Data Quality Metrics

(IN COLLABORATION WITH BMW GROUP)

Developers: Vladana Djakovic, Valari Pai, Ekaterina Shmaneva

Supervisors: Dr Maka Karalashvili (ext.), Prof. Dr Matthias Schubert (int.)

CONTENT

MOTIVATION

When (or after) the car is produced, different defects occur. These defects are recorded and stored in the data source, known as the “Knowledge base”. Its purpose is to summarize similar quality defects and assign them to the prebuilt defect cluster.

OUR GOAL

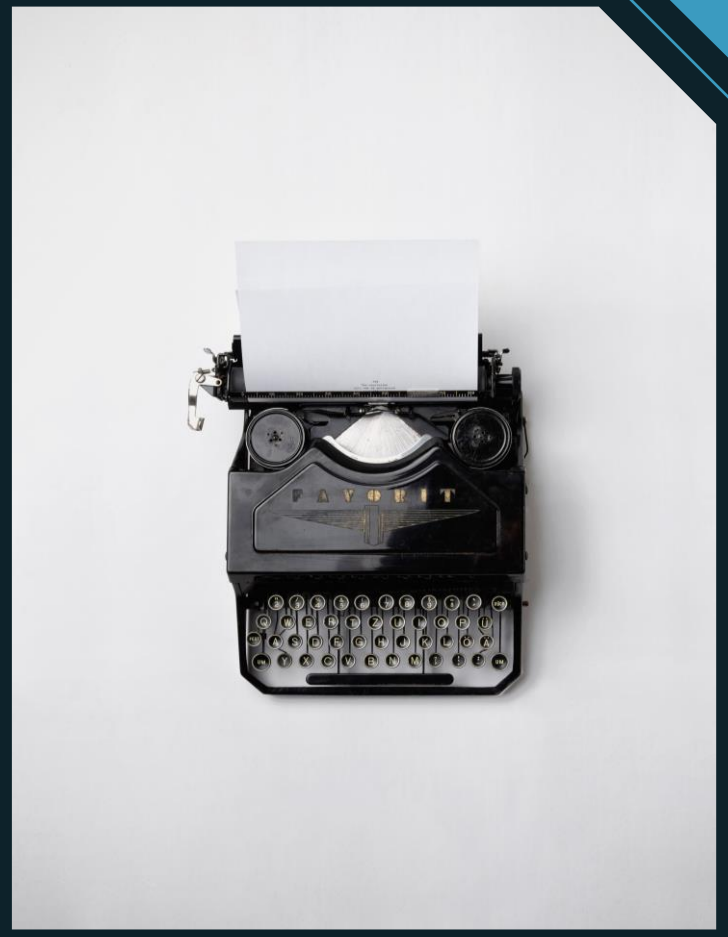
Build a model, that will preprocess the text data, create a summary of it, classify it, based on the „mood“ of the generated summary and evaluate the quality of it

Summarization

– a text that is produced from one or more texts, that contains a significant portion of the information in the original text(s).

