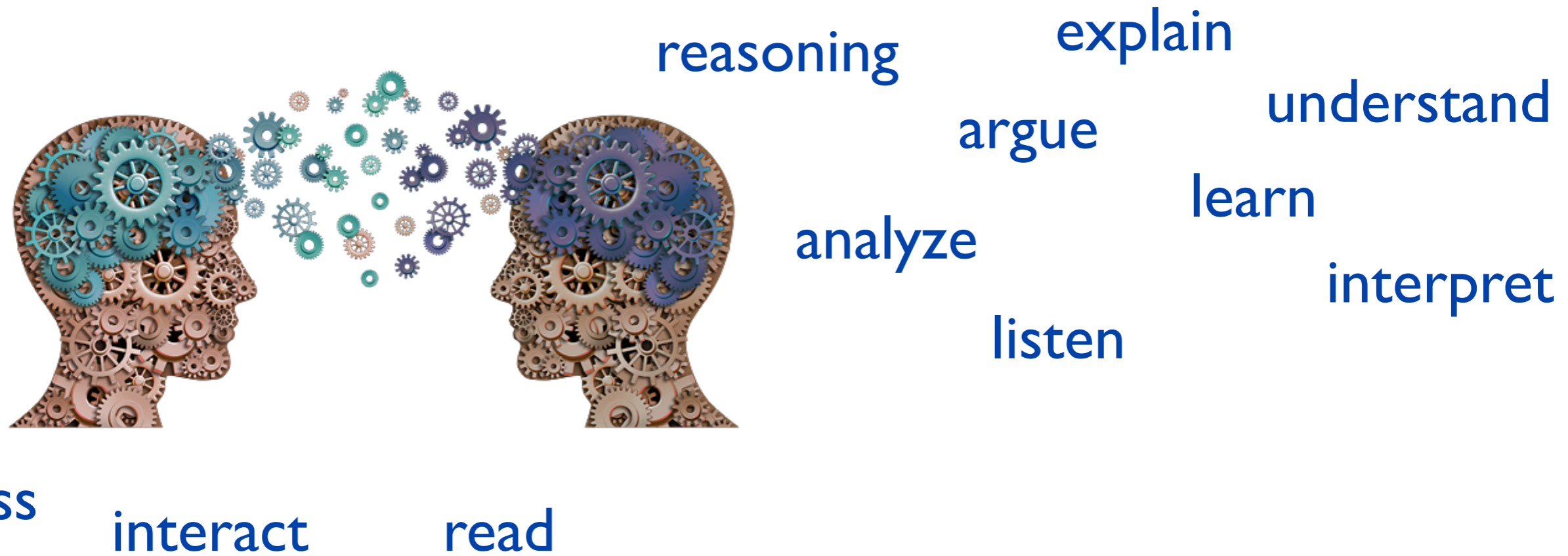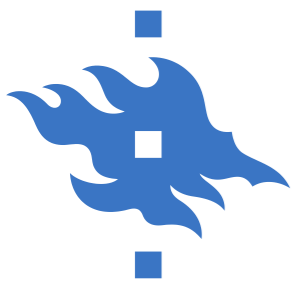# *The potential of Language Technology and AI –*

where we are, where we should be heading
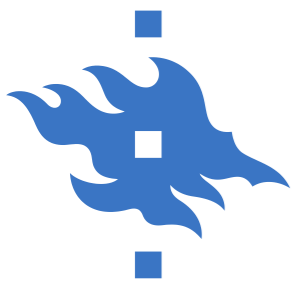
Jörg Tiedemann
Department of Digital Humanities
University of Helsinki
jorg.tiedemann@helsinki.fi
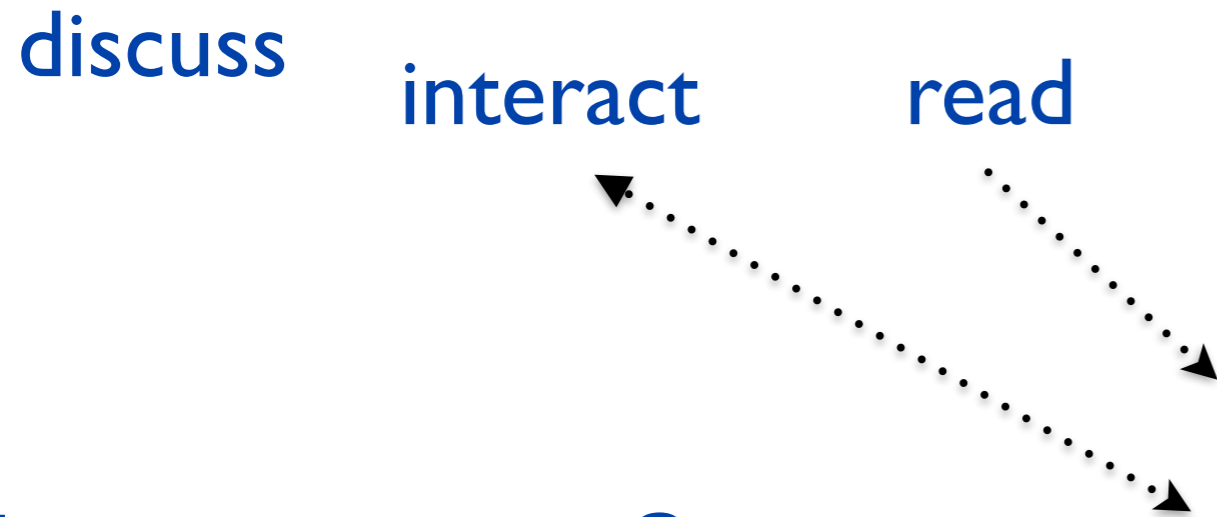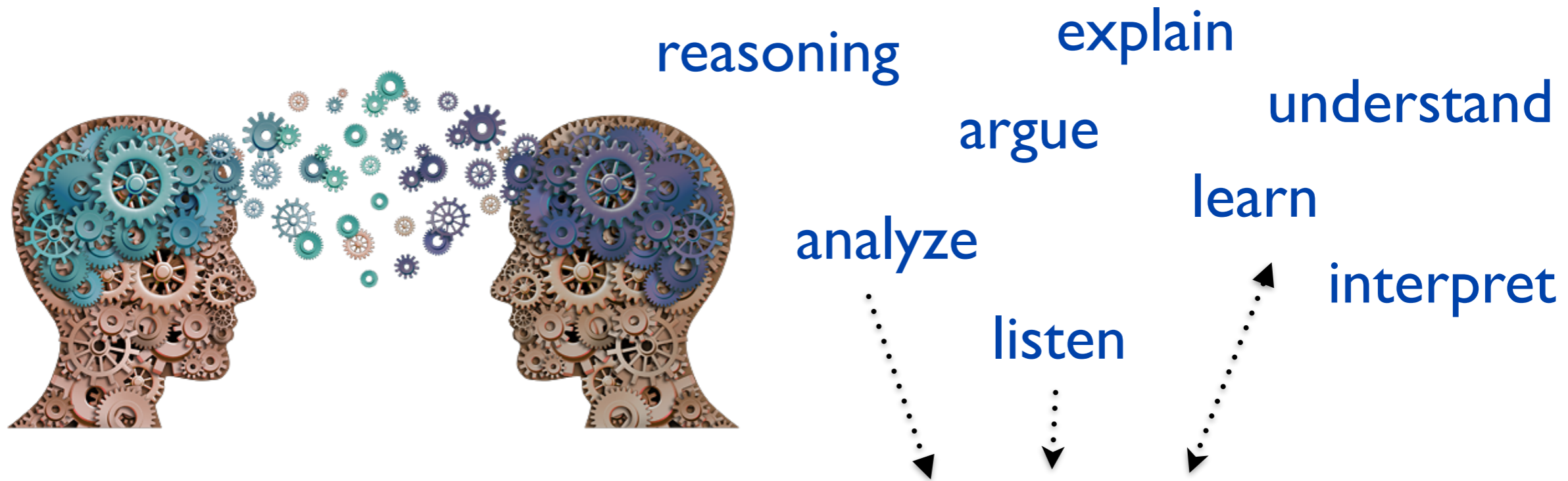
Jörg Tiedemann
Department of Digital Humanities
University of Helsinki
jorg.tiedemann@helsinki.fi

How do humans become intelligent?

