NOVARTIS

Explainable Machine Learning for Assessing Treatment Effect Heterogeneity in Clinical Trials

Kostas Sechidis
Advanced Methodology & Data Science

Workshop: Methods for Explainable Machine Learning in Healthcare,
Amsterdam University Medical Centers
04/02/2026

Overview of today's presentation

Motivating the problem

- Why assessing treatment effect heterogeneity (TEH) matters and why it is a challenging problem

Introducing WATCH

- A workflow for a structured exploration of TEH

Using explainable ML/AI for assessing TEH

- Using Shapley values derive effect modifiers.

Imaginary scenario …

The trial failed!

... but for women over 50, the response rate is 70% in our drug and only 30% in placebo

Example inspired by a talk of Prof Richard Samworth

Treatment Effect Heterogeneity

- Subgroup analysis ⟺ Assessing Treatment Effect Heterogeneity (TEH)
- TEH: Non-random variation in treatment effects across levels of baseline patient characteristics, which in this talk we as features or covariates.

- It is a hard problem, because of …
  - High chances of false negatives due to insufficient sample size
    Clinical trials not designed for assessing subgroup treatment effects or testing interactions (underpowered).
  - High chances of false positives due to multiplicity
    Performing multiple comparisons on unreliable or noisy subgroup treatment effects and selecting “the best” can introduce bias (selective inference).

Issues with replicability of subgroup findings

Article

July 3, 1991

Analysis and Effects in Sub Clinical Trials

Salim Yusuf, DPhil, MRCP; Janet

» Author Affiliations

JAMA. 1991;266(1):93-98. doi:10

Table 2.—Claims of Subgroup Effects on Mortality in the 65 Randomized Trials of β-Blockers in Acute Myocardial Infarction*

Study	Subgroup Benefit Claimed	Prior Hypothesis	Confirmed in Other Trials	Overall *P* Value	Test for Heterogeneity	Correction for Multiplicity	*P*
			Early Initiation of Treatment				
1. Barber et al[20]	Tachycardia at entry >100 beats per min	No	No	Not significant	–	–	–
2. MIAMI[21]	"High-risk" patients	No	No	NS	–	–	–
			Late Initiation of Treatment				
3. Anderson et al[22]	Treatment beneficial in patients <65 y and harmful in those >65 y	Unclear	No, most trials show similar reductions in relative risk among younger and older patients.[23]	NS	–	–	–
4. Hjalmarson et al[24]	Benefit observed only in patients with HR >65 beats per min (never formally published)	No	No, the MIAMI trial included only this group. Overall results were not significant. The impact of HR on effect of treatment was tested in ISIS-1.[25] No differential was found.	<.03	–	–	–
5. Wilhelmsson et al[26]	Benefit only in patients with "electrical" or "mechanical" complications	No	No, although one other study[27] observed a similar result. Many other studies and the Beta-Blocker Pooling Project[28] failed to identify this subgroup as benefiting preferentially.	NS	–	–	–
6. Multicenter International[29]	Benefit only in patients with anterior MI before entry	No	No	<.08	–	–	–
7. Taylor et al[30]	Benefit only among those with treatment initiated within 6 mo of MI, while those treated later appeared harmed	No	No	NS	–	–	–
8. Beta-blocker Heart Attack Trial[27]	Benefit only in patients with "electrical" or "mechanical" complications prior to randomization	No	Not consistently	<.003	–	–	+
9. Yusuf et al[23] (pooled data)	β-blockers without ISA more effective than those with ISA	No	Uncertain. Three new trials appear to contradict this conclusion. A trial of metoprolol was unpromising. Two studies, one of acebutolol and one of oxprenolol, both with ISA, were favorable	<.0001	+(*P*<.02)	–	–

*Only subgroups that are "proper" are included. HR indicates, heart rate; MI, myocardial infarction; and ISA, intrinsic sympathomimetic activity.

Issues with replicability of subgroup findings

Original Investigation

April 2017

Evaluati
and Cor
Random

Joshua D. Wallach,
PhD[1]; Ewout W. Ste

» Author Affiliations | Article Information

JAMA Intern Med. 2017;177(4):554-560. doi:10.1001/jamainternmed.2016.9125

RESULTS Sixty-four eligible RCTs made a total of 117 subgroup claims in their abstracts. Of these 117 claims, only 46 (39.3%) in 33 articles had evidence of statistically significant heterogeneity from a test for interaction. In addition, out of these 46 subgroup findings, only 16 (34.8%) ensured balance between randomization groups within the subgroups (eg, through stratified randomization), 13 (28.3%) entailed a prespecified subgroup analysis, and 1 (2.2%) was adjusted for multiple testing. Only 5 (10.9%) of the 46 subgroup findings had at least 1 subsequent pure corroboration attempt by a meta-analysis or an RCT. In all 5 cases, the corroboration attempts found no evidence of a statistically significant subgroup effect. In addition, all effect sizes from meta-analyses were attenuated toward the null.

Without external replication or plausibility, data-based findings alone are very speculative

Exploratory assessment of the TEH is important

Internally

Influence strategic internal development decisions, such as:

- enrichment,
- generating hypothesis for new (targeted) trials,
- prepare for health authority questions.

" … ignoring the problem, and similarly routinely dismissing results of subgroup analysis, is no scientific solution."

Externally

EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

31 January 2019
EMA/CHMP/539146/2013
Committee for Medicinal Products for Human Use (CHMP)

Guideline on the investigation of subgroups in confirmatory clinical trials

STATISTICAL PERSPECTIVES ON SUBGROUP ANALYSIS: TESTING FOR HETEROGENEITY AND EVALUATING ERROR RATE FOR THE COMPLEMENTARY SUBGROUP

Mohamed Alosh[1], Mohammad F. Huque[2], and Gary G. Koch[3]
[1]*Division of Biometrics III, Office of Biostatistics, OTS, CDER, FDA, Silver Spring, Maryland, USA*
[2]*Office of Biostatistics, OTS, CDER, FDA, Silver Spring, Maryland, USA*
[3]*Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA*

AN OVERVIEW OF STATISTICAL AND REGULATORY ISSUES IN THE PLANNING, ANALYSIS, AND INTERPRETATION OF SUBGROUP ANALYSES IN CONFIRMATORY CLINICAL TRIALS

Robert Hemmings
Medicines and Healthcare Products Regulatory Agency, London, United Kingdom

Workflow for Assessing Treatment effeCt Heterogeneity - WATCH

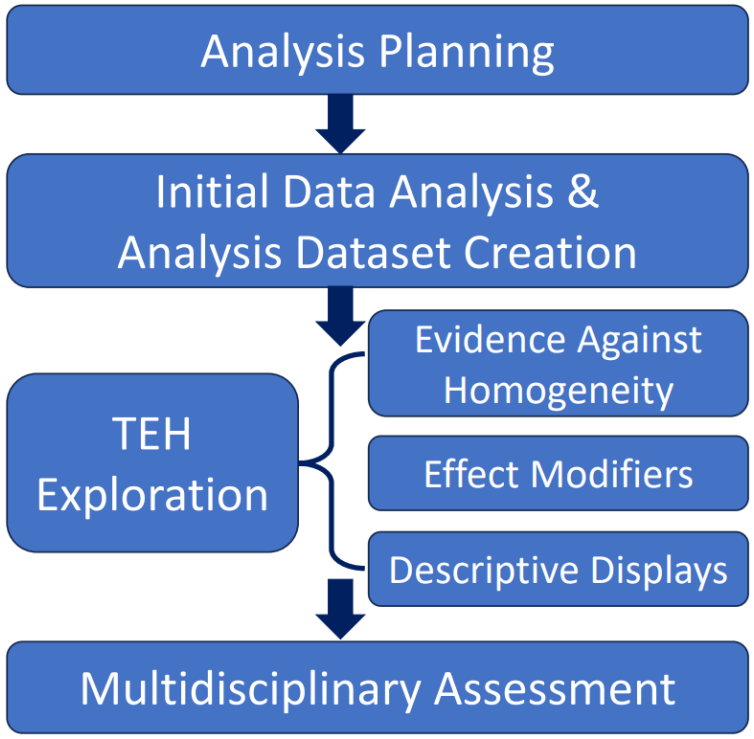
Pharmaceutical Statistics
The Journal of Applied Statistics in the Pharmaceutical Industry
The official journal of PSI

MAIN PAPER | Full Access

WATCH: A Workflow to Assess Treatment Effect Heterogeneity in Drug Development for Clinical Trial Sponsors

Konstantinos Sechidis, Sophie Sun, Yao Chen, Jiarui Lu, Cong Zhang, Mark Baillie, David Ohlssen, Marc Vandemeulebroecke, Rob Hemmings, Stephen Ruberg, Björn Bornkamp

First published: 26 December 2024 | https://doi.org/10.1002/pst.2463 | VIEW METRICS



Aims

- Generate insights on how treatment effects may vary with baseline characteristics in the clinical trial (or pool of trials)
  → Pre-planned and results quickly available after data-base lock
- Combine external and data-based evidence to improve decision making on treatment effect heterogeneity
- A systematic approach that fosters transparency, reducing misunderstanding or errors, and allowing better replication of the findings

Scope

- Exploratory assessment of TEH, applicable to Phase 2 or Phase 3 trial.

Out of scope

- Confirmatory statements on subgroup effects.

TEH exploration

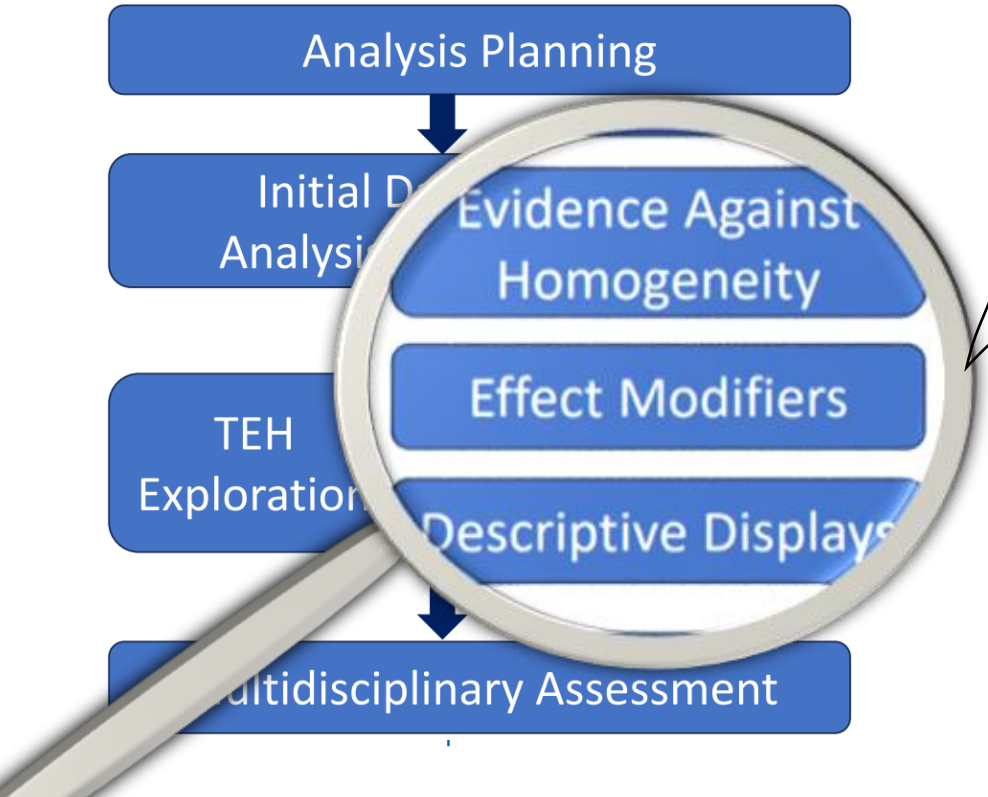
Pharmaceutical Statistics
The Journal of Applied Statistics in the Pharmaceutical Industry
The official journal of PSI

