Machine Learning: Pitfalls and Promise

Kristin K. Nicodemus, Ph.D., M.P.H.
Senior Lecturer and Chancellor's Fellow
University of Edinburgh

MRC Medical Research Council

THE UNIVERSITY OF EDINBURGH

CANCER RESEARCH UK

Machine Learning

- Size of datasets increasing exponentially
- In genome-wide association studies, assume underlying biologic model = network or gene set
- Current approaches consider single markers or additive combinations

Machine Learning

- Brute-force all-possible interaction model search not computationally tractable
  - Members of Epistasis Working Group have software to perform exhaustive 2-way interaction search
- Use machine learning to detect sets of SNPs influencing case status
- Anything to consider before applying ML?

Machine Learning

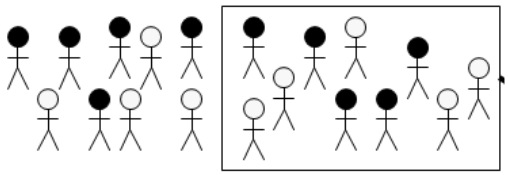
- ML literature is full of statements
  - "X method does not overfit"
  - "Y method is a good 'off-the-shelf' classifier"
- Often no proof (or limited proof) is given

Machine Learning

- Three 'flavours' of ML
  - Ensemble-based: Random Forest, Boosting
  - Regression-based: LASSO, elastic net, Monte Carlo Logic Regression
  - Kernel-based: Support Vector Machines
- Focus on RF methodology issues – happy to focus on other two flavours in upcoming talks

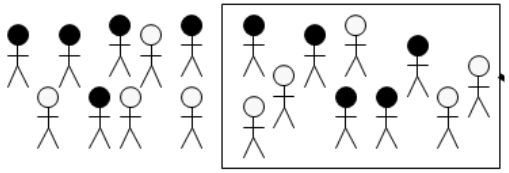
Machine Learning: RF

Step 1: select subsample of cases/controls,
set aside additional samples

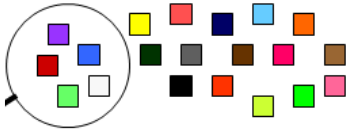


Machine Learning: RF

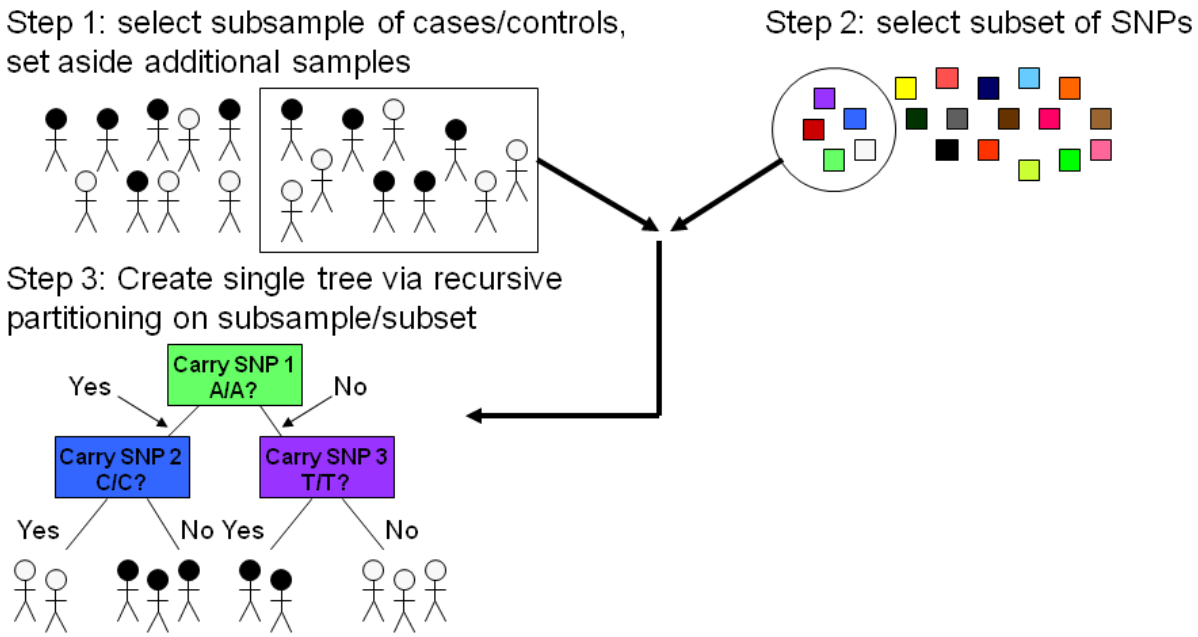
Step 1: select subsample of cases/controls, set aside additional samples