If it was created with the computer, it is called **automatic summarization.**

Can be **abstractive** and **extractive.**

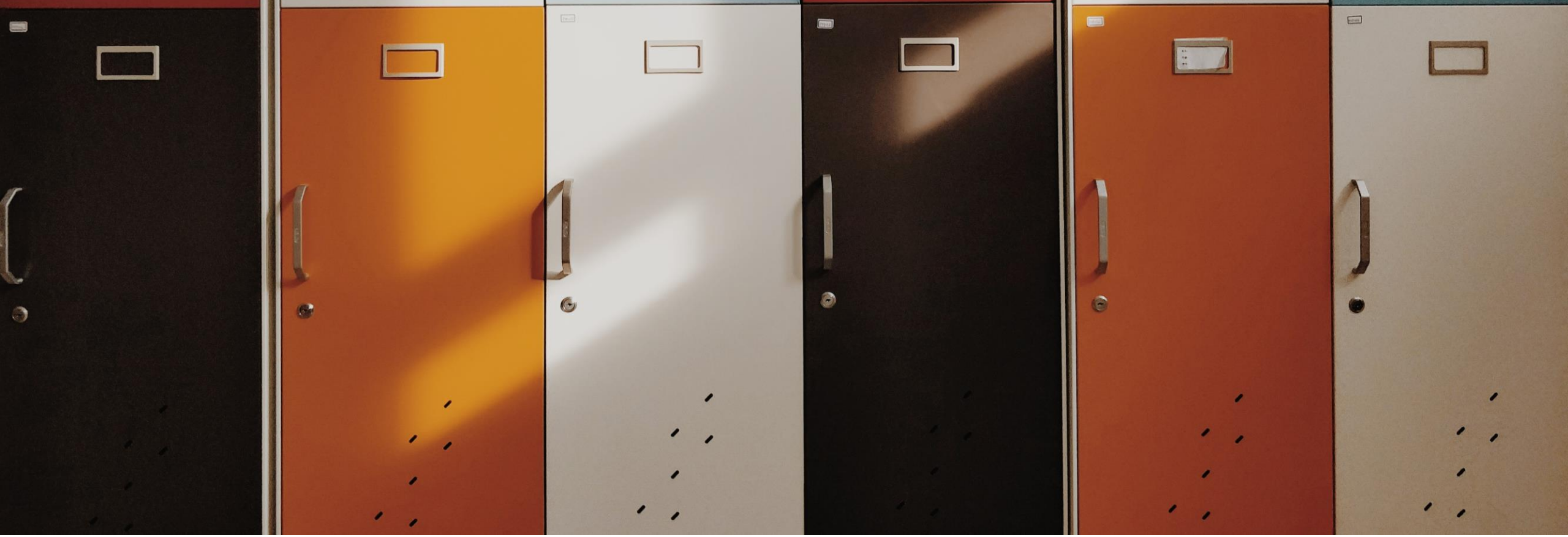


Classification

– categorizing open-ended text into two or more predefined classes based on some rules or similarities between these texts.

Can be performed based on of the three approaches:

- **Rule-based systems**
- **ML-based systems**
- **Hybrid systems**



Models, used only for summarization

(e.g. Sumy)

Models, used only for classification

(e.g. Naive Bayes, SVMs)

Models, used for both tasks

(e.g. Gensim, CNNs, RNNs, BERT-based models, GPT models, XLNet, T5)

Data access and security issues

Insufficient resources issues

Data access and security issues

(new open-source dataset should be found, that would match the original one)

Data

Amazon Product Review Dataset

Information

10 columns: *ID, Product ID, User ID, Profile Name, Helpfulness Numerator, Helpfulness Denominator, Score, Time, Summary, Text*

