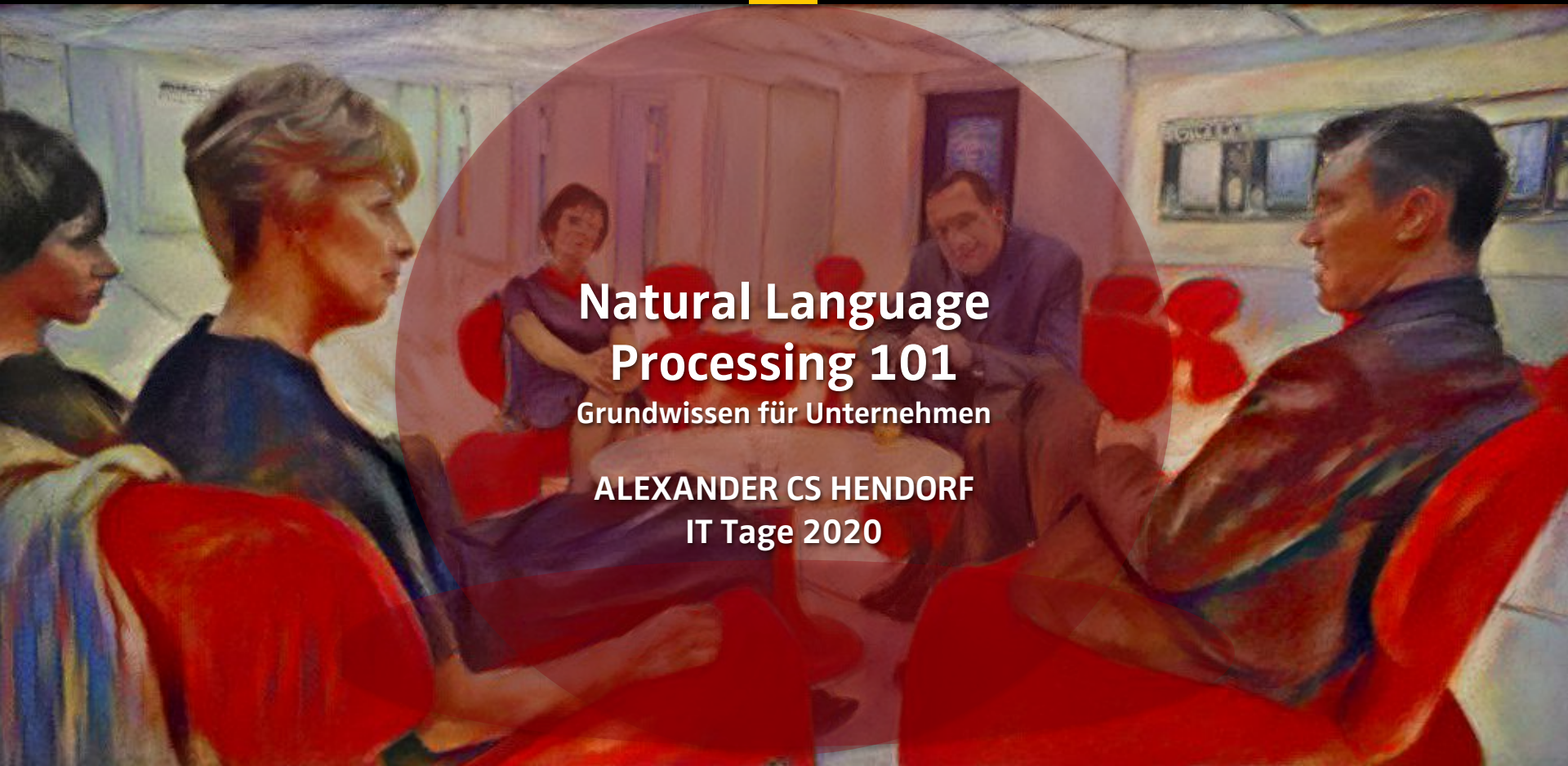
KÖNIGSWEG



Natural Language Processing 101

Grundwissen für Unternehmen

ALEXANDER CS HENDORF

IT Tage 2020

ALEXANDER C. S. HENDORF

MANAGING PARTNER & PRINCIPAL CONSULTANT
DATA SCIENCE & AI AT KÖNIGSWEG.

PYTHON SOFTWARE FOUNDATION FELLOW, PYTHON SOFTWAREVERBAND CHAIR,
PYCONDE & PYDATA BERLIN CHAIR, LOCAL PYDATA COMMUNITY ORGANIZER

