Response Grid Plots for Model-Agnostic Explainable AI

Russell Barton
The Pennsylvania State University

Based on Barton. R. R. (2025). Response grid plots for model-agnostic machine learning insight, *IMA Journal of Management Mathematics* 36(4), pp. 623–648, 2025

Summary

- Machine learning (ML) models provide predictive characterization of response functions based on observed or simulated data.
- **Insight on** the nature of **the approximation** to the underlying **response** function is essential **for** the model output to be **trusted and used** by decision makers.
- This is the motivation for explainable AI (xAI).
- When a response function is well-behaved, methods that identify marginal effects and interactions for important input variables can provide adequate insight on its nature. Alternatively, interpretable ML models can provide an ensemble of simple models to decompose a complex response into interpretable elements.
- These methods fail when the response has local anomalies.
- This talk introduces a model-generated but model-independent visual display of functional information, not interpreted through model form, model coefficients, or sensitivity analysis.
- The value of response grid plots (**RGP**s) is shown through several examples.

Context

- Prediction models have two primary purposes:
  - Given a set of characteristics, *x*, predict a response or outcome *y*;
  - Characterize key aspects (i.e. **the nature**) of the response function: important predictor variables, linearity, interaction, local anomalies.
- ML Examples: prediction model for simulation model output (a metamodel); prediction model for an engineering finite element or fluid flow model (a surrogate model); prediction for weight based on age (a regression model); prediction for probability of part failure within 10 years (a logistic regression or failure rate regression model).
- RGPs can be applied to prediction models in general – if their computational cost is not too high.
- How do model-builders engender TRUST in their models? (Necessary for adoption and use.)

What Follows

(In This Talk)

- ML general form.
- ML Model requirements and the issue of trust identifying three key research streams on trust for ML models, all considerd under the umbrella xAI:
  - interpretability;
  - post-hoc explainability;
  - sensitivity analysis.
- Applied to a virus propagation simulation model.
- Direct visualization of the ML response surface with Response Grid Plots.
- An insurance premium example (if time).

ML Model Form

- Any mathematical model of a response function $g(x)$ that produces $g(x)=(\hat{y})$
  - a predicted quantitative response as a function of one or more quantitative or qualitative predictors.
- Variables, x_k, $k = 1, \ldots, d$. Called ‘predictor variables’ or ‘factors’.
- x_k often scaled to values in the interval $[-1, 1]$; region of interest will be assumed to be a d-dimensional unit hypercube.
- The response that is being predicted (y) is a real number. Some models produce more than one y.
- Common ML models include polynomial regression, Gaussian Process (Bayesian) models, neural network models, radial basis function models

Trust, Interpretability and Explainability

- Three overlapping streams of research for characterizing insight to build trust in ML models: **explainability, interpretability and sensitivity analysis**.
- Research on sensitivity analysis is mostly a separate stream - perhaps because it is applied directly to system models as well as to ML approximations.
- Recent explosion of research on Explainability methods, called xAI. Searching ‘explainable AI’ OR ‘interpretable AI’ OR ‘explainable ML’ OR ‘interpretable ML’ in the Dimensions database gave 1600 documents in 2016 but over 66,000 in 2025.
- Here all three streams lumped together as xAI.

Trust, Interpretability and Explainability

- Retzlaff et al. (2024) define **interpretability** in close alignment with Rudin (2019): interpretability is a method that examines the ML model itself (e.g., regression model coefficients); explainability is applied post-hoc by exercising the metamodel.
- **Sensitivity analysis** methods (think first- and second-order effects and interactions) are post-hoc explainability methods. Borgonovo (2023) identifies four sensitivity analysis goals:
    - identification of variable or **feature importance**;
    - **trend** determination (direction of change);
    - identification of **interactions** among the predictor variables or other model internal structure;
    - **stability**, the lack of extreme changes in response from small changes in variable values.

Interpretable Models

- Model form directly implies nature of response surface:

$$\hat{y} = 3 + 5x$$

- Response is linear function of x, increments 5 units for every unit increase in x.
- Directly interpretable models include linear regression, logistic regression, generalized linear models (GLM), generalized additive models (GAM), tree models and other rule models (Molnar, 2025).

Interpretability for a Virus ML Model

- Since many ML applications are in health care, we illustrate the advantage of RGPs using a virus propagation simulation model (e.g., H1N1, Beeler et al., 2012).
- Model predicts total infections (y) as a function of:
- Two policy variables tested at three levels: vaccination level (**Vac**; 0, 30 and 60%) and quarantine (**Quar**; 0, 30 and 60%).
- Two virus characteristics tested at three levels: infectiousness (**Infec**) was modeled as ±50% above and below a baseline value, and days contagious (**Days**) as ±2 days around a baseline value.

Interpretability: Regression/ANOVA ML Model

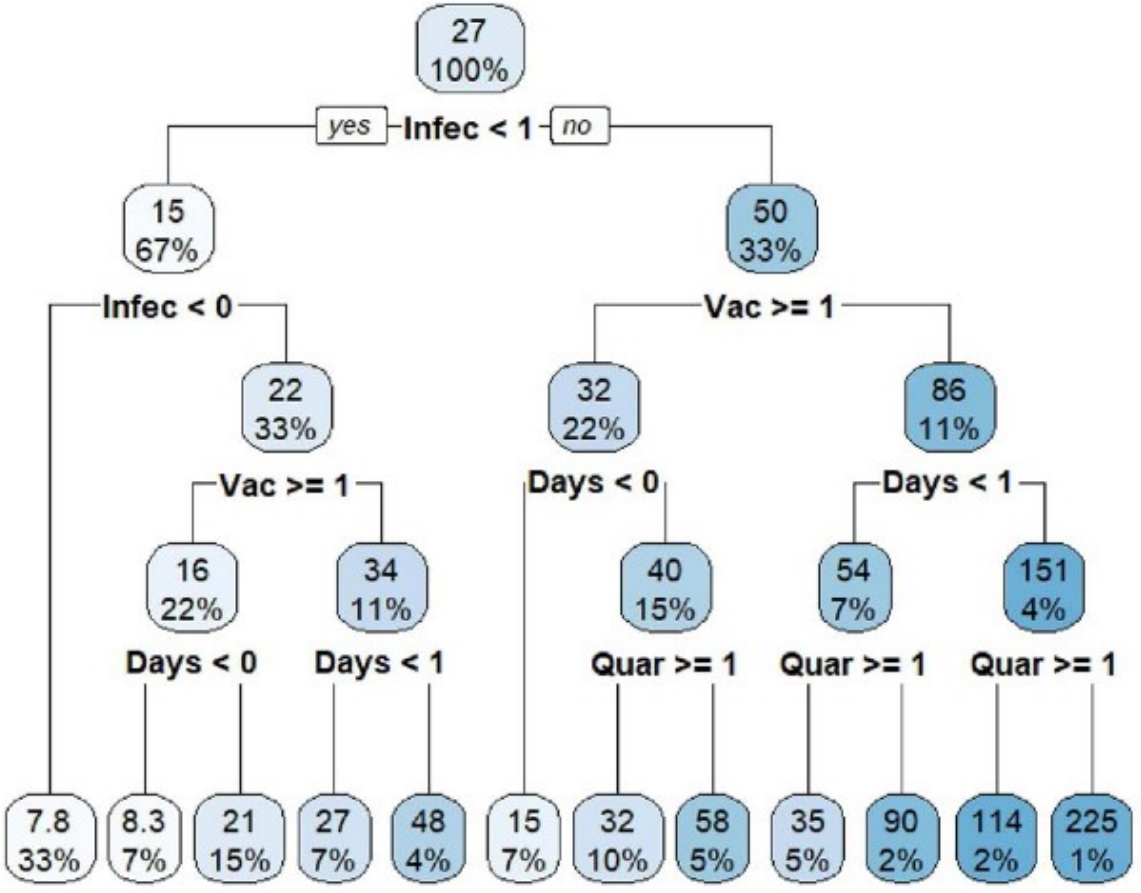
- Our interpretability example suggests regression ML models are interpretable.
- This example suggests otherwise: can you interpret this table? (Infec has highest impact, interactions important)
- Note: predictor variables scaled so coefficient is effect (thousands of cases) in moving predictor variable from middle value to highest value.

