Sprachtechnologien in KMUs

Ihr Sprecher

Gründer **TRADOS**

CEO **coreon** **SEMIOX**

ESTeam

Vice-Chair **LT-INNOVATE.EU**

lebe in

glücklicher Vater von

FIND, STORE & RETRIEVE
Ability to locate and manage data

MACHINE LEARNING
Learn how to respond to the user by analysing human agent responses

INTENT RECOGNITION
Ability to guess what the user is requesting even if phrased unexpectedly

UNDERSTANDING LANGUAGE INTELLIGENCE
Enhancing different functions of the human brain

LEARN

REASON

INTERACT

UNDERSTAND

PERCEIVE

NATURAL LANGUAGE PROCESSING
Ability to read or parse human language text

TRANSLATION
Ability to convert text and speech from one language to another

SPEECH RECOGNITION & GENERATION
Ability to understand & synthetically produce speech

DIALOGUE MANAGEMENT
Ability to follow conversation history, recall & memorize a single conversation and across conversations for human-like back and forth conversation
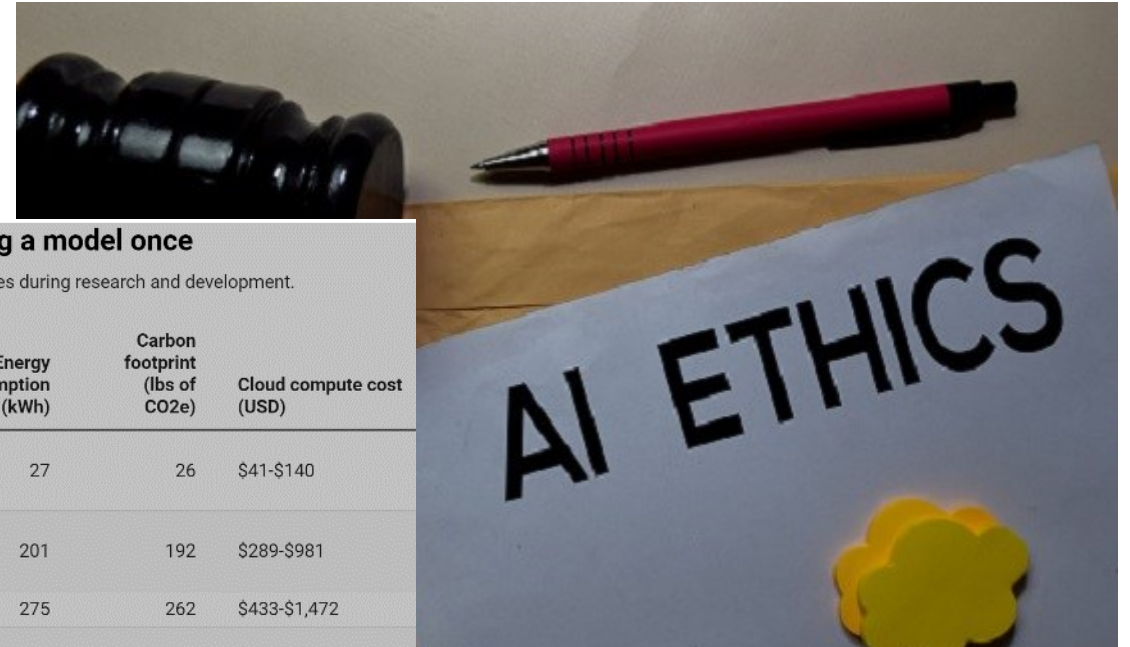
ENTITY RECOGNITION
Understand that some text refers to informative abstract categories (entities) e.g. 15 August = date
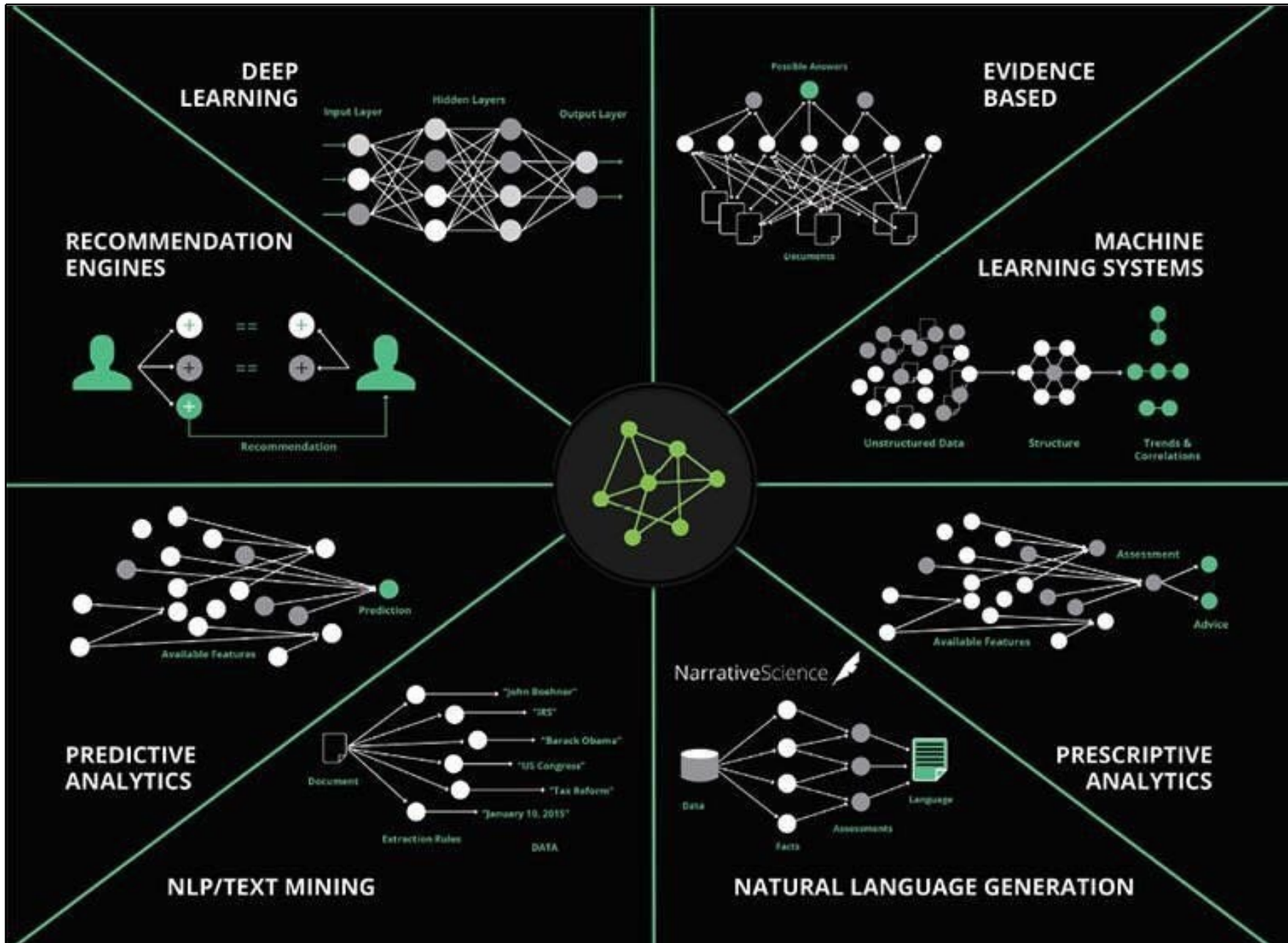
AI = NLP? %?

