NAIC CASTF Book Club Meeting
2026-02-24

Shapley Values and the Reason for Explanation

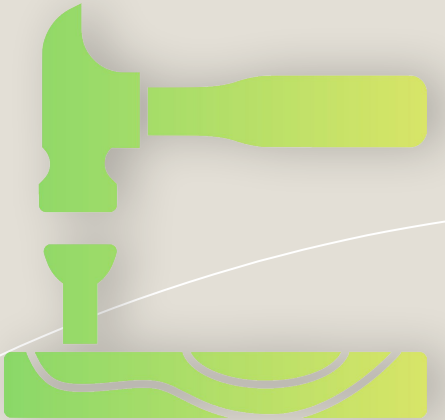
Raymond Sheh | ray@raymondsheh.org

These opinions are my own and not those of NIST, JHU, or any other organization or entity. I'm presenting this in my personal capacity.

Goals of this presentation

To help better understand the evolving toolbox of explainable AI.

- Outline a taxonomy for explanations.
- Provide an intuitive understanding of how Shapley Values and Shapley Additive Explanations (SHAP) work.
- Dive into the math *just* enough to support that intuition.
- Discuss ways of interpreting Shapley Values.
- Suggest some alternative approaches.



These slides are also intended to serve as notes and go into more detail than I will be covering in the talk.

About me ...

- Associate Research Scientist at Johns Hopkins University (JHU).
- Guest Researcher at the National Institute of Standards and Technology (NIST).
- Currently working on AI Risk Management, forensic machine perception, and robot safety.
- Ph.D. in Machine Learning for Robot Control.
- 20+ years teaching/research in AI, robotics, cybersecurity, measurement science, risk management, and governance.
- Global experience in public safety, government, healthcare, academia/education, and the energy sector.

These opinions are my own and not those of NIST, JHU, or any other organization or entity. I'm presenting this in my personal capacity.

Terminology

- **Artificial Intelligence (AI):** A machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions ... *(Adapted from NIST SP 800-218A)*
- **(AI) Model:** A component ... (that) uses computational, statistical, or machine-learning techniques to produce **outputs** from a given set of inputs. *(Adapted from NIST SP 800-218A)*
- **Explanation:** To give the reason for or cause of. *(Merriam-Webster Dictionary)*
- **Instance:** A set of **attribute values** input to an AI model to produce an **output**, such as "Ray's medical record as of 2026-01-01 00:00:00".
- **Attribute:** An atomic property of a generic **instance** that takes on a **value**, e.g. "height".
- **Value:** The numeric (or other) **instantiation** of an **attribute**, e.g. "5.75 feet".

Types of Explanations

Source

Introspective
("The" explanation)

Post-Hoc
Rationalization
("An" explanation)

Depth

Attribute Use (e.g. saliency)

Attribute Value

Model (Generation)

Scope

Justification

Teaching

What are your application's requirements?

Types of Explanations

Source

Introspective
(**The** explanation)

Post-Hoc
Rationalization
(**An** explanation)

- **Introspective** explanations seek to reflect (possibly incompletely, inaccurately, imprecisely, or abstractly) the true underlying decision making process.
- **Post-Hoc Rationalization** explanations only seek to be consistent with the observed behavior of the system.

From the outside, it is usually impossible to distinguish between these explanations.

“Q: Why did you decide to order a pizza?”
“A1: Because it was on special.”
“A2: Because Ray had one yesterday and enjoyed it.”
“A3: Because Ray had pasta yesterday and I figured he might like something different.”

Types of Explanations

Depth

Attribute Use

Attribute Value

Model

- **Attribute Use** tells us **which** parameters were considered (and perhaps how important they were).
  - *“I looked at the toppings and price.”*
- **Attribute Value** tells us something about how the specific **value** of the parameter affected the decision.
  - *“It has pineapple and cost less than $10.”*
- **Model** tells us something about how the part of the model that affected the decision came about (e.g., from the training and any symbolic background data).
  - *“Ray has ordered pineapple on his pizza 90% of the time in the past year and only has a $10 lunch budget.”*

Types of Explanations

Scope

Justification

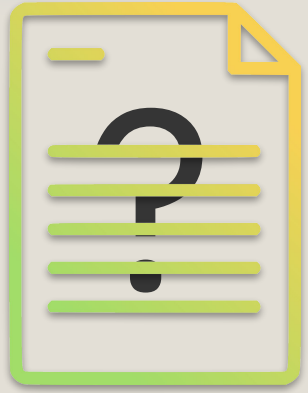
Teaching

- **Justification** explanations are valid for one or a finite set of decisions but does not allow us to predict the behavior of the model.
  - A slight change in the attributes can invalidate the explanation.
- **Teaching** explanations generalize across some subset of the attribute space and allow us to predict behavior (or at least tell us the bounds of where they are valid).
  - For example, the decision boundaries of a decision tree.

From the outside, it **can** be impossible to tell the difference between a **justification** and **teaching** explanation.

“I ordered a pizza for Ray because it’s Thursday. I won’t on Friday.”

Why Explain?



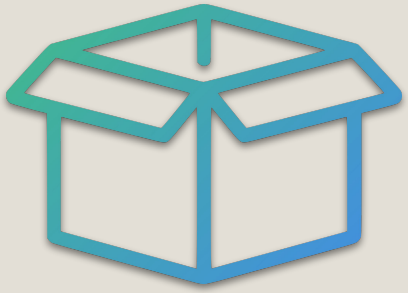
- To help manage risk and improve governance?
  - Identify/Assess/Evaluate/Respond/Monitor
- To improve the system and/or learn something?
- To satisfy obligations/regulations?
- To appropriately calibrate (vs raise) trust?