TABLE 1 *Coefficients (rounded) for the linear regression virus model*

First order	Coefficient	Second order	Coefficient	Third order	Coefficient
(Intercept)	56	I(Infec^2)	7	Days:Infec:Quar	−9
Days	24	I(Quar^2)	6	Days:Infec:Vac	−10
Infec	54	I(Vac^2)	8	Infec:Quar:Vac	9
Quar	−27	Days:Infec	34		
Vac	−37	Days:Vac	−11		
		Infec:Quar	−19		
		Infec:Vac	−23		
		Quar:Vac	7		

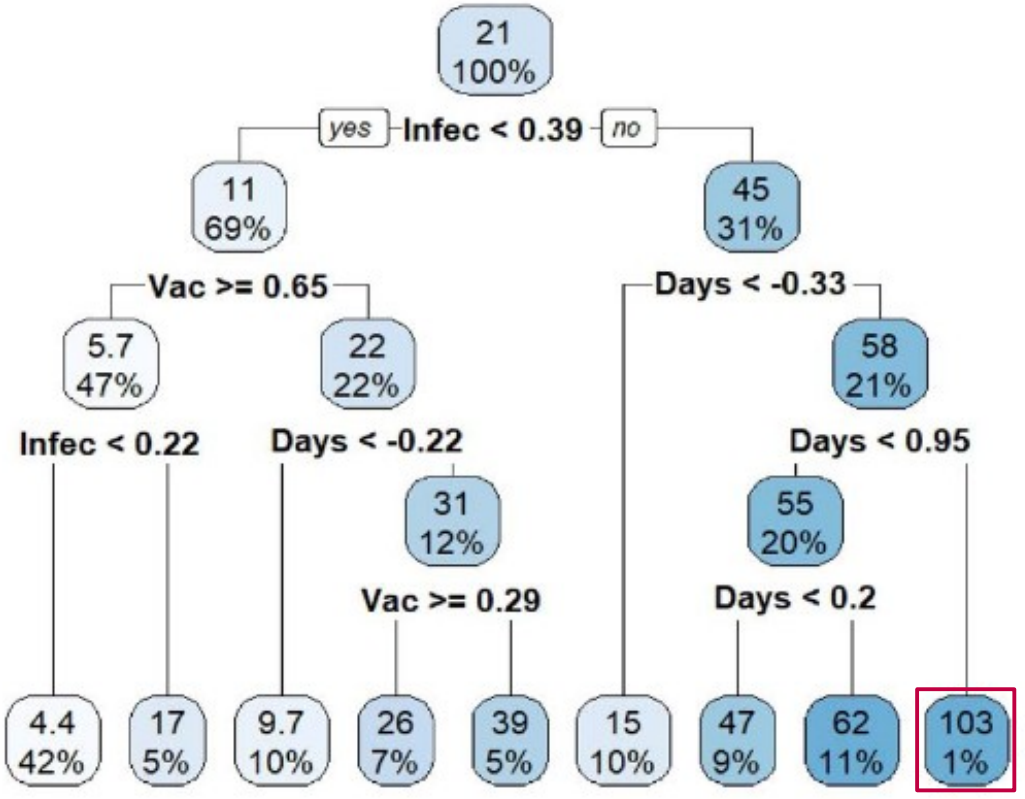
Interpretability: Regression Tree

- Regression trees predict a quantitative *y*. But typically each leaf predicts a specific numerical value, so the ML response function is piecewise constant
- The R package **`rpart`** default parameters produced a difficult-to-interpret plot with five levels and an R^2 of 0.80.
- Reducing the maximum number of levels to four produced the plot in Fig. 1 with an R^2 of 0.91.
- Infec highest impact, interactions exist – vaccination next most important.



Interpretability: Regression Tree 2

- **<u>Interpretability depended on factorial structure of fitting data</u>**. Here is fitted regression tree based on data from an orthogonal array (think fractional factorial). R^2 now .94
- Infectiousness still at top level, but now in one case Days is next most important (interaction) – or is it Vac?
- Predictions differ: predicted number of infections is 103,000, compared with 225,000 for the tree based on grid training data.



Interpretable Regression Model Trees

- Regression model trees replace the single-valued leaves with simple linear regression models, different models for each leaf.
- Applied to the virus simulation model, the results were similar to that for regression trees: for factorial data the models were somewhat interpretable, but not when fitted using data from an orthogonal array.

Post-hoc Explainability/Sensitivity Methods

- Too many to illustrate today. All have similar problems. Will only show ICE and ALE plots, and Tornado plot.