Schlüsseltechnologie

The estimated costs of training a model once

In practice, models are usually trained many times during research and development.

	Date of original paper	Energy consumption (kWh)	Carbon footprint (lbs of CO2e)	Cloud compute cost (USD)
Transformer (65M parameters)	Jun, 2017	27	26	$41-$140
Transformer (213M parameters)	Jun, 2017	201	192	$289-$981
ELMo	Feb, 2018	275	262	$433-$1,472
BERT (110M parameters)	Oct, 2018	1,507	1,438	$3,751-$12,571
Transformer (213M parameters) w/ neural architecture search	Jan, 2019	656,347	626,155	$942,973-$3,201,722

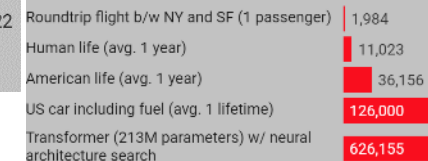## Common carbon footprint benchmarks

in lbs of CO2 equivalent

Roundtrip flight b/w NY and SF (1 passenger)	1,984
Human life (avg. 1 year)	11,023
American life (avg. 1 year)	36,156
US car including fuel (avg. 1 lifetime)	126,000
Transformer (213M parameters) w/ neural architecture search	626,155

Chart: MIT Technology Review • Source: Strubell et al. • Created with Datawrapper

KI-Landschaft
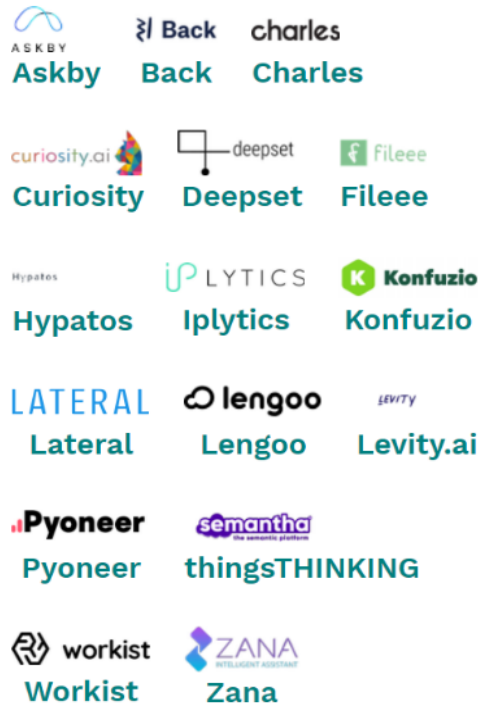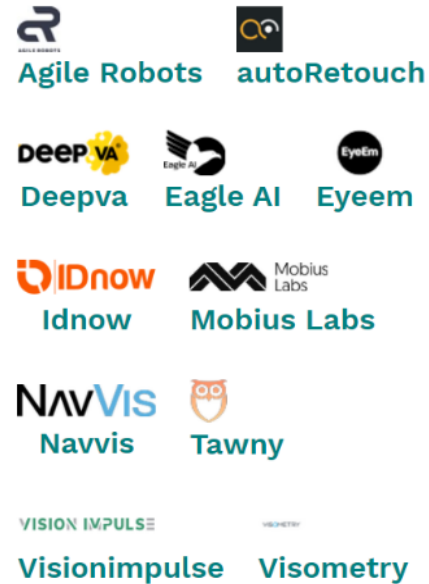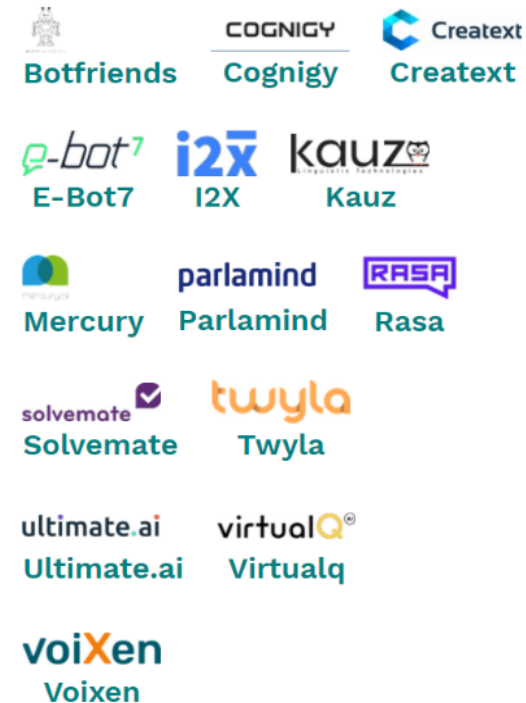
lengoo

20 Millionen Dollar Serie B-Finanzierung: Lengoo expandiert und entwickelt eigenes NMT-Framework weiter

Lengoo erhält 20 Millionen US-Dollar aus Serie B-Finanzierungsrunde. Die nächsten Schritte: expandieren und das firmeneigene NMT-Framework weiterentwickeln.

Lengoo gehört erneut zu den am schnellsten wachsenden Technologieunternehmen

Das Berliner Language-Tech-Unternehmen Lengoo gewinnt den Technology Fast 50 Award und ist damit das dritte Jahr in Folge im renommierten Ranking der Beratungsfirma Deloitte vertreten.

fyrfeed

Einsatz von Künstlicher Intelligenz

Wir entwickeln und nutzen eine Künstliche Intelligenz auf Basis von Machine Learning und NLP. Damit sind wir einer der innovativsten Social Media Anbieter Deutschlands.