reasoning

explain

argue

understand

analyze

learn

interpret

listen

discuss

interact

read

How do **machines** become intelligent?

reasoning

explain

argue

understand

analyze

learn

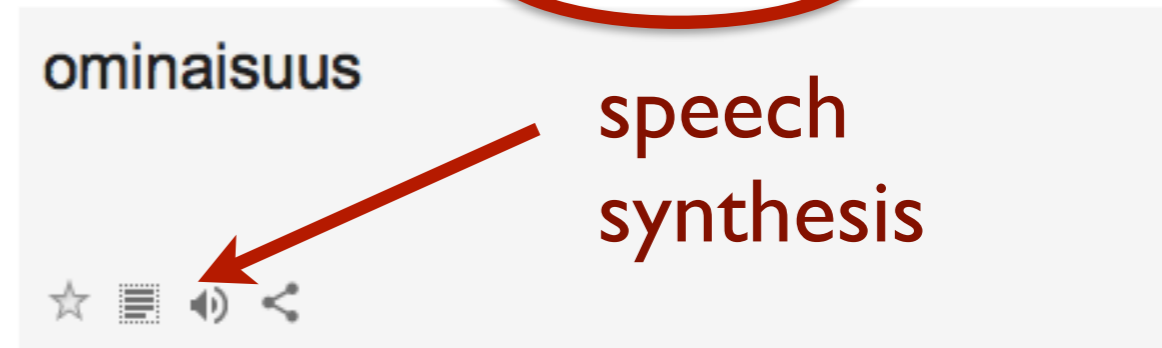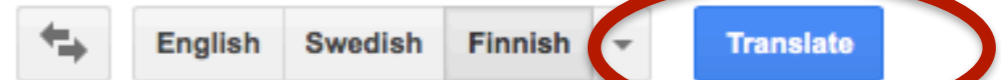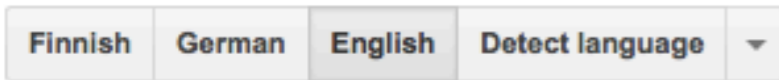interpret

listen

discuss

interact

read

Language & Speech Technology

LT is already all around us

sentiment analysis
spam detection

language identification

recommendation systems

translation

| Finnish | German | English | Detect language | ▼ |

| ⇄ | | English | Swedish | Finnish | ▼ | **Translate** |

atribute ✕

ominaisuus

handwritten text
recognition

speech
synthesis

🔊 ✏️ ▼

☆ ▤ 🔊 ⊰

Showing translation for *attribute*
Translate instead atribute

speech
recognition

Definitions of attribute

noun

a quality or feature regarded as a characteristic or inherent part of someone or something.
"flexibility and mobility are the key attributes of our armed forces"
synonyms: quality, characteristic, trait, feature, element, aspect, property, sign, hallmar...

a piece of information that determines the properties of a field or tag in a database or a
string of characters in a display.
"The destination of the alias is stored in the maildrop attribute"

verb

regard something as being caused by (someone or something).
"he attributed the firm's success to the efforts of the managing director"
synonyms: ascribe to, assign to, attach to, credit to ...
▶ See more definitions

atribute

| ⎵ attribute | ⎵ atribute | ⎵ attributes | ⎵ attributee | a |
| ⌫ | | English | ↵ |

Translations of attribute

noun

▬▬ attribuutti   attribute,
▬▬ ominaisuus   feature, p
▬ tuntomerkki   attribute

verb

▬ lukea jkn ansioksi   attribute
▬ lukea jkn syyksi   impute, a
▬ sanoa jklle kuuluvaksi attribute

spelling correction
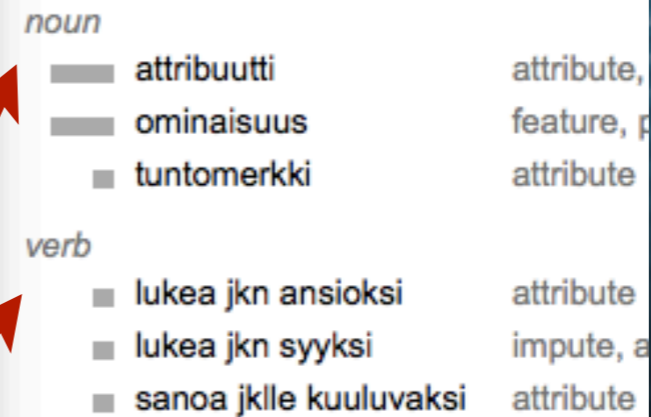
term definition and
disambiguation

linguistic analysis

lexical info

chatbots

❝ On October 18 remind
me it's dad's birthday ❞

Here's your reminder for
October 18, 2012 at 9 am:

18 Thursday
    October 2012

☐ Dad's birthday
  10/18/12, 9:00 AM

Cancel   Confirm

Traditional approaches

**single purpose /
theory-driven /
modular**

apply &
evaluate

task-specific model

annotated
data

learning
algorithms

theory-specific
rules

raw
data

Traditional approaches

**single purpose /
theory-driven /
modular**

apply &
evaluate

my favourite sub-task
(syntactic parsing, …)

task-specific model

for project-
internal use
only
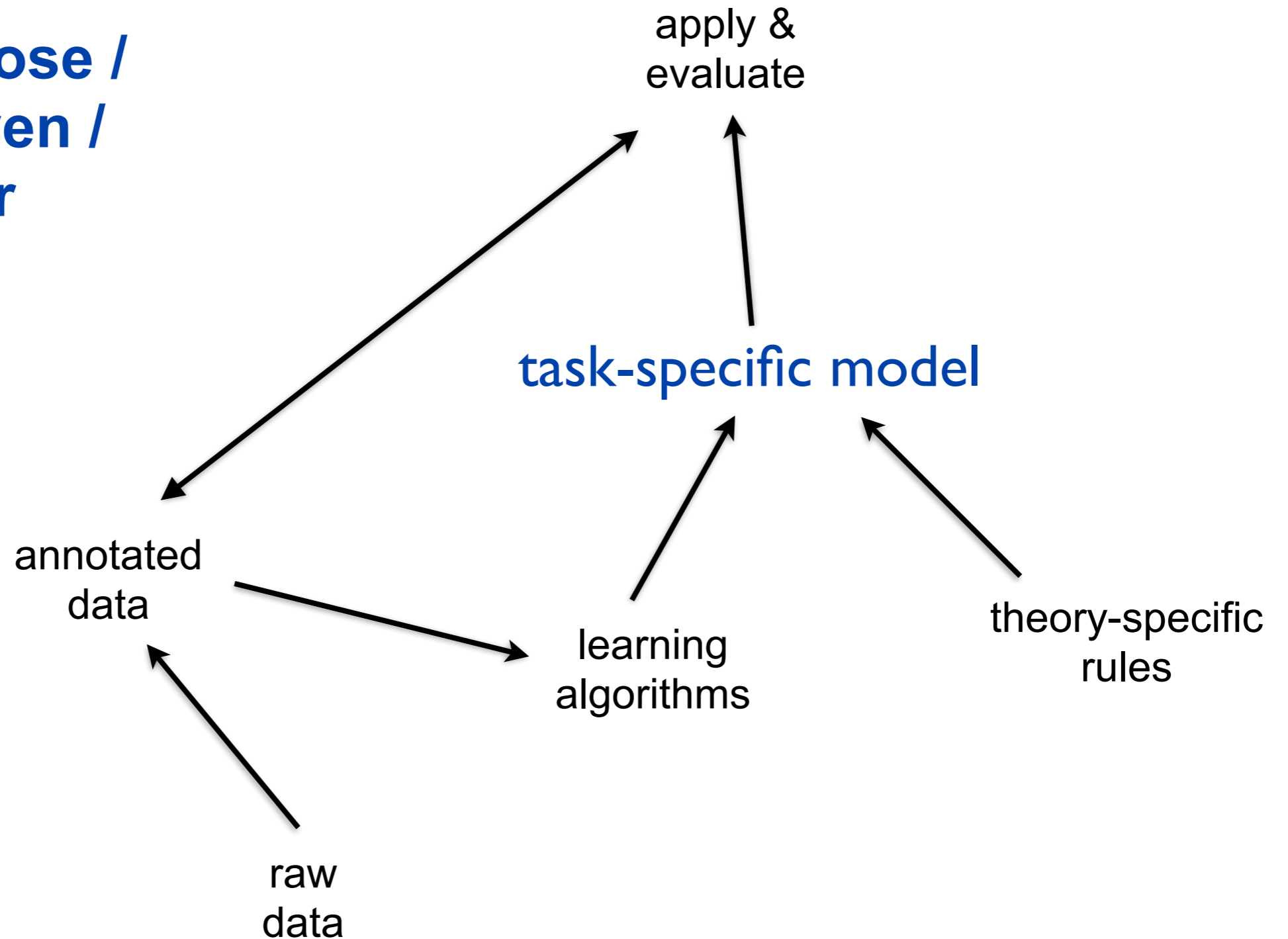
my favourite
theoretical
framework

annotated
data

learning
algorithms

theory-specific
rules

my favourite language
(often English only)

raw
data

my favourite modality
(typically text only)

