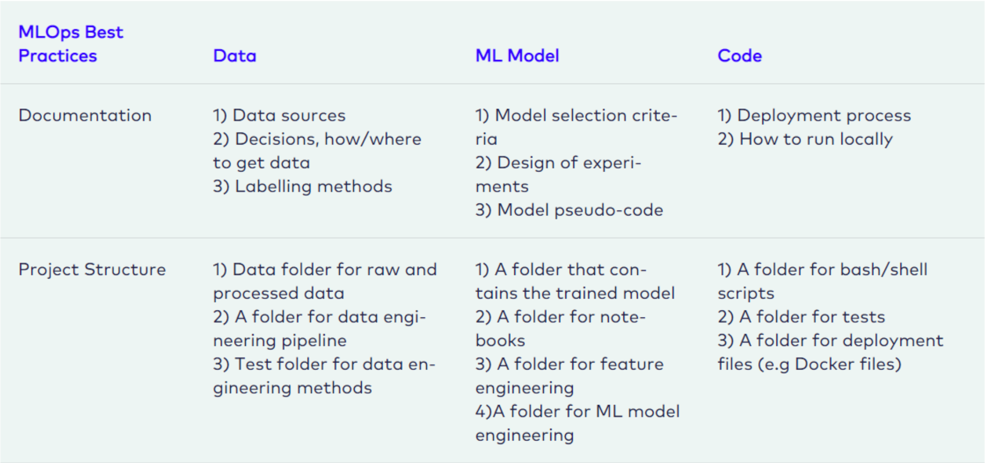
Natural Language Processing

Team

🧪 Catalyst

~ *Corporate Joker* ~

Data Collection	1. *https://www.reddit.com/r/Jokes/* 2. *https://bestlifeonline.com/what-do-you-call-jokes/* 3. *https://www.kaggle.com/code/alohahejahe/what-when-why-how/data* 4. *We scraped different websites and scraped data from reddit by using their API Calls.*
Steps of Data Wrangling	1. *Discovery* a. *We compiled the Disparate, Siloed data sources and configure each of them so they can be understood and examined to find patterns and trends in the data*
Model Training	*We used the **Sec2Sec Deep learning model** because of variable length input and output to train our Prediction System. We also combined an attention module to make our model smarter.*
Model Evaluation	*We followed **92/4/4 distribution** to train our model on test data sets. Although we faced a lot of difficulties in model evaluation, we finally managed to increase the perplexity of our model by 15%.*

<table>
<tr><td><b>Model Deployment</b></td><td><i>The Source code of the Project is hosted on github here and the Project is deployed on to streamlit.io here</i></td></tr>
<tr><td><b>MLOPS Practices</b></td><td>

<table>
<tr><th>MLOps Best Practices</th><th>Data</th><th>ML Model</th><th>Code</th></tr>
<tr><td>Documentation</td><td>1) Data sources
2) Decisions, how/where to get data
3) Labelling methods</td><td>1) Model selection criteria
2) Design of experiments
3) Model pseudo-code</td><td>1) Deployment process
2) How to run locally</td></tr>
<tr><td>Project Structure</td><td>1) Data folder for raw and processed data
2) A folder for data engineering pipeline
3) Test folder for data engineering methods</td><td>1) A folder that contains the trained model
2) A folder for notebooks
3) A folder for feature engineering
4)A folder for ML model engineering</td><td>1) A folder for bash/shell scripts
2) A folder for tests
3) A folder for deployment files (e.g Docker files)</td></tr>
</table>
</td></tr>
</table>