# The challenge

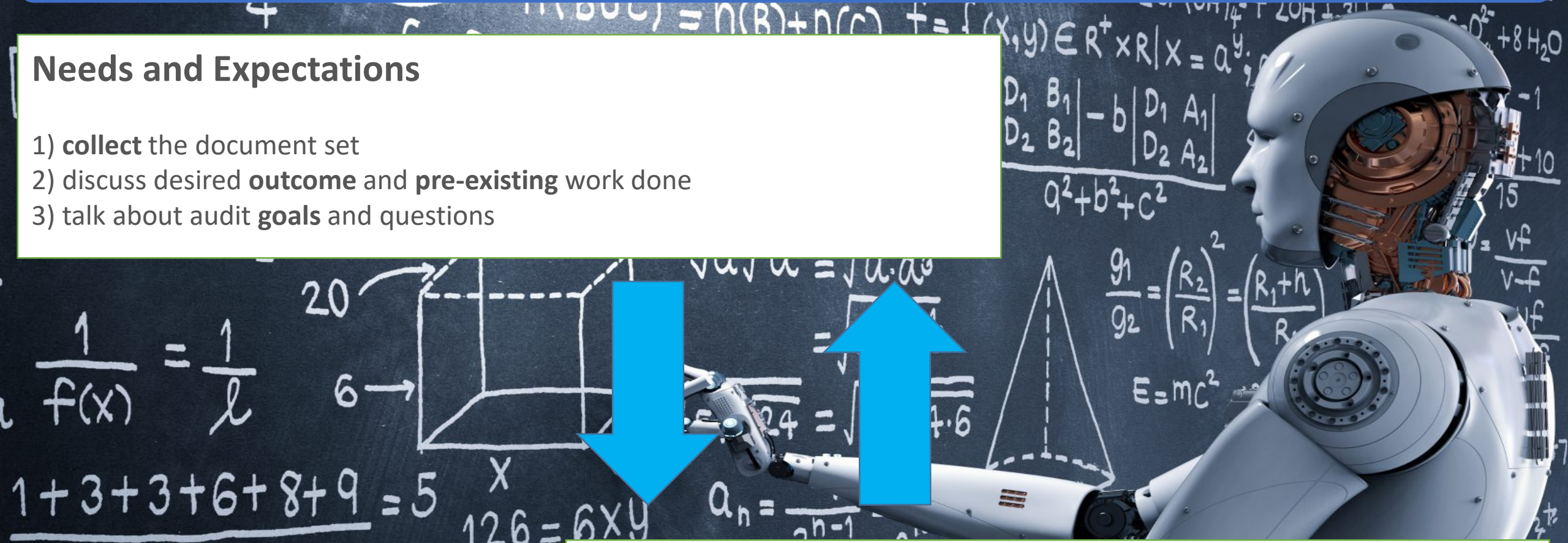
I have a huge number of documents to read and analyse for my Audit Task!

How can I go quickly through them and decide:
1) where to start
2) what may be more useful?
3) what is linked to what?
4) ...

# The process

## Needs and Expectations

1) **collect** the document set
2) discuss desired **outcome** and **pre-existing** work done
3) talk about audit **goals** and questions



## Possibilities

1) select promising **techniques**
2) apply them, try many parameter **combinations**
3) imagine a way to **show** the results to the users