MAIN PAPER | Full Access

WATCH: A Workflow to Assess Treatment Effect Heterogeneity in Drug Development for Clinical Trial Sponsors

Konstantinos Sechidis, Sophie Sun, Yao Chen, Jiarui Lu, Cong Zhang, Mark Baillie, David Ohlssen, Marc Vandemeulebroecke, Rob Hemmings, Stephen Ruberg, Björn Bornkamp

First published: 26 December 2024 | https://doi.org/10.1002/pst.2463 | VIEW METRICS



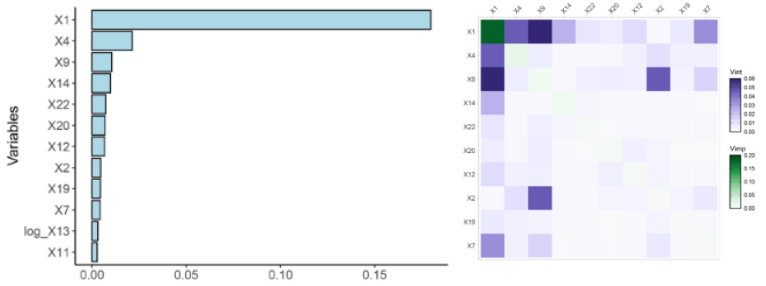
Address questions with flexible statistical modelling

Question 1: How strong is the overall evidence against the scenario of homogeneous treatment effects?
Question 2: Which variables drive observed heterogeneity?
Question 3: How does the treatment effect change for the identified variables?

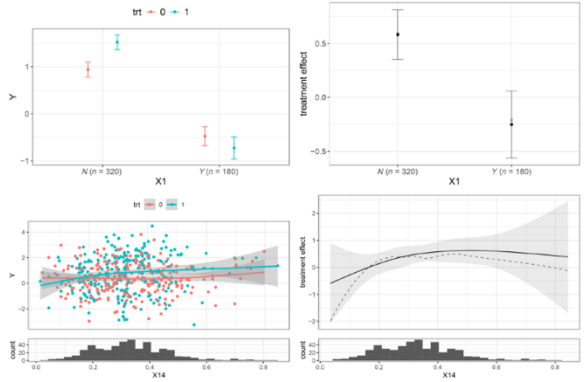
Global heterogeneity test p-value

p	Surprise value ($\log_2(p)$)	Verbal summary of evidence against homogeneity
[0.25,1]	(0, 2]	Low
[0.063,0.25)	(3, 4]	Moderate
[0.008,0.063)	(5, 7]	Noteworthy
[0.001,0.008)	(8, 10]	Strong
< 0.001	> 10	Very strong

Variable importance



Treatment effect plots



TEH exploration methods …

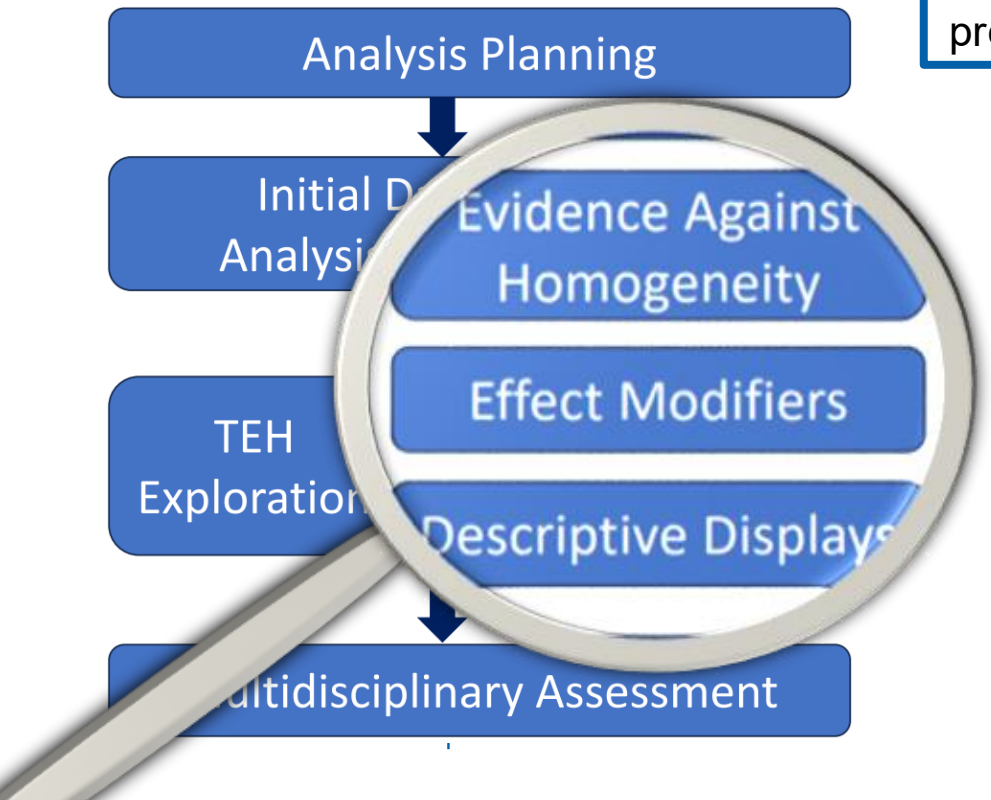
Pharmaceutical Statistics
The Journal of Applied Statistics in the Pharmaceutical Industry
The official journal of PSI

MAIN PAPER | Full Access

WATCH: A Workflow to Assess Treatment Effect Heterogeneity in Drug Development for Clinical Trial Sponsors

Konstantinos Sechidis, Sophie Sun, Yao Chen, Jiarui Lu, Cong Zhang, Mark Baillie, David Ohlssen, Marc Vandemeulebroecke, Rob Hemmings, Stephen Ruberg, Björn Bornkamp

First published: 26 December 2024 | https://doi.org/10.1002/pst.2463 | VIEW METRICS



Statistics in Medicine

RESEARCH ARTICLE | Full Access

Using Individualized Treatment Effects to Assess Treatment Effect Heterogeneity

Konstantinos Sechidis, Cong Zhang, Sophie Sun, Yao Chen, Asher Spector, Björn Bornkamp

First published: 27 November 2025 | https://doi.org/10.1002/sim.70324 | VIEW METRICS

Statistics in Medicine

RESEARCH ARTICLE | Full Access

Comparing Methods to Assess Treatment Effect Heterogeneity in General Parametric Regression Models

Yao Chen, Sophie Sun, Konstantinos Sechidis, Cong Zhang, Torsten Hothorn, Björn Bornkamp

First published: 22 January 2026 | https://doi.org/10.1002/sim.70381

Case 1: For continuous and binary endpoint use conditional average treatment effect (CATE) to derive a proxy of individualized treatment effects

Case 2: For treatment effects based on regression models (e.g. hazard ratio, odds ratio) derive a score residual for each patient

Q1: Use the derived variable to test the null hypothesis of homogenous treatment effect

Q2: Use the derived variable to rank covariates on their strength of modifying the treatment effect

Q3: Visualize how treatment effect changes with the identified covariates

Deriving effect modifiers … using SHAP

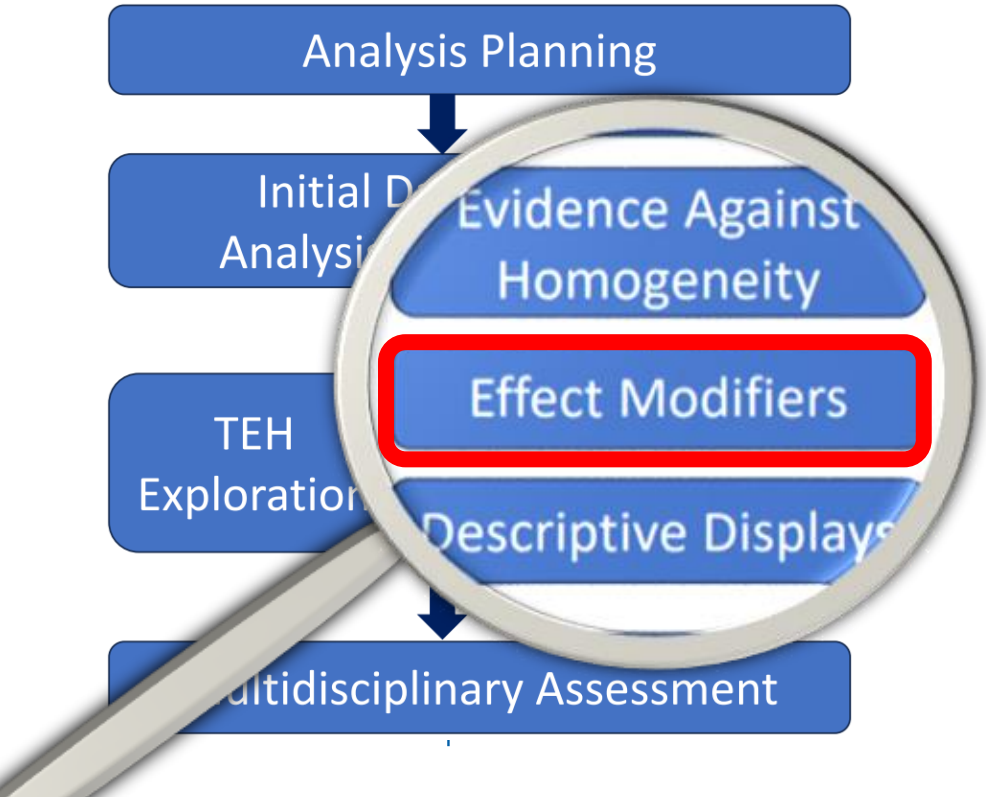
Pharmaceutical Statistics
The Journal of Applied Statistics in the Pharmaceutical Industry
The official journal of PSI

MAIN PAPER | Full Access

WATCH: A Workflow to Assess Treatment Effect Heterogeneity in Drug Development for Clinical Trial Sponsors

Konstantinos Sechidis, Sophie Sun, Yao Chen, Jiarui Lu, Cong Zhang, Mark Baillie, David Ohlssen, Marc Vandemeulebroecke, Rob Hemmings, Stephen Ruberg, Björn Bornkamp

First published: 26 December 2024 | https://doi.org/10.1002/pst.2463 | VIEW METRICS