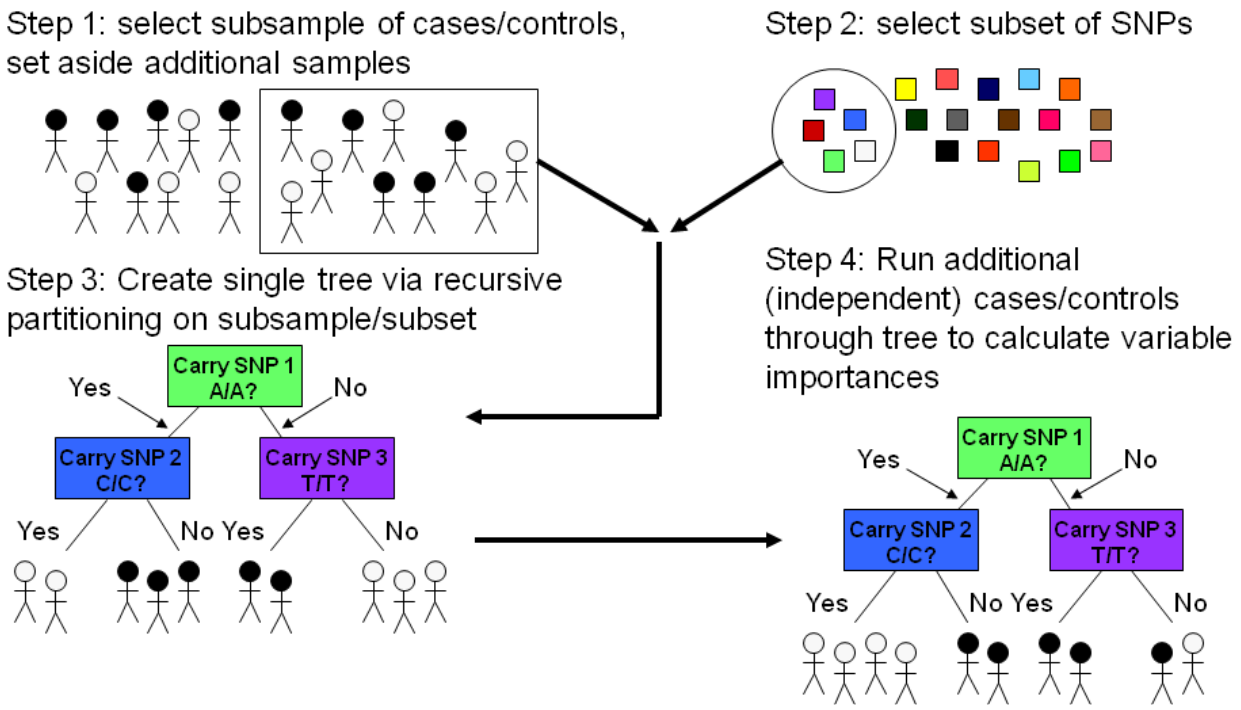
Step 2: select subset of SNPs



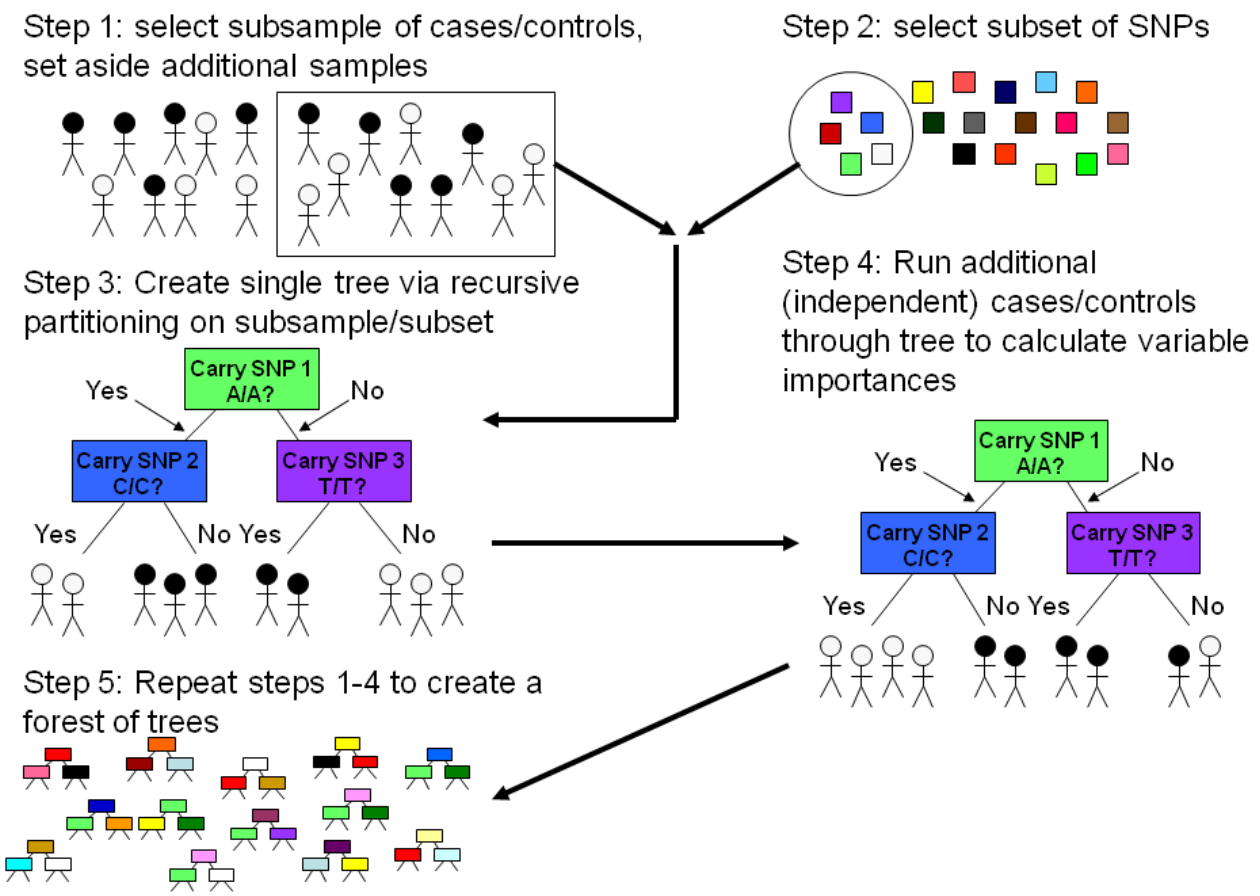
Machine Learning: RF



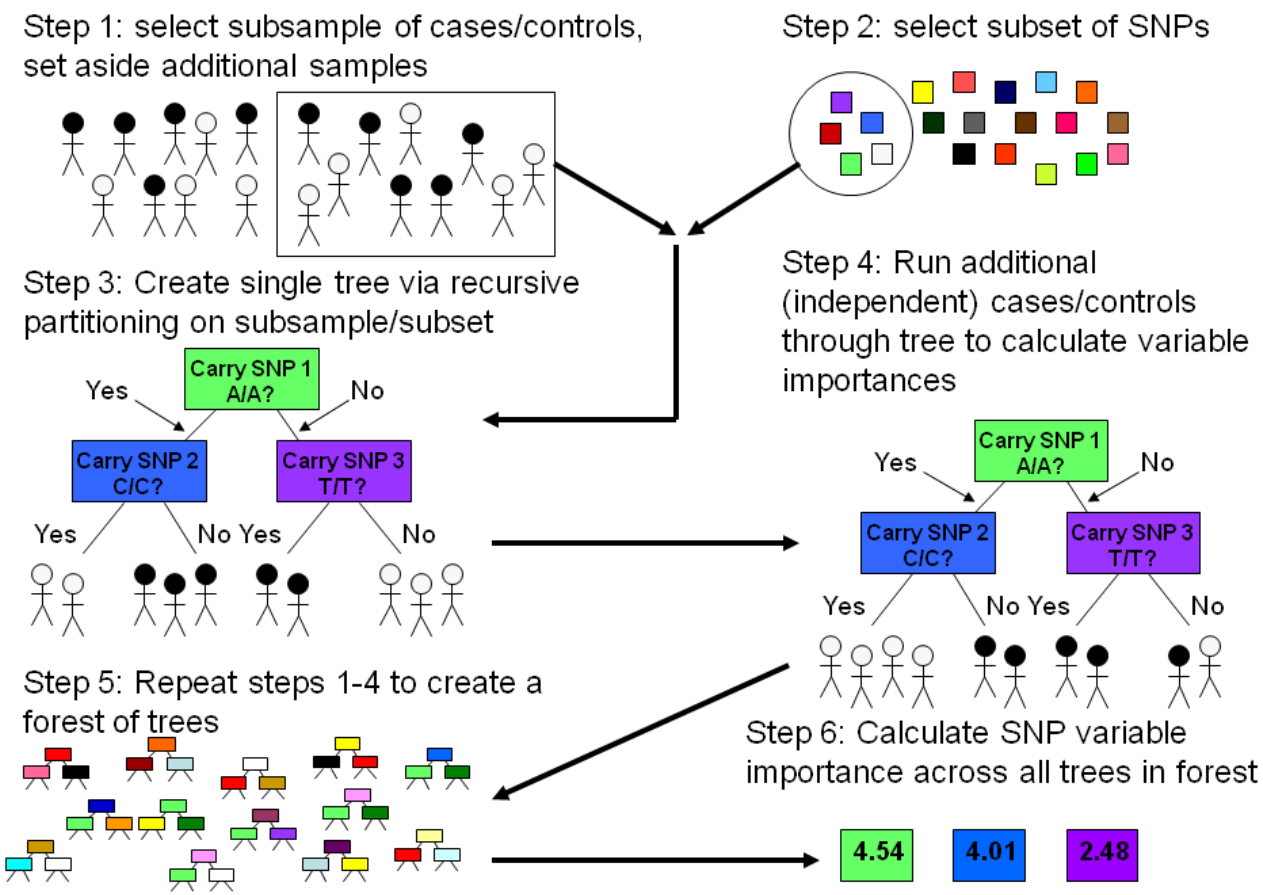
Machine Learning: RF



Machine Learning: RF

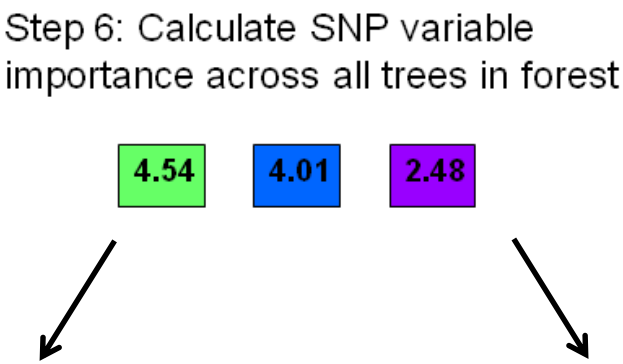


Machine Learning: RF



Machine Learning

- VIMs are the main way to understand RF



Gini index

Average across all trees in the forest of decrease in node impurity for each predictor

Based on **in-bag** samples

Permutation-based

