ECIS European Conference on Information Systems JUNE 2020 Marrakech

Maintenance of Data Richness in Business Communication Data

Muhammed-Fatih Kaya & Mareike Schoop

UNIVERSITY OF HOHENHEIM

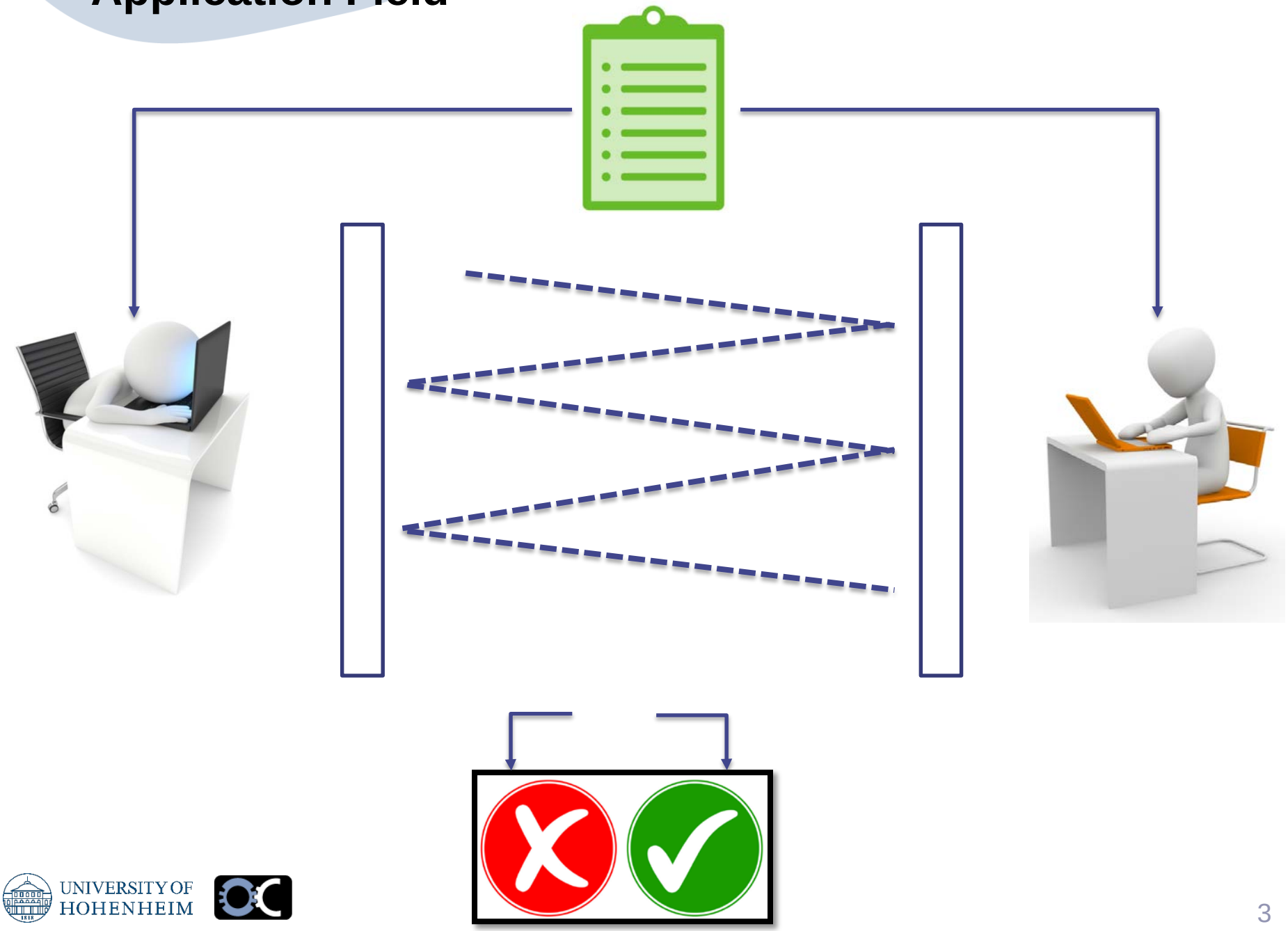
Information Systems Department I (580 A)
University of Hohenheim
Institute of Interorganisational Management & Performance
Prof. Dr. Mareike Schoop

