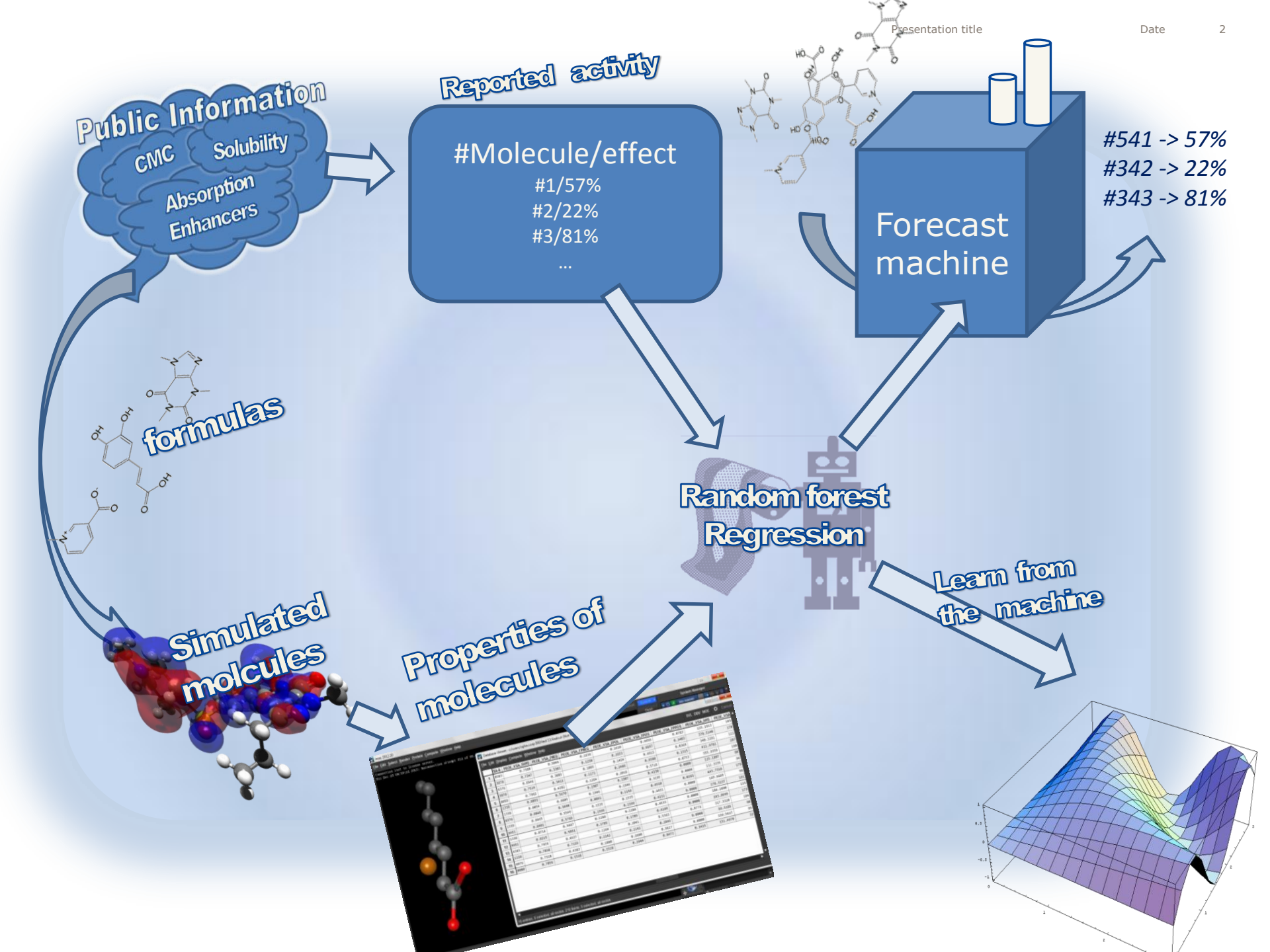
a package to visualize and comprehend the full curvature of random forests

Søren H Welling, Line K Clemmensen, Lars Hovgaard, Per Brockhof, Stephen T Buckley, Hanne F Refsgaard


Public Information
CMC
Solubility
Absorption
Enhancers
Reported activity
#Molecule/effect
#1/57%
#2/22%
#3/81%
...
Forecast machine
#541 -> 57%
#342 -> 22%
#343 -> 81%
formulas
Random forest Regression
Learn from the machine
Simulated molcules
Properties of molecules

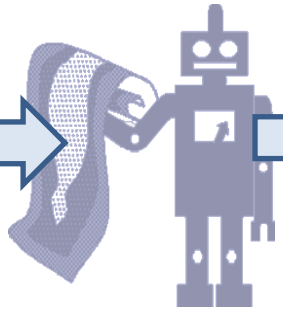
A supervised explorative analysis with RF

Unknown system

Sample data

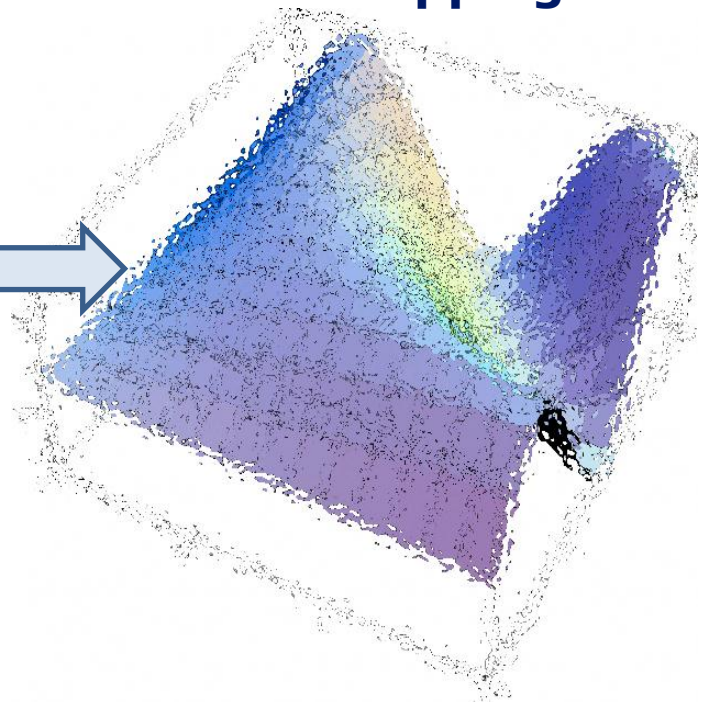
Sample No.	Thickness (cm)	Temperature (°C)	Concentration (g/L)
1	2.1740228	82	0.066
2	1.8774501	77	0.071
3	1.8774704	77	0.072
4	1.9762727	79	
5	2.0266303	80	
6	2.0994529	81	
7	1.9468132	78	0.067
8	1.8972298	77	0.071
9	1.9169798	77	0.07
10	2.0692626	80	0.066
11	2.1292363	82	0.067
12	2.0479427	80	0.067
13	2.0479598	80	0.069
14	1.8972463	77	0.071
15	1.8774795	77	0.066