Statistics in Medicine

TUTORIAL IN BIOSTATISTICS | Full Access

Overview and Practical Recommendations on Using Shapley Values for Identifying Predictive Biomarkers via CATE Modeling

David Svensson, Erik Hermansson, Nikolaos Nikolaou, Konstantinos Sechidis, Ilya Lipkovich

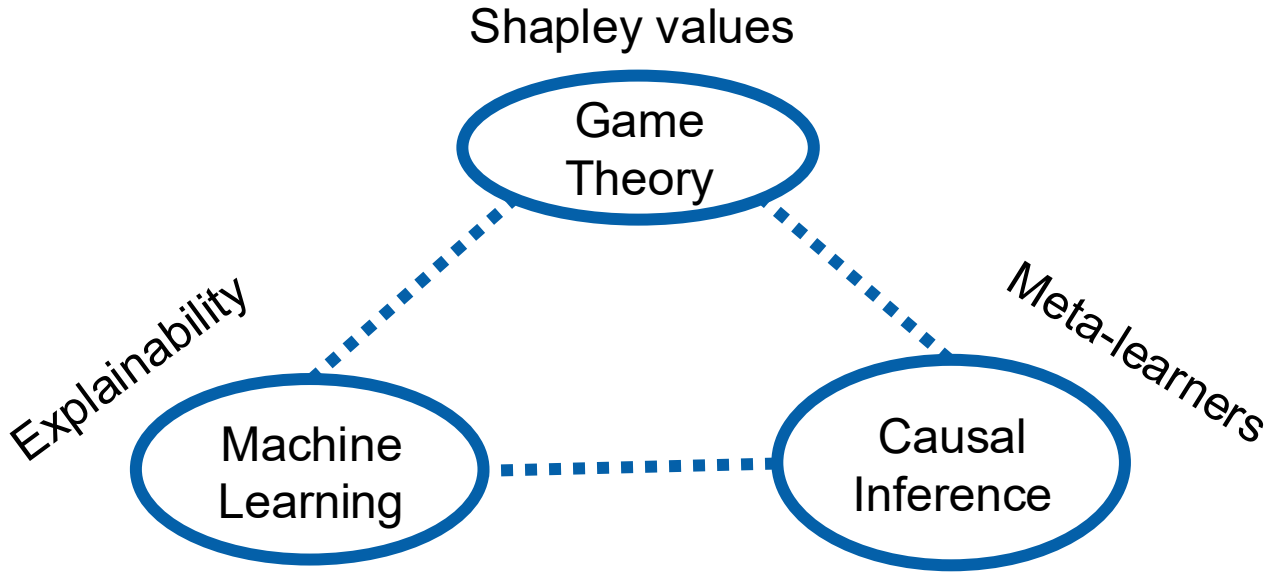
First published: 22 January 2026 | https://doi.org/10.1002/sim.70375

David Svensson (AstraZeneca)

Erik Hermansson (AstraZeneca)

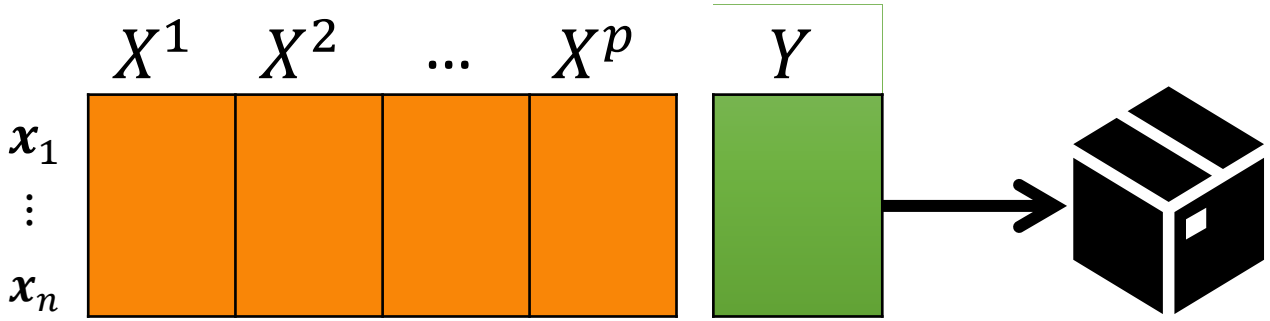
Nikolaos Nikolaou (UCL)

Ilya Lipkovich (Eli Lilly)

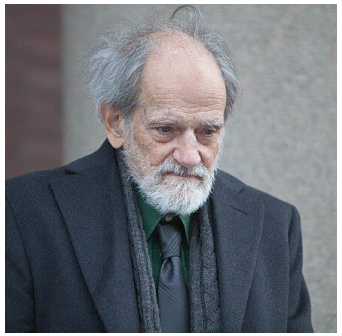


A game theoretic approach for model explainability

SHAP (SHapley Additive exPlanations) is a method that applies the game-theoretic principles to explain the output of ML/AI models.



Shapley values provide a fair method for distributing a game's total payoff among players by averaging their marginal contributions across all possible coalitions.

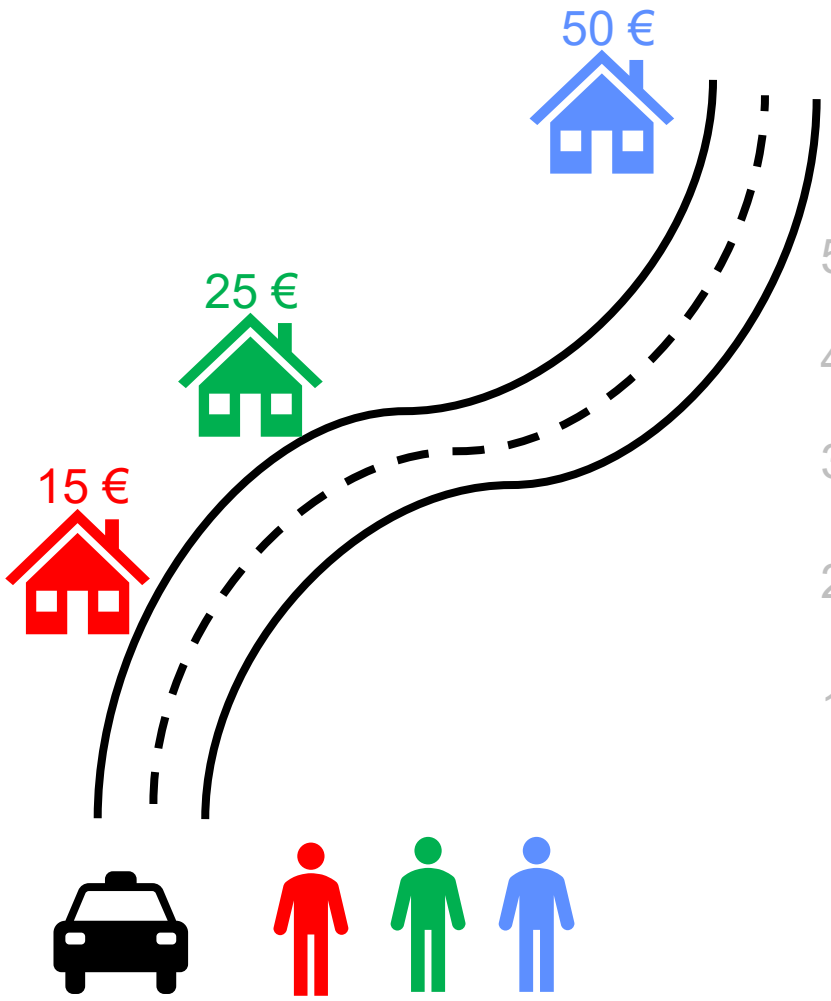


Lloyd Shapley
(Nobel Prize in Economy 2012)

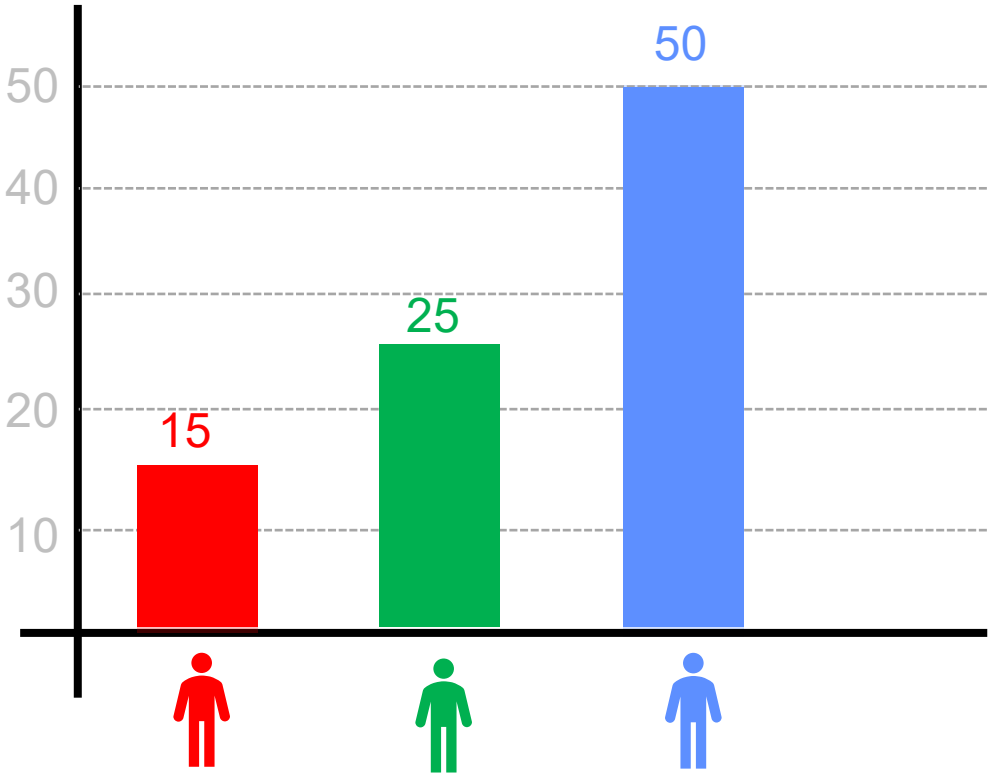
SHAP provide:

- **Global and local interpretability:** It offers a unified approach by explaining both the overall drivers of the model, but also the drivers behind an individual prediction
- **Theoretical grounding**: Because it is rooted in Shapley values, it ensures that feature contributions are mathematically consistent and fairly attributed.