200 1818 2018 YEARS UNIVERSITY OF HOHENHEIM

Agenda

1. Motivation
2. Research Framework
3. Results
4. Discussion & Outlook

Application Field



Challenges

- Communication
  - Unstructured data
  - High dimensional attributes

- Required processing (Reddy and Rajinikanth, 2017)
  - Structured representation
  - Transformation steps
  - Curse of dimensionality

- Maximisation of information richness **(!)**
  (Nahm and Mooney, 2002)

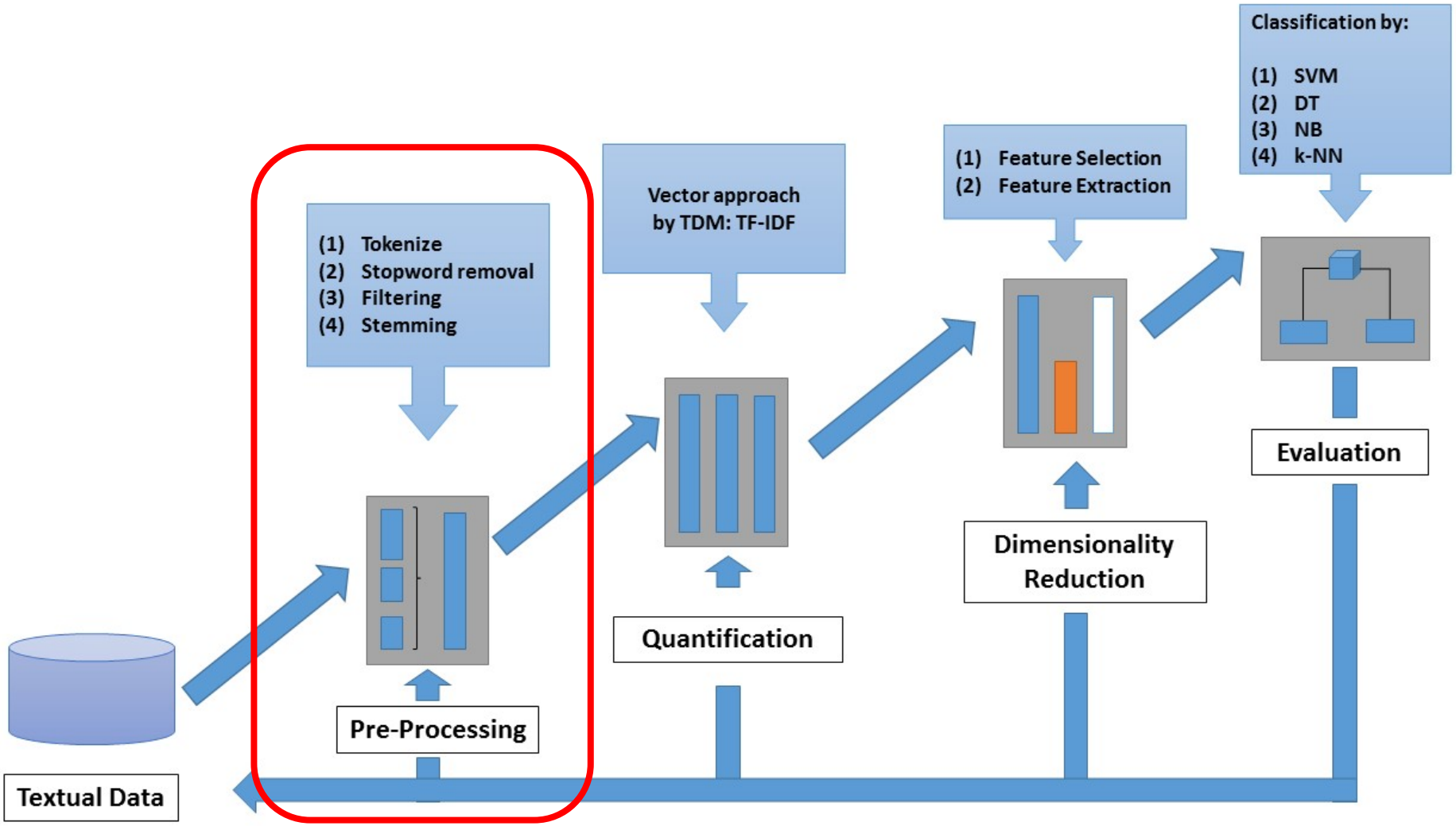
Research Question

Which Text Mining **quantification techniques** and **dimensionality reduction approaches** maintain data richness in business communication data?

Agenda

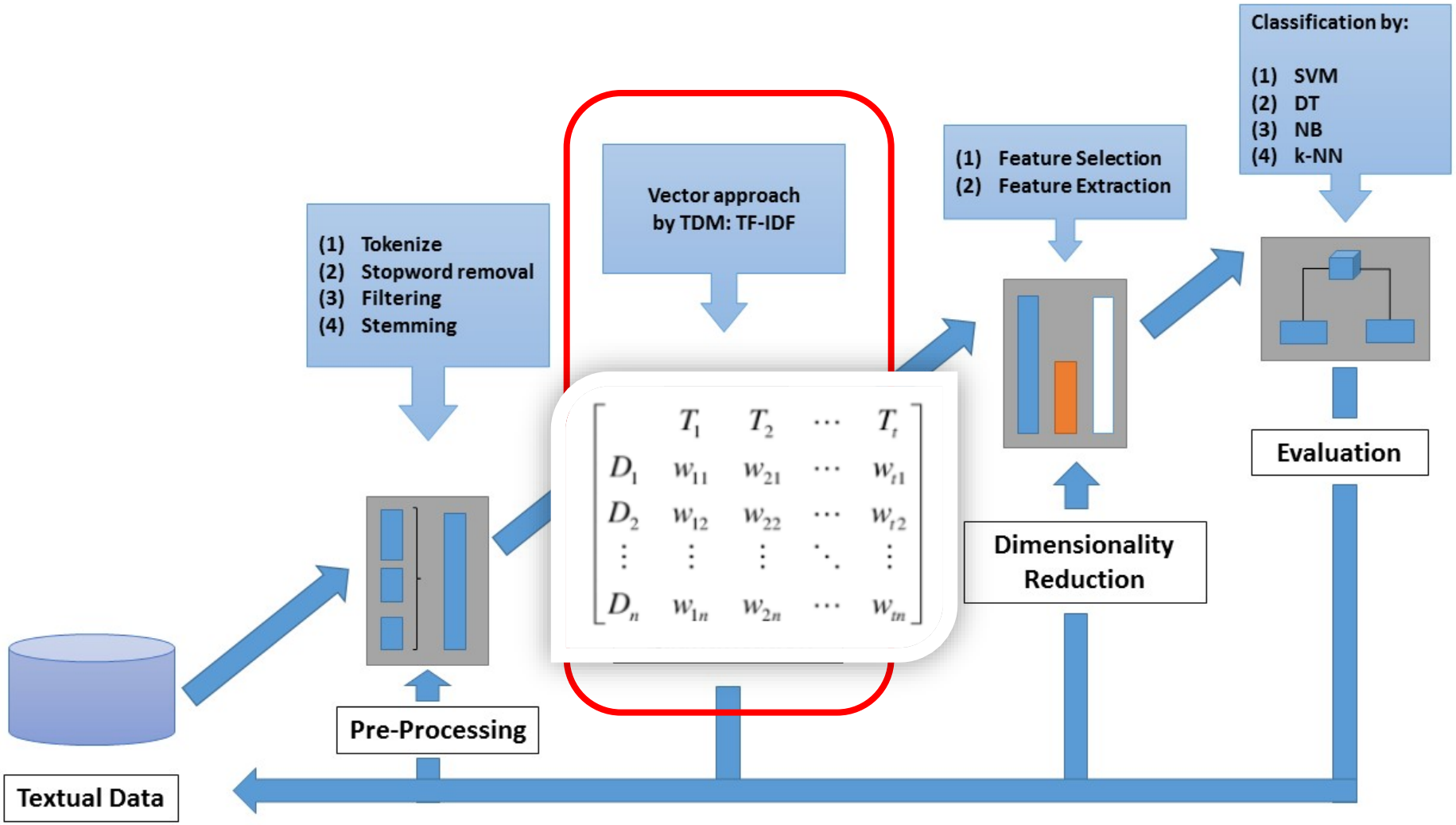
1. Motivation
2. **Research Framework**
3. Results
4. Discussion & Outlook

