

Hybrid Language Models: From Super-Parrots to Human-Like Dialog Understanding



Wolfgang Wahlster



German Research Center for Artificial Intelligence

Berlin / Bremen / Darmstadt / Kaiserslautern / Lübeck / Oldenburg / Osnabrück / **Saarbrücken** / Trier

Email: wahlster@dfki.de

WWW: <http://www.dfki.de/~wahlster>

ICSI and DFKI Were both Established in 1988



Prof. Szyperski (Mannesmann-Kienzle, CEO) and Prof. Schwärtzel (Siemens, CTO) played a key role in founding ICSI and DFKI, Ron Kay was the acting director in 1987, turns 100 on Friday

The German Postdoc Program at ICSI was a Blueprint for other International Programs



Return Rate to Germany 80 %, 50 % are now Professors



ChatGPT compares ICSI and DFKI

Please compare DFKI in Germany and ICSI in Berkeley.



DFKI (Deutsches Forschungszentrum für Künstliche Intelligenz) in Germany and ICSI (International Computer Science Institute) in Berkeley are both research institutions that focus on artificial intelligence and computer science. However, there are some differences between them:

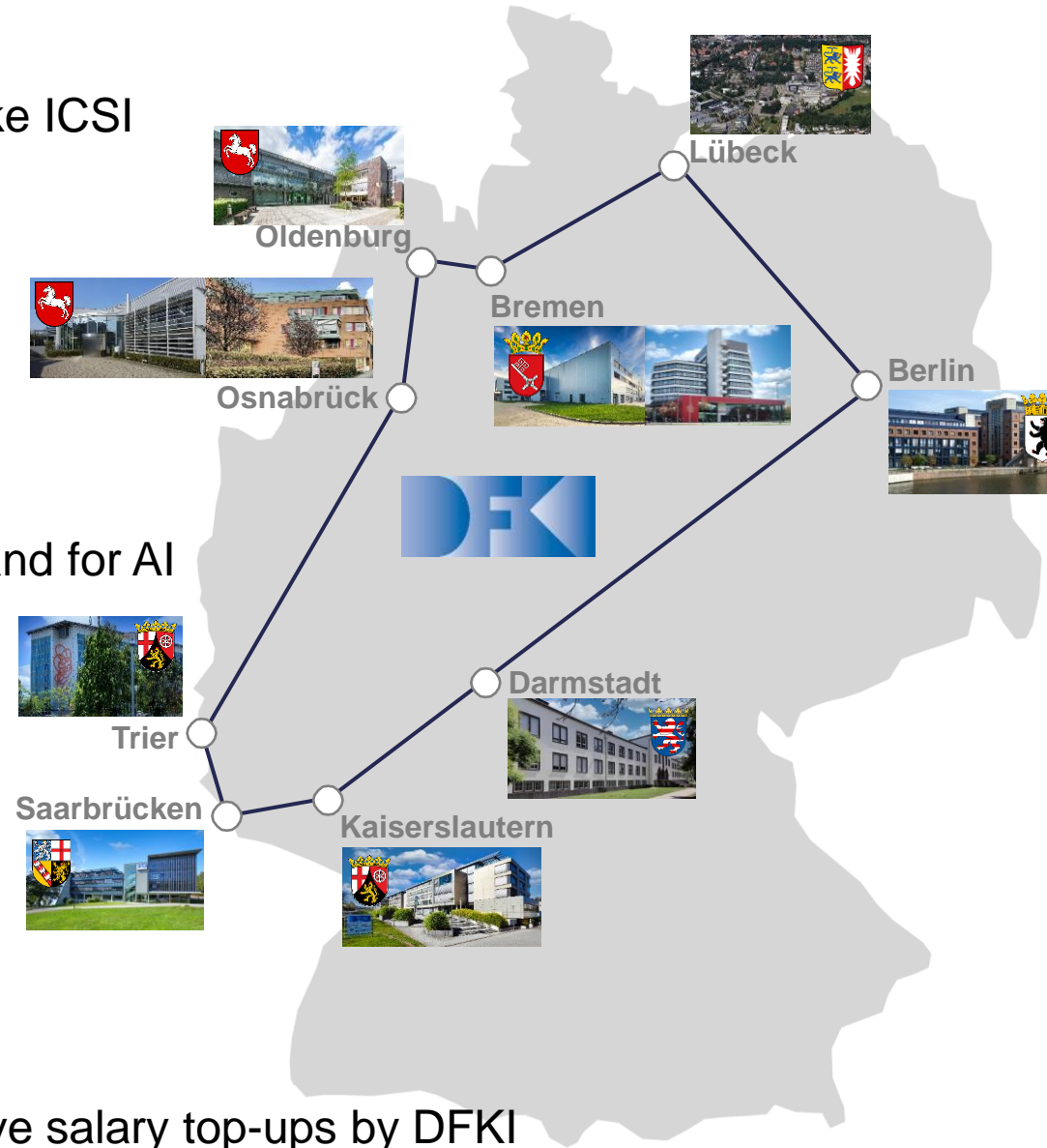
1. Location: DFKI is located in Saarbrücken, Germany, while ICSI is located in Berkeley, California, USA.
2. History: DFKI was founded in 1988 as a non-profit organization, while ICSI was founded in 1988 as a research center affiliated with the University of California, Berkeley.
3. Size: DFKI is a larger organization, with over 1,000 employees and several locations throughout Germany, while ICSI has a smaller team of around 50 researchers.
4. Research areas: Both institutions focus on artificial intelligence, but they have slightly different research areas. DFKI has a particular focus on robotics, language technology, and intelligent user interfaces, while ICSI focuses on natural language processing, computer vision, and networking.
5. Funding: DFKI is funded by both public and private sources, while ICSI is primarily funded by the National Science Foundation and private foundations.

Overall, both DFKI and ICSI are highly respected research institutions in the field of artificial intelligence and computer science, but they have slightly different research focuses and funding models.

DFKI is one of the World's Largest Public AI Research Centers

Established 1988,
35 anniversary this year like ICSI

Network of 9 DFKI-Labs
all over Germany with a
focus on the regional demand for AI



> 1530 employees
> 400 ongoing research projects
Average age of staff: 37 years

Revenues: 81 Million € (for 2022)
Total Contract Volume: 266 Million €

896 research staff
59 professors with attractive salary top-ups by DFKI

DFKI is a non-profit research institute with public and private shareholders



has international shareholders from a wide spectrum of industries:



Mercedes-Benz Group



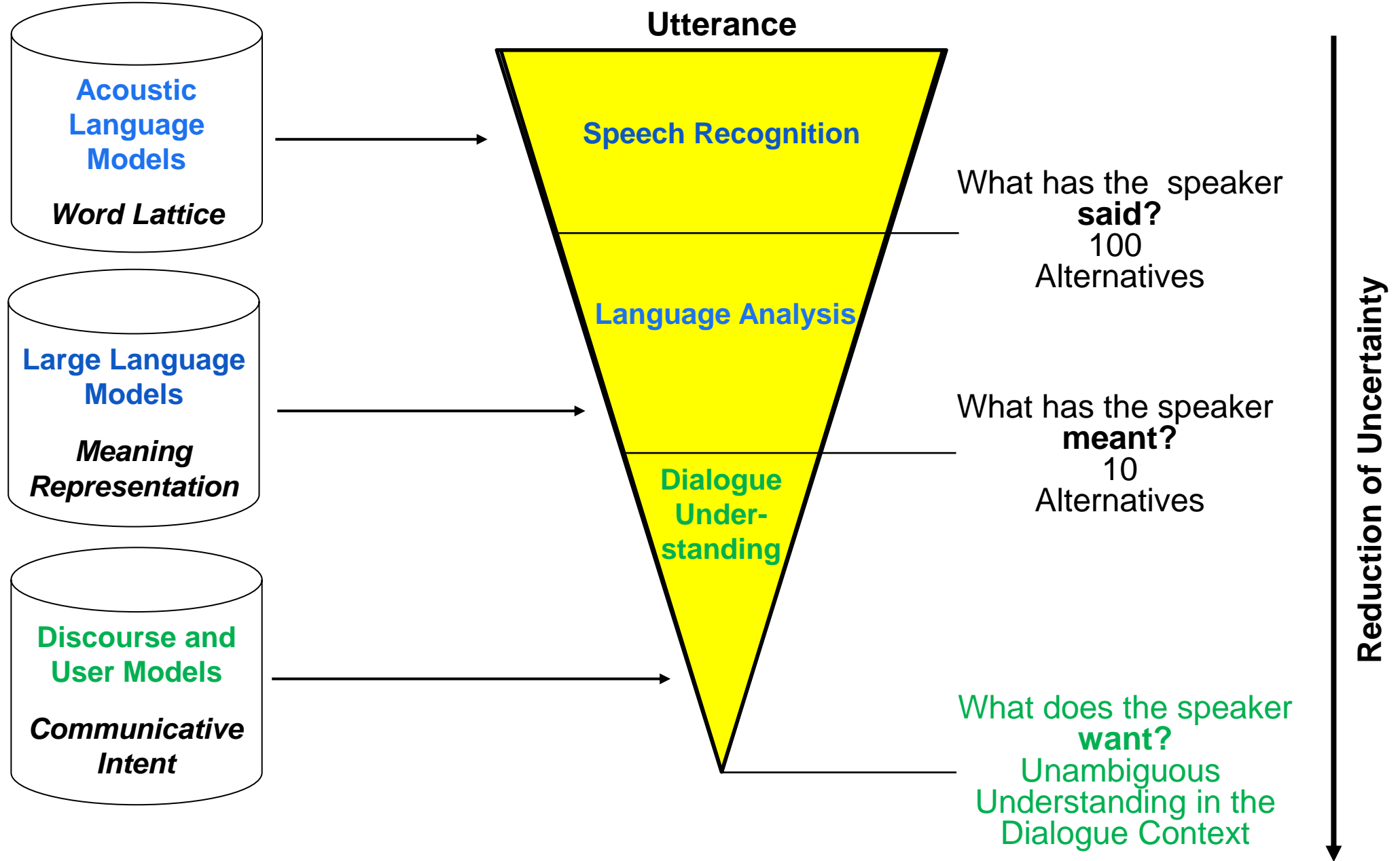
- Software Industry
- Chip Industry
- Manufacturing Industry
- Automotive Industry
- Aerospace Industry
- Retail Industry
- Finance Industry
- Food Industry
- Energy Industry
- Construction Industry



DFKI Has Created 2500 New R&D Jobs in IT Industry and > 100 Startup and Spin-Off Companies