- Exploring the behavior of the model, particularly to new/different data.
- Determining if/when the model is considering the right attributes and if its decision process is reasonable.
- Debugging and correcting mistakes in the model.
- Having the model teach us something new.
- Regulatory compliance.
- Forensic root cause analysis and credit/blame assignment.
- Determining when and how much to use or trust the model.
- For transparency to affected populations.
- ...

Which purposes expect true causality? Which ones are fine with correlation?

What are the other requirements for each purpose?

Scoping note



- We are only considering explanations for the model.
- We are not considering explanations of reality.
- We are not looking for real-world causality.
  - Causality with respect to the model might be nice though.
- An explanation that reflects the model but doesn't make sense in reality is a good explanation for a (potentially) bad model.
  - A bad model might be confused between highly correlated features. A good explanation might highlight the model using the "wrong" one.

Analogy with building models

Building models	Explaining models
Examples of inputs and outputs ⇒ Model parameters that predict outputs from inputs.	Examples of decisions ⇒ parameters that describe how that output could be generated.
Some models have parameters with real-world meaning.	Some explanations reflect the true way that the output was generated.
Some modelling techniques just attempt to fit the data.	Some explanations just attempt to be consistent with the decision(s).
Application requirements and modelling technique capabilities vary widely. A mismatch is a bad idea.	Application requirements and explanation technique capabilities vary widely. A mismatch is a bad idea.
Modelling techniques make different assumptions to constrain interpolation/extrapolation and deal with noise (sometimes called the “inductive bias”).	Explanation techniques make different assumptions to constrain the space of explanations.

What is SHAP?

- “SHapley Additive exPlanation”
- Computes **an estimate** for how much each attribute value in a specific instance contributed to a given output of the model for that instance.
- This estimate is only valid for a specific instance, it doesn’t generalize.
- SHAP is just one way of computing this estimate.
  - Constrained by various assumptions and approximations.
- SHAP only makes use of observations of the model’s behavior in response to hypothetical instances.

What explanation requirements match with this capability?

What is SHAP?

SHAP satisfies some mathematically nice properties. For a given instance:

- **Efficiency**: Sum of estimates add up to the total for all attributes.
- **Symmetry**: Equivalent attributes have the same estimate.
- **Dummy player**: Meaningless attributes have a zero estimate.
- **Linearity**: Estimates for a given attribute on the parts of a task will add up to the estimate for that attribute on the whole task.

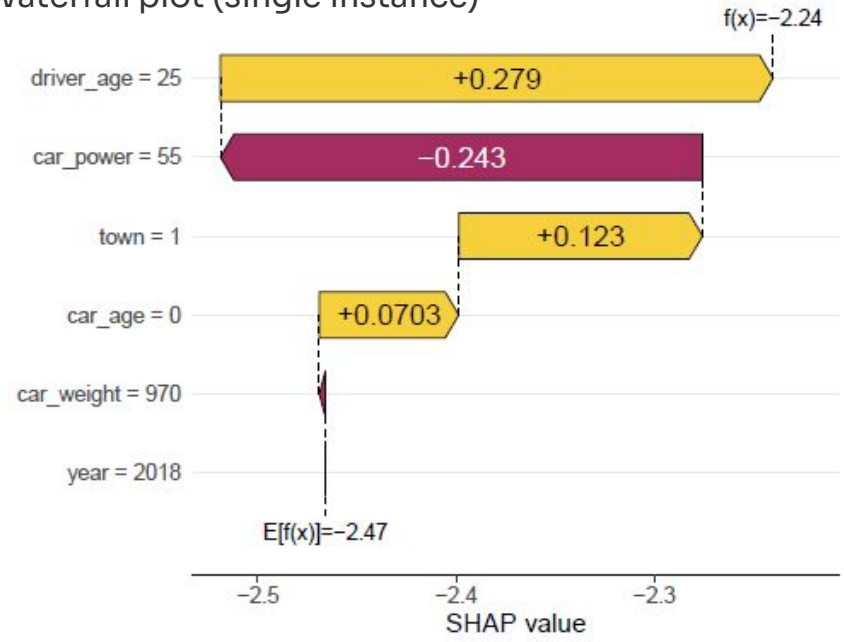
What is SHAP?

In abstract terms:

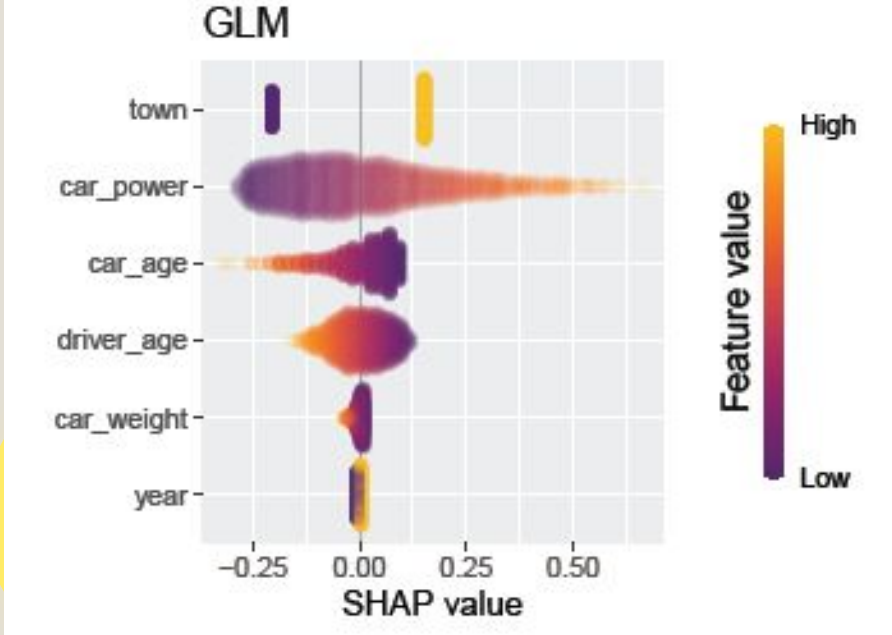
- A task is performed by a team. We want to estimate how much an individual contributed to the output of the task.
  - Prediction, decision making, baking a cake, rowing a kayak, ...
- For each possible sub-team that includes that individual, see how the output changes if we remove them.
  - It can be mathematically convenient to express it the other way round ...
- We take the average of all of the differences across all of these sub-teams, weighted based on the size of each sub-team.
  - We assume it's easier to join a small team and make a difference, and harder to do so when joining a big team where influence may be diluted.

