Randomness of Random Forest

Lipidome Profiling of Finnish Men With Prostate Cancer in a Randomized Clinical Trial – An AI approach

13.5.2019 - Artificial Intelligence & Statistics – Friends or Foes?

Paavo Raittinen / Aalto / SCI / Stochastics & Statistics

Needle, possibly in a haystack

Lost in translation

In machine		In statistics
Learning		Fitting
Weights		Parameters
Features		Covariates
Supervised learning		Classification
N/A		Hypothesis

The field and the haystack

Physical

Socio-Economic

Exposure

Molecular

Immunoprofiling

...

The field and the haystack

Physical

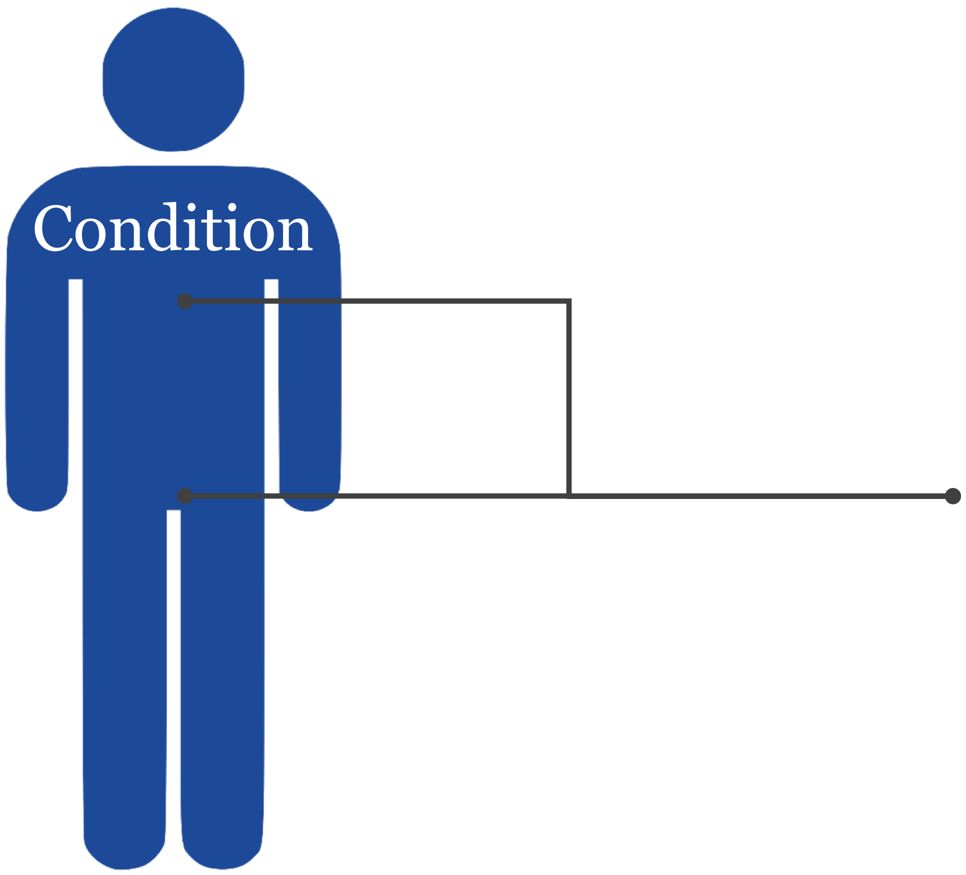
Socio-Economic

Exposure – system-wide

Molecular

Immunoprofiling

The field and the haystack



Physical

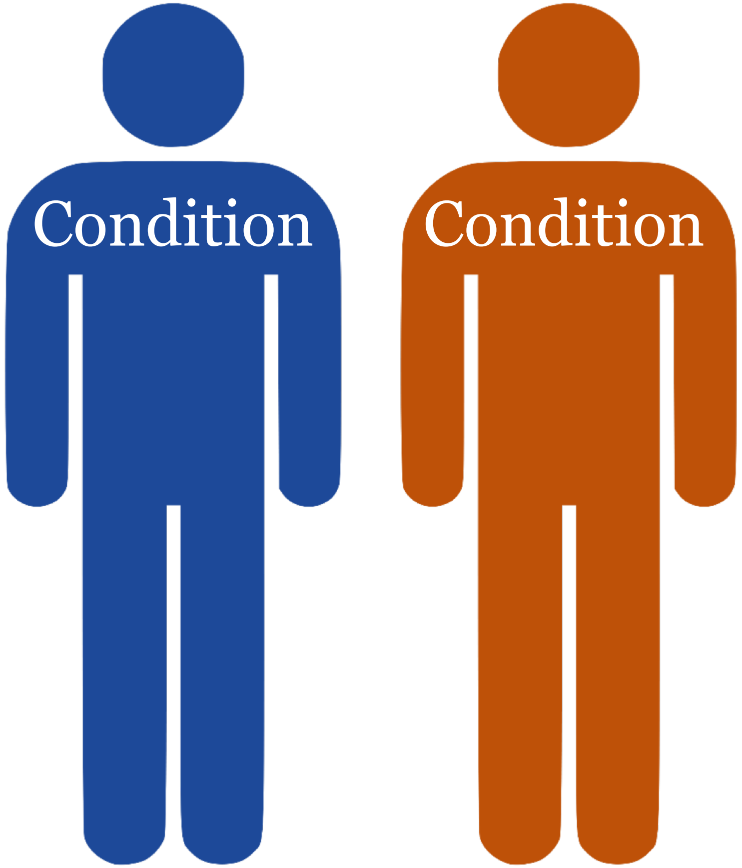
Socio-Economic

Exposure – system-wide

Molecular – system-wide and local

Immunoprofiling

Randomized Clinical Trial



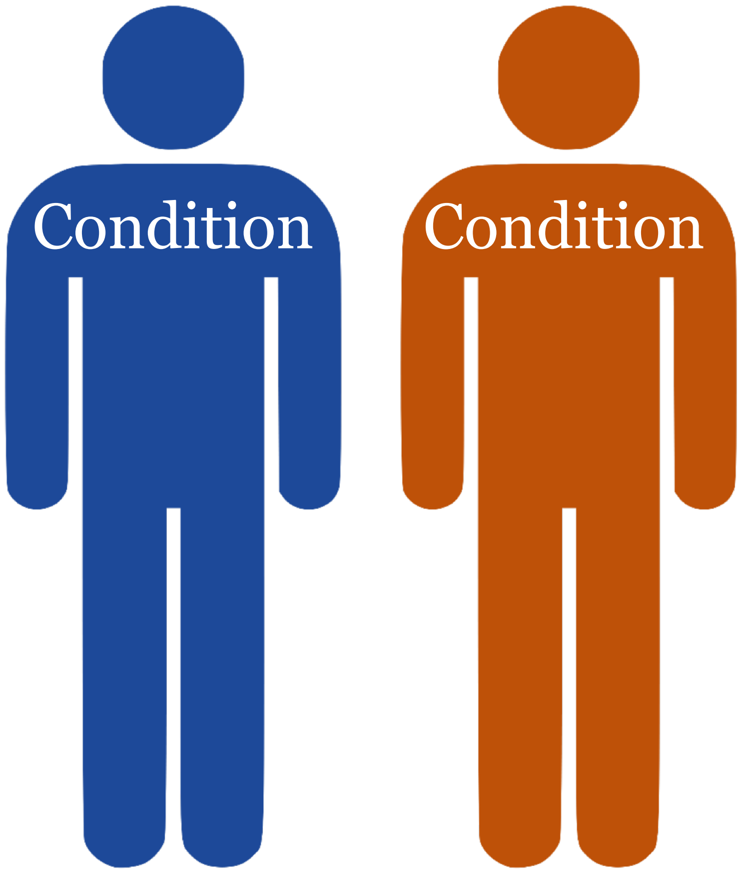
Real drug **Placebo**

Exposure y_i, $y \in \{0,1\}$

Lipidome **X** is n x p data matrix, p >> n

Condition Prostate cancer

Baseline



Statin **Placebo**

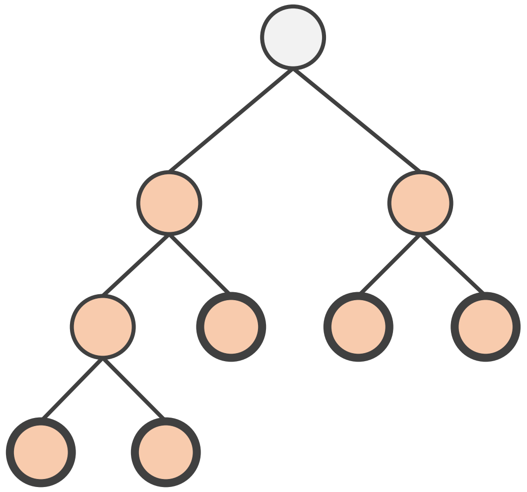
