KŌNIGSWEG

15 THINGS YOU SHOULD KNOW ABOUT SPACY

ALEXANDER CS HENDORF @ EUROPYTHON 2020

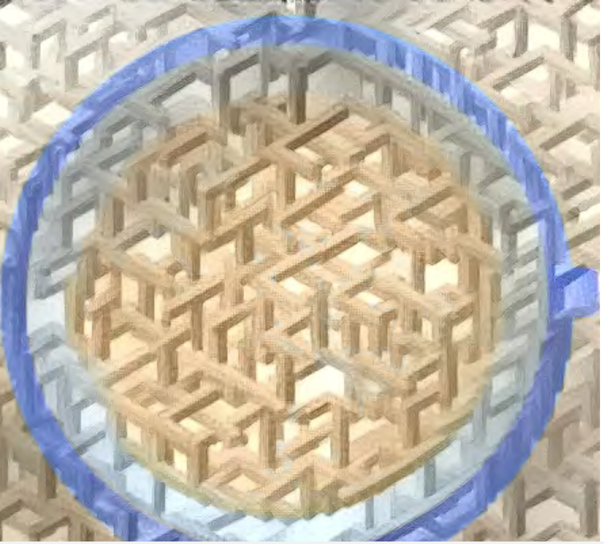
0 -preface-

Natural Language Processing

KŌNIGSWEG

Natural Language Processing (NLP): A Unstructured Data Avalanche

KŌNIGSWEG



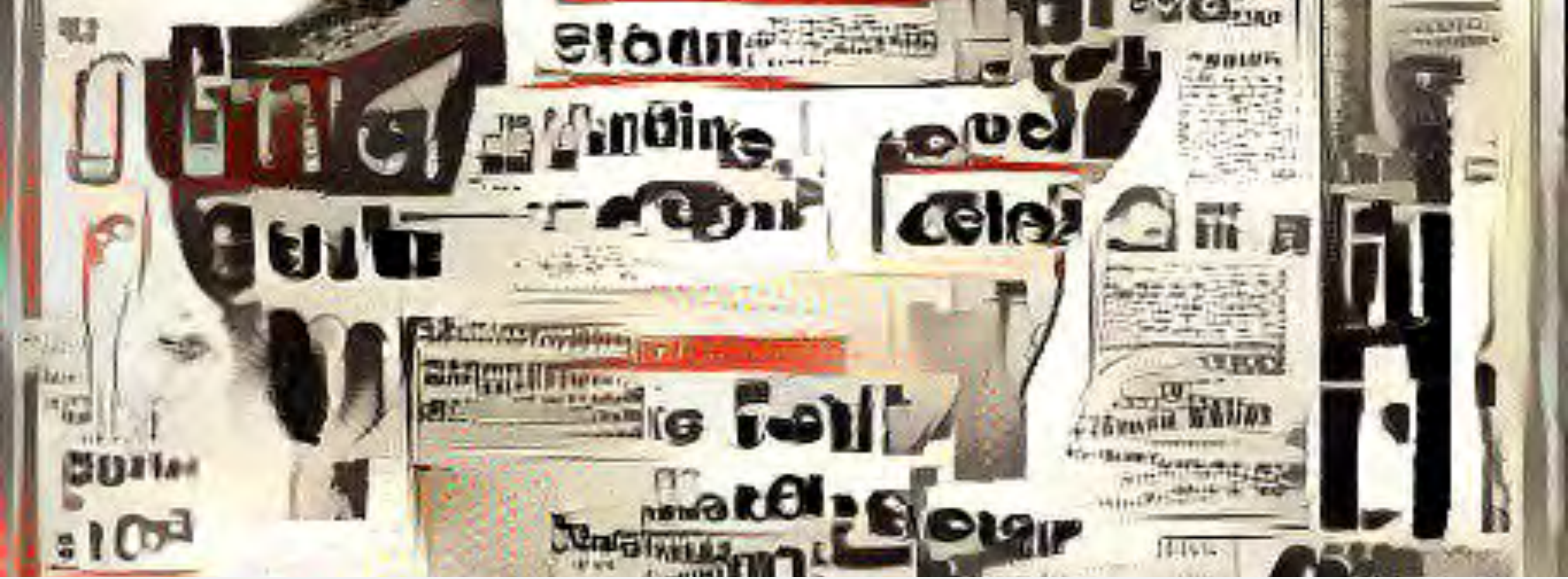
- **Structured Data**
  - Est. 20% of all data
  - Data in databases, format-xyz

Different Kinds of Data

- **Unstructured Data**
  - Est. 80% of all data, requires extensive pre-processing and different methods for analysis
  - Verbal, text, sign- communication
  - Pictures, movies, x-rays
  - Pink noise, singularities

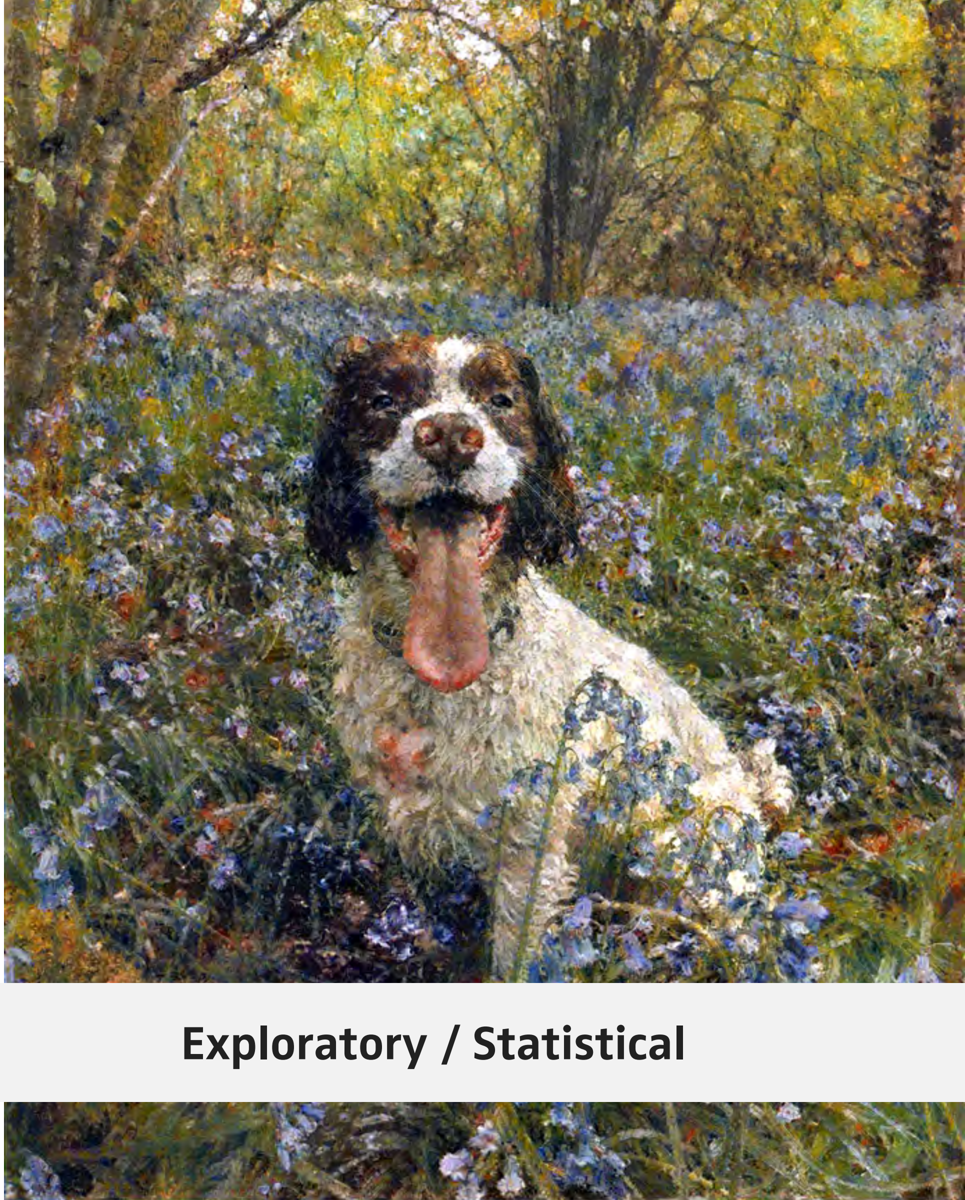
Some Applications of NLP

- Dialogue systems (Chatbots)
- Machine Translation
- Sentiment Analysis
- Speech-to-text and vice versa
- Spelling / grammar checking
- Text completion



Many Smart Things in Text Data!