What is SHAP?

Waterfall plot (single instance)



Beeswarm plot (many instances)



From Mayer, M., Meier, D., Wuthrich, M. V. (2023) “SHAP for Actuaries: Explain any Model”, Fachgruppe “Data Science”, Swiss Association of Actuaries SAV, March 15, 2023

The original Shapley Value formula

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S))$$

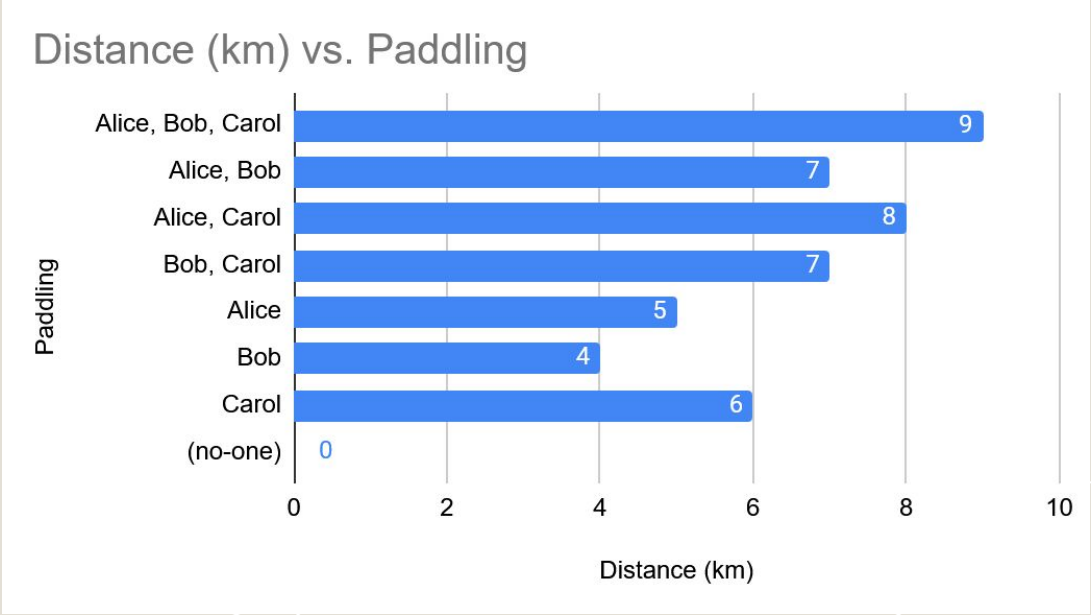
1 2 3 4

1. Estimated effect of individual *i* to the output of task *v.*
2. For every hypothetical sub-team that **doesn't** include individual *i.*
3. Weight proportional to the number of ways we can create this team.
    - This weights bigger teams more than smaller ones.
4. The marginal (additional) output with and without *i*.

There are several variations in form and notation. We're using the one from Wikipedia: https://en.wikipedia.org/wiki/Shapley_value

Example of SHAP

- Imagine we have a team of Alice, Bob, and Carol in a 3 person kayak.
- We want to know how much Alice, Bob, and Carol contribute to the distance the kayak moves in 1 hour.
- We run several experiments to see how much each person contributes to moving the kayak.



Example of SHAP

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S))$$

$$= \frac{1}{n} \sum_{S \subseteq N \setminus \{i\}} \binom{n-1}{|S|}^{-1} (v(S \cup \{i\}) - v(S))$$

- $\varphi_i(v)$ = contribution of individual i to the process/game *v*
- N = the overall team
- n = total number of individuals in the overall team
- S = a sub-team of N that doesn't include i $S \subseteq N \setminus \{i\}$
- |S| = number of individuals in sub-team S

(Don't worry about the math on this slide - pause the recording if you want to follow it!)

Example of SHAP

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S))$$

What is Alice's contribution? Set i = A

S = [BC], |S| = 2 ⇒ (2!(3-2-1)!/3!)*(9-7) = 1/3*2 = ⅔

S = [B], |S| = 1 ⇒ (1!(3-1-1)!/3!)*(7-4) = 1/6*3 = ½

S = [C], |S| = 1 ⇒ (1!(3-1-1)!/3!)*(8-6) = 1/6*2 = ⅓

S = [0], |S| = 0 ⇒ (0!(3-0-1)!/3!)*(5-0) = 1/3*5 = 1 ⅔

Total:

φ_A(v) = ⅔+½+⅓+1 ⅔ = 3 ⅙

- N = [ABC], n = 3
- v([ABC]) = 9 km
- v([AB]) = 7 km
- v([AC]) = 8 km
- v([BC]) = 7 km
- v([A]) = 5 km
- v([B]) = 4 km
- v([C]) = 6 km
- v([0]) = 0 km

(Don't worry about the math on this slide - pause the recording if you want to follow it!)

