Data mining for vaccine process understanding

Matthew Wiener
Department of Applied Computer Science & Mathematics
Merck & Co.

Acknowledgements

- This talk represents the work of many people in many departments at Merck

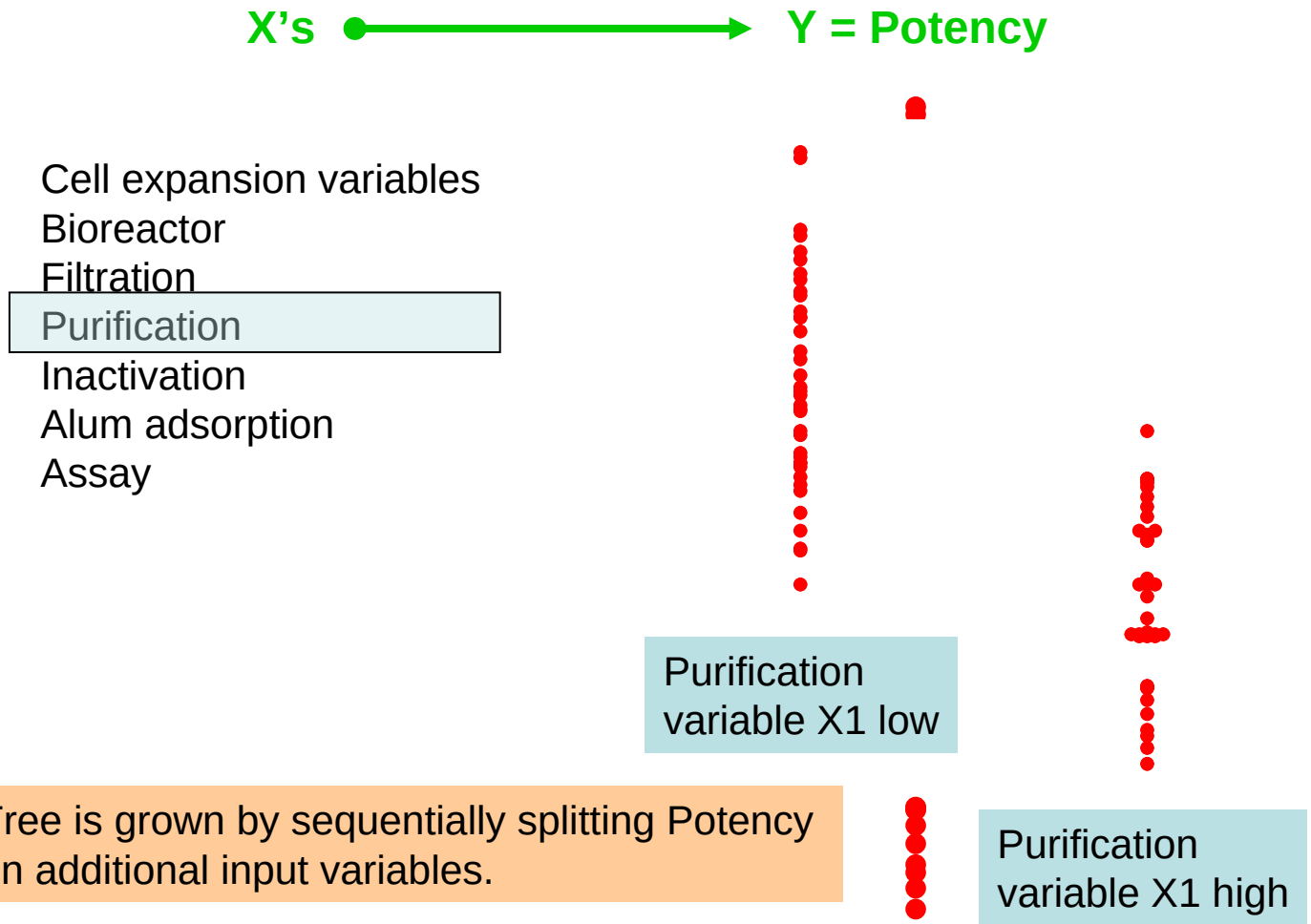
The problem

- Vaccines, once discovered, still have to be manufactured in sufficient quantities.
- This tends to be complicated, especially for viral vaccines, which are frequently grown in cell culture.
- Variable yield can (and does) complicate planning.
  - Low yield can prevent manufacturers from meeting demand, with potentially large consequences for both public health and sales.
  - Even unexpectedly high yield can raise questions about whether we are producing what we expected, preventing the sale of vaccine.

Building Process Understanding

- Generate hypotheses about causes of potency shifts
  - Identify suspect process changes using multivariate data mining
    - » Random forests – tree-based method
- Check hypotheses for scientific reasonableness
  - Check with subject matter experts
    - » Biologists, manufacturing engineers, process supervisors and technicians, etc.
    - » If things go really well, we may \ even be able to explain why those variables are the critical ones.
- Confirm with further data
  - Validate models using new production results
  - Design controlled studies if needed

Tree-based methods
(recursive partitioning based on predictors)



Tree is grown by sequentially splitting Potency on additional input variables.

Random Forests (Breiman 1996, 2001)

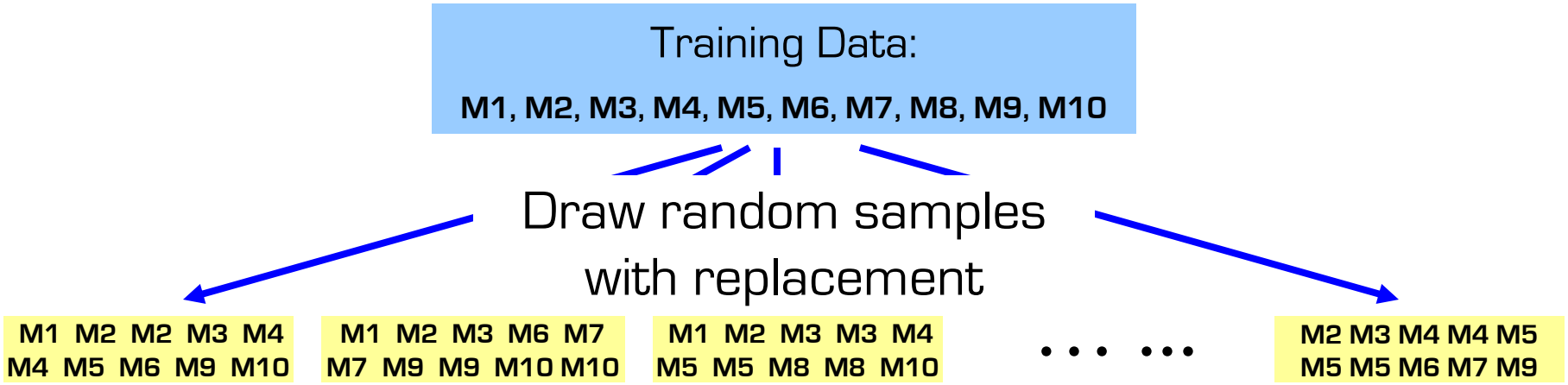
- A collection of trees with controlled variations – two kinds of randomness
- Trees “vote” for the best answers (predictions).
- Advantages:
  - Consistently matches or outperforms accuracy of other data mining methods.
  - Handles a large number of inputs, resistant to over-fitting.
  - Very fast.
  - Not confounded by confounding.
  - Handles non-linear relationships.
  - **Estimates the importance of variables as predictors of the output.**