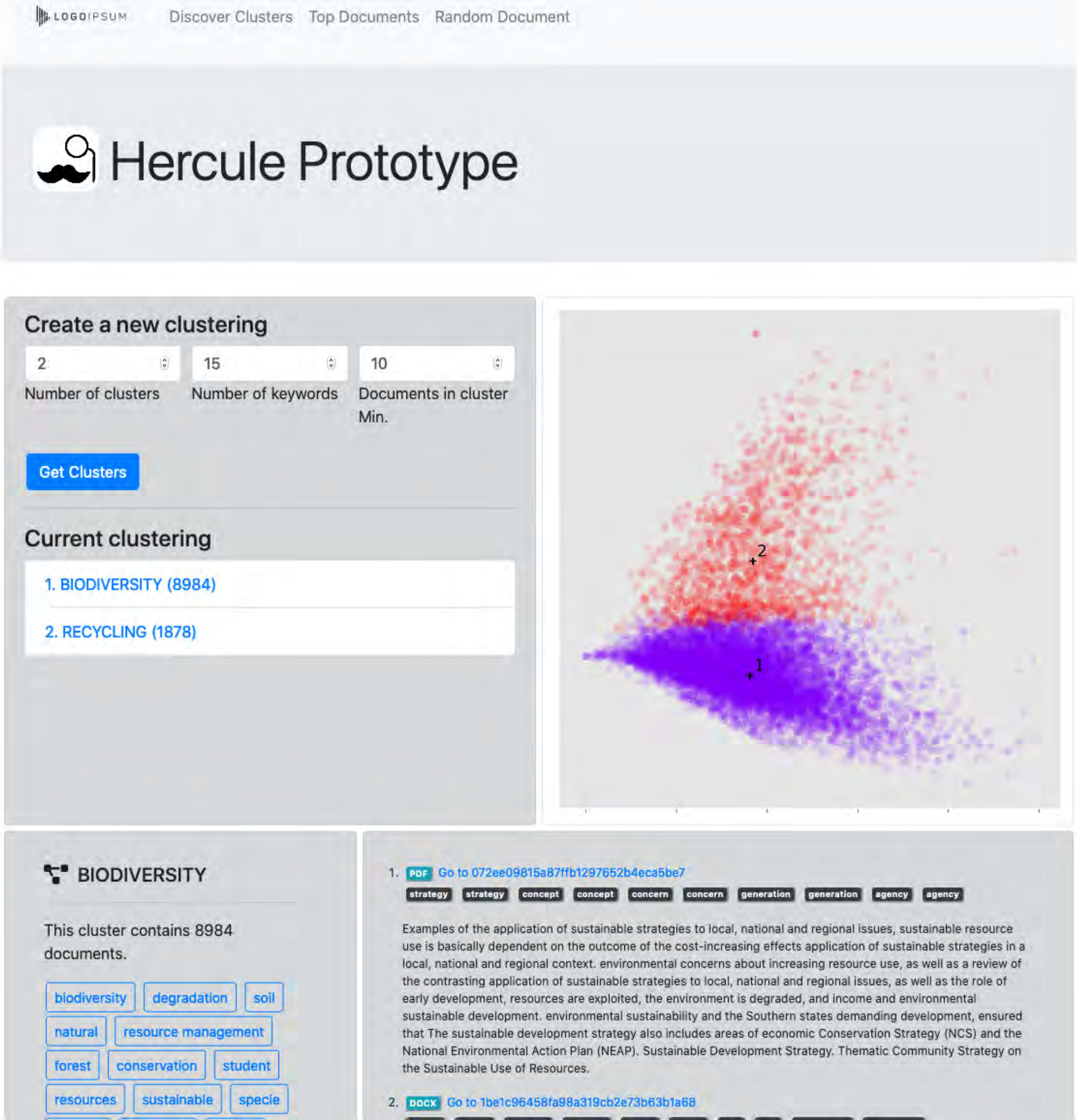
Rule-Based



Exploratory / Statistical

NLP Use Cases

KŌNIGSWEG



Hercule Prototype

Create a new clustering

Current clustering

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

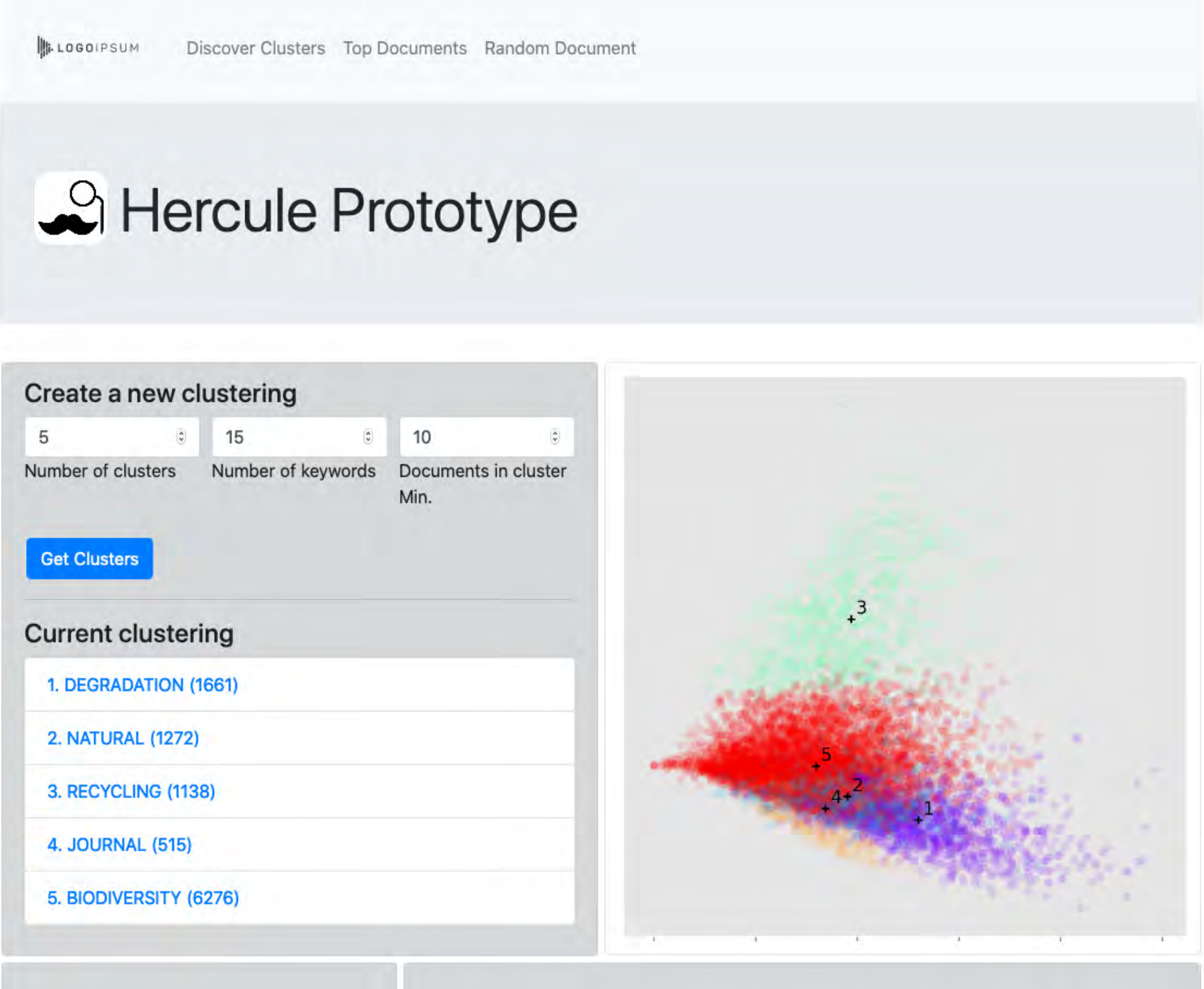
BIODIVERSITY

This cluster contains 8984 documents.

1. PDF Go to 072ee09815a87ffb1297652b4eca5be7

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

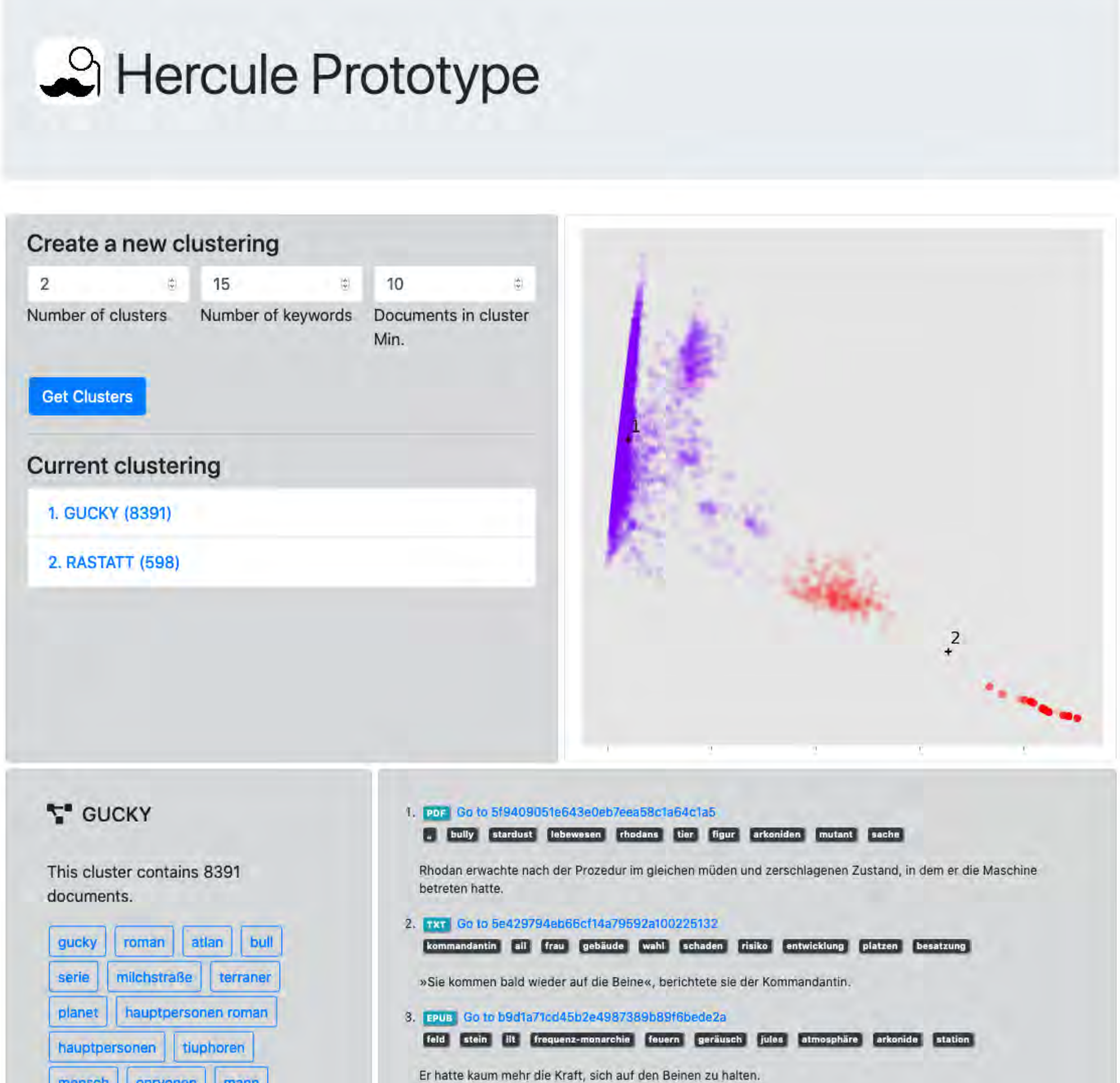
Number of clusters: 5

Number of keywords: 15

Documents in cluster Min.: 10

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.