Example of SHAP

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S))$$

What is Bob's contribution? Set i = B

S = [AC], |S| = 2 ⇒ (2!(3-2-1)!/3!)*(9-8) = 1/3*1 = ⅓

S = [A], |S| = 1 ⇒ (1!(3-1-1)!/3!)*(7-5) = 1/6*2 = ⅓

S = [C], |S| = 1 ⇒ (1!(3-1-1)!/3!)*(7-6) = 1/6*1 = ⅙

S = [0], |S| = 0 ⇒ (0!(3-0-1)!/3!)*(4-0) = 1/3*4 = 1 ⅓

Total:

φ_B(v) = ⅓+⅓+⅙+1 ⅓ = 2 ⅙

- N = [ABC], n = 3
- v([ABC]) = 9 km
- v([AB]) = 7 km
- v([AC]) = 8 km
- v([BC]) = 7 km
- v([A]) = 5 km
- v([B]) = 4 km
- v([C]) = 6 km
- v([0]) = 0 km

(Don't worry about the math on this slide - pause the recording if you want to follow it!)

Example of SHAP

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S))$$

What is Carol's contribution? Set i = C

S = [AB], |S| = 2 ⇒ (2!(3-2-1)!/3!)*(9-7) = 1/3*2 = ⅔

S = [A], |S| = 1 ⇒ (1!(3-1-1)!/3!)*(8-5) = 1/6*3 = ½

S = [B], |S| = 1 ⇒ (1!(3-1-1)!/3!)*(7-4) = 1/6*3 = ½

S = [0], |S| = 0 ⇒ (0!(3-0-1)!/3!)*(6-0) = 1/3*6 = 2

Total:

$\varphi_C(v)$ = ⅔+½+½+2 = 3 ⅔

- N = [ABC], n = 3
- v([ABC]) = 9 km
- v([AB]) = 7 km
- v([AC]) = 8 km
- v([BC]) = 7 km
- v([A]) = 5 km
- v([B]) = 4 km
- v([C]) = 6 km
- v([0]) = 0 km

(Don't worry about the math on this slide - pause the recording if you want to follow it!)

Example of SHAP

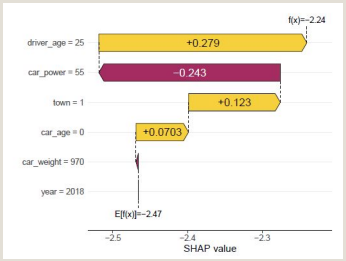
$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S))$$

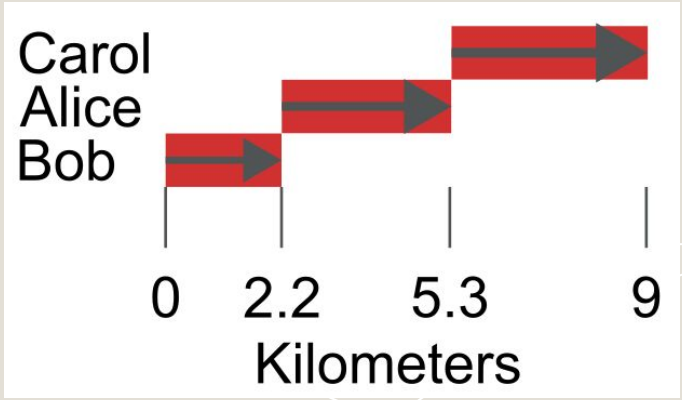
All of the contributions together ...

$\varphi_A(v) = \frac{2}{3} + \frac{1}{2} + \frac{1}{3} + 1\frac{2}{3} = 3\frac{1}{6}$

$\varphi_B(v) = \frac{1}{3} + \frac{1}{3} + \frac{1}{6} + 1\frac{1}{3} = 2\frac{1}{6}$

$\varphi_C(v) = \frac{2}{3} + \frac{1}{2} + \frac{1}{2} + 2 = 3\frac{2}{3}$





- N = [ABC], n = 3
- v([ABC]) = 9 km
- v([AB]) = 7 km
- v([AC]) = 8 km
- v([BC]) = 7 km
- v([A]) = 5 km
- v([B]) = 4 km
- v([C]) = 6 km
- v([0]) = 0 km

Extending to AI models

How much does each attribute's value in a given instance contribute to "rowing the kayak" towards the model's output for that instance?

- The task *v* ⇒ the model's task of producing an output for this instance.
  - Predict claim value based on {age = 30, height = 5' 6", weight = 150 lb, state = MD}
- The team *N* ⇒ all of the attribute values in this instance.
  - {age = 30, height = 5' 6", weight = 150 lb, state = MD}
- A sub-team S ⇒ a subset of the attribute values in this instance.
  - E.g., if we exclude height then {age = 30, weight = 150 lb, state = MD}
- The individual *i* ⇒ the specific attribute/value in this instance that we wish to compute the contribution of.
  - E.g., what is the effect of age = 30

$$\varphi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} (v(S \cup \{i\}) - v(S))$$

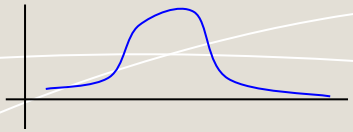
Extending to AI models

How do we figure out the model's output for the given instance if we only include a subset of the attribute values and exclude the rest?

$$(v(S \cup \{i\}) - v(S))$$

We marginalize out the excluded attributes.

- Query the model many times, with the real included attribute values and **every possible combination** of excluded attribute values.
  - {age = 30, height = x, weight = 150 lb, state = MD}
  - x = {1', 1'1", ..., 5'6", 5'7", ...}
- Take the average of all of the outputs, weighted by the probability of each combination of excluded attributes' values.
  (This could be estimated based on training set frequency.)



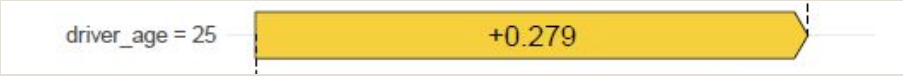
This is often intractable, implementations (Kernal SHAP, etc.) use various approximations.

