Regression with Enriched Random Forest

Martin Otava

Non-Clinical Statistics Conference
Brugge, Belgium

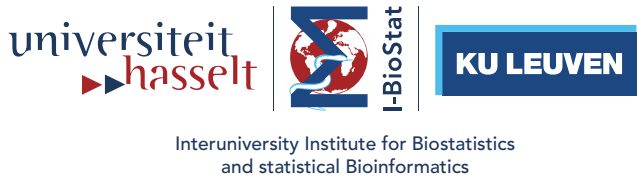
09.10.2014

universiteit hasselt | I-BioStat | KU LEUVEN

Interuniversity Institute for Biostatistics
and statistical Bioinformatics

Research team

- Hasselt University:
  - Martin Otava
  - Ziv Shkedy



- Janssen Pharmaceutica:
  - Dhammika Amaratunga

- Rutgers University:
  - Javier Cabrera

- Dongguk University:
  - Yung-Seop Lee

universiteit
hasselt
I-BioStat
KU LEUVEN
Interuniversity Institute for Biostatistics
and statistical Bioinformatics

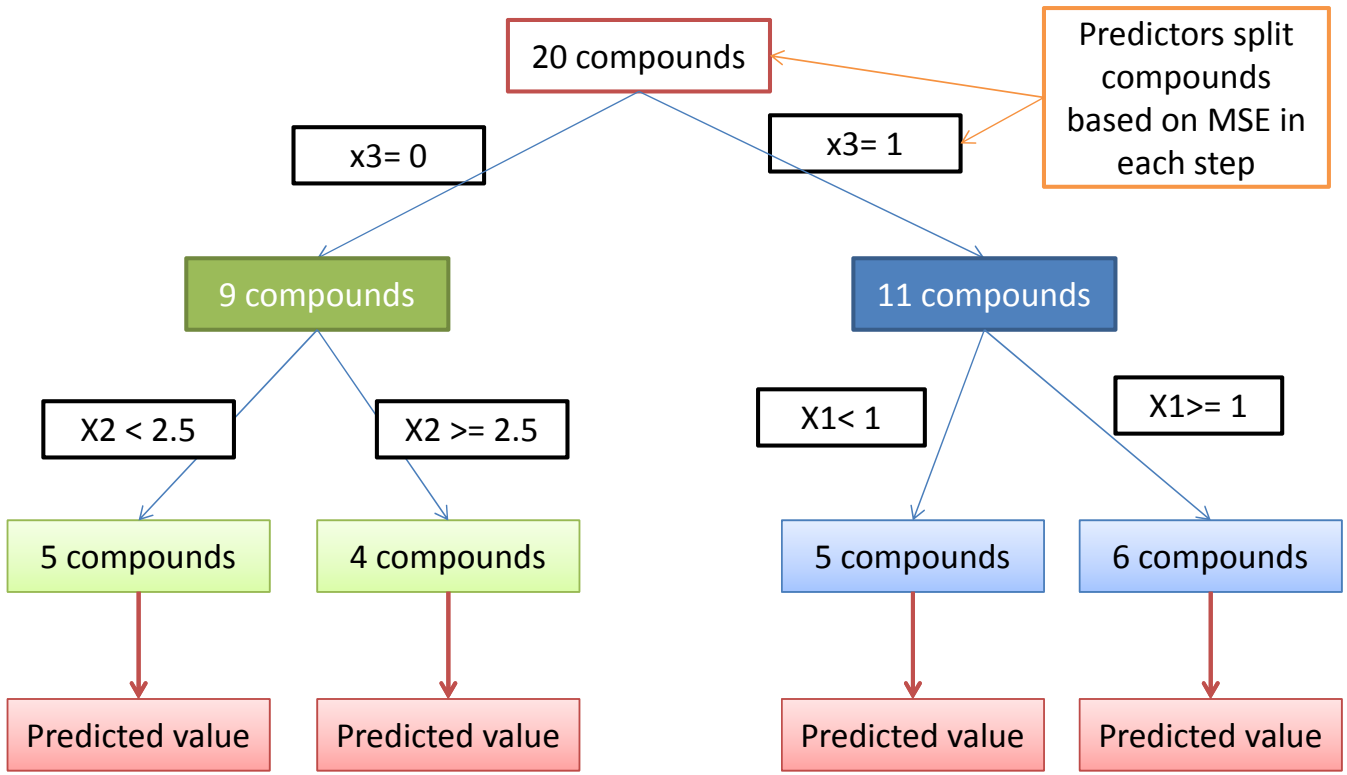
Framework

- Group of compounds as observations
- Response of interest (continuous)
- Large number of predictor variables
- How to select the variables that can predict response?
- Example:
  - Predict value of gene module using chemical representation
  - Predict bioassay value using gene expression