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 mensch onryonen mann

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

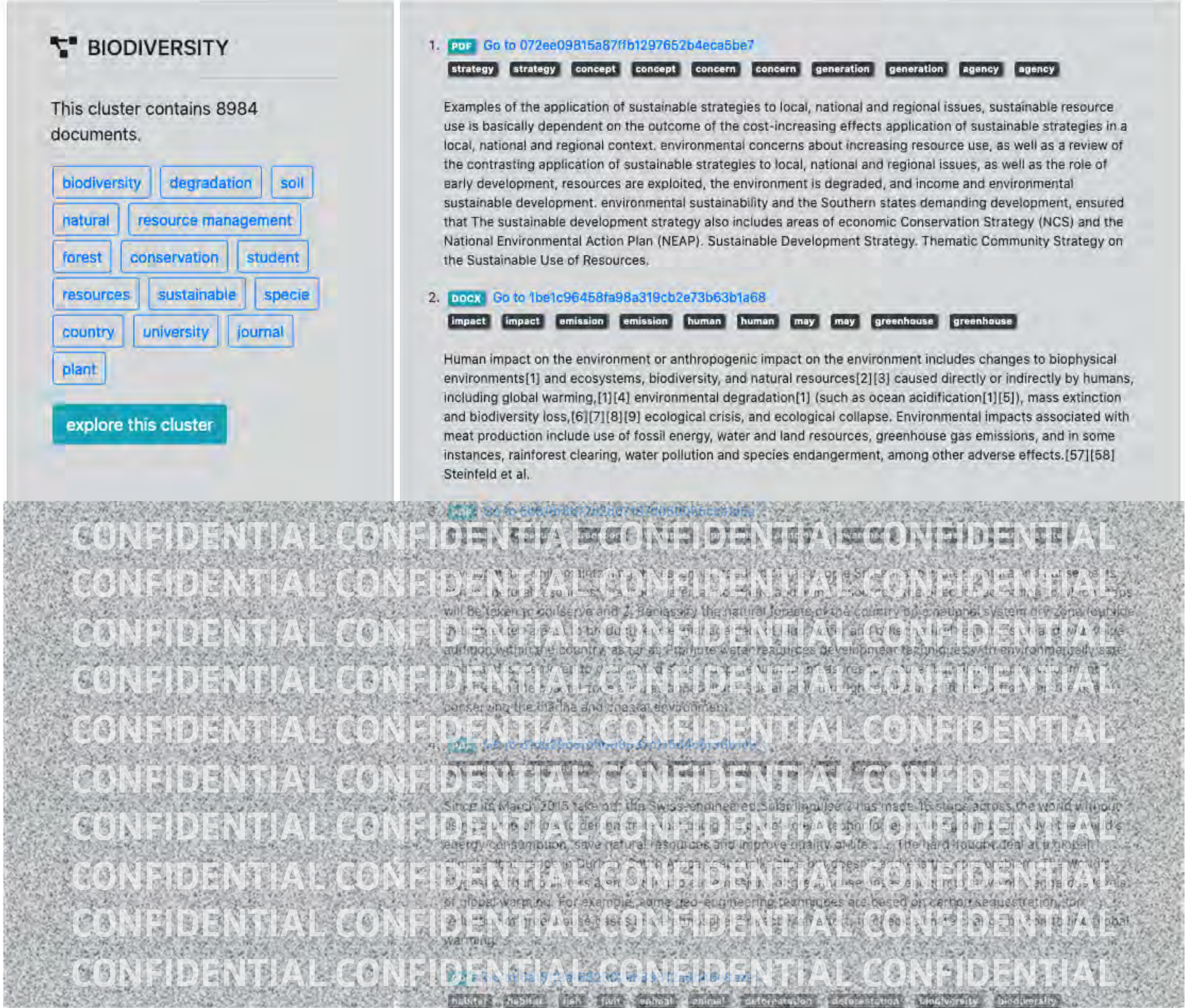
   Rhodan erwachte nach der Prozedur im gleichen müden und zerschlagenen Zustand, in dem er die Maschine betreten hatte.

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.

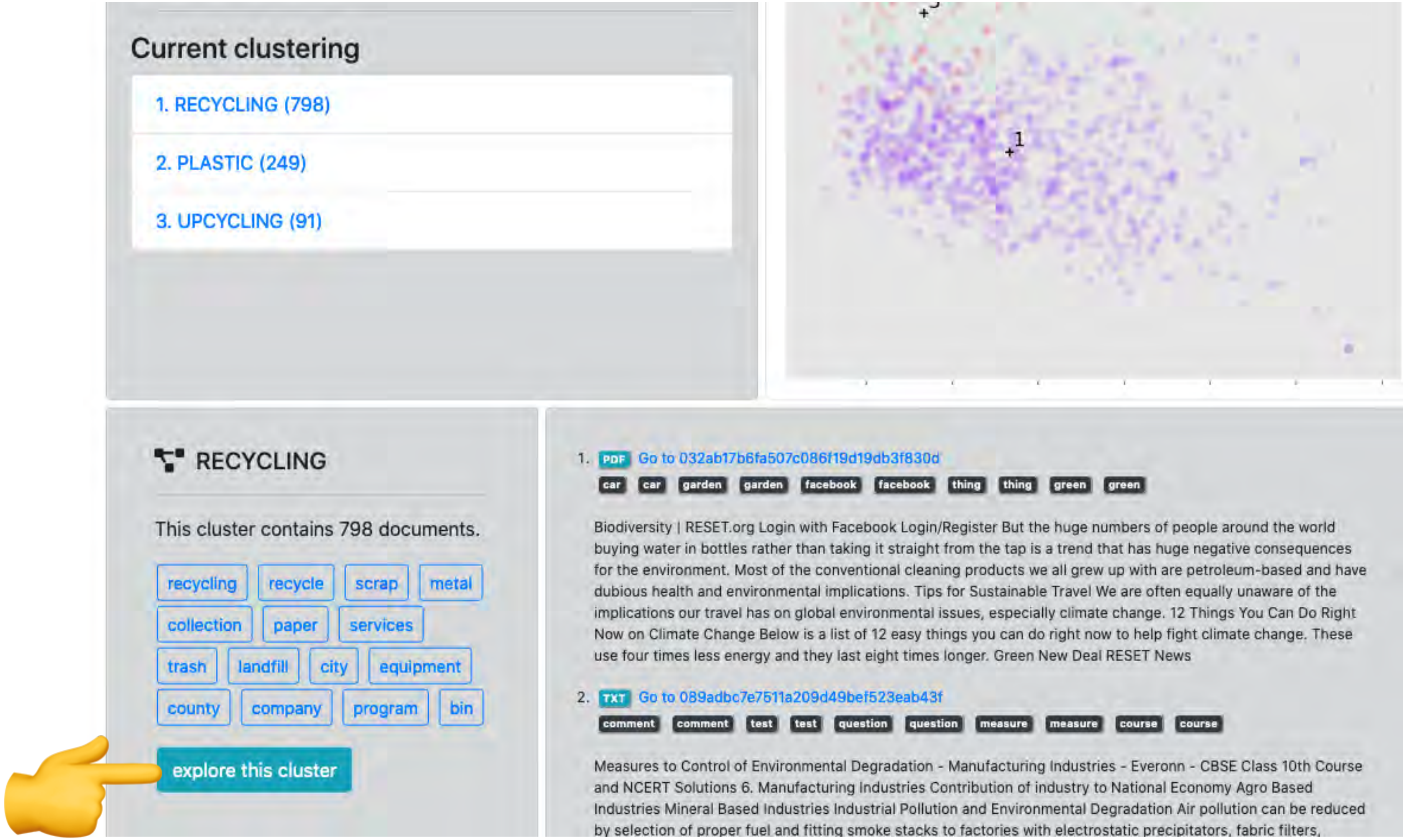
1. PDF Go to 072ee09815a87ffb1297652b4eca5be7

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

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.

[redacted]



Current clustering

1. RECYCLING (798)

2. PLASTIC (249)

3. UPCYCLING (91)

RECYCLING

This cluster contains 798 documents.