@HENDORF

AH@KOENIGSWEG.COM

Back in 2021 🚀😘
Feb/Mar @ Factset

https://www.meetup.com/PyData-Frankfurt

PyCon.DE

PyData Berlin

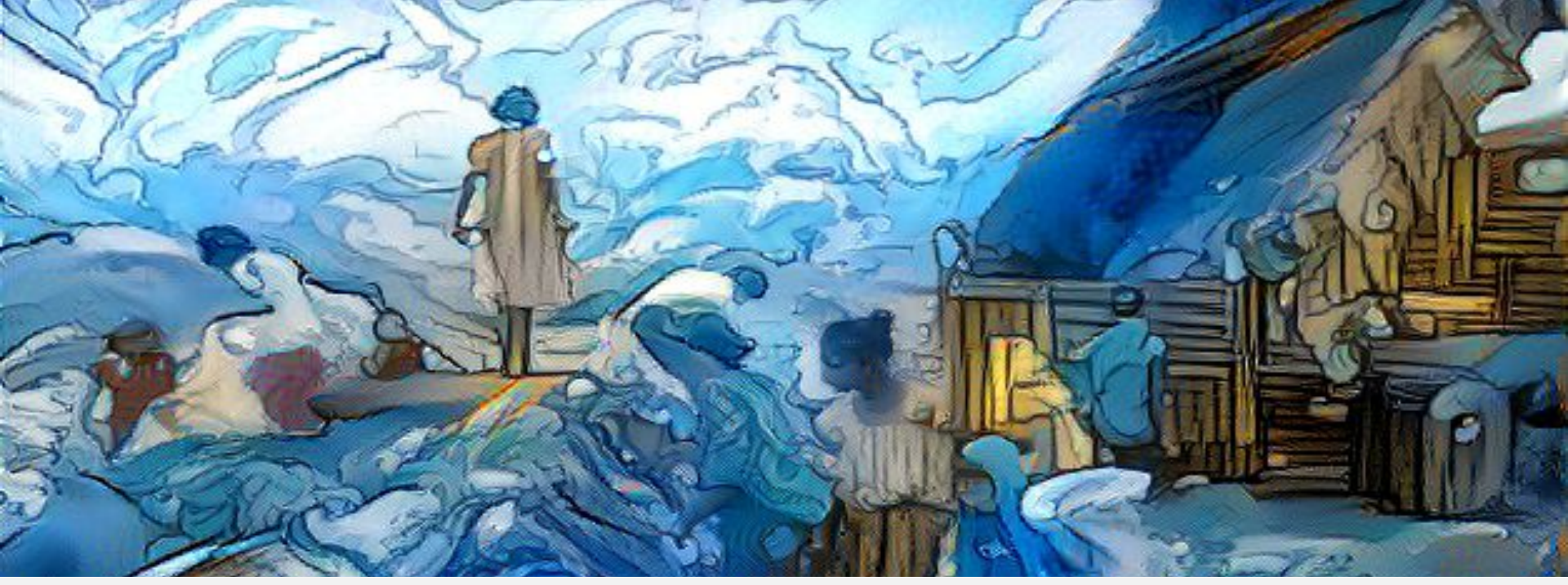
PyConDE & PyData Berlin 2021

Oct 13–15 2021
bcc Berlin Congress Center

Status Quo

Natural Language Processing

KŌNIGSWEG



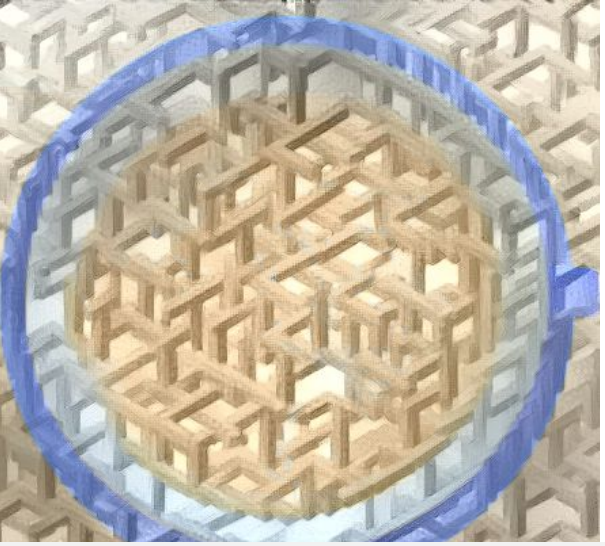
Natural Language Processing (NLP): Eine unstrukturierte Datenlawine

KŌNIGSWEG

Einige Anwendungen von NLP

- Dialogsysteme (Chatbots, Navigation)
- Maschinelle Übersetzung
- Stimmungsanalyse (Sentiment)
- Sprache zu Text – **Text zu Sprache**
- Rechtschreib-/Grammatikprüfung
- Text-Vervollständigung

- **Entity-Recognition**
- **Zusammenfassungen**
- **Knowledge-Graphs**
- **Signal-Extraktion**
- Text-Generierung



- **Strukturierte Daten**
  - Est. 20% of all data
  - Data in databases, format-xyz

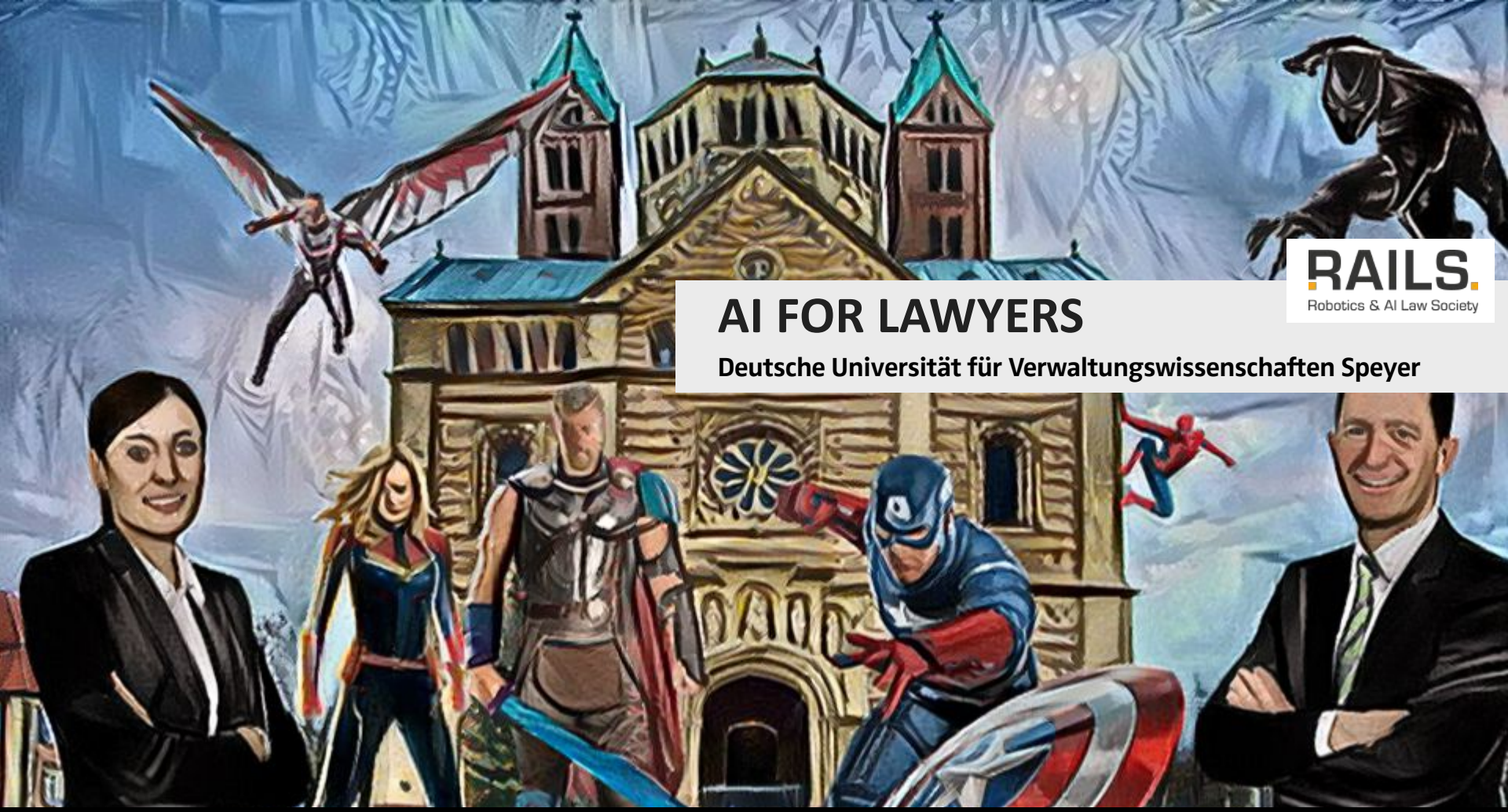
Verschiedene Arten von Daten

- **Unstrukturierte Daten**
  - Est. 80% aller Daten, erfordert eine umfangreiche Vorverarbeitung und verschiedene Analysemethoden
  - Verbale, Text-, Zeichenkommunikation
  - Bilder, Filme, Röntgenaufnahmen

Regelbasiert

Exploratorisch / statistisch

Wissen und Daten aus Textdaten gwinnen.



RAILS.
Robotics & AI Law Society

AI FOR LAWYERS

Deutsche Universität für Verwaltungswissenschaften Speyer

KÖNIGSWEG

One-Size Fits All?

KŌNIGSWEG

HOW TO BUY AI IN 2020

AUFSCHNITT ODER AM STÜCK?

CUT OR PIECE?