Train RF

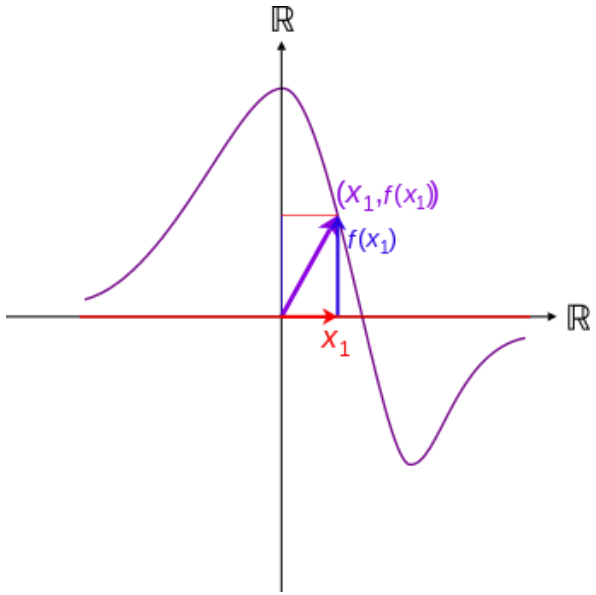


Choose target!

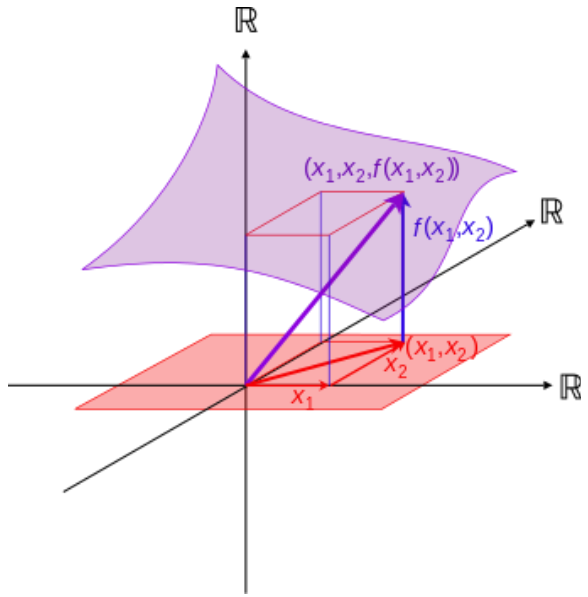
Look at mapping



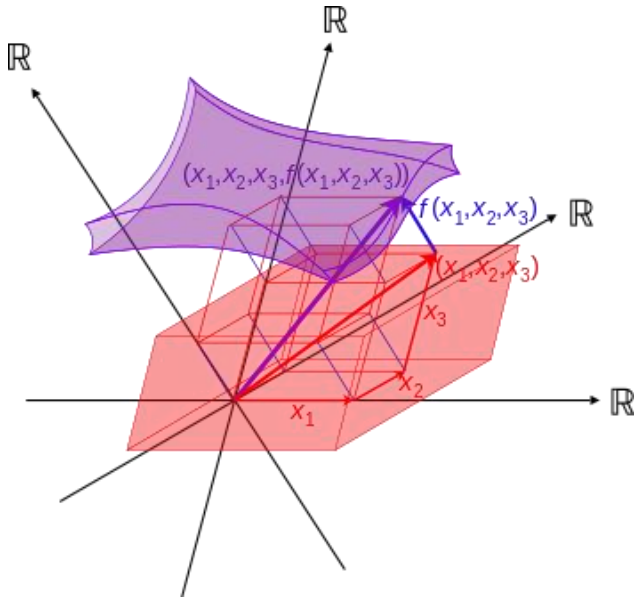
Visualizing mapping



2D



3D



4D

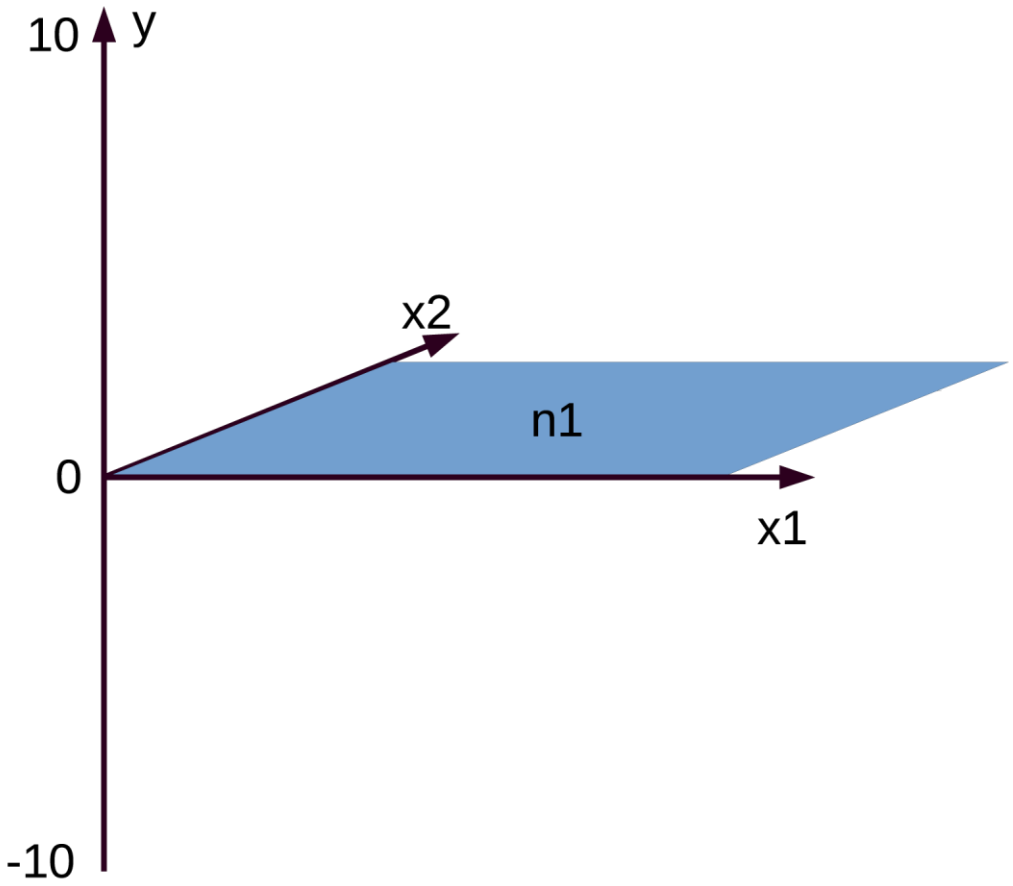