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,

Alexander C. S. Hendorf

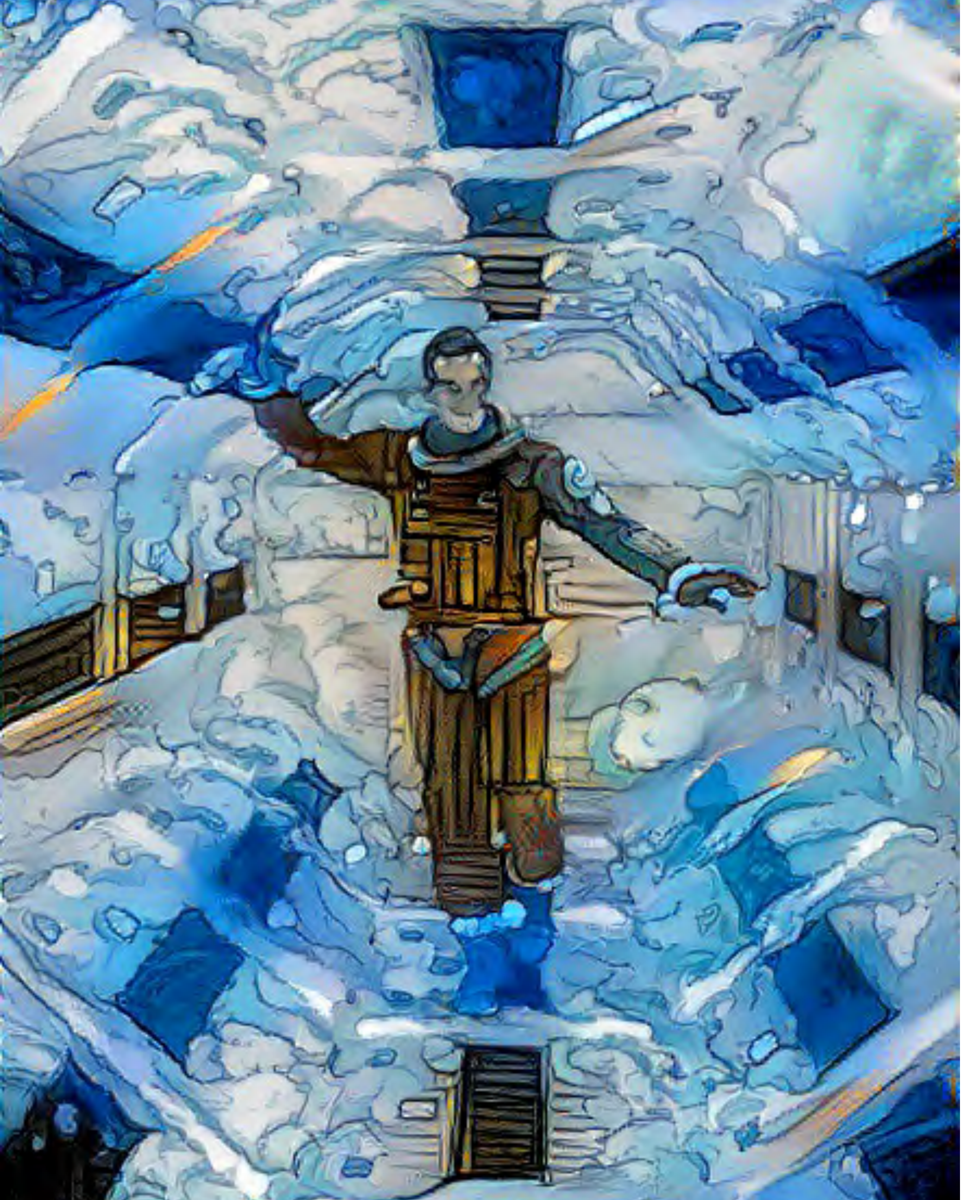
- Partner & Principal Consultant Data Science & AI
  Consulting and building AI & Data Science for enterprises.
- Python Software Foundation Fellow, Python Softwareverband chair, Emeritus EuroPython organizer, Currently Program Chair of EuroSciPy, PyConDE & PyData Berlin, PyData community organizer
- Speaker Europe & USA MongoDB World New York / San José, PyCons, CEBIT Developer World, BI Forum, IT-Tage FFM, PyData London, Berlin, PyParis,... here!

ah@koenigsweg.com

@hendorf

NLP Alchemy Toolset

1 - What is spaCy?



spaCy

- **Stable Open-source library for NLP**:
  - Supports over 55+ languages
  - Comes with many pretrained language models
  - Designed for production usage
- **Advantages**:
  - Fast and efficient
  - Relatively intuitive
  - Simple deep learning implementation
- **Out-of-the-box support for**:
  - Named entity recognition
  - Part-of-speech (POS) tagging
  - Labelled dependency parsing
  - …

