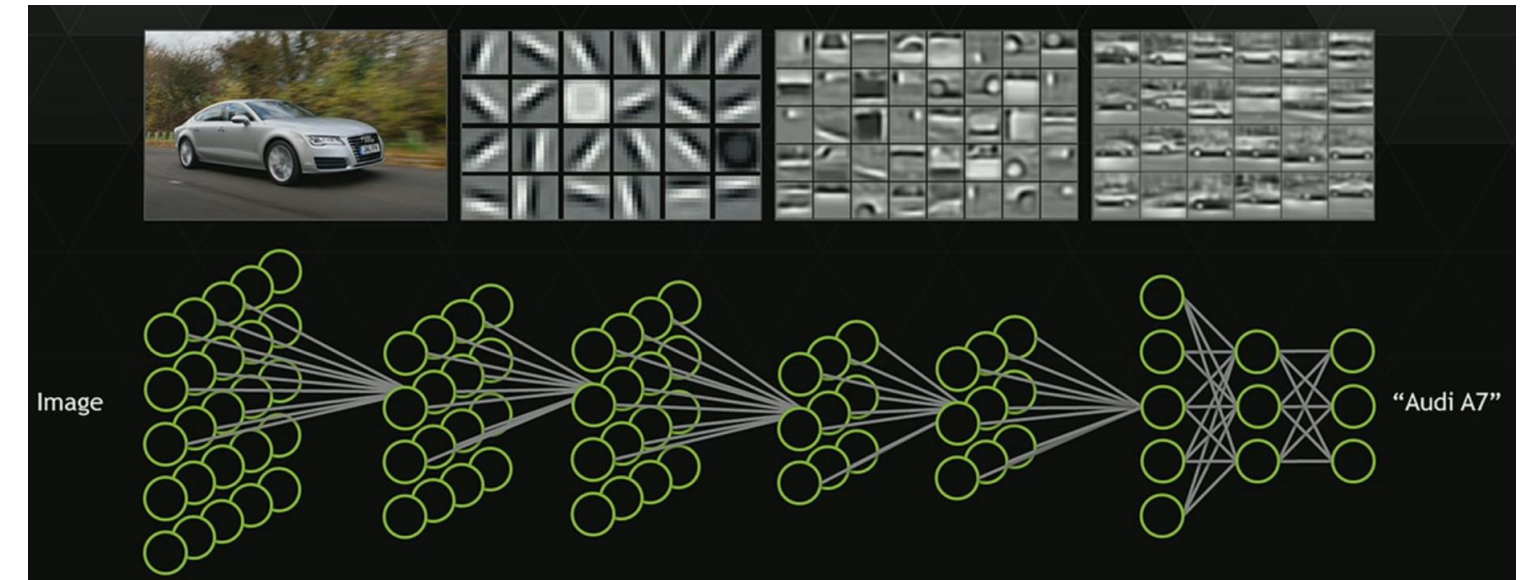
what is considered difficult for humans, tends to be easy for computers,
what seems easy to humans tends to be very challenging for computers

However, what seems easy to humans ...

- tends to be essential in life (survival of the fittest), ...
- ... and hence, has been perfected over millions years of evolution
- is practiced daily, ...
- .. and, did require a fair amount to learn