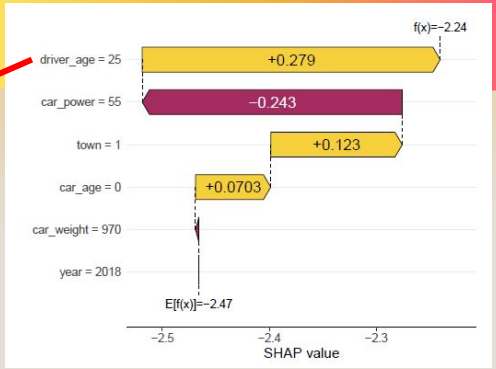
Extending to AI models

In abstract terms:

- We want to estimate how much a specific attribute value in a query instance contributed to the output of a model.
- Determine the difference in output of every possible subset of attribute values in this instance with and without the specific attribute.
    - Excluded attributes are marginalized out, i.e. taken to be the (weighted) average of the effect of all values of that attribute across the training data.
- Take the average of all of these differences across all of these subsets, weighted based on how many of the original instance's attribute values are included (and not marginalized out).

What is SHAP?





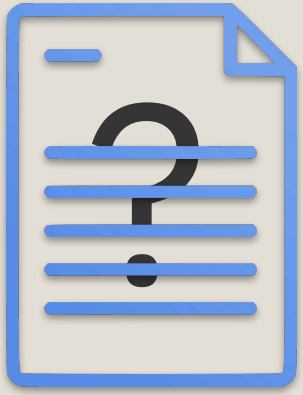
This bar means that for this specific instance:

- Across all possible hypothetical instances where driver_age is excluded (marginalized out) along with every subset of other attributes,
  - Where hypothetical instances with more original instance attribute values are weighted higher,
    - Adding driver_age = 25 added an average of 0.279 to the output of the model.

So is this a good explanation?

Back to the analogy with building models.

- How do you match application requirements to model capabilities?
- How do you decide which modelling technique (e.g., GLMs) to use?
- How do you decide on the appropriate distance/quality metric?
- How do you decide if data transformations are necessary?
- How do you decide when the model is and isn't applicable?
- How do you manage risks associated with the modelling technique's assumptions?
- ...



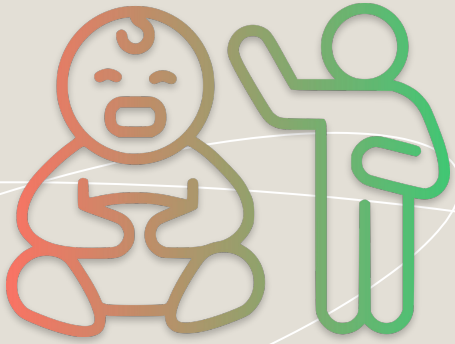
Equivalent questions should be asked when choosing your explanation technique.

For example ...

SHAP assumes attributes are independent.

- Consider a model that includes height and age, and we are calculating the contribution of age to the output for a 1-year-old.
- As part of marginalizing out different attribute values, SHAP will query the model with hypothetical instances of 1-year-olds at **every** value of height observed in training. $(v(S \cup \{i\}) - v(S))$
- SHAP will weight the model's response to a hypothetical 6' tall 1-year-old **more** than a 2' tall 1-year-old.
    - The model clearly has no data on 6' tall 1-year-olds so the extrapolation it returns is likely to be random.

Is the output likely to be sufficiently worthy of trust for your application?



SHAP Explanation Capabilities

Source

Post-Hoc Rationalization

Provides an explanation without claiming to represent the underlying reason.

It is not opening the box. It's a way of poking the box and coming up with a theory for what's in it.

Depth

Attribute Value

Describes not only which attributes were used, but also how the values were used.

Scope

Justification

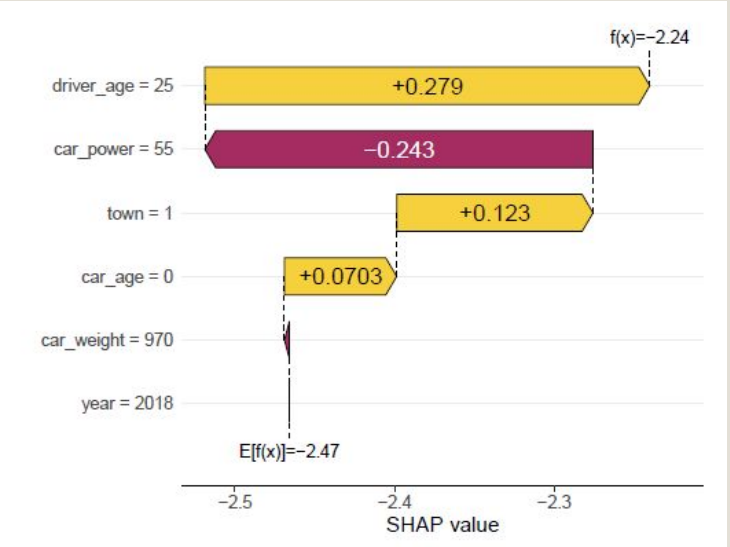
Justifies a single output.