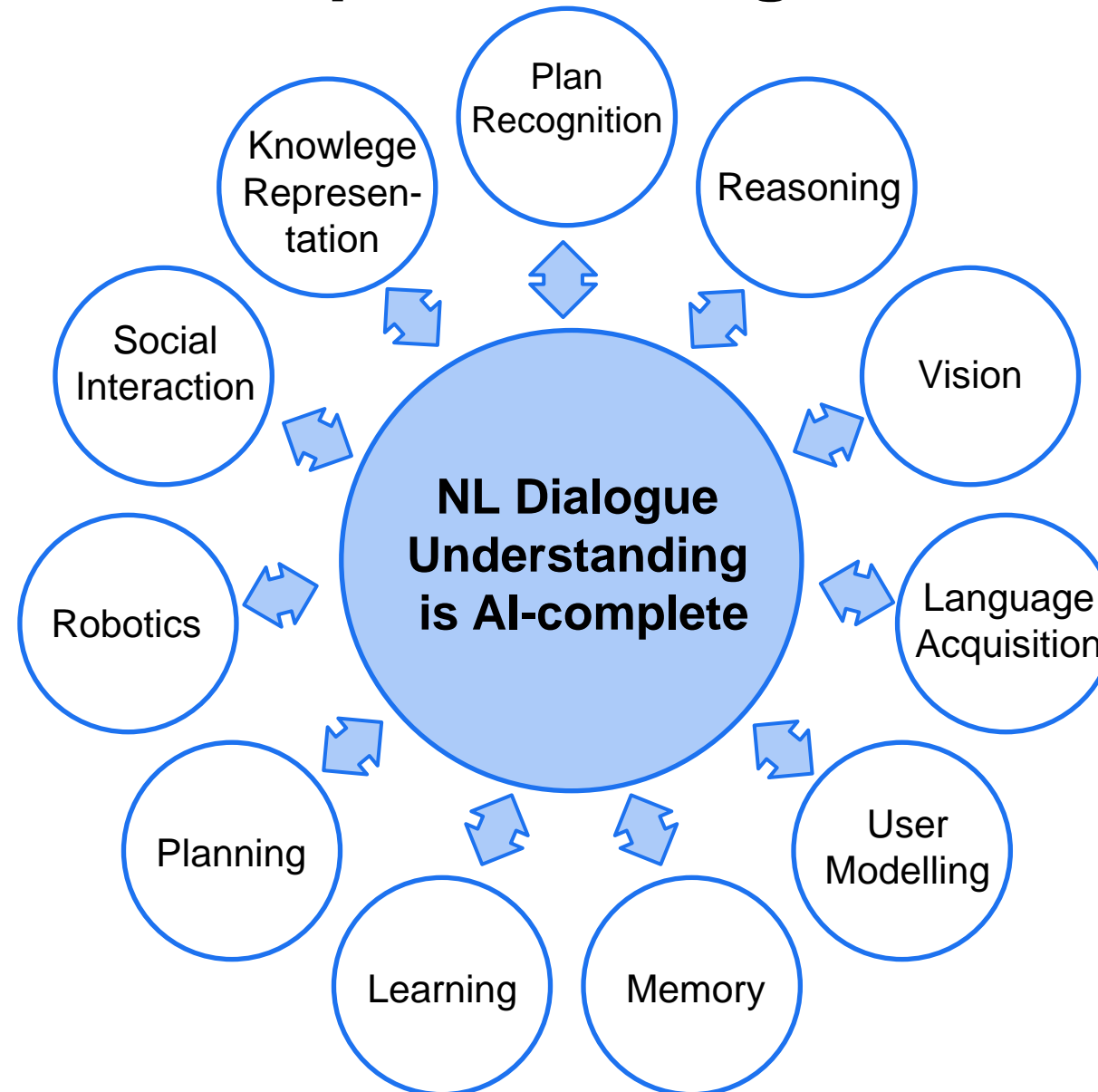
					DFKI maintains business relationships with many of these companies.			

Uncertainty Reduction in Dialogue Understanding Systems



Dialogue Understanding Is not an Isolated Capacity, but Is Embedded into other Aspects of Cognition.

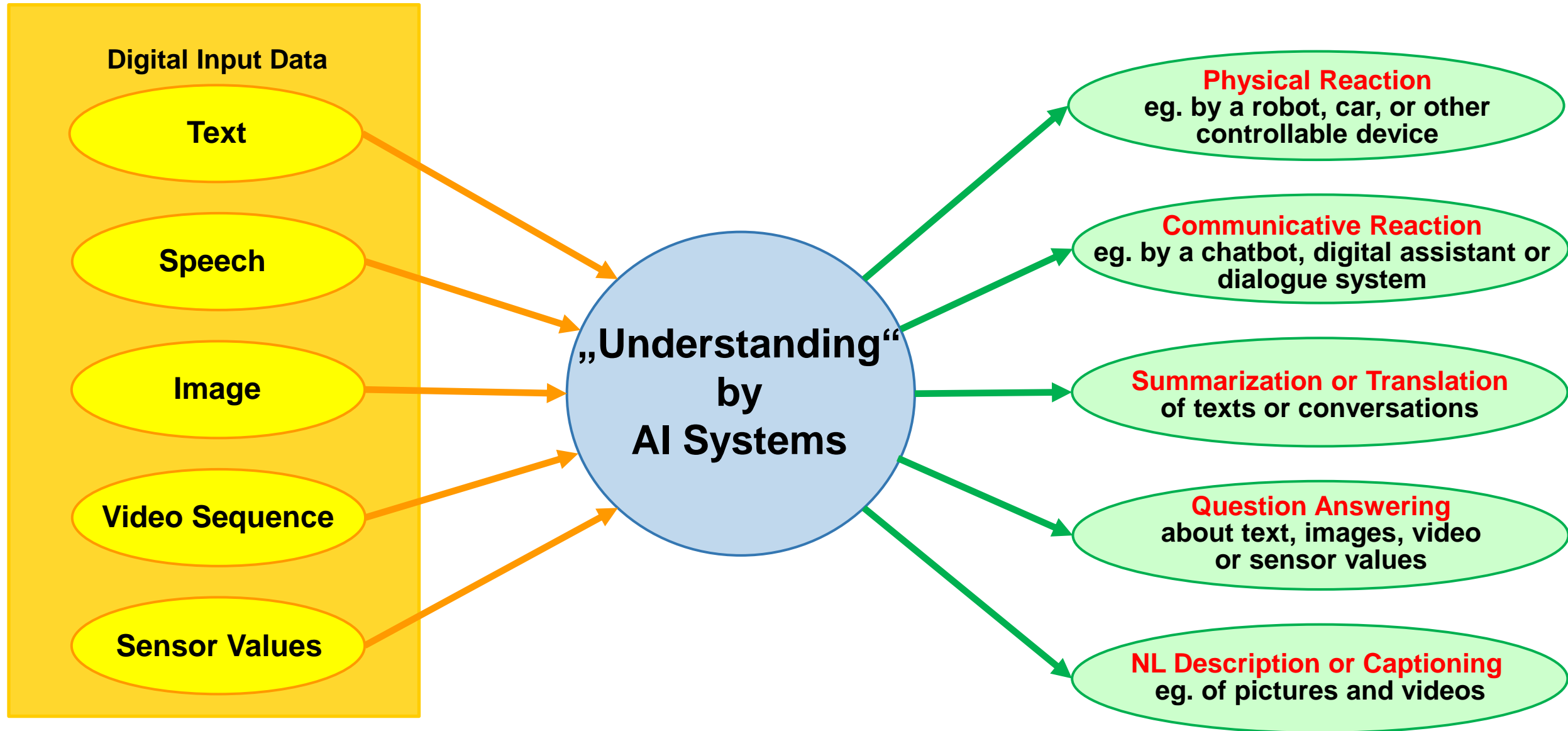
Social interaction based on natural language dialogue is **one of the defining characteristics of our species.**



AI-complete means that AI systems cannot have human-like cognitive competences in a particular subfield of AI without **having all other cognitive competences that are the basis of human intelligence.**

The computer simulation of human dialogue is one of the **most ambitious scientific goals of this millennium.**

How can we test an AI system's „understanding“ of input?



Understanding Test: Adequate Response

Five Decades of Natural Language Dialogue Systems

1976	1986	1996	2006	2016	2026
Closed-Domain Dialogue Systems	Perceptually Grounded Dialogue Systems	Speech-to-Speech Dialogue Translation	Open-Domain Dialogue Systems	Massively Multimodal Dialogue	
	Multimodal Dialogue Systems	Conversational Characters	Empathic Virtual Agents	Hybrid Team Interaction	
	Task-oriented Dialogue Systems	Embodied Dialogue Systems	Multiparty Dialogue Systems	Chatbots based on Large Language Models	
HAM-RPM, HAM-ANS	VITRA, XTRA	VERBMOBIL, SMARTKOM, MSA	SMARTWEB, THESEUS, AMI VIRTUAL HUMAN	MADMACS, HYSOCIATEA, OPENGPT-X	
Selected Natural Language Dialogue Systems for German					

Winograd's SHRDLU: Dialogue Understanding Grounded in a Microworld (50 years ago)

SHRDLU answers questions, executes commands, and accepts information in an interactive English dialog.

It deals in an integrated way with all of the aspects of **language-syntax, semantics, and inference**.

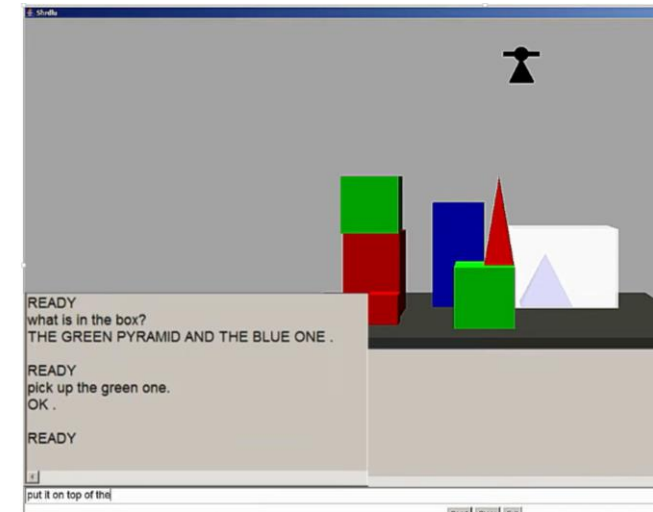
The system contains a parser, a recognition grammar of English, programs for semantic analysis, and a general problem solving, reasoning and planning system.

SHRDLU had a detailed model of a particular domain: **the BLOCKS world**.

It **can remember and discuss its plans and actions as well as carrying them out**.

It enters into a dialog with a person, responding to English sentences with actions and English replies, asking **for clarification** when it cannot understand a sentence through the use of syntactic, semantic, contextual, and physical knowledge.

Terry Winograd (1972): Understanding Natural Language Academic Press, New York and Edinburgh University Press (1972). ISBN 10: 0852242271





(2) U: "What is your price?"

(1) O: "Hi, there. I'm the new camera from Canon with 7 mega pixels."

(3) O: "My price is € 599."