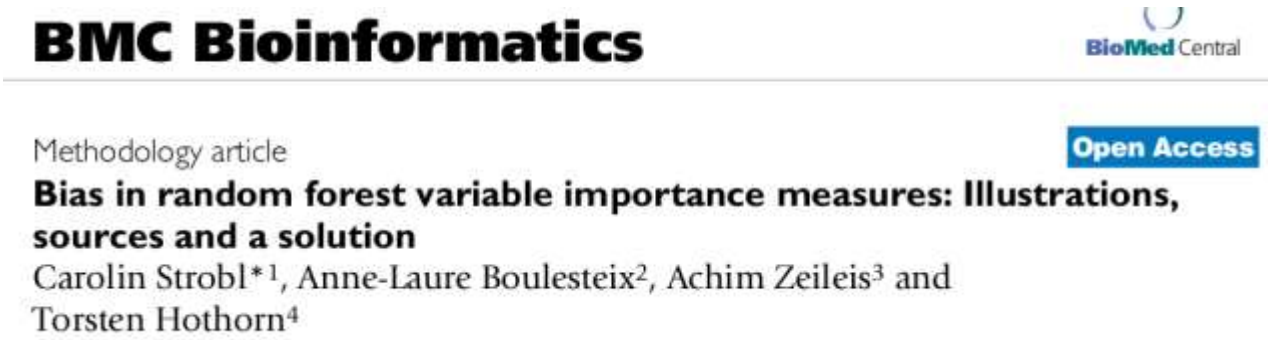
Average across all trees in the forest of the difference between the prediction accuracy using the observed data and permuted. May be scaled by SE.
Based on **out-of-bag** samples

Machine Learning

- Are there conditions where the RF VIM is misleading or the algorithm fails?
  - Different scales of measurement
  - Correlation between predictors
  - Different category frequencies
  - Interaction effects and new VIMs – do they fare better?