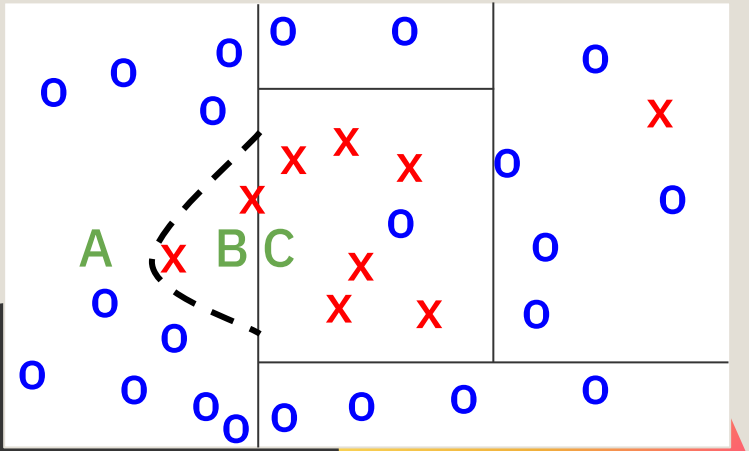
These are SHAP's capabilities.
Do they match your application's requirements?

Interpreting SHAP values

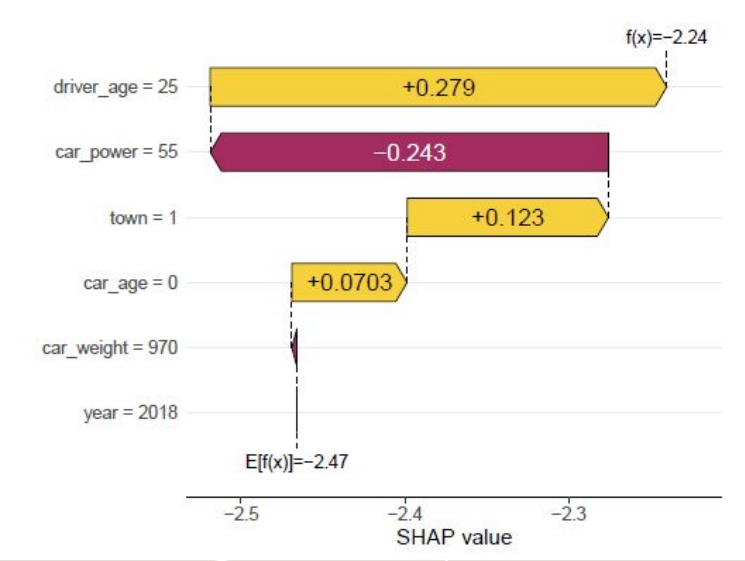
Waterfall charts

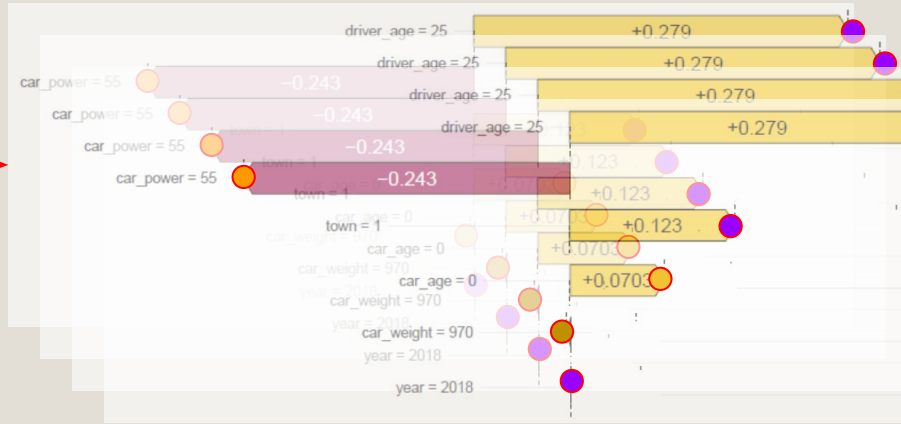
- Intuitively shows contributions of this instance's values.
- Only valid for this single instance.
- Only unique given the aforementioned assumptions.
- Doesn't tell you about what happens in response to a small change.
  - Consider how the waterfall chart might change between points A, B, and C. ⇒
  - Especially consider that this model *might* have gotten B wrong …





