Data Science @UL-FRI

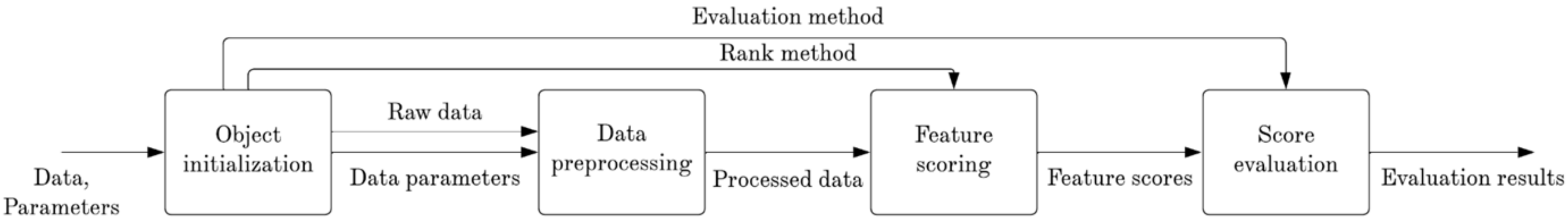
# Feature selection on large and sparse datasets

Leon Hvastja, Amadej Pavšič

**Advisors:** Jure Demšar, Blaž Mramor, Blaž Škrlj

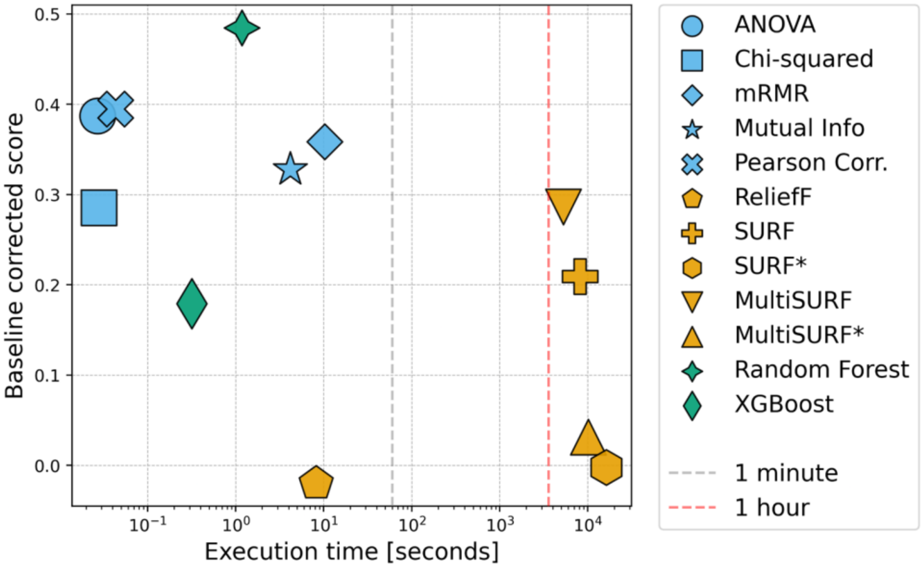
# Problem & dataset

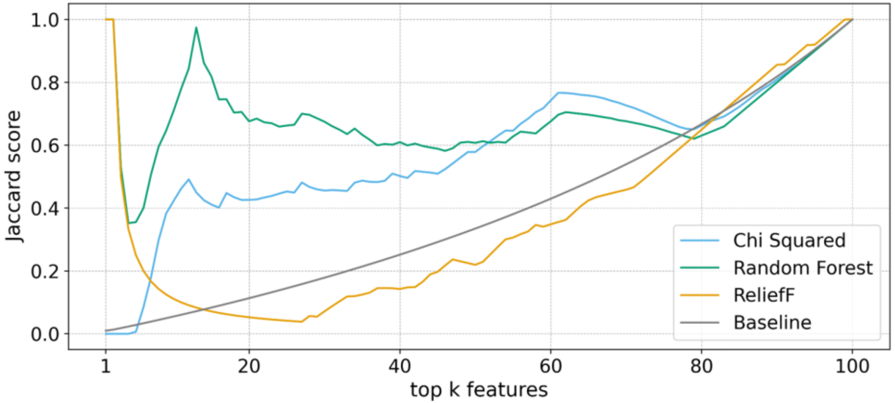
- Outbrain, web recommendation platform
- Sparsity, high cardinality, uninformative features
- Efficient feature selection approach
- Ranking & evaluation pipeline

