

Random forest analysis in vaccine manufacturing

Matt Wiener
Dept. of Applied Computer Science &
Mathematics
Merck & Co.

Acknowledgements

Many people from many departments

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.

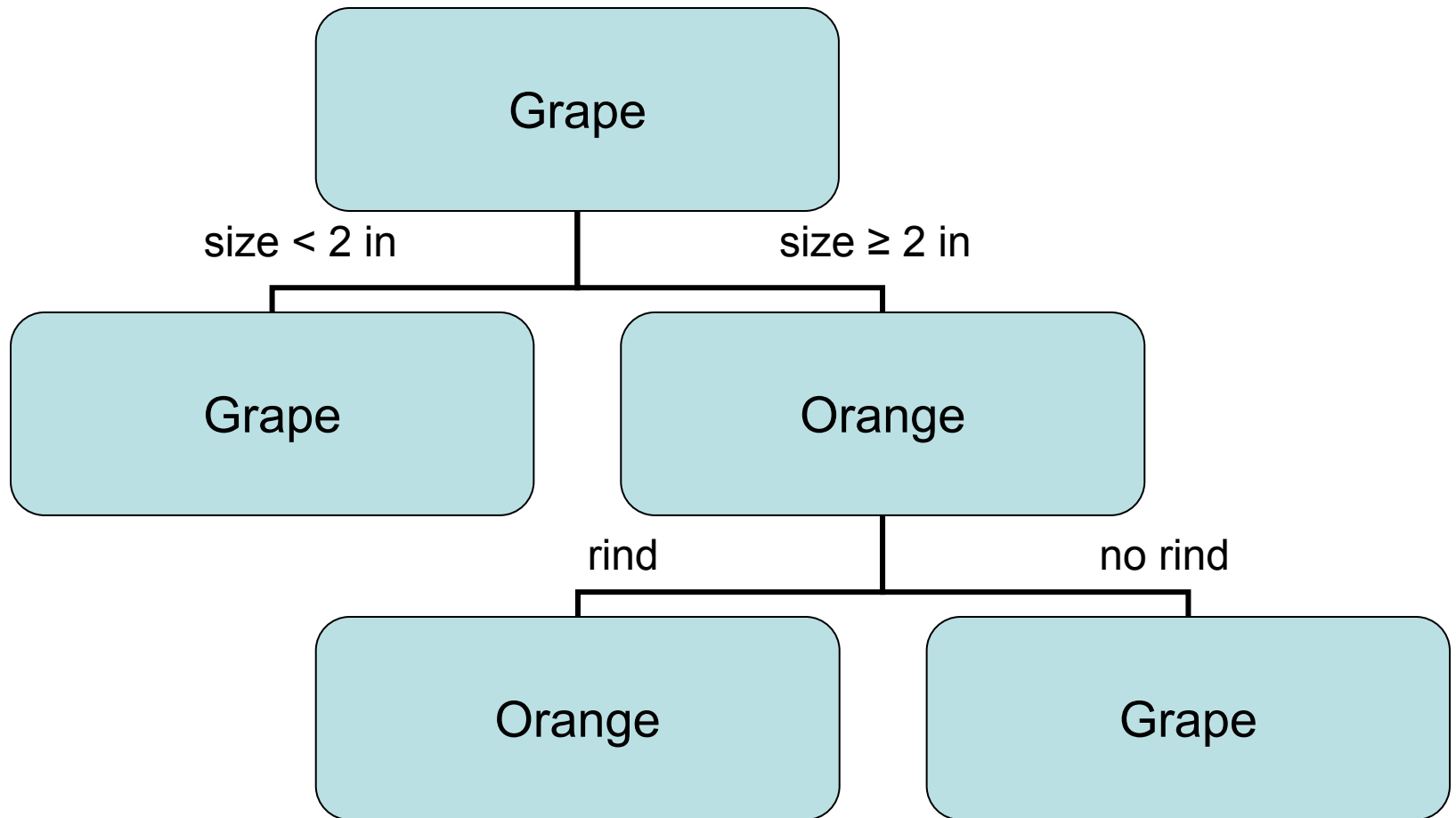
Goal: Prediction and explanation

- The goal is to be able to predict vaccine production (“yield”) based on critical process variables.
 - Many process variables are measured; we have to figure out which are the critical ones.
- If things go really well, we may (in collaboration with process experts) even be able to explain why those variables are the critical ones.
 - Then use that understanding to control production.

Outline

- Tree methods & random forests
- 2 examples
 - High yield investigation
 - Bulk potency investigation

A tree model for classification



Random forests

- Collections of classification or regression trees
 - Introduced by Breiman (1996, 2001)
- Trees vote on the classification or regression value
- Shown to give much better results than single trees
- Two kinds of randomness built into the process to make the trees different from one another

Growing a Forest

Training Data:

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

Growing a Forest

Training Data:
M1, M2, M3, M4, M5, M6, M7, M8, M9, M10

Draw random samples
with replacement

M1 M2 M2 M3 M4
M4 M5 M6 M9 M10

M1 M2 M3 M6 M7
M7 M9 M9 M10 M10

M1 M2 M3 M3 M4
M5 M5 M8 M8 M10

...

M2 M3 M4 M4 M5
M5 M5 M6 M7 M9

Growing a Forest

Training Data:
M1, M2, M3, M4, M5, M6, M7, M8, M9, M10

Draw random samples
with replacement

M1 M2 M2 M3 M4
M4 M5 M6 M9 M10

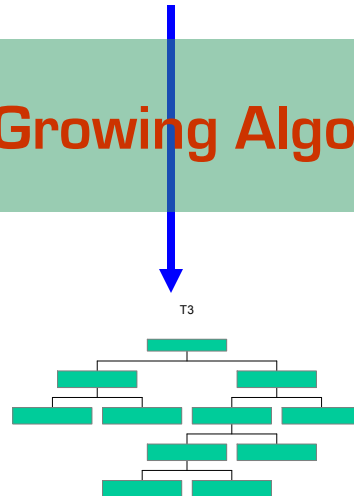
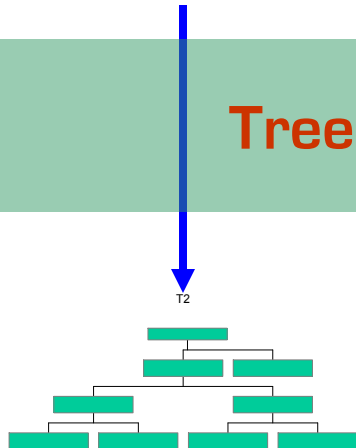
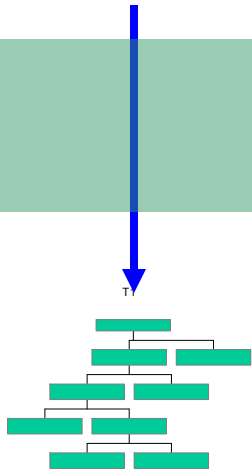
M1 M2 M3 M6 M7
M7 M9 M9 M10 M10

M1 M2 M3 M3 M4
M5 M5 M8 M8 M10

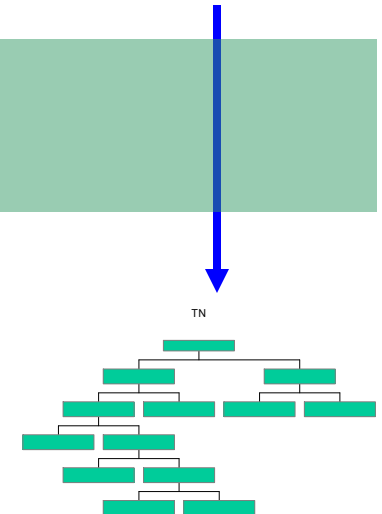
...