Interpreting SHAP values

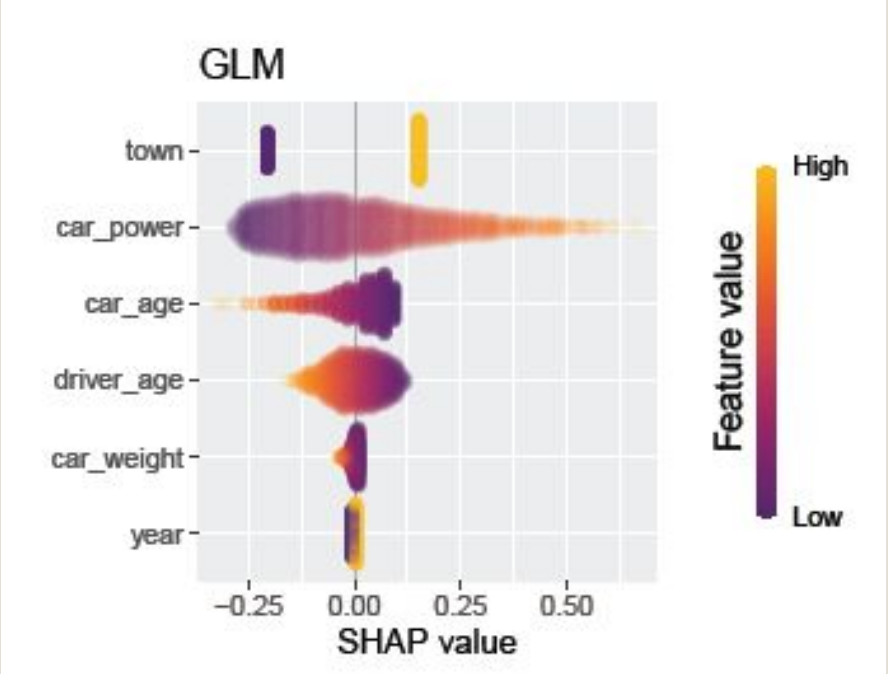




Interpreting SHAP values

Beeswarm plots

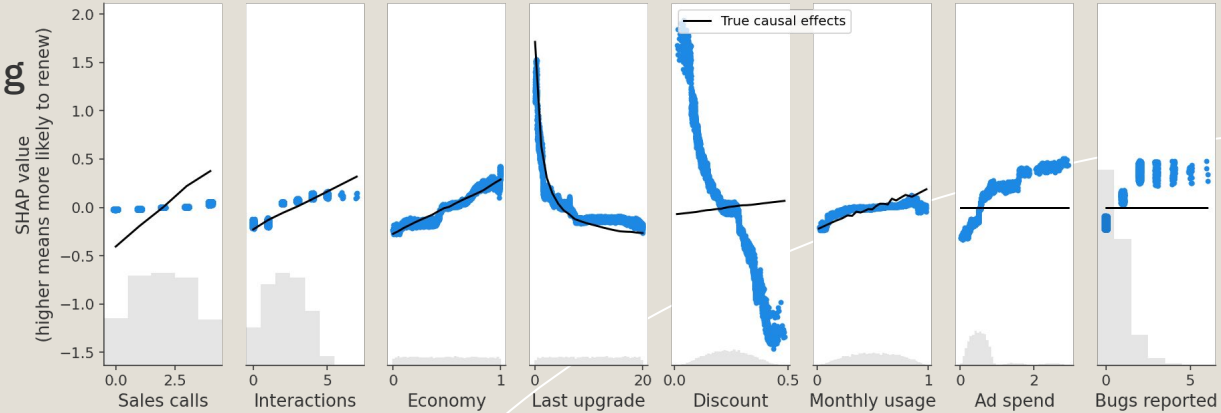
- A sample of instances.
- Like a bunch of waterfall plots with the zeros lined up.
- Shows density and trend.
- Does not show any mutual dependence.
    - Still hard to generalize.



Interpreting SHAP values

Scatterplots

- Same information as a beeswarm plot but turned sideways and with shift rather than color.
- Can still show density as a histogram (on either axis).
    - Depending on the purpose, showing it on the attribute (rather than SHAP) axis might be more informative.
- May be better at highlighting causality problems.



From https://shap.readthedocs.io/en/latest/example_notebooks/overviews/Be%20careful%20when%20interpreting%20predictive%20models%20in%20search%20of%20causal%20insights.html

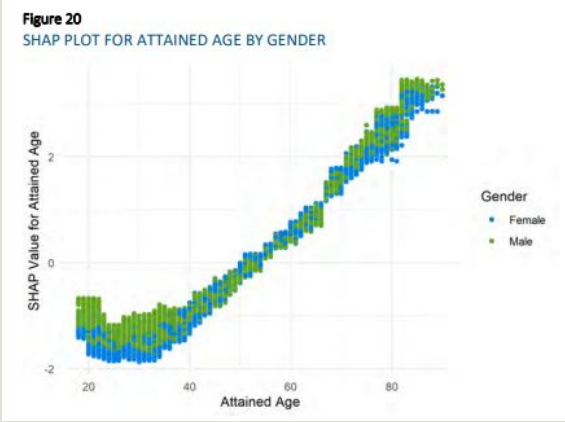
Interpreting SHAP values

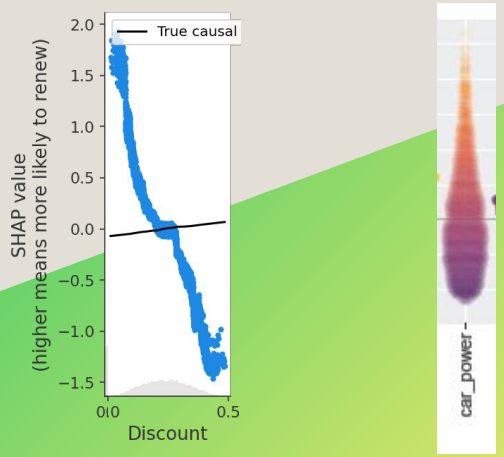
Scatterplot colored by a 2nd attribute.

Or ...

(Sideways) Beeswarm plot with a 2nd attribute shifted right by value.

- Shift and color helps to indicate some mutual dependence.
  - Need many plots to get a full picture.
- Could highlight when independence is (severely) violated.
  - But this also breaks SHAP. Can we trust the resulting SHAP values?





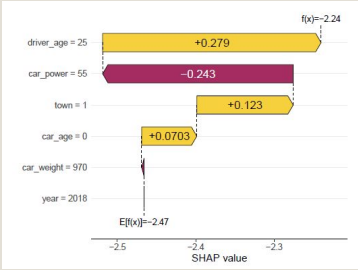
From https://www.soa.org/globalassets/assets/files/resources/research-report/2021/interpretable-machine-learning.pdf

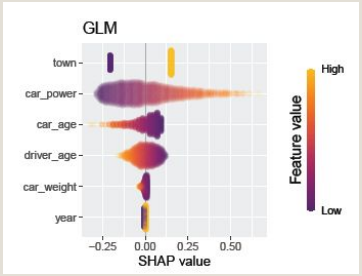
So is this useful?

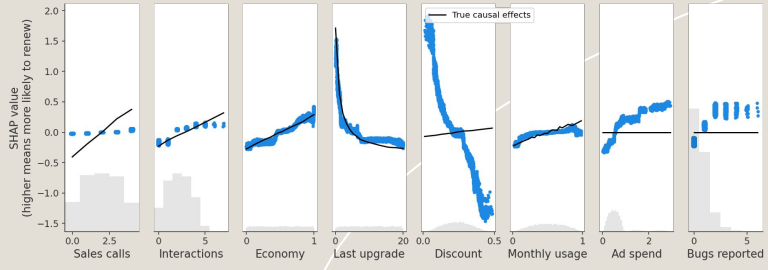
- Could you define what a “good” or “bad” waterfall, beeswarm, or (colored) scatterplot looks like?
- Could a change or violation in any of those assumptions change one from “good” to “bad”?
  - Mutual independence (not just correlation!) is a big one.