Cholesterol-lowering statins are associated with improved survival among prostate cancer patients

The serum lipidome contains **212** lipid aggregates, whereas the intraprostatic lipidome contains **4494** molecules. The RCT has **100** men.

Does the statin intervention cause lipidome shift in the serum and in the prostate?

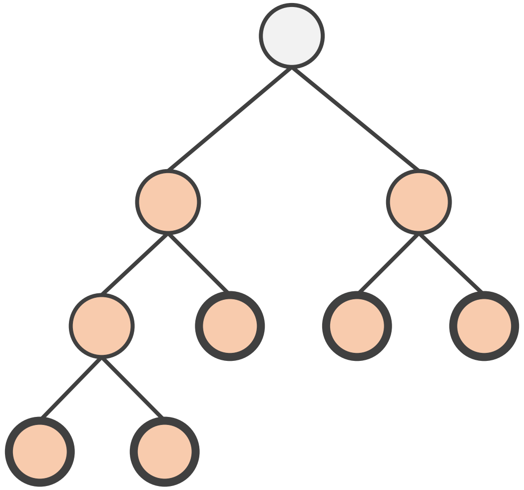
Random Forest Classification

A decision tree



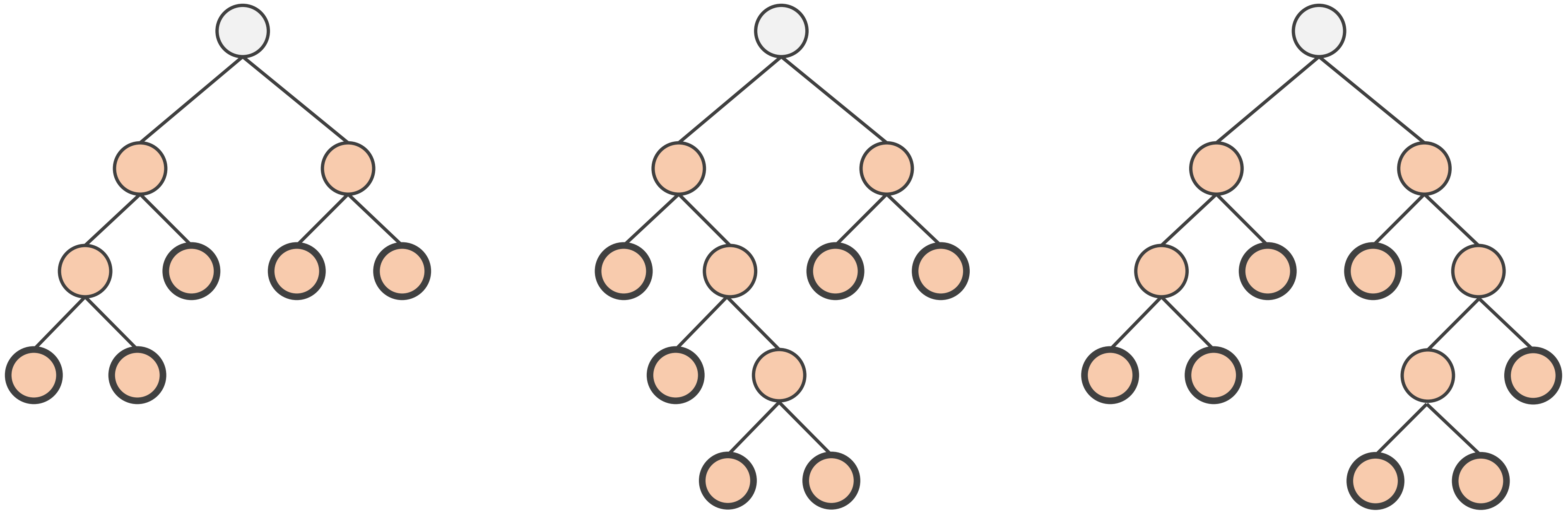
Random Forest Classification

Multiple trees is…

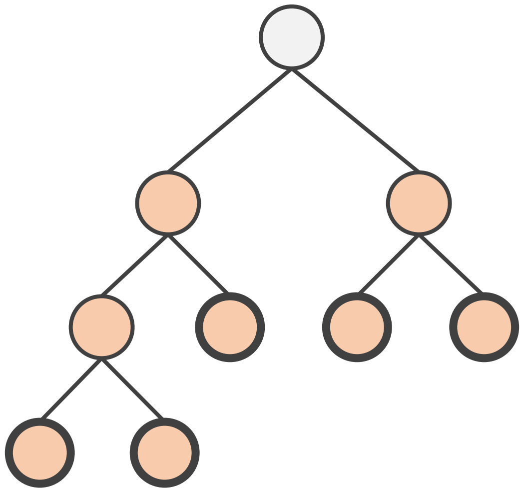


Random Forest Classification

Multiple trees is…a forest



Random Forest Classification



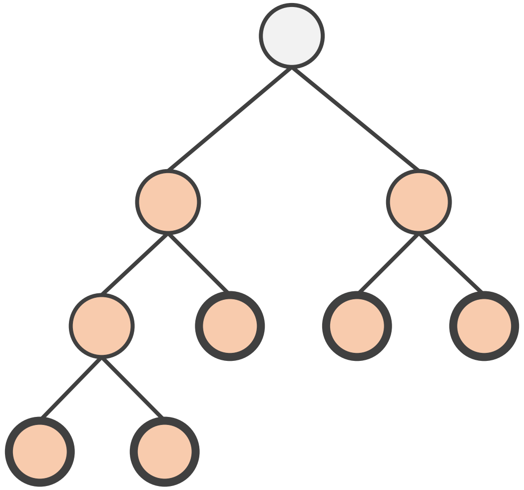
1. Draw a bootstrap sample B of size N from the training data
2. Grow a random forest tree to the bootstrapped data, and repeat:
   i. Select m variables randomly from the p variables
   ii. Pick the best variable/split-point among the m
   iii. Split the node into two daughter nodes
3. Output the ensemble of trees, i.e., the forest
4. Predict the class based on majority vote

Obtain:

1. **Classification error**
2. N x N **proximity matrix**
3. **Variable importance**

Random Forest Classification

How about in practice?



