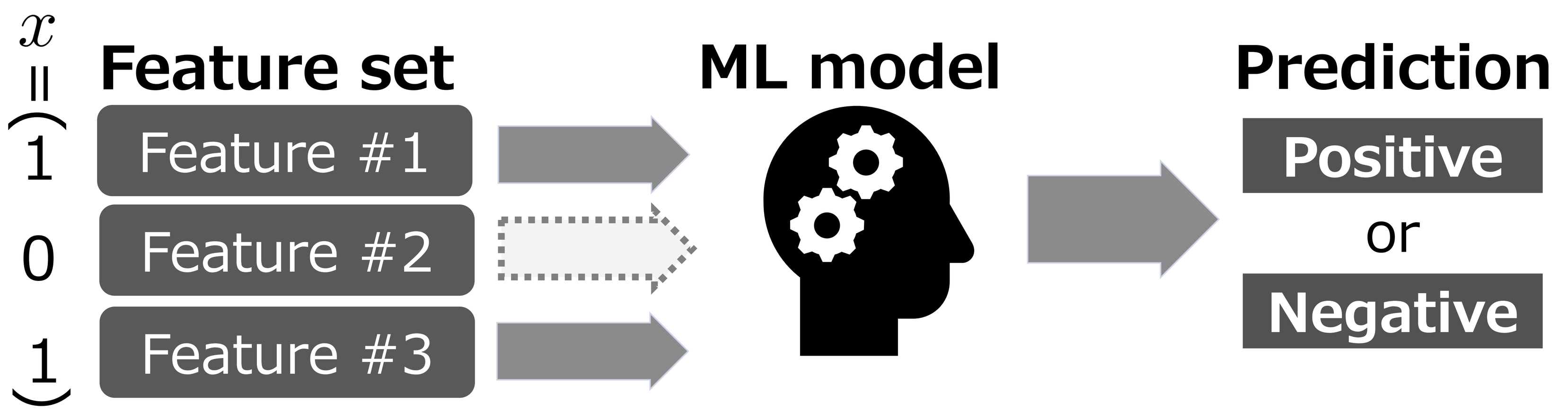
# Contrasting the Landscapes of Feature Selection under Different Machine Learning Models

Arnaud Liefooghe[1], Ryoji Tanabe[2], and Sébastien Verel[1]

1: University of the Littoral Opal Coast, 2: Yokohama National University

## 1. Background

- **Evolutionary wrapper feature selection is a hot topic**
  - Inputs: a feature set of size $n$ and an ML model $M$
  - Output: a subset that maximizes the performance of $M$
  - A binary vector $x \in \{0,1\}^n$ represents a feature subset




- **Disadvantage:** Evaluating $f(x)$ is computationally expensive
  - Each $f$ call requires training an ML model
- The computationally cheap k-nearest neighbors classification (kNN) is used in most previous studies

## 2. Motivation & Contribution

**RQ: How does the choice of an ML model influence the search difficulty of feature selection?**

- **The landscape of feature selection is poorly understood**
  - Only two works by Mostert et al. addressed this topic
  - But, both of their analyses focus only on kNN
- **Is the choice of an ML model influential?**
  - No: Everything is fine! All we need is kNN!
  - Yes: Existing algorithms may overfit feature selection using kNN. Benchmarking should be revisited

## 3. Experimental setup

**6 ML algorithms**

| | |
|---|---|
| kNN | k-near. neigh. classif. |
| SVC | Support vector classif. |
| LR | logistic regression |
| DT | decision tree |
| RF | random forests |
| NB | naive Bayes |

- All $2^n$ subsets were enumerated
- $x = (0, ..., 0)$ is removed
- A 5-fold cross-validation
- 10 runs of DT and RF to minimize the effect of randomness
- Classif. accuracy is used as $f(x)$
- scikit-learn implementation

