

# Semantic PDF Segmentation for Legacy Documents in Technical Documentation

**Jan Oevermann**

jan.oevermann@dfki.de



German  
Research Center  
for Artificial  
Intelligence

## Most common: **PDF** documents

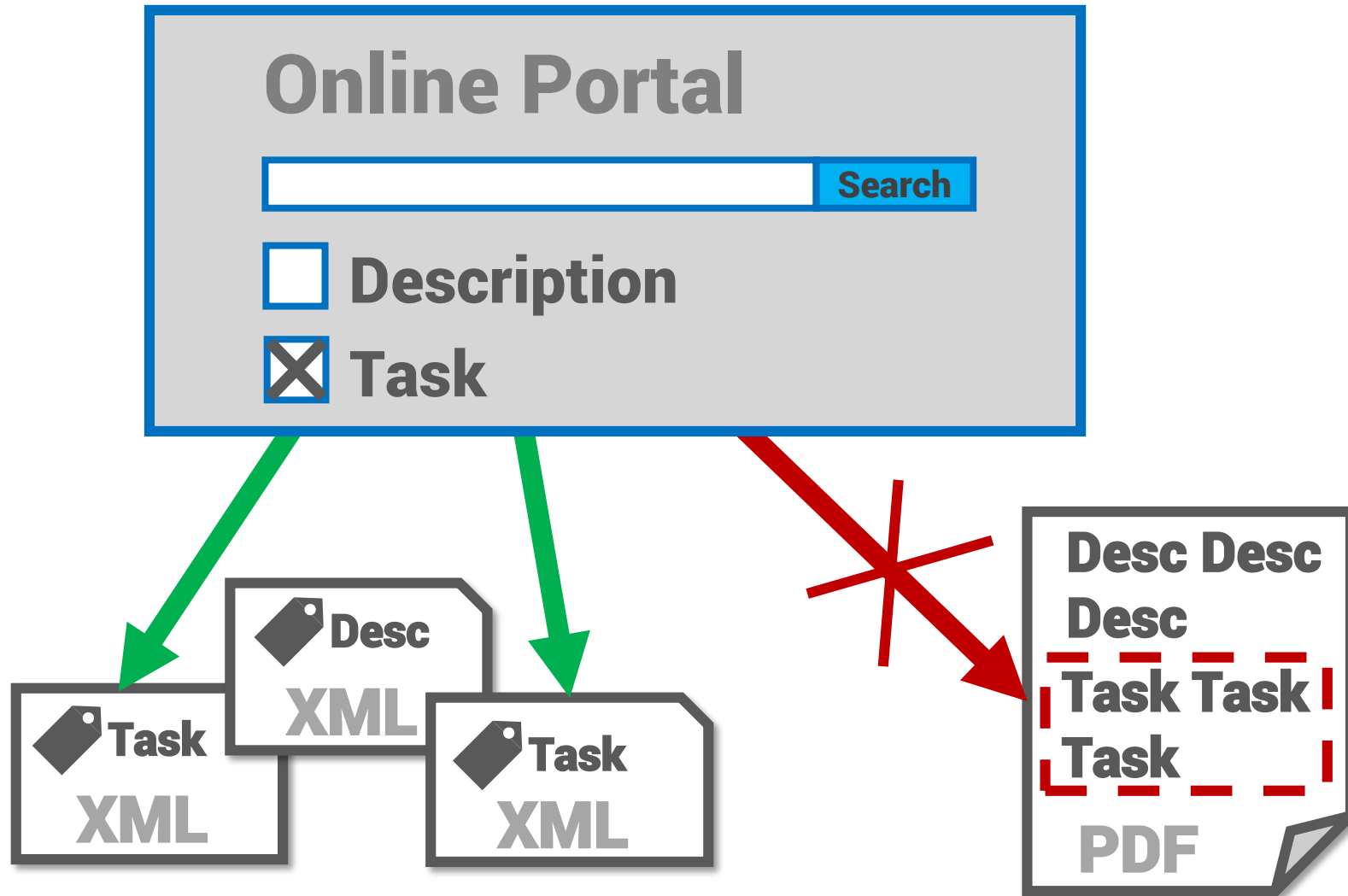
- “Digital Paper”, archival & distribution
- ISO Standard, guaranteed reproduction, ubiquitous support