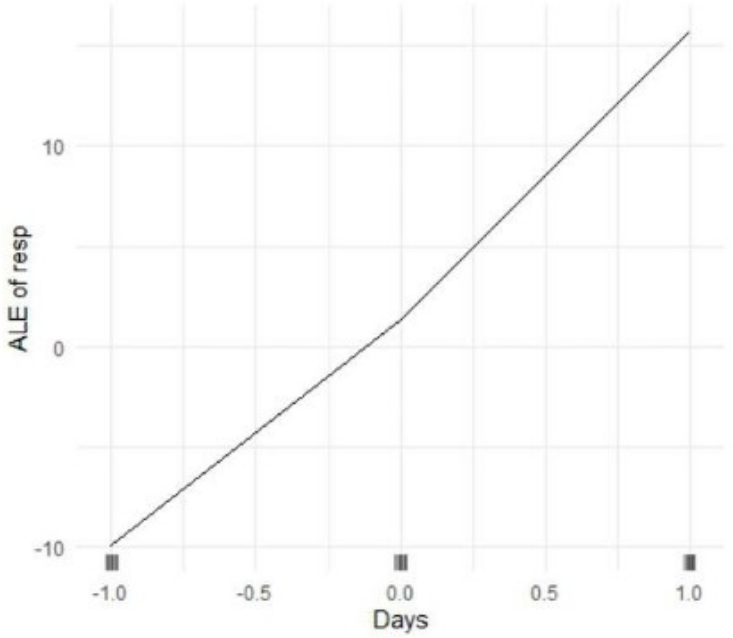
TABLE 2 *Post-hoc explainability methods*

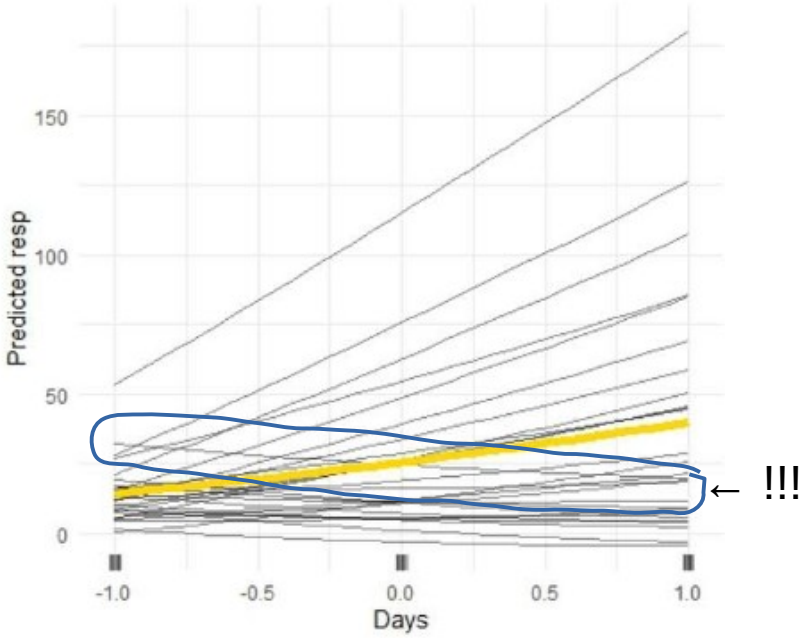
Explainability method	Scope	Output format
CP Plot	Local/global	Visual
ICE plot	Local/global	Visual
PDP	Global	Visual
M-plot	Global	Visual
ALE plot	Local/global	Visual
Interpretable Surrogate	Global	Numerical/rules
LIME, LIVE	Local	Numerical/rules
Sensitivity methods		
Finite change methods and tornado plot	Local/global	Visual/numerical
Friedman H statistic for interactions	Global	Visual/numerical
S-ICE plot	Global	Visual/numerical
Functional ANOVA, Sobol and Shapley indices and variance decomposition	Global	Visual/numerical
Mikado plot	Global	Visual
FAST, RBD-FAST	Global	Visual/numerical
Permutation feature importance	Global	Visual/numerical

Explainability:ALE and ICE Plots for ML Days Infectious

- When x's not orthogonal (random design; perhaps for input uncertainty application), Accumulated Local Effects (ALE, Apley and Zhu 2020) plots show effect of x_k on metamodel response separately from the effects of correlated predictors, but local averaging requires segmenting of the predictor-variable values. There is one plot for each x_k (here: Days Infectious, scaled (-1 to +1)). ALE from R package **`iml`**. Here based on factorial data so ALE not needed.

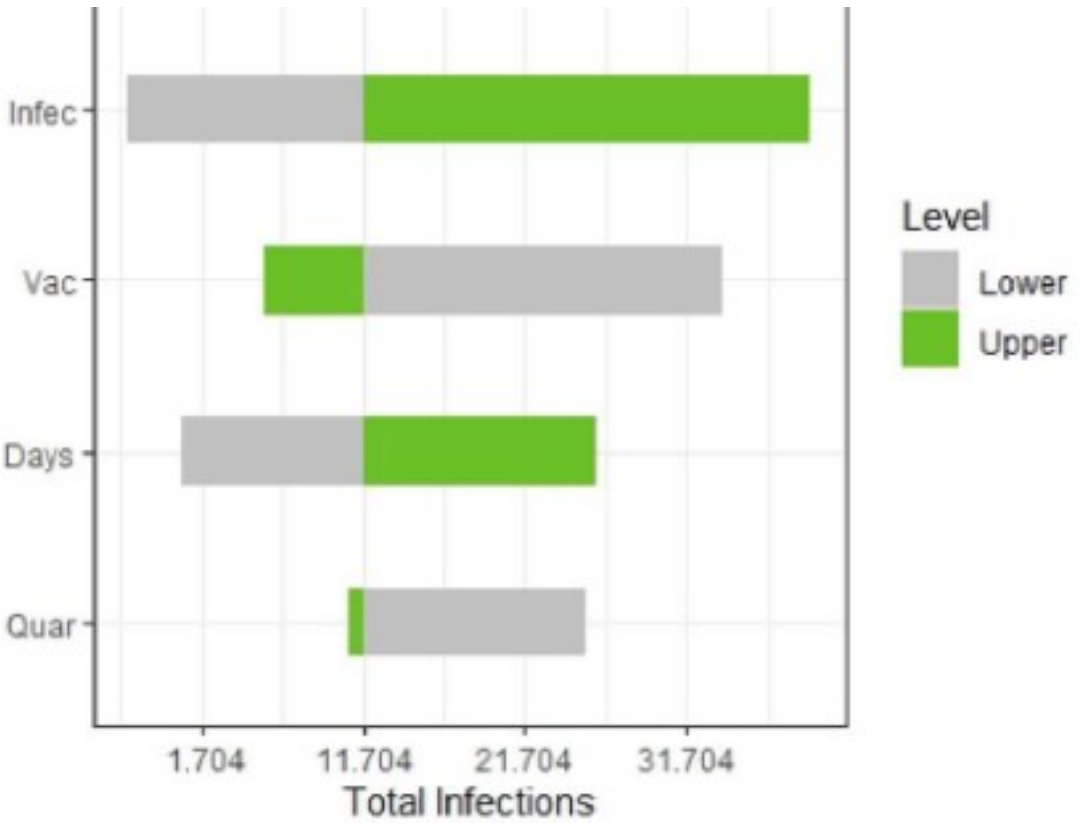
- Individual conditional expectation (ICE) plots show how the metamodel response function varies as one predictor variable x_k is changed about its value for each data point in the training set. There is one plot for each x_k (here: x_k = Days Infectious). ICE from R package **`iml`**.





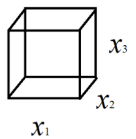
Sensitivity: Tornado Diagram for Virus Model

- Provides order of importance for predictor variables like variable importance plot, with additional useful information. (Borgonovo & Smith 2011, **`tornado`** R package).
- Changes in infectiousness can lead to a very small number of predicted infections at lowest value, or a large number (over 30,000) for the highest.
- Increasing % vaccinated and % adhering to quarantine both LOWER predicted number of infections.
- Increasing days infectious and infectiousness both INCREASE predicted number of infections.