Can we make inference based on:

1. Classification error
2. N x N proximity matrix
3. Variable importance

Random Forest In Practice

Serum lipidome before the intervention: $n = 100$, $p = 212$

Random Forest In Practice

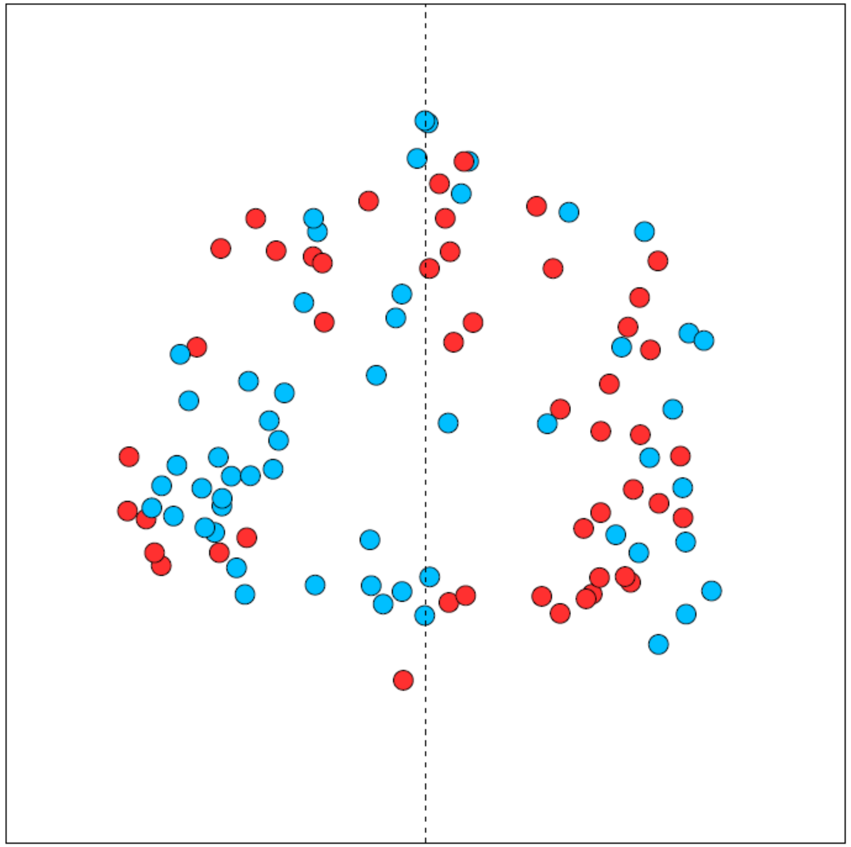
Serum lipidome before the intervention: n = 100, p = 212

1. **Classification error: 44.66 %**
   (Placebo 48 %, Statin 42 %)