Machine Learning

- Study examined random forest vs. conditional inference forest: different splitting criteria



BMC Bioinformatics

BioMed Central

Methodology article

Open Access

Bias in random forest variable importance measures: Illustrations, sources and a solution

Carolin Strobl*[1], Anne-Laure Boulesteix[2], Achim Zeileis[3] and Torsten Hothorn[4]

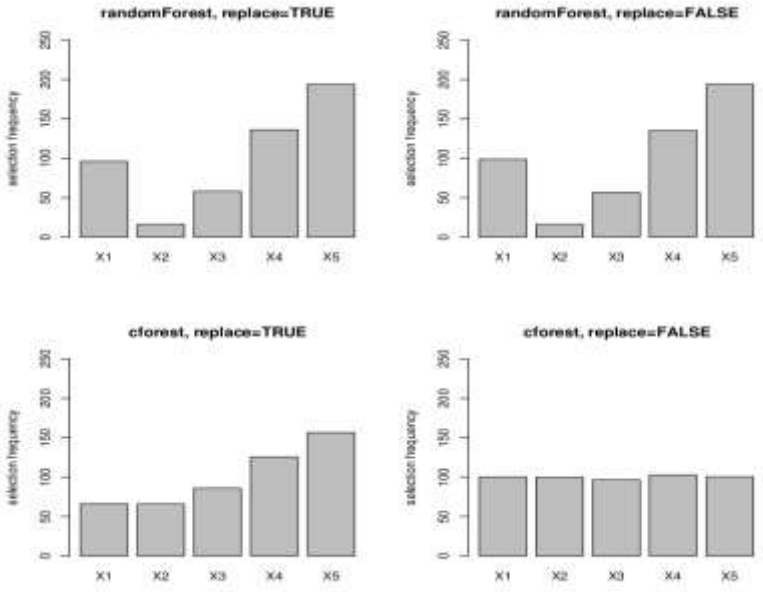
Machine Learning

- Predictors with different scales of measurement (use of bootstrapping vs. subsampling): Under H_0, selection frequencies

Table 1: Simulation design for the simulation studies – predictor variables

Predictor variables	
X_1	~$N(0, 1)$
X_2	~$M(2)$
X_3	~$M(4)$
X_4	~$M(10)$
X_5	~$M(20)$

The predictor variables are sampled independently from the following distributions. $N(0, 1)$ stands for the standard normal distribution, $M(k)$ stands for the multinomial distribution with values in $\{0,..., k - 1\}$ and equal probabilities (discrete uniform distribution on $\{0,..., k - 1\}$), $B(p)$ stands for the binomial distribution (Bernoulli distribution) with probability p, thus $M(2)$ equals $B(0.5)$.



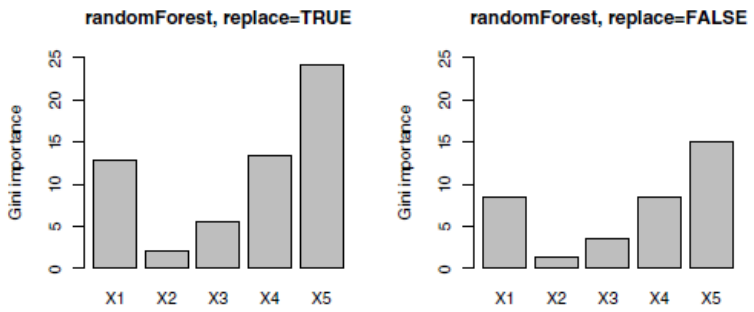
Machine Learning

- Predictors with different scales of measurement (use of bootstrapping vs. subsampling): Under H_0, Gini VIM – also leads to low power in alternative

Table 1: Simulation design for the simulation studies – predictor variables

Predictor variables	
X_1	~$N(0, 1)$
X_2	~$M(2)$
X_3	~$M(4)$
X_4	~$M(10)$
X_5	~$M(20)$

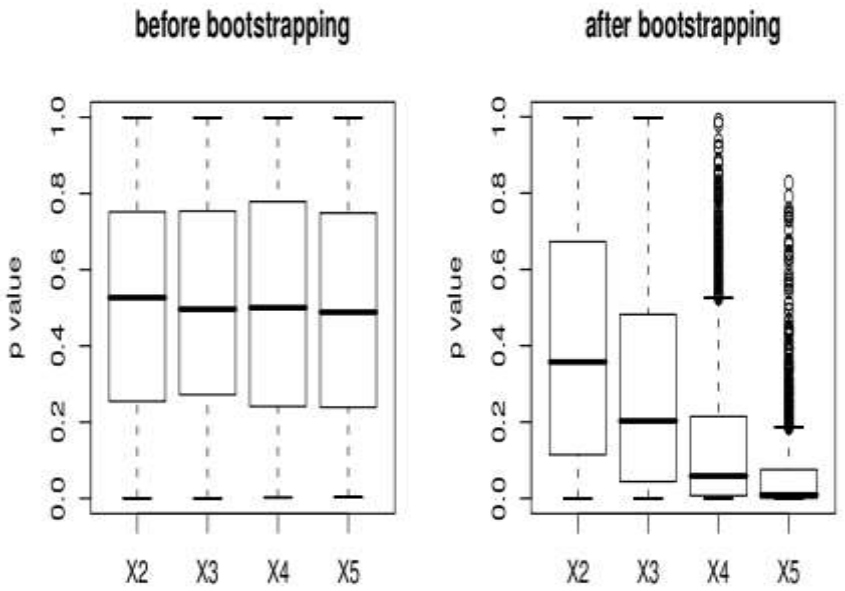
The predictor variables are sampled independently from the following distributions. $N(0, 1)$ stands for the standard normal distribution, $M(k)$ stands for the multinomial distribution with values in $\{0,..., k - 1\}$ and equal probabilities (discrete uniform distribution on $\{0,..., k - 1\}$), $B(p)$ stands for the binomial distribution (Bernoulli distribution) with probability p, thus $M(2)$ equals $B(0.5)$.



Machine Learning

- Predictors with different scales of measurement (use of bootstrapping vs. subsampling): χ^2 tests

more categories. The reason for the shift in the distribution of the p values displayed in Figure 11 is that each original sample, even if sampled from theoretically independent distributions, may show some minor variations from the null hypothesis of independence. These minor variations are aggravated by bootstrap sampling with replacement, because the cell counts in the contingency table are affected by observations that are either not included or are doubled or tripled in the bootstrap sample, and therefore the bootsrap sample deviates notably from the null hypothesis – even if the original sample was generated under the null hypothesis.



