1. How do digital data work? How do you digitize data?
2. Most data is unstructured.
3. Difference between structured and unstructured data.
4. Data in tables with relations: databases.
5. Data revolution.
6. Collecting data: with and without human intervention
7. Processing data: algorithms
8. Machine learning: automatic learning from data. Automate "head labor".
9. Machine learning algorithms and learning techniques.
10. Computers, machine learning and language.
11. Practical examples: TF-IDF, RECOMMENDATION ENGINES
12. Computers, machine learning and image recognition
13. Problems with AI: AI and bias

TEXT MINING & NLP

LANGUAGE PROCESSING AND UNDERSTANDING

Which algorithm can you apply as a human?



1. Looking at characters
2. Align left: LTR
3. Align right: RTL
4. Groups of characters and spaces



A machine can apply the same techniques.

But a machine must also learn to recognize the characters: OCR



TEXT MINING & NLP

TEXT ANALYSIS & UNDERSTANDING TEXT

LANGUAGE + TEXT = for a machine a priori unstructured

LANGUAGE + TEXT = does not consist of rows and columns like a database

“If all the data in the world was equivalent to the water on earth, then textual data is like the ocean, maken up the majority of the volume”

Eric Siegel

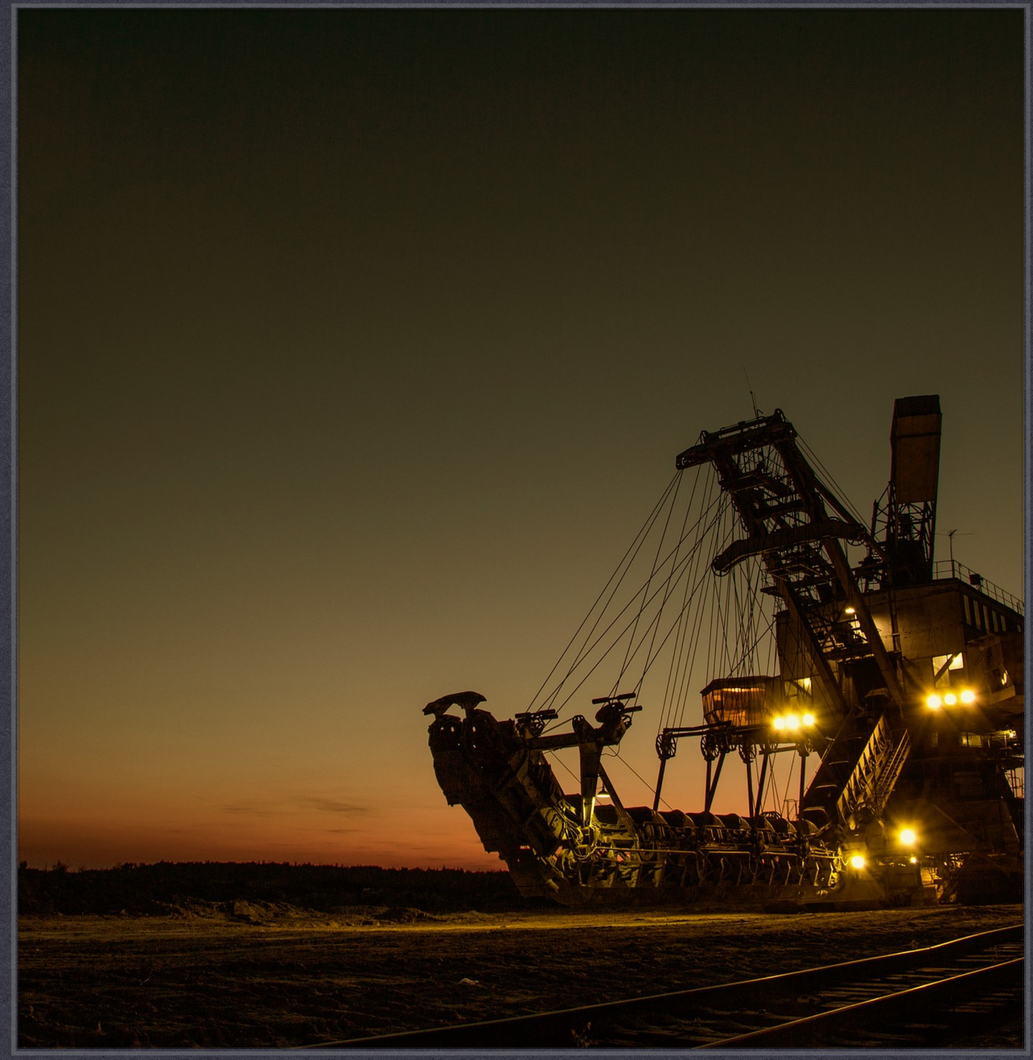
TEXT MINING
“FIND” PATTERNS IN TEXT

TEKST MINING

SEPARATING MEANINGFUL KEY WORDS FROM A LOT OF UNNECESSARY WORDS

TF-IDF

TEXT FEQUENCY - INVERSE DOCUMENT FREQUENCY



Purpose?