Research Framework



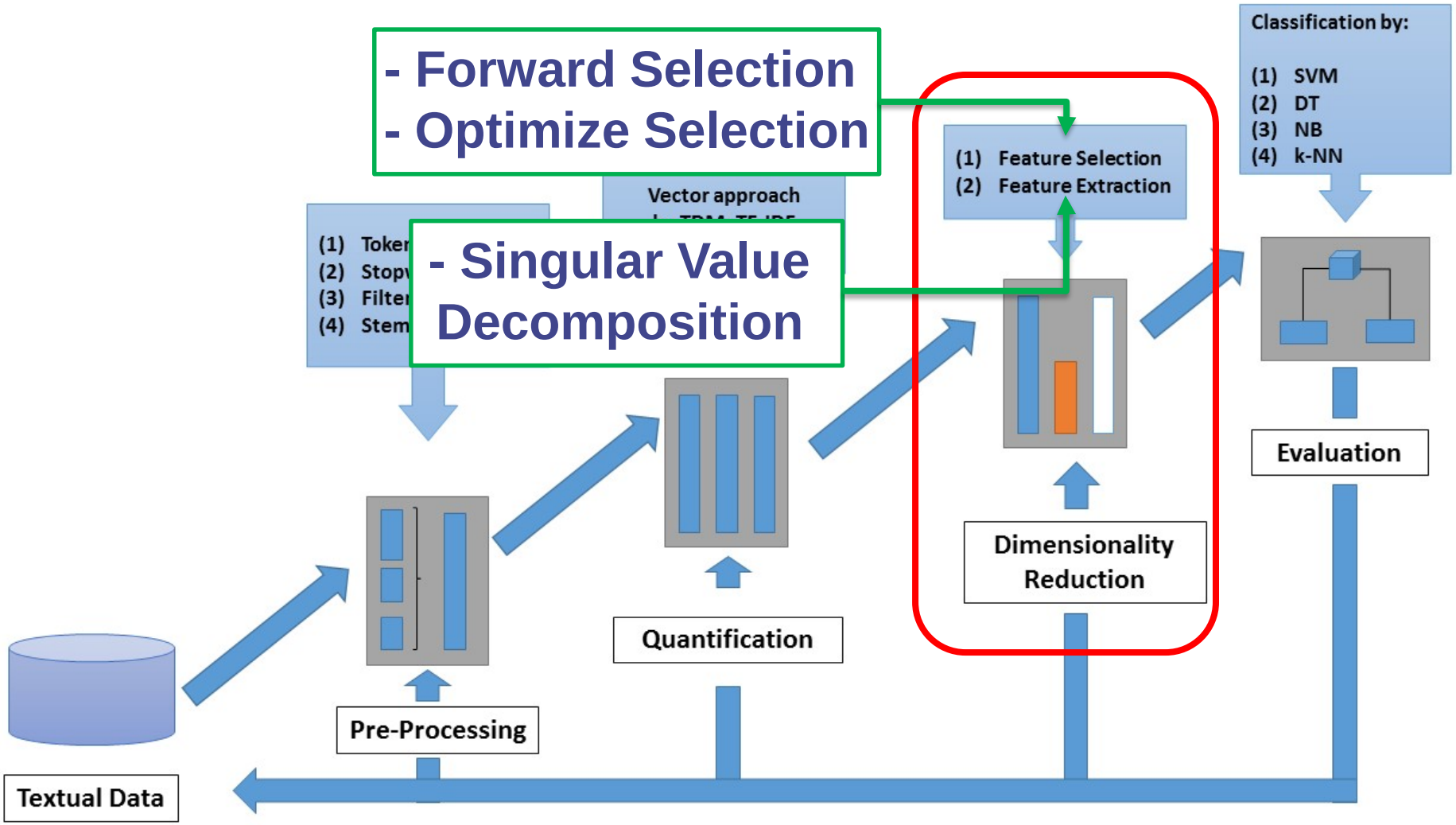
(cf. Fayyad et al., 1996; Kumar and Bhatia, 2013; Vidya and Aghila, 2010)