Growing a Forest

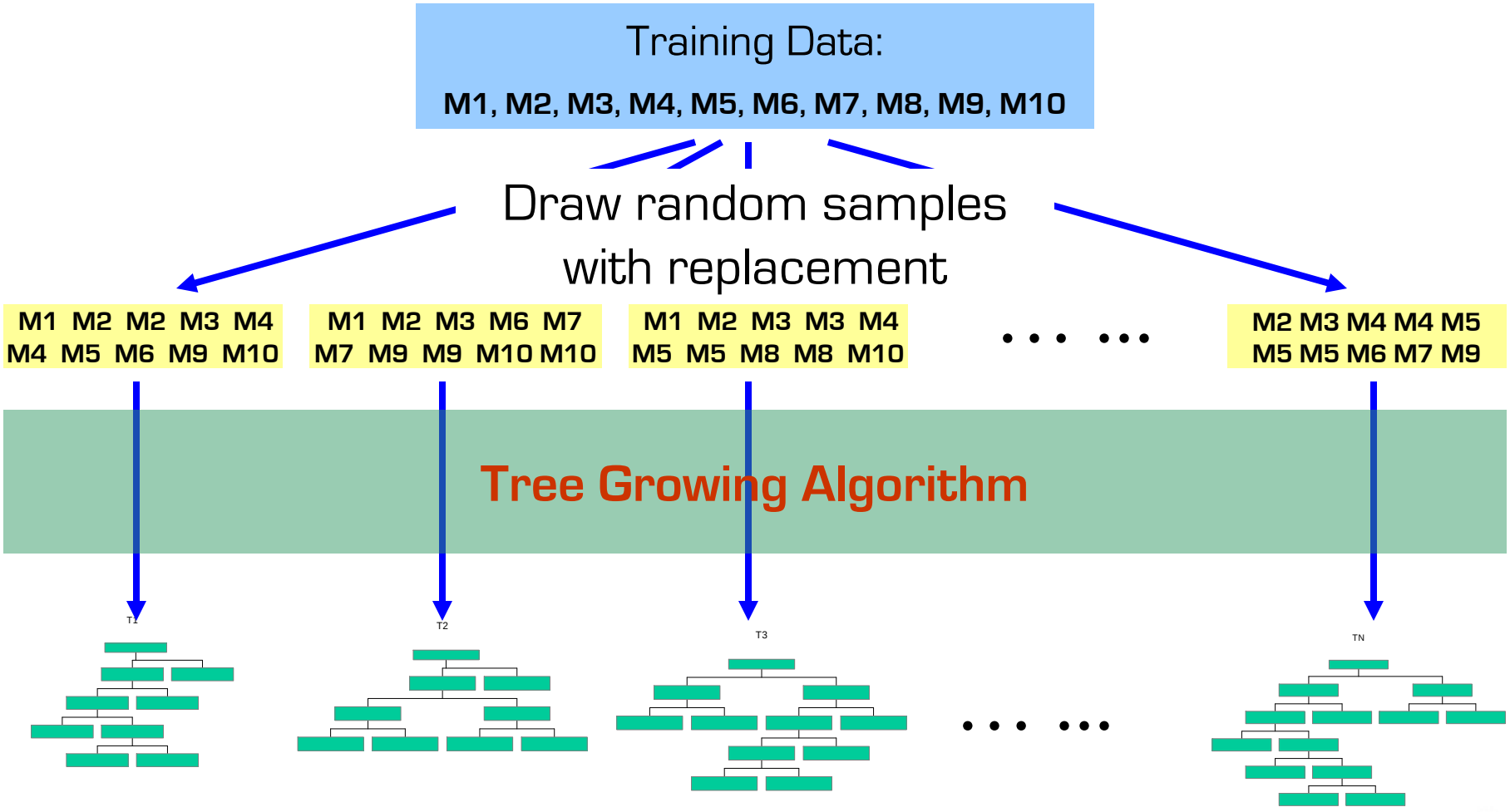
Training Data:

M1, M2, M3, M4, M5, M6, M7, M8, M9, M10

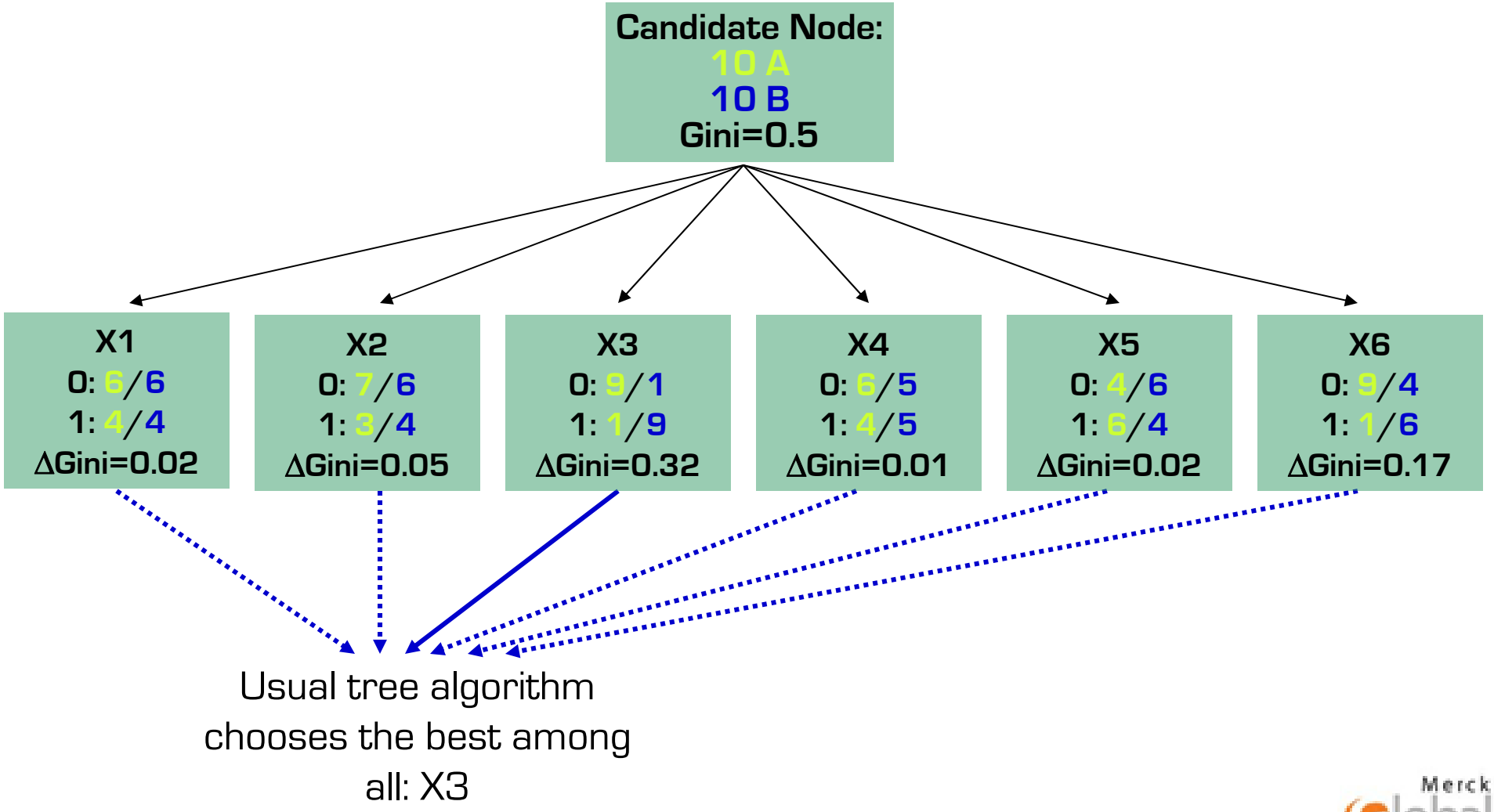
Growing a Forest



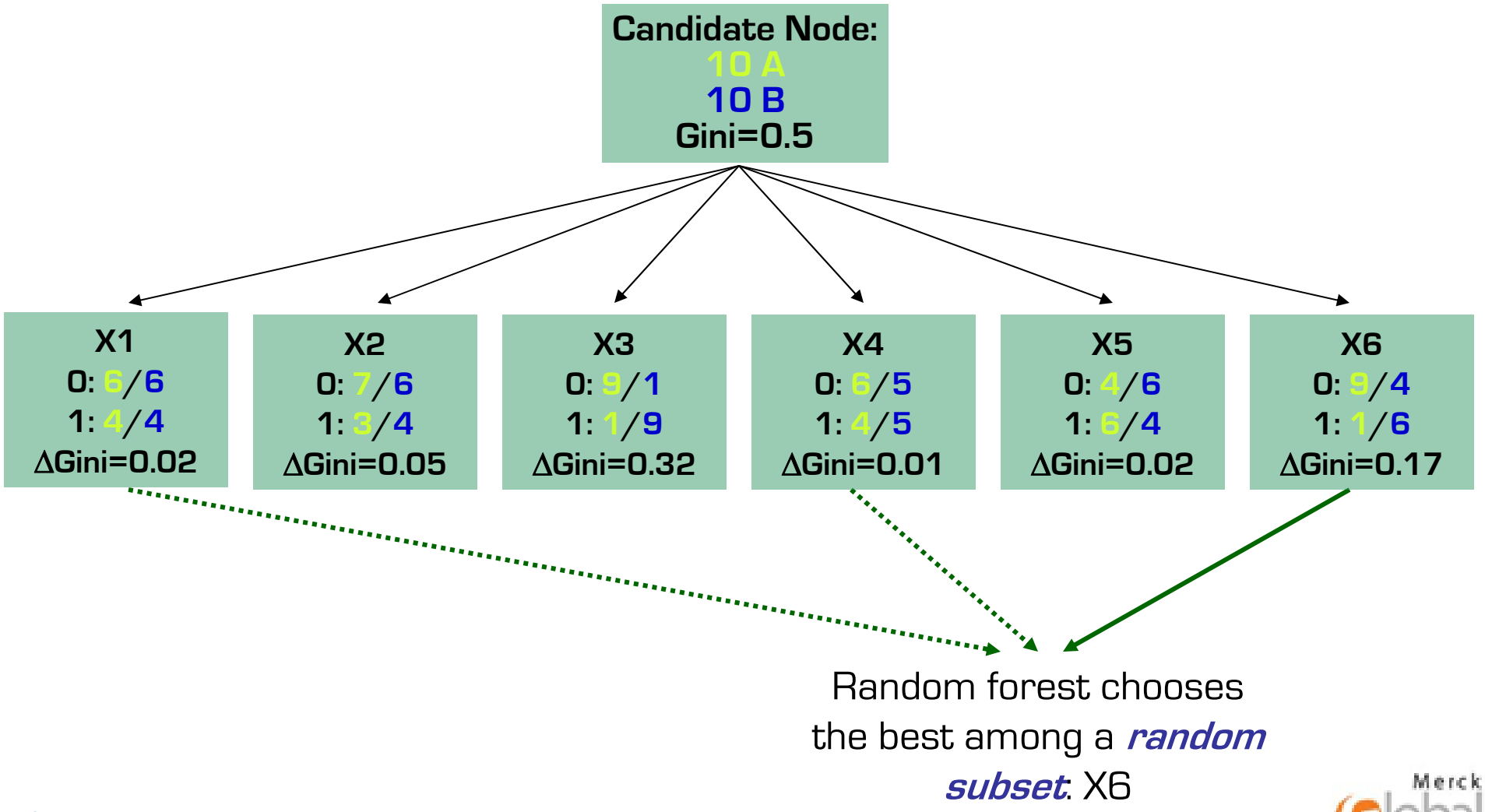
Growing a Forest



Semi-Random Splitting



Semi-Random Splitting



Estimate variable importance by shuffling

- If a variable has no information about the quantity to be explained, it won't be used much in the model, and shuffling it won't make any difference to your predictions.
- If a variable has a lot of information about the quantity to be explained, it will be used a lot. Shuffling the variable will make a big difference to your predictions.
- Random forest estimates importance by checking how much shuffling each variable changes the results of the fitted model.

What random forests give us

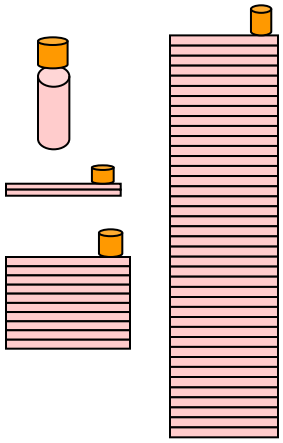
- A measure of variable importance
  - Orders the variables
  - “Consensus builder” in root cause investigations
- Good predictions and error estimates
  - » Consistently among the most accurate methods
  - » Effectively get predictions on cross-validation test set data
  - » Prediction for a point uses only trees without that point in the training set
  - » Resistant to overfitting
- Basically no parameters to fiddle with
  - Number of trees in forest, number of variables checked at each split
  - And not very sensitive to those

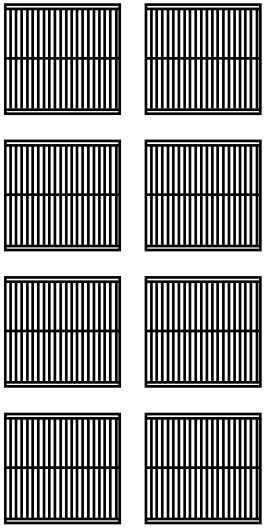
Example 1

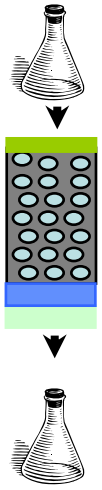
High-level manufacturing process

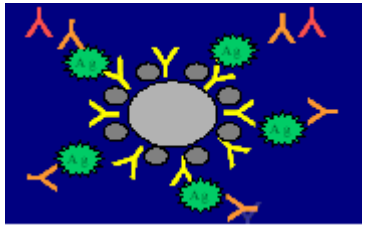
Bioreactors

Downstream









Cell Bank Lot exhausted; new lot introduced.

Virus Stock Seed Lot exhausted; new lot introduced.