M2 M3 M4 M4 M5
M5 M5 M6 M7 M9

Tree Growing Algorithm



...



Semi-Random Splitting

Candidate Node:
10 A
10 B
Gini=0.5

X1
0: 6/6
1: 4/4
 $\Delta\text{Gini}=0.02$

X2
0: 7/6
1: 3/4
 $\Delta\text{Gini}=0.05$

X3
0: 9/1
1: 1/9
 $\Delta\text{Gini}=0.32$

X4
0: 6/5
1: 4/5
 $\Delta\text{Gini}=0.01$

X5
0: 4/6
1: 6/4
 $\Delta\text{Gini}=0.02$

X6
0: 9/4
1: 1/6
 $\Delta\text{Gini}=0.17$

Usual tree algorithm
chooses the best among
all: X3

Semi-Random Splitting

Candidate Node:
10 A
10 B
Gini=0.5

X1
0: 6/6
1: 4/4
 $\Delta\text{Gini}=0.02$

X2
0: 7/6
1: 3/4
 $\Delta\text{Gini}=0.05$

X3
0: 9/1
1: 1/9
 $\Delta\text{Gini}=0.32$

X4
0: 6/5
1: 4/5
 $\Delta\text{Gini}=0.01$

X5
0: 4/6
1: 6/4
 $\Delta\text{Gini}=0.02$

X6
0: 9/4
1: 1/6
 $\Delta\text{Gini}=0.17$

Random forest chooses
the best among a *random
subset*. X6

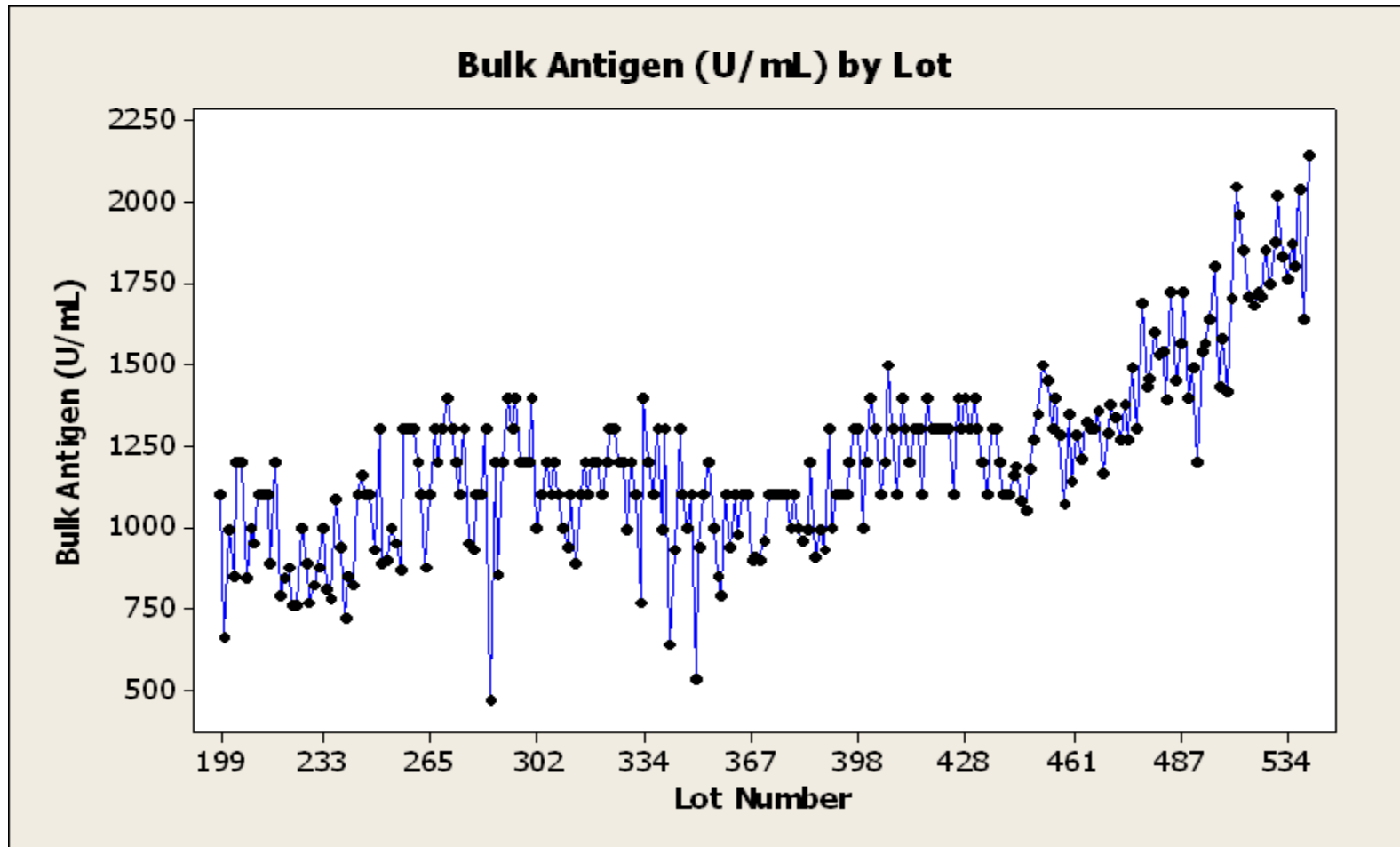
What random forests give us

- 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
 - And shown to be reasonably insensitive to those