Dimensions needed = nvar + 1

Dimensions needed = nvar + nclasses -1

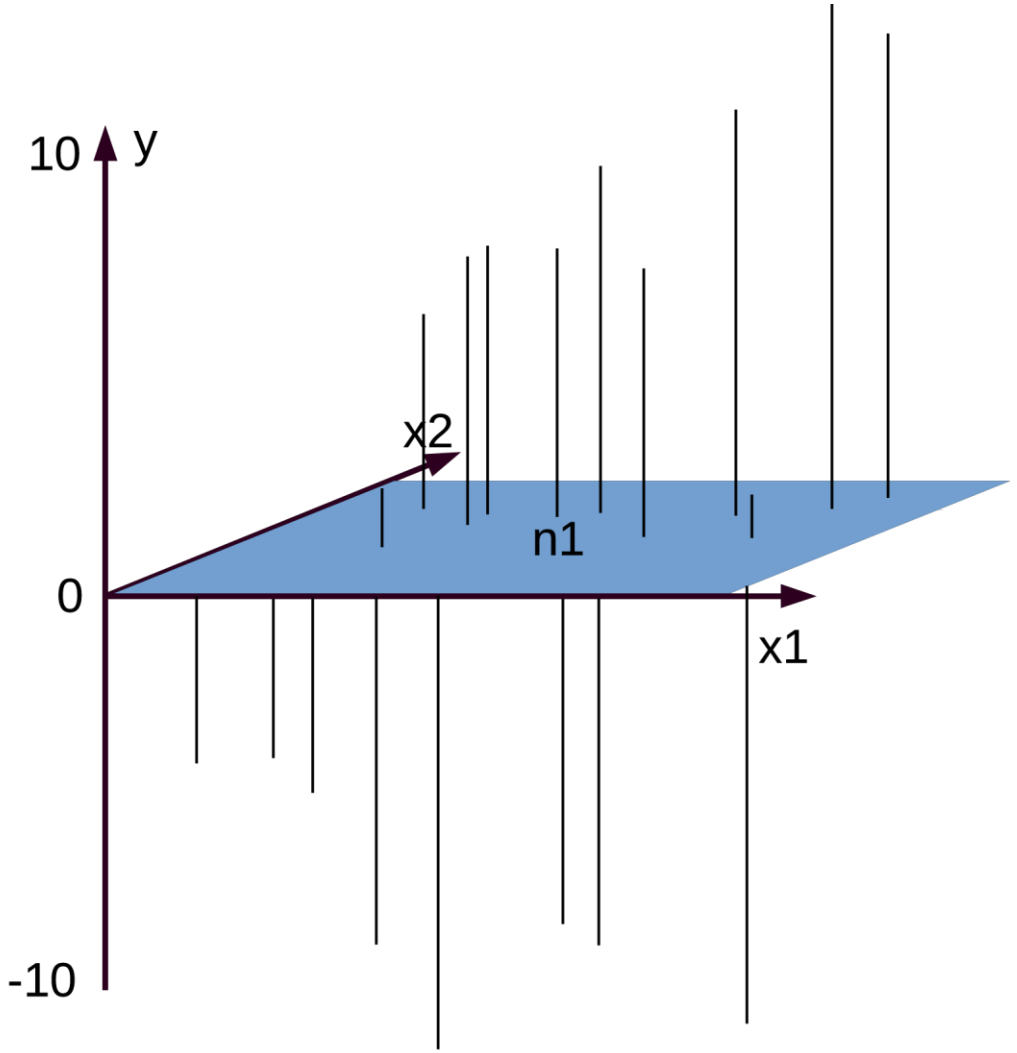
Part 2 – how forestFloor works

- Decision trees for dummies
- Feature contributions
- It is all about context!