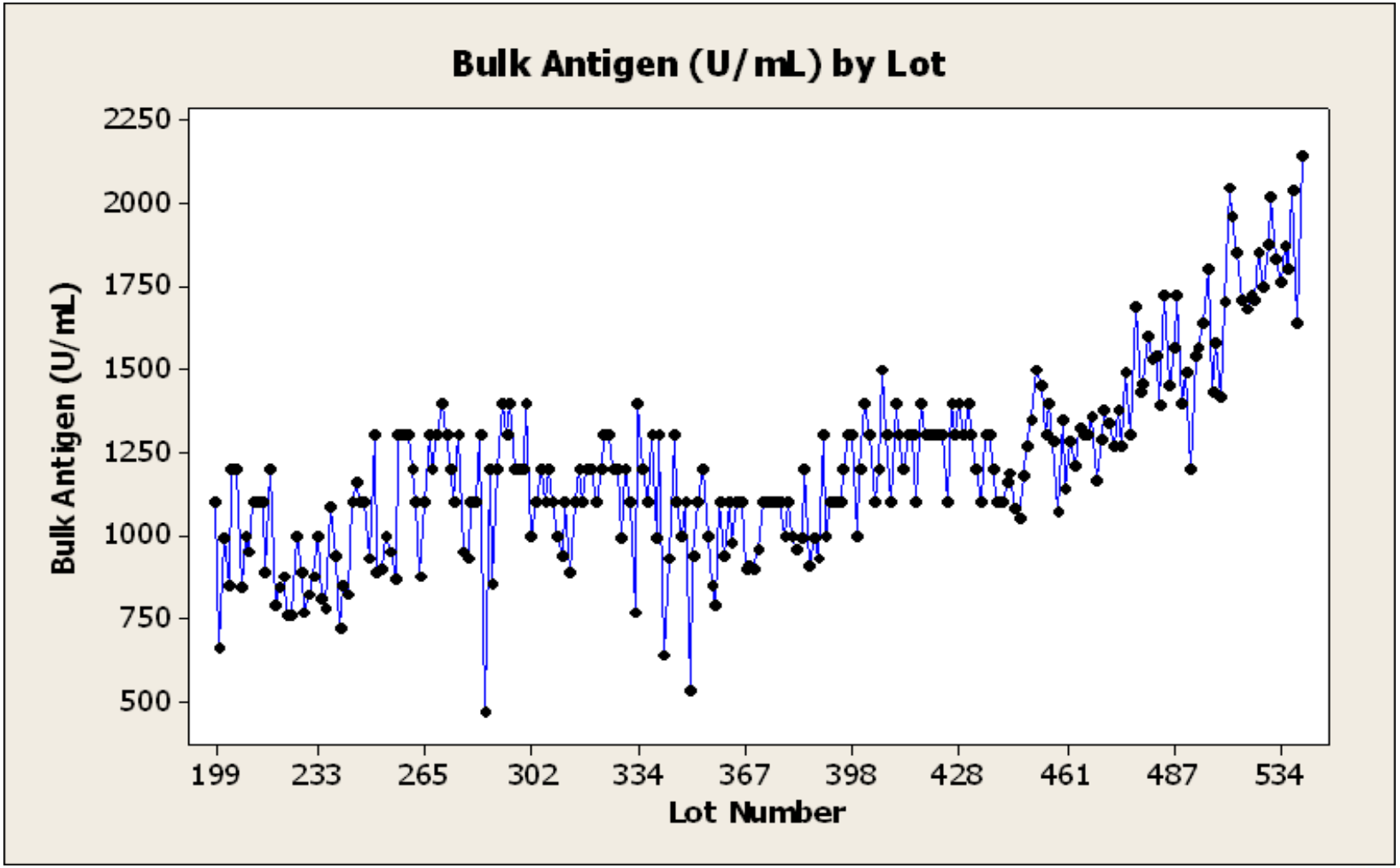
Chromatography resin lots exhausted and replaced.

Improved assay implemented

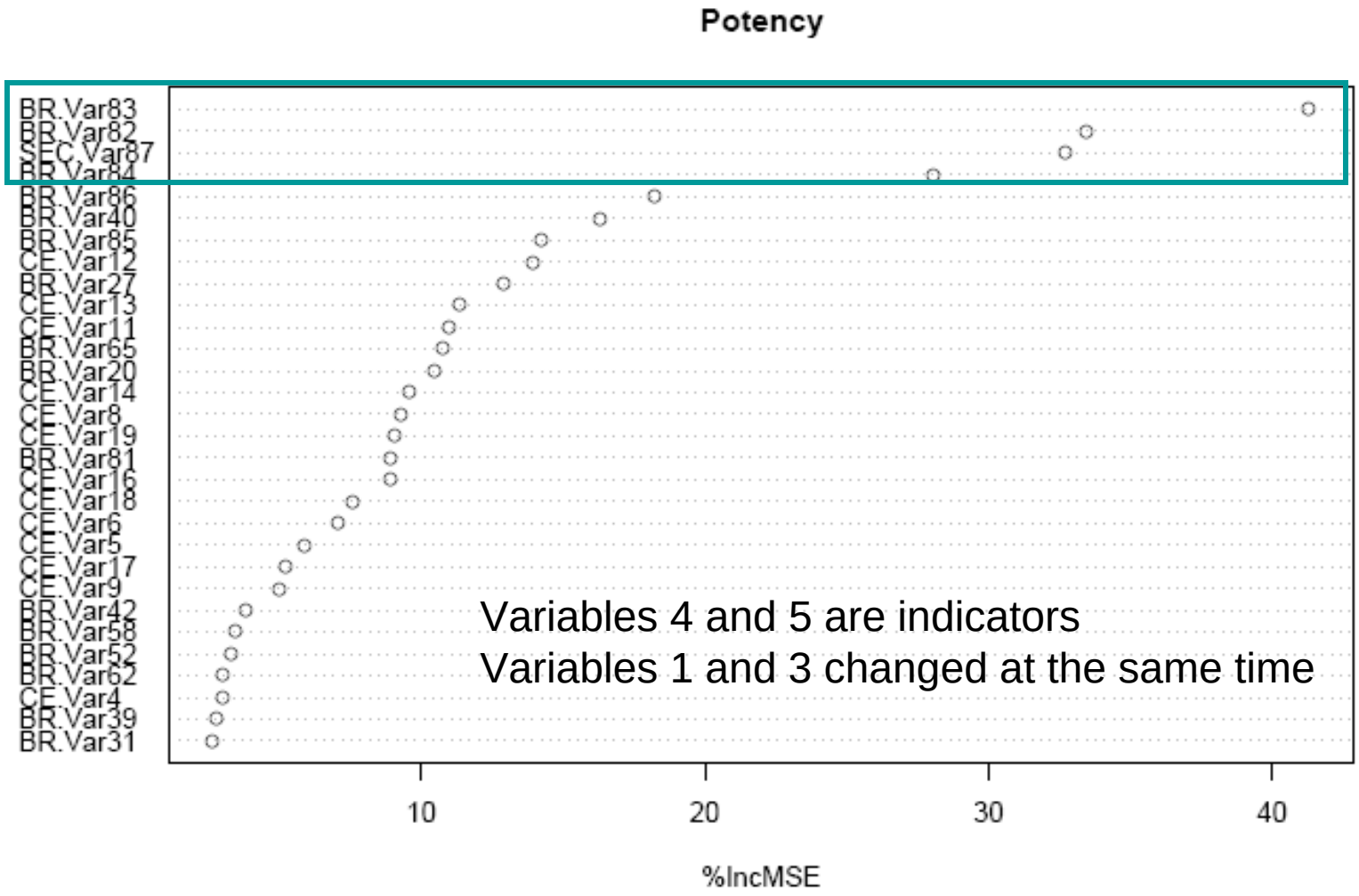
Raw Material preparation methods improved by vendor.

A “fixed process” does not guarantee a fixed product.

How to explain the increase?