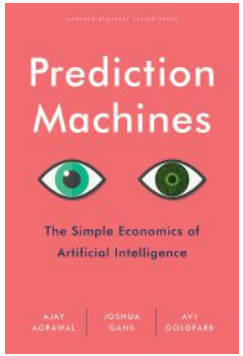
KŌNIGSWEG

Fachidiot KI

KŌNIGSWEG

Besser:
Prediction Machines

- Zutreffenderer Begriff
- Lösen das Problem menschlicher Voreingenommenheit
- Lösen das Problem menschlicher Aufnahmefähigkeit (mehr Parameter)
- Fallende Vorhersagepreise



Treiber

- Leistungssteigerung = fallende Kosten
  Computation (CPU, GPU, et al.)
- Kapazitätsssteigerung = fallende Kosten
  Datenspeicherung & -transport
- Flexible Infrastruktur = Cloud
- Open Source Software
- Offener Zugang zu Information

Erfolg: Nur als Team

Management	Budget, Freiheiten geben, realistische Erwartungen formulieren.
Architect	Planning, Ausführungsüberwachung, Stakeholder & Expectation Management
Domain Expertise	Domänenverständis vermitteln, Mehrwerte hervorheben, Qualitätskontrolle
NLP Skills	Linguistik - Know-How, Chancen und Grenzen versch. Algorithmen
Coding	Umsetzung in production ready-code in Zusammenarbeit mit NLP
Infrastructure	IT Ressourcen
Data Engineering	Datenaufbereitung und Management
Compliance	Datenschutz, Geschäftsgeheimnisschutz, Ethik

KÖNIGSWEG

Vorraussetzungen

- Ordentliche Daten
- Offene Wissenskultur
- Mut zum Experiment
- Allokiertes Budget
- Zugang

Hindernisse

- Halbwissen
- Hype
- Legacy
- Lahmheit

Beispiel-Prozess

Mehrwertgenerierung für eine Domäne

- **Erste Stufe: MVP**
- **Kleines Team**
- **Einfache, stabile Ansätze zuerst**
- **Agil: zweiwöchentliche Entwicklung, Evaluation, ggf. Nachjustierung**
- **Projektdauer 3-6 Monate max.**

- **Überwachter Einsatz**
- **Re-evaluation**

- **Produktion**

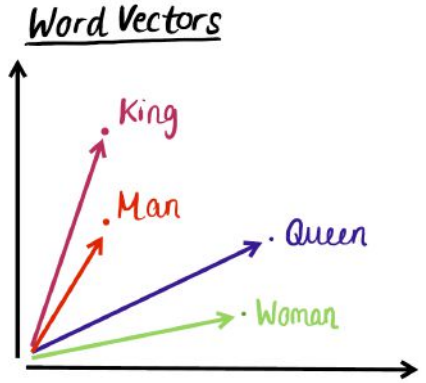

Ab 4. Januar
auf YouTube
#PyData
BETTER CODE FOR
DATA SCIENCE
ALEXANDER CS HENDORF
@ PYDATA GLOBAL 2020

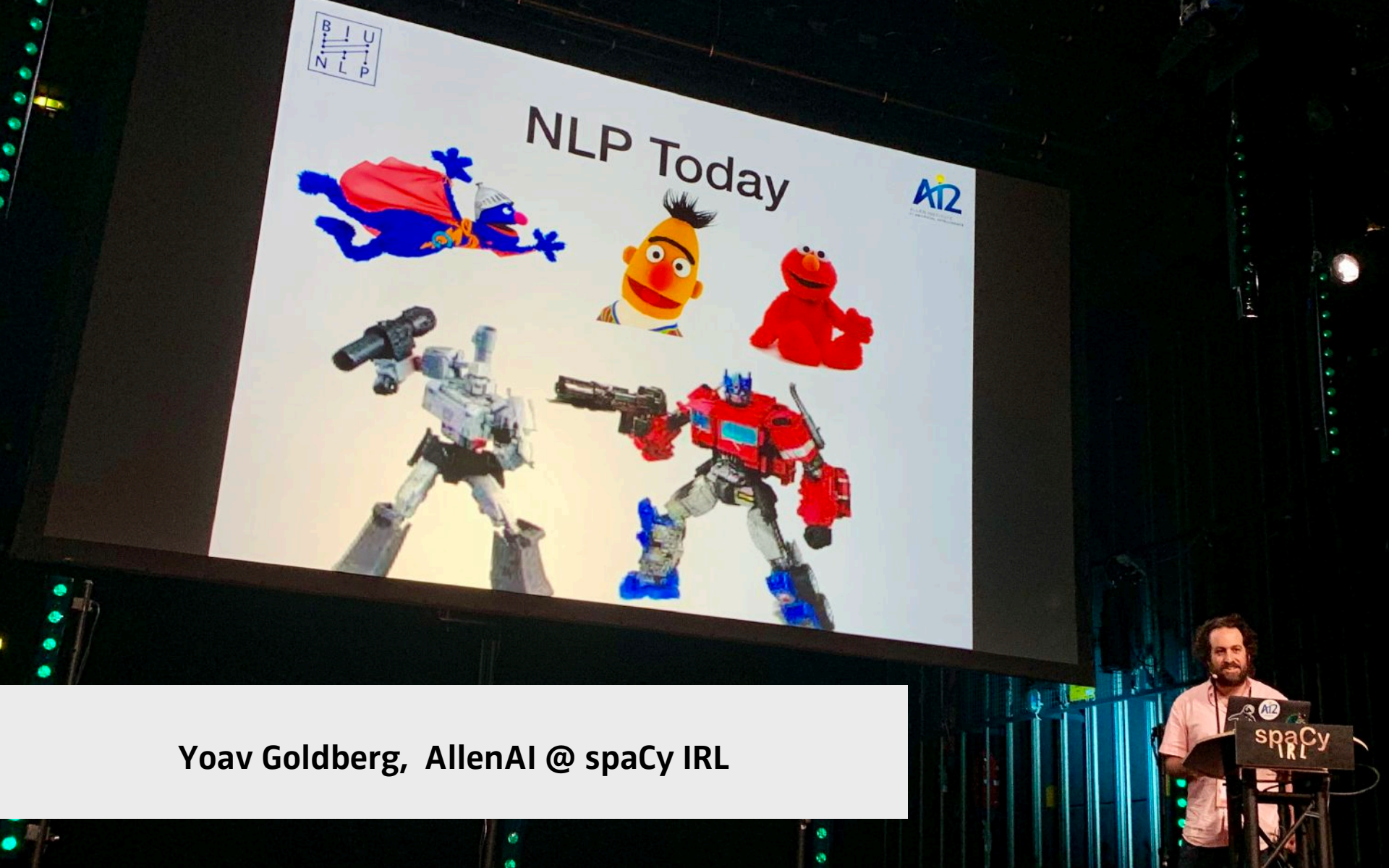
KÖNIGSWEG

Sprachmodelle I

- Vortrainierte statistische Modelle.
  Kontext-frei
  Word2Vec (2013), Glove (2014)

