Empirical Study of a Two-Step Approach to Estimate Translation Quality

J. González-Rubio, J.R. Navarro-Cerdán, F. Casacuberta

jegonzalez@dsic.upv.es, jonacer@iti.upv.es, fcn@dsic.upv.es

Pattern Recognition and Human Language Technology Group

Instituto Tecnológico de Informática

Universitat Politècnica de València (Spain)

Work supported by the EU 7th Framework program (FP/2007-2013) under the CasMaCat project (gran nº 287576)

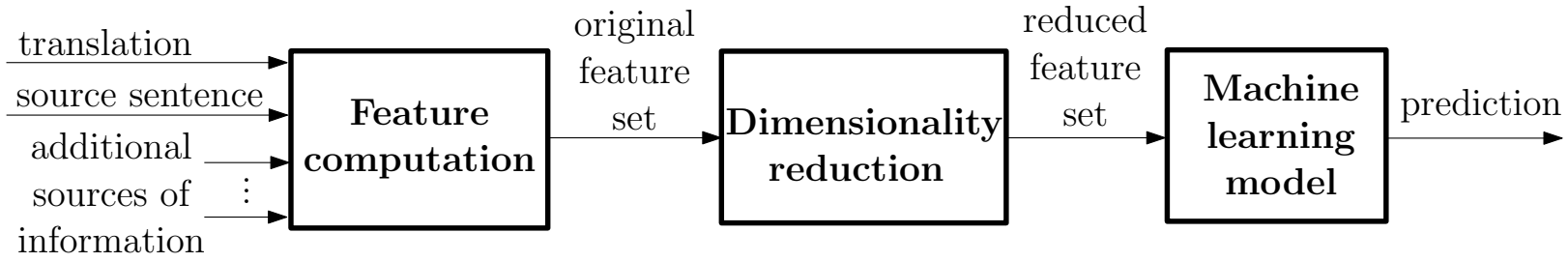
Overview

- Introduction
- Proposed Two-Step Quality Estimation Approach
- Experimental Setup
- Results
- Conclusions

Introduction

Motivation

- Quality estimation (QE) is a key element in practical translation systems
- Usually addressed as a regression problem
  - Predict a quality score from a set of translation features
- Problem: translation features are ambiguous, noisy, and collinear
- Chosen solution: a two-step training methodology



Two-Step Quality Estimation Approach

Dimensionality Reduction

- Based on Partial Least Squares Regression (PLSR)
- Widely-used PCA takes into account only the features
  - Principal components (PCs) contain almost not redundancy...
  - ...but they do not necessarily are the best features for prediction
- In contrast, PLSR does take into account the values to be predicted
  - The new set of Latent Variables (LVs) contain almost no redundancy
  - Additionally, they explain most of the variability in the quality scores

Prediction Model

- Goal: predict the actual quality scores from the LVs
- Model: Support Vector Machines for regression (SVR)
- Good empirical prediction accuracy in a number of tasks
- Widely used in the QE literature