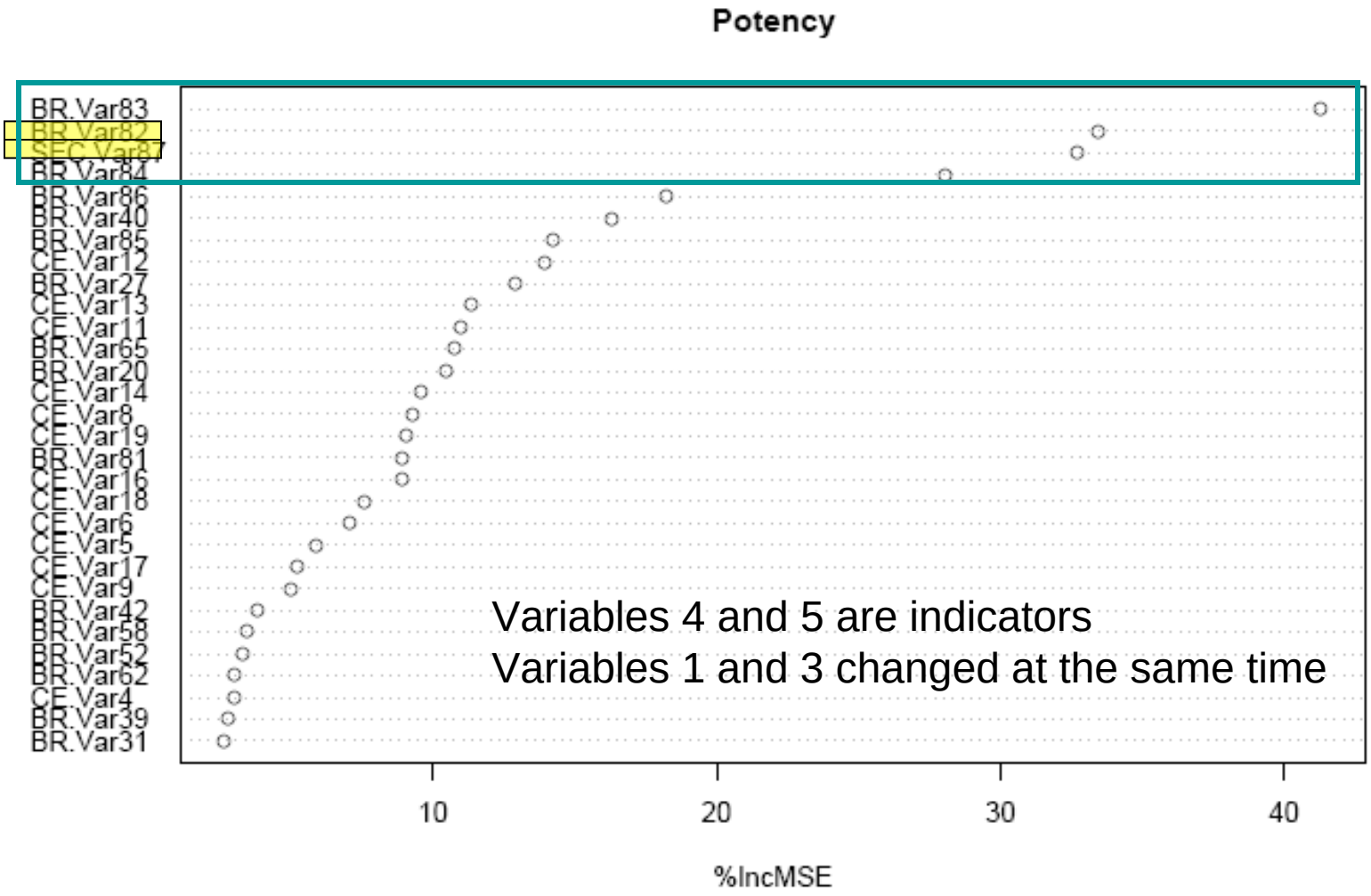


Variable Importance for predicting Potency by Random Forests

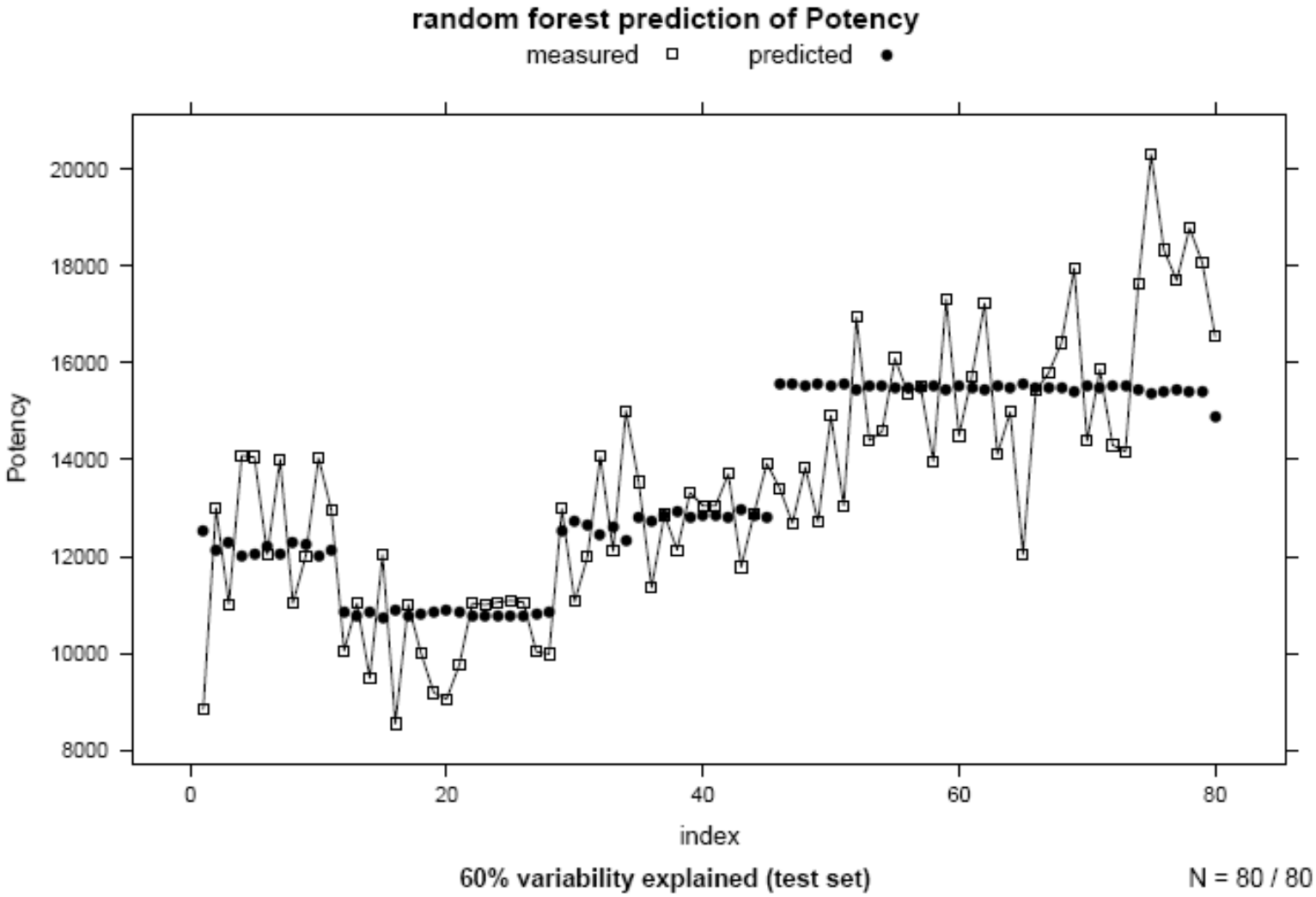


N = 64 / 80

Variable Importance for predicting Potency by Random Forests



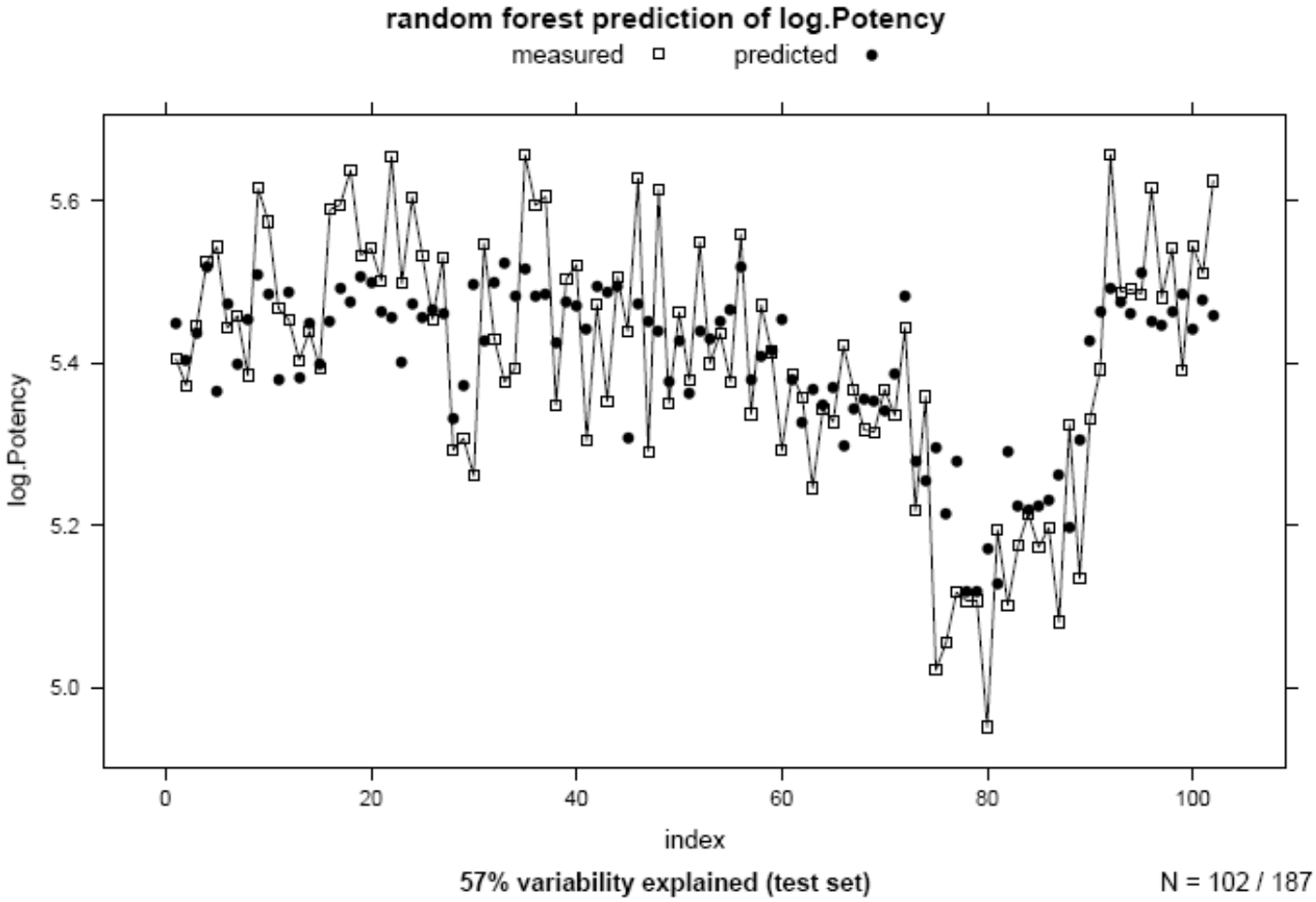
These two root causes explain 60% of the variability.