Large Collaboration Research Projects ICSI-DFKI on Multimodal Dialog Systems 1999-2009



1999-2003 Multimodal Dialog in the Car, BMBF, 25 M €



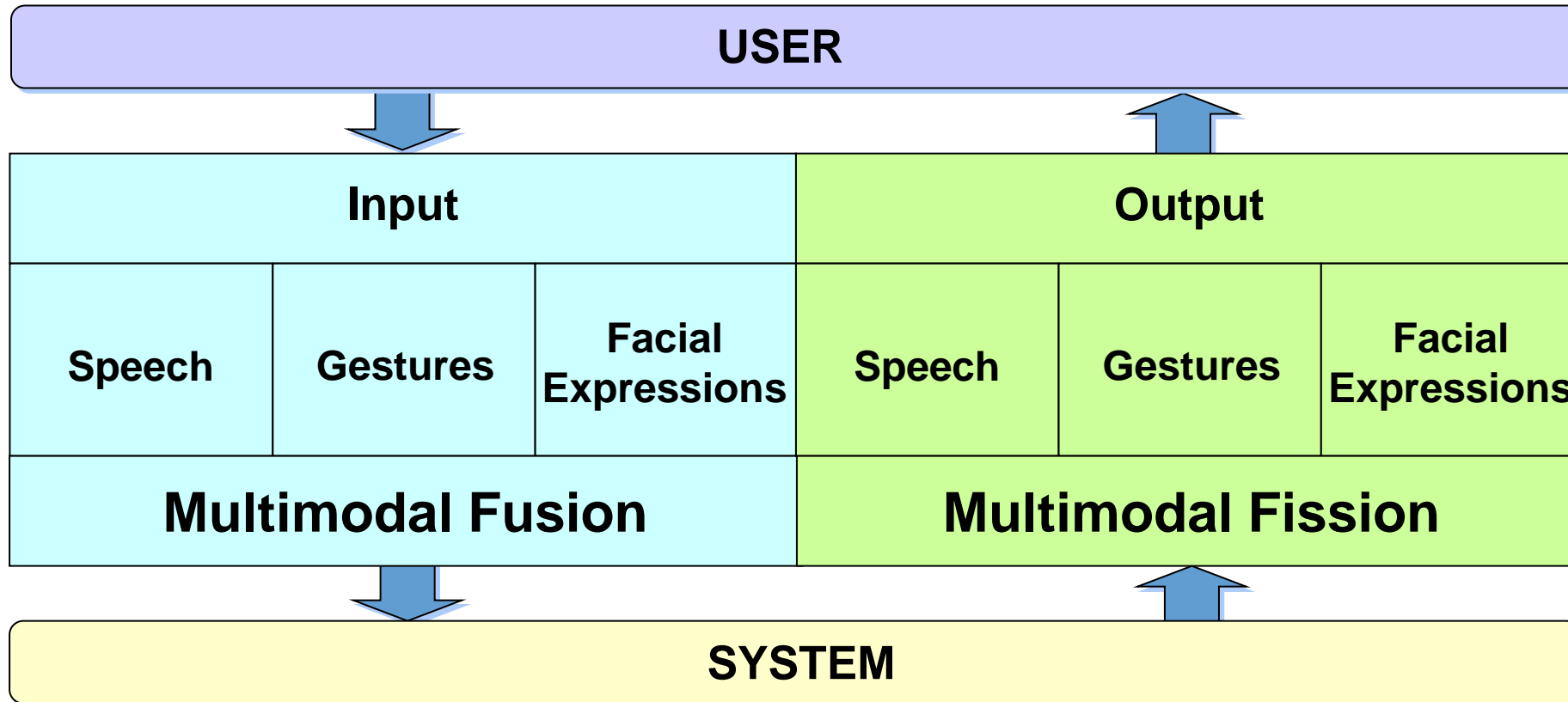
2004-2008 Mobile Speech-based Access to the Semantic Web, funded by BMBF, 20 M €



2004-2009 Augmented Multi/Party Interaction, funded by EU, jointly with Hervé Bourlard, IDIAP, TNO, 10 M € EC funding

Subcontracts for ICSI, paid by German Government or EU

No Presentation without Representation !

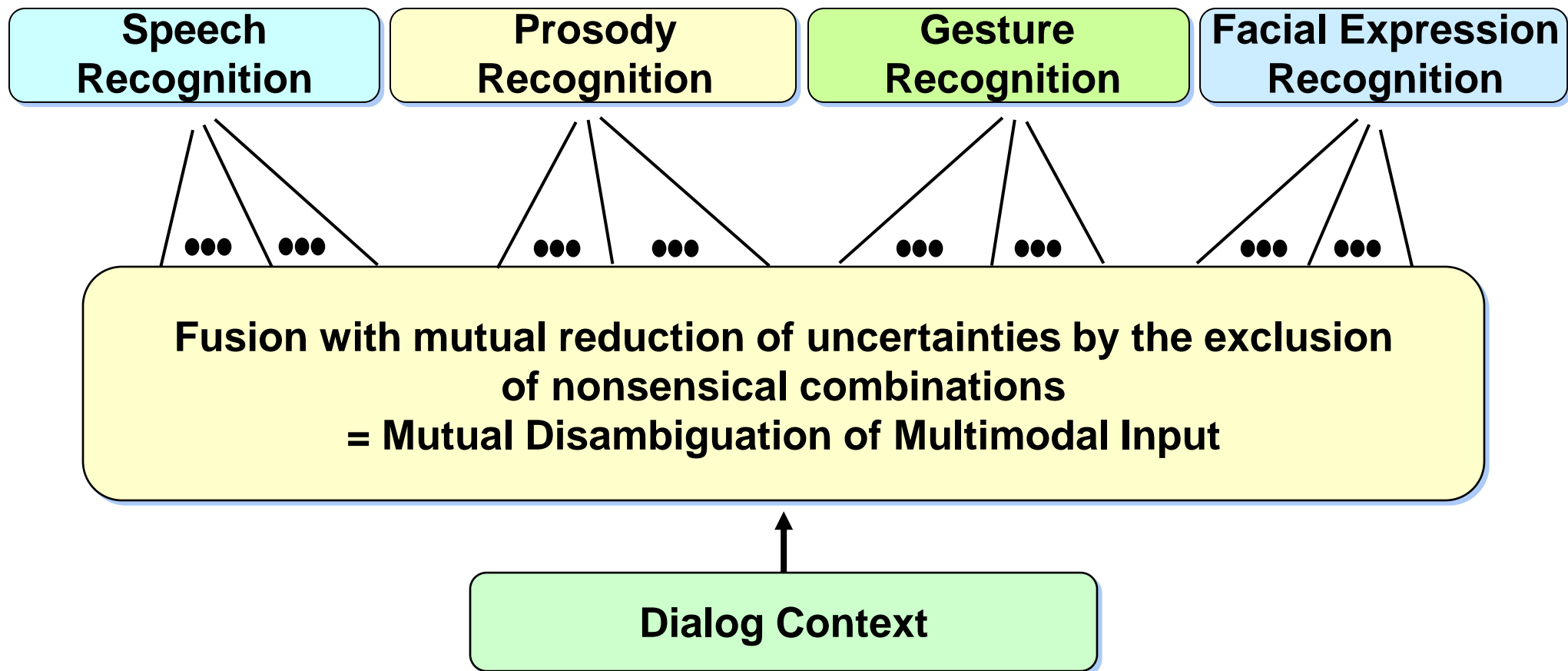


The modality fission component provides the inverse functionality of the modality fusion component.

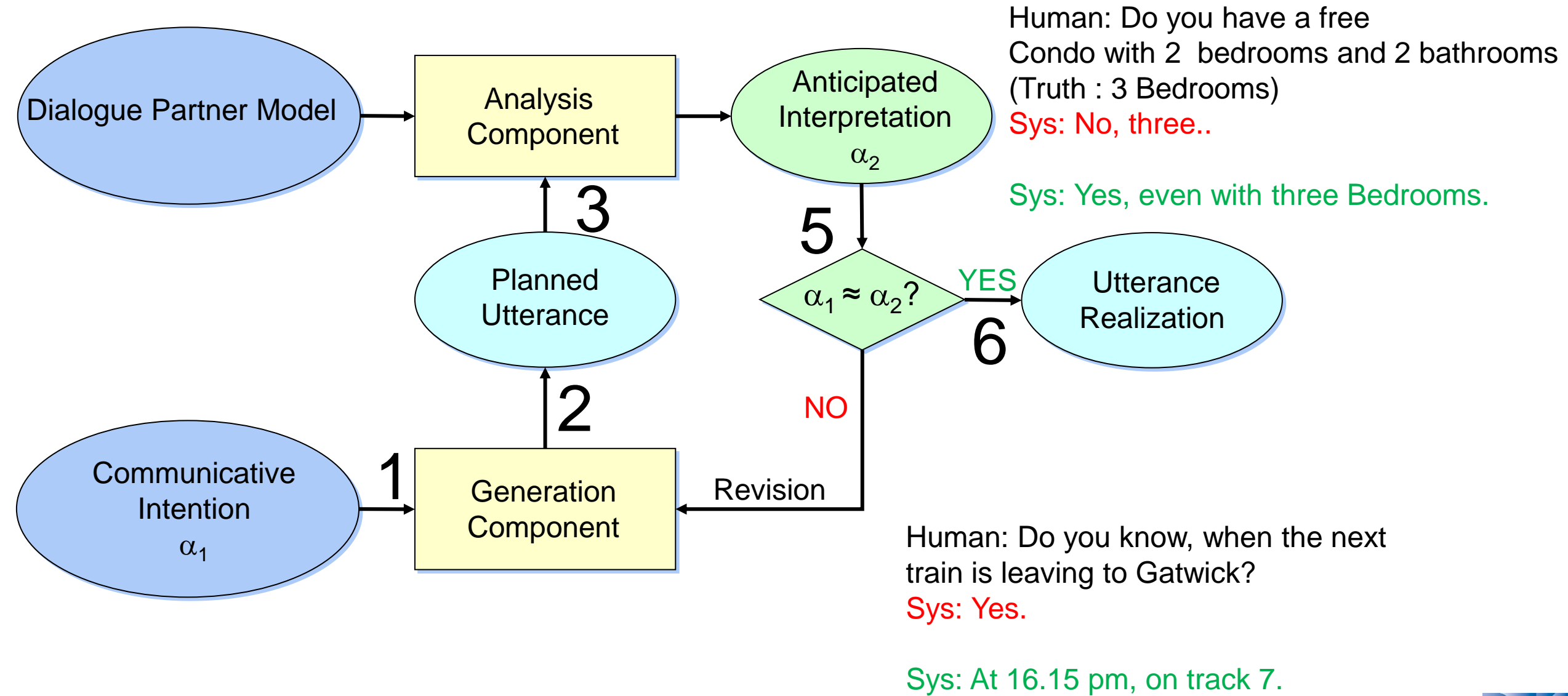
A dialogue system with symmetric multimodality must not only understand and represent the user's multimodal input, but also its own multimodal output.