CART

- **C**lassification **A**nd **R**egression **T**rees
- Greedy algorithm
- Search for variables that best separate response values in two group
- Mean squared error is typically used as criterion
- Output is group of variables used as separators

Regression Tree

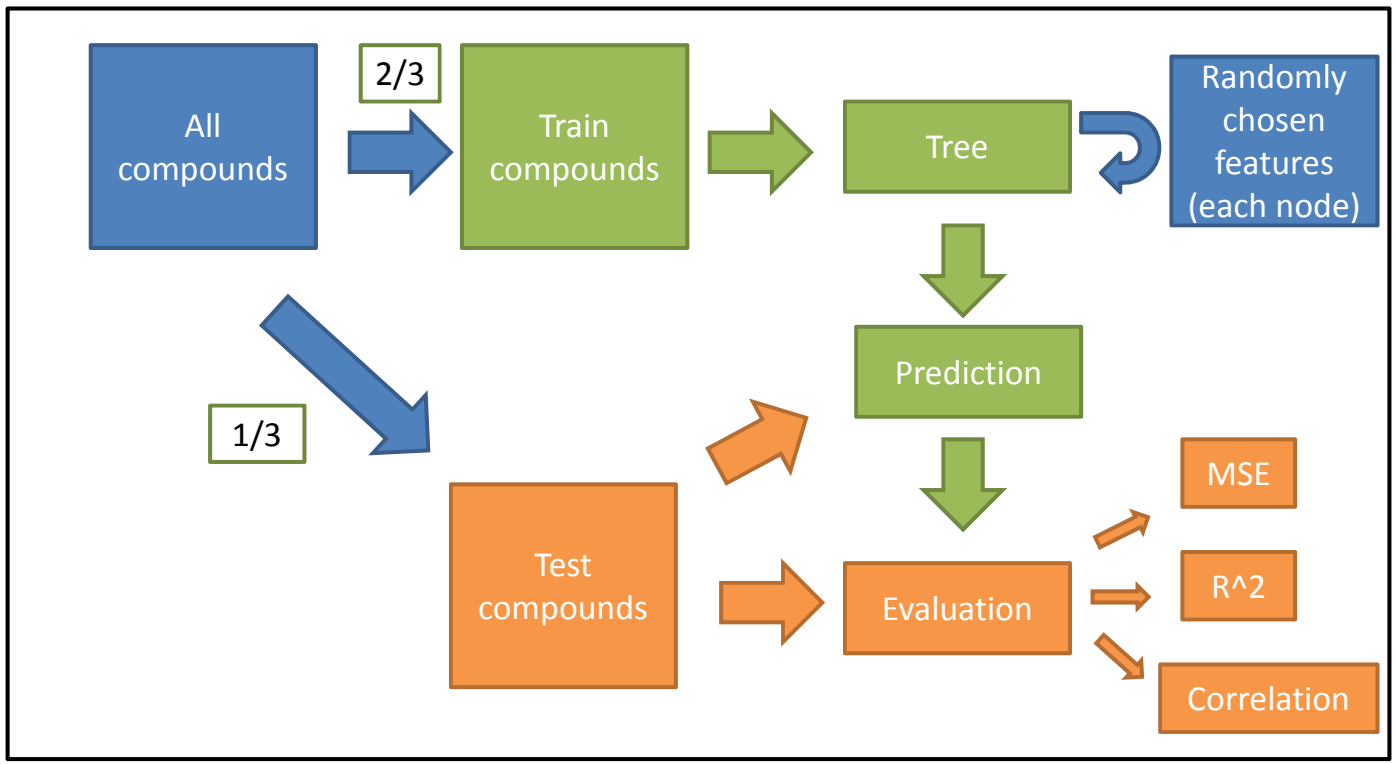


Random forest

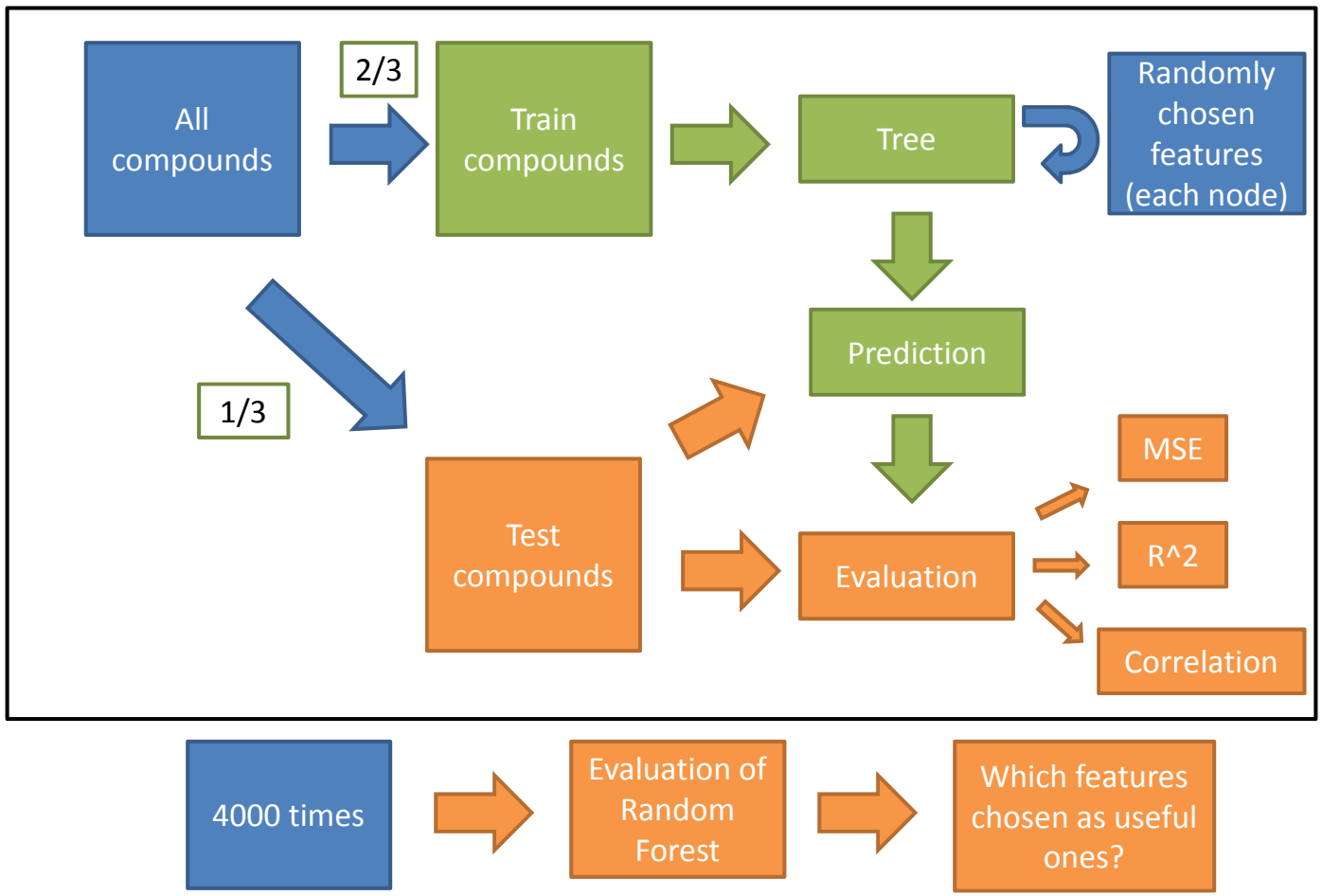
- Issues with CART:
  - Where to cut the tree?
  - Very unstable: small change in observations produce entirely different tree

- Solution: Random forest (RF)
  - Collection of trees
  - Stochasticity added into each of these trees
  - Group of weak predictors collectively predicts well

RF Scheme: single tree



RF Scheme: forest



Example: QSTAR

- 206 fingerprints (FPFs) as predictors
- Gene module as response variable
- Goal: identify FPFs that are related to level of response