Machine Learning

- Use of predictors with different scales of measurement leads to increased selection frequencies in RF due to
  - Bootstrapping
  - Multiple testing

- Conditional inference forest with subsampling can improve performance of the Gini index VIM, but permutation VIM is recommended

Machine Learning

- Within-predictor correlation: effect on resulting variable importance measures

BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 15 2009, pages 1884–1890
doi:10.1093/bioinformatics/btp331

Genetics and population analysis

Predictor correlation impacts machine learning algorithms: implications for genomic studies

Kristin K. Nicodemus[1,2,3,*] and James D. Malley[4]

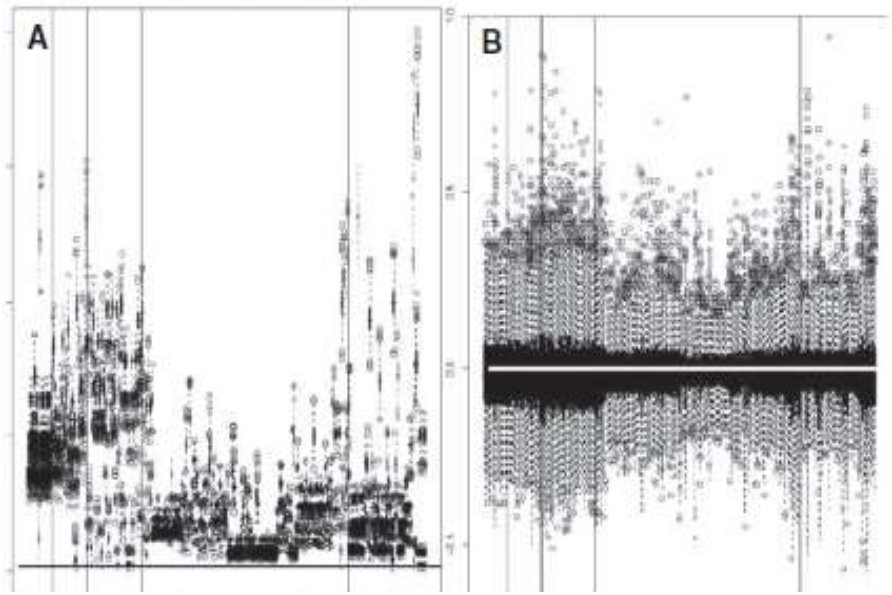
Machine Learning

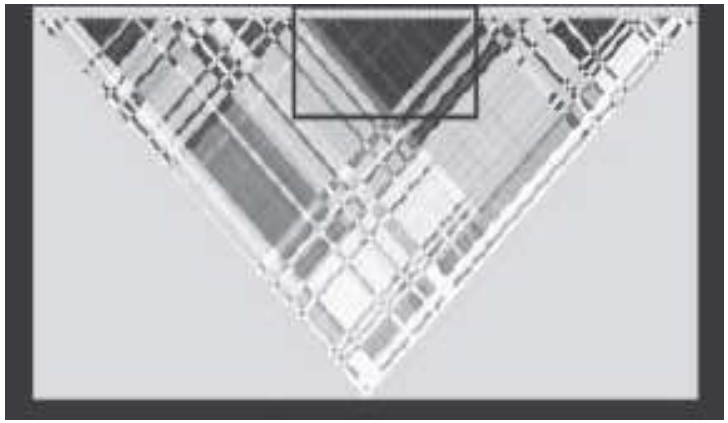
- Simulation study assessed impact of correlation on VIMs under H_0 and H_A

- Genetic: 5 Genes, 199 SNPs, 500 replicates/condition, retained LD structure

Machine Learning

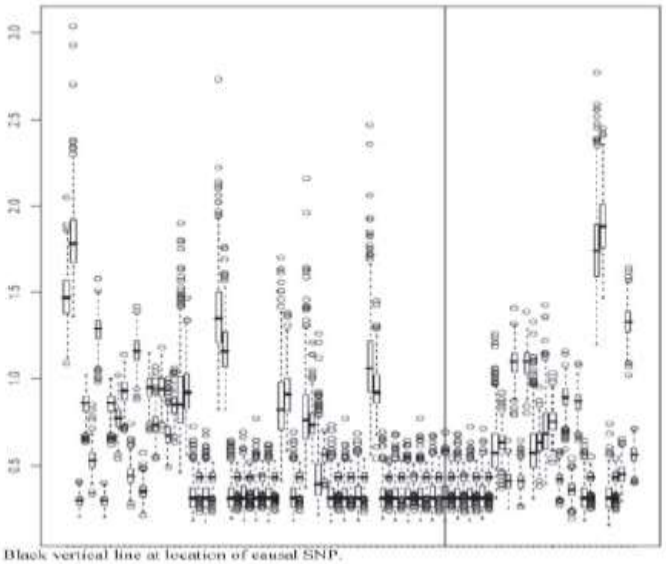
- Within-predictor correlation: effect on resulting variable importance measures: H_0

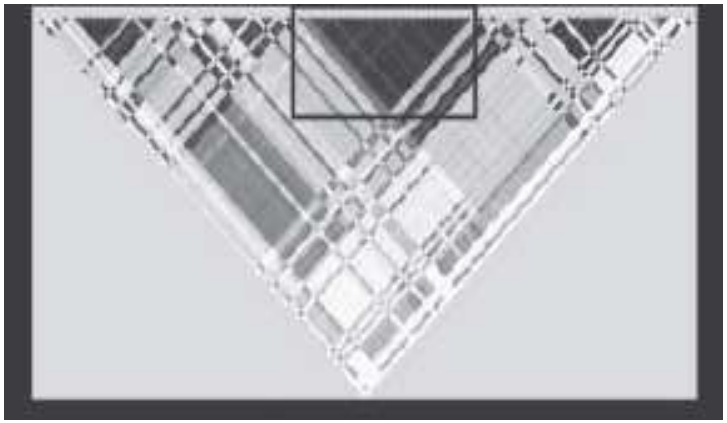




Machine Learning

- Within-predictor correlation: effect on selection frequencies and resulting variable importance measures: H_A (OR = 2.0)





Machine Learning

- Within-predictor correlation leads to spurious results for the Gini VIM under H_0 and H_A