Structure

568.427 reviews

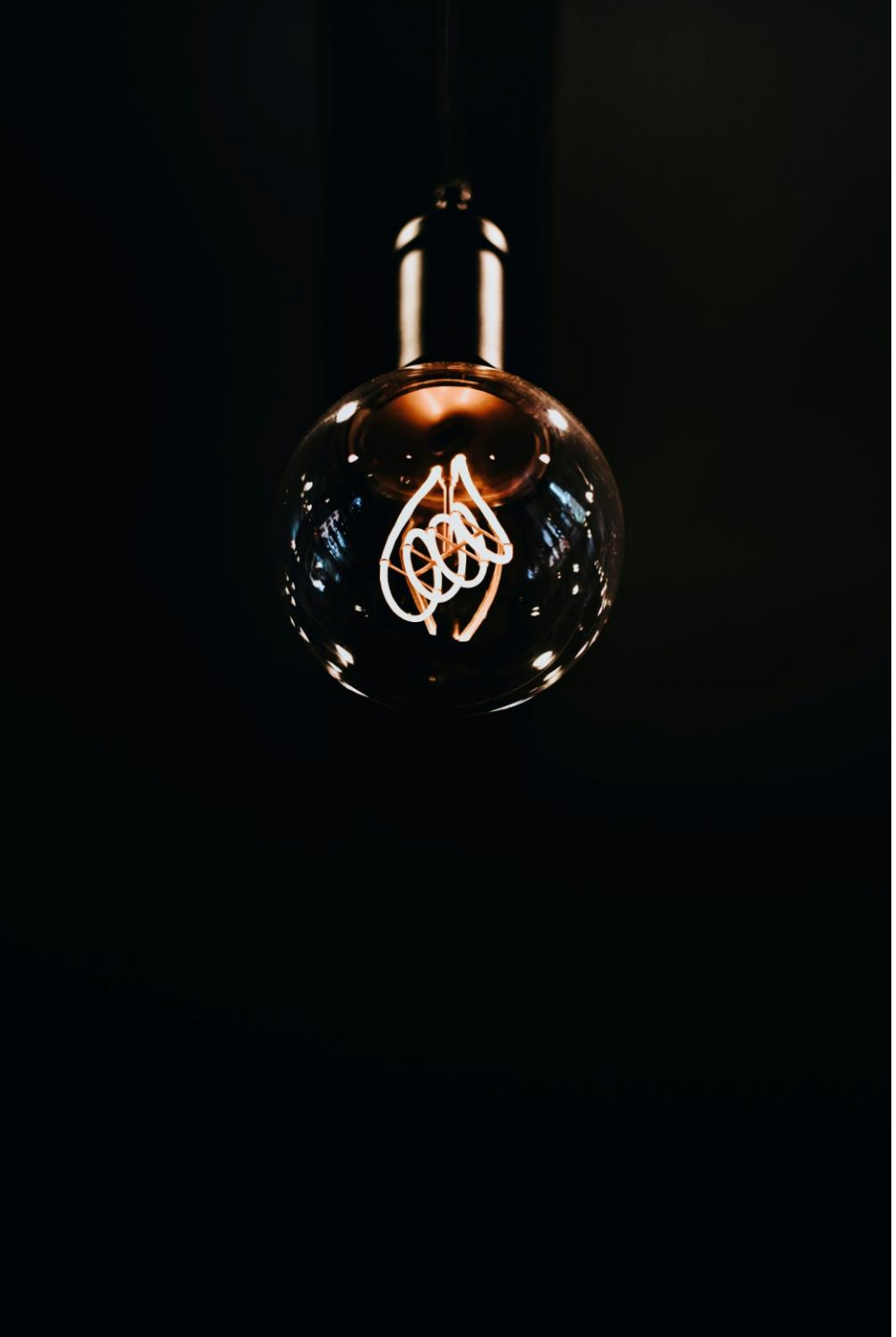
Content

2 columns kept: *Summary, Text*

Useful data

Lack of proper computational resources

(lightweight models should be found to complete the task)