## Best practice: **XML** content components

- Self-contained building blocks, e.g. chapter-sized, ~150-500 words
- Reuse, translation, aggregation, delivery





refine by search term

refine by type  
 Journal Articles (only)  
 Conference and Workshop Papers  
select all | deselect all

refine by coauthor  
Wolfgang Ziegler (2)  
Sebastian R. Bader (1)

refine by venue  
DocEng (1)  
SEMANTICS (1)  
Computational Intelligence (1)



Only **safety information** of the document



I need **maintenance information**  
about the **fuel injection**



Everything about the **hydraulic pump** in  
**technical overview** or **technical data**

Faceted search

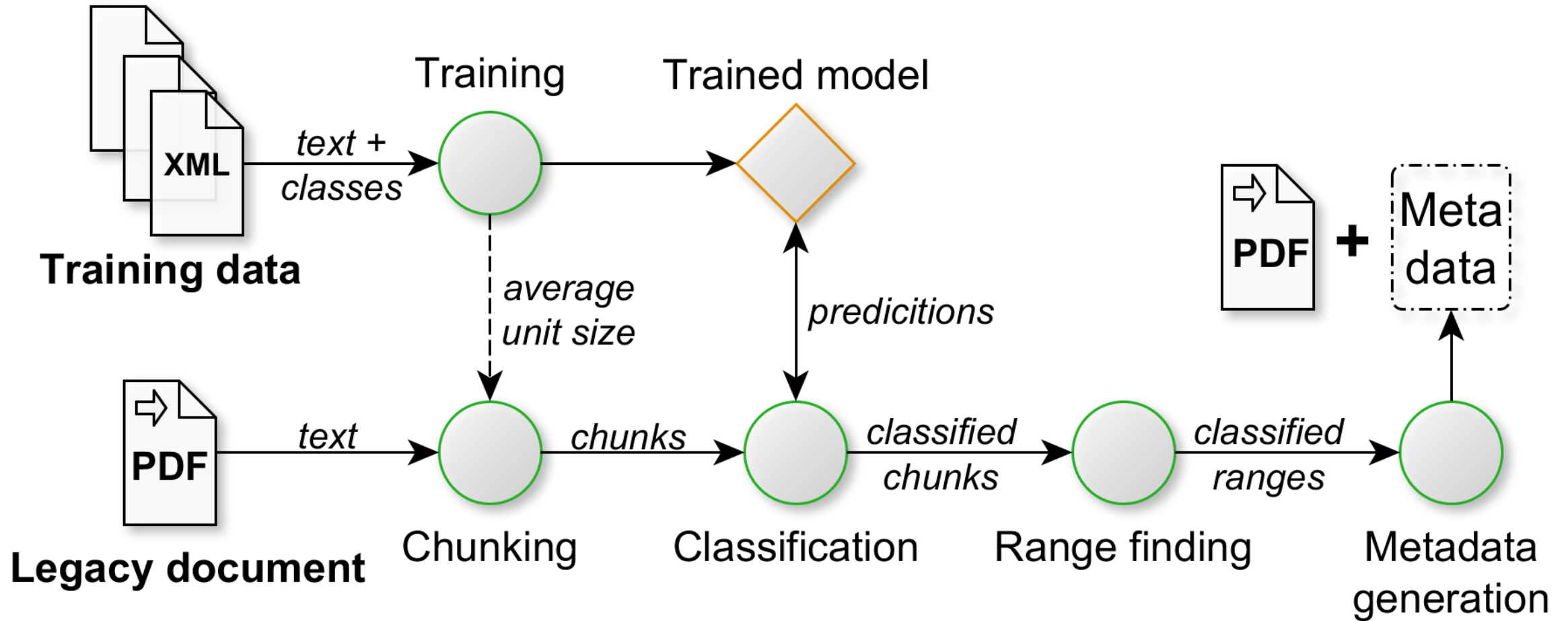
Information request with **semantic concepts**  
which can be used as facets