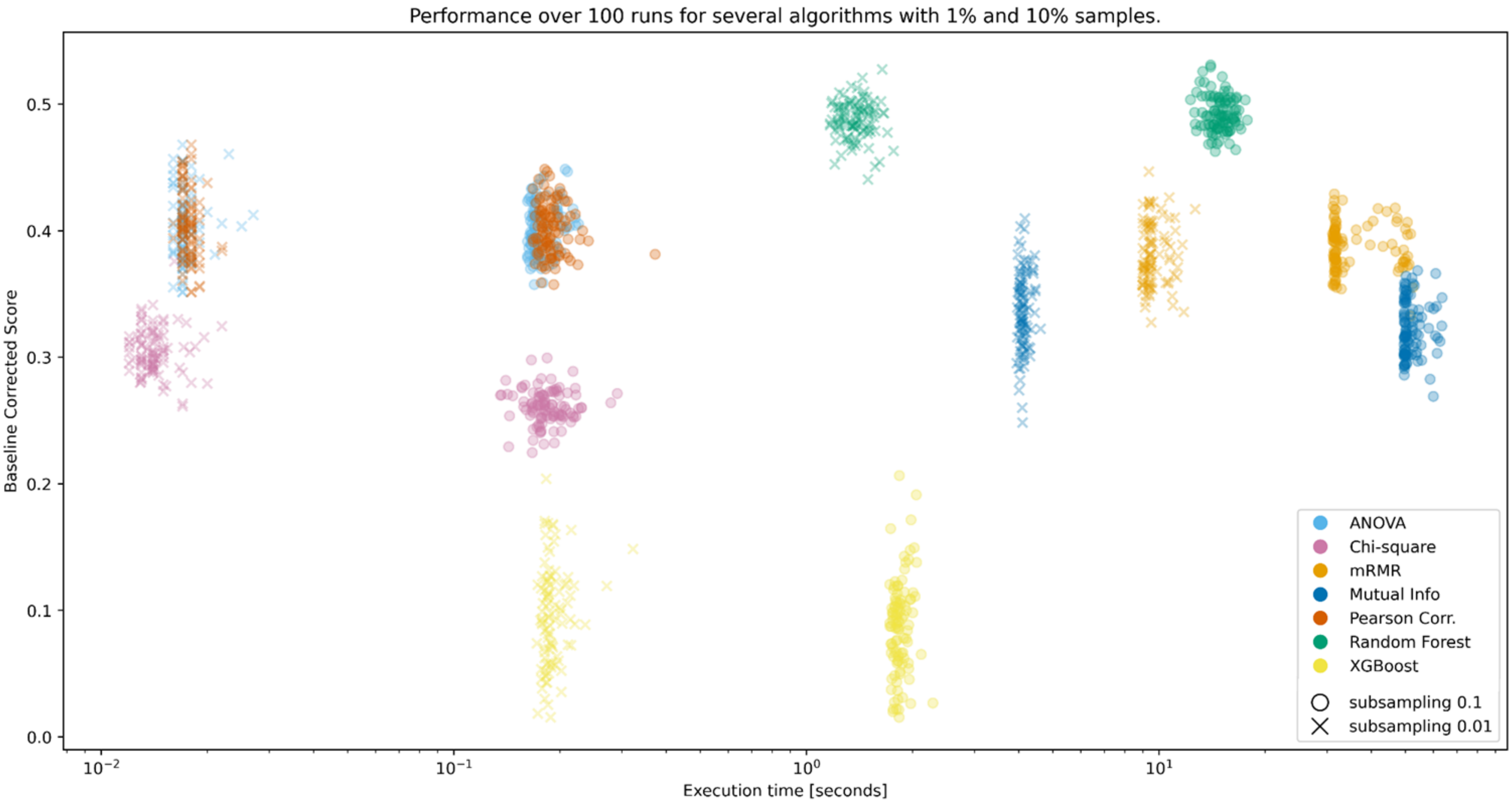


# Algorithms

- Filter and wrapper methods
- Using statistical properties
- Relief based approaches
- Baseline corrected score
- First eliminations