Motivating Shapley values: the taxi sharing problem

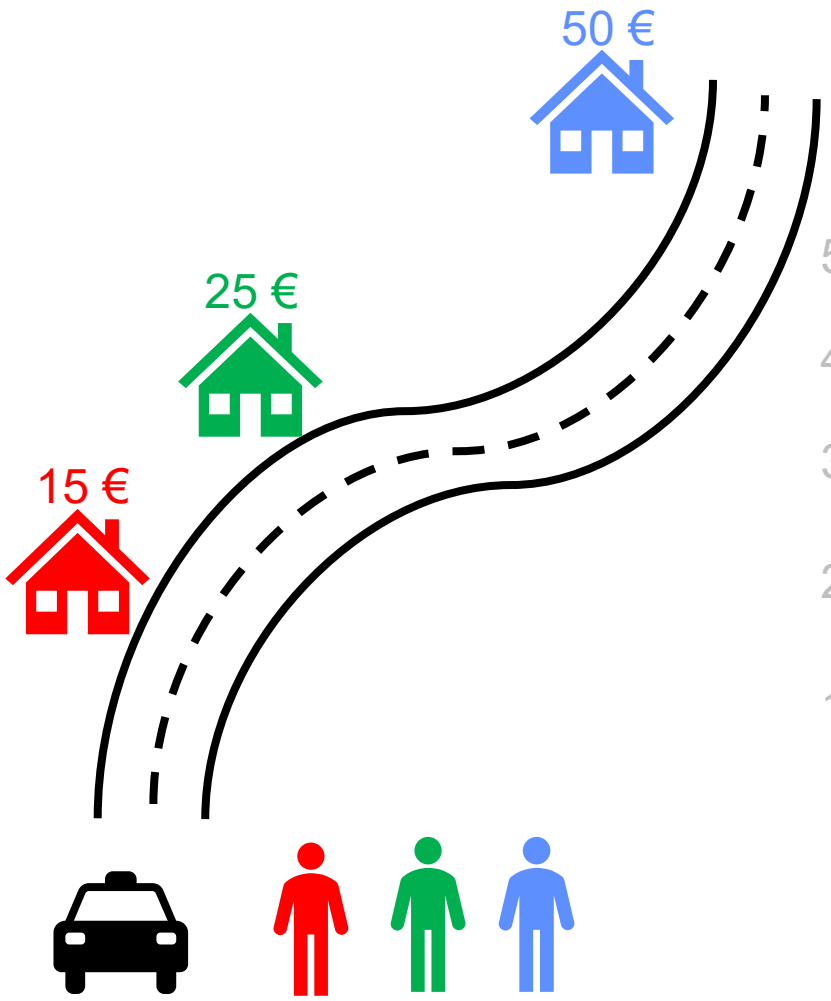


How should these three individuals fairly split a 50 € bill?

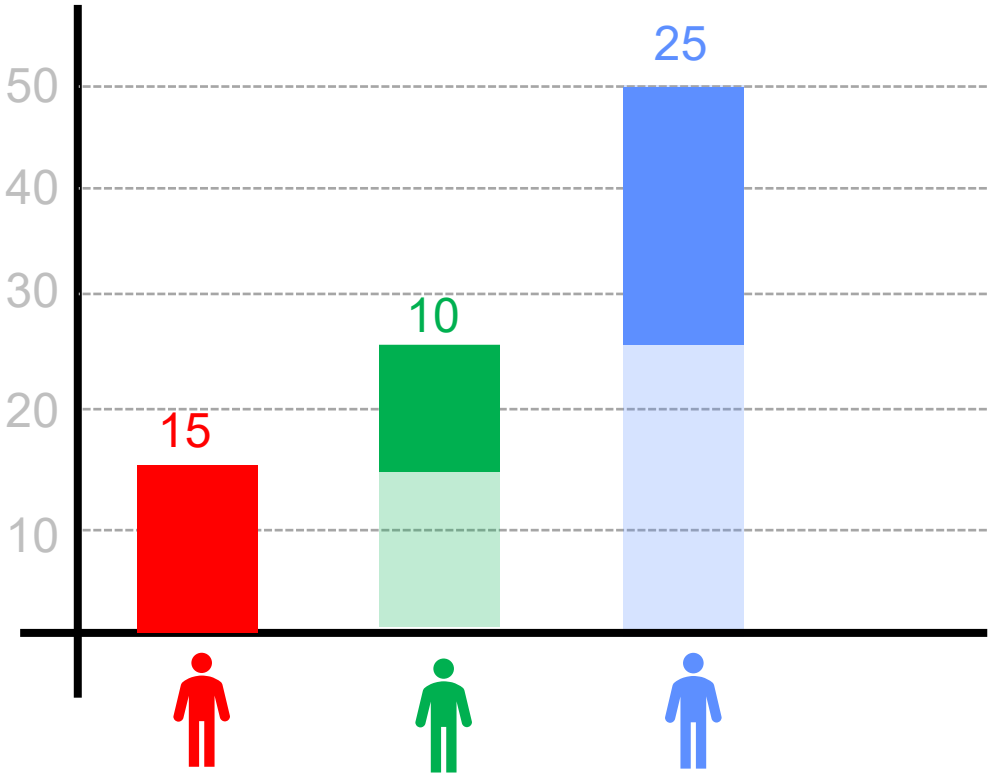


Example inspired by a presentation of Dr. Trefor Bazett

Motivating Shapley values: the taxi sharing problem

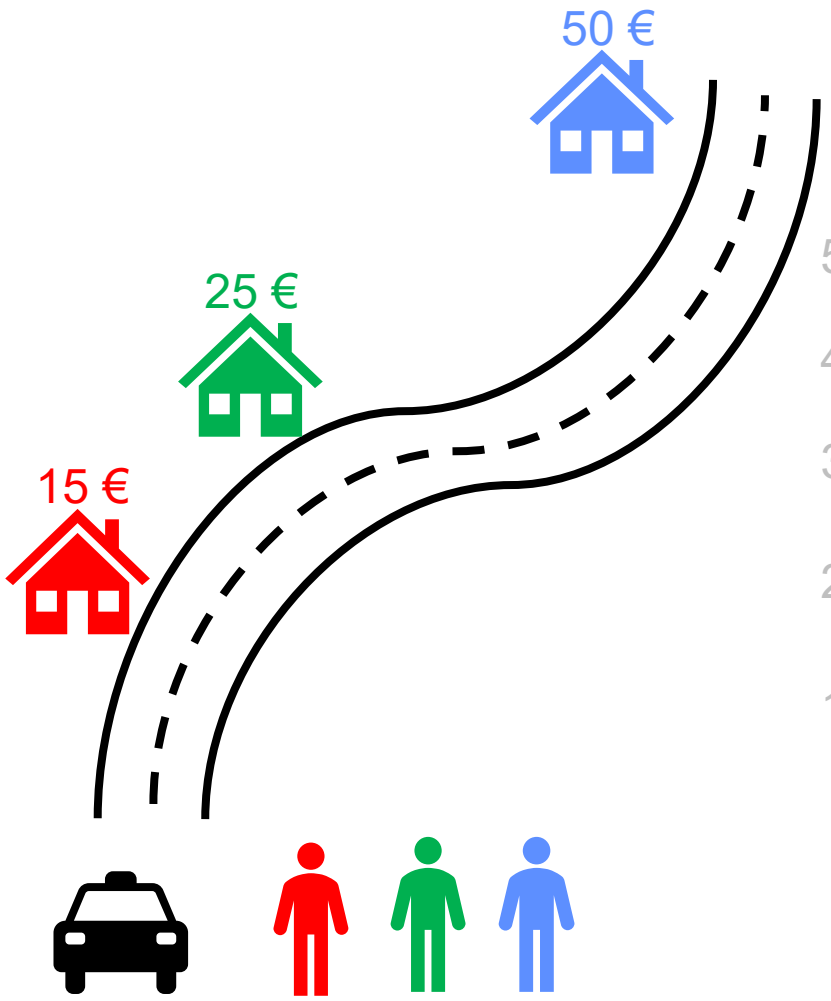


How should these three individuals fairly split a 50 € bill?

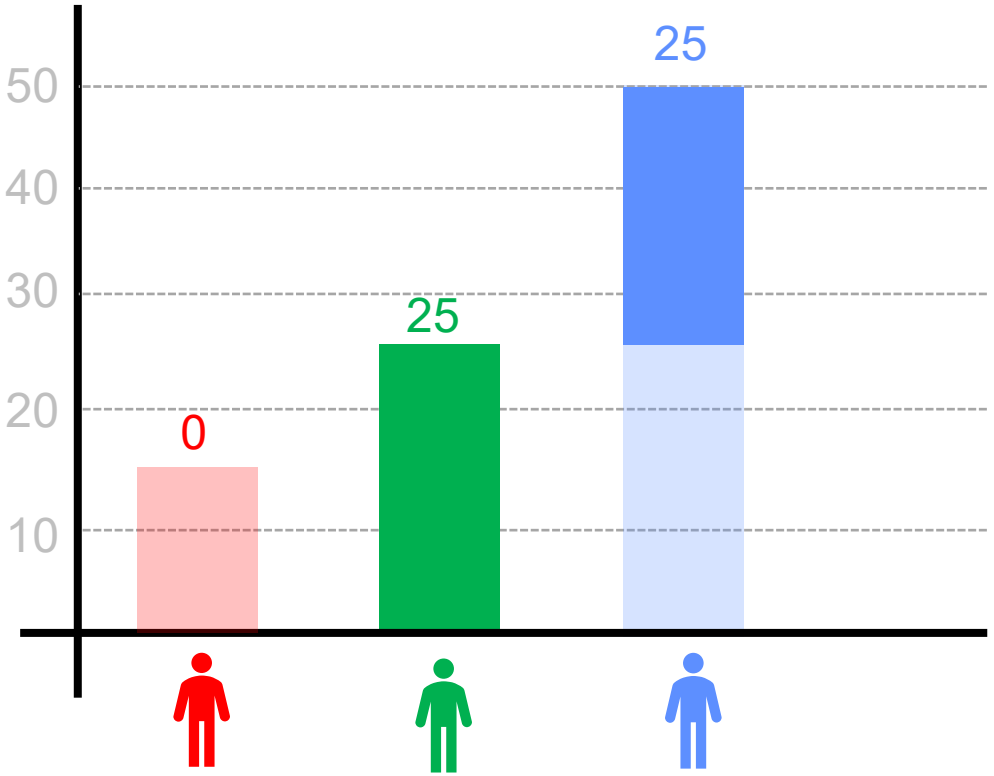


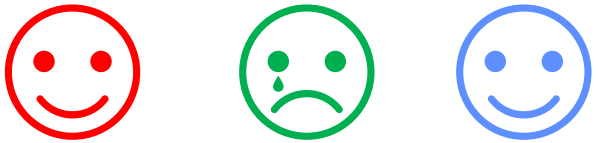
Red is not happy, because all of the benefit of him joining this coalition goes to the green and blue.