But language is a multi-purpose tool

complex problems

communication is multimodal and interactive

languages are dynamic and not independent of each other
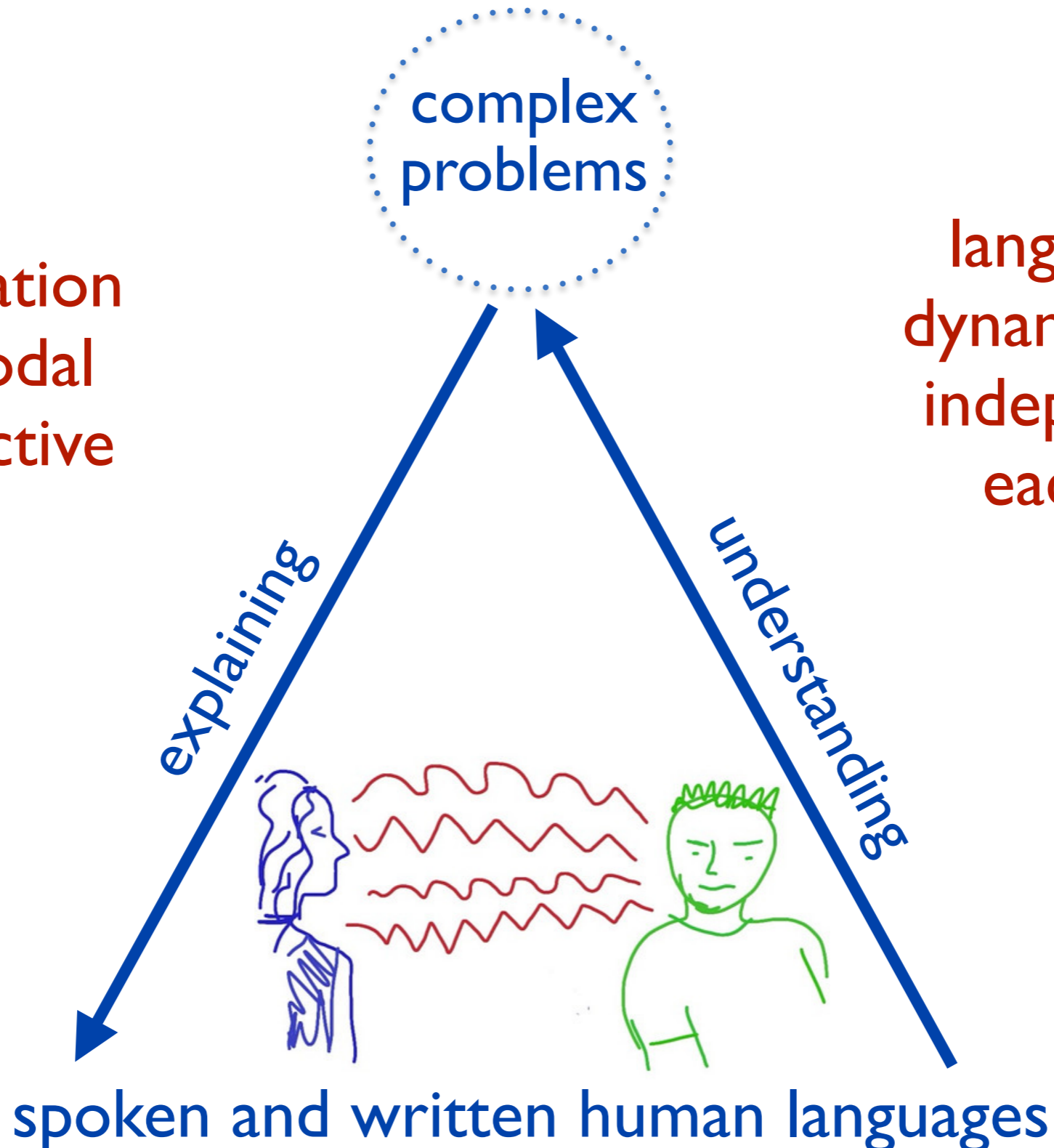
explaining

understanding

spoken and written human languages
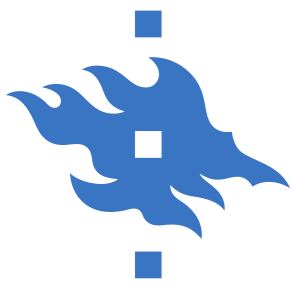
Current trend in language technology

**multi-purpose /
data-driven /
end-to-end**

apply &
evaluate

task-specific model

(big fat) language model

annotated
data

auto-encoder
de-noising, …

raw
data

Current trend in language technology

**multi-purpose /
data-driven /
end-to-end**

a battery of
benchmarks

apply &
evaluate

task-specific model

fine-tuning
transfer learning
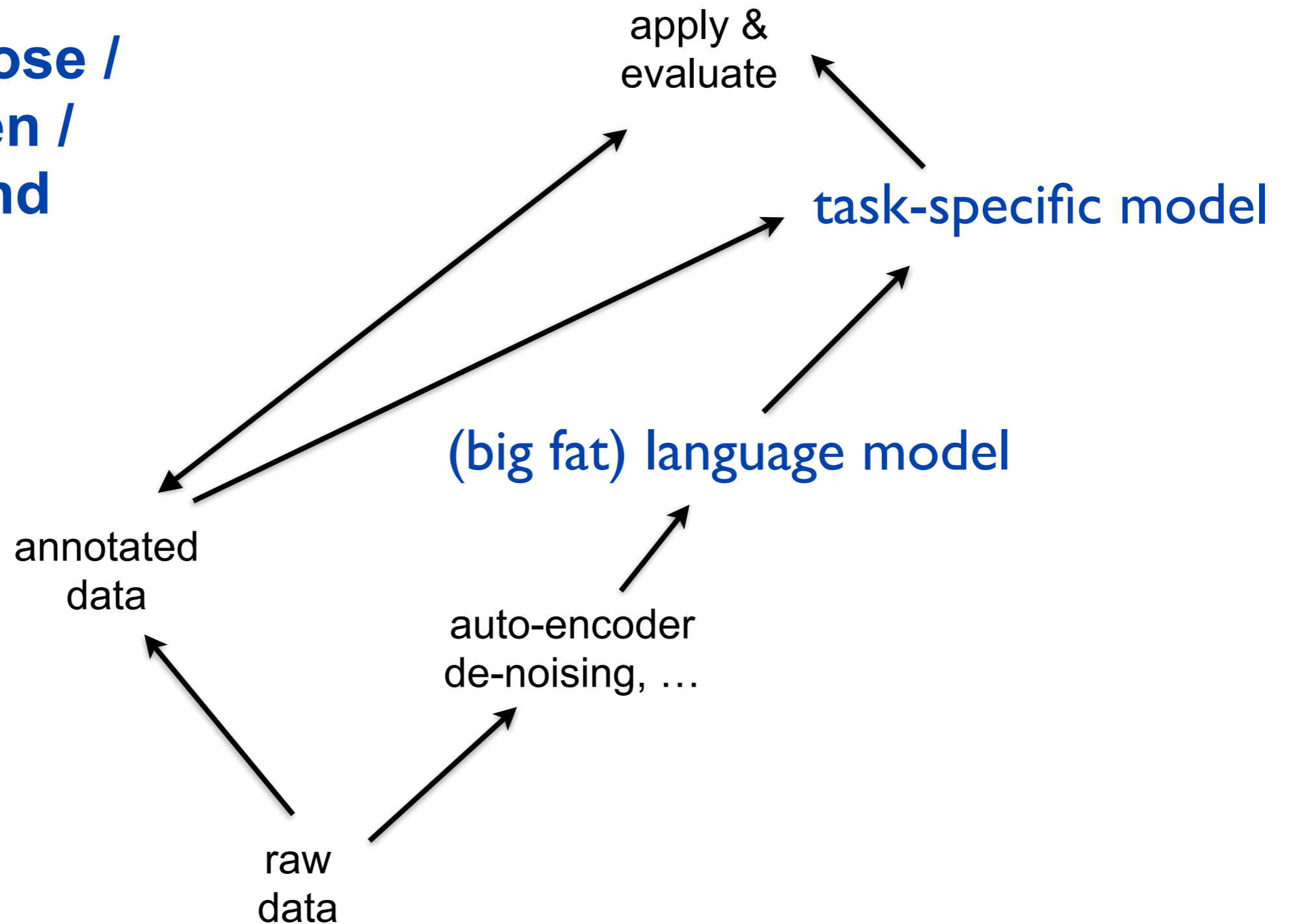
data augmentation
knowledge distillation
self-supervision