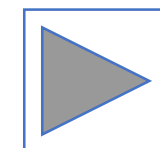
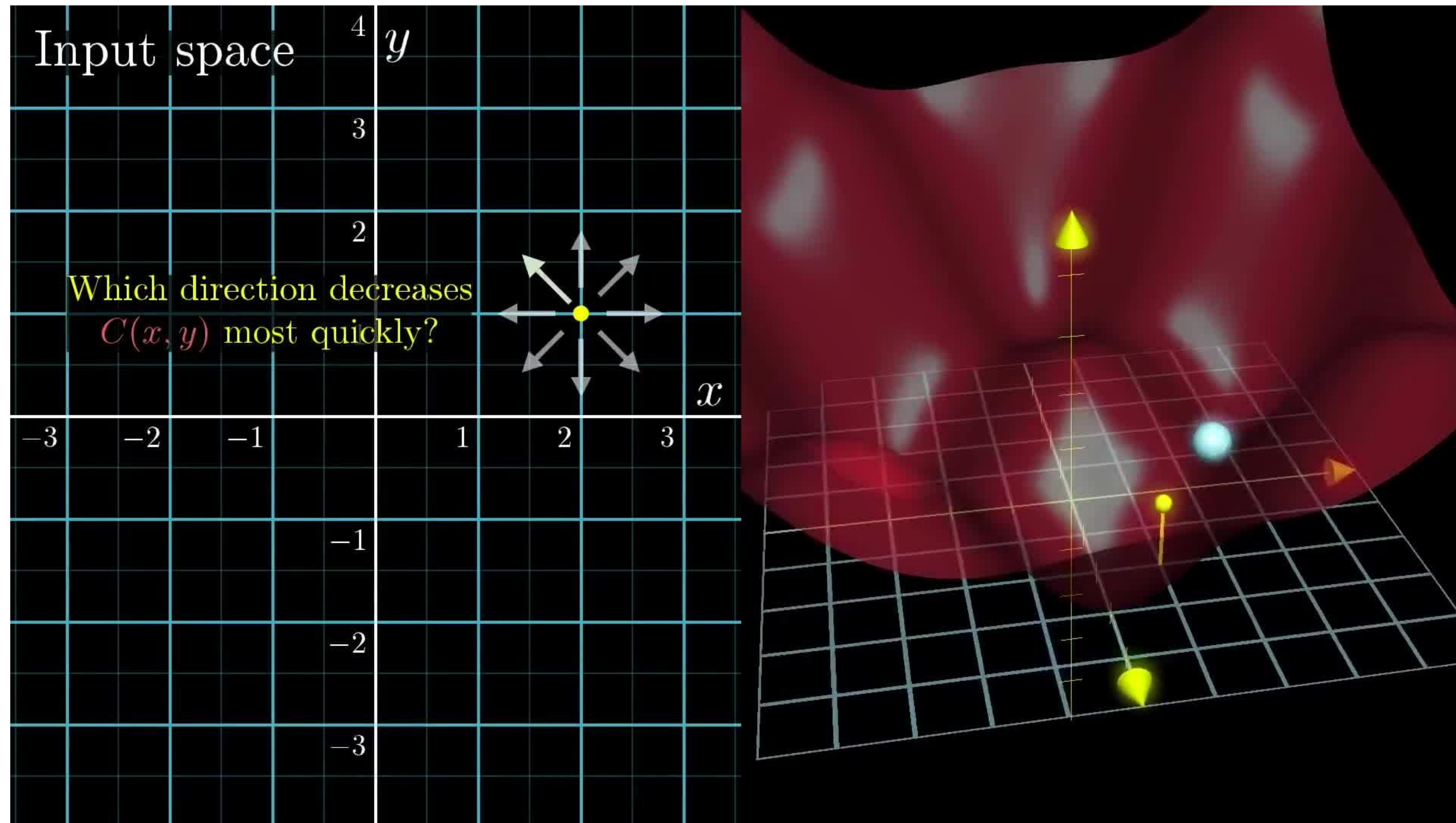
Deep learning & neural networks: what & why now?

- What1: learn from data (examples)
- What2: multiple layers;
automatically learns to decompose
the problem in sub-problems
- Why now1: compute power is now available,
thanks to 3D gaming (GPU)
- Why now2: easy access to large amounts
of data thanks to the Internet



Deep learning & neural networks: how

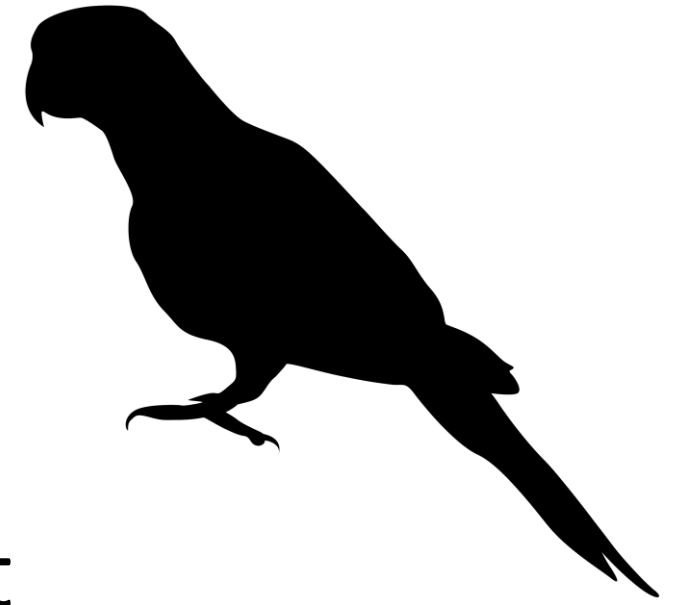
define cost function → math: search minimum (derivatives, matrices)