Motivating Shapley values: the taxi sharing problem



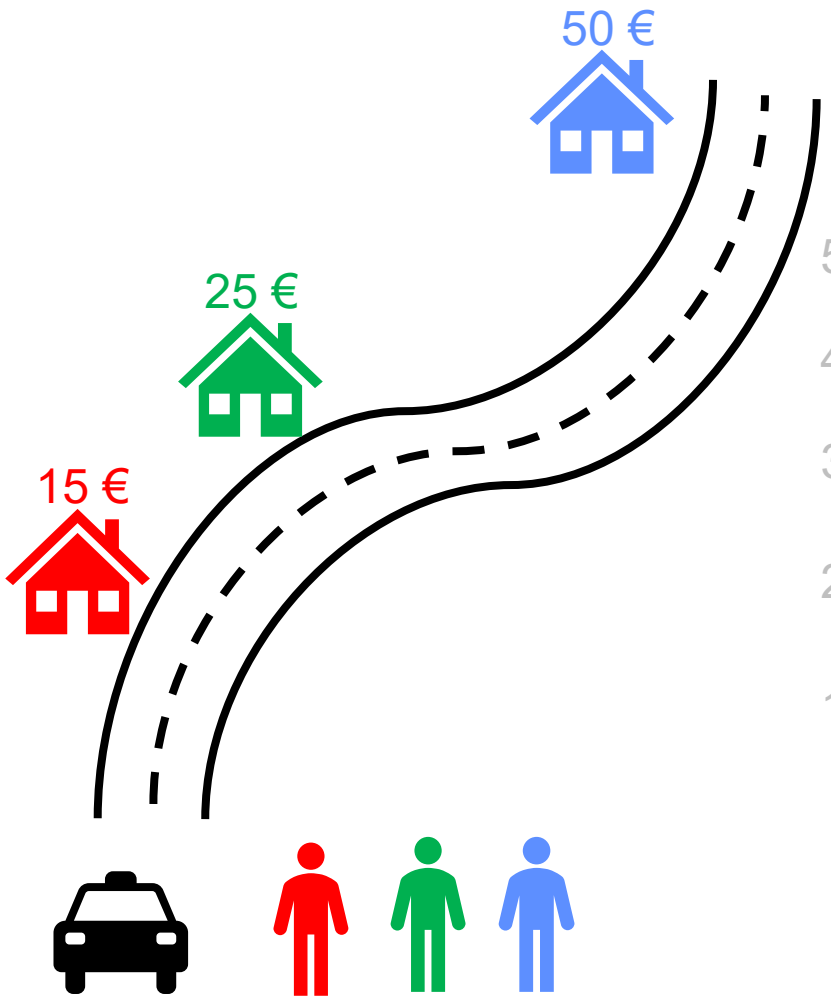
How should these three individuals fairly split a 50 € bill?



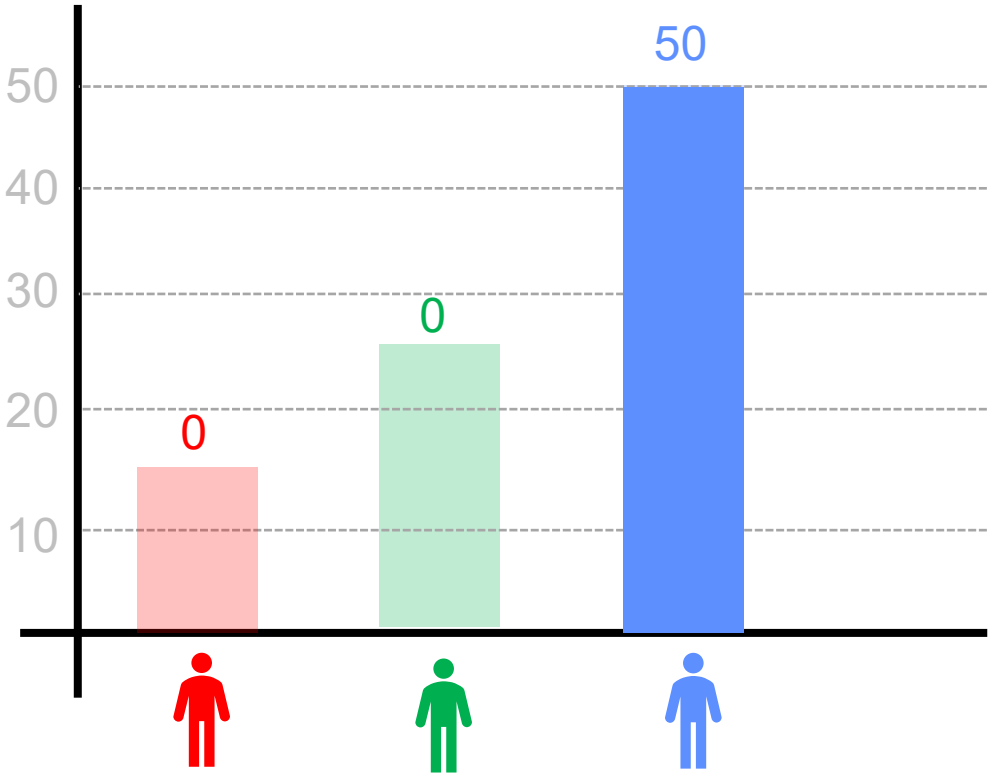


This time green is not happy since he is not sharing the benefits of forming the coalition.

Motivating Shapley values: the taxi sharing problem

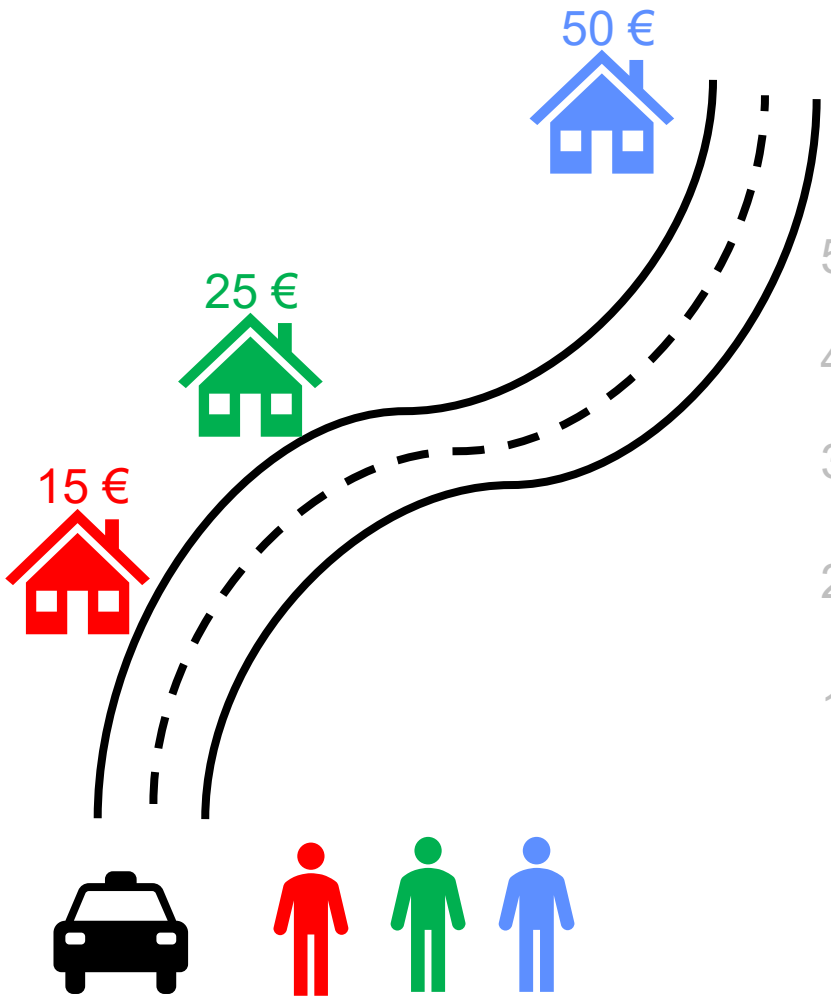


How should these three individuals fairly split a 50 € bill?

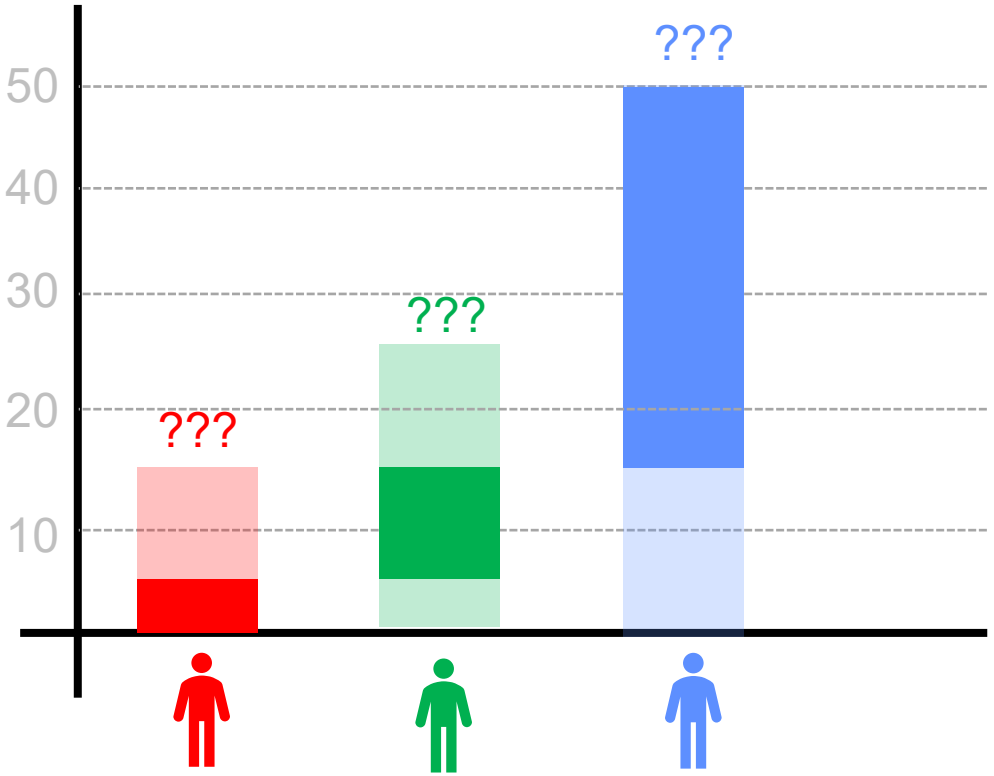


Finally, in this case the blue does not gain anything from this coalition.

Motivating Shapley values: the taxi sharing problem



How should these three individuals fairly split a 50 € bill?



There's a fair solution where each person pays less than they would have individually, offering a clear benefit to forming the coalition.

The idea of marginal contribution …

Order	Red adds	Green adds	Blue adds
Red → Green → Blue	15	10	25

A player's **marginal contribution** is the additional value they bring to a coalition when they join it.