OUR CHOICE: T5 model

summarization

Encoder & Decoder blocks

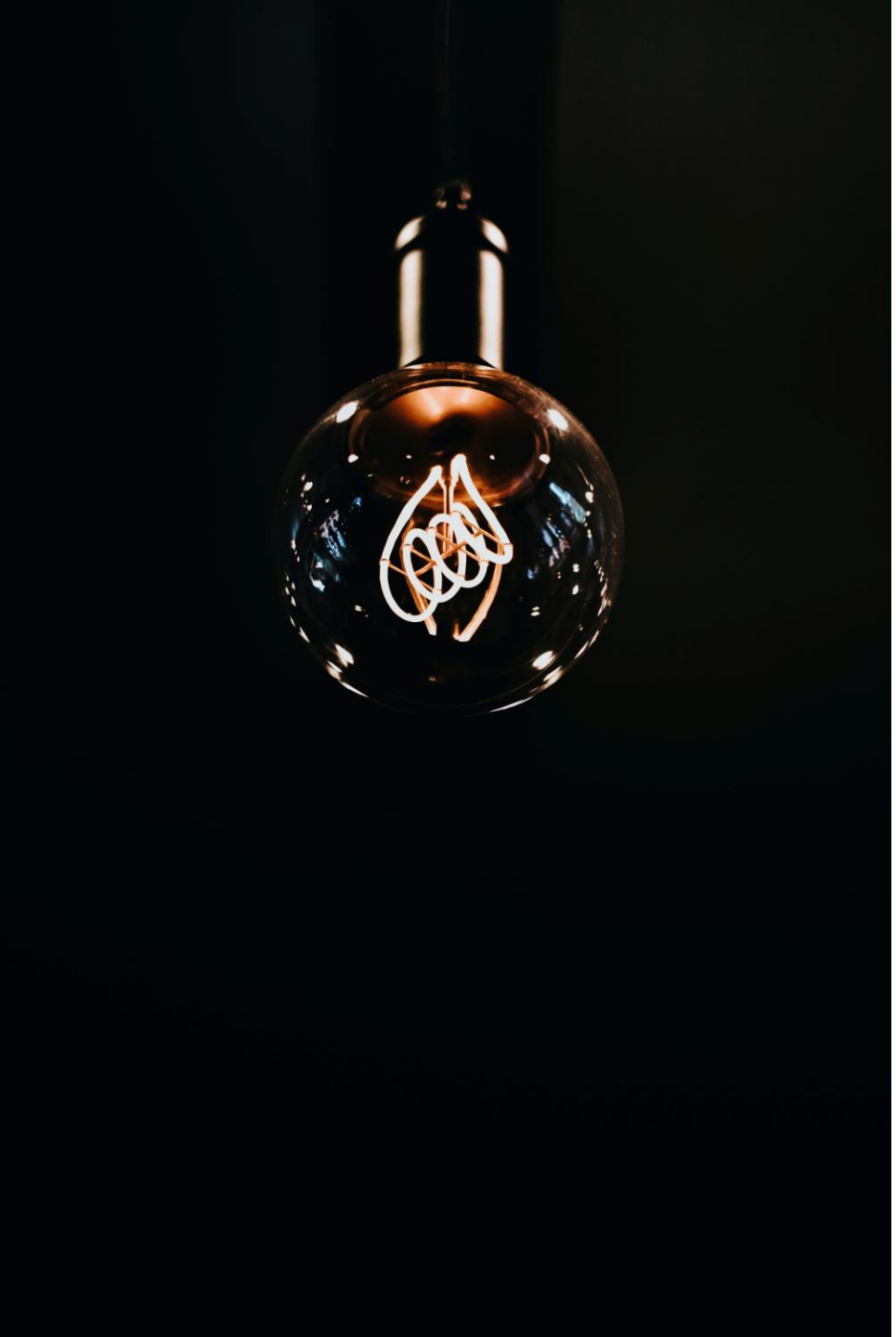
(decoder block helps model to create better summary)

The output is a text string

(many other models have labels/spans as output
→ improper output for summarization task)

Robust and extensible

(weights are assigned more properly,
the model can be easily modified to other tasks)



OUR CHOICE: DistilBERT model
classification

Small, fast, cheap

(40% less parameter than BERT → 60% faster)

Distilled & transfer-learning adapted

(mix of the distillation and transfer-learning
→ Above 90% accuracy on classification)

Open-source & flexible

(model available via HuggingFace,
retains 97% of BERT performance)

PROJECT TIMELINE

Oct, 2021

(Getting to know the supervisor, the project and the goal of it, searching for the data)

Nov, 2021

(Exploration of the dataset, metric extraction & processing ideas, building a data loader)

Dec, 2021

(Research on summarization techniques, exploring necessary packages)

Jan, 2022

(First-choice model research, baseline model building (RoBERTa), research on classification)

PROJECT TIMELINE

Feb, 2022

(RoBERTa issue handling, parameter fine-tuning, classification implementation)

Mar, 2022

(Classification model issue handling, testing and parameter fine-tuning)

Apr, 2022

(Second-choice model research and implementation (Google T5 model))

May, 2022

(New model issue handling, parameter fine-tuning, documentation preparation)

Note on summarization model change

	RoBERTa	vs.	Google T5
Pre-training			
Base of the model			
Parameter set			
Flexibility			
Resources needed			
Avg. performance			

tbd
(happy for any help :P)

2-4 slides with code snipets

2-3 slides with performance analysis

FUTURE

1. Further classification and/or clusterization of the data (based on the information, that summary contains)
2. Score prediction (very good, very bad, neutral)
3. Further fine-tuning for better summary