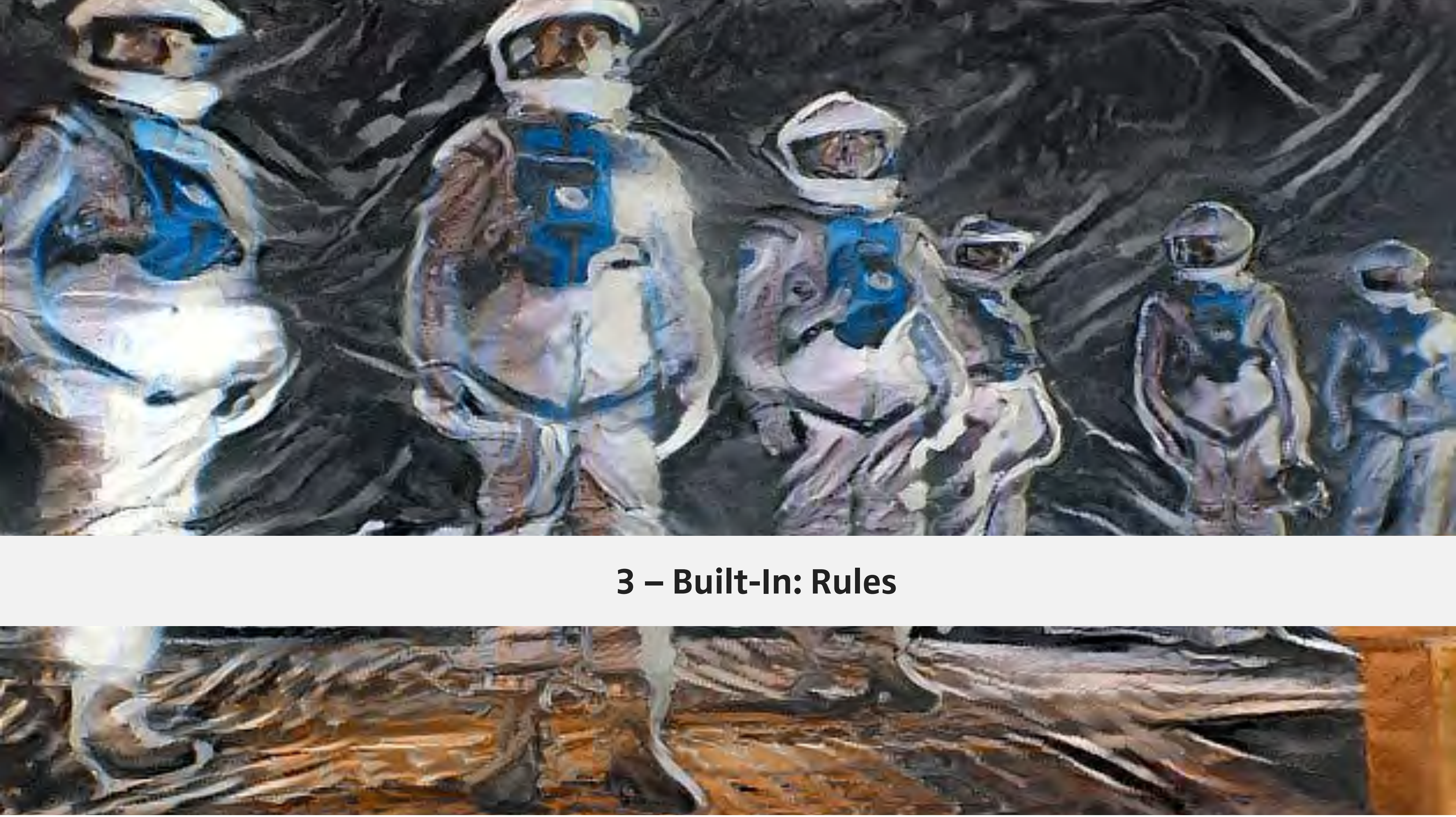
2 – Building Blocks of spaCy

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' : '?'}

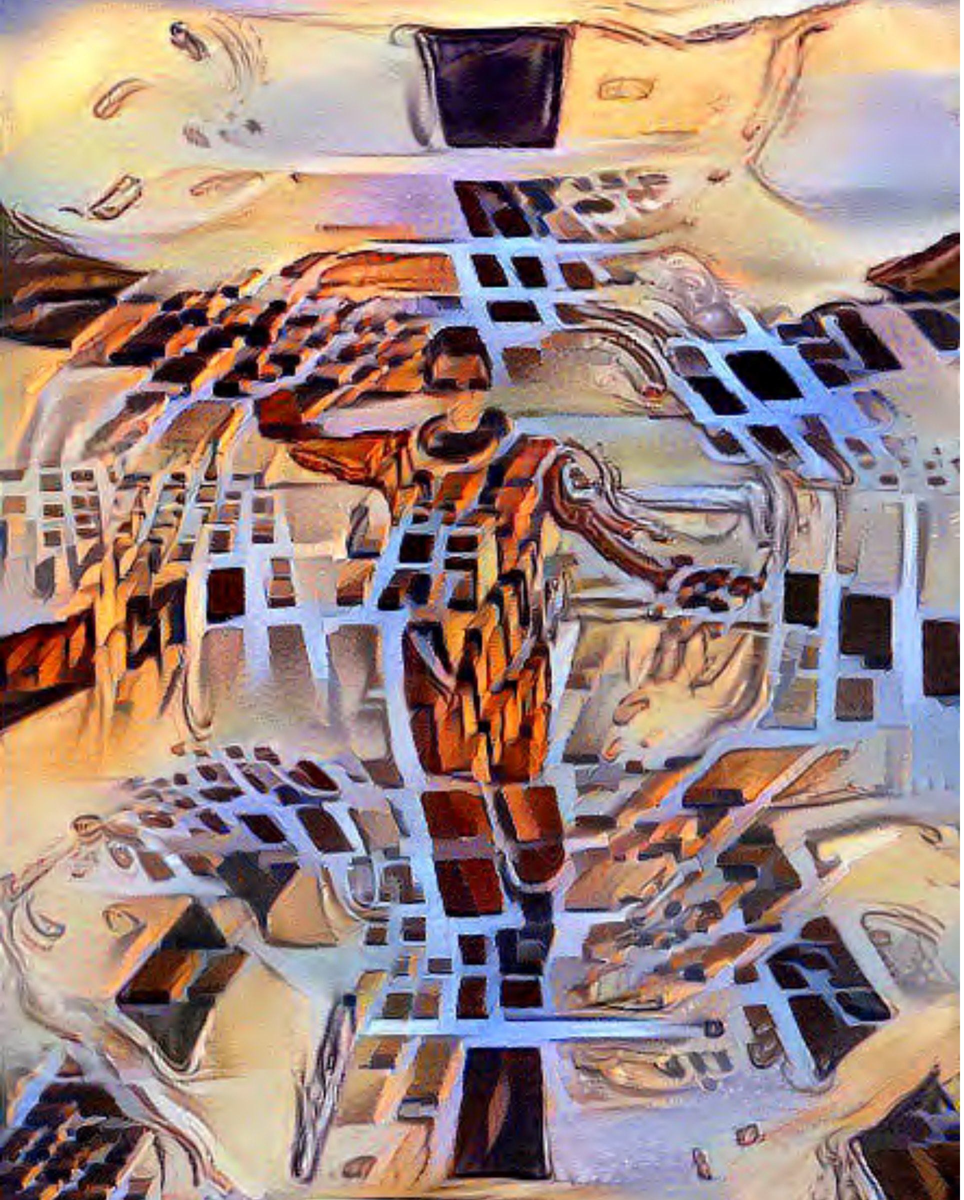


3 – Built-In: Rules

3 – Built-In: Rules

- **Language support**
  spaCy come with simple, rule based language support for many languages.
- **Language support's rules**
  Rules are mostly general, without covering too many exceptions.
- **Language models' features**
  Not all languages are supported with all features: Always refer at the documentation first.

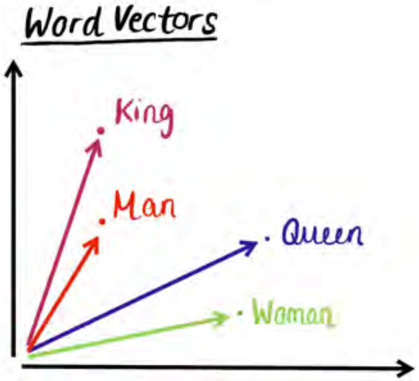
4 – Built-In: Models

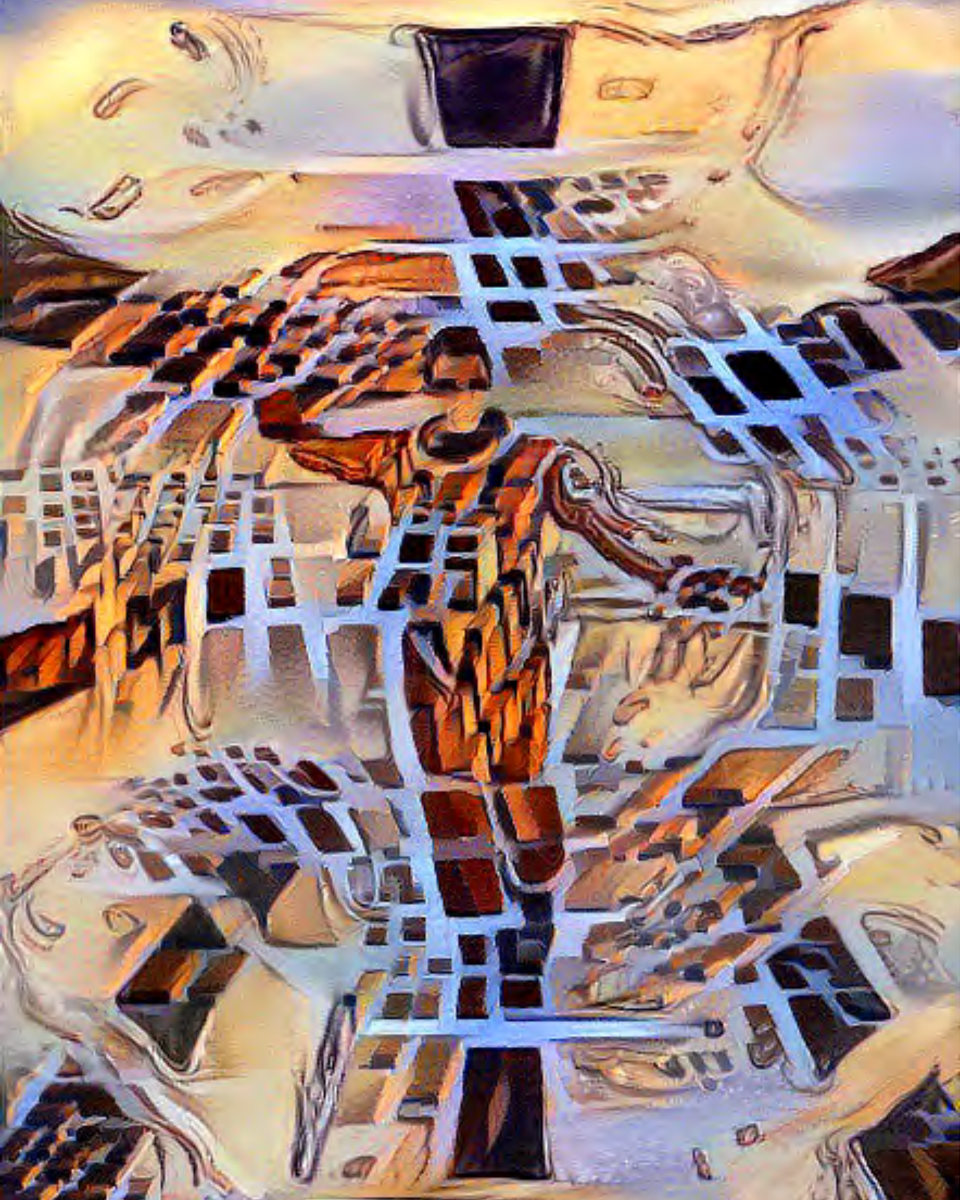


4 – Built-In: Models

- **Language models**
  Pretrained statistical models for only a few as English, German, Spanish, French, Greek, Italian, Lithuanian, Norwegian Bokmål, Dutch, Portuguese, multi-language.
  https://spacy.io/usage/models

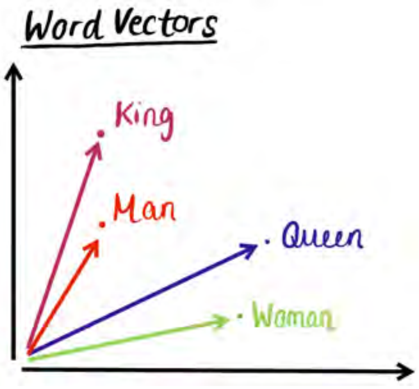
- **Language models’ features**
  Not all models support the same features: Always refer at the documentation first.
- **Training of own models**
  with nlp.update()

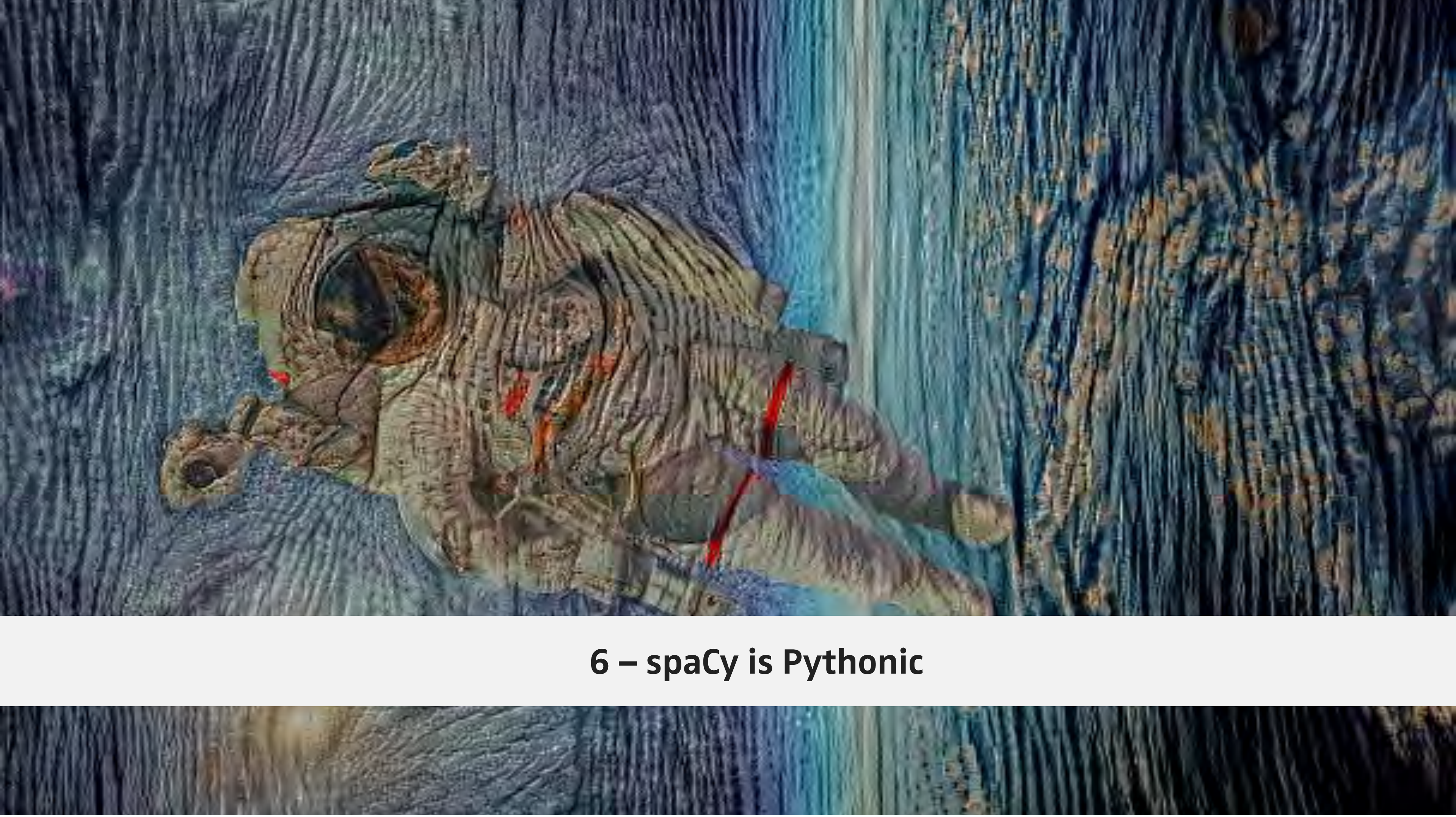




5 – Update Models

- **Training of own models supported with** with nlp.update()



