The Natural Language Processing automatization at JRC for Knowledge Management
Joint Research Centre
Conclusions from the previous presentations

- HTCondor is at the core of the EU regulation for GMO authorizations.
- HTCondor is giving support to EFSA.
- HTCondor is a key player at the JRC research to fight against food fraud and species identification.
Natural Language Processing as the underlying instrument
Natural Language Processing automatization at JRC
IBM Watson It's a suite of tools and services of enterprise-ready AI applications

One of the Ai services provided is NLU
It comes with Application Program Interfaces (APIs)
Natural Language Processing automatization at JRC

Speech-to-text recognition

Diagram showing the process from video sources, video to audio extraction, speech-to-text, text-cleaning, natural language understanding, and features extraction and publish.
Speech-to-text recognition

**Case 1:** Inside the JRC
Vladimir Šucha

**Case 2:** Inside EC 2019 (spitzencadidate)

*In collaboration with JRC.I3 and JRC.A3 units*

**Case 3:** Inside of Science Human Gene Editing 2nd conference
Natural Language Processing automatization at JRC

**Sources**
- Scientific bibliography (text, pdfs, etc)
- Audio Files (mp3, wav, webm, etc)
- Video Files (mp4, mpeg, etc)

**Results**
- JSON files
- CSV files

**Data Processing**
- FFmpeg
- IBM Watson Natural Language Understanding
- IBM Watson speech-to-text API
- Python
- Jupyter
What next?

**Machine reading comprehension**

Machine Reading Comprehension models enable computers to read a document and answer general questions against it. While this is a relatively elementary task for a human, it's not that straightforward for AI models.

We have chosen AllenNLP from AI2, due to to portability and reproducibility. Runs on GPUs/TPUs.
Machine reading comprehension and Artificial assistants

<table>
<thead>
<tr>
<th>TECHNOLOGY</th>
<th>How we use/apply</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IBM Watson</strong></td>
<td>APIs</td>
</tr>
<tr>
<td><strong>AllenNLP</strong></td>
<td>Local Installation for Machine Reading comprehension</td>
</tr>
<tr>
<td><strong>Pytorch and TensorFlow:</strong></td>
<td>Running/training AI models</td>
</tr>
<tr>
<td><strong>Libraries</strong></td>
<td>Extraction of Audio, chunking and editing; Summarization; entity extraction from Biomedicine bibliography</td>
</tr>
<tr>
<td>like Ffmpeg, Gensim, spaCy (scispacy)</td>
<td></td>
</tr>
<tr>
<td><strong>Scripting languages</strong></td>
<td>Programming and automating</td>
</tr>
<tr>
<td><strong>Python Bash, etc.</strong></td>
<td></td>
</tr>
</tbody>
</table>
The way forward: question answering for scientific literature