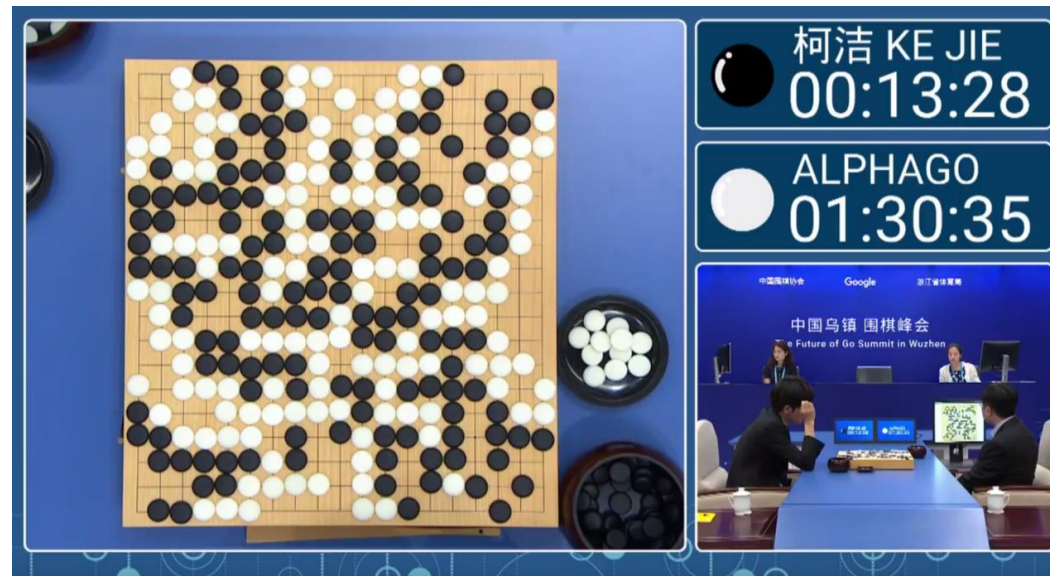
Deep learning & neural networks: cost function

straightforward for problems of the type

mimic this behaviour (reproduce)



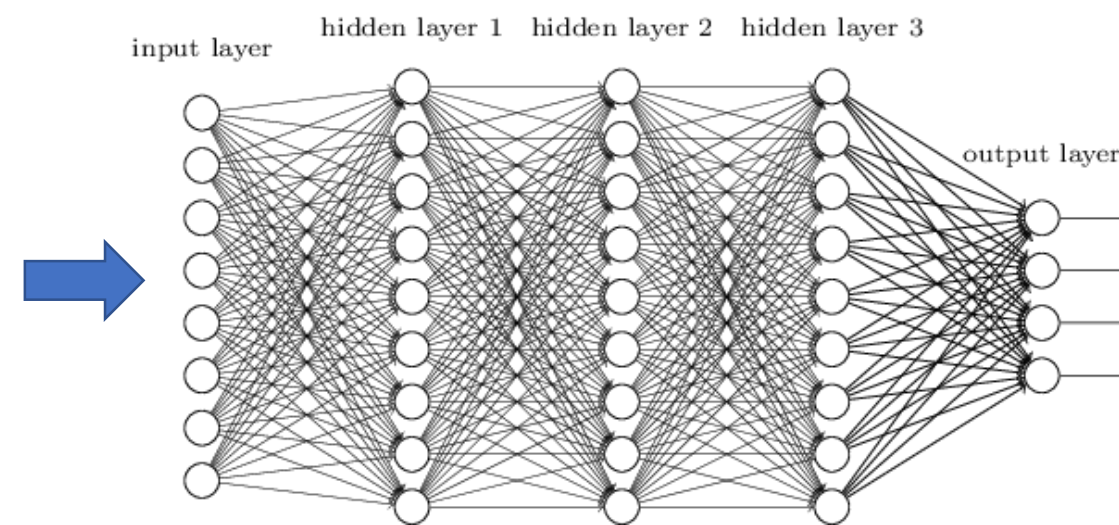
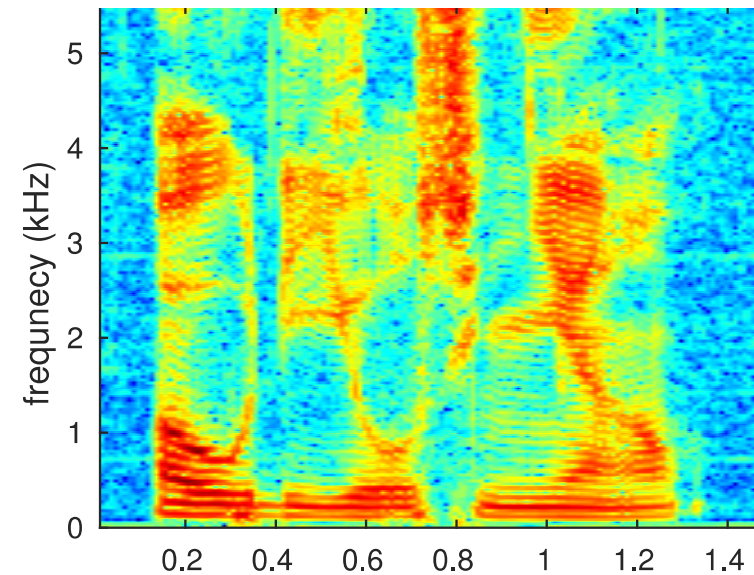
often a challenge to put other problems in this straitjacket



no one knows the cost function for true intelligence **(strong AI)**

Speech → Text (S2T)

easy cost function:



The quick brown fox jumps over the lazy dog.

main challenge: collecting 1000+ hours of annotated speech is costly

- easy:
- read speech: annotations are known (e.g. books on tape)
 - noisy speech: just add noise (lots of variation possible) to clean speech
- difficult:
- spontaneous speech, dialects, ...
 - costly to annotate
- correlate:
- works better for large and rich language groups
 - lagging for smaller or poor language groups
 - problems with dialects, spontaneous speech, ...

Text → Speech (TTS)

easy cost function:

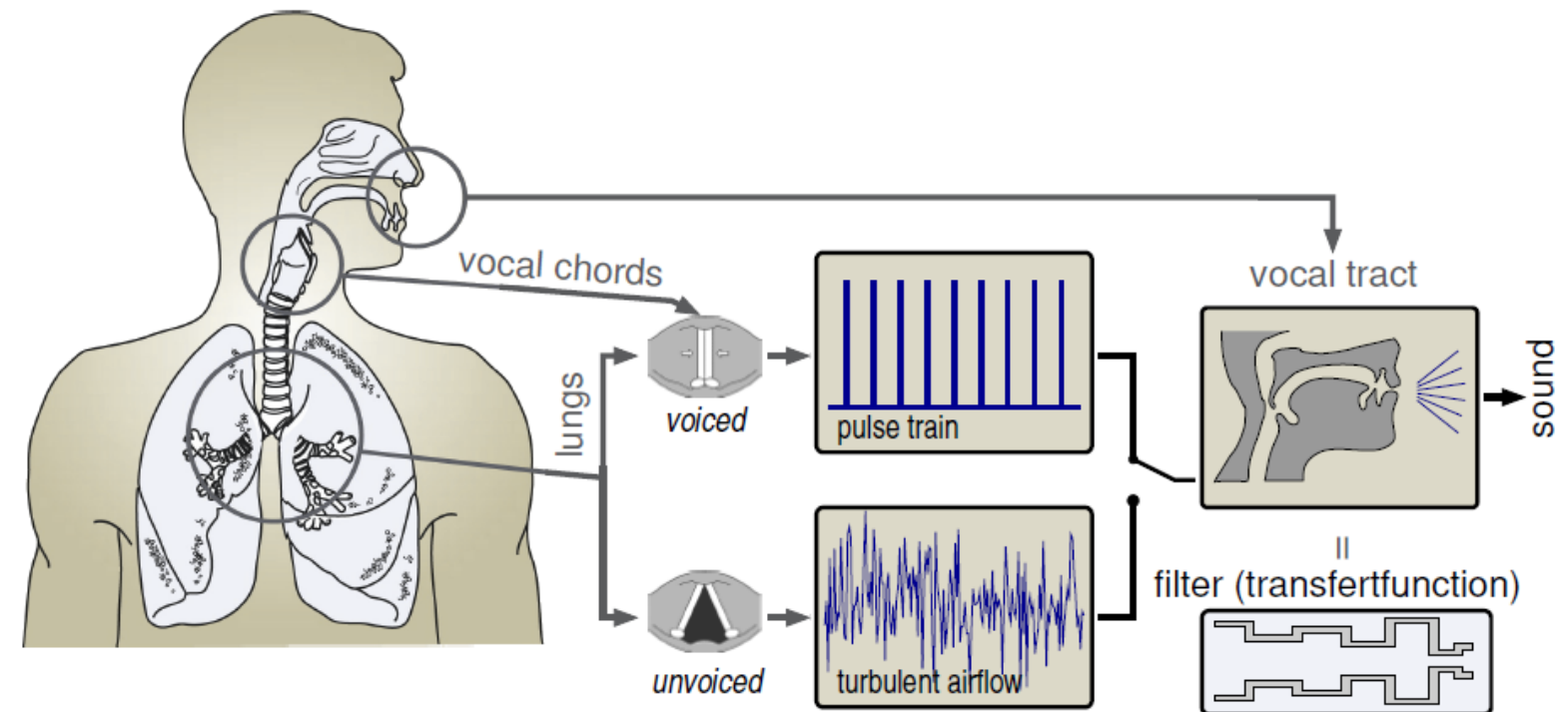


(Very) old approach: “engineer” a voice

- problem: sounds artificial
(no variation/emotions)

Modern approach:

- learn from data
- deep learning: split speaker characteristics (timbre, tempo, ...) and sound generation
 - localize speaker in a huge speaker space → easy to mimic a new speaker ([Lyrebird](#))
- is now a mature technology for most languages



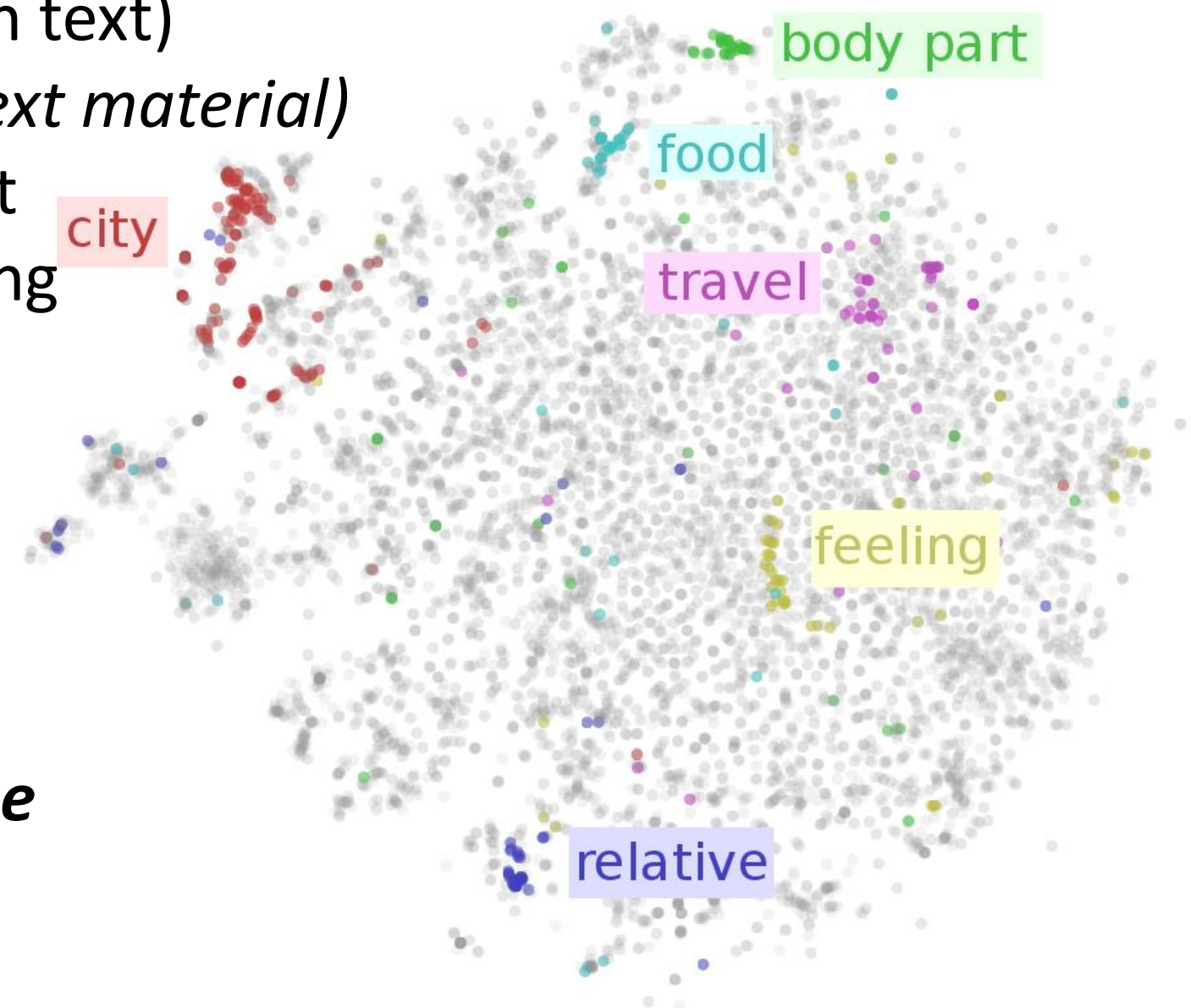
remaining challenges: natural speech with emotions, singing, ...