**14 classification datasets**

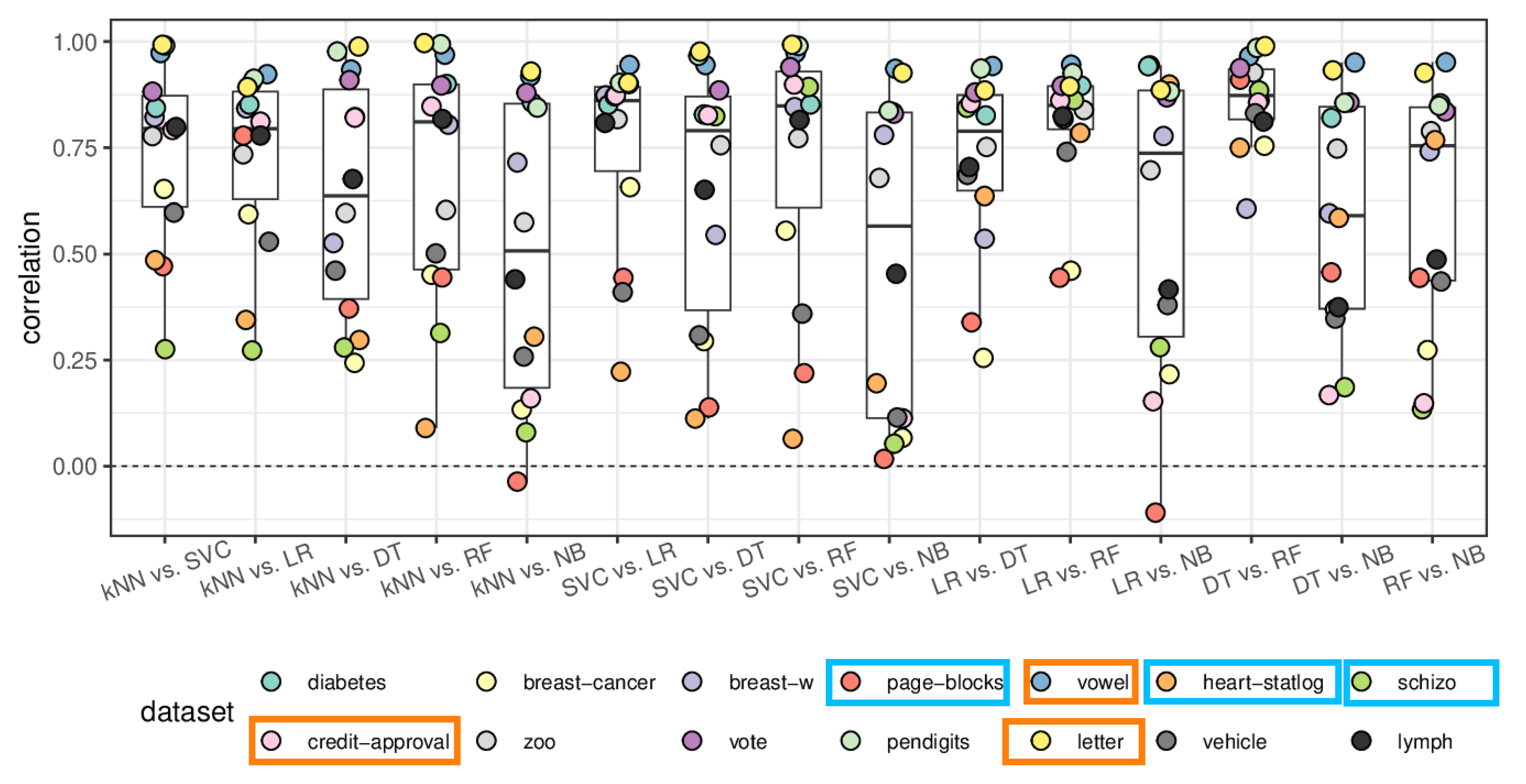
| dataset | class. | data | Feats. |
|---|---|---|---|
| diabetes | 2 | 768 | 8 |
| breast-cancer | 2 | 286 | 9 |
| breast-w | 2 | 699 | 9 |
| page-blocks | 5 | 5473 | 10 |
| vowel | 11 | 5473 | 10 |
| heart-statlog | 2 | 270 | 13 |
| Schizo | 2 | 340 | 14 |

| dataset | class. | data | Feats. |
|---|---|---|---|
| creditapproval | 2 | 690 | 15 |
| zoo | 7 | 101 | 16 |
| vote | 2 | 435 | 16 |
| pendigits | 10 | 10992 | 16 |
| letter | 26 | 20000 | 16 |
| vehicle | 4 | 846 | 18 |
| lymph | 4 | 148 | 18 |