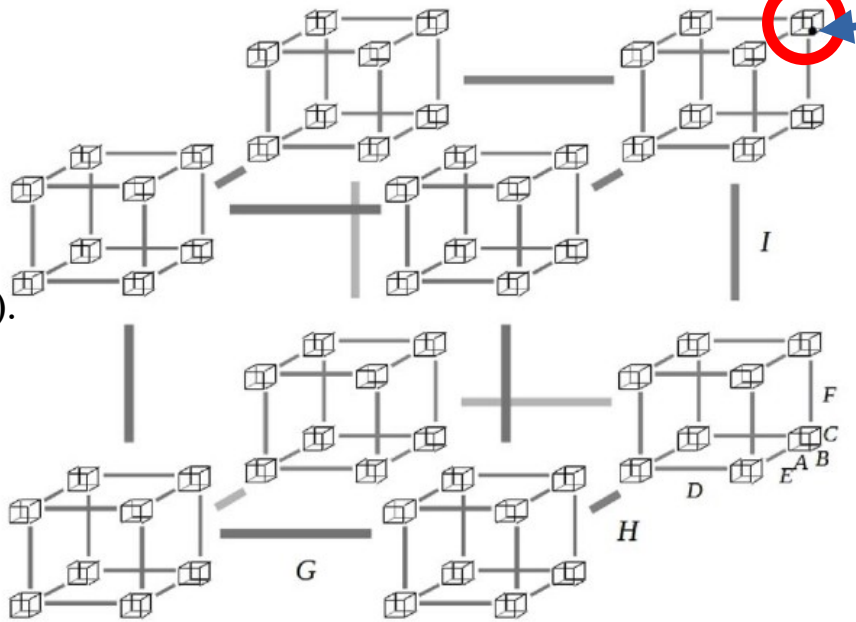


New: Response Grid Plot (RGP) Framework

- Grid structure of factorial design allows visualization of points in up to nine-dimensional space: increases left-to-right, front-to-back, bottom-to-top.
- 3 dimensions:



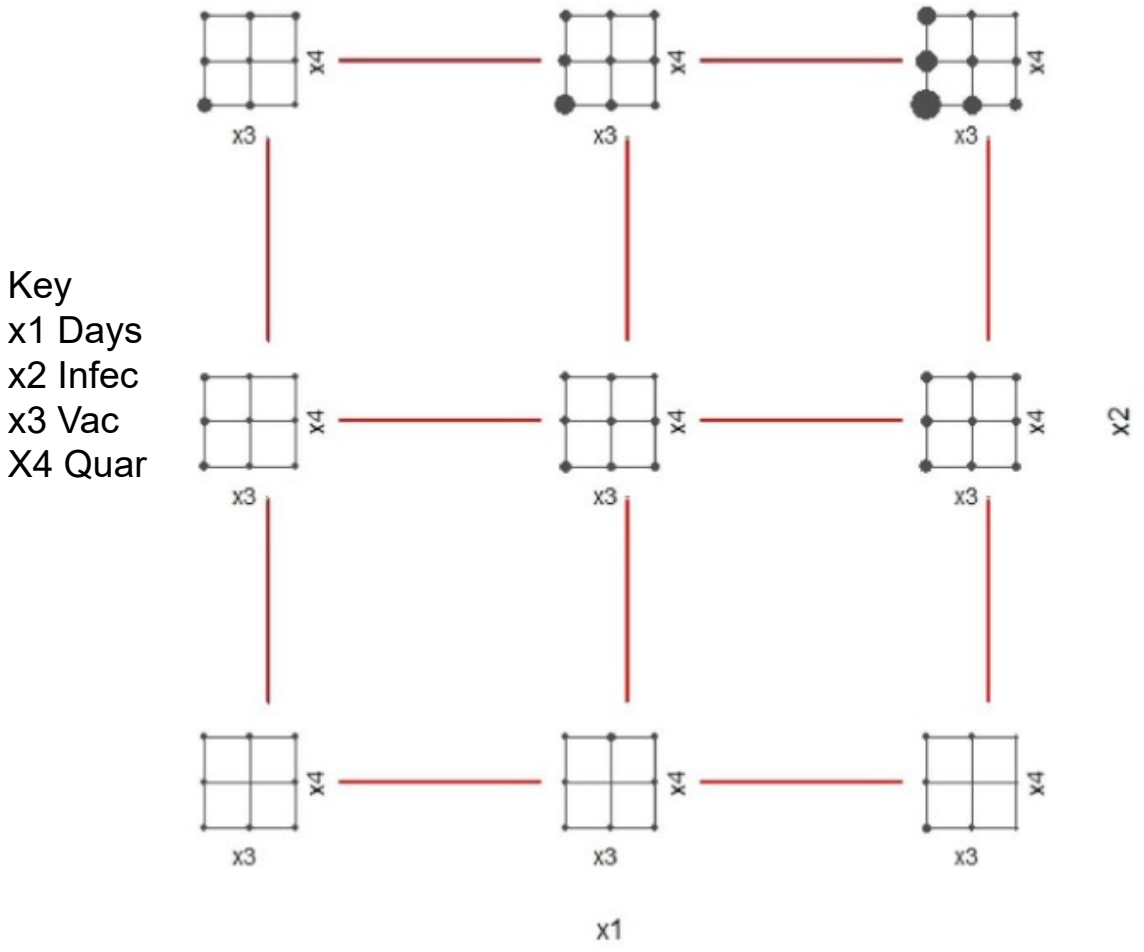
- 9 dimensions:
- 9 not practical if more than two predictor variables plotted at more than two levels (8 may be OK).
- **FOR RGP: add disks, areas scaled to response value**.



Full recovery of 9-dim. location
G: hi
H: hi
I: hi
D: hi
E: hi
F: hi
A: hi
B: lo
C: lo

Response Grid Plot for Virus Model

- A COMPLETE explanation of the ML model (shows local anomalies):
- Vaccination rates (x_3) can have a strong effect on infections, but only when the infectiousness (x_2) is high.
- At high levels of infectiousness (x_2), the effect of vaccination (x_3) decreases as days infectiousness (x_1) decreases and/or quarantining (x_4) percentage increases.
- The effect of vaccination (x_3) does not matter much at low- to mid-levels of infectiousness (x_2), where the number of infections remain relatively low and insensitive to the other predictor variables.