- collecting and annotating example data is difficult and expensive
- example: hire actors

Understanding the meaning of words, ... (NLP)

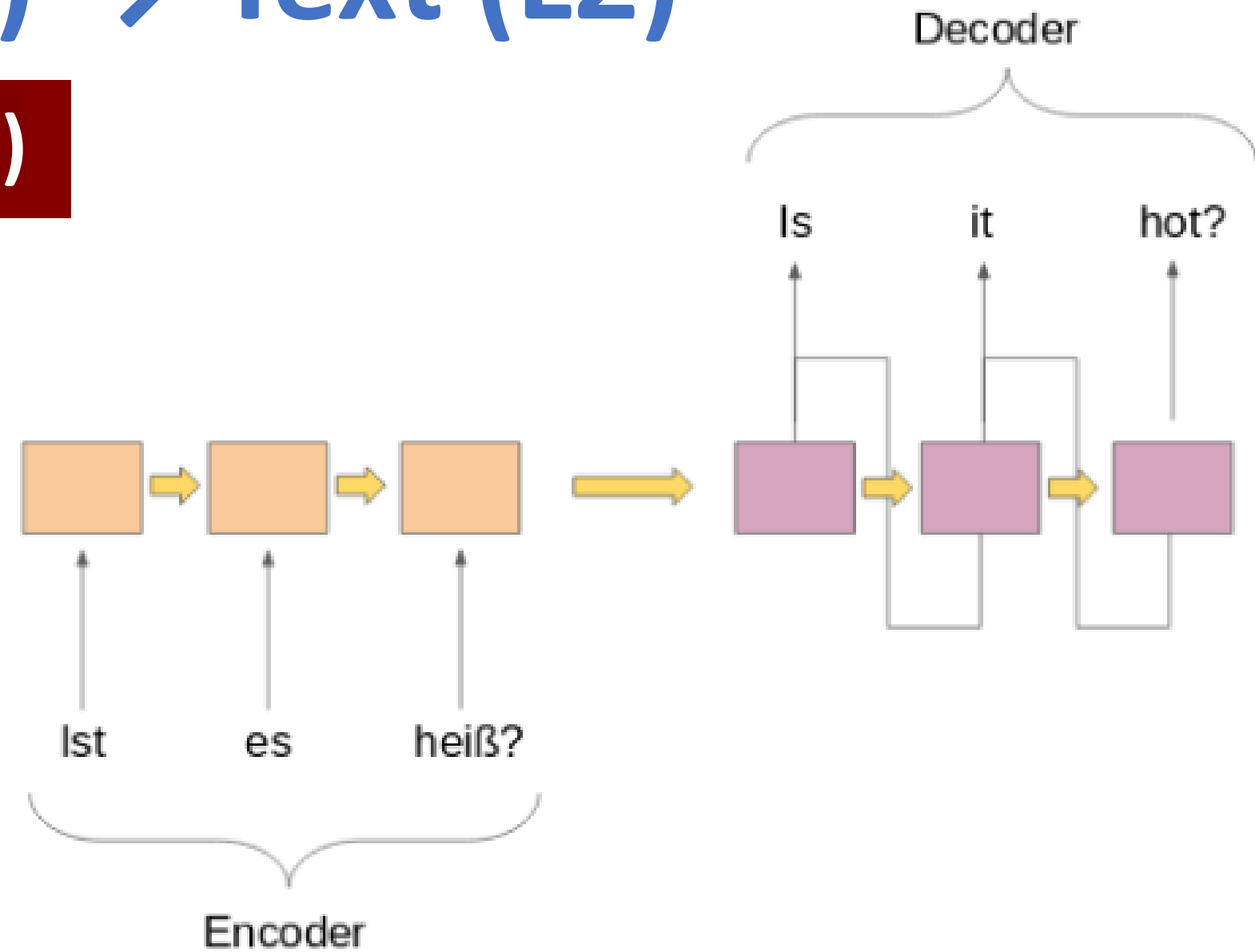
embeddings (same meaning → same location in a huge vector space)

- distributional hypothesis (linguistics):
"a word is characterized by the company it keeps" (Firth, 1950s)
- step1: collect very large text corpora (easy for written text)
(billions of words, i.e. multiple life-times of speech+text material)
- step2: cost function — items (words, phrases, ...) that surrounded by the same words have a similar meaning
→ *cost = distances in a high dim. vector space*
- work very good, even allows math on words:
$$\textit{king} - \textit{man} + \textit{woman} \approx \textit{queen}$$
- applying this to phrases, sentences, paragraphs,
→ *multi-layered approach (transformers)*
- **main challenge: spoken language ≠ written language**
 - *translation from spoken to written is needed*
 - *no (little) example data*



Translation: Text (L1) → Text (L2)

transform (encode → generate)



Dutch ▾

automatische vertaling Edit

Chinese (Simplified) ▾

自动翻译
Zìdòng fānyì

Speech (L1) → Speech (L2)

