TEXT MINING & SEARCH

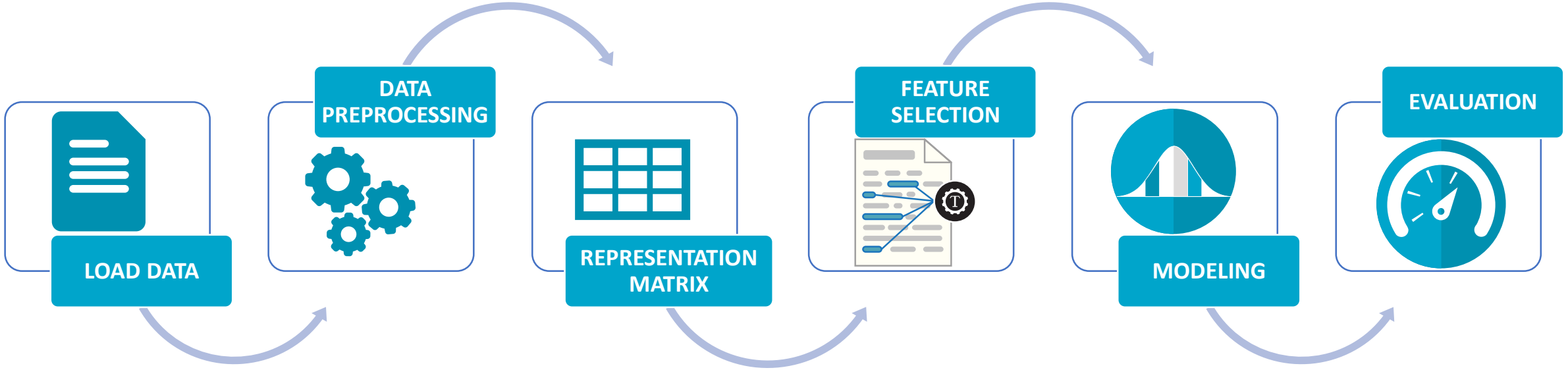
20 Newsgroups dataset Classification

Caronte Martina – 789451 – CLAMSES

Rola Stefano – 790383 – DS

Sisti Sara - 789909 – CLAMSES

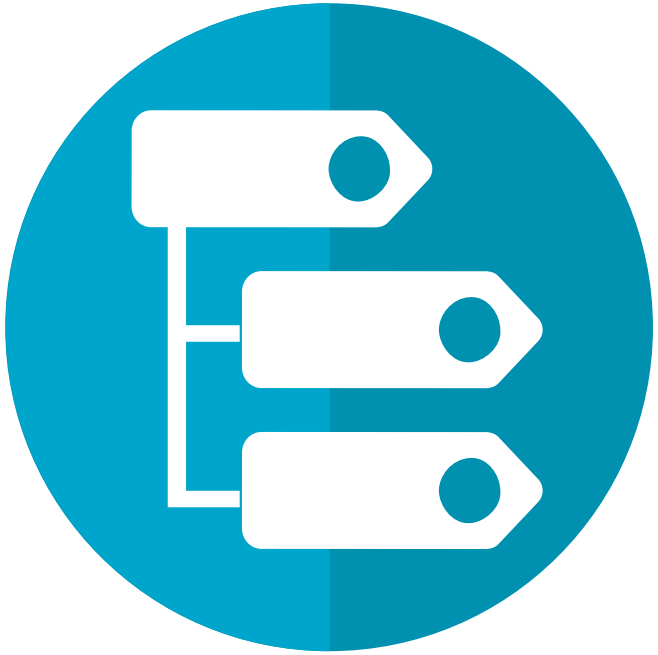
Workflow


LOAD DATA
DATA PREPROCESSING
REPRESENTATION MATRIX
FEATURE SELECTION
MODELING
EVALUATION