Random Forest In Practice

Serum lipidome before the intervention

1. Classification error: 44.66 %
   (Placebo 48 %, Statin 42 %)
2. **Proximity plot**
3. Variable importance N/A



Random Forest In Practice

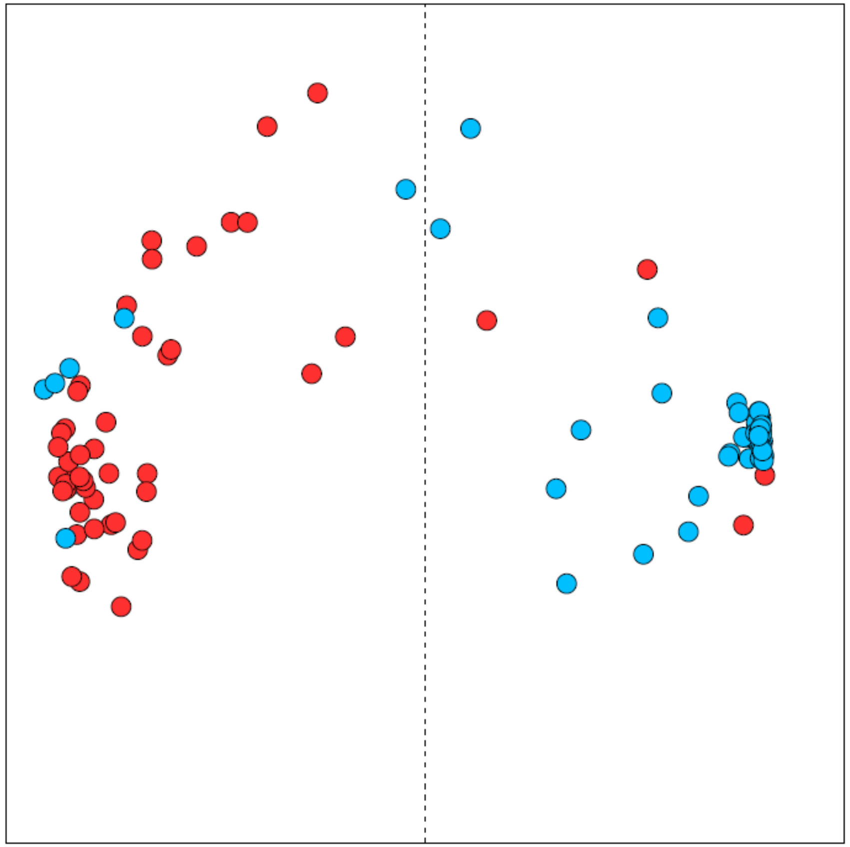
Serum lipidome after the intervention

1. Classification error: 11.65 %
   (Placebo 8.33 %, Statin 14.55 %)

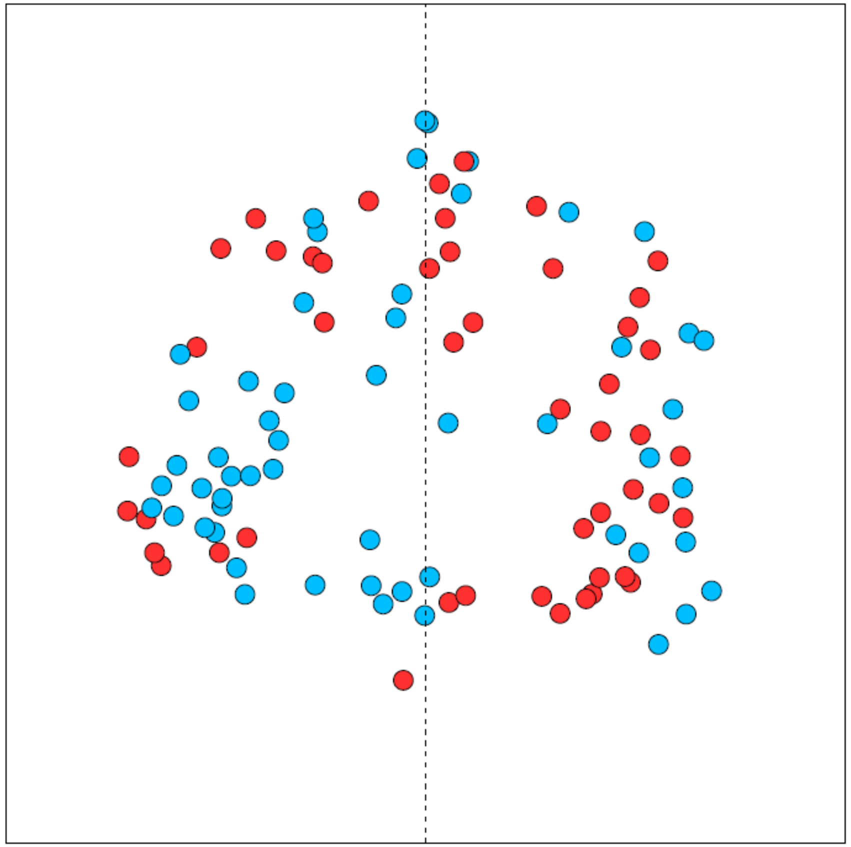
Random Forest In Practice

Serum lipidome after the intervention

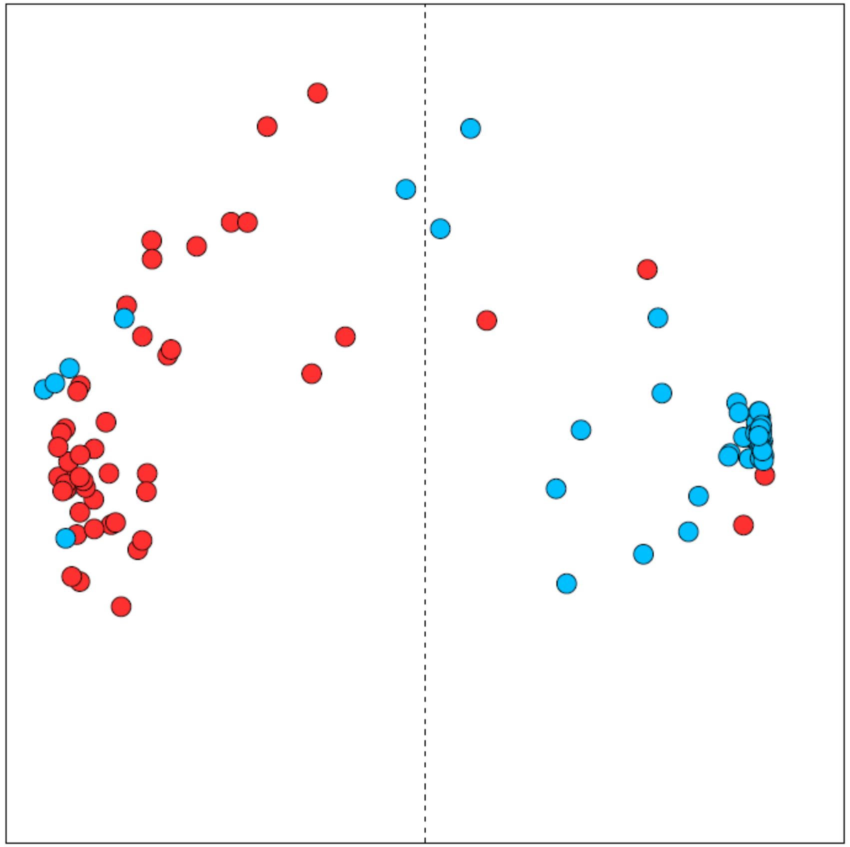
1. Classification error: 11.65 %
   (Placebo 8.33 %, Statin 14.55 %)
2. Proximity plot
3. Variable importance
   1. Total Cholesterol in IDL
   2. Cholesterol esters in IDL
   3. Concentration of Large LDL