- **Tuning auf eigene Use Cases**



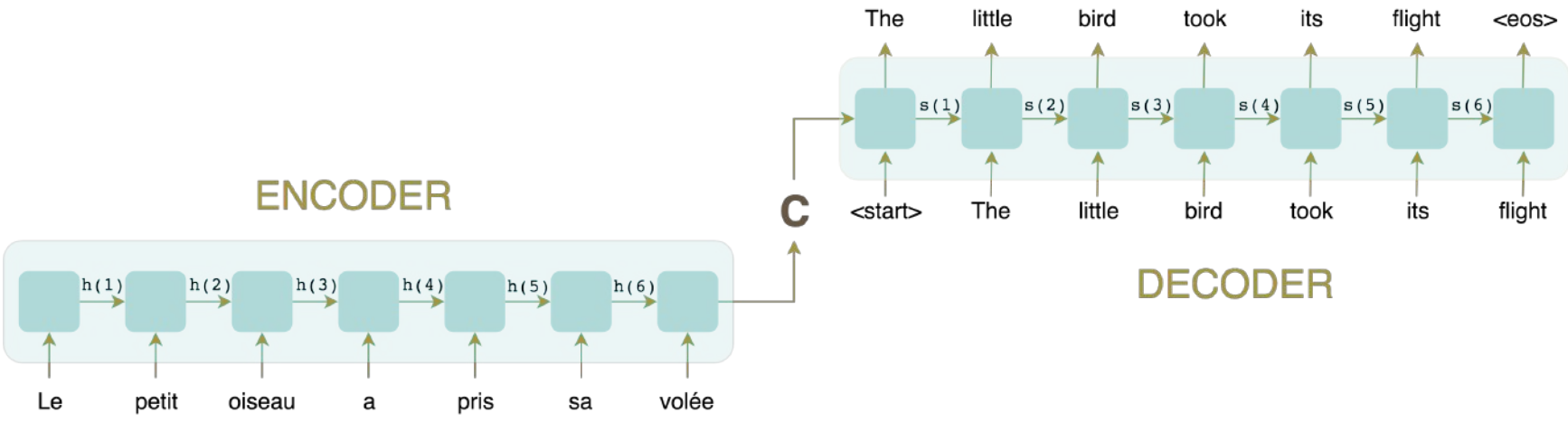


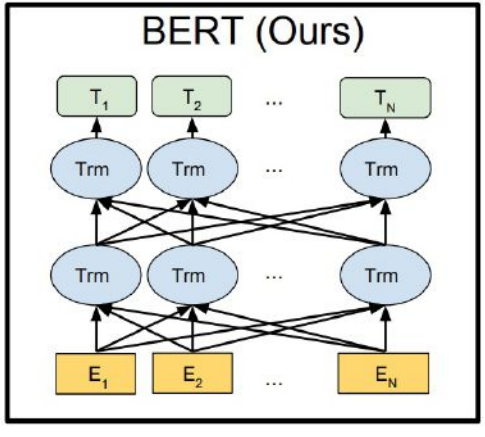
Yoav Goldberg, AllenAI @ spaCy IRL

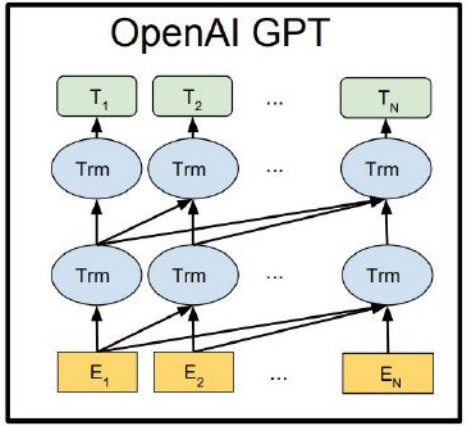
Sprachmodelle II

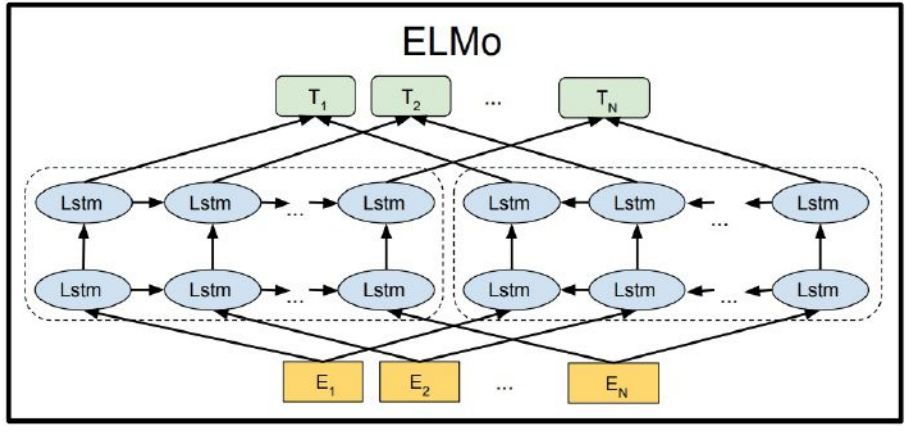
- ***Modelle mit Attention = Kontext***
  ***BERT* (2018)**
- $BERT_{BASE}$ model, a 12-layer, 768-hidden, 12-heads, 110M parameter neural network
- $BERT_{LARGE}$ model, a 24-layer, 1024-hidden, 16-heads, 340M parameter neural network architecture