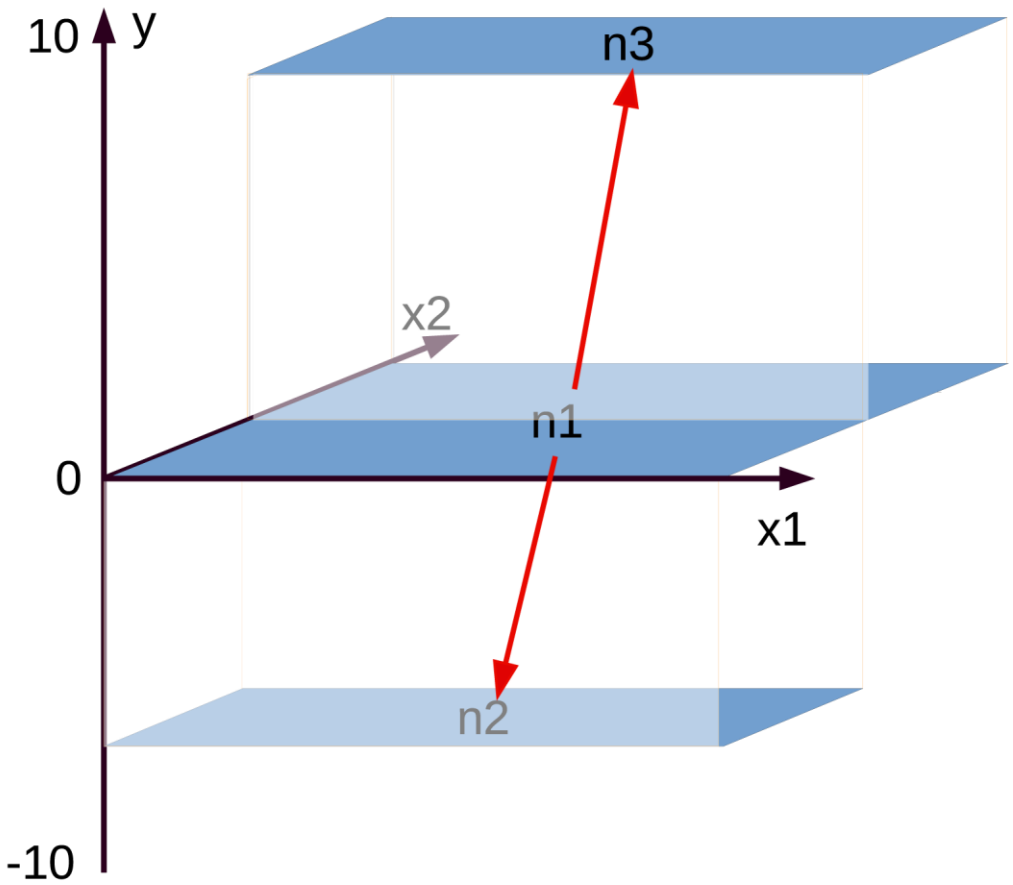
Two variables, one target, one grand mean(n1)