Wir schreiben Ihre Social Media Posts

Erhalten Sie hochwertige Social Media Posts – von Experten für Sie geschrieben. Schnell, professionell und sicher.

plusmeta
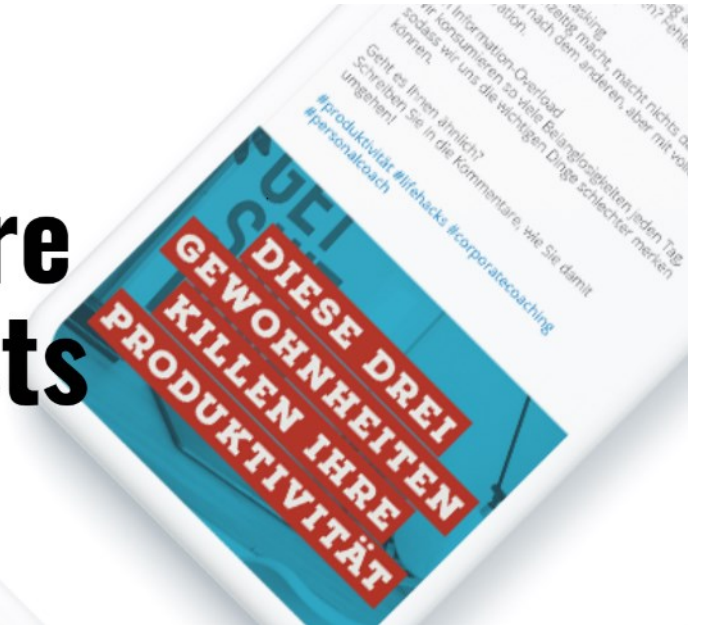
Inhalte einfach klassifizieren

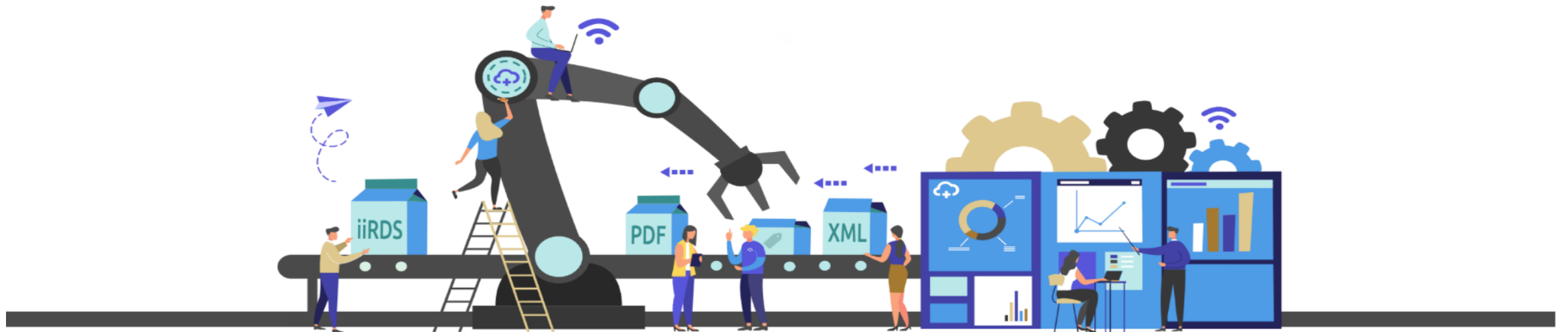Duplikate gezielt identifizieren