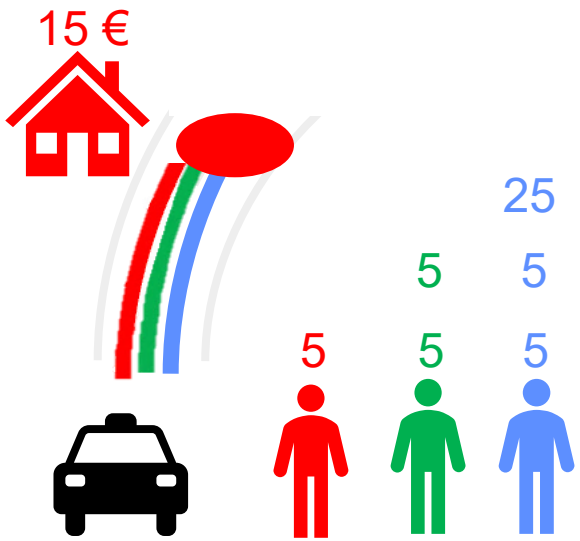
Average marginal contributions over every possible way that we can permute the order

- Red = (15 + 15 + 0 + 0 + 0 + 0) / 6 = 5
- Green = (10 + 0 + 25 + 25 + 0 + 0) / 6 = 10
- Blue = (25 + 35 + 25 + 25 + 50 + 50) / 6 = 35

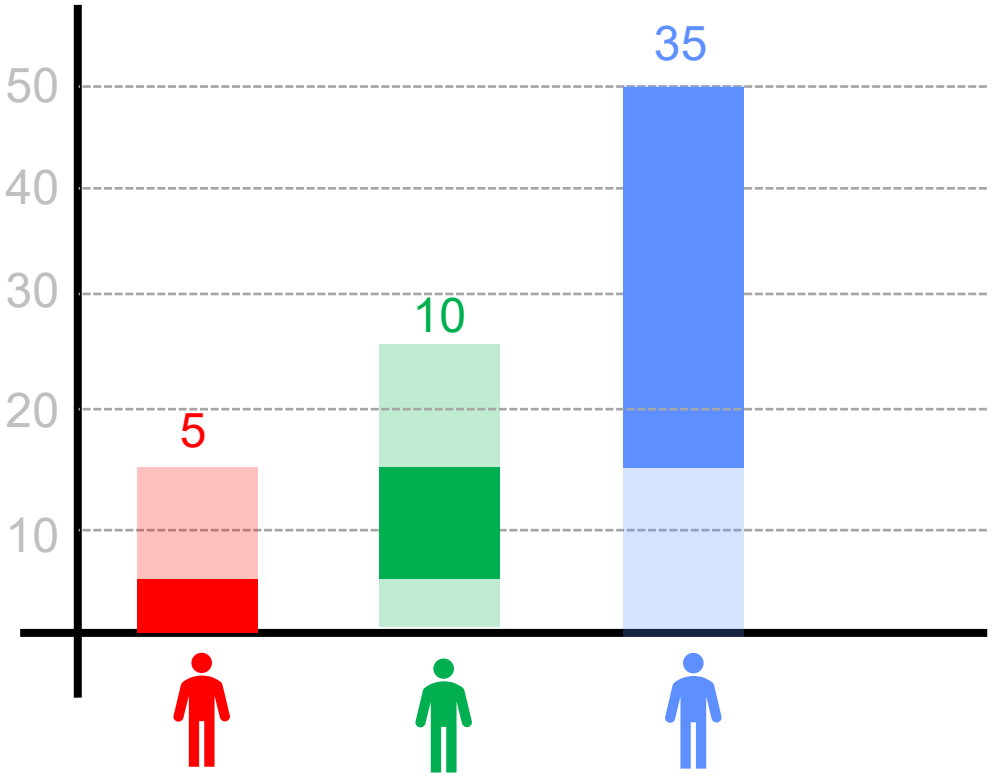
The **average marginal contributions is** the **unique** solution that **satisfies four axioms**, widely recognized as defining a fair allocation.

Efficiency	The contributions from all players exactly add up to the total payoff.
Symmetry	If two players always contribute the same amount, they should get the same share.
Dummy (Null player)	If a player never changes the payoff, their contribution is zero.
Additivity	If we combine two separate games, each player's total share is the sum of their shares in the individual games.

Motivating Shapley values: the taxi sharing problem

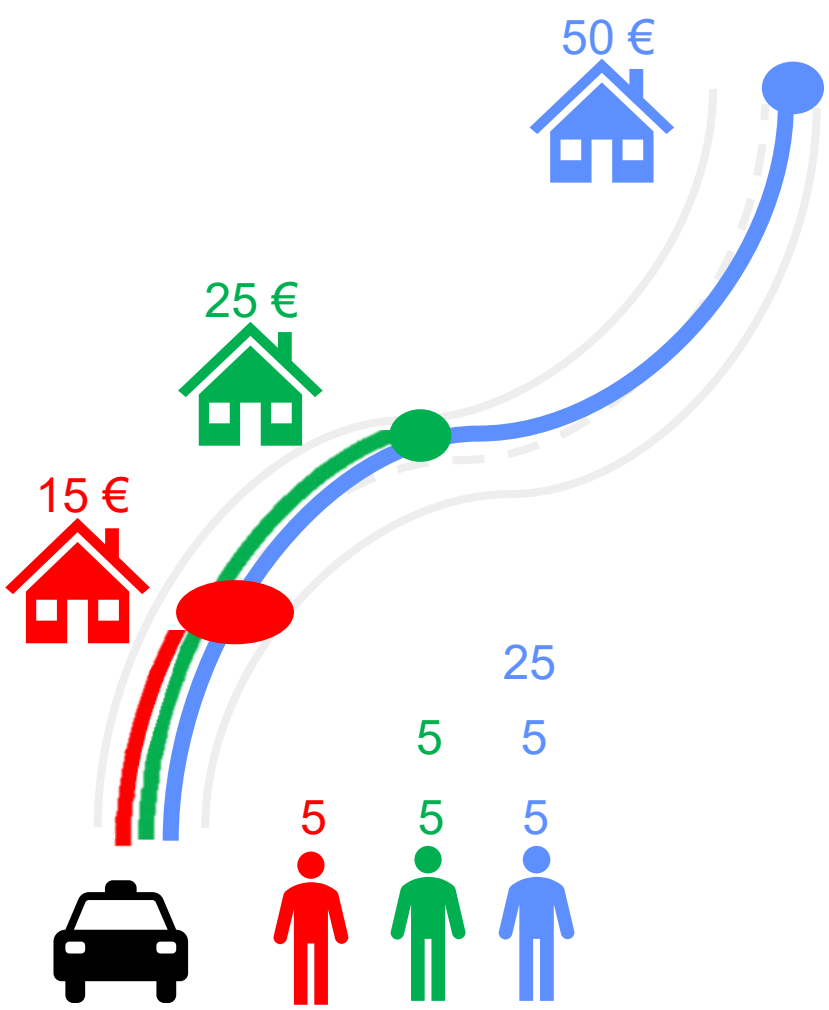


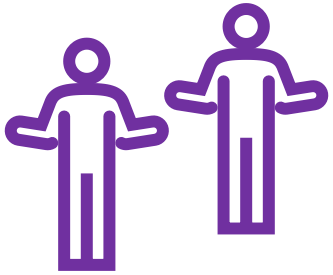
How should these three individuals
fairly split a 50 € bill?

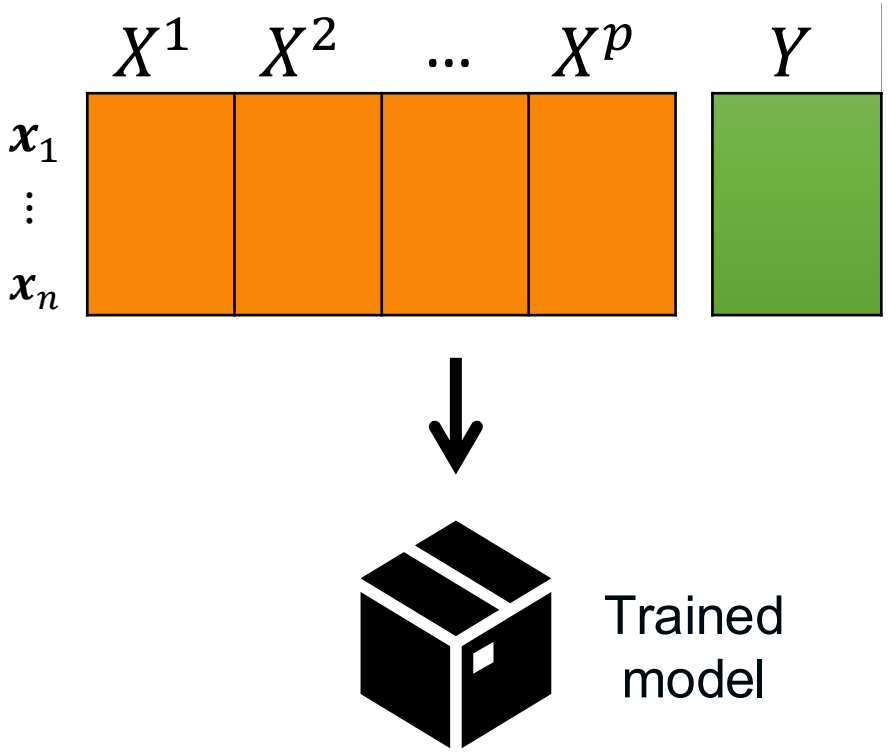


A different way to think about it …

From taxi sharing problem to explaining ML/AI models







From taxi sharing problem to explaining ML/AI models

Taxi sharing story

Players = passengers (Red, Green, Blue)

Total payoff = total taxi fare

Coalition = who's in the taxi so far

Marginal contribution = extra cost when someone joins

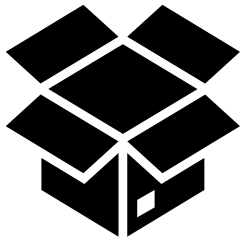
Explaining ML models

Players = features (e.g., Age, Income, Location)

Total payoff = model output for one example

Coalition = which features are "known" to the model

Marginal contribution = how much prediction changes when we reveal that feature

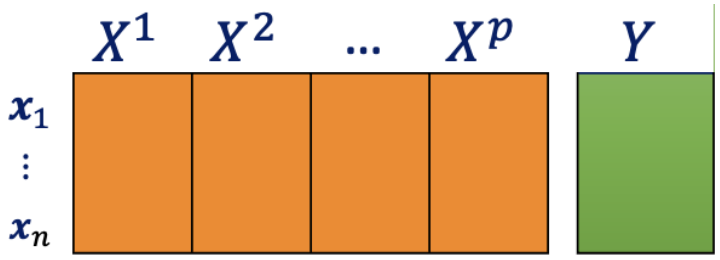


Trained model

Model-agnostic SHAP: Just like we tried every seating order for passengers, we can try every permutation of features to see how much each adds to the model prediction. Testing every order is often impossible, and there are practical approximations like KernelSHAP.

Model-specific SHAP: For certain model types, we can use the model's internal structure to calculate Shapley values exactly and quickly, e.g., TreeSHAP, DeepSHAP.