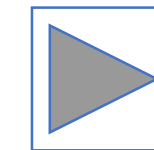
The screenshot shows a presentation slide titled "interACT Sprachbarrieren". On the left, a man in a blue suit stands at a podium in a conference room, addressing an audience. A speech bubble above him contains a string of non-standard characters: "êß*0vúbØijBA¬pysUêj} hÿ5≈fÄ<.y†ëœkû çOF°Ø□kô#â¯«Zeû". On the right, a computer interface displays the German text: "Und haben sozusagen führende, Arbeiten in Deutschland geleistet. Der nochmal man schaut im internationalen Umfeld ist das KIT trotzdem nicht die erste Wahl." Below this, it shows "German-speech" with a green checkmark and an "auto scroll" button. The English translation is displayed below: "And have done, so to speak, leading to work in Germany. The once again you look in the international environment is the KIT still not the first choice." Below the English text, it shows "English-speech" with a green checkmark and an "auto scroll" button. The slide footer includes logos for KIT (Karlsruhe Institute of Technology), MOBILE TECHNOLOGIES, and EU BRIDGE, along with the text "Introducing the First Simultaneous Translation Service by Computer".

Computer: phased approach

1. detect sentence end
2. speech recognition
3. translate sentence
4. speech synthesis

Reason:

handling complete sentences works better (in every step)



Human interpreter: anticipate what will be said

- speaker matches anticipation → low delay translation
- speaker deviates from anticipation → correction (cf. spontaneous speech)

More speech & language applications

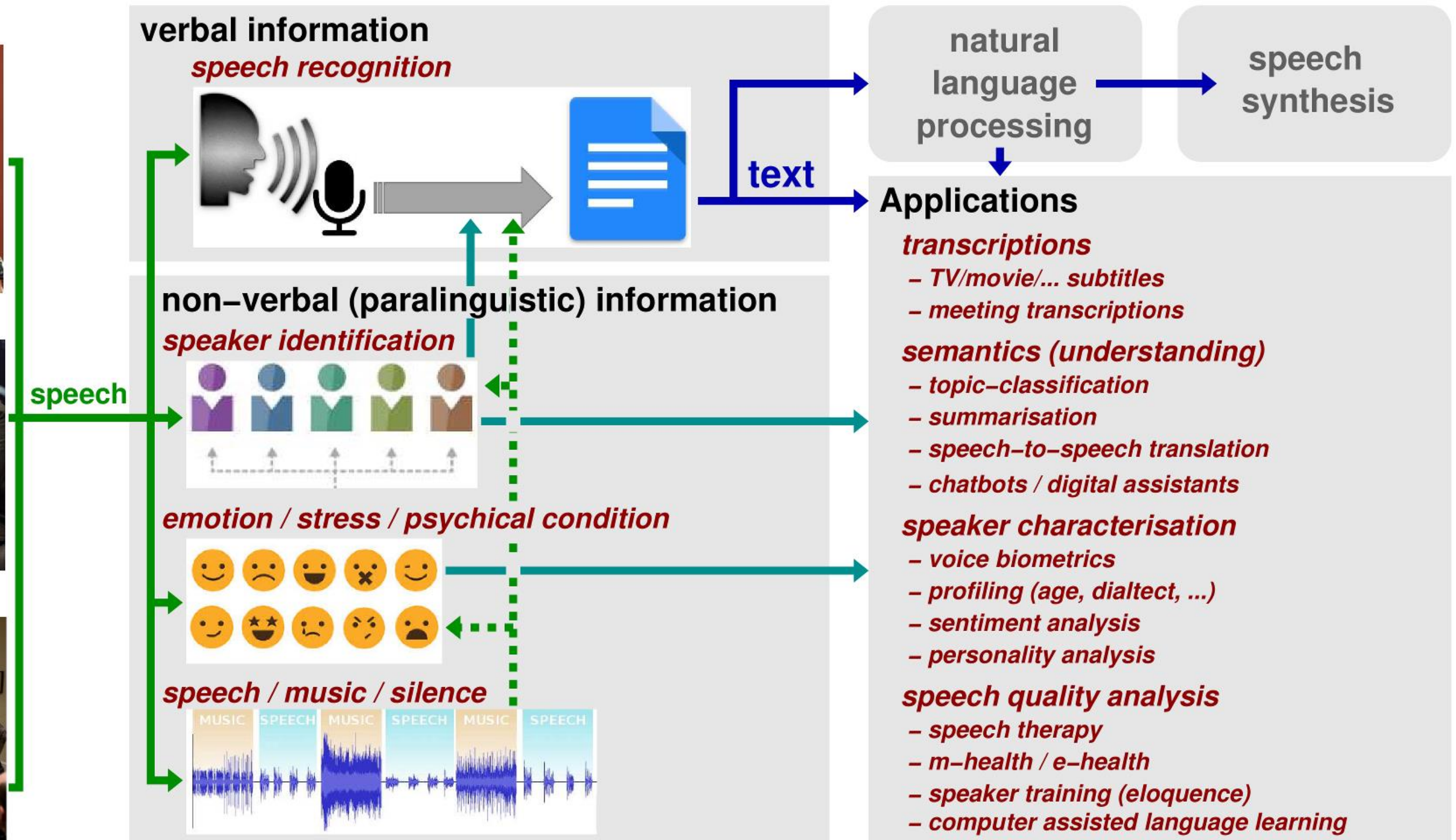
telephone



face-to-face



meetings



A list of technology providers can be found at: <https://cef-at-service-catalogue.eu/>