Transformer

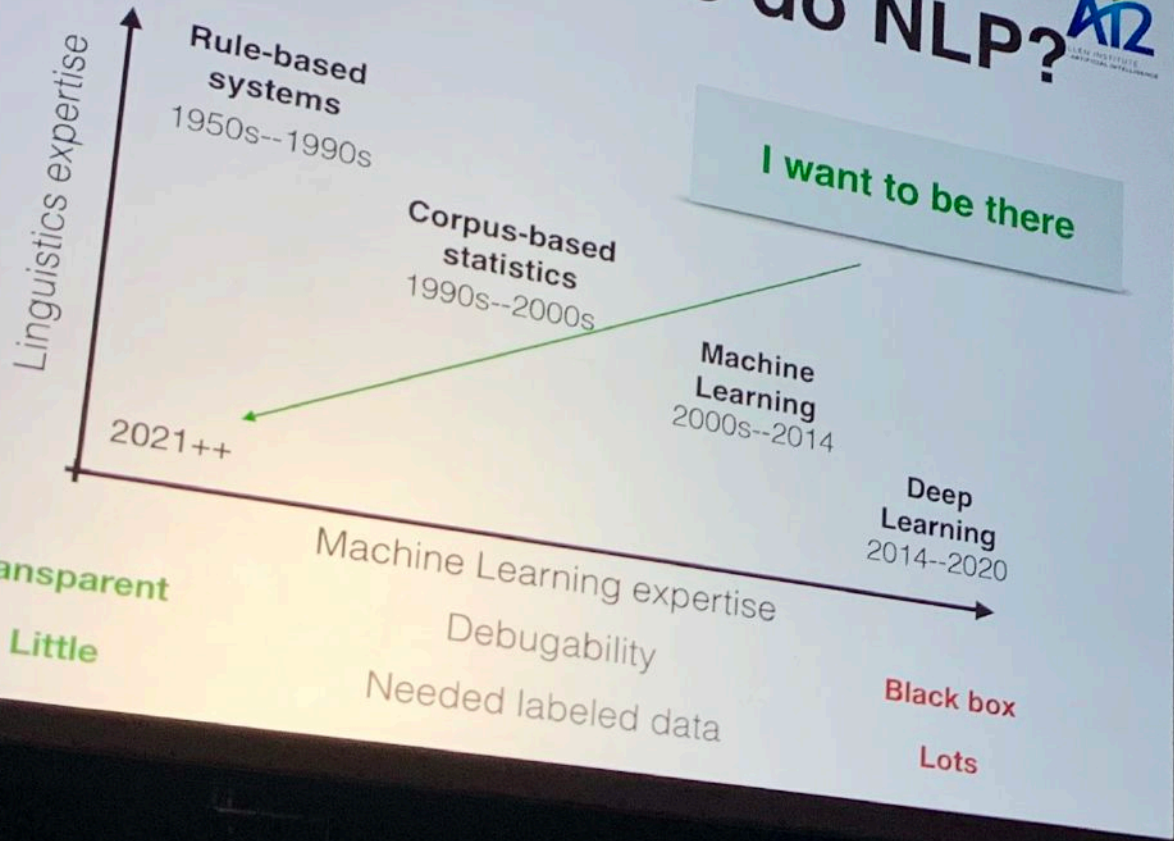






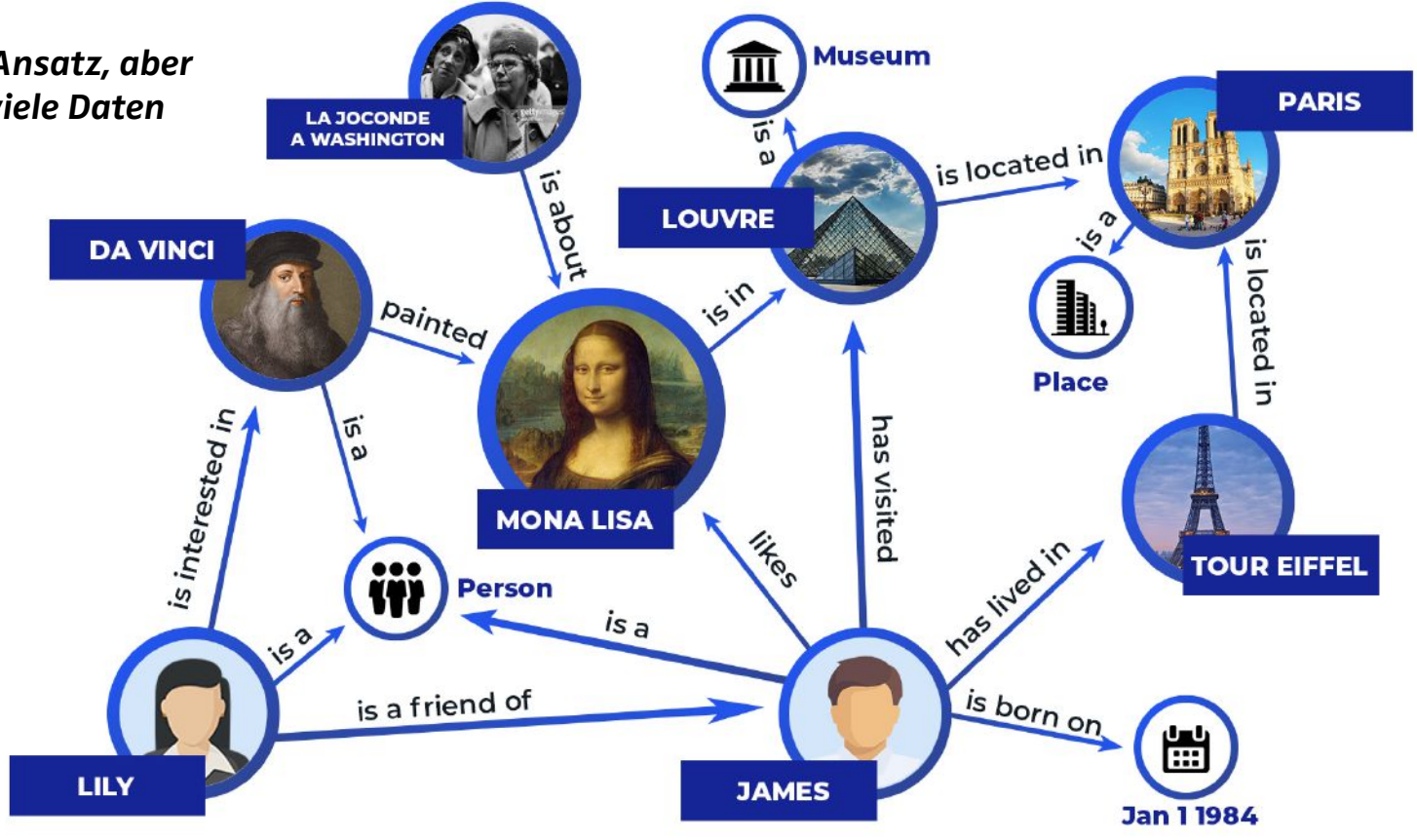


How should we do NLP?



Knowledge Graphs

- *Naheliegender Ansatz, aber benötigt SEHR viele Daten*



NLP Use Cases

Hercule Prototype

Create a new clustering

2 Number of clusters

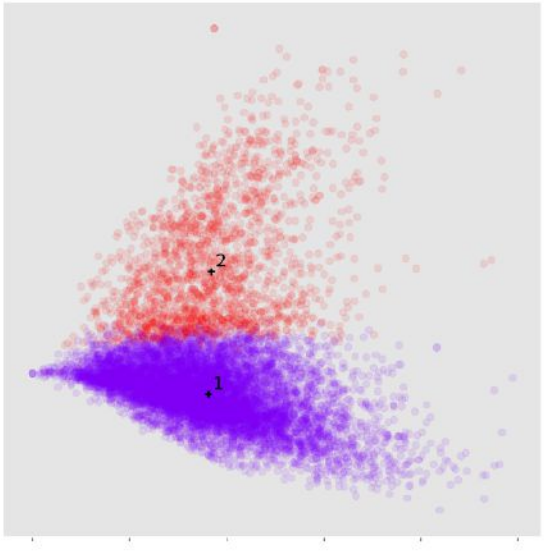
15 Number of keywords

10 Documents in cluster Min.

Get Clusters

Current clustering

1. BIODIVERSITY (8984)
2. RECYCLING (1878)



BIODIVERSITY

This cluster contains 8984 documents.

biodiversity degradation soil
natural resource management
forest conservation student
resources sustainable specie

1. PDF Go to 072ee09815a87ffb1297652b4eca5be7
strategy strategy concept concept concern concern generation generation agency agency

Examples of the application of sustainable strategies to local, national and regional issues, sustainable resource use is basically dependent on the outcome of the cost-increasing effects application of sustainable strategies in a local, national and regional context. environmental concerns about increasing resource use, as well as a review of the contrasting application of sustainable strategies to local, national and regional issues, as well as the role of early development, resources are exploited, the environment is degraded, and income and environmental sustainable development. environmental sustainability and the Southern states demanding development, ensured that The sustainable development strategy also includes areas of economic Conservation Strategy (NCS) and the National Environmental Action Plan (NEAP). Sustainable Development Strategy. Thematic Community Strategy on the Sustainable Use of Resources.