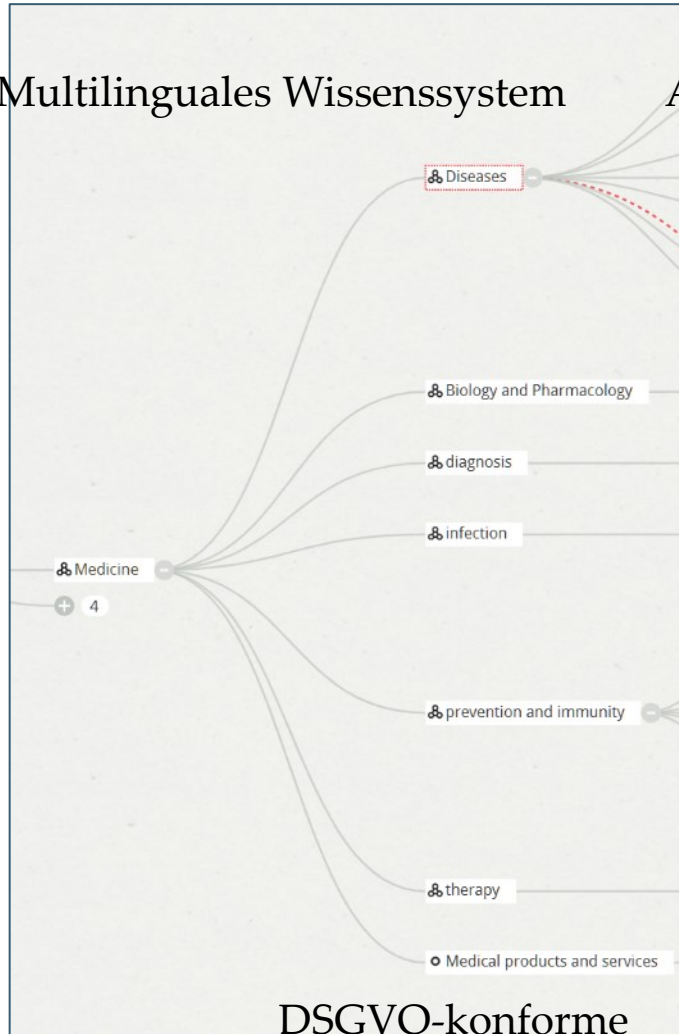
Zulieferdaten integrieren
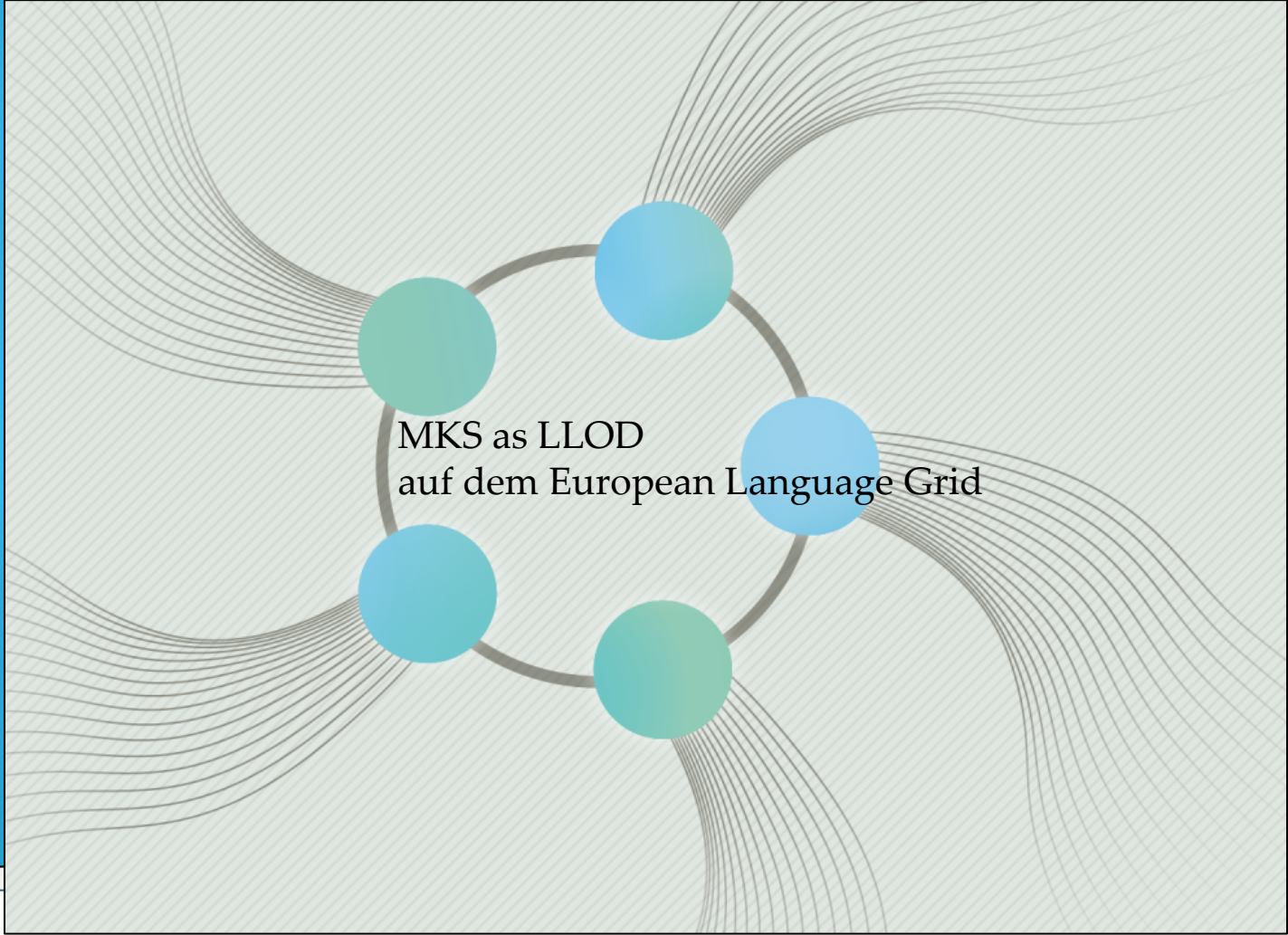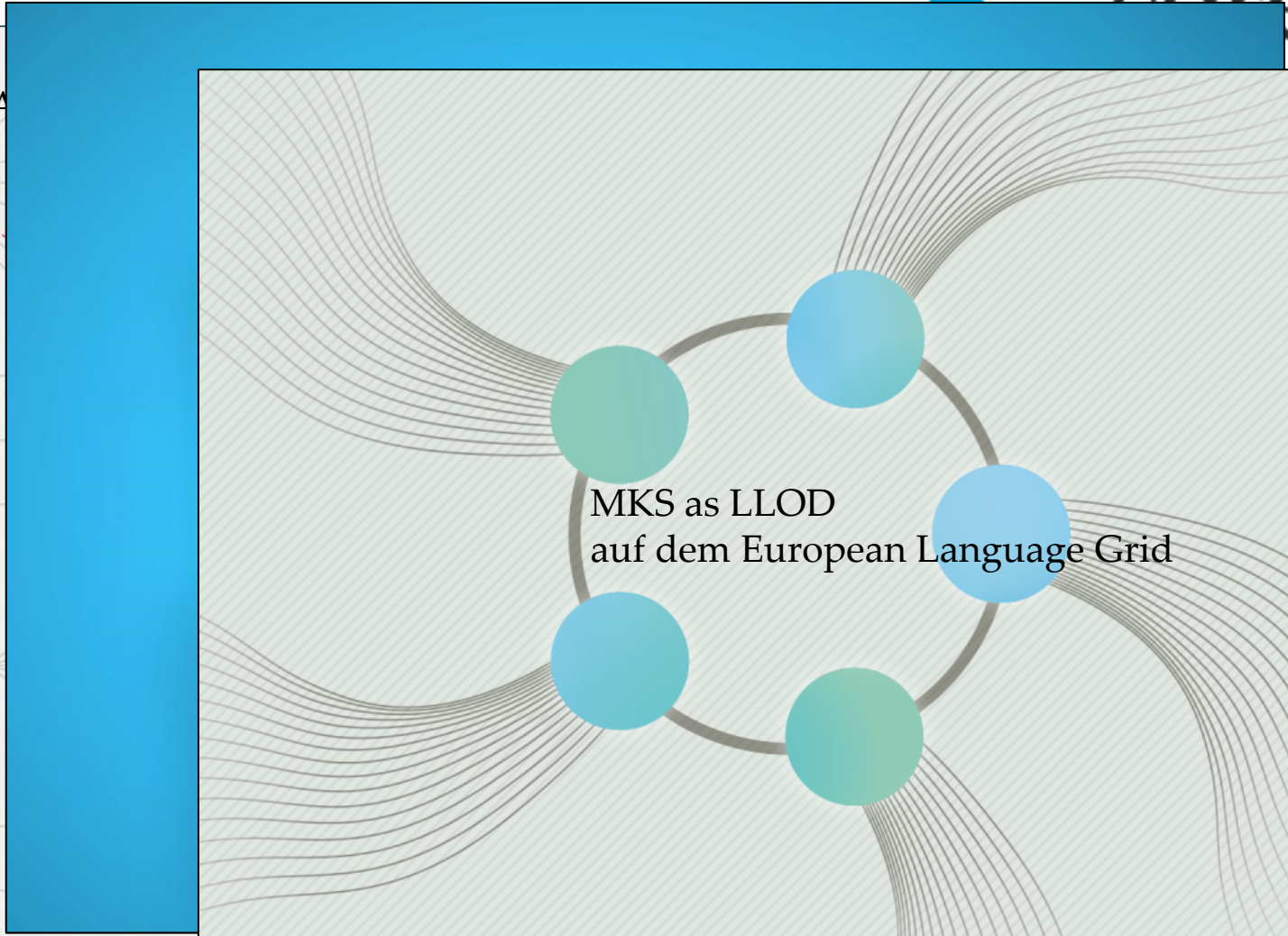
Chatbots initialisieren