(big fat) language model

public release

annotated
data

auto-encoder
de-noising, …

unsupervised pre-training
multi-task training
life-long learning

multi-modal
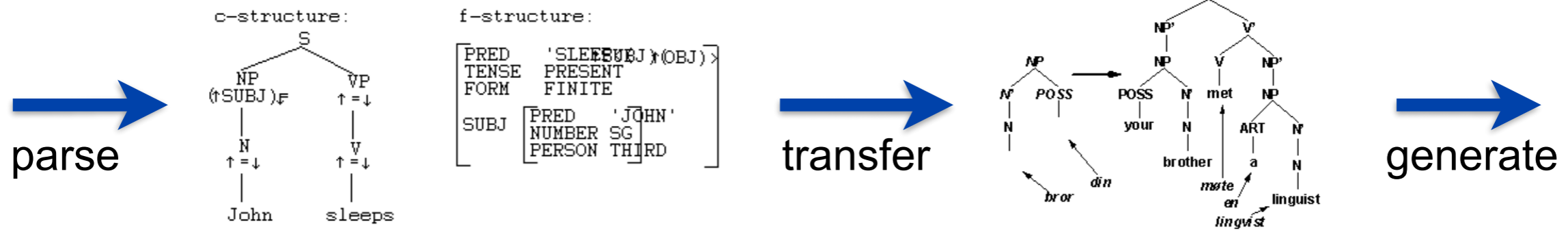multi-lingual
multi-domain

raw
data

Natural language understanding & LT
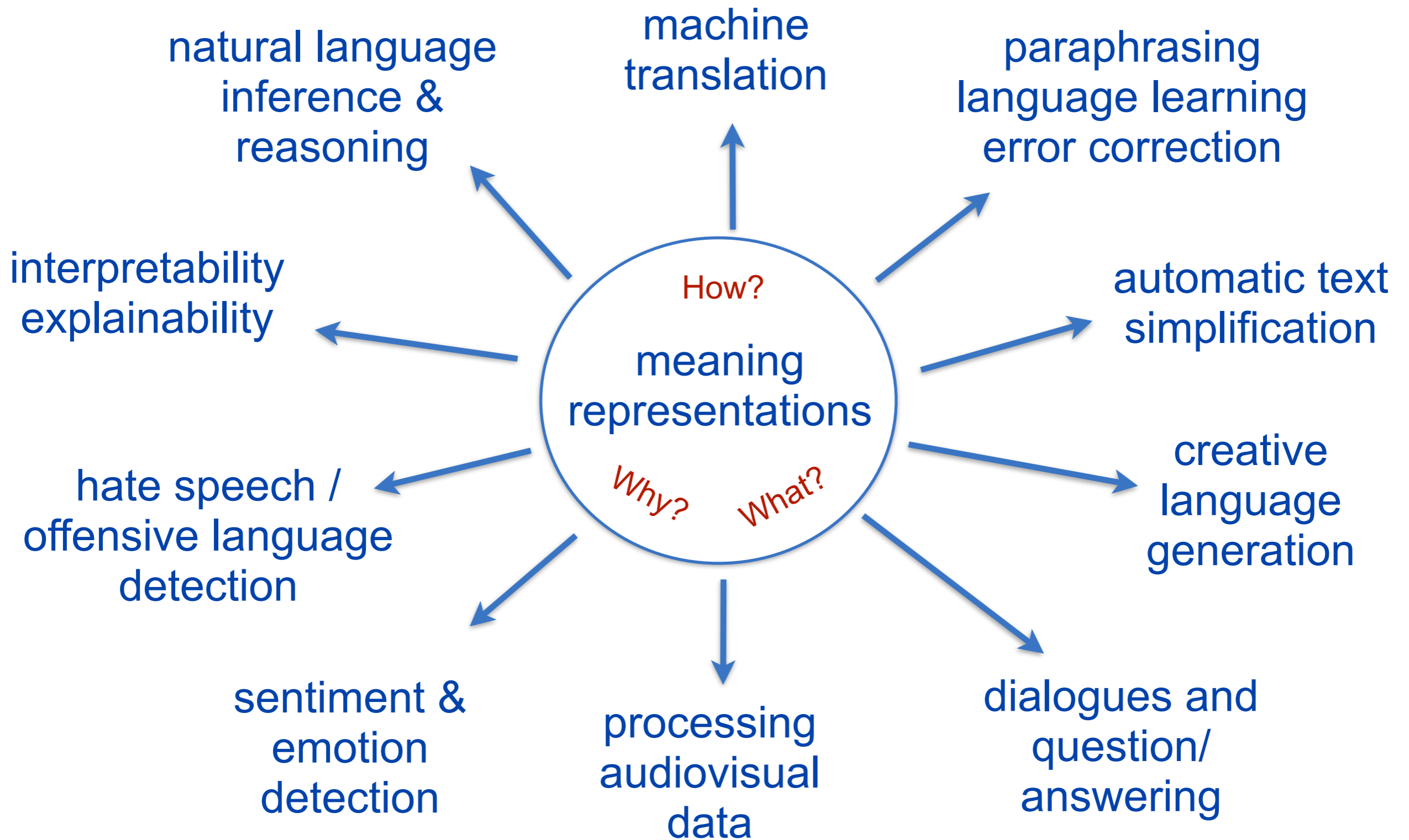
The case of machine translation

Transfer-based MT: rule-based, theory-driven
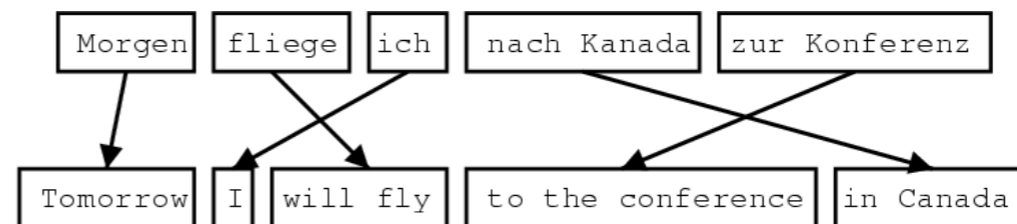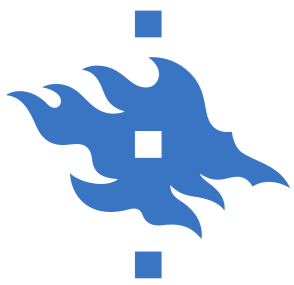

parse → transfer → generate

Statistical MT: rule-based, data-driven

alignment-based rule-extraction
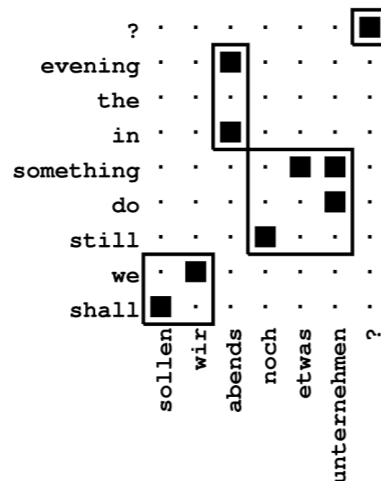
probabilistic decoding