2. DOCX Go to 1be1c96458fa98a319cb2e73b63b1a68

Hercule Prototype

Create a new clustering

5 Number of clusters

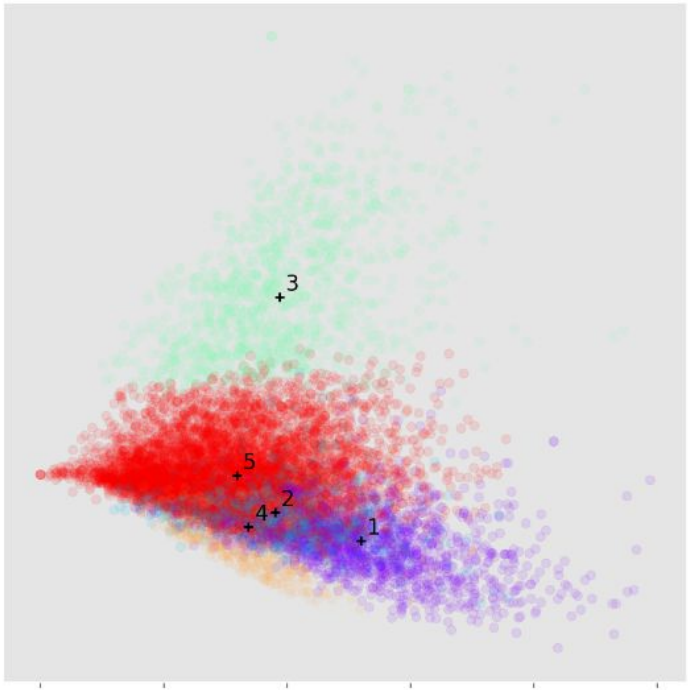
15 Number of keywords

10 Documents in cluster Min.

Get Clusters

Current clustering

1. DEGRADATION (1661)
2. NATURAL (1272)
3. RECYCLING (1138)
4. JOURNAL (515)
5. BIODIVERSITY (6276)



Hercule Prototype

Create a new clustering

2 Number of clusters

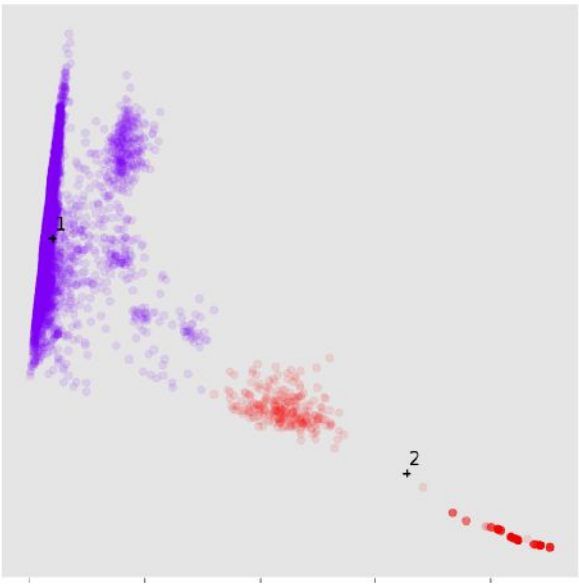
15 Number of keywords

10 Documents in cluster Min.

Get Clusters

Current clustering

1. GUCKY (8391)
2. RASTATT (598)



GUCKY

This cluster contains 8391 documents.

gucky roman atlan bull serie milchstraße terraner planet hauptpersonen roman hauptpersonen tiuphoren

1. PDF Go to 5f9409051e643e0eb7eea58c1a64c1a5
   bully stardust lebewesen rhodans tier figur arkoniden mutant sacha

   Rhodan erwachte nach der Prozedur im gleichen müden und zerschlagenen Zustand, in dem er die Maschine betreten hatte.

2. TXT Go to 5e429794eb66cf14a79592a100225132
   kommandantin all frau gebäude wahl schaden risiko entwicklung platzen besatzung

   »Sie kommen bald wieder auf die Beine«, berichtete sie der Kommandantin.

3. EPUB Go to b9d1a71cd45b2e4987389b89f6bede2a
   feld stein ilt frequenz-monarchie feuern geräusch jules atmosphäre arkonide station

   Er hatte kaum mehr die Kraft, sich auf den Beinen zu halten.

BIODIVERSITY

This cluster contains 8984 documents.

biodiversity | degradation | soil | natural | resource management | forest | conservation | student | resources | sustainable | specie | country | university | journal | plant

explore this cluster

1. PDF Go to 072ee09815a87ffb1297652b4eca5be7

strategy strategy concept concept concern concern generation generation agency agency

Examples of the application of sustainable strategies to local, national and regional issues, sustainable resource use is basically dependent on the outcome of the cost-increasing effects application of sustainable strategies in a local, national and regional context. environmental concerns about increasing resource use, as well as a review of the contrasting application of sustainable strategies to local, national and regional issues, as well as the role of early development, resources are exploited, the environment is degraded, and income and environmental sustainable development. environmental sustainability and the Southern states demanding development, ensured that The sustainable development strategy also includes areas of economic Conservation Strategy (NCS) and the National Environmental Action Plan (NEAP). Sustainable Development Strategy. Thematic Community Strategy on the Sustainable Use of Resources.

2. DOCX Go to 1be1c96458fa98a319cb2e73b63b1a68

impact impact emission emission human human may may greenhouse greenhouse

Human impact on the environment or anthropogenic impact on the environment includes changes to biophysical environments[1] and ecosystems, biodiversity, and natural resources[2][3] caused directly or indirectly by humans, including global warming,[1][4] environmental degradation[1] (such as ocean acidification[1][5]), mass extinction and biodiversity loss,[6][7][8][9] ecological crisis, and ecological collapse. Environmental impacts associated with meat production include use of fossil energy, water and land resources, greenhouse gas emissions, and in some instances, rainforest clearing, water pollution and species endangerment, among other adverse effects.[57][58] Steinfeld et al.

3. PDF [illegible]

[illegible] will be taken to conserve and 2. Reclassify the natural forests of the country [illegible] addition within the country as far as Promote water resources development techniques with environmentally safe [illegible] conserving the marine and coastal environment.

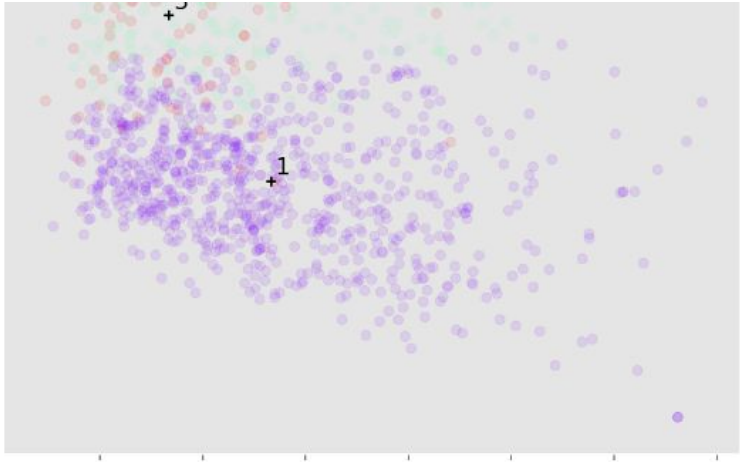
4. [illegible]

Since its March 2015 take off, the Swiss-engineered Solar Impulse 2 has made [illegible] energy consumption, save natural resources and improve quality of life [illegible] of global warming. For example, some geo-engineering techniques are based on carbon sequestration [illegible] warming.

5. [illegible]

Current clustering

1. RECYCLING (798)
2. PLASTIC (249)
3. UPCYCLING (91)



RECYCLING

This cluster contains 798 documents.

recycling | recycle | scrap | metal | collection | paper | services | trash | landfill | city | equipment | county | company | program | bin

explore this cluster

1. PDF Go to 032ab17b6fa507c086f19d19db3f830d

car car garden garden facebook facebook thing thing green green

Biodiversity | RESET.org Login with Facebook Login/Register But the huge numbers of people around the world buying water in bottles rather than taking it straight from the tap is a trend that has huge negative consequences for the environment. Most of the conventional cleaning products we all grew up with are petroleum-based and have dubious health and environmental implications. Tips for Sustainable Travel We are often equally unaware of the implications our travel has on global environmental issues, especially climate change. 12 Things You Can Do Right Now on Climate Change Below is a list of 12 easy things you can do right now to help fight climate change. These use four times less energy and they last eight times longer. Green New Deal RESET News

2. TXT Go to 089adbc7e7511a209d49bef523eab43f

comment comment test test question question measure measure course course

Measures to Control of Environmental Degradation - Manufacturing Industries - Everonn - CBSE Class 10th Course and NCERT Solutions 6. Manufacturing Industries Contribution of industry to National Economy Agro Based Industries Mineral Based Industries Industrial Pollution and Environmental Degradation Air pollution can be reduced by selection of proper fuel and fitting smoke stacks to factories with electrostatic precipitators, fabric filters,

NLP Alchemy Toolset

KŌNIGSWEG

Erfolgsrezept

- Agile Umsetzung, 2 Wochen Sprints
- Kommunikation von Erfolgen als auch Misserfolgen
- Mehrwertorientierung inkl. Stetiger Nachsteuerung
- Erfahrene Domän-Experten bei Bedarf hinzugezogen
- Open Source, Unterstützung durch Community [Give+Take!]
- Zielgerichte Architektur
- Usability

1 - What is spaCy?