Random Forest In Practice



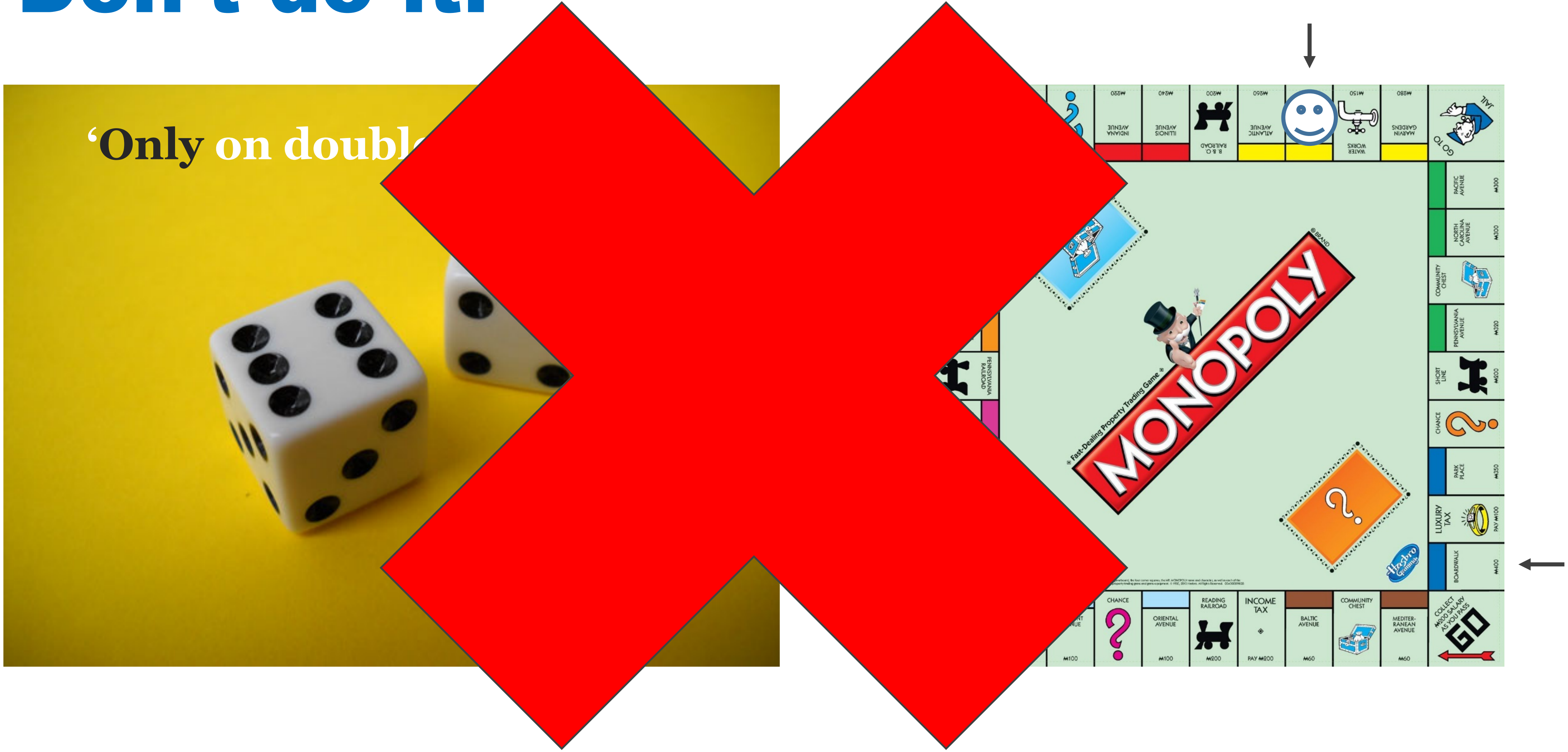
Before = random



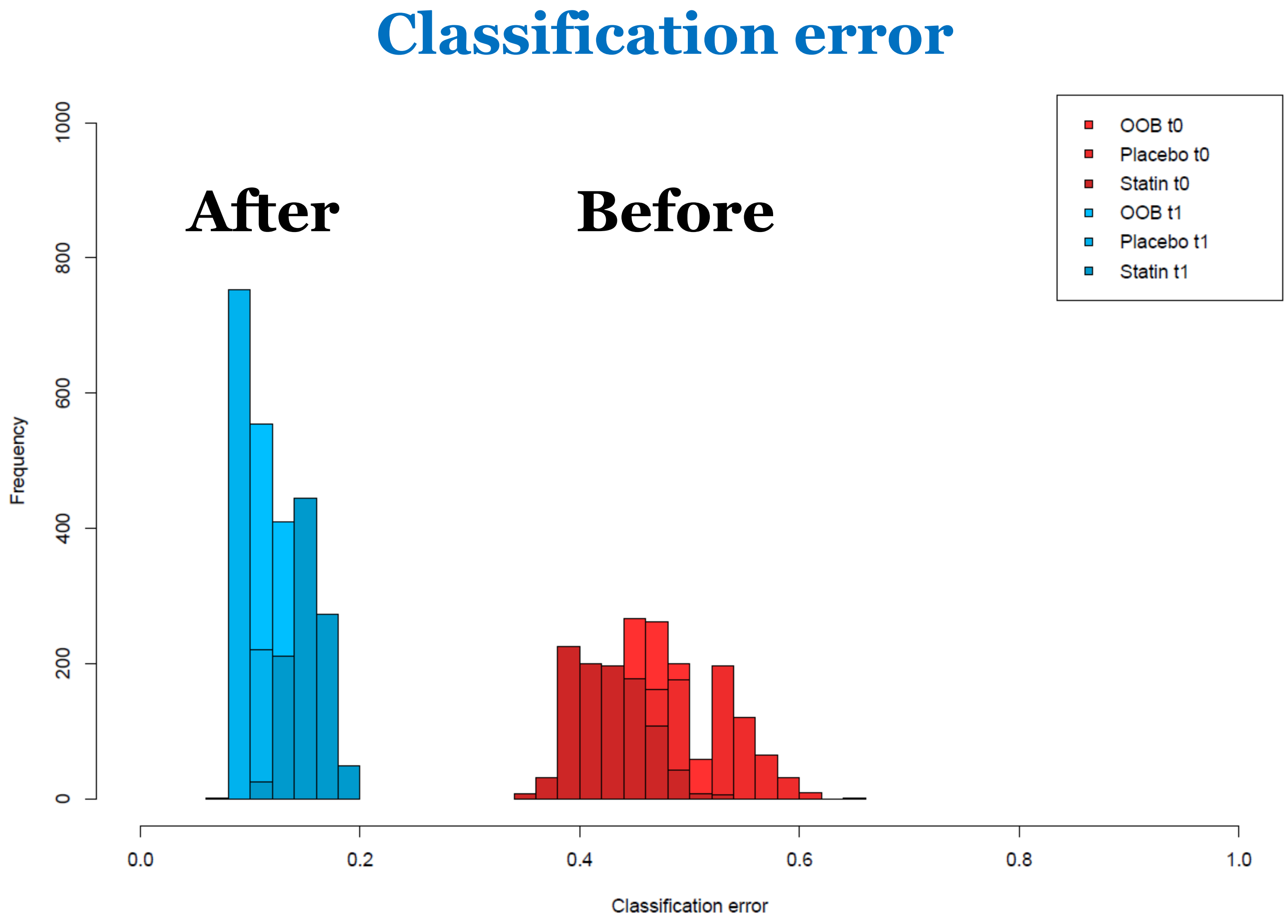
After = systematic

Wait, how about chance?

Don’t do it!


‘Only on doubl
MONOPOLY
Fast-Dealing Property Trading Game
GO
JAIL
INCOME TAX
LUXURY TAX
CHANCE
COMMUNITY CHEST

Heuristic bootstrap confidence interval



Random Forest In Practice

Intraprostatic lipidome after the intervention: $n = 100$, $p = 4494$

Random Forest In Practice

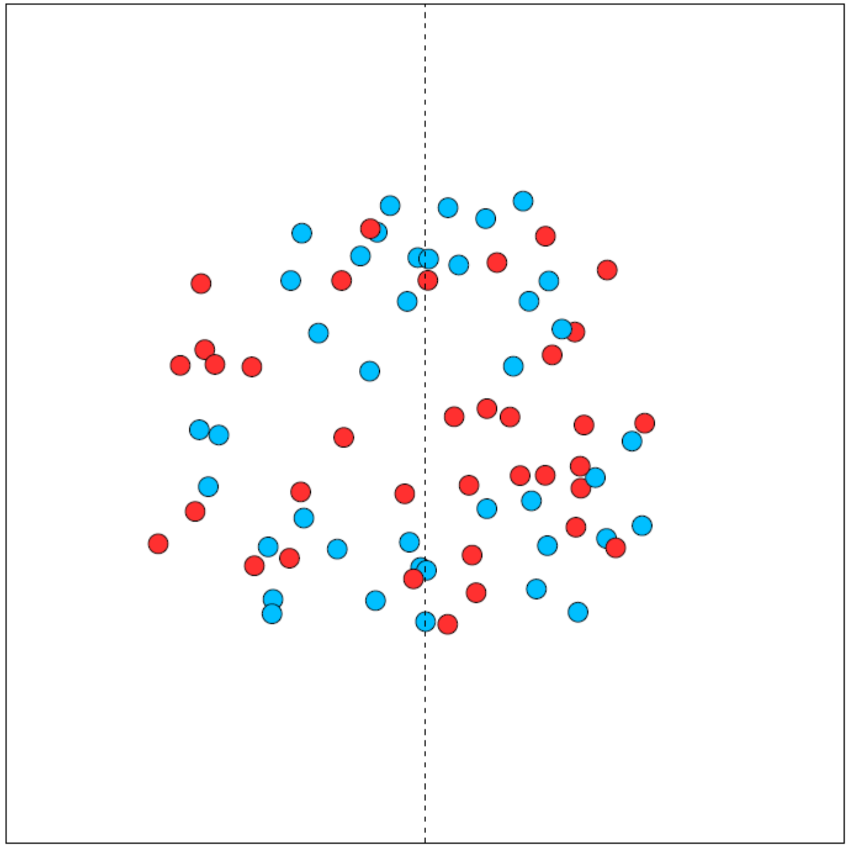
Intraprostatic lipidome after the intervention: n = 100, p = 4494

1. **Median** classification error: 50 %
   (Placebo 55 %, Statin 45 %)

Random Forest In Practice

Intraprostatic lipidome

1. **Median** classification error: 50 %
   (Placebo 55 %, Statin 45 %)
2. Proximity plot