- The same correlation leads to lower variability in calculated VIM for permutation-based measures, but does not lead to spurious results

Machine Learning

- Differing category frequencies within predictors with the same scale of measurement

Briefings in Bioinformatics Advance Access published March 31, 2010

BRIEFINGS IN BIOINFORMATICS, page 1 of 4 doi:10.1093/bib/bbq011

Letter to the Editor: Stability of Random Forest importance measures

M. Luz Calle and Victor Urrea

Submitted: 3rd March 2010; Received (in revised form): 9th March 2010

BRIEFINGS IN BIOINFORMATICS, VOL 12. NO 4. 369–373
Advance Access published on 15 April 2011 doi:10.1093/bib/bbr016

Letter to the Editor: On the stability and ranking of predictors from random forest variable importance measures

Kristin K. Nicodemus

Machine Learning

- Differing category frequencies within predictors with the same scale of measurement

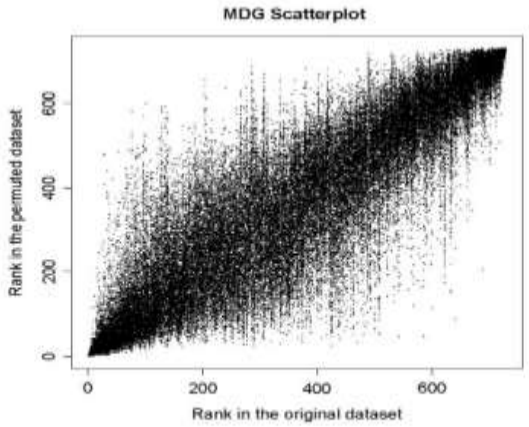


Figure 1: MDG rank in the original dataset against MDG rank in the perturbed datasets (10% left out).

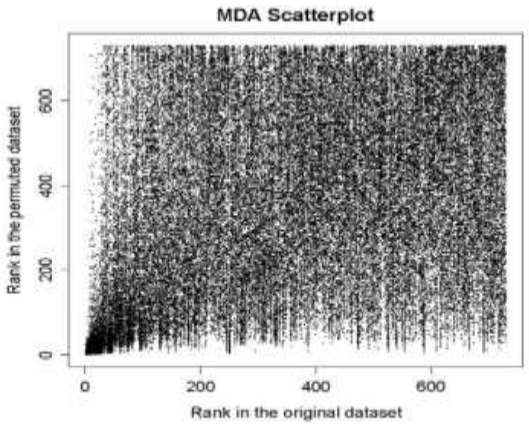


Figure 2: MDA rank in the original dataset against MDA rank in the perturbed datasets (10% left out).

Machine Learning

- Synthetic simulation: 1000 cases and controls, no correlation
- First 5: ORs 3.0, 2.5, 2.0, 1.5, 1.25, minor category frequency of 0.5
- 45 additional variables, no association

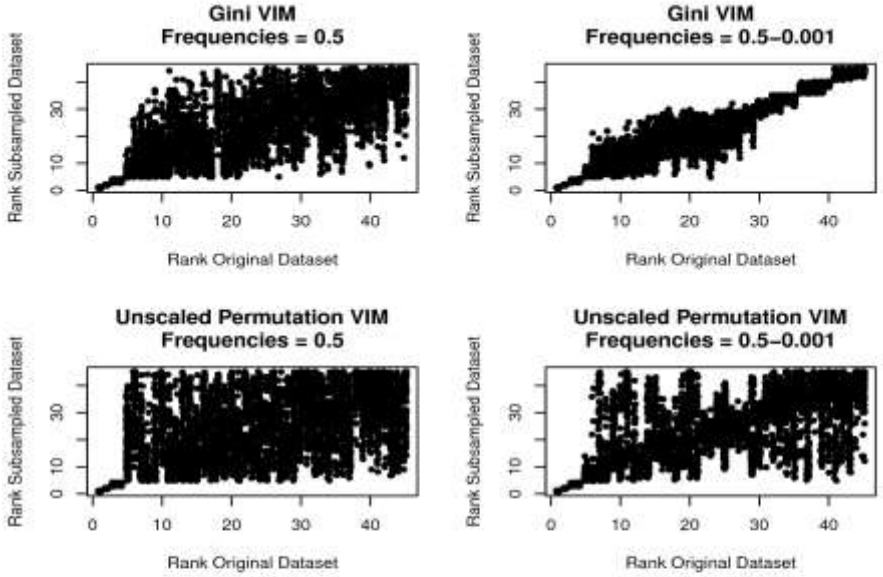
Machine Learning

- Condition 1: all unassociated minor category frequency = 0.5
- Condition 2: varied minor category frequency: 0.5, 0.4, 0.3, 0.2, 0.1, 0.05, 0.01, 0.001

- Same analysis as in Calle et al

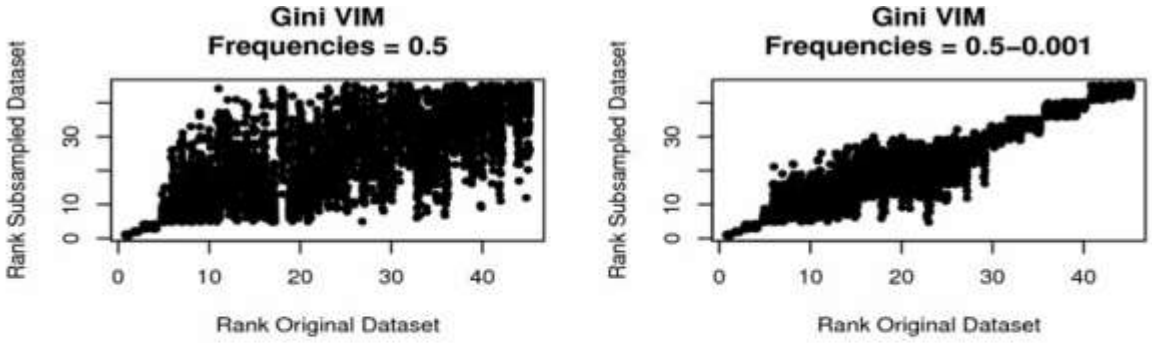
Machine Learning

- Differing category frequencies within predictors with the same scale of measurement



Machine Learning

- Varying minor category frequencies produced a very similar plot to that in Calle et al.