Classification of texts:
find out the theme / topic

Clustering:
classify texts and find connections between texts

In any case, it means that the relevant words are counted (Term Frequency) and how often does the word occur in all texts (Inverse Document Frequency).

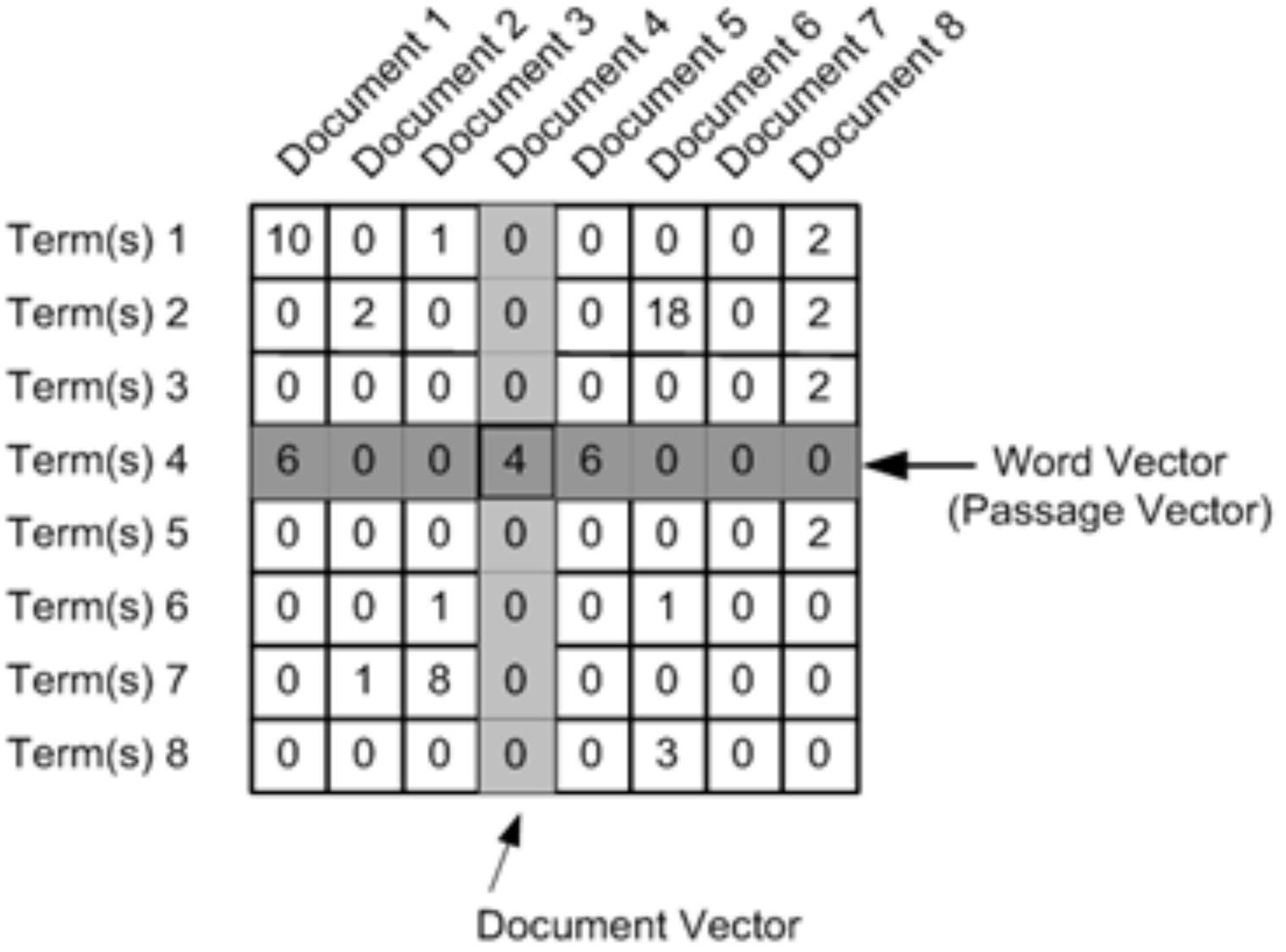
That factor is lower for common terms and higher for less common terms, filtering out words like "the", "it" and "a".