# The dataset enrichment (1/2)

Document Set

Document

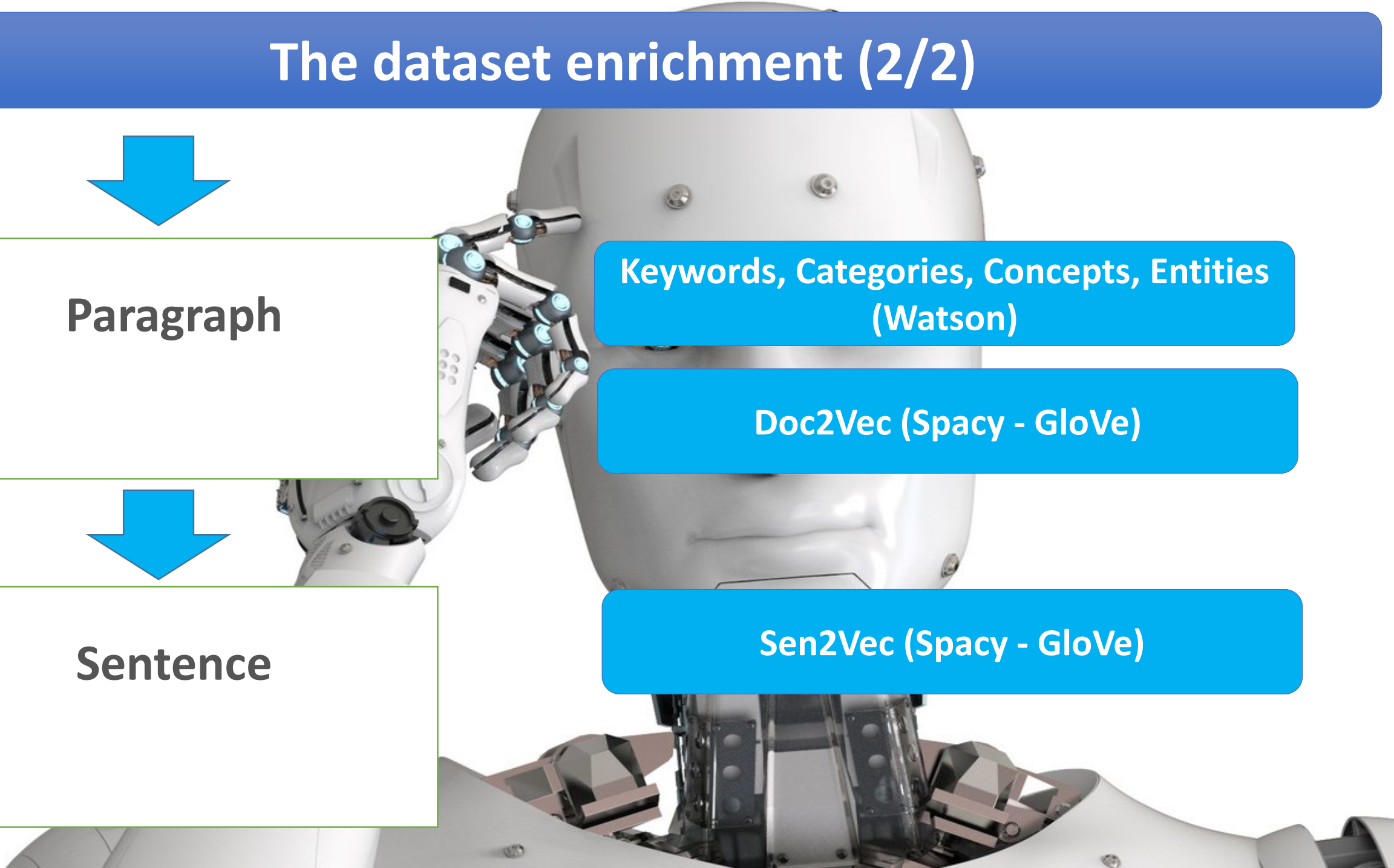
Clustering (K-means)

Topic Modeling (LDA)

Categorisation (JRC EuroVOC)

Summarisation (Gensim)

# The dataset enrichment (2/2)



**Paragraph**

- Keywords, Categories, Concepts, Entities (Watson)
- Doc2Vec (Spacy - GloVe)