## 4. Correlation between the relative rankings of all subsets

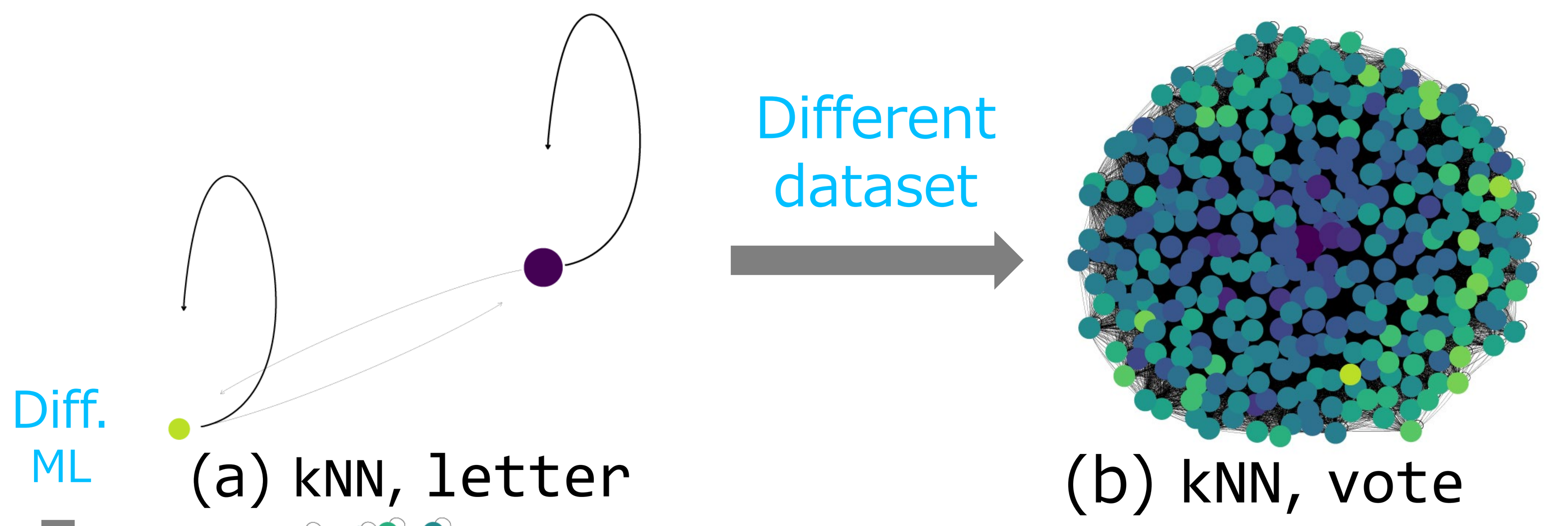
- A point represents a specific dataset
- A boxplot summarizes the distrib. of coefficients
  - for each pair of ML algorithms

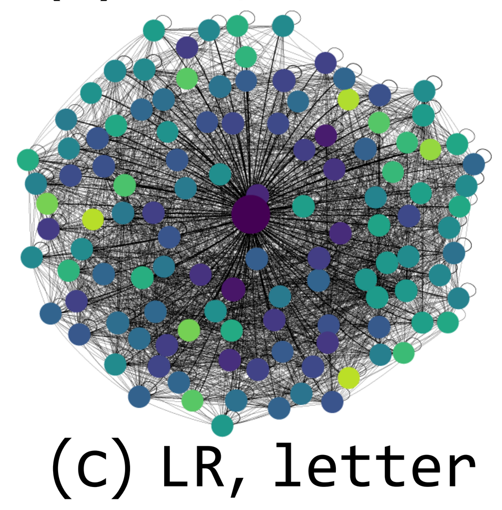
**Small and large corr. are observed in particular datasets**