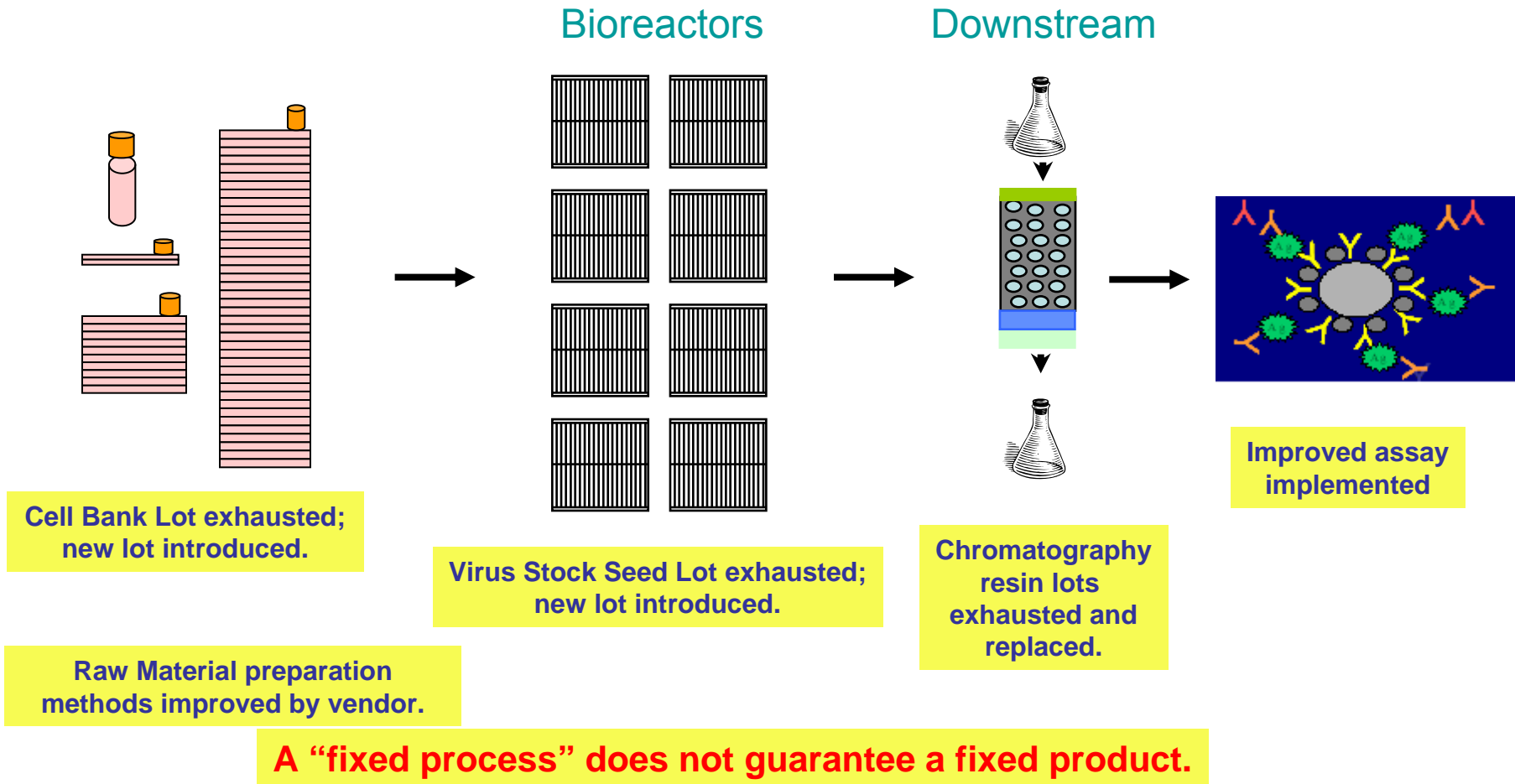
Example 1

High yield investigation

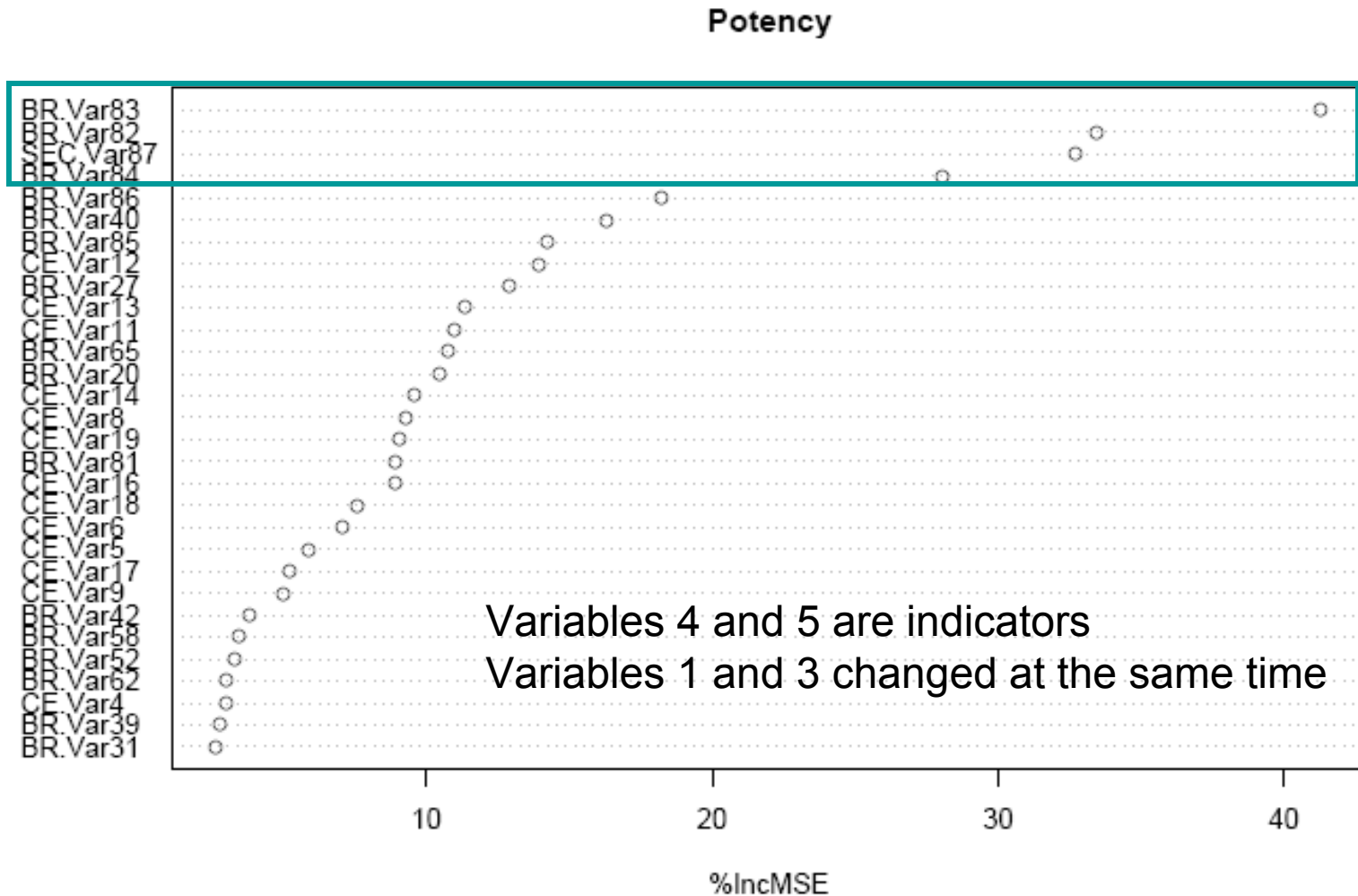
How to explain the increases? (simply diluting not good enough.)