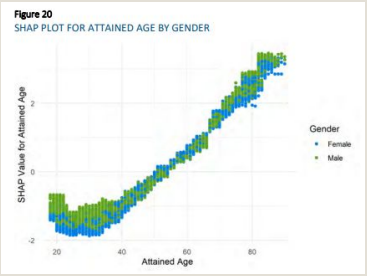
More fundamentally, do the capabilities of SHAP satisfy your requirements?

If I gave you one of these and told you “This came from a GLM” … and after you gave your opinion I said “Oops sorry, I got mixed up, this came from a memory based system” … would your opinion change?





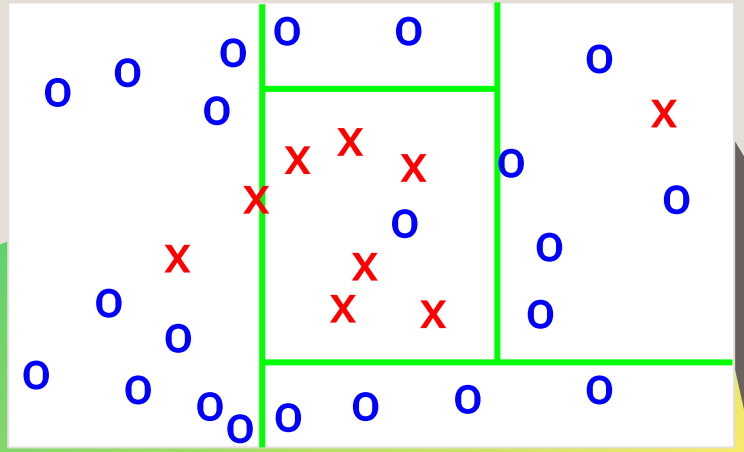




Examples of possible alternatives ...

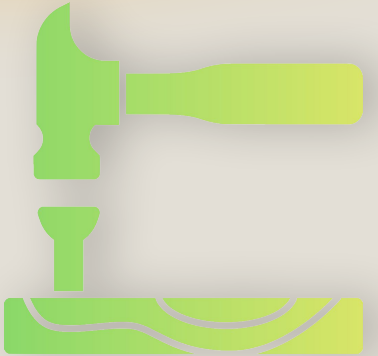
Assuming our goal is to understand the population-wide behavior of the model:

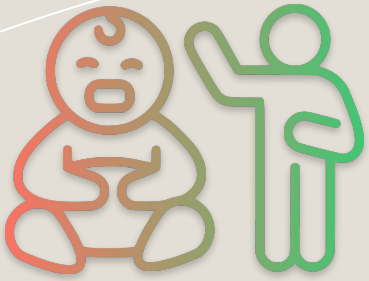
- Variants of SHAP that avoid unrealistic hypotheticals when marginalizing.
  - Only using real instances.
  - Changing the weighting probabilities during marginalization to reflect plausibility of the hypothetical instances (how?).
- Modelling techniques that are inherently interpretable.
  - Introspective explanations!
  - Maybe you don't need an opaque model.
- Transform the data to remove mutual dependence (including correlations).
- Only consider really high SHAP values.



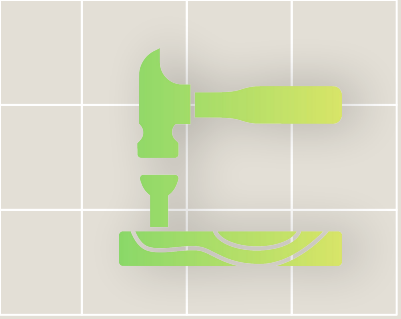
Closing thoughts ...

- SHAP is just another tool, with some capabilities and issues.
- In an ideal world, we would have many tools with different, well understood capabilities and choose based on requirements.
- We are still developing the “toolbox” for AI and understanding the capabilities is often an open research challenge.
- We can find analogies with “toolboxes” that are better understood and try and adopt some of the same risk management strategies.
- If the right tool doesn’t exist, be sure that an alternative tool really is “close enough” and that risks are managed accordingly.
    - There are situations when it’s OK to hammer in a screw.
    - I can’t think of any situation where it’s OK to screw in a nail.





Watch out for 6’ tall babies.

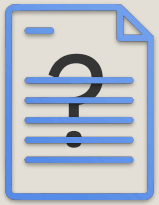


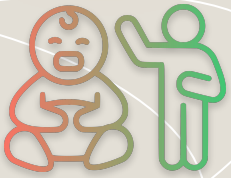
SHAP and the Reason for Explanation

Thank you for listening!

Questions welcome! Do get in touch to discuss further.

Raymond Sheh | ray@raymondsheh.org





Further reading ...

- Be careful when interpreting predictive models in search of causal insights (Dillon, E. et. al. 2018)
- Defining Explainable AI for Requirements Analysis (Sheh, R., Monteath, I. 2018)
- Explaining black box decisions by Shapley cohort refinement (Mase, M., Owen, A., Seiler, B. 2020)
- Explaining individual predictions when features are dependent: More accurate approximations to Shapley values (Aas, K., Jullum, M., Loland, A. 2021)
- Interpretable Machine Learning for Insurance (Baeder, L, Brinkmann, P., Xu, E. 2021)
- Using SHAP for explainability – Understand these Limitations First!! (Durgia, C. 2021)
- The NIST AI Resource Center (NIST 2026)