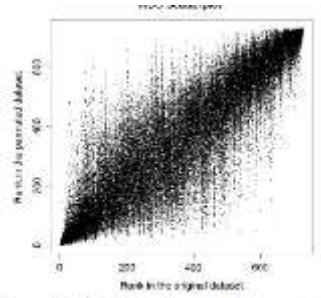


Figure 1: MDG rank in the original dataset against MDG rank in the perturbed datasets (10% left out).

Machine Learning

- Permutation-based VIMs are robust to correlated predictors
- They can be unstable
- Re-run RF using different random number seeds, take average or median VIM

- Gini VIMs can lead to spurious results

- Several new VIMs have been developed – how do they fare?

Machine Learning

- Expanded set of VIMs:
- RF: Standard Gini and permutation, plus scaled permutation VIMs: Breiman and Liaw
- RF: Conditional permutation VIM
- Conditional Inference Forest VIMs: permutation and AUC
- RF: Minimum depth

- Work led by Lara Neira Gonzalez

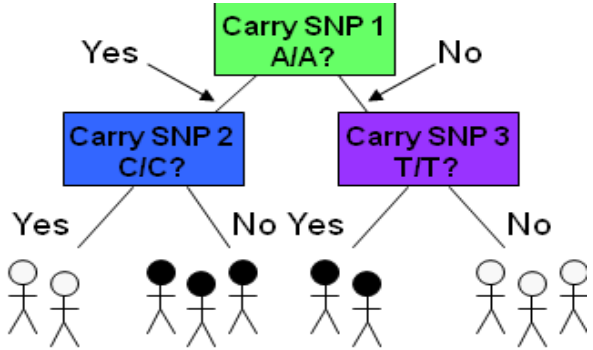
Machine Learning

- Scaled VIMs: permutation VIM scaled by SE
- CIF VIM: same as RF permutation VIM
- AUC VIM: calculates the AUC in out-of-bag samples before and after permutation

Machine Learning

- Conditional VIM: permutes variable within strata of other correlated predictors

- Minimal depth VIM: average across forest of the minimal depth of predictor in trees



Machine Learning

- Simulation conditions: H_0 and H_A, 500 replicates, 100 variables, continuous outcome/predictors

- Varied N correlated predictors (5, 20, 40) and strength of correlation (0.1, 0.4, 0.8)

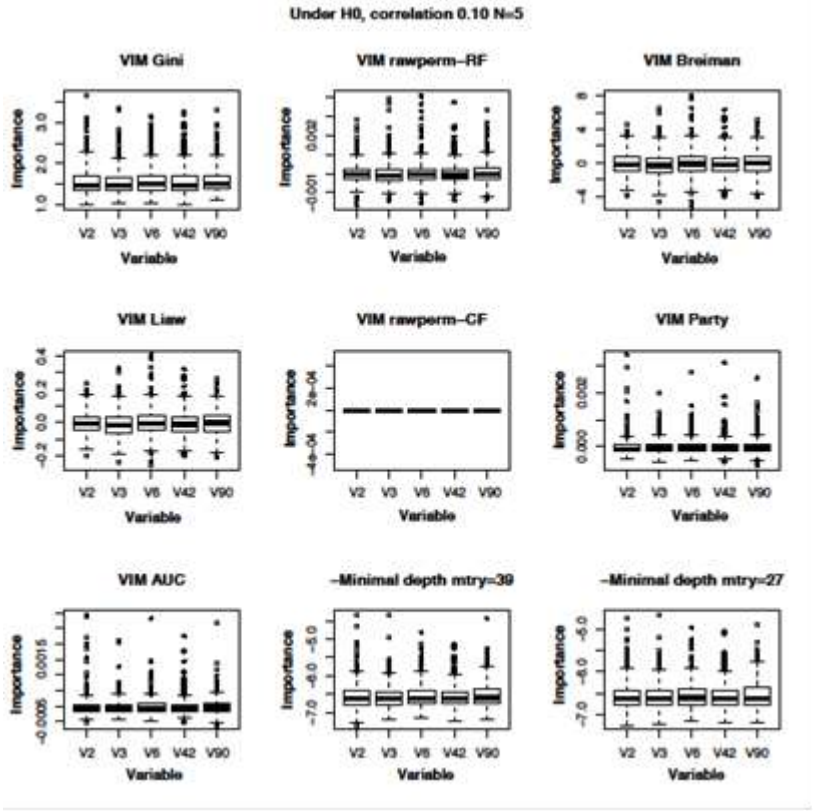
- Mtry for minimal depth varied as previous studies showed correlation required smaller values of mtry

Machine Learning

- Weak association/interaction under H_A: One correlated predictor and one uncorrelated predictor participate in interaction