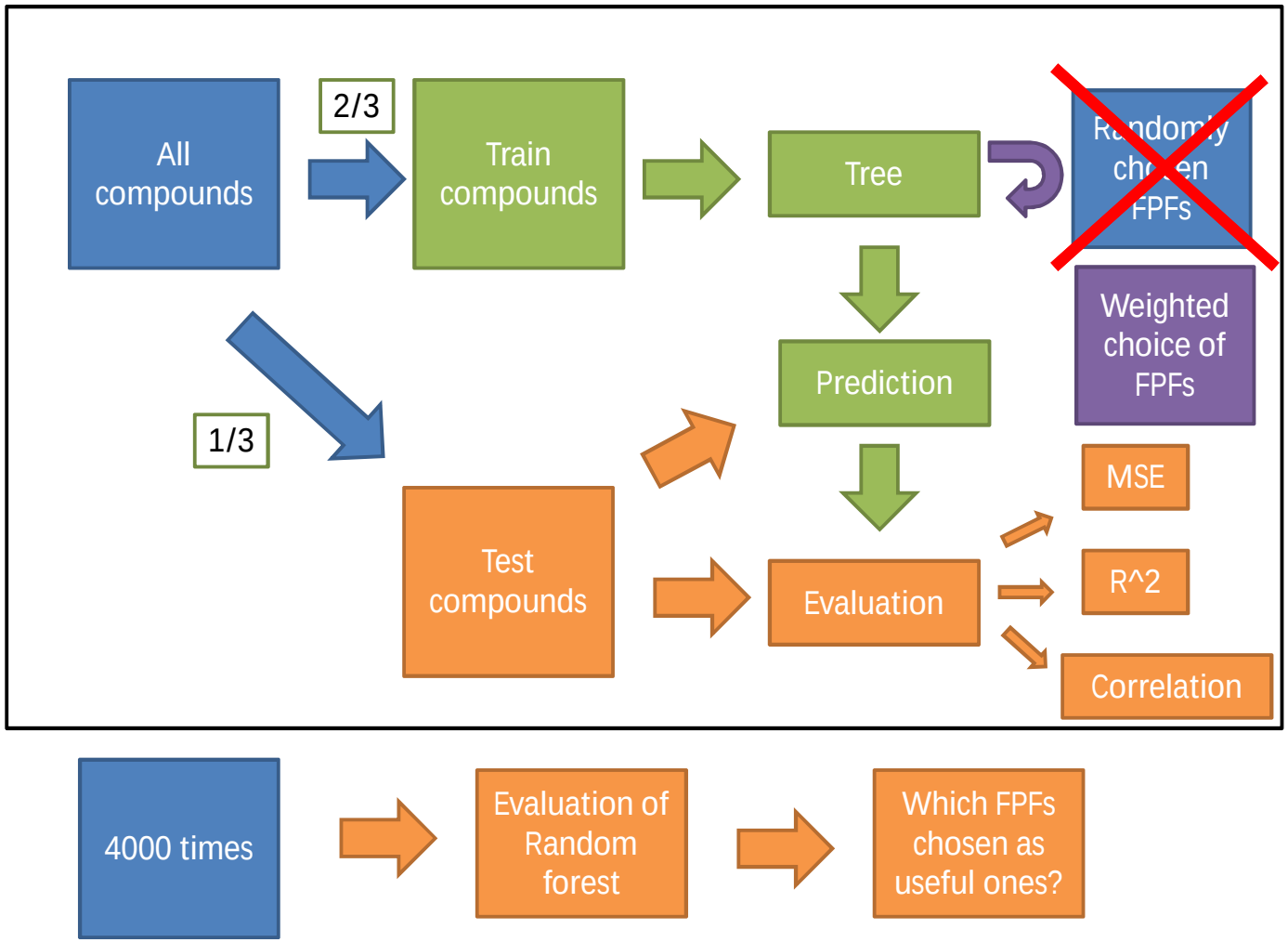
FPF	Selection	MSE Rank	Purity Rank
FPF17	1749 (0.44)	206	206
FPF202	1126 (0.28)	130	126
FPF19	1100 (0.28)	174	161
FPF22	1063 (0.27)	1	174
FPF148	1062 (0.27)	205	204

Enriched random forest

- Issues with RF:
  - Random selection of considered predictors
  - What if most of the predictors are useless? (E.g. $50,000$ vs 50 useful)
  - We are mostly creating trees of no predictive power

- Solution: Enriched Random forest (RF)
  - Replace random sampling with weighted sampling
  - Relate the weights to the relationship with response
  - Which relationship to choose?