Biologics mantra: “the product is the process”

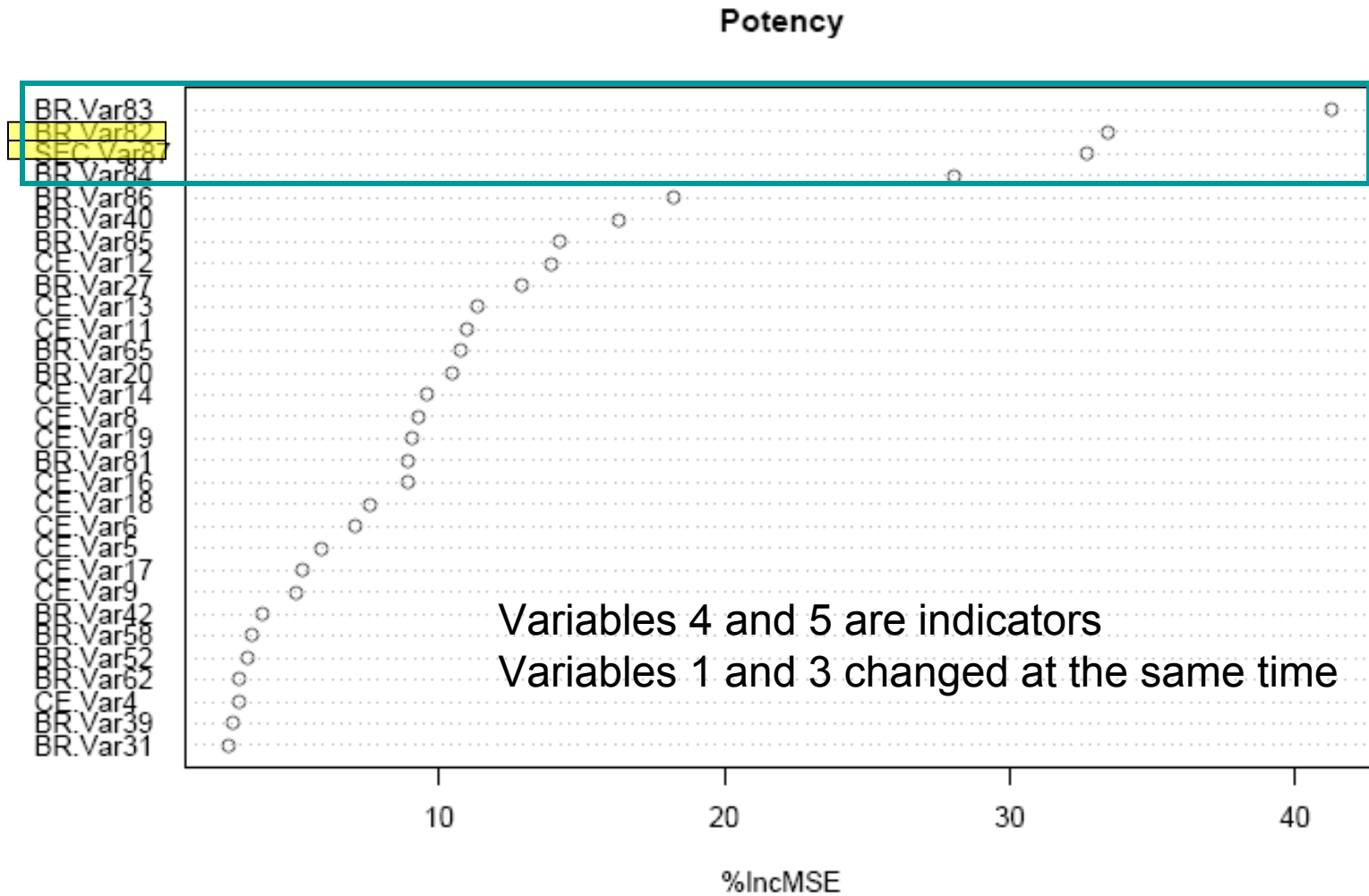


Variable Importance for predicting Potency by Random Forests

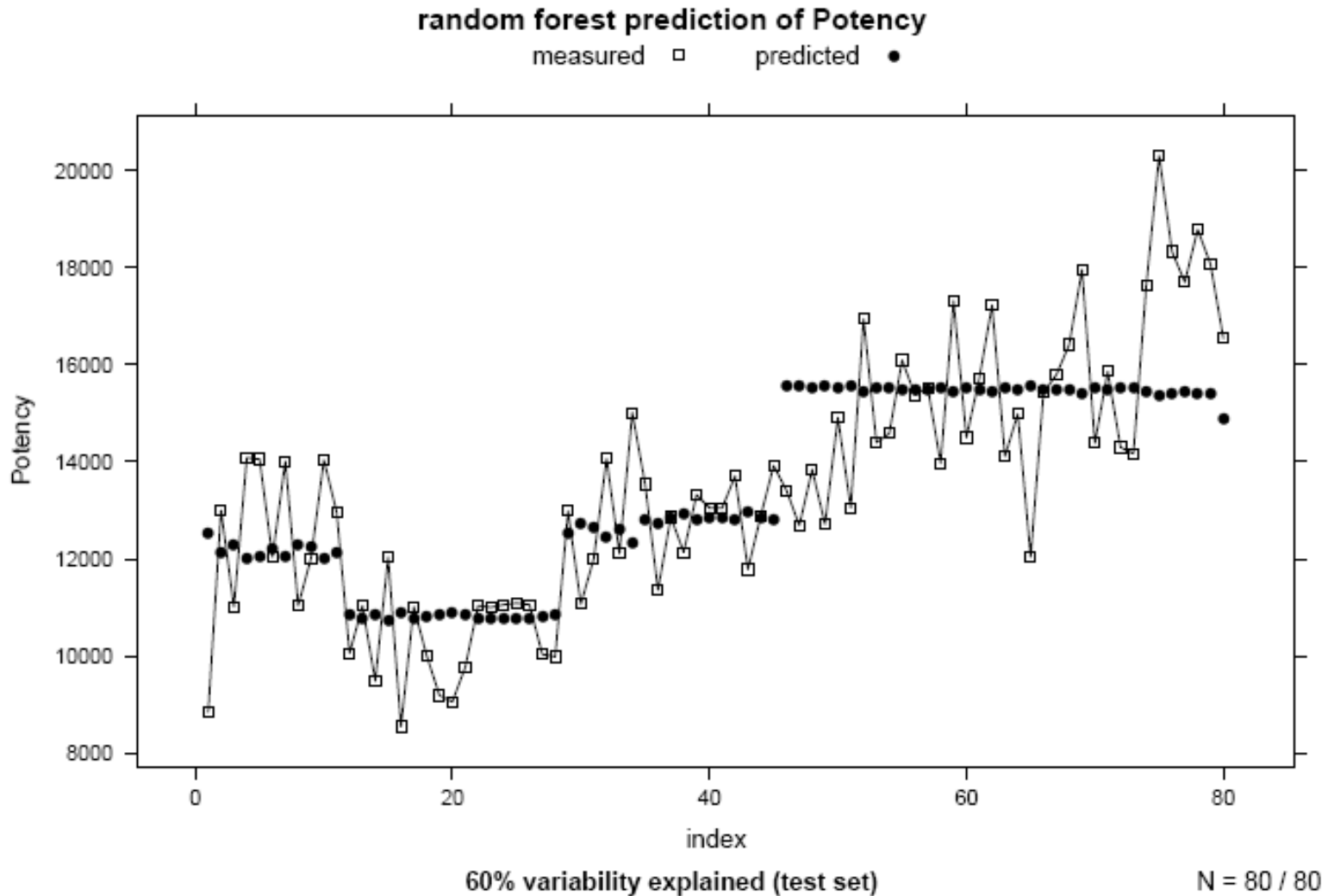


N = 64 / 80

Variable Importance for predicting Potency by Random Forests



These two root causes explain 60% of the variability.