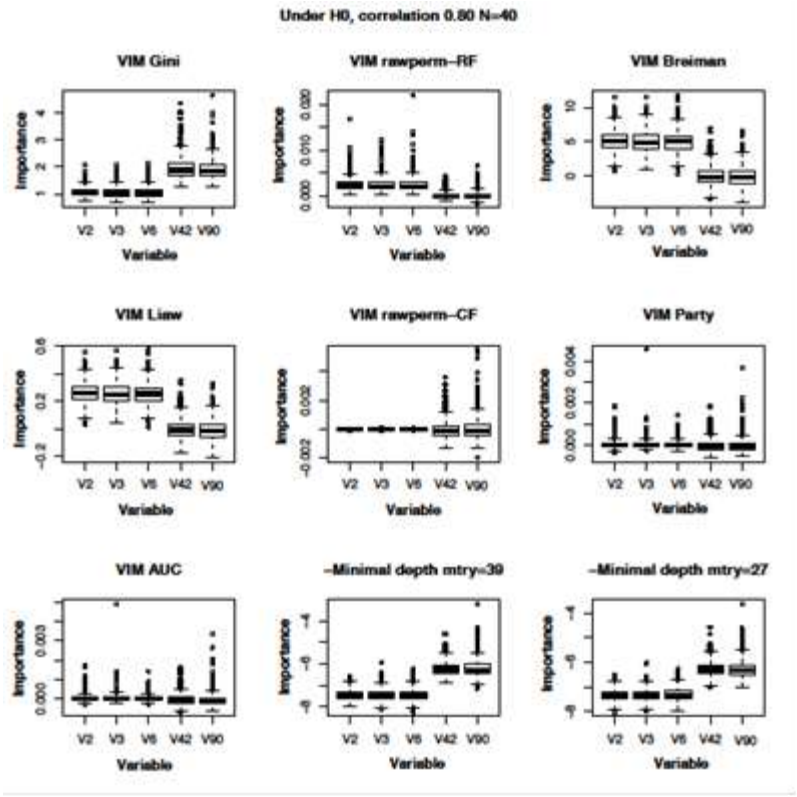
Machine Learning

- H_0, N correlated = 5, correlation = 0.1



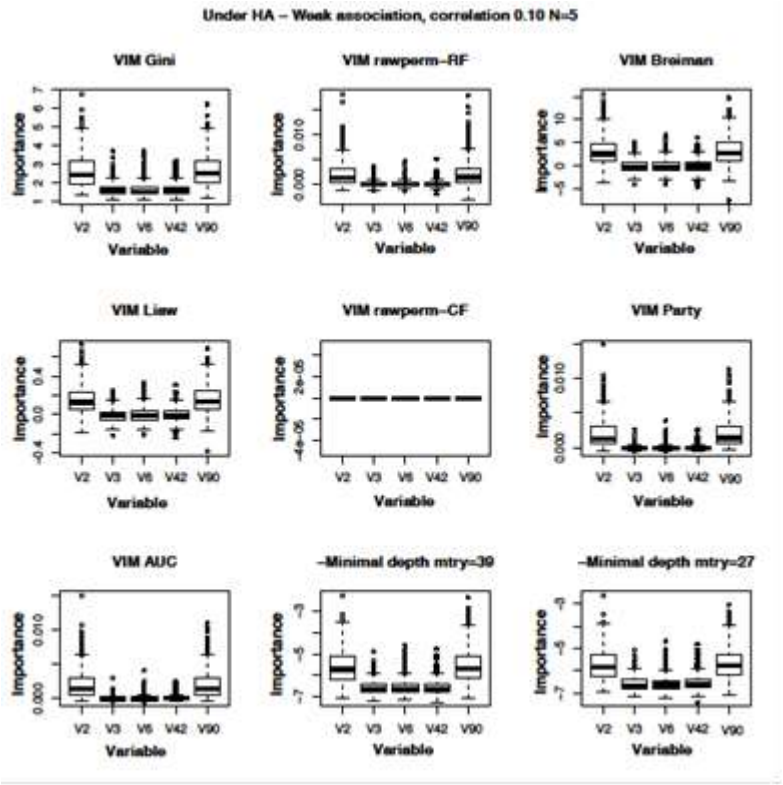
Machine Learning

- H_0, N correlated = 40, correlation = 0.8



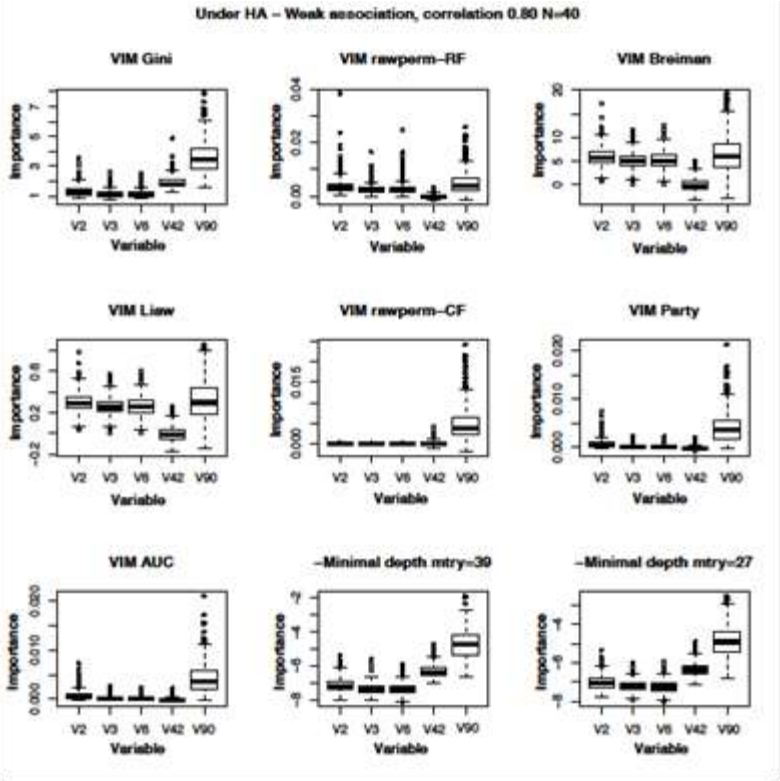
Machine Learning

- H_A, N correlated = 5, correlation = 0.1



Machine Learning

- H_A, N correlated = 40, correlation = 0.8



Machine Learning

- Under H_0 with small N correlated/low correlation, all VIMs except conditional performed adequately

- Under H_0 with large N correlated/high correlation, Gini, scaled permutation, conditional and minimal depth produced spurious results

Machine Learning

- Under H_A with small N correlated/low correlation, all VIMs were able to detect the true signal except conditional

- Under H_A with large N correlated/high correlation, all methods could detect the uncorrelated interacting predictor

- The ability of all methods to detect the correlated interacting predictor varied

Machine Learning

- The work on RF is fairly extensive: most MLs require further study
- One of the goals of the Epistasis Working Group is to perform similar simulation studies and compare performance across algorithms (see epistasis proposal)
- The EWG is open to all interested – email me to join the googlegroup!

Acknowledgements:

Nicodemus Group: Lara Neira Gonzalez

MRC Confidence in Concept (x2), The Carnegie Trust, Science Foundation Ireland

Email: kristin.nicodemus@igmm.ed.ac.uk