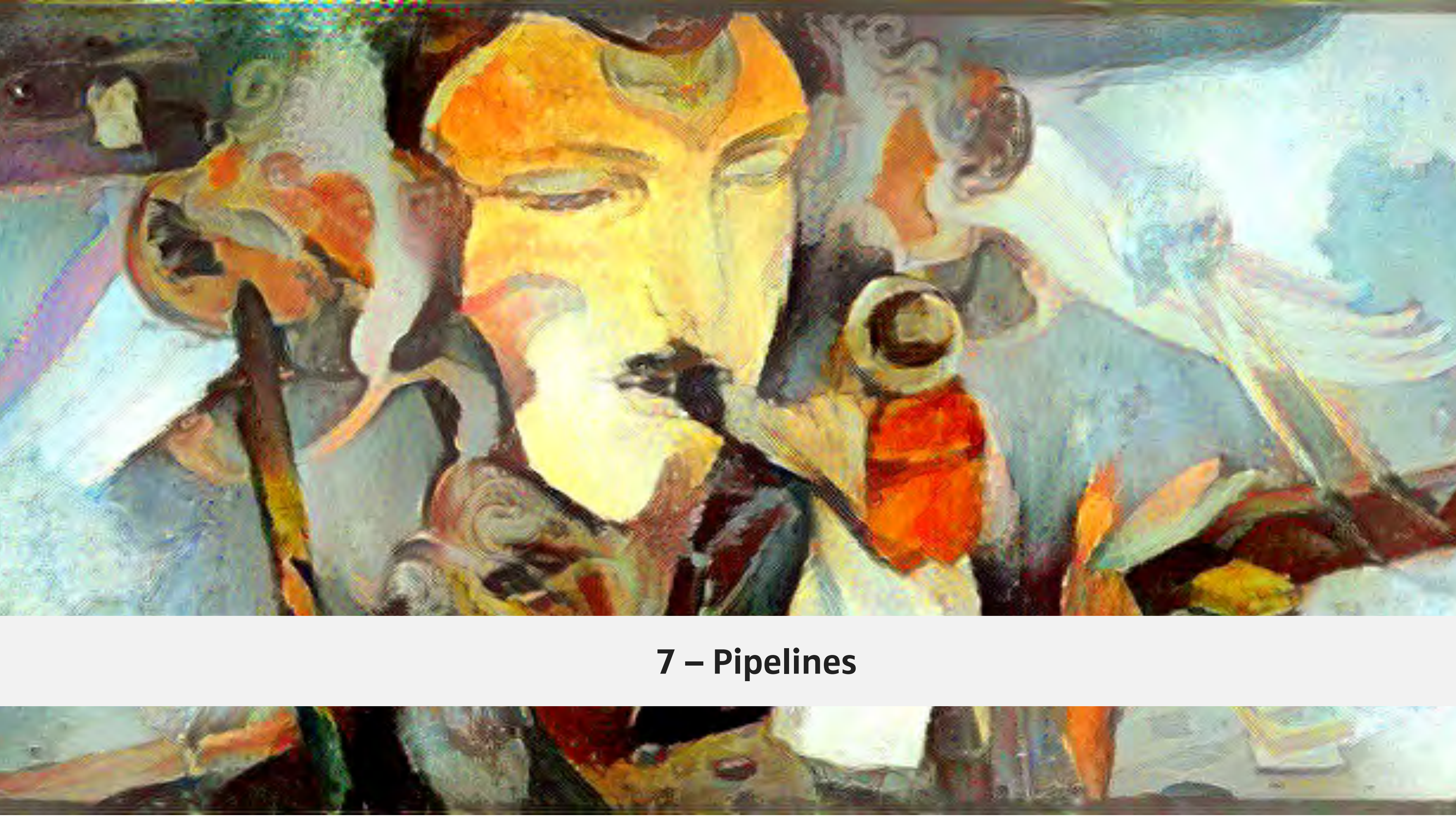


6 – spaCy is Pythonic

6 – spaCy is Pythonic

- **Extensive API**
  spaCy features an extensive API that might be overwhelming at first
  spaCy offers many access points to it’s internal data structures

- **Solid Python skills help**
  One should understand mechanics
  - objects
  - iteration/comprehensions
  - classes / methods