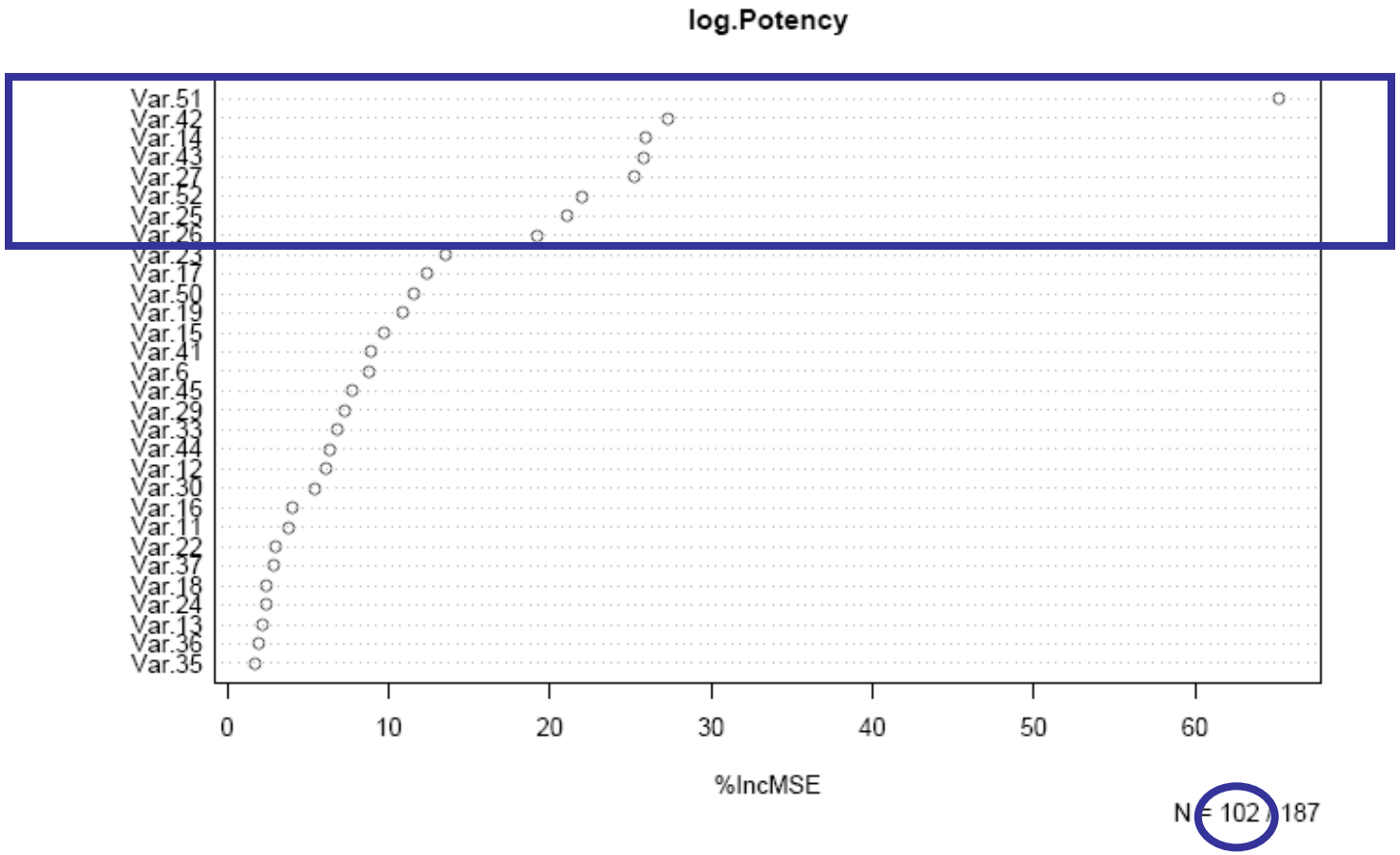
Conclusions – example 1

- Random forest analysis identified two variables, each with only two values, that together explained 60% of the variability in yield.
- The analysis helped bring consensus on root causes.
- The root causes were confirmed using direct experimentation.