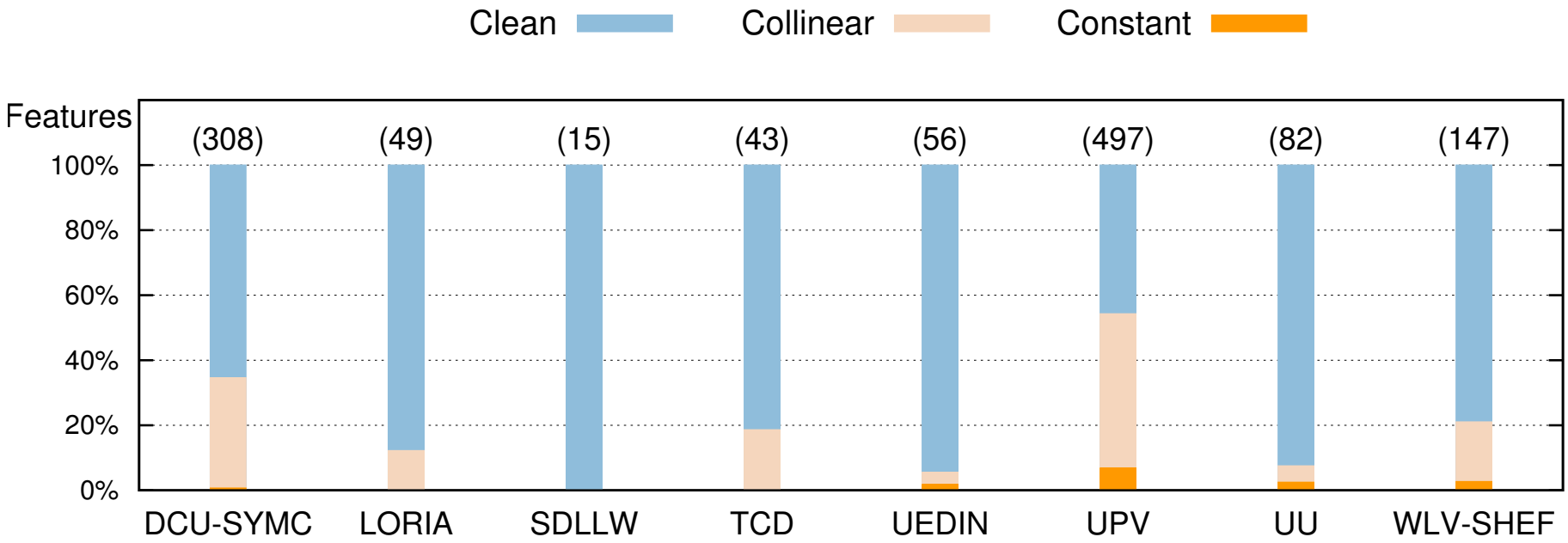
Experimental Setup

Corpus

- English-Spanish news texts from WMT 2012 QE task
- 1832 translations for training and 422 for test
- Each translation has a real-valued score between one and five
- Post-edition effort likert scale:

  5: The translation requires little editing to be publishable