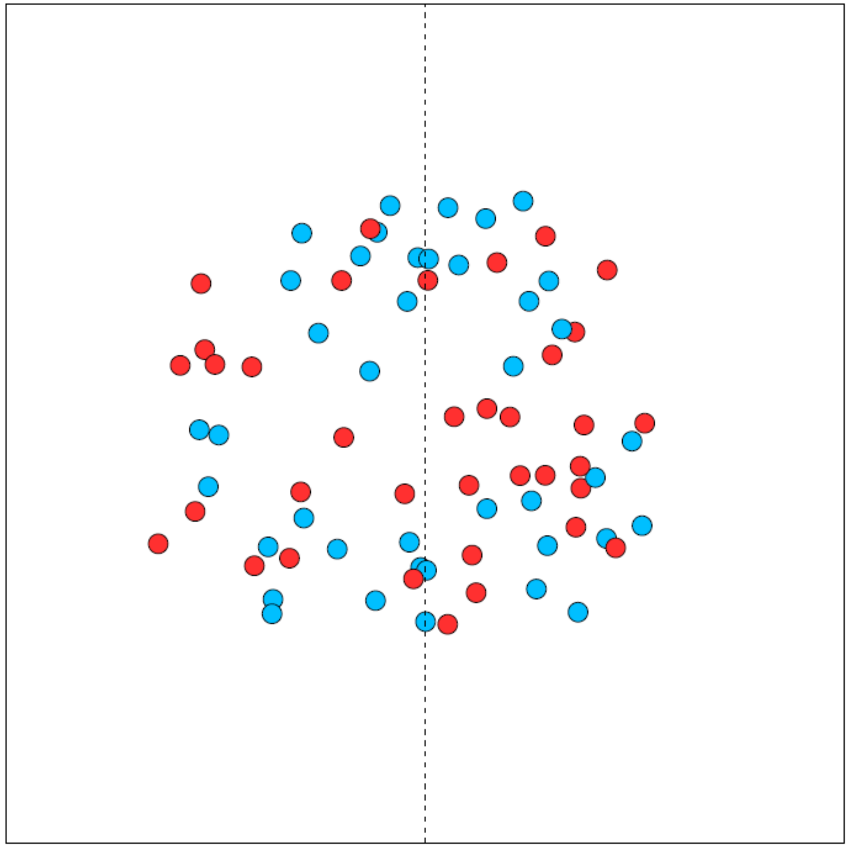


Random Forest **In Practice**

Intraprostatic lipidome

1. **Median** classification error: 50 % (Placebo 55 %, Statin 45 %)
2. Proximity plot

→ Too much hay in the stack

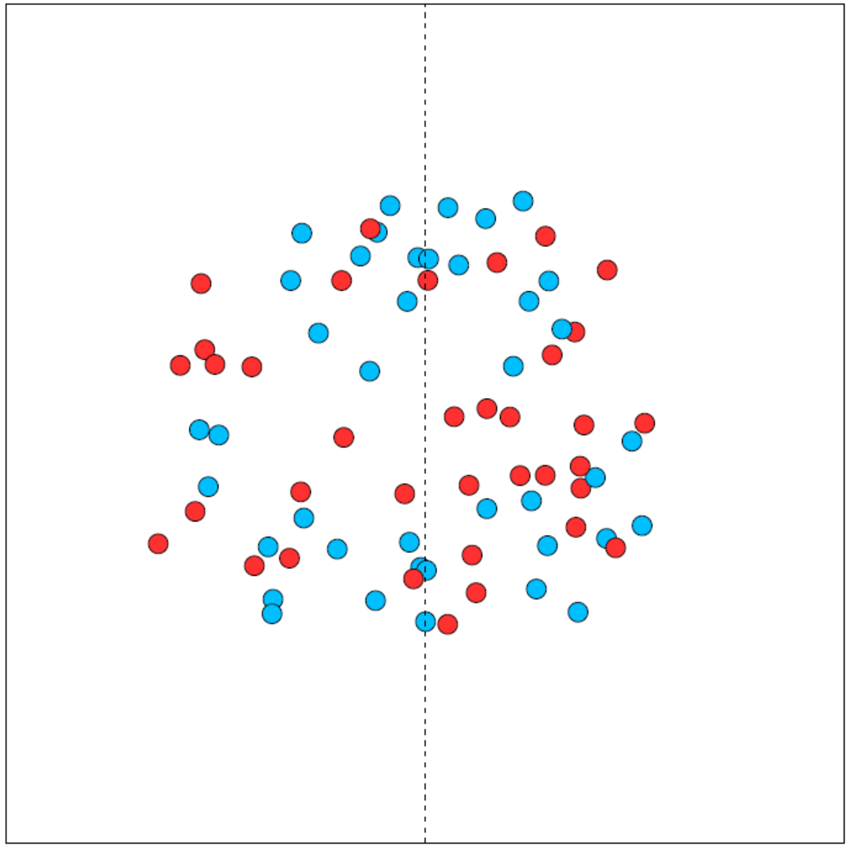


Random Forest In Practice

Intraprostatic lipidome

1. **Median** classification error: 50 %
   (Placebo 55 %, Statin 45 %)
2. Proximity plot

→ Too much hay in the stack
→ Need brain…and "t-test"
  → Roughly search for statistically significant difference in the lipid levels between the study arms, discard non-significant from the analysis.



Random Forest In Practice

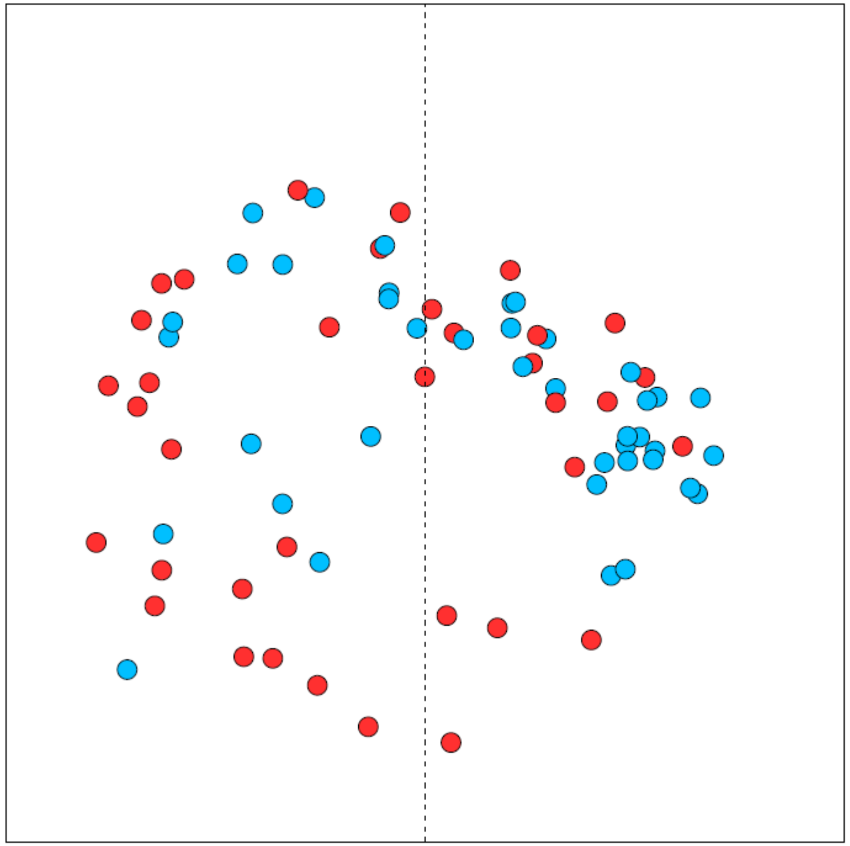
Intraprostatic lipidome after the intervention: $n = 100$, $p = 22$