ERF Scheme


All compounds
2/3
Train compounds
Tree
Randomly chosen FPFs
Weighted choice of FPFs
Prediction
1/3
Test compounds
Evaluation
MSE
R^2
Correlation
4000 times
Evaluation of Random forest
Which FPFs chosen as useful ones?

ERF: choice of weights

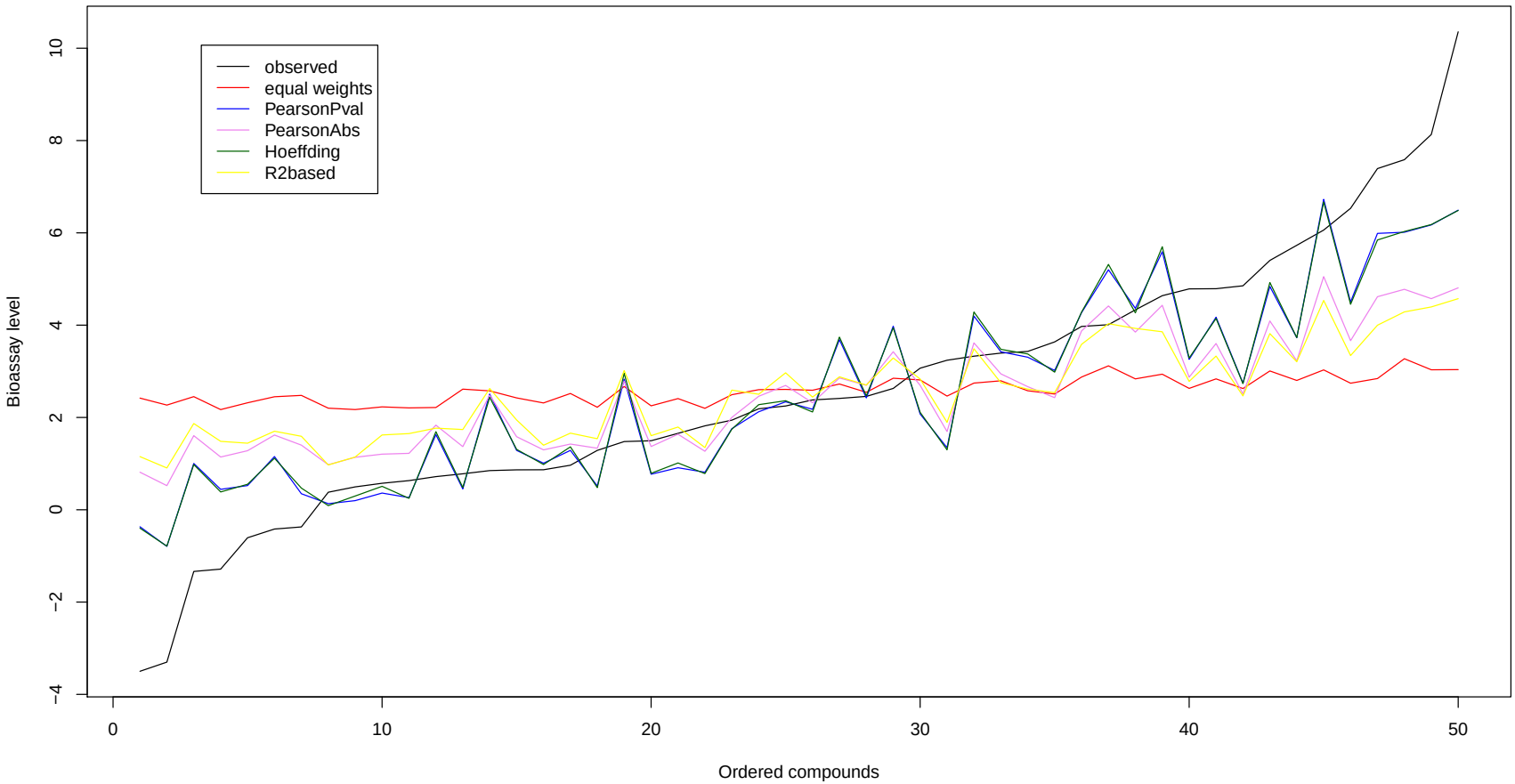
- Represent relationship response-predictor
- To filter out completely unrelated predictors
- In general: lost of information
- based on absolute measure, p-vals or q-vals