The Fusion of Multimodal Input

Multiple modalities increase the uncertainty of interpretation but: the semantic fusion of multiple modalities in the dialog context enables an unambiguous interpretation



Anticipation Feedback Loops for Deep Dialogue



Grounding Referential Meaning

by the Driver's Eyepointing to Landmarks in the Visual Context Combined with Speech Input in Multimodal Dialogues

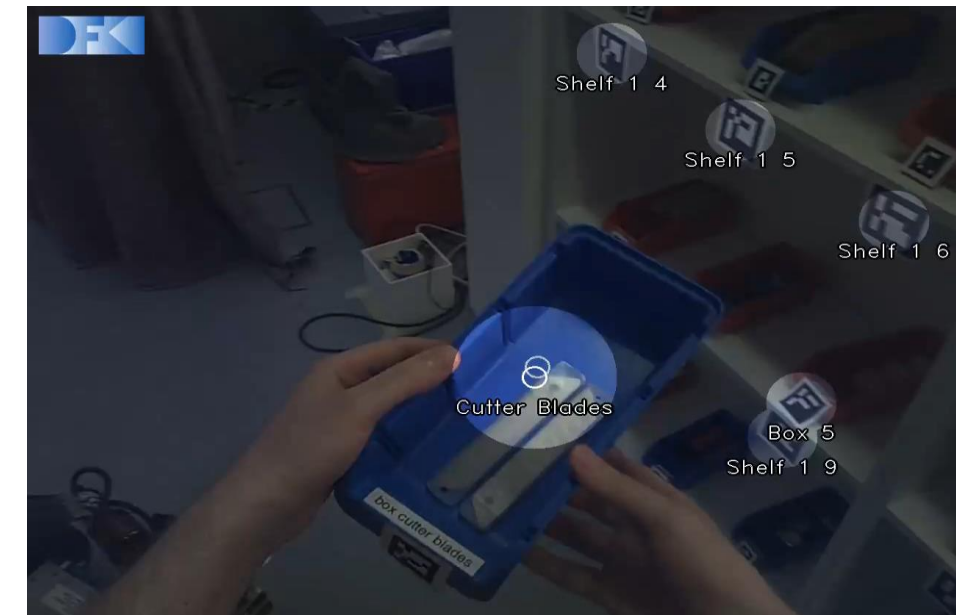


2012: Early Prototype by DFKI in the Carmina/Siam project (PI W. Wahlster)



2021: Commercial Product for Mercedes Cars World Novelty first shown at CES 2021

Multimodal Discourse with 2 Humans, 3 Robots and 2 Virtual Assistant Softbots with Grounding by Computer Vision



Deep Learning-based Object Recognition on the Job Floor

Virtual Humans: Multimodal Multiparty Interaction with Two Humans and Three Virtual Characters, Wahlster et al.



World's first computational model of complex dialogue behavior and multimodal turn-taking in multiparty dialogs

Reithinger et al.: DOI:10.1145/1180995.1181007



Stochastic Parrots: The Power and the Limitations of NLP Systems based on Gigantic Neural Language Models



Talking birds
(mostly parrots)
are birds that
can mimic the
speech of
humans (up to
2000 words)

The past 4 years of research in natural language processing have produced ever larger language models (from BERT to LamDA open-sourced by hyper-scalers such as Microsoft, Google, and Baidu).

Brute-force approach:

- Dramatically increased size of training data sets (LamDA 1.56 trillion words)
- Extremely large parameter spaces
- Simple variations of transformer-based deep learning architectures
- Very successful in language generation benchmarks

But no natural language understanding in the semantic & pragmatic sense and often no grounding in the real world to enable human-like dialogue understanding

No model of language acquisition: human language learning is not only grounded in the physical world, but also in interaction with other people in that world. Kids won't pick up a language from passive exposure such as TV or radio.

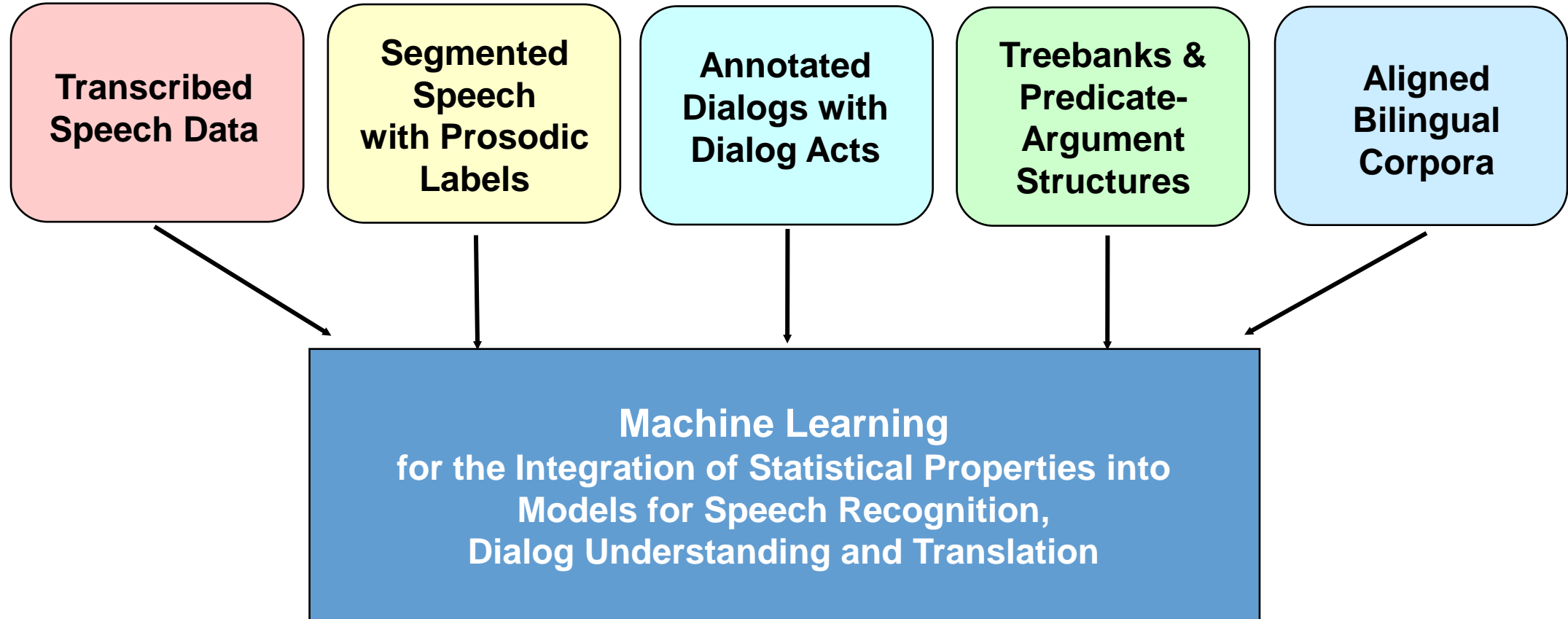
Emily M. Bender et al (2021).: On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?

FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency March 2021 Pages 610–623 <https://doi.org/10.1145/3442188.3445922>

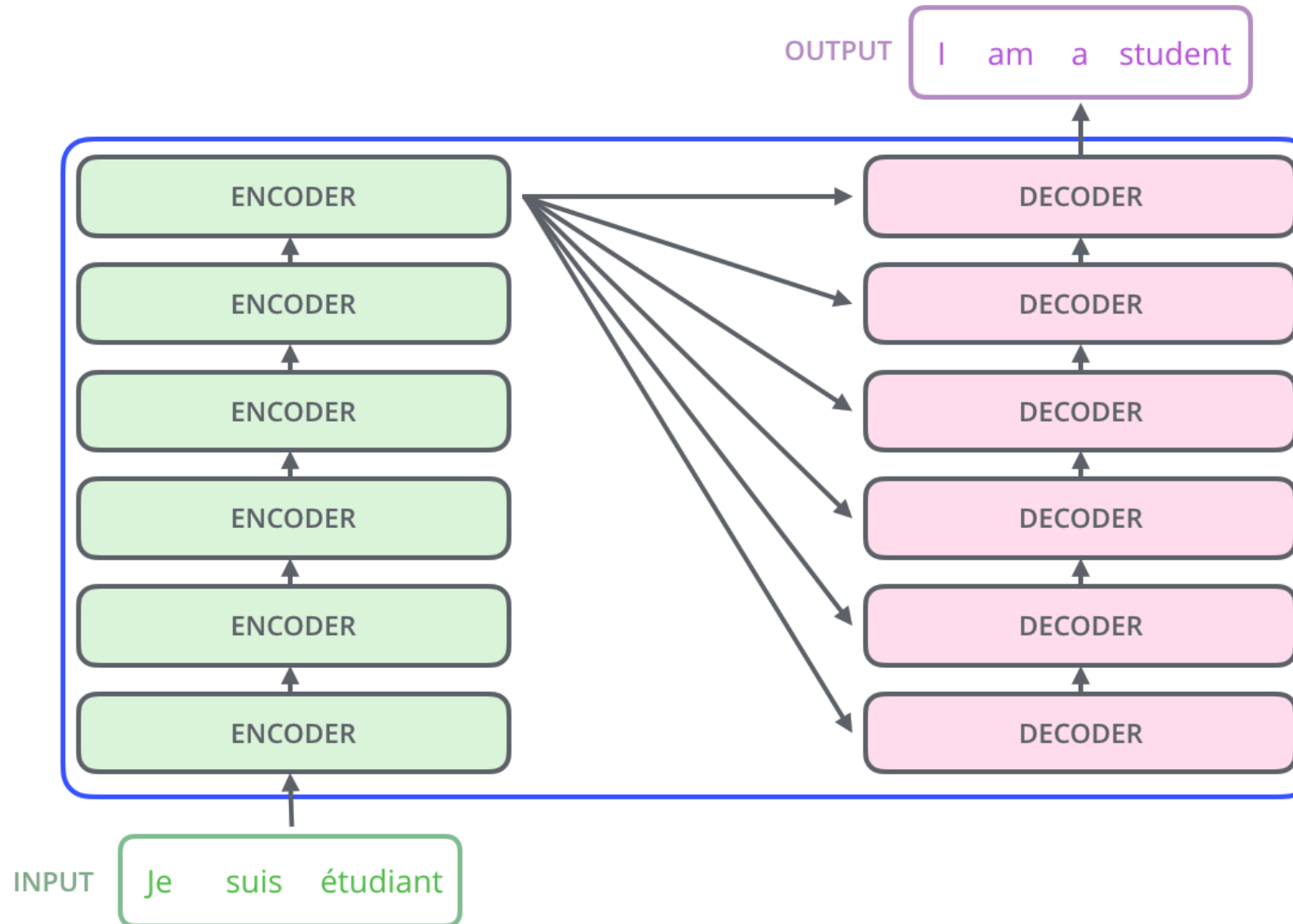
Emily M. Bender, Alexander Koller (2020) Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics

DOI: : 10.18653/v1/2020.acl-main.463

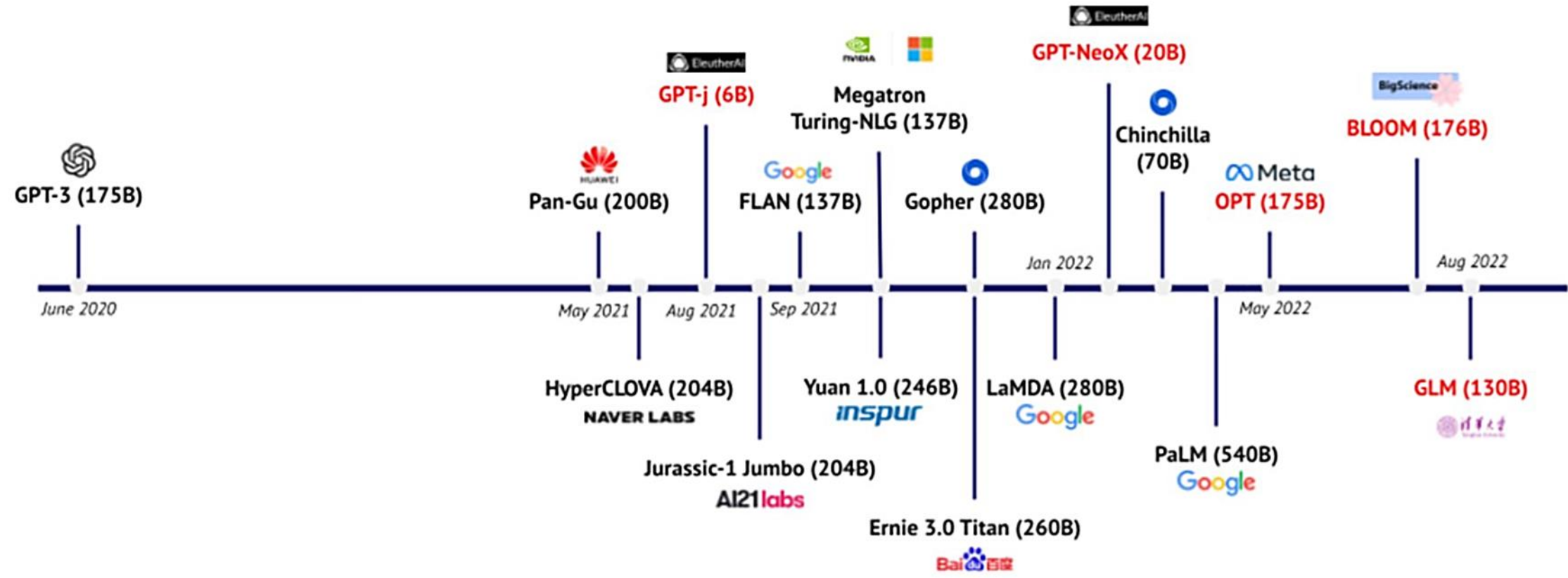
Extracting Statistical Properties from Large Corpora



Basic Transformer Architecture for Deep End-to-End Learning of Large Neural Language Models



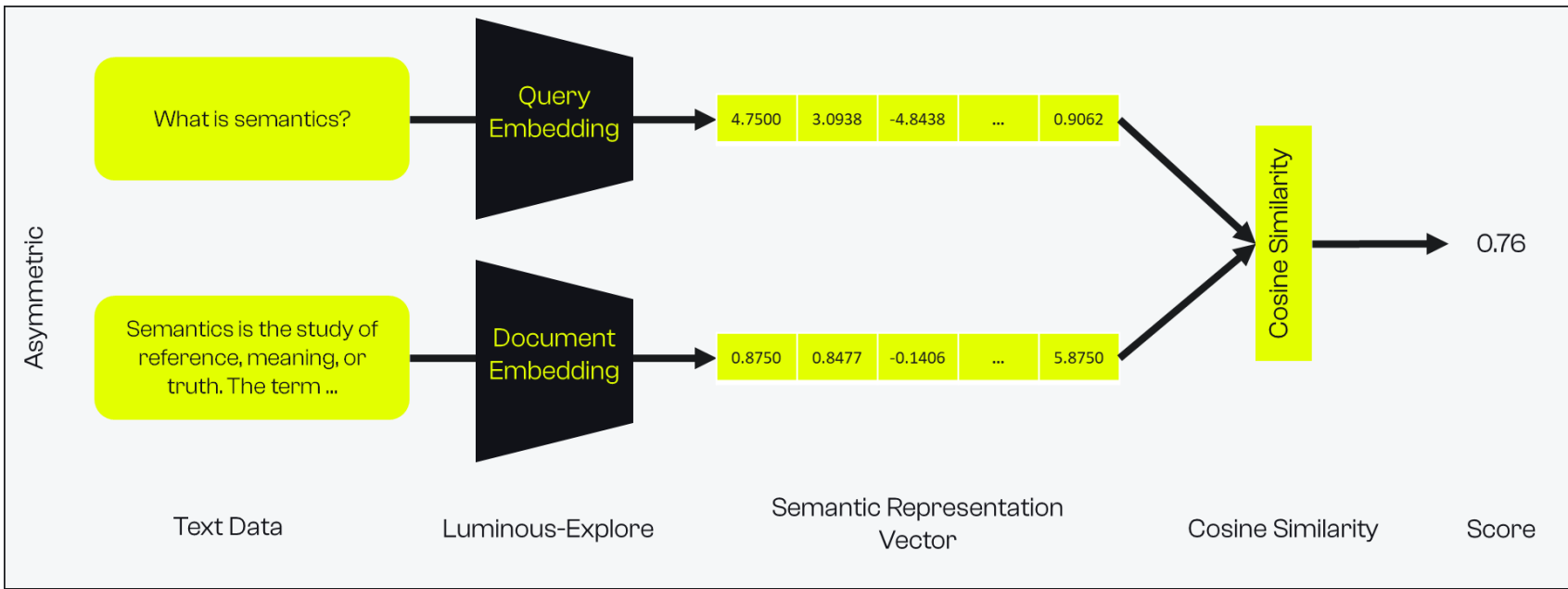
Large Language Models with Billions of Parameters (red = open source)



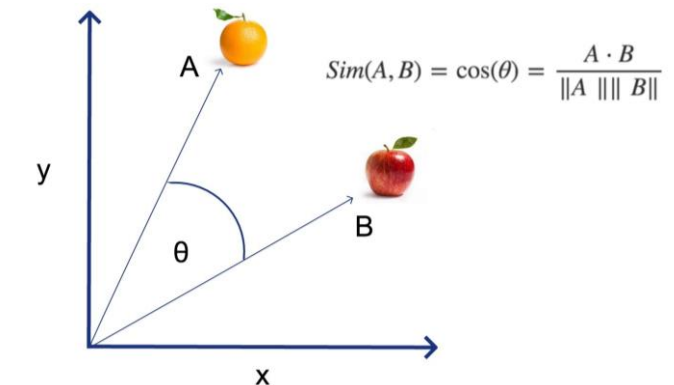
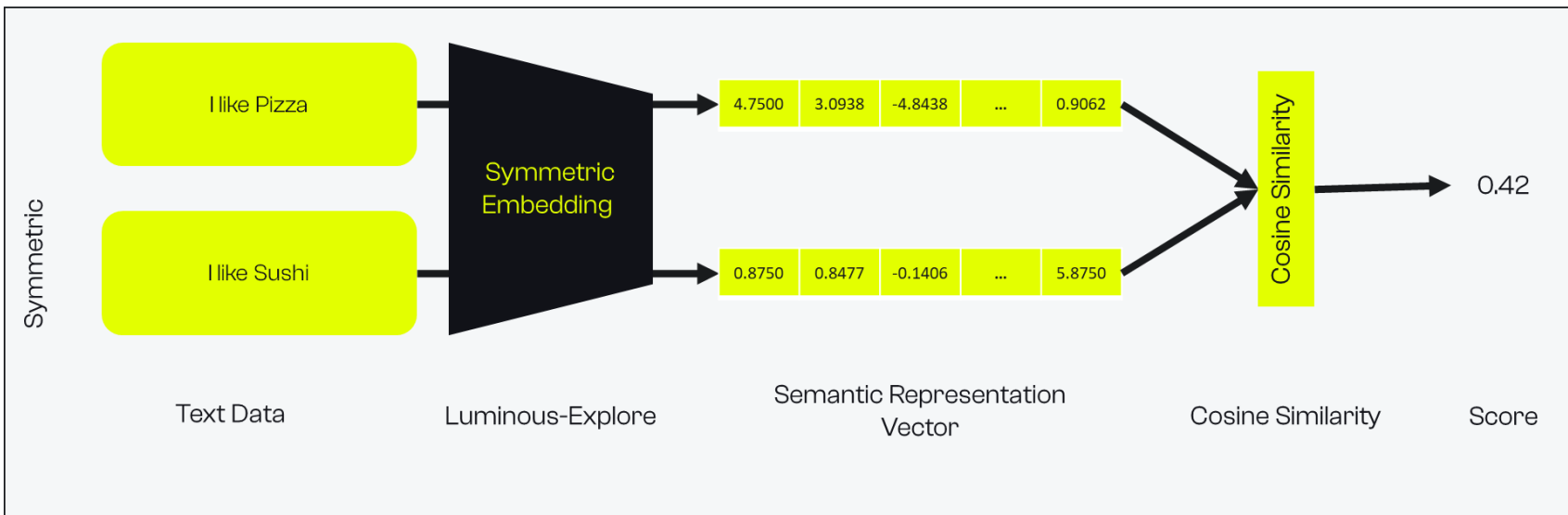
ChatGPT (MS, OpenAI), Ernie Bot (Baidu), YouChat (You.com), Bard (Google), Luminous (Aleph Alpha), LLaMA (Facebook)

Source: State of AI Report 2022 (Benaic & Hogarth, 2022)