1. **Median** classification error: 36.8 %
   (Placebo 41.6 %, Statin 35 %)

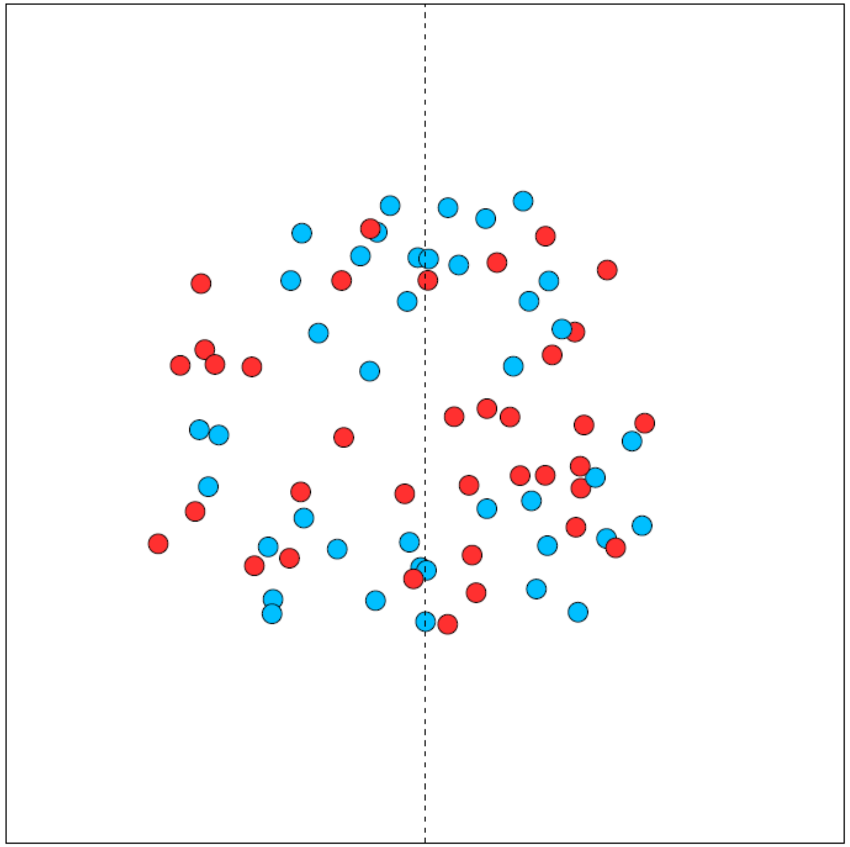
Random Forest In Practice

Intraprostatic lipidome

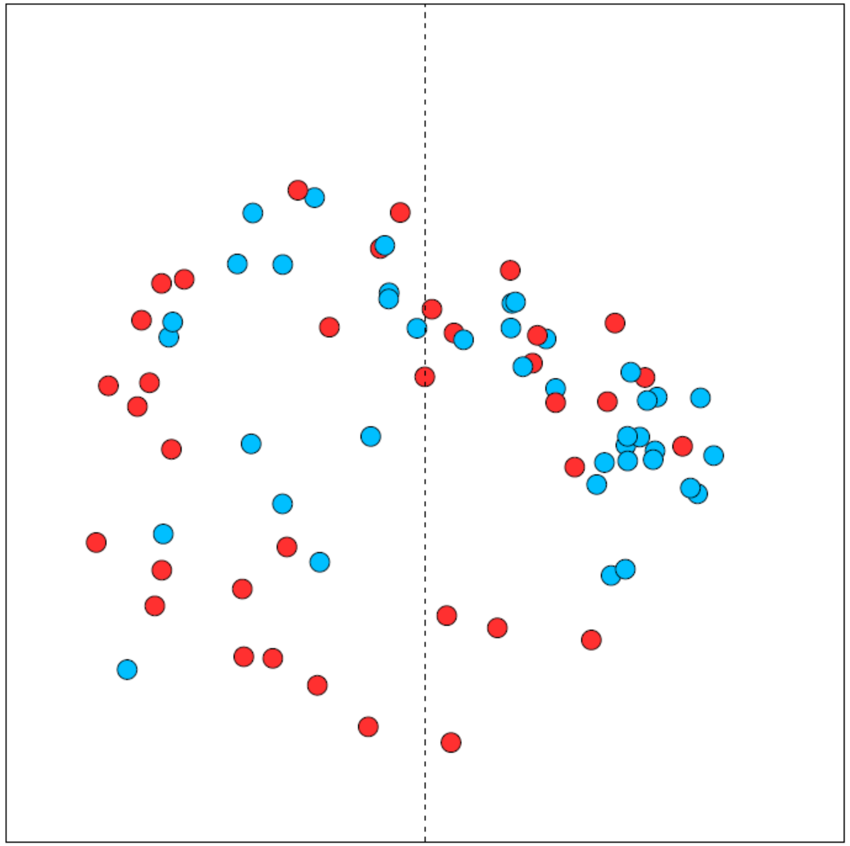
1. **Median** classification error: 36.8 % (Placebo 41.6 %, Statin 35 %)
2. Proximity plot
3. Variable importance:
    1. Vitamin-D like compounds
    2. LPC 20:4
    3. PC 20:1_18:1