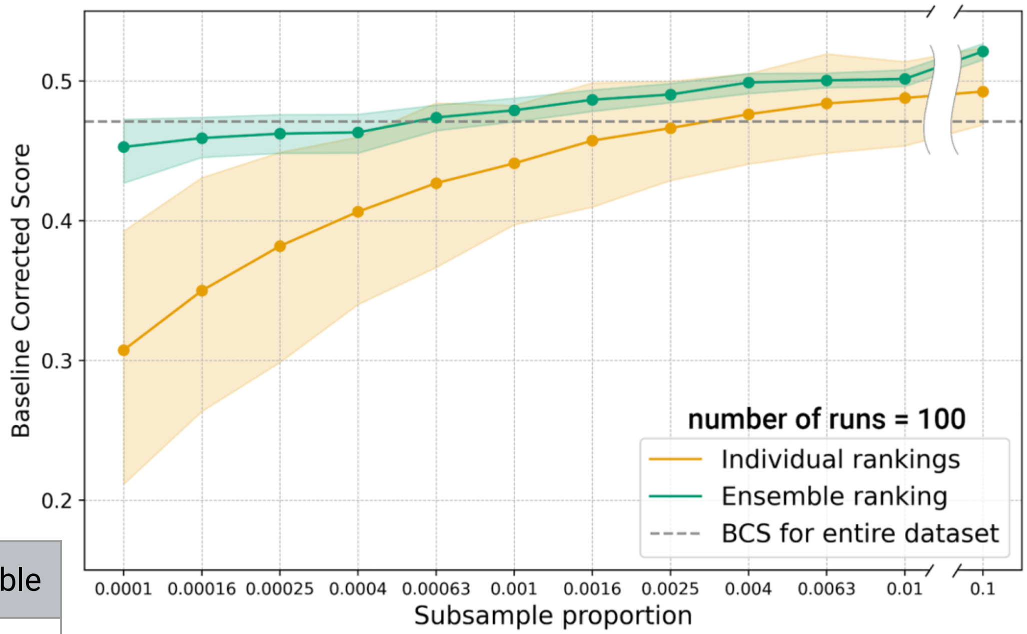





Performance over 100 runs for several algorithms with 1% and 10% samples.
Baseline Corrected Score
0.0
0.1
0.2
0.3
0.4
0.5
$10^{-2}$
$10^{-1}$
$10^{0}$
$10^{1}$
Execution time [seconds]
ANOVA
Chi-square
mRMR
Mutual Info
Pearson Corr.
Random Forest
XGBoost
○ subsampling 0.1
× subsampling 0.01

# RF performance & ensemble

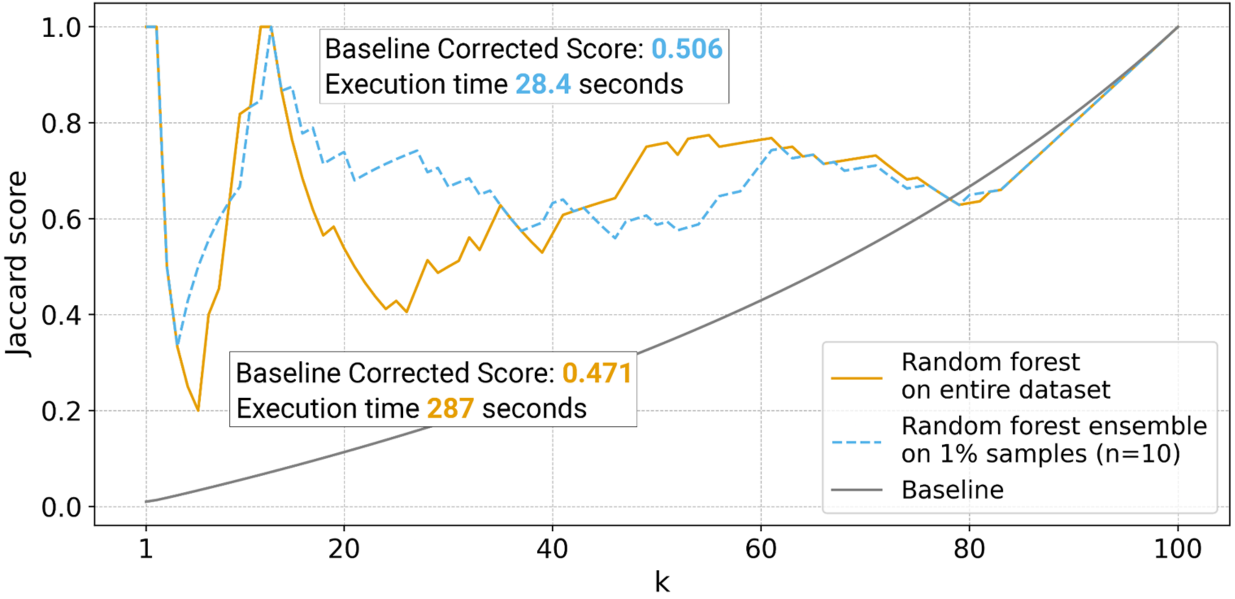
- Subsampling optimization
- Ensemble approach
- Borda count voting system
- Performance & uncertainty



| Feature | Run 1 | Run 2 | Run 3 | Borda | Ensemble |
|---|---|---|---|---|---|
| A | 1 | 2 | 1 | 4 | 1 |
| B | 2 | 1 | 3 | 6 | 2 |
| C | 3 | 3 | 2 | 8 | 3 |

# Conclusions

- Better scores
- Faster execution
- Less data
- Smaller uncertainty


Baseline Corrected Score: 0.506
Execution time 28.4 seconds
Baseline Corrected Score: 0.471
Execution time 287 seconds
Jaccard score
k
Random forest on entire dataset
Random forest ensemble on 1% samples (n=10)
Baseline