$$TF = \frac{\text{NUMBER OF TIMES A TERM APPEARS IN A DOCUMENT}}{\text{THE NUMBER OF TERMS IN THE DOCUMENT}}$$

$$IDF = \frac{\text{TOTAL NUMBER OF DOCUMENTS}}{\text{THE NUMBER OF DOCUMENTS CONTAINING A PARTICULAR TERM}}$$

$$\text{TF-IDF} = n_k/n * \log_2(N/N_k)$$

TDM
term document matrix
&
document/word vector



1. Tokenization
2. Stop word filtering
3. Filtering (vb. Leestekens)
4. Stemming
5. N-grams: Groups of “tokens / terms”

TEXT MINING
NOT NECESSARILY AI

NATURAL LANGUAGE PROCESSING & SPEECH RECOGNITION & SPEECH SYNTHESIS

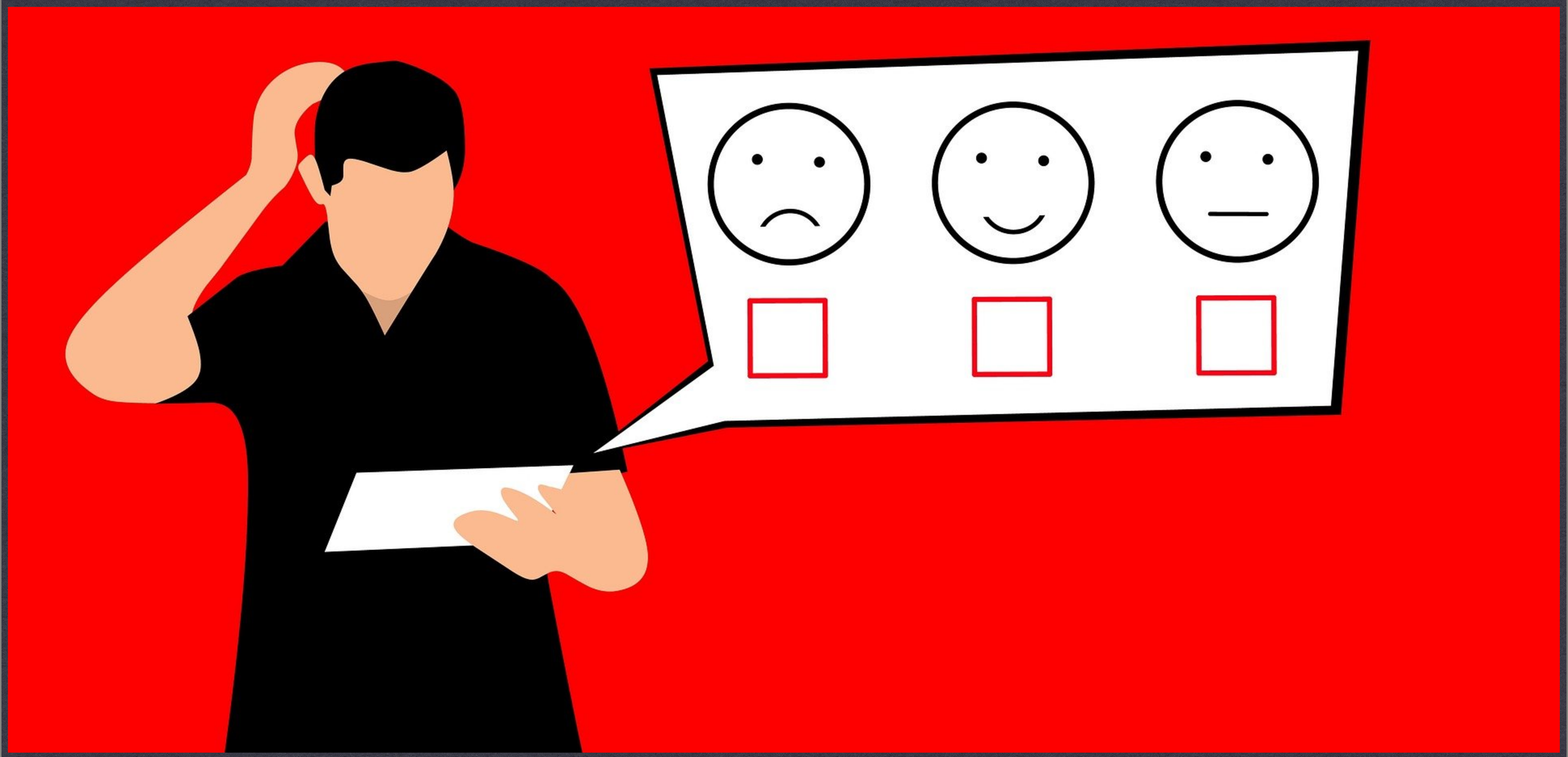
LANGUAGE IS A DIFFICULT DOMAIN



Assignment

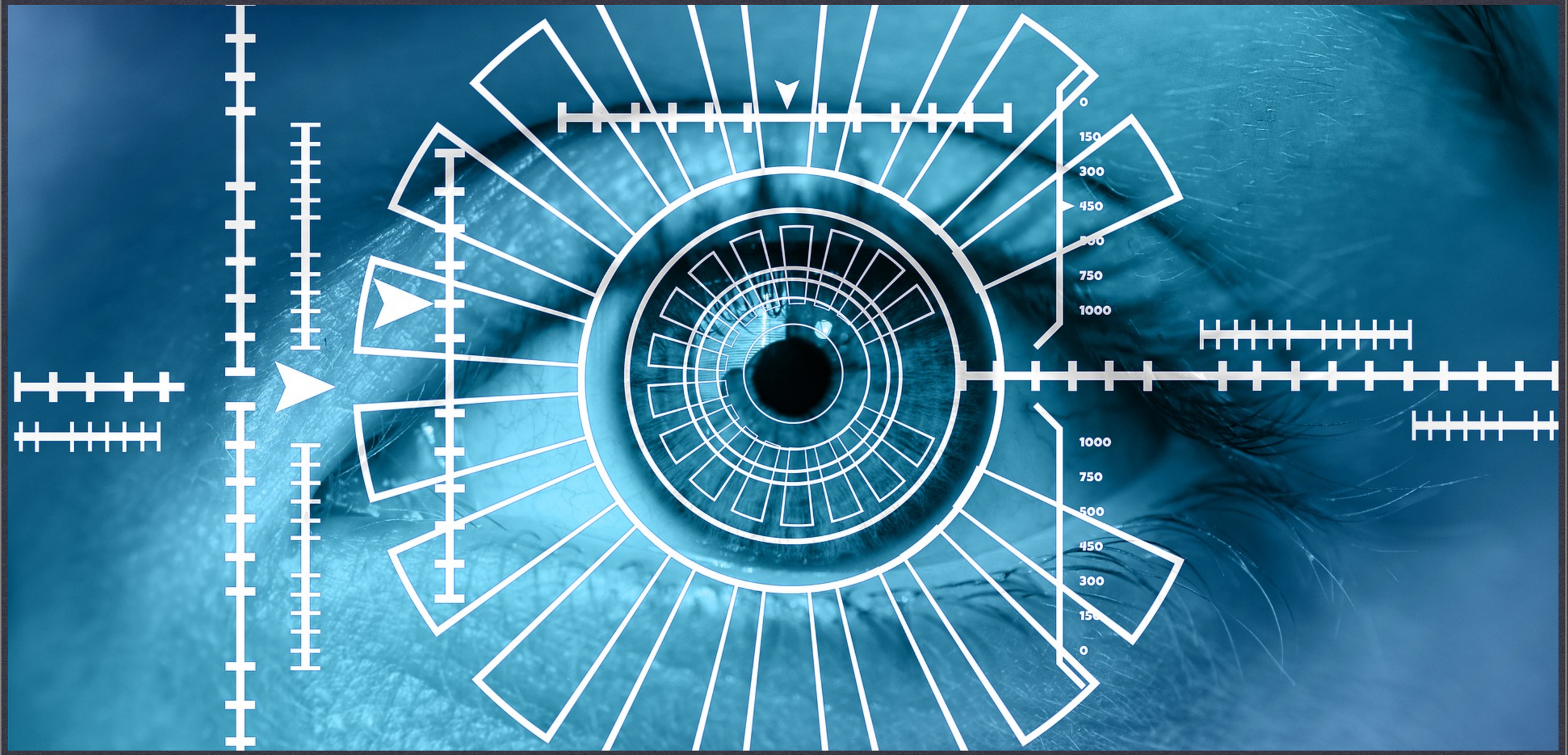
- Give at least 10 reasons why it’s difficult for machines to understand written or spoken text. Make a presentation file or Word document and explain it in a very clear way. Someone who doesn't know about the problem needs to understand what you mean.
- Robot Journalism. How could you build an algorithm to teach software to write football match reports? Work out a “step-by-step plan”.

NEXT LESSONS



1. RECOMMENDATION ENGINES & OTHER PRACTICAL EXAMPLES

RECOMMENDATIONS FOR THE CUSTOMER



2. IMAGE RECOGNITION

IMAGE RECOGNITION, FACE RECOGNITION AND RELATED



3. AI AND PREJUDICE

THE BLACK BOX PROBLEM AND WHY AI CAN BE BIASED