The case of machine translation

meaning

understanding

generating

latent semantic representation

encoder

decoder

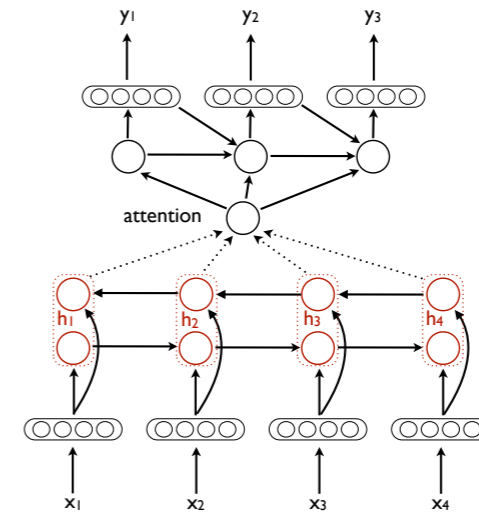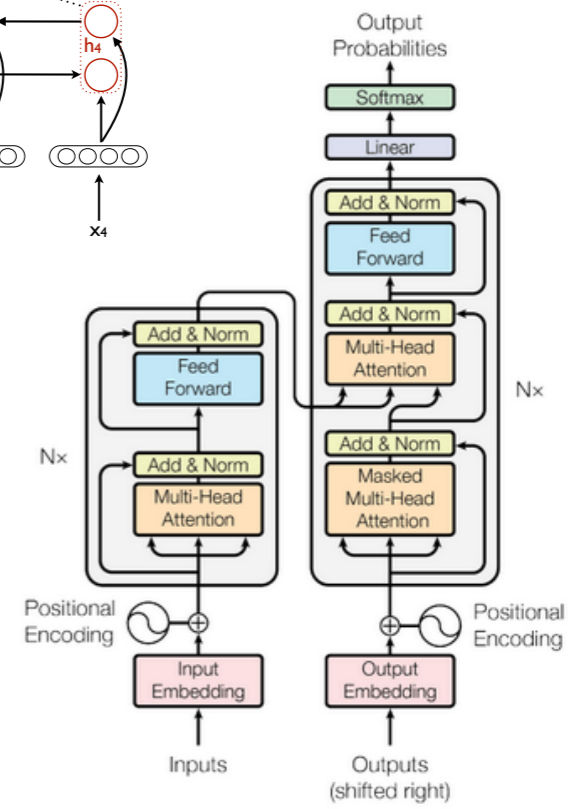source language

neural machine translation
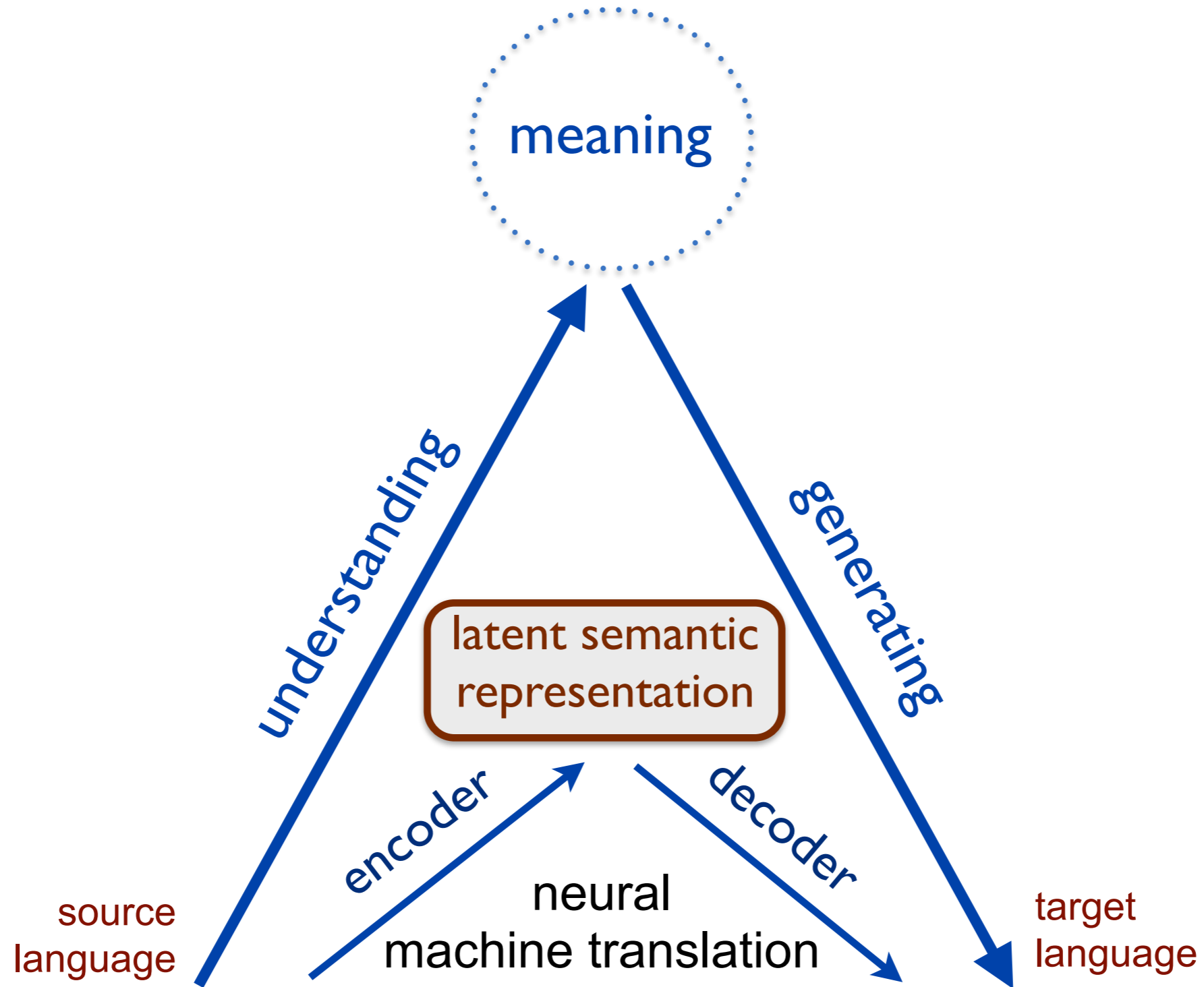
target language

The case of machine translation
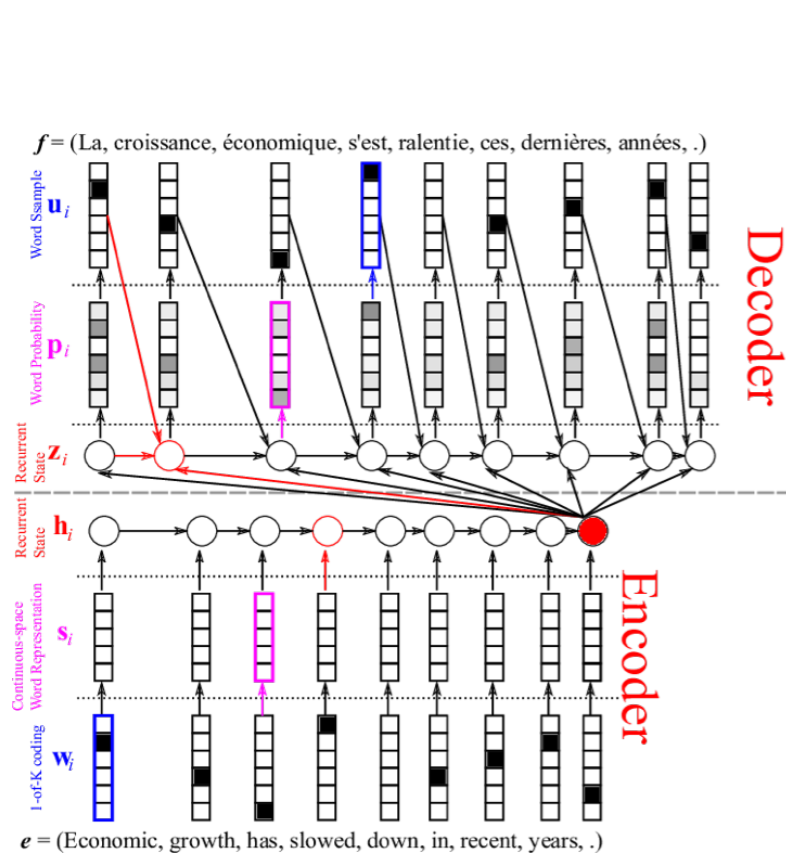
Recurrent NNs

attention

transformers

meaning

understanding

generating

latent semantic representation

encoder

decoder

neural
machine translation

source
language

target
language

Multilingual, multimodal, multitask ...