Workflows automatisieren

Metadaten standardisieren

Multilinguales Wissenssystem

DSGVO-konforme
Übersetzung

MKS as LLOD
auf dem European Language Grid

spaCy

```
Edit the code & try spaCy                                spaCy v3.0 · Python 3 · via Binder

# pip install -U spacy
# python -m spacy download en_core_web_sm
import spacy

# Load English tokenizer, tagger, parser and NER
nlp = spacy.load("en_core_web_sm")

# Process whole documents
text = ("When Sebastian Thrun started working on self-driving cars at "
        "Google in 2007, few people outside of the company took him "
        "seriously. "I can tell you very senior CEOs of major American "
        "car companies would shake my hand and turn away because I wasn't "
        "worth talking to," said Thrun, in an interview with Recode earlier "
        "this week.")
doc = nlp(text)

# Analyze syntax
print("Noun phrases:", [chunk.text for chunk in doc.noun_chunks])
print("Verbs:", [token.lemma_ for token in doc if token.pos_ == "VERB"])

# Find named entities, phrases and concepts
for entity in doc.ents:
    print(entity.text, entity.label_)

RUN
```

Features

- ✅ Support for **64+ languages**
- ✅ **55 trained pipelines** for 17 languages
- ✅ Multi-task learning with pretrained **transformers** like BERT
- ✅ Pretrained **word vectors**
- ✅ State-of-the-art speed
- ✅ Production-ready **training system**
- ✅ Linguistically-motivated **tokenization**
- ✅ Components for **named entity** recognition, part-of-speech tagging, dependency parsing, sentence segmentation, **text classification**, lemmatization, morphological analysis, entity linking and more
- ✅ Easily extensible with **custom components** and attributes
- ✅ Support for custom models in **PyTorch**, **TensorFlow** and other frameworks
- ✅ Built in **visualizers** for syntax and NER
- ✅ Easy **model packaging**, deployment and workflow management
- ✅ Robust, rigorously evaluated accuracy

Ressourcen, Ressourcen...

Is there something wrong in general with the German model?

kamiwa Kai M. Wadsack — 1✏ Sep '19

I've now had the time to experiment with spaCy and prodigy for several weeks and am very frustrated with the german model's NER accuracy.

Be it ORG, PERSON or LOC entity, out of the box it by far detects too many false positives.
For ORG entities, I'd say its predictions

honnibal ✓ Matthew Honnibal — Sep '19

The models we distribute for spaCy are limited by what training data is available. We've paid licensing fees to get better data for the English parsing and NER data, and somewhat better data for German dependencies. We distribute these models for free, just as we've made the spaCy library free.

However, no resources are available for us to license for German NER --- so we haven't even had the option to buy better data for German. The same is true for the NER data for most of the other languages.

In order to provide *some* sort of free NER model for German, we've ha annotations derived from Wikipedia text semi-automatically. We've tri is unideal in the docs :

When I finally found spaCy with its extensive documentation website, German NER out of the box and prodigy as a tool to improve the shipped German model should it not suffice, I finally believed that the quote

`NER today is regarded as a solved problem in NLP` which I had come across during my web researches, must be true.

While this seems to be true for English it seems that for other languages it is not, at least not without investing considerable time in training and creating a language dependent model. prodigy comes in here as a very handy tool.

Ask me two Questions!

coreon

Knowledge meets language.

ESTeam

Language Technology Software

SEMIOX

M: jochen@coreon.com
T: @JochenHummel
C: +49 172 766 66 33
L: Berlin-Mitte