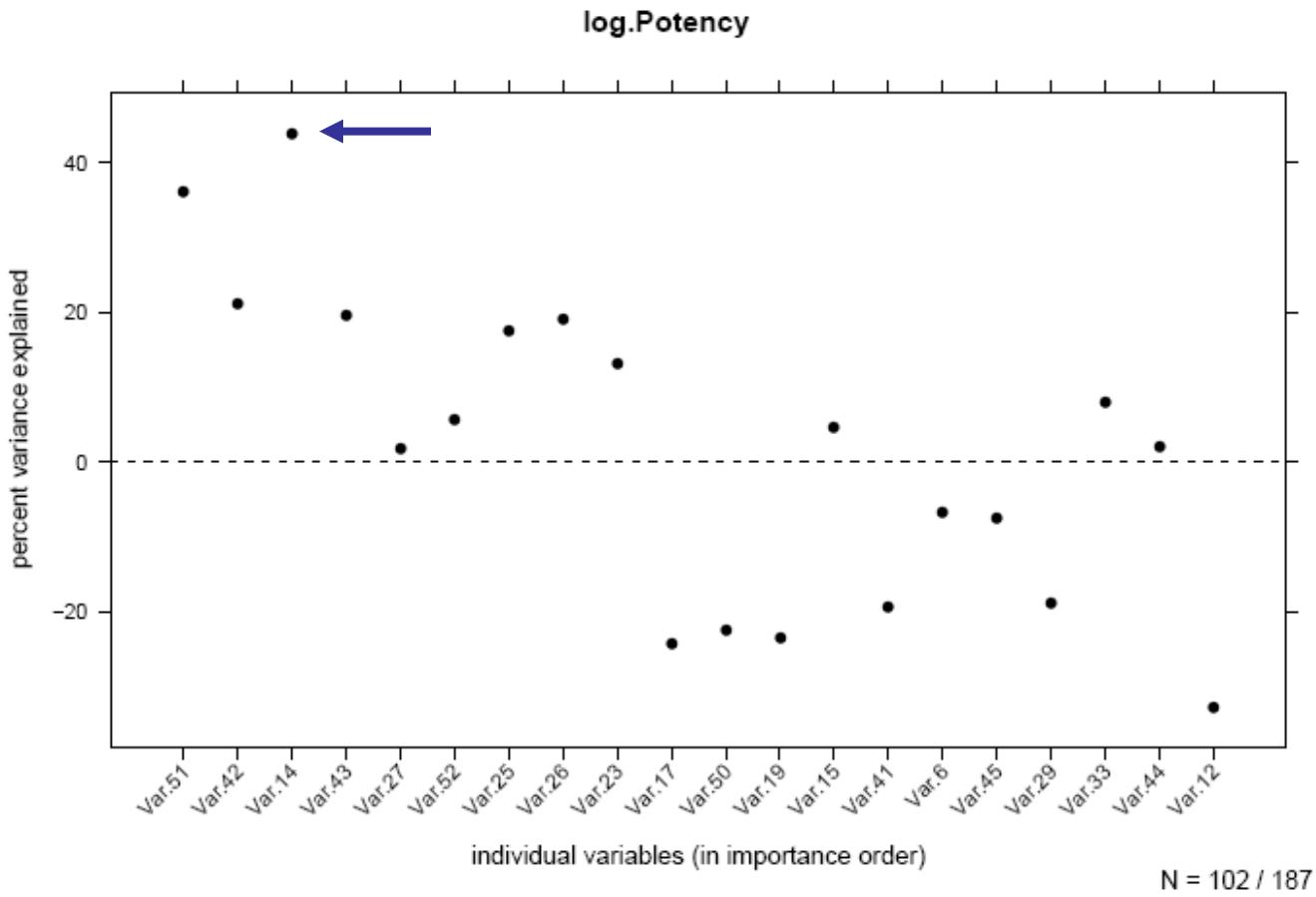
Example 2



Random forest indicates 1 variable – different units of a type of equipment – is most important

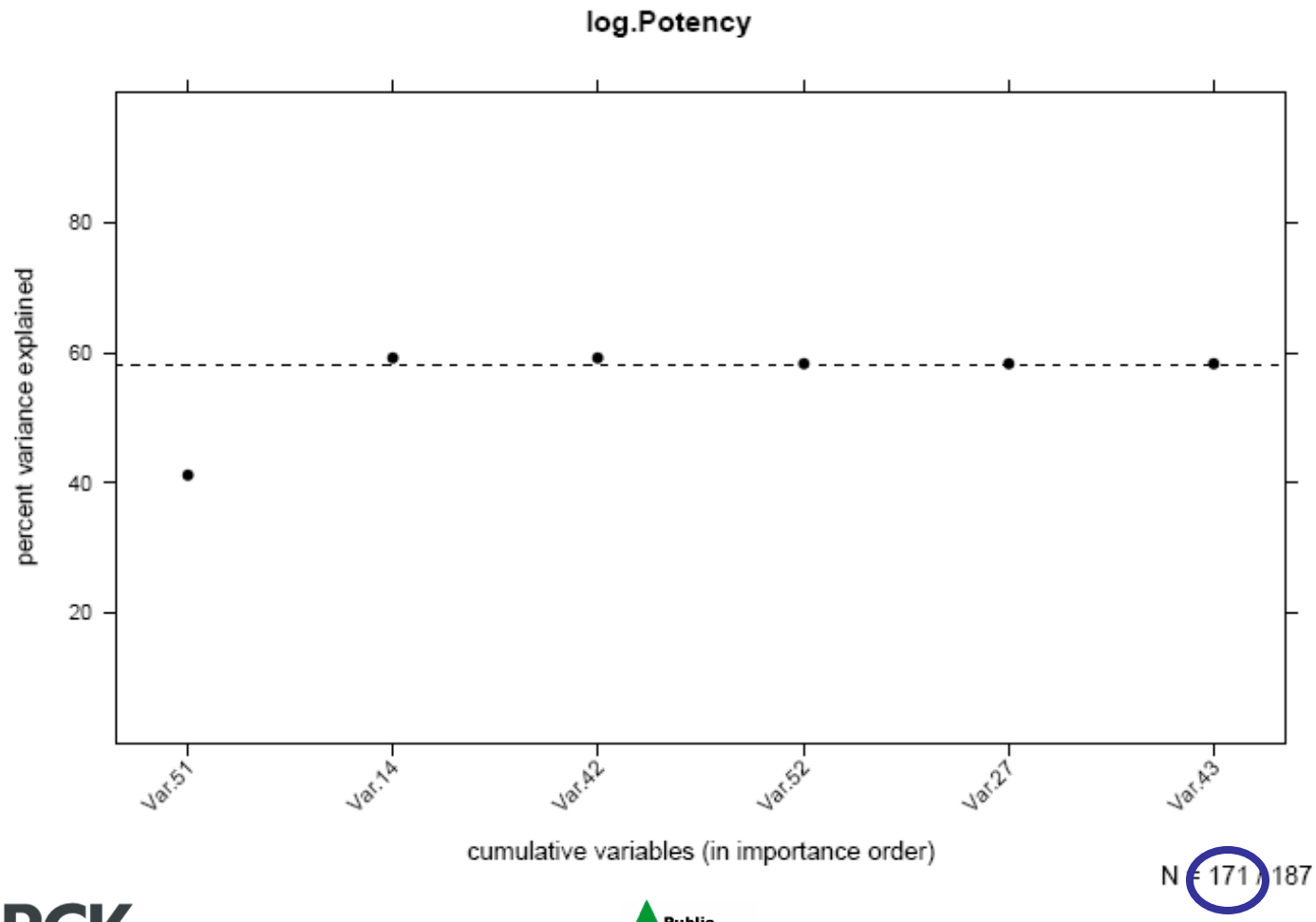


One of the second-tier variables (lot of a raw material) actually explains more variability