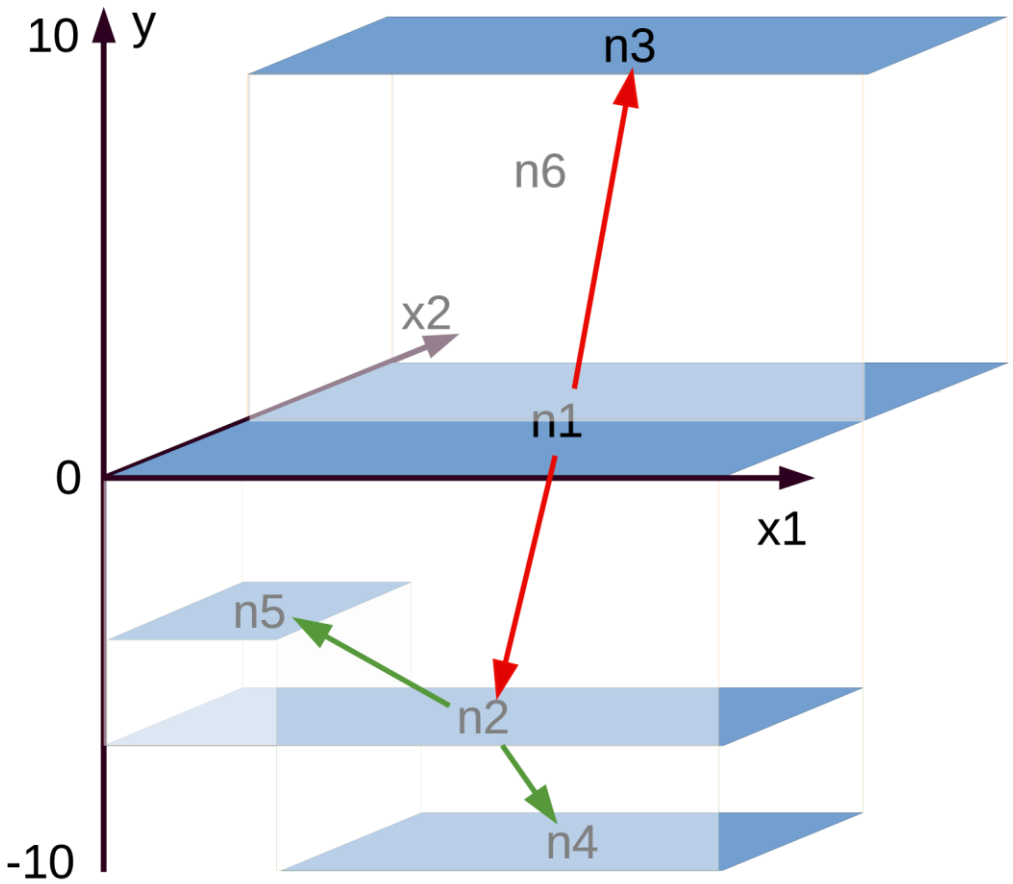
… and residuals of the grand mean



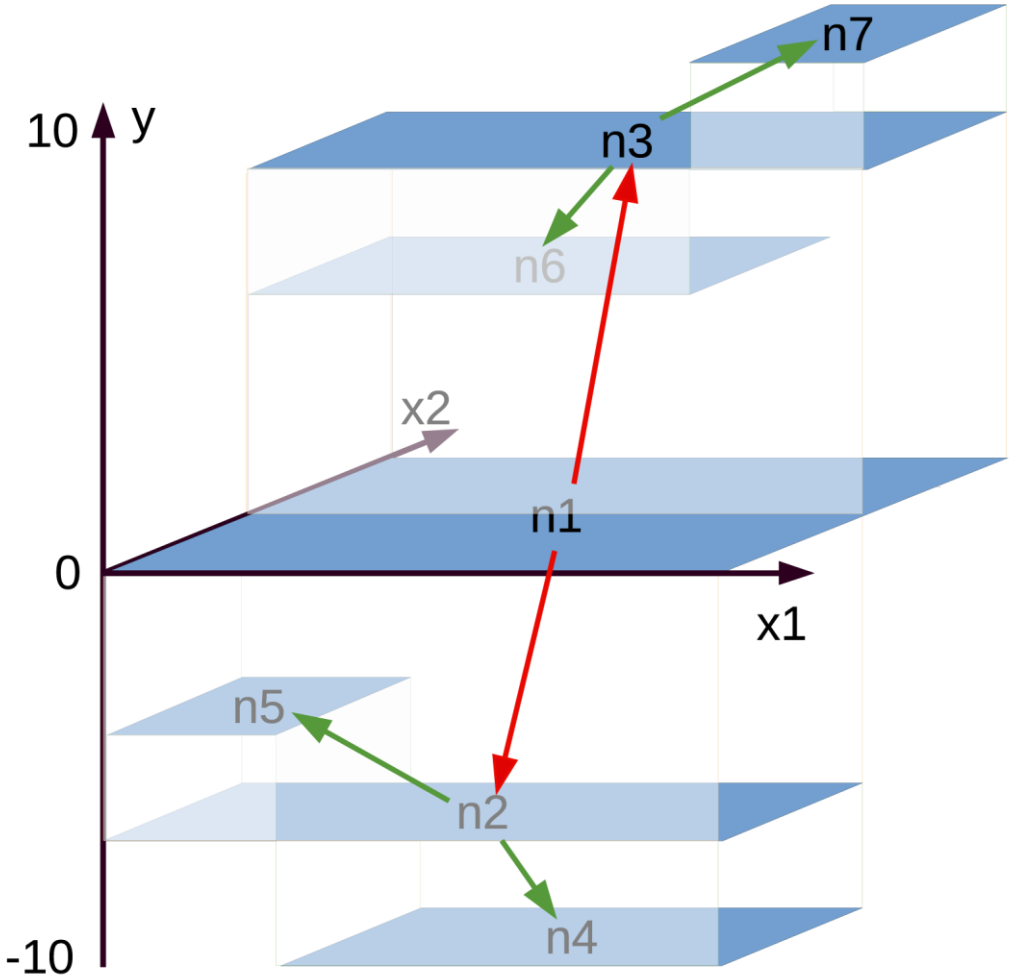
Best split of n1 by x2...



Best split of n2...



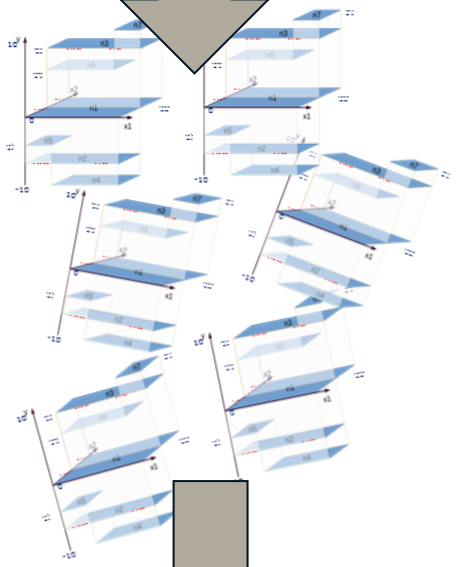
Best split of n3



Random forest

Training data

Bootstrap and train many trees



Forest

Aggregate models by voting

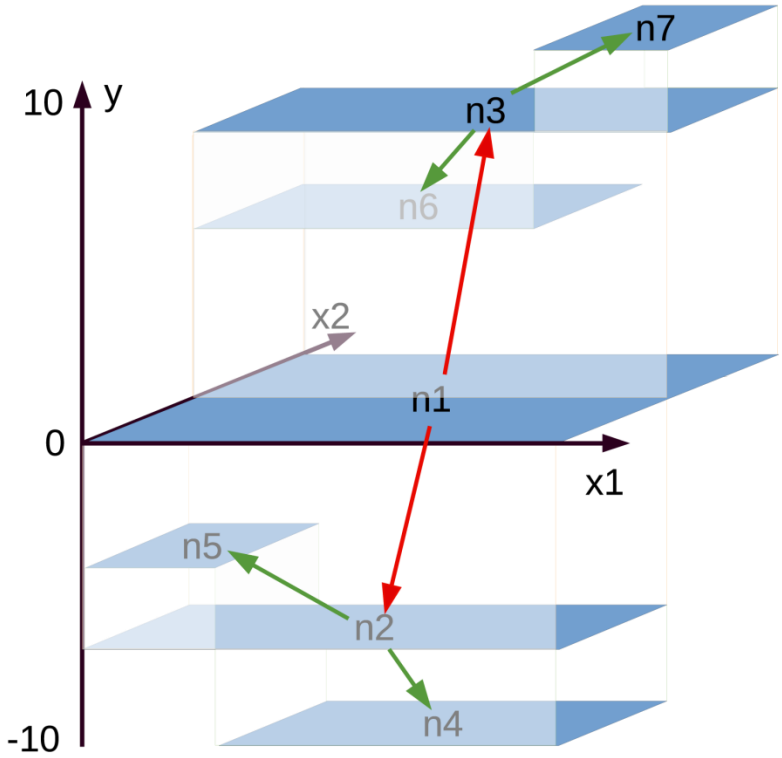
Robust predictions

F, Feature contributions

F is a decomposition of the predicted targets…

$$\hat{y}_i = \bar{u}_0 + \sum_{j=1}^{nvar} F_{ij} = \bar{u}_0 + F_{i1} + F_{i2}$$

y target
i^{th} sample
j^{th} variable out of nvar



…how to compute F for a whole forest

$$F_{ij} = \frac{1}{ntree} \sum_{k=1}^{ntree} \sum_{l=1}^{nnodes_{ik}} L_{i_{j_{k_l}}}$$

k^{th} tree
l^{th} node in the unique pathway of i^{th} sample in the k^{th} tree

Part 3

Plot feature contributions

How to learn this hidden function?