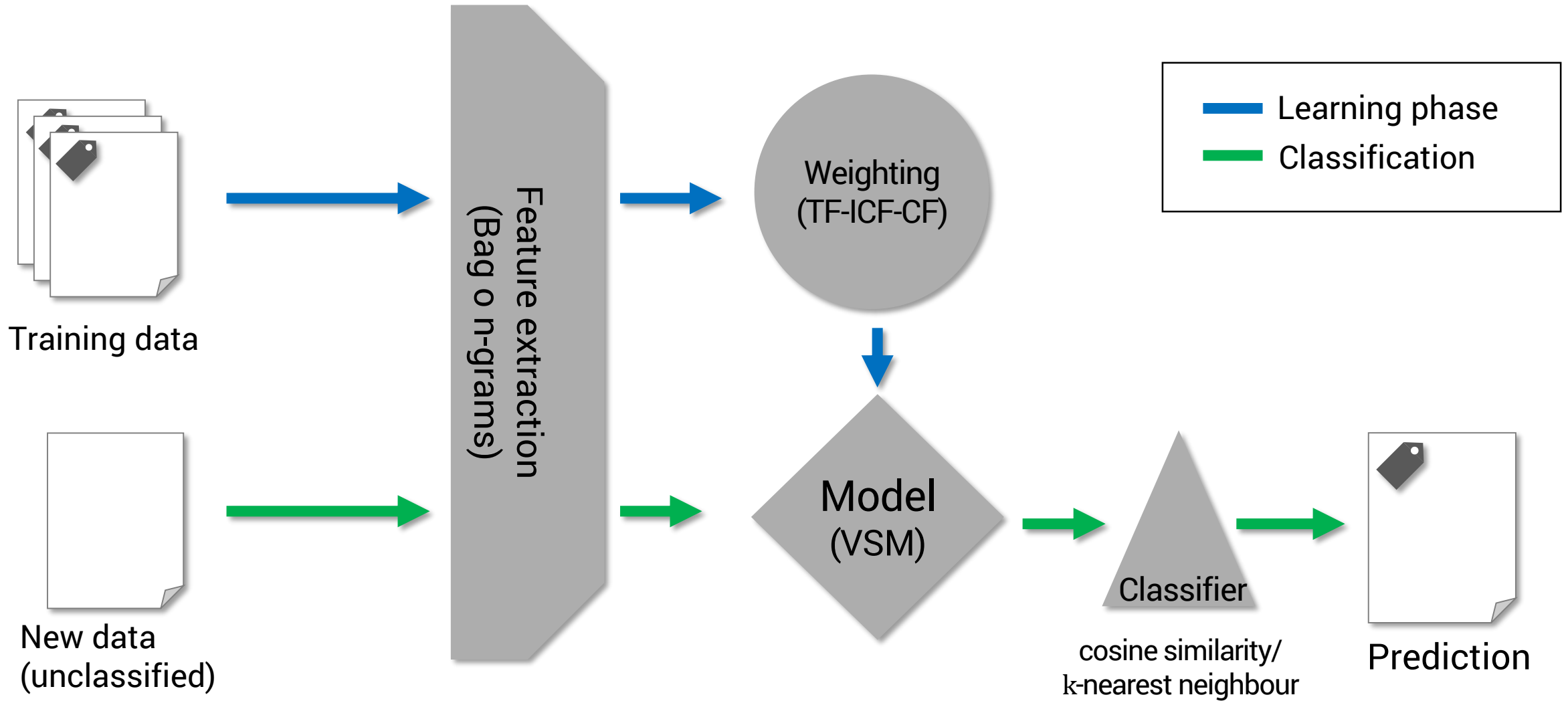
## Limitations of PDF

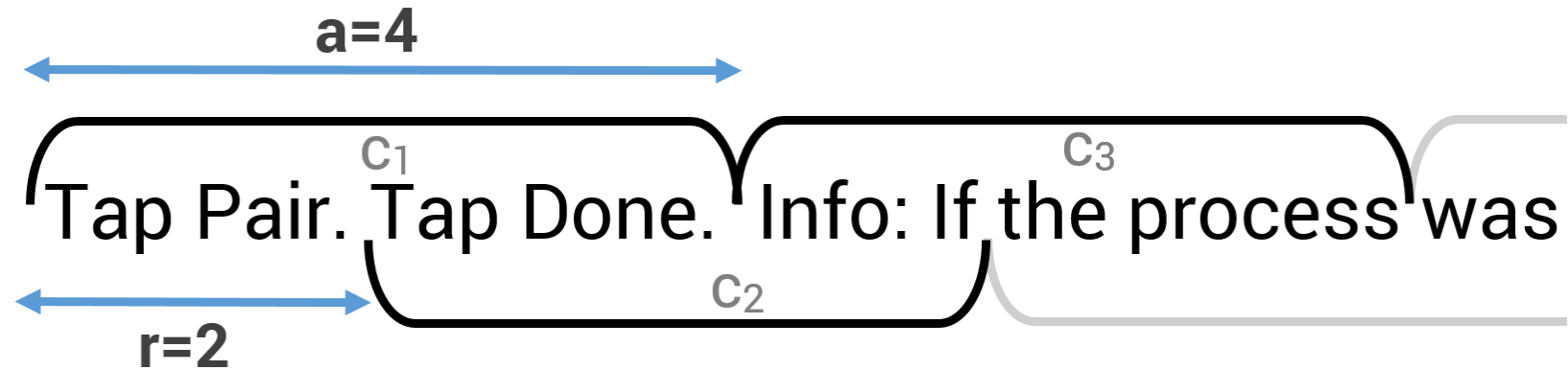
- Semantic structure gets lost
- No metadata for (overlapping) segments
- Large documents (>200p) only accessible via full text search

## Idea

- Use knowledge from structured XML content components
  - Manually annotated semantic concepts / metadata
- Apply trained model on text extracted from PDF
- Find segments which are semantically relevant