CONCLUSIONS

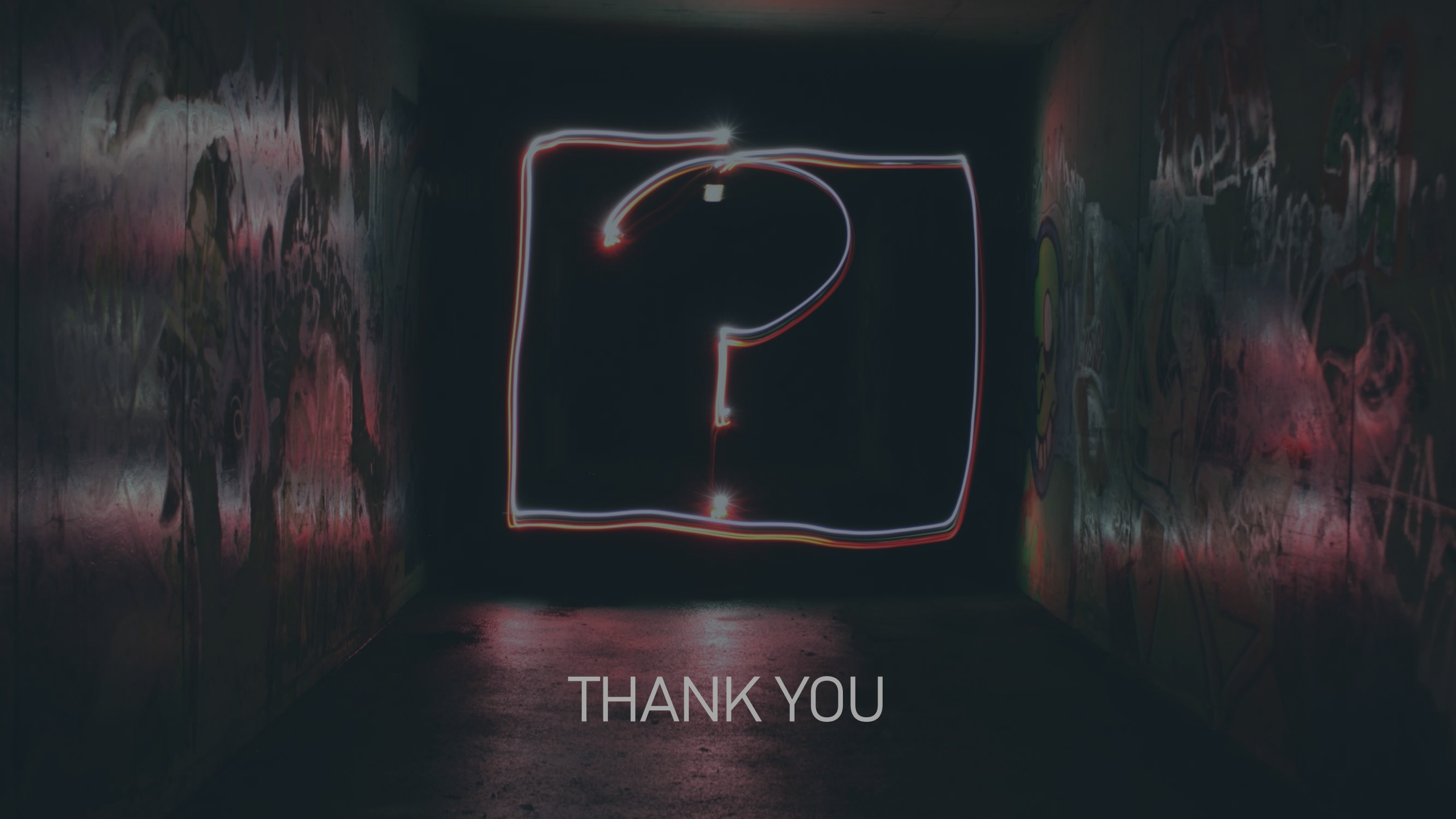
I Why task is important

II What models exist

III What model we've chosen

IV Performance analysis

V What else can be done


THANK YOU

References

Literature:

- Ref 1
- Ref 2
- Ref 3
- Ref 4
- Ref 5
- Etc.

Imagery:

- unsplash.com
- pinterest.de
- behance.net

Graphics:

- icons8.com

Additional info?
Tables?
Graphics?
Code links?