2 – Building Blocks I.

- **Tokenization**
  Segmenting text into words, punctuations marks
- **POS (Part-of-speech tagging)**
  cat -> noun, scratched -> verb
- **Lemmatization**
  cats -> cat, scratched -> scratch
- **Sentence Boundary Detection**
  Hello, Mrs. Poppins!
- **NER (Named Entity Recognition)**
  Marry Poppins -> person, Apple -> company
- **Serialization**
  Saving NLP documents

2 – Building Blocks II.

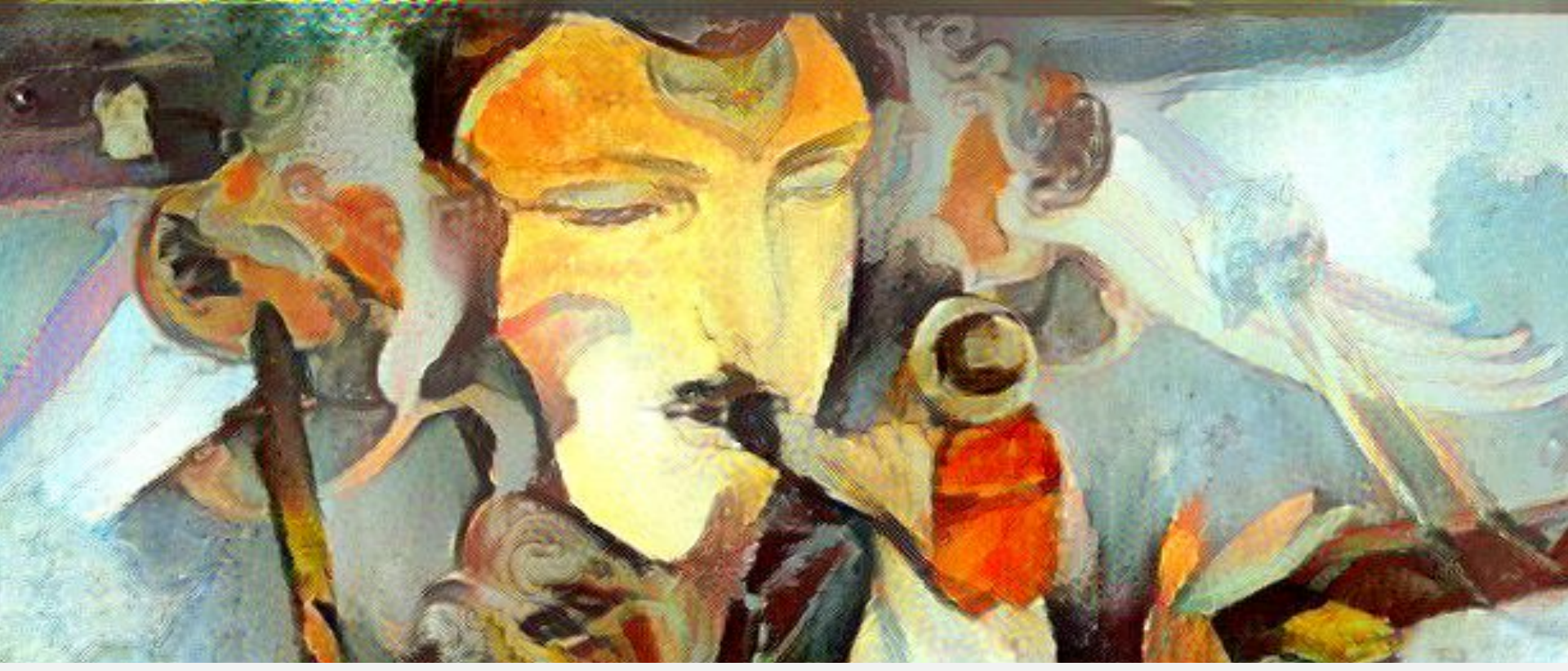
- **Dependency Parsing**
  Dependencies between tokens in a sentence

- **Entity Linking**
  Connecting multiple textual entities to a unique identifier

- **Training**
  Train and update statistical models

- **Text Classification**
  Label whole or parts of a document

- **Rule-based Matching**
  {"label" : "PRODUCT",
  "pattern": [{"LOWER": 'Apple', 'OP' : '?'}