## 5. Local Optima Networks

**The structure of LONs significantly depends on the combination of ML algorithms and datasets**

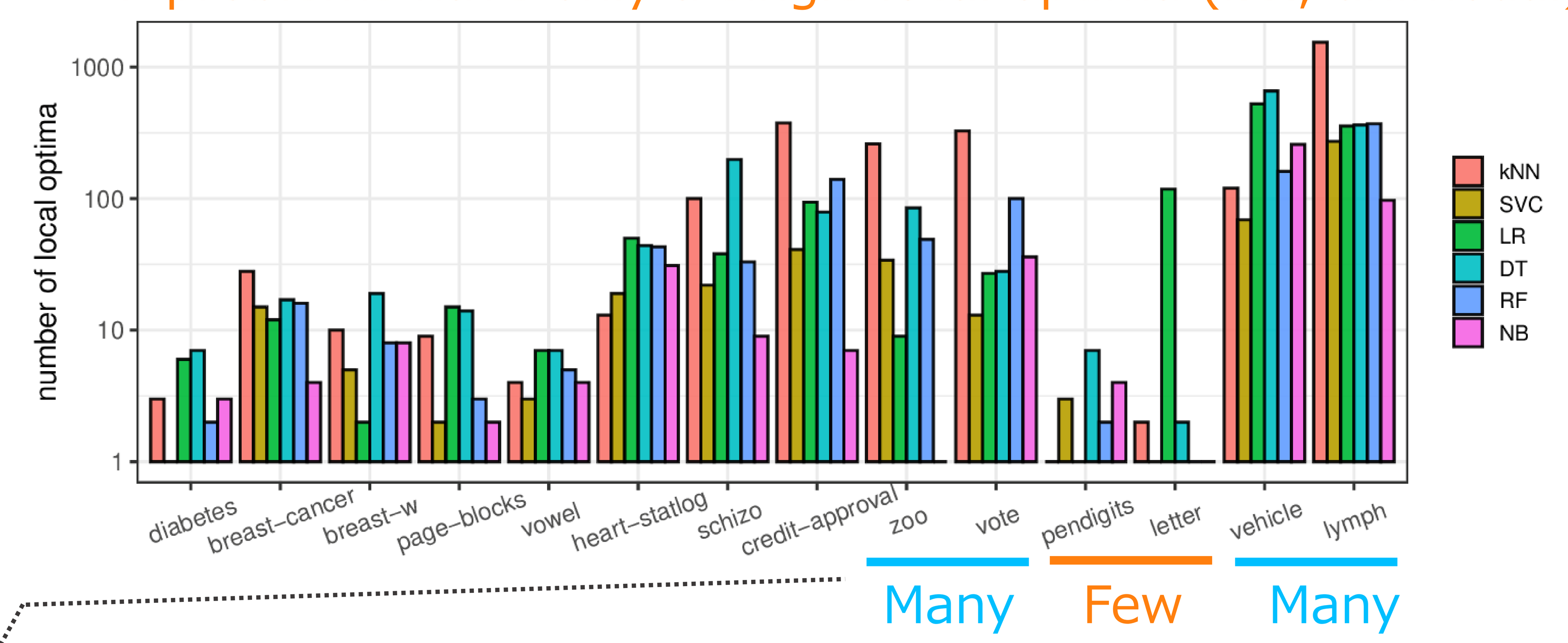



- A node represents a local optima
- A node size represents the size of the basin of attraction
- A node color represents the accuracy $f(x)$
  - A darker color represents a better solution
- An edge width represents the transition probability between two nodes

## 6. Number of local optima

- kNN and LR tend to produce more local optima
- By contrast, NB often produces the fewest local optima
  - This could explain the lower correlation between NB and other ML models (see 4 ↙)
- 7 problems have only a single local optima (i.e., unimodal)