A toy example for deriving SHAP values



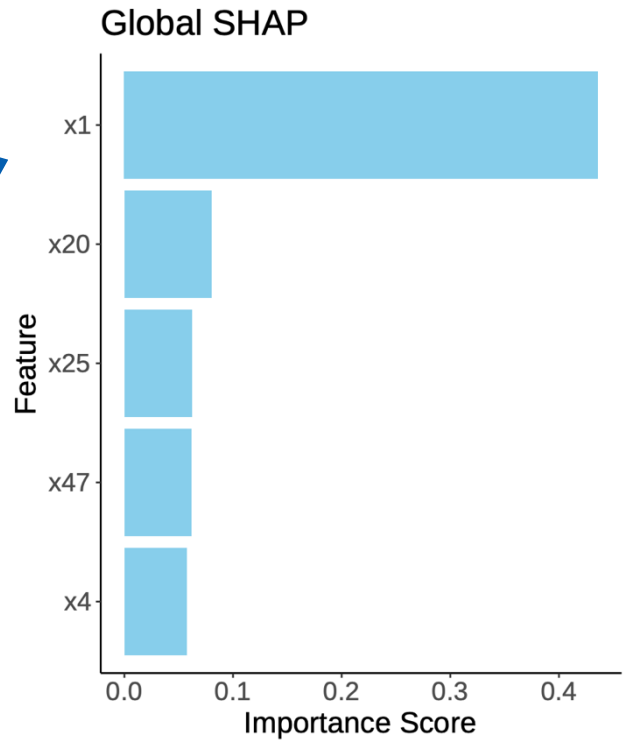
A toy simulated scenario:

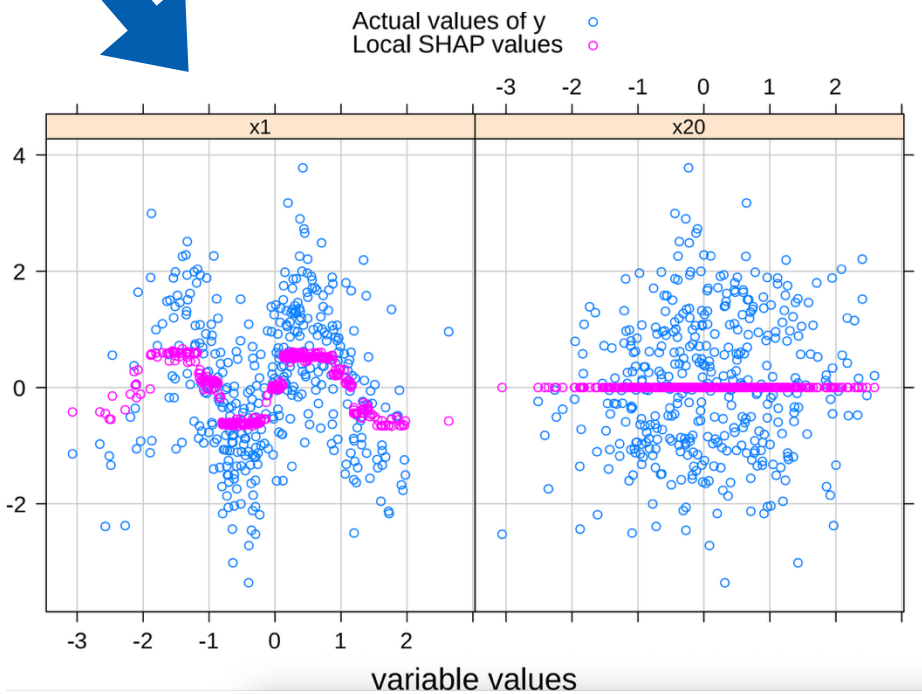
$$x_1, \ldots, x_{50} \sim N(0,1),$$

$$y = \sin(\pi x_1) + \epsilon$$

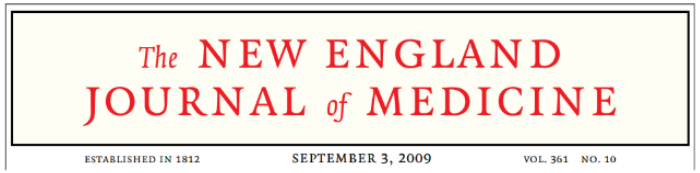
An XGBoost fitted to data;
& then derive TreeSHAP.

Generate both global ranking and local importance (dependency plot)

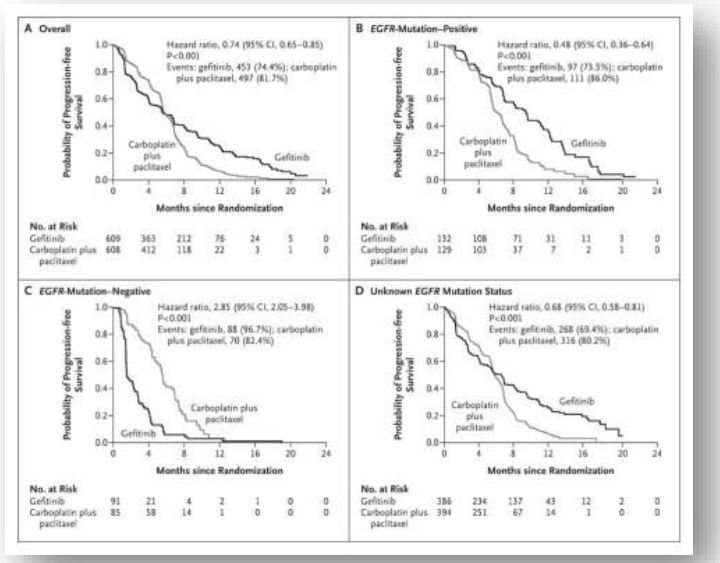




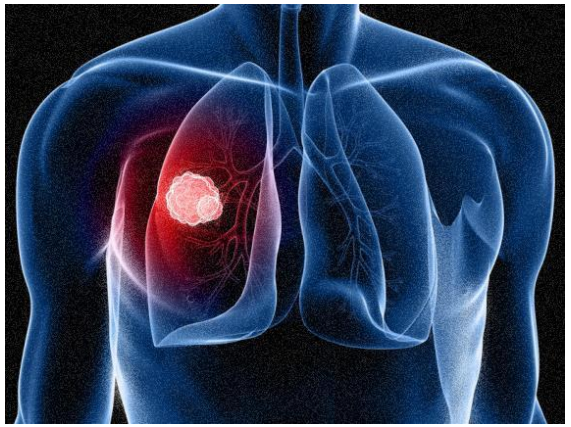
Motivating the problem: treatment effect heterogeneity and drug development

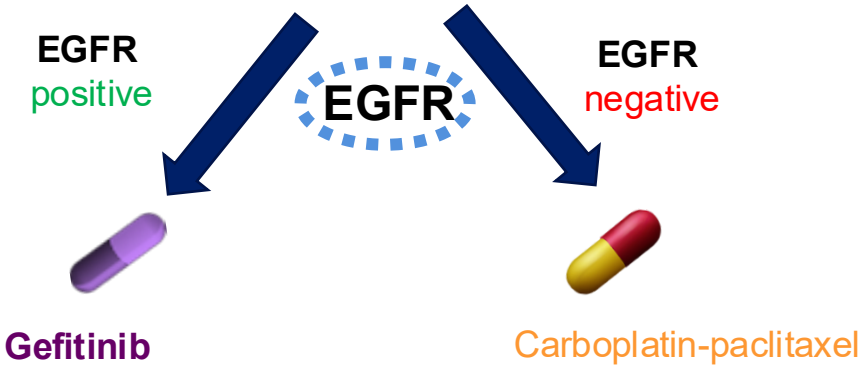


Gefitinib or Carboplatin–Paclitaxel in Pulmonary Adenocarcinoma



EGFR mutation is predictive ...



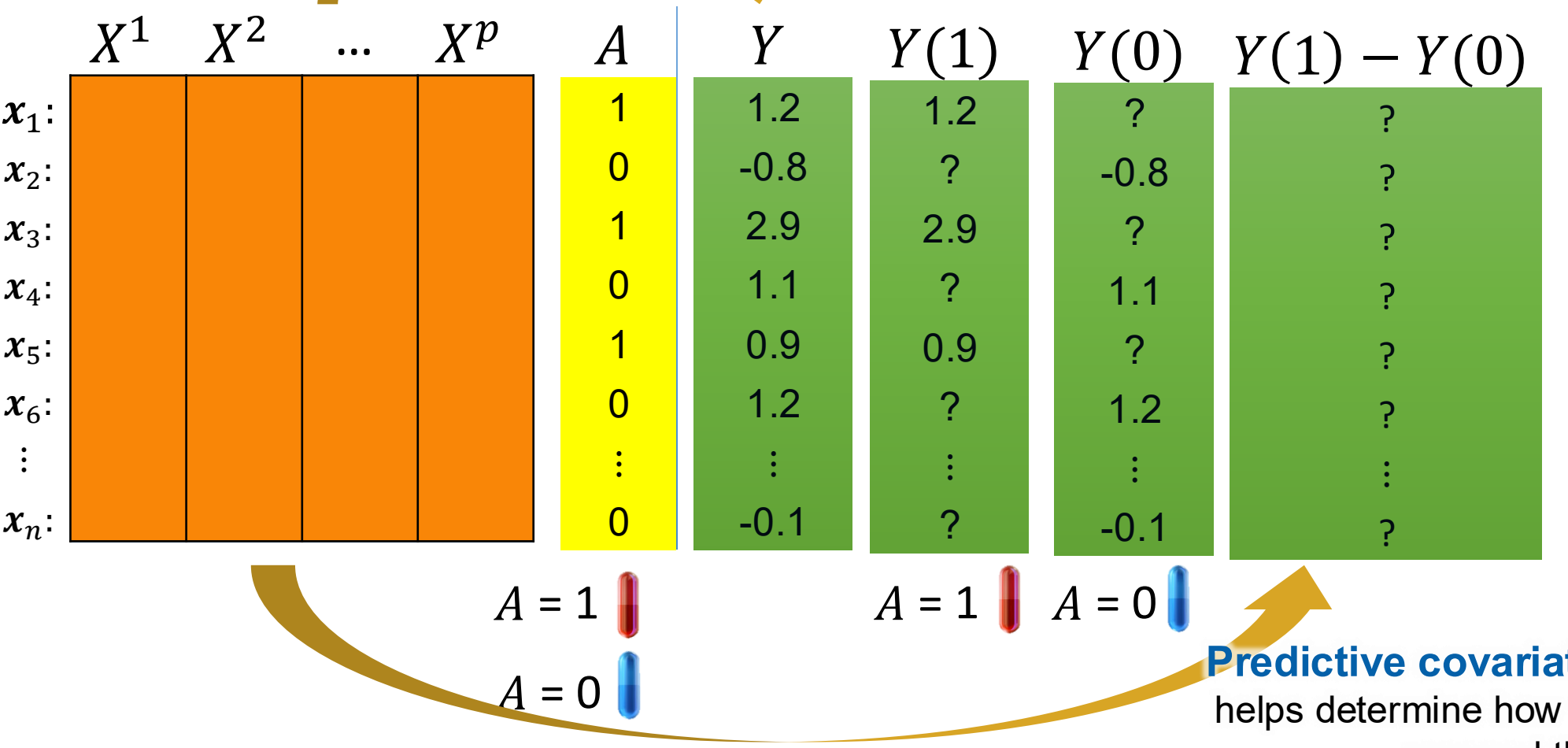


A framework for discovering predictive biomarkers based on SHAP values.

EGFR: Epidermal Growth Factor Receptor

Estimating individualized treatment effects

Prognostic covariates
are patient characteristics that predict the likelihood of an outcome, regardless of treatment.



Conditional Average Treatment Effect (CATE):
$\text{CATE} = \mathbb{E}[Y(1) - Y(0)|X]$

Predictive covariates (effect modifiers)
helps determine how well a patient is likely to respond the treatment

Estimating individualized treatment effects

T-Learner

Fit one model for the outcome if treated, another if not treated; subtract predictions to get CATE.

S-Learner

Fit a single model with treatment as just another feature; predict outcomes for both treatment values and subtract.