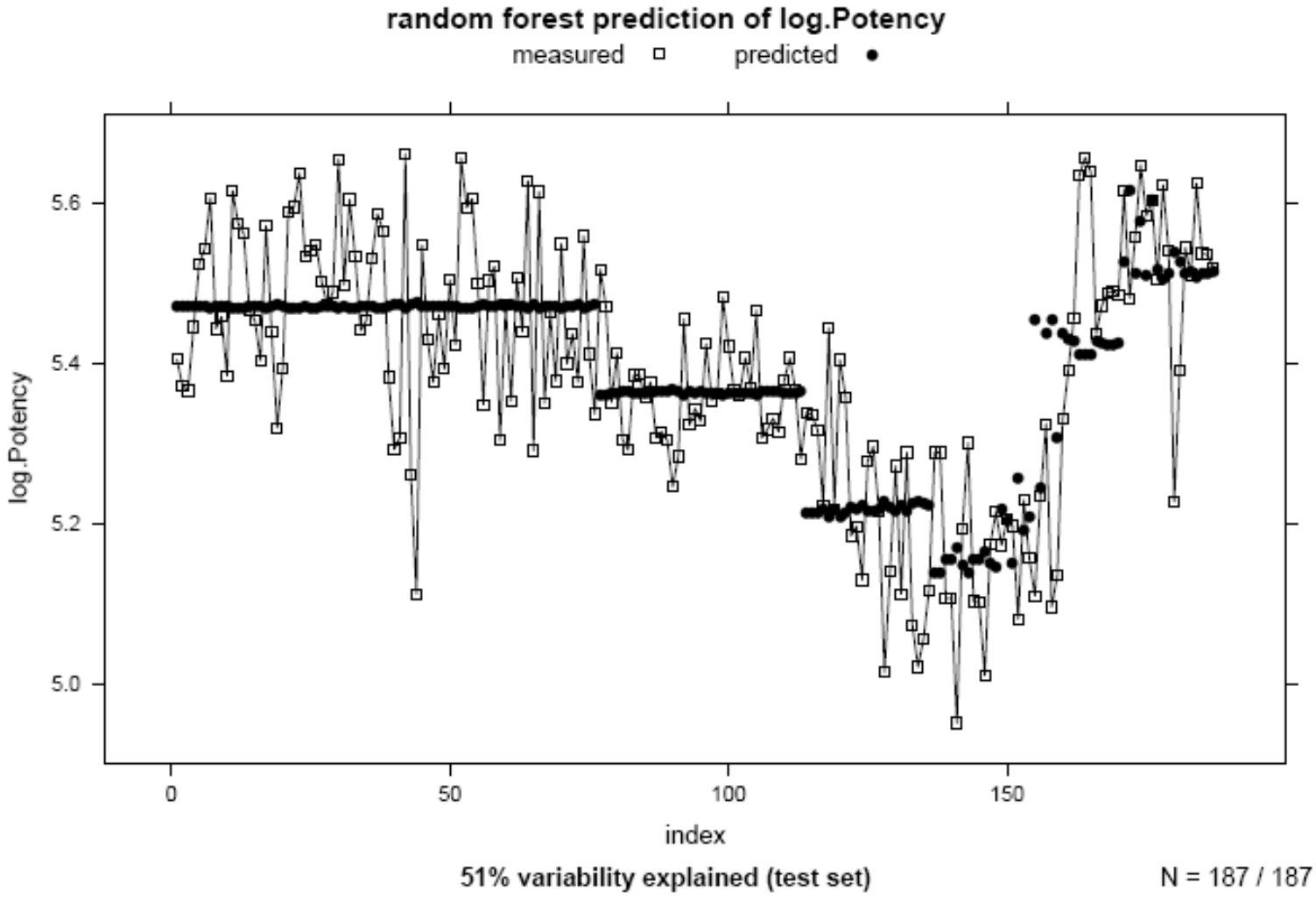


Use fewer variables, more cases

Just two variables – equipment and raw material lot – account for all variability we can explain



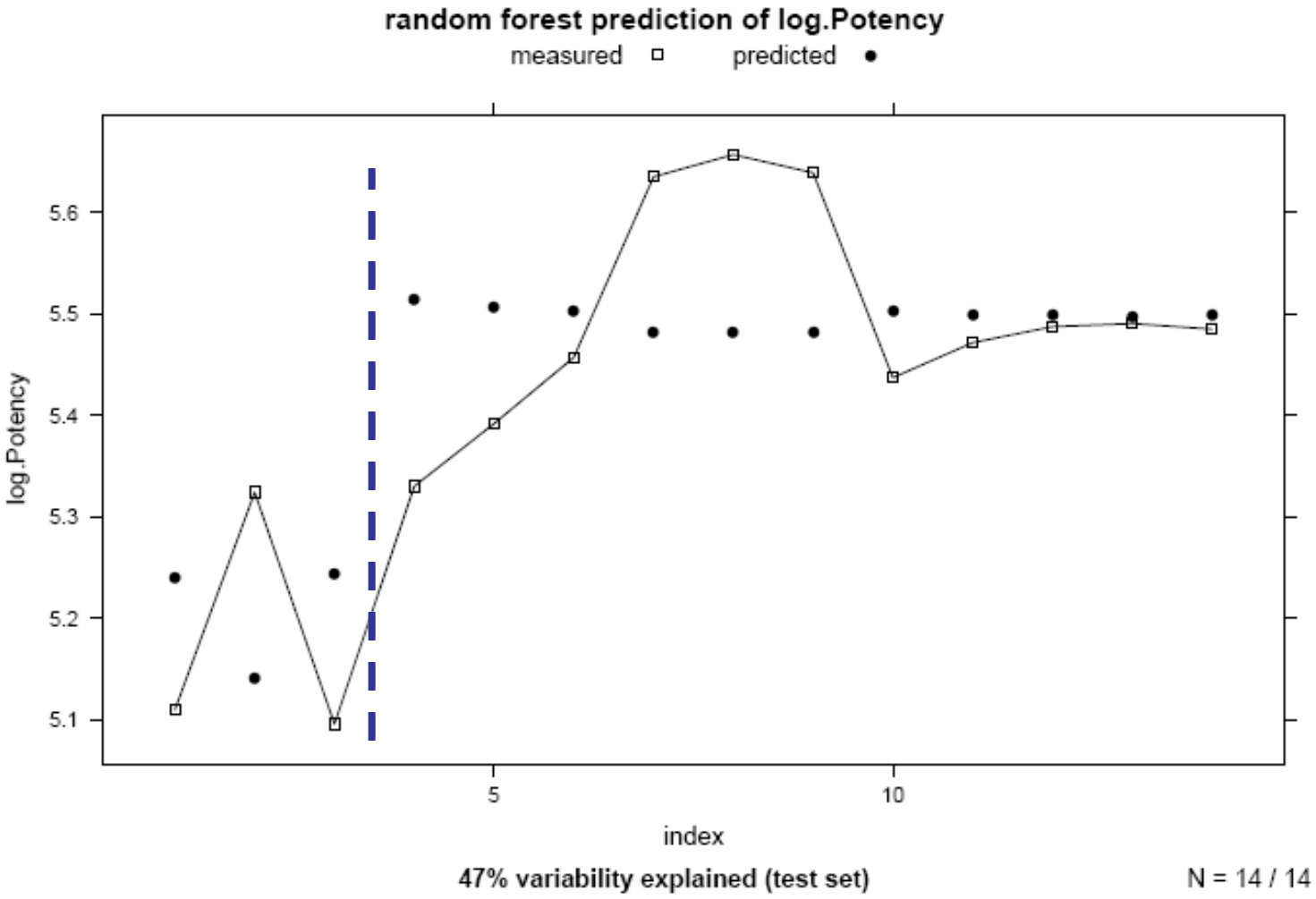
Raw Material Lot Number explains about half of variability



Equipment and raw material lots mostly changed at the same time

Equipment accounts for ~50% of variability for raw material lot 2115900 (accounting for almost all of its overall influence)

AND large time gap in the middle of these data



Conclusions: example 2

- We confirmed one main contributor accounting for about 50% of variability
  - And it's not the one that jumps out in univariate analysis
  - Could have wasted a lot of time chasing that down
- Measurement variability accounted for a small additional portion (5%) of variability
- No other variable currently in the data set explains a substantial amount of the remaining variability
  - Including some that are apparently explanatory if looked at alone, but are actually simply confounded with the most important one
- Further work can address:
  - Identifying what about the different lots of raw material was important
  - Finding additional variables that could account for the remaining variability

Random forest pros and cons

Pros

- Fast and simple to use
  - Available in R and JMP
- Can focus attention on most important variables
- Very high accuracy
- Handles non-linear relationships

Cons

- Something of a black box
  - Does not immediately reveal how each variable influences the outcome, unlike (say) linear regression
- Like most (all?) tree-based methods, dependent on the parameters in which the data are presented
  - Unlike, for example, principal component regression or partial least squares regression