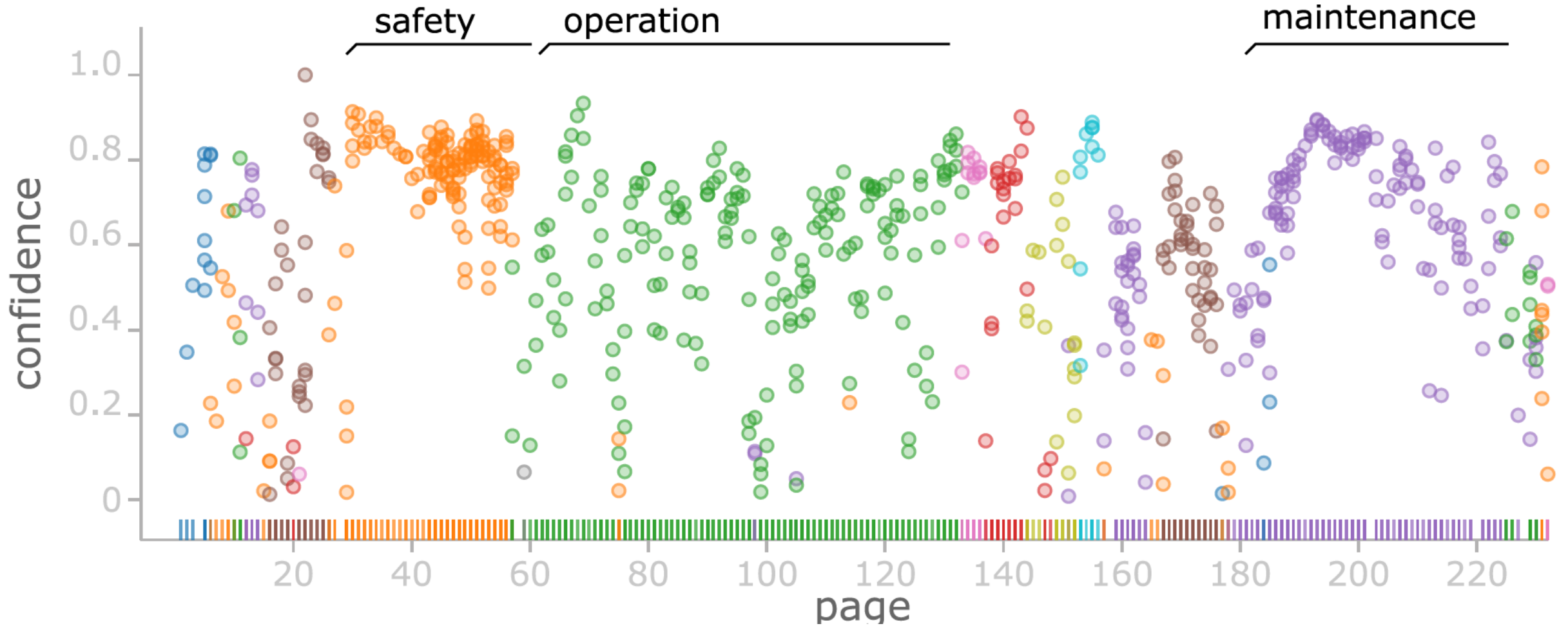
Class: Task  
Confidence: **High**

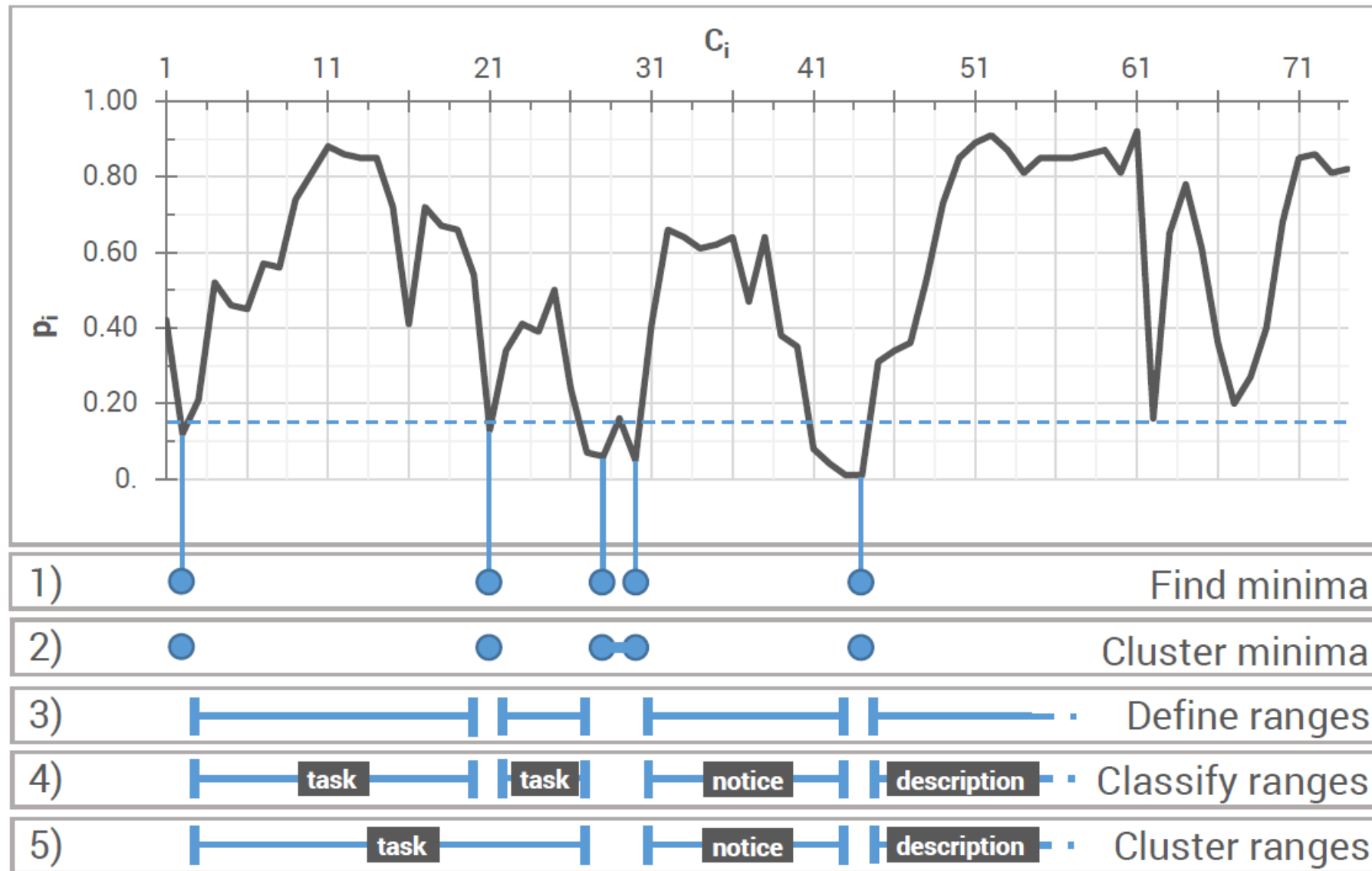
Class: Notice  
Confidence: **High**



Class: Task  
Confidence: **Low**









```
<iirds:Fragment rdf:about="urn:uuid:0b86fd8a-76b7-4cb9-ad41-2725edbf94c2">
  <iirds:has-subject
    rdf:resource="http://iirds.tekom.de/iirds#Safety"/>
  <iirds:has-rendition>
    <iirds:Rendition>
      <iirds:format>application/pdf</iirds:format>
      <iirds:source>files/manual.pdf</iirds:source>
      <iirds:has-selector>
        <iirds:RangeSelector>
          <iirds:has-start-selector>
            <iirds:FragmentSelector>
              <dcterms:conformsTo
                rdf:resource="http://tools.ietf.org/rfc/rfc3778"/>
              <rdf:value>page=15</rdf:value>
            </iirds:FragmentSelector>
          </iirds:has-start-selector>
          <iirds:has-end-selector>
            <iirds:FragmentSelector>
              <dcterms:conformsTo
                rdf:resource="http://tools.ietf.org/rfc/rfc3778"/>
              <rdf:value>page=63</rdf:value>
            </iirds:FragmentSelector>
          </iirds:has-end-selector>
        </iirds:RangeSelector>
      </iirds:has-selector>
    </iirds:Rendition>
  </iirds:has-rendition>
</iirds:Fragment>
```