- Classification: T statistics, Conditional T
- Regression:
  - Correlation: Pearson, Spearman
  - R^2 of single CART
  - Hoeffding's D
  - many more possibilities

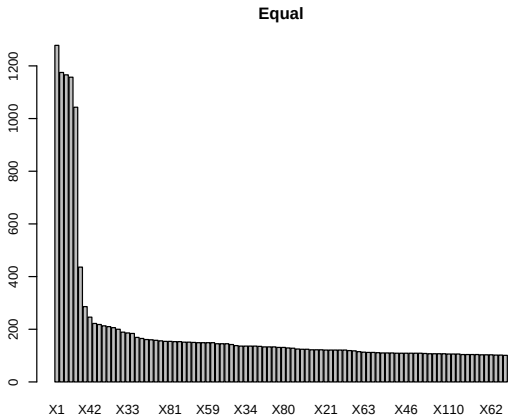
ERF Example: Toy data set

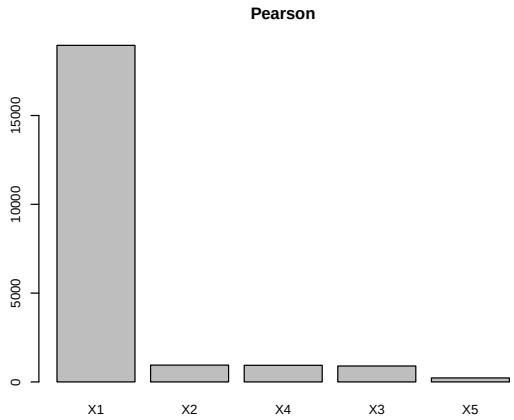
- 1000 predictors
- Only first 5 related to response
- 50 observations
- ERF with different weights
- 4000 trees used

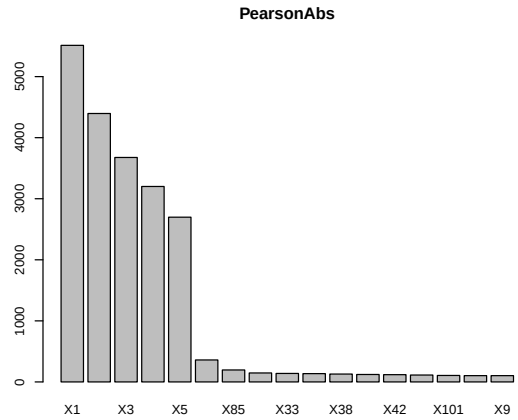
ERF Example: Toy data set

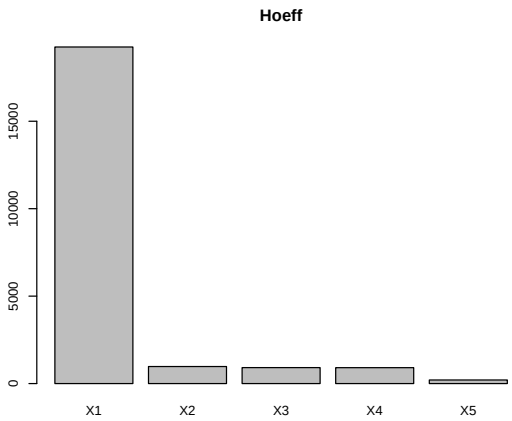


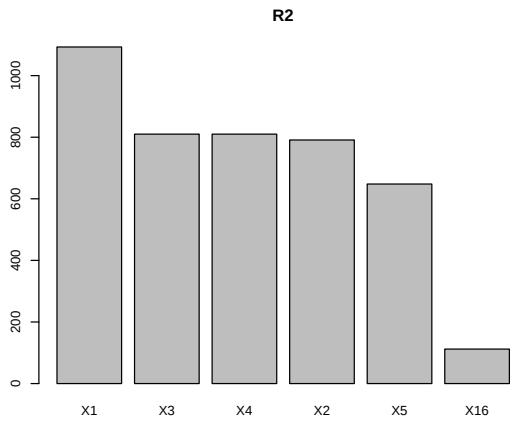
ERF Example: Toy data set











Conclusion

- CART:
  - Unstable, weak predictor
  - Not really good idea
- RF:
  - Great prediction
  - Higher computational time
  - Fail if most predictors useless
- ERF:
  - Good prediction
  - Works in any setting
  - Necessity of setting the weights

Thank you for your attention!