Example 2

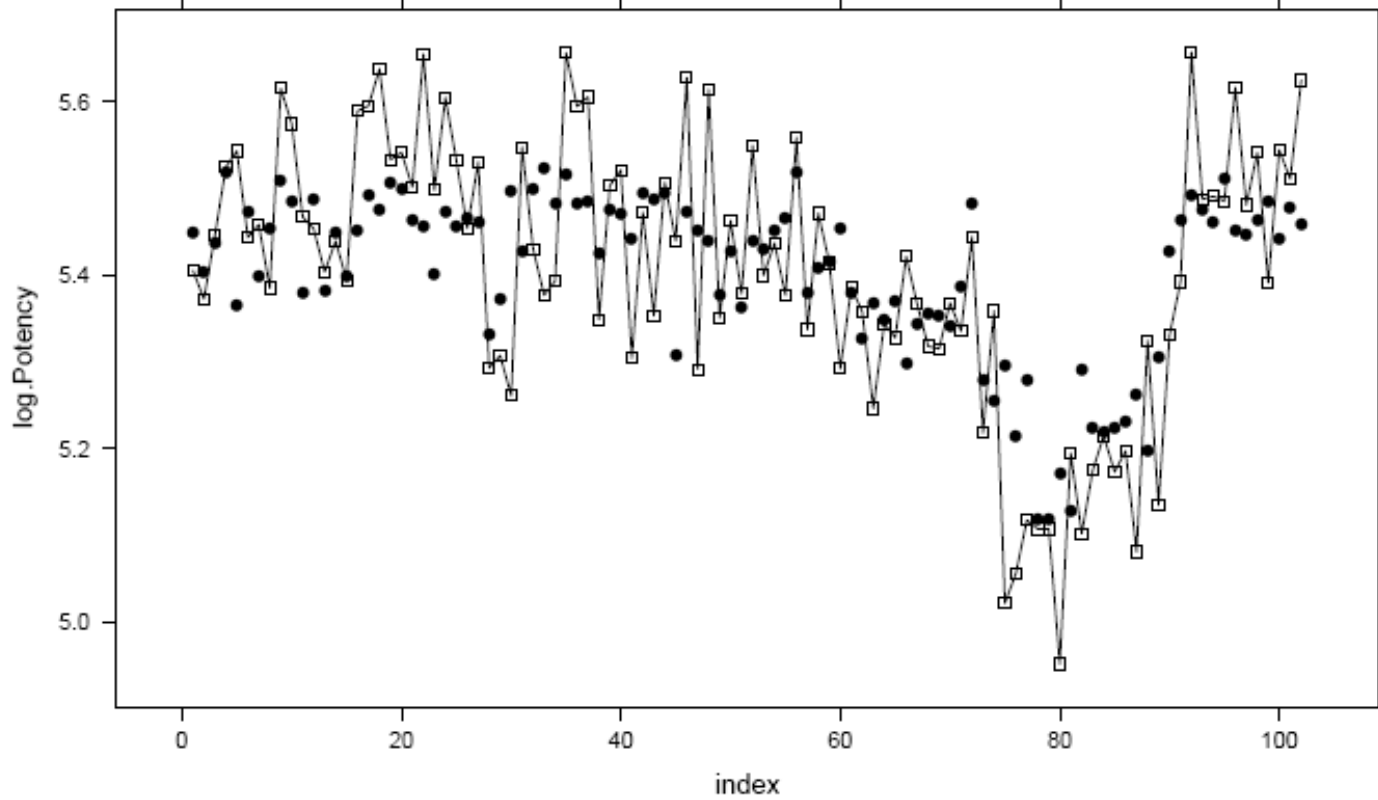
Bulk potency investigation

The problem

- Yield of bulk for a vaccine has been variable and sometimes not high enough, causing a shortfall in production.
- Hard to do experiments on vaccine processes
 - It takes many weeks to manufacture a lot of vaccine.
 - Experiments take a long time (and could potentially interfere with production).
 - Thus historical data analysis is especially valuable.