Outcome model 1:

$$\mu_1(x) = \mathbb{E}(Y \mid A = 1, X = x)$$

Outcome model 2:

$$\mu_0(x) = \mathbb{E}(Y \mid A = 0, X = x)$$

Propensity model:

$$\pi(x) = \mathbb{E}(A \mid X = x)$$

Received: 20 November 2023 | Revised: 28 May 2024 | Accepted: 21 June 2024

DOI: 10.1002/sim.10167

TUTORIAL IN BIOSTATISTICS

Statistics in Medicine WILEY

Modern approaches for evaluating treatment effect heterogeneity from clinical trials and observational data

Ilya Lipkovich[1] | David Svensson[2] | Bohdana Ratitch[3] | Alex Dmitrienko[4]

Electronic Journal of Statistics
Vol. 17 (2023) 3008–3049
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2157

Towards optimal doubly robust estimation of heterogeneous causal effects

Edward H. Kennedy
Department of Statistics & Data Science, Carnegie Mellon University

Example: DR-learner

- STEP 1: first estimates "pseudo-outcomes" $\hat{\psi}_{DR}$:

$$\hat{\psi}_{DR}(\mathbf{x}_i, y_i, a_i) = \frac{a_i - \hat{\pi}(\mathbf{x}_i)}{\hat{\pi}(\mathbf{x}_i)(1-\hat{\pi}(\mathbf{x}_i))} y_i + \left(1 - \frac{a_i}{\hat{\pi}(\mathbf{x}_i)}\right)\hat{\mu}_1(\mathbf{x}_i) - \left(1 - \frac{1-a_i}{1-\hat{\pi}(\mathbf{x}_i)}\right)\hat{\mu}_0(\mathbf{x}_i)$$

- STEP 2: regress *X*s on these "pseudo-outcomes" to derive the CATE:

$$\widehat{\mathrm{CATE}}_{\mathrm{DR}} = \hat{\boldsymbol{\tau}}_{\mathbf{DR}}(\boldsymbol{x}) = \hat{\mathbb{E}}\left(\hat{\psi}_{DR} \middle| X = \boldsymbol{x}\right)$$

Outcome model 1:
$\boldsymbol{\mu_1}(\boldsymbol{x}) = \mathbb{E}(Y \,|A = 1, X = \boldsymbol{x})$

Outcome model 2:
$\boldsymbol{\mu_0}(\boldsymbol{x}) = \mathbb{E}(Y \,|A = 0, X = \boldsymbol{x})$

Propensity model:
$\boldsymbol{\pi}(\boldsymbol{x}) = \mathbb{E}(A \mid X = \boldsymbol{x})$

CATE model:
$\boldsymbol{\tau}_{\mathbf{DR}}(\boldsymbol{x}) = \mathbb{E}\left(\hat{\psi}_{DR} \middle| X = \boldsymbol{x}\right)$

In the DR-Learner, CATE estimation is reducible to fitting a single final model on pseudo-outcomes — turning the problem into a standard prediction task.
This final model can be interpreted using tools like SHAP.

Generic idea[1] (applicable to any CATE approach)

Approach: “explaining CATE in terms of the baseline predictors” by a surrogate Model:

Regressing $\hat{\tau}(\boldsymbol{x})$ against $x_1, \ldots, x_p$, and derive SHAP from the fitted model.

Our implementation: $\hat{\tau}(x) \sim x$ using XGBoost (→ TreeSHAP).

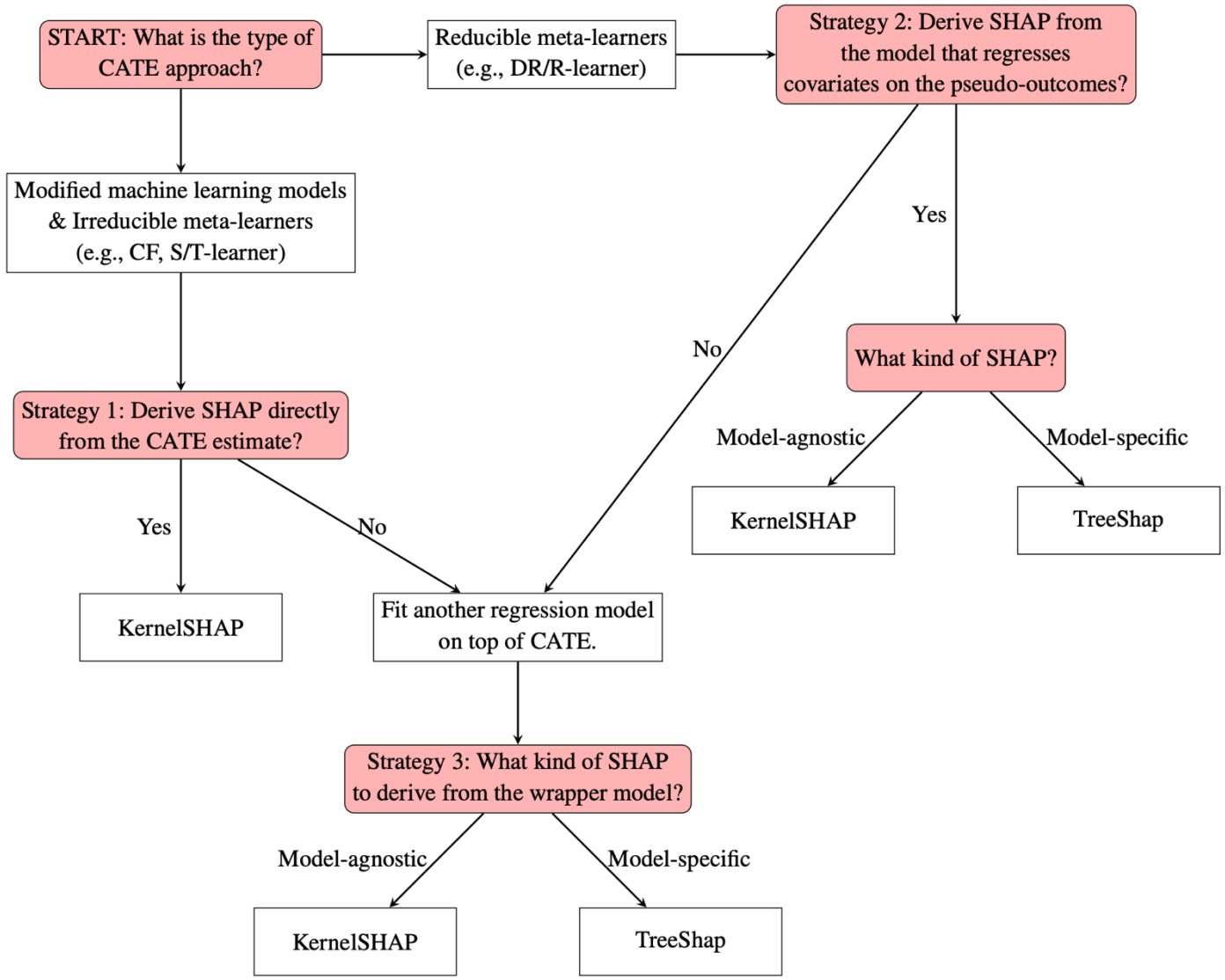
Bypasses issues (related to multistage approaches),
i.e., agnostic to CATE approach details.

Scales up well.

1: This approach shares ideas with early approaches in Subgroup Identification such as ”Virtual Twins” (Foster et al 2011) where novel subgroups are derived via regressing estimates of ITE against x (pruned CART tree)

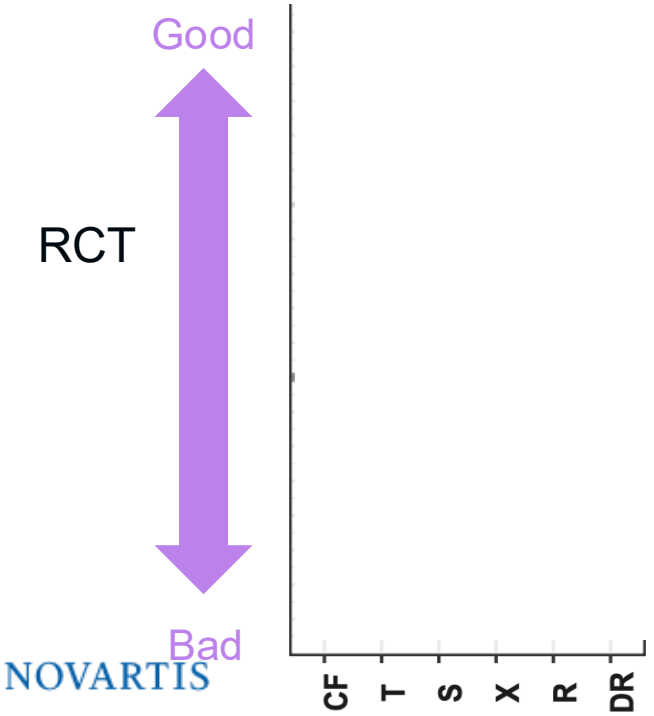
Using SHAP values to explain CATE models

A roadmap illustrating the various principled approaches to derive SHAP values for CATE models. This flow chart provides a guide through the different methods of CATE estimation and stages of SHAP implementation.



Simulation study to answer various questions

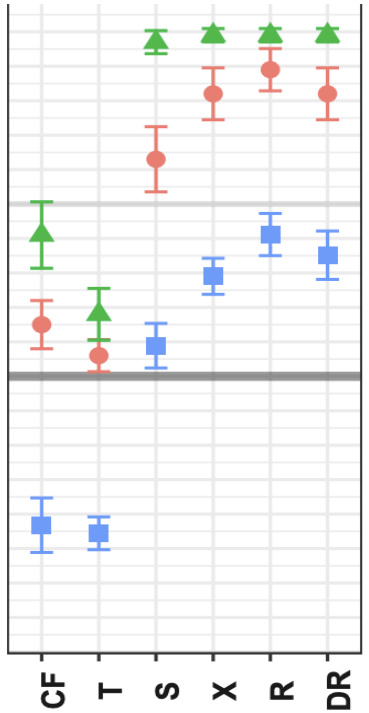
(1) Does the meta-learner choice affect performance of SHAP biomarker discovery?
(2) How the prognostic strength affects the performance of the different methods?
(3) How SHAP methods perform in comparison to model-specific VIP?
(4) Do the strategies for deriving SHAP for R- and DR-learner differ?
(5) Can SHAP values help us to derive the marginal predictive effect?



3 evaluation measures

- Top1
- Net3
- Margin

Observational



- Top1
- Net3
- Margin

NOVARTIS

Conclusions

- ✓ Assessing TEH in a challenging problem, and needs to be approached through some structured workflows, like WATCH
- ✓ Explainable ML/AI can play an important role on assessing TEH.
- ✓ SHAP values provide a powerful framework for explaining CATE models; however, the method of deriving SHAP values can vary across different CATE modeling strategies.
- ✓ S-Learning, DR-Learning, and R-Learning showed good performance across metrics and scenarios.
- ✓ The surrogate approach is generic, and scales up well.

Thank you!!!

Code available in:
https://github.com/DaveJSvensson/SHAP_CATE