Summary

- ML models require TRUST for implementation.
- Existing interpretable models and explainability methods can fall short if there are local anomalies in the response surface.
- Response Grid Plots (RGPs) can be an effective tool for understanding the nature of an ML (or real system) response function.
- More details (including references and dual response plots) in the *IMA J. Mgt. Math.* Article.
- Code for the RGP examples in this article can be found at https://github.com/rrbarton51/ResponseGridPlots.git

Acknowledgments

- Thanks to the 2026 ENBIS organizers and volunteers, particularly Bertrand Iooss, Chair of the Program Committee, Rossella Berni, Chair of the Organising Committee and the QSR organizers, Mostafa Reisi Gahrooei and Christian Capezza. And to Kamran Paynabar for facilitating my participation.

- Thanks to Lee Schruben and Stu Hunter for inspiration and encouragement.

- Links:
  - PowerPoint and link to github: https://sites.psu.edu/russellbarton/ (Google "Russell Barton's personal page")
  - WSC 26 RGP code with more examples:https://github.com/rrbarton51/Response-Grid-Plots-for-WSC-Tutorial

References

D. M. Aleman, T. G. Wibisono, and B. Schwartz. A nonhomogeneous agent-based simulation approach to modeling the spread of disease in a pandemic outbreak. *Interfaces* 41, 301–315. 2011.

D. W. Apley and J. Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82, 1059–1086, 2020.

R. R. Barton. Response grid plots for model-agnostic machine learning insight. *IMA Journal of Management Mathematics* 36(4):623–648, 2025.

M. F. Beeler, D. M. Aleman, and M. W. Carter. A large simulation experiment to test influenza pandemic behavior. In *Proceedings of the 2012 Winter Simulation Conference* (WSC), pages 828–834, Piscataway, NJ, USA, 2012. IEEE.

E. Borgonovo. Sensitivity analysis. In *Tutorials in Operations Research: Advancing the Frontiers of OR/MS: From Methodologies to Applications*, 52–81. INFORMS, 2023.

E. Borgonovo, E. Plischke, and G. Rabitti. The many Shapley values for explainable artificial intelligence: A sensitivity analysis perspective. *European Journal of Operational Research* 318, 911–926, 2024.

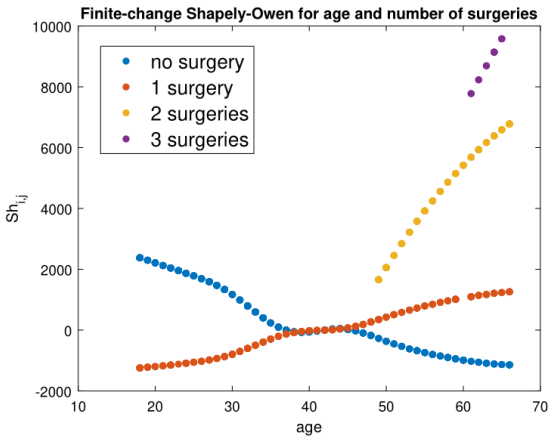
P. Diaconis. Bayesian numerical analysis. In Statistical Decision Theory and Related Topics IV(J. Berger & S. Gupta, eds), vol.1. New York, NY: Springer, 163–175 1988.

C. O. Retzlaff, A. Angerschmid, A. Saranti, D. Schneeberger, R. Röttger, H. Müller, and A. Holzinger. Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists. *Cognitive Systems Research* 86, 101243, 2024.

C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 206–215, 2019.

Shapley for xAI: Medical Insurance Premium Data

- Online Kaggle **observational dataset** used by Borgonovo et al. (2024)
- 986 records, # by category value.
- Example finite change Shapley-Owen age-surgeries interaction plot is below:

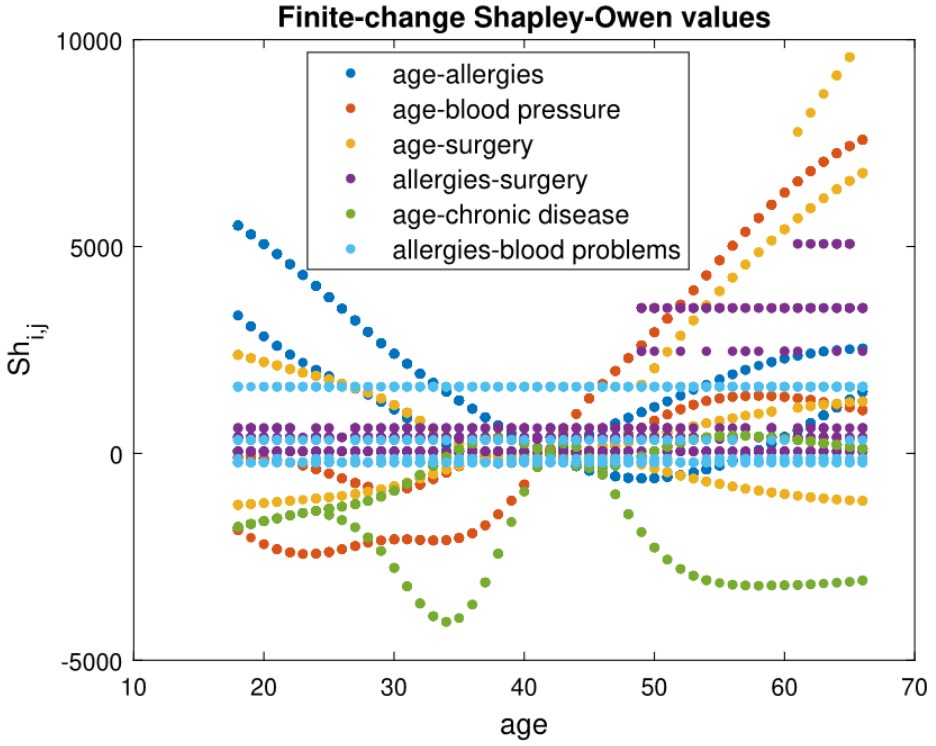


# Age	# Diabetes	# BloodPressurePr...	# AnyTransplants	# AnyChronicDisea...
Age Of Customer	Whether The Person Has Abnormal BloodSugar Levels	Whether The Person Has Abnormal Blood Pressure Levels	Any Major Organ Transplants	Whether Customer Suffers From Chronic Ailments Like Asthama, Etc.
18 66	0 1	0 1	0 1	0 1

# Height	# Weight	# KnownAllergies	# HistoryOfCancerI...	# NumberOfMajorS...
Height Of Customer	Weight Of Customer	Whether The Customer Has Any Known Allergies	Whether Any Blood Relative Of The Customer Has Had Any Form Of Cancer	The Number Of Major Surgeries That The Person Has Had
145 188	51 132	0 1	0 1	0 3

Medical Insurance Premium Neural Network Model:

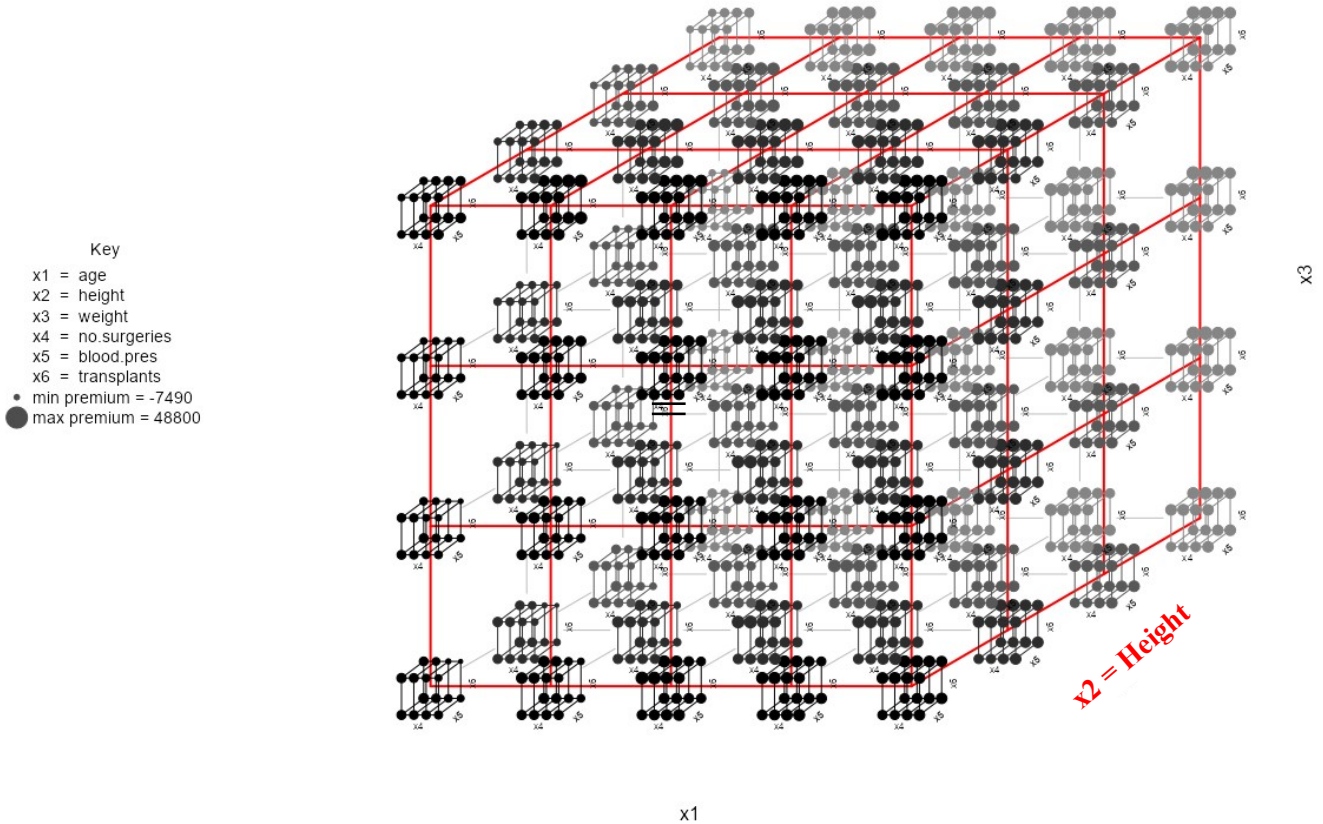
- Global neural network model needed to compute finite-change Shapley-Owen interactions from non-factorial observational data (Borgonovo et al. 2024).
- Plot shows change in finite-change Shapley-Own values as a function of Age, indicating interactions.
- 'Knowledge' of response is informative, but still marginal. No 3-, 4-, …, 9-factor interactions.



(a) Finite-change Shapley-Owen values for individual policyholders with $\bar{\mathbf{x}}$ as base case value, computed with the neural network.

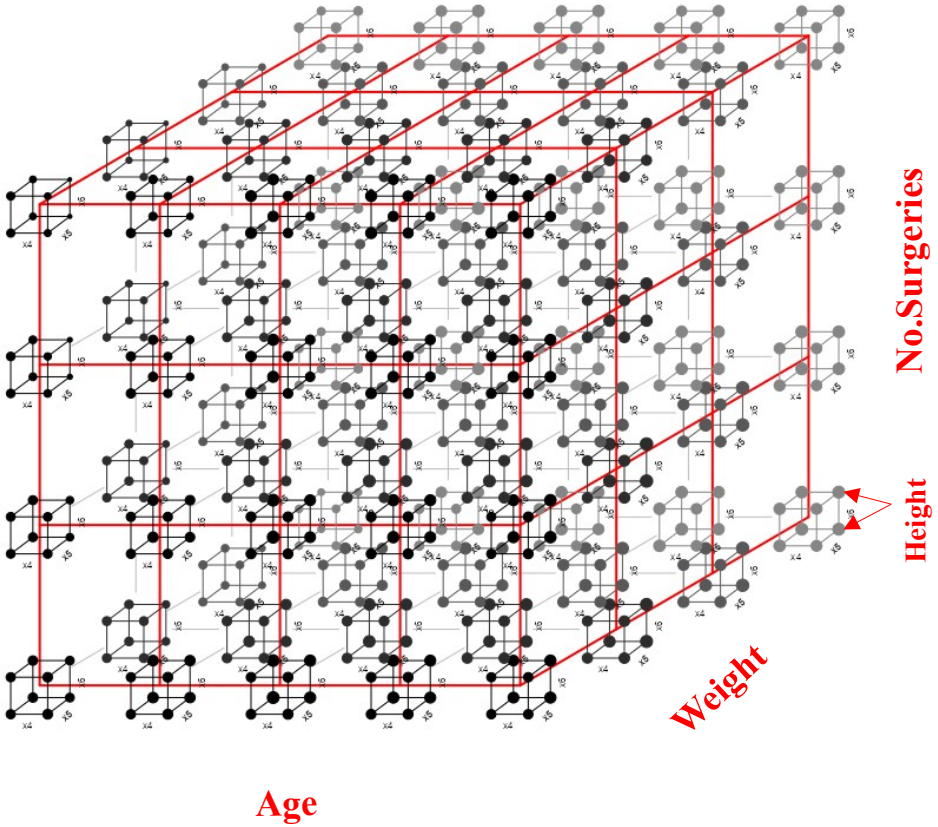
Medical Insurance Premium NN: RGP

- What does global neural network model response function look like? Factorial grid with ranges set to ranges in the Kaggle data.
- RGP of NN response, 5 levels age (most significant), 4 levels height, 4 levels weight, given discrete levels of surgeries (4), blood pressure (2), transplants (2).
- Q: This RGP too hard to view. Can it be reduced?
- A: x2 (Height) has little or no effect; reduce to two levels and re-plot.