random forest prediction of log.Potency

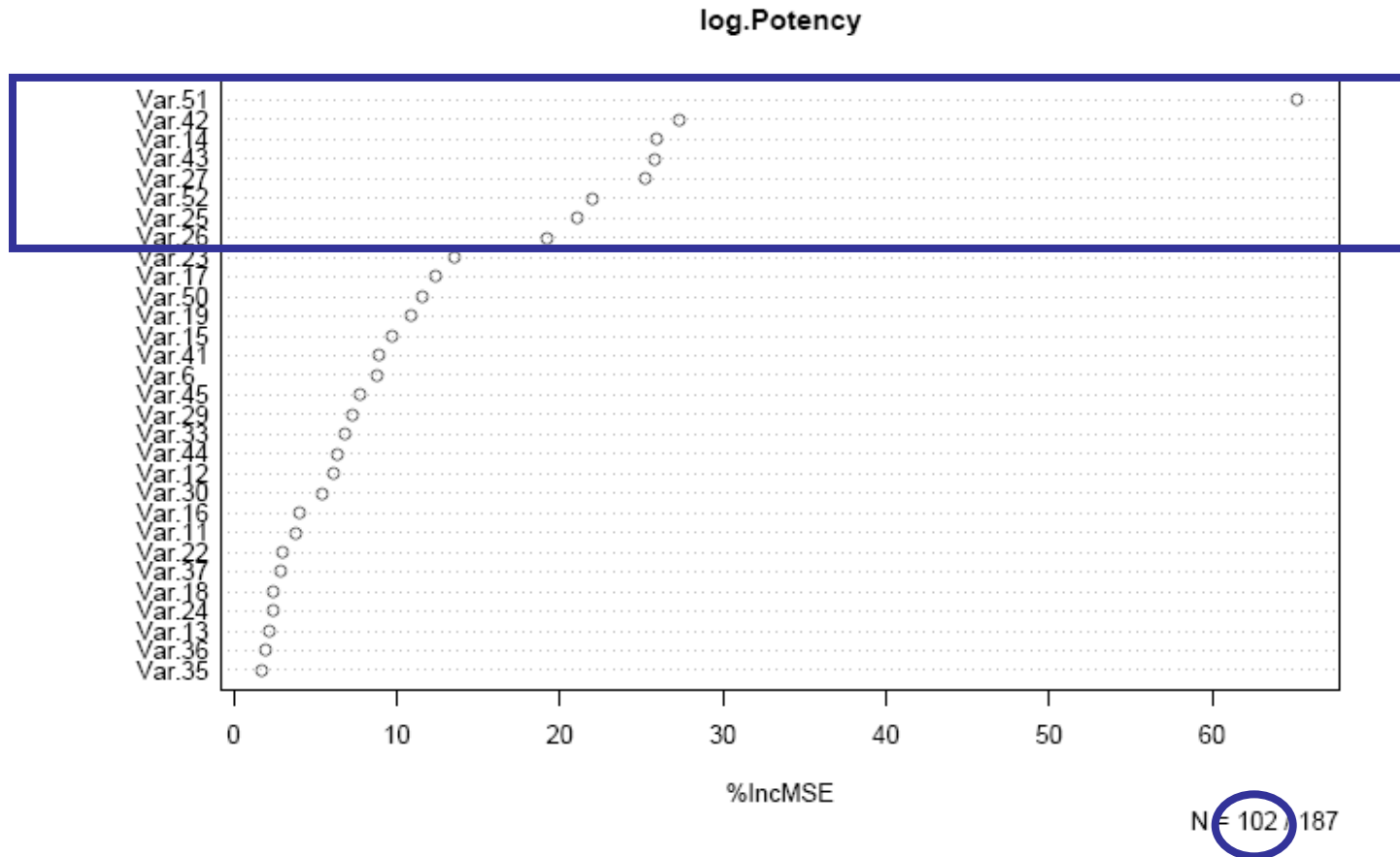
measured \square predicted \bullet



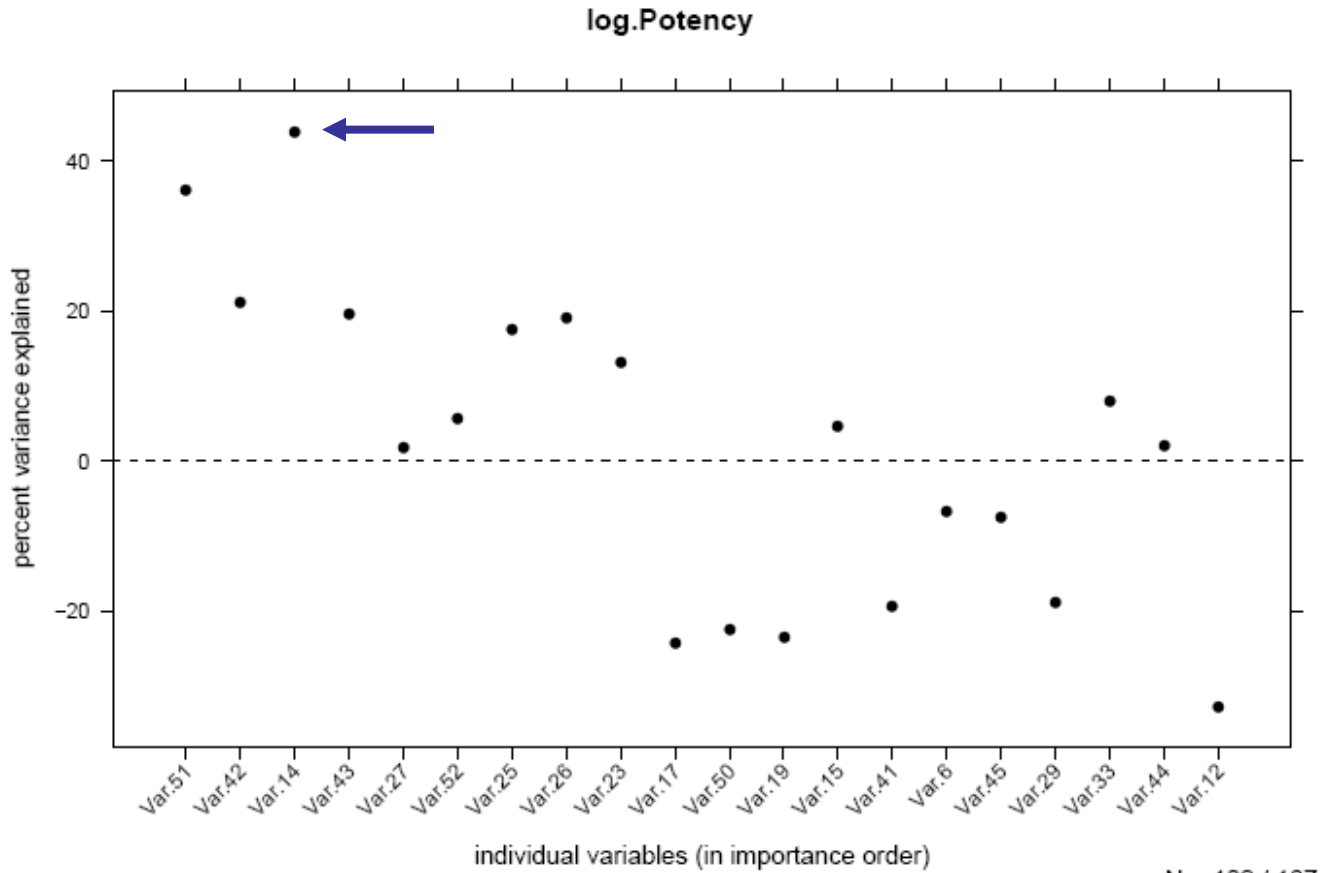
57% variability explained (test set)