NN Decoders:

transfer /
zero-shot /
few-shot
learning

EN   FI   CZ

universal meaning representation

NN Encoders:

EN   FI   CZ   audio   video   ...
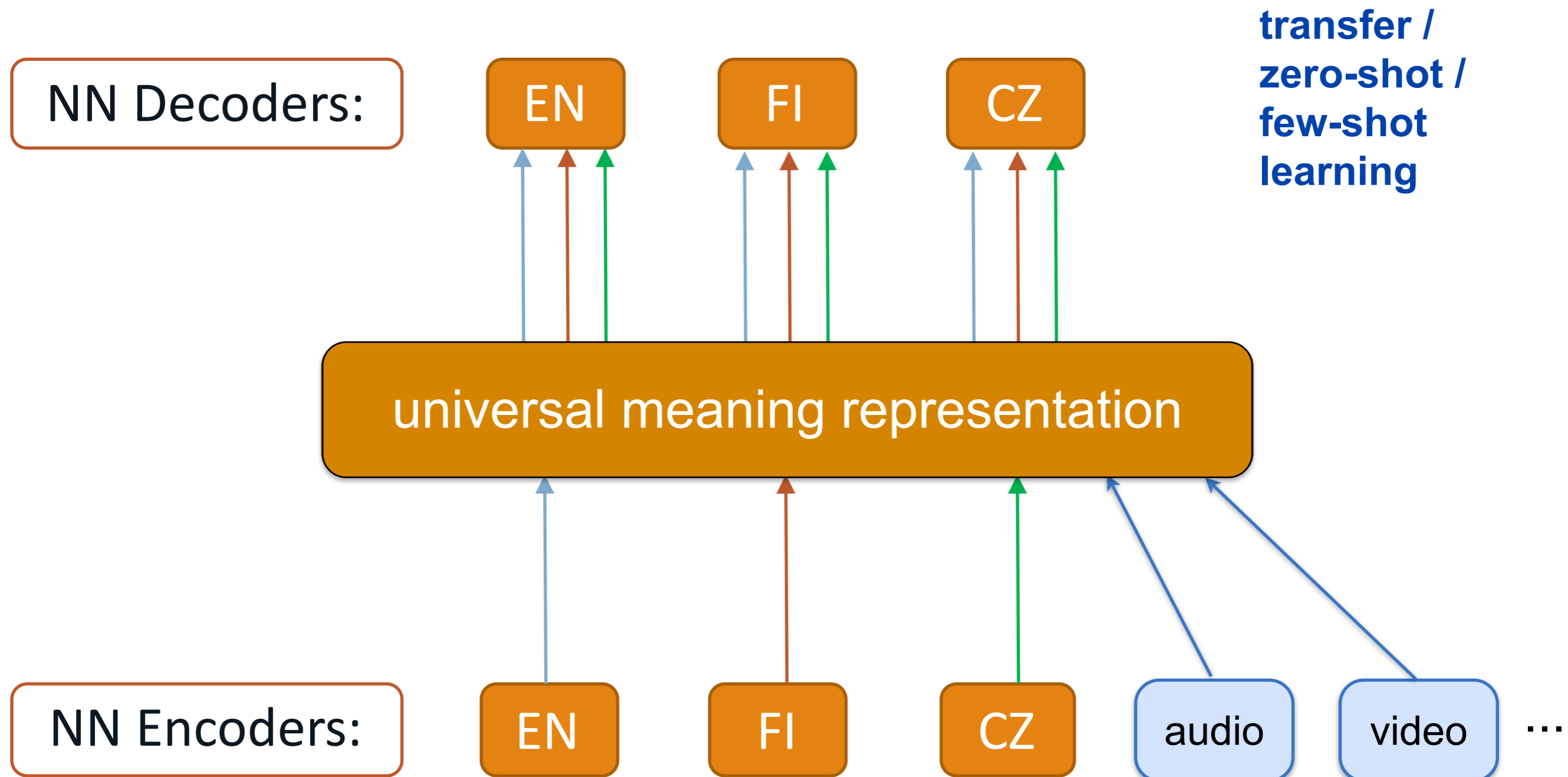
MeMAD: Access to audio-visual content

automatic transcription of videos accessible in many languages
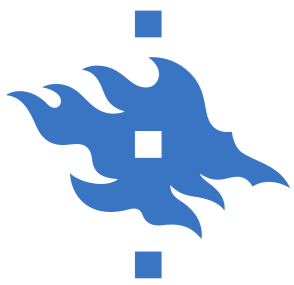
https://memad.eu

MeMAD
Methods for Managing Audiovisual Data

The importance of data

Neural models are data hungry

- millions of sentence pairs required for good performance
- serious disadvantage for stakeholders without data access
- many under-resourced languages and domains

The importance of data

Neural models are data hungry

- millions of sentence pairs required for good performance
- serious disadvantage for stakeholders without data access
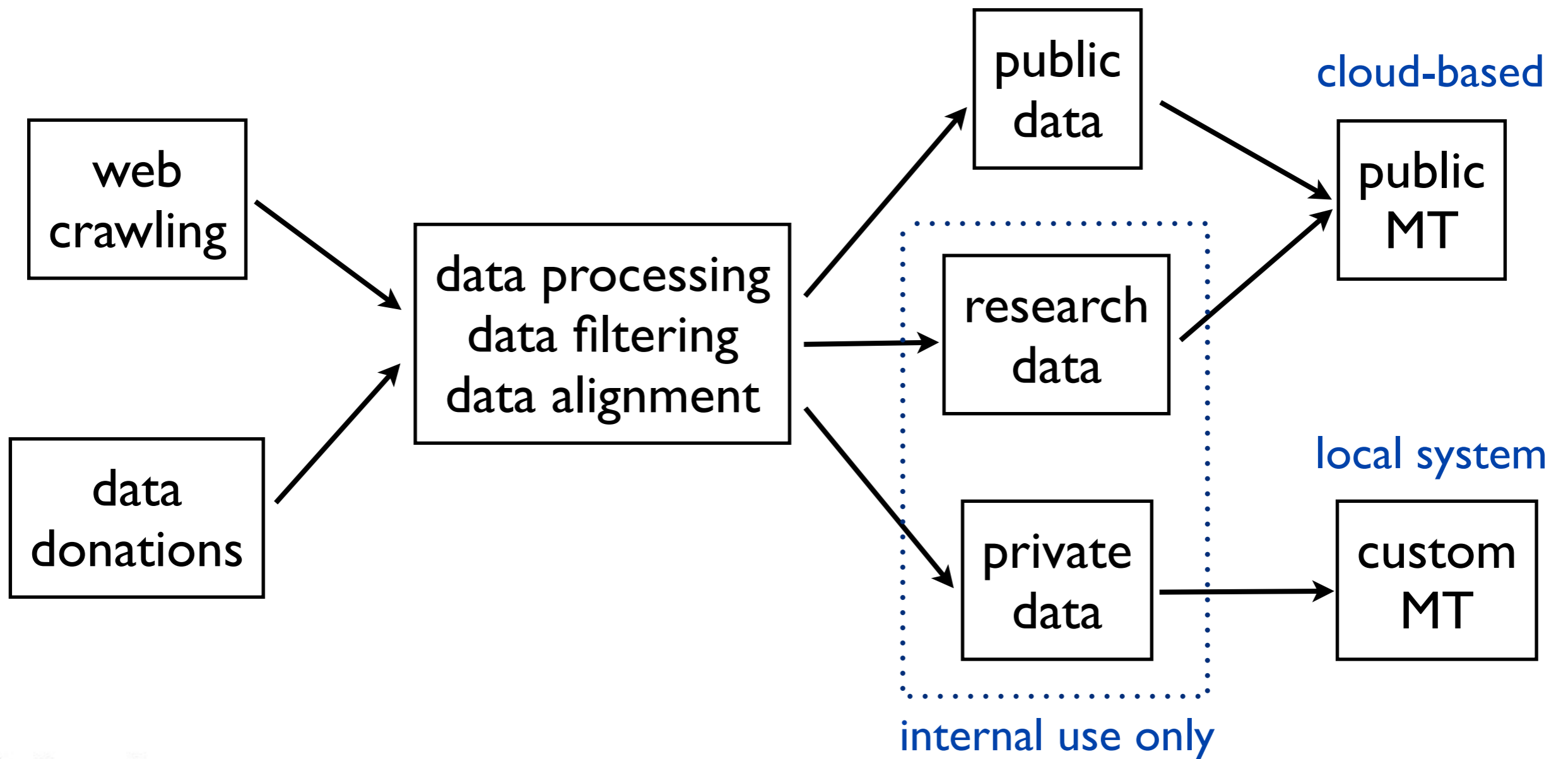- many under-resourced languages and domains