  4: 10%–25% of the translation needs to be edited
  3: 25%–50% of the translation needs to be edited
  2: 50%–70% of the translation needs to be edited
  1: The translation must be translated from scratch

Feature Sets



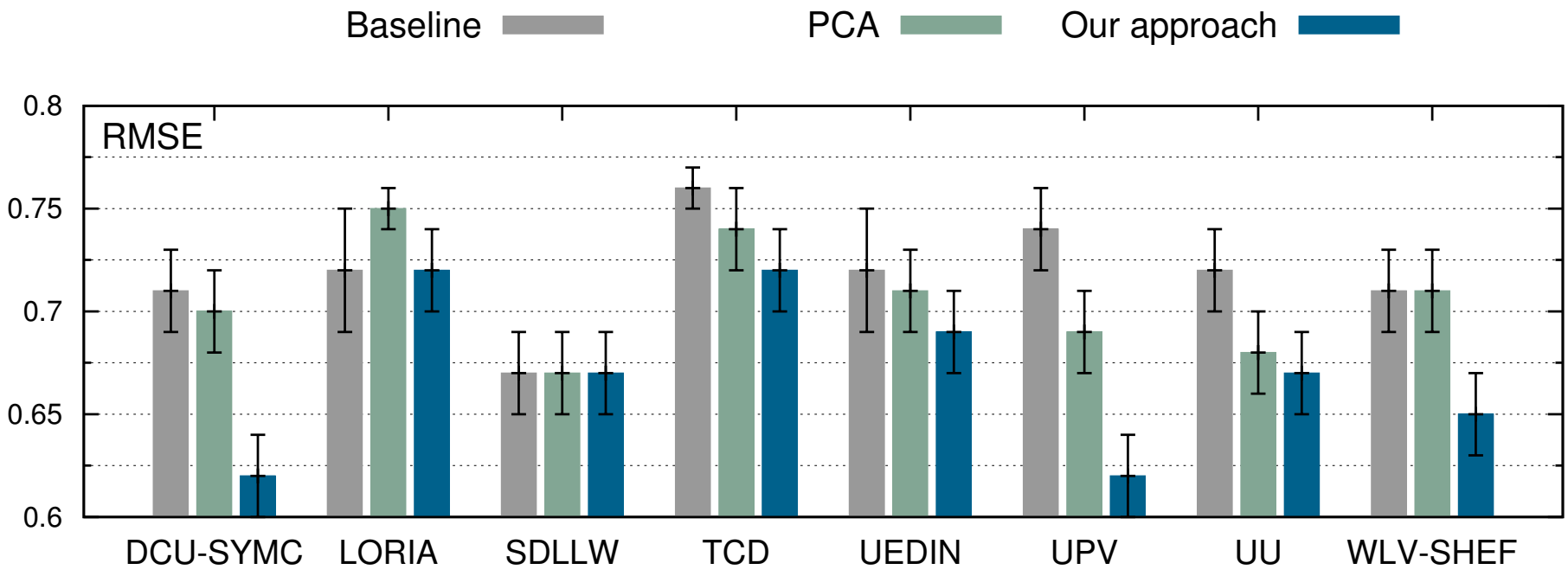
- Wide variety of experimental conditions