Random Forest In Practice

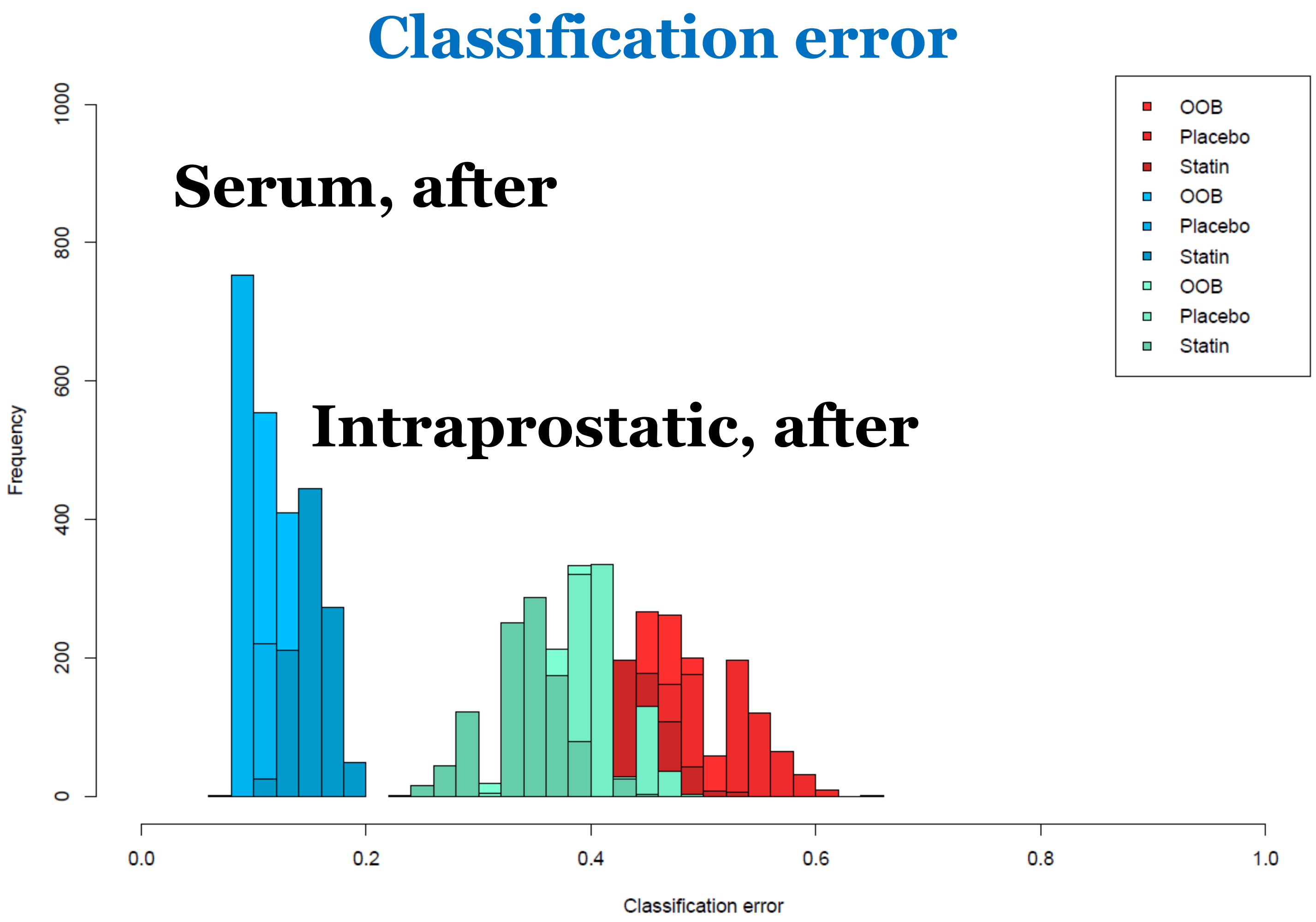


Too much hay

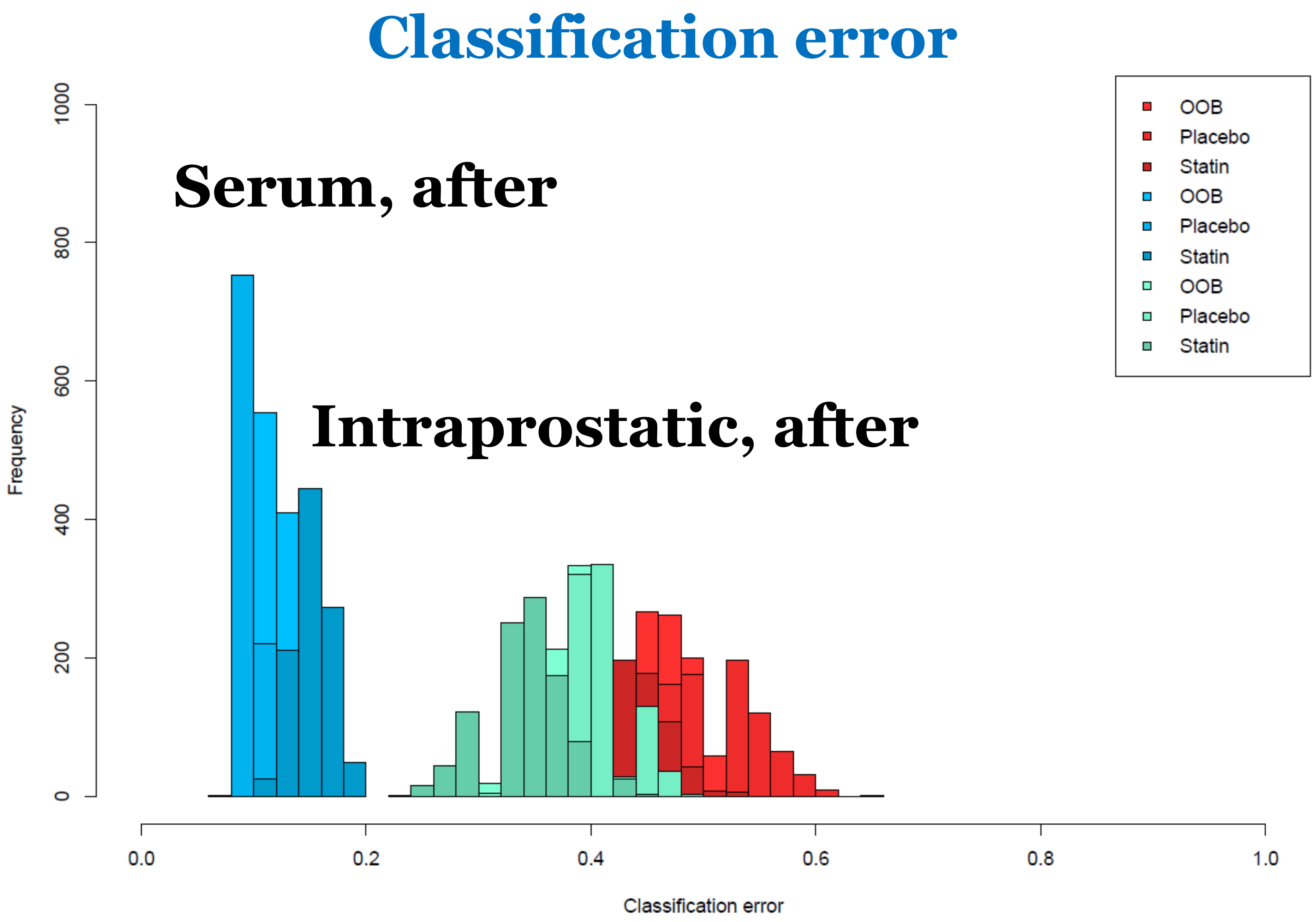


Reduced hay

Heuristic bootstrap confidence interval



Beats the coin flip...



Conclusion statement

1. **Statin intervention causes clear lipidome shift in the serum, as expected.**
2. **Furthermore, we observe a slight shift in the intraprostatic lipidome profile as well.**

Therefore, any benefit statin use might display, can be partly mediated by lipids.

Wrap-up

- This time, the needle was in the haystack

Wrap-up

- **This time, the needle was in the haystack**
- **The friendly trio, AI, Machine Learning, and statistics are all every-day tools in multiple fields…**

Wrap-up

- This time, the needle was in the haystack
- The friendly trio, AI, Machine Learning, and statistics are all every-day tools in multiple fields…
- …They are also really good tools when they are interpretable and help you to explain the underlying mechanism

Wrap-up

- This time, the needle was in the haystack
- The friendly trio, AI, Machine Learning, and statistics are all every-day tools in multiple fields
- They are also really good tools when they are interpretable and help you to explain the underlying mechanism
- Furthermore, it is really helpful if you can communicate what you do, as an expert, to another expert

Wrap-up

- This time, the needle was in the haystack
- The friendly trio, AI, Machine Learning, and statistics are all every-day tools in multiple fields
- They are also really good tools when they are interpretable and help you to explain the underlying mechanism
- Furthermore, it is really helpful if you can communicate what you do, as an expert, to another expert
- You should not trash t-test

References

- **Breiman, Leo**. "Random forests." Machine learning 45.1 (2001): 5-32.
- **Friedman, Jerome, Trevor Hastie, and Robert Tibshirani.** The elements of statistical learning. Vol. 1. No. 10. New York: Springer series in statistics, 2001.

Thank you!

This is the end of the presentation.

Artificial Intelligence & Statistics – Friends
13.5.2019 - Paavo Raittinen