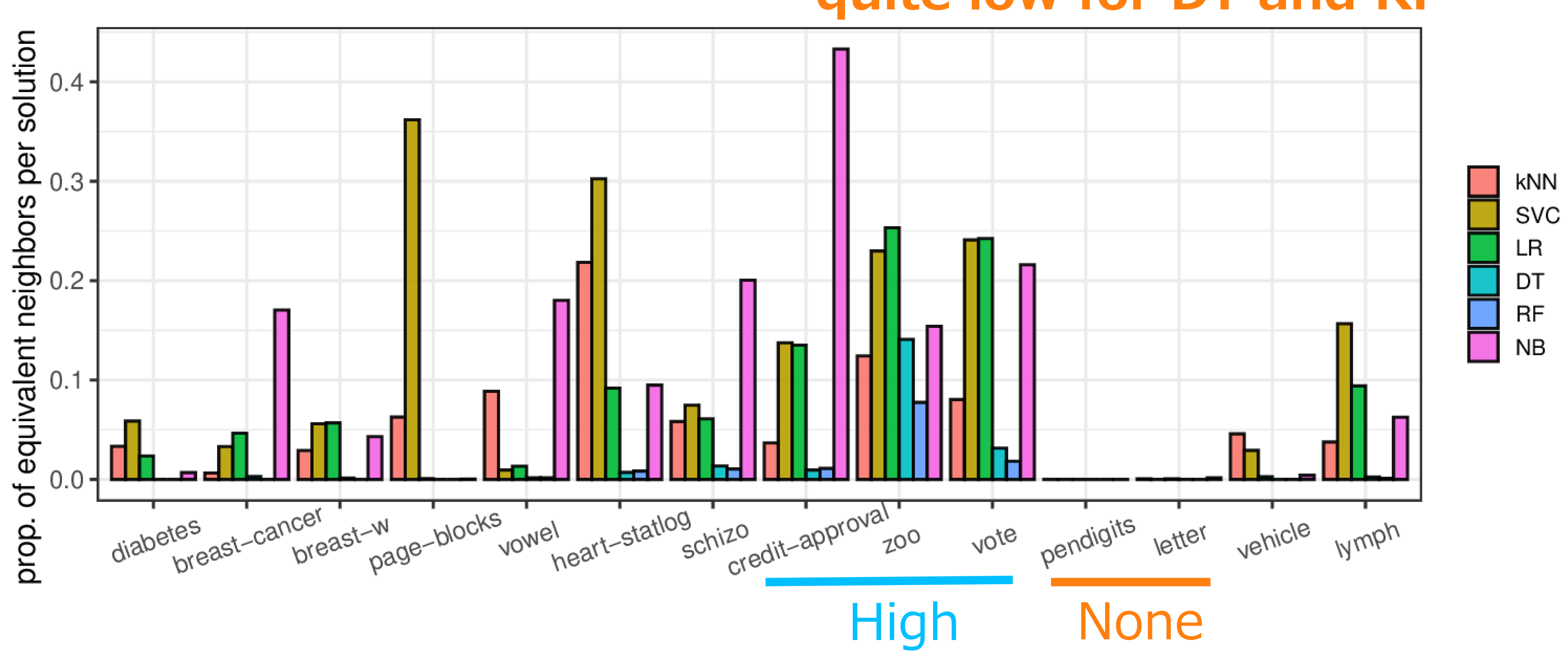



- The two datasets with the largest number of observations produce a small number of local optima, and vice versa
  - Large datasets are seldom used for benchmarking EAs
  - Our findings suggest the importance of considering them

## 7. Neutrality

- The average proportion of equivalent solutions in the neighborhood of each solution

**The level of the neutrality is** high for SVC and NB, quite low for DT and RF




## 8. Conclusion

- Significant differences across ML models were observed
  - This highlights the need to explore ML models beyond kNN
  - It is better not to use an ML model as a proxy for another
  - We highlight the importance of considering large datasets
- **Results *not* shown in this poster**
  - Analysis by 1. the fitness distribution, 2. n. of global optima, 3. FDC, 4. ruggedness, and 5. basins of attractions
  - Explaining the perf. of SFS and GA by landscape metrics
- **Future work**
  - Using other scores and datasets with a larger n. of features