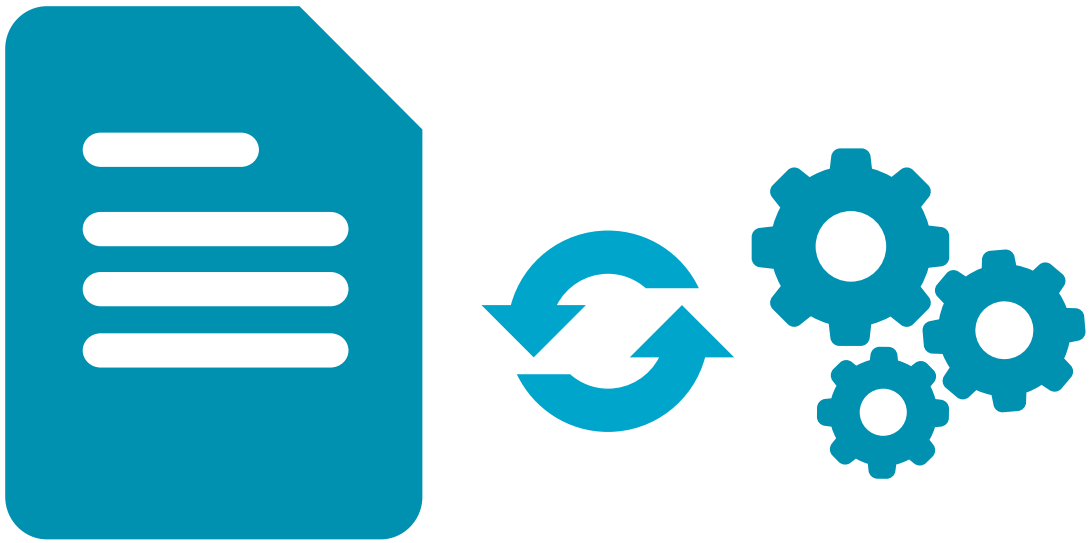
Data

20,000 documents
partitioned
(nearly) evenly
in 20 newsgroups

Splitting
in training set (200x20)
e test set (150x20)

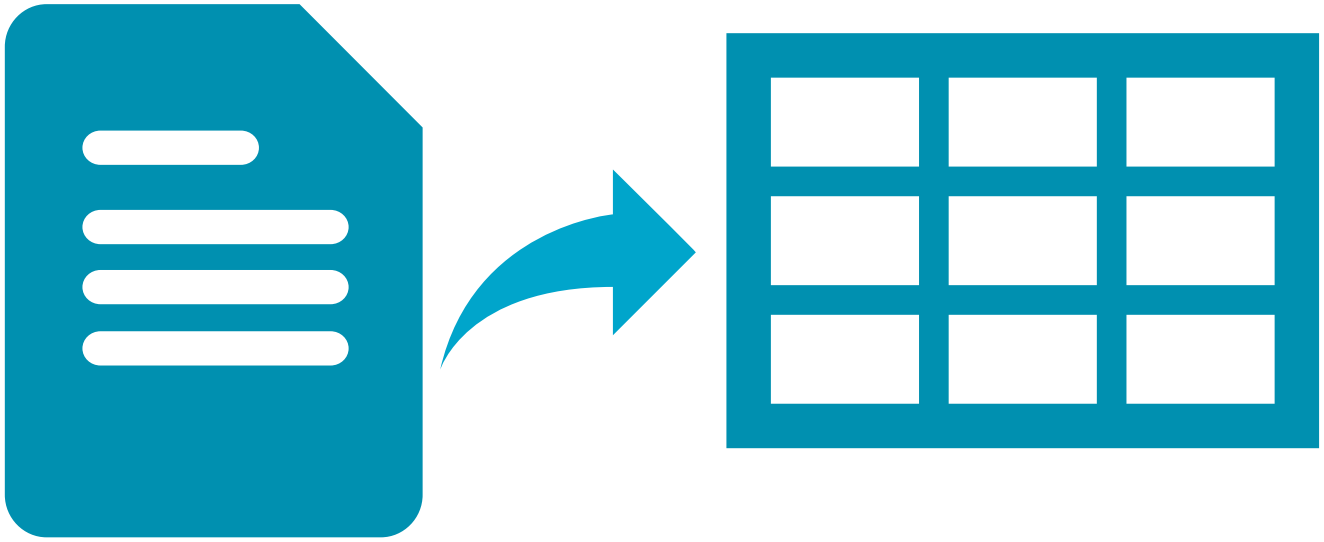


Preprocessing



- **Replace contraction**
- **To lowercase**
- **Remove punctuation and numbers**
- **Stemming**
- **Remove Stopwords**