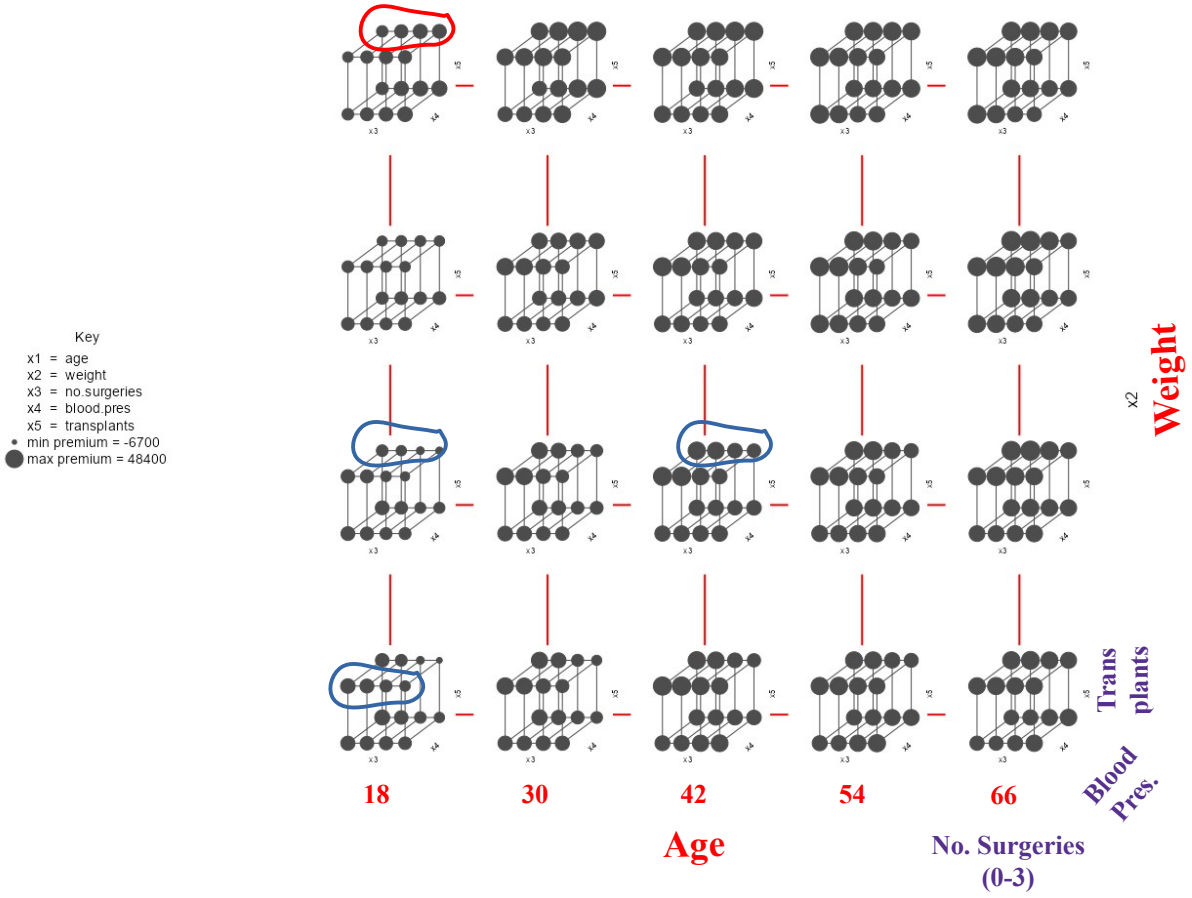
Medical Insurance Premium: RGP

- Height at 2 levels makes it clear that Height is not influential (top and bottom of each small cube approximately the same) – plot again without Height.
- Note: this RGP uses only half of the grid values in the NN model prediction data set.



Medical Insurance Premium: RGP

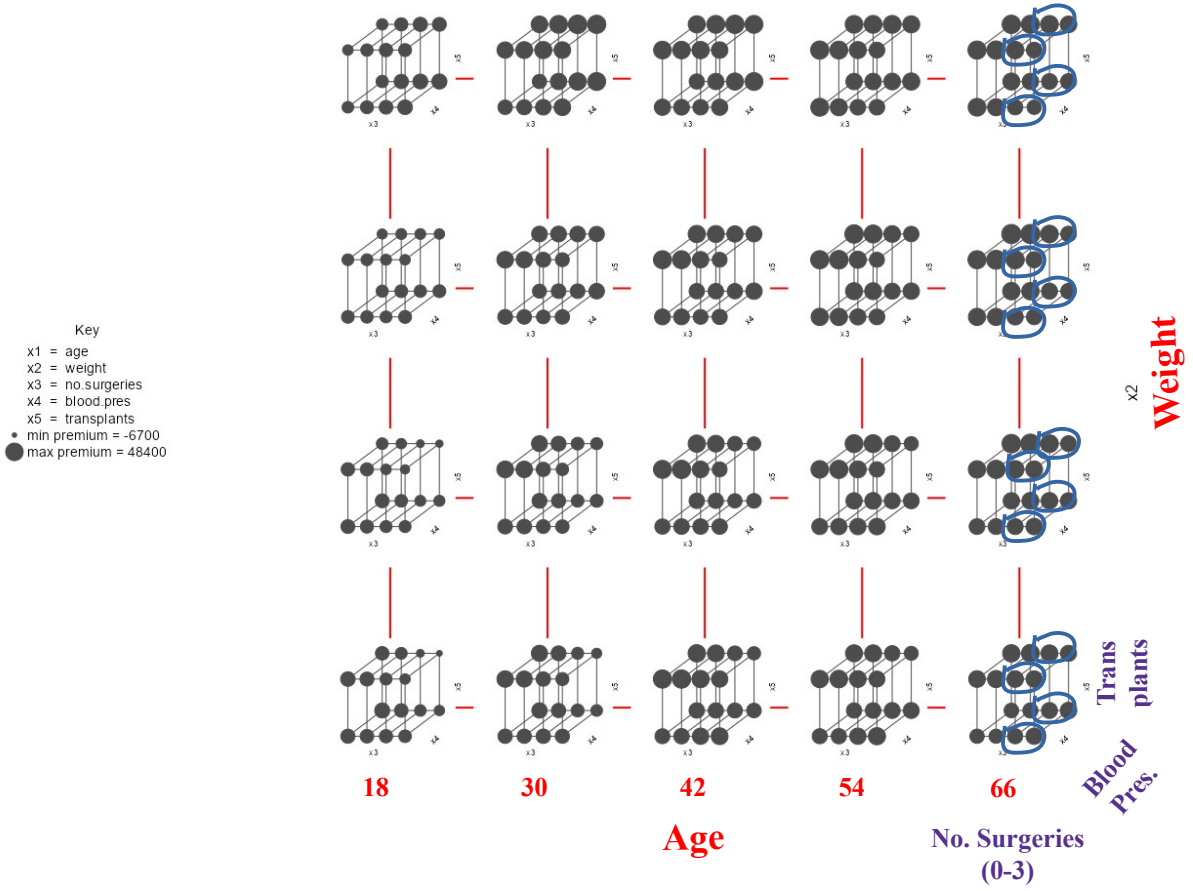
- Note: plot uses ALL of the grid values in the NN model prediction data set. Response averaged over heights.
- Age >= 30 strongest effect – sudden jump in predicted premium.
- Weight a gradual and smaller increase in predicted premium.
- Strange: predicted premiums decrease with increasing # surgeries generally (left to right on small cubes) except upper left: high weight and young age.
- Strange: at 18 (left column), having transplants DECREASES predicted premium (except for high weight).