Positive developments

- increasing number of public data sets
- open data and data sharing policies
- more efficient transfer learning and data augmentation
- release of models and tools

finsk-svensk korpus & maskinöversättning

https://blogs.helsinki.fi/fiskmo-project/

web crawling

data donations

data processing
data filtering
data alignment

public data

research data

private data

cloud-based

public MT

local system

custom MT

internal use only

Svenska kulturfonden

OPUS mt

OPUS-MT at huggingface
https://huggingface.co/Helsinki-NLP

🤗 HUGGING FACE

⬆ Back to all models
Model: `Helsinki-NLP/opus-mt-ROMANCE-en`

pytorch rust marian lm-head seq2seq roa en translation

Hosted inference API ⓘ
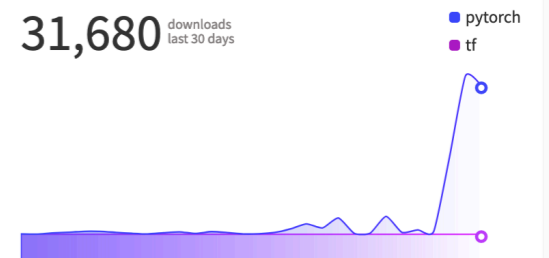
translation

Your sentence here...

Compute

This model can be loaded on the Inference API on-demand.

EUROPEAN LANGUAGE GRID

https://github.com/Helsinki-NLP/Opus-MT

Monthly model downloads

🌐 Helsinki-NLP/opus-mt-ROMANCE-en

31,680 downloads last 30 days

● pytorch
● tf

```
{
    "alignment": [
        "0-0 0-2 1-1 2-3",
        "0-0 1-1 3-2 4-3 5-4"
    ],
    "result": "How are you? The translation is fun.",
    "server": "192.168.1.18:20001",
    "source": "fi",
    "source-segments": [
        "Mit\u00e4 kuuluu ?",
        "K\u00e4\u00e4@@ nn\u00f6@@ s on hauskaa ."
    ],
    "source-sentences": [
        "Mit\u00e4 kuuluu?",
        "K\u00e4\u00e4nn\u00f6s on hauskaa."
    ],
    "target": "en",
    "target-segments": [
        "How are you ?",
        "The translation is fun ."
    ],
    "target-sentences": [
        "How are you?",
        "The translation is fun."
    ]
}
```

OPUS-CAT - plugins and local MT engines
https://github.com/Helsinki-NLP/OPUS-CAT

How to use this model directly from the 🤗/`transformers` **library:**

```
from transformers import AutoTokenizer,
AutoModelWithLMHead

tokenizer = AutoTokenizer.from_pretrained("Helsinki-
NLP/opus-mt-ROMANCE-en")

model = AutoModelWithLMHead.from_pretrained("Helsinki-
NLP/opus-mt-ROMANCE-en")
```
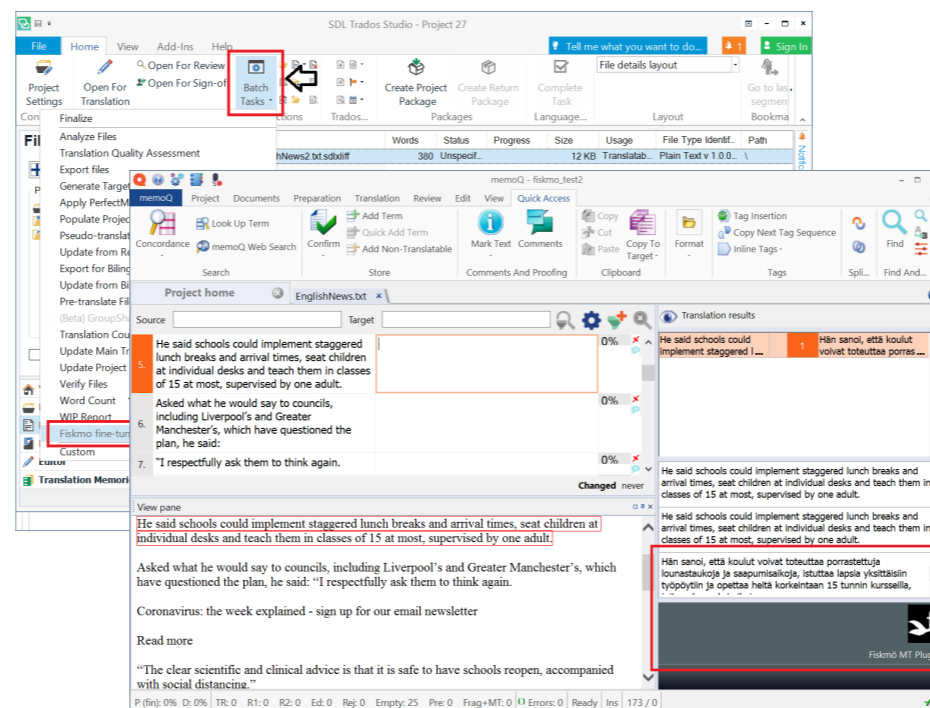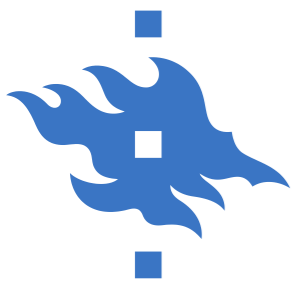
List all files in model · See raw config file

demo interface:
https://translate.ling.helsinki.fi

HELSINGIN YLIOPISTO
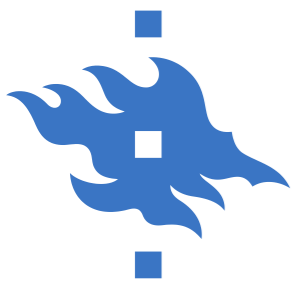HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Directions for future developments

FCAI mission statement

Directions for future developments

Understandability

- understand people

- make yourself understood

FCAI mission statement

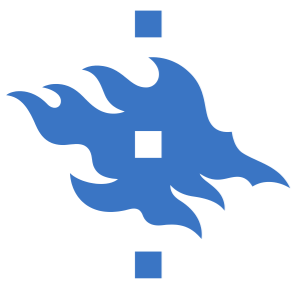# Directions for future developments

Understandability

- understand people

- make yourself understood

Trust

- support your claims and predictions

- able to argue and debate but also to adopt and to be convinced

FCAI mission statement

Directions for future developments

Understandability

- understand people
- make yourself understood

Trust

- support your claims and predictions
- able to argue and debate but also to adopt and to be convinced