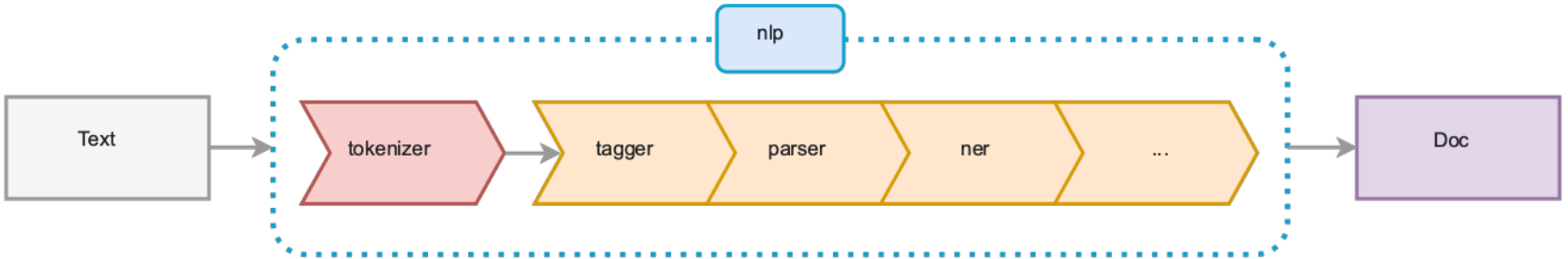
Pipelines

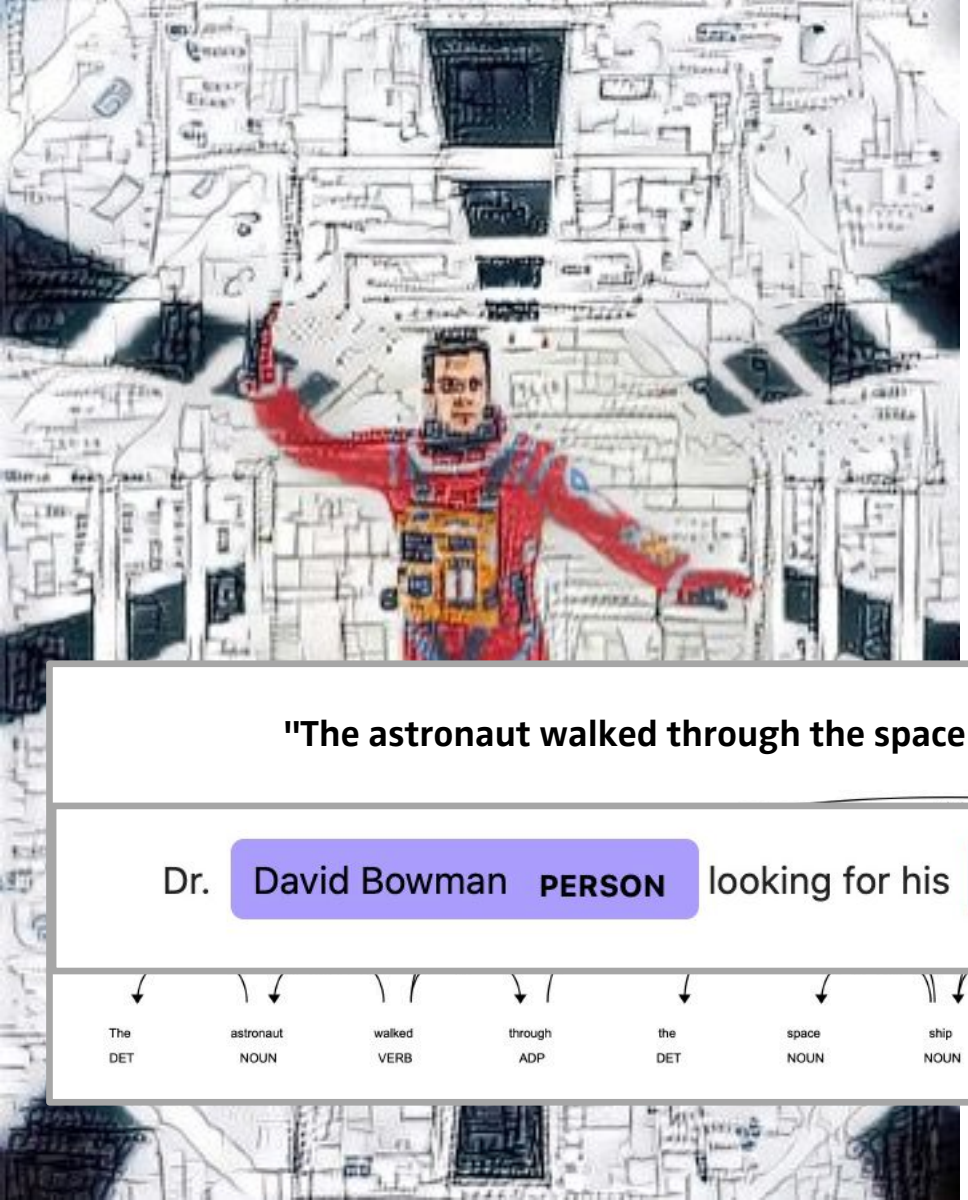
KŌNIGSWEG

Pipelines

- **Default Pipeline:**
  tagger, parser and entity recognizer

- **Custom Pipelines**
  One may add, remove pipeline steps permanently or temporarily





Displacy

- Comes now with spaCy
- Visualize dependencies
- Visualize entities

Speech Synthesis with Tacotron2 and PyTorch

PyData Amsterdam 2019 @ booking.com

- Sprachsynthese
- Geringe Erfolgserwartung
- Transkiption auf Deutsch über Cloudservices unter Erwartung
- Datenbeschaffung auf deutsch schwierig
- Ergebnis überraschend gut

Talk auf Youtube:

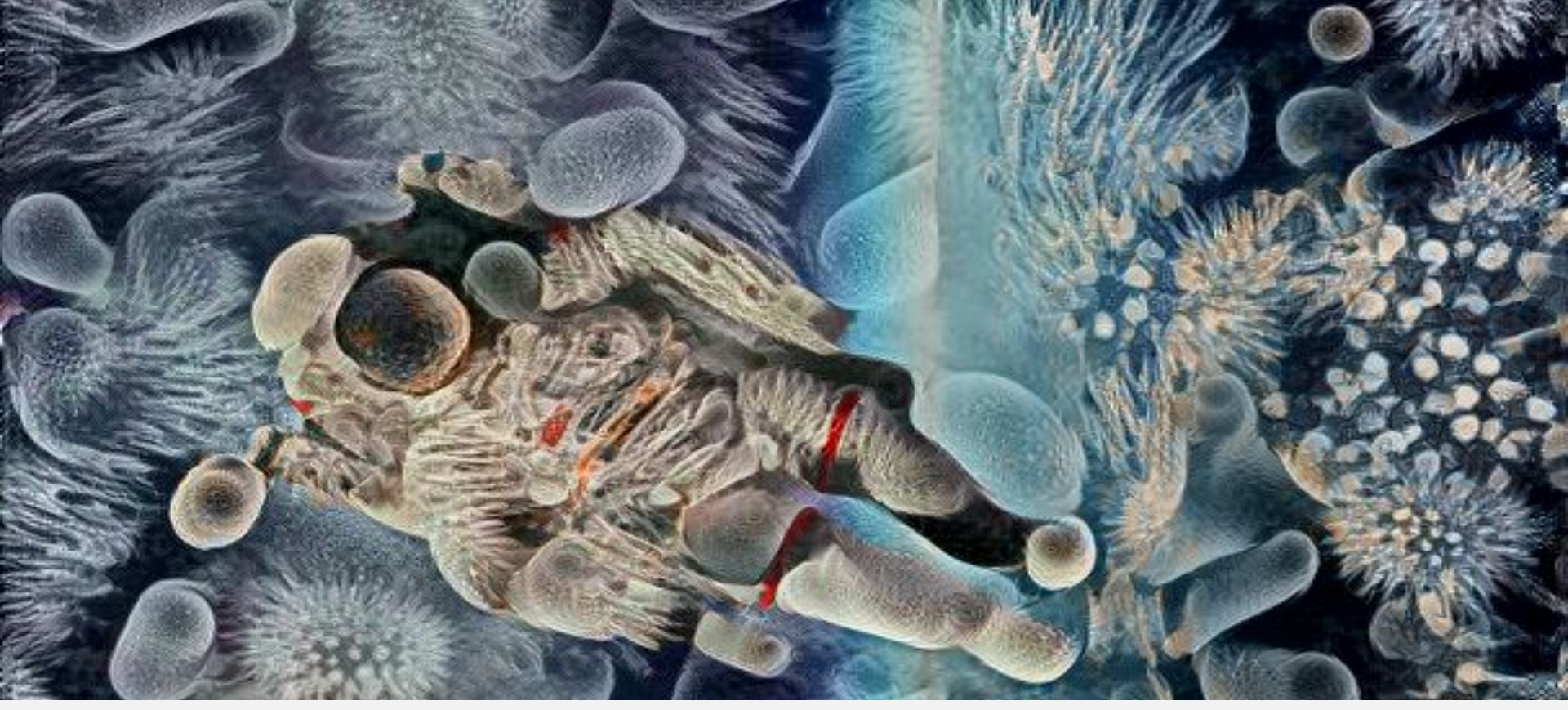
https://youtu.be/ijhZR43TOwc

Once upon a time there was a little mermaid named Siren, who lived with her step mother under the sea, She didn't get to go out of the Sea like any other."

— 10 hrs 

— 9 days 

— 14 hrs 



Thank you & 👋 stay healthy!

KŌNIGSWEG

Thank you & 👋 stay healthy!

KŌNIGSWEG

KŌNIGSWEG

Thank you!

We're hiring

koenigsweg.com

ah@koenigsweg.com

@hendorf