Medical Insurance Premium: RGP

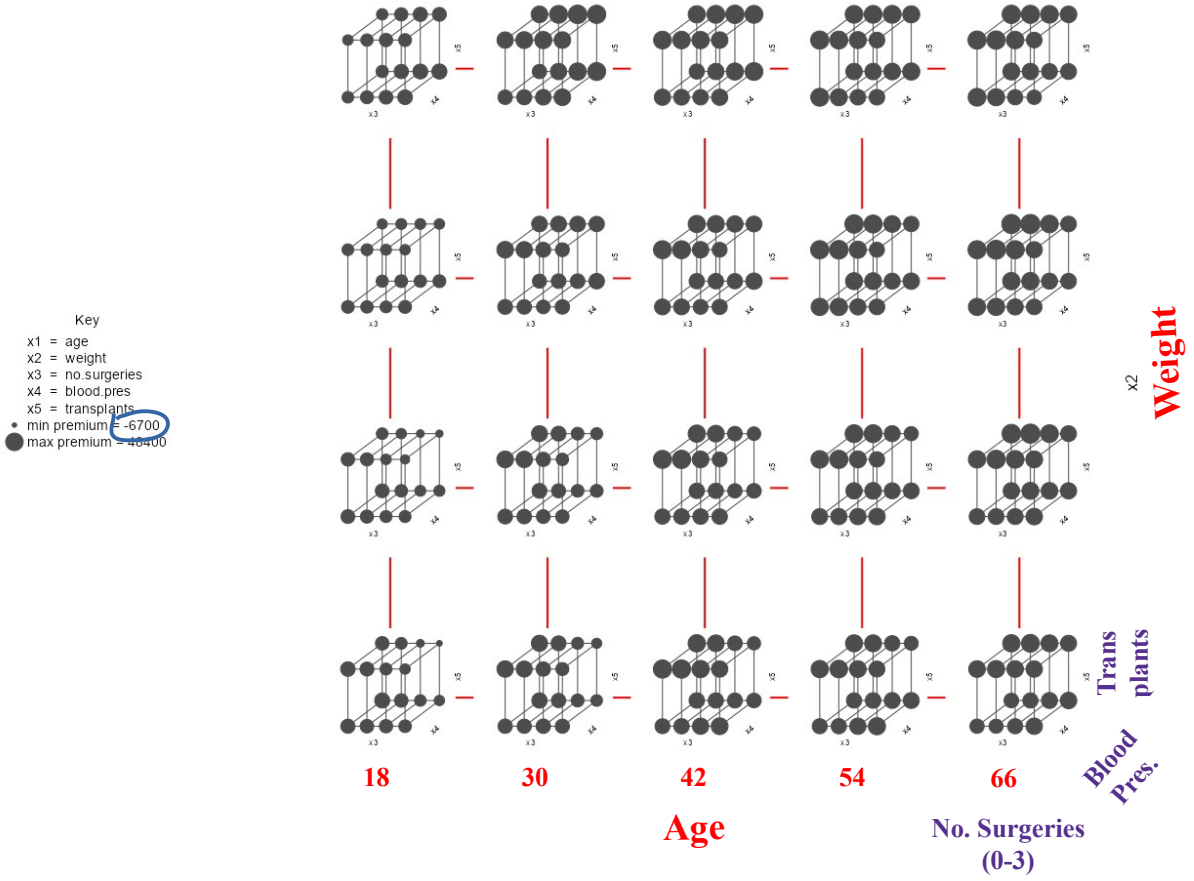
- Compare with Finite-change Shapley interaction interpretation:

"Age and number of surgeries have a very negative explanatory power when considered together, meaning that these features provide globally redundant information to the pricing model. [By squared cohort figure] **the joint contributions of age and two or three surgeries are highly negative at ages greater than 60. Thus, for these ages, age and the occurrence of two or three surgeries provide redundant information for insurance pricing**."



Medical Insurance Premium: RGP

- Before we leave this non-simulation example…
- SOME NN PREDICTED PREMIUMS ARE NEGATIVE!
- Cause – evaluate response on a grid over the domain.



Medical Insurance Premium: RGP

- Negative minimum for premium!
- What are predictor *x* values?
  - Young (Less than 10% age 18-21 in Kaggle dataset)
  - with transplant (Only 1 age 18 in Kaggle dataset)
  - multiple surgeries (Youngest in Kaggle dataset is 49!)
- What data nearest to these DOE points in the original Kaggle dataset used to fit the NN model?

	A	B	C	D	E	F	G
1	age	blood.pres	transplants	height	weight	no.surgeries	premium
2	18	1	1	187	51	3	-7485.18599264282
3	18	1	1	173	51	3	-6991.81869720127
4	18	1	1	159	51	3	-6456.81427683236
5	18	1	1	145	51	3	-5880.25503437547
6	18	1	1	187	78	3	-4016.84050217206
7	18	1	1	173	78	3	-3507.9729307439
8	18	1	1	159	78	3	-2977.16197179891
9	18	1	1	187	78	2	-2714.97656589627
10	18	1	1	187	51	2	-2579.20094934104
11	18	1	1	145	78	3	-2423.6265585094
12	18	1	1	173	78	2	-2271.67796300734
13	18	1	1	173	51	2	-2215.55655151791
14	18	1	1	159	51	2	-1799.61938682241
15	18	1	1	159	78	2	-1789.91833880295
16	18	1	1	145	51	2	-1334.24117595892
17	18	1	1	145	78	2	-1269.7535400859
18	18	0	1	187	78	3	1438.49430779712
19	18	0	1	173	78	3	1826.54159253231

Medical Insurance Premium: RGP

- What is nearest to these DOE points in the original Kaggle dataset used to fit the NN model?
- Histogram of DOE point distances to three nearest neighbors in actual Kaggle dataset. Scaling is +/-1, $x \in R^{10}$, so a distance of 2 would be generated if near neighbor distance was .6 in each dimension.
- DOE points with negative NN predictions a bit farther from data than those with positive NN predictions.
- Therefore may be poor fit ... but may be real. Not conclusive.
- Critic: Medical Insurance Premium example is for model of DATA not a model of a SIMULATION RESPONSE. How does this relate to simulation metamodels?
- A: 0) Finite-change Shapley-Owen vs RGP comparison still applies;
    1) Location of odd predictions NOT noticed with Shapley-Owen;
    2) Metamodels fitted to 'space filling' OA or other designs may have some RGP grid points far from any simulation data - warning.