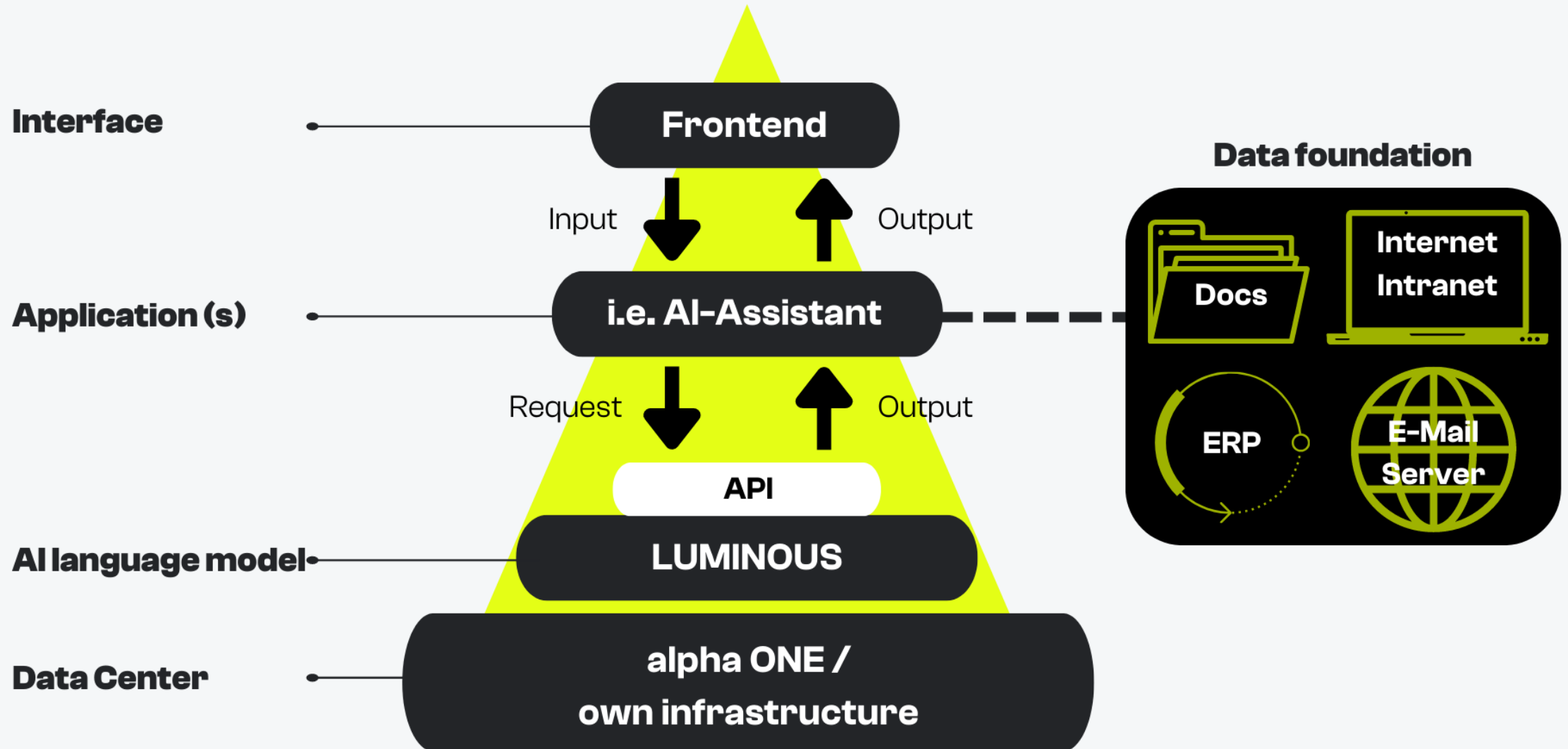
The German LLM-System LUMINOUS by Aleph Alpha includes symmetric and asymmetric Embeddings.



Semantic search: Given a set of sentences, we can search using a 'query' sentence and identify the most similar records. This enables search to be performed on concepts (rather than specific words).



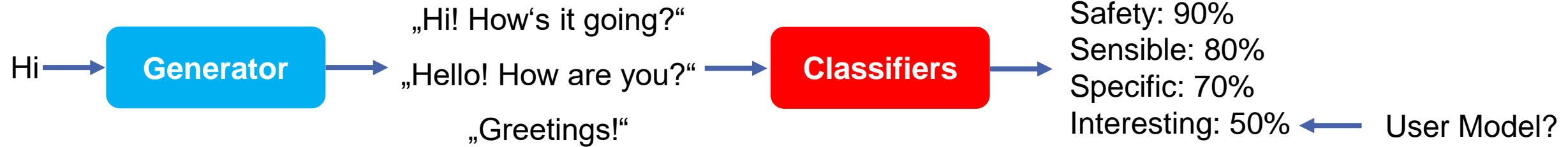
Secure and Sovereign ChatBots Made in Germany based on Luminous LLM



Response Generation Algorithm in LaMDA: Human-Like?

The LaMDA generator is trained to predict the next token on a dialog dataset

The LaMDA classifiers are trained to predict (SSI) ratings for the response in context using annotated data



During a dialog, the LaMDA generator first generates several candidate responses given the current multi-turn dialog context.

1. LaMDA classifiers predict the SSI and Safety scores for every response candidate.
2. Candidate responses with low Safety scores are first filtered out.
3. Remaining candidates are re-ranked by their SSI scores
4. The top result is selected as the response.



Heng-Tze Cheng, Romal Thoppilan; LaMDA: Towards Safe, Grounded, and High-Quality Dialog Models for Everything, 2022, Google Research, Brain Team

Solving Math Word Problems with the PaLM Language Model of Google Brain Research

prompt

Q: Tom's ship can travel at 10 miles per hour. He is sailing from 1 to 4 PM. He then travels back at a rate of 6 mph. How long does it take him to get back?
A:

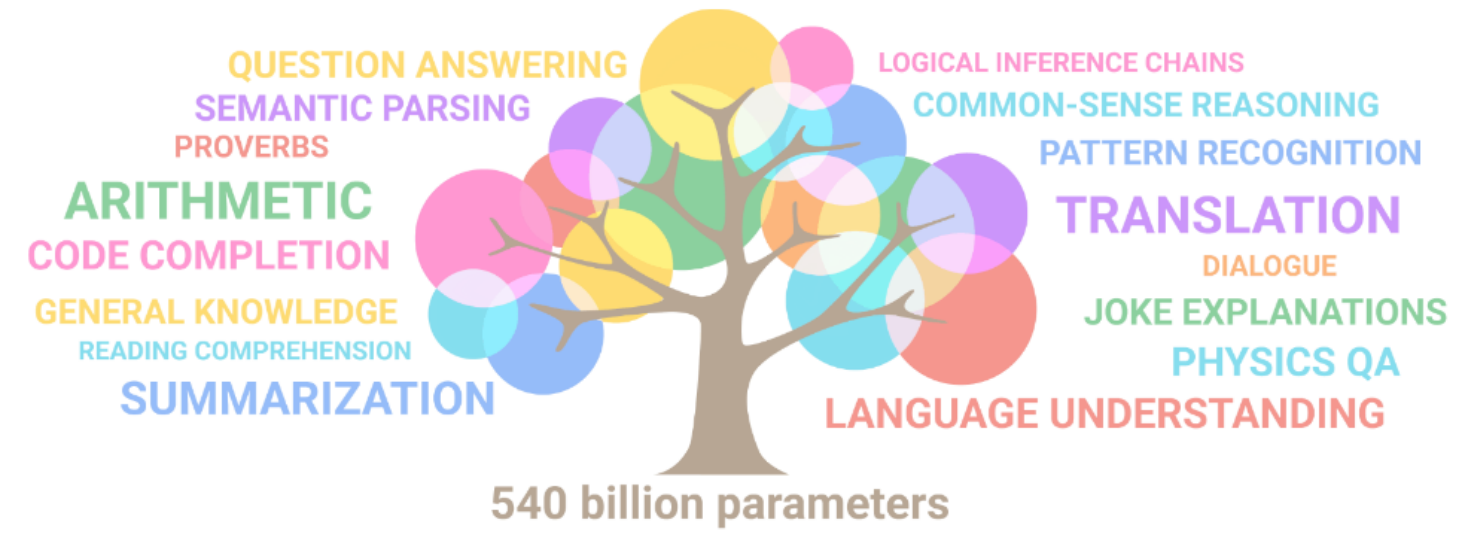
62b model

He travels from 1 to 4 PM. That means he travels for 3 hours. He travels back at 6 mph. That means he travels for $3 * 6 = 18$ miles. The answer is 18. ✗

540b model

He travels at 10 mph from 1 to 4 PM. This means he travels 3 hours. 3 hours at 10 mph means he travels $3 * 10 = 30$ miles. He then travels back at 6 mph. This means he travels 6 miles per hour. He has to travel 30 miles, so it takes him $30 / 6 = 5$ hours. The answer is 5. ✓

The 540 Billion parameter model leads to a reduction of errors compared to the 62 Billion parameter model.



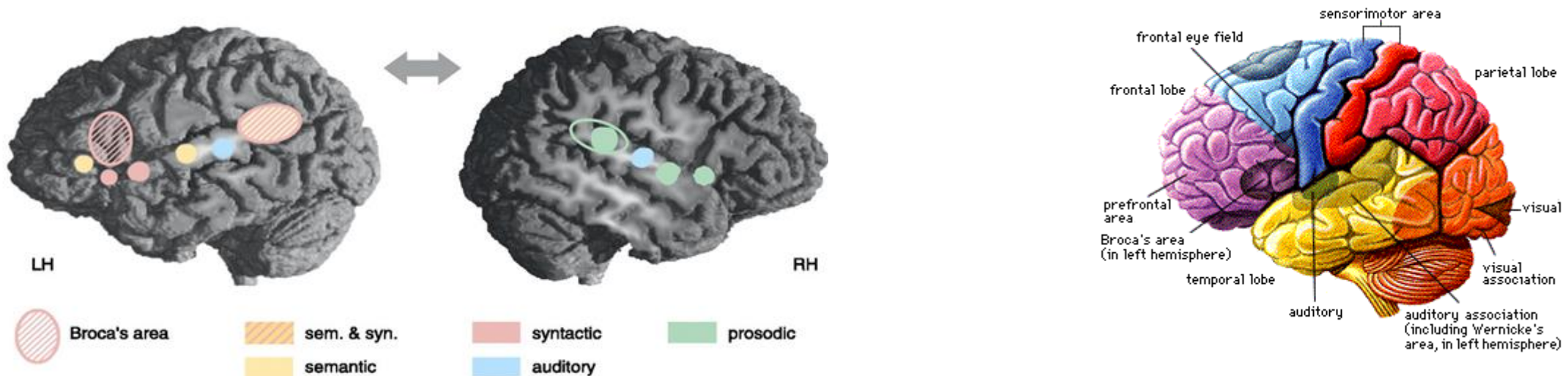
Sharan Narang and Aakanksha Chowdhery (2022) Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for Breakthrough Performance
<https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html>

Language Models with more Parameters than the Human Brain?