$$y = f(X) =$$
$$(x_1)^2 + sin(x_2) + (x_3 \cdot x_4) + 0x_5 + 0x_6$$

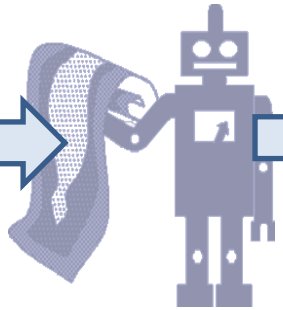
A explorative analysis with RF

Unknown system

Sample data

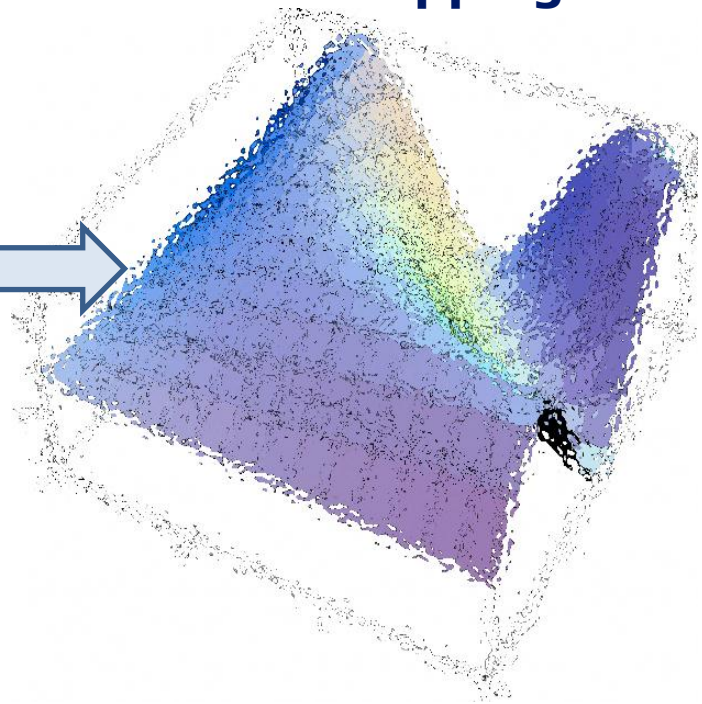
Sample No.	Thickness (cm)	Temperature (°C)	Concentration (g/L)
1	2.1740228	82	0.066
2	1.8774501	77	0.071
3	1.8774704	77	0.072
	1.9762727	79	
	2.0266303	80	
	2.0994529	81	
	1.9468132	78	0.067
8	1.8972298	77	0.071
9	1.9169798	77	0.07
10	2.0692626	80	0.066
11	2.1292363	82	0.067
12	2.0479427	80	0.067
13	2.0479598	80	0.069
14	1.8972463	77	0.071
15	1.8774795	77	0.066

Train RF



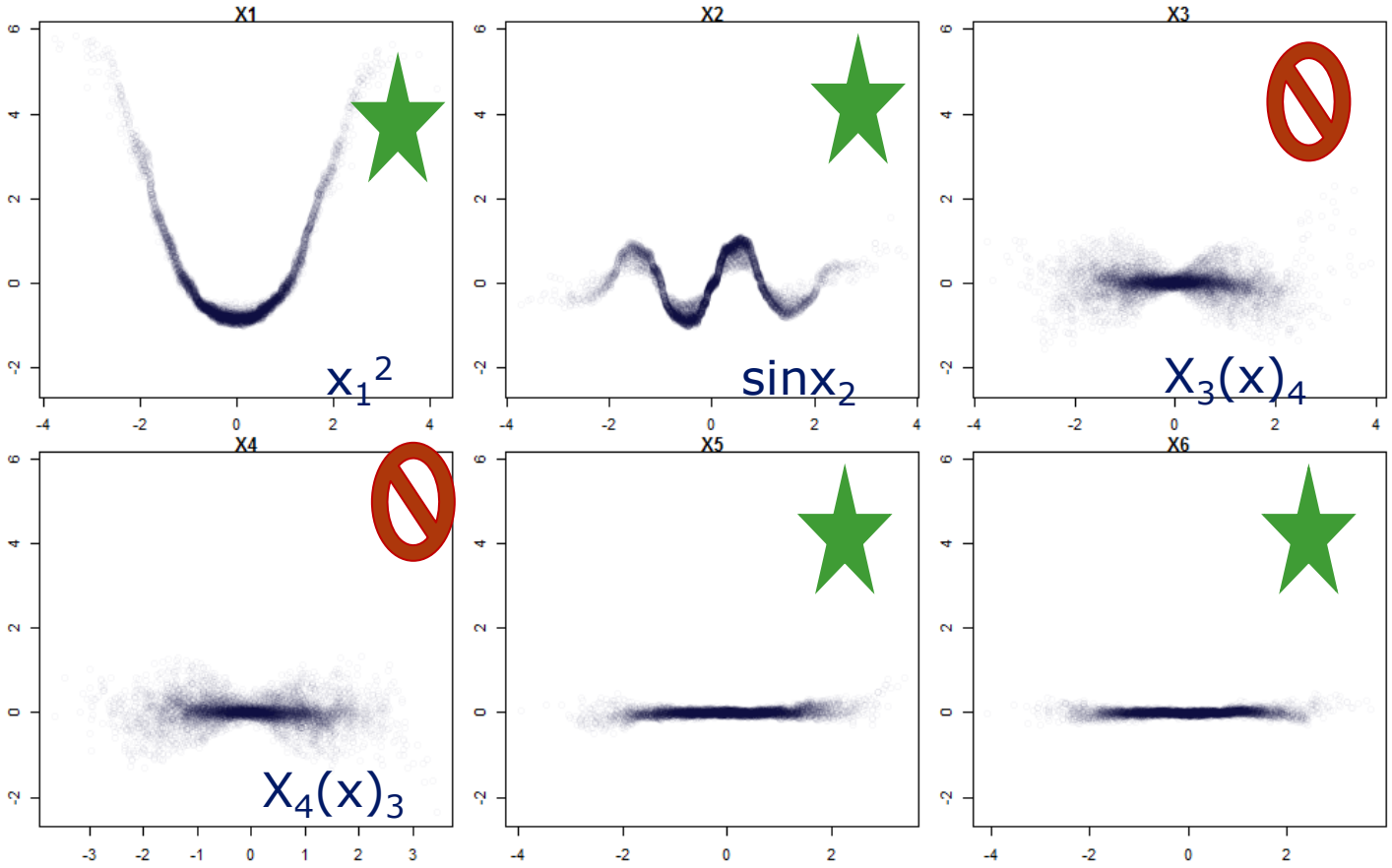
Choose target!