N = 102 / 187

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



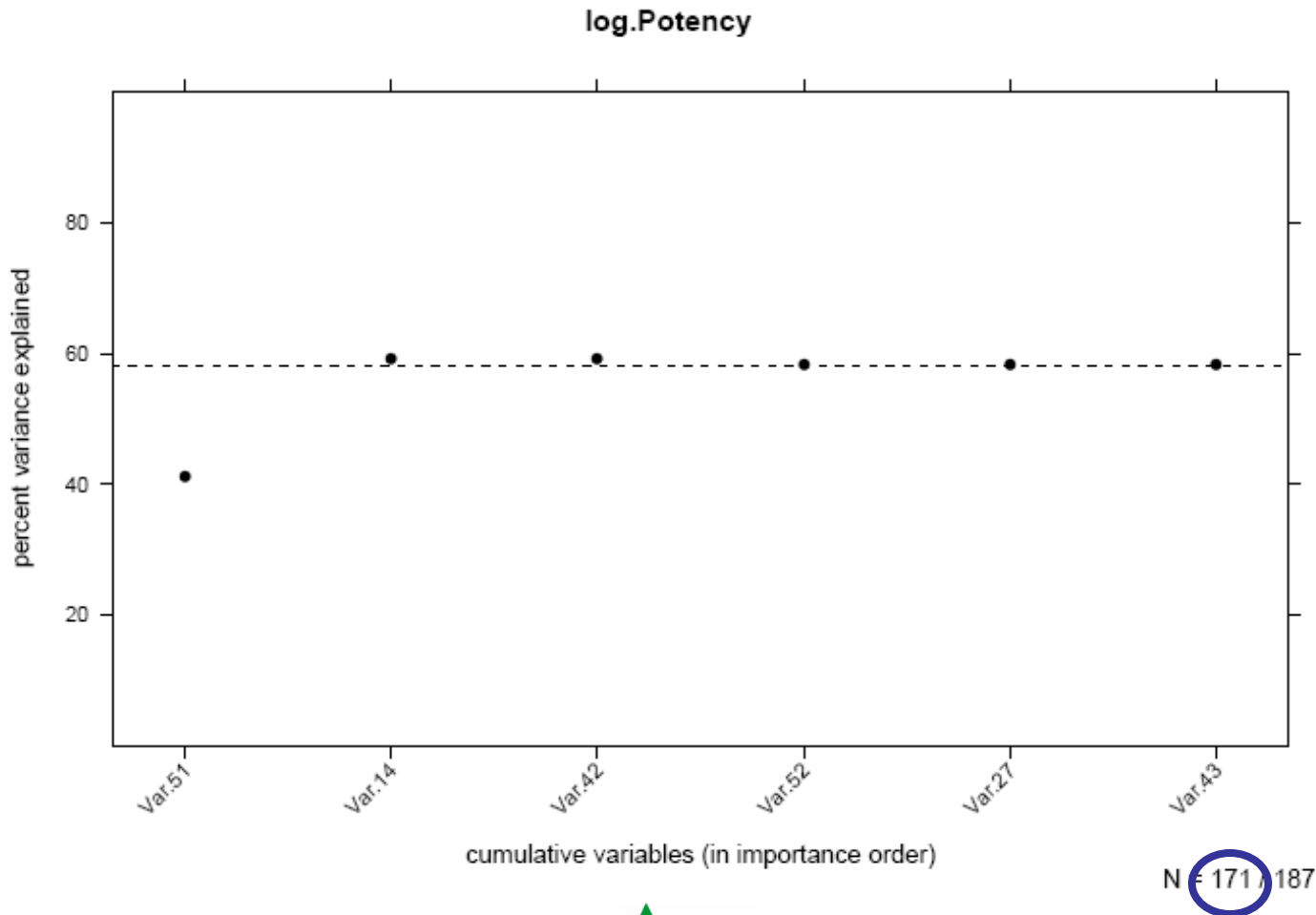
One of the second-tier variables (lot of a raw material) actually explains more variability