Research Framework



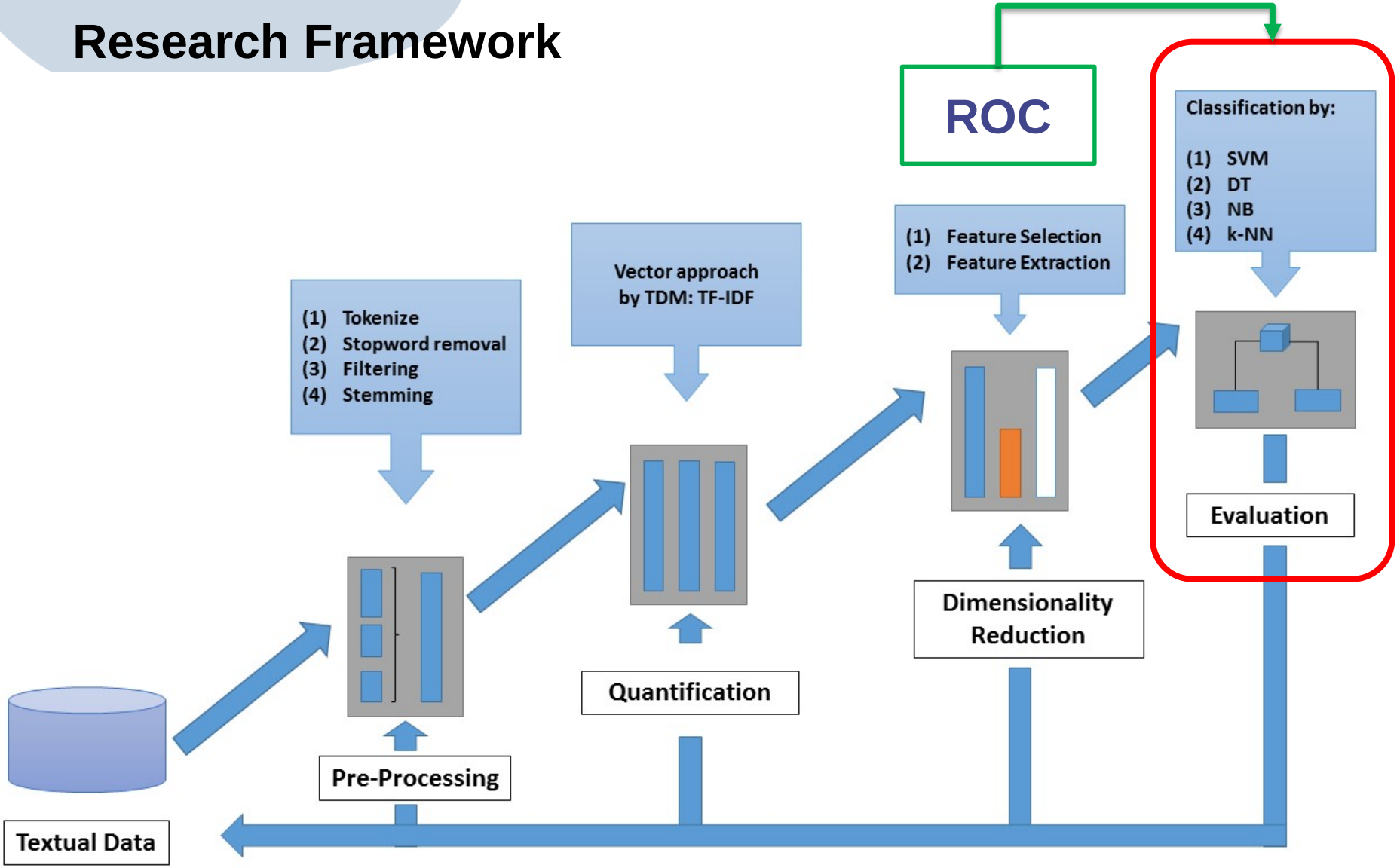
(cf. Fayyad et al., 1996; Kumar and Bhatia, 2013; Vidya and Aghila, 2010)