# Live demo

manual (Typ: PDF)  
classified with model: SEMANTiCS Model

Save results ▾

New classification

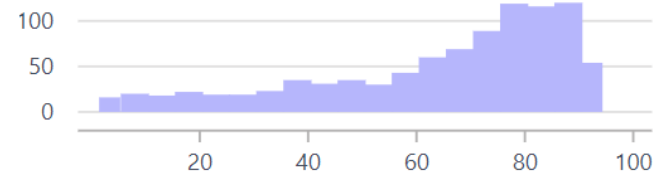
## Classes

Number of classes found

8

of 8 classes in model

## Confidence distribution



## Time

Processing time

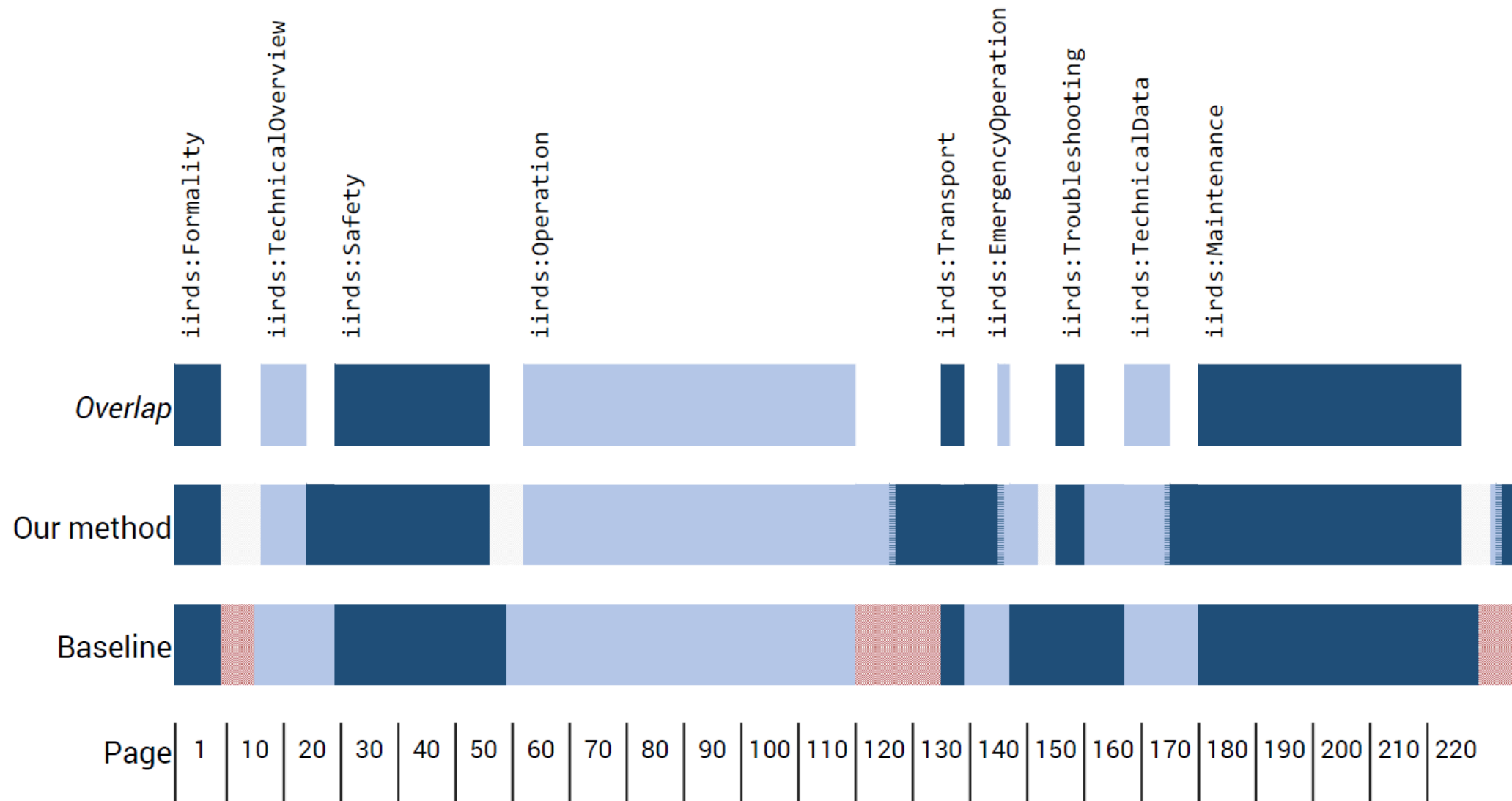
11.7s

for 938 classified objects

## Class sequence

Shows the sequence of classes over the document compared to confidence.





## Outlook

- Other text sorts (e.g. patents) or document types (e.g. Word)
- Combination with other techniques (formatting / heuristics)

## Conclusion

- Method relies on text and is formatting-independent
- No splitting of PDF, just additional metadata
- Good results in detecting semantic segments
- Identified ranges can be provided in a standardized format

# Contact

Jan Oevermann

[jan.oevermann@dfki.de](mailto:jan.oevermann@dfki.de)

[www.janoevermann.de](http://www.janoevermann.de)

Code & Demo

[github.com/j-oe/segments](https://github.com/j-oe/segments)

[segments.fastclass.de](http://segments.fastclass.de)