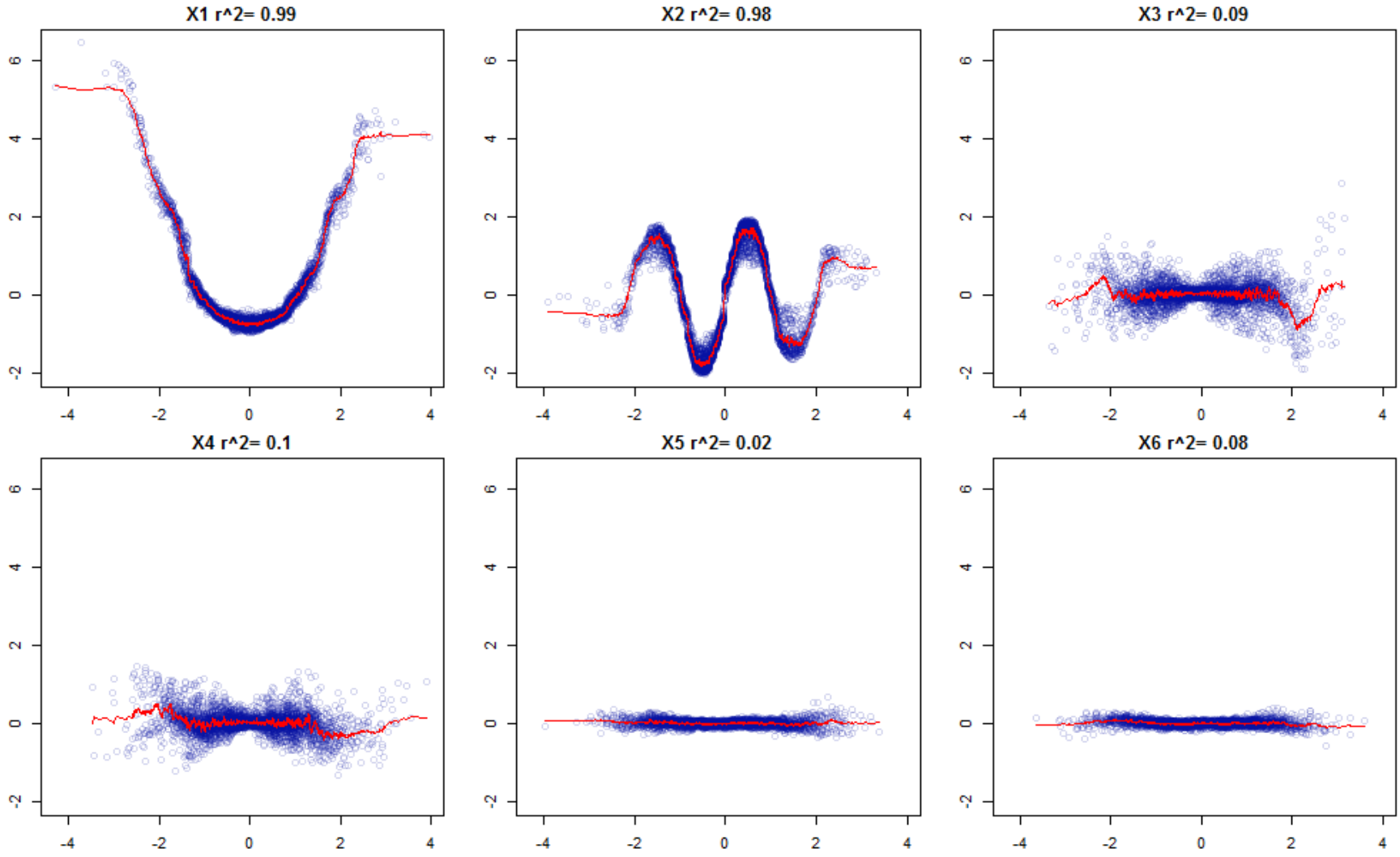
Look at mapping

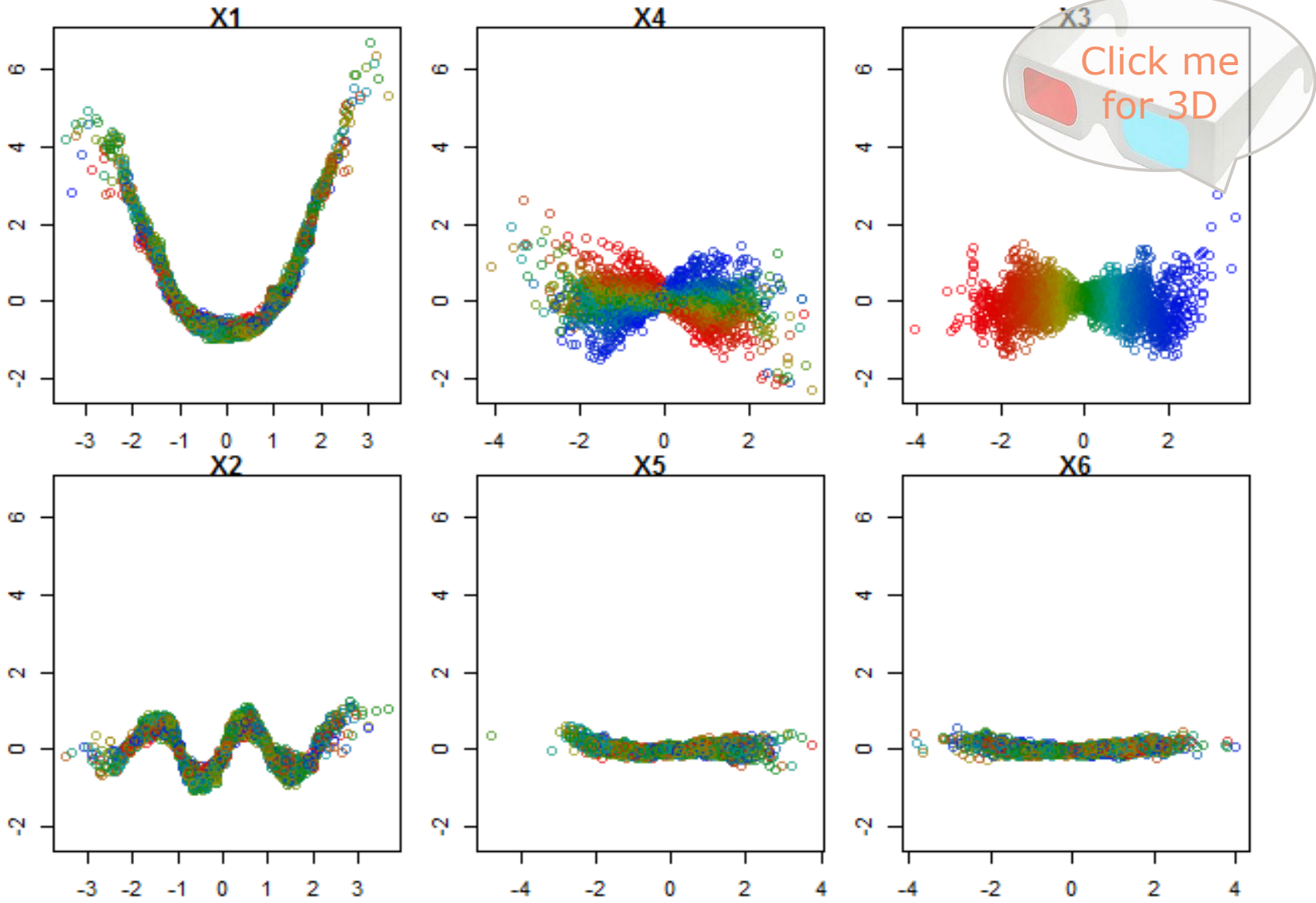


The solution is...

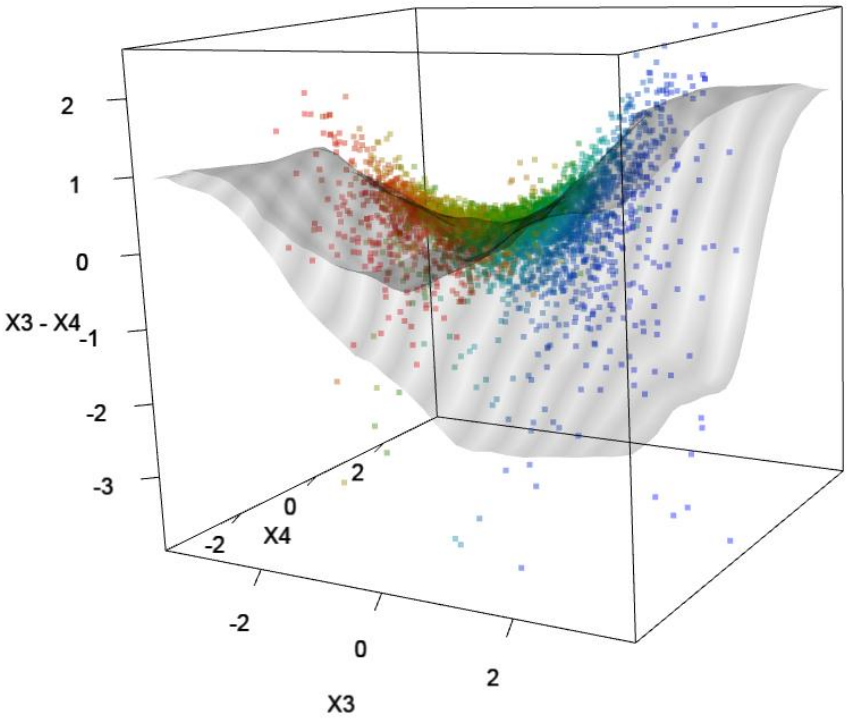
$$y = f(X) = (x_1)^2 + sin(x_2) + (x_3 \cdot x_4) + 0x_5 + 0x_6 + \text{noise}$$

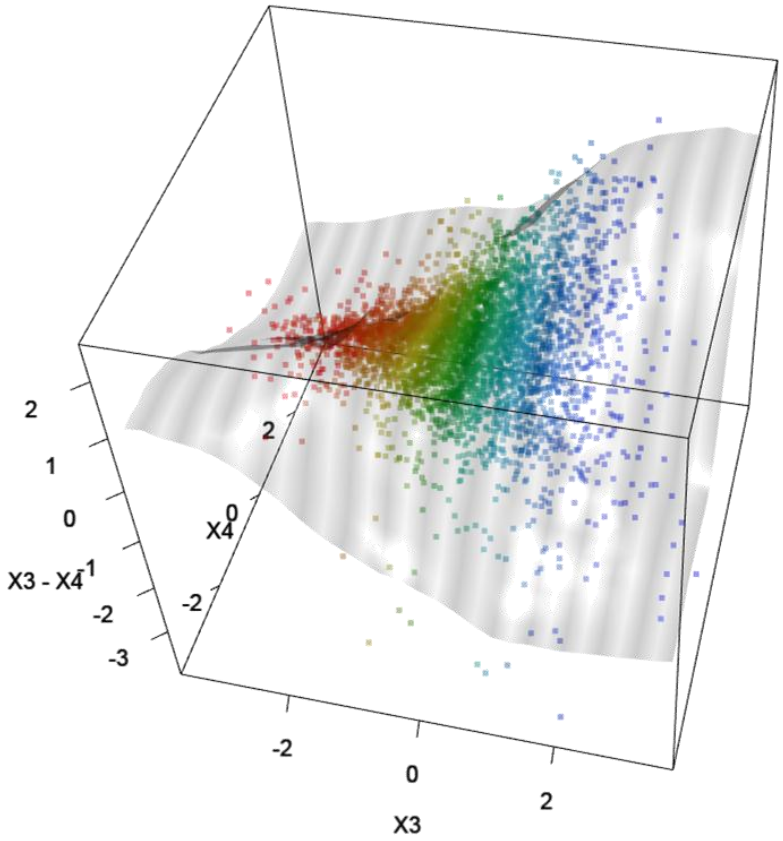



X1 r^2= 0.99
X2 r^2= 0.98
X3 r^2= 0.09
X4 r^2= 0.1
X5 r^2= 0.02
X6 r^2= 0.08


X1
X4
X3
X2
X5
X6
Click me
for 3D

...rgl 3d plot, printed screens





Thank you

References:
Feature contributions:
Victor E. Kuz'min et al, Interpretation of QSAR Models Based on Random Forest Methods, Molecular informatic 2011
Feature contributions for multi-classification.
Anna Palczewska et al, Interpreting random forest classification models using a feature contribution method, arXiv Dec 2013
Feature contributions plotted against feature values:
Søren Welling et al, In silico modelling of permeation enhancement potency in Caco-2 monolayers based on molecular descriptors and random forest, Eur J Pharm Biopharm May 2015