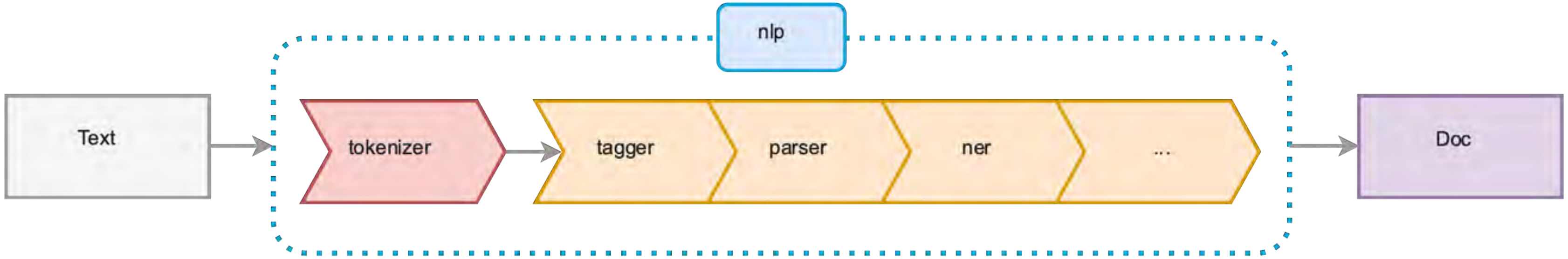


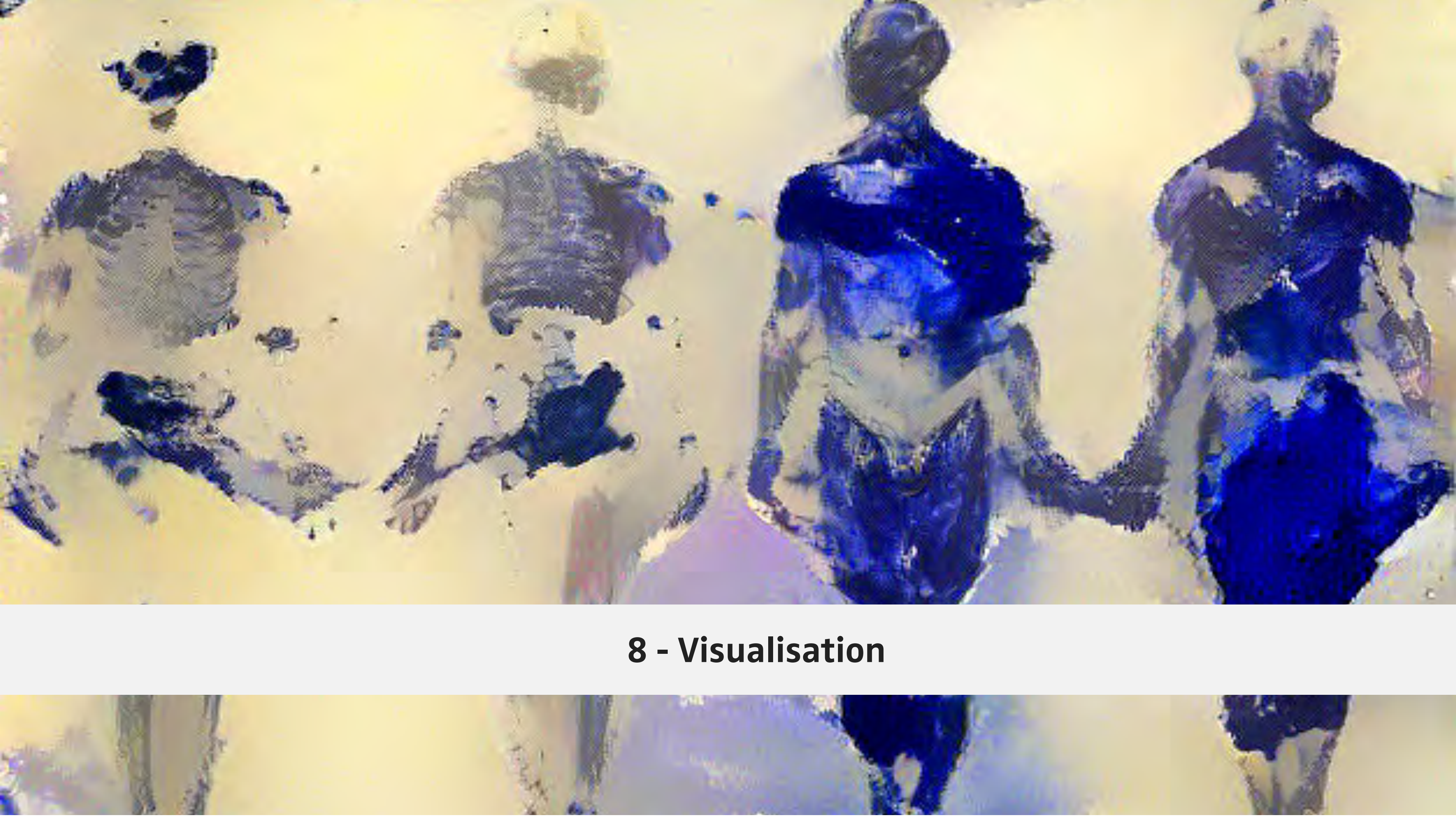
7 – Pipelines

7 - Pipelines

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

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

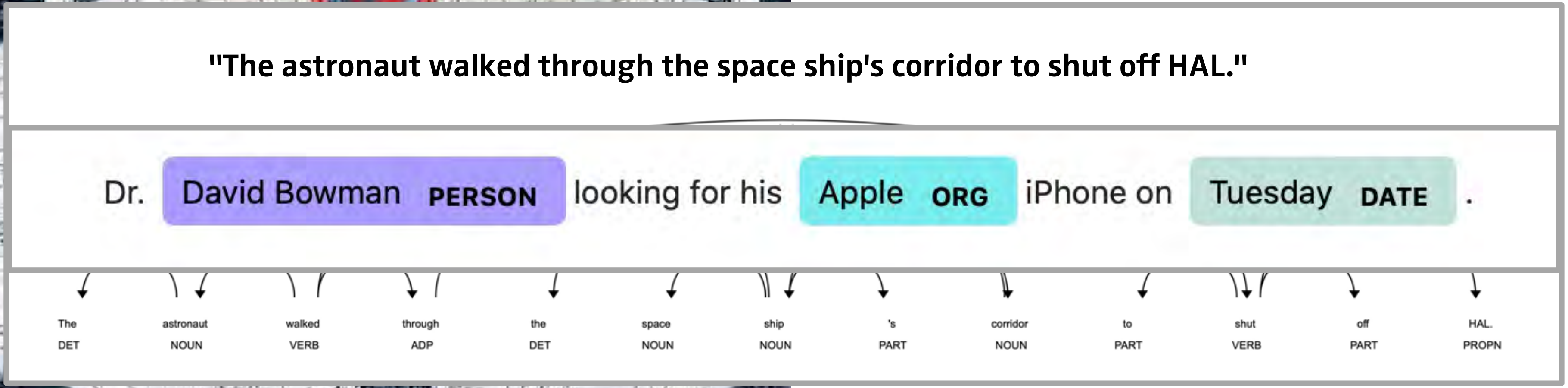




8 - Visualisation

8 – displacy

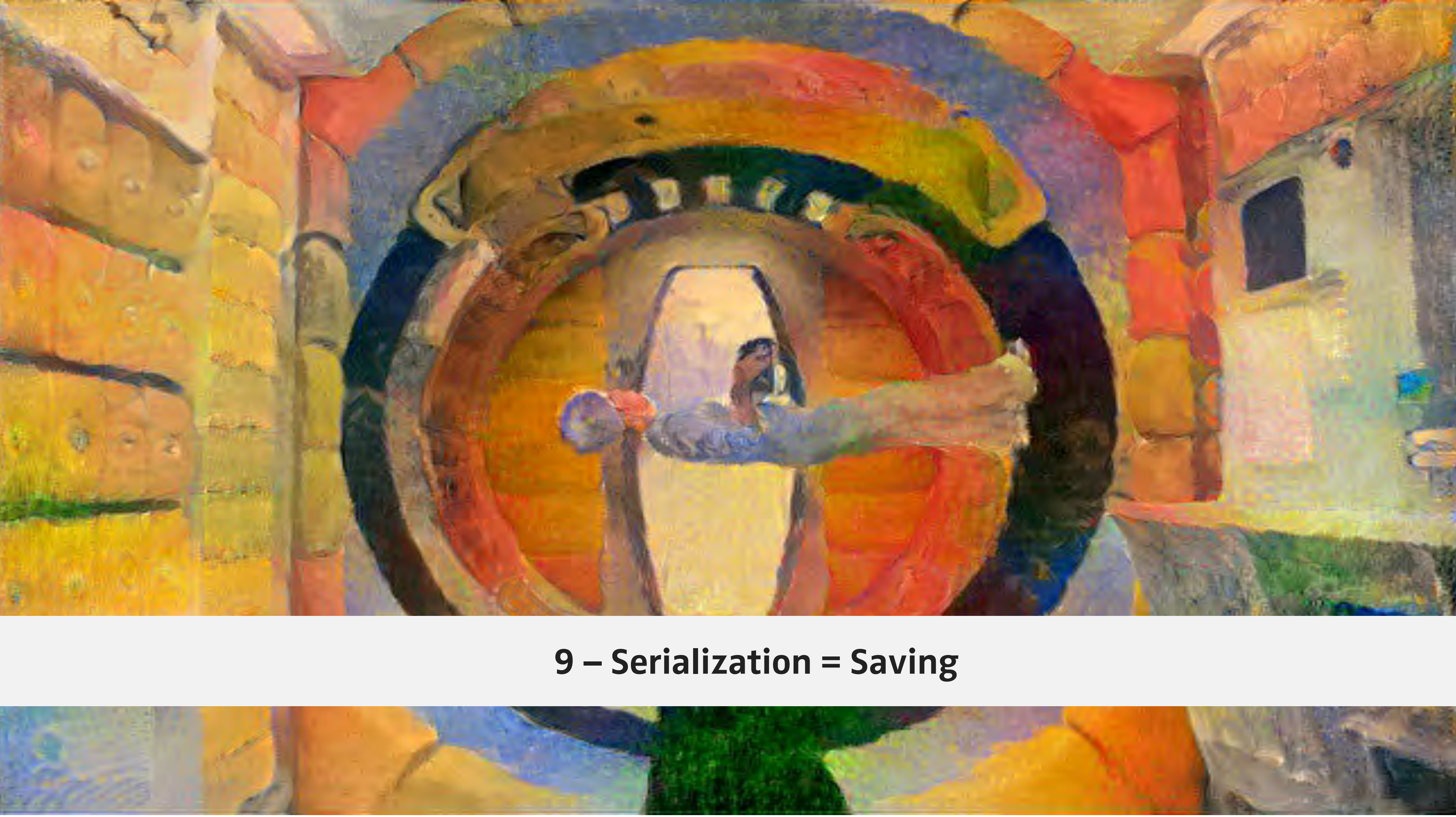
- Comes now with spaCy
- Visualize dependencies
- Visualize entities



"The astronaut walked through the space ship's corridor to shut off HAL."

Dr. David Bowman PERSON looking for his Apple ORG iPhone on Tuesday DATE .

9 – Serialization = Saving



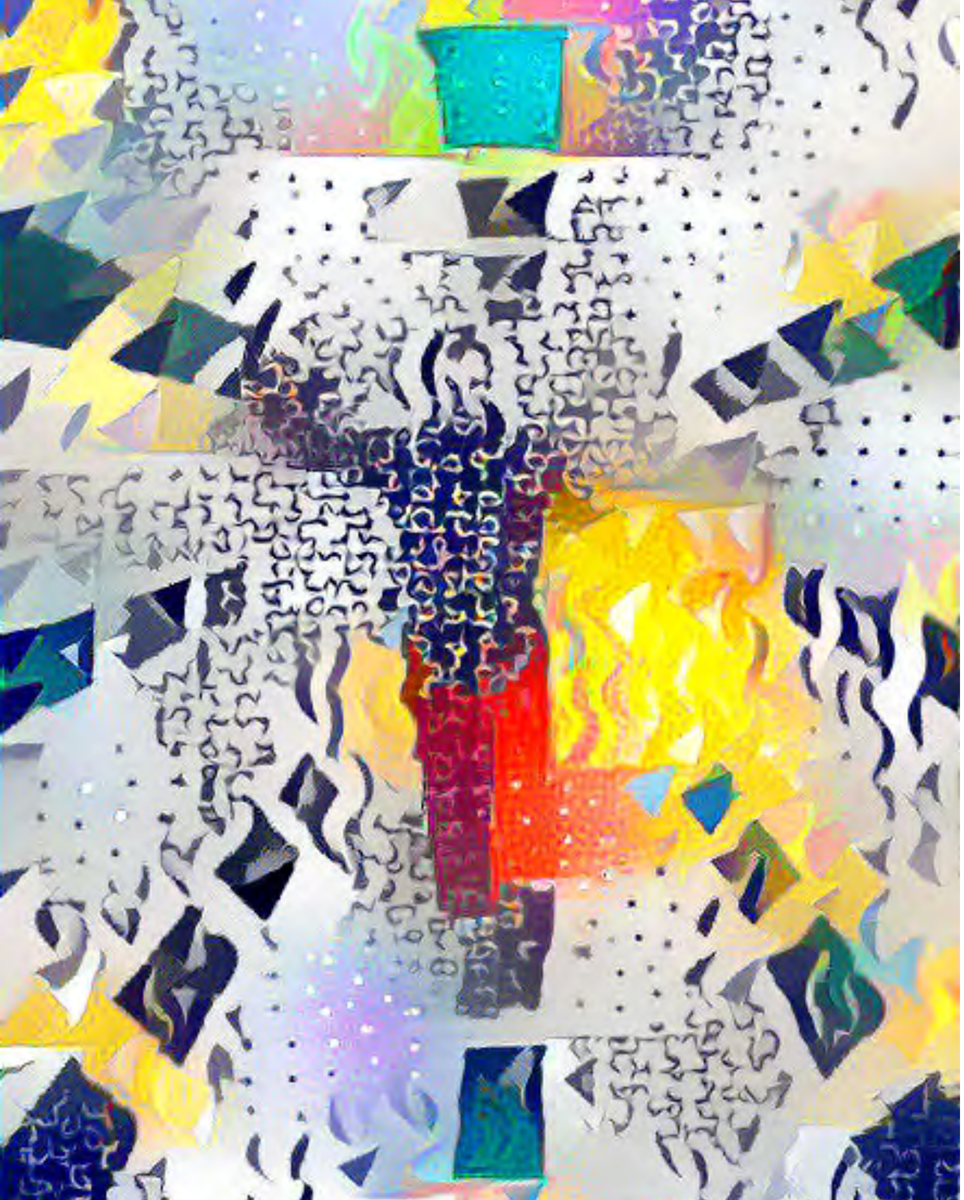
9 – Serialization = Saving

- **What to Save:**
  NLP Documents
  Vocabulary
  Model

- **Built-in**
  spaCy uses **pickle** (persisting Python objects)
  .to/from_bytes .dumps(path)/loads(path)
  We’re saving/reading bytes (not text e.g. as json) to/from disk