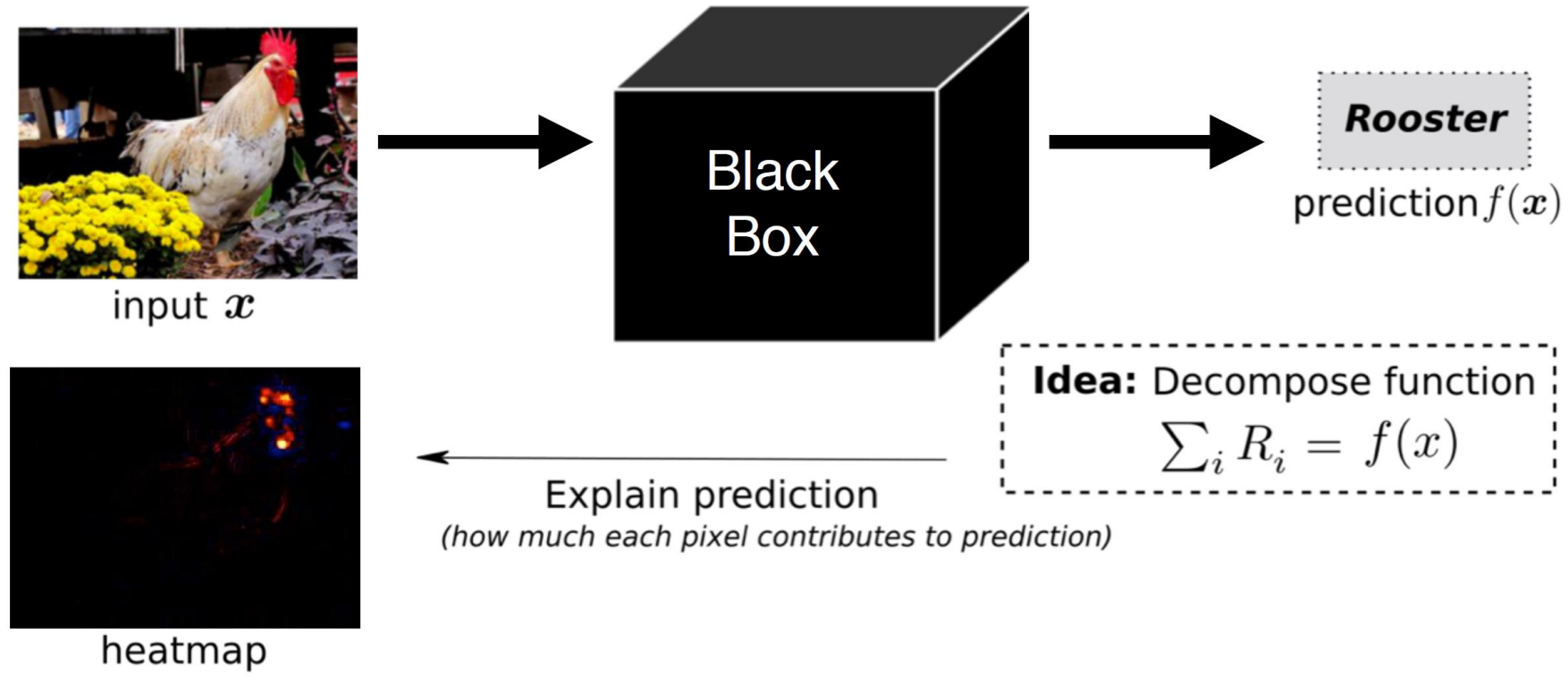
- **LaMDA** is based on up to **137 Billion model parameters**.
- **The adult male human brain** contains on average **86.1 +/- 8.1 Billion neurons**.
- The number of neurons is generally assumed to be a determinant of computational power
- We often forget that the brain is more than neurons (e.g. hormones).

Azevedo FA, Carvalho LR, Grinberg LT, Farfel JM, Ferretti RE, Leite RE, Jacob Filho W, Lent R, Herculano-Houzel S. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *J Comp Neurol*. 2009 Apr 10;513(5):532-41. doi: 10.1002/cne.21974. PMID: 19226510.

The brain basis for language understanding

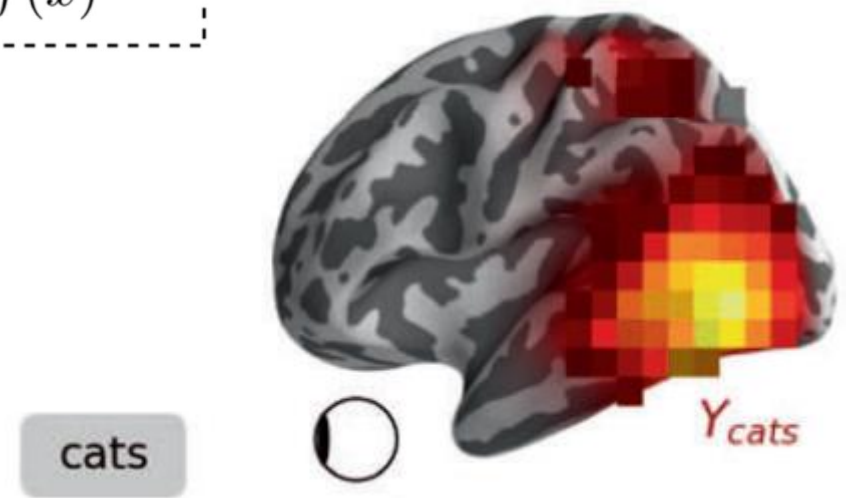


Activity Heatmaps don't fully explain the interpretation process on a transparent semantic level for ANNs and fMRI for the human brain.



Layer-wise Relevance Propagation (LRP)
(Bach et al., PLOS ONE, 2015)

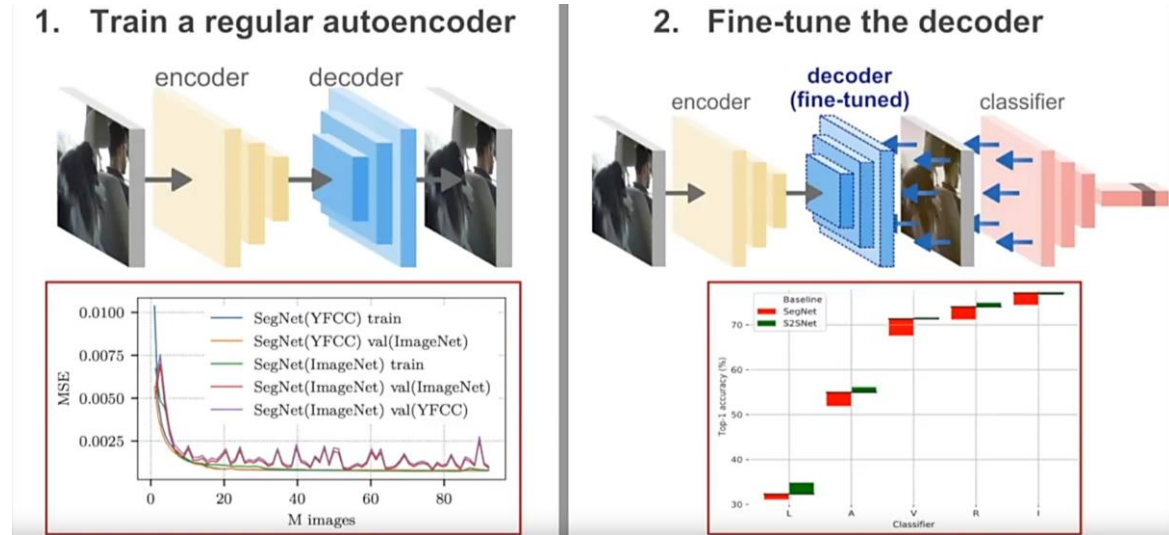
Charlotte Caucheteux, Jean-Rémi King
Brains and algorithms partially converge in natural language processing



Pioneer Award for DFKI at CVPR for Breakthrough Result on Explainable Deep Networks



Sebastian Palacio from DFKI's Deep Learning Center with CEO of NVIDIA



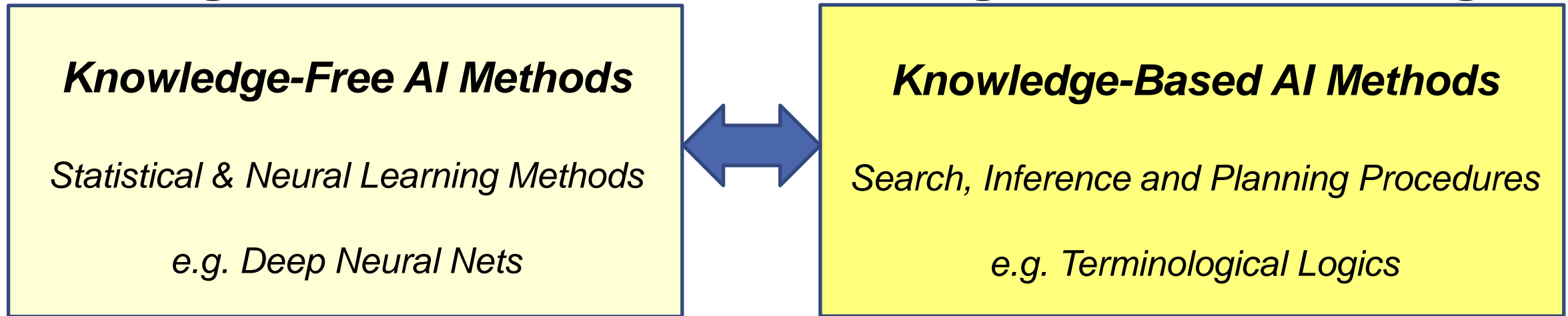
What do Deep Networks Like to See? In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City,

We invented a novel way to understand convolutional neural networks **by quantifying the amount of input signal they let in**. An autoencoder (AE) was fine-tuned on gradients from a pre-trained classifier with fixed parameters.

The AE learns **which aspects of the input space to preserve and which ones to ignore**, based on the information encoded in the backpropagated gradients.

Measuring the changes in accuracy when the signal of one classifier is used by a second one, a relation of total order emerges.