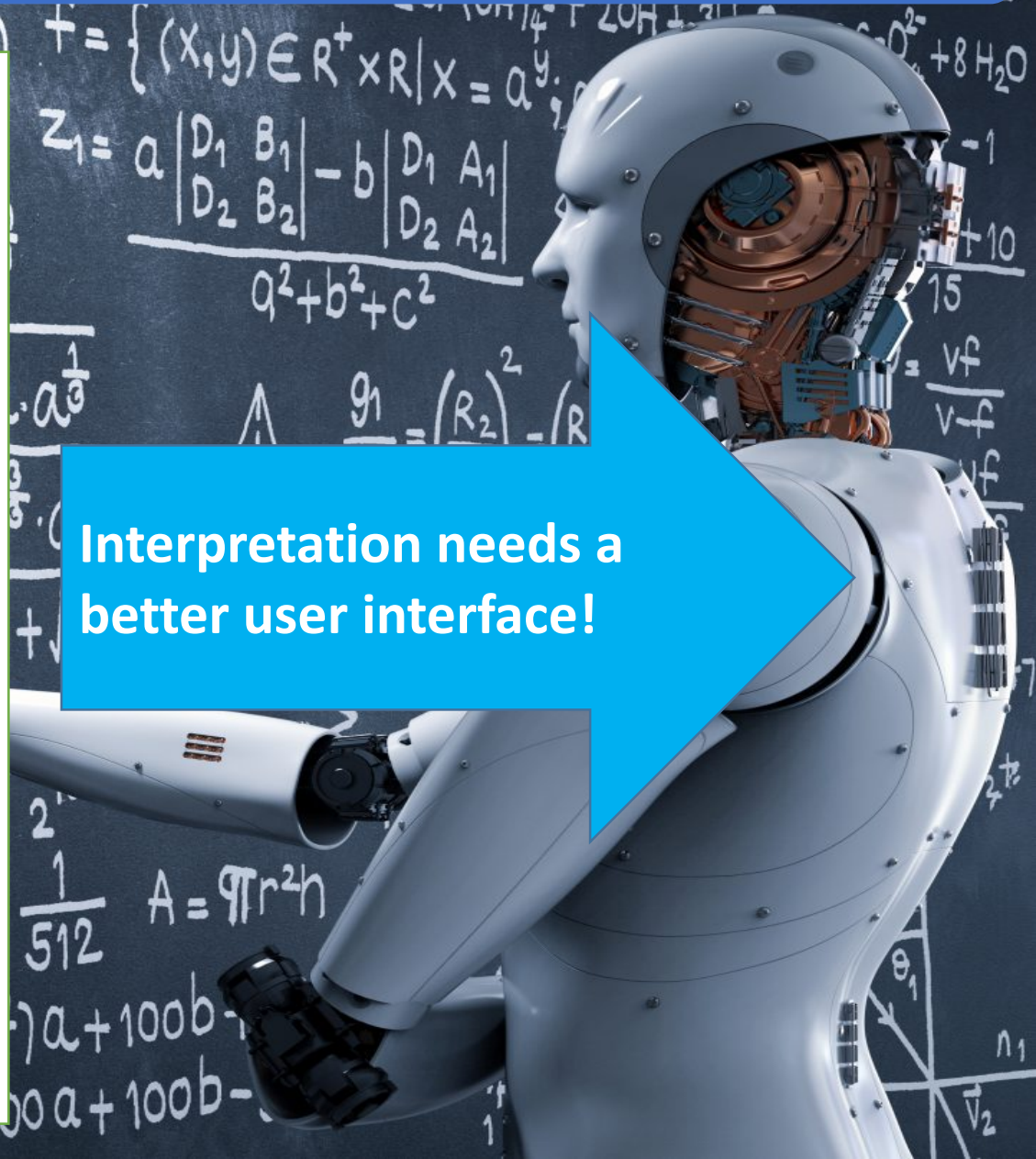
**Sentence**

- Sen2Vec (Spacy - GloVe)

# The evaluation

**Final result is way too complex!**

(ex: 15 documents, just the main category...)



**Interpretation needs a better user interface!**