Research Framework



(cf. Fayyad et al., 1996; Kumar and Bhatia, 2013; Vidya and Aghila, 2010)

Research Framework



(cf. Fayyad et al., 1996; Kumar and Bhatia, 2013; Vidya and Aghila, 2010)

Agenda

1. Motivation
2. Research Framework
3. **Results**
4. Discussion & Outlook

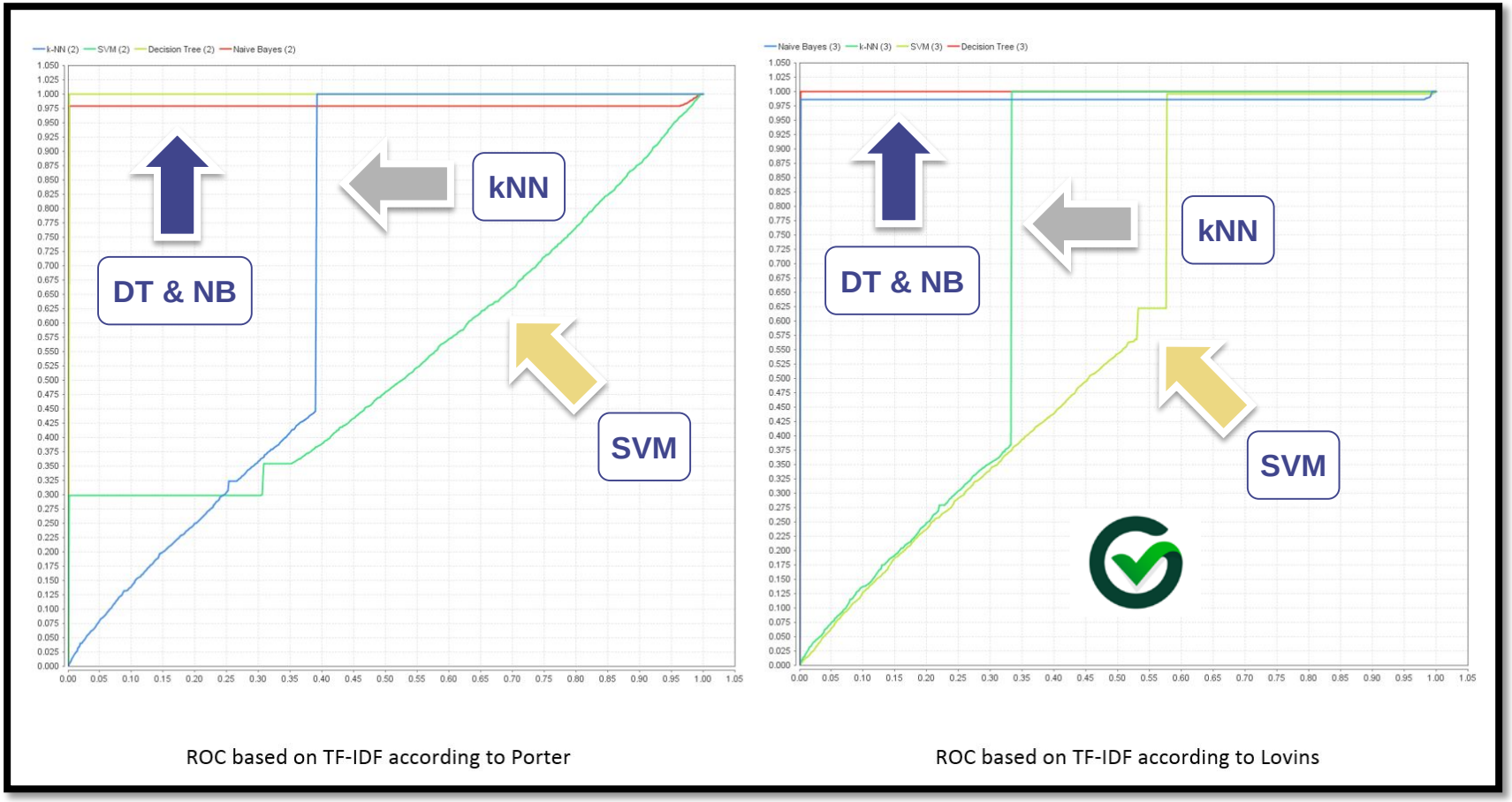
Results – Dimensions

Dimensionality Reduction	Model Type	Stemming Base	Resulting Dimensions
Feature Selection	Forward Selection	Porter	16 of 9661
Feature Selection	Forward Selection	Lovins	17 of 8880
Feature Selection	Optimize Selection	Porter	4864 of 9661
Feature Selection	Optimize Selection	Lovins	4399 of 8880
Feature Extraction	Singular Value Decomposition	Porter	658 of 9661
Feature Extraction	Singular Value Decomposition	Lovins	639 of 8880

- Quantity of resulting dimensions:

Forward Selection < Singular Value Decomposition < Optimize Selection

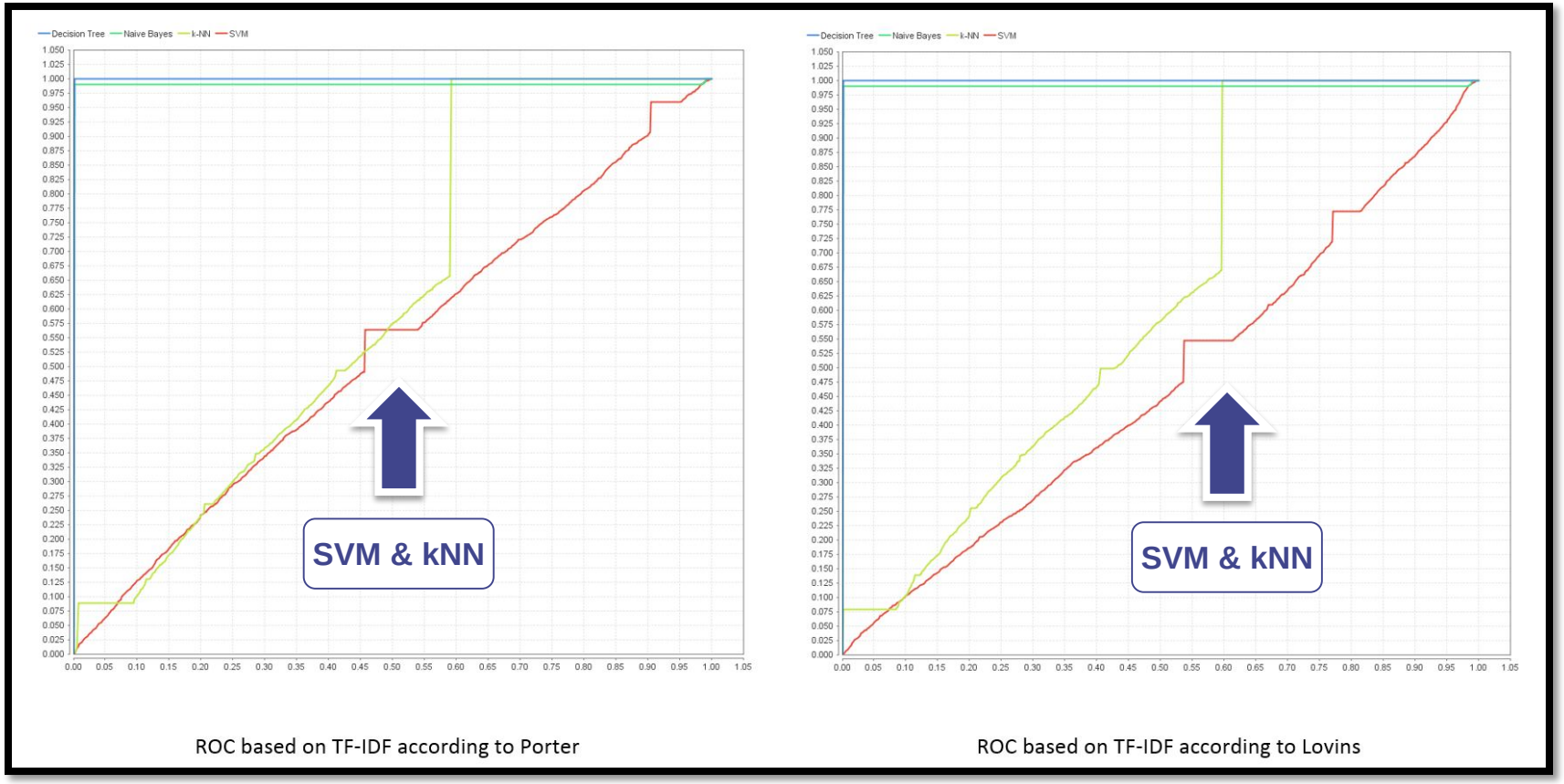
Results – ROC of Optimized Selection



ROC based on TF-IDF according to Porter

ROC based on TF-IDF according to Lovins

Results – ROC of Singular Value Decomposition



ROC based on TF-IDF according to Porter

ROC based on TF-IDF according to Lovins

Agenda

1. Motivation
2. Research Framework
3. Results
4. **Discussion & Outlook**

Discussion

- Stemming
  - Precision of Lovins' stemming (Willett, 2006)
  - 781 fewer dimensions

- Dimensionality Reduction
  - Forward Selection:
    - Problem of local optima (Gheyas and Smith, 2010)
  - Optimize Selection:
    - Approximation to global optimum (Venkatesh and Anuradha, 2019)
  - Singular Value Decomposition:
    - Linearly dependent dimensions (Guyon and Elisseeff, 2006)
    - Similar number of dimensions

Discussion

- Evaluation by ROC analysis
  - Forward Selection
    - No learning possible → Overfitting (Flach, 2016)
    - Stagnation in the local optimum
  - Optimize Selection
    - SVM makes the difference
    - No learning process for DT and NB
  - Singular Value Decomposition
    - No learning process for DT and NB

OS-Lovins* > *OS-Porter* > *(SVD-Porter = SVD-Lovins)

Outlook

- Pattern Recognition on behavioural patterns
  - Clustering
  - Topic Modelling
  - Association Rule Discovery
- Additional aggregation levels
  - Message level
  - Negotiation phases
  - […]

Contact

Thank you for your attention!

Muhammed-Fatih Kaya M. Sc.
University of Hohenheim
Information Systems Group (580 A)

Web: https://wi1.uni-hohenheim.de
E-Mail: mkaya@uni-hohenheim.de