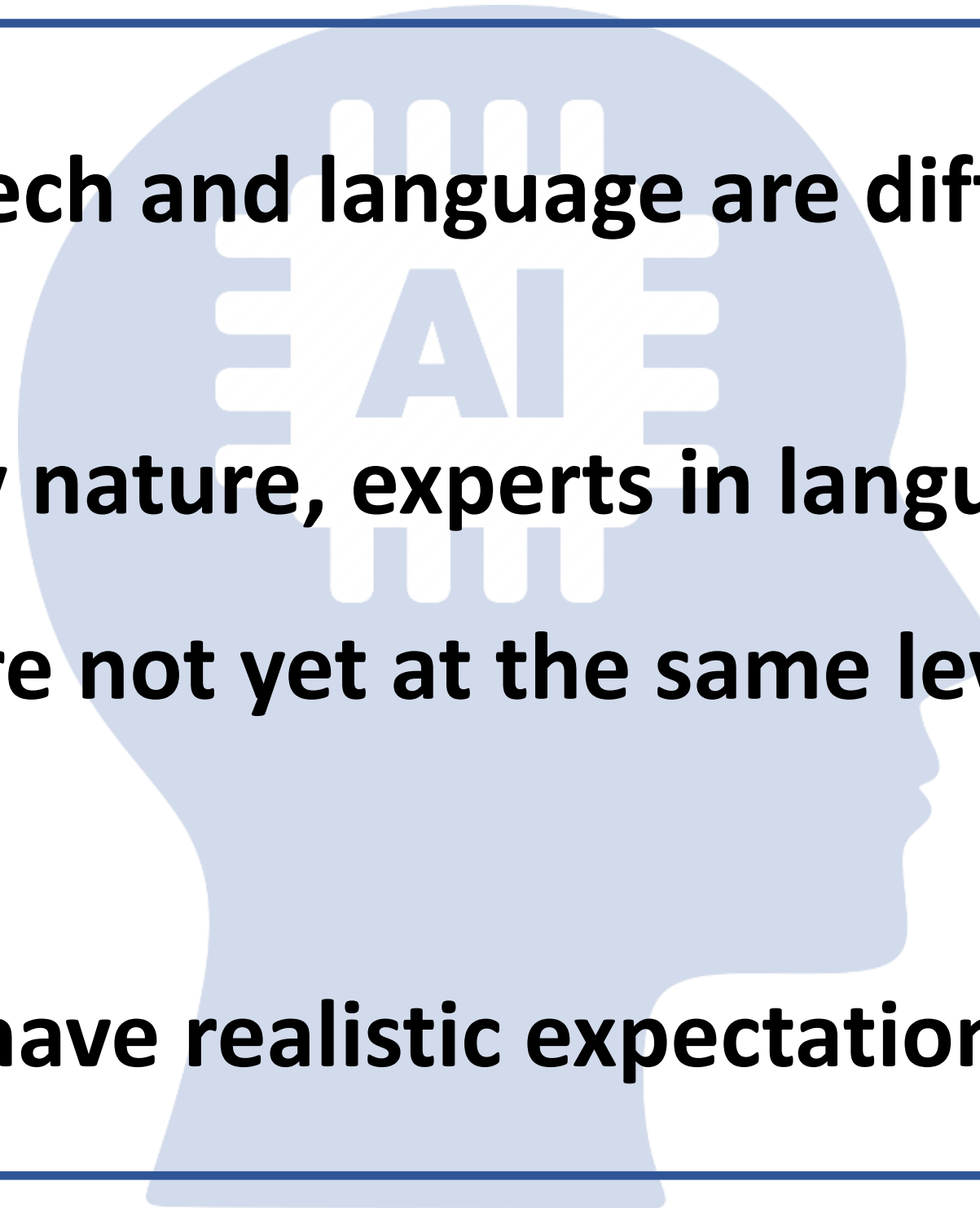
Conclusion (main take-away message)

speech and language are difficult

humans are, by nature, experts in language and speech

computers are not yet at the same level as humans

have realistic expectations



Links

- <https://www.youtube.com/watch?v=vjSohj-Iclc>
- <https://www.youtube.com/watch?v=9QLNutP-rNo&t=381>
- <https://www.youtube.com/watch?v=IHZwWFHWa-w?t=416>
- <https://www.youtube.com/watch?v=tclQEigamq4>
- <https://cef-at-service-catalogue.eu/>