The Combination of Machine Learning and Knowledge-Based Inference for Dialogue Understanding



Mutual Support

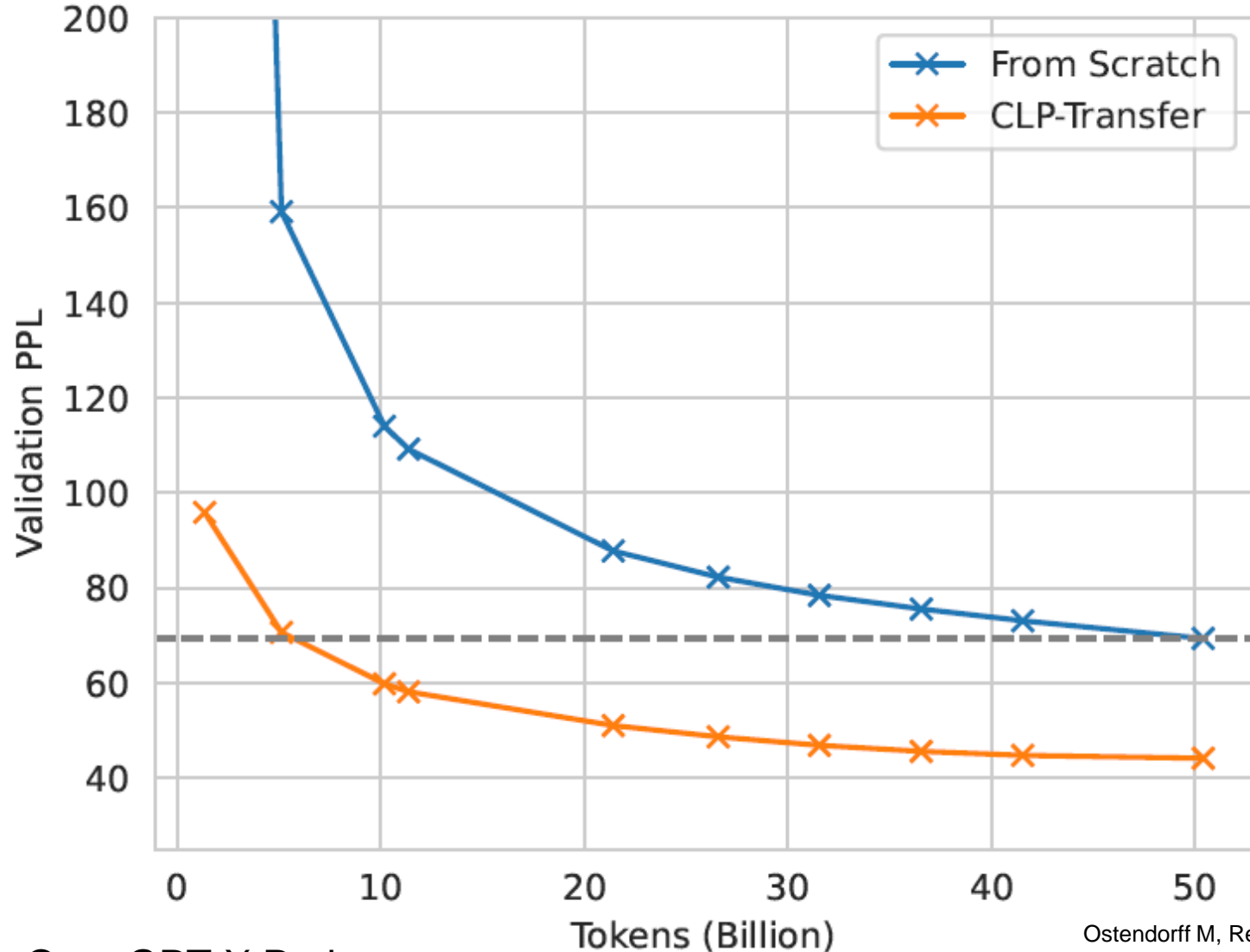
Deep learning methods can be used,

- to control search processes in knowledge-based systems and thus make them more efficient.
- to learn basic operators for knowledge-based inference and planning procedures.

Knowledge-based AI methods can be used,

- to filter, combine, complete or correct the results of machine learning.
- to make the results of machine learning processes plausible or explain them to end users.

DFKI's CLP-Transfer approach achieves the same Perplexity as from-scratch training but after 20% of tokens (dashed line).



When a token is not part of the overlapping vocabulary $v \in V_s \cap V_t$, we initialize its embedding v_t as the weighted average over the embeddings \hat{v} of the overlapping token:

$$v_t^{(large)} = \sum_{\hat{v} \in V_s \cap V_t} \frac{\hat{v}_s^{(large)}}{\delta(v_t, \hat{v}_t)}$$

if $v \notin V_s \cap V_t$

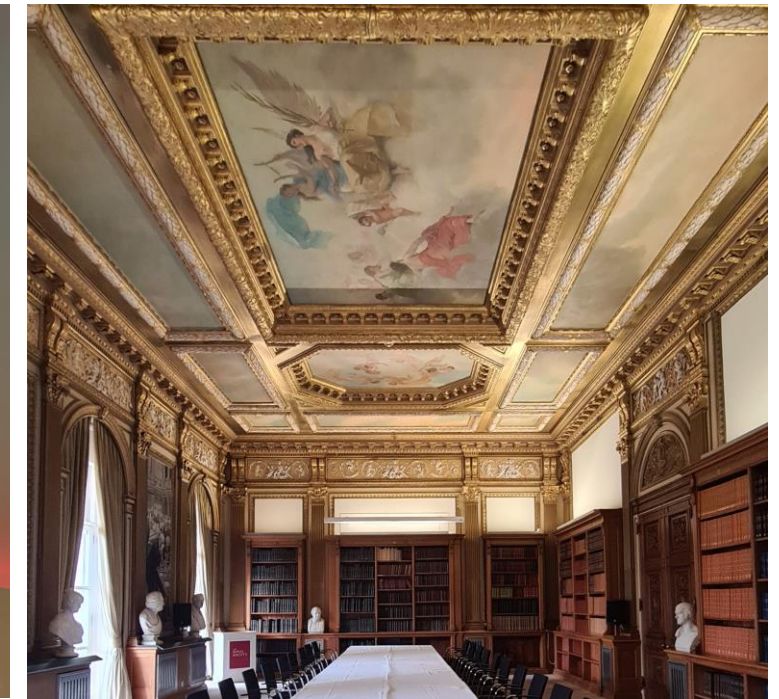
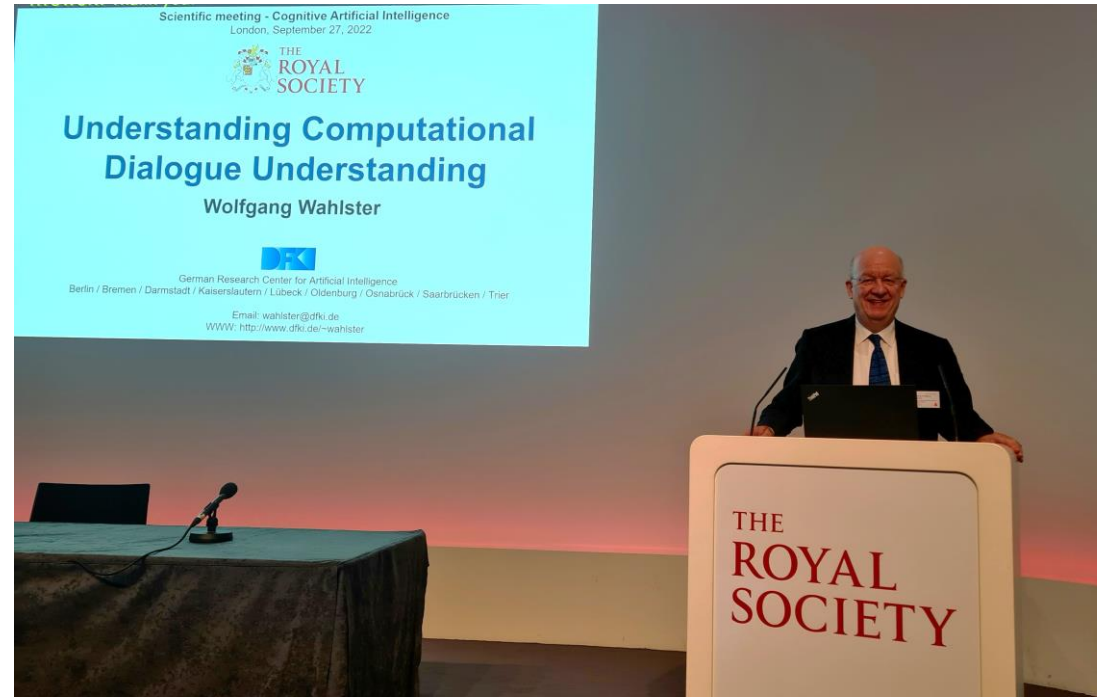
The weight function δ has the objective to transfer the properties from the small model to the large model and is defined as the normalized cosine similarity of the small embeddings of overlapping v and missing \hat{v} tokens:

$$\delta(v, \hat{v}) = \frac{\cos(v_t^{(small)}, \hat{v}_t^{(small)})}{\sum_{\substack{\hat{v}' \in V_s \cap V_t, \\ v' \in V_s \cup V_t}} \cos(v_t'^{(small)}, \hat{v}_t'^{(small)})}$$

DFKI's approach relies on the tokenizers of source and target languages sharing

Robert Hooke Meeting - Cognitive Artificial Intelligence

London, September 26-27, 2022, at RS Founded in 1660, German Leopoldina in 1652



Leopoldina 178 nobel prize winners are members

- Language models lack key features of what we know about humans. Specifically, they are not (Gary Marcus, NYU, 2018)
 - Symbolic: Humans compose symbols and apply rules
 - Grounded: Humans engage with non-linguistic world
- + Nonetheless, they might plausibly learn representations that are functionally equivalent (Elli Pavlik, Stanford 2022). This leaves open the question of whether text-only models could serve as useful models of multimodal discourse despite their different acquisition stories

Peer-Reviewed Special Issue of the Philosophical Transactions A



Since 1665: The oldest continuously-published science journal in the world



Renowned Professors as Speakers at the International Hooke Meeting at the Royal Society:

Europe:

**Alan Bundy (Edinburgh)
Stephen Muggleton (Imperial)**

**Ulrike Hahn (LMU Munich)
Wolfgang Wahlster (DFKI Berlin)**

US:

**Leslie Pack Kaelbling (MIT)
Josh Tenenbaum (MIT)**

**Noah Goodman (Stanford)
Hyowon Gweon (Stanford)**



Seven major research trends for the next generation of computational dialogue systems

- from closed-domain to **open-domain** dialogue systems
- from single-initiative to **mixed-initiative** dialogue systems
- from unimodal to **multimodal** dialogue systems
- from single task to **multitask** dialogue systems
- from monolingual **to multilingual** dialogue systems
- from dyadic dialogues **to multi-party** conversations
- from emotionless dialogues to **emotionally charged** conversations



Two other trends of a more methodological nature have been increasingly adopted in dialogue research over the last five years:

- The move from black-box dialogue processing to **transparent dialogue systems**, that can explain their own processing architecture, knowledge sources and their limitations in order to become a **trustworthy conversational assistant**
- Moving from purely symbolic or neural methods **to hybrid neuro-symbolic methods** that combine the best of these two approaches for advanced dialogue systems

Conclusions

The **scalability** of language models for dialogue systems based on deep learning is a clear advantage for **open-domain** applications. They clearly **outperform** previous approaches for most international benchmarks.

Empirically validated results and **insights from psycholinguistics, neurolinguistics, discourse analysis and other cognitive sciences** about human dialogue understanding are **not explicitly considered as constraints** for deep learning approaches.

The **high compute and memory requirements** and the **huge associated energy consumption** for the training phase of deep learning makes the current approaches **implausible** from the perspective of **cognitive science and brain science**.

Successful end-to-end learning approaches eg. for speech translation using deep learning architectures are **not easily applicable to task-oriented multimodal dialogues** since **training sets** with annotated speaker intent, dialogue partner model and situational context cause **prohibitive costs**.

We believe that hybrid approaches combining statistical approaches with symbolic approaches are the way to achieve deep understanding.

Thank you very much for your attention.