- **Larger projects may require a saving strategy**
  Document databases might become handy

10 – Danger Zones



10 – Danger Zones

- **Privacy**
  Not respecting people privacy and data protection

- **BIAS**
  Data is always biased, try to minimize and act accordingly to know biases

- **Legal**
  Some text must not be mined, some countries have forbidden e.g. to mine judges' rulings

- **Language**
  Language is never perfect, language is always dynamic

11 – One to Rule All Languages?

11 – One to Rule All Languages?

- **NO**

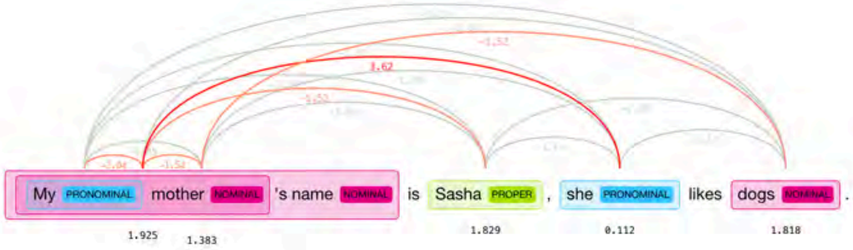
- Main research in English and Chinese
- Languages differ in complexity

12 – Extensions

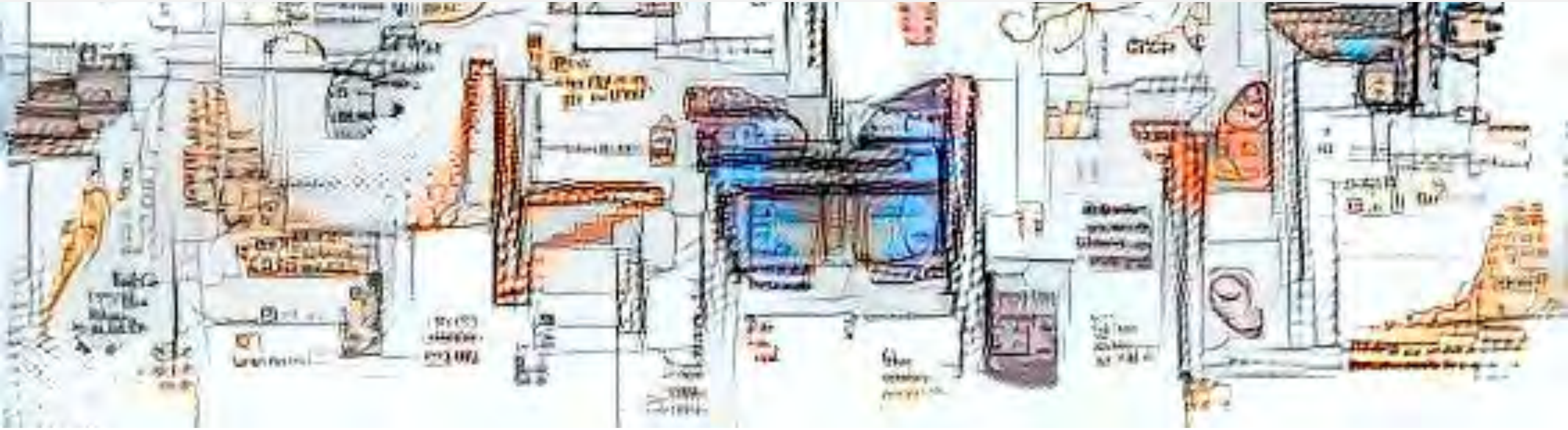
12 – Extensions

- **spaCy Universe**
  There are many projects building on, extending spaCy
  https://spacy.io/universe

- E.g. displaCy or Hugging Face’s NeuralCoref 4.0: Coreference Resolution



13 – Bugs



13 - Bugs !?!

- **spaCy**:
  - Well maintained
  - Good response time on bug reports
  - Constant, well documented updates
- **Extensions**:
  - Third party extensions may be harder to integrate
  - They may not work with all spaCy versions including spaCy updates
  - Consider different processing strategies for different tasks
  - Replicable documentation of library versions used is advised

14 – Development Status / Comparison to Others

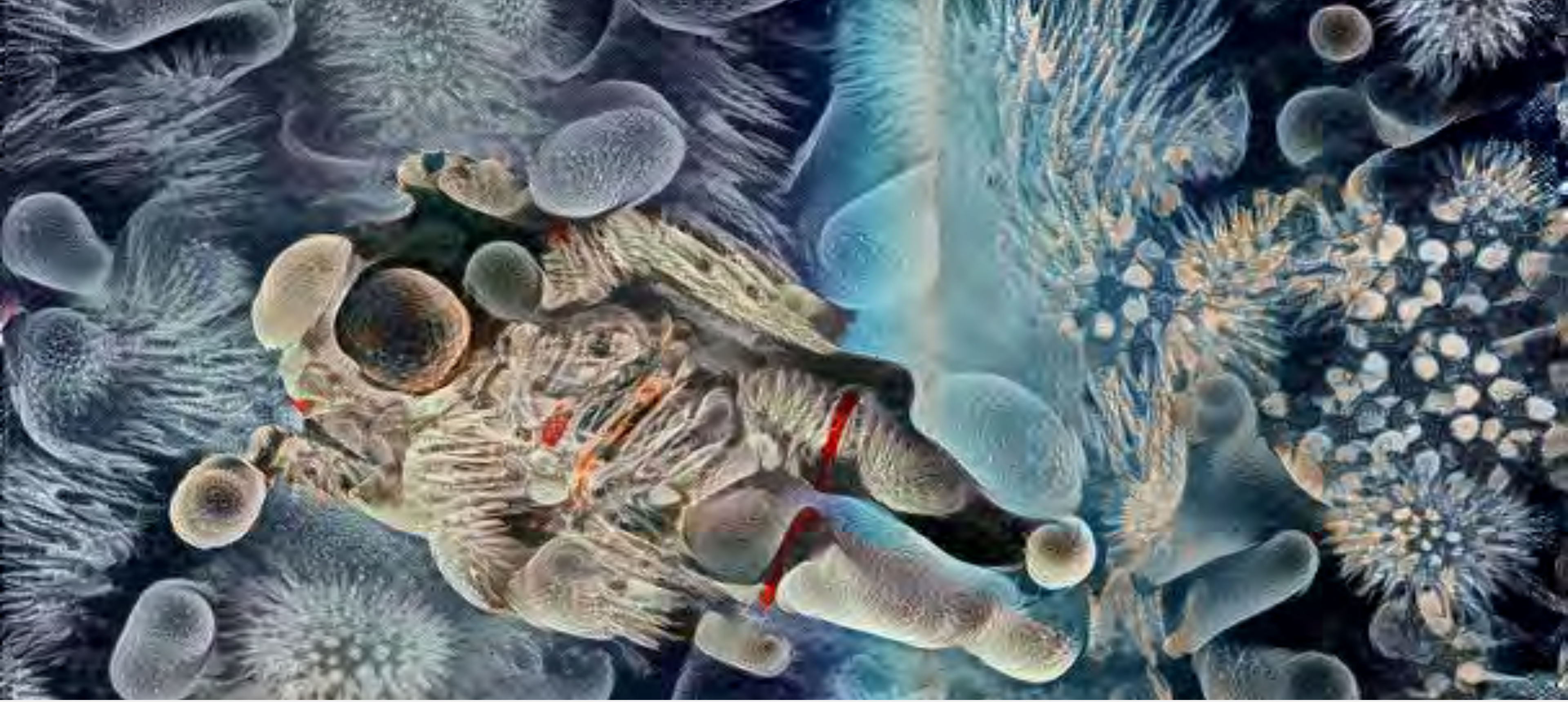
KŌNIGSWEG

14 – Status / Comparison

- **Other noticeable libraries are:**
  - NLTK
  - Gensin
  - TextBlob
  - Pattern
  - ...
- **spaCy is**:
  - Usually close to the State-of-the-Art
  - Esp. for language models, e.g. spacy-transformers
  - Flexible
  - Extendable via spacy universe
  - Fast (often powered by cython)
- For **specialized cases**, e.g. contextual assistants & chatbots check e.g. RASA

15 – Last but not Least

❤ spaCy



Thank you & 👋 stay healthy!

Thank you & 👋 stay healthy!

KŌNIGSWEG

KŌNIGSWEG

Thank you!

We're hiring

koenigsweg.com

ah@koenigsweg.com

@hendorf