Experimental Methodology

- Evaluation metric: Root Mean Squared Error (RMSE)
- Free parameters optimized by 10-fold cross-validation
  - Number of LVs and SVR meta-parameters
- 8 dev-train folds, one dev-tuning fold, and one dev-test fold
  - Result: averaged prediction accuracy for the separated dev-test folds
- Final models built with the whole training using best parameter values

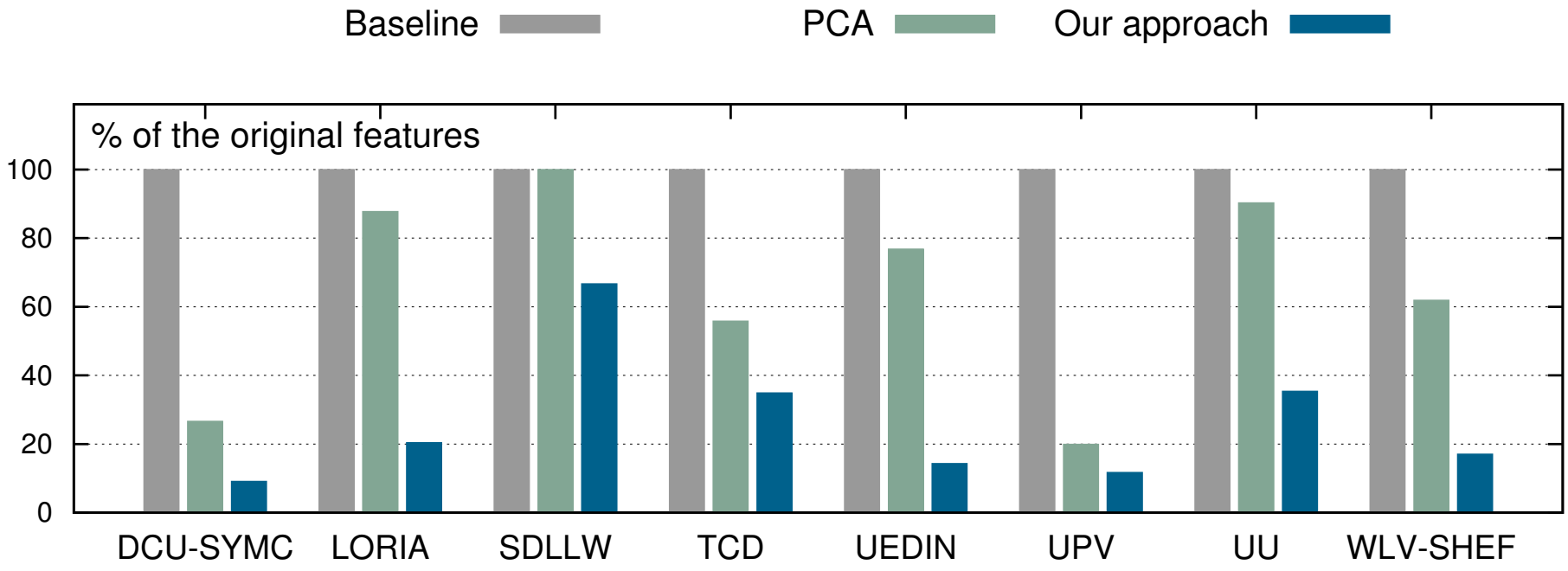
Results

Cross-Validation RMSE Results



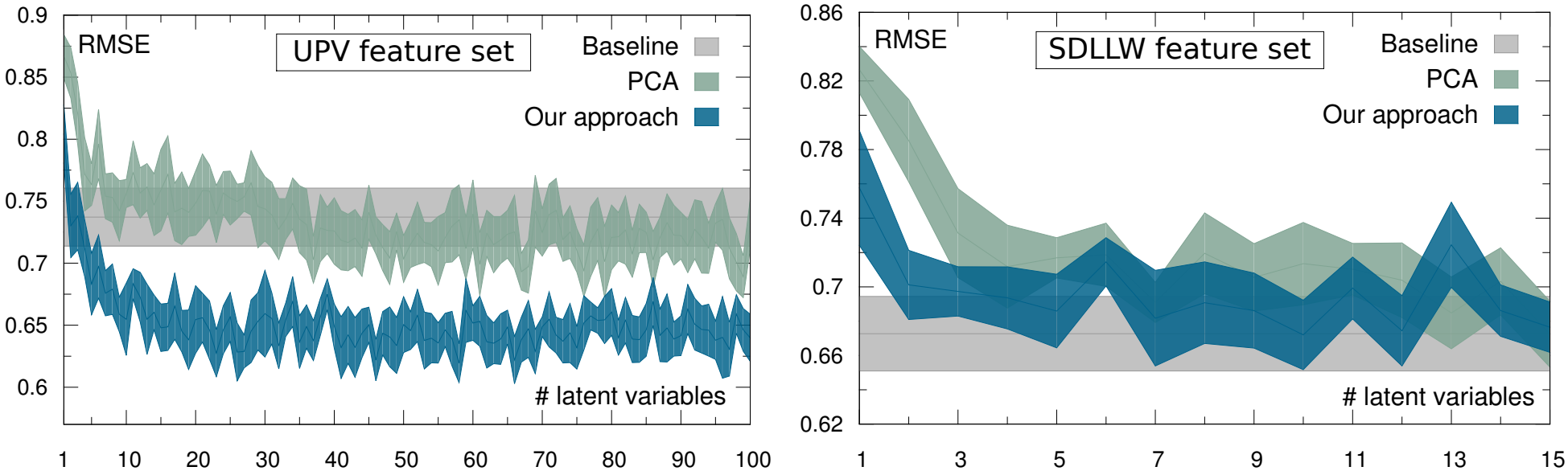
- Equal or lower prediction error than Baseline and PCA
  - Error reduction correlated with the number of noisy features

Cross-Validation Feature Reduction Ratio



- About half the number of LVs than PCs
- Operational time of the QE system largely reduced