Self-awareness

- recognise lack of knowledge
- curiosity-driven learning

FCAI mission statement

Directions for future developments

Understandability

difficult without language

- understand people

- make yourself understood

Trust

- support your claims and predictions

- able to argue and debate but also to adopt and to be convinced

Self-awareness

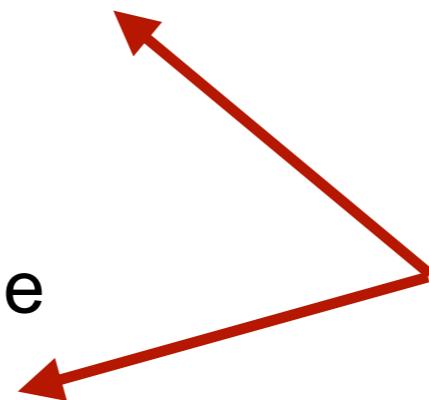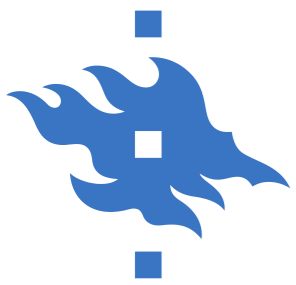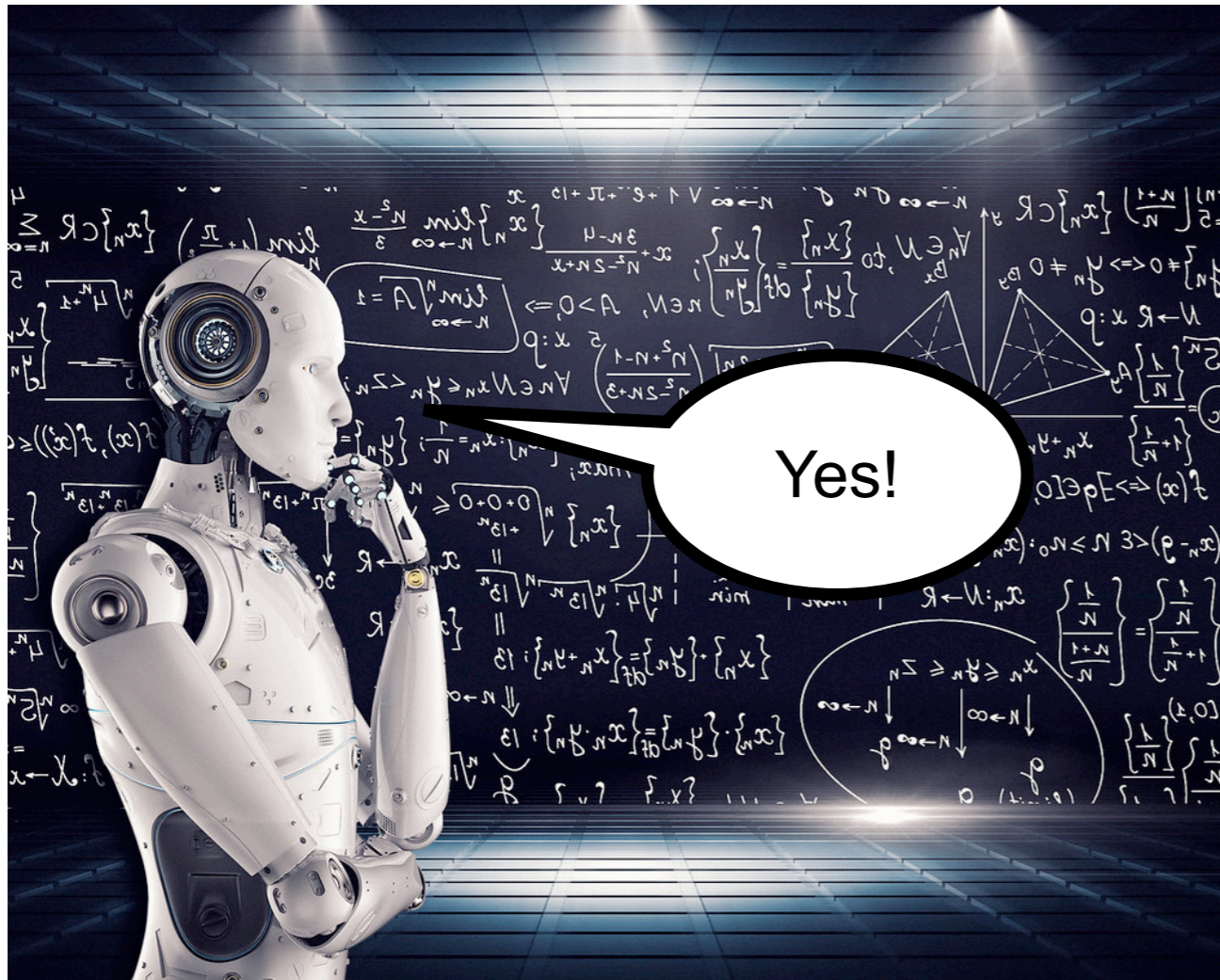- recognise lack of knowledge

- curiosity-driven learning

requires interactive communication

DATA EFFICIENCY
UNDERSTANDABILITY

REAL AI
REAL PEOPLE
REAL WORLD

TRUST & ETHICS

FCAI mission statement

Complex decision require explanations

Should we join NATO?

Yes!

Language, knowledge and problem solving
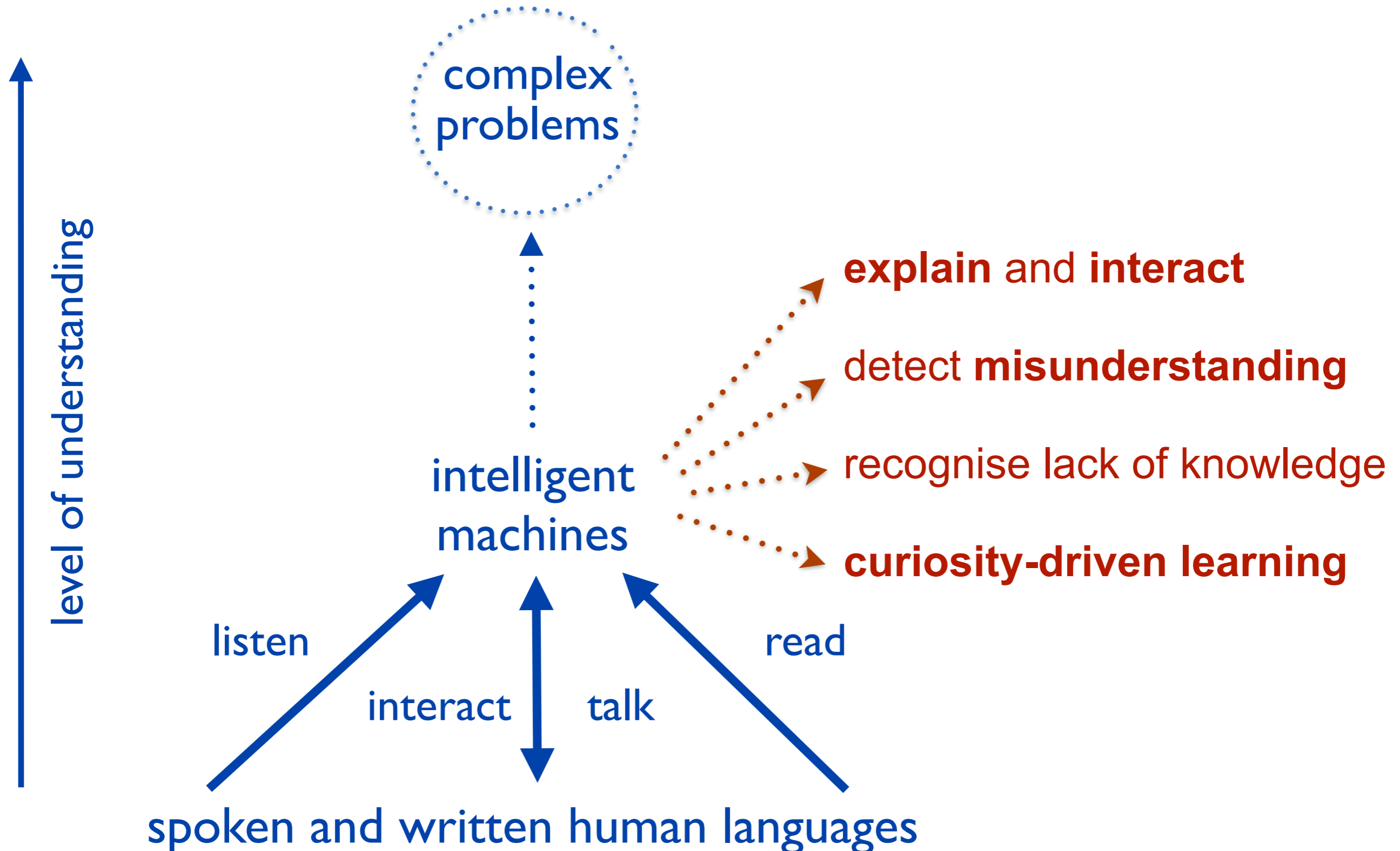
Thank you for your attention!

Website: www.lr-coordination.eu

Twitter: @LR_Coordination

Email: info@lr-coordination.eu

https://blogs.helsinki.fi/language-technology/