Representation matrix

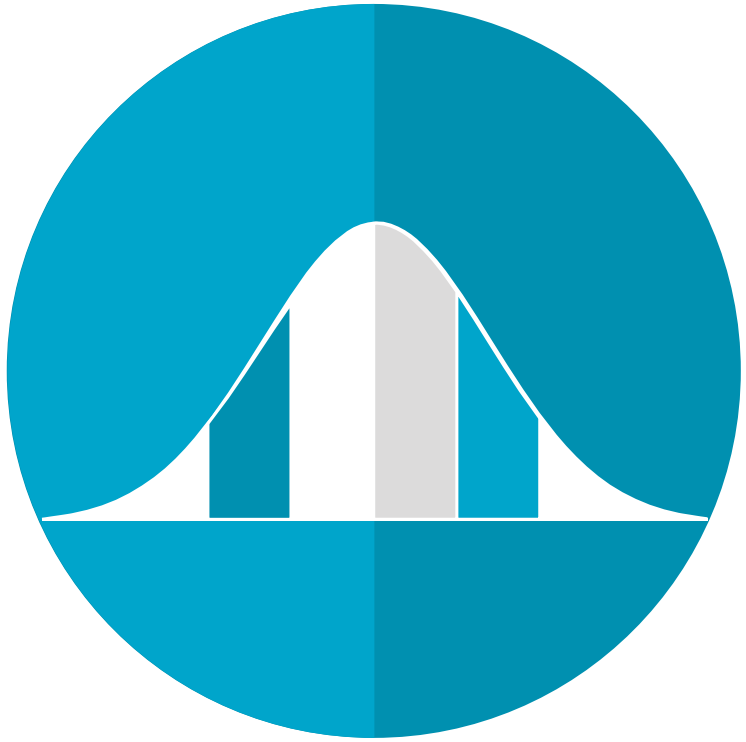


- **After many attempts the best performing (accuracy) representation is unigrams with TF-IDF weighting.**
- **Removed terms with sparsity > 0.99**

Feature selection

- **In each group selected the 60 most frequent terms.**
- **Removed duplicated terms.**
- **Final matrix representation: about 350 terms.**

Modeling



- **Decision Tree**
- **Support Vector Machine**
- **K-nearest neighbors**
- **Random Forest**
- **Neural Network**

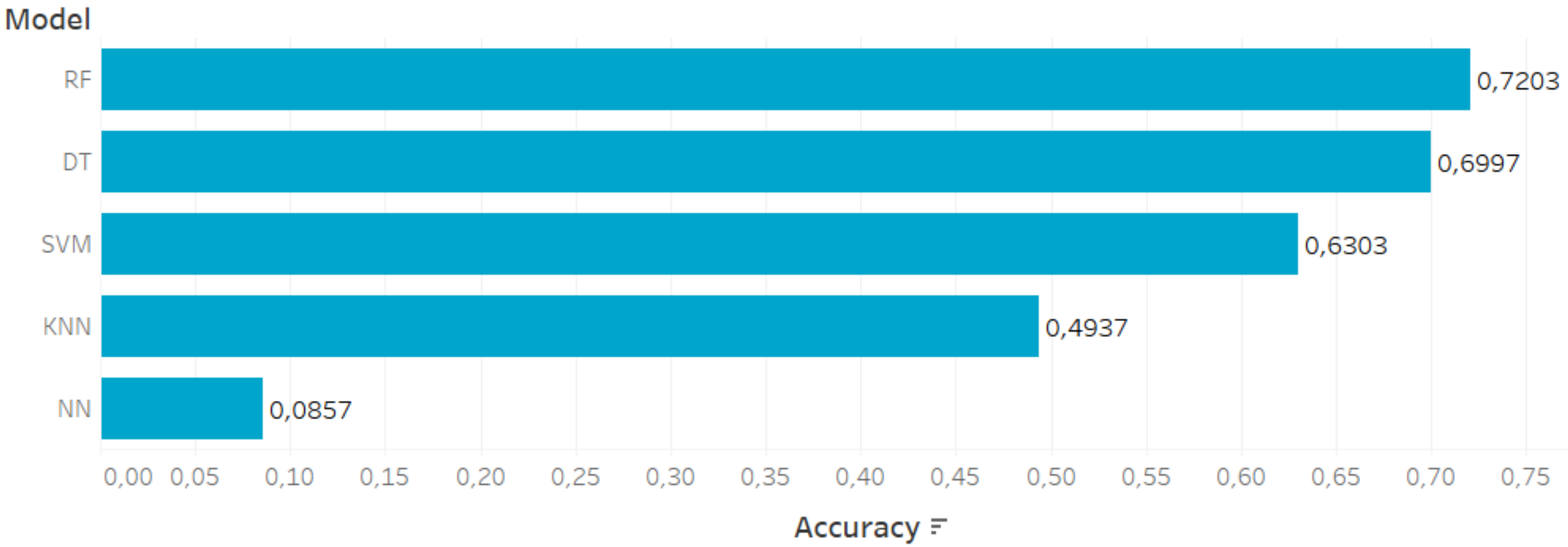
Evaluation

Training set

```
## Accuracy
##        Min. 1st Qu.  Median    Mean 3rd Qu.    Max. NA's
## dt  0.66125 0.66375 0.67625 0.67375 0.67750 0.69000    0
## svm 0.57125 0.58625 0.59750 0.59225 0.60125 0.60500    0
## knn 0.44750 0.45375 0.47500 0.47050 0.48750 0.48875    0
## rf  0.67375 0.70125 0.70875 0.70325 0.71000 0.72250    0
## nn  0.08875 0.09625 0.09750 0.09675 0.10000 0.10125    0
```

Evaluation

Test set



GRAZIE PER L’ATTENZIONE