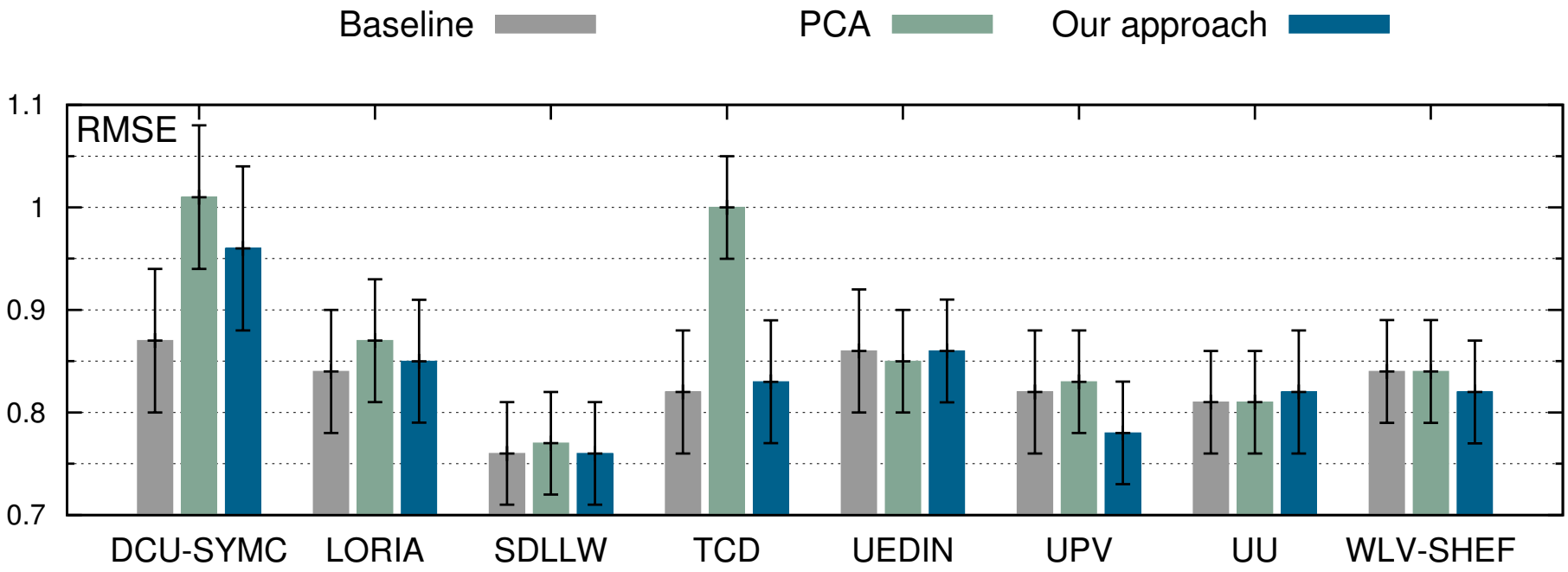
Cross-Validation learning curves



Band indicates the 95% confidence interval of prediction accuracy (RMSE)

- Larger and faster error reduction for higly-redundant sets (left plot)
- Same accuracy with less features for concise sets (right plot)

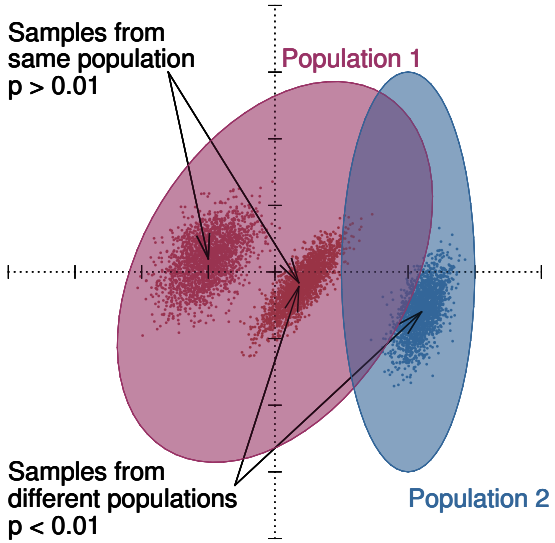
Test Results



- Different result respect to cross-validation, why?

Analysis of Test Results

- Hypothesis: training partition did not adequately represent test
- Studied by a series of Hotelling's two-sample T^2 tests
  - Multivariate analog of Student's t-test
  - Compares two independently drawn samples
    - E.g., training and test partitions
  - Do they belong to the same population?



Analysis of Test Results II

- T^2 tests indicated that training and test were from different populations
  - Main reason: data scarcity (only 1832 training samples)
- Further analysis of each individual feature:
  - Most had statistically different values in test
  - Between one quarter and three quarters depending on the set
- In contrast, only about only 1% between dev-train and dev-test folds

Conclusions

Conclusions

- Empirical results showed the soundness of the proposed approach
  - Improvements in prediction accuracy
  - Large feature reduction ratios
- Not so good test results due to data scarcity
- Feature reduction boosts QE scalability and time-efficiency
  - Suitable to be applied in scenarios with temporal restrictions
  - Allows the use of thousands of features

Thank you,

questions?