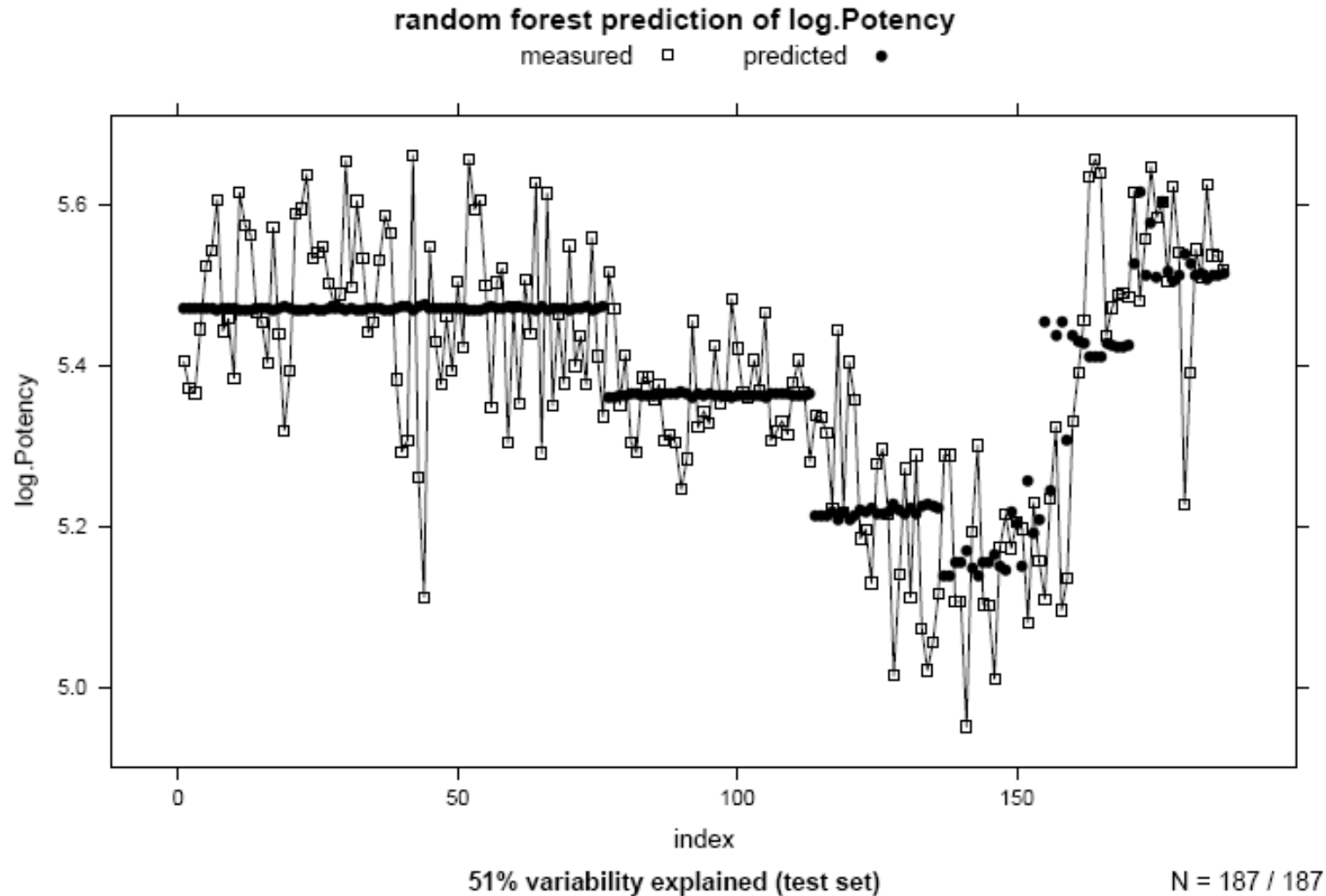
N = 102 / 187

Use fewer variables, more cases

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

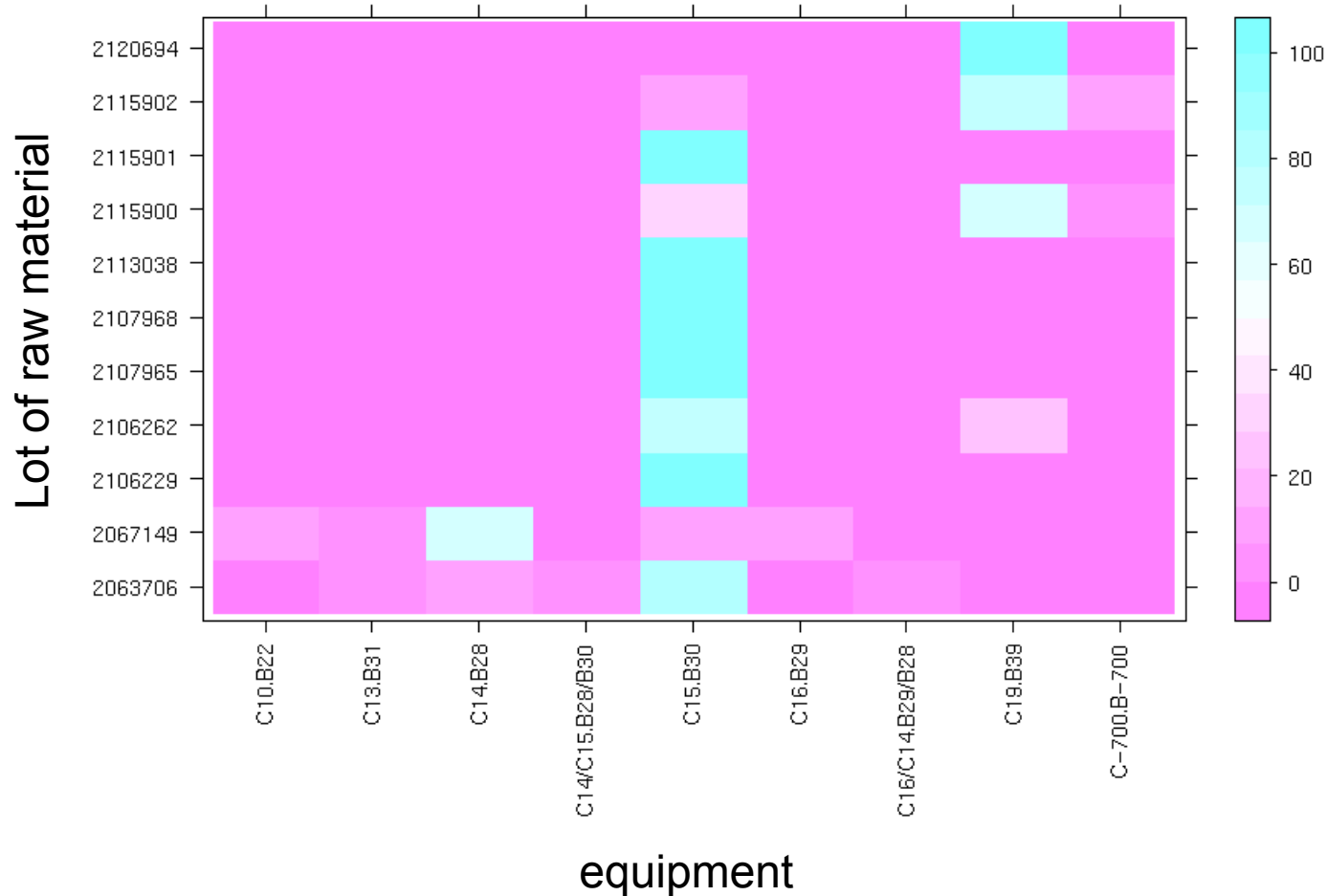


Raw Material Lot Number explains about half of variability



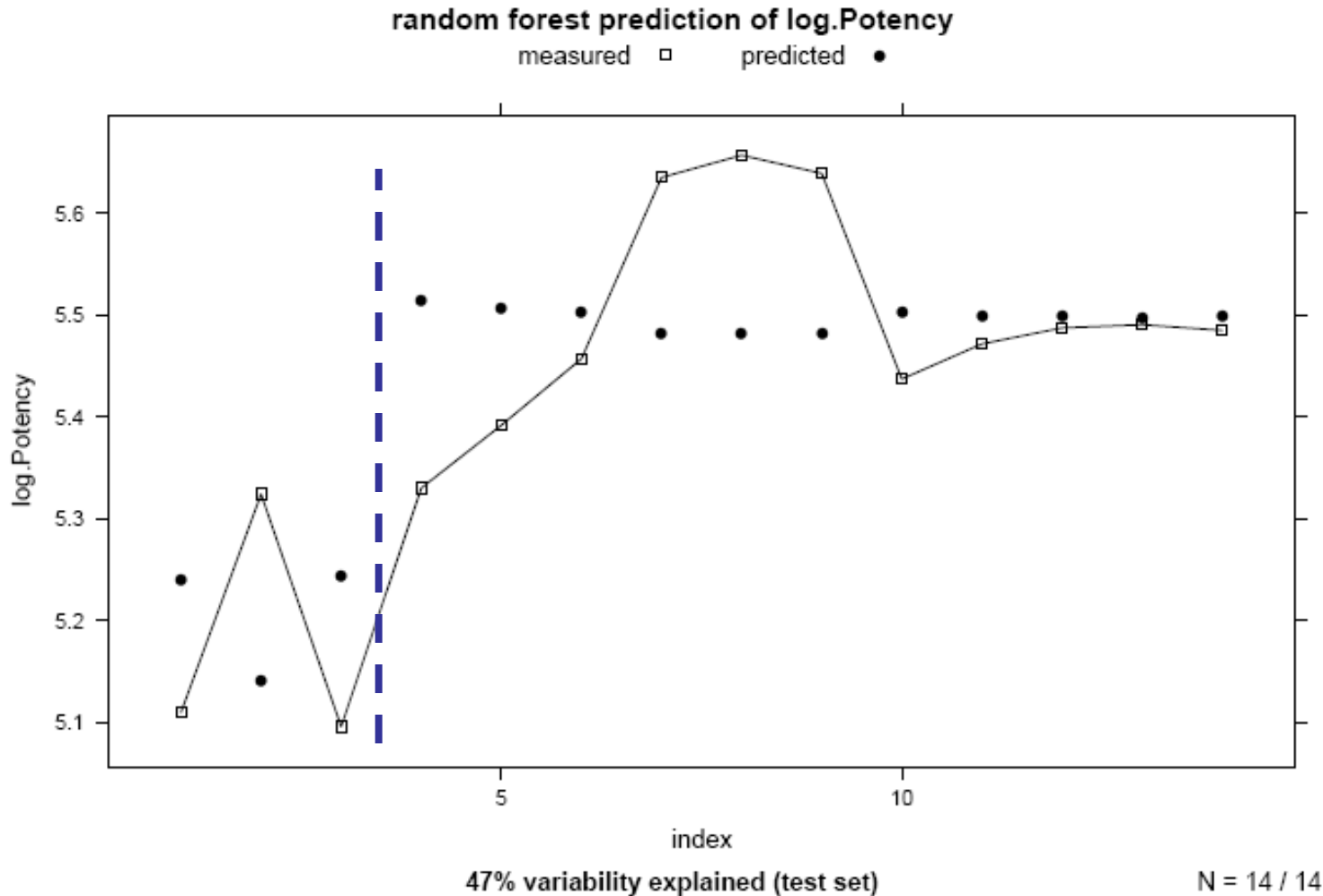
Equipment and raw material lots mostly changed at the same time

percent of lots processed on each piece of equipment
(by raw material lot)



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

- 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 is ongoing to identify additional variables contributing to variability.

Summary

- Random forest analysis can sift through a large number of variables to help find the most important ones.
- This analysis can help identify key process variables in vaccine manufacturing.
- randomForest in R makes these analyses easy to perform.